USING DATA FOR GOOD: A MATTER OF GEOGRAPHY

Andee Rubin
TERC
Andee_Rubin@terc.edu

The purpose of this research is to learn how youth aged 11–14 understand highly aggregated data about social and economic conditions, especially related to the United Nations Sustainable Development Goals. We designed an after-school program that introduces students to data broadly in the context of "using data for social good." Using CODAP, an educational data visualization tool, students explore data about countries' health and education indicators. We observe that youth are highly engaged with these data yet sometimes struggle to make sense of the aggregate values that hide variability within countries. Using examples from student projects, this paper reports preliminary findings about how youth analyze data aggregated by governmental entities such as countries, states, or cities.

INTRODUCTION AND RELATED WORK

Using publicly available governmental data in statistics education has become more popular in recent years with the rise of the open data movement and multiple portals at many levels of government. The United States maintains a site called data.gov (U.S. General Services Administration, n.d.) that provides a uniform interface to data from most U.S. states, a group of U.S. cities and counties, over 50 international countries, and over 150 international regions; as of May 2022, it provided access to almost 350,000 datasets. The OECD (Organisation of Economic Co-operation and Development) has an "open government data" philosophy that "promotes transparency and accountability … by making government data available to all" (OECD, n.d., first paragraph). In general, it has been assumed that making such data accessible will lead to increased transparency and people's ability to discover patterns and trends that they may use to advocate for causes they deem important.

No one would argue against the basic premise of open government data, but effective use of such data in educational settings requires examining how youth are able to make meaning of such data, what kinds of supports can facilitate their understanding, and what special considerations might be relevant because of the structure and content of these data sets. The ProCivicStat project, a multi-national collaboration that advocated for increased inclusion of civic statistics in schools, pointed out in its Call for Action "the multivariate, dynamic and aggregated nature of social phenomena" (ProCivicStat Partners, 2018, p. 3). The same report further noted that the social context of civic statistics makes interpreting them particularly complex, requiring knowledge of the vast interrelated web of correlates, causal factors, and outcomes.

Other projects that involved civic data have worked with census microdata (e.g., Erickson, 2012; Louie et al., 2021). However, using census microdata is quite different from working with quantities such as "average income" or "percent of families in poverty." With census microdata, the case is a person, not a country, state, or city. Students can imagine a person and invent a story about them, based on data about their household, education, income, etc. It is more difficult for students to get an image of a country, based on its median income, average household size, and average educational level because there is so much variability hidden in aggregate values. Some of these difficulties are described below.

THE CONTEXT: NETAPP DATA EXPLORERS

The research reported here was carried out in the context of the NetApp Data Explorers project, funded by NetApp (https://www.netapp.com/), a cloud storage company. The project developed an out-of-school program that introduces students aged 11 to 14 to data in the context of "using data for social good" (Dorsey et al., 2022). Data Explorers focuses on the 17 United Nations Sustainable Development Goals (SDGs), which present a "blueprint to achieve a better and more sustainable future for all" (United Nations, n.d., first paragraph) and introduces participants to the data the UN has collected to track progress in the SDGs. Using the educational data visualization tool, CODAP (Common Online Data Analysis Platform; The Concord Consortium, 2014), students explore health and education indicators from 195 countries that belong to the United Nations (n.d.). The attributes include life

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

expectancy (overall, for males, and for females), average years of school attended (overall, for males, and for females), teen birthrate, population in millions, percent of the population that is urban, medical doctors per 100,000 people, and percent of the population who use the internet.

Students then analyze more local data (county-level data in the United States and similar administrative structures in other countries), focusing on health and education. In the United States, the attributes include a set of health outcomes (e.g., life expectancy), health factors (e.g., percent of people who smoke and medical doctors per 100,000 people), social and economic factors (e.g., median income and high school graduation rate), and racial demographic data.

In the final step in the Data Explorers program, teams of students carry out a project, "Digging Deeper," to dig into an attribute they find particularly interesting. They then create a presentation describing their discoveries and issue a call to action, either to find out more about the situation, take action to improve it, or increase awareness around it.

CODAP has several affordances that made it a good choice for Data Explorers. In addition to a simple way to create graphs, CODAP includes a map capability that integrates seamlessly with graphs. Users can see the relative values of attributes by country by dragging the attribute onto a map. All graphs and maps connected to a dataset are linked; if a point representing a country or county is highlighted in one representation, it is similarly highlighted in the others. Regional patterns that would not have been discernable in a table or graph can become obvious when attributes are displayed on a map.

RESEARCH FOCUS AND SUBJECTS

The research reported on here was based on a pilot implementation of the Data Explorers program in the United States as an after-school program at a middle school. The author taught the pilot via Zoom with assistance from several teachers at the school. The program involved approximately 20 students who worked on projects in teams of three to five people. Data consisted primarily of observations of the students and their final projects. Observations were in the form of written notes, triangulated with recordings of the Zoom sessions; data was stored on a password-protected Google drive.

