Automatic Selection of Partitioning Variables for Small Multiple Displays

Anushka Anand, Justin Talbot

Presented by Yujie Yang, CPSC 547 Information Visualization

Agenda

- Introduction
- Goodness-of-Split Criteria
- Algorithm
- Validation
- Conclusion
- Comments

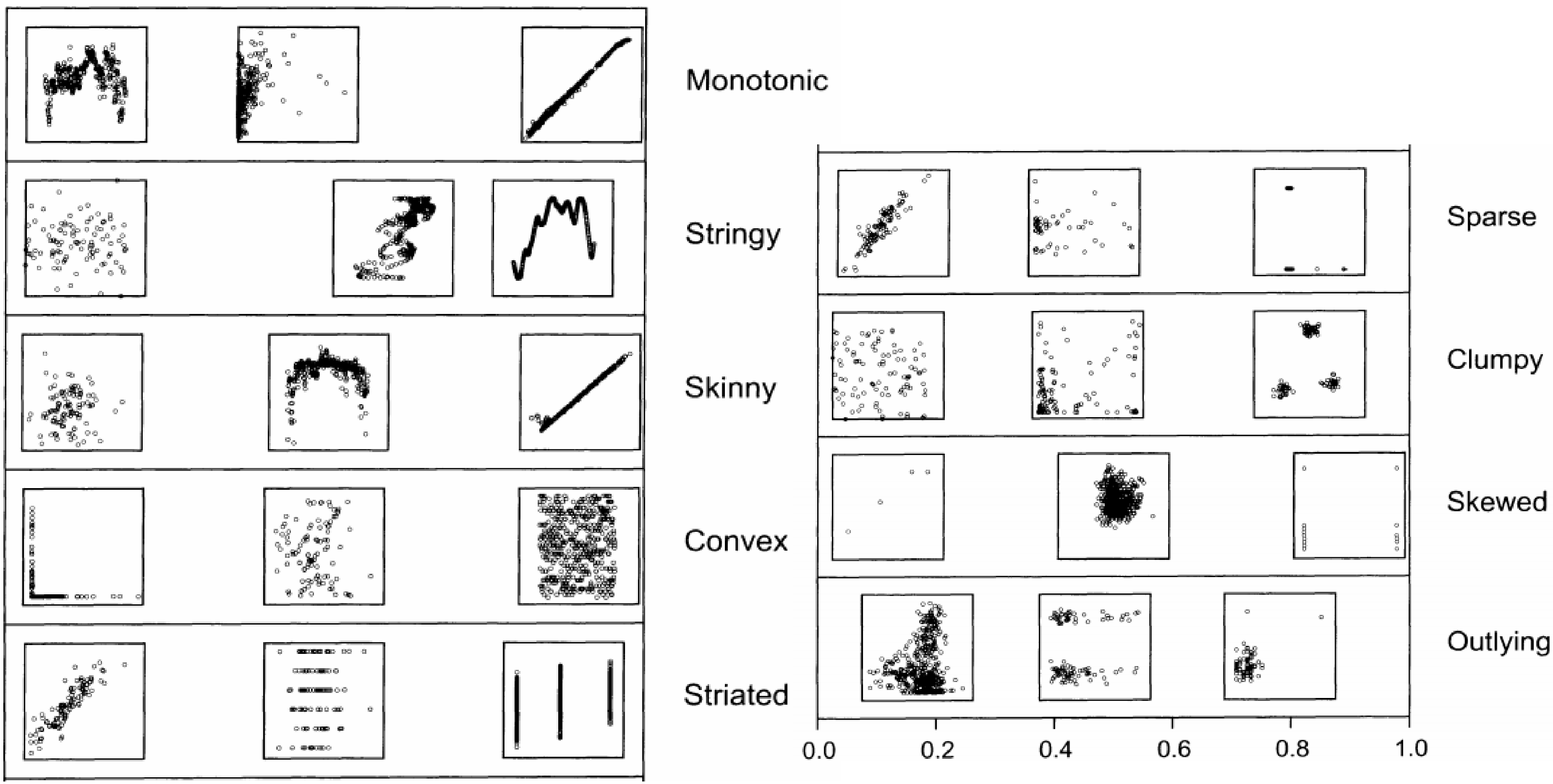
Introduction

- Authors – from Tableau Research
  - Anushka Anand
  - Justin Talbot
- IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS(TVCG)
- January 2016

Introduction

- What: multidimensional data sets
- Why: For small multiples, automatically select the partitioning variables?
- How?
- Cognostics
  - Firstly introduced by John and Paul Tukey
  - Wilkinson extended original idea
  - "Judge the relative interest of different displays"
  - Scagnostics – scatterplot diagnostics

