École Polytechnique Fédérale de Lausanne

Modeling migration intentions worldwide

Hilda Abigél Horváth

Semester project

March 5, 2023

Contents

Abstract

Migration flows have been on the rise in recent years, and this trend has significant impacts on both the origin and destination countries, as well as the individuals involved. There are many factors that influence an individual's decision to migrate, and their choice of destination. Previous research has attempted to model these decisions, but often these models are based on assumptions about which countries should be grouped together. This report aims to understand which countries can be modeled together using a dataset of Gallup World Poll (GWP) and country-specific attributes. The methods used in this report range from decision trees to regression models, and include different clustering techniques. These methods produce different clusters, with some being more effective in grouping together neighboring countries. However, the ultimate test of these clusters will be to use them in predictive models to see if they improve performance.

1 Introduction

Migration has become a hot topic in recent years, as the number of people moving from one place to another has increased significantly. This trend has raised a number of concerns, particularly in industrialized countries where many migrants choose to settle. Integration, for example, is often cited as a major concern, as migrants may struggle to adapt to their new surroundings and find their place in their new communities. There is also the potential for a loss of cultural identity as migrants adapt to their new environments and assimilate into their host societies. In addition, the financial costs associated with migration can be a source of worry for both migrants and their host communities, as the burden of supporting new arrivals may be perceived as a strain on resources.

On the other hand, there are also fears that low-opportunity countries may experience a brain drain as highly-skilled workers migrate to more developed nations in search of better opportunities. This phenomenon can have a negative impact on the long-term development prospects of these countries, as they lose some of their most talented and educated citizens.

The study of migration is a complex and multifaceted endeavor, as it involves understanding the various factors that influence both migration aspirations and destination choices. In this report, we aim to shed light on this topic by examining the similarities between countries from different points of view. Specifically, we will group countries based on the ways in which their citizens decide to migrate, as well as the general characteristics that make them similar in terms of migration patterns. By understanding these similarities, we can gain valuable insights into the motivations and decision-making processes behind migration, and how they vary across different countries.

The report is structured as follows: first, we will provide an overview of previous work in the topic of migration flows, and highlight the research gap that this study aims to fill. Next, we will describe the two datasets used for the analysis: the Gallup World Poll, and the country attributes dataset. We will then describe how countries of origin can be clustered based on the responses of residents. The responses are aggregated by means or modes for each country, and K-Means and K-Modes clustering algorithms are applied. In the following section, we will present two methods for clustering countries based on residents' decision making regarding permanent migration to another country. The first method involves training a decision tree and clustering countries based on the distribution of responses on the leaves. Five different variants are compared. The second method involves training a decision tree for each country and performing hierarchical clustering. We will also demonstrate how countries can be clustered based on their popularity as destinations for migration. Finally, we will summarize the findings of the study and provide conclusions.

2 Previous work

2.1 Background

The existing literature is rich in studies that aim to understand the movement of people between different countries, and can provide valuable insights into the factors that influence this flow. As the main focus of the report is finding similarities between countries whose residents tend to take similar decisions regarding migration, this review includes previous work both on migration aspiration and destination choice.

In all mentioned studies, the Gallup World Poll dataset is used, with the difference that the available data increased with the course of time, so more years are used in the more recent reports.

Table 1 ans 2 below summarize existing studies on the topic. The table 1 indicates the choice model and the specific research question examined in each study. Most of the studies included in the table consider both aspiration and destination choice, while two papers only examine destination choice. All studies use a logit model, with the exception of two articles that employ nested logit model and a cross-nested logit model.

Table 2 summarizes which countries are considered in each study. In all but two cases, only a smaller subset of countries is used for origin or destination.

To understand better why the work of this report is relevant, it is important to summarize the previous works on aspiration and destination choice modeling.

Lovo's paper [1] provides evidence that in European countries potential migrants prefer those destinations where the average life satisfaction is relatively higher. However, the results can not be generalized to other countries.

Gubert and Senne's article [2] investigates what factors are determinants for potential migrants when choosing the European Union as a destination and comparing the resulting profiles of migrants with those who have recently resided in OECD countries.

Bertoli and Ruyssen's paper [3] use data from the Gallup World Polls to examine the relationship between an individual's connections to migrant networks in various countries and their preference for a specific destination country.

Docquier et al.'s article [4] conducts an analysis to determine if emigrants from the MENA region tend to have certain cultural traits, such as religiosity and gender-egalitarian attitudes when making their emigration decisions.

Bekaert et al.'s study [5] uses individual-level data to examine the relationship between an individual's reported experience of environmental stress and their intention to migrate, as well as their preferred destination country.

Beine et al. [6]'s study presents a case study of migration aspiration data from India and shows that the used cross-nested logit model outperforms other standard approaches in terms of the quality of fit and predictive power.

Langella and Manning [7] focus on the influence of household and country-level personal income on both the desire to emigrate and the preferred destination country.

Studies	**Choice**	**Model**
Lovo, 2014 [1]	asp., dest.	logit
Gubert and Senne, 2016 [2]	dest.	logit
Bertoli and Ruyssen, 2018 [3]	dest.	logit
Docquier et al., 2020 [4]	asp., dest.	logit
Bekaert et al., 2021 [5]	asp., dest.	logit
Beine et al., 2021 [6]	asp., dest.	cross-nested logit
Langella and Manning, 2021 [7]	asp., dest.	nested logit

Table 1: Previous studies on migration aspiration and destination

Studies	**Origin**	**Destination**
Lovo, 2014 [1]	25 EU c.	24 c.
Gubert and Senne, 2016 [2]	150 c.	EU, US, OECD, no OECD
Bertoli and Ruyssen, 2018 [3]	147 c.	6-58 c.
Docquier et al., 2020 [4]	17 MENA c.	EU, US, OECD, no OECD
Bekaert et al., 2021 [5]	90 c.	OECD, dom., intra-reg.
Beine et al., 2021 [6]	India	86 c.
Langella and Manning, 2021 [7]	159 c.	199 c.

Table 2: Previous studies on migration aspiration and destination

2.2 Research objective

There are several research gaps in the field of migration aspiration and destination choice studies that need to be addressed in order to deepen our understanding of these patterns.

As, each article either focuses on a specific point of view of the effects of the variables and provides evidence for their influence on migration aspiration and decision choice, or uses only a subset of the countries to build models, one key objective could be to generalize the findings of previous works and build a more complex model. For example, one could generalize the framework of Beine et al. [6] to a wider range of countries.

However, this project aims to address other research gaps and aims to understand which countries of origin can be modeled together and which countries of destination can be nested. As in the previous works, the authors assume that some countries share some characteristics and use these assumptions in their models' structures. To elaborate on this, they are using only a subset of the countries as origin or destination, or grouping them together as destination by their involvement in the EU, OECD or US. This help them to reduce computational time and model complexity. However, by identifying an ideal grouping of countries the key difficulty of computational time and a large number of countries can be overcome, since these countries can be modeled together.

In this project we aim to understand how these countries can be clustered by the means of exploration. These clusters of countries could then inform the formulation or structure of choice models. We uncovered valuable insights. However, in the scope of this semester's project, the clustering of countries is not tested with the logit model. This would be the next step of future work.

3 The Gallup World Poll dataset

3.1 Introduction

The Gallup World Poll (GWP) is a yearly worldwide survey, conducted in more than 150 countries. The respondents are asked several questions regarding their opinion on various topics and their life situation. Approximately 1000 people are questioned in each country in each wave of the survey. The distribution of respondents across countries can be seen on Figure 1.

