4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15100

Cracking the Code of Geo-Identifiers: Harnessing Data-Based Decision-Making for the Public Good

Patricia Snell Herzog
Indiana University Lilly Family School of Philanthropy; Department of Human-Centered Computing, School of Informatics & Computing; Department of Sociology, IUPUI, USA

Abstract

The accessibility of official statistics to non-expert users could be aided by employing natural language processing and deep learning models to dataset lexicons. Specifically, the semantic structure of FIPS codes would offer a relatively standardized data dictionary of column names and string variable structure to identify: two-digits for states, followed by three-digits for counties. The technical, methodological contribution of this paper is a bibliometric analysis of scientific publications based on FIPS code analysis indicated that between 27,954 and 1,970,000 publications attend to this geo-identifier. Within a single dataset reporting national representative and longitudinal survey data, 141 publications utilize FIPS data. The high incidence shows the research impact. Yet, the low proportion of only 2.0 percent of all publications utilizing this dataset also shows a gap even among expert users. A data use case drawn from public health data implies that cracking the code of geo-identifiers could advance access by helping everyday users formulate data inquiries within intuitive language.

Keywords: *Geospatial data; Big data; Official statistics; Bibliometrics.*

Cite as: Herzog, P. (2022). Cracking the Code of Geo-Identifiers: Harnessing Data-Based Decision-Making for the Public Good. Proceedings of the 2022 International Conference on Advanced Research Methods and Analytics: 245-252. DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15100

1. Introduction

Many official statistics provide publicly available datasets of social patterns that could be harnessed in making informed decisions. The funding to collect and share these data is often supported through public entities due to the potential for data to have broad applications that benefit the public good. However, the inaccessibility of the data structure is a barrier to broader use. While data can readily be downloaded, users need to understand the data lexicon, including the meta-data, data dictionary, and most importantly the meaning that can be extracted from data variable names and labels.

A fundamental problem that prevents broader accessibility of publicly available data is the expert-level vocabulary embedded within the syntax of complex datasets. Everyday people and knowledge workers are required to decode this syntax in order to understand the information the data offer. Datasets that have existed for a long time carry a layered legacy of complex codes and dictionary structures that are difficult to make sense of and disentangle. This data syntax is crucial for understanding the meaning of the available variables, and ultimately the kinds of answers that a dataset can provide. However, the complexity of the data syntax obscures the meaning of the data for non-expert users. Explicating this syntax can lessen the expert-level barriers inhibiting broader data usage.

This paper focuses on a common attribute of datasets that is crucial for extracting actionable insights: geospatial data. The geographic location of data are often coded within a relatively controlled vocabulary of geo-identifiers. GeoIDs are frequently coded within a fairly standardized and finite set of codes, and thus the lexicon of geospatial data is a prime syntax to detect in automation procedures. Machine learning can be utilized to detect semantics of GeoIDs by developing data dictionaries with common location attributes.

2. Geospatial Data

Geospatial data can connect across otherwise disparate facets the data pipeline, from data acquisition to analysis and visualization (Breunig et al., 2020). Moreover, geospatial data aid replicability of information and analysis techniques across distinct datasets, questions, researchers, and stakeholders: from academia to urban planning (Lee and Kang, 2015). Yet, despite shared semantic foundations in geospatial data ontology, heterogeneity in data lexicons remains a barrier to broader sharing and accessibility (Sun et al., 2019).

2.1. Controlled Vocabularies

Rieder (2020) compares column names to the contracts or promises that software products make with users, with the caveat that published data tables provide a service to users that is more ambiguous. Data consumers are offered information but without clear parameters. Within this context, column names provide an informal contract with the data user

regarding what information can be harnessed from which variable. As contracts benefit from a relatively standard set of vocabulary to express the promises that users can expect, Rieder asserts that engaging a controlled vocabulary within column names can serve as a latent contract between data producers and consumers, with accessibility, integration, and transferability promised within a recognizable lexicon. For example, ID can be used to indicate a uniquely identified entity in the dataset, and N can be used for sample counts.

