Different data types

Recoding before CA

Ratings

Rankings

Preferences

Continuous data

- Correspondence analysis (CA) can also be applied to other types of data:
  - ratings
  - preferences
  - paired comparisons
  - distances
  - measurement data
- The “art” is in the recoding of the data to be suitable for CA.
- Remember that CA analyses profiles, weighted by masses and with inter-profile distances measured by chi-squared distance.
- If the data can be put into a form for which these concepts makes sense, then CA is a valid method for visualizing the data

ISSP 1993: Environment

Q.4 SCIENCE AND ENVIRONMENT

How much do you agree or disagree with each of these statements?

Q.4a We believe too often in science, and not enough in feelings and faith.
Q.4b Over all, modern science does more harm than good.
Q.4c Any change humans cause in nature - no matter how scientific - is likely to make things worse.
Q.4d Modern science will solve our environmental problems with little change to our way of life.

1. Strongly agree
2. Agree
3. Neither agree nor disagree
4. Disagree
5. Strongly disagree
8. Can't choose, don't know
9. NA, refused

Recall the indicator matrix definition of MCA

Original responses

```
2 2 1 2
2 2 2 5
4 3 2 5
2 5 4 2
4 2 1 5
1 4 1 5
1 2 2 3
1 3 2 4
3 2 2 4
3 5 5 2
. . . .
. . . .
etc. . .
```

Recoded indicator matrix

```
0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0
0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1
0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1
0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0
0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1
1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1
1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0
1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0
0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0
0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0
. . . . . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . . . . .
etc. . (N rows)
```

The resulting graphical display of the column categories has a separate point for each level of the scale. Each category can find its own position in the map and the ordering of the scale points is not taken into account.

There are methods for forcing the ordering of the scale points on, say, the first principal axis: for example, categorical PCA (implemented in SPSS)

Alternatively, we can constrain the scale points to lie at equal distances along straight lines in the map, leading to a much simpler graphical display.

Doubling of ratings

Original responses

2	2	1	2
2	2	2	5
4	3	2	5
2	5	4	2
4	2	1	5
1	4	1	5
1	2	2	3
1	3	2	4
3	2	2	4
3	5	5	2
.	.	.	.
.	.	.	.
etc.	.	.	

Doubled ratings

1	3	1	3	0	4	1	3
1	3	1	3	1	3	4	0
3	1	2	2	1	3	4	0
1	3	4	0	3	1	1	3
3	1	1	3	0	4	4	0
0	4	3	1	0	4	4	0
0	4	1	3	1	3	2	2
0	4	2	2	1	3	3	1
2	2	1	3	1	3	3	1
2	2	4	0	4	0	1	3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
etc.	.	(N rows)					

Each categorical variable is converted into a pair of columns: one is called the ***positive pole*** of the rating scale and the other the ***negative pole***. Deciding which is positive or negative depends on the context.

Assuming a rating scale that starts at 1 (e.g., the 1-to-5 Likert scale here), the first column consists of all the ratings minus 1. Since "strongly agree" is a 1 in this case, this column will measure the strength of disagreement, so we should give call it the negative pole. The second column is in this case 4 minus the first one, and measures agreement (positive pole). The two columns sum to 4.

Doubling of ratings

Original responses

A	B	C	D
2	2	1	2
2	2	2	5
4	3	2	5
2	5	4	2
4	2	1	5
1	4	1	5
1	2	2	3
1	3	2	4
3	2	2	4
3	5	5	2
.	.	.	.
.	.	.	.
etc.	.	.	

Doubled ratings

A		B		C		D	
−	+	−	+	−	+	−	+
1	3	1	3	0	4	1	3
1	3	1	3	1	3	4	0
3	1	2	2	1	3	4	0
1	3	4	0	3	1	1	3
3	1	1	3	0	4	4	0
0	4	3	1	0	4	4	0
0	4	1	3	1	3	2	2
0	4	2	2	1	3	3	1
2	2	1	3	1	3	3	1
2	2	4	0	4	0	1	3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
etc.	.	(N rows)					

The rational for this coding is as follows:

CA analyses frequency data. So when a person gives a response of "2" on the 5-point agreement-disagreement scale, then he/she has 1 scale point "below" and 3 scale points "above": 1 **2** 3 4 5 , i.e. has a "count" of 1 towards disagreement and a "count" of 3 towards agreement. Hence the corresponding pair of doubled ratings is [1 3].

