
LONGITUDINAL EMPLOYER - HOUSEHOLD DYNAMICS

TECHNICAL PAPER NO. TP-2002-17

The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers

Date : This version: January 2003
Prepared by : John M. Abowd and Lars Vilhuber
Contact : U.S. Census Bureau, LEHD Program
FB 2138-3
4700 Silver Hill Rd.
Suitland, MD 20233 USA

This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. [This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress.] This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging, and the Alfred P. Sloan Foundation. The views expressed herein are attributable only to the author(s) and do not represent the views of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau is preparing to support external researchers' use of these data; please contact U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA.

Abstract

In this paper, we describe the sensitivity of small-cell flow statistics to coding errors in the identity of the underlying entities. Specifically, we present results based on a comparison of the U.S. Census Bureau's Quarterly Workforce Indicators (QWI) before and after correcting for such errors in SSN-based identifiers in the underlying individual wage records. The correction used involves a novel application of existing statistical matching techniques. It is found that even a very conservative correction procedure has a sizable impact on the statistics. The average bias ranges from 0.25 percent up to 15 percent for flow statistics, and up to 5 percent for payroll aggregates.

KEYWORDS: Flow statistics, Probabilistic matching, Transitions, Tenure, Job flows, Job creation, QWI

1 Introduction

As governmental information technology systems have improved, measuring employment and job flows using administrative data has become an important tool for official statistics and social science research (Abowd, Lane & Prevest 2000). Administrative data have long been used to enhance the quality of official statistics and to maintain sampling frames. The Bureau of Labor Statistics uses the ES-202 data to maintain the sampling frame for the Current Employment Statistics (CES), and firm surveys derived therefrom (Bureau of Labor Statistics 1997a, chapter 2), and the U.S. Census Bureau uses administrative records for the initial sampling frame for the Economic Censuses (U.S. Census Bureau 2000, pg. 60). The potential biases in the estimation of counts, totals, and averages that arise from errors in the underlying administrative data are well understood (Little & Rubin 1990). Social science researchers using administrative data to measure employment flows have acknowledged that there are different biases in flow measures that arise from errors in the underlying data and have developed a variety of methods for addressing these problems (Anderson & Meyer 1994, Burgess, Lane & Stevens 2000, Davis, Haltiwanger & Schuh 1996, Jacobson, LaLonde & Sullivan 1993, Haltiwanger, Lane & Spletzer 1999, Lane, Miranda, Spletzer & Burgess 1999). In this paper we address one of the most important potential sources of bias in flow measures developed from administrative data; namely, the upward bias in transitions that results from errors in the period-to-period linking of the records. Such transitions are a crucial component of a new series of U.S. Census statistics, the Quarterly Workforce Indicators (QWI), and have also been used in other recent research (Bowlus & Vilhuber 2002).

The paper is organized as follows. Section 2 describes the data used for this paper, and documents the extent of the problem. Section 3 provides a detailed description of the theory and implementation of the probabilistic matching procedure used to correct errors in the data. The effect of the correction on a number of individual-level, firm-level, county-level and industry-level statistics is described in Section 4. Section 5 concludes.

2 Data and problem description

The UI database used here contains quarterly reports on earnings for all workers who worked for a covered employer¹ in the state of California between 1991 and 1999. The basic identifier for an “individual” is a

¹Exceptions of coverage are to be found in agriculture and among the self-employed. See Abowd, Lengermann & Vilhuber (2002) for some details.

SSN, but the definition of an SSN varies at each stage of processing, as it is modified along the way. Let $SSN(0)$ denote the SSN as on file before the start of processing, and $SSN(i)$ the SSN on the file after processing in Stage i . On one hand, before the editing process, the number of “individuals” as identified by the $SSN(0)$ is presumably larger than the actual number of human beings, due to the coding errors that this process is designed to address. On the other hand, improper use of SSNs by multiple workers would render the number of observed $SSN(0)$ s an underestimate of the actual number of workers. The process described here will be able to address coding errors, but will not be able to address the misuse issue.

A record is completed by information on first and last name as well as middle initial, and the actual earnings information. Processing relies on name information, and name information may and does vary from one record to another for the same $SSN(0)$. Define a unique identifying key (UID) as a unique combination of $SSN(0)$, first name, middle initial and last name. A single digit or letter difference will lead to a second entry among those combinations. Table 1 on page 32 describes the number of records per year, as well the number of unique identifying keys (UID). This is a measure of how homogeneous or how diverse name coding is on these files.

“Employers” or “firms” are identified on the basis of their UI account number, called a “State Employer Identification Number” (SEIN) here, which is used by the state to track their unemployment insurance tax payments.² Reports by firms are verified for consistency and compliance by the state agency, and a missing or substantially changed firm report is clerically investigated. As a result, we assume throughout that the SEIN does *not* suffer from coding issues.

The data set contains 57,393,771 UIDs associated with 28,431,008 unique $SSN(0)$ s. Note that a UID is associated with one and only one $SSN(0)$, but it will be potentially associated with a different $SSN(i)$ after Stage i of the probabilistic matching procedures.

2.1 Construction of job histories

We will refer to “employment histories” as the observed employment pattern of an individual (a $SSN(i)$) across all employers. A “job” is a match between an individual and an employer, the latter identified by SEIN. The history of such a match in all available quarters is termed a “job history”, which typically corresponds to the notion of employment tenure, although interruptions within a job history may be of substantial length.³ Interruptions are succinctly called “holes.”

²A single “firm” might have multiple SEINs.

³Job and employment histories are recomputed after every stage of processing. Here, we only compare job histories before the start and after the end of processing. An ap-

Coding errors occur for a variety of reasons. A survey of 53 state employment security agencies in 1996-1997 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. California, whose data was used in this paper, had one of the lowest rates in the use of key entry, relying more heavily on OCR and magnetic media, which tend to be less prone to errors. The types of errors will differ by the source of the error. When a record is manually transcribed by an employer onto a paper form, or scanned or entered by hand when entering the state agency's data warehouse, the most likely error to occur is a random 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 SSN in an employer's data system, which is then repeatedly transmitted to the state agency. Thus, to select potentially miscoded records, we use job histories to identify observed holes, and to identify short job histories which can serve to "plug" these holes. For reference, we also compute employment histories before the start of processing. Selection for matching occurs only based on job histories.

Table 2 on page 33 presents baseline patterns of job histories for the uncorrected data. The unit of observation is a job, potentially interrupted. For each such observation, the longest interruption is tabulated if there is one. If no interruption was observed during the worker's tenure with the employer, then the type of continuous job spell is tabulated. By definition, the absence of a hole implies continuous tenure, but that spell may have been ongoing in the first (*left-truncated*) or last (*right-truncated*) quarter of the data, or in both (*Entire period*). If the spell was continuous, with both the beginning and the end of the job spell observed within the data, then the default code of C is assigned.⁴

Most interruptions are short: holes of not more than one quarter account for nearly 41 percent of all interruptions. Furthermore, not reported in the table is the fact that 87 percent of those having interruptions of at most one quarter have only one interruption of that length. Given the quarterly frequency of reporting, many of these are likely to be caused by simple coding errors in the SSN. On the other hand, over 85 percent of all job spells are observed to be uninterrupted.⁵ The matching process described in this paper addresses the single-quarter

pendix available from the authors also contains intermediate job and employment histories.

⁴Of course, an interruption could be ongoing, and the worker return to the employer after the end of the data, or have come back from an interruption that started before the begin of the data. We ignore potential interruptions at the end points, and such patterns are counted towards continuous spells.

⁵Of course, interruptions of less than one full calendar quarter are unobservable in the data.

interruptions tabulated in Table 2.

Table 3 presents tabulations of the longest continuous employment spells with a given SEIN. It is well known that while most workers are in long employment relationships, most job spells are short, as shown by this table. On the other hand, short job spells, in particular those of exactly one quarter length, could well be due to coding error in SSN and/or SEIN. In this paper, the at-risk records to be matched to observed holes are the short, single-quarter spells, or plugs.

2.2 Previous results

Bureau of Labor Statistics (1997b) also presents results from a SSN validation project. Eight states sent a sample of wage records to the BLS, which then sent a 1 in 216 sample to Social Security Administration (SSA) for verification. Verification consisted of comparing the name on the submitted wage record to the name associated with the SSN on the SSA records. The overall error rate was 7.8 percent (Bureau of Labor Statistics 1997b, Table 3, pg.87), but varied substantially across states. Minnesota, with similar data collection methods as California (which was not included in the project), had an error rate of 4.7%.

The method proposed in this paper is both more extensive and less complete than the BLS/SSA validation project. It relies exclusively on information already present in the wage record data, but for a much longer period. Thus, although the procedure cannot verify that the name information associated with the name actually matches the record on the original SSN request, it can ascertain that the information is consistent across up to 9 years of wage record data. This implies that the most likely error to be corrected using this procedure is a random coding error, as would occur when a record is scanned or entered by hand when entering the state agency's data warehouse. The procedure will not be able to address errors that occur persistently over time, as would occur if the SSN on an employer's data system was mistyped when entered, and is repeatedly transmitted to the state agency.

The method in this paper is capable of addressing a much larger number of records. Whereas the BLS/SSA project only handled one in 216 records, with at the most 60,000 records for any given state, we have processed half a billion records. Finally, the matching procedure uses prior, contemporaneous, and future earnings information in the matching procedure, and thus complements procedures in place in some state agencies, that check quarter-to-quarter consistency of names.

3 Matching process

3.1 Concepts

Matching software is based on concepts developed by Newcombe, Kennedy, Axford & James (1959) and formalized by Fellegi & Sunter (1969). Concepts relating in the actual software implementation used in this research⁶ are described in Jaro (1989). An excellent overview of matching and probabilistic record linkage is provided elsewhere (Winkler 1993, Winkler 1999a, Winkler 1999b).

Probabilistic linkage⁷ of administrative records is distinct from statistical matching. In the latter, two unrelated datasets drawn from the same or similar populations are linked by common non-identifying variables, by combining records with the highest similarity. The probabilistic matching used in this research, on the other hand, combines records from datasets that contain a common identifier, although this identifier may contain errors, which need to be taken into account. Thus, using terminology consistent with Fellegi & Sunter (1969) (see also Winkler (1993)), such linkage is based on two files A and B . In their product space $A \times B$, a “match” is a pairing of records that represent the same persons, and a “nonmatch” is a pair of records that represent two different persons. When relying on a single file with product space $A \times A$, a “duplicate” is a record representing the same person as another record within the same file.

