# The LEHD Infrastructure Files

# *and the Creation of the Quarterly Workforce Indicators*

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

♠ Cornell University

♣ U.S. Census Bureau, LEHD Program

Introduction

- ➲ What are QWI?
- ➲ What is it?
- ➲ In this paper

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

# Introduction

# What are QWI?

Introduction
➲ What are QWI?
➲ What is it?
➲ In this paper
Input Files
Infrastructure Files
Forming Aggregated Estimates: QWI
Disclosure-proofing the QWI
Publicly available files
Conclusion

- ▲ Since 2003: publication of Quarterly Workforce Indicators
- ▲ The first 21st century statistical system
  - No additional burden
  - Extensive use of modern statistics to integrate and improve the data
  - State-of-the-art confidentiality protection methods
  - Innovative use of wage records to constitute a frame to integrate data
  - The first statistical system to use “jobs” as a frame

# What is it?

- Combines
  - (state) administrative records data on workers (UI Wage records)
  - (state) administrative records data on firms (QCEW aka ES-202)
  - administrative information on demographics
  - surveys on people and firms collected by Census Bureau
- careful longitudinal edit of person identifiers and economic firm units
- careful longitudinal edit of person and firm characteristics

# In this paper

Introduction
➲ What are QWI?
➲ What is it?
➲ In this paper

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

- ▲ Describe the construction of the LEHD infrastructure
  - ... in particular the imputation mechanisms used
- ▲ Describe the computation of the QWI statistics
  - ... in particular the imputation mechanisms used
- ▲ Describe the disclosure-proofing mechanism
- ▲ Describe researcher access to infrastructure files and confidential QWI files

Introduction

Input Files

- ➲ Wage records: UI
- ➲ Employer reports: ES202
- ➲ Demographics

Infrastructure Files

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

# Input Files

# Wage records: UI

Introduction

Input Files
➲ Wage records: UI
➲ Employer reports: ES202
➲ Demographics

Infrastructure Files

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

- report of an individual’s UI-covered earnings by an employing entity
- appears if at least one dollar was earned by that individual during the quarter
- identifies EARNINGS, EMPLOYER, TIME PERIOD
- some limited other state-dependent information available
- in particular, for Minnesota, the ESTABLISHMENT is reported

# Employer reports: ES202

- ▲ collected as part of the Covered Employment and Wages (CEW) (administered by the BLS)
- ▲ Also used as the inputs to the Business Employment Dynamics (BED)
- ▲ collects from employers covered by state unemployment insurance programs:
  - ● employment
  - ● payroll
  - ● geographic information
- ▲ fundamental unit: ’reporting unit’ (≈ establishment)
- ▲ One report per establishment per quarter is filed

# Demographics

- ▲ Demographics are taken from a number of Census-internal files derived from administrative data:
  - ● Person Characteristics File (PCF)
  - ● Census Numident
- ▲ Where available, more detailed data on individuals is also extracted from surveys and censuses:
  - ● CPS
  - ● SIPP
  - ● ACS
  - ● 1990 Census
  - ● 2000 Census

Introduction

Input Files

Infrastructure Files

- EHF: Employment History Files
- Individual Characteristics File: ICF
- The Employer Characteristics File: ECF
- The Geocoded Address List: GAL
- Flow so far

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

# Infrastructure Files

# EHF: Employment History Files

Introduction

Input Files

Infrastructure Files

➲ EHF: Employment History Files

➲ Individual Characteristics File: ICF

➲ The Employer Characteristics File: ECF

➲ The Geocoded Address List: GAL

➲ Flow so far

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

- ▲ Job-level EHF
  - complete in-state work history for each individual on UIwage records.
  - one record for each employee-employer combination – a job
  - earnings and employment patterns
- ▲ Employer and establishment-level employment history
  - QCEW-based employment-activity history for every SEIN (employer) and SEINUNIT (establishment)
- ▲ Comparison of employment and activity of SEINs between UI and QCEW files is done for QA purposes, and in preparation of weighting.

