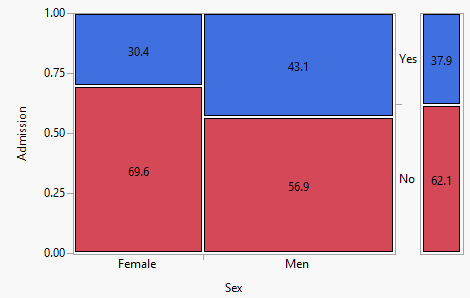
**6.4 - Multiple Correspondence Analysis (MCA) on Higher Dimensional Tables**Not all tables are two dimensional. As a simple example consider the following data dealing with graduate school admissions at the University of California – Berkeley. The categorical variables are gender of the applicant, the program they are applying to, and whether or not they were admitted to the program. Thus we have three categorical/nominal variables to consider in our analysis that can stored in a data *array*.

Example 6.2: - Gender Discrimination in Graduate School Admissions at Berkeley

> UCBAdmissions

, , Dept = A


Research Question: Is there evidence of gender discrimination in graduate admissions at UCB?

Gender

Admit Male Female

Admitted 512 89

Rejected 313 19

, , Dept = B

Gender

Admit Male Female

Admitted 353 17

Rejected 207 8

, , Dept = C

Gender

Admit Male Female

Admitted 120 202

Rejected 205 391

, , Dept = D

Gender

Admit Male Female

Admitted 138 131

Rejected 279 244

, , Dept = E

Gender

Admit Male Female

Admitted 53 94

Rejected 138 299

, , Dept = F

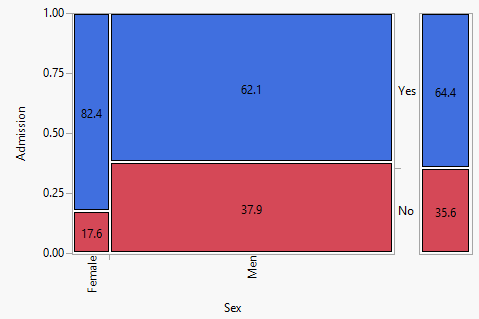
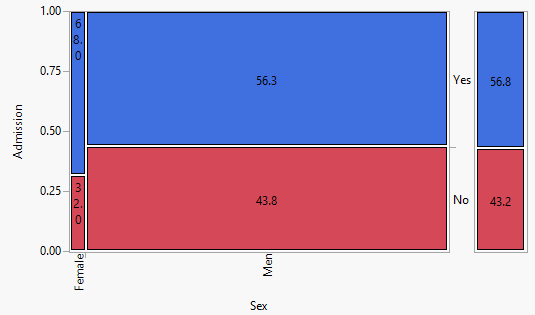
Gender

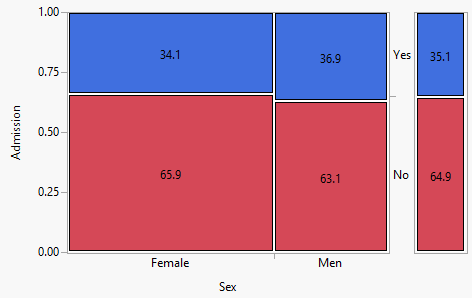
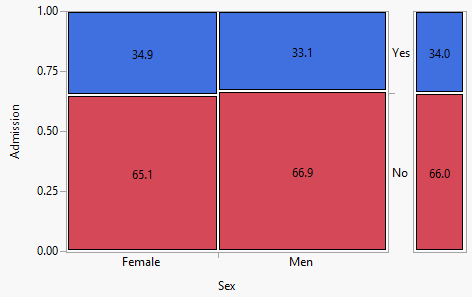
Admit Male Female

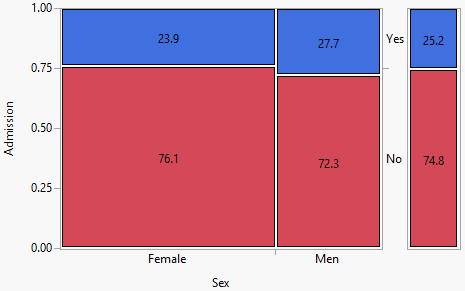
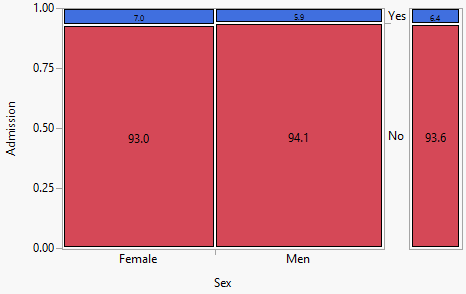
Admitted 22 24

Rejected 351 317

It appears that there is gender discrimination in graduate school admissions at UCB. However there is third dimension that is being completely ignored here, namely the programs (departments) the students are applying to. If we take department applied to into account this what we see when examining 2-D mosaic plots for each department separately.

Conditioning on the program applied to, it is not clear there is gender discrimination. A higher percentage of female applicants are admitted in 4 of the 6 programs! This is an example of what is commonly referred to as *Simpson’s Paradox*. Multiple correspondence analysis will allow us to construct a 2-D display, which represents a dimension reduction, showing the relationship between the three categorical variables in these data. The package ca available from CRAN contains functions for conducting correspondence analysis, including multiple correspondence analysis (MCA), from data stored in a contingency table format.

> ucb.mca = mjca(UCBAdmissions)

> summary(ucb.mca)

Principal inertias (eigenvalues):

dim value % cum% scree plot

1 0.114945 80.5 80.5 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2 0.005694 4.0 84.5 \*

3 00000000 0.0 84.5

4 00000000 0.0 84.5

5 00000000 0.0 84.5

-------- -----

Total: 0.142840

Columns:

name mass qlt inr k=1 cor ctr k=2 cor ctr

1 | Admit:Admitted | 129 911 93 | 365 875 150 | 74 36 123 |

2 | Admit:Rejected | 204 911 59 | -231 875 95 | -47 36 78 |

3 | Gender:Female | 135 863 95 | -399 845 187 | 59 19 84 |

4 | Gender:Male | 198 863 65 | 272 845 127 | -40 19 57 |

5 | Dept:A | 69 838 117 | 512 837 156 | 13 1 2 |

6 | Dept:B | 43 829 124 | 573 824 123 | -45 5 15 |

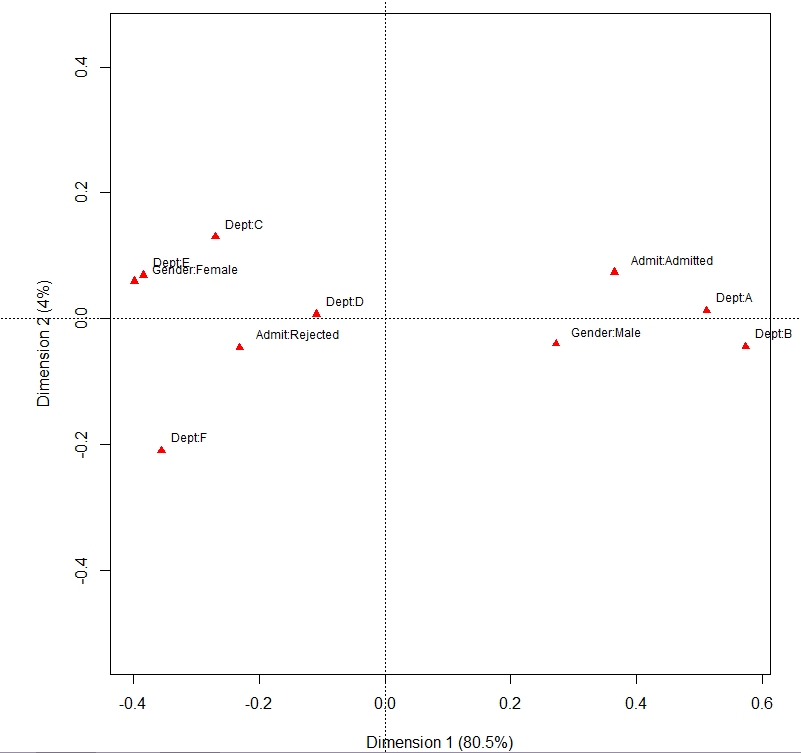
7 | Dept:C | 68 731 108 | -270 594 43 | 130 137 199 |

8 | Dept:D | 58 832 106 | -110 828 6 | 7 3 0 |

9 | Dept:E | 43 812 117 | -384 787 55 | 69 25 35 |

10 | Dept:F | 53 737 116 | -355 547 58 | -210 190 406 |

> plot(ucb.mca)



How does this plot help explain the apparent admission bias?

**6.5 – Multiple Correspondence Analysis (MCA) for Individual Level Data**

Multiple Correspondence Analysis (MCA) is also suited for individual level data where the rows of the data table correspond to individuals and the columns correspond to the observed categories for the individuals on multiple qualitative variables. The format of data appropriate for this type of MCA consists of qualitative (nominal or ordinal) variables measured on individuals. Continuous variables could also be used if they are to converted non-overlapping intervals, e.g. age ranges (20-29 years, 30-39 years, etc.).   
  
Thus the data format appropriate for conducting MCA with data of this type might be something like:

ID Reading Listening music Cinema Show Exhibition … Sex Age Occupation

1 1 1 1 1 1 … M (45,55] Professional

2 1 0 0 0 1 … F (55,60] Professional

3 1 1 0 0 1 … M (40,45] Laborer

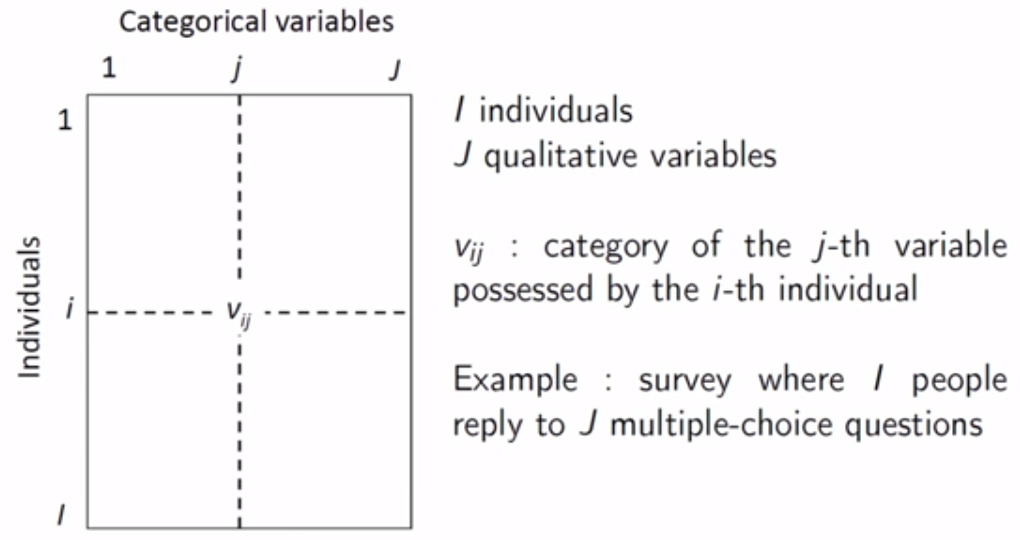
4 1 0 0 0 1 … F (60,65] Retired

5 1 1 1 1 0 … F (20,25] Student

6 0 0 0 0 0 … M (50,55] Laborer

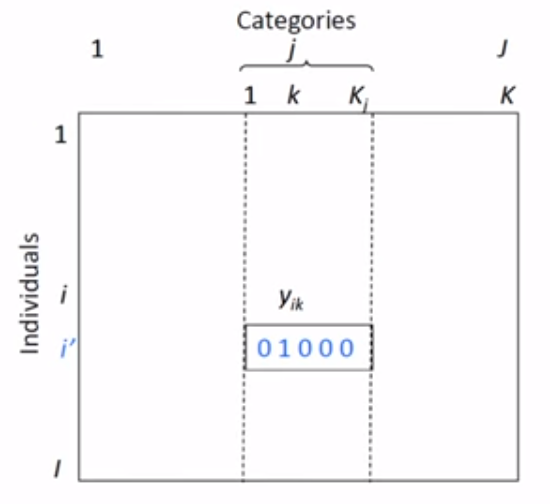
Here the activity participation variables are coded as: 1 = Yes, 0 = No

More generally, the data structure is as shown below.



As these data are non-numeric, in order to reduce the dimensionality using SVD (i.e. PCA) we need to convert it to numeric, which we do by using dummy variables for each possible level of each the categorical variables.

Suppose we have qualitative variables measured for each of individuals and that the variable has levels. We can form a matrix based on these data by using dummy variables to represent each of the possible levels for all of the qualitative variables. The total number of columns in this matrix will be and the number of rows will obviously be the number of individuals . A generic matrix of this form is shown below.



This matrix is called the **complete disjunctive table** or CDT. Performing SVD (PCA) on a properly scaled version of this matrix will produce a lower dimensional representation of the individuals and the variables in this table. The elements of this matrix are all 0 or 1 as defined below:

To standardize the we need only center them as the dummy variables only take on values . To do this first define,

If we take the mean of the values (i.e. summing down the columns) we obtain,

thus to center the we need only to subtract 1, therefore we finally define which will produce “standardized” versions of the variables in the CDT.

