Choosing Size Classes for Industry Employment Estimates by Firm-Size Class

October 2012

Jeffrey A. Groen[1], Lowell G. Mason[2]

[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Room 4945, Washington, DC 20212; Groen.Jeffrey@bls.gov.

[2]U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Room 4945, Washington, DC 20212; Mason.Lowell@bls.gov.

Abstract

The Bureau of Labor Statistics is conducting research on the feasibility of producing industry employment estimates by size of firm using the Current Employment Statistics (CES) survey, a monthly survey of business payrolls. The Office of Management and Budget has designated 12 size classes to serve as the basis for official estimates produced by federal statistical agencies, but a smaller number of categories must be used for CES estimates in order to reduce sampling error. This paper analyzes the problem of collapsing the 12 classes into a smaller number of categories for use in constructing CES estimates. We develop a methodology based on K-means cluster analysis modified to require that clusters retain the ordering of classes. We also require that each cluster has sufficient employment share and adequate sample in each industry. We use 4 characteristics of size classes in the cluster analysis: seasonality of employment, long-term trend in employment, cyclicality of employment, and share of employment related to business births and deaths. We also consider size categories used in related data and reclassifications of firms over time across size categories.

Key Words: Employment, firm size, classification, cluster analysis, K-means

1. Introduction

The Bureau of Labor Statistics (BLS) is conducting research on the feasibility of producing industry employment estimates by size of firm using the Current Employment Statistics (CES) survey, a monthly survey of business payrolls. There are many conceptual and methodological issues in moving towards publication of size-class estimates. One issue is how size class is defined. The Office of Management and Budget (OMB) has designated 12 size classes to serve as the basis for official estimates produced by federal statistical agencies, but a smaller number of categories must be used for CES estimates in order to reduce sampling error.

This paper analyzes the problem of collapsing the 12 classes into a smaller number of categories for use in constructing CES estimates. The next section of the paper describes the OMB standards for business size and summarizes the size categories used in related data. Section 3 provides background on the CES and presents our methodology for collapsing the OMB size classes using K-means cluster analysis. We apply our method using historical data from the Quarterly Census of Employment and Wages (QCEW). Section 4 presents the results, and Section 5 discusses some implications of the results for the development of size-class estimates from the CES.

2. Business-Size Classes: Guidelines and Uses

2.1 OMB Standard

In 1982, OMB established a standard on size classes used by federal agencies to create statistics on business size (OMB 1982). The purpose was “to provide a standard means of comparing various business size series prepared by various Federal agencies” (OMB 1982, p. 21362). The standard was endorsed in the President’s 1982 report on small business (President of the United States 1982). This was the U.S. government’s first standardization of business-size statistical data, and it remains in force.

The standard provides size categories to be used to classify businesses by employment, revenues, or assets. Twelve classes are to be used for employment, based on the number of employees: 1–4, 5–9, 10–19, 20–49, 50–99, 100–249, 250–499, 500–999, 1,000–2,499, 2,500–4,999, 5,000–9,999, and 10,000 or more. The standard provides agencies discretion to combine adjacent size classes because of “the limited scope of the data, the need to assure the confidentiality of individual response, or very large sampling variability at the recommended level of detail” (OMB 1982, p. 21362). Agencies are also allowed to subdivide the categories as long as detail adds to totals and sensible breaks are used.

The standard applies to employment statistics by business size regardless of the statistical unit used for businesses. In practice, data producers use a variety of ways to define businesses. At one extreme is an establishment, which is an economic unit (such as a factory or store) that produces goods or services at a single location and is engaged predominantly in a single type of economic activity. At the other extreme is a firm, which contains all establishments operating under common ownership and control. In between these extremes are units defined for tax purposes by identifiers such as state Unemployment Insurance (UI) account numbers and federal Employer Identification Numbers (EINs). Such units may contain multiple establishments but they often do not comprise entire firms, because individual firms may have multiple UI accounts (even within a given state) or multiple EINs. Henceforth, we use “businesses” as a generic term and we indicate for a specific context whether statistical units are defined by establishments, firms, EINs, or UI accounts.

2.2 Size Categories Used by Others

One consideration in selecting size categories for CES estimates is the size categories used by other data producers. This is important because some users will compare CES estimates of employment by size class to similar estimates from other sources. Given this consideration, we summarize the size categories used by other data producers. The sources vary in the combinations of OMB classes used to define size.

The QCEW program at BLS produces estimates of employment for the first quarter (based on March data) of each year separately by establishment-size class and EIN-size class. The QCEW is a quarterly census of all U.S. business establishments subject to UI taxes, covering approximately 9 million establishments nationwide and 98 percent of U.S. wage and salary employment. Monthly employment data are collected using the quarterly contribution reports that employers submit to state agencies responsible for administering UI programs. For the size-based estimates, size is defined using 9 categories. The first 8 categories match the first 8 OMB classes, and the ninth category (employment of 1,000 or more) is the union of the top 4 OMB classes.

The Business Employment Dynamics (BED) program at BLS produces quarterly statistics on gross job flows (e.g., gross job gains and gross job losses). BED data are tabulated by linking QCEW establishment records across quarters to create a longitudinal history. Size-class estimates were initially constructed for EINs, and the methodology has recently been extended to establishments (Butani et al. 2006; Dalton et al. 2011). The BED program uses the same 9 size categories as the QCEW program.

