Classification Analysis

Allocate data to several populations

Situation

- Have multivariate (or univariate) data from one or several populations (the number of populations is unknown)
- Want to determine the number of populations and identify the populations

Example

Table: Numerals in eleven languages

English	Norwegian	Danish	Dutch	German	French	Spanish	Italian	Polish	Hungarian	Finnish
one	en	en	een	ein	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuarto	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seix	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyole	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

Distance Matrix

Distance = # of numerals (1 to 10) differing in first letter

$$
\begin{array}{cc}
 & \begin{array}{ccccccccccc} \text{E} & \text{N} & \text{Da} & \text{Du} & \text{G} & \text{Fr} & \text{Sp} & \text{I} & \text{P} & \text{H} & \text{Fi} \end{array} \\
\begin{array}{c} \text{E} \\ \text{N} \\ \text{Da} \\ \text{Du} \\ \text{G} \\ \text{Fr} \\ \text{Sp} \\ \text{I} \\ \text{P} \\ \text{H} \\ \text{Fi} \end{array} &
\begin{bmatrix}
0 & & & & & & & & & & \\
2 & 0 & & & & & & & & & \\
2 & 1 & 0 & & & & & & & & \\
7 & 5 & 6 & 0 & & & & & & & \\
6 & 4 & 5 & 5 & 0 & & & & & & \\
6 & 6 & 6 & 9 & 7 & 0 & & & & & \\
6 & 6 & 5 & 9 & 7 & 2 & 0 & & & & \\
6 & 6 & 5 & 9 & 7 & 1 & 1 & 0 & & & \\
7 & 7 & 6 & 10 & 8 & 5 & 3 & 4 & 0 & & \\
9 & 8 & 8 & 8 & 9 & 10 & 10 & 10 & 10 & 0 & \\
9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 8 & 0
\end{bmatrix}
\end{array}
$$

Similarity Measures

To produce a group structure from the dataset we need a measure of "closeness". Items (observations, cases) are grouped together based on distance, while variables are grouped based on correlation coefficients or other measures of association.

- Classical Euclidean (straight-line) distance between two numerical p-dimensional observations $\boldsymbol{x}' = [x_1, x_2, \ldots, x_p]$ and $\boldsymbol{y}' = [y_1, y_2, \ldots, y_p]$:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + \left(x_p - y_p\right)^2} = \sqrt{(\boldsymbol{x} - \boldsymbol{y})'(\boldsymbol{x} - \boldsymbol{y})}$$

- Statistical distance:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})'\boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{y})}$$

where $\boldsymbol{S}$ is the sample covariance matrix.

- More general Minkowski distance:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}$$

Similarity Measures

- Canberra distance for nonnegative variables:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}$$

- When measurements are not numerical the observations are compared based on presence or absence of certain characteristics. That is, we introduce binary variables which assume value 0 if the characteristic is absent and 1 if it is present. Then squared Euclidean distance is applied, and it measures the total number of mismatches.

Example:

	Variables				
	1	*2*	*3*	*4*	*5*
Item *i*	1	0	0	1	1
Item *j*	1	1	0	1	0

Then $d(\text{Item } i, \text{Item } j) = (1 - 1)^2 + (0 - 1)^2 + (0 - 0)^2 + (1 - 1)2 + (1 - 0)^2 = 2$

Similarity Measures

- The total count of dissimilarities suffers from weighting the 1-1 and 0-0 matches equally. Very often 1-1 match is a stronger indication of similarity than 0-0 (think about two people who both read Latin).

