# Building the evidence base for productivity policy using business data linking

# Chiara Criscuolo, Jonathan Haskel and Ralf Martin

Centre for Research into Business Activity, Office for National Statistics and University College London, Queen Mary University of London and London School of Economics

Much government policy is aimed at raising productivity. Much of the work that has guided this policy is based on macro data of, for example, industries or countries. This article reports on recent work by the Centre for Research into Business Activity (CeRiBA), using ONS data, to use micro data to inform policy. We describe the construction of an establishmentlevel manufacturing productivity panel dataset, its use and its combination with other datasets.

## Introduction

Raising national productivity is a key aim of the government. But the evidence concerning productivity trends used in successive Budgets and Pre-Budget Reports (PBR) has been largely macroeconomic in nature, most notably in international comparisons of productivity and TFP (see, for example, the 2002 PBR). Some important insights have arisen from these data. For example, O'Mahony and De Boer (2002) argue that much of the productivity gap with the US is due to capital and technology and much of the gap with Germany is due to capital and skills. In an industry study Basu *et al* (2003) show that differences in UK and US productivity growth in retailing explain around two-thirds of the difference between the UK and US productivity accelerations in the late 1990s.

If all firms or plants in an industry or economy had like productivity, there would be no need for a more micro study. However, recent work suggests that productivity differs substantially between businesses. For example, Disney *et al* (2003) show that exiting plants have about 4 per cent lower productivity than survivors. Griffith (1999) shows that foreign-owned plants in the UK car industry have a substantial labour productivity advantage over the set of all UK-owned plants.

These studies above are all based on early attempts to match successive years of *Census of Production* data to form a longitudinal plant-level data set (the Annual Respondents to the Census of Production Database).<sup>1</sup> This data set has been described in two previous *Economic Trends* articles, Oulton (1997) and Barnes and Martin (2002). Recent work at the business data linking project (BDL) has attempted to take this work further by:

- updating the manufacturing part of the ARD to as recent a year as possible;
- including services;
- matching in other data sets to expand knowledge on variables not well measured by the ARD (such as skills and innovation).

This article sets out some of the work that has been carried out in creating these data sets and highlights some of the findings.

# **Creating the ARD**

#### **The Interdepartmental Business Register**

#### Enterprises, enterprise groups and local units

To compile a data set on businesses, the first step is to obtain a list of businesses in the UK. This is the role of the Interdepartmental Business Register (IDBR), where addresses of businesses are compiled using a combination of tax records on VAT

and PAYE, information lodged at Companies House, Dun and Bradstreet data and data from other surveys. The IDBR has been operating since 1994. The IDBR tries to capture the structure of ownership and control of firms and plants or business sites that make up the UK economy using three aggregation categories: local units, enterprises and enterprise groups. Their meaning is best illustrated by means of an example which is also laid out in Figure 1.

Let us suppose that *Brown* is a single firm, operating in a single location, producing goods for a single industry. Suppose now that *Smith and Jones Holdings* are a holding company, registered in London. In turn, they own two firms, *Smith* and *Jones*, who produce in separate plants. *Smith* has four plants, *Smith North*, *Smith South*, *Smith East* and *Smith West. Jones* has a plant, *Jones North* and an R&D lab, *Jones R&D*.

*Brown*, being a firm responsible for a single business activity, is a single plant 'enterprise'. *Smith and Jones Holdings*, being responsible for firms with distinct business activities, is called an "enterprise group".<sup>2</sup> *Smith and Jones* are also enterprises. All plants are called 'local units'. To qualify as a local unit a business entity must only consist of one site at a single mailing address. Consequently if *Jones R&D* is located at a different site than *Jones North* the enterprise *Jones* would consist of two local units. If *Jones R&D* was located at the same site as *Jones North* the two would form one local unit for the IDBR.<sup>3</sup>

# Maintaining information on enterprise groups, enterprises and local units

A major advantage of the IDBR and related datasets such as the ARD is that information is available at many disaggregated levels (whereas company accounts would only be at the enterprise or enterprise group level for example). This is very useful in some cases. For example, job creation by an enterprise that opens a local unit of 100 and closes one of 100 is zero at the enterprise level but 100 at the local unit level. As another example, regional employment data would be unreliable if employment was only recorded at the enterprise level but the enterprise consisted of local units in different regions. It is therefore critical that the IDBR maintains as accurate a record as possible at the different levels.

The Annual Register Inquiry (ARI) maintains this information (Jones, 2000, p.51). It began operation in July 1999 and is sent to large enterprises (over 100 employees) every year, to enterprises with 20-99 employees every four years and to smaller enterprises on an *ad hoc* basis. The ARI currently covers around 68,000 enterprises, consisting of about 400,000 local units. It asks each enterprise for employment, industry activity and the structure of the enterprise. For the Brown enterprise in our example this is straightforward. A multi-site enterprise such as Smith receives a form and is asked to report on its overall activity and employment. It will also be sent four extra forms to report the same for each local unit. If Smith has closed a local unit it must report this on the form. If a local unit has opened, it has to fill out extra forms (which are obtained from ONS by an automated procedure). Returns from the ARI update the IDBR in the summer of each year.

# Maintaining information on employment, turnover and other data

As well as structure of business information, the IDBR holds other data, such as address and SIC code. For productivity we will require independent output, employment and (possibly) other input information. Output information on the IDBR comes from VAT records if the original source of business information was VAT data. Employment information comes from PAYE data if that is the source of the original inclusion. Thus as long as the single-local unit enterprise Brown is large enough to pay VAT ((the threshold was £52,000 in 2000-01) it would have turnover information at the enterprise and local unit level. On the other hand, if Brown does not operate a PAYE scheme, it will have no employment information. However, employment data is required to construct sampling frames and hence it will be interpolated from turnover data. For the multi-local unit enterprise Smith, no turnover information will be available for Smith's local units, since most multi-local unit enterprises do not pay VAT at the local unit level, If the PAYE scheme is operated at the local unit level, it would have independent employment data.

There are two other ways in which employment and output data are gathered. The first is if the business is included in the ARI and the second if it is included in the Annual Business Inquiry (ABI), see below.

