Geovisualization

Luc Anselin

THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

http://spatial.uchicago.edu

from EDA to ESDA

from mapping to geovisualization

mapping basics

multivariate EDA primer

From EDA to ESDA

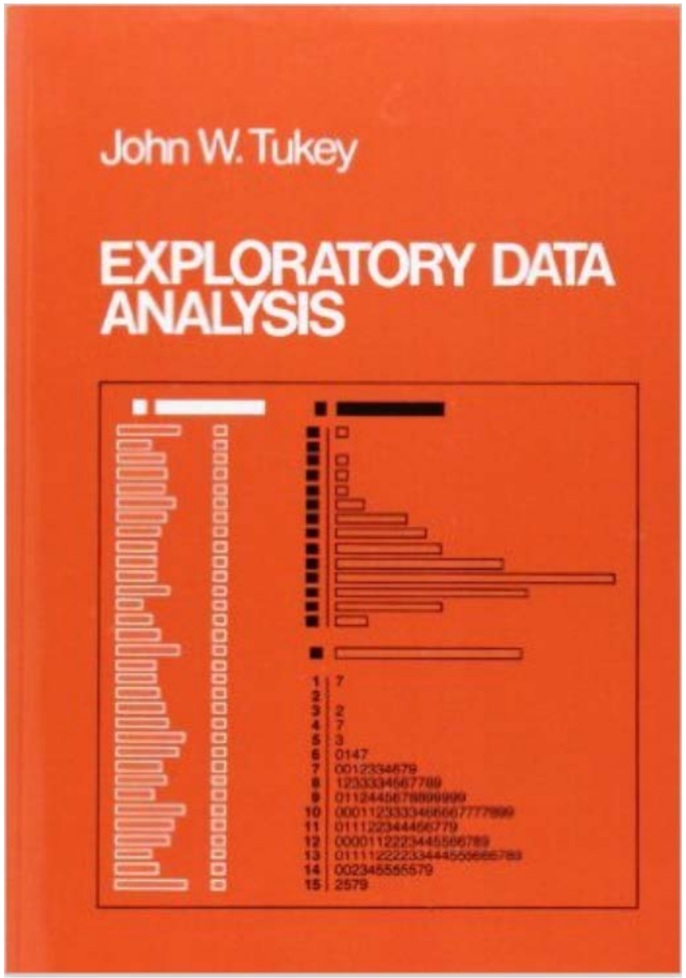
- Exploratory Data Analysis (EDA)

    reaction to modeling without looking at the data

    classic EDA book, Tukey (1977)

    Good (1983), Philosophy of Science

        “discover potentially explicable patterns”



THE PHILOSOPHY OF EXPLORATORY DATA ANALYSIS*

I. J. GOOD†

Statistics Department
Virginia Polytechnic Institute and State University

This paper attempts to define Exploratory Data Analysis (EDA) more precisely than usual, and to produce the beginnings of a philosophy of this topical and somewhat novel branch of statistics.

A *data set* is, roughly speaking, a collection of k-tuples for some k. In both descriptive statistics and in EDA, these k-tuples, or functions of them, are represented in a manner matched to human and computer abilities with a view to finding patterns that are not "kinkera". A *kinkus* is a pattern that has a negligible probability of being even partly potentially explicable. A potentially explicable pattern is one for which there probably exists a hypothesis of adequate "explicativity", which is another technical probabilistic concept. A pattern can be judged to be probably potentially explicable even if we cannot find an explanation. The theory of probability understood here is one of partially ordered (interval-valued), subjective (personal) probabilities. Among other topics relevant to a philosophy of EDA are the "reduction" of data; Francis Bacon's philosophy of science; the automatic formulation of hypotheses; successive deepening of hypotheses; neurophysiology; and rationality of type II.

1. Introduction. Both data analysis (EDA) and confirmatory data analysis (CDA) have existed, under any reasonable definition, for more than a century, but in recent years the distinction between them has been recognized much more consciously by statisticians, partly because of the influence of Tukey (1977).

EDA is concerned with observational data more than with data obtained by means of a formal design of experiments. When data are obtained informally, we are not surprised if the methods for handling them are also often informal, and perhaps EDA is more an art, or even a bag of tricks, than a science. If this is so, it might be difficult or impossible to find a reasonably comprehensive philosophy of EDA. As Cochran (1972) says, in his article on observational studies, "we can claim only to be groping toward the truth".

EDA is an extension of descriptive and graphical statistics so it seems pertinent to quote David Cox (1978, p.5) also. He says "There is a major need for a theory of graphical methods", and goes on to say "Of course, theory is not to be taken as meaning mathematical theory!" Leamer (1978)

*Received October 1982; revised January 1983.
†I am grateful to John W. Pratt for some useful criticisms. This work was supported in part by N.I.H. Grant R01-GM18770.

Philosophy of Science, 50 (1983) pp. 283–295.

- Data Visualization

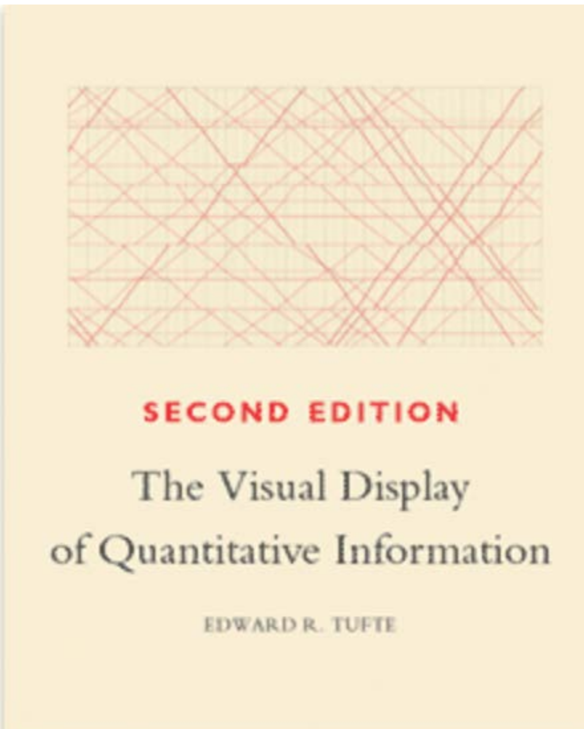
    concept of a "view" (e.g., Buja et al 1996)

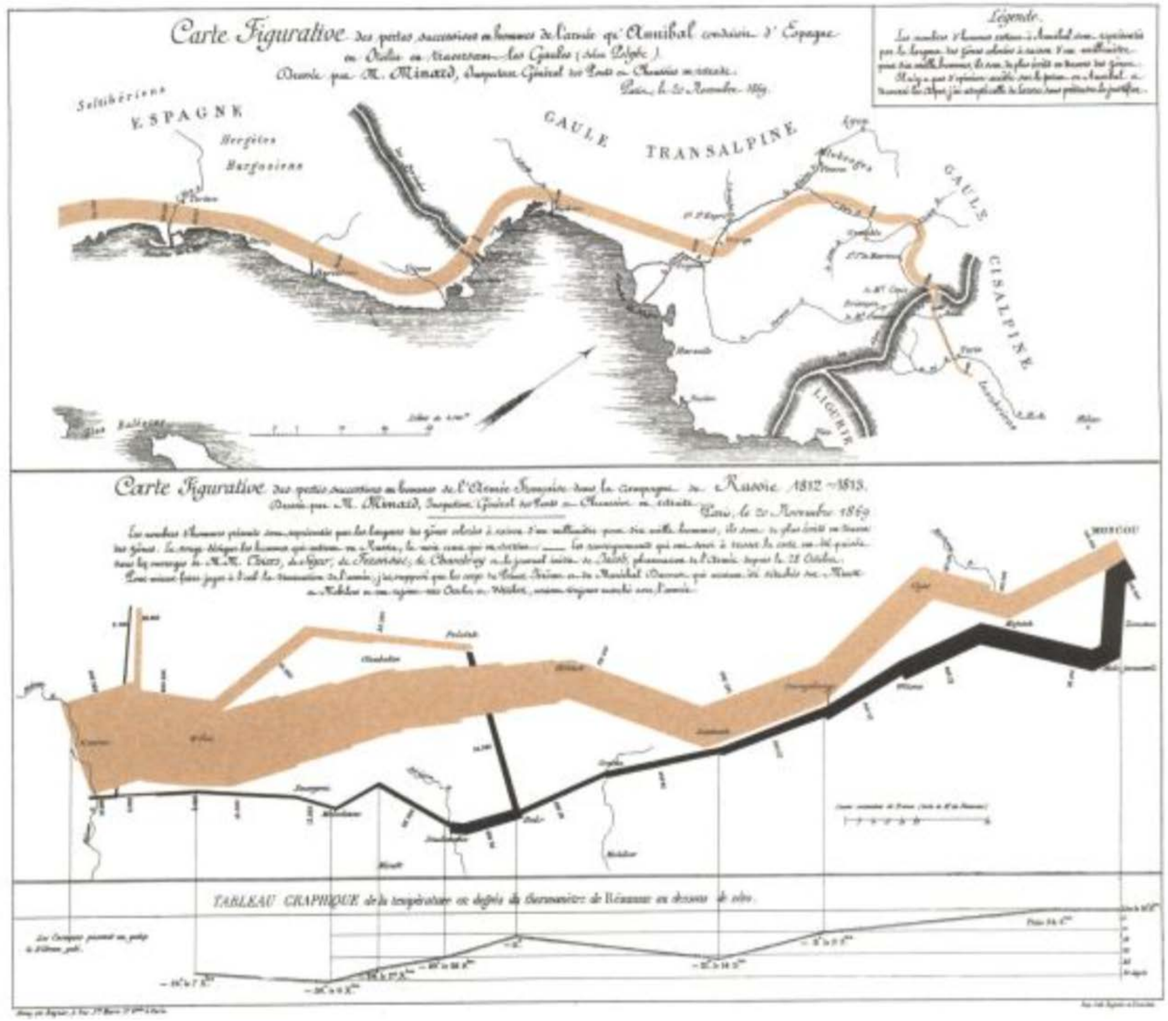
        a graphical representation and summary of the data

        many different views

        chart, table, graph, map

- Visual Explanations
    - Tufte (1997) and later
    - reasoning about evidence and design of graphics
        - multivariate nature of analytic problems
        - document sources (metadata)
        - quantify and show cause and effect
        - evaluate alternative explanations


SECOND EDITION
The Visual Display
of Quantitative Information
EDWARD R. TUFTE


Carte Figurative
ESPAGNE
GAULE
TRANSALPINE
CISALPINE
Carte Figurative
MOSCOU
TABLEAU GRAPHIQUE

- Visual Analytics

Thomas et al (2005)

the science of analytical reasoning facilitated by interactive visual interfaces

“detect the expected and discover the unexpected”

Introduction

Foundations and Frontiers in Visual Analytics

Joe Kielman[a]
Jim Thomas[b,*] and
Richard May[b]

[a]US Department of Homeland Security, Science and Technology Directorate, Washington, DC, USA.
[b]Pacific Northwest National Laboratory, PO Box 999, K7-28, Richland, WA 99352, USA.

