RECOGNITION OF MORE THAN ONE POSSIBLE TREND IN TIME-SERIES: REDESIGNED SCREENING OF MICRODATA IN THE CURRENT EMPLOYMENT SURVEY

Sheila McConnell, William Goodman, Bureau of Labor Statistics, U.S. Department of Labor
William Goodman, 2 Massachusetts Ave., N.E., Suite 4860, Washington, D.C. 20212

Introduction

In June of 1995, The Bureau of Labor Statistics (BLS) announced plans to redesign the method of selecting establishments to be included in the sample of the Current Employment Statistics (CES) survey. The newly designed sample is random, unlike the current sample. Partially in response to the recent sample design, new methods to check the quality of reported data are being developed. The new data editing and screening system will include basic checks for internal consistency of records and will also check data against a variety of plausible patterns in reported data over time. In the latter type of test, an establishment's current reported values are compared to earlier data from the same establishment. The current values are screened using several tests. Passing just one of the tests results in acceptance of the data. This paper presents the editing and screening methods currently being tested and their performance.

Nature of the survey

The CES survey is conducted in cooperation with State employment security agencies in every State and the District of Columbia. Each month, data representing employment, payroll dollars, employee hours, female employees, and production or nonsupervisory workers of approximately 375,000 business establishments are collected. These data are used to derive estimates of employment, hours, and earnings for the nation, the states, and selected large metropolitan areas. The CES survey provides quick release of data and a large amount of industry and geographic detail.

The employment statistics estimated at the national level include all employees, production or nonsupervisory workers, and women workers. Average weekly hours and average hourly earnings of production or nonsupervisory workers are estimated also. Furthermore, in the manufacturing sector, estimates of average overtime hours worked by production workers at premium pay are calculated.

New CES sample design

The Business Establishment List (BEL) is the universe or sample frame for the CES survey. The sample frame is constructed from the quarterly contribution reports filed by employers for each unemployment insurance (UI) account. The data are collected by the state unemployment insurance agency and refined by the State Employment Security Agencies (SESAs) in cooperation with the BLS' Covered Employment and Wages (ES-202) program. The BEL is one of the outputs of the ES-202 program. The BEL contains records of approximately 7,000,000 establishments (generally consisting of individual work sites), providing a count of 98 percent of the employees on nonfarm payrolls. In addition to employment, the records on this database also contain SIC code, state, and county code information broken down at the work site level for each UI account. The first-quarter file of the current year is available for sampling purposes approximately 12 months later. For example, establishment data for January 1999 will be available in December 1999 or January 2000.

The sampling unit is the unemployment insurance (UI) account. UI account numbers are unique within a state. The UI account is broken down to the work site or establishment level. A work site is a single physical location where business is conducted or where services or industrial operations are performed. A UI account can have more than one work site. Employers with more than one work site account for about 3.6 percent of all UI accounts, 16 percent of all work sites, and over 35 percent of total private employment.

In the new sampling and estimating system, each sample member will have a sampling weight associated with its probability of selection. Since the UI account is the unit being sampled, all of the work sites within a UI account will receive the same sample weight. The sample weight is defined as the inverse of the probability of selection. One would expect the reported employment of the unit in the ongoing survey to be reasonably similar to the employment at the time the unit was selected. If not, the reported data may cause distortions in the estimates of employment and other fields. For example, a small firm will have a low probability of selection; thus, its selection weight will be high. If the establishment’s employment is reported as dramatically different from its employment at the time the establishment was selected, the month-to-month change in estimates of employment in industries, states, metropolitan areas, or even the entire nation will be affected. A large change in an establishment’s employment may be valid; but before the respondent’s data are used in estimation, the reported employment must be validated. The screening procedures check whether the respondent’s employment is appropriate for its selection weight.

The current system and its shortcomings

A description of the older system of editing and screening, the one presently in use, will be useful toward understanding some of the decisions made in designing the new system, as the new system was intended not only to fit the new sample design but also to eliminate some problems in the old screening system. The most important problem in the current system is that most of the data identified as outliers are adjudged by analysts to be in fact accurate. Reviewing such falsely rejected or unnecessarily questioned data takes up considerable staff time. The new design is intended to recognize more of the data as accurate.

Editing and screening

There are two broad categories of possible errors: "edit errors" and "screening errors." "Edit errors" are the more obvious and definite errors, such as non-numeric data in a numeric field or contradictory records, such as a report showing all employees as fewer than women employees. While edit errors include a wide variety of definite or gross errors, "screening" comparisons have to do with the relationship between present data and corresponding preceding data. "Screening errors," in the old system, are identified primarily on the basis of comparisons to prior-month data from the same establishment; current data normally should be reasonably comparable to preceding values, with certain exceptions such as strikes or natural disasters.

