VISUALIZING MULTIVARIATE DATA: GRAPHS THAT TELL STORIES

Joachim Engel[1], Pedro Campos[2], James Nicholson[3], Jim Ridgway[3], and Sónia Teixeira[2]
[1]Ludwigsburg University of Education, Germany
[2]University of Porto, Portugal
[3]University of Durham, UK
engel@ph-ludwigsburg.de

Important statistical ideas can be introduced via visualizations without heavy mathematics, hence can become accessible to a broader citizenry. Along a few selected examples, from historical to modern, with technology-based data visualizations, we highlight the potential of data visualizations to enhance students' capacity to reason with complex data and discuss the role of visualization as a tool to strengthen civic participation in democracy.

BACKGROUND

Visual representations are a central means of conveying information, illuminating facts, supporting the user in recognizing patterns and gaining insights into difficult concepts (see, e.g., Chambers et al., 1983; Chance et al., 2007; Tishkovskaja & Lancaster, 2012). A Graph can provide a compelling approach to statistical thinking that focuses on important concepts rather than formal mathematics and procedures (Biehler, 1993; Nolan & Perrett, 2016). Graphical methods provide powerful diagnostic tools for confirming assumptions, or, when assumptions are not met for suggesting corrective actions. Therefore, creating meaningful data visualizations to communicate information is an important skill in its own right. It is an important mean of informing citizens about governance and presenting evidence about the state of the world in order to raise awareness for injustices and structural social inequalities or burning problems like global warming or demographic change.The simulation to illustrate the outbreak of COVID-19 and the effect of social distancing, published by the Washington Post, is another striking example (see https://www.washingtonpost.com/graphics/2020/world/corona-simulator/).

While statistical graphics emerged with the earliest attempts to analyze data (Beniger & Robyn, 1978), and much has been written on best practices for data visualization (e.g., Tufte, 1992; Cleveland, 1994; Yau, 2011; Wainer, 1997), with the rise of data science and an increasingly data-infused society, the teaching and understanding of effective data visualizations has become even more crucial.

Fortunately, technology today provides tools for data visualization (DV) that offer the potential to explore rich sources of information without requiring deep mathematical knowledge. With interactive data visualizations (IDV), conceptual understanding can go even one step further by using technology in diagrams to retrieve and interactively change more detailed information, e.g. what data is displayed and how it is displayed. Suitable visualizations can make a significant contribution to understanding complex relationships. Data on many "burning issues" (e.g., climate change, public health, migration, economic justice) are often communicated via rich, novel data visualizations. Vivid democracies need well-informed citizens who can understand important social issues, discuss them and contribute to public decision-making. Citizens need to be able to develop skills of effective data communication to be engaged in public decision processes (ProCivicStat Partners, 2018, Cukier, 2011). Rich data sets are accessible in abundance. There is a wealth of data collected on a large scale by governmental and non-governmental agencies that can inform about the state of the world. The ProCivicStat project (ProCivicStat 2018, see http://iase-web.org/islp/pcs/) developed an extended concept of statistical literacy called Civic Statistics, which focuses on the exploration and sense making from data about the social and economic well-being of humans and the realization of civil rights. Most data about society are influenced by a multiplicity of factors. Their exploration requires an understanding of multivariate phenomena. For example, to study the impact of factors inside and outside of school on educational success requires to assess the relative impact of social class, gender and ethnicity, and the links between them (Ridgway 2016). Traditional print media are increasingly using interactive and dynamic visualizations as part of data journalism, which are much broader and more sophisticated compared to the limited scope of graphs, histograms and tables used in introductory statistical units at schools and universities where statistical graphs are often limited to

In: P Arnold (Ed.), *New Skills in the Changing World of Statistics Education*
Proceedings of the Roundtable conference of the International Association for Statistical Education (IASE),
July 2020, Online. Voorborg, The Netherlands. ©2020 ISI/IASE iase-web.org/Conference_Proceedings.php

simple one-or two-dimensional representations. The ability to understand statistical graphs is one of the main competencies of statistical education and statistical literacy (Gal 2002). In its recommendations for statistics teaching (Guidelines for Assessment and Instruction in Statistics Education, GAISE 2016), the American Statistical Association emphasizes multivariate thinking and the importance of creating and interpreting graphics as the first step in any data analysis

SOME HISTORIC DATA VISUALIZATIONS

The idea of using images to represent complex quantitative information has been around for centuries, from maps and graphics in the 17th century to the invention of the pie chart in the early 19th century. Data visualizations have been used to promote social change. We highlight briefly four historic milestone examples and recommend the reader to explore the associated graphics by following the provided links.

