Introduction

In this paper for a dataset on Erasmus mobility obtained from the "EU open data portal" was combined (reference 1) with values for cultural dimensions created by GLOBE in 2004 (reference 2). These cultural dimensions were created via a questionnaire that let to values ranging from 1-7. The result of this combination was a matrix which contains information on the cultural difference between the host and home country measured in the cultural dimensions of the GLOBE project. A second output of the preprocessing was a vector of weights which contains the information on how many students were on an exchange from a certain home to a certain host country.

The goal of the analysis was to see if clusters could be found that cluster exchanges by the cultural experience of students going on Erasmus. An exchange, characterized by Home and Host country, can then be seen as similar to another exchange if the differences in cultural dimensions were similar for the pair of home and host countries. A cluster analysis would result in a cluster of Home-Host country pairs that as an exchange gives a similar cultural experience.

If clusters can be found then it could be interesting to characterize these clusters. For example to investigate which variables/cultural dimensions are characteristic for a cluster. Finally we will look at which home countries are part of which cluster.

Before we dive into the analysis, first a word on the cultural dimensions. I will not list all variables contained in the dataset but will limit the explanation to the two variables that were found most relevant in this analysis:

- **In-Group Collectivism Societal Practices (Collectivism 2) (C.II.S.P):** The degree to which individuals express (and should express) pride, loyalty, and cohesiveness in their organizations or families.
- **Uncertainty Avoidance Societal Values (U.A.S.V):** The extent to which a society, organization, or group relies (and should rely) on social norms, rules, and procedures to alleviate unpredictability of future events. The greater the desire to avoid uncertainty, the more people seek orderliness, consistency, structure, formal procedures, and laws to cover situations in their daily lives.

Analysis

Starting the analysis with an Andrews plot can give hints on the presence of clusters. Indeed the pattern of the lines hints at the presence of possibly 2 clusters (figure 1). After this initial observation using the Andrews plot a clustering was performed using average linkage and euclidean distance as dissimilarity measure. Here the vector of weights was included in the analysis. This means that the distance between the observation is no longer treated as a distance between individual observations but each observation is weighted with the frequency this observation (this exchange/home-host-country-pair) occurs. The clustering resulted in the Cluster Dendrogram presented in figure 2.

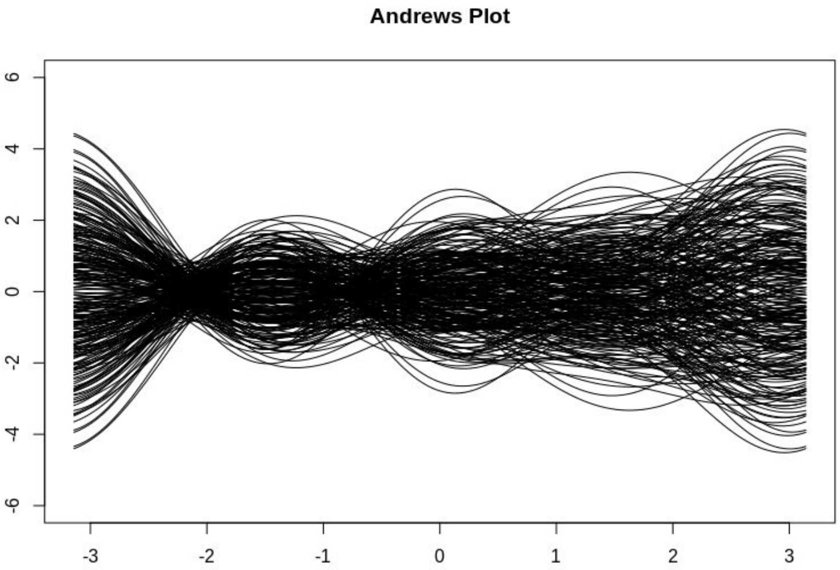

Figure 1: Andrews Plot of the Cultural Difference data. The plot suggests that there are 2 groups present in the dataset because part of the lines (representing the observations) seem to decrease whereas another part of the lines are increasing at the same time and vice versa.

