Machine Learning in the Analysis of Social Problems: The Case of Global Human Trafficking

التعلم الألى في تحليل المشاكل الاجتماعية: دراسة حالة الاتجار بالبشر علي الصعيد العالمي

by

ARSENIO BALUYOT CAOLI, JR.

Dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS

at

The British University in Dubai

September 2019

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

_____

Signature of the Student

COPYRIGHT AND INFORMATION TO USERS

Abstract

Machine learning has been key to significant information technology discoveries in myriad disciplines. However, it has received mixed outlook in the social science field. This study aims to use the methods of learning from real data set on human trafficking, which is a serious social problem of today. The Counter-Trafficking Data Collaborative (CTDC) dataset, which is an initiative of the International Organization for Migration (IOM) for human trafficking was used for the experimental study. The exploration of the dataset revealed 61% of missing data — another incentive for the applicability of machine learning via multiple imputation using chained equations (MICE) instead of single imputation or deletion. Agglomerative hierarchical clustering using Gower's Distance was used for pattern discovery of the categorical type of data in this research, with a comparison to Fuzzy k-mode clustering. Results show that MICE had a level of effectiveness in handling missing data, while agglomerative hierarchical clustering was successful in identifying distinct and describable clusters from three time periods that the imputed dataset was segmented.

Keywords: human trafficking; machine learning; multiple imputation by chained equations; MICE; agglomerative hierarchical clustering; pattern mining

الملخص:

ساعد التعلم الألى بشكل كبير في اكتشافات تكنولوجيا المعلومات في مختلف التخصصا. بيد انها تلقت اراء متباينة في مجال العلوم الاجتماعية .والهدف من هذه الدراسة هو تطبيق أساليب التعلم من مجموعه البيانات الحقيقية المتعلقة بالاتجار بالبشر ، وهي مشكله خطيره في الحاضر. استخدمت مجموعه بيانات CTDC ، التي هي مبادرة من منظمه الفوائد الوطنية للهجرة من أجل الاتجار بالبشر ، في الدراسة التجريبية .كشفت مجموعه البيانات عن أن 61% من المعطيات المفقودة، وهو دافع آخر للمعالجة باستخدام التعلم الألى عن طريق الإسناد المتعدد باستخدام المعادلات المتسلسلة. وقد استخدمت أجلوميراتيفي التسلسل الهرمي باستخدام المسافة Gower لاكتشاف نمط من نوع القطعية من البيانات في هذا البحث ، مع مقارنه لنظام المجموعات الوضع k غامض. وفي الوقت نفسه ، نجحت التكتلات أجلوميراتيفي هرمية في تحديد مجموعات متميزة من ثلاث فترات زمنيه لمجموعه البيانات المنسوبة.

Acknowledgment

My most thoughtful gratitude to all the people who believed, supported, and motivated me throughout the whole process — my family, friends, academic circle, and my colleagues. The fastest way to achieve this dream is on my own, but with all your support I have gone farther.

Prof. Sherief Abdallah — you have been supportive of my research visions from the start and for that I am thankful. *Rian Gabor* — the sleepless nights and sacrifices were not worth it without you in the journey. I won't be able to repay all the insights you have given me right from the start. *Chloe Smith* — thanks for reminding me of the meaning of this dream and that home is not a physical entity. Lastly, I dedicate this to my *mom* — you taught me the value of education and to keep in mind that it is not just the knowledge you gained but what you do with it. May you rest in peace.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
AGNES	Agglomerative Nesting
AUC	Area under curve
CTDC	The Counter-Trafficking Data Collaborative
DARPA	Defense Advanced Research Projects Agency
FCA	Formal Concept Analysis
IOM	International Organization for Migration
KNN	K-Nearest Neighbors
LR	Logistic Regression
MI	Multiple Imputation
MICE	Multiple Imputation by Chained Equations
ML	Machine Learning
NLP	Natural Language Processing
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
UN	United Nations
UNODC	United Nations Convention Against Transnational Organized Crime

1. Introduction

1.1. Background

Machine learning has proven its success in the information and technology field and other disciplines that it encompasses. The uptake on the significant use of machine learning has been evident not just in the information and computer science industry, but also in healthcare, business and finance, government, biological and natural sciences, to name a few (Jordan & Mitchel 2015; Franco-Arcega et al. 2014). Machine learning particularly helps researchers provide more applicable and conclusive solutions using predictive analytics i.e. estimations of future or unknown values using variables in the data; and clusters, patterns, and relationships among data. Relative to previous decades, large numbers of data are more conveniently available nowadays hastening the data-driven decision-making and insight generations, helping different types of organisations in the process (Grimmer 2014).

Data mining and machine learning according to Hindman (2015) was said to have also caused the social sciences field to be revolutionised, broadly affecting its conventional process, methods, and analytical techniques. Machine learning delivers new methods of data analysis that should influence social scientists' methods instead being overlooked (Dhar 2012; Hindman 2015). This transformation in the discipline is incredibly meaningful for the branch of learning that potentially helps in innovating solutions to social problems.

One of the most serious social problems of today is human trafficking, which is also dubbed as "modern day slavery". Human trafficking or trafficking of persons is a concerning crime due to its geographic coverage, threat to public

health, and violation of human rights (Ford, Lyon & van Schendel 2012). Human trafficking encompasses gender, age, and nationality, in which cases are mostly trafficked for forced labour and sexual exploitation (IOM.int 2016). Currently, the main international organisation dealing with human trafficking is the International Organization for Migration or IOM. IOM is the leading international organisation for migration management, and its effects on social and economic development, while maintaining migrant well-being (IOM 2019). One of IOM's key strategic focus in fulfilling their mission is undertaking programmes and expertise in combatting trafficking in persons. This is supported by their campaign in the IOM X platform with the message of encouraging "safe migration and public action to stop exploitation and human trafficking" (IOMX.IOM.int 2019). In line with their strategic focus, the Counter-Trafficking Data Collaborative (CTDC) was created, which is an initiative of IOM as the first collaborative data hub on human trafficking. CTDC is simply put as the "human trafficking data portal" with the intention to help stakeholders identify and respond to trends in human trafficking though their expertise or call for expertise (CTDatacollaborative.org 2019).

The CTDC online platform is a repository of anonymized human trafficking cases or records with multiple variables including age, gender, nationality, exploitation type, among others, containing categorical variables with binary or multinomial values and missing data. This study uses the CTDC data for exploratory and descriptive machine learning, which to the knowledge of the author has not been currently researched empirically and published. Human trafficking has been a topic covered by many social researchers and the UN, but not in a significant amount that used the techniques of machine learning. A

noteworthy number of experimental papers have been published addressing the human trafficking problem using machine learning, but most of these research works used textual data from online advertisements of sexual exploitation type. Thus, these references for human trafficking are technically focused under the text mining and natural language processing (NLP) category.

Missing values in datasets are in a real world inevitable and imposes problems in data analysis (Cismondi et al. 2013). Researchers for the longest time address these via deletion of records with missing values that reduces the generalizability of the data or single imputation techniques that are bias-prone and technically limiting. This study addresses the conundrum of missing values using predictive machine learning or particularly Multivariate Imputation using Chained Equations techniques.

As mentioned by Franco-Arcega et al. (2014) and Dhar (2012), and as generalized by Hindman (2015), experts in social science research conventionally use statistical techniques or for most studies employing regression techniques in understanding their data. However, to benefit from the advent of machine learning, this research work is dedicated to use both supervised and unsupervised learning to generate insights and observe pattern, depicting usability of machine learning in social research context.

1.2. Research objectives

Recognizing its significance and the current status of machine learning in social science, this study aims to assesses the applicability of machine learning techniques using the data by the Counter Trafficking Data Collaborative (CTDC) of the IOM, otherwise treated conventionally by social researchers. By doing so,

this study also aims to identify clusters or discover patterns in the dataset. Specifically, this research aims to:

1. employ multiple imputation techniques using predictive or supervised learning in the treatment of missing values;
2. cluster or use unsupervised learning for the categorical variables in the dataset, with comparison of methods;
3. provide insights on human trafficking out of the CTDC dataset.

From both technical and conclusive standpoints, this study is aimed to demonstrate:

1. that machine learning is highly functional and beneficial in social science research of real-world dataset confirming its potential to the field;
2. that there are distinctive groups or clusters identified in the human trafficking dataset;
3. that there is dynamism in human trafficking trend, as described from the results of the learning techniques.

1.3. Dissertation structure

Coming from establishing the research background, motivation, and objectives in Chapter 1, the Chapter 2 of this paper discusses machine learning and its significance in today's knowledge discovery and current usage in human trafficking research. This is followed by the Chapter 3 on CTDC dataset description, and the methods used. Results of the experiment and the interpretation and discussion of the findings were followed suit in Chapter 4. Finally, the culmination of the objective achievements is in Chapter 5, trailed by the researcher's prospective research outlook on the topic (Chapter 6).

2. Literature Review

The introductory chapter established the popularity of the field of machine learning as a remarkable feat to practical and research aspects of modern day. Indeed, the study of machine learning has purveyed a significant amount of historical contributions by using computers and learning systems (Jordan & Mitchel 2015).

This chapter contains an overview of machine learning applicability and followed by a highlight of the status of machine learning in the social science field. The latter cites how social scientist and experts have mix perceptions on machine learning applicability to social research, which was a crucial point for this study to address. To do so, and by addressing the applicability conundrum, this study as mentioned empirically focuses on human trafficking, with dataset by the CTDC. Rightly so, an overview of the state of human trafficking was included in this chapter, followed by human trafficking-centred research studies using machine learning over the past 10 years.

2.1. Machine learning

In the paper written by Jordan & Mitchel (2015) discussing the trends, prospects, and perspectives in machine learning, it was mentioned that the rapid and significant popularity of machine learning over the years was credited to its sensible applications, extending from the medical field to business and development use. In the field of computer science and information technology, much of its credited applications prevail in robotics, natural language processing (NLP), image recognition and speech recognition. Machine learning has also seen

in a more encompassing fields such as system analysis, supply chain, biology, and social sciences.

The techniques of machine learning are in sum categorized into types namely prediction and clustering. Prediction or supervised learning is the method of estimating values deemed unknown using data containing attributes or variables. On the other hand, descriptive or unsupervised learning is exploratory in nature and a key technique in identifying patterns, finding relationships, or delivering summaries (Han, Pei & Kamer 2011). Supervised learning uses existing data on some problem through training. A labelled dataset is used to learn or train the algorithm in identifying or classifying according to that label. In medical diagnostics, it is commonly used in identifying "with disease" or "without disease" for new patients based on their health records and features. In the credit industry, "fraud" and "not fraud" labelled historical datasets are then used to train and identify new applications on whether there is a risk in committing fraud or a likely candidate for approval. Both medical diagnostics and credit-risk examples are just a few applications in supervised learning techniques, that are then evaluated using a performance metric. Measuring an algorithm's performance is essential to see if the model is identifying records accurately. In unsupervised learning, clustering is the important data mining tool in pattern recognition. In clustering, observations or objects are partitioned into groups or clusters, where the objects in a group are more similar to each other than from the other groups. In the medical field, unsupervised learning can be performed using the attributes to group observations based on gender, age, diagnostic results, treatments, or even genetic predispositions. Unsupervised learning is very helpful in exploring datasets that

are not labelled. The identification of groups has the potential of producing a labelled data set, which can then be used for supervised learning. (Sharma & Gaud 2015; Lison 2015).

Nowadays, there is a myriad range of machine learning algorithms to address any learning needs and data challenges. Algorithms like decision trees, Naïve Bayes, or gradient boosting have been designed to address the different nature and acquisition of data in predictive data mining. On the other hand, algorithms like k-means is a base point for clustering numerical data types, but not for categorical data – in which k-modes or hierarchical clustering type are more useful (Jordan & Mitchel 2015).

2.2. Machine learning and social sciences

As mentioned previously, some of the well-known applications of machine learning that proves its success are prevalent in natural language processing (NLP), robotics, computer applications, among others. It is highlighted by Franco-Arcega et al. (2014) that data mining and machine learning techniques impressively contribute to social environments and research. On the other hand, the emerging fields as stated by Jordan & Mitchel (2015) that are also benefitting from machine learning include psychological studies, educational practices, organisational behaviour, and economics, with the last two being part of the social sciences.

Social science is the study of human society and social relationships that includes politics and economics (SSRN.com 2019). Like any researcher, social scientists rely on data for their studies. In recent years, as the discipline of machine learning and the emergence of "big data" boomed, the promise to

revolutionise the answers to social science inquiries have gone with an interest as much as in other disciplines (Grimmer 2014). In a paper by Grimmer (2014) however, many social scientists negate these claims of Jordan & Mitchel (2015) and Franco-Arcega et al. (2014), finding that data science claims are "over-the-top" and exaggerated. This is because data science is mostly unfitted and inexperienced in the challenges of "social scientific inquiry" to say the least. For instance, social scientists are extensively experienced in using observational data and making inferences from samples out of a social population in which control is absent due to ethical issues. Likewise, the studies social scientists conduct on societal inquiries are generally data extensive rendering large datasets inadequate due to their mostly narrow depth. Another angle where machine learning was criticised is in the paper of Lipton & Steinhardt (2018), citing that machine learning fails to differentiate explanations, an important social science research inquiry, versus speculations delivered by ML-specific experimentations.

In this context, social scientists were said to potentially benefit more from the advent of digital and large number of generated data, highly in demand nowadays from various points more than the techniques of machine learning. Another point of view by Grimmer (2014) suggests that large datasets used in machine learning should be taken as a supplement to observational data in enhancing the causal inferences that social scientists usually make. Overviewed in the previous chapter, Hindman (2015) affirmed how social scientists are traditionally for the longest time accustomed with the mere use of statistical analyses and regression, and as such contrasts the formerly mentioned apprehensions in the data point of view. This is supported by Grimmer (2014), adding that causal inference tools that social scientists traditionally use has an

equivalent in the machine learning field, which already has improved the studies of a few researchers. The combination of machine learning and causal inferences is now a fast-growing method in political science to site an example (Roberts et al. 2014; Fowler et al. 2011). Much like the other industry discussed earlier, the use of machine learning is highly usable in clustering or predicting sample or population responses to certain issues.

