

Visual Statistical Inference for Political Research

Richard Traunmüller, *University of Mannheim & Goethe University Frankfurt*

Abstract

This paper introduces a remedy to the criticism frequently voiced against data visualization and exploration: that it may give rise to an over-interpretation of random patterns. A way to overcome this problem is the realization that “visual discoveries” correspond to the implicit rejection of “null hypotheses”. The basic idea of visual inference is that graphical displays can be treated as “test statistics” and compared to a reference distribution of plots under the assumption of the null. Visual inference helps us answer the question “Is what we see really there?” By so doing, it seeks to overcome long-standing reservations against visualization as merely “informal” approach to data analysis and the fear that beautiful pictures may in fact not correspond to any meaningful patterns of substantive scientific interest. The paper illustrates the application and benefits of this visual method by drawing on examples from political research.

Introduction

Data visualization is an indispensable tool for political science research. Few methods are better able to uncover and communicate structure in quantitative data. Next to the compelling *presentation* of statistical results and quantities of interest (Jacoby & Schneider 2010, Kastlelec & Leoni 2007, King et al. 2001) statistical graphics are used as *analytic* tools for various purposes and at various stages of the data analysis (Jacoby 1997a, 1997b, 2000, Bowers 2004, Bowers and Drake 2005, Gelman 2003, Gelman and Hill 2007, Kerman et al. 2008). Based on an analysis of all articles published in the *American Journal of Political Science* between February 2003 and March 2018, Figure 1 demonstrates that graph use has dramatically increased in political science over the last fifteen years.¹ Whereas the average political science article in the discipline’s flagship journal contained roughly one (.92) graphic in 2003, graph use has grown to an average of three and a half (3.58) graphics per article in 2018.

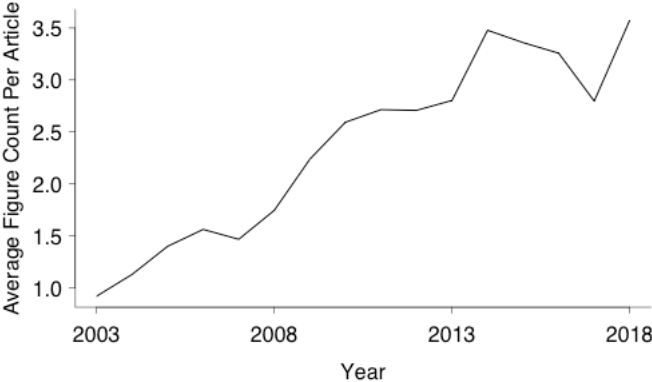


Figure 1: Average Number of Figures in all Articles Published in the AJPS, 2003-2018.

¹ I use figure count as a preliminary proxy for graph count. Of course, not every figure is necessarily a graphic. I am currently in the midst of classifying all figures and based on a random sample of roughly 700 figures I found that roughly 10 percent are not data visualizations.

Despite the widespread use of graphs and the clear benefits of turning abstract data structures into visible patterns, a long-standing reservation against data visualization holds that it is merely an “informal” approach to data analysis (cf. Best et al. 2001, Healy & Moody 2014). The fear expressed in this view is that beautiful pictures may not correspond to any meaningful patterns of substantive scientific interest. Instead, it is argued, serious scientists should base their inferences on more “formal” methods of hypothesis testing to discern signal from noise.

In this paper I seek to overcome these reservations against data visualization. In particular, I introduce *visual statistical inference*, a new visual approach that was only recently developed in statistics and information visualization (Buja et al. 2009, Wickham et al. 2010). The basic idea of visual inference is that graphical displays can be treated as “test statistics” and compared to a “reference distribution” of plots under the assumption of the null hypothesis. The null hypothesis usually posits that there is no systematic structure in the data and that any pattern is really the result of randomness. If the null hypothesis were indeed correct, the plot of the true observed data should not look any different from the plots showing random data. If however the plot of the true data clearly stands out from the rest, this could be taken as a rejection of the null hypothesis of no structure. In other words, visual statistical inference brings the rigor of statistical testing to data visualization.

This paper shares the general spirit of Jacoby (1997a: 12) who argues that “graphical approaches [...] should be very useful in the social sciences, where the robustness characteristics of traditional statistical techniques often are pushed to their limits” (Jacoby 1997a: 12) and Bowers & Drake (2005: 303) who state that given the many challenges of political science data (i.e. small N, non-stochastic data, see also Western & Jackman 1994, Stegmüller 2013) it is best to rely on “graphical presentations [...] rather than formal hypothesis testing.” But this paper goes a step further by showing that visualization and

statistical inference are not at odds with each other. Instead, it demonstrates how visual statistical inference merges statistical testing with data visualization and illustrates the application and benefits of this visual method by drawing on examples from political science research. The hope is to stimulate the use of graphical displays as analytical tools in political science by contributing to the exploding interest in visual methods and catering to discipline's "need to do a better job of data visualization" (Alvarez 2016: 15).

