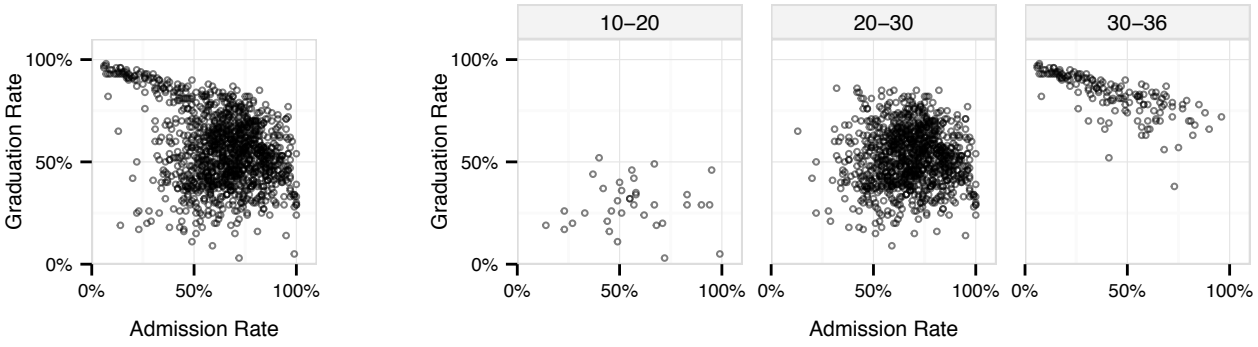
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2015.2467323, IEEE Transactions on Visualization and Computer Graphics

Automatic Selection of Partitioning Variables for Small Multiple Displays

Anushka Anand and Justin Talbot



(a) Input scatterplot

(b) Output small multiple display partitioned on admit ACT scores

Fig. 1: On the left is an input plot showing the complex relationship between admission and graduation rates at US universities. On the right is the top ranked small multiple display automatically picked by our algorithm to help explain this data. It partitions the data on aggregate admit ACT scores, revealing that for universities with very high ACT scores, there is a strong linear relationship between selectivity and graduation, while for other universities, there is no clear relationship.

Abstract— Effective small multiple displays are created by partitioning a visualization on variables that reveal interesting conditional structure in the data. We propose a method that automatically ranks partitioning variables, allowing analysts to focus on the most promising small multiple displays. Our approach is based on a randomized, non-parametric permutation test, which allows us to handle a wide range of quality measures for visual patterns defined on many different visualization types, while discounting spurious patterns. We demonstrate the effectiveness of our approach on scatterplots of real-world, multidimensional datasets.

Index Terms—Small multiple displays, Visualization selection, Multidimensional data.

1 Introduction

Understanding multidimensional data sets is a common challenge in Exploratory Data Analysis [39]. Many techniques have been proposed for visualizing multidimensional data in 2D. Perhaps the two most common techniques are *projective displays*, such as scatterplot matrices (SPLOMs), which consist of a set of 2D projections of the data, and *small multiple displays* (also called collections or trellis displays) [5, 38, 4], which show 2D slices of the data created by partitioning on one or more variables.

Unfortunately, as the number of variables in the data set grows, neither approach scales well since the number of plots that must be displayed increases quickly. This problem can be addressed by showing only the subset of "interesting" variables. However, if the user does not know a priori which variables might be of interest, finding them can be time-consuming since the user must manually iterate through all the variables to find views that help explain their data.

In the context of projective displays, there has been substantial work in automating this process. John and Paul Tukey suggested *cognostics*, visualization metrics that permit computers to "judge the relative interest of different displays" [41]. Perhaps the best known of these are *scagnostics*, cognostics designed to help find interesting scatterplots in a SPLOM. However, there has been little corresponding work in the automatic selection of interesting partitioning dimensions for small multiple displays. In this paper, we address this problem.

To motivate our work, consider the scatterplot in Figure 1(a) which shows the surprisingly complex relationship between acceptance and graduation rates at US universities [28]. In Figure 1(b), a small multiple display partitioned on aggregate ACT scores for admits at each university helps explain this pattern. For universities with low ACT scores (in the range 10-20), the graduation rate is low regardless of the admission rate. For universities with mid-range ACT scores (20-30), there is no clear correlation. And for universities with very high ACT scores (30-36), there is a nearly linear relationship between the acceptance rate and the graduation rate. Thus, the relationship between acceptance and graduation rates is strongly mediated by admit ACT scores. Our goal is to devise a method that examines the variables in a data set and automatically recommends ones, such as the ACT score in this example, that create effective small multiple displays.

In this paper, we:

- describe a set of goodness criteria for evaluating small multiple displays,
- develop a method for measuring the quality of partitioning variables based on a randomized, non-parametric permutation test which allows us to incorporate a wide range of existing cognostics for single visualizations, while adjusting for their pattern detection sensitivities, and
- demonstrate that our method selects small multiple displays that meet our goodness criteria.

- *Anushka Anand is with Tableau Research. E-mail: aanand@tableau.com.*
- *Justin Talbot is with Tableau Research. E-mail: jtalbot@tableau.com.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2015.2467323, IEEE Transactions on Visualization and Computer Graphics

We focus in this paper on small multiples based on scatterplots and use scagnostics as our cognostics, but our method could be applied to other view types by using corresponding cognostics.

The next section summarizes related work on small multiples, cognostics, and non-parametric approaches in visualization. Then we describe our criteria and our method for ranking small multiple displays. We then validate our method against our criteria. We conclude with a discussion of challenges and future work.

2 Previous Work

Our work draws on previous work in three areas of visualization—small multiple displays, cognostics, and the use of non-parametric statistics to improve visual analytics.

2.1 Small Multiple Displays

Small multiple displays are tables of similar visualizations, where each cell visualizes a subset of the data. Such displays allow viewers to make visual inferences about the conditional impact of the partitioning variable(s). Use of small multiples dates back to the work of economist W. S. Jevons in the 19th century [21] who used them to transform tables of time series data into rich graphical displays.

Today, many popular visual analysis tools can generate small multiple displays, such as the Trellis package in S-Plus [3], the ggplot2 library for the R language [43], based on Wilkinson's Grammar of Graphics [46], and the Polaris system (now Tableau) [36]. These systems allow users to rapidly generate small multiple displays to explore their data. Mackinlay's APT system [23] and Tableau's Show Me system [24] implement heuristics for automatically laying out effective small multiple displays based on the data types and functional dependencies in a data set. These tools all require users to manually select the partitioning variables.

