The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators

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

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

January 16, 2008

Abstract

The Longitudinal Employer-Household Dynamics (LEHD) Program at the U.S. Census Bureau, with the support of several national research agencies, has built a set of infrastructure files using administrative data provided by state agencies, enhanced with information from other administrative data sources, 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. Beginning in 2003 and building on this infrastructure, the Census Bureau has published the Quarterly Workforce Indicators (QWI), a new collection of data series that offers 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. We describe the multiple imputation methods used to impute missing data and the statistical matching techniques used to combine and edit data when a direct identifier match requires improvement. Both of these innovations are crucial to the success of the final product. Finally, we pay special attention to the details of the confidentiality protection system used to protect the identity and micro data values of the underlying entities used to form the published estimates. We provide a brief description of public-use and restricted-access data files with pointers to further documentation for researchers interested in using these data.

Contents

iii

List of Tables

List of Figures

Section 1

Introduction

Since 2003, U.S. Census Bureau has published a new and innovative 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 integrated from other data sources at the Census Bureau, these statistics offer unprecedented detail on the local dynamics of labor markets. Despite the fine geographic and industry detail, the confidentiality of the underlying micro-data is maintained by the application of new, state-of-the-art protection methods.

The underlying data infrastructure was designed by the Longitudinal Employer-Household Dynamics (LEHD) Program at the Census Bureau (Abowd et al.; 2004). The Census Bureau collaborates with its state partners, the suppliers of critical administrative records from the state unemployment insurance programs, through the Local Employment Dynamics (LED) cooperative federal-state program. 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 46 currently participating states.[1] When complete, the LEHD Infrastructure Files will be the first nationally comprehensive statistical product developed from a universe that covers jobs–a statutory employment relation between an individual and employer–as distinct from ones that

cover households (*e.g.*, the Decennial Census of Population and Housing) or establishments (*e.g.*, the Economic Censuses or the Quarterly Census of Employment and Wages [QCEW]).

In this article, we describe the primary input data underlying 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. Many of the infrastructure and derivative micro-data files are now available within the Research Data Centers of the U.S. Census Bureau, and we indicate these files during the discussion.

The QWI use a bewildering array of data sources–administrative records, demographic surveys and censuses and economic surveys and censuses. The Census Bureau receives UI wage records and ES-202 (QCEW) establishment records from each state participating in the LED federal/state partnership. 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 exact one-to-one matches based on stable identifiers. In some cases, statistical matching techniques are used, and in other cases critical linking values are imputed. Throughout the process, critical imputations are done multiple times, improving the accuracy of the final estimates and permitting an assessment of the additional variability due to the imputations.

Data integration is a two-way street. Not only do the Census Bureau's surveys and censuses improve the detail on the administrative files, allowing the creation of new statistical products without any increase in respondent burden, but also as a part of its Title 13 mission, the Census Bureau uses the integrated files to improve the its other demographic and economic products. The demographic products that have been improved include the Current Population Survey, the Survey of Income and Program Participation, and the American Community Survey. In addition, the LEHD Infrastructure Files are used for research 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 Censuses.

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, establishment, and firm level, proper identification of the entities is important, and we briefly highlight the steps undertaken to edit and verify the identifiers. A more detailed analysis of the longitudinal editing of individual record identifiers using probabilistic record linking 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 these Infrastructure Files are brought together to create the QWI. 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 micro data supplied by the underlying entities. How their identities and data are 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. A brief description of these files with pointers to more detailed documentation is provided in Section 8. Section 9 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 effort documented in this paper. As far as we are aware, in addition to the authors of this paper, the following individuals, who are or were part of the LEHD Program and other parts of the Census Bureau, contributed to the design, implementation, and dissemination of the Infrastructure Files and the Quarterly Workforce Indicators: Romain Aeberhardt, Charlotte Andersson, Matt Armstrong, B.K. Atrostic, Sasan Bakhtiari, Nancy Bates, Gary Benedetto, Melissa Bjelland, Lisa Blumerman, Holly Brown, Bahattin Buyuksahin, Barry Bye, John Carpenter, Nick Carroll, Pinky Chandra, Hyowook Chiang, Karen Conneely, Rob Creecy, Anja Decressin, Pat Doyle, Lisa Dragoset, Robert Dy, Colleen Flannery, Matt Freedman, Kaj Gittings, Cheryl Grim, Matt Graham, Matt Harlin, Sam Hawala, Sam Highsmith, Tomeka Hill, Rich Kihlthau, Charlene Leggieri,

Section 2

Input files

The LEHD Infrastructure File system is, fundamentally, a job-based frame designed to be represent the universe of individual-employer pairs covered by state unemployment insurance system reporting requirements.[2] Thus, the underlying data are wage records extracted from Unemployment Insurance (UI) administrative files from each LED partner state. In addition to the UI wage records, LED partner states also deliver an extract of the file reported to the Bureau of Labor Statistic's 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 many 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 (recoded to the State Employer Identification Number (SEIN) in the LEHD system). An individual's UI wage record is retained in the processing if at least one employer reports earnings of at least one dollar for that individual during the quarter. Thus, an in-scope job must produce at least one dollar of UI-covered earnings during a given

quarter in the LEHD universe. Maximum earnings reported are defined in a specific state's unemployment insurance system, and observed top-coding varies across states and over time.

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

Current UI wage records are reported for the quarter that ended approximately six months prior to the reporting date at Census (the first day of the calendar quarter). Wage records are also reported for the quarter that the state considers "final" in the sense that revisions to its administrative UI wage record data base after that date are relatively rare. This quarter typically ends nine months prior to the reporting date. Historical UI wage records were assembled by the partner states from their administrative record backup systems.

2.2   Employer reports: ES-202

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 jointly administered by the U.S. Bureau of Labor Statistics (BLS) and the Employment Security Agencies in a federal-state partnership. This cooperative program between the states and the federal government collects employment, payroll, and economic activity, and physical location information from employers covered by state unemployment insurance programs and from employers subject to the reporting requirements of the ES-202 system. The employer and work place reports from this system are the same as the data reported to the BLS as part of the Quarterly Census of Employment and Wages (QCEW), but are referred to in

the LEHD system by their old acronym "ES-202." The universe for these data is a "reporting unit," which is the QCEW establishment–the place where the employees actually perform their work. Most employers have one establishment ("single-units"), but most employment is with employers who have multiple establishments ("multi-units"). One report per establishment per quarter is filed[3].

The information contained in the ES-202 reports 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 includes the federal EIN, geography both at an aggregated civil level (county or MSA) and at a detailed level (physical location street address and mailing address). A recent expansion of the standard report's record layout has increased the informational content substantially.

2.3   Administrative demographic information: PCF and CPR

The UI and ES-202 files are the core data files describing the economic activity of individuals, jobs and employers. Although these files contain a tremendous amount of detail on the economic activity, they contain little or no demographic information on the individuals. Demographic information comes from two administrative data sources–the Person Characteristics File (PCF) and the Composite Person Record (CPR), compiled by the Planning, Research, and Evaluation Division at the Census Bureau.[4]   The PCF contains information on sex, date of birth, place of birth, citizenship, and race, most of which is extracted from the Social Security Administration's Numident file–the data base containing application information for Social Security Numbers sorted in SSN

order. The CPR information contains annual place of residence data compiled from the Statistical Administrative Records System (StARS).

2.4 Demographic product integration

As part of the integration of individual and household demographic information, the LEHD system uses the fact that many individuals were part of respondent households in the Survey of Income and Program Participation (SIPP) or the March Current Population Survey (CPS). Identifier information from the 1984, 1990-1993, and 1996 SIPP panels as well as from March Demographic Supplement to the CPS from 1983 forward have been integrated into the system. See the discussion of the Individual Characteristics File system below.

2.5 Economic censuses and annual surveys integration

The LEHD Infrastructure Files include a crosswalk between the SEIN/SEINUNIT and the federal Employer Identification Number (EIN). This cross walk can be used to integrate data from the 1987, 1992, 1997 and 2002 economic censuses, all annual surveys of manufacturing, service, trade, transportation and communication industries and selected, approved fields from the Census Bureau's Employer and Nonemployer Business Registers. The integration is used for research to improve the economic activity and geocoding information in both the Infrastructure Files and the Business Registers. The integration of these data is based upon exact EIN matches, supplemented with statistical matching to recover establishments. See the discussion of the Business Register Bridge below.

2.6 Identifiers and their longitudinal consistency

Both the wage records and employer reports are administrative data–comprehensive, but some-times less than perfect. Spurious changes in the entity identifiers (Social Security Number for individuals, SEIN/SEINUNIT for employers and establishments) used for longitudinal matching can have a significant impact on most economic uses of the data. This section discusses the proce-dures implemented in the LEHD Infrastructure Files to detect, edit, and manage these identifiers.

2.6.1 Scope of data and identifiers

In the LEHD system, a person is identified initially by Social Security Number, and later by the 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 the reporting require-ments of the state's unemployment insurance system. The prime exclusions are agriculture and some parts of the public sector, particularly federal, military, and postal works. Coverage varies across states and time, although on average, 96% of all private-sector jobs are covered. The BLS Handbook of Methods (Bureau of Labor Statistics; 1997a, p. 42) describes UI coverage as "broad and basically comparable from state to state" and claims "over 96 percent of total wage and salary civilian jobs" were covered in 1994. Stevens (2002) provides a survey of coverage for a subset of the current participant states in the LEHD system.

An employer is identified primarily by its state UI account number (recoded to SEIN). A sin-gle legal employer might have multiple SEINs but regardless of its operations in other states a legal employer has a different unemployment insurance account in each state in which it has statu-tory employees. In particular, the QWI are based exclusively on SEIN-based entities and their associated establishments. Since the SEIN is specific to a state, the QWI does not account for simultaneous activity of individuals across state lines, but within the same multi-state employer.

Such activity appears as distinct jobs in the universe. Time-consistency is also not guaranteed, since the UI account number associated with an employer can also change (see later discussions).

Although the QWI are based on SEIN/SEINUNIT establishments, this restriction does not apply to the Infrastructure Files themselves. Using the federal EIN, reported on the ES-202 extract and stored on the Employer Characteristics File (ECF) and the Business Register Bridge (BRB), research links to the Census Employer and Nonemployer Business Registers (BR) permit analyses that map entities from the QCEW universe to the Census establishment universe even when the employer-entity operates across state lines. (See Section 8.2.3 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; 1997b, pg. ii). The report noted that 38% of all records were entered by key entry, while another 11% were read in by optical character readers. OCR and magnetic media tend to be less prone to errors.

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 harder to identify and to correct, the LEHD system uses statistical matching techniques, primarily probabilistic record linking, to correct for spurious and non-persistent coding errors. The incidence of errors and the success rate of the error correction methods differs widely by state. In particular, it depends critically on the quality of the available individual name information on the wage records.

Abowd and Vilhuber (2005) describe and analyze the LEHD SSN editing process as it was applied to data provided by the state of California. The process verified over half a billion records

for that state and is now routinely applied to all states in the LEHD Infrastructure Files. 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 the 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%.

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, because they are covered by the Privacy Act, are replaced by the Protected Identification Key (PIK) early in the processing of wage records. The PIK is used for all individual data integration. The original SSN and the individual's name are not part of the LEHD Infrastructure Files. 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 employer identifiers (see the next section).

The Census Bureau designed the PIK as a replacement for the Privacy Act protected SSN. The PIK itself is a random number related to the SSN solely through a one-to-one correspondence table that is stored and maintained by the Census Bureau on a computing system that is isolated from all LEHD systems and from most other systems at the Census Bureau. To avoid any commingling of SSN-laden data with PIK-laden data, which might compromise the protection afforded by the PIK, the wage record editing process takes place in a secure and distinct computing area from the rest of the LEHD processing.

2.6.3 Correcting for changes in firm identifiers

Firms in the QCEW system are identified by a UI account number assigned by the state. As with all employer identifiers, an account number can change over time for a number of reasons, not all of which are due to economically meaningful changes. 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 change in ownership of a firm without any change in economic activity may lead to a change in the account number. Often, but not always, such a change is noted in the successor/predecessor fields of the ES-202 record. Other times, without changes in ownership, employees migrate *en masse* from one UI account to another. In this case, one might make a reasonable inference that there were continuous economic operations.

Because changes in the employer identifiers are correlated with some elements of economic choice, albeit imperfectly, these identifiers are managed in the LEHD Infrastructure File system. Because the system is designed to operate from regular reports of the administrative record systems in the partner states, the original employer identifiers must be retained in all files in the system. The LEHD system then builds a database of entity demographics that traces the formal successor/predecessor relations among these identifiers. In addition, entity-level summary inferences about undocumented successor/predecessor relations, which are based on worker flow statistical analysis, are also stored in this entity demography data base. An auxiliary file, the Successor-Predecessor File (SPF), is created from the entity demographic histories and used to selectively apply successor/predecessor edits to the input files for the QWI. Handling the entity identifiers in this manner allows the LEHD system to receive and integrate updates of input data from partner state (because these share common entity identifiers) and to purge statistical analyses of the spurious changes due to non-economic changes in the entity demography over time. Benedetto et al. (2007) provide more detail on the development of the SPF and its validity. The SPF is described in more detail later in this paper.

Section 3

Infrastructure Files

This Section describes the creation of the core Infrastructure Files from the raw input files. These files form the core of the integrated system that supports the job-based statistical frame that LEHD created. Each Infrastructure File is integrated into the system with longitudinally consistent identifiers that satisfy fundamental database rules allowing them to be used as unique record keys. Thus, the core Infrastructure File system can be used to create valid statistical views of data for jobs, individuals, employers, or establishments. The system is programmed entirely in SAS and all files are maintained in SAS format with SAS indices.

