Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation

Roy G. Biv, Ed Grimley, *Member, IEEE*, and Martha Stewart

Abstract—Models of human perception — including perceptual "laws" — can be valuable tools for deriving visualization design recommendations. However, it is important to assess the explanatory power of such models when using them to inform design. We present a secondary analysis of data previously used to rank the effectiveness of bivariate visualizations for assessing correlation (measured with Pearson's r) according to the well-known Weber-Fechner Law. Beginning with the model of Harrison *et al.* [1], we present a sequence of refinements including incorporation of individual differences, log transformation, censored regression, and adoption of Bayesian statistics. Our model incorporates all observations dropped from the original analysis, including data near ceilings caused by the data collection process and entire visualizations dropped due to large numbers of observations worse than chance. This model deviates from Weber's Law, but provides improved predictive accuracy and generalization. Using Bayesian credibility intervals, we derive a partial ranking that groups visualizations with similar performance, and we give precise estimates of the difference in performance between these groups. We conclude with a discussion of the value of data sharing and replication, and share implications for modeling similar experimental data.

Index Terms—TBD

Introduction

TBD, motivation, blah blah blah. ….

Models of human perception — sometimes called laws — are valuable tools for deriving concrete recommendations in visualization design. However, we must be careful to assess the explanatory power of such models before putting them into practice, particularly if they do not model variance in individual performance. We present a secondary analysis of data from Harrison et al. [1], investigating the relationship between the correlation of two variables (measured using Pearson's r) and the precision of people's estimates of that correlation using different visualizations (measured using just-noticeable differences). We begin with a linear model — similar to that of the original work — and walk through a series of refinements to this model. We first address problems of non-constant variance, presenting evidence that a log-linear — rather than linear — model better describes the relationship between just-noticeable differences and objective correlation. We then augment our model with censored regression to include all observations in the analysis, including outliers, data near ceilings and floors resulting from features of the data collection process, and entire visualizations originally dropped due to large numbers of data points worse than chance. Finally, we adopt a Bayesian variant of our model to derive a partial ranking of visualizations of correlation based on the expected precision of people's estimates of correlation on a randomly-drawn dataset. This partial ranking provides concrete guidance to practitioners by grouping visualizations with similar performance and by giving precise estimates of the difference in performance between groups of visualizations. We discuss the applicability of similar models to other problems of estimating the perceptual performance of visualizations from experimental data.

The original paper used a modelling approach that removed individual variation before fitting the model: it fit a linear regression to the means of the just-noticeable differences within each visandsign * r * approach, not to the individual observations directly. By removing a large portion of the variance in the data (individual differences), they could not use their parametric model to make predictive inferences. Instead, they employed non-parametric tests to examine differences between visualization types, which complicates the estimation of effect sizes. Even if we establish that one visualization is better than another, we would like to know by how much in order to judge whether the difference is meaningful in practice. Ideally we would like to know this effect size on some interpretable scale (such as in terms of just-noticeable differences in r), which is made difficult when using non-parametric tests.

Much of this paper focusses on understanding and accounting for individuals' differences in precision of estimation in order to derive parametric models that can predict the expected precision of each visualization technique. Given an appropriate parametric model, we can estimate interpretable differences between visualization types (e.g., as a ratio of just-noticeable differences) in order to judge whether these differences have practical significance. To derive such a model, we first tweak the original linear model to fit it to individual observations. Through a series of refinements, we construct a censored, log-linear regression model that improves on the fit of the linear model and accounts for individual differences and artifacts of the experimental design.

This model allows us to directly and quantitatively answer questions left largely unaddressed by the original paper: given a dataset with unknown correlation, how well would we expect each visualization technique to perform? What are the practical differences in performance? Which visualizations are effectively equivalent? We identify clusters of visualizations with similar precision and quantify the expected difference in precision between clusters, yielding a comprehensive set of practical recommendations in the form of a partial ranking of visualizations of correlation.

In the rest of this paper, we briefly overview the original experimental design and model from Harrison et al. [1]. We then walk through a series of models: a linear model based on the original paper, a log-linear model addressing problems of non-constant variance and skewed residuals, a censored model addressing floors and ceilings in the data caused by artifacts of the experiment, and a Bayesian model that can estimate the performance of each visualization on a class of datasets. From this last model we derive a partial ranking of visualizations of correlation.

