Economics Letters 152 (2017) 83–87

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Validating RefUSA micro-data with the Longitudinal Employer-Household Dynamics Data

Christos A. Makridis*, Michael Ohlrogge

Stanford University, United States

HIGHLIGHTS

- Compare micro-moments between RefUSA and both the LEHD and CBP.
- Means, standard deviations, and growth rates differ dramatically between the two.
- Conditional correlations on the employment effects of corporate tax rates differ.

ARTICLE INFO

Article history:
Received 29 September 2016
Received in revised form
25 November 2016
Accepted 1 January 2017
Available online 3 January 2017

JEL classification:
H25
J30

Keywords:
Micro-data
Employment
Longitudinal Employer-Household Dynamics
County Business Patterns
RefUSA

ABSTRACT

This paper validates the reliability of employment data in a frequently used establishment panel database assembled by InfoGroup by comparing it with employment and establishment data from the publicly available Longitudinal Employer-Household Dynamics (LEHD) and County Business Patterns (CBP) at the three-digit industry-by-state-by-year and two-digit industry-by-county-by-year levels between 1997 and 2013. We document substantial differences in both their cross-sectional and time series properties. Through an application involving the evaluation of the employment effects of state corporate tax rates, we also illustrate that the inclusion of fixed effects does not eliminate the bias associated with the extrapolation and/or other measurement error. These results suggest that both descriptive evidence and causal inference from the RefUSA data are unreliable.

1. Introduction

The emergence of readily available micro-data has accelerated the quality of research and the breadth of questions. One source of information that has been used in recent years is InfoGroup's RefUSA database, which, according to their marketing materials, contains information on approximately 24 million US private establishments over time. It has been licensed to many research universities and applied in various areas of social sciences, ranging from public finance (Suarez Serrato and Zidar, 2016) to industrial organization (McDevitt, 2014) to economic geography (Dai and Jaworski, 2016) to public health (D'Angelo et al., 2016). However, RefUSA relies heavily on extrapolations in constructing their data and, to our knowledge, there has been no external validation of RefUSA's reliability. To give a sense of the impact of this extrapolation, the mean annual employment in RefUSA has a −0.50 correlation with actual employment from the LEHD and a growth rate near zero between 2000 and 2008.

We address this shortcoming in the literature through two comparisons: (i) employment records at the three-digit NAICS industry-by-state-by-year level with the Longitudinal Employer-Household Dynamics (LEHD) publicly accessible data, and (ii) establishment counts at the two-digit NAICS industry-by-county-by-year level with the County Business Patterns (CBP). The LEHD is collected by state unemployment agencies and, since their obligations depend upon the data, they invest significant resources to ensure that their data remains a reliable source of information.[1] The CBP, authorized under Titles 13 and 26 of the US code, is equally reliable since it draws from the Business Register, a

* Correspondence to: Stanford University, Department of Economics, and Department of Management Science, Huang Engineering Center, 475 Via Ortega, Stanford, CA 94305-4121, United States.

E-mail address: cmakridi@stanford.edu (C.A. Makridis).

http://dx.doi.org/10.1016/j.econlet.2017.01.001
0165-1765/© 2017 Elsevier B.V. All rights reserved.

[1] See Abowd et al. (2009) for detailed documentation of construction about the LEHD.

relational database that links administrative, Census, and survey data. CBP records are maintained based on the best available information from the Internal Revenue Service (IRS) employer identification numbers (EINs).

We document that, while there are positive correlations for employment and establishment counts between RefUSA and LEHD & CBP, they have quantitatively important differences. We also provide a simple illustration of the potential biases that may emerge in causal inference by examining the conditional correlations on the effects of state corporate tax rates on employment.

2. Data description

Establishment panel from RefUSA.—InfoGroup produces information on a panel of establishments in the RefUSA database. We use the subset of it that spans between 1997 and 2013. InfoGroup begins by generating an estimate of the number of employees a firm has based on, for instance, the square footage of their office space, as available from public property records data. InfoGroup also collects data from the Department of Commerce on sales per employee for each four-digit SIC and NAICS code. By multiplying this number by the number of employees at each location, they produce a measure of establishment sales. InfoGroup also conducts periodic phone surveys of small samples of firms in their database, asking them for specific figures on employees and other metrics. RefUSA then updates its estimation models accordingly. Although it is not clear to us how the imputation algorithm works, they incorporate information from the parent company and compare it with data from similar locations within the same geography and industry classification. Their marketing materials suggest that the employment model is 95% accurate within two employee size ranges at 100% accurate within three employee size ranges where ranges are between 1–4, 5–10, 10–19, 20–49, 50–99, 100–249, 250–499, 500–999, 1000–4999, 5000–9999, and 10,000+. Because of these imputation techniques, a large majority of firms exhibit zero change in employment and sales from year-to-year in RefUSA.