- To allow for differential treatment of the matches we need some notation and a contingency table:

		Item *j*		**Total**
		1	0	
Item *i*	1	a	b	$a + b$
	0	c	d	$c + d$
Total		$a + c$	$b + d$	$p = a+b+c+d$

Example: In the previous example, $a = 2$, $b = c = d = 1$

Similarity Measures

p. 675

Table 12.1 Similarity Coefficients for Clustering Items*

Coefficient	Rationale
1. $\dfrac{a+d}{p}$	Equal weights for 1–1 matches and 0–0 matches.
2. $\dfrac{2(a+d)}{2(a+d)+b+c}$	Double weight for 1–1 matches and 0–0 matches.
3. $\dfrac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4. $\dfrac{a}{p}$	No 0–0 matches in numerator.
5. $\dfrac{a}{a+b+c}$	No 0–0 matches in numerator or denominator. (The 0–0 matches are treated as irrelevant.)
6. $\dfrac{2a}{2a+b+c}$	No 0–0 matches in numerator or denominator. Double weight for 1–1 matches.
7. $\dfrac{a}{a+2(b+c)}$	No 0–0 matches in numerator or denominator. Double weight for unmatched pairs.
8. $\dfrac{a}{b+c}$	Ratio of matches to mismatches with 0–0 matches excluded.

*[p binary variables; see (12-7).]

Hierarchical Clustering Methods

The following are the steps in the agglomerative hierarchical clustering algorithm for grouping N objects (items or variables).

1. Start with N clusters, each consisting of a single entity and an $N \times N$ symmetric matrix (table) of distances (or similarities) $\mathbf{D} = (d_{ij})$.
2. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between the "most similar" clusters U and V be d_{UV}.
3. Merge clusters U and V. Label the newly formed cluster *(UV)*. Update the entries in the distance matrix by

   a) deleting the rows and columns corresponding to clusters U and V and

   b) adding a row and column giving the distances between cluster *(UV)* and the remaining clusters.

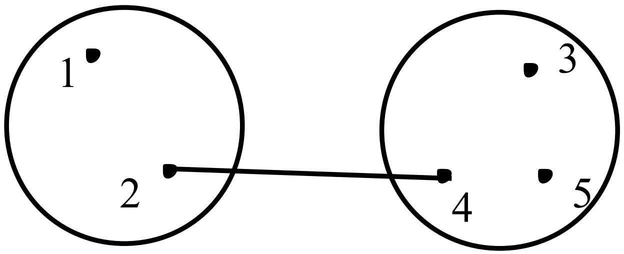
4. Repeat steps 2 and 3 a total of N-1 times. (All objects will be a single cluster at termination of this algorithm.) Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

A reasonable question to ask is how do we measure distance between clusters? This is known as the *linkage* method. Most commonly used ones are presented on the next slide.

Different methods of computing inter-cluster distance

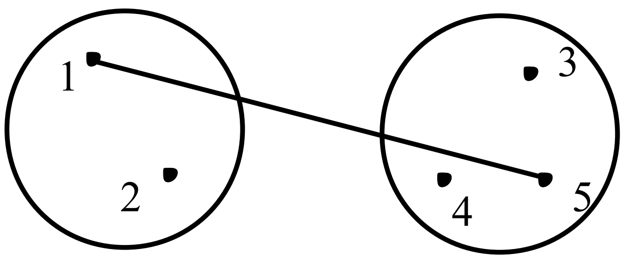
Cluster Distance

Single Linkage



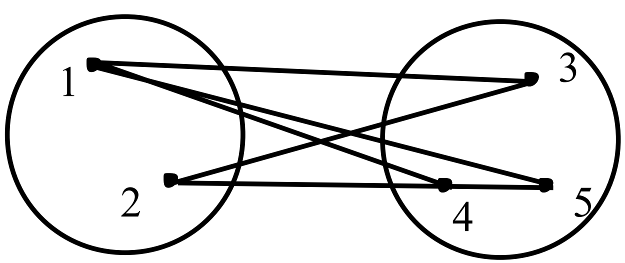
d_{24}

Complete Linkage



d_{15}

Average Linkage



$$\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$$

Example 12.3

To illustrate the **single linkage** algorithm, we consider the hypothetical distance matrix between pairs of five objects given below:

$$
\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{ccccc} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{array}\right] \end{array}
$$

Treating each object as a cluster, the clustering begins by merging the two closest items (3 & 5).

