"Data Analysis on Human Trafficking" Methodology Report

Capstone Course, University of Zurich
in collaboration with
International Organization for Migration

Albiez, Janine
Aus der Au, Patrik
Gibilisco, Saro
Woeffray, Théoda
Yin, Jenny

Zurich, 31.12.18

Content

List of Figures

List of Tables

1 Introduction

Trafficking in persons is a pervasive and destructive phenomenon in our global society often underestimated due to its predominant clandestine nature. Every now and then, its omnipresence becomes visible through scandals in the sex industry, on prominent construction sites or through migrant smuggling along current migration streams. This phenomenon of human trafficking as we know it today bears enormous human costs. It is a grave offence against the human dignity and a violation of human rights. Victims are men, women, children in their country of origin or abroad in almost every country worldwide. They are trafficked for the purpose of exploitation, taking on different forms like sexual exploitation, forced labour, slavery and others (UNODC, 2008). According to the International Labour Organisation (ILO) forced labour generates USD 150.2 billion a year. Human trafficking is one of the most lucrative and fastest growing forms of international crime (FAFT & APG, 2018).

In a global effort the United Nations Office on Drugs and Crime (UNODC) criminalized human trafficking, encouraging national legislations to follow. Such national adoption of a dynamic and flexible definition of human trafficking is key as victim characteristics and organizational structures of trafficking in person are changing (UNODC, 2009). Given that human trafficking is a transnational challenge, UNODC inspired definitions contribute to consistency and consensus in global counter-trafficking efforts (UNODC, 2006).

The 2010 United Nations Global Plan of Action to Combat Trafficking in Person established the "3P" paradigm of prosecution, protection, and prevention. It is used by states, international organizations and non-governmental organizations in their collective counter-trafficking efforts (UNGA, 2010).

Prevention on this issue can be versatile. It can target potential victims through awareness campaigns informing about the risk of trafficking. Efforts can be made through promoting safe alternatives to circumvent the traffickers' opportunities. Or prevention might inform consumers to target the demand side of services and goods whose supply involve trafficking and exploitation of persons (IOM, 2017a). For targeted and effective prevention, the vulnerabilities of individuals have to be addressed before they are exploited. To timely identify victims and potential victims during different phases of the trafficking process a more comprehensive and clearer understanding of their characteristics is crucial (Aronowitz, 2009; Galos et al., 2017). Prevalent victim characteristics in current literature on human trafficking mainly rely on categories like age, gender, country of origin and destination (UNODC, 2006; UNODC, 2016). At this point our work aims to contribute to research by further analysing victim characteristics.

In general, counter-trafficking efforts are still limited through an information deficit about the extent, the nature and new trends of this tragedy (UNODC, 2006). To breach this knowledge gap the coordinated efforts of governmental and non-governmental partners to collect better and more standardized data is important (IOM, 2017a). A more solid data foundation for prevention and protection approaches would not only assure practically more effective but also politically more feasible solutions (IOM, 2018). IOM's efforts to coordinate the collection of individual-level victim data among different non-governmental organizations opens further opportunities to understand and fight trafficking in persons better.

Our analysis of the data from the IOM Global Database on Victims of Human Trafficking aims to draw a more nuanced picture of the characteristics of trafficking victims, foremost, to enable the identification of potential victims and to improve prevention and protection activities. With choosing an explorative approach forgoing previous assumptions, we hope to detect the relevant characteristics of victims and their entrance into the trafficking process less influenced by existing knowledge. For this, we used machine learning techniques such as clustering and classification

tree analysis. Our main findings confirmed some conventional assumptions and detected yet not considered characteristics of trafficking victims.

Our report is structured to first outline our findings with which we aim to add to the overcoming of the information gap in counter-trafficking. Then, we introduce the data and the methodology covering the application of the new analysis techniques. At last, we derived some implications for the improvement of data collection from experiences with the data.

2 Terminology

This section clarifies the concepts, which were most relevant during our work. First, the definition of the concept of trafficking in persons is provided. Second, the section is split into two subsections. Section 2.1 explains the most relevant terms regarding entry into process. Section 2.2 defines the two most frequent types of exploitation, i.e. sexual exploitation and forced labour.

Trafficking in persons

The term trafficking in persons is defined as: “the act of recruiting, transporting, transferring, harbouring or receiving of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation.”

This definition is based on Art. 3(a) of the UN Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, supplementing the UN Convention against Transnational Organized Crime (United Nations 2000). It is also known as the Palermo Protocol.

Entry into process

Victims entering the process of human trafficking are either recruited, kidnapped or sold. Recruitment describes the act of getting in contact with a recruiter and being persuaded to accept an offer, usually a job (UNODC 2006, 59). In other words, the entry into the process of human trafficking is not forced, but a person is deceived about the real circumstances of the offer he or she has accepted. In contrast to recruitment, an entry into process by kidnapping or selling means that the person did not give his or her consent. Kidnapping means unlawfully detaining a person or persons against their will (including through the use of force; threat; fraud or enticement) for the purpose of demanding for their liberation an illicit gain or any other economic gain or other material benefit; or in order to oblige someone to do or not to do something. “Kidnapping” excludes disputes over child custody (UNODC, 2008). Selling mainly affects children (UNODC 2016, 8). Often, poor families sell their children to enable them a more prosperous life abroad (Dottridge 2002, 39). In the view of those families, they do not actually sell their children. Nevertheless, they are either promised to receive money in return of sending their children away or are paid directly. Due to that, the act is considered as “selling”. Families who sell their children are often convinced by agents to do so. Hence, lines between recruitment and selling are blurred. The three forms of entry into process are not mutually exclusive.

Forms of exploitation

Forced labour and sexual exploitation are by far the most frequent forms of exploitation. A victim of forced labour is a person in any form of work or service, which is demanded from him or her under the threat of penalty and for which the person has not offered him- or herself voluntarily (ILO, 1930). The term forced labour is further divided into forced labour exploitation and forced sexual exploitation. While forced sexual exploitation refers to forced labour and services involving commercial sex, forced labour exploitation denominates all other forms of forced labour (ILO,

2017). However, usually the term forced labour is used to describe forced labour exploitation and forced sexual exploitation is simply called sexual exploitation (UNODC, 2016; IOM, 2017b). The same accounts for the data we received from IOM. Hence, the terms forced labour and sexual exploitation are used in the remaining report to distinguish the two forms of exploitation.

Machine learning

Machine Learning is the practice of using algorithms to semi-automatically learn from data. There are two main types of machine learning algorithms, which differ primarily in the type of task that they are intended to solve. In unsupervised learning, the task is to uncover functions, groupings, and patterns in the data, and there is no predefined outcome. In supervised learning, there is an outcome we seek to predict, given the input information we provide. The goal is to formulate a function that best predicts a pre-selected outcome.

3 Overview of the Main Findings

3.1 Descriptive Statistics

The individual-level data provided by IOM encompasses over 52,000 identified victims of human trafficking between 2002 and 2017. Figure 1 visualises the data by region of exploitation and region of citizenship. Regions are defined according to the structure of IOM regional offices. First, the number of identified victims differs across regions. Second, the vast majority of the identified victims in the dataset are exploited in the same region as their region of origin. On the one hand, the large variation in the number of identified victims may be due to differences in data collection across regions. On the other hand, victims' willingness to account their story might be influenced by cultural norms, levels of trust and privacy concerns, as well as traumatic experiences (Aronowitz, 2009; Galos et al., 2017). Using machine learning techniques, we searched for patterns in the data. A general typology of the victims of human trafficking was established through cluster analysis. Classification tree analysis was applied to find regional patterns.

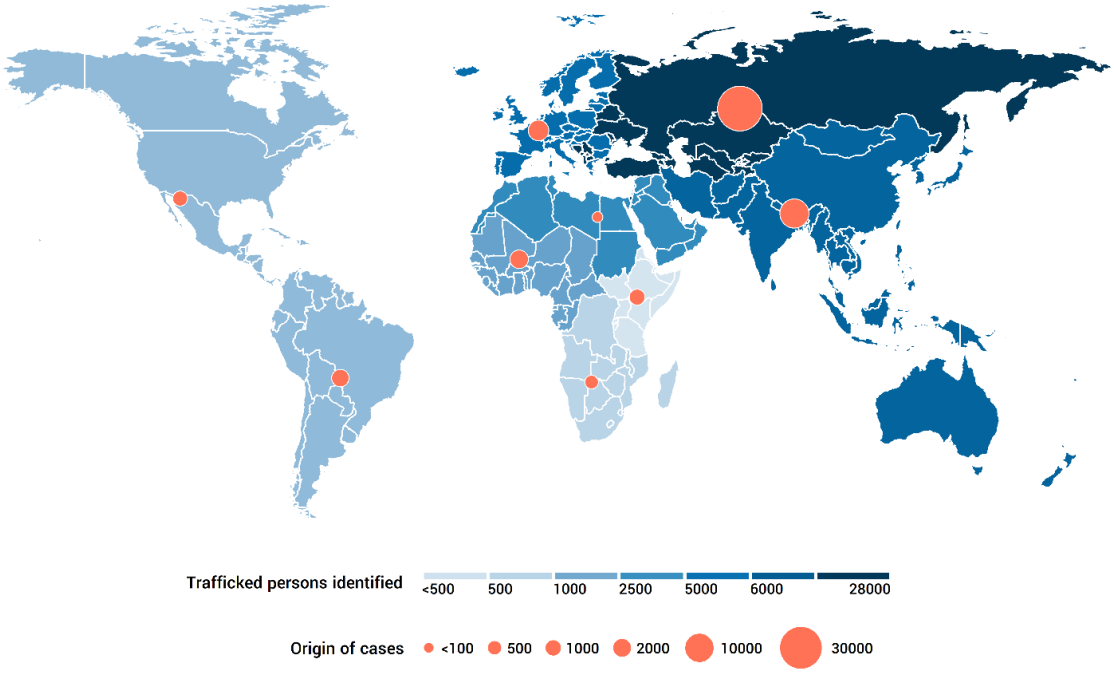


Figure 1. Identified victims of human trafficking by region of citizenship and region of exploitation. Regions are defined according to IOM regional offices.

3.2 General Typology of the Victims of Human Trafficking

Via cluster analysis, we built three subgroups of the data. The first subgroup represents 60% of the data. It contains young adults from South-Eastern Europe and Central Asia or Asia and the Pacific. They are exploited through forced labour, either in their region of origin or in the Middle East and North Africa (MENA). The second subgroup represents identified victims of sexual exploitation in Europe and Central Asia. Most of the victims are young, female and single. The last subgroup is characterized by the transregional exploitation of young women. The majority was exploited in the MENA region or in the European Union and the European Economic Area (EU & EEA). The exploitation type is not a defining characteristic of this subgroup. Figures 2 and 3 display the clusters described.

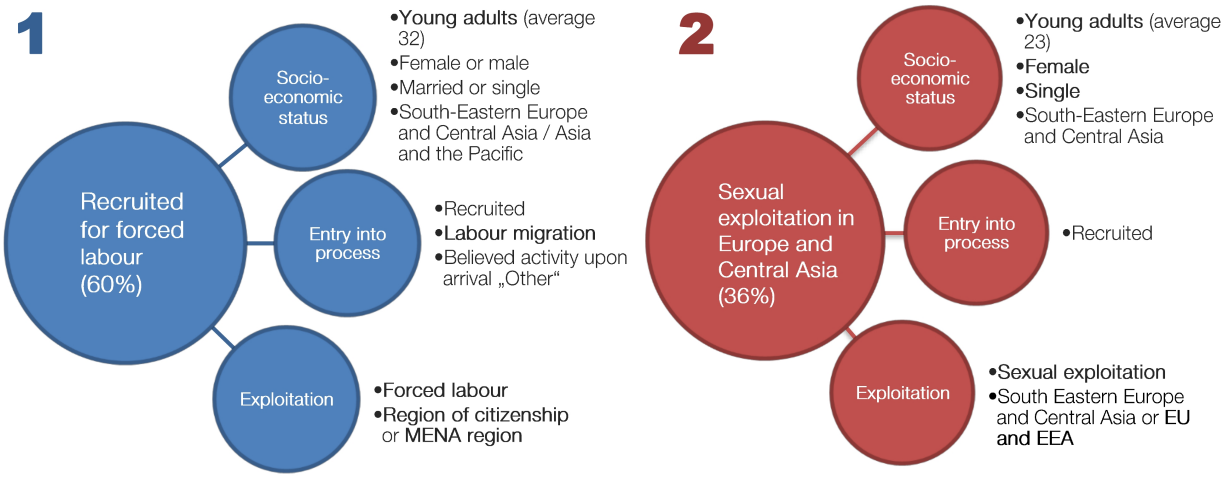


Figure 2. Prototypical characteristics of identified victims of the first (1) and the second (2) subgroup.

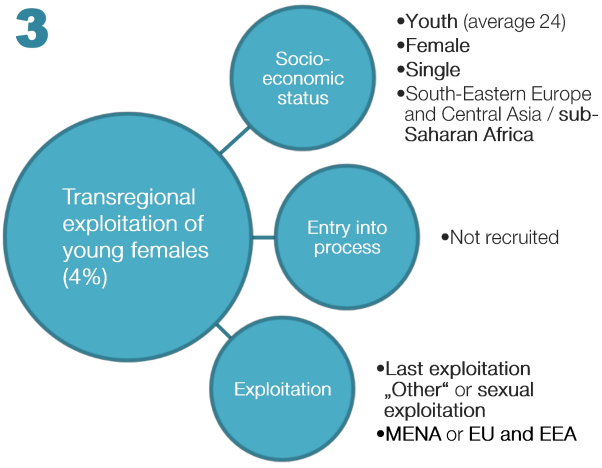


Figure 3. Victim characteristics of the third subgroup.

3.3 Regional Patterns in Determining the Type of Exploitation

The aim of the classification tree analysis was to get a more detailed knowledge about the characteristics of the victims in our dataset. Classification trees allow to identify the input variables which best predict a certain outcome. Our focus was to determine the victim characteristics predicting the type of exploitation. The victims in our dataset were almost exclusively victims of either forced labour or sexual exploitation. Hence, type of exploitation was the outcome variable taking on two distinct values. Regarding the input variables, we considered socio-economic factors and information about the entry into the process of trafficking. Furthermore, separate trees were created for the following regions: South-Eastern Europe and Central Asia, Asia and the Pacific, and West and Central Africa. The three regions were chosen because they are the major regions of origin for most of the identified victims in our data. Classification trees are read top-down. The input variable appearing at the top of the tree is the most powerful to predict the outcome variable. Each input variable in the tree splits the data into two subgroups. Each split indicates the predicted probability of the subgroup and the percentage of the sample it represents. The predicted probability is a measure of how many individuals are expected to fall into the same outcome category, given that they share the same characteristics.