The raw input files, quarterly UI wage records and ES-202 reports, are first standardized.[5] The UI wage record files are edited for longitudinal identifier consistency, and the SSN is then replaced by the PIK. The ES-202 files are standardized but no identifier or longitudinal edits are performed at this stage. Thus, the raw input files with only the edits noted here, are preserved for future research. Beyond these standardizing steps, no further processing of the raw files occurs. Instead, all the editing and imputation are done in the process of building the Infrastructure Files.

The LEHD system builds the Infrastructure Files from the standardized input files augmented by a large number of additional Census-internal demographic and economic surveys and censuses. The Employment History File (EHF) provides a full time-series of earnings at all within-state jobs

for all quarters covered by the LEHD system and provided by the state.[6] It also provides activity calendars at a job, SEINUNIT, and SEIN level. The Individual Characteristics File (ICF) provides time-invariant personal characteristics and some address information.[7] The Employer Characteristics File (ECF) provides a complete database of firm and establishment characteristics, most of which are time-varying. The ECF includes a subset of the data available on the Geocoded Address List (GAL), which contains geocodes for the block-level Census geography and latitude/longitude coordinates for the physical location addresses from a large set of administrative and survey data, including address information in the ES-202 input files. We will describe each of these files in detail in this section.

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 at their employing SEIN (a theoretical possibility) in a given year do not have a record in the EHF for that year. The EHF data are restructured into a file containing one observation per job (PIK-SEIN combination), with all quarterly earnings and activity information available on that record. The restructured file is called the Person History File (PHF).[8] An active job within a quarter, the primary job-level economic activity measure, is defined as having strictly positive quarterly earnings for the individual-employer pair that define the job.

A similar time-series, based on observed activity (positive employment) in the ES-202 records, is computed at the SEINUNIT level (UNIT History File, UHF) and the SEIN level (SEIN History File, SHF).

At this stage of the data processing the first major integrated quality control checks occur. The

14

system performs a quarter-by-quarter comparison of the earnings and employment information from the UI wage records (beginning-of-quarter employment, see Appendix A, also available at www.nber.org/books/dunno5-1, for the definition, and total quarterly payroll) and ES-202 records (month one employment and total quarterly payroll). Large discrepancies in any quarter are highlighted and the problematic input files are passed to an expert analyst for study. Discrepancies that have already been investigated and that will, therefore, be automatically corrected in the subsequent processing of a state's data are allowed to pass. Other discrepancies are investigated by the analyst. The analyst's function is to find the cause of the discrepancy and take one of three course of action:

- arrange for corrected data from the state supplier,

- develop an edit that can be applied to correct the problem

- or flag the data as problematic so that they are not used in the QWI estimation system.

The first two actions result in a continuation of the Infrastructure File processing and no change in the QWI estimation period. The third action results in continuation of the Infrastructure File processing and either the suppression of a state's QWI data until the problem can be corrected or a shortening of the time period over which QWI data are produced for that state. Often, a state-supplied corrected data file is imported into the LEHD system. Equally often, a state-specific edit is built into the data processing. Each time the state's data are reprocessed, this edit is invoked. Unfortunately, not all data discrepancies can be resolved. Then, the third action occurs. In particular, the state's archival historical UI wage record and ES-202 data are sometimes permanently damaged or defective. In these cases, the data have been lost or permanently corrupted. The quality control during the EHF processing identifies the state and quarter when such problems occur. In the current Infrastructure File system, such data are not used for the QWI estimation but may be used by analysts for specific research projects. In the course of such research projects, the analyst

often develops a statistical method for improving the defective data. These improvements are then ported into the Infrastructure File system.[9]

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. Basic demographic information from the PCF is merged using the PIK, and records without a valid match are flagged. PIK-survey identifier crosswalks link the CPS and SIPP ID variables into the ICF. Sex and age information from the CPS is used to complement and verify the PCF-provided information.

3.2.1   Age and sex imputation

Approximately 3% of the PIKs found in the UI wage records do not link to the PCF. Multiple imputation methods are used to impute date of birth and sex for these individuals. To impute sex, 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 the time period spanned by the state's records, for each individual i with strictly positive earnings within state s and non-missing PCF sex. 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 sex within state s, and sex 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 sex is assigned ten independent missing data implicates, all of which are used in the QWI processing.[10]

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 birth date decades and then assigning a birth date within decade based on this probability and the distribution of birth dates within the decade. Again, ten implicates are imputed for birth date.

If an individual is missing sex or birth date in the PCF, but not in the CPS, then the CPS values are used, not the imputed values. Before the imputation model for date of birth is implemented, basic editing of the date of birth variable eliminates obvious coding errors, such as a negative age at the time when UI earnings are first reported for the individual. In those relatively rare cases where the date of birth information is deemed unrealistic, birth date is set to missing and 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 less than ten percent of all cases the geography information is incomplete or missing. Since the QWI estimation relies on completed place of residence information, because this information is a critical conditioning variable in the unit-to-worker (U2W) imputation model (see Section 4.2), all missing residential addresses are imputed.

County of residence is imputed based on a categorical model of the data that is a fully-saturated contingency table. Separately for each state, unique combinations of categories of sex, 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

$$county = j \text{ if } P_{ij-1} \le u_k < P_{ij}$$

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.[11]

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:

$$P(educat) = f(Z_{is}\gamma_s) \tag{3.3}$$

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. The education category is imputed based on

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

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 $i \in EHF$.[12]

3.3 The Geocoded Address List: GAL

The *Geocoded Address List* (GAL) is a file system containing the unique commercial and residential addresses in a state geocoded to the Census block and latitude/longitude coordinates. The file encompasses addresses from the state ES-202 data, the Census Bureau's Employer Business Register (BR), the Census Bureau's Master Address File (MAF), the American Community Survey Place of Work file (ACS-POW), the American Housing Survey (AHS) and others. Addresses from these source files are processed by geocoding software (Group1's Code1), address standardizers (Ascential/Vality), and record matching software (Ascential/Vality) for unduplication. The remaining processing is done in SAS and the final files are in SAS format.

The final output file system 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).

3.3.1 Geographic codes and their sources

A geocode on the GAL is constructed as the concatenation of FIPS (Federal Information Processing Standard) state, county and Census tract:

$$\texttt{FIPS-state(2)} \| \texttt{FIPS-county(3)} \| \texttt{Census-tract(6)}$$

This geocode 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 some block-level analyses unreliable. Geocoding to the block allows the addition of all the higher-level geocodes associated with the addresses. Latitude and longitude coordinates are also included in the file.[13]

3.3.1.1 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 11.1 describes the typical distribution of geocode sources.

[see Table 11.1 on page 82]

In all states processed to date, 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 using 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 codes are unique within the U.S.; Census codes are not. Table 11.2 lists the available higher-level geocodes.

[see Table 11.2 on page 83]

3.3.1.2 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 of their quality is provided. Table 11.3 provides the typical distribution of codes, which range from 1 (highest quality) to 9 (lowest quality).

[see Table 11.3 on page 84]

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. As another example, if a state partner supplies 1991 ES-202 data with no address information, then E1991 will be 0 for all addresses. In a typical GAL year, between 3 and 6 percent of addresses are present on that 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

21

than one percent of addresses are found on the ACS-POW and AHS data, because these are sample surveys. Note that this distribution indicates where the GAL found it geocoding, not the percentage of addresses that could be geocoded.

3.3.2   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, users merge the GAL crosswalk to the GAL by GALID, selecting only observations existing on the required entities 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 11.4 lists the entity identifiers by data set or survey.

3.4   The Employer Characteristics File: ECF

The *Employer Characteristics File* (ECF), which is actually a file system, consolidates most employer and establishment-level information (size, location, industry, etc.) into two files. The employer, SEIN-level, file contains one record for every year-quarter in which an SEIN is present in either the ES-202 or the UI wage records, 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 is an easier and more efficient way to access SEIN-level summary data.

A number of inputs are used to build the ECF. The primary input is the ES-202 data UI wage record summary data are used to supplement information from the ES-202; in particular, SEIN-

level employment (beginning of quarter, see Appendix A, also available at www.nber.org/books/dunno5-1, for definitions) and quarterly payroll measures are built from the wage records. UI wage record data are also used to supplement published BLS county-level employment data, which are used to construct weights for use in the QWI processing. Geocoded address information from the GAL file contributes latitude-longitude coordinates of most establishments, as well as updated Workforce Investment Board (WIB) area and MSA information. The state-provided extracts from the BLS Longitudinal Database (LDB) and LEHD-developed imputation mechanisms are used to backfill NAICS information for periods in which NAICS was not collected. Finally, the QWI disclosure avoidance mechanism is initiated in the ECF. We will describe basic methods for constructing the ECF in Section 3.4.1. Details of the NAICS imputation algorithm are described in Section 3.4.3. The entire disclosure-proofing mechanism described in Section 6.

3.4.1 Constructing the ECF

ECF processing starts by integrating yearly summary files for each SEIN and SEINUNIT in the ES-202 data files. General and state-specific consistency checks are then performed. The county, NAICS, SIC, and federal EIN data are checked for invalid values. The industry code edit goes beyond a simple validity check. If a 4-digit SIC code or NAICS industry code (6-digit) is present, but is not valid, then the industry code undergoes a conditional missing data imputation based on the first 2 and 3 (SIC) or 3,4 and 5 (NAICS) digits.[14] All other invalid or missing industry codes are subjected to the longitudinal edit and missing data imputation described below.

Based on the EHF, SEIN-level quarterly employment (beginning of quarter) and payroll totals are computed. UI wage record data are used as an imputation source for either payroll or employment in the following situations:

- if ES-202 month one employment is missing, but ES-202 payroll is reported, then UI wage record beginning-of-quarter employment is used.

- if ES-202 month-one 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 retroactively paid, even though no employees were actively employed.

- if ES-202 quarterly payroll is zero and E-S202 employment is positive, then UI wage record quarterly payroll is used.

- if ES-202 quarterly payroll and employment are both zero or both missing, then UI wage record quarterly payroll and beginning-of-quarter employment are used.

The ES-202 data contain 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 handled at this stage of the processing. For a single unit SEIN, which has two records—master and establishment, information from the master records is used to impute missing data items directly for the establishment record. For a multi-unit SEIN, a flat prior is used in the allocation process—missing establishment data are imputed assuming that each establishment has an equal share of unallocated employment and payroll. A subsequent longitudinal edit re-examines this allocation and improves it if there is historical information that is better than the equal-size assumption.

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 establishment structure for some periods, but reported a valid multi-unit structure in other periods, are inspected. The absence of information on establishment structure typically occurs when an SEIN record is missing due to a data processing error. An SEIN with a valid multi-unit structure in a previous period is a candidate for structure imputation. The employer's establishment 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 employer's establishment structure (number of establishments per SEIN) is defined for all periods. Geocoded data from the GAL are incorporated to obtain geographic information on all establishments.

Once the multi-unit structure has been edited and the geocoding data have been integrated, the ECF records undergo a longitudinal edit. Geographic data, industry codes (SIC and NAICS), and EIN data from quarters with valid data are used to fill missing data in other quarters 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. All missing data imputations are single draws from posterior predictive distributions that are multinomial based on an improper uniform Dirichlet prior. The imputation probabilities are the ratio of employment in each possible value to total employment in the support of the distribution.[15]

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. Missing SIC and NAICS are randomly imputed with probability proportional to the state-wide 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, provisional weights are calculated. These weights are discussed in the section on QWI processing (Section 5). The disclosure avoidance noise-infusion

factors are also prepared at the SEIN and SEINUNIT level and added to the ECF at this point. Disclosure avoidance methods are discussed in detail in Section 6.

3.4.2 Imputations in the ECF

All employer or establishment data items used in the QWI processing, when missing, are imputed. These items include employer-establishment structure, employment, payroll, geography, industry, ownership, and EIN. This subsection describes these imputations, which are of two types: longitudinal edits–data from another period closest in time to the period with missing data are copied into the missing data items–and probabilistic imputation–missing data are imputed by sampling from a posterior predictive Multinomial distribution based on a uniform Dirichlet prior, conditional on as much sample information as possible. The analyst is responsible for developing the likelihood component of the posterior predictive distribution.

The employer-establishment structure refers to the structure of establishments withn single-unit and multi-unit employers. In the ECF, the SEIN master record summarizes the information from all establisments. This record is either based on the comparable record in the raw ES-202 data (input directly or aggregated from the establishment records) or imputed by calculating summary information on beginning-of-quarter employment and total quarterly payroll directly from all UI wage records in a given quarter that come from the indicated SEIN (in the case where the SEIN does not have a record in the raw ES-202 data for that quarter). In either case, an SEIN master record is always available for every SEIN that exists in a given quarter in either the ES-202 or UI wage record data for that quarter. However, the establishment structure of this SEIN may be missing in a given quarter; that is, the SEINUNITs associated with the SEIN for this quarter are not input directly from the ES-202 data. In this case, the establishment structure is imputed by a longitudinal edit that looks for the nearest quarter in which the establishment structure is not missing and copies this structure to the quarter with the missing structure. Then, the missing SEINUNIT employment and payroll are imputed from the SEIN master record by proportionally

allocating the current quarter SEIN-level values to the SEINUNITs based on the proportions of the same variables in the donor quarter's establishments. Only longitudinal edits are used in this process. If no donor quarter can be found, then the establishment structure is assumed to be single unit and a single SEINUNIT record is built from the SEIN master record.

At this point, the employer-establishment structure is available for all SEINs and all missing employment and payroll data have been imputed for every SEIN and SEINUNIT that exists in a state's complete ECF. The ECF records are then geocoded from the GAL, as described in section 3.3. Hence, the missing geocode items are completed before the remainder of the missing data in the ECF are imputed.

The geography subprocess of the ECF combines information about the entity history with the geocoding information from the GAL. Geocodes in the GAL are determined exclusively by contemporaneous address information, but contain information on the quality of the geocode information–whether a geocode reflects a rooftop geocode, a block, a block group, a tract, or only a county. The ECF geography subprocess takes this information, and applies a longitudinal edit, conditional on the SEINUNIT not changing locations.