# The ABI and the ARD

Whilst the IDBR holds much useful information, more data is required on outputs and other inputs, in order to calculate GDP. Thus the ONS conducts a business survey, based on the IDBR. This is the Annual Business Inquiry (ABI) and the ARD consists of the panel micro-level information obtained from successive cross-sections of the ABI. The ABI covers production, construction and some service sectors, but not public services, defence and agriculture.<sup>4</sup>

# Reporting units, selected and non-selected data

To reduce compliance costs, the ABI is not a Census of all local units. This is in two regards. First, enterprises normally report on all their local units jointly unless the enterprise has local units in both Britain and Northern Ireland. There is a legal requirement for the ONS to keep data for these two areas separately and therefore enterprises are required to report separately as well. Another reason for separate reporting is if a business explicitly requests such a split. So for example, Smith may decide to report on North and South combined and East and West separately. This creates a somewhat different structure of 'reporting units', as opposed to the structure of enterprises and local units, and this reporting unit structure is shown for our example at the bottom of Figure 1. Brown forms one reporting unit (A) only whereas Smith reports on three mutually exclusive parts of its enterprise, B, C and D. These reporting units are consequently the fundamental unit on the ARD.

Second, only reporting units above a certain employment threshold (currently 250<sup>5</sup>) are sent an ABI form every year.

#### Figure 1 Plants and firms in the IDBR

Enterprise group level	Smith and Jones Holding						
Enterprise level	Brown Enterprise	Smith				Jones	
Local unit level	Brown Plant	Smith North	Smith South	Smith East	Smith West	Jones North	Jones R&D
ARD reporting unit	Α	В		С	D	E	

Smaller reporting units are sampled by size-region-industry bands.<sup>6</sup> In the ARD, all data returned from reporting units is held on what is called the 'selected' file. Other data is held on the 'non-selected file'. Since the non-selected RUs are not sent a form, the non-selected data is of course the IDBR data.

Table 1 and Table 2 give an overview of the structure of business units as well as the relation between the selected sample for which ABI returns are available and the total population of businesses as captured in the IDBR for 1998. Consider Table 1 first. The top left cell shows that in 1998 there were 162,477 enterprise groups. Reading diagonally down the top panel, there are 171,271 reporting units and 196,355 local units. Reading down the first column, each enterprise group consists of, on average, 1.43 reporting units and 8.9 local units. Each reporting unit consists of, on average, 4.15 local units. Thus it would seem that the ABI consists of very many multi-unit businesses. The second panel explores this in more detail. The top line shows there are 158,399 enterprise groups with only one RU. The rest of the rows document the disposition of the rest of the sample: LUs in an enterprise group with 2-3 RUs, 7,342 have between 2 and 3 RUs. The lower panel shows a similar picture for local units in enterprise groups and RUs. The vast majority of enterprise groups and RUs consist of just one local unit (149,326 and 158,727 respectively).

The second part of the table shows the same descriptive statistics for selected businesses. The first panel shows that in the 1998 selected file there were 11,088 enterprise groups, 13,264 RUs and 28,765 LUs. Since selected data consist mainly of larger reporting units, the prevalence of multi-unit businesses is greater (2.54 RUs per enterprise for example). The second panel shows the distribution of RUs per enterprise group. Finally, the bottom panel shows how local units are distributed across reporting units and enterprise groups. 9,279 LUs belong to RUs that consist of only this single LU.

Table 2 focuses on RUs and sets out the number of RUs that consist of different numbers of LUs (Table 1 among other things showed the distribution of LUs that belonged to RUs of different size). The top panel of Table 2 refers to the population whereas the lower panel to the selected sample. The top row of the top panel shows that in all our data (again for 1998), the average RU has 15 employees. The final column shows the fraction of total employment accounted for by a particular type of RU, in this case 57 per cent of total employment. Looking at the far right bottom cell of

#### Table 1

# Structure of enterprises reporting units (RUs) and local units (LUs) in the ARD 1998

(Number of RUs by enterprise groups and number of LUs by enterprise group reporting units)

Population of all bu		Numbers	
Ву	Enterprise groups	RUs	LUs
Enterprise group	162,477		
RUs	1.43	171,271	
LUs	8.91	4.15	196,355
1 RU	158,399		
2–3 RUs	7,342		
4–5 RUs	1,729		
6–10 RUs	1,575		
11 plus RUs	2,226		
1 LU	149,326	158,727	
2–3 LUs	19,237	17,632	
4–5 LUs	5,167	4,346	
6–10 LUs	5,214	4,209	
11 plus LUs	17,411	11,441	

#### Selected sample (ABI) 1998

Ву	Enterprises groups	RUs	LUs
Enterprise group	11,088		
RUs	2.54	13,264	
LUs	36.55	18.13	28,765
1 RU	10,311		
2–3 RUs	1,238		
4–5 RUs	461		
6–10 RUs	567		
11 plus RUs	687		
1 LU	7,674	9,279	
2–3 LUs	4,482	5,498	
4–5 LUs	2,139	2,392	
6–10 LUs	2,780	2,795	
11 plus LUs	11,690	8,801	

Notes: RU = ARD reporting unit; LU = IDBR local unit. Source: Author's calculations based on ARD.

### Table 2 Number of RUs by number of LUs

	Number of	Average size	Row share of
	RUs	of RUs	total emp
RUs with		all	
1 LU	158,727	15.14	57%
2–3 LUs	6,091	149.65	22%
4–5 LUs	641	486.73	8%
6–10 LUs	320	889.89	7%
11 plus LUs	206	1,197.31	6%
		selected	
1 LU	9,279	109.49	42%
2–3 LUs	1,993	329.73	29%
4–5 LUs	368	658.46	11%
6–10 LUs	225	997.52	10%
11 plus LUs	147	1,389.86	9%

this top panel, we see that 6 per cent of total employment is in RUs with 11 or more plants. The bottom panel shows the analogous data for the selected sample. For the selected sample we find that 42 per cent of employment is in RUs that have only 1 LU and 29 per cent in RUs with two or three LUs. Thus 71 per cent (42 per cent + 29 per cent) of employment is in RUs with between one and three LUs.

Because the RU level is the most disaggregated level for which extensive data on production inputs and outputs and other data is available, much productivity analysis using the ARD is conducted at this level. One argument in favour for this practice can be made on the basis of Table 2: 77 per cent of RUs (9,279 out of the total RU number, 12,012) representing almost half (42 per cent) of selected employment consist of one LU. Hence for these RUs analysis at the reporting unit level is analysis at the local unit level. Finally, note that many studies have, to simplify discussion, referred to reporting units as 'plants' and enterprise groups as 'firms'.

