EUKLEMS – Linked Data: Sources and Methods

Mary O'Mahony, Carolina Castaldi, Bart Los, Eric Bartelsman, Yasheng Maimaiti and Fei Peng

University of Birmingham

October, 2008

Introduction

The EUKLEMS technology indicators satellite database contains variables derived from additional sources that can be linked to the EUKLEMS basic data in analytical research. The data can be divided into three broad groups:

1. Patents and R&D
2. Series derived from plant level data(Distributed microdata indicators, DMD)
3. Series derived from company accounts

This note contains a description of the sources and methods used to derive the estimates, plus contact details of authors.

The data vary in their country, time period and industry coverage. The data in group cover large EU countries plus additional OECD countries. The DMD data cover only 10 countries, the US and nine EU countries. The company data cover all 25 EU countries but data are not reported for some smaller countries and some sectors where the number of companies was too small to be considered reliable. The time periods also vary considerably from 1970-99 for patents, 1980-2003 for R&D stocks and 1997 to 2006 for the company data. The time coverage of DMD data vary by country but mostly cover the 1980s and the early to mid 1990s but some extend as far forward as 2004. Finally the extent of industry detail also varies across groups of data but all variables cover manufacturing and most some services. Industry coverage is summarized in Table A.1 below while country coverage is summarized in Table A.2.

Table A.1. Linked Data: Industry Coverage

desc	code	Patents	R&D	DMD	Company
AGRICULTURE, HUNTING, FORESTRY AND FISHING	AtB			X	
MINING AND QUARRYING	C			X	
FOOD , BEVERAGES AND TOBACCO	15t16	X	X	X	
Food and beverages	15				X
Tobacco	16				X
TEXTILES, TEXTILE , LEATHER AND FOOTWEAR	17t19	X	X	X	
Textiles	17				X
Wearing Apparel, Dressing And Dying Of Fur	18				X
Leather, leather and footwear	19				X
WOOD AND OF WOOD AND CORK	20		X	X	X
PULP, PAPER, PAPER , PRINTING AND PUBLISHING	21t22		X	X	
Pulp, paper and paper	21				X
Printing, publishing and reproduction	22				
Publishing	221				X
Printing and reproduction	22x				X
Coke, refined petroleum and nuclear fuel	23	X	X	X	X
Chemicals and chemical products	24	X	X	X	
Pharmaceuticals	244				X
Chemicals excluding pharmaceuticals	24x				X
Rubber and plastics	25	X	X	X	X
OTHER NON-METALLIC MINERAL	26	X	X	X	X
BASIC METALS AND FABRICATED METAL	27t28	X	X	X	
Basic metals	27				X
Fabricated metal	28				X
MACHINERY, NEC	29	X	X	X	
ELECTRICAL AND OPTICAL EQUIPMENT	30t33	X	X	X	
Office, accounting and computing machinery	30				X
Electrical machinery and apparatus, nec	31				
Insulated wire	313				X
Other electrical machinery and apparatus nec	31x				X

Radio, television and communication equipment	32				
Electronic valves and tubes	321				**X**
Telecommunication equipment	322				**X**
Radio and television receivers	323				**X**
Medical, precision and optical instruments	33				
Scientific instruments	331t3				**X**
Other instruments	334t5				**X**
TRANSPORT EQUIPMENT	34t35	**X**	**X**	**X**	
Motor vehicles, trailers and semi-trailers	34				**X**
Other transport equipment	35				
Building and repairing of ships and boats	351				**X**
Aircraft and spacecraft	353				**X**
Railroad equipment and transport equipment nec	35x				**X**
MANUFACTURING NEC; RECYCLING	36t37	**X**	**X**	**X**	
Manufacturing nec	36				**X**
Recycling	37				**X**
ELECTRICITY, GAS AND WATER SUPPLY	E		**X**	**X**	
ELECTRICITY AND GAS	40				
Electricity supply	40x				
Gas supply	402				
WATER SUPPLY	41				
CONSTRUCTION	F		**X**	**X**	
MARKET SERVICES	MSERV		**X**		
WHOLESALE AND RETAIL TRADE	G			**X**	
Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of fuel	50				**X**
Wholesale trade and commission trade, except of motor vehicles and motorcycles	51				**X**
Retail trade, except of motor vehicles and motorcycles; repair of household goods	52				**X**
HOTELS AND RESTAURANTS	H			**X**	**X**
TRANSPORT AND STORAGE AND COMMUNICATION	I				
TRANSPORT AND STORAGE	60t63			**X**	
Inland transport	60				**X**
Water transport	61				**X**
Air transport	62				**X**

Supporting and auxiliary transport activities; activities of travel agencies	63				**X**
POST AND TELECOMMUNICATIONS	64			**X**	**X**
FINANCE, INSURANCE, REAL ESTATE AND BUSINESS SERVICES	JtK				
FINANCIAL INTERMEDIATION	J			**X**	
Real estate activities	70			**X**	
Renting of m&eq and other business activities	71t74			**X**	
Renting of machinery and equipment	71				**X**
Computer and related activities	72				**X**
Research and development	73				**X**
Other business activities	74				
Legal, technical and advertising	741t4				**X**
Other business activities, nec	745t8				**X**
COMMUNITY SOCIAL AND PERSONAL SERVICES	LtQ				**X**
PUBLIC ADMIN AND DEFENCE; COMPULSORY SOCIAL SECURITY	L			**X**	
EDUCATION	M			**X**	
HEALTH AND SOCIAL WORK	N			**X**	
OTHER COMMUNITY, SOCIAL AND PERSONAL SERVICES	O			**X**	