To implement the next level of clustering we need to compute the distances between cluster (35) and the remaining objects:

$$d_{(35)1} = \min\{3,11\} = 3$$

$$d_{(35)2} = \min\{7,10\} = 7$$

$$d_{(35)4} = \min\{9,8\} = 8$$

The new distance matrix becomes:

The new distance matrix becomes:

$$
\begin{array}{c}
\\ (35) \\ 1 \\ 2 \\ 4
\end{array}
\begin{array}{c}
\begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\
\begin{bmatrix}
0 & & & \\
3 & 0 & & \\
7 & 9 & 0 & \\
8 & 6 & 5 & 0
\end{bmatrix}
\end{array}
$$

The next two closest clusters ((35) & 1) are merged to form cluster (135). Distances between this cluster and the remaining clusters become:

Distances between this cluster and the remaining clusters become:

$$d_{(135)2} = \min\{7,9\} = 7$$
$$d_{(135)4} = \min\{8,6\} = 6$$

The distance matrix now becomes:

$$\begin{array}{c} \\ (135) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

Continuing the next two closest clusters (2 & 4) are merged to form cluster (24).

Distances between this cluster and the remaining clusters become:

$$d_{(135)(24)} = \min\{d_{(135)2}, d_{(135)4}) = \min\{7,6\} = 6$$

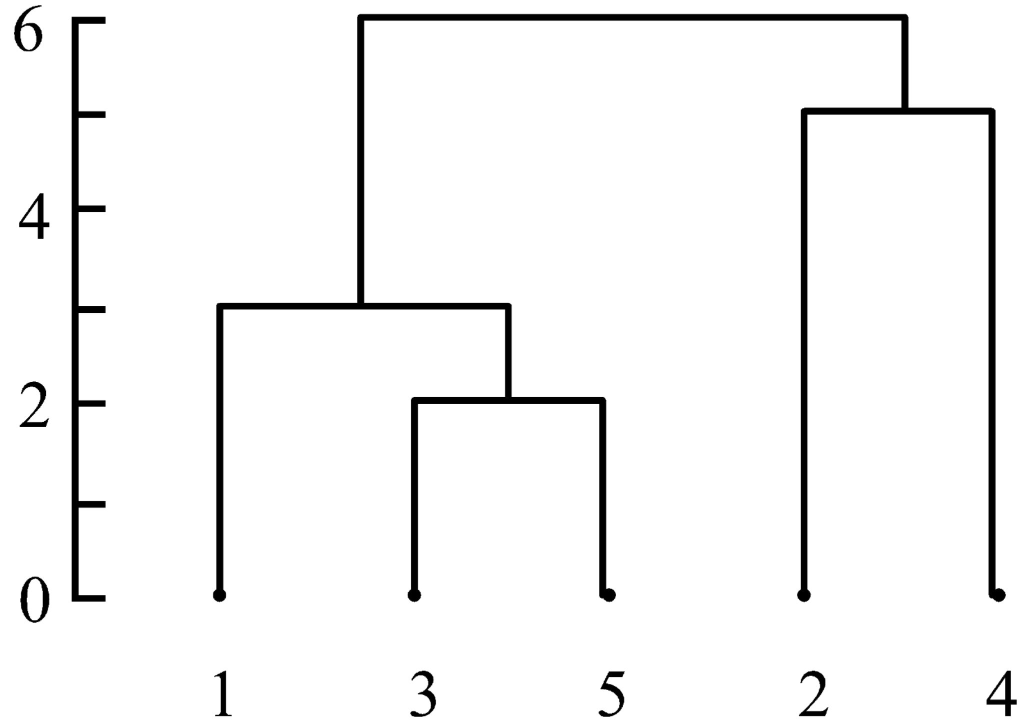
The final distance matrix now becomes:

$$\begin{array}{c c} & \begin{array}{cc} (135) & (24) \end{array} \\ \begin{array}{c} (135) \\ (24) \end{array} & \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{array}$$

At the final step clusters (135) and (24) are merged to form the single cluster (12345) of all five items.

