The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators
Preliminary - Do not cite or quote

John M. Abowd, Bryce E. Stephens and Lars Vilhuber,

with Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock[1]

April 7, 2005

[1]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. This document is based in part on a presentation first given at the NBER Summer Institute Conference on Personnel Economics, 2002, by John Abowd, Paul Lengermann, and Lars Vilhuber. 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 Grant R01 AG018854-02, 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, Cornell University, 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 supports external researchers' use of these data through the Research Data Centers (see www.ces.census.gov). For other questions regarding the data, please contact Jeremy S. Wu, Director, U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA. (Jeremy.S.Wu@census.gov http://lehd.dsd.census.gov).

Abstract

The Longitudinal Employer Household Dynamics program at the U.S. Census Bureau, with funding from several national funding agencies, has built a set of infrastructure files using administrative data provided by state agencies, enhanced with information culled from demographic and economic (business) surveys and censuses. The LEHD Infrastructure Files provide a detailed and comprehensive picture of workers, employers, and their interaction in the U.S. economy. Building on this infrastructure, the Quarterly Workforce Indicators (QWI), a new dataseries published since 2003 by the U.S. Census Bureau, are computed. The QWI offer unprecedented detail on the local dynamics of labor markets. Despite the fine detail, confidentiality is maintained due to the application of state-of-the-art confidentiality protection methods. This article describes how the input files are compiled and combined to create the infrastructure files. The multiple imputation mechanisms that are used to fill in missing data, and the statistical matching techniques used to combine data where a direct match is not possible are both crucial to the success of the final product, and described in detail here. Finally, special attention is paid to the confidentiality protection mechanisms used to hide the identity of the underlying entities in the final published data. A brief description of public-use and restricted-access data files is also provided, with pointers to further documentation for researchers interested in using these data.

Contents

List of Tables

List of Figures

Section 1

Introduction

Since 2003, U.S. Census Bureau has published a new and novel statistical series: the Quarterly Workforce Indicators (QWI). Compiled from administrative records data collected by a large number of states for both jobs and firms, and enhanced with information culled from other data sets at the Census Bureau, these statistics offer unprecedented detail on the local dynamics of labor markets. Despite the fine detail, confidentiality is maintained due to the application of state-of-the-art protection methods.

The underlying data infrastructure was designed by the Longitudinal Employer-Household Dynamics Program at the Census Bureau (Abowd et al.; 2004). Although the QWI are the flagship statistical product published from the LEHD infrastructure files, the latter have found a much more widespread application. The infrastructure constitutes an encompassing and almost universal data source for individuals and firms of all 31 currently participating states.[1]

In this article, we describe the primary input data underlying the the LEHD Infrastructure Files, the methods by which the Infrastructure Files are compiled, and how these files are integrated to create the Quarterly Workforce Indicators. We also provide details about the statistical models used to improve the basic administrative data, and describe enhancements and limitations imposed by both data and legal constraints. Some of the infrastructure and derivative microdata files have recently been made available within the Research Data Centers of the U.S. Census Bureau, and we point out these files during the discussion.

The QWI use a bewildering array of data sources, both from administrative records and from survey and census data. The Census Bureau receives UI wage records and ES-202 establishment records from each state participating in the program. The Bureau then uses these products to integrate information about the individuals (place of residence, sex, birth date, place of birth, race, education) with information about the employer (place of work, industry, employment, sales). Not all of the integration methods are straight one-to-one matches. In some cases, statistical matching techniques are used, and in others variable values are imputed. Throughout, critical imputations are done multiple times, improving the precision of the final estimates.

It should be noted that the data integration is a two-way street. Not only do the Census Bureau's

[1]The number of participating states still increases regularly as new Memoranda of Understanding are signed and new states begin shipping data. As of April 5, 2005, there are 31 states in production (shipping data to Census) of which 27 are available at http://lehd.dsd.census.gov/

surveys and censuses improve the detail on the administrative files: As a part of its Title 13 mission, the Census Bureau uses the integrated files to in turn improve the Census Bureau's demographic surveys, like the Current Population Survey, the Survey of Income and Program Participation, and the American Community Survey. They are also used to improve the Census Bureau's Business Register, which is the sampling frame for all its economic data and the initial contact frame for the Economic Census.

We give an overview of the different raw data inputs and how they are treated and adjusted in Section 2. In a system that focusses on the dynamics at the individual and firm level, proper identification of the entities is important, and we briefly highlight the steps undertaken to this end. A more detailed analysis of the probabilistic editing of individual record has been published elsewhere (Abowd and Vilhuber; 2005). The raw data are then aggregated and standardized into a series of component files, which we call the "Infrastructure Files", as described in Section 3.1. Finally, Sections 4 and 5 illustrate how they are brought together to create the QWI statistics. It will soon become clear to the reader that the level of detail potentially available with these statistics requires special attention to the confidentiality of the the underlying entities. How their identity is protected is described in Section 6. Many of the files described in this paper are accessible in either a public-use or restricted-access version, and a brief description with pointers to more detailed documentation is provided in Section 7. Section 8 concludes and provides a glimpse at the ongoing research into improving the infrastructure files.

We should note that this paper has far too *few* authors. Over the years, many individuals have contributed to the creation of these files. [complete list here]

Section 2

Input files

The underlying data are wage records extracted from UI administrative files from each participating state, as well as from the (typically independently created) files from the Quarterly Census of Employment and Wages (QCEW, formerly known as 'ES-202'). These data are received by LEHD on a quarterly basis, with historical time series extending back to the early 1990s for some states.

2.1 Wage records: UI

Wage records correspond to the report of an individual's UI-covered earnings by an employing entity, identified by a state UI account number (called State Employer Identification Number, or 'SEIN' in the LEHD system). An individual's UI wage record appears if at least one employer reports earnings of at least one dollar for that individual during the quarter. Thus, the job must produce at least one dollar of UI-covered earnings during a given quarter to count in the LEHD system. Maximum earnings reported are defined in a specific state's unemployment insurance system, and observed top-coding varies across states and time.

A record is completed with information on the individual's Social Security Number (later masked within the LEHD system), first name, last name, and middle initial. A few states include additional information: the firm's reporting unit, or establishment (called SEINUNIT in the LEHD system), available for Minnesota, and a crucial component to the Unit-to-Worker impute described later; weeks worked, available for some years in Florida; hours worked, available for Washington state.

2.2 Employer reports: ES202

The employer reports are based on information from each state's Department of Employment Security. The data are collected as part of the Covered Employment and Wages (CEW) program, also known as the ES-202 program, which is administered by the U.S. Bureau of Labor Statistics (BLS). This cooperative program between the states and the federal government collects employment, payroll, and location information from employers covered by state unemployment insurance programs. These are the same records that form the basis of the Quarterly Census of Employment

and Wages (QCEW), but are referred to in the LEHD system by their old acronym "ES202". The fundamental unit of these files is a 'reporting unit', typically taken to be equivalent to a 'establishment'. Most firms only have one establishment ('single-units'), but most employment is in firms with multiple establishments ('multi-units'). One report per establishment per quarter is filed[1].

The information contained in these files has increased substantially over the years. Employers report wages subject to statutory payroll taxes on this form, together with some other information. Common to all years, and critical to LEHD processing, are information on the employer's identity (the SEIN), the reporting unit's identify (SEINUNIT), ownership information, employment on the 12th of each month covered by the quarter, and total wages paid over the course of the quarter. Additional information pertains to industry classifications (initially SIC, and later NAICS). Other information include the federal EIN, geography both at a high level (county or MSA) and low level (street address). A recent expansion of the record layout has increased the informational content substantially.

2.3 Administrative demographic information: PCF

The UI and ES202 files are the core data files describing the economic activity of individuals and firms. Although combined, these files contain a tremendous amount of detail on the economic activity, they contain little or no demographic information on the individuals. This information comes from a third administrative data source, compiled by ARRS/PRED. The Person Characteristics File (PCF) contains information on gender, date of birth, place of birth, citizenship, and race, most of which is extraced in turn from the Social Security Administration's Numident file. Other information contains place of residence for several years culled from other administrative sources.

2.4 Demographic products

Many individuals have appeared in at least one of the eligible Census demographic products, and their detailed demographic information from the 1984, 1990-1993, and 1996 SIPP panels as well as from March Demographic Supplement to the Current Population Survey (CPS) from 1983 onwards can be linked to the extensive longitudinal data gleaned from the state records. They are used by the ICF.

2.5 Economic censuses and annual surveys

These data include the complete 1987, 1992 and 1997 economic censuses, all annual surveys of manufacturing, service, trade, transportation and communication industries and selected, approved fields from the Census Bureau's Business Register.

Linkage to these data is based upon exact EIN matches, supplemented with statistical matching to recover establishments.

[1] These data are also used to compile the Business Employment Dynamics (BED) data at the BLS.

2.6 Identifiers and their longitudinal consistency

Both the wage records and employer reports are administrative data - comprehensive, but sometimes less than perfect. In particular, spurious changes in the numerical entity identifers used for longitudinal matching can have a significant impact on most economic uses of the data.

2.6.1 Scope of data and identifiers

In the LEHD system, a person is identified initially by the Social Security Number, and later by their Protected Identification Key (PIK). This identifier is national in scope, and individuals can be tracked across all states and time periods. Not all individuals are in-scope at all times. To be included in the wage record database, an individual's job must be covered by a state's unemployment insurance system. The prime exclusions are agriculture and to some extent the public sector. Coverage varies across states and time, although on average, 98% of all private-sector jobs are covered. Stevens (2002) provides a survey of coverage for a subset of the current participant states in the LEHD system.

A 'firm' is identified primarily by their state UI account number (SEIN). A single legal "firm" might have multiple SEINs but regardless of its operations in other states a legal firm has a different unemployment insurance account in each state in which it has statutory employees. In particular the QWI are based exclusively on SEIN-based entities and their associated establishments. Since a SEIN is specific to a state, the QWI cannot account for movements of individuals across state lines, but within the same company. Time-consistency is also not guaranteed, since the tax number associated with a firm can change (see later discussions). Again, the coverage caveat mentioned also applies.

The restriction to SEIN does not apply to the Infrastructure Files. For some states, the federal Employer Identification Number, used for federal tax purposes, is available, and reported on the Employer Characteristics File (ECF). Links to the Census Business Register (BR) allow to map entities from the QCEW to larger companies across state lines (see Section 7.2.3 for more information on the Business Register Bridge).

2.6.2 Error correction of person identifiers

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

Errors can be random digit coding errors that do not persist, typically generated when data are transferred from one format (paper) to another (digital), or can be persistent, typically occurring when a firm's payroll system contains an erroneous SSN. While the latter is hard to identify and to correct, the LEHD system uses statistical matching techniques to correct for spurious and non-persistent coding errors. Both the incidence of errors and the success rate of the error correction

methods differs widely by state. In particular, it depends critically on the availability of name information on the wage records.

Abowd and Vilhuber (2005) describe and analyze the process as it was applied to data provided by the state of California. The process verified over half a billion records. The number of records that are recoded is slightly less than 10 percent of the total number of unique individuals appearing in the original data, and only a little more than 0.5% of all wage records. The authors estimate that the true error rate in their data is higher, in part due to the conservative setup of the process. Over 800,000 job history interruptions in the original data are eliminated, representing 0.9% of all jobs, but 11% of all interrupted jobs. Despite the small number of records that are found to be miscoded, the impact on flow statistics can be large. Accessions in the uncorrected data are overestimated by 2%, and recalls are biased upwards by nearly 6%. Payroll for accessions and separations are biased upward by up to 7 percent.

The wage record editing occurs prior to the construction of any of the Infrastructure Files, for two reasons. First, the wage record edit process requires access to the original Social Security Numbers as well as to the names on the wage records, both of which are replaced by Personal Identifier Numbers (PIK) or stripped off very early in the processing of wage records. The wage record editing process takes place in a secure and separate area from the rest of the LEHD processing, to avoid any commingling of SSN-laden data with anonymized data. Second, because the identifier changes underlying the wage record edit are deemed spurious, and because individuals have no economic reason at all to change Social Security Numbers, there is little ambiguity about the applicability of the edit. This is different from the editing of firm identifiers (see the next section).

2.6.3 Correcting for changes in firm identifiers

Firms in the QCEW system are identified by a (UI tax) account number attributed by the state. As with all firm identities, an account number can change for a number of reasons over time, not all of which are distinguishable economic entities for the purpose of these statistics. State administrative units take great care to follow the legal entities in their system, but account numbers may nevertheless change for reasons which economists may not consider legitimate economic reasons. For instance, a simple change in ownership of a firm may lead to a change in the account number.

Because changes in the firm identifiers are correlated with some elements of economic choice, albeit imperfectly, they are not imposed on the entire LEHD Infrastructure Files. Rather, an auxiliary file, the Successor-Predecessor File, is created that allows for the selective application of such edits. This file is produced after the first of the Infrastructure Files have been created, and is described later in this document.

Section 3

Infrastructure files

Once received, the UI and ES202 files are standardized.[1]The UI files have been edited for longitudinal consistency, and the SSN replaced by the Protected Identification Key (PIK). Beyond that, no further processing has occurred.

The core Infrastructure Files are built from the core input files, and augmented from a large number of additional Census-internal demographic and economic (firm) surveys and censuses. The Employment History File (EHF) provides a full time-series of earnings at all within-state jobs for all time periods covered by the LEHD data, and activity calendars at a job, SEINUNIT, and SEIN level. The Individual Characteristics File (ICF) provides time-invariant personal characteristics and some address information.[2]The Employer Characteristics File (ECF) provides a complete database of firm and establishment characteristics, most of which are time-varying. It includes a subset of the data available on the Geocoded Address List (GAL), which contains geocoded at the block-level and latitude/longitude coordinates for addresses from a large set of administrative and survey data. We will describe each in detail.

3.1 Employment History File: EHF

The *Employment History File* (EHF) is designed to store the complete in-state work history for each individual that appears in the UI wage records. The EHF for each state contains one record for each employee-employer combination – a job – in that state in each year. Both annual and quarterly earnings variables are available in the EHF. Individuals who never have strictly positive earnings (a theoretical possibility) are dropped.