Besides the ABI, the ONS runs a large number of surveys based on the IDBR. It is important to bear in mind that the boundaries of a reporting unit might vary from survey to survey. For example, the Annual Inquiry into Foreign Direct Investment (AFDI) asks about FDI. FDI activities can typically not be attributed to a particular LU or RU but are decided upon at the level of the holding company. For the purposes of the AFDI survey therefore the RU would be the holding company. Similarly information about R&D activities is gathered at R&D enterprises which are separate from the establishments reporting on production activity of a large enterprise group (compare with the case of *Jones R&D* in our example). This implies that one has to be careful when matching other surveys to the ABI.

### Information quality of the non-selected data

Non-selected data is the IDBR data. Selected data consists of the responses from firms to the ABI. Non-selected data

comes either from the IDBR administrative sources, i.e. the VAT or PAYE, or other data that brought the business onto the register in the first place, or the ARI. Not all of this data is equally reliable especially for smaller business units that are typically not included in the ARI. The quality of this data is important in a number of areas including the construction of sample weights for the selected data and studies conducted at the local unit level. The following points are worth noting.

First, since some of the input data is interpolated from sales data and vice versa, one cannot do productivity studies. Second, there is a specific problem with employment data on the IDBR. According to ONS (2001), when a business first arrives on the register, its employment, if present, is frozen at its first reported point until updated. Turnover is updated however. Thus productivity for these businesses is unreliable unless their employment is updated. Updating is done from the results of the ARI, or before the ARI was introduced, if the firm was in one of the Annual Employment Surveys (AES). We can get some impression of the problem by considering Table 3. The table shows when the employment data of enterprises in the IDBR in year 2000 were last updated. The first 4 columns of Table 3 refer to different size bands. The final column shows that 8.5 per cent of total employment had not been updated since 1993. 1993 is the year when there was last a Census of Employment. Looking at columns 1 and 2 we see that the updating problem is concentrated in the smallest enterprises. 28.7 per cent of employment in enterprises of size 0-9 and 40.2 per cent of employment in enterprises of size 10-19 had not been updated since 1993. Indeed row 11 of Table 3 also reports that 56.9 per cent and 21.8 per cent of enterprises of size 0-9 and 10-19 have never been sent an ARI form or included in the AES. By contrast, larger enterprises are updated more frequently. An additional problem is that the ONS (2001) also states that even larger enterprises in the ARI or AES, may not have fully reported on their local units (see also Partington, 2001).<sup>7</sup> This suggests that the employment and therefore productivity data on these smaller enterprises, who are overwhelmingly in the non-selected sample, is likely to be very unreliable.

# Timing, types of forms and processing

The IDBR is updated using ARI data in the summer. The ABI sample is drawn in the autumn and the forms are sent out at the end of the year. The ABI consists of two sets of forms. The ABI1 form asks for employment information for December and is collected by March. The ABI2 form asks for accounting information and is collected by September to allow firms to use their accounting information (the accounting year ends 5 April).<sup>8</sup>

RUs sent an ABI form might receive a short or a long form. This is again in the interest of reducing compliance costs. A short ABI1 form is for businesses who have already provided employment information on the ARI or for the 4th quarter of the quarterly employment survey. A short ABI2 form is sent to a proportion of businesses. The short form asks for the main aggregates, but does not ask for breakdowns of some of the variables. For example, it asks for data on intermediates, but not on components of intermediates such as electricity etc. Finally, forms also differ slightly between sectors. There are three basic form types for ABI1 and 21 for ABI2.

When data is received from reporting units it is checked for consistency relative to previous responses. If it is not consistent, the contributor is phoned to check the data. Contact with contributors is recorded on a separate database. Non-responders are contacted with two reminders and phone calls. If persistent non-responders have provided data to other inquiries their data is imputed from these sources.

For the RUs sent short forms, the more detailed data asked on the long forms is imputed using the ratios from the longform responses of RUs in similar industry-region-size cells (this imputation process is called expansion). Table 4 sets out the main variables available and their source.

### Before the IDBR

The IDBR was introduced between 1994 and 1995. Before that sampling was on the basis of a business register

#### Table 4 Main ARD variables available and their source

# Table 3

# Percentage distribution of employment by date and enterprise size

	Enterprise size							
Year of update	0–9	10–19	20–99	100 or more	All enter- prises			
1991	0.3	0.6	0.1	0.0	0.1			
1992	0.0	0.1	0.0	0.0	0.0			
1993	28.7	40.2	9.3	0.2	8.5			
1994	0.2	0.4	0.1	0.0	0.1			
1995	1.9	3.8	2.7	0.4	1.1			
1996	1.6	4.7	4.7	1.0	1.8			
1997	3.8	8.4	12.2	5.5	6.2			
1998	3.0	8.2	32.6	12.7	13.2			
1999	3.2	10.8	34.8	47.8	36.6			
2000	0.4	1.1	3.4	32.4	21.7			
Unproven enterprises	56.9	21.8	0.0	0.0	10.7			
Total	100.0	100.0	100.0	100.0	100.0			

Source: ARI, referring to the 2000 IDBR, cited in ONS (2001, Table 10, p.53).

IDBR	Comment	Core questions on ABI1 (by RU)	Comment	Core questions on ABI2 (by RU)	Comment
Business structure	Number of LUs, country of ownership, legal status				
Region	NUTS hierarchy				
Industry					
Employment by local unit	Relates to summer	Employment by reporting unit	For December. Breakdown by male, female, full- and part- time, working proprietors/partners and unpaid workers (e.g. family)		
Turnover by	May be			Turnover	
enterprise	for LUs			Materials costs	Additional breakdown may be available to fuel, electricity etc.
				Inventories	
				Investment	
				Wages cost	Additional breakdown may be available to wages and salaries, pension contributions, social security contributions and redundancy payments.

Legal status means company, partnership, single proprietor business, public corporation, non-profit making body, central or local government. NUTS hierarchy is region (NUTS1), group of counties or unitary authorities (NUTS2), country or unitary authority (NUTS3), district or unitary authority (NUTS4) and ward (NUTS5).