The Census Bureau's Statistics of U.S. Businesses (SUSB) program produces an annual series that provides national and sub-national statistics on the distribution of economic activity by firm size and industry. These statistics cover most of the nation's economic activity and are derived from the Census Bureau's business register, which is based on a variety of administrative-record and survey sources. The SUSB program uses 12 categories to define size. They are not identical to the 12 OMB size classes, although there is substantial overlap. The first size category used in the SUSB is 1–19 employees (OMB sizes 1–3), and the second size category is 20–99 employees (OMB sizes 4 and 5). The next 6 SUSB categories are identical to OMB size classes 6–11, which together cover firms with 100–9,999 employees. For firms with 10,000 or more employees, the SUSB subdivides the largest OMB size class into 4 categories: 10,000–24,999, 25,000–49,999, 50,000–99,000, and 100,000 or more.

Another Census Bureau program, Business Dynamics Statistics (BDS), also produces employment estimates by firm-size class. The BDS is similar in scope to the BED program at BLS in its focus on business dynamics (Haltiwanger, Jarmin, and Miranda 2009). The BDS series, which is based on the same source data as the SUSB, includes annual statistics on job creation and job destruction by firm size. The statistics published from the BDS classify firms into sizes using the 12 OMB size classes.

There are at least 2 private companies that produce employment statistics based on business size. The forecasting firm Macroeconomic Advisors developed the ADP Employment Report, which is based on a sample of roughly 500,000 U.S. business clients of the payroll firm Automatic Data Processing (ADP). Macroeconomic Advisors uses methodologies that are designed to be similar to those used by BLS in producing monthly employment estimates, and the ADP Employment Report for a given reference month is released to the public 2 days before BLS releases its first CES employment estimates for that month (Macroeconomic Advisers 2008). The ADP Employment Report includes estimates of total private nonfarm employment by broad industry sector (goods-producing and service-producing sectors) and size of payroll. Three categories are used for payroll size: 1–49 employees (corresponding to OMB sizes 1–4), 50–499 employees (OMB sizes 5–7), and 500 or more employees (OMB sizes 8–12). Payroll size appears to be a hybrid of establishment size and firm size because "in some cases small and medium-sized payrolls belong to businesses employing more workers than indicated by the size grouping" (Automatic Data Processing 2012).

The payroll firm Intuit produces the Intuit Small Business Employment Index. The index is based on a sample of about 70,000 Intuit clients, each with fewer than 20 employees (Intuit 2012). This range matches the combined range of the first 3 OMB size classes. One of the inputs to the Intuit index is national employment estimates from CES. The reporting unit used in the Intuit database (as with the ADP database) is likely to be somewhere between an establishment and a firm because the reporting unit is determined by clients.

The Small Business Administration has established size standards that define whether a business is “small” and thus eligible for federal government programs and preferences reserved for small businesses. A size standard is the maximum size that a firm may be to qualify for federal programs that provide a benefit to small businesses. In most cases, size standards are based on the average annual receipts or the average employment of a firm (Small Business Administration 2010). Size standards vary by industry. Of the industries for which size is determined by employment, the most common maximum is 500, which covers 64 percent of the cases. The other common ones are 100 (12 percent), 750 (11 percent), and 1,000 (12 percent). The thresholds of 100, 500, and 1,000 are consistent with OMB size classes, but the threshold of 750 is not.

3. Methodology and Data

3.1 Background on CES

The CES survey is a large-scale, nationwide survey of nonagricultural establishments. Each month the CES program surveys about 141,000 businesses and government agencies, representing about 486,000 individual worksites, in order to provide detailed industry data on employment, hours, and earnings of workers on nonfarm payrolls. The private sector is covered by a probability sample, which is a random sample of UI account numbers (clusters of worksites) within strata. Sample strata are defined by state, industry, and business size. Size is based on the number of employees in the UI account, specifically the maximum monthly employment over the previous 12 months. The government sector is covered by a quota sample. The federal-government sample is a virtual census; the state-government sample covers about 80 percent of employment; and the local-government sample covers about 50 percent of employment.

Sampling rates for each stratum are determined through optimal allocation, which distributes a fixed number of sample units across the set of strata in a state to minimize the variance of the estimate of total employment in the state. The sampling frame is the QCEW, which is a quarterly census of all U.S. business establishments subject to UI taxes. The frame and the CES sample are updated once a year. Additionally, the CES sample is supplemented in mid-year with a sample of births (i.e., new businesses).

Employment, the key data item in CES, is defined as the number of employees who worked or received pay during the pay period that includes the 12th day of the month. Once a year, the CES obtains a total employment figure (or “benchmark”) as of March for each estimating cell from universe employment counts derived mainly from the QCEW. The CES estimate for March is determined by this benchmark, and employment estimates for subsequent months are computed using a ratio derived from CES respondents who provided data in both the current month and the previous month. The numerator of this ratio is the weighted employment in the current month for all such respondents in the estimating cell, and the denominator is the weighted employment in the previous month for these respondents. The employment estimate for a given month is computed by multiplying this ratio by the estimate for the prior month and then adding an estimate of residual net birth-death employment (Mueller 2006). Seasonally adjusted series are published monthly for select employment series.