Rightly so, data mining and machine learning has had success in research papers for social problems, including demographics and population segmentation, education, labour, and unemployment. Franco-Arcega et al. (2014) had also proven the usefulness of data mining and machine learning in the analysis of social problems in their migration study. To start with, a social problem usually involves a large population deeming a certain social situation undesirable, and thus covers the mentioned successes. Explicitly, human migration is a socio-spatial occurrence, resulting from several changes in the to-and-from places, social and spatial make-ups. In their research, Franco-Arcega et al. (2014) focused on a migration case study where a collected dataset from a government institution in the State of Hidalgo in Mexico about external migrant workers, describing their data as more "real-word" and non-digital. To the researchers, the study of migration is an important social issue, as it is highly influential in the shaping culture, economy, and demography. A descriptive learning through cluster analysis through simple k-means, Density-Based Clustering, and Self-Organising Maps, uncommon to usual social science research. The results of this migration case-study were intended for helping State of Hidalgo and their government in identifying patterns and the creation of social programs to cushion the negative effects of migration. Method-wise the acquired

results were deemed by the researchers to have pronounced importance to migration subject matter specialists, as it veers away from conventional or on the very least, manual analysis. However, with the premise of machine learning having also the automation capability, it is noteworthy feat (Franco-Arcega et al. 2014). More than the case study region of the State of Hidalgo, machine learning process and automation process lends faster generation of results as more cases of migration grows, not just in the state but also in the country, and more so in the global scale. As a whole, data mining and machine learning are significantly helpful in the characterization of behaviours, and the discovery and understanding of patterns related to crime, trafficking, poverty, among others.

In this study, the feasibility of machine learning is focused on human trafficking, which is another serious social issue of the modern day, using human trafficking casework dataset from the CTDC.

2.3. Human trafficking

Human trafficking is one of the more serious social problems in our modern day. In fact, many have coined human trafficking as "modern day slavery" (Ford, Lyon & van Schendel 2012). To formally define, the United Nations Convention Against Transnational Organized Crime (UNODC.org 2004, p. 42) (eventually named as Transnational Crime Convention) assigned a protocol called the UN Trafficking Protocol, defining human trafficking or "trafficking in persons" as:

> recruitment, transportation, transfer, harbouring or receipt 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. Exploitation shall include, at a minimum, the exploitation of the prostitution of others or other forms of sexual exploitation, forced labour or services, slavery or practices similar to slavery, servitude or the removal of organs;

This exploitation is often the commercialisation of humans for the monetary benefit of the abusers. The process of trafficking humans has become an intricate societal problem, that violates human rights, a threat to the global public health crisis and undermines sense of global order (Caraway 2006; Ford, Lyon & van Schendel 2012).

The incidences of human trafficking according to (International Organization for Migration 2016) transcends economic status of a country, thus are prevalent in both wealthy and poor countries. The Global Slavery Index of 2016 and Trafficking in Persons Report of 2016 have estimated about 27-45.8 million individuals worldwide are under some form of human trafficking or modern-day slavery and is rapidly increasing (IOM.int 2016). According to Ford, Lyon & van Schendel (2012) the prevalence of human trafficking is mostly a covert activity, and as such difficult to measure hence the huge incidence range.

This incidence estimate is not unforeseen, since the commercialization of humans is a highly lucrative business to most exploiters. The two most common types of exploitations, which are forced labour and sexual exploitation alone are estimated to value for about USD 150 Billion of black-market profits globally per year (Konrad et al. 2017). This figure pegs human trafficking behind illegal substances or drugs and weapons trading when it comes to profit from organised crime (Haken 2011). Research has estimated that forced labour and domestic trafficking constitute about 68% of the human trafficking victims globally

(Konrad et al. 2017). Human trafficking victims range from minors to adults, subjected to domestic work or housekeeping, labour in commercial industries, pornography, exotic dancing.

Traffickers hold victims under their command to carry out jobs or tasks unwillingly and out of their consent. As these jobs or tasks are considered illegal, traffickers resort to controlling their victims with abusive means that include threats and other psychological control, physical violence or restraints, financial restrictions, and coercion (Konrad et al. 2017).

Human trafficking, by definition, is closely linked to migration due to the transport nature of committing the trafficking (Ahmad 2008; Asis 2008; Mahmoud & Trebesch 2010; Ryazantsev et al. 2015). In a paper by Ryazantsev et al. (2015) on the problems of human trafficking and illegal migration focusing on Russia, the researchers concluded that there is a strong link between human trafficking and irregular labour migration. The researchers stated that aside from the intrinsic fondness of a large quantity of people from neighbouring Russian countries into their country, Russia's geographical location has also been a transit strip between Asia and Europe through which trafficking cases prevail.

Another paper by Mahmoud & Trebesch (2010) also supported the migration and human trafficking link by arguing that trafficking and exploitation are apparent effects of migration in an opportunistic and global scale. The researchers expanded on this growing social consequence citing legal courses of migration have limited the options of people seeking movement and relocation, which exploiters and criminal organisations use to take advantage by profiting. All these exploiters and criminals need to do is capitalize on the desire to work

abroad of these prospective migrants who are willing to an extent of sacrificing human rights and freedom of movement.

2.4. Machine learning in human trafficking research

Over the last 10 years, seven studies were found to have experimental research on human trafficking employing the tools of machine learning. Of the seven studies, five of them were all experimental research specifically tackling sexual trafficking (Kejriwal et al. 2017; Dubrawski et al. 2015; Hundman et al. 2018; Alvari, Shakarian & Snyder 2017; Kejriwal & Szekely 2018). These five studies fulfilled their experimental research through data from online advertisements perpetrating sexual or escort services. Three of the five studies mentioned (Kejriwal et al. 2017; Hundman et al. 2018; Kejriwal & Szekely 2018) were prompted by the DARPA MEMEX program. DARPA or Defense Advanced Research Projects Agency is a government agency of the United States Department of Defense which collaborates with various industries and academe for research and development for national security (DARPA.mil 2019). MEMEX on the other hand is a program under DARPA which aims to develop advance online search capabilities beyond the common web or into the deep web, which usually contains information that threatens public security (DARPA.mil 2019). The other two studies, scoping on general human trafficking are experimental (Poelmans et al. 2012) and a position paper/meta-analysis (Konrad et al. 2017). All six of the experimental research papers used text mining and natural language processing in their studies with a using a variety of machine learning techniques, models, and architecture.

As one of the sexual trafficking and text mining studies, Kejriwal et al. (2017) emphasized that the convenience of sharing and usability of the internet and the Web has also tipped the rise of illicit activities including human trafficking, manifested through online advertisements particularly for sex exploitations. A study by Kejriwal et al. (2017) used natural language processing and semi-supervised learning in addressing the uptake on these online advertisements. Natural language processing is the use of machine learning algorithms to read and understand human languages, often with the help of a collection of texts related to a domain called corpus or a deciphering dictionary. In this study, the researchers used a corpus collected under the DARPA MEMEX program which has the collection of hundreds of millions of online sex advertisements with significant content related to human trafficking. The researchers' solution is to tag advertisements that has the human trafficking potential using what they call as the Flag-It system, which eventually guides authorities for investigation. Results of the study showed that on preliminary evaluations, Flag-It has exhibited promising performance compared to its alternatives. The researchers established their research and other studies under the DARPA MEMEX program is an example of social and computer science interdisciplinary success over recent years.

Another study similar to the previously discussed is by Kejriwal & Szekely (2018), recognizing also that although the Web has positive impact to many, it also paved the way to an influx of illegal activities advertising including sexual trafficking. Extending the work of Kejriwal et al. (2017) on the DARPA MEMEX program, this study aimed to identify budding victims of trafficking latently identifying missing people and runaway cases in the process. From an

institutional authority and investigative point of view, a normal search in Google or any search engine is inadequate in identifying these illicit advertisements. These due to factors of ambiguity inherent to our machine and the deliberate attempts of exploiters to dodge any investigation in their advertisements. To solve this problem, the researchers, who are from the USC Information Sciences Institute have proposed and developed a system called Domain-specific Insight Graphs (DIG) to provide effective and efficient identification. DIG is composed of information extraction technologies, indexing techniques, caching, featuring or attribution algorithm, deep neural network (for images), and big data architecture. DIGs results were compared to an existing model and architecture called TellFinder, also developed under the MEMEX program. DIG and TellFinder in the end have varying results compared to each other, which the authors attribute to issues on accuracy evaluation and were deemed an interesting problem to be tackled on the baseline.

Like Kejriwal et al. (2017) and Kejriwal & Szekely (2018), Hundman et al. (2018) also leveraged the DARPA MEMEX program in their more recent study. Aside from looking at the importance of human trafficking cases detection in the escort related services, the researchers' study also examines methods in alleviating biases that are inherent to human trafficking detection with textual data. The researchers developed an intelligent detection system for over three years of research under the DARPA MEMEX program to incorporate data from various institutions. An architecture was developed by the Hundman et al. (2018) to achieve their binary classification system model (human trafficking-related and non-human trafficking related). The researchers' model follows a standard text mining approach, which includes: data collection, feature extraction, manual

labelling, sampling to address scarcity, clustering for training, featurisation through vectorisation, binary classification, and evaluation through looking at ROC AUC. The researchers also recognized that biases due to algorithmic violation or training data being too small/inherently partial can also arise. In such cases, mitigating biases were applied and presented in their paper. Improvements of the model can therefore aid the New York law enforcement and district attorney's office in their operations against human trafficking.

Another study on human trafficking using text mining and NLP, not to mention a focus on sexual trafficking, is by Dubrawski et al. (2015) that uses publicly available online advertisements to classify human trafficking by aiming for the contexts. The researchers proposed different models of supervised machine learning to detect escort services related advertisements. This was a challenging task for the researchers since online advertisements alone lack the distinguishable words that will flag the escort services related advertisements. In the natural language processing term, this challenge is the features sparsity. This is because of the illicit nature of human trafficking and thus leads to concealment of highly palpable keywords in the perpetrator's advertisements (Konrad et al. 2017). In their methods, keywords, regular expressions, and an unsupervised feature created by a natural language processing technique were used to train using Random Forest classifier. Validation of results were done by looking at the AUC, which gave a high score of 96.6% and a recall of 79%. Similar to Kejriwal et al. (2017), the study's goal is also to help enforcers in isolating and tracking leads in human trafficking of sexual and escort services nature and wherein the machine learning model is being continuously developed to discern illicit advertisements better.

Alvari, Shakarian & Snyder (2017) is the last in this chapter to tackle sexual trafficking also using text mining in their experimental work, leveraging data from a website called "Backpage" which is a classified site for advertisements. Their objective was to discern human trafficking activities, particularly escort services-related, from these advertisements using semi-supervised learning techniques. Faced with a corpus challenge in their study, the researchers relied on hand-labelled part of the sourced data by an analyst from the law enforcement field (manual tagging). S3VM-R or the extension of the Laplacian SVM was used to improve the features, and then trained the data for learning using KNN and other algorithms like Logistic regression, AdaBoost and Random Forest for the supervised learning portion of their process. The models were then evaluated using a 10-fold cross validation, looking at the AUC. The study's result has proven a better learner in comparison to both existing supervised and semi-supervised approaches.

The study by Poelmans et al. (2012) is different from the previous studies. The researchers' premise on the status quo of the field begins with although automation of discovering patterns and knowledge is highly important, it has a drawback of undefined fundamental concepts and features. This, according to the researchers, is simply because automated tools are often not human-centred, meaning that interpretations should still be established from the natures of human complex human thought. The central idea and its importance were verified by the researchers through an experimental study using text documents containing police reports in Amsterdam, Netherlands to aid researchers' model in detecting human trafficking cases. Formal concept analysis (FCA) or the formal ontology in text mining was used in the automation of knowledge discovery on

provide interactive approach to gaining insight. This semi-automated method ensures that the results are still subject matter-supervised and human-centred. In their experiment's case, Poelmans et al. (2012) verified the method with the police officers in Amsterdam-Amstelland police in Netherlands and were given initial indicators to proactively isolate suspects instead of reactive.

The paper by Konrad et al. (2017) is the only non-experimental research in the selection. The researchers made a critical survey and meta-analysis of the potential use of data science in overcoming human trafficking. The researchers recognise the difficulty of detecting and tracking human trafficking in general due to its covert nature. However, with the high potential of the quantitative fields, particularly operations research and analytics, the researchers highlighted how the problem on detecting and tracking human trafficking cases can be eased up. In their paper, Konrad et al. (2017) pointed out that these quantitative tools can provide understanding of the occurrence of human trafficking, prevent cases on the enforcement operations side, and eventually amend policies. There are on the other hand some limiting factors to the idea of using analytics in addressing human trafficking. One of which is the lack of access to the cases or human trafficking victims since the traffickers are running operations covertly. Another hindrance is that traffickers' behaviour and pattern of operations are becoming increasingly dynamic to avoid exposure. The insufficiency of resources and general lack of comprehensive data were also discussed to be a limiting factor. The research of Konrad et al. (2017) provided a literature standpoint for challenging the quantitative and analytics field in battling human trafficking and integrate the social science discipline in the process.

Author	Domain	Special Program	Technique
Poelmans et al. 2012	General	-	text mining; semi-automated learning
Dubrawski et al. 2015	Sexual Trafficking	-	NLP, Supervised Learning
Konrad et al. 2017	General	-	Critical Survey
Kejriwal et al. 2017	Sexual Trafficking	MEMEX	text mining; NLP; architecture
Alvari et al. 2017	Sexual Trafficking	Backpage	text mining; semi-supervised learning
Hundman et al. 2018	Sexual Trafficking	MEMEX	text mining; NLP
Kejirawal & Szekely 2018	Sexual Trafficking	MEMEX	text mining; NLP; architecture

Table 2.1. Summary of papers on human trafficking using ML

3. Methodology

This chapter discusses the dataset used in this experiment, and the methods from data processing, techniques, and concepts behind the methods that will be applied to the dataset. It is therefore important that the dataset was examined first on an exploratory level, where in necessary pre-processing tasks were taken. This was especially imperative due to the presence of missing values, which will be discussed in the next portion. The assessment of the data provided the direction on the treatment of the missing values on a predictive level. Ultimately, an imputation process was deemed necessary to process the data, and the rationale behind the chosen imputation technique. After which, imputed dataset was sampled before proceeding with unsupervised learning or clustering for three different time periods of the dataset to analyse trends.

