Parallel Coordinates

Shengying Li
CSE591 Visual Analytics
Professor Klaus Mueller
March 20, 2007

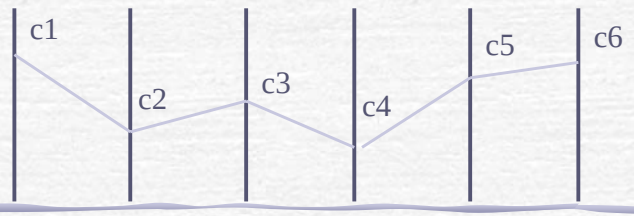
Background

- Proposed in 80's by Alfred Insellberg
- Good for multi-dimensional data exploration
- Widely used in information visualization of multi-dimensional data

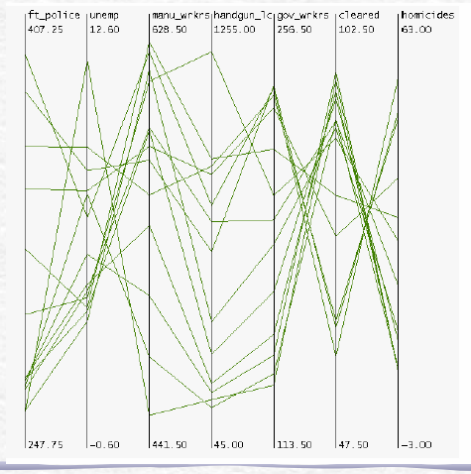
Basics

- To represent N dimensional data
  - Set N vertical axes in parallel
  - Put data to intersects on corresponding axes
  - Connect intersects

Example: (c1, c2, c3, c4, c5, c6)

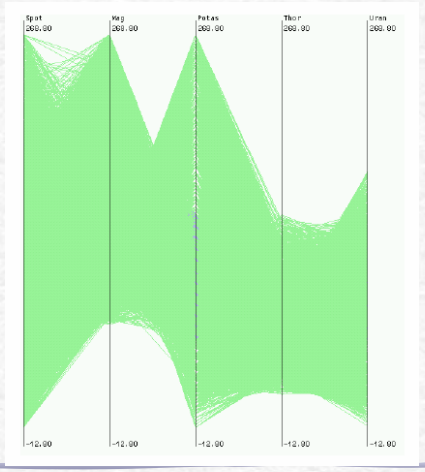


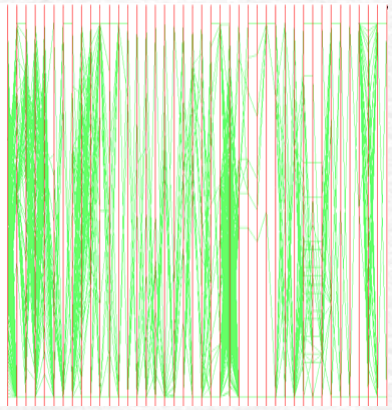
Good Example



- Data structure
- Data trend
- Correlations

Bad Examples

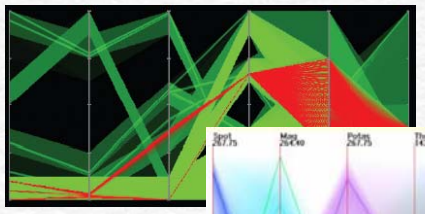




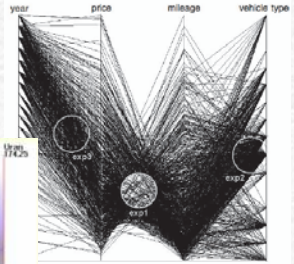
Common Solutions

- Sampling
- Filtering
- Aggregation and Summarization
- Dimensionality Reduction (PCA, MDS)
- Binning
- Multiresolution Methods