We can then perform PCA via SVD (or spectral decomposition) of the scaled data matrix to perform dimension reduction. We can then use the lower dimensional representation to explore inter-individual variability in terms of the categories of the qualitative variables. By examining category loadings we can explore associations between the qualitative variables, visualize the associations between the categories, and interpret the synthetic variables (i.e. dimensions created) as quantitative indices based upon the qualitative variables.

Example 6.3: INSEE “History of Life” Survey Extract Pertaining to Hobbies

There were 8,303 individuals surveyed on their involvement in a variety of hobbies listed on the survey. Also the sex, age, marital status, and profession of the individuals was recorded. Our goal is understand inter-individual variability in terms of their hobby choice profiles, how the hobbies relate to one another, and how the demographics of the individuals potentially relate to their hobby choices.

These data are part of the FactoMineR package which has functionality for conducting a thorough multiple correspondence analysis and achieving the goals stated above. The package ade4 also has functions for creating the complete disjunctive table and performing MCA. There also other packages that have MCA.

> library(FactoMineR)  
> head(hobbies)

Reading Listening music Cinema Show Exhibition Computer Sport Walking

11000210 1 1 1 1 1 0 1 1

11000410 1 0 0 0 1 0 1 1

11000610 1 1 0 0 1 0 0 0

11000710 1 0 0 0 1 0 1 0

11000810 1 1 1 1 0 0 0 1

11000910 0 0 0 0 0 0 0 0

Travelling Playing music Collecting Volunteering Mechanic Gardening

11000210 1 0 0 1 1 0

11000410 0 0 1 1 1 1

11000610 1 0 0 0 0 0

11000710 1 0 0 0 0 0

11000810 0 0 0 0 0 0

11000910 0 0 0 0 1 0

Knitting Cooking Fishing TV Sex Age Marital status Profession

11000210 0 0 0 2 F (55,65] Married Management

11000410 0 0 0 4 M (45,55] Married <NA>

11000610 0 0 0 4 F (25,35] Remarried Management

11000710 0 0 0 1 M (75,85] Married <NA>

11000810 0 0 0 3 M (55,65] Married Employee

11000910 0 0 0 3 M (45,55] Married Manual labourer

nb.activitees

11000210 11

11000410 9

11000610 5

11000710 5

11000810 6

11000910 2

> summary(hobbies)

Reading Listening music Cinema Show Exhibition Computer Sport Walking

0:2757 0:2456 0:5044 0:5978 0:5808 0:5245 0:5308 0:4228

1:5646 1:5947 1:3359 1:2425 1:2595 1:3158 1:3095 1:4175

Travelling Playing music Collecting Volunteering Mechanic Gardening Knitting

0:5040 0:6943 0:7541 0:7118 0:4864 0:5047 0:6990

1:3363 1:1460 1: 862 1:1285 1:3539 1:3356 1:1413

Cooking Fishing TV Sex Age Marital status

0:4717 0:7458 0:1017 M:3787 (45,55]:1837 Single :2140

1:3686 1: 945 1:1223 F:4616 (35,45]:1646 Married :4333

2:2156 (25,35]:1302 Widower : 734

3:1775 (55,65]:1257 Divorcee : 792

4:2232 (65,75]: 937 Remarried: 404

[15,25]: 857

(Other): 567

Profession nb.activitees

Employee :2552 Min. : 0.000

Manual labourer :1161 1st Qu.: 4.000

Management :1052 Median : 7.000

Unskilled worker: 792 Mean : 6.866

Foreman : 735 3rd Qu.: 9.000

(Other) : 613 Max. :16.000

NA's :1498

We can see that Reading – Fishing are dichotomous (1 = yes, 0 = no), TV is ordinal (0 = no TV, …, 4 = Lot’s of TV), there are additional self-explanatory demographic variables, and a variable called nb.activities which is the number of activities/hobbies the individual says they participate in. We will first perform MCA using only the activity/hobby choices along with their degree of TV viewing.

> names(hobbies)

[1] "Reading" "Listening music" "Cinema" "Show"

[5] "Exhibition" "Computer" "Sport" "Walking"

[9] "Travelling" "Playing music" "Collecting" "Volunteering"

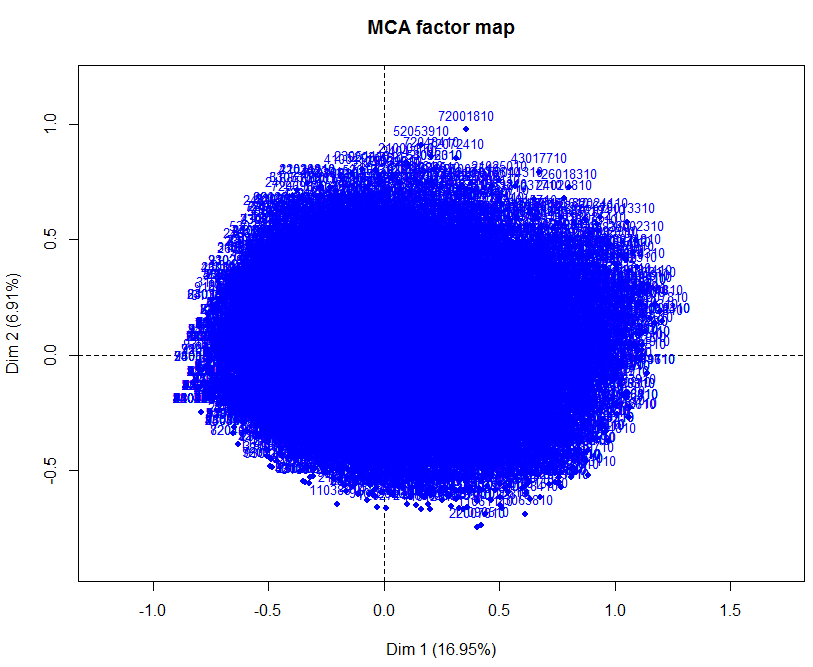
[13] "Mechanic" "Gardening" "Knitting" "Cooking"

[17] "Fishing" "TV" "Sex" "Age"

[21] "Marital status" "Profession" "nb.activitees"

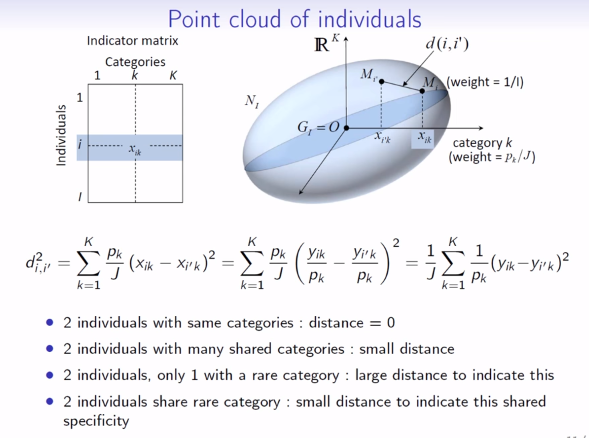
> X = hobbies[,1:18]

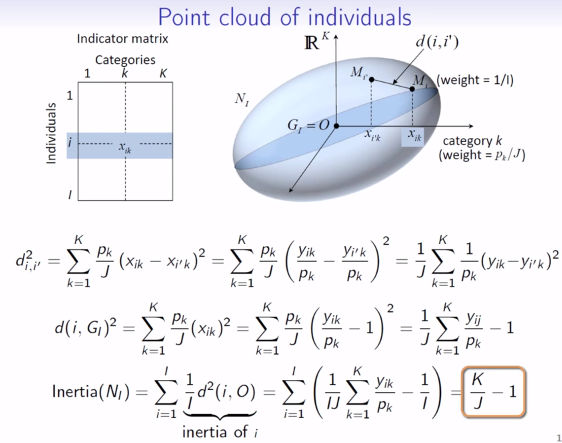
> X.mca = MCA(X)



The plot above shows the point cloud of the individuals in this survey. Nothing particularly stands out in this plot. However, at this point, we don’t understand what determines the location of an individual within this plot.

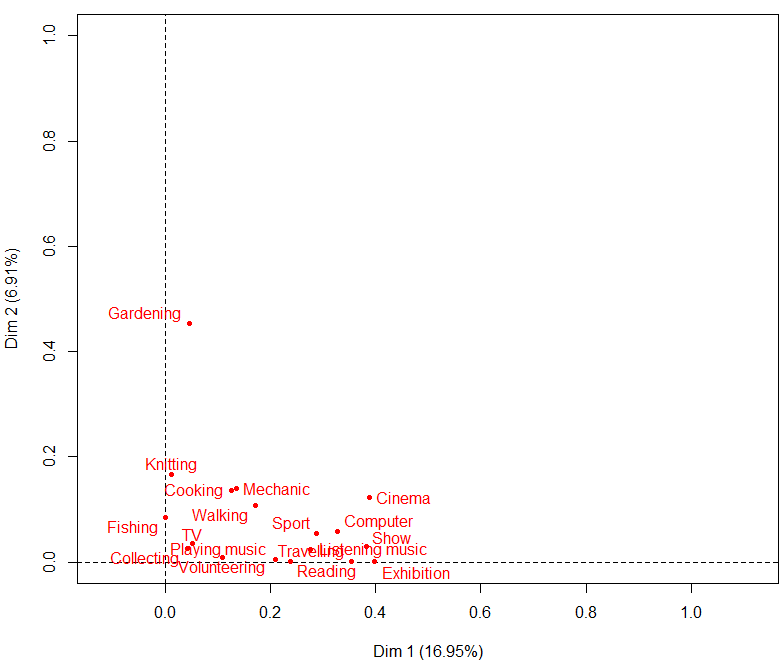
To better understand the individual point cloud we consider the distance between the individuals in the full dimensional data space and the two dimensional representation above. Recall the scaled indicator matrix is made up of the and we define the distance between observations using the formula shown below.



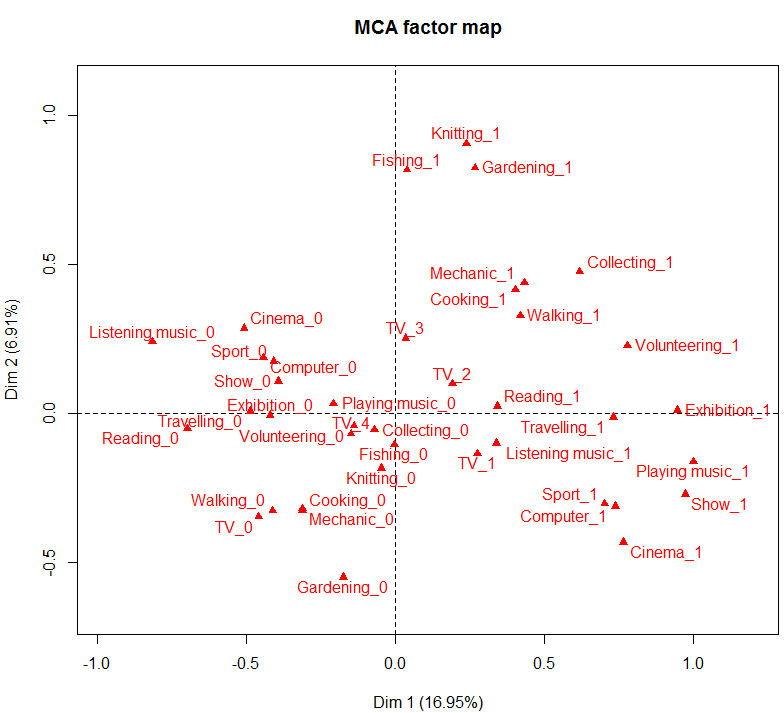


Images taken from MCA part 2 – visualizing the point cloud of individuals by Francois Husson (<https://www.youtube.com/watch?v=b4kRAt4mkB8&index=2&list=PLnZgp6epRBbTVjKd_-KPhaGWLE7K7InL6> )

Considering now the qualitative variables and their levels which are the columns of the CDT, we obtain the following plots from multiple correspondence analysis.



Correlations between the qualitative variables and the first two dimensions from the MCA dimension reduction are displayed in this plot. The qualitative variables with the highest correlations with a given dimension, indicate the most important variables in determining that dimension.



Visualization of the individual levels of the qualitative variables on the first two dimensions. This is akin to the biplot from a PCA. This plot helps explain the location of individuals in the quadrants shown in point cloud of the individuals above.

To superimpose the individuals and qualitative variable categories in the same plot we use the SVD details of the scaled complete disjunctive table (CDT) with entries . To construct such a plot we first extract the SVD information from our X.mca object.

> X.svd = X.mca$svd

> attributes(X.svd)

$names

[1] "vs" "U" "V"

> delta = X.svd$vs

To plot the columns, i.e. the levels of the qualitative variables, we plot the scaled columns of .   
  
Note: in order to label the levels of the categorical variables we have extract the column labels from different component of the MCA results stored in X.mca.