The inference of a geographical move for an SEINUNIT occurs whenever the geocode delivered by the GAL is different for two different time periods in a way that is not due to variations in the quality of geography coding. For example, a rooftop and a block group geocode will always necessarily have different geocodes. However, if the block groups corresponding to each entity differ then, the system assumes that the entity has physically moved. If the two SEINUNITs have been geocoded to the same block group, the difference in geocodes is considered a change in geography quality, not a move. Finally, the GALID associated with the best quality geography is copied to all quarters within the non-move time period for that SEINUNIT.

SEINUNITs with missing geography are excluded from the longitudinal edit. These units are assigned geography by a probabilistic imputation based on employment shares across counties given SIC, if the industry for the SEINUNIT is available, or by unconditional employment shares

27

across counties, if not. Each SEINUNIT with missing geography is assigned a pseudo-GALID reflecting the imputed county's centroid. Additional geographic information (MSA or CBSA, and WIB area) is attached to the ECF based on the GALID or pseudo-GALID. At this point all SEINUNIT-level records have completed geocoding.

When the records are returned from geocoding, missing industry codes, ownership, and EIN are imputed by longitudinal edit, if possible. The values of SIC, NAICS, ownership, and EIN are copied from the nearest non-missing quarter. No further editing occurs for ownership or EIN.

Industry codes that are still missing after the longitudinal edit are imputed with probabilistic methods based on empirical distribution 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.

Geocoding and industry coding are supplied for the SEIN-level record based on the following edit. The unweighted and employment-weighted modal values across SEINUNITs from the same SEIN are computed for WIB, MSA/CBSA, state, county, SUBCTYGEO, ownership, SIC, NAICS, and EIN. All SEIN-level records get assigned the both modal values (weighted and unweighted) and a researcher or analyst may choose the appropriate value.[16]

3.4.3 NAICS codes on the ECF

Enhanced NAICS variables on the ECF can be differentiated by the sources and coding systems used in their creation. There are two sources of data – the ES-202 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 commu-

nity. The ESO variables use information from the ES-202 exclusively and ignore any information that may be available on the LDB. We provide in Section 7.2.1 an analysis on why this may be preferred. The FNL variables incorporate information from both the ES-202 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.

3.4.3.1 NAICS algorithm precedence ordering

Four basic sources of industry information are available on the ECF: NAICS and NAICS_AUX as well as SIC from ES-202 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 employment-weighted SEIN modal 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 propagated to all SEINUNITs.)

3.4.3.2 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.

Section 4

Completing the Missing Job-level Data

The Infrastructure Files contain most of the information necessary to compute the QWI. However, there are two important sources of missing job-level data that must be addressed before those estimates can be formed using sub-state levels of geography and detailed levels of industry: spurious employer-level identifier changes and missing establishment-level geography and economic activity data. We discuss the edits and imputations associated these problems in this section.

Fundamentally, the QWI are based on the job-level employment histories. Dynamic inconsistencies in these histories that are caused by individual identifier breaks are handled by the wage record edit described above. Dynamic inconsistencies in these histories that are caused by employer or establishment identifier breaks that are not due to real economic activity are handled by the creating the Successor-Predecessor File from the entity demographics, then extracting information from this file to suppress spurious employment and job flows. We describe this process in subsection 4.1.

Job histories that are part of a multi-unit SEIN do not contain the establishment (SEINUNIT) associated with the job, except for the state of Minnesota. This means that establishment-level characteristics–geography and industry, in particular–are missing data for job histories that relate to multi-units. The missing imputation associated with this problem is discussed in subsection 4.2.

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

The firm identifier used in all of LEHD's files is a state-specific account number from that state's unemployment insurance accounting system, used, in particular, to administer the tax and benefits of the UI system. These account numbers recoded and augmented by a state identifier become the entity identifier called the SEIN in the Infrastructure File system. SEINs can, and do, change for a number of reasons, including a change in legal form or a merger. Typically, the separation of a worker from an employer is identified by a change in the SEIN on that worker's UI wage records. If an employer changes SEINs, but makes no other changes, the worker would appear to have left the original firm even though his or her employment status remains unchanged from an economic viewpoint. These spurious apparent employer changes are known to induce biases in both employment and job flow statistics. For example, a simple change in account numbers would lead to the observation of a firm closing even though all workers remain employed.

To identify such events, the Successor Predecessor File (SPF) tracks large worker movements between SEINs. Benedetto et al. (2007) used the SPF for an early analysis in one particular state of the impact of such an exercise. The SPF provides 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 employer 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. This is accomplished by coding in the QWI processing, not by changing any of the information in the infrastructure files.[17]

Of importance to the unit-to-worker imputation, described in subsection 4.2 is a similar measure, computed within an SEIN. For most states, and employers within states, the breakout of units

32

into SEINUNITs is at the discretion of the employer, and the employer may decide to change such a breakout. The SPF, by following groups of workers as they move between SEINUNITs, identifies spurious intra-SEIN flows, which are then ignored when doing the unit-to-worker imputation for multi-unit job histories.

4.2   Allocating workers to workplaces: Unit-to-worker imputation (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 (SEIN-UNIT) on its UI wage records.

A primary objective of the QWI is to provide employment, job and worker flows, and wage measures at a very detailed levels of geography (place-of-work) and industry. 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 legal employer (SEIN) of the workers. The ES-202 micro-data, however, are comprised of establishment-level records which provide the geographic and industry detail needed to produce the QWI. For employers operating only one establishment within a state, the assignment of establishment-level characteristics to UI wage records is straightforward because there is no distinction between the employer and the establishment. However, approximately 30 to 40 percent of state-level employment is concentrated in employers that operate more than one establishment in that state. For these multi-unit employers, the SEIN on workers' wage records identifies the legal employer in the ES-202 data, but not the employing establishment (place-of-work). Thus, establishment level characteristics–geography and industry,

in particular–are missing data for these multi-unit job histories.

In order to impute establishment-level characteristics to job histories of multi-unit employers, a non-ignorable missing data model with multiple imputation was developed. The model imputes establishment-of-employment using two key characteristics available in the LEHD Infrastructure Files: 1) distance between place-of-work and place-of-residence and 2) the distribution of employment across establishments of multi-unit employers. The distance to work model is estimated using data from Minnesota, where both the SEIN and SEINUNIT identifiers appear on a UI wage record. Then, the posterior distribution of the parameters from this estimation, combined with the actual SEIN and SEINUNIT employment histories from the ES-202 data, are used for multiple imputation of the SEINUNIT associated with for workers in a given SEIN in the data from states other than Minnesota.[18] 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-unit employer. 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-unit employers is identified. For these employers, the ECF also provides establishment-level employment, date-of-birth, and geocodes (which are 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, job 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 employers (SEINs), and $t = 1, ..., T$ index time (quarters). Let R_{jt} denote the number of active establishments at employer j in quarter t, let

$\mathfrak{R} = \max_{j,t} R_{jt}$, and $r = 1, ..., \mathfrak{R}$ index establishments. Note that the index r is nested within j. Let N_{jrt} denote the quarter t employment of establishment r in employer j. Finally, if worker i was employed at employer j in t, denote by y_{ijt} the establishment at which the worker was employed.

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

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

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

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

where

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

35

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 employer 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 posterior distribution of the estimated β parameters is combined with the entity specific posterior distribution of the α parameters 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 the imputation method

Ignoring temporal considerations, 10 implicates are generated as follows. First, using the posterior 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 ES-202 employment counts for the establishments, we compute 10 values of α_{jt} based on the hierarchical model for these parameters. Note that 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 asymptotic 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 employers (SEINs) that ever operate more that one establishment in a given quarter is identified; these SEINs represent the set of ever-multi-unit employers defined above as the set $\mathcal{J}_t$. For each of these employers, its establishment-level records are identified. For each establishment, latitude and longitude coordinates, parent employer (SEIN) employment, and ES-202 month-one employment[19] for the entire history of the establishment are retained. Those establishments with positive month-one employment in a given

quarter characterize $\mathcal{R}_{jt}$, the set of all active establishments. An establishment birth date is identified and, in most cases, is the first quarter in the ES-202 time series in which the establishment has positive month-one employment. For some employers, 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-unit employers. For each in-scope job (a worker-employer 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 employer during which there are no more than 3 consecutive periods of non-positive earnings.[20] 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.[21]   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 ES-202 first month employment) establishment of the employing employer (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 $\mathcal{R}_{jt}^i \subset \mathcal{R}_{jt}$ for a job spell (worker) at a given employer.[22]

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,[23] thereby creating no spurious labor market transitions.

Imputation and output data   Once the input data are organized, a set of 10 imputed establish-ment identifiers are generated for each job spell ending in every quarter for which both ES-202 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 (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, employers, and individuals. Next, for each employer j at time t, a set of $\hat{\alpha}_{jrt}$ are drawn from

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

which are based on the ES-202 month-one employment totals (N_{jrt}) for all candidate establish-ments $r_{jt} \subset \mathcal{R}_{jt}$ at employer j within the state (ST) being processed. The initial draws of $\hat{\alpha}_{jrt}$ from this distribution vary across time and employers but not across job spells. Combining (4.7) and (4.8) yields

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

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, employer, 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 employer. 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 employer. 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.

The imputation process yields a data file containing a set of 10 imputed establishment identifiers for each job spell. In a very small set 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 employer-level characteristics are used.

Section 5

Forming Aggregated Estimates: QWI

5.1 What are the QWI?

The Quarterly Workforce Indicators (QWI) provide detailed local estimates of a variety of employment and earnings indicators. Employment, earnings, gross job creation and destruction, and worker turnover are available at different levels of geography, including the county, Workforce Investment Area, and Core Based Statistical Area.[24] At each level of geography, the QWI are available by detailed industry (SIC and NAICS), sex, and age of workers. At the time of publication of this article (January 2008), QWI for 43 states had been published, 3 additional states were in pre-release analysis, and a total of 46 states, including the District of Columbia, had signed MOUs. The program was still expanding with the goal of national coverage.

5.2 Computing the estimates

The establishment of the LEHD Infrastructure Files was driven in large part, although not exclusively, by the needs of the QWI computations. Completed and representative job-level data, with worker and workplace characteristics, are the primary input for the QWI. The ICF (Section 3.2)

and the ECF (Section 3.4) draw on a large number of data sources, and use a set of editing and imputation procedures described above, to provide a 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 longitudinal edits and probabilistic matching rules to the improve the longitudinal linking of entities. The U2W (Section 4.2) completes the picture, by multiply imputing an employing establishment to each job reported by the multi-unit employers. Figure 10.1 provides a graphical overview of how these data sources are used in QWI processing.

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 had positive earnings at the SEIN-SEINUNIT, and when the worker did not have positive earnings–was recorded. Note that for job history associated with multi-unit SEINs, there are 10 implicate SEINUNITs (possibly non-unique) for each job, and these implicates each get a weight of 0.1 in the downstream processing.[25]

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. Exam-

ples 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 8.2.2 for details.

4. The variables necessary for applying the QWI disclosure avoidance algorithm–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.2.5). 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 processed by the disclosure avoidance procedure (see Section 6), using a comparison with the undistorted version of each indicator and appropriate cell counts. If necessary, items in some cells are suppressed, and noisy estimates are flagged as such.

5.3   Weighting in the QWI

The QWI are estimates formed from weighted sums where the weights have been controlled to state-level QCEW statistics for all private employers as published by the BLS. The control is approximate, however, because the weights are calculated from the unfuzzed beginning-of-quarter employment data whereas the publication estimates are based on the weighted sums of the noise-infused data.

When building the ECF, weights are computed such that the measured beginning-of-quarter UI employment of in-scope units, when properly weighted, is equal to the published QCEW state-wide employment in the first month of the quarter for all private employers. A preliminary weight

43

is computed as part of the ECF processing. An adjustment factor that accounts for system-wide missing data imputation and other edits, is computed in the downstream (UFF_B) processing. This adjustment factor is computed for all private establishments. The final weight is computed in the UFF_B processing to control the product of the initial weight and the adjustment factor to the state total for all private employment in that quarter's QCEW data. The same overall adjustment factor that was calculated for all private establishments is used to produce the final weights for the all establishments QWI estimates.

Selection, editing, longitudinal linking, and disclosure avoidance procedures in the micro data used to build the QWI all change the in-scope units' data somewhat–causing the preliminary and final weights to disagree. When the final weight is used for all published QWI statistics, the difference between the published QCEW statistic and the appropriate statistic in the QWI system is less than 0.5%.

Section 6

Disclosure Avoidance Procedures for the QWI

The disclosure avoidance procedures for the QWI consist of a set of methods used to protect the confidentiality of the identity and attribute of the individuals and businesses that form the underlying data in the system. In the QWI system, disclosure avoidance 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. The QWI disclosure avoidance mechanism is described and analyzed in more detail in Abowd et al. (2006); we present an overview here.

6.1 Three layers of confidentiality protection

There are three layers of confidentiality protection and disclosure avoidance protections in the QWI system. The first layer occurs when job-level estimates (computed from the EHF) are aggregated to the establishment level. The QWI system infuses specially constructed noise into the estimates of all of the workplace-level measures. We will describe the noise-infusion process in more detail

in Section 6.2. After this noise infusion, the distorted micro data item is used as the source for all published QWIs.

A second layer of confidentiality protection occurs when the workplace-level measures are aggregated to higher levels, *e.g.,* sub-state geography and industry detail. The data from many individuals and establishment are combined into a (relatively) few estimates using a dynamic weight that controls the state-level beginning of quarter employment for all private employers to match the first month in quarter employment as tabulated from the QCEW. The weighting procedure introduces an additional difference between the confidential data item and the released data item, and in combination with the noise infusion, the published data are moved away from the value contained in the underlying micro-data, contributing to the protection of the confidentiality of the micro-data.

