Cluster Analysis

Martin Boroš
Methodology and Statistics
University of Groningen

May 24, 2011

Outline

- What is cluster analysis?
- How does it work?
- Data
- Application on data
- Validation and Interpretation of results
    - Average silhouette width

What is Cluster Analysis?

- Set of methods for grouping or classifying objects
  - maximalization of within group similarity
  - minimalization of between group similarity
  - finding structure in data

- Main approaches
  - Hierarchical algorithms
    - clustering from previously established clusters
      - Sequence of nested clusters
    - agglomerative ("bottom-up") or divisive ("top-down")
  - Partitional algorithms
    - typically determine all clusters at once

How does it work?

- 1. Generating similarity (distance) matrix
  - depends on information value and nature of the variables describing the objects to be clustered
- 2. Choosing the linkage criteria
  - Single-linkage clustering
    - the distance between two clusters is computed as the distance between the two closest elements in the two clusters
    - produces clusters with good local coherence
  - Complete-linkage clustering
    - the distance between two clusters is computed as the maximum distance between a pair of objects, one in one cluster, and one in the other
    - focuses on global cluster quality

How does it work?

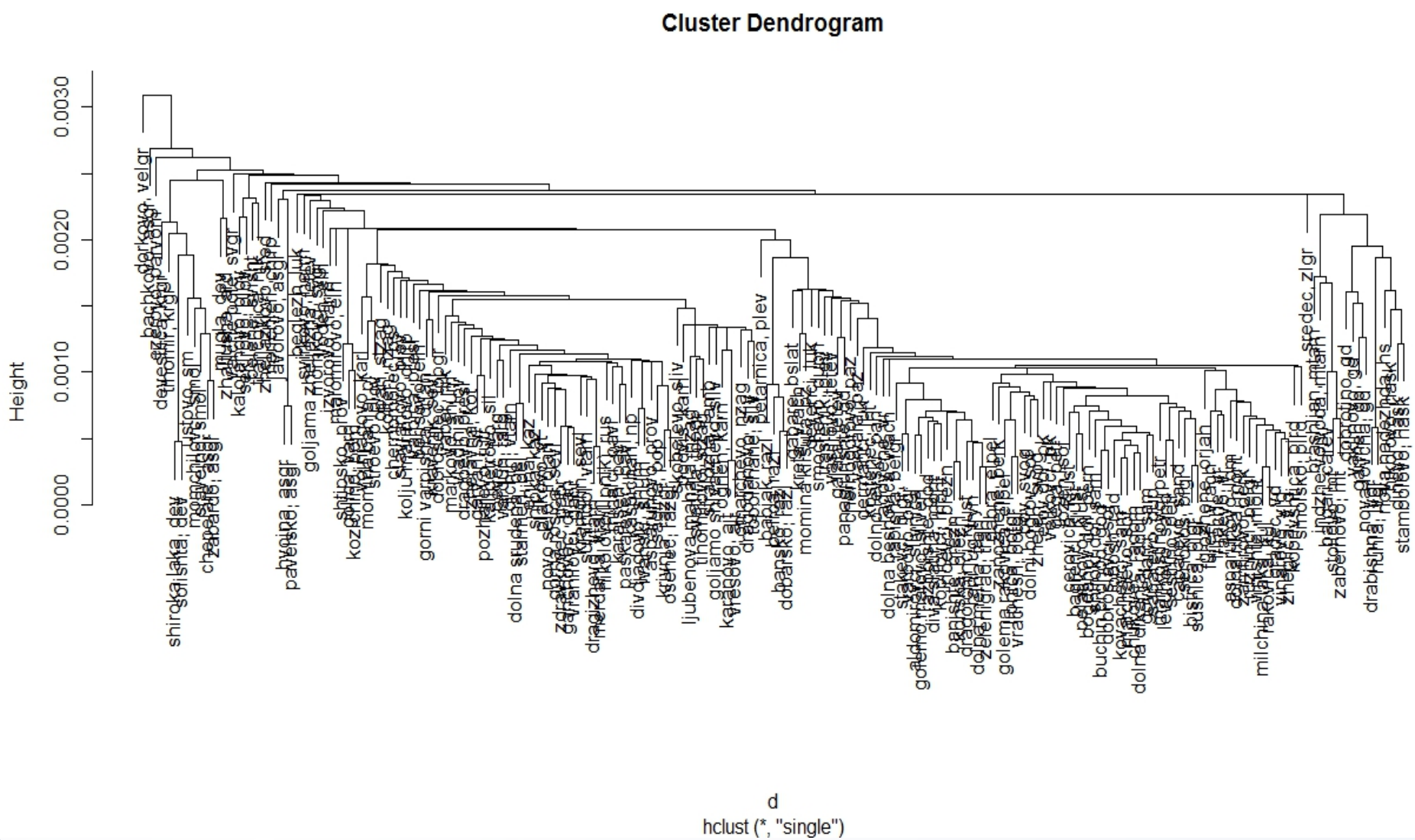
- 2. Choosing the linkage criteria
  - Average linkage clustering
    - looks for the average similarity between the objects in different clusters
    - creates clusters with similar variances
  - Ward's method
    - minimize information loss associated with grouping
    - creates small and even sized clusters
    - at each step, considers union of every possible cluster pair
      - merge those two elements, whose merging least increases their sums of squad difference from the mean
- 3. Interpreting the results
- 4. Validating the results