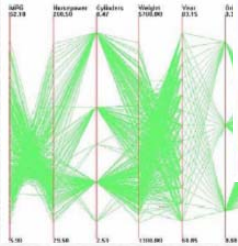
Solutions in This Talk

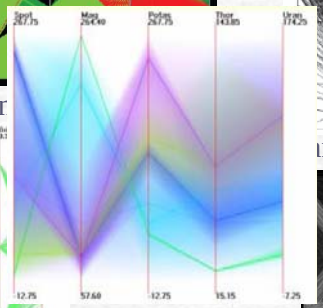


Outlier-preservir



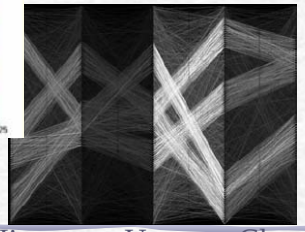
ndom Sample Lens





Hierarchical Viewing

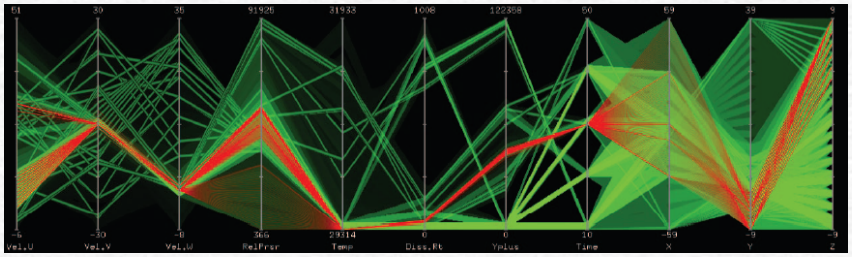
Dimension Reordering



Histogram Uncover Cluster

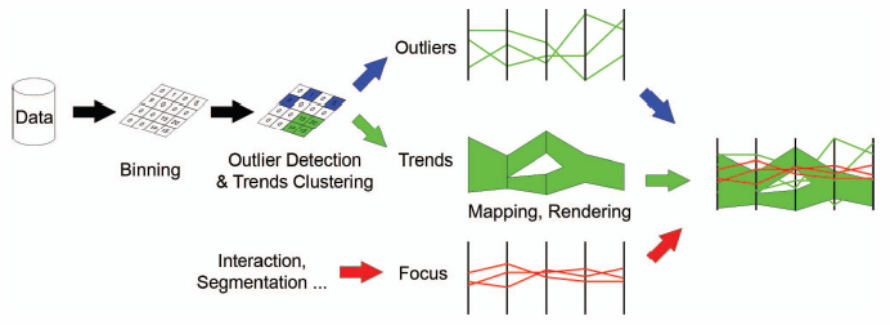
Outlier-preserving Focus+Context

- InfoVis 2006
  - “Outlier-preserving Focus+Context Visualization in Parallel Coordinates”, by Matej Noveotny, Helwig Hauser



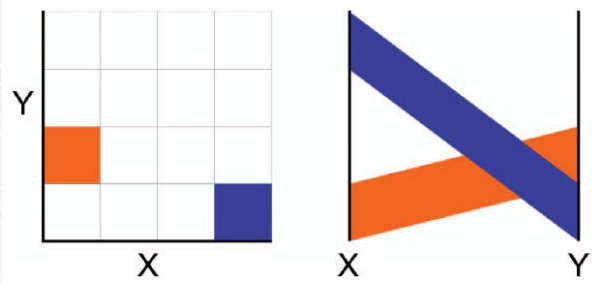
General Idea

- To get focus in very large dataset
  - Limit thousands of data per view
- To show context
  - Preserve trend together with outlier



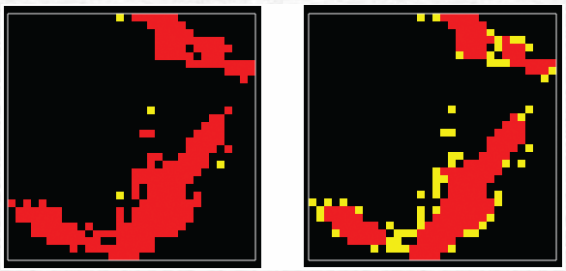
2D Bin Map

- One 2D Bin map for every neighboring pair of axes
- Every polyline across axes is a bin
- K-axes combines (k-1) 2D bin map



Outlier Detection in Bin Map

- Outlier definition
  - Data points which don't belong to any cluster
- Outlier detection
  - First apply a low-pass filter
  - Next compare difference to the original bin



