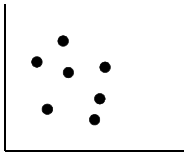
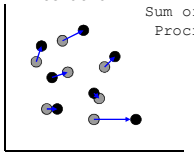
Procrustes Distance

- metaMDS uses a **Procrustean analysis** instead of stress values to assess differences in configurations for each iteration
  - Procrustes – son of Poseidon, had an iron bed that all visitors had to sleep in. He would stretch or cut limbs so they all fit.
  - Procrustes analysis – how much "cutting" and "stretching" is necessary to get one configuration (ordination) to match another.

Iteration 1



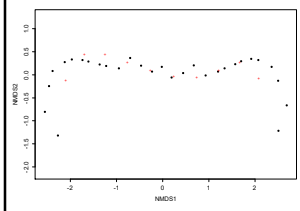
Iteration 2

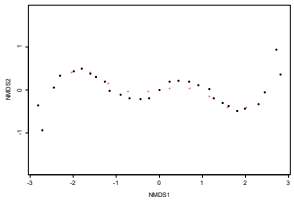


Sum of line lengths ~ Procrustes distance

Procrustes Statistic

- Any two configurations can be compared and their differences quantified with a Procrustes statistic –sum of squares of the residuals.

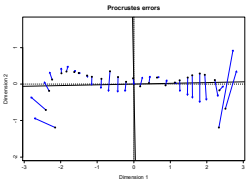




- How different are these two configurations?

Procrustes Statistic

- You can also test the significance of Procrustes values (function protest compares observed vs. permuted values). This is testing for non-random concordance between two configurations.



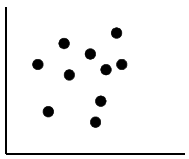

```
Call:
procrustes(X = metanmds$points, Y = metanmds_without_noshare$points)

Number of objects: 29     Number of dimensions: 2

Procrustes sum of squares:
 9.477623
Procrustes root mean squared error:
 0.571677
Quantiles of Procrustes errors:
       Min         1Q     Median         3Q        Max
0.03398588 0.14887655 0.25752112 0.53496706 2.15886845

Rotation matrix:
             [,1]         [,2]
[1,] 0.99978430 -0.02076904
[2,] 0.02076904  0.99978430

Translation of averages:
              [,1]          [,2]
[1,] 3.450472e-17 -4.280952e-18

Scaling of target:
[1] 0.9371033
```

Protest – analysis of congruence

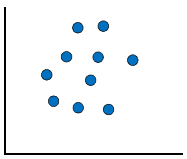
- Similar to Mantel tests, noise removed from data by working in fewer dimensions
- Generally more powerful than Mantel

Pedro R. Peres-Neto · Donald A. Jackson

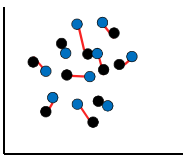
How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test



Morphological Data



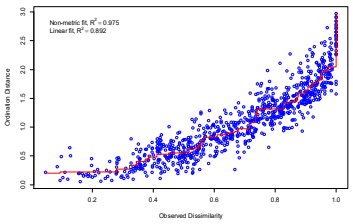
Performance Data



Permutation test of congruence

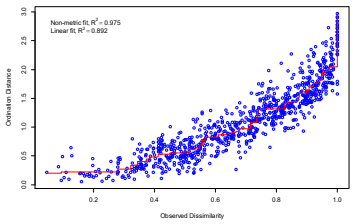
Stress

- Stress is calculated from the fit (residuals) of the original distance matrix and Euclidian distance in ordination space.
- The Non-metric fit is
  - 1-S^2
  - `1-(nmds$stress^2)`
  - `[1] 0.974788`
- The linear fit (R^2) is not the same as stress, but is analogous.
- Using R^2 as a measure of % variance explained is not precise because of some things nmds does to improve fit (noshare option etc.).