Visual Statistical Inference

Exploratory data analysis, according to one of its founders "is about looking at data to see what it seems to say. It concentrates on [...] easy-to-draw pictures. [...] Its concern is with appearance, not with confirmation" Tukey (1977: V). Consequently, a criticism that frequently arises with data visualization is that it is merely an informal tool for exploration and that it lacks the rigor of formal tests found in confirmatory analysis or conventional statistical inference. Exploratory data analysis and graphical displays may thus give rise to an over-interpretation of patterns that are in fact due to mere randomness: "Humans' pattern recognition skills are amazing and the source of great insights, but sometimes they're too good. We are so adept at finding patterns that we sometimes detect ones that aren't really there" (Few 2009: 139). This is where visual statistical inference steps in and "allows us to uncover new findings, while controlling for apophenia, the innate human ability to see pattern in noise" (Wickham et al. 2010). In other words, visual inference brings formal statistical testing to data visualization.

The key idea to overcoming the seeming opposition of exploratory and confirmatory analysis is in the realization that graphical displays can be considered as implicit comparisons to a reference distribution or a baseline model (Gelman 2003, 2004). For instance, any trend in a line chart is essentially an implicit comparison to a flat line or any pattern in a scatter plot

an implicit comparison to a random cloud of points. At the same time, a visual trend or pattern is considered “surprising” or “interesting” if compared to and contradicting implicit prior expectations. If we are able to make these implicit models explicit, we may formalize visual discoveries as any systematic difference to what we expected to see and the rejection of null hypotheses (Buja et al. 2009).

The Logic of Hypothesis Testing

Formal testing involves the comparison of a test statistic to its reference distribution under the assumption of the null hypothesis (cf. Gill 1999). If the test statistic is reasonably unlikely to have occurred under the null assumption, say $p < 0.05$, then the null hypothesis is rejected and one has a “statistically significant” result. These basic principles remain the same in visual inference – with the exception that the “test statistic” is now a graphical display of the data and the “reference distribution” made up of a collection of plots showing the null assumption. The “rejection” of the null involves a human viewer able to discern the true plot. This correspondence is illustrated in figure 2.

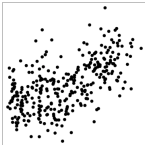
Formal Test	Visual Inference
Null hypothesis H_0	Null hypothesis H_0
Test statistic $T = h(x)$	Visual feature in a plot 
Test: Reject? $T(x) > c ?$	Human viewer: Discovery?

Figure 2: The Correspondence of Formal Hypothesis Testing to Visual Inference. Adapted from Buja et al. 2009.

Plots as Test Statistics

In visual inference plots take the role of test statistics. Using a graphical display as a “test statistic” in the sense of a data summary has several potential advantages over numerical test statistics (Anscombe 1973, Jacoby & Schneider 2010). First, in contrast to classical test statistics, graphs make little or no assumptions about the nature of the data such as their scale, functional form or distribution. Second, graphs can be used to describe complex data patterns for which simply no test statistic exists. Third, graphical displays of data retain more information about the data and may encourage further investigation by not only indicating if, but by also showing how the data deviate from the reference distribution. Of course, the choice of a particular graphical format depends on both the nature of the data and the pattern it is supposed to reveal. Line charts are commonly used to show time trends, bar charts to show the distribution over discrete categories and histograms for continuous variables. Scatter plots are incredibly versatile in not only revealing relations between two variables of any functional form, but also clusters, gaps and outliers in the data.

Generating Reference Distributions Using Variable Permutations

Visual inference is closely related to permutation tests in the way reference distributions under the null are generated (see figure 3). In permutations tests, one repeatedly permutes the “labels” of observations and calculates the test statistic under each of those permutations, resulting in the sampling distribution under the null and conditional on the observed data (Good 2005, Berry et al. 2016). The “labels” could indicate any attribute relevant to the analysis, such as treatment condition, time point or group membership. Under the null hypothesis, these labels are unrelated to the outcome of interest (i.e. it does not differ across treatment conditions, time or group) and therefore any random permutation would be equally likely. The only assumption needed for the re-labeling is that

the data are exchangeable under the null, i.e. their distribution remains the same whatever the particular labeling. The statistical significance of the observed test statistic can then be evaluated by comparing it to sampling distribution of the test statistics created by the random permutations.

Constructing reference distributions using permutations is useful for at least two reasons. First, it is not restricted to large sample sizes and specific distributional assumptions concerning the outcome (such as normality). Instead, it is extremely flexible and can be applied in a wide range of different settings, which are typical in political science research (small N, sparse or ill-behaved data). Second, permutation produces an exact description of the sampling distribution under the null and does not rely on approximations.

While for small N all possible permutations can be produced, for larger N one would choose a random subsample of all possible permutations using a Monte Carlo approach. Permutation tests ignore sampling variability in the data which is fine for many political science applications dealing with full population data. If sampling variability is an issue one could use a bootstrapping approach to building the reference distribution.

