The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, stata, Stata Press, Mata, mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2015)
15, Number 1, pp. 247–274

dynemp: A routine for distributed microdata analysis of business dynamics

Chiara Criscuolo
OECD
Paris, France
chiara.criscuolo@oecd.org

Peter N. Gal
OECD
Paris, France
Tinbergen Institute and
VU University Amsterdam
Amsterdam, Netherlands
peter.gal@oecd.org

Carlo Menon
OECD
Paris, France
carlo.menon@oecd.org

Abstract. In this article, we introduce a new command, `dynemp`, that implements a distributed microdata analysis of business and employment dynamics and firm demographics. As its data source, `dynemp` requires business registers or comparable firm- or establishment-level longitudinal databases that cover the (near-)universe of companies in all business sectors. Access to such confidential data is usually restricted, and the microlevel data cannot be brought together to a single platform for cross-country analysis. To solve this confidentiality problem while also maintaining a high level of harmonization of the key economic concepts, `dynemp` can be distributed in a network of researchers who have access to the national confidential microdata. This way, the rich firm-level employment dynamics can be analyzed from new angles (such as firm age and size), significantly expanding the scope of analyses relying only on more aggregated data.

Keywords: st0379, dynemp, employment dynamics, job flows, firm demographics, administrative data, data analysis

1 Introduction

The `dynemp` command produces a set of statistics based on microlevel (firm or plant) employment data.[1] The microlevel information is aggregated to the level of industrial activities (sectors), age classes, size classes, and segments of the employment growth distribution. While most industrialized countries—as well as many emerging economies—now maintain comprehensive business registers containing information on the universe of active firms in the economy in a longitudinal format for relatively long time periods, the cross-country, comparative analysis of this rich source of data is often limited by confidentiality rules and by the lack of appropriate statistical platforms that would bring together national databases from different countries.

We developed the `dynemp` routine with the aim of providing a tool to produce a nonconfidential, microaggregated cross-country dataset, while exploiting the richness of the firm-level databases used as the underlying sources. It follows the methodol-

[1] The routine can be run on databases at either the firm/enterprise or the plant/establishment level. For simplicity throughout the article, we use the term “unit” for the longitudinal unit of analysis.

ogy of distributed microdata analysis, which was initially developed and implemented by Bartelsman, Haltiwanger, and Scarpetta (2004) for comparing firm demographics across countries. Since then, it has become increasingly used for wide-ranging purposes, including the microeconomic sources of job creation (Anyadike-Danes et al. 2013; Bravo-Biosca, Criscuolo, and Menon 2013), the impact of information and communication technologies on firm-level outcomes (Hagsten et al. 2012; Van Leeuwen and Polder 2013), and to analyze productivity dispersion and allocation of resources across firms (Bartelsman, Haltiwanger, and Scarpetta 2013). A similar methodology has been used to compare price- and wage-setting practices across Euro-area countries in the context of the inflation persistence and wage flexibility networks of the European Central Bank (Dhyne et al. 2006; Dickens et al. 2007). These analyses typically require a network of researchers or statistical offices that share a centrally written routine and run it on the national microdata. Sharing those routines with the broader research community can help with similar future projects, and it also makes previous findings more easily replicable or extendable to different countries.

In particular, `dynemp` serves several purposes within the broad theme of characterizing the dynamics of employment:

- it allows computing several indicators and summary statistics from microlevel business data;
- it provides researchers in national statistical offices with a tool for creating, and possibly publishing, detailed summary statistics on employment and business dynamics, with the possibility of blanking cells that do not comply with primary disclosure rules;[2] and
- it can serve as a platform to create harmonized cross-country databases.

The output data allow for the analysis of many policy-relevant issues on enterprise dynamics, in particular:

- identifying the contribution of different groups of firms to job creation and destruction, and the margins underlying these different contributions (for example, entry versus post-entry growth or contraction versus exit);
- characterizing the transition dynamics of cohorts of young firms;
- assessing the heterogeneous response of firms of different age, size, and sector over the business cycle and in particular during the recent international financial crisis; and
- exploring the extent to which firms differ in their employment growth performance within the same sector, size class, or age class, and within sector-size-age class.

[2] The program does not control for secondary disclosure or for cases of predominance.

The economic literature exploiting the richness of business-level data to explore employment dynamics was spurred by seminal publications in the 1990s. These publications focused mainly on the United States (Dunne, Roberts, and Samuelson 1989; Davis and Haltiwanger 1990, 1999; Davis, Haltiwanger, and Schuh 1998) and presented evidence of significant heterogeneity across different types of firms, which implies that the assumption of a "representative firm" is problematic when analyzing questions related to employment and productivity.[3]

Because of the inherent difficulties in accessing business-level data simultaneously in several countries, the first rigorous cross-country analysis of heterogeneous firm dynamics was just undertaken in the 2000s, published in Bartelsman, Scarpetta, and Schivardi (2005). The article studies firm demographics and survival across 10 OECD countries, by collecting microaggregated data from national business registers based on a common data protocol. Further, a recent contribution by Haltiwanger, Jarmin, and Miranda (2013) explores job creation and destruction dynamics in the United States, to find that young firms disproportionately contribute to job creation; once firm age is controlled for, there is no systematic relationship between firm size and growth.

`dynemp` allows updating and expanding this stream of research.[4] `dynemp` requires a unit-level (firm or establishment) input dataset containing a longitudinal unit identifier, the calendar year, the three-digit sector of activity, the birth year of the unit, and employment. Its output consists of three sets of Stata databases: the first one reports variables on gross job flows (job creation, job destruction) and employment growth by groups of firms classified by age class, size class, and employment growth percentile; the second set of output files contains the transition matrices of selected groups of firms, classified along the age and size dimension, over a 3-, 5-, or 7-year time horizon; and the third set of results consists of `.xml` (Extensible Markup Language) tables reporting the output of regressions of employment growth and the probability of exit on size class, age class, sector, and year dummies.[5]

2 Required structure of the input data, syntax, and options

2.1 Input dataset

The data source (input data) must be a longitudinal, annual, firm-, or establishment-level database with information on the number of employees, sectoral activity, and birth year of the unit. Ideally, the source should be the national business register (or possibly social security records or tax repositories) covering the universe of units in the private business sector. The calculated statistics on job dynamics for a given year

[3] On the wide dispersion of productivity across firms and its possible causes, see the reviews by Bartelsman and Doms (2000) and Syverson (2011).

[4] A first step in that direction, using a simplified set of statistics (called DynEmp Express) compared with what is contained in this routine, is presented in Criscuolo, Gal, and Menon (2014).

[5] The `.xml` tables are produced using the user-written `outreg2` command (Wada 2005). The user must install this package before running `dynemp`.

also involve information from the previous year, for instance, when calculating gross job flows. Individual units must be identified by a unique longitudinal identifier (`id()`) that is constant over time. If the unit exits the firm-level dataset, its identifier must not reappear.

