Explorative methods

Introduction to Statistics
Carl von Ossietzky Universität Oldenburg
Fakultät III - Sprach- und Kulturwissenschaften

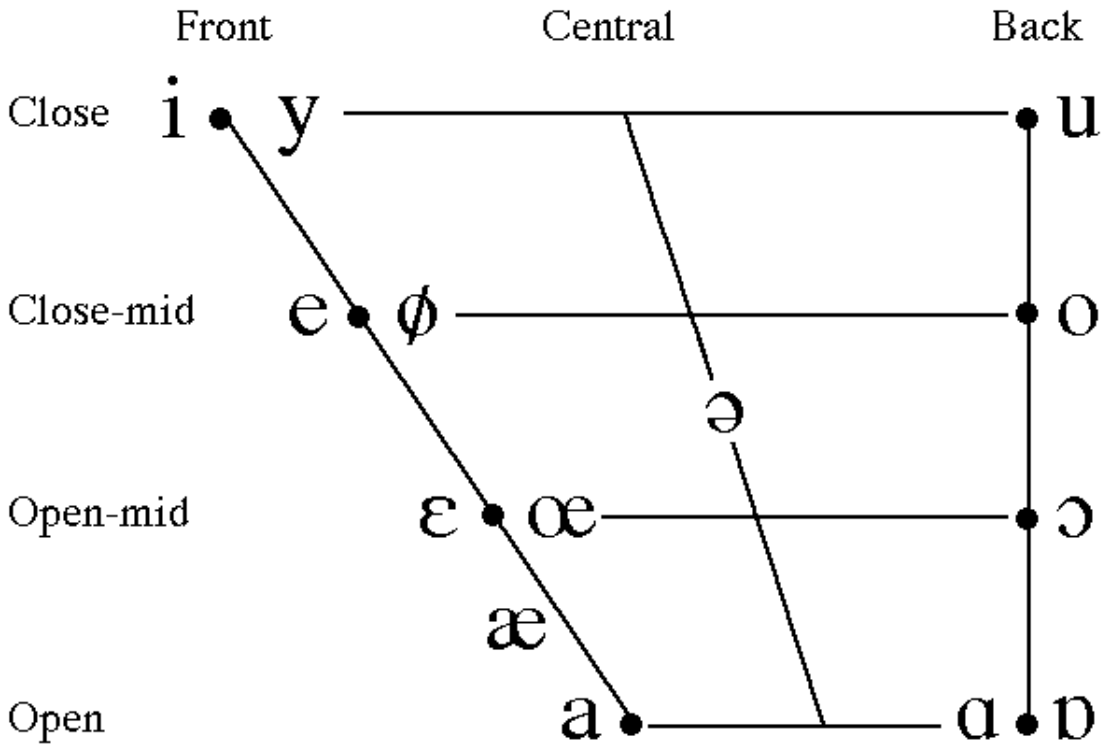
Introduction

- Explorative research:
  goal is to explore the reality in order to find patterns that are not predicted by the experimenters current knowledge or pre-conceptions.
- We focus on:
  - Principal components analysis
  - Multidimensional scaling
  - Cluster analysis
- Exploratory research is not typically generalizable to the population at large.

Example

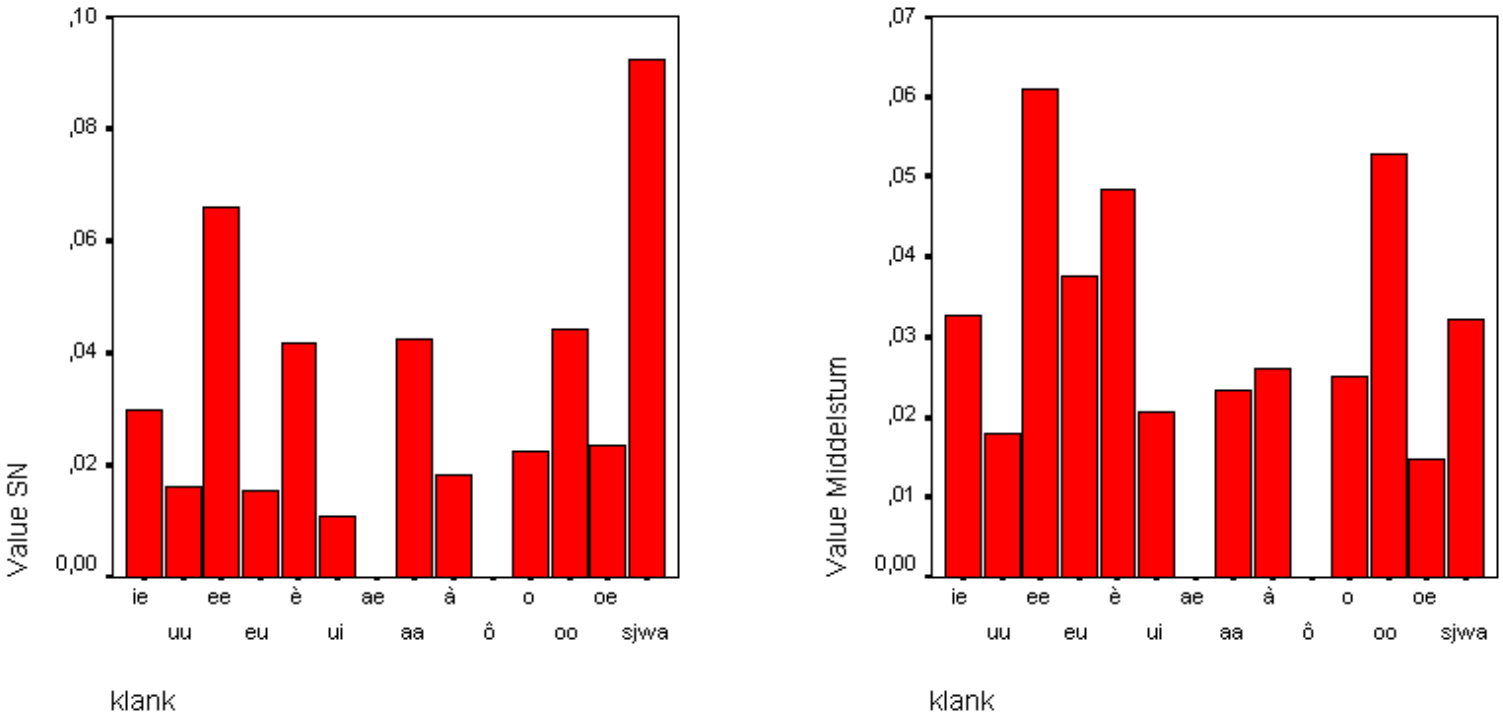
- We compare 16 local Dutch dialects, one for each province.
- Data taken from the *Reeks Nederlandse Dialectatlassen*, a series of Dutch dialect atlasses compiled by E. Blancquaert and W. Pée.
- The material is recorded in the period 1922–1975.
- For each dialect the same set of 139 sentences is translated and pronounced by the informants and transcribed by the field workers.
- We choose a subset of 125 words.
- For each dialect we calculated the relative frequencies of vowels and consonants in the set of 125 words (method of Hoppenbrouwers & Hoppenbrouwers, 1988).
- We focus on the relative frequencies of the *vowels*.
- Standard Dutch is also added.

For each province we choose one local dialect (except for Flevoland).



Overview of the vowels used in the dialects. For each dialect we measured the relative frequency of the vowels in the set of 125 words.

Example



Vowel frequencies of standard Dutch (left) and the Groningen dialect of Middelstum (right).

Example

- Each dialect is characterized 13 relative sound frequencies, for each of the sounds one relative frequency.
- Could the 13 variables be reduced to a smaller set of underlying variables?
- We can find out this by using **principal component analysis**.

Principal components analysis

- Principal Component Analysis (PCA) allows researchers to investigate concepts that are not easily measured directely.
- PCA collapses a larger number of variables into a few interpretable factors.
- Example:
  people may respond similarly to questions about income, education, and occupation, which are all associated with the latent variable socioeconomic status.
- Latent variable:
  a variable that cannot be directly measured, but is assumed to be related to several variables that can be measured.
- Key concept:
  multiple observed variables have similar patterns of responses because of their association with an underlying latent variable or **factor**.
- First a correlation matrix (or: R-matrix) is generated for all possible pairings of the variables.

Correlation Matrix

		ie	uu	ee	eu	e	ui	ae	aa	a	o	oo	oe	schwa
Correlation	ie	1,000	,265	-,841	-,369	-,160	-,201	,582	-,586	,181	,027	-,363	,456	,259
	uu	,265	1,000	-,362	-,047	,300	,427	,046	-,145	-,100	-,102	-,030	-,261	-,048
	ee	-,841	-,362	1,000	,572	-,069	,135	-,516	,215	-,065	,030	,412	-,528	-,422
	eu	-,369	-,047	,572	1,000	,074	,013	-,204	-,226	-,314	,111	,427	-,583	-,604
	e	-,160	,300	-,069	,074	1,000	,604	-,614	,423	-,475	,260	-,054	-,331	-,153
	ui	-,201	,427	,135	,013	,604	1,000	-,463	,294	-,178	,169	-,027	-,540	-,492
	ae	,582	,046	-,516	-,204	-,614	-,463	1,000	-,528	,069	-,204	-,282	,426	,402
	aa	-,586	-,145	,215	-,226	,423	,294	-,528	1,000	-,154	-,186	,075	-,030	,014
	a	,181	-,100	-,065	-,314	-,475	-,178	,069	-,154	1,000	-,494	,060	,359	,042
	o	,027	-,102	,030	,111	,260	,169	-,204	-,186	-,494	1,000	-,377	-,216	-,310
	oo	-,363	-,030	,412	,427	-,054	-,027	-,282	,075	,060	-,377	1,000	-,300	-,127
	oe	,456	-,261	-,528	-,583	-,331	-,540	,426	-,030	,359	-,216	-,300	1,000	,557
	schwa	,259	-,048	-,422	-,604	-,153	-,492	,402	,014	,042	-,310	-,127	,557	1,000

Correlations for all possible pairings of the variables.