- *Roy G. Biv is with Starbucks Research. E-mail: roy.g.biv@aol.com.*
- *Ed Grimley is with Grimley Widgets, Inc.. E-mail: ed.grimley@aol.com.*
- *Martha Stewart is with Martha Stewart Enterprises at Microsoft Research. E-mail: Martha.stewart@marthastewart.com.*

- *Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.*

For information on obtaining reprints of this article, please send *e-mail to: tvcg@computer.org.*

1 Background

- Should briefly describe original paper's experimental setup
- Might be some useful pieces about perceptual laws to put in here. Might also be out of scope for this.

2 Model 1: Linear Model

Harrison *et al.* [1] used an analysis approach common in prior work on Weber's law, such as Rensink & Baldridge [2] (also Stevens' Power Law). First, they took the mean of all individual observations of JND within each condition (where each condition is defined as a unique combination of visualization × direction × approach × *r*). They then modelled the relationship between the value of *r* and that within-condition mean JND. Thus, their model can describe the relationship between the mean performance of a group of people from the population, but not the performance of any individual.

What this omits is any sense of the variance in individual performance, which diminishes the explanatory power of such models. For example, it may be that visualization *A* exhibits high precision of estimation (low JND) in the average case — but that its variance is higher than visualization *B*, which performs slightly worse on average but which is more consistent across individuals. Without considering variance, we have no way of knowing whether such differences exist, and we may be led (for example) to choose to deploy a visualization that has slightly better average-case performance but which elicits much worse performance for some substantial portion of the population. From a design perspective, this is not unlike an architect who designs every home for the average family of 2.6 people. Individuals, not group means, digest visualizations.

Such an analysis also obscures problems with model fit by discarding large portions of the variance (essentially all individual variation) and reducing a large sample of data to comparatively few data points. This explains why Harrison *et al.* [1] (like Rensink & Baldridge [2]) found very high R^2 values describing the fit of their models (as high as 0.98 for one condition). But when we attempt to interpret these values of R^2 — for example, as the percent of variation explained by the model — something is missing. 98% of individual variation is not explained by this model, as individual variation was discarded before the model was fit. We might instead interpret this as indicating 98% of the variation in the location of the mean was explained, but this is a much less useful thing to know if we wish to understand how *individuals* perceive visualizations. As we will see below, if we try to fit linear models to individual observations directly, the linear model does not exhibit the best fit.

Finally, because of the poor fit of the linear model to individual observations, we cannot use the fitting error to estimate significant differences between conditions. Indeed, despite being a paper that proposes a parametric model to describe the performance of each visualization, Harrison et al. do not use their parametric models to estimate differences between conditions; they use the nonparametric Wilcoxon rank-sum test. In this paper we propose a model of sufficient specificity that parametric estimation of differences becomes straightforward; this allows us to not only examine the differences between conditions but to clearly describe the expected magnitude of those differences (i.e., effect sizes) using parameters from the model. By employing parametric models, we have the advantage of interpretable effect sizes — for example, ratios of just-noticeable differences, from which we can say, "*visualization A* is *x* times more precise than *visualization B*" — that are not easily gleaned from nonparametric tests.

2.1 Incorporating individual differences

A first pass at incorporating individual differences would be to simply use a linear regression of the JND based on *r*. Such a model might look like:

$$\begin{aligned} y_{i,v} &= \beta_{v,1} + \beta_{v,2} r_i + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_v^2) \end{aligned}$$

This is a fairly standard linear regression. Here we say that, for each visualization × direction pair *v*, each JND ($y_{i,v}$) is equal to a linear function of r_i with intercept $\beta_{v,1}$ and slope $\beta_{v,2}$ plus some normally-distributed error ϵ_i. Note that the intercept, slope, and variance of the error (σ_v) are all dependent on *v,* the particular visualization × direction pair.

Unfortunately, this straightforward model leaves out consideration of *approach* — half of the JNDs were determined by a procedure having people compare the reference *r* to higher values of *r* (an approach *from above*), and half compared to lower values of *r* (*from below*). When the approach is from above, the values of JND are underestimated (because higher values of *r* tend to have lower JND), and when the approach is from below, JND is overestimated. This effect is visible in Figure X: note the two roughly-parallel (but systematically different) estimates of JND depending on approach. Harrison et al. used the correction described by Rensink & Baldridge [2] to address this: they adjusted the value of *r* by moving it up by half the mean JND at that value of *r* when from above, and down by half the mean JND when from below.