More specifically, the required variables are as follows:

- the number of employees, preferably measured in full-time equivalents, averaged over the year.
- the calendar year to which the time-varying variables refer.
- the birth year of the unit. This can be earlier than the period covered by the business register. It should be constant over the entire period during which the unit is observed. It can also be missing for some units, in which case the first year of appearance is assumed to be the birth year. For those units where this coincides with the first year of the database—and birth year is missing—the birth year is left missing and age is not defined (see more below).
- the three-digit-or-lower-level sector identifying the main economic activity of the unit, following the ISIC Rev. 4 (NACE Rev. 2) classification.[6] If the sectoral classification is at a finer level than three digits, it is automatically converted to three digits. The program can also deal with the dataset being partially or completely classified according to the ISIC Rev. 3.1 (NACE Rev. 1.1) classification.[7] In such cases, the options `sectorchange`, `isic3()`, `isic4()`, and `newindyear()` need to be correctly specified (see below). The sector variable must be an integer in numeric format, and it is preferred that the sector is held fixed over time. If this is not the case, the program will attribute to the unit its modal sector (selecting the most recent modes in cases of multiple modes). See more details on this in section 3.1.
- (optional variable) the year of left-censoring for the birth variable. The left-censoring variable may change across units, but it must be constant within units. If not, the user must replace the left-censoring variable with its minimum value. For cases where birth year predates the censoring year, the program assumes that the reported birth year is correct and does not apply any correction.

[6] ISIC stands for International Standard Industrial Classification of All Economic Activities, developed by the United Nations. NACE is the Statistical Classification of Economic Activities in the European Community.

[7] When only the earlier NACE Rev. 1.1 classification is available, an external look-up table contained in the command package will be used (the file is named `changeover_database.txt`, which should be saved in the directory where the input data are stored). A similar converter for ISIC Rev. 3.1 classification is available from the authors upon request.

2.2 The dynemp command

The syntax of the dynemp command is as follows:

dynemp [*if*] [*in*], country(*country*) unit(*unit*) id(*varname*)
employment(*varname*) year(*varname*) birth(*varname*)
{isic3(*varname*)|isic4(*varname*)} [sectorchange newindyear(#)
outputdir(*string*) blank conf(#) express leftcensoring(*varname*)
yeart(*numlist*) transyears(*numlist*) extraformat(*string*) levels(*numlist*)
exitdeath(*varname*) exitchange(*varname*) noreg regyear(*numlist*)
turnover(*varname*)]

2.3 Options

country(*country*) specifies the name of the country. country() is required.

unit(*unit*) specifies the unit of analysis (for example, plant or firm). unit() is required.

id(*varname*) indicates the variable containing the unique longitudinal unit identifier. It can be either string or numeric. id() is required.

employment(*varname*) indicates the variable containing the unit's employment. It can be either an integer or a noninteger. employment() is required.

year(*varname*) indicates the year variable. It must be an integer. year() is required.

birth(*varname*) indicates the variable containing the unit's year of birth. It must be an integer. birth() is required.

isic3(*varname*) indicates the variable containing the unit's industry; this must follow the ISIC Rev. 3.1 classification at the three- or four-digit level. It must be an integer. Either isic3() or isic4() must be specified; both may be specified in a case where there is a change in classification over the sample period.

isic4(*varname*) indicates the variable containing the unit's industry; this must follow the ISIC Rev. 4 classification at the three- or four-digit level. It must be an integer. If isic4() is left empty, the external conversion table, changeover_database.txt (which is included in the command package), is required. Either isic4() or isic3() must be specified; both may be specified in a case where there is a change in classification over the sample period.

sectorchange specifies there is a change in classification over the sample period, that is, a change in sectoral classification from ISIC Rev. 3.1 to ISIC Rev. 4 happens at some point in the dataset. In such a case, both of the industry variable options (isic3() and isic4()) must be specified, although they can refer to the same variable. sectorchange requires that the option newindyear() also be specified.

`newindyear(`*#*`)` specifies the year in which the industrial classification changes from ISIC Rev. 3.1 to ISIC Rev. 4. It must be an integer. `newindyear()` requires that the option `sectorchange` also be specified.

`outputdir(`*string*`)` specifies the output directory (for example, `C:\OECD\output`). If not specified, the `OUTPUT_TOSEND` folder is created in your Stata working directory on Windows or in your home directory on Mac or Unix.

`blank` sets to missing all the records referring to cells containing fewer units than the confidentiality level (which is set in option `conf()`; see below).

`conf(`*#*`)` sets a confidentiality level, that is, the minimum number of units in a given cell. The command also shows the number of cells below such level on screen, as a preview of the number of cells that are likely to be blanked. It can be any positive integer. The default is `conf(5)`.

`express` runs a faster version of the code that excludes the calculation of percentiles.

`leftcensoring(`*varname*`)` indicates the variable reporting the year of left-censoring in the business register. It must be an integer.

`yeart(`*numlist*`)` specifies the years over which the program will run. The default is to start in 1998 and end in 2011 (or the latest available year).

`transyears(`*numlist*`)` specifies the starting years of transition matrices. The default is `transyears(2001 2004 2007)`.

`extraformat(`*string*`)` specifies additional formats for the output datasets. *string* may be `txt` (tab-separated) or `csv` (comma-separated), which correspond to the file extensions.

`levels(`*numlist*`)` limits the yearly flow datasets to the selected aggregation levels (see table 3).

`exitdeath(`*varname*`)` identifies a binary variable (0/1) where 1 flags exit events due to the closing down of the business. The variable should equal 1 only in the unit's last year of appearance.

`exitchange(`*varname*`)` identifies a binary variable (0/1) where 1 corresponds to exit events due to a change in legal status, for example, mergers and acquisitions. The variable should equal 1 only in the unit's last year of appearance.

`noreg` tells the program not to run distributed regressions.

`regyear(`*numlist*`)` specifies the years over which the program will run the regressions. The default is to run them for all available years in the data. For example, if the chosen period is 2004–2008, the option would be `iregyear(2004(1)2008)`.

`turnover(`*varname*`)` identifies the variable containing turnover values. It must be numeric.

2.4 System requirements

`dynemp` does not require much more memory than the amount needed to load the input dataset. The computation time with a standard PC is less than one hour for smaller datasets (for example, fewer than 1 million units) and within five hours for larger ones (4–5 million units), assuming a temporal extension of around 10 years.

3 Input data harmonization and output datasets

3.1 Data cleaning

The program performs basic consistency checks of the data and corrects observations that are considered implausible: it replaces negative values for employment with missing; it interpolates employment records that are disproportionately smaller or bigger than those of the previous and following years (threshold values are set to $+/-1.5$ and are calculated as in (1); moreover, the correction only affects units with at least 20 employees on average over the years $t-1$, t, $t+1$); and it replaces industry classification that varies over time with the modal three-digit sector by which the unit's activity is classified. In case of multiple modes, the program chooses the most recent mode.

Probabilistic industry conversion

Industrial classification systems such as ISIC or NACE are revised regularly to reflect structural changes in the economy. Typically, services become more specialized and gain more importance, thus requiring a more detailed breakdown, while other activities that become less important may be classified in less detail. A recent major change occurred in 2008–2009, where many former industries were split into several parts, and others merged into a single industry. For example, the activities classified under printing and publishing (code 22) in NACE Rev. 1.1 became split into five different two-digit industries in NACE Rev. 2 (some of them in manufacturing, some in services). As such, changes were not one-to-one but n-to-m types, and this applies to all levels of industry classification (that is, two-, three-, and four-digit levels).

Moreover, units also change their activity from time to time irrespective of classification system changes. However, researchers typically find it more convenient to work with a constant industry identifier over time for each unit because it simplifies many types of analyses that use the industry dimension. Finally, a constant industry classification per unit simplifies entry and exit definitions because there is no need to follow which activity the unit enters or exits. To accommodate these needs, and to work with the latest available classification system, we designed the following probabilistic conversion system:[8]

[8] We are grateful for Eric J. Bartelsman, who highlighted this idea during our discussions.

