Designing Graphics Requires Useful Experimental Testing Frameworks and Graphics Derived From Empirical Results

Susan VanderPlas¹

¹Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America

Published on: Jul 30, 2021

DOI: <https://doi.org/10.1162/99608f92.7d099fd0>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

[Hullman and Gelman \(2021, this issue\)](#) have provided a very thorough discussion of the premise that interactive exploratory data analysis requires a theoretical framework for graphical inference to effectively support the analyst and counter any tendencies towards assuming all results are real and not just due to sample variability. They specifically mention a common problem with papers that theorize about interactive analysis:

Even if some activities fall outside of the predictions of any specific model, without an underlying theoretical framework to guide the design of tools, we are hard pressed to identify where our expectations have been proven wrong and can easily end up with the sort of piecemeal and mostly conceptual theories that dominate much of the literature on interactive analysis. This lack of formalization makes it difficult to falsify or derive clear design implications from theoretical work.

Unfortunately, the theoretical framework for model checking during exploratory and confirmatory data analysis proposed in this paper is just another conceptual and theoretical framework that is difficult to test or falsify as presented.

The authors clearly support empirical validation of graphical analysis, but not enough to empirically examine their proposals via testing even a simple mock-up implementation of such a system with a fixed, relatively simple, data set. Without this empirical analysis, it is very difficult to see what this proposal adds to the two empirical methods discussed within as sub-cases of the model-check system, Bayesian Cognition and Visual Inference. These two formulations of empirical graphical testing arise out of different goals—to understand the mathematical reasoning used when processing charts, and to assess the perception of charts in the presence of randomization based on a null model. Unfortunately, the model-check integration proposal mentions, but does not actually address or even propose concrete solutions, to several major challenges that would need to be overcome in order to integrate either option into an automatic system for on-the-fly model checking. Of the two empirical processes ‘subsumed’ into the Bayesian model-check formulation, visual inference is perhaps the closest to the goals of model-checking integration proposed in this paper; as this is where I have done the majority of my work, I will address the shortcomings of the proposed system primarily from this perspective.

Issue 1: Automatic Null Model Generation

As the authors point out in the paper, visual inference, as proposed by Buja et al. (2009) and expanded upon in Wickham et al. (2010), requires that the analyst specify a null model for each lineup. While this process may be simple enough for simple queries (e.g. if looking for a linear relationship, resampling y values may suffice), it very quickly becomes complex because quite often the signal in the data is not sufficiently described in a way that simple resampling can fix. Ceding Hullman and Gelman’s proposition that we learn more from rejecting complex models than simple models, then, we can expect that in any meaningful test of a complex model, it will be relatively difficult to come up with a null generation method which adequately mimics the data characteristics which are not under examination. Speaking from experience, when writing VanderPlas and

Hofmann (2017), we spent approximately 80% of the time developing an adequate null plot generation model, 15% of the time collecting and analyzing the data, and the remaining 5% of the time actually writing the paper. Even then, we did not successfully create a null generation method which was sufficient to not distract from the target plots with visual characteristics that were not of interest to the investigation.

Many visual inference projects have the same problem—the null model is exceptionally difficult to generate in a way that tests only a single hypothesis about the data. In part, this is the curse of visual inference—our visual systems are so efficient at examining multiple properties of the data simultaneously that a null model which is effective at controlling extraneous properties is very difficult (if not impossible) to create with complicated data. This problem of calibrating visual or mathematical patterns in structured data is one that has been previously addressed on Dr. Gelman’s blog (2019). It is clear that the authors are aware that this is an issue, but choose not to address it with the depth it deserves, in part because (acceptable) automatic solutions may not be feasible given a novel dataset with unspecified dependency structure. Obviously, it is likely possible to automate some null model comparison, but whether the results from that would actually be useful to test the proposition at hand is another thing entirely. In any case, for the Bayesian model checking proposal, the software would not only have to define a sampling/data model, it would also have to elicit reasonable priors in order to arrive at the suggested posterior predictive visual model checks.

