

This PDF is a selection from a published volume from the National Bureau of Economic Research

Volume Title: The Structure of Wages: An International Comparison

Volume Author/Editor: Edward P. Lazear and Kathryn L. Shaw, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-47050-4; 978-0-226-47050-4

Volume URL: <http://www.nber.org/books/laze08-1>

Conference Date:

Publication Date: January 2009

Chapter Title: Adjusting Imperfect Data: Overview and Case Studies

Chapter Author: Lars Vilhuber

Chapter URL: <http://www.nber.org/chapters/c2366>

Chapter pages in book: (p. 59 - 80)

Adjusting Imperfect Data

Overview and Case Studies

Lars Vilhuber

1.1 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 uses for the data contained in these systems: from tax administration in countries such as France and Sweden, to calculation of unemployment benefits in the United States, to administration of centralized bargaining agreements in Norway and Sweden. The size of the countries and, thus, the data sets, ranges 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 coincidentally 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 data sets and handled them in what their experience told them was the best manner, given the constraints both of their data and the

Lars Vilhuber is a senior research associate at the Cornell University School of Industrial and Labor Relations (ILR), and a senior research associate in the Longitudinal Employer-Household Dynamics program at the U.S. Census Bureau.

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 preceding, John Abowd, and Julia Lane for comments on drafts of this text. All errors, of course, remain mine.

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 are 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. However, although the starting points vary widely, the end result of the authors' efforts is a very high degree of comparability of the analytical data sets across countries, as evidenced by the different chapters in this book.

In this chapter, instead of using the similarity in the cleaned data sets to investigate economic fundamentals, we focus on the steps undertaken by each chapter's authors and their respective data providers to render the data both usable and comparable. We describe two data elements that remain fundamentally different across countries—the sampling or data collection methodology and the basic unit of analysis (establishment or firm)—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.

1.2 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 data sets.

1.2.1 Sampling Schemes

The data sets used in this book were derived from administrative sources that essentially have three “sampling schemes”: worker-and-firm universe files, worker-based samples, and firm-based samples (see tables I.1 and I.2 in the introduction to this book, by Edward P. Lazear and Kathryn L. Shaw, for more details). In some cases, the sampling scheme is the result of the underlying fundamental structure of the administrative data upon which the data set is based; in other cases, researchers’ access mode is restricted to data sets derived using the sampling scheme observed.

Worker-and-firm universe files are not samples, though there may be some smaller coverage issues. Workers and firms appear in the data set because they are covered by a universal entitlement or tax system. The Danish Center for Corporate Performance (CCP) data, the Swedish RAMS data, and the U.S. Longitudinal Employer-Household Dynamics (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 U.S. data are 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; Swedish Employer’s Federation] data set) 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 more closely 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 subsample 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 *Déclaration Annuelles de Données Sociales* (DADS) and the Italian work histories (WHIP) data are worker-based samples. For some

fraction of workers, all jobs with all employers are tracked. Whenever a worker changes firms, his or her move to another employer is included in the database, 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, instead of using firms as the basic unit of analysis for within-unit wage levels and wage changes, the Italy chapter chose to compute these statistics on cells, where cells are defined by unique combinations of four geographic areas by six firm size classes by eight industry sectors. The technical appendix to the Italy 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, nonemployment 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 ten-person firm, and one-in-twenty-five worker sample. Naive 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—that is, 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, both of which use firm-based 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.

1.2.2 Aggregation Levels and the Concept of an Employer

The data sets 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 data sets may allow for higher-level aggregations. Furthermore, because some of the data are merged from different sources, not all the detail may be available at the lowest aggregation level. Table I.1 in the introduction to this book, by Edward P. Lazear and Kathryn L. Shaw, 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 following discussion.

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 data sets can be recombined by firm, this is not feasible in the Swedish SAF data, and is 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 may be 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 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.

1.3 Longitudinal Linking (Identifier Issues)

1.3.1 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.