The data

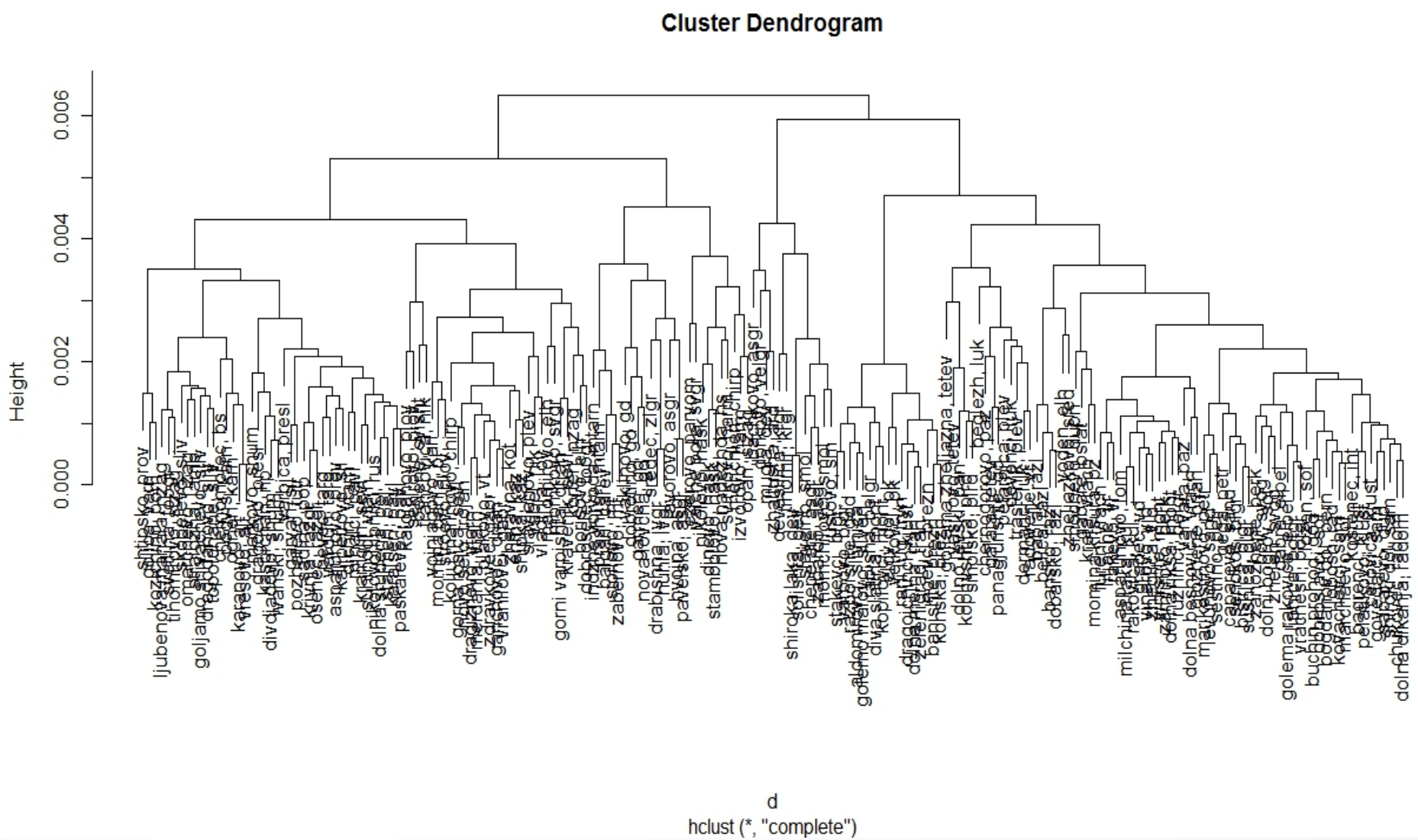
- Archive of the Ideographic Dialect Dictionary of Bulgarian
  - phonetic transcriptions of words
  - collected from 197 sites all over Bulgaria