Isolate filter

Median filter

Yellow bin
is outlier

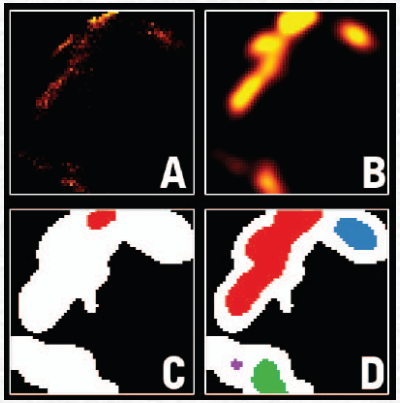
Outlier Effect

Original

Traditional

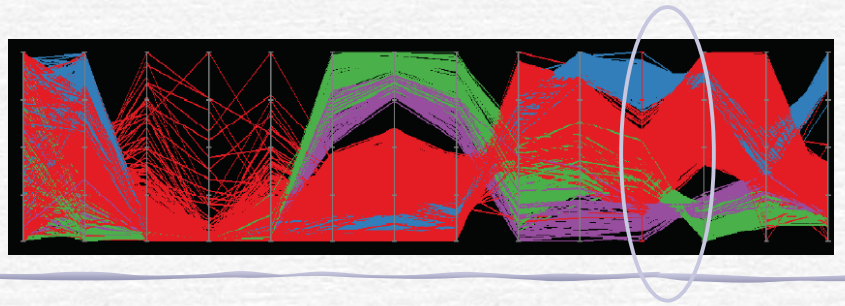
With outlier

Clustering

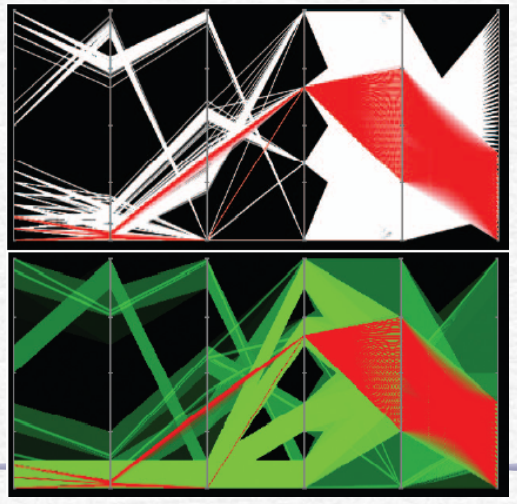


A. 2D bin map
B. Use a Gaussian filter
C. D. :
Start from the bin with highest occupancy, grow the region until reaches threshold.
C: 50%; D: 10%

Clustering Effect



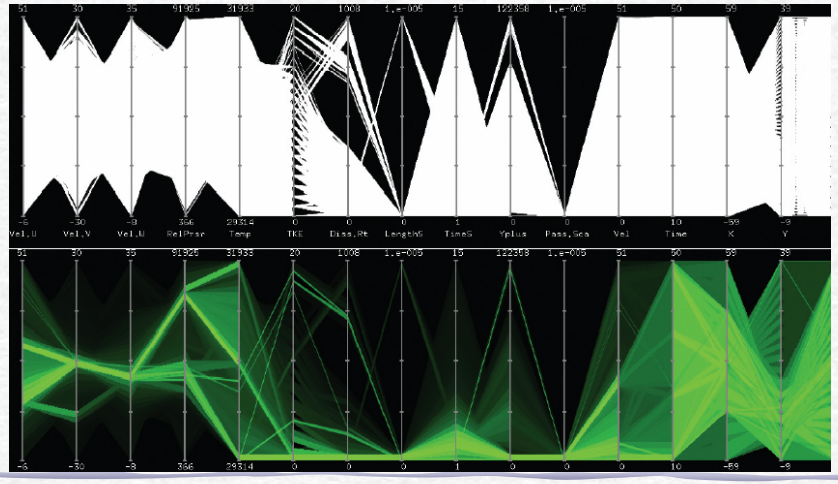
Focus



Traditional focus+context

Outlier-preserving focus+context

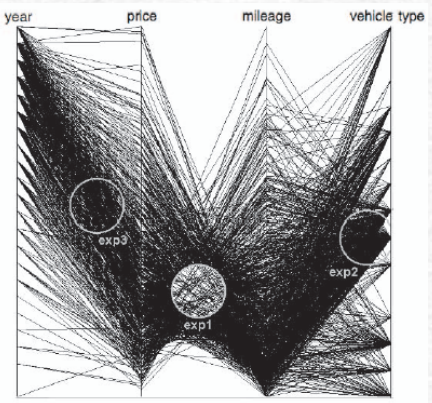
More Results



Performance

- Large dataset: 3M in 16 dimension
- Interactive frame rate, 0.1 or 0.2 sec
- Bin map is time cost, 70 seconds

