DynEmp
CIIE-WPIA project on employment and firm dynamics

Stata program *dynemp3*: Information and Guidance for Participating Countries[1]

This version: March 2024

Productivity, Innovation and Entrepreneurship Division
Directorate for Science, Technology and Innovation

dynemp@oecd.org

[1] Main contributors: Flavio Calvino, Chiara Criscuolo, Isabelle Desnoyers-James, Carlo Menon, Martin Reinhard and Rudy Verlhac.

Table of Contents

Brief summary of the outcomes

The Stata command *dynemp3* produces a set of statistics based on micro-level (firm/enterprise or plant/establishment) employment dynamics. The information is aggregated to the level of sectors, age classes, and size classes. In addition, there are two further optional levels of aggregation at the level of ownership status (independent, domestic or foreign group) and engagement in foreign trade (no trade, importer, exporter, or both).

The output files are stored in the "OUTPUT_TOSEND_*" folder within the declared output folder, if specified in the program option, or otherwise in the folder in which the input data are stored.

The output files can be classified into five groups:

1. One dataset containing transition matrices.
2. 15 datasets containing different aggregations of annual flows data, including statistics on gross job creation and destruction, average growth rate, average and median size, and age.
3. Up to nine excel tables (*.xlm) containing the output of the distributed regressions of growth rate and probability of exit on size and age dummies; average growth, share of stable, growing, shrinking firms and overall frequencies on size and age dummies; coefficients of the Pareto distribution of size (employment), descriptive statistics on three 3-digit sectors (activities of holding companies, head offices, and temporary employment agencies).
4. Up to eight text files reporting a tabulation of gaps in the data by macro sector of activity.
5. A number of graphs reporting density estimations and histograms of the employment growth distribution at different points in time.

Main policy relevant issues that *dynemp3* is designed to address

The output datasets will allow the analysis of a wide number of policy-relevant issues on enterprise dynamics, in particular they will allow to:

- Identify the groups of firms contributing most to job creation and destruction;
- Identify transition dynamics of cohorts of young firms;
- Assess the heterogeneous response of firms of different age, size and sector to the 2008 international financial crisis;
- Explore the extent to which firms differ in their employment growth performance within the same sector, size class, or age class;
- Assess whether firms engaged in international trade and/or member of a domestic or foreign owned business group tend to perform differently from other firms;
- Compare size and age distributions across countries; and the role of policies that might affect country specificities;
- Focus on the dynamics of ICT sectors.

Stata routine syntax

Below we present and explain the syntax of the *dynemp3* command.

dynemp3 [if exp] [in range], country(*string*) unit(*string*) id(*varname*) employment(*varname*) year(*varname*) birth(*varname*) empltype(*string*) empltypenr(*string*) emplperiod(*string*)

startyear(*string*) startmonth(*integer*) [forceyear isic3(*varname*) isic4(*varname*) sectorchange newindyear(*integer*) outputdir(*string*) blank conf(*integer*) dominance ntopdom(*integer*) express leftcensoring(*varname*) extraformat(*string*) levels(*numlist*) exitchange(*string*) entrychange(*string*) denovo noreg nograph turnover(*varname*) ownership(*varname*) dommne(*varname*) tradestatus(*varname*)]

Required options in the dynemp3 command:[2]

country(string)	Specifies the name of the country (please use a 3-digit ISO code).
unit(string)	Specifies the unit of analysis (e.g., plant or firm, please use firm if possible).
id(varname)	Indicates the variable containing the unique longitudinal unit identifier.
employment(varname)	Indicates the variable containing the unit's employment.
year(varname)	Indicates the year variable.
birth(varname)	Indicates the variable containing the unit's year of birth.
empltype(string)	Indicates if the employment variable is headcount (specify "HC") or full time equivalent (specify "FTE"). Headcount should be prioritised if available.
empltypenr(string)	Indicates if the employment variable records the number of employees (specify "EMPLOYEES") or the total number of persons engaged/total employment (specify "PERSENG"). Please use capital letters. Total number of persons engaged/total employment should be prioritised if available. If "EMPLOYEES" is specified, a correction is implemented on the employment variable (persons engaged = employees + 1).
emplperiod(emplperiod)	Indicates if the employment variable is recorded at a point in time ("PT") or as average over a period ("AV"). Average should be prioritised if available.
startyear(startyear)	*startyear* must be either "t" or "t-1". If *emplperiod* = PT, this option should be used to indicate which point in time the employment variable refers to (normally *t*). If *emplperiod* = AV, this option should be used to indicate whether the average is calculated starting from a month in *t* or *t–1*. See also the examples below.
startmonth(startmonth)	If *emplperiod* = PT, this option indicates the month to which the employment variable refers to. If *emplperiod* = AV, this option indicates the starting month over which the average is calculated. Only integer numbers between 1 and 12 are allowed.
Examples:	If employment reported in year *t* is an average between employment in January *t* and December *t*, please indicate emplperiod(AV) startyear(t) startmonth(1); if employment reported in year *t* is an average between employment in December *t–1* and November *t*, please indicate

[2] *Dynemp3* also requires the *egenmore*, *outreg2* and *tabout* Stata packages to be installed.

emplperiod(AV) startyear(t-1) startmonth(12); if the employment variable is recorded as a point in time in March of year *t*, indicate emplperiod(PT) startyear(t) startmonth(3).