Table A.2: Country Coverage

Countries		Patents	R&D	DMD	Company
AUS	Australia	**X**	**X**		
AUT	Austria	**X**			**X**
BEL	Belgium	**X**	**X**		**X**
CAN	Canada	**X**	**X**		
CZE	Czech Republic	**X**	**X**		**X**
DNK	Denmark	**X**	**X**		**X**
EST	Estonia			**X**	**X**
FIN	Finland	**X**	**X**	**X**	**X**
FRA	France	**X**	**X**	**X**	**X**
GER	Germany	**X**	**X**	**X**	**X**
GRC	Greece	**X**			**X**
HUN	Hungary	**X**		**X**	**X**
IRL	Ireland	**X**	**X**		**X**
ITA	Italy	**X**	**X**	**X**	**X**
JPN	Japan	**X**	**X**		
KOR	Korea	**X**	**X**		
LTA	Latvia			**X**	**X**
LTU	Lithuania				**X**
NLD	Netherlands	**X**	**X**	**X**	**X**
NOR	Norway	**X**	**X**		
POL	Poland	**X**	**X**		**X**
PRT	Portugal	**X**			**X**
ESP	Spain	**X**	**X**		**X**
SWE	Sweden	**X**	**X**		**X**
SVK	Slovak Republic	**X**			**X**
SVN	Slovenia			**X**	**X**
TWN	Taiwan	**X**			
UK	United Kingdom	**X**	**X**	**X**	**X**
USA	United States	**X**	**X**	**X**	

1.a. EU KLEMS Raw Patent Counts: A Note on Construction Issues

Carolina Castaldi & Bart Los.
Contact: B.Los@rug.nl

Introduction

It is an almost uncontested fact that technological progress has been one of the main drivers of productivity growth. In economics, the purposeful search for innovations has become a prominent feature of growth theories. Policymakers are also concerned about innovation performance, as is, for example, reflected in the Barcelona objectives.

Innovation is hard to measure, in particular because the output of innovation processes has a strongly intangible nature. Among the available innovation performance indicators, a rough distinction can be made between input indicators and output indicators. The most well-known input indicators are derived from statistics on R&D expenditures. Firms, industries, regions and countries that devote higher shares of their resources on R&D projects are likely to see innovation as a (relatively) more important issue than comparable entities with lower R&D efforts. Nevertheless, their efforts might be less productive, as a consequence of which the innovation output could well be overstated. Data on R&D efforts is included in the technology indicators satellite database to the core EU KLEMS database, see below.

Concerning output indicators, patent statistics are often used, in particular in manufacturing. In principle, patents are only granted if the process or product concerned is a true improvement over the existing state-of-the-art (note that many inventions are not patented, often because other mechanisms to protect intellectual property rights are preferred). Hence, every patent granted can be seen as proof of an invention. However, inventions reflected in patents imply innovations having impacts that vary considerably in their technological and economic importance. One of the few options to recover information on really important innovations from patent data rely on citations to previous patents. Heavily cited patents have probably been more important than patents that were

barely cited. Data on citations will be added in the technology indicators satellite database to the core EU KLEMS database at a later stage. This note briefly describes the procedures followed to arrive at raw patent counts (by industry by country), i.e. numbers of patents granted without citation-based information on their importance.

Relating data in the core EU KLEMS database to data in its technological indicators satellite database will allow for studying several interesting question, such as:

- "Do national industries that generate relatively many patents enjoy higher productivity growth than competing industries abroad?"

- "Are national industries that spend relatively much on R&D in general more profitable than competing industries abroad?"

- "Do numbers of patents per euro of R&D spend increase faster for industries for which average skill levels of employees rise faster than in competing industries abroad?"

The raw patent counts by industry, by country and by year provide the first part of this satellite database that will be completed over the next months and is expected to be ready by June 2008.

Methodology

Data source

Two databases were used to arrive at the figures included in the database. We used the NBER Patent Citations Data File (Hall *et al*., 2001), updated by Bronwyn Hall (http://elsa.berkeley.edu/~bhhall/bhdata.html). This database contains data on patent citations to utility patents granted by the U.S. Patent Office in the period 1963-2002. For the present analysis, we used the large subset of these patents applied for in 1970 and

later. To deal with some issues regarding lags between applications for patents and the actual grants (see below), we could not include data for patents applied for after 1999. The remaining subset contains over 2.5 millions of patents, of which more than 1.1 millions related to processes and products invented outside the U.S. These patents include patents granted to individuals and governments, but more than 75% were awarded to non-governmental organizations (corporations and universities).

Assignment to industries

The assignment of patents to industries consisted of two steps.