Automatic Clutter Reduction



InforVis 2006

"Enabling Automatic Clutter Reduction in Parallel Coordinate Plots", by Geoff Ellis, Alan Dix

General Idea

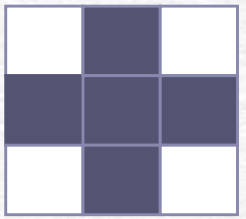
- Random sample lens is helpful
  - Random sample
    - Remove overlapping clutter
    - Preserve trends or pattern in data
    - No requirement of user interactions
  - Sampling lens
    - Get interesting trend while preserve context
    - Has freedom of location, size, shape, sampling rate of the lens.
- How to effectively calculate the "clutter" during interactive movement of lens?

Overplotted%

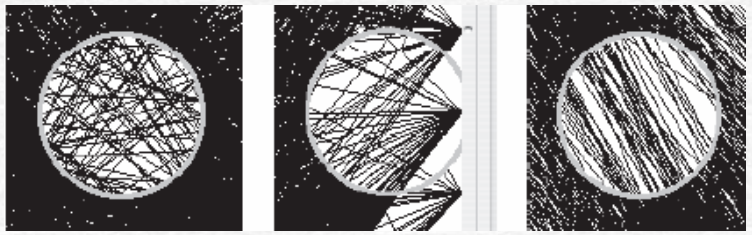
- M1- number of plotted points on their own pixel
- Mn – number of plotted points sharing a pixcel
- S0 – number of empty pixels
- S1 – number of pixels with 1 plotted point
- Sn – number of pixels with more than 1 plotted point

M1 = 4, Mn = 2
S0 = 4, S1 = 4, Sn = 1

Overplotted% = 1/(4+1)
= 20%



Line Patterns



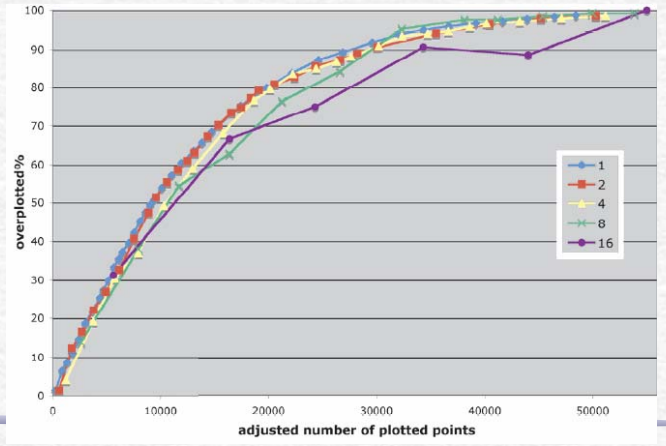
Crossed | Same end point | Small angles

Clutter Calculation Methods

- Raster algorithm
  - Counts number of plotted points on grid cell
- Random algorithm
  - Every plotted points randomly placed in pixels
  - Binomial distribution
- Lines algorithm
  - Estimate intersection volumes of all lines crossing lens