1.) The year 1786 is considered the founding year of modern data visualization when William Playfair published his “Commercial and Political Atlas”, designed to make complex evidence accessible to a wide audience (Playfair 1786, 2005). Playfair wanted to enlighten in times when statistics were not published but rather kept a top secret by the state. With his atlas, Playfair made the state of society tangible for broad audiences. (see https://archive.org/details/PLAYFAIRWilliam1801TheCommercialandPoliticalAtlas/page/n35/mode/2up)

2.) A few decades after Playfair, Charles Minard mapped Napoleon's invasion of Russia and created one of the most frequently cited examples of statistical graphics (see Figure 1). The card contains a lot of information. The variables include the location and size of the army. The graphic also catalogs the date and temperature. The direction of the army is given by the color. In addition, cities and rivers are used as reference markers along the path. Napoleon's catastrophic Russian campaign 1812-1813 can be seen, two years before he experienced his proverbial Waterloo. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates. The light area or line on top shows the number of Napoleons troops when attacking, the black area or line shows that when retreating. The size indicates the number of troops, and the viewer recognizes where auxiliary troops have been added. In 1812 Napoleon started with 422,000 men in Kovno, of whom only 100,000 arrived in Moscow. The course of the light and dark lines shows the route the troops had chosen when attacking and when retreating. You can also see below the temperature curve and the very low degrees. It can also be seen, for example, that the decision to choose the route across the Berezina river halved the number of troops during the retreat.

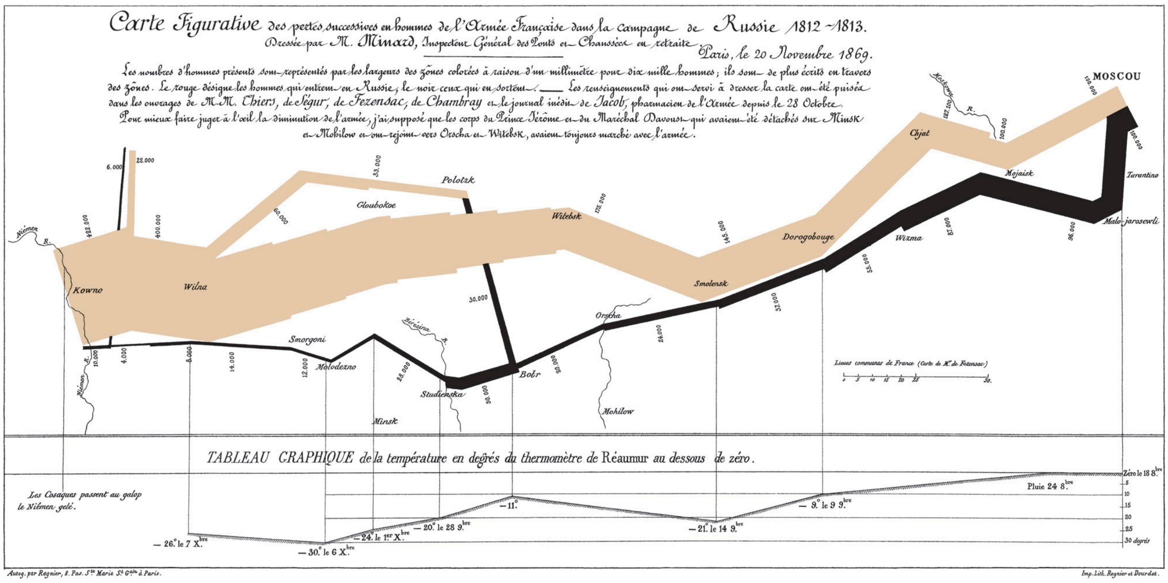

Figure 1. Graphical display of Napoleon’s Russian campaign, 1812-1813 (source: https://commons.wikimedia.org/wiki/File:Minard.png)

3.) Another famous example of displaying information to inform action is illustrated vividly by the work of Florence Nightingale (1820-1910) and her 'coxcomb' plots of deaths (see https://commons.wikimedia.org/wiki/File:Nightingale-mortality.jpg) during the Crimean war attributable to battle and to diseases encountered in military hospitals. This startling data was used by Nightingale to push forward major changes in hospital hygiene and sanitation in general.

4.) Since William Playfair, graphics have often been used to make social data accessible to a wide audience, such as graphical representations for the demonstration of social inequalities by the Austrian national economist Otto Neurath (Neurath, 2010). Isotypes (International System of Typographic Picture Education) were designed in Vienna as a universal visual language (https://isotyperevisited.org/documents/index.html), created in the 1920s by Neurath's group to promote civic statistics, and is grounded in a rich philosophy about the nature of knowledge, and social justice.