Small multiples can also be employed as a visual layout metaphor in user interfaces for exploring the input space of encoding parameters as in Design Galleries [26]. The alternating use of small multiples together with a large single view has also been used an an interaction device for data exploration [42].

2.2 Cognostics

Cognostics and scagnostics were first proposed by Tukey and Tukey [40, 41]. More computationally efficient scagnostics were proposed by Wilkinson et al. [46, 47]. Other scagnostics have been developed to capture properties such as cluster separation [33, 37], class consistency and separation [35, 30], or statistically-motivated measures [20, 34, 29].

Cognostics have also been developed for other plot types. Many authors have suggested quality measures for parallel coordinate plots [2, 12, 19, 49]. Albuquerque et al. [1] offer measures for radial visualizations, pixel-oriented displays, and table lenses. Schneidewind et al. [31] propose Pixnostics, a cognostic based directly on the pixel representation of a visualization. Some measures [6, 10, 49] focus on the level of abstraction, including aggregation, clustering, and sampling in various chart types. For more information on quality measures for visualization, consult the survey by Bertini et al. [7].

Some visual analytic systems leverage these measures to recommend visualizations to their users. For example, ScagExplorer [11] applies scagnostics to cluster and filter through large collections of bivariate relationships automatically. EvoGraphDice [8] uses evolutionary algorithms and a scagnostics-based fitness function to select interesting linear and non-linear 2D projections. AutoVis [48] uses scagnostics to provide users with effective summaries of their data, tuned to highlight patterns that professional statisticians would also identify. MacEachren et al. [22] use conditional entropy to identify pairs of variables in a high-dimensional dataset that are likely to display interesting relationships. These variables are displayed in a matrix of view types. Trelliscope [17] uses scagnostics to organize and filter the large number of panels that result from using a trellis display on complex data.

2.3 Non-parametric statistics in visual analytics

Graphical inference [9, 44, 25] asks viewers to judge whether a visualization of the actual dataset is visually distinguishable from random bootstrapped samples. The result is a non-parametric significance test of a visual pattern. Conversely, Menjoge [27] uses bootstrapping to generate a 95% visual confidence interval that can correctly communicate the sampling variability in a visual pattern.

3 Method

Our algorithm takes three inputs from the analyst:

1. a scatterplot which the analyst wants to partition into a small multiple display,
2. a scagnostic that measures the presence or absence of a visual pattern of interest to the analyst, and
3. a list of potential partitioning variables.

The output is a scoring of the small multiple displays produced by each partitioning variable.

To motivate our algorithm, we first describe four intuitive criteria for effective small multiple displays. We then describe an algorithm that incorporates these criteria to automatically score potential partitioning variables.

3.1 Goodness-of-Split Criteria

We hypothesize that effective small multiple displays conform to the following four criteria:

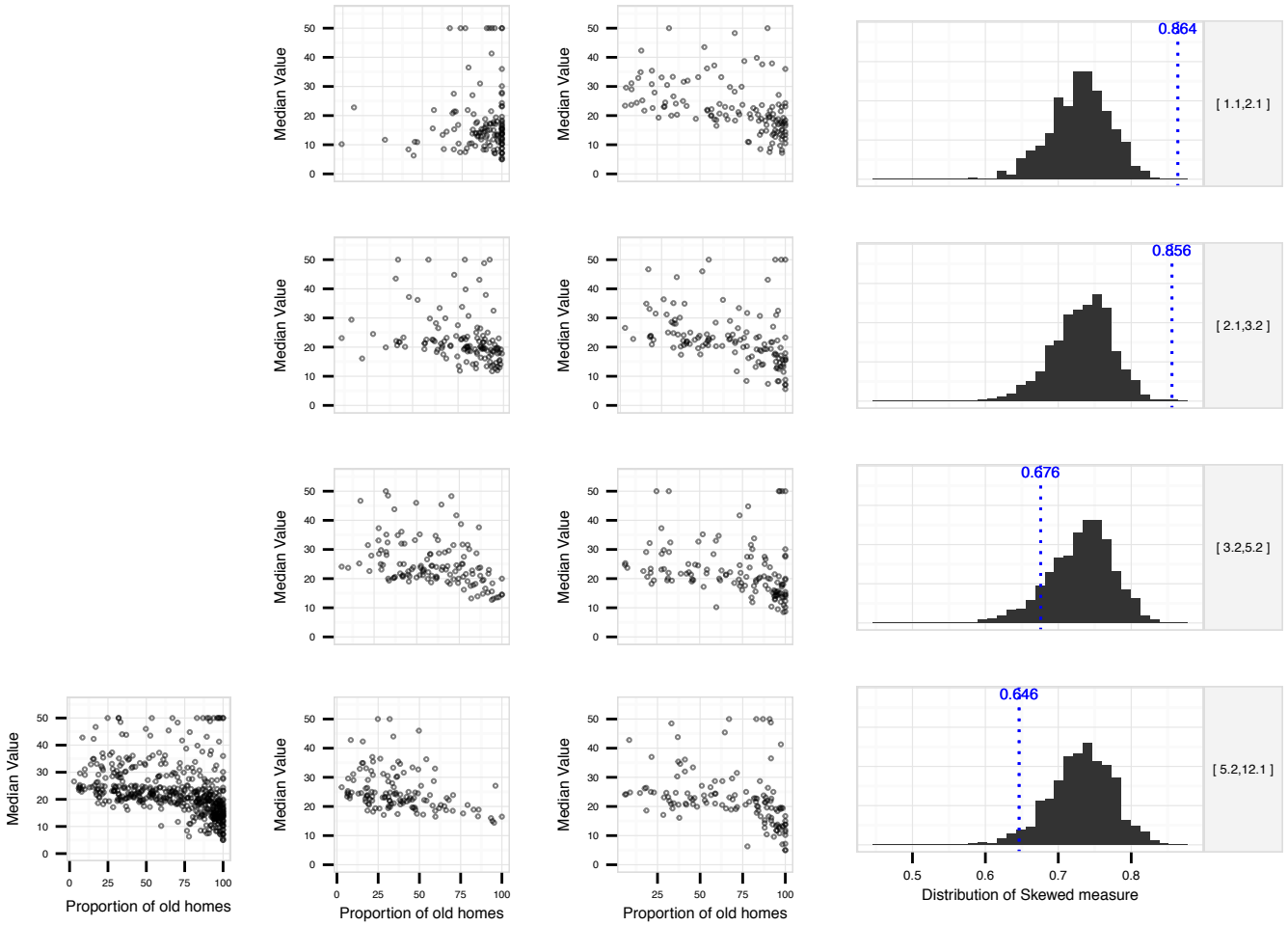
- *Visually rich*: Effective small multiple displays should leverage the capabilities of the human visual system by conveying rich visual patterns. This visual richness is unlikely to be captured by the relatively simple summary statistics used in common analytic methods such as ANOVA.
- *Informative*: Small multiple displays should be more informative than the input visualization, allowing the analyst to deepen their understanding of the data. Small multiple displays that randomly partition the input data are not useful since they contain no more information than the original plot.
- *Well-supported*: For some data sets, particularly those with outliers or with a small number of data points, strong visual patterns can occur by chance. These spurious patterns are misleading; they appear informative, but are not. Good small multiple displays should convey robust patterns, guiding analysts to reliable results.
- *Parsimonious*: A small multiple display with many partitions can be very difficult to read and understand. All things being equal, we should favor fewer partitions.

