Name: Nghi Vo
Student number: 1029365

Multvariate Analysis of subject of study, Stress level and Stressors in postgraduate students

1. Introduction

1.1 Motivation & Goal

Stress has been a constant factor diminishing life quality for students. Stress can cause tiredness, irritation, or inefficiency in mild cases, but also disorders and depression in severe cases. Therefore, studying stress levels and stressors can be considerably beneficial. This is the main motivation for this project.

The project is aimed at studying the correlations between the academic discipline of a student, the stress level, and stressors. The data used for this study is collected as data for a research conducted by University of Bristol and Pukka Herbs, available at: https://figshare.com/articles/dataset/Student_Stress_Survey_Jan2020_OPENDATA_xlsx/11559528. The data is the raw data from a survey evaluating stress in postgraduate university students in the United Kingdom. Students from a wide range of disciplines, included home and international, were asked about how stressed they feel, academic causes and other causes of stress, and how they cope. In this study, the academic disciplines, stress levels and three stressors are studied in details. These stressors are: financial issues, lack of confidence in academic performance, and lack of confidence in subject or career choice.

1.2 Research Questions

The main research question is how academic principles, stress levels and different stressors correlate to one another. This main question could be broken down to more specific questions, including:

- Does academic principle affect the level of stress of a student?
- How do different academic principles correlate with specific stressors?
- To what extent does each stressor contribute to overall stress level?
- Is there some correlation between one stressor and another?

1.3 Dataset

The original dataset includes several data about the respondents' demographic, level of study and subject of study, as well as several answers related to stressors and coping mechanisms. The necessary information is extracted to form the used dataset, which includes 218 observations (218 students) of 5 variables as following:

Variables	Descriptions
subject	Academic principle (major/subject) of the student
stress_level	Stress level in the past 3 months
financial_issues	Frequency in past 3 months of stressor: financial issues
academic_performance	Frequency in past 3 months of stressor: lack of confidence in academic performance
subject_career_choice	Frequency in past 3 months of stressor: lack of confidence in subject or career choice

The variable '**subject'** stores categorical data of the subject that the student studies. In this survey, there are 11 categories: Arts and Humanities; Clinical, pre-clinical and health; Computer Science; Education; Engineering and Technology; Law; Life Sciences; Physical Science; Psychology; Social Sciences; and Other.
Other variables '**stress_level**', '**financial_issues**', '**academic_performance**', and '**subject_career_choice**' store ordinal data of the answers to the survey describing the levels of stress or the frequencies of stressor in the past 3 months from the survey. Answers to the level of stress include: "Not at all", "To a small extent", "Somewhat", "To a large extent", and "Completely". Answers to the frequencies of stressors include: "Never", "Almost never", "Sometimes", "Fairly often", and "Very often".

2. Univariate analysis

The distribution of the modalities for each variable can be visualized as:

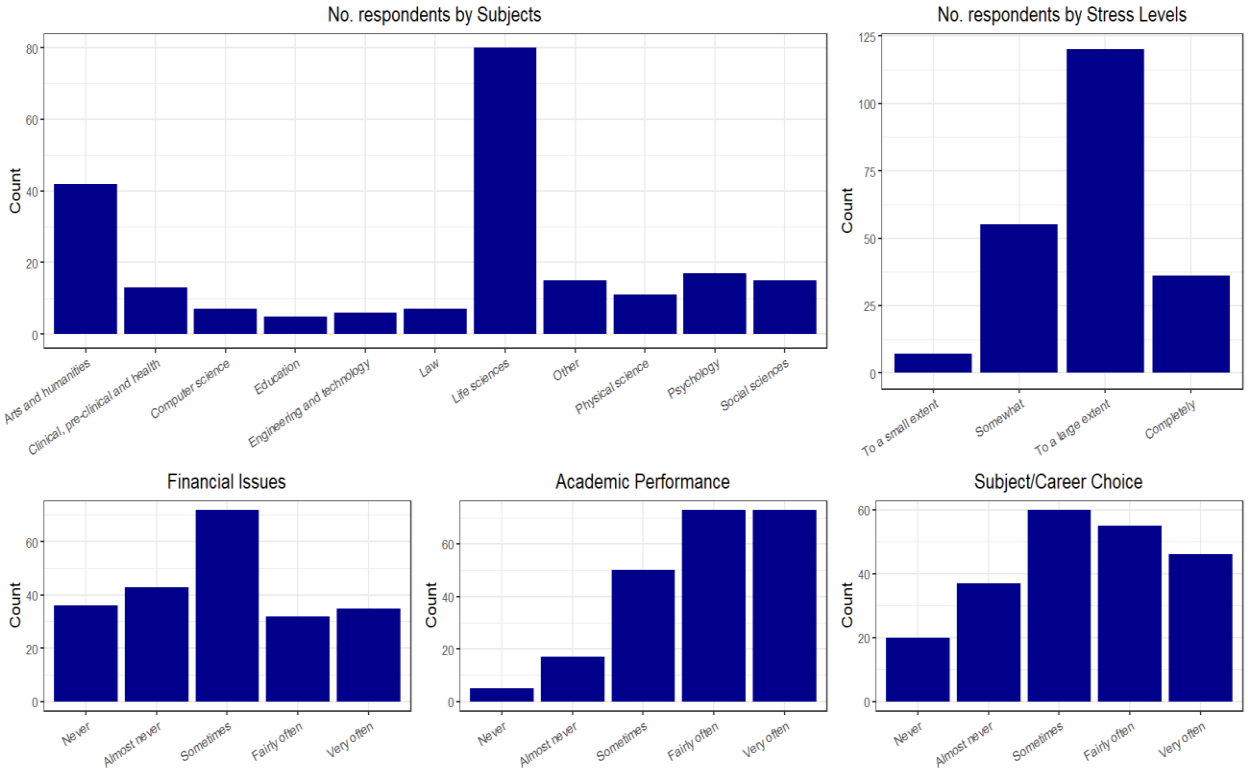


Fig. 1 : Distribution of modalities

There are several observations to be made from the distribution of datapoints for each variable:

- Despite the relatively sufficient number of datapoints (218 respondents), the distribution of respondents by subject is extremely uneven. There are subjects with significantly higher number of respondents (Life Sciences; Arts and Humanities), while there are subjects with fewer than 10 respondents (Computer Science; Education; Engineering and Technology; Law). In subjects with such a small number of respondents, the analysis could be severely affected by outliers. Furthermore, this uneven distribution could also reflect selection bias due to compliance varying in different academic principles, which may affect the validity of the results. This would be discussed further in critical evaluation.
- In the distribution of answers to stress levels in past 3 months, the reluctance of respondents to choose extreme answers is evident. While having the option "Not at all", no respondents chose that answer, and there is a significant drop in the number of respondents choosing "Completely" despite the considerable number of respondents choosing "To a large extent". Furthermore, it is visible that the distribution is skewed to the right, with significantly more respondents choosing higher level of stress rather than lower.
- The distributions of reported frequencies of stressors Financial Issues and Subject/Career Choice are relatively similar and resemble the normal distribution, with the medium frequency ("Sometimes") being the mean. However, the distribution for answers to Subject/Career Choice is slightly skewed to the right (higher frequencies are reported more often than lower frequencies).
- The distribution of reported frequencies of stressor Academic Performance is severely skewed to the right, resembling the distribution of answers to stress levels. This can suggest strong correlation between the two variables.

3. Bivariate Analysis

The bivariate analysis is conducted using the attraction-repulsion matrix between each pair of variables. The attraction-repulsion matrix illustrates the differences between the observed number of and the theoretical number of respondents of each combined category of two variables under the assumption of independence between those variables. If the value indicates such difference is higher than 1.0, the observed number is higher than the theoretical number, signifying 'attraction' between two modalities of two variables. Similarly, if the difference is lower than 1.0, there is 'repulsion' between two modalities, and if there is no difference, the two modalities are independent.

The following graphs demonstrate the attraction-repulsion of modalities in each pair of variables. The colors of the graphs scale with the value of the difference (high values mean attraction, while low values mean repulsion). Generally, "red" denotes attraction, "blue" denotes repulsion, and lighter shades or "white" denote independence or close to independence.

