ANALYZING HUMAN TRAFFICKING DATA

Methodology Report

September 2019

Capstone Course Report

IOM UN MIGRATION

University of Zurich[UZH]

The opinions expressed in the report are those of the authors and do not necessarily reflect the views of the International Organization for Migration (IOM). The designations employed and the presentation of material throughout the report do not imply the expression of any opinion whatsoever on the part of IOM concerning the legal status of any country, territory, city or area, or of its authorities, or concerning its frontiers or boundaries.

The data analysis was conducted from February to June 2018. All data, figures and citations refer to information available from 2002 until 2017.

Details provided by identified and assisted victims have been changed to prevent identification.

ACKNOWLEDGEMENTS

This report was written by Janine Albinez, Patrik Aus der Au, Saro Gibilisco, Théoda Woeffray and Jenny Yin under the guidance and supervision of Dr. Judith Spirig (University of Zurich), Eliza Galos (IOM) and Prof. Marco Steenbergen (University of Zurich).

The report was produced in the context of a joint project between the International Organization for Migration (IOM) and the Department of Political Science at the University of Zurich (UZH). The project was carried out by Master students within a Capstone Course with the aim to use students' research and quantitative data analysis skills to support the evidence base for IOM's counter trafficking efforts. In return, the students had the opportunity to learn about migration and human trafficking and the identification of trafficking victims. This publication was made possible by accessing IOM's global database on victims of human trafficking.

The authors are very thankful to Harry Cook and Mathieu Luciano for their review. They also thank Damiano Toffano for his presentation to the authors on assistance to victims of human trafficking, and to Anh Nguyen who facilitated the final presentation on this work in Geneva, at IOM headquarters.

ACRONYMS

CENA	Central Asia
DTM	Displacement Tracking Matrix
EEA	European Economic Area
EU	European Union
FMP	Flow Monitoring Point
FMS	Flow Monitoring Survey
ILO	International Labour Organization
IOM	International Organization for Migration
MENA	Middle East & North Africa
SEE	South-Eastern Europe
UNICEF	United Nations International Children's Emergency Fund
UNODC	United Nations Office on Drugs and Crime

INTRODUCTION

This report supplements the report "Analyzing Human Trafficking Data – A Machine Learning Approach" with in-depth explanations on the methodology. It informs in detail about the statistical methods applied, limitations and challenges that were faced during the analyses and recommendations for data collection and processing.

To support prevention and protection activities in human trafficking, an in-depth understanding of victim characteristics is essential. It would allow for timely identification of potential victims of human trafficking (Aronowitz, 2009; Galos et al., 2017). Therefore, the analyses from this report – based on IOM's Global Database on Victims of Human Trafficking - aim to get a more nuanced picture of the dynamics at work, foremost, to enable the identification of potential victims and to improve prevention and protection activities. They detect the most distinctive victim characteristics related to their entry into the human trafficking process, and the type of exploitation they are subjected to. The analysis focuses on the likelihood of a victim to be subjected to either forced labour or sexual exploitation, given a set of (pre-trafficking) characteristics. In other words: which victim characteristics predict specific types of exploitation?

The results of this study are based on the application of less conventional methods of analysis in the field of trafficking data. These machine learning techniques like clustering and classification tree analysis allow to follow a highly inductive approach. New insights emerge directly from the data, rather than from empirical analysis guided by pre-defined theoretical assumptions. The application of machine learning techniques enables the uncovering of complex relationships in the data. This is especially valuable for large datasets like the one applied in this analysis.

Further, these machine learning techniques bear the advantage of handling missing values and other weaknesses of the data very well. Such shortcomings in the data are often related to inconsistencies in the data collection and processing. 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 open further opportunities to understand and fight trafficking in persons.

This methodology report starts with introducing the terminology and the data. Subsequently, it explains the main features of the methodology and the application of the new analysis techniques. The report concludes with an assessment of the suitability and the potential of the method for the application to this type of data.

TERMINOLOGY

This section clarifies the concepts, which are most relevant for this work. As most of the definitions and concepts in the field of human trafficking are not uniformly used, this research relies on the following ones.

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 trafficking 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 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 frequently identified 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 used to describe forced labour exploitation, while forced sexual exploitation is simply called sexual exploitation (UNODC, 2016; ILO, 2017; IOM, 2017b). Hence, the terms forced labour and sexual exploitation are used in the remaining report to distinguish the two main 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. Unsupervised learning aims to uncover functions, groupings, and patterns in the data, without any predefined outcome. Supervised learning aims to predict a defined outcome given the input information. The goal is to formulate a function that best predicts a pre-selected outcome.

THE DATA

The data consists of over 50'000 interviews with registered victims of trafficking from more than 70 countries between 01.01.2000 and 30.06.2017. It is a subset of the IOM Global Database on Victims of Human Trafficking collecting primary data from identified victims of trafficking that were assisted in IOM's assistance programs. Caseworkers collect the information during the assistance interviews using IOM's case management tool MiMOSA to read in the data. The data includes quantitative and qualitative information on the individual trafficking experiences.

The structure of the dataset is based on the IOM Victim of Trafficking questionnaires guiding the interviews. Questions with predefined answers are translated into variables with structured values (structured variables). For questions which allow for multiple answers (e.g. type of exploitation), a dummy variable is created for each answer category. Open questions resulting in individual answers are translated into variables with unstructured values (unstructured variables). 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, the data set's representation of a reference population of all victims of trafficking in terms of baseline demographic features has to be considered as limited. Selection bias of assisted and interviewed individuals might be influenced by different factors like the location and size of the IOM country offices and the existence of counter-trafficking experts and programs. 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 responsible handling. To safeguard the victims' identity and privacy and to adhere to data protection principles, the whole dataset needs to be anonymized first. The next section explains this procedure.