The research was exploratory, with the basic question: "How do youth aged 11 to 14 make meaning of highly aggregated civic data using an educational visualization tool?" Data analysis consisted of informal qualitative coding around the themes of understanding of rates, spatial aggregation, and covariation.

REASONING ABOUT SPATIALLY AGGREGATED DATA

One common characteristic of civic data is that it is reported as a rate, either as a percentage or, quite often, as a quantity "per X people." Students in the Data Explorers program had some grasp of percentages but had rarely encountered quantities of this latter form. Several students were interested in the attribute, "number of doctors per 100,000 people," reported by the UN at the country level. Even though most students understood the difference between "number of doctors in a country" and "number of doctors per 100,000 people," some students struggled to imagine how the doctors and people were distributed around the country.

The final project of one student group illustrated how such spatially aggregated measures can be difficult to reason about. This group hypothesized that the availability of doctors might have something to do with the area of a country. One of the members of the group had talked in depth about how doctors are more likely concentrated in cities and why, even in a country with a relatively large number of doctors available per 100,000 people, there might be a need for more doctors in rural areas. Based on this reasoning, the group chose to compare the number of doctors per 100,000 people in a very small country (Qatar) with the same measure in a much larger country (Australia). Their hypothesis was that, because the population is more widely scattered in Australia, the number of doctors per 100,000 people would need to be larger to cover the far-flung population centers, i.e., that this measure would be larger in Australia than in Qatar. However, it turned out that in Qatar, there are 77 doctors per 100,000 people, whereas in Australia, there are only 32. In their final project, the students wrote about how the data did not turn out the way they expected and how they now had a new hypothesis: "Our test ended with the answer that it [area] does not affect availability. Our thought was that it wasn't the size of the country but rather the economic value of the country."

COVARIATION

Most of the final projects from the Data Explorers pilot focused on finding a relationship between two attributes. One of the most compelling was from a group who were interested in causes and effects of teenage pregnancy. They investigated the relationship between average educational level in a country (average years of schooling) and the teen birth rate (reported as number of births for every 100,000 women between the ages of 15 and 19). Their key slide (Figure 1) showed that relationship in a scatterplot. Their commentary displayed rather sophisticated statistical reasoning: "While there is not a perfect correlation, due to various other factors, typically countries with less years of schooling have higher teen birth rates."

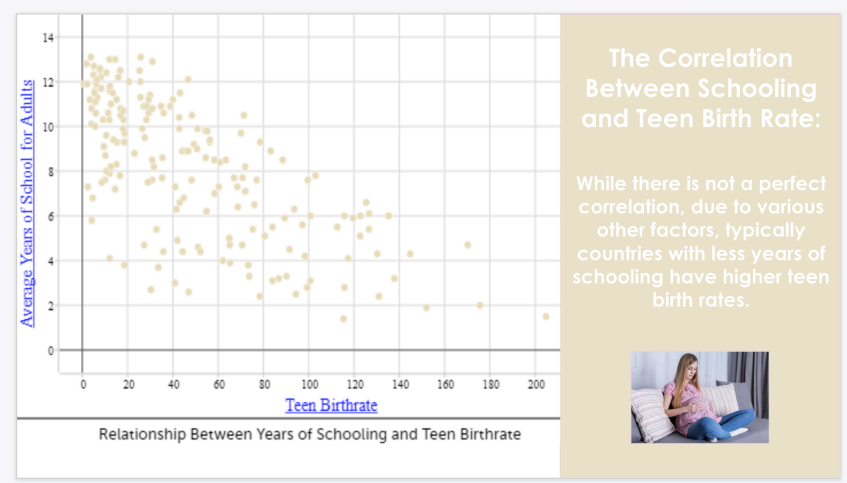

Figure 1. Student graph showing covariation between schooling and teen birth rate

In a creative use of maps, they also visualized the same relationship by plotting each variable separately on a world map and noting that, in their words, "With a few exceptions, these maps are almost opposites. As years of schooling goes up, teen birth rate goes down and vice versa." This juxtaposition is a persuasive example of the power of using maps in the analysis of aggregated civic data: not only do the maps illustrate the relationship in the graph, but they also show which areas of the world have high teen birth rates/low educational levels and vice versa (see Figure 2).