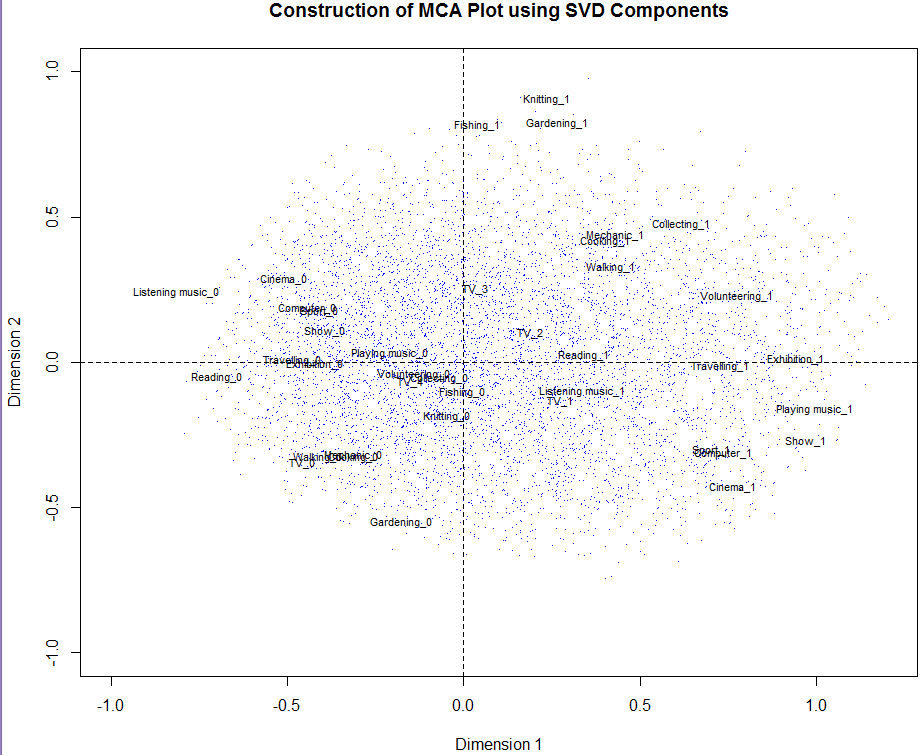
> plot(delta[1]\*V[,1],delta[2]\*V[,2],type=”n”,xlim=c(-1.1,1.1),ylim=c(-1.1,1.1))

> text(delta[1]\*V[,1],delta[2]\*V[,2],labels=row.names(X.mca$var$coord),  
 cex=.7,col=”black”)

Next we add vertical & horizontal axes to the plot and then plot rows, i.e. the individuals, by adding points corresponding to the scaled columns of .

> abline(h=0,v=0,lty=2)

> points(delta[1]\*U[,1],delta[2]\*U[,2], pch=".",col="blue")



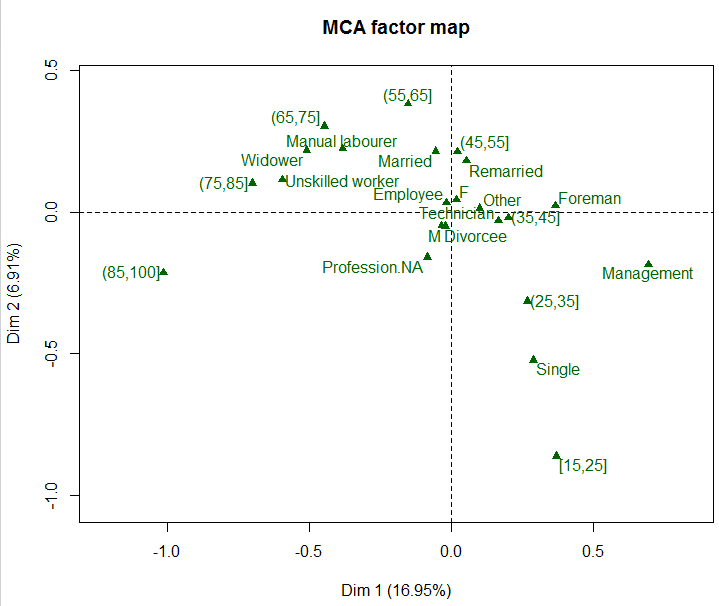
The FactoMineR package has some specialized plotting functions for creating plots similar to one shown above, as well as plots that bring the supplementary demographic variables into the displays. In order to include supplementary variables in our MCA, we need to supply those variables as arguments to the original MCA function call. The supplementary variables of interest for these data are the qualitative variables sex, age (in class intervals), profession, marital status, and a quantitative variable representing the total number of activities the respondent participates in. We will examine several plots to aid in our interpretation of these data below.

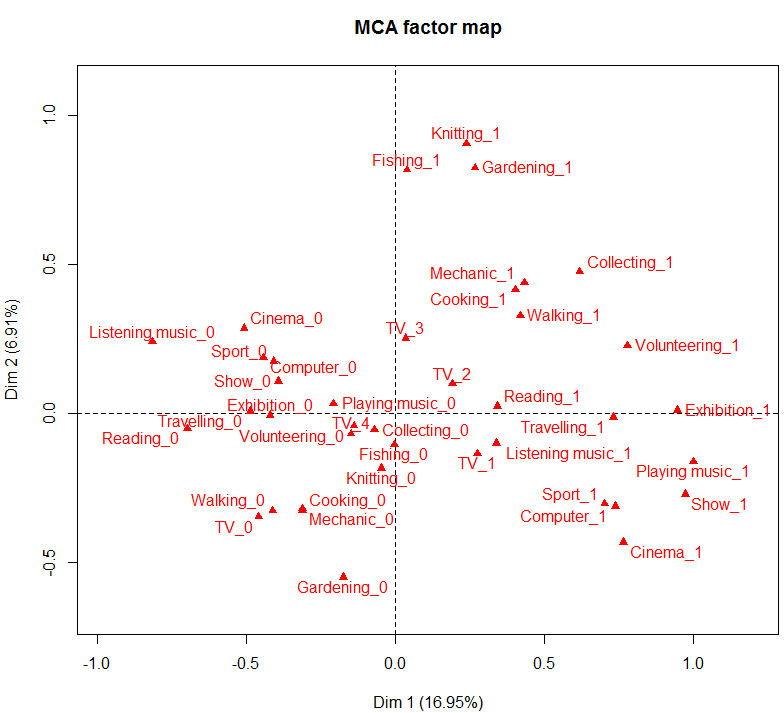
> X.mca.supp = MCA(hobbies,quali.sup=19:22,quanti.sup=23)

To focus only on the supplementary variable we need to make the individuals and the levels of the active variables representing their activity/hobby choices invisible.

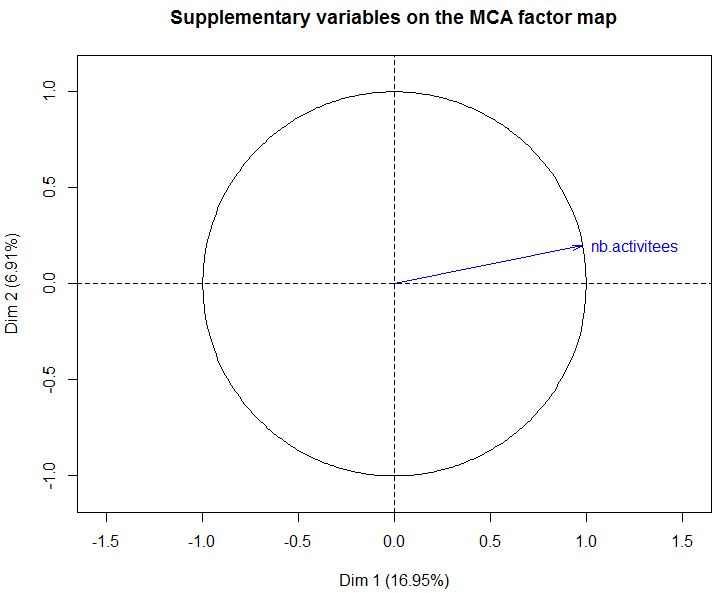
> plot(X.mca.supp,invisible=c("ind","var"))

Below is the resulting the plot displaying the role or association of the supplementary variables with the two dimensions created from MCA.

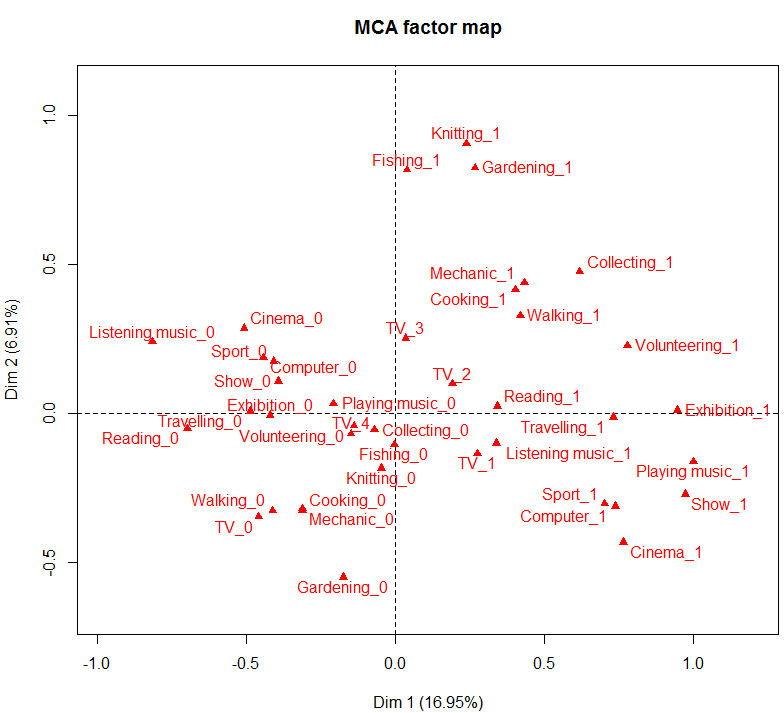




Individuals in the lower right quadrant represent younger single professional (management) individuals. From the plots of the levels of the active variables regarding an individual’s hobby choices we know that these individuals participate in activities such as going to shows, listening to and playing music, watching or participating in sports, and using computers.



The quantitative supplementary variable number of hobby activities participated in is most associated with the first dimension with a correlation near 1. Thus individuals in the two quadrants on the left participate in fewer activities than individuals in the two rightmost quadrants. This is association is not surprising given that all of the hobby choices corresponding to 1 = yes are located in the right side of the factor map.



There are numerous summary statistics and quality measures that can be examined to aid in interpretation of the results from an MCA.

> summary(X.mca)

Eigenvalues

Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6

Variance 0.198 0.081 0.072 0.063 0.058 0.056

% of var. 16.947 6.913 6.173 5.389 5.011 4.784

Cumulative % of var. 16.947 23.859 30.033 35.422 40.433 45.217

Dim.7 Dim.8 Dim.9 Dim.10 Dim.11 Dim.12

Variance 0.056 0.053 0.053 0.049 0.046 0.045

% of var. 4.759 4.569 4.547 4.211 3.985 3.864

Cumulative % of var. 49.976 54.545 59.092 63.303 67.288 71.152

Dim.13 Dim.14 Dim.15 Dim.16 Dim.17 Dim.18

Variance 0.044 0.043 0.041 0.038 0.037 0.036

% of var. 3.730 3.717 3.497 3.256 3.200 3.105

Cumulative % of var. 74.881 78.598 82.095 85.351 88.551 91.655

Dim.19 Dim.20 Dim.21

Variance 0.035 0.032 0.030

% of var. 2.997 2.772 2.575

Cumulative % of var. 94.652 97.425 100.000

The % of variance explained (i.e. inertia) for each dimension in an MCA is presented first. These may seem small as the first two dimensions have a combined inertia of 23.86% and thus explain this percentage of the total variation in these data. This is well short of the 90% rule of thumb we often use when conducting a PCA. In MCA the individuals are originally in dimensional space which implies that generally that the variance explained by the first dimension will be small. In fact one can show that the maximal variance explained by any single dimension in an MCA is given by . Suppose we have qualitative variables with a total number levels, then the maximal variance explained by a dimension is 10/90 = .111 or 11.1%. Here we have variables with all but TV (5 levels) having 2 levels for a total of levels. The maximal variation explained by a single dimension for the MCA of these data is therefore . Our first dimension here is well short of that mark. The average inertia for all dimension is , thus any dimension with a eigenvalue/inertia/variance over .056 or 5.6% could potentially be interpreted.

