CENTRE for ECONOMIC PERFORMANCE

Occasional paper

42

Dynemp: A Stata® Routine for Distributed Micro-data Analysis of Business Dynamics

Chiara Criscuolo
Peter N. Gal
Carlo Menon

June 2014

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Abstract
This paper introduces a new Stata® command, *dynemp*, which implements a distributed micro-data analysis of business and employment dynamics and firm demographics. The data source it requires are business registers or comparable firm- or establishment- level longitudinal databases which cover the (near-) universe of companies in all business sectors. Access to such confidential data is usually restricted and the micro-level 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 harmonisation of the key economic concepts (gross job flows, growth rates of employment, definition of high-growth firms, etc.), dynemp can be distributed in a network of researchers who have access to the national confidential microdata. In such manner, the rich firm-level employment dynamics can be analysed from new angles (such as firm age and size), significantly expanding the scope of the analysis insofar possible using more aggregated data.

JEL classification: D22, L26, E24, L25
Key words: Employment dynamics, job flows, firm demographics

This paper was produced as part of the Centre's Productivity and Innovation Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

Acknowledgements

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

Chiara Criscuolo, economist at OECD and Associate at Centre for Economic Performance, London School of Economics. Peter N. Gal, OECD and a PhD candidate at the Tinbergen Institute and VU University Amsterdam. Carlo Menon, economist at OECD and Bank of Italy.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

TABLE OF CONTENTS

1. Introduction[1]

The Stata® command dynemp produces a set of statistics based on micro-level (firm or plant) employment data.[2] The information is aggregated to the level of industrial activities (sectors), age classes, size classes, and by segments of the employment growth distribution. While most industrialised 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 analysis of this rich source of data is often limited by confidentiality rules and by the lack of appropriate statistical platforms which would bring together the national databases from the different countries. The dynemp routine has been developed with the aim of providing a tool to produce non-confidential micro-aggregated dataset, while exploiting the richness of the firm-level databases used as the underlying sources.

In particular, dynemp may serve several purposes:

- It allows computing a number of indicators and summary statistics from micro-level 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;[3]
- It can serve as a platform to create a harmonised cross-country database.

The output database allows for the analysis of a wide number of 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 (e.g., entry *vs.* post-entry growth; contraction *vs.* exit);
- Characterising 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;
- 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.

The economic literature exploiting the richness of business-level data to explore employment dynamics was spurred by seminal publications in the 1990s, focusing mainly on the United States (Dunne

[1] We thank Eric Bartelsman, Giuseppe Berlingieri, Markus Eberhardt, Dirk Pilat, Stefano Scarpetta, Mariagrazia Squicciarini, Paul Schreyer, Colin Webb, and the national delegates to the Working Party of Industry Analysis (WPIA) of OECD for useful comments and discussions, and Pekka Honkanen for excellent research assistance.

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

[3] The programme does not control for secondary disclosure or for cases of predominance.

et al., 1989; Davis et al., 1998; Davis and Haltiwanger, 1990; 1999), presenting evidence of significant heterogeneity across different types of firms, which implies that the assumption of a "representative firm" is problematic when analysing questions related to employment and productivity.[4] Due to the inherent difficulties in accessing business-level data simultaneously in several countries, the first rigorous cross-country analysis of heterogeneous firm dynamics was only undertaken in the 2000s, published in Bartelsman et al. (2005). The paper studies firm demographics and survival across ten OECD countries, by collecting micro-aggregated data from national business registers based on a common data protocol. Further, a recent contribution by Haltiwanger et al. (2013) explores job creation and destruction dynamics in the United States, to find that young firms disproportionately contribute to job creation; while 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. A first step into that direction, using a simplified set of statistics compared to what is contained in this routine (DynEmp Express), is presented in Criscuolo, Gal and Menon (2014).

Dynemp requires a unit-level (firm or establishment) input dataset containing a longitudinal unit identifier, the calendar year, the 3-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 group of firms, classified along the age and size dimension, over a 3, 5, or 7 year time horizon; finally, the third set of results consists of *.xml* (Extensible Markup Language) tables reporting the output of regressions of the 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 year of birth of the unit. Ideally, the source should be the national business register (or possibly also 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 also involve information from the previous year, for instance when calculating gross job flows. Individual units need to be identified by a unique longitudinal identifier (id) that has to be constant over time. If the unit exits the firm-level dataset, its identifier must not reappear again.