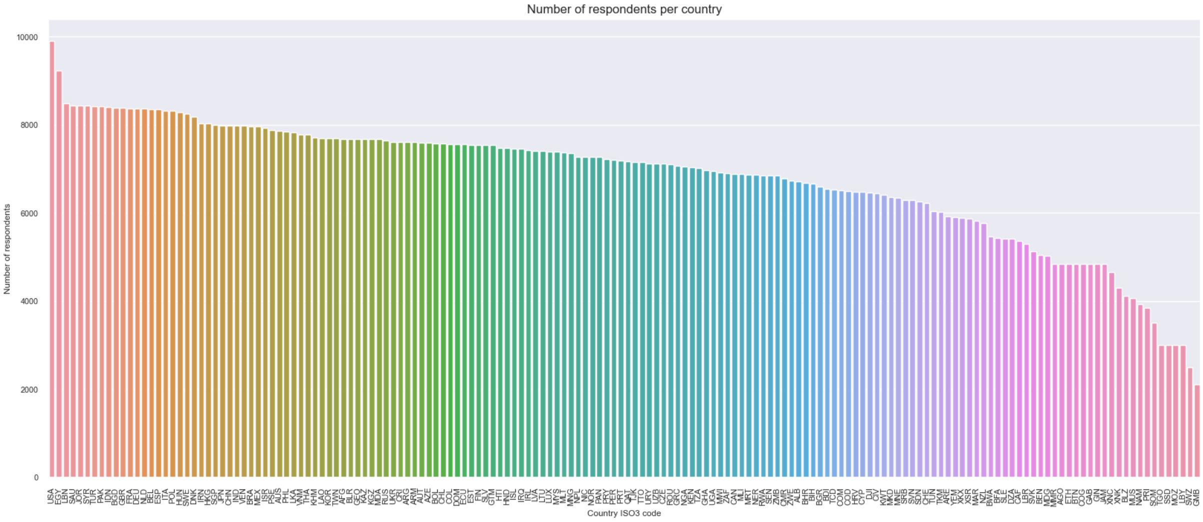


Figure 1: Gallup World Poll: Number of respondents per country (2008–2022)

The three countries with the most respondents are the USA (9234), Egypt (8492), and Lebanon (8442). The countries with the fewest respondents are Gambia (2109), and the Kingdom of Eswatini (2499).

The prepared dataframe includes 1104558 observations and 130 features. There are observations in 162 countries. The surveys are conducted from 2008 to 2022. However, we used the whole set of data only to carry out the results of this section, the results of the other sections are based on observations from 2018 to 2022. We added the functionality in the code, of using different years as the basis of the experiments, so one can change it easily to experiment.

To give an idea about the dataset, here are some basic statistics about it.

- Age: the mean is 41.33 and the standard deviation is 17.71. The values are ranging from 15 to 99.
- Gender: The proportion of women vs. men in the answers is 53,60% and 46,40%, the remaining refused to answer or no answers were given.

3.2 Migration aspiration in GWP

As a basis of the analysis, these two questions related to moving to a different country in the GWP were used.

1. (WP1325) Ideally, if you had the opportunity, would you like to move permanently to another country, or would you prefer to continue living in this country?
2. (WP3120) To which country would you like to move? (Asked only to those who would like to move to another country.)

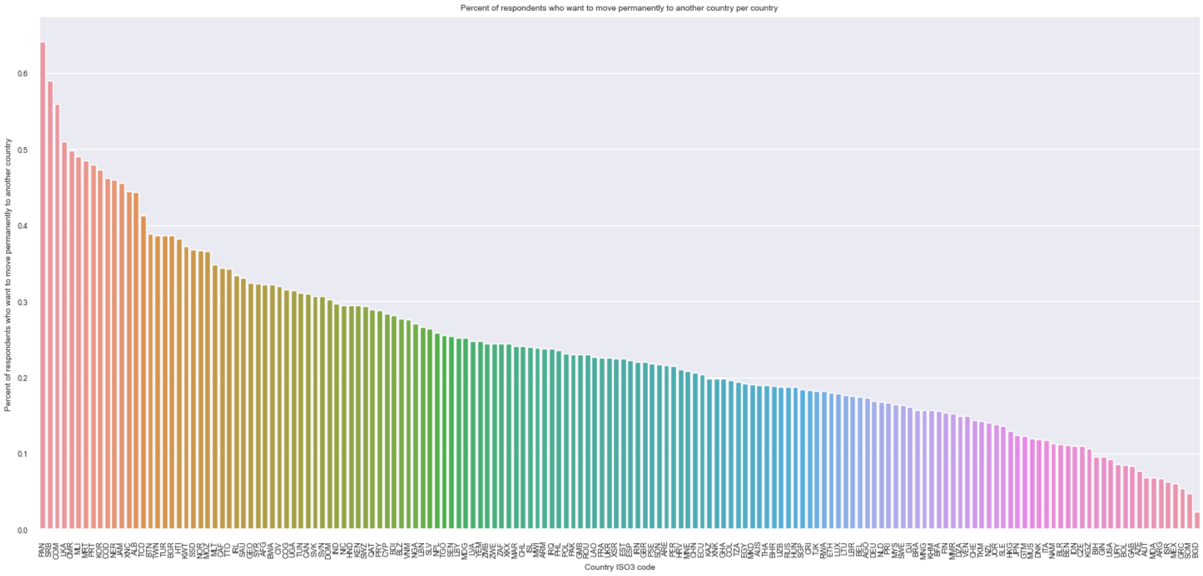


Figure 2: Percentage of respondents aspiring to move permanently to another country

As shown in Figure 2, the percentage of respondents expressing their aspiration to leave their country varies greatly across the world, ranging from 2.39% to 64.17%. By multiplying these percentages by the populations of the corresponding countries, we obtain the actual volumes of migration intentions, which are shown in Figure 3. The two largest volumes belong to India and China, while the smallest non-zero volume belongs to Iceland. Tables 3 and 4 provide additional details on the percentages and volumes of migration intentions.

3.3 Destination choice in GWP

In the Gallup World Poll, respondents who indicated that they would like to move permanently to another country are asked which country they would choose. Table 5 presents the top 10 and bottom 10 destinations for such respondents. Figure 4 shows the percentage of respondents from every country who selected each destination as their preferred location. This information can give us insight into the preferences and priorities of respondents when considering a move to another country. The top 10 destinations are predominantly composed of developed countries

Country	TOP 10 (%)	Country	BOTTOM 10 (%)
Panama	64.17	Gabon	8.38
Serbia	58.98	Azerbaijan	7.71
Comoros	55.92	Austria	6.81
Sri Lanka	50.98	Moldova	6.79
Cameroon	49.84	Argentina	6.74
Mali	49.03	Israel	6.24
Mauritania	48.51	Mexico	6.08
Portugal	47.93	Greece	5.39
South Korea	47.27	Somalia	4.73
Democratic Republic of the Congo	46.17	Bangladesh	2.39

Table 3: Percentage of respondents who want to move permanently to another country, top and bottom 10 countries

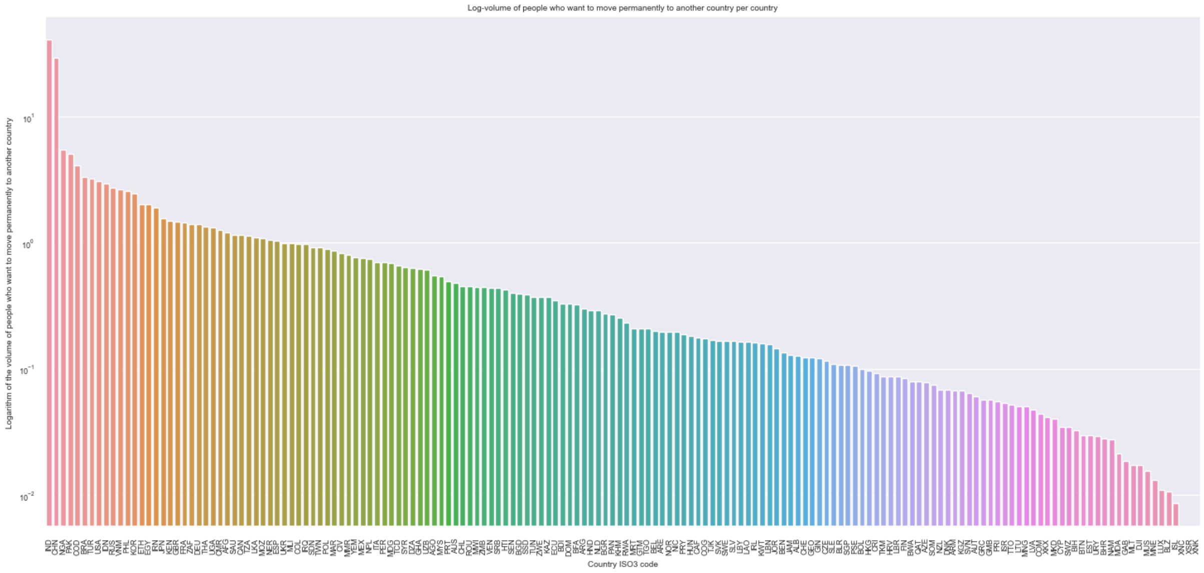


Figure 3: Log-volume of people who want to move permanently to another country per country

with strong economies and high standards of living, while the bottom 10 destinations are largely made up of developing countries. This can suggest that respondents may prioritize economic stability and quality of life when considering a permanent move to another country.

