Similarità, distanza, associazione

Misure di similarità

- Variano da 1 (massima somiglianza, osservazioni identiche) a 0 (nessuna somiglianza, osservazioni completamente diverse)
- Possono essere simmetriche (l'assenza di una specie è considerata informativa) o asimmetriche (l'assenza non è un dato certo)
- Se trasformate in dissimilarità (D=1-S), possono godere di proprietà metriche o meno

Osservazione *j*

Osservazione *k*	**1**	**0**
1	*a*	*b*
0	*c*	*d*

$p = a + b + c + d$

	St. A	St. B
Sp. 1	3	0
Sp. 2	4	2
Sp. 3	0	0
Sp. 4	2	5
Sp. 5	1	16
Sp. 6	0	4
Sp. 7	12	5
Sp. 8	0	1
Sp. 9	0	4
Sp. 10	1	0

a = 4 b = 3
c = 2 d = 1

Alcune misure di similarità

simmetriche

concordanza semplice

$$S_{jk} = \frac{a+d}{p}$$

Rogers & Tanimoto

$$S_{jk} = \frac{a+d}{a+2b+2c+d}$$

asimmetriche

Jaccard

$$S_{jk} = \frac{a}{a+b+c}$$

Sørensen

$$S_{jk} = \frac{2a}{2a+b+c}$$

Gower

$$S_{jk} = \frac{\sum_{i=1}^{p} w_i s_i}{\sum_{i=1}^{p} w_i}$$

- per descrittori binari s_i=1 nei casi di concordanza e s_i=0 altrimenti (la concordanza da doppio zero viene trattata in accordo con il significato dello zero)
- per descrittori semi-quantitativi ordinali e quantitativi:
  s_i=1-|x_{ij}-x_{ik}| R_i^{-1} (dove R_i è l'intervallo di variazione dell'*i*-mo descrittore)

Steinhaus

$$S_{jk} = \frac{2\sum_{i=1}^{p} \min(x_{ij}, x_{ik})}{\sum_{i=1}^{p} x_{ij} + x_{ik}}$$

D=1-S

Bray-Curtis

$$D_{jk} = \frac{\sum_{i=1}^{p} \left| x_{ij} - x_{ik} \right|}{\sum_{i=1}^{p} x_{ij} + x_{ik}}$$

Dissimilarità metriche se…

semimetrica

1. D_{jk}=0 se j=k
2. D_{jk}>0 se j≠k
3. D_{jk}=D_{kj}

4. D_{jk}+D_{kh}≥D_{jh} (assioma della diseguaglianza triangolare)

Misure di distanza

euclidea

$$D_{jk} = \sqrt{\sum_{i=1}^{p} (x_{ij} - x_{ik})^2}$$

Manhattan

$$D_{jk} = \sum_{i=1}^{p} \left| x_{ij} - x_{ik} \right|$$

Minkowski

$$D_{jk} = \sqrt[r]{\left(\sum_{i=1}^{p} \left| x_{ij} - x_{ik} \right|^r \right)}$$

Czekanowski

$$D_{jk} = \frac{1}{p} \sum_{i=1}^{p} \left| x_{ij} - x_{ik} \right|$$

Canberra

$$D_{ij} = \sum_{i=1}^{p} \frac{\left| x_{ij} - x_{ik} \right|}{\left(x_{ij} + x_{ik} \right)}$$

Bray-Curtis

$$D_{ij} = \frac{\sum_{i=1}^{s} \left| x_{ij} - x_{ik} \right|}{\sum_{i=1}^{s} \left(x_{ij} + x_{ik} \right)}$$