Anonymization

Anonymization of the data makes the identification of individuals based on the information in the data impossible. This is important to ensure that identified victims do not endanger themselves by sharing sensitive information about their story and their perpetrators. Information contained in the "decision" variable is particularly critical as it is very detailed. The text contained in this variable has been fully anonymized.

The anonymization of the variable implies the following two steps:

- the named-entity recognition technique extracts names from the text and classifies them into two categories: names of persons and names of locations
- names of persons and of locations in the text are randomly replaced by the classified names of persons and locations, respectively

Consequently, the names in the unstructured "decision" variable no longer correspond to the original, i.e. true, names. Randomly swapping the names was preferred to replacing them with encrypted names since the named-entity recognition technique does not guarantee the classification of all names in the unstructured variable. Therefore, classified and swapped names cannot be distinguished from original not classified names.

Data Cleaning

An important foundation of any analysis is clean data. Cleaning data enables the access to consistently structured information free from errors that might have arisen when the information was entered into the database or during further processing. Especially, Machine learning methods are very sensitive to the structure of a dataset and rely on consistently formatted information. Primarily, this step targets the detection and correction of instance-level[1] errors and inconsistencies to improve the accuracy and consistency of the data. The cleaning included following steps:

- **Standardizing the labels of missing values:** changing all the different labels used to flag missing values (e.g. -99, N/K, n.a., etc.) to one consistent label - ,,NA"; this ensures equal treatment throughout all the variables.
- **Cleaning Value Levels:** Value levels of a variable were fitted to their logical structure. Illogical and invalid values have been removed (e.g. if binary ,,yes/no" variables, answers like ,,child care" or ,,15" were removed).
- **Adjusting Data Type of Variables:** Variable types 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 were removed – so too, variables that do not contain any information (all NA).
- **Adjusting the language:** The language was adjusted by correcting for common spelling mistakes detected through thescanning of a random samples. French and Spanish text passages were manually translated.

[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).*

Open Answer Categories

One source of unstructured information are open answer categories like the „if other, please specify". These answer options allow the respondent to specify the answer if it does not match any of the predefined answer categories. This collection of individual answers translates as an unstructured variable into the dataset. The structuring of that unstructured information is possible in following two ways:

- reducing the degree of informational detail of the individual answer (e.g. alter „cook in a canteen" into „restaurant") and assign it to one of the existing answer categories (marked as A in Figure 1).
- harmonizing and grouping the remaining answers (that could not be assigned to existing categories) to create new additional answer categories (marked as B in Figure 1).

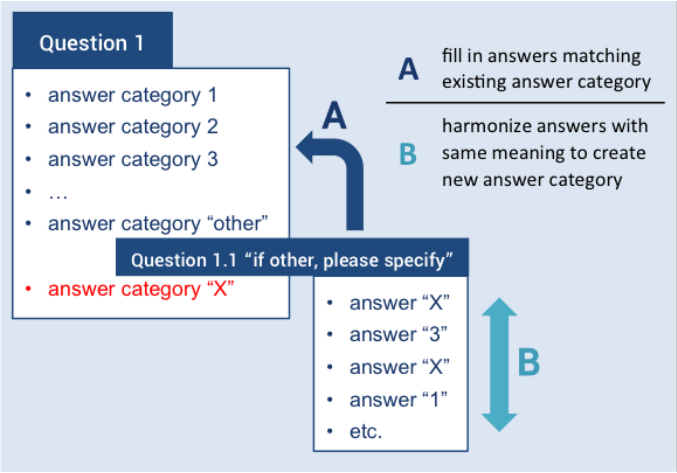

Figure 1. Assigning and grouping individual answers from the „if other, please specify"-variables to existing or new answer categories making unstructured information accessible.

The criteria for the creation of a new category are mutual exclusiveness and comparable size of observations to the existing categories. The following example illustrates the two processes how 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?". As it allows for multiple answers, each predefined answer category translates into a dummy variable[2]. One answer category is „restaurant" representing activities in the gastronomic sector. When for example a respondent did not pick the category „restaurant" (marked as 0 in the structured dummy variable) but answered the „if other, please specify" section with „cook in a canteen" this information will be underrepresented in the analysis. In a wider sense this individual answer could be understood as circumscribing an activity matching the category „restaurant". Through simplifying the information of this answer it could be assigned to the category "restaurant" (marked as 1 in the structured dummy variable). At the costs of losing the additional information about the function of the respondent at the restaurant (cook), an additional observation can be made available for the analysis.

The remaining individual answers, which could not be assigned to an existing category, were grouped to new answer categories, if possible. Among the unstructured answers referring to the believed activity upon arrival, answers like „pole dancer at a night club", „bar hostess", „dancer", „night club" and „masseuse" were summarized under the new answer category „entertainment" counting 742 observations. Creating „entertainment" is justified by the assumption that these activities might be closely linked to the milieu where prostitution is organized.

This procedure was applied to seven questions from the dataset that included an „if other, please specify" open answer option (Table 1). The selection of these questions relies on their relevance to our analysis and their potential through the number of complete individual answers. The following table gives an impression of the informational gain from these seven questions.

The table shows that the potential of informational gain (making unstructured information accessible) varies among the questions. Among the questions with high informational potential it can be distinguished between the ones where the potential is driven by the possibility of filling in existing categories and others where the predefined answers are non-exhaustive and new categories can be created. While the former might result from the hesitant use of the predefined answer categories through the case worker, latter might result from answer categories which do not adequately reflect real situations. The informational gain (row 5 in Table 1) varies widely as sometimes unclear, too generally formulated or niche answers had to be neglected.

