10-5-2005

Adjusting Imperfect Data: Overview and Case Studies

Lars Vilhuber
Cornell University, lv39@cornell.edu

Adjusting Imperfect Data: Overview and Case Studies

Abstract
[Excerpt] In this chapter, instead of using the similarity in the cleaned datasets to investigate economic fundamentals, we focus on the differences in the underlying 'dirty' data. We describe two data elements that remain fundamentally different across countries, and the extent to which they differ. We then proceed to document some of the problems that affect longitudinally linked administrative data in general, and we describe some of the solutions analysts and statistical agencies have implemented, and some that they did not implement. In each case, we explain the reasons for and against implementing a particular adjustment, and explore, through a select set of case studies, how each adjustment or absence thereof might affect the data. By giving the reader a look behind the scenes, we intend to strengthen the reader's understanding of the data. Thus equipped, the reader can form his or her own opinion as to the degree of comparability of the findings across the different countries.

Keywords
economics, dirty data, administrative data, wages, mobility

Adjusting imperfect data: overview and case studies

Lars Vilhuber, Cornell University*

lars.vilhuber@cornell.edu

This paper is a draft chapter for a forthcoming NBER book edited by Ed Lazear and Kathryn Shaw.
This version: 10 November 2005

I am indebted to all the authors of the country-specific chapters for having provided me with detailed data descriptions, allowing me to write this chapter. Juhana Vartiainen, Lia Pacelli, Roberto Leombruni, Claudio Villosio, and Bruno Contini provided valuable contributions beyond their data descriptions. I am thankful to all of the above, John Abowd, and Julia Lane for comments on drafts of this text. All errors, of course, remain mine.

1 Introduction

This book combines an astonishing variety of data sets in a coherent analytical framework. Data have been extracted from the administrative systems of countries with very different approaches to administration, from highly centralized countries such as France to fiercely de-centralized countries such as the United States, from countries with as few as 5 million inhabitants (Finland, Norway, Denmark) to the 293 million inhabitants of the United States. Some statistical agencies conduct coordinated surveys to gather the data; others rely on purely administrative tasks to co-incidentally gather the data.

The ease with which the reader can compare the analyses presented in this book was purchased through much hard work by the authors from each country. In particular, each set of authors had to adjust their data for quirks, problems, and issues that are inevitable when working with large administrative datasets, and handled them in what their experience told them was the best manner, given the constraints both of their data and the homogenization imposed on all by this project. It goes without saying that not all solutions are identical, and how they differ may affect how the data is to be interpreted. Furthermore, some known issues with the data were left untouched, in order to make the data more comparable between countries. Finally, some elements of the data, such as the unit of observation, remain fundamentally different, and it is important to keep that in mind when comparing data across countries. The end result, however, is a very high degree of comparability achieved by the authors.

In this chapter, instead of using the similarity in the cleaned datasets to investigate economic fundamentals, we focus on the differences in the underlying ‘dirty’ data. We describe two data elements that remain fundamentally different across countries, and the extent to which they differ. We then proceed to document some of the problems that affect longitudinally linked administrative data in general, and we describe some of the solutions analysts and statistical agencies have implemented, and some that they did not implement. In each case, we explain the reasons for and against implementing a particular adjustment, and explore, through a select set of case studies, how each adjustment or absence thereof might affect the data. By giving the reader a look behind the scenes, we intend to strengthen the reader’s understanding of the data. Thus equipped, the reader can form his or her own opinion as to the degree of comparability of the findings across the different countries.

The structure of this chapter is as follows. A first section provides an overview of select data elements of all data sets, and discusses the similarities and differences. The next several sections

discuss longitudinal and other linking issues, outline why it is important to properly handle such issues, and provide examples of applications in the data sets underlying the other chapters of this book. Case studies are summarized where appropriate and available. The case studies differ from the applications in that they (typically) do not use the same data sets, but provide a deeper analysis of the same method used on the data sets in this book. They are thus able to provide some empirical insight into the importance of the data adjustment.

2 Overview of data sets

We start out with a brief overview of all data sets used in this book. The reader is referred to each chapter itself for a detailed description of each data set. Some of the data sets used in this book have also been described previously in Abowd and Kramarz (1999), which also contains an exhaustive list of other matched employer-employee datasets.

[Table 1 here : sampling frame, plant unit,]

2.1 Sampling schemes

The data sets used in this book were constructed using several different types of sampling schemes (see column Table 1). Essentially three sampling schemes are present: worker-and-firm universe files, worker-based samples, and firm-based samples.

Worker-and-firm universe files are not samples, though there may be some smaller coverage issues. Workers and firms appear in the dataset because they are covered by a universal entitlement or tax system. The Danish CCP data, the Swedish RAMS data, and the US LEHD data are examples of such files. However, while the Scandinavian data are national registers and thus cover all firms and workers within each country, the US data is compiled from state-level wage record registers and only covers a select number of states within the United States. Within those states, the LEHD covers almost all firms and workers within those states (Stevens, 2002).

The German and the remaining Nordic data sets (Finland, Norway, and the Swedish SAF dataset) are firm-based samples. For a select number of firms, all workers can be identified, but if those workers work for a firm outside of the sample frame, that employment is not captured. Transitions to firms outside of the sampling frame are also not captured. The remaining Nordic data sets are in fact similar to the worker-and-firm universe files used in the United States, with a critical difference. Whereas the LEHD data set covers all workers within a certain geographic area, but does not cover the full geographic area of the United States, the Nordic data sets cover all firms within each respective country that are members of national employer organizations, and have data on all workers working for those firms. From an economy-wide perspective, this much closer

resembles a firm-based sample than a universe file. In fact, if not for the sectoral coverage, the sample obtained looks very similar to the German data, which is an explicitly stratified firm sample of all firms in the economy. The firm-based samples have in common that, in principle, data on all workers working for the firms in the sample are available, although some authors have chosen to work with a sub-sample of workers. Note that a Norwegian national register does exist, and a selection of it is used in conjunction with the selected firm sample in the chapter on Norway, but the full Norwegian register is not used in this book.