In the heatmap below 4, each row represents an origin country and each column represents a destination country. The color of a cell is darker if more people are choosing that destination country as a destination. The heatmap also shows that among the 200 destination countries, few are particularly popular, while the majority are not.

Country	Volume of people who want to move permanently to another country
India	40.77k
China	29.29k
Nigeria	5.44k
Pakistan	5.11k
Democratic Republic of the Congo	4.09
Brazil	3.32k
Turkey	3.21k
United States	3.09k
Indonesia	2.95k
Russia	2.73k

Table 4: Top 10 countries with the highest volume of people who want to move permanently to another country

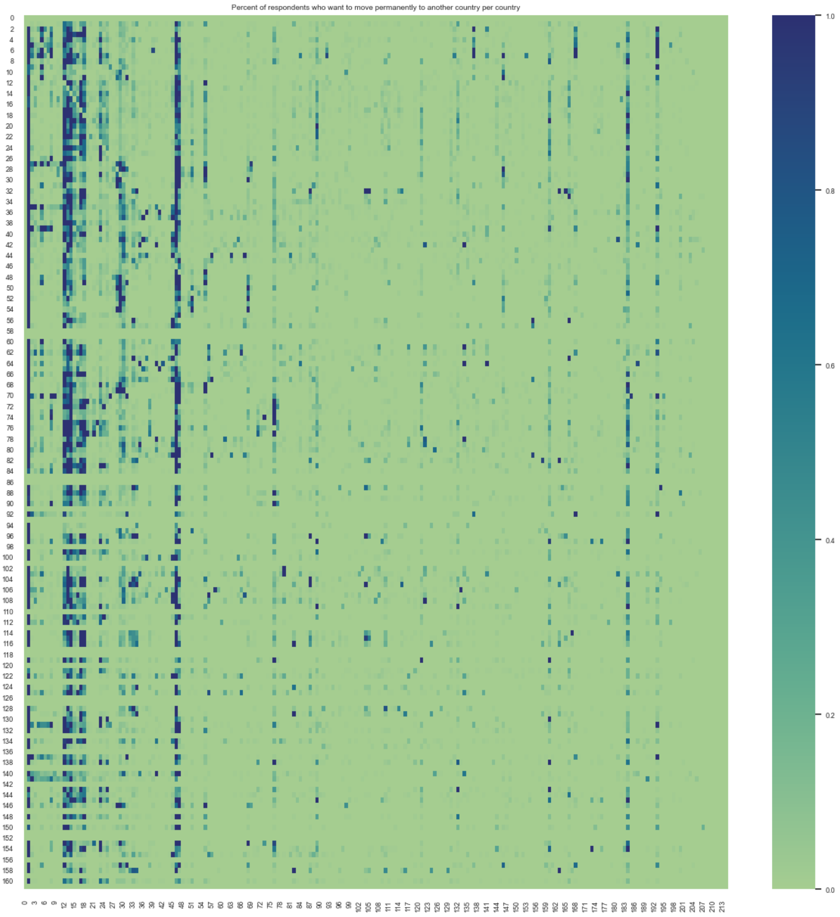


Figure 4: Percent of respondents who want to move permanently to another country per country

Country	**TOP 10 (%)**	**Country**	**BOTTOM 10 (%)**
United States	20.80	Sao Tome & Principe	0.002342
Germany	6.74	Sudan	0.001951
Canada	6.16	St. Vincent & Grenadines	0.001951
France	5.62	Somalia	0.001951
United Kingdom	4.92	Serbia	0.001951
Spain	4.70	Yemen	0.001561
Australia	3.42	Mauritius	0.001561
Italy	2.91	Panama	0.001561
Ukraine	2.38	Trinidad & Tobago	0.000781
Tajikistan	2.04	United Arab Emirates	0.000390

Table 5: Percentage of respondents who want to move to the country, top and bottom 10 countries

4 Country characteristics dataset

For modeling migration, the characteristics of the countries influence the destinations chosen by the individuals. To this aim, we use a dataset that includes different important characteristics of the countries such as the official or used language, the religion, the population, or the climate. These indicators are important to gain meaningful insights into the decision-making process of the GWP respondents.

Feature	Description
GDPPC	GDP per capita between 2007-2021
logGDPPC	Logarithm of GDP per capita between 2007-2021
Christian	The main religion is Christian
Muslim	The main religion is Muslim
english_official	English is an official language
spanish_official	Spanish is an official language
french_official	French is an official language
english_used	English is used
spanish_used	Spanish is used
french_used	French is used
POP	Population between 2017-2021
logPOP	Logarithm of population between 2017-2021
OECD	Country is part of OECD
schengen	Country is part of Schengen
freedom	Freedom categorization between 2005-2022
tropical	Country's climate is tropical
temperate	Country's climate is temperate
dry	Country's climate is dry
polar	Country's climate is polar
continental	Country's climate is continental

Table 6: Country characteristics

5 Clustering countries of origin based on the residents' responses

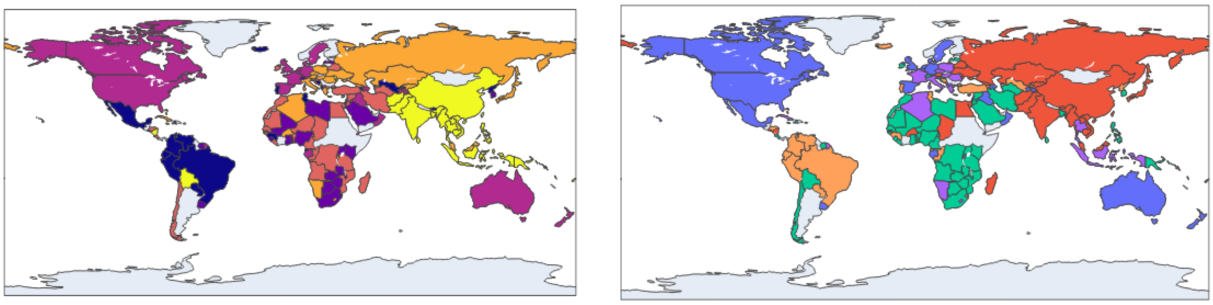
5.1 Introduction

As a first step of the analysis, I started clustering the countries based on the responses of their residents. This helps understand the data in detail and gives a first idea of which countries share some similarities.

5.2 Data preparation

The first step is the data preparation. As the dataset includes a lot of missing values which decision trees can not handle, the first step is to tackle this problem. As some questions are only asked in a specific group of countries, as an initial step I removed them, since these features are not useful to compare countries. When predicting the migration aspiration, it is also necessary to remove those observations where this variable includes as an answer "DK" meaning "Don't know" or the person refused to answer. The remaining missing values are imputed based on the type of data. For categorical values, a new category is introduced for these missing values. Regarding the continuous features, the mean is imputed, and another feature is added to not lose information representing if the feature was missing or not.

As the last step, the prepared dataset is joined with the country characteristics dataset described in Section 4.



(a) Mean aggregation

(b) Mode aggregation

Figure 5: Clustering the countries of origin

5.3 Methodology

To perform the clustering, the following simple methods are used. Firstly, for each country, each feature is aggregated. So, we calculate one value for each feature for each country. After that, an appropriate clustering method is used. As each country is represented as a vector by the aggregation, the clustering can be easily done. The main difficulty is that the data is mixed, so using solely K-Means or K-Modes is not appropriate. However, the data mainly consists of yes or no questions, with a few true categorical and a few continuous features, so I tried to use both ways. First, I assumed that all features are continuous, aggregated them by the mean, and used K-Means for clustering. After this, the features were assumed to be categorical, aggregated by their mode, and the countries were clustered with K-Modes. The elbow method was used to determine the optimal number of clusters, which is 6 in the continuous case and 5 in the categorical case.

