Exploratory Data Analysis
EDA

Luc Anselin

THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

http://spatial.uchicago.edu

from EDA to ESDA

dynamic graphics

primer on multivariate EDA

interpretation and limitations

From EDA to ESDA

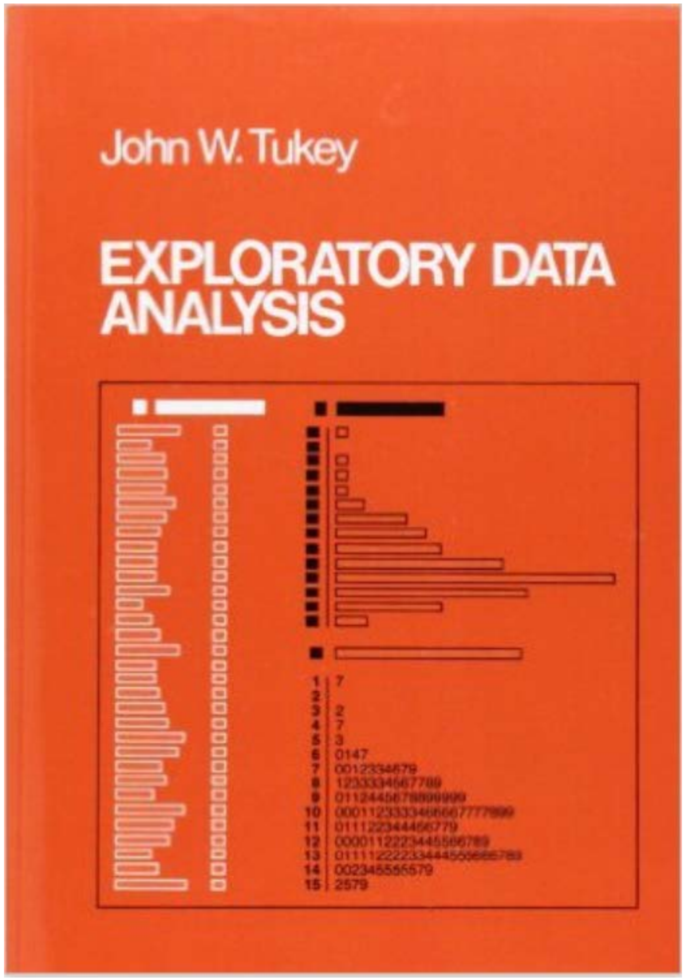
- Exploratory Data Analysis (EDA)

  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.

283

- Data Visualization

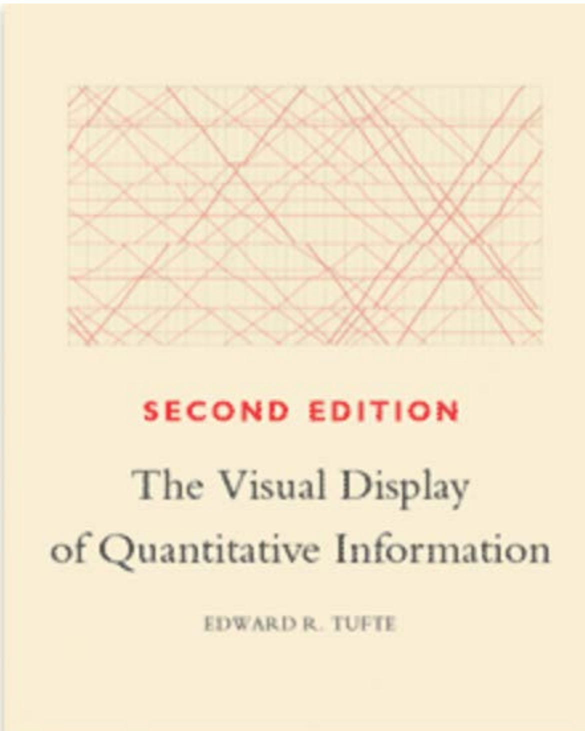
    concept of a "view" (e.g., Buja et al 1996)

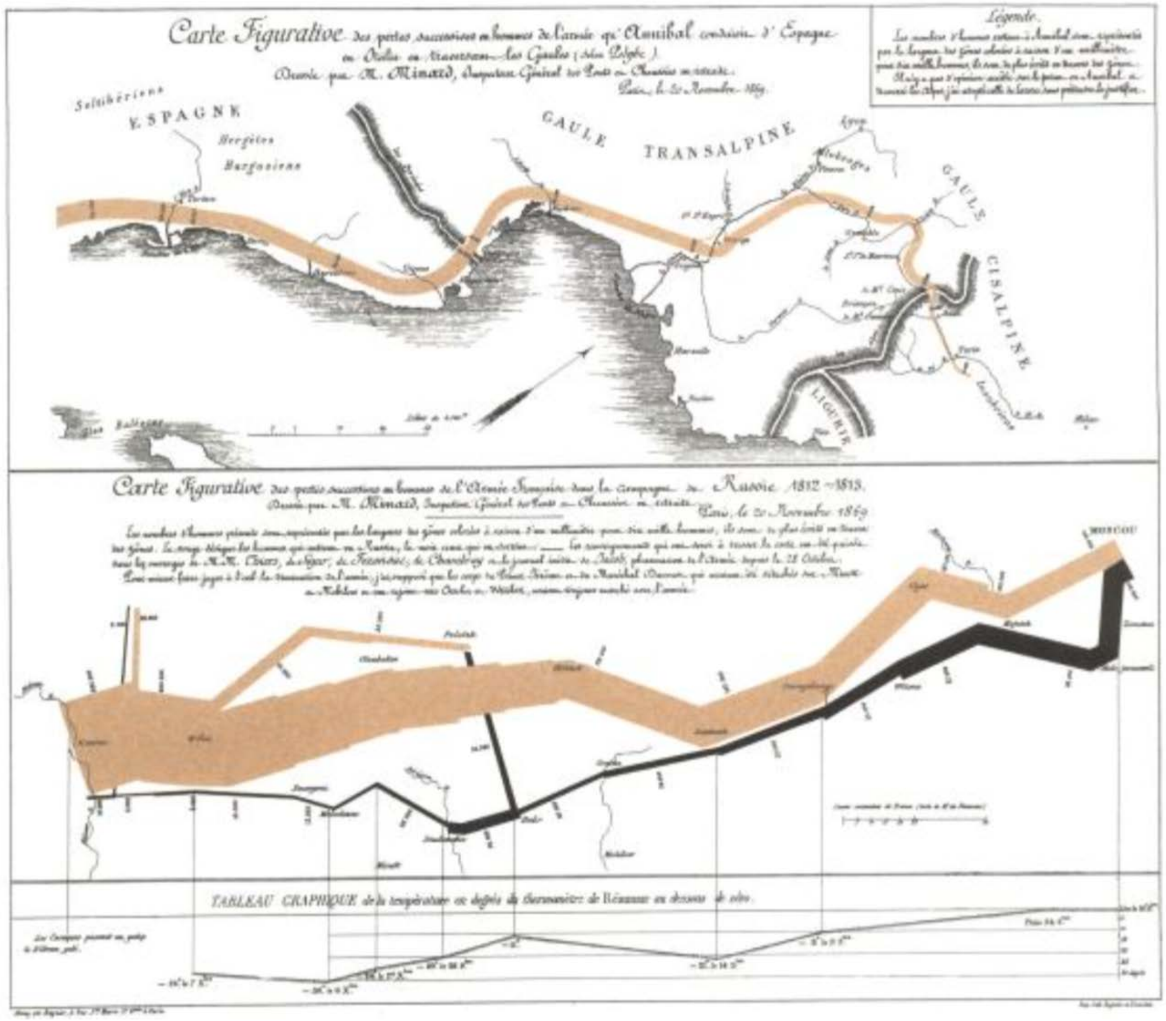
        a graphical representation and summary of the data

        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)
    - appropriate comparisons
    - 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
Russie 1812 – 1813.
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”

- Visual Analytics Tools

  synthesize information

  derive insights

  understandable assessments

  communicate effectively

  focused on policy actions

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

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

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

- Exploratory Spatial Data Analysis (ESDA)

  EDA +

  not just maps to present results, but spatial information as an integral part of the data exploration

  focus on spatial patterns

- ESDA Activities
  - 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
    - regionalization (spatial clustering)

Dynamic Graphics

Concepts

- Interactive View Manipulation

    different views to represent the data

    the analyst interacts with the data

    concept of dynamic graphics

        graphics as a tool in dynamic data exploration

- Dynamic Graphics

  three important classes of tasks

    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

Brushing the Scatter Plot

- Bivariate Scatter Plot

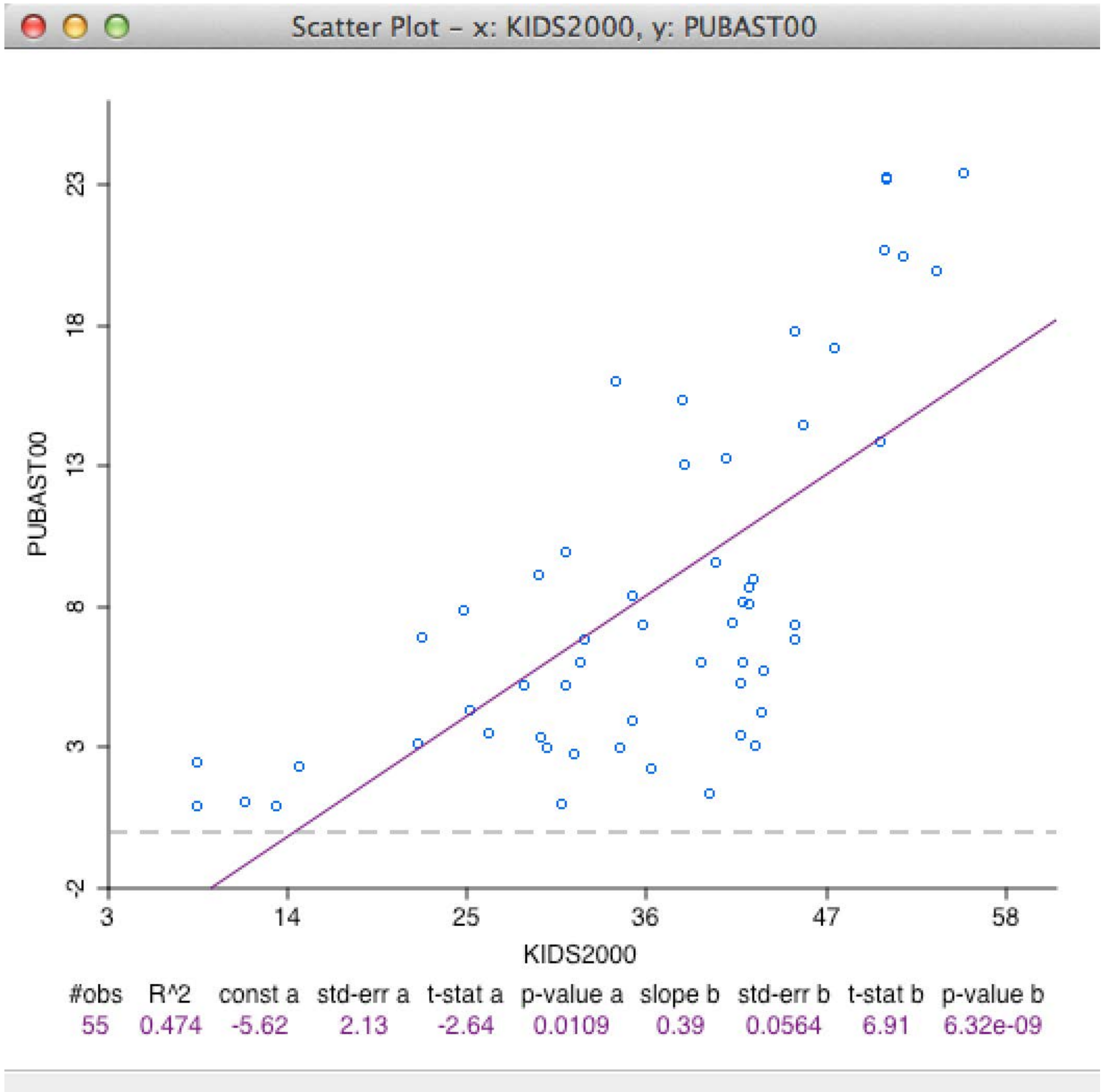
    axes = variables

    points in two-dimensional space

    smoothing the scatter plot

        linear smoother (regression fit)