When these files are linked, a decision rule is implemented, separating all feasible pairs into links, possible links, and nonlinks. The decision rule is thus an attempt to classify pairs of records into the set of true matches M and the set of true non-matches U . Often, this occurs only within a restricted “block” of pairs, and not among the full product space $A \times B$. “Possible links” are those pairs where the decision rule is not sufficient to make a final decision, and a clerical review may follow. Two sets of errors can occur. First, “false matches” are nonmatches that are erroneously designated as links. Second, “false nonmatches” are matches that either are not designated as links within a set of pairs, or are not within the same block of pairs, and thus excluded from the scope of the decision rule.

In Fellegi-Sunter computer-based matching procedures, a decision rule is based on a matching weight, or score, assigned to each pair of records. Let x_Y^i denote the value of field x from a record i on file Y . With

⁶We used the commercially available program “Integrity Data Re-Engineering Environment - Automated Record Linkage System” (Anonymous 2000) from Vality Technology, Inc., now Ascential Software Corporation. It is based on earlier software by MatchWare Technologies, Inc. (Jaro 1997). The version used in this research is Release 3.6.9. See Appendix A for details.

⁷Fellegi & Sunter (1969) simply refer to “record linkage”, whereas Winkler (1993) calls it “exact” matching. Winkler (1999b) calls it “statistical data editing”.

slight abuse of notation, let

$$m_x = P(x_A^i = X_B^j | (i, j) \in M)$$

denote the probability that the field x agrees on records i and j , given that the pair of records (i, j) is a true match. Let

$$u_x = P(x_A^i = X_B^j | (i, j) \in U)$$

denote the probability that the field x agrees on records i and j , given that the pair of records (i, j) is a not a match. A matching weight is then computed for each field or variable used in the matching process as

$$\log_2 \frac{m_x}{u_x} \tag{1}$$

if the fields agree and

$$\log_2 \frac{1 - m_x}{1 - u_x} \tag{2}$$

if the fields disagree. The composite weight for a record pair is computed as the sum of the individual field weights.

In practice, the values for m_x and u_x are taken to be one minus the error rate of the field in matched records, and the unconditional probability that the field agrees at random based on a frequency analysis of all field values on the files (Jaro 1997), but other applications may compute these values differently (Winkler 1999a). Often, the exact values used in the actual application are derived from previous experience, a clerically edited subsample or a training dataset, since the true error rate of a field may not be known (see Winkler & Thibaudeau (1991) for an example comparing the different methods of defining parameters).

3.2 Matching earnings records

Measuring earnings using UI wage records presents interesting challenges. The earnings of employees who are present at the end of a quarter, but not at the beginning of the quarter are the earnings of acceding workers during that quarter. The UI wage records do not provide any information about how much of the quarter such individuals worked.⁸ The range of possibilities goes from 1 day to every day of the quarter. Similarly, the earnings of employees present at the beginning of a quarter who are not present at the end of the quarter represent the

⁸Although strictly true for California, the state providing the data used in this document, as well as for the vast majority of UI record keeping systems in the United States, there are some states that now or in the past have kept information on how much of a quarter was worked. For instance, Florida kept information on weeks worked in the system until the mid-1990s, while Minnesota and Washington state still record information on hours worked.

earnings of separations. Finally, workers present both at the beginning of the quarter and at the end are most likely, though not certain, to have worked continuously during that quarter. Thus, their earnings are closest to a “wage” measure. Workers that are thus observed are called “full-quarter employees” within the QWI system, and the earnings associated with such quarters are “full-quarter earnings”.

To clarify this, let a quarter Q be defined as the a segment of continuous time $[q, q + 1)$, where $q \in \mathbb{R}^+$, and the units are defined appropriately. Whether or not a worker was present at the beginning of a quarter is determined by the presence of a record for that worker and SEIN for the preceding quarter $Q - 1$ and the current quarter Q . By inference, if the worker was present in $Q - 1$ and in Q , she is assumed to have been present at the start of quarter, i.e. at time q .⁹ If a worker was present both at time q and at time $q + 1$, then she is assumed to have been present throughout quarter Q , and is found to be a “full-quarter employee”. Conversely, true single-quarter job spells are generated by workers who were present neither at the start nor at the end of the quarter.

Under reasonable assumptions about when the a job starts within a quarter, the earnings associated with true single-quarter job spells should be systematically and substantially lower than the earnings associated with wage records that have a miscoded SSN. The latter are actually earnings associated with a “full-quarter employee”. By the same token, for a job spell observed to be interrupted in quarter Q , the earnings of the bounding quarters $Q - 1$ and $Q + 1$ are “full-quarter” earnings if the true job spell is uninterrupted, but are the earnings of separations and accessions, respectively, if the job spell is truly interrupted, i.e., the observed job history is the truth.¹⁰

The competing hypotheses can be made more precise. Define time t to be the elapsed fraction of a quarter, $t \in [0, 1]$. Assume that the probability of an accession or a separation is constant throughout a quarter, i.e., $f(t) = c = 1$. Consider earnings in a quarter Q as a time rate times the time worked, $e(Q) = wt$, and denote $e_{FQ}(Q)$ the earnings associated with a full-quarter employee in quarter Q , $e_S(Q)$ those of separators, $e_A(Q)$ those of accessions, and finally $e_1(Q)$ those of true single-quarter job spells. Without loss of generality, normalize $w = 1$, and consider the null hypothesis that a plug and hole stem from the same job history against the alternate hypothesis that the two are

⁹For a precise definition for all point-in-time and other measures, consult Appendix B.

¹⁰In practice, the vast majority of bounding quarters are full-quarter earnings. When they are not, an adjustment is made to the earnings so that they correspond to expected full-quarter earnings.

unrelated. This hypothesis can be stated as:

$$\begin{aligned}
 H_0 : \quad E[e(Q-1)] &= e_{FQ}(Q) \text{ and} \\
 \quad E[e(Q)] &= e_{FQ}(Q) \text{ and} \\
 \quad E[e(Q+1)] &= e_{FQ}(Q) \\
 H_1 : \quad E[e(Q-1)] &= e_S(Q) \text{ and} \\
 \quad E[e(Q)] &= e_1(Q) \text{ and} \\
 \quad E[e(Q-1)] &= e_A(Q)
 \end{aligned}$$

Then the following relations hold:

$$\begin{aligned}
 e(Q-1) > e(Q) \quad \text{and} \quad e(Q+1) > e(Q) \quad \text{under } H_1 \\
 e(Q-1) = e(Q) \quad \text{and} \quad e(Q+1) = e(Q) \quad \text{under } H_0
 \end{aligned}$$

This is easily seen when computing expected earnings under the distributional assumptions above. Under H_0 , $e(Q) = e_{FQ}(Q)$ and $e(Q-1) = e_{FQ}(Q-1)$, and the same for $Q+1$. However, $E[e_{FQ}] = E[wt|FQ] = 1$ because no departure or accession occurred for FQ employees. On the other hand, under H_1 , $e(Q) = e_1(Q)$, $e(Q-1) = e_S(Q)$ and $e(Q+1) = e_A(Q)$. Here, $E[e_S] = E[e_A] = E[wt|separation] = \int_0^1 tf(t)dt = \frac{1}{2}$, whereas $E[e_1] = E[wt|A \text{ and } S] = \int_0^1 t(1-F(t))dt = \frac{1}{6}$.

Under the null hypothesis “The plug and the hole are not related”, earnings both for the plug and the hole are lower than full-quarter earnings, but the earnings for the plug are on average only a third of that of the hole. Under the alternate hypothesis “The plug and the hole stem from the same job history”, earnings of both the plug and the hole are in truth full-quarter earnings, and should match.

3.3 Implementation

In the process described here, we used Vality Integrity software. The software can be configured using a GUI interface from a desktop PC or from configuration files in batch mode on the executing server.

The first stage of the SSN editing starts with a list of unique combinations of SSN, First name, Middle Initial, and Last name (uniquely identified by the variable UID) across all years and quarters (see Table 1 on page 32). This stage verifies the likelihood that the records for a given SSN are actually for the same person, based on name information and weighted by frequency in the data. It is designed to capture false positives, *i.e.*, SSNs miscoded and wrongly attributed to another, valid, SSN.¹¹

¹¹This is not designed to do a full-scale unduplication effort. In particular, there is currently no attempt to standardize names at this stage, nor will this capture consistent miscoding by firms or consistent use of SSNs by multiple persons if that behavior persists for more than one quarter.

In the second stage, eligible records are identified and constructed in two ways. A potential plug is simply a single-quarter job, *i.e.*, the only quarter of employment ever observed for that SSN-SEIN match. The position of the observed wage in the population earnings distribution for that calendar year quarter is computed (expressed in decile positions). Under the hypothesis that these records are plugs, the observed earnings are full-quarter earnings. The data construction for holes is slightly more complex. First, we identify the year, quarter, and SEIN in which a one-period interruption for a given SSN(1) occurs. By definition, a hole is bounded on either side by a wage record. These records are extracted from the UI wage record files, and possibly adjusted.¹² Earnings observations from the two bounding quarters are averaged to obtain an estimate of the earnings which that particular SSN(1) would have had in the hole if he or she actually had worked during that quarter *i.e.*, under the null hypothesis that the hole is due to miscoding of a record from a continuous job spell, and not due to a true absence from work for more than one quarter length. A record containing the SEIN, SSN(1) and name information from the bounding quarters, plus the constructed earnings measure and its decile position, is output. The matching software then uses multiple passes, based on different block and string comparators, to generate match scores. Records above a threshold value, determined by iterative inspection of the data and the match results, are considered matches.

3.4 Results

Table 4 on page 34 shows the number of SSN(0)s reassigned in the first stage because of unreliable name information, expressed as a percentage of UIDs. Note that the number is slightly less than 10 percent of the total number of individuals SSN(0)s ever appearing in the data, and only a little more than 0.5% of all wage records. Trials in the late 1980s, in which the SSN and name information of a small number of wage records were handchecked by the Social Security Administration (SSA), found an average error rate of 7.8 percent, with significant variation across states (Bureau of Labor Statistics 1997b). The matching process implies a much lower error rate. That may in part be due to the conservative setup of the process, in part to the increased use of electronic submission of UI wage records, which substantially reduces error rates. On the other hand, the SSA trials are not feasible on a large scale, and typically involved less than 50,000 records. The process here verified over half a billion records.