Correlation Matrix

		ie	uu	ee	eu	e	ui	ae	aa	a	o	oo	oe	schwa
Correlation	ie	1,000	,265	-,841	-,369	-,160	-,201	,582	-,586	,181	,027	-,363	,456	,259
	uu	,265	1,000	-,362	-,047	,300	,427	,046	-,145	-,100	-,102	-,030	-,261	-,048
	ee	-,841	-,362	1,000	,572	-,069	,135	-,516	,215	-,065	,030	,412	-,528	-,422
	eu	-,369	-,047	,572	1,000	,074	,013	-,204	-,226	-,314	,111	,427	-,583	-,604
	e	-,160	,300	-,069	,074	1,000	,604	-,614	,423	-,475	,260	-,054	-,331	-,153
	ui	-,201	,427	,135	,013	,604	1,000	-,463	,294	-,178	,169	-,027	-,540	-,492
	ae	,582	,046	-,516	-,204	-,614	-,463	1,000	-,528	,069	-,204	-,282	,426	,402
	aa	-,586	-,145	,215	-,226	,423	,294	-,528	1,000	-,154	-,186	,075	-,030	,014
	a	,181	-,100	-,065	-,314	-,475	-,178	,069	-,154	1,000	-,494	,060	,359	,042
	o	,027	-,102	,030	,111	,260	,169	-,204	-,186	-,494	1,000	-,377	-,216	-,310
	oo	-,363	-,030	,412	,427	-,054	-,027	-,282	,075	,060	-,377	1,000	-,300	-,127
	oe	,456	-,261	-,528	-,583	-,331	-,540	,426	-,030	,359	-,216	-,300	1,000	,557
	schwa	,259	-,048	-,422	-,604	-,153	-,492	,402	,014	,042	-,310	-,127	,557	1,000

We find groups of variables with (mostly) strong mutual correlations: a group of high vowels (ie, ee, eu, oe, schwa) and a group of front vowels (ie, e, ae, aa).

Extracting factors

- From the correlation matrix, factors are extracted. The most common method of extraction is called **principal components**.
- We find at least two groups of sounds which strongly correlate to each other. Each group will be represented by a factor.
- If there are n variables, PCA will generate n factors. The factors are always listed in order of how much variation they explain.

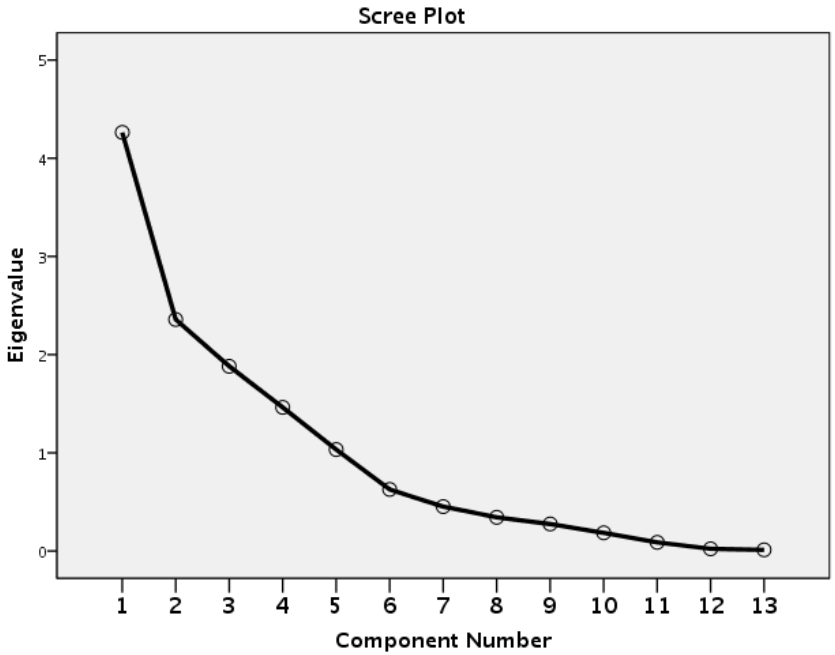
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,263	32,790	32,790	4,263	32,790	32,790	3,186	24,509	24,509
2	2,357	18,133	50,923	2,357	18,133	50,923	2,728	20,983	45,492
3	1,881	14,469	65,392	1,881	14,469	65,392	2,049	15,760	61,252
4	1,463	11,250	76,642	1,463	11,250	76,642	1,605	12,348	73,600
5	1,034	7,951	84,593	1,034	7,951	84,593	1,429	10,993	84,593
6	,626	4,819	89,412						
7	,452	3,477	92,889						
8	,343	2,641	95,529						
9	,275	2,113	97,642						
10	,185	1,423	99,065						
11	,088	,675	99,739						
12	,022	,170	99,909						
13	,012	,091	100,000						

Extraction Method: Principal Component Analysis.

The **eigenvalue** is a measure of how much variance of the observed variables a factor explains. Any factor with an eigenvalue ≥ 1 explains more variance than a single observed variable.

Scree plot



The factors that explain the least amount of variance are usually discarded. A **scree plot** may help to decide about this. In a scree plot each factor is plotted against its eigenvalue.

Factor loadings

- Factor:
  is an latent (unmeasured) variable that expresses itself through its **relationship** with other measured variables.
- Factor loading:
  expresses the relationship of a factor to a variable. It is the correlation coefficient between the variable and the factor.
- For each variable the loadings (correlations) are given with respect to each of the factors.

Factor loadings

Component Matrix[a]

	Component				
	1	2	3	4	5
oe	-,794				
ae	-,776				
ie	-,753	,410			
ee	,732	-,572			
schwa	-,652		,419		,481
eu	,606		-,600		
ui	,586	,558			
e	,524	,660			
oo		-,533		,444	
aa	,404		,832		
uu		,589		,662	
o		,450		-,660	
a		-,416		,435	-,631

Extraction Method: Principal Component Analysis.

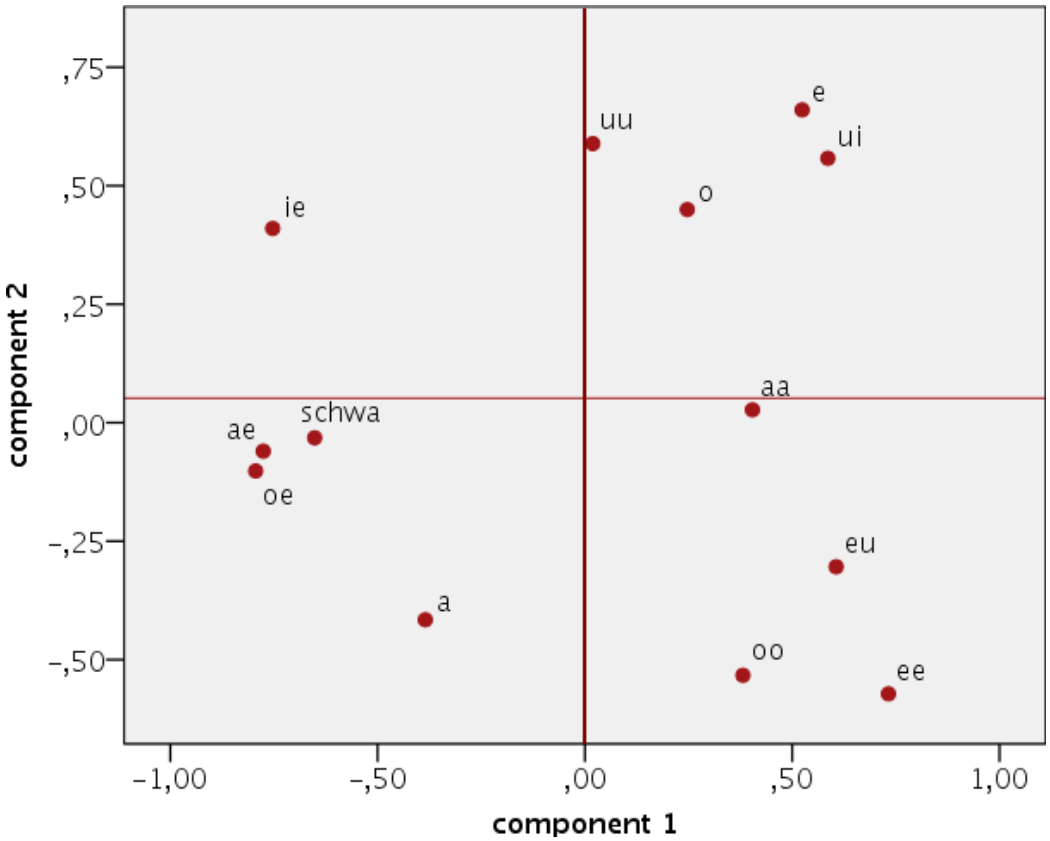
a. 5 components extracted.

Absolute values smaller than 0.4 are omitted.

Factor loadings

- When there are n factors, than the n loadings per variable are the coordinates of the variable in n-dimensional space.
- We show the 13 sounds on the basis of the loadings of the first two factors only.
- The loadings of Factor 1 are the x-coordinates, the loadings of Factor 2 are the y-coordinates.
- The x-axis represents Factor 1, the y-axis represents Factor 2.
- In PCA factors are called components!

Factor loadings



Example: ie has loading -0.753 for Factor 1 and loading 0.410 for Factor 2.

Communalities

Communalities

	Initial	Extraction
ie	1,000	,867
uu	1,000	,837
ee	1,000	,890
eu	1,000	,864
e	1,000	,829
ui	1,000	,845
ae	1,000	,788
aa	1,000	,872
a	1,000	,940
o	1,000	,873
oo	1,000	,749
oe	1,000	,799
schwa	1,000	,843

Extraction Method: Principal Component Analysis.

The **communality** measures the proportion of a variable's variance that is common variance. Varies between 0 (variable shares none of its variance with any other variable) and 1 (variable has no unique variance).

