Chapter 14 – Cluster Analysis 聚类分析

Instructor: Zach Zhizhong ZHOU,
Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Clustering: The Main Idea 聚类主要思想

Goal: Form groups (clusters) of similar records 目标：把类似的记录分组（聚类）

Used for segmenting markets into groups of similar customers 用在市场分割上，将类似的顾客归入几个组别中。

Example: Claritas segmented US neighborhoods based on demographics & income: "Furs & station wagons," "Money & Brains", …

Other Applications 其他应用

- Classification of species 进行种属分类

- Grouping securities in portfolios 投资组合中的股票聚类

- Grouping firms for structural analysis of economy 将企业进行聚类以便对经济做结构性分析

- Army uniform sizes 军服大小

Example: Public Utilities

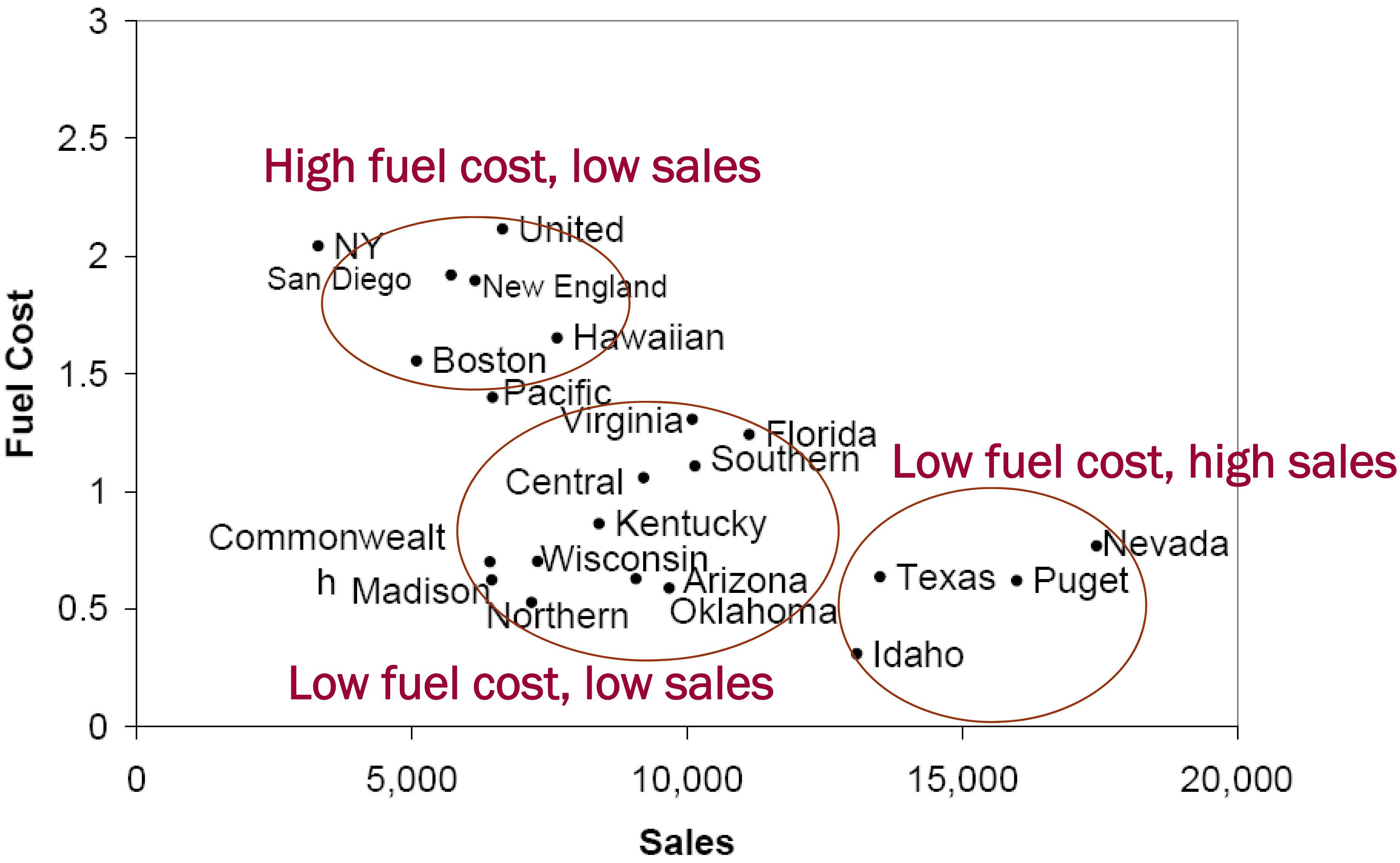
Goal: find clusters of similar utilities