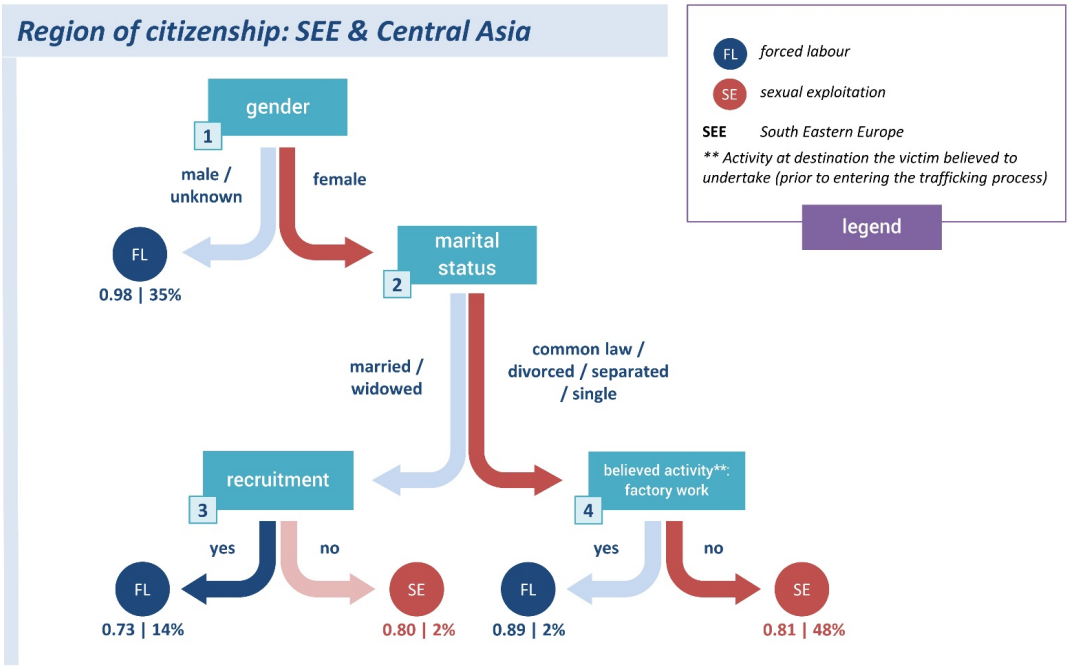


Figure 4. Classification tree predicting type of exploitation for citizens of SEE & CENA.

Figure 4 displays the classification tree for all identified victims whose origin is South-Eastern Europe and Central Asia (SEE & CENA). The tree includes over half of the identified victims in the dataset. The most relevant characteristic determining the type of exploitation is gender. Identified victims whose gender is male or unknown have a predicted probability of 98% to become victims of forced labour (FL) rather than sexual exploitation (SE). 35% percent of the sample from SEE & CENA fall into this subgroup. If gender is female, the data is split into further subgroups. The most important characteristic for female victims is their marital status. Women who are married or widowed and are recruited have a probability of 73% to be subjected to forced labour. The subgroup represents 14% of the sample. In contrast, women whose marital status is common law, divorced, separated, or single, and who did not believe that their activity upon arrival was going to be factory work, show a 81% predicted probability of being victims of sexual exploitation. 48% of the sample are classified into this group.

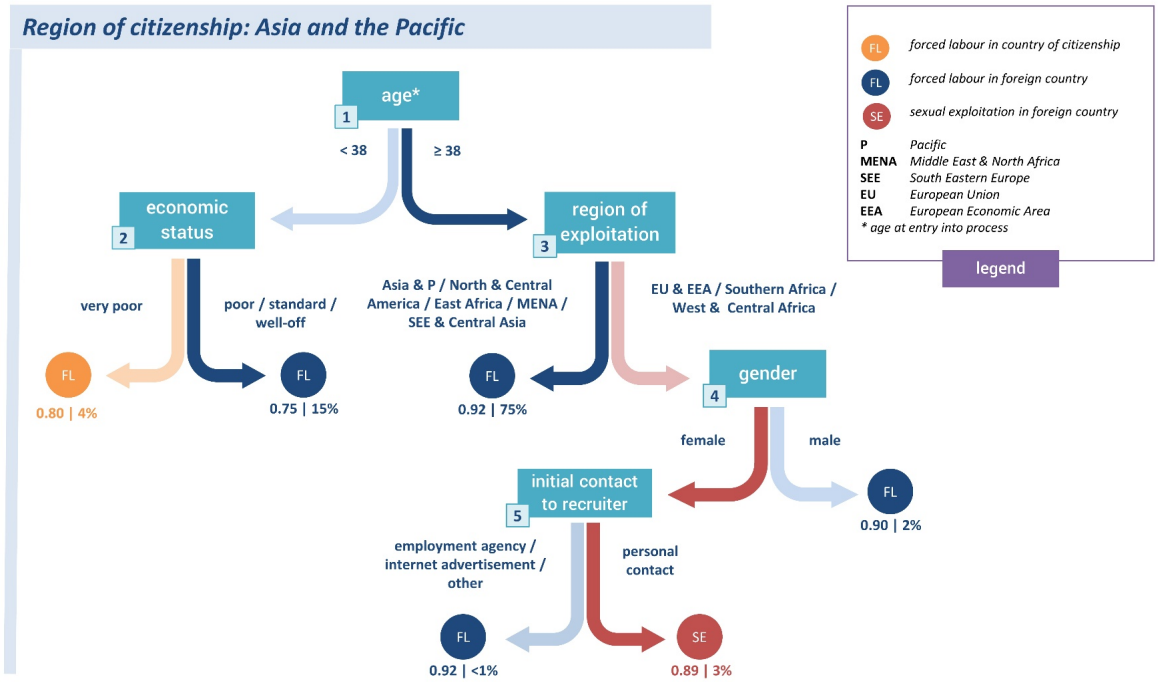


Figure 5. Classification tree predicting type of exploitation for citizens of Asia and the Pacific.

Regarding victims from Asia and the Pacific, most are exploited through forced labour, which is illustrated by Figure 5. Only women at age 38 or older, who are exploited in the EU & EEA, Southern Africa, or West and Central Africa and who are recruited by personal contact, are likely to become victims of sexual exploitation. Moreover, socio-economic status is relevant to determine whether exploitation occurs in a foreign country or in the country of citizenship.

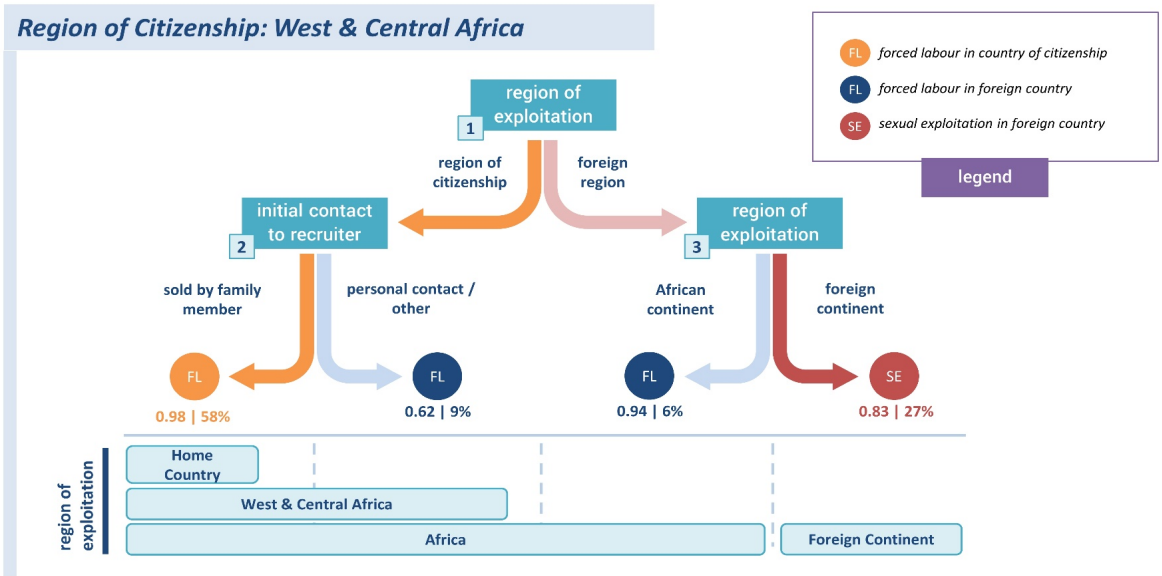


Figure 6. Classification tree predicting type of exploitation for citizens of West and Central Africa.

As visualised by Figure 6, the region of exploitation is particularly relevant for vicitms from West and Central Africa. Victims who are exploited in a foreign region but within the African continent are expected to become victims of forced labour. On the contrary, victims exploited outside the African continent rather become victims of sexual exploitation. For victims exploited within their region of citizenship, a crucial factor is how contact with the recruiter was initiated. Victims sold by family members are unlikely to be trafficked across the borders of their home country. However, victims recruited via personal contact or other methods tend to be exploited in a foreign country.

Having provided an overview of the main findings, we will explain in detail the data and the methodology in the following sections. A more in-depth description and discussion of our main findings can be found in “Analysing Human Trafficking – A Machine Learning Approach”.

4 Data Source

For our analysis we used a subset of the IOM Global Database on Victims of Human Trafficking. It is a database on primary data from identified victims of trafficking that were assisted by IOM since 2000. IOM case workers collect the data during the assistance interviews using IOM's case management tool, MiMOSA. The data from over 50'000 registered victims include quantitative and qualitative information on the individual trafficking experiences.

The structure of the dataset stems from the format of the IOM Victim of Trafficking questionnaires used for the interviews. Most of the questions are translated into single variables in the dataset. For questions, which allow for multiple answers (e.g. type of exploitation), a dummy variable for each answer category is created. Throughout the analysis we label question variables with a clear answer structure, i.e. a defined set of standardized answer categories, as structured variables. Open question and comment variables resulting in individual answers, i.e. values, are labelled as unstructured. The most extensive unstructured variable is the so-called "decision"-variable. It represents the question asking the case worker to justify why the individual should be identified as victim of trafficking. These answers can extend from a few words up to a few pages transcribing the victim's whole story.

However, it has to be considered that this dataset might not represent a reference population of all victims of trafficking in terms of baseline demographic features. The selection of assisted and interviewed individuals is influenced by the geographic assistance capacity like the location and size of the IOM country offices. Further, selection bias could stem from the victims' willingness to share their experience influenced by the sensitivity of traumatic experiences, different cultural norms, level of trust or privacy concerns (Aronowitz, 2009; Galos et al., 2017).

Such individual level data is very sensitive and requires a respectful and cautious handling. To safeguard the confidentiality of personal data and the anonymity of the interviewed victims we first of all anonymised the whole dataset. This procedure is explained in the following part.

4.1 Anonymization

The aim of this task was to anonymize the unstructured "decision"-variable (MIM_00931_S). The anonymization of the variable implies the following two steps: First, the named-entity recognition technique extracted names from the variable and classified them into two categories: names of persons and names of locations. Second, the names of persons and of locations in the unstructured variable were randomly replaced with the classified names of persons and of location, respectively. In doing so, all names in the text variable do not correspond to their 'real' names. We decide to not replace the classified names with encrypting names because in this case, the named-entity recognition technique does not guarantee the classification of all names in the unstructured variable.

4.2 Cleaning the Data

To provide access to good quality data for our analysis, we began with cleaning the dataset. Primarily we targeted to detect and remove instance-level[1] errors and inconsistencies to improve the accuracy and consistency of the data. The following list names the steps we took:

- **standardizing the labels of missing values:** We changed all the different labels used for missing values (e.g. -99, N/K, n.a., etc.) to one consistent label "NA" ensuring equal treatment throughout the variables.
- **cleaning value levels:** Value levels of a variable were fitted to the logic structure they are supposed to have. Values, which did not match any of the predefined answer categories (i.e. value levels), were removed (e.g. if binary "yes/no" variables, answers like "child care" or "15" were removed).
- **adjusting data type of variables:** Variables were converted either to factors, characters or numeric to assure correct treatment during the analysis.
- **removing of duplicates:** To gain clarity and prevent misleading statistics, duplicate variables and variables, which contain only missing values (all NA), were removed.
- **adjusting the language:** To get a solid foundation for the text analysis the language was adjusted by correcting for spelling mistakes with reference to British English. This was done by applying correction for common mistakes detected through a scanning of a random sample. The few French and Spanish text passages were manually translated or taken into account during the analysis when working with text patterns.

Through the cleaning of the dataset we improved the foundation for applying automated analysis methods. Machine learning methods as we used it rely on a consistently formatted dataset. In addition, the cleaning enabled us to get a better overview of the information contained in the dataset and to understand how the questionnaire was translated into variables. The following chapters present how we further improved the dataset to meet the method's requirements.

[1] Instance-level problems occur in actual data contents and are not visible at the structural level of the dataset. Such errors stem from operational mistakes during data entering; e.g. data entry errors like misspellings, duplicates, contradictory values, etc. (Rahm & Do, 2000).

5 Text Analysis and Structuring Information

The IOM Global Database on Victims of Human Trafficking bears great potential to support evidence based counter-trafficking efforts. Though, a closer inspection of the dataset revealed that parts of this potential cannot be tapped into. The data exists of structured and unstructured information. Since our advanced statistical methods depend on structured data we dedicate a part of our analysis to making that unstructured information accessible. Complementing the dataset's structured part with this newly accessible information reduces the number of missing values and therefore raises the predictive power of the concerned variables and adds to the accuracy of our model. The following chapters explain the different sources of unstructured information and the techniques used to make it accessible.

5.1 Open Answer Categories

One type of unstructured variables complementing structured information are the variables representing the "if other, please specify" open answer category. These answer options give the possibility to specify ones answer if it does not match any of the predefined answer categories. Therefore, the variable is a collection of individual answers. We used that unstructured information in two ways. First, reducing the degree of informational detail of such an individual answer (e.g. if stated "cook in a restaurant" we alter it in "restaurant") allowed us to assigning it to one of the existing categories (marked as A in Figure 7). Second, harmonizing and grouping the remaining answers (that could not be assigned to existing categories) enabled us to create new additional answer categories, i.e. variables (marked as B in Figure 7).

