CIIE-WPIA Project on the micro-drivers of aggregate productivity

MultiProd (v2.1)
Information and Guidance for Participating Countries

Last update of document: November 2025

Productivity, Innovation and Entrepreneurship Division

Directorate for Science, Technology and Innovation

MultiProd@oecd.org

Table of contents

Overview

The MultiProd code is used in the context of a *distributed microdata analysis*. This is a method of collecting statistical moments by a centrally written routine that is flexible and automated enough to run across datasets in different countries. It was pioneered in the beginning of the 2000s in a series of cross-country projects on firm demographics and productivity (Bartelsman et al., 2005, 2009), some of them with OECD involvement. The *MultiProd* project contributes to the analysis of productivity and business dynamics, together with the *DynEmp* distributed microdata project, which relies on business registers to investigate employment dynamics and firm demographics (see e.g., Criscuolo et al., 2014a and 2014b).

The latest version of the code is the version 2.1, released in April 2024.

The program *MultiProd* produces a set of statistics based on longitudinal micro-level (firm or plant) production information. The essential data requirements are information on output (gross output and value-added), inputs (labour, investment/capital, intermediates), and sector of activity. These unit-level data, along with information on labour costs, firm age and ownership, are used to generate detailed statistics on the level, dispersion, and dynamics of labour and multifactor productivity, wages, and mark-ups, aggregated by sector, age and size classes as well as detailed productivity groups, among other breakdowns.

The program relies on two key sets of information:

1. Production Data (*ProdData*), from either a production survey or administrative production data (e.g., from firm balance sheets), which contains all the variables needed for the productivity analysis but may be limited to a sample of firms; and
2. Population Data (*PopData*), from either a Business Register or comprehensive administrative data (e.g., social security data), which may contain a more limited set of variables (in particular, employment and sector) but for the entire population of firms.

These two data sources may be kept separate, may be combined into a single data file, or, in the case of comprehensive administrative data, may indeed come from a single original source (e.g., business tax returns).

In the case where full-coverage *ProdData* is not available, the presence of *PopData* substantially improves the representativeness of the results and thus their comparability across countries. In particular, conditional on the coverage and continuity of the information collected for each firm, *PopData* allows for: i) the calculation of the population breakdowns that are necessary for obtaining the weights used in the analysis; ii) a much more precise treatment of entry and exit; and iii) where relevant, improved identification of industries in the case of a change in the underlying industry classification. **All participants in the *MultiProd* project are therefore warmly encouraged to use comprehensive population data from either a business register or administrative sources for the project**.

Ideally, population and production data will be linked together to enable *PopData* to be used to identify firm entry and exit. If the two sources cannot be merged, *PopData* is used independently to inform the creation of weights and the industry classification.

Participants are also kindly asked to fill as comprehensively as possible the metadata questionnaire provided with the code. This metadata information will be used to i) ensure that the most appropriate options of the MultiProd code are used to achieve high output quality; ii) ensure that the resulting output is used correctly and to inform cross-country comparability; iii) publish metadata documents to accompany analyses.

Contacts and support

The MultiProd team is available to help with any queries you have in relation to the code, the data requirements, confidentiality, or any other aspects of the project. Although the code has been tested extensively on the data we hold in-house, and piloted on several national datasets, every dataset has its own quirks that may affect how the code runs. In addition, many countries have specific requirements with respect to confidentiality, beyond the standard approach implemented in the code. If you have any questions or concerns, please get in touch with us at MultiProd@oecd.org.

What's new with respect to MultiProd v2.0

Some new features have been introduced to the MultiProd code, to extend and refine analyses, improve measurement and harmonisation, and to account for feedback received by participants. Please contact the team for further information on these updates or to discuss adjustments relevant to your country/data.

The MultiProd Team welcomes your feedback on these changes and novelties and suggestions for further improvements!

Flexibility of the code, run-time and streamlining

1) First, a number of options and revisions to the code have been implemented to **improve the modularity and flexibility of the code** and to tailor its execution to the needs of the project and the country specificities (including the nature of the data and the resources available to countries).
    - Option *MinimalRun*: This new option can be set to 1 to **produce a minimum set of output**. This option is useful to test the specification and produce a first set of output that can be used to assess data quality. This is recommended for instance for a first run of the code notably on large microdata or for runs in environments with more limited computational capacity. See additional information in the section Modularity and additional settings.
    - Options *runStatistics_toExclude* or *runStatistics_toInclude.* These mutually exclusive options can be used to **select some specific output** to exclude from the run or on the contrary to select only specific output to include in the run. This can be useful to tailor the set of statistics to produce when it is not feasible to compute or to share the entire MultiProd output. Please contact the team if you wish to use these options. These options cannot be used with *MinimalRun*.
    - *varDefinitions_only:* When switched on (i.e., set to 1) this option enables to **create only the micro-data** prepared for the analysis. This can be used to prepare the data for related work on productivity, or to run with complementary modules. Participants can use it to obtain

micro-data ready for productivity analysis according to the MultiProd methodologies. It can also be used to split up the run of MultiProd in case of limitations to computational capacity, in conjunction with the option ResumeRun (see below). Please contact the team if you wish to use this option.

- *ResumeRun* option: when switched on (i.e., set to 1) this option, that enables to **resume a previous run**, has been improved and made more robust. This option can notably be used after a first restricted/minimal run (cf. option *MinimalRun*) or initial definition of variables (e.g., using varDefinitions_only) to produce the rest of the MultiProd output after data have been validated on the restricted run. ResumeRun can also be used to continue a run of MultiProd that unexpectedly stopped, e.g., due to a server time-out or reboot.
  Note that in case markups were not computed in the previous run and are set to be computed in the resumed run, this option enables to skip parts of the data preparation (i.e., computation of population breakdown, data cleaning, computation of other measures and transformations are taken from the previous run) to add only markups. However, at the moment, in this case all the micro-aggregated data output files need to be regenerated to also include markups.
- *threshold_L_PS* option: if your production data comprehensively covers only firms above a certain size threshold in terms of the main labour variable, you can specify this threshold here. With this option specified, it is no longer necessary to perform any related adjustments on your data that you may have been asked to perform in the past (e.g., dropping selectively covered firms below the threshold or imposing a similar threshold to the business register, if available). The MultiProd routine now performs these adjustments internally in a harmonised way. Generally, the use of this option will ensure that the MultiProd output is restricted to firms above the specified threshold. Please contact the team if you wish to use this option.
- *Asset_Deflator* option: this option allows for more precise deflation of capital stocks depending on what is measured by the main capital and investment variables, if available. In the preferred case where the main capital and investment variables specified refer to tangible capital and investment only, this option should be set to "Tangible", in which case price series for tangible asset investments/stocks are used. In the case that tangible and intangible assets cannot be distinguished, this option should be set to "All", in which case price series for investments and stocks of total fixed assets are used.

2) The default **output produced by the routine has been restricted to Manufacturing, Construction and Non-financial market services** (excluding Agriculture, Mining, Utilities, and Non-market services) as these industries are the primary focus of analysis. This is particularly intended at **limiting the run time and easing the sharing of the output** (by limiting the size of individual output files to share, which also reduces the implementation of confidentiality by participants). This default behaviour can be changed through internal settings defined in the _*Multiprod.ado* file, please contact the team should you want to produce output without any sectors restrictions.
   Note that we kindly ask to **please still use data for all sectors available**, as this is crucial to determine the modal industry of a unit.

3) **Additional internal settings** have been added loop over more detailed industries when preparing the microdata or computing relevant statistics, for cases in which the size of the data creates **memory issues**. Note however that these options generally induce longer run time, and are

recommended only when the default options create issues in the current IT environment. Please contact the team to use this option.

Harmonisation across countries and over time

1) The **main focus** of the core set of output will be on **businesses with at least two persons engaged or two employees** (depending on the variable provided) in headcounts (or 1.5 in FTE). The purpose is to achieve better notably regarding non-employer businesses. In particular, a large number of businesses with one employee may also have one person engaged only, the owner-operator. The extent to which owner-operators in one person engaged units are employees or unpaid workers may vary across countries and legal forms, also depending on national incentives. These businesses are therefore considered separately to focus on a more similar set of units relevant to productivity analysis, independently from whether persons engaged or employees are provided, and whether non-employer businesses are included or not.

   Furthermore, additional investigations conducted during the development of the MultiProd update suggest that the focus on units with at least two persons engaged/two employees has the potential to improve the productivity estimation.

   Please note that information on firms with less than two persons engaged/employees is still collected but recorded separately for descriptive statistics that can inform about the role of non-employer businesses where relevant.[1]

2) **Two additional output files restricted to units with at least 10 persons engaged/employees** (henceforth 10+ units) have been added. In particular, output considering quantiles of productivity (LP and MFP-W) for 10+ units provide a complementary picture of the productivity distribution excluding micro-units.[2] This additional output can improve comparability and harmonisation on various aspects:
   a. For some countries data are available for units above a size threshold, 10 being a common threshold. Collecting more detailed information for this restricted population in other countries will enhance comparability in a cross-country setting, in particular when comparisons to a benchmark are performed.
   b. The coverage of micro-units changes over time in some countries, due to changes in the data collection process (e.g., changes to the sampling design) or due to regulatory changes (e.g., changes in reporting obligations). Collecting additional information on a restricted sample excluding micro firms improves the robustness of the analysis to such changes and enhances comparability of data over time within countries.
   c. This additional information may enable improved linkages with information from other data sources that generally exclude micro-units (e.g., results from ICT survey).

[1] See files *MainStats_W_blwthresh_by_year_ind_a7_sizeclass2901* or similar file at A38 level.

[2] See file Moments_Quant_W_10plus_by_year_ind_a38_sizeclass2222_LogLP_VA_Firm or similar for MFP-W

Measurement

Various refinements of the methodology and measurement have been implemented in this revision of the code.

1) First, the **measurement of firm-level capital** has been refined, which may contribute to better measurement of multifactor productivity.[3]
   a. In particular, when information on firm-level capital is available in the data but missing for some firms, the missing values are now imputed based on firm employment and the capital-to-labour ratio of firms within the same size classes, and a separate imputation is implemented for entering firms.[4] This firm-level information is used for the initialisation of the perpetual inventory method (PIM) when investment series are also available, or for productivity estimation when investment is not available.
   b. When the data include information on investment only, the code implements a new imputation based on country-industry-year specific capital-labour ratio from external sources (notably STAN), adjusted based on the cell average ratio of investment to labour ratio relative to investment/labour ratio in STAN multiplied by firm-level employment, where cells refer to sector-size pairs.

   Additional information can be found in the section Capital Stock Calculation.

2) **An alternative measure of MFP-Wooldridge has been introduced.** Further to the main estimates based on the full sample of firms with two or more persons engaged/employees the code also computes a measure of MFP using the same methodology applied to a restricted sample of firms with at least 10 persons engaged/employees (variable *MFP_W_Alt*). This offers a benchmark against which to compare the main estimates, and can also be used to improve comparability with countries for which data on 10+ firms only is available. Note that the threshold can be adjusted through internal settings. Note further that the main MFP-W measure (based on 2+ firms) is used to computed breakdowns based on MFP-W, but this default behaviour can also be adjusted if required.

3) **The code has been equipped with the possibility to account for the re-assignment of IDs in the microdata**. In some participating countries, some years after the exit of an existing unit, its firm identifiers may be re-used for a new entrant. The code can now perform relevant corrections related to such data which are initiated through the use of the option reassignID (see further information at the description of the option below). Activating this option has significant implications for the identification of business dynamics (firm entry and exit, firm age) and it should

[3] Note that revisions are likely to affect more significantly *levels* of multifactor productivity, with less significant revisions of trends. More significant revisions are expected in some sectors relative to others, and depending on the particular data and settings. For instance, revisions may be more notable when only information on investment is available in the data, or when missing values for the available firm-level capital variable are more frequent.

[4] Previous imputation was relying on information on average investment to depreciation ratio, and on firm-level employment combined with capital-to-labour ratio at the industry level from external sources (notably STAN).

thus be used only if there is concrete reason to believe that it can improve results. If you consider using this option, please mention it to the team.

Expanding and refining the analysis

Some output files have been added or extended and some variables added to refine and extend possible analyses.[5]

1) More detailed information on **productivity by age** groups are collected. Previously, information by age was restricted to young (5 years or younger) and old (6 years or more). While these categories are still used in some files (for instance in transition matrices) a new file at the A38 level collects information for 5 age groups: i) entering firms (age 0), ii) start-ups (1-2 years), iii) young (3-5), iv) mature (6-9), v) old (10+). A new file at the A7 level collects information for 7 age groups: i)-iv) as before, v) 10-14 years old, vi) 15-20 years old, vii) 21+ years old. This will provide additional information on productivity over firms' life cycle. Furthermore, a file on micro-entrants has been added.
2) Additional information on **micro-units (L<10) by age (0-5 vs. 6+ years) and productivity groups** is collected to provide a better understanding of this part of the business population. Note that for this file the micro firms are grouped into productivity bins corresponding to the productivity distribution of the 10+ units. This enables a benchmarking of young and old micro units against larger units. It is also relevant to understand the extent to which the productivity distribution is shaped by the firm-size distribution.
3) Breakdowns according to **weighted quantiles** have been introduced, to collect information on productivity or wage groups reflecting the economic weight of firms. For instance, the group 90-100th of the file by weighted LP (MFP) quantiles records information for units at the top of the productivity distribution representing the 10% largest firms in terms of employment (value-added).
4) Additional files have been added to collect information **disaggregated by quantiles of the labour share distribution, of the wage distribution or the value-added distribution**. This will shed further light on the link between productivity and wages, as well as further insights on the dynamics related to the reallocation of value-added shares.
5) **Additional firm-level information** can be used, with the introduction of new variables related to components of capital and investments (**ICT, software, R&D), profits, international trade**. When available, these variables can be specified in the setting file and statistics will be included in some of the main output files.
6) Additional variables recording **lagged growth rates** have been added to the statistics collected to provide a better characterisation of firm dynamics and will also benefit econometric analysis by adding a set of relevant controls (see the new transformation J, where J3 computes changes from *t-5* to *t-3* and J1 compute changes from *t-3* to *t-1*).