# Individual Characteristics File: ICF

- ▲ Demographic information from the PCF is merged with universe of PIKs from wage records
- ▲ records without a valid match flagged
- ▲ CPS and SIPP identifiers are merged on.
- ▲ ... gender, education, and age information from the CPS
- ▲ Data completion
  - ● Age
  - ● Gender
  - ● Education
  - ● County of residence

  are each imputed ten times

# The Employer Characteristics File: ECF

Introduction

Input Files

Infrastructure Files
➲ EHF: Employment History Files
➲ Individual Characteristics File: ICF
➲ The Employer Characteristics File: ECF
➲ The Geocoded Address List: GAL
➲ Flow so far

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

- ▲ Two files: firm and establishment level, quarterly records
- ▲ Inputs:
  1. ES202
  2. UI: supplement information on the ES202, extend published BLS county-level employment data
  3. GAL: establishment geocodes
  4. LDB (BLS) for backfilling NAICS information
- ▲ Longitudinal edits for consistency and data completion
- ▲ Imputation of
  - ● impute of SIC if NAICS non-missing and vice-versa
  - ● unconditional impute of missing SIC and NAICS codes
  - ● geography conditional on industry

# The Geocoded Address List: GAL

Introduction

Input Files

Infrastructure Files
➲ EHF: Employment History Files
➲ Individual Characteristics File: ICF
➲ The Employer Characteristics File: ECF
➲ The Geocoded Address List: GAL
➲ Flow so far

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

- ... is a data set containing unique commercial and residential addresses
- geocoded to the Census Block and latitude/longitude coordinates
- Inputs:
  1. ES202 data
  2. Census Bureau’s Business Register (BR)
  3. Census Bureau’s Master Address File (MAF)
  4. American Community Survey Place of Work file (ACS-POW)
- Addresses are
  1. geocoded
  2. standardized
  3. unduplicated (by firm name)

# Flow so far

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

- Correction of spurious worker flows
- Solution
- Attaching establishment characteristics to jobs
- Attaching establishment characteristics to jobs
- Probability Model
- Implementation
- Implementation
- Computing the statistics

Disclosure-proofing the QWI

Publicly available files

Conclusion

# Forming Aggregated Estimates: QWI

# Correction of spurious worker flows

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

➲ Correction of spurious worker flows
➲ Solution
➲ Attaching establishment characteristics to jobs
➲ Attaching establishment characteristics to jobs
➲ Probability Model
➲ Implementation
➲ Implementation
➲ Computing the statistics

Disclosure-proofing the QWI

Publicly available files

Conclusion

- ▲ Firm identifier:
- ▲ Account numbers can and do change:
  - ● change in legal form
  - ● a merger
- ▲ Change in <u>firm identifier</u>
- ▲ → non-economic change in identifier creates spurious flow

# Solution

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

➲ Correction of spurious worker flows

➲ Solution

➲ Attaching establishment characteristics to jobs

➲ Attaching establishment characteristics to jobs

➲ Probability Model

➲ Implementation

➲ Implementation

➲ Computing the statistics

Disclosure-proofing the QWI

Publicly available files

Conclusion

- ▲ track large worker movements between SEINs
- ▲ → link entities that have different account numbes, but constitute the same economic entitiy
- ▲ 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.
- ▲ QWI: 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.

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

➲ Correction of spurious worker flows

➲ Solution

➲ Attaching establishment characteristics to jobs

➲ Attaching establishment characteristics to jobs

➲ Probability Model

➲ Implementation

➲ Implementation

➲ Computing the statistics

Disclosure-proofing the QWI

Publicly available files

Conclusion

# Attaching establishment characteristics to jobs

- ▲ Goal: achieve a high level of accuracy and detail
- ▲ Problem:
- ▲ 30-40% of state-wide employment in multi-establishment firms
- ▲ Solution: probability model for employment location and imputation
- ▲ Key elements are:
  1. distance between place-of-work and place-of-residence
  2. distribution of employment across establishments of multi-establishment firms.
- ▲ Important practical aspects:
  - ● Non-ignorable missing data imputation
  - ● Several million imputations every quarter

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

➲ Correction of spurious worker flows

➲ Solution

➲ Attaching establishment characteristics to jobs

➲ Attaching establishment characteristics to jobs

➲ Probability Model

➲ Implementation

➲ Implementation

➲ Computing the statistics

Disclosure-proofing the QWI

Publicly available files

Conclusion

# Attaching establishment characteristics to jobs

- workers $i = 1, ..., I$
- firms $j = 1, ..., J$
- active establishments at firm $j$ $R_{jt}$
- quarter $t$ employment of establishment $r$ in firm $j$ $N_{jrt}$
- $y_{ijt}$ establishment at which $i$ was employed
- $\mathcal{J}_t$ firms active
- $\mathcal{I}_{jt}$ individuals employed at firm $j$
- $\mathcal{R}_{jt}$ set of active ($N_{jrt} > 0$) establishments
- $\mathcal{R}^i_{jt} \subset \mathcal{R}_{jt}$ set of active establishments that are feasible for worker $i$.
- Feasibility: an establishment $r \in \mathcal{R}^i_{jt}$ if $N_{jrs} > 0$ for every quarter $s$ that $i$ was employed at $j$.

