SECONDARY DATA IN THE SECONDARY DATA SCIENCE AND STATISTICS CLASSROOM

Anna Bargagliotti and Robert Gould
Loyola Marymount University, Los Angeles, California, USA
University of California, Los Angeles, (UCLA) California, USA
rgould@stat.ucla.edu

The complexity of modern data raises the stakes of what it means to be data literate, challenging statistics educators who wish to engage students with real data. Two authors of the revised Guidelines for Assessment in Instruction and Statistics Education Pre-K–12 report discuss the use of secondary data, illustrating these challenges with an exploration originating from a What's Going On in This Graph exercise. In particular, the authors discuss two important issues when using secondary data for Pre-K–12 students (ages 5–18): the importance of using interrogative questions to determine the provenance of data in order to assess its suitability to the task at hand as well as to consider ethical concerns, and the need to "tame" the data if the data collection scheme is advanced.

INTRODUCTION

The emergence of data science has pushed a demand for an increase in what is sometimes called data literacy (Levitt, 2019) or data acumen (Nae et al., 2018) in K–12 and undergraduate education. We conceive of data acumen as a broad collection of skills, habits of mind, attitudes, and knowledge centered around questioning, collecting, and analyzing data, as well as understanding the cultural, social, and ethical implications of these activities. The recently revised *Guidelines for Assessment in Instruction and Statistics Education Pre-K–12* (Bargagliotti et al., 2020) report, hereafter referred to as GAISE II, sees the discipline of statistics as foundational to data science, particularly at the K–12 level, and accordingly updates the original GAISE pre-K–12 report (Franklin et al., 2007) by providing guidelines for integrating data science concepts into the curriculum. The ubiquity and accessibility of secondary data sources, combined with the need for data scientists who can use these sources to answer important questions, is an important reason the GAISE II report encourages educators to include secondary data in the classroom. We address two issues that are raised by bringing secondary data into the classroom: understanding the origins of the data and dealing with sampling schemes that may be too complex for introductory students.

We consider data to be "secondary" if it is collected by people other than the students. The importance of secondary data in the classroom has been recognized for quite some time, with calls for including secondary data extending back at least to the 1990's. (For example, see Hoerl & Snee, 1995.) More recently, the growing prevalence of public data repositories, data "streams," and the ability to digitize objects and use them as data means that secondary data skills are more important than ever. Secondary data is likely the most common type of data students will encounter outside of the classroom. When we consider data science education, as opposed to statistics education, the role of secondary data becomes particularly important. Wise (2020) defines a data scientist as "one who must build a bridge between ill-defined questions and unstructured messy data that may (or may not) be fit to address them, by assessing, cleaning, organizing, integrating and visualizing data, selecting suitable algorithms and code to be composed into a reproducible workflow, and communicating appropriate inferences in an ethical manner" (p. 166). Thus, teaching students how to evaluate secondary data is essential.

Despite the importance of secondary data, it can be challenging for students to engage in statistical investigations using real data. The first two steps of such an investigation, according to GAISE II, are for students to pose strong statistical investigative questions and then to "consider" data for its appropriateness to address those questions. (Arnold & Franklin, 2021). This consideration phase includes interrogative questions (such as Who collected the data? Who funded the data collection? How were the data collected?) that, once answered, might provide the instructor with an overwhelming level of complexity.

In this paper we discuss one resource, What's Going on in this Graph (WGOITG), that is recommended by GAISE II. Although this resource provides rich materials for beginning and intermediate students, (levels A and B in the GAISE II lexicon), advanced students will, when applying their interrogative skills, discover complexities that are common to real secondary data. After

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

introducing a WGOITG example, the paper discusses some of the challenges classroom teachers will face when confronting these complexities.

SECONDARY DATA IN THE CLASSROOM

One resource that brings rich and engaging secondary data into the classroom is the weekly publication, What's Going On in This Graph (WGOITG), a collaboration between the New York Times (NYT) and the American Statistical Association (ASA). WGOITG invites students to engage in understanding and interpreting a multivariate data visualization and is appropriate for level A and B students (beginning and intermediate, respectively). More advanced students, such as those at GAISE level C (advanced level), should go deeper and, when possible, engage with the original data.

For example, consider the WGOITG post on Sept 15, 2021 (The Learning Network, 2021) providing the graphic in Figure 1. These graphs presume a statistical investigation that compares how our time use has changed before and during the early stages of the COVID-19 pandemic. Students at all levels can begin by posing statistical investigative questions that can be answered with these graphs, but level C students should go deeper by examining whether the data are appropriate for the posed questions.