[5] The MultiProd output encompasses 83 analytical dta files and a graphs folder. The total output including dta files and graphs should typically be around 600MB to 1GB large, depending on the length of the time series. The number of files has increased relative to the previous iteration but the restriction to main industries of focus (manufacturing, construction, non-financial market services) has significantly reduced the size of the output.

7) Following the restriction to firms with at least 2 units of labour input (or >= 1.5 FTE), additional variables have been added to track the share of firms in a given year that were below the threshold in the past, and that move above the threshold in the future, which allows to assess the stability of the sample (see variables with root BT_sh below).

Statistical disclosure control ("confidentiality checks")

Following feedback from participants, the statistical disclosure control and implementation of blanking has been further consolidated to fulfil some country-specific requirements.

- **A new option *blankcount* has been introduced**. By default, count variables (i.e., the variables recording the number of firms with non-missing values for a given variable in a cell) are not blanked. However, it is possible to also blank these count variables by setting the option *blankcount* to 1. If possible, it is recommended to keep the default setting (set to 0) as count variables are relevant for data checks and for computing various statistics. Therefore, the blanking of counts should be used only if this is a requirement.
- A new internal option has been introduced to blank cells also when the variable is always missing and the cell is empty (i.e., missing values "." are replaced with missing values ".n"). By default, if a cell is empty for a given variable, the statistic and the count are reported as missing ("."). Please contact the team should you need to use this option.
- The application of **dominance has been made more robust to possible cases in which GO is missing for all firms** in the cell. To apply dominance based on GO or L to another variable, for instance LogW, the code computes the share of GO of the largest firm in the cell, for the sample of firms with non-missing LogW. However, if GO is missing for all firms in the cell, the dominance statistic based on GO cannot be computed. The default behaviour of the code has been changed to be more conservative and set the corresponding variables to missing (i.e., it applies blanking if the dominance statistic is missing)
- Variable specific dominance: Note that the code also allows the dominance to be variable-specific (for instance blanking of L is based on L, blanking of VA on VA, etc.) but for a more limited set of variables (for instance some transformations are excluded, as well as some core variables such as MFP). See option *DominanceVarSpecific* in section Confidentiality*.*

We kindly ask participants to check that the confidentiality applied by the code fulfil national requirements. The MultiProd Team is available to answer any questions you may have on this, to provide further assistance with the implementation of confidentiality and to discuss adjustments to the standard procedures of the code.

Structure of the program

The *MultiProd* program is composed of a series of Stata ado files:

- `MultiProd_Settings.do`: For simplicity, all inputs and options of the program are specified in `MultiProd_Settings.do`. The user simply fills in a set of options then executes the program, which will launch the `_MultiProd.ado`. This generates a wide range of statistics which are stored as .dta files in the folder specified by the user.

- `_MultiProd.ado`: This is the master file. It specifies a series of internal settings and calls a series of other .ado files to compute all the statistics generated by the program. Neither `_MultiProd.ado` nor the associated analysis files require any user intervention.

Data requirements to run the program

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

1. **Longitudinal data on inputs, outputs, and labour costs**: the data source must be a longitudinal response to a production survey or similar type of data (e.g., tax records). In the case that multiple data sources are available, users should choose the one that provides the best trade-off between coverage of units and quality of information on inputs (employment, investment and/or capital stock and intermediate inputs), output (gross output and value added) and labour costs.
2. Sectoral coverage: the aim is to cover the full non-financial market sector, including manufacturing as well as non-financial market services and construction. A list of the included sectors is provided in Annex 1.[6]
3. Panel dimension: individual units need to be identified by a unique longitudinal identifier (*id*) that must be constant over time. If an active unit exits the production survey, its identifier must not be re-assigned to another firm.[7]
4. The minimum set of required variables is:
    a. A unique firm **identifier**.
    b. A **calendar year** to which the time-varying variables refer.[8] Please provide data until the most recently available year and as many years back in time as possible.
    c. A **3- or 4-digit sector** identifying the main economic activity of the unit.
    d. The firm's **labour** input, preferably as a headcount measure of persons engaged. In case both headcounts and full time equivalent (FTE) are available, it would be greatly appreciated to compile two sets of output based on headcounts and FTE (modalities can be discussed with the team).
    e. The **gross output** (GO) of the firm. If gross output is unavailable, **sales** can be used instead.
    f. **Value added** (VA) of the firm.

[6] The data for all sectors should be merged together into a single file. If this is not possible, please contact us.

[7] Where comprehensive *PopData* is available to identify activity (sales, purchases, or labour input), the *MultiProd* code provides an option to follow the Eurostat convention that a firm which is inactive for two years or more will be counted as an exit, and any future activity will be treated as the entry of a new firm.

[8] If the data are collected on the base of a fiscal year that does not coincide with the calendar year, please assign the data to the most appropriate calendar year. This should be reported in the accompanying metadata questionnaire.

g. **Intermediate inputs.** The code will still run in the absence of intermediates data, but many core outputs (including mark-ups and some MFP measures) will not be produced.
h. Either **investment** or **capital stock** (ideally both).
i. The total **labour costs** of the firm. The code will run in the absence of the labour costs data, but statistics on wages and the labour share of value added will not be calculated.

5. Additional data which will be used by the programme if available:
   a. The **birth year of the unit**, used to identify new entrants and to produce statistics by age class.
   b. **Intangible capital** and/or **intangible investment**, used to generate descriptive statistics and to inform later analysis.
   c. **Foreign ownership.** If available, separate descriptive statistics will be calculated for foreign and domestically owned firms.
   d. Member of a **Business Group** If available, separate descriptive statistics will be calculated for independent enterprises and members of business groups.
   e. Information on profits, used to generate descriptive statistics and to inform later analysis.
   f. Information on imports and exports, used to generate descriptive statistics and to inform later analysis.

Detailed variable definitions and data structure

Data Structure

As noted above, *MultiProd* can accommodate a range of different data structures. Two main structures are considered (please contact the MultiProd team for different data structures):

1) Information on the population characteristics (*PopData*), including industry, birth year, and firm employment, are available from a Business Register or administrative data on the *full* business population (i.e., covering the universe of units considered to the extent possible). Information on productivity components such as capital, labour, labour costs, and value added are available from a sample-based production survey in a separate database. Please flag any potential discrepancies and conceptual differences between the employment variable in the *PopData* and in the *ProdData* as well as other variables available in both datasets.
2) Information on the population characteristics (*PopData*) and on productivity components are based on comprehensive administrative data, and available in the same file.

PopData is used to identify the population of active firms, to identify firm birth and death, and for the allocation of firms to a single ISIC Rev. 4 industries in the case of a change in classification.[9] Where

[9] The code can account for changes in classification system over time (from ISIC revision 3 to ISIC revision 4) and units changing industries over time.

both data sources are available, the code will attempt to merge them, checking that the industry and firm size (Labour) measures for the same firm identifier match. If the *PopData* cannot be merged with the *ProdData* (either due to using different firm identifiers or due to security restrictions), *PopData* will be used only to identify the population of firms and to improve the industry concordance. If you have additional data sources (e.g., comprehensive tax data on firm output or activity) which would help to identify the population of active firms, these can also be combined with the *PopData* to improve precision. Please do this prior to running the MultiProd code.

In order to improve the precision of entry and exit variables, and to allow the correct calculation of weights, please keep all available years and firm observations from the *PopData* in the dataset, even if production data is not available for some observations or years.

If it is not possible to access comprehensive unit-record data on the population of firms, it may be possible to weight based on external, published population data. Please contact us to discuss this option.

Unit of measurement for numeric variables

All numeric variables should be in single units (e.g., not in thousands or millions). All monetary variables should be in nominal values and recorded in local currency. An exception is for countries that joined the Euro (or EZ12 countries before 1999): the series should be transformed in Euro for the entire period using the fixed official exchange rate at the time of accession.

Unit of analysis

The *MultiProd* routine should preferably be executed using statistical units, at the enterprise level. The routine can also be run for administrative or legal units. Please report the units used in the metadata questionnaire.

If establishment level data is also available and resources allow, please also consider re-running the routine to provide an additional set of results at the plant/establishment level. This would enable the MultiProd team to investigate whether observed patterns are robust to the unit definition, and to provide comparable analysis for countries where only plant-level information is available. The unit of analysis should be specified in the option unit. The output databases will be named accordingly.

Firm Identifier

Each firm must have a unique, longitudinal firm identifier. Ideally, these should be consistent between the *ProdData* and the *PopData* (please inform the team about any discrepancies between the two). If an active firm leaves the *ProdData* sample, the same firm identifier should not be reallocated to another firm. The code can accommodate reallocation of firm identifiers in the *PopData*, provided this only occurs after a minimum two complete years of inactivity by the original firm, using the option *reassignID* (see details later).

Year

The calendar year to which the time-varying variables refer. If the data are collected on the base of a fiscal year that does not coincide with the calendar year, please assign the data to the most appropriate calendar year. Please provide data until the most recently available year and as many years back in time as possible.

Industry Code

A 3- or 4-digit sector identifying the main economic activity of the unit, following the ISIC Rev. 4 (NACE Rev. 2), ISIC Rev. 3.1 (NACE Rev. 1.1) or ISIC Rev. 3 classifications.[10] 3-digit industry classification is generally the preferred choice, unless data coverage is sufficient to use more granular information at the 4-digit level.[11]

In addition, it is preferred that the firm is assigned to a unique sector over time. If this is not the case, the program will attribute the modal sector to the unit (selecting the most recent modes in case of multiple modes) over the time period observed.

The code can accommodate data classified either partially or completely according to the ISIC Rev. 3.1 (NACE Rev. 1.1) or ISIC Rev. 3 classifications (see notes below). For data following another industry classification (such as NAICS, etc.), users are requested to convert the industry classification to ISIC Rev. 4 using appropriate conversion tables prior to running the program.

Regarding the industry classification, please note that:

4) In `MultiProd_Settings.do`, one of the two variables must be specified: either IndustryCode_old or IndustryCode_new.
5) If only one variable is specified, the code assumes that the same industry classification system is used in all years.
6) If only the old industry classification is provided, this needs to be either in NACE Rev 1.1 or ISIC Rev 3.1. The reason is that the code needs to be converted into the new classification system using an external conversion table (the file named changeover_database.txt, provided with the code), which is internally available only under those two classification systems.
7) If both variables are specified, then also the IndustryCode_ChangeYear must be specified. If in your dataset the industry identifier is stored in one single variable before and after the change in the industry classification system, you can specify the same name for both IndustryCode_old and IndustryCode_new. The code can also handle this case properly.
8) If a separate *PopData* source (BusinessRegisterFile) is given, it should contain the same variable as industry identifier: same name, type and content. If industry classification changes it has to change in the same year. If not, please contact us.

Birth Year

The birth year of the unit. This can be outside (i.e., earlier than) the period covered by the population data or production data. The birth year can be left-censored i.e., birth is known to occur at the censor year or before but the exact date is unknown, so that the reported birth corresponds to the censoring

[10] If available only at a higher level of aggregation, please let us know. The code will also function with 2-digit classifications.

[11] Note that in case 4-digit level is specified, the population breakdown based on *PopData* used for reweighting will be computed at this level of aggregation, but if the production data has weak coverage of some cells, this may actually induce lower representativeness of the data with respect to the alternative of using 3-digit level information.

value. For instance, all birth occurring prior to 1990 are reported as occurring in 1990. In this case, please provide a variable containing the year of left censoring.[12]

If the birth year or left censor year are not available in the production database (*ProdData*), the users can also provide their variable names as stored in the business register (if available). The code will automatically merge them from there, after checking that firm identifiers and the year variable definitions are the same (see above). If birth year is recorded in *both* the *PopData* and *ProdData* datasets, the code prioritises the birth year found in the *PopData*. If this is not appropriate for your data, please drop the *PopData* birth year variable to ensure that *ProdData* is used instead, or contact the MultiProd team to discuss.

Birth year information may be missing for some units. In this case, the first year in which the unit appears with non-zero employment in *PopData* is considered as the entry year, unless this year coincides with the first year of the *PopData* (in which case birth year is left as missing).

Labour Input

The firm's labour input (employees or persons engaged) in headcount or full-time equivalents. For the sake of comparability across countries, the main employment measure should be based on headcounts if available. If FTE is also available, FTE-based measures will also be used as a robustness check and if resources allow, it would be useful to consider a separate complementary run with FTE (this run could be restricted to priority output to avoid producing all files).

Where both are consistently available, the preferred measure for the labour input variable (**L**) is persons engaged (i.e., including working proprietors and unpaid family members engaged in the business), as this is the most appropriate measure for calculating LP and MFP. See additional information in the Glossary.

Employment can refer either to a yearly average (preferred) or to a precise point in time. 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 (i.e., no rounding to multiples of 10 or 100).

Labour Costs and Count of Paid Employees

Please provide the total labour costs of the firm. The preferred measure is the total gross wage bill including social security contributions. The code will run in the absence of the labour costs data, but statistics on wages and the labour share of value added will not be calculated.

In some cases, the Labour Costs measure may cover only a subset of the people working in a firm (due to, e.g., the presence of working proprietors, unpaid family workers, etc.). If this is the case, please also provide information on the number of workers over which the total labour costs are calculated (e.g., excluding working proprietors and unpaid family members) to enable the calculation of an average wage. This variable should be identified as Employees_Wagebill in

[12] Note that the code allows for the year of left censoring to vary across units (e.g., it may differ by sector), but requires that it does not change over time for the same unit. If this is the case, the user needs to create a new left-censoring variable that is equal to the minimum value for that unit. The user should then input only the name of the latter variable into the program. In addition, please ensure that only a single birth year is supplied for each unit.

`MultiProd_Settings.do`. If Employees_Wagebill is left empty, the program assumes that the wage bill captures the earnings of all workers (i.e., average wage is calculated as LabourCosts/Labour).

For comparability, please use a headcount measure if possible. If FTEs are also available, please also provide this (as Employees_Wagebill_FTE) as it will be used as a robustness check.

Gross Output

The gross output (GO) of the firm. If gross output is unavailable, sales can be used instead. See the Glossary for additional details.

