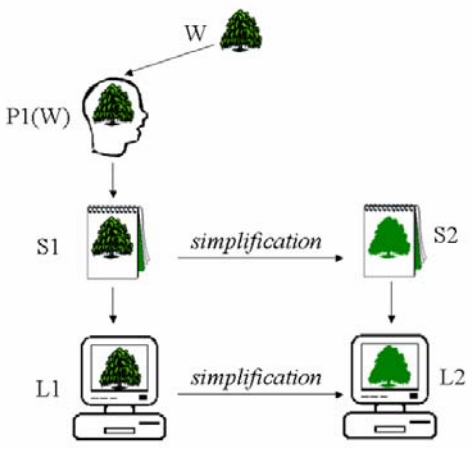
Clustering, Histograms, Sampling, MDS, and PCA

Class 11

1

Recall: The MRV Model



2

Recall: Simplification

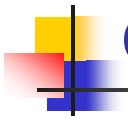
- Simplification operators - today!
- Simplification operands
  - Data space (structure level)
    - Data item space
    - Dimension space
    - Topology space
  - Visualization space (language level)
    - Visualization structure space
    - Visual encoding space
    - Screen space

3

Simplification operators

- Clustering
- Sampling
- Histogram
- Multidimensional Scaling (Jeong, Min)
- Principal Component Analysis (Tom)

4



Clustering

- **Definition**: *Clustering* is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar among themselves and dissimilar to objects of other groups [Ber02].

5

Why use clustering in MRV?

- By visualizing clusters rather than the original data the number of visual elements displayed can be greatly reduced.
- Clustering itself is a pattern discovering process. Thus visualizing clusters explicitly reveals hidden patterns to viewers.

6

Major Categories of Clustering Algorithms

- Hierarchical clustering
- Partitioning clustering
- Grid-based clustering
- Human-computer clustering
- Other approaches

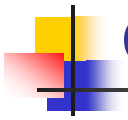
Ref: [Ber02]

7

Hierarchical Clustering

- *Hierarchical clustering* builds a cluster hierarchy (a tree of clusters, or a dendrogram).
  - Every cluster node contains child clusters
  - Sibling clusters partition the objects covered by their common parent
  - Allows exploring data on different levels of granularity

8



Categories of Hierarchical Clustering Approaches

- Agglomerative (bottom-up) Approaches
  - Start with one-object clusters and recursively merges two or more most appropriate clusters.
- Divisive (top-down) approaches
  - Start with one cluster of all objects and recursively splits the most appropriate cluster
  - Continue until a stopping criterion (frequently, the requested number k of clusters) is achieved.
- Distance (or similarity) measures between objects are used in traditional hierarchical clustering approaches

9

Another Categorization

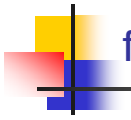
- According to how the distance between subsets of objects is decided when merging or splitting subsets of objects, which is called the *linkage metrics*.
- *Graph methods*
- *Geometric methods*

10

Another Categorization (con.)

- Graph (all point) methods
  - Minimum, maximum, or average of the distances measured for all pairs of objects with one object in the first set and another object in the second set as linkage metrics
  - Example: SLINK [Sib73], CLINK [Def77].
- Geometric (one centroid) methods
  - Central point of a cluster to represent the cluster
- Traditional graph methods suffers from time complexity

11



CURE: An Efficient Clustering Algorithm for Large Databases [Guha at. el. SIGMOD 98]

- Target: datasets with a large number of data items and a low number of numerical attributes
- Graph + geometric
  - Represents a cluster by a certain fixed number of points (not all, not one)
  - The distance between two clusters: the minimum of distances between two representative points
- Random sampling and partitioning

12

CURE [GRS98]

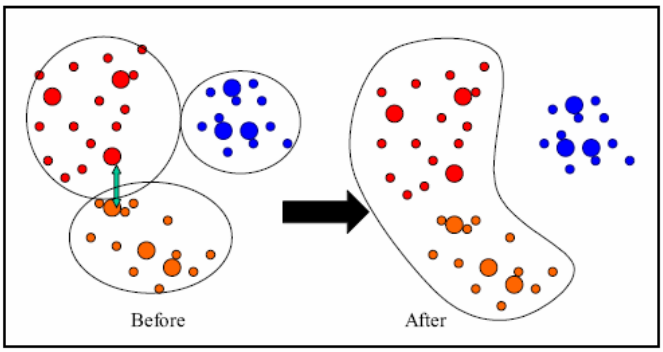


Figure 1: Agglomeration in Cure.