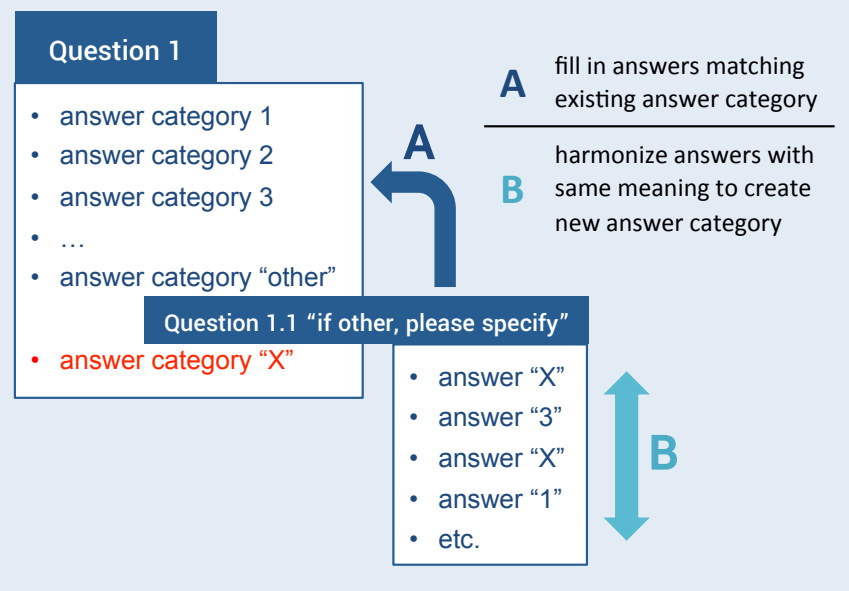


Figure 7. Assigning and Grouping individual answers from the "if other, please specify"-variable to existing or new answer categories making more informational content accessible.

The following two examples illustrate the two processes how the information is made accessible. The questionnaire contains the question "If labour migration, what activity did the individual believe he or she was going to be engaged in following arrival at the final destination?" allowing for multiple answers. Each predefined answer category translates into a dummy variable. Among the set of

answers there is the category "restaurant" representing individual activities in the gastronomic sector. In a case where for example the "if other, please specify" section is answered with "cook in a canteen" and the category "restaurant" is marked with a 0 (i.e. not applicable) we deal with an underrepresentation of this information. In a wider sense this individual answer could be understood as circumscribing an activity matching the category "restaurant". This simplification seems to make sense as the distinction between restaurant and canteen, as well as the additional information about the function at the restaurant, are not of great importance at this stage. Much more, we profit from the additional observations made available for the analysis.

Individual answers to this question, which could not be assigned to an existing category like in the example above, were further examined for categorical grouping. In the case of our example question on the believed activity upon arrival, we summarized individual answers like "pole dancer at a night club", "bar hostess", "dancer", "night club", "masseuse" etc. under the new category "entertainment" counting 742 observations. The criteria for creating a new category are mainly the mutual exclusiveness to existing categories and the comparable size of observations to the existing categories. For this harmonization, we created sets of words for each category, i.e. text patterns, after manually scanning all the individual answers and getting a rough impression of the answers wording and scope. In the "entertainment" example we justify this practice with the idea that it does not become clear from these answers if prostitution was expected, but these activities are still closely linked to the milieu where prostitution is organized.

We applied this procedure to seven questions of the dataset that included the "if other, please specify" answer option (Table 1). We justify this selection by the relevance of the questions for our analysis, as well as by the high number of complete individual answers included. Further, the unstructured variable has to bear much potential to make further information accessible and the filling-in needs to be feasible. The following table gives an impression of the informational gain for these seven questions.

Table 1. Information gained through assigning individual answers from the "if other, please specify"-variable to existing or new answer categories.

		Believed Activity upon Arrival	Last Activity prior Entry into Process	Activity at Destination	Entry into Process	Contact initiated to Recruiter	Movement Phase Travel With
original answers	structured	9944	2883	12192	17210	19623	6451
	unstructured	11742	3372	3234	697	388	1039
additional answers assigned	to existing categories	224	89	2461	56	151	79
	to new categories	7555	358	200	145	88	935
	info gained	78%	16%	22%	1%	1%	16%
	unstructured answers used	66%	13%	82%	29%	62%	98%
new categories		7[a]	4[b]	1[c]	1[d]	1[e]	2[f]

a Agricultural Work, Entertainment, Beauty, Domestic Work, Child Care, Any Work, Service Sector
b Service Sector, Beauty/Lifestyle, Self-Employment/Family Business, Entertainment
c Entertainment
d by Smuggling
e Street Advertisement
f Exploiter, Other Victims

The table shows clearly that there are questions were a lot of additional information could be gained and others where the informational content could hardly be improved. Further, among the questions with high informational potential in their individual answers, we could distinguish between the ones where the potential is driven by the possibility of filling in existing categories and others were the predefined answers are non-exhaustive and new categories can be created. For example, in the case of the question on the activity at destination 2461 additional answers could be filled in already existing answer categories (i.e. variables) compared to 200 answers that were gained by creating a new category. As a contrary, there is the question on the believed activity upon arrival where we gained 7555 answers by harmonizing the unstructured answers to 7 new categories. The sixth row (“unstructured answers used”) shows how much of the unstructured information was used. This varies widely as sometimes answers have to be neglected as they are not understandable, too general or represent a very small niche group for which it is not worth creating an own category.

5.2 "Decision"-Variable: Top Features

Along the internal harmonization and structuring of a question to make more information accessible, we also used an apporach structuring information accross questions. The largest unstructured variable with information covering multiple topics of the interview is the above-mentioned "decision"-variable. This chapter explains how this variable was used to further complement structured variables.

The content of the “decision”-variable had previously not been explored thoroughly; consequently, its content was at that point little known. To get a grasp of the information it contains, we generated word clouds, as seen in Figure 8 below:

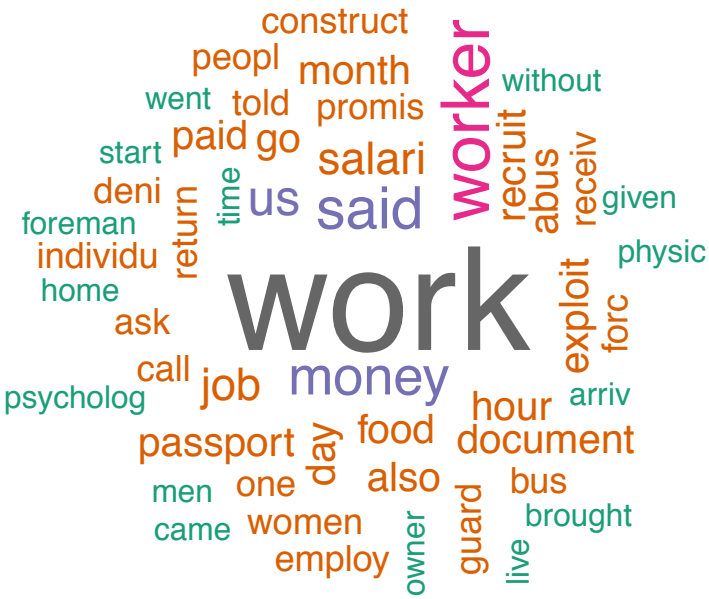


Figure 8. Word cloud of the full "decision"-variable.

This word cloud assembles the 50 most frequent words that appear in the “decision”-variable. The bigger a word, the more frequently it is used in the “decision”-variable. The colours help to better differentiate between the sizes of the words.

The word clouds are based on the top-features procedure. Top-features is a list of the most frequent words of a text, which then can be used to generate word clouds. The “quanteda” package (Benoit et al., 2018) provides the necessary tools to generate these top-features.

In a first step the text in the variable was tokenised. This means, that the computer identifies a text as a composition of words and not as one of sentences or as an entire text. Consequently, words are recognized as single elements. Further, the punctuation was removed. In a next step, “NA” was defined as a word that needs to be removed from the analysis, as the computer also counts it as a word (or element).

Then, a document-feature matrix (dfm) was created. This matrix associates values for certain features with each document (in our case, observations). This step also contains lowering all the characters, stemming the words and removing all English stopwords (for example, “and”, “or”) as well as the before defined “NAs”. The stemming is used in order to group words with the same meaning. For example, “working”, “worked”, “works” all become “work” and just indicate the action of working (the verb). By using the “topfeatures” function in a last step, we could extract the top 50 words that were contained in this variable.

Using the dfm, the word clouds can be generated: we defined the maximum of words that should appear, decided not to have a random order of the words and let the words rotate by ¼. To better differentiate between the sizes of the words, we also defined the colouring.

Additionally, we created top-features for bi-grams. Bi-grams are a form of n-grams, which are “a sequence of tokens from already tokenized text objects” (Kohei and Müller, 2018). This means that the list of the bi-gram top-features shows the two words that appear together most frequently. Getting more targeted ideas of what the “decision”-variable contains, top-features and word clouds were made for subgroups of the observations, such as by regions, exploitation type and means of contact initiation.

This technique is adequate as it is easy to apply and gives an idea of the words that have been used by the case workers to justify their decision. If we take a closer look at this word cloud, we can see that most of the used terms are related to the exploitation phase and are in line with what one would expect in order to justify a decision based on the Palermo Protocol: “work”, “construct”, “money”, “abus”, “exploit”. However, the word cloud also contains words regarding the entry into process and recruitment, as we can see from the following words: “recruit”, “promis” and eventually even regarding the transport of the victims as “bus” might indicate. It is interesting that these words do not seem to give much information about the socio-economic background of the person or the pre-exploitation period.

An important limitation of this technique, however, is that the words are not considered in context. This means that the interpretation of them is very limited. As for example a word like “construct” could indicate that the person was exploited in construction work, or that the work previous to the exploitation was in construction.

5.3 Text Analysis

As mentioned in section 5.2, the top-features analysis revealed that the “decision”-variable contains a lot of information regarding the exploitation phase, the entry into process and recruitment. This indicates that the “decision”-variable includes information, which is already represented by the structured variables. Hence, we concluded that if variables have many missing values, the missing information is written in the “decision”-variable. Using text analysis techniques, we added the information in the “decision”-variable to the structured variables.

The purpose of this procedure was to reduce the missing values in the dataset to improve its quality. 16034 out of 52364 of the observations in the IOM dataset do contain an entry for the “decision”-variable. Of those 16034 observations we randomly selected 100 observations from the “decision”-variable. By reading those 100 observations we were able to see which questions from

the IOM screening and assistance questionnaires had been answered most frequently. For the text analysis, we chose all questions, which were answered more than ten times, which led to 85 remaining questions. Of those 85 variables each one was compared to the “decision”-variable and it was verified whether information from the “decision”-variable could be added to the structured variable. Figure 9 shows a selection of those variables, i.e. the variables which were answered more than 25 times.

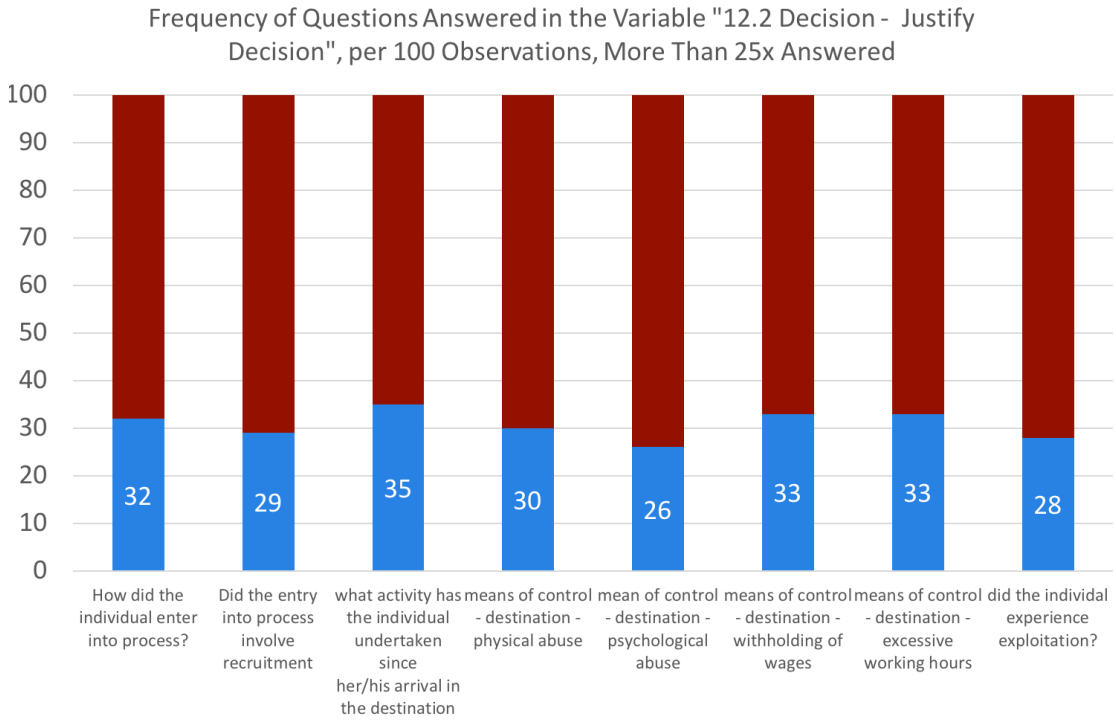


Figure 9. Most answered questions in the "decision"-variable.

The question whether a person was recruited is used as an example to explain the adding of information to the structured variables. Three criteria were defined to decide whether information from the “decision”-variable was inserted into a structured variable.

Criterion 1: Correct term. It had to be verified that a term in the “decision”-variable gives a correct answer to the corresponding structured variable. Regarding our example, this means we had to find terms, which describe a person’s recruitment. We began with the terms “recruit”, “recruited”, “recruitment” and “recruiting”. For those terms there was a high probability that they all refer to the process of recruitment. Moreover, there was the term “lured” which was assumed to denote recruitment as well. As it is less straightforward than the other terms it had to be verified that it indeed refers to recruitment. In such cases we applied a text analysis tool called “keywords in context” (KWIC)[2]. The tool allows assessing the context of predefined words. The meaning of a term can be grasped which leads to a more profound analysis of texts. This is an advantage compared to the top-features tool mentioned earlier. But it has to be kept in mind that KWIC only produces snippets of texts, which can still be misinterpreted. Nevertheless, we chose KWIC because it delivers good results considering its required work input. We set a threshold of accuracy of 85%. This means that in 85% of the cases a term has to refer to the given concept. Coming back to our example it means that if the term “recruitment” appears in the text variable it has to describe the fact that a person was recruited in at least 85% of the cases.

[2] The KWIC tool is also included in the “quanteda” package.

Criterion 2: A substantive number of variables to replace. We determined that the information in the structured variable has to be augmented by at least 10%. For example, the recruitment variable contains 28,555 observations. Through text analysis we discovered additional information meeting the pre-defined criteria. We created a new variable including the existing observations of the "recruitment"-variable as well as the detected information. To this variable, 3,755 observations were added, which is 13.1% of the existing information in the "recruitment" variable.

Criterion 3: Importance for machine learning analysis. The variable is relevant for the analysis described in section 6, meaning that they refer to the pre-exploitation phase.