1. To convert the old classification to the new one, the routine creates a conversion table based on classifiers at the three-digit level. In overlapping years, units are registered with both their old and their new classifications; if such overlapping years do not exist in the database, then `dynemp` relies on units that exist in both systems and creates a link as follows: the observed value in the old system in, for example, 2008 will be paired with the new value in 2009.

2. This procedure may yield n-to-m type pairs. `dynemp` will use them in a probabilistic way, by calculating the frequency at which each industry in the old system occurs in the new system. To make the conversion more tractable, such n-to-m transition pairs are disregarded where the fraction of units classified out of an old industry classification into a new one is less than or equal to 10%. This conversion table is stored along with the frequencies of transition pairs.

3. Returning to the unit-level database, for each unit the following steps are taken:

   a. The first step involves finding the industry where the unit spent most of its observed years. Because part of the industry classifications associated with a unit may come from the old system (that is, before the changeover year) and another part from the new system (that is, after and including the changeover year), one needs to take this into account when finding the most appropriate industry classification for the unit. For this reason, a temporary industry classifier $i_{\text{temp},t}$ is created, which is defined to take the value of the new classification $i_{\text{new},t}$ after the changeover year:

   $$i_{\text{temp},t} \quad = \quad i_{\text{new},t} \text{ if } t \geq t_{\text{change}-\text{year}}$$

   Before the changeover year, $i_{\text{temp},t}$ is driven backwards in time, assuming that in the changeover year there was no real change in the activity of the unit. However, if there is an observed change in the industry classification in the years before the changeover year, the temporary variable is also changed accordingly:

   $$\begin{aligned} i_{\text{temp},t} &= i_{\text{temp},t+1} \text{ if } t < t_{\text{change}-\text{year}} \text{ and } i_{\text{old},t} = i_{\text{old},t+1} \\ i_{\text{temp},t} &= i_{\text{old},t} \text{ if } t < t_{\text{change}-\text{year}} \text{ and } i_{\text{old},t} \neq i_{\text{old},t+1} \end{aligned}$$

   Finally, if the unit does not have an observation in the new system (that is, all its observations are before the changeover year), the temporary variable merely takes the original values from the old system.

   b. Based on this temporary classifier, the routine selects the industry that classifies the activity that the unit has carried out the longest; that is, it chooses the mode industry classification (the value that appears most often). If this value is not unique, the most recent one is selected. The result is a single industry classifier for each unit. For some, it is from the new system; for others (who did not exist at the changeover year or whose most frequent industrial classification is in the years before the changeover), it is from the old system.

c. For those units that have their industrial classification defined in the old system, the routine assigns a value from the new system based on the conversion table's frequencies obtained in step 2. For instance, if the industry classifier X from the old system is split into three industry classifiers Y_1, Y_2, and Y_3 for 25%, 35%, and 40% of the cases, then the units that belong to X will get a randomly assigned new industry classifier, where the probability of being classified into a new industry will equal the observed frequencies in the conversion table—that is, 25%, 35%, and 40% in this example.[9]

Employment, growth, birth, and exit definition

The program calculates a few intermediate unit-level variables, which are subsequently used to calculate summary statistics at different aggregation levels in the final microaggregated ("collapsed") dataset. The program runs regardless of whether the employment data are expressed as an integer or a decimal number (it rounds up in the latter case). It is assumed that no additional rounding beyond that to unity is applied on the data, that is, that employment figures are not rounded to multiples of 5, 10, 100, etc.

The employment growth rate is calculated according to the following formula:

$$\gamma_{i,t}^{L} = \frac{L_{i,t} - L_{i,t-1}}{\frac{1}{2}(L_{i,t} + L_{i,t-1})} \tag{1}$$

where $L_{i,t}$ stands for employment of unit i in year t. The formula is commonly used in the business dynamics literature because it has the advantage of not being biased by mean-reversion dynamics (see Davis and Haltiwanger [1999], among others). The index is also scale neutral (that is, it does not depend on the employment level at the beginning of the period) and is bounded between -2 and $+2$.[10]

Year of birth is the first year of activity of the unit and is needed to calculate the unit's age. If the data are left-censored and the user specifies this, the calculation of the age variable will take this into account. Finally, the exit variable is a dummy equal to 1 in the year following the last time a unit appears in the data with positive employment.

Entering, exiting, and incumbent units

The transition matrices and the yearly job flow statistics are calculated for three different groups of units: entrants, exitors, and incumbents. For each interval $(t-1, t)$, we define entrants, exitors, and incumbents as follows:

[9] To make the procedure replicable, the seed of the random-number generator is always reset to the same number.

[10] Up to a second-order approximation, it is equivalent to taking first differences of the logarithms of the series.

- an entrant is a unit that is not present in the data in $t - 1$ but is present in t;
- an exitor is a unit that is not present in t but is present in $t - 1$;
- an incumbent is a unit that is present in $t - 1$ and t.

3.2 Output datasets

The `dynemp` command creates several new output files. These files are saved to the newly created folder `OUTPUT_TOSEND`.

- The aggregated statistics on yearly job flows are saved in Stata dataset format files named
  - `dynemp_`*country_unit*`_lev1.dta`
  - `dynemp_`*country_unit*`_lev2.dta`
  - `dynemp_`*country_unit*`_lev3.dta`
  - `dynemp_`*country_unit*`_lev4.dta`
- The transition matrices, also containing employment growth volatility estimates over a three-, five-, and seven-year horizon, averaged across years and two-digit structural analysis (STAN) database A38 sectors are saved in Stata dataset format files named
  - `dynemp_`*country_unit*`_trans_mat.dta`
- The distributed regression output tables are saved in Extended Markup Language (`.xml`) files named
  - `dynemp_`*country_unit*`_regexit.xml`
  - `dynemp_`*country_unit*`_reggrowth.xml`
  - `dynemp_`*country_unit*`_sizecont.xml`
  - `dynemp_`*country_unit*`_sizecont2.xml`
- The tabulation of gaps in the data are saved in a plain-text format file named
  - `Dynemp_`*country_unit*`_tabgaps.txt`

Notes: *country* is the country name specified in option `country()`. *unit* corresponds to the selected unit of analysis (for example, plant or firm) specified in option `unit()`. `lev1`–`lev4` identify the four levels of aggregation, which arise from combinations of the sector, age, size, and employment growth classifications. `regexit` identifies regressions of probability of exit on class size, age class, sector, and year dummies. `reggrowth` identifies regressions of growth rate on class size, age class, sector, and year dummies.

`sizecont` identifies regressions of employment growth indices for one-, three-, and five-year horizons on dummies for employment levels corresponding to regulatory thresholds. And `sizecont2` identifies regressions of the share of shrinking, growing, and stable units on employment-level dummies. See table 1 for additional information on the four levels of aggregation.