Source: Jones (2000).

maintained by the ONS (then the Central Statistical Office). The maintenance of the register was generally regarded as being less reliable than the ARI and indeed before 1983 no VAT information was available for this purpose (in 1984 around 30,000 LUs appeared on the register when VAT information was first made available). The structure of the inquiry was the same, in that the basic surveyed unit was the RU, large firms were all sampled, smaller ones were sampled proportionately and the returned data was held on the selected file. Concerning employment, before 1994, employment, if not known, was interpolated using turnover data, using a turnover to employment ratio where turnover and employment were independently observed. The ONS did check employment for plants with imputed employment of over 11, but this was done only for around 20 per cent of the non-selected sample and as for the imputed data due to time lags in the provision of tax data and processing of imputations, such information typically refers to data from two years earlier (Perry, 1985).

As well as these data quality issues, in building up a historical database the following issues arise. First, all data before 1970 appears to have been destroyed. Second, the non-selected data for 1970–79 is missing. Third, the unique RU and LU identification numbers have been changed in 1994 following the introduction of the IDBR. An ONS lookup table relates the two numbers for the selected data and the CeRiBA team have built up a lookup table for the other data using a combination of data from Richard Harris and matching observations by postcode and industry. Fourth, the ARD before 1998 includes only manufacturing.

#### Issues in using the data

# Level of aggregation

A number of issues arise in using the data. The first question is the level of aggregation at which to work. In principle, the ARD panel can be configured for local units, reporting units, or enterprise groups. Which is 'correct' depends on what question one is trying to answer. The spatial pattern of employment for example is likely best investigated at local unit level, since reporting units might report on several local units that are located in different regions.

The correct unit for productivity analysis is more difficult. Production functions describe output-input relations for like technologies. Technology might vary across local units, across reporting units or indeed within local or reporting units, so there is no clear answer here.

Most productivity analysis involving the ARD is done at the RU level simply because it is the most disaggregated level at which all necessary data is actually reported. Some studies have tried to conduct productivity analysis at the local unit level (see, for example, Harris, 2002). This requires distributing the RU level information on a pro rata basis to local units based on the employment information in the IDBR. In order to do this one has to assume that factor proportions are the same for the various local units belonging to a RU and that local unit employment data are reliable.

### Weighting

With selected and non-selected data, we can construct sampling weights. If we wish to report sample averages as conveying information about the population then we must weight observations beforehand. A more difficult issue is whether to weight regressions that are run on the selected sample. The answer seems to depend upon what coefficient one is trying to estimate. DuMouchel and Duncan (1983) consider the following. Suppose one is trying to estimate a marginal effect  $\beta$  from the model  $Y = X\beta + u$ , where *Y* is the outcome variable and X the set of forcing variables, where the data has been sampled and weights  $w_i$  are assigned to the *i*th observation. The OLS estimator of  $\beta$  is  $\beta_{OLS} = (X'X)^{-1}X'Y$ . The weighted least squares estimator is given by  $\beta_{WLS} = (X'WX)^{-1}X'WY$  where W is a diagonal matrix whose *i*th diagonal element is  $w_i$ . If  $\beta$  is constant across size strata, then there is no need to weight to estimate it. In our sample, for example, we observe all large RUs and a sample of smaller ones. Suppose indeed we only observed the large ones and not the small ones. As long as  $\beta$  is constant across large and small RUs, then we do not need the smaller RUs to estimate  $\beta$  nor do we need to weight the larger ones: we can estimate  $\beta$  solely on the large RUs.9

The more complicated case is when  $\beta$  varies across size strata so that the model is  $Y=X\beta(j) + \varepsilon$ . A marginal effect of interest might be the weighted 'average marginal' effect, namely  $\beta_{AVG}=\Sigma w_i\beta(j)/\Sigma w_i$  where the summation is over strata. DuMouchel and Duncan (1983) show that  $\beta_{WLS}$  is a biased estimate of  $\beta_{AVG}$  (unless all the regressors are constant), and so there is no reason to prefer weighting. In fact  $\beta_{OLS}$  is also biased, but there is no general result that one is less biased than another (indeed in this case it would be preferable to estimate using different size strata). Note that in this case it would be preferable to estimate different  $\beta$ s for different size classes, a procedure also recommended by Carrington *et al* (2001).

A final problem occurs if the sampling weights are measured with error, in which case weighted least squares can yield biased coefficient estimates. This is a real concern, because employment in the non-selected data from which weights can be approximated seems to be unreliable.

### Data cleaning

There are a host of data cleaning issues in assembling the ARD panel from the raw cross-section data. Some of the more important are as follows. First, as mentioned above, in 1994 all the LU identifiers changed. Second, in 1984 and again in 1997 the enterprise group reference numbers changed.<sup>10</sup> Third, coding numbers of the variables changes from time to time (i.e. question 406 is gross output one year and inventories in another year) and hence one must be careful to use consistent questions.

### **Prices and capital**

To measure total factor productivity (TFP) we require price deflators and capital stocks. Price deflators are derived from

the ONS PPI inquiries at as disaggregated a level as possible. Capital stock is problematic. There is data on the ABI on investment but not on the capital stocks. With an assumption about starting values, capital stocks may be created using perpetual inventory methods (see Martin 2002). In turn there are a number of issues here.

First, there is clearly some doubt about the allocation of initial values. Martin's approach is to allocate on the basis of RU level material shares. To check the reliability of this, one can look at exit rates from different parts of the TFP distribution. Interestingly, the use of materials to generate initial values is quite important for obtaining plausible results. Allocation of initial values on the basis of an RU's average share in aggregate investment (instead of materials) lead to exit rates that were higher for the top firms in the TFP distribution than for bottom firms.

Second as Harris and Drinkwater (2000) point out, capital stocks based on reporting units suffer from the problem that in multi-plant establishments plant closures could lead to an overestimation of the capital stock with the perpetual inventory method. The reason is that the perpetual inventory method assumes a constant depreciation rate which does not account for the discrete drop in an RU's capital stock with plant closure. The problem however is that to work at the LU level, we need initial values and investment at the LU level. As pointed out above, for multi-plant RUs the only data available at the LU level is employment. Thus to allocate investment and capital we have to assume a constant investment labour ratio across the LUs of an RU and there are worries, as above, about the quality of the small LU employment data.<sup>11</sup>

#### **Measures of productivity**

Our first measure of productivity is simply labour productivity.

$$\ln LP_{it} = \ln (Y_{it} / L_{it}) - \ln (Y_{lt} / L_{lt})$$
(1)

where *i* denotes the RU, *t* time and *I* industry. We have normalised labour productivity on the industry median for compatibility with TFP. To ensure that we calculate TFP in a way that is comparable across RUs we follow Caves, Christensen and Diewert (1981) and calculate the TFP of RU *i* relative to the TFP of the median RU in the industry

$$\ln TFP_{it} = \ln Y_{it} - \ln Y_{lt} - \overline{\alpha}_{K} (\ln K_{it} - \ln K_{lt}) - \overline{\alpha}_{L} (\ln L_{it} - \ln L_{lt}) - \overline{\alpha}_{M} (\ln M_{it} - \ln M_{lt})$$
(2)

where *I* denotes industry and the factor shares are calculated as the average of the RU and industry median RU factor shares.