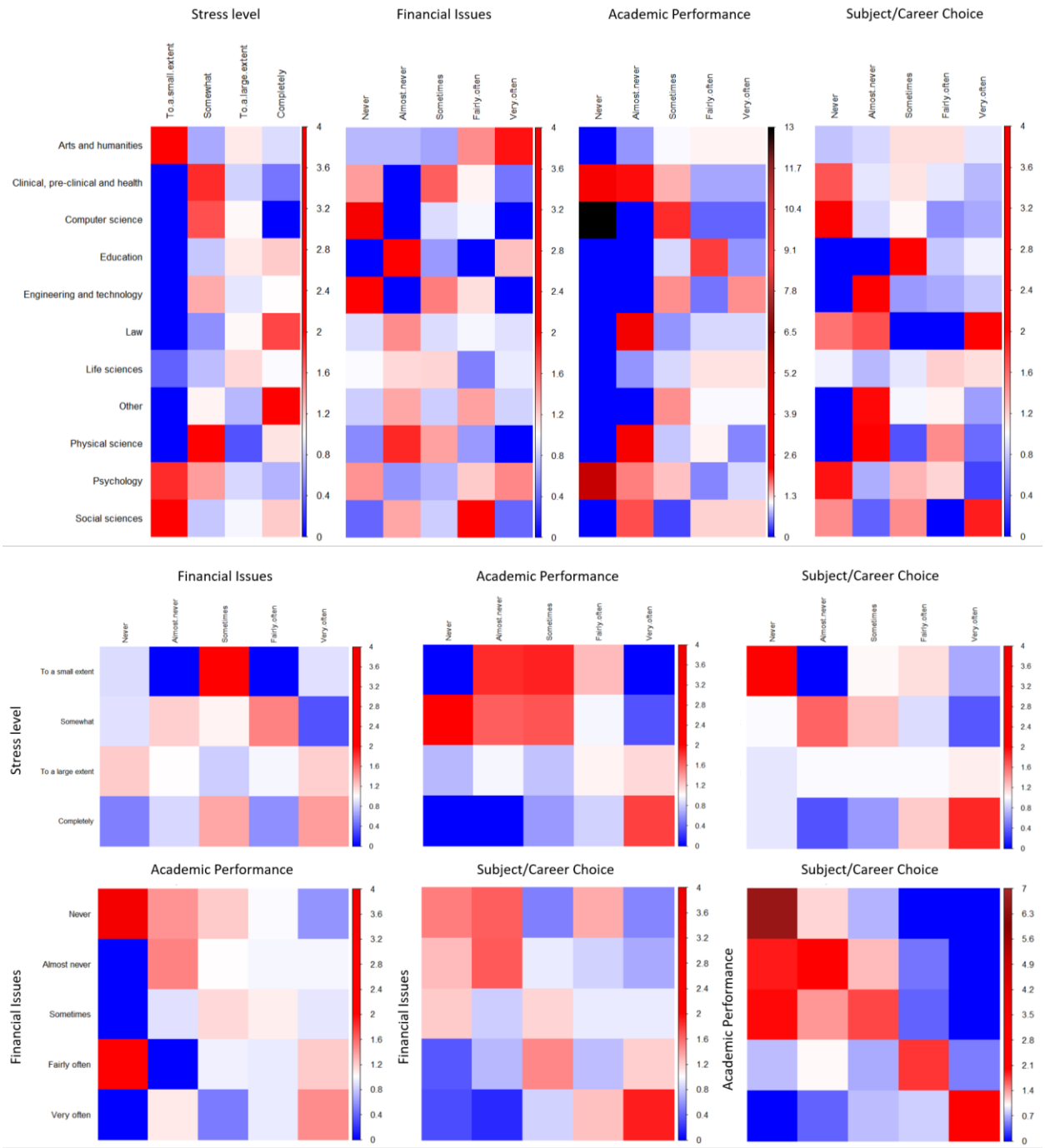


Fig. 2: Bivariate attraction-repulsion matrix

While it is difficult to pinpoint all modalities with strong attraction-repulsion, there are some important observations:

- The subjects of study do seem to be correlated with the stress levels and stressors to certain extents, such as Psychology students tending to report low-medium levels of stress, or Law students tending to report highest level of stress. Even the subjects with the most respondents (of

which results should not be as sensitive to outliers as other subjects), there is still some noticeable correlation: Arts students have a strong tendency to report small extent of stress, and high frequencies of stressor Financial Issues; Life Sciences students is likely to not report small extent of stress, or not reporting low frequencies of stressor Academic Performance.

- Of all other variables, answers to the frequencies of stressor Academic Performance seem to be strongly correlated with subjects of study with many extreme values indicating attraction-repulsion. However, this can be due to the skewed distribution of answers to the frequencies of this stressor as noted in the univariate analysis.
- Of all the stressors, Academic Performance seems to be the most correlated with the level of stress. This is because answers with respect to the stressors tend to have strong attraction or repulsion towards corresponding answers to stress levels, with the exception of answers "Never" to stressor and "To a small extent" to stress level. This also applies to the stressor Subject/Career Choice and the stress level, despite the seeming weaker correlation.