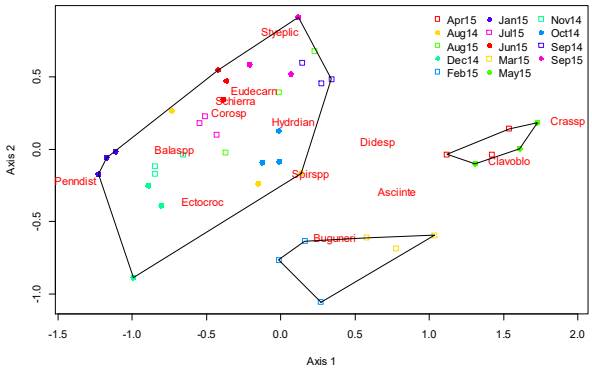
- Each point here represents a pair of sites.
- Stress is calculated from the residuals.
- Thus, you can figure out how much each site is contributing to stress.
- This is the goodness of fit measure.

$$\sum GOF^2 = Stress^2$$



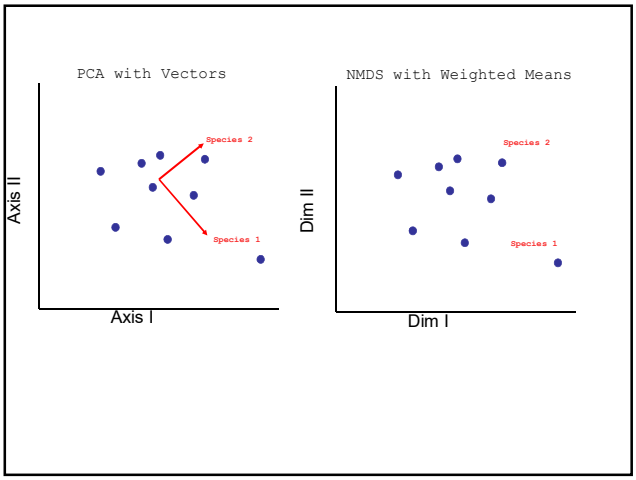

```
> sum(goodness(nmds)^2)
[1] 0.02520983
> nmds$stress^2
[1] 0.02520983
```

NMDS with weighted averages for 9 most abundant species



NMDS Issues and Questions

- Warnings…
  - Zero distance among samples
  - No species scores (ordiplot)
- Despite the lower stress, the K=2 NMDS is better if you were to present a 2D plot. The first two dimensions of K=2 is better than first two of K=4.



Correspondence Analysis (reciprocal averaging)

- Recall some of the properties (potential weaknesses) with PCA:
  - Assumes variables are linearly related with each other and/or gradients
  - **Samples are ordinated in variable space**
  - Results in "horseshoe effect" where ends of ordination axes are distorted (shared 0 seen as a similarity)
- Correspondence analysis allows for non-linear unimodal relationships
  - Both samples and variables handled similarly, axes do not explicitly represent species-space

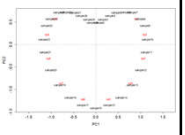
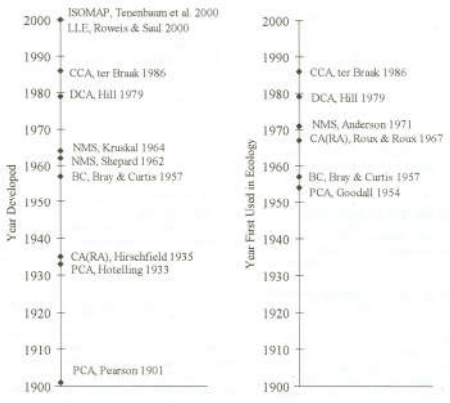


Table 9.1 from Legendre and Legendre (1998)

Method	Distance	Variables
PCA	Euclidean	Quantitative, linear relationships assumed, beware of double-zeroes
PCoA	Any	Quantitative, qualitative or mixed
NMDS	Any	Quantitative, qualitative or mixed
CA	X^2	Non-negative, quantitative or binary
FA	Euclidean	Quantitative, linear relationships assumed, beware of double-zeroes

History of Ordinations in Ecology

- From: McCune, B. and J. B. Grace. 2002. Analysis of Ecological Communities. MJM Software Design.



Correspondence Analysis

- Based on traditional Chi-Square approach to measure **correspondence** between rows and columns.

Observed

site	sp1	sp2	sp3	sp4	Row Total
sam1	1	1	2	2	6
sam2	2	2	4	4	12
sam3	3	3	6	6	18
sam4	4	4	8	8	24
sam5	5	5	10	10	30
Col Total	15	15	30	30	90

$\frac{15*12}{90}$

Expected

Expected	sp1	sp2	sp3	sp4
sam1	1	1	2	2
sam2	2	2	4	4
sam3	3	3	6	6
sam4	4	4	8	8
sam5	5	5	10	10

$\frac{(2-2)^2}{2}$

Chi-Sqr

Chi-Sqr				
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0

Chi Sqr	Inertia
0	0

Inertia = Chi Sqr/Grand Sum

In this simple dataset, rows and columns are not independent. Thus, the contents of each cell are predictable based on row and column totals and the grand total.