As regards output we have gross output and value added. The relative merits of each measure have been discussed by, for example, OECD and Oulton and O'Mahony (1994, pp. 33–36) and following that literature, also at the plant or RU level we prefer to use gross output. For employment we would ideally like to adjust our input measures for human capital and hours (full and part-timers for example). For the period as a whole

we have total employment. For 1980–95 total employment is broken down into administrative, technical and clerical workers and operatives. For 1996–2000, total employment is not broken down this way, but into males, full-time and part time and females, full-time and part time. Thus to have a measure consistent over time we use total employment. Capital is as defined above.

#### Results

#### **Dispersion levels**

Table 5 presents our dispersion results (for two digit industries for ease of reading and to avoid disclosure problems). All columns refer to 2000 data. Tobacco, fuel and recycling are omitted.<sup>12</sup> Column 1 shows the standard deviation of labour productivity (LP), with LP measured by the exponent of (1) using gross output. The numbers vary between 2.00 (radio and TV) and 0.60 (motors). Column 5 shows the standard deviation of the exponent of lnTFP, with lnTFP measured by (2) using gross output data. If other inputs explain part of the productivity distribution we should expect the TFP distribution to be less dispersed than the LP distribution and indeed it is in all cases.

Columns 2 and 6 repeat this analysis using the 90–10 ratios (the ratio of the RU at the 90th percentile of the log productivity distribution and the RU at the 10th percentile), computed for gross output based on lnLP and ln TFP. The LP differential varies between 13.31 (office machinery) and 3.22 (wood). Once again the TFP differential is less than the LP differential. Thus, on average the 'top' manufacturing RU is around five times more productive in labour productivity terms and 1.5 times in terms of TFP.

To examine the dispersion further, the remaining columns show the 90/50 and 50/10 ratios using gross output TFP and LP. The LP90/50 dispersions are sometimes larger and sometimes smaller than the 50/10, but the TFP90/50 dispersions are mostly somewhat bigger than the TFP50/ 10 measures. This latter finding suggests a left skewed productivity distribution.

#### How do plants move in the productivity distribution?

A way of examining the persistence of plant productivity follows Oulton (1998) in using Galton-Markov regressions. The basic regression is

$$p_{it} = \alpha + \beta p_{it-1} + \varepsilon_{it} \tag{3}$$

which, if  $\beta$ <1, implies convergence of plants to a mean industry productivity level  $\alpha$ . Equation (3) implies that convergence is symmetric because convergence speed is the same above and below the mean. A more general version of (3) is therefore

$$p_{it} = \alpha + \beta_1 p_{it-1} + D\beta_2 p_{it-1} + \varepsilon_{;t} \qquad D = 1 \text{ if } p_{it-1} > \overline{p}_{it-1}$$
(4)

which has the following interpretation. The term  $D\beta_{21}$  allows for a different convergence speed if the establishment has

Table 5		
<b>Productivity</b>	<b>Spread</b>	in 2000

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Gross output/employment						
	sd	p90/10	p50/10	p90/50	sd	p90/10	p50/10	p90/50
Food	0.89	4.96	2.35	2.04	0.17	1.52	1.19	1.27
Textile	0.64	3.64	1.60	2.20	0.14	1.41	1.17	1.20
Apparel	1.11	5.99	2.10	2.78	0.19	1.54	1.23	1.25
Leather	0.76	10.95	4.43	2.18	0.17	1.55	1.34	1.15
Wood	0.51	3.22	1.68	1.89	0.12	1.44	1.16	1.23
Paper	0.75	3.75	1.85	2.02	0.14	1.39	1.15	1.21
Publishing	1.05	5.81	1.90	2.81	0.22	1.72	1.25	1.38
Chemicals	1.46	6.46	2.46	2.56	0.20	1.64	1.23	1.33
Rubber	0.47	2.98	1.72	1.75	0.17	1.50	1.22	1.23
Minerals	0.56	3.21	1.86	1.77	0.17	1.54	1.22	1.26
Basic metals	0.74	4.40	2.14	2.03	0.14	1.45	1.21	1.19
Fabricated metals	0.64	3.83	1.85	1.96	0.18	1.52	1.24	1.23
Machinery	1.00	3.53	1.78	1.96	0.17	1.52	1.22	1.24
Office	1.68	13.31	2.68	4.99	0.22	1.85	1.42	1.30
Electrical	0.64	4.29	2.03	2.03	0.23	1.75	1.29	1.35
Radio TV	2.00	7.51	1.86	3.93	0.22	1.61	1.21	1.33
Precision	0.64	3.58	1.97	1.81	0.21	1.69	1.24	1.36
Motor	0.60	4.33	2.25	1.87	0.16	1.46	1.23	1.18
Transport	0.62	3.38	1.68	1.99	0.21	1.63	1.29	1.26
NEC	0.64	4.99	2.55	1.90	0.19	1.64	1.30	1.25
Average	0.87	5.21	2.14	2.32	0.18	1.57	1.24	1.26

Source: Author's calculations based on ARD

# Table 6Galton Markov Regression for TFP

	(1)	(2)	(3)	(4)
Dependant variable:	t TFP		GVA/	emp
Lag	0.520	0.485	0.704	0.527
	(0.007)***	(0.016)***	(0.005)***	(0.010)***
Lag >mean		0.074		0.371
		(0.024)***		(0.014)***
Time dummies	yes	yes	yes	yes
Observations	112290	112290	110874	110874
R-squared	0.32	0.32	0.50	0.51

Robust standard errors in parentheses

\* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%

previous productivity above its industry average. This is to allow establishments below the mean to have a different convergence speed to those above the mean. If competition is important Oulton has argued that we expect convergence to be faster for plants below the mean and hence D should be positive.