Value Added

Value added (VA) of the firm, preferably at basic prices.[13] Please use your most preferred measure. The program does not attempt to calculate VA from its components (e.g., gross output minus intermediate inputs, and possible adjustments), but expects it to be available in the dataset. See the Glossary for additional details.

Intermediate Inputs

If available, please provide information on total intermediate inputs. The code will still run in the absence of intermediates data, but many core outputs (including statistics on firm mark-ups) will not be produced.

(Tangible) Capital and Investment

Either investment or capital stock is required for the calculation of MFP. The preferred measure is annual investment since the program obtains the capital stock through a common approach based on the perpetual investment method (PIM) to increase the comparability across countries (see details in the section on methodological issues below). The measure of investment provided should be total investment across all tangible fixed asset classes (buildings, structures, machinery, etc.). If both investment and capital stock are available, please specify both (the value of the capital stock will be used for the initialisation of the PIM).

If an alternative capital stock measure (i.e., based on book values) is available in the micro data, the program also uses it for carrying out a robustness analysis (or for the main analysis if investment is not available). If the capital stock measure is provided, it should be the net capital stock which reflects current monetary values of capital goods aggregated across all tangible asset classes (book value of capital), where valuation reflects market prices for new and used assets. Importantly, this measure takes into account depreciation.

For comparability across countries, the preferred measure of capital *excludes intangible* capital. Intangible capital is collected separately (see below).

Intangible Capital and Investment

Intangible capital and/or intangible investment. If a separate measure of intangible capital is available, this will be used to generate descriptive statistics and to inform later analysis. Intangible investment will be reported directly in descriptive statistics (PIM is not applied).

[13] If VA at factor costs or producer prices is also available, please inform us.

Investment in ICT, software, R&D

If measures of investments in ICT equipment (computer hardware and telecommunications equipment), or investments in software and databases, or Research and development are separately available, these can be provided separately to generate descriptive statistics and to inform later analysis. Investment intensities will be reported directly in descriptive statistics for these categories of assets (PIM is not applied). Note that in case only a measure of total ICT investment/capital is available (i.e., for tangible and intangible ICT together), this can be specified for the option Investment_ICT and Capital_ICT.

Foreign Ownership

If foreign ownership can be identified, please provide a binary indicator of ownership (0=domestic, 1=foreign).

Business Group

If available, please provide a binary indicator of whether a firm is independent (0=independent), or is part of a group of linked enterprises (1=group member) (e.g., parent-subsidiary group).

Imports and exports

If available, please provide a numeric variable reporting the value of imports and exports (in single units and domestic currency, as recommended above for monetary variables).

System Requirements

MultiProd should not require much more memory than the amount needed to load the input dataset. The computation time required for the full code to run can take multiple days. If lengthy runs are not possible, we recommend running the code initially without the section on mark-ups, as this is particularly computationally intensive. To exclude the mark-ups section, please set the global ComputeMarkups in the settings file to 0. However, we do ask that you re-run the analysis including this section once you have tested that the code runs correctly.

Confidentiality

The program can automatically apply three forms of confidentiality. The user can specify that any or all of these should be applied.

1. Number of units. First, the program can apply a threshold rule based on the number of units per cell. If the MinElementsInCell option is specified, the code blanks cells for statistics based on strictly fewer than X units, where X is specified by the user in the option MinElementsInCell. For example, automatically blanking cells with fewer than five firms. Note that count variables (the number of firms with non-missing values for a specific variable in a given cell) are not blanked by default. To blank these variables as well, users should set the option blankcount to 1.
2. Dominance. Secondly, the program can implement a *Dominance* rule. If option ApplyDominance is set to 1, the program blanks cells in which the largest firm or the largest two firms account for more than Y% of the cell's total, in terms of employment (L) or output (GO). The code program can also implement blanking based on the P-percent rule (see options DominanceShare, DominanceShare_top2, and DominanceP_Rule in `MultiProd_settings.do`.). In order to allow

for the possibility of negative value-added, the p-percent rule is applied after converting to absolute values.

Note that the code also allows the dominance to be variable-specific (for instance blanking of L is based on L, blanking of VA on VA, etc.) but for a more limited set of variables (for instance some transformations are excluded, as well as some core variables such as MFP).

3. Percentiles disturbance or percentiles computed as averages. Finally, if the option DisturbPercentiles is set to 1, the routine applies a random disturbance around median and percentiles values (+/- up to 0.1% of the value). This is done at the end, before saving the statistics in the output file. Random disturbance is used instead of taking the average of several values around the *n*-th percentile, because the weighting complicates that approach substantially. The value of 0.1% is set internally but can be modified.

   The code also allows users to indicate that percentiles should be computed as averages of observations around the percentile value. This requires specifying AveragePercentiles to 1 and AveragePercentilesMinObs to the number of observations that should be used (for instance AveragePercentiles and AveragePercentilesMinObs will compute the percentiles based on the average of 3 observations).

Where confidentiality rules have been applied in the code, the program will also output an additional set of results, saved in the Output directory within a folder named “confidential”. In these additional results, confidentiality is *not* applied, and the output data further include the variables required to confirm that confidentiality rules have been applied correctly (e.g., true percentile values, dominance shares). This output should not be released.

Participants are kindly asked to check the data produced by the program and to ensure that cells are blanked according to their national confidentiality rules. Please report to the team any doubts you may have on the implementation of the confidentiality rules by the code.

The MultiProd team is available to help with this process e.g., if additional diagnostic variables are required in order to apply country-specific confidentiality rules, we can modify the code to create these additional diagnostic variables.

In particular, **please note** that:

(1) **Secondary suppression is not applied.** Dominance checks and counts of units are run ***only*** **for the output cells**. If additional checks are required for your data, please apply the appropriate rules, or contact the team and we can help with this process.

(2) **All variables are treated equally** by the *MultiProd* code. If your country’s confidentiality rules are variable specific (e.g., different rules for levels vs growth rates, or for medians vs averages) please let us know, as we can help with additional ex-post code adapted to country-specific rules.

Program inputs and settings

This section describes the main inputs and settings of the *MultiProd* Stata routine are defined in the `MultiProd_Settings.do` file. Note that additional information on definitions is provided in detailed variable definitions and data structure. The main inputs and settings are:

General Settings

Country

String (Country), required. The 3 letter ISO code in CAPITAL letters for the country of study (e.g., USA)

Unit

String (Unit), required. Specifies the unit at which the analysis is carried out (i.e.: firm or plant). If it is left blank, it is assumed that the unit is the firm.

BR_Only

Integer. Option that allows for a separate run of the code ONLY on the *PopData* (Business Register or comprehensive admin data) to prepare the population breakdown. This option should be set to 1 if *PopData* is in a different secure data repository, or uses a different set of firm identifiers, and cannot be merged with the production survey. For a more detailed explanation, please see the notes below (p.25).

Directories

DataDir

String, required. Directory containing the data file for the productivity analysis (*ProdData*) and, if applicable, the business register data file (*PopData*).

BRDir

String. Directory containing the population data files (*PopData*). If it is left empty, the population data file is assumed to be in the directory DataDir.

CodeDir

String, required. Directory containing the `MultiProd_Settings.do` file and all the ado files that compose the program.

AuxDataDir

String. Directory containing the STAN dataset and `changeover_database.txt` that were sent together with the other inputs of the program by the OECD. By default, it is set to CodeDir/Auxiliary.

OutputDataDir

String, required. Directory where the output results and the log file are stored. It can be the same as DataDir. This folder also stores the detailed population breakdown used for weighting (`PopulationBreakdownExt.dta`). If it is not possible to merge the *PopData* and *ProdData*, please create the population file and put it here before re-running the code (see detailed notes on the BR_Only option below).

TempDataDir

String, required. Directory used for temporary files created and then deleted by the codes. It can be the same as the DataDir.

Files

ProdFile

String, required. Stata data file containing variables for productivity measurement and analysis (*ProdData*). This is usually a production survey, but may also be administrative data (e.g., from tax returns).

BusinessRegisterFile

String. Required if BR_Only is set to 1. Stata data file containing information on firm demographics (*PopData*). This is usually a business register but may also be administrative data. It can be left empty if it is already merged with the productivity data file or if it is unavailable.

Variables Names and Related Settings

Year, Population and Demographic Variables

FirmID

String (FirmID), required. Firm (unit) longitudinal identifier. If a separate BusinessRegisterFile is given, it should be the same variable there (name, type etc.). Ideally, the content of the variable should be numeric (but the code can also handle the case in which FirmID contains non-numeric characters).

reassignID

Integer. Takes value 0 (default) or 1. This option indicates how to treat IDs in case there are "gaps" in the data (for instance there is a one year "gap" if a unit is in the data in *t-2*, not in *t-1* and is in the data again in *t*).

Year

String (Year), required. Calendar year to which the time-varying variables refer. If BusinessRegisterFile is given, it should be the same variable there (name, type etc.). The content of the variable has to be 4-digit numeric.

BirthYear

String (BirthYear). Year of birth variable. Can be left empty if not available.

LeftCensorYear

String (LeftCensorYear). Variable indicating the year of left censoring of the birth variable – i.e., the year before which the exact birth year cannot be obtained (i.e., the birth is known to occur at the censor year or before but the exact date is unknown). It must be constant for a given firm but may vary across sectors or firms.

CensusYearList

Numlist. This option only applies to countries which have the census (population of firms) only in certain years. Please provide a list of census years separated by a space. Example: 2002 2007 2012. Note that this is not a variable contained in the dataset but a list of integers.

Industry Classification

IndustryCode_old

String (IndustryCode_old). Old industrial classification variable name (if ISIC Rev. 3 or 3.1 used in some or all years of data). The variable has to be numeric, and may not include a decimal point or comma. If the industry classification system is the same in all years, either IndustryCode_old or IndustryCode_new can be used and the other variable can be left empty. If BusinessRegisterFile is given, it should be the same variable there (name, type etc.). See note below.

IndustryCode_new

String (IndustryCode_new), required. New industrial classification variable name (ISIC Rev. 4). The variable has to be numeric, and may not include a decimal point or comma. If the industry classification system is the same in all years, either IndustryCode_old or IndustryCode_new can be used and the other variable can be left empty. If BusinessRegisterFile is given, it should be the same variable there (name, type etc.). See note below.

IndustryCode_ChangeYear Integer. First year when the new industrial classification is used. Note that this is not a variable contained in the dataset but a single (4-digit) integer. Can be left empty in case of no classification break. If BusinessRegisterFile is given, the same year should be specified and in exactly the same way (name, type etc.) in both files.

NumberOfDigits Integer (NumberOfDigits), required. Number of digits at which the industry classification of firms is provided. 2 or 3 or 4 digits are all accepted. 3-digit industry classification is generally the preferred choice. Note that this is not a variable contained in the dataset but a single integer.

Labour Input and Labour Costs

Labour String (Labour). It can be left empty only if Labour_FTE is specified. Labour input variable: Employment or Persons Engaged (headcounts). Persons Engaged is the preferred measure. If BusinessRegisterFile is given, the variable should be specified in the same way (name, type etc.) in both files. Ideally, the two should also take the same value. However, due to differences in timing or slight differences in definition (e.g., point in time vs. average over the year) they might differ. If there are significant differences in the specification across sources (e.g., if *PopData* records persons engaged while the *ProdData* records employees), please contact the MultiProd team, as this will affect the weighting calculations.

Labour_FTE String (Labour_FTE). Labour input variable: Employment or Persons Engaged in FTE (full time equivalents). The field can be left empty if FTE measures are unavailable (in this case the setting is Labour required). If BusinessRegisterFile is given, the variable should be exactly the same and specified in the same way (name, type etc.) in both files and ideally the two should also take the same value. However, due to differences in timing or slight differences in definition (e.g., point in time vs. average over the year) they might differ. If there are significant differences in the specification across sources (e.g., if *PopData* records persons engaged while the *ProdData* records employees), please contact the MultiProd team, as this will affect the weighting calculations.

LabourCosts String (LabourCosts). Labour costs: total gross wage bill including social security contributions if available, for the same set of workers who are included in the Employees_Wagebill variable.

Labour_Wagebill String (Labour_Wagebill), required. Number of workers (headcount, if available) underlying the LabourCosts variable. This will differ from the Employees variable if some workers do not receive wages or salaries (e.g., working proprietors, unpaid family members) or are otherwise excluded from the Labour Costs measure.

Labour_Wagebill_FTE String (Labour_Wagebill_FTE). Full-time equivalent (FTE) workers underlying the LabourCosts variable. If both FTE and headcount measures are available, please include the FTE variable here.

Output, Intermediates, Profits

Output String (Output), required. Gross Output variable (can be sales or turnover if gross output is unavailable).

ValueAdded String (ValueAdded), required. Value added variable.

IntermediateInputs String (IntermediateInputs). Intermediate inputs variable. Can be left empty if not available.

Profits String (Profits). A measure of profits, if available (can be left empty if not available). See the Glossary for additional information.

Capital and investment

Investment String (Investment). Investment variable used for the calculation of capital stock. Please include only tangible investments. It can be left empty if it is not available, but in that case, Capital must be specified. Ideally both a measure of investment and capital should be provided.

Capital String (Capital). Tangible capital stock. Must be specified if Investment is left empty, otherwise it is optional (although preferable, as it will be used as a starting value in the PIM and for robustness check). Please include only tangible capital if possible.

Investment_Intangible String (Investment_Intangible). Name of the variable recording intangible investment. Can be left empty if not available.

Capital_Intangible String (Capital_Intangible). Name of the variable recording intangible capital stock. Can be left empty if not available.

Investment_ICT String (Investment_ICT). Name of the variable recording total investment in ICT. Can be left empty if not available.

Capital_ICT String (Capital_ICT). Name of the variable recording total ICT capital. Can be left empty if not available.

Investment_Software String (Investment_Software). Name of the variable recording investment in software and databases. Can be left empty if not available.

Capital_Software String (Capital_Software). Name of the variable recording total capital related to software and databases. Can be left empty if not available.

RnD String (RnD). Name of the variable recording investment in R&D expenditures. Can be left empty if not available.