While a simple model of independence or “no structure” is a natural choice in many situations of exploratory data visualization, it is also possible to construct reference distributions from more specific models. For instance, one could construct a reference distribution by simulating draws from a normal distribution. In a Bayesian context, one could simulate reference data sets by sampling from the posterior predictive distribution (Gelman et al. 2013, Lynch & Western 2004, Kruschke 2013).

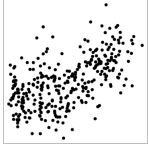
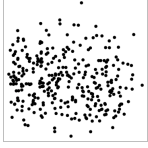
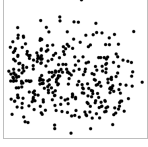
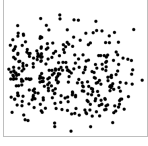
Permutation Test	Visual Inference
Test statistic of true data $T = h(\mathbf{x})$	Plot of true data 
Test statistic simulated data $T_1 = h(\mathbf{x}_{s=1})$ Test statistic simulated data $T_2 = h(\mathbf{x}_{s=2})$... Test statistic simulated data $T_S = h(\mathbf{x}_{s=S})$	Plot of simulated data  Plot of simulated data  ... Plot of simulated data 

Figure 3: The Correspondence of Permutation Testing to Visual Inference. Adapted from Buja et al. 2009.

The “Line-Up” as Visual Inference Tool for Political Research

This section introduces “The Line-Up” protocol, an inferential process based on graphical displays of quantitative information that mimics conventional hypothesis tests. It was developed in the statistics literature by Buja et al. 2009 and introduced to the information visualization community by Wickham et al. 2010. Majumder et al. 2013 established the validity, refined the terminology and present ways to calculate p-values and the power of visual tests. Hoffmann et al. 2012 applied visual inference in the power evaluation of graphical designs, Chowdhury et al. 2015 in a large p , small N data problem of gene expression data, and Widen et al. 2016 in examples from climatology, biogeography, and health geography. To the best of my knowledge this approach has not been used in political science.

The lineup is called “after the ‘police lineup’ of criminal investigations [...], because it asks the witness to identify the plot of the real data from among a set of decoys, the null plots, under the veil of ignorance” (Buja et al. 2009: 4369). In particular, this visual hypothesis test involves the simulation of $m-1$ null plots (for instance using variable permutations as explained in the previous section) and randomly placing the plot of the real observed data among them, resulting in a total of m plots. A human viewer is then asked to choose the plot that looks the most different from the rest. Ideally this human viewer is an impartial observer who has not yet seen the true plot before, such as a colleague, student or crowd worker (see below). If the test person succeeds and picks the plot showing the actual data, then this visual discovery can be assigned a p -value of $1/m$. In other words, the probability of picking the true plot just by chance is $1/m$. Setting $m=20$ and thus simulating $m-1=19$ null plots thus yields the conventional Type I error probability of $\alpha = .05$.

We can further decrease the probability of making Type I errors by either increasing the number of null plots, $m-1$, or by increasing the number of observers, K . For more than one

test person, picking the true plot under the null is a random variable, X , distributed as a binomial variable, $X \sim \text{Binom}_{K,1/m}$, with K trials and success probability $1/m$. Thus the p-value for a line-up with m plots and K (independent) impartial observers is (Majumder et al. 2013):

$$\Pr(X \geq x) = 1 - \text{Binom}_{K,1/m}(x-1) = \sum_{i=x}^K \binom{K}{i} \left(\frac{1}{m}\right)^i \left(\frac{m-1}{m}\right)^{K-i}$$

where x is the number of human viewers picking the true observed plot. In visual inference a type II error occurs if the human viewers fail to identify the true plot and thus to reject the false null hypothesis. The respective error probability is $\Pr(X < x)$. Majumder et al. (2013b) show that the power of visual test can be at least as good as the power of conventional test and even better under some circumstances. Of course, the true strength of visual inference lies in situations where no conventional tests exist.

The “Line-up“ is further aided by general principles and methods of graphical displays that facilitate the comparison between the plots, most notably the idea of “small multiples” (Tufte 1989). This refers to the careful arrangement of graphical displays of the same type, appearance and size that also have constant axis scales. In other words, single displays differ only in the data they present. While “small multiples” usually benefit hugely from a meaningful ordering of the plots, the central idea of the “line up” is of course the random arrangement of the null plots and the plot showing the real data. How the arrangement of the total set of plots as well as how the plot type or format affect the efficiency of the Line-up protocol is part of ongoing research (Hoffmann et al. 2012, Majumder et al. 2013).