[2] *A dummy variable is an indicator variable that takes the value 0 or 1 to indicate the absence or presence of the represented answer category.*

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	9,944	2,883	12,192	17,210	19,623	6,451
	unstructured	11,742	3,372	3,234	697	388	1,039
additional answers assigned	to existing categories	224	89	2,461	56	151	79
	to new categories	7,555	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]

New Categories Legend:

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

„Decision" Variable: Top Features

Another rich source of unstructured information is the "decision" variable. It is the largest unstructured variable containing individual answers covering multiple topics of the interview. This allows for the structuring of information across questions, i.e. variables.

Source: "decision" variable

The content of the "decision" variable bears a high potential on additional information. For some interviews it summarizes the story of the respondent providing detailed information yet not covered by previous questions. The word cloud represents the frequency with which single words appear over all given answers to this question (Figure 2). 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.

This word cloud gives the impression that the answers to the "decision" variable include information on the type of exploitation (e.g. work, salary, exploitation) and the means of control (e.g. passport, guard, food). Taking into account that the Palermo Protocol is the guideline for identifying victims of trafficking, the frequency of words indicating the exploitation phase is not surprising. However, the word cloud also contains words regarding the entry into process, recruitment and transportation (e.g. recruit, promise, bus). Information on the respondents' socio-economic background or the pre-exploitation period seem rare.

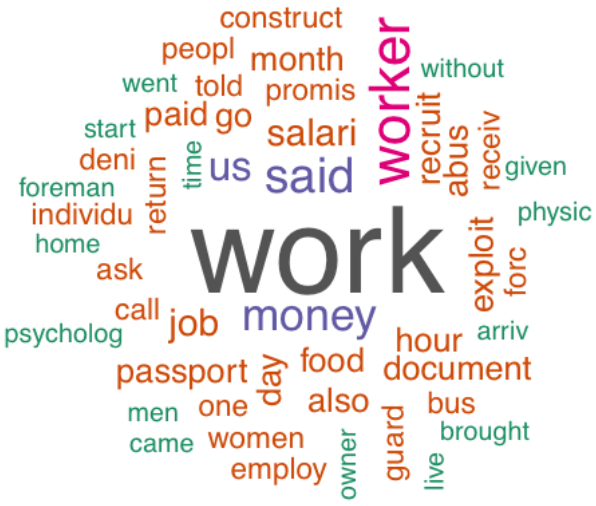

Figure 2. Word cloud of the full „decision"-variable.

Method: top-features

Word clouds are generated based on the top-features procedure – creation of a list of the most frequent words of a text. Through tokenisation answer texts are defined as a composition of words, which are recognized as single elements (tokens) independent from the structure or the semantics of the sentences. Punctuation, English stop words (e.g. "and", "or") and numbers are removed from that list of words. After the stemming of the words[3] a document-feature matrix (dfm)[4] was created. This matrix associates values for certain features – words as basic elements - with each document– answer text. The "topfeatures" function extracts the 50 most frequently appearing words throughout all answers. The "quanteda" package (Benoit et al., 2018) provides the necessary technical tools to generate these top-features.

A variation of top-features is using bi-grams, which are a sequence of two tokens from an already tokenized text objects (Kohei and Müller, 2018). Top features of bi-grams shows word pairs that appear most frequently together. To get more targeted ideas the "decision"-variable's content, top-features and word clouds can be made for subgroups such as regions, exploitation type or forms of contact initiation. Additional word clouds and top-feature lists are presented in the appendix.

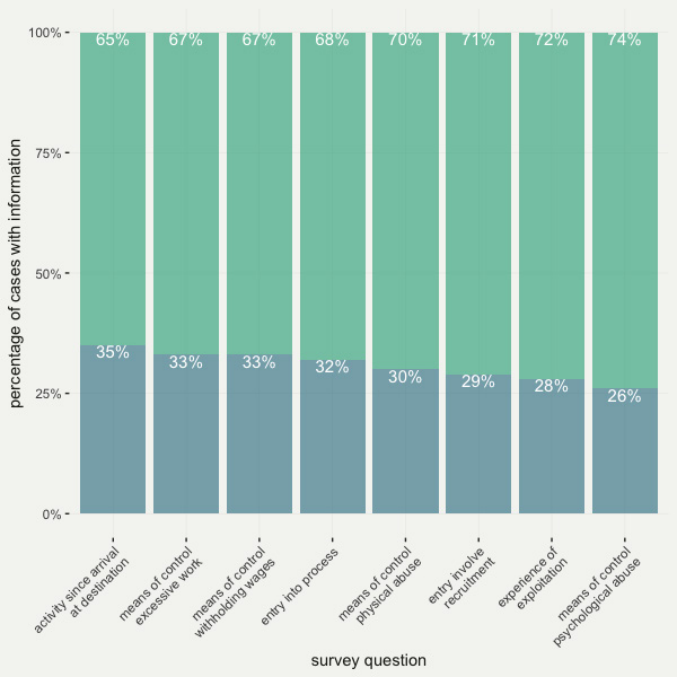

Figure 3. Most answered questions in the „decision"-variable.

This method is useful to get a first impression of the content of this unstructured variable. Knowledge on topics that might be covered can help to conduct additional analysis. However, the fact that the words are not considered in context limits the interpretational strength of this method. A word like "construct" could indicate the exploitation of the respondent on a construction site, but also that he or she worked in construction prior to entering into the process of exploitation.

Text Analysis

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 topics which are covered by questions in the structured part of the dataset. Hence, if variables have many missing values, some of the missing information might be found in the "decision" variable. Through text analysis techniques additional information was detected in the "decision" variable and added to the structured variables.