However, this adjustment is only well-defined if we are using the within-condition means of *r* as our unit of analysis. When fitting a model to individual observations, we must find another way to account for approach. Consider Figure X: because each condition causes a bias in the opposite direction, we could take the average of the two fit lines to approximate the outcome *y* for each *r* (the red line in Figure X). Such a model can be fit by including *approach* and its interaction with *r* in the regression. We code *approach* as a sum-to-zero contrast, defined by the variable a_i:

$$a_i = \begin{cases} -1, & \text{if } approach \text{ is } from\ above \\ 1, & \text{if } approach \text{ is } from\ below \end{cases}$$

We then add the effects of *approach* and *approach* × *r* to the model (new terms in red):

$$\begin{aligned} y_{i,v} &= \beta_{v,1} + \beta_{v,2} r_i + \beta_{v,3} a_i + \beta_{v,4} a_i r_i + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_v^2) \end{aligned}$$

Because these are sum-to-zero contrasts, the original slope ($\beta_{v,1}$) and intercept ($\beta_{v,2}$) for *r* are defined with respect to the mean of the two levels of a_i — in other words, the slope and intercept describe the mean of the slope and intercept of the *above* and *below* levels, exactly the red line in Figure X.

We fit the linear model described above to all individual observations; the results are in Figure X. This and all other non-Bayesian models in this paper were fit using the `gamlss` procedure in R.

2.2 Problems with the linear model

The linear model exhibits several issues of fit that indicate violations of assumptions of the model, illustrated in Figure X. Two such issues in particular are *non-constant variance* and *skewed residuals*, both of which are violations of the assumptions inherent in the distribution of the error term ϵ_i.

Non-constant variance (heteroscedasticity). As defined in the model, the variance of the error, σ_v^2, is constant with respect to *r*. That is, for a given visualization × direction pair *v*, the variance of *y* (the JND) is assumed to be the same no matter what the value of *r* is, nor what the predicted mean JND is. Thus, when the predicted mean JND is high, the variance should be the same as when the predicted mean JND is low. However, as we can see in Figure X, when the predicted JND is low, the variance of the residuals (actual observations minus the prediction) is lower than otherwise. This problem is actually quite common in data with a well-defined lower bound: here, JND cannot be less than 0, and as we approach 0, performance tends to cluster together more tightly.

Skewed residuals. Data with a lower bound also often exhibits the second model violation seen here: skewed residuals (more generally,