1) First, we used OTAF classification codes contained in U.S. patents. These codes are contained in Hall's update of the NBER Patent-Citations Datafile. The OTAF classification assigns patents to one or more industries that are most likely to use the patented process or to manufacture the patented invention. To this end, a concordance was set up that maps 124,000 USPC technology classes onto 41 industry fields, plus one "other industries" category. Thus, at the most detailed level, 42 OTAF- industries are discerned.[1]

An issue we had to deal with is that approximately 30% of the patents examined by USPTO were assigned to multiple OTAF codes. Actually, some patents got as many as seven codes. In studies like these, two approaches can be adopted. If the **"whole counting"** approach is chosen, consideration of a patent assigned to multiple OTAF codes will increase the patent count for all industry codes concerned by one. This approach resembles the partially nonrival nature of knowledge: the usefulness of a patent for a given industry is not necessarily reduced if other industries could also benefit from it. A drawback is that if someone would like to aggregate patent counts over industries, he would end up with more patents than have been granted. This disadvantage is avoided by the second approach, **"fractional counting"**. This approach amounts to adding $1/z$ to

[1] See Hirabayashi (2003) for an overview of issues related to the principles underlying the PATSIC database used by Hall. Griliches (1990, p. 1667) was quite critical about early versions of the OTAF classification, but improvements have been sizeable.

patent counts of z OTAF codes assigned to the patent. In other words, the patent is supposed to be "shared". The database contains results obtained using both approaches.

2) A second stage in the assignment of patents to industries was required to arrive at data that are compatible to data in the core EU KLEMS database. A concordance was set up between OTAF industries and EUKLEMS industries. This involved aggregations that are reported in Table 1. After these procedures were carried out, we ended up with data on the numbers of patents in 20 EU KLEMS industries. The classification is more fine-grained for high-tech industries than for industries known as low-tech.

Table 1.1: Industry classification

Nr.	Description	OTAF codes	EUKLEMS codes
1.	FOOD , BEVERAGES AND TOBACCO	20	15t16
2.	TEXTILES, TEXTILE , LEATHER AND FOOTWEAR	22	17t19
3.	Chemicals and chemical products	28	24
4.	Petroleum and natural gas extraction and refining	1329	11, 23
5.	Rubber and plastics	30	25
6.	OTHER NON-METALLIC MINERAL PRODUCTS	32	26
7.	Basic metals	331+,333+	27
8.	Fabricated metal products	34-	28
9.	MACHINERY, NEC	351,352,353,354,355,356,358,359,348+	29
10.	Office, accounting and computing machinery	357	30
11.	Insulated wire	364	313
12.	Other electrical machinery and apparatus nec	361+,362,363,369	31x
13.	Radio and television receivers	365	323
14.	Electronic components and accessories, and communications equipment	366+	321t322
15.	Motor vehicles, trailers and semi-trailers	371	34
16.	Building and repairing of ships and boats	373	351
17.	Aircraft and spacecraft	372.376	353
18.	Railroad equipment and transport equipment nec	374,375,379-	35x
19.	Medical, precision and optical instruments	38-	33
20.	MANUFACTURING NEC; RECYCLING	99	36t37

Assignment to countries

The assignment of patents to countries was done by means of the records for the variable "country of first inventor" in the NBER Patent-Citations Datafile. The database does not include information about patents or licences transferred or sold by inventors to others. For example, a Japanese multinational might apply for a patent related to an invention

produced in one of its Korean R&D laboratories. In such a case, the patent will be assigned to Korea, although the benefits of the innovation might well accrue to the Japanese firm.

Assignment to years

The patents were assigned to years according to the date of application contained in the patent document. The alternative would have been to classify patents according to their grant dates. The two dates are often a few years apart from each (Hall et al., 2001, report an average lag of two years). Since many users will be most interested in the timing of inventions/innovations (which can have had impacts on productivity and profitability) rather than in the date at which administrative procedures were completed, we opted for linking patents to the year in which the patent was applied for.

This choice has one important drawback. Since the database contains information about patents granted up to 2002, the lag between application date and grant date causes patents with application dates towards the end of the period to be underrepresented. We follow the advice by Hall et al. (2001) not to use data within three years of the final year of the database. Hence, patents applied for in 2000 or later are not included.

1.b. EU KLEMS R&D Data: Sources and Methods

Bart Los
Contact: B.Los@rug.nl

Introduction

Innovation indicators can be classified according to their focus on either inputs or outputs. Patent indicators such as those covered in the previous subsection are output indicators. With regard to input indicators, Research & Development (R&D) statistics are still the most widely employed. R&D indicators like those described here allow for analyses investigating to what extent the differences in productivity growth rates across countries as evident from the core EU KLEMS data can be explained by differences in

investments in innovative activity. Like the productivity data, our R&D data have been compiled at the industry level. This allows for industry-level analyses, which are indispensable given the substantial interindustry differences in technological dynamism characterizing developed countries. Policy-oriented questions that can be addressed using the R&D stocks presented in this EU KLEMS linked data (when linked to the core EU KLEMS data) are:

- How high have the rates of return been to R&D expenditures, and to what extent have these changed over time, across countries and across industries?

- To what extent did industries benefit from R&D expenditures by other industries in the same country? Some industries might play an important role as generators of so-called knowledge spillovers to other industries, and hence play a pivotal role in the medium to long run performance of national economies

- Have industries in a country benefitted from R&D investments in other countries, or did they lose from this? Are such patterns for countries/industries close to the technological frontier different from those for countries/industries with a lot of opportunities for catching up? R&D by foreign countries/industries can lead to productivity-enhancing spillovers, but also to productivity-diminishing "business stealing" effects.