3.2 Algorithm

The key insight of this paper is that these four criteria can be achieved using a simple heuristic: select small multiple displays that have cognostic values that are unlikely to be due to chance. Using a cognostic to evaluate the small multiple displays ensures that we can find *visually rich* patterns. If the cognostic values are different from that of the input plot, then the small multiple is *informative*. If those differences are unlikely to be due to chance, then it is *well-supported*. And if there are redundant partitioning variables, this heuristic will lead to picking the most *parsimonious*.

The key challenge in our approach is determining the likelihood of a small multiple display's cognostic value. This is difficult because the underlying distribution of the cognostic score depends on both the cognostic algorithm itself and on the dataset, so we cannot evaluate the likelihood with a closed form formula. Instead, we propose computing this likelihood using a *randomized permutation test* [16], which is a non-parametric statistical significance test. This procedure builds an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2015.2467323, IEEE Transactions on Visualization and Computer Graphics



(a) Input scatterplot

(b) Partitioned by distance to an employment center

(c) Partitioned by random permutation

(d) Distribution of Skewed scagnostics

Fig. 2: Illustration of our method of evaluating small multiple displays. (a) The input scatterplot of interest. (b) Partitions determined by the mean distance to Boston's five employment centers. (c) Randomly permuted partitions of the data. (d) Distribution of Skewed scagnostics for the randomly permuted partitions. The overlaid blue lines are the corresponding true scores of the partitions in (b). The blue lines are outliers, indicating that they likely did not arise due to chance. Our algorithm will score the small multiple display in (b) highly.

empirical approximation of the underlying cognostic distribution by repeatedly permuting the partitioning variable randomly and computing cognostic values for these random partitions.

To demonstrate how this approach works, consider Figure 2(a) which shows the relationship between the median value of owner-occupied houses (in thousands of US dollars) and the proportion of such houses built prior to 1940 for census tracts in the area of Boston, Massachusetts [18]. We see that as the proportion of older houses increases, the median value decreases. However, the distribution is skewed and an analyst may wonder if partitioning this scatterplot by another variable in the dataset that might reveal more about this relationship. To do so, they run our algorithm using Wilkinson et al.'s *Skewed* scagnostic which can detect a wide range of skewed patterns in scatterplots.

For each possible partitioning variable, we need to compute how unlikely it is that the resulting *Skewed* values are due to chance. For example, consider the "distance to employment center" variable shown in Figure 2(b). This variable produces a small multiple display with four plots, one for each quartile of the distance from the census tract to the nearest employment center. Visually we can see that the top two plots are quite skewed and this is verifed by the *Skewed* scagnostic which produces values of 0.864 and 0.856 for the top two plots, but only 0.676 and 0.646 for the bottom two plots. These values are indicated with blue lines in Figure 2(d).

Next, to evaluate how unlikely it is that these scores arose due to chance, we randomly permute the assignment of data points to partitions, resulting in Figure 2(c). Visually the strong skewness of the top two plots has decreased. This indicates that the true scores are probably not just due to chance. If we randomly permute the data set 1000 times, computing the *Skewed* values for each permutation, we can construct the histograms shown in Figure 2(d) which show the likelihood that a particular *Skewed* score will occur by chance for each component plot of the small multiple.

By comparing the true *Skewed* values (blue lines) to the random results (black histograms), we can see that the top two plots are in fact more skewed than nearly all the random permutations. Furthermore, the bottom two plots are less skewed than many random permutations. This provides strong evidence that the visual patterns seen in Figure 2(b) are not just due to chance, but are rich, informative, and well-supported patterns. In this case, the small multiple display offers a visual explanation of the relationship between employment center locations to the neighborhoods in Boston—areas with older, lower-valued homes are close to employment centers while newer, higher-valued homes tend to be further away.

To compare this partitioning variable to others, we have to summarize this information into a single score. We first need to summarize how unlikely each component plot is. A straightforward non-parametric way to do this would be with order statistics (a count of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2015.2467323, IEEE Transactions on Visualization and Computer Graphics

how many of the random *Skewed* values were more extreme than the true values). However, in practice we found that often the true value would lay well outside the range of the empirical random distribution. To generate useful likelihood values in these cases would require fitting an analytic distribution to the data. But we also want our approach to work with any cognostic measure, thus we have no a priori information about what analytic distribution we should use. So, instead we use Chebyshev's inequality which can give us a very conservative bound on the likelihood with only weak assumptions about the underlying distribution. This bound is inversely proportional to the standard z-score, so minimizing the likelihood is equivalent to maximizing the absolute z-score:

$$|z_i| = \left| \frac{(X_i - \mu_i)}{\sigma_i} \right|$$

where X_i is the true cognostic value of the i-th partition and μ_i and σ_i are the mean and standard deviation of the cognostic measures over the repeated random permutations of the i-th partition.

Finally, to get a score for the whole small multiple display, we use the maximum absolute z-score across all component plots:

$$z = \max_i |z_i|$$

Using the maximum will result in high scores for small multiple displays that have strong, interesting patterns in at least one component plot. This worked well in our experiments.

4 Validation

In this section, we demonstrate by example that our algorithm satisfies our goodness-of-split criteria for effective small multiple displays.

4.1 Visually Rich

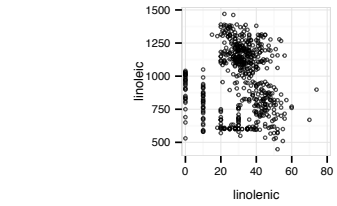
Our first criterion is to prefer small multiple displays that have visually salient patterns. Our algorithm achieves this by using the input cognostic to measure the presence and extent of a visual pattern in the component plots of a candidate small multiple display. This approach works for arbitrary cognostic measures on any type of visualization, allowing us to create small multiple displays targeted at the interests of the analyst.