Correspondence Analysis

- As rows and column deviate (more independent), Chi Sqr values (and inertia) grows.

Observed

site	sp1	sp2	sp3	sp4	Row Total
sam1	4	2	3	2	11
sam2	4	3	7	4	18
sam3	25	10	12	4	51
sam4	18	24	33	13	88
sam5	10	6	7	2	25
Col Total	61	45	62	25	193

$\frac{18*45}{193}$

Expected

Expected	sp1	sp2	sp3	sp4
sam1	3.476684	2.564767	3.533679	1.42487
sam2	5.689119	4.196891	5.782383	2.331606
sam3	16.11917	11.89119	16.38342	6.606218
sam4	27.81347	20.51813	28.26943	11.39896
sam5	7.901554	5.829016	8.031088	3.238342

$\frac{(3-4.1968)^2}{4.1968}$

Chi-Sqr

Chi-Sqr				
	0.07877	0.124363	0.0806	0.232143
	0.501505	0.341336	0.256398	1.193828
	4.892877	0.300778	1.172794	1.028178
	3.462503	0.590862	0.791607	0.224873
	0.557292	0.005016	0.132378	0.473542

Chi Sqr	Inertia
16.44164	0.08519

Inertia = Chi Sqr/Grand Sum 16.44164/193 = 0.0851

Correspondence Analysis

- As rows and column deviate (more independent), chi sqr values (and inertia) grows.

Observed

site	sp1	sp2	sp3	sp4	Row Total
sam1	4	2	3	2	11
sam2	4	3	7	4	18
sam3	25	10	12	4	51
sam4	18	24	33	13	88
sam5	10	6	7	2	25
Col Total	61	45	62	25	193

Expected

Expected	sp1	sp2	sp3	sp4
sam1	3.476684	2.564767	3.533679	1.42487
sam2	5.689119	4.196891	5.782383	2.331606
sam3	16.11917	11.89119	16.38342	6.606218
sam4	27.81347	20.51813	28.26943	11.39896
sam5	7.901554	5.829016	8.031088	3.238342

Chi-Sqr

Chi-Sqr				
	0.07877	0.124363	0.0806	0.232143
	0.501505	0.341336	0.256398	1.193828
	4.892877	0.300778	1.172794	1.028178
	3.462503	0.590862	0.791607	0.224873
	0.557292	0.005016	0.132378	0.473542

Chi Sqr	Inertia
16.44164	0.08519

This matrix describes all the variability in the dataset not explainable by row or column profiles (totals).

Total variance the analysis will attempt to explain.

Correspondence Analysis

Chi-Sqr	sp1	sp2	sp3	sp4	Row totals
sam1	0.07877	0.124363	0.0806	0.232143	***0.515876***
sam2	0.501505	0.341336	0.256398	1.193828	***2.293067***
sam3	4.892877	0.300778	1.172794	1.028178	***7.394627***
sam4	3.462503	0.590862	0.791607	0.224873	***5.069845***
sam5	0.557292	0.005016	0.132378	0.473542	***1.168228***
Col totals	***9.492948***	***1.362354***	***2.433777***	***3.152565***	