- The stressors Academic Performance Subject/Career Choice seems to be strongly correlated with one. Furthermore, the same frequencies of these stressors are strongly attracted to one another, denoted by the high values of attraction on the diagonal of the matrix.

4. Multivariate Analysis

4.1 Multivariate Correspondence Analysis

The data can be analyzed using Multivariate Correspondence Analysis (MCA). Similar to the simple bivariate correspondence analysis, data in MCA will be used to form two data clouds: the row and column profiles. Each individual will be a point in the data clouds, and the distance between two points represent how "different" they are from one another, i.e. how many or few common modalities that they share. Then, PCA type transformation is applied to the data clouds, while the profiles are transformed to be able to center the variables and use Euclidean distance rather than χ^2 distance. Thus, MCA can present the associations at modality level in a lower dimension without substantial information loss.

Performing MCA on the dataset, the data clouds are reduced to 25 dimensions. The below scree plot shows the percentage of total variance that each dimension explains:

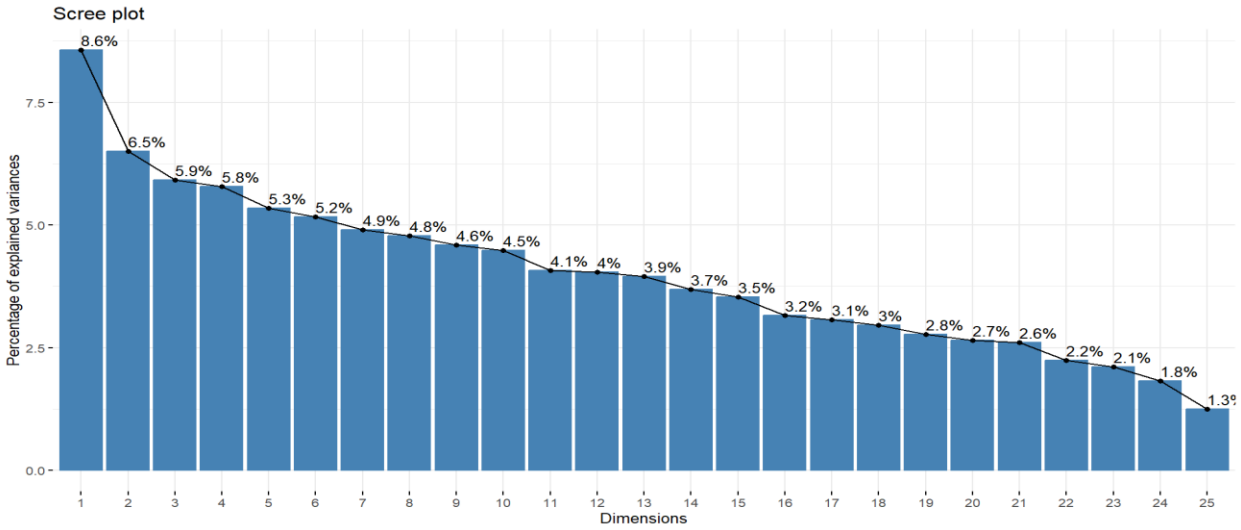


Fig. 3: Scree plot of explained variance of 25 dimensions

The below graphs inspect further into how the modalities are represented in the first and second dimensions, and the third and fourth dimensions. In the graphs, the names of certain variables are abbreviated: “S” – stress level, “FI” – financial issues, “AP” – academic performance, and “SCC” – subject/career choice. Modalities for the level of stress and frequencies of stressors are also presented to their order from least to most (e.g. 1 – “Not at all” of stress level or “Never” of frequencies of stressors).