*Corresponding author.
E-mail: jim.thomas@pnl.gov

Information Visualization (2009) **8,** 239–246. doi:10.1057/ivs.2009.25

Introduction

This introduction and the future vision section for this special issue of *Information Visualization* hopes to set the stage for an emerging worldwide effort to advance the state of the science of visual analytics. We present some of the driving needs followed by some principles and methods for advancing this science through partnerships among national laboratories, academia, industry and the international science community. Also presented is a selection of the many successes the science, engineering and industrial communities have had in taking core scientific research to end users in the field during these early years. These stories are followed by some thoughts on frontiers and the future vision for visual analytics. Finally, we introduce the eight papers in this special issue, each one addressing part of that vision.

Background of Visual Analytics

The formation of the U.S. Department of Homeland Security (DHS) National Visualization and Analytics Center™ (NVAC™)[1] in March 2004 resulted in increased interest in the field of visual analytics. In 2005, a diverse team of academic and laboratory researchers, government managers, and industry scientists turned a vision into a science direction – one published in the book *Illuminating the Path: The R&D Agenda for Visual Analytics*.[2] Shortly after that book's publication, five university-led Regional Visualization and Analytics Centers (RVACs) were established at Stanford University, the University of North Carolina Charlotte with Georgia Tech, Penn State University with Drexel University, Purdue University, and University of Washington. Also, at that same time, many other researchers around the world were developing similar or complementary visions and offering new opportunities for collaboration. Special issues of magazines and journals provided early outlets for emerging research and applications within visual analytics.[3–6] Also in 2005, NVAC began hosting semi-annual Consortiums to bring academia, industry and national laboratories together with end users, government sponsors and international partners to advance this new, potentially significant field of research.

To further build the scientific community, in 2006 IEEE launched the Symposium on Visual Analytics Science and Technology (VAST), the first international symposium dedicated to advances in visual analytics science and technology. Since then, several topical workshops have been held on financial analytics, composition and active products, and mathematic foundations of visual analytics. The latter topic set the stage for the

This article is a product of a workshop on the Future of Visual Analytics, held in Washington, DC on 4 March, 2009. Workshop attendees included representatives from the visual analytics research community across government, industry and academia. The goal of the workshop, and the resulting papers, was to reflect on the first 5 years of the visual analytics enterprise and propose research challenges for the next 5 years. The article incorporates input from workshop attendees as well as from its authors.

Received: 26 May 2009
Revised: 7 July 2009
Accepted: 8 July 2009

- Exploratory Spatial Data Analysis (ESDA)
  - EDA +
  - describe spatial distributions
    - dynamic statistical maps
  - identify atypical spatial observations
    - spatial outliers
  - discover patterns of spatial dependence and spatial heterogeneity
    - spatial clusters, hot spots, cold spots
    - spatial structural breaks

From Mapping to Geovisualization

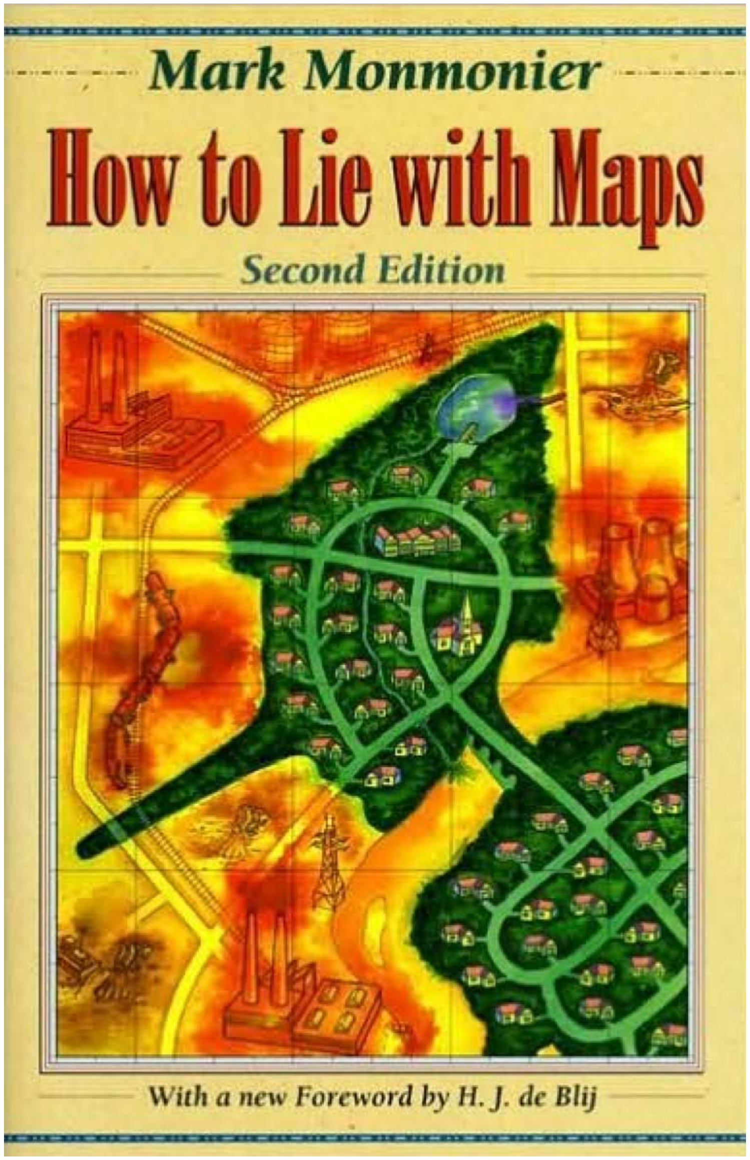
- What Is a Map

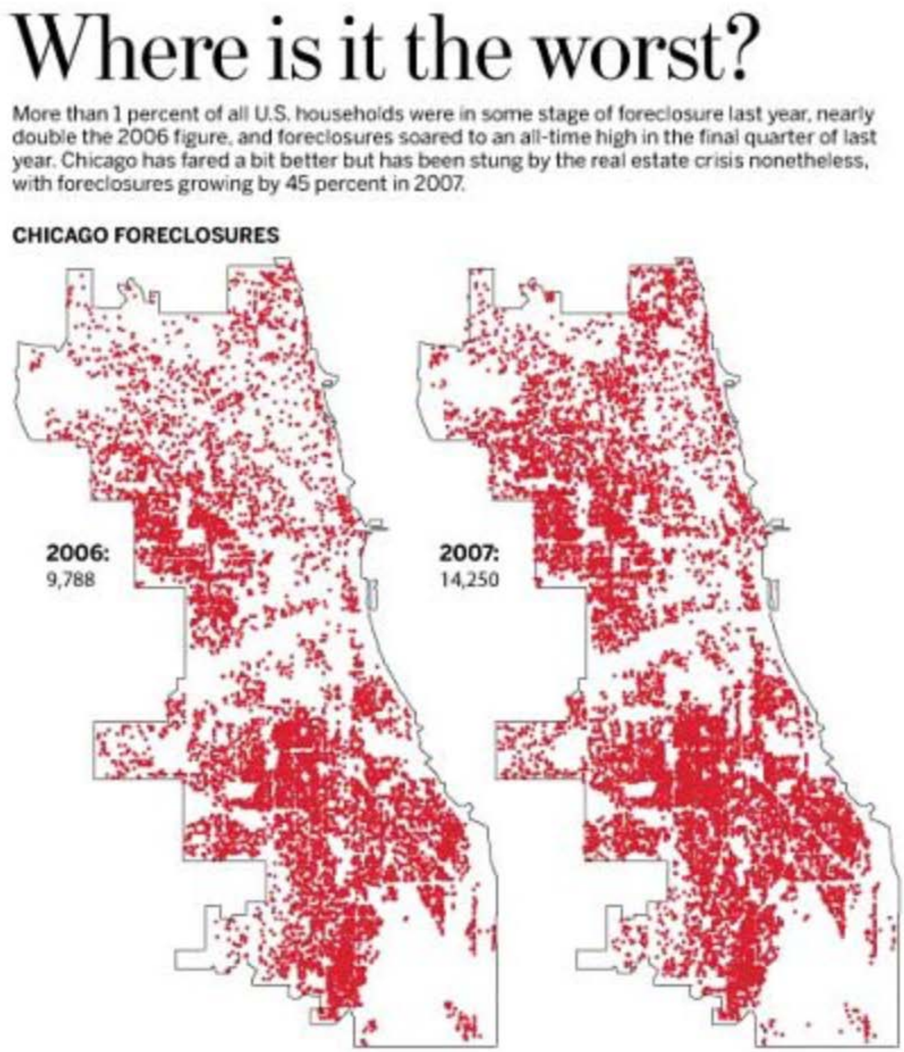
  “a collection of spatially defined objects” (Monmonier)

  importance of depicting location

  importance of representing value

- How to Lie with Maps
    - many design issues
    - legends, colors, intervals
    - projections
    - human perception can be tricked
        - political maps





http://xefer.com//2008/04/maps

- Geovisualization

  map + scientific visualization

  map as presentation vs map as part of the analysis

  interactive mapping

- Maps and Knowledge Discovery

    exploration, synthesis, presentation, analysis

  - visual popout

    abductive approach = pattern discovered along with a hypothesis

    contrast with deductive or inductive

    interaction between data exploration and human perception

- Geovisual Analytics

  leverages both geovisualization and visual analytics

  interactive mapping

  animation

  linking and brushing

PENNSTATE

GeoVISTA Center

SITE SEARCH

Search GeoVISTA | Go

Or Search: Penn State, People, Departments

RESEARCH THEMES

- GeoVisual Analytics
- Knowledge Management & Geocollaboration
- Spatial Cognition & Human Factors
- Risk Assessment & Spatial Decision Support
- Geographic Representation
- GeoSemantics

SOFTWARE TOOLS

- GeoViz Toolkit
- Visual Inquiry Toolkit
- GeoVISTA CrimeViz
- HerbariaViz
- United States Cancer Atlas

RELATED RESOURCES

- Video: Flu Data Analysis with GeoViz Toolkit

GeoVisual Analytics

Tweet | Like 0

Representing, analyzing, modeling and extracting meaning from complex heterogeneous geospatial datasets requires new approaches that can scale up to current and future data complexity and data volume. Our work addresses a wide variety of issues, including:

- the development of 'complex' spatiotemporal systems with emergent properties,
- new techniques for data mining, knowledge discovery, visualization (for application to geospatial and spatiotemporal information about the past, present, and future),
- advanced and semantically aware spatial databases that can represent and integrate both the data and the various higher level knowledge constructs, such as categories and relationships that emerge from the data during knowledge construction and
- developing a geographical agent modeling environment for investigating human activities.

These activities, when integrated, support the entire geo-scientific process, from initial exploration of data, hypothesis generation, concept discovery, model formulation, analysis and validation, and, when fused together seamlessly in GeoVISTA Studio, will form a complete Problem Solving Environment (PSE) for teams of scientists to use, thus supporting our geocollaboration focus. By bringing these activities together in GeoVISTA Studio we avoid many of the integration problems that plague traditional computational analysis. To accomplish this goal, Center affiliates and their collaborators are working to integrate methods and tools that span many disciplines including machine learning, pattern recognition, agent and cellular modeling, data mining, multivariate information visualization and spatial statistics.

