8 Similarity, Scaling, Cluster Analysis

8.1 Dissimilarities

a **Similar Patterns.** Often we want to study observation units on the basis of their similarities. **Similarity** or dissimilarity should therefore be determined on the basis of several or many variables, i.e. multivariately.

Examples are:

- Vegetation: Sample surfaces are similar and thus represent the same ecosystem if the species composition is similar.
- Plant types are similar if they frequently occur at the same locations...
- ... or if they developed evolutionarily from the same type a “shorter” time ago.
- Gene sequences are similar if they have many sections that agree.
- Weather conditions are similar if the weather maps are the same – in which aspects?
- Referenda (voting issues) can be assessed for similarity on the basis of individual results in cantons or municipalities.
- **Customer Management:** Which customers display similar buying behavior?
- Which products are often bought with each other?
- **Pattern Recognition:** Letters are similar if they are often confused for each other.
- Archeological “artifacts” come from similar cultures and time periods if they have characteristic features that are similar.
- Medieval manuscripts or other texts have, on the basis of common sources and through transcription, a kind of family tree, which can be developed through similarities. Anonymous texts can be recognized as similar through style features; similar texts may be attributed to the same authorship.

Generally, it’s *patterns* that are characterized as similar or dissimilar.

b **Goal.** A goal is often the formation of groups of similar objects, locations or individuals. This is the theme of **cluster analysis**.

However, it can also be useful to find a graphical representation through points, in which strong similarities in the observed individuals are reflected by small distances between points and weak similarities through large distances. Such representations give the **scaling methods**.

c **Euclidean Distance.** The most obvious type of dissimilarity to measure is the usual distance – the so-called Euclidean distance – between points. In the case of only two variables between the points in a scatter plot, it can be seen directly; in the case of several variables it is the distance between points in multidimensional space. This measurement is especially sensible for quantitative variables.

The distance d_{hi} between two observed individuals h and i with the variable values $\underline{x}_h$ and $\underline{x}_i$ is the root of

$$d_{hi}^2 = \|\underline{x}_h - \underline{x}_i\|^2 = \sum\nolimits_{j=1}^{m} \left(x_h^{(j)} - x_i^{(j)} \right)^2 .$$

If variables with different units and scatter are calculated with each other, then naturally the differences for those variables that display a large range of values dominate. It is therefore usually necessary to **standardize the data** first, so that all variables have the same scatter.

d ▷ As an illustration, we take a small excerpt of the data of the vegetation study, namely the numbers of the species Nardus stricta (Nardstri), Calluna vulgaris (Caluvulg), Festuca rubra (Festrubr), Deschampsia flexuosa (Descflex) and Agrostis capillaris (Agrotenu) for 9 observations (those in the eastern part, on the Alpboden). They are recorded in Table 8.1.d .

	Nardstri	Caluvulg	Festrubr	Descflex	Agrotenu
54	40	0	10	6	0
58	33	0	5	8	4
59	35	0	5	3	0
60	30	25	0	3	0
64	0	0	13	0	38
66	0	0	20	0	40
67	0	0	10	0	12
68	3	0	13	6	0
70	2	0	0	0	2

Table 8.1.d: Example Data

For the first two the difference is $\sqrt{}((33-40)^2 + 0 + (5-10)^2 + (6-8)^2 + 4^2) = 9.7$. The first four rows give the following distance table:

```
> dist(t.d)
      54    58    59
58  9.70
59  7.68  6.71
60 28.88 26.46 25.98
```

This dissimilarity expresses most strongly the difference between the numbers of the species that cover the largest number range. It is therefore reasonable to standardize the variables from the start. For numbers we can seek equal mean values. We thus divide each column by its mean value. We then get

```
> t.mn <- apply(t.d, 2, mean)
> t.dt <- sweep(t.d, 2, t.mn, "/")
> dist(t.dt[1:4,])
     54   58   59
58 1.08
59 1.24 1.78
60 9.16 9.19 9.02
```

For continuous variables it is often reasonable to transform them (see 8.3.c) and then standardize by variance (or median-deviations). ◁

e **Manhattan Distance.** If two observations are strongly different with respect to one individual variable, then the Euclidean distance is large, even if the other variables correspond perfectly. In this sense, this measure of dissimilarity does not react robustly.

This property is essentially attenuated if we *don't* square the differences for the individual variables, and the sum of the absolute values of the differences

$$d^*_{hi} = \sum\nolimits_{j=1}^{m} \left| x_h^{(j)} - x_i^{(j)} \right|$$

is used as a dissimilarity. We can say that this measures the path length that must be traveled to get from one place to another in a city with strict block structure. This distance is therefore called the **city block distance** or **Manhattan distance**. Mathematicians also call it the L_1 **distance**.

▷ The Manhattan distance of the unstandardized data gives

```
> dist(t.d[1:4,],method="manhattan")
   54 58 59
58 18
59 13 11
60 48 42 35
```

For count data, it is recommended for the formation of dissimilarities, like for other statistical methods, to first do a root transformation(8.3.c). For all 9 observations we then get the following table:

```
>   dist(sqrt(t.d),method="manhattan")
        54    58    59    60    64    66    67    68
58  3.89
59  2.05  3.27
60  9.73 10.60  7.67
64 15.38 14.11 15.18 21.98
66 16.41 15.13 16.21 23.01  1.03
67 12.24 10.96 12.04 18.84  3.14  4.17
68  5.04  7.76  6.27 13.07 10.35 11.37  8.09
70 11.94  9.98  9.88 12.21  9.77 10.80  6.63  7.79
```

◁

* The standardization of the variables by variance 1 also reacts strongly to outliers. It is obvious to also leave out the square here and divide the variables by their mean absolute deviation (mean deviation from the median)
$(1/n)\sum_i |x_i^{(j)} - \text{med}\langle x_i^{(j)}\rangle|$.

f **Common Elements.** In special applications there are other obvious measures for similarity or dissimilarity. For example, determine a similarity for a plot (of land) on the basis of the occurrence of plant species. The variable $x^{(j)}$ is thus a two-valued variable, that indicates the occurrence of species j. It is clear that for each pair h, i of plots, there is a two by two cross table in which species are classified into those occuring in both plots (a), those not occuring in any (d) and the two discordant cases (b and c), see Table 8.1.f. (Note that this table is not appropriate for a test of indendence: the species can hardly be considered as random, so the hypothesis of independence already doesn't make sense.)

Plots			
h	i		
	occurrence	absence	total
occurrence	a	b	$a+b$
absence	c	d	$c+d$
total	$a+c$	$b+d$	$m=a+b+c+d$

Table 8.1.f: Two by two table of the occurrences of the m species on two plots: descriptions.

With these four numbers are defined various similarity measures. The simplest are the **simple matching coefficient** and the **Jaccard coefficient**,

$$s_{hi}^{(s)} = \frac{a+d}{a+b+c+d} \qquad \text{and} \qquad s_{hi}^{(J)} = \frac{a}{a+b+c}\ ,$$

The second assumes that some (many) types that don't occur on both plots say nothing about the similarity of the plots.

g **Directly Determined Dissimilarity.** Sometimes, dissimilarities aren't given on the basis of variables $x^{(j)}$, but directly determined. One can, for example, count times that letters are confused in automatic detection or assess similarities via experts or consumers of products.

We'll come back to possible specifications of similarity and dissimilarity as soon as we have introduced a potential application.

h **(Dis-) Similarity Matrix.** Every definition of similarity s_{hi} or dissimilarity d_{hi} between objects (observations) leads to a symmetric $n \times n$ matrix. This forms the starting point for the following application and for most procedures in cluster analysis that we will discuss in section 8.4.

8.2 Multidimensional Scaling.

a **Basic Idea.** We seek an arrangement of points $\underline{z}_i$ in the plane(or eventually in higher dimensional space) which correspond to the observations (objects, plots), so that the Euclidean distances between points reflects their dissimilarity as exactly as possible.