5.4 Results

As the mixed-type dataset is not handled correctly, we cannot expect really good results. However, in Figure 5, we can see that even with these assumptions the spatial positions of the countries are roughly recognized by the method, so neighboring countries that are actually similar are clustered together. This is good news since it is consistent with the intuition that people in neighboring countries may have similar opinions as they are culturally and economically closer.

More specifically, in the mean aggregated (a) variant, it can be seen by the yellow cluster which clusters together South-East Asia, South Asia and East Asia is cluster, and the purple cluster which includes the USA, Canada, Australia and Western Europe. On the other hand, Africa, the Middle East and Eastern Europe seems random.

In the mode aggregated (b) variant Africa, Asia and South America have their own clusters. On the contrary, the clustering in Europe and Middle East looks unjustifiable. This result is looking better, this might be the because the dataset is closer to a categorical-type dataset.

It is also important to mention that this method assumes that the mode and mean are representative in each country, so the data is collected properly.

6 Clustering countries based on migration aspiration

6.1 Introduction

In this section, I will describe different methods to cluster countries together based on the decision-making of the residents on leaving the country permanently. The main idea behind the first method of clustering countries is to build a decision tree predicting migration aspiration and examine the distribution of observations across leaves for each country. We will show, 5 variants of this method. The other method stems from the idea to cluster together the countries that have a similar decision tree when predicting migration aspiration.

6.2 Data preparation

The previously described data preparation is extended with a few steps. So, the first step is removing questions that are only asked in a specific group of countries. After that, the missing values are imputed. This will be the basis of the further steps.

In one of the previous sections, it was shown that the number of observations in different countries is not the same. In addition to this state, the previous preprocessing steps can also make the distribution of the number of answers between different countries unequal. In the first method, it is a problem as only one decision tree is trained and a country with a lot more observations as others can influence heavily the feature selection during the training. The used solution to overcome this problem is to sample the same amount of observations from each country. The last step of the data preparation was to merge this dataset with the country characteristics.

In the second method, training a different tree as a first step might suggest that we don't need sampling in this case, however without that when merging the trees one country can dominate the other, so it might result in bad results.

6.3 Method 1: Countries on the same decision tree

The first method consists of building a decision tree to predict migration intentions with the full dataset and examine the distribution of observations on the tree leaves. We consider two countries as similar if the observations on the decision tree from the countries are similarly distributed on the leaves on the tree. I will introduce the exact metrics later. The first step for this methodology is to prepare the data to be able to build the decision tree and after that training the tree comes, and the last step is to find the right metrics to examine the observations.

6.3.1 Decision tree

For training the decision tree, scikit-learn (version: 1.1.2) [8] is used, which currently does not support categorical features. This requires further consideration since most of the available features are categorical. For this problem, one-hot encoding of the features is used. The above-mentioned package uses the CART algorithm when training algorithm and Gini impurity to measure the quality of the clustering during the training.

When using decision trees, choosing the maximal depth of the tree is also a crucial part of the training. The depth determines the number of features for prediction as well as the number of clusters in the first two variants of the method. The decisions for choosing this parameter will be elaborated in the following section.

6.3.2 Clustering of countries

After training the decision tree, the variant of the method for clustering the countries should be chosen. There are 5 different variants of this main idea presented in this report. The considered ways are the following:

1. Put those countries in the same cluster, whose observations are represented on the same leaves on the tree.
2. Cluster those countries together, whose top two leaves where they are most represented are the same.
3. Represent as vectors the distribution of answers on the leaves for each country and cluster these vectors with DBSCAN using Euclidean distance.
4. Represent as vectors the distribution of answers on the leaves for each country, and cluster these vectors with K-Means using Euclidean distance.
5. Represent as vectors the distribution of answers on the leaves for each country and cluster these vectors with agglomerative clustering using ward linkage, which minimizes the variance of the clusters being merged at each step.

For the first two variants, the maximum depth of the tree is determining the number of clusters, hence the depth should be chosen to be a small number, in this case, 3 is the used parameter in the case for migration aspiration.

In the 3rd variant, DBSCAN doesn't need an initial number of clusters, only carefully chosen parameters to cluster points.

In the 4th and 5th variants, the number of clusters is chosen to be 8 and 6 clusters based on the elbow method. The clustering is done with the K-Means algorithm with Euclidean distance and agglomeration clustering with Ward-linkage.

The last variant is determining an optimal number of clusters for the minimal variance criteria which is two in this case, but seemingly with other small numbers the method is also working.

6.3.3 Results

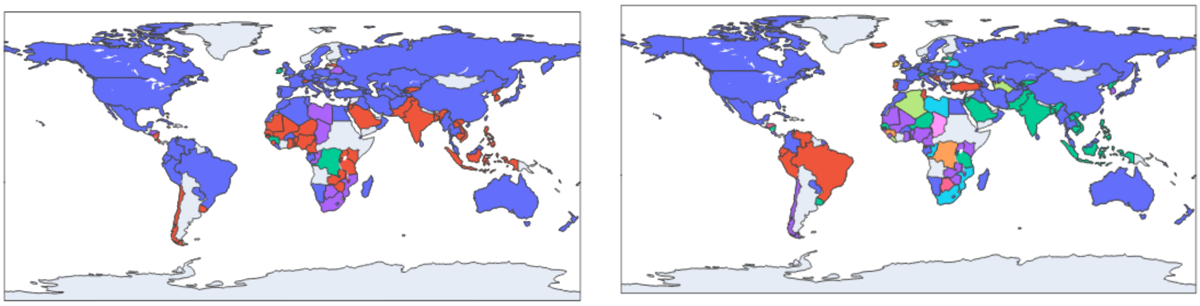
The first experiment used data including REG-GLOBAL and REG2-GLOBAL features as well. The results of these could capture the spatial position of countries really well, so in order to compare experiments without these features were also carried out. There were only minor difference between with and without region results, so these features were presented in not the most important features, we decided to include only one version of the results, the one without region features.

Results for D1 and D2 variants

In the D1 and D2 variants, I received the decision tree described in Appendix 9.2, written as a python code. In this code representation, the if-s represent the right branches and else-s the left branches in the decision tree. In the returns, the number of yes and no samples are written. This tree's predictive power is 76.14% accuracy.

In the case without the region features the same accuracy can be achieved, with the same decision tree. So, the region variables might be present in later levels of the tree, they are not presented in the top 4 most important features. As a result in the two cases, the D1 and D2 variants the clusters are the same with or without region features. In the below sections, the results of the experiments without the region features are introduced.

The clustering based on these trees with the D1 and D2 variants results in the following figures. This variant due to its lack of complexity is unable to capture complex relationships between countries, as it is only taking into account the top 4 features.



(a) D1 variant

(b) D2 variant

Figure 6: Result of the D1 and D2 variants (The list of countries belonging to each cluster is available in Appendix 9.3 9.4.)

As it can be seen in Appendix 9.2 the most essential feature is Age, with a splitting point of 43.5 years. After that, in both branches, if the person is born in the same country as she resides is most important. Another important factor is the quality of water. This is in accordance with the preconception that the age of the residents is highly influential on their willingness to migrate, and if they have already moved to another country.

In both cases in Figure 6 we can observe the following. In the blue cluster, we can see that going down three times to the right in the decision tree classifies well the yes answers, so in these countries, those people who would like to migrate are more likely to be younger than 43.5 years old and did not bear in the country.

In the second most represented cluster the decision rule is also highly dependent on the age and whether the respondent was born in the same country where resides. So, in these countries, if somebody is older than 59.5 and was born in the same country where he resides is a good rule for classifying those people who would like to stay.

As can be seen from this as well, the complexity of the tree can not be enough for the clustering, since the decision rules are rather similar in each case and only uses a few features. Age and the fact that somebody already moved to that country are well-known factors to determine the willingness to move permanently abroad, so they are determinants in the tree is expected, but solely using these can not be enough to observe more sophisticated connections between countries.

Result for D3 variant

In the following cases, the maximal depth of 8 and no restriction on depth is used to compare. As a result, the above tree's expanded form is used, however, due to its large size it can not be included in this report. The predictive power of the depth 8 tree is 76.13% accuracy.