The results of this algorithm can be summarized graphically on the following "**dendogram**"

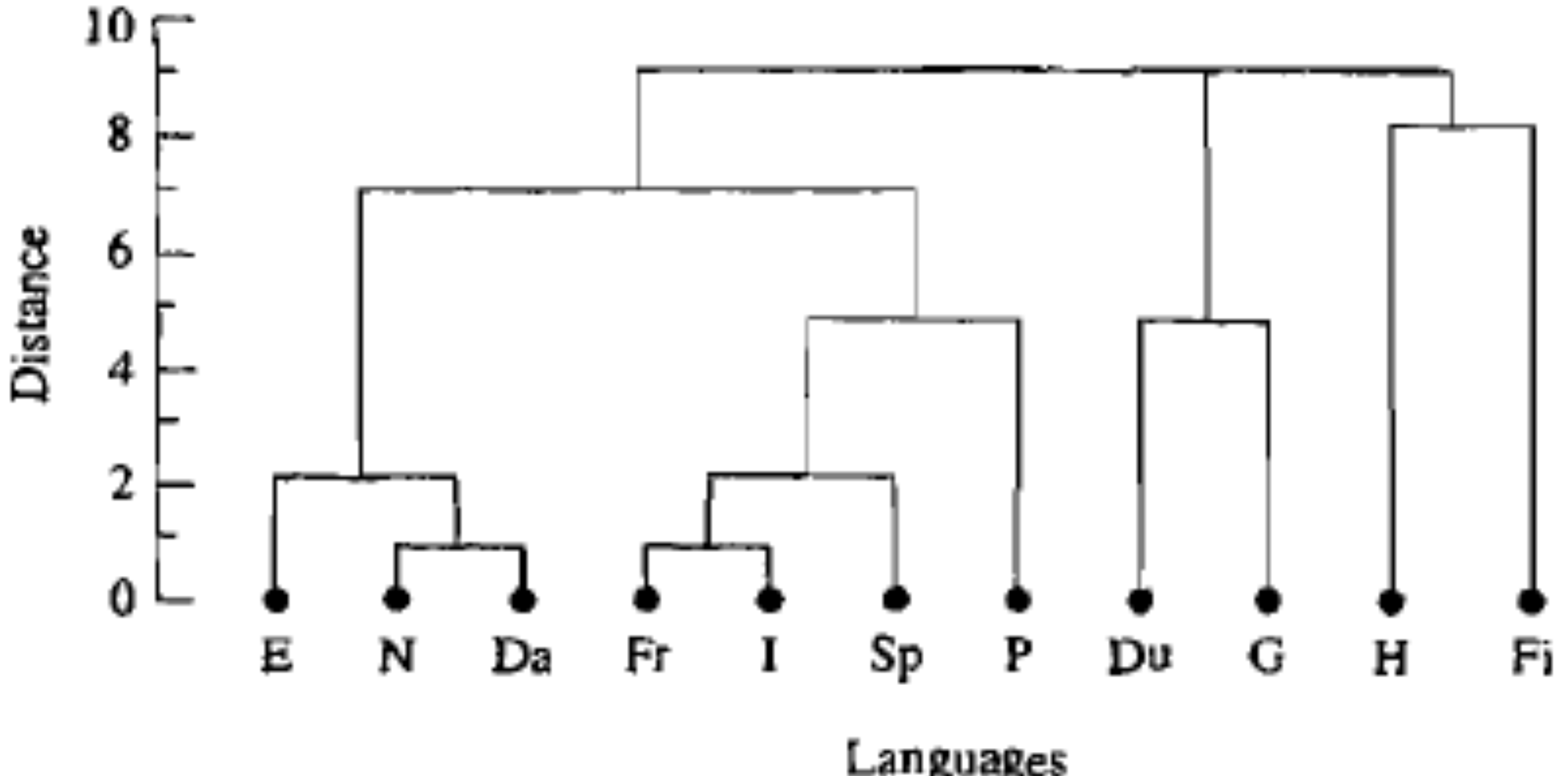
Figure
Single Linkage dendogram for distances between five objects



