Clustering & Bootstrapping

Jelena Prokić
University of Groningen
The Netherlands

March 25, 2009
Groningen

Overview

- What is clustering?
- Various clustering algorithms
- Bootstrapping
- Application in dialectometry

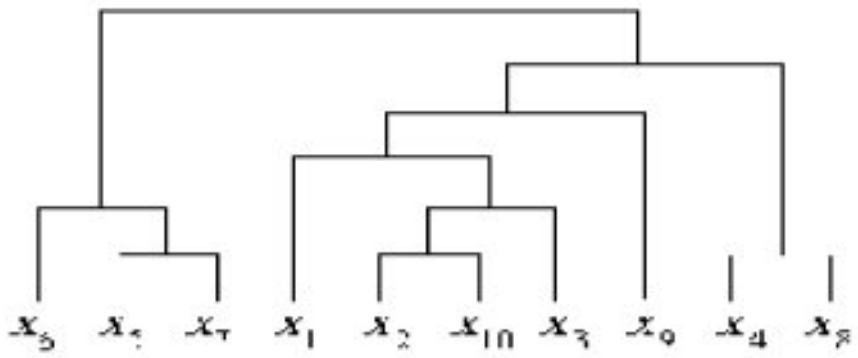
Introduction

- Cluster analysis: study of algorithms and methods for grouping objects
- Objects are classified based on the perceived similarities
- An object is described
  - by a set of measurements or
  - by relationships between the object and other objects
- Clustering algorithms used to find structure in the data

Hierarchical and flat clustering

- Hierarchical clustering:
  - produces a sequence of nested partitions
- Flat clustering:
  - determines a partition of patterns into K initial clusters

Hierarchical and flat clustering (cont.)



X6	X5	X7	X1	X2	X10	X3	X9	X4	X8

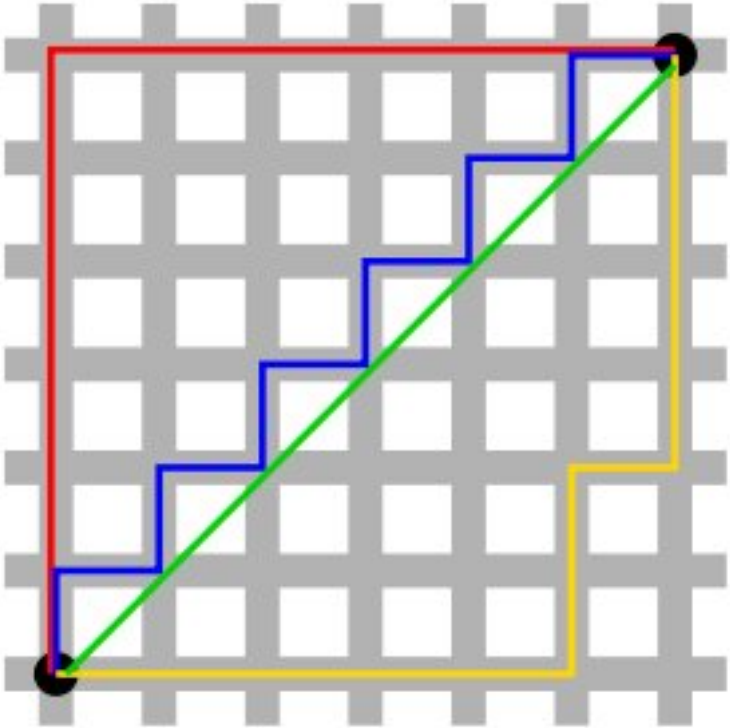
Hard and soft clustering

- Hard clustering:
  - each object is assigned to one and only one cluster
  - hierarchical clustering is usually hard
- Soft clustering:
  - allows degrees of membership and membership in multiple clusters
  - flat clustering can be both hard and soft

Distance measure

- Euclidean distance
  - distance between two points that one would measure with a ruler
  - $d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$
- Manhattan distance
  - the sum of absolute distances between the feature values of two instances
  - $d(p, q) = |p_1 - q_1| + |p_2 - q_2| + ... + |p_n - q_n|$

Euclidean vs Manhattan distance



Hierarchical clustering

- Hierarchical clustering can be top-down and bottom-up
- Top-down
  - starts with one group (all objects belong to one cluster)
  - divides it into groups as to maximize within group similarity
- Bottom-up (agglomerative):
  - starts with separate cluster for each object
  - in each step two most similar clusters are determined and merged into new cluster