A re-ordering of the data into one observation per job, with all quarterly earnings and activity records available within one record, is also available (Person History File, PHF). Activity is defined as active employment within a quarter, requiring a strictly positive value for quarterly earnings. A similar time-series of activity at the SEINUNIT level (UNIT History File, UHF) and the SEIN level (SEIN History File, SHF) is also computed at this time.

[1]The ES202 files in particular have been received in a bewildering array of physical file layouts and formats, reflecting the wide diversity in computer systems installed in state agencies.

[2]A time-varying variant of the ICF is under development.

A comparison of the earnings and employment information from the UI and ES202 files is one of the core quality measures that are computed. Large discrepancies are highlighted, and clarified with the data provider. Often, a corrected data file can be imported into the LEHD system. Not all data discrepancies can be easily resolved. In particular the historical data sometimes are not correctable, because the data has been lost or corrupted.[3]

3.2 Individual Characteristics File: ICF

The *Individual Characteristics File* (ICF) for each state contains one record for every person who is ever employed in that state over the time period spanned by the state's unemployment insurance records.

The ICF is constructed in the following manner. First, the universe of individuals is defined by compiling the list of unique PIKs from the EHF. Demographic information from the PCF is then merged on by PIK, and records without a valid match flagged. PIK-survey identifier crosswalks link the CPS and SIPP ID variables into the ICF, and gender and age information from the CPS is used to complement and verify the PCF-provided information.

3.2.1 Age and gender imputation

Approximately 3% of the PIKs found in the UI wage records do not match to the PCF file. Multiple imputation methods are used to assign date of birth and gender to these individuals. To impute gender, the probability of being male is estimated using a state-specific logit model:

$$P(male) = f(X_{is}\beta_s) \tag{3.1}$$

where X_{is} contains a full set of yearly log earnings and squared log earnings, and full set of employment indicators covering time period spanned by the state's records, for each individual i with strictly positive earnings within state s and non-missing PCF gender. The state-specific $\hat{\beta}_s$ as estimated from Equation (3.1) is then used to predict the probability of being male for individuals with missing gender within state s, and gender is assigned as

$$\text{male if } X_{is}\hat{\beta}_s \geq \mu_l \tag{3.2}$$

where $\mu_l \sim U[0,1]$ is one of $l = 1, \ldots, 10$ independent draws from the distribution. Thus, each individual with missing gender is assigned ten independent implicates.

The imputation of date of birth is done in a similar fashion using a multinomial logit to predict the probability of being in one of eight age categories and then assigning an age based on this probability and the distribution of ages within the category. Again, the imputation occurs ten times.

It should be noted that if an individual is missing gender or age in the PCF, but not in the CPS, then the CPS values are used, not the imputed values. Also, before the imputation model for

[3] A future extension currently being developed will allow to apply imputation models to correct for large discrepancies.

date of birth is implemented, basic editing of the date of birth variable takes place to account for obvious coding errors, such as a negative age at the time when UI earnings is first reported for the individual. In those relatively rare cases where the date of birth information is deemed unrealistic it is set to missing and instead imputed based on the model described above.

3.2.2 Place of residence imputation

Place of residence information on the ICF is derived from the StARS (Statistical Administrative Records System), which for the vast majority of the individuals found in the UI wage records contains information on the place of residence down to the exact geographical coordinates. However, in some 10 percent of all cases this information is incomplete or missing. In particular the QWI computation relies on completed place of residence information is because this is a key conditioning variable in the unit-to-worker (U2W) imputation model (see Section 4.2).

County of residence is imputed based on a categorical model of data that can be represented by a contingency table. In particular, separately for each state, unique combinations of categories of gender, age, race, income and county of work are used to form $i = 1, \ldots, I$ populations. For each sample i, the probability of residing in a particular county as of 1999, π_{ij}, is estimated by the sample proportion, $p_{ij} = n_{ij}/n_i$, where $j = 1, \ldots, J$ indexes all the counties in the state plus an extra category for out-of-state residents.

County of residence is then imputed based on

$$county = j \quad \text{if} \quad P_{ij-1} \leq u_k < P_{ij} \tag{3.3}$$

where P_i is the CDF corresponding to p_i for the ith population and $\mu_{kl} \sim U[0, 1]$ is one of $k = 1, \ldots, 10$ independent draws for the lth individual belonging to the ith population.

In its current version no geography below the county level is imputed and in those cases where exact geographical coordinates are incomplete the centroid of the finest geographical area is used. Thus, in cases where no geography information is available this amounts to the centroid of the imputed county. Geographical coordinates are not assigned to individuals whose county of residence has been imputed to be out-of-state.

3.2.3 Education imputation

The imputation model for education relies on a statistical match between the Decennial Census 1990 and LEHD data. The probability of belonging to one of 13 education categories is estimated using 1990 Decennial data conditional on characteristics that are common to both Decennial and LEHD data, using a state-specific logit model:

$$P(educat) = f(Z_{is}\gamma_s) \tag{3.4}$$

where Z_{is} contains age categories, earnings categories, and industry dummies for individuals age 14 and older in the 1990 Census Long Form residing in the state being estimated, and who reported strictly positive wage earnings.

Education is then imputed based on

$$\text{educat} = j \quad \text{if} \quad cp_{j-1} \leq \mu_l < cp_j \tag{3.5}$$

where $cp_j = Z_{is}\hat{\gamma}_s$ and $\mu_l \sim U[0,1]$ is one of $l = 11, \ldots, 20$ independent draws, and $i \in EHF$.

3.3 The Employer Characteristics File: ECF

The Employer Characteristics File (ECF) consolidates most firm level information (size, location, industry, etc.) into two easily accessible files. The firm or SEIN-level file contains one record for every year-quarter an SEIN is present in either the ES-202 or the UI, with more detailed information available for the establishments of multi-unit SEINs in the SEINUNIT-level file. The SEIN file is built up from the SEINUNIT file and contains no additional information, but should be viewed merely as an easier and/or more efficient way to access SEIN level data.

A number of inputs are used to build the ECF. The ES202 data is the primary input to the ECF file creation process. UI data is used to supplement information on the ES202, in particular SEIN-level employment. UI data is also used to extend published BLS county-level employment data, which is used to construct weights for later use in the QWI process. Geocoded address information from the GAL file contributes latitude-longitude coordinates of most establishments, as well as updated WIB and MSA information. BLS-provided Longitudinal Database (LDB) extracts as well as LEHD-developed imputation mechanisms are used to backfill NAICS information for periods in which NAICS was not collected. Finally, the QWI disclosure mechanism is initiated in the ECF. We will describe in Section 3.3.1, while the details of the NAICS imputation algorithm are described in Section 3.3.3, and the entire disclosure-proofing mechanism described in Section 6.

3.3.1 Constructing the ECF

ECF processing starts by stacking yearly ES202 files. General and state specific consistency checks are then performed. The COUNTY, NAICS, SIC, and EIN data are checked for invalid values. The check for industry codes goes beyond a simple validity check. If a 4-digit SIC code or 5-digit NAICS code is present, but is not valid, then the industry code undergoes a conditional impute based on the first 2 and 3 (SIC) or 3,4 and 5 (NAICS) digits.[4] If the resulting codes are not valid, then the industry code is set to missing, and imputed at a later stage of processing.

Based on the EHF, SEIN-level quarterly employment and payroll totals are computed. UI data is used as an imputation source for either payroll or employment in the following situations:

- if ES202 employment is missing, but ES202 payroll is reported, then UI employment is used.
- if ES202 employment is zero, then UI employment is *not* used, since this may be a correct report of zero employment for an existing SEIN. The situation may arise when bonuses or benefits were retro-actively paid, even though no employees were actively employed.

[4] Both NAICS 1997 and NAICS 2002 are used. The same procedure is later used for LDB data.

- if ES202 payroll is zero and ES202 employment is positive, then UI payroll is used.
- if ES202 payroll and employment are both zero or both missing, then UI payroll and employment are used.

The ES202 data contains a "master" record for multi-unit SEINs, which is removed after preserving information not available in the establishment records. Various inconsistencies in the record structure are also dealt with, such as two records (master and establishment) appearing for a single-unit. Initially, information from the master records is used to impute missing data items for the establishments. A flat prior is used in the allocation process: each establishment is assumed to have equal employment and payroll. This is improved upon later in the process.

The allocation process implemented above (master to establishments) does not incorporate any information on the structure of the SEIN. To improve on this, SEINs that are missing firm structure for some periods, but reported a valid multi-unit structure in other periods, are inspected. The absence of information on firm structure typically occurs when an SEIN record is missing due to a data processing error. A SEIN with a valid multi-unit structure in a previous period is a candidate for structure imputation. The firm structure is then imputed using the last available record with a multi-unit structure. Payroll and employment are allocated appropriately.

From this point on, the firm structure (number of establishments per SEIN) is defined for all periods. Geocoded data from the GAL is incorporated to obtain precise geographic information on all establishments.

Geographic data, industry codes (SIC and NAICS) and EIN data from time periods with valid data are used to fill missing data in other periods for the same establishment (SEINUNIT). If at least one industry variable among the several sources (SIC, NAICS1997, NAICS2002, LDB) has valid data, it is used to impute missing values in other fields. Geography, if still missing, is imputed conditional on industry, if available. Counties with larger employment in a SEINUNIT's industry have a higher probability of being selected.

For SEINs, the (employment and establishment-weighted) modal values of county, industry codes, ownership codes, and EIN are calculated for each SEIN and year-quarter. SEIN-level records with missing data are filled in with data from the closest time period with valid data.

At this point, if an SEIN mode variable has a missing value, then no information was ever available for that SEIN. Additional attention is devoted to industry codes, which are critical for QWI processing. SIC and NAICS are randomly imputed with probability proportional to the statewide share of employment in 4-digit SIC code or 5-digit NAICS code. SIC and NAICS codes with a larger share of employment have a higher probability of selection. If an industry code is imputed, it is done so once for each SEIN and remains constant across time. These industry codes are then propagated to all SEINUNITs as well.

With most data items complete, weights are calculated. These weights are discussed in the section on QWI (Section 5). Furthermore, the disclosure proofing is also prepared at the SEIN and SEINUNIT level. This is discussed in detail in Section 6.

3.3.2 Imputations in the ECF

Many data items, when missing, are imputed. The following is a summary list of such imputations. Imputations can be of two types: algorithmic – data closest in time is copied into the missing data items – and probabilistic – the data is drawn from an empirical distribution, conditional on a maximum of available information.

- Employment and payroll: can be imputed based on information in the SEIN master record, or based on information computed at the SEIN-level from UI data. Imputation is always algorithmic - no employment or payroll is ever imputed through probalistic methods.
- Firm structure (relative size of establishments) can be imputed based on reported firm structure in other periods. Imputation is always algorithmic.
- Geography, industry codes, ownership, and EIN are imputed algorithmically first, if possible.
- Geography, if still missing, is imputed conditional on industry, if available, and unconditionally otherwise. Counties with more employment in an SEINUNIT's industry have a higher probability of being selected.
- Industry codes are imputed probabilistically based on empirical correspondence tables conditional on the same unit's observed other industry data items. For instance, if SIC is missing, but NAICS1997 is available, the relative observed distribution of SIC-NAICS1997 pairs is used to impute the missing data item.
- If all previous imputation mechanisms fail, SIC is imputed unconditionally based on the observed distribution of within-state employment across SIC industries. Once SIC is assigned, the previous conditional imputation mechanisms are again used to impute other industry data items.

3.3.3 NAICS codes on the ECF

Enhanced NAICS variables on the ECF can be differentiated by the source(s) and coding system used in their creation. There are two sources of data – the ES202 and the BLS-created LDB – and two coding systems for NAICS – NAICS1997 and NAICS2002. Every NAICS variable uses at least one source and one coding system.

The ESO (ES202-only) and FNL (final) variables are of primary importance to the user community. The ESO variables use information from the ES202 exclusively and ignore any information that may be available on the LDB. We provide in Section 3.3.3.3 an analysis on why this may be preferred. The FNL variables incorporate information from both the ES202 and the LDB, with the LDB being the primary source. The QWI uses FNL variables for its NAICS statistics. Neither ESO nor FNL variables contain missing values.

3.3.3.1 NAICS algorithm precedence ordering

Four basic sources of industry information are available on the ECF: NAICS and NAICS_AUX as well as SIC from ES202 records, and the LDB-sourced NAICS_LDB codes. The NAICS, NAICS_AUX, and NAICS_LDB, when missing (no valid 6-digit industry code), are imputed based on the following algorithm. SIC is filled similarly. Depending on the imputation used, a $miss$ variable is defined, which is used in building the ESO and FNL variables.

1. Valid 6 digit industry code ($miss = 0$)
2. Imputed code based on first 3,4, or 5 digits when no valid six digit code is available in another period ($miss = 0$)
3. Imputed code based on contemporaneous SIC if SIC changed prior to 2000 ($miss = 1.5$)
4. Valid 6 digit code from another period ($miss = 2$)
5. Valid code from another source (for example if NAICS1997 is missing, NAICS2002 or SIC may be available) ($miss = 3$)
6. Use SEIN mode value ($miss = 5$ if contemporaneous modal value, $miss = 7$ if the modal value stems from another time period)
7. Unconditional impute ($miss = 6$ if only the SEIN-level modal value is imputed unconditionally, $miss = 11$ if the SEIN-level value was unconditionally imputed and propogated to all SEINUNITs.)

3.3.3.2 ESO and FNL variables

The ESO and FNL variables are made up of combinations of the various sources of industry information. The ESO variable uses the NAICS and NAICS_AUX variables as input. Information from the variable with the lowest MISS value is preferred although in case of a tie the NAICS_AUX value is used.

The FNL variable uses the ESO and LDB variables. Information from the variable with the lowest MISS value is preferred although in case of a tie the NAICS_LDB value is used. Keep in mind that although the source of an ESO or FNL variable may be equal to NCS, the actual source can only be ascertained by going back to the original.

3.3.3.3 LDB versus LEHD NAICS backcoding

The LDB algorithm is to some extent a black box and testing has shown that it does a relatively poor job of capturing industry changes of SEINs that occurred during the 1990s. In fact, the LDB appears to be a simple backfill that does not take into account an SEIN's entire SIC history.