Method: text analysis

The purpose of this procedure was to reduce the missing values in the dataset to improve its quality. 16,034 out of 52,364 of the observations in the IOM dataset contain an entry for the "decision" variable. Analysing a random sample of 100 of those observations showed which survey questions had been answered most frequently. A selection of questions or answer variables, which were answered more than ten times, built the basis for the text analysis. This resulted in the selection of 85 variables (including answer category variables). Of the selected 85 variables each was compared to the "decision"-variable to verify whether additional information could be gained from the unstructured part. Figure 3 shows a selection of those variables, which were answered more than 25 times. To explain the structuring of additionally gained information, the question if a respondent was recruited serves as an example. Three criteria were defined to decide whether information from the "decision" variable was added to 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. Therefore terms, which describe a person's recruitment had to be detected. At first, the terms "recruit", "recruited", "recruitment" and "recruiting" were used. The probability that those terms all refer to the process of recruitment

[3] *Stemming is grouping words with the same meaning by their word stem. For example, "working", "worked", "works" are summed under the stem "work" and indicate the action of working (verb).*

[4] https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html

was high. Moreover, there was the term "lured" which was assumed to denote recruitment as well. As it is less straightforward than the other terms, verification that it indeed refers to recruitment was needed. A text analysis tool called "keywords in context" (KWIC)[5] served the purpose. The method allows assessing the context and therefore the meaning of predefined words. Nevertheless, one must be aware of the limitation that these snippets created by the KWIC method can be misinterpreted as well. An accuracy threshold of 85% of the cases a term has to refer to the given concept was chosen. In this 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.

Criterion 2: Substantive number of values to replace. Information of the structured variable has to be augmented by at least 10%. For example, the recruitment variable contains information in 28,555 observations. The additional information that was detected through text analysis and met the pre-defined criteria and added to the recruitment variable. In the case of 3,755 respondents information of the structured variable could be augmented (13.1% of the existing information).

Criterion 3: Relevance for machine learning analysis. The third criterion is related to the focus of the machine learning analysis. A variable was chosen for complementation through additionally detected information if it referred to the pre-exploitation phase.

If the first and either the second or the third criterium were fulfilled, the information gained from the "decision" variable was added to the structured variables. The information gain by variables is outlined in Table 2.

Evaluation

Application of different text analysis methods allowed the harmonization and improvement of the dataset. Information from the unstructured parts of the dataset was detected and structured. Thereby, this additional information could be made accessible for further analysis. This noteworthy gain of information added considerably to the expressive power and the generalizability of the findings from the analysis. Most of the dataset harmonization was triggered by challenges met during the application of the machine learning technique. It was a cyclical process running the analysis and further adapting and exploring the unstructured parts of the dataset.

Nevertheless, informational gain from the unstructured variables was lower than initially expected. Often, text analysis revealed that the unstructured variables do not contain significantly more information than the information already present in the structured variables. The minimum threshold of 10% additional information gain proved to be too high. Therefore, variables were rather chosen based on their relevance for further analysis than on the improvement potential. The top-features analysis was an important part in gaining an overview of the data and the most important topics of the unstructured parts.

However, there are still two major drawbacks in our data. First, the refined dataset does still contain many missing values, which is displayed in Figure 4. The x-axis of the plot displays all the variables in the IOM dataset, while the y-axis shows how much information the variables contain.

Second, the results of the analyses are not generalizable as the dataset covers only identified victims as a subgroup of victims of human trafficking. They were not randomly selected from the victim population. Under these considerations, it is essential to consider that the findings are only valid for the subgroup covered in the IOM dataset. Nevertheless, this analysis finds 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.

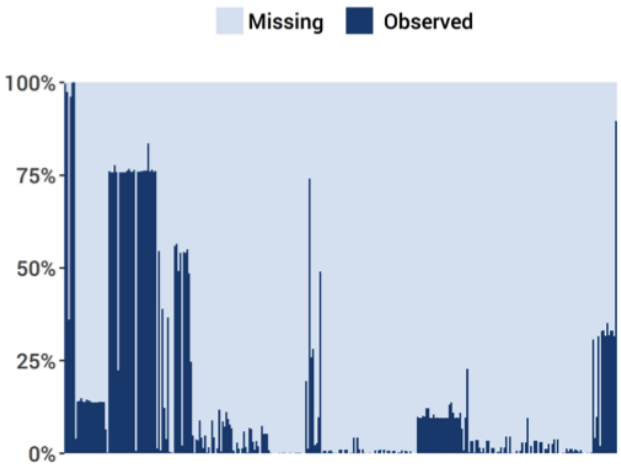

Figure 4. Missing values in the dataset.

[5] *The KWIC tool is also included in the "quanteda" package.*

Table 2. Informational gain through making unstructured variables accessible

Question/Variable	Terms used	Values added	Info gain	Fulfilled criteria	Comments
Did the entry into process involve recruitment?	"recruit*", "lured"	3,755	13%	1-3	* allows different endings like recruitment, recruiting, recruited, etc.
Promised benefit of believed activity upon arrival - Monetary Benefit	"promised salary/wage", "salary promised"	537	263%	1-3	Most frequent answer category of "Promised benefit of believed activity upon arrival"; dummy variable
How did the individual enter the process? Sold by family / Sold by non-family	"sold into/by/for/ as", "was/were sold", "sold her/him", "sold everyone/girls/us/ them", "defrauded/ recruited/ transported/ deceived/ cheated and sold", "sold and exploited"	576	1.4%	1 & 3	Analysis of random sample showed that information in the "decision" variable does not specify if sold by family or non-family; general variable for "sold by" created
How was contact initiated between the individual and the recruiter?	internet", "web" (96% accuracy)	15	3%	1 & 3	Analysis of random sample showed that information in the "decision" variable does not specify if sold by family or non-family; general variable for "sold by" created
Means of control	"newspaper" (96% accuracy)	40	2%	1 & 3	16%
	"personal contact" (78% accuracy)	Not used	-	-	
	See appendix	55 – 3,186	3 – 134%	1 & 2	Answer category dummy variables; no distinction between stages when mean of control was used (movement, destination); combined to one variable per answer category
Type of exploitation - Last Exploitation – Forced Labour	"forced labour", "labour exploitation", "forced to work"	233	1%	1 & 3	KWIC test to verify "forced to work" referred to forced labour (96% accuracy)
Type of exploitation - Last Exploitation – Sexual Exploitation	"sexual exploitation", "sexually exploited", "forced to work as a prostitute", "had to work as a prostitute"	174	<1%	1 & 3	