Doubling of ratings

Original responses

A	B	C	D
2	2	1	2
2	2	2	5
4	3	2	5
2	5	4	2
4	2	1	5
1	4	1	5
1	2	2	3
1	3	2	4
3	2	2	4
3	5	5	2
.	.	.	.
.	.	.	.
etc.	.	.	

Doubled ratings

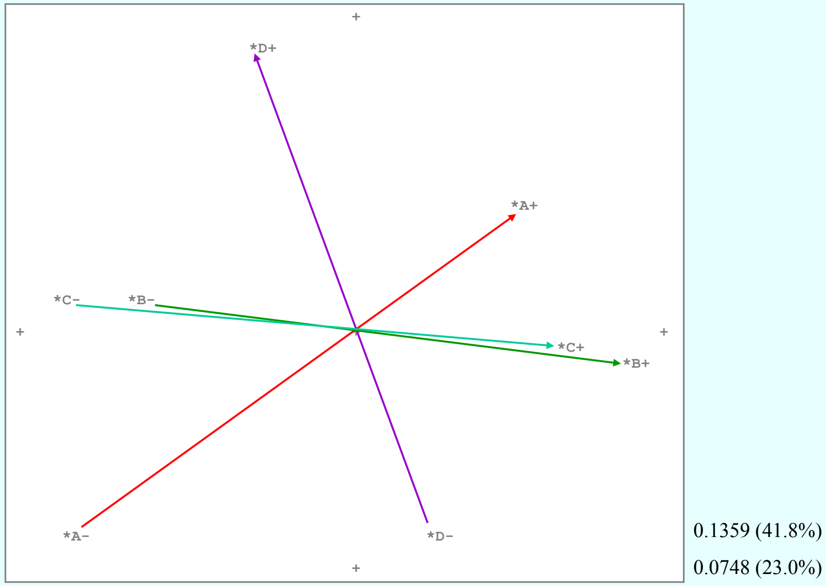
A		B		C		D	
−	+	−	+	−	+	−	+
1	3	1	3	0	4	1	3
1	3	1	3	1	3	4	0
3	1	2	2	1	3	4	0
1	3	4	0	3	1	1	3
3	1	1	3	0	4	4	0
0	4	3	1	0	4	4	0
0	4	1	3	1	3	2	2
0	4	2	2	1	3	3	1
2	2	1	3	1	3	3	1
2	2	4	0	4	0	1	3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
etc.	.	(N rows)					

CA is then applied to the doubled matrix with $2Q$ columns.

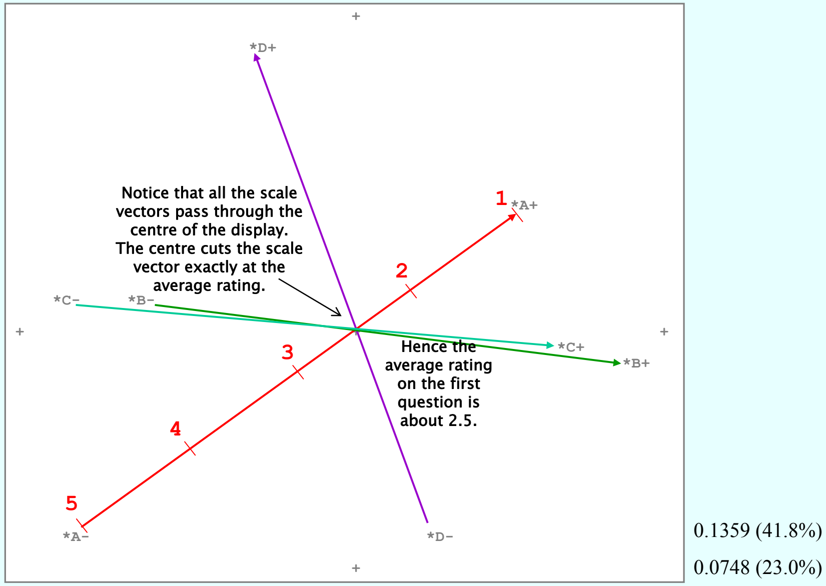
Each pair of points represents the end-points of the rating scale and the intermediate points are at equal intervals between these two extremes.