As soon as we have determined what "as exactly as possible" should mean, an optimization program can determine such an arrangement. This is indeed a task that is not really solvable, since it is a large nonlinear optimization with many variables (the $2n$ coordinates of the desired points). However, with appropriate programs we can get a proposed solution of reasonably good quality, which we cannot guarantee to be the "global" optimum.

b **PCA.** This goal is related to the usual application of Principal Component Analysis, when we use the first two components as a "most informative" representation of the observations in two dimensions. There, we restricted the points $\underline{z}_i$ to being the result of an orthogonal transformation of the original variables $x_i^{(j)}$, and used the sum of the variances as the criterion. Here, we drop the restriction and use another criterion. We come back to this relation at the end of the Section.

c Since only the distances in this representation require optimality, once a solution is arrived at it can be rotated and mirrored without being worsened. So, if we want to compare two representations, we must begin with another procedure to bring the two into congruence as well as possible. An obvious solution is to apply the principal component transformation to a first result to make the solution unique (up to mirror images).

d **Stress.** An obvious measurement for the difference is again a squared sum, which equals

$$Q\langle \underline{z}_1, ..., \underline{z}_n\rangle = c \sum_{h,i} \left(g\langle d_{hi}\rangle - \|\underline{z}_h - \underline{z}_i\|\right)^2$$

with g being the identity function. More generally, g is allowed to be any monotone function and will be determined so that the sum of squares is minimized. The reason for this function is to take into consideration that the exact value for the dissimilarity is unimportant and only the ranking should count.

In general, it is useful to weight this sum differently for smaller and larger dissimilarities. We can then achieve a good representation of either the smaller or the larger dissimilarities. In the second case, for larger data sets we have less hope of success.

e ▷ Fig. 8.2.e (i) shows the result for the small data example of the 9 observations from the vegetation study, using the Manhattan distance of the root-transformed data (8.1.e). In Fig. 8.2.e (ii) are compared the distance and the represented dissimilarity. For the point pair $[68, 58]$ the dissimilarity is much larger (7.76) than the distance in the representation (5.40); for the pair $[68, 54]$ it is much smaller (5.04) than the distance (7.10).

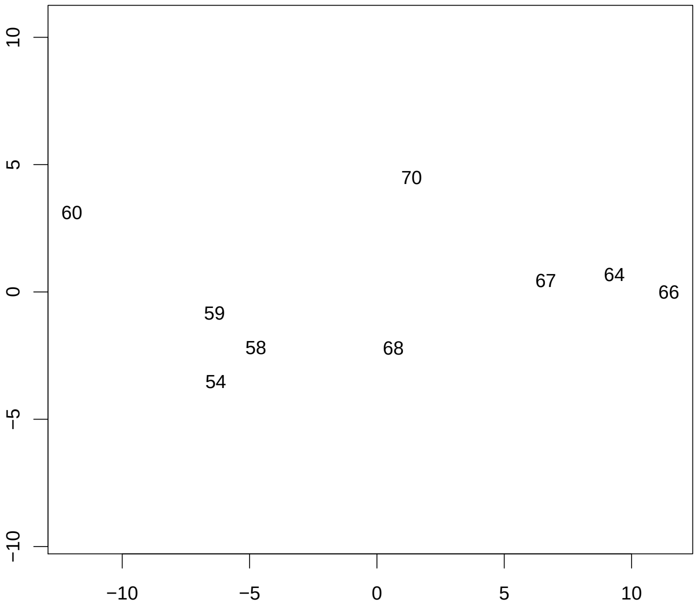

Figure 8.2.e (i): Multidimensional Scaling for the Small Example Dataset

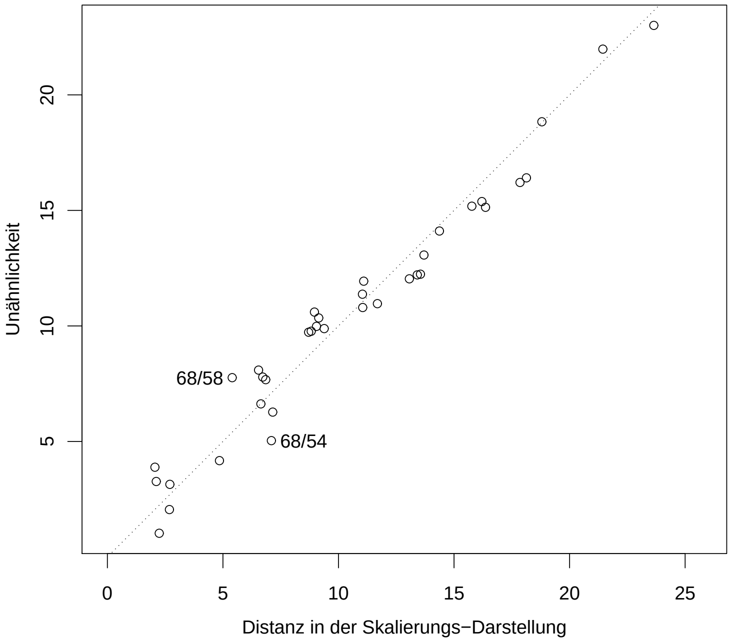

Figure 8.2.e (ii): Comparison of the distance in the multidimensional scaling with the previously given dissimilarities in the small example. Two point pairs for which the agreement is not good are marked.

f ▷ For the whole dataset from the vegetation study we get the representation in Fig. 8.2.f. The 9 observations that also appear in the small example are highlighted to facilitate a comparison of their positions with their previous representation. ◁

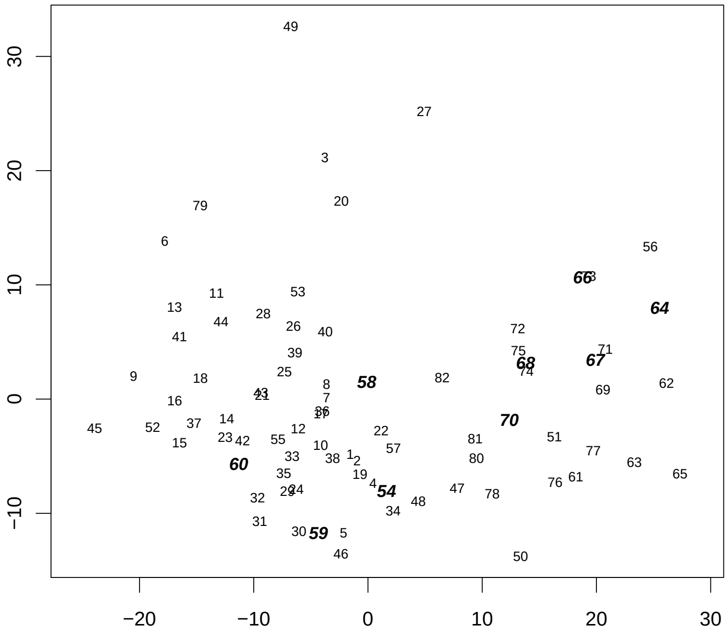

Figure 8.2.f: Multidimensional Scaling with Euclidean distance for the species variables in the vegetation study example.

g **Comparison with Principal Components.** As mentioned above (8.2.b), PCA is used for the same purpose as Multidimensional Scaling. How do these methods compare?

The advantage of multidimensional scaling is that we make fewer restrictions.

- Most importantly, the coordinates that we need for the representation need not depend linearly on the input variables $X^{(j)}$.
- In the stress measurement also, only good agreement of the distances in the representation space corresponds with a monotone transformed version of the dissimilarity.
- Finally, we need no X variables for the multidimensional scaling; a dissimilarity matrix that is obtained is any way is enough as a starting point.

Multidimensional scaling also has a significant disadvantage: It determines a representation only for the existing observations. If new observations are added it is not clear where they would be drawn into the existing figure. In principal components this is clear: We can also calculate the coordinates for new observations without completely repeating the analysis. Therefore, even new observations can be associated with visually determined groups of the pre-existing observations, which with multidimensional scaling is only possible with a repeated analysis, which can go wrong if the new representation is too strongly different from the old.

8.3 Further Considerations about Dissimilarities

a In forming a measurement of dissimilarity, it is worthwhile to use a lot of care, since success (whatever that means) is most dependent on this measure. It should reflect as well as possible appropriate intuition about (dis-) similarity for the data and the goals of the evaluation.

b **Dissimilarity as a Sum of Dissimilarities for Individual Variables.** Most meaningful measures of dissimilarity arise as a sum of contributions of the individual variables – possibly as a weighted sum or a (weighted) mean of the contributions,

$$d\langle \underline{x}_h, \underline{x}_i \rangle = \frac{\sum_j w_j \, d^{(j)} \langle x_h^{(j)}, x_i^{(j)} \rangle}{\sum_j w_j} .$$

In comparison to sums, mean values have the advantage that missing values are treated reasonably, in that the corresponding terms (above and below in the fraction) are left out.