RELATED PROJECTS

- VACCINE: Visual Analytics for Command, Control, and Interoperability Environments
- GAIDD: Geovisual Analytics for Infectious Disease Dynamics
- Vaccine Modeling Initiative
- Geovisualization and Spatial Analysis of Cancer Data
- STNexus: An Integrated Database and Visualization Enviroment for Space-Time Information Exploitation

See all projects . .

RELEVANT PAPERS

Robinson, A. (2009). Needs Assessment for the Design of Information Synthesis Visual Analytics Tools, IEEE International Conference on Information Visualization

Roth, R.E, MacEachren, A., McCabe, C. (2009). A workflow learning model to improve geovisual analytics utility. Proceedings of the

www.geovista.psu.edu

- Dynamic Graphics

  different views to represent the data

  focusing individual views

  linking multiple views

  arranging many views

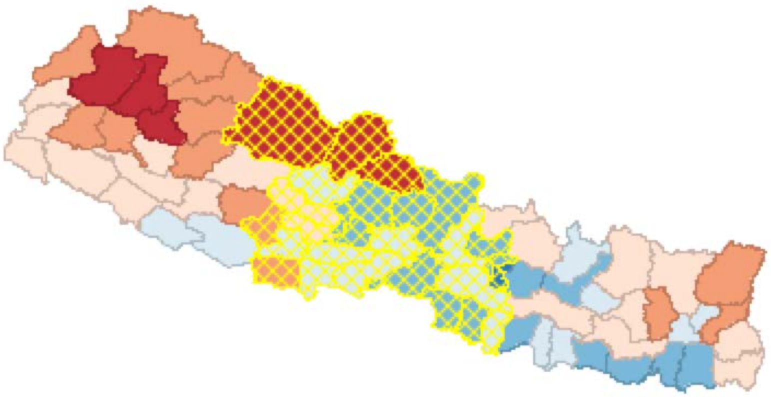
- Linking and Brushing

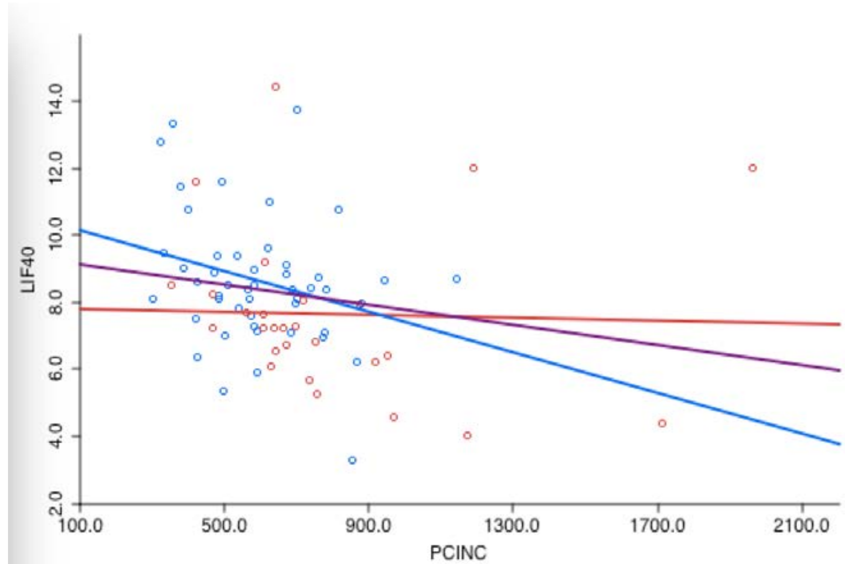
    linking

        selection in one view (graph) is simultaneously selected in all views

    brushing

        dynamically changing the selection updates all views





#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
75	0.0366	9.29	0.658	14.1	0	-0.00152	0.000911	-1.67	0.0999
27	0.000987	7.83	1.17	6.69	5.2e-07	-0.000212	0.00135	-0.157	0.876
48	0.077	10.4	0.966	10.8	3.24e-14	-0.00303	0.00155	-1.96	0.0562

Chow test for sel/unsel regression subsets: distrib=F(2,71), ratio=1.875, p-val=0.1609

linked map and graph

Mapping Basics

- Choropleth Map

    not chloro!

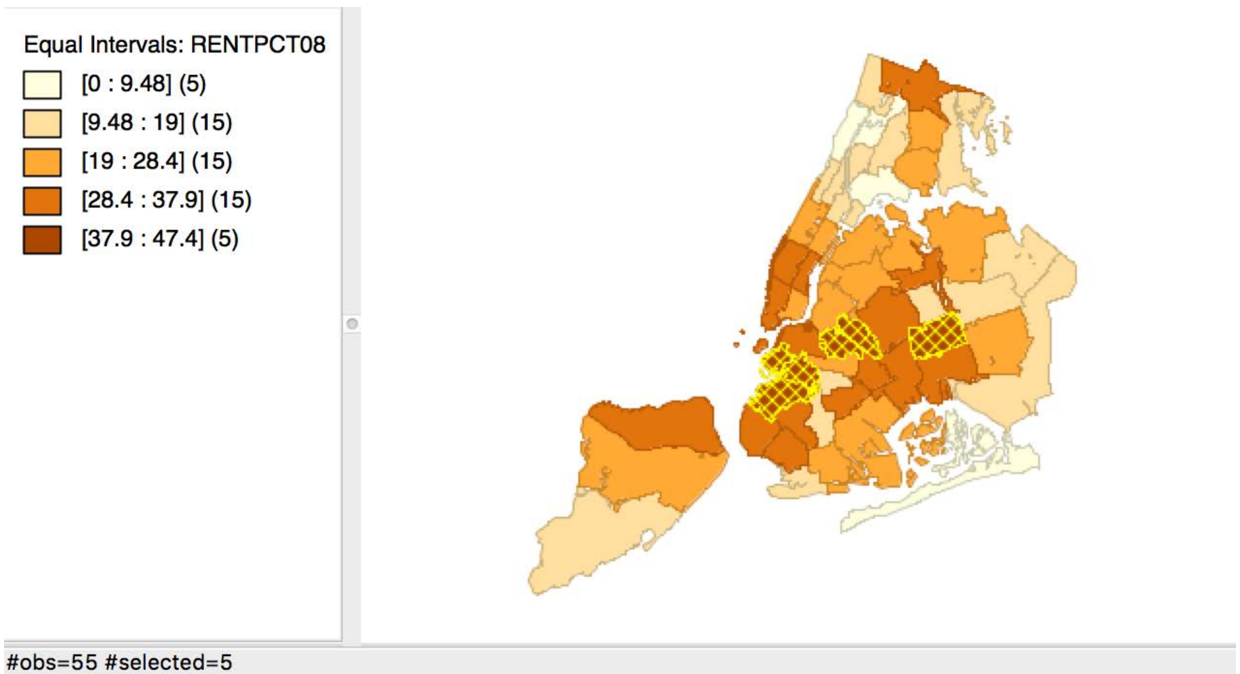
        choros = region

    visualize a spatial distribution

        map counterpart of a histogram

        discrete approximation of the distribution

        all observations in the same value interval get the same color



histogram and equal intervals choropleth map

- Choice of Intervals

  cut points

  equal interval, natural breaks (Jencks), manual

  statistical criteria

  equal share (quantile), standard deviational units

- Map Design Issues
  - choice of colors
    - perception of pattern
    - red = hot, danger; blue = cool
  - misleading role of area
    - larger areas seem more important
  - legends
    - sequential
    - diverging
    - categorical

Statistical Maps

- Quantile Map

    data sorted from low to high

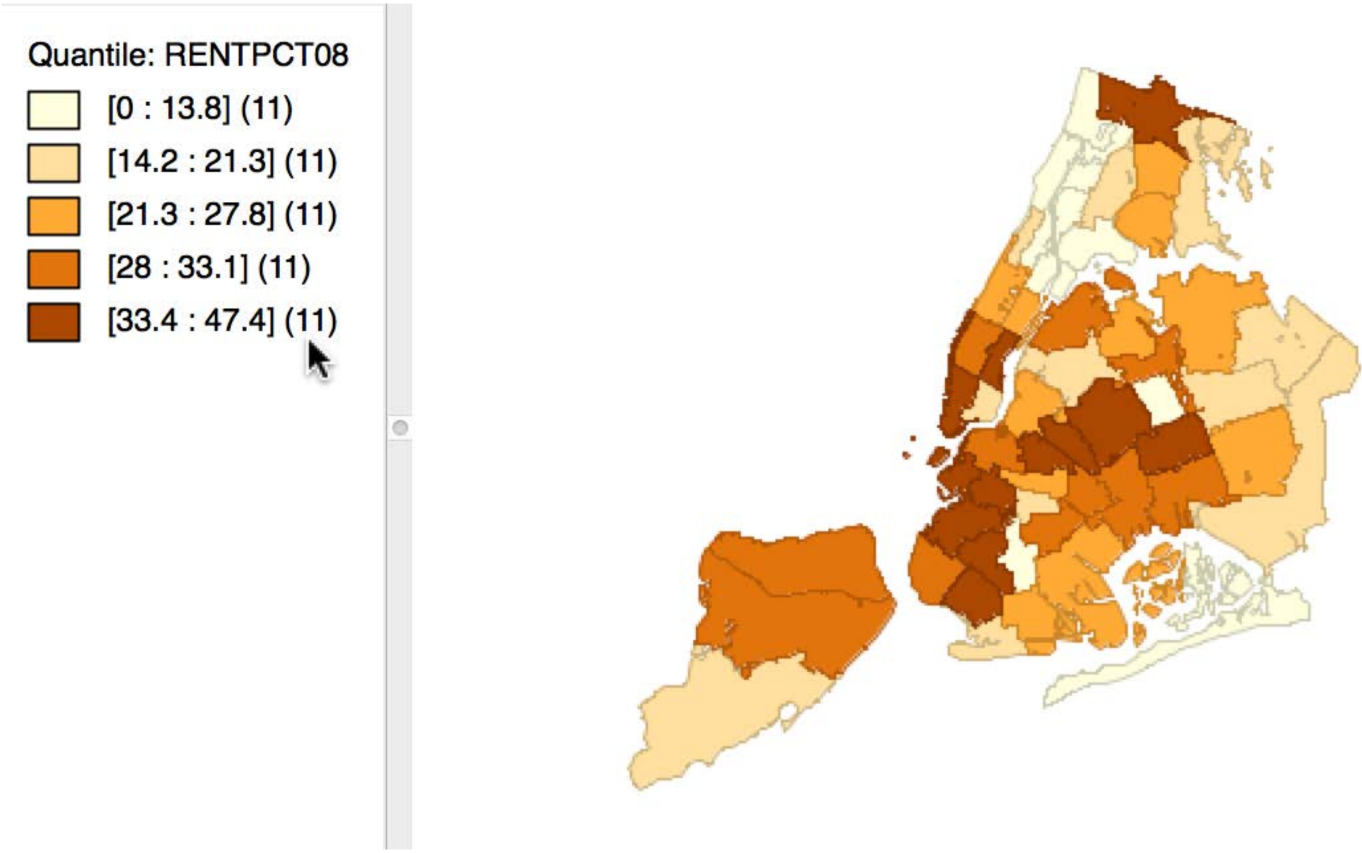
    equal number of observations in each interval

    examples

        quartile map (4 categories)