Table 5 on page 35 tabulates the matching success rate in the second stage, by year and overall. Approximately 21% of at-risk records

¹²Earnings levels may not correspond to “full-quarter earnings” if the job history begins or terminates in the bounding quarters. For the precise adjustment, see Appendix A on page 16.

are matched. The at-risk group is composed of all interrupted job histories with an interruption of at most one quarter (“holes”). Match pairs are all single-quarter “plugs” that match a “hole”. Out of 96 million jobs (Table 2), over 800,000 have an employment history interruption that is eliminated by these matches (slightly less than 0.9%). The number of SSN(0)s, *i.e.*, (apparently) individually identifiable persons, is reduced by over 400,000 (nearly 1.5%).

4 Impact on economic estimates

We proceed in two steps. First, we discuss the effect on individual job histories, this being the most immediate impact of the correction undertaken. We then consider the impact when aggregating individual records to firm, county, or industry level statistics on employment, flows, and earnings. The latter two are precursors to a soon-to-be-released new statistical series by the U.S. Census Bureau, called the Quarterly Workforce Indicators (QWI). All aggregations are done separately by gender and by eight age groups, as well as on the global margin.

4.1 Bias at individual level

The most immediate impact is on individual job histories. In this section, we show both how many records out of all records are impacted, and how this effects job histories.

Table 6 on page 36 compares types of job histories before and after the editing process, comparable to Table 2 on page 33. Since only single-quarter interruptions are at risk of being closed, the most dramatic change is among job histories with single-quarter interruptions. Over 11% (over 600 thousand) of these job histories are eliminated. Most edited job histories are no longer interrupted. The largest absolute increase is among continuous, but not truncated jobs (C). Job histories covering the entire period, and thus truncated both left and right, are increased by over 4%.

Note that Table 6 underestimates the true extent of coding errors. Coding errors at the beginning or the end of a job spell are not captured here, but are likely to be present. Since such errors do not affect the number of interruptions of a job spell, and only the timing of job starts and separations as well as total tenure (and experience), the impact is likely to be minor on individual-level analyses.

4.2 Aggregate-level bias

When aggregating to higher levels, some errors will be averaged out, while others are exacerbated. The variables discussed in this section

are of both kinds. Some are simple person counts within a unit and quarter, where a unit can be a firm, county, or industry. Others are measures of job and person flows in and out of firms, which are then aggregated up to county or industry levels. Whereas the former are unlikely to be much affected,¹³ the impact on the latter is presumably substantial. Consider that every false interruption in a person’s job history will lead to two accessions, two separations, one new hire, and one recall that would otherwise not have occurred. Large biases in variables based on such flow concepts (accessions A , separations S , recalls R , new hires H) are more likely to occur than in stocks. Among stock variables, those based longer periods of employment persistence (FJF , F)¹⁴ are more easily biased by miscoded records than those based on shorter time periods (B , E , \bar{E}). Payroll sums for accessions are going to be biased upwards, because a part of those labelled accessions are actually miscoded long-tenure workers, who typically have higher earnings than true new hires.

The variables we will be using in this analysis are described in Table 7 on page 37 and defined in Appendix B. They are computed over all demographic groups as well as for single-characteristic margins as detailed in Table 8. For example, we will consider the bias in accessions for all individuals, for men and women separately, and for each age group separately, but not for women aged 22-24. The bias is computed as

$$dX = X_{pre} - X_{post} \tag{3}$$

$$pX = \frac{dX}{X_{post}} \tag{4}$$

where X is some variable, and pre and $post$ indicate computation of X before and after the editing procedure.

Both dX and pX are computed for each variable, for each selected demographic group, at all levels of aggregation (firm, county, or SIC division), for all 40 quarters of data. pX is not computed when X_{post} is zero, which should be kept in mind when analyzing distributions of pX .

Table 9 on page 39 tabulates different points in the distribution of the bias for each variable,¹⁵ across the universe of either quarterly firm, county or industry cells, for the overall margin only. Mean biases (in absolute value) among flow and stock variables range from a low of 0.25% to a high of 15.68%, and range between 0.01% and 4.92% for payroll variables. As expected, variables that are based on flows are

¹³In theory, counts of persons ever working for a firm during a quarter (M) should not be affected at all. This is true in a simple aggregation, but some minor sample selection in the preparation of the QWI makes this not quite true in the actual data.

¹⁴See Appendix B for detailed definitions.

¹⁵See Table 7 on page 37 for the full names of the variables, which are omitted for space reasons.

more biased than those based on stocks. Accessions A within industry cells are overestimated by nearly 2%, whereas end-of-quarter employment E is only underestimated by 0.3%. The time frame underlying some variables also influences the mean bias in the expected direction. Full-quarter job flows FJF are overestimated by 4.7% within industry cells, but JF only by 1.7%. Full-quarter employees F , which are required to have at least three consecutive wage records, are underestimated by 0.8% within country cells, but the simple within-period average \bar{E} , which only requires two consecutive wage records, has a downward bias of only 0.46%. New hires H , which count only employees who had no wage records in the past four quarters, and thus exclude most miscoded wage records, are biased upwards by only 1.4% within industry, but recalls R , which almost all miscoded records are taken to be, are biased upwards by nearly 6%.

All measures of (bounded) non-employment preceding the different accession statistics (NA , NH , NR , NS) are biased upwards, as is to be expected, but the again the largest bias is among recalls. Finally, cumulative payroll variables for stocks ($W1$ and $W2$) do not show a large bias. In particular, $W1$ should not show any bias, since summation of records over employers (SEIN) is not affected, and small bias showing up here is probably due to small selection issues when compiling wage records. On the other hand, $W3$ is downward biased more substantially because missing wage records reduce full-quarter employment over three quarters. On the other hand, payroll of accessions WA and separations WS is upward biased for the same reasons mentioned above.

Turning to other points in the distribution of each variable's bias across cells, two things are of note. First, over 80 percent of the over 20 million firm-quarter cells are not biased, since both the 10th and the 90th percentile are zero. Those that are, however, are substantially biased, as witnessed by the mean. When aggregated to the county or industry level, on the other hand, most cells are biased. The median for most variables is close to the mean, with the exception of the jobflow variables JF and FJF . The top and bottom deciles are also typically close to the mean, most often within one standard deviation of the mean.

The net flows JF and FJF differ in another respect. Whereas most flow measures are unidirectional (i.e. separations are by definition negative flows, whereas accessions are by definitions positive flows), JF and FJF can be either positive or negative. The bias also goes in both directions, as shown by the spread between the 10th and 90th percentiles. Given the symmetric distribution, it is thus not surprising that net flows have a mean quite close to zero. Nevertheless, there is a lot of bias in the statistic as evidenced by the tails of the distribution.

Table 10 on page 42 provides the same information, but for dX , the bias expressed in levels, rather than the percentage bias. Note,

in particular, the payroll sums that are hidden behind the percentage bias in Table 9. Table 10 also shows that the biases are larger when aggregated to the industry than when aggregated to the county level. In Table 9, the percentage biases within counties typically, but not always, were larger than within industries. The miscoding of identifiers is essentially random in the universe of wage records, but affects flows non-randomly, since it generates false flows. It is thus natural to expect that small flows are more strongly biased by this than large flows. Furthermore, both Tables 9 and 10 only tabulate the distribution of biases for the overall margin. Given the likely dependency on the size of the underlying population, gender and age-specific statistics are even more likely to be biased.

To further explore the relation between the bias and the flow, we turn to some straightforward regressions. Table 11 on page 44 tabulates results from regressions of the form

$$pX_{jkt} = \beta_0 + \beta_1' Z_{jkt} \quad (5)$$

for some unit j , either a county (Table 11a) or an industry (Table 11b), some margin k (see Table 8) and some quarter t . Note that contrary to the results in the previous tables, these regressions take into account statistics for all margins, not just the overall margin. The means reported in Table 9 corresponds to such a regression, with constrained $k = 0$ and $\beta_1 = 0$. For the flow and stock statistics, Z contains XR , the rate associated with X , defined relative to the appropriate basis (\bar{E} or F). For all non-employment counts $X = NY$ and payroll sums $X = WY$, except for $X = W3$, Z contains the accession rate AR and the separation rate SR . For $W3$, Z contains $FJFR$. Thus, each regression controls for the size of the associated statistics. Furthermore, a second equation of the form

$$pX_{jkt} = \beta_{3j} + \beta_4' Z_{jkt} \quad (6)$$

is also estimated, to condition on the different sizes of industries and in particular counties. The last row of each block in Table 11 reports an F-test for the joint significance of these fixed unit effects.

As before, the results can be split into four groups: gross flows, net flows (JF , FJF), non-employment counts, and payroll sums. For nearly all flows (exception being $H3R$), the relationship between the percent bias and its associated rate is the same. The bias is negatively related to the size of the flow rate, but even when controlling for the rate, is significantly different from zero. This is true whether doing cross-sectional or within cell analysis. Generally, controlling for cell-specific average bias improves the explanatory power of the regression significantly, as evidenced by the F tests. Net flows, on the other hand, do not have significant average bias, even when controlling for the size of the rate, and do not vary substantially across cells. This pattern is consistent with random errors in the *stock* of wage records. All errors

by definition increase flows, and this bias increases as the relative flow for the cell decreases. For instance, job turnover is generally lower for workers in the middle age brackets. The regression results tell us that the flow estimates for this group of workers will be more biased by the errors than for young people.

The number of non-employment periods for new hires, recalls, and separations is typically negatively related to both accession and separation rates of the particular cell. Those for accessions, on the other hand, are positively related to a cell's separation rate.

The bias for the payroll for all workers (in a cell) for a particular quarter $W1$ is not very large, and at least within and across industry cells, not systematically different from zero. Remember that since records are not re-allocated across SEINs or quarters, these numbers only differ because of some small sample selection issues at the global margin. Within specific age or gender categories, though, this is no longer true. Since most miscoded SSNs do not have associated information on gender or age, this is imputed. Re-assignment to its true cell will most likely also change age and gender information, and thus change the value in two cells. The bias is more systematic when concentrating on more selective measures. The number of end-of-period employees for any given SEIN and quarter will be reduced by errors, and the error in the associated payroll $W2$, even when controlling for flows in and out of the cell, is still significantly negative, between 0.2 and 0.7 percent. This effect is even stronger for payroll sums of full-quarter workers $W3$.