13

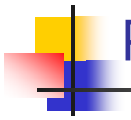
Partitioning Clustering

Partitioning algorithms divide data into several subsets

Major categories:

- Relocation algorithms
  - Iteratively reassign points between the clusters until an objective function is optimized
    - Probabilistic clustering
    - K-medoid and k-mean methods
- Density-based partitioning
  - A cluster is defined as a connected dense component growing in any direction the density leads
    - Density connectivity methods
    - Density function methods

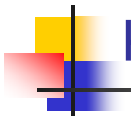
14



Probabilistic Clustering

- Data: a sample independently drawn from a mixture model of several probability distributions [MB88]
- The area around the mean of each distribution constitutes a natural cluster
- Goal: maximize the overall probability or likelihood of the data, given the final clusters
- Examples: EM (Expectation-Maximization) method [MB88], FREM [OO02]
  - EM methods accommodate categorical variables as well as numeric variables.

15



K-Medoids Methods

- A cluster is represented by one of its points, which is called a *medoid*
- When medoids are selected, clusters are defined as subsets of points close to respective medoids
- Objective function is defined as the averaged distance or another dissimilarity measure between a point and its medoid.
- Resist outliers well since peripheral cluster points do not affect the medoids.
- Example: CLARANS [NH94] for large datasets

16

K-Mean Methods

- Represent a cluster by the mean (weighted average) of its points, the so called *centroid*
- Objective function: the sum of discrepancies between the points and their centroids expressed through appropriate distance
- Example: Forgy's algorithm [For65]

17

Density-Based Partitioning Clustering

- A cluster is defined as a connected dense component growing in any direction the density leads
- Capable of discovering clusters of arbitrary shapes that are not rectangular or spherical
- Book: [Han & Kamber 2001]

Figure 4. Irregular shapes difficult for *k*-means are

18

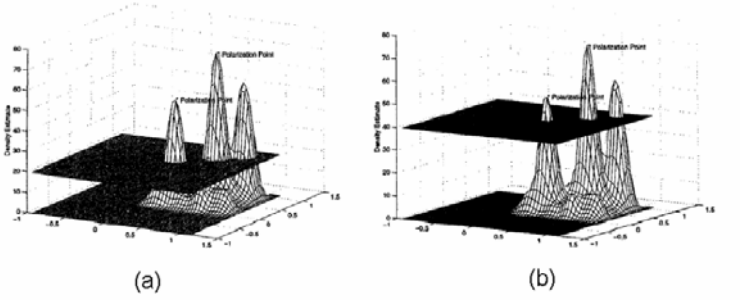
Grid-based Clustering

- Data partitioning is induced by points' membership in segments (cubes, cells, or regions) resulted from space partitioning
- Space partitioning is based on grid-characteristics accumulated from input data
- Independent of data ordering
- Different attribute types
- Contains features of both partitioning and hierarchical clustering

19

Human-Computer Clustering

- Meaningfulness and definition of a cluster are best characterized with use of human intuition [Agg01].



(a) (b)

Figure 3.8: The IPCLUS algorithm allows users to interactively select density value thresholds to separate clusters [Agg01]. (a) Two clusters are formed with a lower density value threshold. (b) Three clusters are formed for the same data projection with a higher density value threshold. This figure is used without the authors' permission.

20

Reference

- Survey of Clustering Data Mining Techniques Pavel Berkhin

21

Histograms

- A *histogram* partitions the data space into buckets. In each bucket, the data distribution is often assumed uniform and recorded using simple statistic data.
  - The distribution of each bucket can also be approximated using more complex functions and statistical data.
- Histograms are used to capture important information about the data in a concise representation [WS03]
- Selectivity estimation

22

One-Dimensional Histograms

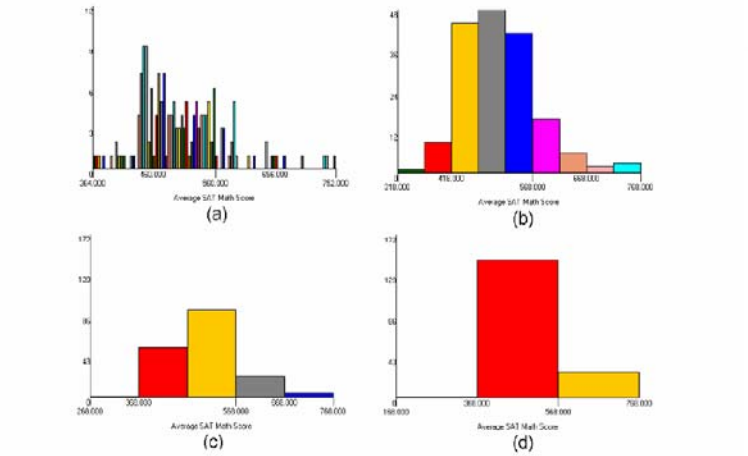


Figure 3.13: Equi-width histograms of a data set that containing the SAT math scores of students in some colleges with various bin sizes. These figures were captured from [NIS05].