The results of estimating these equations are set out in Table 6. Column 1 estimates (3) with the dependent variable lnTFP, finding a coefficient on the lagged dependent variable of 0.52. Column 2 estimates (4) . The estimate of the  $\beta_2$  is positive, indicating that convergence is indeed faster for plants with below mean industry productivity. Column 3 repeats

column 1 using value added (to be compatible with Oulton, 1998). The coefficient on the lagged dependent variable is 0.704, which compares very closely with Oulton's 0.74 for independent plants and 0.77 for subsidiaries. Column 4 shows results for the value added version of equation 4. Again the  $\beta_2$  coefficient is positive and with a value of 0.371 somewhat higher than Oulton's 0.17. Thus in both this and Oulton's study, on different data sets, we can conclude that firms or – respectively – plants with productivity below the mean for their industry converge quicker to mean productivity levels.

### Working with matched data

The above is an example of what can be done with the ARD by itself. Here we present some results using matched data with the Community Innovation Survey (CIS).

### The Community Innovation Survey (CIS)

Innovation is seen as an important source of productivity growth. However, whilst the ARD has good data to measure productivity growth, it has no innovation data. A body of work has therefore matched the ARD with R&D data. The great advantage of R&D data is that measures of R&D are reasonably well codified, but R&D is an input to the innovation process and not an output (also, firms might generate technological advance outside formal R&D laboratories which R&D expenditure might not capture). To attempt to overcome these problems, the OECD developed company surveys that measure innovations directly. Such

#### Table 7 CIS 3 and ARD

		CIS 3	Successfully merged with ARD
1	Number of Reporting Units	8,172	3,397
2	Number of Reporting Units in distribution and Services	3,605	98
3	Number of Reporting Units in production (Mining, manufacturing and Construction)	4,567	3,299
4	Number of Reporting Units in Manufacturing excluding sector 23	3,425	3,277
5	Number of Reporting Units in Manufacturing (exc. sec23) after cleaning <sup>+</sup> the CIS3		2,389
6	Number of Reporting Units in Manufacturing (exc. sec23) selected $^*$ in 2000		1,593
7	Number of Reporting Units in Manufacturing (exc. sec23) selected <sup>*</sup> in 1996 or 1997 or 1998 and in 2000		827
8	Number of Reporting Units in Manufacturing (exc. sec23) after cleaning <sup>†</sup> the ARD		716
9	Number of Reporting Units in Manufacturing (exc. sec23) in the final sample		520

+ cleaning means excluding missings and implausible observations.

\* selected means that the reporting units are in the ARD sample with full Census information.

t cleaning means excluding implausible productivity growth values.

§ final is the sample obtained after excluding outliers and missing observations from both the CIS3 and the ARD data sets.

surveys set out a definition of innovations and ask companies to report the output of the innovation process (introduction of innovative new products, new processes, percentage of sales arising from new and improved products, and 'soft' innovations, such as organisational change), the inputs to innovations (R&D, scientists, sources of knowledge) and the obstacles to innovation (finance, bad luck etc.). The Oslo manual (1992) codifies such survey models and the Community Innovation Survey (CIS), carried out in EU countries in the early, mid and late 1990s, implemented the questions. For the UK there have been three CIS surveys, CIS1 (covering the period 1991–93), CIS2 (1994–96) and CIS3 (1998–2000).

The problem is that the CIS data does not contain labour productivity<sup>13</sup> or TFP information. However the UK CIS is carried out using the IDBR as the same sampling frame. We have therefore matched the CIS data with the ARD and here we report our results using CIS3.<sup>14</sup>

The UK CIS is a voluntary postal survey carried out by ONS on behalf of the DTI. ONS randomly selects a stratified sample of reporting units with more than 10 employees drawn from the IDBR. CIS3 covers innovations between 1998-2000. Of the total 19,625 enterprises to which the survey was sent, 8,172 responded (Table 7, row 1), achieving a response rate of 42 per cent.<sup>15</sup> The results of the matching of these 8,172 RUs are set out in Table 7. Of the 8,172 RUs which responded to the survey, 3,397 were successfully matched with the ARD manufacturing data. In 98 cases there was a discrepancy between the industrial classification in the Innovation survey and that of the Production survey. In these ambiguous cases, since the innovation survey is the same for both sectors, we decided to include these RUs in the sample, using also the direct information from the RUs available in CIS3.16

The number of reporting units that are in the manufacturing sector excluding sector 23 (nuclear fuel) according to the ARD are 3,277 as shown in the second column of row 4. Row

5 shows that 1,593 were surveyed in 2000. Since we need longitudinal information to be able to draw growth profiles, we report in rows 6 the number of reporting units that are selected in 2000 and in previous years. In row 7 we report the number of firms for which we can construct an average annual growth rate after having cleaned the ARD dataset from outliers in productivity growth. In the last row we report the number of firms after having dropped missing values and 'problematic' observations in CIS3. This leaves us with a cleaned sample of 520 enterprises. In the productivity growth analysis we choose to use this sample.

# The samples: how representative are they of the whole population?

When using matched data sets, one might be concerned about the fact that the characteristics of the matched sample differ from the full data sets from which the matched sample derives. We look at how two key characteristics of the reporting units in the matched sample, size – measured as number of employees – and sector, compare with those of the overall population of firms in the manufacturing sectors, in the ARD data and in the CIS3 sample.

Figure 2 compares the sectoral composition of CIS3, the ARD selected sample, the IDBR population and the CIS3–ARD matched sample (for which we can construct productivity growth profiles). The food sector (15) appears to be overrepresented and the publishing and media, and the metal products sectors (22 and 28, respectively) underrepresented in our sample. The graph does not report any statistics on the coke, refined petroleum products and nuclear fuel (SIC 23) and the tobacco (SIC 16) industries, since we exclude them from the analysis.

Figure 3 describes the size distribution of firms in the four samples according to five size categories: 10 to 49 employees; 50 to 249; 250 to 499; 500 to 999 and 1,000 or more employees. The graph shows that our matched sample

# Figure 2

Differences in sectoral composition: matched sample, CIS3, Selected ARD, IDBR by sector of manufacturing Per cent



overrepresents medium size enterprises. Also the graph shows that in our samples we do not have firms smaller than 10 employees because these were not surveyed in the CIS3.