Chi Sqr	Inertia
16.44164	0.08519

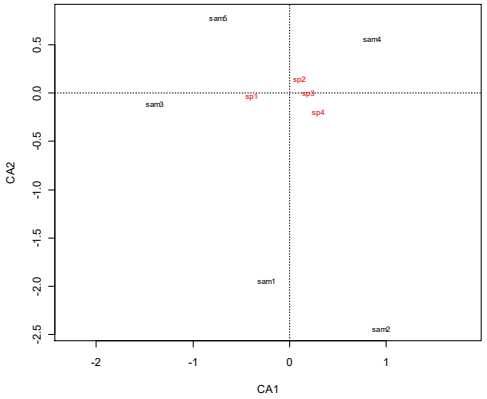
Look at where the variability is in the Chi Sqr matrix…

```
Partitioning of mean squared contingency coefficient:
            Inertia Proportion
Total       0.08519          1
Unconstrained 0.08519        1
..
Importance of components:
                          CA1    CA2      CA3
Eigenvalue             0.0748 0.0100 0.000414
Proportion Explained   0.8776 0.1176 0.004850
Cumulative Proportion  0.8776 0.9951 1.000000
..
Species scores
             CA1         CA2         CA3
sp1 -0.39131 -0.030492 -0.0008905
sp2  0.09946  0.141064  0.0219980
sp3  0.19632  0.007359 -0.0256591
sp4  0.29378 -0.197766  0.0262108

Site scores (weighted averages of species scores)
        CA1     CA2     CA3
sam1 -0.2405 -1.9357  3.4903
sam2 -0.9431 -2.4310 -1.6574
sam3 -1.3920 -0.1065 -0.2535
sam4  0.8520  0.5769  0.1625
sam5 -0.7355  0.7884 -0.3974
```

CA does an eigenvalue decomposition to summarize this variability in fewer axes (components).

Species and sites that contribute most to the inertia have the largest magnitude CA1 scores.

Scores are centered and scaled to be directly comparable.



Correspondence Analysis

- Output
  - Row and column sums, total Chi Square
  - Species and sample scores that can be plotted in the same space. Interpretation is similar to sample scores and species weighted averages in NMDS.
  - # axes = n-1 for whichever dimension of the data matrix is lower (samples or species).
  - Eigenvalues – relative importance of each axis, interpreted as the percentage of total **inertia** explained.

```
Partitioning of mean squared contingency coefficient:
             Inertia Proportion
Total          1.780          1
Unconstrained  1.780          1

Eigenvalues, and their contribution to the mean squared contingency coefficient

Importance of components:
                            CA1    CA2    CA3    CA4    CA5     CA6     CA7     CA8     CA9
Eigenvalue            0.850 0.530 0.252 0.0953 0.0303 0.01097 0.00638 0.00307 0.00179
Proportion Explained  0.478 0.298 0.142 0.0536 0.0170 0.00616 0.00358 0.00173 0.00101
Cumulative Proportion 0.478 0.775 0.917 0.9705 0.9875 0.99368 0.99727 0.99899 1.00000
```

0.85/1.780=0.478

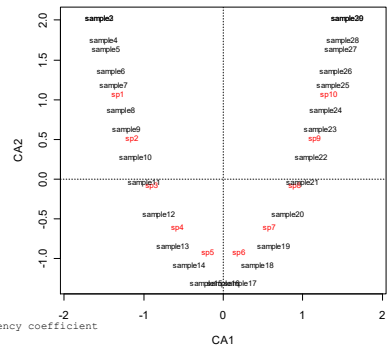
Correspondence Analysis

- Unlike NMDS
  - Not iterative, no local minima problem
  - First axis always most informative
  - Number of axes produced is set by the dimensionality of the data (n-1), not a user option
  - Not distance based, data transformations typically more important
  - Row and column sums must be >0. If your dataset includes negatives, think carefully about whether CA is appropriate.
  - No missing data allowed (like PCA)
  - Data expected to be frequency-based (contingency table)
- Ordinates both samples and variables directly
- Also the bases of the most popular direct gradient analysis (CCA)

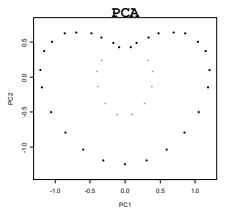
CA Code

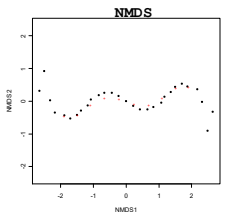
- Correspondence Analysis is run as an **unconstrained** cannonical correspondence analysis (CCA)
- CCA function with no environmental matrix
  - Code
    - `ca<-cca(community)`
  - Options
    - All options for this function apply to CCA
    - Transformations to raw data are often used before
      - Log or eliminating rare species
      - Scale option – scale species data to unit variance
      - Na.action – how to handle missing data
    - Function `downweight(community, fraction=x)`

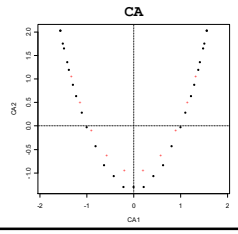
CA Example




```
cca(X = community)

              Inertia Rank
Total            1.78
Unconstrained    1.78    9
Inertia is mean squared contingency coefficient

Eigenvalues for unconstrained axes:
     CA1      CA2      CA3      CA4      CA5      CA6      CA7      CA8
0.849792 0.529798 0.252153 0.095300 0.030295 0.010971 0.006376 0.003073
     CA9
0.001793
```