A bit more flexible is the following formula, that works with transformed contributions $d^{(j)}$:

$$\ell \langle d \langle \underline{x}_h, \underline{x}_i \rangle \rangle = \frac{\sum_j w_j \, \ell \Big\langle d^{(j)} \Big\langle x_h^{(j)}, x_i^{(j)} \Big\rangle \Big\rangle}{\sum_j w_j} .$$

If $\ell\langle d \rangle = d^2$ and $d^{(j)} \Big\langle x_h^{(j)}, x_i^{(j)} \Big\rangle = |x_h^{(j)} - x_i^{(j)}|$, for equal weights we get the usual Euclidean distance, up to the factor $1/\sqrt{m}$.

c So, underlying the total dissimilarity is the definition of the dissimilarities $d^{(j)} \Big\langle x_h^{(j)}, x_i^{(j)} \Big\rangle$ for individual variables.

For **continuous variables** we choose as dissimilarity the absolute value of the difference of the values. In order to achieve that the same differences really indicate the same dissimilarities, so we often have to start by **transforming the variables**. If, for example, the number of individuals on a plot is counted, the difference between 0 and 2 means something totally different than the difference between 10 and 12.

A reasonable proposal for a transformation is usually

- the log transformation of **concentrations and amounts** – so for continuous variables that can only take on positive values –
- the root transformation for **count data** and
- the so genannte arcsine transformation $\ell\langle x \rangle = \arcsin \sqrt{x}$ for **Proportions** (Percentages/100).

J. W. Tukey named these transformations the **first aid transformations** and **for such data should always be applied** if there are no contraindications – but, in relation to dissimilarities, it is not forbidden to use another transformation if it causes **equal differences of the transformed variables to mean equal dissimilarities**.

In summary, for the dissimilarity of a single variable $x^{(j)}$ we can write

$$d^{(j)} \Big\langle x_h^{(j)}, x_i^{(j)} \Big\rangle = \left| g^{(j)} \langle x_h^{(j)} \rangle - g^{(j)} \langle x_i^{(j)} \rangle \right| .$$

d The most obvious **dissimilarity for ordered variables** is the absolute difference of the ranks. This definition follows the just mentioned scheme of transformation and differencing. The rank-difference of the observations of the observations h and i, if all values are different, is the number of observations with values between $x_h^{(j)}$ and $x_i^{(j)}$, plus 1. It thus depends on all of the observations. This idea can also be applied for quantitatively interpretable variables (previous case).

e For **binary variables** it is obvious to only differentiate between agreement ($x_h^{(j)}$ and $x_i^{(j)}$ both 0 or both 1) and disagreement. The dissimilarity then only takes the values 0 and 1 and we can write these as for quantitative variables: $d^{(j)}\left\langle x_h^{(j)}, x_i^{(j)} \right\rangle = |x_h^{(j)} - x_i^{(j)}|$.

However, "positive agreement" (species occurring on both plots) often means more similarity than negative agreement. Then, we should set $d^{(j)}\langle 0, 0\rangle \neq 0$. In this case we can also reduce the weight w_j – though this violates the scheme introduced above (8.3.b), in which w_j does not depend on the $x^{(j)}$ values.

f Finally, for nominal or **categorial variables** we can set $d^{(j)}\langle x_h^{(j)}, x_i^{(j)}\rangle = 0$ as for binary variables, if $x_h^{(j)}$ and $x_i^{(j)}$ agree, and =1 if they disagree. If we want to differentiate, here there are many possibilities to take into account the frequency of the possible values and their similarities.

g **Standardization and Weighting of Variables.** Continuous variables usually have different units. So that in the definition of dissimiliarity (8.3.b) they have, in some sense, "equal weight" we standardize them to get equal (robust) standard deviation $\widehat{\sigma}_j$. This is equivalent to setting $w_j = 1/\widehat{\sigma}_j$ in formula 8.3.b.

h ▷ In the example, all of the recommendations lead to first taking the root of the numbers of individuals of the species, then standardizing them by the mean absolute deviation and finally taking the Manhattan distance. We thus get

```
      1     2     3     4
2  2.60
3 11.37 13.49
4  4.14  3.41 13.35
5  4.63  3.90 11.70  1.65
```

(The root transformation does little in this small example.)

For the case of numbers of individuals we can also do without the standardization. We then give the uncommon species little influence on the dissimilarity – their number of individuals is indeed less reliably determined. ◁

i **Explicit Weighting.** It can be very useful to explicitly choose a weighting in order to define a meaningful dissimilarity. A strong statement for this point is formulated by Anderberg (1973, p.13):

"Some investigators recommend reducing all variables to standard form (zero mean and unit variance) at the outset. Such suggestions simplify the mechanics of analysis but constitute a complete abdication of the analyst's responsibilities and prerogatives to a mindless procedure."

j **Principal Components.** An unequal weighting of variables seems to be suitable if certain variables are strongly correlated. We have seen principal component analysis, which makes uncorrelated variables out of any given variables. We may therefore consider to use the Euclidean distance on principal component coordinates. However, this is only effective if we limit ourselves to a few principal components. If we use all m, the distance doesn't change; this was indeed an important property of principal component analysis.

Neglecting the last principal components, however, leads to dropping "unique" variables, that are only weakly correlated with all others. Whether this is desired must be carefully considered. Maybe the unique variables are also important variables that should have a contribution to the dissimilarity.

k **Mahalanobis Distance.** A distance measurement that takes into consideration correlations between variables is the Mahalanobis distance, that is based on the covariance matrix of all observations

$$d\langle \underline{x}_h, \underline{x}_i \rangle = (\underline{x}_h - \underline{x}_i)^T \widehat{\underline{\Sigma}}^{-1} (\underline{x}_h - \underline{x}_i) \; .$$

At first look, it appears very appropriate. The following reflections, however, lead to some thought: If we initially perform a principal component analysis and then standardize the obtained values (scores) to have variance 1, then the usual Euclidean distance is the same as the Mahalanobis distance of the original observations. Here the principal components with small variances (small eigenvalues) are thus not neglected, but are, on the contrary,"inflated" to the same variance as the first principal component. This is usually not reasonable.

l A more reasonable way of dealing with**strongly correlated variables** is to replace groups of two or more such variables with a single one, chosen out of the group or newly defined – for example the sum, the mean value, or some other "index".

m **Similarities.** Here the definition of *dissimilarity* has been discussed in detail. More intuitive is to talk about similarities. The previous analysis is done for dissimilarities; to transfer it to similarities seems hard.

However, from every dissimilarity, we can get a similarity, for example by

$$s_{hi} = 1/(1 + d_{hi}) \; .$$

The result then lies between 0 and 1 – which is common for a measure of similarity. Conversely, from a similarity s we can get a dissimilarity by $d = 1/s - 1$ or, more simply, by $d = 1 - s$.

n **Similarities of Variables.** We can also introduce (dis)similarities among variables in order to represent them graphically or to distribute them into clusters. The prototype of similarity of variables is their correlation. It is

$$s\left\langle \underline{x}^{(j)}, \underline{x}^{(k)} \right\rangle = \frac{1}{n-1} \sum\nolimits_i x_i^{(j)} x_i^{(k)} \; ,$$

if the variables are standardized. It is then

$$d\left\langle \underline{x}^{(j)}, \underline{x}^{(k)} \right\rangle \;=\; \frac{1}{n-1} \sum\nolimits_i (x_i^{(j)} - x_i^{(k)})^2$$

$$
\begin{aligned}
&= \frac{1}{n-1}\sum\nolimits_i (x_i^{(j)})^2 + \frac{1}{n-1}\sum\nolimits_i (x_i^{(k)})^2 - \frac{2}{n-1}\sum\nolimits_i x_i^{(j)} x_i^{(k)} \\
&= 2 - 2s\Big\langle \underline{x}^{(j)}, \underline{x}^{(k)} \Big\rangle \ .
\end{aligned}
$$

In use, the absolute value of the correlation can be more reasonable, since a negative correlation also indicates a “close relationship”.