The French DADS and the Italian WHIP data are worker-based samples. For some fraction of workers, all jobs with all employers are tracked. Whenever a worker changes firms, his move to another employer is included in the data base, no matter what the activity or sector of the next employer. The constraint of worker-based sampling is that not all workers within a given firm show up in the sample, imposing some restrictions on analysis of the within-firm structure. For this reason Italian statistics on within-firm wage levels and wage changes are computed on cells instead of on firms. The technical appendix to the Italian chapter addresses the issue of cell-level versus firm-level statistics.

How does the sampling frame affect the analysis? Worker-based samples provide excellent data to provide worker-based statistics. The amount of work experience a worker accumulates is well documented, non-employment is well captured, and the earnings trajectory, and earnings changes associated with employer changes, can be followed accurately. On the other hand, firm-based statistics are less well defined. The computation of firm size, if not reported by the firm itself, can be noisy for small firms, and small firms themselves may not be well represented in the data. For instance, consider a 10-person firm, and 1-in-25 worker sample. Naïve estimates for the size of this firm can range from 25 to 250 workers, conditional on at least one worker being sampled, and there is a 66 percent probability that the firm never appears in the data, i.e., none of its workers get sampled. Proper adjustments can be made, but for small firms, firm size estimates will remain noisy. Turnover rates, where employment enters the denominator, and estimates of the within-firm variation can be particularly noisy. The chapter on French data contains some further discussions and analysis of the bias introduced by worker-based sampling.

On the other hand, firm-based samples capture most of the firm measures well, while performing less ably for the worker-based statistics. For instance, the earnings of workers who switch firms can only be measured if workers stay within the sample. Differences will arise here between the German and Scandinavian samples. The latter will most likely capture only workers who stay within the sector, whereas the former will capture more of the industry-switchers, while missing some of the industry-stayers. How this will affect the estimates of earnings changes will depend on whether industry switchers predominantly have larger or smaller earnings differentials than stayers.

Neal (1995), using U.S. data, reports average wage losses for industry switchers of 14 percent, for industry stayers of 6 percent (see also Parent, 2000). For France and Germany, the literature seems to indicate that workers tend to have earnings gains rather than losses (Bender et. Al, 2002), but whether there is a differential gain between stayers and switchers is unknown. For Italy, Leombruni and Quaranta (2002) report average wage changes for industry switchers about 3 percentage points lower than those of job movers within the same industry, although Contini and Villosio (2003) find small wage gains for industry switchers.

2.2 Aggregation levels and the concept of an employer

The datasets also differ substantially in another dimension. Whereas the unit of observation on the person side is always well defined, the entity represented in the data on the employer side is not as clearly specified. Although all entities are "employers" in the sense that they have hired the workers, different levels of aggregation are present in the data. Some correspond to physical plants, some are administrative units that may be smaller or larger than any single plant, and some are "firms" or "enterprises" that are better defined by ownership relationships than physical location. Each aggregation level typically has a different identifier, though link files may exist.

The administrative data for each country typically have observations on one specific level of aggregation, although additional variables or links to other datasets may allow for higher-level aggregations. Furthermore, since some of the data are merged from different sources, not all the detail may be available at the lowest aggregation level. Table 1 tabulates what the lowest level of aggregation is for data on employer characteristics for each country. The aggregation level on the files containing job characteristics, if different, is pointed out in the discussion below.

The Nordic data, for the most part, report employer and job characteristics at the level of a physical plant or establishment. This allows allocating an individual to a particular plant, at least once a year. However, the data obtained from employer associations may have the feature that only the workers of a particular type (blue- or white-collar) are identifiable. While in Norway, blue- and white-collar datasets can be recombined by firm, this is not feasible in the Swedish SAF data, and an imperfect process in Finland. On the other hand, the Norwegian, Finnish, and the Swedish RAMS data can typically identify both the firm and the plant a worker is associated with, and can thus explore both within-firm and within-plant variation in wages and other measures.

In Italy and the United States, the smallest unit of observation on the employer characteristics file is a reporting unit respectively for the social security pension system and for the unemployment insurance system, which typically corresponds to a plant, but maybe either larger or smaller. In both countries, the choice is up to the employer, and some employers report all establishments within a

large geographic area (a state) on a single record. Furthermore, the records can be aggregated up to a "firm." In the American data, only a state-specific identifier identifies this firm, and the data used in this book cannot identify which firms in different states are actually the same employer. In Italy, as in France, the firm is defined at the national level. However, whereas in Italy, a worker's job can be associated with a particular reporting unit, this is not possible for the United States and France.

Finally, in Belgium, as in France, only a "employer" can be identified, and associated both with additional details on the firm as well as with the worker.

In summary, the Nordic data, in general, report statistics calculated at the both the firm and the plant level in this book, whereas France, Belgium, Italy, the Swedish SAF data, and the United States calculate statistics at the firm level. This aspect of the data needs to be taken into account when comparing "firm" size statistics, turnover, and the variability of earnings within a "firm" or "plant." The difference between plant and firm is a critical distinction and is discussed in more detail in many other locations in this book.

3 Longitudinal linking (Identifier issues)

3.1 Coding errors in person identifiers

The data used in this book are typically used for administrative purposes, and the widespread perception is that administrative data are objective and comprehensive. However, that does not ensure that they are perfect. One particular problem affecting the millions of person records in each of the administrative databases is coding errors. And although coding errors can occur in every item on the "wage record," the variable analysts are most worried about is the person identifier.