        lowess or loess smoother (local regression)



scatter plot - linear smoother

- LOWESS Smoother

    local regression

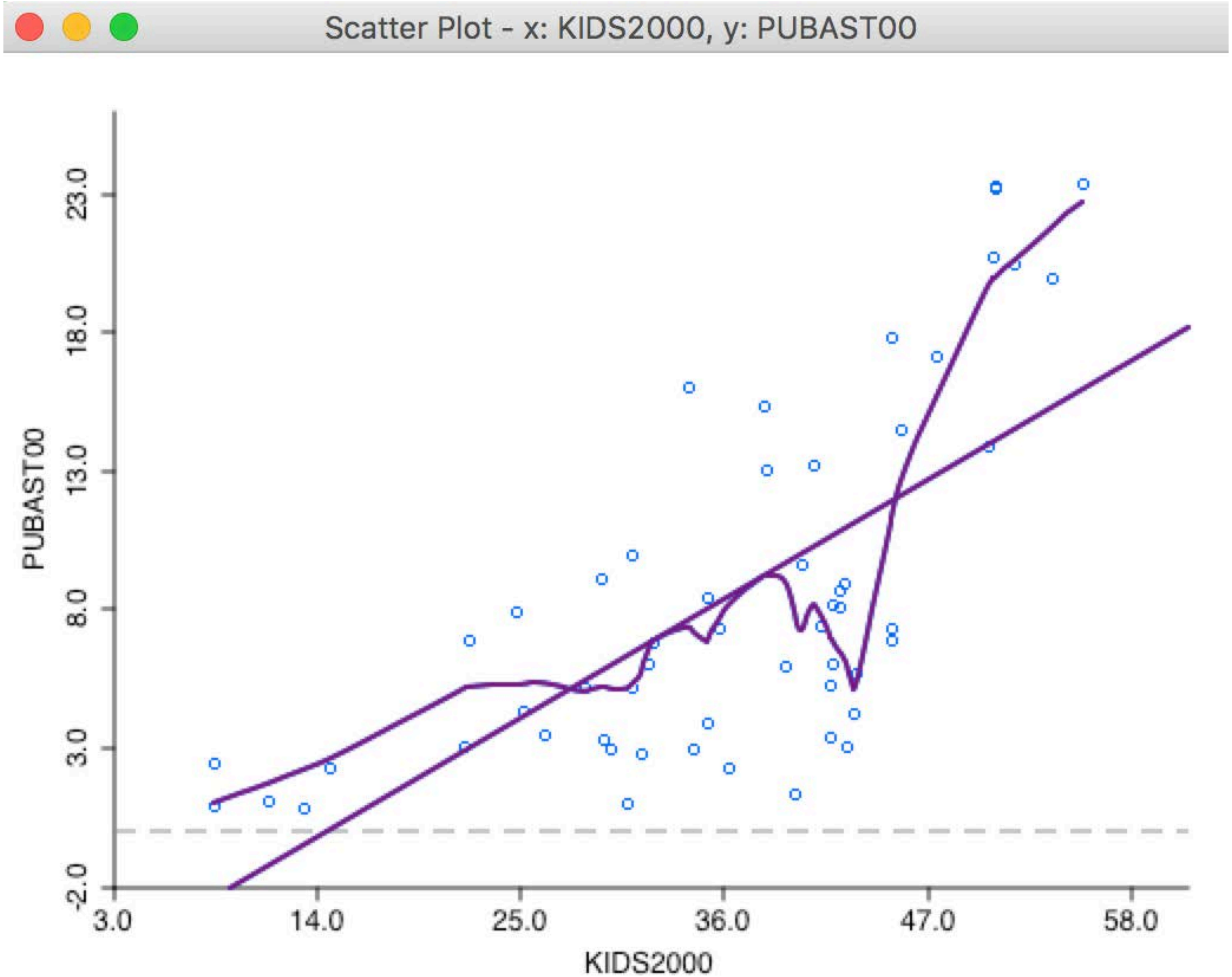
    slope based on a subset of the observations

