



RUG

Hierarchical agglomerative **Cluster Analysis**

Christine Siedle

19-3-2004



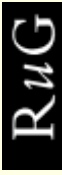
Classification

- Basic (unconscious & conscious) human strategy to reduce complexity
- Always biased
 - Cluster analysis
 - to find or confirm types in data
 - to uncover relations between objects
 - The more entities and the more attributes – the more difficulties classifying them manually
 - Computer-based cluster analysis



Cluster analysis – overview

- Selection of objects to be classified
- Selection of relevant attributes of these objects
- Calculation of distances between objects
- Cluster analysis
- Check of results
- (Modifications + rerun analysis)



Objects

- Selection of objects depends on intention
- If clusters are expected:
 - Number of objects should be balanced
- Many objects = large distance matrix
 - $\frac{n \times (n-1)}{2}$ values (e.g. 200 objects = 19900 distance values)



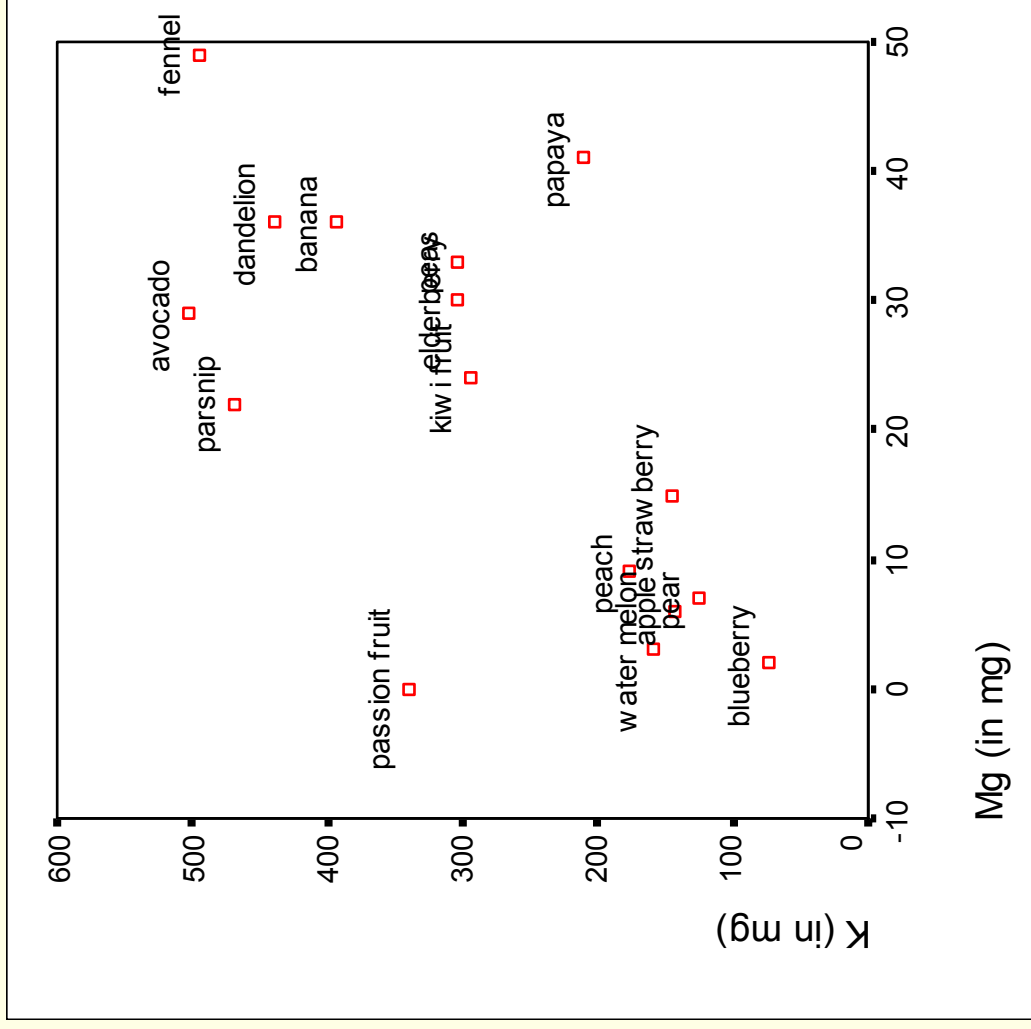
Attributes

- Selection of attributes depends on intention
- *Not:* The more attributes the surer groups will appear
- Avoid correlations between attributes
- Values of attributes have to be comparable
- Treat missing values
- (Weight attributes to influence clustering)



Attributes – example

Position of selected fruits/vegetables in the 2 dimensions of magnesium & potassium