23

Selectivity Estimation for M-D Datasets

- 1-D histograms
  - Dimension independent assumption
  - Project data set to each dimension and construct a 1-D histogram upon each projection
  - For a given multi-dimensional range (*query*), the number of data items falling into this range (*selectivity* of the query) is estimated in this way:
    - Project the query on each dimension, estimate the selectivity of each one-dimensional query using the 1-D histograms
    - Multiply the selectivities from all one-dimensional queries to get the selectivity of the multi-dimensional query

24

Selectivity Estimation for M-D Datasets

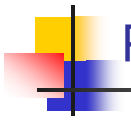
- M-D histograms
  - A multi-dimensional histogram partitions the data space into buckets in the multi-dimensional space. In each bucket, the data distribution is often assumed to be uniform and recorded using simple statistic data.

25

Accuracy Comparison of 1-D and M-D approaches

- Assumption: All histograms are accurate
- Comparison: M-D ☺ 1-D ☹
- Reason: 1-D approach is based on attribute independent assumption, which is often false in M-D space
- Example:
  t.A = t.B for all tuples
  1O% tuples have t.A = t.B = 0.5
  Query "count tuples whose t.A = 0.5 & t.B = 0.5"
  - Answer of 1-D approach: 1% ☹
  - Answer of M-D approach: 10% ☺
- However, accuracy of M-D histograms degrades fast with increasing dimensionality!

26



Finite VS. Real Domains

- Finite domains:
  - Data distribution can be accurately expressed.
  - Number of distinct values in a bucket is often recorded so that frequency of each distinct value combination can be estimated.
- Real domains:
  - Infinite possible distinct values
  - Not many values will appear more than once.
  - Data distribution can hardly be accurately expressed.
  - Density of a bucket is often recorded and used.

27

Paper: Approximating Multi-Dimensional Aggregate Range Queries over Real Attributes

(D. Gunopulos, G. Kollios, V. Tsotras, and C. Domeniconi. 2000)

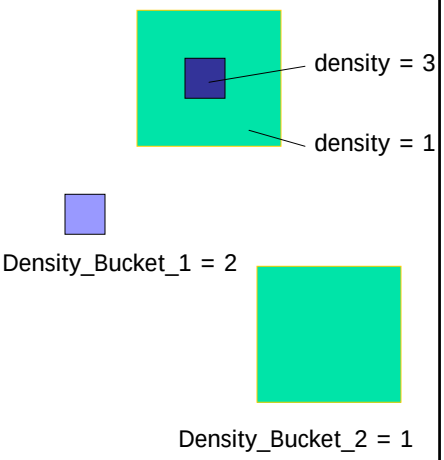
- Target: Selectivity estimation for real datasets (GENHIST)
- Motivation:
  - Smaller bucket estimates density better but increases bucket number
  - Large bucket number causes more partially intersected buckets

28

Paper: Approximating Multi-Dimensional Aggregate Range Queries over Real Attributes

(D. Gunopulos, G. Kollios, V. Tsotras, and C. Domeniconi. 2000)

- Key ideas:
  - Use rectangular buckets
  - Allow buckets to overlap and have variable sizes
  - density = sum of density of overlapping buckets
- Construction:
  - Iteratively smooth data density by removing some data points from denser areas to form buckets
  - Regular grids are used in each iteration. Grids are coarser in later iterations.



29

Paper: Approximating Multi-Dimensional Aggregate Range Queries over Real Attributes

(D. Gunopulos, G. Kollios, V. Tsotras, and C. Domeniconi. 2000)

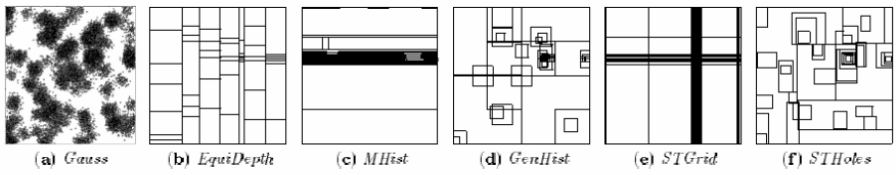


Figure 3.17: The Guass data set and its histograms [BCG03]. (a) The Guass data set (b) EquiDepth histogram (c) Mhist histogram (d) GenHist histogram (e) STGrid histogram (f) STHoles histogram. This figure is used without the authors' permission.

30

Paper: Approximating Multi-Dimensional Aggregate Range Queries over Real Attributes

(D. Gunopulos, G. Kollios, V. Tsotras, and C. Domeniconi. 2000)

- Advantages:
  - Compact - more regions with different densities can be expressed using the same number of buckets
  - True multi-dimensional – buckets are constructed according to density of the M-D space
- Disadvantages:
  - Uniform parameter setting degrades performance for some datasets.
  - It requires multiple passes of the whole dataset.
  - It can't adapt to underlying data distribution change as dynamic histograms.