If the first and either the second or the third criterion were fulfilled, we added the information from the "decision"-variable to the structured variables. These variables are described hereafter:

A. Entry into process - Did the entry into process involve recruitment?

- Terms used: "recruitment", "recruiting", "recruit", "recruited", "lured"
- Number of values added: 3'755
- % added information: 13.1
- Fulfilled all three criteria

B. Entry into process - Promised benefit of believed activity upon arrival

Dummy variables were created for each of the subcategories of this variable. Due to that, the variable became more convenient for the analysis. The subcategories are "basic needs", "education", "job", "material benefit" "monetary benefit", "registration/visa" and "transportation". In this repsect, the most frequent term in the „decsion"-variable was related to money. Hence, we filled the variable "Promised benefit of believed activity upon arrival - Monetary Benefit".

- Terms used: "promised salary", "promised wage", "salary promised"
- Number of values added: 537
- % added information: 263
- Fulfilled all three criteria

C. Entry into process - How did the individual enter the process? Sold by family / Sold by non-family

A random sample of 50 observations from the "decision"-variable was analysed to understand the context of the term "sold". The IOM contains separate variables for people sold by a family member and for those sold by a person not belonging to the family. However, this was not mentioned in the "decision"-variable in most of the cases. Therefore, the two variables "Sold by Family" and "Sold by non-Family" were combined to one variable.

- Terms used: "sold into", "sold by", "was sold", "were sold", "sold her", "sold him", "sold everyone", "defrauded and sold", "recruited and sold", "sold and exploited", "transported and sold", "deceived and sold", "cheated and sold", "recruited, sold", "sold for", "sold as", "sold girls", "sold us", "sold them"
- Number of values added: 576
- % added information: 1.4
- Fulfilled criteria 1 and 3

D. Questions asking about the activity of a person

Three types of activities are covered by the IOM questionnaires: pre-trafficking activities, activities promised to the victim of trafficking upon arrival and the arrival activities, which are the activities the victims of trafficking were actually exploited in. A random sample of 50 observations was analysed using the KWIC method for the activities “agriculture”, “begging”, “construction”, “domestic servitude”, “factory”, “prostitution”, “restaurant”, “study” and “trade”. Most of the activities mentioned were related to arrival activities. Through the reading of the 100 random samples of the “decision”-variable mentioned at the beginning of this section, we learned that the arrival activity is identical to the believed activity except for the activities prostitution and restaurant as well as study and domestic servitude.[3] Consequently, if an arrival activity was mentioned in an observation of the “decision”-variable we assumed it was also the believed activity of this observation and vice versa. Due to that, we complemented the variables relating to arrival activities and promised activities by that information. These variables are displayed in the appendix. For calculating the accuracy threshold, we considered both activities relating to believed and arrival activities. The procedure is explained in detail in the appendix as well.

- Terms used: See appendix, section 8 for details
- Number of values added: Ranges from 81 to 321
- % added information: Between 0.4 and 19.5; see appendix for details
- Fulfilled criteria 1 and 3

E. Entry into process - How was contact initiated between the individual and the recruiter?

We applied the KWIC method for the following terms: personal contact, newspaper advertisement and internet advertisement. The other terms (street advertisement, TV advertisement, radio advertisement, employment agency, travel agency) appeared in the “decision”-variable in very small numbers. Hence, we did not apply KWIC for those terms since it was clear beforehand that only a neglectable number of answers could be added to the structured variable.

- Terms used: Internet advertisement («internet» has an accuracy level of 96%); Newspaper advertisement («newspaper» has a 96% accuracy level); Personal contact («recruited by» has an 78% accuracy level, term is not used to fill the variable)
- Number of values added: 15 for “internet” and “web”, 40 for “newspaper”
- % added information: 2.5 for “internet” and “web”, 2.4 for “newspaper”
- Fulfilled criteria 1 and 3

F. Means of control

For most identified victims of human trafficking exploitation is accompanied by different types of means of control. For each mean of control there are two different variables referring to the two distinct stages of the human trafficking process, i.e. the movement phase and the destination phase. However, this temporal distinction was hardly mentioned in the “decision”-variable. For this reason, for every mean of control the two variables were combined to one “means of control”-variable. The variables concerned are displayed in the appendix.

[3] Victims ending up in prostitution usually believed they would be working in a restaurant. Victims being exploited in domestic work very often thought they would be able to study.

- Terms used: See appendix, section 9 for details
- Number of values added: Ranges from 55 to 3,186
- % added information: Between 2.9 and 133.6; see appendix for details
- Fulfilled criteria 1 and 2

G. Type of exploitation

The two exploitation types “sexual exploitation” and “forced labour” are the output factors we chose to conduct the classification tree analysis (see section 4.2.2). Therefore, we decided to add information concerning those exploitation types to the variables the variables “23 - Exploitation - Last Exploitation - Sexual Exploitation” and “23 - Exploitation - Last Exploitation – Forced Labour”.

- Terms used: Sexual exploitation (“sexual exploitation”, “sexually exploited”, “forced to work as a prostitute”, “had to work as a prostitute”); Forced labour (“forced labour”, “labour exploitation”, “forced to work”)[4]
- Number of values added: 174 for terms related to sexual exploitation, 233 for terms related to forced labour
- % added information: 0.6% for terms related to sexual exploitation, 0.8% for terms related to forced labour
- Fulfilled criteria 1 and 3

5.4 Evaluation

Chapter 5 outlined the techniques used to detect and exhaust the dataset's informational potential. Through cleaning and structuring parts of the dataset, we were able to make its information accessible for statistical analysis. Escpecially, text analysis and the ensuing completion of the structured variables served to reduce the number of missing values significantly.

In particular, the need of cleaning the dataset emerged from challenges in the machine learning analyses. Through the cleaning, we aimed at arranging the data more clearly and therefore making it more accessible to statistical analysis. In regard to the structuring technique, we expected to add more information to the structured variables in order to produce more generalizable findings. During the process of complementing the structured variables, we often dealt with the issue that the unstructured variables did not contain significantly more additional information than already present. Thus, for the selection of variables to be completed we set the minimum threshold of additional information to be gained to 10%. However, we selected some variables relevant for the subsequent analyses, regardless of the threshold. Furthermore, the top-features analysis, with the help of the sampling technique, was an important basis to gain an overview of the most important topics in the “decision” -variable.

Nevertheless, there are still two major drawbacks in our data. First, our refined dataset does still contain many missing values, which is displayed by Figure 10. The x-axis displays all the variables in the IOM dataset.

[4] We analysed a sample of 50 observations using KWIC to verify the expression “forced to work”. In 48 out of 50 observations the term referred to forced labour. Thus, the expression has an accuracy level of 96%.

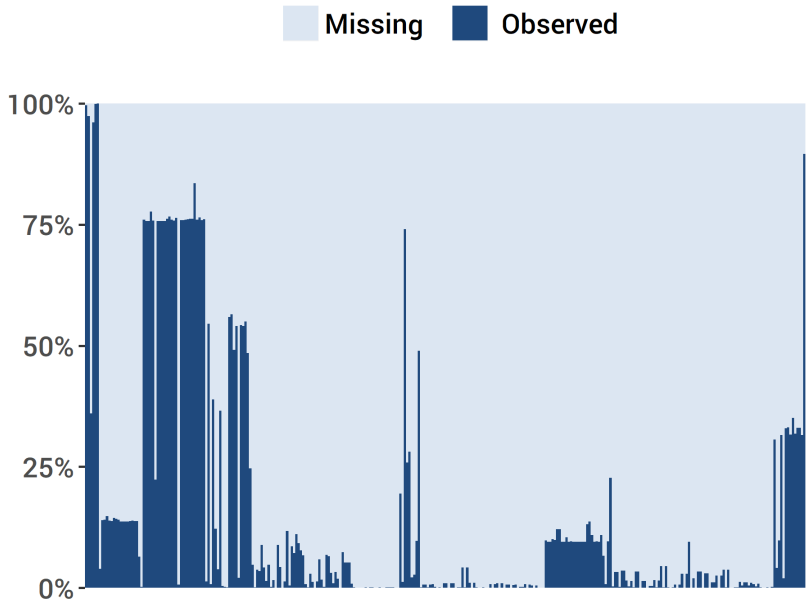


Figure 10. Missing values in the dataset.

Second, the results of our analyses are not generalizable as the dataset covers a subgroup of victims of human trafficking, which was not randomly selected from all victims involved in human trafficking. Under these considerations, it is essential to consider that our findings are only valid for the subgroup covered in the IOM dataset. Nevertheless, we found a way to deal with the many missing values in the dataset by using the machine learning technique, which is explained in the next chapter.

6 Advanced Statistical Analysis Framework

A crucial basis of human trafficking prevention is the development of a reliable set of characteristics that can be used to identify potential victims. The task of detecting potential characteristics is difficult because the process that drives persons into human trafficking is associated with complex patterns. To approach this challenge, we used machine learning techniques for the following three reasons.

Firstly, this approach allows us to map the relative importance of a vast number of characteristics and to display how they interact with each other (Breiman et al. 1984). Secondly, some machine learning approaches have ad-hoc built-in procedures that permit to handle missing data. This is an important aspect for our analysis as our dataset contains many missing values (Ibidem). Thirdly, this technique allows a highly inductive approach. That is, we are interested in new insights emerging directly from the data, rather than letting our empirical analysis to be guided (and potentially biased) by pre-defined theoretical assumptions. Moreover, machine learning techniques do not make any assumption about the distribution of the population. Hence, the pattern of data distribution would not affect the performance of the machine learning. This may be relevant if factors are not normally distributed, and also if different groups of victims show significantly different degrees of variance (Ibidem).

In the next sections we explain two types of machine learning techniques: cluster analysis and classification tree analysis. Then, we conclude with a discussion of the main limitations of our models that we have detected during the analysis. The purpose of this methodological part is to provide an overview of these two types of machine learning techniques, emphasizing its practical use rather than its underlying statistical theory.

6.1 Cluster Analysis

To get a first impression of the dataset, we applied cluster analysis. Cluster analysis is a useful tool to find patterns and groupings in the data and therefore a popular form of exploratory data analysis. It is an unsupervised machine learning technique. The goal of clustering is to build meaningful groups (or clusters) with observations that share common characteristics. Ideally, objects in the same cluster are very similar (high intra-cluster similarity) to each other and very dissimilar to those in other clusters (high inter-cluster dissimilarity).

6.1.1 Model: Hierarchical Clustering

Clustering can be achieved by various algorithms, which all differ in their understanding of what constitutes a cluster and how to find them. Generally, there are two standard clustering strategies: partitionial clustering (e.g. k-means) and hierarchical clustering. We decided to use the latter, since it is more of an exploratory type, constituting a better fit to the nature of our analysis. Hierarchical clustering creates clusters, which are nested in a predetermined ordering, i.e. a hierarchy, while partitional clustering divides data objects into groups that are non-overlapping, such that each observation is in exactly one cluster.

One of the advantages of hierarchical clustering is that the algorithm does not require a specification of the total number of clusters beforehand but allows us to define which number of clusters looks best after the clustering process. In addition, any valid measure of distance (or similarity) can be used in hierarchical clustering. This is especially important in our case because we have a mixed dataset, containing both numerical and categorical data, and are therefore limited in our choice of distance measure. Generally, the clustering process can be summed up in four steps.

4 Steps to Hierarchical Cluster Analysis

Step 1	Preparing the data: Remove missing values
Step 2	Choosing a similarity measure: Which metric is appropriate to calculate the distance between data points?
Step 3	Choosing the clustering method: How should observations be grouped?
Step 4	Assessing clusters: How many clusters are appropriate?

Step 1: Preparing the data

Before performing cluster analysis in R, missing values must be removed. The received subset of the IOM Global Database on Victims of Human Trafficking serves as a basis for the following analysis. After the cleaning and adding variables (e.g. variables classifying observations according to regions defined by IOM) coded by us, the dataset covers 52,364 observations and 321 variables in total. After deleting all the unstructured variables and the variables containing information on the country level, 210 variables remain. Due to the large proportion of missing values in the dataset provided (see chapter 5.4), removing each observation, which contains at least one missing value (NA), would result in a dataset of zero observations. For this reason, we decided to only include variables, which have a maximum of 60 percent NAs. Once the missing values are removed, we are left with a dataset containing 17,385 observations of 43 variables. The threshold of 60 percent was chosen based on a trade-off between keeping as many variables as possible, whilst still maintaining a large enough sample size (Table 2).

Table 2. NA-theshold for cluster analysis.

NA-Threshold	No. of variables, once NA's removed	No. of observations, once NA's removed
None, whole dataset	210	0
Max. 70%	55	503
Max. 65%	47	503
Max. 60%	43	17'385
Max. 50%	41	18'413

One drawback of hierarchical clustering is that its advantages come at the cost of lower efficiency. Because hierarchical clustering algorithms require a NxN distance matrix (where N is the number of observations) to be calculated, they are computationally intensive for large samples. For this reason, we decided to perform the analysis on a random sample of a thousand observations.

Step 2: Choosing a similarity measure

In order to decide which observations should be combined or divided into a cluster, a distance metric for measuring the similarity between objects is needed. There are many methods to calculate the distance between objects, in our case we needed a distance metric that can handle mixed data types. We decided to use the Gower distance, also known as Gower's Similarity Coefficient (Gower, 1971). Gower's Similarity Coefficient is one of the most popular measures of (dis)similarity for mixed data types. For each variable type, a separate distance metric adjusted to that variable type is used and scaled to fall between 0 (= identical) and 1 (= maximally dissimilar).
For two cases x_i and x_j, the similarity measure is defined as:

$$s_{ij} = \sum_{k=1}^{p} S_{ijk}\delta_{ijk} \,/\, \sum_{k=1}^{p} \delta_{ijk}$$

Where δ_{ijk} indicates whether a comparison of i and j is possible based on variable $\mathrm{k}, \mathrm{whereas}\ \mathrm{k} = 1, \dots, \mathrm{p}$. *Hence,* δ_{ijk} equals to 1 when variable k can be compared for i and j, and 0 otherwise.[5] The score s_{ijk} captures the similarity between individuals i and j with respect to variable k. The scores s_{ijk} are assigned as follows:

- For binary variables, the presence of attribute k is denoted by $+$ (present) and $-$ (absent), as shown in Table 3. $s_{ijk} = 1$ if cases i and j both have attribute k "present", or 0 otherwise, and δ_{ijk} causes negative matches to be ignored (Gower 1971, p. 859).

Table 3. Scores and Validity of Binary Variable Comparisons

	Values of attribute k			
Individual i	+	+	-	-
Individual j	+	-	+	-
s_{ijk}	1	0	0	0
s_{ijk}	1	1	1	0

- For nominal variables, $s_{ijk} = 1$ if individuals i and j agree in attribute k, and $s_{ijk} = 0$, if they differ in k.

- For quantitative variables with values $x_1, x_2, \dots, x_n$ of attribute k for the total sample of n individuals, $s_{ijk} = 1 - \frac{|x_i - x_j|}{R_k}$, where R_k is the observed range of attribute k.