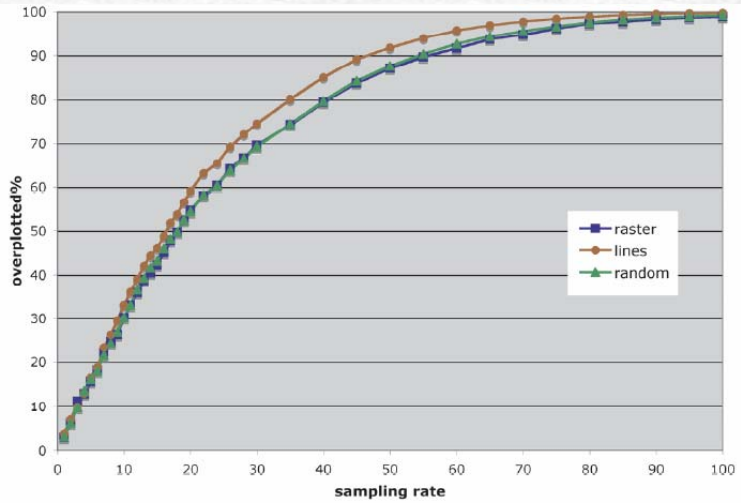
Raster Algorithm

- Adjust grid cell size for speed

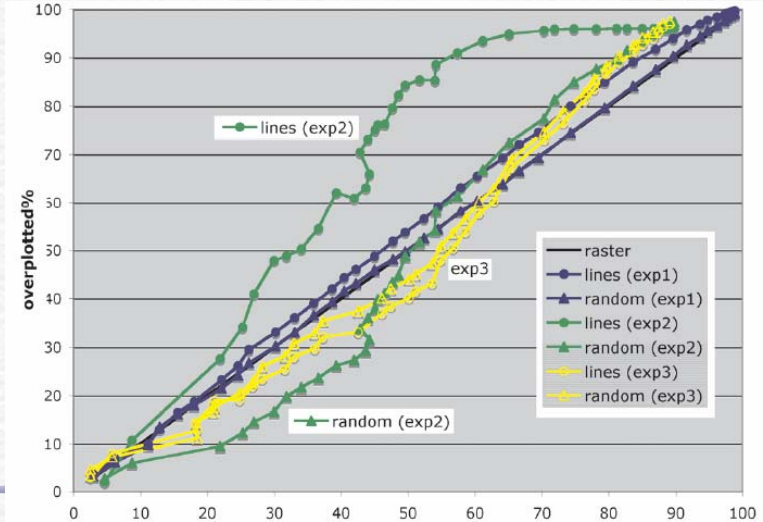


Accuracy Measurement

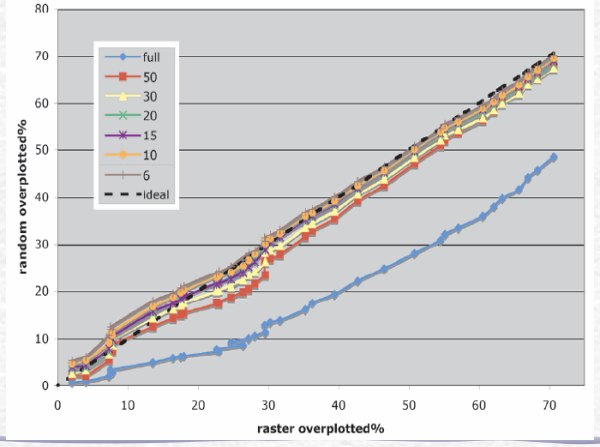
For exp 1 dataset:



- For exp1, 2, 3 normalised against raster



Multiple-bin Random Algorithm

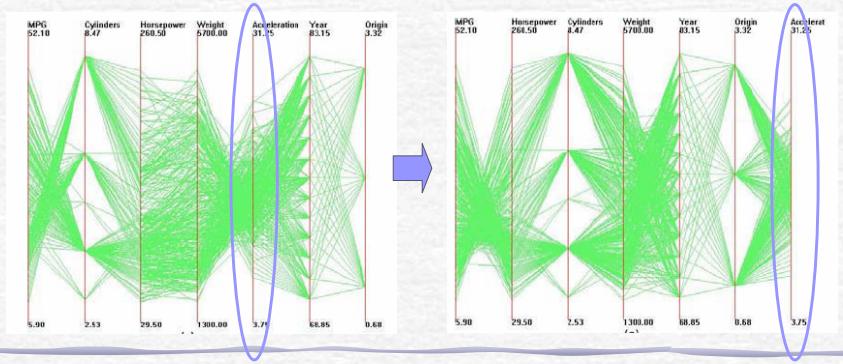


Steps:

- Split the lens into smaller areas or bins
- Calculate the random overpoltted% for each bin
- Perform a weighted average

Dimension Reordering

- InfoVis 2004
  - "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering", by Wei Peng; Ward, M.O.; Rundensteiner, E.A



General Idea

- Neighboring advantages
  - Inter-dimensional relationship between two neighboring dimensions is more clear
- Outlier causing clutter
  - A lot of outlier means a little relationship
- Key idea
  - To maximally reduce outliers between neighboring dimensions

Clutter Defined with Outlier

- C = $S_{outlier}$ / (n-1) / S_{total}
  - C: clutter measurement
  - $S_{outlier}$: total number of outliers between neighbors
  - S_{total} : total number of data points
- Find outlier
  - Calculate Euclidean distances between each pair of data
  - If a point has no neighbor whose distance to it is less than a threshold, then it is an outlier.