3.3 Annual flow datasets

The flow datasets contain annual statistics on gross job flows (gross job creation and job destruction, defined as the total job variation of growing and shrinking units, respectively) and on several moments of unit-level employment growth (mean, median, and standard deviation); the latter three statistics are also calculated for the turnover variable if available. To simplify confidentiality clearing, all median values are calculated as the average value of the three central values in the reference group distribution. The flow output datasets also report the total number of units in the cell, their median and average age, the number of units never growing above one employee, the units that appear just for one year, and statistics on the high-growth units based on the OECD–Eurostat definition (Eurostat and OECD 2007).[11]

The aggregation levels considered are summarized in table 1. Macrosectors are manufacturing (10–33), nonfinancial business services (45–63 and 68–75) and construction (41–43) (NACE Rev. 2 two-digit classifications in parentheses; STAN produced and maintained by the OECD). The STAN sector aggregation is generally done on the basis of the A38 level; see the list of industries and macrosectors summarized in table 2.

Table 1. Aggregation levels in transition matrices

Level	Sector	Growth percentiles	Size	Age
1	3 macrosectors	5 growth percentiles		3 classes
2	3 macrosectors		6 classes	3 classes
3	27 STAN A38 (two-digit ISIC/NACE2)		4 classes	3 classes
4	27 STAN A38 (two-digit ISIC/NACE2)	5 growth percentiles		

[11] "All enterprises with average annualized growth greater than 20% per annum over a three-year period should be considered as high-growth enterprises. Growth can be measured by the number of employees or by turnover" (Eurostat and OECD 2007).

Table 2. Industries included in `dynemp`

Macrosectors	Included in macro-sectors	Covered NACE2/ ISIC4	Included in two-digit breakdown industries	Name
		01–03	•	agriculture, forestry, and fishing [A]
		05–09	•	mining and quarrying [B]
Manufacturing	•	10–12	•	food products, beverages, and tobacco [CA]
	•	13–15	•	textiles, wearing apparel, leather, and related products [CB]
	•	16–18	•	wood and paper products; printing [CC]
	•	19	•	coke and refined petroleum products [CD]
	•	20	•	chemicals and chemical products [CE]
	•	21	•	basic pharmaceutical products and pharmaceutical preparations [CF]
	•	22–23	•	rubber and plastics products, and other nonmetallic mineral products [CG]
	•	24–25	•	basic metals and fabricated metal products, except machinery and equipment [CH]
	•	26	•	computer, electronic, and optical products [CI]
	•	27	•	electrical equipment [CJ]
	•	28	•	machinery and equipment n.e.c. [CK]
	•	29–30	•	transport equipment [CL]
	•	31–33	•	furniture; other manufacturing; repair and installation of machinery and equipment [CM]
		35	•	electricity, gas, steam, and air conditioning supply [D]
		36–39	•	water supply; sewerage, waste management, and remediation activities [E]
Construction	•	41–43	•	construction [F]

Continued on next page

Macrosectors	Included in macro-sectors	Covered NACE2/ ISIC4	Included in two-digit breakdown industries	Name
Nonfinancial business services	•	45–47	•	wholesale and retail trade/repair of motor vehicles and motorcycles [G]
	•	49–53	•	transportation and storage [H]
	•	55–56	•	accommodation and food service activities [I]
	•	58–60	•	publishing, audiovisual, and broadcasting activities [JA]
	•	61	•	telecommunications [JB]
	•	62–63	•	IT and other information services [JC]
		64-66	•	financial and insurance activities [K]
	•	68	•	real estate activities [L]
	•	69–71	•	legal and accounting activities [MA]
	•	72	•	scientific research and development [MB]
	•	73–75	•	advertising and market research; other professional, scientific, and technical activities; veterinary activities [MC]
	•	77–82	•	administrative and support service activities [N]
		85	•	education [P]
		86	•	human health activities [QA]
		87–88	•	residential care and social work activities [QB]
		90–93	•	arts, entertainment, and recreation [R]
		94–96	•	other service activities [S]

Note: The list and definition of industries are based on the OECD's STAN A38 industry classification.

Source: http://www.oecd.org/sti/ind/2stan-indlist.pdf

Size classes considered in aggregation level 2 are 0–9, 10–49, 50–99, 100–249, 250–499, and 500+. Size classes considered in aggregation level 3 are 0–9, 10–49, 50–249, and 250+. Age classes considered in aggregation levels 1, 2, and 3 are 0–2, 3–5, 6+, and 99 (missing). Size is defined according to the average of employment at time $t-1$ and t for incumbents, employment at time $t-1$ for exitors, and employment at time t for entrants. Employment growth classes are defined on five intervals of the growth distribution. These data are only available, and hence computed, for incumbents.

The classes are divided according to the following percentile thresholds: bottom 10% of the distribution, 11th to 25th percentile, 26th to 75th percentile, 76th to 90th percentile, and top 10% of the distribution. This classification, however, may be problematic if a significant share of units in the reference group has zero growth, because all these units would end up in the same percentile group. To avoid this, the percentile allocation is based on a growth rate that is increased or decreased by a random small number if the actual growth rate is equal to 0. The random number is drawn from a uniform distribution with the maximum value set to the minimum (in absolute value) nonzero growth rate in the same country and calendar year.

Variables in annual flow datasets

Using the breakdowns above, the variables created are summarized in table 3. Gross job flows are defined as follows:

- Job creation (JC_{jt}) captures the gross amount of jobs created in year t by unit in group j, and it is defined as

  $$\mathrm{JC}_{jt} = \sum_{i \in j} \Delta L_{it}^{+}$$

  where i indexes units and ΔL_{it}^{+} is a positive employment change from the previous year.

- Job destruction (JD_{jt}) measures the gross amount of jobs lost from period $t-1$ to t:

  $$\mathrm{JD}_{jt} = \sum_{i \in j} |\Delta L_{it}^{-}|$$

  where $|\Delta L_{it}^{-}|$ is the negative employment change in absolute terms.

Table 3. Variables in the annual job flow datasets

Variable name	Description
`macrosector`	manufacturing, services, construction; computed in levels 1 and 2
`ageclass`	aggregation according to `ageclass` in levels 2 and 3—only incumbents considered in levels 1 and 4; computed in levels 2 and 3
`sizeclass`	aggregation according to `sizeclass` (note: different size classifications in levels 2 and 3—see section *Entering, exiting, and incumbent units*); computed in levels 2 and 3
`prc`	aggregation according to percentiles of employment growth; computed in levels 1 and 4
`group`	whether the firm is an incumbent, entrant, or exitor
`meangrowthemp`	average growth in employment from time $t-1$ to t
`meanemp`	average employment at time t for firms in the group
`meantrn`	average turnover at time t
`meangrowthtrn`	average growth in turnover from time $t-1$ to t
`meanturnovemp`	mean turnover per employee
`medianage*`	median age of firms in the group
`emp1emp`	total employment of one-employee units
`emp1year`	total employment of one-year firms
`grosscreatemp`	gross job creation from time $t-1$ to t
`grossdestremp`	gross job destruction from time $t-1$ to t
`grosscreattrn`	gross turnover growth from time $t-1$ to t
`grossdestrtrn`	gross turnover loss from time $t-1$ to t
`medianemp*`	median employment of firms in the group
`medianempt_1*`	median employment of firms in the group at time $t-1$
`mediangrowthemp*`	median growth in employment from time $t-1$ to t
`mediangrowthtrn*`	median growth in turnover from time $t-1$ to t
`mediantrn*`	median turnover at time t
`medianturnovemp*`	median turnover per employee
`mediantrnt_1*`	median turnover at time $t-1$
`nrunit_posemp`	number of units with employment greater than 0
`nr1emp`	number of units never growing over one employee
`nr1year`	number of units appearing for just one year
`nrunit`	number of units in the group
`p90p10turnovemp`	difference between the 90th and 10th percentiles in turnover per employee
`sdemp`	standard deviation of employment at time t
`sdtrn`	standard deviation of turnover at time t
`sdtrnovemp`	standard deviation of turnover per employee at time t
`totemp`	total employment at time t
`tottrn`	total turnover at time t
`nrunit_hgf`	number of high-growth firms
`medianage_hgf`	median age of high-growth firms
`totemp_hgf`	total employment in high-growth firms
`meanemp_hgf`	mean employment in high-growth firms
`grosscreat_emp_hgf`	gross job creation of high-growth firms
`grossdestr_emp_hgf`	gross job destruction of high-growth firms
`meangrowth_emp_hgf`	mean growth of employment in high-growth firms
`year`	reference year