▷ The representation for the species in the vegetation study is shown in Fig. 8.3.n. ◁

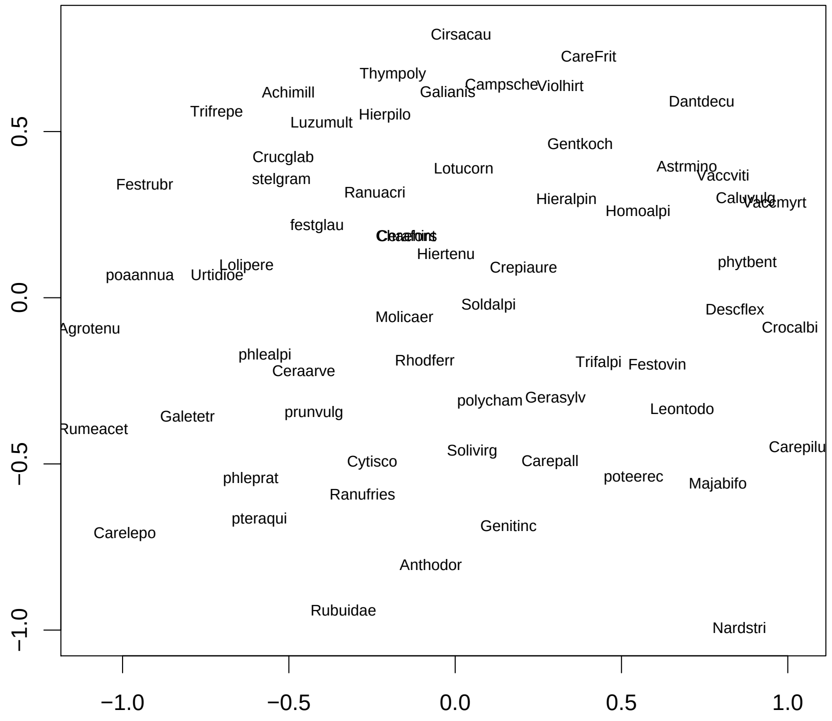

Figure 8.3.n: Multidimensional Scaling for the Species in the Vegetation Study on the Basis of their Correlations

o In summary: **In the definition of the dissimilarity, as much intuition or precise knowledge about the variables as possible should be included.**

8.4 Cluster Analysis: Optimal Partitions

a **Basic Ideas.** We want a division of the observations into homogeneous groups, i.e. groups of observed units that are as similar as possible. A division into groups is called in mathematics a **partition**. The groups $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_q$ are subsets of the total set $\mathcal{S}$. A partition is characterized such that the union of all $\mathcal{G}_k$ covers the whole original set $\cup_{k=1}^{q} \mathcal{G}_k = \mathcal{S}$ and the $\mathcal{G}_k$ exclude each other, $\mathcal{G}_k \cap \mathcal{G}_\ell = \emptyset$ for all $k \neq \ell$.

If we now ask for *homogeneous* groups, for "mathematically socialized" people the obvious guiding idea is to define a **quality measure** Q for the homogeneity of a partition into q groups and then find the partition with the best quality measure for the given q. If the quality measure is appropriate for comparing partitions with different group numbers, we can also optimize over q.

b How do we get a quality measure Q? As for *dis*similarities, it is easier to define an ***in*homogeneity or heterogeneity measure**. The total heterogeneity is defined as a sum of the heterogeneities of the clusters,

$$Q = \sum\nolimits_k h\langle \mathcal{G}_k \rangle \ ,$$

where the heterogeneity h is usually based on a dissimilarity definition for observations, as we have introduced in the last section.

There is a common simple way of following these ideas.

c **K-means.** For a cluster $\mathcal{G}$ we form the center or the "centroid"

$$\underline{x}_{\mathcal{G}} = \operatorname*{ave}_{i \in \mathcal{G}} \langle x_i \rangle$$

($\text{ave}_i\langle . \rangle$ denotes the mean value over the arguments, as $\sum_i$ denotes the sum.) The heterogeneity measure is then the sum of the squared distances between the observations of the cluster and its center,

$$h\langle \mathcal{G} \rangle = \sum\nolimits_{i \in \mathcal{G}} d\langle \underline{x}_i, \underline{x}_{\mathcal{G}} \rangle^{\,2} \ .$$

The search for a minimum of the corresponding quality measure $Q = \sum_k h\langle \mathcal{G}_k \rangle$ is called the **K-means algorithm**.

By replacing the mean value by the "componentwise median" in determining the centroid $\underline{x}_{\mathcal{G}} = [\text{med}_i\left\langle x_i^{(1)} \right\rangle, ..., \text{med}_i\left\langle x_i^{(m)} \right\rangle]^T$ and using the Manhattan distance in the definition of h, then we get the **K-medians** algorithm.

d **Optimization.** The optimization of such a criterium is really calculable only for very small datasets. In fact, for $n = 25$ objects and $q = 3$ clusters, there are 1.4×10^{11} partitionings. It is therefore impossible to evaluate all of them.

As usual in such cases, we content ourselves with a **"local optimization"**: We begin with a partitioning into k groups and define a change that leads to an improvement according to the criterion. This step is repeated until no more improvement can be achieved. Depending on the starting partition, different "local solutions" can result – with k-means and k-medians there are often many. It is therefore recommended to determine the local optimum with as many starting points as possible, then choose the solution that gives the best value for the critera. However, even this doesn't guarantee that there does not exist a better solution. (It remains: What I don't know doesn't hurt me...)

		k-means			
	A	B	C	D	total
1	0	3	0	4	7
Arti- 2	4	17	0	0	21
kel 3	6	0	22	2	30
4	2	0	0	22	24
total	12	20	22	28	82

Table 8.4.e: Comparison of the vegetation groups according to the article with the results of the k-means algorithm

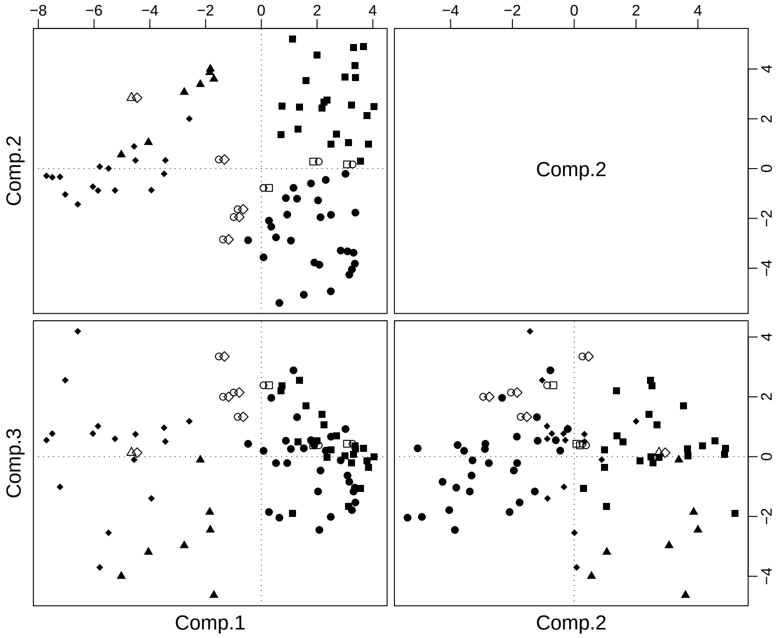

Figure 8.4.e (i): Grouping according to k-means and according to the article, represented in principal components

e ▷ In the **vegetation study example**, in the authors' analysis four groups were formed with the help of the vegetation data. The k-means algorithm with root-transformed, unstandardized individual count numbers for all 63 species and $k = 4$ gives clusters that are compared with the groups found in the article in Table 8.4.e .

We can assess the obtained grouping by showing the observations in principal components (Fig. 8.4.e (i)) or with multidimensional scaling (Fig. 8.4.e (ii)) and symbolizing the group membership. The figures show that the distances can be represented quite well in two dimensions – in any case the groups found are separated quite well in these dimensions.

In the article is also given a vegetation map that is based on a mapping independent of the study. Fig. 8.4.e (iii) shows that the partitioning derived from the data in the article does not agree entirely with the mapping. In particular, Groups 1 and 2 are very similar, and therefore their distinction is difficult – and probably also less important.