3.1.1 When do coding errors occur

Coding errors occur for a variety of reasons. A survey of 53 state employment security agencies in the United States over the 1996-1997 time period found that most errors are due to coding errors by employers, but that when errors were attributable to state agencies, data entry was the culprit (Bureau of Labor Statistics 1997b, pg. ii). The report noted that 38% of all records were entered by key entry, while another 11% were read in by optical character readers. OCR and magnetic media tend to be less prone to errors. Similar errors are known to be present in European data as well, but the extent will vary considerably from one country to the next.

The types of errors differ by the source of the error. When a record is manually transcribed by an employer onto a paper form, scanned, or entered by hand when entering the state agency's data warehouse, the most likely error is a random digit coding error for a single record in a worker's job

history. Errors that occur persistently over time will typically be the result of recording a wrong or mistyped person identifier in an employer's data system, which is then repeatedly transmitted to the state agency.

Check digits allow the verification of the validity of a person identifier without reference to any external data, and generally prevent, or at least highlight the presence of coding errors, allowing for easy correction at data entry. However, most person identifiers do not incorporate check digits. In the data used in this book, only the Norwegian and Italian data are known to have checksums on the person identifier.

One last reason why such errors persist, and are not corrected, is that none of the involved parties has a strong incentive to actively search for and obtain more accurate records on an ongoing basis. The primary focus of the data collection is typically cross-sectional. In the United States, the wage records are collected in the course of the administration of payroll taxes and unemployment insurance systems. Only the sum of wages by firm is used by the collecting agency at the time of collection, ensuring that the firm identifier is generally considered very reliable (but see the next section for exceptions to that statement). In Italy, the primary purpose of collecting the contribution data is for the national pension system. In both cases, for the ultimate beneficiary, the worker, longitudinal consistency only becomes relevant when filing an unemployment or pension benefit claim, possibly years after the coding error was entered into the database. Absences in contributions are corrected using workers' copies of contribution reports, and at least in the U.S. are known not to flow back into the actual wage record database. The Italian data typically does not have the coding problem, in part because incentives may be properly aligned, in part because of the presence of a check digit on the person identifier number (Revelli 1996).

3.1.2 What is the impact

Most flow variables (accessions, separations, length of tenure at a firm, etc.) are constructed by associating a person entity – a human being – with a particular person identifier, and constructing job histories based on that identifier. Continuity of employment for a given person is inferred from the presence of two records at different points in time bearing the same person identifier. Coding errors in the person identifiers will generate spurious job interruptions that affect all flow variables. Systematic and random errors in the person identifier will generally bias upwards flow statistics, but tenure will be biased downwards.

The necessity of making a valid longitudinal integration of information for the same individual collected at two different points in time with incomplete linking information is not a new problem in economic measurement. Indeed, probabilistic record linking applications have flourished as a

part of research programs that seek to improve such measures. For example, there is a large literature discussing the difficulty of inferring the continuing employment status of an individual between two reference dates using consecutive months of the CPS (Fienberg & Stasny 1983, Abowd & Zellner 1985, Poterba and Sommers, 1985, and others). Flows into employment, unemployment and nonparticipation are biased by incorrect longitudinal linkage for exactly the same reason as the accession and separation statistics based on the UI wage records are potentially biased.

3.1.3 Solutions

Methods exist that can avoid or correct for such coding errors. The national person identifiers in some countries have a check digit, which allows the identification at data entry of whether or not the person identifier was correctly entered. However, for many countries and administrative systems, changing a pervasive identifier without a checksum to a more stable identifier system is not a feasible alternative, or at least a very costly one, and certainly won't work with historical data files.

The practical solution to most coding error problems is automatic and manual editing procedures. Most wage record data bases contain names, and inspection and matching of records based on names is a reliable, though not perfect method of linking records into one consistent job or employment history. Additional information, such as demographic information on the file with the miscoded record and matching information on other files, may facilitate the matching exercise and improve the match rate. The problem is the sheer number of records, which at least for person identifiers makes regular manual editing impossible, and has only made automatic editing procedures feasible in the last couple of years, at great computational cost. Often, the simplest solution is to simply drop records that are identifiable as being miscoded. This is the case in Finland, whereas most other countries continue to include such records.

3.1.4 Application: Imputation to correct for coding errors in France

In the French data, a different approach was taken to tackle the same problem (coding errors in the person identifier NNI due to key punch error). As before, as a consequence of coding errors, some job observations, identified by a NNI-SIREN (firm identifier) combination, appear only for a single year in the data. Furthermore, this job is the only one ever registered for this particular NNI. Other job histories will present a single one-period interruption. Consider now the case of a worker with observations in, say, 1978 and 1980 in the same firm (SIREN) but no observation for 1979. If true, this history would mean that the worker was employed until some date in 1978 (depending on the number of days worked, December 31 most likely) and also employed after some date in 1980

(depending on the number of days worked, January 1 most likely) in this firm but not employed at all during year 1979. This is very improbable. INSEE thus adopted the following solution: Whenever an observation was missing in a given year while the same NNI-SIREN combination exists for the preceding and the following year, an observation is created for the missing year with the same NNI-SIREN combination. Earnings are computed as the geometric mean of the preceding and following wages (in real terms). All other variables are taken at lagged values. For the entire French data set, this procedure added 193,148 observations, or about 1.2% of all records.