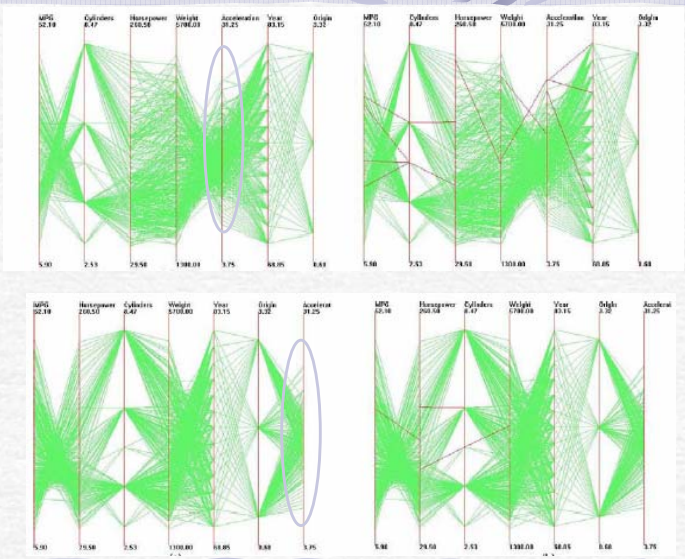
Dimension Swap

- Brute-force
- Random swap
- Nearest-Neighboring
- Greedy

Time Cost

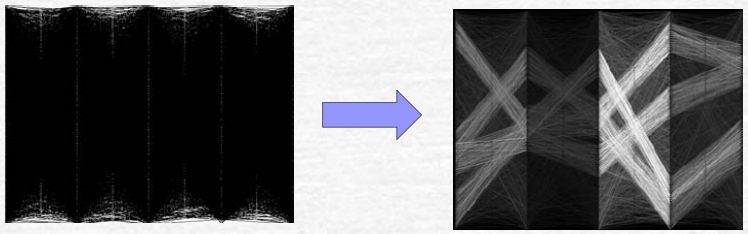
- Dataset size: m, in n dimension
  - Find outlier: O(m*m*n)
  - Find the optimal dimensional order: O(n*n!)

Dataset	Data Number	Dimensionality	Algorithm	Time
Census-Income	200	42	Nearest-Neighbor Algorithm	2 sec.
			Greedy Algorithm	3 sec.
			Random Swapping	2 sec.
AAUP	1161	14	Nearest-Neighbor Algorithm	7 sec.
			Greedy Algorithm	9 sec.
			Random Swapping	6 sec.



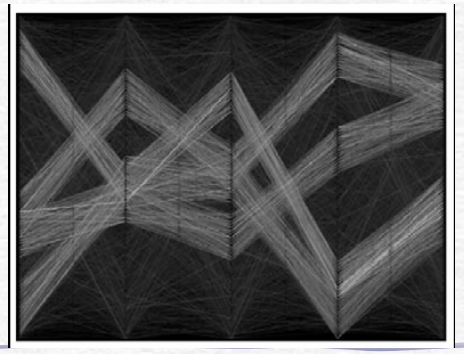
Histogram Uncovering Cluster

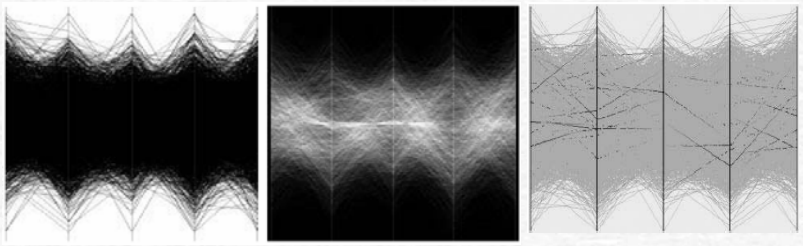
- InfoVis2004
  - "Uncovering Clusters in Crowded Parallel Coordinates Visualizations", by Almir Olivette Artero, Maria Cristina Ferreira de Oliveira, Haim Levkowitz



Frequency/Density Histogram

- Key idea
  - Create bi-dimensional frequency histogram for each pair of data
  - Line intensity is proportional to the frequency/density