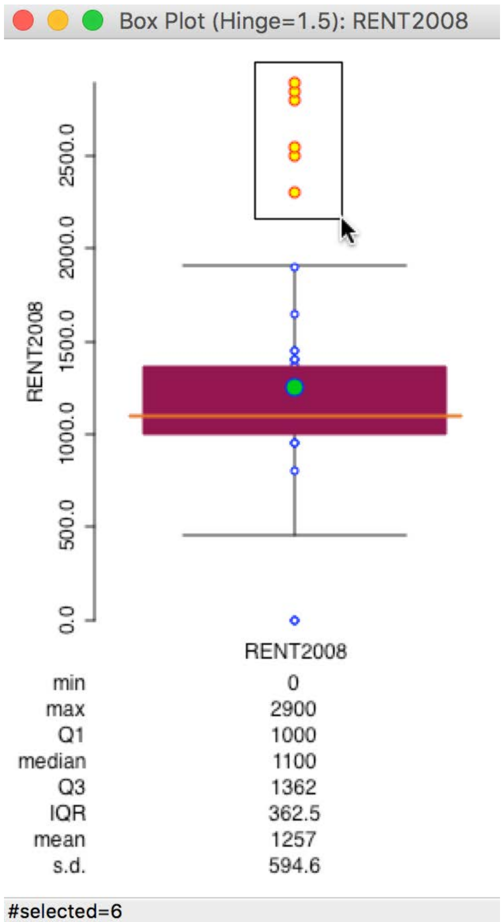
        quintile map (5 categories)

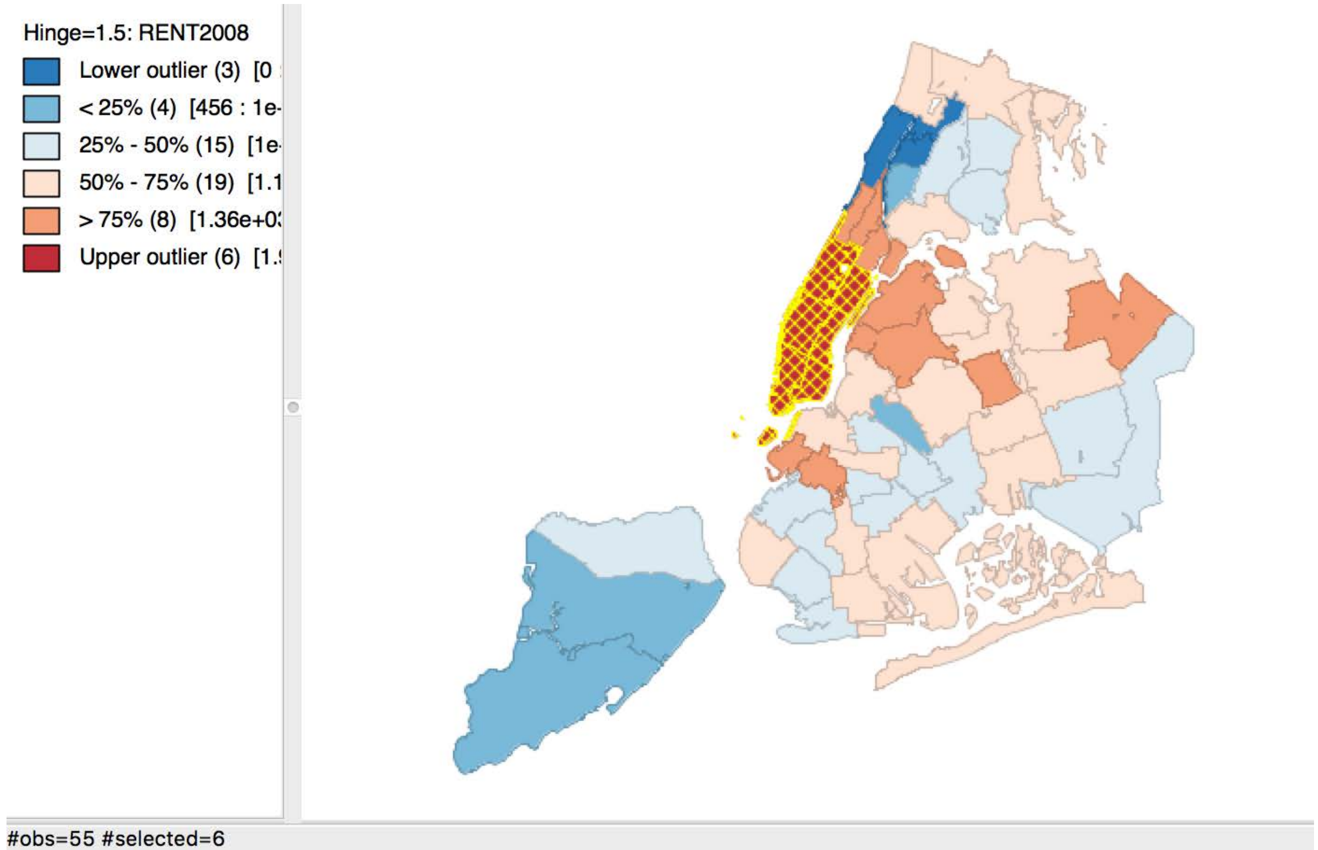
    possible issues with ties



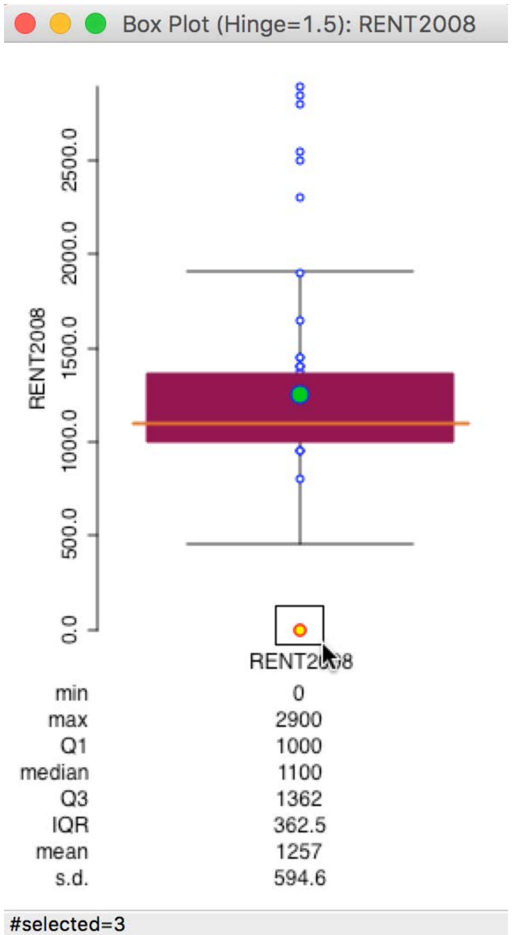
quintile map (NYC % rental units)

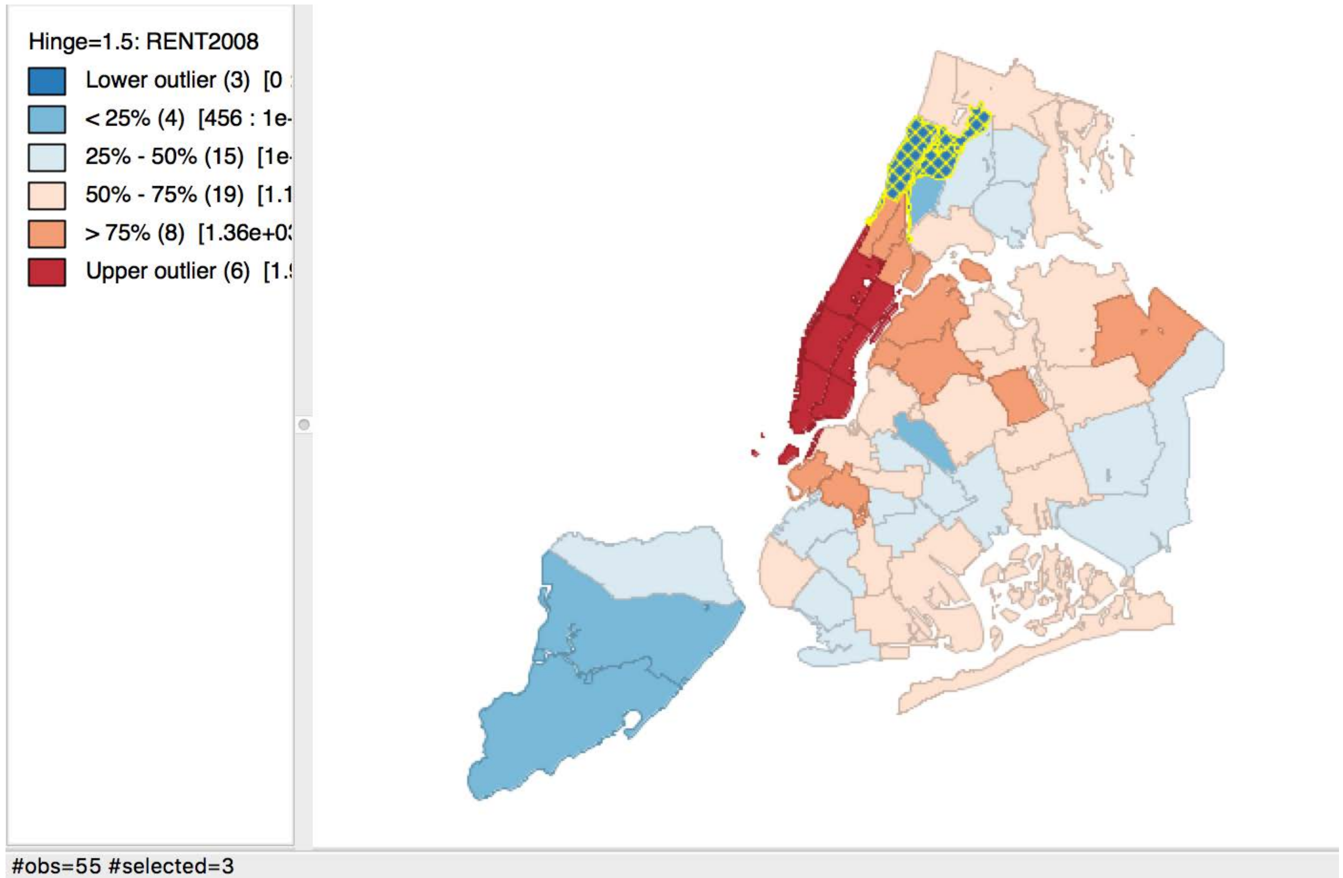
- Box Map
    - identifying outliers
        - same principle as in box plot
        - fence = median + 1.5 IQR or + 3 IQR
            - IQR = inter quartile range, 25% to 75%
    - six intervals
        - same principle as quartile map
        - outliers identified as a separate category





upper outliers in box plot and box map
(NYC median rent 2008)





lower outliers in box plot and box map
(NYC median rent 2008)