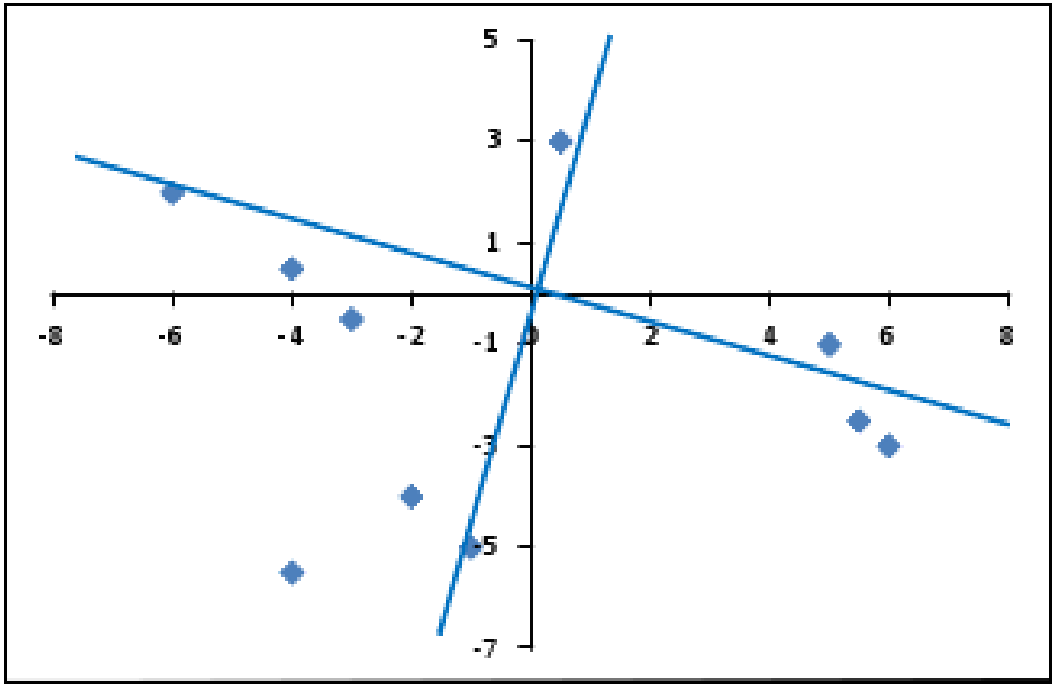
Rotation

- Imagine we have 10 variables that go into a factor analysis.
- The program finds the group of strongest mutually correlating variables and calls it Factor 1.
- Then the program looks for the group of second strongest correlating variables and calls it Factor 2, and so on.
- When there are two factors, then the 10 variables can be visualized in two-dimensional space, where Factor 1 is represented by the x-axis and Factor 2 by the y axis.

Rotation

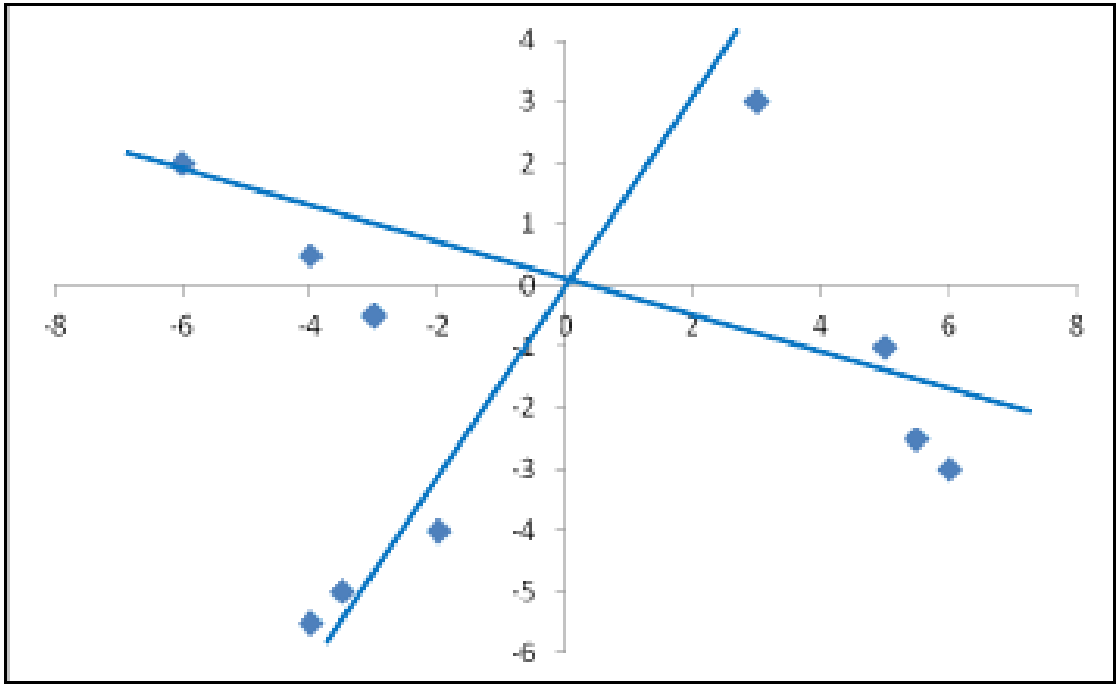
- Sometimes the initial solution results in strong correlations of a variable with **several** factors or in a variable that has no strong correlation with **any** of the factors.
- The program can rotate the axes so that it is ensured that groups of mutually correlating variables are intersected by the factor to which they relate most.
- After rotation, the loadings of the variables are maximized onto one factor (the factor that intersects the group) and minimized on the remaining factor(s).

Rotation



Orthogonal rotation. The axes have an angle of 90 degrees. Source: `http://www.theanalysisfactor.com/rotations-factor-analysis/`.

Rotation



Oblique rotation. The axes have an angle smaller than 90 degrees. Source: `http://www.theanalysisfactor.com/rotations-factor-analysis/`.

Rotation

- Orthogonal rotation:
  rotation that assumes the factors are not correlated.
- In SPSS: choose **varimax**.
- Oblique rotation:
  rotation that allow for correlation between factors.
- In SPSS: choose **direct oblimin**.

Factor loadings

Component Matrix[a]

	Component				
	1	2	3	4	5
oe	-,794				
ae	-,776				
ie	-,753	,410			
ee	,732	-,572			
schwa	-,652		,419		,481
eu	,606		-,600		
ui	,586	,558			
e	,524	,660			
oo		-,533		,444	
aa	,404		,832		
uu		,589		,662	
o		,450		-,660	
a		-,416		,435	-,631

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

Rotated Component Matrix[a]

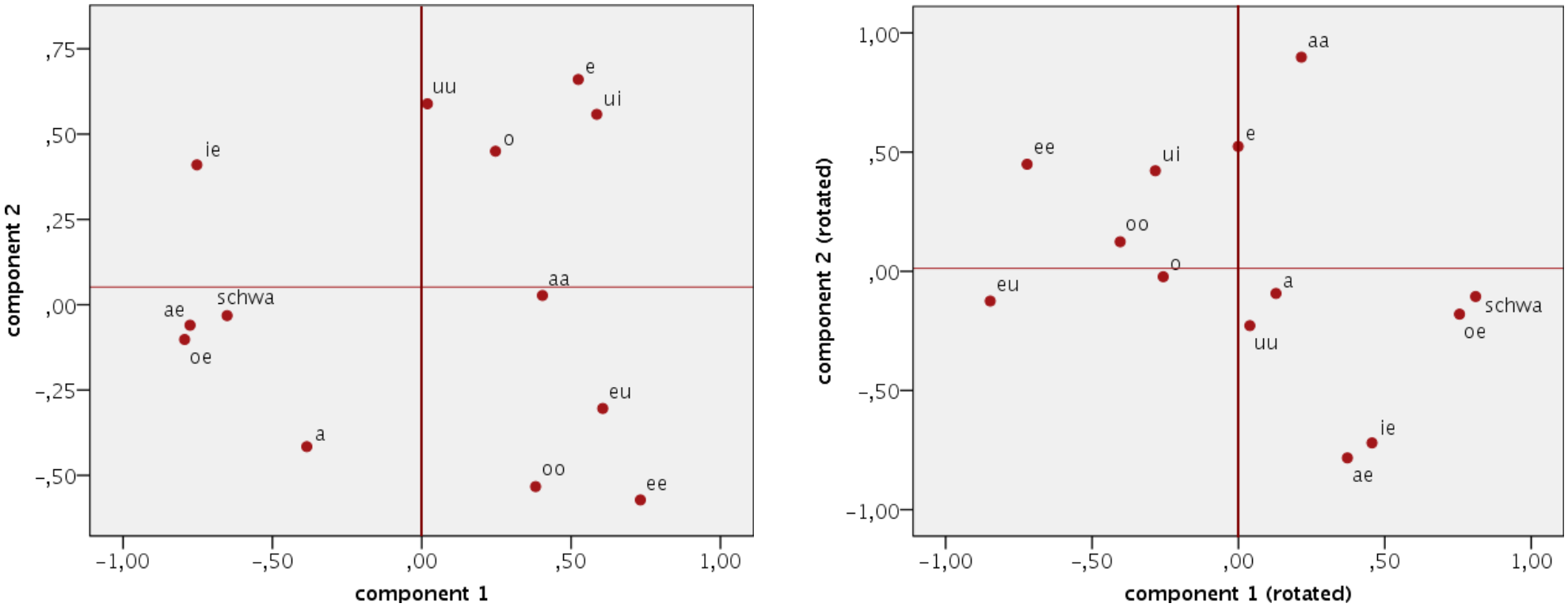
	Component				
	1	2	3	4	5
eu	-,848				
schwa	,810				
oe	,755				
ee	-,722	,449			
aa		,898			
ae		-,783			
ie	,456	-,720			
uu			,866		
ui		,422	,730		
e		,524	,576		,429
o				-,814	
oo	-,404			,754	
a					-,936

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 13 iterations.

Absolute values smaller than 0.4 are omitted.

Factor loadings



Factor plots showing the first two factors before (left) and after varimax rotation (right).