Although some of the SIC changes over time may be spurious, an SEIN's SIC code history contains valuable information that we have attempted to preserve in our imputation algorithm.

Overall, the effect of the different approaches is relatively small, since very few SEINs change industry, in particular relative to the proportion of SEINs that change geography.

In the following, we present a summary of research done on a comparison of the ESO and FNL NAICS codes on the Illinois ECF. The LDB-sourced NAICS variable is used for about 85% of the records for Illinois, the rest are filled with information from the ES202. It is unclear why only 85% of ES202 records are in the LDB. The results weighted by employment are about the same suggesting that activity was not a criterion for being included on the LDB.

First and not surprisingly, in later years and quarters (1999+) when NAICS is actively coded by the states, the ESO and FNL codes look almost identical when available.

Second, there is little variation in the LDB NAICS codes over time compared with SIC. Among all of the active SEIN-SEINUNITs over the period covered by the Illinois data, only slightly more than 8% experience at least one SIC change compared with about 1.5% on the LDB. Almost all NAICS code changes occur after 1999. While this is not entirely unexpected, it is something to keep in mind when comparing NAICS_FNL versus SIC or NAICS_ESO employment totals. Many of these changes in industry appear to be real and are not captured on the LDB.

One effect of this is that as we go back in time a larger portion of employment can be found in NAICS_FNL codes that are different than one would expect given the SIC code on the ECF. For example, in 1990 about 13% of employment is in a NAICS_FNL code that is different than what we would expect based on the SIC. By 2001, the proportion of employment that is in a NAICS code outside of the set of possible values predicted by the SIC-NAICS crosswalk falls to 3%. The ES202-based NAICS variable does a better job tracking SIC, since more SIC information is used in putting it together.

The main source of the discrepancy is due to entities that experience a change in their SIC code prior to 2000. The LDB appears to ignore this change, while the ES202-based NAICS variable uses an SIC-based impute for these SEINUNITS. The result is a series that exhibits similar patterns of change over time as SIC, while still preserving the value added in the NAICS codes for entities that did not experience a change.

Also, users should keep in mind that for early years (< 1997) some of the NAICS industries have yet to come into existence. The prevalence of this problem has not been investigated yet.

3.4 The Geocoded Address List: GAL

The Geocoded Address List (GAL) is a data set containing unique commercial and residential addresses in a state geocoded to the Census Block and latitude/longitude coordinates. The file encompasses addresses from the state ES202 data, the Census Bureau's Business Register (BR), the Census Bureau's Master Address File (MAF), the American Community Survey Place of Work file (ACS-POW), and others. Addresses from these source files go through geocoding software (Group1's Code1), address standardizers (Ascential/Vality), and matching software (Ascential/Vality) for unduplication.

The final output consists of the address list and a crosswalk for each processed file-year. The GAL contains each unique address, identified by a GAL identifier called `galid`, its geocodes, a flag for each file-year in which it appears, data quality indicators, and data processing information,

including the release date of the Geographic Reference File (GRF). The GAL Crosswalk contains the ID of each input entity and the ID of its address (`galid`).

3.4.1 Geographic codes and their sources

A geocode on the GAL is constructed as

`FIPS-state(2)||FIPS-county(3)||Census-tract(6)`,

and it uniquely identifies the Census tract in the U.S. The tract is the lowest level of geography recommended for analysis. The Census block within the tract is also available on the GAL, but the uncertainties in block-coding make block-level analysis questionable. However, geocoding to the block allows us to add all the higher-level geocodes to the addresses.

3.4.1.1 Block coding

Block coding is achieved by a combination of geocoding software (Group1's Code1), a match to the MAF, or an imputation based on addresses within the tract. Table 3.1 describes the typical distribution of geocode sources.

Table 3.1: Sources of geocodes on GAL

Value	Typical Percent	Meaning
C	12.20	Code1, or the address matches an address for which Code1 supplied the block code
M	81.86	The MAF - the address is a MAF address or matches a MAF address
E	0.00	The MAF, the street address is exactly the same as a MAF address in the same tract
W	0.03	The MAF, the street address is between 2 MAF addresses on the same block face
O	1.23	Imputed using the distribution of commercial addresses in the tract
S	1.17	Imputed using the distribution of residential addresses in the tract
I	0.01	Imputed using the distribution of mixed-use addresses in the tract
D	0.00	Imputed using the distribution of all addresses in the tract
missing	3.50	Block code is missing
	100.00	

In all states observed so far except California, no address required the 'D' method. That is, almost every tract where an address lacks a block code contains commercial, residential, and mixed-use addresses.

The Census Bureau splits blocks to accommodate changes in political boundaries. Most commonly, these are place boundaries (a place is a city, village, or similar municipality). The resulting block parts are identified by 2 suffixes, each taking a value from A to Z. The GAL assigns the block part directly from the MAF, or by adopting the one whose internal point is closest to the address by the straight-line distance.

The GAL also provides the following components of the geocodes as separate variables, for convenience: FIPS code (5 digits), FIPS state code (the first 2 digits of the FIPS code), FIPS county code within state (the right-most 3 digits of the FIPS code), and Census tract code (a tract within the county, a 6-digit code).

Higher-level geographic codes originate from the Block Map File (BMF). The BMF is an extract of the GRF-C (Geographic Reference File - Codes). All geocodes are character variables. FIPS (Federal Information Processing Standard) codes are unique within the U.S.; Census codes are not. Table 3.2 lists the available higher-level geocodes.

Table 3.2: Higher-level geocodes on GAL

`a_fipsmcd`	5-digit FIPS Minor Civil Division (a division of a county)
`a_mcd`	3-digit Census Minor Civil Division (a division of a county)
`a_fipspl`	5-digit FIPS Place
`a_place`	4-digit Census Place
`a_msapmsa`	Metropolitan-Statistical-Area(4)——Primary-Metropolitan-Statistical-Area(4)
`a_wib`	6-digit Workforce Investment Board area

3.4.1.2 Geographic coordinates

The geographic coordinates of each address available as latitude and longitude with 6 implied decimals. The coordinates are not always as accurate as 6 decimal places implies. An indicator flag of their quality is provided. Table 3.3 provides the typical distribution of codes, which range from 1 (highest quality) to 9 (lowest quality).

Table 3.3: Quality of geographic coordinates

Value	Typical Percent	Meaning
1	80.15	Rooftop or MAF (most accurate)
2	1.59	ZIP4 or block face, block face is certain
3	10.12	Block group is certain
4	4.65	Tract is certain
9	3.50	Coordinates are missing
	100.00	

Variables indicating the source of the geographic coordinates (Block internal point, geocoding software, MAF, or otherwise derived) are also available. Most coordinates are provided by either commercial geocoding software or the MAF.

Finally, a set of flags also indicates, for each year and source file, whether an address appears on that file.

For example, the flag variable `b1997` equals 1 if the address is on the 1997 BR; otherwise it equals 0. If a state partner supplies 1991 ES202 data with no address information, `e1991` will be 0 for all addresses. Typically, between 3 and 6 percent of addresses are present on any given year's ES202 files, between 4 and 10 percent are present on a specific BR year file, and between 80 and 90 percent are present on the MAF. Less than 1 percent of addresses are found on the ACS-POW and AHS data, because these are sample surveys.

3.4.2 Accessing the GAL: the GAL Crosswalks

The GAL Crosswalks allow data users to extract geographic and address information about any entity whose address went into the GAL. Each crosswalk contains the identifiers of the entity, its galid, and sometimes flags. To attach geocodes, coordinates, or address information to an entity, merge the GAL Crosswalk to the GAL by galid, outputting only observations existing on the GAL Crosswalk. Then merge the resulting file to the entities of interest using the entity identifiers. An entity whose address wasn't processed (because it's out of state or lacks address information) will have blank GAL data. Table 3.4 lists the entity identifiers by dataset or survey.

Table 3.4: GAL crosswalk entity identifiers

Dataset	Entity identifier variables	Note
AHS	`control and year`	
ES202	`sein, seinunit, year, and quarter`	$e\_flag = p$ for physical addresses, $e\_flag = m$ for mailing addresses as source of address info
ACS-POW	`acsfileseq, cmid, seq, and pnum.`	
BR	`cfn, year, and singmult`	$singmult$ indicates whether the entity resides in the single-unit (su) or the multi-unit (mu) data set. $b\_flag = P$ if physical address, $b\_flag = M$ for mailing address.
MAF	`mafid and year`	

Section 4

Auxiliary data

The infrastructure files in principle have all the information necessary to compute the QWI statistics. However, to be able to compute statistics at a low level of geographical aggregation, we need to be able to place individuals at specific work locations. And since the focus is on dynamics – flows of workers in and out of firms and workplaces – the economic definition of a workplace and a firm needs to be well-defined.

4.1 Connecting firms intertemporally: the Successor-Predessor File (SPF)

The firm identifier used in all of LEHD's files is a state-specific account number in that state's unemployment insurance system, used in particular to collect payroll taxes. These account numbers, here called 'SEIN', can and do change for a number of reasons, including a simple change in legal form or a merger. Typically, the separation of a worker from a firm is identified by a change in the SEIN on that worker's wage records. If a firm changes SEINs, but makes no other changes, the worker would seem to have left the original firm, when in fact his employment status remains unchanged. Thus, a simple change in account numbers would lead to the observation of a firm closing, when in fact, all workers remain employed.

To identify such events, the Successor Predecessor File (SPF) tracks large worker movements between SEINs. Benedetto et al. (2003) used the SPF for an early analysis in one particular state of the impact of such an exercise. The SPF provides for a variety of link characteristics, based on the number of workers leaving an SEIN, in both absolute and relative terms, and the number of workers entering an SEIN, again in absolute and relative terms.

For the QWI, only the strongest links are used to filter out spurious firm identifier changes: If 80% of an SEIN's workers (the predecessor) are observed to move to a single successor, and that successor absorbs 80% of its employees from a single predecessor, then all flows between those two account numbers are filtered out, and treated as if they had never existed.

An evaluation of the impact of the SPF on the aggregate QWI statistics is currently under way.

Of importance to the Unit-to-Worker impute described in the next section is a similar measure, computed within an SEIN. For most states and firms within states, the breakout of units into SEI-

NUNITs is at the discretion of the firm, and the firm may decide to change such a breakout. Again, by following groups of workers as they move between SEINUNITs, spurious intra-SEIN flows can be detected.

4.2 Allocating workers to workplaces: Unit-to-worker impute (U2W)

Early versions of the QWI (then called the Employment Dynamics Estimates, EDE), were computed only at the SEIN level, with employment allocated to a single location per SEIN. This approach was driven by the absence of workplace information on almost all state-provided wage records. Only the state of Minnesota requires the identification of a worker's workplace (SEINUNIT) on a wage records.

A primary objective of the QWI is to provide employment, job and worker flow, and wage measures at a very fine level of geographic (place-of-work) and industry detail. The structure of the administrative data received by LEHD from state partners, however, poses a challenge to achieving this goal. QWI measures are primarily based on the processing of UI wage records which report, with the exception of Minnesota, only the employing firm (SEIN) of workers. The QCEW micro-data, however, are comprised of establishment-level records which provide the level of geographic and industry detail needed to produce the QWI. For firms operating only one establishment, the attachment of establishment-level characteristics is trivial. However, approximately 30 to 40 percent of state-level employment is concentrated in firms that operate more than one establishment. For these multi-establishment firms, the SEIN on workers' wage records identifies the employing firm in the QCEW data, though, not the employing establishment.

In order to attach establishment-level characteristics to workers of multi-establishment firms, a probability model for employment location and imputation was developed. The model explains establishment-of-employment using two key characteristics available in the LEHD data: 1) distance between place-of-work and place-of-residence and 2) the distribution of employment across establishments of multi-establishment firms. The model is estimated using data from Minnesota, where both the firm (SEIN) and establishment identifiers appear on a worker's UI wage record. Then, parameters from this estimation are used to multiply impute establishment-of-employment for workers in the data from other states. Emerging from this process is an output file, called the Unit-to-Worker (U2W) file, containing ten imputed establishments for each worker of a multi-establishment firm. These implicates are then used in the downstream processing of the QWI.

The U2W process relies on information from each of the four infrastructure files – ECF, GAL, EHF, and ICF – as well as the auxiliary SPF file. Within the ECF, the universe of multi-establishment firms is identified. For these firms, the ECF also provides establishment-level employment, date-of-birth, and location (which is acquired from the GAL). The SPF contains information on predecessor relationships which may lead to the revision of date-of-birth implied by the ECF. Finally, individual work histories in the EHF in conjunction with place-of-residence information stored in the ICF provide the necessary worker information needed to estimate and apply the imputation model.

4.2.1 A Probability Model for Employment Location

4.2.1.1 Definitions

Let $i = 1, ..., I$ index workers, $j = 1, ..., J$ index firms (SEINs), and $t = 1, ..., T$ index time (quarters). Let R_{jt} denote the number of active establishments at firm j in quarter t, let $\mathfrak{R} = \max_{j,t} R_{jt}$, and $r = 1, ..., \mathfrak{R}$ index establishments. Note the index r is nested within j. Let N_{jrt} denote the quarter t employment of establishment r in firm j. Finally, if worker i was employed at firm j in t, denote by y_{ijt} the establishment at which she was employed.

Let $\mathcal{J}_t$ denote the set of firms active in quarter t, let $\mathcal{I}_{jt}$ denote the set of individuals employed at firm j in quarter t, let $\mathcal{R}_{jt}$ denote the set of active ($N_{jrt} > 0$) establishments at firm j in t, and let $\mathcal{R}^i_{jt} \subset \mathcal{R}_{jt}$ denote the set of active establishments that are feasible for worker i. Feasibility is defined as follows. An establishment $r \in \mathcal{R}^i_{jt}$ if $N_{jrs} > 0$ for every quarter s that i was employed at j.

4.2.1.2 The Probability Model

Let $p_{ijrt} = \Pr(y_{ijt} = r)$. At the core of the model is the probability statement:

$$p_{ijrt} = \frac{e^{\alpha_{jrt} + x'_{ijrt}\beta}}{\sum_{s \in \mathcal{R}^i_{jt}} e^{\alpha_{jst} + x'_{ijst}\beta}} \tag{4.1}$$

where α_{jrt} is a establishment- and quarter-specific effect, x_{ijrt} is a time-varying vector of characteristics of the worker and establishment, and β measures the effect of characteristics on the probability of being employed at a particular establishment. In the current implementation, x_{ijrt} is a linear spline in the (great-circle) distance between worker i's residence and the physical location of establishment r. The spline has knots at 25, 50, and 100 miles.