3.1. The Data

The dataset was obtained from the Counter-Trafficking Data Collaborative (CTDC) site, which is an initiative by the International Organization for Migration of the United Nations launched back in November 2017. CTDC is the first collaborative data hub initiative on human trafficking that publishes from various counter-trafficking organisations globally, including the International Organization for Migration, Polaris and other partners. With the global drive to make data more publicly available, CTDC's aim has aligned with the current trend, as well as provide researchers and other stakeholders with human trafficking

data that is updated, reliable, and has a broad coverage of access. The CTDC initiative addresses the previous challenges of having difficulty in accessing human trafficking related data for researchers, policy-makers, and scholars and provide resolutions ahead of the social problem at a faster rate. Historically, human trafficking data are difficult to access due to its covert nature and the sensitivity that comes along with it (CTDC, 2018 and Konrad et al. 2016). The latter issue of sensitivity often raises privacy, human rights, and civil autonomy concerns, which could negatively impact the victims even more. The sensitivity issue was therefore addressed by CTDC with de-identification of records or anonymization. The human trafficking dataset by CTDC gets updated cumulatively and therefore grows.

The dataset was downloaded from the CTDC website last February 3, 2019, which contains the September 2018 update, or the most updated dataset during that time. The downloaded data contains 55,434 records of reported human trafficking victims and information about these victims. It had 62 variables originally, capturing socio-demographic profile as well as trafficking process and exploitation types.

The variables included are *year* which is the registration of the case when the organisation assisted the victims or when the case report was received. The *year* variable's values range from 2002 to 2018. The next variable is the *datasource* which is a string type having values of either case management (social service assisted case) or hotline (case reported or serviced through phone, text message, online form report, or email). Both *year* and *datasource* do not contain any missing or unknown values.

Gender is another string variable that contains information about whether the victim is male or female and contains some missing values and unknown values due to any caseworker data collection reason. The *ageBroad* are categorized string variables of victim's age and is the registered ages during the time of assistance or casework. The variable *majorityStatus,* with string type values*,* denotes whether the individual is a minor (under the age of 18) or adult (age 18 and above). Like *ageBroad* variable, *majorityStatus* identifies the individual's age status during registration to the IOM or Polaris's assistance. A sixth variable in the extracted dataset was called *majorityStatusAtExploit*, also a string type, contains the values minor and adult, denoting the individual's age status when exploitation started. On the other hand, a variable with string values called *MajorityEntry*, is an indicator of individual's age status (minor or adult) during entry to the trafficking process, in which exploitation does not necessarily happened yet. Variables related to age or age status all have missing or unknown values. The eighth variable labelled *citizenship* contains string values based on ISO two alphabetical codes that denote nationality interchangeably or country of origin as proxy of the individual. This variable also contains missing and unknown values.

The next 18 variables are the means or form of control variables to the human trafficking victims, and containing binary numeric variable (0, 1). The first three variables are financial-related means of control. The *meansOfControlDebtBondage* indicates whether the individual is subjected to forced work in payment for a perceived debt with little to no pay, and with no monitoring of debt payment progress. There is also the *meansOfControlTakesEarnings* which indicates whether exploiters have taken

individual's compensation for control purposes. Another financial-related attribute is *meansOfControlRestrictsFinancialAccess*, which denotes if individuals have experienced prohibition or restriction of access to their personal finances via controlling bank cards and account or stealing personal funds by the exploiters.

The variable *meansOfControlThreats* indicates if individual has experienced being threatened with an intent to be harmed or imposed some loss by the exploiter. Another variable under means of control is *meansOfControlPsychologicalAbuse*, which is indicated of whether and emotional abuse or deceitful tactics have been used by the exploiter to influence the individual. Verbal abuses, shaming, and manipulation of power are just some of the examples of this form of abuse. Another form of abuse that is of physical nature, indicated under *meansOfControlPhysicalAbuse* variable. This includes exploiters inducing physical pain or injury to the individuals, up to the extent of causing death, injury, or trauma. On the other hand, an abuse of non-consenting sexual contact by exploiters to control the individual is categorized under the variable *meansOfControlSexualAbuse.* Using sexual contacts and assault as punishment or manipulation of the individual, coercion to terminate or continue pregnancy, exposure of individual to sexually transmitted infections are just some examples of this control. By definition, *meansOfControlSexualAbuse* is controlling means and is not a purpose for the trafficking.

The variable *meansOfControlFalsePromises* denotes deception in which exploiters lead individuals to a different outcome, in this case to an exploitative condition than what was promised. Another variable is the *meansOfControlPsychoactiveSubstances*, which denotes if exploiters have brought

the individual into taking substances to comply or have their behavior be altered to comply. The variable *meansOfControlRestrictsMovement* on the other hand is the confinement or isolation of the individual by the exploiter to restrict physical and social movement. Detainment, accompaniment, or threaten to impose negative results of the individual's movement are some of the examples of this control. There is also the variable *meansOfControlRestrictsMedicalCare* which denotes restriction of the individual's access to necessary health or medical services by the exploiter.

Another means of control variable is *meansOfControlExcessiveWorkingHours* which indicates exploiters requiring the individual to work excessive number of hours that what was contracted. These are often used by exploiters to keep the individual isolated socially or from any form of seeking help. The *meansOfControlUsesChildren* variable denotes impeding or limiting the individual's access to their children by the exploiter through separation, physical removal of children, or manipulation of custody. Another variable in this group is the *meansOfControlThreatOfLawEnforcement* indicates whether the individual has been threatened by the exploiter to be subjected to law enforcement and immigration authorities to adversely affect the individual.

The *meansOfControlWithholdsNecessities* variable denotes if the exploiter having denied or restricted the individual's basic living necessities including food, shelter, clothing, water, hygiene, among others.

On the other hand, *meansOfControlWithholdsDocuments* indicates whether the exploiter has restricted or controlled the individual's access to valuable documents like passport, work permit, identification card and other certifications, government benefits, legal papers, among others. There is also the

meansOfControlOther is the catch variable for other means of control by the exploiter to construct and maintain power over the individual not fitting any control variable above.

Also, there is a variable named *meansOfControlNotSpecified* which indicates if the control type was not provided by the responder or caseworker about the individual. Finally, for the means of control variables, the variable *meansOfControlConcatenated* is the concatenated or combined list of all means of control variable in string value form and delimited.

The succeeding eight variables are the exploitation type, or the purpose for which the individual was trafficked. These are main categories set as variable containing binary values (0, 1). Under this group there is *ForcedLabour* indicates if the individual was trafficked for the purpose of work or labour that is not voluntary. Sexual services are not included in this variable. Another one is *isSexualExploit* indicates if the individual was trafficked by the exploiters for the purpose of sexual services such as prostitution or pornography using fraud or coercion. There is also the *isOtherExploit* variable which indicates other type of exploitation not under forced labour or sexual exploitation. The variable named *isSexAndLabour* indicates whether the individual was trafficked for both force labour and sexual services. There are also special variables or categories for which the individual was trafficked. Under this group is the variable *isForcedMarriage*, which indicates if the individual was trafficked for the imposition of marriage through coercion, normally under penalty threats. Another one under the special purpose is i*sForcedMilitary* indicates if the purpose of the individual being trafficked was for the imposition of military service through coercion that is normally under penalty threats. Last special

variable is the *isOrganRemoval* which indicates if the purpose of trafficking the individual was for the illicit removal of internal organs, through deceit or threat and without consent. All of these variables contain missing values. There is the concatenated or combined list of all exploitation type variable in string value form and is delimited called *typeOfExploitConcatenated*.

As mentioned previously, two of the main reasons or purposes for exploiting and individual are through forced labour and sexual exploitation. The next variables are the sub-categories for those two main exploit types. Under force labour, we have *typeOfLabourAgriculture* variable which denotes if the individual exploited for labour or work is forced under "crop and animal production, hunting, and related service activities" as defined by CTDC. Next is the *typeOfLabourAquafarming* variable which denotes if the individual exploited for labour or work is forced under "fishing and aquaculture" as defined by CTDC. There is also the variable *typeOfLabourBegging* which denotes if the individual exploited for labour or work is forced to solicit money, material goods, or valuable items from other people with no provided service or products in exchange. The variable *typeOfLabourConstruction* is also included in this main category, which denotes if the individual exploited for labour is forced under to work in the construction industry or construction in general. The next one is *typeOfLabourDomesticWork* which denotes if the individual exploited for labour or work is forced under "Activities of households as employers; undifferentiated goods and services producing activities of households for own use" as defined by CTDC.

There is also the *typeOfLabourHospitality* which denotes if the individual exploited for labour or work is forced under "Accommodation and food service

activities" or "Food and beverage service activities" as defined by CTDC. Another variable is the *typeOfLabourIllicitActivities* which indicates if the individual was forced to work on illegal businesses including human smuggling, illegal substances trading, illegal substance production, and arms or weaponry dealing. The *typeOfLabourManufacturing* variable on the other hand denotes if individual was forced to work in the manufacturing industry or manufacturing in general. The next one is *typeOfLabourMiningOrDrilling* which indicates if individual was subjected to work in the mining or quarrying industry or mining or quarrying in general by means of coercion. There is the *typeOfLabourPeddling* variable which indicates if individual was forced to work on activities that relates to informal businesses in the street or public venues where items are being sold in a small scale. The *typeOfLabourTransportation* however indicates if individual was forced to work on "transportation and storage" as defined by CTDC. Finally, there is the *typeOfLabourOther* which indicates if the individual was forced to work on other types of labour that are not classified under previous labour type categories. If the labour type was not provided by the responder or caseworker about the individual, that will fall under the variable *typeOfLabourNotSpecified*. Like it its main category, all of these variables contain missing values. The concatenated or combined list of all forced labour exploitation type variable in string value form and delimited is calle *typeOfLabourConcatenated*

The second main category is sexual exploitation. Five sub-categories are under this type, including *typeOfSexProstitution* which indicates if the individual was forced to provide sexual acts of service for payment. The next variable is *typeOfSexPornography* which indicates if the individual was forced into the production of materials that are intended for the enticement of sexual excitement

to users without any participation from said users. Also included is the variable *istypeOfSexRemoteInteractiveServices* which is defined as the variable indicating whether individual was forced to engage in live sexual acts or for sexual excitement having commercial value through webcams, messaging platforms or chats, and phone sex lines. Last in this category is *typeOfSexPrivateSexualServices* which indicates if the individual was highly controlled or forced for providing personal sexual service for only one person with commercial value.

For the concatenated or combined list of all forced sexual exploitation type variable in string value form and delimited *typeOfSexConcatenated*.

An indicator variable if the individual was forced into the exploitation situation through unlawful removal is also in the dataset and is called *isAbduction*.

In the dataset, there is the variable *CountryofExploitation* which indicates where the individual was exploited. This is the country of destination in the context of trafficking, and in cases where no data of last country of exploitation was provided, the casework is the de facto value for this variable. The values are string types based on ISO two alphabetical codes for countries.

The last set of variables are the recruiter relationship type, denoting the relationship of the individual to the exploiter during the time the individual was first involved in the activities of the exploiter. The first variable is *recruiterRelationIntimatePartner* which indicates if the person who initially lured or took the individual to the exploitation situation has a current or former romantic relationship with each other. The second one is the *recruiterRelationFriend* which indicates if the person who initially lured or took the individual to the exploitation situation are familiar with each other without romantic or familial forms of relationship. The next one is the variable

recruiterRelationFamily which indicates if the person who initially lured or took the individual to the exploitation situation are connected to each other biologically, legally through marriage, or by custody. Finally, there is the *recruiterRelationOther* which indicates if the person who initially lured or took the individual to the exploitation situation have other notable relationships like contractors, former employers, or smugglers. If the person who initially lured or took the individual to the exploitation situation was not provided by the responder or case worker, this will fall under *recruiterRelationUnknown*. On the other hand, the concatenated or combined format of the recruiter relationship or the above variables that is delimited in string values is called *RecruiterRelationship*.

A dataset like the CTDC data is highly desirable in a learning from data study like this, since the amount of records is highly advantageous to learn from. On the other hand, the whole host of variables and the time span are advantageous in a research for social problem, which is more than what observational data could provide. Thus, it was motivational for this research to use this dataset on experimental viewpoint.