We finally consider how the key innovation variables in the matched data set compare with that in the 'cleaned' CIS3.<sup>17</sup> Table 8 shows that our matched sample is more innovative than the whole CIS3 according to process innovation (row 5, columns 2 and 5) and the patents<sup>18</sup> (row1, columns 2 and 5) variables but does not appear to present significantly different characteristics from the 'cleaned' CIS3 sample, for all the other innovation indicators.

In sum, we have the following concerns about our matched sample. First, it is more skewed to medium sized firms. Second it is more 'innovative'. Third, in respect to the whole population, it overrepresents medium-tech sectors.

#### Innovation and productivity growth

Using the matched CIS3-ARD sample we are in a position to investigate the impact of innovation on productivity growth. To investigate this we construct TFP-growth as (Criscuolo and Haskel, 2003)

$$\Delta \ln TFP_{it} = \Delta \ln Y_{it} - \sum_{j=1}^{n} \bar{s}_{ji} \Delta \ln X_{jit}$$
(5)

the bar over the *s* denotes the time average share of input *X* in total output and the *j* inputs *X* are *K*, *L* and *M* from (2).

# Figure 3 Size composition of the matched dataset, CIS3, selected ARD and IDBR

Per cent



Firms are asked if innovations are new to the industry or new to the firm. We take this as a measure of whether firms are novel innovators (i.e. an innovation new to the industry) or innovation imitators (i.e. an innovation not new to the industry but to the firm). Thus we can estimate

$$\Delta \ln TFP_{it} = a_1 \Delta \ln K_{it} + a_{21} Proc\_novel_{it} + a_{22} Proc\_imitate_{it}$$
$$a_{31} Prod\_novel_{it} + a_{32} Prod\_imitate_{it} + a_4 \Delta y_1 + v_{it}$$
(6)

### Table 8

#### Characteristics of innovation measures in the regression samples

		CIS3–ARD matched sample			CIS3 clean		
		1	1 2 3	3	4	5	6
		Median in sample	Mean in sample	Mean for prod. inn.	Median in sample	Mean in sample	Mean for prod. inn.
1	Number of patents	0	3	6	0	1	3
2	R&D intensity (per cent)	0	0.63	1.32	0	0.45	1.23
3	Total innovation expenditure (per cent)	0.86	2.78	4.77	0.55	2.87	5.89
4	Percentage sales new products	0	10.51	25.31	0	8.41	28.72
5	Process innovator (per cent)	0	39.42	59.26	0	25.41	52.57
	Observations		520			2,389	

Source: Authors' calculations.

#### Table 9 Output production function, CIS 3

Product innov	Product innovations measured as % turnover					
	(1)	(2)				
	1998-2000	1998-2001				
Process innovation	0.0084	0.0039				
	(0.0127)	(0.0076)				
Novel process innovation	-0.037	-0.0269				
	(0.0189)**	(0.0108)**				
Product innovation	0.0647	0.017				
	(0.0328)**	-0.0207				
Novel product innovation	0.0347	0.0667				
	(0.0519)	(0.0379)*				
$\Delta lnk_{it}$	-0.1211	-0.1141				
	(0.0697)*	(0.0436)***				
Observations	480	631				
R-squared	0.1	0.11				

where Proc\_novel and Prod\_novel denote innovationsprocess and product respectively- new to the industry and Proc\_imitate and Prod\_imitate denote innovations not new to the industry but to the firm and  $\Delta$ InK is included to control for non-constant returns and/or imperfect competition.

The results of estimating are set out in Table 9. Column 1 measures TFP between 1998–2000 and column 2 from 1998 to the average of 2000 and 2001. Consider first process innovation. Column 1 shows that novel process innovation has a negative effect on productivity growth with a positive (but insignificant) effect of imitative process innovation. Column 2 sheds some light on this; as the post survey period is extended the negative coefficient falls (in absolute value). This suggests that novel process innovations take time to be implemented, leading to a fall in measured TFP growth initially. Such a preliminary dip is the basis of the macro work by Basu et al (2003). The results for product innovation are set out in rows 3 and 4. The effects are positive and sometimes significant for CIS3.

### **Conclusions**

This article draws from our experience in working with the firm and plant level micro data provided by the ONS. It provides an overview of the main issues in making this rich data source usable for economic research. The article focused on definitions and concepts of the Interdepartmental Business Register which is the sampling frame for the Annual Business Inquiry (ABI) – the main source for input and output information – as well as most of other Business level surveys run by the ONS including the Annual Inquiry into Foreign Direct Investment (AFDI), the Community Innovation Survey (CIS).

The article also provides some examples of economic analysis using the ARD and the ARD matched with the CIS. The CeRiBA team has conducted more research using the ARD in conjunction with other micro level datasets. The results of this work can be found on the CeRiBA web page (http://www.ceriba.org.uk). The richness of the data and the possibilities for matching new datasets suggest that there will be much more research in the future which will be informative for policy-makers and academic audiences.

#### Notes

- 1. ARD: Annual Respondents to the Census of Production Database; ABI: Annual Business Inquiry.
- 2. A holding company responsible for a number of enterprise groups is called an 'apex enterprise'.
- 3. The two could nevertheless be separate local units if, for example, an R&D survey which collects data just for the R&D part of the business would identify them as distinct.
- 4. The ABI replaces Annual Employment Survey, Annual Census of Production and Construction (ACOP/ACOC) and the six following Annual Inquiries: wholesale, retail, motor trades, catering, property and service trades.
- 5. The threshold was lower in the past. See Barnes and Martin (2002) for more details.