Methodology

Industry classification

The data on annual R&D expenditures were taken from the "OECD Research and Development in Industry Database" (ANBERD). The private sector has spent the amounts included in this dataset. It covers two periods: for 1973-1997 the data have been organized according to ISIC2, and for 1987-2004 the expenditures have been classified according to ISIC3. The most important difference is in the treatment of the furniture

industry. In ISIC2, it is part of the "Wood products and furniture" industry, while it is classified under "Other Manufacturing" in ISIC3. To solve this discrepancy between the two sets of data, we have taken the most recent data as our point of departure. We used the annual proportional changes in R&D expenditures in the ISIC2 industry "Wood products and furniture" and applied these to R&D expenditure levels of the ISIC3 industry "Wood and products of wood and cork" for 1987 to obtain estimates for R&D in the ISIC3 industry for 1973-1986. In a similar vein, we used the annual proportional changes in R&D expenditures in ISIC2 industry "Other manufacturing, n.e.c." and applied these to the R&D expenditure levels of the ISIC3 industry "Furniture; Manufacturing, n.e.c." for 1987.

Deflation

The R&D expenditures described above have been expressed in current prices (in national currencies). To arrive at R&D stocks, R&D expenditures in constant prices are needed. To obtain these, GDP deflators were used. Expenditures have been expressed in prices for the year 2000.

Construction of R&D stocks

The common Perpetual Inventory Method was used to construct R&D stocks using the R&D expenditures data. This implies that initial stocks had to be estimated. This was done by computing average annual growth rates of these expenditures (at industry level, per country) over the first seven years for which expenditures were available. If we denote these by g_{ic} (i=industry, c=country), the estimated initial R&D stocks K_{ic}^{0} can be represented by

$$K_{ic}^{0} = \frac{R_{ic}^{0}}{g_{ic} + 0.12},$$

in which the numerator stands for the R&D expenditures in the first year for which these are available. For most countries, this is 1973. Exceptions are Belgium, the Czech Republic, Korea and Poland.

To arrive at annual stocks for year *t*, the stock for year *t*-1 was multiplied by the value (1-0.12) and the R&D expenditures for year *t* were added. The rate of depreciation was set at 0.12, in line with the estimate by Nadiri & Prucha (1996). Finally, the first seven annual stocks were deleted for each country and industry, because these might be very sensitive to the estimates of the initial stocks. The estimated stocks for 2004 were also deleted, since the R&D expenditures for 2004 were very preliminary.

2. EU KLEMS DMD indicators: Sources and Methods

Eric Bartelsman

Contact: ebartelsman@feweb.vu.nl

Distributed micro-data (DMD) analysis

The firm-level projects financed by the OECD, the World Bank, and various grants of EU member countries, have generated micro-aggregated data that form the basis of the datasets delivered to the EU-KLEMS project. The World Bank work involves 14 countries (Estonia, Hungary, Latvia, Romania, Slovenia; Argentina, Brazil, Chile, Colombia, Mexico, Venezuela, Indonesia, South Korea and Taiwan.(China)) The earlier OECD study was based on firm-level data for : Canada, Denmark, Germany, Finland, France, Italy, the Netherlands, Portugal, United Kingdom and United States. Further updates for EU countries have been made under various arrangements with member countries. The work made use of a common analytical framework and was conducted by active experts in each of the countries. The framework involves the harmonization, to the extent possible, of key concepts (e.g. entry, exit, or the definition of the unit of measurement) as well as the definition of common methodologies for studying firm-level data. The methodology for collecting the country/industry/time panel dataset built up from underlying micro-level datasets has been referred to as 'distributed micro-data

analysis' (Bartelsman, 2004). This section starts with a brief description of this methodology and comparisons with other methods of analysis, and continues with a description of the underlying data and the resulting panel dataset

Micro-data is unique in allowing researchers to identify behavioral changes of individual agents owing to policy changes, but it does not help in identifying the effects of general policy changes affecting all agents. Datasets such as the EU-KLEMS database, with information that varies by industry and time as well as across countries, have been used frequently by academic and policy researchers to identify effects of changes in the general policy environment. However, because the economic outcomes, such as employment, output or productivity, are for industries and not for agents, behavioral response cannot be identified.

A dataset consisting of 'stacked' micro-level datasets from multiple countries will contain the necessary information lacking from either single-country micro datasets or multiple-country sectoral datasets. Unfortunately, owing to the legal necessity of maintaining confidentiality of firms' responses in many countries, micro datasets from individual countries cannot be stacked for analysis. Creating 'public use' data from the underlying sources is a possible workaround for collecting the necessary information. For firm-level data, a public-use dataset made through randomization or micro-aggregation often is not feasible without the loss necessary information.

Another possible workaround is to create a dataset consisting of results from single-country studies that become the input for 'meta-analysis.' For example, a collection of results from single-country studies on the link between ICT and growth at the firm-level, were presented in a recent volume of the OECD (2003b). However, the combination of results of analyses from single-country studies will not provide a solution if the effect of interest is not identifiable within a single country. Further, meta-analysis becomes difficult when the details of the underlying micro-datasets are not well documented. The construction of longitudinal firm-level data is often complex, and requires researchers to have specialized knowledge and experience of the data sources. For example, tracking firms through business registers requires an in-depth understanding of how registers are designed and changes that occur to them over time. Firm-level data are also subject to various protocols (often embodied in legal requirements) relating to

the protection of information. The data are typically only accessible to designated individuals and output prepared for wider circulation usually has to be vetted before being released

In the firm-level projects, a hybrid approach was followed that mitigates many of the discussed problems. The project does not use 'stacked' data. Instead, identical analyses were conducted separately by experienced researchers using micro-level datasets residing in participating countries. The analysis was designed and programmed after face-to-face meetings with country experts and collection of meta-data describing each country's datasets. The analysis was run in each country separately and produced output that could be collected centrally. The combined output provided the information necessary for cross-country analysis. This approach was first developed for the OECD firm-level growth project and is known as 'distributed micro-data analysis.' This method requires tighter co-ordination and less flexibility in research design in each country than for 'meta-analysis,' where the methodology and output may vary across samples. Further, the approach is fairly costly in terms of the coordination effort involved in keeping the analyses in the separate countries on track. The method of distributing work to participating countries and analyzing the comparable output centrally arguably provides better results than meta-analysis of single country micro-level studies or multi-country studies with aggregated data.