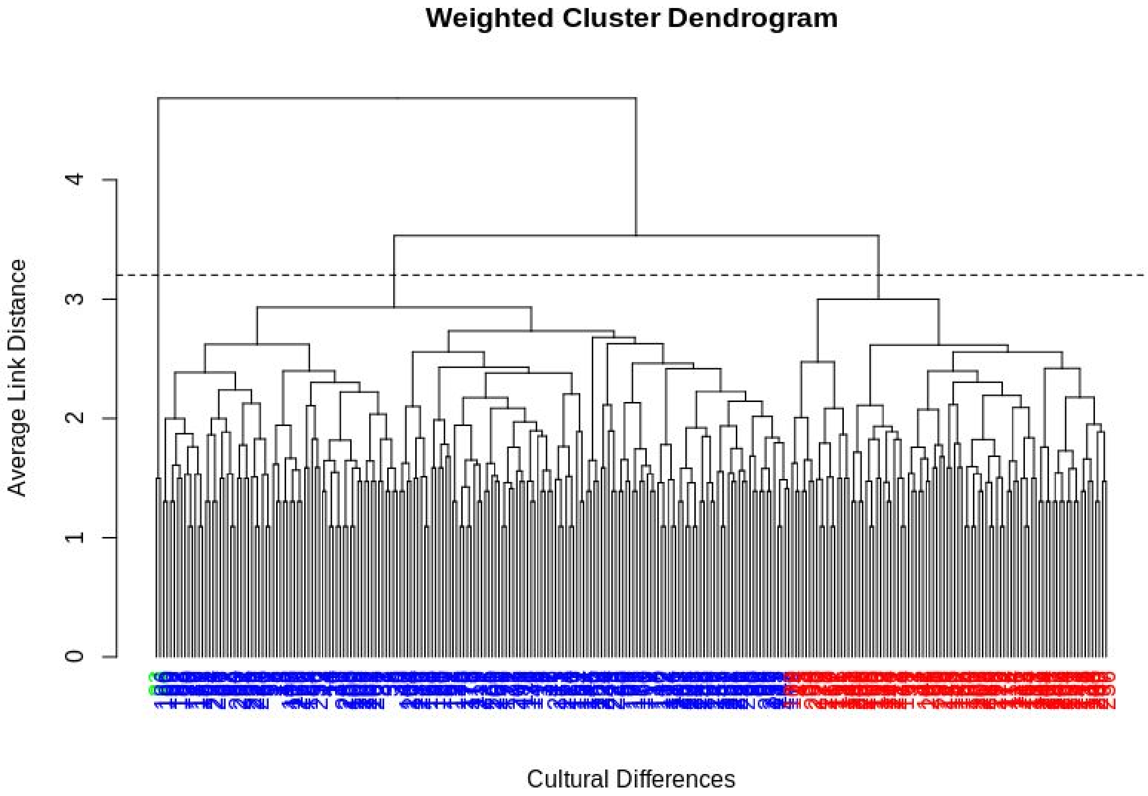

Figure 2: Cluster Dendrogram using the average linkage method and euclidean distance. The two groups suggested by the Andrews plot are visible also in the Dendrogram. Interestingly, we also see a third group which is very different from the other two and hardly any exchanges fall into that group. The clusters are highlighted in colour and the dashed line indicates the line at which the cuttree

functions cuts to create a non-hierarchical representation and assigning the clusters to the different observations.

The impression we got from the Andrews plot is confirmed by the Cluster Dendrogram. We again see 2 groups but also a very small third group which might be interesting to look at. Next, to characterize the clusters a tree was created using the rpart and rpart.plot libraries.

