Stanford and Purdue Fodava Research

Fodava Core Group

Purdue: Ashrith Barthur, Bill Cleveland, Saptarshi Guha, Jeff Li, Bowei Xi, Jin Xia

Stanford: John Gerth, Pat Hanrahan, Justin Talbot

Research Partners

David Anderson, Xavier University

Carter Bullard, Qosient, LLC

Paul Kidwell, Lawrence Livermore Labs

Ryan Hafen and William Pike, Pacific Northwest National Laboratories

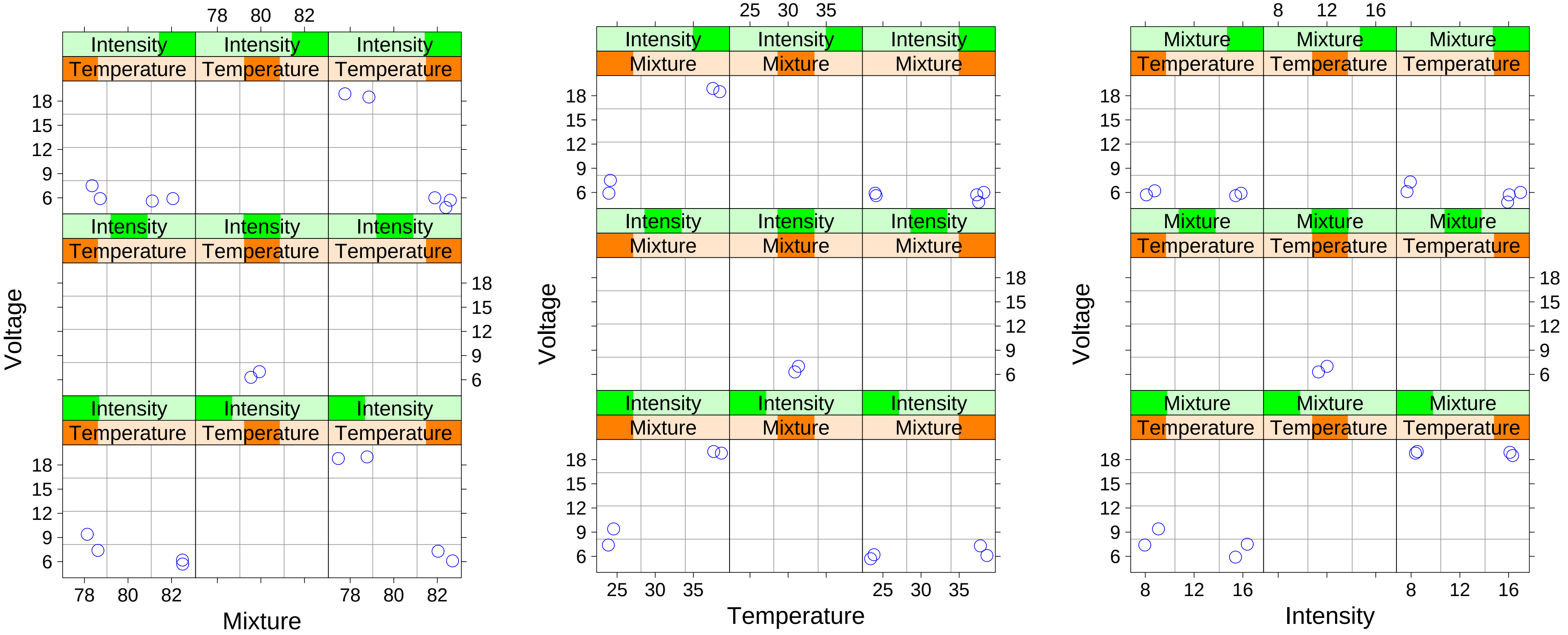
Montse Fuentes, North Carolina State University

John Chambers, Stanford

Tyson Condie and Joe Hellerstein, UC Berkeley

Trellis Display: Model Building for Data from Designed Experiments

Voltage against an explanatory variable given two others



Trellis developed in the context of moderate and large datasets

Industrial experiments with a small number of runs are very common

Discovered that trellis display is very useful

The experiments, to succeed, must be designed to keep the noise small, and this enables Trellis to succeed in the model building phase