Third, some of the aggregate estimates turn out to be based on fewer than three persons or establishments. These estimates are suppressed and a flag set to indicate suppression. Suppression is only used when the combination of noise infusion and weighting may not distort the publication data with a high enough probability to meet the criteria laid out above. Estimates such as employment are subject to suppression. Continuous dollar measures like payroll are not. All published estimates are influenced by the noise that was infused in the first layer of the protection system. When the distortion exceeds certain limits, the estimates are still published, but flagged as substantially distorted. Each observation on any one of the published QWI tables thus has an associated flag that describes its disclosure status. Table 11.5 lists all possible flags in the published QWI tables.

6.2 Details of the QWI noise infusion process

The noise infused into the QWI data is designed to have three very important properties. First, every data item is distorted by some minimum amount. Second, for a given workplace, the data

are always distorted in the same direction (increased or decreased) by the same percentage amount in every period, and in every revision of the QWI series. Third, 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, preserving both cross-sectional and time-series analytical validity. We describe below the algorithms by which the above goals are achieved. A statistical analysis providing evidence of the third goal is provided in Section 7.2.5.

6.2.1 Disclosure avoidance using noise infusion factors

To implement the multiplicative noise model in Section 6, a random fuzz factor δ_j is drawn for each establishment j according to the following process:

$$p\left(\delta_j\right) = \begin{cases} (b-\delta)/(b-a)^2, & \delta \in [a,b] \\ (b+\delta-2)/(b-a)^2, & \delta \in [2-b, 2-a] \\ 0, & \text{otherwise} \end{cases}$$

$$F\left(\delta_j\right) = \begin{cases} 0, & \delta < 2-b \\ \left[(\delta+b-2)^2\right]/\left[2\left(b-a\right)^2\right], & \delta \in [2-b, 2-a] \\ 0.5, & \delta \in (2-a, a) \\ 0.5 + \left[(b-a)^2 - (b-\delta)^2\right]/\left[2\left(b-a\right)^2\right], & \delta \in [a,b] \\ 1, & \delta > b \end{cases}$$

where $a = 1 + c/100$ and $b = 1 + d/100$ are constants chosen such that the true value is distorted by a minimum of c percent and a maximum of d percent (the exact numbers are confidential). Note that $1 < a < b < 2$. This produces a random noise factor centered around 1 with distortion of at

least c and at most d percent. Figure 10.2 depicts such a distribution. A fuzz factor is drawn for each employer and for each of the establishments associated with that employer. Although fuzz factors vary across establishments of the same employer, the fuzz factors attached to all establishments of the *same* employer are drawn from the same (upper or lower) tail of the fuzz factor distribution. Thus, if the fuzz factor associated with a particular employer (SEIN) is less than unity, then all that employer's establishments (SEINUNITs) will also have fuzz factors less than unity. It is also important to point out that a fuzz factor is attached to each SEIN and SEINUNIT only once and retained for all time periods after the initial assignment.

[Figure 10.2 about here]

6.2.2 Applying the fuzz factors to estimates

Although all estimates are distorted based on the multiplicative noise model, the exact implementation depends on the type of estimate that is computed. For completeness we show all the relevant formulas here, referring the reader to Abowd et al. (2006) for details. In all cases, the micro data noise infusion occurs at the level of an establishment estimate. However, for QWI involving ratios and changes, the basic fuzzed and unfuzzed values are combined at the publication level of aggregation to produce the released estimates. In what follows, distorted values are distinguished from their undistorted counterparts by an asterisk, *i.e.*, the true (unfuzzed) value of beginning-of-quarter employment is B, its noise-infused (fuzzed) counterpart is B^*.

Fuzzing of estimates of employment   The fuzz factor δ_j is used to fuzz all estimates of employment totals by scaling of the true establishment level statistic according to the formula:

$$X_{jt}^* = \delta_j X_{jt}, \tag{6.1}$$

48

where X_{jt} is an establishment level employment estimate: $B, E, M, F, A, S, H, R, FA, FS,$ and FH. All variable definitions are provided in A.2.1 also available at www.nber.org/books/dunno5-1.

Fuzzing of averages of magnitude estimates where the denominator is an employment estimate Ratios of magnitude estimates to employment estimates are protected by using fuzzed numerators and unfuzzed denominators according the formula:

$$ZY_{jt}^* = \frac{Y_{jt}^*}{B(Y)_{jt}} = \delta_j \frac{Y_{jt}}{B(Y)_{jt}},$$

where ZY_{jt} is a ratio of a magnitude estimate, Y_{jt}, (dollars or quarters) and $B(Y_{jt})$ is an estimate of employment. The ratio has the interpretation of an average in most cases. The variables protected according to this method are: $ZW_2, ZW_3, ZWFH, ZWA, ZWS, ZNA, ZNH, ZNR,$ and ZNS. The relevant values of Y_{jt} and $B(Y_{jt})$ are shown in the establishment-level statistics above. In the actual QWI processing the numerator and denominator of these confidentiality-protected ratios are tabulated separately for each publication category ($ownership \times state \times substate\text{-}geography \times industry \times age\ group \times sex$). Then, the publication ratio is computed when the public-use release files are created.

Fuzzing of differences of counts and magnitudes Fuzzed net job flow (JF) is computed at the aggregate level for $k = ($ $ownership \times state \times substate\text{-}geography \times industry \times age\ group \times sex$) cell as the product of the aggregated, unfuzzed rate of growth of net jobs and the aggregated fuzzed employment:

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

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 (JC) and job destruction $(JD)$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 method was used to protect estimates of wage changes for different employment estimates. The unfuzzed estimated total changes were divided by the unfuzzed denominators then multiplied by the ratio of the fuzzed denominator to the unfuzzed denominator, as in the formula:

$$Z\Delta WY^*_{kt} = \frac{\Delta WY_{kt}}{Y_{kt}} \times \frac{Y^*_{kt}}{Y_{kt}}.$$

where, again, Y denotes a particular employment, ΔWY denotes the estimated change in wages for that employment estimate and $Z\Delta WY^*$ is the confidentiality protected estimate of the ratio. This method is used for $Z\Delta WA, Z\Delta WS, Z\Delta WFA,$ and $Z\Delta WFS$. The ratio FT involves three QWI that are also in the release file. In order to protect the ratio of the fuzzed to unfuzzed estimate of full-quarter employment, the release value of FT is protected by the formula:

$$FT^*_{kt} = \frac{(FA^*_{kt} + FS^*_{kt})/2}{F_{kt}} \frac{F^*_{kt}}{F_{kt}}$$

In the actual QWI processing the numerator and denominator of these confidentiality-protected changes and ratios are tabulated separately for each publication category ($ownership\times state\times substate\text{-}geography\times industry\times age\ group\times sex$). Then, the publication change or ratio is computed when the public-use release files are created.

Section 7

Analysis of the QWI Files

In this section, we will provide some basic analysis highlighting the usefulness of the QWI as time series data on local labor market conditions and measuring the impact of the various corrections that are applied to the series.

7.1   Basic trends of some variables

The QWI are uniquely positioned to provide a picture of a dynamic workforce at a highly disaggregated level with both demographic and economic detail. In this section, we consider three variables, and provide examples of analyses that can be easily produced with the QWI. We consider employment (more precisely: begininning-of-quarter employment), job creation, and recalls We have picked the states of Illinois and Montana to illustrate the analyses.

Figures 10.3 and 10.4 show the basic data trends for the three variables, stated in thousands of workers, for both sexes combined and separately. In general, all three time-series show considerable seasonality, but job creations and recalls are considerably more variable. However, when looking at the time-series by sex, there appears to be less volatility in job creations and recalls for women than for men. Figures 10.5 and 10.6 restate these series as the percentage of women in

51

the total for each variable. In Illinois, the percentage of jobs created that are filled by recalls is significantly lower for women (46.2%) than it is for men (53%), and persistently so over time (Figure 10.3), although there is strong seasonality in this pattern as well. A similar pattern, although not quite as stable, emerges in Montana (Figure 10.3). Thus, it would seem that although women participate as much as men in job creation, they are more likely to have found a new job than to have been recalled to an old job.

Of course, this is a very simple analysis. A further breakdown by industry (also feasible using the public-use QWI) might reveal that the lower recall rate of women is a phenomenon specific to certain industries that employ a higher fraction of women for other reasons. However, it is an example that serves to highlight the utility of the demographic, geographic, and industry breakdown that is possible with the QWI.

Disaggregating statistics by geography is one of the more common strategies for policy analysts, and several data sources are available to perform such an analysis. With the QWI, geographic analysis can be extended to distinct demographic groups. In Figure 10.7, the geographic distribution of job creation is plotted for young workers 19-21 years of age, by counties in Illinois. A policy analyst could perform such an analysis for eight age groups and both sexes, using the complete QWI. Note that net job creation is computed with both the numerator and the denominator computed for workers aged 19-21, so it is not simply a decomposition of an aggregate job creation statistics–it is computed from the ground up using data only on workers between 19 and 21 years of age.

7.2 Importance of LEHD adjustments to raw data

There are numerous data edits, corrections, and imputations performed in the processing of the QWI data series. We summarize the effect of some of these adjustments here.

7.2.1 Choosing between LDB and LEHD coding of NAICS variables

As noted in Section 3.4.3, the ECF provides enhanced NAICS variables that expand on the information available on the ES-202 files. Information is imputed based on all available industry information, and backcoded to time periods that precede the introduction and widespread implementation of NAICS coding on ES-202 data. The creation of the enhanced NAICS variables was described in Section 3.4.3. In this section, we present a summary of research done on a comparison of the ESO (ES-202-only) and FNL (final) NAICS codes on the Illinois ECF.

The imputation algorithem used by the BLS to create the LDB stably backfills NAICS codes when once it has imputed a code for a later year; that is, once an establishment has received a backcoded NAICS, that code is used for all prior years of data for the establishment. The LEHD algorithm allows the backcoded NAICS to change if the contemporaneously coded SIC changes. Thus we expect the two backcodes to have different statistical properties for historical NAICS-based QWI. Although some of the SIC changes over time may be spurious, an SEINUNIT's SIC code history contains valuable information that we have attempted to preserve in the LEHD imputation algorithm. Overall, the effect of the different approaches is relatively small, since very few SEINUNITs change industry, in particular relative to the proportion of SEINUNITs that change geography.

The LDB-sourced NAICS variable is used for about 85% of the records for Illinois, the rest are filled with information from the ES-202. It is unclear why only 85% of ES-202 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

53

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.

As we go back in time a larger portion of employment can be found in NAICS FNL codes that are different from what 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 from 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 ES-202-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 ES-202-based NAICS variable uses an SIC-based imputation 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. Users should keep in mind that for early years (before 1997) some of the NAICS industries have yet to come into existence. The prevalence of this problem has not yet been investigated.

7.2.2   Correcting for coding errors in personal identifiers

Abowd and Vilhuber (2005) describe and analyze the method used at LEHD to identify coding errors in the person identifier (Social Security Number, SSN), and provide an analysis of the impact that correcting for such errors has on statistics generated from the corrected and the uncorrected data for one state (California). A simplified version of the same analysis is used as a quality assurance method during the wage record edit, and the results are similar for other states, but vary with length of available data and with prior state processing of name and SSN fields.

For California, the process verified over half a billion records. 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 require some corrective measures, which is considered conservative relative to other analyses done (see Abowd and Vilhuber (2005) for further references).

Table 11.7 on page 89 presents patterns of job histories for uncorrected and corrected data. The unit of observation is a worker-employer match (a "job"), potentially interrupted. For each such observation, the longest interruption is tabulated if there is one. If no interruption was observed during the worker's tenure with the employer, then the type of continuous job spell is tabulated. By definition, the absence of a hole implies continuous tenure, but that spell may have been ongoing in the first (*left-truncated*) or last (*right-truncated*) quarter of the data, or in both (*entire period*). If the spell was continuous, with both the beginning and the end of the job spell observed within the data, then the default code of C is assigned.

[see/insert Table 11.7 on page 89 about here]

Over 800,000 job history interruptions in the original data are eliminated by the corrections, representing 0.9% of all jobs, but 11% of all interrupted jobs (Table 11.7). 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%. On the other hand, and as expected, overall payroll W_1 are not biased, but payroll for accessions (W_A) and separations (W_S) are biased upward by up to 7% percent (Table 11.8).

[see/insert Table 11.8 on page 90 about here]

7.2.3 Identification of successor-predecessor links of firms and establishments

As noted in Section 4.1, care is taken when tracking firms and establishments over time, by tracking worker movements between firms. These corrections should have little or no impact on the time

55

series of pure stock measures (total wage bill W_1), but should influence a number of flow measures. In particular, separations (S) and accessions (A) will be reduced when between-firm (successor-predecessor) links are identified.

For this article, a small experiment was run using the standard processing stream for the QWI for a single state. Transitions associated with observed successor-predecessor flows as identified by the SPF, which are normally suppressed, were left intact. In other words, the SPF was removed from the processing stream. Comparing the resultant (unreleased) QWI with published QWI from the same time period provides an estimate of the bias due to firm links that unadjusted QWI would otherwise have.

The suppression of flows due to successor-predecessor links also affects B, beginning-of-quarter employment, which in turn is used to weight the QWI (Section 5.3). Thus, all statistics will be affected, either directly through the statistic itself, or indirectly through a change in the weights.

Analysis performed on Montana reveals that earnings and separations are 4 percent lower if successor-predecessor transitions are filtered out. Beginning-of-period employment estimates are 0.4 percent lower.

For more results, consult Benedetto et al. (2007), who have used the successor-predecessor flows in the analysis of the firm.

