Austrian Journal of Statistics
Volume 33 (2004), Number 1&2, 89-97

Business Data Linking – Recent UK experience

Felix Ritchie
Office for National Statistics, London

Abstract: This paper discusses recent developments in the UK Office for National Statistics in linking business data. The regulatory background, the history of the Business Data Linking Project, and a number of important results arising are described. Creating good microdata from survey data collected for aggregate analysis causes some problems, and other difficulties with generating usable data from micro studies are considered. The paper notes recent developments which improve the analytical process and offers suggestions to improve the effectiveness of microdata linking.

Zusammenfassung: In dieser Studie werden neuere Entwicklungen der nationalen Statistikbehörde (UK Office for National Statistics) zum Zusammenführen von Wirtschaftsdaten diskutiert. Die Studie beschreibt den rechtlichen Hintergrund bezüglich Regulierungsgesichtspunkten, die geschichtliche Entwicklung des Projekts "Business Data Linking Project" sowie einige wichtige bisher erzielte Ergebnisse. Thema sind etwa das Generieren guter Mikrodaten aus Daten, die zur aggregierten Analyse erhoben wurden und Probleme, bei der Generierung von nutzbaren Daten aus Mikrostudien. Die Studie nimmt Bezug auf neuere Entwicklungen zur Verbesserung der analytischen Verarbeitung und bietet Vorschläge zum verbesserten „microdata linking".

Keywords: Business data linking, microdata linking

Business data in the UK

1 Collection of business data

The Office for National Statistics (ONS), like most NSIs, holds a register of businesses which is uses to create the sampling frame for surveys. The UK version, the Inter-Departmental Business Register (IDBR), has now been in operation since 1994 and is intended to provide a common thread for all government surveys.

Most business surveys carried out by the ONS are collected under the Statistics of Trade Act 1947 (STA), which makes completion of the survey compulsory but limits the use to statistical purposes for the benefit of ONS. This means a high response rate and high-quality data but severe restrictions on access to and use of the microdata.

In theory other government departments should also use the IDBR when surveying business. This is primarily so that the burden on business can be managed. The UK has an independent Survey Control Unit whose task is to ensure that, amongst other things,

- the compliance burden is identified
- the survey need is justified
- alternatives are thoroughly explored

The aim is to manage the sample so that businesses are not overburdened with forms. The sample fraction increases with firm size, so small firms are less likely to be sampled. Small firms which have recently been sampled may be excluded from other surveys for a period. In addition, the "Osmotherly Rules" try to ensure that the very smallest firms would not face more than one statutory survey every three years.

1.1 The regulatory environment

The UK has a devolved political and statistical system which makes the legislative framework complex. There is no single "Statistics Act", and government bodies are all separate legal entities. There is no right to distribute data amongst parts of government, and no overarching legislation. Instead, use of micro-data is controlled by:

- specific pieces of legislation (for example, Data Protection Act, STA)
- case law, which has defined confidentiality and which means that law may be undefined until tested
- limits on administrative powers of authorities

The combination of a devolved government system and the use of the STA to collect business data means that access to microdata for research is limited to analysis on-site at ONS premises for researchers working under ONS control.

2 The Business Data Linking (BDL) project

2.1 History of business data linking

Creation of linked data sets for business demography and productivity work in manufacturing has been under way in ONS - in different forms - for five years. Work was begun in the late 1990s to create a longitudinally linked data set from the 1970s based on the Annual Business Inquiry (ABI) and its predecessors which constitute the UK Structural Business Surveys. This project linked, via the business register, data from successive surveys for the manufacturing sector, and tackled a number of difficult problems due to changes in register structure, sampling strategy and survey design.