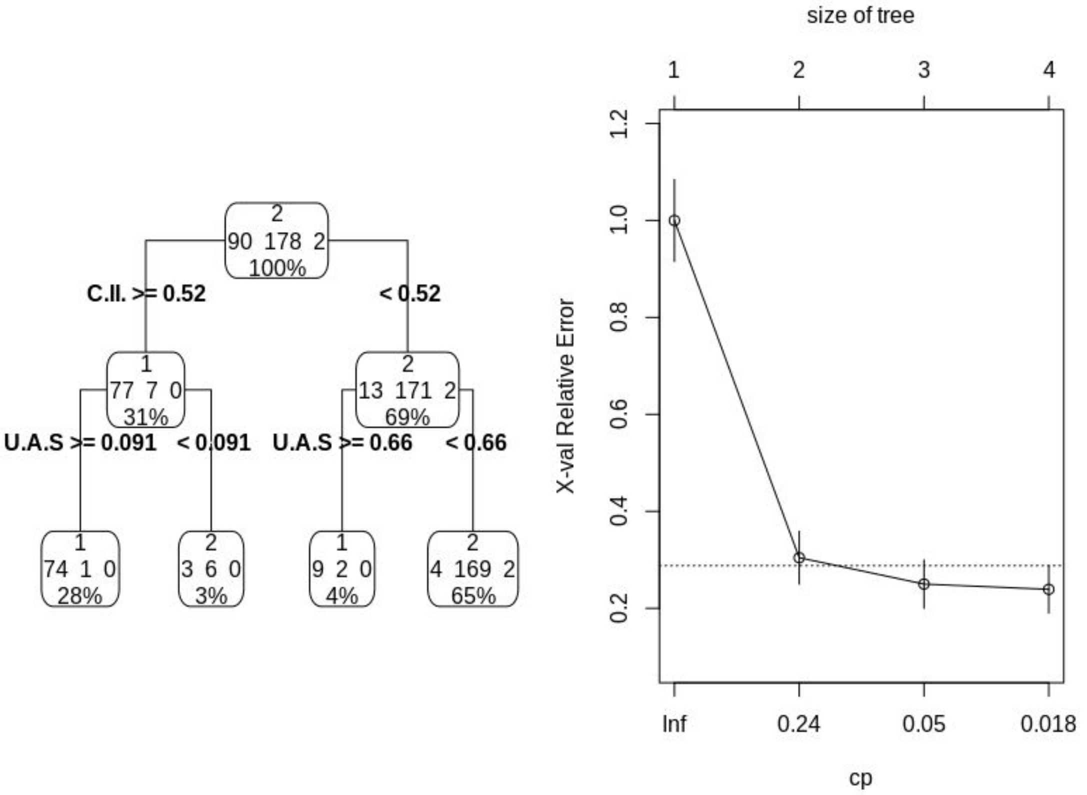

Figure 3: Decision tree predicting class labels obtained from weighted cluster analysis (left) and Complexity Parameter (cp) values leading to different tree sizes (right) and the corresponding relative error. On the left the decision tree displaying the variables that are most relevant in the classification of an observation in class 1 or 2, class 3 is in this decision tree behaving similar to class 2. The relative error versus cost complexity and tree size is plotted on the right. This plot suggests that a tree of size 2 as displayed on the left or of size 3 is the best tree to choose.

The output from the left graph in figure 3 can also be represented by decision rules as displayed in table 1.

Table 1: Rpart output. Terminal nodes are highlighted with an asterix. Most observations of cluster 2 are present in node 7 and most observations of cluster 1 are present in node 4.

```
n= 270
node), split, n, loss, yval, (yprob), *  denotes terminal node
1) root 270 92 2 (0.3333 0.6593 0.0074)
        2)  Collectivism.II.Societal.Practices..In.group.Collectivism.>=0.52  84    7  1  (0.9167
0.0833 0.0000)
        4) Uncertainty.Avoidance.Societal.Values>=0.091 75  1 1 (0.9867 0.0133 0.0000) *
        5) Uncertainty.Avoidance.Societal.Values< 0.091 9  3 2 (0.3333 0.6667 0.0000) *
        3) Collectivism.II.Societal.Practices..In.group.Collectivism.< 0.52 186 15 2 (0.0699 0.9194 0.0108)
        6) Uncertainty.Avoidance.Societal.Values>=0.66 11  2 1 (0.8182 0.1818 0.0000) *
        7) Uncertainty.Avoidance.Societal.Values< 0.66 175  6 2 (0.0229 0.9657 0.0114) *
```

We see that the exchanges in cluster 1 are characterized by going to countries with a lower degree of C.II.S.P and U.A.S.V whereas exchanges in cluster 2 are characterized by going to countries with more similar and possibly higher degrees of C.II.S.P and U.A.S.V in the host than in the home countries.

The right part of figure 3 can also be represented in tabular form as done below. The complexity parameter punishes additional end nodes, this is to prevent overfitting by growing too large trees. In the plot on the right in Figure 3 we see a sharp drop in relative error by including one internal node. Adding extra splits improves the cross-validation error slightly but applying the 1-SE rule (in table 1 xstd) a tree of size 2 or 3 is the best choice.

Table 2: CP values, cross-validation error (xerror) and tree size (nsplit).

```
Classification tree:
Root node error: 92/270 = 0.34074
        CP nsplit rel error  xerror      xstd
1 0.760870      0   1.00000 1.00000 0.084651
2 0.076087      1   0.23913 0.30435 0.054452
3 0.032609      2   0.16304 0.25000 0.049859
4 0.010000      3   0.13043 0.23913 0.048862
```

To further characterize the identified clusters a biplot was created. It confirms the result from the tree that home countries for exchanges in cluster one have higher levels of C.II.S.P and U.A.S.V.