Using (4.1), the following likelihood is defined

$$p(y|\alpha, \beta, x) = \prod_{t=1}^{T} \prod_{j \in \mathcal{J}_t} \prod_{i \in \mathcal{I}_{jt}} \prod_{r \in \mathcal{R}^i_{jt}} (p_{ijrt})^{d_{ijrt}} \tag{4.2}$$

where

$$d_{ijrt} = \begin{cases} 1 & \text{if } y_{ijt} = r \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

and where y is the appropriately-dimensioned vector of the outcome variables y_{ijt}, α is the appropriately-dimensioned vector of the α_{jrt}, and x is the appropriately-dimensioned matrix of characteristics x_{ijrt}. For α_{jrt}, a hierarchical Bayesian model based on employment counts N_{jrt} is specified.

The object of interest is the joint posterior distribution of α and β. A uniform prior on β, $p(\beta) \propto 1$ is assumed. The characterization of $p(\alpha, \beta|x, y, N)$ is based on the factorization

$$\begin{aligned} p(\alpha, \beta|x, y, N) &= p(\alpha|N)\, p(\beta|\alpha, x, y) \\ &\propto p(\alpha|N)\, p(\beta)\, p(y|\alpha, \beta, x) \\ &\propto p(\alpha|N)\, p(y|\alpha, \beta, x)\,. \end{aligned} \tag{4.4}$$

Thus the joint posterior (4.4) is completely characterized by the posterior of α and the likelihood of y in (4.2). Note (4.2) and (4.4) assume that the employment counts N affect employment location y only through the parameters α.

4.2.1.3 Estimation

The joint posterior $p(\alpha, \beta|x, y, N)$ is approximated at the posterior mode. In particular, we estimate the posterior mode of $p(\beta|\alpha, x, y)$ evaluated at the posterior mode of α. From these we compute the posterior modal values of the α_{jrt}, then, maximize the log posterior density

$$\log p(\beta|\alpha, x, y) \propto \sum_{t=1}^{T} \sum_{j \in \mathcal{J}_t} \sum_{i \in \mathcal{I}_{jt}} \sum_{r \in \mathcal{R}_{jt}^i} d_{ijrt} \left(\alpha_{jrt} + x'_{ijrt}\beta - \log \left(\sum_{s \in \mathcal{R}_{jt}^i} e^{\alpha_{jst} + x'_{ijst}\beta} \right) \right) \tag{4.5}$$

which is evaluated at the posterior modal values of the α_{jrt}, using a modified Newton-Raphson method. The mode-finding exercise is based on the gradient and Hessian of (4.5). In practice, (4.5) is estimated for three firm employment size classes: 1-100 employees, 101-500 employees, and greater than 500 employees, using data for Minnesota.

4.2.2 Imputing Place of Work

After estimating the probability model using Minnesota data, the estimated parameters are applied in the imputation process for other states. A brief outline of the imputation method, as it relates to the probability model previously discussed, is provided in this section. Emphasis is placed on not only the imputation process itself, but also the preparation of input data.

4.2.2.1 Sketch of Imputation Method

Ignoring temporal considerations, 10 implicates are generated as follows. First, using the mean and variance of β estimated from the Minnesota data, we take 10 draws of β from the normal approximation (at the mode) to $p(\beta|\alpha, x, y)$. Next, using QCEW employment counts for the establishments, we compute 10 values of α_{jt} based on the hierarchical model for these parameters. Note these are draws from the exact posterior distribution of the α_{jrt}. The drawn values of α and β are used to draw 10 imputed values of place of work from the normal approximation to the posterior predictive distribution

$$p(\tilde{y}|x, y) = \int \int p(\tilde{y}|\alpha, \beta, x, y)\, p(\alpha|N)\, p(\beta|\alpha, x, y)\, d\alpha d\beta. \tag{4.6}$$

4.2.2.2 Implementation

Establishment Data Using state-level micro-data, the set of firms (SEINs) that ever operate more that one establishment in a given quarter are identified; these SEINs represent the set of ever-multi-establishment firms defined above as the set $\mathcal{J}_t$. For each of these firms, its establishment-level records are identified. For each establishment, latitude and longitude coordinates, which

emerge from GAL processing, parent firm (SEIN) employment, and QCEW first month employment[1] for the entire history of the establishment are retained. Those establishments with positive first-month employment in a given quarter characterize $\mathcal{R}_{jt}$, the set of all active establishments. An establishment date-of-birth is identified and, in most cases, is the first quarter in the QCEW time series in which the establishment has positive first-month employment. For some firms, predecessor relationships are identified in the SPF; in those instances, the establishment date-of-birth is adjusted to coincided with that of the predecessor's.

Worker Data The EHF provides the earnings histories for employees of the ever-multi-establishment firms. For each in-scope job (a worker-firm pair), one observation is generated for the *end* of each job spell, where a job spell is defined as a continuum of quarters of positive earnings for worker at a particular firm during which there are no more than 3 consecutive periods of non-positive earnings[2]. The start-date of the job history is identified as the first quarter of positive earnings; the end-date is the last date of positive earnings[3]. These job spells characterize the set $\mathcal{I}_{jt}$

Candidates Once the universe of establishments and workers is identified, data are combined and a priori restrictions and feasibility assumptions are imposed. For each quarter of the date series, the history of every job spell that *ends in that quarter* is compared to the history of *every* active (in terms of QCEW first month employment) establishment of the employing firm (SEIN). The start date of the job spell is compared to the birth date of each establishment. Establishments that were born after the start of a job spell are immediately discarded from the set of candidate establishments. The remaining establishments constitute the set $\mathcal{R}^i_{jt} \subset \mathcal{R}_{jt}$ for a job spell (worker) at a given firm[4].

Given the structure of the pairing of job spells with candidate establishments, it is clear that within job spell changes of establishment are ruled-out. An establishment is imputed once for each job spell[5], thereby creating no false labor market transitions.

Imputation and Output Data Once the input data are organized, a set of 10 imputed establishment identifiers are generated for each job spell ending in every quarter for which both QCEW and UI wage records exist. For each quarter, implicate, and size class, $s = 1, 2, 3$, the parameters on the linear spline in distance between place-of-work and place-of-residence $\hat{\beta}^s$ are sampled from the normal approximation of the posterior predictive distribution of β^s conditional on Minnesota

[1] In rare instances where no QCEW employment is available, an alternative employment measure based on UI wage record counts may be used.

[2] A new hire is defined in the QWI as a worker who acceeds to a firm in the current period but was not employed by the same firm in any of the 4 previous periods. A new job spell is created if, for example, a worker leaves a firm for 4 or more quarters and is subsequently re-employed by the same firm.

[3] By definition, an end-date for a job spell is not assigned in cases where a quarter of positive earnings at a firm is succeeded by fewer than 4 quarters of non-employment and subsequent re-employment by the same firm.

[4] The sample of UI wage and QCEW data chosen for processing of the QWI is such that the start and end dates are the same. Birth and death dates of establishments are, more precisely, the dates associated with the beginning and ending of employment activity observed in the data. The same is true for the dates assigned to the job spells.

[5] More specifically, an establishment is imputed to a job spell only once within each implicate.

(MN)

$$p(\beta^s|\alpha_{MN}, x_{MN}, y_{MN}) \tag{4.7}$$

The draws from this distribution vary across implicates, but not across time, firms, and individuals. Next, for each firm j at time t, a set of $\hat{\alpha}_{jrt}$ are drawn from

$$p\left(\alpha_{ST}|N_{ST}\right) \tag{4.8}$$

which are based on the QCEW first-month employment totals (N_{jrt}) for all candidate establishments $r_{jt} \subset \mathcal{R}_{jt}$at firm j within the state (ST) being processed. The initial draws of $\hat{\alpha}_{jrt}$ from this distribution vary across time and firms but not across job spells. Combining (4.7) and (4.8) yields

$$\begin{aligned} & p\left(\alpha_{ST}|N_{ST}\right) p(\beta^s|\alpha_{MN}, x_{MN}, y_{MN}) \\ \approx \; & p\left(\alpha_{ST}|N_{ST}\right) p(\beta^s|\alpha_{ST}, x_{ST}, y_{ST}) \\ = \; & p\left(\alpha_{ST}, \beta_{ST}|x_{ST}, y_{ST}, N_{ST}\right), \end{aligned} \tag{4.9}$$

an approximation of the joint posterior distribution of α and β^s (4.4) conditional on data from the state being processed.

The draws $\hat{\beta}^s$ and $\hat{\alpha}_{jrt}$ in conjunction with the establishment, firm, and job spell data are used to construct the p_{ijrt} in (4.1) for all candidate establishments $r \in \mathcal{R}^i_{jt}$. For each job spell and candidate establishment combination, the $\hat{\beta}^s$ are applied to the calculated distance between place-of-residence (of the worker holding the job spell) and the location of the establishment, where the choice of $\hat{\beta}^s$depends on the size class of the establishment's parent firm. For each combination an $\hat{\alpha}_{jrt}$ is drawn which is based primarily on the size (in terms of employment) of the establishment relative to other active establishments at the parent firm. In conjunction, these determine the conditional probability p_{ijrt} of a candidate establishment's assignment to a given job spell. Finally, from this distribution of probabilities is drawn an establishment of employment.

Emerging from the imputation process is a data file containing a set of 10 imputed establishment identifiers for each job spell. In a minority of cases, the model fails to impute an establishment to a job spell. This is often due to unanticipated idiosyncrasies in the underlying administrative data. Furthermore, across states, the proportion of these failures relative to successful imputation is well under 0.5%. For these job spells, a dummy establishment identifier is assigned and in downstream processing, the employment-weighted modal firm-level characteristics are used.

Section 5

Forming Aggregated Estimates: QWI

5.1 What are the QWI statistics?

The Quarterly Workforce Indicators (QWI) provide detailed local statistics for a variety of indicators. Employment, earnings, gross job creation and destruction, and worker turnover are available at different levels of geography, typically down to the county or metro area. At each level of geography, they are available by detailed industry (SIC and NAICS), sex, and age of workers. At the time of writing of this article, QWI statistics for 31 states had been tabulated, and the program is still expanding with the goal of national coverage.

5.2 Computing the statistics

The establishment of the LEHD Infrastructure Files was driven in large part, although not exclusively, by the needs of the QWI statistics. Completed and representative data were and are the primary concern for the QWI. The ICF (Section 3.2) and the ECF (Section 3.3) draw on a large number of data sources, and use a set of imputation procedures, to provide a complete and detailed picture of each economic actor. The ECF also provides the input data for the weighting, which is explained in more detail in Section 5.3. The Wage Record Edit (Section 2.6.2) and the SPF (Section 4.1) apply algorithmic and statistical matching rules to the proper longitudinal linking of entities. The U2W (Section 4.2) completes the picture, by attributing an employing establishment to each individual employed at some point during the time period covered by the multi-unit UI account under which the data were reported.

These data are then combined and aggregated to compute the QWI statistics. The aggregation is a four step process:

1. A "job" – a unique PIK-SEIN-SEINUNIT combination – is identified, and the job's complete activity history – when the worker worked, and when not – recorded. Note that each job history stems from an implicate of the U2W, and is weighted accordingly.

2. Job-level variables are computed as a set of indicators. The computation of each of these variables is described in detail in Section A.2.2.

3. Job-level variables are aggregated to the establishment level (SEINUNIT), using appropriate implicate weights. The aggregation is done using formulae described in Section A.2.3. For many variables, aggregation to the establishment-level is achieved by summing the job-level variables (beginning-of-period employment, end-of-period employment, accessions, new hires, recalls, separations, full-quarter employment, full-quarter accessions, full-quarter new hires, total earnings of full-quarter employees, total earnings of full-quarter accessions, and total earnings of full-quarter new hires). Some aggregate flow variables are computed using the beginning- and end-of-quarter employment estimates for that workplace. Examples are net job flows (see Equation (A.43) in Appendix A.2), average employment (A.44), job creations (A.46) and job destructions (A.48).

   The file created in this step, internally known as the Unit Flow File (UFF_B), is also available in the RDC system, see Section 7.2.2 for details.

4. The variables necessary for disclosure-proofing – SEINUNIT-specific noise infusion called "fuzz factors" – are attached, and the establishment-level file is summed to the desired level of geographic and demographic detail, using the noise-infused values. Some flow variables are computed directly from other aggregated variables (see Section A.2.4). An undistorted version of all aggregates is also created. All aggregations use weights (see Section 5.3).

5. The tables created in the previous step are disclosure-proofed (see Section 6), by comparison with the undistorted version and in comparison with cell counts. If appropriate, items in some cells are suppressed, and noisy estimates are flagged as such.

5.3 Weighting in the QWI

The QWI statistics are weighted to conform, along one dimension, to published BLS QCEW statistics. The fit is, however, not exact, since the weights are applied before statistics are calculated from the noise-distorted data.

When building the ECF, weights are computed such that measured beginning-of-quarter UI employment of in-scope units, when properly weighted, is equal to published QCEW state-wide employment in the first month of the quarter. The overall adjustment factor is calculated for private establishments and later applied to public-sector establishments as well.

Selection and longitudinal linking in the QWI changes the in-scope units somewhat, and a weight-adjustment is recalculated. This weight is then used for all published QWI statistics. For almost all states and periods, the post-disclosure difference between the published QCEW statistic and the appropriate statistic in the QWI system is less than 0.5%.

Section 6

Disclosure-proofing the QWI

Disclosure proofing is the set of methods used by statistical agencies to protect the confidentiality of the identity of and information about the individuals and businesses that form the underlying data in the system. In the QWI system, disclosure proofing is required to protect the information about individuals and businesses that contribute to the UI wage records, the ES-202 quarterly reports, and the Census Bureau demographic data that have been integrated with these sources. There are two layers of and disclosure proofing in the QWI system.