My concern with this prior and data model elicitation requirement is that if we minimize the challenges associated with solving this problem, we may end up with a solution integrated into software that is reductive and distracting. I will confess at this point that I have even reduced the problem’s complexity somewhat in the above paragraphs—because graphics provide us with a suite of possible visual comparisons to make, it is not clear how one might determine that the analyst is drawing conclusions based on one visual feature of the data, rather than another. VanderPlas and Hofmann (2017) demonstrated quite well that while we designed the experiment around clustering and strength of linear trends, viewers also incorporated features like cluster size, cluster dispersion, and the presence of outliers. Software which is capable of discerning the specific features the analyst is using to declare a graphic ‘surprising’ or ‘significant’ would need to either explicitly ask (which has its own problems) or be psychically linked to the analyst in order to design a null generation model which is suitable for investigating the likelihood of the finding being “real.” In this case, the suite of visual tests run when looking at a chart in the process of EDA or even rough CDA is a curse as much as a blessing, in that in order to actually implement the proposed system in a way that is not reductive, we must design an AI system for model generation which is even more complex than the not-yet-fully-understood visual system between our ears.

Issue 2: The Analyst’s Cognitive Load

One clear risk is that the additional cognitive load of interacting with reference distributions overwhelms some users, for example, distracting them from paying as much as attention to the data as they might have (Section 5.2).

It is worthwhile here to consider exactly what demands we are placing on the analyst during a relatively simple interactive data analysis.

1. The analyst must load the data and navigate to the relevant view and variables. At this point a good analyst will be thinking about how the data were generated, what the measurement process was, reviewing expectations as to missingness, and deciding what graph would best represent the variables.
2. The analyst must create the graph that answers any exploratory questions about the data related to e.g. range of the variables, missingness, relationship to other measured or predetermined variables, and so on. This is the start of the classic ‘EDA’ step, where there are not particularly firm hypotheses and the goal is to understand the data.
3. The analyst refines the graph generated in the previous step, potentially including sub-plots and mapping additional variables to features like color or shape. These additions may be due to a desire to reduce visual complexity of the plot (for instance, if the data is heavily over-plotted) rather than an explicit desire to test whether the sub-plots are all similarly shaped or whether there is an interaction between the faceted variable and the variables plotted on the main axes.
4. At the point where the analyst is relatively satisfied with their initial graph, the system might offer the chance to enter a model-check subroutine. The analyst might be asked to record their initial conclusions from the graph in some sort of data log before entering this subroutine, for observational purposes, or perhaps the subroutine would be launched without asking for the analyst’s consent or current conclusions.
 - a. Initially, the subroutine would have to elicit the analyst’s prior beliefs about the variables (perhaps this step might need to be included after step 1 of the main loop for accuracy’s sake) or the analyst would have to use system defaults of weakly informative priors drawn from other datasets (if available). In the latter case, the analyst might need to look at one or more prior predictive distributions to ensure that the automatically generated priors are representative of the analyst’s beliefs.
 - b. Next, the subroutine would have to elicit the analyst’s working graphical model. Perhaps we could start with a full model detailing the x and y axis variables, the subplot variables, and the color and shape mappings, and then allow the analyst to remove any combinations which are not relevant. Then, the analyst would have to select appropriate data-generating models/distributions for each variable. This process might include 4-5 variables at a minimum for a relatively simple graphic.
 - c. Then, the subroutine could combine the priors and the data model to generate posterior predictive versions of the graphic the analyst created. The analyst would have to decide among a panel of 8-10 already faceted, relatively complex graphics, to decide which one is unique.
 - d. The model check interface might generate a report showing the analyst that their conclusions from the data are not necessarily specific to the data itself, depending on the results of (c).

From a cognitive load perspective, steps (a) and (b) would seem to distract from the “flow” of exploratory data analysis. That is, by deviating from the focus on the specific aspect of the data the analyst wants to explore to requiring thoughts on the marginal effects, the analyst’s working memory and executive function resources are

diverted to a different (if related) set of problems. Then, in step (c), the analyst is asked to handle the original problem, but in a context which is 8-10 times as visually complex (by virtue of the additional posterior predictive graphics the analyst is comparing to). Given that pairwise comparisons between successive charts are necessary, this task is actually on the order of $\left(\frac{9}{2}\right) = 36$ times more cognitively demanding, assuming that the analyst knows exactly which feature is most important for comparing the panels, and that comparisons between sub-panels of each of the 8-10 main panels are not necessary (which is not necessarily likely). Incidentally, here, the analyst has an advantage that the typical users of visual inference don't have—the analyst saw the initial data plot first, which might leave them with some preconceived notions of what to look for in any subsequent visual inference style comparison plots. At any rate, if more than one graphical feature is notable in a chart, the visual complexity of the pairwise comparisons and the resulting cognitive load increases exponentially.