You can think of this analysis as an MCA where the 5 scale points of each variable are forced to lie on a straight line at equal intervals.

CA of doubled ratings



CA of doubled ratings – origin and scale



CA of doubled ratings

INERTIAS AND PERCENTAGES OF INERTIA

1	0.135890	41.85%	***
2	0.074808	23.04%	******************************
3	0.065729	20.24%	**************************
4	0.048311	14.88%	*******************
5	0.000002	0.00%	
6	0.000001	0.00%	
7	0.000000	0.00%	
	0.324740		

COLUMN CONTRIBUTIONS

J	NAME	QLT	MAS	INR	k=1	COR	CTR	k=2	COR	CTR
1	A-	662	89	171	-504	407	166	-399	255	190
2	A+	662	161	95	279	407	92	221	255	105
3	B-	597	137	112	-395	590	158	44	7	4
4	B+	597	113	136	480	590	191	-54	7	4
5	C-	580	103	150	-519	573	205	58	7	5
6	C+	580	147	106	366	573	145	-41	7	3
7	D-	766	142	99	132	77	18	-395	689	297
8	D+	766	108	131	-175	77	24	522	689	393

Further remarks on geometry of doubling

Doubled ratings

A −	A +	B −	B +	C −	C +	D −	D +	total
1	3	1	3	0	4	1	3	16
1	3	1	0	1	3	4	0	16
3	1	2	2	1	3	4	0	16
1	3	4	0	3	1	1	3	16
3	1	1	3	0	4	4	0	16
0	4	3	1	0	4	4	0	16
0	4	1	3	1	3	2	2	16
0	4	2	2	1	3	3	1	16
2	2	1	3	1	3	3	1	16
2	2	4	0	4	0	1	3	16
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

etc. . (*N* rows)

- Each respondent has the same mass (as in MCA).
- Distances between column points are (unweighted) Euclidean between column profiles.
- Distances between row points have weights which are related to a measure of **polarisation**. Differences between respondents on a highly polarised variable (i.e., mean near the end-points) will be weighted more than usual.
- The column masses occur in pairs which are proportional to the doubled means of the ratings, and sum to a constant.

Doubling of preferences (rankings)

Suppose N individuals rank-order p objects in order of preference.

For example, in a marketing survey about bottled mineral water, 400 respondents rank-ordered 6 different attributes of this type of product.

Usually the data are coded as the ranking given to the attributes, where 1 indicates first choice, 2 second, and so on... This is just like a rating scale except no scale values can be repeated by a respondent.

A	B	C	D	E	F
6	1	5	4	3	2
5	2	3	1	6	4
4	3	2	5	1	6
5	2	4	3	6	1
5	3	2	1	4	6
.	.	.	.		
.	.	.	.		

etc. . .

Original responses

A −	A +	B −	B +	C −	C +	D −	D +	E −	E +	F −	F +
5	0	0	5	4	1	3	2	2	3	1	4
4	1	1	4	2	3	0	5	5	0	3	2
3	2	2	3	1	4	4	1	0	5	5	0
4	1	1	4	3	2	2	3	5	0	0	5
4	1	2	3	1	4	0	5	3	2	5	0
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

etc. . (400 rows)

Doubled rankings (doubled columns)

	A	B	C	D	E	F
1−	5	0	4	3	2	1
2−	4	1	2	0	5	3
3−	3	2	1	4	0	5
4−	4	1	3	2	5	0
5−	4	2	1	0	3	5
.	.	.	.	.	.	.
.	.	.	.	.	.	
1+	0	5	1	2	3	4
2+	1	4	3	5	0	2
3+	2	3	4	1	5	0
4+	1	4	2	3	0	5
5+	1	3	4	5	2	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Doubled rankings (doubled rows)

Coding of paired comparisons

Once again, we can consider the doubled data as counts: for example...