2.2. Geographic Identifiers (GeoIDs)

Geo-identifiers indicate to which geographic units the data can be aggregated, and geo-identifiers are nearly ubiquitous within publicly available datasets. Many public entities are geopolitically structured at the level of country, state, city, county, and thus associated data are also imbued with these geographic units. The United States Census Bureau (2021) states that: "GEOIDs are numeric codes that uniquely identify all administrative/legal and statistical geographic areas for which the Census Bureau tabulates data. From Alaska, the largest state, to the smallest census block in New York City, every geographic area has a unique GEOID." In official statistics, there are several different GeoID coding systems, such as the American National Standards Institute (ANSI), Geographic Names Information System (GNIS), and Federal Information Processing Series (FIPS) codes.

2.3. Federal Information Processing Series Codes (FIPS)

There are two primary appeals of focusing on the lexicon for Federal Information Processing Series (FIPS) codes. First, FIPS codes have a high degree of standardization and nested structure in the data dictionary (US Census Bureau 2020; see Figure 1). Second, there is broad utilization of FIPS codes across a range of analyses and within subsidiary datasets (see for example: Mullen and Bratt, 2018; Brown et al., 2020; Boland et al., 2017; Roberts et al., 2014). The next section quantifies the scientific research impact of FIPS.

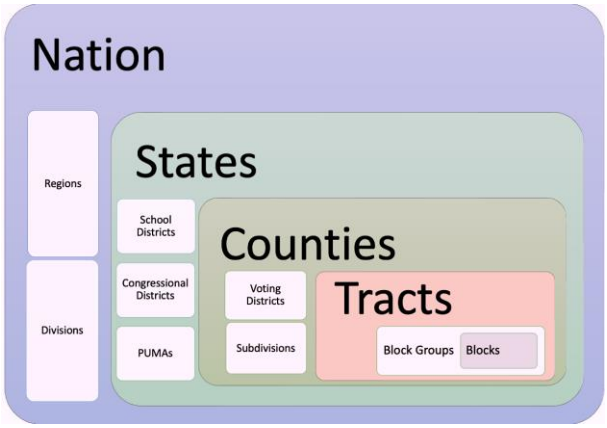

Figure 1. Nested Geographic Entities. Source: Author creation based on US Census (2020).

3. Bibliometric Data

Bibliometrics is a methodological approach that focuses on scientific literature as the subject of analysis (Ball, 2017). Many of these techniques focus on citation analysis, including content analysis of titles, keywords, abstracts, and full text of published journal articles, books, conference proceedings, dissertations, and reports (Zhao and Strotmann, 2015). Applying scientific techniques to citation analysis facilitates a statistical evaluation and measurement of influence within the scientific community (Iftikhar et al., 2019). Citation analyses have been utilized in studying research impact from social work (Holden et al., 2012) to the humanities (Ochsner et al., 2016). This paper presents a bibliometric analysis of FIPS code impact within scientific literatures and official statistical datasets.

3.1. Census Data

Searching census data within a popular scholarly bibliometric database Google Scholar returns 4,230,000 results, and 1,970,000 of these entries were published since 2010. Moreover, 6,400 publications cite FIPS codes within census data analysis. Computing the same set of analyses within Scopus respectively returns: 47,761; 27,954; and 27. Combined, these results indicate the high degree of research impact that FIPS codes have.

3.2. PSID Data

Additionally, FIPS codes have been utilized within subsidiary datasets. For example, the Panel Study of Income Dynamics (PSID) is a longitudinal survey of a nationally representative sample of U.S. based more than 18,000 individuals within 5,000 families. Data have been collected in over 40 waves of data spanning multiple generations of descendants from the original respondents. In the PSID bibliographic database of citations (PSID, 2022), there are a total of 7,033 publications in this database, of which 4,892 are journal articles, 785 book chapters, 92 books, 1,180 dissertations, and 84 reports. The PSID uses FIPS codes, and Table 1 displays the bibliometric data respective to each geography.

Table 1. Table captions should appear *above* tables.

Geographic Entity	Count	Percent
Tract	31	0.4
County	03	0.0
Metropolitan Area	22	0.3
Region	20	0.3
Urban / Rural	65	0.9
TOTAL	141	2.0

Source: Author creation based upon the PSID (2022).