3.4 Transition matrices

The transition matrices summarize the growth trajectories of cohorts of units from year t to year $t+j$, where t takes by default the values 2001, 2004, and 2007 if not otherwise specified by the option `transyears()`, and j is equal to 3, 5, or 7 (therefore, if data are available, transition matrices are calculated for the periods 2001–2004, 2001–2006, 2001–2008; 2004–2007, 2004–2009, 2004–2011; and 2007–2010, 2007–2012, 2007–2014). The matrices contain a few basic statistics (number of units in the cell, median employment at t and at $t+j$, total employment at t and at $t+j$, and mean growth rate) for several different combinations of age classes and size classes at times t and $t+j$, plus a focus on the dynamics of high-growth units. The different aggregation levels are reported in table 4.

Table 4. Aggregation levels in transition matrices

Size class at time t	Age class at time t	Size class at time $t+j$	Sectors
all (nonmissing)	0, 1–2, 3–5, 6–10, 11+	all surviving, missing employment, exit	
0–9, 10–19, 20–49, 50–99, 100–249, 250+, missing employment	all	all surviving, missing employment, exit	
0–9, 10–19, 20–49, 50–99, 100–249, 250+, missing employment	1–2, 3–5, 6–10, 11+	all surviving, missing employment, exit	manufacturing, services, construction, all private sector
0–9, 10–19, 20–49, 50–99, 100–249, 250+, missing employment	entrants	0–9, 10–19, 20–49, 50–99, 100–249, 250+, exit, missing employment	
all (nonmissing)	entrants	all surviving, missing employment, exit	
all (nonmissing)	entrants, all others	all surviving, missing employment, exit	two-digit STAN A38

Variables in transition matrices

The variables contained in the transition matrices are listed in table 5. In addition to the standard set of variables computed in the flow datasets, `dynemp` constructs an average measure of unit-level volatility of employment growth. The measure is calculated in two steps. In the first step, for each unit i and period t, the program computes the unit-level standard deviations of the employment growth rate over rolling windows of length S (with $S = 3, 5$, and 7),

$$\sigma_{it}^{S} = \sqrt{\sum_{i \in j, s=1}^{S} \left(\gamma_{i,t+s}^{L} - \overline{\gamma}_{it}^{L}\right)^2}$$

where $\gamma_{i,t+s}^{L}$ is the annual growth rate of employment in unit i over the period $t+s$,

$$\gamma_{i,t+s}^{L} = \frac{L_{i,t+s} - L_{i,t+s-1}}{\frac{1}{2}(L_{i,t+s} + L_{i,t+s-1})}$$

and $\overline{\gamma}_{it}^{L}$ is the average employment growth over period $(t+1, t+S)$.

The second step is to average these unit-level volatilities over the group of units $i \in j$ in period t,

$$\sigma_{jt}^{vol,S} = \sum_{i \in j} w_{it}^{S} \sigma_{it}^{S}$$

where weights w_{it}^{S} are defined as the average shares of the group employment in unit i $(t, t+S)$:

$$w_{it}^{S} = \frac{\sum\limits_{s=0}^{S} L_{i,t+s}}{\sum\limits_{i \in j} \left(\sum\limits_{s=0}^{S-1} L_{i,t+s}\right)}$$

Table 5. Variables in the transition matrices datasets

Variable name	Description
macrosect	macrosector classification (manufacturing, services, construction, or all)
ageclass4	age class
sizeclass6	size class at time t
f_sizeclass6	size class in the forward period
totemp	total employment at time t
f_totemp	employment in the forward period
medianemp*	median employment at time t
f_medianemp*	median employment in the forward period
nrunit	number of units in the group
meangrowth	mean growth rate
volat_emp	employment growth volatility, calculated at firm level and averaged at sector level
volat_trn	turnover growth volatility, calculated at firm level and averaged at sector level
JC_surv	gross job creation from time t to $t+j$
JD_surv	gross job destruction from time t to $t+j$
JC_surv_top10	gross job creation from time t to $t+j$; top 10% firms for employment growth
Jobvar_top10	net job variation from time t to $t+j$; top 10% firms for employment growth
j	number of years ahead of time t to which the forward period refers
year	reference year

3.5 Distributed regressions

dynemp runs a series of unit-level regressions on the full sample.

The first set of estimates consists of five ordinary least-squares regressions with the growth rate as the dependent variable and the following sets of dummies on the right-hand side of the equation: i) size; ii) age; iii) size–age; iv) size–age interacted with the "big recession" (2008–2009) dummy; v) size–age interacted with the "hi-tech sector" dummy.[12]

The second set of regressions is based on a linear probability model where the dependent variable is the "exit" dummy. This set of regressions follows the same structure as the first (although the model with age dummies only is excluded). Year and three-digit sector fixed effects are included in all specifications. The output dataset contains only the coefficients on the age and size dummies, the number of observations, and statistics on the quality of fit.

The third set of regressions is aimed at analyzing the effects of size-contingent policies on firm or establishment growth. This is done in two different ways. First, the employment growth index over a one-, three-, and five-year horizon is regressed over a set of dummies for different employment levels (8–9, 13–14, 18–19, 23–24, 48–49, 98–99)

[12] The hi-tech dummy is based on the 2009 Eurostat classification of "High-technology" manufacturing activities and "Knowledge-based services". The big recession dummy is equal to 1 in the years 2008 and 2009, when the peak of the downturn was reached by most OECD countries.

corresponding to possible regulatory thresholds in certain countries. Second, the share of shrinking, growing, and stable units for each employment level from 1 to 50 is regressed over a full set of employment-level dummies. The output dataset contains only the coefficients on the age and size fixed effects, the number of observations, and some statistics on the quality of fit.

3.6 Confidentiality

The program deals with confidentiality only if the `blank` option is specified. In that case, it performs a simple blanking of cells containing fewer units than the set number (the default of which is 5 and can be changed in the `conf()` option). All percentile values are calculated as the average of the two units around the percentile value and the percentile value itself in the distribution of interest. In such a way, no information referring to an individual unit is disclosed.

The program does not deal with more complex issues such as residual confidentiality or concentration.

