TOOLS FOR VISUALISING DATA: A REVIEW

Jim Ridgway*, James Nicholson*, Pedro Campos** and Sónia Teixeira**
*Durham University, England
**University of Porto, Portugal
jim.ridgway@durham.ac.uk

There has been an explosion in the range of tools available for presenting data, many of which are available to support statistics teaching. These include tools that allow users to 'drag and drop' data sets (e.g. RAW), tools designed to display particular data sets (e.g. eXplorer) and software libraries (e.g. D3.js). We report on a review of visualisation tools, where we have described the sorts of visualisations facilitated by each tool, along with features such as ease of use and cost. Data visualisations can give new insights into complex data sets, and can be used directly to reshape teaching. We map out teaching opportunities facilitated by different tool types. Understanding novel data visualisations has become an important element of statistical literacy, and so curricula should expose students to a wide variety of examples.

INTRODUCTION

Statistics courses should help develop skills necessary for statistical enquiry in realistic settings. The 2016 GAISE guidelines recommend that statistics teachers should:

- Teach statistical thinking.
- Teach statistics as an investigative process of problem-solving and decision-making.
- Give students experience with multivariable thinking.
- Focus on conceptual understanding.
- Integrate real data with a context and purpose.
- Foster active learning.
- Use technology to explore concepts and analyze data.
- Use assessments to improve and evaluate student learning.

An on-going project, ProCivicStat, is an ERASMUS+ funded project that brings together 6 partner universities from 5 countries which addresses these goals. In particular, we are developing curriculum materials for high schools and undergraduate classes that require students to work with authentic data on topics relevant to social progress such as the gender pay gap, racism in football, and demographic change. These are inherently multivariable topics. The overall ambition of ProCivicStat is to engage students in the processes of investigation, problem solving and decision making in order to develop conceptual understanding and teach statistical thinking. We are presenting challenges based on authentic data from a wide variety of sources – GAISE's "real data with a context and purpose", and we make extensive use of technology to explore concepts and analyze data (the only GAISE element missing from our current work is assessment - we will be devoting time and resource to assessment later in the project).

In this paper, we describe a review of data visualisation tools we have conducted, along with brief descriptions of some of the excellent tools reviewed. Links to extensive collections of data (e.g. the World Bank, UN, and OECD) can be found on the ProCivicStat website. We also describe the ways in which working with data visualisations can support the GAISE goals.

FEATURES OF THE REVIEW OF VISUALISATION TOOLS

The review is contained in an Excel spreadsheet on the ProCivicStat website (http://www.procivicstat.org). For each tool, we provide the following information:

- Tool name and website address.
- Cost (free, free to educational users, pay).
- Accessibility (open source, free trial, online, cloud-based).
- Ease of use.
- Display functionality (a tick list for lines, tables, scatter graphs, bar charts, maps, combined charts, relational diagrams, boxplots; and notes on other functions such as radar plots, bubble graphs and funnel plots).

In: A Molnar (Ed.), *Teaching Statistics in a Data Rich World*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2017, Rabat, Morocco. ©2017 ISI/IASE iase-web.org/Conference_Proceedings.php

- Available data sources (pre-loaded with data, facilities to upload data).
- Data types handled (microdata, macrodata).
- Notes on interesting features (e.g. tutorial material available).

FUNCTIONS OF VISUALISATION TOOLS

In this section, we describe some different functions that can be served by data visualisation tools, with examples.