Variable	**Data Type**	**Missing Data**
yearOfRegistration	Numeric	X
Datasource	Categorical	X
gender	Categorical	✓
ageBroad	Categorical	✓
majorityStatus	Categorical	✓
majorityStatusAtExploit	Categorical	✓
majorityEntry	Categorical	✓
citizenship	Categorical	✓
meansOfControlDebtBondage	Binary	✓
meansOfControlTakesEarnings	Binary	✓
meansOfControlRestrictsFinancialAccess	Binary	✓
meansOfControlThreats	Binary	✓
meansOfControlPsychologicalAbuse	Binary	✓
meansOfControlPhysicalAbuse	Binary	✓

meansOfControlSexualAbuse	Binary	✓
meansOfControlFalsePromises	Binary	✓
meansOfControlPsychoactiveSubstances	Binary	✓
meansOfControlRestrictsMovement	Binary	✓
meansOfControlRestrictsMedicalCare	Binary	✓
meansOfControlExcessiveWorkingHours	Binary	✓
meansOfControlUsesChildren	Binary	✓
meansOfControlThreatOfLawEnforcement	Binary	✓
meansOfControlWithholdsNecessities	Binary	✓
meansOfControlWithholdsDocuments	Binary	✓
meansOfControlOther	Binary	✓
meansOfControlNotSpecified	Binary	✓
meansOfControlConcatenated	Categorical	✓
isForcedLabour	Binary	✓
isSexualExploit	Binary	✓
isOtherExploit	Binary	✓
isSexAndLabour	Binary	✓
isForcedMarriage	Binary	✓
isForcedMilitary	Binary	✓
isOrganRemoval	Binary	✓
typeOfExploitConcatenated	Binary	✓
typeOfLabourAgriculture	Binary	✓
typeOfLabourAquafarming	Binary	✓
typeOfLabourBegging	Binary	✓
typeOfLabourConstruction	Binary	✓
typeOfLabourDomesticWork	Binary	✓
typeOfLabourHospitality	Binary	✓
typeOfLabourIllicitActivities	Binary	✓
typeOfLabourManufacturing	Binary	✓
typeOfLabourMiningOrDrilling	Binary	✓
typeOfLabourPeddling	Binary	✓
typeOfLabourTransportation	Binary	✓
typeOfLabourOther	Binary	✓
typeOfLabourNotSpecified	Binary	✓
typeOfLabourConcatenated	Categorical	✓
typeOfSexProstitution	Binary	✓
typeOfSexPornography	Binary	✓
typeOfSexRemoteInteractiveServices	Binary	✓
typeOfSexPrivateSexualServices	Binary	✓
typeOfSexConcatenated	Categorical	✓
isAbduction	Binary	✓
RecruiterRelationship	Categorical	✓
CountryOfExploitation	Categorical	✓
recruiterRelationIntimatePartner	Binary	✓
recruiterRelationFriend	Binary	✓
recruiterRelationFamily	Binary	✓
recruiterRelationOther	Binary	✓

recruiterRelationUnknown	Binary	✓

Table 3.1. Summary of variables and data types of the CTDC dataset

The data processing for any analysis, and specifically for this research considered the presence of missing data or values as described in the presentation of the data set. The next chapter will discuss how the data processing is taken care of as well as dealing with missing values.

3.2. Data exploration and pre-processing

Initial pre-processing for the CTDC dataset has revealed a necessity to streamline variables from their current number. First, as discussed in the presentation of the data chapter, some variables that were concatenate forms or combined values of variables under the same category or line of question. For example, *RecruiterRelationship* was a concatenated variable for all values of *recruiterRelationIntimatePartner*, *recruiterRelationFriend*, *recruiterRelationFamily*, *recruiterRelationOther*, and *recruiterRelationUnknown*, and is therefore a repetition of the binary or dummy variables as enumerated. As most of these variables are categorical, the dummy variables were kept and the concatenated variables with string values were eliminated. The other variables eliminated under this premise were *meansOfControlConcatenated* (a concatenate of 18 variables), *typeOfExploitConcatenated* (concatenate of 7 variables), *typeOfLabourConcatenated* (concatenate of 13 variables), and *typeOfSexConcatenated* (concatenate of 4 variables).

Unknown or unspecified variables, as defined by CTDC as values for the category that were not provided by the responder or case worker, do not provide

any insight to the exploration of the dataset and were therefore eliminated as well. These are *meansOfControlNotSpecified*, *typeOfLabourNotSpecified*, and *recruiterRelationUnknown*.

Some attributes, although were defined by the CTDC and have a variable heading, upon examination in the dataset do not contain any presence value of 1, and only contain missing values and 0. Such variables also do not provide insights and were therefore eliminated as well. These variables are *typeOfLabourMiningOrDrilling*, *typeOfLabourTransportation*, i*sForcedMilitary*, *isOrganRemoval* and *istypeOfSexRemoteInteractiveServices*. In the context of CTDC data collection, these variables were kept following protocol and may have succeeding values in the future. A total of 48 variables remain after the elimination of variables.

3.3. Handling missing values

The presence of significant number of missing values according to the description of dataset in the previous chapter has led to a deliberate pre-processing for missing values before machine learning. These missing values, as presented in the previous chapter about the CTDC dataset, are either a result of contributors not collecting the data for all variables or due to the anonymization resulting to data loss. As machine learning tools or statistical data processes are hard to be done when dataset has missing values, necessary pre-processing must be done prior. This section discusses common practice when dealing with missing values, as well as the method of machine learning level of imputation, which fills out the missing values.

Missing values or missing data is the incomplete data matrix or the absence of data point in a record (Newman 2014). In social science, especially in surveys, it is a statistical problem when respondents do not respond to survey questions (Newman 2009). These are essentially called nonresponse which according to Newman (2009) can be deliberate, meaning a complete disregard to the survey or avoidance of sensitive items, or can be an inadvertent act of forgetting to respond. However, missing data may also occur due to collection or technical errors in the data gatherer side. The CTDC dataset's occurrence of missing data falls under the latter. Missing data problems occur commonly in all fields dealing with data processing or analyses. The amount of missing values could vary depending on the domain and field of study. In the medical field for instance, the consolidation of different sources of data from manually collected documents, laboratory results, to monitor and digital data could lead to missing values (Cismondi et al. 2013). In psychiatric and social research, it is a common problem due to the variation in collection protocol and human error (Stuart & Leaf, 2011). Missing values in machine learning can adversely affect modelling and most of the time could not be executed. Hence, missing values should be addressed first before processing (Cismondi et al. 2013).

The conventional methods in dealing with missing values are either single imputation or deletion. Researchers typically run their data processing and analysis to only complete cases or no missing values (Wulff & Elskov 2017) also called complete-case processing or analysis, due to the general adverse effect of missing values according to Cismondi et al. (2013). *Deletion* corresponds to the removal of all variables of a certain observation if one of the values for a variable is missing, and thus leaving only the complete records. *Single imputation* on the

other hand is the replacement or substitution of the apparent missing data, completing the values in the process. Imputation causes bias while deletion causes bias and a general "loss of statistical power" (Cismondi et al. 2013). In addition, the effect of missing data in affected variables snowballs into excluding sizeable amount of the sample, thus losing precision (Wulff & Elskov 2017).

Some of the most common methods of imputation is replacement using mean average or mean-value imputation for numerical values or mode-value (frequency substitution) for discrete and categorical values (Newman 2014). This imputation method is also coined as single imputation. Mean and mode imputation often results to bias since both methods tends to alter the distribution of values especially if the amount of missing values is significant and should only be limited to up to 10% of missing values. Deletion on the other hand takes out the observation or record in the dataset context. This method renders the record null and therefore does not affect the data processing or estimates. This is a common practice in a lot of statistical estimates. Like imputation, deletion could result to bias due to its distribution altering effects, as well as significant reduction of sample size or population. Thus, Newman (2014) and Wulff & Elskov (2017) both cautioned that these two common methods of imputation, although popular and commonly used, to be used only in rare cases. However, as discussed by Newman (2014) in his paper on guidelines to dealing with missing values, social scientists are still choosing the two common methods due to its acceptability in industry practice and lack of familiarity on "less biased" and "less error-prone" techniques.

A more advanced treatment to missing values is the use of classification techniques or machine learning alternatively called multiple imputation. Multiple

imputation (MI) uses the values from other variables from the observation to predict the missing value. This method is highly suggested by researchers especially in addressing significant amounts of missing values (Sterne et al. 2009). MI was explored in the CTDC human trafficking data pre-processing to have a more successful discovery of knowledge and assess the applicability of classification or predictive learning in real-world dataset.

3.4. Multiple Imputation by Chained Equations

One of the most popular MI packages is MICE or Multivariate Imputation by Chained Equations. MICE has transformed the solutions to imputation problems, relying on successive estimations of the variables using the other variables (regression) (White, Royston & Wood 2011). This prediction of the missing values in a variable takes dependence from the other variables, providing flexibility since each variable with missing values will have a more fitting probable value and distribution. MI methods substitute the missing data with several likely values, that will be solitary if it was single imputation. This considers the uncertainties that are usually underlying in the estimation process (Zhang 2016). The production of numerous and varied possible imputed data sets are ultimately combined, producing a consolidated estimate that accounts uncertainty through underestimation resulting from the consolidation that otherwise single imputation misses (Buuren & Groothuis-Oudshoorn 2010, Zhang 2016) (see Figure 3.1 for the schematic illustration of MICE).

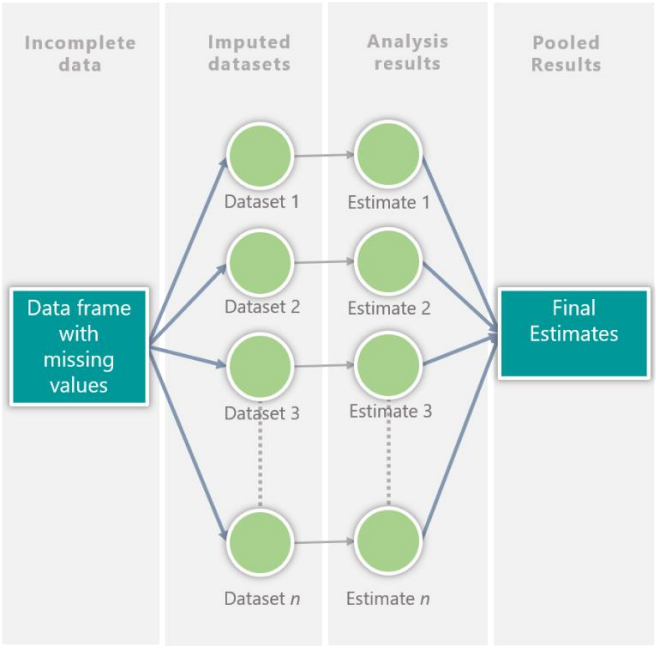

Figure 3.1. Schematic illustration of MICE process derived from Zhang (2016)

Even though suggested to be used rarely, single imputations are only acceptable in missing rates lower than 10%, while MICE or multiple imputations in general is especially beneficial on variables with missing rate of 10% or higher (Newman 2014). The resulting difference between running both methods are therefore at that level significant to resort to one technique or the other. This means that at a dataset with missing values greater than 10 percent will have significantly less bias compared to resorting to single imputation. Arguments could vary on the amount of missing data by proportion for multiple imputations to be still employable. Several studies have exhibited that multiple imputation is still impartial at a 50% rate and begins to be unstable at higher than 50% especially skewed distributions (Haji-Maghsoudi et al. 2013, Lee & Carlin 2012). However, continuing with the research done by Wulff & Elskov (2017), it was found that multiple imputations beats deletion method even at missing data rate of 75%, and thus by no means can outperform multiple imputation.

Despite the significant benefits of multiple imputation, many researchers in the disciplines of social science (Newman 2014), medical literature (Zhang 2016) and psychiatric research (Azur et al. 2011) have yet to learn this technique in handling missing values. This study demonstrated how multiple imputation was applied in a social problem setting with real casework dataset. MICE was used for the imputation of the missing values in the CTDC dataset using two models. In the application of MICE for this research, a chain or sequence of regression models was performed, where in a variable with missing values is regressed conditional with other variables in the whole data frame (Azur et al. 2011). This procedure was done in every variable with missing data, and modelling is rendered based on the data distribution. As an example, for mixed data types, logistic regression for binary variables and linear regression for continuous variables.

For the CTDC dataset's all categorical data-type variables, the MI model is logistic regression —applicable for the binary categorical data (variables with 2-level values).

Imputation of variables with missing values of categorical or nominal data type is feasible using generalized linear models. Especially for binary variables, logistic regression is one of the most usable models. *Logistic regression* (LR) or logit as a learning tool is a classifier using which uses one (simple logistic regression) or more (multiple logistic regression) variables or attributes, having the log-odds (logarithm of the odds) ratio as its expected algorithm (values 0 and 1). In a general overview, this makes logistic regression a probabilistic classifier that straightforwardly models the class discriminatively. Logistic regression is frequently carried out in machine learning tasks that requires discovery of class (dependent variable) and predictor (independent variables) relationship

(Aggarwal 2015). In a perception of multiple imputation, the idea is the approximation of the probability model on the observed data, wherein synthetic data can be drawn from the fitted probabilities, providing imputations to the missing data (Van Buuren 2012).

Multiple imputation in a universal perspective is offered and accessible via statistical and data mining tools as packages. Using R in this research, the package is the **`mice`** under method = **`logreg`** (Rdocumentation.org 2019).

In the end, multiple imputation should account for the process that created the missing data, preserve the relations in the data, and preserve the uncertainty about these relations (Buuren & Groothuis-Oudshoorn 2010). Thus, the proportion and distribution of the variables and values were checked and monitored, comparing pre and post MI for validation.

3.5. Clustering techniques

As a self-organised type of learning, clustering was considered fitting for the CTDC dataset having been unexplored. Unsupervised learning like this enables researchers to find knowledge, insights, and patterns that are otherwise unknown from having unlabelled dataset. Clustering is a vastly popular machine learning or data mining technique performed generally either by distance-based or model-based method (Van den Hoven 2015). Model-based method is where clustering is probability distribution-based and where parameters of the model are estimated from data. Distance-based on the other hand is clustering based on the computed similarity or dissimilarity measures of the data points. Both types of approaches support hierarchical clustering.

In this study, the focus on distance-based approach, also useful for categorical data. Naturally, categorical data lacks the natural ordering compared to numerical data within the variables (Saha, Sarkar & Maukik 2019; Sharma & Gaud 2015). Because of this intrinsic absence of natural ordering, outright distance-based approach cannot be performed until data configuration. Conventional and popular clustering techniques like k-means are not applicable for categorical data, since these traditional methods compute for the mean-value.