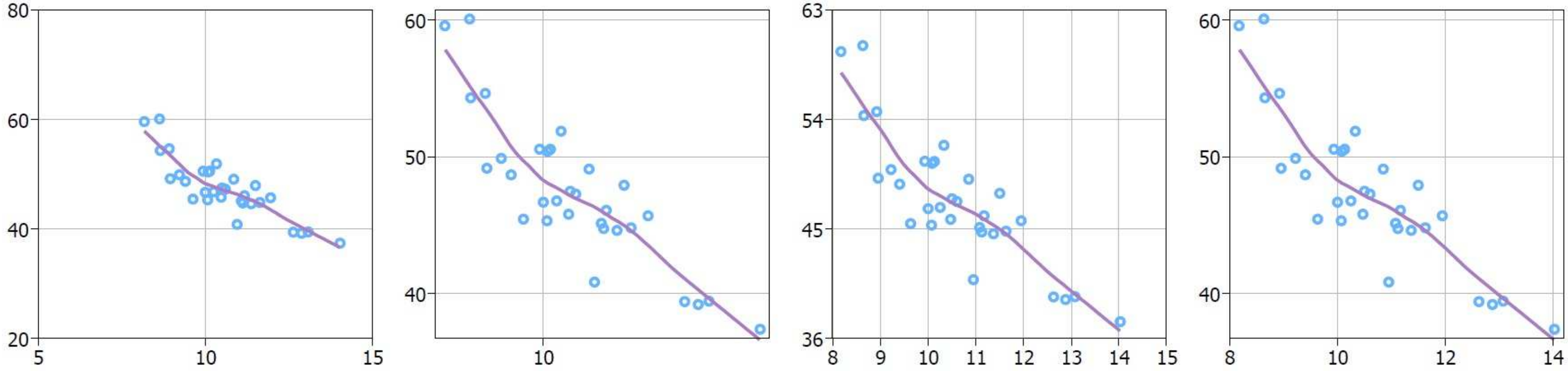
An Extension of Wilkinson’s Algorithm for Positioning Tick Labels on Axes

New, automated system for choosing positions and labels for axis tick marks.

Extends Wilkinson’s optimization-based labeling approach

Creates a more robust, full-featured axis labeler

Example of how ideas from automated graphic design can be applied to information visualization.