Tools designed specifically for statistics education

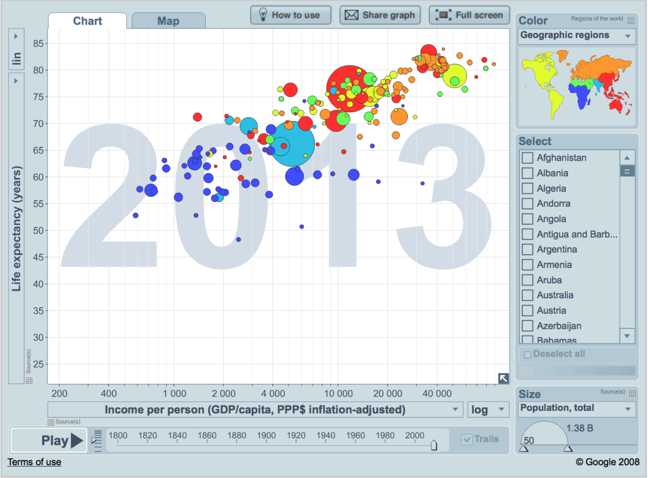
This category refers to software written explicitly to demonstrate statistical phenomena, such as: linear regression and correlation, and sampling and estimation (e.g. Hunt and Tyrrell, n.d.); Mittag's (n.d.). interactive statistical experiments; the instability of *p*-values with small samples and modest effect sizes (e.g. Cumming 2012, n.d.); permutation analysis (e.g., Lock, Lock, Lock-Morgan, Lock, & Lock, 2013); e-books on a variety of statistical topics (e.g. Sterling's CAST, n.d.); and tools for visual inference (e.g. Wild's iNZight, n.d.). These tools are well known to the statistics education community, and will not be discussed here.

Tools for graphical exploration

Graphing tools make it easy for users to drag and drop data, and then to explore it in a variety of ways. RAW (http://rawgraphs.io/) is a good example. RAW is open source. Users can drag and drop csv and tsv files (and text pasted from e.g. Excel) into the RAW interface, and can create a wide variety of graphs. Graphs can be exported as svg files and edited appropriately. There are a number of pedagogical uses for such tools. Graphing tools facilitate visual exploration of the structure of data sets before more detailed analysis is conducted; insights from the visual exploration might well obviate the need for some avenues of analysis. A large library of potential data visualisations facilitates discussions with students about the match between different sorts of data and appropriate displays. New graphing tools continue to be invented, and students need experiences in interpreting unfamiliar data visualisations. There is a large literature on misleading graphics in print media (e.g. Tufte, 1997, 2001; Wainer, 2000). New data visualisations will offer even more scope for misleading graphics, and students should acquire skills in understanding and critiquing novel visualisations (Sutherland and Ridgway, 2017).

Tools for exploring specific data sets

A number of organisations such as OECD, UN and Eurostat have large collections of data that they want to make accessible and intelligible to a broad public. Figure 1 shows Gapminder (https://www.gapminder.org). Here, life expectancy is plotted against income per person for a number of countries in 2013. Population size is shown by circle size, and colour shows world region. Gapminder software provides an animated display to show changes over time.



Income per person (GDP/capita, PPP$ inflation-adjusted) log

Color: Geographic regions

□ Afghanistan
□ Albania
□ Algeria
□ Andorra
□ Angola
□ Antigua and Barb...
□ Argentina
□ Armenia
□ Aruba
□ Australia
□ Austria
□ Azerbaijan
□ Bahamas

Size: Population, total

Figure 1: Gapminder screenshot.

In Figure 2, OECD's Regional eXplorer (http://www.oecd.org/cfe/regional-policy/oecdexplorer.htm) shows multiple representations of data on elderly dependence. The chloropleth map provides a spatial component, and the Gapminder-like bubble chart facilitates the exploration of covariates (here the proportion of the population aged between 15 and 64 years); a variety of charts are available for further exploration (see the lower right panel).