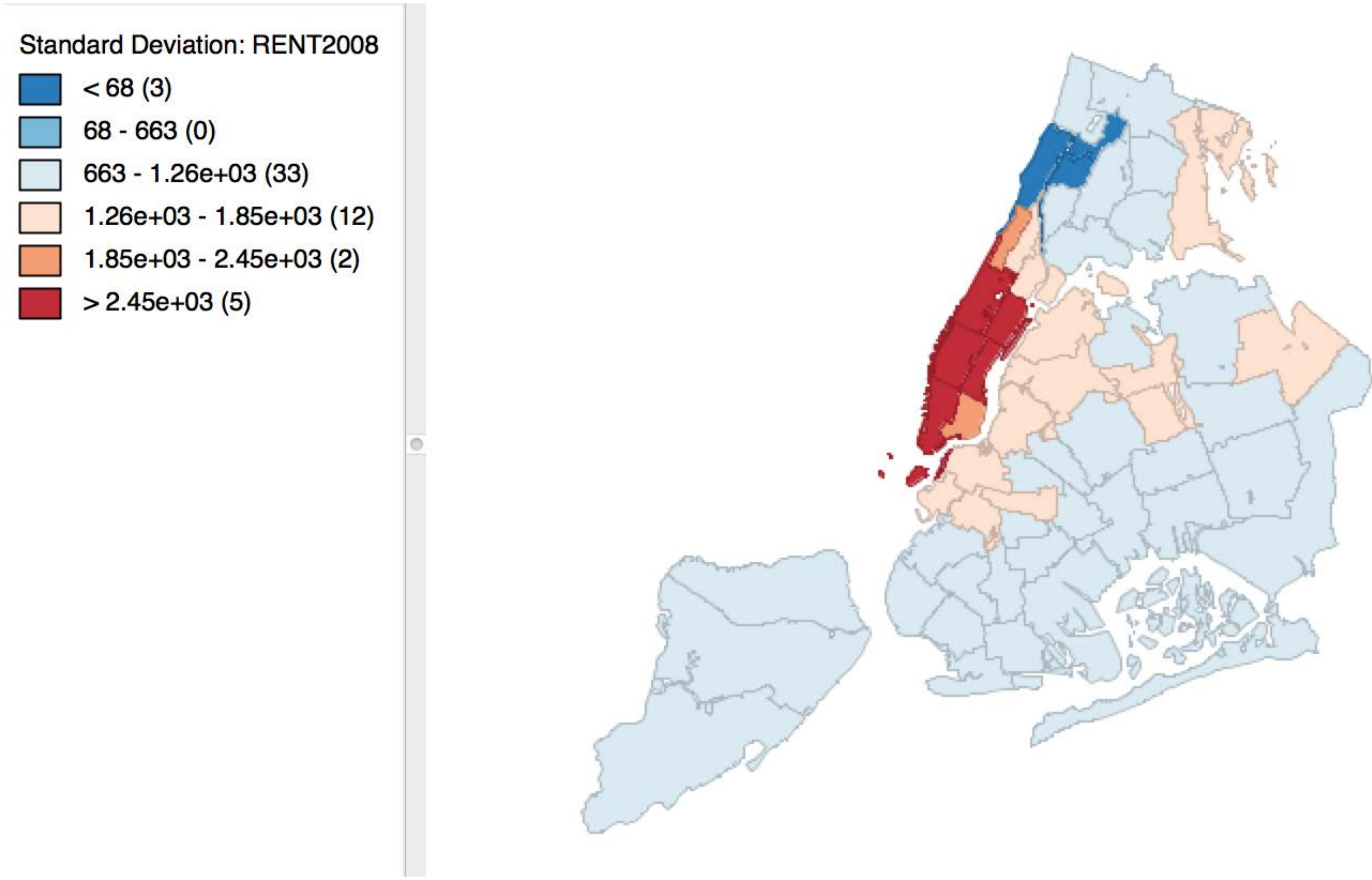
- Standard Deviational Map

  based on standardized data values

    mean = 0, standard deviation = 1

  intervals correspond to one standard deviation

  outliers are more than 2 standard deviations from the mean



standard deviational map
(NYC median rent 2008)

- Cartogram

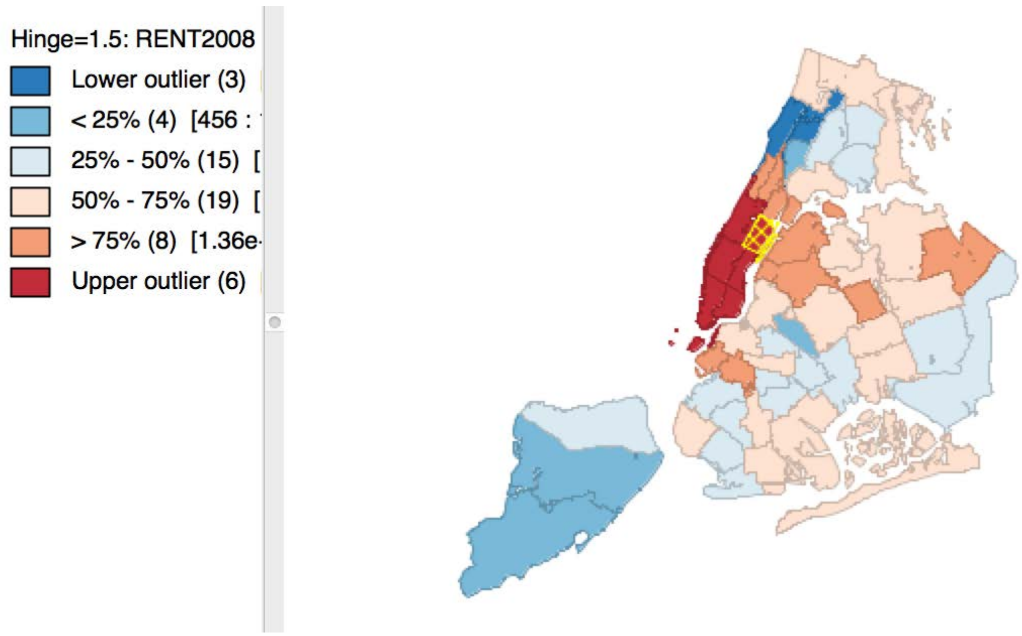
    areal unit proportional to variable of interest