Payroll for accessions WA and separations WS are biased upward, by up to 7 percent. One explanation can be found in the usual hazard rate pattern for a worker's tenure, which is downward sloping, implying that separations are composed mostly of short-tenure workers. Equally, accessions typically are at the start of a career, and earn less than high tenure workers. And trivially, as explained above (Section 3.2), the earnings of true separations and accessions are on average for a shorter time period than that of full-quarter workers. All this leads to misallocated high-tenure full-quarter, and thus high-earning workers being classified as low-earnings separations and accessions, inflating those payroll sums.

5 Conclusion

In this paper, we propose and implement an algorithm particularly suited for the probabilistic matching of UI wage records. We consider the chosen methods to be conservative, *i.e.*, the achieved match rates are substantially lower than the true error rate, but are likely to be the best that can be done with this type of data. Nevertheless, the bias revealed by this procedure is substantial in a number of important vari-

ables both at the individual level and at all levels of aggregation. Job spells observed to be interrupted are decreased by 11 percent. Average biases in some flow statistics are around 7 percent, with substantial variation around that mean.

The potential uses of administrative data are vast. However, this study highlights the pitfalls that researchers and statisticians may encounter when using UI wage records. Probabilistic matching in this context greatly enhances the value of such data.

Acknowledgement

This research is a part of the LEHD Program at the U.S. Census Bureau, which is partially supported by the National Science Foundation Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging, and the Alfred P. Sloan Foundation. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Census Bureau, Cornell University, or the National Science Foundation. Confidential data from the LEHD Program were used in this paper. The U.S. Census Bureau is preparing to support external researchers' use of these data under a protocol to be released in the near future. For further information, contact Ronald C. Prevost LEHD Program Director (ronald.c.prevost@census.gov).

The authors acknowledge the substantial contributions of the staff and senior research fellows of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program. Corresponding author: Lars Vilhuber (lars.vilhuber@cornell.edu).

A Description of matching algorithms

The SSN editing procedure used here is split into two stages. The first stage starts with a list of unique combinations of SSN, First name, Middle Initial, and Last name (uniquely identified by the variable UID) from all files across all years and quarters. This stage verifies the likelihood that the records for a given SSN are actually for the same person, based on name information and weighted by frequency in the data. It is designed to capture “false positives” (SSNs miscoded and wrongly attributed to another valid SSN), and is not designed to do a full-scale unduplication effort. In particular, there is currently no attempt to standardize names at this stage, nor will this capture consistent miscoding by firms or consistent use of SSNs by multiple persons if that behavior persists for more than one quarter. At the end of Stage 1, records deemed not to pertain to the SSN(0) with which they were associated are assigned a temporary SSN, which together with all retained SSN(0) becomes SSN(1).

The second stage does the actual probabilistic matching, based on the SSN(1) and SEIN information. Plugs that are successfully matched to holes obtain an SSN(2), which corresponds to the SSN(0) of the records bounding the hole. A record with a SSN(1) that is not matched to any hole is reassigned its SSN(0). These records, whose allocation to a specific SSN(0) employment history seems doubtful based on a comparison of names, cannot be associated with any existing holes with sufficient confidence (*i.e.*, a probability score below the threshold value), and is put back into its original employment history.

A.1 Stage 1: Unduplication

In total, four passes are used in order to accommodate different scenarios (constellations of name information) in the data.¹⁶ All passes “block” on SSN(0), *i.e.* a record’s name information is only compared to name information on other records with the same SSN(0). In all passes, name information is weighted by the number of UI wage records that have that name information on file. The matching software identifies names that are associated with no other wage record for that particular SSN(0). Thus, in the following example, records 51 and 52 are similar, whereas records 53 and 54 are sufficiently different to be deemed “miscoded”, and rejected in Stage 1.¹⁷

Example 1

¹⁶The actual match parameters are available on demand. They are specific to the realized data, and require modifications if applied to data from a different source.

¹⁷All names and SSNs used in this and other examples are purely fictitious.

Records with SSN(0)=123-45-6789

Name	Info on file				
	UID	first_name	middle_name	last_name	Quarter
John C. Doe	51	JOHN	C	DOE	92Q1
John C. Doe	51	JOHN	C	DOE	92Q2
John C. Doe	52	JOHN		DOE	93Q1
John C. Doe	52	JOHN		DOE	94Q1
Robert E. Lee	53	ROBERT	E	LEE	94Q2
Ulysses S. Grant	54	ULYSSES	S	GRANT	94Q2

On the other hand, none of the passes will capture repeated use of SSN(0)s by different people, potentially illegally, or because some employer has miscoded the SSN in her files for several quarters. In the following example, John C. Adam might be the legitimate holder of SSN 123-45-6789, whereas Robert E. Benjamin's employer miscoded his true SSN (723-45-6709) when Robert starting working for her, and nobody noticed this for two quarters. Robert's records will *not* be rejected by the matching software at this stage, because he has multiple records using the same, wrong SSN(0).

Example 2**Records with SSN(0)=123-45-6789**

Name	Info on file				
	UID	first_name	middle_name	last_name	Quarter
John C. Adam	151	JOHN	C	ADAM	92Q1
John C. Adam	151	JOHN	C	ADAM	92Q2
John C. Adam	152	JOHN		ADAM	93Q1
John C. Adam	152	JOHN		ADAM	94Q1
Robert E. Benjamin	153	ROBERT	E	BENJAMIN	94Q2
Robert E. Benjamin	153	ROBERT	E	BENJAMIN	94Q3

The second case, not solved in Stage 1, can be solved in different ways. First, validation of each UID by SSA would yield a validated SSN for John, but an invalid SSN for Robert. Second, the miscoding of Robert's SSN(0) will yield a short employment spell for that SSN, which

could be linked up to the employment spell associated with Robert's true SSN, based on start and end dates, and the name information on the file.

A.2 Pass 1

The first pass captures the bulk of the differences. It is based on a straight comparison of all components of the name: the first name of a record is compared only to first names on other records, the last name only to last names on other records, and the middle name only to other middle names.

A.3 Pass 2

A second pass was added to allow for switched first and middle names. Inspection of the data reveals that many of these switches are actually part of a more general problem, presumably rooted in some historical data processing problem. In part of the data, composite family names are written as one word. However, in other years, this same information is miscoded in the data received at Census. The family name is written with spaces, but some parsing on systems has allocated the first part of the last name to `last_name`, but the second part to the `first_name`, with the first name being relegated to `middle_name`.¹⁸

Example 3

<i>Name</i>	<i>Info on file</i>			
	<i>Recnum</i>	<i>first_name</i>	<i>middle_name</i>	<i>last_name</i>
<i>Al DiMeola</i>	<i>1</i>	<i>AL</i>		<i>DIMEOLA</i>
<i>Al DiMeola</i>	<i>2</i>	<i>MEOLA</i>	<i>A</i>	<i>DI</i>
<i>Joe DiMaggio</i>	<i>3</i>	<i>JOE</i>		<i>DIMAGGIO</i>
<i>Joe DiMaggio</i>	<i>4</i>	<i>MAGGIO</i>	<i>J</i>	<i>DI</i>

Another frequent scenario is also attributable to data entry problems (and cannot be captured by standardizer programs). In this second scenario, parts of the first name are coded into the last name field:

Example 4

¹⁸All names and SSNs used in this and other examples are purely fictitious.

<i>Info on file</i>				
<i>Name</i>	<i>Recnum</i>	<i>first_name</i>	<i>middle_name</i>	<i>last_name</i>
<i>John C. Doe</i>	<i>5</i>	<i>JOHN</i>	<i>C</i>	<i>DOE</i>
<i>John C. Doe</i>	<i>6</i>	<i>C</i>		<i>DOEJOH</i>

Note that these cases seem to occur in the earlier years of the data, where last name information was restricted to six and first name information to one character.

Since both scenarios are interspersed in the data, and seem to occur concurrently, it is difficult to post-process these names before running them through the unduplication process. In particular, most of these cases would not get changed by using standardizer software. In fact, it is likely that an incorrect parametrization of a standardizer lies at the root of these problems.

However, permitting a switch between first and middle names, while controlling for an uncertainty match on family names, captures most of these cases, matching Joe with J and DiMaggio with Di. Nevertheless, this might turn out to be a problem in later stages of the matching process.

A.4 Pass 3

Next, a third pass was added to allow for switched middle and last names. This is necessary for two observed scenarios: First, some women seem to move the maiden name to middle initial, and this pass captures that well. Second, data entry errors pop up here again, with *last_name* containing both the last name and the middle initial.

Example 5

<i>Info on file</i>				
<i>Name</i>	<i>Recnum</i>	<i>first_name</i>	<i>middle_name</i>	<i>last_name</i>
<i>Nicole M. Kidman</i>	<i>7</i>	<i>NICOLE</i>	<i>M</i>	<i>KIDMAN</i>
<i>Nicole K. Cruise</i>	<i>8</i>	<i>NICOLE</i>	<i>K</i>	<i>CRUISE</i>
<i>Nicole M. Kidman</i>	<i>9</i>	<i>NICOLE</i>		<i>M KIDM</i>

A.5 Pass 4

Finally, the fourth pass allows for switched first and last names, with a straight match on middle initials (if existant). Again, the most likely

source for this are data entry errors. This last pass is the most tenuous comparison, since it reduces to a simple comparison of the first letters of first and last names if one or the other are single-character. However, remember that all comparisons are done within the same observed SSN, so that these are not randomly combined individuals from the general population based on first and last initial concordance.

A.6 All passes

All passes also use a matching field created by concatenating first and middle initials with the first six digits of the last name, and taking out all blank spaces (variable *CONCAT*). This is a frequent error in the data, similar to the following example:

Example 6

<i>Name</i>	<i>Info on file</i>				
	<i>Recnum</i>	<i>first_name</i>	<i>middle_name</i>	<i>last_name</i>	<i>concat</i>
<i>Nicole M. Kidman</i>	<i>7</i>	<i>NICOLE</i>	<i>M</i>	<i>KIDMAN</i>	<i>NMKIDMAN</i>
<i>Nicole M. Kidman</i>	<i>11</i>			<i>NMKIDM</i>	<i>NMKIDM</i>

When this occurs, matching on any individual name components may not provide enough concordance. However, a match gets extra weight assigned if the *CONCAT* matches on both records. Thus, in the above example, the *CONCAT* of both records are a much better comparison than the other variables.