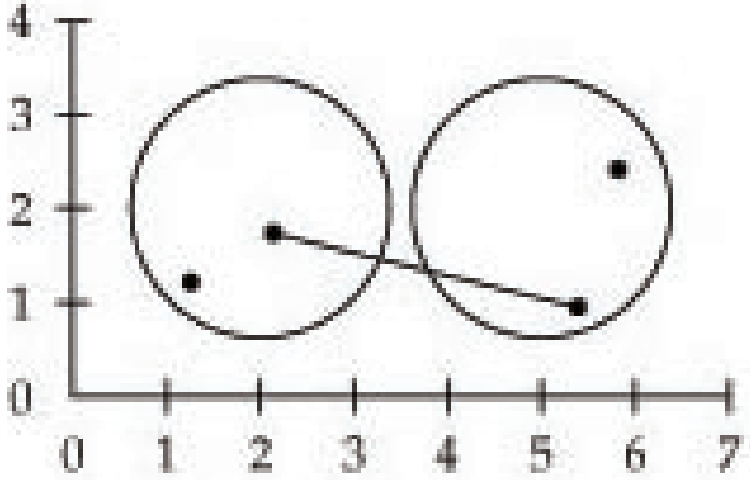
Cluster similarity

- How do we determine the similarity between two clusters?
- Single-link clustering
  - the similarity between two clusters is the similarity of the two closest objects in the clusters
  - checks all pairs of objects that belong to different clusters and selects the pair with greatest similarity
  - produces clusters with good local coherence

Cluster similarity (cont.)

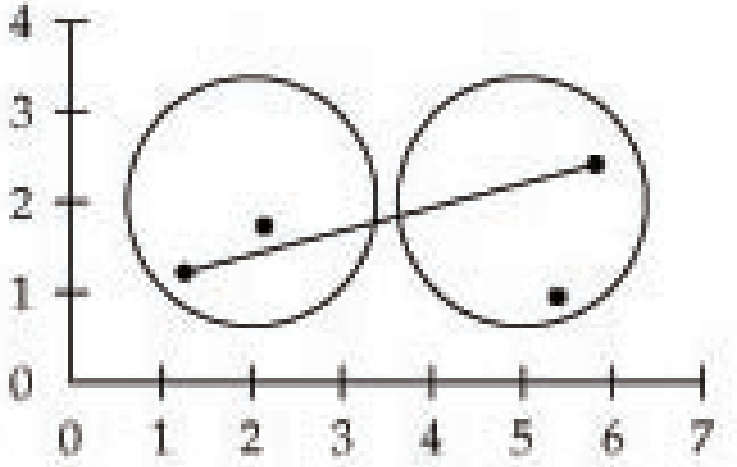
- Complete-link clustering:
  - focuses on global cluster quality
  - the similarity between two clusters is the similarity of the two most dissimilar objects in the clusters
  - merges the two clusters with the smallest maximum pairwise distance
- Group-average agglomerative clustering:
  - in each iteration merges the pair of clusters with the highest cohesion
  - looks for the average similarity between the objects in different clusters

Single link clustering



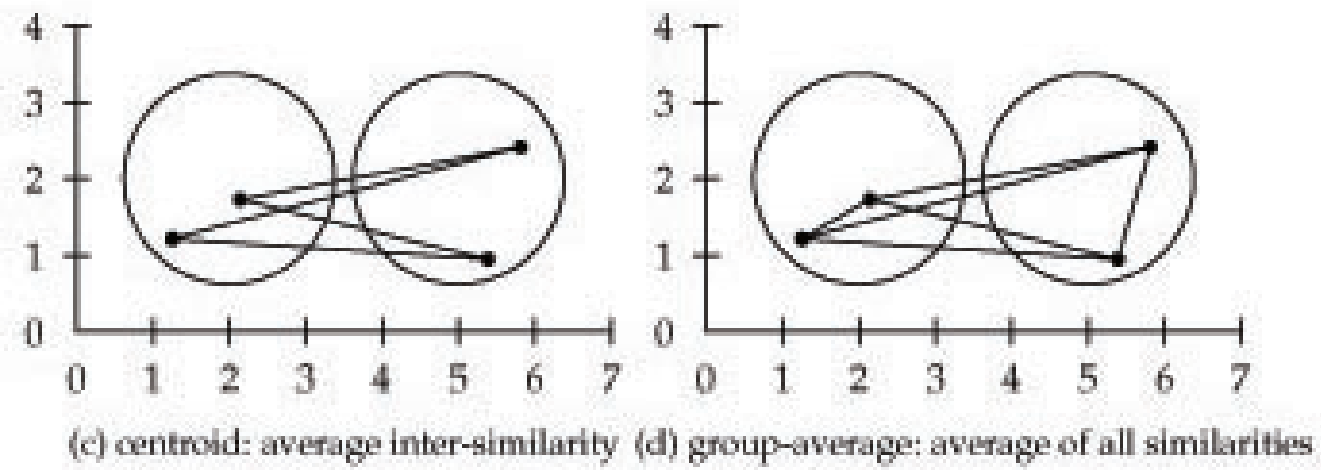
(a) single link: maximum similarity

Complete link clustering



(b) complete link: minimum similarity

Average similarity clustering



(c) centroid: average inter-similarity (d) group-average: average of all similarities

General scheme

- Estimate pairwise distances
- Put information on distances into matrix

	A	B	C	D
A	0	0.00717223	0.003664	0.00628
B		0	0.00299	0.006288
C			0	0.00066
D				0

General scheme (cont.)