Distance measures

- Based on the attribute values the distances between the objects have to be determined.
- Distance measures have to ensure:
 - Symmetry
 - Triangle inequality
 - Distinguishability of nonidenticals
 - Indistinguishability of identicals

$$d(x, y) = d(y, x) \geq 0$$

$$d(x, y) \leq d(x, z) + d(y, z)$$

if $d(x, y) \neq 0$, then $x \neq y$

$$d(x, x') = 0$$



Distance measures – examples

- Distance measures

- (squared) Euclidian distance
- Manhattan distance

$$\delta(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$

$$\delta(X, Y) = \sum_{i=1}^n |X_i - Y_i|$$

- Similarity measures

- Pearson's correlation coefficient

$$r(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$



RUG

Squared Euclidian distance – example

Distances of selected fruits/vegetables based on (standardized) content of Mg & K

Proximity Matrix

Case	Squared Euclidean Distance			
	1:banana	2:avocado	3:parsnip	4:dandelion
1:banana	,000	1,250	1,477	,183
2:avocado	1,250	,000	,346	,578
3:parsnip	1,477	,346	,000	1,070
4:dandelion	,183	,578	1,070	,000

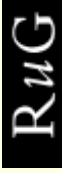
This is a dissimilarity matrix



Cluster analysis

- Here discussed (because most common):
 - Sequential **Agglomerative Hierarchical Nonoverlapping** (SAHN)

- Other approaches for clustering:
 - Hierarchic divisive
 - Iterative partitioning
 - Factor analytic
 - Clumping
 - ...



Cluster analysis

- Iterative process
- $n - 1$ steps necessary to cluster all objects
- At every step the two most similar objects or clusters will be merged until all are aggregated in one cluster



Cluster analysis – example

	banana	avocado	parsnip	dandelion
banana		1.25	1.477	0.183
avocado			0.346	0.578
parsnip				1.07
dandelion				

$$d_{avocado[banana, dandelion]} = \frac{d_{avocado, banana}}{2} + \frac{d_{avocado, dandelion}}{2}$$

$$d_{avocado[banana, dandelion]} = \frac{1.25}{2} + \frac{0.578}{2} = 0.914$$



Cluster analysis – example



	banana-dandelion	avocado	parsnip		banana-dandelion	avocado-parsnip
banana-dandelion		0.914	1.2735	banana-dandelion		1.09375
avocado			0.346	avocado-parsnip		
parsnip						

$$d_{[banana, dandelion][avocado, parsnip]} = \frac{d_{[banana, dandelion], avocado}}{2} + \frac{d_{[banana, dandelion], parsnip}}{2}$$

$$d_{[banana, dandelion][avocado, parsnip]} = \frac{0.914}{2} + \frac{1.2735}{2} = 1.09375$$



Matrix updating algorithms

- Several SAHN clustering algorithms
- They differ in how they calculate the distances of new formed clusters to the other elements.
- Not every algorithm equally suitable for every situation
 - **Results can be very different!**



Matrix updating algorithms

- Single linkage
- Complete linkage
- Unweighted average linkage
- Weighted average linkage
- (Un)Weighted centroid linkage
- Ward's method

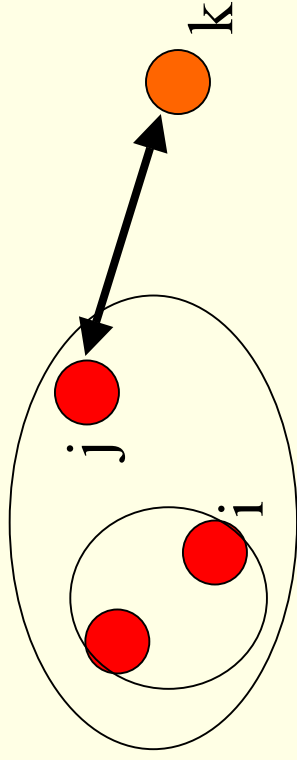


RUG

Single linkage

$$d_{k(ij)} = \min(d_{ki}, d_{kj})$$

- Nearest neighbor
- Distance between new cluster and other elements equals the *smallest* in the cluster occurring distance to the other elements
- Tendency to very different sized clusters (outliers!)



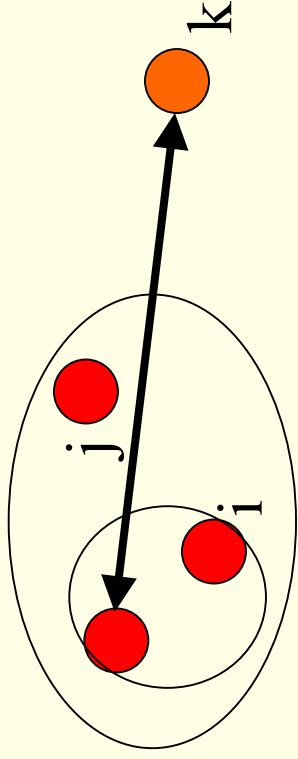


RUG

Complete linkage

$$d_{k(ij)} = \max(d_{ki}, d_{kj})$$

- Furthest neighbor
- Distance between new cluster and other elements equals the *largest* in the cluster occurring distance
- Clusters are only merged when dissimilarity is small.
 - Balanced and equally sized clusters