Adapting Daniel and Wood’s Modeling Approach to Interactive Visual Analytics

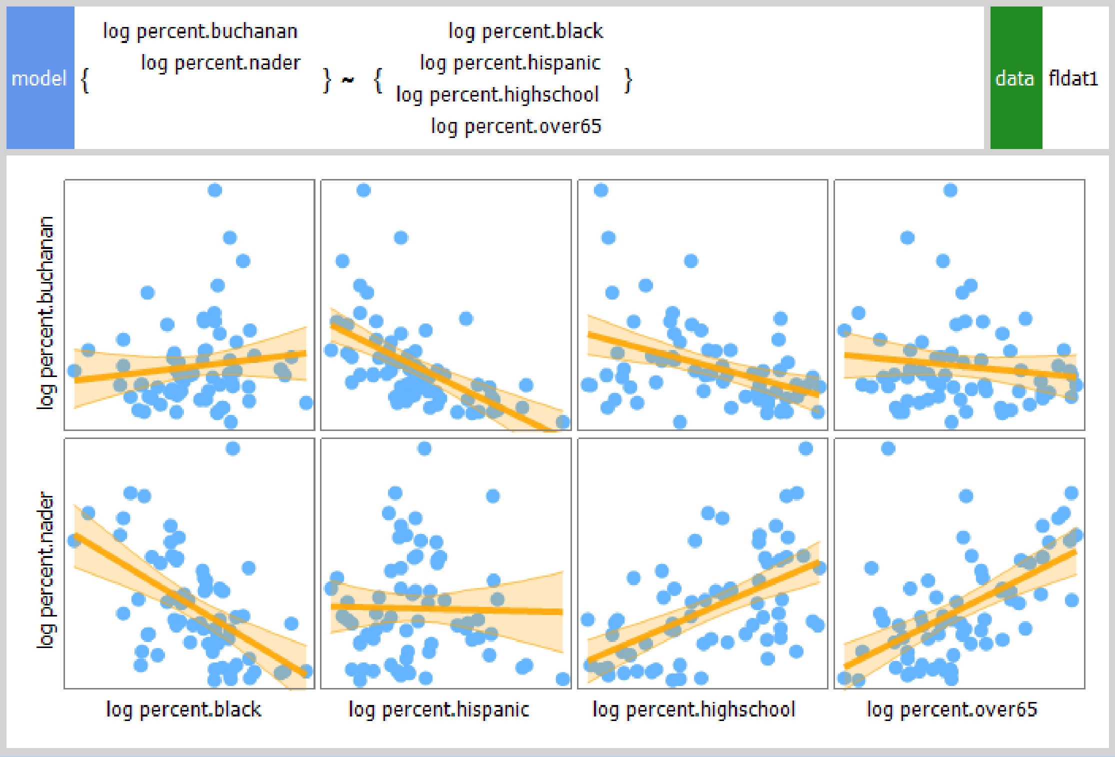
See the poster

Modeling often most challenging part of analysis of a data set

- application of model to achieve the goal can be easy by comparison

A visual interface for building structured visual comparisons of linear models

Leverages Wilkinson’s model formula notation to permit construction of multiple linear models for comparison



model { log percent.buchanan
log percent.nader } ~ { log percent.black
log percent.hispanic
log percent.highschool
log percent.over65 }

data fldat1

The ed Method for Nonparametric Density Estimation and Diagnostic Checking

Two parts to ed

1. Method of *estimation*

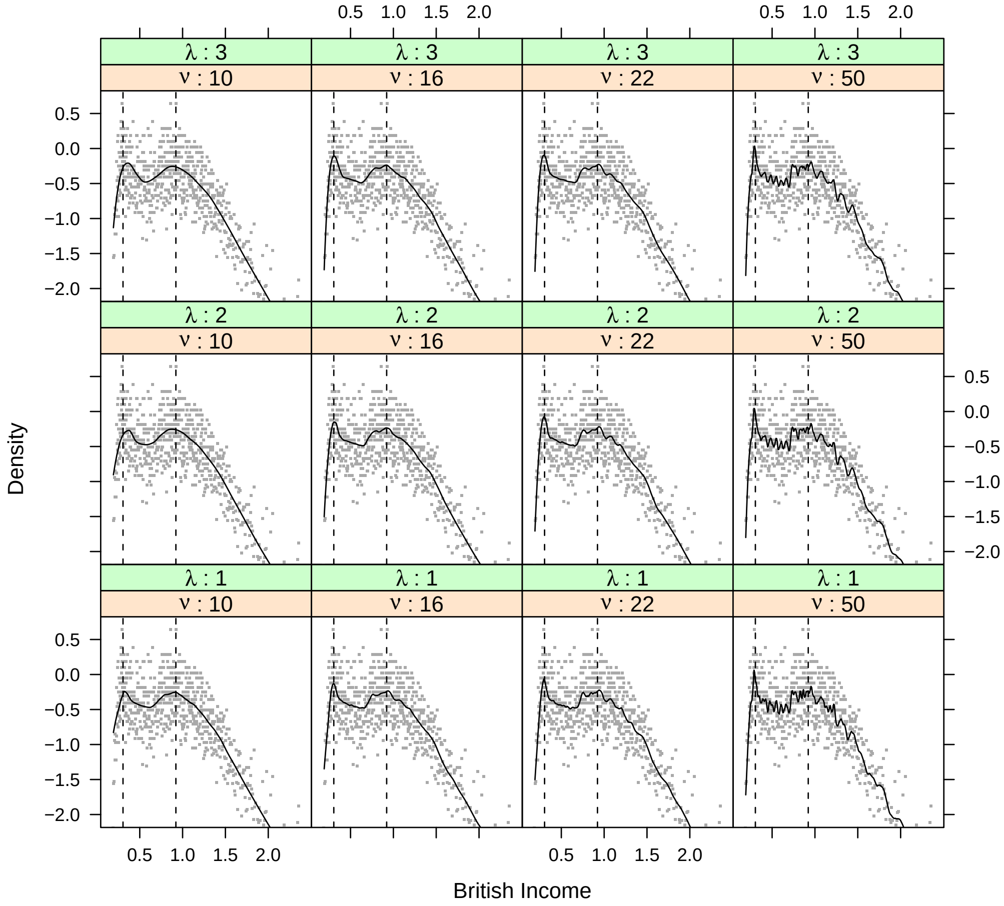
2. Design of the estimation method enables for *diagnostic checking* of the estimates