        for each x_i, y_i, fit based on x,y pairs with x in neighborhood of x_i

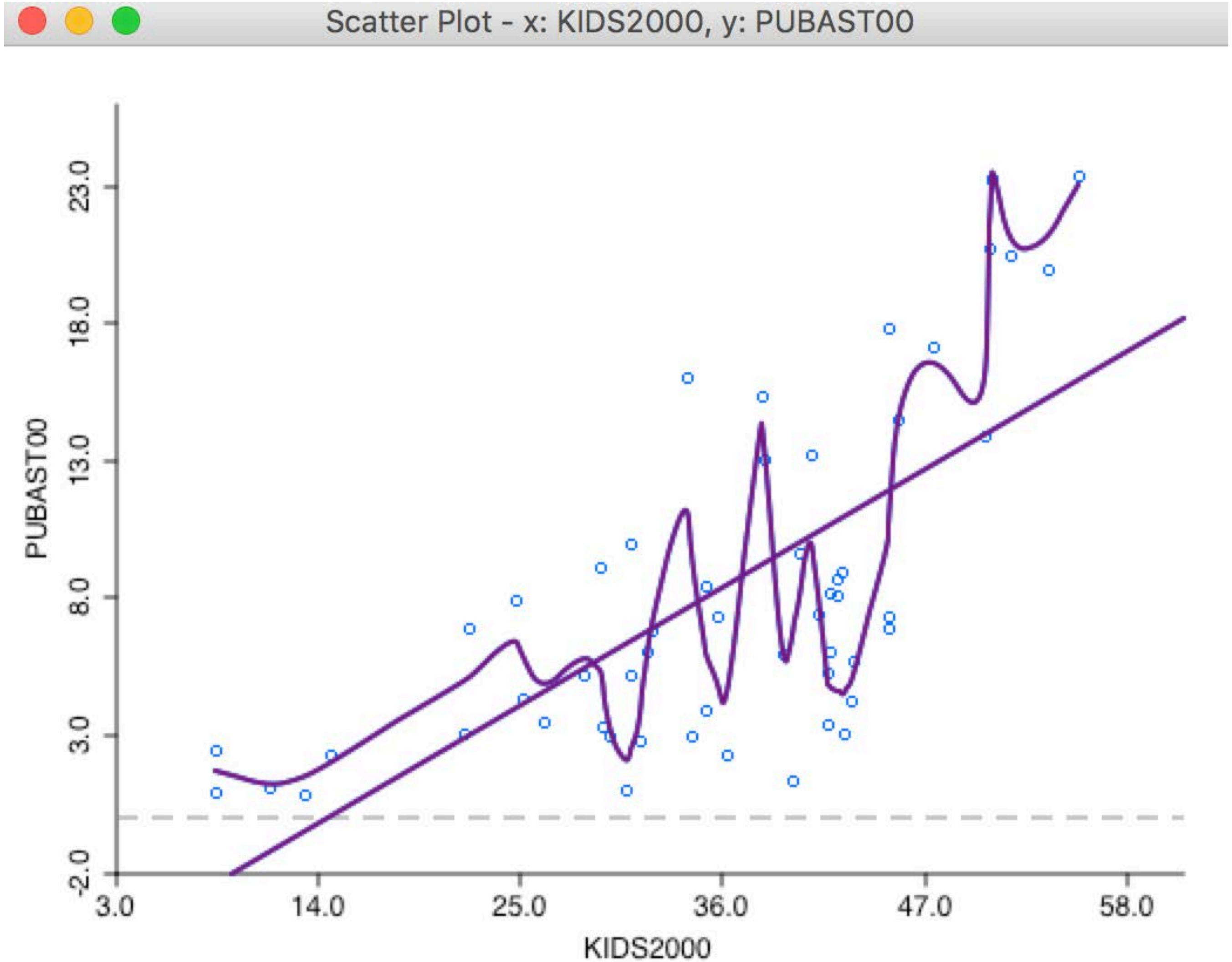
    choice of bandwidth

        short bandwidth yields spiky curve

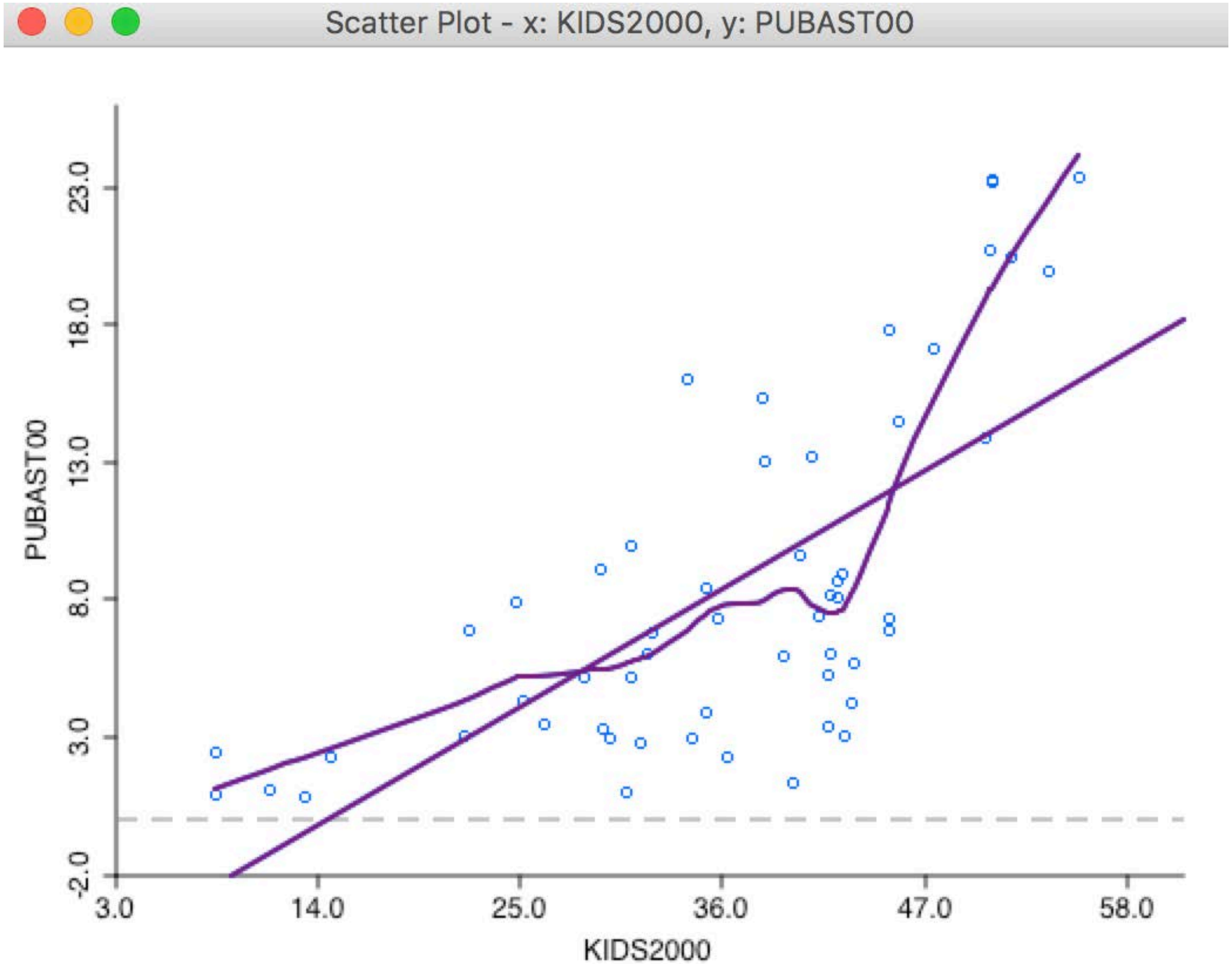
        wide bandwidth yields smoother curve



lowess smoother



effect of bandwidth - shorter bw



effect of bandwidth - wider bw

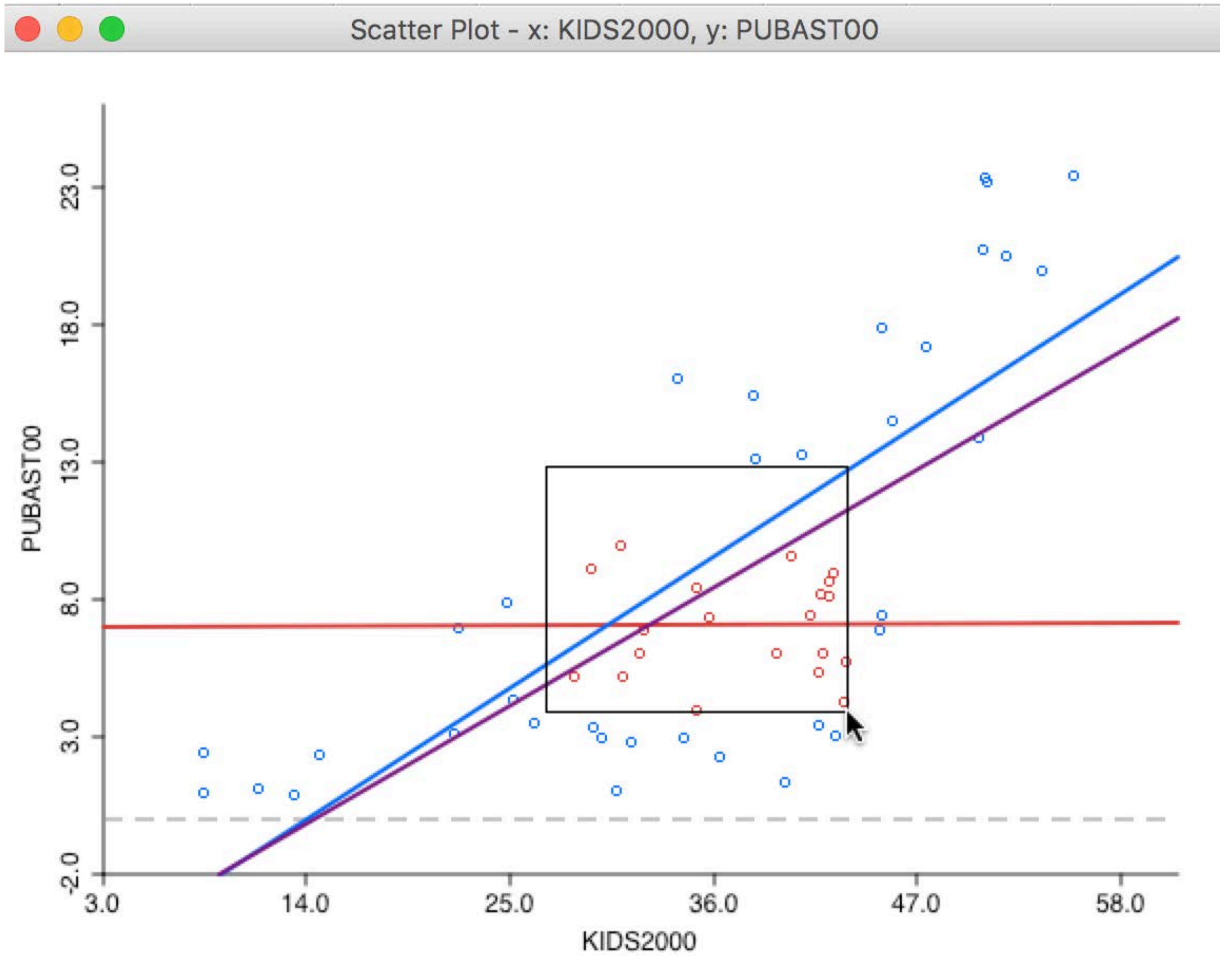
- Brushing the Scatter Plot

  a brush is a selection shape

  two slopes: selected and unselected

  as the brush moves, the slopes are recalculated in a dynamic way = dynamic brushing

  the matching observations in other windows are also selected = dynamic brushing and linking

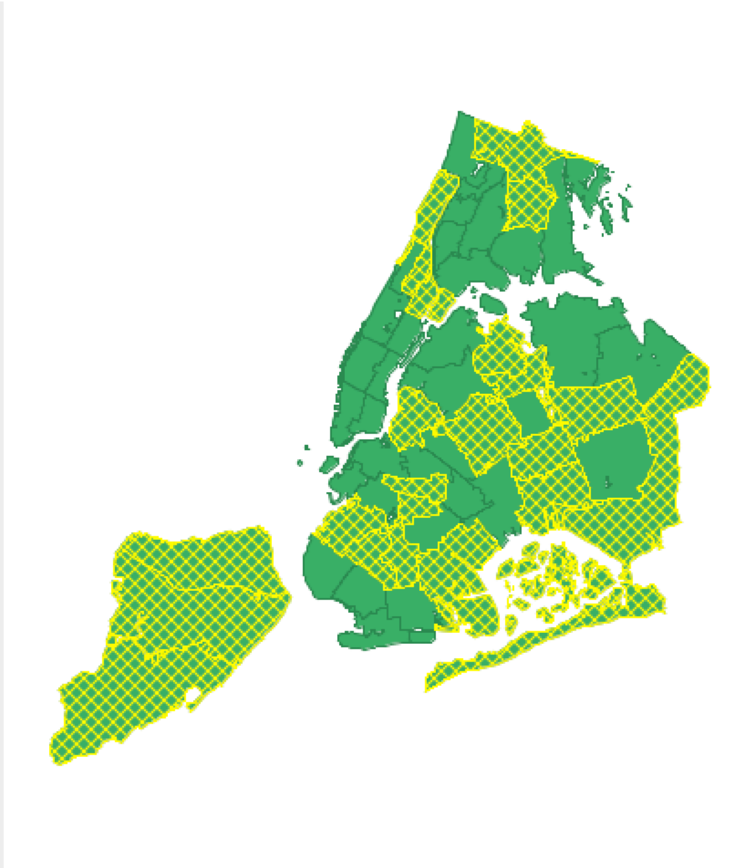


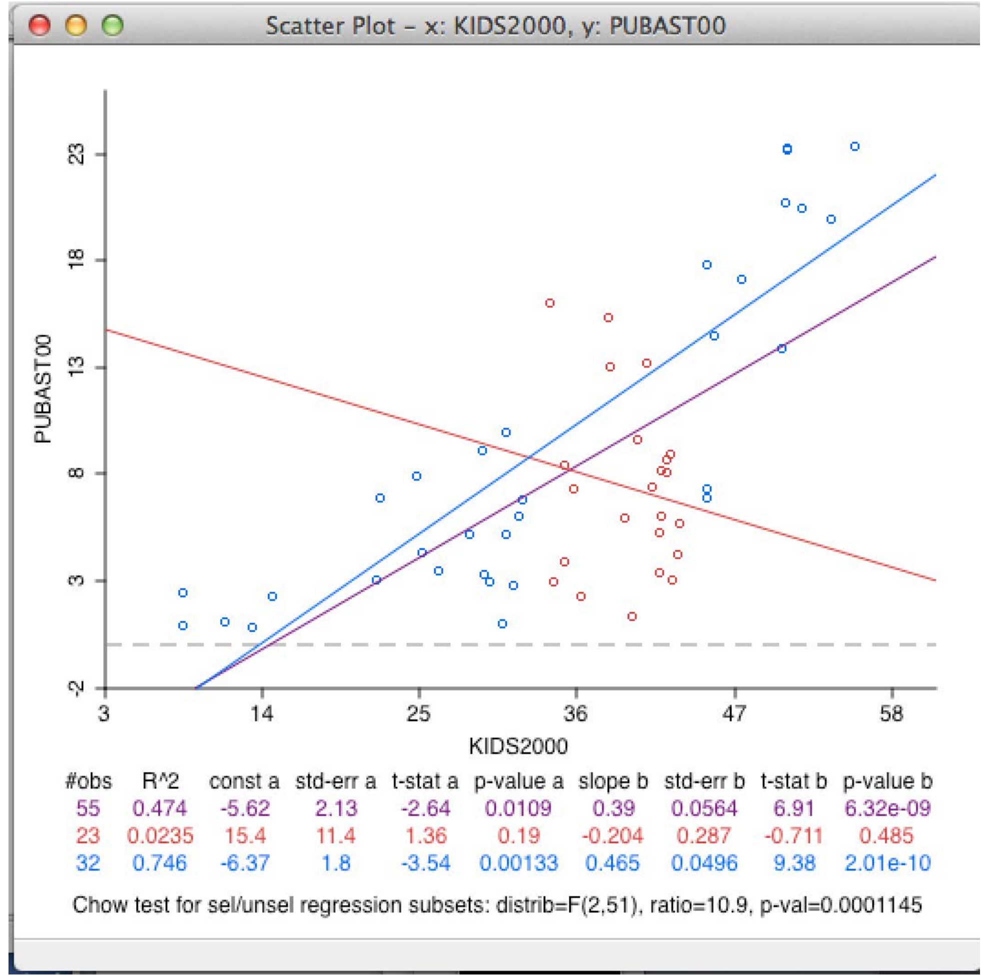
#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
55	0.474	-5.62	2.13	-2.64	0.0109	0.39	0.0564	6.91	6.32e-09
20	6.99e-05	6.92	3.09	2.24	0.0378	0.0029	0.0818	0.0355	0.972
35	0.572	-6.09	2.47	-2.47	0.0189	0.435	0.0654	6.65	1.46e-07