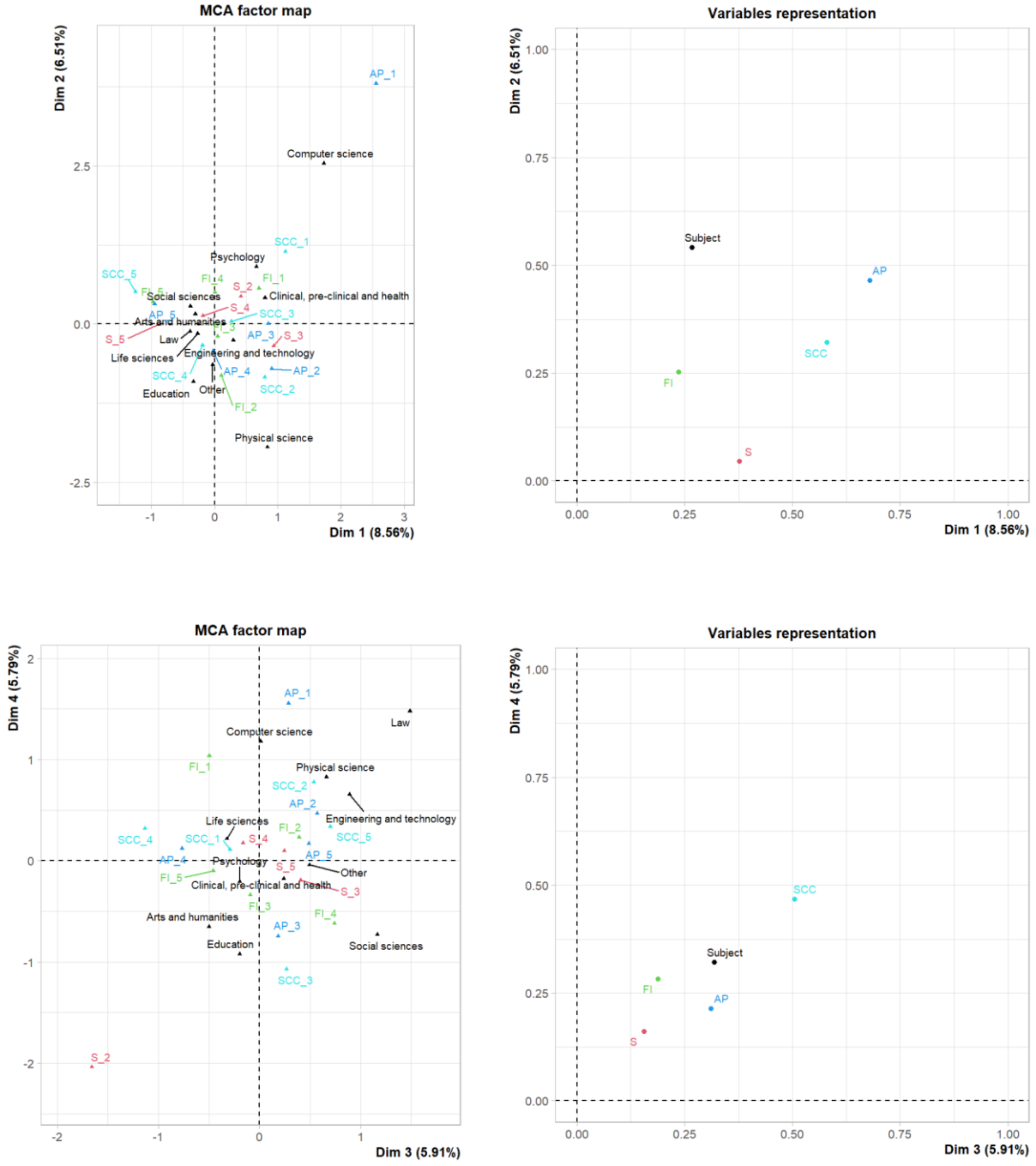


Fig. 4: Biplots of dimensions 1&2 (top-left) and 3&4 (bottom-left) and quality of representation

There are some insights to be gained with the presented graph:

On dimensions 1 and 2:

The variables best represented in dimension 1 and 2 are “AP” (stressor: Academic Performance), “SCC” (stressor: Subject/Career Choice), and “Subject” (subject of study), respectively. Some associations between the modalities of these variables can be inferred from the graph as:

- The modalities “Computer science”, “Psychology”, “AP_1” and “SCC_1” seem to have very strong positive association as the angles connecting those modalities are close to 0, while these modalities are relatively far from the origin compared to other modalities, especially the modalities “AP_1” and “Computer Science”.
- The modalities indicating the same frequencies of stressors between Academic Performance and Subject/Career Choice seem to have strong positive associations as their points are generally very close on the graph.
- The modalities “Arts and humanities” and “Social sciences”, as well as “Law” and “Life sciences” of the variable “Subject” are very close to each other, suggesting some similarities between the individuals within these modalities.