Data: 22 firms, 8 variables

- Fixed-charge covering ratio
- Rate of return on capital
- Cost per kilowatt capacity
- Annual load factor
- Growth in peak demand
- Sales
- % nuclear
- Fuel costs per kwh

Company	Fixed_charge	RoR	Cost	Load	Δ Demand	Sales	Nuclear	Fuel_Cost
Arizona	1.06	9.2	151	54.4	1.6	9077	0	0.628
Boston	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
Central	1.43	15.4	113	53	3.4	9212	0	1.058
Commonwealth	1.02	11.2	168	56	0.3	6423	34.3	0.7
Con Ed NY	1.49	8.8	192	51.2	1	3300	15.6	2.044
Florida	1.32	13.5	111	60	-2.2	11127	22.5	1.241
Hawaiian	1.22	12.2	175	67.6	2.2	7642	0	1.652
Idaho	1.1	9.2	245	57	3.3	13082	0	0.309
Kentucky	1.34	13	168	60.4	7.2	8406	0	0.862
Madison	1.12	12.4	197	53	2.7	6455	39.2	0.623
Nevada	0.75	7.5	173	51.5	6.5	17441	0	0.768
New England	1.13	10.9	178	62	3.7	6154	0	1.897
Northern	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
Oklahoma	1.09	12	96	49.8	1.4	9673	0	0.588
Pacific	0.96	7.6	164	62.2	-0.1	6468	0.9	1.4
Puget	1.16	9.9	252	56	9.2	15991	0	0.62
San Diego	0.76	6.4	136	61.9	9	5714	8.3	1.92
Southern	1.05	12.6	150	56.7	2.7	10140	0	1.108
Texas	1.16	11.7	104	54	-2.1	13507	0	0.636
Wisconsin	1.2	11.8	148	59.9	3.5	7287	41.1	0.702
United	1.04	8.6	204	61	3.5	6650	0	2.116
Virginia	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

Sales & Fuel Cost: 3 rough clusters can be seen



Extension to More Than 2 Dimensions

In prior example, clustering was done by eye 上面的例子中，我们通过观察得到聚类。

Multiple dimensions require formal algorithm with 对更高维度的记录进行聚类要求正式的算法，该算法包括

- A distance measure 距离的度量
- A way to use the distance measure in forming clusters 一种使用距离度量进行聚类的方法。

We will consider two algorithms: hierarchical and non-hierarchical 我们考虑2种算法：层次和非层次算法

