Delaware Valley Regional Planning Commission

Employment Data Evaluation

Addison Larson

September 26, 2018

1 Overview

This paper evaluates differences in the 2013 vintage of two employment data sources available for purchase—the National Establishment Time-Series (NETS) and InfoUSA—for Conshohocken, Montgomery County, PA. Specifically, these two datasets are evaluated based on summations of employment over census blocks, block groups, and tracts; the overall composition of establishments by number employed; establishment-level differences in employment; differences among large employers; and overall data quality. See Table 1, *General Properties of NETS and InfoUSA Data for Conshohocken*, for a brief overview of the two datasets.

The Delaware Valley Regional Planning Commission (DVRPC) purchases NETS data when it becomes available. This occurred most recently in 2013. However, NETS data requires extensive manual cleanup for DVRPC's nine-county region. If InfoUSA data requires substantially less correction, then it may be worth the additional cost to purchase it instead of NETS. This paper assumes NETS data is already clean and consistent.

There are critical differences between NETS and InfoUSA data sources, as well as between these sources and LEHD (Longitudinal Employer-Household Dynamics) data, which is used for comparative purposes when aggregating employment across Census geographies.[1]

1. **NETS** is... (***HALP!*** I plan to fill in more on this section once I hear answers to the question: what do you want people to know about these data sources?)
2. **InfoUSA** is...
3. **LEHD** is a Census Bureau program providing free and publicly-available data on workers, workplace characteristics, and commuting flows at the census block level. LEHD data is

[1] For more on employment data sources, see NCHRP 08-36, Task 127, *Employment Data for Planning: A Resource Guide*. Retrieved from http://onlinepubs.trb.org/onlinepubs/nchrp/docs/NCHRP08-36(127)_EmployDataGuide.PDF.

synthesized from multiple imputation of tax and employment records.[2] Because of privacy concerns, some noise is introduced in the data. Even though LEHD data is derived from tax records, it is likely geocoded using the Census Geocoder, which is less spatially accurate than other geocoding options ***HALP!*** Is this true, or it is more accurately that something like Google can better handle idiosyncrasies in address quality?

Results. There is no clear winner from this analysis. None of the datasets gives an indication that it is more truthful than the others. This is especially true when the data is spatially aggregated to the census tract level, as all three datasets present separate qualms to the analyst:

1. **NETS.** Why should all 5,780 jobs be concentrated in a single census tract, when it is unlikely that the bounds of Conshohocken align perfectly with this tract? Why does this source report so many more jobs than the others?

2. **InfoUSA.** Why are 655 jobs present in a tract where the other data sources report only 0 and 22 jobs, respectively? Does this indicate widespread errors in geocoding?

3. **LEHD.** In the same vein, why are 178 jobs present in a tract where the other data sources report only 0 and 48 jobs, respectively?

Here are some basic findings:

1. NETS has more than twice the number of records and many more jobs than InfoUSA. Two possible scenarios include that NETS fails to clean its records when establishments close their doors for good, or that NETS manages to record establishments that InfoUSA misses.

2. InfoUSA has a higher percentage of larger employers. Even when NETS and InfoUSA records are matched one-to-one by name and address, InfoUSA systematically reports more employees for the establishment.

3. When NETS and InfoUSA are subset for establishments with only 50 or more employees, there are not many establishments shared among the datasets. An analysis of the records common to both NETS and InfoUSA reveals that InfoUSA's jobs counts may be more accurate. This is important, as large employers skew the employment totals of individual geographies and drive the overall employment totals for the region.

4. All datasets require cleaning, and InfoUSA is no exception.

2 Summations of Employment Counts by Geographic Level

When NETS and InfoUSA point-level employment data are aggregated to the block, block group, and census tract level, the counts differ greatly within each spatial unit. This section uses 2013 LEHD Workplace Area Characteristics as an employment comparison for NETS and InfoUSA.

[2] For more on the creation of LEHD, see Technical Paper TP-2006-01, *The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators.* Retrieved from https://lehd.ces.census.gov/doc/technical_paper/tp-2006-01.pdf.