7.2.4 Analytical validity of the unit-to-worker imputation

This subsection presents some results of the assessment of the analytical validity of the unit-to-worker imputation process (U2W, Section 4.2). For five QWI measures–beginning-of-quarter employment (B), full-quarter employment (F), accessions (A), separations (S), and total payroll (W_1)–percentiles of the distribution of the bias induced by the imputation process for two levels of industry aggregation are presented. A complete evaluation of the validity of the unit-to-worker imputation process is provided in Stephens (2006, chapter on "Imputation of Place-of-Work in the

Quarterly Workforce Indicators").

To assess the analytic validity of the imputation process, two sets of QWI measures for the 1994:1 to 2003:4 time period were generated using the Minnesota data. The first set– $\mathbb{X}_{True}$ –is produced using the establishment of work reported on the Minnesota UI wage records. The second set– $\mathbb{X}_{Imputed}$ –is generated treating the establishment of work as unknown; thus, $\mathbb{X}_{Imputed}$ is generated using the same imputation process that is applied to other states in the QWI system. Measures for both sets were tabulated using data for all establishments in Minnesota and produced for two levels of industry aggregation–SIC Division and 2-digit SIC–as well as by county, sex, and age. For each measure, the discrepancies between values X, prior to the application of multiplicative noise factors, contained in $\mathbb{X}_{Imputed}$ and $\mathbb{X}_{True}$ for each interior quarter $\times$ industry $\times$ county $\times$ age $\times$ sex cell are calculated as:

$$bias = \frac{X_{Imputed} - X_{True}}{X_{True}}$$

Table 11.9 presents percentiles of both the weighted and unweighted distribution of the bias statistic for five QWI measures across interior data cells for SIC Division and 2-digit SIC industry aggregations. For all distributions, the median discrepancy never exceeds 0.005 in absolute value, suggesting that on average the bias induced by the imputation of place of work is relatively insignificant. For all bias statistics, the unweighted distribution is tighter than the weighted distribution, illustrating that the bias is less severe in data cells with higher levels of employment. This is expected as fewer establishments and workers contribute data to cells with relatively low levels of employment, and the lower are the number of establishments and workers contributing data, the more detectable are outliers that emerge from the imputation process. Also expected is the relative tightness of the distributions of the bias when comparing across levels of industry aggregation. The SIC Division level distributions of the bias are tighter than the 2-digit SIC distributions as more establishments and workers contribute more data to each SIC Division level cell. The tightening

distributions are clear when examining the 90-10 differential.For B at the SIC Division level, for example, the spread between the 90th and 10th percentile falls by 0.19 when the distribution of the bias is weighted. It is also clear that the spread between the 90th and 10th percentiles is smaller for the SIC Division level as compared to the 2-digit SIC level of aggregation.

[see Table 11.9 on page 91]

7.2.5 Time-series properties of disclosure avoidance system

The disclosure avoidance algorithm described in Section 6 has the dual goals of the preserving confidentiality and maintaining a high level of analytical validity of the public-use data. This section draws on Abowd et al. (2006), who provide an in-depth analysis of the extent of disclosure protection, and the degree to which analytical validity is maintained.

The analysis presented in this subsection focusses on the time-series properties of the published QWI, after noise-infusion and suppressions. Abowd et al. (2006) also show the cross-sectional un-biasedness of the published data. In each case, data from two states (Illinois and Maryland) were used. The unit of analysis is an interior sub-state geography $\times$ industry $\times$ age $\times$ sex cell kt. Sub-state geography in all cases is a county, whereas the industry classification is SIC. Analytical validity is obtained when the data display no bias and the additional dispersion due to the confidentiality protection system can be quantified so that statistical inferences can be adjusted to accommodate it.

To analyze the impact on the time series properties of the distorted data, we estimated an AR(1) for the time series associated with each cell kt, using county-level data for all counties in each state. Two AR(1) coefficients are estimated for each cell-time series. The first order serial correlation coefficient computed using undistorted data is denoted by r. The estimate computed using the distorted data is denoted by r^*. For each cell, the error $\Delta r = r - r^*$ is computed. Table 11.10 on page 92 shows the distribution of the errors Δr across SIC-division $\times$ county cells,

for B, A, S, F, and JF when comparing raw (confidential) data to published data, which excludes suppressed data items. The table shows that the time series properties of all variables analyzed remain largely unaffected by the distortion. The maximum bias (as measured by the median of this distribution) is never greater than 0.001 The error distribution is tight: the semi-interquartile range of the distortion for B in Maryland is 0.010, which is less than the precision with which estimated serial correlation coefficients are normally displayed. The maximum semi-interquartile range for any variable in either of the two states is 0.012[26].

Although the overall spread of the distribution is slightly higher when considering two-digit SIC $\times$ county and three-digit SIC $\times$ county cells, which are sparser than the SIC-division $\times$ county cells, the general results hold in these cases as well (Abowd et al.; 2006, Tables 7 and 8). Abowd et al. (2006) thus conclude that the time series properties of the QWI data are unbiased with very little additional noise, which is, in general, economically meaningless.

Section 8

Public-use and Restricted-access Files

In this section, we briefly describe the public-use release files and those files available at Census Research Data Centers. We focus on how these files differ from the corresponding internal files discussed in the rest of the article.

8.1 Public-use files

Three public-use products, fully or partially based on QWI data, are currently available on a regular basis: the QWI distribution files, the Older Worker Reports, and OnTheMap.[27] A subset of eight variables from the full public-use release is available at QWI Online (http://lehd.did.census.gov/). Additional variables are used in other applications accessible from the same Census Bureau web site. The complete set of QWI public-use variables is available from the Cornell Virtual Research Data Center (VirtualRDC) as of January 2008. The VirtualRDC is partially funded by grants from the National Science Foundation (NSF) and the National Institute on Aging. Computing resources to manipulate complete QWI are also available on the VirtualRDC for qualified researchers (http://www.vrdc.cornell.edu/). Other distribution options for the full QWI data may be available when this article appears. Up-to-date information on all access options is posted at

http://lehd.did.census.gov/. The public-use QWI data differ from the Census-internal version because the public-use version has been subjected to the disclosure avoidance methods described in previous Section 6. In order to preserve the integrity of these disclosure avoidance algorithms, all special tabulations released from the QWI must follow the same procedures.

8.2 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 removed from RDC versions of QWI and LEHD Infrastructure files relative to their internal-use counterparts is the information specifically used to do confidentiality protection of the QWI–the fuzz factors and the fuzzed data items. All of RDC-accessible LEHD files can be used for research purposes by submitting a research proposal to the Center for Economic Studies (CES) at the U.S. Census Bureau[28].

8.2.1 ECF

The version of the ECF available in the RDC environment is called the 'LEHD-ECF' on the CES RDC documentation. It is identical to the one described in Section 3.4 except for the removal of the QWI fuzz factors.

8.2.2 Unit flow files: Establishment-level QWI data

The SEINUNIT-level input files to the final aggregation step of the QWI, internally known as UFF_B, are available in the RDC environment under the name 'LEHD-QWI'. These files are identical to final establishment level flow files documented in Section 5 except that they contain only the unfuzzed raw aggregates.

8.2.3 Establishment crosswalks: Business Register Bridge

The Census Bureau maintains lists of establishments to develop the frames for economic censuses and surveys. These lists are called the Employer and Non-employer Business Registers (BR). The research version of the Employer BR is maintained by CES, which produces a new set of files 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 (SEIN/SEINUNIT). As a consequence, there is no single best way to form linkages between these data sources.

The LEHD Business Register Bridge (LEHD-BRB) that is available in the RDC network provides several ways to integrate the economic censuses and surveys with LEHD-provided data. The choice of linking strategy is left to researchers, who must determine the best definition of an entity on both side of the linkage considering data sources and the stated research objectives. Available identifiers on the LEHD-BRB that are common to both the LEHD Infrastructure Files and the BR are the EIN, geographic information, and 4-digit SIC. These variables may be used to construct pseudo-establishments that are aggregates of SEIN/SEINUNIT establishments at different levels of aggregation. These identifiers can also be linked to sets of ALPHA/CFN establishments on the BR and other Census economic data products. A more detailed guide is available on the CES or LEHD web site (Chiang et al.; 2005).

8.2.4 Household and establishment geocoding: GAL

The GAL (Geocoded Address List) described in Section 3.3, is available in the RDC environment under the reference 'LEHD-GAL'. Access to the GAL is predicated on the project having permission to use business or residential address information from other RDC-available source files. Once that permission has been properly established, the researcher is granted access to the GAL for the purpose of obtaining a consistent set of geocodes.

8.2.5   Wage decomposition data: Human capital files

These files will contain employer-level distributions of human capital measures as initially developed in Abowd et al. (2002). They are expected to become available in 2008.

8.2.6   Remaining LEHD Infrastructure Files

The remaining LEHD Infrastructure Files outlined in this document should be available in the RDC environment at time of publication of this article. In general, unless explicitly mentioned above, these files are provided to researchers as-is, and are subject to the same Title 13 use restrictions as all other data on the RDC network. The LEHD Infrastructure data are also subject to usage restriction in the MOU that governs the Census Bureau and state participation in the LED partnership. The most important of those restrictions is the one that requires the written consent of the state's signatory official on the MOU before state-specific results based on the LEHD Infrastructure Files may be released. Results may be released from analyses performed on multiple states.

Section 9

Concluding Remarks

9.1 Future projects

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

9.1.1 Planned improvements to the ICF

Currently researchers at LEHD are developing an enhanced, longitudinal version of the ICF, internally named ICF version 4 because the current system was the third version of most of the infrastructure files. The improved ICF is the first national LEHD Infrastructure File system. Individuals appearing in any state, including those that have not yet joined the LED federal/state partnership have their ICF data on a set of annual records.

Additional data sources will be integrated with this 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 imputation 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.

9.1.2 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 probable existence 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.

9.1.3 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.

9.1.4 Integration of data from missing parts of the universe

9.1.4.1 Non-employer data

The job universe currently used by all LEHD Infrastructure Files is legal employment with an employer that has mandatory reporting to the state UI wage record system. Non-employer businesses are out-of-scope for this universe but are of intrinsic interest in the economic analysis of sources of labor income. In addition, the income to the sole proprietor of an employer business is of interest as a source of labor income. The LEHD Program and CES are collaborating in developing enhancements to the Business Register to account for non-employer income sources and to better track sole proprietor employers. The non-employer enhancements will also affect the LEHD Infrastructure Files because the information on the identity of the non-employer, the identity of the non-employer business, and the income from the non-employer business provides a "job" record for this activity, which can then be integrated with the EHF, ICF and ECF file systems.

9.1.4.2 Federal government employment

The LEHD program has completed an MOU with the Office of Personnel Management in the federal government to obtain historical and ongoing information from the OPM databases that permits construction of LEHD Infrastructure File system records that correspond to "job" histories for federal employees in the EHF and employer/establishment records in the ECF. Records already exist for these individuals in the new ICF.

9.1.5 Creation of public-use synthetic data

As a part of a new National Science Foundation Information Technology Research grant (SES-0427889), awarded to a consortium of Census Research Data Centers, researchers at LEHD and other parts of Census are collaborating with social scientists and 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.

9.1.6 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 estimates 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., Stephens, B. E. and Vilhuber, L. (2006). Confidentiality protection in the Census Bureau's Quarterly Workforce Indicators, *Technical Paper TP-2006-02*, 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* **23**(2): 133–152.

Benedetto, G., Haltiwanger, J., Lane, J. and McKinney, K. (2007). Using worker flows in the analysis of the firm, *Journal of Business and Economic Statistics* **25**(3): 299–313.

Bureau of Labor Statistics (1997a). *BLS Handbook of Methods*, U.S. Bureau of Labor Statistics, Division of Information Services, Washington DC. http://www.bls.gov/opub/hom/.

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

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.

Longitudinal Employer-Household Dynamics Program (2002). The Longitudinal Employer-Household Dynamics program: Employment Dynamics Estimates Project Version 2.2 and 2.3, *Technical paper TP-2002-05-rev1*, U.S. Census Bureau.

Stephens, B. (2006). *Firms, Wage Dispersion, and Compensation Policy: Assessment and Implications*, PhD thesis, University of Maryland College Park.

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

Notes

[1]The number of participating states still increases regularly as new Memoranda of Understanding are signed and new states begin shipping data. As of January 15, 2008, there are 46 states with signed MOUs, and 43 states with public use data available at http://lehd.did.census.gov/

[2]The frame is intended to be comprenhensive for legal employment relations and self-employment. Current development efforts include the addition of federal employment via records provided by the Office of Personnel Management and the addition of self-employment via records constructed from the Employer and Non-employer Business Registers.

[3]These data are also used to compile the Covered Employment and Wages (CEW) and and Business Employment Dynamics (BED) data at the BLS.

[4]This Division has now been reconstituted as part of the Data Integration Division in the Demographic Programs Directorate at the Census Bureau.

[5]The ES-202 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.

[6]The earliest data accepted by the LEHD system are 1990, quarter 1. Most states provided data beginning some time in the early 1990s. All partner states provide data beginning in 1997, quarter 1. Current input raw data files are delivered six months after the close of the quarter. QWI data are produced within three months of the receipt of the raw input files from the unemployment insurance system. The LEHD system maintains all of the data reported by a partner state (or nationally for the national files). The QWI system uses as much of these data as possible.

[7]A longitudinal enhancement of the ICF, which updates residential address information annually and contains some data from 2000 Census of Population and Housing, is under development.

[8]It should be noted that the actual file structure is at the PIK-SEIN-SEINUNIT-YEAR level for the EHF, and at the PIK-SEIN-SEINUNIT level for the PHF. Although only one state (Minnesota) has non-zero values for SEINUNIT, this allows the file structure to be homogeneous across states.