The method of distributed micro-data analysis maintains the advantages of multi-country studies with aggregated data, because the output provided by each country consists of indicators aggregated to a pre-specified level of detail that passes disclosure in all countries. The method also maintains information on behavior of agents residing in micro data because the computed indicators on the (joint) distribution of variable(s) are designed to capture hypothesized behavior. While not allowing the full flexibility of research design available with multi-country stacked micro data, distributed micro-data analysis provides a skilled researcher the ability to use cross-country variation to identify behavioral relationships.

Description of the data

The distributed micro-data analysis was conducted for two separate themes. The first set of analysis gathered data relating to firm demographics, such as entry and exit, jobs flows, size distribution and firm survival. The second theme gathered indicators of productivity distributions and correlated of productivity. In particular, information was collected on the distribution of labour and/or total factor productivity by industry and year, on the decomposition of productivity growth into within-firm and reallocation components. Further, information is provided on the means of firm-level variables by productivity quartile, industry, and year. The key features of the micro-data underlying the analysis are as follows:

Unit of observation: Data used in the study refer to the firm as the unit of reference, with a few exceptions (see below for country details). More specifically, most of the OECD data used conform to the following definition (Eurostat, 1995) "an organizational unit producing goods or services which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources". Generally, this will be above the establishment level. However, firms that have operating units in multiple countries in the EU will have at least one unit counted in each country. Of course, it may well be that the national boundaries that generate a statistical split-up of a firm, in fact split a firm in a 'real' sense as well. Also related to the unit of analysis is the issue of mergers and acquisitions. Only in some countries does the business register keep close track of such organizational changes within and between firms. In addition, ownership structures themselves may vary across countries because of tax considerations or other factors that influence how business activities are organized within the structure of defined legal entities.

Size threshold: While some registers include even single-person businesses (firms without employees), others omit firms smaller than a certain size, usually in terms of the number of employees (businesses without employees), but sometimes in terms of other measures such as sales (as is the case in the data for France). Data used in this study exclude single-person businesses, although the data were tabulated for all firms in countries where available. However, because smaller firms tend to have more volatile

firm dynamics, remaining differences in the threshold across different country datasets should be taken into account in the international comparison.[2]

Period of analysis: Firm-level data are on an annual basis, with varying time spans covered.

Sectoral coverage: Special efforts have been made to organize the data along a common industry classification (ISIC Rev.3) that matches the STAN database. This corresponds very closely to the industry definition of the EUKLEMS, with a notable exception being the split of high-tech machinery and equipment in EUKLEMS. In the panel datasets constructed to generate the tabulations, firms were allocated to the detailed sector that most closely fit their operations over the complete time-span. In countries where the data collection by the statistical agency varied across major sector (e.g., construction, industry, services), a firm that switched between major sectors could not be tracked as a continuing firm but ended up creating an exit in one sector and an entry in another. For industrial and transition economies, the data cover the entire non-agricutural business sector (see below for details).

Unresolved data problems: An unresolved problem relates to the 'artificiality' of national boundaries to a business unit. As an example, say that the optimal size of a local activity unit is reached when it serves an area with 10 million inhabitants. In smaller nations, one activity unit must be supported by the administrative activities of a business unit. If the EU boundaries were to disappear, the business unit could potentially serve 25 activity units. This geographic logic may contribute to the larger average firm size observed in the large countries in the sample (U.S. and Brazil). From a policy perspective, this difference may point towards aligning regulations in a manner that would allow business units to enjoy transnational scale economies in meeting administrative requirements. Also related to the unit of analysis is the issue of mergers and acquisitions. No attempt has been made to follow these in a systematic and comparable manner. In some countries, the business registers have been keeping track of

---

[2] The productivity data are collected at different levels of aggregation in different countries and very few are able to work at more than one level. A sensitivity analysis of the productivity decompositions suggests, however, that this issue does not significantly affect the results.

such organizational changes within and between firms in the most recent years, but this information is not used in the present study

- Reflections on robustness of data, how to make cross-country comparisons with different types of data.

The source of the data: Firm demographics

The analysis of firm demographics is based on business registers (Canada, Denmark, Finland, Netherlands, United Kingdom and United States, Estonia, Latvia, Romania, Slovenia), social security databases (Germany, Italy, Mexico) or corporate tax roles (Hungary). . Data for Portugal are drawn from an employment-based register containing information on both establishments and firms. All these databases allow firms to be tracked through time because addition or removal of firms from the registers (at least in principle) reflects the actual entry and exit of firms.[3]

3 . In most countries, these data are confidential and cannot leave the confines of the statistical agency. See above under the heading "Research Protocol".