- Find the shortest distance in the matrix
- Fuse two closest points
- Calculate the distance between the newly formed node and the rest of the nodes (matrix updating algorithms)
- Repeat until there are no more nodes to be fused

Matrix updating algorithms

- Single link

$$d_{k[ij]} = minimum(d_{ki}, d_{kj})$$

- Complete link

$$d_{k[ij]} = maximum(d_{ki}, d_{kj})$$

- Unweighted Pair Group Method using Arithmetic averages

$$d_{k[ij]} = (n_i/(n_i + n_j)) \times d_{ki} + (n_j/(n_i + n_j)) \times d_{kj}$$

- Weighted Pair Group Method using Arithmetic averages

$$d_{k[ij]} = (\frac{1}{2} \times d_{ki}) + (\frac{1}{2} \times d_{kj})$$

- Unweighted Pair Group Method using Centroids

$$d_{k[ij]} = (n_i/(n_i + n_j)) \times d_{ki} + (n_j/(n_i + n_j)) \times d_{kj} -$$

$$((n_i \times n_j)/(n_i + n_j))^2 \times d_{ij}$$

- Weighted Pair Group Method using Centroids

$$d_{k[ij]} = (\frac{1}{2} \times d_{ki}) + (\frac{1}{2} \times d_{kj}) - (\frac{1}{4} \times d_{ij})$$

- Ward's method

$$d_{k[ij]} = ((n_k+n_i)/(n_k+n_i+n_j))\times d_{ki}+((n_k+n_j)/(n_k+n_i+n_j))\times d_{kj}-$$

$$((n_k/(n_k+n_i+n_j)\times d_{ij}$$

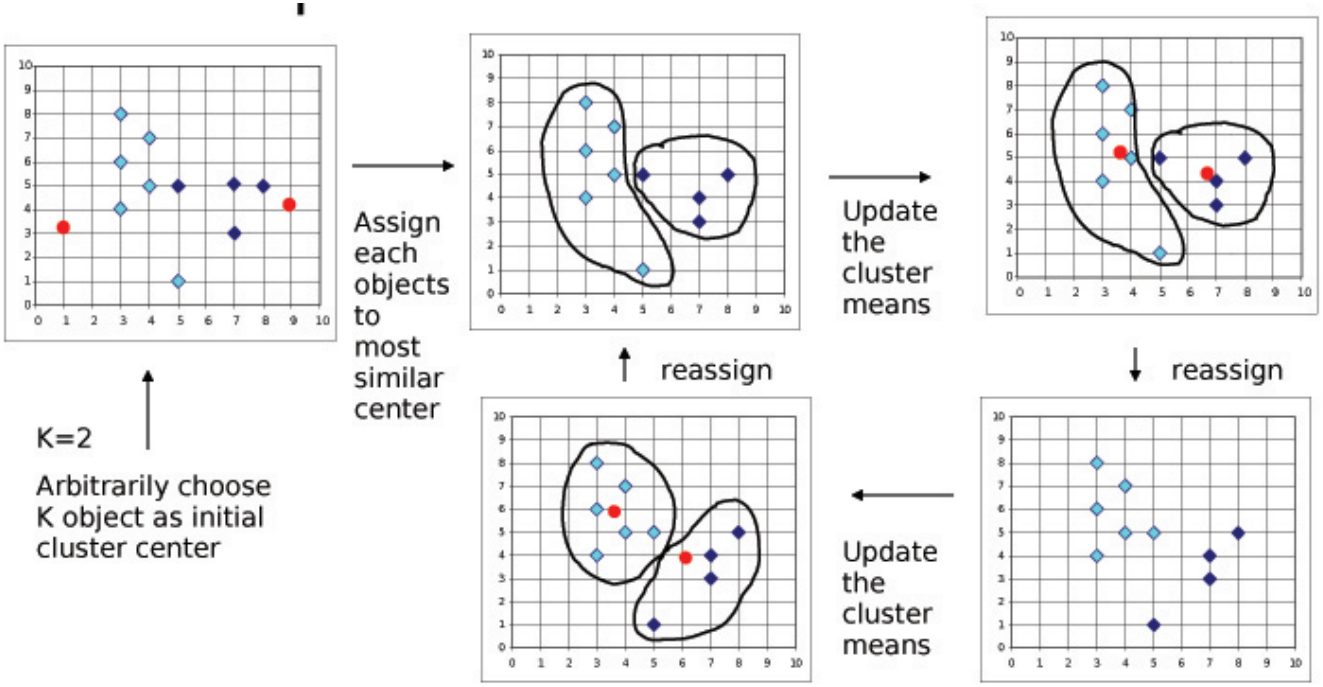
Flat clustering

- Starts with a partition based on randomly selected seeds
- Several passes of reallocating objects to the currently best cluster
- Number of clusters can be given in advance
- More often the optimal number of clusters has to be determined
  - Minimum Description Length
  - measure of goodness: how well the objects fit into the clusters and how many clusters there are

K-means

- Hard clustering algorithm
- Starts by partitioning the input points into k initial sets
- Calculates the mean point, or centroid, of each set
- Constructs a new partition by associating each point with the closest centroid
- Repeats last two steps until the objects no longer switch clusters

K-means (cont.)