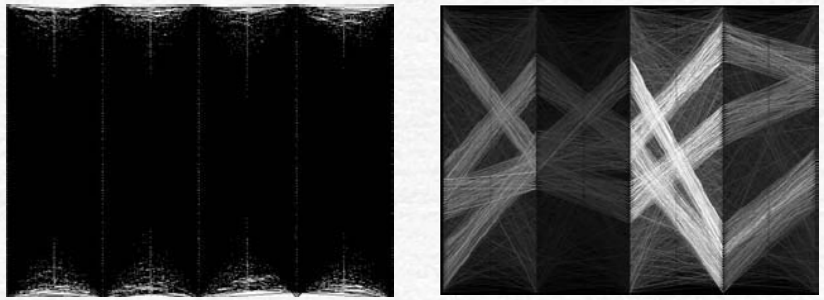




Original By intensity By histogram

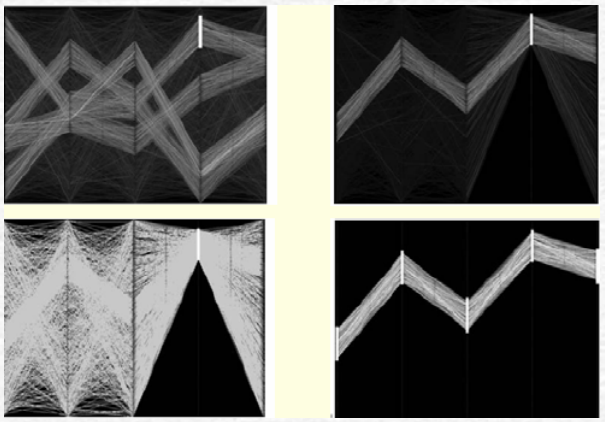
Scale Factors

It can adjust scale factor for intensity



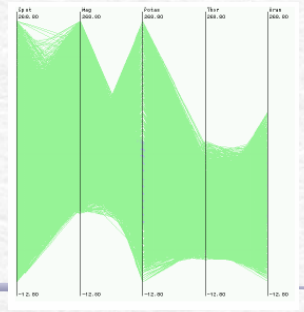
S=1.0 S=0.5 S=2.0 S=1.0

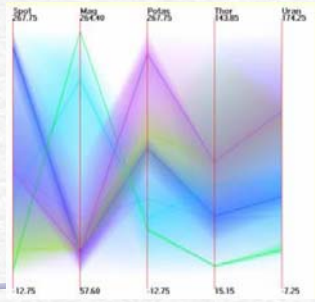
Cluster Selection



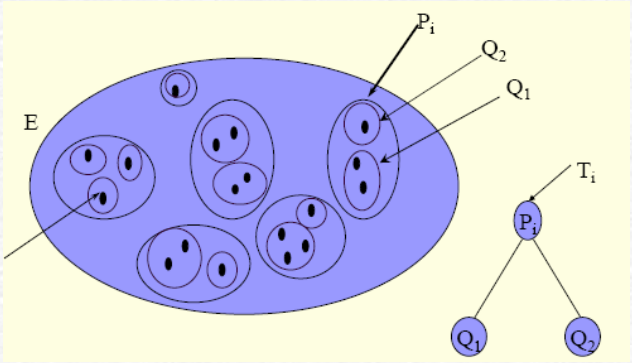
Hierarchical Parallel Coordinates

- Vis 99
  - "Hierarchical Parallel Coordinates for Exploration of Large Datasets", by Ying-Huey Fua, Matthew O. Ward and Elke A. Rundensteiner



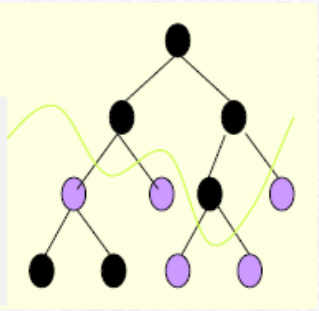


Cluster Tree



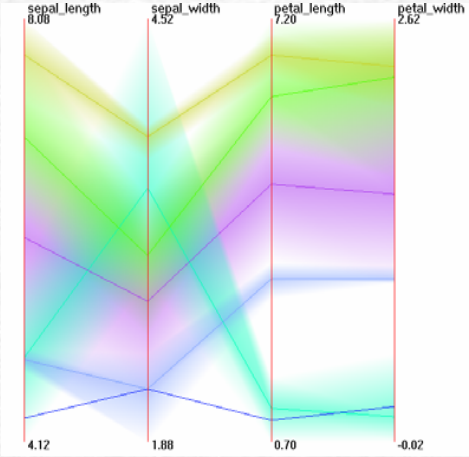
Cluster Cut

A cut across the tree is a partition on E if it intersects any given path exactly once.