Ownership

Group

String (Group). Variable indicating whether the firm belongs to a business group. This must be a binary variable (0=independent or 1=part of a business group). The field can be left empty if this information is not available.

Foreign

String (Foreign). Variable indicating whether the owner is foreign (possibly the ultimate owner if available). This must be a binary variable (0=domestic or 1=foreign). The definition of owner requires that it has controlling stake/interest in the firm (e.g., 50% of shares, depending on legislation). The field can be left empty if this information is not available.

Trade

Exports

String (Exports). Variable indicating the total exports of the firm, preferably referring to exports of goods and services and at factor cost. Can be left empty if not available.

Imports

String (Imports). Variable indicating the total imports of the firm, preferably referring to exports of goods and services and at basic price cost. Can be left empty if not available.

Confidentiality

MinimalRun

Integer. Binary value 0 or 1 (0 is the default). If set to 1, the code will **produce a minimum set of output**. This option is useful to test the specification and produce a first set of output that can be used to assess data quality. This is recommended for instance for a first run of the code notably on large microdata or for runs in environments with more limited computational capacity. In case no additional data modifications are necessary after a run with MinimalRun = 1 to generate the full MultiProd output, the code can be run again with MinimalRun = 0 and ResumeRun = 1 (below) [and otherwise identical settings] to complete the run without duplication of result production and runtime.

MinElementsInCell

Integer. Minimum number of required observations in a cell to report a statistic. Any output cells using (strictly) fewer observations will be blanked (set to missing).

Blankcount

Integer. Binary value 0 or 1 (0 is the default). Set to 1 if count variables (i.e., the count of the number of firms in a cell for a specific variable) should be blanked according to the threshold MinElementsInCell. This should be set to 1 only if necessary.

DisturbPercentiles

Integer. Binary value 0 or 1 (0 is the default). Set to 1 if percentile values (e.g., the median) cannot be reported for confidentiality reasons. In that case, percentile values will be randomly disturbed by a small amount. The default disturbance parameter is 0.1%. Please contact the MultiProd team if a higher parameter is required.

AveragePercentiles

Integer. Binary value 0 or 1 (0 is the default). Set to 1 in case percentile values (e.g., the median) cannot be reported for confidentiality reasons and you want them to be computed as average of a certain number of observations. If set to 1 please see also the option AveragePercentilesMinObs .

AveragePercentilesMinObs

Integer. Minimum number of required observations to calculate a percentile as an average if the option AveragePercentiles is set to 1. Please set it to an odd number if possible (the default value corresponds to MinElementsInCell).

ApplyDominance

Integer. Binary value 0 or 1 (0 is the default). Set to 1 if you want to apply a dominance rule in the confidentiality checks. Largest firm or top two firms in terms of employment or sales cannot have a share larger than defined in DominanceShare or DominanceP_Rule (see below).

DominanceVar

String. Determines whether dominance is measured in terms of labour (L) or sales/gross output (GO).

DominanceType

How to calculate the dominance share [only for the case of DominanceVarSpecific=0]. Four options exist: default, unrestricted, restricted, combined.

- default: for a variable VAR, the dominance share corresponds to the share of the top firms (in terms of L or GO, as specified in DominanceVar) with a non-missing value of VAR in the total of DominanceVar
- unrestricted: for a variable VAR, the dominance share corresponds to the share of the top firms (regardless of VAR missing or not) in the total of DominanceVar
- restricted: for a variable VAR, the dominance share corresponds to the share of the top firms with a non-missing value of VAR in the sum of DominanceVar among firms with non-missing VAR
- combined: for a variable VAR, the dominance share corresponds to the maximum share of unrestricted and restricted option above (in case P-rule is specified, blanking is triggered as soon as it would be required according to one of the unrestricted and restricted P-rules)

DominanceVarSpecific

Integer. Binary value 0 or 1 (0 is the default). Set to 1 if to compute dominance on variable-specific basis (e.g., L is blanked on dominance statistics computed for L, W based W etc.). Note however that this is implemented for a more limited set of variables. Please contact the MultiProd team for additional information. This option is not compatible with DominanceVar.

DominanceShare

Numeric. Number between 0 and 1, the maximum share that the largest firm can represent without the cell's content being blanked (e.g., 0.2).

DominanceShare_top2
Numeric. Number between 0 and 1, the maximum share that the largest two firms can represent without the cell's content being blanked. It can be left empty is there is no such a rule (i.e., dominance only for largest firm).

DominanceP_Rule
Minimum P-percent value for which a cell can be disclosed, where P = (T – (X1 + X2))/X1. T is the total value, and X1 and X2 are the top two individual firm values in a cell (e.g., 0.2).

Modularity and additional settings

Compute_Markups
Binary value 0 or 1. If set to 0, markups will not be computed. This is used to save computation time (especially in test runs), but it is preferable to execute the code with Compute_Markups 1.

ResumeRun
Binary value 0 or 1 (0 is the default). If it is set to 1, files already generated are not re-created. Please use this option only to resume the generation of output when the underlying data have not changed. Note that importantly, the use of this option requires to keep the file Temp_Main_VarsDefined.dta from the previous run and to execute the code with the same settings. Note that this option will also skip most of the data preparation for sectors for which relevant variables have been already fully computed.

runStatistics_toExclude
String. List of output to exclude from the run (see MultiProd_Settings file for additional information). This is used to execute the output in a more modular way, by ignoring the compilation of some output files.

runStatistics_toInclude
String. List of output to include from the run (see MultiProd_Settings file for additional information). This is used to execute the output in a more modular way, by restricting the output to specific output files.

indlevels_MainStats
String which can be empty, take the value "A7", "A38" or "A7 A38". Industry breakdowns for which the "Mainstat" files (see section Program output) are computed. The default should be "A7 A38".

indlevels_Moments_TM
String which can be empty, takes the value "A7", "A38" or "A7 A38". Industry breakdowns for which the "moments" files and the transition matrices (see section Program output) are computed. The default should be "A7 A38".

KdensNum
Minimum number of observations required in order to produce kernel density graphs. Set to 200 by default but can be changed by the user.

KdensPerc
Percentage for trimming of kernel density graphs. The kernel density graphs will drop the top and bottom *x%* of firms before computing the kernel densities. Set to 1 by default but can be changed by the user.

Notes

If the *ProdData* and the *PopData* use different firm identifiers or are in two separate locations and they cannot be merged together, the program allows for a separate run of the code only on the *PopData* to prepare the population breakdown.

To do this, please set the option BR_Only to 1 and then specify: all directories (except DataDir); the BusinessRegisterFile (*PopData*); and the following variables: FirmID, Year, IndustryCode_old/new, NumberOfDigits, IndustryCode_ChangeYear, Employees. If the OutputDataDir changes between this separate preparatory run and the full run on the production survey, then please copy the two output `dta` files containing `PopulationBreakdownExt` in their name (with suffixes `_firm` and `_firm_restricted` if the unit of analysis is the firm, or `_plant and _plant_restricted` if the unit of analysis is the plant) into the OutputDataDir that you will specify in the full run (i.e. when the option BR_Only is set to 0).

Program Outputs

File names and dimensionality of the output

The `_MultiProd.ado` program specifies various levels of aggregation at which data are collected and, for each of those levels, the types of aggregated data that are collected.

Statistics are not collected for individual firms. Instead, the program splits firms along various dimensions, into cells, and for each cell collects aggregate yearly data. Each output file produced by *MultiProd* ends with the list of variables defining what constitutes a cell in the output file.

For instance, the file `MainStats_UW_by_year_ind_a38.dta` contains data at the year/ind_a38 level, where ind_a38 is the OECD sectoral classification used in STAN. Within an output file, firms can be aggregated based on firm-specific dimensions, such as sector, age class, or size class; or relative dimensions, where a given firm’s classification will depend on other firms too: for instance, quantiles of productivity. In order to preserve the confidentiality of the data, not all dimensions are used together, especially at high levels of disaggregation.

Several types of aggregated statistics are collected and defined in the `_statistics.ado` file. The name given to the statistics type is also reflected in the output file name (defined in parenthesis):

1. Basic moments (`Moments`): mean, median, standard deviation, and number of non-missing values, for a series of variables.
2. Distribution measures (percentiles) for a series of variables and related transformations: productivity levels, productivity growth rates, and wages (`MainStats`).
3. Descriptive statistics of firm characteristics (including growth rate and wage dispersion) by quantiles of the productivity distribution in levels and growth, and by quantiles of the sales and mark-ups distributions (e.g., `Moments_Quant`).
4. Characteristics (productivity, age, persistence) of firms at the productivity frontier (`Frontier`).
5. Employment dynamics by quantiles of the productivity distribution (`Moments_Quant_NJC`)
6. Transition Matrices collecting the number and characteristics of firms which transition between quantiles of the productivity distribution over various time horizons (`transmats`).

7. Detailed population breakdown used in weighting (`PopulationBreakdown`).
8. Basic statistics for firms with particular characteristics, such as negative value added (SumStats)
9. Kernel density graphs of sectoral productivity distributions (in a /graphs/ subfolder).

The *MultiProd* program output is saved as `.dta` Stata dataset, either at the path specified in the `OutputDataDir` option, or in the `DataDir` folder otherwise. The filename of these outputs follows the general syntax, where the middle part of the name indicates whether the output is produced using sampling (inverse probability) weights (`W`) or not (`UW`):

`[Type_of_Statistics]_[W/UW]_[Dimensions_used_to_split_firms].dta.`

Cell definition

The table below details the firm-specific variables that can be used, sometimes in combination, to define a cell.

Table 1. Cell definition variables

Variable	Description	Maximum number of categories
ind_a7	Macro-sector classification based on a coarser OECD STAN A*10 classification	7
ind_a38	Sectoral classification based on the OECD STAN A*38 classification	38
group_nat	Whether firm belongs to a domestic group or not	2
foreign	Whether a firms is under foreign or domestic ownership	2
ageclass21	Age classes: Young (0-5 years), Old (6+ years)	2
ageclass51	Age classes: Entry, Start-ups (1-2 years old), Young (3-5 years), Mature (6-10 years), Old (10+ years)	5
sizeclass22	Size classes: Small (<10 employees), MediumLarge (10 employees or more)	2
sizeclass2222	Same as sizeclass22, but no data is collected on Micro firms (9 employees or less), i.e. only data for 10+ is collected.	1
sizeclass52	Size classes: Micro (1-9 employees), Small (10-19), MediumSmall (20-49), Medium (50-249) and Large (250+)	5
sizeclass5222	Same as sizeclass52, but no data is collected on Micro firms (9 employees or less)	4
sizeclass71	Size classes: vMicro (<5 employees), Micro (5-9), Small (10-19), MediumSmall (20-49), Medium (50-99), MediumLarge (100-249), Large (250+)	7
demographics_L	Demographic of the firm: incumbent, entrant, future exiter, blip firm (appears one year only) – only used for measures of job creation	4

Quantile classes

In addition to groups that depend on firm-specific characteristics, some statistics are also computed based on firms' position relative to the other firms in their sectors, based on a given variable. Firms are then grouped into quantile classes corresponding to their relative positions.

Several variables are used to create these quantile classes: three measures of productivity (LogLP_VA, LogMFP_SW and LogMFP_W) —both in levels and growth— as well as gross output GO, value added

(VA) and wages (W), labour share (L_Sh) and Markups.[14] Variables that record to which quantile class a given cell is referring to are of the format `p_[var]`, where quantiles are computed using the variable `var`.

These quantile classes are cut off at the 10th, 40th, 60th and 90th percentiles. For instance, a cell where `p_GO` takes the value "90_100" denotes data relating to firms in the top decile of the gross output distribution i.e., the 10% biggest firms.[15]

The resulting files are called `Moments_Quant_W_by` followed by the variables used to define the cell. For example, the file `Moments_Quant_W_by_year_ind_a7_LogMFP_W.dta` contains statistics by year, macro-sectors (ind_a7) and quantiles of LogMFP_W (MFP à la Wooldridge).

General Syntax of Variable Names

Most variables are of the form: `[QUANTITY]_[statistic]`, where `QUANTITY` is the economic quantity being measured and `statistic` is the type of statistic being collected. For example: quantity `L` stands for the number of employees (persons engaged) in a given firm, so `L_av` measures the average number of employees (persons engaged) across all firms in the cell.

Table 2. List of statistics computed for a given quantity

Type of statistics	Description
_av	Mean
_sd	Standard deviation
_pXY	XYth percentile of the variable. In particular, p50 is the median.
_N	Number of observations (see below).
_sum_w/_sumw	Sum of weights (see below)
_pareto_b; _pareto_se	Estimated Pareto coefficient (and associated standard error) for the distribution of the variable.

Importantly, for any quantity, the variable `[quantity]_N` records the number of firms on which the other statistics were computed. For instance, in a given cell, if `L_N` = 10, this means that variables `L_av` (average number of employees by firm) and `L_sd` (std. dev. of the number of employees by firm) were computed on the basis of 10 firms. The number of observations used for calculations is unique to each measure, as it is possible that different data are present or missing in the original dataset; taking the previous example, it is possible that, in a given cell, data on the number of employees by firm (L) is available for 10 firms, so that `L_N` = 10, but that data on gross output (GO) is only available for 4 firms, so that `GO_N` = 4.

[14] See Annex 1 for a detailed description of the MFP measures used for *MultiProd.*

[15] In addition to the quantile classes, some outputs (e.g., transition matrices) also allow `p_[var]` to reflect other characteristics of the firm, for example, in the case of VA as "neg" for "negative value added".

In order to be able to aggregate up from the average to the total, the code also collects the sum of the weights assigned to each observation, as **_sum_w**. This variable identifies the size of the underlying population for the cell, and (unlike **_N**) should be approximately constant across related variables.[16]

A note on confidentiality

In the confidentiality parameters, if the option MinElementsInCell is set to a number X, any cell containing statistics based on less than X observations will be blanked (set to missing value *.n*). Given that the number of observations used is variable dependent, it is possible that *in the same cell* some variables are blanked while others are not.

Continuing our previous example, let's assume that option MinElementsInCell is set to 5. In a given cell where **L_N** = 10 but **GO_N** = 4, it means that in that cell, variables such as **GO_av**, **GO_sd** will be blanked (set to *.n*) while variables will **L_av** and **L_sd** will be reported in the results.