- Our distance matrix
  - transcriptions of 156 words
  - Levensthein distance between strings (words)
  - **site-to-site distance**
    - mean of all word distances calculated for those two sites

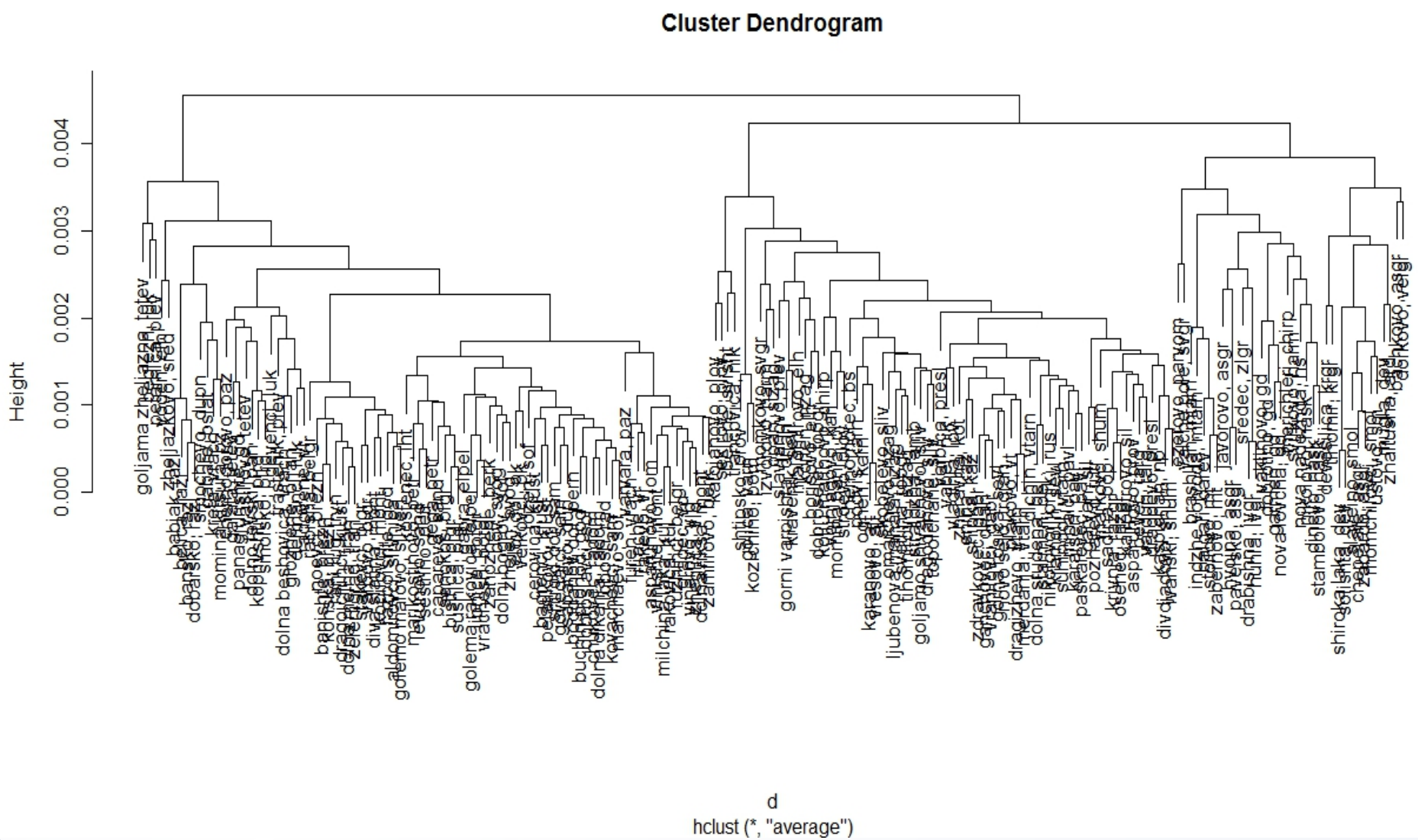
Hierarchical agglomerative single linkage clustering