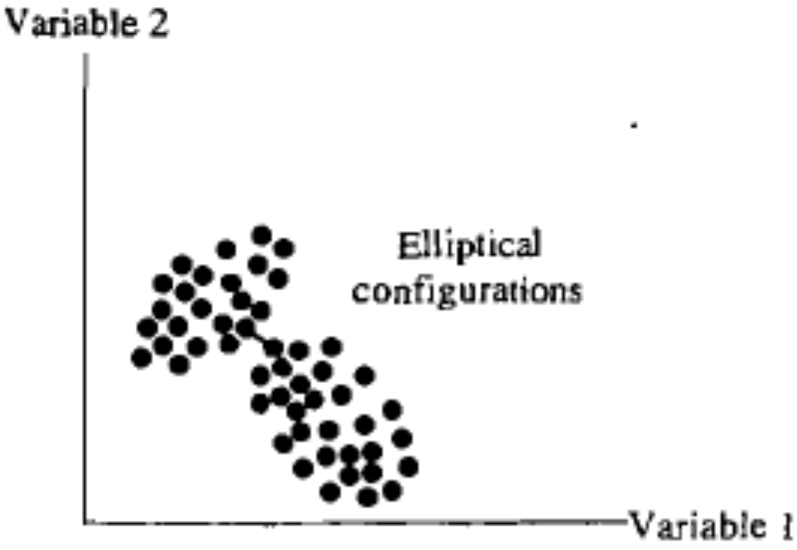
Dendograms

for clustering the 11 languages based on the ten numerals

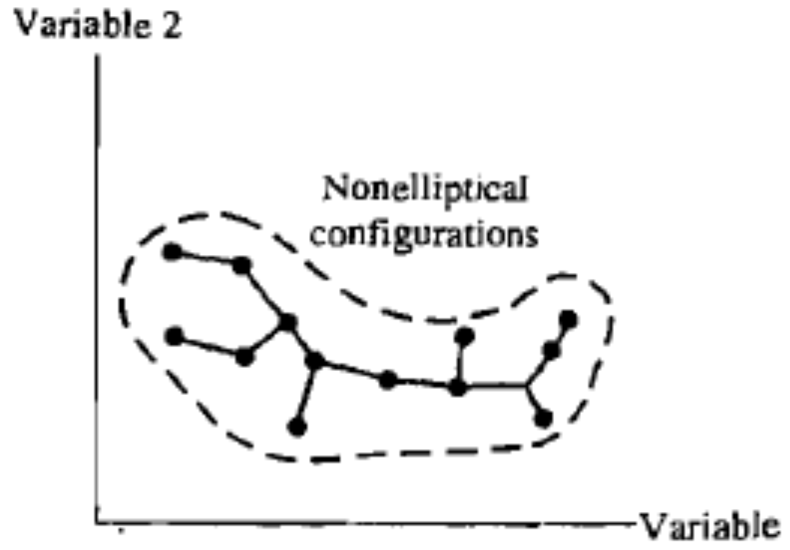
Single linkage dendrograms



Single linkage *chaining* effect

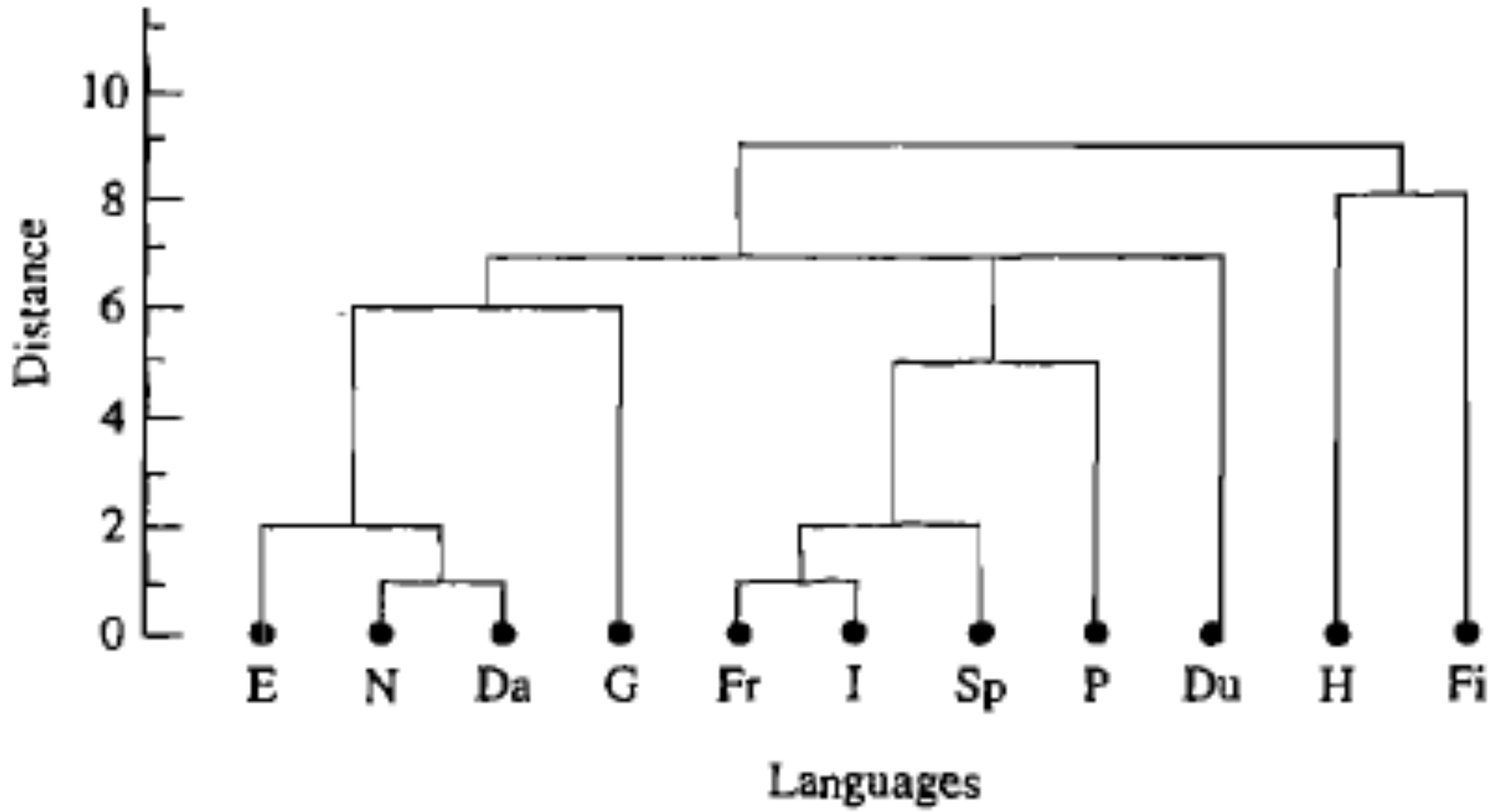


(a) Single linkage confused by near overlap



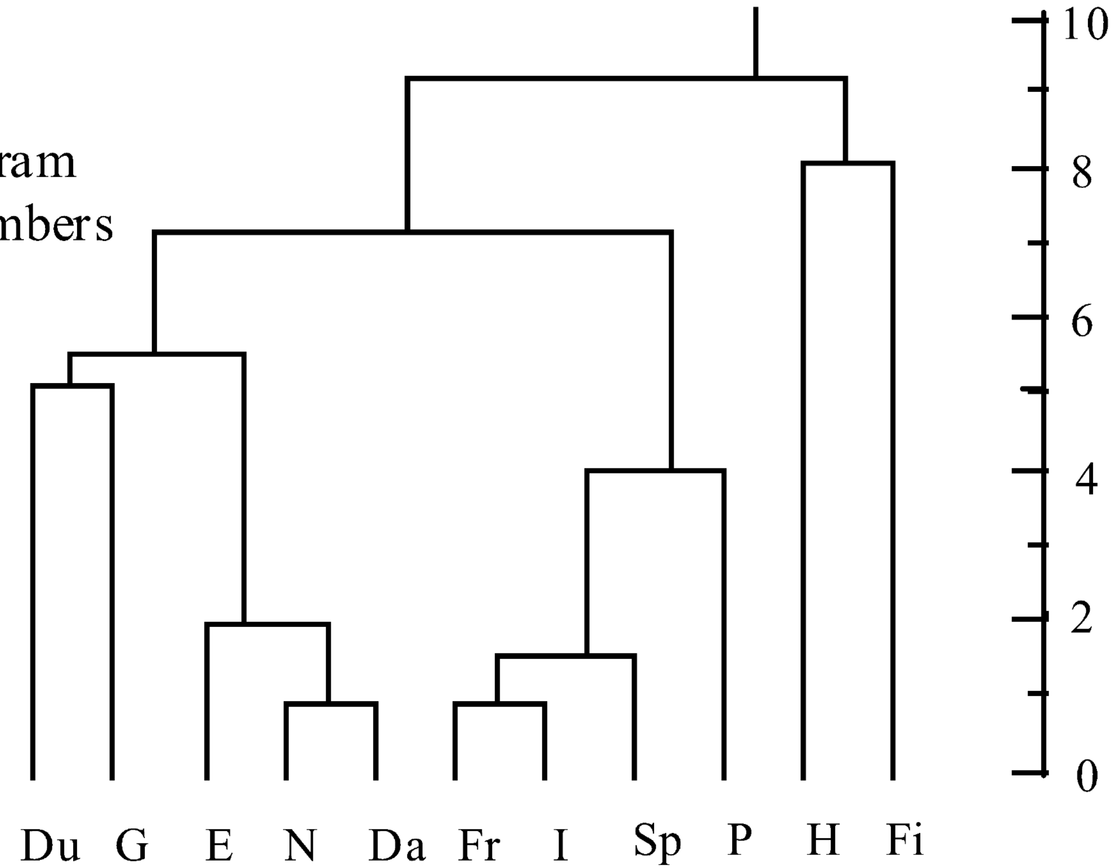
(b) Chaining effect

Complete linkage dendrogram



Figure

Average Linkage dendogram for distances between numbers in 11 languages



Example 2: Public Utility data

	Company	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
		variables							
1	Arizona Public Service	1.06	9.2	151	54.4	1.6	9077	0.0	0.628