3.1.5 Case study: The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers in the USA (Abowd and Vilhuber, 2005)

Abowd and Vilhuber (2005) describe the method used by the U.S. Census Bureau to identify coding errors in the person identifier (Social Security Number, SSN), and provide an analysis of the impact that correcting for such errors has on statistics generated from the corrected and the uncorrected data. Their analysis only covered one of the states that are used in the U.S. data chapter of this book, but are generalizable.

First, job histories (the unique combination of an employer identifier SEIN and a person identifier SSN) are constructed. It is posited that the most likely coding error (random transposition of identifiers) results in (a) a single-period job history for some SSN-SEIN combination and (b) a job history with a single-period interruption. Records are extracted from the wage record database that fit one or the other of the job history profiles. A probabilistic matcher is then used to compare names, the miscoded SSNs themselves, as well as earnings to identify matches.

The process verified over half a billion records. The number of records that are recoded is slightly less than 10 percent of the total number of unique individuals appearing in the original data, and only a little more than 0.5% of all wage records. Trials in the late 1980s using Unemployment Insurance wage records found an average error rate of 7.8 percent, with significant variation across states (Bureau of Labor Statistics 1997b). The authors estimate that the true error rate in their data is higher, in part due to the conservative setup of the process. Over 800,000 job history interruptions in the original data are eliminated, representing 0.9% of all jobs, but 11% of all interrupted jobs.

Despite the small number of records that are found to be miscoded, the impact on flow statistics can be large. Accessions in the uncorrected data are overestimated by 2%, and recalls are biased upwards by nearly 6%. Payroll for accessions and separations are biased upward by up to 7 percent.

3.2 Quality of firm links, measures undertaken and not undertaken to improve links.

3.2.1 Errors in links.

The same mechanisms that generate coding errors in person-level data can work on employer data as well. Whether mistyping a person or a firm identifier when transferring information from paper to electronic format, the result is a break in an individual time-series. However, several factors combine to make this a problem both less pervasive and more difficult to correct for in firm-level data.

3.2.2 Administrative vs. economic identifier changes: the concept of a 'firm' in administrative data

People can and do change names, and possibly other theoretically permanent physical attributes, but they always remain a single human entity. The numeric person identifier attributed to a person is only changed in very rare and exceptional instances.

That same intertemporal uniqueness does not necessarily hold for firms. Tracking firms in data, and in particular administrative data, thus poses additional challenges. Firms can be born, split, merge, and disappear. Changes in ownership, of legal and organizational form, and changes in products and services offered can all lead to legitimate and legal changes in administrative identifiers. The very boundaries of what constitutes a single economic entity called a "firm" are often fluid.

For the purposes of the papers in this book, the fundamental focus is on firms as places of work for workers, i.e., the firm as employer. As such, the fundamental economic activity that the authors have attempted to isolate is the employment of a set of individuals which, taken together, constitute the "firm". Under that premise, the tenure of a worker should not be affected by purely administrative changes of the employer identity. But should it be affected by a merger or the transfer of a plant from one firm to another? The identification of an economic, rather than legal successor to a firm becomes an important distinction.

3.2.3 Impact of failure to properly link firms

One of the focal statistics used in this book is the "average change in wage from workers who change firms". The failure to properly link firms that change administrative identifiers without an underlying economic event can, under certain circumstances, bias that statistic. Consider an economy with strongly hierarchical firms having few ports of entry and positive returns to tenure. The literature describes various theories, and provides examples of firms, that have at least some aspects of such a personnel policy (Baker, Gibbs, and Holmström, 1994; Lazear, 1995). By definition, these ports of entry will be at the lower end of the firm-internal pay scale. A worker

entering this firm will typically do so at one of the ports of entry, and thus receive a wage that is below the firm average. As a consequence, the average wage of *all* workers entering this firm will most likely be below the firm-average wage.

Now consider a firm that changes legal form, thus changing its administrative ID in the system, and for some reason, this is not captured in the administrative follow-up. No workers leave the firm, and no workers join the firm. The average change in wage from workers changing firms calculated from this particular subset will be equal to the firm-average wage, substantially higher than if firm changes were measured accurately. If such occurrences are frequent enough, the entire statistic can be biased upwards by a significant margin.

Other research would also be affected. An extensive literature shows that a large fraction of workers that are part of a mass layoff have some difficulty re-entering the labor market, showing significantly negative effects on the earnings history (Jacobsen, Lalonde and Sullivan, 1993) or difficulties in finding a new job (Margolis 1999). An identifier change results necessarily in the observation of a mass layoff, albeit not a real one. However, the workers of such an identifier-induced "mass layoff" do not suffer any earnings problems since in fact they are never laid off. Measures of turnover – separation and accession rates – are also driven by the quality of the link, with missing links biasing both measures upwards (Spletzer, 2000, Benedetto et al., 2003, Vartiainen, 2004).

3.2.4 Following up with firms

Administrations are also interested in linking firms for other reasons. In particular in the United States, payroll taxes can be experience rated, and firms with a higher payroll tax rate have an incentive to change identity, and become an apparently new entity not subject to the predecessor's higher tax rates. Administrations follow-up on firms, and the US administrative data contain a field that identifies a possible legal or legally obligated successor. In Italy, the Italian Institute for Social Security (INPS) distinguishes "insurance records" (the basic entity on the firm-level file) from "firms", identified by a single (firm) social security number. The level of disaggregation, i.e., the number of insurance records that a firm decides to have, is arbitrary, and at the discretion of the firm. But all insurance records can be linked back to the same legal entity, identified by a social security number.

Thus, administrations typically have incentives to properly identify the firm, both at any point in time and across time periods. Most administrative data sets on firms contain some information about a firm's legal predecessors and successors, and this can, if so desired, be used to link firms. This mechanism is known to be used in the United States (Spletzer 2000). In other jurisdictions,

administrative follow-up may simply mean that no new identifier gets assigned if the firm or establishment is economically the same (Vartiainen, 2004).