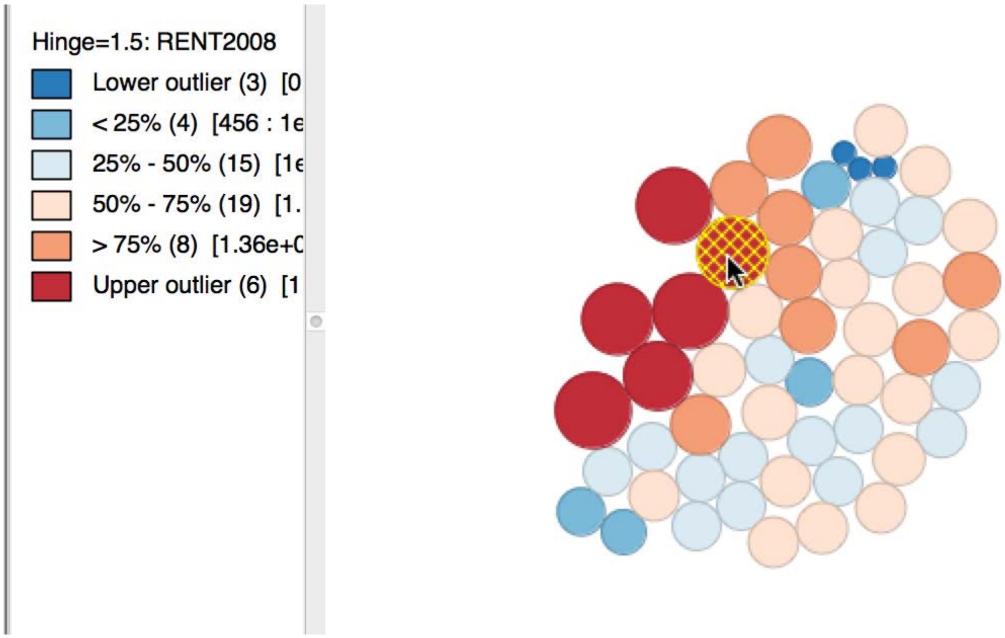
    avoid misleading effect of area

    use transformed shapes

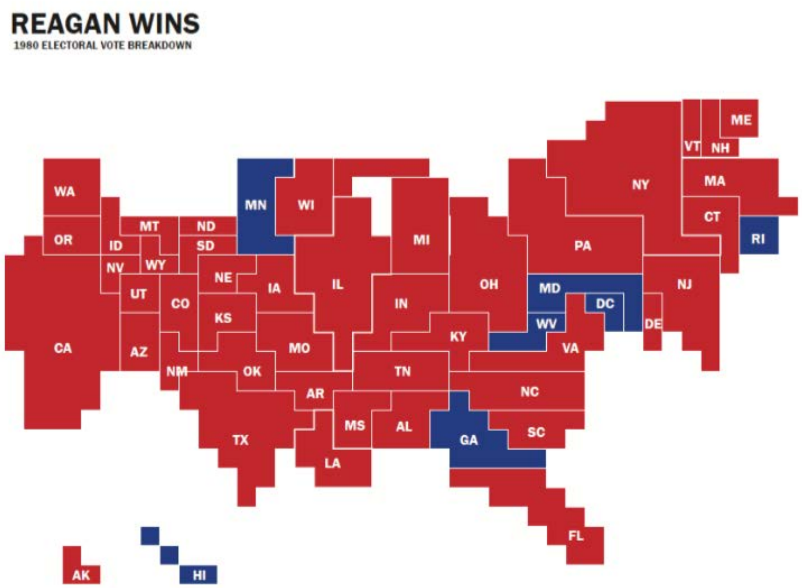
        circular cartogram

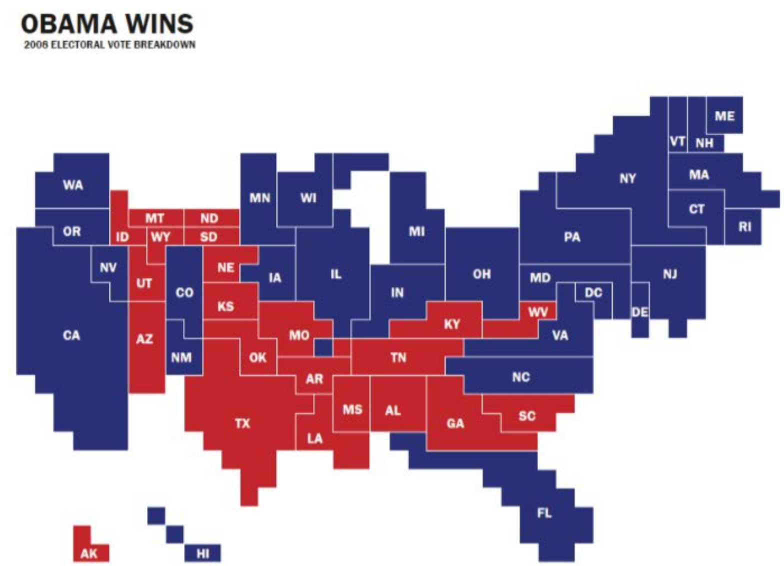
        contiguous cartogram





box map and circular cartogram





contiguous cartogram

area = number of votes in electoral college

source: Sarah Williams

- Conditional Maps

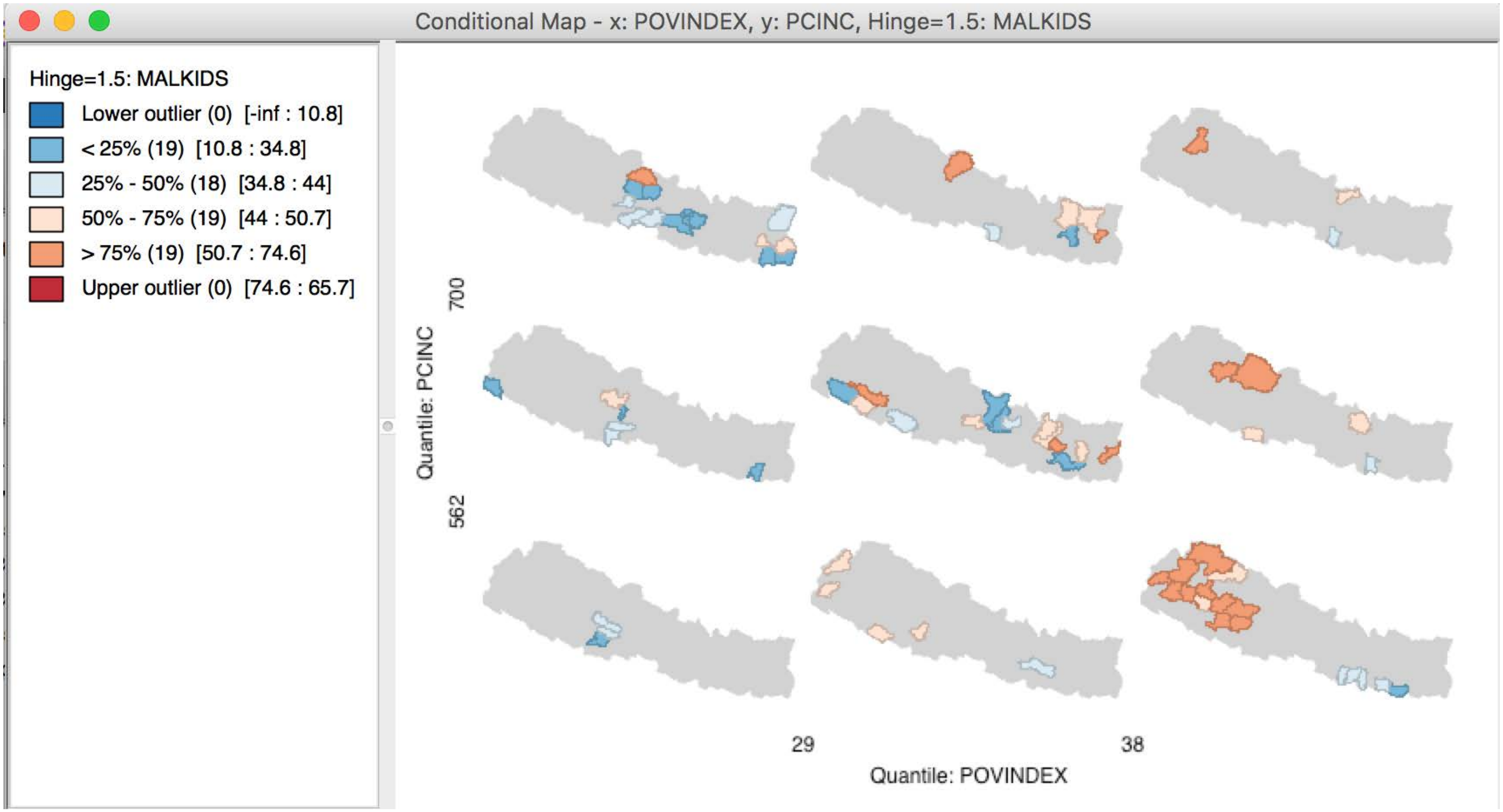
    cc maps, conditioned choropleth maps (Carr)

    special case of trellis graphs

    micromap matrix

        conditioning variables on the axes

        matrix of mini maps for the variable of interest conditioned by the values on the axes



child malnutrition cc map conditioned on poverty index and per capita income (Nepal districts)

- Map Animation

    map movie

    highlight observations in increasing or decreasing order

        one at a time

        cumulative

    visual impression of patterning/clustering

Multivariate EDA Primer

- Objectives of Multivariate EDA

  represent multi-dimensional data in two dimensions

  dimension reduction

  projection

  discover structure, interaction, patterns

- 3-D Scatter Plot

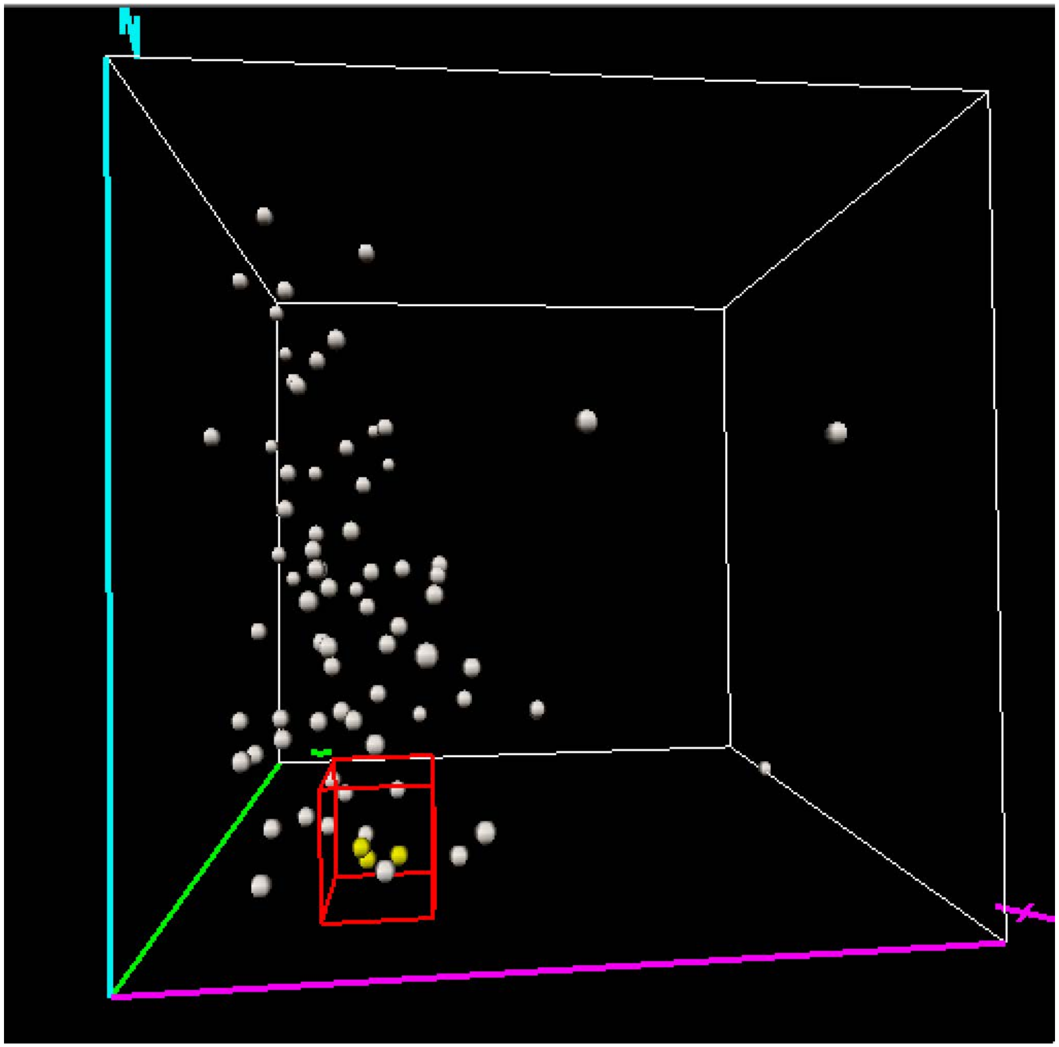
  points in a 3-D data cube

  two-dimensional analysis on side panels

  issues of perspective

    zooming, rotating

  brushing the 3-D data cube



selection in a 3D scatter plot

- Parallel Coordinate Plot (PCP)

    due to Inselberg (1984)