From the similarity values s_{ij} the distances $d_{ij} = 1 - s_{ij}$ can be calculated and a dissimilarity matrix is derived. These pairwise dissimilarities now serve as the basis for the clustering algorithm applied.

[5] Sometimes, a comparison is not possible due to missing information, or for binary variables, when an attribute is missing for both i and j.

Step 3: Choosing the clustering method

Now that the distance matrix has been calculated, a cluster algorithm has to be selected. As already mentioned, we chose to apply a hierarchical clustering algorithm. Hierarchical clustering can be divided into two main types: agglomerative and divisive. Agglomerative clustering works in a bottom-up manner, meaning that each observation is initially considered to be a single cluster. At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster. This procedure is repeated until only one cluster remains. Divisive clustering works the other way around, splitting the data in a top-down manner. It starts with a single cluster containing all observations, and the most dissimilar ones are divided into smaller clusters, until all objects are in their own cluster.

Agglomerative clustering is the more popular algorithm, since divisive clustering is conceptually more complex and requires more computational power. On that account, we decided to go with agglomerative clustering.

To determine how newly formed clusters are linked to each other to form bigger clusters, a linkage method needs to be specified. Based on the dissimilarity matrix calculated in step 2, the linkage function decides which rules should apply to determine how close or similar two clusters are. Again, there are various different linkage methods. For our analysis, we chose the complete linkage method. This linkage method defines the distance between two clusters by looking at the maximum distance between their individual observations. The shortest of these maximum distances that remains at each step then causes the fusion of two clusters. This method is also known as furthest-neighbor distance (PennState Science, 2018).

Mathematically, the complete linkage distance between clusters X and Y can be described by the following expression:

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y)$$

where $d(x,y)$ is the distance between elements $x \in X$ and $y \in Y$; and X and Y are two sets of elements or clusters.

Step 4: Assessing Clusters

As mentioned earlier in this chapter, hierarchical clustering allows us to select the number of clusters after the clustering process. Assessing the number of clusters is more of a judgment call, since there is no objectively correct way of doing it. In general, the clusters should make sense, the distance within clusters should be small and the distance between clusters should be large. Often, there is a trade-off between homogeneity and generalizability: too few clusters might be too general, while too many clusters might be overly complex and difficult to understand. Two approaches to assess the cluster validity, which might produe different results, are introduced here: the elbow method and the silhouette method.

The elbow method is a measure for intra-cluster-similarity, or the compactness of a cluster. The elbow graph in Figure 8 shows the total intra-cluster variation, measured by the total within-cluster sum of squares (WSS), as a function of the number of clusters. The smaller the WSS, the higher the compactness of a cluster. One should choose a number of clusters so that adding another cluster does not improve the total WSS in a significant way. This drop in marginal gain can be seen by an angle, or elbow, in the graph (Kassambara, 2017). Looking at Figure 11 one could select 3, 6, or 8 clusters.

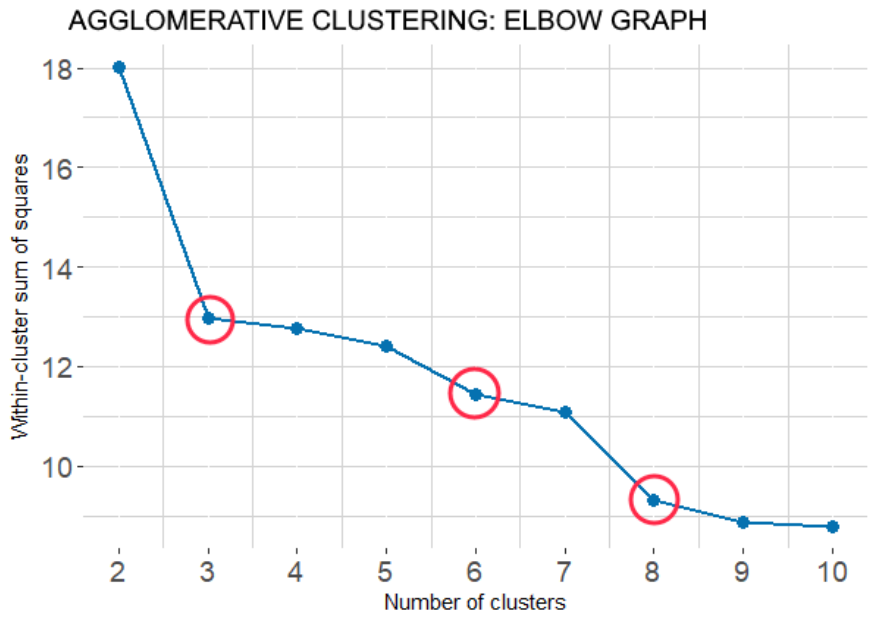


Figure 11. Elbow graph.

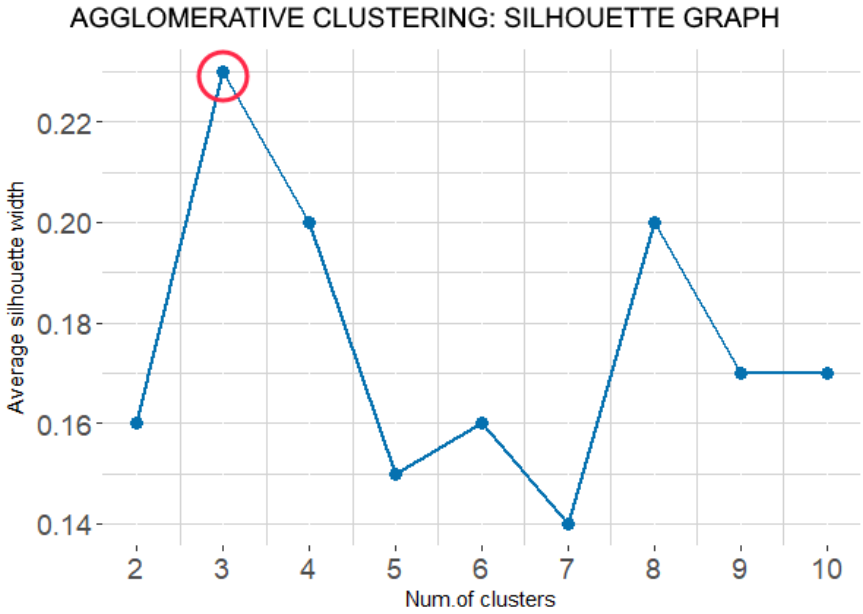


Figure 12. Silhouette graph.

The silhouette method measures how well an observation is clustered based on the average distance between clusters. A high average silhouette width indicates a good clustering. Figure 9 shows the average silhouette width as a function of the number of clusters. When assessing the silhouette coefficient, the highest values should be chosen. Judging by Figure 12, 3 clusters should be most appropriate.

Based on the information gained by the elbow and the silhouette plots, we chose to group our data in three clusters.

6.1.2 Results

The decisions and procedures discussed in the previous sections resulted in three clusters. The main characteristics of the first two larger clusters are depicted below in Figure 13.

The first cluster contains 599 out of 1000 observations, making up 60% of the total sample. 97% of the cases belonging to the first cluster were subjected to forced labour and 70% of them entered the process by labour migration. Gender and marital status only play a minor role in defining this group: the male-female-ratio is 53%:47%, with 44% of the cases being single and 39% married, and 17% reporting another marital status. The individuals were recruited (81%) and a striking 71% of identified victims belonging to this first cluster believed that their activity upon arrival would be "other", meaning an activity that was not pre-defined in IOM's forms and case management system. These victims were predominantly young adults (average age 32) from South-Eastern Europe and Central Asia (59%), or Asia and the Pacific (34%). The majority was exploited in their region of citizenship (48% in South-Eastern Europe and Central Asia; and 18% in Asia and the Pacific), although some victims were exploited in the MENA region (15%).

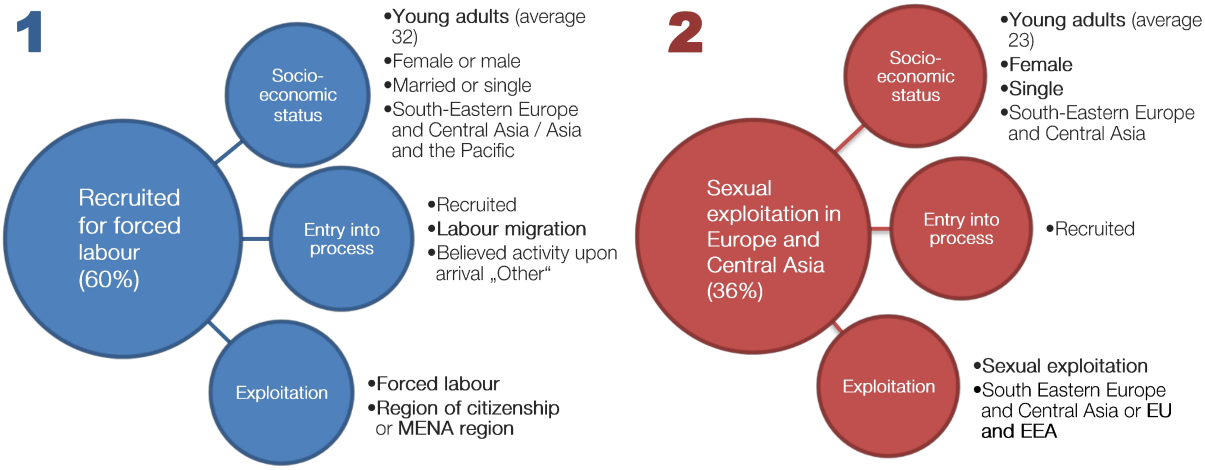


Figure 13. First and second cluster.

Cluster two covers 365 out of 1000 observations. Again, the type of exploitation is a defining factor, 88% of all the observations belonging to the second cluster are identified victims of sexual exploitation. Compared to cluster one, members of this group are almost exclusively female (98%) and single (72%). Also, they tend to be younger (on average 23) than the first group. There is little information regarding the entry into process for this group, except that the majority has been recruited (79%). The vast majority of the victims are citizens of South-Eastern Europe and Central Asia (77%), some of the come from the EU and EEA (25%). The exploitation mainly took place in the region of citizenship (64% in South-Eastern Europe; and 25% in Asia and the Pacific).

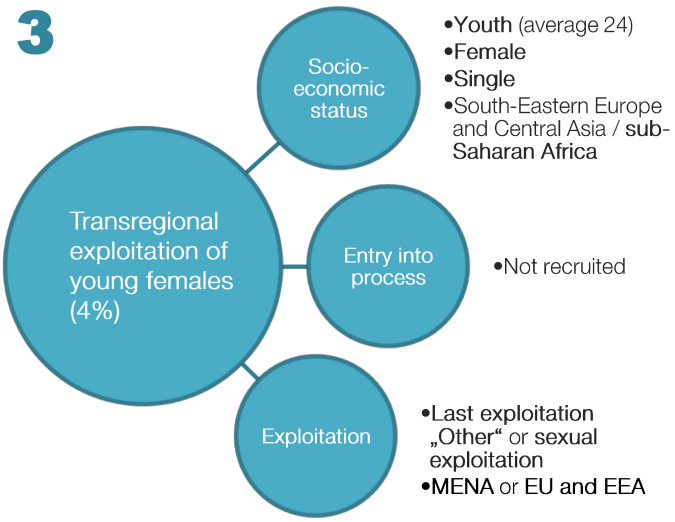


Figure 14. Third cluster.

The third cluster, seen in Figure 14, is much smaller than the first two, making up only 4 percent of the total sample. What sets this group apart is that the identified victims were not recruited (72%) and the majority was transregionally exploited (75%).
The victims mainly come from sub-Saharan Africa (44%, either East Africa and the Horn of Africa or West and Central Africa) and South-Eastern Europe and Central Asia (36%) and are predominantly young women. Over half of the identified victims assigned to this cluster were exploited in the MENA region, about one third suffered exploitation in the EU and EEA.
Another main difference between this third group compared to the other two is that the exploitation type is not a defining feature. Almost half of the cases belonging to this group was subjected to sexual exploitation (47%). 55% of the identified victims were subjected to an exploitation type “other” than the ones predefined by IOM’s forms and case management system. In addition, 30% were identified as victims of forced labour.

6.1.3 Summary

Cluster analysis involves various steps, requiring a lot of decisions. These decisions depend on the available data, the aim of the cluster analysis, and entail some subjective judgement. We believe that using hierarchical agglomerative cluster analysis has helped us gain interesting insights into the main patterns in our dataset and allowed us to build three prototypical victim profiles in a first, exploratory step. The clustering results have shown that the type of exploitation, along with region of citizenship, and region of exploitation, are essential determinants in dividing identified victims into subgroups. Moreover, gender (female) and marital status (single) seem to be highly associated with sexual exploitation compared to labour exploitation. These insights led us to the following questions: how do socio-economic factors interact with other features to predict the type of exploitation? Are there regional differences? The following section describes how we constructed more nuanced, dynamic victim profiles with the help of classification trees.

6.2 Classification Tree Analysis

Based on the results of the cluster analysis, the goal of our classification tree analysis is to identify victim chrachteristics according to types of exploitation.

6.3 Model

The classification tree is a type of supervised learning algorithm, which means that the algorithm learns from pre-selected input factors to predict a pre-selected output factor. The learning process of the algorithm then approximates a predictive decision tree model (Breiman et al. 1984). We have organised our analysis in the following steps.

4 Steps to Classification Tree Analysis	
Step 1	Preparing the data: Divide the dataset in two parts and select variables
Step 2	Choosing an algorithm: Which is the appropriate metric for the classification tree building process?
Step 3	Determining the number of splits: The pruning technique
Step 4	Assesisng decision tree model: How well does our decision tree model fit?

Step 1: Preparing the dataset

In order to assess our model's performance, we divided the dataset into two parts: a training set and a test set. The first is used to train the data, while the second is used to evaluate the learned or trained data. In practice, we randomly divided the dataset into a test and a training set. The most common splitting choice is to take 80% of the original dataset as the training set, while the remaining 20% will compose the test set (Gareth et al. 2013).

Furthermore, we generated two new variables as output factors in order to measure the different types of exploitation. The first measure is binary and stands either for sexual exploitation or for forced labour. The second measure represents the types of exploitation according to national citizenship. For this variable, the output factor includes four possible categories: sexual exploitation in the country of citizenship, sexual exploitation in a foreign country, forced labour in the country of citizenship, or forced labour in a foreign country. These are the prevailing types of exploitation emerged in the dataset. As input factors, we included all socio-economic characteristics and various other characteristics.