ADVANCED STATISTICAL ANALYSIS FRAMEWORK

A crucial basis of human trafficking prevention is the development of reliable sets of charachteristics, which can be used to identify potential victims. The task of detecting potential charachteristics is difficult because the process that drives persons into human trafficking is associated with complex patterns. Machine learning techniques allow to approach this challenge. The selection of this approach is justified by following three reasons:

Firstly, this approach allows the mapping of the relative importance of as many charachterstics as possible and display how these interact with each other (Breiman et al. 1984).

Secondly, some machine learning approaches have ad-hoc procedures built-in to handle missing data. This is an important aspect for the analysis as the data set contains many missing values (Ibidem).

Thirdly, this technique allows a highly inductive approach. The interest lies on new insights emerging directly from the data, rather than letting the empirical analysis be guided (and possibly 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 if different groups of victims may have markedly different degrees of variance (Ibidem)

The next sections explain two types of machine learning technique: cluster anaylsis and classification tree analysis. It concludes with a discussion of the model's main limitations. The purpose of this methodological part is to provide an overview of these two types of machine learning methodology, emphasizing its practical application rather than its underlying statistical theory.

Cluster Analysis

The application of cluster analysis served to get a first impression of the data. 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).

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 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 this analysis because it deals with a mixed dataset containing both numerical and categorical data what limits the 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 initial cleaning of the dataset and the adding of variables (e.g. variables classifying observations according to regions defined by IOM) coded for the analysis, 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, removing each observation, which contains at least one missing value (NA), would result in a dataset of zero observations. For this reason, only variables, which have a maximum of 60% NAs, were considered for the analysis. 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 3).

Table 3. NA-threshold for cluster analysis.

NA-Threshold	No. of variables, once NA's removed	No. of observations, once NA's remove
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, the analysis was conducted 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 this analysis a distance metric that can handle mixed data types was needed. The decision fell for the use of 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 k, whereas k = 1, …, p. Hence, δ_{ijk} equals to 1 when variable k can be compared for i and j, and 0 otherwise[6]. 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 4. 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).
- For nominal variables, s_{ijk} = 1 if individuals i and j agree in attribute k, and s_{ijk} = 0, if they differ in k.

Table 4. Scores and Validity of Binary Variable Comparisons

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

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

[6] *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, a hierarchical clustering algorithm was chosen for application. 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, the use of agglomerative clustering was decided.

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. This analysis uses 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-neighbour 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 ϵ X and y ϵ 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 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 produce 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 is 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 5 one could select 3, 6, or 8 clusters.

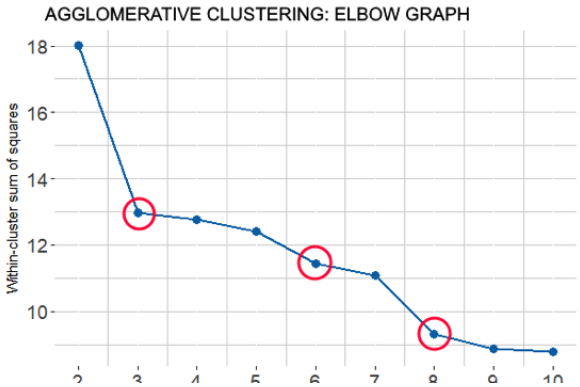

Figure 5. Agglomerative Clustering: Elbow Graph

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 6 shows the average silhouette width as a function of the number of clusters. When assessing the silhouette coefficient, the highest values should be chosen. Thus, 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.