2	Boston Edison Co	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
3	Central Louisiana Electric Co	1.43	15.4	113	53.0	3.4	9212	0.0	1.058
4	Commonwealth Edison Co	1.02	11.2	168	56.0	0.3	6423	34.3	0.700
5	Consolidated Edison Co (NY)	1.49	8.8	192	51.2	1.0	3300	15.6	2.044
6	Florida Power & Light Co	1.32	13.5	111	60.0	-2.2	11127	22.5	1.241
7	Hawaiian Electric Co	1.22	12.2	175	67.6	2.2	7642	0.0	1.652
8	Idaho Power Co	1.10	9.2	245	57.0	3.3	13082	0.0	0.309
9	Kentucky Utilities Co	1.34	13.0	168	60.4	7.2	8406	0.0	0.862
10	Madison Gas & Electric Co	1.12	12.4	197	53.0	2.7	6455	39.2	0.623
11	Nevada Power Co	0.75	7.5	173	51.5	6.5	17441	0.0	0.768
12	New England Electric Co	1.13	10.9	178	62.0	3.7	6154	0.0	1.897
13	Northern States Power Co	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
14	Oklahoma Gas & Electric Co	1.09	12.0	96	49.8	1.4	9673	0.0	0.588
15	Pacific Gas & Electric Co	0.96	7.6	164	62.2	-0.1	6468	0.9	1.400
16	Puget Sound Power & Light Co	1.16	9.9	252	56.0	9.2	15991	0.0	0.620
17	San Diego Gas & Electric Co	0.76	6.4	136	61.9	9.0	5714	8.3	1.920
18	The Southern Co	1.05	12.6	150	56.7	2.7	10140	0.0	1.108
19	Texas Utilities Co	1.16	11.7	104	54.0	-2.1	13507	0.0	0.636
20	Wisconsin Electric Power Co	1.20	11.8	148	59.9	3.5	7287	41.1	0.702
21	United Illuminating Co	1.04	8.6	204	61.0	3.5	6650	0.0	2.116
22	Virginia Electric & Power Co	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

