Rebecca Stewart

Prof. Matthew Mayernik

LIS 545 B: Data Curation I

14 March 2024

Term Project - Final Report

I. Data and Metadata Profile

The dataset I selected for review is *School Neighborhood Poverty Estimates, 2020-2021*, which is based on school locations from the Common Core of Data (CCD) school file as well as on income data from families with children 5-18 in the U.S. Census Bureau's 2017-2021 American Community Survey (ACS). This dataset uses IPR (income-to-poverty ratio) values calculated by the Census Bureau to establish its poverty estimates. As such, the estimates included here describe the conditions for families with children living in neighborhoods around specific schools, not necessarily the conditions of students enrolled in those schools.

The publisher for this dataset is the National Center for Educational Statistics (NCES). As the NCES is the "primary federal entity for collecting and analyzing data related to education," the U.S. federal government and related federal agencies such as the U.S. Department of Education are stakeholders for the dataset, along with the NCES. The dataset draws on U.S. census data, making the U.S. Census Bureau a stakeholder as well. Other stakeholders include U.S. public elementary and secondary schools (included in the CCD school file), students and their families, and other people living in the neighborhoods around public schools; these entities and individuals have not only contributed data to this NCES project but may also be impacted by policies, research, and other projects drawing on this dataset in the future. The people and organizations whose work involves education policy, social work, and/or relates to income inequality and who may directly or indirectly work with this dataset in the future are additional stakeholders.

The *School Neighborhood Policy Estimates, 2020-2021* dataset comprises seven downloads and links found in the "downloads & resources" section of the landing page. Five of these files can be downloaded directly from the landing page: the original ISO-19139 metadata, the GeoJSON, a CSV file, a KML file, and a zipped “Shapefile” folder including an XML file. Data visualizations in the form of ArcGIS maps can be accessed in two different locations, which are also linked to from the landing page; there is “the ArcGIS Hub Dataset” hosted by Edge Open Data / NCES Open Data, and there is a link to the “ArcGIS GeoService” landing page which connects to the same ArcGIS Hub Dataset (but one which is hosted directly on the ArcGIS website, not via Edge Open Data). The GeoJSON, CSV, and XML files do not require special software to open and analyze. The KML file is used specifically with mapping applications such as Google Earth.

The information collected in this dataset is in the public domain, though the landing page specifies that “data users are advised to review NCES program documentation and feature class metadata to understand the limitations and appropriate use of these data.”

This dataset comes with a couple of different types of metadata. It links to an XML file under “Original ISO-19139 metadata,” which is listed along with the other “Downloads & Resources” on the main page for the data set. Further down the landing page, under “Metadata Source,” a download link for “Data.json Metadata” is included. Under “Other Data Resources,” a link is provided to “Geoplatform Metadata Information,” but this link was broken as of March 2024. (The metadata is noted to have been updated on November 4, 2023.) The data files themselves (e.g., the CSV file) include basic metadata as well since that is necessary for interpreting the data tables: school names, IPR values, academic year surveyed, and geographical coordinates are provided.

The XML file specifies that as of July 2023, it conforms to ISO 19139 Geographic Information - Metadata - Implementation Specification (v. 2007). According to ISO.org, the status of this metadata standard is “withdrawn,” and it has since been revised by: ISO/TS 19139-1:2019. Comparing the two standards is not within the scope of this assignment, but simply as a matter of compliance and in making the data more usable for the broader community, the metadata would be enriched by updating the standard to the 2019 version.

Another recommendation is to update the dataset landing page to reflect the current (2020-2021) version of the dataset; the metadata files correctly specify 2020-2021 collection dates and 2017-2021 ACS data, but the abstract on the landing page refers to the dataset as the *2018-2019 School Neighborhood Poverty Estimates* and refers to 2015-2019 ACS data. The landing page title on data.gov reads "School Neighborhood Policy Estimates – Current" and it appears that a single identifier is being re-used every year when the files are updated (past years' datasets are archived and available on the NCES site). I am uncertain of best practices here because the identifier is not a DOI, but assigning a unique identifier and landing page to each year's dataset might also be an option, as this would make citation a lot clearer as well.

Beyond this issue, there are no apparent problems with the metadata, and it seems fairly comprehensive. For instance, data quality information and provenance information are included, as is information on maintenance and on the frequency of updates. The XML was last updated in November 2023, which is in keeping with the annual updates specified in the document.

