Journal of Data Science, Statistics, and Visualisation

MMMMMMM YYYY, Volume VV, Issue II. *doi: XX.XXXXX/jdssv.v000.i00*

Visual narratives of the Covid-19 pandemic

Susan Vanderplas
University of Nebraska–Lincoln

Adalbert F.X. Wilhelm
Jacobs University Bremen

Abstract

Covid-19 has created a world-wide interest in understanding the dynamics of a pandemic. Hence, numerous media outlets as well as researchers have produced comprehensive data visualizations to illustrate the relevant trends and figures. In this paper, we will look at an elective choice of Covid-19 data visualizations to evaluate and discuss currently established visualization tools in their capacity to provide a communication channel both within the data science team and also between data analyst, domain experts and a more general interested audience. While there is no fixed catalogue of evaluation criteria for data visualizations we will try to provide an overview on the different core aspects of visualization evaluation and their competing principles.

Keywords: exploratory data visualisation, logarithmic scales, visual comparisons, R.

1. Introduction

Here some general introduction

Effective participation in the social dynamics of a democracy requires clear understanding of quantitative evidence and rules by the people. As the medical community, governments, business, and the public come together to fight the COVID-19 pandemic, scientists have a special role in helping to communicate what the data actually mean. For data analysis and modeling to have an impact as a component of decision making, appropriate reporting and communication is key. There are numerous standards for statistical reporting in the application areas, such as the ESS standard for quality reporting, the CONSORT, PRISMA, CHEERs guidelines, and others (see https://equator-network.org). These standards are based on commonly accepted core quality principles and values such as accuracy, relevance, timeliness, clarity, coherence, and reproducibility. For measures to restrain and overcome an epidemic effectively, communication among experts that follows highest professional and ethical standards is not sufficient. In a democratic society, policy measures can only be implemented if they are based on the acceptance of the wider population. This puts high demands on skills associated with communicating statistical evidence on the side of scientists, governments and media, and a citizenry able to understand statistical messages.

In recent decades, there have been numerous publications, initiatives, and ideas to improve the communication of quantitative and statistical information, see Hoffrage et al. (2000); Tufte (2001); Rosling and Zhang (2011); Otava and Mylona (2020), to name only a few. Data journalism has recently taken off as an innovative component of news publishing, and COVID-19 provides numerous excellent examples, often using an interactive visual format on the Internet, such as dashboards. A fundamental problem in assessing probabilities lies, for example, in the intuitive conflation of subjective risks ("how likely am I to become infected") and general risks ("how likely is it that some person will become infected"). Another issue is that of equating sensitivity of a diagnostic test and the positive predictive value (Eddy 1982; Gigerenzer et al. 2007; McDowell et al. 2019; Binder et al. 2020). In particular, the prevalence (or base rate) is often neglected leading to this confusion. As **?** Hoffrage et al. (2000) point out "... statistics expressed as natural frequencies improve the statistical thinking of experts and nonexperts alike". Smooth sliding back and forth between probabilities and natural frequencies is easy once figures are related to a unifying reference value. A well proven method to effective communication lies in representing information in accessible ways (Gigerenzer et al., 2007; Gigerenzer and Edwards, 2011) by saying, for instance, "one in 10" instead of 10%. Using absolute in place of relative numbers, representing information in appealing graphic forms, and summarizing most relevant facts in "fact boxes" are some of the methods developed and advocated for increasing transparency in communication between stakeholders such as health providers and patients (See, for example: https:// www. hardingcenter.de) Fact boxes combined with icon arrays are recommended for the presentation of test results. Both representations are based on natural frequencies (Gigerenzer 2011; Krauss et al. 2020) and present case numbers as simply and concretely as possible. Many scientific studies show that icon arrays help people understand numbers and risks more easily (e.g. McDowell et al. (2019)). The Harding Center for Risk Literacy shows many other examples of transparent communication of risks, including COVID-19[1].