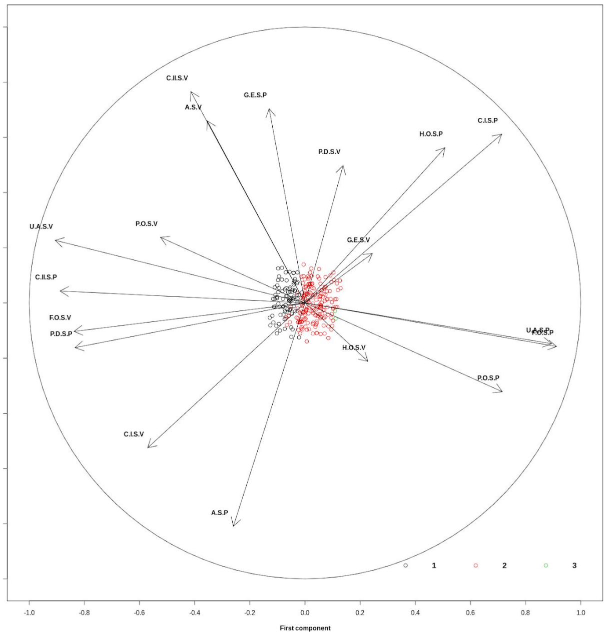

Figure 4: Biplot - Arrows represent variables as factor loadings (eigenvectors multiplied by the square root of the eigenvalues/singular values). The observations are represented by open circles, these are the factor scores (standardized PC scores).

Finally, we move on to see which home countries are over/under represented in which cluster. For this a new dataframe was created containing the percentages of exchanges per country belonging to either one of the three clusters. That means that the percentages per country sum to 1.

The barplot in Figure 4 shows that e.g. students from Hungary, Slovenia and Greece mostly go on exchanges belonging to cluster 1 which means to countries with lower degree of in-group collectivism and lower uncertainty avoidance. Denmark and the Netherlands send students mostly on exchanges where the cultural experience is characterized by the profile of cluster 2, meaning going to countries with more similar or higher degrees of C.II.S.P and U.A.S.V . Cluster 3 interestingly has solely Switzerland as home country.

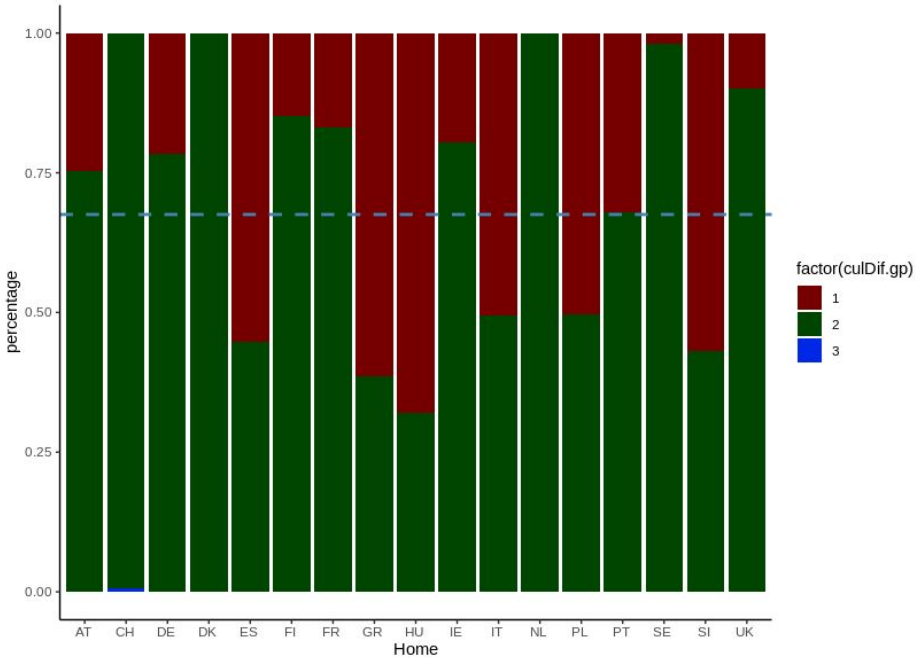

Figure 4: Barplot of class distribution by country. The information of the data frame in table 2 was then displayed in a bar plot using the ggplot2 library. The dashed line represents average distribution of classes 1 and 2 over all countries.

Conclusions

Two main cultural profiles were found which characterize the cultural experience of Erasmus exchanges. However, a few critical remarks are required.

Since the difference of cultural values was used, cluster 1 will be dominated by host countries which have high degrees of in-group collectivism and uncertainty avoidance. This is simply the case because students from these countries can only go to countries for which these values are less important. Another point is that these cultural values may not play a main role in the choice of host country. Major factors were not included such as host university ranking, climate and cost of living but nevertheless some interesting findings could be made.

References

1. https://data.europa.eu/euodp/data/dataset/erasmus-mobility-statistics-2012-13 (retrieved on 14-Nov-2019)
2. https://globeproject.com/study_2004_2007?page_id=data#data (retrieved on 03-Jan-2020)