More specifically, the required variables are:

- The number of employees, preferably measured in full-time equivalents, averaged over the year;
- The calendar year to which the time-varying variables refer to;
- The birth year of the unit. This can be of course outside (i.e. earlier than) the period covered by the business register. It should be constant over the whole 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

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

[5] The *.xml* tables are produced by using the user-written outreg2 Stata® command (Wada, 2008). For this reason, it is required that the user installs this package before running dynemp.

database - and birth year is missing -, the birth year is left missing and age is not defined (see more below);

- The 3-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 3 digit, it is automatically converted to 3 digit. The programme can also deal with the dataset being partially or completely classified according to the ISIC Rev. 3.1 (NACE Rev. 1.1) classification. In such cases, the options sectorchange, isic3, isic4, and newindyear need to be correctly specified (see below). In case only ISIC rev 3 or 3.1 is available, an external classification will be used, contained in the command package (the file named changeover_database.txt, which should be saved into the directory where the input data are stored). The sector variable must be an integer in numeric format. In addition, it is preferred that the sector is held fixed over time. If this is not the case, the programme will attribute to the unit its modal sector (selecting the most recent modes in case of multiple modes). See more details on this in Section 3.1.

- An optional variable is the year of left censoring for the birth variable. Note that the left-censoring variable may change across units, however it must be constant within units. If it is not the case, the user must replace the left-censoring variable with its minimum value. Note also that for those cases where birth year predates the censoring year, the programme assumes that the reported birth year is correct and does not apply any correction.

2.2. Stata® routine syntax

Below we present and explain the syntax of the dynemp command:

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

2.3. Options

- country(*string*) is required. It specifies the name of the country.

- unit(*string*) is required. It specifies the unit of analysis (e.g. plant or firm) id(*varname)* is required. It indicates the variable containing the unique longitudinal unit identifier. It can either be string or numeric.

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

- year(*varname*) is required. It indicates the year variable. It has to be an integer.

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

[6] ISIC stands for International Standard Industrial Classification of All Economic Activities, developed by the United Nations. NACE stands for *Nomenclature générale des Activités économiques dans les Communautés européennes*, i.e. the European industrial classification as used by Eurostat.

- isic3(*varname*) indicates the variable containing the unit's industry; this must follow the ISIC v.3 classification at the 3 or 4-digit level. It has to be an integer.

- isic4(*varname*) indicates the variable containing the unit's industry; this must follow the ISIC v.4 classification at the 3 or 4-digit level. Note that either this one or isic3(.) needs to be specified – or both, in case there is a change in classification over the sample period. If *isic4()* is left empty, the external conversion table named *changeover_database.txt*, include in the command package, is required. It has to be an integer.

- *sectorchange* must be specified in case some parts of the dataset are classified according to different classifications, i.e., a change in sectoral classification from ISIC v.3 to ISIC v.4 happens at a certain point in time. In such a case, both the industry variable options (isic3 and isic4) must be specified, although they can refer to the same variable. It requires the option newindyear to be also specified.

- newindyear*(#)* specifies the year in which the industrial classification changes from ISIC v. 3 to ISIC v.4. It requires the option sectorchange to be also specified. It has to be an integer.

- outputdir*(string)* specifies the output directory (e.g. C:\OECD\output). If not indicated, the output files will be saved in the Stata® working directory.

- *blank* sets to missing all the records referring to cells containing less units than the confidentiality level (option conf, see below).

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

- *express* runs a faster version of the code, which excludes the calculation of percentiles.

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

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

- extraformat(*string*) specifies additional formats for the output datasets. The allowed options are "txt" (tab-separated) and "csv"(comma-separated), which correspond to the file extensions.

- levels(*numlist*) limits the yearly flow datasets to the selected aggregation levels (see Table 1).

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

- exitchange(*varname*) identifies a binary variable (0/1) where 1s correspond to exit events due to change in legal status, e.g. M&A. The variable should be equal to one only in the last year of appearance of the unit.

- *noreg* tells the programme not to run distributed regressions.

- regyear(t*extitnumlist*) specifies the years over which the programme will run the regressions. The default is to run them for all available years in the data. E.g., if the chosen period is 2004-2008, the user should write 2004(1)2008.

- turnover(*varname*) identifies the variable containing turnover values. It has to be numeric.

2.4. System requirements

Dynemp should 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 (e.g., less than a million units), and within 5 hours for larger ones (4-5 million units), assuming a temporal extension of around 10 years.