Industry-by-state panel from LEHD.—We compare RefUSA with data over the same time period from the Longitudinal Employer-Household Dynamics (LEHD) dataset, specifically the Quarterly Workforce Indicators (QWI), which is publicly accessible at an aggregated level from the Census Bureau website (http://lehd.ces.census.gov/data/).[2] The LEHD covers over 95% of jobs in the US and consists of a unique federal-state data sharing collaboration called the Local Employment Dynamics (LED) partnership. It is a partnership whereby all state agencies voluntarily submit quarterly data files from existing administrative records. These administrative records combine information from employers' quarterly earnings reports that are required for state unemployment insurance agencies, the Quarterly Census of Employment and Wages, the Business Dynamics Statistics, and other demographic sources from the Census Bureau and Social Security Administration.

Industry-by-county panel from CBP.—We also compare RefUSA with data from the County Business Patterns (CBP), which is publicly accessible from the Census Bureau website (http://www.census.gov/programs-surveys/cbp.html). The CBP covers businesses with paid employees throughout the whole US, Puerto Rico, and Island Areas at a detailed geographic and industry level and covers most NAICS industries (with the exception of crop and animal production, rail transportation, the National Postal Service, pension, health, welfare, and vacation funds, trusts, estates, and agency accounts, private households, and public administration). We also note that the CBP does not cover establishments with government employees. The data is extracted from the Business Register, which contains the most complete, current, and consistent data for business establishments consolidated by the IRS, Census Bureau, and surveys (e.g., Annual Survey of Manufacturers).

3. Comparing RefUSA and LEHD

In our validation exercises, we focus on the total number of employees and establishments at a three-digit industry-by-state-by-year and two-digit industry-by-county level, respectively, i.e., $E_{ilt} = \sum_{j\in\mathcal{J}} E_{jilt}$ where $\mathcal{J} = \{1, 2, \ldots, J\}$ denotes the index on establishments in an industry (i), location (l), and year (t).[3] Our fundamental question is whether RefUSA consistently represents the same underlying facts about employment and establishments that appear in administrative data. If so, then even if there are certain differences in methodology and/or coverage, we should still expect to see a variety of stable and sensible trends between the data.

3.1. Descriptive differences

We begin by presenting evidence from a static perspective that pools all years together. The first column in Fig. 1 begins by plotting the quantiles between the two datasets. That is, if the total number of employees in a given industry-by-state-by-year is in the ρth quantile of the distribution in RefUSA, is it also near the ρth quantile in the LEHD? Unfortunately, the wedge between the quantiles in each dataset is increasing in the number of employees in a given industry and/or location. This monotonic relationship is likely to carry over into the establishment-level, meaning that quantiles in the RefUSA are less likely to match quantiles in the LEHD as the position in the firm size distribution grows. The divergence is significant for causal inference in light of the fact that firm size is linked with productivity (e.g., Lucas, 1978).

The second column in Fig. 1 examines the similarities in a related way by simply plotting the logged total employment in each industry-by-year combination in the RefUSA with the corresponding industry-by-year in the LEHD. We subsequently fit a regression line through these points.[4] A regression of logged employment in the LEHD on logged employment in RefUSA produces a coefficient of 0.54 and an R-squared of 0.502. Since the regression coefficient and R-squared provide two measures of closeness of fit between the datasets, we interpret this as evidence that, on average, the RefUSA micro-data is explaining at most half of the variation in actual employment at an industry-level.

Fig. 2 subsequently plots the distribution of logged employment and the change in logged employment between the two datasets.

[2] While their micro-data requires working in one of the Census Bureau's restricted data research centers after having a research proposal approved, their public extraction tool provides enormous value to researchers who can instead work with detailed industry-by-geography levels of aggregation. The QWI provides information on employment, hires, separations, and earnings at a quarterly frequency starting from 1990. One limitation is that not all states were initially covered by the data. However, all states are now part of the database, and the bulk of them were in by the late 1990s.

[3] We were concerned about implementing an accurate comparison, which is why we use the same measure between the two datasets. We initially experimented with using the mean number of employees across establishments within a state and industry and year since we were concerned that sampling variability in RefUSA. The rationale behind this approach involves the assumption that level differences between the mean and total would be absorbed into an intercept term or washed out when working in first-differences. While both methods produced similar divergences between the datasets, we choose to present the results in the former format because of the comparability and ease of interpretation.

[4] We collapsed to the industry-by-year level since there plot is not as visually helpful with the full sample (including state heterogeneity).