Individuals (the 10 first)

Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3

11000210 | 0.667 0.027 0.336 | -0.191 0.005 0.027 | 0.147

11000410 | 0.140 0.001 0.011 | 0.434 0.028 0.108 | 0.163

11000610 | -0.155 0.001 0.032 | -0.244 0.009 0.079 | -0.293

11000710 | -0.108 0.001 0.011 | -0.285 0.012 0.073 | 0.000

11000810 | -0.022 0.000 0.001 | -0.268 0.011 0.087 | -0.225

11000910 | -0.636 0.024 0.449 | 0.019 0.000 0.000 | 0.192

11001010 | -0.206 0.003 0.046 | -0.239 0.008 0.063 | 0.319

11001110 | 0.284 0.005 0.065 | -0.611 0.055 0.304 | -0.066

11001210 | 0.598 0.021 0.261 | -0.577 0.049 0.243 | 0.028

11001310 | 0.204 0.003 0.033 | -0.015 0.000 0.000 | 0.089

ctr cos2

11000210 0.004 0.016 |

11000410 0.004 0.015 |

11000610 0.014 0.113 |

11000710 0.000 0.000 |

11000810 0.008 0.061 |

11000910 0.006 0.041 |

11001010 0.017 0.111 |

11001110 0.001 0.004 |

11001210 0.000 0.001 |

11001310 0.001 0.006 |

As there are I = 8,403 individuals in these data in-depth analyses of the quality of representation (cos2) and contribution (ctr) for each individual would be an exercise in futility. That being said, let’s consider what these quantities represent. The cos2 for an individual can be added across the first two or three dimensions to determine how well the individual is represented in the lower dimensional space. It measures the angle between the vector from the origin to observation in original data space and the vector from the origin to point in the projected space. If the point is well represented in the lower dimensional space this angle will be near 0 and thus the cos2 will be close to 1. For example for the individuals highlighted above we see that for individual 11001210 the sum of the cos2 value is which indicates this point is fairly well represented and individual 11000810 is poorly represented with a total of .088.

The contributions are the squared coordinate on the dimension of interest divided by the eigenvalue (i.e. variance) for that dimension and is therefore a percentage. We don’t generally expect a single individual to have a large contribution to a particular dimension, particularly when the number of individuals is large, but when it does happen we know that without the individual in the data the dimension in question may change. We can explore this by performing an MCA without the individual in question and note any substantive changes in our results.

Categories (the 10 first)

Dim.1 ctr cos2 v.test Dim.2 ctr

Reading\_0 | -0.699 4.503 0.239 -44.766 | -0.051 0.058

Reading\_1 | 0.341 2.199 0.239 44.766 | 0.025 0.028

Listening music\_0 | -0.817 5.478 0.275 -48.111 | 0.241 1.170

Listening music\_1 | 0.337 2.262 0.275 48.111 | -0.100 0.483

Cinema\_0 | -0.509 4.369 0.389 -57.170 | 0.287 3.398

Cinema\_1 | 0.764 6.561 0.389 57.170 | -0.430 5.103

Show\_0 | -0.394 3.109 0.383 -56.753 | 0.109 0.586

Show\_1 | 0.972 7.663 0.383 56.753 | -0.270 1.444

Exhibition\_0 | -0.422 3.461 0.399 -57.885 | -0.005 0.001

Exhibition\_1 | 0.945 7.745 0.399 57.885 | 0.012 0.003

cos2 v.test Dim.3 ctr cos2 v.test

Reading\_0 0.001 -3.255 | 0.460 5.367 0.104 29.496 |

Reading\_1 0.001 3.255 | -0.225 2.621 0.104 -29.496 |

Listening music\_0 0.024 14.202 | 0.231 1.207 0.022 13.630 |

Listening music\_1 0.024 -14.202 | -0.096 0.498 0.022 -13.630 |

Cinema\_0 0.123 32.200 | -0.045 0.093 0.003 -5.030 |

Cinema\_1 0.123 -32.200 | 0.067 0.139 0.003 5.030 |

Show\_0 0.029 15.735 | 0.032 0.056 0.002 4.582 |

Show\_1 0.029 -15.735 | -0.078 0.137 0.002 -4.582 |

Exhibition\_0 0.000 -0.748 | 0.076 0.306 0.013 10.384 |

Exhibition\_1 0.000 0.748 | -0.169 0.684 0.013 -10.384 |

The interpretation of the cos2 and contribution measures for the variables is essentially the same as for individuals. In the case of contribution (ctr) we are not concerned when certain variables have relatively high contributions on a particular dimension, rather we use this to aid in our dimension interpretations. The v-test conducts a two-sample t-test for comparing the dimension coordinate means for dichotomous qualitative variables and a one-way ANOVA F-test for qualitative variables with three or more levels.

Categorical variables (eta2)

Dim.1 Dim.2 Dim.3

Reading | 0.239 0.001 0.104 |

Listening music | 0.275 0.024 0.022 |

Cinema | 0.389 0.123 0.003 |

Show | 0.383 0.029 0.002 |

Exhibition | 0.399 0.000 0.013 |

Computer | 0.327 0.058 0.041 |

Sport | 0.287 0.053 0.062 |

Walking | 0.172 0.107 0.002 |

Travelling | 0.355 0.000 0.000 |

Playing music | 0.209 0.005 0.002 |

These squared correlations speak to the importance of each variable on the dimensions.

Example 6.4: International Attitudes about Science  
International survey on attitudes towards science in the form of four Likert scale items (1 = strongly disagree,…, 5 = strongly agree (items A-D)) and demographic info for the respondents (sex, age, and education level).

> library(ca) 🡨 contains wg93 data frame

> library(FactoMineR)

> head(wg93) # this data set is the ca library – notice there is a line for each respondent!

A B C D sex age edu

1 2 3 4 3 2 2 3

2 3 4 2 3 1 3 4

3 2 3 2 4 2 3 2

4 2 2 2 2 1 2 3

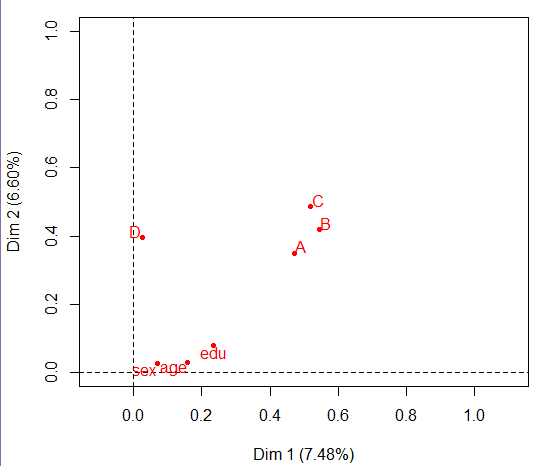
5 3 3 3 3 1 5 2

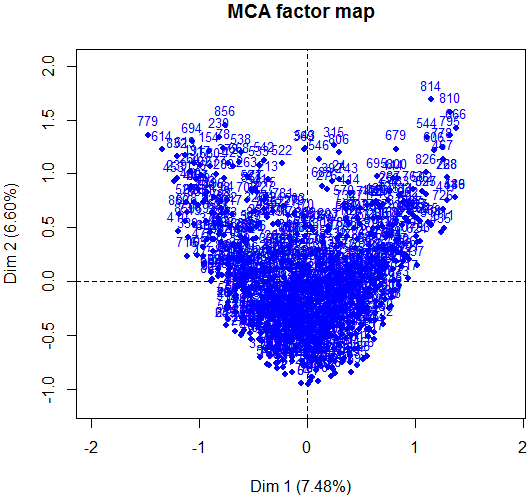
6 3 4 4 5 1 3 2

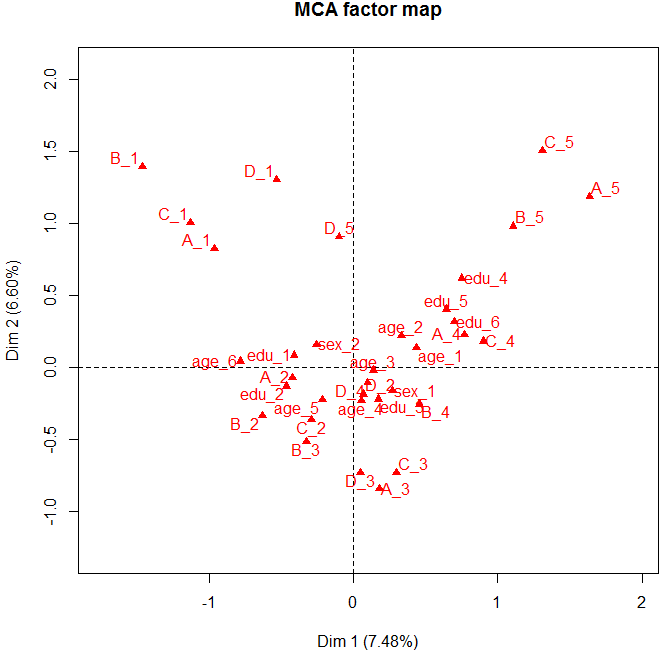
> ?wg93 🡨 get information about these data  
> dim(wg93)

[1] 871 7

> wg93.mca = MCA(wg93)







> summary(wg93)

Eigenvalues

Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6

Variance 0.289 0.255 0.208 0.197 0.183 0.168

% of var. 7.483 6.604 5.400 5.117 4.732 4.350

Cumulative % of var. 7.483 14.087 19.488 24.605 29.337 33.687

Dim.7 Dim.8 Dim.9 Dim.10 Dim.11 Dim.12

Variance 0.165 0.160 0.157 0.151 0.147 0.144

% of var. 4.281 4.139 4.078 3.912 3.824 3.730

Cumulative % of var. 37.968 42.107 46.184 50.097 53.921 57.651

Dim.13 Dim.14 Dim.15 Dim.16 Dim.17 Dim.18

Variance 0.139 0.133 0.129 0.127 0.127 0.122

% of var. 3.599 3.458 3.357 3.297 3.280 3.163

Cumulative % of var. 61.250 64.708 68.065 71.362 74.643 77.806

Dim.19 Dim.20 Dim.21 Dim.22 Dim.23 Dim.24

Variance 0.117 0.112 0.107 0.098 0.094 0.092

% of var. 3.023 2.893 2.764 2.530 2.442 2.398

Cumulative % of var. 80.829 83.722 86.486 89.016 91.458 93.856

Dim.25 Dim.26 Dim.27

Variance 0.085 0.082 0.069

% of var. 2.210 2.138 1.796

Cumulative % of var. 96.066 98.204 100.000

Individuals (the 10 first)

Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2

1 | 0.115 0.005 0.005 | -0.278 0.035 0.028 | 0.377 0.078 0.051 |

2 | 0.410 0.067 0.049 | -0.499 0.112 0.072 | 0.363 0.073 0.038 |

3 | -0.409 0.067 0.072 | -0.298 0.040 0.038 | -0.112 0.007 0.005 |

4 | -0.138 0.008 0.008 | -0.318 0.045 0.042 | -0.233 0.030 0.022 |

5 | -0.059 0.001 0.001 | -0.945 0.403 0.289 | -0.243 0.033 0.019 |

6 | 0.364 0.053 0.043 | -0.093 0.004 0.003 | -0.032 0.001 0.000 |

7 | -0.133 0.007 0.007 | -0.501 0.113 0.096 | -0.039 0.001 0.001 |

8 | 0.578 0.133 0.112 | -0.428 0.082 0.061 | -0.105 0.006 0.004 |

9 | -0.358 0.051 0.030 | -0.158 0.011 0.006 | -0.537 0.159 0.069 |

10 | -0.117 0.005 0.005 | -0.630 0.179 0.155 | -0.572 0.181 0.128 |

Categories (the 10 first)

Dim.1 ctr cos2 v.test Dim.2 ctr cos2 v.test

A\_1 | -0.967 6.325 0.148 -11.348 | 0.823 5.186 0.107 9.654 |

A\_2 | -0.426 3.321 0.106 -9.624 | -0.071 0.105 0.003 -1.611 |

A\_3 | 0.180 0.377 0.010 2.943 | -0.844 9.362 0.218 -13.772 |

A\_4 | 0.769 5.976 0.152 11.491 | 0.228 0.594 0.013 3.403 |

A\_5 | 1.639 7.324 0.157 11.672 | 1.183 4.323 0.082 8.424 |

B\_1 | -1.466 8.671 0.191 -12.882 | 1.390 8.837 0.172 12.217 |

B\_2 | -0.635 3.987 0.101 -9.358 | -0.335 1.256 0.028 -4.934 |

B\_3 | -0.330 1.268 0.034 -5.399 | -0.515 3.505 0.082 -8.432 |

B\_4 | 0.454 3.289 0.098 9.238 | -0.256 1.186 0.031 -5.211 |

B\_5 | 1.105 9.712 0.234 14.262 | 0.979 8.645 0.184 12.641 |

Dim.3 ctr cos2 v.test

A\_1 0.209 0.411 0.007 2.457 |

A\_2 -0.082 0.170 0.004 -1.849 |

A\_3 -0.069 0.076 0.001 -1.120 |

A\_4 0.292 1.193 0.022 4.361 |

A\_5 -0.760 2.183 0.034 -5.414 |

B\_1 0.251 0.353 0.006 2.208 |

B\_2 -0.067 0.062 0.001 -0.991 |

B\_3 -0.064 0.067 0.001 -1.055 |

B\_4 0.366 2.970 0.064 7.457 |

B\_5 -0.685 5.169 0.090 -8.839 |

Categorical variables (eta2)