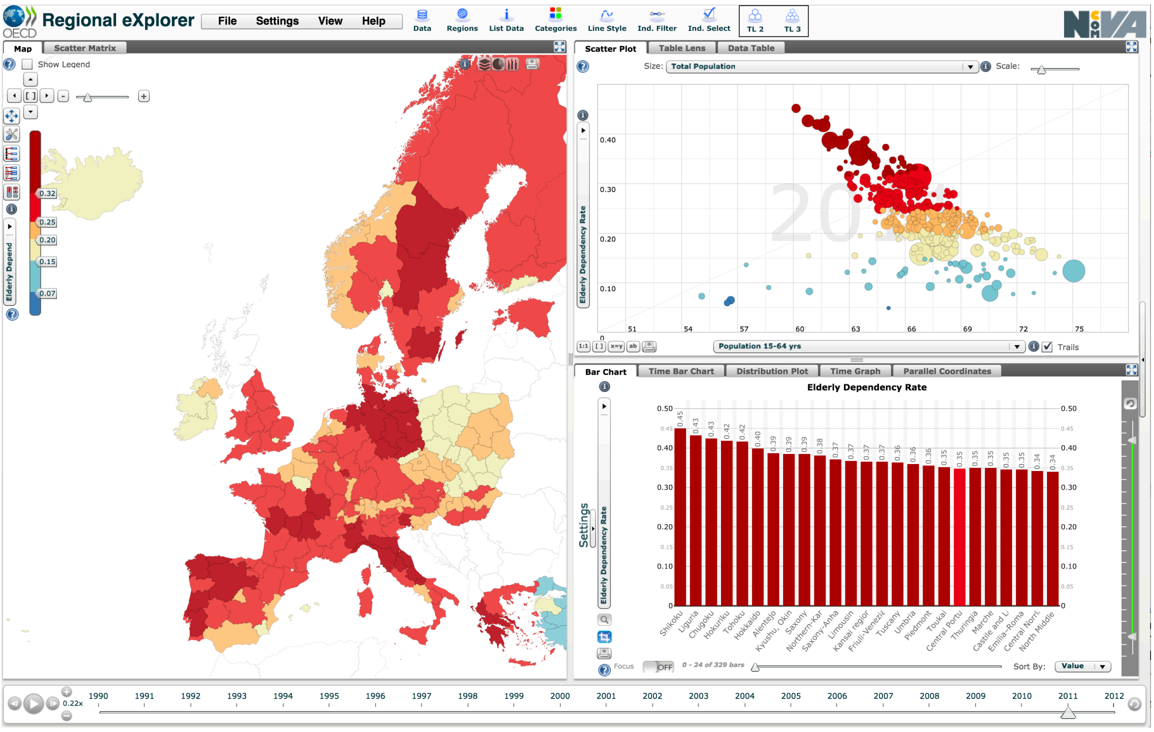
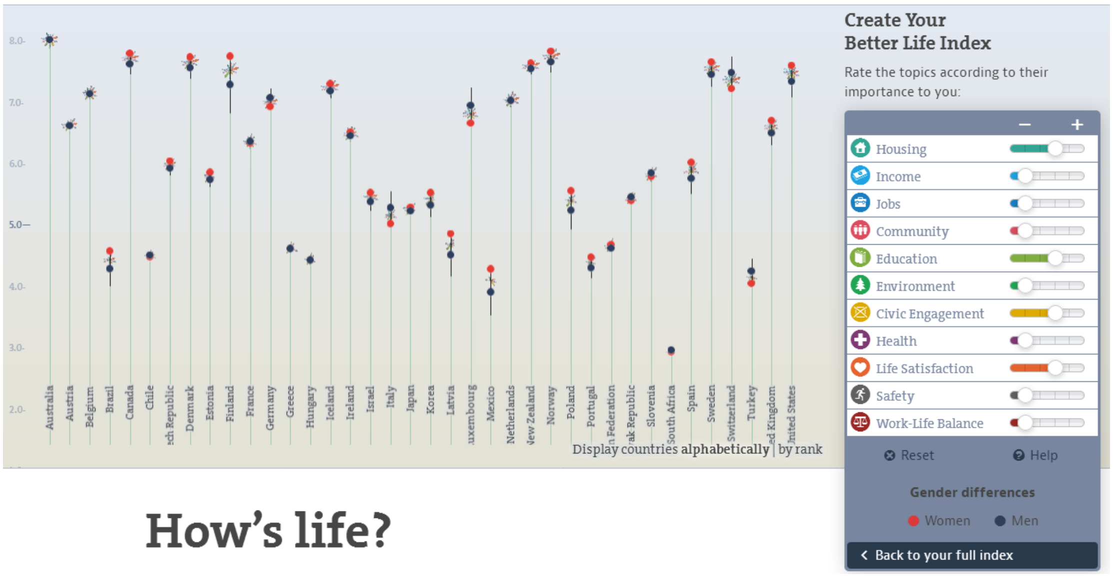


Figure 2: OECD's Regional eXplorer.