4 Example

The business register of the DynEmp Republic presents the following structure:

```
. use randomdata_sj

. describe

Contains data from randomdata_sj.dta
  obs:        78,360
 vars:             9                          1 Sep 2014 18:17
 size:     1,410,480
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
idimp           int     %9.0g                 longitudinal unit identifier,
                                                numeric variable
empl            int     %9.0g                 headcount employees point-in-time
rsect           int     %9.0g                 3-digit ISIC 3.1 industry
                                                classification
birthyear       int     %9.0g                 year of birth of the unit
yyear           int     %9.0g                 year
bankrupt        byte    %9.0g                 exit by bankruptcy/liquidation:
                                                0/1 dummy
MeA             byte    %9.0g                 exit by merger or acquisition: 0/1
                                                dummy
sales           float   %9.0g
censor          int     %9.0g
--------------------------------------------------------------------------------
Sorted by:  idimp  yyear
```

```
. summarize
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       idimp |     78360    5001.079    2865.986          1      10000
        empl |     78360     200.447     209.194          0       2609
       rsect |     78360    400.2432    179.2493        100        630
   birthyear |     78360    1993.881    4.066733       1992       2011
       yyear |     78360    2005.962    3.158269       2001       2011
-------------+--------------------------------------------------------
    bankrupt |     78360    .0102986    .1009589          0          1
         MeA |     78360    .0105411    .1021279          0          1
       sales |     78360    210.3576    167.3552         50     2137.2
      censor |     78360        1992           0       1992       1992
```

Here `idimp` is the longitudinal unit identifier that denotes calendar year, `empl` is the unit's total employment in the year indicated by the `yyear` variable, `rsect` is the three-digit industry code (based on ISIC Rev. 3.1 through year 2007 and on ISIC Rev. 4 from 2008 onward), `birthyear` is the year of birth of the company, `bankrupt` is a dummy variable equal to 1 if the unit is last appearing in the dataset in that year because it is closing down, `MeA` is a dummy variable equal to 1 if the unit is last appearing in the dataset in that year because of being acquired by or merged with another unit, `sales` is the unit's turnover in the year indicated by the `yyear` variable, and `censor` is a variable that indicates the year of left-censoring for the `birthyear` variable in the business register.

The output datasets will be stored in an empty directory the user has created. If the data followed the NACE Rev. 1.1 sectoral classification for the entire period, the `changeover_database.txt` file—which is part of the routine package—would need to be saved in the folder containing the input data. To do this, open the input dataset with the command `use`, and then change the Stata working directory to the one that will contain the output datasets (unless the path is specified in the `outputdir()` option). It is also advisable to open a log file before executing the program. Then, the `dynemp` command can be launched:

```
. dynemp, id(idimp) year(yyear) employment(empl) country(DYNEMPREP) unit(unit)
> birth(birthyear) isic3(rsect) isic4(rsect) sectorchange newindyear(2008)
> exitchange(MeA) exitdeath(bankrupt)
  (output omitted)
```

As explained in section 3.1, because both the `isic3()` and the `isic4()` options are specified, the command converts ISIC Rev. 3.1 (or NACE Rev. 1.1) industry classification to ISIC Rev. 4 (NACE Rev. 2), creating a probabilistic conversion table.

The program may need a few hours to run in a standard personal computer if the input data contain information on a few million units, as is the case for business registers of large industrialized countries. During its execution, `dynemp` first noisily displays some summary statistics of the input dataset before and after the data-cleaning part. Subsequently, it prints on screen the tasks that it is performing. When the program has finished, the following files are stored in the output folder:

1. dynemp_DYNEMPREP_unit_lev1.dta
2. dynemp_DYNEMPREP_unit_lev2.dta
3. dynemp_DYNEMPREP_unit_lev3.dta
4. dynemp_DYNEMPREP_unit_lev4.dta
5. dynemp_DYNEMPREP_unit_trans_mat.dta
6. dynemp_DYNEMPREP_unit_regexit.txt
7. dynemp_DYNEMPREP_unit_regexit.xml
8. dynemp_DYNEMPREP_unit_reggrowth.txt
9. dynemp_DYNEMPREP_unit_reggrowth.xml
10. dynemp_DYNEMPREP_unit_sizecont.txt
11. dynemp_DYNEMPREP_unit_sizecont.xml
12. dynemp_DYNEMPREP_unit_sizecont2.txt
13. dynemp_DYNEMPREP_unit_sizecont2.xml
14. Dynemp_DYNEMPREP_unit_tabgaps.txt

Files 1 to 4 contain the yearly flow data, with the variables listed in table 3; file 5 contains the transition matrices, with the variables listed in table 5; files 6 to 13 contain the regression tables produced by the distributed regressions; the last file contains a tabulation of gaps in the sample.

Each file contains a number of observations which is equal to or smaller than the total number of possible combinations of the several dimensions (age class, size class, percentiles, etc.) along which the data are aggregated. For example, the level 1 dataset is broken down by 10 years, 4 macrosectors (including the "all" aggregation), 1 group (incumbents only), 3 age classes (including "missing"), and 6 percentiles (including "missing") of the growth distribution. Therefore, the maximum number of cells is 720 (resulting from $10 \times 4 \times 1 \times 3 \times 6$). The actual number can be lower, however, because some combinations—for example, those with missing age or growth—are empty; in such a case, the cell is not defined and does not appear in the dataset.

```
. use dynemp_DYNEMPREP_unit_lev1.dta

. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        year |       547    2006.523    2.891336       2002       2011
   macrosect |       547    2.744059    2.326849          1          9
       group |         0
   ageclass3 |       547    2.789762    2.315271          1          9
         prc |       547    11.24863    26.96267          1         99
-------------+--------------------------------------------------------
      nrunit |       547    255.1554    563.3662          1       3539
nrunit_pos~p |       547    254.4351    563.6834          0       3539
   medianage |       501    8.147705     6.25349          1         19
   medianemp |       501    136.8762    60.60189          0        478
     meanemp |       547    187.3392    82.09152          0      659.5
-------------+--------------------------------------------------------
      totemp |       547    51118.19    112580.5          0     709529
     emp1emp |       547    1.400366    5.979715          0         40
      nr1emp |       547    2.120658    6.587539          0         40
    totemp_b |       547    51118.19    112580.5          0     709529
    nrunit_b |       547    255.1554    563.3662          1       3539
-------------+--------------------------------------------------------
grosscreat~p |       547    2008.464     4865.04          0      26569
grossdestr~p |       547    2037.254    4921.767          0      26155
mediangrow~p |       471   -.0109749    .1201746  -.2660104    .210655
meangrowth~p |       500   -.0101818    .1273944  -.2660104   .2222222
       sdemp |       522      184.31    77.79493          0   510.1482
-------------+--------------------------------------------------------
medianempt_1 |       501     137.511    58.79141          0        440
```

The level 2 dataset is bigger: now units are also classified along the group dimension (entering, exiting, incumbents, and possibly also `exitchange()` and `exitdeath()`) and along size class. However, it is important to be aware of the risks of double counting when collapsing the dataset; for instance, if `exitchange()` and `exitdeath()` are defined, exiting units are included in two groups—the "exiting" one and either the `exitchange()` or the `exitdeath()` group. Also, the macrosector "all" (codified with 9) is equal to the sum of the values of the three macrosectors (conditional on not having blanked values in the sample). Therefore, summing the four aggregates would lead to double counting.

The user should also keep in mind that while some variables can be aggregated simply by summing them because they are simple counts (for example, `nrunit`, `totemp`, `grosscreatemp`, and `grossdestremp`), others need to be weighted by the number of units on which they are calculated (for example, `meangrowthemp` should be weighted by `nrunit_b`, that is, the number of units with nonmissing employment at both time t and time $t-1$). Still other variables—namely, the median value—cannot be aggregated at all.