The question arises whether to aggressively weed out “false” positives, potentially also eliminating some valid matches. We have chosen to be aggressive at this stage. Any valid matches that are eliminated at this stage are reintegrated fairly easily at a later stage with more matching information available. Furthermore, identification of discontinuities created purely the Stage 1 procedure, and subsequent readjustment of records, is straightforward, and done before release for data processing. The only downside is that the number of records needing to be matched at later stages increases.

A.7 Post-processing

For records that were identified as close duplicates of each other, the “best” name as determined by the matching software is retained. UID records determined not to match other records for a given SSN (“residuals”) are assumed to be false positives. They are assigned a unique

identifier, based on the SSN(0), using an algorithm that ensures assignment of a unique SSN while retaining information on the original SSN(0). The following table provides a quick reference into how the original SSN(0) digits are transposed by the algorithm to yield the new identifiers SSN(1).

Original digit	Replacement characters					
0	A	K	U	e	o	y
1	B	L	V	f	p	z
2	C	M	W	g	q	
3	D	N	X	h	r	
4	E	O	Y	i	s	
5	F	P	Z	j	t	
6	G	Q	a	k	u	
7	H	R	b	l	v	
8	I	S	c	m	w	
9	J	T	d	n	x	

For instance, in Example 1, above, UID 53, the record for “Robert E. Lee”, which is associated with SSN(0)=123-45-6789 on the original wage records, gets assigned SSN(1)=123-45-6H89. UID 54 gets assigned SSN(1)=123-45-6R89.

A.8 Stage 2: Correcting broken job histories

The first step for the within-job matching stage is to select the eligible records. We use the information on job histories, based on SSN(1)-SEIN matches, to select eligible histories, i.e. those that have a single interruption of one quarter length (potential holes) at that employer, as well as job histories that are exactly one quarter long with that employer (potential plugs). We then perform a statistical match, conditional on eligibility, based on name information and the decile of the earnings distribution a given record is associated with.

A.8.1 Construction of earnings information

The extraction of data of the earnings information for potential plugs is straightforward, once one-period job histories have been identified: They correspond strictly to the wage records as found on the UI wage record files. The data construction for holes is slightly more complex.

First, we identify the year and quarter in which a one-period interruption for a given SSN(1)-SEIN combination occurs. By definition, a “hole” is bounded on either side by a wage record. These records are extracted from the UI wage record files.

However, earnings levels may not correspond to “full-quarter earnings” if the job history begins or terminates in the bounding quarters. For instance, in the following example, all SSN(1)s have a “hole” in quarter Q5. For SSN(1)=123-45-1234 (case A) and SSN(1)=123-45-1235 (case B), one of the bounding quarters is the bounding quarter of a job spell.

Example 7

SSN(1)	Job history							
	Q1				Q5			
A 123-45-1234	0	0	0	1	0	1	1	1
B 123-45-1235	1	1	1	1	0	1	0	0
C 123-45-1236	1	1	1	1	0	1	1	1

Thus, earnings in those quarters do not correspond to “full-quarter” earnings. Matching on the earnings decile based on the raw earnings information would fail.

Here, an adjustment is made to the wage data to make the assigned earnings deciles correspond more closely to “full-quarter” earnings of the hypothesized plugs. We verify whether the SSN(1) in question had positive earnings *two* quarters on either side of the hole. Thus, in the above example, we verify whether cases A through C have positive earnings in Q3 and Q7. If this is not the case, then the earnings of the corresponding bounding quarter are upweighted by a factor of two, based on the fact that the expected accession (separation) time within a known interval is its midpoint. In the above example, the earnings corresponding to Q3 for case A, and the earnings corresponding to Q7 for case B, are doubled. Case C is not adjusted.

After adjustment of the earnings for any of the bounding quarters, the earnings observations from the two bounding quarters are averaged to obtain an estimate of the earnings which that particular SSN(1) would have had in the “hole” if he or she actually had worked during that quarter (i.e. under the null hypothesis that the “hole” is due to miscoding of a record from a continuous job spell, and not due to a true absence from work for more than one quarter length).

The SEIN, name and SSN(1) information from the two bounding quarters correspond by virtue of definition and homogenization in Stage

1. We thus output a record containing the SEIN, SSN(1) and name information from the bounding quarters, plus the constructed earnings measure and its decile position. Note that this layout corresponds exactly to the layout of the potential plugs.

A.8.2 Restrictions

A technical constraint is imposed on the process by the software used. The efficiency of most matching software declines with the square of the number of items within a block, i.e. the number of records that match exactly on a select number of variables. In VALITY, this is around 1000 records.

There are also fundamental reasons to concentrate on blocks with fewer records. Large blocks of job histories with interruptions may reflect systematic, rather than random coding error, or may reflect a prolonged strike or similar economic event. Large blocks of one-quarter employment spells may reflect firms with particularly high turnover. In either case, it becomes more difficult to distinguish similar records based on poor name information and concordance of dates alone.

For practical purposes, we have restricted blocks to not be larger than 750 elements, both for plugs and holes.

B Definition of statistics

The variable t refers to the sequential quarter, and runs from $qmin = 1$ corresponding to 1985:1 to $qmax$ defined for the latest quarter available (here: 1999:4), regardless of the state being processed. The quarters are numbered sequentially from 1 (1985:1) to the latest available quarter. The variable $qfirst$ refers to the first available sequential quarter of data (here: 23, corresponding to 1991:3). The variable $qlast$ refers to the last available sequential quarter of data for a state (here identical to $qmin$). Unless otherwise specified a variable is defined for $qfirst \leq t \leq qlast$.

Statistics are computed from individual-level job movements, and then aggregated to higher levels. The following will define individual and firm level statistics; higher levels of aggregations are straightforward.

B.1 Individual concepts

Flow employment (m): for $qfirst \leq t \leq qlast$, individual i employed (matched to a job) at some time during period t at employer j

$$m_{ijt} = \begin{cases} 1, & \text{if } i \text{ has positive earnings at employer } j \text{ during quarter } t \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Beginning of quarter employment (b): For $qfirst < t$, individual i employed at the end of $t - 1$, beginning of t

$$b_{ijt} = \begin{cases} 1, & \text{if } m_{ijt-1} = m_{ijt} = 1 \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

End of quarter employment (e): For $t < qlast$, individual i employed at j at the end of t , beginning of $t + 1$

$$e_{ijt} = \begin{cases} 1, & \text{if } m_{ijt} = m_{ijt+1} = 1 \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Accessions (a_1): For $qfirst < t$, individual i acceded to j during t

$$a_{1ijt} = \begin{cases} 1, & \text{if } m_{ijt-1} = 0 \text{ \& } m_{ijt} = 1 \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Separations (s_1): For $t < qlast$, individual i separated from j during t

$$s_{1ijt} = \begin{cases} 1, & \text{if } m_{ijt} = 1 \text{ \& } m_{ijt+1} = 0 \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Full quarter employment (f): For $qfirst < t < qlast$, individual i was employed at j at the beginning and end of quarter t (full-quarter job)

$$f_{ijt} = \begin{cases} 1, & \text{if } m_{ijt-1} = 1 \text{ \& } m_{ijt} = 1 \text{ \& } m_{ijt+1} = 1 \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

New hires (h_1): For $qfirst + 3 < t$, individual i was newly hired at j during period t

$$h_{1ijt} = \begin{cases} 1, & \text{if } m_{ijt-4} = 0 \text{ \& } m_{ijt-3} = 0 \text{ \& } m_{ijt-2} = 0 \text{ \& } m_{ijt-1} = 0 \text{ \& } m_{ijt} = 1 \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

New hires to full quarter status Hires!New!to full quarter status

(a_3): For $qfirst+4 < t < qlast$, individual i transitioned from consecutive-quarter hired to full-quarter hired status at j at the start of $t + 1$ (hired in $t - 1$ and full-quarter employed in t)

$$h_{3ijt} = \begin{cases} 1, & \text{if } h_{1ijt-1} = 1 \text{ \& } f_{ijt} = 1 \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

Recalls (r_1): For $qfirst + 3 < t$, individual i was recalled from layoff at j during period t

$$r_{1ijt} = \begin{cases} 1, & \text{if } m_{ijt-1} = 0 \text{ \& } m_{ijt} = 1 \text{ \& } h_{ijt} = 0 \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

Total earnings during the quarter (w_1): for $qfirst \leq t \leq qlast$, earnings of individual i at employer j during period t

$$w_{1ijt} = \sum \text{all UI covered earnings by } i \text{ at } j \text{ during } t \quad (16)$$

Earnings of end-of-period employees at employer j during period t

$$w_{2ijt} = \begin{cases} w_{1ijt}, & \text{if } e_{ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (17)$$

Earnings of full-quarter individual i at employer j during period t

$$w_{3ijt} = \begin{cases} w_{1ijt}, & \text{if } f_{ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (18)$$

For $qfirst \leq t \leq qlast$, total earnings of individual i during period t

$$w_{1i \bullet t} = \sum_{j \text{ employs } i \text{ during } t} w_{1ijt} \quad (19)$$

Total earnings of end-of-period employees i during period t

$$w_{2i \bullet t} = \begin{cases} w_{1i \bullet t}, & \text{if } e_{ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (20)$$

Total earnings of full-quarter employees i during period t

$$w_{3i \bullet t} = \begin{cases} w_{1i \bullet t}, & \text{if } f_{ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (21)$$

For $qfirst < t$, change in total earnings of individual i between periods $t - 1$ and t . The goal is to produce statistics based on:

$$\Delta w_{1i \bullet t} = w_{1i \bullet t} - w_{1i \bullet t-1} \quad (22)$$

Earnings of accessions to employer j during period t

$$wa_{1ijt} = \begin{cases} w_{1ijt}, & \text{if } a_{1ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (23)$$

Earnings of separations from employer j during period t

$$ws_{1ijt} = \begin{cases} w_{1ijt}, & \text{if } s_{1ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (24)$$

Periods of non-employment prior to an accession by i at employer j during t during the previous four quarters (defined for $qfirst + 3 < t$)

$$na_{ijt} = \begin{cases} \sum_{1 \leq s \leq 4} n_{it-s}, & \text{if } a_{1ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (25)$$

where $n_{it} = 1$ if $m_{ijt} = 0 \forall j$.