Needless to say the effectiveness of the “Line-up” also rests on properties of the individual graphical displays that make up the total set of plots. Here, the principles of good data visualization concerning the choice of graphical displays and their design, discussed in many

classic texts, apply (e.g. Cleveland 1993, 1994, Few 2009, Tufte 1984). The most important advice is arguably to increase the “data-ink ratio” by focusing on showing the data and reducing any auxiliary graph elements. In particular, plot annotations such as titles, axis or tick labels, and legends can usually be omitted.

Figure 4 gives a first impression of how this inferential process works. Try for yourself:
Which of the 20 histograms stands out from the rest?

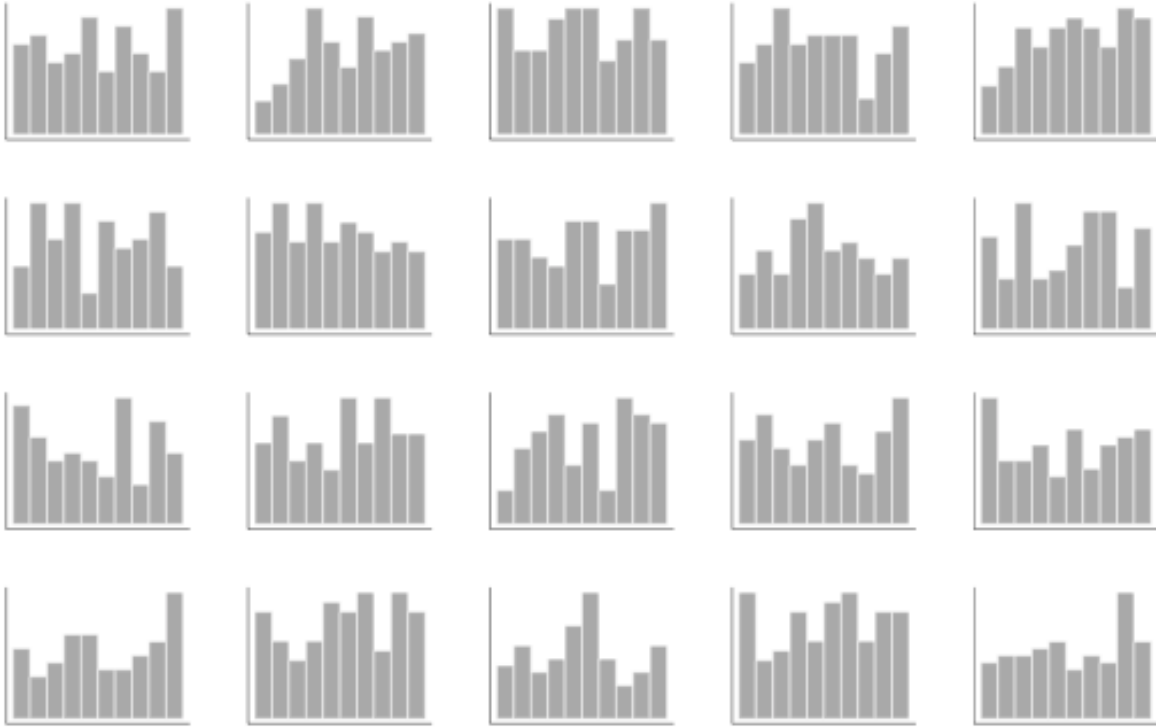


Figure 4: Line-up with 20 histograms. Which plot is the most different?

In fact, none is the true plot. All 20 histograms show 100 random draws from a uniform distribution $U(0, 1)$. However, this demonstrates how easy it is to over-interpret patterns that are in fact due to mere randomness. Buja et al. 2009 call this the “Rorschach protocol” and suggest it be used as a calibration exercise.

Examples from Political Science


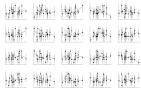


In this section I illustrate the workings of graphical inference by applying it to examples that face some of the methodological challenges found in political science research: a) tracking policy change over many countries and several time points using a heat map, b) studying context effects with few context units using a scatter plot of regression results, c) identifying clusters in data using a scatter plot with colored dots and d) finding spatial patterns in a dot map.

For each example, I used a small sample of N=9 political scientists² as well as a small sample of N=10 crowds-sourced respondents³ as impartial observers. In addition to achieving more reliable visual inference, this allows me to see to what degree experts differ from non-experts in their visual pattern finding skills. Using an online survey tool, respondents were presented with a total of five different line-ups (four “real” line-ups plus the “Rorschach test” presented in figure 1) and asked each time to pick the plot, which is “the most different.” While I randomized the order in which the line-ups were shown, the position of the true plot among the null plots was kept constant. Respondents could choose the plot they thought to be the most different by directly clicking on them. On average respondents took about three and a half minutes (221 seconds) to complete the task, although the crowd-workers were almost significantly faster than the political scientists (167 vs. 281 seconds). The results are documented in table1 and I will refer back to them during the discussion of each line-up.

² I asked 10 colleagues at different career stages (PhD students, postdocs as well as assistant and full professors) to participate. All of them have a strong quantitative research background and use plots in their daily work. Still, several mentioned that it was „hard“ to pick the true plot – which is a good thing and demonstrates that this exercise is not trivial.