Since the inception of k-mode algorithm in 1997, it has been possible to cluster data with categorical values (Ji et al. 2012; Sharma & Gaud 2015) and has since been a widespread clustering technique in various application domains (Ng & Jing 2009). Conceptually it was developed under the framework of k-means, using modes for clusters, which are computed based on frequency of values in a variable (frequency-based), and using a straightforward matching dissimilarity measure. This frequency-based method has then since eliminated the prior strict clustering through numerical values, easing out the limitation (Ng & Jing 2009). One of the variants of k-mode clustering is *Fuzzy k-mode* which is the inclusion of fuzzy logic introduced in 1999. Fuzzy logic in computing generally refers to the "degrees of truth" instead the usual "true or false" or what is known as Boolean logic (1 or 0 values). In fuzzy logic, the values are between 0 and 1, in which 0 and 1 are the extreme values, meaning the in between values represent different states of truth. This logic extends to Fuzzy k-mode clustering, in which membership value of each data point can vary from 0 to 1, contrary to the more conventional hard clustering of only one cluster. Below is the iteration of Fuzzy k-mode clustering by Pang et al. (2012):

$$E= \sum_{l=1}^{k} \Sigma_{i=1}^{n} (u_{il})^{\alpha} \, d(x_i, Q_L)$$

where:
$\propto$ indicates fuzziness coefficient, $1 < \propto < \infty$
u_{il} indicates membership degree of data point, *i* to *l;* subjected to $\sum_{l=1}^{k} u_{il} = 1$, $0 \leq u_{il} \leq 1$.

In a paper by Saha, Sarkar & Maukik (2019), it was mentioned however that users of fuzzy k-mode could still find difficulty in interpreting membership results of the algorithm, since it failed to address indiscernibility and vagueness of clustering.

As mentioned in the presentation of data section, this study will be dealing with categorical data all the way through. One of the most effective ways, to cluster categorical data is hierarchical clustering. *Hierarchical clustering* is a part of the unsupervised learning for which clusters are built like a tree-type construction based on hierarchy, typically using distance measures (Aggarwal 2015). To expound further, it is more explicit to look at the two types of hierarchical clustering algorithms — agglomerative and divisive.

Agglomerative or bottom-up approach begins with the individual data points that are grouped into higher-level clusters. An individual data point or leaf is agglomerated with another leaf that is its most similar forming a node, and where each node is combined into its most similar one forming bigger clusters. The procedure is repeated until one big cluster or root is formed. On the other hand, *divisive* or top-down approach is the antithesis of agglomerative clustering, where in the beginning of the clustering is from the root or a single big cluster. The proceeding steps is partitioning one big group into two clusters depending on their heterogeneity or dissimilarity. The end process is until data points are membered to their own clusters. Although both types can be represented in a

tree-type structure, agglomerative clustering does not need pre-specification of the number of clusters contrary to divisive (Kamande, Miriti & Ahishakiye 2018).

As agglomerative clustering tends to be more informative due to its specific-to-general grouping ability, this study used agglomerative clustering to explore imputed dataset. Furthermore, as there is an exploratory objective in this research, the number of clusters to be found is a discovery and thus, need not be specified.

As the timescale of the dataset expands from 2002 to 2018 (17 years), a lot of trends and insights have changed over these years that a single cluster analysis for the entire data set does not isolate progression of human trafficking victims' profile. Thus, the dataset was split into 3 subsets: 2014-2018, 2009-2013, and 2002-2008. The first two mentioned subsets each contain a 5-year time span, while the last-mentioned time span is a 7-year period, to take advantage of the available data.

To infer and approach the characteristics of the dataset, a stratified proportional sampling was done in each subset before running a cluster analysis. Stratified sampling was done using the year variable as the layer, and a 30% sample was obtained per year to emulate proportion throughout the entire dataset. This sampling design gave a final sample count of 2,161 records for 2002-2008, 2,295 records for 2009-2013, and 11,551 records for 2014-2018. In the simplest sampling approach, each time period subset required different minimum sample due to varying number of total records. From the social science sampling industry norm of 95% confidence level and $\pm$5% margin of error, a minimum requirement for each time period are as follows: 379 for 2002-2008, 382 for 2009-2013, and 396 for 2014-2018. Having identified the minimum

sample required to infer from data justifies the stratified sampling plan. The method of selection after sample counts were decided was the Bernoulli's sampling, which is a suitable random sampling for finite population just like the CTDC dataset. In Bernoulli sampling, all records in the dataset are given equal probability to be part of the sample.

Each data subsets with the sample records were processed similarly using hierarchical clustering of agglomerative approach. This is to avoid the subjectivity and a guarantee of consistency when approaching the variables. To do so, Gower's Distance measure was accomplished first. *Gower's Distance* is a more suitable distance measure for categorical data, unlike its counterparts Euclidean Distance or Manhattan Distance. It is a distance measure that is practical for calculating distances even between mixed type of data (Akay & Yüksel 2018; Pavoine et al. 2009; Tsivelikas et al. 2009).

Year	Total	30% Sample
2002	1,116	335
2003	379	114
2004	250	75
2005	1,617	486
2006	1,534	461
2007	1,469	441
2008	828	249
2009	792	238
2010	1,521	457
2011	1,823	547
2012	1,395	419
2013	2,113	634
2014	3,129	939
2015	6,660	1,998
2016	17,606	5,282
2017	9,836	2,951
2018	1,270	381

Table 3.2. Stratified Sampling per Year

Period	Sample	MOE at 95% CL
2002-2008	2,161	1.80%
2009-2013	2,295	1.86%
2014-2018	11,551	0.78%

Table 3.3. Sampling per Time Period

Gower's Distance is the average of partial dissimilarities across records in the dataset. In an iteration of Van den Hoven (2015) derived from Gower (1971), the Gower's Distance formula is:

$$d(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^{(f)}$$

where: $d_{ij}^{(f)}$ is the partial dissimilarity depending on the variable type

Numeric type observations $y_i$and y_j for example will have a partial dissimilarity between the difference of the observations and the maximum range from all the records in the dataset. However, since the CTDC dataset deals with categorical variables, the partial dissimilarity between these type of observations x_i and x_j for instance will either be 1 if they have difference in value, or zero if otherwise:

$$d_{ij}^{(f)} = \begin{cases} 0 \; if \; x_{if} = x_{jf} \\ 1 \; if \; x_{if} \neq x_{jf} \end{cases}$$

In hierarchical clustering, and advisably before performing the clustering, appropriate linkage should be identified first. Clustering linkages are proximity matrices singular, complete, and average. *Single linkage* measures the shortest distance between observations, and tends to produce long, "loose" clusters. *Complete linkage* on the other hand measures the distance between the longest observations and tends to produce more compact clusters. *Average linkage*

measures the average distance between all points of the observation or instance. (Kamande et al. 2018; Van den Hoven 2015).

In R, this is done using the **`agnes`** function under the **`cluster`** library. Agnes or "agglomerative nesting" is an essential process in the data processing of this study, with methods to be tried using singular, complete, or average, since agglomerative nesting provides a coefficient that determines the linkage to be used during clustering. An agglomerative coefficient closest to 1 was the chosen link method to proceed. The agglomerative nesting is also important as it is most useful for multivariate data like the dataset in this study (Sabitha, Mehrotra & Bansal 2014; Górecki, Hofert & Holeňa 2017).

The resulting data processing using Gower's Distance provides a dissimilarity matrix, which will then be used for agglomerative hierarchical clustering. In addition, the appropriate link method that was chosen will be incorporated as a parameter in agglomerative hierarchical clustering using the dissimilarity matrix produced after running the Gower's Distance method. In R, this function is called the **`daisy`**, still under the **`cluster`** library. As mentioned in the earlier portion of this methodology section, agglomerative clustering will group the HT cases on their own and get paired and merged with other cases, moving up hierarchies.

To identify the optimal number of clusters, both the elbow chart agglomerative measures were considered for the heuristics. The elbow method provides a graphical visualization of where the elbow of the chart stops and where the rest of the values have low rate of movement. The *elbow method* explains the discrepancies in data as a function of the cluster numbers having positive relationship. The higher the cluster assignments *k,* the more varied the

discrepancies. However, the returns of an additional cluster in terms of additional information or value has a negative function to the number of clusters, and thus, is diminishing (Zambelli 2016; Bholowalia & Kumar 2014). For example, a cluster number k=2 has more information than a more varied clustering number of k=3. Essentially, the optimal number of clusters should be chosen where there is a balance of information and variance. In elbow method, this is done by looking at the angle of the line chart corresponding to cluster number *k,* and the marginal gain has dramatically drop (the elbow) or in diminishing rate (Bholowalia & Kumar 2014). Another form of verifying the optimal number of clusters considered in this research was by looking at the agglomerative measures in the statistics when agglomerative nesting was done. These are the within cluster sum of squares from the distance matrix, average distance within and between clusters, and the ratio of within and between distances.

After determining the number of optimal clusters, a visualization using coloured dendrogram was used to illustrate the cluster size. A dendrogram is a classic visual representation and output of hierarchical clusters, showing relationships between data points. The cluster labels were then obtained and used to describe and find insights from the data according to their grouping. As additional descriptive items, the income class of country and regional grouping was added based on *citizenship* and *CountryOfExploit*ion. Regional grouping was intended for pattern discovery in a geographical sense, while the income class is to verify human trafficking having to encompass country's wealth, as mentioned by the International Organization for Migration (2016). Both information on regional group (UN.org 2019) and income class (data.worldbank.org 2018) were merged with the clustering dataset for description purposes only after clustering.

The agglomerative clustering procedure described so far in this chapter is part of the hierarchical approach. For comparative purposes, a Fuzzy k-mode, which is a non-hierarchical cluster, is also performed to the imputed dataset. The validation of clustering was done by looking at the Dunn Index, wherein a value equal to 1 means an identification of the best possible partition (Misuraca, Spano & Balbi 2019). In R, this package is under **`fanny`**, in which cluster labels were also obtained for comparative descriptive pattern results.

Summary of methods:

- Exploration of missing values in the CTDC dataset
- Data transformation and normalization (binarisation)
- Multiple imputation using chained equations
- Segmentation of imputed dataset into three time periods
- Data sampling per time period
- Gower's Distance measurement
- Agglomerative hierarchical clustering
- Selection of optimal clusters and "tree cutting"
- Cluster description

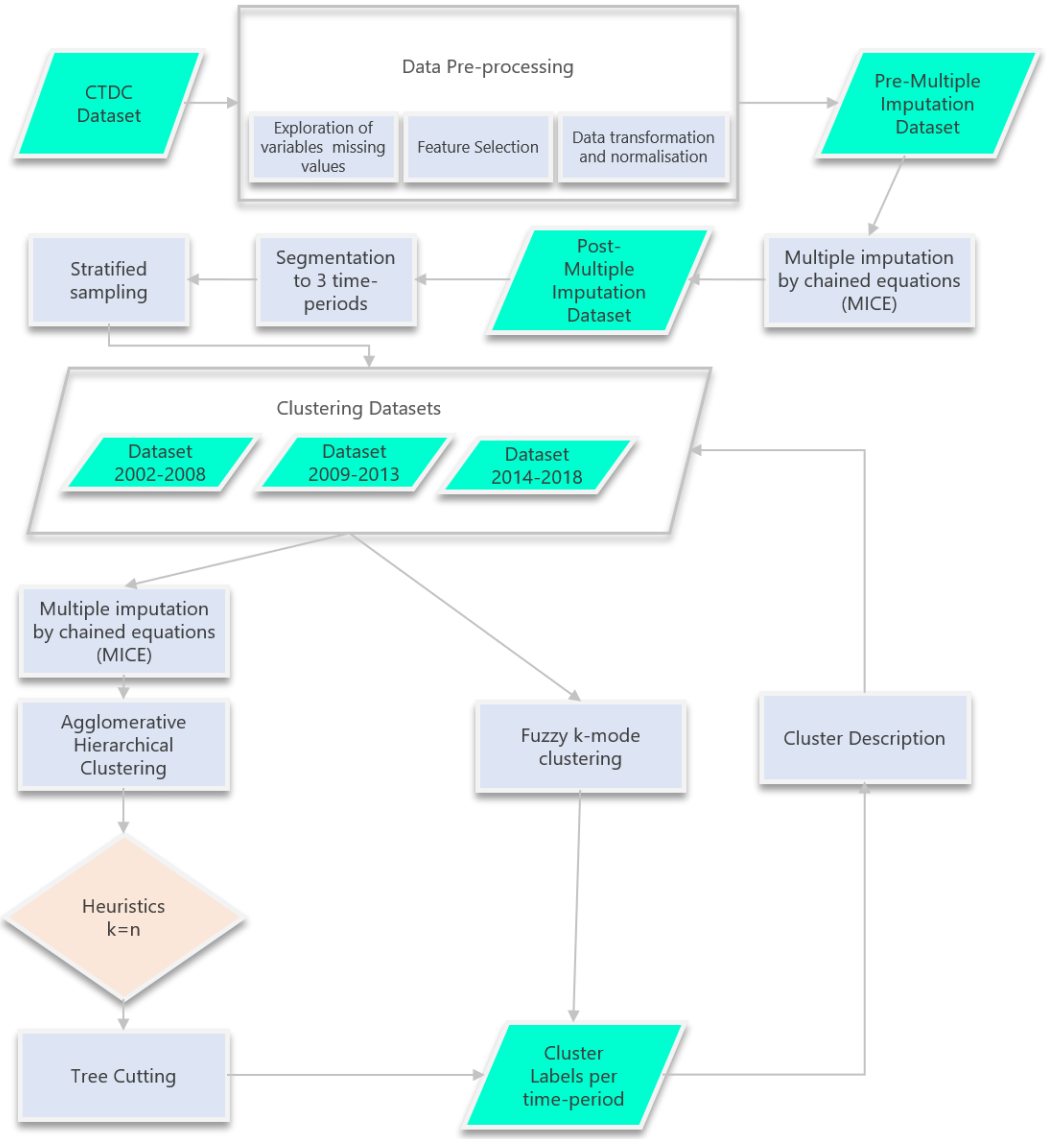

Figure 3.2. Process flow (summary of methods)

4. Results and Discussion

4.1. Multiple Imputation Results

Multiple imputation was done after discovering existence of missing values in the CTDC dataset. MICE was instrumental in the treatment of missing data using logistic regression, and classification and regression trees algorithm. The entire dataset containing records from 2002-2018 was used to run the multiple imputation processing. Hence, sub-setting to three datasets was done only before clustering.