Attribute A is in 6th (last) position: 5 attributes are preferred to it, and it is preferred to none.

This idea can be extended to paired comparisons. Suppose that, instead of rank-ordering, respondents consider each pair of attributes and then decide which they prefer. Thus each of the N individuals has to make $\frac{1}{2}p(p-1)$ decisions.

A	B	C	D	E	F
6	1	5	4	3	2
5	2	3	1	6	4
4	3	2	5	1	6
5	2	4	3	6	1
5	3	2	1	4	6
.	.	.	.		
.	.	.	.		
etc.	.	.			

Original responses

A		B		C		D		E		F	
−	+	−	+	−	+	−	+	−	+	−	+
5	0	0	5	4	1	3	2	2	3	1	4
4	1	1	4	2	3	0	5	5	0	3	2
3	2	2	3	1	4	4	1	0	5	5	0
4	1	1	4	3	2	2	3	5	0	0	5
4	1	2	3	1	4	0	5	3	2	5	0
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
etc. . (400 rows)											

Doubled rankings (doubled columns)

	A	B	C	D	E	F
1−	5	0	4	3	2	1
2−	4	1	2	0	5	3
3−	3	2	1	4	0	5
4−	4	1	3	2	5	0
5−	4	2	1	0	3	5
.	.	.	.	.	.	.
.	.	.	.	.	.	
1+	0	5	1	2	3	4
2+	1	4	3	5	0	2
3+	2	3	4	1	5	0
4+	1	4	2	3	0	5
5+	1	3	4	5	2	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Doubled rankings (doubled rows)

Coding of paired comparisons

Once again, we can consider the doubled data as counts: for example...

Again all 5 other attributes have been preferred to A and A has been preferred to none.

There can be some inconsistencies in the paired comparisons, which means that we can get this type of repetition in the doubled data which we never had for rank-orderings.

A/B	A/C	A/D	...
B	C	D	...
B	C	D	...
B	C	A	...
B	C	D	...
B	A	A	...
.	.	.	...
.	.	.	...
etc...			

Original responses to paired comparisons

A		B		C		D		E		F	
−	+	−	+	−	+	−	+	−	+	−	+
5	0	0	5	4	1	3	2	3	2	2	3
4	1	1	4	2	3	0	5	5	0	3	2
3	2	2	3	1	4	4	1	0	5	5	0
4	1	1	4	3	2	2	3	5	0	1	4
4	1	2	3	1	4	0	5	3	2	5	0
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
etc. . (400 rows)											

Doubled preference counts (doubled columns)

	A	B	C	D	E	F
1−	5	0	4	3	3	2
2−	4	1	2	0	5	3
3−	3	2	1	4	0	5
4−	4	1	3	2	5	1
5−	4	2	1	0	3	5
.	.	.	.	.	.	.
.	.	.	.	.	.	
1+	0	5	1	2	2	3
2+	1	4	3	5	0	2
3+	2	3	4	1	5	0
4+	1	4	2	3	0	4
5+	1	3	4	5	2	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Doubled preference counts (doubled rows)

Continuous measurement scales...

I (12)

Q

A	B	C	. .
3.2	15.7	1.7	. .
5.1	10.3	2.0	. .
4.2	7.8	3.1	. .
2.0	12.3	5.2	. .
1.9	13.2	2.0	. .
.	.	.	. .
.	.	.	. .

Q

A	B	C	. .
4	12	3	. .
11	8	5	. .
6	4	7	. .
2	10	10	. .
1	11	4	. .
.	.	.	. .
.	.	.	. .