The XML file contains a link to a U.S. Department of Education publication entitled *Education Demographic and Geographic Estimates Program (EDGE): School Neighborhood Poverty Estimates - Documentation (NCES 2018-027)*. This document outlines the purpose and process for the data collection, as well as a basic overview of file format and variables. This publication can also be found on the NCES landing page for School Neighborhood Poverty Estimates, where users can access available data sets for each academic year of data collection.

To get a quick overview of some publications that have used this dataset, I conducted a simple keyword search using Google Scholar, using the terms "NCES 2020-2021 data school neighborhood poverty estimates." "School neighborhood poverty estimates" might require an exact phrase search, as it is the most specific way to refer to this dataset. Based on my initial search, I could see there are a number of recent academic articles referencing this dataset. Many are in the field of education policy, including "School deserts: Visualizing the death of the neighborhood school" (Alexander and Massaro, 2020), which draws on the 2018-2019 SNP dataset. Other publications that are more recent or slightly broader in scope, such as "Measuring School Economic Disadvantage" (Spiegel et al., 2024) might cite either the

2020-2021 dataset, data from multiple school years, or the methods of this annual data collection more generally. As expected, many of the records are academic articles published in journals related to education policy. Using the UW Libraries databases would also work for a more targeted search on related topics and would typically be preferable to Google Scholar.

II. Repository Profile: Kaggle

For this part of the project, I investigated Kaggle as a possible repository for redistribution of the *School Neighborhood Poverty Estimates* dataset. Kaggle is an open data repository but also describes itself as “the world’s largest data science and machine learning community”; it is designed to support current and aspiring data scientists and those with an interest in machine learning and artificial intelligence algorithms. Kaggle has more of a community focus, allowing users to explore a wide range of datasets, to work with those datasets directly on the platform if they wish, and to share their own datasets. The range of content is also much broader on Kaggle than it is on data.gov.

Since the target audiences for Kaggle and for data.gov do not completely overlap, it seems reasonable that the NCES data might reach new users on this new platform. Even as a non-specialist in education policy, I find the content of this government dataset to be of personal interest, and I think it might likewise appeal to Kaggle users browsing by content category or by other sorting options for a dataset to work (or practice) with. It might also be put to new kinds of uses, even more informal ones, since the site makes it easy to explore data files within the web browser and to “play” with them in the Jupyter notebook environment. The GIS aspect of the dataset might interest users who want practice with mapping software, and it might also provide a point of comparison to other datasets with geographically mapped values. Finally, because the overall quality of a dataset is a factor in its usefulness to community members, the high quality of the NCES dataset and trustworthiness of the publisher should also give it some appeal to Kaggle users.

This repository is open to data submissions from anybody with a Kaggle account, which requires an email address but no other personal information or banking information. Though there are some technical specifications and accessibility recommendations, there are not any stated limits to what can be deposited in terms of the content of datasets. The most common and best supported file types on Kaggle include CSVs, JSON, SQLite, Archives, and BigQuery; as discussed on their “How to Use Kaggle” support page, any file type can be uploaded, but formats aside from those listed will be less well-supported and likely less familiar to Kaggle users (therefore less accessible). Kaggle notes that there are use cases for alternative data formats including NPZ, PNG, and HDF5. In such cases, their suggestion is to upload a Notebook including an explanation of what files are included and how to work with them.

To the best of my understanding, since Kaggle is a repository and also an online community, that community aspect means that high-quality datasets shared in accessible formats will get the most attention. Dataset topics that are “interesting” to a broader audience for whatever reason (sometimes this might be a very niche dataset topic, like “Temperature and Ice Cream Sales,” or “The Complete Pokemon Dataset”) might also have a better chance of trending on the site. Some examples of content categories for Kaggle datasets include Education, Law, Transportation, Earth and Nature, Retail and Shopping, and Online Communities.

Kaggle does not specifically refer to a “Submission Information Package (SIP)” per the OAIS Reference model, but it provides minimum required fields for publishing a dataset. These include the title, URL, and data files, which must be uploaded from a single source (either a local machine, remote files via public URL(s), a Github Repository, or Notebook Outputs.) Datasets are set to private by default, but if the visibility is changed to public, then license and owner information must be added as well. Kaggle recommends adding a cover image, subtitle, tags, and a description as well, and there is the option to publish a public Notebook.

While reviewing the site, I could not find any way of connecting directly with support staff, but there is a very active community on the forums, which is likely able to fulfill the role of “human assistance” for most queries. Kaggle also has a YouTube channel with a number of “how-to” videos, as well as a very thorough set of online tutorials and FAQ pages.

