Rebecca Stewart

Prof. Matthew Mayernik

LIS 545 B: Data Curation I

March 2024

Term Project - Final Project First Draft

1. Data and Metadata Profile

The data I selected for review are the [2018-2019 School Neighborhood Poverty Estimates](https://catalog.data.gov/dataset/school-neighborhood-poverty-estimates-current-c7e05), which are based on school locations from the [Common Core of Data (CCD) school file](https://nces.ed.gov/ccd/files.asp) as well as on income data from families with children 5-18 in the US Census Bureau’s 2015-2019 American Community Survey (ACS). This data set 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 data set is the [National Center for Educational Statistics (NCES)](https://nces.ed.gov/). In addition to the NCES, stakeholders for this data set include the U.S. federal government, the U.S. Census Bureau, 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. They also include people and organizations whose work involves education policy, social work, and/or relates to income inequality (to give just a few examples).

The 2018-2019 School Neighborhood Policy Estimates data comprise 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”](https://data-nces.opendata.arcgis.com/maps/nces::school-neighborhood-poverty-estimates-current) hosted by Edge Open Data / NCES Open Data, and there is a link to the [“ArcGIS GeoService”](https://services1.arcgis.com/Ua5sjt3LWTPigjyD/arcgis/rest/services/School_Neighborhood_Poverty_Estimates_Current/FeatureServer/0) 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 data set 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 data set comes with a couple of different types of metadata. It links to an XML file under “[Original ISO-19139 metadata](https://catalog.data.gov/dataset/school-neighborhood-poverty-estimates-current-c7e05/resource/9ade020f-025a-45d9-a562-5f7b9f040c92),” 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](https://catalog.data.gov/harvest/object/8d07d786-4147-4ae2-8a6a-ed7a2bc90448)” is included. Under “Other Data Resources,” a link is provided to “[Geoplatform Metadata Information](https://geoplatform.gov/metadata/9885828b-f40f-5213-8112-448278adedcf),” but this link was broken as of February 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](https://www.arcgis.com/sharing/rest/content/items/45baf5309f384bb98cb50d31526be680/info/metadata/metadata.xml?format=iso19139) specifies that as of July 2023, it conforms to ISO 19139 Geographic Information - Metadata - Implementation Specification (v. 2007). According to ISO.org, the current status of this metadata standard is “withdrawn,” and it has since been revised by: [ISO/TS 19139-1:2019](https://www.iso.org/standard/32557.html). 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. 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)*](https://nces.ed.gov/programs/edge/docs/EDGE_SIDE_PUBSCH_FILEDOC.pdf)*.* 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](https://nces.ed.gov/programs/edge/Economic/NeighborhoodPoverty), 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 data set, I conducted a simple keyword search using Google Scholar, using the terms “NCES 2018-2019 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 data set. Based on my initial search, I could see there are a number of recent academic articles referencing this data set. Many are in the field of education policy, including [“School deserts: Visualizing the death of the neighborhood school”](https://doi.org/10.1177/1478210320951063) (Alexander and Massaro, 2020), which draws on the 2018-2019 SNP data set. Other publications that are more recent or slightly broader in scope, such as [“Measuring School Economic Disadvantage”](https://doi.org/10.3102/01623737231217683) (Spiegel et al., 2024) might cite either the 2020-2021 data set, 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.

References:

**Dataset selected:**

National Center for Educational Statistics. (2022). *School Neighborhood Policy Estimates - 2018-2019*

(1.1) [Dataset]. *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](https://d.docs.live.net/cc19913528a6ad64/Documents/%20https:/doi.org/10.3102/01623737231217683)

1. Repository Profile: Kaggle

For assignment 2, I chose to review the [2018-2019 School Neighborhood Poverty Estimates](https://catalog.data.gov/dataset/school-neighborhood-poverty-estimates-current-c7e05) dataset, published by the National Center for Educational Statistics and retrieved from data.gov. For the repository profile, I am looking at Kaggle, which is an open data repository but also describes itself as “the world’s largest data science and machine learning community”; it appears to be a plausible location for the government data set that is still different from where I initially retrieved it. Kaggle has more of a community focus, allowing data scientists and those new to the field to explore a wide range of datasets and to work with those datasets directly on the platform if they wish. 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 so easy to explore data files within the web browser and to “play” with them in the Jupyter notebook environment.

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”](https://www.kaggle.com/docs/datasets) 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,”](https://www.kaggle.com/datasets/raphaelmanayon/temperature-and-ice-cream-sales) or [“The Complete Pokemon Dataset”](https://www.kaggle.com/rounakbanik/pokemon)) might also have a better chance of trending on the site. Some examples of content categories for [Kaggle datasets](https://www.kaggle.com/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.

My understanding is that per the OAIS reference model, the contents of a Dissemination Information Package or DIP should be ready to be processed using the designated access software once downloaded by the user. I think the dataset files that are available for direct download on Kaggle fit this definition. They should already include at least some basic metadata if a readme file is not provided alongside them. As with many of the other specifications for dataset creation on Kaggle, a lot of freedom is left up to the creator here and it will vary from dataset to dataset.

1. Recommended data citation
2. Considerations for long-term preservation
3. Copyright license statement
4. Human subject considerations