Consider Figure 3(a), which shows the relationship between linolenic and linoleic in the olive oils dataset [14], which includes eight chemical measurements on different specimens of olive oil produced in various regions in Italy. There are visually striking clumps and striation patterns in the data. An analyst might wonder whether these patterns can be isolated and explained by any of the other variables in the dataset.

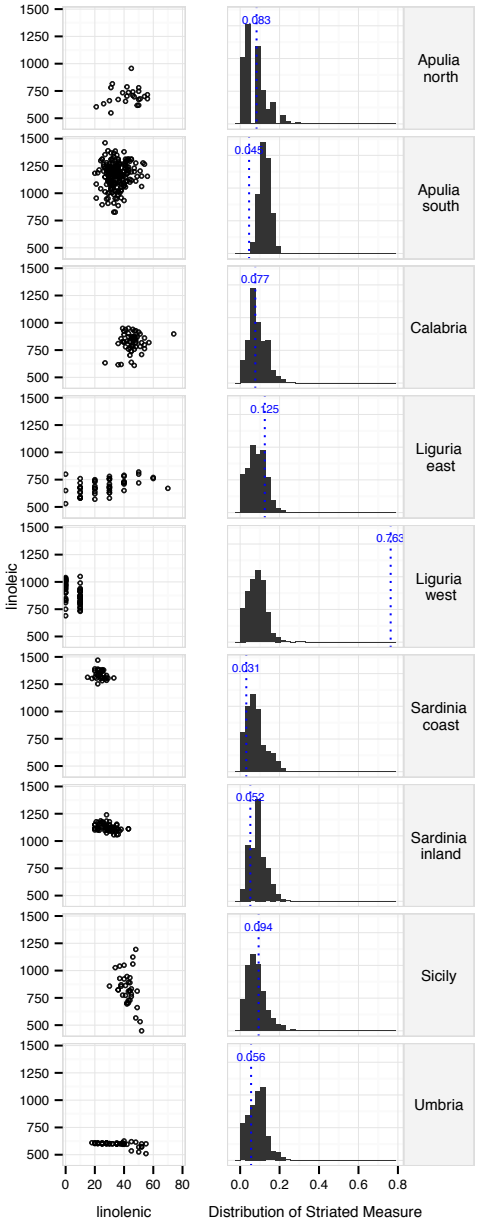
To do so, the analyst can use our approach with Wilkinson et al.'s *Striated* scagnostic which detects banding of points in a scatterplot. With this cognostic, our approach ranks the Region variable as the best partitioning variable for isolating striated patterns from the six remaining variables in the data set. This is the partitioning shown in the left half of Figure 3(b), which reveals the clean isolation of the striated pattern for olive oils from the Liguria region and the distinctive measurement structure of linoleic values for the Umbria region.

As before, the right side of Figure 3(b) shows the true Striated scores for this partitioning in blue and the randomized permutation scores in the black histogram. We can see that our algorithm has identified the striated pattern in Liguria west as being very unlikely to have arisen due to chance, leading to the top ranking for this small multiple display. Note, however, that the striation in Liguria east is not identified as being an outlier. This is because the Striated scagnostic itself fails to catch this case. Thus, the visual richness of the small multiples our approach selects depends on the quality of input cognostic, an issue we revisit in the discussion (Section 5).

The highest ranked partitioning variable changes with the choice of scagnostic. For the Boston dataset, the example in Section 3 shows the highest ranked variable ("distance to an employment center") on the *Skewed* scagnostic. Wilkinson et al.'s *Outlying* scagnostic detects the presence of shapes with a high number of outlier points. We see



(a) Input scatterplot

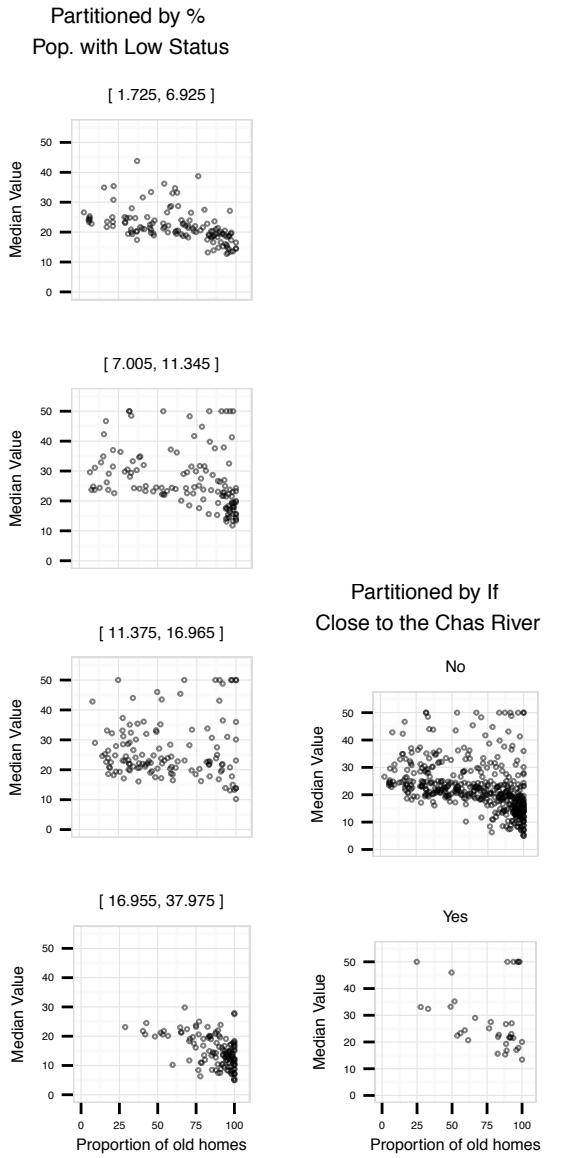


(b) Highest-ranked small multiple display, partitioned by region

Fig. 3: The highest ranked small multiple display of the olive oils data set using the Striated scagnostic. Our algorithm detects that the striation pattern in Liguria west is very unlikely to be due to chance and recommends this small multiple display.

evidence of this in the highest ranked Outlying small multiple seen in Figure 4(a) based on the variable, "% of the population with low status". The *Clumpy* scagnostic detects shapes with multiple dense regions or clumps of points. The most "clumpy" small multiple display is determined by the boolean variable "close to the Chas river"