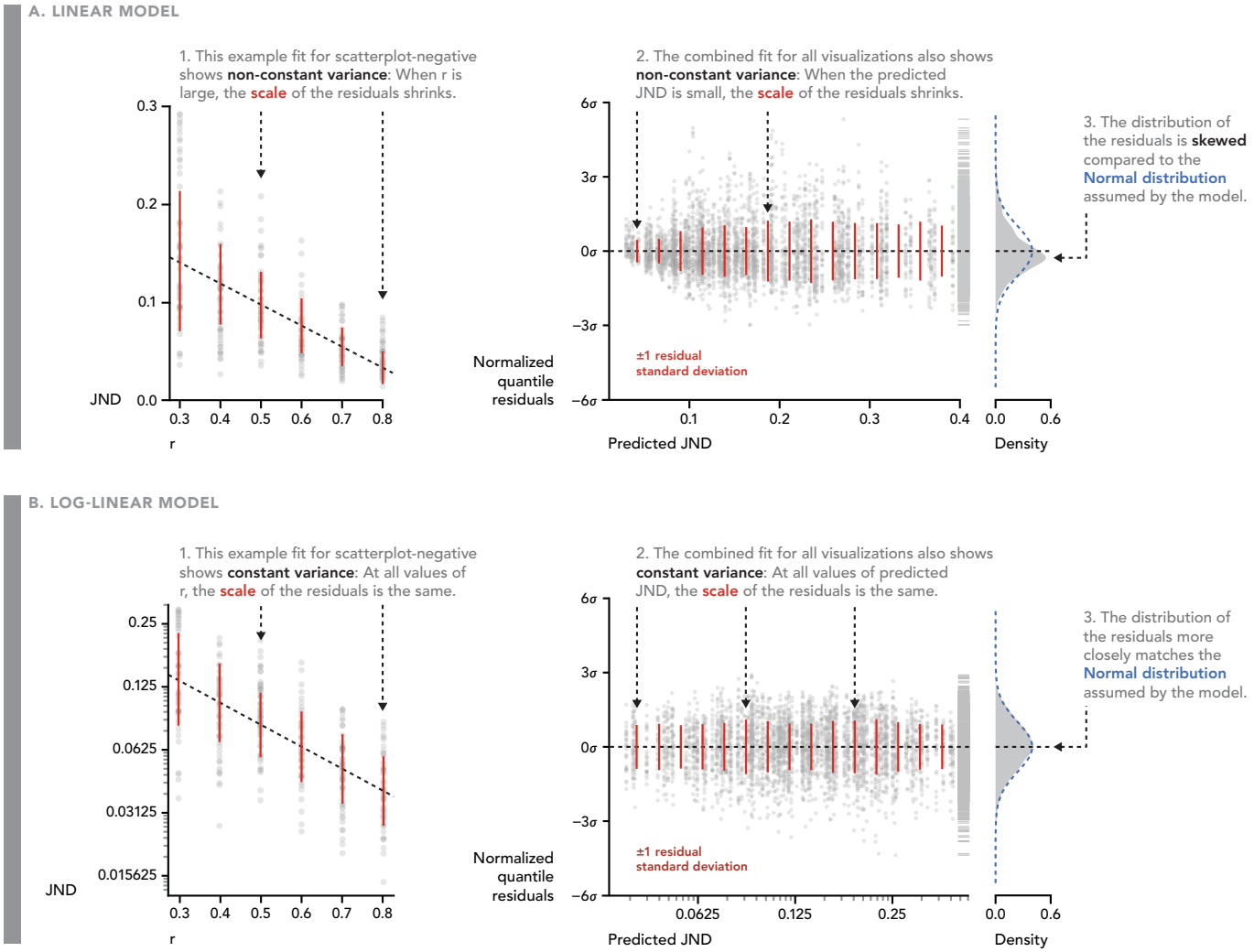

Fig. 1 This will be a comparison of linear and log-linear model fits (currently just linear fit shown)

non-normal residuals). We can think of JND as "bunching up" the closer it gets to 0; besides resulting in less variance, this also explains the skew in the residuals seen in Figure X. The residuals do not follow a normal distribution, which is not unexpected given the bounded nature of the data. While it is sometimes the case that we can get away with assuming bounded data is normally-distributed, such simplifications tend to break down the closer we get to the boundaries; here, the assumptions are clearly violated and suggest we should consider other models.

3 Model 2: Log-Linear Model

Fortunately, a log transformation of the response is often sufficient in cases of non-constant variance and skewed residuals to solve both problems simultaneously, and often shows up in models of human performance. The applicability of such a transformation is hinted at here as the residual distribution has the approximate appearance of a log-normal distribution.[1] This transformation also has the useful property that the resulting model retains some interpretability: coefficients of this model that describe additive differences on the log-scale correspond to multiplicative differences on the original data scale (in other words, we will be able to use this model to make claims like, "*visualization A* yields x times the precision of *visualization B* for estimating correlation"). The log-linear model is as follows (differences from previous model in red):

$$\begin{aligned} \log(y_{i,v}) &= \beta_{v,1} + \beta_{v,2} r_i + \beta_{v,3} a_i + \beta_{v,4} a_i r_i + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_v^2) \end{aligned}$$

Comparing the residual fit of the log-linear model to the linear model (Figure X), we can see that the fit no longer suffers from problems of non-constant variance or skewed residuals. This fit also exhibits lower AIC than the linear model (-11683 versus -10037), indicating greater predictive validity.[2] The residual distribution more closely matches the normal distribution assumed by the model. In addition, because all values in $(-\infty, +\infty)$ are mapped onto $(0, +\infty)$ by the log transformation, we have solved another problem not previously discussed: the linear model can make nonsensical predictions, such as JNDs that are less than 0, that the log-linear model does not.

[1] We can more systematically justify this transformation by fitting a Box-Cox transformation to the data, whose parameter λ describes a power transformation of JND. The Box-Cox procedure for this data estimates $\lambda = 0.0292$ with a 95% confidence interval of [-0.005, 0.0635], which includes 0 (the log transform) and excludes 1 (identity; i.e. the linear model) at $p < 0.00001$ (LR $\chi^2(1) = 2756.77$).

[2] Model comparison by AIC is asymptotically equivalent to leave-one-out cross validation [3]. The log-linear model was fit using a log-normal error distribution (rather than the equivalent log transformation of responses with a normal error distribution shown here) so that its AIC can be compared to the linear model.

3.1 Data dropped from the analysis so far

So far, we have restricted our analyses to those data points analysed in the original work. The original work used two criteria to exclude data from analysis:

Outliers. Within each condition, observations outside of 3 median absolute deviations from the median were dropped from the analysis. The original paper justified this as a way to address non-normality in the data (although as we have seen above, it did not). Since we have addressed the issue of normality through log transformation, this criteria is no longer particularly relevant. Since our goal is to explain as much of the data as possible, we believe there is no additional need to drop outliers from the analysis.