4. Deep Learning

In order to harness the power of geo-identifiers to unlock the information in the thousands of scientific publications summarized in the previous section, it is necessary to correctly detect the semantic structure. Though the geo-identifier lexicon is fairly complex and typically embedded in dirty data, it is also finite and more standardized than unstructured text. The controlled vocabulary of FIPS codes offers an opportunity to apply a multi-input deep neural network for detecting semantic types, such as Sherlock (Hulsebos et al., 2019).

4.1. GeoID Structure

As displayed in Figure 1, the GeoID structure of FIPS codes is nested. Specifically, the structure begins with a 2-digit state code, such as 18 for Indiana. This is followed by a 3-digit county code, such as 097 for Marion County, which includes the city of Indianapolis. The nested structure to FIPS codes is such that these identifiers can be combined into a 5-digit code = 2 for state + 3 for county: 18097. The standardization of the digit format in FIPS codes lends itself to a detectable lexicon, as the data dictionary can be trained to identify recurring combinations of 2-digit, 3-digit, and 5-digit string data as a geography. Moreover, the deep learning process can be improved by harnessing column names, which would be a finite variety of for instance: state, STATE, sta, county, COUNTY, cty.

Continuing to smaller geographies, the Census approximation of a neighborhood is a tract, which has a 6-digit code. For example, one tract within Marion County is 310104, and a neighboring tract is 310105. These are sometimes designated with a period after the first four digits, as: 3101.04 and 3101.05. This indicates that a broader set of neighborhoods can be grouped together within the 3101 identifier. However, some datasets would omit the period and others would not. This presents a complication to automatically detecting the semantic lexicon, yet again the dictionary in column names is limited: tract, trc, tra, ct. Moreover, an 11-digit string variable (state+county+tract) is readily identifiable.

4.2. Public Health Data

Applied to a specific data use case, the Indiana Department of Health provides a data hub of publicly available data regarding health issues (IN.MPH 2022a). Currently, this data hub has a prevalent array of datasets reporting Covid-19 rates: tests, cases, deaths, trends over time. One available dataset is for Covid-19 county statistics (IN.MPH 2022b). The meta-data include this additional information: Spatial/Geographic Coverage – State of Indiana; Granularity – Aggregate, County-Level. The data dictionary reports four fields: County (Reported county where patient resides), Case_Count (Number of reported Covid-19 cases), Death_Cases (Number of reported Covid-19 deaths), and Lab_Tests (Number of reported individuals with a resulted Covid-19 test).

Yet, the dataset actually contains six fields, with the addition of _id and Location_ID. Plus, County is not labeled as described in the data dictionary but is rather labeled: County_Name. Thus, to train a machine learning model well, it is necessary to identify the concactanated column name of county to match it to the data dictionary. More importantly, in terms of harnessing the power of FIPS codes, it is crucial to identify the string pattern of Location_ID. In this example, all the values in this field begin with 18, which is the state FIPS for Indiana. This is readily visually detectable to a human, at least one with the necessary expertise to recognize the state digits. Another human-detectable clue is that there are a total of 92 counties in the state of Indiana, and the dataset has 92 entries.

If a deep learning model is trained to recognize Location_ID as a FIPS code, then the power of the dataset can be harnessed through identification of its semantic structure. Only then could a non-expert user be automatically provided with a natural language description of the information: Covid-19 tests, cases, and deaths by Indiana county. Building upon the intuitive human curiosity structure of questions, this data could generate answers to the question: Which counties in Indiana have the highest Covid-19 rates? Even more complex, the dataset could also offer answers to an inquiry regarding the preventive contributions of testing through responding to the question: Do counties in Indiana that have a lower deaths to cases ratio have a higher test rate?

5. Conclusion

In conclusion, the accessibility of official statistics to non-expert users could be aided by employing natural language processing and deep learning models to dataset lexicons. Specifically, the semantic structure of FIPS codes would offer a relatively standardized data dictionary of column names and string variable structure to identify. The bibliometric analysis indicated that decoding this structure would enable access to the insights included in thousands of scientific publications. The datasets embedding FIPS codes span from macro-level geopolitical units in census data, to public health data aggregated to counties, to individual and family survey data aggregated to tracts, counties, stages, and regions. Thus, cracking the code of geo-identifiers could advance access by everyday people by helping users to formulate data-based inquiries within their intuitive language.