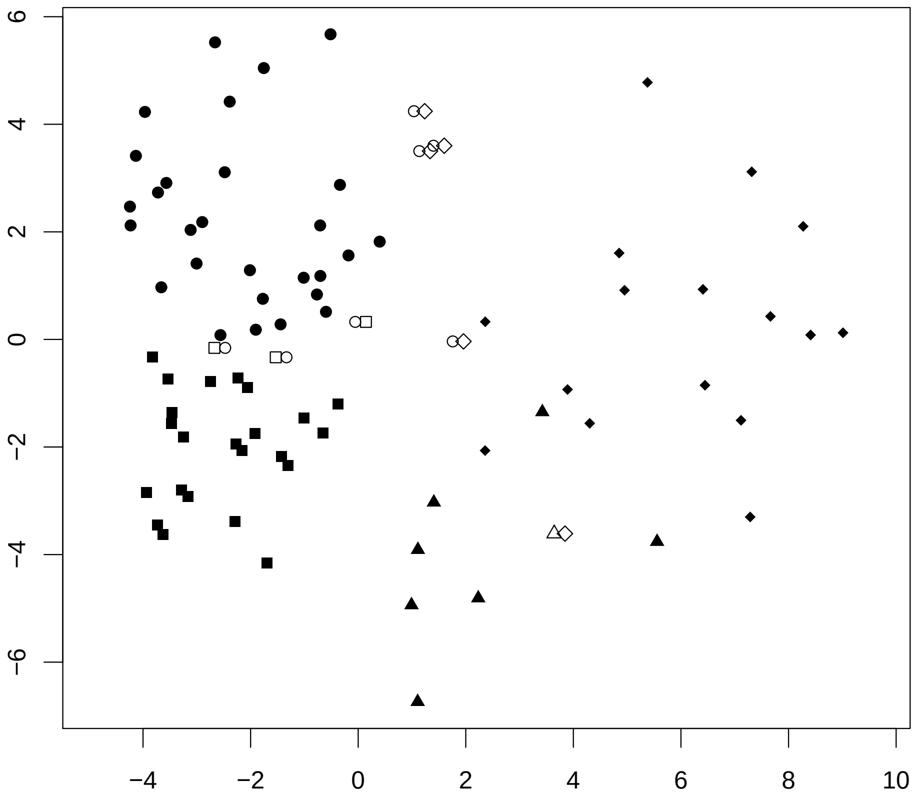

Figure 8.4.e (ii): Grouping according to k-means and according to the article, shown with multidimensional scaling

Fig. 8.4.e (iv) shows the grouping found with k-means. ◁

f **PAM.** Instead of allowing the centroids of the groups to be any point in space, we can restrict them to the observations. Then the grouping is defined by the choice of the centroids from the observations. We must therefore find the choice that optimizes the quality criterion for the corresponding grouping. This method is known as the **k-medoid** method. In R it is called PAM, corresponding to the promoter Kaufman and Rousseeuw (1990). Since no centers have to be calculated, it is on the one hand applicable if only distances are provided, and on the other hand it is applicable for larger datasets thanks to simpler calculation.

The algorithm implemented in the S function `pam` has two parts. In the first, the k observations that will serve as an initial solution are determined according to an ad-hoc method. In the second, iterative part, these are improved according to the criterion. – We can, as with k-means and k-median, study which solutions we get if we choose a random k observations as an initial solution multiple times and then determine the corresponding "local solutions".

▷ In the example, the use of Manhattan distances and PAM gives the grouping that is shown in the map Fig. 8.4.f. It seems to prove less here than k-means with Euclidean distances. ◁

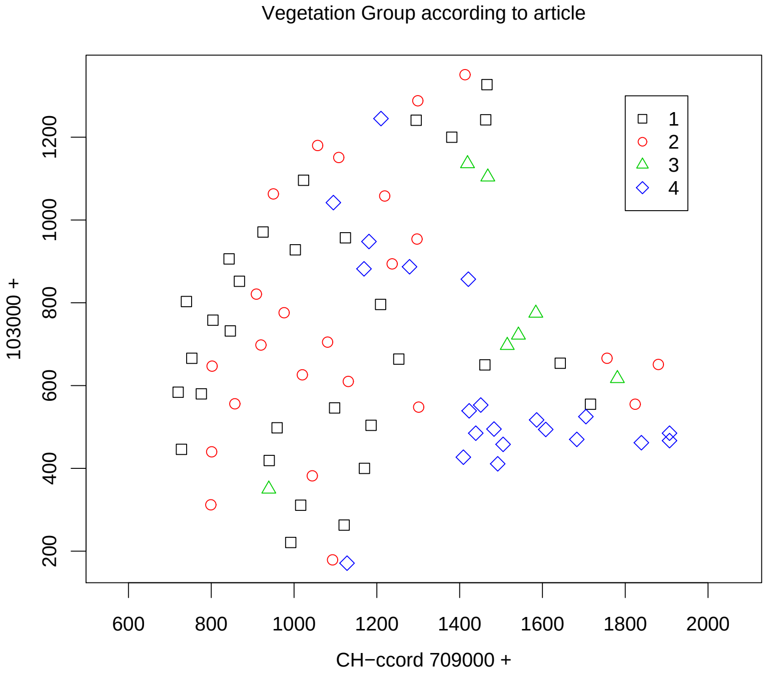

Figure 8.4.e (iii): Geographical Locations of the Plots with Grouping According to the Article

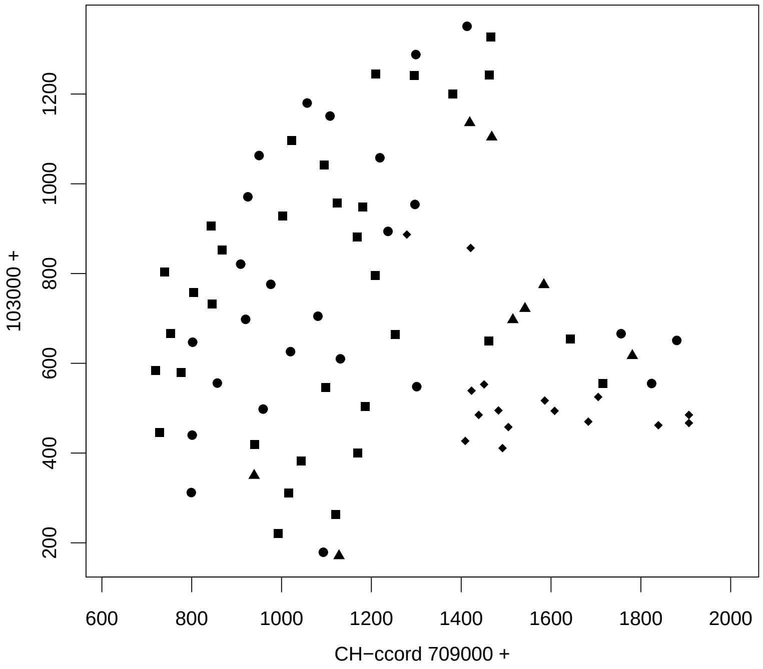

Figure 8.4.e (iv): Geographical Locations of the Plots with Grouping According to k-means

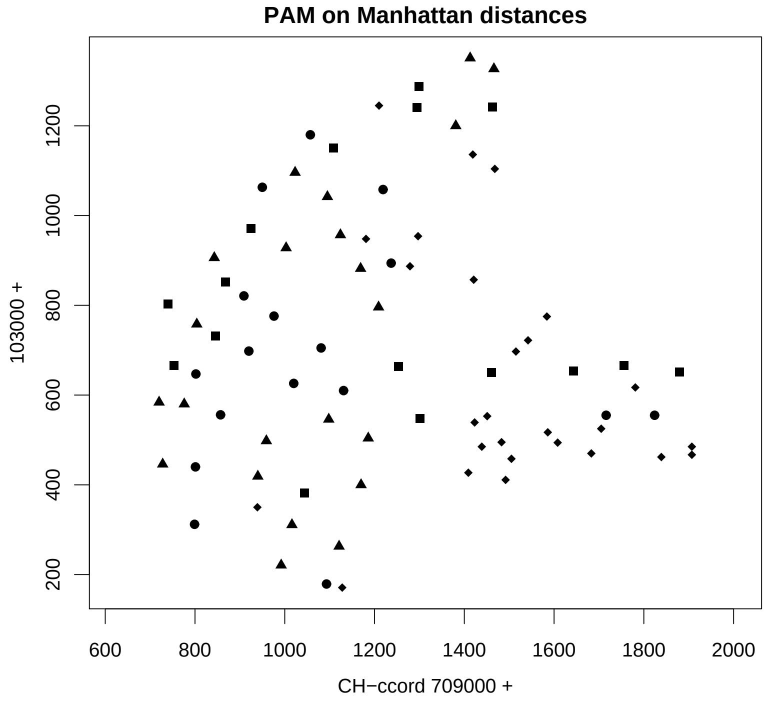

Figure 8.4.f (iv): Geographical Locations of the Plots with Grouping According to PAM with Manhattan Distances