K=2
Arbitrarily choose K object as initial cluster center
Assign each objects to most similar center
reassign
Update the cluster means
reassign
Update the cluster means

Problems

- There is no one best clustering algorithm
  - every algorithm has its own bias
- The success depends on the data set it is used on
- Small differences in input can lead to substantial differences in output

Traditional division of sites

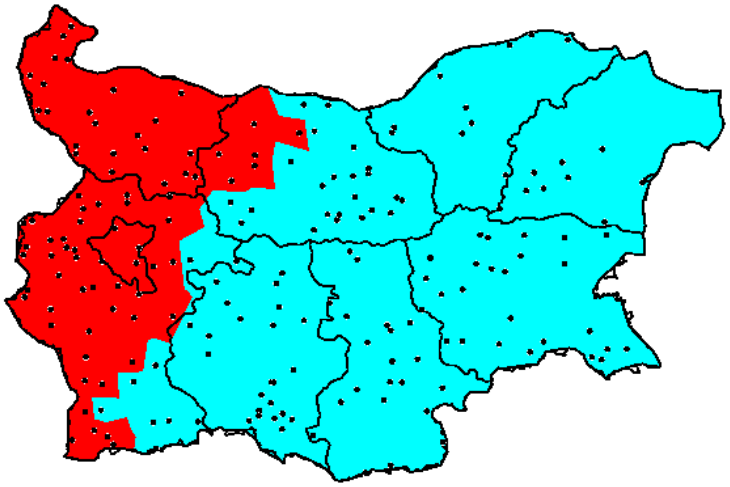


Figure 1: Two-fold division

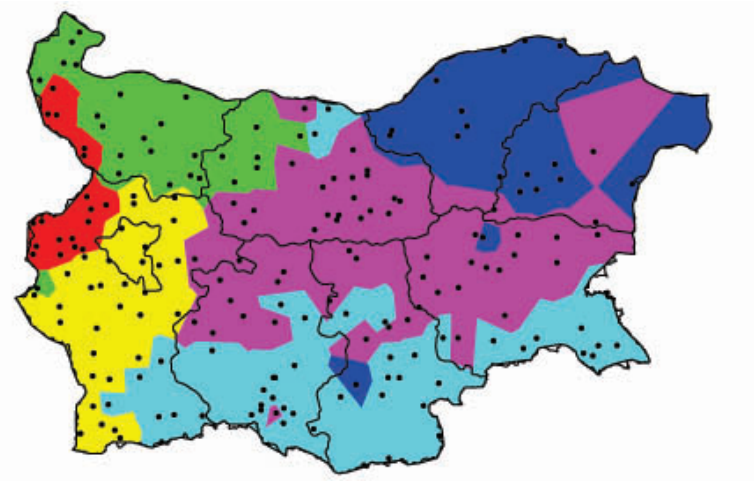
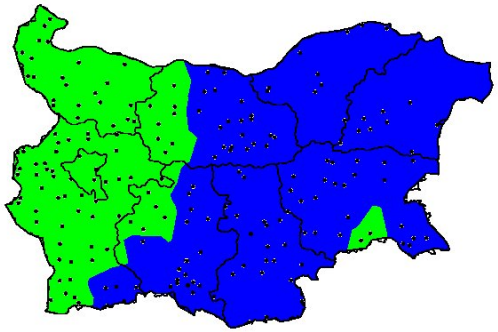
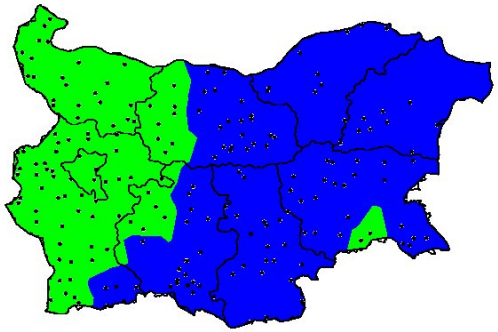