Chow test for sel/unsel regression subsets: distrib=F(2,51), ratio=5.582, p-val=0.006425

#selected=20

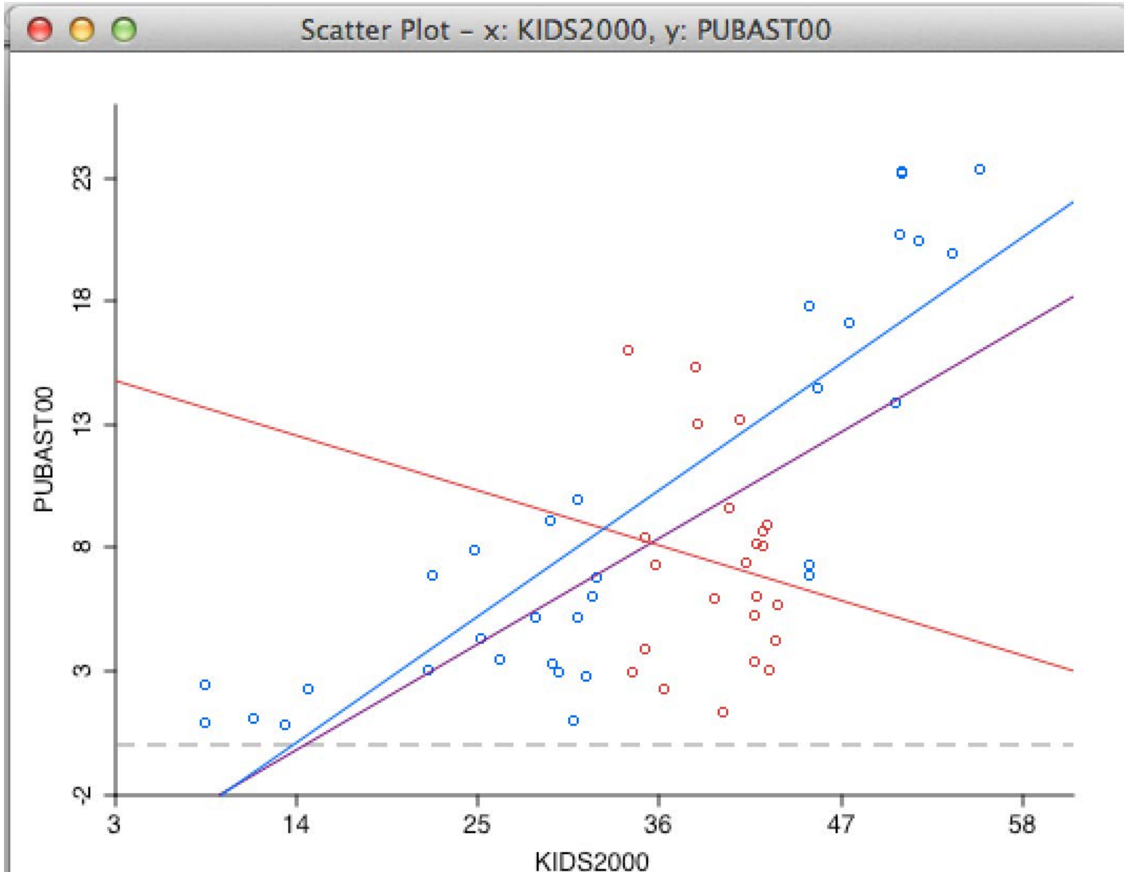
brushing the scatter plot





linked map selection

- Chow Test on Homogeneity of Slopes
    - overall regression slope
    - slope for selected
    - slope for unselected
    - hypothesis test on equality of slopes
        - reject H_0 = evidence of structural instability
    - linking Chow test with map view
        - insight into spatial heterogeneity



#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
55	0.474	-5.62	2.13	-2.64	0.0109	0.39	0.0564	6.91	6.32e-09
23	0.0235	15.4	11.4	1.36	0.19	-0.204	0.287	-0.711	0.485
32	0.746	-6.37	1.8	-3.54	0.00133	0.465	0.0496	9.38	2.01e-10

Chow test for sel/unsel regression subsets: distrib=F(2,51), ratio=10.9, p-val=0.0001145

Chow test

Primer on Multivariate EDA

- Objectives of Multivariate EDA

  represent multi-dimensional data in two dimensions

  dimension reduction

  projection

  discover structure, interaction, patterns

- Methods

  3-D scatter plot

  parallel coordinate plot (PCP)

