Micro Data Infrastructure (MDI)
Methodological Report on Cross-country Analysis of Newly Developed Firm-level Indicators*

Eric Bartelsman
Stichting VU
Tinbergen Institute

Mirja Hälbig
Halle Institute for Economic Research (IWH)

Filippo di Mauro
NUS Singapore
CompNet

Abstract This report summarizes the lessons drawn from experimenting with new data sources to improve our understanding of observed productivity patterns. Based on pilot studies with several National Statistical Institutes (NSIs), we set up a prototype Micro Data-Infrastructure (MDI) that allows cross-country comparative analysis of firm-level micro data. This infrastructure provides researchers with (i) harmonization tools to run common code at multiple sites, even if the details of the data, the technical infrastructure and confidentiality practices differ across NSIs, and (ii) a set of analytical tools that simplify research design to study productivity dynamics. While the technical components of the infrastructure are in place, details of the institutional set-up and funding to ensure longevity of the infrastructure are still under development.

*We thank Márta Bisztray, Peter Bøegh Nielsen, Mirosław Błażej, Kalle Emil Holst Hansen, Wolfhard Kaus, Sini Liukonnen, Tim Peeters, Andreas Poldahl, Michael Polder, Jan Olav Rørhus, Sébastien Roux and Markus Zimmermann for useful discussions and ongoing support. Special thanks go to Jing Chen, Sergio Inferrera and Alessandro Zona Mattioli for their valuable contributions to the construction of the infrastructure. We acknowledge funding by the European Union's Horizon 2020 research and innovation program, grant agreement No 822390 (MICROPROD). Email: e.j.bartelsman@vu.nl (Bartelsman), mirja.haelbig@iwh-halle.de (Hälbig), bizfdm@nus.edu.sg (di Mauro)

1 Introduction

Micro data—particularly when longitudinally linked and broadly covering transactions, behaviors, and characteristics—have become an essential tool for economic research and policy analysis. Data from administrative processes and from micro-level surveys are increasingly being integrated by National Statistical Institutes (NSI) to generate published statistical indicators. Owing to confidentiality rules, the micro data themselves are not published. At individual NSIs, some of the micro-level data are available to qualified (national) researchers, under legal conditions that safeguard confidentiality of persons, households, and firms.

The purpose of the work detailed in this report is twofold. To start, research has been undertaken to experiment with indicators of intangible assets and new technology and their relation to productivity, using linked micro data. Next, development has taken place to set up a prototype Micro Data-Infrastructure (MDI) that allows cross-country comparative analysis of firm-level micro data. While the research applications being piloted in the MDI—in collaboration with the NSIs of Denmark, Finland, Netherlands, Norway and Sweden—pertain to the relations between new technology, globalisation, and productivity developments, many different areas of research and policy analysis can benefit from the proposed statistical infrastructure[1].

The main objective of the proposed MDI is to improve the ability of researchers to conduct cross-country comparative analysis using firm-level micro data. The infrastructure will harmonize the processing of underlying firm-level surveys and administrative sources across countries by using a meta-data translation layer. Further, a suite of analytical tools are made available through the MDI that will simplify research design to study productivity dynamics and experiment with new determinants and conceptual measures of productivity. Further, the research infrastructure is designed in a way that shields the external researchers from the cumbersome process of gaining legal access to firm-level data at each individual NSI by making use of a network of NSI partners.

[1]The French NSI *INSEE* has agreed on participating in the MDI in next phases, and we initiated discussions with further NSIs (among others Germany, Hungary, and Poland).

This document is structured as follows: We start with describing issues related to the measurement of intangible assets and present novel findings and lessons drawn regarding modeling intangibles in the production process. Next, we describe the current set up of the MDI, the underlying firm-level data sources, the analytical modules designed by Microprod, and prospects for external researchers. We conclude with an outlook, remaining challenges and a proposal for a permanent MDI that can be continued after completion of the Microprod project.

2 Measuring Intangible Assets

While early research on productivity growth mainly focused on technological assets built up through R&D expenditures or measured by patents (Griliches 1979), Corrado et al. (2005) developed a systematic framework to categorize the firm-level expenditures that account for investments into a complete stock of intangible capital. In this framework intangible capital comprises investments in research and development (R&D), software, patents, as well as branding and organizational capital (Corrado et al. 2005; Corrado et al. 2009).

According to business accounting standards, such as IFRS or local GAAP, some components of what productivity researchers now consider as firm-level intangible investment are classified as expenditure. Further, comparisons of measures of intangible capital from public balance sheet filings show differences in interpretation of these standards across firms (Covarrubias et al, 2019). According to national accounting standards, SNA 2008 and ESA 2010, only a limited range of investments are included in economy-wide intangible assets: R&D, mineral exploration, computer software and databases, and entertainment, literary and artistic originals. In Microprod, we triangulate between accounting standards, available data sources, and growth theory to experiment with proxies for intangible assets that can be used to study productivity developments. The following paragraphs present these experiments developed in Microprod that operationalise certain aspects of intangibles.