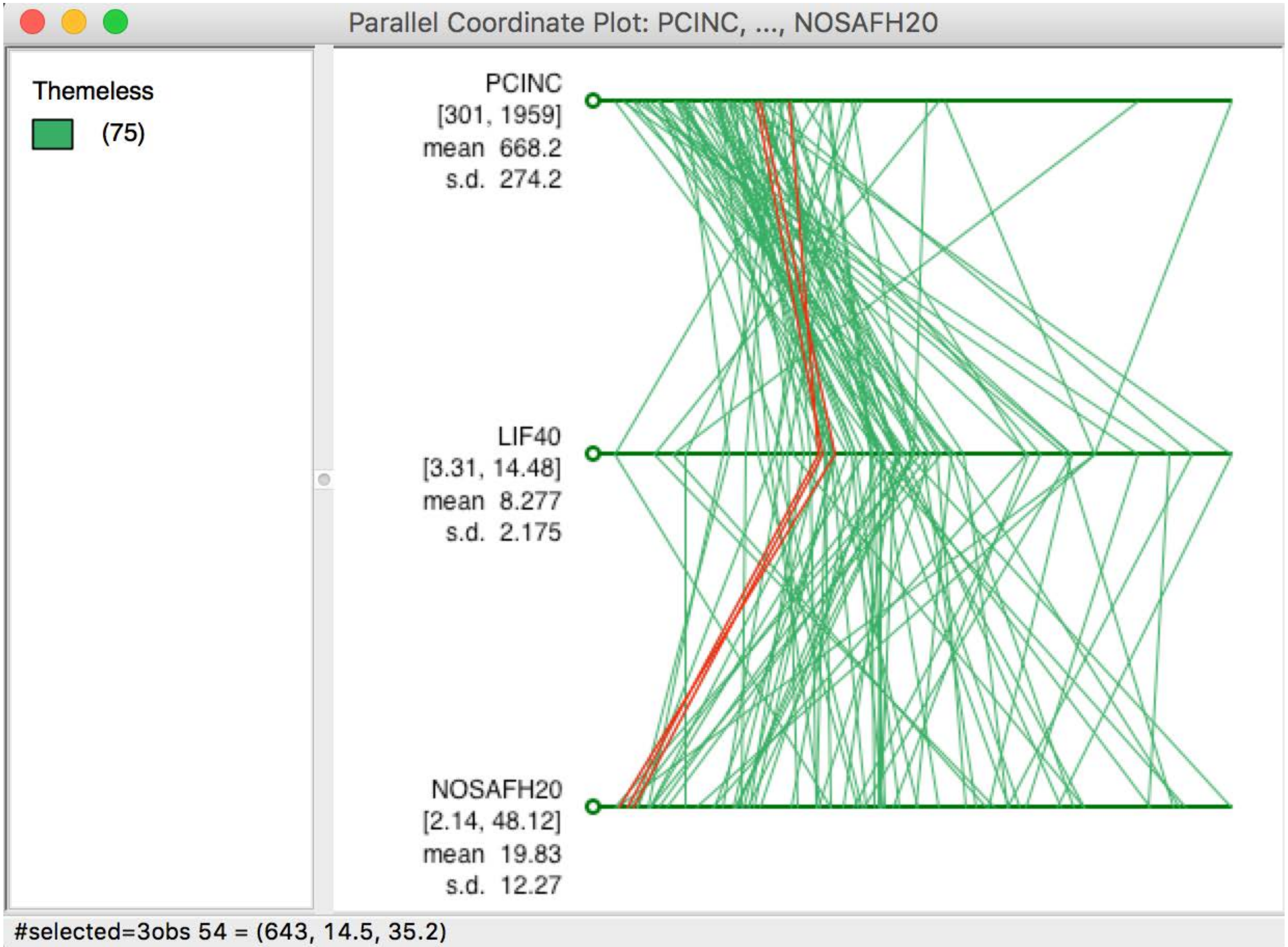
    variables

        one parallel line for each variable

    observations

        a line connecting points on the parallels

        the line is the counterpart of a point in the multidimensional data cube



selected points in PCP

- Clusters in PCP

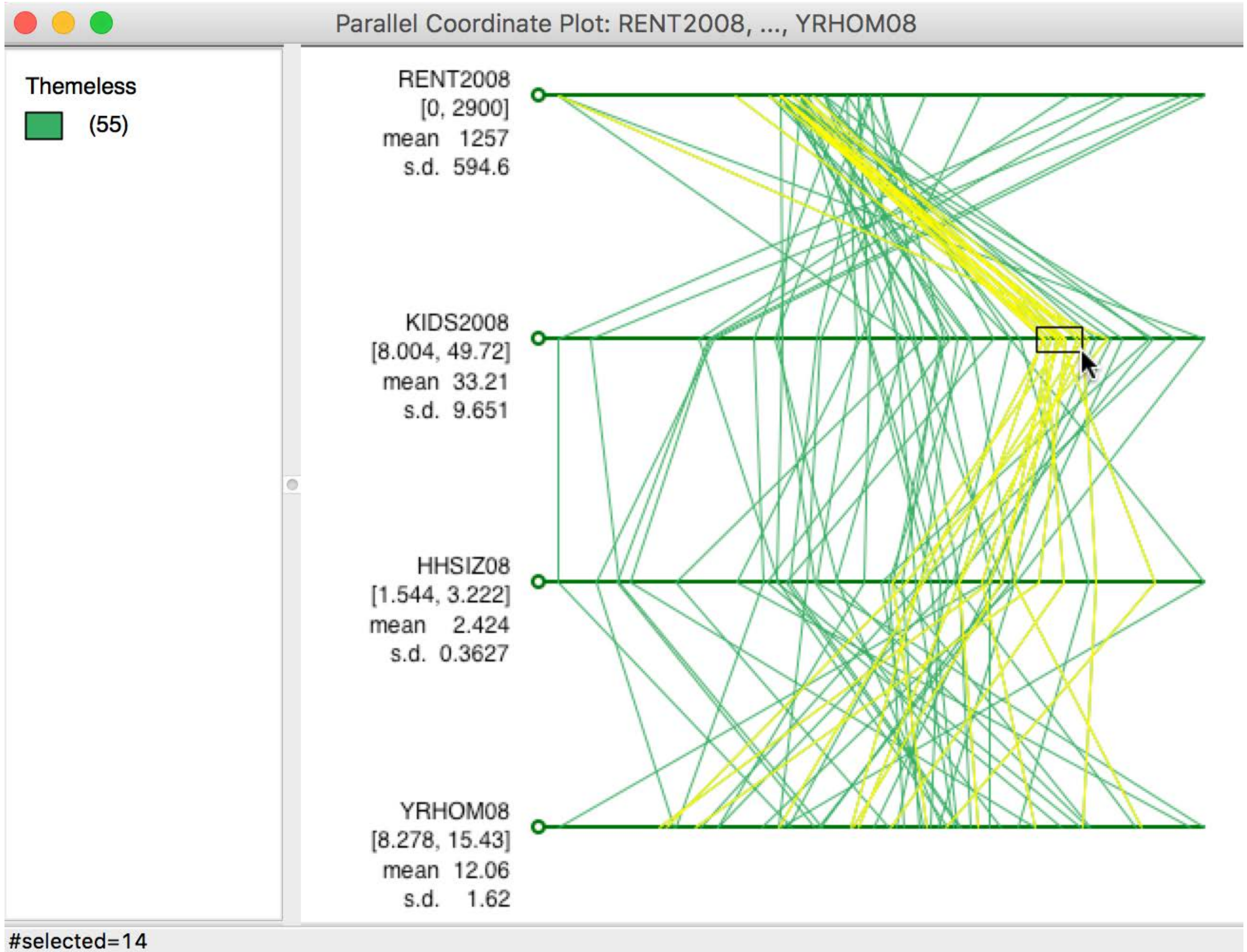
  lines that move closely together correspond to points closely together in multidimensional space