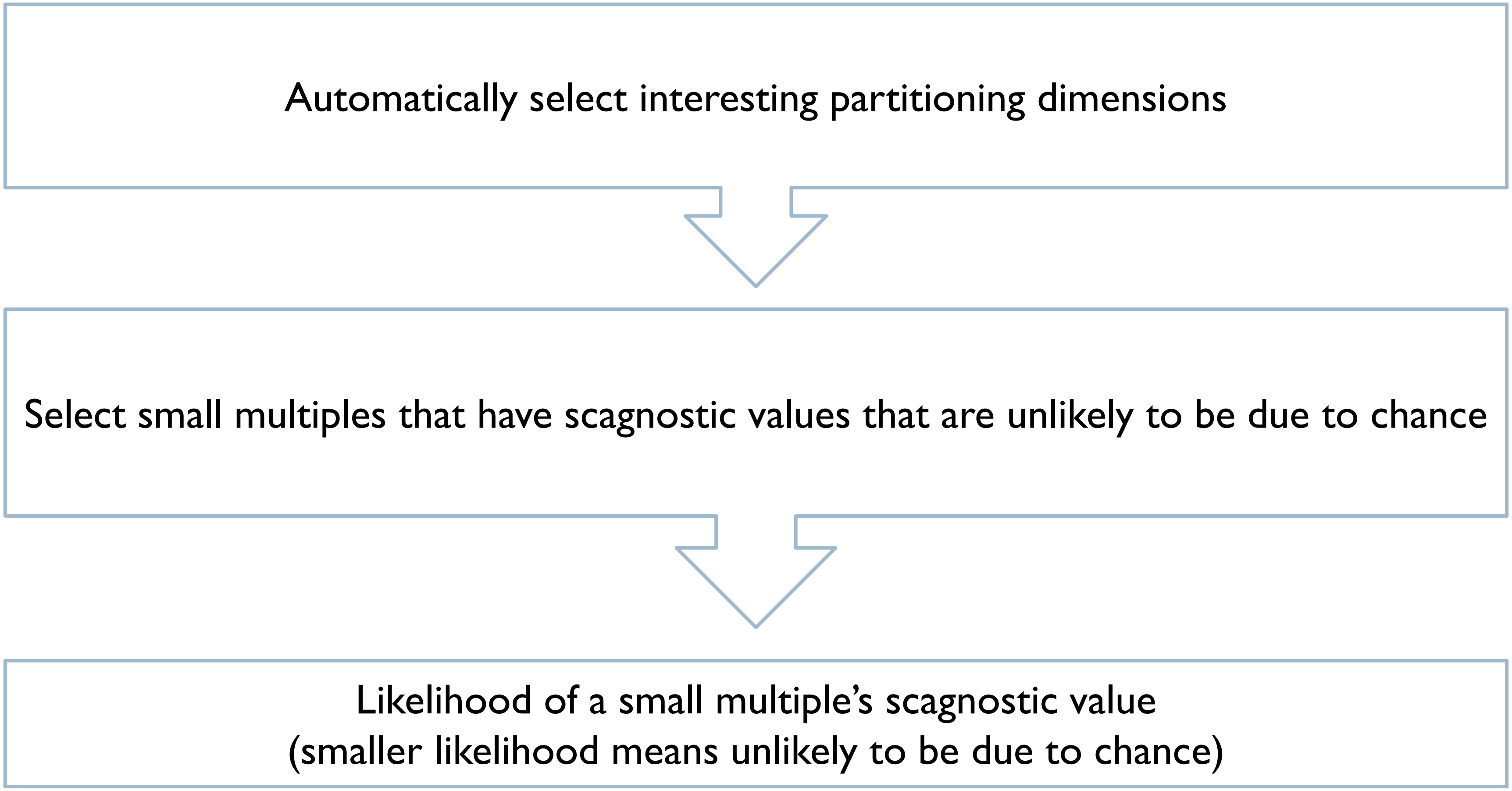
Introduction - Scagnostics



Goodness-of-Split Criteria

- Visually rich
  - Convey rich visual patterns
- Informative
  - More informative than the input
- Well-supported
  - Convey robust and reliable patterns
- Parsimonious
  - All things being equal, then fewer partitions