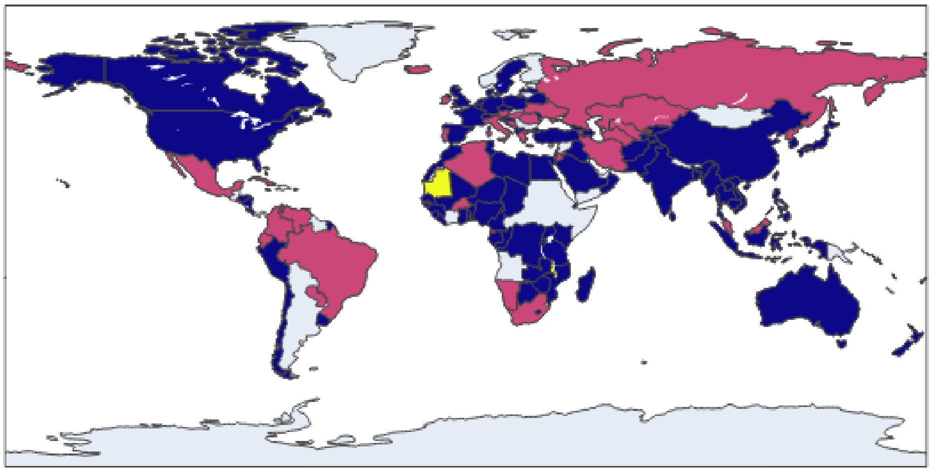


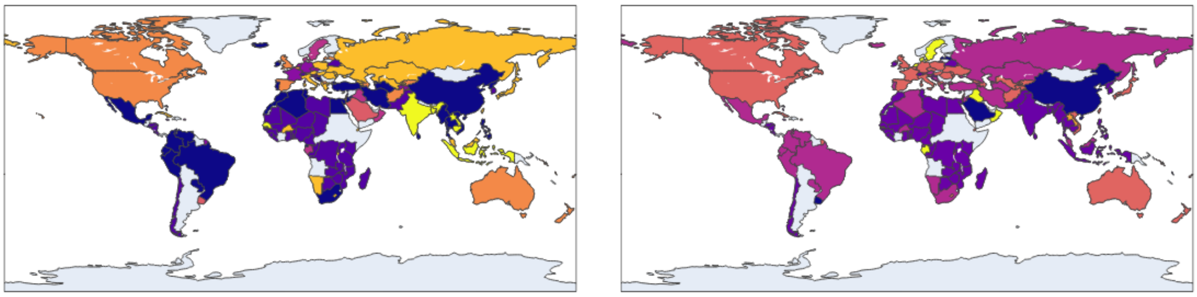
Figure 7: Result of the D3 variant (The list of countries belonging to each cluster is available in Appendix 9.5.)

In the resulting clustering shown in Figure 7, there are three clusters where one of which only includes Malawi Mauritania. The other countries are distributed between the remaining two clusters. Unfortunately, interpreting the clusters in more detail is impossible due to the behavior of the algorithm. What we can see in the plot, is that in this variant of the method, the spatial organization of the clusters is seemingly random, but the neighboring countries are more likely to be clustered together as previously. The reason for the randomness can be that DBSCAN is not robust in the case when the clusters are not well-defined in the high-dimensional space. It is interesting to note that as previously Africa seems the most varying the terms of the position of different clusters, but here neighboring countries are more likely to be clustered together there as well. As a result, there can be still some hope to fine-tune the hyper-parameters.

Result for D4 and D5 variants with 3 clusters

The variants were first tried out with 8 and 6 clusters. As can be seen, hierarchical clustering and K-Means are better at capturing the spatial position of countries, as they are clustering neighboring countries together with higher probability.

In both cases, training the decision tree to the maximum depth was not a good idea as the clustering is overfitting. It can be seen from that it is putting a single country into one of the clusters. This justifies that in both variants it is reasonable not to train the trees until the end and use a cut-off threshold. In the experiments, the maximum depth was chosen to be 8 based on empirical reasons.



(a) D4 variant, 8 clusters, 8-depth

(b) D5 variant, 6 clusters, 8-depth

Figure 8: Result of the D4 and D5 variants with an 8-depth tree. (The list of countries belonging to each cluster is available in Appendix 9.6, 9.7.)

In the K-Means (D4) variant in Figure 8a, we can observe that there are some countries that are similar to our preconceptions. For example, the USA, Canada, and Australia are clustered together with a few European countries. Former USSR countries are clustered together with some Eastern European countries. Moreover, the purple cluster only contains countries from Africa. On the other hand, the dark blue cluster seems illogical, it contains South America, a few African countries, a few countries from the Middle East, and from East Asia.

It is interesting, that in the hierarchical clustering in Figure 8b we can see similar patterns. The orange cluster with the USA, Canada, Australia with European countries is an explainable cluster. The purple one also, with African countries, South Asia, and South-East Asia. On the contrary, the yellow and the blue clusters are looking random.

As the clustering algorithms are working on high-dimensional feature vector representations of the countries the interpretations is not meaningful. On the other hand, the results of these methods are the most promising, as neighboring countries are more likely to be clustered together. We can conclude from these clusterings in Figure 8 that the clusters are in accordance with the expectation that similar countries in terms of culture, climate, and wealth are clustered together. It is also reasonable, that there are some other factors that are influential in migration aspiration, and making the clusters less dense on the map.

Result for D4 and D5 variants with other number of clusters

For K-Means the optimal number of clusters can be determined using the elbow or the silhouette variant. With these variants, the optimal number of clusters is above 8 and 6. However, this is questionable if this is indeed the optimal number since it is not tested for the aim presented in the introduction.

6.4 Method 2: Hierarchical clustering for decision trees

6.4.1 Introduction

The main idea behind this method is that first for all countries a different decision tree is trained to predict migration aspiration. After that with the method of hierarchical clustering in each step pairs of countries are merged and their associate tree is trained again until the desired number of clusters is achieved.

6.4.2 Decision tree

In this case also scikit-learn (version 1.1.2) [8] is used, with CART algorithm and Gini impurity. So again, the categorical values are encoded with one-hot encoding.

The maximum depth of the trees is chosen to be different numbers: 6,8 and 10, so for every country, the first 6,8 or 10 most important features are represented in the decision trees.

6.4.3 Hierarchical clustering of trees

We train a different tree for each country. The next step is merging them to achieve the desired number of clusters. For this aim hierarchical clustering is used, where in each step we choose the pair of trees that are most similar. We retrain a tree on the data of both associated countries and get one tree with the set of data. We measure the similarity between two trees as the average accuracy of classifying the data associated with the other tree. In each step in the hierarchical clustering, we choose the pair of trees, with the highest average accuracy and merge the two datasets and retrain a tree on the resulting dataset.

6.4.4 Results

Unfortunately, this method did not show good results. The phenomena we could observe was that each country were put in the same cluster until the algorithm were stopped. This can be due to merging more trees results in a tree generalizing better. So in every step, the algorithm is putting the countries in the biggest cluster of countries, where the tree is good for capturing more general decision rules.

In order to solve this issue, different depths were tried out. The other method used was to start from a state where pairs or 3-sized groups are clustered together. However, none of them solved the issue, in all cases a big cluster were produced.

Another issue with this method is that it is computationally heavy since in each step with every cluster the new cluster's distance should be computed.

On the other hand, this method can be still fine-tuned, but it would take a lot more time due to the previously mentioned computational heaviness.

7 Clustering countries based on migration destination choice

7.1 Introduction

In this section, the methods used for clustering countries based on the individuals' destination choice will be described. When choosing the destination country it is reasonable to use the characteristics of the destination country. Here, the dataset is used introduced in Section 4.

7.2 Pipeline

7.2.1 Data preparation

The data includes each country's main characteristics for each year, such as GDP, or whether the English language is used as the main language. These pieces of information weren't part of the GWP dataset.

The characteristics are extended with a popularity score, which is computed from the GWP dataset. For each country, for each year the popularity score is the percentage of respondents indicating that country as their preferred destination. (As the numbers are pretty small, this popularity score is multiplied by 1000000, but this doesn't affect the results.)

The dataset is also extended with aggregated features from the GWP dataset. Each feature is aggregated by mean or median depending on its type. We assumed here the mean or the median is representative of each country, so the responses are sampled correctly.