3.2.5 Reverse-engineering code changes

In some cases, changes to the coding system have radically altered the identifying codes, resulting in a discontinuity in the time series. One of the reasons this may arise is that the agency collecting the data is not obligated to provide continuity, as in the case of the Finnish employer organization (Vartiainen, 2004). Also, the purpose of collecting data may (again) be primarily cross-sectional, with little benefit to the agency of maintaining longitudinal consistency. Finally, extraction and transcription problems when accessing or retrieving historical data series may introduce errors to all records of certain time periods.

When such coding changes occur, researchers do not always have access to the historical documentation detailing the code changes, and need to reverse-engineer the coding changes. Many of the methods described subsequently in this section (probabilistic matching, flow analysis) can be used as intermediate tools, rather than actual corrective measures, to identify the way in which coding conventions have changed. Vartiainen (2004) used flow measures to identify pairs of likely "stayers," workers who did not change employers despite a change in their employer's identification code. The resulting pairs of consecutive-year records for the same worker combined with visual inspection allowed the researchers to correct algorithmically for the changes in the establishment codes that had occurred in several years of the Finnish data (Vartiainen, 2004).

3.2.6 Using probabilistic matching again

Statistical agencies and researchers also employ probabilistic name matching techniques to link firms. The U.S. Bureau of Labor Statistics attributes about one third of the quarter-to-quarter matches that are not directly linked through firm identifiers to each of (a) the use of the administrative follow-up described in the preceding paragraph, (b) probabilistic matching (c) clerical review of otherwise unmatched records (Pivetz, Searson, and Spletzer, 2001; Clayton and Spletzer, 2004). Davis, Haltiwanger, and Schuh (1996), and Abowd, Corbel, and Kramarz (1999), among others, have also used probabilistic name matching in research using US and French data, respectively.

3.2.7 Correcting by sample selection

The fundamental problem with not correcting for administrative ID changes remains pollution of statistics based on changes in employers. In order not to misclassify the disappearance of administrative numbers as plant closings, most researchers in this book only include plants that

existed in two consecutive years when studying changes. Thus, the calculated exit rates will not include plant closings, but will also not include administrative ID changes misclassified as plant closings. To the extent that misclassified plant closings bias the statistics upwards, the usual bias described earlier is eliminated. However, to the extent that the earnings changes of workers that are part of true plant closings differ from workers separating for other reasons, a new bias is introduced.

3.2.8 Using worker flows to correct for firm identifier changes

Most of the merged employer-employee data used in all these papers allow for a further solution to the problem. If workers can be followed from one employer to the next, then worker flows can be used to identify firms that are economically identical despite changing administrative identifiers. At the extreme, if all workers of firm A simultaneously "separate", to then be collectively hired by firm B, where they constitute the totality of employment, then firms A and B are very likely to be the same firm having changed administrative identifiers. More generally, in order for a firm B to be the economic successor of firm A, some fraction f(A) of workers leaving firm A must be linkable to firm B, and possibly some fraction f(B) of workers at firm B must have come from firm A. How to set the cutoff levels f(A) and f(B) is the subject of academic discussion, and no clear consensus arises.

This solution helps address more clearly the problem of economic successor versus legal successor, mentioned above, but conditional on the cutoff levels chosen. Among the data sets used in this book, Denmark, Italy, Finland, and the United States are known to apply such mechanisms. The cases of Finland, Italy and the United States are described in more detail below, but Table 2 describes how each of these countries handles linking firms based on the cutoff levels just described.

[Table 2 about here]

Several researchers have also linked firms longitudinally into time-coherent "hiring entities." Because the technique is richer than the simple longitudinal linking which will be described here, it can also be used to identify changes in firm relationships such as mergers, acquisitions, and outsourcing. Using worker flows to identify predecessor-successor links has been used in Italian data (Revelli 1996, Contini 2002), French data, Swedish data, Finnish data (Vartiainen 2004) and US data (Benedetto et al. 2003). We will discuss some of these approaches and their results in the next sections.

3.2.9 Application: Using worker flows to link firms in Italy

Among the data sets presented in this book, the Danish, Finnish, and Italian data have implemented this approach (the US data set used in this book pre-dates the implementation of the worker flow method at the U.S. Census Bureau). The Italian WHIP data, which is a 1 in 90 extract of the underlying universe, uses weighting and a cutoff in absolute numbers to define flow-based links. The fundamental hypothesis is that it is unlikely to observe large numbers of workers simultaneously and within a short period of time (one month) moving between two different firms. Since each record in the WHIP worker file represents 90 workers, observing two workers move between firms in the WHIP extract is equivalent to the movement of approximately 180 workers. Such an event is defined a link, and all movements in preceding and succeeding months between the firms linked in this way are classified as spurious movements. About 3.4% of all job spells have been corrected according to the spurious movements identifier.

3.2.10 Case study: Firm identifier changes in the USA and the concept of the firm (Benedetto, Haltiwanger, Lane, and McKinney, 2003)

This section draws on Benedetto, Haltiwanger, Lane, and McKinney (2003, henceforth BHLM). For 18 states, some of which are also used in the U.S. chapter of this book, BHLM track all movements between firms with more than 5 employees at the time of the movement, between 1992 and 2001. About 2.5 billion such movements are observed. Four conditions are defined. Two characterize the life-cycle of predecessor and successor, two characterize the movements between predecessors and successors. The link quality is defined on how many of these conditions are met. A predecessor-based link is of the highest rank if (1) the predecessor exits within two quarters of the movement that defines the link and (2) 80% of the predecessor's pre-link employees move to the successor. Not meeting one or the other condition reduces the quality rank attributed to the link. Equivalent conditions characterize the successor link. BHLM use these link variables to identify successor-predecessor relationships related to a change in administrative ID, merger-acquisition activity, and the presence of outsourcing. The relation type of relevance to the analyses presented in this book is the "ID change" relation, depends only on the second condition for both predecessor and successor based links. Thus, an ID change occurs when at least 80% of a predecessor's employees move to a successor, where they constitute at least 80% of the successor's employees (BHLM, Table 2).