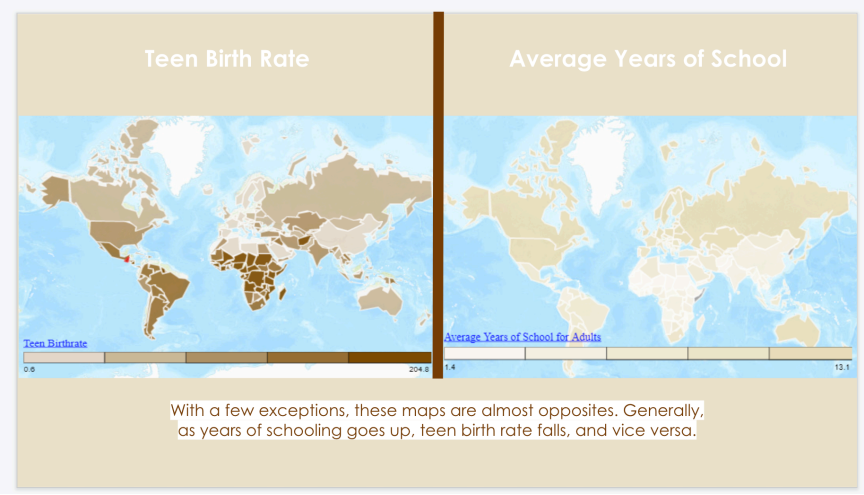

Figure 2. Maps showing the relationship between schooling and teen birth rate

These students were more fortunate than many of their peers: there was an easily discernable and understandable relationship between the attributes they chose. One group of students investigated

the county-level attribute "% insufficient sleep," the percent of people in each county who got an average of less than 7 hours of sleep a night. They tried valiantly to find a relationship, but the best they could come up with were two rather weak relationships—one with the availability of doctors and one with median income. In the youths' words, "We can concur that in general if you have more doctors, people sleep more" but they also warned that "there are some outliers." It is worth noting here that even understanding this variable was challenging for some students because a high value on "% insufficient sleep" actually means that people slept less on average.

PEDAGOGICAL AND RESEARCH IMPLICATIONS

Civic data is here to stay, and hopefully, increasing numbers of students will have access to such data combined with tools like CODAP. This pilot study suggests several pedagogical steps.

- *Attention to measurement and rates:* The processes by which civic data are produced are complex. Students need to learn how to interpret attributes such as "% insufficient sleep" that are generated by a process of transforming and combining many individual reports of "hours slept." Even when students understand rates, they are curious about the absolute number of people or households that comprise a given percentage; it is beneficial to provide an opportunity for them to engage with both the rates and the counts.
- *Explicit discussion of spatial aggregation:* Students have an intuitive understanding of how data aggregated over space conceals variability. They know that any country is a mix of cities, rural areas, and even national parks; they know that there are poorer areas and richer ones, and that some communities have more racial diversity than others. How do we recognize and honor this knowledge while still helping students find meaning in the aggregated data?
- *Support inquiries into civic data other than correlation:* Without explicit examples of the kinds of useful knowledge that might be gleaned from civic data, most youths' instincts are to look for a correlation, especially if they think there might be a cause-and-effect relationship. Such clear correlations, however, are unlikely to emerge given the number of inter-related factors in civic data. What other models of the analysis of civic data can we share with youth?

Although these three points are pedagogical recommendations, they also suggest research directions from the questions they raise. What are the best ways to support students' understanding of rates, especially where large numbers (such as populations in the millions) are concerned? What do students understand about spatial variability in general, beyond the anecdotal evidence included in this paper? How might we superimpose maps at different aggregation levels to help students gain more insight into data?

CONCLUSION

Taking seriously the increased availability of civic data, its particular statistical characteristics, and youths' interest in social issues, it is important to look deeply at the challenges and opportunities such data might present and to study students' work to discern how we can best support them in their analyses. Both research and pedagogical innovations will be necessary to make the best use of such data.

REFERENCES

Dorsey, C., Rubin, A., & Mann, M. (2022). *Preparing youth for the data-filled future.* The Concord Consortium. https://www.netapp.com/pdf.html?item=/media/70713-Data_Explorers_eBook.pdf

Erickson, T. (Ed.). (2012). *Signs of change: History revealed in U.S. census data.* eeps Media.

Louie, J., Roy, S., Chance, B., Stiles, J., & Fagan, E. (2021). Promoting interest and skills in statistical and multivariable thinking with social justice data investigations. In R. Helenius & E. Falck (Eds.), *Statistics education in the era of data science. Proceedings of the Satellite conference of the International Association for Statistical Education*. ISI/IASE. https://doi.org/10.52041/iase.ohylj

Organisation for Economic Cooperation and Development. (n.d.) *Open government data*. https://www.oecd.org/digital/digital-government/open-government-data.htm

ProCivicStat Partners. (2018): *Engaging civic statistics: A call for action and recommendations.* https://iase-web.org/islp/pcs/documents/ProCivicStat_Report.pdf?1545118071

The Concord Consortium. (2014). Common Online Data Analysis Platform (Version 2.0). [Computer software]. https://codap.concord.org/app/static/dg/en/cert/index.html

United Nations. (n.d.) *Take action for the sustainable development goals*. Sustainable development goals. https://www.un.org/sustainabledevelopment/sustainable-development-goals/

U.S. General Services Administration. (n.d.). *Open government*. Data.gov. https://data.gov/open-gov/