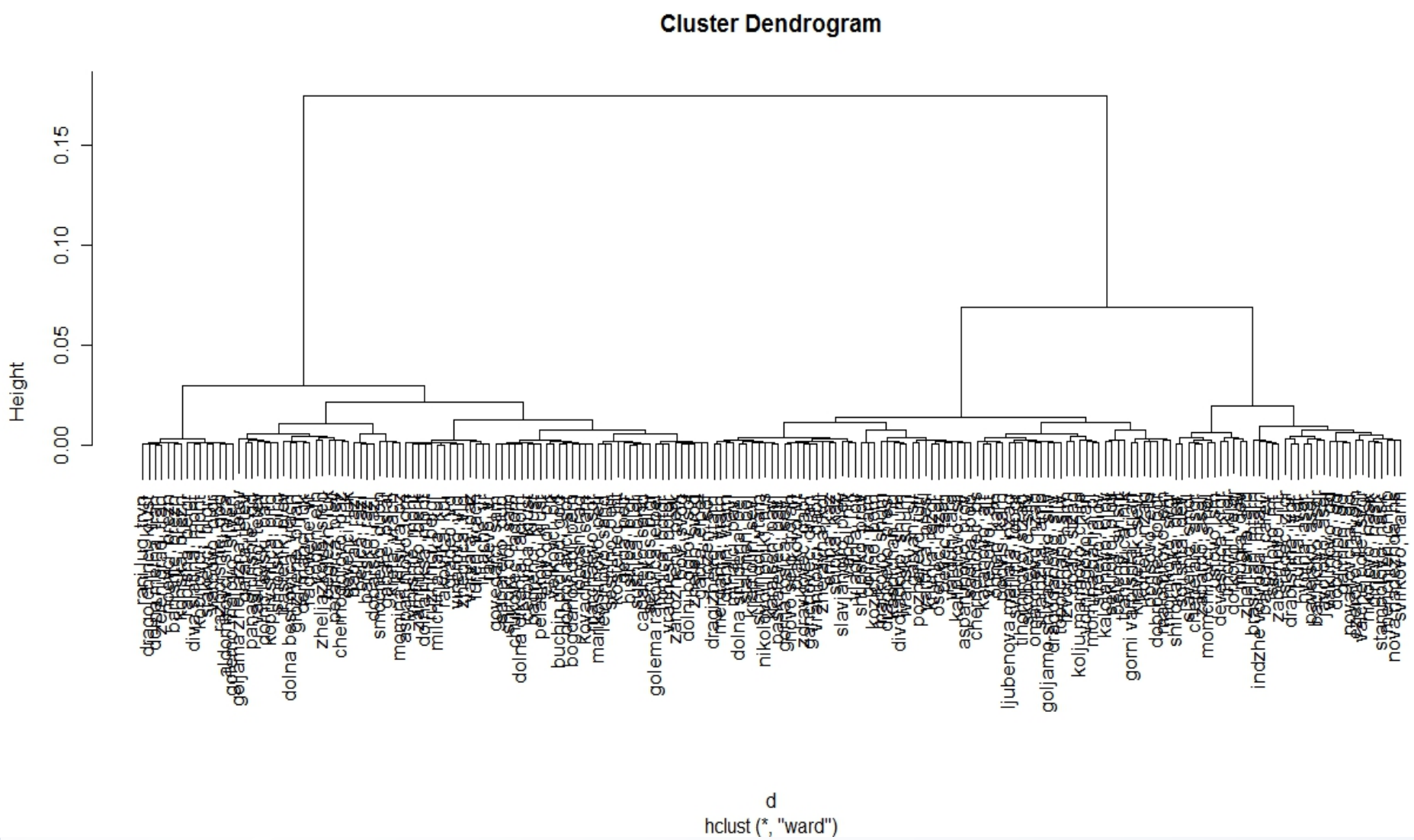
Hierarchical agglomerative complete linkage clustering



