Are Large-Scale Data from Private Companies Reliable? An Analysis of Location Data on Financial Establishments in a Popular Dataset

Keywords: large-scale data; data quality; classification; reliability; algorithms

Extended Abstract

Large-scale data from private companies offer unique opportunities to examine important topics, such as racial inequality, partisan polarization, and mobility segregation. However, because these data are algorithmically generated from private companies with their own purposes, their accuracy and reliability for social science research remain unclear (Grigoropoulou and Small 2022). Classification problems are a prime example of purpose discord, wherein private companies construe them as predictive tasks, while social scientists repurpose the outputs of these classifications to examine human behavior. A classification can be sufficiently accurate from a predictive perspective, given the practical limits to prediction (Hofman, Sharma, and Watts 2017), while being inaccurate from a measurement perspective, given the needs of scientific research (Bailey 1994). The present study examines how quality issues in large-scale data from private companies can afflict the reporting of even ostensibly uncomplicated values. We assess the reliability with which the company, SafeGraph, sorted sourced data on financial institutions, such as banks and payday lenders, into industry categories based on a standard classification system.

In early 2022, we acquired data from SafeGraph, a popular device-tracking company that has informed many important recent studies (Chang et al. 2021; Chen and Rohla 2018; Huang et al. 2023; Massenkoff and Chalfin 2022). For each establishment, SafeGraph provides, among others, the establishment's name, addresses, and North American Industry Classification System (NAICS) codes, a widely-used classification standard based on precise descriptions of primary business activities. We downloaded location data for a total of 202,750 establishments SafeGraph had classified in the six 6-digit NAICS codes that most closely matched our research interest in consumer-facing conventional vs alternative financial institutions in the U.S. (NAICS codes in parenthesis): Commercial banks (522110), Savings institutions (522120), Credit unions (522130), Consumer lending (522291), All other nondepository credit intermediation (522298), and Other activities related to credit intermediation (522390).

Our objective for the present study was to assess how well SafeGraph's algorithm classified its establishment data into the six 6-digit NAICS codes, by independently classifying the 202,750 establishments ourselves and comparing results. To do so, we used two main strategies: First, we classified each establishment automatically based on high-confidence inclusion keywords indicating clear industry markers, such as "savings bank," "car title loan," "check* cash*," "payday loan," etc. in the establishment's location name in the dataset. We also searched for exclusion terms signaling other business activities that may have been wrongly classified into one of the six codes (e.g. "bitcoin). Second, we classified all establishments of the same company based on information for the company's activities from company websites and various government documents. Finally, we probe for potential flaws in our procedures by conducting individual online searches for several thousand of establishments, and for selected establishments, by doing fieldwork in New York City.

In addition to classification problems, we examined the possibility of duplicate records or overlooked establishment closures.

Our results suggest extensive problems in the classification of the location data. First, with respect to data coverage, SafeGraph undercounts savings institutions, establishments on all other nondepository credit intermediation, credit unions, and commercial banks, by 99.91%, 25.57%, 15.42%, and 6.36%, respectively, compared to the establishment counts in the County Business Patterns (CBP) from the U.S. Census; it overcounts establishments in other activities related to credit intermediation by an extraordinary 475.13%. Second, with respect to reliability, we calculate the probability of detection and precision of the establishments' classification by SafeGraph into what we determined to be the appropriate, primary NAICS code based on our classification procedures (Table 1). The probability of detection suggests that a researcher using the data would detect almost all (99.8%) banks, and a vast majority of credit unions (88.6%), while missing a large portion, or a majority, among the other four industry categories. For precision, the estimates range from a low 4.1% for other activities related to credit intermediation to 89.9% for commercial banks.

Finally, we find that approx. 16.7% of the establishments in the data were not in operation in 2022, with wide variation in unidentified closures by industry category. Our inspection of duplicate records also revealed that 10.6% of financial establishments in the data are likely duplicates.

In conclusion, we find evidence of major classification problems that vary by type of institution, and considerably high rates of unidentified closures and duplicate records. The problems are of sufficient magnitude to have affected the conclusions drawn in our empirical research. Moreover, the processes through which the company produced the data suggest that similar problems are likely to affect other data sources were algorithms play a major role in classification systems. We conclude with recommendations for scholars working with company data.

References

Bailey, Kenneth D. 1994. *Typologies and Taxonomies: An Introduction to Classification Techniques*. Thousand Oaks, Calif: Sage Publications.

Chang, Serina, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2021. "Mobility Network Models of COVID-19 Explain Inequities and Inform Reopening." *Nature* 589(7840):82–87. doi: 10.1038/s41586-020-2923-3.

Chen, M. Keith, and Ryne Rohla. 2018. "The Effect of Partisanship and Political Advertising on Close Family Ties." *Science* 360(6392):1020–24. doi: 10.1126/science.aaq1433.

Grigoropoulou, Nikolitsa, and Mario L. Small. 2022. "The Data Revolution in Social Science Needs Qualitative Research." *Nature Human Behaviour* 1–3. doi: 10.1038/s41562-022-01333-7.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. "Prediction and Explanation in Social Systems." *Science* 355(6324):486–88. doi: 10.1126/science.aal3856.

Huang, Justin T., Masha Krupenkin, David Rothschild, and Julia Lee Cunningham. 2023. "The Cost of Anti-Asian Racism during the COVID-19 Pandemic." *Nature Human Behaviour*. doi: 10.1038/s41562-022-01493-6.

Massenkoff, Maxim, and Aaron Chalfin. 2022. "Activity-Adjusted Crime Rates Show That Public Safety Worsened in 2020." *Proceedings of the National Academy of Sciences* 119(46):e2208598119. doi: 10.1073/pnas.2208598119.

Table 1. Estimates of classification agreement between SafeGraph and Authors.

NAICS Code	Assigned by SafeGraph	Assigned by Authors	Agreements	Probability of Detection	Precision
522110 Commercial banks	82,039	73,876	73,755	99.8%	89.9%
522120 Savings institutions	6	3,331	5	0.2%	83.3%
522130 Credit unions	16,236	16,119	14,285	88.6%	88.0%
522291 Consumer lending	15,180	5,768	4,265	73.9%	28.1%
522298 All other nondepository credit intermediation	7,926	11,552	6,769	58.6%	85.4%
522390 Other activities related to credit intermediation	81,363	8,041	3,301	41.1%	4.1%
Total	202,750	118,687	102,380	86.3%	50.5%