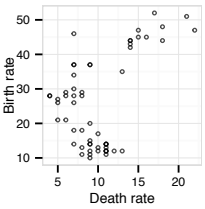
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2015.2467323, IEEE Transactions on Visualization and Computer Graphics



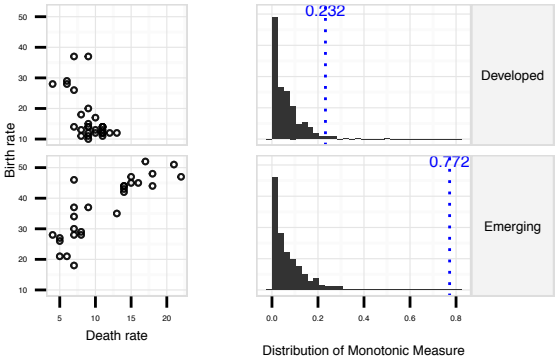
(a) Partitioned by % of population with low status in the area

(b) Partitioned by whether the home is close to the Chas river

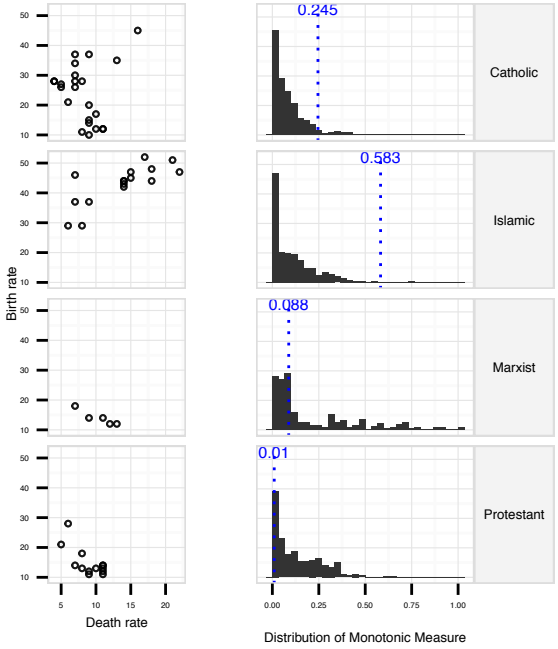
Fig. 4: The choice of scagnostic changes the ranking of the partitioning variables. Above are the highest ranked small multiple displays for the Boston housing dataset on: (a) The Outlying scagnostic which detects shapes with many outlier points. (b) The Clumpy scagnostic which detects shapes with multiple dense regions of points.



(a) Input scatterplot



(b) Partitioned by GDP category



(c) Partitioned by the dominant religion

Fig. 5: Our algorithm picks informative small multiple displays that diverge from the user-selected input plot. (a) User-selected relationship between birth and death rates for countries around the world. (b) The highest ranked small multiple display shows partitions that reveal strong opposite trends that were not seen in the original view. (c) The lowest ranked small multiple display that has more partitions with fewer points that look like random subsets of the input plot.

seen in Figure 4(b) – these component plots do not quite have multiple clumps, but rather, a single clump in the overall shapes. These partitioned views reveal different slices of the original view with visually salient features that align with the scagnostics selected and the analyst's task.

4.2 Informative

Our second criterion is that we want small multiple displays that reveal informative structures when compared to the original view. This criterion is incorporated in our algorithm by the comparison between the true cognostic score and that of the randomized permutations. Randomly permuting the partitioning variable results in partitions that are random subsets of the data in the original plot. Visual patterns in these random subsets are likely to be similar to those in the original plot, and thus, not informative. Thus, a high absolute z-score for the true cognostic value is associated with a small multiple display that shows a pattern *different* from the original plot.

An illustration of this behavior involves the Ourworld dataset of six UN statistics on world countries [45]. We want to determine how to partition the scatterplot showing the relationship between birth rate and death rate seen in Figure 5(a) to isolate the increasing and decreas-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2015.2467323, IEEE Transactions on Visualization and Computer Graphics

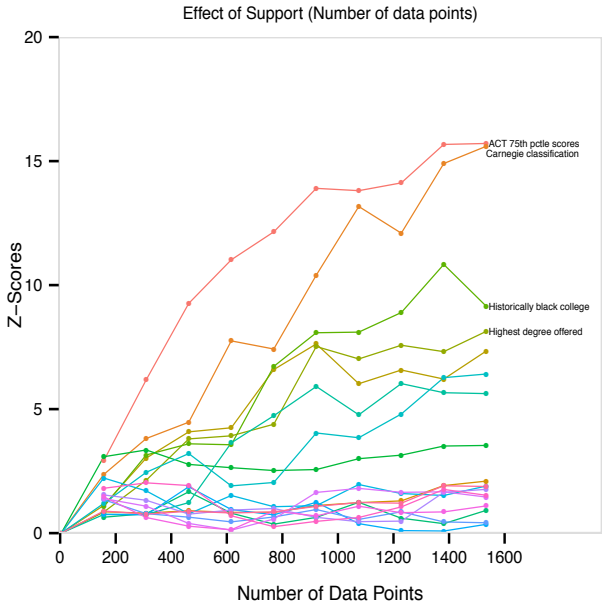


Fig. 6: The effect of support on the combined z-scores ranking of partitioning variables for the data about US universities. As the number of points in the dataset increases, the importance of the variable determined by the z-scores increases too.

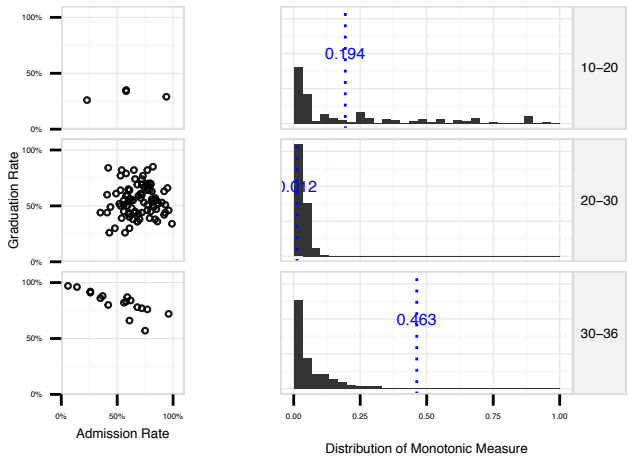
ing trends that seem to be overlaid. So we use the *Monotonic* scagnostic [46] to find informative small multiples. The highest ranked small multiple is partitioned into two GDP categories, "Developed" with 30 points and "Emerging" with 27 points, as seen in Figure 5(b). This partitioning reveals that GDP is a confounding covariate. While the main view shows an overall positive trend between the birth and death rates, the small multiple display shows that in developed regions there is only a strong negative relationship as expected for countries in that category. As shown by the histograms on the right, this informative display arises because the patterns in the small multiple display have true Monotonic scores, seen in blue lines, that are substantially outside the black histogram of scores for random subsets of the original plot. Therefore, the monotonic patterns are significantly different from those seen in the original plot.