Data worse than chance. In the original work, visualization × direction pairs with more than 20% of JNDs greater than 0.45 were dropped (6 out of 18 pairs). The 0.45 threshold represents the *chance* threshold for this experiment: values of JND near or beyond this threshold indicate a failure on a participant's part to judge degree of correlation better than could be done by answering at random. However, removing visualizations with large numbers of observations worse than chance addresses only part of the problem. As can be seen in Figure X, many of the remaining tested visualization × direction pairs still have observations at or beyond the chance boundary. The problem is that we have excluded certain visualization × direction pairs for having too many observations worse than chance, but have done nothing to address those observations worse than chance that remain in the visualizations we *do* analyse.

Importantly, in the case of points near or beyond this boundary, we can say that these observations probably represent JNDs of .45 or worse, but that we do not know the exact JND due to the constraints of the experiment. This type of data can be analysed using censored regression.

4 Model 3: Censored Log-Linear Model

Censored regression can be used when some of the observed data points do not have a known value, but instead are known to lie above (or below) a certain threshold. While we do not know the exact value of points beyond the threshold, we still know how many points were observed beyond the threshold, and it is this information that we can use to fit the model. In this case, while we cannot reliably observe certain values of JND — either because the setup of the experiment makes them indistinguishable from chance, or because of ceilings and floors in observable JND due to the bounds on r — we can use observations close to or beyond those thresholds to estimate the proportion of values we might expect to see above them.

- Ceiling in JND (when approach is from above) and floor in JND (when from below) also candidates for censoring. Deriving thresholds

- Include example of censored distribution as explanation?

- Show example of downward bias at low r in a viz near the chance threshold to demonstrate value of censored models

To incorporate censoring into our model, we first define a censoring threshold, $c_{i,v}$. This threshold varies depending on r and the approach, (see Figure X):

$$c_{i,v} = \begin{cases} \min(0.95 - r_i, 0.4), & a_i = -1 \\ \min(r_i - 0.05, 0.4), & a_i = 1 \end{cases}$$

We use the log-linear model to predict a latent variable y^* instead of y (differences from previous model in red):

$$\log(y^*_{i,v}) = \beta_{v,1} + \beta_{v,2} r_i + \beta_{v,3} a_i + \beta_{v,4} a_i r_i + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma_v^2)$$

We redefine y as being equal to the censoring threshold at the corresponding value of r if its observed value is greater than that threshold.[3] The model then predicts y based on the latent variable y^* and the censoring threshold c:

$$y_{i,v} = \begin{cases} y^*_{i,v}, & y^*_{i,v} \le c_{i,v} \\ c_{i,v}, & y^*_{i,v} > c_{i,v} \end{cases}$$

4.1 Bias in uncensored model

The censored model allows us to address problems of bias caused by JND being underestimated near the ceilings described above. See Figure X, which compares the censored and uncensored models for visualization XXX. Note that where large amounts of observations are worse than chance, the uncensored model estimates people as having *higher* precision (lower JND) than we should expect! This bias conspires to make a low-performing visualization seem better than it is, motivating our use of censored regression here. This underscores the problem with excluding some visualization × direction pairs based on the chance criteria without accounting for chance in the pairs we do analyse.

5 Model 4: Bayesian Censored Log-Linear Model

- Derive priors
- Demonstrate model
- Possibly add participant effect (though probably not --- it mostly just complicates things at this point without gaining much)
- Possibly talk about differences in variance
- Simulate drawing random distribution to derive expected differences given unknown correlation

6 Partial Ranking of Visualizations

- Given a problem space with datasets having some known/estimated distribution of r, easy to re-compute rankings from the model (possibly put in discussion)

7 Discussion

- Censoring probably useful in other similar experiments with thresholds that are artifacts of the data collection process.
- Might be something in here about perceptual "laws", model fitting, individual differences, etc
- Other approaches could have tried --- other transformations, other error distributions, truncation. None fit as well or were as parsimonious as the log transformation.
- Reproducibility, code, data

[3] As a result of this transformation of the responses and the inclusion of data not included in previous models, the censored model cannot be compared to the previous models using AIC. However, we believe the theoretical justification based on ceilings caused by the structure of the experiment and the ability of these models to accommodate data dropped previously motivate the use of censored regression here, and for visualization × direction pairs far from those ceilings the fit is similar to the non-censored log-linear model.

8 Conclusion

Acknowledgments

- Original authors?

References

[1] L. Harrison, F. Yang, S. Franconeri, and R. Chang, "Ranking Visualizations of Correlation Using Weber's Law," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1943–1952, 2014.

[2] R. A. Rensink and G. Baldridge, "The perception of correlation in scatterplots," *Comput. Graph. Forum*, vol. 29, no. 3, pp. 1203–1210, 2010.

[3] M. Stone, "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 44–47, 1977.