The resulting Annual Respondents Database (ARD) is now added to each year from the ABI. This has been the core of BDL research output: inherently rich in its own right, the ARD has been linked (through the IDBR reference) to several different datasets to increase the level of analysis possible. To date, surveys linked to the ARD include

- the Annual Inquiry into Foreign Direct Investment (AFDI), to identify all reporting units linked to multinationals, over several years from 1995-2002
- the Business Enterprise Research and Development survey, to analyse social return on R&D expenditure by enterprises during the ten years to 2000
- the e-Commerce Survey , on use of ICT and electronic processes, for 2000-2002
- the New Earnings Survey (NES), which provides detailed information on employee earnings and occupations, from 1986 to 2003
- the Community Innovation Survey, which measures inputs to and outputs from enterprise innovation, and which is now linked for the two most recent rounds, 1996 and 2000
- the Employer Skills Surveys and Learning and Training at Work surveys, which covers in some detail the skills of employees and employer investment in training programmes from 1999 until 2001

In addition, enterprise-level „Permanent Inventory Model" estimates have been built up from the ARD to provide feasible capital stock estimates for productivity work. A plant-level equivalent capital stock has been commissioned.

Most datasets are cross-sectional in construction, being collected from periodic surveys. However, in many cases the same firms are sampled at different periods, and so it has been possible to create some form of longitudinal data, particularly for large firms.

Work is also underway to examine possibilities for linking data from surveys based on other company lists and sample frames. Probabilistic matching using these surveys has been shown to be possible in two cases using combinations of names, address, and telephone numbers.

Finally, a separate project to link the NES employee data into a true panel has been ongoing since the mid-1980s. From 2004 BDL will provide a permanent home for the resulting longitudinal dataset of almost five million observations.

2.2 Method of operation

The concrete form of collaboration has been for ONS to provide infrastructure and data for academic and research institute experts working under contract, and under supervision, at its premises in London. In addition to providing facilities, ONS provides advice in statistical interpretation through its survey experts, who also check the output to guard against disclosure. The researchers are under contract to provide two types of outputs:

- research results which will be published, and may be used by government departments and others as evidence to inform policy development on productivity
- linked, documented, data sets on which this research is based, which can later be used by others to check and extend this first round of analysis.

In areas which are of specific statistical interest, such as multinationals, ICT effects, and the labour market, ONS has also provided analytical input.

In the past, BDL has allowed users to work as contractors sharing standard ONS facilities. This has now been changed so that researchers now work on secondment, and a purpose-built "safe setting" was introduced in January 2004. This safe setting is based upon thin-client technology similar to a system employed by Statistics Denmark, and allows data to be accessed efficiently at ONS offices throughout the UK. Due to legal restrictions there is little possibility of general internet access to this data, but some extension of activity across the country using secure networks is being investigated in the medium term.

The data made available to researchers is not identified but is identifiable. It is not possible to anonymise this data while still retaining its usefulness. BDL therefore has stringent disclosure control procedures designed to ensure the confidentiality of contributor data. These procedures are based upon the standard ONS rules but adapted for the peculiar features of a research environment . In addition, BDL requires that all researchers (including internal staff) undergo a training session which includes a practical guide to disclosure control.

2.3 Research Outputs

A large part of BDL's research has been sponsored, directly or indirectly by other government departments, with ONS' internal research mainly concentrating on analysis of the ICT sector and labour markets.

Examples of outputs include:

- linking the ARD and AFDI permitted researchers to identify enterprise productivity effects associated with multinational operations, and to separate them from the issue of foreign ownership. The results show (as similar work in the US and Sweden has done) large productivity advantages associated with multinational operations, irrespective of country of origin, after taking account of sector, scale, and capital input and other relevant factors. It suggests that multinationals are able to exploit shared intellectual capital, not captured by current surveys. These results have had a major impact on the productivity agenda in the UK.
- early analysis of buying or selling over the internet from linked ARD/e-commerce data appears to show significant price effects and gains in welfare (producer and consumer surplus); but the gains are asymmetric with buyers coming out as clear winners in most industries. However, in monopolistic or oligopolistic industries there does appear to be some gain to sellers, suggesting that effect of ICT is to attenuate existing market conditions.
- linking the New Earnings Survey, the ARD and the Employer Skills Survey shows a significant link between skill levels and productivity; but also that firms with higher productivity tend to hire workers with more formal schooling, implying a genuine return to the firm from general human capital.