³ I used the commercial crowdsourcing platform Microworkers which is similar to Amazon’s MTurk. Each worker was paid 30 cents for completing the task. The overall cost for this visual inference test therefore amounts to three US dollars.

Table 1: Experimental Results: Correct Identification of True Observed Plots

	All (N=19)		Political scientists (N=9)		Crowd Workers (N=10)	
	#	%	#	%	#	%
	7	37	6	67	1	10
	0	0	0	0	0	0
	18	95	9	100	9	90
	10	53	7	78	3	30

a) Time Series Cross Sectional Data: Blasphemy Legislation Across the Globe

A very common data structure in political science is time series cross sectional (TSCS) data that provide information on several states over a period of several years. For this example I draw on a TSCS dataset of the *Religion and the State Project* (Fox 2008) which provides data on religious policy by the state for 177 countries across the globe and for the period 1990-2008. For this application I am interested in blasphemy laws that protect religious figures and groups from insulting and discriminatory public speech and in particular how this type of regulation has changed over the time period. The natural null hypothesis therefore simply states that there was no change over time.

The visual “test statistic” used here is a heatmap that shows the number of blasphemy regulations (0-3, where darker red means more regulation) across countries (y-axis) and years (x-axis). Heat maps are a good alternative to line charts when the outcomes are discrete and thus over-plotting becomes a serious concern in the visualization of comparative trends.

The heat maps were sorted across countries according to the amount of regulation. To generate the “reference distribution” under the null, I generated 19 random permutations of the columns of the heat map. This is equivalent to randomly and repeatedly scrambling the year variable in the data set and thus breaking any dependence between time and blasphemy laws. Under the null the year should not matter for regulation and each permutation would be equally likely. The 19 null plots, i.e. the 19 heat maps of the permuted data, were then randomly arranged in a 4 x 5 matrix, that also includes the true heat map of the actual data (see figure 5). Can you detect the true plot?



Figure 5: Line-up of heat maps for the change in blasphemy laws across 177 countries and the period 1990-2008 .

As it turns out, only 37 percent of the human observers in my sample (67 percent of political scientists and only 10 percent of crowd-sourced respondents) correctly identified the true heatmap.⁴ Yet, even that only 7 out of 19 respondents would pick the true plot just by chance

⁴ The true heat map is the one in the fourth row and first column.

is extremely unlikely ($p \approx .00002$). Thus, we may reject that the null hypothesis that there is no change in blasphemy laws in our plot.

b) Contextual Effects: Education and Political Participation in the US

The next empirical example follows Bowers and Drake (2005) and looks at the relation between education and political participation in the US and how this individual level relation is conditioned by state-level educational context. A typical concern with this kind of analysis is that the number of contextual units is too small to rely on asymptotic assumptions of classical statistical inference. Therefore, Bowers and Drake (2005) suggest visual methods instead of formal tests. Yet their visual inference remains purely informal: “when we detect a feature with our eyes, we will try to only report it as a feature rather than noise if we feel that any reasonable political scientist in our field would also detect this feature“ (Bowers & Drake 2005: 17). Applying the visual inference approach introduced in this paper, we can swap our assumptions concerning the reasonableness of political scientists for a formal visual test. The null hypothesis in this example is that there is no relationship between the educational context in a state (i.e. the share of highly educated) and the effect of individual education on political participation.

The “test statistic” is a scatter plot version, where each dot is a state-specific individual-level effect of education on participation which is plotted along with vertical lines for the 95% confidence intervals. The size of this individual-level effect is on the y-axis. On the x-axis is the share of highly educated in the state. In addition, the plot includes a non-parametric scatter-plot smoother to help reveal any relation between state-level feature and individual-level effect. To construct a “reference distribution” under the null, I randomly re-shuffle the state-level education variable and create 19 new data sets that will have no systematic relation between this variable and the coefficient by repeating this process 19 times. Figure

6 below shows the 19 null plots based on this simulated data along with the true plot. Which one stands out?