If employment is recorded in a heterogeneous way across firms or the information is not available, please specify the *forceyear* option (see below) that excludes automatic corrections of year-related variables. However, the other options still need to be filled in. Please use information associated with the majority of firms in the database if available; if information is not available you can use emplperiod(AV) startyear(t) startmonth(1) forceyear. Please inform the DynEmp team about this.

Facultative options in the *dynemp3* command:

forceyear — The variables *year*, *birth, leftcensoring* are adjusted to proxy as much as possible the calendar year. If *emplperiod*=AV and *startyear*=t and *startmonth*>=8 these variables are shifted one year forward. If *emplperiod*=AV and *startyear*=t–1 and *startmonth*<=6 or if *emplperiod*=PT and *startyear*=t and *startmonth*<=3 these variables are shifted one year backwards. The option *newindyear* is adjusted consequently. The option *forceyear* excludes the automatic adjustment of these variables.

isic3(*varname*) — Indicates the variable containing the unit's industry classified according to the ISIC rev.3 classification at the 3 or 4-digit level (please see further details below).

isic4(*varname*) — Indicates the variable containing the unit's industry classified according to the ISIC rev.4 classification at the 3 or 4-digit level. Note that the user can specify *isic3()* only, *isic4()* only, or can specify 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* is required. This conversion table is sent together with the code, and it should be saved into the directory where the firm level data is stored.

sectorchange newindyear

Integer (sectorchange newindyear). Must be specified if the dataset is classified according to different industry classifications, i.e., a change in sectoral classification from ISIC rev.3 to ISIC rev.4 happens at a certain point in time within the sample period. In such a case, both *isic3()* and *isic4()* must be specified, although they can refer to the same variable; newindyear(*integer*) specifies the year in which the industrial classification changes from ISIC rev. 3 to ISIC rev.4 in the input data (i.e., the first year in which ISIC rev.4 is effective).

outputdir(string) — Specifies the output directory.

blank	Tells the program to automatically set to missing all the records referring to cells containing less units than the confidentiality level (option *conf*, see below).
conf(integer)	Sets a confidentiality level, i.e., the minimum number of units in a given cell. The command also makes the program to show the number of cells below such level on screen, as a preview of the number of cells that are likely to be blanked. The default value is 5.
dominance	Includes in the output an additional variable with dominance statistics (i.e., the share of employment and turnover, if inputted, of the N leading units in the cell total; option *ntopdom*, see below).
ntopdom(integer)	Indicates the number of leading units for which the dominance statistics should be calculated; the default value is 1. It can only be specified if the option *dominance* is also included.
express	Runs a faster version of the code, which excludes the calculation of percentiles. To be avoided unless really necessary!
leftcensoring(*varname*)	Indicates the variable reporting the year of left censoring in the business register.
extraformat(string)	Specifies the format for an additional output dataset that may facilitate confidentiality checks (the default is to produce only one output dataset in Stata .dta format). The allowed options are “txt” (tab-separated) and “csv” (comma-separated), which correspond to the file extensions.
levels(*numlist*)	Limits the calculations to the selected aggregation levels (available levels are 1 11 12 13 2 3 4 41 42 43 5 6 7 72 8).
exitchange(*varname*)	Identifies a binary variable (0/1) where one corresponds to an exit 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.
entrychange(*varname*)	Identifies a binary variable (0/1) where one corresponds to an entry due to change in legal status, e.g., M&A. The variable should be equal to one only in the first year of appearance of the unit.
denovo	Indicates that the inputted birth data has already been corrected for *de alio* entry events (e.g., acquisitions) and therefore the birth year should not be further corrected by the program.
noreg	Forces the program to skip distributed regressions.
nograph	Forces the program to skip histograms and density graphs of the employment growth distribution.
turnover(*varname*)	Identifies the variable containing sales (or turnover) values.
ownership(*varname*)	Is a numeric variable which takes the value of 1 if the firm is domestic and not part of a group (i.e., it is independent); it takes value of 2 if the firm is

part of a domestic-owned group; finally it is equal to 3 if the firm is part of a foreign-owned group.

dommne(*varname*) Is a binary variable that further refines the definition of domestic-owned groups. It should be equal to 1 if the domestic group is a domestic multinational group (i.e., it has affiliates abroad).

tradestatus(*varname*) Is a numeric variable indicating the unit's trade status: (1) no trade, (2) importer only, (3) exporter only, (4) both importer and exporter.

An auxiliary do file called *dynemp3_start.do* is also provided to show the data prerequisites, the necessary steps and the syntax to run *dynemp3*.

Basic requirements to run the program

The program requires minimal intervention by the participant, but it is strictly required that the input data respect a number of basic requirements.