Dim.1 Dim.2 Dim.3

A | 0.471 0.349 0.059 |

B | 0.544 0.418 0.126 |

C | 0.519 0.486 0.203 |

D | 0.026 0.395 0.160 |

sex | 0.069 0.026 0.055 |

age | 0.158 0.030 0.391 |

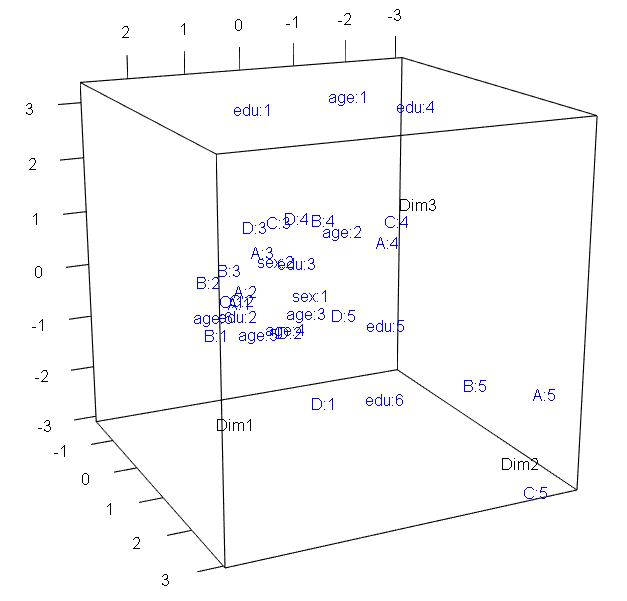
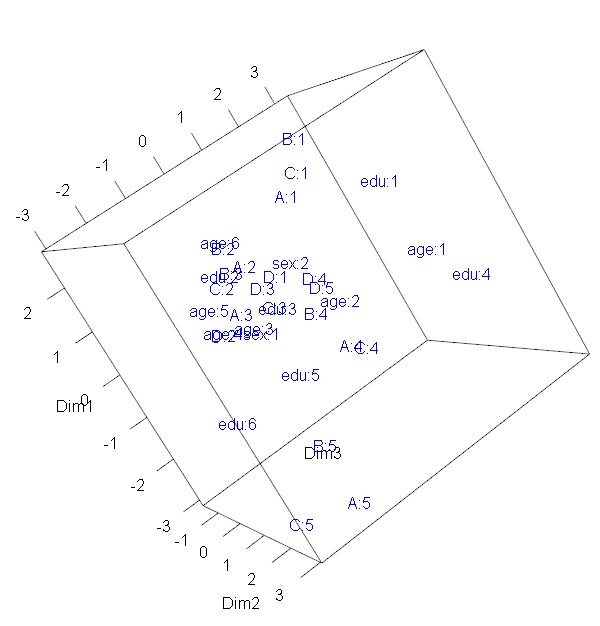
edu | 0.233 0.080 0.464 |

3-D Multiple Correspondence Analysis

> library(rgl)

> plot3d(wg93.mca$var$coord[,1:3],type="n",xlab="Dim1",ylab="Dim2",zlab="Dim3")

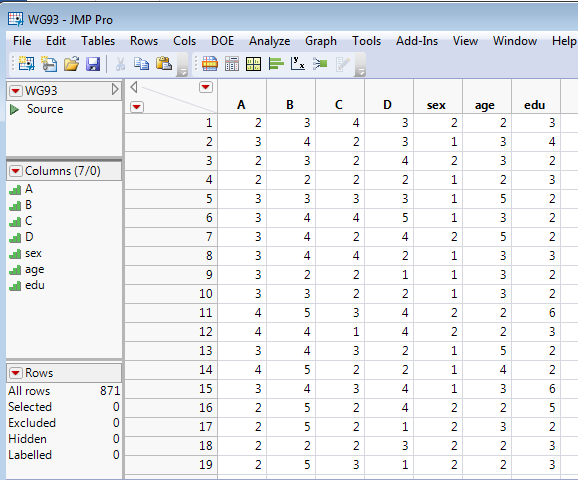
> text3d(wg93.mca$var$coord[,1:3],texts=row.names(wg93.mca$var$coord),col="blue")

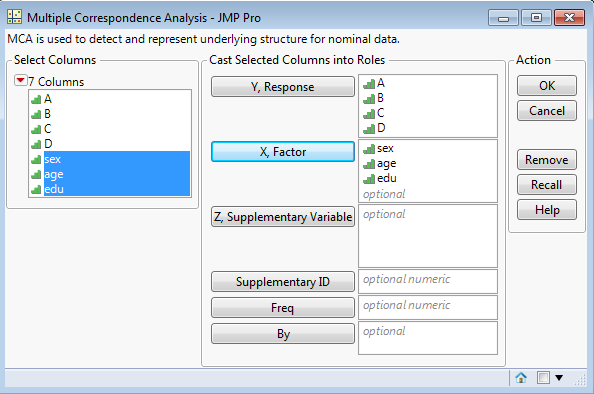
**6.6 – Multiple Correspondence Analysis in JMP**

Example 6.4: International Attitudes about Science (cont’d)

The results of this survey are stored in the JMP Data Table shown below. Remember that each row in these data corresponds to a single respondent, thus there are people surveyed.

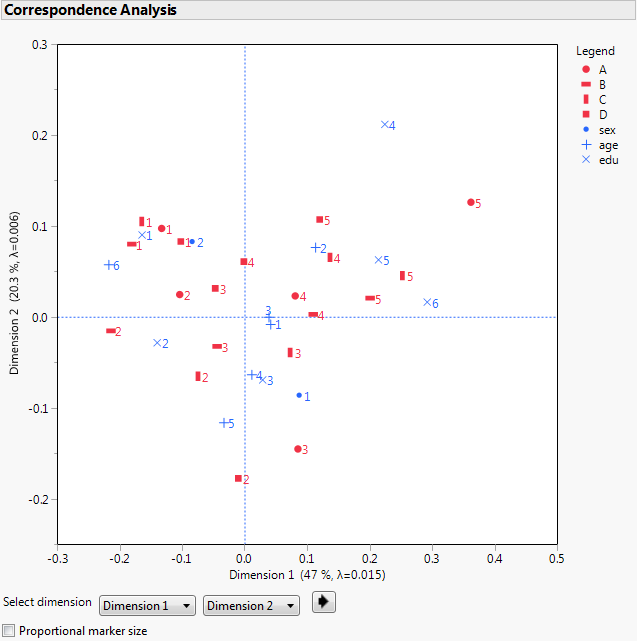
  
ISSP (1993) - International Social Survey Program: Environment, http://www.issp.org

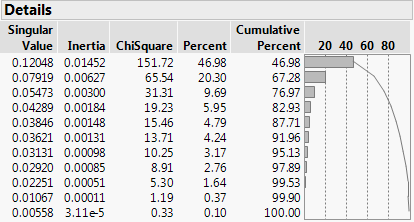
To conduct a multiple correspondence analysis in JMP select Analyze > Consumer Research > Multiple Correspondence Analysis to open the dialog box shown below.

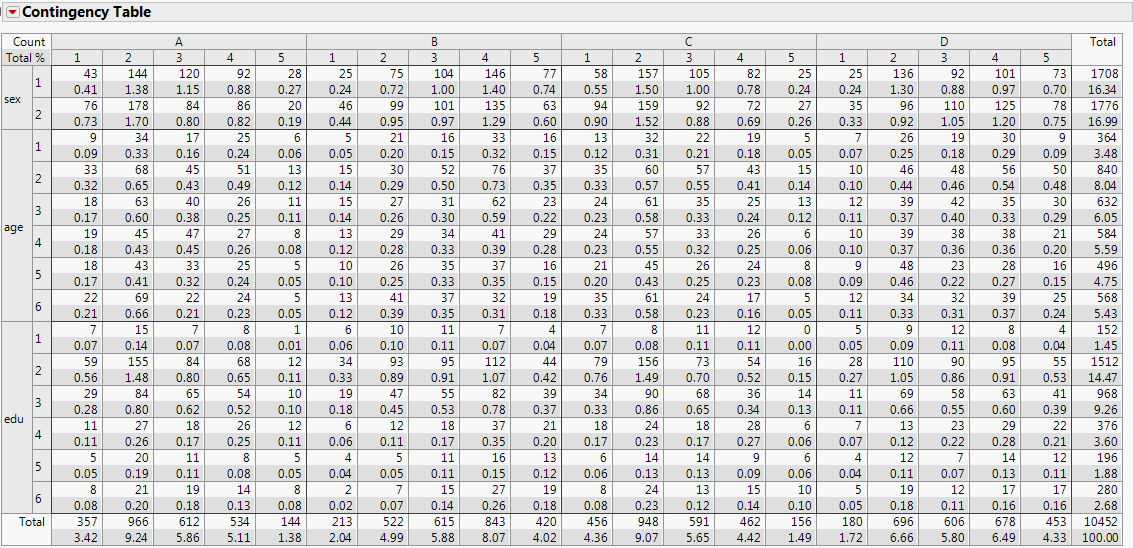
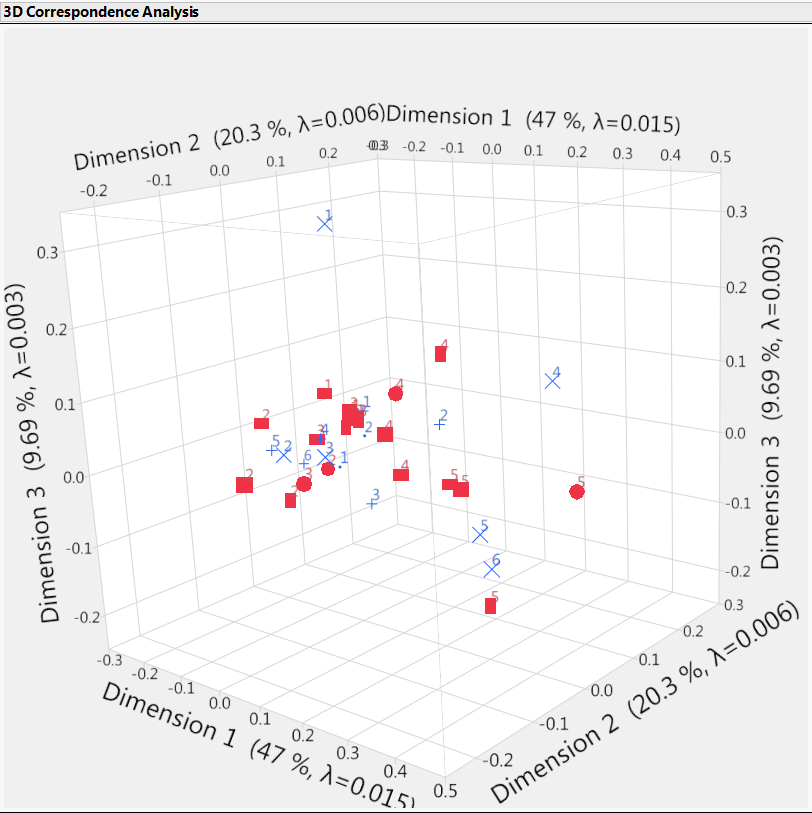


Demographic factors that may help explain attitudinal variation go here.

Survey responses reflecting science attitudes.

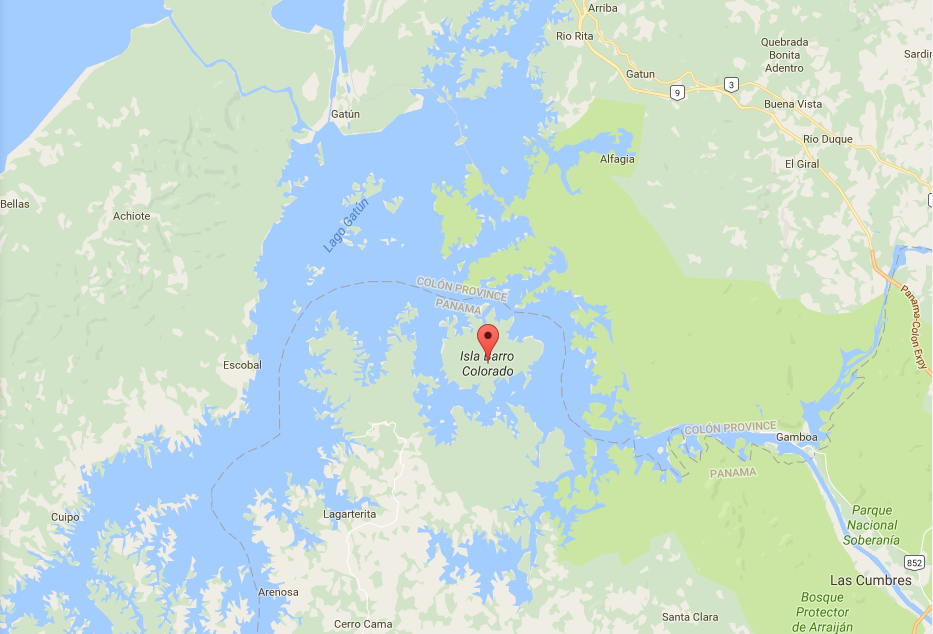




Example 6.5: Barro Colorado Island Tree Counts   
(**Datafiles: Barro Island Species.JMP, Barro Island Environment.JMP, BCI with Environmentals.JMP**)

These data come from a study where tree counts on 1-hectare plots in the Barro Colorado Island and associated site level information were collected. Barro Colorado Island is located in the man-made Gatun Lake in the middle of the Panama Canal. The island was formed when the waters of the Chagres River were dammed to form the lake in 1913. A collected data from 50 plots (rows) of 1 hectare with counts of trees on each plot with total of 225 species (columns) are contained the file **Barro Island Species.JMP**. Full Latin names are used for all tree species. The names were updated against <http://www.theplantlist.org> and Kress et al. (2009) which allows matching 207 of species against <http://datadryad.org/resource/doi:10.5061/dryad.63q27> (Zanne et al., 2014).