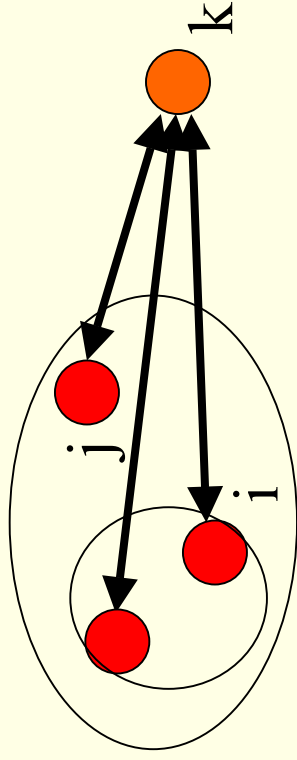




Unweighted average linkage

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

- UPGMA, Baverage, linkage *between* groups
- Uses averages instead of extreme values
- Number of elements in clusters is taken into account

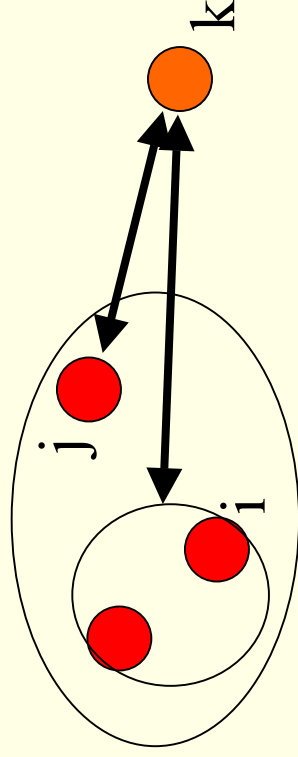




Weighted average linkage

$$d_{k[ij]} = \frac{d_{ki}}{2} + \frac{d_{kj}}{2}$$

- WPGMA, Waverage, linkage *within* groups
- Equals UPGMA but the number of elements in clusters is *not* taken into account
- Can be necessary when the size of supposed clusters or the object density in them differs



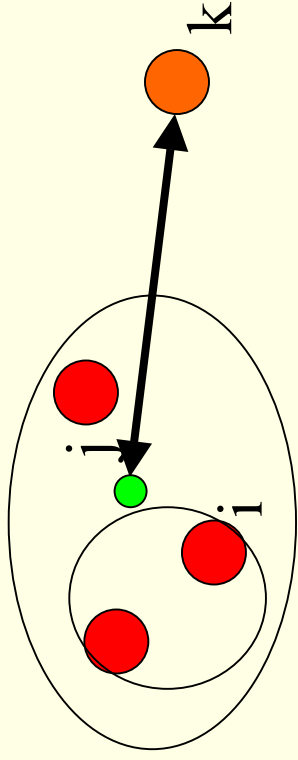


(Un)Weighted centroid linkage

$$d_{k[ij]} = \frac{n_i}{n_i + n_j} \times d_{ki} + \frac{n_j}{n_i + n_j} \times d_{kj} - \frac{n_i \times n_j}{(n_i + n_j)^2} \times d_{ij}$$

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

- Centroid of cluster is calculated
- Distance to new cluster equals distance to centroid





Ward's method

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

- Minimum variance
- Idea: Heterogeneity is not a reasonable feature of clusters
 - Minimize variance
 - To be used only with quantitative attributes and squared Euclidian distance!