Tract- and Block Group-Level Differences. Table 2, *Differences in Employment, Tract Level*, and Table 3, *Differences in Employment, Block Group Level*, show the employment tallies for NETS, InfoUSA, and LEHD data at the census tract and block group levels, respectively. It is suspicious that InfoUSA has 655 employees recorded for Block Group 420912041021 and Tract 42091204102, given that NETS and InfoUSA have nearly zero employees recorded for the same area. In addition, it is suspicious that LEHD has 178 employees recorded for Block Group 420912059062 and Tract 42091205906, when neither NETS nor InfoUSA have similar counts. The discrepancies may point to the need to manually correct both InfoUSA *and* LEHD data.

Block-Level Differences. Table 4, *Differences in Employment, Block Level*, includes the counts by block and three additional columns: NETS x IG, NETS x LEHD, and IG x LEHD. (IG stands for InfoGroup, the parent company of InfoUSA.) A value of "Yes" in any of these columns indicates absolute percentage difference[3] between employment counts in the datasets equal to or exceeding 100%. For example, if the sum of employment over a block were 64 and 20 for NETS and InfoUSA, respectively, the NETS x IG column would read "Yes," because there is a 104.7% absolute difference between 64 and 20. The results of the absolute percentage difference calculations indicate that InfoUSA and LEHD counts are the most similar at the block level, with 16 of 46 blocks differing over 100% in job counts.

Overall Count Differences. In general, NETS has little in common with either InfoUSA or LEHD. InfoUSA has the fewest job counts overall, at 4,870. Though the LEHD counts lie closer to InfoUSA than to LEHD for the borough of Conshohocken, there is no clear evidence which of these datasets is the most accurate.

3 Composition of Establishments by Data Source and Number Employed

For the borough of Conshohocken, NETS has 637 observations and InfoUSA has 304. Because of the discrepancy in the number of overall records, Figure 1, *Percentage of Establishments at Six Employment Levels*, shows the percentage of establishments at each employment level. The results indicate that InfoUSA has a higher percentage of large employers. NETS has more employees in Conshohocken partially because it has more records overall.

However, this information gives rise to the question: does InfoUSA systematically report more employees than NETS for the same business establishments?

4 One-to-One Comparison of InfoUSA Sample to NETS Data

For the random sample of 143 InfoUSA records, we manually matched the InfoUSA firms to their counterparts in the NETS dataset. We found 72 matches in total. Of these, there were:

- 44 instances where InfoUSA reported more employees than NETS for the same establishment;

[3] The absolute percentage difference is calculated as:

$$\frac{|v_1 - v_2|}{\frac{v_1+v_2}{2}} \times 100$$

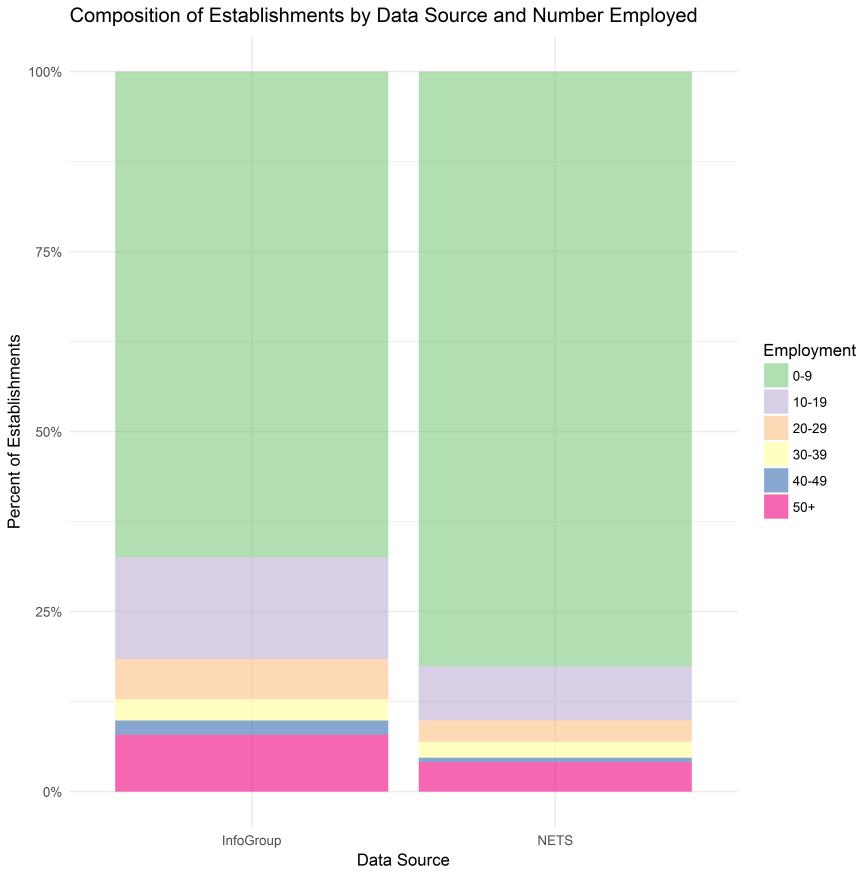

Figure 1: Percentage of Establishments at Six Employment Levels.