Periods of non-employment prior to a new hire by i at employer j during t during the previous four quarters

$$nh_{ijt} = \begin{cases} \sum_{1 \leq s \leq 4} n_{it-s}, & \text{if } h_{1ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (26)$$

Periods of non-employment prior to a recall by i at employer j during t during the previous four quarters

$$nr_{ijt} = \begin{cases} \sum_{1 \leq s \leq 4} n_{it-s}, & \text{if } r_{1ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (27)$$

Periods of non-employment following a separation by i from employer j during t during the next four quarters, (defined for $t < qlast - 3$)

$$ns_{ijt} = \begin{cases} \sum_{1 \leq s \leq 4} n_{it+s}, & \text{if } s_{1ijt} = 1 \\ \text{undefined,} & \text{otherwise} \end{cases} \quad (28)$$

B.2 Employer concepts

For statistic x_{cijt} denote the sum over i during period t as $x_{c.jt}$. For example, beginning of period employment for firm j is written as:

$$b_{.jt} = \sum_i b_{ijt} \quad (29)$$

All individual statistics generate employer totals according to the formula above. The key employer statistic is the average end-of-period employment growth rate for employer j , the components of which are defined here.

Beginning-of-period employment (number of jobs)

$$B_{jt} = b_{.jt} \quad (30)$$

End-of-period employment (number of jobs)

$$E_{jt} = e_{.jt} \quad (31)$$

Employment any time during the period (number of jobs)

$$M_{jt} = m_{.jt} \quad (32)$$

Full-quarter employment

$$F_{jt} = f_{.jt} \quad (33)$$

Net job flows (change in employment) for employer j during period t

$$JF_{jt} = E_{jt} - B_{jt} \quad (34)$$

Average employment for employer j between periods $t - 1$ and t

$$\bar{E}_{jt} = \frac{(B_{jt} + E_{jt})}{2} \quad (35)$$

Net change in full-quarter employment for employer j during period t

$$FJF_{jt} = F_{jt} - F_{jt-1} \quad (36)$$

Accessions for employer j during t

$$A_{jt} = a_{1.jt} \quad (37)$$

Separations for employer j during t

$$S_{jt} = s_{1.jt} \quad (38)$$

New hires for employer j during t

$$H_{jt} = h_{1.jt} \quad (39)$$

Full Quarter New hires for employer j during t

$$H_{3jt} = h_{3.jt} \quad (40)$$

Recalls for employer j during t

$$R_{jt} = r_{1.jt} \quad (41)$$

Total payroll of all employees

$$W_{1jt} = w_{1.jt} \quad (42)$$

Total payroll of end-of-period employees

$$W_{2jt} = w_{2.jt} \quad (43)$$

Total payroll of full-quarter employees

$$W_{3jt} = w_{3.jt} \quad (44)$$

Total payroll of accessions

$$W_{Ajt} = wa_{1.jt} \quad (45)$$

Total payroll of separations

$$W_{Sjt} = ws_{1.jt} \quad (46)$$

Total periods of non-employment for accessions

$$N_{Ajt} = na.jt \quad (47)$$

Total periods of non-employment for new hires (last four quarters)

$$N_{Hjt} = nh.jt \quad (48)$$

Total periods of non-employment for recalls (last four quarters)

$$N_{Rjt} = nr.jt \quad (49)$$

Total periods of non-employment for separations

$$N_{Sjt} = ns.jt \quad (50)$$

B.3 Aggregation of flows

We calculate the aggregate job flow as

$$JF_{kt} = \sum_{j \in \{K(j)=k\}} JF_{jt}. \quad (51)$$

for some county (or industry division (SIC)) k for some group of firms, where the function $K(j)$ indicates the classification into counties (or industries) associated with firm j .

References

Abowd, J. M., Lane, J. I. & Prevost, R. (2000). Design and conceptual issues in realizing analytical enhancements through data linkages of employer and employee data, *Proceedings of the Federal Committee on Statistical Methodology* .

- Abowd, J. M., Lengermann, P. A. & Vilhuber, L. (2002). The creation of the employment dynamics estimates, *Technical paper TP-2002-13*, LEHD, U.S. Census Bureau.
- Anderson, P. & Meyer, B. (1994). The extent and consequences of job turnover, *Brookings Paper on Economic Activity: Microeconomics* pp. 177–248.
- Anonymous (2000). *The INTEGRITY Data Re-engineering Environment, SuperMATCH Concepts and Reference*, version 3.0 edn, Vality Technology Inc. Vality Technology Inc was bought by Ascential Software, Inc. in 2002.
- Bowlus, A. & Vilhuber, L. (2002). Displaced workers, early leavers, and re-employment wages, *Technical paper TP-2002-18*, LEHD, U.S. Census Bureau.
- Bureau of Labor Statistics (1997a). *BLS Handbook of Methods*, U.S. Bureau of Labor Statistics, Division of Information Services, Washington DC. <http://www.bls.gov/opub/hom/>.
- Bureau of Labor Statistics (1997b). Quality improvement project: Unemployment insurance wage records, *report*, U.S. Department of Labor.
- Burgess, S., Lane, J. & Stevens, D. (2000). Job flows, worker flows and churning, *Journal of Labor Economics* **18**(3).
- Davis, S. J., Haltiwanger, J. C. & Schuh, S. (1996). *Job creation and destruction*, MIT Press, Cambridge, MA.
- Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage, *Journal of the American Statistical Association* **64**: 1183–1210.
- Haltiwanger, J. C., Lane, J. I. & Spletzer, J. R. (1999). Productivity differences across employers: The role of employer size, age, and human capital, *American Economic Review* **89**(2).
- Jacobson, L. S., LaLonde, R. J. & Sullivan, D. G. (1993). Earnings losses of displaced workers, *American Economic Review* **83**(4): 685–709.
- Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida, *Journal of the American Statistical Association* **89**: 414–420.
- Jaro, M. A. (1997). *AUTOMATCH Generalized Record Linkage System*, version 4.2 edn, MatchWare Technologies, Inc., Burtonsville, Maryland, 20866. Matchware Technologies Inc. was bought by Vality Technology Inc, and is now owned by Ascential Software, Inc.

- Lane, J., Miranda, J., Spletzer, J. & Burgess, S. (1999). The effect of worker reallocation on the earnings distribution: Longitudinal evidence from linked data, North-Holland, Amsterdam, pp. 345–74.
- Little, R. J. A. & Rubin, D. B. (1990). *The Analysis of Social Science Data with Missing Values*, Modern methods of data analysis, Sage, Newbury Park, Calif.
- Newcombe, H. B., Kennedy, J. M., Axford, S. J. & James, A. P. (1959). Automatic linkage of vital records, *Science* **130**: 954–959.
- U.S. Census Bureau (2000). *History of the 1997 Economic Census*, number POL/00-HEC, U.S. Census Bureau, Washington DC.
- Winkler, W. E. (1993). Matching and record linkage, *Research Report Series 93/08*, U. S. Bureau of the Census, Washington, D.C.
- Winkler, W. E. (1999a). The state of record linkage and current research problems, *Research Report Series 99/04*, U. S. Bureau of the Census, Washington, D.C.
- Winkler, W. E. (1999b). State of statistical data editing and current research problems, *Research Report Series 99/01*, U. S. Bureau of the Census, Washington, D.C.
- Winkler, W. E. & Thibaudeau, Y. (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. Decennial Census, *Research Report Series 91/9*, U. S. Bureau of the Census, Washington, D.C.

C Tables

Table 1: Unique combinations of SSNs and Names

Year	Observations	Unique Keys
1991	29 138 811	14 656 899
1992	56 356 832	16 508 875
1993	56 006 335	16 352 185
1994	56 992 314	17 084 002
1995	58 066 989	17 158 021
1996	60 157 386	20 021 727
1997	62 604 006	19 179 948
1998	64 524 103	22 476 213
1999	66 270 481	22 883 341
1991-1999	510 117 257	57 393 771

Table 2: "Holes" in job and employment histories

Pattern in job history	Job histories			Employment histories		
	(a)	(b)	(c)	(d)	(e)	(f)
	Frequency	Percent	Cumul. Percent	Frequency	Percent	Cumul. Percent
<i>Non-continuous,</i>						
<i>length of longest interruption</i>						
1 quarter	5 315 869	5.50%		3 461 297	12.17%	
2 quarters	2 357 942	2.44%		1 924 432	6.77%	
3 quarters	1 764 701	1.83%		1 514 519	5.33%	
4 quarters	750 910	0.78%		981 093	3.45%	
5 quarters	532 174	0.55%		759 555	2.67%	
6 quarters	466 301	0.48%		654 760	2.30%	
7 quarters	430 549	0.45%		558 690	1.97%	
8 quarters	241 573	0.25%		417 023	1.47%	
9 or more quarters	1 172 039	1.21%	13.49%	2 389 404	8.40%	44.53%
<i>Continuous</i>						
C Not present in 1st or last quarter	59 990 419	62.08%		6 347 998	22.33 %	
F Entire period	1 735 340	1.80%		3 577 269	12.58 %	
L Left-truncated	9 871 084	10.22%		2 721 446	9.57 %	
R Right-truncated	12 001 245	12.42%	86.51%	3 123 522	10.99 %	55.47%
	96 630 146	100.00%	100.00%	28 431 008	100.00 %	100.00%

NOTE: Data covers 1991Q3 through 1999Q4.