Basic Firm-Level Quantities

The following table details the basic data and their description. This pertains to firm-level data that is being aggregated into different statistics. Further information on particular variables can be found in the discussion below.

Table 3. List of Basic Quantities

Quantity	Description
age	Firm age
GO	Gross Output
I	Investment in tangible capital
I_ICT	ICT investment (if available)
I_Intan	Intangible investment (if available)
I_Soft	Software and database investment (if available)
II	Intermediates
K	Capital: tangible capital, either book values or using PIM (preferred)
K_ICT	ICT capital stock (if available)
K_intan	Intangible capital stock (if available)
K_L	Capital to labour ratio
K_Rob	Capital robustness check: book values of tangible capital stock, only used when K is PIM and both book values and investment are available
K_Soft	Software and database capital stock (if available)
L	Labour input: number of employees or, if available, persons engaged (headcount)
L_FTE	Labour input: number of employees or, if available, persons engaged (full time equivalent)
L_Sh	Firm-level labour share of income: WL/VA[17]
LP_VA	Labour Productivity: Value added per unit of labour, headcounts
LP_VA_FTE	Labour Productivity: Value added per unit of labour, FTE
M	Imports

[16] **_sum_w** will differ in certain cases: between VA-based variables and others, due to the inclusion of negative value-added firms; for differences or growth rates across two time periods, where the population refers to those active in both time periods; or where variable-specific data is missing for all firms in one or more of the narrowly defined weighting cells within an aggregate cell.

[17] The labour share of income (**L_Sh**) is calculated only for firms with positive value added.

Markup	Firm level mark-ups, computed using De Loecker and Warzynski (2012) methodology.[18]
MFP_ACF	Multi-Factor Productivity, computed using Ackerberg, Caves and Fraser (2015) methodology.
MFP_SW	Multi-Factor Productivity, computed as a Solow residual, using external, industry specific labour shares from OECD STAN (the cross-country-year median, one value for each industry)
MFP_SW_Rob	Same as above, but using the alternative measure of capital K_Rob (if available)
MFP_W	Multi-Factor Productivity, computed using Wooldridge (2009) instrumented production function estimation performed on the main analysis sample of firms with 2+ units of labour input
`MFP_W_Alt`	Multi-Factor Productivity, computed using Wooldridge (2009) instrumented production function estimation performed on the restricted sample of firms with 10+ units of labour input
Pi	Profits (if available)
RnD	R&D investment (if available)
VA	Value Added
W	Average labour costs: WL/number of paid employees
WL	Total labour costs: total employee earnings + employer social security contributions (if available)
X	Exports
BT_sh	Share of firms below the sampling threshold (L<2 or L<1.5 if L refers to FTEs). The variable only exists in forward- and backward-looking growth rates as its current level is zero in any file (only firms currently above the threshold are included). FjBT_sh is the share of currently active firms above the threshold that are below the threshold after j years. (-1)*DjBT_sh is the share of currently active firms above the threshold that were below the threshold j years ago. For more information on transformations, see the next section.

The notation for the basic firm-level quantities is combined with the notation for the statistics to define particular variables available in the MultiProd output. For instance, `LP_VA_av` denotes the average Labour Productivity (value-added per worker). Once again, the `_N` suffix designates the number of observations on which these are based, so that `LP_VA_N` would record the number of firms used to compute `LP_VA_av` in that cell.

Variables Derived from Basic Quantities

Based on the basic quantities listed in Table 3, a number of additional variables are computed. These transformed variables follow the syntax `[TRANSFORMATION][QUANTITY]_[statistic]`.

[18] The code estimates the mark-up of price over marginal cost at the firm level, adopting the methodology proposed by De Loecker and Warzynski (2012). This method identifies mark-ups as the ratio of an input's output elasticity ($\frac{\partial GO}{\partial II} \cdot \frac{II}{GO}$) and its revenue share ($\frac{II}{GO}$), where the output elasticity is estimated using the Ackerberg et al. (2015) production function.

Logged variables

The most common variable transformation is Log, which gives the logged value of a quantity. For instance, `LogGO_av` represents the average of logged gross output in that cell.

Time transformations

Other transformations typically involve a time-dimension e.g., using values of a given quantity in past or future periods.

Below is a list of the main transformations. In each case, *t* refers to the current year and *j* refers to the number of years over which the transformation is applied.

Table 4. Main transformations

Transformation	Description
L*j*	Lagged value at time *t-j* (e.g., L1=value last period)
A*j*	Lead value at time *t+j* (e.g., A1=value next period)
D*j*	Backward-looking difference, normalised by the number of periods: $(x_t – x_{t-j})/j$
F*j*	Forward-looking difference: $(x_{t+j} – x_t)/j$
J*j*	Lagged backward-looking difference between two periods. E.g. J1=(xt-1 – xt-3)/j and J3=(xt-3 – xt-5)/j
M*j*	*j*-period moving average (forward-looking) (e.g., M2=mean of x_t and x_{t+1})
P2	Squared value
N*j*	Only applicable to the variable L. N*j*L is the net job creation between *t* and *t+j*, by currently incumbent firms.
NP*j*	Only applicable to variable L. NP$_j$ is the net job creation between *t* and *t+j*, by currently incumbent firms, restricted to those firms where net job creation over the period is positive.

These other transformations are often compound with Log, since log-differences represent (approximate) percentage growth rates. For instance, `D1LogL` represents the first difference of logged number of employees in a firm and hence represents a growth rate between last and current period, so `D1LogL_av` represents the average growth rate in the number of employees of firms in that cell between *t-1* and *t*.

Note on variable transformations

Note that transformed variables are computed at the firm-level *before* being aggregated into a statistic. For instance, `LogLP_VA_av` is the average of the logged labour productivity, not the log of the average labour productivity.

Examples

`LogLP_VA_p90` represents the 90th percentile (top decile) of logged labour productivity. `P2GO_av` is the average level of squared gross output. `p_LogMFP_W` represents the quantile class of logged MFP, where for instance "0_10" would denote the bottom decile in the distribution of the logged MFP, and "90_100" would represent the top decile.

Exceptions to the general syntax

Some variables do not follow the general syntax of `[quantity]_[statistic]`. We detail the exceptions below.

Correlations

Some correlations are computed. Resulting variables follow the syntax `corr_[VariableX]_[VariableY]`. For instance, `corr_W_LP_VA` represents the correlation between real wage W and labour productivity LP_VA.

Concentration ratios

Statistics are also computed for the output share (GO and VA) of the top 8 and top 20 firms in each industry. These are computed at the 4-digit industry level and the statistics reported refer to the average, median, and standard deviation across 4-digit industries within the aggregated industry category. For example, `GO_CR8_av` for the Manufacturing sector refers to the average of the share of the top 8 firms in total gross output for each 4-digit industry within Manufacturing, weighted by the number of firms in each 4-digit industry.

Weighted variables

The *MultiProd* code includes two distinct types of weighting. Further details on the weighting are given in the section on Methodology below.

First, where production data is only available for a sample of the population, inverse probability weighting is applied to all variables to ensure that results are representative of the population of firms. The sum of the inverse probability weights is collected and saved as `[Variable]_sum_w`**.** This value should approximately match the total number of firms in the relevant population and is used for aggregation across cells.

Secondly, for certain variables (e.g., wages or productivity), firms are also weighted according to a measure of their size, in order to provide an estimate of the aggregate. In this case, the variable name follows the syntax: `[QUANTITY]_w_[WEIGHT]_[statistic]`. Different weighting variables are used for different quantities as appropriate. For instance, `LogLP_VA_w_L_av` is the average of logged labour productivity weighted by the number of employees. For MFP, a number of aggregation weights are used, depending on the MFP measure in question. The aggregation weight most used is denoted `_KLII`, and represents an input-based aggregation weight computed following Van Biesebroeck (2008).[19] Variables ending `_wsum` (e.g. `MFP_W_w_KLII_wsum)` record the sum of the aggregation weights.

Melitz-Polanec (2015) dynamic decomposition of aggregate productivity changes

For each core measure of productivity (LogLP_VA, LogMFP_W and LogMFP_SW), the code implements a decomposition of the changes in aggregate productivity à la Olley-Pakes (2015) that accounts for the dynamic contribution of entry and exit of firms. The decomposition shows the contribution of resource allocation to aggregate or industry productivity growth. The decomposition results in four elements: a within-firm change (suffixed by `_OPDyn_W`), a covariance term (reallocation of market shares, suffixed by `_OPDyn_Cov`), the contribution of entrants (suffixed by `_OPDyn_E`) and that of exiters (suffixed by `_OPDyn_X`). The decomposition is done over three different time frames of 1, 3 and 5 years, with the suffix of the variable indicating the time frame. For example, `LogLP_VA_w_L_OPDyn_E_5` denotes the

[19] KLII weights are calculated using elasticities from a constant returns to scale production function, and capture only firms with positive value added.

contribution of entrants to the change in aggregate (logged) labour productivity, calculated over 5 years.

In turn, the OP Gap (**_opgap**) reflects the gap between the weighted average of productivity and the unweighted average. A large OP gap implies that firms which are more productive capture a larger share of inputs (or a larger share of the market), and is hence a proxy for allocative efficiency.

Petrin-Levinsohn (2012) decomposition of productivity change

The code also implements a decomposition of firm-level change in productivity (MFP_W) following the Petrin and Levinsohn (2012) methodology. This method decomposes aggregate productivity growth (APG) into improvements in technical efficiency and improvements in resource reallocation. The technical efficiency term captures the hypothetical increase in aggregate productivity caused by an increase in output, if firm-level inputs were held constant. The reallocation terms for different inputs describe the increase in APG caused by a reallocation of inputs from firms with a low value of marginal product to a high value of marginal product for these inputs, while assuming aggregate inputs and technical efficiency remain constant.

The resulting variables are **D1LogMFP_w** followed by **_APG** to denote aggregate productivity growth (**_APG_aggr** is an alternative measure computed with a different aggregation strategy, and **_APG_inc** is aggregate productivity growth computed on incumbents only), **_TE** for the within-firm change in technical efficiency, and **_RE_L** and **_RE_K** for the reallocation terms for labour and capital respectively.

Foster, Haltiwanger and Krizan (2001) decomposition

The code also includes a decomposition of aggregate productivity changes into the shares associated with continuing firms, entrants, and exiters, following the methodology of Foster, Haltiwanger and Krizan (2001). In this measure, changes in industry level (weighted average) productivity are decomposed into five components: the within-firm component **_FHK_W_**,which captures within-firm changes in productivity, weighting by the initial year market share; the between-firm component (**_FHK_B_**) which captures the reallocation of market shares between firms, based on their initial relative productivity levels; the covariance component (**_FHK_Cov_**), which captures the extent to which changes in firm productivity are associated with changes in market share; **_FHK_E_** which captures the contribution to aggregate productivity growth of entering firms; and **_FHK_X_** which captures the contribution of exiting firms.

Diewert and Fox (2010) decomposition

Finally, the code also includes a decomposition of aggregate productivity change following the methodology of Diewert and Fox (2010). While closely related to the Melitz and Polanec (2015) decomposition, this methodology places lower weight on small firms in calculating the within-firm component of the decomposition, leading to lower volatility in the within and between components for continuing firms. It also applies a symmetric treatment across time periods, such that levels differences in productivity, and the various decomposition terms, reverse in sign when the time periods are interchanged but are otherwise identical. Analogous to the other decompositions, these are denoted by **_DF_W_, _DF_E_** etc.

Transition matrices

The code also produces transition matrices that analyse the dynamics of productivity, mark-ups, labour share, and firm size within industries. In the style of Kehrig and Vincent (2018), these transition matrices give the probability that a firm will be in a certain quantile of the distribution of the variable of interest in year *t+j* based on its quantile in time *t* (or its employment-based size class in the case of firm size). Analogous to the OECD *DynEmp3* database, the transition matrices contain information on transitions between time *t* and time *t+j*, where *t* takes the values of year 1995, 1998, 2001, 2004, 2007, 2008, 2009, 2010, 2012, 2015, 2018 and *j* is equal to 3, 5, 7, 10 or 14. Transition matrices are calculated for all possible combinations of *t* and *j*, wherever the relevant data is available. The matrices also contain averages of initial values, final values, and growth rates over the selected period, of variables like firm employment, output, and capital. These summary statistics are also calculated in selected intermediate years for each transition group (e.g., for the group which transitions from the bottom quantile in 2000 to the top quantile in 2005, we collect summary stats at 2000, 2003, and 2005). Statistics are also collected in the intermediate years for those firms which exit between *t+k* and *t+j* where *t+k* refers to the intermediate year immediately prior to *t+j*. For example, if *t*=2000 and *j*=10, the category for exit refers to firms which exit between 2007 and 2010. The resulting outputs are included in the files labelled `transmats_year_ind_a38_by_` followed by the variable used to define the cell over which transitions are calculated.

Net and gross job creation

Net and gross job creation statistics by productivity quantile use labour from the *ProdData* where available, to minimise false entry and exit (due, e.g., to sampling or missing data in production surveys) and for consistency with the DynEmp project. Job creation outputs are reported in files labelled `Moments_Quant_W_NJC`.

Kernel densities

In order to provide a visual comparison across industries, countries and time, the code also produces and plots kernel densities of key productivity measures, by year and industry. Both the graphs and the underlying kernel densities are saved in the `/graphs/` subfolder). By default, the top and bottom 1% of each distribution are trimmed and kernel densities are output only where there are at least 200 observations in the cell. These settings can be changed in `MultiProd_Settings.do`.

Methodology

Capital Stock Calculation

For the MFP calculations, a measure of capital stock is needed. In the baseline case, the program defines the capital stock variable through the perpetual inventory method (in order to increase the comparability of results across countries). Then, in case an alternative measure of capital stock is directly available in the micro data, the program uses it to initialise the capital stock for the PIM, and further carries out a robustness analysis using the capital series provided, as discussed below.