While human thinking tends towards pattern simplification and political communication also prefers a simple cause-effect relationship, real phenomena are often multivariate. Thus, when studying COVID-19 and predicting its spread, it is not only important to consider its symptomatology, the incidence and geographic distribution of diseases, population behavior patterns, government policies and impacts on the economy, on schools, on people in nursing homes and on social life as a whole, but to integrate these into the data analyses and the communication of results. Associations observed in the data can often be caused by third-party variables (confounders). In addition, much of

[1] https://www.hardingcenter.de/de/mrna-schutzimpfung-gegen-covid-19-fuer-aeltere-menschen

Figure 1: Choropleth map of the incidence figures for Germany by district. Source: Robert-Koch-Institute https://app.23degrees.io/export/oCRP768wQ3mCswE7-choro-corona-faelle-pro-100-000/image.

the data comes from observational studies, which usually makes a robust causal attribution problematic. Statisticians calling out these limitations, however, face the danger that their statements might be pulled out one side in a polarized debate (McConway and Spiegelhalter 2021).

2. The global perspective

Visual representations take a central position in public communication and aim to represent the corresponding dynamics and contents in a quickly understandable way. Usually either time-dependent parameters or data with a spatial reference are visualised. For spatially distributed data, choropleth maps are predominantly used, in which administrative regions defined by the responsible health authorities are coloured according to the distribution density of the infection figures or variables derived from them (see Figure 1). Their visual perception problems - such as the visual dominance of the area of administrative regulatory frameworks that have no direct relation to infection events - are well known but still widespread. In addition, the use of ordinance thresholds as the basis for color scaling is often at odds with color schemes that emphasize real spatial distributional differences.

For time-dependent parameters, different variants of time series diagrams are used, predominantly line and column diagrams. Classic errors of graphical representation, such as overemphasising temporal variability by reducing the value axis to a small section, have now largely disappeared from the media. Switching between line and bar charts for purely design-related reasons in order to produce corresponding graphical diversity nevertheless seems questionable. The use of logarithmic scales in time series diagrams should be evaluated with caution. On the one hand, they tempt superficial readers to underestimate dynamic growth processes; on the other hand, they increase the demands on the mathematical and statistical literacy of the readership without corresponding advantages of visual representation. Figure 2 shows the time course of the 7-day incidence per 100,000 people between 24 January and 4 February 2021 for some selected countries. While the differences appear relatively small on the logarithmic scale, the linear scale shows considerable differences.

Simulations were used particularly illustratively in the media in the course of the pandemic. An inspiring example illustrating the spread of the epidemic appeared as early as 14 March 2020 in the Washington Post[2] with the title "Why outbreaks like coronavirus spread exponentially, and how to flatten the curve". The Washington Post made this simulation available free of charge and in all major languages, which led to it being distributed worldwide, including repeatedly on German television[3]. The New

[2] `https://www.washingtonpost.com/graphics/2020/world/corona-simulator/`
[3] `https://web.br.de/interaktiv/corona-simulation/`

incidence1.png

incidence2.png

Figure 2: The 7-day incidence for different countries over the course of time. On the logarithmic scale (left graphic), differences seem small. The linear scale (right graphic), however, shows considerable differences. Source: Our World in Data, https://ourworldindata.org/covid-cases?country=IND USA GBR CAN DEU FRA.

York Times[4] published a dynamic graphic entitled "How the Virus Won", which maps the spread of COVID-19 cases from February to June 2020 in the USA. It shows how an analysis of the associations between different COVID-19 strains and travel patterns can help understand the spread of the disease.

Another illustrative example is a simulation from ZEIT Online[5], which - based on models developed by a group of researchers at the Max Planck Institute for Chemistry - estimates the probability of an infected person infecting other people in closed rooms in various scenarios. While the visualisations present the simulated infection processes in a catchy way, the dependence of the simulations on parameter assumptions and settings is usually not addressed. Simulations should also always make transparent on which model assumptions and which data basis the simulations were created.

3. Comparisons and rankings

4. To log or not to log

5. Correspondence Analysis

[4] https://www.nytimes.com/interactive/2020/us/coronavirus-spread.html

[5] https://www.zeit.de/wissen/gesundheit/2020-11/coronavirus-aerosole-ansteckungsgefahr-infektion-

6. Summary and discussion

Computational Details

If necessary or useful, information about certain computational details such as version numbers, operating systems, or compilers could be included in an unnumbered section. Also, auxiliary packages (say, for visualizations, maps, tables, ...) that are not cited in the main text can be credited here.