- 18 instances where NETS reported more employees than InfoUSA for the same establishment; and
- 10 instances where employment was the same in both NETS and InfoUSA.

The results indicate that, while InfoUSA has fewer overall records, it tends to report more employees than NETS for identical establishments. The average absolute percentage difference (see Footnote 3) between NETS and InfoUSA employment was 79.32%, and the median was 66.67%.

5 Large Employers

It is important to compare large employers in Conshohocken for at least three reasons: 1) They are the most visible employers in the area and therefore the easiest to verify; 2) They skew employment totals for individual geographic units, making it important that they are correctly geocoded; and 3) Small percentage differences between data sources for these establishments (i.e., between NETS and InfoUSA) can translate into large aggregate differences in employment for the area. Figure 2, *Distribution of Establishments by Employment and Source*, shows the NETS and InfoUSA records with over 50 employees. Accounting for the overall difference in the number of records between NETS and InfoUSA, NETS reports establishments with approximately 50 employees more often than InfoUSA. This may drive the overall difference in employment tallies between NETS and InfoUSA for the study area.

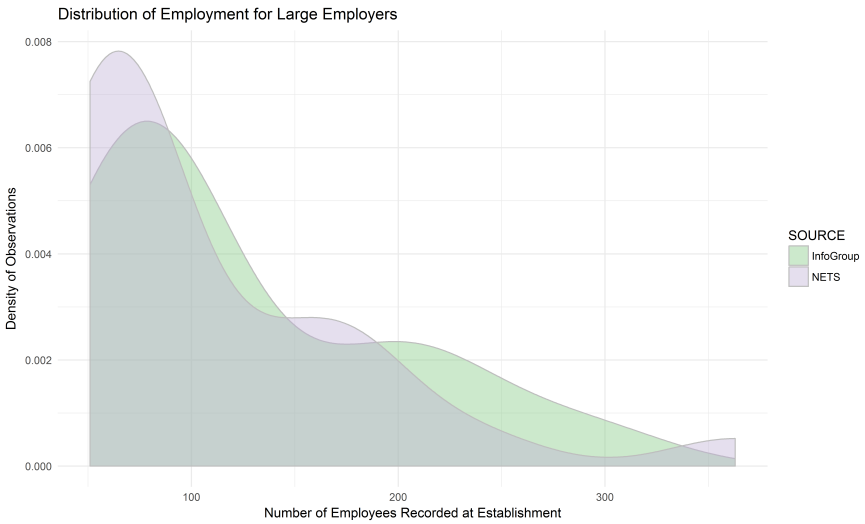

Figure 2: Distribution of Establishments by Employment and Source.

24 establishments in NETS and 19 in InfoUSA have a total reported number of employees exceeding 50. Of these, 7 large employers are present in both datasets. View Table 5, *Differences in Employment for Seven Large Employers* for employment counts by employer and data source. In four instances, InfoUSA reported more employees than NETS for the same establishment. However, it is worth highlighting two employers: Employer D, where NETS reported 113 more employees than InfoUSA; and Employer E, where InfoUSA reported 118 more employees than NETS. Employer D occupies only one floor of a modest-sized office building: it is likely that the InfoUSA record is closer to the true number of employed in this instance. Furthermore, an online search shows that Employer E self-reports as having over 250 employees at this particular location, pointing again to the InfoUSA record as the more accurate record. Further investigation of the veracity of large employer records might indicate which of these data sources is more reliable in general.

6 Overall Data Quality

This section, as with the rest of the paper, assumes NETS data is already clean and consistent. To evaluate the overall quality of InfoUSA data, we manually cleaned the random sample of 143 InfoUSA records for Conshohocken.