Figure 2: Six-fold division

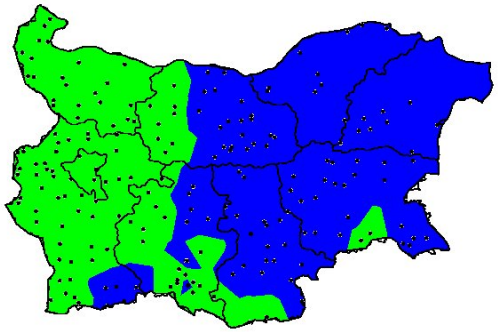
Two-fold division of sites



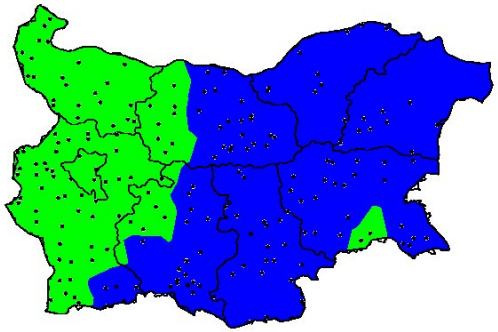
UPGMA



WPGMA

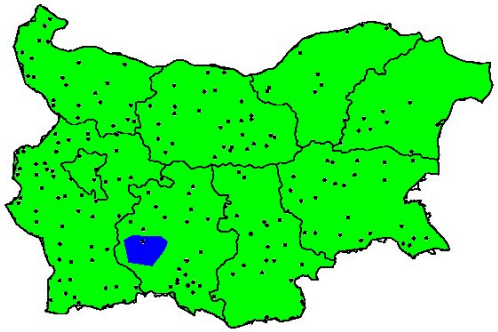


Complete link

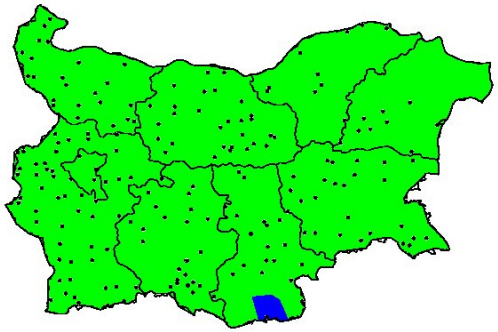


Ward's method

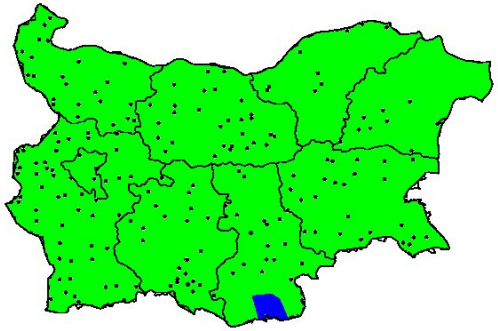
Two-fold division of sites (cont.)



Single link

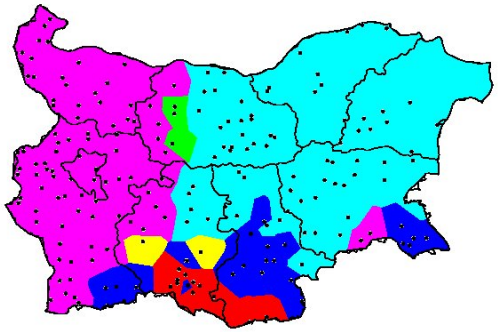


UPGMC

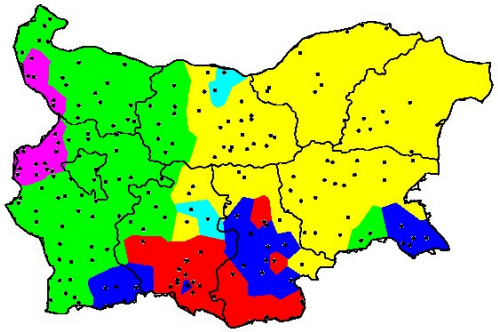


WPGMC

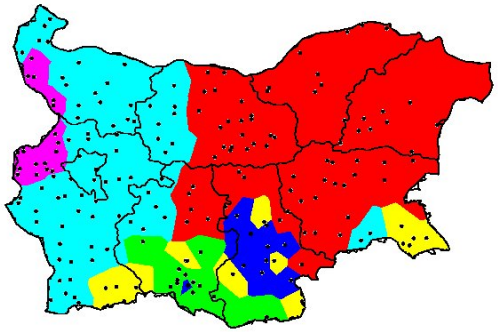
Six-fold division of sites



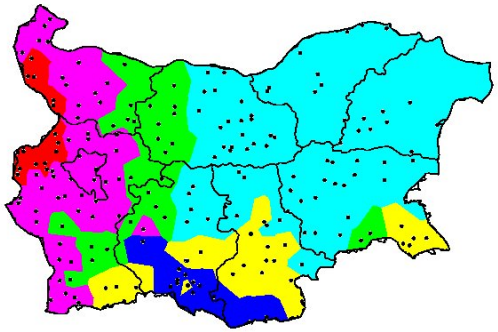
UPGMA



WPGMA

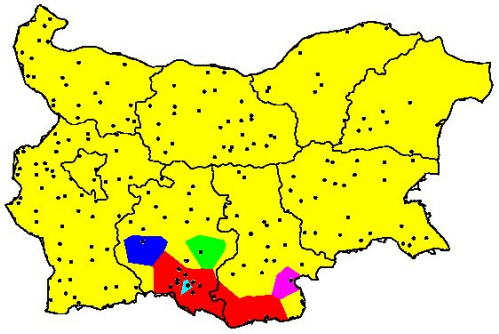


Complete link

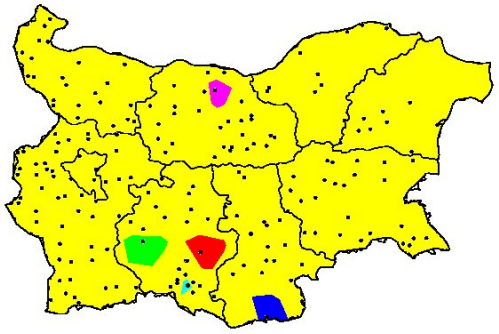


Ward's method

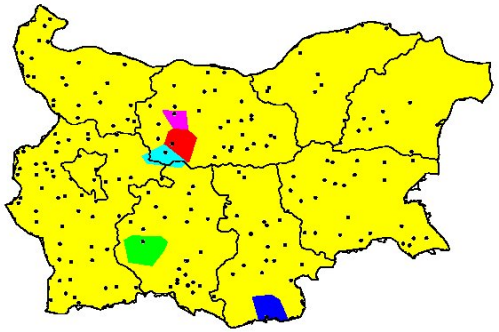
Six-fold division of sites (cont.)