Intangibles from financial data Both balance sheet data as well as profit & loss statements can be consulted to calculate measures of intangibles. Bisztray et al. (2020) use the accounting measure for intangible fixed assets that consists of "...mineral exploration, computer software, entertainment, literary or artistic originals and other intangible fixed assets intended to be used for more than one year"[2]. While the accounting measure aims to capture the entire stock of intangible capital, this measure does not perfectly capture the economic concept of intangible capital. First, asset book values do not necessarily reflect the economic value of the incorporated assets. Second, in order to enter the firm's balance sheet, the respective intangible asset's value and lifetime need to be quantifiable. This has the effect that assets developed in-house in contrast to acquired assets seldomly stated in the balance sheet, but rather show up in profit & loss statements as expenses. Using income statements, Altomonte et al. (2020) derive firm intangible investments from firm expenditure on fixed costs, calculated as net revenues minus operating profits.[3]

Intangibles from investment data Using cost structure and investment surveys, Kaus et al. (2020) compute an intangible capital stock for each firm, consisting of expenditure for R&D, concessions, licenses, patents and trademarks and acquired software for 14,000 German manufacturing firms per year from 2009-2015. The authors use the Perpetual Inventory Method (PIM) to transform yearly investment flows into capital stocks as follows

$$K^{\theta}_{ijt} = (1 - \delta^{\theta}_{jt}) \times K^{\theta}_{i,j,t-1} + I^{\theta}_{ijt} \tag{1}$$

where δ^{θ}_{jt} denots the yearly depreciation rate and I^{θ}_{jt} yearly real investment for capital good $\theta \in$ (machines, buildings, software, patents, R&D) for firm i in industry j in year t. To transform nominal to real investment flows, the authors use price deflators provided by National Accounts, which also include separate deflators for investments in machines,

[2]Eurostat, "European System of Accounts - ESA 1995", Office for Official Publications of the European Communities, Luxembourg, 1996

[3]This approximation is related to measuring expenditures on intangibles based on *Selling, General and Administrative Expenses (SG&A)* which is – besides *Costs of Goods Sold (COGS)* – the second major component of costs and includes all intangible-building activities (e.g., R&D, Advertising and IT staff expense) (Gutiérrez and Philippon 2017; Covarrubias et al. 2019).

buildings, and intellectual capital. While depreciation rates for tangible capital can be derived from the National Accounts, for intangible capital the authors use fixed rates for all industries and years, that is 33% for software, and 20% for patents and R&D (Corrado et al. 2009). The authors construct an initial capital stock using the average investment during the firm's first three years in the data

$$K^{\theta}_{i,j,t=0} = \frac{1}{3} \times \sum_{t=1}^{3} \frac{I^{\theta}_{ijt}}{\delta^{\theta}_{jt} + g^{\theta}_{j}} \tag{2}$$

where r^{θ}_{ij} is the geometric mean of the annual growth rates of the different investment types in the National Accounts. The results show that although the dispersion of productivity decreases slightly when intangible capital is accounted for in the production function, a large part of productivity dispersion remains. The authors conclude that other intangibles not contained in the investment data such as organisation and branding capital or management quality are additional factors explaining productivity dispersion.

Innovation, R&D and ICT usage surveys In order to complement the information derived from accounting statistics and administrative investment data, microprod experiments with additional survey data. Bisztray et al. (2020) use the community innovation survey (CIS) and ICT use survey, which provides them with a rich set of indicators on qualitative and quantitative firm innovative activity and innovation output by category, and detailed qualitative data on ICT use. The key conceptual advantage of the CIS is that it includes different innovative outputs and asks for the inputs used. This allows to estimate a knowledge production function that directly links innovative outputs to inputs of the following form

$$\ln N_{it} = \omega_i + \beta_L \ln L_{it} + \beta_L \ln K_{it} + \beta_R \ln R_{it} + \eta_{it} \tag{3}$$

where N_{it} is firm innovative output measured by sales from innovative products, and L_{it} and K_{it} are labor and capital inputs related to innovation, proxied by R&D expenditures (in-house and purchased) and other innovative investments (including investment in machines, software, purchased other external knowledge, design, training and marketing). The knowledge stock R_{it} is proxied by the stock of patents. The residual term ω_i captures

how effectively the firm transforms innovative inputs into innovative outputs. This measure of firm innovativity proves to be highly correlated with conventional measures of firm total factor productivity (TFP), suggesting that (i) both are confounded by basic firm capabilities or (ii) that innovativity captures the unmeasured part of intangible capital.

The ICT use survey contains various dimensions of ICT usage, thus capturing organizational capital in terms of ICT capabilities in the firm's internal operations as well as in the relationship with buyers and suppliers. Bisztray et al. (2020) classify ICT usage in six categories suspected to be related to firm productivity: providing IT training to employees, ICT use in within-firm processes and in communication with buyers or suppliers, use of website, social media and cloud computing. For each of the categories, he authors create indices by counting positive answers to the related questions on the firm-year level. In addition, the authors use principal component analysis to further reduce the dimensionality of the data, resulting in one principal component capturing the intensity of firm ICT usage. Their results show that this measure of firm ICT capability is highly correlated with firm output, suggesting that firm capabilities captured by information technology provide additional information on firm intangible capital.

While the ICT use survey only contains qualitative information on firm ICT use, Smeets and Warzynski (2020) use a novel dataset on firms' ICT investment for Denmark, which allows to disentangle ICT investment into three categories—hardware, software and communication equipment—and analyse their respective effects on firm growth and productivity. Their findings show that all three components of ICT spending at the firm level correlate strongly with firm growth and productivity, but also suggest a strong selection effect and little variation over time in the spending heterogeneity across firms.

3 Microdata Infrastructure

In this section we describe the technical details of the development of the research infrastructure that has been built at the NSIs, and the pilot analysis embarked upon during the Microprod project.

3.1 Micro level datasets

In the EU, Eurostat regulations mostly harmonize (aggregate) output of statistical indicators in each country. In recent years there has been some progress in harmonizing micro-level data, for example by regulations on Business Registers (Regulation (EC) No 177/2008) and surveys on ICT usage in business (Regulation (EC) NO 808/2004), as well as by Eurostat model questionnaires, .e.g. for the Community Innovation Survey (with voluntary participation). With the Business Register as a 'backbone', NSIs have been able to link information from these datasets and other survey or register-based information at the individual enterprise-level (in this document loosely referred to as 'firm-level'). The result is an incredibly rich set of information which allows us to understand for instance how a variety or disparate of factors affect productivity at the firm level. In the following, the underlying data sources available at the MDI are introduced.