31

Self-Tuning Histograms

- Paper 1: Self-tuning histograms: Building histograms without looking at data

  (A. Aboulnaga and S. Chaudhuri. 1999)

  STGrid

- Paper 2: STholes: A multidimensional workload-aware histogram

  (N. Bruno, S. Chaudhuri, and L. Gravano. 2003)

  STHoles

32

Self-Tuning Histograms

- Drawbacks of traditional (static) M-D histograms that access whole datasets to construct the histograms
  - High construction and maintaining cost
  - Waste resources (use assumption that all queries are equally likely)
- Key ideas of Self-Tuning (S-T) histograms
  - Construct and refine histograms using query results without accessing the dataset
  - Allocate more resources to heavily queried areas

33

Self-Tuning Histograms

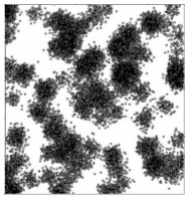
- Life circle of S-T histograms has two stages:
  1. A coarse histogram is built without accessing the data.
  2. The coarse histogram is refined using query feedbacks

     Compare estimation and actual query results to find inaccurate buckets and adjust them
- Advantages of S-T histograms:
  - Cheaper
  - (May be) more accurate
  - Adaptive to changes

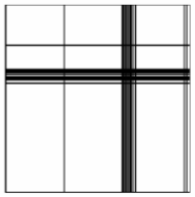
34

Differences between STGrid and STHoles

- STGrid: keeps rigid grid partitioning
  - Refine – split and merge rows of buckets
  - Unrelated buckets will be split.
  - Refine based on unidimensional statistic
- STHoles: Allows a bucket to be inside another one
  - Refine – drill holes and merge similar buckets
  - Unrelated buckets won't be affected
  - Refine based on M-D information



(a)



(b)

(c)

Example:

(a) Original data set (b) STGrid (c) STHoles

Figures are used without users' permission

Sampling

- The process of selecting some part of a population to observe so that one may estimate something about the whole population [Tho92]
- Input
  - $A = (a1, a2, \ldots, ak)$
- Output:
  - $B = (ai1, ai2, \ldots, aim)$
    - $1 <= ij <= k$, for $1 <= j <= m < k$.

Simple Random Sampling

- *Simple random sampling* - a sampling procedure that assures that each element in the population has an equal chance of being selected.

37

Paper 1: A Bi-Level Bernoulli Scheme for Database Sampling (P.J. Haas and C. König. 2004)

- Motivation:
  - Row-level sampling is accurate but expensive.
  - Page-level sampling is cheap but less precise.
- Bi-level sampling with sampling rate q
  - Sample pages at rate p (p >= q)
  - Sample rows of sampled pages at rate r = q/p
- Challenges:
  - How to set p and r to trade off speed and precision?

38

Paper 1: A Bi-Level Bernoulli Scheme for Database Sampling
(P.J. Haas and C. König. 2004)

- Solution for aggregation queries:
  - Calculate precision using standard error of unbiased estimator of an aggregation query
  - Calculate speed using I/O cost models
  - Solve the optimization problems, such as

    For a given sampling rate and maximum processing cost, minimize the standard error

39

Paper 1: A Bi-Level Bernoulli Scheme for Database Sampling
(P.J. Haas and C. König. 2004)

- Conclusion:
  - PHI, the measure of variability of values within a page relative to the variability of values between pages, plays a center role in the optimal solution of p and r.
  - PHI < 1, the sampling is as row-like as possible
  - PHI >= 1, the sampling is as page-like as possible
- This paper proposes a heuristic approach to estimate p and r using a small number of statistic data

40

Paper 1: A Bi-Level Bernoulli Scheme for Database Sampling
(P.J. Haas and C. König. 2004)

- Sampling rate is distributed to row-level and page-level sampling rate
- Bernoulli sampling is used – if a page or row is selected is decided by the sampling rate and is independent from other pages or rows.
  - It requires to rescan the whole dataset if sampling rate changes.
  - It is discrete in nature

41

Paper 2: Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Data Sets
(G. Kollios, D. Gunopulos, N. Koudas, and S. Berchtold. 2003)

- Motivation:
  - A large sample is needed for detecting small clusters using uniform random sampling.
- Key ideas:
  - Estimate the density of the dataset
  - Bias sampling rate according to local density so that clusters or outliers can be detected using a relatively small sample

42

Paper 2: Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Data Sets

(G. Kollios, D. Gunopulos, N. Koudas, and S. Berchtold. 2003)

- Strategies:
  - Oversample dense regions to identity clusters
  - Oversample sparse regions to identify outliers
- Biased probability calculation function:

  $c * f^{a}$

  a: biasing factor
  - a = 0: uniform sampling
  - a < 0: favors sparse regions
  - a > 0: favors dense regions

43

Scenarios to Use Biased Sampling in MRV

- If outliers or clusters need to be detected
- If a data set contains a number of unreliable objects (for example, data items that contain many missing values
- If a user finds and marks interesting objects in interactive exploration and wants
- In a focus + context display

44

Paper 2: Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Data Sets

(G. Kollios, D. Gunopulos, N. Koudas, and S. Berchtold. 2003)

- Information loss measure:
  - Number of missed clusters
- Sampling rate changes:
  - Sampling rate changes with data density in M-D space
  - If transition of density in M-D space is continuous, transition of sampling rate is continuous.