Note that `totemp` is reported for entrants and exitors only, while `grosscreatemp` and `grossdestremp` are set to missing. For most applications, it would be appropriate to replace `grosscreatemp` of entrants equal to `totemp`, `grossdestremp` of exitors equal to `totemp`, and `totemp` of exitors equal to missing (because at the end of year t they have exited already). Particular care is necessary when calculating growth statistics at the cell level, because cell population varies over time as units get older and change size class—that is, a cell should not be interpreted as a cohort of units that can be followed over time, but rather as a snapshot of a group of units over the biennium $t-1$, t. Therefore, the cell growth rate can be calculated as the difference between the total employment of the cell at time t minus the total employment of the same cell at time $t-1$ over the average of the two values. However, given that the cell composition changes over time, the total employment of the same cell at time $t-1$ needs to be calculated using the gross flows. Operationally, this leads to the following formula:

$$\mathrm{GrowthRate}_{t-1,t} = \frac{\texttt{totemp}_t - (\texttt{totemp}_t - \texttt{grosscreatemp}_t + \texttt{grossdestremp}_t)}{0.5 \times \{\texttt{totemp}_t + (\texttt{totemp}_t - \texttt{grosscreatemp}_t + \texttt{grossdestremp}_t)\}}$$

where `totemp` minus `grosscreatemp` plus `grossdestremp` corresponds to the total cell's employment at year $t-1$. The resulting cell growth rate is normally different from the average growth rate (`meangrowthemp`) in the cell, because the former is implicitly weighted by each unit's average employment level over the period.

```
. use dynemp_DYNEMPREP_unit_lev2.dta

. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        year |      2200    2006.315    3.072442       2001       2011
   macrosect |      2200    4.013182    3.342821          1          9
       group |         0
   ageclass3 |      2200        2.06    .8636145          1          3
  sizeclass6 |      2200    3.581818    1.612998          1          6
-------------+--------------------------------------------------------
      nrunit |      2200    74.20182    236.0856          1       2023
nrunit_pos~p |      2200    74.00182    235.9684          0       2023
   medianage |      1509    7.945659    6.429558          0         19
   medianemp |      1509    208.8721    217.7851          1        930
     meanemp |      2200    219.9599    238.5172          0       1064
-------------+--------------------------------------------------------
      totemp |      2200    14889.42    55692.68          0     447760
     emp1emp |      2200    .3981818     3.00153          0         37
      nr1emp |      2200    .5981818    4.552797          0         55
    emp1year |      2200    27.91364    114.7474          0       1118
     nr1year |      2200    .1336364    .4193538          0          3
-------------+--------------------------------------------------------
    totemp_b |       708    39493.86    88380.17          0     447288
    nrunit_b |       708    197.1328    365.8306          1       2023
grosscreat~p |       708    1551.737    3523.801          0      18554
grossdestr~p |       708     1573.98    3473.278          0      18342
mediangrow~p |       669   -.0022592    .0252195  -.1498299   .0950634
-------------+--------------------------------------------------------
meangrowth~p |       707   -.0070959     .033665  -.3157895   .2222222
       sdemp |       759    60.09909    80.77371   1.154701   480.6484
medianempt_1 |       671    208.9225     221.065          2        759
  nrunit_hgf |       495    .0525253    .2570242          0          2
medianage_~f |       468    6.405983    5.192414          1         16
-------------+--------------------------------------------------------
  totemp_hgf |       495     12.4202    77.92159          0        806
 meanemp_hgf |        22       260.5    260.5565         51        806
```

The level 3 and level 4 datasets contain a larger number of observations because the information is now aggregated at the two-digit industry level. The peculiarity of these datasets is that generally the employment and the total number of units varies substantially across the different sectors, which in turn can significantly impact the quality of the resulting statistics. This is particularly true for those two-digit sectors in which only a handful of companies operate (for example, Coke and refined petroleum products). Level 3 and 4 datasets also include those sectors that are excluded from the macrosector classification, if data are available. Apart from that, the level 3 dataset substantially mirrors the level 2 dataset, and the level 4 mirrors the level 1 (although the age and size classification is less detailed in level 3 and absent in level 4 to avoid creating cells that are too scarcely populated).

```
. use dynemp_DYNEMPREP_unit_lev3.dta

. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        year |      4444    2006.271    3.075952       2001       2011
       group |         0
     ind_a38 |      4444      34.223    16.86608         10         62
   ageclass3 |      4444    2.161116    .8725489          1          3
  sizeclass4 |      4444    2.781503    1.011396          1          4
-------------+--------------------------------------------------------
      nrunit |      4444    17.26238    44.82848          1        504
nrunit_pos~p |      4444    17.21535    44.82607          0        504
   medianage |      2261    8.489164    6.314367          0         19
   medianemp |      2261    175.9876    164.5633          2        942
     meanemp |      4444    185.3234    200.6043          0       1700
-------------+--------------------------------------------------------
      totemp |      4444    3453.497    11052.55          0     125703
     emp1emp |      4444    .0915842    .5148537          0          7
      nr1emp |      4444    .1386139     .739639          0         10
    emp1year |      4444    6.666967    50.99793          0        841
     nr1year |      4444    .0310531    .1810979          0          2
-------------+--------------------------------------------------------
    totemp_b |      2077    6307.842    14847.41          0     125703
    nrunit_b |      2077    31.57246    59.01785          1        504
grosscreat~p |      2077    248.2956    606.2264          0       5450
grossdestr~p |      2077    251.9316    585.4253          0       5558
mediangrow~p |      1547   -.0060604    .0372372  -.2039604   .1485276
-------------+--------------------------------------------------------
meangrowth~p |      2068   -.0083272    .0536718        -.4   .2222222
       sdemp |      1937      77.768    99.41074          0   1197.132
 medianempt_1 |      1548    169.3346    165.4219          2        984
```

```
. use dynemp_DYNEMPREP_unit_lev4.dta

. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        year |      1090     2006.53    2.879187       2002       2011
       group |         0
     ind_a38 |      1090    33.24587    16.42128         10         62
         prc |      1090    10.92661    26.46886          1         99
      nrunit |      1090    60.16147    73.56073          1        550
-------------+--------------------------------------------------------
nrunit_pos~p |      1090    59.98991    73.69725          0        550
   medianage |      1027    14.46641    2.863971          9         19
   medianemp |      1027    133.2308      42.715          0        301
     meanemp |      1090    185.4678    69.63459          0   403.6667
      totemp |      1090    12019.62    14613.95          0     115252
-------------+--------------------------------------------------------
     emp1emp |      1090    .3275229    1.014167          0          7
      nr1emp |      1090    .4990826    1.189369          0          7
    totemp_b |      1090    12019.62    14613.95          0     115252
    nrunit_b |      1090    60.16147    73.56073          1        550
grosscreat~p |      1090    473.1284    666.0539          0       4374
-------------+--------------------------------------------------------
grossdestr~p |      1090    480.0569    668.8003          0       4268
mediangrow~p |      1000   -.0086147     .121486  -.2401266   .1858432
meangrowth~p |      1000   -.0091986    .1276161  -.2363445   .1917897
       sdemp |      1040    193.8029    64.97567          0   425.0154
medianempt_1 |      1027    133.4167    39.95901          0        336
```