Hierarchical Clustering 层次聚类

Hierarchical Methods 层次聚类方法

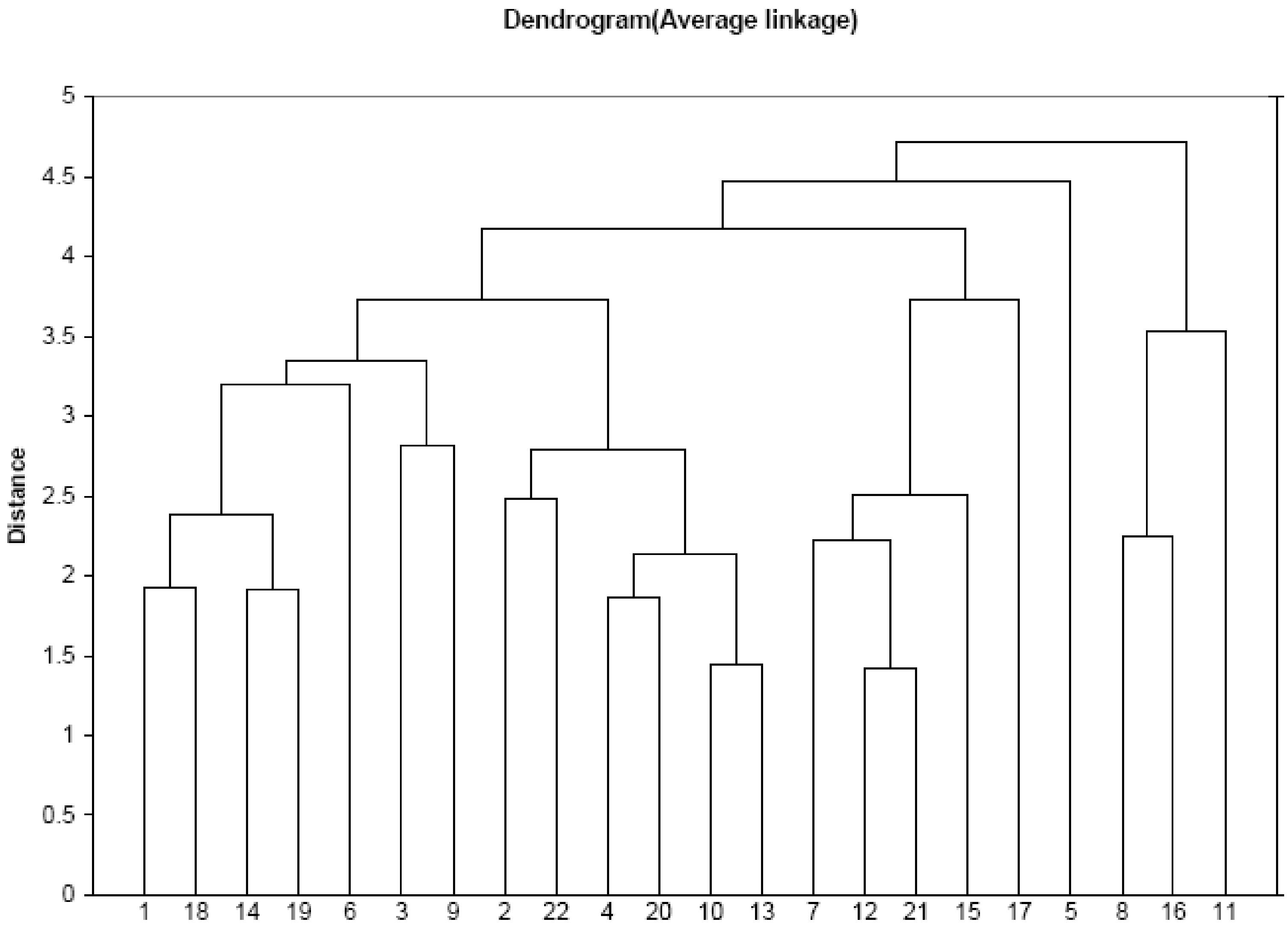
Agglomerative Methods 凝聚法

- Begin with n-clusters (each record its own cluster) 从n个簇开始（每个记录视为一个簇）
- Keep joining records into clusters until one cluster is left (the entire data set) 不断将记录加入簇中直到剩下一个簇（整个数据集）
- Most popular 最流行

Divisive Methods 分裂法

- Start with one all-inclusive cluster 从一个包括所有记录的簇开始
- Repeatedly divide into smaller clusters 不断分割成更小的簇

A Dendrogram shows the cluster hierarchy
一个显示簇的层次的树状图



Measuring Distance 衡量距离

- Between records 记录之间的距离
- Between clusters 簇之间的距离

Measuring Distance Between Records

Distance Between Two Records

Euclidean Distance is most popular:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

Normalizing

Problem: Raw distance measures are highly influenced by scale of measurements

Solution: normalize (standardize) the data first

- Subtract mean, divide by std. deviation
- Also called *z*-scores

Example: Normalization

For 22 utilities:

Avg. sales = 8,914

Std. dev. = 3,550

Normalized score for Arizona sales:

(9,077-8,914)/3,550 = 0.046

For Categorical Data: Similarity

To measure the distance between records in terms of two 0/1 variables, create table with counts:

	0	1
0	a	b
1	c	d

Similarity metrics based on this table:

- Matching coef. = (a+d)/p, p=a+b+c+d
- Jaquard's coef. = d/(b+c+d)
  - Use in cases where a matching "1" is much greater evidence of similarity than matching "0" (e.g. "owns Corvette")

Other Distance Measures

- Correlation-based similarity
- Statistical distance (Mahalanobis)
- Manhattan distance (absolute differences)
- Maximum coordinate distance
- Gower’s similarity (for mixed variable types: continuous & categorical)

Measuring Distance Between Clusters

Minimum Distance (Cluster A to Cluster B)

- Also called single linkage

- Distance between two clusters is the distance between the pair of records A_i and B_j that are closest

Maximum Distance(Cluster A to Cluster B)

- Also called complete linkage

- Distance between two clusters is the distance between the pair of records A_i and B_j that are farthest from each other

Average Distance

- Also called average linkage

- Distance between two clusters is the average of all possible pair-wise distances

Centroid Distance

- Distance between two clusters is the distance between the two cluster centroids.