Approach turns nonparametric density estimation into nonparametric regression

Many visualization tools from regression diagnostics

ed estimates: can accurately fit a wide range of density patterns

First comprehensive diagnostic checking to determine if a density estimate follows density patterns of the data



Divide & Recombine (D&R) for The Analysis of Large Complex Datasets

See the poster

Approach to analyzing large, complex datasets that parallelizes the data

- divide the data into subsets
- analyze subsets using visualization methods and mathematical methods
- recombine using visualization methods and mathematical methods

Visualization

- apply a visualization method across a large representative sample of subsets (100s or even 1000s)
- rapid scan displays viewed by a controlled animation that allows effective study

RHIPE: R and Hadoop Integrated Programming Environment

- Saptarshi Guha
- R and Hadoop merger
- enables analysis, wholly from within R, of very large datasets using D&R

Multifractal Fractional Sum-Difference (MFSD) Models for Internet Traffic

Aggregate packet-level best-effort Internet traffic

Finding a valid model began in mid 90s, but no success until now

Valid Model: can generate traffic of any packet rate with the same statistical properties as live traffic

MFSD model class succeeds in this

Also very simple and mathematically tractable

Modeling based on analyzing large, complex packet-level datasets

D&R with extensive visualization critical to success

A Gaussian fractional sum-difference (GFSD) model for a time series z_u $z_u = \sqrt{(1-\theta)}s_u + \sqrt{\theta}n_u$

$0 \leq \theta \leq 1$

s_u and n_u are independent processes

n_u is Gaussian white noise with mean zero and variance 1

$(I - B)^d s_u = \epsilon_u + \epsilon_{u-1}$

B is the backward shift operator

$0 < d < 0.5$

ϵ_u is Gaussian white noise with mean 0 and variance $\{(1-d)\Gamma^2(1-d)\}/\{2\Gamma(1-2d)\}$

A multifractal fractional sum-difference model (MFSD) for a time series t_u

t_u is a nonlinear strictly monotone transformation of a GFSD, z_u.

A Streaming Statistical Algorithm for Detection of SSH Keystroke Packets in TCP Connections

See the poster

Part of our work in A Science of Cyber Security: Monitoring and Forensics

First algorithm that detects SSH client keystroke packets in a TCP connection on any port

Uses only packet headers and arrival timestamps, no packet content

Algorithm succeeds because a keystroke creates an identifiable dynamical pattern in the packets

Use for cyber security: detect backdoor logins

Algorithm demonstrates the potential for use of detailed packet dynamics to classify connections, important for cyber security

D&R with extensive visualization critical to success

Visual Analytics for Hadoop Performance

See the poster

Hadoop to handle large datasets increasingly popular (e.g., RHIPE)

Hadoop extension developed at UC Berkeley (HOP): streams back partial results

Developed a web-based monitoring tool for HOP

Stream back meta-information about the job

- processor utilization
- memory consumption
- task scheduling information
- error conditions

Users can see overall completion information, drill down into specific tasks on specific machines, see machine usage statistics

The web interface helped experts find and debug a couple of major problems in the HOP task scheduler

Future work

- combine HOP and RHIPE to allow data analysts to study partial results
- interesting issues serial display of streaming information