3. Input data harmonisation and output datasets

3.1. Data cleaning

The programme carries out some basic consistency checks of the data and corrects observations which are considered implausible: it replaces negative values for employment when missing; it interpolates employment records that are disproportionally smaller/bigger than those of the previous and following year (threshold values are +/−1.5 change calculated as in formula (2) and at least 20 employees on average over the years $t - 1$, t, $t + 1$); it replaces industry classification that varies over time with the modal 3-digit sector the unit's activity is classified by. In case of multiple modes, the programme 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 are becoming more and more specialised and gain more importance, thus requiring a more detailed breakdown, while other activities which are becoming 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 some 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 2-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 detail (i.e. 2-, 3- and 4-digit).

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 as it simplifies many types of analyses which make use of the industry dimension. Finally, a constant industry classification per unit simplifies entry and exit definitions as 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:[7]

1. To convert the old classification to the new one, the routine creates a conversion table based on classifiers at the 3-digit detail. In overlapping years, i.e. typically in 2008 or 2009 or in both years, units are registered with both their old and their new classification; if such overlapping years do not exist in the database, the routine relies on units which exist in both systems, and create a link as follows: the observed value in the old system in 2008 will be paired with the new value in 2009.

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

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. In order to make the conversion more tractable, such *n* to *m* transitions 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. Since part of the industry classifications associated with a unit may come from the old system (i.e. before the changeover year) and another part from the new system (i.e. 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 $i_{temp,t}$ is created, which is defined to take the value of the new classification $i_{new,t}$ after the changeover year:

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

    while before the changeover year, it 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:

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

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

    Finally, if the unit does not have an observation in the new system (i.e. 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, i.e. 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 some 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 who 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 who 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.[8]

[8] In order to make the procedure replicable, the seed of the random number generator is always reset to the same number.

Employment, growth, birth, and exit definition

The programme calculates a few intermediate unit-level variables, which are subsequently used to calculate summary statistics at different aggregation levels in the final micro- aggregated ('collapsed') dataset. The programme runs regardless of whether the employment data is 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, i.e. that employment figures are not rounded to multiples of 5, 10, 100, etc.

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

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

where $L_{i,t}$ stands for employment of unit *i* in year *t*. The formula is commonly used in the business dynamics literature as 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 (i.e., it does not depend on the employment level at the beginning of the period) and is bounded between –2 and +2.[9]

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 in the programme, the calculation of the age variable will take this into account. Finally, the exit variable is dummy equal to one 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 flows statistics are calculated for three different groups of units: entrants, exitors, and incumbents. For each interval (t − 1, t), we define incumbents, entrants and exitors as follows:

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

3.2. Output datasets

The programme stores eight output databases in the output folder in .dta format:

- The aggregated statistics on yearly job flows, 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 3, 5, and 7 years horizon, averaged across years and 2-digit STAN A38 sectors. This is named *dynemp_'country'_'unit'_trans_mat.dta*.
- An excel file containing the distributed regression output tables, named

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

dynemp_'country'_ols.xls.

- An excel file containing the tabulation of gaps in the data, named *dynemp_'country'* _'*unit*'_tabgaps.txt.

Note: '*country*' is the country name selected in the required option; '*unit*' corresponds to the selected unit of analysis (e.g., plant/firm) selected in the required option; lev1 to lev4 identify the four different levels of aggregation, which arise from combinations of the sector, age, size, and employment growth classifications. The classification levels for the job flow datasets are reported in Table 1 while for the transition matrices datasets in Table 4.

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 four statistics are also calculated for the turnover variable if available. In order 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-OECD, 2007).[10]

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

Table 1. Aggregation levels in the annual job flow datasets

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 (~2 digit ISIC4/NACE2)		4 classes	3 classes
4	27 STAN A38 (~2 digit ISIC4/NACE2)	5 growth percentiles		

[10] "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-OECD, 2007).

Table 2: Industries in DynEmp

Macro-sectors	Included in DynEmp	Covered NACE2 / ISIC4 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 prodcuts [CB]
	●	16-18	Wood and paper products, and 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 non-metallic 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]
Non-financail 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, etc [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]
		84	Public administration and defence; compulsory social security [O]
	●	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]
		97-98	Activities of households as employers; undifferentiated activities of households for own use [T]
		99	Activities of extraterritorial organizations and bodies [U]

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

Source: 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; 500+; size classes considered in aggregation level 3 are: 0-9; 10-49; 50-249; 250+. Age classes considered in aggregation levels1, 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; top 10% of the distribution. This classification, however, may be problematic if a significant share of units in the reference group has zero growth, as all these units would end up in the same percentile groups. In order to avoid this, the percentile allocation is based on a growth rate which is increased or decreased by a random small number if the actual growth rate is equal to zero. The random number is drawn from a uniform distribution with maximum (minimum) value the (negative of) the minimum growth rate in the same country and calendar year.