Table 2.1 Data sources used for firm demographics

Country	Source	Period	Sectors	Availability of survival data	Threshold
Canada	Business register	84-98	All Economy	No	1+
Denmark	Business register	81-94	All	No	Emp ≥ 1
Finland	Business register	88-98	All	Yes	Emp ≥ 1
France	Fiscal database	89-95	All	Yes	Turnover: Man: Euro 0.58m Serv: Euro 0.17m
Germany (West)	Social security	77-97	All but civil service, self employed	Yes	Emp ≥ 1
Italy	Social security	86-94	All	Yes	Emp ≥ 1
Netherlands	Business register	87-97	All	Yes	None
Portugal	Employment-based register	83-98	All but public administration	Yes	Emp ≥ 1
Sweden	Business Register	90-04	All	Yes	Emp≥ 1
UK	Business register	80-97	Manufacturing	Yes	Emp ≥ 1
USA	Business register	88-97	Private businesses	Yes	Emp ≥ 1
Estonia	Business Register	95-04	All	Yes	Emp ≥ 1
Hungary	Fiscal register (APEH)	92-01	All	Yes	Emp ≥ 1
Latvia	Business register	96-06	All	Yes	Emp ≥ 1
Romania	Business register	92-01	All	Yes	Emp ≥ 1
Slovenia	Business register	92-04	All	Yes	Emp ≥ 1

The source of the data: Productivity decompositions

The other major component of the firm-level projects concerns productivity and its components. The data sources used for the analysis of productivity differ from those

used for firm demographics in many countries. For productivity measures, data are needed on output, employment and possibly other productive inputs such as intermediate materials and capital services. Using these source data, indicators are calculated on labour and/or total factor productivity by industry and year, and on the decomposition of productivity growth into within-firm and reallocation components. The underlying source data and availability of the indicators are provided in table 2.2. A full list of indicators is displayed in Table 2.4.

Table 2.2 Summary of the data used for productivity decompositions

Country	Source	Periods		Coverage		Productivity			Unit	Threshold
		First	Last	Mfg	Serv	LPV, LPQ	TFP	MFP		
Finland	Census	95-00	97-02	✓		✓	✓	✓	Estab	Emp>5
France	Fiscal database	85-90	90-95	✓		✓	✓	✓	Firm	Turnover €0.58m
Italy	Survey	82-87	93-98	✓	✓	✓	✓		Firm	Turnover €5m
Netherlands	Survey	83-88	99-04	✓	Some	✓	✓	✓	Firm	Emp>20, emp<20 →S
Portugal	Employment-based register	86-91	93-98	✓	✓	✓	×		Firm	Emp>1
UK	Survey	80-85	96-01	✓		✓	✓	✓	Estab	Emp > 100, emp<100 →Sample
USA	Census	87-92	92-97	✓		✓	✓	✓	Estab	Emp>1
Estonia	Business Register	95-00	99-04	✓	✓	✓	✓	✓	Estab	Emp ≥ 1
Latvia	Business register	96-01	00-05	✓	✓	✓			Estab	Emp ≥ 1
Sweden	Business register	90-95	99-04	✓		✓	✓	✓	Firm	Emp ≥ 1
Slovenia	Business register	92-97	99-04	✓		✓	✓	✓	Estab	Emp ≥ 1

Source: OECD Firm Level Study participants, OECD 2001b, see Appendix for more details.
Note: Key to Method column: C: Census, R: Register, S: Sample, (PW): Population weighted. Most countries impose some restriction on the size of firm/plant that is included in their data. LP is labor productivity, TFP includes labor and capital inputs, and MFP includes labor, capital and materials.

Indicators collected

Depending on the availability of output and input measures, productivity could be calculated in a variety of ways. The methods used are shown in the table below:

Table 2.3 Methods used for the calculations of the different productivity measures

Productivity Measure	Gross output (or sales)	Value Added	Labor	Capital	Intermediate inputs
LPQ	✓		✓		✓
LPV		✓	✓		
or					
TFP		✓	✓	✓	
MFP	✓		✓	✓	✓

Definitions of variables:

- **Labor input**: generally number of employees.
- **Sales, gross output**: No correction made for inventory accumulation.
- **Capital stock**: in countries where available, book value.
- **TFP**, at the firm level is the log of deflated output minus the weighted log of labor plus capital, where the weights are industry specific and the same for all countries. The weights are calculated using the expenditure shares of inputs for an industry using the cross-country average from the OECD STAN database. In the World Bank project, TFP also is computed by using average expenditures shares of firms in an industry observed in each country's dataset.
- **MFP** calculations use expenditure shares for labor, capital and materials.
- **LPQ** and **MFP** are based on deflated gross output, while LPV and TFP are based on deflated value added.

Using common factor shares across countries for a particular industry in principle allows cross-country comparison of productivity levels. However, differing units of measurement for the inputs, notably capital, make the cross-country comparison of TFP or MFP levels meaningless. To 'benchmark' the levels of TFP and MFP, the measured units of capital are adjusted with a multiplicative factor, such that value added minus payroll (or gross output minus payroll and materials expenditures) represents a return to capital of eight percent. This adjustment is similar to the arbitrary adjustments to TFP made by Bernard and Jones (1996) in order to compare 'apples and oranges'.

Firm level nominal values of output, value added and materials are deflated at the industry level using deflators supplied by the team members.

Based on the files collected by the 'demography' theme, the 'productivity decomposition' theme and the 'industry distributions' theme from the distributed micro data projects, a dataset has been generated that will accompany the EUKLEMS database, the so-called DMD dataset.