The first layer occurs when workplace-level estimates are aggregated to higher levels. At this stage, the QWI system infuses specially constructed noise into the estimates of all of the workplace-level measures. This noise is designed to have two very important properties. First, for a given workplace, the data are always distorted in the same direction (increased or decreased) by the same percentage amount in every period. Second, the statistical properties of this distortion are such that when the estimates are aggregated, the effects of the distortion cancel out for the vast majority of the estimates.

The second layer of confidentiality protection occurs after the workplace-level measures are aggregated to the higher levels. The data from many individuals and businesses are combined into a (relatively) few estimates. This aggregation helps to conceal the exact information about any of the individuals or businesses that underlie the estimate. At this level of confidentiality protection, some of the estimates turn out to be based on fewer than three persons or firms. These estimates are suppressed. In addition, some of the estimates are based on data that are still substantially influenced by the noise that was infused in the first layer. These estimates are flagged as substantially distorted. Table 6.1 lists the possible flags for these and other cases. Each observation on any one of the published QWI tables has an associated flag.

6.1 Multiplicative noise model

To implement the multiplicative noise model, a random fuzz factor δ_j is drawn for each employer j according to the following process:

Table 6.1: Disclosure flags in the QWI

Flag	Explanation
-2	no data available in this category for this quarter
-1	data not available to compute this estimate
0	no employment in this cell, or no positive denominator (OK to disclose a 0 for sum or count, missing for ratio)
1	OK, fuzzed value released
2	less than 3 employees (value suppressed in publications)
3	less than 3 employers (value suppressed in publications)
4	for ratio and change variables, the value could not be computed because a denominator rounds to zero.
9	data significantly distorted, fuzzed value released

$$p\left(\delta_j\right) = \begin{cases} (b-\delta)/(b-a)^2, \ \delta \in \ [a,b] \\ (b+\delta-2)/(b-a)^2, \ \delta \in [2-b, 2-a] \end{cases}$$

$$F\left(\delta_j\right) = \begin{cases} 0.5 + \left[(b-a)^2 - (b-\delta)^2\right]/\left[2\,(b-a)^2\right], \ \delta \in \ [a,b] \\ \left[(\delta+b-2)^2\right]/\left[2\,(b-a)^2\right], \ \delta \in [2-b, 2-a] \end{cases}$$

where a and b are constants chosen such that $1 < a < b < 2$.[1] This produces a random noise factor centered around 1 with distortion of at least $a-1$ and at most $b-1$.

Fuzzing of totals The δ_j fuzz factor is used to fuzz all employer totals according to the multiplicative formula $B^*_{jt} = \delta_j \times B_{jt}$. Statistics fuzzed by this method are B, E, M, F, A, S, H, R, FA, FH, FS, W_1, W_2, W_3, WFH, NA, NH, NR, and NS.

Fuzzing of averages of magnitude variables The fuzzed totals are used to construct the following averages: $ZW_2, ZW_3, ZWFH, ZWA, ZWS, ZNA, ZNH, ZNR$, and ZNS. The averages are constructed from fuzzed numerators with unfuzzed denominators according to the formula $ZW^*_{2jt} = \frac{W^*_{2jt}}{E_{jt}} = \frac{\delta_j \times W_{2jt}}{E_{jt}}$.

Fuzzing of differences of counts and magnitudes Fuzzed net job flow is computed at the aggregate (k = geography, industry, or combination of the two for the appropriate age and sex categories) level as the product of the aggregated (unfuzzed) rate of growth and the aggregated fuzzed employment:

[1] The exact numbers are confidential.

$$JF^*_{kt} = G_{kt} \times \bar{E}^*_{kt} = JF_{kt} \times \frac{\bar{E}^*_{kt}}{\bar{E}_{kt}}.$$

This method of fuzzing net job flow will consistently estimate net job flow because it takes the product of two consistent estimators. The formulas for fuzzing gross job creation and job destruction are similar:

$$JC^*_{kt} = JCR_{kt} \times \bar{E}^*_{kt} = JC_{kt} \times \frac{\bar{E}^*_{kt}}{\bar{E}_{kt}}$$

and

$$JD^*_{kt} = JDR_{kt} \times \bar{E}^*_{kt} = JD_{kt} \times \frac{\bar{E}^*_{kt}}{\bar{E}_{kt}}.$$

The same logic was used to fuzz wage changes: total change in earnings for accessions (all jobs), total change in earnings for full-quarter accessions (all jobs), total change in earnings for separations (all jobs), and total change in earnings for full-quarter separations (all jobs). (Symbols used below: $\Delta WA, \Delta WS$.) The unfuzzed total changes were divided by the unfuzzed denominators then multiplied by the ratio of the fuzzed denominator to the unfuzzed denominator for the computation of average change in earnings for accessions (all jobs), average change in earnings for full-quarter accessions (all jobs), average change in earnings for separations (all jobs), and average change in earnings for full-quarter separations (all jobs). (Symbols used below: $Z\Delta WA, Z\Delta WS$.) Averages are fuzzed by multiplying by the ratio of the fuzzed denominator to the true denominator. For example:

$$Z\Delta WA^*_{kt} = \frac{\Delta WA_{kt}}{A_{kt}} \times \frac{A^*_{kt}}{A_{kt}}.$$

6.2 Item suppression

Despite the noise infusion described in the previous sections, some disclosure risk remains for counts based on very few entities in a cell. For counts based on data from fewer than three individuals or employers, the fuzz factors may not provide sufficient protection. This condition applies to the variables B, E, M, F, A, S, H, R, FA, FH, FS, JC, JD, and JF. The QWIs therefore also implement item suppression based on the number of either workers or the number of employers that contribute data for that item in a particular geography $\times$ industry $\times$ age $\times$ sex cell. Because of the noise infusion used previously, however, no complementary suppressions are needed since all of the values based on three or more individuals or employers are adequately protected. Any estimated of the suppressed item computed by subtraction is also protected.

The algorithm for item suppression for the variables B, E, M, F, A, S, H, R, FA, FH, FS, JC, JD, and JF is as follows:

- For the variables listed above, check the conditions leading to a disclosure flag of -2 or -1 (data availability). If met, set the item to missing in the release file.

- For the variables JC, JD, and JF, check whether the denominator $\bar{E}_{kt}$ in the relevant cell kt rounds to zero. If so, set the disclosure status to 4 and set the item to missing in the release file.
- Check whether the item in cell kt rounds to zero. If so, set the disclosure status to 0 and set the item to 0 in the release file.
- Check whether the data used to construct the cell kt value were based on 1 or 2 individuals. If so, set the disclosure status to 2 and set the item to missing in the release file.
- Check whether the data used to construct the cell kt value were based on 1 or 2 employers. If so, set the disclosure status to 3 and set the item to missing in the release file.
- Check whether the distortion of cell kt value exceeds the limit set by the Census Disclosure Review Board. If so, set the disclosure status to 9 and copy the fuzzed value to the release file.
- Otherwise, set the disclosure status to 1 and copy the fuzzed value to the release file.

6.3 Analysis of the distortion due to the use of noise in the disclosure proofing process

Table 6.2 on the following page shows the distribution of the error in the first order serial correlation coefficient based on estimating an AR(1) using the multiplicatively distorted data (r^*) and using the undistorted data (r) for all counties in Illinois. The table shows that none of our variables is seriously affected by the distortion. In particular, the semi-interquartile range of the distortion is less than the precision with which estimated serial correlation coefficients are normally displayed—generally less than 2%, which means that distortion is economically meaningless.

Table 6.2: Distribution of the Error in the First Order Serial Correlation Coefficient Due to Multiplicative Noise Distortion ($r^* - r$)

Quantile	Beginning of Quarter Employment	Accessions	Separations	Full Quarter Employment	Net Job Flows
99%	0.07894	0.07153	0.06711	0.06644	0.01104
95%	0.04338	0.04253	0.04070	0.03465	0.00503
90%	0.02610	0.03043	0.02826	0.01972	0.00314
75%	0.00946	0.01387	0.01326	0.00718	0.00124
50%	-0.00043	0.00103	0.00004	-0.00003	0.00000
25%	-0.01026	-0.01271	-0.01179	-0.00641	-0.00096
10%	-0.02520	-0.03012	-0.02592	-0.01720	-0.00281
5%	-0.03695	-0.04100	-0.03569	-0.02806	-0.00471
1%	-0.06984	-0.06863	-0.06645	-0.06185	-0.01038

Section 7

Publicly available files

In this section, we describe the publicly available files, and how they differ from their internal correspondent files.

7.1 Public use files

The only public use product currently available on a regular basis are the QWI files proper. The public-use version differs from the Census-internal versions only in that the public-use version has been subject to the disclosure-proofing methods (coarsening and suppression) described in a previous section.

7.2 Restricted-access files

A larger set of files are available within the protected environment provided by the Census Research Data Centers (RDCs). The only information missing on these files relative to their internal-use counterparts is any information related to the confidential portions of the disclosure-proofing methods. All of these files can be accessed for research purposes by submitting a research proposal to the Center for Economic Studies at the U.S. Census Bureau[1].

7.2.1 ECF

The version of the ECF available in the RDC environment, referred to as 'LEHD-ECF', differs only minimally from the internal use ECF. Only variables used in the disclosure-proofing of the QWI have been suppressed. More information, including a detailed description of the LEHD-ECF, is available on the CES website (McKinney and Vilhuber; 2005).

[1] http://www.ces.census.gov

7.2.2 Unit Flow Files - Firm-level QWI

The SEINUNIT-level input files to the final aggregation step of the QWI, internally referred to as UFF(b), is available in the RDC environment under the reference 'LEHD-QWI'. The actual state-specific file is called `qwi_STATE_SEINUNIT`. While the internal-use version contains all information necessary to compute the disclosable QWI statistics, these variables have been suppressed from the RDC version. All statistics available at aggregated levels in the public-use QWI are available on the LEHD-QWI for the establishment. More information is available on the CES website.

7.2.3 Business Register Bridge

The Census Bureau maintains a list of establishments to develop the frame for economic censuses and surveys. This list is called the Business Register (BR), and is updated annually. The BR contains very reliable information on business identifiers, business organizational structure, and business location. Unfortunately, the establishment identification system for the Business Register differs from the LEHD establishment identifier (SEINUNIT). As a consequence, there is no single best way to form linkages between these data sources.

The LEHD Business Register Bridge (LEHD-BRB) available in the RDC network provides several ways to integrate the economic censuses and surveys with LEHD-provided data. The choice of link record is left to researchers, and the optimal choice will depend on the research objective. Available identifiers on the LEHD-BRB are the EIN, geographic information, and 4-digit SIC, which are linked to SEIN and SEINUNIT at different levels of precision. A more detailed guide is available on the CES website (Chiang et al.; 2005).

7.2.4 Human Capital files

These files will contain firm-level distributions of human capital measures as initially developed in Abowd et al. (2002). It will become available in 2005.

Section 8

Concluding remarks

8.0.1 Future projects

This section describes some of the ongoing efforts to improve the LEHD Infrastructure Files.

8.0.1.1 Planned improvements to the ICF

Currently researchers at LEHD are developing an enhanced, longitudinal version of the ICF, referred to as the LICF. It improves on the current version of the ICF in a number of ways. The current ICF is a collection of state-specific files. Individuals appearing in multiple states are treated independently for each state in all missing data analyses and in the computation of employment statistics. The LICF will be national in scope, with a single set of missing data implicates for any PIK found on any of the UI wage records, regardless of state and with integrated national geography.

Additional data sources will be integrated with the enhanced version of the ICF using direct links. The statistical link to the 1990 Decennial Census will be replaced by a direct link to the 2000 Decennial Census, and additional links to the ACS will be incorporated. The existing education impututation will greatly benefit from this enhancement. The additional links, as well as improved links to currently integrated data, will also allow for additional time-invariant characteristics to be incorporated and completed, including information on race and ethnicity and additional time-varying characteristics such as TANF recipiency.

Longitudinal residence information will be appended to the ICF based on the information available from the StARS. Where appropriate, residence will be imputed based on a change in residence imputation model and Bayesian methods for imputing geography at the block level, replacing the current residential address missing data imputation model. In fact, all imputation models will be based on the most up-to-date imputation engines developed at LEHD.

8.0.1.2 Planned improvements to the EHF

The UI wage records in several states suffer from defects in the historical records. These defects can be detected automatically when they produce a big enough fluctuation in certain flow statistics, typically beginning of period employment as compared to total flow employment. Algorithms

have been developed to detect the presence of missing wage records using the posterior predictive distribution.of employment histories given the available data and an informative prior on certain patterns. Once detected, the missing wage records are imputed, again using appropriate Bayesian methods. The same imputation engines are also being used to impute top-coded UI wages. These improvements are in the testing stage and should be implemented within the next year.

8.0.1.3 Planned improvements to the ECF

Two major enhancements to the ECF are in development. The first is a probabilistic record link to the Census Bureau's Business Register in order to improve the physical addresses on the ECF. This enhancement is currently in the testing phase. The second improvement to the ECF is a long-term project to incorporate information on non-employer businesses. The non-employer enhancements will affect both the ECF and the EHF because the information on non-employers also includes earnings from the non-employing business.

8.0.1.4 Creation of public-use synthetic data

As a part of a new, National Science Foundation Information Technology Research grant awarded to a consortium of Census Research Data Centers, researchers at LEHD and other parts of Census will collaborate with statisticians working in the RDCs to create and validate synthetic micro-data from the LEHD infrastructure files. Such synthetic micro-data will be confidentiality protected so that they may be released for public use. They will also be inference valid–permitting the estimation of some statistical models with results comparable to those obtained on the confidential micro-data.

8.0.2 The first 21st century statistical system

The goal of the development of the Quarterly Workforce Indicators was to create a 21st century statistical system. Without increasing respondent burden, the LEHD infrastructure permits the creation of extremely detailed statistics that, for the first time in the U.S., provide integrated demographic and economic information about the local labor market. The same techniques will work for other areas of interest–transportation dynamics and welfare-to-work dynamics to name just two examples. The two essential features of 21st century statistical systems will be their heavy reliance on existing data instruments (surveys, censuses and administrative records that are already in production) and their extensive use of data-intensive statistical modeling to enhance and summarize this information. In these regards, we think the LEHD infrastructure and the QWI system are worthy pioneers.