Variables in annual flow datasets

Using the breakdowns above, the set of variables created are summarised in Table 3. Gross job flows are defined as below:

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

$$JC_{jt} = \sum_{i \in j} \Delta L_{it}^{+}$$

where *i* indexes units, ΔL_{it}^{+} is positive employment change from the previous year.
Job destruction (JD_{it}) measures the gross amount of jobs lost from period *t*-1 to *t*:

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

where $|\Delta L_{it}^{-}|$ is 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 – please see section on *Groups of Firms Considered*). Computed in levels 2 and 3.
prc	Aggregation according to percentiles of employment growth. Computed in levels 1 and 4.
group	Indicates which group the firm belongs to; incumbents, entrants, exitors.
meangrowthemp	Average growth in employment from t–1 to t.
meanemp	Average employment at time t for firms in the group.
meantrn	Average turnover in time t.
meangrowthtrn	Average growth in turnover from t–1 to t.
meanturnovemp	Mean turnover per employee.
medianage*	Median age of firms in the group.
emp1emp	Total employment of 1-employee units.
emp1year	Total employment of 1-year firms.
grosscreatemp	Gross job creation from t–1 to t.
grossdestremp	Gross job destruction from t–1 to t.
grosscreattrn	Gross turnover growth from t–1 to t.
grossdestrtrn	Gross turnover loss from t–1 to t.
medianemp*	Median employment of firms in the group.
medianempt_1*	Median employment of firms in the group at t–1.
mediangrowthemp*	Median growth in employment from t–1 to t.
mediangrowthtrn*	Median growth in turnover from t–1 to t.
mediantrn*	Median turnover in t.
medianturnovemp*	Median turnover per employee.
mediantrnt_1*	Median turnover at t–1.
nrunit_posemp	Number of units with employment greater than zero.
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.
nrunit_b	Number of units in the group with non-missing employment in both t and t-1 (defined for incumbents only).
p90p10turnovemp	Difference between the 90^{th} and 10^{th} 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.
totemp_b	Total employment at time t of units in the group with non-missing employment in both t and t-1 (defined for incumbents only).
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.
year	Reference year.

3.4. Transition matrices

The transition matrices summarise the growth trajectories of cohorts of units from year t to year t + j, where *t* takes the values 2001, 2004, and 2007, 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, 2004-2011; 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 a number of different combinations of age classes and size classes at time *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*	Ageclass at time *t*	Size class at time *t+j*	Sectors
All (non missing)	Entrants, 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 (non missing)	Entrants	All surviving, Missing employment, Exit	
All (non missing)	Entrants, all others	All surviving, Missing employment, Exit	2-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 also in the flow datasets, the command 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 programme computes the unit-level standard deviations of the employment growth rate over rolling windows of length *S* (with S = 3, 5, and 7):

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

where γ^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 γ^{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* over the period [*t*, *t + S*]:

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

Table 5. Variables in the transition matrices datasets

Variable name	Description
macrosect	Macrosector classification (manufacturing, services, construction, or all)
ageclass4	Age class
sizeclass6	Sizeclass in time t
f_sizeclass6	Sizeclass in the forward period
totemp	Total employment in time t
f_totemp	Employment in the forward period
totemp_hgf	Total employment in time t of high-growth firms
f_totemp_hgf	Employment in the forward period of high-growth firms
medianemp*	Median employment in time t
f_medianemp*	Median employment in the forward period
medianemp_hgf*	Median employment in time t of high-growth firms
f_medianemp_hgf*	Median employment in the forward period of high-growth firms
nrunit	Number of units in the group
nrunit_hgf	Number of high-growth firms in the group
meangrowth	Mean growth rate
meangrowth_hgf	Mean growth rate of high-growth firms
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 t to t+j
JD_surv	Gross job destruction from t to t+j
JC_surv_top10	Gross job creation from t to t+j - top 10% firms for employment growth
Jobvar_top10	Net job variation from t to t+j - top 10% firms for employment growth
j	Number of years ahead of t the forward period refers to
year	Reference year

* Medians are defined as the average value of the three central units.

3.5. Distributed regressions

Dynemp runs a series of unit-level regression on the full sample.