Step 2: Choosing an algorithm

The classification tree analysis is based on the Gini-index algorithm, which is by far the most common strategy for learning decision trees from data (Breiman et al. 1984). It constructs the decision trees in a top-down process, by choosing an input factor at each step that best splits the observations according to the output factor. Generally, the Gini-Index algorithm measures the homogeneity of the output factor within the subgroups. In more detail, it measures how many observations, for each socio-economic characteristics and entry-into-the-process characteristics, were misclassified according to the type of exploitation. The mathematical formalisation of the algorithm is the following:

$$\text{Gini Index (C)} = \sum_{i=1}^{n} p_i\,(1 - p_i) = \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} p_i^2 = 1 - \sum_{i=1}^{n} p_i^2$$

Where p_i is the proportion of observations of a particular category C of the input factor classified in each category of types of exploitation. In our analysis, n is either equal to two or equal to four categories depending on the selected output factor for the decision tree model. The split of recruitment in the classification tree model of South Eastern Europe (henceforth SEE) and Central Asia as regions of citizenship can be taken as an example for all our illustrated splits in decision tee models. The Gini-Index for recruitment characteristic includes two categories "yes" and "no", calculates the proportion of observation of "yes" and "no" which fall under the category either of forced labour or of sexual exploitation.

We then compute the weighted average of observations for each i-th category C over all categories resulting from the split of an input factor I:

$$\text{Weighted Gini Index}(C, I) = \sum_{i=1}^{n} \frac{C_i}{C} \cdot \text{Gini Index } (C_i)$$

In our example, the algorithm computes the proportion of two categories' observations, yes and no of the recruitment characteristic, compared to the total observations for the recruitment factor; and then multiplies them by their equivalent Gini-Index.

Under this consideration, the algorithm calculates the Gini-Index for each socio-economic and entry-into-the-process characteristic. It then chooses the split that optimally reduces the Gini index algorithm. This process is applied to each resulting split in the classification tree model. Consequently, the approach allows assessing the best split based on socio-economic and entry-into-the-process characteristics that results in the greatest homogeneity in each subgroup regarding the output factor.

Step 3: Determing the number of splits

To calculate the number of splits, we use the pruning method of "bias-variance". The bias term refers to the error rate which means how much the predicted values differ on average from the actual value. The variance term refers to a complex parameter which denotes the difference between the predictions of the model and different samples from the same population. This approach measures how much accuracy an additional split could add to the entire tree in order to accept the additional complexity. More specifically, as the complexity parameter increases without significant decreases in the predicted misclassification, splits are pruned away, resulting in simpler trees (Friedman et al. 2001).

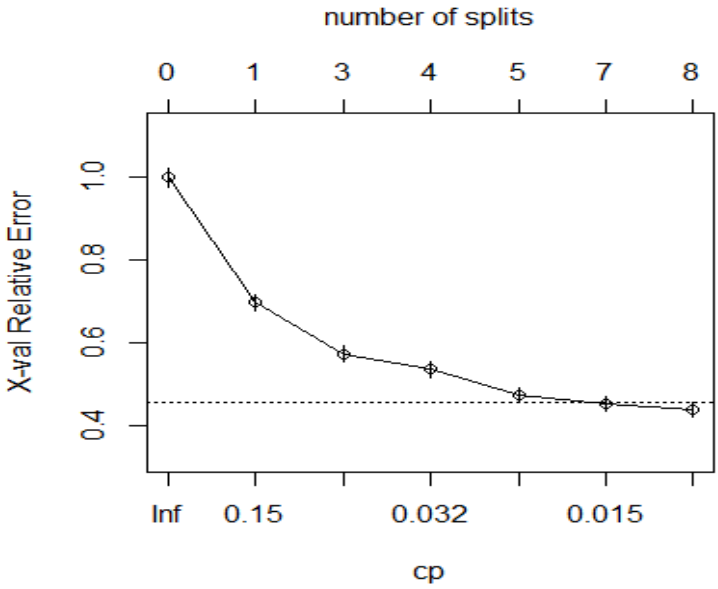


Figure 15. Pruning process regarding model for country of citizenship SEE and Central Asia.

We attempted to maintain a balance between bias and variance regarding our decision tree models. Figure 15 shows the pruning process for the decision tree model of SEE and Central Asia as regions of citizenship that can be taken as an example for all our illustrated decision tree models. The error rate is shown on the vertical axis, while on the horizontal axis depicts the complexity parameter and the equivalent number of splits. The Figure shows that as we increase the number of splits, we see a reduction in the error rate. The reason is that there is an increase of homogeneity within subgroups as the number of splits increases. However, as we continue to increase the number of splits, we would generate a very complex model with many splits. This means that our model would tend to be over fitted. Accordingly, we pruned all our decision tree models according to this horizontal line, which represents 1 standard error above the minimum of the error rate, horizontal line, which represents 1 standard error above the minimum of the error rate, leaving us in this case with 8 splits.

After the pruning process, all our victim characteristics that we presented in the main findings report have to fulfil further four criteria. Firstly, we consider only splits which have at least a probability of 0.80 according to the types of exploitation in order to be highly reliable results. Secondly, we cover only splits which are adequately represented in the sample in order to avoid overt-fitting. Thirdly, we selected victim characteristics, which are described by clear categories in each split (i.e. category not labelled as “other”) in order to interpret them as precise as possible. Fourthly, we chose victim characteristics representing an interesting case in respect to existing assumptions on vulnerability among prevention organisations (ILO, 2017).

Step 4: Assesisng decion tree models

Finally, the last step refers to the assessment of the decision tree model. This can be understood as the degree of correctness of our models’ predictions. Generally, we evaluated the performance of each decision tree models by means of a confusion matrix. A confusion matrix is a convenient way to examine the error rates for all subgroups. In other words, the confusion matrix assesses the predictions from the model with the real outcomes. Moreover, we assess the performance of our decision tree model by using the test dataset in order to validating our classification tree model based on observations that are not part of the training dataset (Stehman 1997).

Table 4. Confusion matrix of the model for country of citizenship SEE and Central Asia.

		Actual Values: Forced Labour	Actual Values: Sexual Exploitation
Predicted Values	Forced Labour	1331	118
	Sexual Exploitation	203	1122
		Accuracy Rate: 88.43 %	

Table 4 shows the confusion matrix for the classification tree model of SEE and Central Asia as regions of citizenship. The matrix with two rows and two columns reports the number of correctly predicted and the number of incorrectly predicted victims. With regard to victims of forced labour, 1331 cases are correctly predicted, whereas 203 cases are incorrectly predicted. With regard to victims of sexual exploitation, the number of correctly predicted is 1122, whereas the number of incorrectly predicted is 118. The accuracy indicator below the matrix evaluates the performance of our model based on this matrix, which is equal to 88.43% of correct classifications. The high rate of accuracy shows that our decision tree model constantly predicted our types of exploitation.

In the next section, we explain our findings with the help of an exemplary decision tree model. For readability reasons, we chose to discuss only the classification tree model of country of SEE and Central Asia as regions of citizenship as an example. The decision tree models for the other regions are presented in the appendix and can be understood with the same framework of analysis as presented here.

6.3.1 Results: Decision Tree Model for Country of Exploitation SEE and Central Asia

Figure 16 illustrates the classification tree for South Eastern Europe (henceforth SEE) and Central Asia as regions of citizenship. The names inside the light blue rectangles are the characteristics of victims. The related types of exploitation are displayed inside each circle. FL stands for forced labour and SE stands for sexual exploitation. Below the circles, there are two values indicated. The one, on the left side, indicates the predicted probability of the displayed subgroup of victims to end up in either forced labour or sexual exploitation. The value, on the right side, indicates the percentage of the sample for the selected regions that fall in a particular type of exploitation. In our example, the sample size for SEE and Central Asia includes 11251 observations. The highlighted path shows the dominant type of exploitation in the sample.

At the top of the tree, the algorithm has placed the most influential characteristics that permit to determine how to classify a victim according to types of exploitation. At the bottom of the tree, the algorithm has put the less influential characteristics. It becomes apparent that the most influential characteristic of this model is gender: a male has a probability of 0.98 to be involved in forced labour rather than in sexual exploitation, whereas, for a female, the probability to be pushed either into sexual exploitation or forced labour is less obvious (0,68 for sexual exploitation)

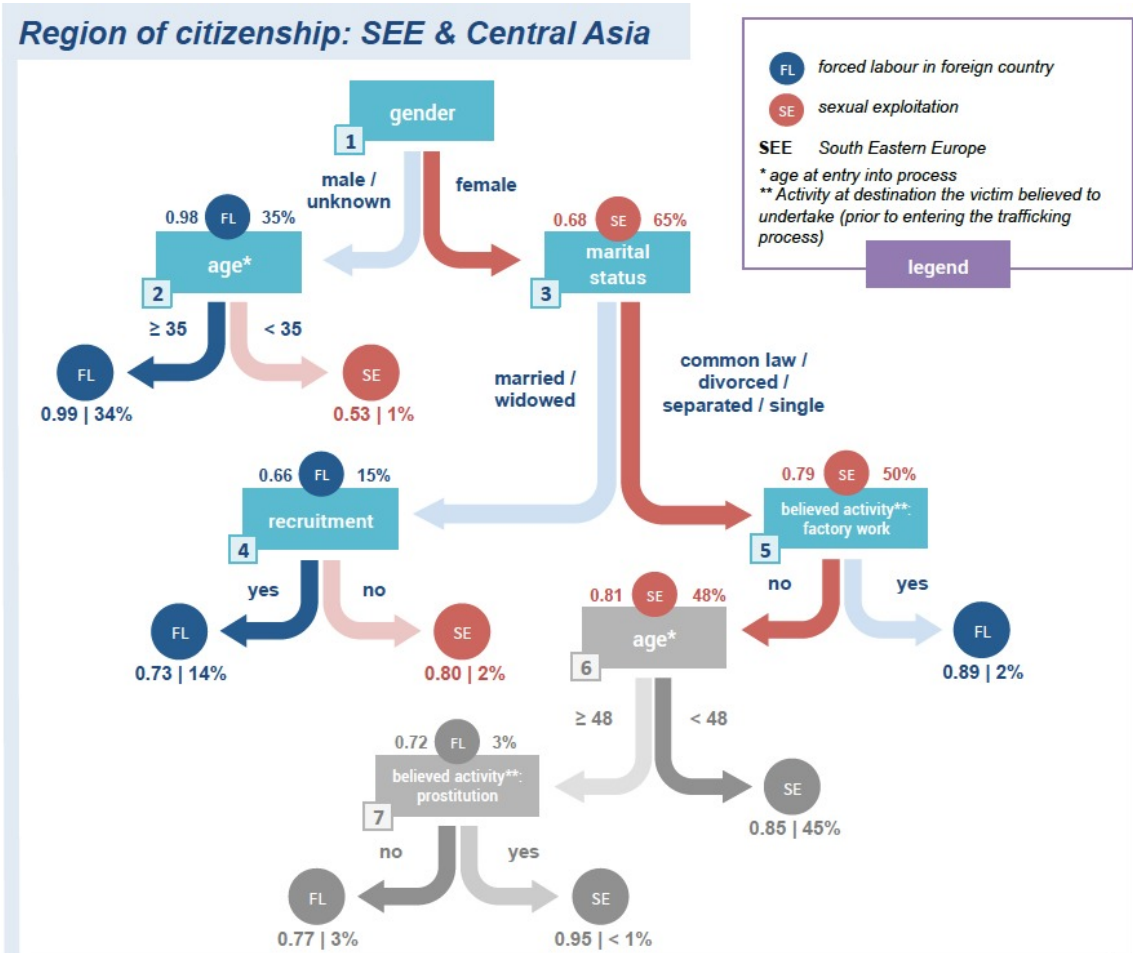


Figure 16. Decision tree model for region of citizenship SEE & Central Asia.

However, the decision tree model illustrates a more complex pattern of victim characteristics for females in comparison to males (and the category of unknown). With regard to males, the split number 2 shows that a male aged 35 years or older has a higher likelihood to end up in forced labour than in sexual exploitation, whereas males younger than 35 years old have a slightly higher probability (0.53) to end up in sexual exploitation than in forced labour. Considering the size of these two subgroups, males aged 35 years or older represent a large subgroup of 34% of cases in the sample, whereas males younger than 35 years old represent only 1% of the sample.

With regard to females, the split number 3 shows that the most decisive characteristic is the marital status in determining the type of exploitation. It is worth to mention that the probability of a female to be pushed into sexual exploitation rather than into forced labour increases to 0.79 if she is in either common law, divorced, separated or single. By contrast, a female who is married or widow has a slightly higher probability to end up in forced labour rather than in sexual exploitation. For this subgroup, whether a female is recruited or not becomes a decisive characteristic for predicting the type of exploitation (split number 4). More specifically, the probability of a married or widow female to be pushed into forced labour rather than into sexual exploitation increases to 0.73 if she is recruited; otherwise, if she is not recruited, she has a probability of 0.80 to be pushed into sexual exploitation than into forced labour. The former subgroup represents 14% of the sample size, whereas the latter subgroup represent only 2% of the sample.

Going back to the split number 3 (maritial status) and following the highlighted path further down (split number 5), the decision tree model shows that the believed activity of factory work is an important characteristic for the subgroup of females who are either common law, divorced, separated or single. Most notably, the probability of a common law, divorced, separated or single

female to be pushed into sexual exploitation, rather than into forced labour, increases to 0.81 if their believed activity was not factory work before the entry in the process. By contrast, the probability of this subgroup to be pushed into forced labour rather than, sexual exploitation, is 0.89 if their believed activity was factory work. The former subgroup represents 48% of the sample, while the latter case counts for only 2% of the sample.

Splits number 6 and 7 are coloured in grey since we want to emphasize that further splits would tend to be too specific of the selected sample and wound not add significant information. Indeed, the decision tree model ends up with subgroups with are represented by small number of cases. Thus, we prefer seemingly less accurate trees, or otherwise altering the accuracy when adopting the decision tree model for predicting observations that are not part of the training dataset.

6.3.2 Summary

In this section, we have discussed the classification tree analysis as a supervised machine learning approach. In the first part, we provided an in-depth insight into the methodology of classification tree approach in order to elucidate how we have generated our main findings. As in the cluster analysis, classification tree analysis involves various steps, which have help us to refine our main findings. In the second part, we showed how to interpret the classification tree analysis with the help of an example for South Eastern Europe and Central Asia as regions of citizenship. For these regions, the results confirm the prevalence of common characteristics as age and gender strongly deterring the types of exploitation. On the other side, our analysis reveals additional victim characteristics such as marital status, believed activity of factory work and recruitment which give a nuanced understanding of victims pushed into human trafficking according to the type of exploitation.

6.4 Limitations

As the previous sections of this chapter have shown, our empirical strategy was based on considerations of how to best deal with the available data, the structure of our data, and the complexity of our research object. Nevertheless, the chosen methods are bound to certain methodological limitations.

While cluster analysis is a great tool to get an idea of the main patterns and grouping in our data, performing hierarchical clustering on a dataset as large as ours is computationally very time expensive. For this reason, the results of our cluster analysis are based on a random sample of a thousand observations, limiting the representativeness and stability of our findings. Furthermore, only 43 out of 210 variables were included, due to the large amount of missing values in the dataset.