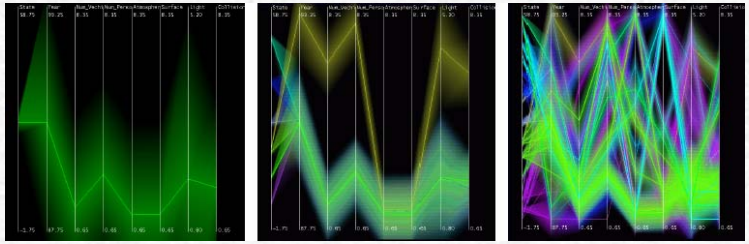


Hierarchical Parallel Coordinates

- Opacity conveys cluster population
- Color similarity indicates proximity in hierarchy
- Mean stretches has deepest opacity



Levels of Detail



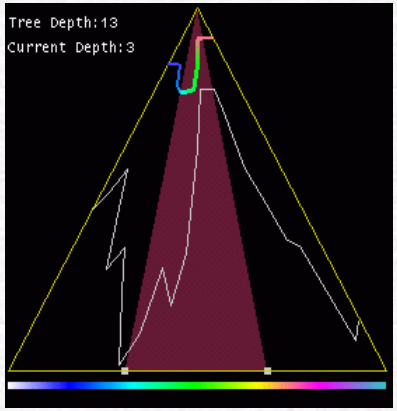
Root level Intermediate level Leaf level

Structure-Based Brushing

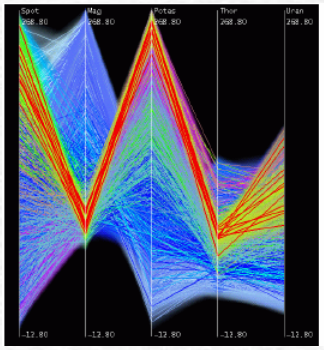
- Enhancement to screen-based and data-based methods
- Specify focus, extents, and level of detail
- Intuitive - wedge of tree and depth of interest
- Implemented by labeling/numbering terminals and propagating ranges to parents

Structure-Based Brush

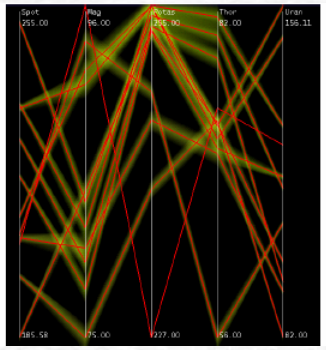
- White contour links terminal nodes
- Red wedge is extents selection
- Color curve is depth specification
- Color bar maps location in tree to unique color
- Direct and indirect manipulation of brush



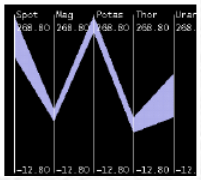
Dimension Zooming



Big picture

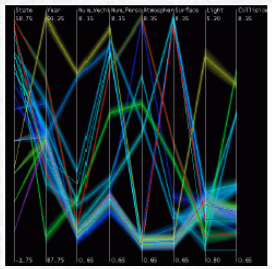


Magnified brush

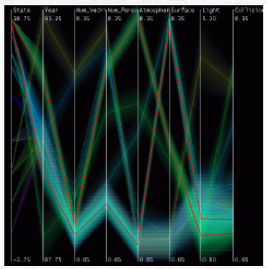


Mini map

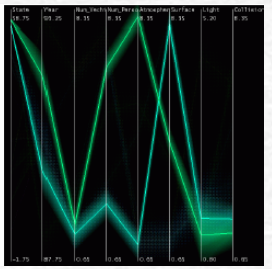
Dynamic Masking



Original



Partial fading



Complete fading

Conclusion

- Parallel Coordinate creates parallel dimensions
  - To explore clusters, trends or outlier of high dimensions
- Large dataset with high dimension
  - Cluttered display
  - Difficult user interactions
- Approaches of solution
  - Sampling, focus+context, histogram, dimension reduction & reorganization, hierarchical presentation...