Events characterized as "ID changes" account for about 12% of all events that meet at least one of the conditions (BHLM, Table 3). More importantly, movements of this type account for 1-2% of all accessions in the data. It can be speculated that this symmetrically holds for separations as well. BHLM also find significant number of smaller clusters moving between firms. Such movements

can be due to small portions, possibly individual establishments, being transferred between firms, or for movements of workers across divisions of a firm that appears under multiple identifiers. Additional linkage to the Census Business Register, which allows the identification of more complex firm relationships, indicates that about a fifth of all ID changes occur within the same firm.

Using worker movements to identify predecessor-successor relationships is not the only way to establish such links. BHLM compare the worker flow-based links with information present on the ES202 data establishing such links from administrative information. Among "ID changes", more than half of all link events prior to 1998, and approximately half of link events after 1998 are not identified in the administratively defined links. Independent research by other researchers at the U.S. Census Bureau has shown that some of the links defined administratively do not have a corresponding flow.

Overall, the research reported in BHLM highlights that using flow-based links as well as administrative information is an important element in accurately defining flows in US data. In the absence of such controls, the bias in separation rates can be as high as 2.5% in state-level aggregates.

3.2.11 Case study: Firm identifier changes in the USA, zero employment, and establishment turnover (Spletzer, 2000)

Spletzer (2000) used establishment-level data from West Virginia to look at the contribution of establishment turnover – births and deaths – to total employment growth. The context required him to accurately measure establishment births and deaths. He contrasted two definitions of an establishments "birth." The first definition categorizes the administrative birth of an establishment by the appearance of its administrative identifier in the data, and the absence of any administratively captured predecessor. The second definition defines the economic birth of an establishment by the first quarter with strictly positive employment. An equivalent set of definitions is made for establishment deaths. Spletzer (2000) finds that 23% of establishments have zero employment at their administrative birth, and 76% have zero employment at their administrative death. Thus, although the firms may already or still exist legally, no economic activity involving employees is occurring. In Spletzer's analysis, this matters. Using administrative birth and death would bias downwards the contribution of establishment births and deaths to employment growth.

3.2.12 Application: Firm identifier changes in Finland, flow-based identifiers, and worker separation rates (Vartiainen, 2004)

Vartiainen (2004) describes the impact of using flow-based identifier correction on the computation of worker separation rates from firms and establishments in Finish data. In a first step, flow-analysis was used to help in reverse-engineering identifier code changes (see Section 3.2.4). One particularly problematic problem was the change in the coding systems between 1989 and 1990. Although about two-thirds of all firms as identified by their 1990 identifier have a clear flow of workers from a single firm as identified by the 1989 firm identifier, and could thus be readily classified as being the same firm, enough problematic firms, both with multiple predecessors and multiple successors, remain. These problematic firms are likely true mergers or firm splits commingled with the identifier change. As a result, the Finish authors in this book decided not to report exit rates for 1989-1990, since such data would have been too unreliable.

For the remainder of the Finish employer organization data, Vartiainen analyzes the impact of two different firm and establishment identification strategies on separation rates. The administrative identify of a firm is defined as a unique identifier in the data. In the sample of worker records, a worker is recorded as having separated from a firm in year *Y*, and thus contributing to the separation rate, if the code of his or her employer in year *Y*+1 is different than the employer code in year *Y*, or if the worker no longer appears in the data in the following year. Yearly separation rates based on this criterion are tabulated in Column (1) of Table 3, adapted from Vartiainen (2004).

[Table 3 about here]

The flow-based identity of a firm is established by considering the movements of groups of workers. A link between two firms is established if at least 80% of ABC's worker in *Y* re-appear at a single firm DEF in year *Y+1*, and constitute 80% of DEF's employment in year *Y+1*. Note that DEF might or might not be called ABC – the flow criterion ignores the actual administrative identifier code. A worker *I* at firm ABC is considered a stayer if he then also is observed working for DEF in year *Y+1*. All other workers are considered to have separated. Separation rates using only the flow-based criterion are tabulated in Column (2) of Table 3.

The difference between the two columns varies between 1 and 10 percentage points. For the bulk of separating workers, whether administrative or economic entities are tracked is irrelevant. However, for a significant fraction of workers, it does matter. The reason for the difference is broken out in Columns (3) and (4). Column (3) tabulates the portion of Column (2) that is due to a worker being qualified as a stayer by the flow criterion, but as a separating worker by the administrative criterion. A worker is observed changing identifiers between two years, but moves with over 80% of his old and new colleagues to the new identifier. This may be a pure administrative code change, or it could

be a large spin-off or de-merger. It turns out that only a small portion of the separation rate is due to such movements.

Column (4) considers the portion of the separation rate in Column (2) that is due to workers being classified as stayers by the administrative criterion, but exiters by the flow criterion. Such a situation may arise when a large layoff, affecting over 20% of a firm's workforce, occurs. By the flow criterion, no successor firm exists, since no group, including the surviving workforce, qualifies for the double-80% criterion. Thus, by the flow criterion, such a firm died. The successor or survivor by the administrative criterion is a new firm by the flow criterion, and all workers, whether part of the layoff or part of the surviving workforce, are classified as separators. This situation accounts for almost the entire difference between Columns (1) and (2).