7.2.2 Method

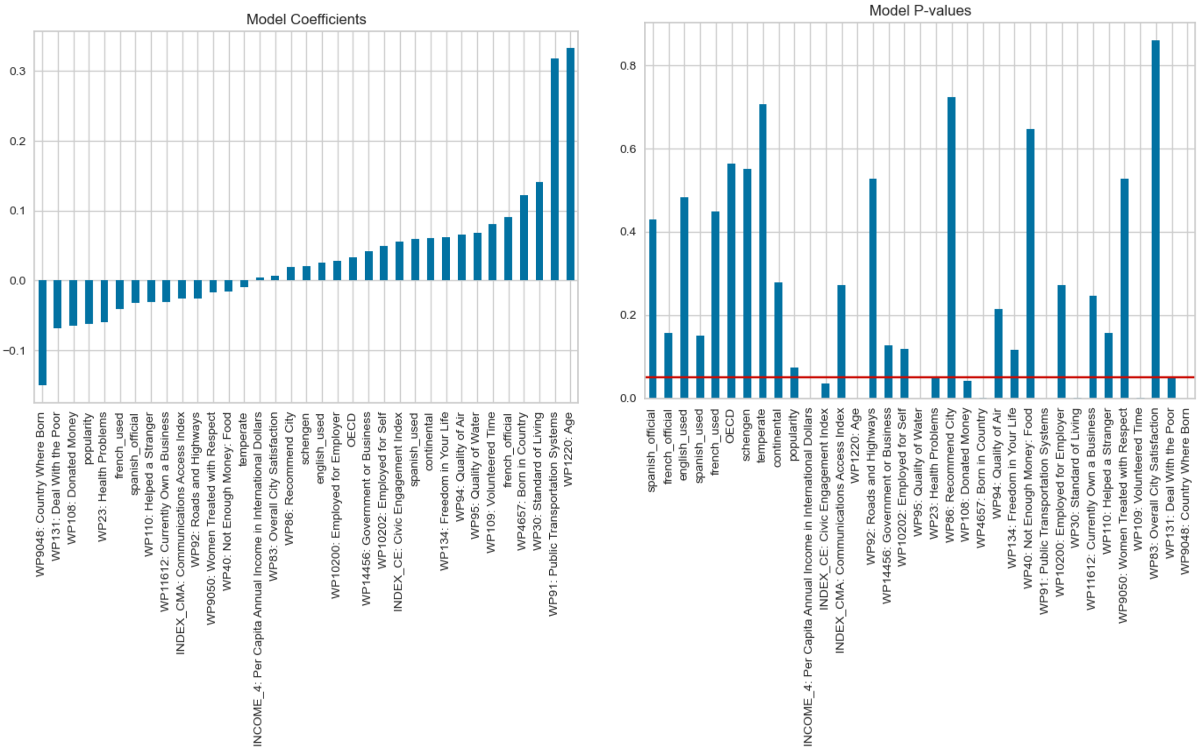
The first step in this method is predicting the popularity score from the other characteristics of the countries by linear regression. The coefficients of the regressions represent how much a characteristic influences the popularity number. As there are a lot of features, recursive feature elimination (RFE) is used with 35 resulting features. The resulting coefficient and the p-values are in the plots below 9. Unfortunately, there are still a lot of features that are not significant. Including the cross-products of features can solve this problem, but it is not included in the scope of this project.

After getting these coefficients, the original data is weighted by these values. As in the original table, the values are between 0 and 1, there is no need for normalization.

By weighting the original values, the countries in each year are represented in a multidimensional space, where the clustering is done with agglomeration clustering. The optimal number of clusters is 5, determined by the elbow method.

7.3 Results

The linear regression results yielded few statistically significant features that were found to be related to destination choice. These features include: age, annual income, civic engagement



(a) Model coefficients

(b) P-values

Figure 9: Results of the linear regression

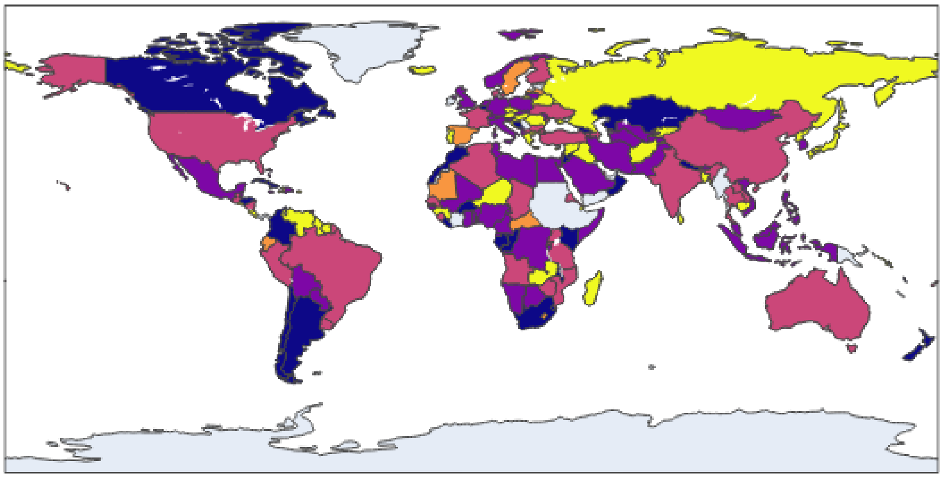


Figure 10: Result of the method with linear regression.

index, quality of water, donated money, whether the individual was born in the country, the availability of public transportation, the standard of living, and the amount of time that the individual volunteered.

Despite the limited number of statistically significant features, it makes sense that these particular factors would be influential in the decision-making of potential migrants. Age, annual income, and the standard of living are all factors that may impact an individual's ability to relocate to a different country. The civic engagement index and the quality of water are also relevant, as they may indicate the overall stability and well-being of the country of origin, which could influence an individual's destination choice. The amount of donated money and time volunteered also would likely be positively related to socio-economic factors. And being born in the country is also a factor that may influence whether or not an individual feels a strong connection to their country of origin and is less likely to want to leave.

Unfortunately, it seems that the results of the clustering are uninterpretable. The clusters that were generated appear to be random and do not provide any clear insight clustering of the data. This may be caused by several issues, such as the choice of the clustering algorithm, the similarity measure used, or the amount of data used. Also, the data itself may not be suitable for clustering, or there may be other unknown factors at play.

To address this issue, it may be beneficial to conduct further investigation in order to understand the underlying cause of the uninterpretable results. This could include trying different clustering algorithms, experimenting with different similarity measures, increasing the amount of data used, or possibly even collecting additional data to supplement the existing set.

8 Conclusion

In the process of clustering countries based on various criteria, a multitude of different methods was attempted. The results of these attempts are promising, indicating that the methods employed are able to cluster countries in meaningful ways. However, determining the best clustering is difficult without testing them in a predictive model. Additionally, interpreting the results of these clustering methods can be challenging.

Despite these challenges, we observed that the important features used in the clustering process made sense, and the clusterings that were produced also appeared to be logical. One key factor that was considered in the clustering process was spatial dependency, as countries that are geographically close to one another may have similar characteristics that make them suitable to be clustered together.

However, it was also noted that there are inconsistencies in the results when using different random seeds for the clustering process. This highlights the need for further investigation and refinement of the methods used.

Moving forward, there are several areas of exploration that may help to further enhance the value of these clusterings. Firstly, it would be beneficial to test the clusterings in a predictive model to better understand their usefulness. Secondly, the refinement of the clustering countries of origin based on the destination choice of their residents would also give us more insights into migration on a global scale. Additionally, it may be informative to consider geographical regions separately, such as America and Europe, to gain a more detailed understanding of the clusterings. Another important aspect to investigate is random seed consistency, as resolving inconsistencies in results could greatly improve the accuracy of the clustering methods. Lastly, it could also be beneficial to explore other clustering methods and fine-tune the hierarchical clustering method that was used in this process.

In conclusion, while the results of clustering countries based on various criteria were promising, there is still much work to be done in order to fully understand and utilize these clusterings. Further testing and refinement of the methods used will be important in determining the true value of these clusterings, and in finding ways to utilize them in a meaningful and effective manner.