These associations are noteworthy as they support previous analysis or general assumptions. For instance, from bivariate correspondence analysis, there is very strong attraction between the modality “Computer Science” and the modality “Never” of stressors Academic Performance and Subject/Career Choice. This is also reflected when projecting the modalities on the first and second dimensions as reflected in the graphs. Similarly, the strong attraction between the same frequencies of stressors Academic Performance and Subject/Career Choice is visible in the graphs. Additionally, these graphs denote the possible similarities between individuals studying in “Arts and humanities” and “Social sciences”, which are quite related fields of study. On the other hand, new insights could be gained, such as the potential similarities regarding stress levels and stressors between individuals studying “Law” and “Life sciences”.

However, these observed associations can be undermined by the low quality of representation of the variables. Despite choosing the best represented variables, these variables could hardly reach 0.5 quality of representation in both dimensions. Hence, there are high chances that speculations about the associations as observed in the biplot are incomplete due to the low quality of representation of these variables.

On dimensions 3 and 4:

The problem of low quality of representation is even more severe for dimensions 3 and 4, as presented in the graph of variables representation. Despite of the low quality of representation, some associations corresponding to previous analysis could still be speculated:

- The modalities “Computer science” and “AP_1” are still presented to have positive association, but not as strong as shown in dimensions 3 and 4.
- With the exception of the modalities “AP_1” and “SCC_1”, the modalities indicating the same frequencies of stressors between Academic Performance and Subject/Career Choice still exhibit strong positive associations as their points are generally very close on the graph.

- The modalities “Arts and humanities” and “S_2” are shown to have very strong positive association, which is not visible in the biplot of dimensions 1 and 2. This association is supported by the bivariate analysis.

Despite being able to show new associations which are supported by previous analysis, there are some missing associations as well as some associations represented inaccurately, such as the shown repulsion between the modalities “Psychology” and “AP_1” instead of attraction as in the bivariate analysis and in graph of dimensions 1 and 2. Such a problem is understandable as the quality of representation of the variables is very low.

4.2 Hierarchical Clustering on Principal Components

Hierarchical Clustering on Principal Components allows performing cluster analysis on categorical data by using the results of MCA. Based on the distances between the points in the dimensions calculated by MCA, the most “similar” clusters are grouped together, building each level of the hierarchical tree. The agglomerative hierarchical clustering starts by assigning each datapoint to be a separate cluster. Then, it repeated identify the clusters with the smallest distance (most “similar” cluster) and merge them together until all clusters are merged. For this analysis, agglomerative hierarchical clustering is used, with the Euclidean distance to measure similarity (utilizing results from MCA), and Ward’s method as the linkage criteria which minimizes the total within-cluster variance and merges the pair of clusters with minimum between-cluster distance at each step. Clustering analysis is conducted on the results of MCA on the first 20 dimensions to retain 90.1% of the total variance.

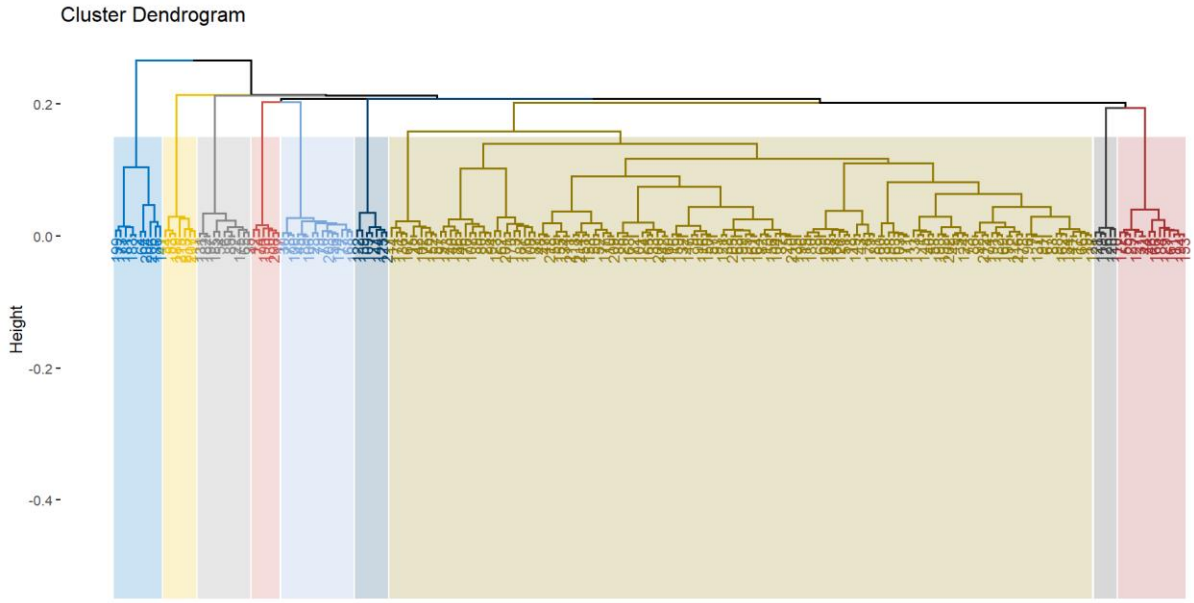


Fig. 5: Dendrogram of hierarchical clustering

Based on this hierarchical tree, it is reasonable to form 9 clusters from the data. This number of clusters will be used for k-means clustering, which partitions the datapoints into k clusters by minimizing total within-cluster variation. From k-means clustering with 9 clusters, the dimensions with the strongest correlation to the clusters are:

	Eta2	P-value		Eta2	P-value
Dim.12	0.74970332	1.178188e-58	Dim.3	0.36325383	3.317316e-17
Dim.2	0.61874716	8.371371e-40	Dim.4	0.34151296	9.252948e-16
Dim.13	0.60330474	4.924606e-38	Dim.16	0.31552303	4.197548e-14
Dim.5	0.52861369	2.250217e-30	Dim.19	0.25214898	2.288767e-10
Dim.14	0.51026765	1.096101e-28	Dim.1	0.24413353	6.353231e-10
Dim.10	0.46148012	1.664892e-24	Dim.6	0.18330475	9.113682e-07
Dim.7	0.41622847	5.645044e-21	Dim.15	0.17928619	1.429357e-06
Dim.11	0.40497977	3.828142e-20	Dim.20	0.14650169	4.849537e-05
Dim.8	0.40233852	5.965887e-20	Dim.17	0.13907167	1.037449e-04
Dim.9	0.38126675	1.904769e-18	Dim.18	0.09924861	4.635724e-03

Fig. 6: Correlation of each dimension to the clusters

As such, the clusters can be are visualized on dimensions 12 and 2 as:

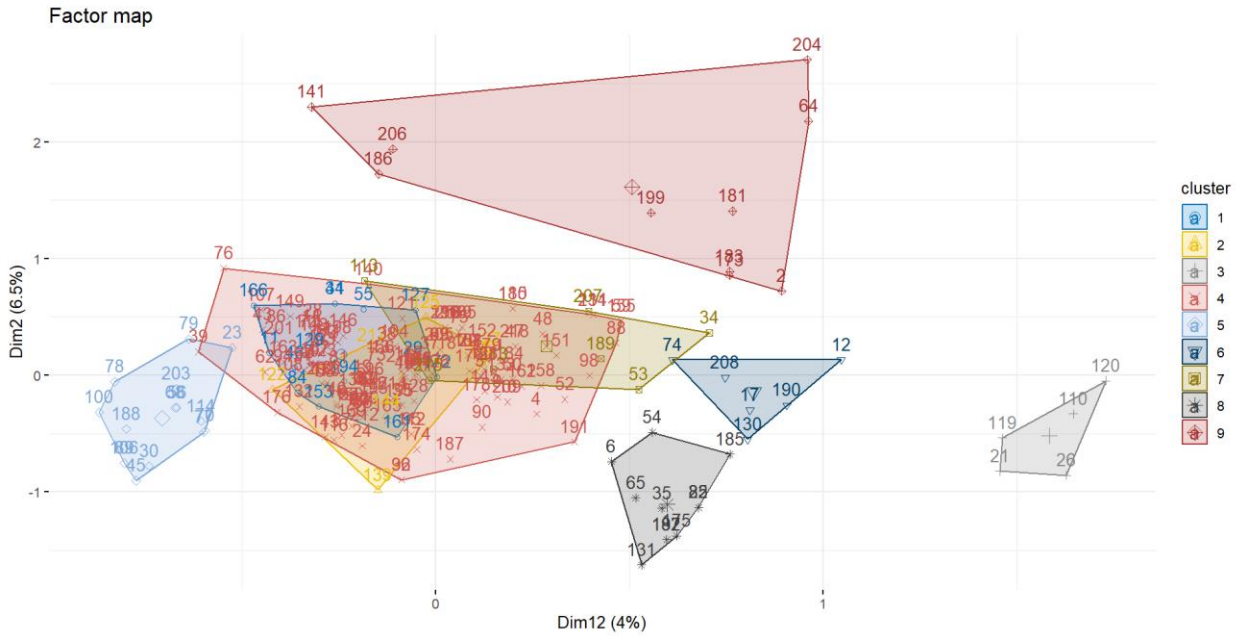
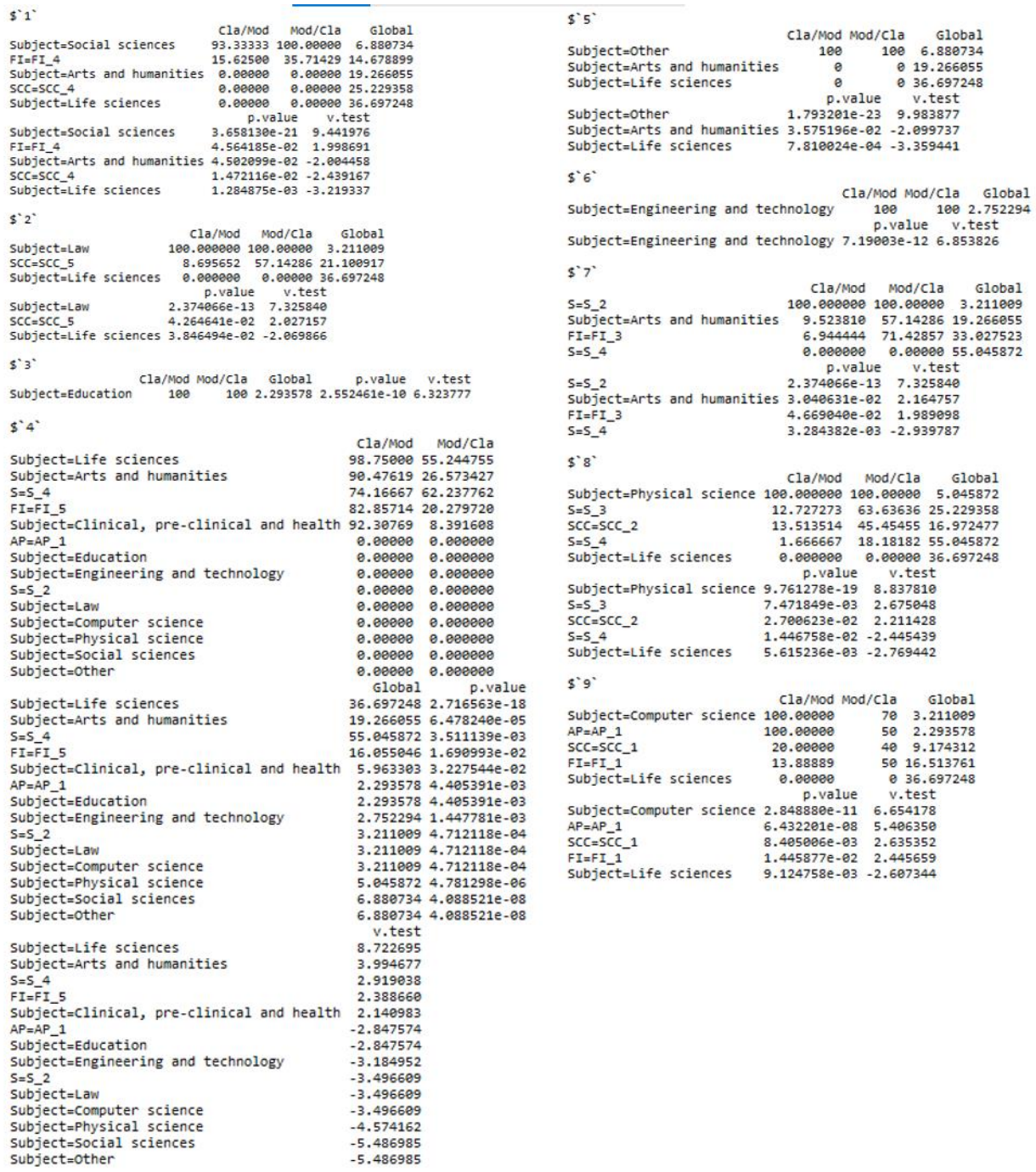