Perpetual Inventory Method

In the baseline case, the code obtains capital using the Perpetual Inventory Method (PIM) methodology. As common in the literature (e.g., Olley and Pakes, 1996; Levinsohn and Petrin, 2003), the initial value is set to the capital stock reported by the firm in the initial year, whenever this is available.[20] If information on capital stock is available in the data but missing for some firms, the missing values are imputed based on firm employment and the capital-labour ratio in firms within the same size classes. Separate imputation is implemented for entering firms.

If on the other hand the capital stock is not available in the data, the code calculates the initial capital stock based on country-industry-year specific capital-labour ratio (K/L) from STAN, adjusted based on the cell average ratio of investment/labour relative to investment/labour ratio in STAN multiplied by firm-level employment (L), where cells refer to sector-size pairs. Alternatively, if this value is missing for a firm-year pair, the code will use the average between i) average firm investment in all years divided by the depreciation rate (from STAN), borrowing a similar idea from Mueller (2008); and ii) country-industry-year specific capital-labour ratio (K/L) from STAN multiplied by firm-level employment (L).

The code has an option to set the first *n* years of the calculated capital stock to missing, in order to drop the initial noise due to the PIM method.

Note: the code simply treats consecutive observations as consecutive years; this implies that whenever the firm is missing in some years, the PIM method essentially assumes that capital is constant in the years in between. The same is true if only investment is missing, because investment is set to zero whenever is missing (unless it is missing for all years for that firm, in which case investment is set to missing).

Alternative measure

For those countries whose micro data already include a measure of capital stock (from book values or other measures of capital stock obtained by country researchers), the code creates a second measure of capital (**K_Rob**). The code will fill in missing values by interpolating the provided information, this will make **K_Rob** coherent with the measure of capital (**K**) built using the PIM. In the micro data used for testing, the two capital stock measures have a correlation of 0.96. For the countries that have the capital stock information but not the investment data, **K_Rob** becomes the main (and only) measure of capital.

Deflation, PPP and auxiliary variables

MultiProd uses external sources (ISIC Rev.4 version, including preliminary data) for obtaining deflators, depreciation rates, capital-labour ratios and labour productivity levels and growth rates at the country, industry and year level. The main priority source is the OECD STAN database, but other sources are also used (e.g., OECD and Eurostat national accounts data, EU-KLEMS). As a preparatory step, the relevant databases have been checked for possible inconsistencies and, in order to have a balanced panel for each series, missing values at the 2-digit industry level have been filled up with

[20] Conversely to Olley and Pakes (1996), rented capital is not considered to compute the total capital stock because this information is not available in all countries. Consequently, the capital stock might be underestimated, particularly for small firms.

values from industries at a more aggregate level (except labour productivity). Data are then combined, striking a balance between source priority and level of aggregation.

Every nominal monetary variable is transformed into real 2005 local currency. To ensure international comparability, exchange rates and PPPs are therefore applied in that year. Since most of the manufacturing sector is usually traded internationally, the series for manufacturing are simply adjusted using the nominal exchange rate (as average over 2005), taken from Eurostat. All other sectors are adjusted using country-level PPPs for 2005 from the Eurostat-OECD PPP Programme.[21]

Weighting

MultiProd applies two distinct types of weighting. First, for countries where production data is only available for a sample of the population, and to correct for missing data for specific observations, the code applies weights based on the inverse probability that a firm is observed in a given firm size-industry stratum. This weighting is applied to all variables to make the results more representative for the entire population of firms and more comparable across countries, and relies on access to a Business Register or other comprehensive information about the population of firms (*PopData)*. Secondly, for selected variables (e.g., wages or productivity), firms are also weighted according to a measure of their size, in order to provide an estimate of the aggregate.

The weighting strategy entails the following two main steps:

1) Preparation of the population structure from the population data

The code prepares the population structure (number of firms by year, industry, size class) from *PopData* or from census years. The industry used to obtain the population structure is the most detailed available, and the size class is determined by the internal option `sizeclass` (see below). For consistency with the productivity calculations, the population structure uses modal industry over time. Industry composition at a point in time may therefore differ from official statistics.

The program also adopts a special treatment for variables like lags and growth rates, where multiple years of data are required to calculate a statistic. In these cases, the population is defined as the number of firms active in both the relevant years, not just the current one. This is done to reflect the actual "availability of growth rates" in the population of firms (for instance, in case of a "true" exit, a firm would have missing growth rates even if the population of firms was available).

There are two key options for this first step (they are set internally but can be modified), which are:

- `preferred_digit`. This is the disaggregation level at which the industry classification transformation and the weighting is performed. In the baseline case, it is set at 3 digits, to balance granularity and data availability, but the code is also robust to 4 or 2 digits.
- `sizeclass`. This is the number of size classes that are used to define the population breakdowns. The baseline case is `sizeclass 71`, which corresponds to seven size classes with

[21] For more information, see Eurostat-OECD Methodological Manual on Purchasing Power Parities.

thresholds at 5, 10, 20, 50, 100 and 250. The size classes for the weighting are obtained using employment from *PopData* (if available).[22]

2) Calculation of inverse probability weights

The weights are based on the population information contained in the business register (whenever available). Application of the weights takes place after the outlier filtering (see below) and just before the final output statistics are calculated, so that variable specific weights can be computed.[23] The weights are calculated as the total number of firms in the population divided by the number of firms in the survey, after the cleaning. All calculations are performed by industry and size class, which are the same as used in the first step described above (hence industry at the 3-digit level and `sizeclass` 71 in the baseline case). The main issues that are worth highlighting in the calculation of the weights are the following:

- The code always checks that the number of firms in the population is at least as large as in the *ProdData*, otherwise it sets the weight to 1. This is a non-zero probability event, in particular in the case of external population breakdowns (e.g., separate run on the *PopData* only) together with a break in the industry classification, or where the population structure is missing in some years. The latter case is possible when the population is available only in some Census years (for countries like Japan or the US). The program then obtains interpolated values in the missing years and a firm might change size group between Census and non-Census years. The former case, despite being unlikely in general, is still possible because the program adopts a random assignment of firms to new industries in case of a change in the classification, hence firms can be assigned to different industries in the separate run on the *ProdData* and on the *PopData*.
- When VA-based MFP is calculated, the inverse probability weighting is implemented on value added. Thus, ceteris paribus, an increase in the number of negative value added firms in the sample will lead to a fall in the estimated average MFP. The rationale for this choice is that the probability of observing negative VA cases should be a priori the same in the population and in the sample.
- Then missing variables for specific measures of productivity are taken into account (e.g. missing capital for MFP). There is no need to take weights on the levels for any other log variable because negative or zero values are dropped in data cleaning. For the specific case of MFP_W the number

[22] Note that, on the other hand, employment from *ProdData* is used for all other calculations. A consequence of this choice is that when the size classes are defined for the output statistics (which are based on *ProdData*), the weights may not be entirely correct in case of inconsistencies between the *PopData* and *ProdData* employment measures, or in case the population breakdown is obtained through a separate run. Despite this shortcoming, this choice is better than defining the size classes for the output statistics on *PopData* (with *ProdData* employment for calculations), because in that case there would be an inconsistency between the size classes and actual firms contained in the class.

[23] The weighted means and percentiles do not depend on the type of Stata weights that are used. In principle, *pweights* should be used but these types of weights do not allow the user to calculate standard deviations. Thus, we use the *aweights* option, which does what one would expect from *pweights* and exactly the same thing as the *svy* commands. For further information, refer to: http://www.stata.com/support/faqs/statistics/weights-and-summary-statistics/.

of firms in the sample is obtained by counting the non-missing observations for the variable itself and adding the observations with negative or zero VA.

- The weights used for calculating lagged values, annual growth rates or 3/5 year annualised growth rates are based on specific population structures obtained to map the availability of firms over a 2, 3, 5 year horizon in the population. For countries that have the census every X years (e.g., JPN, USA etc.), the annual growth rates or 3-year annualised growth rates are going to take a zero value (apart from years that span two consecutive censuses, e.g. 5 year annualised growth rates for the USA). In this particular case (when all values are zero or missing), the code sets all weights to missing and the weighted results will not be reported.

Aggregation weights

In addition to the inverse probability weights, certain outputs also rely on aggregation weights. These are applied in a second stage, by multiplying the firm-level values by their aggregation weights (e.g., multiplying firm-level LP_VA by employment to calculate aggregate LP_VA).

Summary of data cleaning

The program carries out a series of basic consistency checks and cleaning of the data:

- **Duplicates**. The program checks for duplicates in terms of firm identifier and year. In case duplicates are found, they are not treated but the program simply stops and warns the user.
- **Population and productivity data match** After merging the *PopData* with the *ProdData*, the code checks that all firms in the *ProdData* have been matched, and that *PopData* and *ProdData* measures of firm size are the same or closely correlated. If not, the program stops and warns the user. Note that observations that are in the *ProdData* but not in the *PopData* are dropped to ensure that cleaning based on the productivity data is also applied to the *PopData.*
- **Units of measurement**. A pop-up question at the very beginning of the code asks whether all monetary values are in local currency, nominal values, and single unit (not thousands).
- **Variable availability**. If they are not available in the production survey, the program checks whether they are in the business register; if so it takes them from there. The code also checks the availability of the variables used for MFP calculation and verifies whether the robustness measures of MFP (using deflated capital stock) can be calculated. If neither capital nor investment is detected, the program stops because MFP cannot be calculated. If you do not have capital or investment measures but would like to participate using labour productivity measures only, please contact us.
- **Industry variables**. The program checks for the consistency of the industry variables. If the industry variable is available only in the old classification system, it asks whether it is NACE Rev. 1.1 or ISIC Rev 3.1, because our external transformation file works only for those two systems. If the code detects that the industry classification is at 2 digits, it explicitly asks the user whether a more detailed industry classification is available.
- **Birth**. The code initially looks for the year of birth in the *PopData*, then in the *ProdData*. Similarly to *DynEmp*, if both are missing or birth > first year of non-missing L, then birth is set to the first year of non-missing L (unless it is the first year in the database). The code also allows for left censoring, and the birth is set to missing if it coincides with the left censor year. For all cases where birth year pre-dates the censoring year, the program assumes that the reported birth year is the correct one and does not apply any correction. For reported birth

years that occur within the sample period, the code checks that no activity is observed prior to the supplied birth year, and adjusts the birth year accordingly if prior activity is observed. Finally, if comprehensive activity data is available (e.g., via Business Register or tax records) the program sets the birth year to the first year of appearance with positive employment or observed sales activity, as long as this is posterior to the first year of the sample used if the birth year is missing or prior to the first recorded activity. Otherwise, it leaves it as missing or to the recorded birth year.

- **Firm exit.** If comprehensive activity data is available, the code provides the option to apply the Eurostat convention whereby a firm that is inactive for two complete years or more is treated as having exited. If the same firm identifier is observed to be active in a later year, they are classified as a new firm.
- **Implausible jumps**. This applies to the core variables inputted to the code only and it is aimed at excluding observations with implausible jumps (consecutive large one-off increases and large one-off decreases or vice versa) in the main variables of interest (employment, value added, total revenues, wage bill). The code excludes observations in case of one-off jumps by factor of 50 (30 in case of labour costs). In the case of employment, the jump factor parameter is 10 and the adjustment is performed only for firms with more than 20 employees. Moreover, the code excludes observations in case of large falls after entry; this is done symmetrically for all variables whenever they drop by a factor of 10 in the two years after entry. In these cases, all information from the observation is dropped instead of simply setting the relevant variable to missing because according to experience on the data it is difficult to assess whether it is a problem related to the single variable or to the entire record. A very similar strategy is adopted to clean the capital stock, once computed by the program. The jump parameter takes again a value of 50.
- **Missing or zeros**. The code drops observation with missing, zero or negative employment. Moreover, it sets FTE employment, wage bill, total revenue, and the capital stock (if available) to missing whenever they take zero or negative values (this is not performed for value added because the program keeps also negative values of VA to characterise the overall distribution in levels of VA/L). The code does not drop for missing or negative values of these other variables because the observation might still contain useful information for other useful variables (e.g., if employment in full time equivalents is missing, we might still have it in headcounts).
- **Cleaning of labour share**. Values of the labour share are censored to 9 to avoid important distortions of unweighted averages of the labour share in case of extreme values (for instance due to very low value-added). This threshold is chosen to be conservative and implements a mild cleaning.

Examples of contributions using MultiProd output

Berlingieri, G., P. Blanchenay and C. Criscuolo (2017), "The great divergence(s)", *OECD Science, Technology and Industry Policy Papers*, No. 39, OECD Publishing, Paris, https://doi.org/10.1787/953f3853-en.

Berlingieri, G., S. Calligaris and C. Criscuolo (2018a), "The productivity-wage premium: Does size still matter in a service economy?", *OECD Science, Technology and Industry Working Papers*, No. 2018/13, OECD Publishing, Paris, https://doi.org/10.1787/04e36c29-en.

Berlingieri, G., Calligaris, S., and C. Criscuolo (2018b), "The productivity-wage premium: Does size still matter in a service economy?", *AEA Papers and Proceedings,* Vol. 108, pp. 328–333, https://www.aeaweb.org/articles?id=10.1257/pandp.20181068.

Berlingieri, G., S. Calligaris, C. Criscuolo and R. Verlhac (2020), "Laggard firms, technology diffusion and its structural and policy determinants", *OECD Science, Technology and Industry Policy Papers*, No. 86, OECD Publishing, Paris, https://doi.org/10.1787/281bd7a9-en.

Calligaris, S., F. Calvino, R. Verlhac and M. Reinhard (2023), "Is there a trade-off between productivity and employment?: A cross-country micro-to-macro study", *OECD Science, Technology and Industry Policy Papers*, No. 157, OECD Publishing, Paris, https://doi.org/10.1787/99bede51-en.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015), “Structural Identification of Production Functions”, *Econometrica*, Vol. 83, Issue 6, pp. 2411-2451, https://doi.org/10.3982/ECTA13408.

Ackerberg, D. A., K. Caves, and G. Frazer (2006), “Structural Identification of Production Functions,” mimeo.

Arnold, J., G. Nicoletti, and S. Scarpetta (2008), “Regulation, Allocative Efficiency and Productivity in OECD Countries: Industry and Firm-Level Evidence”, *OECD Economics Department Working Papers*, No. 616, OECD Publishing, Paris, https://doi.org/10.1787/241447806226.

