Detailed Data Steps for "Financial Reporting Quality, Turnover Risk, and Wage Differentials: Evidence from Worker-level Data"

Jung Ho Choi
Brandon Gipper
Stanford University Graduate School of Business

Sara Malik
University of Utah David Eccles School of Business

February 2023

This document summarizes, in words, the preparation of data sets used in our analyses. Our paper relies on confidential microdata from the U.S. Census Bureau, which also includes federal tax information provided by the IRS. Access to this data is restricted to U.S. based researchers with approved research projects. Any disclosure about the content of the data to non-approved individuals constitutes is not allowed. In light of this restriction, we describe our data processing steps.[1]

1. Data Sets

We utilized the following data sets managed by the U.S. Census Bureau:

- Longitudinal Employer-Household Dynamics (LEHD) (2014 version) which includes
  - Employer Characteristics Files (ECF, establishment-level data)
  - Employment History Files (EHF, establishment-employee matches and wages)
  - Individual Characteristics Files (ICF, individual-level data including gender and date of birth)
  - Quarterly Workforce Indicators (QWI, establishment level data on hiring and separations)
- Compustat-SSEL Bridge (CSB, a direct link between Compustat and LEHD identifiers)
- Standard Statistical Establishment List/Business Register (SSEL)

We also utilized the following data sets available for purchase: Audit Analytics SEC 404, Compustat Fundamentals Annual, CRSP Stock Security Files, the Panel Study of Income Dynamics (PSID), Bloomberg, and Refinitiv ESG. Audit Analytics SEC 404, Compustat Fundamentals Annual, and CRSP Stock Security Files are downloaded originally in October 2017 and updated in April 2019. The PSID data were collected and processed in May 2020. The Bloomberg and Refinitiv ESG data were accessed in March 2022.

We process the LEHD data following the instruction included in CES 18-27: LEHD Infrastructure S2014 files in the FSRDC. We highly encourage all researchers to familiarize themselves with this document, as it is the U.S. Census Bureau standard for LEHD processing. CES 18-27 is very dense and can be overwhelming for new researchers. Thus, we view our memo not just as a description

[1] Should a reader gain approved access to the datasets used in our analysis, we will be happy to discuss our methodology in detail. Code cannot be transferred between project drives unless the author of the code is on both projects. Thus, we would be unable to share our code even with approved researchers.

of our processing but also as a guide to CES 18-27. For example, we use more intuitive names for variables (as opposed to the official variable names) for ease of understanding. For example, the employee ID in LEHD is PIK, but we refer to it as "employee."

We used the LEHD data to generate two datasets that are subsequently used in regressions. The first dataset's observations represent an employee-firm-state-year tuple. The second dataset's observations represent a firm-state-year tuple. The LEHD data primarily uses an establishment (not a firm) key.

2. Data Processing

Employee-Firm-State-Year Data Set

- The ICF key is employee, which is analogous to an employee's social security number. ICF includes time-invariant employee-level characteristics like sex, date of birth, and education. We process the ICF data to rename variables, check for missing values, and generate dummy variables.
    - Researchers are encouraged to select a random sample of individuals when generating employer-employee matched data. The leading two digits of the employee ID variable represent a uniform distribution over 100 buckets. Following Census recommendations, we sampled 1% of employees.
- The ECF includes establishment characteristics including industry and location. Most importantly, the ECF has both (internal) establishment and firm identifiers. We use the ECF to generate an establishment-state-year-firm map.
- The EHF key is establishment-employee-state-year. Relevant variables include wage. We merge the ICF data onto EHF using the employee key. Note that this reduces the size of the EHF data set substantially, which we expected as we only sample 1% of workers. We then merge on the firm variable using the establishment-state-year-firm map.
- The resulting dataset's key is employee-firm-state-year but is not yet limited to Compustat firms. Using internal Census resources, we create a map from firms to gvkey (through CUSIP) to limit our analysis to public firms.
    - We use three steps of merging Compustat to the LEHD data. The first step is to use the Compustat-SSEL Bridge (CSB). The CSB has the map between CUSIP and Census Bureau (internal) firm identifiers in every year. The first match is constructed based on CUSIP-Internal Firm Identifier-Year.
    - Any unmatched observations in the first procedure are subject to the second method: Employer Identification Number (EIN). The Compustat has information about firms' EINs. The Standard Statistical Establishment List/Business Register (SSEL) has an establishment's EIN and (internal) firm identifiers. Using EIN, we add additional matches between gvkey and a Census Bureau internal firm identifier to the merged dataset.
    - Finally, we merge observations in the SSEL and Compustat if those observations have the same name, industry, and location in both datasets while deleting any duplicates.
- Data Restrictions
    - Employee requirements: We require that workers are between 20 and 60 years old; this requirement generally limits the sample to workers who are (or desire to be)

full-time participants in the workforce. We also require that the worker's annual real wages are higher than $10,000 to exclude temporary workers.
    - Firm requirements: We limit our sample to US publicly listed firms that report in USD currency. We require a unified sample, meaning that we only include those observations that have all non-missing values in the requisite variables (i.e., size, Tobin's Q, leverage, return on asset, tangible assets, M&A activity, and returns volatility).
- Connected Sets
    - In order to correctly estimate worker and firm fixed effects, we limit our data set to a connected set (Abowd et al., 2002). A connected set is the set of all workers who ever worked for any of the firms in the data set, as well as all firms that ever employed any of the workers in the data set. The connected set is time invariant. A single large data set can be composed of multiple, smaller disjoint connected sets. We use the largest connected set in the Employee-Firm-State-Year Data Set.
    - To identify the connected set, we use a union-find algorithm, a standard workhorse computer science algorithm. We refer readers to the computer science literature for further explanation and example code.