  scatter plot matrix

  conditional plots

3-D Scatter Plot

- 3-D Scatter Plot

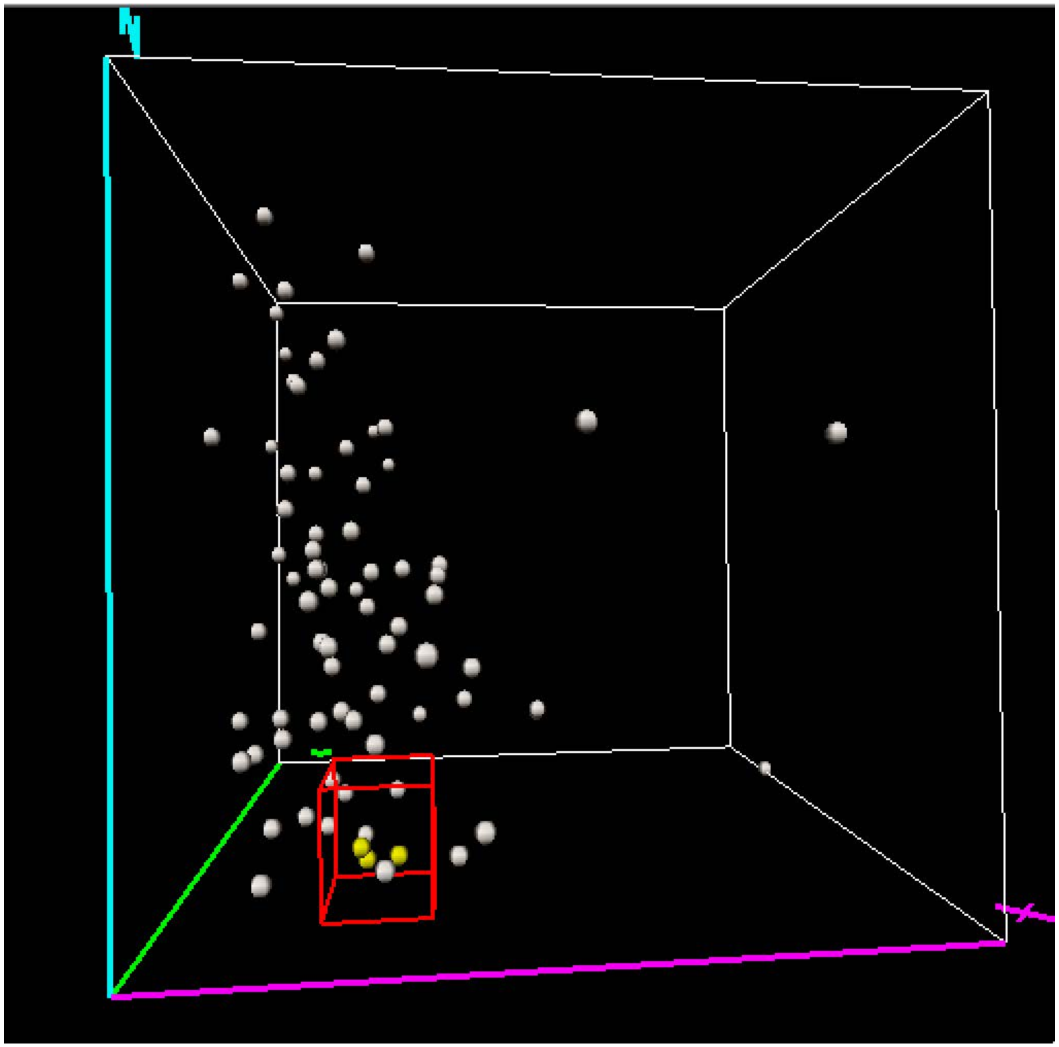
  points in a 3-D data cube

  two-dimensional analysis on side panels

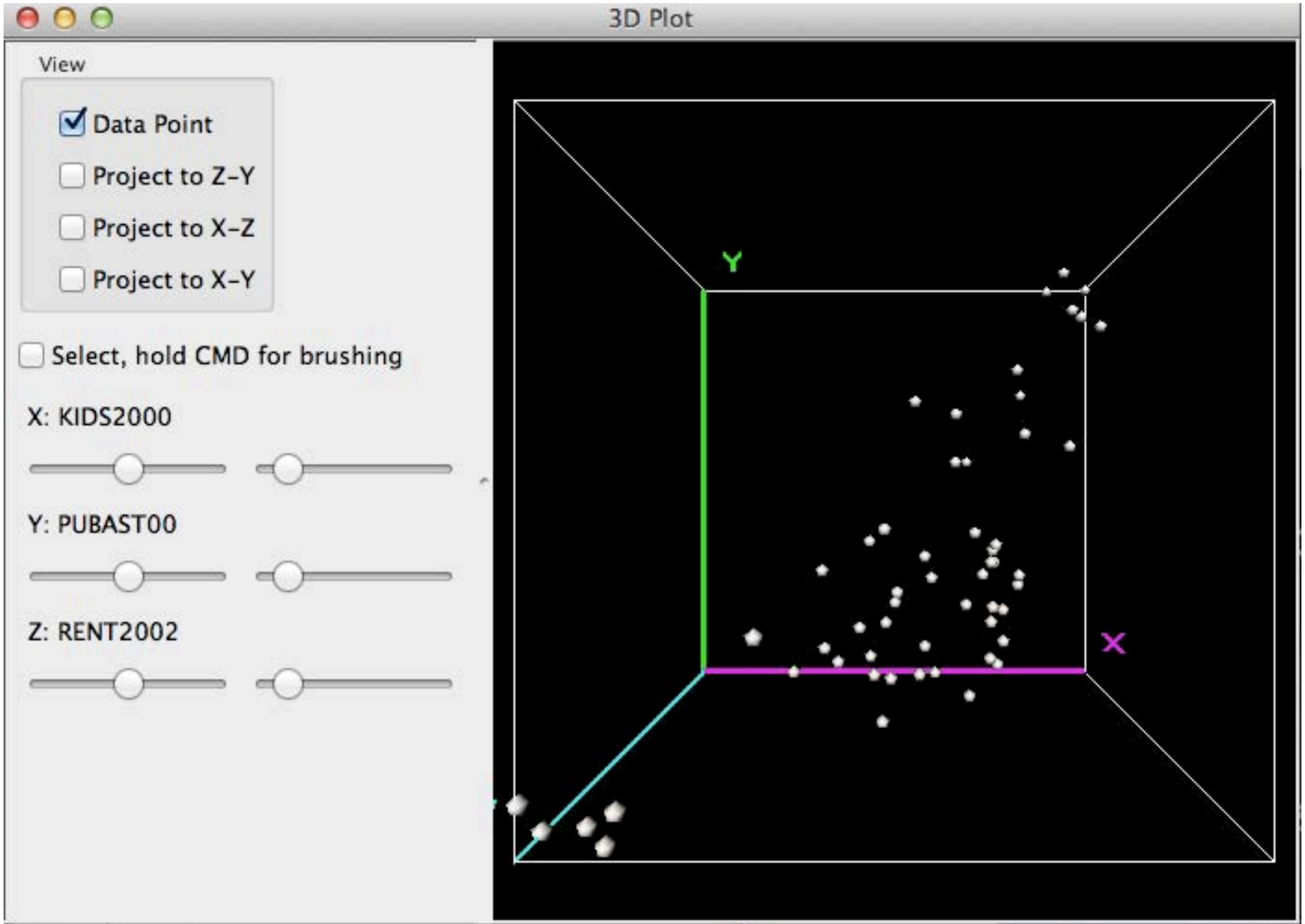
  issues of perspective

    zooming, rotating

  brushing the 3-D data cube



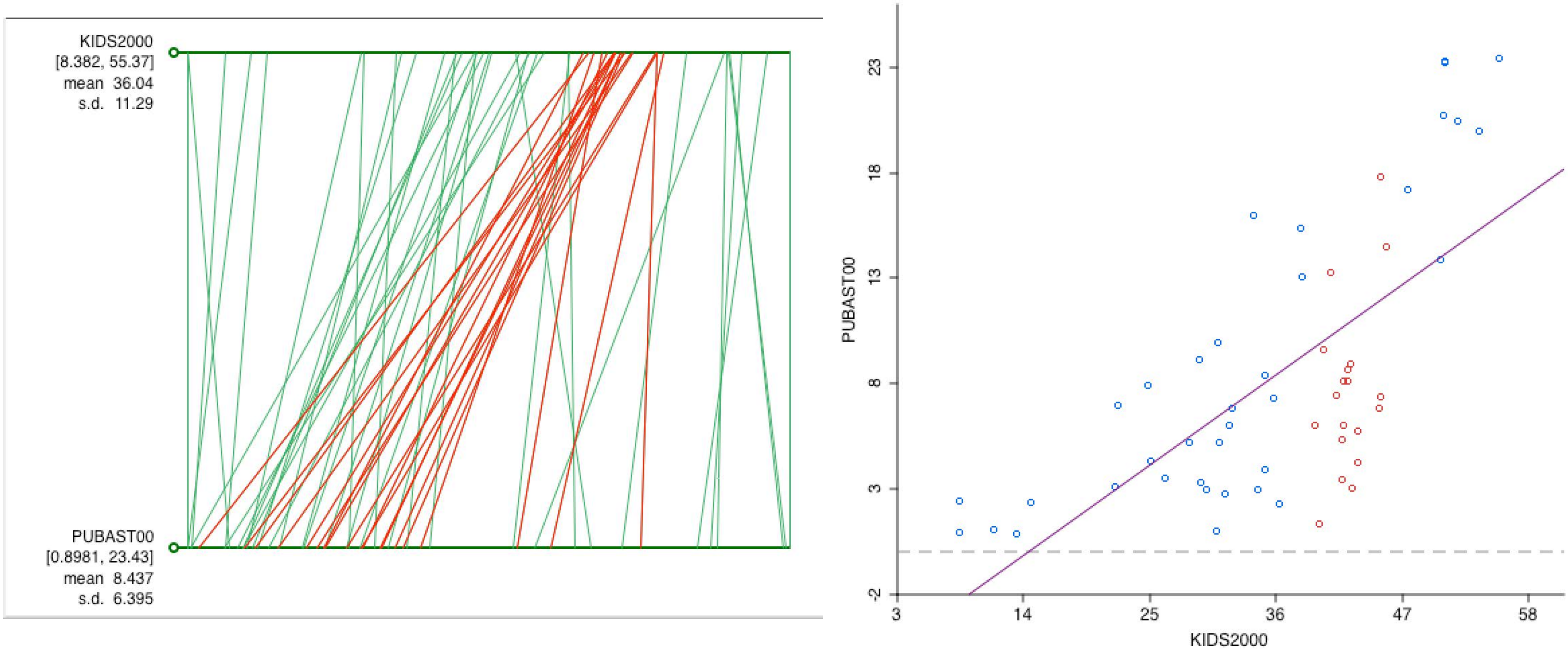
selection in a 3D scatter plot



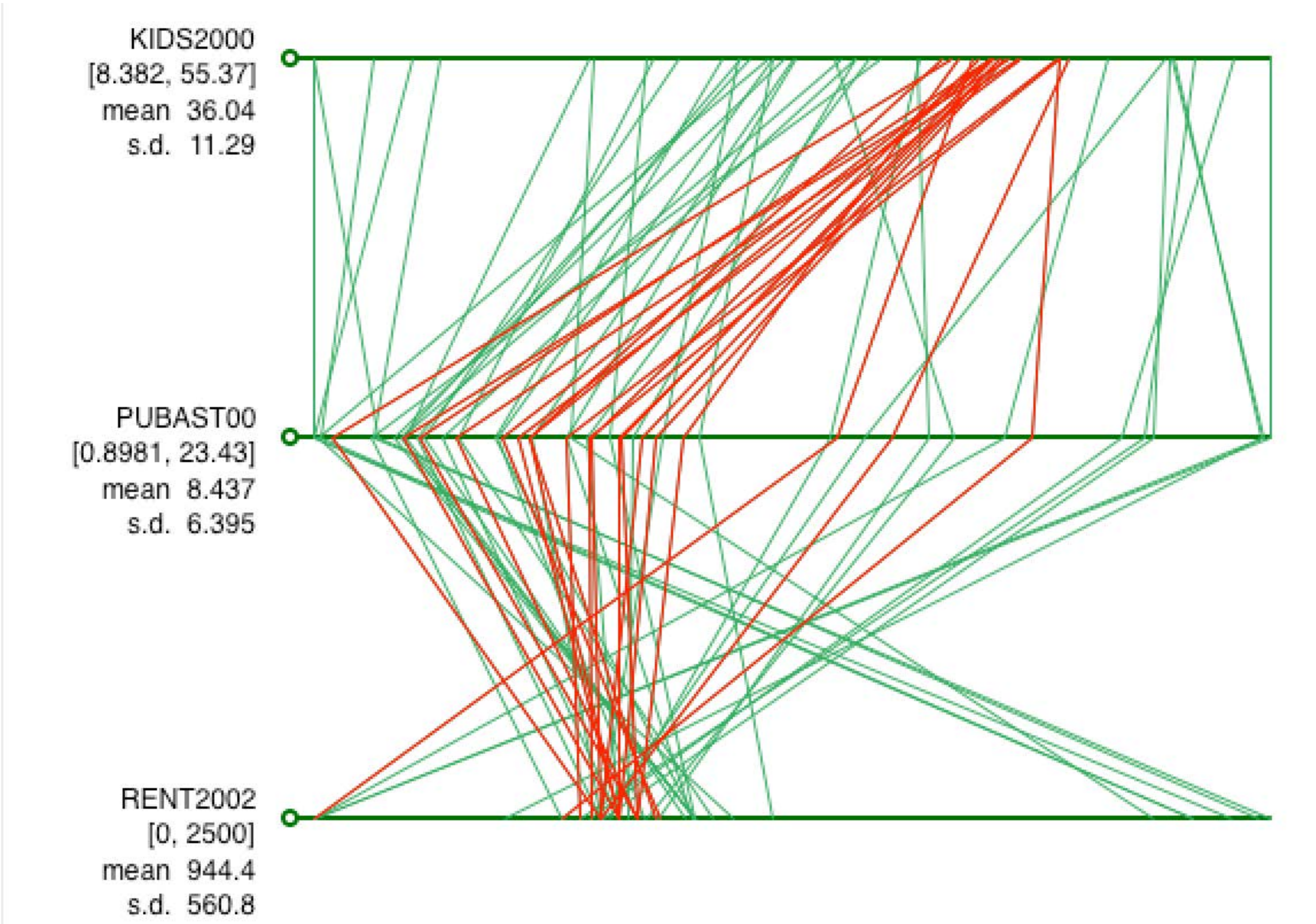
manipulating a 3-D scatter plot

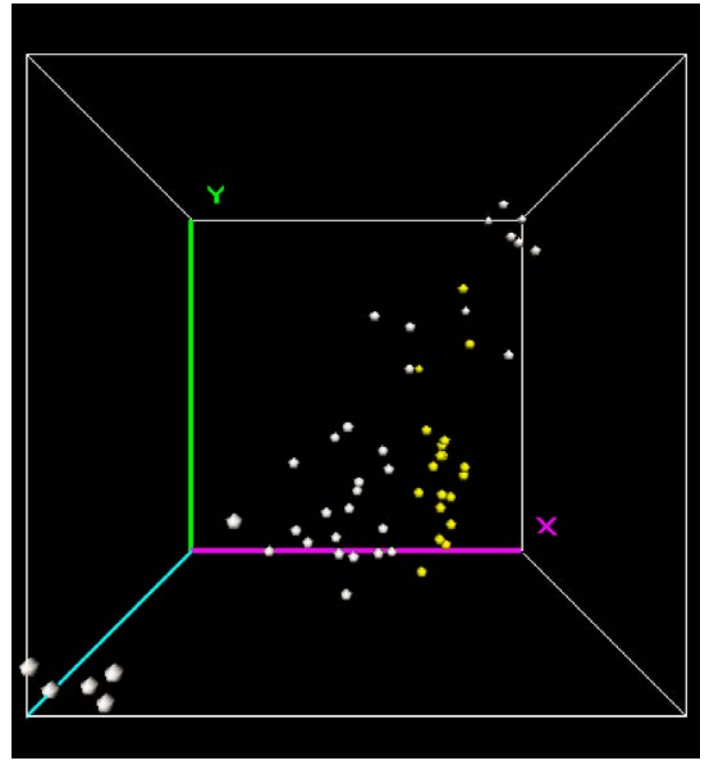
Parallel Coordinate Plot

- 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

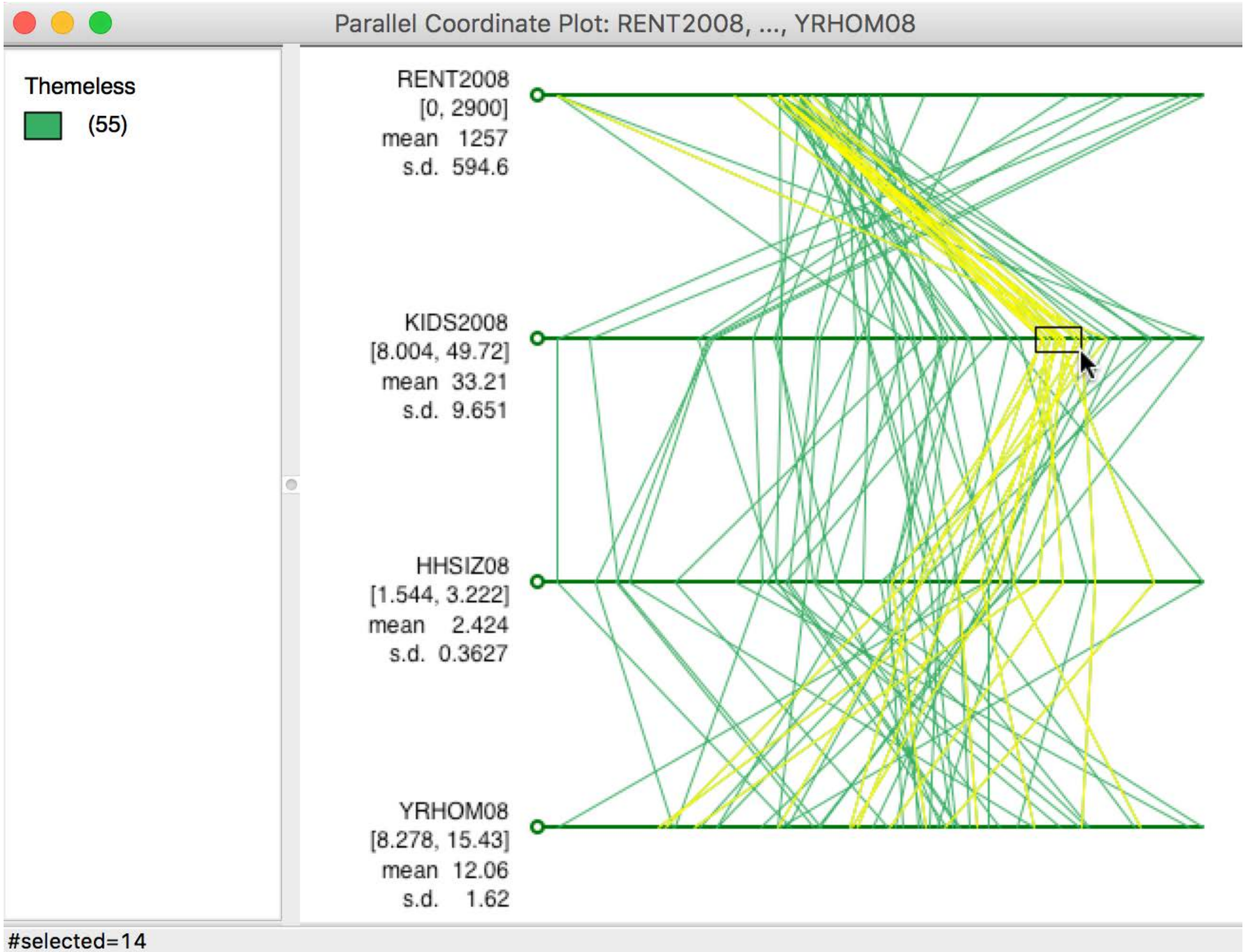


selected lines in pcp
match selected points in scatterplot





selected lines in pcp