1. The data source must be business register data including the universe of units belonging to the covered sectors (or similar data, e.g., based on social security or tax records).
2. Panel dimension: individual units need to be identified by a unique longitudinal identifier (*id*) that has to be constant over time. If the unit exits the business register, its identifier must not reappear again, since the statistics on job dynamics for a given year are based also on information on contiguous years.
3. Other required variables are:
    a. **Calendar year** to which the time-varying variables refer to.
    b. **Birth year of the unit**. This can be outside (i.e., earlier than) the period covered by the business register. It can also be missing for some units. In the latter case the first year in which the unit appears with non-zero employment in the register is considered as the entry year, unless this year coincides with the first year of the business register (in which case birth year is left as missing).
    c. **3 digit sector** identifying the main economic activity of the unit, following the ISIC rev. 4 or NACE rev. 2 classification.[3] If the sectoral classification is at 4-digit, it is automatically converted to 3-digit.
        i. The program can also deal with the dataset being partially or completely classified according to the ISIC rev. 3.1 (or NACE rev. 1.1) classification. In such cases, the options *sectorchange, isic3, isic4,* and *newindyear* need to be correctly specified (see above).
        ii. In case only NACE rev. 1.1 is available, an external classification will be used, sent together with the code (the file named *changeover_database.txt*, which should be saved into the directory where the firm level data are stored).
        iii. Participants are kindly required to convert the industry classification they use with appropriate conversion tables if their BR, or part of it, is not originally classified following the ISIC or NACE classification, but another

[3] NACE and ISIC have a one-to-one correspondence at 3-digit level.

system (such as NAICS, etc.). If only ISIC rev. 3 or 3.1 is available, please contact the DynEmp team, as a customised conversion table is required.

iv. The sector variable must be an integer and should not contain any letters, commas, or points.

v. In addition, it is preferred that the sector is held fixed over time, if this is not the case the program will attribute to the unit the modal sector (selecting the most recent modes in case of multiple modes) over the time period observed.

vi. Please see Table 1 for a summary of the sectoral classification options.

4. An optional variable is the year of left-censoring to be specified if the BR is left censored as *leftcensoring()*. Note that the left-censoring variable must not change over time for the same unit. If this is not the case, the user needs to create a new left-censoring variable that is constant over time within the unit and equal to the minimum value for that unit.4 The user should then input only the name of the latter variable into the program. For all those cases where birth year predates the censoring year, the program assumes that the reported birth year is the correct one and does not apply any correction.

5. System requirements: *dynemp3* should not require much more system memory (RAM) than the amount needed to load the input dataset. The computation time with a standard PC is around few hours for smaller datasets (around one million unit for 8-10 years), and within 5 to 15 hours for larger ones (e.g., 4-5 millions of units over several years).

Table 1 - Sectoral classification and syntax

Sectoral classification	Option syntax
NACE rev. 1.1 until year *N*; NACE rev. 2 from year *N+1*	isic3(*varname*) isic4(*varname*) sectorchange newindyear(*N+1*)
NACE rev. 2 only	isic4(*varname*)
NACE rev. 1.1 only	isic3(*varname*) + changeover_database.txt saved in the input data directory
ISIC rev. 3.1 until year *N*; ISIC rev. 4 from year *N+1*	isic3(*varname*) isic4(*varname*) sectorchange newindyear(*N+1*)
ISIC rev. 3.1 only	isic3(*varname*) + contact DynEmp team
ISIC rev. 4 only	isic4(*varname*)

Output

The program produces the following output files:

- The transition matrices, named *dynemp_'country'_'unit'_trans_mat.dta*
- The aggregated statistics on yearly job flows, named:
    - *dynemp_'country'_'unit'_lev1.dta*
    - *dynemp_'country'_'unit'_lev11.dta*
    - *dynemp_'country'_'unit'_lev12.dta*
    - *dynemp_'country'_'unit'_lev13.dta*

[4] In Stata, this can be easily done with the command 'bys *unit_id*: egen *newvar*=min(*oldvar*)'.

  - *dynemp_'country'_'unit'_lev2.dta*
  - *dynemp_'country'_'unit'_lev3.dta*
  - *dynemp_'country'_'unit'_lev4.dta*
  - *dynemp_'country'_'unit'_lev41.dta*
  - *dynemp_'country'_'unit'_lev42.dta*
  - *dynemp_'country'_'unit'_lev43.dta*
  - *dynemp_'country'_'unit'_lev5.dta* [optional]
  - *dynemp_'country'_'unit'_lev6.dta* [optional]
  - *dynemp_'country'_'unit'_lev7.dta*
  - *dynemp_'country'_'unit'_lev72.dta*
  - *dynemp_'country'_'unit'_lev8.dta*

- Up to nine excel files containing the distributed regression output tables, named:

  - *dynemp_'country'_'unit'_agedist.xml*
  - *dynemp_'country'_'unit'_pareto.xml*
  - *dynemp_'country'_'unit'_regexit.xml*
  - *dynemp_'country'_'unit'_reggrowth.xml*
  - *dynemp_'country'_'unit'_sizecont.xml*
  - *dynemp_'country'_'unit'_sizecont2.xml*
  - *dynemp_'country'_'unit'_regsect642.xml*
  - *dynemp_'country'_'unit'_regsect701.xml*
  - *dynemp_'country'_'unit'_regsect782.xml*

- Up to eight .txt files containing the tabulation of gaps in the data, named *dynemp_'country'_'unit'_tabgaps_ms*.txt*
- Several graphs in Portable Network Graphics (PNG) format depicting the shape of employment growth distribution in different years, both with histograms and kernel density estimations. Nine different graphs are drawn for each year in 1995, 1998, 2001, 2004, 2007, 2010 and 2012, when available. These are saved into the "GRAPHS" folder. See Annex 3 for further details.