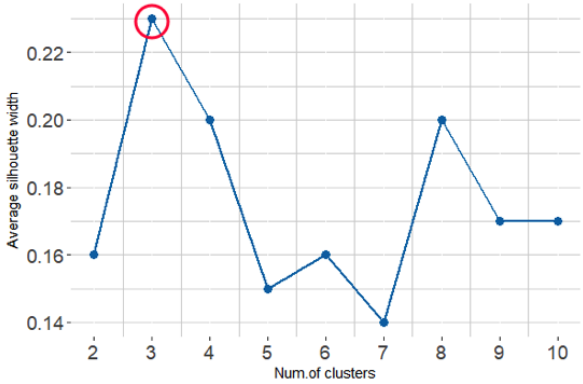

Figure 6. Agglomerative Clustering: Silhouette Graph

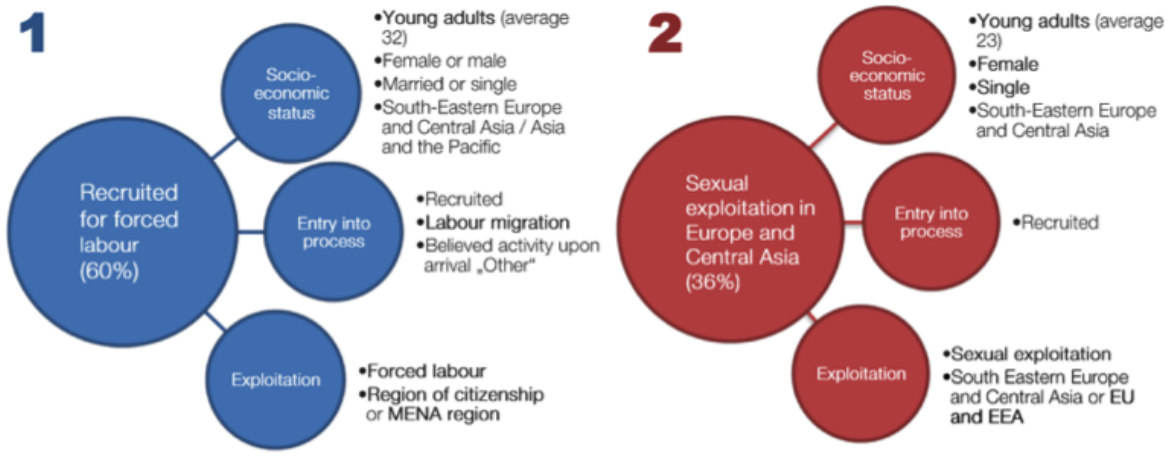

Figure 7. First and second cluster.

Results: Cluster Analaysis

The first cluster contains 599 out of 1,000 observations, making up 60% of the total sample (Figure 7). 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% to 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%).

The second cluster covers 365 out of 1,000 observations (Figure 7). 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).

The third cluster, seen in Figure 8, is much smaller than the first two, making up only 4% 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 difference between this third group and 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.

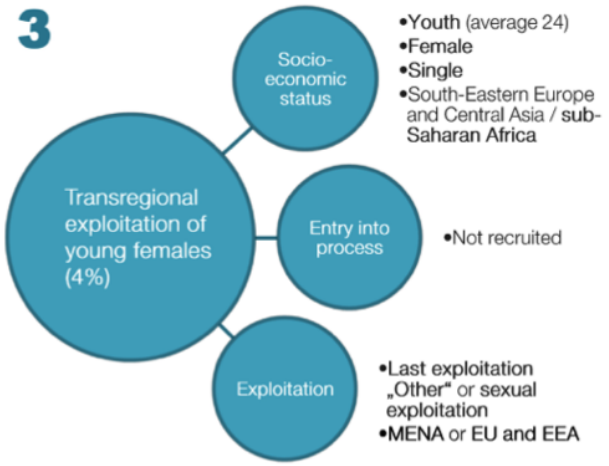

Figure 8. Third Cluster

Summary: Cluster Analysis

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. The authors believe that using hierarchical agglomerative cluster analysis has helped gain interesting insights into the main patterns in the dataset and allowed 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 compel to ask further 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 more nuanced, dynamic victim profiles were constructed, with the help of classification trees.

Classification Tree Analysis

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

Model: Classification Tree Analysis

The classification tree method 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). The analysis was organised 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 function is appropriate for the classification tree building process?

Step 3 Determining the number of splits: The pruning technique

Step 4 Assesisng decision tree model: How good is our decision tree model fitted?

Step 1: Preparing the data

In order to assess the model's performance, the dataset is divided 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, the dataset is randomly divided 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 20% that remains will compose the test set (Gareth et al. 2013).

Furthermore, two new variables as output factors were generated, representing different types of exploitation. The first measure is binary and stands either for sexual exploitation or forced labour. The second measure represents the types of exploitation according to national citizenship. For this variable, the output factor comprehends four possible categories: sexual exploitation took place in the country of citizenship, sexual exploitation took place in foreign country, forced labour took place in country of citizenship, or forced labour took place in foreign country. The focus on sexual exploitation and forced labour stems from the fact that these are the prevailing types of exploitation that emerged from the dataset. The input factors included all socio-economic attributes and various attributes of entry into process.

Step 2: Choosing an algorithm

The classification tree building process is based on the Gini-index measure, 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 and entry into the process attributes were misclassified according to the types of exploitation attribute. 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 i-th 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 a region of citizenship can be taken as an example for all our illustrated splits in decision tee models. The Gini-Index of this characteristics, which comprehends two categories yes or no, calculates the proportion of observation of "yes" and "no" which fall into either forced labour category or sexual exploitation category.

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

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

Step 3: Determing the number of splits

To calculate the number of splits the pruning method of "bias-variance" is used. The bias term refers to the error rate, which means how much the predicted values differ in average from the actual value. The variance term refers to a complexity 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 must add to the entire tree to warrant 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 9 shows the pruning process for the decision tree model of SEE and Central Asia as a region of citizenship, which can be taken as an example for all our illustrated decision tree models. The error rate is depicted on the vertical axis, while the horizontal axis depicts the complexity parameter and the equivalent number of splits. The Figure shows that as the number of splits increases, a reduction in the error rate occurs. The reason is that there is an increase of homogeneity within subgroups as the number of splits increases. However, a continually increasing number of splits generates an overtly complex model, also known as the problem of overfitting. Accordingly, all the decision tree models were pruned according to this horizontal line, which represents one standard error above the minimum of the error rate resulting in 8 splits.

After the pruning process, all victim characteristics presented in the main findings report have to fulfil further four criteria.

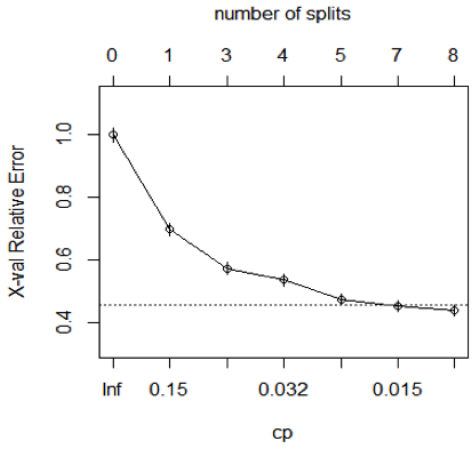

Figure 9. Pruning process regarding model for country of exploitation EU and EEA.