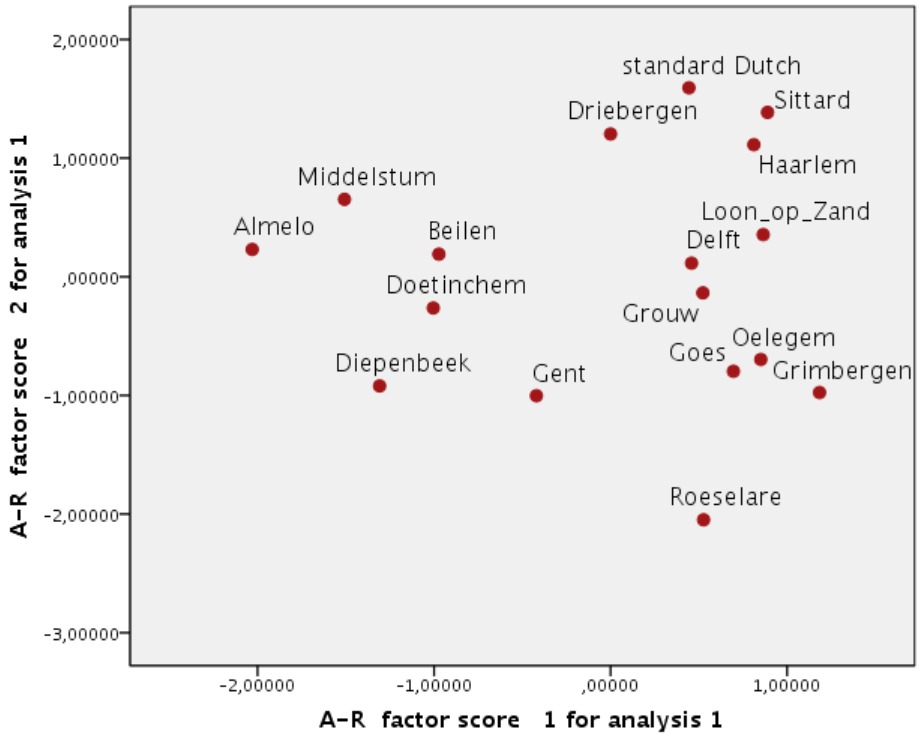
Factor scores

- **Factor scores** are the scores of each case (row) on each factor (column). In our example: what are the scores of each dialect on each of the factors?
- Simplest approach:
  - To compute the factor score for a given case for a given factor, first standardize each of the variables, i.e. per variable transform the values to z-scores.
  - Multiply each of the standardizes values of a case (one value from each variable) with the corresponding loading of the variable for the given factor.
  - Calculate the sum of the products.

Factor scores

- SPSS offers several more sophisticated alternatives:
  - Regression:
    if correlations between factor scores are acceptable.
  - Anderson-Rubin:
    ensures that factor scores are uncorrelated (to be prefered).
- When correlating variables to factor scores, we find the loadings!

Factor scores



The first factor represents variation in the higher vowels, and the second factor represents variation in the front vowels.

Assumptions

- The variables should be measured at an interval level.
- The variables should be rougly normally distributed. Use normal quantile plots and run the Shapiro-Wilk test.
- The assumption of normality is most important if you wish to generalize the results of your analysis beyond the sample collected.
- We test the normality of each of the 13 variables.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ie	,179	17	,148	,948	17	,425
uu	,163	17	,200*	,932	17	,239
ee	,161	17	,200*	,923	17	,167
eu	,217	17	,032	,929	17	,210
e	,099	17	,200*	,979	17	,950
ui	,147	17	,200*	,931	17	,229
ae	,287	17	,001	,754	17	,001
aa	,223	17	,024	,897	17	,061
a	,147	17	,200*	,957	17	,568
o	,193	17	,093	,951	17	,476
oo	,195	17	,084	,881	17	,033
oe	,184	17	,128	,905	17	,084
schwa	,222	17	,026	,884	17	,037

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Results of the Shapiro-Wilk test in the right columns. Consider also normal quantile plots.

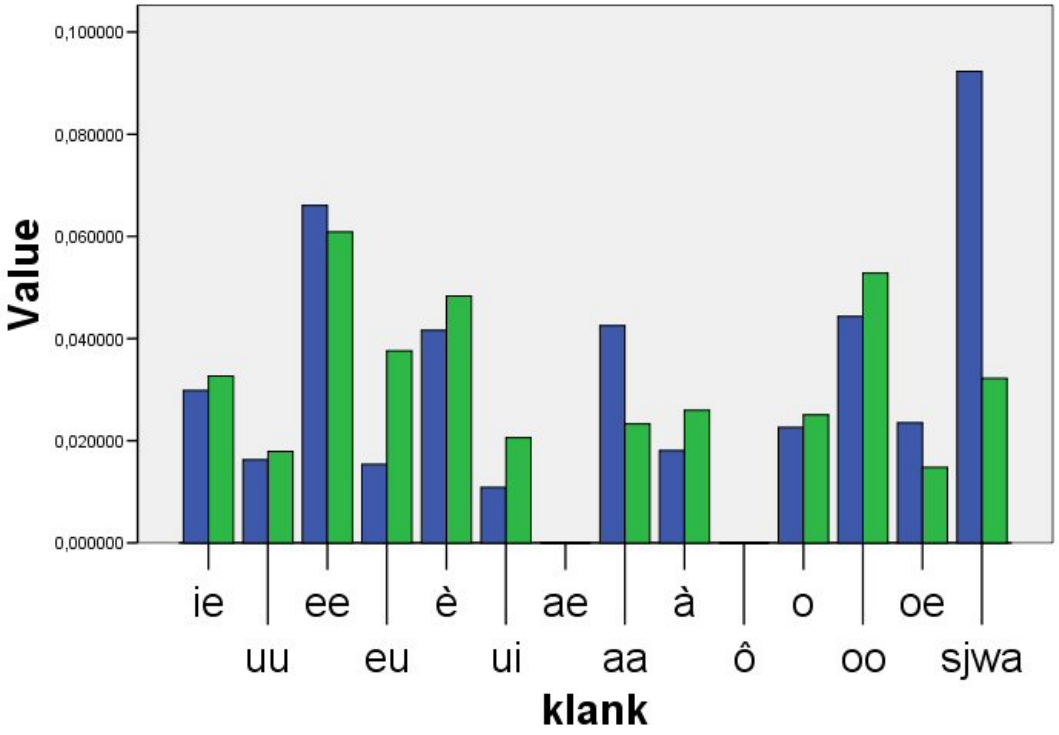
Multidimensional scaling

- In a geographic map distances between locations can be measured with a ruler.
- Multidimensional scaling: distances are given, on the basis of them coordinates in the map are determined.
- We have 17 language variaties (16 dialects and standard Dutch). We can measure the mutual distances. On the basis of those distances we determine coordinates in k-dimensional space.
- The distances between the points in the k dimensional space should approximate the original distance in the distance matrix as close as possible.
- The higher k, the better the approximation. The lower k, the easier the result can be interpreted. Usually we would choose k=2.
- Two types: metric MDS (Togerson 1952) and nonmetric MDS (Shepard 1962, Kruskal 1964).

The measurements of distances

- MDS expects distances between objects.
- Our data consists of a table where the columns represent the dialects, the rows represent the sounds, and each cell contains the relative frequency of a sound for a dialect.
- When performing MDS in SPSS see to it that SPSS calculates proximities (distances) between the dialects.

The measurements of distances



Comparison of relative vowel frequencies of Standard Dutch (blue) and the dialect of Middelstum (green).

The measurements of distances

- Assume we measure the distance between dialect X and dialect Y with $n = 13$ relative frequencies per dialect.
- Block:

$$\delta(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- Euclidean distance:

$$\delta(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- The Euclidean distance emphasizes larger differences.

The measurements of distances

Relative frequency SD	Relative frequency Middelstum	Difference	Squared difference
.0298600	.0326800	-.002820	.000008
.0162900	.0179100	-.001620	.000003
.0660600	.0608800	.005180	.000027
.0153800	.0376000	-.022220	.000494
.0416300	.0483400	-.006710	.000045
.0108600	.0205900	-.009730	.000095
.0000000	.0000000	.000000	.000000
.0425300	.0232800	.019250	.000371
.0181000	.0259600	-.007860	.000062
.0226200	.0250700	-.002450	.000006
.0443400	.0528200	-.008480	.000072
.0235300	.0147700	.008760	.000077
.0923100	.0322300	.060080	.003610

Euclidean distance: $\sqrt{0.0049} = 0.0698$

Metric MDS

- Given a distance matrix D with mutual distances between the objects.
- We want to put objects in a k dimensional space, so that the Euclidean distances between the points (calculated on the basis of the coordinates which were found by the MDS procedure) approximate the original distances in the distance matrix as close as possible.
- Given objects $x_1, x_2, \ldots, x_n$, then:

$$Stress = \sum_{i,j} (\delta(x_i, x_j) - d_{ij})^2$$

  is minimized.
- $\delta(x_i, x_j)$ is the Euclidean distance between the objects x_i and x_j in k dimensional space.
- The d_{ij}'s are distances between the objects in k dimensional space which are predicted on the basis of the original distances D_{ij} with *linear regression analysis*.