$2Q$

A		B		C		. .
−	+	−	+	−	+	. .
3	8	11	0	2	9	. .
10	1	7	4	4	7	. .
5	6	3	8	6	5	. .
1	10	9	2	9	2	. .
0	11	10	1	3	8	. .
.	.	.	.	.	.	. .
.	.	.	.	.	.	. .

- Convert all data to rank-orders within the variable.
- Double the ranks across the cases.
- "Nonparametric" CA...

Continuous measurement scales...

I (12)

Q

A	B	C	. .
3.2	1	1.7	. .
5.1	2	2.0	. .
4.2	2	3.1	. .
2.0	1	5.2	. .
1.9	1	2.0	. .
.	.	.	. .
.	.	.	. .

$2Q$

A		B		. .
+	−	1	2	
-0.26	1.26	1	0	. .
0.83	0.17	0	1	. .
0.36	0.64	0	1	. .
-0.56	1.56	1	0	. .
-0.58	1.58	1	0	. .
.	.	.	.	. .
.	.	.	.	. .

- Standardise the data by centring with respect to mean and dividing by standard deviation: $z = \frac{x - \bar{x}}{s}$
- Calculate doubled entries for variable: $A+ = (1+z)/2 \quad A- = (1-z)/2$
- This is a good coding when categorical variables, especially binary variables, are present.

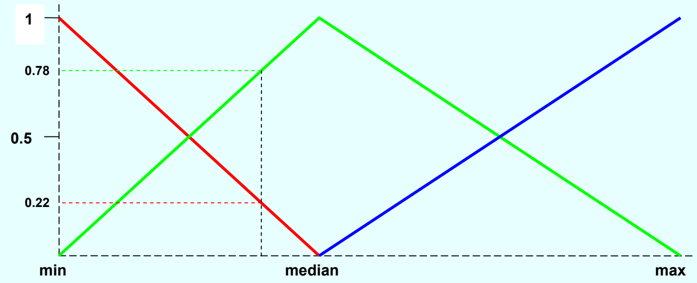
Data set “meteo”:

Annual averages of five meteorological variables in 40 Turkish cities

Data all on different scales. One way to homogenize the data is to code into categories, either “crisply” or “fuzzily”

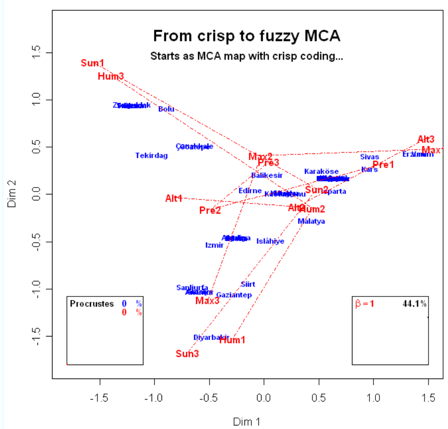
	SUN	HUM	PRE	ALT	MAX	Sun1	Sun2	Sun3	Sun1	Sun2	Sun3
Adana	7.55	66	647.1	27	45.6	0	1	0	0.000	**0.634**	0.366
Afyon	7.09	64	434.4	1034	39.8	0	1	0	0.003	**0.997**	0.000
Anamur	8.33	69	993.5	5	44.2	0	0	1	0.000	0.000	**1.000**
Ankara	7.19	60	377.7	891	40.8	0	1	0	0.000	**0.927**	0.073
Antakya	7.15	70	1124.1	100	43.9	0	1	0	0.000	**0.959**	0.041
Antalya	8.28	64	1052.3	54	45.0	0	0	1	0.000	0.041	**0.959**
Aydın	7.42	63	857.7	57	44.6	0	1	0	0.000	**0.740**	0.260
Balıkesir	6.56	70	588.5	147	43.7	0	1	0	0.182	**0.818**	0.000
Bolu	5.49	73	536.4	742	39.4	1	0	0	**0.544**	0.456	0.000
Bursa	6.35	69	696.3	100	43.8	0	1	0	0.253	**0.747**	0.000
Çanakkale	7.31	73	615.4	6	38.8	0	1	0	0.000	**0.829**	0.171
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Siirt	7.43	51	726.5	896	46.0	0	1	0	0.000	**0.732**	0.268
Sivas	6.43	64	417.0	1285	40.0	0	1	0	0.226	**0.774**	0.000
Tekirdağ	5.40	76	575.4	549	46.8	1	0	0	**0.574**	0.426	0.000
Trabzon	4.36	72	833.8	3	38.4	1	0	0	**0.926**	0.074	0.000
Şanlıurfa	8.28	49	463.1	30	38.2	0	0	1	0.000	0.041	**0.959**
Van	7.43	59	380.4	1661	37.5	0	1	0	0.000	**0.732**	0.268
Zonguldak	5.54	72	1220.2	137	40.5	1	0	0	**0.527**	0.473	0.000