The CES program is conducting research on the feasibility of producing monthly employment estimates by business size within an industrial sector. Employment estimates by business size would be developed for the nation as a whole and would cover all private nonfarm industries. Business size would be based on EINs, because EIN is the closest approximation of firm that is available for CES. Although UI accounts are used for sampling due to the state-based design, EINs are closer to firms because an EIN may include establishments in multiple states, whereas a UI account is specific to a state. EIN size would be determined when a given sample is drawn and based on maximum employment over the previous 12 months. Because the entire CES sample is updated annually, the size of a particular EIN in the sample would be allowed to change once a year. For EINs with establishments in multiple industries, size would be determined for the EIN as a whole but employment would be assigned to industries based on the industry designations of the EIN's establishments. The size categories chosen for CES are to be the same for all industries in order to facilitate comparisons across industries and to allow the construction of aggregate employment series.

3.2 Methodology

The problem is to collapse the 12 OMB size classes into a smaller number of categories for use in producing CES employment estimates by business size. In solving this problem, we are guided by 5 general objectives. First, we would like the chosen size categories to consist of size classes with similar characteristics (as defined below). Second, we wish to achieve balance in employment shares across the chosen size categories. Third, we would like to have adequate sample in each size category in order to reduce sampling error. Fourth, we would like to minimize the extent of reclassifications of businesses over time into different size categories. Fifth, we would like the CES size categories to be comparable to the size categories in related data.

Our method directly addresses the first 3 objectives and provides a key input for addressing the remaining 2 objectives. Our method is based on K-means cluster analysis, which breaks a set of observations into a distinct number (K) of nonoverlapping groups based on a vector of characteristics (e.g., Kaufman and Rousseeuw 1990). However, our problem is different from the usual problem solved by K-means because the clusters of size classes must be formed by adjacent size classes. We also depart from the usual K-means approach by defining constraints on the types of clusters that are allowed.

We measure the characteristics of each OMB size class using historical data from the QCEW. Following the construct to be used in the CES size-class estimates, we define businesses using EINs. We use the monthly employment data from 1992 to 2010 and aggregate the establishment-level data in the QCEW into EIN-level observations. This period is used because it is the longest period of QCEW micro data that is available to us for constructing the size-class characteristics. The size of an EIN in a particular year is based on its maximum employment over the 12 months of the previous year. This definition, which parallels the one used for UI size in CES sampling, is used because employment varies greatly over the course of the year in many industries but the seasonal pattern varies by industry.

For our constraints, we require data broken down by size class within industry. We measure industrial sectors using the 14 NAICS "supersectors" defined by the U.S. Economic Classification Policy Committee and used by the CES program in constructing estimates of private nonfarm employment. For EINs with establishments in multiple sectors, size is based on total employment across all sectors and employment is allocated to sectors based on the industry codes of the establishments that comprise the EIN.

3.2.1 Characteristics

We aggregate the EIN-level information into cells defined by the 12 OMB size classes and the 14 sectors. For defining characteristics, we further aggregate the sector detail into 12 series with private-sector totals by size class. For each cell, we have a time series of employment covering 19 years with 12 monthly observations per year. We use these time series to define 4 characteristics of a size class. For each characteristic, a case can be made that it is desirable to combine size classes with similar values of the characteristic when forming size categories.

We use a time-series decomposition to derive 3 of the 4 characteristics. We apply the same method separately to each of the 12 series. We use the SAS X12 procedure (an adaptation of the U.S. Census Bureau's X-12-ARIMA seasonal adjustment program) to decompose each series (O_t) into a seasonal component (S_t), a trend-cycle component (C_t), and an irregular component (I_t) such that $O_t = S_tC_tI_t$. The seasonal component is the monthly variation that is repeated constantly from year to year. The trend-cycle component includes variation due to the long-term trend and the business cycle.[1] The irregular component is the residual variation.

To allow us to construct separate characteristics for long-term trend and cyclicality, we further decompose the trend-cycle component into a trend component and a cycle component. To identify the trend, we estimate a linear regression with dependent variable C_t and independent variable t, where t is a continuous variable equal to 1 for the first month of the time series (January 1992), 2 for the second month (February 1992), ..., and 228 for the last month (December 2011). Denoting the estimated coefficient on t as $\hat{\beta}$, the trend component is defined as $\hat{C}_t = \hat{\beta}t$. The cycle component is then defined as $\tilde{C}_t = C_t/\hat{C}_t$. The complete decomposition is $O_t = S_t\hat{C}_t\tilde{C}_tI_t$. In terms of magnitude, S_t, $\tilde{C}_t$, and I_t are all in the neighborhood of 1, while $\hat{C}_t$ is of the same order of magnitude as O_t. Figure 1 provides an example of the decomposition, for size class 5.

Our first characteristic, the degree of seasonality in the time series, is measured as the standard deviation of S_t. A series with a high degree of seasonality will have some months with S_t much greater than 1 and some months with S_t much less than 1. We use seasonality as a characteristic in the cluster analysis because grouping size classes with similar seasonality can improve seasonal adjustment. Indeed, this is the reason that the CES program stratifies estimation cells by geographic region for the construction industry (Manning and Mullins 2006). In most industries, basic estimation cells for national estimates are defined by industry. Within the construction industry, estimation cells are further stratified by region because the seasonality of construction employment varies by region due to differences in weather. As a result, stratifying by region improves the fit of seasonal adjustment for construction.

The second characteristic, the long-term trend of employment from 1992 to 2010, is measured as $\hat{\beta}/\bar{C}t$, where $\bar{C}t$is the mean of C_t over the entire time period. This measure is similar to a percentage change because $\hat{\beta}$ captures the absolute change and $\bar{C}t$captures the level. The third characteristic, the degree of cyclicality in the time series, is measured as the standard deviation of $\tilde{C}_t$. The trend and cycle characteristics are desirable for

[1] There were 2 recessions over the 1992–2010 period: March 2001–November 2001 and December 2007–June 2009 (dates according to the National Bureau of Economic Research).

clustering because data users may examine time-series patterns in the CES estimates by size class.

