A Technical Note to Impute Workers' Industry Clusters Using Publicly Available Microdata

(Working draft, please do not cite)

Fernando A. Lozano[*]

Morris B. and Gladys P. Pendleton Professor of Economics, Pomona College
Research Fellow, Lowe Institute for Political Economy, Claremont McKenna College

June 2024

[*] This research is funded by a summer research grant from Claremont McKenna College's Lowe Institute for Political Economy. Gus Albach CMC'23 provided excellent research assistance. For correspondence, please e-mail fl004747@pomona.edu.

1. Introduction

Understanding industry clusters is essential to understand regional economic development. Clusters are composed of the firms and workers who interact on industries linked by their output, by their inputs, or by their workers' skills. Identifying clusters in economic data is important as clusters play an important role on the productivity of workers, firms, and regions (Moretti, 2014). Clusters have long been present in discussions of local economic development, and they were identified as early as 1890 by Alfred Marshall's in his Principles of Economics book. In fact, clusters increase productivity by three mechanisms: first, by creating thick labor markets with many workers and many employers. Second, by offering firms and workers a common set of suppliers and shared infrastructure. Third, and perhaps most importantly, by creating knowledge spillovers across workers within the cluster (Marshall, 1920). Thus, it is essential to empirically define clusters, and create a map of clusters that can be used by researchers and policy makers.

Unfortunately, consistent definitions of clusters in the economy do not exist, and most analysis of clusters has relied on regional case studies or at the aggregate industry level. To overcome this shortcoming, Delgado *et al*. (2014) develop a methodology that links industries in the six-digit North American Industry Classifications System (NAICS) with different clusters. Six-digit NAICS industry definitions is the most granular level in which researchers can identify industries. In fact, because of its greater detail, the authors show that there are substantial gains in measurements when using six-digit NAICS codes, relative to using broader industry categories. Unfortunately due to confidentiality concerns, public use data, such as the American Community Survey (ACS) or the Current Population Survey (CPS) available from the integrated Micro Public Use Data (IPUMS), do not provide industry codes at the six-digit level. For smaller industries, IPUMS collapses industry categories into broader categories to assure confidentiality of the respondents who work on these broader industries.

When six-digit industry data is available, like for example on the Quarterly Census of Employment and Wages (QCEW) most cluster analysis using these data is aggregated into the industry level, basically transforming each industry cluster as a black-box, where we know little about the characteristics and outcomes of the workers within the cluster. This approach leaves researchers and policy makers without the availability to identify workers within a cluster and learn which workers benefit from the productivity gains in the cluster and which workers remain at the margin of these gains. Therefore, it is essential to map the workers within a cluster using micro data sets to understand their labor market outcomes such as occupation, hours of work, or income. In addition, mapping clusters into workers will allow us to understand their demographic characteristics, such as gender, age, race, or ethnicity. Again, we can learn who benefits from the cluster and who is on the margin from these benefits.

In this paper, we create an algorithm that disaggregates public use industry data and imputes them into a six-digit NAICS category. This algorithm would allow researchers who

do not have access to confidential data using six-digit NAICS codes to still create cluster definitions that are consistent across data and across time. To illustrate the importance of accessing data of workers within a cluster, we present a descriptive analysis of wage inequality across different productivity clusters, and show preliminary data that suggest that high-productivity clusters increase productivity of all workers, but they significantly reward their most productive workers. In contrast, low-productivity workers, seem to penalize workers in the bottom of the distributions and workers with little formal education.

2. Describe the Algorithm

Our exercise will map respondents' industry in the 2019 American Community Survey (ACS) whose NAICS was aggregated into a three, four, or five-digit category to a six-digit NAICS using the distribution from the Quarterly Census of Employment and Wages (QCEW, a national census collected from employers' survey).

We define an industry code in the ACS as w_{inst} that refers to workers in industry *i* collapsed into *n* digits in region *s* and year *t* (where *n* is usually 3, 4, or 5). Also, we define $\widehat{q_{\imath 6st}}$ as the imputed six-digit industry on the ACS data that corresponds to the six-digit industry in the QCEW q_{inst} such that w_{instt}={ $\widehat{q_{16st}}$, $\widehat{q_{26st}}$ ....$\widehat{q_{N6st}}$} .