[9]For example, research on wage dynamics associated with estimates of firm-level human capital use, has produced a statistical missing data edit for the UI wage records that detects missing wage records and imputes them by drawing from an appropriate posterior predictive distribution. The statistical models that detect and correct this problem will be imported into a future version of the EHF Infrstructure File.

[10]Note that this imputation does not account for estimation error in $\hat{\beta}$. This was one of the first missing data imputations developed at LEHD. At the time, techniques for sampling from the posterior predictive distribution of a binary outcome where the likelihood function is based on a logistic regression were not feasible on the LEHD computer system. Since only three percent of the observations in the ICF are subject to this missing data edit, it was implemented as described in the text. A longitudinal, enhanced ICF is under development (see Section 9). All missing data imputations in the new ICF will be performed by sampling from an appropriate posterior predictive distribution. This will properly account for estimation error.

[11]The longitudinal, enhanced ICF that is under development augments the model in the text with a Dirichlet prior distribution for the P_{ij}s. The imputations are then made by sampling from the posterior predictive distribution, which is also Dirichlet.

[12]In the longitudinally enhanced ICF that is under development, this imputation is replaced by a probablistic record link to Census 2000 long form data. Approximately one person in six acquires directly reported educational attainment as of 2000. The remaining individuals get 10 multiple imputations from a Dirichlet/Multinomial posterior predictive distribution.

[13]An enhanced geocoding system was developed for the newer LEHD product called OnTheMap, which published to the block level. These enhancements are being integrated into an enhanced version of the GAL, which will be used for both QWI and OnTheMap.

[14]NAICS 1997 are updated to NAICS 2002 Then, NAICS 2002 are used for the imputation. The same procedure is later used for LDB data.

[15]The posterior predictive distribution is multinomial because the employment proportions are

derived from the population of employing establishments in the quarter, which is assumed to be nonrandom. Only a single imputation is performed because the unit-to-worker missing data model imputes 10 establishments to each job in a multi-unit SEIN. Multiple imputation of the missing data in those establishment would have meant that 100 implicates would have to be processed for each multi-unit job. This processing requirement was deemed impractical for the current QWI system.

[16]The employment weighted modal values that are on the SEIN-level record are only used in the QWI processing when the unit-to-work imputation described in the next section fails to impute an SEINUNIT to a job history.

[17]A more extensive evaluation of the impact of the SPF on the aggregate QWI statistics is currently under way.

[18]The actual SEINUNIT coded on the UI wage records is used for Minnesota, and would be used for any other state that provided such data. Note that there are occasional, and rare, discrepancies between the unit structure on the Minnesota wage records and the unit structure on the Minnesota ES-202 data for the same quarter. These discrepancies are resolved during the initial processing of the Minnesota data in its state-specific readin procedures.

[19]In rare instances where no ES-202 employment is available, an alternative employment measure based on UI wage record counts may be used.

[20]A new hire is defined in the QWI as a worker who accedes 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 more than 4 quarters and is subsequently re-employed by the same firm.

[21]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 4 or fewer quarters of non-employment and subsequent re-employment by the same firm.

[22]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.

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

[24]The original QWI release used Metropolitan Statistical Areas. The older MSA definitions were replaced with CBSA definitions in 2005.

[25]In the underlying frame, a job is a PIK-SEIN pair. For single-unit employers, this is equivalent to a PIK-SEIN-SEINUNIT triple. For multi-unit employers within a single state, the original pair is completed to a triple by the unit-to-worker multiple imputation.

[26]Abowd et al. (2006) also report that the maximum semi-interquartile range for SIC2-based variables is 0.0241, and for SIC3-based variables, 0.0244.

[27]The Older Worker Reports are based entirely on the QWI public use files. OnTheMap uses the QWI micro-data to produce a QWI report for the user-defined geographic area.

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

Section 10

Figures
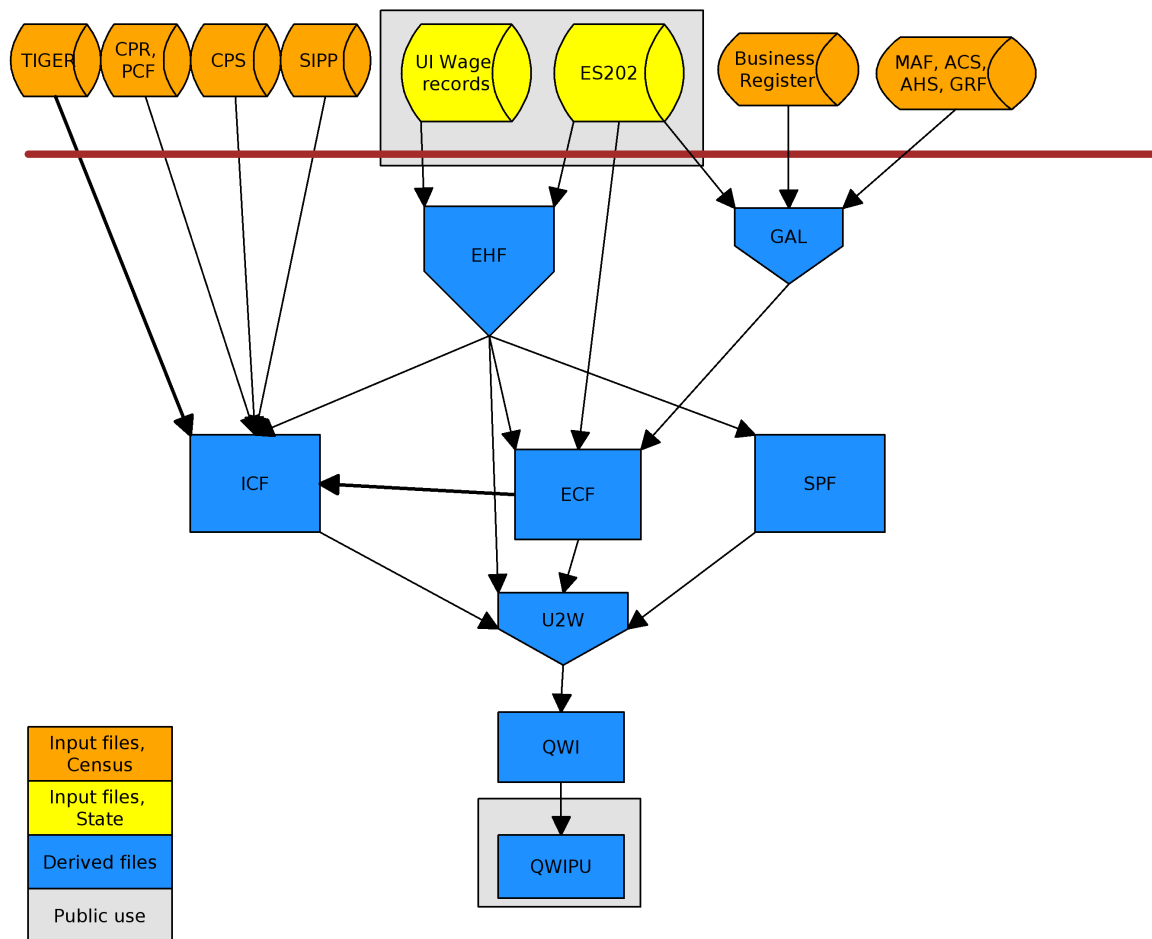
Figure 10.1: Overview of LEHD data flow

Figure 10.2: Distribution of Fuzz Factors
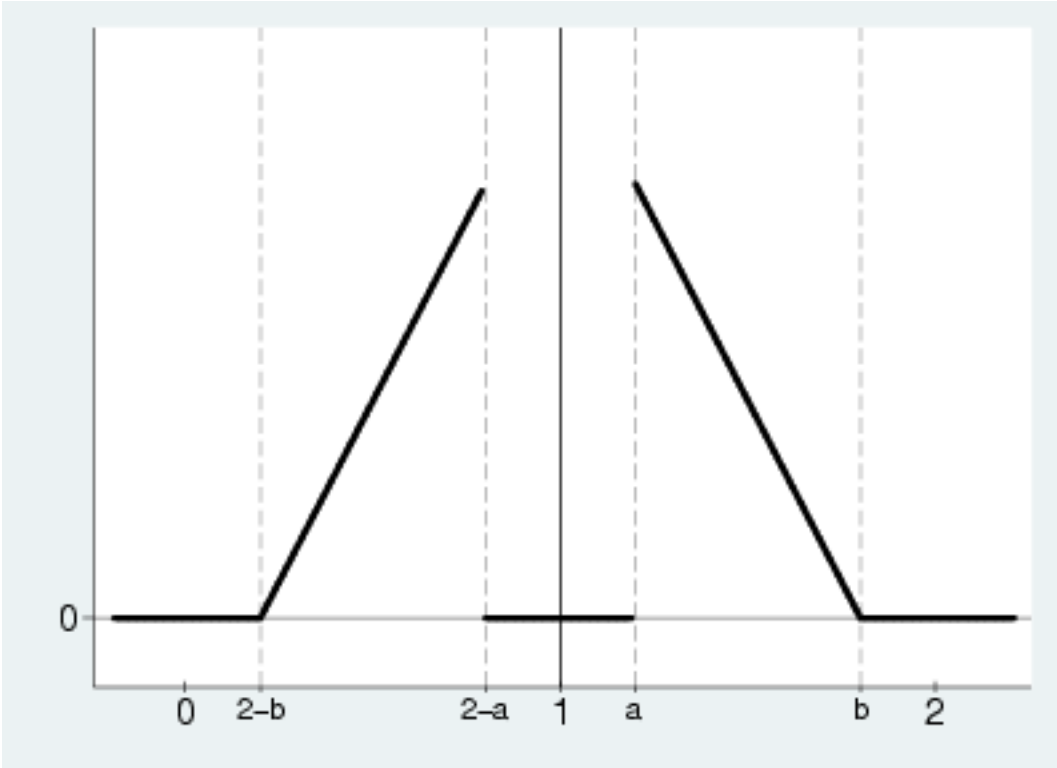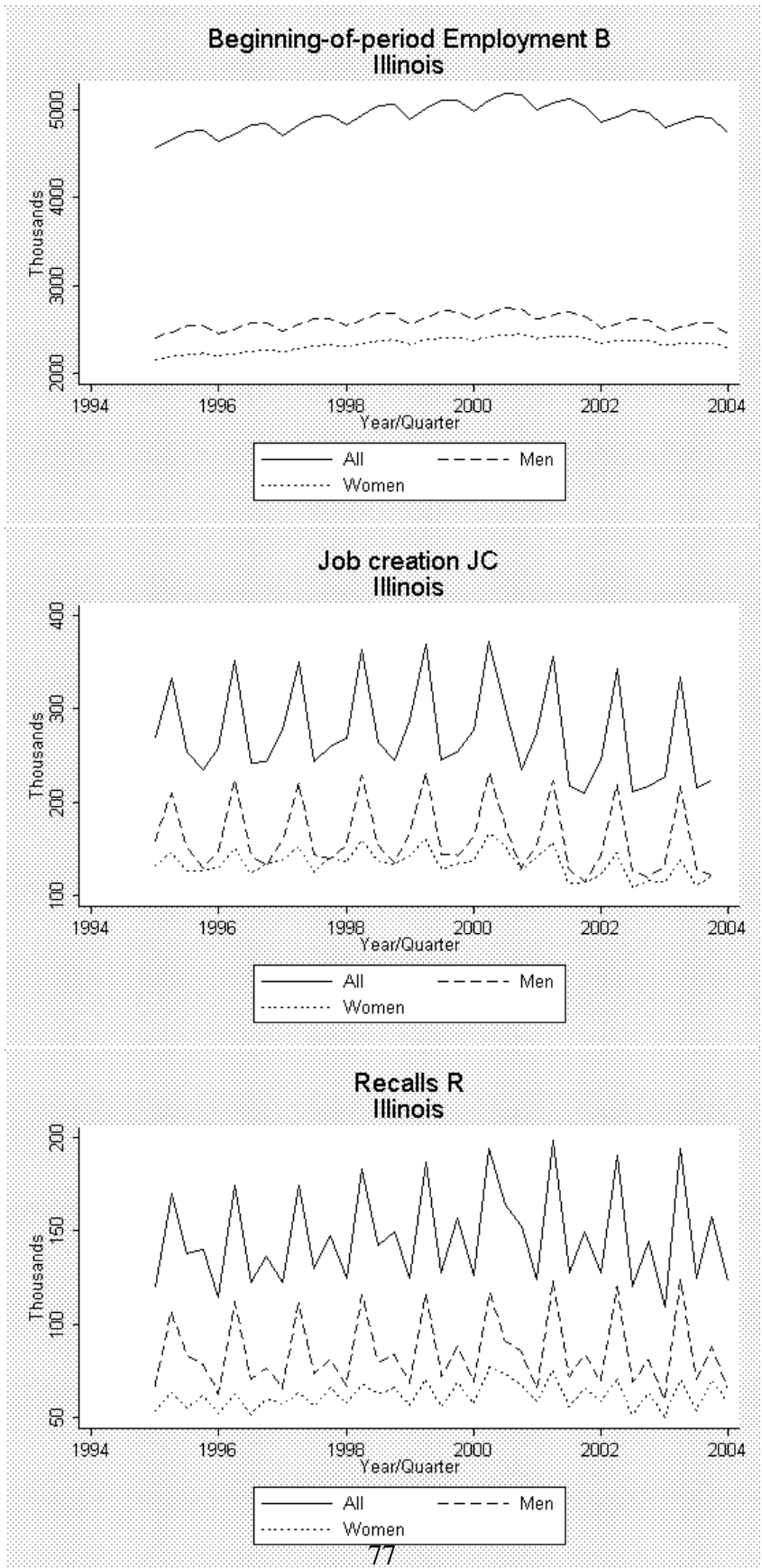
Figure 10.3: Basic Data Trends, Illinois