Bibliography

Abowd, J. M., Haltiwanger, J. C. and Lane, J. I. (2004). Integrated longitudinal employee-employer data for the United States, *American Economic Review* **94**(2).

Abowd, J. M., Lengermann, P. A. and McKinney, K. L. (2002). The measurement of human capital in the U.S. economy, *Technical paper TP-2002-09*, LEHD, U.S. Census Bureau.

Abowd, J. M. and Vilhuber, L. (2005). The sensitivity of economic statistics to coding errors in personal identifiers, *Journal of Business and Economic Statistics* **forthcoming**.

Benedetto, G., Haltiwanger, J., Lane, J. and McKinney, K. (2003). Using worker flows in the analysis of the firm, *Technical paper TP-2003-09*, LEHD, U.S. Census Bureau.

Bureau of Labor Statistics (1997). Quality improvement project: Unemployment insurance wage records, *report*, U.S. Department of Labor.

Chiang, H., Sandusky, K. and Vilhuber, L. (2005). LEHD Business Register Bridge technical documentation, *Internal Document IP-LEHD-BRB*, U.S. Census Bureau - LEHD.

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

McKinney, K. and Vilhuber, L. (2005). LEHD-ECF technical documentation, *Internal Document IP-LEHD-ECF*, U.S. Census Bureau - LEHD.

Stevens, D. W. (2002). Employment that is not covered by state unemployment insurance laws, *Technical paper TP-2002-16*, LEHD, U.S. Census Bureau.

Appendix A

Definitions of fundamental LEHD concepts

A.1 Fundamental Concepts

A.1.1 Dates

The QWI is a quarterly data system with calendar year timing. We use the notation YYYY:Q to refer to a year and quarter combination. For example, 1999:4 refers to the fourth quarter of 1999, which includes the months October, November, and December.

A.1.2 Employer

An employer in the QWI system consists of a single Unemployment Insurance (UI) account in a given state's UI wage reporting system. For statistical purposes the QWI system creates an employer identifier called an State Employer Identification Number (SEIN) from the UI-account number and information about the state (FIPS code). Thus, within the QWI system, the SEIN is a unique identifier within and across states but the entity to which it refers is a UI account. All QWI statistics are produced at the establishment level.

A.1.3 Establishment

For a given employer in the QWI system, an SEIN, each physical location within the state is assigned a unit number, called the SEINUNIT. This SEINUNIT is based on the reporting unit in the ES-202 files supplied by the states. All QWI statistics are produced by aggregating statistics calculated at the establishment level. Single-unit SEINs are UI accounts associated with a single reporting unit in the state. Thus, single-unit SEINs have only one associated SEINUNIT in every quarter. Multi-unit SEINs have two or more SEINUNITS associated for some quarters. Since the UI wage records are not coded down to the SEINUNIT, SEINUNITs are multiply imputed as described in the section on unit-to-worker imputation above. A feature of this imputation system is that it does not permit SEINUNIT to SEINUNIT movements within the same SEIN. Thus, for multi-unit SEINs, the definitions below produce the same flow estimates at the SEIN level whether the definition is applied to the SEIN or the SEINUNIT.

A.1.4 Employee

Individual employees are identified by their Social Security Numbers (SSN) on the UI wage records that provide the input to the QWI. To protect privacy and confidentialty of the SSN and the individual's name, a different branch of the Census Bureau removes the name and replaces the SSN with an internal Census identifier called a Protected Identity Key (PIK).

A.1.5 Job

The QWI system definition of a job is the association of an individual (PIK) with an establishment (SEINUNIT) in a given year and quarter. The QWI system stores the entire history of every job that an individual holds. Estimates are based on the definitions presented below, which formalize how the QWI system estimates the start of a job (accession), employment status (beginning- and end-of-quarter employment), continuous employment (full-quarter employment), the end of a job (separation), and average earnings for different groups.

A.1.6 Unemployment Insurance wage records (the QWI system universe)

The Quarterly Workforce Indicators are built upon concepts that begin with the report of an individual's UI-covered earnings by an employing entity (SEIN). An individual's UI wage record enters the QWI system if at least one employer reports earnings of at least one dollar for that individual (PIK) during the quarter. Thus, the job must produce at least one dollar of UI-covered earnings during a given quarter to count in the QWI system. The presence of this valid UI wage record in the QWI system triggers the beginning of calculations that estimate whether that individual was employed at the beginning of the quarter, at the end of the quarter, and continuously throughout the quarter. These designations are discussed below. Once these point-in-time employment measures have been estimated for the individual, further analysis of the individual's wage records results in estimates of full-quarter employment, accessions, separations (point-in-time and full-quarter), job creations and destructions, and a variety of full-quarter average earnings measures.

A.1.7 Employment at a point in time

Employment is estimated at two points in time during the quarter, corresponding to the first and last calendar days. An individual is defined as employed at the beginning of the quarter when that individual has valid UI wage records for the current quarter and the preceding quarter. Both records must apply to the same employer (SEIN). An individul is defined as employed at the end of the quarter when that individual has valid UI wage records for the current quarter and the subsequent quarter. Again, both records must show the same employer. The QWI system uses beginning and end of quarter employment as the basis for constructing worker and job flows. In addition, these measures are used to check the external consistency of the data, since a variety of employment estimates are available as point-in-time measures. Many federal statistics are based upon estimates of employment as of the 12th day of particular months. The Census Bureau uses March 12 as the reference date for employment measures contained in its Business Register and on the Economic

Censuses and Surveys. The BLS "Covered Employment and Wages (CEW)" series, which is based on the ES-202 data, use the 12th of each month as the reference date for employment. The QWI system cannot use exactly the same reference date as these other systems because UI wage reports do not specify additional detail regarding the timing of these payments. QWI research has shown that the point-in-time definitions used to estimate beginning and end of quarter employment track the CEW month one employment estimates well at the level of an employer (SEIN). For single-unit SEINs, there is no difference between an employer-based definition and an establishment-based definition of point-in-time employment. For multi-unit SEINs, the unit-to-worker imputation model assumes that unit-to-unit transitions within the same SEIN cannot occur. So, point in time employment defined at either the SEIN or SEINUNIT level produces the same result.

A.1.8 Employment for a full quarter

The concept of full quarter employment estimates individuals who are likely to have been continuously employed throughout the quarter at a given employer. An indivdual is defined as full-quarter-employed if that individual has valid UI-wage records in the current quarter, the preceding quarter, and the subsequent quarter at the same employer (SEIN). That is, in terms of the point-in-time definitions, if the individual is employed at the same employer at both the beginning and end of the quarter, then the individual is considered full-quarter employed in the QWI system.

Consider the following example. Suppose that an individual has valid UI wage records at employer *A* in 1999:2, 1999:3, and 1999:4. This individual does not have a valid UI wage record at employer *A* in 1999:1 or 2000:1. Then, according to the definitions above, the individual is employed at the end of 1999:2, the beginning and end of 1999:3, and the beginning of 1999:4 at employer *A*. The QWI system treats this individual as a full-quarter employee in 1999:3 but not in 1999:2 or 1999:4. Full-quarter status is not defined for either the first or last quarter of available data.

A.1.9 Point-in-time estimates of accession and separation

An accession occurs in the QWI system when it encounters the first valid UI wage record for a job (an individual (PIK)-employer (SEIN) pair). Accessions are not defined for the first quarter of available data from a given state. The QWI definition of an accession can be interpreted as an estimate of the number of new employees added to the payroll of the employer (SEIN) during the quarter. The individuals who acceded to a particular employer were not employed by that employer during the previous quarter but received at least one dollar of UI-covered earnings during the quarter of accession.

A separation occurs in the current quarter of the QWI system when it encounters no valid UI wage record for an individual-employer pair in the subsequent quarter. This definition of separation can be interpreted as an estimate of the number of employees who left the employer during the current quarter. These individuals received UI-covered earnings during the current quarter but did not receive any UI-covered earnings in the next quarter from this employer. Separations are not defined for the last quarter of available data.

A.1.10 Accession and separation from full-quarter employment

Full-quarter employment is not a point-in-time concept. Full-quarter accession refers to the quarter in which in individual first attains full-quarter employment status at a given employer. Full-quarter separation occurs in the last full-quarter that an individual worked for a given employer.

As noted above, full-quarter employment refers to an estimate of the number of employees who were employed at a given employer during the entire quarter. An accession to full-quarter employment, then, involves two additional conditions that are not relevant for ordinary accessions. First, the individual (PIK) must still be employed at the end of the quarter at the same employer (SEIN) for which the ordinary accession is defined. At this point (the end of the quarter where the accession occured and the beginning of the next quarter) the individual has acceded to continuing-quarter status. An accession to continuing-quarter status means that the individual acceded in the current quarter and is end-of-quarter employed. Next the QWI system must check for the possibility that the individual becomes a full-quarter employee in the subsequent quarter. An accession to full-quarter status occurs if the individual acceded in the previous quarter, and is employed at both the beginning and end of the current quarter. Consider the following example. An individual's first valid UI wage record with employer *A* occurs in 1999:2. The individual, thus acceded in 1999:2. The same individual has a valid wage record with employer *A* in 1999:3. The QWI system treats this individual as end-of-quarter employed in 1999:2 and beginning of quarter employed in 1999:3. The individual, thus, acceded to continuing-quarter status in 1999:2. If the individual also has a valid UI wage record at employer *A* in 1999:4, then the individual is full-quarter employed in 1999:3. Since 1999:3 is the first quarter of full-quarter employment, the QWI system considers this individual an accession to full-quarter employment in 1999:3.

Full-quarter separation works much the same way. One must be careful about the timing, however. If an individual separates in the current quarter, then the QWI system looks at the preceding quarter to determine if the individual was employed at the beginning of the current quarter. An individual who separates in a quarter in which that person was employed at the beginning of the quarter is a separation from continuing-quarter status in the current quarter. Finally, the QWI system checks to see if the individual was a full-quarter employee in the preceding quarter. An indivdidual who was a full quarter employee in the previous quarter is treated as a full-quarter separation in the quarter in which that person actually separates. Note, therefore, that the definition of full-quarter separation preserves the timing of the actual separation (current quarter) but restricts the estimate to those individuals who were full-quarter status in the preceding quarter. For example, suppose that an individual separates from employer *A* in 1999:3. This means that the individual had a valid UI wage record at employer *A* in 1999:3 but did not have a valid UI wage record at employer *A* in 1999:4. The separation is dated 1999:3. Suppose that the individual had a valid UI wage record at employer *A* in 1999:2. Then, a separation from continuing quarter status occured in 1999:3. Finally, suppose that this individual had a valid UI wage record at employer *A* in 1999:1. Then, this individual was a full-quarter employee at employer *A* in 1999:2. The QWI system records a full-quarter separation in 1999:3.

A.1.11 Point-in-time estimates of new hires and recalls

The QWI system refines the concept of accession into two subcategories: new hires and recalls. In order to do this, the QWI system looks at a full year of wage record history prior to the quarter in which an accession occurs. If there are no valid wage records for this job (PIK-SEIN) during the four quarters preceding an accession, then the accession is called a new hire; otherwise, the accession is called a recall. Thus, new hires and recalls sum to accessions. For example, suppose that an individual accedes to employer *A* in 1999:3. Recall that this means that there is a valid UI wage record for the individual 1 at employer *A* in 1999:3 but not in 1999:2. If there are also no valid UI wage records for individual 1 at employer *A* for 1999:1, 1998:4 and 1998:3, then the QWI system designates this accession as a new hire of individual 1 by employer *A* in 1999:3. Consider a second example in which individual 2 accedes to employer *B* in 2000:2. Once again, the accession implies that there is not a valid wage record for individual 2 at employer *B* in 2000:1. If there is a valid wage record for individual 2 at employer *B* in 1999:4, 1999:3, or 1999:2, then the QWI system designates the accession of individual 2 to employer *B* as a recall in 2000:2. New hire and recall data, because they depend upon having four quarters of historical data, only become available one year after the data required to estimate accessions become available.

A.1.12 New hires and recalls to and from full-quarter employment

Accessions to full-quarter status can also be decomposed into new hires and recalls. The QWI system accomplishes this decomposition by classifying all accession to full-quarter status who were classified as new hires in the previous quarter as new hires to full-quarter status in the current quarter. Otherwise, the accession to full-quarter status is classified as a recall to full-quarter status. For example, if individual 1 accedes to full-quarter status at employer *A* in 1999:4 then, according to the definitions above, individual 1 acceded to employer *A* in 1999:3 and reached full-quarter status in 1999:4. Suppose that the accession to employer *A* in 1999:3 was classified as a new hire, then the accession to full quarter status in 1999:4 is classified as a full-quarter new hire. For another example, consider individual 2 who accedes to full-quarter status at employer *B* in 2000:3. Suppose that the accession of individual 2 to employer *B* in 2000:2, which is implied by the full-quarter accession in 2000:3, was classified by the QWI system as a recall in 2000:2; then, the accession of individual 2 to full-quarter status at employer *B* in 2000:3 is classified as a recall to full-quarter status.

A.1.13 Job creations and destructions

Job creations and destructions are defined at the employer (SEIN) level and not at the job (PIK-SEIN) level. To construct an estimate of job creations and destructions, the QWI system totals beginning and ending employment for each quarter for every employer in the UI wage record universe, that is, for an employer who has at least one valid UI wage record during the quarter. The QWI system actually uses the Davis et al. (1996) formulas for job creation and destruction (see definitions in Appendix A.2 on page 45). Here, we use a simplified definition. If end-of-quarter employment is greater than beginning-of-quarter employment, then the employer has created jobs.

The QWI system sets job creations in this case equal to end-of-quarter employment less beginning-of-quarter employment. The estimate of job destructions in this case is zero. On the other hand, if beginning-of-quarter employment exceeds end-of-quarter employment, then this employer has destroyed jobs. The QWI system computes job destructions in this case as beginning-of-period employment less end-of-period employment. The QWI system sets job creations to zero in this case. Notice that either job creations are positive or job destructions are postive, but not both. Job creations and job destructions can simultaneously be zero if beginning-of-quarter employment equals end-of-quarter employment. There is an important suptelty regarding job creations and destructions when they are computed for different sex and age groups within the same employer. There can be creation and destruction of jobs for certain demographic groups within the employer without job creation or job destruction occuring overall. That is, jobs can be created for some demographic groups and destroyed for others even at enterprises that have no change in employment as a whole.