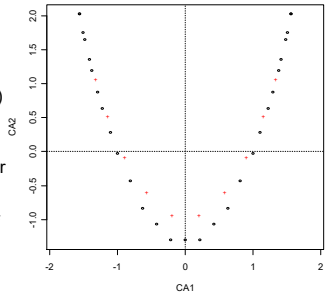




Review of unconstrained ordinations of our sample dataset. Recall that the sample dataset had one strong gradient.

CA Example

- The "arch effect" here is unwanted.
- The ends of the axes are also compressed
- Detrending (detrended correspondence analysis, DCA) deals with the arch by:
  - 5 segment smoothing of 1st axis. Divide into segments (weights of 1,2,3,2,1), center each at 0.
  - Rescaling of axis into "standard deviation" units of species turnover.
- Assumptions
  - Same as for CA
  - DCA is not really an analysis. It is a post hoc modification of a CA.
  - "vague bag of tricks"



DCA

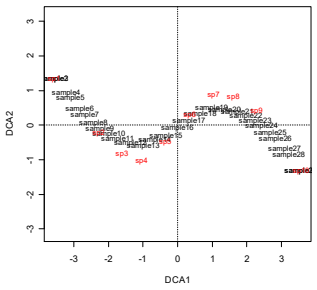
- DCA (function decorana) – detrending and rescaling of CA
- Only first 4 axes are adjusted, the rest are discarded
- Units on axes are in SD of species turnover, beta diversity in samples measured in length of 1st axis
- Code
  - `decorana(community)`
- Options
  - Downweight rare species (uses downweight function)
  - number of rescaling iterations
  - number of detrending segments to use
  - Whether or not to detrend at all (avoid detrending = regular CA)

DCA Output

- For first 4 axes only:
  - Eigenvalues
  - Variable scores
  - Site scores
  - "Decorana Values" – values estimated before detrending…interpretation unclear

DCA Example

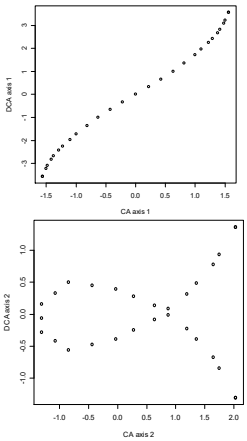
- Arch effect removed
- First axis a good representation of the original gradient
- Species evenly distributed along first axis
- First axis length = 6, indicates complete turnover in variables (species)
- Second axis length ~ 2

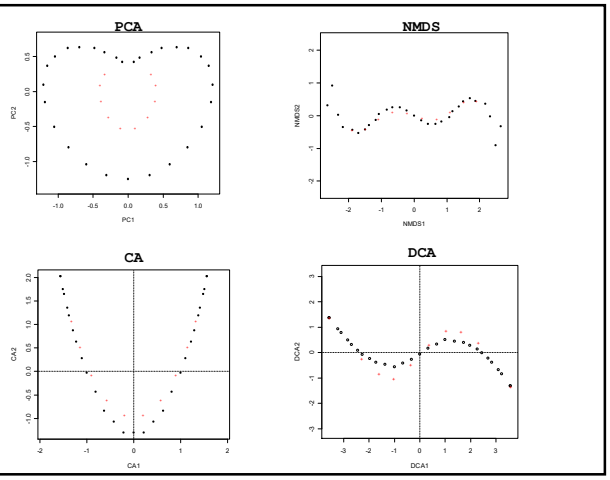


Beta diversity ~ length of 1st axis

CA and DCA

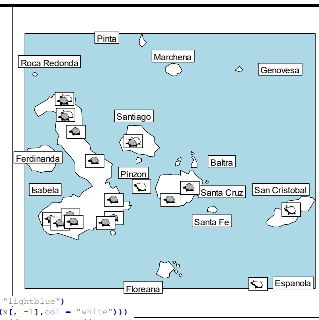
- First axis is usually fine (not very different from CA axis 1), second is often problematic.
- Problems – adjustments made during detrending are arbitrary.