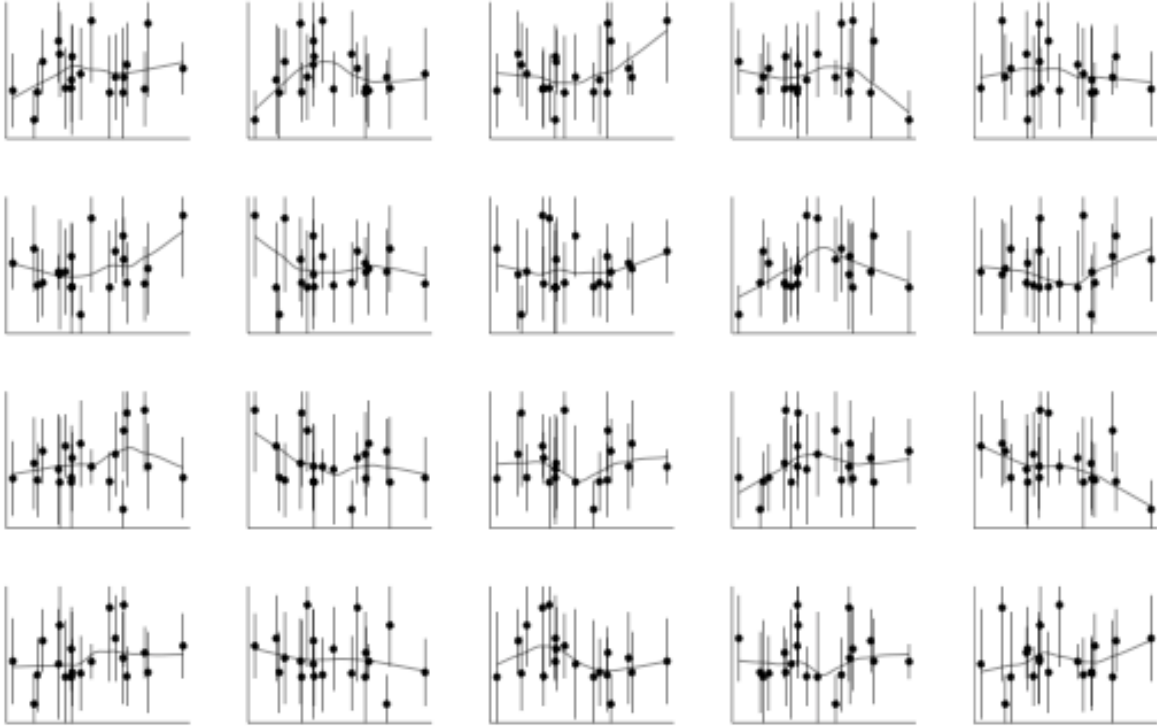


Figure 6: Line-up for the relation between the individual education effect on political participation (y-axis) and state-level education (x-axis).

No respondent in my sample – neither political scientist, nor crowd worker – managed to identify the plot showing the real data.⁵ With the resulting p-value of one, we clearly cannot reject the null hypothesis that individual educational effects are unrelated to state-level education.

⁵ The true plot is in row three and column two.

c) Clustered Data: World Values Survey Cultural Map

The next example comes from political culture research and is inspired by the famous *World Values Survey Cultural Map* which displays value orientations related to human development and democracy for a range of societies across the globe (see for instance Inglehart & Welzel 2005). The „map“ is really a scatter plot that does not show geographic, but cultural proximity by plotting countries along two value dimensions derived by factor analysis. The dimension of so-called survival vs. self-expression values is plotted on the x-axis and the dimension of traditional vs. secular-rational values on the y-axis. In addition, countries are colored according to their cultural zone or civilizational heritage: African, Islamic, Latin American, South Asian, Protestant European, Catholic European, Orthodox, and English Speaking.

One finding of theoretical interest suggested by the plot (and indeed the literature, see Welzel et al. 2003), is that cultural zones form more or less distinct clusters with similar value orientations: culture matters. The question is whether this pattern is really systematic? The null hypothesis in this case would be that there are in fact no such civilizational clusters and that societies belonging to the same cultural zone do in fact not show similar survival vs. self-expression and traditional vs. secular-rational values. The reference null distribution can be constructed by a simple random permutation of the vector of cultural zones and thus the color of the dots in the scatter plot. Figure 7 presents 19 such null plots along with the true data plot. Can you pick the true cultural map?



Figure 7: Line-up for the relation between survival vs. self-expression values (x-axis) and traditional vs. secular-rational values (x-axis).

The true plot clearly stands out.⁶ Indeed, all of the political scientists and 90 percent of the crowd-sourced respondents correctly identified the observed cultural map, yielding a p-value of essentially zero. This allows us to reject the null hypothesis of no cultural value clusters around the world.

d) Spatial Data: Interviewer Behavior in the German Longitudinal Election Study

The final example comes from the *German Longitudinal Election Study* (GLES) a large-scale survey project of voter behavior based on face-to-face interviews of the general population (Schmitt-Beck et al. 2009). The involved researchers were worried that some of their

⁶ The true cultural map is in row two and column four.

interviewers might selectively contact households, e.g. avoid low income areas and/or areas with high shares of foreigners where it could be hard to obtain successful interviews. The final methodological report provided by the polling firm did not give any information on this problematic interviewer behavior. What would we expect to see if interviewers indeed were to avoid certain areas? Clearly, selective interviewer behavior would show up as some kind of spatial pattern: interviewers would only visit and complete interviews in certain areas and avoid others. Consequently, the null hypothesis in this example states that there is no spatial pattern. Re-shuffling the variable vector for interviewer behavior in a data set with household addresses 19 times easily creates the respective null distribution. It breaks its relation to the location of sampled households given by longitude and latitude. A suitable “test statistic” would be a dot map which indicates the locations where interviewers failed to make contact. Figure 8 shows the line-up for the dot maps of interviewer behavior. Which one is the true map?