The old screening design is based on an expected value for each current-month datum and ranges centered around the expected value. The expected value is normally based on the preceding month's value adjusted for seasonality. That is, the seasonal influence of the prior month (say, December) is removed and the seasonal influence of the current month (say, January) is put in. For example, in an industry like department stores, with a Christmas-time buildup and subsequent layoff, the expected value for January employment would be considerably less than the reported value in December. Seasonal factors developed for estimates are used to make the adjustment.

Designing new editing and screening

The new editing design is almost identical to the old one. The changes include, first, for contradictory data, the new edit check now rejects both of the items that conflict. For example, when the all-employee count is less then the women-worker count, both items are rejected. The old design would have rejected only the women-worker figure, retaining the all-employee datum. The old design's implicit philosophy was to retain as much valid data as possible. The change of policy is in anticipation of rapid communication with Data Collection Centers in the field; they should be able to respond to queries quickly.

The old design, in its calculation of an expected value, uses as a model only one pattern of values over time in microdata, a seasonal pattern. The expected value differs from the previous-month value only through seasonal variations. The seasonal variations recognized are those of the universe to which the respondent belongs. Seasonal factors for estimates of that universe are used to calculate the expected value from the previous-month value.

Years of experience in reviewing microdata have convinced the analysts in charge that several possible trends in reported data can be valid. For example, some reports show seasonal movements that are different from those of the respective universe. The miscellaneous amusements and recreation industry is a good example. In the same month of the year when some recreational establishments are being reduced in employment or shut down entirely for the winter, ski resorts may be expanding for their busy season. Thus an individual establishment can have an annual movement that is opposite that of its universe. Also, establishments that show little or no change from one month to another, despite a general large seasonal shift in the industry, are generally considered to be reporting accurately. Still more patterns of movement are also accepted.

Early in the project to design new screening, the team in charge decided that the new design should avoid the rejection of so much valid data by more closely imitating the decision process that analysts use to accept or reject data. The new system, then, tests changes in data against several models of movement by sequentially applying a number of tests. Passing any one of the tests by approximately matching any one of the models results in acceptance of the data.

The following patterns in microdata were mentioned in discussions toward the design:

- No movement except small random fluctuation over a one-month span
- No movement except small random fluctuation over a two-month span (The intervening month may have been anomalous.)
- About the same change as the datum showed a year earlier
- About the same change as *approximately* a year ago:
  - Eleven months ago
  - Thirteen months ago
- Seasonal change typical of the industry (a change proportionate to the seasonal change of the industry as a whole)

- Fluctuation typical of those shown by the datum throughout the past 6 to 12 months (A history of variability justifies a larger current change.)
- Fluctuation *over a 12-month span* in line with fluctuations throughout the latest 6 to 12 months (a history of variability combined with a seasonal pattern)
- Large change accompanied by appropriate comment code (for strike, shutdown, natural disaster, etc.)

Each of the situations just mentioned corresponds to a test in the completed design.

Amount of tolerance

After discussing the various models of movement, we can now turn to the amount of tolerance to allow around the model movement: How closely must a reported change approximate a model change to be acceptable? This question brings up the underlying further question of whether the change should be regarded as an absolute quantity (for example, a gain of 10 employees) or as a percentage (in an establishment of 100 employees, a gain of 10 more is 10 percent). In general, it seems reasonable that as the value of a datum increases, its maximum plausible absolute change tends to increase. For example, if an establishment employs 50 people, it is not ordinarily likely to hire 50 more (100 percent) in one month if no large seasonal change is anticipated. But if the establishment employs 2,000 people, it might more plausibly gain 50 more (2.5 percent) over a one-month span. Because the previous-month values of the various data to be tested will always vary, one can provide more appropriate tolerances for the differing levels by providing tolerances in percentage form.

A percentage tolerance, however, does not necessarily fit all situations. If the level of a datum is low enough, an extremely small change can be entirely plausible yet be a high percentage. For example, an establishment that has exactly one employee one month can easily gain two more. The percentage increase, however, would be 200 percent, far too high a tolerance to use for larger establishments.

Our solution was to use both a percentage tolerance and a minimum absolute tolerance. The absolute tolerance is to accommodate low levels of previous data, such as one employee in an establishment. The percentage tolerance allows larger, but proportionately reasonable, changes in larger establishments. Most of the tests, then, have two versions, one using an absolute tolerance, and one using a percentage tolerance. For example, dealing with the model of only small, random month-to-month changes, test 1 provides that

$$\text{if } |D_c/D_{c-1} - 1| < X$$

where D is the datum being tested, c is the current month, c-1 is the preceding month, and X is a percentage tolerance ,

then D_c is accepted.