The fourth characteristic is employment attributable to business births and deaths as a share of total employment in the EIN size class, i.e., (birth employment + death employment) / (total employment). Births and deaths are identified at the establishment level using QCEW information on the first and last quarter of non-zero employment. Birth employment is the monthly employment of establishments that existed in a particular year but did not exist in the previous year. Deaths are establishments that permanently cease operations at some point during a particular year, and death employment is based on average monthly employment in the prior year. We use this fourth characteristic in clustering because it captures the extent to which the birth-death model, as opposed to the sample of continuing establishments, is used to estimate employment change.

3.2.2 Constraints

For the cluster analysis, we wish to group the 12 OMB size classes ($i = 1, \dots, 12$) into K clusters. We consider 3 values of K: 2, 3, and 4. For a given K, we first determine all possible sets of K clusters in which each cluster contains at least 1 OMB class and clusters are formed by adjacent OMB classes. (An example of a set of 3 clusters is OMB classes 1–4 in cluster 1, classes 5–9 in cluster 2, and classes 10–12 in cluster 3.) Then we determine which of these sets of clusters satisfy 2 constraints, one based on employment shares and the other based on the number of reporting UI accounts. (The second constraint is based on UI accounts rather than EINs because CES selects its samples based on UI accounts.) Each constraint is applied separately for each sector.

The first constraint requires that in each sector the share of employment in each size cluster exceeds a threshold. This constraint ensures some balance across sectors in the employment shares. The threshold varies by the number of clusters: 20 percent for $K = 2$, 10 percent for $K = 3$, and 5 percent for $K = 4$. As explained in Section 4, we relax these constraints somewhat for particular sectors because the size distribution of employment is unusual in these sectors.

The second constraint, which is designed to control sampling error, requires that in each size cluster and sector the number of UI accounts that report CES data exceeds a threshold. This constraint is evaluated using data on the number of reporting UI accounts for each of the 12 months from April 2009 to March 2010. For most sectors, we require that the number of reporting UI accounts is at least 100 for each month. For the utilities sector, we use a lower threshold of 50 because the number of reporting UI accounts for many size clusters is much lower for utilities than for other industries. This is because (1) utilities represents a small share of total private employment (less than 1 percent) and (2) the distribution of employment in utilities is skewed towards the largest OMB size classes. The CES program generally requires a minimum of 50 reporting UI accounts for national employment estimates by industry.

3.2.3 Distance

To discriminate among the feasible sets of clusters (i.e., those satisfying both constraints), we use the values of the 4 characteristics for each of the OMB size classes (w_i, x_i, y_i, z_i). For each cluster ($k = 1, \dots, K$) in a feasible set of clusters, we compute the average value of each characteristic among the classes in the cluster. The vector of average values

$(\bar{w}_k, \bar{x}_k, \bar{y}_k, \bar{z}_k)$ is called the cluster center. We determine the distance of each class to the center of the cluster to which it belongs using Euclidean distance with equal weights: $\sqrt{(w_i - \bar{w}_k)^2 + (x_i - \bar{x}_k)^2 + (y_i \ - \bar{y}_k)^2 + (z_i - \bar{z}_k \ )^2}$. The distances for each of the 12 size classes are added together to create the measure of total distance. Prior to computing distance, we standardize the characteristics so that each has mean 0 and standard deviation 1.

Recall that our cluster analysis directly addresses 3 of our 5 objectives. We address the remaining 2 objectives using the feasible sets of clusters. The first of these objectives, having size categories for CES that are comparable to those in related data, is addressed by comparing each feasible set of clusters to the size categories in related data, relying on the summary of related data in Section 2.2. The other remaining objective, minimizing the extent of reclassifications over time, is addressed by computing, for each feasible set of clusters, the percent of EINs (or employment) that are classified in different size clusters in consecutive years.

We examine reclassifications of EINs over time using the QCEW data that is the basis for the cluster analysis. For measuring classification changes, EIN size in a given year is based on the maximum monthly employment over the 12 months of that year. From this information we construct a transition matrix showing the relationship between OMB size classes of EINs in consecutive years from 1992 to 2010. For a given feasible set of clusters, we aggregate the size categories in the transition matrix from OMB size classes to the clusters in the given grouping. Then we compute 2 percentages: the percent of EINs that changes clusters from year to year and the percent of employment that changes clusters from year to year. These percentages allow us to compare across feasible sets the extent of reclassification.

4. Results

4.1 Distributions by Size

The first column of Table 1 shows the distribution of total private employment by OMB size class over the 1992–2010 period. The highest size classes represent a relatively large share of employment: 17 percent of employment is at EINs with 10,000 or more employees, 38 percent at EINs with 1,000 or more, and 52 percent at EINs with 250 or more. At the other end of the size spectrum, 10 percent of employment is at EINs with less than 10 employees, 17 percent is at EINs with less than 20 employees, and 28 percent is at EINs with less than 50 employees.