In my test sample roughly half of the impartial observers (53 percent, political scientists: 78 percent, crowd-workers: 30 percent) correctly identified the dot map showing the true interviewer behavior.⁷ This yields a p-value very close to zero and again suggests that we can reject the null that there is no spatial pattern in interviewer behavior.

⁷ The true dot map is in row three and column three.

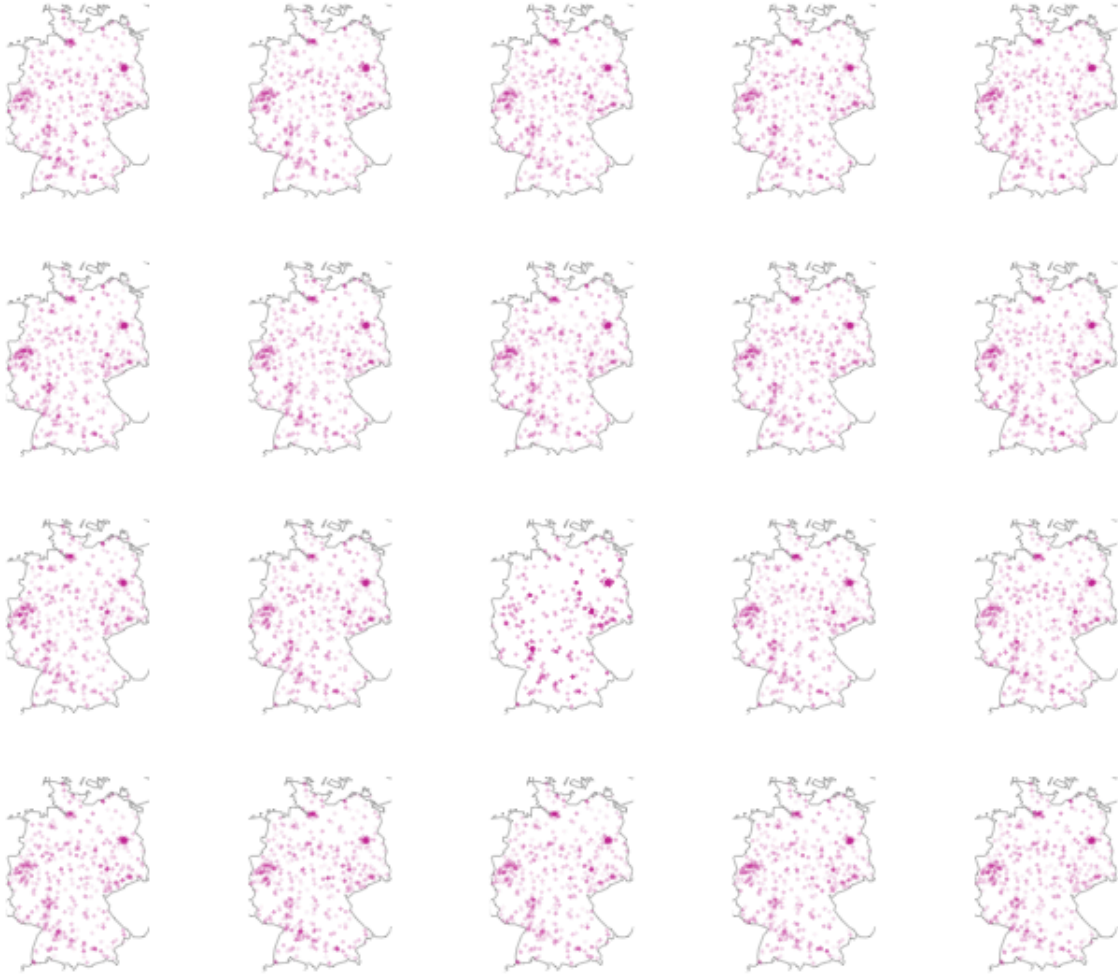


Figure 8: Line-up for the spatial pattern of interviewer behavior (“no contact”) in the German Longitudinal Election Study (GLES).

Conclusion

This paper introduces visual inference to political science and offers a remedy to the criticism frequently voiced against data visualization and exploration: that it may give rise to an over-interpretation of random patterns. By treating graphical displays as "test statistics" and comparing them to a "reference distribution" of plots under the assumption of the null, visual inference helps us answer the question "Is what we see really there?" By so doing, it seeks to overcome long-standing reservations against visualization as merely "informal" approach to data analysis and the fear that beautiful pictures may in fact not correspond to any meaningful patterns of substantive scientific interest.