45

Paper 3: Dynamic Sample Selection for Approximate Query Processing

(B. Babcock, S. Chaudhuri, and G. Das. 2003)

- Motivation:

  An appropriately biased sample provides good accuracy for a particular set of queries but not for the others.
- Key ideas:
  - Pre-process the dataset to build and store a set of biased samples
  - Construct an appropriately biased sample from them for a give query
- Gain: processing speed and accuracy
- Cost: pre-processing and disk space

46

Paper 3: Dynamic Sample Selection for Approximate Query Processing
(B. Babcock, S. Chaudhuri, and G. Das. 2003)

Proposed technique:

Small group sampling for aggregation queries with group-bys.

- Pre-process: Build a small group table for each eligible column, and an overall sample table
- Query: Decompose a query to the overall sample table and one or more small group tables, and compose the final answer from these answers.

47

Paper 3: Dynamic Sample Selection for Approximate Query Processing
(B. Babcock, S. Chaudhuri, and G. Das. 2003)

- Information loss measure:
  - Group loss
  - Average squared relative error of the estimator
- Sampling rate is different for overall sample table and small groups:
  - Sampling rate of overall sample table is decided by input parameters
  - Sampling rate of small groups is 100%.

48

Paper 4: By Chance - Enhancing Interaction with Large Data Sets through Statistical Sampling (A. Dix and G. Ellis. 2002)

- Goal: summarize the relationships between random sampling and visualization.
- Why random sampling benefits visualization?
  - Fasten interactions
  - Reduce clutter
- Why random sampling is acceptable in visualization?
  - Gestalt visual processing often depends on approximate rather than exact properties of data.
  - A dataset itself is often a sample from the real world.

49

Paper 4: By Chance - Enhancing Interaction with Large Data Sets through Statistical Sampling (A. Dix and G. Ellis. 2002)

- When random sampling can be applied in visualization?
  - Visualization based on aggregate or summary statistics can use sampled data to give approximations
  - Visualization containing points or lines that could saturate the display can use sampled data to avoid saturation and reveal features and relationships.
  - Sampling can be used to reduce the data set to a size that allows detailed visualization such as thumbnails of individual items.

50

Paper 4: By Chance - Enhancing Interaction with Large Data Sets through Statistical Sampling (A. Dix and G. Ellis. 2002)

- Information loss:
  - Measure - how the visualization obtained from the sample is distinguishable from that obtained from the full data set.
  - Users should be aware of the information loss – use error bars, blurred edges, and ragged displays
- Sampling rate changes:
  - Smooth transitions between different resolutions are beneficial
  - Example: when sampling rate increases, keep the data points in the previous view on the screen.

51

Paper 4: By Chance - Enhancing Interaction with Large Data Sets through Statistical Sampling (A. Dix and G. Ellis. 2002)

Example - Astral Telescope Visualizer

- Users can interactively change the zoom value of a 2D scatterplot display.
- The sampling rate increases automatically with the increase of the square of the zoom value.
- Extra points are sampled and previously sampled data points in the zoomed area are remained after a zoom in operation.
- It permits smooth transition between sampling rates.

52

Distinct Measures of Information Loss

- Standard error (sqare root of the variance) of the unbiased estimator
- Cluster loss
- Group loss
- Average squared relative error
- How the visualization obtained from the sample is distinguishable from that obtained from the full data set.