Algorithm

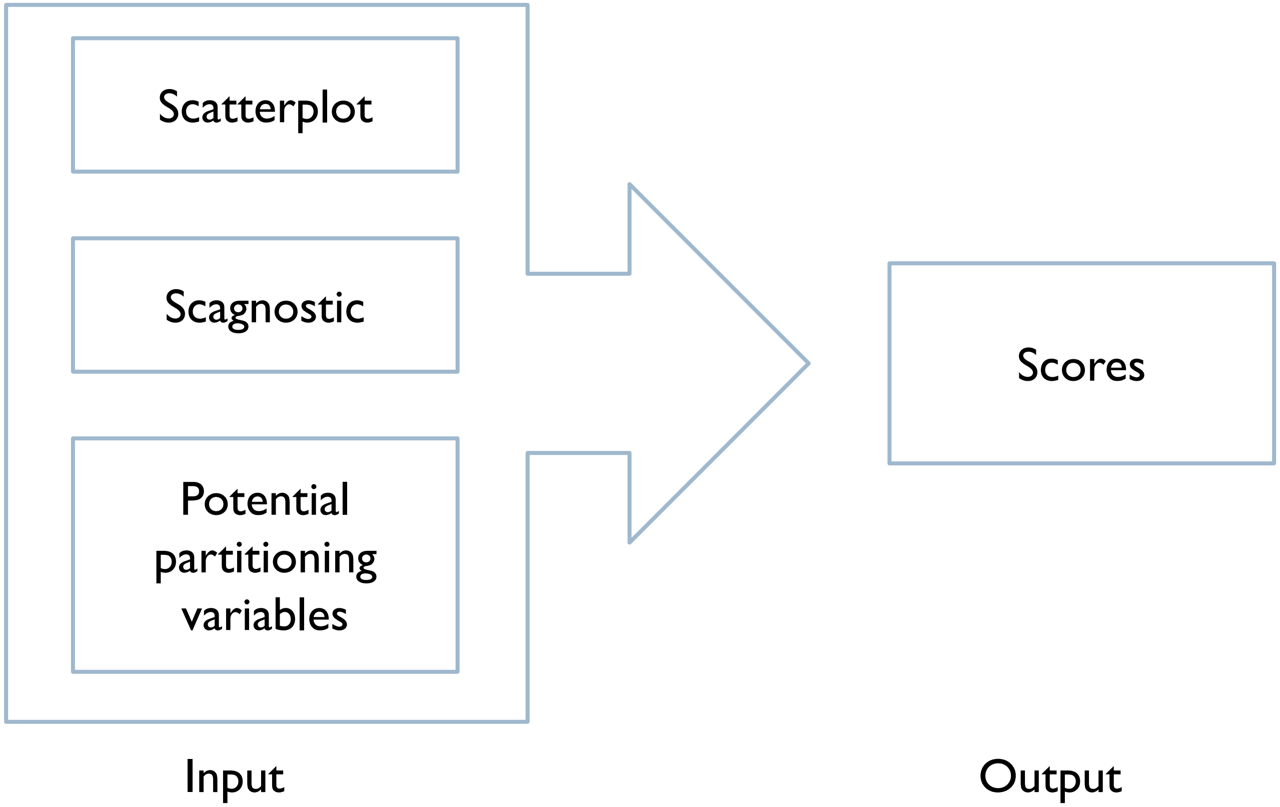


Automatically select interesting partitioning dimensions

Select small multiples that have scagnostic values that are unlikely to be due to chance

Likelihood of a small multiple’s scagnostic value
(smaller likelihood means unlikely to be due to chance)

Algorithm



Scatterplot

Scagnostic

Potential partitioning variables

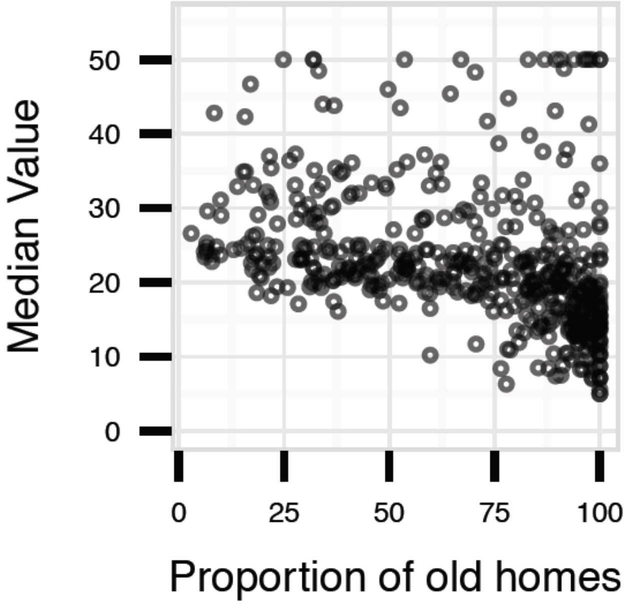
Scores

Input

Output

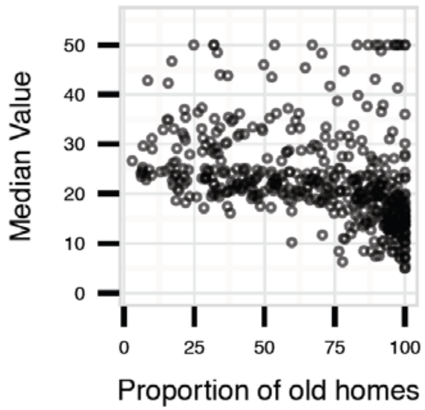
Algorithm

- Input:
  - Scatterplot
  - Scagnostic: skewed
  - Partitioning Variable: distance to employment center