78
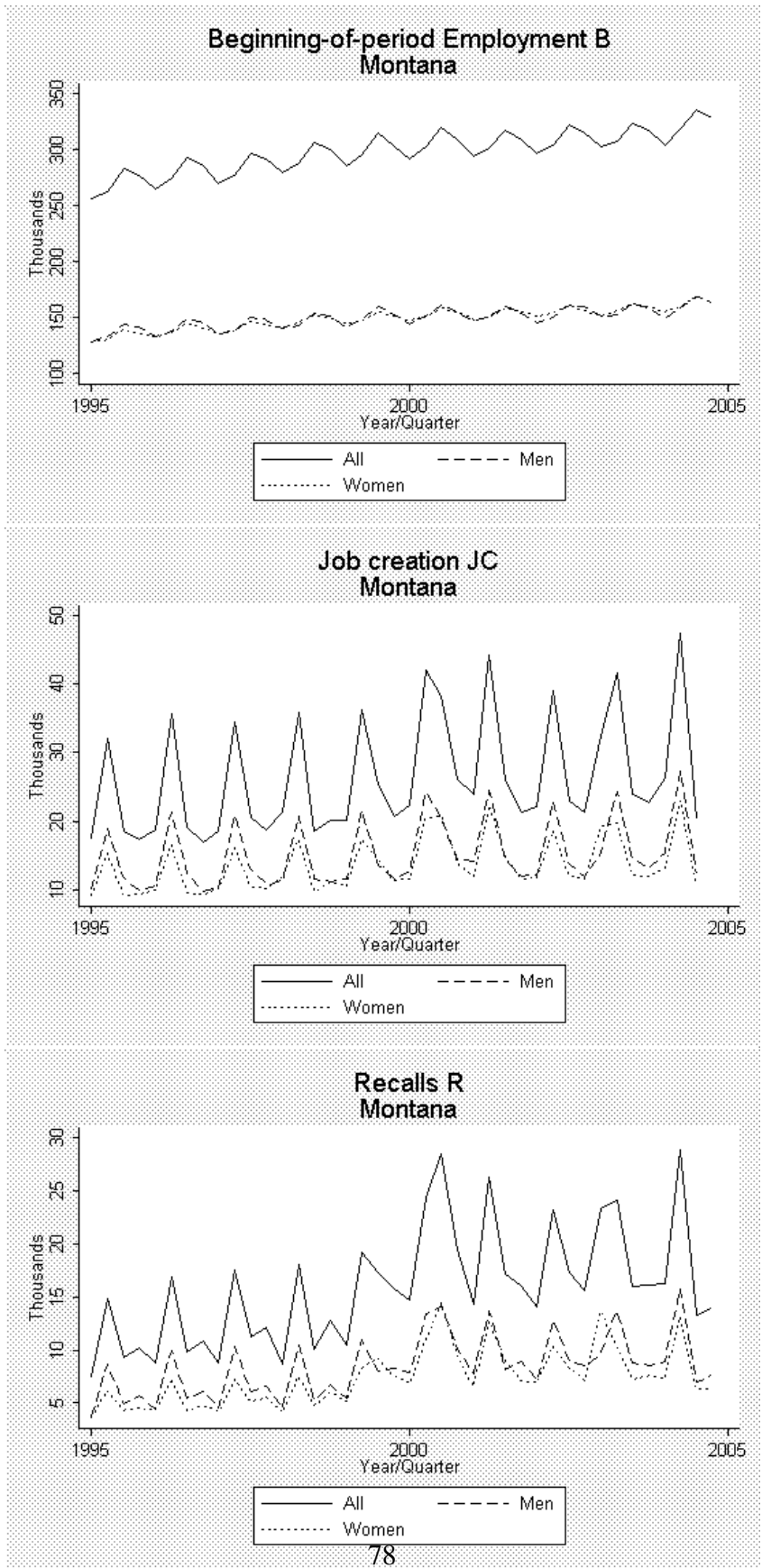
Figure 10.4: Basic Data Trends, Montana
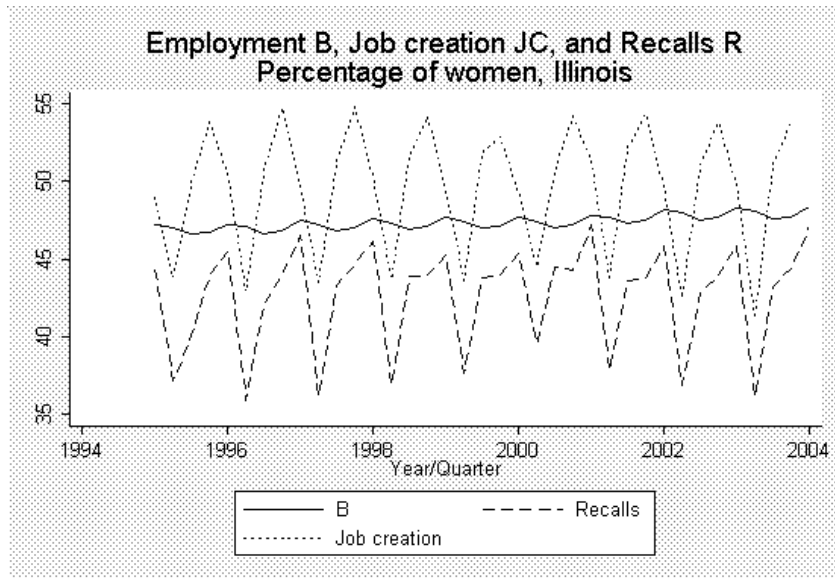
Figure 10.5: Proportion of women, select variables, Illinois
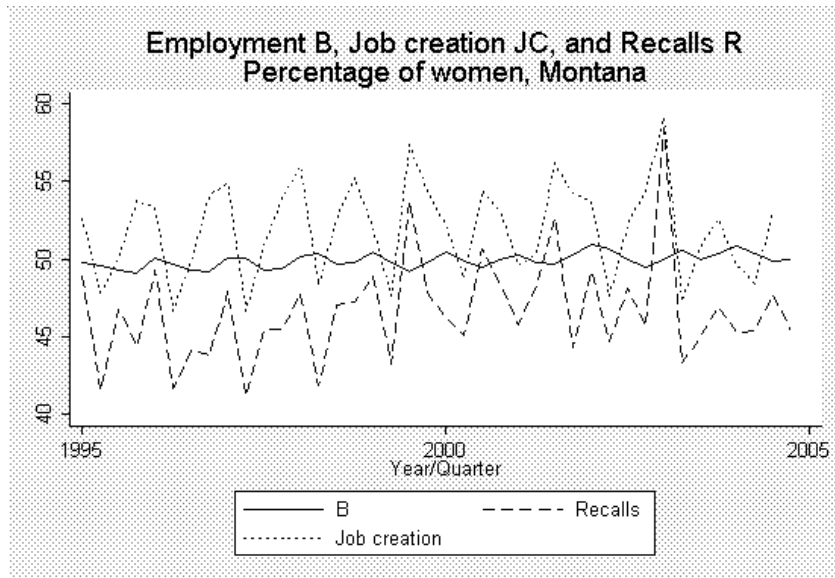
Figure 10.6: Proportion of women, select variables, Montana
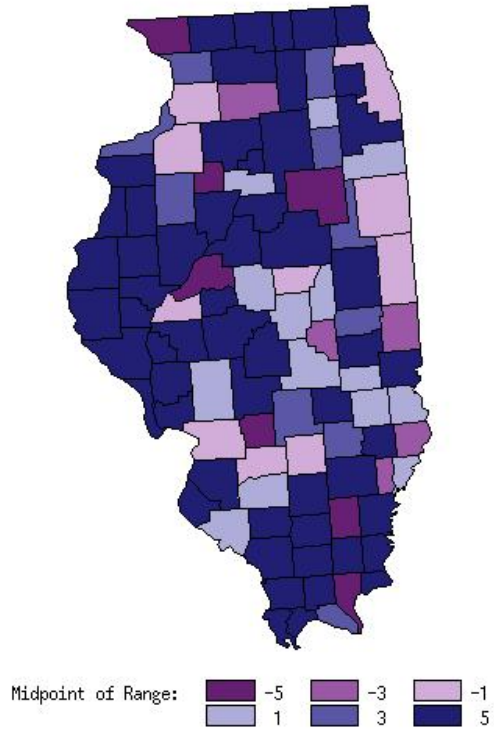
Figure 10.7: Job creation for young workers, by county, Illinois

Section 11

Tables

Table 11.1: 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	

Table 11.2: 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

Table 11.3: 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	

Table 11.4: 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`	

Table 11.5: 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	zero employment estimated or zero estimated denominator in a ratio, zero released
1	OK, distorted value released
5	Value suppressed because it does not meet US Census Bureau publication standards.
9	data significantly distorted, distorted value released

Table 11.6: Comparison of QWI and QCEW estimates

Deviation of 'All Private Employment' (B) QWI Estimates
from the 'Month 1 Employment' QCEW Estimates, in percent

Year	Quarter	Montana	Idaho
1995	1	0.289	0.333
1995	2	0.250	0.342
1995	3	0.215	0.287
1995	4	0.253	0.305
1996	1	0.244	0.224
1996	2	0.264	0.170
1996	3	0.195	0.223
1996	4	0.196	0.281
1997	1	0.294	0.111
1997	2	0.342	0.196
1997	3	0.307	0.160
1997	4	0.337	0.194
1998	1	0.264	0.079
1998	2	0.303	0.090
1998	3	0.257	0.067
1998	4	0.294	0.121
1999	1	0.129	0.106
1999	2	0.137	0.049
1999	3	0.233	0.073
1999	4	0.155	0.256
2000	1	0.118	0.369
2000	2	0.359	0.382
2000	3	0.370	0.473
2000	4	0.265	0.493
2001	1	0.090	0.562
2001	2	0.044	0.644

(cont.)

Table 11.6 (cont): Comparison of QWI and QCEW estimates

Deviation of 'All Private Employment' (B) QWI Estimates
from the 'Month 1 Employment' QCEW Estimates, in percent

Year	Quarter	Montana	Idaho
2001	3	0.070	0.623
2001	4	0.105	0.601
2002	1	-0.00	0.701
2002	2	-0.00	0.712
2002	3	0.047	0.709
2002	4	0.183	0.691
2003	1	0.114	0.731
2003	2	-0.18	0.780
2003	3	-0.12	0.801
2003	4	-0.10	0.769
2004	1	-0.19	0.667
2004	2	-0.28	0.711
2004	3	-0.22	0.670
2004	4	-0.17	0.656
Average		0.145	0.440

Table 11.7: Wage Record Edit: Comparing job histories before and after editing process

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

NOTE: For definitions of job history patterns, see text on page 55.

Table 11.8: Distribution of percentage bias in aggregate QWI statistics

All age groups, both sexes, SEIN-level micro-data

Variable (bias)	Unit	Mean	Std	N	P10	P50	P90
A	Firm	2.17%	13.98%	11,755,355			
	County	1.56%	1.01%	2,006	0.62%	1.42%	2.64%
	Industry	1.97%	2.29%	374	0.51%	1.47%	3.40%
B	Firm	-0.74%	6.14%	20,717,508			
	County	-0.46%	0.31%	1,947	-0.75%	-0.45%	-0.25%
	Industry	-0.31%	0.31%	363	-0.59%	-0.34%	-0.14%
F	Firm	-1.23%	8.05%	18,454,708			
	County	-0.78%	0.36%	1,888	-1.21%	-0.74%	-0.43%
	Industry	-0.53%	0.31%	352	-0.90%	-0.53%	-0.24%
R	Firm	4.71%	26.86%	3,242,186			
	County	5.26%	3.61%	1,888	1.70%	4.59%	9.18%
	Industry	5.95%	3.49%	352	1.93%	5.46%	10.29%
S	Firm	2.31%	14.29%	11,161,916			
	County	1.66%	1.11%	1,947	0.67%	1.46%	2.72%
	Industry	2.01%	2.08%	363	0.63%	1.53%	3.41%
W_1	Firm	-0.01%	4.96%	23,229,843			
	County	-0.01%	0.15%	2,006	-0.05%	-0.02%	0.00%
	Industry	0.04%	0.35%	374	-0.04%	-0.02%	0.08%
W_A	Firm	15.57%	1111.78%	11,755,355			
	County	4.92%	3.34%	2,006	1.89%	4.38%	8.44%
	Industry	3.95%	4.94%	374	0.77%	3.35%	6.79%
W_S	Firm	18.77%	1094.50%	11,161,916			
	County	4.87%	3.17%	1,947	2.02%	4.31%	8.06%
	Industry	3.64%	4.48%	363	1.00%	3.18%	5.71%

Note: From Table 9, Abowd and Vilhuber (2005). 23232068 firm-quarter cells, 2006 county-quarter cells, and 374 industry-quarter cells. Percentiles for firm-quarter cells are all zero, and not reported for simplication.

