Scientific Reform and Visual Data Science: Retiring the EDA/CDA dichotomy

Roger Beecham[1]

[1]*School of Geography, University of Leeds, UK. e-mail: r.j.beecham@leeds.ac.uk*

Abstract

Concerns around the replicability of published scientific findings has prompted much introspection into the way in which scientific knowledge is produced. To address issues of data fishing, searching exhaustively for discriminating patterns in a dataset, picking and then publishing those that are statistically significant, an argument is made that research findings should only be claimed through pre-registered confirmatory data analyses. Pre-registration studies are, though, somewhat inimical to the more informal research environments typical of modern applied data analysis ('Data Science'). In this talk I enumerate some of these challenges and demonstrate, through an analysis of road crash data in the UK, how nascent visualization techniques can be used to navigate and inject statistical rigour into contemporary data analysis environments.

Keywords. Exploratory data analysis; confirmatory data analysis; data visualization; replicability crisis; graphical inference, uncertainty visualization.

1 Scientific Reform and Data Science

A key tenet of Scientific Reform is that research findings are claimed through out-of-sample hypothesis tests, pre-registered in advance (Open Science Collaboration, 2015; Amrhein et al., 2019; Devezer et al., 2021; Szollosi and Donkin, 2021). As an organising framework for this, data analysis is separated into two discrete phases: Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA). EDA makes heavy use of graphical methods in order to discover high-level properties and structure in a dataset and to help formulate plans for further analysis. Once this scoping activity has been completed, a set of hypotheses are recorded (pre-registered) and an entirely new dataset is collected. A formal data analysis, CDA, is then conducted to evaluate whether the new data are consistent with the pre-registered hypotheses. Only at this point can empirical findings can be claimed and submitted for publication.

This $EDA \rightarrow preregistration \rightarrow CDA$ workflow is best suited to mature research settings where the theories and methods are well-developed and where prior studies and

data exist that can be used to judge and evaluate observed effect sizes (McIntosh, 2017). In modern applied data analysis settings, where administrative and passively-collected data are variously combined and repurposed (Singleton and Arribas-Bel, 2021), these certainties are difficult to achieve: the potential analysis space is broad and there is no stand-out approach to formulating research questions, selecting datasets and statistical techniques. Additionally, too strict an adherence to separating EDA (informal pattern-finding) from CDA (statistically-grounded enquiry) may not always lead to strong theory and knowledge development. Attempts to impose pre-specification designs onto research settings that rely on secondary or 'found data' may result in knowledge statements that are overly specific or that do not express sufficient detail or richness in outcomes (Szollosi and Donkin, 2021).

2 Visualization and knowledge development

In a recent discussion in the Harvard Data Science Review, Hullman and Gelman (2021b) make the case for a larger role, or a realignment, of 'exploratory' approaches in knowledge generation. They argue that model development should be intrinsic to EDA. Rather than simply displaying descriptive summaries of observed data, an EDA should support inferential thinking by encouraging analysts to consider, and subsequently model for, the processes that might have generated that structure. A key aspect of exploratory analysis is then enabling detailed comparisons against those reference distributions. If the ambition of a CDA is to accept/reject pre-specified hypotheses, then EDA is concerned with '*the particularities of the discrepancies between model and data*' (Hullman and Gelman, 2021b) — or locating and characterising *where* the data depart from the reference distribution.

Data visualization is instrumental to this sort of activity. Hullman and Gelman (2021b) and others (Hullman and Gelman, 2021a; Heer, 2021; Cook et al., 2021) envision data graphics supporting iterative comparison to models of increasing complexity and sophistication. This may happen either explicitly, for example in graphical inference where plots are compared to others generated under simulated data (Buja et al., 2009; Wickham et al., 2010; Beecham et al., 2017; Morris et al., 2019), or indirectly by the way in which the plot is composed and read.

3 Visualization and spurious discovery

A charge against elevating the status of exploratory analysis in this way is that visual approaches are vulnerable to over-interpretation and false discovery. Data graphics emphasise data patterns, but the complexities around those patterns may be overlooked (Hullman and Gelman, 2021b). It is worth mentioning two counter-arguments here.