- Centroid is the vector of variable averages for all records in a cluster

The Hierarchical Clustering Steps (Using Agglomerative Method)

1. Start with *n* clusters (each record is its own cluster)

2. Merge two closest records into one cluster

3. At each successive step, the two clusters closest to each other are merged

Dendrogram, from bottom up, illustrates the process

Table 15.1 Example of Single-Linkage Clustering

Original 5×5 distance matrix, with subsequent single-linkage clustering

$$\begin{array}{c} a \\ b \\ c \\ d \\ e \end{array} \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \mathbf{2} & 8 & 0 \end{bmatrix}$$

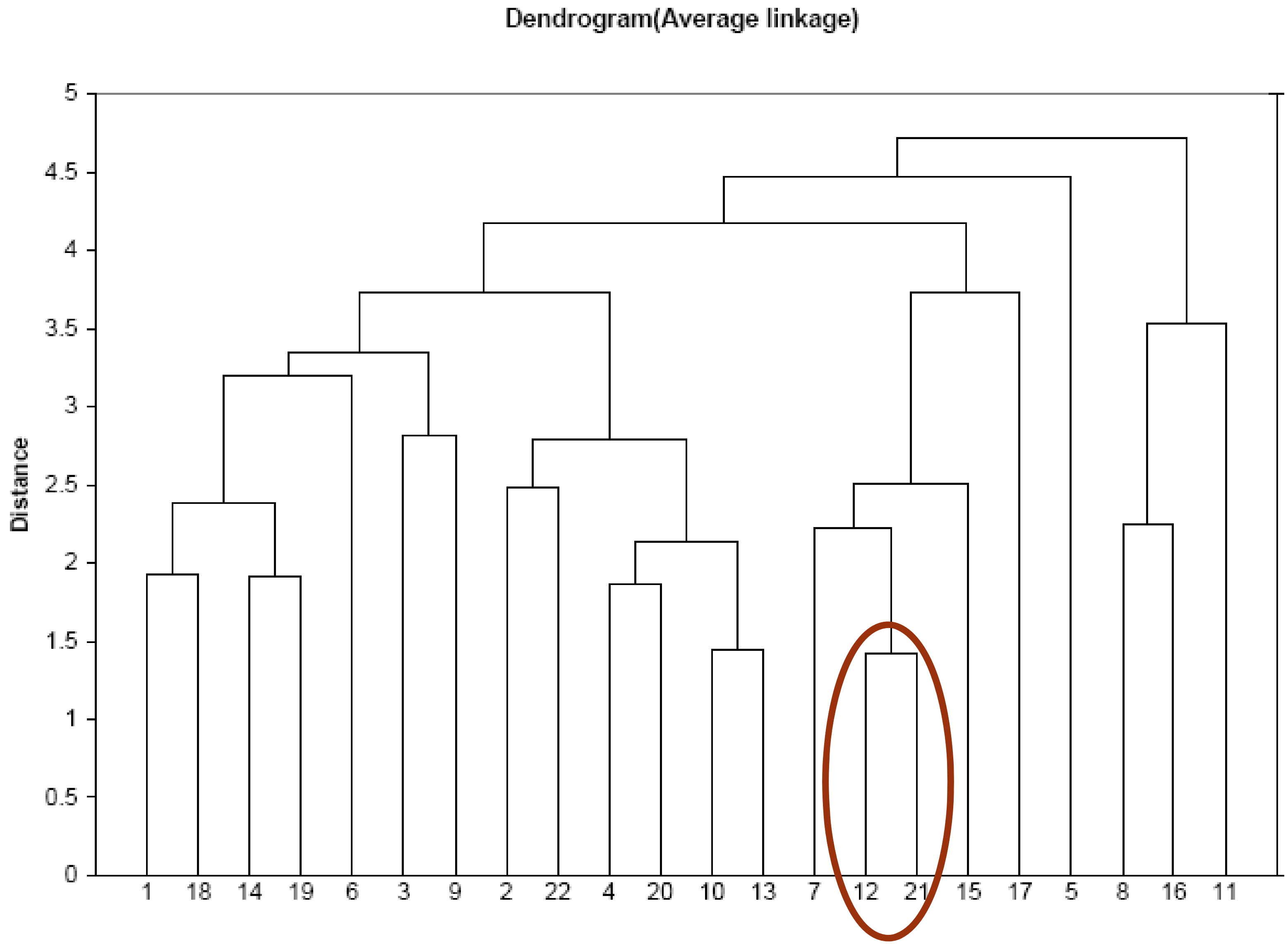
$$\begin{array}{c} (ce) \\ a \\ b \\ d \end{array} \begin{bmatrix} 0 & & & \\ \mathbf{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix}$$