1. 2 of 143 (1.4%) InfoUSA establishments had falsely duplicated names. In this case, the InfoUSA sample had one completely misplaced duplicate. A record for a similarly named business located in Newtown Square—not Conshohocken—somehow made it into the Conshohocken dataset. This happened because the Newtown Square record *has the wrong latitude and longitude*. In addition, the business does not appear to exist in Newtown Square, though it does exist in Conshohocken.

2. 14 of 143 (9.8%) establishments required address formatting cleanup, e.g. changing "Cnshohckn" to "Conshohocken."

3. 4 of 143 (2.8%) of addresses were P.O. boxes.

4. 14 of 143 (9.8%) had the exact same address and suite number. In one instance, 12 establishments shared the same suite, which is suspicious.

7 Tables

	NETS	InfoUSA
Total Number of Establishments	637	304
No. Establishments > 0 Employees	631	286
No. Establishments > 50 Employees	24	19
Proposed Sample Size (90% CI, MOE ± 5%)	143	—

Table 1: General Properties of NETS and InfoUSA Data for Conshohocken.

Tract GEOID	NETS	InfoUSA	LEHD
42091204102	0	655	22
42091204200	5780	4167	4931
42091205906	0	48	178
TOTAL	5780	4870	5131

Table 2: Differences in Employment, Tract Level.

Block Group GEOID	NETS	InfoUSA	LEHD
420912041021	0	655	22
420912042001	3913	3031	3996
420912042002	1867	1136	935
420912059062	0	48	178
TOTAL	5780	4870	5131

Table 3: Differences in Employment, Block Group Level.

Block GEOID	NETS	InfoUSA	LEHD	NETS x IG	NETS x LEHD	IG x LEHD
420912041021003	0	655	22	Yes	Yes	Yes
420912042001006	72	0	0	Yes	Yes	No
420912042001008	430	1071	621	No	No	No
420912042001009	880	427	475	No	No	No
420912042001011	1056	551	417	No	No	No
420912042001012	858	24	0	Yes	Yes	Yes
420912042001014	33	34	161	No	Yes	Yes
420912042001015	1	0	0	Yes	Yes	No
420912042001016	41	44	55	No	No	No
420912042001018	110	24	33	Yes	Yes	No
420912042001019	2	0	0	Yes	Yes	No
420912042001021	60	57	101	No	No	No
420912042001022	44	44	0	No	Yes	Yes
420912042001034	63	176	2	No	Yes	Yes
420912042001035	93	493	835	Yes	Yes	No
420912042001037	4	5	1263	No	Yes	Yes
420912042001038	8	0	0	Yes	Yes	No
420912042001039	3	0	1	Yes	Yes	Yes
420912042001040	25	17	7	No	Yes	No
420912042001041	1	0	0	Yes	Yes	No
420912042001048	126	64	25	No	Yes	No
420912042001054	1	0	0	Yes	Yes	No
420912042001056	2	0	0	Yes	Yes	No
420912042002000	317	287	155	No	No	No
420912042002002	86	27	0	Yes	Yes	Yes
420912042002005	16	2	22	Yes	No	Yes
420912042002006	2	5	0	No	Yes	Yes
420912042002008	284	3	28	Yes	Yes	Yes
420912042002009	3	0	0	Yes	Yes	No
420912042002010	2	0	3	Yes	No	Yes
420912042002012	746	547	404	No	No	No
420912042002013	54	0	0	Yes	Yes	No
420912042002014	257	175	223	No	No	No
420912042002015	6	0	0	Yes	Yes	No
420912042002016	6	2	0	Yes	Yes	Yes
420912042002017	36	36	19	No	No	No
420912042002018	2	0	0	Yes	Yes	No
420912042002019	10	7	0	No	Yes	Yes
420912042002020	2	0	0	Yes	Yes	No
420912042002021	3	0	0	Yes	Yes	No
420912042002023	3	0	0	Yes	Yes	No
420912042002025	2	0	0	Yes	Yes	No
420912042002027	1	19	33	Yes	Yes	No
420912042002028	19	22	48	No	No	No
420912042002029	10	4	0	No	Yes	Yes
420912059062045	0	48	178	Yes	Yes	Yes
TOTAL	5780	4870	5131	27	34	16

Table 4: Differences in Employment, Block Level.

Employer	NETS	InfoUSA
A	56	75
B	56	64
C	171	165
D	363	250
E	182	300
F	138	110
G	60	75

Table 5: Differences in Employment for Seven Large Employers.