Statistical Business Register (BR) The statistical business register (BR) plays a central role in the production of business statistics and is the starting point for establishing statistical survey frames. The BR contains information on identifying characteristics such as ID numbers, names and addresses, demographic characteristics, economic activity, legal form and institutional sector code as well as information on control and ownership relations for enterprises, their local and legal units and enterprise groups. In MDI, the BR serves as a 'backbone' or connection between various surveys and administrative datasets.

Structural Business Statistics (SBS) The Structural Business Statistics (SBS) describe the economic activities within the business economy, including industry, construction, distributive trade and services. SBS indicators at the detailed sector level are transmitted to Eurostat and published by all European Statistical System (ESS) members (EU Member States, Norway and Switzerland, some candidate and potential candidate countries). Harmonization of the SBS has taken place regarding the detail and coverage of the sectors (now NACE 2.1) and the statistical definition of the transmitted indicators (Commission Regulation (EC) No 250/2009). Generally, the SBS indicators in each country are collected at the level of individual enterprises engaged in economic activity.

The firm-level sources for each of the SBS indicators vary, across indicators, sectors and across countries, but possibly also across statistical units. For example, business surveys could be used to collect data for the indicator 'production value' for manufacturing firms while administrative data could be used to collect production value for firms in the telecommunications industry. The source for the indicator 'wages and salaries' could be administrative data on payroll taxes, or could be collected through a statistical survey. For gross investment expenditures the source data frequently are investment surveys amongst large firms with the small firms' contribution to the sector aggregate imputed or estimated, for example using a supply-use framework. The SBS includes information on various types of tangible investment as well as the following intangible investments (i) concessions, patents, licenses, trade-marks and similar rights, (ii) purchased software and (iii) total intra-mural R&D expenditure.

For the purpose of the micro-data infrastructure, a firm-level SBS is created in each country for a common set of output and input indicators, using the underlying firm-level sources available in each country. Care is taken to flag for the researcher when data are observed rather than imputed. The main variables in the firm-level SBS are monetary values, or as counts (for example, persons employed).

Community Innovation Survey (CIS) The Community Innovation Survey (CIS) is part of the EU science and technology statistics and provides mostly qualitative information on firm innovative activity. Surveys are carried out every two years by EU member states and a number of ESS member countries on a voluntary basis. The harmonized survey contains information on the types of innovation and various aspects of the development of an innovation, such as the type of funding and innovation expenditures. The CIS covers both innovation outputs and the innovative process and inputs (type of funding, R&D expenditure) and distinguishes four innovation types: process, product, organizational, marketing, thus covering both innovative property as well as capabilities and organizational capital. Additionally, the CIS asks about the novelty of the innovation, i.e. whether it is new for the market, new to the country, developed by the firm or was adopted, and thus provides information about the innovative value.

ICT usage/ E-Commerce Survey (ICTEC) The Community survey on ICT usage and e-commerce in enterprises is an annual survey conducted since 2002, which collects information on the use of information and communication technology, the internet, e-government, e-business and e-commerce in enterprises. Like the CIS, the EC survey contains mostly qualitative data. The ICT use survey measures various dimensions of firm technology use. Besides software and databases being considered as an integral part of intangibles, the adoption of certain technologies also provides information about firms' organizational capital and ICT capabilities both in the firms' internal operations and regarding the firms' supplier and buyer relationships. The qualitative information in the survey can be used to construct an ICT intensity index which allows for variation in the underlying source variables, thereby overcoming the issue with changing survey questions and the saturation of certain variables over time (Bartelsman et al. 2018).

International Trade Statistics Firm-level statistics concerning exports and imports are the International Trade in Goods Statistics (ITGS) and International Trade in Services Statistics (ITSS). International trade in goods statistics (ITGS) measure the value and quantity of goods traded between EU Member States (intra-EU trade) and goods traded by EU Member States with non-EU countries (extra-EU trade) broken down by types of goods (Combined Nomenclature) and by partner countries. The providers of statistical information differ between intra and extra EU-trade. In the first case, it corresponds to all taxable persons reporting transactions exceeding a certain threshold fixed by member states; in the second one, it corresponds to administrative data from the customs declarations lodged by natural or legal persons in the customs administration. International Trade in Services Statistics (ITSS) typically cover trade in services, i.e. transactions paid for the services that have taken place between the residents and non-residents.

Foreign Affiliate Statistics (FATS) The Foreign Affiliate Statistics is distinguished into inward FATS, i.e. the activity of foreign affiliates resident in the compiling country, and the outward FATS, that is, the activity of foreign affiliates abroad but controlled by the compiling country. The FATS allows to qualitatively assess the degree of economic

activity of a domestic enterprise abroad and identify foreign-controlled firms.

Other sources Further data sources are available at the NSIs, but not yet included in the MDI due to being less harmonized. The sources and their possible contribution are briefly described in the following.

A promising and interesting source is **linked employer-employee data (LEED)** that cover the working populations' characteristics like employment relations, income and education and socio-demographic characteristics. For example, linked employer-employee data can be used to analyse complementarity between firm human capital and intangible assets (Piekkola 2016).

The **International Sourcing Survey (ISS)** gathers data on international organisations and sourcing of business functions in 16 European countries, covering the period 2014-2016 or 2015-2017, depending on the country. The survey results cover nearly 60,000 businesses each with more than 50 persons employed. However, since the survey is still in pilot stage, the survey design varies across countries.