Firstly, only splits, which have at least a probability of 80% according to the types of exploitation, are considered in order to have highly reliable results. Secondly, only splits which are adequately represented in the sample are covered. Thirdly, victim characteristics, which are described by clear categories in each split (i.e. category not labelled as "other"), are selected in order to assure a more telling interpretation. Lastly, victim characteristics representing an interesting case in respect to existing assumptions on vulnerability (according to ILO, 2017) were highlighted.

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 a models' predictions. Generally, the performance of each decision tree model was evaluated 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 in relation to the real outcomes. Moreover, the performance of the decision tree model is assessed by using the test dataset in order to validating the classification tree model based on observations that are not part of the training dataset (Stehman 1997).

Table 5. Confusion matrix of the model for country of exploitation EU and EEA.

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

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.

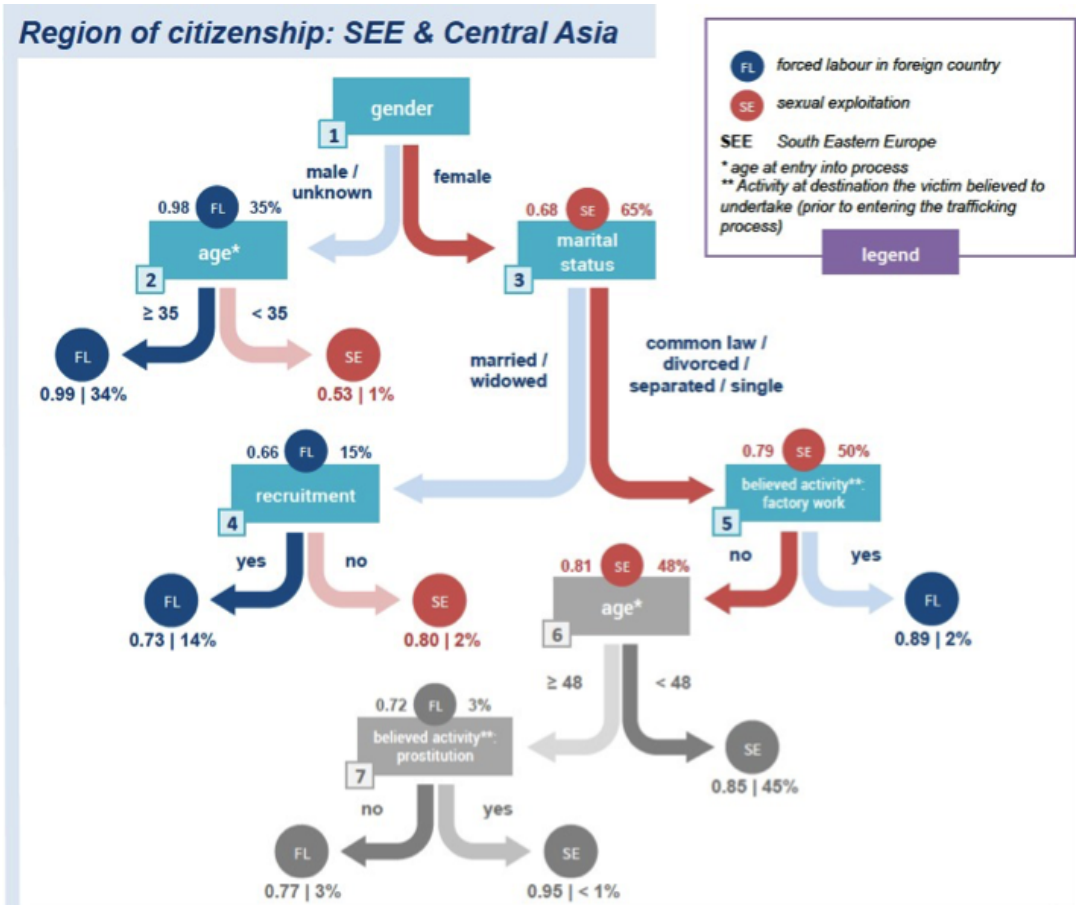

Figure 10. Decision tree model for region of citizenship SEE & Central Asia

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

Figure 10 illustrates the classification tree for South Eastern Europe (henceforth SEE) and Central Asia as a region of citizenship. For readability reasons only this decision tree model is discussed. The other decision tree models can be downloaded from the website.

The names inside the light blue rectangles are the characteristics of victims and the related types of exploitation are displayed inside each circle. FL stands for forced labour and SE stands for sexual exploitation. Below the circles two values are indicated. The left one indicates the predicted probability of the displayed subgroup of victims. The right one indicates the percentage of cases of the sample population for the selected regions, which fall in a particular group. In our example, the sample size for SEE and Central Asia is 11251 observations. The marked path emphasises the dominant type of exploitation in the sample.

At the top of the tree, the most influential characteristics to determine how to classify a victim according to types of exploitation are shown, while at the bottom of the tree, the less influential can be seen. It becomes apparent that the most influential attribute of this model is gender: a male has a probability of 0.98 to be involved in forced labour, whereas a female has a moderately high probability to be pushed into sexual exploitation.

However, the decision tree model illustrates a more complex pattern of victim characteristics for females in comparison to males (and unknown). With regard to males, split number 2 shows that a male aged 35 years or older has a higher likelihood to end up in forced labour than sexual exploitation, whereas a male who is younger than 35 years old has slightly higher probability of 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 cases.