When Do Coding Errors Occur?

Coding errors occur for a variety of reasons. A survey of fifty-three state employment security agencies in the United States over the 1996 to 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 1997, ii). The report noted that 38 percent of all records were entered by key entry, while another 11 percent were read in by optical character readers (OCRs). The 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 United States, are known not to flow back into the actual wage record database. The Italian data typically do 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).

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 lead to an upward bias in flow statistics, but result in a downward bias in tenure measures.

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 Current Population Survey (CPS; Fienberg and Stasny 1983; Abowd and 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 unemployment insurance (UI) wage records are potentially biased.

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 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 databases 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.

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 keypunch 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, most likely December 31) and also employed after some date in 1980 (depending on the number of days worked, most likely January 1) in this firm but not employed at all during year 1979. This is very improbable. The National Institute for Statistics and Economic Studies (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 fol-

lowing 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 percent of all records.

Case Study: The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers in the United States (Abowd and Vilhuber 2005)

Abowd and Vilhuber (2005) describe the method used by the LEHD Program at 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 percent of all wage records. Trials in the late 1980s using UI wage records found an average error rate of 7.8 percent, with significant variation across states (Bureau of Labor Statistics 1997). 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 percent of all jobs, but 11 percent 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 percent, and recalls are biased upward by nearly 6 percent. Payroll for accessions and separations are biased upward by up to 7 percent.

1.3.2 Quality of Firm Links, Measures Undertaken and Not Undertaken to Improve 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 the time series for the affected entity. However,

several factors combine to make this a problem both less pervasive and more difficult to correct for in firm-level data.

*Administrative versus 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 chapters in this book, the fundamental focus is on firms as places of work for workers, that is, 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 that, 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.

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 Holmstrom 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 upward by a significant margin. For instance, Jacobson, LaLonde, and Sullivan (1993a) estimate that in Pennsylvania administrative data for the 1980s, such “bogus job changes” would have been the primary source of all worker movements between employers within industries (Jacobson, LaLonde, and Sullivan 1993a, 39).

Other research would also be affected. An extensive literature shows that a large fraction of workers who are part of a mass layoff have some difficulty reentering the labor market, showing significantly negative effects on the earnings history (Jacobson, Lalonde and Sullivan 1993b) 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 because, 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 upward (Spletzer 2000; Benedetto et al. 2007; Vartiainen 2004).

Methods of Linking Firms

To counter the linkage problem identified in the preceding, a variety of methods are used. All corrective methods reduce the linkage failure to varying degrees, but it matters who performs the corrective measure—the researcher or the administrative. Because their motives differ, the goals achieved by corrective measures will also vary.

Following up with firms. Administrations are interested in linking firms for reasons other than longitudinal consistency of research data sets. In particular in the United States, payroll taxes are often experience rated; that is, the tax rate increases as firms lay off large numbers of workers. 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. To prevent this form of tax evasion, administrations follow up on firms, and the U.S. 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, that is, 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 administra-

tive 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).

The method relies on accurate information on the firms involved being available and (typically) substantial expenses for clerical review. In particular in the United States, where the data are administered by individual states, the success rate of such linkages will vary across states. It is likely to work better in smaller states and for larger firms, where information about individual firms is more easily obtained. It is known to be used by the Finnish and Italian data sources used in this book.

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).

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, and (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 U.S. and French data, respectively. In this book, the French data incorporate knowledge gained from the previous probabilistic name-matching exercise.

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 upward, 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.

Using worker flows to correct for firm identifier changes. Most of the merged employer-employee data used in all these chapters 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 considered the economic successor of firm A, most researchers require that at least one of the following two conditions holds: (1) A large fraction $f(A)$ of workers leaving firm A can be traced to firm B, and (2) a large fraction $f(B)$ of workers at firm B must have come from firm A. How high to set the cutoff levels $f(A)$ and $f(B)$ is the subject of academic discussion, and no clear consensus arises.