Financial data provides information on firms' financial assets and liabilities. While for nearly all pilot countries, firm financial data is available at the NSI, for some countries it is only available at the respective National Central Bank (NCB) (e.g. Germany). Balance sheet data contain an accounting measure which aims to capture the entire stock of intangible capital. However, intangibles can only appear on the balance sheet of a company if their value is clearly identifiable, with the shortcomings that (i) acquired assets are much more likely to enter the firm's balance sheet, (ii) the item covers only certain aspects of the economic concept of intangibles and (iii) does not necessarily reflect the economic value of the incorporated intangible assets due to accounting principles and depreciation rules (Bisztray et al. 2020). Conversely, the profit & loss statement includes information on expenditures on intangibles such as *Sales, General and Administrative Costs (SG&A)*.

3.2 Research Infrastructure

In this section we describe the technical details of the research infrastructure developped to deploy at the NSIs during the Microprod project.

Metadata, Modules and Tools In addition to the above mentioned firm-level datasets, the microdata infrastructure consists of tools and modules to harmonize the data across countries by using meta-data translation, and will simplify research design through a common set of data preparation and analytical tools. Thus, we provide researchers with the ability to generate one research design that can be executed in each country, even if details of the underlying datasets vary.

Comprehensive metadata, to be maintained by NSI staff, allow linking of the appropriate datasets and the mapping of indicators in each country to present a common format across countries. The user needs to choose which one or more of the available linked datasets to use, and must refer to the available variables from a common data listing. Selection and filtering of the data takes place via a file in which the researcher indicates the variables she needs. Appendix A.1 shows a mapping from the variables from the underlying datasets to a common name to be used by program code of infrastructure users. The NSI will maintain the mapping of the indicator description to a variable name and datasource (columns NSIname and DatasSource), while the user can refer to the variablename MPname that will be common across countries.

Based on the selection of variables, the code reads the required firm-level data sources and links them to the BR. At this point, firm-level sample weights could be generated, using a re-weighting algorithm that compares firms available in a linked dataset with the universe of firms in the BR. Upon reading the datasets, the code maps the variables available in the different datasets to a common nomenclature. This procedure allows each NSI to use their own variable names and classifications and the program code translates the NSI specific name for the variable name available for the common code. In addition, the code ensures the mapping of classifications (e.g. industry, product, region) to a common hierarchy. The program code reads in common industry-level deflator timeseries for each country for output, value added and intermediate inputs sourced from Eurostat and applies them in a uniform manner to the firm-level data.

Further, the MDI features data preparation and cleaning tools such as outlier detection programs, and aggregation tools for the high-dimensional product level trade data and foreign affiliate statistics. Additional analysis tools (for example for productivity estimation

or clustering) facilitate and ensure a comfortable use of the data.

The output of a researcher-written code module can consist of output datasets, aggregated to industry or other firm-level characteristics to avoid breaking confidentiality, or tables of analytical results (ie regression coefficients and diagnostics). Pre-programmed aggregation and output and documentation tools and obligatory disclosure routines, customized to each country's confidentiality practice, further reduce NSI staff workload for disclosure analysis. As an example, the first pilot module run by Microprod using the infrastructure prepared in the Netherlands, Denmark, Sweden, Norway and Finland, generated output datasets as documented in Appendix A.2.

Taking into account heterogeneity across NSIs and users with respect to the technical storage facilities and available analytical software, the tools are presently available in two different software languages (R and SAS).

Data Access At present, the research designs (computer code) prepared and tested by Microprod are run by staff at each NSI site, such that iterations between coding and viewing intermediate results is not possible. To decrease the burden to NSIs, we strive for remote access, if legally available.

To guarantee micro data access also to external researchers while avoiding overburdening the NSIs, we suggest that a consortium, for example CompNet, takes on an intermediary position between external researchers requiring cross-country comparative work and NSIs in filtering and assisting in the data application process. Access to the data could be structured in two ways.

- An external user could write an original module, partly based on modules and tools developed by the Microprod/CompNet team, if needed. Such routine could then be remotely executed on the NSIs' data by the Microprod/CompNet team and the output - checked for confidentiality - could then be provided to the user. The user could test the code on a mock sample database provided by the Microprod/CompNet team.

- In case the user is not interested in coding herself a routine, some already made packages could be provided by Microprod/CompNet to produce descriptives, gen-

eral trends, standard estimations, etc. In this case the user would only need to adjust some basic parameters, e.g. the aggregation level, the time coverage and the particular variable in mind. The output would then be some moments of the distribution of a variable of interest, cross country trends in productivity, etc. In perspective, we strive to provide a user interface for potential users to search relevant metadata and aggregate statistics.

In this way we could adapt to the needs of more advanced users, who have a very specific project in mind, and some more basic users or policy makers, who need more standard statistics.

3.3 Micro-aggregated data (CompNet)

As a second path towards cross-country analysis, a central part of the research infrastructure consists of a set of micro-aggregated indicators at different levels of aggregation - the CompNet vintages - that include measures of business performance and industry dynamics. These measures include typical aggregates, such as sums and means, but also higher moments of distributions of variables of interest, as well as moments from multivariate distributions. New data and productivity concepts generated by Microprod are and will continue to be implemented in the data collection process for the updates of the CompNet datasets.

4 Conclusion and Outlook

We have developed a research infrastructure for the decentralized execution of program code on longitudinally linked firm-level data in multiple countries. This infrastructure will allow custom research projects to tap into confidential firm-level data in multiple countries, and to extract and combine harmonized results for each country. The technical components of the infrastructure are in place. The next steps will entail designing an institutional set-up as well as exploring sources of funding to expand the infrastructure to more ESS member countries and to continue operating the infrastructure for use by researchers and policy analysts in the future.