Here is a simple example. Suppose employer *A* has 250 employees at the beginning of 2000:3 and 280 employees at the end of 2000:3. Then, employer *A* has 30 job creations and zero job destructions in 2000:3. Now suppose that of the 250 employees 100 are men and 150 are women at the beginning of 2000:3. At the end of the quarter suppose that there are 135 men and 145 women. Then, job creations for men are 35 and job destructions for men are 0 in 2000:3. For women in 2000:3 job creations are 0 and job destructions are 5. Notice that the sum of job creations for the employer by sex (35 + 0) is not equal to job creations for the employer as a whole (30) and that the sum of job destructions by sex (0 + 5) is not equal to job destructions for the employer as a whole.

A.1.14 Net job flows

Net job flows are also only defined at the level of an employer (SEIN). They are the difference between job creations and job destructions. Net job flows are, thus, always equal to end-of-quarter employment less beginning of quarter employment.

Returning to the example in the description of job creations and destructions. Employer *A* has 250 employees at the beginning of 2000:3 and 280 employees at the end of 2000:3. Net job flows are 30 (job creations less job destructions or beginning-of-quarter employment less end-of-quarter employment). Suppose, once again that employment of men goes from 100 to 135 from the beginning to the end of 2000:3 and employment of women goes from 150 to 145. Notice, now, that net job flows for men (35) plus net job flows for women (-5) equals net job flows for the employer as a whole (30). Net job flows are additive across demographic groups even though gross job flows (creations and destructions) are not.

Some useful relations among the worker and job flows include:

- Net job flows = job creations - job destructions
- Net job flows = end-of-quarter employment - beginning-of-period employment
- Net job flows = accessions - separations

These relations hold for every demographic group and for the employer as a whole. Additional identities are shown in Appendix A.2.

A.1.15 Full-quarter job creations, job destructions and net job flows

The QWI system applies the same job flow concepts to full-quarter employment to generate estimates of full-quarter job creations, full-quarter job destructions, and full-quarter net job flows. Full-quarter employment in the current quarter is compared to full-quarter employment in the preceding quarter. If full-quarter employment has increased between the preceding quarter and the current quarter, then full-quarter job creations are equal to full-quarter employment in the current quarter less full-quarter employment in the preceding quarter. In this case full-quarter job destructions are zero. If full-quarter employment has decreased between the previous and current quarters, then full-quarter job destructions are equal to full-quarter employment in the preceding quarter minus full-quarter employment in the current quarter. In this case, full-quarter job destructions are zero. Full-quarter net job flows equal full-quarter job creations minus full-quarter job destructions. The same identities that hold for the regular job flow concepts hold for the full-quarter concepts.

A.1.16 Average earnings of end-of-period employees

The average earnings of end-of-period employees is estimated by first totaling the UI wage records for all individuals who are end-of-period employees at a given employer in a given quarter. Then the total is divided by the number of end-of-period employees for that employer and quarter.

A.1.17 Average earnings of full-quarter employees

Measuring earnings using UI wage records in the QWI system presents some interesting challenges. The earnings of end-of-quarter employees who are not present at the beginning of the quarter are the earnings of accessions during the quarter. The QWI system does not provide any information about how much of the quarter such individuals worked. The range of possibilities goes from 1 day to every day of the quarter. Hence, estimates of the average earnings of such individuals may not be comparable from quarter to quarter unless one assumes that the average accession works the same number of quarters regardless of other conditions in the economy. Similarly, the earnings of beginning-of-quarter who are not present at the end of the quarter represent the earnings of separations. These present the same comparison problems as the average earnings of accessions; namely, it is difficult to model the number of weeks worked during the quarter. If we consider only those individuals employed at the firm in a given quarter who were neither accessions nor separations during that quarter, we are left, exactly, with the full-quarter employees, as discussed above.

The QWI system measures the average earnings of full-quarter employees by summing the earnings on the UI wage records of all individuals at a given employer who have full-quarter status in a given quarter then dividing by the number of full-quarter employees. For example, suppose that in 2000:2 employer *A* has 10 full-quarter employees and that their total earnings are $\$300,000$. Then, the average earnings of the full-quarter employees at *A* in 2000:2 is $\$30,000$. Suppose, further that 6 of these employees are men and that their total earnings are $\$150,000$. So, the average earnings of full-quarter male employees is $\$25,000$ in 2000:2 and the average earnings of female full-quarter employees is $\$37,500$ $(= \$150,000/4)$.

A.1.18 Average earnings of full-quarter accessions

As discussed above, a full-quarter accession is an individual who acceded in the preceding quarter and acheived full-quarter status in the current quarter. The QWI system measures the average earnings of full-quarter accessions in a given quarter by summing the UI wage record earnings of all full-quarter accessions during the quarter and dividing by the number of full-quarter accessions in that quarter.

A.1.19 Average earnings of full-quarter new hires

Full-quarter new hires are accessions to full-quarter status who were also new hires in the preceding quarter. The average earnings of full-quarter new hires are measured as the sum of UI wage records for a given employer for all full-quarter new hires in a given quarter divided by the number of full-quarter new hires in that quarter.

A.1.20 Average earnings of full-quarter separations

Full-quarter separations are individuals who separate during the current quarter who were full-quarter employees in the previous quarter. The QWI system measures the average earnings of full-quarter separations by summing the earnings for all individuals who are full-quarter status in the current quarter and who separate in the subsequent quarter. This total is then divided by full-quarter separations in the subsequent quarter. The average earnings of full-quarter separations is, thus, the average earnings of full-quarter employees in the current quarter who separated in the next quarter. Note the dating of this variable.

A.1.21 Average periods of non-employment for accessions, new hires, and recalls

As noted above an accession occurs when a job starts; that is, on the first occurance of an SEIN-PIK pair following the first quarter of available data. When the QWI system detects an accession, it measures the number of quarters (up to a maximum of four) that the individual spent non-employed in the state prior to the accession. The QWI system estimates the number of quarters spent non-employed by looking for all other jobs held by the individual at any employer in the state in the preceding quarters up to a maximum of four. If the QWI system doesn't find any other valid UI-wage records in a quarter preceding the accession it augments the count of non-employed quarters for the individual who acceded, up to a maximum of four. Total quarters of non-employment for all accessions is divided by accessions to estimate average periods of non-employment for accessions.

Here is a detailed example. Suppose individual 1 and individual 2 accede to employer *A* in 2000:1. In 1999:4, individual *A* does not work for any other employers in the state. In 1999:1 through 1999:3 individual 1 worked for employer *B*. Individual 1 had one quarter of non-employment preceding the accession to employer *A* in 2000:1. Individual 2 has no valid UI wage records for 1999:1 through 1999:4. Indivdiual 2 has four quarters of non-employment preceding the accession

to employer *A* in 2000:1. The accessions to employer *A* in 2000:1 had an average of 2.5 quarters of non-employment in the state prior to accession.

Average periods of non-employment for new hires and recalls are estimated using exactly analogous formulas except that the measures are estimated separately for accesions who are also new hires as compared with accession who are recalls.

A.1.22 Average number of periods of non-employment for separations

Analogous to the average number of periods of non-employment for accessions prior to the accession, the QWI system measures the average number of periods of non-employment in the state for individuals who separated in the current quarter, up to a maximum of four. When the QWI system detects a separation, it looks forward for up to four quarters to find valid UI wage records for the individual who separated and other employers in the state. Each quarter that it fails to detect any such jobs is counted as a period of non-employment, up to a maximum of four. The average number of periods of non-employment is estimated by dividing the total number of periods of non-employment for separations in the current quarter by the number of separations in the quarter.

A.1.23 Average changes in total earnings for accessions and separations

The QWI system measures the change in total earnings for individuals who accede or separate in a given quarter. For an individual accession in a given quarter, the QWI system computes total earnings from all valid wage records for all of the individual's employers in the preceding quarter. The system then computes the total earnings for the same individual for all valid wage records and all employers in the current quarter. The acceding individual's change in earnings is the difference between the current quarter earnings from all employers and the preceding quarter earnings from all employers. The average change in earnings for all accessions is the total change in earnings for all accesions divided by the number of accessions.

The QWI system computes the average change in earnings for separations in an analogous manner. The system computes total earnings from all employers for the separating indivdiual in the current quarter and subtracts total earnings from all employers in the subsequent quarter. The average change in earnings for all separations is the total change in earnings for all separations divided by the number of separations.

Here is an example for the average change in earnings of accessions. Suppose individual 1 accedes to employer *A* in 2000:3. Earnings for individual 1 at employer *A* in 2000:3 are $\$8,000$. Individual 1 also worked for employer *B* in 2000:2 and 2000:3. Individual 1's earnings at employer *B* were $\$7,000$ and $\$3,000$ in in 2000:2 and 2000:3, respectively. Individual 1's change in total earnings between 2000:3 and 2000:2 was $\$4,000$ $(= \$8,000 + \$3,000 - \$7,000)$. Individual 2 also acceded to employer *A* in 2000:3. Individual 2 earned $\$9,000$ from employer *A* in 2000:3. Individual 2 had no other employers during 2000:2 or 2000:3. Individual 2's change in total earnings is $\$9,000$. The average change in earnings for all of employer *A*'s accessions is $\$6,500$ $(= (\$4,000 + \$9,000)/2)$, the average change in total earnings for individuals 1 and 2.

A.2 Definitions of Job Flow, Worker Flow, and Earnings Statistics

A.2.1 Overview and basic data processing conventions

For internal processing the variable t refers to the sequential quarter. The variable t runs from $qmin$ to $qmax$, regardless of the state being processed. The quarters are numbered sequentially from 1 (1985:1) to the latest available quarter. These values are $qmin = 1$ (1985:1) and $qmax = 78$ (2004:2), as of April 5, 2005. For publication, presentation, and internal data files, all dates are presented as (year:quarter) pairs, *e.g.* (1990:1) for first quarter 1990. The variable $qfirst$ refers to the first available sequential quarter of data for a state (*e.g.*, $qfirst = 21$ for Illinois). The variable $qlast$ refers to the last available sequential quarter of data for a state (*e.g.*, $qlast = 78$ for Illinois). Unless otherwise specified a variable is defined for $qfirst \leq t \leq qlast$. Statistics are produced for both sexes combined, as well as separately, for all age groups, ages 14-18, 19-21, 22-24, 25-34, 35-44, 45-54, 55-64, 65+, and all combinations of these age groups and sexes. An individual's age is measured as of the last day of the quarter.

A.2.2 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 establishment j

$$m_{ijt} = \begin{cases} 1, \text{ if } i \text{ has positive earnings at establishment } j \text{ during quarter } t \\ 0, \text{ otherwise.} \end{cases} \tag{A.1}$$

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

$$b_{ijt} = \begin{cases} 1, \text{ if } m_{ijt-1} = m_{ijt} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.2}$$

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

$$e_{ijt} = \begin{cases} 1, \text{ if } m_{ijt} = m_{ijt+1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.3}$$

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

$$a_{1ijt} = \begin{cases} 1, \text{ if } m_{ijt-1} = 0 \ \& \ m_{ijt} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.4}$$

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

$$s_{1ijt} = \begin{cases} 1, \text{ if } m_{ijt} = 1 \ \& \ m_{ijt+1} = 0 \\ 0, \text{ otherwise.} \end{cases} \tag{A.5}$$

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)

$$f_{ijt} = \begin{cases} 1, \text{ if } m_{ijt-1} = 1 \ \& \ m_{ijt} = 1 \ \& \ m_{ijt+1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.6}$$

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

$$h_{1ijt} = \begin{cases} 1, \text{ if } m_{ijt-4} = 0 \ \& \ m_{ijt-3} = 0 \ \& \ m_{ijt-2} = 0 \ \& \ m_{ijt-1} = 0 \ \& \ m_{ijt} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.7}$$

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

$$r_{1ijt} = \begin{cases} 1, \text{ if } m_{ijt-1} = 0 \ \& \ m_{ijt} = 1 \ \& \ h_{ijt} = 0 \\ 0, \text{ otherwise.} \end{cases} \tag{A.8}$$

Accessions to consecutive quarter status (a_2): For $qfirst < t < qlast$, individual i transited from accession to consecutive-quarter status at j at the start of $t + 1$ (accession in t and still employed at the end of the quarter)

$$a_{2ijt} = \begin{cases} 1, \text{ if } a_{1ijt} = 1 \ \& \ m_{ijt+1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.9}$$

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

$$a_{3ijt} = \begin{cases} 1, \text{ if } a_{2ijt-1} = 1 \ \& \ m_{ijt+1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.10}$$

New hires to consecutive quarter status (h_2): For $qfirst+3 < t < qlast$, individual i transited from newly hired to consecutive-quarter hired status at j at the start of $t + 1$ (hired in t and still employed at the end of the quarter)

$$h_{2ijt} = \begin{cases} 1, \text{ if } h_{1ijt} = 1 \ \& \ m_{ijt+1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.11}$$

New hires to full quarter status (a_3): For $qfirst + 4 < t < qlast$, individual i transited 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)

$$h_{3ijt} = \begin{cases} 1, \text{ if } h_{2ijt-1} = 1 \ \& \ m_{ijt+1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.12}$$

Recalls to consecutive quarter status (r_2): For $qfirst + 3 < t < qlast$, individual i transited from recalled to consecutive-quarter recalled status at j at the start of $t + 1$ (recalled in t and still employed at the end of the quarter)

$$r_{2ijt} = \begin{cases} 1, \text{ if } r_{1ijt} = 1 \ \& \ m_{ijt+1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.13}$$

Recalls to full quarter status (r_3): For $qfirst + 4 < t < qlast$, individual i transited from consecutive-quarter recalled to full-quarter recalled status at j at the start of $t+1$ (recalled in $t-1$ and full-quarter employed in t)

$$r_{3ijt} = \begin{cases} 1, \text{ if } r_{2ijt-1} = 1 \ \& \ m_{ijt+1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.14}$$