Figure 5(c) shows the lowest ranked partitioning variable which categorizes countries by its dominant religion. This variable produces twice as many partitions with less visually salient monotonicity, particularly in those determined by "Protestant" and "Marxist" countries. For those two categories, the blue lines of true Monotonic scores fall within the wide black distributions making the patterns seem more likely due to chance. The other two partitions ("Catholic" and "Islamic" countries) have true Monotonic scores that do not fall far outside the black histogram, making the whole small multiple display not as informative as the top ranked GDP small multiple.

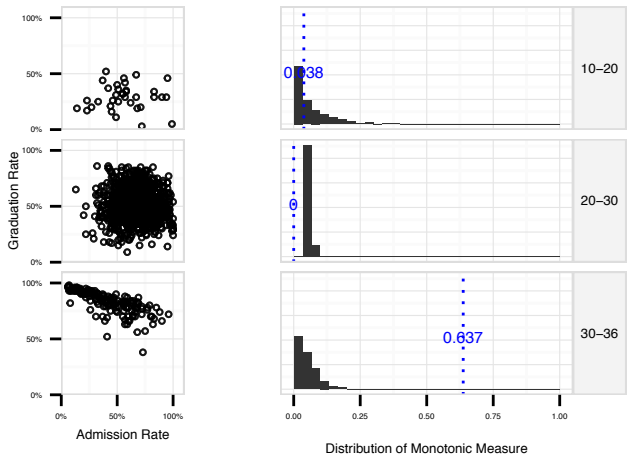
4.3 Support

Our third criterion for good small multiple displays is that they have patterns that are well-supported by the data. This property is incorporated into our approach through the use of the z-score normalization which adjusts for the variability in the simulated null distributions of the randomized cognostic scores. If the data set is small or there are outliers in the data set, this distribution will have high variance, which will downweight the resulting z-scores.

To examine how our algorithm behaves with different amounts of support, we experiment by varying the size of the input data set. Using the US university data set discussed in the introduction and the Monotonic scagnostic, we compute the z-scores for all 19 partitioning variables in the dataset. Figure 6 shows these rankings for different data set sizes, generated by randomly subsetting the full data set. As can



(a) Random 10% of the full dataset partitioned by admit ACT scores.

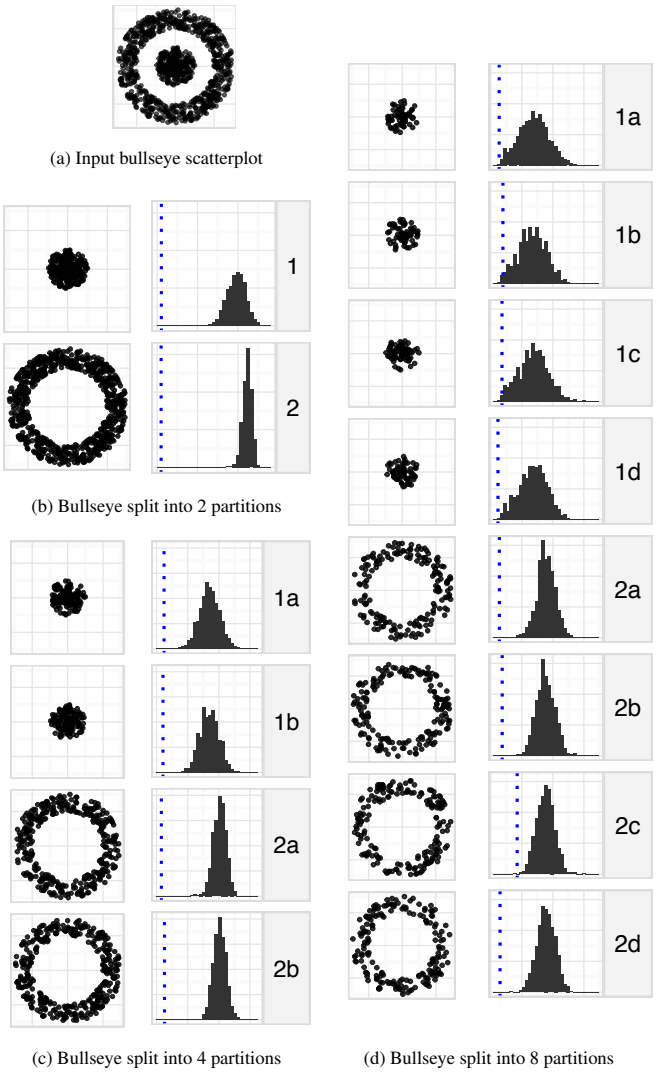


(b) Full dataset partitioned by admit ACT scores.

Fig. 7: The highest ranked small multiple display with random subsets of the full US university dataset show the effect of lower support on the simulated null distributions. (a) A random 10% of the full dataset has a combined z-score of 3.6. The true Monotonic scores fall within the wider null distributions. (b) The full dataset has a combined z-score of 16.4 as we get more confident with more data about the monotonic pattern in the 30 bin.

be seen, for small data sets, the scores are small and there is no clear ranking. The patterns in these small multiple displays are weaker and we correctly require more data to be confident in their ranking. As the data set grows, we become more confident in the rankings, with "ACT 75th Percentile Scores" and "Carnegie Classification" separating from the other variables. We can see the effect of the smaller number of data points in the component plots of the small multiple display produced by the binned ACT scores in Figure 7(a). This shows a 10% random subset (150 points) of the full dataset (1500 points). An analyst should be skeptical about the reliability of any visual pattern in these partitions with small data sizes. The simulated null distributions, seen as the black histograms, have large variance and the true scores, seen as blue overlaid lines, fall within them making it likely these visual patterns look similar to those in random subsets of the input scatterplot. The small multiple on the full dataset, seen in Figure 7(b), reveals a strong monotonic pattern in the partition determined by the 30–36 bin. The analyst should be more confident that this pattern is informative as the true scores, seen as blue lines, fall far outside the black histogram of Monotonic scores from random subsets of the input scatterplot.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2015.2467323, IEEE Transactions on Visualization and Computer Graphics



(a) Input bullseye scatterplot

(b) Bullseye split into 2 partitions

(c) Bullseye split into 4 partitions

(d) Bullseye split into 8 partitions

Fig. 8: Our ranking of small multiple displays of an artificially generated data pattern respects the parsimony criterion. (a) The input bullseye pattern. (b) The best small multiple display determined by the clumpy scagnostic. (c) The second best partitioning variable redundantly halves the two partitions from (b). (d) The lowest ranked small multiple display with eight partitions.