References

Altomonte, Carlo, Domenico Favoino, Monica Morlacco, and Tommaso Sonno (2020). "Markups, Intangible Capital and Heterogenous Financial Frictions".

Bartelsman, Eric, Eva Hagsten, and Michael Polder (2018). "Micro Moments Database for cross-country analysis of ICT, innovation, and economic outcomes". In: *Journal of Economics & Management Strategy* 27.3, pp. 626–648.

Bisztray, Marta, Balazs Muraközy, and Dzsamila Vonnak (2020). "Analysis of the importance of intangible capital and knowledge for productivity measurement".

Corrado, Carol, Charles Hulten, and Daniel Sichel (2005). "Measuring Capital and Technology: An Expanded Framework". In: *Measuring Capital in the New Economy*. National Bureau of Economic Research, Inc, pp. 11–46.

— (2009). "Intangible Capital and U.S. Economic Growth". In: *Review of Income and Wealth* 55.3, pp. 661–685.

Covarrubias, Matias, Germán Gutiérrez, and Thomas Philippon (2019). *From Good to Bad Concentration? U.S. Industries over the past 30 years*. Working Paper 25983. National Bureau of Economic Research.

Griliches, Zvi (1979). "Issues in Assessing the Contribution of Research and Development to Productivity Growth". In: *The Bell Journal of Economics* 10.1, pp. 92–116.

Gutiérrez, Germán and Thomas Philippon (2017). *Declining Competition and Investment in the U.S.* Working Paper 23583. National Bureau of Economic Research.

Kaus, Wolfhard, Viktor Slavtchev, and Markus Zimmermann (2020). *Intangible capital and productivity: Firm-level evidence from German manufacturing*. IWH Discussion Papers 1/2020. Halle (Saale).

Piekkola, Hannu (2016). "Intangible Investment and Market Valuation". In: *Review of Income and Wealth* 62.1, pp. 28–51.

Smeets, Valerie and Frederic Warzynski (2020). "ICT, Firm Growth and Productivity".

A Appendix

A.1 Data and Variables

Table A.1 depicts the data and variables available in the MDI. Column *NSIname* depicts the exemplary variable name as in the original NSI dataset, while column *MPname* depicts the variable name homogenized across NSIs by common code developed by Microprod. Column *uniqueDim* shows the unique (combined) keys of each dataset (set to 1). These, together with the year variable that is added while reading and merging the datasets, uniquely identifies the observations. The column *indexDim* indicates the respective datasets index dimensions that can be used for the aggregation of the data.

Table A.1: Description of Variables

MPname	NSIname	DataSource	Description	Format	uniqueDim	indexDim	Type
firmid	ENT_ID	br	Unique enterprise identification	Character	1	0	
entgrp	ENTgrp_ID	br	Enterprise Group ID	Character	0	0	
admid	AD_ID	br	Administrative ID	Character	0	0	
birthyr	Start_Ent	br	Start date for the enterprise ID	Character	0	0	
	End_ent	br	End date for the enterprise ID	Character	0	0	
	START_ENTgr	br	Start date for the enterprise Group ID	Character	0	0	
	END_ENTgr	br	End date for the enterprise Group ID	Character	0	0	
lfo	LEGAL	br	Legal form of the enterprise ID	Number	0	0	
nace	NACE_M	br	Main activity of the enterprise (NACE 4-digit)	Character	0	1	
soe	PUB	br	Ownership of the enterprise (private/public)	Number	0	0	
	START_NACE_M	br	Start date for the main activity	Character	0	0	
	DEMO_REL	br	Information on demographic relations (mergers and acquisitions etc.)	Character	0	0	
firmid	ENT_ID	sbs	Enterprise ID (identifikation number)	Character	1	0	
nq	SBS_12110	sbs	Turnover	Number	0	0	
nv	SBS_12150	sbs	Value added at factor cost	Number	0	0	
ngos	SBS_12170	sbs	Gross operating surplus	Number	0	0	
nm	SBS_13110	sbs	Total purchases of goods and services	Number	0	0	
pay	SBS_13310	sbs	Personnel costs	Number	0	0	
wages	SBS_13320	sbs	Wages and salaries	Number	0	0	
persons	SBS_16110	sbs	Number of persons employed	Number	0	0	
emp	SBS_16130	sbs	Number of employees	Number	0	0	
fte	SBS_16140	sbs	Number of employees in full-time equivalents	Number	0	0	
imputesbs	SBS_Type	sbs	Code to show if data (unit) is observed or imputed in SBS	Number	0	0	

Continued on next page

Table A.1: Description of Variables

MPname	NSIname	DataSource	Description	Format	uniqueDim	indexDim	Type
ni_intan	SBS_15429	sbs	Gross investment in concessions, patents, licences, trade marks and similar rights	Number	0	0	
ni_sw	SBS_15441	sbs	Investment in purchased software	Number	0	0	
ni_rd	SBS_22110	sbs	Total intra-mural R & D expenditure	Number	0	0	
rdemp	SBS_22120	sbs	Total number of R & D personnel	Number	0	0	
nnrg	SBS_20110	sbs	Purchases of energy products (in value)	Number	0	0	
ni_eq	SBS_15150	sbs	Gross investment in machinery and equipment	Number	0	0	
ni_tan	SBS_15110	sbs	Gross investment in tangible goods	Number	0	0	
firmid	ENT_ID	itgs	Unique enterprise identification	Character	1	0	
ntrade	STAT_VALUE_ITGS	itgs	Trade amount	Number	0	0	
exim	Flow_ITGS	itgs	Code to distinguish between import and export	Number	1	0	
ctry	CL_AREA_GEO_ITGS	itgs	Partner country (country of origin/destination) (2 letter code)	Character	1	1	
ctrygrp	Country_grp_ITGS	itgs	Aggregation of destination countries to groups	Number	0	1	
cno08	CN08	itgs	Product nomenclature CN08 8-digit	Character	1	1	
	BEC	itgs	Products aggregated to BEC-codes	Character	0	1	
imputeitgs	ITGS_type	itgs	Code to show if data (unit) is observed or imputed in ITGS	Number	0	0	
firmid	ENT_ID	its	Unique enterprise identification	Character	1	0	
ntrade	STAT_VALUE_ITS	its	Trade amount	Number	0	0	
exim	Flow_ITS	its	Code to distinguish between import and export	Number	1	0	