Notes:

- 'country' is the country name selected in the required option.
- 'unit' corresponds to the selected unit of analysis (e.g., firm or plant) selected in the required option.
- *lev1* through *lev8* identify the different levels of aggregation, which arise from combinations of the sector, age, size, and employment growth classifications. A detailed description of the aggregation levels is given below. The level 5 and 6 datasets are produced only if the ownership status and trade status variables, respectively, are inputted.

Unit of analysis

For BRs containing information on different units of analysis (e.g., firm/enterprise or plant/establishment), the DynEmp routine should preferably be executed at least at the level of the firm. If data are available at the level of other units (plants; groups etc.), the program could be run on these datasets as well. The unit of analysis should be specified in the option unit. The output databases will be named accordingly. Statistical units are the preferred option.

Employment, growth, birth, and exit definition

For the sake of comparability, employment records should preferably be based at least on **headcounts** recording the **total number of persons engaged / total employment** if available. Note that if FTE is also available, it would be of great value if the countries could run the code twice, separately using HC and FTE measures of employment in the two separate runs. Employment can refer to either a **yearly average** or to a precise point in time, with a preference for yearly average if available. If the employment variable measures employees, the program adjusts it to proxy total employment by adding one proprietor / owner.

All this information should be specified in the options required when running the program, as discussed above. If more than one employment variable is available, participants are kindly asked to perform multiple runs, if possible.

The program will run regardless of whether the employment data is expressed as an integer or a decimal number. It is assumed that no additional rounding beyond that to unity is applied on the data.

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

$$GR_{it} = \frac{Employment_t - Employment_{t-1}}{0.5*(Employment_t + Employment_{t-1})} \qquad (1),$$

which is commonly used in the business dynamics literature as it has the advantage of not being biased by mean-reversion dynamics. 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. The index is calculated only on units that have non-missing employment in both years.

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 program the calculation of the age variable will take this into account. Exit is defined as the year after the last appearance of the unit with a positive employment level.

Definition of the employment growth percentiles

Employment growth classes are defined on five intervals of the growth distribution. The classes are divided as:

- units in the bottom 10% of the distribution (prc1);
- units between the 11th and 25th percentile (prc2);
- units between the 26th and 50th percentile (prc3);
- units between the 51st and 75th percentile (prc4);
- units between the 76th and 90th percentile (prc5);
- units in the top 10% of the distribution (prc6).

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. 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.

The employment growth percentiles are calculated both on the unweighted and weighted employment growth distributions, where the weights are the average employment level in the two periods (year *t-1* and *t*). This increases comparability with other related studies in the economic literature and allows for results cross-checking. The percentile values are then used to define cells in the year flow datasets (level 1, 11, 12, 13, 4, 41, 42 and 43), in combination with the age, macro sector, or A38 sector variables.

In addition, the percentiles groups are defined on both the full sample of all units, and on incumbent units only. In the former case, an assumption on the growth rate of entrants and exiting firms is necessary, as in the data the employment level of these units is missing either in *t-1* (entrants) or *t* (exiting units). Following a common practice in the relevant literature, the missing employment level is implicitly assumed to be zero. It follows from Equation (1) that entrants always have a growth rate of +2, and symmetrically exiting units have a growth rate of –2.

Employment growth percentiles are also used to compute employment growth dispersion variables collected in level 7 and 72. In these output databases, percentiles on which employment growth dispersion variables are based are calculated at the 3-digit industry level and then re-aggregated at A38 sector level. Aggregation of the employment growth dispersion variables is based on the average employment shares of 3-digit industries within each A38 sector.

Transition matrices: structure and statistics

The transition matrices contain information on units' size evolution from time *t* to time *t+j*, where *t* takes the values of year 1995, 1998, 2001, 2004, 2007, 2008, 2009, 2010, 2012, 2015 and 2018 and *j* is equal to 3, 5, 7, 10, or 14. Therefore, if data are available, the transition matrices are calculated for all the possible combinations of starting years *t* with the length intervals *j*. 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 employment growth rate) for a number of different aggregation and combination of age classes and size classes at time *t* and *t+j*.

The different combinations of the aggregation levels are reported in Table 2. Please see Annex 2 for a table with all statistics computed.

Table 2 - Size and age classes in transition matrices

Block	Size class at time *t*	Ageclass at time *t*	Size class at time *t+j*	Sectors	Ownership status	Trade status
1	0-1, 2-9, 10-19, 20-49; 50-99; 100-249, 250+**,** Missing employment	Entrants only	0-1,2-9, 10-19, 20-49; 50-99; 100-249, 250+, Exit, Missing employment	Macro sector classification (see Table 4); All macro sectors	All	All
2	0-1, 2-9, 10-19, 20-49; 50-99; 100-249, 250+**,** Missing employment	1-2, 3-5, 6-10, 11+, missing age	All surviving (excl. missing and 0-1), 0-1 Missing employment, Exit		All	All
2B	0-1, 2-9,10-19, 20-49; 50-99; 100-249, 250+**,** Missing	Entrants only	All surviving (excl. missing and 0-1), 0-1,		All	All