The basic methodological idea behind our decision tree technique is easy to understand and simple to use, offering many advantages compared to other statistical techniques. However, we noticed some limitations while applying our decision tree on the dataset.

One the one hand, decision tree models are to some degree instable. This is not desirable as we want our decision tree model to be pretty robust to noise and be able to generalize well to future observed data. One solution that we applied is to carry out the analysis for slightly different versions of the dataset and select decision tree models that are sufficiently robust. We experienced that decision tree models based on well-covered regions with a large sample in the dataset result in a higher robustness.

On the other hand, decision tree models are faced with a complexity problem, with a tendency to generate overfitted models. This is not desirable as the complexity problem hinders the generalization of decision tree models across the population. As already illustrated above, we applied the pruning technique in order to reduce the complexity of the tree models. Moreover, we combined the pruning mechanism with the confusion matrix in order to have an optimal trade-off between accuracy of the models and the complexity of the models.

7 Improvement of Data Collection

Throughout our work with the dataset we gathered a variety of experiences and also faced challenges and difficulties with the data. Dealing with those, revealed some potential for improvement in the dataset. We discovered this potential along three characteristics: exhaustiveness, completeness and consistency. From those characteristics, we derived recommendations for the data collection. Our recommendations are minor changes in the data collection practice that have relatively high impact on the possibilities of the data analysis. They should be understood as complementary remarks in consideration of the dataset's valuable contribution to counter-trafficking efforts. Furthermore, comparing them with the newest changes in the MiMOSA system (e.g. user interface, trainings, etc.) shows that some of the issues leading to our challenges were already addressed.

7.1 Exhaustiveness

A quality of a survey or interview to portray an issue realistically is the ability to cover the entire realm of possible cases. This requires a set of questions with a range of possible answers that are collectively exhaustive. If exceptional cases are expected, an open option like the category "other" can be used.
Among the first outputs of our explorative analysis we received a cluster including such an "other" category from the question "believed activity upon arrival" to be a decisive factor (Figure 17).

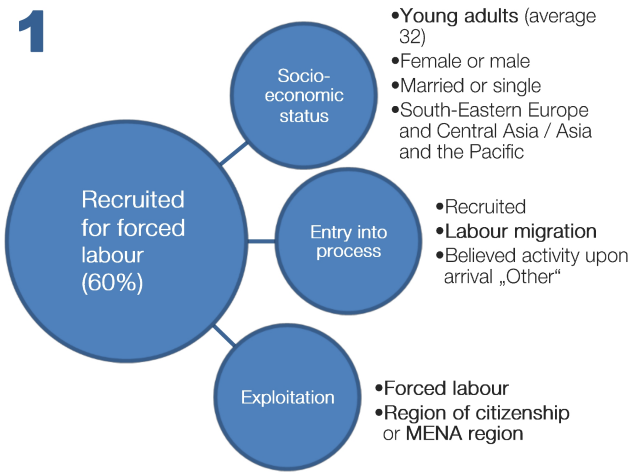


Figure 17. Believed activity upon arrival "Other" in the first cluster.

In relation to the cluster it means that this group is characterized by believing, before entering the process, to work in an activity other than any of the predefined answers. Analyzing further what activities "other" contains reveals that many cases covered by this category can be assigned to an existing category or harmonized into new categories (see chapter 5.1). This proves that the "other" category carries too much weight as it is represented as characteristic in the cluster. Therefore, the predefined set of answers is not yet able to represent some of the interviewed cases of human trafficking, i.e. it lacks exhaustiveness.

Consequently, we recommend to improve the data collection in this particular case by extending the question "If labour migration, what was the believed activity upon arrival" in the interview questionnaire by the newly created categories (see chapter 5.1). More in general, it would also seem meaningful to harmonize answer categories to one single set applied to all the questions relating to activities.

The following Table 5 informs about the comparison of newly created categories over three questions related to "activities". It would be more useful if the categories rely on an international standard guaranteeing that possible answers compare on the same categorical rank.

Table 5. New categories and numbers of cases filled-in.

Number of cases filled in newly created categories	Believed Activity upon Arrival	Last Activity prior to Entry into Process	Activity at Destination
Agricultural Work	794		
Entertainment	742	11	200
Beauty & Lifestyle	132	49	
Domestic Work	2906		
Child Care	187		
Any Work	2075		
Service Sector	719	52	
Self-Employment		246	

7.2 Completeness

A challenge we faced working with the dataset was its completeness, i.e. an overwhelming amount of missing values in the dataset. Some variables hardly contained any information. One possible reason for this lack of information we assume in the interview and data registration phase. In particular, questions where multiple answers are possible each predefined answer translates into the dataset as a dummy variable. We noticed that the dummy variables were covered differently depending on the case. The dummy variables which meet the victim's situation contain the value of "1". The remaining dummy variables (belonging to the same questions), which did not meet the victim's situation, were in some cases marked with "0" and in others marked as missing. Consequently, the significance of the missing value is unclear, i.e. if it has to be interpreted as real missing value or as a not meeting with the victim situation.

Table 6. Ignored observations in the subsamples due to missing values

	Citizen Asia and the Pacific	Exploitation in EU & EEA	Exploitation in South-Eastern Europe and Central Asia	Exploitation in MENA
Total Observations	8070	4189	23138	2355
Included Observations	4311	3034	8920	1920
Ignored Observations [%]	46.6%	27.6%	61.4%	18.5%

In our analysis missing values are either ignored in favor of simplicity or replaced with substituted values estimated with a statistical method. Table 6 presents the 4 classification tree models we used for the identification of our victim profiles. The last column shows the percentage of lost cases, which could not be used in the classification tree analysis, due to lack of information, i.e. the amount of missing values. However, the high presence of missing values in our dataset leads to a smaller sample size eventually compromising the reliability of our results. Generally, the high presence of missing values combined with the uncertainty regarding the effective missing values in our dataset is serious, leading to loss of information, decreased statistical power, and eventually weakened generalizability of findings. These severe limitations should be taken into account by the interpretation of our findings.

Relating to the mentioned example, there is a lot of potential to reduce the amount of missing values by clarifying and standardizing how information from the interviews has to be translated into the MiMOSA system.

7.3 Consistency

A further challenge we faced working with the dataset was the consistent representation of information, so called consistency, throughout the whole dataset. Under the term consistency we understand that the data does not contain contradicting information across the variables. By means of three examples we illustrate the difficulties we had due to lack in consistency.

As a first challenge in consistency we saw a harmonization problem. Initially, we assumed the many missing values to be a consequence of the lack of information. For some characteristics this proved partially wrong as we discovered the wrongly considered missing information in the unstructured variable. An example for this arose from the text analysis on the “decision”-variable where we tried to detect cases with evidence of recruitment (explained in the chapter 5.3). Comparing the victim cases we identified in this analysis to the cases documented as recruited in the concerned structured variable, differences could be seen. As stated in the chapter on refining the dataset, we detected 3’755 cases where recruitment was not marked in the structured variable. Here we assume a harmonization problem, which could originate from the information not being entered completely into the MiMOSA system by the IOM caseworker.

As a second challenge in consistency, we faced a contradiction problem within the dataset. There were two types how the contradiction occurred, again between a structured and an unstructured variable and within distinct structured variables. The former type is well represented by the example of the cases where the believed activity at arrival was prostitution. This attribute showed up as dominant in some machine learning outputs we initiated further verification. Reading the justified decisions of a random sample of 50 cases with believed activity prostitution we detected 10 cases with a clear contradiction between the variables. This is illustrated by the following text extract from a concerned case:

"*the victim was lured by a lucrative job opportunity far away from her home and when she arrived in her workplace, she realized that she actually works in a prostitution [...]*"

The latter type of the contradiction problem is illustrated along the example of the structured variables "recruitment" and "initial contact to the recruiter". The challenge became obvious in a classification tree where "initial contact to the recruiter" was a dominant attribute among a group of cases, which were identified as not recruited (structured "recruitment" variable marked with "no"). An illustration thereof can be seen in Figure 18.

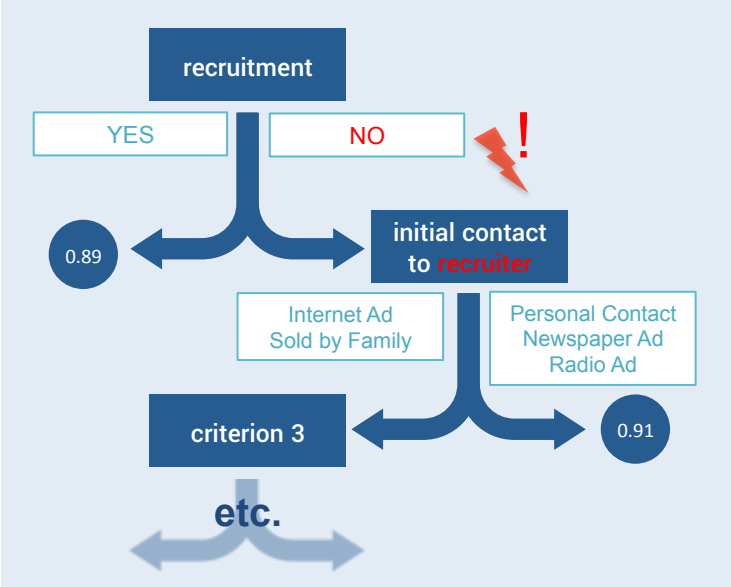


Figure 18. Schematic extract of classification tree showing contradiction between two attributes related to recruitment.

Generally, these examples may imply two causes of the harmonization and the contradiction problem. The first cause we imagine to be inherent to the interview questionnaire's structure and the entering of the information into MiMOSA. The second cause we see in the clarity of the concepts used in the interview as for example believed activity or recruitment. IOM caseworkers but also the assisted victims might understand the concepts differently whereby the information would not be comparable.

A way to prevent such challenges might be to either emphasize stronger on the clarification of central concepts for all the caseworkers. Or to reduce the use of such complex concepts using circumscriptions of the concepts in simple language that are understandable to the victims of which some do not have a clear understanding of those concepts.

7.4 Recommendations

On the basis of the above-discussed challenges we want to stress the importance of the improvement of the data collection. These challenges limit the explanatory power of the results of our analysis. The limitation raised by different understanding of the used concepts implies the importance of the MiMOSA trainings synchronizing the caseworkers usage of the concepts. Moreover, all of the challenges show that improvement can be reached by adapting the MiMOSA interface by for example creating mandatory fields or clear dependencies among the questions. In addition, the interview questions can also be better adapted to the current nature of the issue to reach better exhaustiveness, i.e. better covering the different victim's situations. Generally, we noticed that there is a big margin of improvement where with simple changes in data collection big impacts on the quality of the dataset can be achieved.

8 Conclusion

In the fight against the phenomenon of human trafficking the necessity of an in-depth understanding of victim characteristics becomes evident. We examine a new individual-level dataset from trafficking victim surveys to shed light on characteristics determining the entry of an individual into the trafficking process. The scope of our analysis is to refine existing knowledge on victims for developing potential prevention strategies of counter-trafficking.

The classification tree analyses presented within the scope of this report revealed empirical evidence on the perception of potential victims. On the one side it confirms the prevalence of the dominant characteristics as age, gender, origin and the individual's economic status. But on the other side our analysis emphasized additional characteristics that proofed to be influential. Among these we find marital status, what the individual believed to work as at destination and how the contact to the recruiter was initiated. These factors are more far-reaching than classical socio-economic characteristics and describe aspects of entering into the trafficking process. But most importantly the classification tree models allow the assessment of the relative importance of a characteristic among a wide range of characteristics. Therefore, the identification of groups with distinct, intersected characteristics is possible.

Besides complementing the understanding of victim characteristics, our study shows how the application of advanced analysis techniques can be advantageous in the field of human trafficking research. The inductive approach enables the identification of victim characteristics moving beyond pre-assumed social groups. Also, through machine learning strategies we were able to map complex interactions of such characteristics. Our approximated models were generated through a learning process directly from the data itself. Most notably, the two machine learning approaches we apply, cluster and classification tree analysis, can illustrate which particular characteristics determine variation across groups of victims respecting a context of various characteristics. Consequently, machine learning techniques show that structured quantitative empirical analysis has much potential to support counter-trafficking efforts. Classification tree analysis has further the advantage of dealing well with missing values like in the case of the Database on Victims of Human Trafficking.

Nevertheless, the importance of conducting high quality surveys plays a crucial role for gaining explanatory power and validity of the analysis' findings. Although, information collected from victim assistance interviews is already very valuable due to its closeness to the issue and because it is disaggregated at the individual-level. The translation into a complete and consistent dataset is one of the fundamental interests for developing effective prevention strategies on potential victims. We have to bear in mind that human trafficking is of clandestine nature with a large part of its victims remaining unidentified. The fact that IOM human trafficking data stems from a distinct group of identified and assisted victims limits the deduction of our findings to trafficking victims in general. Moreover, as data on human trafficking is still quite limited in its scope further emphasis should be put on the enlargement and optimization of data collection.

Despite these limitations, an elaborated data collection strategy combined with advanced statistical techniques is the starting point for designing reliable counter-trafficking solutions, their implementation and the monitoring of the policies effect.

Literature

Aronowitz, A. (2009). Guidelines for the collection of data on trafficking in human beings, including comparable indicators. (Vienna: Federal Ministry of the Interior of Austria & International Organizaion for Migration (IOM)). Retrieved from http://publications.iom.int/system/files/pdf/guidelines_collection_data_iomvienna.pdf

Benoit, K. et al. (2018). quanteda: Quantitative Analysis of Textual Data. R package version 1.3.14. Retrieved from https://cran.r-project.org/web/packages/quanteda/quanteda.pdf

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Wadswort: Chapman & Hall/CRC.

Dottridge, M. (2002). Trafficking in Children in West and Central Africa. Gender and Development, 10 (1), 38-42.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27 (4), 857-871.

Financial Action Task Force (FATF) and Asia/Pacific Group on Money Laundering (APG). (2018). Financial Flows from Human Trafficking. (Paris.)

Retrieved from https://www.fatf-gafi.org/media/fatf/content/images/Human-Trafficking-2018.pdf

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning.* New York: Springer

Galos, E., Bartolini, L., Cook, H. & Grant, N. (2017). Migrant Vulnerability to Human Trafficking and Exploitation: Evidence from the Central and Eastern Mediterranean Migration Routes. (Geneva: International Organization for Migration publication). Retrieved from https://publications.iom.int/system/files/pdf/migrant_vulnerability_to_human_trafficking_and_exploitation.pdf

Gareth, J., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning.* New York: Springer.

International Labour Organization (ILO). (1930). Forced Labour Convention, 1930 (No. 29). Retrieved from https://www.ilo.org/dyn/normlex/en/f?p=NORMLEXPUB:12100:0::NO::P12100_ILO_CODE:C029

International Labour Organization (ILO). (2017). *Methodology of the global estimates of modern slavery: Forced labour and forced marriage*. Geneva: International Labour Organization publication.