Single link



UPGMC



WPGMC

K-means

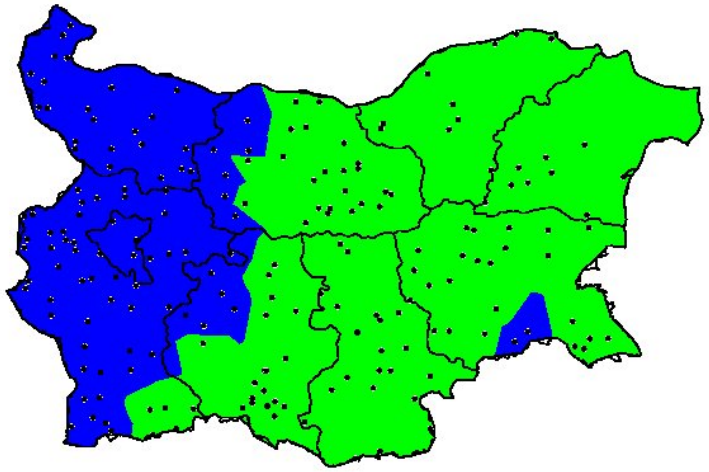


Figure 3: Two-fold division

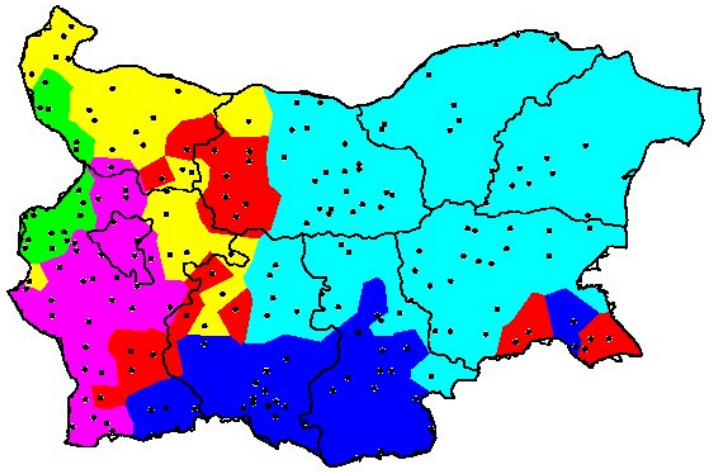


Figure 4: Six-fold division

Jackknife and bootstrapping

- Two general-purpose techniques for empirically estimating the variability of an estimate
- Jackknife: involves dropping one observation at a time from one's sample and calculating the estimate each time
- Bootstrapping: involves resampling from one's sample with replacement and making the fictional sample of the same size
- Set us free from the need for Normal data and large samples

Jackknife

- Compute the desired sample statistics St based upon the complete sample (of size n)

- Compute the corresponding statistics St_{-i} based upon the sample data with each of the observations i ignored in turn

- Compute the so-called pseudo values ϕ_i as follows:

$$\phi_i = nSt - (n-1)St_{-i}$$

Jackknife

- The jackknifed estimate of the statistics is:

$$\widehat{St} = \frac{\sum \phi_i}{n} = \overline{\phi}$$

- The approximate standard error of $\widehat{St}$ is:

$$s_{\widehat{St}} = \sqrt{\frac{s_\phi^2}{n}} = \sqrt{\frac{\sum(\phi_i - \overline{\phi})^2}{n(n-1)}}$$

Bootstrapping

- Related technique for obtaining standard errors and confidence limits
- Set of observations is from independent and identically distributed population

Step 1: Resampling

- In place of many samples from the population, create many resamples
- Each resample is obtained by random sampling with replacement from the original data set
- Each resample is the same size as the original random sample
- Sampling with replacement: after we randomly draw an observation from the original sample we put it back before drawing the next observation