  
There lots of potential analyses we could conduct using these data, however let’s first consider questions that can be answered using the methods discussed above. Some potential research questions might be:

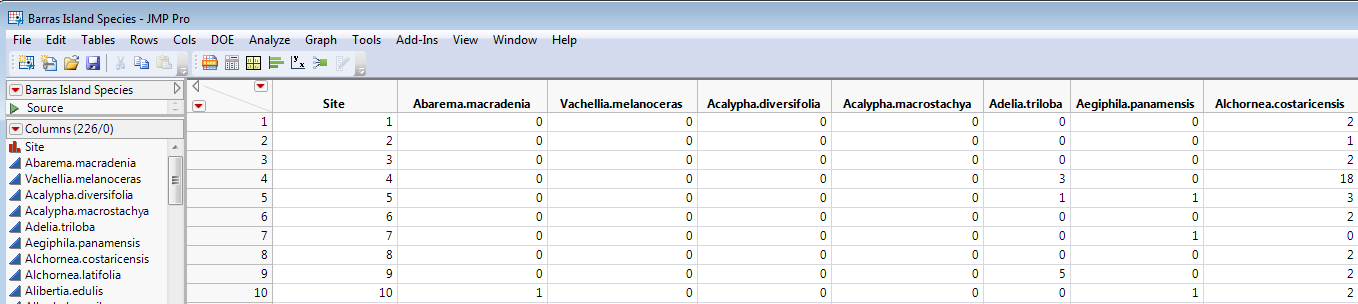
* Are there site differences in the distribution of tree species found?
* If there are differences, what are the nature of the between the sites and the species found?
* What are the most prevalent species found? Least?
* How do the sites differ in terms of number of trees found?

Other questions that we might have that will require additional statistical methods are:

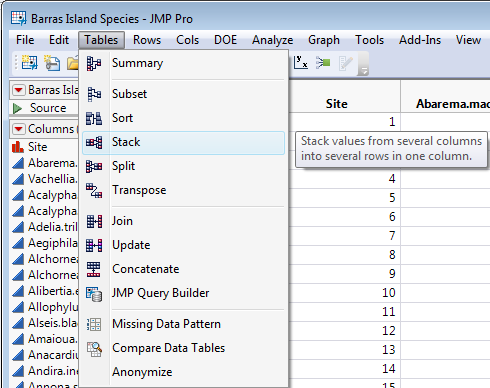
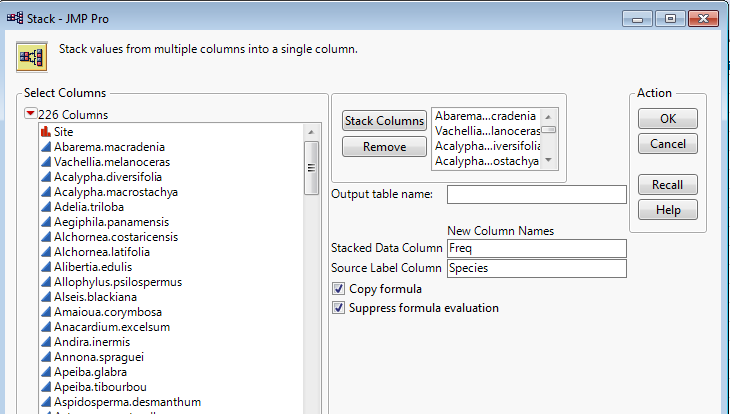
* How diverse are the sites in terms of species found? How similar?
* What sites are most similar in terms of their species distribution? Are there groups of sites with similar species patterns?
* Is there evidence of dominance of certain species at the site level?
* How are the species found related to the environmental measures at each site (note: these measures are contained in the file **BCI with Environmentals.JMP)**

We will begin by examining the first set of questions and discussing the data format.

The file **Barro Island Species.JMP** looks like this.

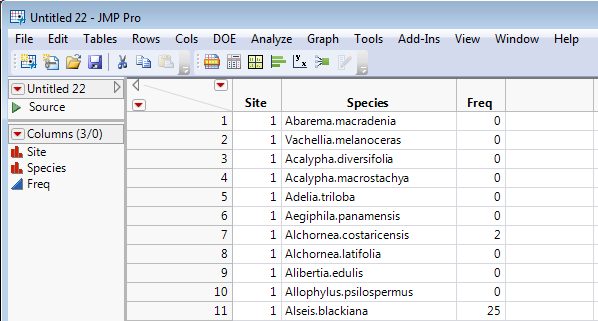


Each row corresponds to one of the 50 1-hectare sites examined and each of the 225 columns corresponds to the species found in the study. The numeric entries in each row are the number of trees of each of the 225 species found at the corresponding site. While this might make data entry easy it is not optimal for answering the questions posed above in JMP. We will **Tables > Stack** the 225 species columns so we have two nominal variables for site and species along with the associated frequencies in a third column.

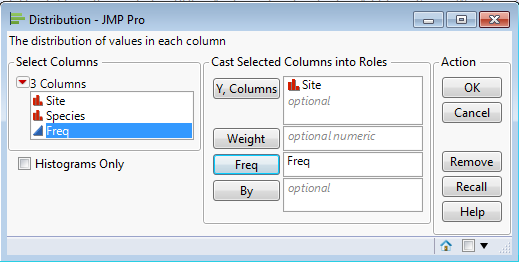
 

Select Tables > Stack and put all 225 species columns the Stack Columns section. Then we can change the name of the stacked columns as shown above. The Source Label Column is determined by the **Species** columns and the Stacked Data Column will contain the frequencies of each species at each site thus will label it **Freq**.

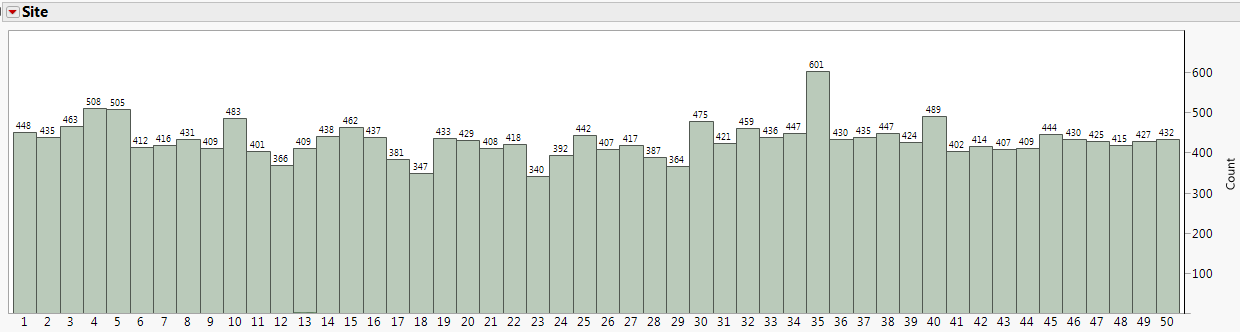
The resulting untitled table contains these data in layout similar to the examples above (i.e. Hodgkin’s and Blister Rust examples). **(BCI Site Species (stacked).JMP)**

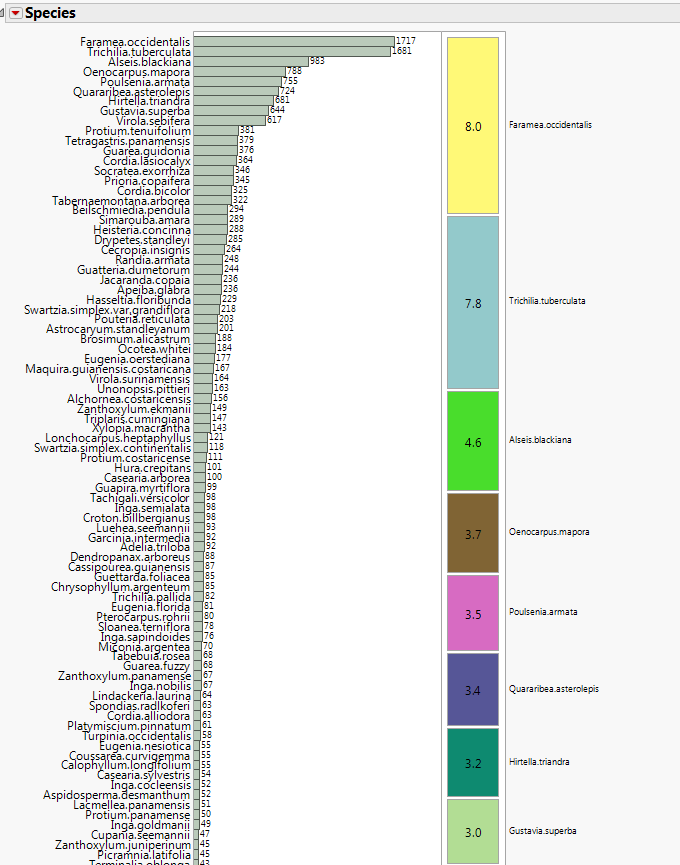


We can begin by examining the univariate distributions of Site and Species by using **Analyze > Distribution** and adding graphical enhancements to the resulting bar graph.



**Number of trees across all species found at each site**



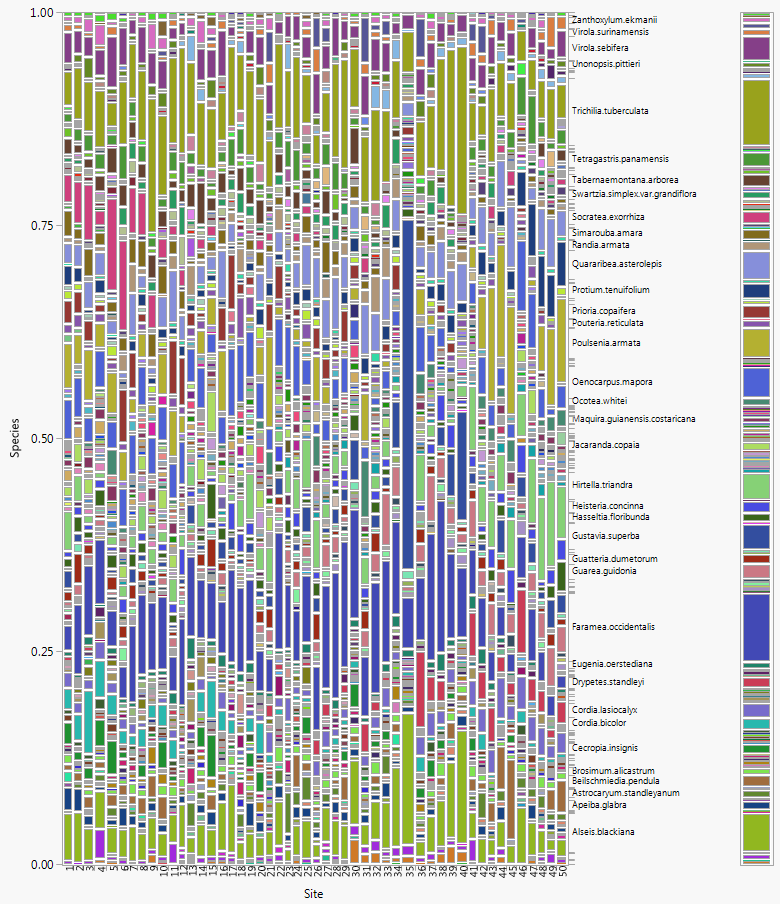
**Distribution of tree species found across all sites**

Certainly some species are far more prevalent than others. In order for tree species to be added it was only necessary to find one tree at one of the 50 sampled sites. There were 19 species with a count of 1, 13 species with a count of 2, and 54 species with a count of 5 or less.

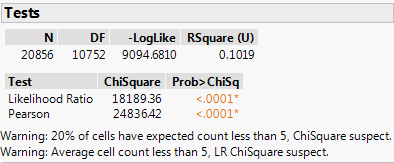
The most prevalent species were Faramea Occidentalis with a count of 1717, Trichilia Tuberculata (1681), and Alseis Blackiana (983) – shown from left to right below.  
  


We can attempt to look at species differences across site by using a chi-square test of homogeneity and a 2-D mosaic plot.