g* It can be shown that also for the k-means method, no X values need to be known. We calculate

$$\begin{aligned}
\sum\nolimits_{h,i\in\mathcal{G}} d_{hi}^2 &= \sum\nolimits_{h,i\in\mathcal{G}} (\underline{x}_i - \underline{x}_h)^T(\underline{x}_i - \underline{x}_h) \\
&= \sum\nolimits_{h,i\in\mathcal{G}} \big((\underline{x}_i - \bar{x}_{\mathcal{G}}) - (\underline{x}_h - \bar{x}_{\mathcal{G}})\big)^T(...) \\
&= \sum\nolimits_{h,i\in\mathcal{G}} \big((\underline{x}_i - \bar{x}_{\mathcal{G}})^T(\underline{x}_i - \bar{x}_{\mathcal{G}}) - 2(\underline{x}_i - \bar{x}_{\mathcal{G}})^T(\underline{x}_h - \bar{x}_{\mathcal{G}}) + (\underline{x}_h - \bar{x}_{\mathcal{G}})^T(\underline{x}_h - \bar{x}_{\mathcal{G}})\big) \\
&= 2n_{\mathcal{G}} \sum\nolimits_{i\in\mathcal{G}} \big((\underline{x}_i - \bar{x}_{\mathcal{G}})^T(\underline{x}_i - \bar{x}_{\mathcal{G}}) - 2\sum\nolimits_{i\in\mathcal{G}}(\underline{x}_i - \bar{x}_{\mathcal{G}})^T \sum\nolimits_{h\in\mathcal{G}}(\underline{x}_h - \bar{x}_{\mathcal{G}})\big) \\
&= 2n_{\mathcal{G}} \sum\nolimits_{i\in\mathcal{G}} d(\underline{x}_i, \bar{x}_{\mathcal{G}})^2
\end{aligned}$$

Thus

$$h\langle\mathcal{G}\rangle = \sum\nolimits_{i\in\mathcal{G}} d\langle\underline{x}_i, \underline{x}_{\mathcal{G}}\rangle^{2} = \frac{1}{n_{\mathcal{G}}} \sum\nolimits_{h,i\in\mathcal{G}} d_{hi}^2 \,.$$

This formula can be applied if only the distances are known. For large datasets this is, however, not an advantage, since the distance matrix is then much larger than the data matrix. On the other hand, we can also take this formula for the determination of more possible heterogeneity measures, in which instead of d_{hi}^2 some other dissimilarity is used.

h **Silhouettes.** How well are the clusters separated? The procedure always gives a solution, and the question arises whether it consists of natural, well defined clusters or rather reflects more or less arbitrary borders drawn throughout the evenly distributed whole.

To answer this question, Kaufman and Rousseeuw (1990) have introduced variables that determine for each observation how clearly it belongs to "its" cluster, Let $\widetilde{d}\langle i,\mathcal{G}\rangle$

be the distance from object i to cluster $\mathcal{G}$ – in k-means and PAM the distance of the observation i to the center or centroid of the cluster $\mathcal{G}$. For this, $\mathcal{G}\langle i \rangle$ denotes the Cluster to which i belongs ($i \in \mathcal{G}$), and $\mathcal{B}\langle i \rangle = \arg\min_{\mathcal{G}|i \notin \mathcal{G}} \left\langle \widetilde{d}\langle i, \mathcal{G} \rangle \right\rangle$ the neighbor cluster of i. The silhouette value for i is then 1− the ratio of the dissimilarities of i to these two clusters

$$\widetilde{s}\langle i \rangle = 1 - \frac{\widetilde{d}\langle i, \mathcal{G}\langle i \rangle \rangle}{\widetilde{d}\langle i, \mathcal{B}\langle i \rangle \rangle}$$

(The original definition is here simplified for negative values.) The value will be 0 if the observation lies on the border between two groups.

The silhouette values laid out in a graphical representation against an appropriate arrangement of the i: The members of each cluster are drawn together and sorted in descending order of the silhouette value.

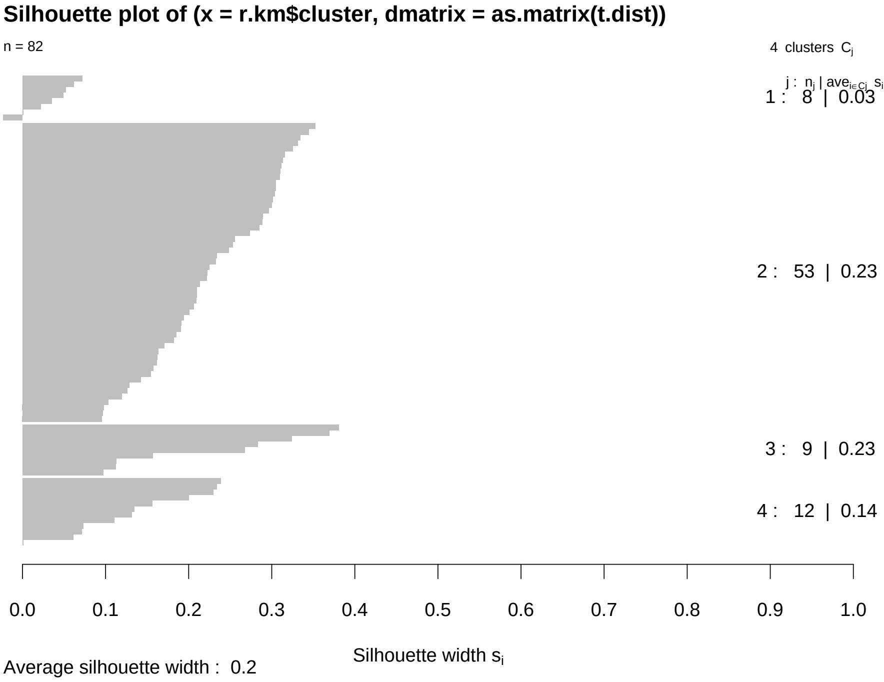

Figure 8.4.h: Silhouette for the k-means Groups in the Example

Fig. 8.4.h shows such a silhouette for our example. The clusters are very well defined; only a few observations show small silhouette values.

i For the **choice of the number of clusters**, ad-hoc ideas are reasonably used, which depend on the problem in the individual application. Silhouettes are also helpful here. We compare silhouettes for different q, which may show that raising q by 1 leads to the division of one of the clusters into two. That suggests well defined clusters.

j **Large Datasets.** Today, groupings in large datasets (n large) are often desired. Then, the calculation and storage of all dissimilarities is expensive and should be avoided.

A suggestion for an algorithm is called CLARA, which also comes from Kaufman and Rousseeuw (1990): It starts with a randomly chosen sample on which PAM is executed. Then, all objects are classed using the resulting centroids. This is repeated 5 times and finally the best of the 5 solutions is given as the result.

8.5 Hierarchical Methods, Dendrograms

a The so-called **agglomerative** procedures of cluster analysis are popular because, on the one hand, they are based on very simple calculation steps and therefore are already a half century old and, on the other hand, because they produce an interesting graphical representation in the form of a "dendrogram".

b ▷ The basis of a hierarchical procedure is a dissimilarity measure (or similarity measure) for the observations, which gives a corresponding dissimilarity matrix as a starting point. To illustrate the procedure we proceed from the dissimilarity table in 8.1.e.

Individuals that are similar should, if possible, be in one cluster. So, we begin by merging the two individuals h_1 and h_2 with the smallest dissimilarity into a "mini-cluster". In the example these are the observations 64 and 66 with a dissimilarity of 1.03.