Clearly, the situation captured by Column (4) is not necessarily the desired outcome, since most analysts would consider the administratively surviving firm to be legitimately in continuous existence. Column (5) thus adopts the following strategy. A firm is in continuous existence if a continuous administrative identifier exists (administrative criterion). If a firm death occurs by the administrative criterion, but a successor entity exists by the flow criterion, then the firm is still in continuous existence. Only if no administrative and no flow successor can be found does a firm cease to exist. A worker is only counted as a exiter if leaving a firm for an administrative entity that is not a successor either by the administrative or by the flow criterion. In essence, Column (5) is obtained by combining Columns (1) and (3). Since the Finnish data seems to track administrative successors quite well by maintaining a single identifier throughout time, the difference between a purely administratively based "firm death criterion" and one moderated by worker flows is insignificant. One can conclude that in the Finnish data, the administrative codes seem to track the economic entity quite accurately.

3.3 Crossing borders and boundaries: The concept of a firm again

In most of the data sets used here, the firm and person identifiers are national identity numbers. This defines a particular concept of a firm. Both finer and broader definitions of a "firm" typically exist, but are invisible in this data. For instance, most data sets do not allow the connection of firms in a parent company – subsidiary relationship.

Exceptions, however, appear even here. The Swedish SAF data has person and firm identifiers that are internal to each of the two distinct data sets (blue and white-collar) it encompasses. Thus, a worker can be followed within the sample of firms reporting data for blue-collar workers even when they switch firms. But a worker switching from a blue-collar occupation to a white-collar occupation within the same firm will appear as an exit from the blue-collar sample and an accession

to the white-collar sample. Neither the firm nor the worker can be linked between the blue and white-collar samples.

In the US data, firms are represented by a state-specific account number within the state unemployment insurance system. Thus, although workers can be traced across state lines, firms cannot be linked across states using the data from the unemployment insurance system (it is feasible to do this using Census-internal data links). A worker transferring from one unit in state A to another unit owned by the same "company" in a different state will be identified as a separation.

Again, as before, the interpretation of certain statistics depends on the granularity of the entity defnition, *i.e.*, whether a firm or an establishment is the basic unit of accounting on the employer side. Intra-firm transfers between establishments show up in countries that are able to pinpoint employment to an establishment, but are hidden in data that can only identify worker movements at the firm level. Thus, turnover statistics – separation and accession rates – will appear higher in establishment data than in firm-level data.

4 Missing data and related issues

4.1 Lost records and unavailable data

Data captured by most data sources goes back up to three decades. Inevitably, computer systems are no longer the same today as they were at the start of the data collection period. The same applies to the legal environment in which data is collected.

One manifestation of the changing environment is that in many cases, certain portions of the data are no longer available today for reasons outside of the control even of the data collectors. Norway, France, and others all had to face this problem, and there are as many solutions as there are problems.

4.1.1 Application: Tackling unavailable data in France

The French DADS (Déclaration Annuelles de Données Sociales) was not collected in some years surrounding the 1982 and 1990 Censuses. As a result, data for 1981, 1983, and 1990 were missing. Data were imputed in the same way as in the case of the person identifier miscoding described in Section 3.1.4, thus adding 759,017 observations to the data, equivalent to approximately 4.7% of all records.

4.1.2 Application: Tackling unavailable data in Norway

In the Norwegian NHO data, the year 1987 is no longer available. However, all years of the NHO data contain lagged values, and so most of the 1987 data can be reconstructed from 1988 data. The only records that cannot be reconstructed are those for workers who left the data in 1987, and for which no lagged values are available

4.2 Censoring issues

4.2.1 Earnings

For the most part, the data used in this book are collected not to produce data for researchers, but to administer a government program, to collect payroll taxes or unemployment insurance contributions from firms, or to compute income taxes for workers. In particular the insurance contributions often have a top-code, beyond which contributions are no longer computed. The data then only record that top-code value, rather than the true income or wage earned by that worker.

4.2.2 Application: Correcting for right-censored earnings in German data

In the German IAB data, gross monthly earnings are censored at a time-varying threshold defined by the limitation of payments into the social security system. The following procedure was used to circumvent the censoring problem. A Mincerian earnings equation is estimated, including sector and occupation dummies. From the parameters of this regression, predicted earnings are computed and replace the top-coded values. Across time, between ten and fifteen percent of all observations are imputed, but within some more narrowly defined demographic groups, this percentage increases dramatically. Among workers with a university degree, about 50% of all observations are found to be censored.

4.2.3 Tenure

Although the most frequent censoring issues affect earnings, other variables can be incompletely recorded as well. In particular the tenure variable suffers from such problems. In most data sets, tenure is computed as the number of years an individual appears in the data, starting at the earliest date. However, if an individual is present in the first year of the data, his or her actual start date is unknown, and thus tenure is censored at a point that varies by individual.

4.2.4 Application: Correcting for left-censored tenure in French data

Individuals for whom the first year of observation was in 1976, the start of the data set, and who had worked 360 days in that year, the actual start date is unknown. Abowd, Kramarz, and Margolis

(1999, AKM hereafter) estimated the expected length of the in-progress employment spell by regression analysis using a supplementary survey, the 1978 Enquête sur la Structure des Salaires (ESS, Salary Structure Survey). In this survey, respondent establishments provided information on seniority, occupation, date of birth, industry, and work location for a scientific sample of their employees. Separate regressions were used for men and women. The coefficients were then used to predict seniority for the in-progress spells in 1976 with 360 days worked.