$$\begin{array}{c} (ace) \\ b \\ d \end{array} \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & \mathbf{5} & 0 \end{bmatrix}$$

$$\begin{array}{c} (ace) \\ (bd) \end{array} \begin{bmatrix} 0 & \\ \mathbf{6} & 0 \end{bmatrix}$$

The numbers in bold face refer to the minimum distances.

Records 12 & 21 are closest & form first cluster



Reading the Dendrogram

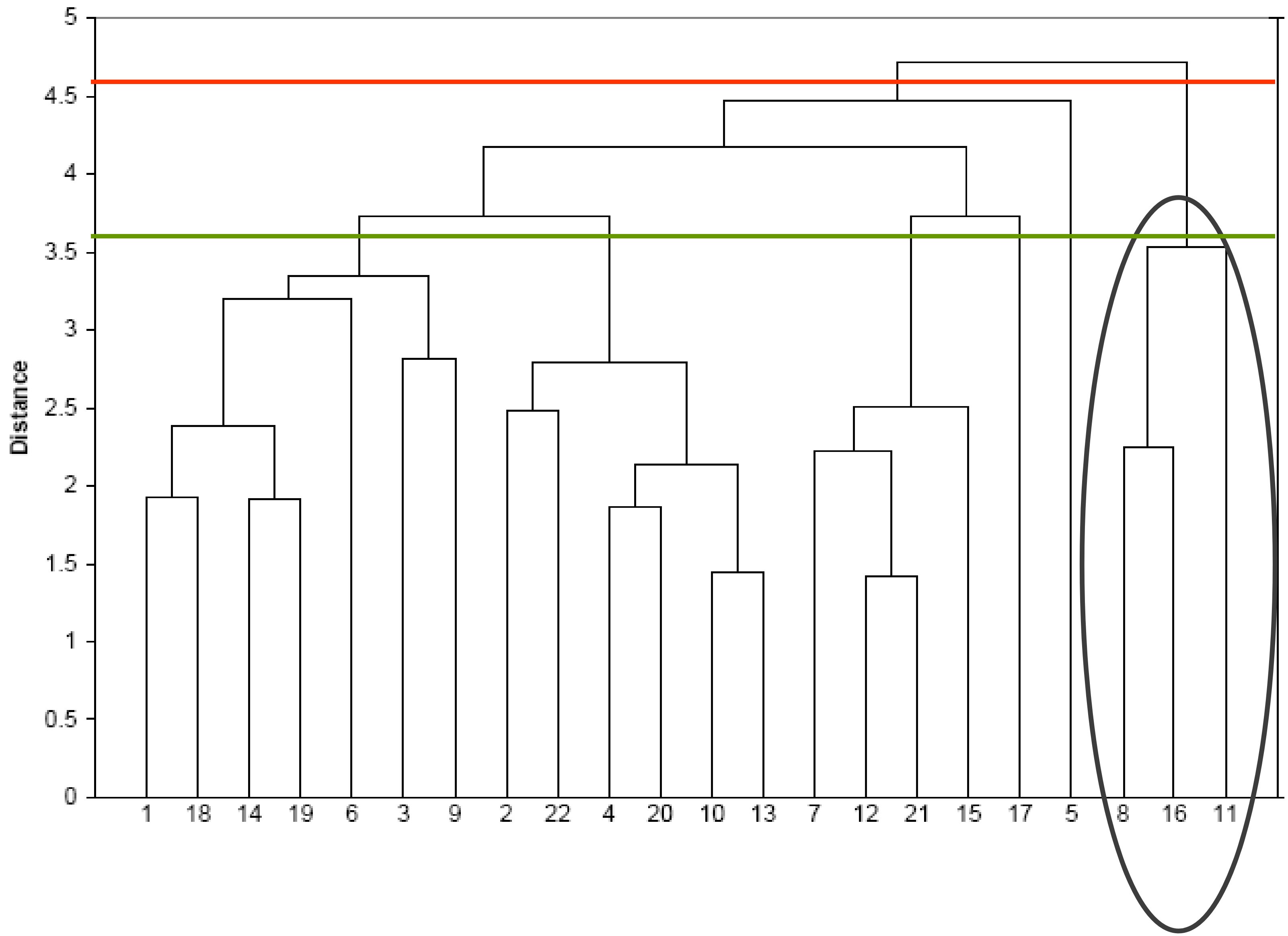
See process of clustering: Lines connected lower down are merged earlier