For information concerning the data collected from the underlying firm level micro datasets, see the documentation for demography and decomp, in the appendix.

In the EUKLEMS DMD release, the data are provided at the '30 industry' level. The data are available for 10 countries, and for selected years between 1990 and 2004.
The following variables are distinguished:

Table 2.4: Indicators in the DMD database

Variable	Source	Suffix variation	Description
Num_*sz*	demog	sz in (1,2,3,4,5)	Number of firms
Empshr_ *sz*	demog	Sz = 1: <20 2: 20-50 3: 50-100 4:100-500 5:>500	Share of employment in size class
Num_*stat*	demog	Stat in ('CO' 'OY' 'EN' 'EX')	Number of firms: COntinuer, One-Year, ENtrant, EXiter
Empshr_ *stat*	demog	Stat in ('CO' 'OY' 'EN' 'EX')	Share of Employment in: COntinuer, One-Year, ENtrant, EXiter
JDR_CO	demog		Job destruction rate among continuers (denominator: avg empl in t and t-1)
JCR_CO	demog		Job creation rate among continuers (denominator: avg empl in t and t-1)
OPC_*typ*	DC	typ in (lpv; lpq; tfp; mfp): productivity measure: VA/E; Q/E; VA/F(K,E); Q/F(K,E,M)	Olley-Pakes Cross-term: gap between weighted and unweighted productivity in log-points.
DC*x*_*typ*	DC	x in (w b c x n), within, between, cross, exit, and entry decomposition terms (see Foster, Haltiwanger, Krizan)	Productivity Decompositions: contribution to growth in 5 years (not annualized)
P4_*typ*	ST	typ in (lpv; lpq; tfp; mfp):	Average of log-productivity of top quartile of firms minus average log-productivity of bottom quartile
CV_*typ*	ST	typ in (lpv; lpq; tfp; mfp):	Coefficient of variation of distribution of log-productivity
STD_DE			Standard deviation of across-firm distribution of 5-year employment growth
STD_DQ			Standard deviation of across-firm distribution of 5-year output growth

3. EU KLEMS Company data: Sources and Methods

Yasheng Maimaiti, Mary O'Mahony and Fei Peng
Contact: m.omahony@bham.ac.uk

Introduction

The Amadeus database contains information on about 120,000 companies in the EU-25, and so the primary advantage of using this resource is its extensive country coverage. In the analysis it was decided to exclude sectors dominated in some countries/time periods by either state run or regulated monopolies (mining, utilities, post and telecommunications), sectors where the average firm size is too small for a reliable assessment from Amadeus (Agriculture, Construction and personal services), and non-market services and financial companies whose reporting differs from other companies. The number of firms varies considerably by country with the highest coverage in the UK followed by Germany, France, Italy and Spain. In most EU countries about 60% of reporting companies are located in service sectors, the exceptions were Germany and Italy among the group of EU-15 countries with about 50% in manufacturing, and some East European New Member States (Poland, Czech Republic, Romania and Slovenia) also with more than 50% in manufacturing. Denmark had the least coverage of manufacturing with only 20% of firms.

Pros and cons of company accounts

There are a number of well known problems in using company accounts data in productivity analysis. The first is that they have often been criticised as being too skewed in company coverage to produce meaningful industry level results. Thus these accounts omit small unincorporated firms. It is well known that small firms account for only a very small fraction of output. Since small firms are most often found in service sectors, there is an argument that company accounts databases tend to over report manufacturing firms.

Table 3.1 shows the share of turnover in manufacturing and services for selected countries with relatively large samples and contrasts this with the corresponding shares of gross output from EUKLEMS. This shows that when the sample in restricted to only manufacturing and market services, the differences in shares are not as pronounced as the

above observation would lead us to believe, and, if anything, manufacturing is under-represented. Thus there does not appear to be a coverage bias in favour of manufacturing. However in smaller countries the differences between turnover and gross output shares tends to be larger.

Table 3.1 Manufacturing and services output shares, Amadeus and EUKLEMS

	Amadeus output share manufacturing	Amadeus output share services	EUKLEMS output share manufacturing	EUKLEMS output share services
Austria	30.8	69.2	54.0	46.0
Belgium	34.0	66.0	52.1	47.9
Czech Republic	47.8	52.2	60.2	39.8
Denmark	14.4	85.6	48.1	51.9
Estonia	21.4	78.6	44.2	55.8
Finland	60.8	39.2	65.0	35.0
France	28.4	71.6	49.6	50.4
Germany	47.8	52.2	61.6	38.4
Greece	43.0	57.0	37.6	62.4
Hungary	31.9	68.1	61.3	38.7
Ireland	31.1	68.9	60.0	40.0
Italy	42.3	57.7	52.1	47.9
Latvia	21.1	78.9	37.4	62.6
Lithuania	32.8	67.2	53.2	46.8
Netherlands	41.2	58.8	49.1	50.9
Poland	52.0	48.0	52.6	47.4
Portugal	21.1	78.9	47.7	52.3
Slovakia	45.1	54.9	59.7	40.3
Spain	37.2	62.8	54.7	45.3
Sweden	23.9	76.1	53.4	46.6
UK	37.8	62.2	40.4	59.6

Source: AMADEUS and EU KLEMS, University of Birmingham calculations.