OECD's Better Life Index (BLI) allows users to define their own BLI by weighting different components. Survey data are available from different countries on satisfaction with each component. Figure 3 shows scores for different countries on a Better Life Index which places increased importance on Housing, Education, Civic Engagement and Life Satisfaction. Here, the display shows male and female ratings for different countries - illustrating a gender gap in the BLI.



Rate the topics according to their importance to you:

Figure 3: OECD's Better Life Index.

Other examples include Eurostat's Country Profiles and the UN's Devinfo. Most of these providers make it easy to download data for detailed exploration. Many allow users to upload data and make use of the data visualisation functions available, and then to export the finished product.

From the viewpoint of teaching statistics, these sites have a number of advantages. Data is authentic and multivariable. Multivariable displays are available. For example, Gapminder can display 5 variables simultaneously (x and y, time, country size, and country location). The SMART Centre interface allows up to 7 variables to be displayed. Data on these sites is (usually) up to date, and covers a number of countries. This facilitates the sharing of teaching resources; good teaching ideas developed in one country using these materials can be applied elsewhere, because it is easy to use the same ideas, but to apply them to different data sets. For example, activities designed to explore gender inequalities in one country or region can be customized to explore the same issues in other countries, where teachers can download data sets relevant to their context.

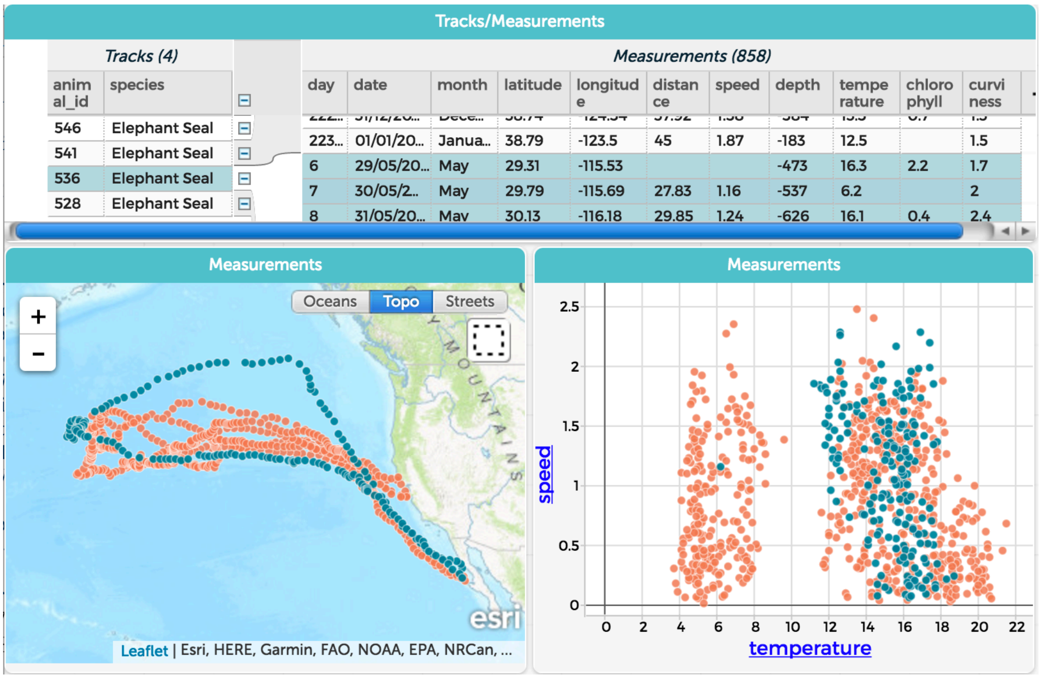
Presentation and analysis of data relevant to global issues such as poverty, migration and inequality can involve statistical techniques that might not be covered in conventional statistics courses. Examples include: the use of indicator systems (e.g. the UN Sustainable Development Goals, and OECD's Better Life Index); presentation and analysis of spatial patterns; and important measurement issues such as creating measures for hard-to-define concepts such as employment or poverty (and associated clear metadata), and techniques to ensure data quality. These data sources are often the basis for policy decisions at national and international levels, and so students can see the relevance of statistical thinking to policy.