Bartelsman, E. and M. Doms (2000), “Understanding Productivity: Lessons from Longitudinal Microdata”, *Journal of Economic Literature*, Vol. 38, No. 3, pp. 569–594, American Economic Association, https://www.aeaweb.org/articles?id=10.1257/jel.38.3.569.

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*, Vol. 14, Issue 3, pp. 365–391, https://doi.org/10.1093/icc/dth057.

Bartelsman, E., J. C. Haltiwanger, and S. Scarpetta (2009), “Cross-Country Differences in Productivity: The Role of Allocation and Selection”, *NBER Working Papers*, No. 15490, National Bureau of Economic Research, Cambridge MA, https://www.nber.org/papers/w15490.pdf.

Blundell, R. and S. Bond (2000), “GMM Estimation with persistent panel data: an application to production functions”, *Econometric Reviews*, Vol. 19, Issue 3, pp. 321–340, https://doi.org/10.1080/07474930008800475.

Bond, S. and M. Söderbom (2005), “Adjustment costs and the identification of Cobb Douglas production functions”, *IFS Working Papers*, WP 05/04, Institute for Fiscal Studies, https://www.ifs.org.uk/wps/wp0504.pdf.

Caves, D. W., L. R. Christensen and W. E. Diewert (1982), “The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity”, *Econometrica*, Vol. 50, No. 6, pp. 1393-1414, https://www.jstor.org/stable/1913388.

Criscuolo, C., P. N. Gal and C. Menon (2014a), “The Dynamics of Employment Growth: New Evidence from 18 Countries”, *OECD Science, Technology and Industry Policy Papers*, No. 14, OECD Publishing, Paris, https://doi.org/10.1787/5jz417hj6hg6-en.

Criscuolo, C., P. Gal and C. Menon (2014b), “DynEmp: A Stata® Routine for Distributed Micro-data Analysis of Business Dynamics”, *OECD Science, Technology and Industry Working Papers*, No. 2014/02, OECD Publishing, Paris, https://doi.org/10.1787/5jz40rscddd4-en.

De Loecker, J. and F. Warzynski (2012), “Markups and Firm-Level Export Status”, *American Economic Review*, Vol. 102, No. 6, pp. 2437–2471, American Economic Association, https://www.aeaweb.org/articles?id=10.1257/aer.102.6.2437.

Diewert, W. E. and K. J. Fox (2010), “On measuring the contribution of entering and exiting firms to aggregate productivity growth”, chapter 3, pp. 41–46 in Diewert, W. E., B. M. Balk, D. Fixler, K. J. Fox and A. O. Nakamura (2010), *Price and Productivity Measurement: Volume 6 – Index Number theory*, Trafford Press, www.indexmeasures.ca/V6_Ch3_10,09,26_%20Diewert_Fox_%20EnterExit.doc.

Foster, L., J. C. Haltiwanger, and C. J. Krizan (2001), “Aggregate Productivity Growth: Lessons from Microeconomic Evidence”, in *New Developments in Productivity Analysis*, National Bureau of Economic Research, pp. 303–372, https://www.nber.org/chapters/c10129.

Gal, P. (2013), “Measuring Total Factor Productivity at the Firm Level using OECD-ORBIS”, *OECD Economics Department Working Papers*, No. 1049, OECD Publishing, Paris, https://doi.org/10.1787/5k46dsb25ls6-en.

Griffith, R., S. Redding and H. Simpson (2009), “Technological Catch-Up And Geographic Proximity”, *Journal of Regional Science*, Vol. 49, No. 4, pp. 689–720, https://personalpages.manchester.ac.uk/staff/Rachel.Griffith/pdf/Publications/GriffithReddingSimpson2009.pdf.

Hulten, C. R. (2001), “Total Factor Productivity. A Short Biography”, in *New Developments in Productivity Analysis*, National Bureau of Economic Research, pp. 1–54, https://www.nber.org/chapters/c10122.pdf.

Kehrig, M. and N. Vincent (2018), “The Micro-Level Anatomy of the Labor Share Decline”, *NBER Working Papers,* No. 25275, National Bureau of Economic Research, Cambridge MA, https://www.nber.org/papers/w25275.

Levinsohn, J. and A. Petrin (2003), “Estimating Production Functions Using Inputs to Control for Unobservables”, *Review of Economic Studies*, Vol. 70, No. 2, pp. 317–341, https://www.jstor.org/stable/3648636.

Melitz, M. J. and S. Polanec (2015), “Dynamic Olley-Pakes decomposition with entry and exit”, *The RAND Journal of Economics*, Vol. 46, Issue 2, pp. 362–375, https://doi.org/10.1111/1756-2171.12088.

Mueller, S. (2008), “Capital stock approximation using firm level panel data”, *BGPE Discussion Papers*, No. 38, Bavarian Graduate Program in Economics (BGPE), Nürnberg, http://hdl.handle.net/10419/73391.

Olley, G. S. and A. Pakes (1996), “The Dynamics of Productivity in the Telecommunications Equipment Industry”, *Econometrica*, Vol. 64, No. 6, pp. 1263–1297, https://www.jstor.org/stable/2171831.

Petrin, A. and J. Levinsohn (2012), “Measuring aggregate productivity growth using plant-level data”, *The RAND Journal of Economics,* Vol. 43, No. 4, pp.705–725, https://www.jstor.org/stable/41723351.

Rovigatti, G., and Mollisi, V. (2018), “Theory and Practice of Total-Factor Productivity Estimation: The Control Function Approach using Stata”, *The Stata Journal*, Vol. 18, Issue 3, pp. 618-662. https://doi.org/10.1177/1536867X1801800307.

Syverson, C. (2011), “What Determines Productivity?”, *Journal of Economic Literature*, Vol. 49, No. 2, pp. 326–65, American Economic Association, https://www.aeaweb.org/articles?id=10.1257/jel.49.2.326.

Van Biesebroeck, J. (2008), “Aggregating and decomposing productivity”, *Review of Business and Economics*, Vol. 53, No. 2, pp.122–146, https://feb.kuleuven.be/public/N07057/CV/vb08rbe.pdf.

Van Biesebroeck, J. (2007), “Robustness of Productivity Estimates”, *Journal of Industrial Economics*, Vol. 55, No. 3, pp. 529–569, https://doi.org/10.1111/j.1467-6451.2007.00322.x.

Wooldridge, J. M. (2009), “On estimating firm-level production functions using proxy variables to control for unobservables”, *Economics Letters*, Vol. 104, No. 3, pp. 112–114, https://ideas.repec.org/a/eee/ecolet/v104y2009i3p112-114.html.

Annex 1: Productivity Definitions

Productivity, in its broadest interpretation, is meant to capture the efficiency by which inputs are turned into outputs (Hulten, 2001). The *MultiProd* program computes a series of productivity measures that go from the least to the most data-demanding methodologies, summarised in Table 5. MultiProd 2.0 collects outputs for only the following four measures: LP_VA, MFP_SW, MFP_W, MFP_ACF. A detailed discussion of the construction and relative merits of the different measures is provided below. This section draws heavily on Gal (2013).

The gross output or total revenue (GO) based labour productivity is the most crude measure but potentially offers the highest coverage among all productivity variables. Note that even though gross output differs from total revenues in case of changes in inventories, either measure will be admitted in *MultiProd* since their correlation tends to be very high at annual frequency (the more preferred one is gross output). Of course, one of the main problems with a gross output based measure is that it does not control for intermediate input usage.[24] As a result, a company with significant reselling activity will probably rank very high in this measure.

Value added (VA) based labour productivity alleviates this problem, as value added itself is the difference between output (or sales) and intermediate inputs (including resold goods, typical in retail trade). Of course, neither type of labour productivity control for differences in capital intensity across firms or differences in other inputs. Multi-factor productivity (MFP) measures control for a wider set of inputs, using various methodologies. In *MultiProd*, two broad types of MFP measures are considered: index number approaches and methods based on production function estimation.

MFP measures take either gross output or value added as the output variable, and capture both labour (measured by the number of employees), capital (measured by the total depreciated capital stock), and intermediates (when gross output is used, measured as total cost of intermediates) as inputs. Introducing more sophisticated measures either on the labour side (e.g., using the number of hours, workers with different skills) or on the capital side (e.g., using different types of assets, relying on capital services) would put additional constraints in terms of country, sector and time coverage and comparability, as their availability is more limited. The choice of simpler but cruder input measures is driven by the aim of collecting comparable statistics across countries while achieving the broadest possible coverage. Some refinements to the input measures have been implemented in MultiProd 2.0, but these remain as options for those countries where the data is available.

Index number measures relate output to a weighted sum of inputs generally assuming constant returns to scale.[25] They are relatively easy to calculate. Two widely used index numbers are the Solow-

[24] Note that if one considers within-sector patterns, this is less of a problem, since firms in the same sector tend to have similar intensities of intermediate input usage. Revenue based labour productivity has been used, for instance, in Foster et al. (2001), where the authors mention that the results in terms of their decompositions and relative productivity are very similar across productivity measures.

[25] Note that usually all MFP measures assume that firms are price takers on both output and input markets and maximise profits. Also note that assumptions about returns to scale are not too critical from an empirical point

residuals and the Superlative-index measures.[26] The first measure weights inputs using external, industry specific shares. The latter uses as weights the average between firm-level factor shares and a reference value in the industry (an average across firms or a specific firm e.g., the median), and implicitly compares productivity levels to that reference value. As such, it is more prone to measurement error than the Solow-residual measure (see Van Biesebroeck, 2007). *MultiProd* provides four index-number productivity measures:

a) MFP_S, a Solow-residual based MFP using external, country-industry specific labour and intermediate shares;
b) MFP_SW, a Solow-residual based MFP using external, industry specific labour and intermediate shares (the cross-country-year median);
c) MFP_SI: a Superlative index-based MFP using labour and intermediate shares calculated as the average between the labour/intermediate share of the firm (averaged over time) and the geometric mean of firm labour/intermediate shares in the industry;
d) MFP_GR: a Superlative index-based MFP using firm-level labour and intermediate shares calculated as a Törnqvist discrete time approximation of Divisia shares (average over two consecutive years, MFP_GR is obtained for MFP growth only).

As suggested by Syverson (2011), the labour and intermediate shares are taken as constant over time in all cases but d).[27,28] This makes the empirical implementation more consistent with the underlying model, which assumes away any factor adjustment costs. Moreover, it makes cross-country comparison easier, particularly in the case of MFP_SW.

Turning to the production function-based MFPs, the simplest benchmark is the residual from an OLS regression of the production function, usually run separately for each industry. However, the estimates obtained through this exercise are likely to be inconsistent and biased. In the case of the labour coefficient, for instance, firms with higher productivity hire more workers. Productivity is not

of view; according to the experiments on productivity measurement performed by Van Biesebroeck (2007), all approaches described score relatively well irrespective of such assumptions.

[26] For the original idea, see Caves et al. (1982) and for applications, see Griffith et al. (2009) and Arnold et al. (2008).

[27] The external aggregate labour and intermediate shares are taken from the OECD STAN database. Whenever the labour and intermediates' shares are missing (or their sum larger than 1), the value is taken from sectors at a more aggregated level, as for deflators (see below). Note that the aggregate labour share might be underestimated because the income of self-employed workers is missing. The share of labour input costs published in the STD Productivity database is adjusted for self-employed income but unfortunately, the database does not have the country-industry-time coverage needed. Adjustments like the one suggested in the OECD Productivity Manual have not been carried out because this might end up overestimating the labour share, and the micro-data are more likely to underestimate it as well (e.g.: temporary workers from employment agencies are not always included).

[28] The firm level labour and intermediate shares are winsorized at the top and bottom 1% to remove outliers. In case the resulting capital share is negative, it is set to missing and productivity is not computed (incidence is very low: only 0.1% of observations in the micro data used for testing).

directly observed, and it enters the error term. Firms' behaviour will therefore introduce a positive correlation between the error term and the labour input, yielding standard OLS estimates that are inconsistent and biased.[29] Fixed effect estimators can only partially solve the problem since they take into account only a time-invariant firm-specific productivity effect and cannot control for cases where productivity shocks vary across firms and time.

A well-known solution for these problems was proposed by the semi-parametric approach of Olley and Pakes (1996), who took investment as a proxy variable for unobserved MFP.[30] However, investment is proven to be a 'lumpy' variable and hence potentially a poor proxy for productivity.[31] The method of Levinsohn and Petrin (2003) takes a similar approach, but instead of investment, it uses intermediate inputs as proxies. Later, Wooldridge (2009) developed a technique that builds upon the Levinsohn and Petrin (2003) method. It is a quicker one-step procedure with consistent standard errors without bootstrapping and it also overcomes the critique of Ackerberg et al. (2006).[32] This is the methodology adopted in *MultiProd* and is implemented using a value-added based MFP only, as in Ackerberg et al. (2006). The reason for this approach is that in a GO setting there are more severe identification problems (see Bond and Söderbom, 2005).

Finally, outputs are also computed using the Ackerberg, Caves and Frazer (2015) methodology. Ackerberg et al. (2015) argue that the Levinsohn and Petrin method suffers from a problem of collinearity, and propose a two-stage alternative which accounts for endogeneity while allowing the identification of all the input coefficients of the production function. This method relies on an assumption that the timing of input choices can be further disaggregated: capital inputs are largely fixed, labour is chosen in response to the firm's previous observation of productivity but prior to observing the current year's shock, and intermediate inputs are flexible and chosen simultaneously with or after labour input.

Overall, the theoretical advantage of the estimation-based methods (no assumption of constant returns to scale needed) comes at a cost: their data requirement is usually higher, both in terms of the number and the quality of the available variables. Finally, it is interesting to note that, irrespective of the sophistication of the measures, it has been shown that firms' productivity levels and dynamics vary a lot, even in narrowly defined industries (see Bartelsman and Doms, 2000 and Syverson, 2011).

[29] Endogeneity arises because as productivity changes, optimising firms react by adjusting their inputs (right hand side variables), and productivity also directly affects value added or gross output (the left-hand side variable).