Hierarchical agglomerative average linkage clustering



Hierarchical agglomerative Ward linkage clustering



Interpreting the results

- What is the optimal number of clusters?

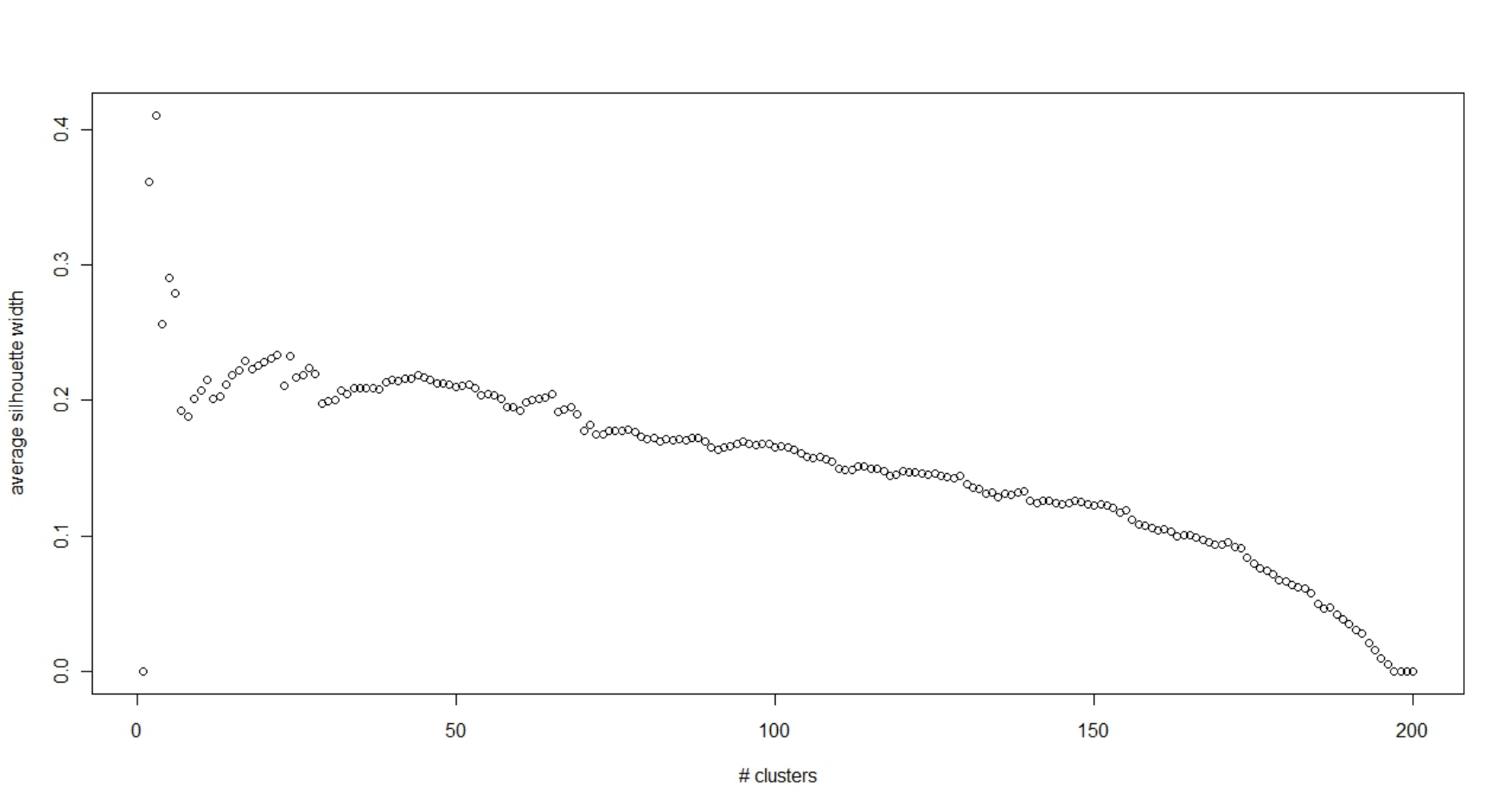
- Silhouette width
  - way of measuring the strength of clusters
  - or how well one element was clustered