Firm-State-Year Data Set

- The QWI key is establishment-state-year and includes multiple measures of hiring, separation, and turnover. We keep the measures specific to full-time employees.
    - We calculate turnover rates for each establishment. We observe full-time workers' joins and separations, as well as their employment at the beginning and end of each year.
    - Turnover rates are winsorized at the top at one (100%). Observations with negative turnover rates are removed.
- In this case, we do not limit the sample because the data are aggregated to establishment-state-year and do not require as much processing power as employer-employee matched data. We merge on the firm variable using the establishment-state-year-firm map. We perform the same exercise as above to limit the sample to Compustat firms.

Structural Approach of Financial Reporting Quality

- We closely follow Nikolaev (2019). We use cash-flow based accruals, earnings, and earnings measures. We require firms to be formally listed in US exchanges and to have at least 7 observations in our sample period. We winsorize all the variables at top and bottom 1%.
- In the estimation models, we use firm accrual and cash flow levels, within firm accrual and cash flow changes, and firm accrual and cash flow levels adjusted for firm characteristics (i.e., changes in size, non-current assets, and market-to-book ratio).

Panel Study of Income Dynamics (PSID)

- The PSID is maintained by the University of Michigan. We were approved for access by the Institute for Social Research.
- We downloaded PSID data spanning 1976 to 2013. We used the provided code books to interpret the variables. The first code book covers 1976 to 2001. The second codebook covers 2003 to 2013.

- Whenever possible, we process the PSID using the methodology in Lemieux, MacLeod, and Parent (2009). Below, we detail our important methodology choices and the associated assumptions.
    - Starting from 1999, the survey is conducted every other year. Lemieux et al.'s (2009) method is valid only if we assume that between two surveys, there is no gap between the current job and the last job if the individual is employed by the time of the survey or there is no other job between the job that was recorded from the last survey and the most recent job if the individual is unemployed by the time of the survey. In other words, the job variable we created is saying that the respondent at least changed his job once. If he changed his job more than once, we would not be able to capture the other job.
    - Following Lemieux et al. (2009), we drop those who are non-sample observations. We drop those who were not head of household and only keep those who were head of household for at least three years. We use the education variable from individual data for all years.

Miscellaneous Data Sets

- In Tables 8 and 9, we use variables from other data sets to augment our regression specifications. These other data sets do not require significant processing.
- Refinitiv ESG Performance Data
    - We access Refinitiv ESG performance data directly from Refinitiv Workplace. Refinitiv provides a summary ESG performance score by firm year starting in 2002. We use ticker to merge Refinitiv ESG performance data with Employee-Firm-State-Year Data Set. We require our observations to have ticker in both data sets. We require our observations to have non-missing relevant ESG score.
- Bloomberg ESG Score Data
    - We access Bloomberg ESG scores using a Bloomberg terminal. The Bloomberg scores measure ESG disclosure quality by firm year starting in 2008. We use CIK to merge Bloomberg ESG score data with Employee-Firm-State-Year Data Set. We then merge unmatched observations using CUSIP. We require our observations to have either CIK or CUSIP in both data sets. We require our observations to have non-missing relevant ESG scores.
- Former Arthur Andersen Clients Data Sets / Internal Control Weaknesses Data Sets
    - We use Audit Analytics Auditor Opinion and SOX 404 Internal Controls data to identify firms that used Arthur Andersen in the fiscal year ending 2001. We use CIK for merging. We also use Audit Analytics data to identify firms that announced Internal Control Weaknesses between 2004 and 2012. We require internal control weakness firms to issue the weakness report only once and resolve it in a subsequent year; this is to avoid "double" treatment. We also require that the firm have data between 2004 and 2012.

3. Data Sets for Tables

Table and Figure	Data Sets
Table 1	Employee-Firm-State-Year Data Set
Table 2	Employee-Firm-State-Year Data Set
Table 3	Employee-Firm-State-Year Data Set
Table 4	Employee-Firm-State-Year Data Set
Table 5	Employee-Firm-State-Year Data Set; PSID
Table 6	Firm-State-Year Data Set
Table 7	Employee-Firm-State-Year Data Set
Table 8	Employee-Firm-State-Year Data Set; Firm-State-Year Data Set; Refinitiv ESG Performance Data; Bloomberg ESG Disclosure Data
Table 9	Employee-Firm-State-Year Data Set; Arthur Andersen Client List; Internal Control Weakness Data
Figure 1	Employee-Firm-State-Year Data Set
Figure 2	Employee-Firm-State-Year Data Set