With regard to females, the split number 3 shows that the most decisive characteristic for women is the marital status according to types of exploitation. It is necessary to mention that the probability of a female to be pushed into sexual exploitation rather than forced labour increases to 79% if she is in either common law, divorced, separated or single. By contrast a female who is divorced or widowed has a slightly higher probability to be pushed into forced labour rather than sexual exploitation. For this subgroup, whether a woman 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 widowed female to be pushed in forced labour rather than sexual exploitation increases to 0.73 if she is recruited; otherwise she has a probability of 0.80 to be forced into sexual exploitation rather than forced labour. The former case is covered by 14% of the sample size, whereas the latter case is covered by only 2% of the sample size. Note that the probability of a not recruited married or widowed female to be pushed in sexual exploitation rather than in forced labour is slightly higher when compared to the probability of a female who is in common law, divorced, separated or single.

Going back to the split number 3 (marital status) and following the marked path further down (split number 5), the decision tree model shows that believed activity regarding factory work is an important characteristic for this 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 forced labour increases to 81% if her believed activity was not factory work before entry into the process. By contrast, the same female has a probability of 89% to be pushed in forced labour rather than sexual exploitation if her believed activity was factory work. The former subgroup is represented by 48% of cases in the sample, while the latter case is represented by a small subgroup of 2% of cases.

Splits number 6 and 7 are coloured in grey since we want to emphasis that further splits would tend to design an overfit of all cases in the sample. The decision tree model ends up with subgroups which are represented by a 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.

Summary: Classification Tree Analysis

This section discussed the classification tree analysis as a supervised machine learning approach. Its first part provided an in-depth insight into the methodology of the classification tree approach in order to elucidate how the main findings were generated. As in the cluster analysis, classification tree analysis involves various steps. The second part showed how to interpret the classification tree analysis with the help of an example for South Eastern Europe and Central Asia as a region of citizenship. For these regions, the results confirm the prevalence of common characteristics, as age and gender, which strongly determine the type of exploitation. Furthermore, our analysis emphasized additional victim characteristics such as marital status, what the individual believed to work as at destination and how the contact to the recruiter was initiated. These findings give a nuanced understanding of victims who are pushed into human trafficking according to their type of exploitation.

LIMITATIONS

As the previous sections of this chapter have shown, the empirical strategy was based on considerations of how to deal best with the available data, the structure of the data, and the complexity of this 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 the data, performing hierarchical clustering on a dataset as large as the one applied here is computationally very time intensive. For this reason, the results of the 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 the decision tree technique is easy to understand and simple to use, offering many advantages compared to other statistical techniques. However, some limitations became obvious during the application of the decision tree method.

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

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. The applied pruning technique in order to reduce the complexity of the tree models was illustrated in previous sections. Moreover, the pruning mechanism was combined with the confusion matrix in order to have an optimal trade-off between accuracy of the models and the complexity of the models.

CONCLUSION

In the fight against the phenomenon of human trafficking, the necessity of an in-depth understanding of victim characteristics becomes evident. Besides complementing the understanding of such victim characteristics and how they relate to type of exploitation, the analysis showed how the application of advanced analysis techniques could be advantageous in the field of human trafficking research.

The applied methods revealed empirical evidence on the perception of potential victims. The highly inductive approach enabled the identification of new victim characteristics beyond the confirmation of pre-known characteristics. Further, the machine learning strategies enable the mapping of complex interactions among such characteristics. It reveals the relative importance of characteristics depending on different contexts and in various combinations with other characteristics. Depending on the combinations a certain characteristic can have diverging effects on the type of exploitation. Therefore, the method takes account of changing contexts over time, like changing business models and recruitment methods by traffickers.

Classification tree analysis has also the advantage of dealing well with datasets containing many missing values. In the field of human trafficking research with often very limited access to information, this might be of special interest. Consequently, machine learning techniques showed that structured quantitative empirical analysis has much potential to support counter-trafficking efforts.

However, the importance of extensive, high quality, and standardized data in assessing the explanatory power and validity of empirical findings through these methods cannot be stressed enough. Limiting factors faced during the analysis were the high amount of missing information, non-standardized categories, non-exhaustiveness of the question design and inconsistency in the processing and understanding of the information.

The translation into a complete and consistent dataset is one of the fundamental interests to perform precise analysis, which then provide information relating to the development of effective prevention strategies on potential victims. Following recommendations would facilitate the application of machine learning techniques and improve the applicability of the findings in prevention work:

- reducing the scope of the survey to the important and highly relevant questions to increase the focus on them and ensure their complete answering
- standardizing and complementing of answer categories of certain questions to improve their ability to reflecting all the cases correctly on a distinct level of detail. For example, all questions on „activities“ (professional occupation) should include standardized and unified answer categories to ensure a comparison on the same categorical level.
- clarifying key concepts (e.g. recruitment, initial contact to recruiter, etc.) in the case worker trainings to make sure interviewers and interviewees have the same understanding throughout the whole survey
- adapting the registration software to read in the data with mandatory fields, drop-down menus and conditions to better guide the processing of the interview information and prevent unnecessary errors
- avoiding unranked multiple answer questions as it is much more difficult to interpret machine learning outputs based on them. Design questions by ranking first, second and third experience, reasons or occupations.

Keeping these limitations in mind, an elaborated data collection strategy combined with advanced statistical techniques constitutes a promising starting point for designing reliable counter-trafficking solutions, their implementation, and the monitoring of the policies' effect. Research can only benefit and evolve from the application of new, innovative methods. In the context of human trafficking in particular, the promising potential of the synergy between field expertise and modern analytic strategies derived from the world of big data remains yet untapped. For example, countries or regions could be grouped according to various policy frameworks to measure which socio-economic factors interact with macro conditions and how they affect policy effectiveness. Machine learning methods can also be a helpful tool to optimize the data handling process and improve common issues such as duplicated or inaccurate data. Machine learning algorithms can serve as a warning system by providing ongoing assessments of the data and detect anomalies. Furthermore, machinine learning can facilitate the data entry process by automating time-intensive tasks.

REFERENCES

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.

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 Drugs 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). (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