match selected points in 3-D scatterplot



brushing the PCP

- Clusters in PCP

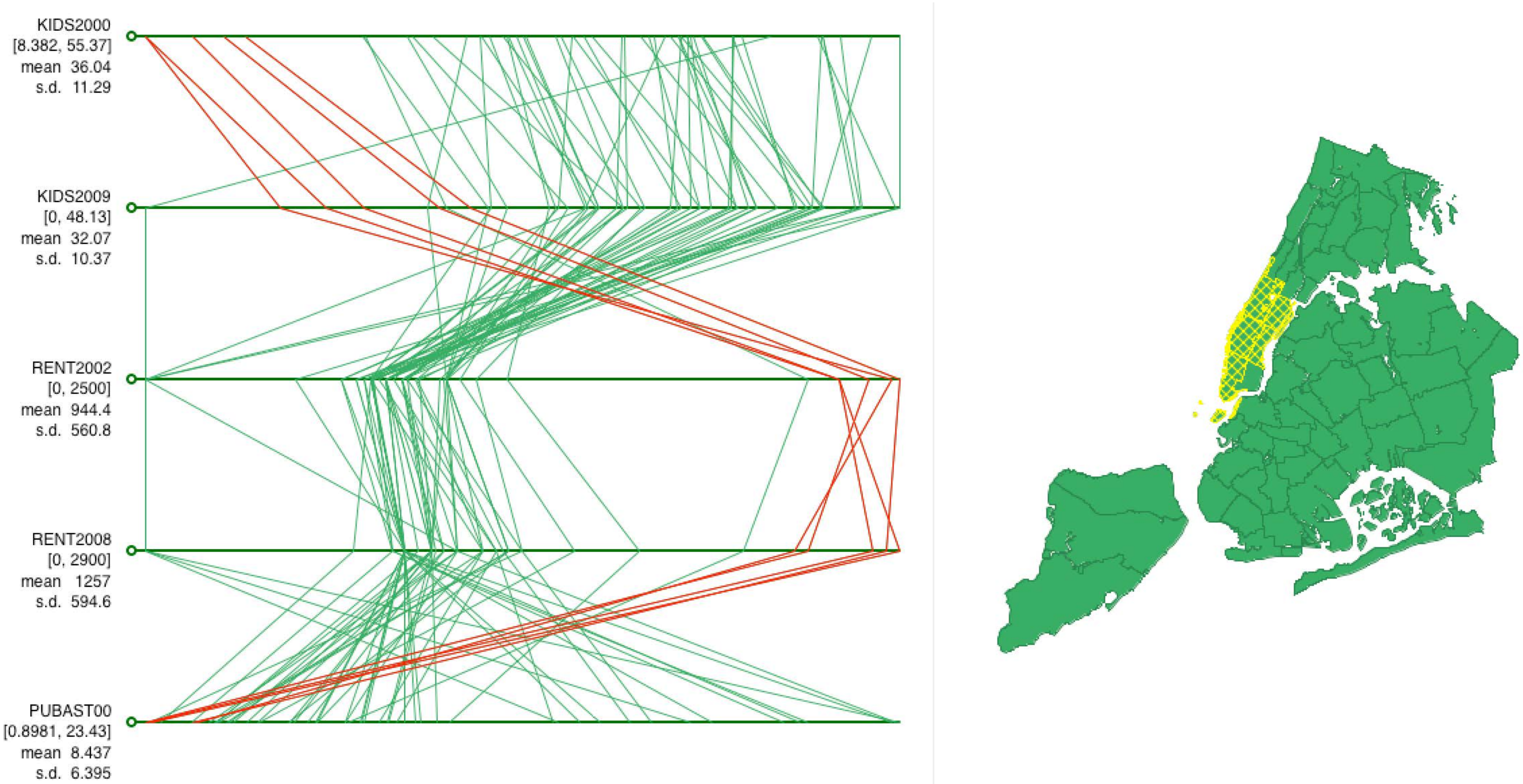
lines that move closely together correspond to points closely together in multidimensional space

= clusters

visual cluster identification

problems with large data sets

remove clutter


KIDS2000
[8.382, 55.37]
mean 36.04
s.d. 11.29
KIDS2009
[0, 48.13]
mean 32.07
s.d. 10.37
RENT2002
[0, 2500]
mean 944.4
s.d. 560.8
RENT2008
[0, 2900]
mean 1257
s.d. 594.6
PUBAST00
[0.8981, 23.43]
mean 8.437
s.d. 6.395

clusters

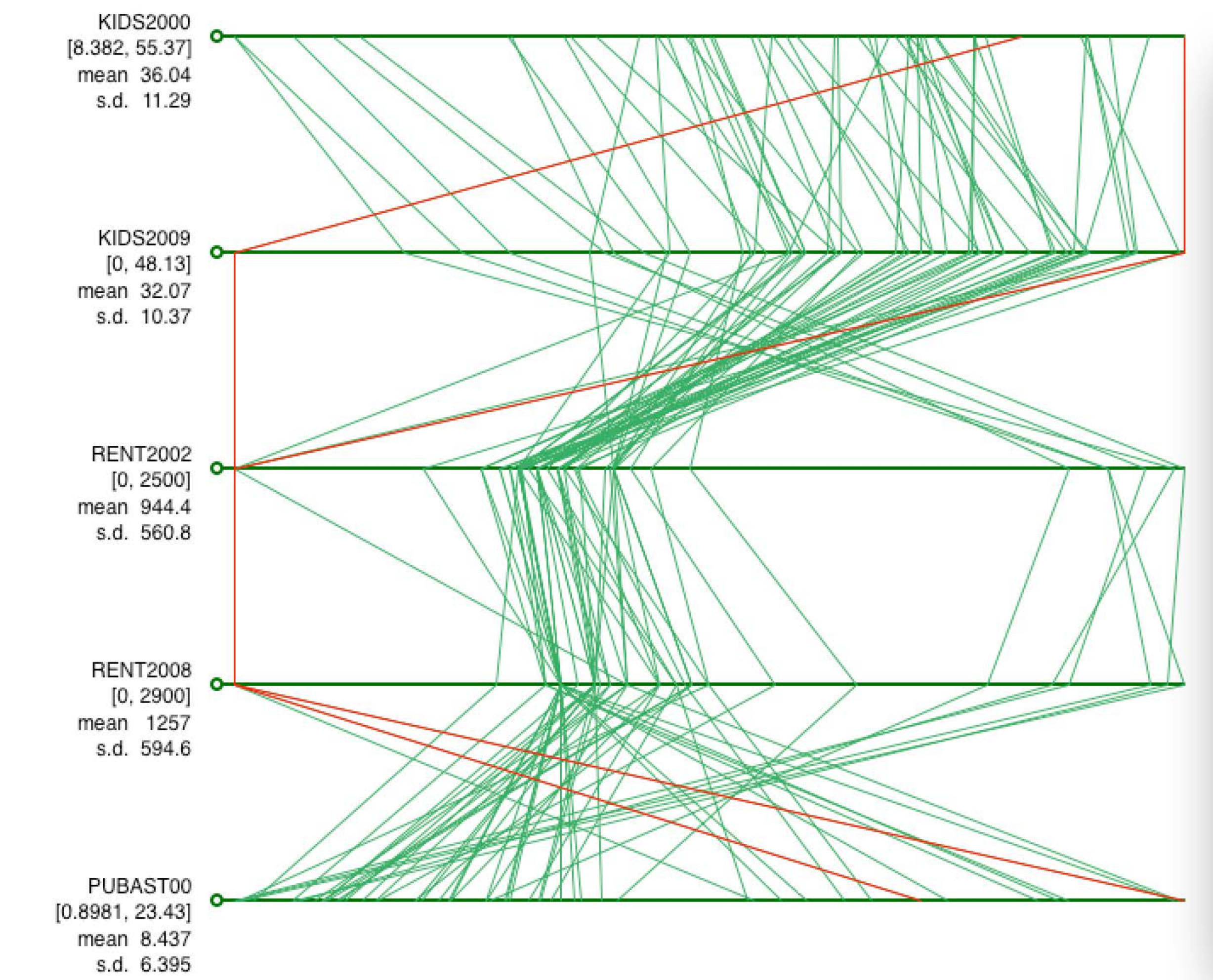
- Outliers in PCP

    lines that are far from the main pack correspond to outlying points in multi-dimensional hyperspace