Resampling idea

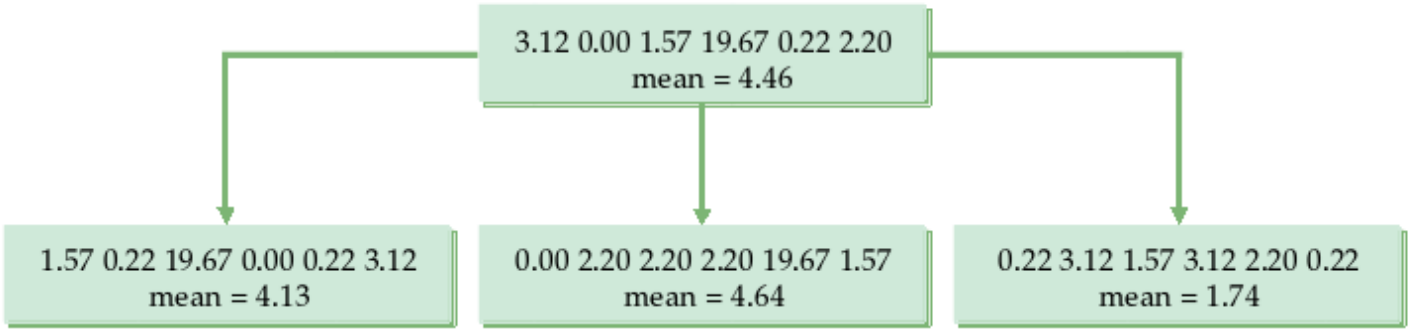
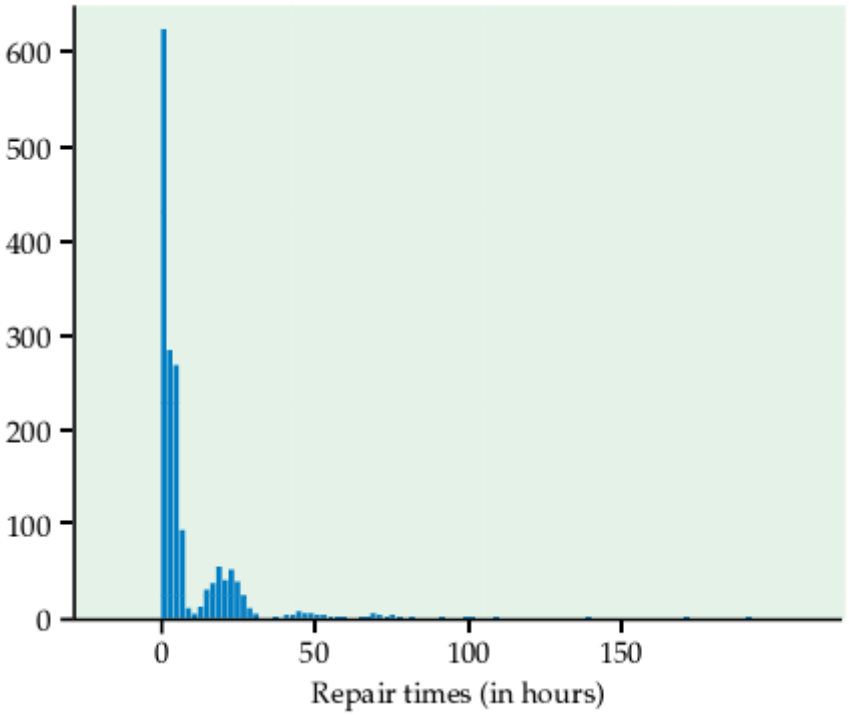


FIGURE 16.2 The resampling idea. The top box is a sample of size $n = 6$ from the Verizon data. The three lower boxes are three resamples from this original sample. Some values from the original are repeated in the resamples because each resample is formed by sampling with replacement. We calculate the statistic of interest—the sample mean in this example—for the original sample and each resample.

Step 2: Bootstrap distribution

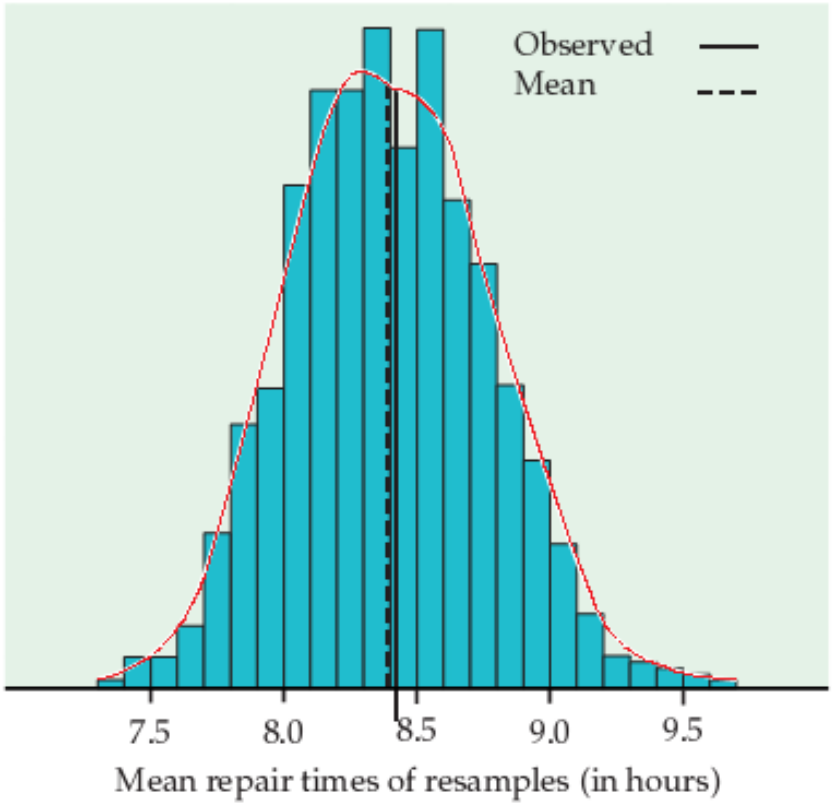
- The bootstrap distribution of a statistic collects its values from the many resamples.
- The bootstrap distribution gives information about the sampling distribution.
- Statistically bootstrapped data sets contain variation that you would get from collecting new data sets.