Fig. 7: Clusters on dimensions 12&2

The variables most correlated with forming the clusters are “Subject” and “S” (stress levels). Specific details about the correlation of the modalities and each cluster are shown as following:

	p.value	df
Subject	9.971208e-252	80
S	4.399567e-37	24
AP	5.793651e-14	32
SCC	4.397101e-03	32

Fig. 8: Correlation of variables to the clusters



```
$`1`
                            Cla/Mod   Mod/Cla    Global
Subject=Social sciences     93.33333 100.00000  6.880734
FI=FI_4                     15.62500  35.71429 14.678899
Subject=Arts and humanities  0.00000   0.00000 19.266055
SCC=SCC_4                    0.00000   0.00000 25.229358
Subject=Life sciences        0.00000   0.00000 36.697248
                                p.value    v.test
Subject=Social sciences     3.658130e-21  9.441976
FI=FI_4                     4.564185e-02  1.998691
Subject=Arts and humanities 4.502099e-02 -2.004458
SCC=SCC_4                   1.472116e-02 -2.439167
Subject=Life sciences       1.284875e-03 -3.219337

$`2`
                       Cla/Mod   Mod/Cla    Global
Subject=Law          100.000000 100.00000  3.211009
SCC=SCC_5              8.695652  57.14286 21.100917
Subject=Life sciences  0.000000   0.00000 36.697248
                          p.value    v.test
Subject=Law          2.374066e-13  7.325840
SCC=SCC_5            4.264641e-02  2.027157
Subject=Life sciences 3.846494e-02 -2.069866

$`3`
                Cla/Mod Mod/Cla   Global      p.value   v.test
Subject=Education     100     100 2.293578 2.552461e-10 6.323777

$`4`
                                                    Cla/Mod   Mod/Cla
Subject=Life sciences                              98.75000 55.244755
Subject=Arts and humanities                        90.47619 26.573427
S=S_4                                              74.16667 62.237762
FI=FI_5                                            82.85714 20.279720
Subject=Clinical, pre-clinical and health          92.30769  8.391608
AP=AP_1                                             0.00000  0.000000
Subject=Education                                   0.00000  0.000000
Subject=Engineering and technology                  0.00000  0.000000
S=S_2                                               0.00000  0.000000
Subject=Law                                         0.00000  0.000000
Subject=Computer science                            0.00000  0.000000
Subject=Physical science                            0.00000  0.000000
Subject=Social sciences                             0.00000  0.000000
Subject=Other                                       0.00000  0.000000
                                                      Global      p.value
Subject=Life sciences                              36.697248 2.716563e-18
Subject=Arts and humanities                        19.266055 6.478240e-05
S=S_4                                              55.045872 3.511139e-03
FI=FI_5                                            16.055046 1.690993e-02
Subject=Clinical, pre-clinical and health           5.963303 3.227544e-02
AP=AP_1                                             2.293578 4.405391e-03
Subject=Education                                   2.293578 4.405391e-03
Subject=Engineering and technology                  2.752294 1.447781e-03
S=S_2                                               3.211009 4.712118e-04
Subject=Law                                         3.211009 4.712118e-04
Subject=Computer science                            3.211009 4.712118e-04
Subject=Physical science                            5.045872 4.781298e-06
Subject=Social sciences                             6.880734 4.088521e-08
Subject=Other                                       6.880734 4.088521e-08
                                                      v.test
Subject=Life sciences                               8.722695
Subject=Arts and humanities                         3.994677
S=S_4                                               2.919038
FI=FI_5                                             2.388660
Subject=Clinical, pre-clinical and health           2.140983
AP=AP_1                                            -2.847574
Subject=Education                                  -2.847574
Subject=Engineering and technology                 -3.184952
S=S_2                                              -3.496609
Subject=Law                                        -3.496609
Subject=Computer science                           -3.496609
Subject=Physical science                           -4.574162
Subject=Social sciences                            -5.486985
Subject=Other                                      -5.486985

$`5`
                        Cla/Mod Mod/Cla    Global
Subject=Other               100     100  6.880734
Subject=Arts and humanities   0       0 19.266055
Subject=Life sciences         0       0 36.697248
                           p.value    v.test
Subject=Other           1.793201e-23  9.983877
Subject=Arts and humanities 3.575196e-02 -2.099737
Subject=Life sciences   7.810024e-04 -3.359441

$`6`
                                    Cla/Mod Mod/Cla   Global
Subject=Engineering and technology     100     100 2.752294
                                    p.value   v.test
Subject=Engineering and technology 7.19003e-12 6.853826

$`7`
                          Cla/Mod   Mod/Cla    Global
S=S_2                   100.000000 100.00000  3.211009
Subject=Arts and humanities   9.523810  57.14286 19.266055
FI=FI_3                     6.944444  71.42857 33.027523
S=S_4                       0.000000   0.00000 55.045872
                             p.value    v.test
S=S_2                   2.374066e-13  7.325840
Subject=Arts and humanities 3.040631e-02  2.164757
FI=FI_3                 4.669040e-02  1.989098
S=S_4                   3.284382e-03 -2.939787

$`8`
                             Cla/Mod   Mod/Cla    Global
Subject=Physical science 100.000000 100.00000  5.045872
S=S_3                     12.727273  63.63636 25.229358
SCC=SCC_2                 13.513514  45.45455 16.972477
S=S_4                      1.666667  18.18182 55.045872
Subject=Life sciences      0.000000   0.00000 36.697248
                             p.value    v.test
Subject=Physical science 9.761278e-19  8.837810
S=S_3                    7.471849e-03  2.675048
SCC=SCC_2                2.700623e-02  2.211428
S=S_4                    1.446758e-02 -2.445439
Subject=Life sciences    5.615236e-03 -2.769442

$`9`
                          Cla/Mod Mod/Cla    Global
Subject=Computer science 100.00000      70  3.211009
AP=AP_1                  100.00000      50  2.293578
SCC=SCC_1                 20.00000      40  9.174312
FI=FI_1                   13.88889      50 16.513761
Subject=Life sciences      0.00000       0 36.697248
                             p.value    v.test
Subject=Computer science 2.848880e-11  6.654178
AP=AP_1                  6.432201e-08  5.406350
SCC=SCC_1                8.405006e-03  2.635352
FI=FI_1                  1.445877e-02  2.445659
Subject=Life sciences    9.124758e-03 -2.607344
```

Fig. 9: Correlation of modalities to the clusters

In the above figure, the column “Cla/Mod” describes the percentages of all datapoints in the modality which are included within the cluster, and the column “Mod/Cla” describes the percentages of datapoints within the cluster. It is visible that the variable “Subject” is very relevant for forming the clusters. For instance, the clusters 3 and 6 are formed entirely based on the modalities “Education” and “Engineering and technology” of “Subject”. This can suggest that datapoints within each of these two categories are relatively “similar”. Likewise, most datapoints of the modalities “Life Sciences” and “Clinical, pre-

clinical and health" belongs to the same cluster, suggesting "similarity" between datapoints of these two modalities.

5. Conclusions & Critical Evaluation

Analysis on the data supports the conclusion that certain subjects of study could affect the stress levels and frequencies of stressors to some extent, and that there can be some correlation between the stressors and the stress levels. However, the analysis can hardly conclusively determine any specific correlation with the currently used methods. For instance, analyzing biplots from Multivariate Correspondence Analysis can be insufficient due to the low quality of representation and the low percentages of total variance explained by the first few categories. Similarly, analyzing the clusters from hierarchical clustering can be difficult due to the large number of datapoints; and while analysis of correlation between modalities and clusters reveals some insights regarding the subject of study, it hardly contains information about the correlation between stressors and stress levels. Therefore, follow-up analysis could investigate more methods for analyzing the data, or using different parameters (e.g. different cutting-height or number of clusters).

Furthermore, it can be beneficial if the data is distributed more evenly, especially regarding the variable "Subject". From univariate analysis, data are very unevenly distributed across the modalities of this variable, which can cause analysis of modalities with small number of datapoints more influenced by outliers. On the other hand, the nature of this data can lead to certain biases. As the data is self-reported, (students answering the survey from their own knowledge and evaluation), the same levels of stress or frequencies of stressors reported by different students can differ on an objective scale. Additionally, as the students can choose whether to answer the survey, there could also be compliance bias: students who choose to answer differ from students who do not, so the results of the analysis cannot be generalized to the whole population. As such, the new direction should be finding or collecting a new dataset without these problems and potential biases.