For more detailed discussions, see Barnes and Martin (2002) and Criscuolo and Waldron (2003).

3 Problems in creating linked microdatasets

3.1 Dominance of aggregate statistics

ONS' primary purpose in collecting business data is to provide timely and accurate macroeconomic aggregates. Microanalytical research is a relatively new phenomenon and therefore ONS' systems do not yet fully incorporate this new development. The consequences of this are that data appropriate for macro analysis may not be adequate for micro analysis:

- **microdata quality** suffers. If the primary use of survey data is to present macro statistics, it is not cost-effective to check all returns. Many NSIs now employ selective (or significance) editing, where tolerances and automating editing tools are used to ensure that macro statistics are accurate at a minimal cost. Only survey values that are thought to make a significant difference to macro statistics are reviewed in detail. This is inappropriate at a micro level where variable correlation is at the core of analysis.
- **redefinition of data** is more of a problem for micro research than macro research. For example, the change in industrial classification from SIC80 to SIC92 is not an exact split and requires a small amount of probabilistic matching. At the macro level this gives similar results to deterministic reclassification, by construction. However, at a micro level different allocation methods can affect microanalytical results.
- **interpretation** may change. For example, there seem to be significant inconsistencies in "innovation" as measured in the two Community Innovation Survey waves. However, it may be that activity which was classified as innovative in the mid-1990s is now seen as part of the ordinary competitive process by firms. Again, tabular aggregates can handle this adjustment more easily than microdata
- the **longitudinal integrity** of the data is largely a micro concern. For example, many of the same firms are surveyed each year for the ABI. Firms that are taken over, demerge, or otherwise change their corporate status may be assigned new identifiers. This does not affect the aggregate statistics, but it can play havoc with attempts to create a panel of firms from the identifiers alone.
- **documentation** may be limited as the primary focus is on explaining aggregates rather than differentiating between (for example) imputed, returned and calculated values. This is particularly true for electronic documentation.

3.2 Sampling frame

Most statutory surveys have an element of sampling. Therefore, when linking data over time or across datasets there is an issue of overlapping samples. This is particularly relevant for SMEs, where the sampling fractions tend to be small. Particular problems are

- as noted, surveys are explicitly designed to avoid repeatedly **sampling small firms**. Hence by construction there is little linkable data available for the smallest firms unless several years worth of surveys are available. In theory „data fusion“ or similar techniques could fill these holes but there has been some concern over this (see Chesher and Nesheim (2004) for a literature review).
- even for surveys with a census element, the **census band** may change. For example, in 1970 the ABI's forerunner was a census of all firms employing over 25 people; by 2000 this had been reduced to firms employing over 250.
- for **voluntary surveys** the response rates are lower and there is some evidence of bias in respondents. For example, the Community Innovation Survey response rates are noticeably lower in large firms; this is problematic as these firms are thought to be significant innovators.
- some surveys select from **non-IDBR populations**. For example, linking the New Earnings Survey's tax reference codes to the IDBR introduces an extra element of uncertainty as registered tax points and business units may not correspond exactly.

3.3 Inconsistencies across datasets

There are significant inconsistencies across datasets. This may be due the wording or interpretation of questions: about 30% of firms which appear in both the eCommerce Survey and the ARD claim to use electronic networks in one survey and deny it in the other. It may also be due to different definitions: for example, the AFDI and ARD descriptions of foreign involvement do not tally, but the AFDI includes subsidiaries and associates whereas the ARD only records a controlling stake.