Tools to support data science

Data science is a loosely defined discipline where people with skills in statistics and computer science (and specialist domain knowledge) come together to make sense of data that is often large-scale and messy. Increasingly, there has been a call for statistics curriculums to pay more attention to programming (Nolan and Temple Lang, 2010). Libraries of routines facilitate programming (both in terms of time taken, and the likely robustness of the code produced). For general programming in statistics, R is probably the best known example. Programming data visualisations is greatly aided by Bostock's Data Driven Documents (D3.js) – a powerful JavaScript library that underpins tools such as RAW. A considerable investment of time is needed to become fluent in R, or with software written to create data visualisations. However, data science can be introduced in simpler ways. Tableau (http://www.tableau.com) has been designed primarily for business use, but can be used in education free of charge. It is based on VizQL, a Visual Query Language, which can be used to query relational databases, cubes, cloud databases, and spreadsheets, and to generate a variety of graph types. Tableau combines a structured query language for databases with a descriptive language for rendering graphics. Facility with Tableau requires a investment of time.

In contrast, CODAP (https://codap.concord.org/) is a resource also designed to facilitate data wrangling, but it has been designed to be accessible to a broad range of users, including school children. It is easy to import data, to create a variety of visualisations, and to conduct some statistical analyses. The CODAP display in Figure 4 facilitates an analysis of journeys of 4 elephant seals. Each seal has its own track; one is highlighted in the figure. Variable names in the table can be dragged to graph axes to enable patterns and conjectures to be explored.

From a pedagogic viewpoint, CODAP offers a way to move evidence to the centre of the curriculum, and to encourage engagement from a variety of disciplinary perspectives with statistical thinking. This is particularly valuable in the context of GAISE's "real data with a context and purpose" (2016). Problems do not come in silos; for example, human activity has a dramatic effect on the biosphere; changes in the biosphere are likely to have a dramatic effect on human activity.



Tracks (4)

animal_id	species
546	Elephant Seal
541	Elephant Seal
536	Elephant Seal
528	Elephant Seal

Measurements (858)

day	date	month	latitude	longitude	distance	speed	depth	temperature	chlorophyll	curviness
223...	01/01/20...	Janua...	38.79	-123.5	45	1.87	-183	12.5		1.5
6	29/05/20...	May	29.31	-115.53			-473	16.3	2.2	1.7
7	30/05/2...	May	29.79	-115.69	27.83	1.16	-537	6.2		2
8	31/05/20...	May	30.13	-116.18	29.85	1.24	-626	16.1	0.4	2.4

Figure 4: CODAP Screenshot.

Free standing demonstrations

This group represents tools that are often seen in media outlets to demonstrate some particular phenomenon. They are often designed to capture attention, as well as to give access to data in novel ways. A dramatic example can be found at https://earth.nullschool.net/ where sensor data is used to map a wide variety of variables (wind speed, ocean currents, pollutant levels etc.) dynamically onto various projections of the surface of the Earth. More prosaically, there are displays that simulate stochastic processes such as death (such as Flowing Data's years-you-have-left-to-live-probably from 2015) that are inherently interesting, and can be used to introduce ideas about distributions, probability density functions and conditional probability. Such displays can have interesting features. For example, Figure 5 is a static shot from an animation showing the size of the arctic ice-cap between 1978 and 2016 ("Arctic Sea Ice", n.d.). Months are plotted on a circle, so the reader sees both large seasonal effects, and a steady decline in the size of the ice-cap over time. This is immediately evident in the animated version, but by representing time periods by colour (here, from purple to yellow), shrinkage over time can be seen in the static display.