- 10 and 13 will be merged next, after 12 & 21

Determining number of clusters: For a given "distance between clusters", a horizontal line intersects the clusters that are that far apart, to create clusters

- E.g., at distance of 4.6 (red line in next slide), data can be reduced to 2 clusters -- The smaller of the two is circled
- At distance of 3.6 (green line) data can be reduced to 6 clusters, including the circled cluster

Dendrogram(Average linkage)



Validating Clusters

Interpretation

Goal: obtain meaningful and useful clusters

Caveats:

(1) Random chance can often produce apparent clusters

(2) Different cluster methods produce different results

Solutions:

- Obtain summary statistics
- Also review clusters in terms of variables not used in clustering
- Label the cluster (e.g. clustering of financial firms in 2008 might yield label like "midsize, sub-prime loser")

Desirable Cluster Features

Stability – are clusters and cluster assignments sensitive to slight changes in inputs? Are cluster assignments in partition B similar to partition A?

Separation – check ratio of between-cluster variation to within-cluster variation (higher is better)

Nonhierarchical Clustering: K-Means Clustering

K-Means Clustering Algorithm

1. Choose # of clusters desired, *k*

2. Start with a partition into k clusters

   Often based on random selection of k centroids

3. At each step, move each record to cluster with closest centroid

4. Recompute centroids, repeat step 3

5. Stop when moving records increases within-cluster dispersion

K-means Algorithm: Choosing k and Initial Partitioning

Choose *k* based on the how results will be used

e.g., "How many market segments do we want?"

Also experiment with slightly different *k*' s

Initial partition into clusters can be random, or based on domain knowledge

If random partition, repeat the process with different random partitions

XLMiner Output: Cluster Centroids

Cluster	Fixed_charge	RoR	Cost	Load_factor
Cluster-1	0.89	10.3	202	57.9
Cluster-2	1.43	15.4	113	53
Cluster-3	1.06	9.2	151	54.4

We chose $k = 3$

4 of the 8 variables are shown

Distance Between Clusters

Distance between	Cluster-1	Cluster-2	Cluster-3
Cluster-1	0	5.03216253	3.16901457
Cluster-2	5.03216253	0	3.76581196
Cluster-3	3.16901457	3.76581196	0

Clusters 1 and 2 are relatively well-separated from each other, while cluster 3 not as much

Within-Cluster Dispersion

Data summary (In Original coordinates)

Cluster	#Obs	Average distance in cluster
Cluster-1	12	1748.348058
Cluster-2	3	907.6919822
Cluster-3	7	3625.242085
Overall	22	2230.906692

Clusters 1 and 2 are relatively tight, cluster 3 very loose

Conclusion: Clusters 1 & 2 well defined, not so for cluster 3

Next step: try again with k=2 or k=4

Summary

- Cluster analysis is an exploratory tool. Useful only when it produces meaningful clusters
- Hierarchical clustering gives visual representation of different levels of clustering
  - On other hand, due to non-iterative nature, it can be unstable, can vary highly depending on settings, and is computationally expensive
- Non-hierarchical is computationally cheap and more stable; requires user to set k
- Can use both methods
- Be wary of chance results; data may not have definitive "real" clusters