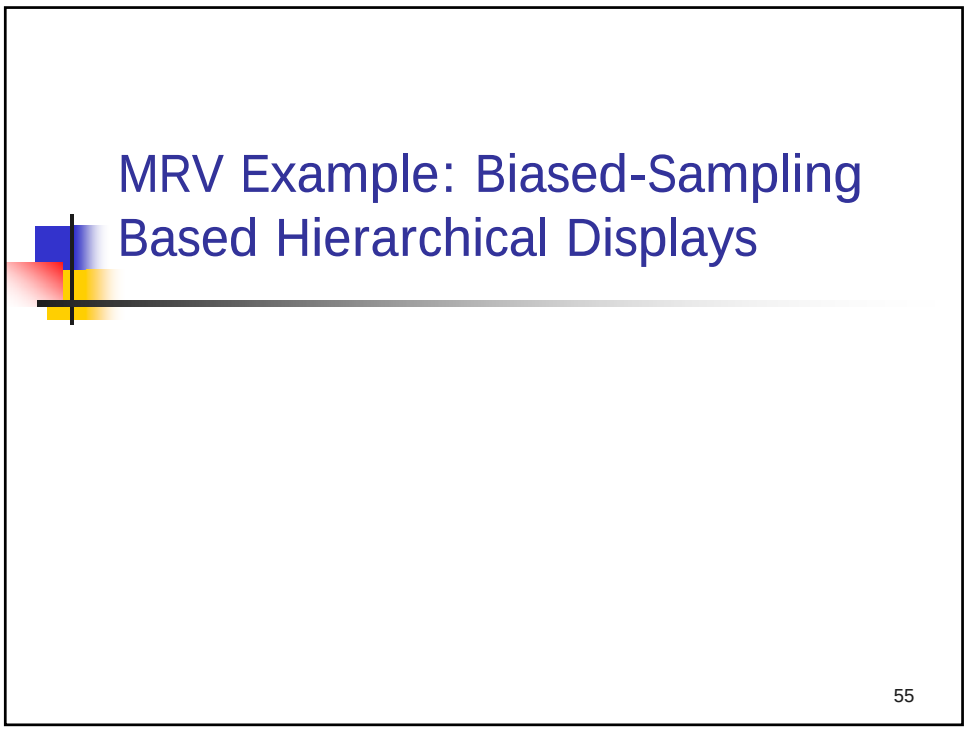
53

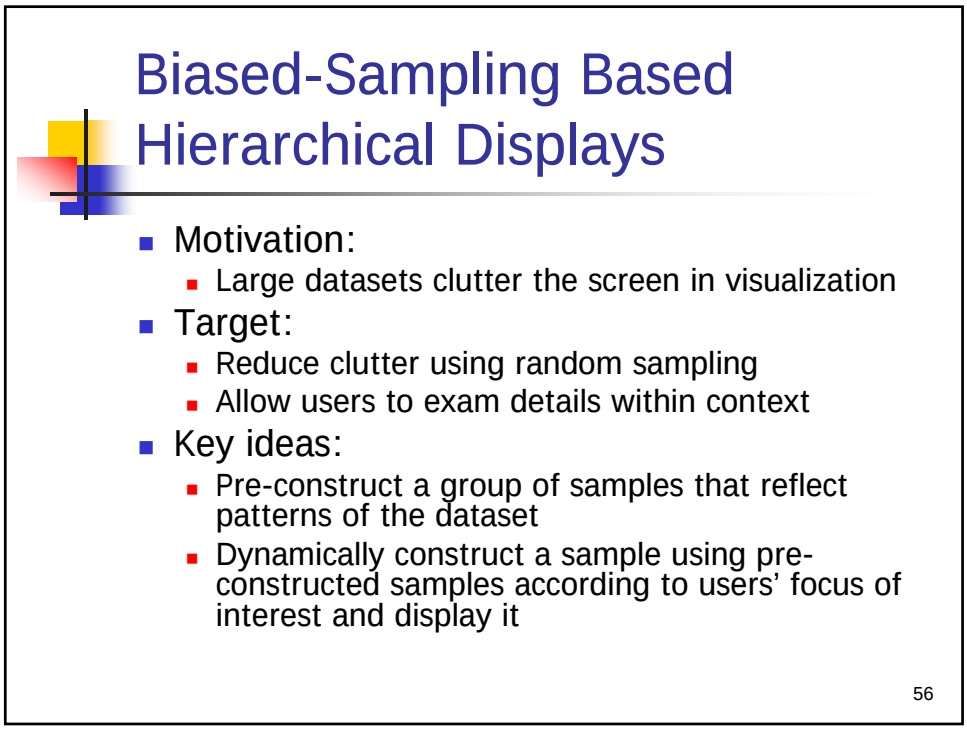
A Scenario: Small Change in Sampling Rate Causes Significant Information Loss

Page-level Bernoulli sampling with large pages

- Bernoulli sampling cause randomness in sample size
- Page-level Bernoulli sampling magnifies this fluctuations by a factor equal to the number of rows per page

54


MRV Example: Biased-Sampling Based Hierarchical Displays
55


Biased-Sampling Based Hierarchical Displays
Motivation:
Large datasets clutter the screen in visualization
Target:
Reduce clutter using random sampling
Allow users to exam details within context
Key ideas:
Pre-construct a group of samples that reflect patterns of the dataset
Dynamically construct a sample using pre-constructed samples according to users' focus of interest and display it
56