Table 11.9: Distribution of proportional bias in unit-to-worker imputation

$$\text{Bias} = \frac{X_{Imputed} - X_{True}}{X_{True}}$$

	Beginning of Period Employment		Full-Quarter Employment		Accessions		Separations		Total Payroll	
	Unwgt.	Wgt.	Unwgt.	Wgt.	Unwgt.	Wgt.	Unwgt.	Wgt.	Unwgt.	Wgt.
SIC Division										
90	0.21484	0.09165	0.21274	0.09021	0.27966	0.14072	0.26821	0.13265	0.28185	0.11456
75	0.07426	0.03697	0.07287	0.03709	0.09597	0.04312	0.09332	0.04043	0.09184	0.04159
50	0.00002	0.00415	0.00000	0.00482	0.00001	-0.00449	0.00000	-0.00414	0.00262	0.00405
25	-0.03475	-0.02232	-0.03696	-0.02159	-0.01610	-0.04274	-0.02105	-0.04015	-0.02998	-0.02219
10	-0.15197	-0.08075	-0.15879	-0.08020	-0.17476	-0.11827	-0.18000	-0.10717	-0.16044	-0.08178
P90-P10	0.36680	0.17239	0.37152	0.17040	0.45441	0.25898	0.44820	0.23981	0.44229	0.19633
2-Digit SIC										
90	0.25999	0.13802	0.25257	0.13494	0.30875	0.17390	0.29985	0.17117	0.36038	0.16118
75	0.04425	0.03965	0.04218	0.04002	0.05200	0.03683	0.04938	0.03622	0.05296	0.04267
50	-0.00004	-0.00010	-0.00004	-0.00005	-0.00004	-0.00111	-0.00004	-0.00075	-0.00002	-0.00004
25	-0.00232	-0.04169	-0.00232	-0.04061	-0.00075	-0.07000	-0.00075	-0.06709	-0.00127	-0.04034
10	-0.17605	-0.13951	-0.18858	-0.13865	-0.20977	-0.20784	-0.22191	-0.19383	-0.16911	-0.14427
P90-P10	0.43603	0.27752	0.44114	0.27359	0.51851	0.38173	0.52175	0.36500	0.52948	0.30544

Estimated using Minnesota data. A statistic X_{True} is calculated using job histories coded to the SEIN/SEINUNIT level; a statistic $X_{Imputed}$ is calculated using job histories coded to the SEIN level with imputed SEINUNITs. See text for further details.

Table 11.10: Distribution of the error in first order serial correlation of QWI

Estimated from undistorted (r) and published data (r*)

$$\Delta r = r - r^*$$

Percentile	Beginning of Quarter Employment	Accessions	Separations	Full Quarter Employment	Net Job Flows
		IL County x SIC Division			
01	-0.085495	-0.092455	-0.098770	-0.079205	-0.008447
05	-0.047704	-0.046665	-0.045208	-0.046830	-0.004959
10	-0.034558	-0.031767	-0.032898	-0.033607	-0.003186
25	-0.015317	-0.014197	-0.015077	-0.015533	-0.001189
50	-0.000512	-0.000997	-0.000707	-0.001000	-0.000049
75	0.013438	0.011536	0.012457	0.011670	0.000861
90	0.030963	0.027037	0.028835	0.027970	0.002489
95	0.044796	0.037906	0.041862	0.040096	0.004801
99	0.080282	0.079122	0.083824	0.077419	0.007537
		MD County x SIC Division			
01	-0.065342	-0.072899	-0.072959	-0.058021	-0.009081
05	-0.035974	-0.036995	-0.040314	-0.030985	-0.004540
10	-0.024174	-0.027689	-0.028577	-0.021361	-0.002823
25	-0.010393	-0.013686	-0.012505	-0.009401	-0.001243
50	0.000230	-0.000542	0.000797	0.000279	-0.000025
75	0.011382	0.012628	0.013034	0.009429	0.001045
90	0.025160	0.026325	0.025272	0.022027	0.002799
95	0.035176	0.034114	0.034999	0.030152	0.004321
99	0.060042	0.056477	0.055043	0.049213	0.009208

Unit of observation is a county$\times$ SIC division$\times$ age-group$\times$ sex cell for all private employment, interior cells only. For more details, see text and Abowd et al. (2006).

Appendix A

Definitions of Fundamental LEHD Concepts

A.1 Fundamental concepts

A.1.1 Dates

The QWI are 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 a State Employer Identification Number (SEIN) recoded 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.

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 recoded from 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 Section 4.2 on the unit-to-worker imputation. 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 the privacy 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 Identification 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 individual 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 QCEW, formerlyES-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. LEHD 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 individual 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 an 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 occurred and the beginning of the next quarter) the individual has acceded to continuing-

96

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 individual 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 occurred 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 accessions 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. For single-unit employers, there is never more than one SEINUNIT per quarter, so the definition at the employer level and the definition at the establishment level are equivalent. For multi-unit employers, the QWI system performs the calculations at the establishment level (SEINUNIT); however, the statistical model for imputing establishment described in Section 4.2 does not permit establishment-to-establishment flows. Hence, although the statistics are estimated at the establishment level, the sum of job creations and destructions at a given employer in a given quarter across all establishments active that quarter is exactly equal to the measure of job creations that would have been

estimated by using employer-level inputs (SEIN) directly.

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 104, also available at www.nber.org/books/dunno5-1.). 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 positive, 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 subtlety 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 occurring 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). Once again, the QWI system computes these statistics at the establishment level but does not allow establishment-to-establishment flows. Hence, the estimates for a given employer (SEIN) are the sum of the estimates for that employer's establishments (SEINUNIT) that are active in the given quarter. Net job flows 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 workers 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 employer 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 achieved 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 occurrence 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. Individual 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 accessions 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 among 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 accessions 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 individual 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 $qmax = 88$ (2006:4), as of December 2007. 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.*, $qlast = 88$ 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 beginning of t (and the end of $t-1$),

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

End of quarter employment (e): for $t < qlast$, individual i employed at j at the end of t (and the 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 end of t and the beginning 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 during period t (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 end of t and the beginning 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 during period t (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 end of t and beginning 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 during period t (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 (w_2): for $qfirst \leq t < qlast$, earnings of individual i at establishment j during period t

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

107

Earnings of full-quarter individual (w_3): for $qfirst < t < qlast$, earnings of individual i at establishment j during period t

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

Total earnings at all employers $(w_{1\bullet})$: 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 at all employers for of end-of-period employees $(w_{2\bullet})$: for $qfirst \leq t < qlast$, total earnings of individual 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 at all employers of full-quarter employees $(w_{3\bullet})$: for $qfirst < t < qlast$, total earnings of individual i during period t

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

Change in total earnings at all employers $(\Delta w_{1\bullet})$: for $qfirst < t \leq qlast$, change in total earnings of individual i between periods $t-1$ and t

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

Earnings of accessions (wa_1): for $qfirst < t \leq qlast$, earnings of individual i at 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 (wa_2): for $qfirst < t < qlast$, earnings of individual i at employer 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 (wa_3): for $qfirst+1 < t < qlast$, earnings of individual i at employer 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 (wh_3): for $qfirst+4 < t < qlast$, earnings of individual i at employer 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 (Δwa_1): for $qfirst + 1 < t \leq qlast$, earnings change of individual i at employer j during period 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 (Δwa_3): for $qfirst + 2 < t < qlast$, earnings change of individual i at employer j during period 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 (ws_1): for $t < qlast$, earnings of individual i separated from j during 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 (ws_3): for $qfirst + 1 < t < qlast$, individual i separated from j during $t+1$ with full-quarter status during 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 (Δws_1): for $t < qlast$, earnings change in period $t+1$ of individual i separated from 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 (Δws_3): for $t < qlast$, earnings change in period $t+1$ of individual i full-quarter separated from j during t, last full-quarter employment was $t-1$

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

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

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

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

Periods of non-employment prior to a new hire (nh): for $qfirst + 3 < t$, periods of non-employment during the previous four quarters by i prior to a new hire at establishment j during t

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

Periods of non-employment prior to a recall (nr): for $qfirst + 3 < t$, periods of non-employment during the previous four quarters by i prior to a recall at establishment j during t

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

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

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

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 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

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

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

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

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

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

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

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

Accessions   for establishment j during t

$$A_{jt} = a_{1 \cdot jt} \qquad\qquad (A.57)$$

Average accession rate for establishment j during t

$$AR_{jt} = A_{jt} \big/ \bar{E}_{jt} \qquad\qquad (A.58)$$

Separations for establishment j during t

$$S_{jt} = s_{1 \cdot jt} \qquad\qquad (A.59)$$

Average separation rate for establishment j during t

$$SR_{jt} = S_{jt} \big/ \bar{E}_{jt} \qquad\qquad (A.60)$$

New hires for establishment j during t

$$H_{jt} = h_{1 \cdot jt} \qquad\qquad (A.61)$$

Full-quarter new hires for establishment j during t

$$H_{3jt} = h_{3 \cdot jt} \qquad\qquad (A.62)$$

Recalls for establishment j during t

$$R_{jt} = r_{1 \cdot jt} \qquad\qquad (A.63)$$

Flow into full-quarter employment for establishment j during t

$$FA_{jt} = a_{3 \cdot jt} \qquad\qquad (A.64)$$

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

$$FH_{jt} = h_{3 \cdot jt} \qquad\qquad (A.65)$$

114

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

$$WS_{jt} = ws_{1 \cdot jt} \qquad \text{(A.83)}$$

Total change in total earnings for separations

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

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

$$WFS_{jt} = ws_{3 \cdot jt} \qquad \text{(A.85)}$$

Total change in total earnings for full-quarter separations

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

Total periods of non-employment for separations

$$NS_{jt} = ns_{\cdot jt} \qquad \text{(A.87)}$$

Average earnings of end-of-period employees

$$ZW_{2jt} = W_{2jt} \,/\, E_{jt} \qquad \text{(A.88)}$$

Average earnings of full-quarter employees

$$ZW_{3jt} = W_{3jt} \,/\, F_{jt} \qquad \text{(A.89)}$$

Average earnings of accessions

$$ZWA_{jt} = WA_{jt} \,/\, A_{jt} \qquad \text{(A.90)}$$

Average change in total earnings for accessions

$$Z\Delta WA_{jt} = \Delta WA_{jt} \,/\, A_{jt} \qquad \text{(A.91)}$$

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)

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

118

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   Identities

The identities stated below hold at the establishment level for every age group and sex subcategory. These identities are preserved in the QWI processing.

Definition 1 Employment at beginning of period t equals end of period $t-1$

$$B_{jt} = E_{jt-1}$$

Definition 2 Evolution of end of period employment

$$E_{jt} = B_{jt} + A_{jt} - S_{jt}$$

Definition 3 Evolution of average employment

$$\bar{E}_{jt} = B_{jt} + (A_{jt} - S_{jt})/2$$

Definition 4 Job flow identity

$$JF_{jt} = JC_{jt} - JD_{jt}$$

Definition 5 Creation-destruction identity

$$E_{jt} = B_{jt} + JC_{jt} - JD_{jt}$$

Definition 6 Creation-destruction/accession-separation identity

$$A_{jt} - S_{jt} = JC_{jt} - JD_{jt}$$

Definition 7 Evolution of full-quarter employment

$$F_{jt} = F_{jt-1} + FA_{jt} - FS_{jt}$$

Definition 8 Full-quarter creation-destruction identity

$$F_{jt} = F_{jt-1} + FJC_{jt} - FJD_{jt}$$

Definition 9 Full-quarter job flow identity

$$FJF_{jt} = FJC_{jt} - FJD_{jt}$$

Definition 10 Full-quarter creation-destruction/accession-separation identity

$$FA_{jt} - FS_{jt} = FJC_{jt} - FJD_{jt}$$

Definition 11 Employment growth rate identity

$$G_{jt} = JCR_{jt} - JDR_{jt}$$

Definition 12 Creation-destruction/accession-separation rate identity

$$JCR_{jt} - JDR_{jt} = AR_{jt} - SR_{jt}$$

Definition 13 Full quarter employment growth rate identity

$$FG_{jt} = FJCR_{jt} - FJDR_{jt}$$

Definition 14 Full quarter creation-destruction/accession-separation rate identity

$$FJCR_{jt} - FJDR_{jt} = FAR_{jt} - FSR_{jt}$$

Definition 15 Total payroll identity

$$W_{1jt} = W_{2jt} + WS_{jt}$$

Definition 16 Payroll identity for consecutive-quarter employees

$$W_{2jt} = W_{1jt} - WCA_{jt} - WS_{jt}$$

Definition 17 Full-quarter payroll identity

$$W_{3jt} = W_{2jt} - WCA_{jt}$$

Definition 18 New hires/recalls identity

$$A_{jt} = H_{jt} + R_{jt}$$

Definition 19 Periods of non-employment identity

$$NA_{jt} = NH_{jt} + NR_{jt}$$

Definition 20 Worker-jobs in period t are the sum of accessions and beginning of period employment.

$$M_{jt} = A_{jt} + B_{jt}$$

Definition 21 Worker-jobs in period t are the sum of accessions to consecutive quarter status, separations, and full quarter workers.

$$M_{jt} = CA_{jt} + S_{jt} + F_{jt}$$

Definition 22 Consecutive quarter accessions in period $t-1$ are the sum of consecutive quarter separations in period t and full quarter accessions in period t

$$CA_{jt-1} - CS_{jt} = FA_{jt} - FS_{jt}$$

A.2.5   Aggregation of job flows

The aggregation of job flows is performed using growth rates to facilitate confidentiality protection. The rate of growth JF for establishment j during period t is estimated by:

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

For an arbitrary aggregate $k = ($ $ownership\times$ $state\times$ $substate\text{-}geography\times$ $industry\times$ $age\ group \times sex)$ cell, we have:

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

where the function $K(j)$ indicates the classification associated with firm j. We calculate the aggregate net job flow as

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

Substitution yields

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

so the aggregate job flow, as computed, is equivalent to the aggregate growth rate times aggregate employment. Gross job creation/destruction aggregates are formed from the job creation and destruction rates by analogous formulas substituting JC or JD, as appropriate, for JF (Davis et al.; 1996, p. 189 for details).

A.2.6   Measurement of employment churning

The QWI measure employment churning (also called turnover) using the ratio formula:

$$FT_{kt} = \frac{(FA_{kt} + FS_{kt})/2}{F_{kt}} \tag{A.108}$$

for an arbitrary aggregate $k = ($ $ownership\times$ $state\times$ $substate\text{-}geography\times$ $industry\times$ $age\ group \times sex)$ cell. In the actual production of the QWI, the three components of this ratio are computed as separate estimates and are released.

This version: $Id: qwi-overview.tex 442 2008-01-16 20:30:55Z vilhu001 $