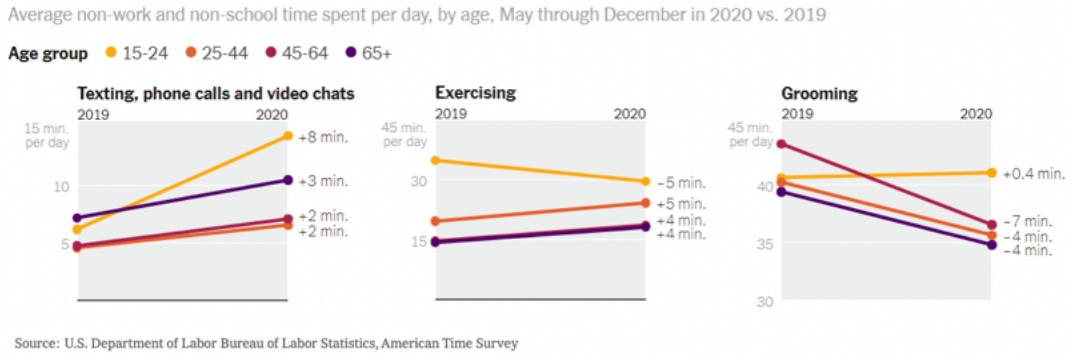

Figure 1. WGOITG on time spent per day on texting, phone calls, and video chats; exercising; and grooming (The Learning Network, 2021). The data visualizations compare time spent for 2019 with time spent for 2020. (Reprinted with permission.)

As a starting point, students pose the interrogative questions of who, what, when, where, and how the data were collected, questions that aren't answered by the graphics themselves in this example. Answering the "who collected the data" question requires students to go to the source of the data, the American Time Use Survey (ATUS), collected by the U.S. Department of Labor and Bureau of Labor Statistics (2022). Students can examine the web pages and discover the original source of the data. Why were the data collected? (Data were collected to understand the manner in which people in the United States spend their time to better understand the "hidden economy.") What data were collected and how? (American citizens aged 15 years and older kept diaries and underwent substantial interviews in order to determine which activities they did and how long they did them in the 24 hours prior to the survey interview.)

Students will realize that they can go beyond the investigations presented in WGOITG and pose additional statistical investigative questions that might be of personal interest, such as questions concerning sleep, recreation, work, and more. Students will also see the care, complexity, and plethora of details involved in carefully documenting a data set intended to answer research questions for a broad public. When reading about the sampling scheme, they will realize that although the participants were chosen "randomly," they were not chosen using the sampling schemes usually discussed in GAISE level C statistics but, instead, data were collected using a fairly complex weighted sampling scheme.

CHALLENGES

The ATUS data provides a context that many students will find interesting and has the potential to provoke meaningful and relevant investigative questions that students can address with tools and

analytic approaches accessible to GAISE II level C. Despite this promise, the ATUS "raw" data set is challenging for teachers to bring into the classroom and for students to engage with. Two prominent challenges are the wealth of documentation and the sampling scheme used to collect the data. These challenges are not unique to ATUS; government data sets are often extensively documented and use sampling designs beyond those taught in introductory classrooms.

The ATUS website provides links to ten separate documentation files, including data dictionaries, lists of variables, detailed descriptions of sampling procedures, and more. A teacher must spend considerable time deciding which of these documents would help students achieve the learning goals related to understanding the data provenance and which parts of the documentation could be skipped over. Even then, some documents are quite long, and the teacher will likely need to draw the students' attention to particular sections. The task sounds daunting, but the complexity itself should be part of the take-home lesson: important data requires great care to document, and the analyst must be prepared to invest time to understand the data. This is the foundation of the notion of *reproducibility*, a foundational topic in science and data science that is essential for anyone seeking to understand whether the data are suitable for their investigative questions.

The sampling scheme provides another challenge. ATUS is not collected using the simple random sampling scheme commonly taught in most introductory statistics courses. Instead, the scheme uses multiple layers with weights. Students examining the racial distributions in ATUS may notice, for instance, that it does not reflect the actual population distribution (due to oversampling of some groups within the population). This disparity might cause students to distrust the data. Indeed, if the sampling scheme is not accounted for by an inferential procedure, the results *should* be distrusted. This complexity is quite common with large-scale survey data and provides a dilemma for the educator who wishes to prepare students for "real" data. How much do we "soften the blow" for the student? One could re-sample the data to better match population characteristics, but this sacrifices some of the "reality" of the data and perhaps disempowers the students. Our belief is that this is one area the statistics and data science education community should study further: at which stage of their education should students learn to approach "raw" data in its original form, particularly for data sets that are commonly used and readily available? When, in a students' education, should they learn to handle data collected through weighted sampling schemes, and what should they be taught?

When considering bringing secondary data into the classroom, teachers either need to spend time examining the documentation at length and preparing scaffolded, targeted lesson plans for students to engage in data cleaning and wrangling, or they need to find access to interesting secondary data sets that have already been prepared for classroom use. Secondary data have great potential to motivate and engage students; somewhere, there is a data set that matches a student's interests. However, that data set may require considerable work before it can be brought into the classroom. As such, the statistics and data science education community must increase efforts to prepare rich, complex, and interesting data sets at a variety of levels of "cleanness." Without this assistance, most classroom teachers will not have time to use realistic secondary data in their classes.