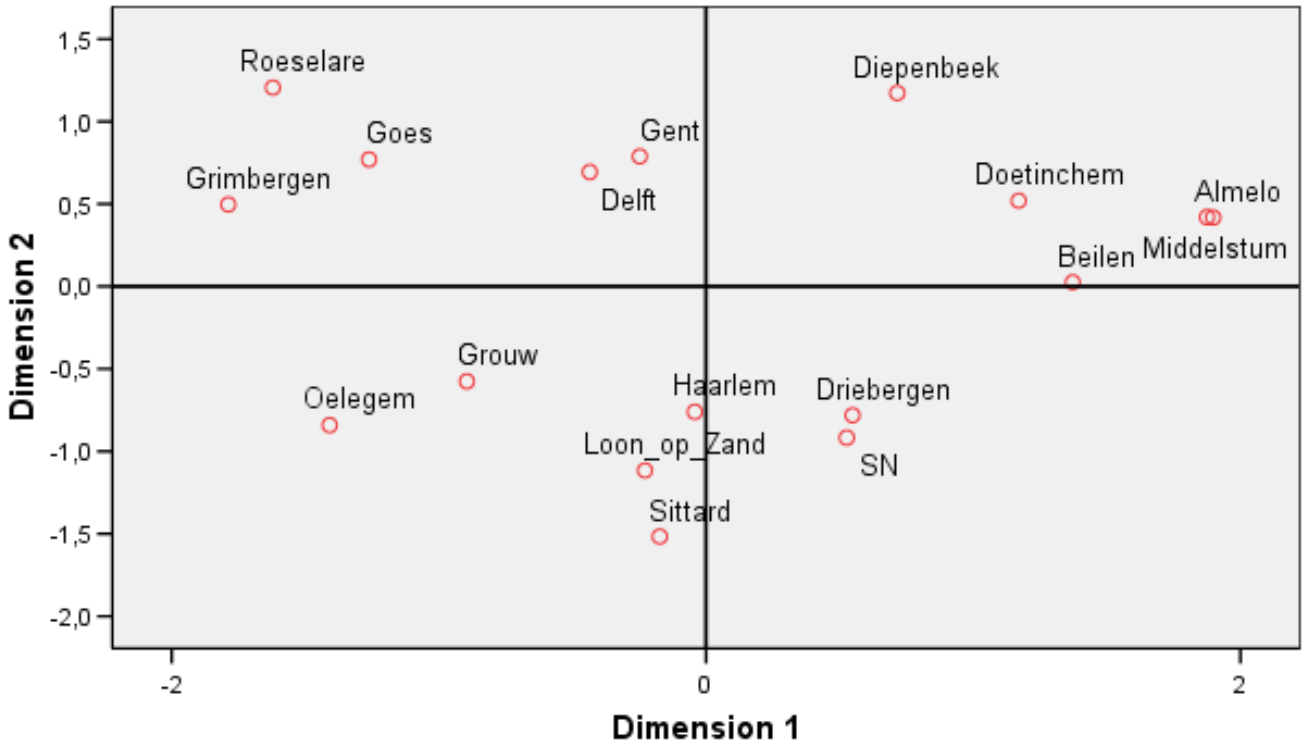
Metric MDS

- In our example, each dialect is represented by a vector of relative sound frequencies. Choose options 'Create distances from data' (Alscal) or 'Create proximities from data' (Proxscal) in SPSS. Use the Euclidean distance measure.
- Choose the 'level of measurement' (Alscal) or 'proximity transformations' (Proxscal) in SPSS. Choose 'Ratio'.

Metric MDS

- Squared correlation coefficient: correlate the distances between the objects in k dimensional space (i.e. Euclidean distances between the corresponding coordinates found by the MDS procedure) with the original distances in distance matrix D, and calculate the squared correlation coefficient.
- The squared correlation coefficient gives the amount of variance in the original distances explained by the distances between the objects in k dimensional space. In SPSS: RSQ = R-SQuare.
- RSQ=0.6 is considered to be an acceptable lower boundary.

Metric MDS



Metric MDS-plot (ratio) created in SPSS (Alscal). Stress = 0.24072, and RSQ = 0.74110.

Nonmetric MDS

- Given a distance matrix D with mutual distances between the objects.
- We want to put objects in a k dimensional space, so that the $order$ of the objects (as suggested by the Euclidean distances between the corresponding coordinates found by the MDS procedure) reflects the $order$ of the objects as suggested by the original distances in the distance matrix D.
- For example: when the distance between object x_1 and x_2 is smaller than between object x_1 and x_3, then this should also be reflected in k dimensional space. The $extent$ to which one distance is smaller than another does not play a role.
- Given objects $x_1, x_2, \ldots, x_n$, then

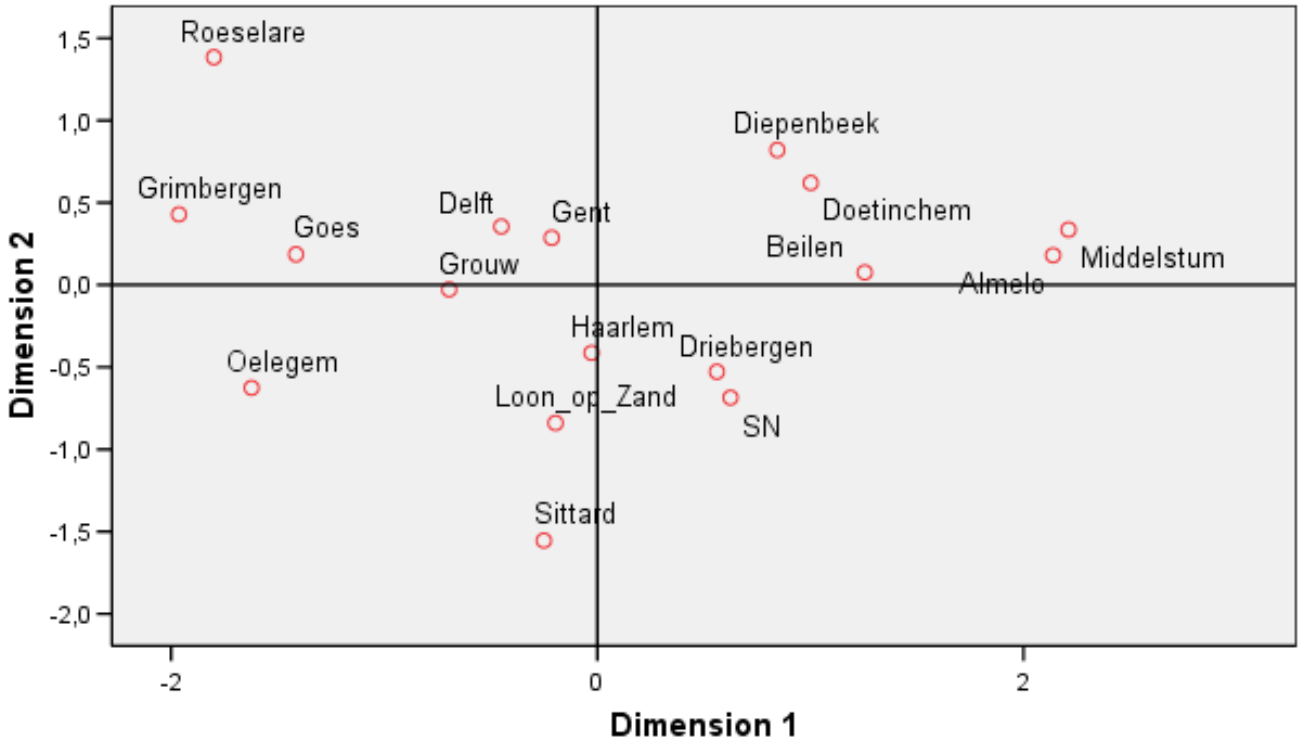
$$Stress = \sqrt{\frac{\sum_{i,j} \left(\delta(x_i, x_j) - d_{ij}\right)^2}{\sum_{i,j} d_{ij}^2}}$$

  is minimized.

Nonmetric MDS

- $\delta(x_i, x_j)$ is the Euclidean distance between objects x_i and x_j in k dimensional spacel
- The d_{ij}'s are distances between objects in k dimensional space which are predicted on the basis of the original distances D_{ij} with $weak\ monotonous\ regression\ analysis$: the order of the distances is kept.
- Choose the 'level of measurement' (Alscal) or 'proximity transformations' (Proxscal) in SPSS. Choose 'Ordinal'.

Nonmetric MDS

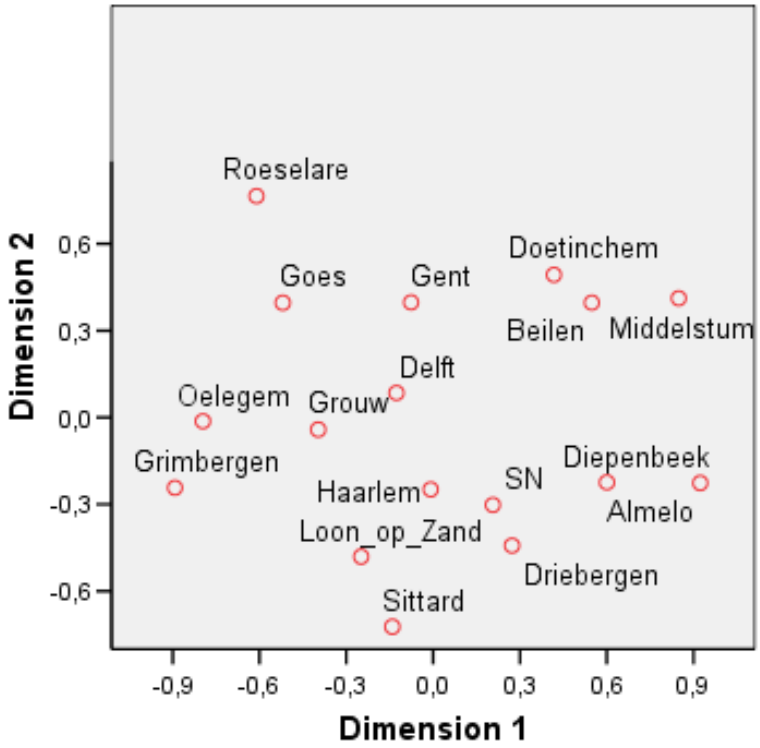


Nonmetric MDS-plot (ordinal) created in SPSS (Alscal). Stress = 0.16190 en RSQ = 0.87035.