MODERN APPROACHES TO VISUALIZING STATISTICS

Data visualizations can look back on a long history. However, it is technology and the possibilities of modern computer graphics that lit the fire under data visualizations at the beginning of the 21st century. Computers have made it possible to process large amounts of data at lightning speed. Today, data visualization has become a rapidly evolving mix of science and art. The major data providers offer powerful visualizations in the hope of making their data more accessible. Important political goals such as the United Nations' Millennium Development Goals (MDG) are presented in the form of an interactive dashboard (e.g. http://datatopics.worldbank.org/mdgs/)to promote public engagement. Documenting the explosion of exciting data visualizations facilitated by modern computers is a task beyond the scope of this short paper (but see Ridgway et al., 2017). Here we highlight just a few examples. A variety of tools have been designed to facilitate data visualization. These can be classified as:

- ***Tools for displaying specific data sets*** – illustrative examples include *Arctic Ice (see below)* or reference to the prediction of mortality (http://flowingdata.com/2015/09/23/years-you-have-left-to-live-probably) and demographic change (e.g. https://www.worldlifeexpectancy.com/ brazil-population-pyramid) which underpin the fundamental idea of statistics: individual events (when will *you* die?) can only predicted with very limited accuracy, but events aggregated across individuals (distribution of mortality) can be predicted with high accuracy.
- ***Tools for facilitating the exploration of specific data sets*** - examples include: *Gapminder* (https://www.gapminder.org/*)*; *Our World in Data* (OWID, https://ourworldindata.org); the *Constituency Explorer*; and *OECD's Better Life Index (see below).* These can facilitate statistical thinking in constrained domains
- ***Software packages with good functionality for visualizing data*** - examples include, *RAW*, R *Datawrapper; Tableau*; *JMP*; *CODAP*; *Tinkerplots*; and *Fathom.* An illustration of the scale of recent graphical innovations can be found at Selva Prabhakan's website, where 'top 50 visualizations' (along with R code) are offered (see http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html)
- ***Programming tools that can be used to create visualizations*** - a key library is *D3.js*. These can support statistics classes where computing is an important element (see https://wattenberger.com/blog/d3).

INTERACTIVE VISUALIZATIONS

A key development has been the emergence of sophisticated ways to communicate visually, apparent in both video and print media. The emergence of interactive data visualization (IDV) has the potential to cause a revolution by democratizing statistics; explorations that would once have required considerable technical skills can now be performed easily by statistically naïve users. This is analogous to the introduction of statistics packages such as BMDP, SAS and SPSS in the 1960s, and offers similar opportunities (namely insights into data structures with little effort involved in data manipulation), and similar risks (namely explorations conducted without understanding, which lead to misinterpretation and bad decisions). We now need to attend not just to 'statistical results that permeate our daily lives' but to the whole process of drawing evidence-based conclusions.

Perhaps the most important demands on users of IDV relate to *dispositions*. Gal (2002) proposed two distinct dimensions of statistical literacy: one relating to knowledge elements (e.g., big ideas such as sampling, mathematical knowledge and context knowledge); the other relating to dispositions (e.g., a willingness to be critical and to ask questions, and having positive feelings about being able to understand situations with statistical elements). These dispositions are central to the new visual approach – without a disposition to have confidence in one's own ability to make sense of the complexities presented, and the willingness and ability to ask questions and seek answers within the IDV, the approach collapses to the static visualization of its first appearance.

EXAMPLE: ARCTIC SEA ICE VOLUME

Here, we show how free-standing DVs taken from the web can be used both as starting points for explorations of substantive issues, and as the focus for the introduction of important statistical ideas, see Figure 2 (http://www.climate-lab-book.ac.uk/files/2016/06/icevol.gif)