Biased-Sampling Based Hierarchical Displays

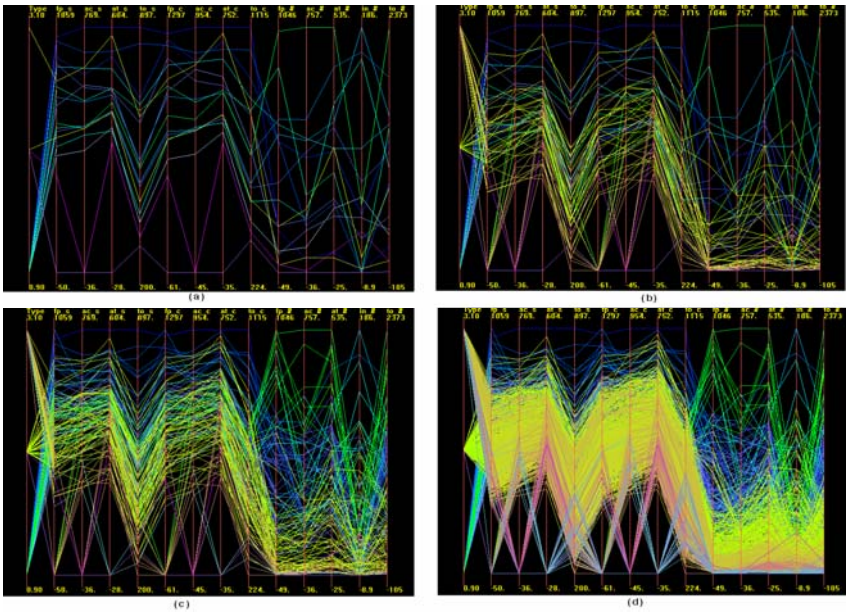
- Approach:
  - Construct a hierarchical cluster tree upon a dataset according to the similarity among the data items
  - Each node of the tree carries a sample
  - Provide users an interface to select focus of interests and preferred levels of detail from the tree
  - Dynamically construct a sample from the samples in the nodes according to users' selection and display it

57

Biased-Sampling Based Hierarchical Displays

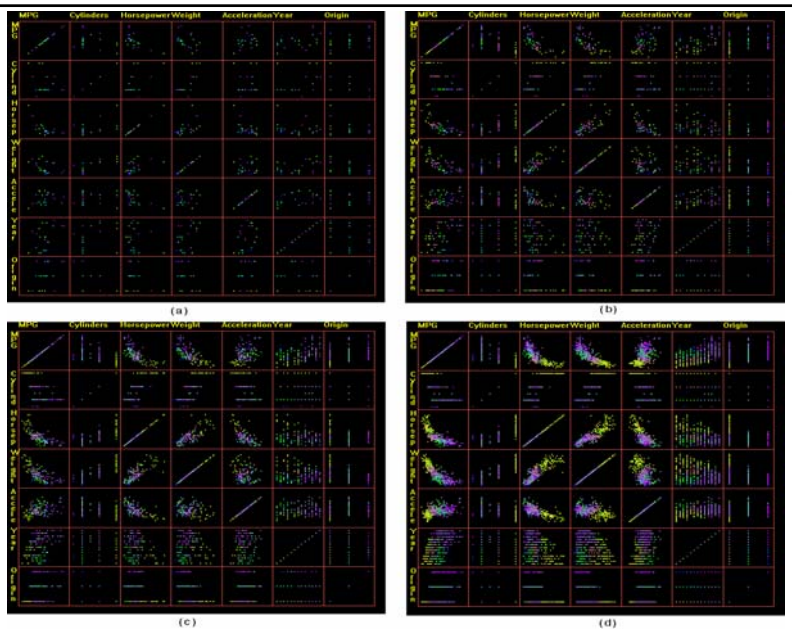
- Sample of a node:
  - Contains data items belongs to the cluster
  - Biased to clusters or outliers
  - Do not contain data items in samples of its ascendant nodes
- Dynamic sample:
  - The sum of samples of selected nodes and their ascendant nodes