CONCLUSION

The need to bring real secondary data into the classroom generates a need to prepare that data so that it is suitable for addressing learning objectives. Kim et al. (2018) discuss the continuum from "raw" to "clean," and note that, for some objectives (such as learning to clean data), perfectly clean is too much and raw is not enough. The happy medium is "tame" data (their terminology) that is just clean enough to fill its intended educational role. However, with many available secondary data sets, finding the correct level of tame is difficult and requires technological skills that not all teachers have.

One characteristic of real data is that it may be large. The size of the data set, however, is often less of a pedagogical challenge than a technological challenge. Some software may restrict the size of the data sets that can be uploaded (particularly if the software is cloud-based), and some data sets are so large that even the hardware may be inadequate. (The ATUS-derived data discussed in this paper has roughly 18,000 rows, which is not large for many software packages.) Instead, the pedagogical challenge lies in the number of columns; the ATUS provides several hundred variables (the WGOITG example merges several of these together in its graphics). Students (and others) are often overwhelmed by data sets with large numbers of columns. According to Erickson (2021), this feeling of being "awash in data" is a defining characteristic of a data science educational experience. Students must learn to

confront this sensation by fine-tuning their analytic objectives and, in some cases, further taming the data by applying "data moves" (Erickson & Chen, 2021). For example, after students have formulated statistical investigative questions that can be answered with the ATUS data, they can apply a data move to extract a subset of the data that contains only the variables that are relevant to their investigation.

This paper presents a timely and relevant data set that was used recently in the news and that can be tamed for the classroom. The complexities of the ATUS data are far from unique and represent common challenges to statistics and data science educators. Arnold et al. (2022) use this data set to illustrate the process of learning to prepare data for analysis. As a template to assist those who wish to bring complex data into their classroom, a sample lesson plan for implementing an investigative activity around this WGOITG is provided by the GAISE II team (Arnold et al., 2022). This sample lesson plan includes links to the original data as well as a link to a relatively "tame" data set based on a random sample of 4,000 cases from the data set and about 70 variables (https://www.amstat.org/docs/default-source/amstat-documents/gaise-ii-data-sets.zip).

As the emphasis of acquiring data literacy/acumen in education continues to grow, we can strive to define and investigate student learning outcomes related to cleaning and structuring data with students. Overall, there is a need for finding better ways to bring level-appropriate and tame secondary data into the classroom. There is also a need for a collection of data sets derived from open data sources that are tailored to the GAISE levels A, B and C. As demonstrated above, the WGOITG NYT and ASA collaboration can provide a springboard for discussion about using complex secondary data in the classroom.

REFERENCES

Arnold, P., Bargagliotti, A., Franklin, C., & Gould, R. (2022). Bringing complex data into the classroom. *Harvard Data Science Review, 4*(3). https://doi.org/10.1162/99608f92.4ec90534

Arnold, P., & Franklin, C. (2021). What makes a good statistical question? *Journal of Statistics and Data Science Education, 29*(3)*,* 122–130. https://doi.org/10.1080/26939169.2021.1877582

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 guideline for assessment and instruction in statistics education II (GAISE II). A framework for statistics and data science education*. American Statistical Association; National Council of Teachers of Mathematics. https://www.amstat.org/docs/default-source/amstat-documents/gaiseiiprek-12_full.pdf

Erickson, T. (2021). *Awash in data.* eeps media. https://codap.xyz/awash/

Erickson, T. & Chen, E. (2021). Introducing data science with data moves and CODAP. *Teaching Statistics*, *43*(S1), S124-S132. https://doi.org/10.1111/test.12240

Franklin, C., Kader, G., Mewborn, D., Moreno, J. U., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A Pre-K-12 curriculum framework.* American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/gaiseprek-12_full.pdf

Hoerl, R., & Snee, R. D. (1995). *Redesigning the introductory statistics course*. University of Wisconsin-Madison Technical Reports. http://digital.library.wisc.edu/1793/69159

Kim, A. Y., Ismay, C., & Chunn, J. (2018). The fivethirtyeight R package: "Tame data" principles for introductory statistics and data science courses. *Technology Innovations in Statistics Education*, *11*(1). https://doi.org/10.5070/T5111035892

Levitt, S. D. (2019, October 2). *American's math curriculum doesn't add up* (Ep. 391) [Audio podcast episode]. *Freakonomics*. https://freakonomics.com/podcast/math-curriculum/

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options.* The National Academies Press. https://doi.org/10.17226/25104

The Learning Network. (2021, September 15). *What's Going On in This Graph? September 15, 2021: How did the pandemic change how we spend our free time?* The New York Times. https://www.nytimes.com/2021/09/09/learning/whats-going-on-in-this-graph-sept-15-2021.html

United States Bureau of Labor Statistics. (2022). *American time use survey*. https://www.bls.gov/tus/

Wise, A.F. (2020). Educating data scientists and data literate citizens for a new generation of data, *Journal of the Learning Sciences, 29* (1),165-181. https://doi.org/10.1080/10508406.2019.1705678