Acknowledgments

The author is grateful to Davide Bolchini, Rama Sai Arun Varma Pensmatsa, Anshuman Dixit, and Rahul Yadav for collaborations on the question-generation project. Additionally, the author is grateful to Laurie Paarlberg for her collaborations in conceiving of data as a public and philanthropic resource; Una Okonkwo Osili and Mark Ottoni-Wilhelm for their

contributions to the Panel Study of Income Dynamics; and the National Science Foundation for funding a human-technology frontier workshop that informed this project (1934942).

References

Ball, R. (2017). *An Introduction to Bibliometrics: New Development and Trends*. Chandos Publishing.

Boland, M. R., Parhi, P., Gentine, P., & Tatonetti, N. P. (2017). Climate Classification is an Important Factor in Assessing Quality-of-Care Across Hospitals. *Scientific Reports*, *7*(1), 4948. https://doi.org/10.1038/s41598-017-04708-3

Breunig, M., Bradley, P. E., Jahn, M., Kuper, P., Mazroob, N., Rösch, N., Al-Doori, M., Stefanakis, E., & Jadidi, M. (2020). Geospatial Data Management Research: Progress and Future Directions. *ISPRS International Journal of Geo-Information*, *9*(2), 95. https://doi.org/10.3390/ijgi9020095

Brown, C. C., Moore, J. E., Felix, H. C., Stewart, M. K., & Tilford, J. M. (2020). County-Level Variation in Low Birthweight and Preterm Birth: An Evaluation of State Medicaid Expansion under the Affordable Care Act. *Medical Care*, *58*(6), 497–503. https://doi.org/10.1097/MLR.0000000000001313

Holden, G., Rosenberg, G., & Barker, K. (Eds.). (2012). *Bibliometrics in social work*. Routledge.

Hulsebos, M., Hu, K., Bakker, M., Zgraggen, E., Satyanarayan, A., Kraska, T., Demiralp, Ç., & Hidalgo, C. (2019). Sherlock: A Deep Learning Approach to Semantic Data Type Detection. *ArXiv:1905.10688*. http://arxiv.org/abs/1905.10688

Iftikhar, P. M., Ali, F., Faisaluddin, M., Khayyat, A., De Gouvia De Sa, M., & Rao, T. (2019). A Bibliometric Analysis of the Top 30 Most-cited Articles in Gestational Diabetes Mellitus Literature (1946-2019). *Cureus*, *11*(2), e4131. https://doi.org/10.7759/cureus.4131

Lee, J.-G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, *2*(2), 74–81. https://doi.org/10.1016/j.bdr.2015.01.003

Mullen, L. A., & Bratt, J. (2018). USAboundaries: Historical and Contemporary Boundaries of the United States of America. *Journal of Open Source Software*, *3*(23), 314. https://doi.org/10.21105/joss.00314

Ochsner, M., Hug, S. E., & Daniel, H.-D. (Eds.). (2016). *Research Assessment in the Humanities*. Springer International Publishing. https://doi.org/10.1007/978-3-319-29016-4

PSID. (2022). *Panel Study of Income Dynamics Bibliography Search*. https://psidonline.isr.umich.edu/Publications/Bibliography/search.aspx

Riederer, E. (2020, September 9). *Column Names as Contracts: Using controlled dictionaries for low-touch documentation, validation, and usability of tabular data*. GitHub. https://github.com/emilyriederer/website/issues/9

Roberts, J. D., Voss, J. D., & Knight, B. (2014). The Association of Ambient Air Pollution and Physical Inactivity in the United States. *PLOS ONE*, *9*(3), e90143. https://doi.org/10.1371/journal.pone.0090143

Sun, K., Zhu, Y., Pan, P., Hou, Z., Wang, D., Li, W., & Song, J. (2019). Geospatial data ontology: The semantic foundation of geospatial data integration and sharing. *Big Earth Data*, *3*(3), 269–296. https://doi.org/10.1080/20964471.2019.1661662

US Census Bureau. (2020, November). *Standard Hierarchy of Census Geographic Entities*. https://www2.census.gov/geo/pdfs/reference/geodiagram.pdf

US Census Bureau. (2021, October 8). *Understanding Geographic Identifiers (GEOIDs)*. https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html

Zhao, D., & Strotmann, A. (2015). *Analysis and visualization of citation networks*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00624ED1V01Y201501ICR039