Alscal, Proxscal, Prefscal

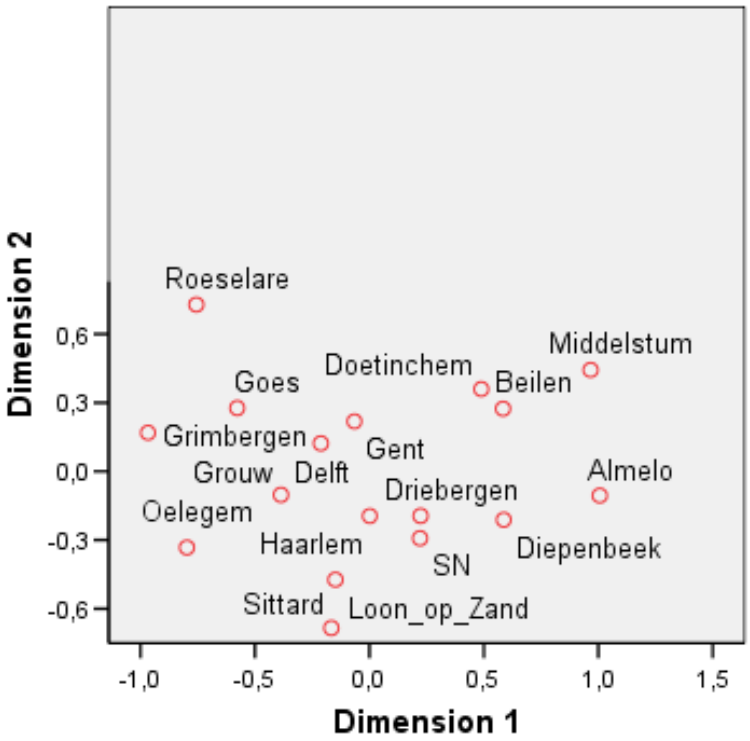
- SPSS has three MDS procedures: Alscal, Proxscal and Prefscal.
- Alscal minimizes the $S\text{-}stress$, a measure based on squared distances which emphasizes large distances relatively more strongly.
- Proxscal minimizes $Normalized\ Raw\ Stress$, a measure based on the 'common' (non-squared) distances: same weighing for small and large distances.
- Prefscal is meant for multidimensional unfolding. The SPSS procedure expects a proximity matrix as input.

Metric MDS

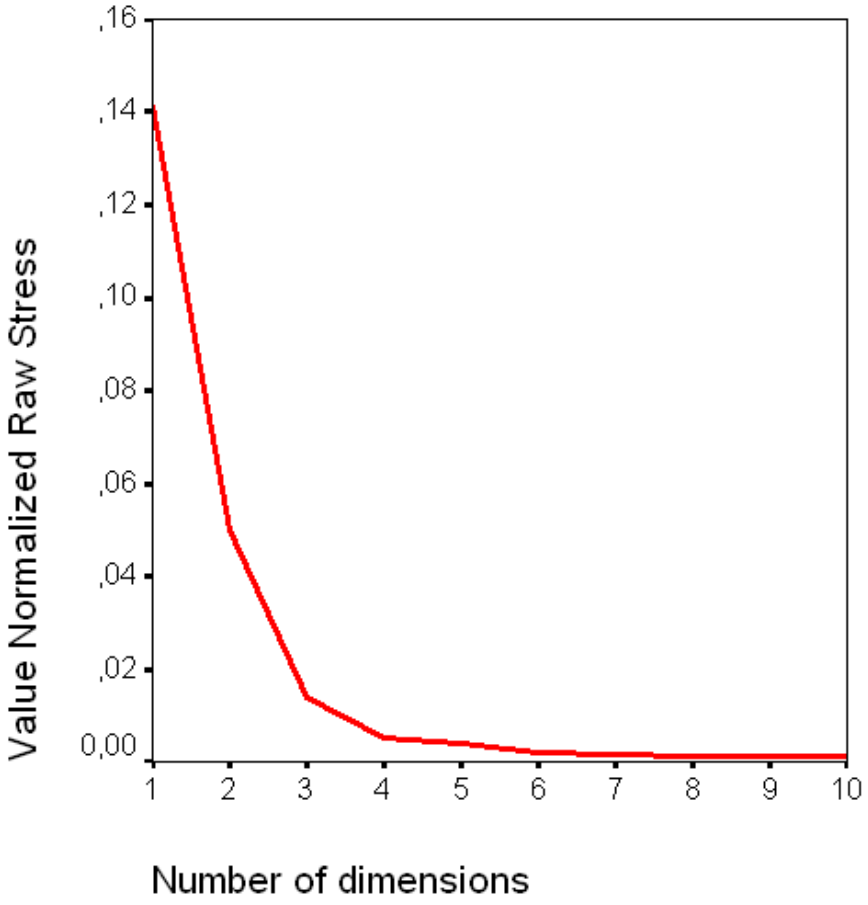


Metric MDS-plot (ratio) created in SPSS (Proxscal). Normalized Raw Stress = 0.04994, and Tucker's Coefficient of Congruence = 0.97471.

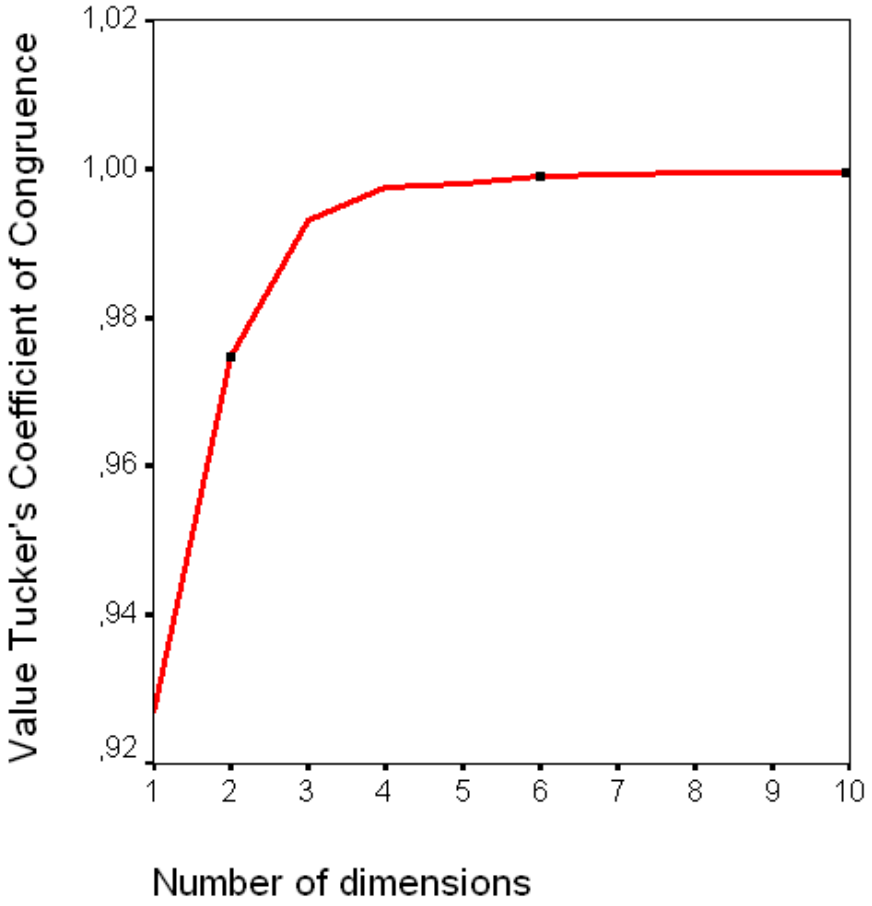
Nonmetric MDS



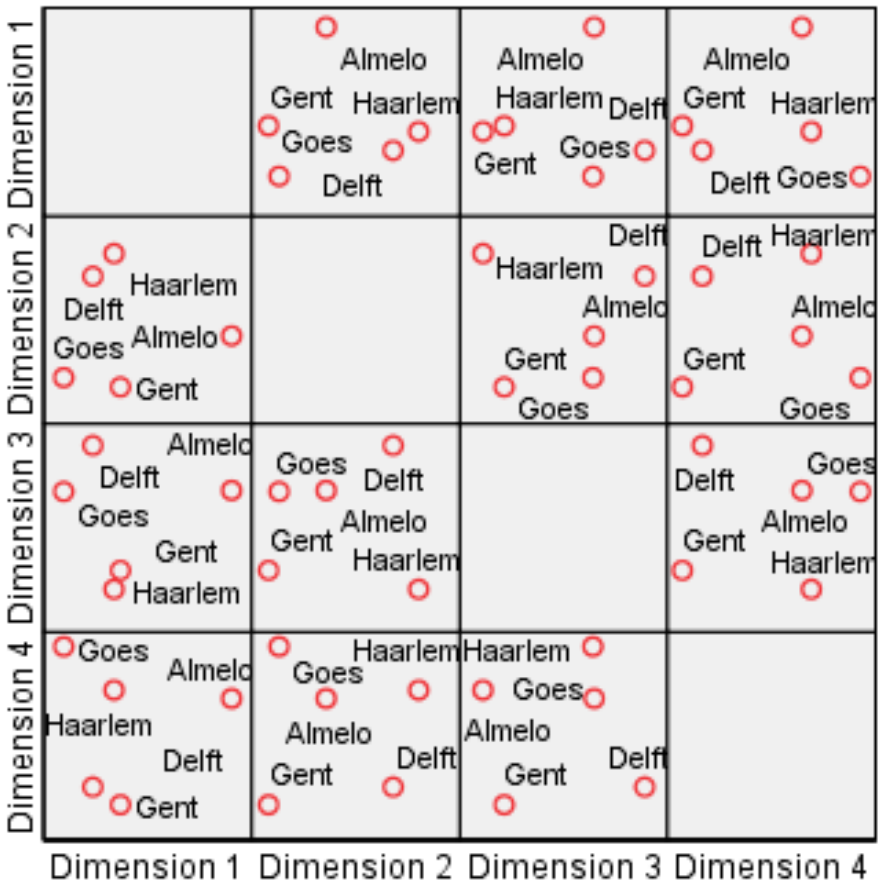
Nonmetric MDS-plot (ordinal) created in SPSS (Proxscal). Normalized Raw Stress = 0.02176, and Tucker's Coefficient of Congruence = 0.98906.



Dimensionality. Scree plot showing the Normalized Raw Stress for each of the 10 dimensions (metric MDS, Proxscal). After four dimensions hardly any improvement is found.



Dimensionality. Elbow graph showing Tucker's Coefficient of Congruence for each of the 10 dimensions (metric MDS, Proxscal). This graph also shows that hardly any improvement is found after four dimensions.



Each of the four dimensions is drawn against the other three dimensions, resulting in 12 two-dimensional plots. A subset of five dialects is used.