Also in terms of coverage, companies in some countries are more likely to report financial information than in others so the country coverage may vary between company accounts and industry output measures. Table 3.2 compares output shares in company accounts versus EU KLEMS and shows that indeed the UK is more highly represented in company accounts than its share of industry output warrants, while Italy, Spain and the East European new member states seem to be under represented.

Table 3.2: Country output shares

	Amadeus output share	EUKLEMS output share
Austria	1.36	2.13
Belgium	3.54	3.40
Czech Republic	0.92	1.59
Denmark	2.57	1.45
Estonia	0.09	0.12
Finland	1.87	1.45
France	17.93	14.69
Germany	18.69	21.00
Greece	0.72	1.22
Hungary	0.84	1.00
Ireland	2.08	1.71
Italy	7.54	15.41
Latvia	0.08	0.11
Lithuania	0.09	0.18
Netherlands	7.14	4.22
Poland	1.54	2.89
Portugal	1.13	1.38
Slovakia	0.28	0.56
Spain	5.81	7.68
Sweden	4.03	2.93
UK	21.18	14.22
	100.00	100.00

Note: Non-euro zone converted to euro at market exchange rates

Source: AMADEUS and EU KLEMS, University of Birmingham calculations.

One reason why company accounts may be less useful than other sources of firm data is that they include large conglomerate firms that cross country borders and operate in many industries. In terms of country allocation, accounts are shown separately for subsidiaries of multi-national companies so there is unlikely to be a serious bias in terms of the country location of production. Industry allocation is a potentially more serious problem and one that cannot be easily dealt with.

Another issue is that for many companies data were missing for one or more years. Given this, the concerns regarding coverage and the considerable effort involved in extracting data for so many companies, it was decided not to attempt a time series analysis of company productivity but instead we use these data to calculate summary industry

measures that can be used to describe industry market structures and used in analysis of productivity growth. These measures can only be used in analysing how market structure in the incorporated part of each industry might be related to economic performance. That said, the company data cover the majority of industry output and so we should not underestimate their usefulness. Annual data on industry indicators (discussed in the next section) were constructed for the period 1997-2006. Missing values were interpolated or extrapolated to ensure that the series did not have artificial jumps. The exceptions were when the date of incorporation indicated that the firm started after 1997 or there were sufficient missing data at the end of the period to suggest the firms exited the market.

Summary measures from firm level data

This section considers how we might use data from company accounts to describe industries in terms of the dynamic behaviour of firms operating within them. Two measures are included – one summarising the concentration level of the industry and the second looking at the date of incorporation of the firms in the industry. The first measure used is the Herfindahl-Hirschman Index (HHI) index, defined as:

$$H = \sum_i (S_i)^2$$

Where S is the share of firm i in industry sales (turnover). The closer this is to 1, the more concentrated the industry. When using company accounts data it is useful to calculate a normalised index, given by:

$H^* = (H-1/N) / (1- 1/N)$

where N is the number of companies in the industry. If N is relatively large H and H* are approximately equal while with fewer companies there is a greater difference. This measure to some extent adjusts for reporting bias where some industries/countries have a low number of firms reporting financial information.

A second potentially useful measure is the extent to which sales are dominated by new or old firms. The date of incorporation (DoI) of each company is available in the Amadeus data. This can be used to construct a simple measure of the (weighted) average age of firms, calculated as:

$AGE = \Sigma_i (2005-DoI)* S_i$

where, as above, S_i is the share of firm i in industry sales.

Average age can be employed as a crude measure of the extent to which older firms dominate an industry. However it is dependent on the history of industrial production and so tends to be much longer in traditional manufacturing such as textiles (where typical average company age is more than 50 years) than in newer industries such as office machinery (typically less than 20 years in EU countries). In addition average age of firms in many service sectors tend to be lower than in manufacturing, although inland transport frequently had some of the oldest firms. Obviously average age is also relatively low in East European new member states.

In order to render such a measure useful in picking up the dynamics of industry structure, it is therefore useful to adjust for this production history. We therefore also calculated an unweighted average age of firms as:

$AGEU = \Sigma_i (2005\text{-}DoI)/N$

A useful measure then is:

$$AGE^* = AGE/AGEU$$

Sales in industries where AGE* is greater than one are dominated by older firms, whereas a ratio less than one indicates that younger firms account for most sales.

References

Bartelsman, E.J., 2004. The Analysis of Microdata from an International Perspective. *OECD Statistics Directorate*, (STD/CSTAT 12).

Foster, Lucia, Haltiwanger, John C., Krizan, C. J., (2001), "Aggregate Productivity Growth: Lessons from Microeconomic Evidence," Chicago: University of Chicago Press.

Griliches, Z. (1990), "Patent Statistics as Economic Indicators", Journal of Economic Literature, vol. 28, pp. 1661-1707.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools" NBER Working Paper 8498.

Hirabayashi, J. (2003), "Revisiting the USPTO Concordance between the U.S. Patent Classification and the Standard Industrial Classification Systems", Paper presented at the WIPO-OECD Workshop on Statistics in the Patent Field (Geneva, 18-19 September). (http://www.wipo.int/patent/meetings/2003/statistics_workshop/en/presentation/statistics_workshop_hirabayashi.pdf)

Nadiri, M.I. and I.R. Prucha (1996), "Estimation of the Depreciation Rate of Physical and R&D Capital in the US Total Manufacturing Sector", *Economic Inquiry*, vol. 34, pp. 43-56.