International Organization for Migration. (2017a). Combating Trafficking in Persons and Contemporary Forms of Slavery. Retrieved from https://www.iom.int/sites/default/files/our_work/ODG/GCM/IOM-Thematic-Paper-Trafficking-in-persons.pdf

International Organization for Migration. (2017b). Global Trafficking Trends in Focus IOM Victim of Trafficking Data 2006 – 2016. (August 2017). Retrieved from https://www.iom.int/sites/default/files/our_work/DMM/MAD/A4-Trafficking-External-Brief.pdf

International Organization for Migration (IOM) and McKinsey & Company. (2018) More than numbers: How migration data can deliver real-life benefits for migrants and government. (Vienna.) Retrieved from https://publications.iom.int/system/files/pdf/more_ than_numbers.pdf

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*. Marseille: STHDA.

Kohei, W. and Müller, S. (2018). Quanteda Tutorials. Retrieved from https://tutorials.quanteda.io

PennState Eberly College of Science. (2018). Applied Multivariate Statistical Analysis: 14.4 - Agglomerative Hierarchical Clustering. Retrieved from https://newonlinecourses.science.psu.edu/stat505/node/143/

Rahm, E., Do, H. (2000). Data Cleaning: Problems and Current Approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 23(4), 3-13.

Stehman, S. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62 (1). 77–89.

United Nations. (2000). Protocol to Prevent, Suppress and Punish Trafficking in Persons Especially Women and Children, supplementing the United Nations Convention against Transnational Organized Crime. Retrieved from https://www.ohchr.org/Documents/ProfessionalInterest/ProtocolonTrafficking.pdf

United Nations General Assembly (UNGA). (2010). United Nations Global Plan of Action to Combat Trafficking in Persons, (A/RES/64/293. New York.) Retrieved from www.refworld.org/docid/4caadf8a2.html

United Nations Office on Drugs and Crime (UNODC). (2006). Trafficking in Persons: Global Patterns. (Vienna: United Nation publication) Retrieved from http://www.unodc.org/documents/human-trafficking/HT-globalpatterns-en.pdf

United Nations Office on Drug and Crime (UNODC). (2008). Kidnapping at the national level, number of police-recorded offences. (Vienna: United Nation publication) Retrieved from https://www.unodc.org/documents/ data-and.../Kidnapping.xls

United Nations Office on Drugs and Crime (UNODC). (2008). An Introduction to Human Trafficking: Vulnerability, Impact and Action. (Vienna: United Nation publication) Retrieved from http://www.unodc.org/documents/human-trafficking/An_ Introduction_to_Human_Trafficking_-_Background_Paper.pdf

United Nations Office on Drugs and Crime (UNODC). (2009). Global Report on Trafficking in Persons. (Vienna: United Nation publication) Retrieved from http://www.unodc.org/documents/Global_Report_on_TIP.pdf

United Nations Office on Drugs and Crime (UNODC). (2009). Global Report on Trafficking in Persons. (Vienna: United Nation publication) Retrieved from http://www.unodc.org/documents/Global_Report_on_TIP.pdf

United Nations Office on Drugs and Crime (UNODC). (2016). Global Report on Trafficking in Persons. (Vienna: United Nation publication.) Retrieved from https://www.unodc.org/documents/data-and-analysis/glotip/2016_Global_Report_on_Trafficking_in_Persons.pdf

Appendix

1. Text Analysis: Questions asking about the activity of a person

Variable Name	Terms used	Accuracy [%]	Absolute Number of Values added	Added Values [%]
16 - Entry_into_process - Believed activity upon arrival - Agriculture	Agriculture/agricultural	92	156	19.5
25 - Destination phase - Arrival Activity - AgriculturalWork	Agriculture/agricultural	92	113	2
25 - Destination phase - Arrival Activity - Begging	Begging	94	81	1.5
25 - Destination phase - Arrival Activity - DomesticWork	Domestic work/ domestic servitude/ domestic servant	98	135	2.1
16 - Entry_into_process - Believed activity upon arrival - Construction	Construction	88	164	0.4
25 - Destination phase - Arrival Activity - Construction	Construction	88	307	4.5
25 - Destination phase - Arrival Activity - Prostitution	Prostitution	96	321	4

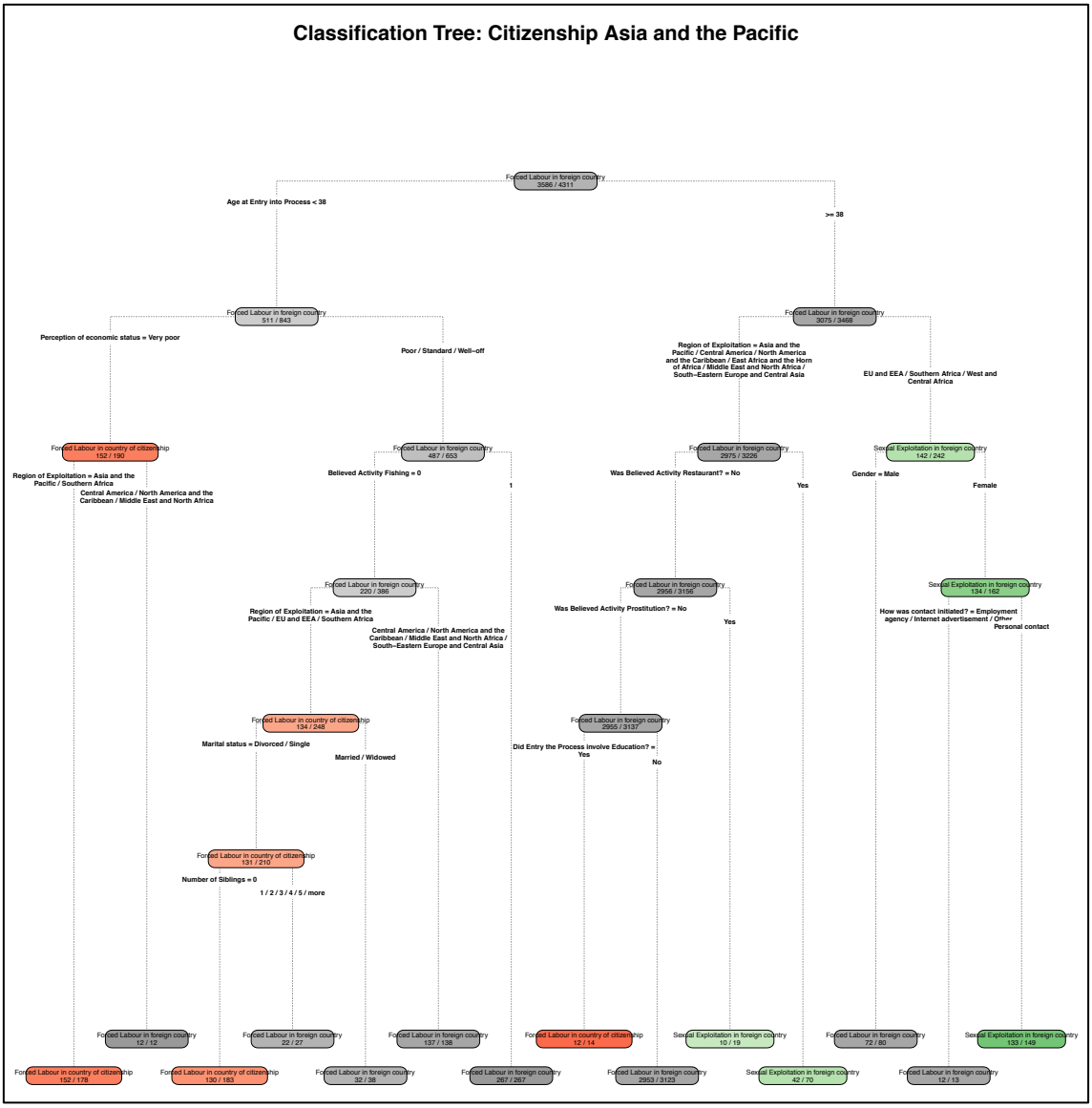
Believed activity chosen because important for analysis

Terms used	N of Terms used relating to	Calculation	Accuracy [%]
Agriculture/agricultural	Pre-trafficking activity: 4 Believed Activity: 28 Arrival Activity: 18	$\frac{28+18}{50}$	92
Begging	Pre-trafficking activity: 1 Arrival activity: 47 Unclear/other: 2	$\frac{47}{50}$	94
Construction	Pre-trafficking activity: 4 Believed activity: 8 Both believed and pre-trafficking activity: 3 Arrival activity: 33	$\frac{8+3+33}{50}$	88
Domestic work/ domestic servitude/ domestic servant	Believed activity: 1 Arrival activity: 49	$\frac{49}{50}$	98
Factory	Pre-trafficking activity: 11 Believed activity: 18 Arrival activity: 20 Unclear/other: 1	$\frac{18+20}{50}$	76
Prostitution/prostitute	Arrival activity: 48 Unclear/other: 2	$\frac{48}{50}$	96
Restaurant	Pre-trafficking activity: 5 Believed activity: 6 Arrival activity: 12 Unclear/other: 27	$\frac{6+12}{50}$	36
Study/ studying/ student/ studied/ studies	Pre-trafficking activity: 22 Believed activity: 13 Unclear/other: 15	$\frac{0}{50}$	0
Trade	Believed activity: 1 Arrival activity: 1 Unclear/other: 48	$\frac{1}{50}$	2

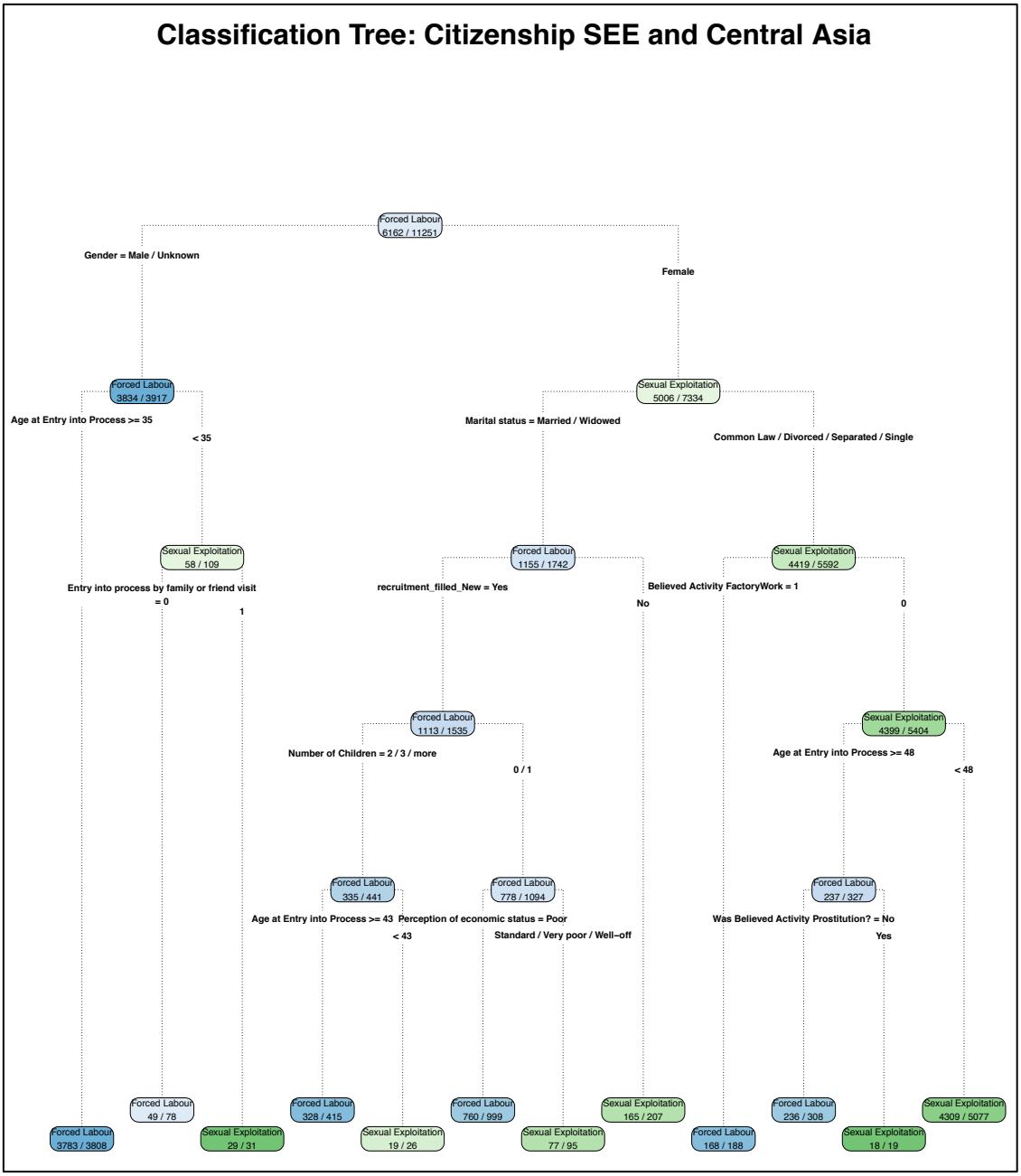
2. Text Analysis: Means of Control

Variable Name (Before Filling)	Terms used	Accuracy [%]	Absolute Number of Values added	Added Values [%]
S11A - Means of control - Physical abuse	Physical abuse/ physically abused	100	1214	68.6
S11B - Means of control - Psychological abuse	Psychological abuse/ psychologically abused	100	2135	110.3
S11C - Means of control - Sexual abuse	Sexual abuse/ sexually abused	100	314	41.5
S11D - Means of control - Threats to individual	Threat(s) to individual	100	55	2.9
S11G - Means of control - False promises/deception	False promise(s)/ deception/ deceit/ lured	100	3186	133.6
S11H - Means of control - Denied freedom of movement	Freedom of movement	100	2171	95
S11K - Means of control - Denied medical treatment	Denied medical treatment/ denied in medical treatment/ medical treatment denied	100	202	13.2
S11L - Means of control - Denied food/drink	Denied food/ denied drink / food was not given	100	244	50.2
S11M - Means of control - Withholding of wages	Withholding of wage(s)/ withholding wage(s)/ wages were withheld/ wage was withheld/ wage(s) withheld/ not paid	100	2134	118.9
S11P - Means of control - Debt bondage	Debt bondage/ bondage by a debt	100	657	48.1
S11Q - Means of control - Excessive working hours	Excessive working hour(s)/ Excessive hour(s)	100	1894	95.1

3. Classification Tree: Citizenship Asia & the Pacific (Counts)



4. Classification Tree: Citizenship SEE and Central Asia (Counts)



5. Classification Tree: Citizenship West and Central Africa (Counts)