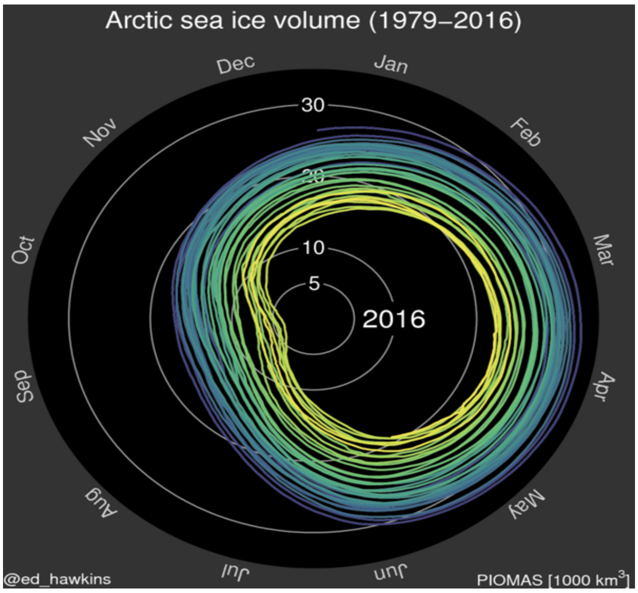

Figure 2. Arctic Sea Ice Volume as a function of time

Global warming is perhaps the greatest threat facing humanity today. This graphic comes from the Polar Science Center at the University of Washington; the observations are collected at the Unified Sea Ice Thickness Climate Data Record. The graphic can support the development of some key statistical ideas. It offers an opportunity for students to interpret graphical information presented in a way they might not have seen before. Plotting monthly data in a circular display reduces the problem of separating out seasonal changes from long-term changes. Students can simply be asked what they see: there are marked seasonal trends, a good deal of variability in year-on-year data, and a decline in the volume of sea ice over time. Is the world getting warmer? Students can be provoked to look for other evidence – is the Antarctic ice volume decreasing? What about glaciers around the globe? Or sea temperatures in different locations? These questions provoke reflections about sampling and representativeness.

- Q1: Which variables are visualized?
- Q2: Which trend do you perceive within each year?
- Q3: Which trend do you perceive across the years?
- Q4: Is our planet getting warmer?

Students can be asked about their predictions for the future – when will there be little or no Arctic ice in summer months? Can they offer a bounded estimate? This takes them into the realms of modelling, which can be addressed at different levels of sophistication. Is a simple linear model likely to apply? (probably not). What factors are likely to accelerate the change? To what extent do they believe that the phenomenon can be predicted with any degree of accuracy (past estimates have almost always underestimated the speed of change)?

EXAMPLE: THE OECD BETTER LIFE INDEX

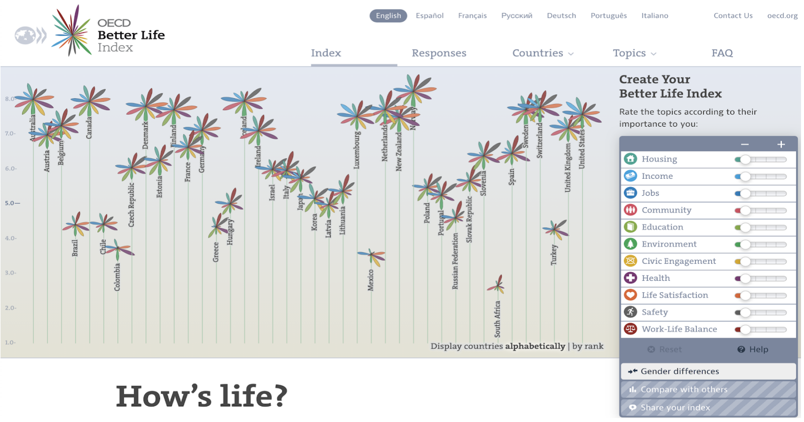

Figure 3. OECD's Better Life Index

'Quality of Life' is an important but problematic concept. The idea is simple: poverty, disease, awful neighbors affect the quality of life; Good health, a clean environment and safe community, all add to the quality of life. Measuring is more difficult – some concepts are hard to measure and people will value different things, and will probably change their ideas about quality of life over the course of their lives. The OECD have created a tool (see Figure 3) to compare different countries in terms of the quality of life offered from *your* point of view (see http://www.oecdbetterlifeindex.org/). You are invited to move sliders to reflect the balance of different factors relevant to your quality of life, and the display will rank countries in terms of how well they score against your criteria. The website allows users to compare their own region to their own country in terms of every indicator, and to make comparisons between countries. This software offers an excellent place to start discussions about the politics and pragmatics of measurement. OECD commissioned a report by some leading economists into the over-use of Gross Domestic Product (GDP) as a measure of the success of a country. Stiglitz *et al* (2018) show that GDP cannot tell us everything we need to know about economic performance, and that it is woefully inadequate as a measure of social progress. This activity provokes questions about what *you* value in life and questions the composition of indices such as GDP or the United Nations Human Development Index HDI to compare countries.