We want to continue by looking for the next smallest dissimilarity value. To do this in a reasonable way, for each value we have to determine a dissimilarity between the mini-cluster $\{h_1, h_2\}$ and each of the remaining observed individual i. There are different variants for this that are discussed below. An obvious suggest is to choose the dissimilarity of the observation i to the more similar of the two "grouped observations" h_1 and h_2. In the example, observation number 70 has to 64 the dissimilarity 9.77 and to 66 the dissimilarity 10.80. So, its dissimilarity to the mini-cluster is 9.77.

In the dissimilarity matrix we replace the two rows with labels h_1 and h_2 with one with the values $d\langle\{h_1, h_2\}, i\rangle$, and do the same for the two columns with these labels. In the example, observation 64 is more similar to every other observation than 66. The new matrix is

```
>   dist(sqrt(t.d),method="manhattan")
      54    58    59    60    C1    67    68
58  3.89
59  2.05  3.27
60  9.73 10.60  7.67
C1 15.38 14.11 15.18 21.98
67 12.24 10.96 12.04 18.84  3.14
68  5.04  7.76  6.27 13.07 10.35  8.09
70 11.94  9.98  9.88 12.21  9.77  6.63  7.79
```

Now in the new matrix we can again determine the smallest dissimilarity values and combine the corresponding elements to a cluster. In the example the observations 54 and 59 have dissimilarity 2.05 and are a new mini-cluster.

In the further steps, not only observed individuals are combined into new mini-clusters, but also individuals to existing clusters and clusters to each other. Next in the example, the first mini-cluster (64 and 66) is merged with observation 67. ◁

c **Scheme.** This procedure can be summarized in an algorithm:

0 Begin with the “partition” in which every observed individual is a cluster.

1 Merge the two clusters with the smallest dissimilarity into one cluster.

2 Calculate the dissimilarity of the new clusters with each remaining cluster. Different formulas for this step, so-called **update-**formulas – siehe 8.5.g – lead to the different methods of agglomerative cluster analysis.

(It) Repeat 1 and 2 until all objects are merged into one cluster.

The result is a **hierarchy**, i.e. a series of “nested” partitions. For every association step ℓ there is an **index** value d_ℓ – the value of the “smallest dissimilarity”.

d **Dendrogram.** All steps of this procedure can be recorded graphically in a dendrogram. Lowermost in Fig. 8.5.d (i) the two observations that were first merged in the example, namely 64 ad 66, are “roots”, which, at a height of 1.03, merge into one root – into the first mini-cluster. Later, at height 3.14, observation 67 joins this “merged root”. The value 3.14 was the dissimilarity between observation 67 and the cluster $\{64, 66\}$.

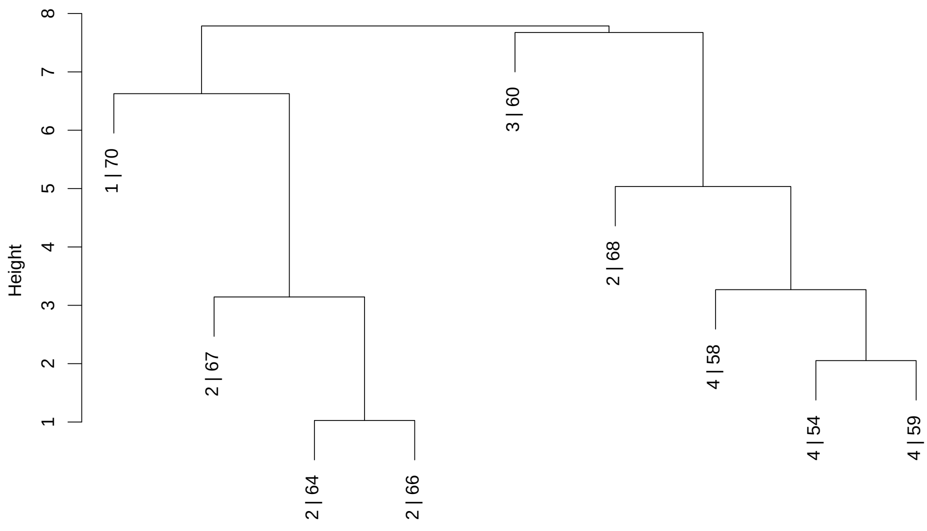

Figure 8.5.d (i): Dendrogram for the 9 Observation Example

The sequence for the objects to be used on the horizontal axis is not obvious. The same hierarchy can be represented by dendrograms that appear very different, as Fig. 8.5.d (ii) shows. The freedom of choice corresponds to a “Mobile”, in which decoration objects hang on a kind of hierarchy of hanging balances from a ceiling.

e **Groups Assignment.** Cluster analysis has the goal to form groups of similar observations. A dendrogram shows many possible assignments. If we remember how the underlying process works, it is clear that each possible number of clusters is achieved in a particular step. At the beginning there are n existing “clusters” , at the end only one, and if we stop after k steps, we have $n-k$ clusters. Thus, cutting the dendrogram at a certain height produces groups of disconnected small dendrograms. The related roots then form a cluster. If we cut the shown dendrogram at height 4, then we have the groups $\{64, 66, 67\}$ and $\{54, 59, 58\}$; the three remaining observations each form another cluster for themselves, so all together $q = 5$ clusters.

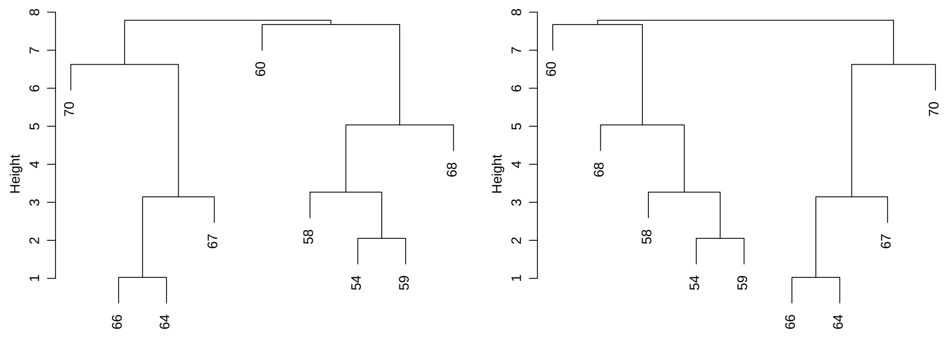

Figure 8.5.d (ii): Two More Dendrograms for the Example, of the Same Significance

f **Dissimilarity Measures Between Clusters.** To determine the dissimilarity between clusters (and between individual observations and clusters), in the example we have used a simple rule that corresponds to the first of the following dissimilarity measures. There are two more useful measures of this type. The definitions are:

- single linkage (nearest neighbor): $d\langle \mathcal{G}_1, \mathcal{G}_2 \rangle = \min\langle d\langle i_1, i_2 \rangle \mid i_1 \in \mathcal{G}_1, i_2 \in \mathcal{G}_2 \rangle$
- complete linkage (farthest neighbor): $d\langle \mathcal{G}_1, \mathcal{G}_2 \rangle = \max\langle d\langle i_1, i_2 \rangle \mid i_1 \in \mathcal{G}_1, i_2 \in \mathcal{G}_2 \rangle$
- average linkage: $d\langle \mathcal{G}_1, \mathcal{G}_2 \rangle = \text{ave}\langle d\langle i_1, i_2 \rangle \mid i_1 \in \mathcal{G}_1, i_2 \in \mathcal{G}_2 \rangle$
- * Another obvious measure for the dissimilarity between clusters would be the distance between their mean points. This turns out to be inappropriate for forming dendrograms, see 8.5.h.

g **Updating.** In step 2 of the basic scheme 8.5.c, dissimilarities between a newly created cluster and the each of the already existing ones must be calculated. For the stated dissimilarity measures, this can be calculated from the previous:

- single linkage:
$$d\langle \mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{H} \rangle = \min\langle d\langle \mathcal{G}_1, \mathcal{H} \rangle , d\langle \mathcal{G}_2, \mathcal{H} \rangle \rangle$$
- complete linkage: analogously
- average linkage:
$$d\langle \mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{H} \rangle = \frac{n_1 d\langle \mathcal{G}_1, \mathcal{H} \rangle + n_2 d\langle \mathcal{G}_2, \mathcal{H} \rangle}{n_1 + n_2}$$
- * Generally:
$$d\langle \mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{H} \rangle = \text{Funktion}\langle d\langle \mathcal{G}_1, \mathcal{H} \rangle , d\langle \mathcal{G}_2, \mathcal{H} \rangle , d\langle \mathcal{G}_1, \mathcal{G}_2 \rangle , n_1, n_2, n_H \rangle$$