It was essential to see the amount of missing data in the CTDC dataset before MI. In the entirety of the dataset, 61% of the data points were missing. At this point of the dataset, the variables with the greatest number of missing values are the "means of control" attributes, at a range of 88.8%-99.7%. Whereas the "exploit type" variables are in middle, at a range of 45% to 85%. The variables *citizenship*, *CountryOfExploitation*, and *gender* have the least amount of missing values, while *DataSource* and *yearOfRegistration* do not contain any missing data. Figure (4.1) details the amount of missing data per variable, while Figure (4.2) provides a plotted outlook on the distribution of missing data in the entire dataset.

As "means of control" variables are highly critical in their amount of missing data, a mitigation was implemented by merging variables of similar grouping under researcher context. The means of control restructuring stemmed to five variables replacing their sub-variables, namely:

1. ***FinControl*** from *meansOfControlTakesEarnings* and *meansOfControlRestrictsFinancialAccess;*

meansOfControlUsesChildren	99.7%	typeOfLabourBegging	54.9%
meansOfControlRestrictsFinancialAccess	99.7%	typeOfLabourAquafarming	54.9%
meansOfControlThreatOfLawEnforcement	95.9%	typeOfLabourAgriculture	54.8%
meansOfControlRestrictsMedicalCare	95.6%	typeOfLabourManufacturing	54.7%
meansOfControlDebtBondage	95.2%	typeOfLabourConstruction	54.3%
meansOfControlSexualAbuse	95.1%	typeOfLabourHospitality	52.7%
meansOfControlWithholdsNecessities	95.0%	typeOfLabourOther	52.7%
meansOfControlOther	94.0%	typeOfLabourDomesticWork	51.0%
meansOfControlPsychoactiveSubstances	93.9%	isSexAndLabour	49.8%
meansOfControlWithholdsDocuments	93.9%	isSexualExploit	49.0%
meansOfControlExcessiveWorkingHours	93.8%	isForcedLabour	45.3%
meansOfControlFalsePromises	92.5%	ageBroad	38.3%
meansOfControlTakesEarnings	91.9%	isOtherExploit	34.8%
meansOfControlPhysicalAbuse	91.0%	majorityStatus	27.6%
meansOfControlThreats	90.0%	recruiterRelationFamily	23.7%
meansOfControlRestrictsMovement	89.2%	recruiterRelationFriend	23.7%
meansOfControlPsychologicalAbuse	88.8%	recruiterRelationIntimatePartner	23.7%
majorityEntry	87.4%	recruiterRelationOther	23.7%
typeOfSexPrivateSexualServices	85.2%	citizenship	22.8%
majorityStatusAtExploit	84.0%	CountryOfExploitation	21.5%
typeOfSexPornography	83.3%	gender	1.2%
isForcedMarriage	78.9%	Datasource	0.0%
isAbduction	67.8%	yearOfRegistration	0.0%
typeOfSexProstitution	60.0%		
typeOfLabourPeddling	55.0%		

Figure 4.1. Frequency of missing data per variable

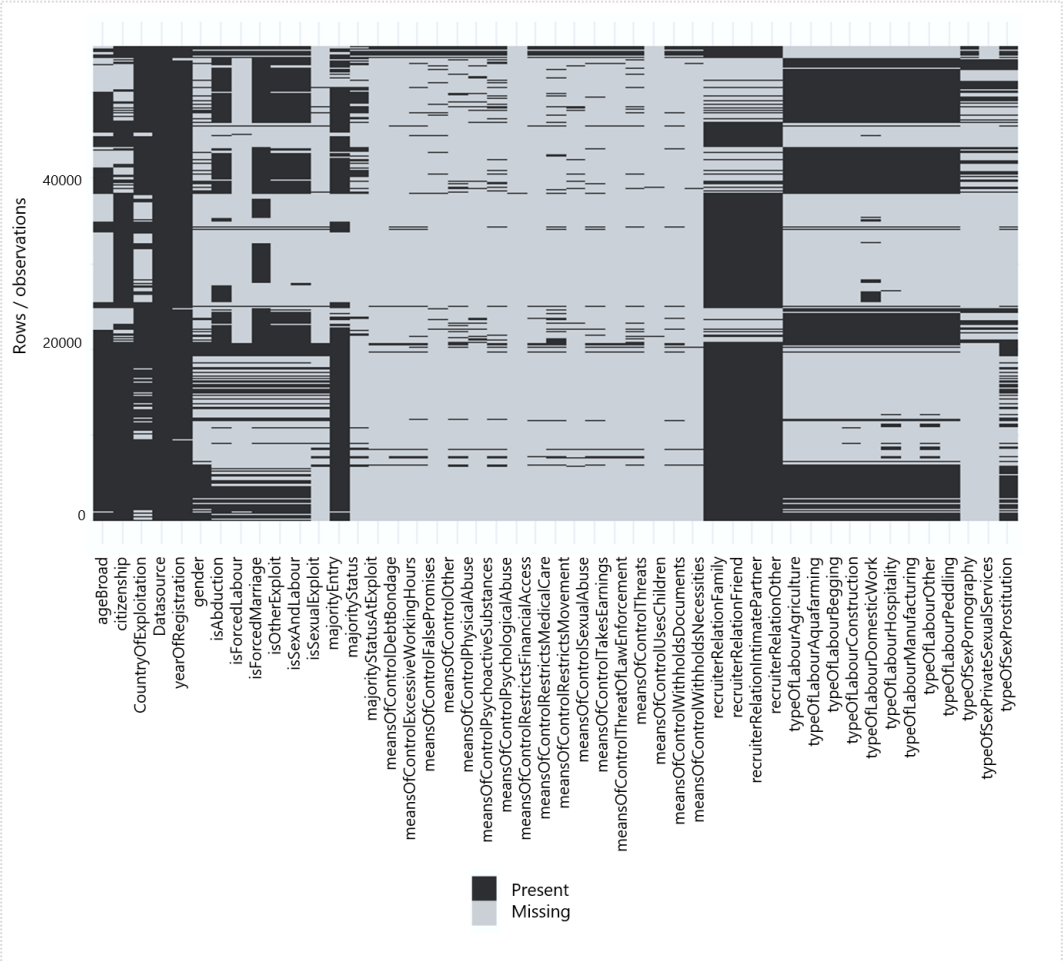

Figure 4.2. Present vs missing data plot

2. ***PsychControl*** from *meansOfControl*Threats, *meansOfControlPsychologicalAbuse*, *meansOfControlFalsePromises*; *meansOfControlUsesChildren*; *meansOfControlThreatOfLawEnforcement*;
3. ***PhysSexRest*** from *meansOfControlPhysicalAbuse*, *meansOfControlPsychoactiveSubstances*, *meansOfControlRestrictsMovement*, *meansOfControlRestrictsMedicalCare*, *meansOfControlExcessiveWorkingHours*, *meansOfControlWithholdsNecessities* and *meansOfControlSexualAbuse*;
4. ***WithholdDocs*** from *meansOfControlWitholdsDocuments* and
5. ***Other.***

This merging of "means of control" variables has reduced the percentage of missing data to 75.5%.

Datasource does not have a missing data since it is outright identifiable whether the human trafficking record has been dealt as a case or hotline. *Datasource*, however, was kept in the imputation for explanatory variable purposes. The human trafficking control variables (*FinControl*, *PsychThreats*, *PhySexRest*, *WithholdDocs*, *Other*) are notably even in the amount of missing values since these are binary variables from a concatenated form.

The pre-MI dataset was transformed and normalised into dummy variables prior to the multiple imputation of attributes, containing binary data type, into 138 variables in total.

After the multiple imputation process, the frequencies of the resulting imputed dataset were observed. It is important to maintain the proportion of the

responses throughout the variable to maintain the distribution based on the existing dataset. Figures 4.3, 4.4 and 4.5 illustrates the before and after imputation. The result value mostly maintained the proportion of data, comparing missing to imputed per variable indicating success in MI as cited by Buuren & Groothuis-Oudshoorn (2010). *Gender* was perfectly maintained due to having the least amount of missing values. On the other hand, *majorityStatusAtExploit* and *majorityEntry* having the greatest number of missing values after the "means of control" variables merging, have categories that were now different in proportion compared to their raw form specially the former. The variables *dataSource* and *yearOfRegistration* are not shown in the results since they do not contain missing data but were kept for the multiple imputation process.

The resulting imputed dataset was segmented to its three time periods as discussed in the previous chapter according to the sampling design.

Variable Source	Attribute	WM 1	WM 0	MI 1	MI 0
gender	Male	26.3%	73.7%	26.3%	73.7%
gender	Female	73.7%	26.3%	73.7%	26.3%
ageBroad	0 - 17	27.4%	72.6%	25.8%	74.2%
ageBroad	18 - 29	40.8%	59.2%	42.5%	57.5%
ageBroad	30 - 47	27.4%	72.6%	27.6%	72.4%
ageBroad	48+	4.4%	95.6%	4.1%	95.9%
majorityStatus	Minor	26.2%	73.8%	22.9%	77.1%
majorityStatus	Adult	73.8%	26.2%	77.1%	22.9%
majorityStatusAtExploit	Minor	63.1%	36.9%	42.7%	57.3%
majorityStatusAtExploit	Adult	36.9%	63.1%	57.3%	42.7%
majorityEntry	Minor	19.0%	81.0%	30.9%	69.1%
majorityEntry	Adult	81.0%	19.0%	69.1%	30.9%
typeOfExploit	Forced Marriage	0.4%	99.6%	0.4%	99.6%
typeOfExploit	Slavery and Similar Practices	1.0%	99.0%	1.2%	98.8%
typeOfExploit	Other	19.2%	80.8%	20.7%	79.3%
typeOfExploit	FL in Agriculture	2.6%	97.4%	2.4%	97.6%
typeOfExploit	FL in Aquafarming	0.5%	99.5%	0.7%	99.3%
typeOfExploit	FL in Begging	0.9%	99.1%	0.9%	99.1%
typeOfExploit	FL in Construction	6.6%	93.4%	6.2%	93.8%
typeOfExploit	FL in Domestic Work	12.2%	87.8%	12.7%	87.3%
typeOfExploit	FL in Hospitality	0.6%	99.4%	1.2%	98.8%
typeOfExploit	FL in Manufacturing	2.1%	97.9%	2.2%	97.8%
typeOfExploit	FL in Peddling	0.4%	99.6%	0.9%	99.1%
typeOfExploit	Other FL	2.1%	97.9%	1.9%	98.1%
typeOfExploit	SEX in Pornography	0.6%	99.4%	0.8%	99.2%
typeOfExploit	SEX in Private Sexual Services	0.1%	99.9%	0.6%	99.4%
typeOfExploit	SEX in Prostitution	50.7%	49.3%	47.3%	52.7%
RecruiterRelationship	Family/Relative	17.6%	82.4%	14.3%	85.7%
RecruiterRelationship	Friend	16.8%	83.2%	17.5%	82.5%
RecruiterRelationship	Intimate Partner	18.2%	81.8%	14.6%	85.4%
RecruiterRelationship	Other	47.4%	52.6%	53.6%	46.4%
meansOfControl	FinancialControl	26.7%	73.3%	27.4%	72.6%
meansOfControl	PsychThreats	62.0%	38.0%	63.2%	36.8%
meansOfControl	PhysSexRest	69.2%	30.8%	70.5%	29.5%
meansOfControl	WithholdsDocuments	17.9%	82.1%	19.6%	80.4%
meansOfControl	Other	17.2%	82.8%	15.1%	84.9%

FL – Forced Labour; SEX – Sexual Exploitation; WM 1– With Missing Data 1-values; WM 1– With Missing Data 0-values; MI 1– Multiple Imputed Data 1-values; WM 1– Multiple Imputed 0-values

Figure 4.3. Comparison with missing data and after multiple imputation

Variable Source	Attribute	WM 1	WM 0	MI 1	MI 0
citizenship	AFGHANISTAN	0.6%	99.4%	0.5%	99.5%
citizenship	ALBANIA	0.1%	99.9%	0.1%	99.9%
citizenship	BANGLADESH	0.0%	100.0%	0.0%	100.0%
citizenship	BURKINA FASO	0.1%	99.9%	0.1%	99.9%
citizenship	BULGARIA	0.8%	99.2%	0.6%	99.4%
citizenship	BOLIVIA	0.0%	100.0%	0.0%	100.0%
citizenship	BELARUS	3.7%	96.3%	2.8%	97.2%
citizenship	REPUBLIC OF THE	0.1%	99.9%	0.1%	99.9%
citizenship	COTE DIVOIRE	0.1%	99.9%	0.1%	99.9%
citizenship	CHINA	0.3%	99.7%	0.5%	99.5%
citizenship	COLOMBIA	0.3%	99.7%	0.5%	99.5%
citizenship	ERITREA	0.0%	100.0%	0.0%	100.0%
citizenship	GHANA	1.3%	98.7%	1.0%	99.0%
citizenship	GUINEA	0.0%	100.0%	0.0%	100.0%
citizenship	GUINEA-BISSAU	0.3%	99.7%	0.3%	99.7%
citizenship	HAITI	0.8%	99.2%	0.7%	99.3%
citizenship	INDONESIA	4.7%	95.3%	3.7%	96.3%
citizenship	INDIA	0.0%	100.0%	0.0%	100.0%
citizenship	KYRGYZSTAN	1.1%	98.9%	1.0%	99.0%
citizenship	CAMBODIA	4.3%	95.7%	3.5%	96.5%
citizenship	KOREA, REPUBLIC OF	0.0%	100.0%	0.1%	99.9%
citizenship	KAZAKHSTAN	0.1%	99.9%	0.1%	99.9%
citizenship	REPUBLIC	0.3%	99.7%	0.2%	99.8%
citizenship	SRI LANKA	0.2%	99.8%	0.2%	99.8%
citizenship	MOLDOVA, REPUBLIC OF	18.1%	81.9%	14.1%	85.9%
citizenship	MADAGASCAR	0.2%	99.8%	0.2%	99.8%
citizenship	MALI	0.1%	99.9%	0.1%	99.9%
citizenship	MYANMAR	3.1%	96.9%	2.4%	97.6%
citizenship	MEXICO	1.1%	98.9%	1.2%	98.8%
citizenship	NIGER	0.1%	99.9%	0.1%	99.9%
citizenship	NIGERIA	0.3%	99.7%	0.2%	99.8%
citizenship	NEPAL	0.1%	99.9%	0.1%	99.9%
citizenship	PHILIPPINES	25.0%	75.0%	30.0%	70.0%
citizenship	ROMANIA	1.6%	98.4%	1.3%	98.7%
citizenship	SIERRA LEONE	0.2%	99.8%	0.2%	99.8%
citizenship	SENEGAL	0.2%	99.8%	0.1%	99.9%
citizenship	EL SALVADOR	0.0%	100.0%	0.0%	100.0%
citizenship	THAILAND	0.3%	99.7%	0.3%	99.7%
citizenship	TAJIKISTAN	0.2%	99.8%	0.1%	99.9%
citizenship	TURKMENISTAN	0.1%	99.9%	0.0%	100.0%
citizenship	UKRAINE	18.8%	81.2%	14.6%	85.4%
citizenship	UGANDA	0.2%	99.8%	0.1%	99.9%
citizenship	UNITED STATES OF AMERICA	10.0%	90.0%	18.0%	82.0%
citizenship	UZBEKISTAN	0.6%	99.4%	0.5%	99.5%
citizenship	VIETNAM	0.4%	99.6%	0.3%	99.7%