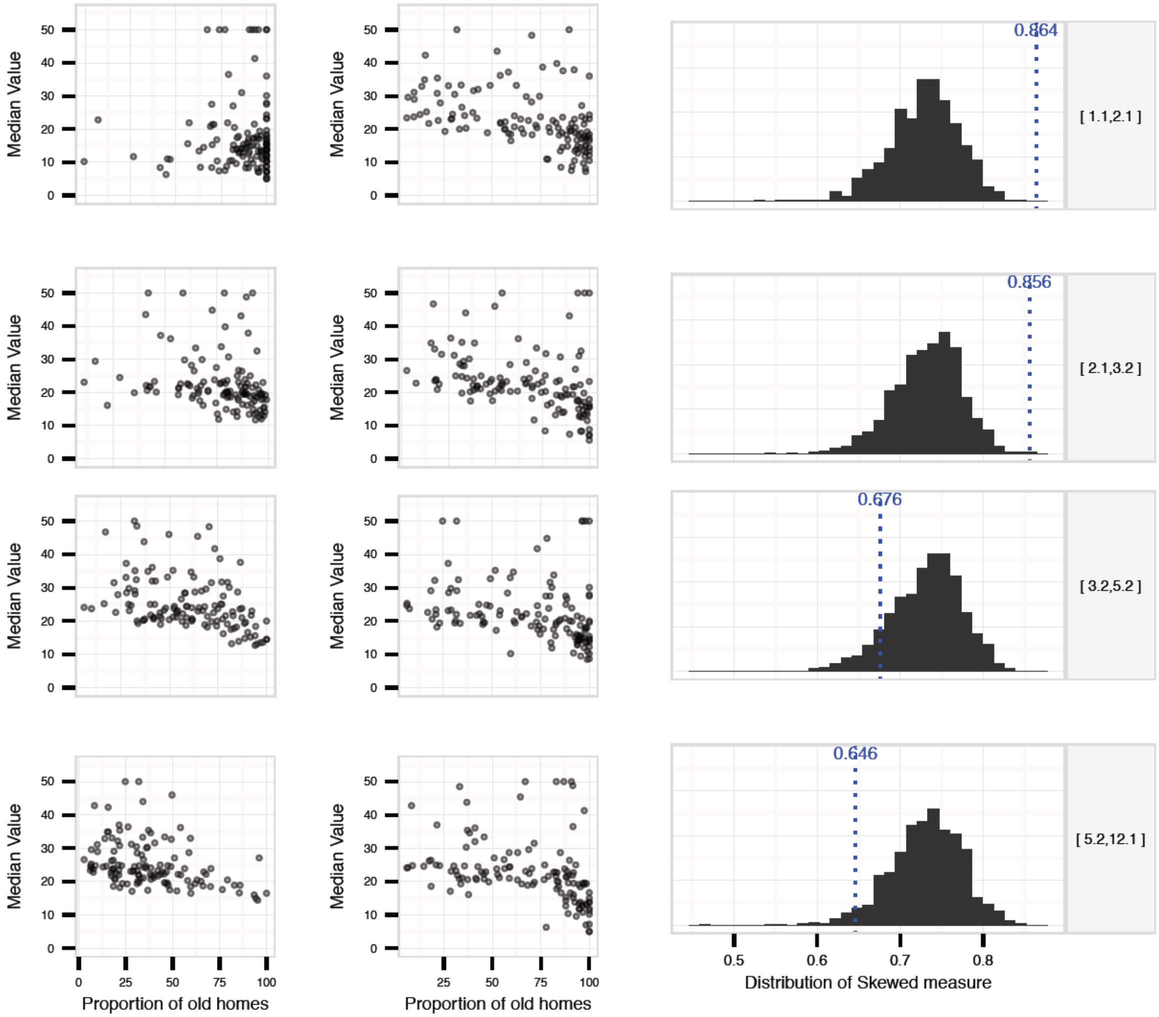


Data:
X: proportion of old houses built before 1940 for census tracts in Boston
Y: median value of owner-occupied houses

Algorithm



(a)



(b) (c) (d)

(a)Input scatterplot
(b)Partitioned by distance
(c)Partitioned by random permutation
(d)Distribution of Skewed value

Algorithm

- Permutation test
- Chebyshev's inequality:

$$\Pr\left(\left|\frac{(X-\mu)}{\sigma}\right| \geq k\right) \leq \frac{1}{k^2}.$$

- z-score:

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

- Output:

$$z = \max_i |z_i|$$

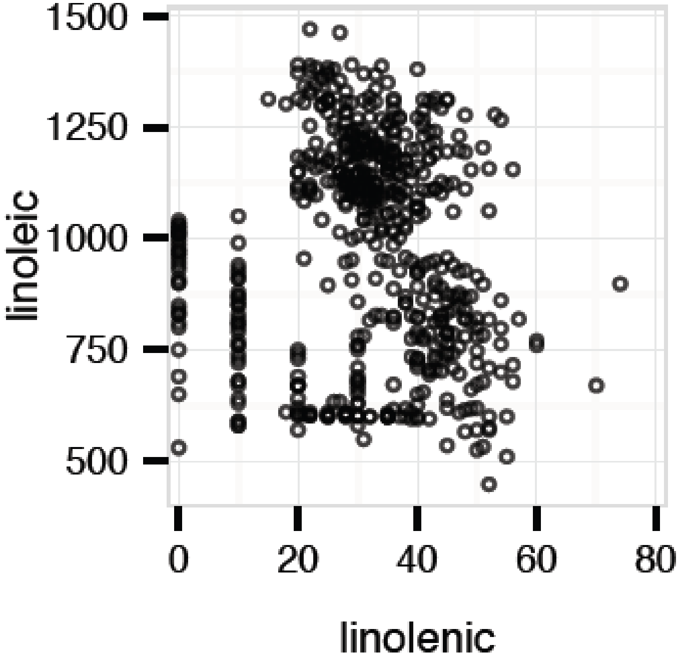
Where *Xi* is the true scagnostic value of the *i*-th partition and *μi* and *σi* are the mean and standard deviation of the scagnostic measures over the repeated random permutations of the *i*-th partition.

Algorithm

Algorithm	Automatic Selection of partitioning variables
What: Data	multidimensional data sets; scatterplot
Why: Task	Automatically select variables to divide scatterplot into small multiples
How: Facet	Small multiples
How: Input	Scatterplot; scagnostic; partitioning variables
How: Output	Max of z-scores
Scale	Items: thousands; dimensions: dozens

Validation - Visually rich

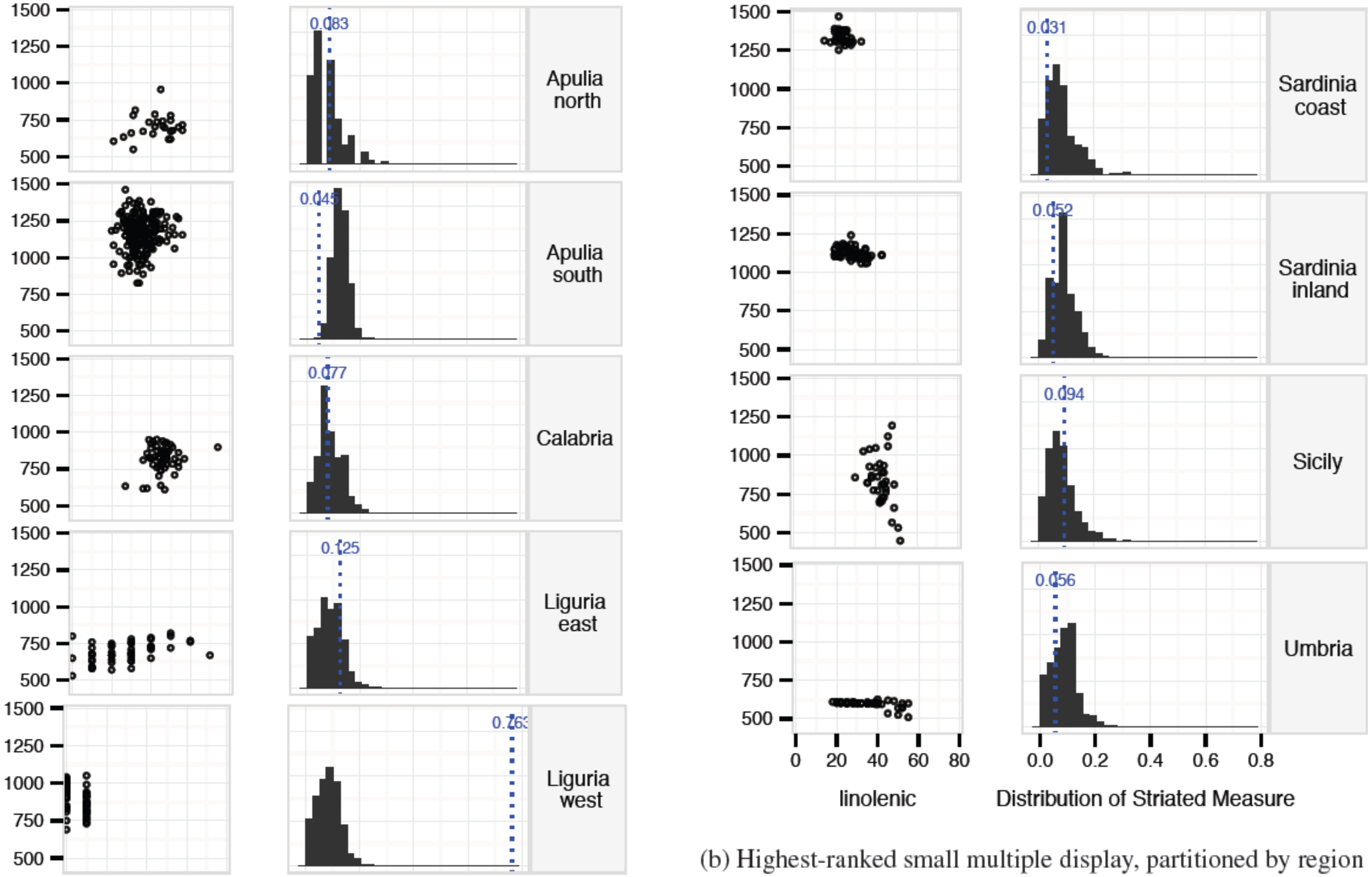
- Visually striking clumps and striation patterns