References

- Alvarez, Michael R. (2016). Introduction. In: Alvarez, Michael R. (Ed.): *Computational Social Science. Discovery and Prediction*. Cambridge University Press.
- Anscombe, F. J. (1973): Graphs in Statistical Analysis. *American Statistician* 27(1): 17-21.
- Berry, Kenneth J., Mielke, Jr., Paul W., Johnston, Janis E. (2016). *Permutation Statistical Methods. An Integrated Approach*. New York: Springer.
- Best, L. A., Smith, L. D., & Stubbs, D. A. (2001). Graphs use in psychology and other sciences. *Behavioural Processes* 54, 155-165.
- Bowers, Jake (2004). Using R to Keep it Simple. Exploring Structure in Multilevel Datasets. *The Political Methodologist* 12: 17-24.
- Bowers, Jake and Drake, K. W. (2005). EDA for HLM: Visualization when Probabilistic Inference Fails. *Political Analysis* 13(4): 301-326.
- Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne and Hadley Wickham (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A* 367: 4361-4383.

- Cleveland, William S. (1993). *Visualizing Data*. Summit, NJ: Hobart Press.
- Cleveland, William S. (1994). *Elements of Graphing Data*. Revised edition. Summit, NJ: Hobart Press.
- Few, Stephen (2009). *Now you see it. Simple Visualization Techniques for Quantitative Analysis*. Oakland: Analytics Press.
- Fox, Jonathan (2008). *A World Survey of Religion and the State*. Cambridge University Press.
- Gelman, Andrew (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* 71: 369–382.
- Gelman, Andrew (2004). Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics* 13(4): 755-779.
- Gelman, Andrew and Jennifer Hill. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin (2013). *Bayesian Data Analysis, 3rd Edition*. Chapman & Hall/CRC Texts.
- Gill, Jeff (1999). The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly*, 52(3), 647-674.
- Good, Philipp (2005). *Permutation, parametric, and bootstrap tests of hypotheses* (3rd edn). New York: Springer.
- Healy, Kieran and James Moody. Data visualization in sociology. *Annual review of sociology* 40: 105-128.
- Hofmann, H., Follett, L., Majumder, M., & Cook, D. (2012). Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics* 18(12): 2441-2448.
- Inglehart, Ronald and Christian Welzel (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.
- Jacoby, William G. (1997). *Statistical Graphics for Univariate and Bivariate Data*. Thousand Oaks: Sage.
- Jacoby, William G. (1998). *Statistical Graphics for Visualizing Multivariate Data*. Thousand Oaks: Sage.
- Jacoby, William G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies* 19: 577-613.
- Jacoby, William G. and Sandra K. Schneider (2010). *Graphical Displays for Political Science Journal Articles*. Unpublished Manuscript: Michigan State University.

- Kastellec, Jonathan and Eduardo Leoni (2007). Using graphs instead of tables in political science. *Perspectives on Politics* 5(4): 755–771.
- Kerman, J., Gelman, A., Zheng, T. and Ding, Y. (2008). Visualization in Bayesian Data Analysis. In: Chen, C., Härdle, W. K. & Unwin, A (Eds.): *Handbook of Data Visualization*. Berlin: Springer. 709-724.
- King, G., Tomz, M. and Wittenberg, J. (2001): Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science* 44(2): 341-355.
- Kruschke, John K. (2013). Posterior predictive checks can and should be Bayesian: Comment on Gelman and Shalizi, ‘Philosophy and the practice of Bayesian statistics.’ *British Journal of Mathematical and Statistical Psychology* 66: 45–56.
- Lynch, Scott M. and Western, Bruce (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods & Research* 32(3): 301-335.
- Majumder, Mahbubul, Heike Hofmann and Dianne Cook (2013). Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association* 108(503): 942-956.
- Chowdhury, N. R., Cook, D., Hofmann, H., Majumder, M., Lee, E. K., & Toth, A. L. (2015). Using visual statistical inference to better understand random class separations in high dimension, low sample size data. *Computational Statistics* 30(2): 293-316.
- Schmitt-Beck, R., Bytzeck, E., Rattinger, H., Roßteutscher, S., and Weßels, B. (2009). The German longitudinal election study (GLES). *Annual Meeting of International Communication Association (ICA), Chicago, 21, 25.*
- Stegmueller, Daniel (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science* 57(3): 748-761
- Tufte, Edward (2001). *The Visual Display of Quantitative Information*. (Second Edition). Graphics Press.
- Tukey, John W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.
- Welzel, C., Inglehart, R., and Kligemann, H. D. (2003). The theory of human development: A cross-cultural analysis. *European Journal of Political Research* 42(3): 341-379.
- Western, Bruce and Jackman, Simon (1994). Bayesian inference for comparative research. *American Political Science Review* 88(2): 412-423.
- Wickham, H., Cook, D., Hofmann, H., and Buja, A. (2010). Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics* 16(6): 973-979.
- Widen, Holly M., James B. Elsner, Stephanie Pau, Christopher K. Uejio (2016). Graphical Inference in Geographical Research. *Geographical Analysis* 48: 115–131