	employment					
3	All (excl. missing and 0-1), 0-1	Entrants, 1-2, 3-5, 6-10, 11+, missing age	All surviving (excl. missing and 0-1), 0-1, Missing employment, Exit		All	All
4	0-1, 2-9, 10-19, 20-49; 50-99; 100-249, 250+, Missing employment	All	All surviving (excl. missing and 0-1), 0-1, Missing employment, Exit		All	All
5	All (excl. missing and 0-1), 0-1	Entrants, all others	All surviving (excl. missing and 0-1), 0-1, Missing employment, Exit	STAN A38	All	All
6	All (excl. missing and 0-1), 0-1	Entrants, all others	All surviving (excl. missing and 0-1), 0-1, Missing employment, Exit	STAN A38	Independent, dom. group, foreign group, missing info.	All
7	All (excl. missing and 0-1), 0-1	Entrants, all others	All surviving (excl. missing and 0-1), 0-1, Missing employment, Exit	STAN A38	All	No trade, importer, exporter, both, missing info.

Note: the statistics computed for each transition group are also listed in Annex 2.

Aggregations in flow databases

The yearly job flows statistics are calculated for three different groups of units:

- entrants
- exitors
- incumbents

For each interval (*t–1, t*), an entrant is a unit that is not in *t–1* but is there in *t*; an exitor is a firm that is not there in *t* and is there in *t–1*. An incumbent is a unit that is there in *t-1* and *t*.5 Optionally, if the *entrychange* or the *exitchange* variables are inputted, in the yearly flow datasets the program aggregates units into four additional groups: entrychange, entrynew, exitchange, and exitdeath.

Size is defined on the average of employment at time *t–1* and *t* for incumbents, on employment at time *t–1* for exitors, and on employment at time *t* for entrants.

The aggregation levels considered in flow databases and the classification of incumbents in age and size classes are reported in Table 3. Please see Annex 2 for a list of statistics computed in the flow databases.

Table 3 – Aggregation levels in the flow datasets

Level	Sector	Growth percentiles	Group of units	Size	Age	Ownership	Trade status
1	Macro sector classification (see Table 4); All	6 growth percentiles unweighted	Incumbents	All	0-2; 3-5; 6+; All; Missing age	All	All
11	Macro sector classification (see Table 4); All	6 growth percentiles unweighted	All	All	0-2; 3-5; 6+; All; Missing age	All	All
12	Macro sector classification (see Table 4); All	6 growth percentiles weighted	Incumbents	All	0-2; 3-5; 6+; All; Missing age	All	All
13	Macro sector classification (see Table 4); All	6 growth percentiles weighted	All	All	0-2; 3-5; 6+; All; Missing age	All	All
2	Macro sector classification (see Table 4); All		Entering, Exiting, Incumbents	0-1; 2-9; 10-49; 50-99; 100-249; 250-499; 500+; All (excl. 0-1); missing	0-2; 3-5; 6+; All; Missing age	All	All
3	27 STAN A38 (~2 digit ISIC rev.4/NACE rev.2)		Entering, Exiting, Incumbents	0-1; 2-9; 10-49; 50-249; 250+; All (excl. 0-1); missing	0-2; 3-5; 6+; All; Missing age	All	All
4	27 STAN A38 (~2 digit ISIC rev.4/NACE rev.2)	6 growth percentiles unweighted	Incumbents	All	All	All	All
41	27 STAN A38 (~2 digit ISIC rev.4/NACE rev.2)	6 growth percentiles unweighted	All	All	All	All	All
42	27 STAN A38 (~2 digit ISIC rev.4/NACE rev.2)	6 growth percentiles weighted	Incumbents	All	All	All	All
43	27 STAN A38 (~2 digit ISIC rev.4/NACE rev.2)	6 growth percentiles weighted	All	All	All	All	All
5	27 STAN A38 (~2 digit ISIC rev.4/NACE rev.2)		Entering, Exiting, Incumbents	0-1; 2+; missing	0-5; 6+; All; Missing age	Independent, dom. group,[a] foreign group, missing info.	All
6	27 STAN A38 (~2 digit ISIC rev.4/NACE		All	0-1; 2+; missing	0-5; 6+; All; Missing	All	No trade, importer, exporter,

	rev.2)				age		both, missing info.
7	27 STAN A38 (~2 digit ISIC rev.4/NACE rev.2)	based on growth percentiles unweigthed	Incumbents	All	All	All	All
72	27 STAN A38 (~2 digit ISIC rev.4/NACE rev.2)	based on growth percentiles weigthed	Incumbents	All	All	All	All
8	ICT and content/media sectoral classification		Entering, Exiting, Incumbents	0-1; 2-9; 10-49; 50-249; 250+; All (excl. 0-1); missing	0-5; 6+; All; Missing age	All	All

[a] If information is available, this level will include an additional variable (dom_MNE) indicating whether the domestic group is a domestic multinational group (i.e., it has affiliates abroad).

Distributed regressions

The *dynemp3* code runs a series of firm-level regression on the full sample of units.

The first set of estimates is composed by seven Ordinary Least Squares (OLS) regressions with the growth rate as dependent variable, and the following sets of dummies as independent variables: size; age; size-age; size-age interacted with the "great recession" (2008-2009) dummy; size-age interacted with the "hi-tech sector" dummy; age interacted with the "great recession" (2008-2009) dummy. Two different age classifications are used. Year and 3-digit sector fixed effects are included in all specifications. These estimates are stored in the file *dynemp_'country'_'unit'_reggrowth.xml*.