The results in this paper were obtained using R 3.5.1. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Acknowledgments

All acknowledgments should be collected in this unnumbered section before the references. It may contain the usual information about funding and feedback from colleagues/reviewers/etc. Furthermore, information such as relative contributions of the authors may be added here (if any).

References

Binder, K., Krauss, S., and Wiesner, P. (2020). A new visualization for probabilistic situations containing two binary events: the frequency net. *Frontiers in psychology*, 11:750.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In Kahneman, D., Slovic, P., and Tversky, A., editors, *Judgment under Uncertainty: Heuristics and Biases*, pages 249–267. Cambridge University Press, Cambridge, DOI: 10.1017/CBO9780511809477.019.

Gigerenzer, G. (2011). What are natural frequencies? *BMJ*, page 343:d6386, DOI: 10.1136/bmj.d6386.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, 8(2):53–96.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science*, 290(5500):2261–2262, ISSN: 0036-8075, DOI: 10.1126/science.290.5500.2261, https://science.sciencemag.org/content/290/5500/2261.

Krauss, S., Weber, P., Binder, K., and Bruckmaier, G. (2020). Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit–Forschungsdesiderate und einige Antworten. *Journal für Mathematik-Didaktik*, 41(2):485–521.

McConway, K. and Spiegelhalter, D. (2021). Sound human, steer clear of jargon, and be prepared. *Significance*, 18(2):32–34, DOI: 10.1111/1740-9713.01508.

McDowell, M., Gigerenzer, G., Wegwarth, O., and Rebitschek, F. G. (2019). Effect of tabular and icon fact box formats on comprehension of benefits and harms of prostate cancer screening: a randomized trial. *Medical Decision Making*, 39(1):41–56.

Otava, M. and Mylona, K. (2020). Communicating statistical conclusions of experiments to scientists. *Quality and Reliability Engineering International*, 36(8):2688–2698, DOI: https://doi.org/10.1002/qre.2697, https://onlinelibrary.wiley.com/doi/abs/10.1002/qre.2697.

Rosling, H. and Zhang, Z. (2011). Health advocacy with gapminder animated statistics. *Journal of Epidemiology and Global Health*, 1:11–14, ISSN: 2210-6014, DOI: https://doi.org/10.1016/j.jegh.2011.07.001, https://doi.org/10.1016/j.jegh.2011.07.001.

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

A. More Technical Details

Appendices can be included after the bibliography (with a page break). Each section within the appendix should have a proper section title (rather than just *Appendix*).

B. Using BibTEX

References need to be provided in a BibTEX file (`.bib`). All references should be made with `\cite`, `\citet`, `\citep`, `\citealp` etc. (and never hard-coded). This commands yield different formats of author-year citations and allow to include additional details (e.g., pages, chapters, ...) in brackets.

Cleaning up BibTEX files is a somewhat tedious task – especially when acquiring the entries automatically from mixed online sources. However, it is important that informations are complete and presented in a consistent style to avoid confusions. JDSSV requires the following format.

- Specific markup (`\proglang`, `\pkg`, `\code`) should be used in the references.
- Titles should be inserted in title case.
- Journal titles should not be abbreviated and in title case.
- DOIs should be included where available.
- Software should be properly cited as well. For R packages `citation("pkgname")` typically provides a good starting point.

Affiliation:

email: susan.vanderplas@unl.edu Susan Vanderplas
Department of Statistics
University of Nebraska–Lincoln
349A Hardin Hall
Lincoln, NE 68583-0963, USA
E-mail: susan.vanderplas@unl.edu
URL: https://statistics.unl.edu/susan-vanderplas

Adalbert F.X. Wilhelm
Department of Psychology and Methods
Jacobs University Bremen gGmbH
Campus Ring 1
28759 Bremen, Germany
E-mail: a.wilhelm@jacobs-university.de
URL: http://www.jacobs-university.de/directory/wilhelm

https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/
MMMMMMM YYYY, Volume VV, Issue II
doi:XX.XXXXX/jdssv.v000.i00
Submitted: yyyy-mm-dd
Accepted: yyyy-mm-dd