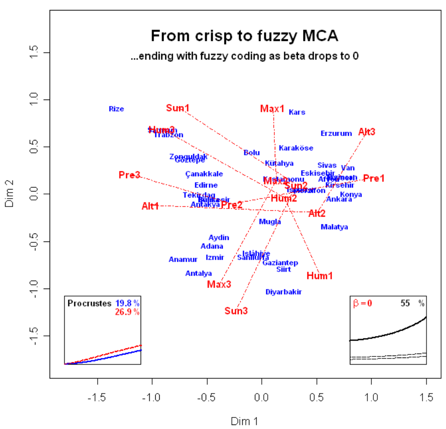
There are several ways of performing fuzzy coding. As an example, we chose the triangular membership function system shown here:



An example is given of a value below the median on the continuous scale which is coded as [0.22 0.78 0].

Crisp → fuzzy MCA

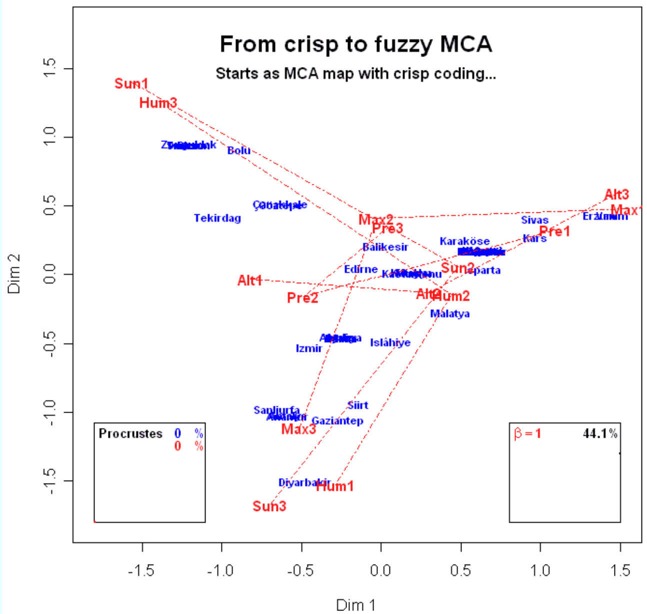




Crisp → fuzzy MCA

Data: Turkish meteorological data

The symmetric map is shown. Cities that are at identical positions in the crisp coding (having the same 5 catgories) are separated by the fuzzy coding which retains more of the information in the original data.



Distance matrices

- Consider the following table from an environmental survey (this is one of the data sets from my book *Correspondence Analysis in Practice: Second Edition*, 2007 – it is given on **www.carme-n.org**)
- The columns refer to 13 sampling sites, the first 11 labelled "1" to "11" are in the vicinity of an oil-platform in the North Sea, while the last two, R1 and R2, are reference sites far from the oil-platform
- The rows are 10 different species of benthic (sea-bed) marine life, labelled s1 to s10.