X_1: Fixed charge coverage ratio (income/debt)

X_2: Rate of return on capital

X_3: Cost per KW capacity in place

X_4: Annual load factor

X_5: Peak KWH demand growth from 1974 to1975

X_6: Sales (KWH per year)

X_7: Percent Nuclear

X_8: Total fuel costs (cents per KWH)

Table: Distances between 22 Utilities

Firm number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0.00																					
2	3.10	0.00																				
3	3.68	4.92	0.00																			
4	2.46	2.16	4.11	0.00																		
5	4.12	3.85	4.47	4.13	0.00																	
6	3.61	4.22	2.99	3.20	4.60	0.00																
7	3.90	3.45	4.22	3.97	4.60	3.35	0.00															
8	2.74	3.89	4.99	3.69	5.16	4.91	4.36	0.00														
9	3.25	3.96	2.75	3.75	4.49	3.73	2.80	3.59	0.00													
10	3.10	2.71	3.93	1.49	4.05	3.83	4.51	3.67	3.57	0.00												
11	3.49	4.79	5.90	4.86	6.46	6.00	6.00	3.46	5.18	5.08	0.00											
12	3.22	2.43	4.03	3.50	3.60	3.74	1.66	4.06	2.74	3.94	5.21	0.00										
13	3.96	3.43	4.39	2.58	4.76	4.55	5.01	4.14	3.66	1.41	5.31	4.50	0.00									
14	2.11	4.32	2.74	3.23	4.82	3.47	4.91	4.34	3.82	3.61	4.32	4.34	4.39	0.00								
15	2.59	2.50	5.16	3.19	4.26	4.07	2.93	3.85	4.11	4.26	4.74	2.33	5.10	4.24	0.00							
16	4.03	4.84	5.26	4.97	5.82	5.84	5.04	2.20	3.63	4.53	3.43	4.62	4.41	5.17	5.18	0.00						
17	4.40	3.62	6.36	4.89	5.63	6.10	4.58	5.43	4.90	5.48	4.75	3.50	5.61	5.56	3.40	5.56	0.00					
18	1.88	2.90	2.72	2.65	4.34	2.85	2.95	3.24	2.43	3.07	3.95	2.45	3.78	2.30	3.00	3.97	4.43	0.00				
19	2.41	4.63	3.18	3.46	5.13	2.58	4.52	4.11	4.11	4.13	4.52	4.41	5.01	1.88	4.03	5.23	6.09	2.47	0.00			
20	3.17	3.00	3.73	1.82	4.39	2.91	3.54	4.09	2.95	2.05	5.35	3.43	2.23	3.74	3.78	4.82	4.87	2.92	3.90	0.00		
21	3.45	2.32	5.09	3.88	3.64	4.63	2.68	3.98	3.74	4.36	4.88	1.38	4.94	4.93	2.10	4.57	3.10	3.19	4.97	4.15	0.00	
22	2.51	2.42	4.11	2.58	3.77	4.03	4.00	3.24	3.21	2.56	3.44	3.00	2.74	3.51	3.35	3.46	3.63	2.55	3.97	2.62	3.01	0.00

Dendogram

Cluster Analysis of $N = 22$ Utility companies

Euclidean distance, Average Linkage