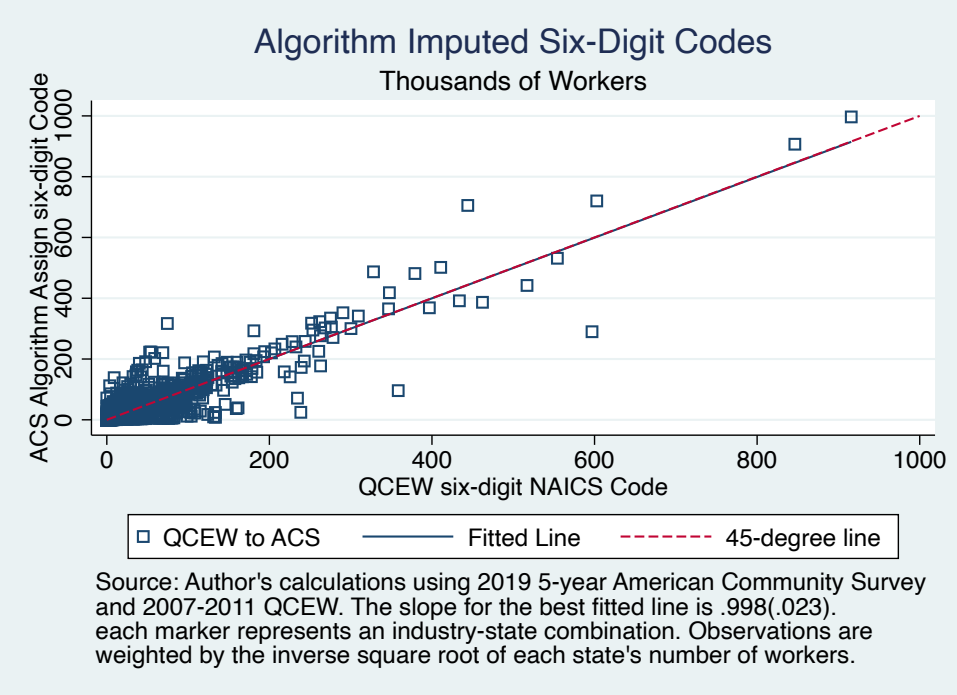
The empirical challenge we face is how to assign each worker to the correct $\widehat{q_{\imath 6st}}$ that to the researcher is unknown. In here, we do so randomly within a defined geographical area (in this exercise we use states). Random selection would likely minimize any errors in the data as most workers within the q_{i6st} category would aggregate into the same industry cluster. Our intent in doing so lies in not assigning category $\widehat{q_{\imath 6st}}$ to every worker in the ACS, but to replicate the distribution according to the QCEW six-digit q_{i6st} industry code. To achieve this, we must (1) save the newly recorded category for each worker, and (2) make sure that we do not oversample within each category. Hence once that the right number of workers is appointed to the six-digit $\widehat{q_{\imath 6st}}$ is assigned, we move to the next six-digit category. In total we distribute workers from w_{inst}=(w_{1nst} ,w_{2nst}..., w_{253nst}) categories into $\widehat{q_{\imath 6st}} = (\widehat{q_{1,6st}}, \widehat{q_{2,6st}}$ ....$\widehat{q_{1356,6st}}$). That is, from 253 categories in the American Community Survey's NAICS variable to 1563 six-digit industry variables[1] that reflect the distribution of the QCEW data.

Figure 1 shows the performance of our mapping at the industry level. The number of workers on each QCEW six-digit industry and each state are represented on the horizontal axis, and the number of workers in its corresponding imputed six-digit industry code is represented on the vertical axis. Each marker represents a state-industry combination. The blue solid line represents the best linear fit between the QCEW data and the ACS data. Recall that the QCEW is a census collected from all employers in the United States. The

[1] Finally, in this piece we randomize the data within each w_{inst} at the state level, but it can be done for other geographic areas, such as metropolitan areas. The Stata programs are available from the authors upon request.

ACS data is a survey representative of households in the United States. While there will be some differences between the QCEW and the ACS, a slope of one suggests that both data sets are equivalent. The data on Figure 1 shows that this is the case, with a slope of the best fitted line of 0.99 that is statistically indistinguishable from one.