Random sample distribution



- random sample
- 1644 telephone repair times
- mean: 8,41 hours

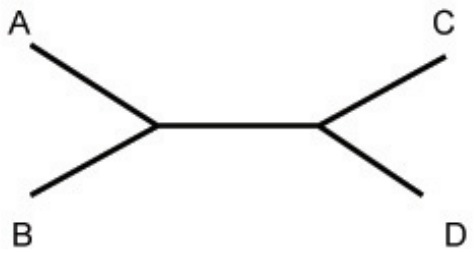
Bootstrap distribution



- nearly Normal distribution
- we get the distribution of the estimator
- we get statistics of the estimator
- bootstrap standard error: 0.367
- theory based estimate: 0.360

Bootstrapping in phylogenetics

		1	2	3	4	5	6	7
a		A	T	A	T	A	A	A
b		A	T	T	A	T	A	A
c		T	A	A	A	A	T	A
d		T	A	T	A	A	A	T



Bootstrapping in phylogenetics

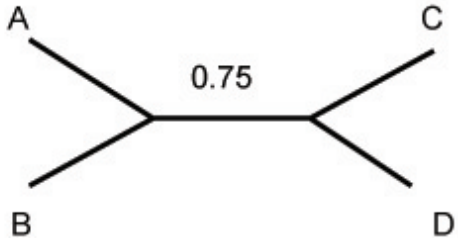
		1	2	3	4	5	6	7
a		A	T	A	T	A	A	A
b		A	T	T	A	T	A	A
c		T	A	A	A	A	T	A
d		T	A	T	A	A	A	T

		1	2	2	4	5	6	7
a		A	T	T	T	A	A	A
b		A	T	T	A	T	A	A
c		T	A	A	A	A	T	A
d		T	A	A	A	A	A	T

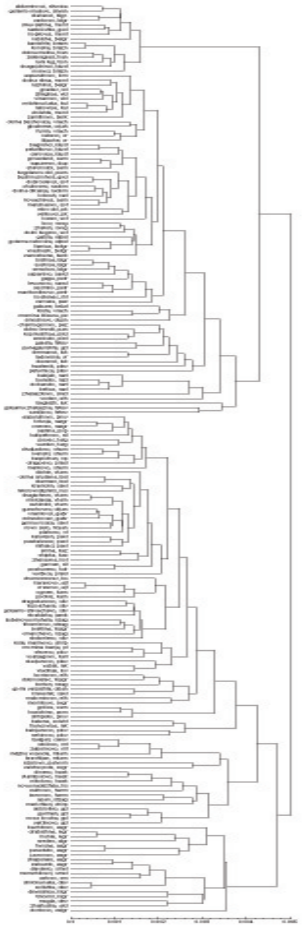
		1	3	3	4	5	6	7
a		A	A	A	T	A	A	A
b		A	T	T	A	T	A	A
c		T	A	A	A	A	T	A
d		T	T	T	A	A	A	T

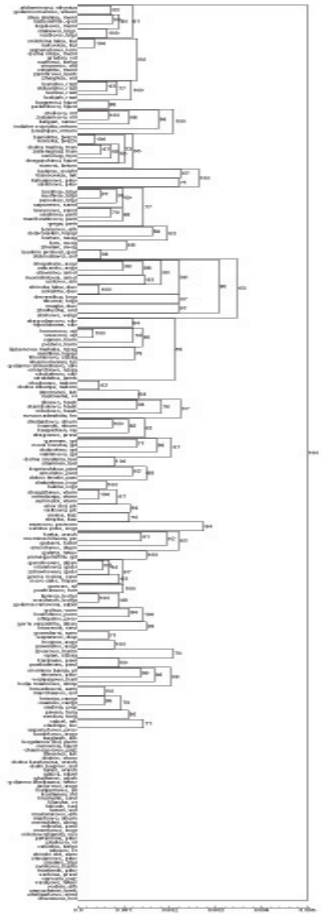
		1	2	3	4	5	6	7
a		A	T	A	T	A	A	A
b		A	T	T	A	T	A	A
c		T	A	A	A	A	T	A
d		T	A	T	A	A	A	T

		1	2	4	4	5	6	7
a		A	T	T	T	A	A	A
b		A	T	A	A	T	A	A
c		T	A	A	A	A	T	A
d		T	A	A	A	A	A	T



Bootstrapping in dialectometry





References

- Anil K. Jain and Richard C. Dubes (1988). *Algorithms for Clustering Data*. Prentice Hall: New Yersey.

- David S. Moore and George McCabe (1993). *Introduction to the Practice of Statistics*. 5th edition. Freeman: New York.

- Robert R. Sokal and F. James Rohlf (1995). *Biometry. The Principles and Practices of Statistics in Biological Research*. 3rd edition. Freeman: New York.