5 Conclusion

In this chapter, we have taken a look at the data underlying the other chapters in this book, with an eye for the adjustments that needed to be made in order to make the data both usable and comparable. Each administrative data set, in each state, country, or other organization, has its particularities, including differences in coverage, basic definitions of entities, and data quality. These differences can have a significant impact on the comparability of results obtained from such data. Precisely because the data collection is administrative in nature, and beyond the control of most researchers, any attempt to make the actual data collection comparable across countries is bound to fail. An exception to this rule is the collection of administrative surveys coordinated by Eurostat ("Structure of Earnings Survey"), the Belgian portion of which was used in the chapter by Thierry Lallemand, Robert Plasman and François Rycx. Such specially administered surveys are costly to produce and coordinate. To wit, the "Structure of Earnings Survey," while providing comprehensive cross-sectional coverage, is administered only every four years, and releases can take up to 3 years to become available to the public.

Researchers accessing the longer time series of conventional administrative data thus need to take extra steps in order to make the data meaningful for analysis, and comparable to the data used by others. For the data used in this book, this chapter has outlined their methods, and provided, where available, the results of comparing the data used in this book to data produced using alternate scenarios and processing methods. The reader of this chapter should take away a better appreciation of the methods needed to make the data comparable across so many countries, and the reassurance that the data can be combined and compared in meaningful ways because of the application and use of these methods.

6 Bibliography

Abowd, John M. and Arnold Zellner (1985) "Estimating Gross Labor Force Flows," *Journal of Business and Economics Statistics*, 3, pp. 254-283.

Abowd, John M. and Corbel, Patrick and Kramarz, Francis (1999) “The Entry and Exit of Workers and the Growth of Employment: An Analysis of French Establishments”, Review of Economics and Statistics, 81(2) pp. 170-87.

Abowd, John and Francis Kramarz (1999) ”The Analysis of Labor Markets Using Matched Employer-Employee Data”, in: Orley Ashenfelter and David Card (eds), *Handbook of labor economics, Vol. 3*, North-Holland, Amsterdam, New York, chapter 40, pp. 2629-2710

Abowd, John M. and Lars Vilhuber (2005) ”The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers,” *Journal of Business and Economics Statistics*.

Anderson, Patricia and Meyer, Bruce (2000), “The Effects of the Unemployment Insurance Payroll Tax on Wages, Employment, Claims and Denials”, *Journal of Public Economics*, v78, n1-2 (October 2000): 81-106

Baker, George and Michael Gibbs and Bengt Holmstrom (1994) “The Internal Economics of the Firm: Evidence from Personnel Data,” *Quarterly Journal of Economics*, 109(4), pp. 881-919

Bender, Stefan, and Christian Dustmann, David Margolis and Costas Meghir (2002), “Worker Displacement in France and Germany,” in Peter Kuhn and Randal Eberts (eds.), *Losing Work, Moving On: International Comparisons of Worker Displacement*, Upjohn Institute, Kalamazoo, MI.

Benedetto, Gary and John Haltiwanger and Julia Lane and Kevin McKinney (2003), “Using Worker Flows in the Analysis of the Firm,” LEHD, U.S. Census Bureau, Technical paper TP-2003-09.

Bureau of Labor Statistics (1997), “Quality Improvement Project: Unemployment Insurance Wage Records”, U.S. Department of Labor, report.

Clayton and Spletzer (2004), Presentation made at the NBER Summer Institute.

Contini, Bruno (2002) (Edited by) *Labour mobility and wage dynamics in Italy,* Rosenberg and Sellier, Torino

Contini, Bruno and Claudia Villosio (2003) ” Worker mobility, job displacement and wage dynamics: Italy 1985-91”, LABORatorio Revelli Working Paper, n. 24

Davis, Steven J. and John C. Haltiwanger and Scott Schuh (1996) *Job creation and destruction*, MIT Press,Cambridge, MA.

Fienberg, Stephen E. and Stasny, Elizabeth A. (1983) “Estimating monthly gross flows in labour force participation”, *Survey Methodology*, 9, pp. 77-102.

Jacobson, Louis S. and Robert J. LaLonde and Daniel G. Sullivan (1993) “Earnings Losses of Displaced Workers,” American Economic Review 83(4), pp. 685-709.

Lazear, Edward P. (1995) *Personnel Economics*, MIT Press, Cambridge, MA.

Leombruni, Roberto and Roberto Quaranta (2002), “The Unemployment Route to Versatility”, LABORatorio Revelli Working Paper, n. 16

Neal, Derek (1995), “Industry-Specific Human Capital: Evidence from Displaced Workers,” *Journal of Labor Economics*, 13 (4), pp. 653-77.

Parent, Daniel (2000), “Industry-Specific Capital and the Wage Profile: Evidence from the National Longitudinal Survey of Youth and the Panel Study of Income Dynamics”, *Journal of Labor Economics* 18(2), pp. 306-23

Pivetz, Timothy R. and Searson, Michael A. and Spletzer, James R. (2001) “Measuring job and establishment flows with BLS longitudinal microdata,” *Monthly Labor Review* 124(4), pp. 13-20

Poterba, James M. and Lawrence H. Summers (1986) “Reporting Errors and Labor Market Dynamics,” *Econometrica* 54(6), pp. 1319-1338.

Revelli R. (1996), "Statistics on Job Creation: Issues in the Use of Administrative Data", in OECD *Job Creation and Loss. Analysis, Policy, and Data Development*, OECD, Paris

Spletzer, James R. (2000) “The Contribution Of Establishment Births And Deaths To Employment Growth,” *Journal of Business and Economic Statistics* 18, pp. 113-26.

Stevens, David (2002) “Employment that is not covered by state unemployment” LEHD Technical Paper No. TP-2002-16.

Vartiainen, Juhana (2004) “Measuring interfirm mobility with an administrative data set”, mimeo, November 2004.