Figure 1 Number of Workers using Imputed Industry



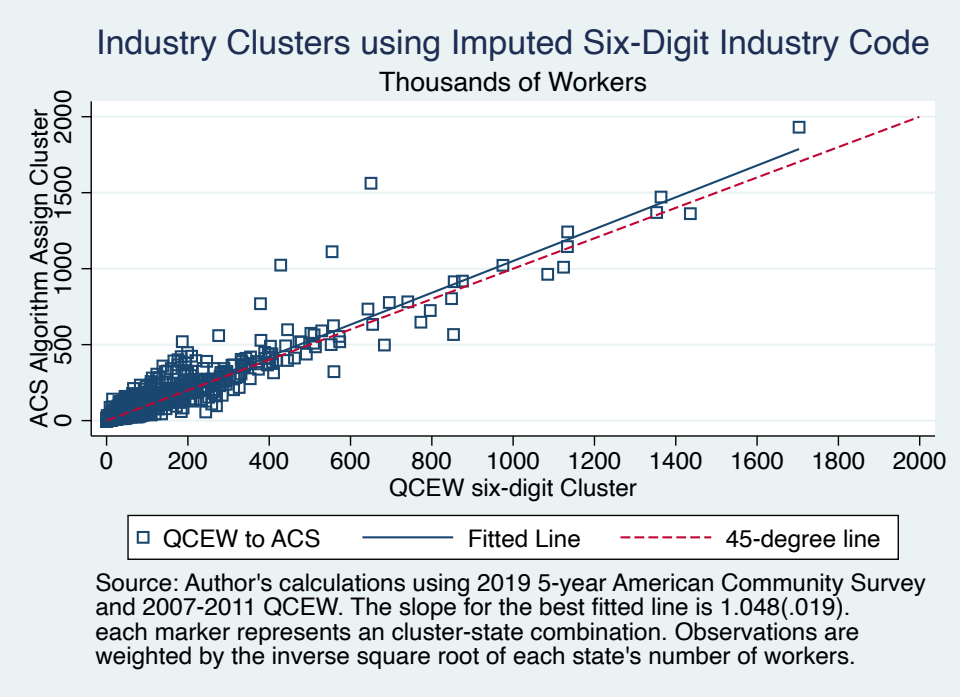
Source: Author's calculations using 2019 5-year American Community Survey and 2007-2011 QCEW. The slope for the best fitted line is .998(.023). each marker represents an industry-state combination. Observations are weighted by the inverse square root of each state's number of workers.

As a second step is to collapse the imputed ACS industry data into industry clusters as defined by Delgado *et al.* (2014). Again, our algorithm works well if, in general, the size of each industry cluster using the QCEW is similar to the size of each imputed industry cluster using the ACS data. Figure 2 shows the data where each marker is a cluster-state combination. Figure 2 suggests that, when aggregating workers from six-digit industry NAICS into clusters, our algorithm tends to over impute the larger clusters in the ACS data. While the empirical test should be that the best-fit line to be equal to one, in this case the best-fit line's slope is 1.04 and statistically different from one at conventional levels.

3. An Application: Income Distribution Within Clusters

As an application of the importance of identifying workers within clusters, we next look at different labor market outcomes within each cluster. Economic theory suggests that workers and firms in clusters learn from each other enhancing productivity (Marshall, 1890). But not all clusters are the same, while different clusters agglomerate workers with different skills, within clusters workers also have different skills.

Figure 2. Number of Workers on Each Cluster using Imputed Industry



In this exercise, we argue that researchers should not only focus on across cluster productivity differentials, but also within cluster differences in productivity. This would allow researchers to understand which workers benefit from the productivity gains within the cluster and which workers remain at the margin. While we know that clusters differ in their agglomeration economies, we do not know whether these economies are available to all workers.

Table 1 presents labor market data from the ACS data across different clusters aggregated by productivity level using average hourly wages as provided by the QCEW: the first column shows hourly wages for different workers in clusters on the bottom productivity quarter, the fourth column shows workers in clusters on the top productivity quarter. The second and third columns represent workers on their respective productivity quarter. In this table we focus on three different measures of income distribution: wage and standard deviation, which are parametric measures of distribution and that we use to compute the coefficient of variation. Second, we focus on different percentile measures such as income at the lowest decile, income at the highest decile, and its difference. Finally, we present average wages for each cluster quartile for workers with college and without college, and the corresponding college wage gap.

The data on Table 1 suggest, not surprisingly, that average wage increase as we move from lower productivity clusters to higher productivity clusters, yet this relationship is not linear, but convex: the highest earners are workers on the most productive clusters, and the difference in wages is the largest compared to workers on the adjacent quartiles. The same relationship is true about the standard deviation, it increases across productivity clusters, and does so on an increasing way. The coefficients of variation, presented on the third row and defined as the ratio of the wage's standard deviation over the mean, show more clear information of the distribution across clusters: low productivity clusters have the largest hourly wage variation, high productivity clusters have the second largest hourly wage variation. Not only do workers in high-productivity clusters earn more, but the dispersion of wages is largest in the low-productivity clusters. Yet, the fact that wages raise convexly across productivity clusters suggest that high productivity clusters reward all workers, but disproportionally the most productive workers.