Continued on next page

Table A.1: Description of Variables

MPname	NSIname	DataSource	Description	Format	uniqueDim	indexDim	Type
ctry	CL_AREA_GEO_ITS	its	Partner country (country of origin/destination)	Character	1	1	
ctrygrp	Country_grp_ITS	its	Aggregation of destination countries to groups	Number	0	1	
ebops3	Bopitem	its	Service nomenclature EBOPS 3-digit	Character	1	1	
	BUS_service	its	Aggregation of the EBOPS nomenclature to business functions	Number	0	1	
imputeits	ITS_type	its	Code to show if data (unit) is observed or imputed in ITS	Number	0	0	
firmid	ENT_ID	cis	Unique enterprise identification	Character	1	0	
ho	HO	cis	Country of head office(From IFATS)	Character	0	1	
mareur	MAREUR	cis	Other EU/EFTA/CC market	Number	0	0	bool
maroth	MAROTH	cis	All other countries	Number	0	0	bool
inpdgd	INPDGD	cis	Introduced onto the market a new or significantly improved good	Number	0	0	bool
inpdsv	INPDSV	cis	Introduced onto the market a new or significantly improved service	Number	0	0	bool
newmkt	NEWMKT	cis	Did the enterprise introduce a product new to the market	Number	0	0	bool
turnmar	TURNMAR	cis	% of turnover in new or improved products introduced during 2006-2008 that were new to the market	Number	0	0	pct
inpspd	INPSPD	cis	Introduced onto the market a new or significantly improved method of production	Number	0	0	bool

Continued on next page

Table A.1: Description of Variables

MPname	NSIname	DataSource	Description	Format	uniqueDim	indexDim	Type
inpslg	INPSLG	cis	Introduced onto the market a new or significantly improved logistic, delivery or distribution system	Number	0	0	bool
inpssu	INPSSU	cis	Introduced onto the market a new or significantly improved supporting activities	Number	0	0	bool
inpsdv1	INPSDV1	cis	Who mainly developed these processes – your enterprise by itself	Number	0	0	bool
inpsdv2	INPSDV2	cis	Who mainly developed these processes – your enterprise together with other	Number	0	0	bool
inpsdv3	INPSDV3	cis	Who mainly developed these processes – your enterprise by adapting or modifying processes	Number	0	0	bool
inpsdv4	INPSDV4	cis	Who mainly developed these processes – other enterprises or institutions	Number	0	0	bool
rrdin	RRDIN	cis	Engagement in intramural R&D	Number	0	0	bool
rdeng	RDENG	cis	Type of engagement in R&D	Number	0	0	
rrdinx	RRDINX	cis	Expenditure in intramural R&D (in national currency)	Number	0	0	
rrdexx	RRDEXX	cis	Extramural R&D (in national currency)	Number	0	0	
rmacx	RMACX	cis	Expenditure in acquisition of machinery (in national currency)	Number	0	0	
rtot	RTOT	cis	Total of these four innovation expenditure categories (in national currency)	Number	0	0	

Continued on next page

Table A.1: Description of Variables

MPname	NSIname	DataSource	Description	Format	uniqueDim	indexDim	Type
funloc	FUNLOC	cis	Public funding from local or regional authorities	Number	0	0	bool
fungmt	FUNGMT	cis	Public funding from central government	Number	0	0	bool
funeu	FUNEU	cis	Public funding from the EU	Number	0	0	bool
funrtd	FUNRTD	cis	Funding from EU's 6th or 7th Framework Programme for RTD	Number	0	0	bool
co	CO	cis	Cooperation arrangements on innovation activities	Number	0	0	bool
orgbup	ORGBUP	cis	New business practices for organising work or procedures	Number	0	0	bool
orgwkp	ORGWKP	cis	New methods of workplace organisation	Number	0	0	bool
orgexr	ORGEXR	cis	New methods of organising external relations	Number	0	0	bool
mktdgp	MKTDGP	cis	Significant changes to the aesthetic design or packaging	Number	0	0	bool
mktpdp	MKTPDP	cis	New media or techniques for product promotion	Number	0	0	bool
mktpdl	MKTPDL	cis	New methods for product placement or sales channels	Number	0	0	bool
mktpri	MKTPRI	cis	New methods of pricing goods or services	Number	0	0	bool
imputecis	CIS_TYPE	cis	Code to show if data is observed or imputed in SBS	Number	0	0	
roekx	ROEKX	cis	Acquisition of existing knowledge from other enterprises or organisations	Number	0	0	bool
rotrx	ROTRX	cis	All other innovation activities including design, training, marketing, and other relevant activities	Number	0	0	bool