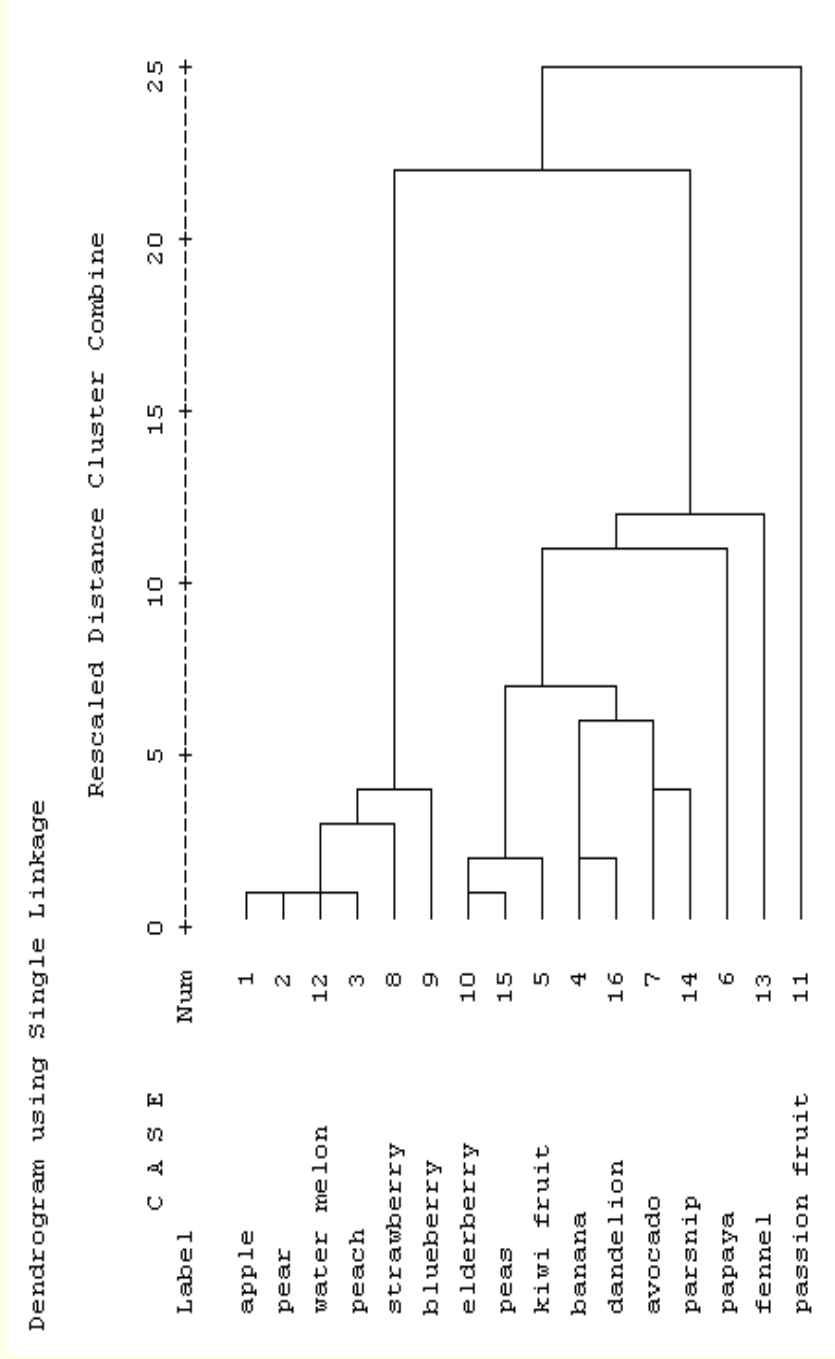


Matrix updating algorithms

- Types of algorithms:
 - Space-contracting (Single & Centroid (?) Linkage)
 - Unequally sized clusters
 - Outliers visible
 - Space-dilating (Complete linkage & Ward's method)
 - Balanced clustering
 - Clusters are often not easy to interpret
 - Space-conserving (Average linkage)
 - No unnaturally blown up clusters
 - Appearing clusters are often interpretable