	1	2	3	4	5	6	7	8	9	10	11	R1	R2
s1	193	79	150	72	141	302	114	136	267	271	992	5	12
s2	49	30	57	34	39	63	58	71	39	68	76	25	48
s3	19	39	11	38	18	20	11	22	30	40	3	55	65
s4	9	26	5	30	35	2	11	13	5	63	1	0	1
s5	17	7	15	8	10	13	21	10	8	18	5	8	3
s6	2	12	4	12	6	7	3	10	8	12	4	2	6
s7	4	2	0	3	4	11	8	1	3	3	29	2	3
s8	7	1	6	1	3	4	2	1	8	6	6	4	6
s9	4	5	2	11	1	2	3	3	2	2	2	3	1
s10	1	5	7	1	5	4	0	1	0	4	0	0	0

Bray-Curtis dissimilarity

- The chi-square distance is used implicitly in CA, but ecologists like to use a non-Euclidean distance called the Bray-Curtis dissimilarity.
- This is a much simpler dissimilarity function to understand in the context of environmental sampling:

$$d(j,j') = 100 \times \frac{\sum_i |x_{ij} - x_{ij'}|}{\sum_i (x_{ij} + x_{ij'})}$$

- B-C = 0 if identical abundances, = 100 if no common species; B-C index of similarity is 100 minus above dissimilarity.

	1	2	3	4	5	6	7	8	9	10	11	R1	R2
s1	193	79	150	72	141	302	114	136	267	271	992	5	12
s2	49	30	57	34	39	63	58	71	39	68	76	25	48
s3	19	39	11	38	18	20	11	22	30	40	3	55	65
s4	9	26	5	30	35	2	11	13	5	63	1	0	1
s5	17	7	15	8	10	13	21	10	8	18	5	8	3
s6	2	12	4	12	6	7	3	10	8	12	4	2	6
s7	4	2	0	3	4	11	8	1	3	3	29	2	3
s8	7	1	6	1	3	4	2	1	8	6	6	4	6
s9	4	5	2	11	1	2	3	3	2	2	2	3	1
s10	1	5	7	1	5	4	0	1	0	4	0	0	0

Bray-Curtis dissimilarity values

- Since the measure is greatly affected by the variance difference between common and rare species, ecologists often take fourth roots ($\sqrt[4]{\ }$) of the abundances, and then calculate the B-C values.

$$d(j,j') = 100 \times \frac{\sum_i \left|\sqrt[4]{x_{ij}} - \sqrt[4]{x_{ij'}}\right|}{\sum_i \left(\sqrt[4]{x_{ij}} + \sqrt[4]{x_{ij'}}\right)}$$

- Part of the 13 x 13 matrix of dissimilarities is shown here:

	1	2	3	4	...	R1	R2
1	0.00	11.94	9.52	11.14	...	21.46	17.74
2	11.94	0.00	14.82	3.31	...	21.53	18.61
3	9.52	14.82	0.00	17.42	...	27.86	22.21
4	11.14	3.31	17.42	0.00	...	21.49	17.72
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
R1	21.46	21.53	27.86	21.49	...	0.00	11.33
R2	17.74	18.61	22.21	17.72	...	11.33	0.00

CA of a distance matrix

- CA has been shown to be applicable to distance or dissimilarity matrices if we convert the dissimilarities to similarities using a transformation of the form $s = k - d$ where k is a large value, at least as large as the maximum dissimilarity. Since the dissimilarities have as maximum of 100, the obvious choice is $k = 100$.
- Hence the Bray-Curtis similarities (in fact, the off-diagonal elements are now exactly the Bray-Curtis indices):

	1	2	3	4	...	R1	R2
1	100.00	89.06	90.48	88.86	...	78.54	82.26
2	89.06	100.00	85.18	96.69	...	78.47	81.39
3	90.48	85.18	100.00	82.58	...	72.14	77.79
4	88.86	96.69	82.58	100.00	...	78.51	82.28
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
R1	78.54	78.47	72.14	78.51	...	100.00	88.67
R2	82.26	81.39	77.79	82.28	...	88.67	100.00

CA map of Bray-Curtis indices