As provided by the absolute-tolerance version of the test,

$$\text{if } D_{c-1}-K < D_c < D_{c-1}+K$$

where K is a constant,

then D_c is accepted.

A datum D_c that passes either version is accepted.

The exact values of both the absolute tolerances and the percentage tolerances were set through a consensus of experienced analysts. The group addressed the issues of what tolerances seemed appropriate for each test in each detailed industry, based on their experience in reviewing microdata. The factors and constants were selected, then, on a somewhat subjective basis.

Special screening tests for specific fields

In addition to the various screening tests already discussed, the team also included in its design a few tests for specific fields in certain situations. One such test recognizes the fact that the change in production or nonsupervisory workers is, in general, approximately equal to the change in all employees. The test, therefore, allows production or nonsupervisory workers to change approximately as much as all employees. The test calculates the number of non-production or supervisory workers by subtracting production or nonsupervisory workers from all employees, resulting in the quantity NPW. Then the test looks for consistency of the NPW number from month to month, allowing variation up to a certain small limit. Little change in NPW implies that the change in all employees is approximately equal to the change in production or nonsupervisory workers. If so, the production- or nonsupervisory-worker datum is accepted. Two versions of the test are applied. One uses a small constant as the limit of change in NPW. The other version uses a percentage of previous-month NPW as the limit of change in NPW.

Another test allows average weekly hours to vary more than usual if employment changes considerably. Newly hired employees may have hours not typical of the existing staff; for example, sometimes large numbers of part-time employees are hired. Conversely, laid-off employees may have hours quite different from the overall average hours in the establishment. Either situation would change average hours in the

establishment. Therefore, when employment changes substantially, one screening test allows hours to change to a greater extent than is otherwise allowed.

A similar test allows average hourly earnings to change more when employment changes substantially. New employees or laid-off employees may be paid at a rate quite different from the average for the entire establishment. Therefore large numbers of hires or layoffs can change the average earnings in the establishment. If employment changes by a large enough percentage, a change of up to a certain constant number of dollars is permitted.

Empirical results: testing new screening

In mid-1998, the new screening system was tested with 24,000 reports from the wholesale trade industry. The test entailed running the new screening system and having analysts examine the rejected reports, plus the most extreme changes among the data the system accepted, to determine the extent to which the new system's acceptance or rejection of data was accurate. Some 0.4 percent of the data (366 items) failed screening; as expected, the rate of rejection was reduced from that of the old screening system, which rejected over twice as much, 1.0 percent, of a separate group of reports for another reference month. (A test of both systems using the same data would have been preferable but was not arranged.) Of the 366 data rejected by the new system, 216 or 59 percent were found to be correct by the analysts. Although over half of the rejections were inappropriate, the new system improved on the old one; 87 percent of the data rejected by the old system was considered correct by the analyst in charge of wholesale trade. (See table 1.) The new system also greatly reduced unnecessary review. While the old system failed 1,028 data that therefore had to be examined, the new screening failed 366. The new system was also slightly superior in uncovering genuinely erroneous data (to be exact, data judged to be erroneous by the analysts). By means of the new system, 150 bad data were found; the old screening captured 134 real errors according to analyst opinion. Analysts assigned to various industries will evaluate the new system and procedures for further improvements.

The data accepted by the new screening system was sorted by month-to-month change so that the most extreme changes in accepted data could be checked for errors. Some 55 accepted data (0.06 percent) were considered erroneous by analysts. An exact corresponding statistic for the old screening was not available, but partial records showed that at least 0.06 percent of data accepted by the old screening was found to be in error. In regard to false acceptances, the new system was at least no worse than the old one.

Tests to see if selected unit is the one reporting

As noted, the new survey design employs a probability sample. Each sample unit has a sampling weight that reflects its probability of selection. It is essential that the employment reported to the CES survey is within an acceptable range of its employment when selected as a sample member.

The first month a respondent reports to the CES survey, the BEL employment history is used to validate the current-month employment to ensure that severely changing employment is identified for examination. Changes in the data are tested using several screening tests. Passing one of the tests by approximately matching any one of the models results in acceptance of the data. The first month a respondent reports to CES, the data is compared to the most current month available on the BEL. In addition, a t-test that takes into account a unit's variability is used.

Simple modifications were made to the screening tests. For example, test 2, which identifies reasonable movements over a 12-month span, is modified as:

if $|D_c/H_{12} - 1| < X$,
then accept D_c.

H_{12} is the most recent month available on the BEL database. There is a 12-month lag between the BEL employment data and the current month in the CES survey. Failing this test means that the respondent's reported employment is significantly different from what it was a year ago.