Space-contracting – example 1

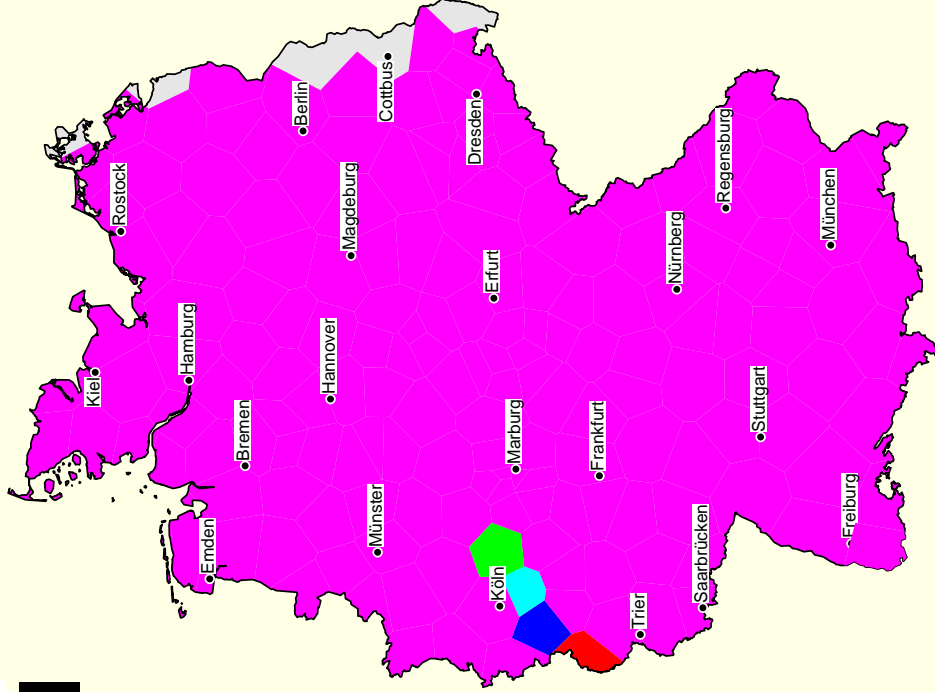


Dendrogram generated by Single-linkage

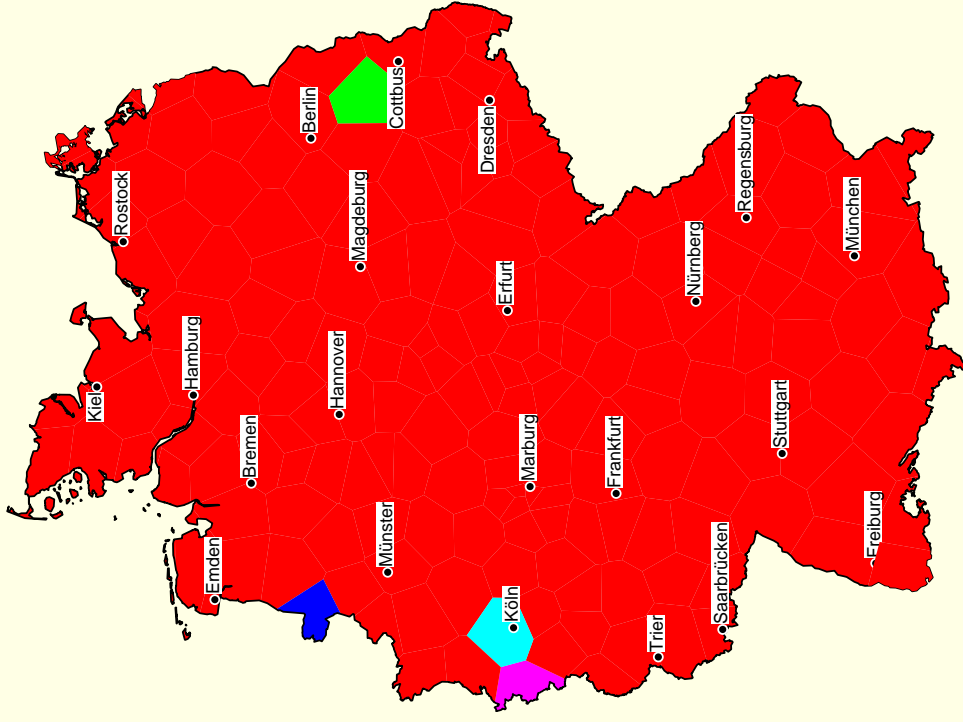


RUG

Space-contracting – example 2



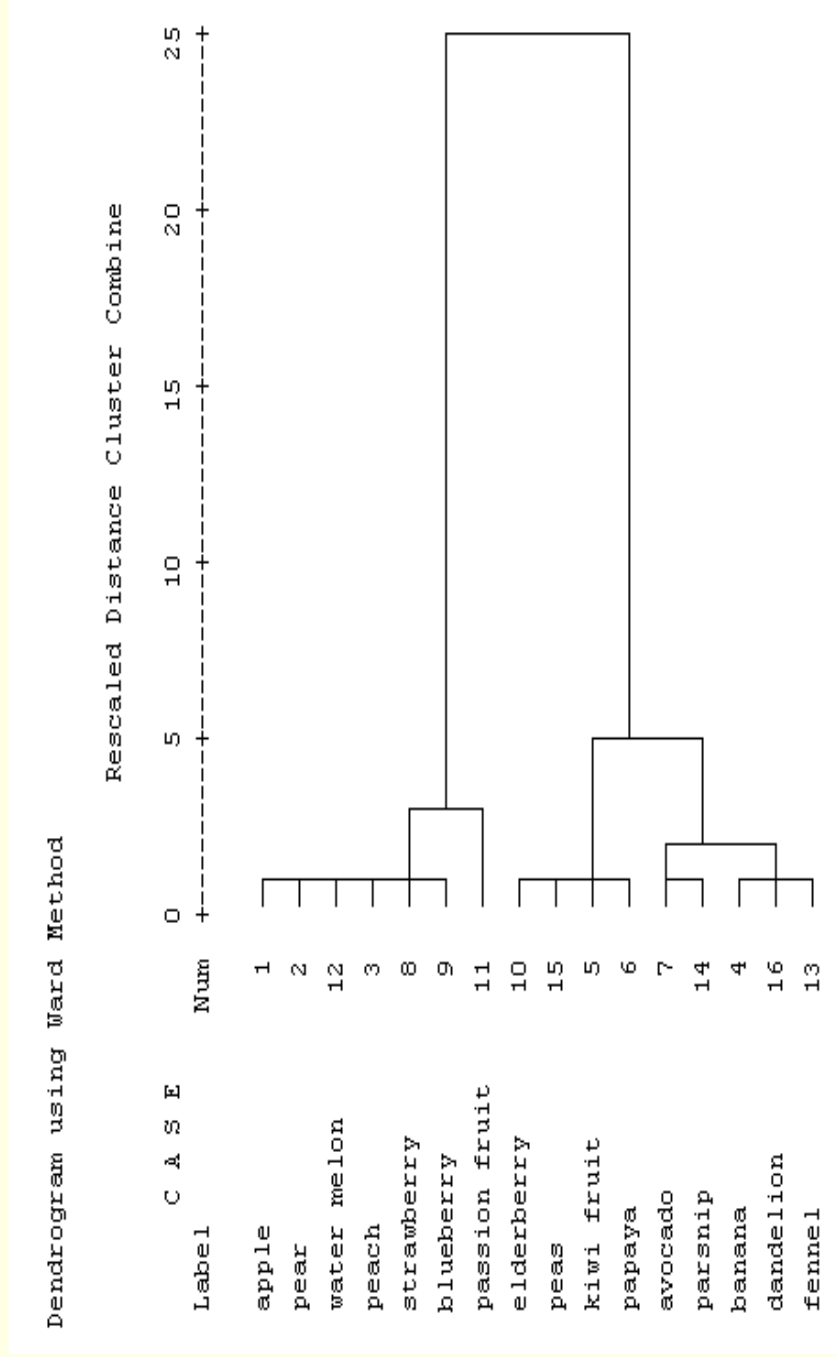
Single linkage



WPGMC

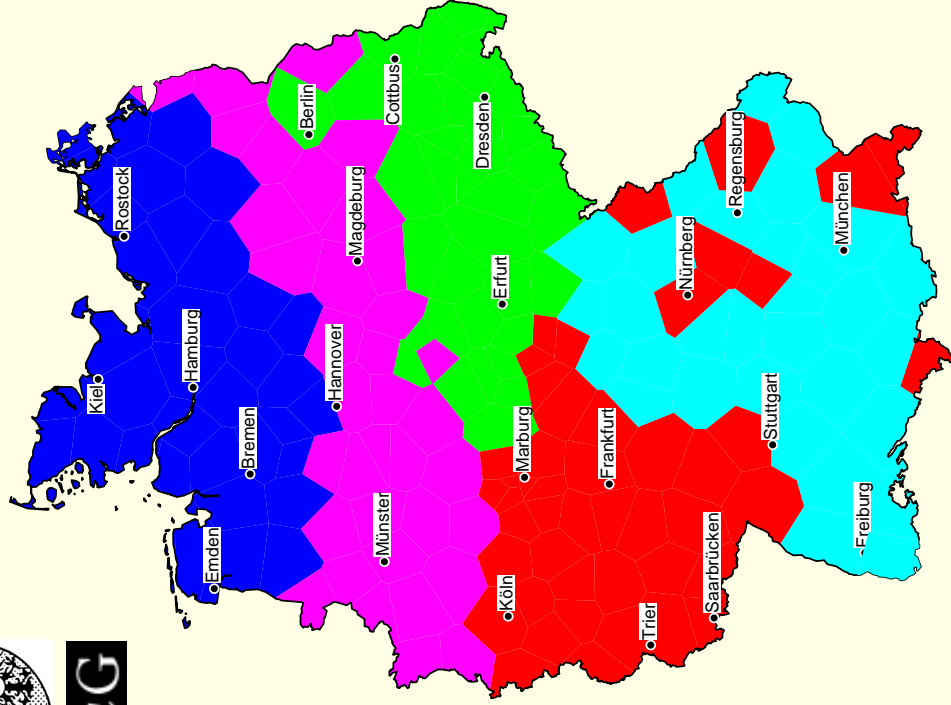


Space-dilating – example 1

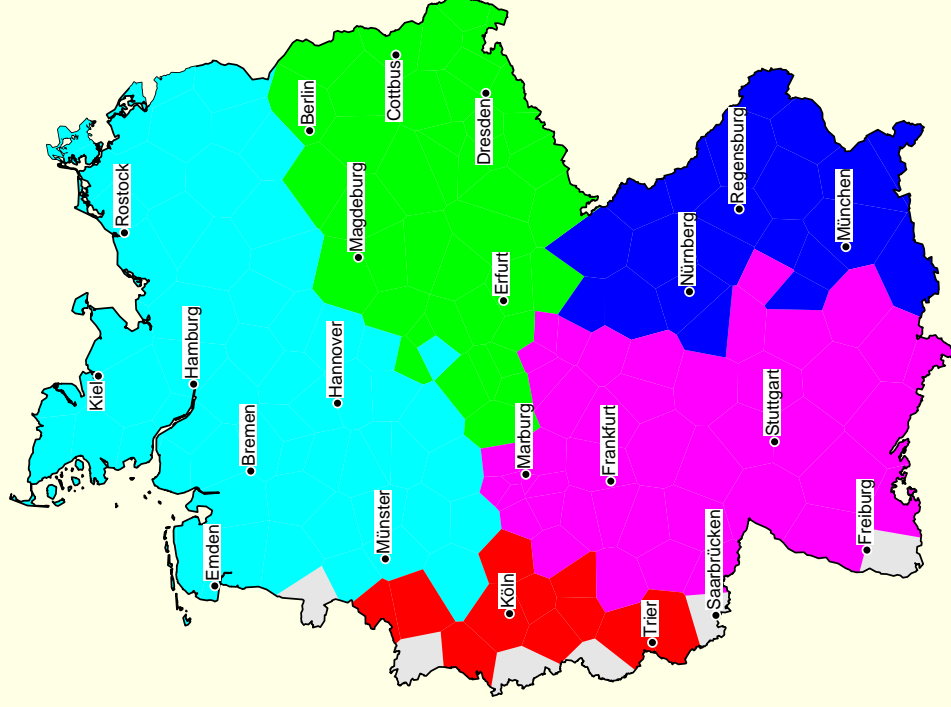


Dendrogram generated by Ward’s method

Space-dilating – example 2



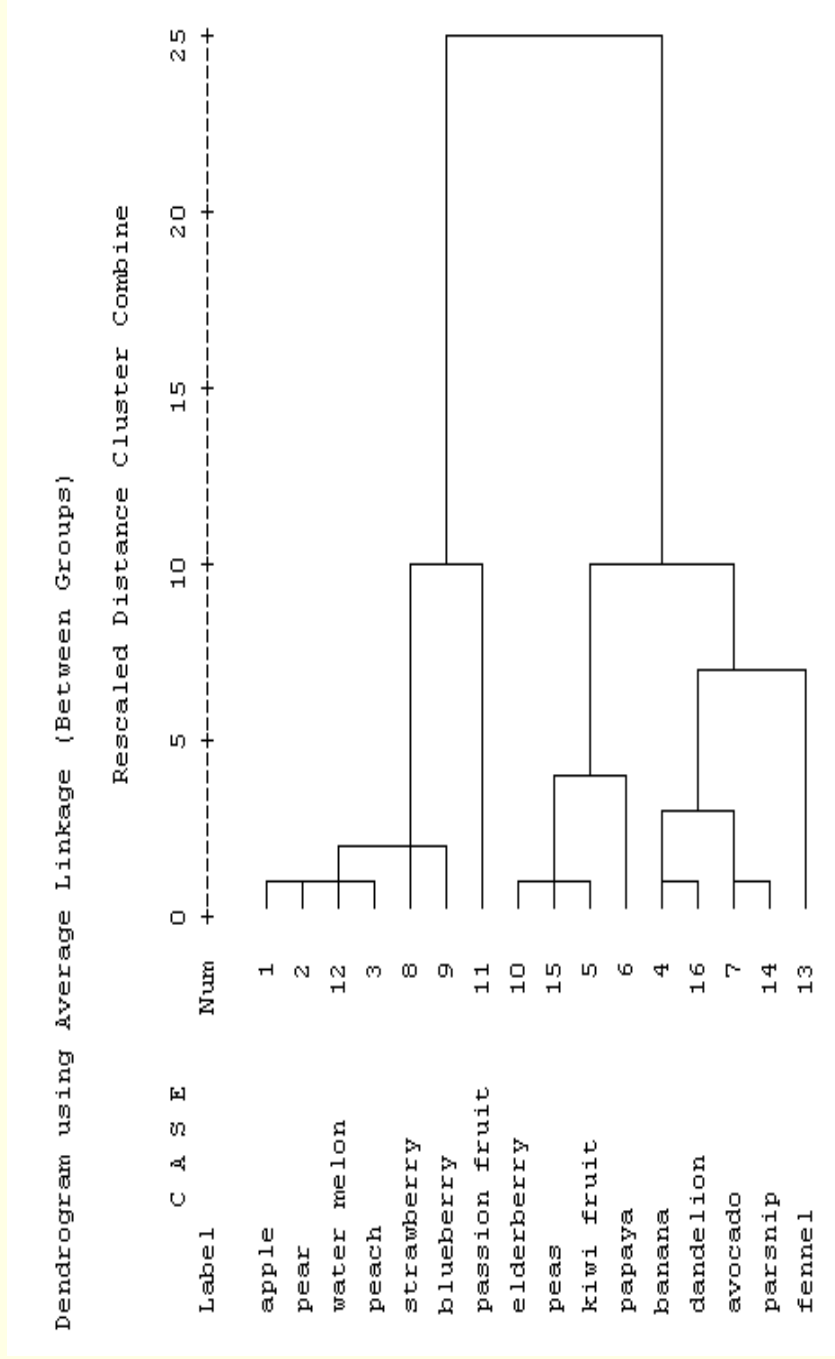
Ward's method



Complete linkage

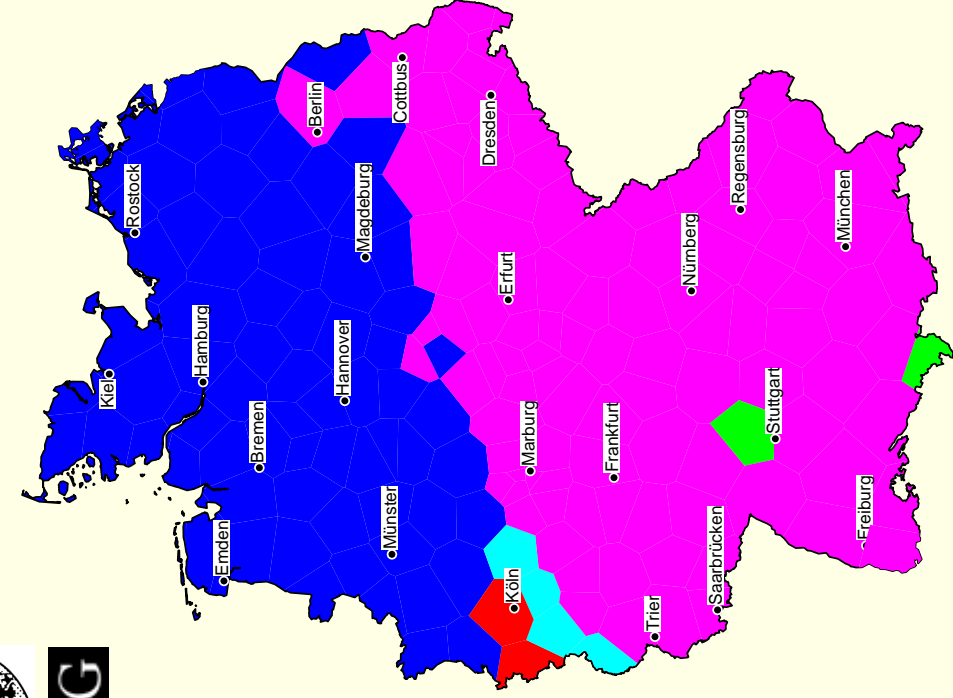


Space-conserving – example 1

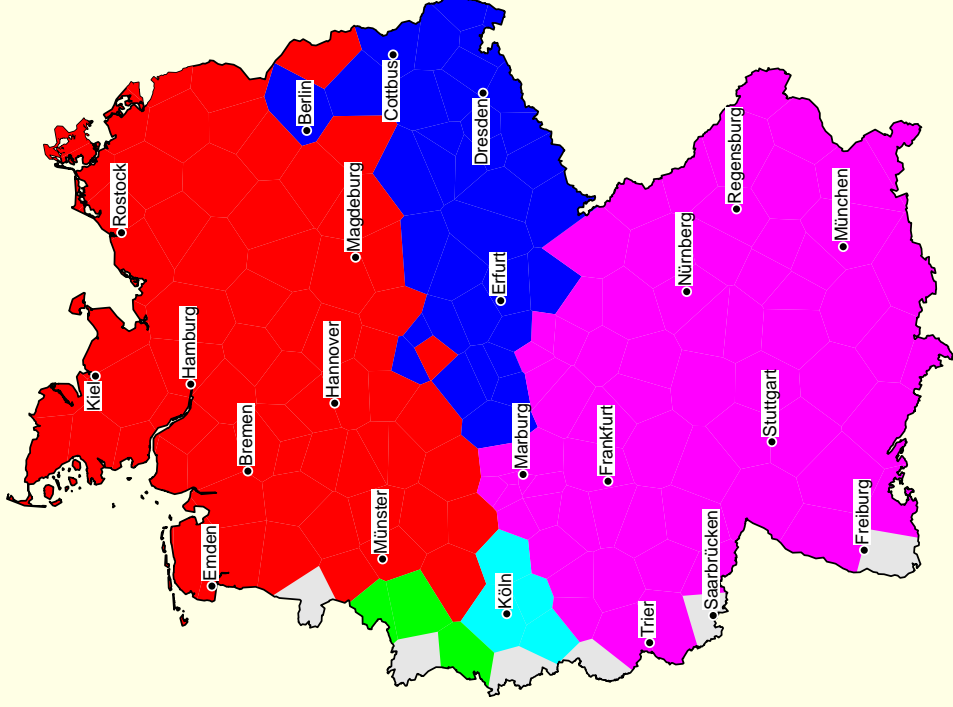