Cluster analysis

- **Cluster analysis**: the process of classification of objects in subsets which have a meaning in the context of a particular problem (Jain & Dubes, *Algorithms for Clustering Data*, 1988, p. 55).
- **SAHN**: Sequential, Agglomerative, Hierarchical, Nonoverlapping
- *Sequential*: objects are processed one after another, not simultaneously.
- *Agglomerative*: when we have n objects, initially create n clusters, each cluster having an object, and each object found in one cluster. Combine small clusters to larger ones until all objects are found in one large cluster.
- *Hierarchical*: we obtained a nested structur: a cluster contains subclusters, a subcluster contains subsubclusters, etc.
- *Nonoverlapping*: At each division in the hierarchie, each object belongs to exactly one cluster.

Cluster analysis

- SPSS expects a table like ours, where each dialect is represented by a vector of relative sound frequencies.
- In our case: each dialect is represented by a column which contains the relative frequencies.
- Under Method choose the Interval (i.e. distance measure). Choose Block or Euclidean distance (as with MDS) or may be another one.

Johnson's algorithm

	Almelo	Haarlem	Delft	Gent	Goes
Almelo	.000	.078	.084	.069	.100
Haarlem	.078	.000	.040	.057	.062
Delft	.084	.040	.000	.057	.049
Gent	.069	.057	.057	.000	.043
Goes	.100	.062	.049	.043	.000

Apply Johnson's algorithm to the top right half of the matrix (blue values):

- Iteratively,
  1. select the smallest distance in the matrix
  2. fuse the two data points that give rise to it to one cluster
- In order to iterate, calculate the distance of the newly formed cluster to the other clusters. For example calculate the average between a cluster and the two data points in the newly formed cluster.
- Repeat until one cluster is left.

Johnson's algorithm

- Iteratively,
  1. select the smallest distance in matrix: between Haarlem and Delft: .040
  2. fuse the two data points to one cluster: Haarlem & Delft
- In order to iterate, calculate the distance between the newly formed cluster and the other clusters:

	Haarlem	Delft	Haarlem & Delft
Almelo	.078	.084	.081
Gent	.057	.057	.057
Goes	.062	.049	.056

Johnson's algorithm

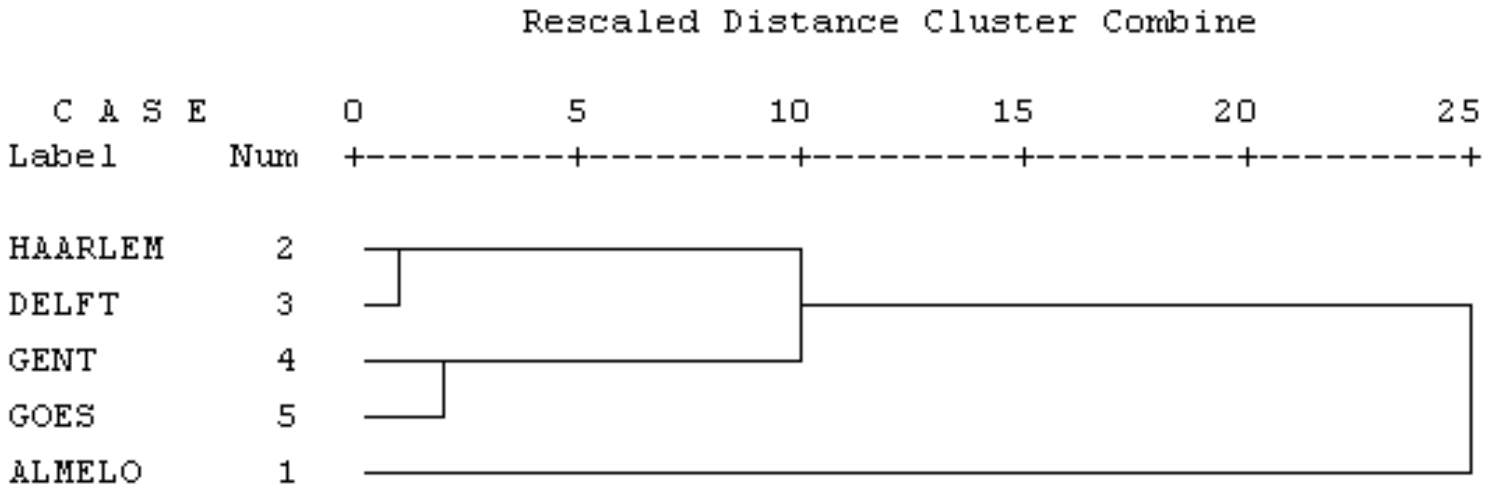
- New table:

	Almelo	Haarlem & Delft	Gent	Goes
Almelo	.000	.081	.069	.100
Haarlem & Delft	.081	.000	.057	.056
Gent	.069	.057	.000	.043
Goes	.100	.056	.043	.000

- Repeat until one cluster is left. In this example this is the cluster (((Haarlem & Delft), (Gent & Goes)), Almelo).

Dendrogram

- The same structure is also visualized by means of a dendrogram. In addition a dendrogram represents the distance between the clusters by the length of the branches.



Matrix updating algorithm

- When two clusters are fused, one larger cluster arises.
- The distances between this larger cluster and the other clusters need to be calculated. This is done by a $matrix\ updating\ algoritme$.
- SPSS has seven algorithms, we discuss five of them.
- Assume clusters i and j are fused to one cluster ij. In order to calculate the distance between cluster ij and a cluster k we need (partially) the following data:

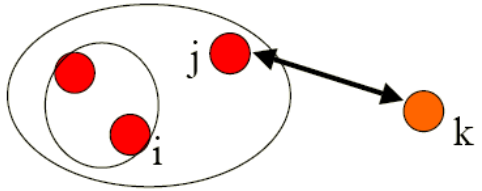
n_i: number of varieties in cluster i;	d_{ki}: the distance between k and i;
n_j: number of varieties in cluster j;	d_{kj}: the distance between k and j;
n_k: number of varieties in cluster k;	d_{ij}: the distance between i and j;

Nearest neighbor

Single-link

- Choose the smallest distance:

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



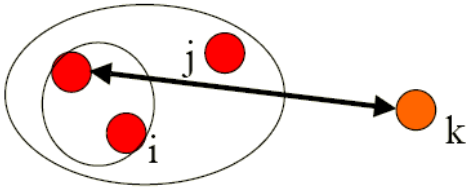
- Outliers are clearly recognizable shown.

Furthest neighbor

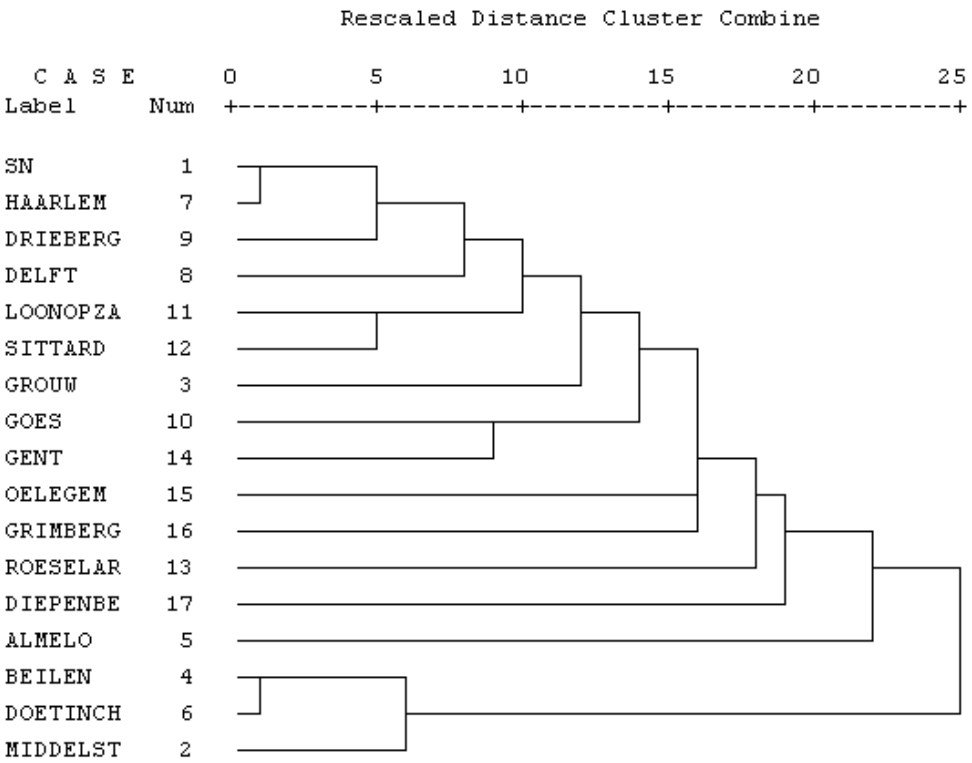
Complete-link

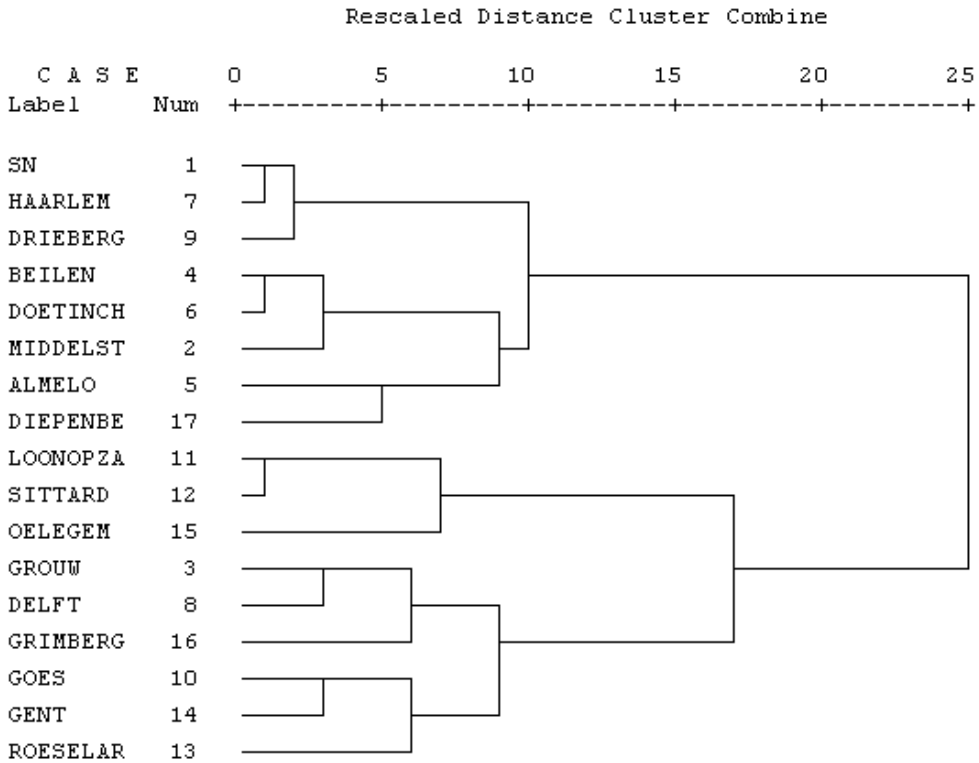
- Choose the largest distance:

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



- Gives a well-belanced dendrogram, clusters have about the same size.