Continued on next page

Table A.1: Description of Variables

MPname	NSIname	DataSource	Description	Format	uniqueDim	indexDim	Type
propat	PROPAT	cis	During the three years xxx to xxx, did your enterprise apply for a patent	Number	0	0	bool
protm	PROTM	cis	During the three years xxx to xxx, did your enterprise Register a trademark		0	0	bool
firmid	ENT_ID	ictec	Unique enterprise identification	Character	1	0	
BROAD	BROAD	ictec	Firm has broadband	Character	0	0	bool
AEBUY	AEBUY	ictec	Firm orders through computer networks (websites or EDI)	Character	0	0	bool
AEBVALPCT	AEBVALPCT	ictec	% of orders through internet	Number	0	0	
AESELL	AESELL	ictec	Firm sells through computer networks (websites or EDI)	Character	0	0	bool
AESVALPCT	AESVALPCT	ictec	% of sales through computer networks (websites or EDI)	Number	0	0	pct
IACC	IACC	ictec	Firm has internet	Character	0	0	bool
INTRA	INTRA	ictec	Firm has intranet	Character	0	0	bool
WEB	WEB	ictec	Firm has website	Character	0	0	bool
MOB	MOB	ictec	Firm has mobile access to internet	Character	0	0	bool
ITERP	ITERP	ictec	Enterprise Resource Planning	Character	0	0	bool
ADE	ADE	ictec	Automated Data Exchange	Character	0	0	bool
ADESU	ADESU	ictec	to suppliers	Character	0	0	bool
INVREC	INVREC	ictec	receiving e-invoices	Character	0	0	bool
ADECU	ADECU	ictec	receiving orders	Character	0	0	bool
INVSND	INVSND	ictec	sending e-invoices	Character	0	0	bool
ADEINFO	ADEINFO	ictec	sending product information	Character	0	0	bool
ADETDOC	ADETDOC	ictec	sending transport documents	Character	0	0	bool
ADEPAY	ADEPAY	ictec	Use of ADE for sending payment instructions to financial institutions	Character	0	0	bool
ADEGOV	ADEGOV	ictec	Use of ADE for sending or receiving data to/from public authorities	Character	0	0	bool

Continued on next page

Table A.1: Description of Variables

MPname	NSIname	DataSource	Description	Format	uniqueDim	indexDim	Type
SISU	SISU	ictec	Sharing SCM data with suppliers	Character	0	0	bool
SICU	SICU	ictec	Sharing SCM data with customers	Character	0	0	bool
CRMSTR	CRMSTR	ictec	share of information with other business functions	Character	0	0	bool
CRMAN	CRMAN	ictec	analyse information for marketing purposes	Character	0	0	bool
SISAINV	SISAINV	ictec	sales: management of inventory levels	Character	0	0	bool
SISAACC	SISAACC	ictec	sales: accounting	Character	0	0	bool
SISAPROD	SISAPROD	ictec	sales: production or services management	Character	0	0	bool
SISADIST	SISADIST	ictec	sales: distribution management	Character	0	0	bool
SIPUINV	SIPUINV	ictec	purchases: management of inventory levels	Character	0	0	bool
SIPUACC	SIPUACC	ictec	purchases: accounting	Character	0	0	bool
imputeictec	ICT_Type	ictec	Observed or not observed data (unit)	Number	0	0	
EMPCUSEPCT	EMPCUSEPCT	ictec	% of workers using computers	Number	0	0	bool
CUSE	CUSE	ictec	Persons employed using computers	Number	0	0	bool
ITSP	ITSP	ictec	firm employs IT specialists	Number	0	0	bool
ITSPT	ITSPT	ictec	firm provides IT training	Number	0	0	bool
XFSP	XFSP	ictec	firm outsource IT functions		0	0	bool
IUSE	IUSE	ictec	Persons employed using computers with access to World Wide Web	Number	0	0	
IT_MEXT	IT_MEXT	ictec	ICT functions are mainly performed by external suppliers	Number	0	0	bool
ISPDF_GE30	ISPDF_GE30	ictec	The maximum contracted download speed of the fastest internet connection is at least 30 Mb/s	Number	0	0	bool

Continued on next page

Table A.1: Description of Variables

MPname	NSIname	DataSource	Description	Format	uniqueDim	indexDim	Type
EMPMD1	EMPMD1	ictec	Persons employed, which were provided a portable device that allows a mobile connection to the internet for business use	Number	0	0	bool
SM1_ANY	SM1_ANY	ictec	Use any social media	Number	0	0	bool
BD	BD	ictec	firm analyses big data	Number	0	0	bool
BDOWN	BDOWN	ictec	Big data analysis for the enterprise is done by the enterprise's own employees	Number	0	0	bool
BDEXT	BDEXT	ictec	Big data analysis for the enterprise is done by an external service provider	Number	0	0	bool
CC	CC	ictec	Buy cloud computing services used over the internet	Number	0	0	bool
CC_DS	CC_DS	ictec	Buy CC services delivered from servers of service providers exclusively reserved for the enterprise	Number	0	0	bool
CC_HI	CC_HI	ictec	Buy high CC services (accounting software applications, CRM software, computing power)	Number	0	0	bool
CC_ME	CC_ME	ictec	Buy only medium CC services (e-mail, office software, storage of files, hosting of the enterprise's database)	Number	0	0	bool
RA	RA	ictec	Provide to the persons employed remote access to the enterprise's e-mail system, documents or applications	Number	0	0	bool
firmid	ENT_ID	ofats	Unique enterprise identification	Character	1	0	
entgrp	ENTgrp_ID	ofats	Unique enterprise group ID	Character	0	0	
nument	ENT	ofats	Number of foreign affiliates	Number	0	0	
empofats	EMP	ofats	Number of persons employed in foreign affiliates	Number	0	0	