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 in the following, but table 1.1 describes how each of these countries handles linking firms in terms of the cutoff levels $f(A)$ and $f(B)$ just described.

Several researchers have also linked firms longitudinally into time-coherent “hiring entities.” Because the technique is richer than the simple longitudinal linking that 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 U.S. data (Benedetto

Table 1.1 Worker flows between firms, as a percentage of firm employment required to link firms

Country	Predecessor cutoff level: No. of workers leaving predecessor firm and moving to the same successor firm, as fraction of employment level (firm A)	Successor cutoff level: No. of workers coming from predecessor firm and newly hired at successor firm, as fraction of successor's employment level (firm B)
Denmark (CCP)	50%	any
Finland	75%	any
United States (LEHD)	80%	80%
Italy (WHIP)	~100 workers	any
France (DADS)	n.a.	n.a.
Germany (IAB Panel)	n.a.	n.a.
Sweden (SAF)	n.a.	n.a.
Norway (NHO)	n.a.	n.a.
Belgium	n.a.	n.a.
Sweden (RAMS)	n.a.	n.a.

Notes: Firms A and B are linked (firm B is classified as a successor of firm A) if flows of workers between A and B satisfy both the predecessor cutoff level and the successor cutoff level criteria. The predecessor cutoff level is the flow of workers between the two firms, expressed as a fraction of firm A's employment before the flow occurs. The successor cutoff level is the flow of workers between the two firms, expressed as a fraction of B's employment after the flows occurred. Example: 85 workers move from A to B in the United States. Prior to the move, firm A has 100 workers. After the move, firm B has 170 workers. The flow of workers satisfies the predecessor cutoff level, but does not surpass the successor cutoff level. It is not classified as a link between firms A and B. If the same movement were to occur in Finland, it would be classified as a link. n.a. = not available.

et al. 2007). We will discuss some of these approaches and their results in the next sections.

Application: Using Worker Flows to Link Firms

Among the data sets presented in this book, the Danish, Finnish, and Italian data have used worker flows to link firms. In this section, we describe results for these three data sets and summarize the detailed analysis performed by Benedetto et al. (2007). We point out that the U.S. data set used in this book predates the implementation of the worker flow method at the U.S. Census Bureau.

The Italian Case

The Italian WHIP data, which is a one-in-ninety 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. Because each record in the WHIP worker file represents ninety 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 percent of all job spells have been corrected according to the spurious movements identifier.

Case Study: Firm Identifier Changes in the United States and the Concept of the Firm (Benedetto, Haltiwanger, Lane, and McKinney 2007)

This section draws on Benedetto, Haltiwanger, Lane, and McKinney 2007 (Benedetto et al. 2007). For eighteen states, some of which are also used in the U.S. chapter of this book, Benedetto et al. (2007) track all movements between firms with more than five 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 percent of the predecessor's prelink 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. Benedetto et al. (2007) 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, which depends only on the second condition for both predecessor- and successor-based links. Thus, an ID change occurs when at least 80 percent of a predecessor's employees move to a successor, where they constitute at least 80 percent of the successor's employees (Benedetto et al. 2007, table 2).

Events characterized as "ID changes" account for about 12 percent of all events that meet at least one of the conditions (Benedetto et al. 2007, table 3). More important, movements of this type account for 1 to 2 percent of all accessions in the data. It can be speculated that this symmetrically holds for separations as well. Benedetto et al. (2007) 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 of firms is not the only way to establish such links. Benedetto et al. (2007) 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 Benedetto et al. (2007) highlights that using flow-based links as well as administrative information is an important element in accurately defining flows in U.S. data. In the absence of such controls, the bias in separation rates can be as high as 2.5 percent in state-level aggregates.

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 Finnish data. In a first step, flow analysis was used to help in reverse engineering identifier code changes (see the section “Application: Using Worker Flow to Link Firms.”) 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 Finnish authors in this book decided not to report exit rates for 1989 to 1990 as such data would have been too unreliable.