I have left out any difficulties with the user interface, assuming that the user-centered design approach proposed in the paper is in fact effective, but realistically, by the time the analyst gets to step (c) of the model check subroutine, it seems likely that they would be visually and mentally overstimulated to an extreme degree. Then, presumably, the analyst would be expected to continue the EDA/CDA/rough CDA process by moving on to another set of variables, another plot, another refinement of their ideas. At that point, however, I suspect that I am not alone in thinking that I'd probably take a walk, get some coffee, and then pick a different task upon my return to my desk, because of the strain of the extreme demands of cognitive load under step 4 of the EDA loop using the proposed model-check system.

I do not think the authors do anyone any favors by deferring the cognitive load question:

A key question that often remains unstated in research in interactive visual analysis is: How much of the statistical inference process that an analyst engages in should be left implicit in order to preserve cognitive load? We acknowledge that it is difficult to answer this question without first making concerted attempts in research to realize the forms of integration we describe above.

Fundamentally, graphics are about leveraging the visual bandwidth available for data exploration, discovery, and communication. Deferring cognitive load questions to philosophical discussion does no favors to the researchers or analysts who may decide to implement or use a system like the one proposed, in part because by using graphics, we are already increasing the cognitive bandwidth available beyond what can be communicated in a simple table. If graphics are useful for exploration and model diagnostics, it is because there is a human brain at the receiving end of the graphics, and that brain has limitations (limitations that cannot be expected to change with hardware improvements, as a computational limitation might). When proposing a setup like the model-check framework, it is essential to have even some limited evidence that the human in the loop is capable of fulfilling their part of the proposed tasks. Instead, the authors duck the essential question of whether this proposal is actually feasible for the analyst, focusing instead on the statistical benefits.

It would have been relatively straightforward for the authors to mock-up an interface which would allow an “analyst” to perform these tasks for a single data set, and to record the analyst’s thoughts via a think-aloud protocol paired with e.g. an eye tracking system (used in this case to monitor the analyst’s gaze and/or focus—I am in no way suggesting eye tracking be integrated with typical analysis software). The authors could even have used a measure of stress to track analysts’ reactions to the additional demands. If carefully combined with a basic questionnaire before and/or after, and possibly a comparison to analysis of a similar dataset using the current protocol, this process could have provided at least some basic information about whether analysts have enough remaining cognitive resources to entertain a model-check formulation as proposed in the paper, and whether any promised benefits of understanding variability and reducing overconfidence in observed effects were actually noticeable from the analyst’s perspective.

Conclusion

In short, while the proposed Bayesian model-check framework is an interesting idea, I do not think that it is ready for implementation without consideration of the technical and human limitations to success. The authors neatly dodge any responsibility for actually implementing or even developing concrete examples of how this process might work in practice, and the lack of those factors is extremely obvious when we consider what it might take to use the proposed process from the analyst’s perspective. Graphics are useful in part because humans can test many hypotheses at once using a visual representation of the data; if we overload and “short-circuit” that essential capacity, then we actively hamper the statistical modeling and analysis process, rather than augmenting it with new capabilities. If what interactive analysis needs is empirically testable propositions rather than falsifiable theories (and I would agree that we do), then we can use (and are using) Bayesian Cognition and Visual Inference to explore many different factors in the perceptual and cognitive factors affecting visualization. Further attempts to create grand unified theories of model-checking for visualization research would do well to accompany those theories with empirical results that support their benefit to the community.

Disclosure Statement

Susan VanderPlas has no financial or non-financial disclosures to share for this article.

References

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A*, 367(1906), 4361–4383. <https://doi.org/10.1098/rsta.2009.0120>

Gelman, A. (2019). Calibrating patterns in structured data: No easy answers here. *Statistical Modeling, Causal Inference, and Social Science*. <https://statmodeling.stat.columbia.edu/2019/07/12/calibrating-patterns-in->

[structured-data-no-easy-answers-here/](#)

VanderPlas, S., & Hofmann, H. (2017). Clusters beat trend!? Testing feature hierarchy in statistical graphics. *Journal of Computational and Graphical Statistics*, 26(2), 231–242. <https://doi.org/10.1080/10618600.2016.1209116>.

Wickham, H., Cook, D., Hofmann, D., & Buja, A. (2010). Graphical inference for Infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 973–979. <https://doi.org/10.1109/TVCG.2010.161>

©2021 Susan VanderPlas. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.