Finally, the transition matrix dataset has a more complex structure: it embeds many different aggregation combinations (those listed in table 4) in the same file. When a classifying variable (for example, `ageclass4`, `f_sizeclass6`, or `sizeclass6`) is not used to aggregate units in a given aggregation combination, then it is set equal to "all" (codified with 9). The `ind_a38` row is instead set to missing when the sectoral classification is at macrosector level. Again, take particular care in identifying the desired aggregation level and in avoiding double counting.

```
. use dynemp_DYNEMPREP_unit_trans_mat.dta

. summarize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   macrosect |       2093     3.91591    3.251945          1          9
   ageclass4 |       2597    3.410089    3.407837          0          9
f_sizeclass6 |       2597    9.050058    11.65324          1         99
  sizeclass6 |       2597    5.228725    2.831839          1          9
   volat_emp |       2597    .0681643    .0355494          0   .2209131
-------------+---------------------------------------------------------
      totemp |       2597    25505.74    94151.18          0    1124745
    f_totemp |       2597    22905.52    94054.63          0    1124197
     JC_surv |       2597    1899.531    7925.003          0     109358
     JD_surv |       2597    1974.752    8114.145          0     108177
JC_surv_t~10 |       2597    949.0866    3965.031          0      56621
-------------+---------------------------------------------------------
jobvar_top10 |       2597    948.1482    3965.262       -141      56621
  nrunit_hgf |       1745    .0114613    .1167477          0          2
  totemp_hgf |       1745    2.370201    40.96046          0        893
f_totemp_hgf |       1745    4.116905    72.28561          0       1578
   medianemp |       2081    135.3123    121.5213          1        559
-------------+---------------------------------------------------------
 f_medianemp |       1235    129.0769    116.2504          1        527
medianemp_~f |          0
f_medianem~f |          0
      nrunit |       2597    127.4852    394.6773          1       5637
  meangrowth |       1433   -.0322861    .1111614  -.6365688   .5022831
-------------+---------------------------------------------------------
meangrowth~f |         18    .4665045    .1154008   .2909091   .5648855
           j |       2597     4.74047    1.670587          3          7
        year |       2597    2003.243    2.099802       2001       2007
    ind_a38 |        504    33.09127    16.45911         10         62
```

5 Acknowledgments

We thank Eric Bartelsman, Giuseppe Berlingieri, Timothy Destefano, Markus Eberhardt, Javier Miranda, Dirk Pilat, Stefano Scarpetta, Mariagrazia Squicciarini, Paul Schreyer, Colin Webb, and the national delegates to the Working Party of Industry Analysis of the OECD for useful comments and discussions, and Pekka Honkanen for excellent research assistance. Usual disclaimers apply.

6 References

Anyadike-Danes, M., C.-M. Bjuggren, S. Gottschalk, W. Hölzl, D. Johansson, M. Maliranta, and A. Myrann. 2013. Accounting for job growth: Disentangling size and age effects in an international cohort comparison. HUI Working Papers. http://www.hui.se/BinaryLoader.axd?OwnerID=3a577059-d869-4d1b-bff1-5673eefc3461&OwnerType=0&PropertyName=EmbeddedFile_fa365caa-458e-47e7-a992-12e92ecad11c&FileName=HUIwp84.pdf.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta. 2004. Microeconomic evidence of creative destruction in industrial and developing countries. Working Paper Series 3464, Policy Research, The World Bank.
http://elibrary.worldbank.org/doi/pdf/10.1596/1813-9450-3464.

———. 2013. Cross-country differences in productivity: The role of allocation and selection. *American Economic Review* 103: 305–334.

Bartelsman, E., S. Scarpetta, and F. Schivardi. 2005. Comparative analysis of firm demographics and survival: Evidence from micro-level sources in OECD countries. *Industrial and Corporate Change* 14: 365–391.

Bartelsman, E. J., and M. Doms. 2000. Understanding productivity: Lessons from longitudinal microdata. *Journal of Economic Literature* 38: 569–594.

Bravo-Biosca, A., C. Criscuolo, and C. Menon. 2013. What drives the dynamics of business growth? OECD Science, Technology and Industry Policy Papers: No. 1. OECD Publishing.

Criscuolo, C., P. N. Gal, and C. Menon. 2014. The dynamics of employment growth: New evidence from 18 countries. OECD Science, Technology and Industry Policy Papers: No. 14. OECD Publishing.

Davis, S. J., and J. Haltiwanger. 1990. Gross job creation and destruction: Microeconomic evidence and macroeconomic implications. In *NBER Macroeconomics Annual 1990*, ed. O. J. Blanchard and S. Fischer, 123–186. Cambridge, MA: National Bureau of Economic Research.

———. 1999. Gross job flows. In *Handbook of Labor Economics*, ed. O. Ashenfelter and D. Card, vol. 3B, 2711–2805. Amsterdam: Elsevier.

Davis, S. J., J. C. Haltiwanger, and S. Schuh. 1998. *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Dhyne, E., L. J. Álvarez, H. L. Bihan, G. Veronese, D. Dias, J. Hoffmann, N. Jonker, P. Lünnemann, F. Rumler, and J. Vilmunen. 2006. Price changes in the Euro area and the United States: Some facts from individual consumer price data. *Journal of Economic Perspectives* 20: 171–192.

Dickens, W. T., L. Goette, E. L. Groshen, S. Holden, J. Messina, M. E. Schweitzer, J. Turunen, and M. E. Ward. 2007. How wages change: Micro evidence from the international wage flexibility project. *Journal of Economic Perspectives* 21: 195–214.

Dunne, T., M. J. Roberts, and L. Samuelson. 1989. Plant turnover and gross employment flows in the U.S. manufacturing sector. *Journal of Labor Economics* 7: 48–71.

Eurostat and OECD. 2007. *Eurostat–OECD Manual on Business Demography Statistics*. Paris: OECD Publishing.

Hagsten, E., M. Polder, E. Bartelsman, G. Awano, and P. Kotnik. 2012. ESSnet on linking of microdata on ICT usage. http://ec.europa.eu/eurostat/documents/341889/725524/2010-2012-ICT-IMPACT-2012-Final-report.pdf.

Haltiwanger, J., R. S. Jarmin, and J. Miranda. 2013. Who creates jobs? Small versus large versus young. *Review of Economics and Statistics* 95: 347–361.

Syverson, C. 2011. What determines productivity? *Journal of Economic Literature* 49: 326–365.

Van Leeuwen, G., and M. Polder. 2013. Linking ICT related innovation adoption and productivity: Results from micro-aggregated data versus firm-level data. MPRA Paper 46479, University Library of Munich. http://mpra.ub.uni-muenchen.de/46479/1/MPRA_paper_46479.pdf.

Wada, R. 2005. outreg2: Stata module to arrange regression outputs into an illustrative table. Statistical Software Components S456416, Department of Economics, Boston College. http://ideas.repec.org/c/boc/bocode/s456416.html.

About the authors

Chiara Criscuolo is a senior economist at the OECD, in the Science Technology and Innovation Directorate, working on enterprise dynamics, entrepreneurship, productivity, and innovation. She is also a research associate with the Centre for Economic Performance at the London School of Economics.

Peter N. Gal is an economist at the OECD, working on the micro- and macroeconomic aspects of labor markets, productivity, and innovation. He is also a PhD candidate at the Tinbergen Institute and VU University Amsterdam.

Carlo Menon is an economist at the OECD in the Science Technology and Innovation Directorate. His main fields of activity are enterprise dynamics, innovation, and entrepreneurship. He is also an affiliate with the Spatial Economics Research Centre at the London School of Economics.