CRITICAL EVALUATION AND REFLECTION AROUND DATA VISUALISATION

Students need to be aware of ways in which data visualizations can be misleading, as well as the potential power of DV to illuminate complex topics. Tufte (1997) and Wainer (2000) pioneered work exposing the perils of poor (or deliberately misleading) DV. *Flowingdata* have a piece on why people make bad charts (see https://flowingdata.com/2018/06/28/why-people-make-bad-charts-and-what-to-do-when-it-happens/). The Financial Times (behind a pay wall) offers guidance on the science behind good data visualization, common mistakes, and how to lie with maps. The Economist, in an article entitled *mistakes, we've made a few* (see https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368) illustrates good design principles by presenting graphics published in the Economist that are misleading, alongside improved versions. Data can be downloaded, so these examples can form the basis for student activities. A richer source still is provided by the New York Times. In collaboration with the American Statistical Association. They publish *What's Going On in This Graph?* (see https://www.nytimes.com/column/whats-going-on-in-this-graph) on a weekly basis. There is a facility for students to discuss these graphs in a public forum, moderated by statisticians.

CONCLUSIONS

Increasingly, information is presented in visual forms, often via interactive displays. It follows that an important aspect of statistical literacy concerns the ability to work with, critique, and learn from visual displays. Citizens need to be critical consumers, and lies can be found in visualizations, as well as in text. Novel IDV continue to be created; learning to read and critique unfamiliar IDV has become an important life skill. However, we can make two stronger claims for skilled interpretation and use of IDV. DV can communicate information effectively in ways that calculations cannot: complex interactions between variables provide an example. Second is that IDV promote active engagement with evidence – users can explore their own conjectures and hypotheses – they are not passive recipients of others' stories. Disposition to engage is a key educational goal for civic statistics, and IDV provide a vehicle for inculcating the thrill of data exploration in students.

REFERENCES

Beniger, J. R. & Robyn, D. L. (1978). Quantitative Graphics in Statistics: A Brief History. *The American Statistician, 32*(1), 1-11.

Biehler, R. (1993). Software tools and mathematics education: The case of statistics. In C. Keitel & K. Ruthven (Eds.), *Learning from computers: Mathematics education and technology* (pp. 68-100). Berlin: Springer.

Chambers, J., Cleveland, W., Kleiner, B. & Tukey, P. (1983). *Graphical Methods for Data Analysis.* New York: Chapman & Hall.

Chance, B., Ben-Zvi, D., Garfield, J., and Medina, E. (2007). The Role of Technology in Improving Student Learning of Statistics. *Technology Innovations in Statistics Education*, 1(1). Retrieved from: *http://www.escholarship.org/uc/item/8sd2t4rr*

Cleveland, W. (1994). *The Elements of Graphing Data*. Summit, NJ: Hobart Press.

Cukier, J, (2011). Can Data Exploration help build Democracy? Crossroads 18(2):26-30

GAISE College Report ASA Revision Committee (2016). Guidelines for Assessment and Instruction in Statistics Education College Report 2016, http://www.amstat.org/education/gaise.

Gal I (2002). Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review* 70(1), 1–51.

Neurath, O. (2010). *From Hieroglyphics to Isotype: A Visual Autobiography*. London: Hyphen Press.

Nolan, D., & Perrett, J. (2016). Teaching and Learning Data Visualization: Ideas and Assignments. *American Statistician, 70*(3), 260-269. doi:10.1080/00031305.2015.1123651

Playfair, W. (1786, 2005). *The Commercial and Political Atlas and Statistical Breviary.* Cambridge University Press.

ProCivicStat Partners (2018). Engaging Civic Statistics: a Call for Action and Recommendations. A Product of the Procivicstat project. *http://IASE-web.org/ISLP/PCS*.

Ridgway, J. (2016). Implications of the Data Revolution for Statistics Education. *International Statistical Review*, 84(3), 528-549.

Ridgway, J., Nicholson, J., Campos, P. & Teixeira, S. (2017). Tools for Visualising Data: a Review. In A. Molnar (ed.), *Teaching Statistics in a Data Rich World*, *Proceedings of the IASE satellite conference*. Rabat, Morocco. Retreived from *http://iase-web.org/documents/papers/sat2017/IASE2017%20Satellite%20R16_RIDGWAY.pdf*

Svetlana Tishkovskaya & Gillian A. Lancaster (2012) Statistical Education in the 21st Century: A Review of Challenges, Teaching Innovations and Strategies for Reform. *Journal of Statistics Education*, 20:2, Retrieved from: DOI: 10.1080/10691898.2012.11889641

Tufte, E. R. (1992). *The Visual Display of Quantitative Information*. Creshire, CT: Graphics Press

Wainer, H. (1997). *Visual Revelations*. New York: Copernicus, Springer.

Yau, N. (2011). *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*. Hoboken, NJ: John Wiley & Sons.