Table 1. Within Cluster Quartile Distribution Measures of Hourly Wage

	(1) Lowest Productivity	(2) Second Quartile	(3) Third Quartile	(4) Highest Productivity
Mean	21.15	25.91	30.20	38.62
Standard Deviation	31.89	32.83	34.97	48.20
Coefficient Variation	1.51	1.27	1.16	1.25
Lowest Decile	6.17	8.30	9.62	11.06
Highest Decile	38.50	45.17	53.14	72.12
90-10 Differential	32.33	36.87	43.52	61.06
No College	19.52	24.80	29.05	34.72
College Grad	28.29	29.24	34.00	47.47
College Premium	0.37	0.16	0.16	0.31

Source: Authors calculations using 2017-2021 American Community Survey, 2017-2021 Quarterly Census of Employees and Wages, and Delgado et, al (2014) definitions of clusters. Wages are computed by dividing annual earnings by usual hours of work multiplied by 52. Sample: all employed, not-self-employed workers aged 25-65.

Similarly, looking at different percentiles across the distribution, the data on Table 1 suggests that across clusters, workers employed in the low productivity clusters and on the lowest wage decile earn fifty-five cents for every dollar that workers on the same decile but in the highest productivity cluster do. Similarly, workers on the highest wage decile show

similar large differences across different productivity clusters in hourly earnings. Workers on the top decile but in the low-productivity clusters earn fifty-three cents for every dollar that workers in high productivity clusters do. The 90-10 differentials suggest that workers in both the high-productivity clusters and low-productivity clusters experience larger wage dispersion that workers on the middle of the distribution. The big difference that these data suggest is that high-productivity clusters reward their highest paid workers, while low-productivity clusters penalize their lowest paid workers.

Finally, we look at wages of college graduates and workers without a college degree across different productivity clusters. The first datum that is striking is that workers *without* a college degree on the high-productivity clusters earn higher wages than workers *with* a college degree on the low-productivity cluster[2]. This result supports the important insight that high-productivity clusters afford productivity gains to all workers within a cluster, regardless of degree of formal education. Yet, and alike the data for the percentile distribution within clusters, the differences in wages for workers in high- and low-productivity clusters across college attainment confirms the hypothesis that high-productivity clusters reward high-skill workers, low-productivity clusters penalize low-skill workers,

4. Conclusion

In this paper we proposed an algorithm that allows researchers to observe data on labor market outcomes of workers within industry clusters. While the literature on regional studies has long established the productivity gains for workers that are part of different industry clusters, most research either focuses on industry-region-specific case studies or on research that use aggregate data. Understanding each worker's outcome within an industry cluster is important as it would allow researchers to understand which workers benefit from the agglomeration economies provided by the cluster and which workers remain on the margins of these benefits.

To do so, we constructed an algorithm that disaggregates 3,4, and 5-digit NAICS codes available in public use data and imputes each observation a six-digit NAICS code that mimics the industry distribution in the QCEW and allows researchers to map industries into clusters as suggested by Delgado et al. (2014). Our algorithm seems to do a good job of imputing disaggregated six-digit industries to public use data, such as the American Community Survey or the Current Population Survey. Yet, when we aggregate imputed industries into industry clusters, our algorithm tends to over-assign observations to the largest clusters, slightly overestimating the number of workers on each one of these clusters. In future work we will refine the algorithm to improve performance over the cluster aggregation.

[2] These means are unconditional, and we expect that some of these differences are explained with regional productivity differences, but this result is still striking.

Finally, we do an exercise to measure the within cluster distribution of wages by dividing clusters into high productivity clusters and low-productivity clusters. The data in this exercise suggest that wage distribution widens on both high-productivity clusters and on low productivity clusters. But because the relationship between cluster productivity and workers' wage is convex, the causes for the wider distribution across clusters seem to be different: the highest productivity clusters reward the most productive workers, yet all workers benefit from the agglomeration economies of the cluster. In contrast, the lowest paid workers in the low-productivity workers seem to remain at the margins of any productivity gains from belonging to the cluster. The consequences and causes of this convexity remain a topic of future research.

5. References

Mercedes Delgado & Michael E. Porter & Scott Stern, 2016. "Defining clusters of related industries," Journal of Economic Geography, vol 16(1), pages 1-38

Marshall, Alfred, (1920) 1842-1924. Principles of Economics; an Introductory Volume. London :Macmillan.

Moretti, Enrico (2012). The new geography of jobs. Boston, Houghton Mifflin Harcourt.

Porter, Michael E. (1998) The Competitive Advantage of Nations. New York: Free Press, 1990. (Republished with a new introduction, 1998.)

Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. IPUMS USA: Version 15.0 [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D010.V15.0