        point(s) far from the main point cloud

    = outliers

    visual outlier identification


KIDS2000
[8.382, 55.37]
mean 36.04
s.d. 11.29
KIDS2009
[0, 48.13]
mean 32.07
s.d. 10.37
RENT2002
[0, 2500]
mean 944.4
s.d. 560.8
RENT2008
[0, 2900]
mean 1257
s.d. 594.6
PUBAST00
[0.8981, 23.43]
mean 8.437
s.d. 6.395

outliers

Scatter Plot Matrix

- Scatter Plot Matrix

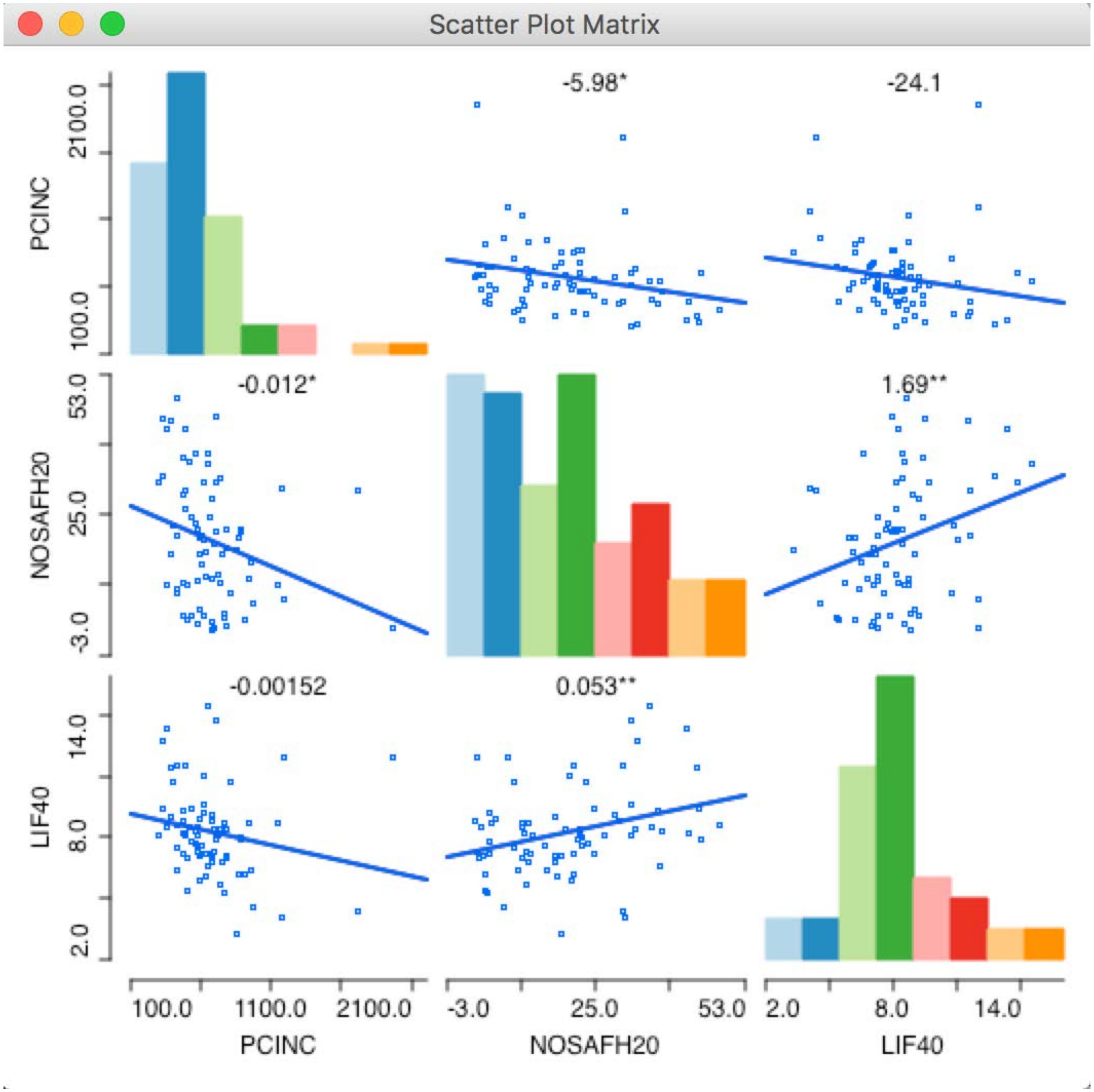
    - matrix of bivariate scatter plots

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

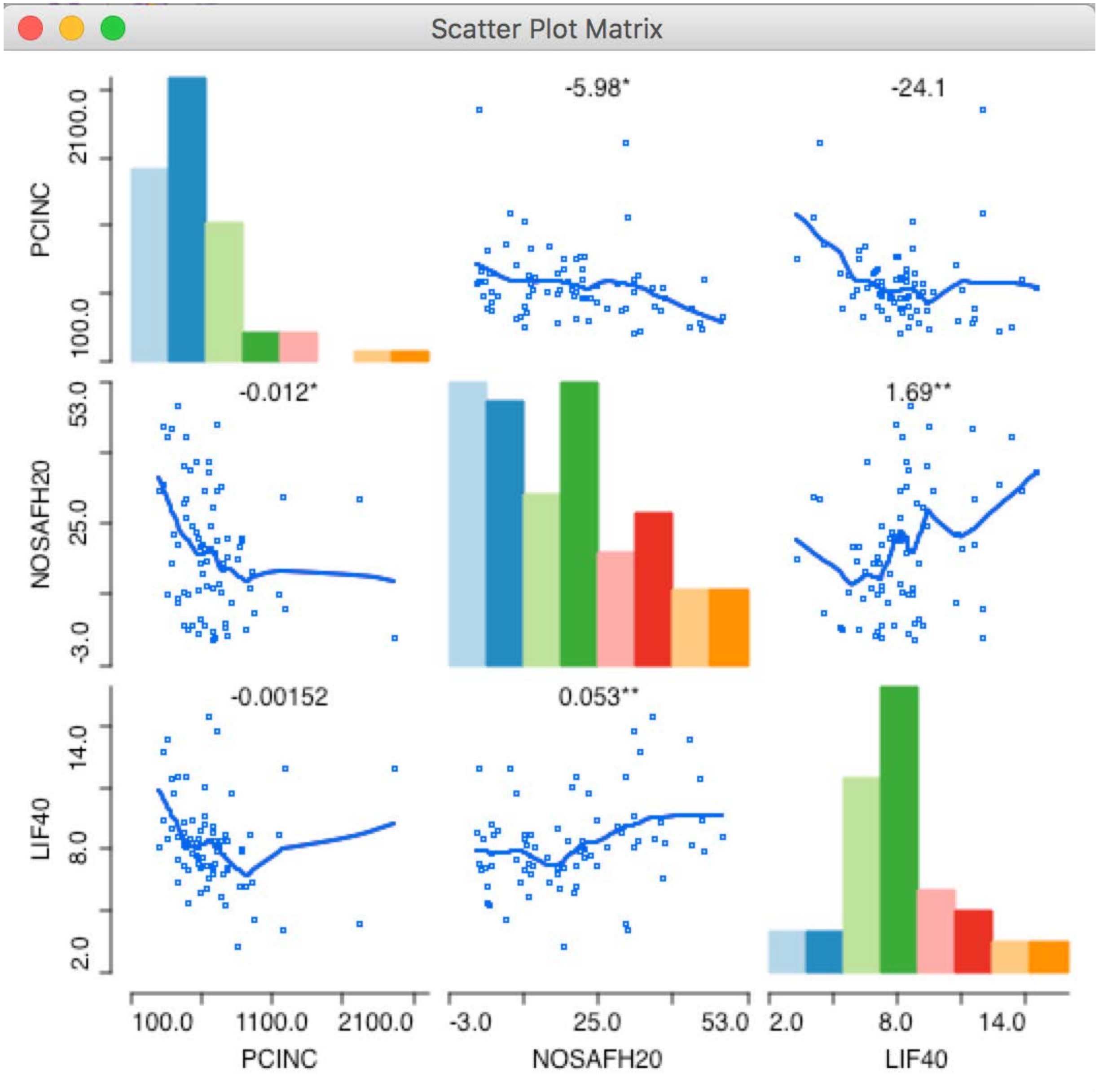
        - not true multivariate, but pairwise bivariate