The first set of estimates consists in five Ordinary Least Squares (OLS) regressions with the growth rate as 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.[11]

[11] 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 one in the year 2008 and 2009, when the peak of the downturn was reached by most OECD countries.

The second set of regressions is based on a Linear Probability Model where the dependent variable is the "exit" dummy, and follows the same structure (although the model with age dummies only is excluded). Year and 3-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 the fit.

The third set of regressions is aimed at analysing 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 1, 3, and 5 years horizon is regressed over a set of dummies for different employment levels (8-9, 13-14, 18-19, 23-24, 48-49, 98-99), 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 the fit.

3.6. Confidentiality

The programme deals with confidentiality only if the "blank" option is specified. In that case, it performs a simple blanking of cells containing less units than the set number (default is 5 which can also be changed, see Section 2.3).

Also, 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 programme 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:

```
  obs:        15,675
 vars:             9                          20 Dec 2013 10:31
 size:       282,150
-------------------------------------------------------------------------------
              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                 turnover
censor          int     %9.0g                 censor year birth
-------------------------------------------------------------------------------
Sorted by:  idimp  yyear
```

where *idimp* is the longitudinal unit identifier, *yyear* denots calendar year, *empl* is the unit's total employment in the year indicated by the *yyear* variable, *rsect* is the 3- digit industry code (based on ISIC 3.1 until year 2007; and on ISIC 4 from 2008 onward), *birthyear* is the year of birth of the company, *bankrupt* is a dummy variable equal to one if the unit is last appearing in the dataset in that year because is

closing down, *MeA* is a dummy variable equal to one if the unit is last appearing in the dataset in that year because is being acquired or is merging with another unit, sales is the unit's turnover in the year indicated in the yyear variable, and censor is a variable that indicates the year of left censoring for the birth year variable in the business register.

The user should create an empty directory where she has writing permissions, in which the output datasets will be stored. If the data followed the ISIC 3 or NACE 1.1 sectoral classification for the entire period, the Changeover_database.txt file - which is part of the routine package - would need be saved in the folder containing the input data. The user should open the input dataset with the command use, then change the Stata® working directory to the one which 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 programme. Then, the dynemp command can be launched:

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

As explained in Section 3.1, the command converts ISIC 3 (or NACE 1) industry classification to ISIC 4 (NACE 2), creating a probabilistic conversion table. Note that the programme may need a few hours to run in a standard personal computer if the input data contains information on a few million units, as it is the case for business registers of large industrialised countries. During its execution, the programme 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 programme has finished, the following files are stored in the output folder:

1. *dynemp_DYNEMPREP_plant_lev1.dta*
2. *dynemp_DYNEMPREP_plant_lev2.dta*
3. *dynemp_DYNEMPREP_plant_lev3.dta*
4. *dynemp_DYNEMPREP_plant_lev4.dta*
5. *dynemp_DYNEMPREP_plant_trans_mat.dta*
6. *dynemp_DYNEMPREP_plant_regexit.txt*
7. *dynemp_DYNEMPREP_plant_regexit.xml*
8. *dynemp_DYNEMPREP_plant_reggrowth.txt*
9. *dynemp_DYNEMPREP_plant_reggrowth.xml*
10. *dynemp_DYNEMPREP_plant_sizecont.txt*
11. *dynemp_DYNEMPREP_plant_sizecont.xml*
12. *dynemp_DYNEMPREP_plant_sizecont2.txt*
13. *dynemp_DYNEMPREP_plant_sizecont2.xml*
14. *dynemp_DYNEMPREP_plant_tabgaps.txt*

The first four files contain the yearly flow data, with the variables listed in Table 3; the 5th file 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.

REFERENCES

Bartelsman, E., and M. Doms (2000), "Understanding Productivity: Lessons from Longitudinal Microdata" *Journal of Economic Literature*, American Economic Association, vol. 38(3): 569-594.

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(3): 365-391.

Criscuolo et al. (2014) The Dynamics of Employment Growth: New Evidence from 18 Countries. *OECD Science, Technology and Industry Policy Papers* no. 14, OECD Publishing. doi: 10.1787/5jz417hj6hg6-en

Davis, S. J., and J. Haltiwanger (1999), "Gross job flows" in: O. Ashenfelter & D. Card (ed.), *Handbook of Labor Economics*, edition 1, volume 3, chapter 41: 2711-2805. Elsevier.