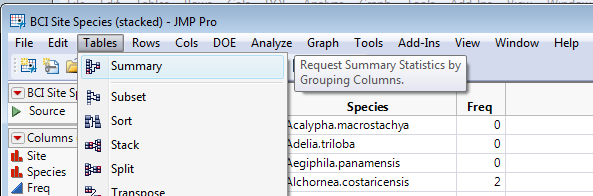
**Mosaic plot showing species breakdown within each site**

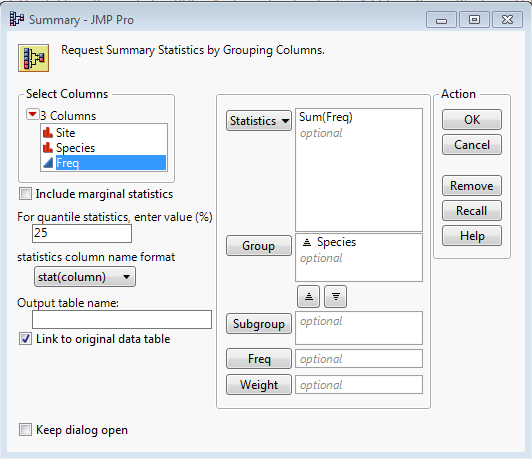


We really don’t need a chi-square test to confirm that the distribution of species found at each of the 50 sites probably differ, but we might want to look at the results anyway.

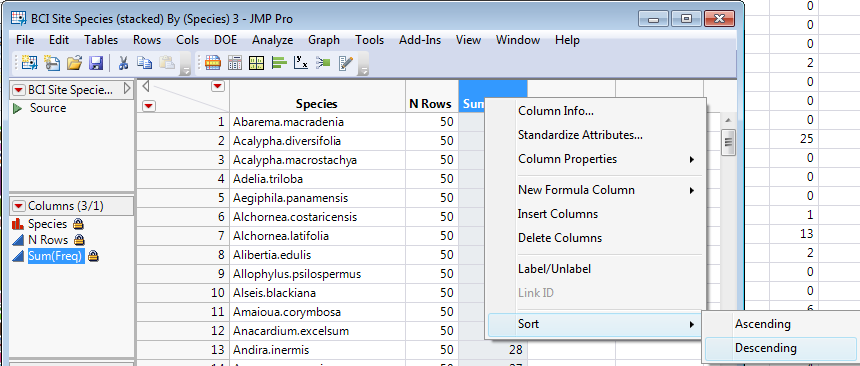
  
Notice we are getting two warnings which caution us against using the chi-square test with these data. This is because many of the cell counts in our contingency table are very small, with many being 0! One way we can deal with this problem is to focus on the most prevalent species found across the sites. There are several ways to do this in JMP but the easiest I think is to create a table of species counts, sort the counts, and select a species with a minimum overall frequency above some cutoff.

**1** – Select Tables > Summary

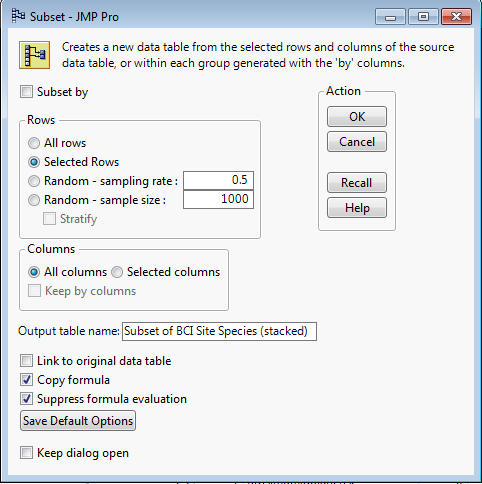


**2** – Put Species in the Group box, highlight Freq and Sum from Statistics pull-down menu. When finished it should look like this – click OK.  


**3** – Right-click on Sum(Freq) column and Sort > Descending. Once sorted highlight all the rows above whatever frequency we want as a cutoff. Using a cutoff frequency of 200 gives 30 species per site.

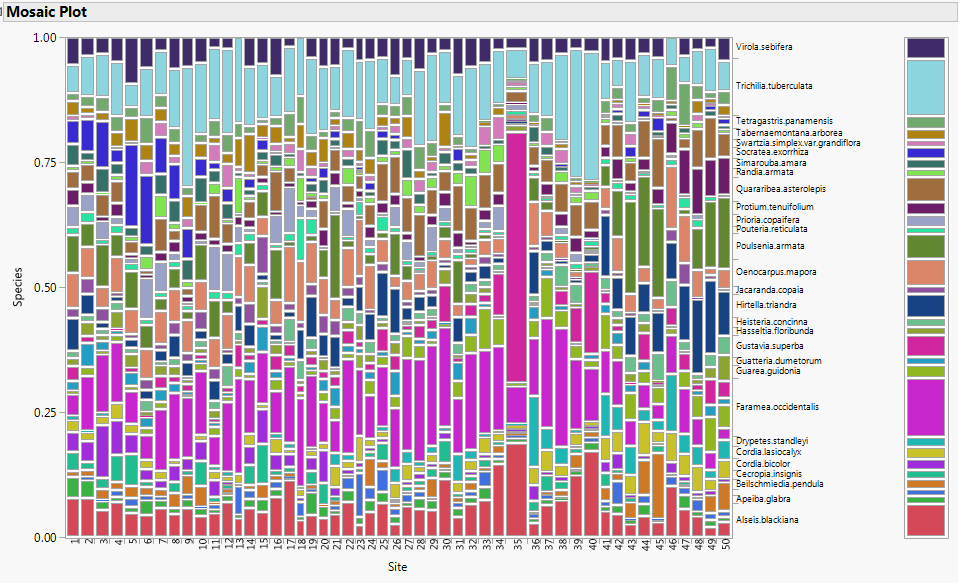


**4** – After highlighting the most frequent species, you carefully go back to the original stacked data set and select **Tables > Subset**.

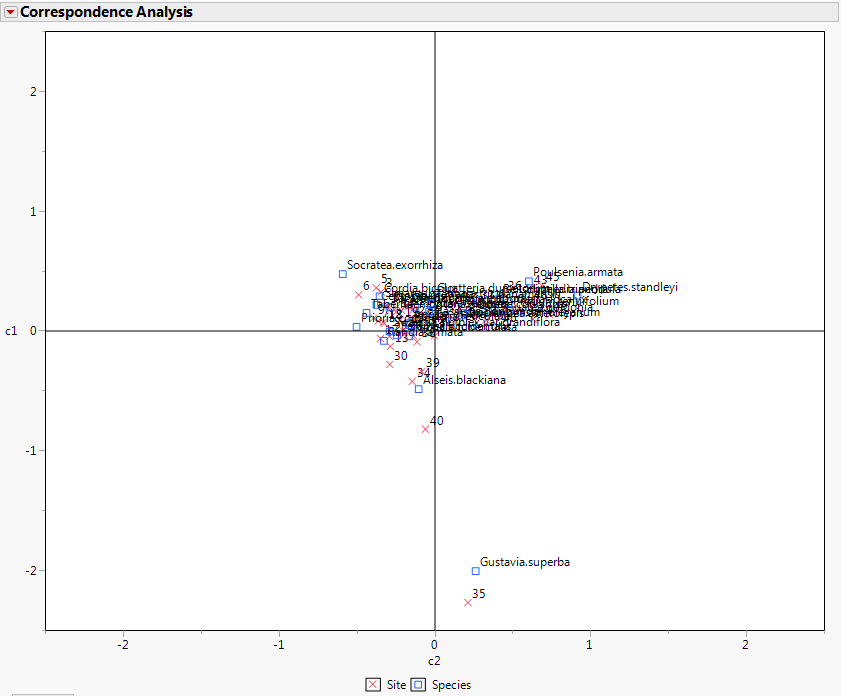


**5** – The subset data set will contain a by site breakdown of the 30 most prevalent species in this study. We can proceed to look at the differences across site again by using contingency table analyses.

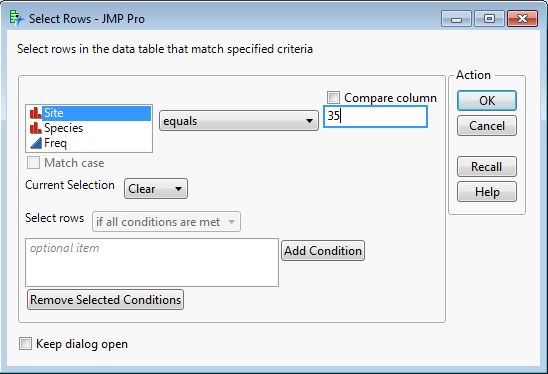
After extracting the data subset here are the 2-D mosaic plot and chi-square test.



Site #35 certainly stands out for two reasons – one it is dominated by the species Gustavia Superba and it also has the highest number of trees found across the 30 species we are now focusing on. The correspondence analysis for these data reflects the strong association between site #35 and Gustavia Superba presence.

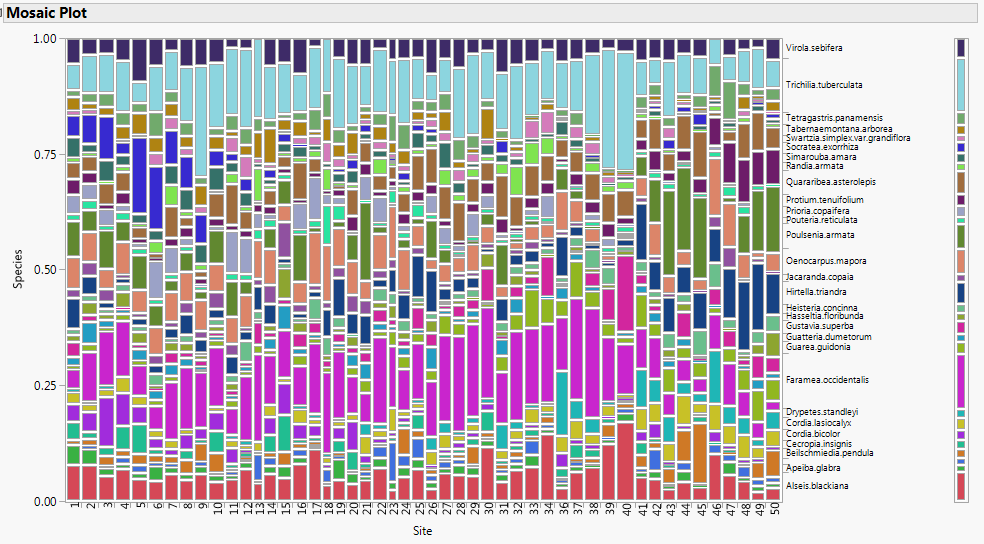


To better see less obvious relationships in these data we will temporarily hide and exclude site #35.

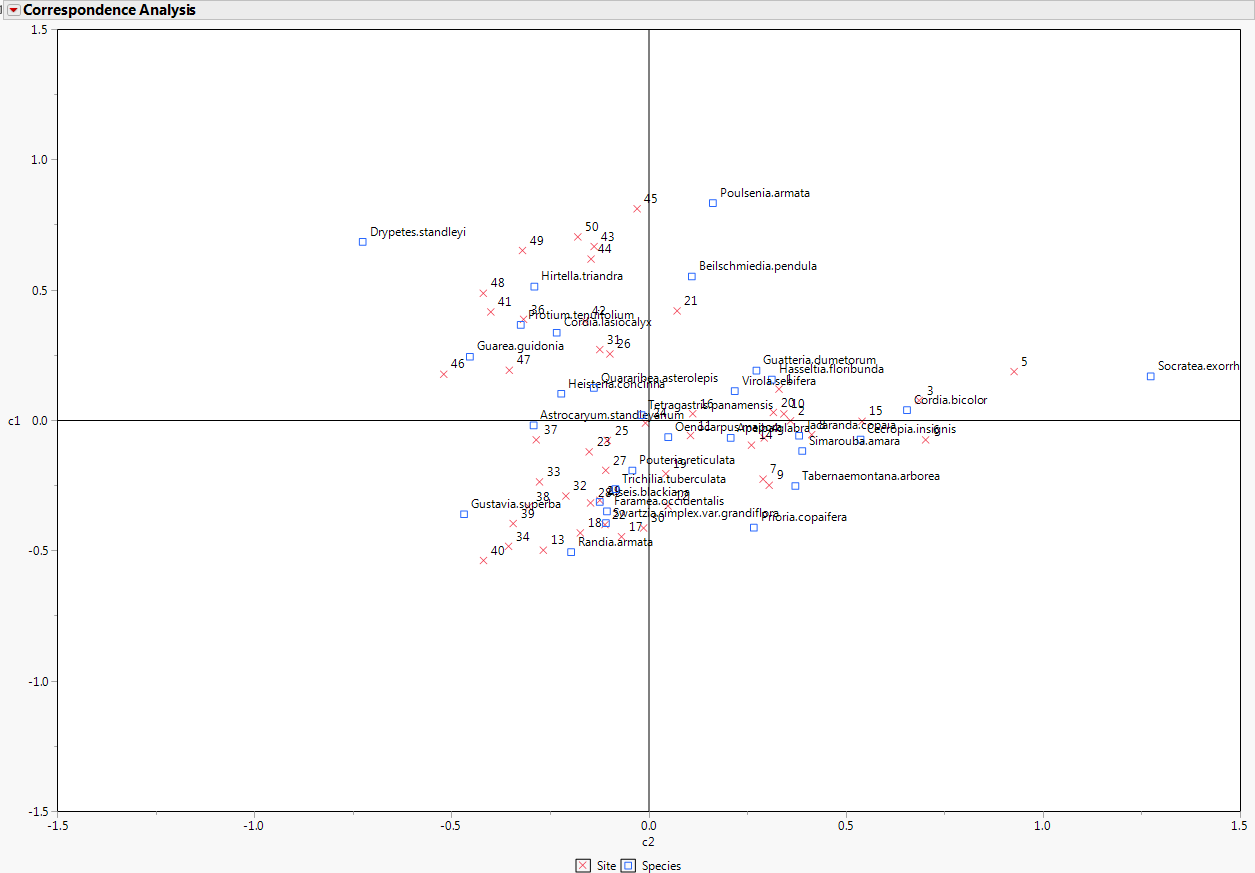
 

It should show back in the spreadsheet that 30 rows have been selected corresponding to the 30 species counts for site #35. Now select **Rows > Hide and Exclude**.