First is that Hullman and Gelman (2021a) invoke the idea of graphical inference as model check (Gelman, 2004) – where observed data are compared graphically to reference

data replicated under a model. They signpost to recent work on uncertainty visualization and a growing repertoire of empirically tested techniques for representing these reference distributions, deliberately designed to protect against spurious discovery. These techniques have gained widespread use in scientific communication (Data Journalism) and for which there are software libraries to support easy implementation (cf. Kay, 2021a,b).

Second is more fundamental to Hullman and Gelman (2021b)'s central thesis. Rather than constraining graphical inference as model check to narrow null hypothesis tests, as in graphical line-up tests (Buja et al., 2009), Hullman and Gelman (2021a) locate graphical model checks within a Bayesian framework. Their re-thinking of graphical inference accepts flexibility around the features in the observed data being tested – there may be several irregularly-defined estimators – and that expectation will be contingent on analysts' prior knowledge and experience. This is particularly compelling for data analysis settings that rely on 'found data', as conventional statistical procedures may not transfer well to the observations being claimed and therefore requiring scrutiny.

4 SFDS 2022

Motivating this extended abstract was a real data analysis scenario (Beecham and Lovelace, 2022, see): a project where pedestrian casualties in STATS19 road crash data, the canonical road safety dataset for the UK, were analysed in order identify and justify, with statistical evidence, investment decisions. It was possible to derive many interesting geographic patterns in crash rates from detailed information on crash context and the vehicles and individuals involved. These were nevertheless patterns which could be reported with limited levels of confidence. Uncertainty around datasets, the selection of context variables and of appropriate statistical techniques and reporting mechanisms, meant it was difficult to make crisp statements around priority areas using the sorts of research designs required by the Scientific Reform movement (Devezer et al., 2021). At SFDS 2022 I will outline some of these difficulties, and demonstrate how the model-based visual techniques advocated by Hullman and Gelman (2021b) can be practically implemented to inject greater rigour into exploratory visual analysis activities that remain widespread, and necessary, to modern data analysis.

References

Amrhein, V., D. Trafimow, and S. Greenland
2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(sup1):262–270.

Beecham, R., J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood

2017. Map line-ups: effects of spatial structure on graphical inference. *IEEE Transactions on Visualization & Computer Graphics*, 23(1):391–400.

Beecham, R. and R. Lovelace
2022. A framework for inserting visually-supported inferences into geographical analysis workflow: application to road safety research. *Geographical Analysis*.

Buja, A., D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham
2009. Statistical inference for exploratory data analysis and model diagnostics. *Royal Society Philosophical Transactions A*, 367(1906):4361–4383.

Cook, D., N. Reid, and E. Tanaka
2021. The foundation is available for thinking about data visualization inferentially. *Harvard Data Science Review*.

Devezer, B., D. J. Navarro, J. Vandekerckhove, and E. O. Buzbas
2021. The case for formal methodology in scientific reform. *R. Soc. open sci*, 8:200805.

Gelman, A.
2004. Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13(4):755–779.

Heer, J.
2021. Exploratory analysis and its malcontents. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/vszs87oj.

Hullman, J. and A. Gelman
2021a. Challenges in incorporating exploratory data analysis into statistical workflow. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/2ym7zm34.

Hullman, J. and A. Gelman
2021b. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/w075glo6.

Kay, M.
2021a. *ggdist: Visualizations of Distributions and Uncertainty*. R package version 3.0.1.

Kay, M.
2021b. tidybayes: Tidy data and geoms for Bayesian models. R package version 3.0.0.

McIntosh, R. D.
2017. Exploratory reports: A new article type for Cortex. *Cortex*, 96:A1–A4.

Morris, T. P., I. R. White, and M. J. Crowther
2019. Using simulation studies to evaluate statistical methods. 38(11):2074–2102.

Open Science Collaboration
2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Singleton, A. and D. Arribas-Bel
2021. Geographic Data Science. *Geographical Analysis*, 53:61–75.

Szollosi, A. and C. Donkin
2021. Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, 16(4):717–724.

Wickham, H., D. Cook, H. Hofmann, and A. Buja
2010. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '10)*, 16(6):973–979.