The second set of regressions is based on a Linear Probability Model and the dependent variable is the "exit" dummy, and follows a similar structure (although the model with age dummy only is excluded and only one age classification is used). Year and 3-digit sector fixed effects are included in all specifications. These estimates are stored in the file *dynemp_'country'_'unit'_regexit.xml*.

The third set of regressions is aimed at analysing the effect of size-contingent policies on firm 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 threshold in certain countries. Year and 3-digit sector fixed effects are included in these specifications. Second, for each employment level from 1 to 100, the share of shrinking, growing, and stable firms; the average employment growth rate; the count of firms, and of firms with non-missing growth rate in A38 industries are regressed over a full set of employment-level dummies. Units with employment = 1 are excluded. Regressions are estimated separately for every year. These estimates are stored in the files *dynemp_'country'_'unit'_sizecont.xml* and *dynemp_'country'_'unit'_sizecont2.xml*.

The fourth set of regressions is similar in spirit to the latest models presented. However, it focuses on the age distribution rather than the size distribution of units. In particular, for each age level from 1 to 30, the share of shrinking, growing, and stable firms; the average employment growth rate; the average size; sectoral total employment; the count of firms, and of firms with non-missing growth

rate in A38 industries are regressed over a full set of age-level dummies. This is done for available age bins. Regressions are estimated separately every year. Units with employment = 1 are excluded. These estimates are stored in the file *dynemp_'country'_'unit'_agedist.xml*.

The fifth set of regressions focuses on three 3-digit sectors of the economy (Activities of holding companies, Activities of head offices, and Temporary employment agency activities; sectors 642, 701 and 782, respectively) and extracts some descriptive statistics for the units in these sectors. Namely, for each of these sectors, the employment and turnover of units in the sector are regressed against a set of year dummies. Furthermore, for each of the sectors, the sectoral number of units and Herfindahl-Hirschman (HH) employment concentration index (both perturbed by a small shock) are also regressed over a set of year dummies. This is repeated using all units first, and second all units with employment higher than 1. These estimates are stored in the files *dynemp_'country'_'unit'_regsect642.xml*, *dynemp_'country'_'unit'_regsect701.xml*, and *dynemp_'country'_'unit'_ regsect782.xml*.

The sixth set of regressions focuses on the shape of the employment distribution. It estimates a Pareto coefficient focusing on different parts of the employment distribution (estimates are computed for all firms with more than 1 employee; for firms that have employment larger than the median employment; larger than the 90th percentile; larger than the 95th percentile). The estimation technique used is based on an OLS log rank corrected methodology. Estimates are computed at different points in time (the years considered are the same years available in the transition matrix database). These results are stored in the file *dynemp_'country'_'unit'_pareto.xml*.

The output tables contain only the coefficients on the relevant fixed effects, the number of observations and some statistics on the quality of the fit.

Summary of data cleaning

The program carries out some adjustments, basic consistency checks and cleaning of the data:

- It adjusts the employment variable if it records employees (and not the total number of persons engaged / total employment). The adjustment implemented is total employment = employees + 1. This corresponds to assuming on average one proprietor / owner.
- It shifts year-related variables to match the calendar year as much as possible. See the description of the *forceyear* option for additional details.
- It replaces negative employment to missing.
- It drops units that have missing or zero total employment (PERSENG) in all periods.
- 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 (1) 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 program chooses the most recent mode.
- It replaces the birth year variable with the minimum value in case it changes over time for the same unit, and it changes the birth year variable to the modal value of the same unit if it is missing in a few years.

- It switches the birth year to the first year with positive employment when the birth year is missing or within the observed period but with zero or missing employment (and the *denovo* option is omitted).

Confidentiality

The program deals with confidentiality only if the "blank" option is specified; in such a case, it performs a simple blanking of cells which have less than the minimum threshold of units in the "levels" and transition matrices. The blanking is based on the number of units and accounts for sub-populations within a cell (e.g., high-growth firms, top 10% of firms). In the case of sub-populations, the unit counts themselves are not blanked.

Participants are kindly asked to check the data produced by the program and to blank cells according to their national confidentiality rules. Code to produce additional diagnostic statistics (e.g., additional variable specific counts and dominance statistics) or to blank unit counts can be provided on request.

All median values are calculated as the average of the five "central" units in the distribution of interest. In such a way, no information referring to an individual unit is disclosed. The number of central units can be increased at the request of those participants who may require it.

If the option *dominance* is included, the program can also compute statistics on dominance, i.e., on the share of employment or turnover account for the biggest N unit in the cell, where N can be inputted by the user (option *ntopdom*) and it is set to one by default.

Further information

For further information or support to run the program, please write to dynemp@oecd.org.

Thank you very much for your participation in the OECD DynEmp project.