Assignment

- Reading – section 5.4
- Luczkovich,J., G. Ward, J. Johnson, R. Christian, D. Baird, H. Neckles and W. Rizzo. 2002. Determining the trophic guilds of fishes and macroinvertebrates in a seagrass food web. Estuaries and Coasts 25(6): 1559-2723.
- Assignment
  - Use the Galapagos Island microsatellite dataset in the ade4 package.
    - `library(ade4)`
    - `data(ggtortoises)`
  - Data:
    - `ggtortoises$tab`
  - Subspecies:
    - `ggtortoises$pop$subspe`
  - Eliminate alleles with less than 2 occurrences
  - Do a correspondence analysis
    - Which markers are most closely associated with the first two axes?
    - Which populations are most distinct (describe how populations are related based on this dataset)?
  - Plot populations with a few of the most important markers.




```
library(ade4)
data(ggtortoises)
a1 <- ggtortoises$area
area.plot(a1)
rect(min(a1$x), min(a1$y), max(a1$x), max(a1$y), col = "lightblue")
invisible(lapply(split(a1, a1$id), function(x) polygon(x[, -1],col = "white")))
s.label(ggtortoises$misc, grid=FALSE,include.ori=FALSE,addaxes=FALSE,add.p=TRUE)
listico <- ggtortoises$ico[as.character(ggtortoises$pop$carap)]
s.logo(ggtortoises$pop,listico,add.p=TRUE)

library(vegan)

msat_data<-ggtortoises$tab
```

Allele frequency for 10 msat loci
Load the vegan library last – package ade4 has a cca function that works differently

Microsatellite analysis of genetic divergence among populations of giant Galápagos tortoises

CLAUDIO CIOFI, MICHEL C. MILINKOVITCH, JAMES P. GIBBS, ADALGISA CACCONE and JEFFREY R. POWELL

Table 1 Characteristics of microsatellite loci in giant Galápagos tortoises. All loci used in the analysis were simple CA repeat arrays. T_A is the optimal annealing temperature

Locus	No. of dinucleotide repeats in the clone	Forward primer sequence (5′–3′)	Reverse primer sequences (5′–3′)	No. of alleles amplified	Allele size range (base pairs)	T_A (°C)
GAL 45	17	[illegible]	[illegible]	13	87–123	62
GAL 50	24	[illegible]	[illegible]	37	96–182	62
GAL 73	24	[illegible]	[illegible]	20	78–126	57
GAL 75	24	[illegible]	[illegible]	22	73–149	57
GAL 94	18	[illegible]	[illegible]	13	85–111	54
GAL 100	26	[illegible]	[illegible]	19	100–156	57
GAL 127	21	[illegible]	[illegible]	31	97–175	57
GAL 136	20	[illegible]	[illegible]	12	73–101	54
GAL 159	24	[illegible]	[illegible]	19	83–123	57
GAL 263	17	[illegible]	[illegible]	25	80–164	62

	GAL45.87	GAL45.95	GAL45.103	GAL45.105	GAL45.107	GAL45.109	GAL45.111	GAL45.113	GAL45.115	GAL45.117	GAL45.119	GAL45.121	GAL45.123
PBL	11	0	0	0	23	6	0	0	0	0	0	0	0
PBR	7	12	0	0	35	0	0	0	0	0	0	0	0
VD	2	0	0	1	27	9	0	0	7	0	2	0	0
VA	8	0	0	0	21	2	9	0	0	0	0	0	0
CAZ	3	0	0	0	5	0	6	0	14	0	0	10	2
RU	4	0	0	0	11	0	13	2	5	0	0	5	0
CR	6	0	0	0	8	2	7	3	9	2	0	3	0
LP	7	0	0	0	18	0	7	2	2	0	0	4	0
LT	14	0	0	0	12	0	4	0	8	0	0	2	0
LC	11	0	0	0	6	0	16	1	6	0	0	0	0
PEG	13	0	0	0	14	1	7	2	0	0	0	1	0
CRU	31	0	0	0	3	0	5	1	0	0	0	0	0
CF	27	0	0	0	5	0	0	0	0	0	0	0	0
PZ	7	0	0	4	28	1	0	0	0	0	0	0	0
AGO	2	0	4	0	20	8	5	0	1	0	0	0	0
SCR	7	10	4	0	16	0	3	0	0	0	0	0	0
ESP	1	10	0	0	19	0	0	0	0	0	0	0	0