4.4 Parsimonious

Our final criterion is parsimony. This criterion is indirectly included in our approach. High-cardinality variables create a large number of partitions, which are likely to have low support as the observations get distributed among more partitions. Thus, we will tend to reject such partitionings if more parsimonious options are available.

We illustrate this behavior using an artificially generated dataset so we can hold the visual patterns across partitioning variables equal as far as possible. The input visualization is the bullseye pattern shown in Figure 8(a). This artificial data set includes a partitioning variable that cleanly separates the ring from the core as seen in Figure 8(b). It also includes two other partitioning variables that separate the ring and core, but they further split them into two and four random partitions (Figures 8(c) and 8(d)). The separation between the core and ring are equally visible in all three variants, however the first is the most parsimonious—it shows the ring/core separation in the fewest number of component plots. To discover such partitionings of the visual pattern, we could use Wilkinson et al.'s *clumpy* scagnostic that has high values for scatterplots with multiple tight clumps of points. Plots of random samples of the bullseye pattern should produce a distribution of clumpy scores that are high, while the plots that split out the core and the ring will have clumpy scores that are much lower.

In the histograms, we can see that the width of the simulated null distributions increases as the number of partitions increases. Thus, the computed z-score will be highest for the first partitioning variable. The actual z-scores are 19.507, 9.274 and 5.129 for Figures 8(b), 8(c) and 8(d) respectively. Thus, our approach does prefer more parsimonious partitioning variables.

5 Discussion and Future Work

We have presented a method for selecting good partitioning variables for small multiple displays. An advantage of combining cognostics with non-parametric statistical approaches is that they can easily be extended to solve a variety of visual analytic problems. For example, we have described our algorithm in terms of a permutation test, which ignores sampling error in the data set. This is correct in many common analytic scenarios where the data set contains the entire population. If, however, the user wants to account for possible sampling error when scoring small multiple displays, they could instead use bootstrapping [13] to build the simulated null distributions. The structure of the approach would be unchanged.

Another natural extension of our method would be to handle continuous variables. Our method works in a straightforward manner on discrete partitioning variables. For continuous variables, discrete partitions can be created through disjoint binning techniques [15, 32], or, overlapping bins (shingles) [4]. In either case, our approach can be extended to handle binning by first permuting the continuous variable and then applying the binning algorithm to partition the data. We used the equal count binning algorithm for non-overlapping shingles in the example described in Section 3. Future work could investigate the use of this process to find interesting bins for continuous variables given a particular partitioning variable. The parameters to the binning algorithm could be varied while the partitioning variable was held constant. This would allow us to pick out the setting of bins that maximize the cognostic for that partitioning variable.

While we frame our algorithm in terms of scoring single variables, it is trivial to combine two discrete variables into a new discrete variable by crossing or nesting the levels of each variables [45, 36]. Doing so would allow our algorithm to consider combinations of variables. Another common use case is creating small multiples by drilling down into aggregated data. A variation of our approach could be used to detect if potentially interesting visual information would be revealed by a change in level of detail. Visualization tools could use this to recommend a drill down or roll up dimension. We could also extend our approach to consider sequences of partitionings. This could be used to develop a decision tree based exploratory data analysis interaction mechanism guided by our algorithm. At each decision level, we could apply our algorithm to select a partitioning variable given a single view of the data at that level. This would produce a small multiple display where each component plot could be further partitioned to reveal interesting structure. Considering the tree structure, each choice of a partitioning variable would be conditional on the other previously used variables, as in model selection methods.

One weakness with our approach is that we do not correct for possible correlation between the patterns in the input visualization and partitioning variables. As a result, we may redundantly choose a small multiple display that shows a pattern that was already clearly visible in the original plot. While exposing highly correlated variables can be useful, it is likely not what the user wants in an effective small multiple display. Statistical methods for variable selection, such as ridge or lasso regression, can downweight highly correlated variables. Our approach would be improved by incorporating similar behavior.

Our use of Chebyshev's inequality produces a very conservative bound on the likelihood of a cognostic. Better results might be achieved if more information is available about the underlying distribution of the cognostic. For example, Wilkinson and Wills have suggested that the distributions of their graph-theoretic scagnostics are well-modeled by a beta distribution [47]. Fitting a beta distribution would capture the skew and truncation visible in some of our empirical

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2015.2467323, IEEE Transactions on Visualization and Computer Graphics

cognostic distributions. We also suggest using the maximum absolute z-score across all component plots to score the overall small multiple display. This allows us to pick up single partitions with strong patterns. However, it may discount small multiples with weaker patterns in many or all the component plots. Averaging the z-scores across component plots might help address this, but may miss strong individual plots. Our choice of the maximum has worked well in practice, but more exploration is needed.

Both cognostics and non-parametric methods are computationally demanding. In our approach we compute the scagnostics for each partition of each variable and then for the randomly permuted partitions for each variable. For a moderate sized dataset with thousands of rows, our R implementation takes about ten seconds on average to evaluate each partitioning variable. More work on computationally efficient cognostics is needed. Also, in our work with Scagnostics, we have found that they sometimes miss very obvious visual patterns. More work is needed to develop cognostics that are robust to properties such as sample size, the amount of noise in the data set, and the location and scale of the axes.

6 Conclusion

Small multiple displays are a powerful mechanism to analyze a visual relationship conditioned on other variables. Multidimensional datasets offer a challenge due to the combinatorics in the choice of partitioning variables. In this paper, made a first step in addressing this problem by describing a method for automatically ranking the small multiple displays created by the partitioning variables in a data set. Our use of a randomized permutation test allows our method to detect and discount non-informative or spurious patterns in small multiple displays. We also described a set of goodness criteria for small multiple displays that favors fewer partitions, visually rich patterns that are well-supported by data observations and are different from the patterns seen in the unpartitioned view of the same data.