[30] Another strand of production function estimators comes from the dynamic panel data literature (e.g., Blundell and Bond, 2000), but they have lost popularity compared to the semi-parametric approaches discussed here.

[31] Their method also controls for selection bias. Selection can cause a problem since firms with higher capital stock will more easily weather negative productivity shocks and are more likely to stay in the sample than others.

[32] The Ackerberg et al. (2006) critique refers to a co-linearity problem: the joint inclusion of the nonparametric, polynomial terms of the variable input (e.g., labour) together with its structural coefficient in the production function makes the latter potentially unidentified.

Table 5. Productivity measures computed in MultiProd

Productivity measures		Definitions[i]
Labour productivity (*LP*)	Based on gross output: *LP_GO*	$\frac{\text{Real } gross\ output\ (or\ sales, GO)}{Employment\ (L)}$
	Based on value added: *LP_VA*	$\frac{\text{Real value added (VA)}}{\text{Employment (L)}}$
Multi-factor productivity (*MFP*): index-numbers[ii]	Solow-residual *MFP_S* with country and industry specific labour/intermediate shares (β_l^S, β_{ii}^S: median of shares across each year by country and industry)	$va - \beta_l^S l - (1 - \beta_l^S)k$ *or* $go - \beta_l^S l - \beta_{ii}^S ii - (1 - \beta_l^S - \beta_{ii}^S)k$
	Solow-residual *MFP_SW* with industry specific labour/intermediate shares ($\beta_l^{Sw}, \beta_{ii}^{Sw}$: median across countries and years by industry)	$va - \beta_l^{Sw} l - (1 - \beta_l^{Sw})k$ *or* $go - \beta_l^{Sw} l - \beta_{ii}^{Sw} ii - \left(1 - \beta_l^{Sw} - \beta_{ii}^{Sw}\right)k$
	Superlative index *MFP_SI*, following Griffith et al. (2009) [iii]	$\widetilde{va} - \beta_l^{SI}\tilde{l} - (1 - \beta_l^{SI})\tilde{k}$ *or* $\widetilde{go} - \beta_l^{SI}\tilde{l} - \beta_{ii}^{SI}\widetilde{ii} - (1 - \beta_l^{SI} - \beta_{ii}^{SI})\tilde{k}$
	Superlative index MFP growth: *MFP_GR*[iv]	$\Delta va - \beta_l^{GR}\Delta l - (1 - \beta_l^{GR})\Delta k$ *or* $\Delta go - \beta_l^{GR}\Delta l - \beta_{ii}^{GR}\Delta ii - (1 - \beta_l^{GR} - \beta_{ii}^{GR})\Delta k$
Multi-factor productivity (*MFP*): production function based[ii]	Residuals from Ordinary Least Squares estimates *MFP_OLS*	$va - \beta_l^{OLS}\, l - \beta_k^{OLS} k$ *or* $go - \beta_l^{OLS} l - \beta_{ii}^{OLS} ii - \beta_k^{OLS} k$
	Residuals from Ordinary Least Squares with firm fixed effect estimates *MFP_FE*	$va - \beta_l^{FE}\, l - \beta_k^{FE} k$ *or* $go - \beta_l^{FE} l - \beta_{ii}^{FE} ii - \beta_k^{FE} k$
	Residuals from Wooldridge (2009) type estimates *MFP_W*	$va - \beta_l^{W}\, l - \beta_k^{W} k$
	Residuals from Ackerberg et al. (2015) type estimates *MFP_ACF*	$va - \beta_l^{ACF}\, l - \beta_k^{ACF} k$

Notes: All output and input variables are firm and year specific, but for simpler exposition, firm and time indices are omitted. i) Small letters (*va, l, k*) denote natural logarithms. ii) For simpler exposition, MFP measures are defined in natural logarithms. iii) Variables with a tilde on top ($\widetilde{va}, \tilde{l}, \tilde{k}$) indicate log-differences from the industry and year mean value. $\beta_l^{SI}, \beta_{ii}^{SI}$ denote the average of the labour and intermediate share in the reference firm (taken to be the geometric average across firms in the same country*industry) and the current firm. iv) Variables with a Δ indicate log-growth. $\beta_l^{GR}, \beta_{ii}^{GR}$ denote the two-period average (e.g.: $\beta_l^{GR} = \frac{\beta_{l,t}+\beta_{l,t-1}}{2}$) of the firm-level labour and intermediate share.

Annex 2: Industry, Age and Size class definitions

Table 6. Sectoral Aggregation

STAN A38 aggregation based on ISIC Rev.4 classification	STAN A7 aggregation
01 to 03 AGRICULTURE, FORESTRY AND FISHING	**Agriculture**
05 to 09 Mining and quarrying	**Mining**
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 equipment	
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	**Utilities**
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	**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	**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	

Table 7. Size classes[33]

Size Class Name	Description
Size Class 21	Two size classes with threshold at 5 employees: vMicro (<5), SmallMediumLarge (>=5)
Size Class 2121	Two size classes with threshold at 5 employees but statistics not reported for the first group: SmallMediumLarge (>=5)
Size Class 22*	Two size classes with threshold at 10 employees: Small (<10), MediumLarge (>=10)
Size Class 2222*	Two size classes with threshold at 10 employees but statistics not reported for the first group.
Size Class 2901*	Two size classes, with firms with 0 labour input and with labour input below the main threshold for analysis (1.5FTE or 2HC by default).
Size Class 2999*	Two size classes, used for micro-entrants only, with firms labour input below the main threshold for analysis (1.5FTE or 2HC by default) and firm with labour input above the main threshold and below 10.
Size Class 42	Four size classes with thresholds at 10, 50, and 250 employees: Micro (1-9), Small (10-49), Medium (50-249), Large (250+)
Size Class 4222	Four size classes with thresholds at 10, 50, and 250 employees but statistics not reported for the first group: Small (10-49), Medium (50-249), Large (250+)
Size Class 52*	Five size classes with thresholds at 10, 20, 50, and 250 employees: Micro (1-9), Small (10-19), MediumSmall (20-49), Medium (50-249), Large (250+)
Size Class 5222*	Five size classes with thresholds at 10, 20, 50, and 250 employees but statistics not reported for the first group: Small (10-19), MediumSmall (20-49), Medium (50-249), Large (250+)
Size Class 71*	Seven size classes with thresholds at 5, 10, 20, 50, 100 and 250 employees: vMicro (<5), Micro (5-9), Small (10-19), MediumSmall (20-49), Medium (50-99), MediumLarge (100-249), Large (250+)
Size Class 81	Eight size classes with thresholds at 5, 10, 20, 50, 100, 250 and 500 employees: vMicro (<5), Micro (5-9), Small (10-19), MediumSmall (20-49), Medium (50-99), MediumLarge (100-249), Large (250-499), vLarge (500+)

Table 8. Age classes

Age Class Name	Description
Age Class 21*	Two age classes with threshold at 5 years: Young (0-5), Old (6+)
Age Class 31	Three age classes with thresholds at 0 and 10 years: Birth (0), Young (1-10), Old (10+)
Age Class 32	Three age classes with thresholds at 2 and 5 years: Startup (0-2), Young (3-5), Old (5+)
Age Class 51*	Five age classes with thresholds at 1, 2, 5 and 10 years: Entry, Start-ups (1-2 years old), Young (3-5), Mature (6-10), Old (10+)

[33] Classes marked with an asterisk are used in the output files, other classes are defined in the code (e.g., for weighting purposes) but are not used for outputs in the current version of the code.

Annex 3: Thematic Glossary

Units

- **Legal unit** – A legal unit is a legal entity of public or private law. This legal entity can be: 1) A unit that is recognised by law or society independently of the persons or institutions that own it, or 2) natural persons who are engaged in an economic activity in their own right.
- **Administrative unit** – a unit specifically designed for the purposes of conforming to an administrative regulation, for example VAT or Social Security.
- **Statistical unit** – unit defined for statistical purposes, for which information is sought and for which statistics are compiled.

Types of statistical units

- **Enterprise** (statistical unit): A legal unit (or the smallest set of legal units) that produces goods or services and that has autonomy with respect to financial and investment decision-making.
- **Establishment** (statistical unit): An enterprise or part of an enterprise that is situated in a single location and in which only a single (non-ancilliary) productive activity is carried out or in which the principal productive activity accounts for most of the value added.
- **Kind-of-activity unit** (statistical unit): An enterprise or part of an enterprise that engages in only one kind of productive activity or in which the principal productive activity accounts for most of the value added.
- **Local unit** (statistical unit): An enterprise or part of an enterprise (e.g., a workshop, factory, warehouse, office, mine or depot) that is engaged in productive activity at or from one location.
- **Enterprise group** (statistical unit): An enterprise group is an association of enterprises bound together by legal and/or financial links. A group of enterprises can have more than one decision-making centre, especially for policy on production, sales and profits. It may centralise certain aspects of financial management and taxation. It constitutes an economic entity which is empowered to make choices, particularly concerning the units which it comprises.

Labour input

- **Employees**: Persons who work for an employer (corporate enterprise or sole proprietorship) on the basis of a contract of employment and receive compensation in the form of wages, salaries, fees, gratuities, piecework pay or remuneration in kind.

  The contract is an agreement between an enterprise (the employer) and a person (the employee), which may be formal or informal, normally entered into voluntarily by both parties, whereby the person works for the enterprise in return for compensation in cash or in kind.

  A worker is considered to be an employee of a particular unit if he or she receives a wage or salary from the unit regardless of where the work is performed (even from remote locations). A worker from a temporary employment agency is considered to be the agency's employee and not that of the business unit to which he or she is assigned.

  Part-time workers, seasonal workers, persons on strike or on short-term leave are considered employees, while volunteers or workers on long-term leave are excluded.

- **Persons engaged** (or “persons employed” in SBS terminology): The total number of persons who work in the observation unit including wage-earners and self-employed persons (i.e. working proprietors, partners working regularly in the unit and unpaid family workers), as well as persons who work outside the unit but who belong to it and are paid by it (e.g. sales representatives, delivery personnel, repair and maintenance teams). It excludes manpower supplied to the unit by other enterprises, persons carrying out repair and maintenance work in the enquiry unit on behalf of other enterprises, as well as those on compulsory military service.
- **Self-employed**: Persons who are the sole owner, or joint owners, of the unincorporated enterprises in which they work. Self-employed persons are classified here if they are not also in paid employment which constitutes their principal activity: in the latter case they are classified as employees. Self-employed persons also include the following categories: unpaid family workers, outworkers and workers engaged in production undertaken entirely for their own final consumption or capital formation, either individually or collectively.

Output

- **Gross output** (or production): The production value is defined as turnover or revenue from sales of goods and rendering of services; plus or minus the changes in stocks of finished products, work in progress and goods and services purchased for resale; minus the purchases of goods and services for resale (only for the goods and services sold during the reporting period and excluding the costs of storage and transport of the goods purchased for resale); plus capitalised production; plus other (operating and extra-ordinary) income (excluding subsidies). Income and expenditure classified as financial or as revenue in the form of interests and dividends in company accounts is excluded from production value. Included in purchases of goods and services for resale are the purchases of services purchased in order to be rendered to third parties in the same condition.
- **Turnover** (or sales): Turnover comprises the totals invoiced by the business during the reference period, and it corresponds to market sales of goods and services supplied to third parties.

  Turnover includes all duties and taxes on the goods or services invoiced by the unit with the exception of the value added tax (VAT) invoiced by the unit vis-a vis its customer, and other deductible taxes directly linked to turnover.

  It also includes all other charges (transport, packages, etc.) passed on to the customer. Reduction in prices, rebates and discounts as well as the value of returned packaging must be deducted.
- **Value added at basic prices**: Value added at basic prices is calculated from the production value plus subsidies on products less the purchases of goods and services (other than those purchased for resale in the same condition) plus or minus the change in stocks of raw materials and consumables less other taxes on products which are linked to turnover but not deductible. It represents the value added by the various factor inputs in the operating activities of the unit concerned.
- **Value added at market prices**: Value added at market prices is output valued at market prices less intermediate consumption valued at purchasers’ prices.

- **Value added at factor cost**: Value added at factor cost is the gross income from operating activities after adjusting for operating subsidies and indirect taxes. It can be calculated from turnover, plus capitalised production, plus other operating income, plus or minus the changes in stocks, minus the purchases of goods and services, minus other taxes on products which are linked to turnover but not deductible, minus the duties and taxes linked to production.

  Summary of VA:

  > Factor cost
  > + Taxes on production, less subsidies on production
  > = Basic prices
  > + Taxes on products, less subsidies on products
  > = Market prices

Other variables

- **Labour costs** (or compensation of employees or "employee benefits expense" in SBS terminology): The total remuneration, in cash or in kind, payable by an employer to an employee (regular and temporary employees, as well as home-workers, but not temporary agencies workers) in return for work done by the latter during the reference period.
  Labour costs are made up of wages, salaries and employers' social security costs. They include taxes and employees' social security contributions retained by the employer, as well as the employer's compulsory and voluntary social contributions.
- **Operating profit**: In accounting terms, EBIT ('earnings before interest and taxes') includes all incomes and expenses (operating and non-operating) except interest expenses and income tax expenses.
  EBIT according to the "cost of goods sold approach" = Revenues – Costs of goods sold + Other income – Distribution costs – Administrative expenses – Other expenses
  EBIT according to the "nature of expense method" = Revenues + Other income +/- Changes in inventories of finished goods and work in progress – Raw materials and consumables used – Employee benefits expense – Depreciation and amortisation expense – Other expenses (including purchased services)
  Alternative proxy measures: EBITA / EBITDA = Revenues (turnover) – Intermediate inputs – Labour costs – Depreciation (– Amortisation)
- **Exports valued at factor cost:** Nominal export turnover (see definition of turnover; unadjusted exports) excluding indirect taxes, tariffs etc., but including subsidies on products and production. (The unadjusted value represents the value from the balance sheet or customs source that depending on the source may already be adjusted by the country specific annual threshold, but not the country specific maximum threshold that will be applied by the code.)
- **Import value:** Expenses for imported products and services acquired valued at basic prices, i.e., excluding non-VAT taxes or tariffs on products but including subsidies on products. Imports include purchases of goods intended for resale.