  - $SW_i = (b_i - a_i) / \max(a_i, b_i)$
  - Where a is the average distance from point a_i i to all other points in i's cluster, and b_i is is the minimum average distance from point i to all points in another cluster
  - $-1 < SW_i < 1$

Average silhouette width

- Optimal ratio
  - maximize inter-clusters distance
  - minimize intra-clusters distance
- Measures global goodness of clustering
  - ASW = ($\sum$i SWi) / n
  - 0 < ASW < 1
  - the larger ASW the better the split
- Interpretation
  - 0.71 – 1.00 excellent split
  - 0.51 – 0.70 reasonable structure has been found
  - 0.26 – 0.50 weak structure, could be artificial
  - ≤0.25 horrible split

Example: Average silhouette width, determining number of clusters in K-means clustering



Maximum value is for 3 clusters

Validation techniques

- Monte Carlo
  - uses random number generators to generate data sets with general characteristics matching the overall characteristics of original data
  - same clustering methods are applied
  - results are compared
- Replication
  - split up your data set into random subsamples and apply the same methodologies
  - if a cluster solution is repeatedly discovered across different sample from the same population, then it is plausible to conclude that this solution has some generality

Closing remarks

- Cluster analysis can be used for
  - development of a typology
  - finding a structure in data
- Most methods are simple procedures
  - different methods – different solutions
- Strategy of clustering is structure-seeking, althought the operations are structure-imposing
- Different methods and approaches are suitable for different tasks and data

References

- Keith Johnson (2008): Quantitative Methods in Linguistics. Wiley-Blackwell

- Peter Houtzagers, Jonh Nerbonne and Jelena Prokić (2010): Quantitative and traditional classifications of Bulgarian dialects compared. Scando-Slavica 59(2), pp.163-188.

- Other presentations
  - Daniel Wiechmann (2008): Cluster Analysis
  - Jelena Prokić (2009): Clustering & Bootstrapping