Table 3: Longest continuous job spell

(in quarters)	Percent	Cumul. Percent
1	35.96 %	
2	20.42 %	56.38%
3	9.74 %	
4	6.05 %	72.17%
5	4.23 %	
6	3.39 %	
7	2.58 %	
8	2.12 %	84.49%
more than 8	15.51 %	100.00%

Table 4: Re-assignment of SSN(0), by UID, in Stage 1

	Frequency	Percent
SSN has unique UID (out-of-scope)	14 042 405	24.47%
SSN(0) of UID not reassigned	40 636 312	70.80%
SSN(0) of UID reassigned	2 715 054	4.73%
	57 393 771	100.00

Table 5: Match rates: Stage 2

Year	Holes	Match pairs	Fraction patched
	(a)	(b)	(c)
Across all match passes, by year			
1991	127869	30743	24 .04%
1992	507335	101874	20 .08%
1993	423721	93337	22 .03%
1994	496937	147142	29 .61%
1995	489793	109978	22 .45%
1996	456878	98804	21 .63%
1997	439520	60804	13 .83%
1998	536123	112325	20 .95%
1999	464643	78708	16 .94%
All	3942819	833715	21 .15%

Notes: (c)= (b)/(a)

Table 6: Comparing job histories before and after editing process

Pattern in job history	Original data		Edited data		Change	
	(a)	(b)	(c)	(d)	(e)	(f)
	Freq.	Percent	Freq.	Percent	Freq.	Percent
<i>Non-continuous,</i>						
<i>length of longest interruption</i>						
1 quarter	5 315 869	5.50%	4 710 673	4.87%	-605 196	-11.38%
2 quarters	2 357 942	2.44%	2 359 374	2.44%	1 432	0.06%
3 quarters	1 764 701	1.83%	1 755 814	1.82%	- 8 887	- 0.50%
4 quarters	750 910	0.78%	747 707	0.77%	- 3 203	- 0.42%
5 quarters	532 174	0.55%	529 777	0.55%	- 2 397	- 0.45%
6 quarters	466 301	0.48%	463 878	0.48%	- 2 423	- 0.51%
7 quarters	430 549	0.45%	429 179	0.44%	- 1 370	- 0.31%
8 quarters	241 573	0.25%	240 214	0.25%	- 1 359	- 0.56%
9 or more quarters	1 172 039	1.21%	1 163 420	1.20%	- 8 619	- 0.73%
<i>Continuous</i>						
C Continuous	59 990 419	62.08%	60 311 626	62.37%	321 207	0.53%
F Entire period	1 735 340	1.80%	1 807 775	1.87%	72 435	4.17%
L Left-truncated	9 871 084	10.22%	10 032 149	10.37%	161 065	1.63%
R Right-truncated	12 001 245	12.42%	12 144 959	12.56%	143 714	1.19%
	96 630 146	100.00%	96 696 545	100.00%	66 399	0.06%

NOTE: For definitions of job history patterns, see text on page 3.

Short name	Long name
A	Accessions
B	Beginning-of-period employment
E	End-of-period employment
\bar{E}	Average employment
F	Full-quarter employment
FJF	Net change in full-quarter employment
H	New hires
H3	Full-quarter new hires
JF	Net job flows
R	Recalls
S	Separations
NA	Periods of non-employment for accessions
NH	Periods of non-employment for new hires
NR	Periods of non-employment for recalls
NS	Periods of non-employment for separations
W1	Total payroll of all employees
W2	Total payroll of end-of-period employees
W3	Total payroll of full-quarter employees
WA	Total payroll of accessions
WS	Total payroll of separations

Table 7: Name mapping for variables used in aggregated analysis

A variable will be named *VARNAME_GA* where *VARNAME* is defined in Table 7, and *G* and *A* are defined as follows:

G: Gender		A: Age	
0	All	0	All
F	Female	1	14-18
M	Male	2	19-21
		3	22-24
		4	25-34
		5	35-44
		6	45-54
		7	55-64
		8	65+

Table 8: Demographic group definitions

Table 9: Distribution of percent bias in aggregate statistics
No demographics, Unit x Quarter cells

Variable (bias)	Unit	Mean	Std	N	P10	P50	P90
pA	Firm	2.17%	13.98%	11755355			
pA	County	1.56%	1.01%	2006	0.62%	1.42%	2.64%
pA	Industry	1.97%	2.29%	374	0.51%	1.47%	3.40%
pB	Firm	-0.74%	6.14%	20717508			
pB	County	-0.46%	0.31%	1947	-0.75%	-0.45%	-0.25%
pB	Industry	-0.31%	0.31%	363	-0.59%	-0.34%	-0.14%
pE	Firm	-0.74%	6.14%	20717507			
pE	County	-0.47%	0.31%	1947	-0.75%	-0.45%	-0.25%
pE	Industry	-0.30%	0.33%	363	-0.59%	-0.34%	-0.13%
p \bar{E}	Firm	-0.71%	5.29%	21954411			
p \bar{E}	County	-0.46%	0.30%	1947	-0.74%	-0.46%	-0.26%
p \bar{E}	Industry	-0.31%	0.30%	363	-0.57%	-0.34%	-0.15%
pF	Firm	-1.23%	8.05%	18454708			
pF	County	-0.78%	0.36%	1888	-1.21%	-0.74%	-0.43%
pF	Industry	-0.53%	0.31%	352	-0.90%	-0.53%	-0.24%
pH	Firm	1.18%	10.18%	9784872			
pH	County	0.94%	0.80%	1888	0.31%	0.81%	1.63%
pH	Industry	1.43%	2.79%	352	0.28%	0.82%	2.45%
pH3	Firm	-0.94%	8.42%	6233024			
pH3	County	-0.77%	0.63%	1770	-1.30%	-0.71%	-0.30%
pH3	Industry	-0.25%	2.54%	330	-1.01%	-0.52%	0.04%
pR	Firm	4.71%	26.86%	3242186			
pR	County	5.26%	3.61%	1888	1.70%	4.59%	9.18%
pR	Industry	5.95%	3.49%	352	1.93%	5.46%	10.29%
pS	Firm	2.31%	14.29%	11161916			
pS	County	1.66%	1.11%	1947	0.67%	1.46%	2.72%
pS	Industry	2.01%	2.08%	363	0.63%	1.53%	3.41%

For notes, see end of table on page 41.

(cont.)

Table 9 (cont.): Distribution of percent bias in aggregate statistics
No demographics, Unit x Quarter cells

Variable (bias)	Unit	Mean	Std	N	P10	P50	P90
pFJF	Firm	-0.57%	22.01%	9968752			
pFJF	County	-15.68%	675.94%	1886	-11.97%	-0.28%	11.39%
pFJF	Industry	4.77%	44.16%	352	-11.65%	-0.05%	12.86%
pJF	Firm	-0.04%	20.44%	11280086			
pJF	County	-0.27%	44.31%	1945	-7.77%	-0.04%	8.06%
pJF	Industry	1.68%	107.29%	363	-8.83%	0.03%	9.19%
pNA	Firm	2.89%	22.68%	9097310			
pNA	County	1.99%	1.17%	1888	0.81%	1.81%	3.33%
pNA	Industry	2.56%	1.75%	352	0.93%	2.18%	4.44%
pNH	Firm	2.06%	19.75%	8179091			
pNH	County	1.64%	1.11%	1888	0.59%	1.46%	2.77%
pNH	Industry	2.25%	1.92%	352	0.69%	1.80%	4.35%
pNR	Firm	4.57%	29.02%	2562640			
pNR	County	5.09%	3.76%	1888	1.52%	4.42%	8.77%
pNR	Industry	5.52%	3.30%	352	1.90%	5.10%	9.05%
pNS	Firm	2.83%	22.02%	8273801			
pNS	County	2.13%	1.26%	1770	0.85%	1.96%	3.48%
pNS	Industry	2.51%	1.65%	330	0.93%	2.18%	4.15%

For notes, see end of table on page 41.

(cont.)

Table 9 (cont.): Distribution of percent bias in aggregate statistics
No demographics, Unit x Quarter cells

Variable (bias)	Unit	Mean	Std	N	P10	P50	P90
pW1	Firm	-0.01%	4.96%	23229843			
pW1	County	-0.01%	0.15%	2006	-0.05%	-0.02%	0.00%
pW1	Industry	0.04%	0.35%	374	-0.04%	-0.02%	0.08%
pW2	Firm	-0.75%	7.73%	20717507			
pW2	County	-0.45%	0.27%	1947	-0.74%	-0.42%	-0.21%
pW2	Industry	-0.27%	0.33%	363	-0.57%	-0.29%	-0.08%
pW3	Firm	-1.21%	12.16%	18454708			
pW3	County	-0.71%	0.36%	1888	-1.12%	-0.65%	-0.36%
pW3	Industry	-0.46%	0.35%	352	-0.85%	-0.43%	-0.17%
pWA	Firm	15.57%	1111.78%	11755355			
pWA	County	4.92%	3.34%	2006	1.89%	4.38%	8.44%
pWA	Industry	3.95%	4.94%	374	0.77%	3.35%	6.79%
pWS	Firm	18.77%	1094.50%	11161916			
pWS	County	4.87%	3.17%	1947	2.02%	4.31%	8.06%
pWS	Industry	3.64%	4.48%	363	1.00%	3.18%	5.71%

Note: There are a total of 23232068 firm-quarter cells, 2006 county-quarter cells, and 374 industry-quarter cells. Percentiles for firm-quarter cells are all zero, and not reported for simplification.

Table 10: Distribution of level bias
No demographics, Unit x Time cells

Variable (bias)	Unit	Mean	Std	N	P10	P50	P90
dA	County	586	1432	2006	13	172	1305
dA	Industry	3141	3680	374	26	1988	7614
dB	County	-618	1567	2006	-1346	-178	-11
dB	Industry	-3317	4326	374	-8792	-2181	0
dE	County	-637	1585	1947	-1370	-192	-16
dE	Industry	-3418	4557	363	-8796	-2266	-16
$d\bar{E}$	County	-632	1572	1947	-1374	-188	-16
$d\bar{E}$	Industry	-3389	4378	363	-8736	-2332	-8
dF	County	-840	2089	1947	-1799	-242	-18
dF	Industry	-4507	5440	363	-11515	-2907	-30
dH	County	281	700	1888	5	81	633
dH	Industry	1507	2223	352	14	983	3575
dH3	County	-79	198	1888	-180	-22	0
dH3	Industry	-424	1124	352	-1322	-268	0
dR	County	330	797	1888	10	102	711
dR	Industry	1770	1991	352	26	1087	4298
dS	County	603	1426	1947	18	184	1343
dS	Industry	3236	3498	363	55	2101	7614
dFJF	County	-20	447	1947	-109	0	87
dFJF	Industry	-106	1570	363	-824	4	645
dJF	County	-11	399	1947	-112	-1	89
dJF	Industry	-57	1640	363	-1041	6	710
dNA	County	1359	3312	1888	37	406	2987
dNA	Industry	7291	8066	352	121	4534	17603
dNH	County	982	2377	1888	23	280	2177
dNH	Industry	5266	5832	352	82	3396	12687
dNR	County	378	1019	1888	10	110	799
dNR	Industry	2025	2536	352	29	1178	4883
dNS	County	1321	3226	1770	35	388	2828
dNS	Industry	7083	7796	330	117	4367	16894

For notes, see end of table on page 43.

(cont.)