3.4 Confidentiality restrictions

As noted above, the data used is identifiable survey returns. The linked datasets are extremely disclosive and access is strictly controlled. This raises its own problems:

- **clearing results** for publication is more difficult, as the “owners” of all datasets need to be involved.
- when linking across datasets, small numbers arising from **limited sampling overlap** can lead to low cell frequencies and hence unacceptably disclosive results

- for **non-ONS data** there is no automatic right to share information, so access to non-ONS surveys has sometimes required complex legal negotiations

4 New developments

ONS has recently been changing to accommodate the expanding use of its data for micro research:

- documentation is more timely and available electronically
- automatic matching across datasets has significantly increased the number of datasets that can be linked to the ONS data. The ONS Sources directorate has developed tools to match on a variety of criteria with a high success rate.
- research is being used to provide feedback for the surveys. For example, questions about ICT on the ARD have been changed directly as a result of research on the microdata.
- a long-term program has been launched to provide an integrated metadata system which, as part of its outputs, is meant to provide ONS-wide consistency in definitions, questions, phrases, variables names, and so on.
- a body of documentation and expertise on statistical disclosure control in a research environment has been developed
- generally there is an increasing awareness inside and outside ONS of the value of microdata research, particularly as a source of added value in ONS datasets.

5 Conclusions – what have we learnt?

Several key lessons can be learnt from the UK experience:

- **a good relationship with data providers is essential**. Getting them involved early and enthusiastically makes a large difference. As well as co-operating with the supply of data, they can offer useful comments on the quality of the data and can suggest alternative sources of data, and they are well placed to audit cleaning and linking procedures.
- in return, there should be opportunities to feed back to the data providers useful information about the dataset.
- if **disclosure checking** is also required than this is much easier when all parties involved know the use to which the data is put and the aim of the research; again, procedures should be in place from an early stage.
- arrange **specific training** on disclosure control in a research environment for those responsible

- **track data closely**. Survey data is often subject to revision. The manager of a microdata resource needs to be clear which are the “definitive” source files and how they relate to other versions and to published results.
- **never assume data is “clean”**. Check within the dataset, look for duplicates and cross check across sources and over time. Data collected for macrostatistics is not checked in the same way as micro data. Linking of datasets across time or sources can throw up inconsistencies which don’t appear in macro totals.
- **be aware of the organisational setting**. National statistics institutes (NSIs) are designed to produce macrostatistics. Data management systems, dissemination programmes, metadata schemas and so on may not be suitable for microdata access, or to provide cross-dataset linking. In a typical NSI microanalytical research is a minor function of the organisation. The microdata manager must allow for this, and may need to take a very proactive stance to get results.

The potential benefits of using linked microdatasets for analysis are enormous, but working with microdata is a novel experience for many NSIs. This is certainly the case in the UK, where a collaborative approach has led to the creation of a significant new research resource – and an awareness of the systems that generated the data has made the resource practical.

References

M. Barnes and R. Martin. Business Data Linking: an introduction, *Economic Trends* no. 581 (April) pp34-41; London: Office for National Statistics. http:/www.statistics.gov.uk/cci/article.asp?id=135 . 2002.

A. Chesher and L Nesheim. Review of the literature on the statistical properties of linked datasets, Report for the UK Department of Trade and Industry. May 2004.

C. Criscuolo and K. Waldron. E-commerce and productivity“, *Economic Trends* no. 600 (November) pp52-57; London: Office for National Statistics. http:/www.statistics.gov.uk/cci/article.asp?id=597. 2003.

F.J. Ritchie. Access to business data in the UK: the regulatory context for government collections; mimeo, Office for National Statistics; presented to the 2003 Conference on Comparative Analysis of Enterprise Microdata, London. http://www.statistics.gov.uk/events/caed/abstracts/downloads/ritchie.pdf .2003.

Author's address:

Dr. Felix Ritchie
Business Data Linking
Office for National Statistics
1 Drummond Gate,
London SW1V 2QQ
Great Britain

Tel. +44 (0) 207 533 5975
Elec. Mail: felix.ritchie@ons.gov.uk
http://www.ons.gov.uk