corda

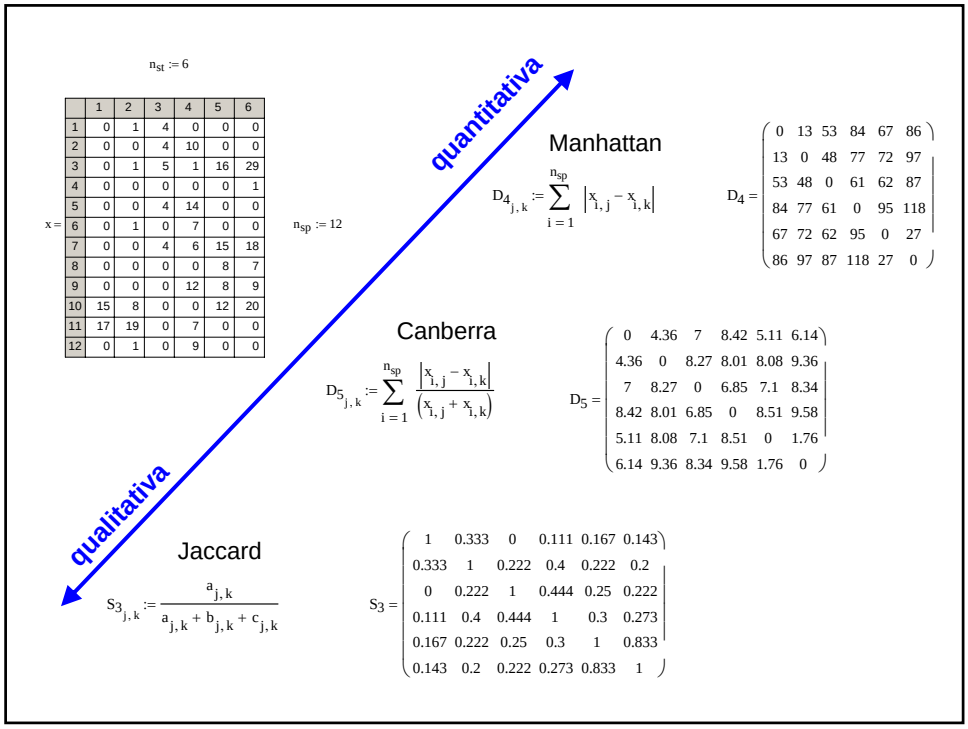
$$D_{jk} = \sqrt{2 \left(1 - \frac{\sum_{i=1}^{p} x_{ij} x_{ik}}{\sqrt{\sum_{i=1}^{p} x_{ij}^2 \sum_{i=1}^{p} x_{ik}^2}} \right)}$$

Misure di associazione

Fager & McGowan

$$S_{jk} = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2 \cdot \sqrt{a+c}} \qquad (c \geq b)$$

…ma possono essere utilizzati
anche i coefficienti di correlazione.


$n_{st} := 6$
x =
	1	2	3	4	5	6
1	0	1	4	0	0	0
2	0	0	4	10	0	0
3	0	1	5	1	16	29
4	0	0	0	0	0	1
5	0	0	4	14	0	0
6	0	1	0	7	0	0
7	0	0	4	6	15	18
8	0	0	0	0	8	7
9	0	0	0	12	8	9
10	15	8	0	0	12	20
11	17	19	0	7	0	0
12	0	1	0	9	0	0
$n_{sp} := 12$
quantitativa
qualitativa
Manhattan
$D4_{j,k} := \sum_{i=1}^{n_{sp}} |x_{i,j} - x_{i,k}|$
$D_4 =$
0	13	53	84	67	86
13	0	48	77	72	97
53	48	0	61	62	87
84	77	61	0	95	118
67	72	62	95	0	27
86	97	87	118	27	0
Canberra
$D5_{j,k} := \sum_{i=1}^{n_{sp}} \frac{|x_{i,j} - x_{i,k}|}{(x_{i,j} + x_{i,k})}$
$D_5 =$
0	4.36	7	8.42	5.11	6.14
4.36	0	8.27	8.01	8.08	9.36
7	8.27	0	6.85	7.1	8.34
8.42	8.01	6.85	0	8.51	9.58
5.11	8.08	7.1	8.51	0	1.76
6.14	9.36	8.34	9.58	1.76	0
Jaccard
$S3_{j,k} := \frac{a_{j,k}}{a_{j,k} + b_{j,k} + c_{j,k}}$
$S_3 =$
1	0.333	0	0.111	0.167	0.143
0.333	1	0.222	0.4	0.222	0.2
0	0.222	1	0.444	0.25	0.222
0.111	0.4	0.444	1	0.3	0.273
0.167	0.222	0.25	0.3	1	0.833
0.143	0.2	0.222	0.273	0.833	1