Kaggle allows a lot of flexibility with dataset uploads and requires only minimal metadata from users. The required and recommended metadata it does request are collected by filling in certain fields during the dataset creation process and file upload process (eg., title and license information). Some basic specifications are provided for handling metadata in individual files according to file type, but these are very standard; for instance, Kaggle recommends including a header row for CSV files. To the best of my knowledge after exploring the site, the decision to adopt a particular metadata standard is up to individual users on the Kaggle platform who are publishing data there, and Kaggle does not make changes to datasets so they will conform to a certain standard.

To share and download datasets, and to access other features of the community, a Kaggle account and login are required. Users can register with Google or with their email. No other personal information is collected aside from the email address, and the account is free. Direct downloads are available for all files, and supported file types can be explored within the web browser. Users can also work directly with the data in a Jupyter Notebook on the Kaggle site, and they can create a new dataset from a Notebook's output files.

The specific contents of a Dissemination Information Package or DIP will vary from dataset to dataset on Kaggle, as a lot of freedom is left up to individual "curators" on what files and file types are included. However, users have the option to download data files directly from each dataset. I selected a few datasets at random to test the download process, and the simplest one included a single CSV file, while a more complex one included several nested folders and various data files. Metadata can be viewed in the web browser, and Kaggle provides the option of downloading the metadata for any dataset as a JSON file in Croissant format.

III. Recommended data citation

National Center for Educational Statistics. (2022). *School Neighborhood Policy Estimates, 2020-2021* (1.1) [Data set]. Data.gov, 4469bbf9-c98b-47e6-b0f1-bbd38ab9d0e0. https://catalog.data.gov/dataset/school-neighborhood-poverty-estimates-current-c7e05

IV. Considerations for long-term preservation

As noted above, the GeoJSON, CSV, and XML files included in this dataset are formats openly accessible and usable by anyone. They are very common formats and do not pose concerns related to long-term preservation. The KML file format is also widely used and openly accessible; it too should be a viable format over the long term, as it is no longer a proprietary format. Documentation is available for all four of these data types.

Up to this point, the owners of the dataset have followed through with timely maintenance of the dataset and its metadata. It is very reasonable to count on the long-term stability of the data.gov repository and on the publisher (NCES), as these are tied to federal government agencies. Because there is a plan in place for regular maintenance of the data, it can be assumed that any future issues related to obsolescence will be taken care of as they arise.

V. Copyright license statement

School Neighborhood Policy Estimates, 2020-2021 is a government dataset and has been placed into the public domain using the CC0 Public Domain Declaration per the U.S. Open Data Action Plan. This is indicated on the original dataset landing page (on data.gov) with the license description "us-pd." This license allows for redistribution of the dataset.

VI. Human subject considerations

This dataset has been made publicly available by the U.S. government, and its IPR estimates are based on publicly available data, namely, the American Community Survey (ACS) administered by the U.S. Census Bureau and the Common Core of Data (CCD) for U.S. schools. Research drawing on public use data such as this (and/or the redistribution of this data) does not typically involve human subject considerations, under the federal regulations for human subjects (45 CFR Part 46).

Of the three related government datasets, the ACS is the only one with potential human subject considerations, as it is based on surveys of individual families over a 5-year period. However, no personally identifiable information is included in the ACS dataset, and the *School Neighborhood Policy Estimates* dataset does not introduce any additional human subject considerations. Confirming this, the U.S. Census Bureau "Census Datasets" page specifies that in regard to its public use datasets, "all personally identifiable information [has been] removed to ensure confidentiality."

VII. References

Dataset selected:

National Center for Educational Statistics. (2022). *School Neighborhood Policy Estimates – 2020-2021* (1.1) [Data set]. Data.gov, 4469bbf9-c98b-47e6-b0f1-bbd38ab9d0e0. https://catalog.data.gov/dataset/school-neighborhood-poverty-estimates-current-c7e05

Publications related to this data:

Alexander, M., & Massaro, V. A. (2020). School deserts: Visualizing the death of the neighborhood school. *Policy Futures in Education*, *18*(6), 787–805. https://doi.org/10.1177/1478210320951063

Geverdt, D. (2018). Education Demographic and Geographic Estimates Program (EDGE): School Neighborhood Poverty Estimates - Documentation (NCES 2018-027). U.S. Department of Education. Washington, DC: National Center for Education Statistics. http://nces.ed.gov/pubsearch/

Spiegel, M., Clark, L. R., Domina, T., Radsky, V., Yoo, P. Y., & Penner, A. (2024). Measuring School Economic Disadvantage. *Educational Evaluation and Policy Analysis*, 01623737231217683. https://doi.org/10.3102/01623737231217683