9 Appendix

9.1 Used features in the migration aspiration

Code	Short name
WP5	Country
EMP_FTEMP	Employed Full Time for an Employer Index (workforce)
EMP_FTEMP_POP	Payroll to Population Index (P2P)
INCOME_4	Per Capita Annual Income in International Dollars
INDEX_CB	Community Basics Index
INDEX_CE	Civic Engagement Index
INDEX_CMA	Communications Access Index
INDEX_CMU	Communications Use Index
INDEX_FL	Financial Life Index
INDEX_FS	Food and Shelter Index
INDEX_JC	Job Climate Index
INDEX_LE	Life Evaluation Index
INDEX_OT	Optimism Index
INDEX_SL	Social Life Index
INDEX_ST	Struggling Index
INDEX_SU	Suffering Index
INDEX_TH	Thriving Index
WP10200	Employed for Employer
WP10202	Employed for Self
WP10248	Opportunities to Make Friends
WP108	Donated Money
WP109	Volunteered Time
WP110	Helped a Stranger
WP11612	Currently Own a Business
WP12	Residents 15+ in Household
WP1219	Gender
WP1220	Age
WP1223	Marital Status
WP1230	Children Under 15
WP128	Work Hard, Get Ahead
WP129	Children Respected
WP130	Children Learn and Grow
WP131	Deal With the Poor
WP132	Preserve the Environment

WP1325	Move Permanently to Another Country
WP134	Freedom in Your Life
WP14456	Government or Business
WP15862	Used the Internet in Past Seven Days
WP16	Life Today
WP16056	Access to the Internet
WP17626	Mobile Phone for Personal Calls
WP18	Life in Five Years
WP23	Health Problems
WP2319	Feelings About Household Income
WP27	Count On to Help
WP30	Standard of Living
WP31	Standard of Living Better
WP40	Not Enough Money: Food
WP43	Not Enough Money: Shelter
WP4657	Born in Country
WP83	Overall City Satisfaction
WP86	Recommend City
WP88	City Economy Getting Better
WP89	Local Job Market
WP9042	Move to Country in Last 5 Years
WP9048	Country Where Born
WP9050	Women Treated with Respect
WP91	Public Transportation Systems
WP92	Roads and Highways
WP94	Quality of Air
WP95	Quality of Water
WP98	City: Good, Affordable Housing

9.2 Result: decision tree for D1 and D2 with regions

```
def predict():
    if 'WP1220: Age' <= '43.5':
        if 'WP9048: Country Where Born_True' <= '0.5':
            if 'WP95: Quality of Water_4' <= '0.5':
                return '[[ 50958. 162924.]]'
            else:  # if 'WP95: Quality of Water_4' > '0.5'
                return '[[32128. 50854.]]'
        else:  # if 'WP9048: Country Where Born_True' > '0.5'
            if 'WP95: Quality of Water_4' <= '0.5':
                return '[[22216. 29888.]]'
            else:  # if 'WP95: Quality of Water_4' > '0.5'
                return '[[27241. 21379.]]'
    else:  # if 'WP1220: Age' > '43.5'
        if 'WP9048: Country Where Born_True' <= '0.5':
            if 'WP1220: Age' <= '60.5':
                return '[[ 19958. 109417.]]'
            else:  # if 'WP1220: Age' > '60.5'
                return '[[ 7590. 97107.]]'
        else:  # if 'WP9048: Country Where Born_True' > '0.5'
            if 'WP1220: Age' <= '59.5':
                return '[[ 9727. 18565.]]'
            else:  # if 'WP1220: Age' > '59.5'
                return '[[ 4021. 15002.]]'
```

9.3 Result: D1

======= cluster: LLL =======

Saudi Arabia Jordan Pakistan Indonesia Bangladesh India Nigeria Kenya Tanzania Ghana Uganda Benin Malawi Philippines Vietnam Cambodia Laos Mali Mauritania Niger Rwanda Senegal Zambia South Korea Kyrgyzstan Cameroon Sierra Leone Zimbabwe Costa Rica Andorra Bahamas Barbados Chile Guinea-Bissau Jamaica Latvia Lichtenstein Maldives Malta Nicaragua North Korea St. Kitts Nevis Switzerland The Gambia Togo Tuvalu Uruguay Vanuatu Kosovo Nagorno-Karabakh Region

======= cluster: LLR =======

Palestinian Territories Congo (Kinshasa) Guinea Ireland

======= cluster: LRR =======

South Africa Botswana Mozambique Belarus Chad Congo Brazzaville Croatia Grenada Honduras Libya Nauru Tonga Macau

======= cluster: RRR =======

United States Egypt Morocco Lebanon Turkey United Kingdom France Germany Netherlands Belgium Spain Italy Poland Hungary Czech Republic Romania Sweden Greece Denmark Iran Hong Kong Singapore Japan China Venezuela Brazil Mexico Israel Madagascar Canada Australia Sri Lanka Thailand Myanmar New Zealand Taiwan Afghanistan Georgia Kazakhstan Moldova Russia Ukraine Burkina Faso Albania Algeria Armenia Austria Azerbaijan Bosnia and Herzegovina Brunei Burundi Colombia Comoros Cuba Djibouti Ecuador El Salvador Equatorial Guinea Fiji Gabon Iceland Iraq Island Nations (11) Kiribati Lesotho Liberia Luxembourg North Macedonia Malaysia Marshall Islands Micronesia Montenegro Namibia Oman Paraguay Peru Portugal Puerto Rico Seychelles Slovenia Solomon Islands Tajikistan Tunisia Turkmenistan Uzbekistan Somaliland region

9.4 Result: D2

======= cluster: RRRLLL =======

United States Egypt Morocco United Kingdom France Germany Netherlands Belgium Spain Italy Poland Hungary Czech Republic Romania Sweden Greece Denmark Iran Hong Kong Singapore Japan China Mexico Israel Madagascar Canada Australia Sri Lanka Thailand Myanmar New Zealand Taiwan Afghanistan Georgia Kazakhstan Moldova Russia Ukraine Burkina Faso Albania Austria Azerbaijan Brunei Burundi Colombia Cuba Djibouti El Salvador Fiji Gabon Iraq Island Nations (11) Lesotho Luxembourg North Macedonia Malaysia Marshall Islands Micronesia Montenegro Namibia Oman Paraguay Puerto Rico Seychelles Slovenia Solomon Islands Tajikistan Uzbekistan Somaliland region

======= cluster: LLLRRR =======

Saudi Arabia Jordan Pakistan Indonesia Bangladesh India Tanzania Philippines Vietnam Cambodia Laos Niger Rwanda Senegal Kyrgyzstan Andorra Bahamas Barbados Latvia Maldives Malta Nicaragua North Korea St. Kitts Nevis Switzerland The Gambia Togo Tuvalu Uruguay Vanuatu Nagorno-Karabakh Region

======= cluster: LLLLRR =======

Nigeria Kenya Ghana Uganda Benin Malawi Mali Mauritania Zambia South Korea Cameroon Sierra Leone Zimbabwe Costa Rica Chile Guinea-Bissau Jamaica Lichtenstein Kosovo

======= cluster: RRRLLR =======

Lebanon Turkey Venezuela Brazil Armenia Bosnia and Herzegovina Comoros Ecuador Equatorial Guinea Iceland Kiribati Peru Portugal Tunisia

======= cluster: LRRLLL =======

South Africa Mozambique Belarus Congo Brazzaville Grenada Libya Nauru Tonga Macau

======= cluster: LLRRRR =======

Palestinian Territories Congo (Kinshasa) Guinea

======= cluster: LRRRRR =======

Botswana Croatia Honduras

======= cluster: RRRLRR =======

Algeria Liberia Turkmenistan

======= cluster: LRRLLR =======

Chad

======= cluster: LLRLRR ======= Ireland

9.5 Result: D3

======= cluster: 0 =======

Lebanon Jordan Italy Hungary Romania Greece Iran Venezuela Brazil Mexico Israel South Africa Kazakhstan Russia Ukraine Burkina Faso Algeria Armenia Austria Bosnia and Herzegovina Brunei Colombia Comoros Cuba Ecuador El Salvador Equatorial Guinea Iceland Ireland Malaysia Montenegro Namibia North Korea Paraguay Portugal Seychelles Turkmenistan Uzbekistan Somaliland region