WM 1– With Missing Data 1-values; WM 1– With Missing Data 0-values; MI 1– Multiple Imputed Data 1-values; WM 1– Multiple Imputed 0-values

Figure 4.4. Comparison with missing data and after multiple imputation (citizenship attribute)

Variable Source	Attribute	WM 1	WM 0	MI 1	MI 0
CountryOfExploitation	UNITED ARAB EMIRATES	1.2%	98.8%	1.0%	99.0%
CountryOfExploitation	AFGHANISTAN	0.2%	99.8%	0.3%	99.7%
CountryOfExploitation	ALBANIA	0.1%	99.9%	0.1%	99.9%
CountryOfExploitation	ARGENTINA	0.0%	100.0%	0.0%	100.0%
CountryOfExploitation	AUSTRIA	0.1%	99.9%	0.0%	100.0%
CountryOfExploitation	BOSNIA AND HERZEGOVINA	0.3%	99.7%	0.3%	99.7%
CountryOfExploitation	BULGARIA	0.8%	99.2%	0.7%	99.3%
CountryOfExploitation	BAHRAIN	0.1%	99.9%	0.0%	100.0%
CountryOfExploitation	BELARUS	0.9%	99.1%	0.9%	99.1%
CountryOfExploitation	CHINA	0.2%	99.8%	0.1%	99.9%
CountryOfExploitation	CYPRUS	0.0%	100.0%	0.0%	100.0%
CountryOfExploitation	CZECH REPUBLIC	0.1%	99.9%	0.1%	99.9%
CountryOfExploitation	DENMARK	0.0%	100.0%	0.0%	100.0%
CountryOfExploitation	ECUADOR	0.0%	100.0%	0.0%	100.0%
CountryOfExploitation	EGYPT	0.0%	100.0%	0.0%	100.0%
CountryOfExploitation	GHANA	1.3%	98.7%	1.0%	99.0%
CountryOfExploitation	HONG KONG	0.1%	99.9%	0.1%	99.9%
CountryOfExploitation	HAITI	0.8%	99.2%	0.6%	99.4%
CountryOfExploitation	INDONESIA	4.1%	95.9%	3.6%	96.4%
CountryOfExploitation	ITALY	0.1%	99.9%	0.1%	99.9%
CountryOfExploitation	JORDAN	0.3%	99.7%	0.2%	99.8%
CountryOfExploitation	JAPAN	0.3%	99.7%	0.4%	99.6%
CountryOfExploitation	CAMBODIA	2.3%	97.7%	2.1%	97.9%
CountryOfExploitation	KUWAIT	0.5%	99.5%	0.7%	99.3%
CountryOfExploitation	KAZAKHSTAN	0.5%	99.5%	0.5%	99.5%
CountryOfExploitation	LEBANON	0.2%	99.8%	0.2%	99.8%
CountryOfExploitation	MOLDOVA, REPUBLIC OF	13.3%	86.7%	13.5%	86.5%
CountryOfExploitation	MADAGASCAR	0.2%	99.8%	0.2%	99.8%
CountryOfExploitation	YUGOSLAV REPUBLIC OF	0.5%	99.5%	0.4%	99.6%
CountryOfExploitation	MAURITIUS	0.0%	100.0%	0.0%	100.0%
CountryOfExploitation	MALAYSIA	2.1%	97.9%	2.3%	97.7%
CountryOfExploitation	OMAN	0.2%	99.8%	0.2%	99.8%
CountryOfExploitation	PHILIPPINES	4.6%	95.4%	8.4%	91.6%
CountryOfExploitation	POLAND	0.8%	99.2%	0.7%	99.3%
CountryOfExploitation	QATAR	0.6%	99.4%	0.6%	99.4%
CountryOfExploitation	ROMANIA	0.6%	99.4%	0.6%	99.4%
CountryOfExploitation	RUSSIAN FEDERATION	6.5%	93.5%	5.6%	94.4%
CountryOfExploitation	SAUDI ARABIA	0.6%	99.4%	0.6%	99.4%
CountryOfExploitation	SINGAPORE	0.1%	99.9%	0.1%	99.9%
CountryOfExploitation	SIERRA LEONE	0.2%	99.8%	0.2%	99.8%
CountryOfExploitation	SENEGAL	0.9%	99.1%	0.7%	99.3%
CountryOfExploitation	SYRIAN ARAB REPUBLIC	0.1%	99.9%	0.1%	99.9%
CountryOfExploitation	THAILAND	1.1%	98.9%	0.9%	99.1%
CountryOfExploitation	TAJIKISTAN	0.2%	99.8%	0.2%	99.8%
CountryOfExploitation	TURKMENISTAN	0.1%	99.9%	0.0%	100.0%
CountryOfExploitation	TURKEY	0.6%	99.4%	0.5%	99.5%
CountryOfExploitation	TRINIDAD AND TOBAGO	0.2%	99.8%	0.2%	99.8%
CountryOfExploitation	TAIWAN, PROVINCE OF CHINA	0.1%	99.9%	0.1%	99.9%
CountryOfExploitation	UKRAINE	12.4%	87.6%	10.0%	90.0%
CountryOfExploitation	UGANDA	0.3%	99.7%	0.2%	99.8%
CountryOfExploitation	UNITED STATES OF AMERICA	38.4%	61.6%	39.6%	60.4%
CountryOfExploitation	UZBEKISTAN	0.4%	99.6%	0.4%	99.6%
CountryOfExploitation	VIETNAM	0.0%	100.0%	0.0%	100.0%
CountryOfExploitation	MONTENEGRO	0.5%	99.5%	0.5%	99.5%
CountryOfExploitation	SOUTH AFRICA	0.1%	99.9%	0.1%	99.9%

WM 1– With Missing Data 1-values; WM 1– With Missing Data 0-values; MI 1– Multiple Imputed Data 1-values; WM 1– Multiple Imputed 0-values

Figure 4.5. Comparison with missing data and after multiple imputation (CountryOfExploitation attribute)

4.2. Linkage Results and Initial Dendrograms

As discussed in the methodology section, an optimal linkage must be discovered before running an actual clustering algorithm. This part of the process included the sampled subsets of the data were used based on time periods. After running an agglomerative nesting using average, complete, and single link methods individually, the coefficients were obtained for comparative purposes. The results show that *complete linkage* is the best link type for the hierarchical clustering task, providing the highest agglomerative coefficient compared to *average* and *single* linkage types throughout all period types.

	2002-2008	**2009-2013**	**2014-2018**
Average	0.9719977	0.9726484	0.9863223
Complete	0.9832705	0.9834034	0.9917859
Single	0.8569962	0.8499709	0.9127882

Table 4.1. Agglomerative linkage coefficients result

The Gower's Distance was then computed per time period, in which the resulting dissimilarity matrices were used to run individual agglomerative hierarchical clusters using *complete linkage*. The resulting dendrograms presented in Figures 4.6, 4.7 and 4.8 provide discernments of the optimal clusters from each time period, as the height values at the Y-axis of the charts provides the distance between node to node or cluster to cluster. For instance, the dendrogram of 2002-2008 seem to have three major clusters, 2009-2013 would have four and 2014-2018 would have three. However, to provide a better measurement of the optimal number of clusters, the elbow chart and agglomerative measures, discussed in previous chapter, were performed as they

were established essential determinants of the number of clusters (cutting the tree).

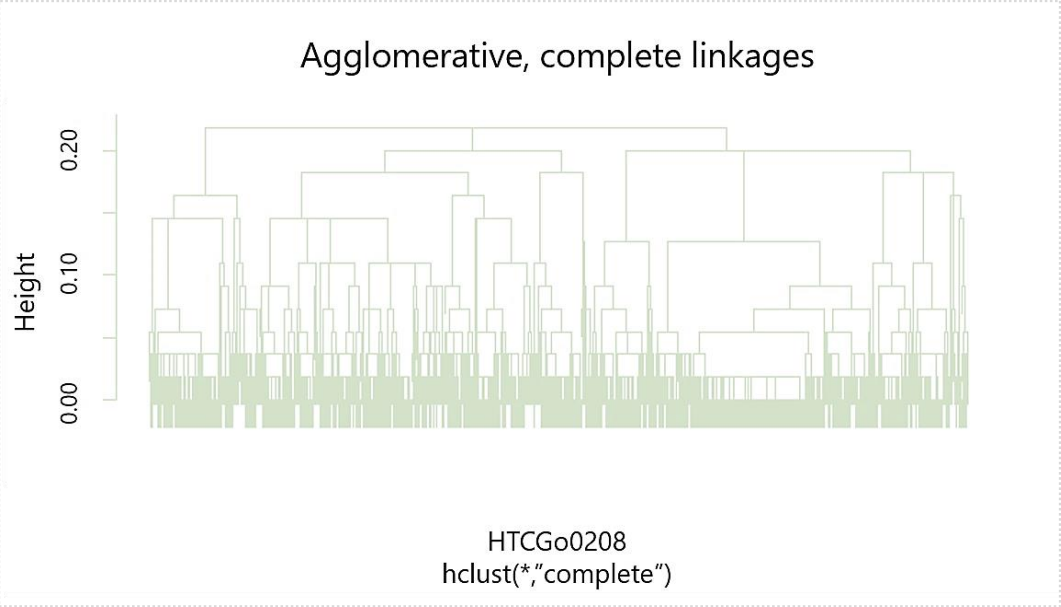

Figure 4.6. Initial dendrogram 2002-2008

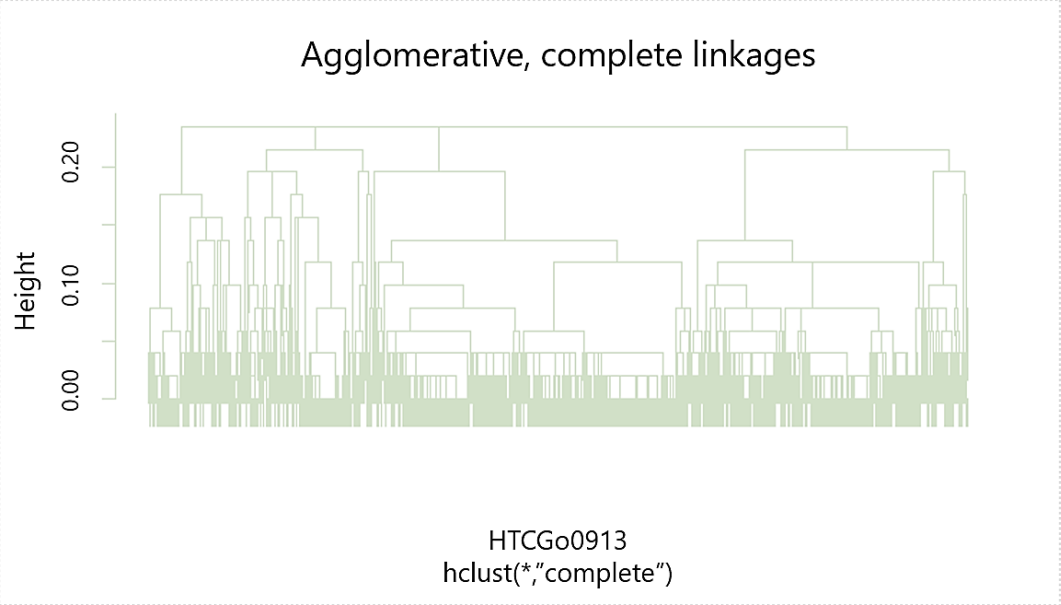

Figure 4.7. Initial dendrogram 2009-2013

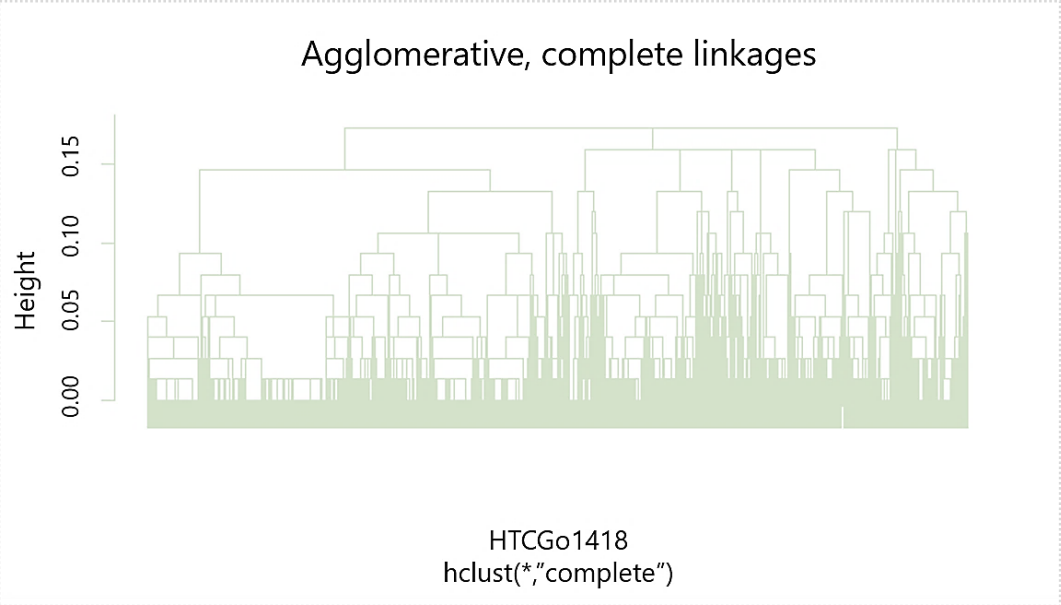

Figure 4.8. Initial dendrogram 2014-2018

4.3. Elbow method and agglomerative measures: tree cutting

The elbow method and the statistics from the agglomerative clustering are the essential heuristics for determining the optimal number of clusters in each time period in this study. The 2002-2008 agglomerative clustering Elbow chart suggests that the "elbow" is at with number of clusters of 3 (Fig 4.9). At this cluster number, the average within distance of leaves within the clusters start to diminish at a flatter rate. The same can be observed with the average between distance of clusters (Table 4.2).