  = clusters

  visual cluster identification

  problems with large data sets

    remove clutter



brushing the PCP

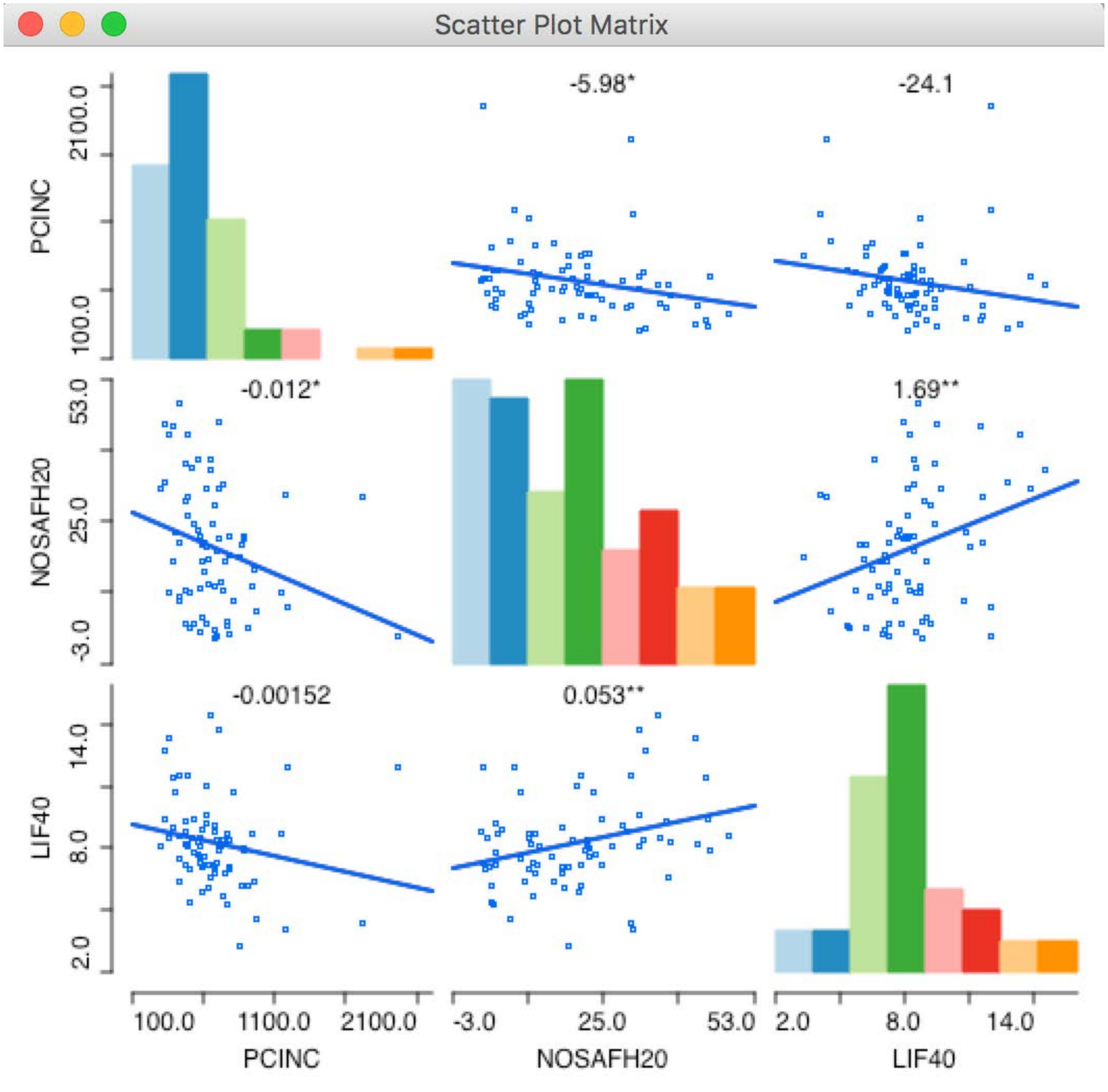
- Scatter Plot Matrix

  matrix of bivariate scatter plots

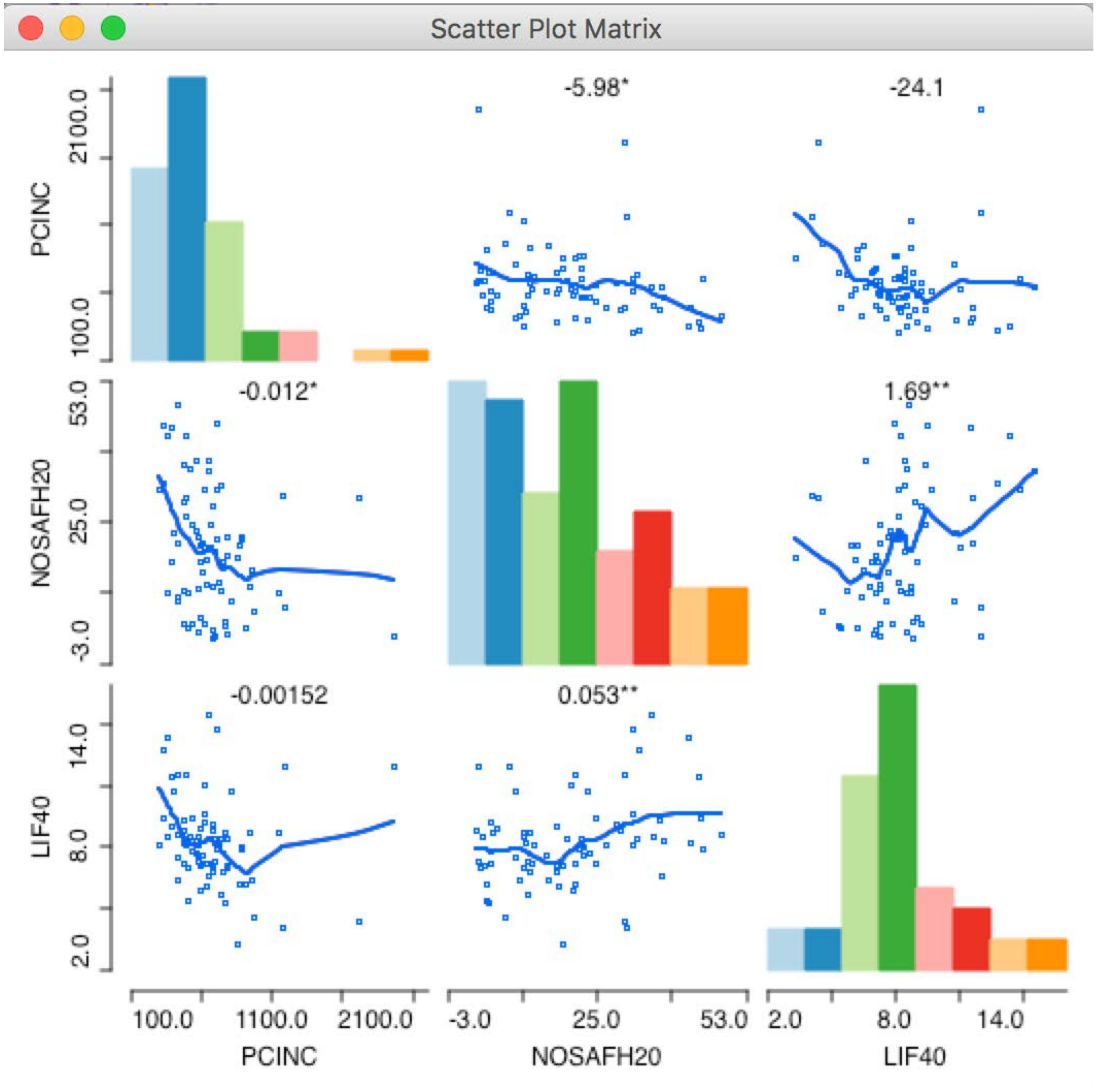
    each variable once on x-axis and once on y-axis

  univariate description on diagonal

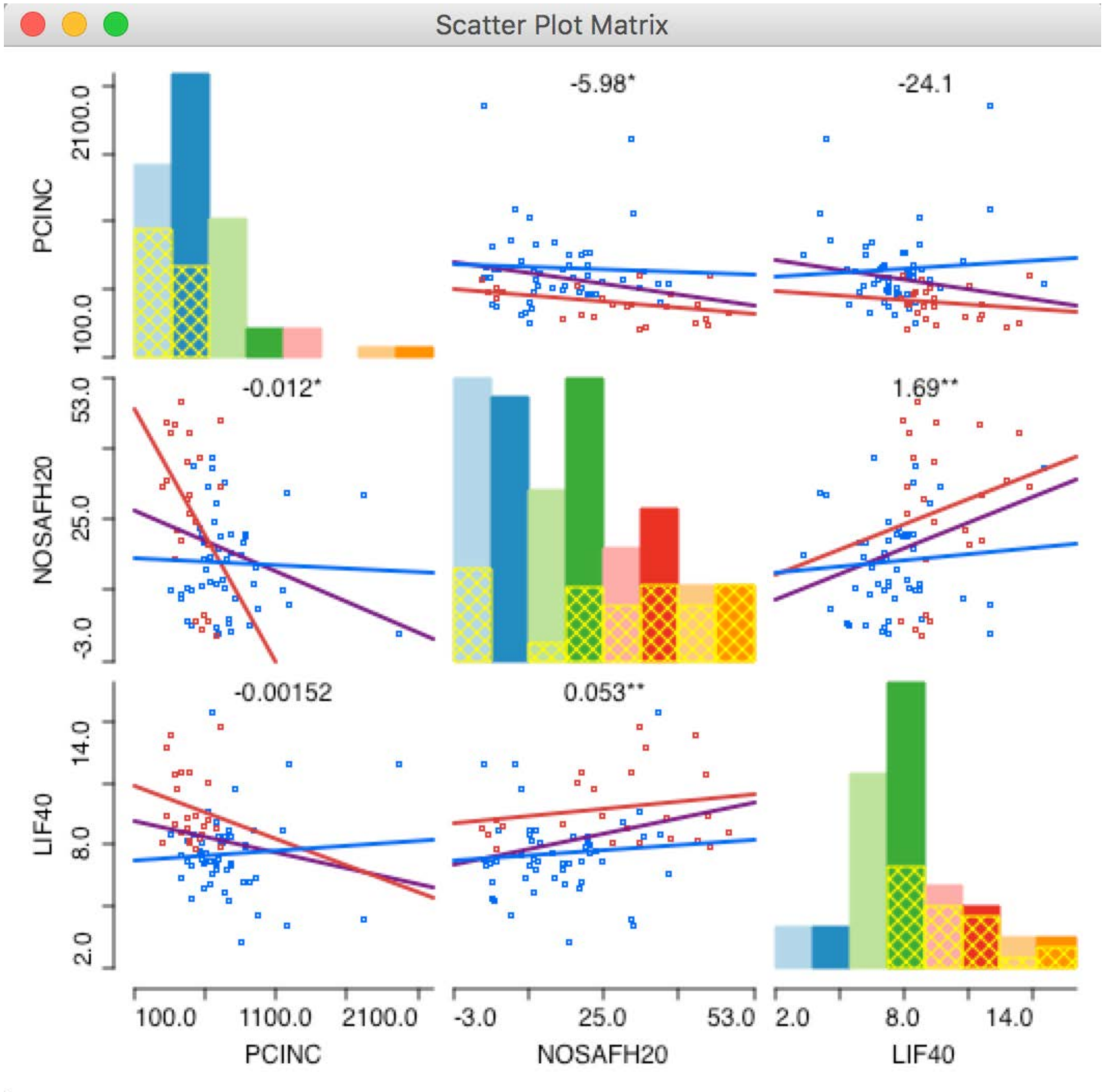
  focus on interaction effects



scatter plot matrix (Nepal districts)



scatter plot matrix with lowess smoother



brushing the scatter plot matrix

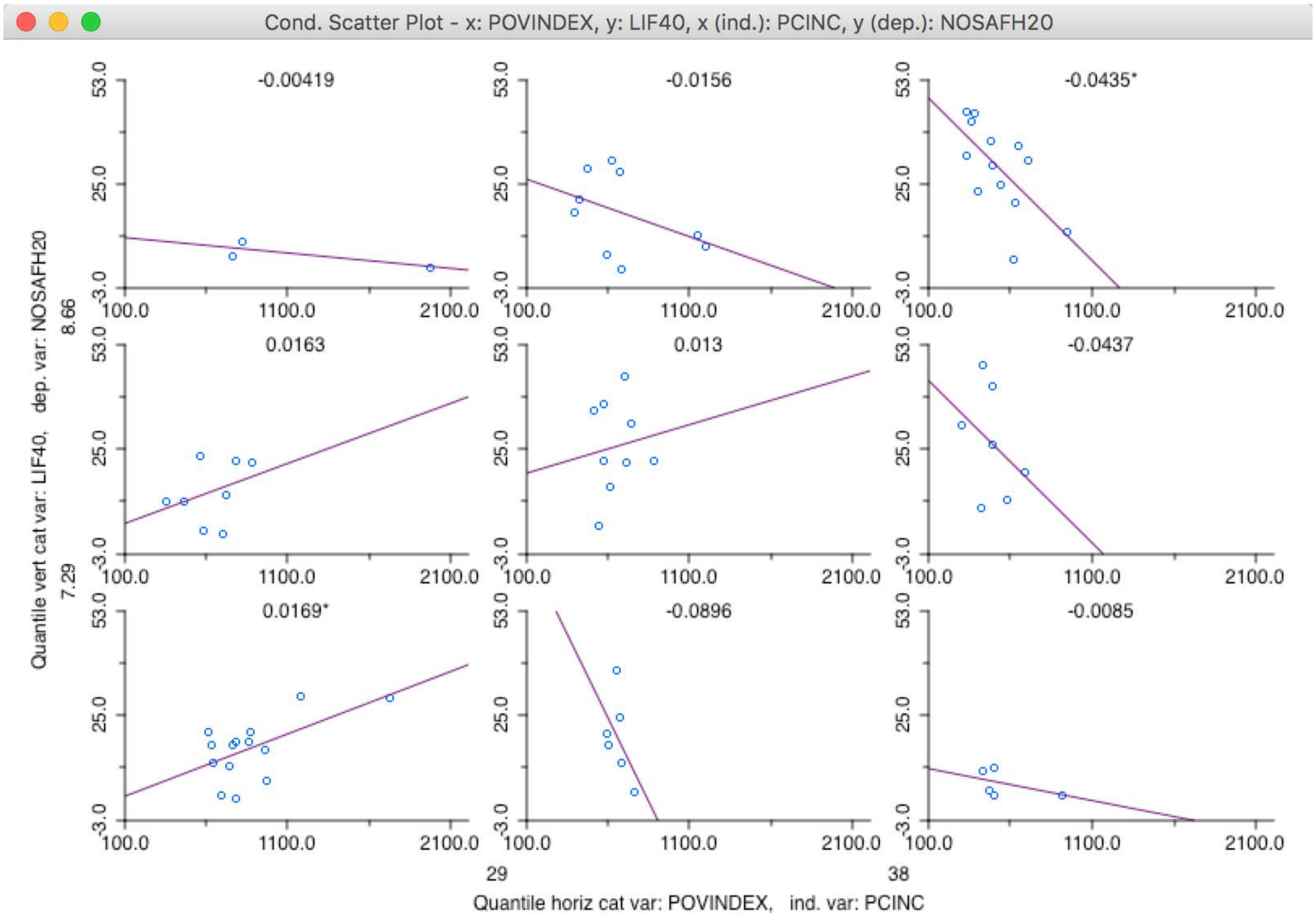
- Conditional Plots

  trellis display

  conditioning variables on the axes

  matrix of micro plots for subsets of observations that match the axes conditions

    data intervals in two dimensions



scatter plot trellis graph
scatter of per capital income on no safe water
conditioned on poverty index and life expectancy

- Interpretation of Conditional Plots

    micro plots are similar

        no effect of conditioning variables

    micro plots are different

        conditioning variables interact with variable under consideration

        effect of conditioning variables