Separations from consecutive quarter status (s_2): For $qfirst < t < qlast$, individual i separated from j during t with consecutive-quarter status at the start of t

$$s_{2ijt} = \begin{cases} 1, \text{ if } s_{1ijt} = 1 \ \& \ m_{ijt-1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{A.15}$$

Separations from full-quarter status (s_3): For $qfirst + 1 < t < qlast$, individual i separated from j during t with full-quarter status during $t - 1$

$$s_{3ijt} = \begin{cases} 1, \text{if } s_{2ijt} = 1 \ \ \& \ \ m_{ijt-2} = 1 \\ 0, \text{ otherwise} \end{cases} \tag{A.16}$$

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

$$w_{1ijt} = \sum \text{all } UI \text{ covered earnings by } i \text{ at } j \text{ during } t \tag{A.17}$$

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

$$w_{2ijt} = \begin{cases} w_{1ijt}, \text{ if } e_{ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.18}$$

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

$$w_{3ijt} = \begin{cases} w_{1ijt}, \text{ if } f_{ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.19}$$

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

$$w_{1i\bullet t} = \sum_{j \text{ employs } i \text{ during } t} w_{1ijt} \tag{A.20}$$

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

$$w_{2i\bullet t} = \begin{cases} w_{1i\bullet t}, \text{ if } e_{ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.21}$$

Total earnings of full-quarter employees i during period t

$$w_{3i\bullet t} = \begin{cases} w_{1i\bullet t}, \text{ if } f_{ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.22}$$

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

$$\Delta w_{1i\bullet t} = w_{1i\bullet t} - w_{1i\bullet t-1} \tag{A.23}$$

Earnings of accessions to employer j during period t

$$wa_{1ijt} = \begin{cases} w_{1ijt}, \text{ if } a_{1ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.24}$$

Earnings of consecutive-quarter accessions to establishment j during period t

$$wa_{2ijt} = \begin{cases} w_{1ijt}, \text{ if } a_{2ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.25}$$

Earnings of full-quarter accessions to establishment j during period t

$$wa_{3ijt} = \begin{cases} w_{1ijt}, \text{ if } a_{3ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.26}$$

Earnings of full-quarter new hires to establishment j during period t

$$wh_{3ijt} = \begin{cases} w_{1ijt}, \text{ if } h_{3ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.27}$$

Total earnings change for accessions to establishment j during t

$$\Delta wa_{1ijt} = \begin{cases} \Delta w_{1i\bullet t}, \text{ if } a_{1ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.28}$$

Total earnings change for full-quarter accessions to establishment j during t

$$\Delta wa_{3ijt} = \begin{cases} \Delta w_{1i\bullet t}, \text{ if } a_{3ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.29}$$

Earnings of separations from establishment j during period t

$$ws_{1ijt} = \begin{cases} w_{1ijt}, \text{ if } s_{1ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.30}$$

Earnings of full-quarter separations to establishment j during period t

$$ws_{3ijt} = \begin{cases} w_{1ijt}, \text{ if } s_{3ijt+1} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.31}$$

Total earnings change for separations from establishment j during t

$$\Delta ws_{1ijt} = \begin{cases} \Delta w_{1i\bullet t+1}, \text{ if } s_{1ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.32}$$

Total earnings change for full-quarter separations from establishment j during t

$$\Delta ws_{3ijt} = \begin{cases} \Delta w_{1i\bullet t+1}, \text{ if } s_{3ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.33}$$

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

$$na_{ijt} = \begin{cases} \sum\limits_{1 \leqslant s \leqslant 4} n_{it-s}, \text{ if } a_{1ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.34}$$

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

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

$$nh_{ijt} = \begin{cases} \sum\limits_{1 \leqslant s \leqslant 4} n_{it-s}, \text{ if } h_{1ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.35}$$

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

$$nr_{ijt} = \begin{cases} \sum\limits_{1 \leqslant s \leqslant 4} n_{it-s}, \text{ if } r_{1ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.36}$$

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

$$ns_{ijt} = \begin{cases} \sum\limits_{1 \leqslant s \leqslant 4} n_{it+s}, \text{ if } s_{1ijt} = 1 \\ \text{undefined, otherwise} \end{cases} \tag{A.37}$$

A.2.3 Establishment concepts

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

$$b_{\cdot jt} = \sum_i b_{ijt} \tag{A.38}$$

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

Beginning-of-period employment (number of jobs)

$$B_{jt} = b_{\cdot jt} \tag{A.39}$$

End-of-period employment (number of jobs)

$$E_{jt} = e_{\cdot jt} \tag{A.40}$$

Employment any time during the period (number of jobs)

$$M_{jt} = m_{\cdot jt} \tag{A.41}$$

Full-quarter employment

$$F_{jt} = f_{\cdot jt} \tag{A.42}$$

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

$$JF_{jt} = E_{jt} - B_{jt} \tag{A.43}$$

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

$$\bar{E}_{jt} = \frac{(B_{jt} + E_{jt})}{2} \tag{A.44}$$

Average employment growth rate for establishment j between periods $t-1$ and t

$$G_{jt} = \frac{JF_{jt}}{\bar{E}_{jt}} \tag{A.45}$$

Job creation for establishment j between periods $t-1$ and t

$$JC_{jt} = \bar{E}_{jt} \max\left(0, G_{jt}\right) \tag{A.46}$$

Average job creation rate for establishment j between periods $t-1$ and t

$$JCR_{jt} = \frac{JC_{jt}}{\bar{E}_{jt}} \tag{A.47}$$

Job destruction for establishment j between periods $t-1$ and t

$$JD_{jt} = \bar{E}_{jt} \operatorname{abs}\left(\min\left(0, G_{jt}\right)\right) \tag{A.48}$$

Average job destruction rate for establishment j between periods $t-1$ and t

$$JDR_{jt} = \frac{JD_{jt}}{\bar{E}_{jt}} \tag{A.49}$$

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

$$FJF_{jt} = F_{jt} - F_{jt-1} \tag{A.50}$$

Average full-quarter employment for establishment j during period t

$$\bar{F}_{jt} = \frac{F_{jt-1} + F_{jt}}{2} \tag{A.51}$$

Average full-quarter employment growth rate for establishment j between $t-1$ and t

$$FG_{jt} = \frac{FJF_{jt}}{\bar{F}_{jt}} \tag{A.52}$$

Full-quarter job creations for establishment j between $t-1$ and t

$$FJC_{jt} = \bar{F}_{jt} \max\left(0, FG_{jt}\right) \tag{A.53}$$

Average full-quarter job creation rate for establishment j between $t-1$ and t

$$FJCR_{jt} = FJC_{jt} \,/\, \bar{F}_{jt} \tag{A.54}$$

Full-quarter job destruction for establishment j between $t-1$ and t

$$FJD_{jt} = \bar{F}_{jt} abs\left(\min\left(0, FG_{jt}\right)\right) \tag{A.55}$$

Average full-quarter job destruction rate for establishment j between $t-1$ and t

$$FJDR_{jt} = FJD_{jt} \,/\, \bar{F}_{jt} \tag{A.56}$$

Accessions for establishment j during t

$$A_{jt} = a_{1 \cdot jt} \tag{A.57}$$

Average accession rate for establishment j during t

$$AR_{jt} = A_{jt} \,/\, \bar{E}_{jt} \tag{A.58}$$

Separations for establishment j during t

$$S_{jt} = s_{1 \cdot jt} \tag{A.59}$$

Average separation rate for establishment j during t

$$SR_{jt} = S_{jt} \Big/ \bar{E}_{jt} \tag{A.60}$$

New hires for establishment j during t

$$H_{jt} = h_{1 \cdot jt} \tag{A.61}$$

Full Quarter New hires for establishment j during t

$$H_{3jt} = h_{3 \cdot jt} \tag{A.62}$$

Recalls for establishment j during t

$$R_{jt} = r_{1 \cdot jt} \tag{A.63}$$

Flow into full-quarter employment for establishment j during t

$$FA_{jt} = a_{3 \cdot jt} \tag{A.64}$$

New hires into full-quarter employment for establishment j during t

$$FH_{jt} = h_{3 \cdot jt} \tag{A.65}$$

Average rate of flow into full-quarter employment for establishment j during t

$$FAR_{jt} = FA_{jt} \Big/ \bar{F}_{jt} \tag{A.66}$$

Flow out of full-quarter employment for establishment j during t

$$FS_{jt} = s_{3 \cdot jt} \tag{A.67}$$

Average rate of flow out of full-quarter employment for establishment j during t

$$FSR_{jt} = FS_{jt} \Big/ \bar{F}_{jt} \tag{A.68}$$

Flow into consecutive quarter employment for establishment j during t

$$CA_{jt} = a_{2 \cdot jt} \tag{A.69}$$

Flow out of consecutive quarter employment for establishment j during t

$$CS_{jt} = s_{2 \cdot jt} \tag{A.70}$$

Total payroll of all employees

$$W_{1jt} = w_{1 \cdot jt} \tag{A.71}$$

Total payroll of end-of-period employees

$$W_{2jt} = w_{2 \cdot jt} \tag{A.72}$$

Total payroll of full-quarter employees

$$W_{3jt} = w_{3 \cdot jt} \tag{A.73}$$

Total payroll of accessions

$$WA_{jt} = wa_{1 \cdot jt} \tag{A.74}$$

Change in total earnings for accessions

$$\Delta WA_{jt} = \sum_{i \in \{J(i,t)=j\}} \triangle wa_{1ijt} \tag{A.75}$$

Total payroll of transits to consecutive-quarter status

$$WCA_{jt} = wa_{2 \cdot jt} \tag{A.76}$$

Total payroll of transits to full-quarter status

$$WFA_{jt} = wa_{3 \cdot jt} \tag{A.77}$$

Total payroll of new hires to full-quarter status

$$WFH_{jt} = wh_{3 \cdot jt} \tag{A.78}$$

Change in total earnings for transits to full-quarter status

$$\Delta WFA_{jt} = \sum_{i \in \{J(i,t)=j\}} \Delta wa_{3ijt} \tag{A.79}$$

Total periods of non-employment for accessions

$$NA_{jt} = na_{\cdot jt} \tag{A.80}$$

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

$$NH_{jt} = nh_{\cdot jt} \tag{A.81}$$

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

$$NR_{jt} = nr_{\cdot jt} \tag{A.82}$$

Total earnings of separations

$$WS_{jt} = ws_{1 \cdot jt} \tag{A.83}$$

Total change in total earnings for separations

$$\Delta WS_{jt} = \sum_{i \in \{J(i,t)=j\}} \Delta ws_{1ijt} \tag{A.84}$$

Total earnings of separations from full-quarter status (most recent full quarter)

$$WFS_{jt} = ws_{3 \cdot jt} \tag{A.85}$$

Total change in total earnings for full-quarter separations

$$\Delta WFS_{jt} = \sum_{i \in \{J(i,t)=j\}} \Delta ws_{3ijt} \tag{A.86}$$

Total periods of non-employment for separations

$$NS_{jt} = ns_{\cdot jt} \tag{A.87}$$

Average earnings of end-of-period employees

$$ZW_{2jt} = W_{2jt} \,/\, E_{jt} \tag{A.88}$$

Average earnings of full-quarter employees

$$ZW_{3jt} = W_{3jt} \,/\, F_{jt} \tag{A.89}$$

Average earnings of accessions

$$ZWA_{jt} = WA_{jt} \,/\, A_{jt} \tag{A.90}$$

Average change in total earnings for accessions

$$Z\Delta WA_{jt} = \Delta WA_{jt} \,/\, A_{jt} \tag{A.91}$$

Average earnings of transits to full-quarter status

$$ZWFA_{jt} = WFA_{jt} \,/\, FA_{jt} \tag{A.92}$$

Average earnings of new hires to full-quarter status

$$ZWFH_{jt} = WFH_{jt} \,/\, FH_{jt} \tag{A.93}$$

Average change in total earnings for transits to full-quarter status

$$Z\Delta WFA_{jt} = \Delta WFA_{jt} \,/\, FA_{jt} \tag{A.94}$$

Average periods of non-employment for accessions

$$ZNA_{jt} = NA_{jt} \,/\, A_{jt} \tag{A.95}$$

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

$$ZNH_{jt} = NH_{jt} \,/\, H_{jt} \tag{A.96}$$

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

$$ZNR_{jt} = NR_{jt} \,/\, R_{jt} \tag{A.97}$$

Average earnings of separations

$$ZWS_{jt} = WS_{jt} \,/\, S_{jt} \tag{A.98}$$

Average change in total earnings for separations

$$Z\Delta WS_{jt} = \Delta WS_{jt} \,/\, S_{jt} \tag{A.99}$$

Average earnings of separations from full-quarter status (most recent full quarter)

$$ZWFS_{jt-1} = WFS_{jt-1} \,/\, FS_{jt} \tag{A.100}$$

Average change in total earnings for full-quarter separations

$$Z\Delta WFS_{jt} = \Delta WFS_{jt} \,/\, FS_{jt} \tag{A.101}$$

Average periods of non-employment for separations

$$ZNS_{jt} = NS_{jt} \,/\, S_{jt} \tag{A.102}$$

End-of-period employment (number of workers) [Aggregate concept not related to a business]

$$N_t = n_{\cdot t} \tag{A.103}$$

A.2.4 Aggregation of flows

The rate of growth is equal to the ratio of net job flows to total employment:

$$G_{jt} = JF_{jt} \Big/ \bar{E}_{jt} \tag{A.104}$$

So, to impute the aggregate growth rate in a county (or sic) for some group of firms, let

$$G_{kt} = \frac{\sum\limits_{j \in \{K(j)=k\}} \bar{E}_{jt} \times G_{jt}}{\bar{E}_{kt}} \tag{A.105}$$

for county k where the function $K(j)$ indicates the classification associated with firm j.

We calculate the aggregate job flow as

$$JF_{kt} = \sum_{j \in \{K(j)=k\}} JF_{jt}. \tag{A.106}$$

Substitution yields

$$JF_{kt} = \sum_{j} (\bar{E}_{jt} \times G_{jt}) = G_{kt} \times \bar{E}_{kt}, \tag{A.107}$$

so the aggregate job flow, as computed, is equivalent to the aggregate growth rate times aggregate employment. Gross job creation/destruction are related to job creation/destruction rates by similar logic (Davis et al.; 1996, p. 189 for details).