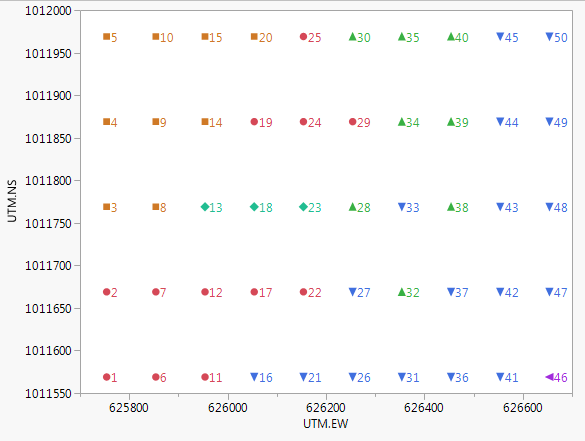
Mosaic plot with site #35 excluded



The correspondence analysis results allow us to see much more structure in these data.



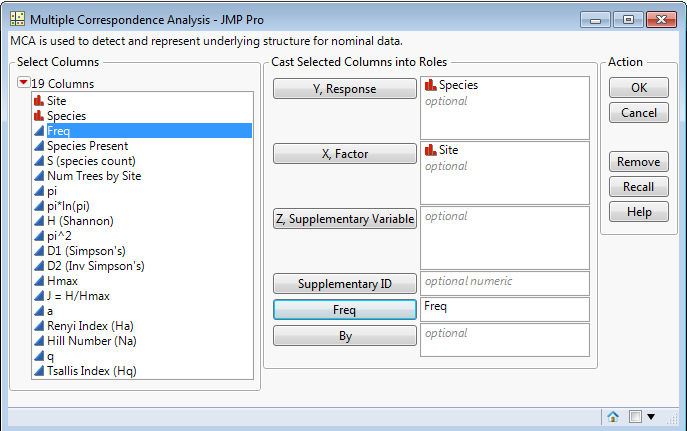
To restore site #35 observation we simply select **Rows > Clear Row States** to take away the exclusion and hiding of these data. From this plot you should be able to see which sites have similar species distributions, which species are most associated with certain sites, which species tend to found together.



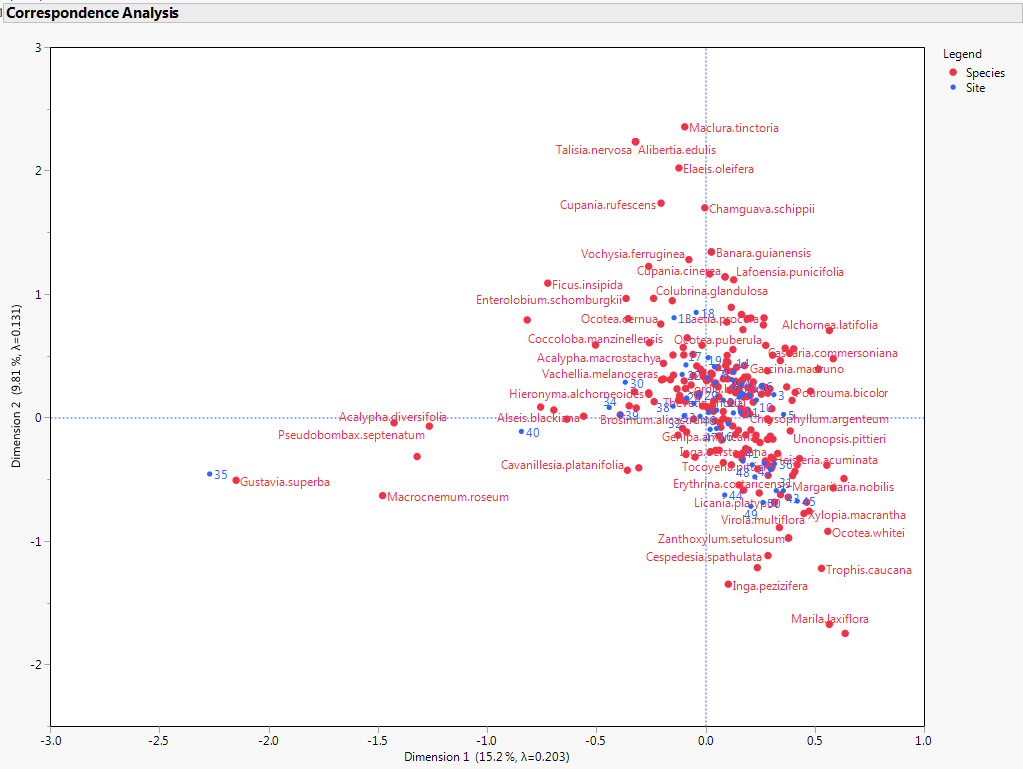
This plot shows how the 50 sites are distributed spatially. The color-coding and marker type were determined by ***clustering*** sites with similar species profiles.

These data are very rich, there must be more things to things we can learn from these data!

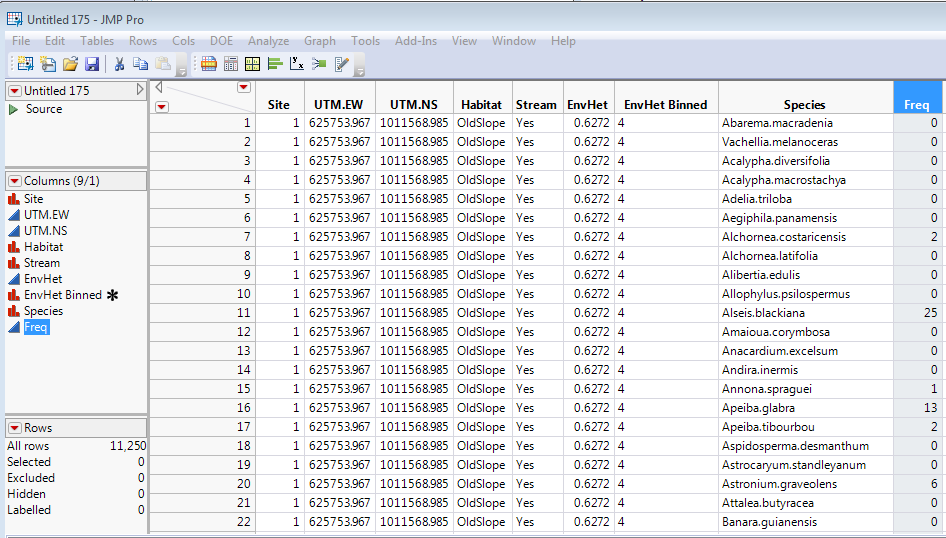
We now consider multiple correspondence analysis for these data using the **Consumer Research > Multiple Correspondence Analysis** approach in JMP. Again working with our stacked version of these data where we have columns for Site, Species, and Freq we can set up our dialog box as follows.



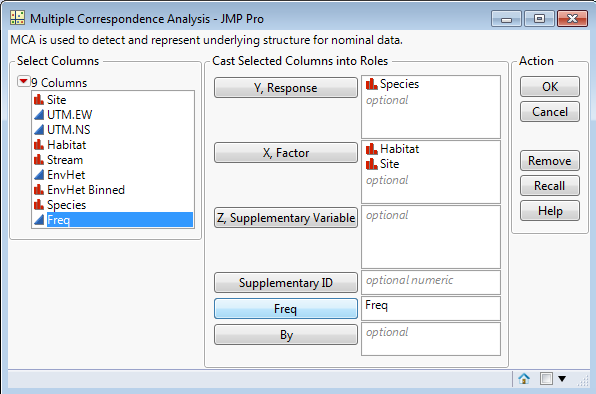
The resulting plot is shown below.



We can add environmental information to this analysis as well by working with the file **BCI with environmentals.JMP**. We again stack the species columns into one column for Species along with a column for the Frequencies. All other columns are automatically extended to maintain site level information. Note some of the environmental values are the same for all 50 sites so I deleted those (e.g. Age.cat and Geology), so I have deleted them.

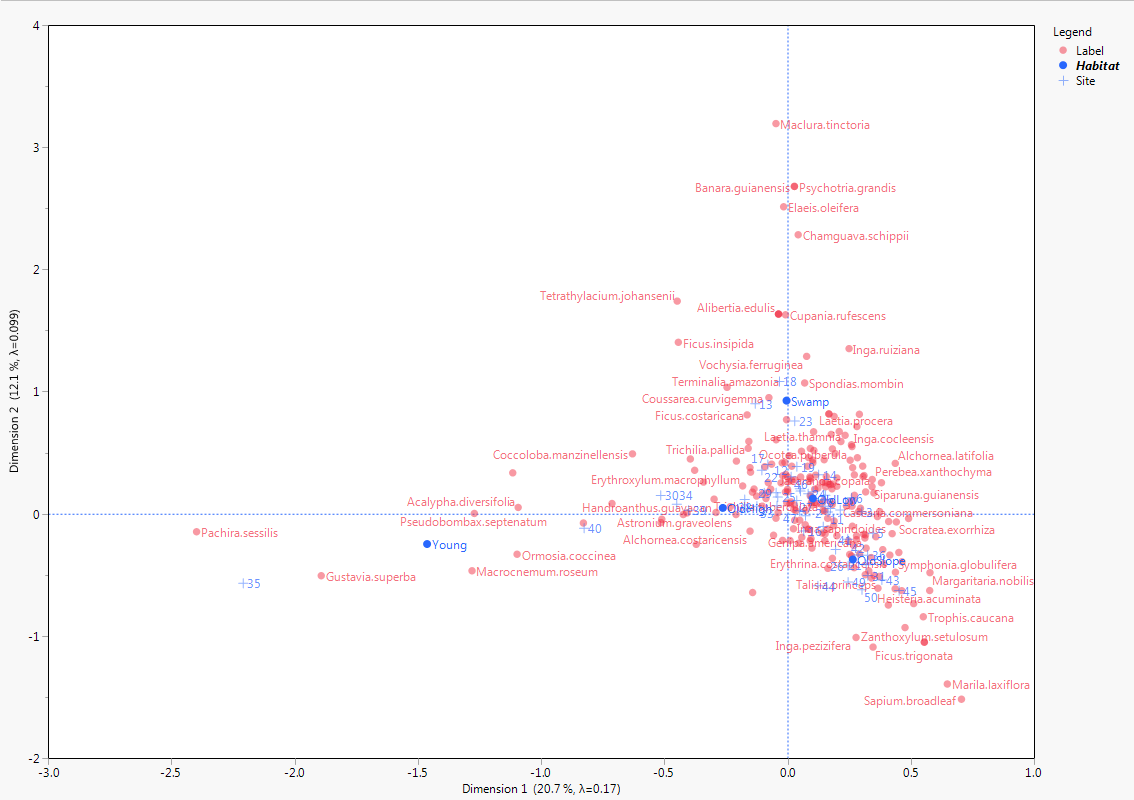


We can add Habitat to our correspondence analysis by adding it to the X, Factor box along with Site in the Multiple Correspondence Analysis window.



The resulting plot is shown on the following page.

Multiple Correspondence Species vs. Site & Habitat Type



I explored adding other factors such as whether a stream flows through the site or not, but it did not produce much delineation between their levels in the plot.

Multiple correspondence analysis (MCA) allows the representation higher dimensional data where the variables are non-numeric (i.e. categorical/nominal or ordinal) in a lower dimensional space. As with PCA, MCA attempts explain as much of the total variation in the data in this lower dimensional representation. We can use the resulting lower dimensional displays to explain the variation between individuals in our data and to uncover the relationship between the variable. Methods for mixing both numeric (PCA) and non-numeric (MCA) variables are an active area of research. Much of the original work in this area falls under the heading Factor Analysis for Mixed Data (FAMD) good resources for dimension reduction methods are “Multiple Factor Analysis by Example Using R” by Jerome Pages and “Principal Component Methods in R: Practical Guide” by Alboukadel Kassambara.

A great resource for all of the FactoMineR and factoextra options for visualizing the results of an MCA conducted using FactoMineR is:

<http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/>

Additional great video resources for PCA and MCA in FactoMineR can be found on YouTube by searching “Francois Husson”. I highly recommend watching any of his videos recorded that are English unless you also speak French.