58



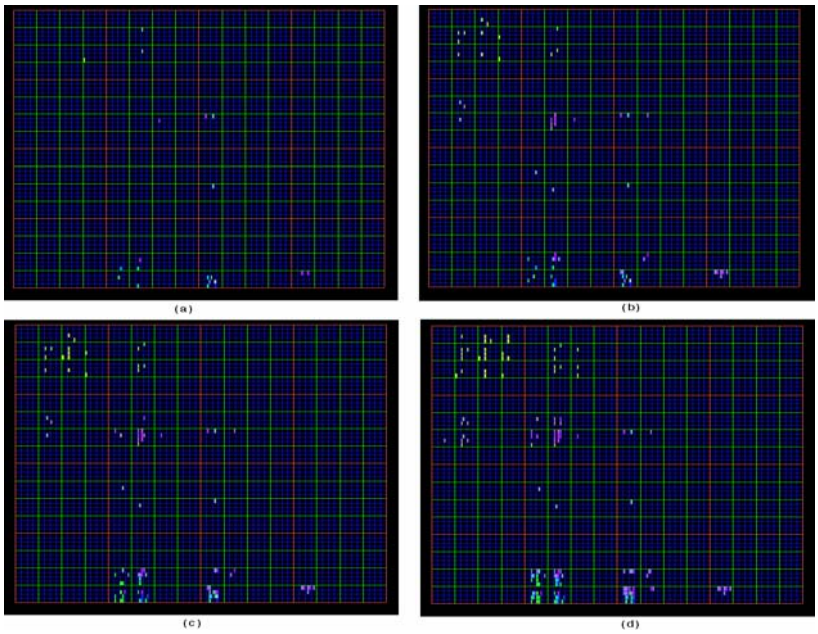
(a)-(d): The BSBH parallel coordinates displays of the AAUP dataset (14 dimensions, 1161 data items) with increasing LODs.

59



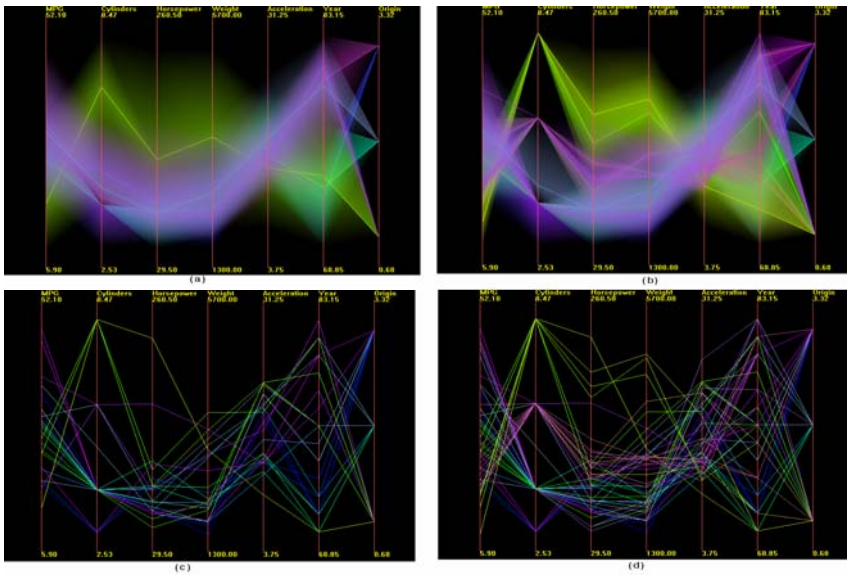
(a)-(d): The BSBH scatterplot matricies displays of the Cars dataset (7 dimensions, 392 data items) with increasing LODs.

60



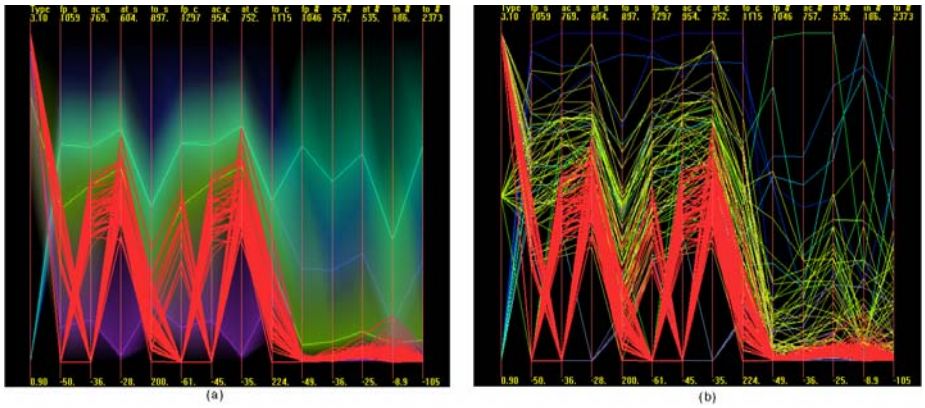
(a)-(d): The BSBH dimensional stacking displays of the Cars dataset with increasing LODs.

61



(a) and (b): The Hierarchical Parallel Coordinates displays of the Cars dataset with two adjacent LODs. There is a visual jitter between (a) and (b). For example, the cluster in the top left of (a) disappears in (b). (c) and (d): The counterpart of (a) and (b) in the BSBH parallel Coordinates displays. The display changes smoothly from (c) to (d).

62



(a): A focus-in-context view of the AAUP dataset in the Hierarchical Parallel Coordinates display. There is a jump of LOD between the focus and context.
(b): The counter part of (a) in the BSBH parallel coordinates display. The LOD changes smoothly from the focus to the context.

63