Davis, S.J., J. Haltiwanger, R. Jarmin and J. Miranda (2007), "Volatility and Dispersion in Business Growth Rates: Publicly Traded versus Privately Held Firms" in *NBER Macroeconomics Annual 2006*, ed. D. Acemoglu, K. Rogoff, and M. Woodford, 107–55. MIT Press.

Davis, S. J., and J. Haltiwanger (1990), "Gross Job Creation and Destruction: Microeconomic Evidence and Macroeconomic Implications". NBER Chapters, in: *NBER Macroeconomics Annual 1990* Volume 5, pages 123-186

Dunne, T., M. J. Roberts, and L. Samuelson. (1989), "Plant turnover and gross employment flows in the U.S. manufacturings sector". *Journal of Labor Economics* 7 (1): 48-71.

Eurostat and OECD (2007), *Eurostat-OECD Manual on Business Demography Statistics*, OECD Publishing, Paris.

Haltiwanger, J., R.S. Jarmin, and J. Miranda. (2013). Who Creates Jobs? Small Versus Large Versus Young. *The Review of Economics and Statistics 95(2)*: 347-361

OECD/Eurostat (2008), *Eurostat-OECD Manual on Business Demography Statistics*, OECD Publishing. doi: 10.1787/9789264041882-en

Syverson, C. (2011), "What Determines Productivity?" *Journal of Economic Literature 49(2)*: 326-65.

Wada, R. (2008), outreg2: Stata ® module to arrange regression outputs into an illustrative table. Statistical Software Components, Boston College Department of Economics. Revised 19 March 2013. Downloadable from http://fmwww.bc.edu/repec/bocode/o.

CENTRE FOR ECONOMIC PERFORMANCE
Occasional Papers

41	Nicholas Bloom Renata Lemos Raffaela Sadun Daniela Scur John Van Reenen	The New Empirical Economics of Management
40	Ghazala Azmat Barbara Petrongolo	Gender and the Labor Market: What Have We Learned from Field and Lab Experiments?
39	Christopher J. Boyce Alex M. Wood James Banks Andrew E. Clark Gordon D.A. Brown	Money, Well-Being, and Loss Aversion: Does an Income Loss Have a Greater Effect on Well-Being than an Equivalent Income Gain?
38	Nicholas Bloom	Fluctuations in Uncertainty
37	Nicholas Oulton	Medium and Long Run Prospects for UK Growth in the Aftermath of the Financial Crisis
36	Phillipe Aghion Nicholas Bloom John Van Reenen	Incomplete Contracts and the Internal Organisation of Firms
35	Brian Bell John Van Reenen	Bankers and Their Bonuses
34	Brian Bell John Van Reenen	Extreme Wage Inequality: Pay at the Very Top
33	Nicholas Oulton	Has the Growth of Real GDP in the UK been Overstated Because of Mis-Measurement of Banking Output?
32	Mariano Bosch Marco Manacorda	Social Policies and Labor Market Outcomes in Latin America and the Caribbean: A Review of the Existing Evidence
31	Alex Bryson John Forth Minghai Zhou	What Do We Know About China's CEO's? Evidence from Across the Whole Economy

30	Nicholas Oulton	Horray for GDP
29	Stephen Machin	Houses and Schools: Valuation of School Quality through then Housing Market – EALE 2010 Presidential Address
28	John Van Reenen	Wage Inequality, Technology and Trade: 21st Century Evidence
27	Barry Anderson Jörg Leib Ralf Martin Marty McGuigan Mirabelle Muûls Laure de Preux Ulrich J. Wagner	Climate Change Policy and Business in Europe Evidence from Interviewing Managers
26	Nicholas Bloom John Van Reenen	Why do Management Practices Differ Across Firms and Countries?
25	Paul Gregg Jonathan Wadsworth	The UK Labour Market and the 2008-2009 Recession
24	Nick Bloom Raffaella Sadun John Van Reenen	Do Private Equity Owned Firms Have Better Management Practices?
23	Richard Dickens Abigail McKnight	The Impact of Policy Change on Job Retention and Advancement
22	Richard Dickens Abigail McKnight	Assimilation of Migrants into the British Labour Market
21	Richard Dickens Abigail McKnight	Changes in Earnings Inequality and Mobility in Great Britain 1978/9-2005/6
20	Christopher Pissarides	Lisbon Five Years Later: *What future for European Employment and Growth?*

For more information please contact the Publications Unit
Tel: +44 (0)20 7955 7673; Fax: +44 (0)20 7955 7595; Email: cep.info@lse.ac.uk
Website http://cep.lse.ac.uk