The wholesale trade industry was the first industry converted using the new probability-based procedures. Using the new screening tests, 13,000 reports were screened and reviewed for the November 1998 reference month. BEL data was used to verify the reported employment. The original employment factors and constants selected by the industry analysts were replaced with a fixed 15-percentage tolerance factor and a five-employee absolute tolerance constant. Some 6.3 percent (817 reports) of wholesale trade sample employment was rejected. An analyst reviewed the rejected reports to determine the types of inconsistencies that existed between the CES and the BEL. In some instances, the employment change is valid and can be attributed to a business trend experienced by the respondent since the UI account was selected. However, in most instances, the respondent may be reporting for a unit that was not initially selected to be in the sample. Of the 817 units rejected, 628 units were found to have important differences between the CES and the historical data taken from the BEL. The differences are attributed to the following errors:

- The respondent is reporting for more than one UI account.
- The respondent is reporting for a different UI account.
- The respondent is reporting for the correct UI account, but is including work sites that are located outside the state sampled.
- For a multi-establishment UI account, the respondent may not have been reporting for all the work sites as expected.

In these cases, all efforts are made to have the respondent report the UI account as originally sampled. The respondent must be re-contacted to resolve discrepancies between the employment reported to the CES survey and the employment reported to the UI program.

The remaining 189 units also reported differences between the CES and the historical data on the BEL but these differences were not significant and were not identified as a respondent failing to report for the sampled UI account. These respondents may have experienced employment growth or layoffs since the time they were selected for the CES sample. Although the respondent is reporting for the correct UI account, its reported employment must be verified before it can be used in estimation.

Once a unit has passed the check against the BEL, it is screened only against its CES reported history. With appropriate screening parameters, large changes in employment will be identified. Distortions in the estimates due to inappropriate combinations of reported employment and sample weight should thus be identified.

The other data items screened are: women workers, production or nonsupervisory workers, average weekly hours, average hourly earnings, and average overtime hours in manufacturing. The BEL provides total employment and wage data. It does not provide data for the other fields produced by CES, and the total wage data does not meet the CES definition. Fields other than all employees can only be validated using the establishment's CES reported history.

Absent previous values

Because screening is by definition a comparison over time (testing a field by comparing the current value to an earlier one), when no earlier corresponding value is available for comparison, screening in the conventional sense cannot be performed. For employment, the BEL database is available for comparison; however, the other data items are not available on that database. The team responsible for designing new screening decided that even data with no preceding values should be checked in some way for unlikely levels. In the absence of preceding values from the establishment, the mean of the field among establishments with the same SIC and fine size class (a classification according to the number of employees) is substituted. Because considerable variation among the establishments within an SIC and fine size class is expected, this test is relatively permissive. It allows the tested datum to vary by up to ± 1.96 within-group standard deviations from the group mean. This particular screening check, however, has not yet been tested.

Plan for automatic adjustment of ranges

Although ranges of accepted or rejected values in the new screening were originally set manually, using an informed but somewhat subjective approach, a proposal and program specification have been written to automate the adjustment of the ranges. The program has not yet been coded, but it should adjust the acceptance and rejection ranges, using as a basis a file of analysts' past decisions to reject data the system accepted and to accept data the system rejected. The program is to try expanding a range to include each datum accepted by the analyst and to try reducing a range to exclude data rejected by an analyst and thus eventually optimize the combination of automatic rejections and automatic acceptances. After each trial change corresponding to an analyst's action, the overall effects of the change on acceptance or rejection of all relevant data are to be automatically assessed. The anticipated automatic improvement of ranges should minimize the future changes that analysts would have to make to the status of microdata as accepted or rejected.

Summary and conclusions

The new editing and screening procedures eventually will be used by all monthly collection systems, national and state, within the program. Once implemented, the CES survey will have a consistent approach to microdata editing and screening.

The transition from a quota-based sample to a probability-based sample prompted new decision rules for the review of reported data. Editing and screening now takes on a greater role. The edit and screening system now must ensure that the CES sample members reflect the universe as each industry division is phased into a probability design.

A single model of movement over time in microdata from the Current Employment Survey is not adequate. At least several patterns of change over time in such data are considered plausible. If the results of automated screening are to agree with analysts' acceptance or rejection of microdata, the program must test data against several models of change. Therefore an adequate screening program for the purpose must include several tests and accept microdata that pass any one of those tests.

Table 1: Number of data rejected by screening but determined to have been correctly reported

Field	Old screening system		New screening system	
	Number	**Percent of rejects**	**Number**	**Percent of rejects**
All employees	114*	89*	38	79
Women employees	86*	91*	17	49
Production workers	127*	92*	14	48
Average weekly hours	N/A separately, 568* combined	N/A separately, 82%* combined	118	69
Average hourly earnings			29	35
All fields	**895**	**87**	**216**	**59**

*Estimate based on proportions in May 1998 first-run old-system screening.