Table 10 (cont.): Distribution of level bias
No demographics, Unit x Time cells

Variable (bias)	Unit	Mean	Std	N	P10	P50	P90
dW1	County	-182007	2113940	2006	-615949	-45498	0
dW1	Industry	-976219	20156108	374	-5671457	-937210	2397960
dW2	County	-4100232	10692303	1947	-9237482	-949286	-73447
dW2	Industry	-21992151	32041187	363	-50304427	-16112377	-225622
dW3	County	-5767583	14958853	1947	-13275143	-1332735	-86395
dW3	Industry	-30935219	39286057	363	-66737552	-21838598	-351640
dWA	County	3861399	10001149	2006	58050	917757	8795154
dWA	Industry	20711138	24981897	374	308185	14529046	46051029
dWS	County	3912211	9912881	1947	76847	973909	8841267
dWS	Industry	20983679	23739414	363	550575	14904299	47933876

Note: 59 counties and 11 SIC divisions.

Table 11a: Regression results, percentage bias
County cells

Dependent Variable	R ²		Parameter Estimate		Standard Error
pA	0.0439	Intercept	0.0228	**	0.0002
		AR	-0.0108	**	0.0003
	0.0853	AR	-0.0107	**	0.0003
		<i>F Test p-value</i>	<0.0001		
pH	0.0215	Intercept	0.0196	**	0.0004
		HR	-0.0217	**	0.0010
	0.0490	HR	-0.0247	**	0.0011
		<i>F Test p-value</i>	<0.0001		
pH3	0.0001	Intercept	-0.0068	**	0.0005
		H3R	-0.0039		0.0032
	0.0103	H3R	-0.0106	**	0.0033
		<i>F Test p-value</i>	<0.0001		
pR	0.1009	Intercept	0.0763	**	0.0007
		RR	-0.3480	**	0.0073
	0.1374	RR	-0.3544	**	0.0086
		<i>F Test p-value</i>	<0.0001		
pS	0.0444	Intercept	0.0262	**	0.0003
		SR	-0.0206	**	0.0007
	0.0958	SR	-0.0221	**	0.0007
		<i>F Test p-value</i>	<0.0001		
pJF	0.0000	Intercept	0.0003		0.0085
		JFR	-0.0017		0.0226
	0.0022	JFR	-0.0016		0.0226
		<i>F Test p-value</i>	0.8361		
pFJF	0.0000	Intercept	-0.0138		0.0165
		FJFR	0.0025		0.0628
	0.0028	FJFR	0.0057		0.0639
		<i>F Test p-value</i>	0.4463		

For notes, see end of table on page 46.

(cont.)

Table 11a (cont.): Regression results, percentage bias
County cells

Dependent Variable	R ²		Parameter Estimate		Standard Error
pNA	0.0547	Intercept	0.0365	**	0.0005
		AR	-0.0381	**	0.0018
		SR	0.0049	*	0.0020
	0.0841	AR	-0.0399	**	0.0018
		SR	0.0044	*	0.0020
			<i>F Test p-value</i>	<0.0001	
pNH	0.0250	Intercept	0.0327	**	0.0007
		AR	-0.0444	**	0.0026
		SR	0.0164	**	0.0028
	0.0420	AR	-0.0462	**	0.0026
		SR	0.0159	**	0.0028
			<i>F Test p-value</i>	<0.0001	
pNR	0.0418	Intercept	0.0739	**	0.0010
		AR	-0.0541	**	0.0037
		SR	-0.0074	*	0.0040
	0.0704	AR	-0.0527	**	0.0037
		SR	-0.0008		0.0040
			<i>F Test p-value</i>	<0.0001	
pNS	0.0372	Intercept	0.0326	**	0.0004
		AR	-0.0013	*	0.0006
		SR	-0.0229	**	0.0013
	0.0715	AR	-0.0012	*	0.0006
		SR	-0.0240	**	0.0014
			<i>F Test p-value</i>	<0.0001	

For notes, see end of table on page 46.

(cont.)

Table 11a (cont.): Regression results, percentage bias
County cells

Dependent Variable	R ²	Parameter	Estimate	Standard Error
pW1	0.0154	Intercept	-0.0010	** 0.0002
		AR	-0.0014	** 0.0003
		SR	0.0095	** 0.0006
	0.0349	AR	-0.0014	** 0.0003
		SR	0.0097	** 0.0007
		<i>F Test p-value</i>	<0.0001	
pW2	0.0103	Intercept	-0.0068	** 0.0002
		AR	-0.0014	** 0.0003
		SR	0.0085	** 0.0007
	0.0523	AR	-0.0015	** 0.0003
		SR	0.0084	** 0.0007
		<i>F Test p-value</i>	<0.0001	
pW3	0.0532	Intercept	-0.0054	** 0.0003
		FJFR	-0.0333	** 0.0010
	0.0789	FJFR	-0.0298	** 0.0010
		<i>F Test p-value</i>	<0.0001	
pWA	0.0222	Intercept	0.0707	** 0.0012
		AR	-0.0265	** 0.0019
		SR	-0.0072	* 0.0037
	0.0448	AR	-0.0257	** 0.0019
		SR	-0.0106	** 0.0039
		<i>F Test p-value</i>	<0.0001	
pWS	0.0166	Intercept	0.0712	** 0.0011
		AR	0.0035	* 0.0017
		SR	-0.0486	** 0.0033
	0.0398	AR	0.0049	** 0.0017
		SR	-0.0565	** 0.0035
		<i>F Test p-value</i>	<0.0001	

Note: Each block represents two regressions, with reported R^2 and coefficients. The first block is estimated by OLS, the second block by OLS on demeaned, where means are taken with respect to the SIC division. The F test reports test score and p-value for joint significance of the implicit SIC division dummies.

Table 11b: Regression results, percentage bias

Dependent Variable	R ²	Industry cells		Standard Error	
		Parameter	Estimate		
pA	0.0467	Intercept	0.0291	**	0.0007
		AR	-0.0149	**	0.0011
	0.1606	AR	-0.0116	**	0.0011
		<i>F Test p-value</i>	<i><0.0001</i>		
pH	0.0301	Intercept	0.0290	**	0.0011
		HR	-0.0345	**	0.0032
	0.1333	HR	-0.0234	**	0.0037
		<i>F Test p-value</i>	<i><0.0001</i>		
pH3	0.0053	Intercept	0.0039	**	0.0015
		H3R	-0.0465	**	0.0106
	0.0365	H3R	-0.0297	**	0.0107
		<i>F Test p-value</i>	<i><0.0001</i>		
pR	0.1360	Intercept	0.0789	**	0.0013
		RR	-0.4003	**	0.0165
	0.2015	RR	-0.4699	**	0.0264
		<i>F Test p-value</i>	<i><0.0001</i>		
pS	0.0591	Intercept	0.0298	**	0.0006
		SR	-0.0228	**	0.0014
	0.2581	SR	-0.0146	**	0.0017
		<i>F Test p-value</i>	<i><0.0001</i>		
pJF	0.0000	Intercept	0.0390		0.0423
		JFR	-0.0228		0.1167
	0.0030	JFR	-0.0238		0.1167
		<i>F Test p-value</i>	<i>0.3018</i>		
pFJF	0.0000	Intercept	-0.0061		0.0219
		FJFR	0.0043		0.1149
	0.0018	FJFR	0.0033		0.1150
		<i>F Test p-value</i>	<i>0.7368</i>		

For notes, see end of table on page 49.

(cont.)

Table 11b (cont.): Regression results, percentage bias
Industry cells

Dependent Variable	R ²		Parameter Estimate		Standard Error
pNA	0.1097	Intercept	0.0436	**	0.0008
		AR	-0.0453	**	0.0051
		SR	0.0062		0.0056
	0.2305	AR	-0.0455	**	0.0048
		SR	0.0171	**	0.0056
		<i>F Test p-value</i>	<0.0001		
pNH	0.0684	Intercept	0.0424	**	0.0010
		AR	-0.0520	**	0.0066
		SR	0.0146	*	0.0072
	0.1807	AR	-0.0528	**	0.0062
		SR	0.0289	**	0.0073
		<i>F Test p-value</i>	<0.0001		
pNR	0.0932	Intercept	0.0742	**	0.0014
		AR	-0.0251	**	0.0087
		SR	-0.0426	**	0.0096
	0.1417	AR	-0.0214	*	0.0085
		SR	-0.0400	**	0.0100
		<i>F Test p-value</i>	<0.0001		
pNS	0.0540	Intercept	0.0357	**	0.0007
		AR	-0.0024	*	0.0013
		SR	-0.0205	**	0.0024
	0.1842	AR	-0.0050	**	0.0013
		SR	-0.0051	*	0.0029
		<i>F Test p-value</i>	<0.0001		

For notes, see end of table on page 49.

(cont.)

Table 11b (cont.): Regression results, percentage bias
Industry cells

Dependent Variable	R ²	Parameter	Estimate	Standard Error	
pW1	0.0026	Intercept	0.0015	0.0012	
		AR	0.0010	0.0023	
		SR	0.0073	*	0.0040
	0.0184	AR	-0.0016		0.0023
		SR	0.0221	**	0.0051
		<i>F Test p-value</i>	<0.0001		
pW2	0.0053	Intercept	-0.0020	**	0.0003
		AR	0.0014	*	0.0006
		SR	-0.0045	**	0.0010
	0.0576	AR	0.0009		0.0006
		SR	-0.0019		0.0012
		<i>F Test p-value</i>	<0.0001		
pWA	0.0014	Intercept	0.0602	**	0.0044
		AR	-0.0130		0.0086
		SR	-0.0015		0.0154
	0.0167	AR	-0.0205	*	0.0088
		SR	0.0406	*	0.0194
		<i>F Test p-value</i>	<0.0001		
pWS	0.0002	Intercept	0.0546	**	0.0064
		AR	0.0103		0.0123
		SR	-0.0125		0.0221
	0.0118	AR	0.0017		0.0126
		SR	0.0350		0.0278
		<i>F Test p-value</i>	<0.0001		
pW3	0.0001	Intercept	-0.0052	**	0.0002
		FJFR	0.0006		0.0011
	0.0493	FJFR	0.0005		0.0011
		<i>F Test p-value</i>	<0.0001		

Note: Each block represents two regressions, with reported R^2 and coefficients. The first block is estimated by OLS, the second block by OLS on demeaned, where means are taken with respect to the SIC division. The F test reports test score and p-value for joint significance of the implicit SIC division dummies.