The size distribution of employment varies across sectors. In some sectors, there is a relatively small share of employment in one tail of the size distribution. In construction, the largest EINs account for a relatively small share of total employment: EINs with 1,000 or more employees account for only 9 percent of employment. In utilities, the smallest EINs account for a relatively small share of total employment: EINs with less than 50 employees account for only 8 percent of employment. We account for the unusual distribution of employment in these sectors by using a lower threshold for the minimum employment share in a given cluster. For construction (higher size clusters) and utilities (lower size clusters), the thresholds are set at 12 percent for the case of 2 clusters ($K = 2$), 7.5 percent for $K = 3$, and 4 percent for $K = 4$.

The second column of Table 1 shows the distribution of EINs by size class for the private sector. Most EINs are small: 46 percent have less than 5 employees and 84 percent have less than 20 employees. The largest size classes contain a very small share of all EINs; for example, only 0.24 percent of EINs have 1,000 or more employees. The small size classes have a large share of EINs but a small share of employment whereas the large size classes have a small share of EINs but a large share of employment.

4.2 Characteristics of Size Classes

The final 4 columns of Table 1 contain the 4 characteristics used as inputs to the cluster analysis. The values are standardized so that each characteristic has mean 0 and standard deviation 1. The pattern of the characteristics by size class is shown in Figure 2.

Seasonality. Seasonality of employment is highest in classes 1–4 (together covering EINs with less than 50 employees) and lowest in classes 7–11 (250–9,999 employees).

Trend. The long-term trend in employment is generally increasing in size. Trend is lowest for classes 2–3 (5–19 employees) and highest for classes 11–12 (5,000 or more employees).

Cyclicality. Cyclicality of employment is increasing in size until class 8 (500–999 employees) and decreasing in size after class 10 (2,500–4,999 employees). Cyclicality is lowest for classes 1–2 (less than 10 employees) and highest for classes 8–10 (500–4,999 employees).

Birth-death. Employment attributable to business births and deaths as a share of total employment is highest for class 1 (1–4 employees) and falls monotonically over the first 7 classes before rising modestly for the last 2 size classes.

4.3 Feasible Sets of Clusters

Table 2 contains the results of our cluster analysis. The table shows the feasible sets of size clusters sorted by minimum distance, where feasible sets are those in which the employment-share and sample-adequacy constraints are met for each cluster and sector. For each feasible set, the table reports the range of OMB size classes (and the associated employment range) in each cluster, the total distance, and the percent of total private employment in each cluster. There are 3 feasible sets of 2 clusters, 5 feasible sets of 3 clusters, and 1 feasible set of 4 clusters. The number of feasible sets is low relative to the number of possible sets (11 for $K = 2$, 55 for $K = 3$, and 165 for $K = 4$). Even though the thresholds in our constraints are relatively low, the number of feasible sets is small because the constraints are applied to each sector, some sectors are relatively small, and the size distribution of employment varies across sectors.

Among the feasible sets of 2 clusters, the set with the smallest distance has size classes 1–5 (1–99 employees) in the first cluster and classes 6–12 (100 or more employees) in the second cluster. The first cluster comprises 37 percent of employment and the second cluster comprises the remaining 63 percent. Compared with this set of clusters, the other feasible sets of 2 clusters have larger first clusters. The set with the next largest distance has classes 1–6 (1–249 employees) in the first cluster, and the remaining set has classes 1–7 (1–499 employees) in the first cluster.

Among the feasible sets of 3 clusters, the set with the smallest distance has classes 1–4 (1–49 employees) in the first cluster, classes 5–6 (50–249 employees) in the second cluster, and classes 7–12 (250 or more employees) in the third cluster. The employment shares for these 3 clusters are 28 percent, 19 percent, and 52 percent. The remaining feasible sets of 3 clusters involve a break between the first and second clusters at 50 or 100 employees and a break between the second and third clusters at 250, 500, or 1,000.

There is only 1 feasible set of 4 clusters. It involves breaks between the clusters at 50, 250, and 1,000 employees. The employment shares are higher for the first (28 percent) and fourth (38 percent) clusters than for the second (19 percent) and third clusters (14 percent).

Table 2 also reports, for each feasible set of clusters, the percent of EINs and the percent of employment that change clusters from one year to the next. The rank order of the feasible sets of clusters is the same for both measures of reclassification. The percent of EINs is relatively small for each of the feasible sets of clusters, never exceeding 1.4 percent. By contrast, the percent of employment is larger, ranging from 2.2 percent to 6.2 percent. The transition percentages are larger for employment than for EINs because employment is more evenly distributed across size classes than are EINs. Measured using employment or EINs, the transition percentages are lower with 2 clusters than 3 clusters and lower with 3 clusters than 4 clusters. This is because with fewer clusters there are fewer boundaries between size clusters and therefore fewer reclassifications.

Next, we address whether each feasible set of clusters is compatible with size categories in related data. Because they are closely aligned with the OMB size classes, the size categories used in the QCEW, BED, and BDS are compatible with each of the feasible sets of clusters. That is, a data user could aggregate the categories used in the QCEW, BED, or BDS to match those in any feasible set of clusters from the CES. On the flip side of comparability, the Intuit data, which are based on clients with less than 20 employees, are not consistent with any of the feasible sets of clusters. This is because the lowest size cluster in each of the feasible sets covers employment up to 49 employees, or more.

There are 3 sources of related data in which the size categories are compatible with some but not all of the feasible sets of clusters in Table 2. Whether a given set of clusters is compatible with the SUSB data, SBA standards, or ADP data is indicated in the last 3 columns of the table. The SUSB data are compatible with each of the feasible sets of clusters except for sets of clusters that have a break between OMB size classes 4 and 5 (i.e., at 50 employees), because the SUSB combines these classes into a single category for 20–99 employees. The SBA standards, using the most common threshold of 500, are compatible with 3 of the feasible sets—those having the highest size cluster as 500 or more employees. The ADP data involve a particular set of 3 clusters (involving breaks and 50 and 500 employees), and this set of clusters is 1 of the feasible sets for CES.