(a) Input scatterplot

Data:
X: linolenic measurement in olive oil specimens in Italy
Y: linoleic measurement in olive oil specimens in Italy

Validation - Visually rich

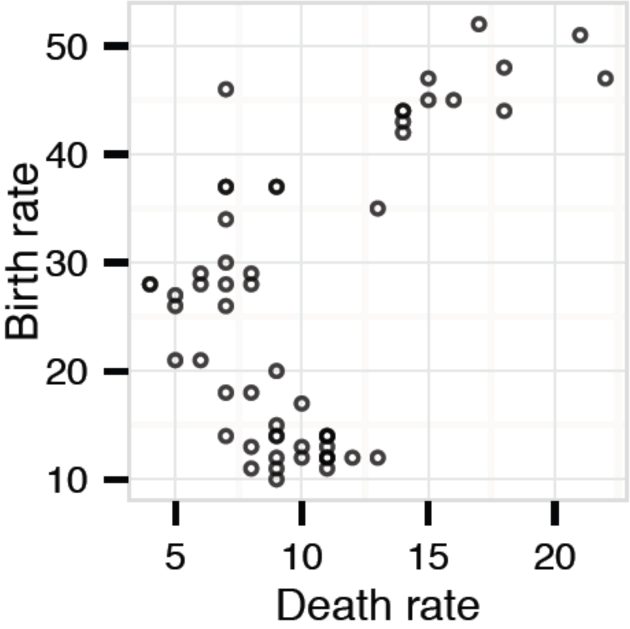


(b) Highest-ranked small multiple display, partitioned by region

- Scagnostic: striated
- Partitioning Variable: region

Validation - Informative

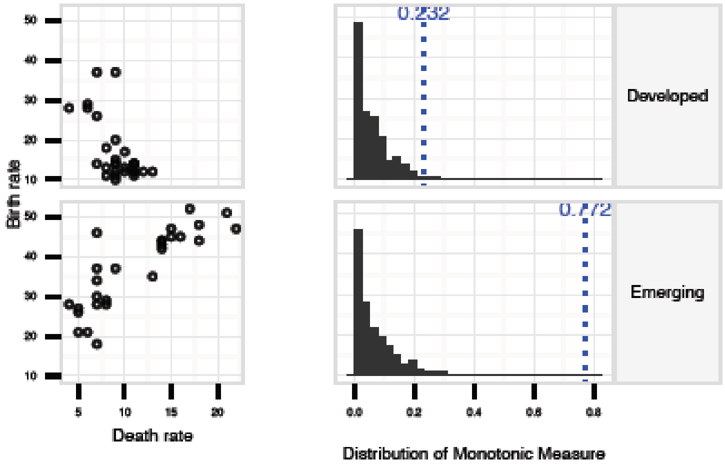
- Increasing and decreasing trends seem to be overlaid