# Probability Model

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

➲ Correction of spurious worker flows

➲ Solution

➲ Attaching establishment characteristics to jobs

➲ Attaching establishment characteristics to jobs

➲ Probability Model

➲ Implementation

➲ Implementation

➲ Computing the statistics

Disclosure-proofing the QWI

Publicly available files

Conclusion

$$p_{ijrt} = \Pr\left(y_{ijt} = r\right)$$

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

$\alpha_{jrt}$ establishment- and quarter-specific effect

$x_{ijrt}$ time-varying vector, worker and establishment

$\beta$ effect on probability of being employed at a particular establishment

Currently:

- $x_{ijrt}$ is linear spline in distance between residence and establishment
- $\alpha_{jrt}$ is a hierarchical Bayesian model based on $N_{jrt}$ is

# Implementation

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

➲ Correction of spurious worker flows

➲ Solution

➲ Attaching establishment characteristics to jobs

➲ Attaching establishment characteristics to jobs

➲ Probability Model

➲ Implementation

➲ Implementation

➲ Computing the statistics

Disclosure-proofing the QWI

Publicly available files

Conclusion

Using Minnesota data,

compute posterior modal value of $\alpha_{jrt}$

evaluate the posterior mode of $p(\beta|\alpha, x, y)$

maximize

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

(2)

# Implementation

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI
➲ Correction of spurious worker flows
➲ Solution
➲ Attaching establishment characteristics to jobs
➲ Attaching establishment characteristics to jobs
➲ Probability Model
➲ Implementation
➲ Implementation
➲ Computing the statistics

Disclosure-proofing the QWI

Publicly available files

Conclusion

- ▲ use mean and variance of $\beta$ from Minnesota data
- ▲ take 10 draws of $\beta$ from the normal approximation (at the mode) to $p\left(\beta|\alpha, x, y\right)$.
- ▲ use QCEW employment counts, compute 10 values of $\alpha_{jt}$
- ▲ The drawn values of $\alpha$ and $\beta$ are used to draw 10 imputed values of place of work from to the posterior predictive distribution

$$(3)\; p\left(\tilde{y}|x, y\right) = \int\int p\left(\tilde{y}|\alpha, \beta, x, y\right) p\left(\alpha|N\right) p\left(\beta|\alpha, x, y\right) d\alpha d\beta$$

- ▲ $\rightarrow$ 10 establishment identifiers associated with a job spell

# Computing the statistics

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

➲ Correction of spurious worker flows

➲ Solution

➲ Attaching establishment characteristics to jobs

➲ Attaching establishment characteristics to jobs

➲ Probability Model

➲ Implementation

➲ Implementation

➲ Computing the statistics

Disclosure-proofing the QWI

Publicly available files

Conclusion

- We now have:
  - Jobs identified
  - Jobholder’s demographics
  - Establishment’s characteristics
- Now compute
  1. For each job, the relevant variables, defined at the person-level (indicators)
  2. Aggregate (typically sum) to the establishment level
  3. → establishment-level statistics, available in RDC
  4. Attach weights to each establishment
  5. Attach ’fuzz’ factors to each establishment
  6. Final aggregation to desired geography-industry-demographic detail
- Disclosure-proof

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

➲ Noise-infusion

➲ Cell suppression

Publicly available files

Conclusion

# Disclosure-proofing the QWI

# Noise-infusion

- First layer: workplace-level aggregation
  - infusion of specially constructed noise:
  -

$$(4)\ p\left(\delta_j\right) = \begin{cases} (b-\delta)\big/(b-a)^2, & \delta \in \ [a,b] \\ (b+\delta-2)\big/(b-a)^2, & \delta \in [2-b, 2-a] \end{cases}$$

  - Result: random noise factor centered around 1 with distortion of at least $a-1$ and at most $b-1$.
- Important properties:
  1. for a given workplace, distortion is always distorted in the same direction (increased or decreased) by the same percentage amount in every period.
  2. when estimates are aggregated, the effects of the distortion cancel out for the vast majority of the estimates.

# Cell suppression

Introduction

Input Files

Infrastructure Files

Forming Aggregated
Estimates: QWI

Disclosure-proofing the QWI
➲ Noise-infusion
➲ Cell suppression

Publicly available files

Conclusion

▲ Second layer: after aggregations
- Some estimates are based on fewer than three persons or firms.
- → suppression of these estimates
- Some of the estimates are based on noisy data
- → flagged as “substantially distorted”

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

# Publicly available files

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

➲ Flow so far

# Conclusion

# Flow so far

Introduction

Input Files

Infrastructure Files

Forming Aggregated Estimates: QWI

Disclosure-proofing the QWI

Publicly available files

Conclusion

➲ Flow so far