This discussion of compatibility has assumed that producers of related data would not change their size categories in response to the choice of size categories for CES. This assumption seems appropriate for producers in the government, such as the Census Bureau and SBA. Alternatively, one could assume that a producer would change its size categories to match whatever BLS chooses for CES. This assumption is plausible for ADP because Macroeconomic Advisors attempts to follow BLS methodology in producing the ADP Employment Report (Macroeconomic Advisors 2008).

5. Discussion

The constraints used in our cluster analysis are very effective at restricting attention to a small number of sets of clusters. There are only 9 feasible sets of clusters, so it is relatively easy to compare sets. We provide several measures with which to compare sets: characteristic similarity (distance), employment shares, reclassification, and compatibility with related data.

In order to select a preferred set of clusters from the feasible sets, one must decide which measure (or set of measures) is most important. We leave these judgments to the CES program, but we provide some examples to illustrate the choices involved. One way to narrow the sets of clusters is to consider the choice of the number of clusters. Our results make clear that choosing 2 clusters minimizes the extent of year-to-year reclassification. By contrast, choosing 4 clusters maximizes the degree of characteristic similarity (minimizing distance) but produces employment shares in the middle 2 clusters that are relatively small.

If the top priority is compatibility with related data, the range of choices narrows considerably. For example, there are only 3 feasible sets of clusters that are compatible with SBA size standards (in the sense of having a break at 500 employees). Among these sets, the one with the greatest degree of characteristic similarity (smallest distance) is the 3-cluster case with breaks at 50 and 500 employees. By contrast, the set that minimizes reclassification is the 2-cluster case with a break at 500 employees.

To be somewhat more specific about the choices involved, consider the size categories used by the CES program in the preliminary size-class estimates that BLS released in February 2012 (Bureau of Labor Statistics 2012). Those size categories involved the set of 3 clusters with breaks at 50 and 500 employees. That set of clusters is among our feasible sets of 3 clusters, and it is the only feasible set that is compatible with both SBA size standards and the size categories currently used in the ADP Employment Report. However, in terms of the other measures we provide to compare sets, this set of clusters is not the "best" set of 3 clusters. Among the other feasible sets, the set with the greatest degree of characteristic similarity (smallest distance) has breaks at 50 and 250 employees. The set of clusters with the best balance of employment shares across size clusters has breaks at 50 and 1,000 employees. The set of clusters with the smallest amount of year-to-year reclassification has breaks at 100 and 1,000 employees.

According to our results, a potential alternative to the set of clusters used in the preliminary size-class estimates has breaks at 50 and 250 employees. This set of clusters has greater characteristic similarity than the set used in the preliminary estimates. However, this set has worse balance of employment shares and more reclassification, and it is not consistent with SBA size standards.

Another potential alternative to the set used in the preliminary estimates has breaks at 100 and 1,000. This set of clusters has better balance of employment shares and less reclassification than the set used in the preliminary estimates, and it is compatible with the SUSB data. However, this set has lower characteristic similarity and is not compatible with SBA size standards.

As this discussion illustrates, there is no set of 3 clusters that dominates the other feasible sets on all measures. This result also holds in the case of 2 clusters. Therefore, the set of clusters that is optimal for CES size-class estimates depends on the CES program's objectives in producing these estimates. We have provided a comprehensive approach that narrows the range of feasible clusters and provides several measures with which to compare these sets. We leave it to the CES program to determine how much weight to place on each measure. This will determine the optimal set of clusters.

In future research, we would like to explore potential applications of our clustering methodology to other contexts. In the same way that BLS is combining OMB size classes into clusters, other statistical agencies confront situations where an ordered variable taking on a discrete number of values must be collapsed into a set of categories for sampling, seasonal adjustment, or estimation. Examples of such variables in business data include assets and revenues, which are covered by the 1982 OMB directive on business size. Examples in individual data include age and income.

Our methodology may also be useful for solving spatial problems such as creating Metropolitan Statistical Areas (MSAs) from counties. Just as the size clusters must retain the ordering in the OMB-designated size classes, MSAs must be formed by adjacent counties. Commuting patterns to different cities may the basis for the characteristics used in the cluster analysis. Constraints on the sets of clusters (MSAs) may be based on population levels or shares of the regional population.

Acknowledgements

We thank Nathan Clausen, Larry Huff, Mark Loewenstein, Anne Polivka, and participants at the Joint Statistical Meetings session on "Labor Estimates by Firm Size" for helpful discussions. The views expressed in this paper are solely those of the authors and do not reflect the views of the U.S. Bureau of Labor Statistics.

References

Automatic Data Processing. 2012 (March 7). *February 2012 ADP National Employment Report.* http://www.adpemploymentreport.com (accessed March 2012).

Bureau of Labor Statistics. 2012 (February 12). *Experimental Size Class Employment, Hours, and Earnings Series from the Current Employment Statistics Survey.* Washington, DC: Bureau of Labor Statistics. http://www.bls.gov/ces/cessizeclass.htm (accessed February 2012).