    - univariate description on diagonal

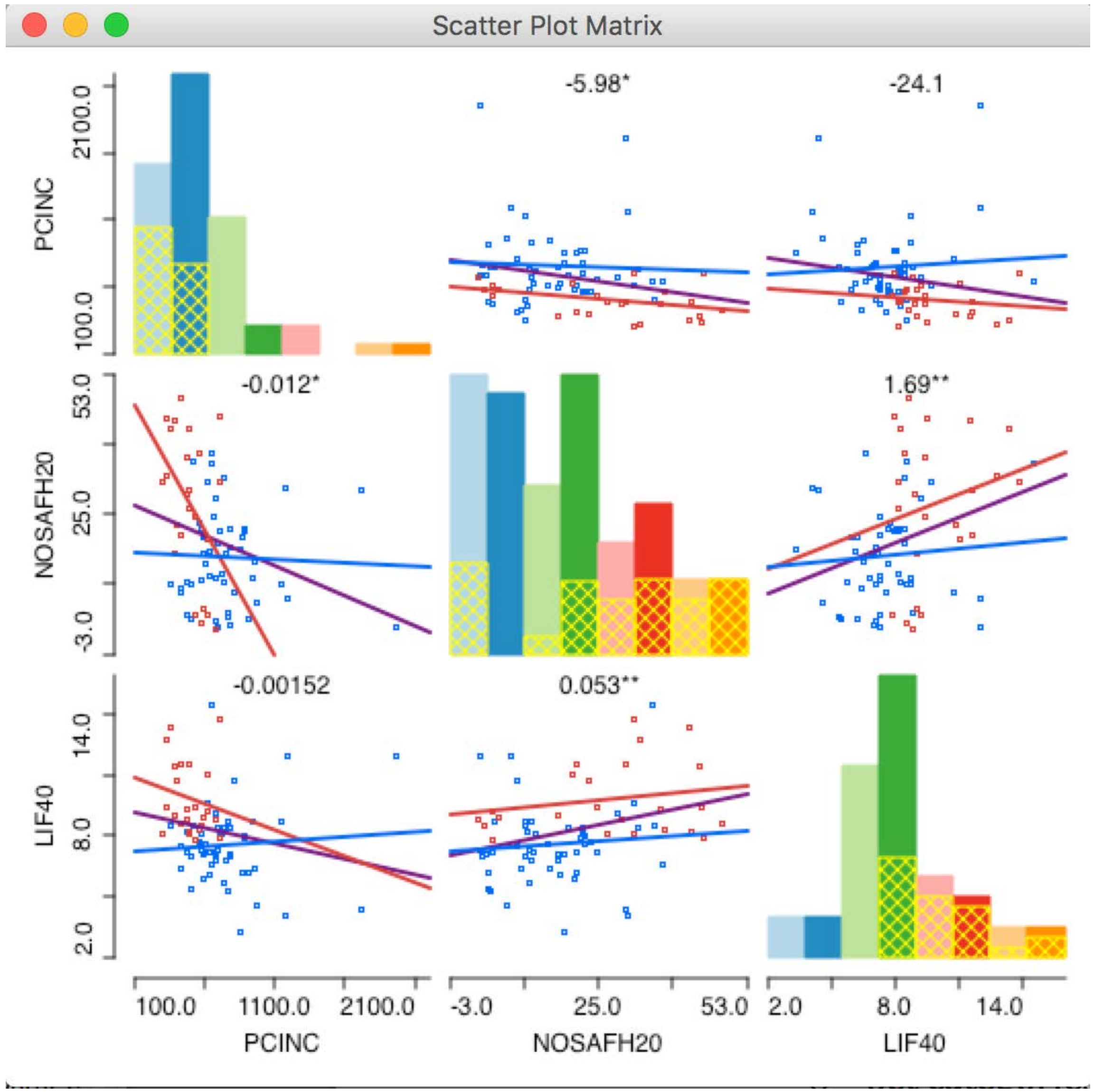
    - focus on interaction effects



scatter plot matrix (Nepal districts)



scatter plot matrix with lowess smoother



brushing the scatter plot matrix

Conditional Plots

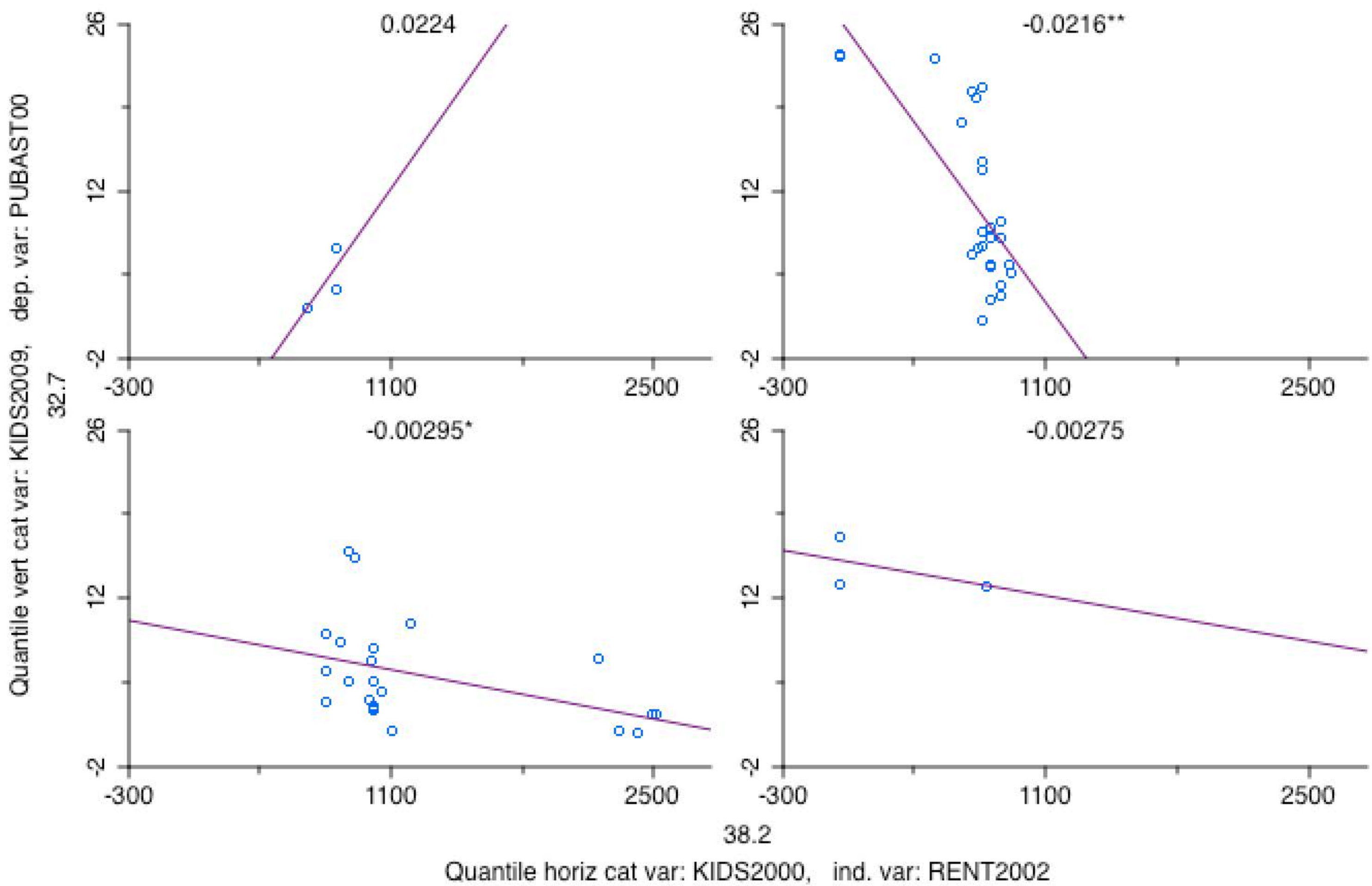
- Conditional Plots

  trellis display

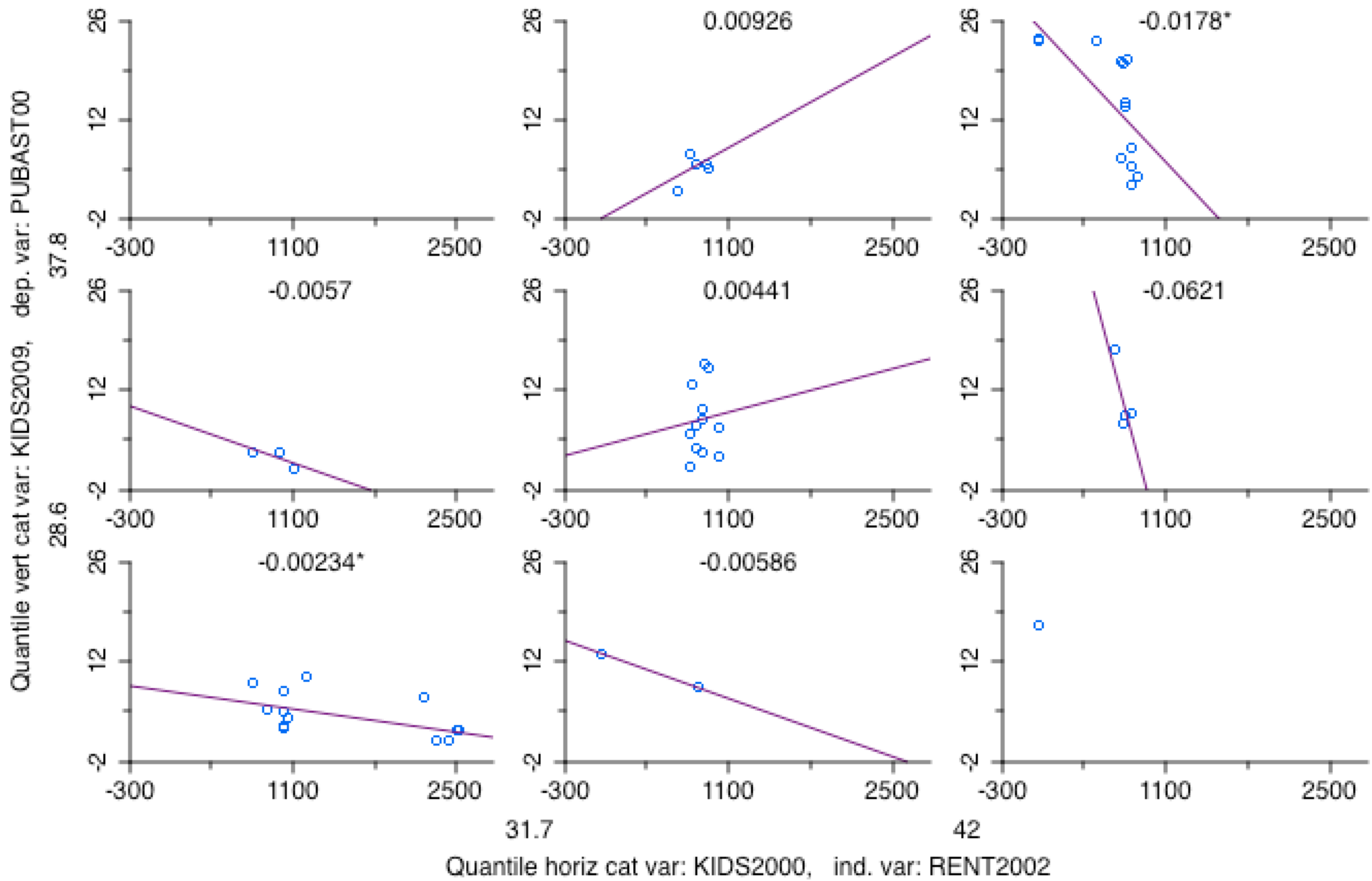
  conditioning variables on the axes

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

    data intervals in two dimensions



conditional scatter plot
cut-point are median



conditional scatter plot
cut-point are third quantile

- 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

Interpretation and Limitations

- No Formal Hypothesis Tests

    exploratory methods do not explain

        suggest hypotheses

        suggest potentially interesting patterns

    no quantification of uncertainty

        no p-values

- Cluster and Outliers

  potentially spurious

  visual inspection — no quantification

  importance/danger of perception

  difficult to extend to multiple dimensions