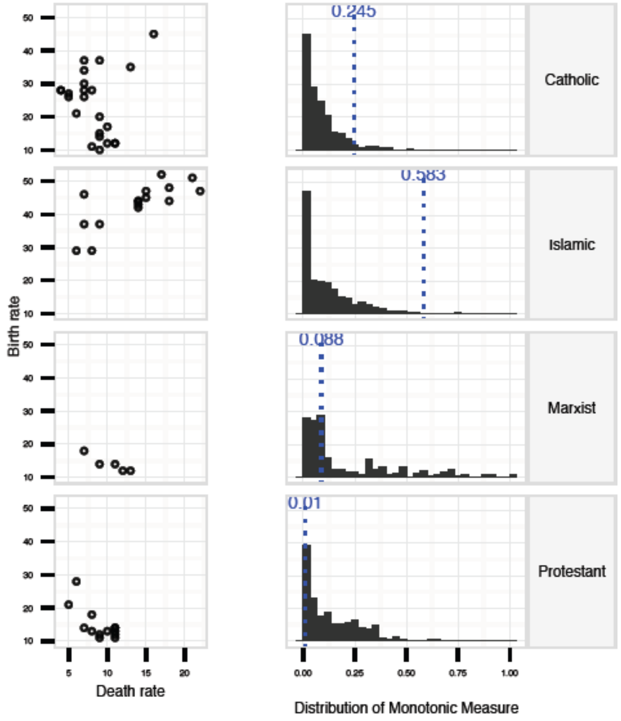
(a) Input scatterplot

Data:
X: death rate of world countries
Y: birth rate of world countries

Validation - Informative



(b) Partitioned by GDP category

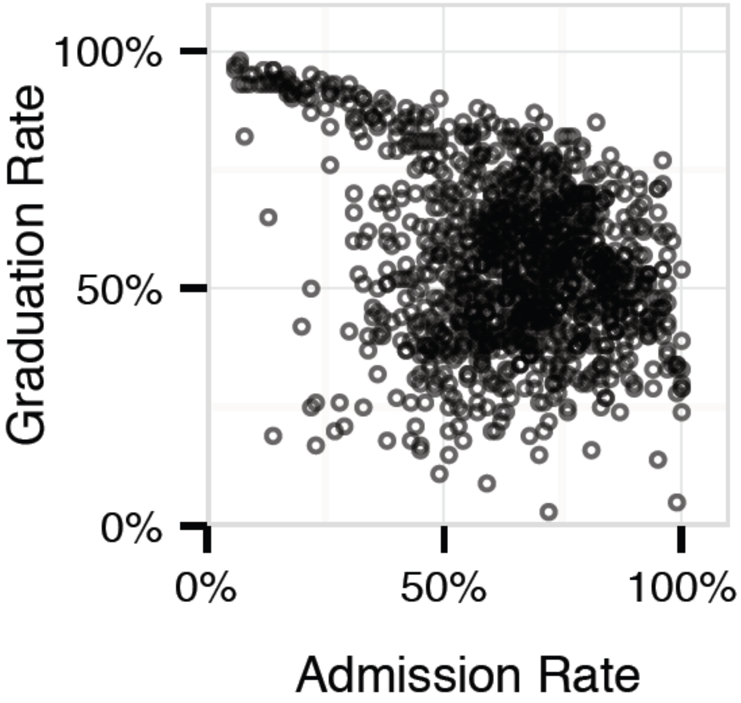


(c) Partitioned by the dominant religion

- Best case
  - Scagnostic: monotonic
  - Partitioning Variable: GDP category
- Worst case
  - Scagnostic: monotonic
  - Partitioning Variable: dominant religion

Validation – Well-supported

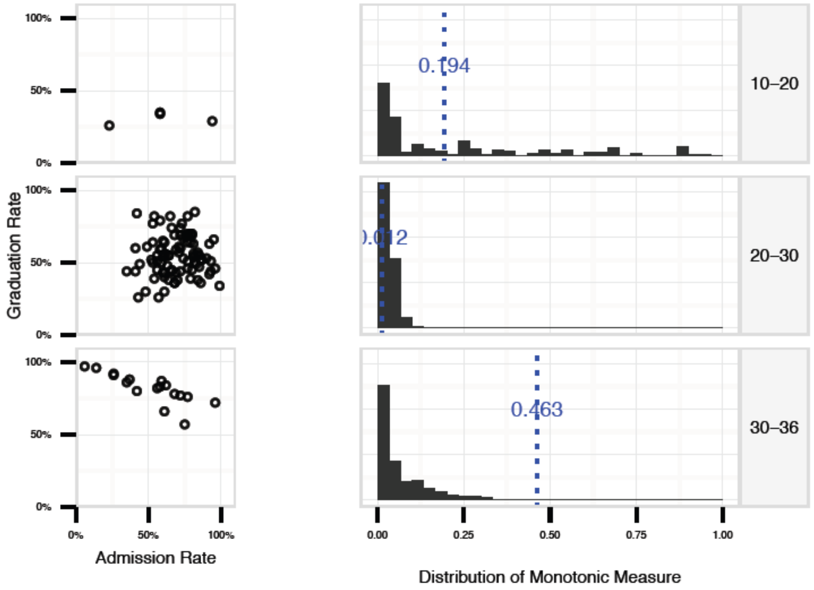
- Run the algorithm for different size of the input data