Butani, Shail J., Richard L. Clayton, Vinod Kapani, James R. Spletzer, David M. Talan, and George S. Werking Jr. 2006. "Business Employment Dynamics: Tabulations by Employer Size." *Monthly Labor Review* 129 (2): 3-22.

Dalton, Sherry, Erik Friesenhahn, James Spletzer, and David Talan. 2011. "Employment Growth by Size Class: Firm and Establishment Data." *Monthly Labor Review* 134 (12): 3-12.

Haltiwanger, John, Ron Jarmin, and Javier Miranda. 2009. *Business Dynamics Statistics: An Overview.* Kansas City, MO: Ewing Marion Kauffman Foundation.

Intuit. 2012 (March 5). *Intuit Index Shows Small Businesses Creating Jobs, but Trend Slowing.* Mountain View, CA: Intuit. http://index.intuit.com (accessed March 2012).

Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley & Sons.

Macroeconomic Advisers. 2008 (December 18). *The ADP National Employment Report.* http://www.adpemploymentreport.com (accessed March 2012).

Manning, Christopher, and John P. Mullins. 2006. “Two New Construction Employment Series for Specialty Trade Contractors.” *Monthly Labor Review* 129 (10): 14-22.

Mueller, Kirk. 2006. “Impact of Business Births and Deaths in the Payroll Survey.” *Monthly Labor Review* 129 (5): 28-34.

Office of Management and Budget [OMB]. 1982 (May 18). “New Statistical Standard on Comparability of Statistics on Business Size.” *Federal Register* 47 (96), 21362-21363.

President of the United States. 1982. *The State of Small Business: A Report of the President.* Washington, DC: Government Printing Office.

Small Business Administration. 2010. Table of Small Business Size Standards Matched to North American Industry Classification System Codes. Washington, DC: Small Business Administration.

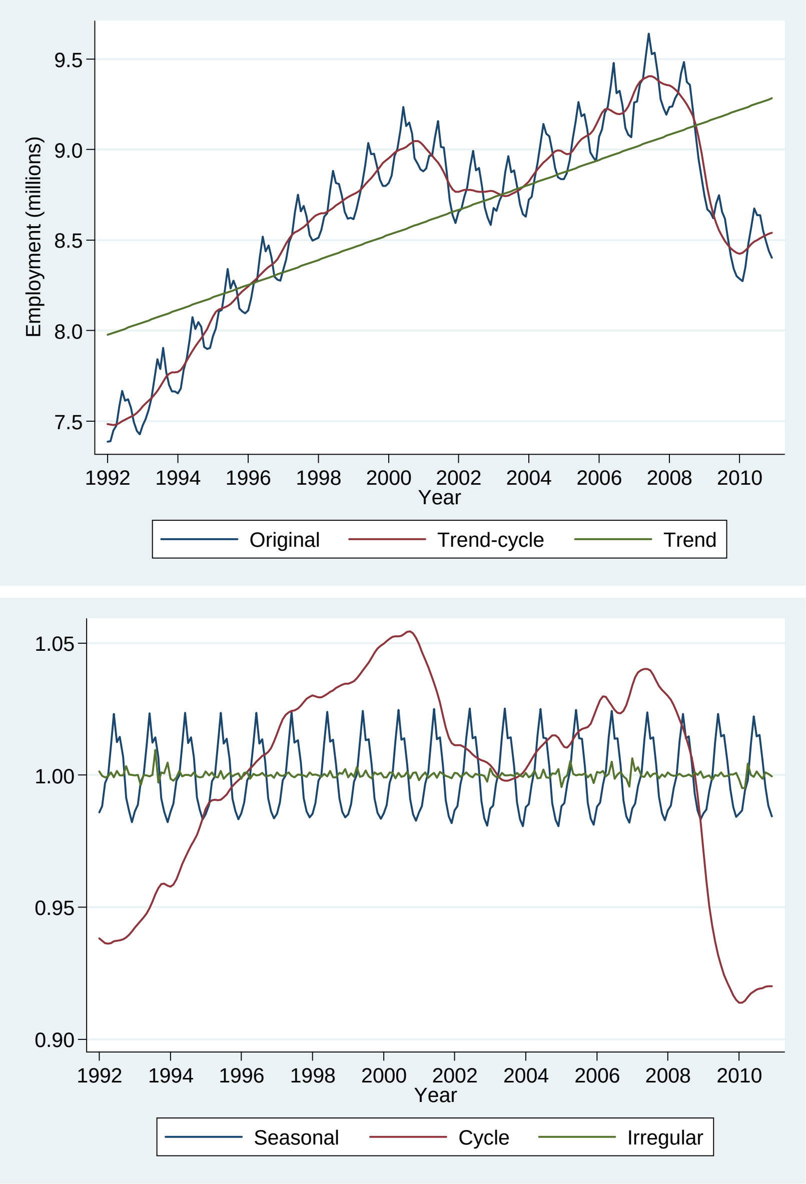

Figure 1: Example of time-series decomposition (size class 5)