The basis of our approach—the combination of cognostics and non-parametric tests—is very general and, as we have outlined, there is much more work to be done exploring this area. We focused on scatterplots, as the primary data view, and scagnostics, as measures of visual patterns, to illustrate our method of evaluating small multiple displays. But, our method can incorporate a wide range of quality measures allowing it to be used on different visualization type and to address different analytic goals. We believe that the development of new cognostics and the application of them to visual analytics using non-parameteric approaches will provide analysts with a new generation of tools that will help them explore their data faster and more accurately.

Acknowledgments

Acknowledgements blinded for review.

References

[1] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor. Improving the visual analysis of high-dimensional datasets using quality measures. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 19–26. IEEE, 2010.

[2] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Information Visualization, 1998. Proceedings. IEEE Symposium on*, pages 52–60. IEEE, 1998.

[3] R. A. Becker and W. S. Cleveland. *S-PLUS Trellis Graphics User's Manual*. Murray Hill: Bell Labs, 1996.

[4] R. A. Becker, W. S. Cleveland, and M.-J. Shyu. The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*, 5(2):123–155, 1996.

[5] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983.

[6] E. Bertini and G. Santucci. Give chance a chance: modeling density to enhance scatter plot quality through random data sampling. *Information Visualization*, 5(2):95–110, 2006.

[7] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: an overview and systematization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2203–2212, 2011.

[8] N. Boukhelifa, W. Cancino, A. Bezerianos, and E. Lutton. Evolutionary visual exploration: Evaluation with expert users. *Computer Graphics Forum*, 32(3pt1):31–40, 2013.

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

[10] Q. Cui, M. O. Ward, E. A. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):709–716, 2006.

[11] T. N. Dang and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *Pacific Visualization Symposium (PacificVis), 2014 IEEE*, pages 73–80, March 2014.

[12] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1017–1026, 2010.

[13] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC Press, 1994.

[14] M. Forina, C. Armanino, S. Lanteri, and E. Tiscornia. Classification of olive oils from their fatty acid composition. In M. Martens and H. J. Russwurm, editors, *Food Research and Data Analysis*. Applied Science Publishers, London, 1983.

[15] D. Freedman and P. Diaconis. On the histogram as a density estimator: L2 theory. *Probability Theory and Related Fields*, 57(4):453–476, Dec. 1981.

[16] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Series in Statistics, 2000.

[17] R. Hafen, L. Gosink, J. McDermott, K. Rodland, K.-V. Dam, and W. Cleveland. Trelliscope: A system for detailed visualization in the deep analysis of large complex data. In *Large-Scale Data Analysis and Visualization (LDAV), 2013 IEEE Symposium on*, pages 105–112, 2013.

[18] D. Harrison and D. Rubinfeld. Hedonic prices and the demand for clean air. *Journal of Environ. Economics & Management*, 5:81–102, 1978.

[19] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):993–1000, 2009.

[20] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 547–554. ACM, 2012.

[21] A. M. Kelley. *Papers and correspondence of William Stanley Jevons. v.1*. Macmillan for the Royal Economic Society, 1973.

[22] A. MacEachren, D. Xiping, F. Hardisty, D. Guo, and G. Lengerich. Exploring high-d spaces with multiform matrices and small multiples. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 31–38. IEEE, 2003.

[23] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, 1986.

[24] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007.

[25] M. Majumder, H. Hofmann, and D. Cook. Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, 108(503):942–956, 2013.

[26] J. Marks, B. Andalman, P. A. Beardsley, W. Freeman, S. Gibson, J. Hodgins, T. Kang, B. Mirtich, H. Pfister, W. Ruml, et al. Design galleries: A general approach to setting parameters for computer graphics and animation. In *Proceedings of the 24th annual conference on Computer Graphics and Interactive Techniques*, pages 389–400. ACM Press/Addison-Wesley Publishing Co., 1997.

[27] R. Menjoge. *New Procedures for Visualizing Data and Diagnosing Regression Models*. PhD thesis, MIT, 2010.

[28] National Center for Education Statistics. Integrated Postsecondary Education Data System (IPEDS). `https://nces.ed.gov/ipeds/`, 2013-14. Dataset retrieved from `http://public.tableau.com/s/resources?qt-overview_resources=1`.

[29] H. Piringer, W. Berger, and H. Hauser. Quantifying and comparing features in high-dimensional datasets. In *Information Visualisation, 2008. IV '08. 12th International Conference*, pages 240–245, July 2008.

[30] M. Schäfer, L. Zhang, T. Schreck, A. Tatu, J. A. Lee, M. Verleysen, and D. A. Keim. Improving projection-based data analysis by feature space transformations. In *IS&T/SPIE Electronic Imaging*, pages 86540H–

86540H. International Society for Optics and Photonics, 2013.

[31] J. Schneidewind, M. Sips, and D. A. Keim. Pixnostics: Towards measuring the value of visualization. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pages 199–206. IEEE, 2006.

[32] D. W. Scott. Sturges' rule. *WIREs Computational Statistics*, 1:303–306, 2009.

[33] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Comp. Graph. Forum*, 31(3pt4):1335–1344, June 2012.

[34] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, July 2005.

[35] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.

[36] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Trans. Vis. Comput. Graph.*, 8(1):52–65, 2002.

[37] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 59–66. IEEE, 2009.

[38] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986.

[39] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[40] J. W. Tukey and P. A. Tukey. Some graphics for studying four-dimensional data. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, 1982.

[41] J. W. Tukey and P. A. Tukey. Computer graphics and explaoratory data analysis: An introduction. In *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics*, 1985.

[42] S. van den Elzen and J. J. van Wijk. Small multiples, large singles: A new approach for visual data exploration. *Computer Graphics Forum*, 32(3pt2):191–200, 2013.

[43] H. Wickham. ggplot: An implementation of the grammar of graphics. *R package version 0.4. 0*, 2006.

[44] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. *IEEE Trans. Vis. Comput. Graph.*, pages 973–979, 2010.

[45] L. Wilkinson. *The Grammar of Graphics*. Springer-Verlag New York, Inc., 2005.

[46] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 157–164, Oct 2005.

[47] L. Wilkinson and G. Wills. Scagnostics distributions. *Journal of Computational and Graphical Statistics*, 17(2):473–491, 2008.

[48] G. Wills and L. Wilkinson. AutoVis: Automatic visualization. *Information Visualization*, 9(1):47–69, 2010.

[49] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the Symposium on Data Visualisation 2003*, pages 19–28, 2003.