With this formula we can even obtain the mean points distance via "updating".

h It makes sense to respect the fact that the new dissimilarity $d\langle\mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{H}\rangle$ is at least as large as the smaller of the two initial values $d\langle\mathcal{G}_1, \mathcal{H}\rangle$ and $d\langle\mathcal{G}_2, \mathcal{H}\rangle$ – otherwise "convoluted" dendrograms result. For the first three measures this is guaranteed, but for the mean points distance it is not!

i ▷ In the vegetation study example, starting from the Euclidean distance, with "average linkage" we get the dendrogram in Fig. 8.5.i. The groups according to the original article are given as the first digits of the "roots". We see that the grouping agrees very well with one that can be obtained by cutting certain branches – although with different cutting heights. The groups 3 and 4 appear to be very homogeneous, while the second group represents a collection of poorly fitting individuals. ◁

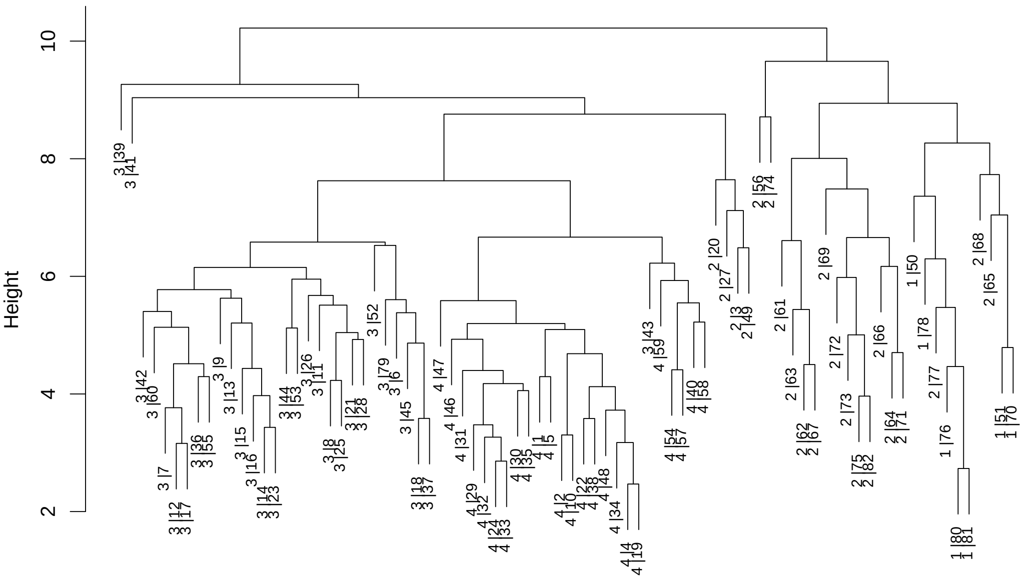

Figure 8.5.i: Dendrogram of the Hierarchical Cluster Formation with "Average Linkage", Starting from the Euclidean Distance

j **Properties.** Which of these measures makes most sense? To answer this question, we imagine a situation with two groups of different values and an observation that should be associated with one or the other of the groups. In Figur 8.2.e (i) observation 68 lies nearer to the group $\{54, 58, 50\}$ than to $\{64, 66, 67\}$, regardless of which of the three measures is used. However, if it were a little further right, so that the distance to point 67 were smaller than to 58, then according to the single linkage rule it would beat the other group, while according to the complete linkage measure it would still be connected to the left – until its distance to 66 were smaller than to 54. Further considerations of this time lead to the following judgments:

- With single linkage, chain-like clusters can arise, whose “ends” are very dissimilar.
- Complete linkage guarantees the opposite, that all members of a cluster “are close together”. This results in compact clusters which tend to be “spherical” with similar “diameters”.
- Average linkage provides a good compromise between these extremes.

k **Conclusion.** Dendrograms are graphical representations that contain a lot of information and also have aesthetic appeal. There are therefore very well liked. However, if we are to consistently purse the goal of dividing the observations into groups, we have to determine an optimal partition. Groupings via cutting a dendrogram are not optimal – how much worse they are depends on the dataset and on the agglomeration method used.

As always, if multiple methods are available, the temptation is great to carry out several or all of them and search for common conclusions, in this case common groupings. This can be reasonable if we choose certain methods that often give very different results, here, for example, single linkage and average linkage (the author does not recommend complete linkage). If the two procedures then give the same groups, they are probably very clearly delimited from each other.

Comparing dendrograms graphically is very difficult; we would have to use the freedoms that we have for the dendrogram representation (8.5.d) to make the grouping of the observations as similar as possible. The author is not familiar with any program for this.

It’s simpler to define a numerical measure for the agreement of two dendrograms. Here, though, we don’t want to carry that out.

l **Divisive Procedures.** Instead of from below, we can also try to develop a dendrogram from above. We start with a single cluster that contains all observations and divide this into two clusters according to an appropriate rule. In each further step, one of the existing cluster is split into two. There again arises a hierarchy of possible group distributions that can be represented by a dendrogram. The divisions are each made so that a dissimilarity measure between groups is as small as possible.

- **Polythetic Methods.** If this measure, like the previously discussed, is based on multiple variables, we meet – except for with very small observation numbers – computational difficulties in the optimization, since there are too many divisions of a cluster into two groups. We can find a way out like, for example, applying k-means or k-medians with $k = 2$ clusters in each step.
- **Monothetic Methods.** Division in each step can be done on the basis of a single variable. This has the advantage that a optimization problem is computationally easy to deal with and that there are very simple rules for its application as to where an observation is to be assigned – one from the underlying dataset or even a new one. The procedure gives a **decision tree**.

L Literature for Cluster Analysis

a Kaufman and Rousseeuw (1990): Concentrated on 5 programs that also are in R (library(cluster)). Use-oriented, simple. Also contains good hints about other methods. Further methods in R: Ripley, 1996, Kap. 9

b German Books: Bock, 1974; Steinhausen and Langer, 1977; Späth, 1977; Späth, 1983; Deichsel and Trampisch, 1985

More English Books: Sokal and Sneath, 1963; Hartigan, 1975; Everitt, 1980; Gordon, 1981

8.S S-Functions

a **Dissimilarity.** The function `dist` gives the Euclidian and Manhattan distances (and a few others),

```
> t.dist  <− dist(scale(sqrt(d.vegenv[,19:82])), method="manhattan")
```

The function `daisy` from the package `library(cluster)` is appropriate for data of mixed type.

b **Multidimensional Scaling.** The function

```
> t.mds  <− isoMDS(t.dist)
```

from the package `library(MASS)` unfortunately does not permit weighting of the stress measures. It minimizes

$$Q = \sum\nolimits_{h,i} (g\langle d_{hi}\rangle - \|\underline{z}_h - \underline{z}_i\|)^2 \Big/ \sum\nolimits_{h,i} \|\underline{z}_h - \underline{z}_i\|^2$$

The function `sammon` optimizes

$$Q = \sum\nolimits_{h,i} \frac{d_{hi} - \|\underline{z}_h - \underline{z}_i\|)^2}{d_{hi}} \Big/ \sum\nolimits_{h,i} d_{hi} \quad .$$

It therefore permits no monotone transformation g of the dissimilarity.

c **Cluster Analysis: Optimal Partition**

```
> kmeans(t.d, k=4)
R> t.cl  <− pam(t.d, k=4, metric="manhattan"); plot(t.cl)
```

Silhouettes are automatically supplied by `pam`, and `plot(t.cl)` shows them after a representation of the principal components. For `kmeans` we must first calculate silhouettes via

```
R> t.dist <- dist(t.d,method="manhattan")
R> t.sh <- silhouette(r.km$cluster,dmatrix=as.matrix(t.dist))
```

Finally we get the graphic via `plot(t.sh)`

d **Agglomerative Methods.**

```
> t.cl  <− hclust(t.dist, method="average")
```

Dendrogramm:

```
> plot(t.cl)
```

e **Divisive Methods.**

```
> t.cl  <− diana(t.dist)
```