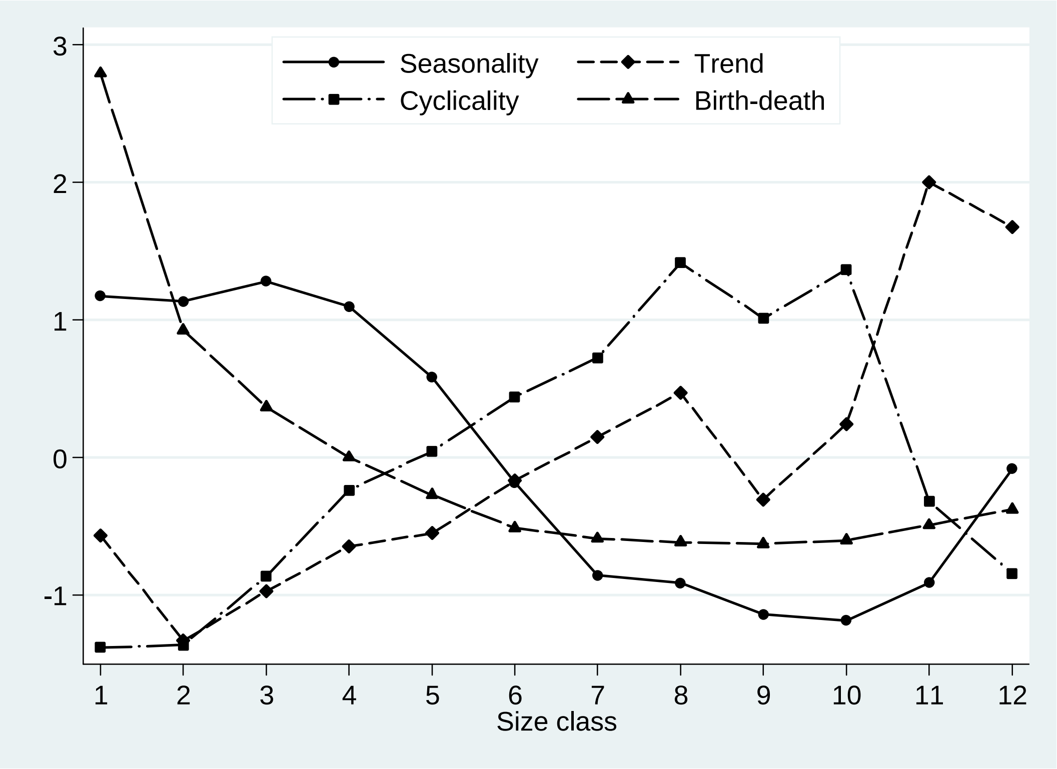

Figure 2: Characteristics of size classes

Table 1: Distributions by Size and Characteristics of Size Classes

	Distributions		Characteristics			
Size Class	Percent of Emp	Percent of EINs	Seasonality	Trend	Cyclicality	Birth-death
1 [1–4]	4.3	46.37	1.2	-0.6	-1.4	2.8
2 [5–9]	5.6	22.89	1.1	-1.3	-1.4	0.9
3 [10–19]	7.4	14.92	1.3	-1.0	-0.9	0.4
4 [20–49]	11.1	9.72	1.1	-0.6	-0.2	0.0
5 [50–99]	8.6	3.25	0.6	-0.6	0.0	-0.3
6 [100–249]	10.8	1.83	-0.2	-0.2	0.4	-0.5
7 [250–499]	7.3	0.54	-0.9	0.1	0.7	-0.6
8 [500–999]	6.9	0.25	-0.9	0.5	1.4	-0.6
9 [1k–2.5k]	9.0	0.15	-1.1	-0.3	1.0	-0.6
10 [2.5k–5k]	6.5	0.05	-1.2	0.2	1.4	-0.6
11 [5k–10k]	5.8	0.02	-0.9	2.0	-0.3	-0.5
12 [10k+]	16.6	0.02	-0.1	1.7	-0.8	-0.4
Total	100.0	100.00	0.0	0.0	0.0	0.0

Note: k=thousands (e.g., “2.5k”=2,500).

Table 2: Feasible Sets of Size Clusters Ranked by Distance

	Size Classes [Employment Range]					Percent of Employment				Percent of EINs	Percent of Emp	Compatible With		
Number of Clusters	1	2	3	4	Rank (Distance)	1	2	3	4	Moving	Moving	SUSB	SBA	ADP
2	1–5 [1–99]	6–12 [100+]			1 (13.77)	37	63			0.51	2.68	Y		
2	1–6 [1–249]	7–12 [250+]			2 (14.84)	48	52			0.18	2.40	Y		
2	1–7 [1–499]	8–12 [500+]			3 (16.72)	55	45			0.08	2.16	Y	Y	
3	1–4 [1–49]	5–6 [50–249]	7–12 [250+]		1 (12.45)	28	19	52		1.29	4.79			
3	1–4 [1–49]	5–7 [50–499]	8–12 [500+]		2 (12.81)	28	27	45		1.20	4.68		Y	Y
3	1–5 [1–99]	6–8 [100–999]	9–12 [1,000+]		3 (13.05)	37	25	38		0.54	4.24	Y		
3	1–5 [1–99]	6–7 [100–499]	8–12 [500+]		4 (13.16)	37	18	45		0.58	4.35	Y	Y	
3	1–4 [1–49]	5–8 [50–999]	9–12 [1,000+]		5 (13.19)	28	34	38		1.16	4.53			
4	1–4 [1–49]	5–6 [50–249]	7–8 [250–999]	9–12 [1,000+]	1 (11.74)	28	19	14	38	1.32	6.20			

Notes: All feasible sets are compatible with the size categories in QCEW, BED, and BDS data. No feasible set is compatible with the size categories in Intuit data. SUSB=Statistics of U.S. Businesses. SBA=Small Business Administration. ADP=ADP Employment Report. QCEW=Quarterly Census of Employment and Wages. BED=Business Employment Dynamics. BDS=Business Dynamics Statistics.