Annex 1 – Sectoral classification

Table 4 – Sectoral aggregation

STAN A38 aggregation based on ISIC rev.4 classification	STAN A7 macro sector
01 to 03 AGRICULTURE, FORESTRY AND FISHING	**Agriculture, forestry and fishing**
05 to 09 Mining and quarrying	**Mining and quarrying**
10 to 12 Food products, beverages and tobacco	**Manufacturing**
13 to 15 Textiles, wearing apparel, leather and related products	
16 to 18 Wood and paper products, and printing	
19 Coke and refined petroleum products	
20 Chemicals and chemical products	
21 Basic pharmaceutical products and pharmaceutical preparations	
22 to 23 Rubber and plastics products, and other non-metallic mineral products	
24 to 25 Basic metals and fabricated metal products, except machinery and	
26 Computer, electronic and optical products	
27 Electrical equipment	
28 Machinery and equipment n.e.c.	
29 to 30 Transport equipment	
31 to 33 Furniture; other manufacturing; repair and installation of machinery and equipment	
35 Electricity, gas, steam and air conditioning supply	**Electricity, gas, water and waste**
36 to 39 Water supply; sewerage, waste management and remediation activities	
41 to 43 CONSTRUCTION	**Construction**
45 to 47 Wholesale and retail trade, repair of motor vehicles and motorcycles	**Non-financial market services**
49 to 53 Transportation and storage	
55 to 56 Accommodation and food service activities	
58 to 60 Publishing, audiovisual and broadcasting activities	
61 Telecommunications	
62 to 63 IT and other information services	
64 to 66 FINANCIAL AND INSURANCE ACTIVITIES	*Excluded*
68 REAL ESTATE ACTIVITIES	**Non-financial market services** (continued)
69 to 71 Legal and accounting activities; activities of head offices; management consultancy activities; architecture and engineering activities; technical testing and analysis	
72 Scientific research and development	
73 to 75 Advertising and market research; other professional, scientific and technical activities; veterinary activities	
77 to 82 Administrative and support service activities	
84 Public administration and defence	*Excluded*
85 Education	**Non-market services**
86 to 88 Human health and social work activities	
90 to 93 Arts, entertainment and recreation	
94 to 96 Other service activities	
97 to 99 Households and extraterritorial activities	*Excluded*

Table 5 – ICT, content and media sector classification

Group of sectors	Sector 3–digit code (ISIC rev.4)
ICT manufacturing	261-64; 268
ICT services	582; 611-19; 620; 631; 951
Content and media	581; 591; 592; 601-602; 639
Other manufacturing	See the definition in Table 4 excluding ICT manufacturing
Other non-financial market services	See the definition in Table 4 excluding ICT services
Others	Other sectors in none of the groups above (see Table 4 for excluded sectors)

Note: See "Information Economy – Sector Definitions based on the International Standard Industry Classification (ISIC rev.4)", OECD document DSTI/ICCP/IIS(2006)2/FINAL and UN ISIC rev.4 manual (United Nations, "International Standard Industrial Classification of All Economic Activities, revision 4", Statistical Paper, Series M No. 4/Rev.4, 2008). ICT trade industries are not considered separately as this would require a finer level of aggregation (4 digit).

Annex 2 – Calculated variables

The tables below list the calculated variables in the flow databases and in the transition matrices, respectively.

The content of the specified options *emplperiod, empltype, emlptyenr, startmonth, startyear* are also stored as variables in the output databases. Some statistics on the censoring year are also incorporated as variables (see *meancensoring* and *sdcensoring*).

Flow data:

Table 6 – Variables contained in the flow datasets

Variable name	Description
ageclass*	age class (note: different age classifications, see Table 3).
dispersion*IJ*	employment growth dispersion (percentile *I* – percentile *J*) - unweigthed
dispersion*IJ*_w	employment growth dispersion (percentile *I* – percentile *J*) - weigthed
dom_MNE	whether units are members of a domestic group which is a multinational entity
emp1emp	total employment of 1-employee units
emp1year	total employment of 1-year firms
grosscreatemp	gross job creation from t–1 to t
grosscreattrn	gross turnover growth from t–1 to t
grossdestremp	gross job destruction from t–1 to t
grossdestrtrn	gross turnover loss from t–1 to t
group	indicates which group the firm belongs to: incumbents, entrants, exitors, etc.
hhemp	Herfindahl–Hirschman concentration index of employment
hhtrn	Herfindahl–Hirschman concentration index of turnover
ict_sect	ICT and content/media sectoral classification (see Table 5)
ind_a38	2-digit sector STAN A38 classification (see Table 4)
macrosector	macro sector classification (see Table 4)
meanage	average age of firms
meancensoring	mean of year of left censoring in the cell
meanemp	average employment at time t for firms in the group