- 6. The employment size bands are 1–9, 10–19, 20–49, 50–99, 100–249. The regions are England and Wales combined, Scotland and Northern Ireland. Within England and Wales, industries are stratified at 4 digit level, NI is at two digit level and Scotland is at a hybrid 2/3/4 digit level (oversampling in Scotland and NI is by arrangement with local executives). See Partington (2001).
- 7. Partington (2001) states that the AES sent x LU forms to each multi-LU enterprise with x based on the expected number of LUs according to administrative sources. Enterprises with less LUs disposed of excess forms, but since there was no systematic method of obtaining more forms, RUs with more LUs than expected simply did not report on these 'excess' LUs.
- 8. The ABI1 is sent to 78,500 enterprises (in 1998) and ABI2 sent to 75,000 businesses (since it covers slightly fewer sectors; relative to the ABI1 it omits forestry, fishing, financial services, public administration, education, health and social work, doctors and dentists.
- 9. This relates to what Deaton (1997, p701) calls the fundamental argument used by econometricians against weighting. Weighting gives consistent estimators of the parameters that one would have estimated using census data. If the true problem is population heterogeneity, weighting will not solve the problem; neither would indeed the availability of population data. But if the population is homogeneous, unweighted LS gives the 'best' estimates (BLUE) and therefore must be preferred to WLS.
- 10. We have created a lookup table to deal with this.
- 11. An interesting study by Harris (2002) sheds some light on this. He calculates capital stock at the local unit level, thereby taking account of plant closures when calculating the RU capital stock. He finds that his results on the productivity difference between two foreign owned and domestic firms differ from a similar study by Griffith (1999) who worked on the RU level. Comparison of his Tables 2 and A2 suggests that the differences were mainly driven by using weighted regressions rather than local unit data. In Table A2 for example, the coefficients from unweighted regressions of log output on log employment, materials and capital are very similar. In Table 2 using unweighted regressions they are rather different.
- 12. Fuel and tobacco are hard to measure with tax distortions and recycling is a recently recoded sector in the SIC system which is small and presents some disclosure problems.
- 13. Measured as value added per employee.
- 14. Harris (2001) has matched the UK CIS2 with the UK Census of Production. Examples of matched CIS/Census data are, for France, Crepon, Duguet, Mairesse (1998); for Holland, Klomp and van Leeuwen (2002); for Sweden, Lööf and Heshmati (2001); and for Finland, Leiponen (2002).
- An interesting question is how representative the responses are of the underlying population, see Criscuolo and Haskel (2003). We confine ourselves here to the matched sample.
- 16. The relevant question in the CIS survey reads as follows: "please briefly describe your enterprise's main product".
- 17. 'Cleaned' defines the sample of 2,389 observations.
- 18. The number of patents in the matched sample is on average double that in the cleaned CIS3 sample. Such a difference for this particular measure of innovation is probably due to the strong skewness of this variable.

#### References

Baily M N, Hulten C and Campbell D (1992) *Productivity Dynamics in Manufacturing Plants.* Brookings Papers: Microeconomics

Basu S, Fernald J, Oulton N, Srinivasan S, (2003) The case of the missing productivity growth: or, does information technology explain why productivity accelerated in the United States but not the United Kingdom? July 2003, paper for NBER Macroeconomics Annual 2003, Volume 18, Mark Gertler and Kenneth Rogoff, Editors, presented at a Conference held April 4–5 2003. Available at http://www.nber.org/books/macro18/

Barnes M. and Martin R (2002) Business data linking: an introduction. *Economic Trends*, No. 581, pp 34–41. Available at http://www.statistics.gov.uk/ccl/article.asp?id=135

Carrington W, Eltinge J, McCue K (2000) *An economist's primer on survey samples*. Centre for Economic Studies, US Census Bureau, CES-0015.

Caves DW, Christensen L R, and Diewert WE (1982) Mulitlateral Comparisons of Output, Input and Productivity Using Superlative Index Numbers. *Economic Journal* 92, pp 73–86.

Crepon B, Duguet E and Mairesse J (2000) *Research innovation and productivity: an econometric analysis at firm level*. National Bureau of Economic Research Working Paper, No. 6696.

Criscuolo C, Haskel J, (2003) Innovations and Productivity Growth in the UK: Evidence from CIS2 and CIS3. CeRiBA Working Paper.

Deaton A (1997) The analysis of household surveys. World bank. Johns Hopkins University Press: Baltimore and London

Disney R, Haskel J and Heden Y (2003) Restructuring and productivity growth in UK manufacturing. *Economic Journal*.

DuMouchel W H and Duncan G J (1983) Using sample weights in multiple regression analysis of stratified samples. Journal of the American Statistical Association, Volume 78, Issue 383 (September 1983), pp 535–543.

Griffith R (1999) Using the ARD establishment level data to look at foreign ownership and productivity in the UK. *Economic Journal* 109, pp F416-F442.

Harris R (2002) Foreign ownership and productivity in the United Kingdom - some issues when using the ARD establishment level data. *Scottish Journal of Political Economy*, vol. 49, August 2002, pp 318–335.

Harris R and Drinkwater S (2003) UK plant and machinery capital stocks and plant closures. Draft paper, available at http://www.dur.ac.uk/richard.harris/

Haskel J and Martin R (2002) The UK Manufacturing Productivity Spread. Ceriba, mimeo.

Jones G (2000) The development of the Annual Business Inquiry. *Economic Trends*, No. 564, pp 49–57. Available at http://www.statistics.gov.uk/ccl/article.asp?id=74

Klomp L and van Leeuwen G (2001) Linking innovation and firm performance: a new approach. *International Journal of the Economics of Business* 8, pp 343–364.

Leiponen A (2000) Competencies, innovation and profitability of firms. *Economics of Innovation and New Technology* 9(1), pp 1–24.

Lööf and Heshmati (2001) On the relationship between innovation and performance: a sensitivity analysis. SSE/EFI Working Paper Series in Economics and Finance No. 446. Martin R (2002) Building the Capital Stock. Ceriba, mimeo.

O'Mahony M and de Boer W (2002). Britain's relative productivity performance: updates to 1999. Final report to dti/hm treasury/ons, NIESR, March.

Office for National Statistics (2001) *Review of the Inter-Departmental Business Register*. National Statistics Quality Review Series Report No. 2. Available at

http://www.statistics.gov.uk/statbase/Product.asp?vlnk=6367&Mor e=N.

Organisation for Economic Co-operation and Development (1992). Oslo Manual – proposed guidelines for collecting and interpreting technological innovation data (first edition). OECD: Paris.

Oulton N (1997) The ABI respondents database: a new resource for industrial economics research. *Economic Trends* No. 528, pp 46–57.

Oulton N and O'Mahony M (1994). Productivity and growth. A study of British industry. Cambridge University Press: Cambridge.

Partington J (2001) The Launch of the Annual Business Inquiry. *Labour Market Trends*, Vol. 109 No. 5, pp 259–268.

Perry J (1995) The Inter-Departmental Business Register. *Economic Trends* No. 505, pp 27–30.

Perry J (1985) The Development of a New Register of Businesses. *Statistical News* 70, pp 13–16.