For the remainder of the Finnish 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 1.2, adapted from Vartiainen (2004).

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 percent of ABC’s workers in Y reappear at a single firm DEF in year $Y + 1$, and constitute 80 percent 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 or she then also is observed working for DEF in year $Y + 1$. All other workers are considered to have separated. Separation

Table 1.2 Separation rates, based on administrative vs. flow-based firm linkage in Finnish data (%)

Year	Administrative criterion (1)	Pure flow-based (2)	Stayers by flow, exiter by code (3)	Exiter by flow, stayer by code (4)	Administrative criterion + flow criterion (5)
1980	16	17	-0.01	1.46	16
1981	15	17	0.00	1.94	15
1982	16	19	0.00	2.88	16
1983	17	21	0.00	3.82	17
1984	18	20	-0.01	2.37	18
1985	19	25	0.00	6.37	19
1986	23	33	-0.01	9.77	23
1987	22	29	0.00	6.44	22
1988	24	32	-0.56	8.29	24
1989	n.a.	n.a.			
1990	26	29	-0.26	3.07	26
1991	23	26	-0.38	3.64	23
1992	33	37	-0.38	4.20	33
1993	21	25	-0.13	4.52	21
1994	22	26	-1.82	5.51	22
1995	21	23	-0.06	1.81	21
1996	16	17	-0.04	1.55	16
1997	15	15	-0.34	0.82	15
1998	19	20	-0.60	1.83	19
1999	18	19	-1.26	2.31	18
2000	24	27	-0.22	3.74	24
2001	22	24	-0.04	2.08	22

Notes: Adapted from Vartiainen (2004), workers separating from his or her employer in Year + 1, based on different definitions of firm identifiers. n.a. = not available. Blank cells indicate data not available.

rates using only the flow-based criterion are tabulated in column (2) of table 1.2.

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 percent of his or her 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 demerger. 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 percent of a firm's workforce, occurs. By the flow criterion, no successor firm exists because no group, including the surviving workforce, qualifies for the double-80 percent 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 columns (column 4) is not necessarily the desired outcome as 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 an 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). Because the Finnish data seem 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.

1.3.3 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-subsidary 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 U.S. 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 definition, that is, whether a firm or an establishment is the basic unit of accounting on the employer side. Intrafirm 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.

1.4 Missing Data and Related Issues

Data captured by most data sources go 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 are 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.

1.4.1 Application: Tackling Unavailable Data in France

The French DADS 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 the section “Application: Imputation to Correct for Coding Errors in France,” thus adding 759,017 observations to the data, equivalent to approximately 4.7 percent of all records.

1.4.2 Application: Tackling Unavailable Data in Norway

In the Norwegian Confederation of Norwegian Business and Industry (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

1.4.3 Censoring Issues

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.

In the German Institute for Employment Research (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 10 and 15 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 percent of all observations are found to be censored.

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.

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) 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 (impute) seniority for the in-progress spells in 1976 with 360 days worked. The procedure has been used for most subsequent research papers using the French data, in particular the research reported in this book's chapter on France. The impact of the procedure greatly improves estimates requiring knowledge of the full tenure of workers.

1.5 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 three 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.

References

- Abowd, John M., Patrick Corbel, and Francis Kramarz. 1999. The entry and exit of workers and the growth of employment: An analysis of French establishments. *Review of Economics and Statistics* 81 (2): 170–87.
- Abowd, John M., and Francis Kramarz. 1999. The analysis of labor markets using matched employer-employee data. In *Handbook of labor economics*. Vol. 3, ed. Orley Ashenfelter and David Card, 2629–2710. Amsterdam: North-Holland.
- Abowd, John M., Francis Kramarz, and David Margolis. 1999. High-wage workers and high-wage firms. *Econometrica* 67 (2): 251–333.
- Abowd, John M., and Lars Vilhuber. 2005. The sensitivity of economic statistics to coding errors in personal identifiers. *Journal of Business and Economics Statistics* 23 (2): 133–52.
- Abowd, John M., and Arnold Zellner. 1985. Estimating gross labor force flows. *Journal of Business and Economics Statistics* 3:254–83.
- Baker, George, Michael Gibbs, and Bengt Holmstrom. 1994. The internal economics of the firm: Evidence from personnel data. *Quarterly Journal of Economics* 109 (4): 881–919.
- Bender, Stefan, Christian Dustmann, David Margolis, and Costas Meghir. 2002. Worker displacement in France and Germany. In *Losing work, moving on: International comparisons of worker displacement*, ed. Peter Kuhn and Randal Eberts, 375–470. Kalamazoo, MI: Upjohn Institute.