Dendrogram generated by UPGMA

Space-conserving – example 2



UPGMA



WPGMA



Matrix updating algorithms

- Which should be used?
 - Outliers shall be visible
 - Single linkage
 - Unequally sized clusters expected
 - *Not* space-dilating methods
 - Differing object density in expected clusters
 - WPGMA
- No-idea-just-try-order:
 - Space-conserving > space-dilating > space-contracting



Number of clusters

- How many ,natural‘ classes has cluster analysis generated?
 - Subjective decision of researcher
 - Analysis of merging values
 - Large step = rather dissimilar clusters = stop
 - Plot number of clusters against merging values
 - Graph flattens = no new information = stop
 - Ward’s method: Significance test possible



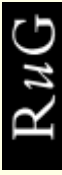
Validation of results

- Results should be stable
- Plausible interpretation possible
- Repeat cluster analysis with different samples of the same population
 - Different results = both invalid, but
 - Same results = not necessarily valid and
 - not always possible due to lack of data
- Cophenetic correlation, but
 - Normal distribution (wrongly?) assumed
 - In dendrogram fewer (different) values



Validation of results

- Significance tests
 - Used attributes: Useless because always significant
 - Not used (but relevant) attributes: Useful but only possible when knowledge about classes already exists...
- Monte Carlo procedures
 - Data set is created which has the same global properties as original data but contains no classes
 - Both sets are clustered & results compared
 - Significant differences => results valid



Attention!

- A lot of factors determine the results of cluster analysis
 - *Very careful selection of objects, attributes, (dis)similarity measure, cluster method and matrix updating algorithm*
- **Cluster analysis will *always* output clusters – if there are natural classes or not!**