meanemp_hgf	mean employment in high-growth firms
meangrowthemp	average growth in employment from t–1 to t
meangrowthemp_w	weighted average growth in employment from t–1 to t
meangrowthtrn	average growth in turnover from t–1 to t
meangrowthtrn_w	weighted average growth in turnover from t–1 to t
meangrowthtrnovemp	average growth of turnover/employment ratio (labour productivity)
meantrn	average turnover in time t
meantrn_hgf	average turnover in time t in high-growth firms
meanturnovemp	mean turnover per employee
medianage	median age of firms in the group
medianage_hgf	median age of high-growth firms
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
mediantrnt_1	median turnover at t–1
medianturnovemp	median turnover per employee
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 with empl. defined at both t and t-1 in the group
nrunit_c	number of units with empl. >0 defined at both t and t-1 in the group
nrunit_hgf	number of high-growth firms
nrunit_posemp	number of units with employment greater than zero
owner_group	ownership status
p90p10turnovemp	difference between the 90th and 10th percentiles in turnover per employee
prc	aggregation according to percentiles of employment growth.
sdage	standard deviation of age of firms
sdcensoring	standard deviation of year of left censoring in the cell
sdemp	standard deviation of employment at time t
sdgrowthemp	standard deviation of employment growth
sdgrowthtrn	standard deviation of turnover growth
sdtrn	standard deviation of turnover at time t
sdtrnovemp	standard deviation of turnover per employee at time t
sizeclass*	size class (note: different size classifications, see Table 3).
totemp	total employment at time t
totemp_b	total employment at time t of units with empl. defined at both t and t-1
totemp_c	total employment at time t of units with empl.>0 defined at both t and t-1
totemp_domN	dominance statistics for employment based on the top N units in the cell
totemp_hgf	total employment in high-growth firms
tottrn	total turnover at time t
tottrn_hgf	total turnover in high growth firms
trade_group	trade status
trnovemp_hgf	turnover/employment ratio (labour productivity) of high-growth firms
turnover_domN	dominance statistics for turnover based on the top N units in the cell

year	reference year

Transition matrices:

Table 7 – Variables contained in the transition matrices

Variable name	Description
ageclass4	age class
block	identifier of the transition matrix partition
dommne	only for domestic group - whether MNE or not
employment_domN	share of employment at time t of N leading units
f_employment_domN	share of employment at time t+j of N leading units
f_medianemp	median employment in the forward period
f_medianemp_hgf	median employment of high growth firms in the forward period
f_sizeclass6	size class in the forward period
f_totemp	employment in the forward period
f_totemp_hgf	employment of high growth firms in the forward period
f3_intemp	employment at time t+3 of survivors from t to t+5, t+7, or t+10 or t+14
f5_intemp	employment at time t+5 of survivors from t to t+7 or t+10 or t+14
f7_intemp	employment at time t+7 of survivors from t to t+10 or t+14
f10_intemp	employment at time t+10 of survivors from t to t+14
ind_a38	2-digit sector STAN A38 classification (see Table 4)
j	number of years ahead of t the forward period refers to
JC_surv	gross job creation from t to t+j
JC_surv_top10	gross job creation from t to t+j - top 10% firms for employment growth
JD_surv	gross job destruction from t to t+j
jobvar_top10	net job variation from t to t+j - top 10% firms for employment growth
macrosect	macro sector classification (see Table 4)
meancensoring	mean of year of left censoring in the cell
meangrowth	mean employment growth rate
meangrowth_hgf	mean employment growth rate of high growth firms
medianemp	median employment in time t
medianemp_hgf	median employment of high growth firms in time t
nrunit	number of units in the group
nrunit_hgf	number of high growth firms in the group
nrunit_posemp	number of units in the group with >0 employment
ownership	ownership status
sdcensoring	standard deviation of year of left censoring in the cell
sdgrowth	standard deviation of the growth rate
sizeclass6	size class in time t
totemp	total employment in time t
totemp_hgf	total employment of high growth firms in time t
tradestatus	trade status
volat_emp	employment growth volatility, calculated at firm level and averaged at cell level
volat_trn	turnover growth volatility, calculated at firm level and averaged at cell level
volatunw_emp	employment growth volatility, calculated at firm level and averaged (unweighted) at cell level
volatunw_trn	turnover growth volatility, calculated at firm level and averaged (unweighted) at cell level
year	reference year

Annex 3 – List of figures

Table 8 - List of figures

Graph name	Description
density_growth_job_flows_`year'_all	Density (kernel estimation) of employment growth for all firms (including entry and exit). Employment growth is computed with the job flows measure.
density_growth_job_flows_`year'_inc_nozero	Density (kernel estimation) of employment growth for incumbents only and excluding zero employment. Employment growth is computed with the job flows measure.
density_growth_log_diff_`year'_inc_nozero	Density (kernel estimation) of employment growth for incumbents only and excluding zero employment. Employment growth is computed as log difference.
hist_unw_growth_job_flows_`year'_all	Histogram of employment growth for all firms (including entry and exit), with equal weight for all firms. Employment growth is computed with the job flows measure.
hist_unw_growth_job_flows_`year'_inc_nozero	Histogram of employment growth for incumbents only and excluding zero employment, with equal weight for all firms. Employment growth is computed with the job flows measure.
hist_unw_growth_log_diff_`year'_inc_nozero	Histogram of employment growth for incumbents only and excluding zero employment, with equal weight for all firms. Employment growth is computed as log difference.
hist_w_growth_job_flows_`year'_all	Histogram of employment growth for all firms (including entry and exit), weighted by firms' employment. Employment growth is computed with the job flows measure.
hist_w_growth_job_flows_`year'_inc_nozero	Histogram of employment growth for incumbents only and excluding zero employment, weighted by firms' employment. Employment growth is computed with the job flows measure.
hist_w_growth_log_diff_`year'_inc_nozero	Histogram of employment growth for incumbents only and excluding zero employment, weighted by firms' employment. Employment growth is computed as log difference.