======= cluster: 1 =======

Malawi Mauritania

======= cluster: -1 =======

United States Egypt Morocco Saudi Arabia Turkey Pakistan Indonesia Bangladesh United Kingdom France Germany Netherlands Belgium Spain Poland Czech Republic Sweden Denmark Hong Kong Singapore Japan China India Nigeria Kenya Tanzania Palestinian Territories Ghana Uganda Benin Madagascar Canada Australia Philippines Sri Lanka Vietnam Thailand Cambodia Laos Myanmar New Zealand Botswana Mali Mozambique Niger Rwanda Senegal Zambia South Korea Taiwan Afghanistan Belarus Georgia Kyrgyzstan Moldova Cameroon Sierra Leone Zimbabwe Costa Rica Albania Andorra Azerbaijan Bahamas Barbados Burundi Chad Chile Congo (Kinshasa) Congo Brazzaville Croatia Djibouti Fiji Gabon Grenada Guinea Guinea-Bissau Honduras Iraq Island Nations (11) Jamaica Kiribati Latvia Lesotho Liberia Libya Lichtenstein Luxembourg North Macedonia Maldives Malta Marshall Islands Micronesia Nauru Nicaragua Oman Peru Puerto Rico Slovenia Solomon Islands St. Kitts Nevis Switzerland Tajikistan The Gambia Togo Tonga Tunisia Tuvalu Uruguay Vanuatu Kosovo Nagorno-Karabakh Region Macau

9.6 Result: D4, K=8, depth=8

======= cluster: 0 =======

Egypt Morocco Lebanon Jordan Turkey Iran China Venezuela Brazil Mexico Palestinian Territories South Africa Philippines Sri Lanka Vietnam Thailand Myanmar Algeria Andorra Bahamas Bosnia and Herzegovina Colombia Comoros Ecuador El Salvador Equatorial Guinea Iceland Ireland Kiribati Lichtenstein Malta Micronesia Montenegro North Korea Paraguay Peru Portugal Tunisia Turkmenistan Somaliland region

======= cluster: 1 =======

Pakistan Nigeria Kenya Tanzania Ghana Uganda Benin Madagascar Malawi Botswana Mali Mauritania Mozambique Niger Rwanda Zambia South Korea Belarus Cameroon Sierra Leone Zimbabwe Costa Rica Chad Chile Congo (Kinshasa) Congo Brazzaville Croatia Grenada Guinea Guinea-Bissau Honduras Jamaica Liberia Libya Nauru Nicaragua Switzerland Togo Tonga Kosovo Macau

======= cluster: 2 =======

France Germany Netherlands Czech Republic Marshall Islands Puerto Rico Slovenia Tajikistan

======= cluster: 3 =======

Sweden Denmark Gabon Iraq Oman

======= cluster: 4 =======

Saudi Arabia Uruguay

======= cluster: 5 =======

United States United Kingdom Belgium Spain Hong Kong Singapore Israel Canada Australia New Zealand Afghanistan Kyrgyzstan Azerbaijan Barbados Fiji Island Nations (11) Latvia North Macedonia Solomon Islands Vanuatu

======= cluster: 6 =======

Italy Poland Hungary Romania Greece Japan Taiwan Georgia Kazakhstan Moldova Russia Ukraine Burkina Faso Albania Armenia Austria Brunei Burundi Cuba Djibouti Lesotho Luxembourg Malaysia Maldives Namibia Seychelles Uzbekistan Nagorno-Karabakh Region

======= cluster: 7 =======

Indonesia Bangladesh India Cambodia Laos Senegal St. Kitts Nevis The Gambia Tuvalu

9.7 Result: D5, K=6, depth=8

======= cluster: 0 =======

Saudi Arabia Singapore China Barbados Latvia Maldives Uruguay Vanuatu Nagorno-Karabakh Region

======= cluster: 1 =======

Egypt Morocco Pakistan Indonesia Bangladesh India Nigeria Kenya Tanzania Ghana Uganda Benin Madagascar Malawi Philippines Sri Lanka Thailand Cambodia Myanmar Botswana Mali Mauritania Mozambique Niger Rwanda Senegal Zambia South Korea Belarus Cameroon Sierra Leone Zimbabwe Costa Rica Bahamas Chad Chile Congo (Kinshasa) Congo Brazzaville Croatia Grenada Guinea Guinea-Bissau Honduras Ireland Jamaica Liberia Libya Lichtenstein Malta Nauru Nicaragua North Korea Switzerland Togo Tonga Tunisia Kosovo Macau

======= cluster: 2 =======

Jordan Turkey Greece Iran Venezuela Brazil Mexico Palestinian Territories South Africa Kazakhstan Russia Burkina Faso Algeria Andorra Armenia Austria Bosnia and Herzegovina Brunei Burundi Colombia Comoros Ecuador El Salvador Equatorial Guinea Iceland Kiribati Malaysia Micronesia Montenegro Namibia Paraguay Peru Portugal Seychelles Turkmenistan Somaliland region

======= cluster: 3 =======

United States Lebanon United Kingdom France Germany Belgium Spain Italy Poland Hungary Czech Republic Romania Hong Kong Japan Israel Canada Australia Vietnam New Zealand Taiwan Afghanistan Georgia Kyrgyzstan Moldova Ukraine Albania Azerbaijan Cuba Djibouti Fiji Island Nations (11) Lesotho Luxembourg North Macedonia Marshall Islands Puerto Rico Slovenia Solomon Islands Tajikistan Uzbekistan

======= cluster: 4 =======

Laos St. Kitts Nevis The Gambia Tuvalu

======= cluster: 5 =======

Netherlands Sweden Denmark Gabon Iraq Oman

9.8 Result: linear regression method

======= cluster: 0 =======

NPL MNE BEL MWI AND JOR ARG KAZ COL VUT COG ZAF SWZ STP CHL GAB OMN LBN MAR GEO LBR KEN BFA CUB HND NZL BIH MCO CAN

======= cluster: 1 =======

CHE TJK EGY PAK SOM COD SAU LTU GBR COM TUN ISR LBY ALB POL BRB MEX TUV KNA BEN NRU LUX MLI JAM GHA TTO TON SVK BWA BTN CMR DEU NLD BOL IDN BHS DOM PHL VNM NOR GNQ MYS KOR MDA IRN BDI UZB NGA PRY NAM TGO SVN MNG BLZ GNB ITA TKM

======= cluster: 2 =======

GTM URY THA NIC GMB WSM BGR DZA MLT TUR REU FRA PSE ERI ZWE QAT FSM UGA SEN TCD IND RWA PER TWN FIN FJI UKR CHN USA TZA AZE ARM SGP MKD AGO DNK LAO LIE AUS SLE BRA MOZ

======= cluster: 3 =======

ECU CAF SWE GRD MRT LSO ESP LVA

======= cluster: 4 =======

BGD HUN PRK HKG CZE GIN ATG HTI NER LKA SLV KHM IRQ BHR ZMB GRC MHL DJI ISL ROU AUT AFG MDV SLB JPN KIR BLR LCA GUY HRV SYC VEN BRN MDG CRI SUR PLW CYP KGZ PRT RUS MAC PRI DMA SYR

References

[1] Stefania Lovo. Potential migration and subjective well-being in europe. *IZA Journal of Migration*, 3(1):1–18, 2014.

[2] Flore Gubert and Jean-Noël Senne. Is the european union attractive for potential migrants?: An investigation of migration intentions across the world. 2016.

[3] Simone Bertoli and Ilse Ruyssen. Networks and migrants' intended destination. *Journal of Economic Geography*, 18(4):705–728, 2018.

[4] Frédéric Docquier, Aysit Tansel, and Riccardo Turati. Do emigrants self-select along cultural traits? evidence from the mena countries. *International Migration Review*, 54(2):388–422, 2020.

[5] Els Bekaert, Ilse Ruyssen, and Sara Salomone. Domestic and international migration intentions in response to environmental stress: A global cross-country analysis. *Journal of Demographic Economics*, 87(3):383–436, 2021.

[6] Michel AR Beine, Michel Bierlaire, and Frédéric Docquier. New york, abu dhabi, london or stay at home? using a cross-nested logit model to identify complex substitution patterns in migration. *Using a Cross-Nested Logit Model to Identify Complex Substitution Patterns in Migration*, 2021.

[7] Monica Langella and Alan Manning. Income and the desire to migrate. 2021.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.