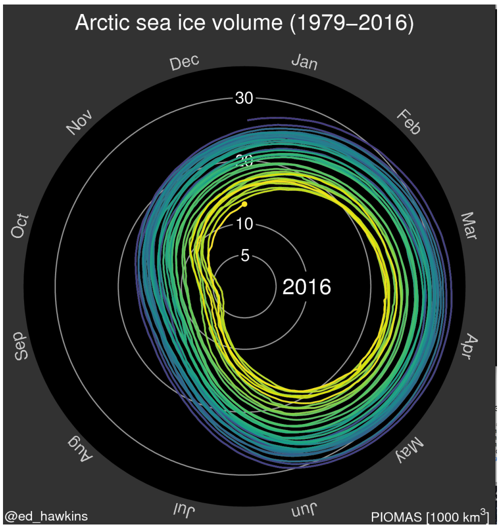


Figure 5: Arctic Sea Ice Volume Over Time.

CONCLUSIONS

The GAISE guidelines for teaching statistics emphasise active learning, statistical thinking, conceptual understanding, and the processes of problem solving and decision making. They also encourage the use of technology to explore concepts and analyse data, using real (multivariable) data with a context and purpose. Using data visualisations of large scale authentic data can help achieve the GAISE ambitions. Data visualisations can be categorised in a number of ways. Here we distinguish between: tools that provide free-standing demonstrations of particular phenomena; tools that offer ways to present data in a variety of displays; tools designed to present large-scale data sets flexibly; and tools that support data science (conceived broadly). Examples can often be classified under more than one of these headings.

Data visualisations have important implications for both practice and theory. At the practical level, data visualisations can facilitate exploratory learning and can be used directly to reshape teaching. At a conceptual level, understanding data visualisations has become an important element of statistical literacy. Students need experiences working with and critiquing novel visualisations. If we are to promote statistical literacy in the broader community, we need a better understanding of the cognitive processes involved in working with complex visual displays.

We are planning some experimental investigations where users at different levels of statistical sophistication engage with different displays. We hope that these investigations will help identify requisite interpretive skills, and will inform interface design (our own, at least!).

Describing, critiquing and researching the range of visualisations available will be an on-going process, so our review, and suggestions for teaching, are steps in providing a summary of what is available, and mapping out some opportunities.

REFERENCES

Arctic Sea Ice (n.d.). http://www.climate-lab-book.ac.uk/spirals/

Cumming, G. (2012). *Understanding the new statistics*. New York, NY: Routledge.

Cumming, G. (n.d.). Dance of the *p*-values. *Intro Statistics*. Melbourne, Australia: La Trobe University. [Online: www.youtube.com/watch?v=5OL1RqHrZQ8]

Flowing Data (2015). http://flowingdata.com/2015/09/23/years-you-have-left-to-live-probably/

GAISE College Report ASA Revision Committee (2016), *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*, Alexandria, VA: American Statistical Association.

Hunt, N. & Tyrrell, S. (n.d.) Regression and Correlation. http://www.icse.xyz/discuss/regression/

Hunt, N. & Tyrrell, S. (n.d.) Sampling and Estimation. http://www.icse.xyz/discuss/estimation/

Lock, R., Lock, P., Lock-Morgan, P., Lock, E., and Lock, D. (2013). *Statistics: unlocking the power of data.* Hoboken, NJ: Wiley.

Mittag, H-J. (n.d.). *Interactive statistical experiments*. http://www.mittag-statistik.de/app/

Nolan, D. and Temple Lang, D. (2010). Integrating computing and data technologies into the statistics curricula. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute.

SMART Centre (n.d.). https://www.dur.ac.uk/smart.centre/freeware/

Sterling, D. (n.d.). *CAST.* http://cast.massey.ac.nz/collection_public.html

Sutherland, S., and Ridgway, J. (2017). Interactive visualisations and statistical literacy. *Statistics Education Research Journal, 16*(1), 26–30.

Tufte, E.R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative.* Cheshire, CT: Graphics Press.

Tufte, E.R. (2001). *The Visual Display of Quantitative Information* (2nd edition). Cheshire, CT: Graphics Press.

Wainer, H. (2000). *Visual Revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* Mahwah, NJ: Erlbaum.

Wild, C. (n.d.). *iNZight*. https://www.stat.auckland.ac.nz/~wild/iNZight/index.php