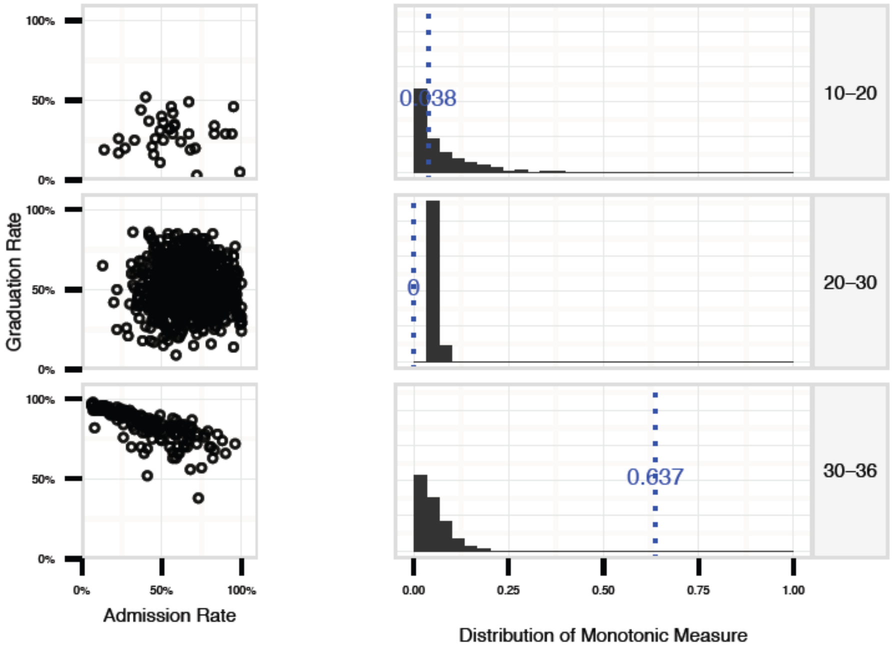
(a) Input scatterplot

Data:
X: admission rate at US universities
Y: graduation rate at US universities

Validation – Well-supported



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



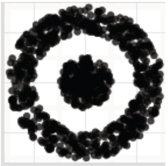
(b) Full dataset partitioned by admit ACT scores.

- Random 10% of full dataset
- Scagnostic: monotonic
- Partitioning variable: admit ACT scores
- Z-score: 3.6

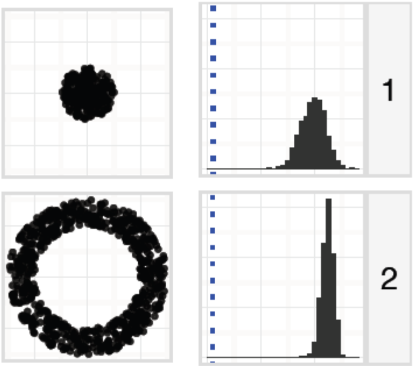
- Full dataset
- Scagnostic: monotonic
- Partitioning variable: admit ACT scores
- Z-score: 16.4

Validation - Parsimonious

- Artificially generated dataset
- Scagnostic: clumpy

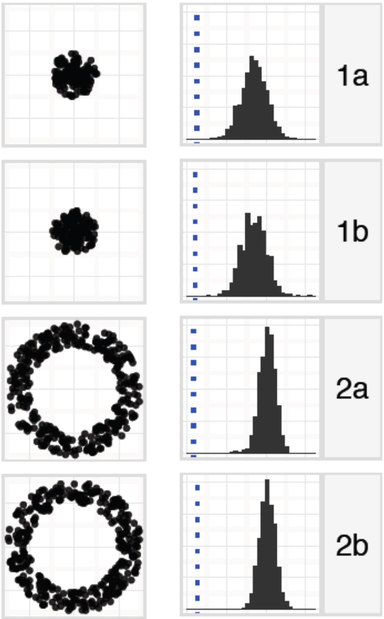


(a) Input bullseye scatterplot



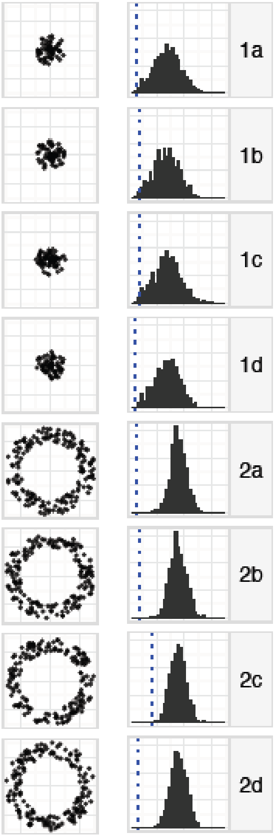
(b) Bullseye split into 2 partitions

Best case



(c) Bullseye split into 4 partitions

Second best case



(d) Bullseye split into 8 partitions

Worst case

Conclusion

- Described a set of goodness criteria for evaluating small multiples
- Proposed a method for automatically ranking the small multiple displays created by the partitioning variables in a data set
- Demonstrated the method meets the criteria

- Future:
  - Scatterplot -> different visualization type
  - Scagnostics -> wide range of quality measures
  - Evaluating small multiple -> different analytic goals

Comments

- As mentioned in their discussion:
  - Lack of examples about different visualization types or analytic goals
  - Not deal with correlation between input and partitioning variables
  - Max of z-scores VS average of z-scores
- More critiques:
  - *Their* method meets *their* criteria?
  - Use the idea of permutation test, but lack of exact likelihood (or p-value) of the cognostic score in the examples
  - Weak proof of the support to the criterias

Thank you!

Reference

[1] Anand A, Talbot J. Automatic Selection of Partitioning Variables for Small Multiple Displays[J]. 2016.

[2] Friedman J H, Stuetzle W. John W. Tukey's work on interactive graphics[J]. Annals of Statistics, 2002: 1629-1639.

[3] Wilkinson L, Anand A, Grossman R L. Graph-Theoretic Scagnostics[C]//INFOVIS. 2005, 5: 21.

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