Continued on next page

Table A.1: Description of Variables

MPname	NSIname	DataSource	Description	Format	uniqueDim	indexDim	Type
nqofats	TUR	ofats	Turnover of foreign affiliates	Number	0	0	
ctryofats	CL_AREA_EE	ofats	Host country of affiliates	Character	0	1	
	Country_grp_OFATS	ofats	Aggregation of destination countries to groups	Number	0	1	
imputeofats	OFATS_type	ofats	Code to show if data (unit) is observed or imputed in OFATS	Number	0	0	
firmid	ENT_ID	ifats	Unique enterprise identification	Character	1	0	
ctryifats	UCI_CO	ifats	Country of ownership	Character	0	1	
imputeifats	IFATS_type	ifats	Code to show if data (unit) is observed or imputed in IFATS	Number	0	0	

A.2 Description of country output generated by the Microprod Pilot Project

ccc : 2-digit country code supplied by NSI, version number embedded in code)

Variable names in *Italics* are the unique (combined) keys of the dataset

- **cccMetaDataDB:** Information on the classifications in each data source. For each source, the unique values of the cartesian product of the 'index dimensions' of each dataset are given.

DB
Variables:
Index_dimension_1
Index_dimension_2
...
Index_dimension_n

Where metadata is generated for each DB
DB = {BR, CIS, IFATS, OFATS, ITGS, ITS}
Index dimensions
BR = {nace}
CIS = {ho: country of headquarters}
IFATS = {ctryifats: country of ownership}
OFATS = {ctryofats: host country of affiliates}
ITGS = {ctry, cn08: product codes}
ITS = {ctry, ebops3: service type code}

- **cccCoverage:** Information on linked business register, production survey, E-commerce survey.

  Variables:

YEAR	Year to which data pertain
IND	Industry classification (2-digit NACE Rev2 industry)
SIZECLASS	Size class (1-5, see below for mapping)
N_BR	Number of firms (from Business Register)
N_SBS	Number of firms (from Structural Business Survey)
N_ICTEC	Number of firms (from E-Commerce Survey)
N_CS	Number of firms (from CIS Survey)
N_BRSBS	Number of firms (merged BR, SBS)
N_BRCIS	Number of firms (merged BR, CIS)
N_BRICTEC	Number of firms (merged BR, ICTEC)
N_SBSCIS	Number of firms (merged SBS, CIS)
N_SBSCISICTEC	Number of firms (merged SBS, ICTEC)
N_ICTECCIS	Number of firms (merged ICT, CIS)
Emp_xx	Number of employees (from single or linked datasets, see above for N_)
SRC[1]	Coding for different calls of coverage sas

[1] Dataset varies by industries (IND, YEAR) and industries × sizeclass (IND, SIZECLASS, YEAR). Coding for SRC

1 = tabulation by IND, YEAR

2 = tabulation by IND, YEAR, SIZECLASS

- **cccSBSPANEL:** Information on Panel length from Structural Business Survey.

  Variables:

SPAN	number of years with continuous non-missing values for sales, va and payroll data
COUNT	number of firms with given span

- **cccSBSAttrition:** Information on Panel attrition from Structural Business Survey.

  Variables:

YEAR1	first year firm occurs
YEAR2	year in which number of remaining firms of cohort year 1 are counted
COUNT	Number of firms remaining in year 2 of cohort of year 1

- **cccDEMOGR:** Firm demographics data from Business Register.

  Variables:

YEAR	Year to which data pertain
IND	Industry classification (EU-KLEMS ALT definition, bottom nodes)
SIZECLASS	Size Class
STATUS[1]	Entrant, Exiter, Continuer, or One-year firm
COUNT	Number of firms underlying each cell of unique dimensions

  [1] Coding for STATUS:
  CO if firm in year = t-1, t, t
  EN in t, t+1, not t-1
  EX in t-1, t, not t+1
  OY in t

- **cccDBSTAT:** Summary Statistics for single or merged DB variables

TABLE	Sumvars; Avgvars	Industry	Samples	Subnames
SBSSTAT	nq nv nm emp ni_intan ni_tan ni_eq pay; nlp w irat	IND	SBS	-
""	nq nv nm emp ni_intan ni_tan ni_eq pay; nlp w irat	IND	SBS	SZ_CLS[1]

  [1] Coding for SZ_CLS
  1 - emp $<$ 10
  0 - 10 $\leq$ emp $<$ 20
  1 - 20 $\leq$ emp $<$ 50
  2 - 50 $\leq$ emp $<$ 250
  3 - 250 $\leq$ emp

- **cccDBst:** File with moments of distributional of variables in single or merged DB

  Variables:

YEAR	Year to which data pertain
IND	Industry classification (EU-KLEMS or ALT definitions)
VNAME[1]	Name of variable whose moments are computed
QRT	Quartile of the distribution (1=lowest, 4=highest, 0=overall mean
Mean	Mean of variable VNAME in quartile=QRT of distribution
STD	Standard deviation of variable in quartile
NOBS	Number of firms in year, ind, vname, quartile

  [1] SBSst: {w, irat, nlp, dnq, demp}

- **cccDBcr:** File with moments of joint distribution of two variables in single or merged DB

  Variables:

YEAR	Year to which data pertain
IND	Industry classification (EU-KLEMS or ALT definitions)
QNAME[1]	Name of variable used for quartile distribution
VNAME[1]	Name of variable whose moments are computed
QRT	Quartile of the distribution (1=lowest, 4=highest, 0=overall mean
Mean	Mean of variable VNAME in quartile=QRT of distribution
STD	Standard deviation of variable in quartile
PCC	Pearson correlation coefficient between VNAME and YNAME in quartile
NOBS	Number of firms in year, ind, vname, quartile

[1] VNAME × QNAME: PScr: {emp, dnq demp} × {lagged nlp}