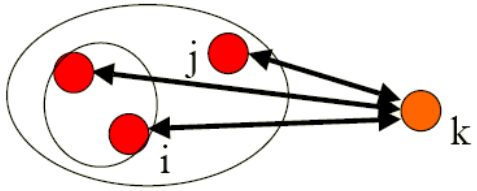
The dendrogram on the left is obtained with *nearest neighbor*, and the one on the right with *furthest neighbor*.

Between-groups linkage

Unweighted Pair Group Method using Arithmetic averages (UPGMA)

- Choose the average distance between all varieties in the two clusters:

$$d_{k[ij]} = \begin{array}{lllll} (n_i & / & (n_i + n_j)) & \times & d_{ki} \\ (n_j & / & (n_i + n_j)) & \times & d_{kj} \end{array} +$$



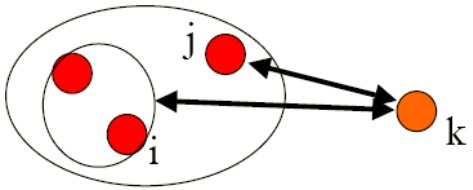
- Number of elements per cluster has influence.
- Dendrograms obtained by this method reflect the original distances in the distance matrix most closely (cophenetic correlation coefficient).

Within-groups linkage

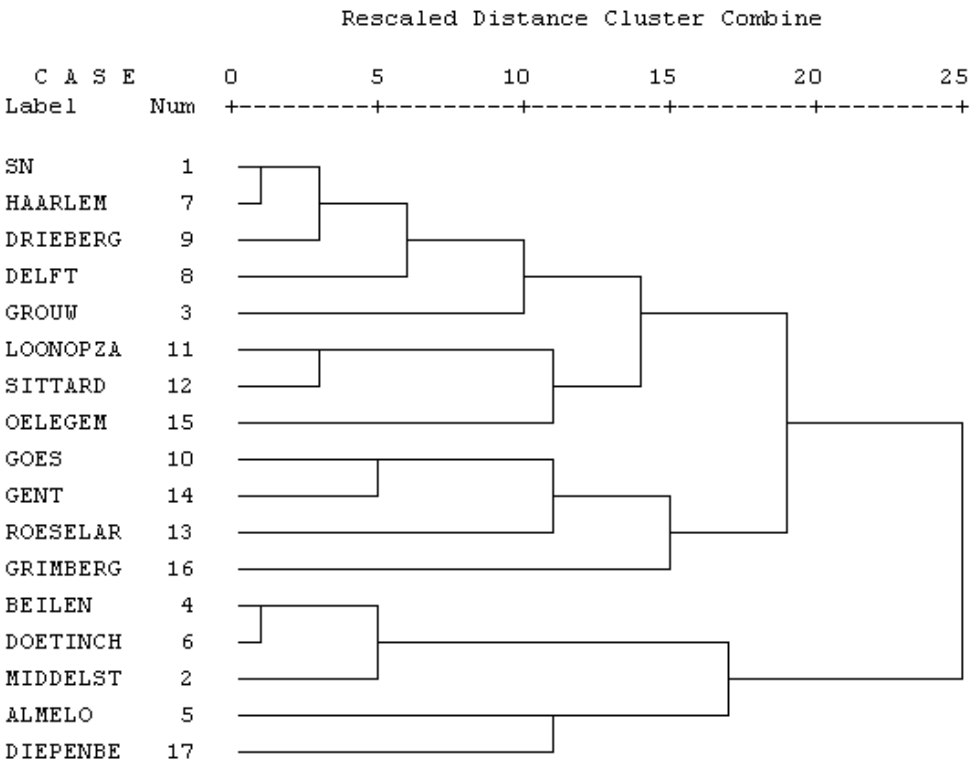
Weighted Pair Group Method using Arithmetic averages (WPGMA)

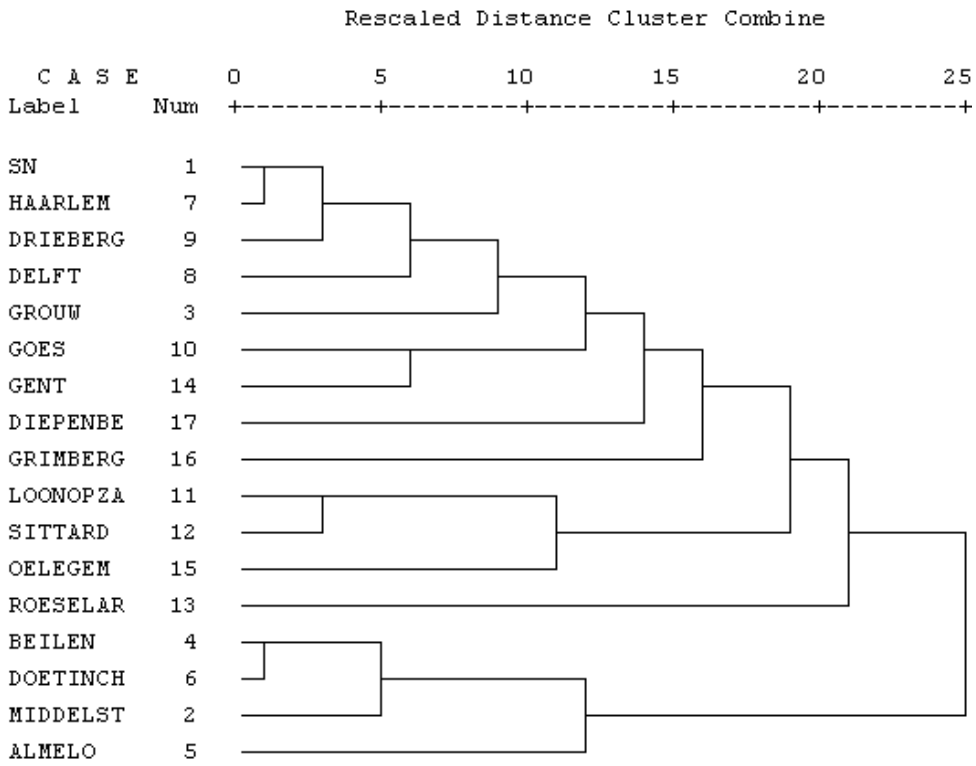
- Choose the average distance between the two clusters:

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



- Number of elements per cluster does not play a role.
- Is recommended in case of an irregular sampling distribution.





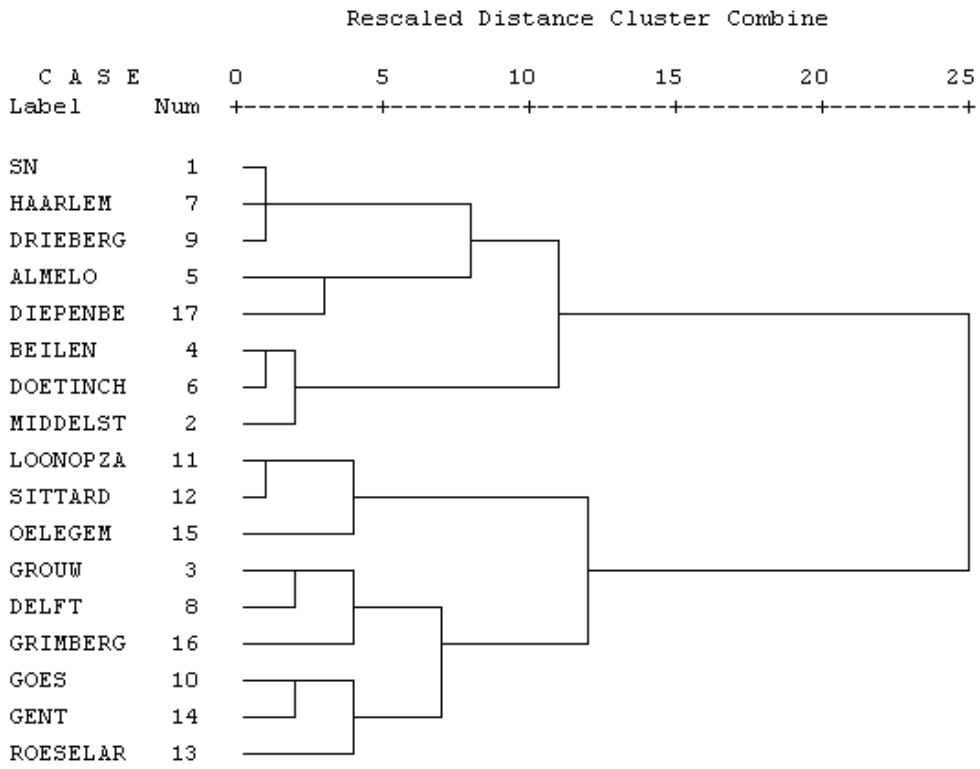
The dendrogram on the left is obtained on the basis of *between-groups linkage*, and the one on the right is obtained on the basis of *within-group linkage*.

Ward's method

- Minimize the variance in the clusters:

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

- Results in a well-belanced dendrogram, all clusters have about the same size.



Dendrogram obtained on the basis of *Ward's method*.