- Benedetto, Gary, John Haltiwanger, Julia Lane, and Kevin McKinney. 2007. Using worker flows in the analysis of the firm. *Journal of Business and Economic Statistics* 25 (3): 299–313.
- Bureau of Labor Statistics. 1997. *Quality improvement project: Unemployment insurance wage records*. Washington, DC: U.S. Department of Labor.
- Clayton, Richard, and James Spletzer. 2004. Presentation made at the 2004 NBER Summer Institute. July, Boston.
- Contini, Bruno, ed. 2002. *Labour mobility and wage dynamics in Italy*. Turin, Italy: Rosenberg.
- Contini, Bruno, and Claudia Villosio. 2003. Worker mobility, job displacement and wage dynamics: Italy 1985–91. LABORatorio Revelli Working Paper no. 24. Turin, Italy: LABORatorio R. Revelli.
- Davis, Steven J., John C. Haltiwanger, and Scott Schuh. 1996. *Job creation and destruction*. Cambridge, MA: MIT Press.
- Fienberg, Stephen E., and Elizabeth A. Stasny. 1983. Estimating monthly gross flows in labour force participation. *Survey Methodology* 9:77–102.
- Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan. 1993a. *The costs of worker Dislocation*. Kalamazoo, MI: W.E. Upjohn Institute.
- . 1993b. Earnings losses of displaced workers. *American Economic Review* 83 (4): 685–709.
- Lazear, Edward P. 1995. *Personnel economics*. Cambridge, MA: MIT Press.
- Leombruni, Roberto, and Roberto Quaranta. 2002. The unemployment route to versatility. LABORatorio Revelli Working Paper no. 16. Turin, Italy: LABORatorio R. Revelli.
- Margolis, David. 1999. Part-year employment, slow reemployment, and earnings losses: The case of worker displacement in France. In *The creation and analysis of employer-employee matched data*, ed. J. Haltiwanger, J. Lane, J. Spletzer, and J. Theeuwes, 375–416. Amsterdam: North-Holland.
- Neal, Derek. 1995. Industry-specific human capital: Evidence from displaced workers. *Journal of Labor Economics* 13 (4): 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): 306–23.
- Pivetz, Timothy R., Michael A. Searson, and James R. Spletzer. 2001. Measuring job and establishment flows with BLS longitudinal microdata. *Monthly Labor Review* 124 (4): 13–20.
- Poterba, James M., and Lawrence H. Summers. 1986. Reporting errors and labor market dynamics. *Econometrica* 54 (6): 1319–38.
- Revelli, Ricardo. 1996. Statistics on job creation: Issues in the use of administrative data. In *Job creation and loss: Analysis, policy, and data development*. Paris: Organization for Economic Cooperation and Development.
- Spletzer, James R. 2000. The contribution of establishment births and deaths to employment growth. *Journal of Business and Economic Statistics* 18:113–26.
- Stevens, David. 2002. Employment that is not covered by state unemployment. LEHD Technical Paper no. TP-2002-16. Washington, DC: U.S. Census Bureau, Longitudinal Employer-Household Dynamics.
- Vartiainen, Juhana. 2004. Measuring interfirm mobility with an administrative data set. National Institute for Economic Research. Mimeograph.