For 2009-2013, the "elbow" was apparent on the number of clusters of 4 (Fig 4.10). This was also confirmed by the average within and between clusters, which both started to decline at a diminishing rate after the 4th cluster (Table 4.3). Hence, for this time period the optimal number of clusters is four.

The last time period (2014-2018), has an apparent 2 elbows at number of cluster 4 and 10 (Fig 4.11). However, looking at the agglomerative statistics (Table 4.4), average within and between distances for 10 clusters are not substantial, while the cluster sizes have very diffused grouping. Thus, the optimal number of clusters chosen was 4, in which the dendrogram tree will be cut.

2002-2008	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Test 11
cluster.number	2	3	4	5	6	7	8	9	10	11	12
n	2161	2161	2161	2161	2161	2161	2161	2161	2161	2161	2161
within.cluster.ss	16.47	12.77	11.36	9.83	8.92	8.58	8.03	7.78	6.66	6.59	6.5
average.within	0.12	0.1	0.09	0.08	0.08	0.08	0.08	0.07	0.07	0.07	0.07
average.between	0.19	0.16	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
wb.ratio	0.62	0.64	0.6	0.56	0.54	0.53	0.51	0.5	0.47	0.46	0.46
dunn2	1.56	1.3	1.18	1.14	1.22	1.22	1.19	1.16	1.21	1.06	1.06
Cluster- 1 size	1868	877	877	877	719	719	719	719	486	486	486
Cluster- 2 size	293	991	826	254	254	254	146	95	95	95	95
Cluster- 3 size	0	293	293	572	572	572	572	572	572	572	572
Cluster- 4 size	0	0	165	293	158	29	29	29	233	233	233
Cluster- 5 size	0	0	0	165	293	293	293	293	29	29	29
Cluster- 6 size	0	0	0	0	165	165	165	165	293	293	293
Cluster- 7 size	0	0	0	0	0	129	108	51	165	165	165
Cluster- 8 size	0	0	0	0	0	0	129	108	51	43	27
Cluster- 9 size	0	0	0	0	0	0	0	129	108	108	108
Cluster- 10 size	0	0	0	0	0	0	0	0	129	129	16
Cluster- 11 size	0	0	0	0	0	0	0	0	0	8	129
Cluster- 12 size	0	0	0	0	0	0	0	0	0	0	8

Table 4.2. Agglomerative statistics 2002-2008

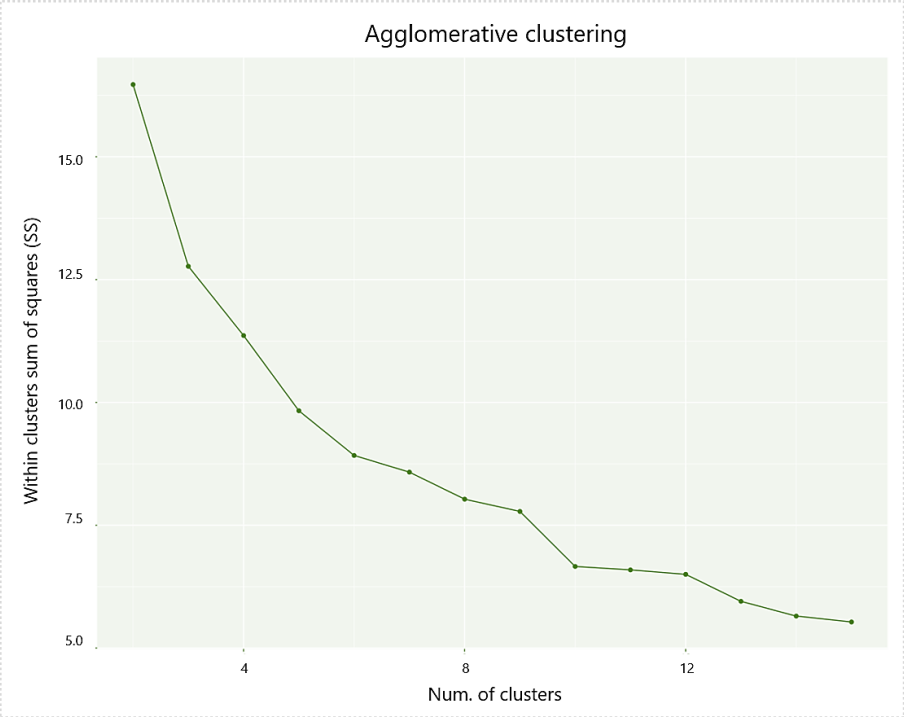

Figure 4.9. Elbow method chart 2002-2008

2009-2013	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Test 11
cluster.number	2	3	4	5	6	7	8	9	10	11	12
n	2295	2295	2295	2295	2295	2295	2295	2295	2295	2295	2295
within.cluster.ss	24.18	20.43	12.24	11.5	9.83	9.67	9.52	9.17	8.98	8.81	8.73
average.within	0.13	0.12	0.09	0.09	0.08	0.08	0.08	0.08	0.08	0.08	0.08
average.between	0.19	0.18	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
wb.ratio	0.72	0.69	0.54	0.52	0.48	0.48	0.47	0.47	0.46	0.46	0.45
dunn2	1.37	1.35	1.26	1.28	1.26	1.24	1.16	1.09	0.93	0.93	0.93
Cluster- 1 size	2029	1665	777	777	650	650	650	650	650	650	650
Cluster- 2 size	266	364	888	888	888	888	888	888	888	878	878
Cluster- 3 size	0	266	364	306	306	306	306	253	178	178	178
Cluster- 4 size	0	0	266	266	266	266	266	266	266	266	266
Cluster- 5 size	0	0	0	58	58	58	21	53	75	75	75
Cluster- 6 size	0	0	0	0	127	16	16	21	53	53	53
Cluster- 7 size	0	0	0	0	0	111	111	16	21	21	21
Cluster- 8 size	0	0	0	0	0	0	37	111	16	16	6
Cluster- 9 size	0	0	0	0	0	0	0	37	111	111	111
Cluster- 10 size	0	0	0	0	0	0	0	0	37	10	10
Cluster- 11 size	0	0	0	0	0	0	0	0	0	37	10
Cluster- 12 size	0	0	0	0	0	0	0	0	0	0	37

Table 4.3. Agglomerative statistics 2009-2013

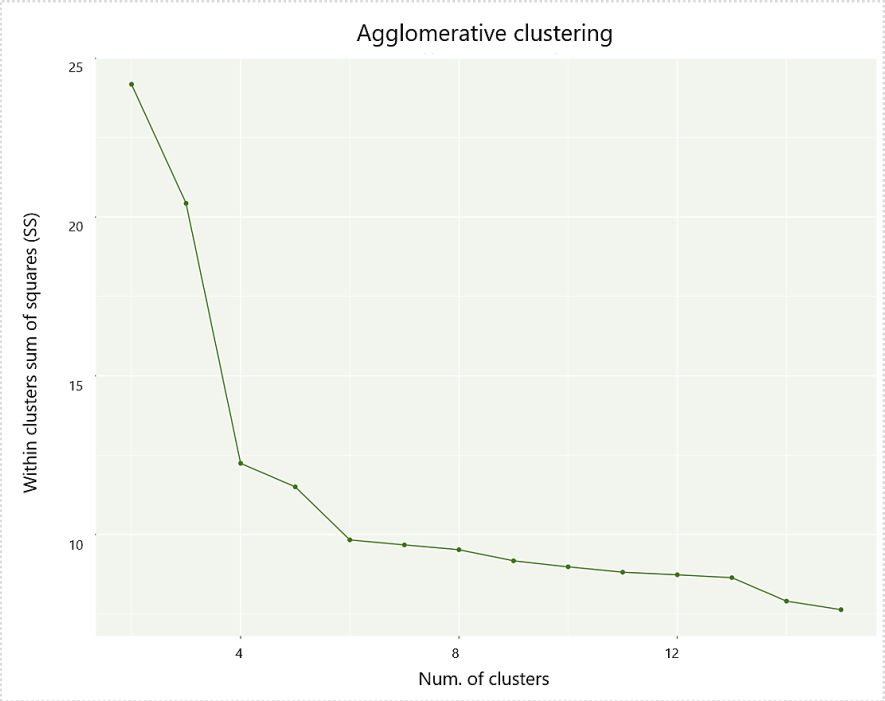

Figure 4.10. Elbow method chart 2009-2013

2014-2018	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Test 11
cluster.number	2	3	4	5	6	7	8	9	10	11	12
n	11551	11551	11551	11551	11551	11551	11551	11551	11551	11551	11551
within.cluster.ss	53.83	49.88	46.36	46.13	43.79	38.93	36.36	32.89	23.83	23.51	23.28
average.within	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.07	0.06	0.06	0.06
average.between	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.11	0.11	0.11
wb.ratio	0.75	0.72	0.69	0.69	0.67	0.63	0.61	0.58	0.5	0.5	0.49
dunn2	1.11	1.15	0.71	0.71	0.74	0.74	0.81	0.99	0.91	0.92	0.92
Cluster- 1 size	5627	4265	4265	4265	3823	2061	1635	1161	1161	1161	1133
Cluster- 2 size	5924	1362	1036	998	998	998	426	474	474	474	474
Cluster- 3 size	0	5924	5924	5924	5924	5924	998	426	426	426	426
Cluster- 4 size	0	0	326	38	38	38	5924	998	998	935	28
Cluster- 5 size	0	0	0	326	326	326	38	5924	3085	3085	935
Cluster- 6 size	0	0	0	0	442	1762	326	38	2839	2839	3085
Cluster- 7 size	0	0	0	0	0	442	1762	326	38	38	2839
Cluster- 8 size	0	0	0	0	0	0	442	1762	326	326	38
Cluster- 9 size	0	0	0	0	0	0	0	442	1762	1762	326
Cluster- 10 size	0	0	0	0	0	0	0	0	442	63	1762
Cluster- 11 size	0	0	0	0	0	0	0	0	0	442	63
Cluster- 12 size	0	0	0	0	0	0	0	0	0	0	442

Table 4.4. Agglomerative statistics 2014-2018

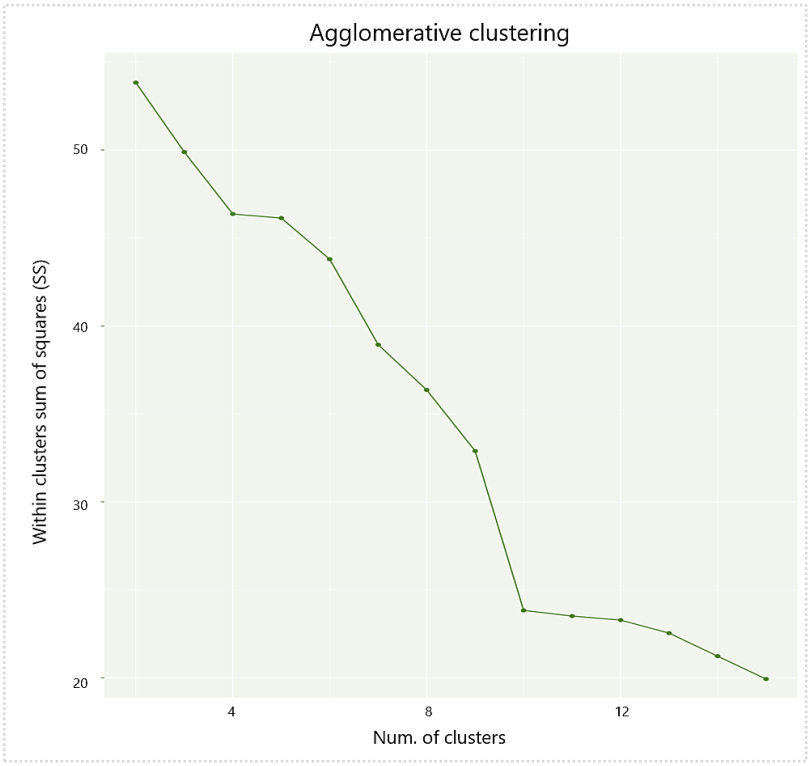

Figure 4.11. Elbow method chart 2014-2018

Initial agglomerative cluster trees (Figures 4.6, 4.7, 4.8) were partitioned according to the number of optimal clusters identified from each time period through a coloured dendrograms (Figures 4.12, 4.13, 4.14). These dendrograms also graphically illustrate the cluster size per time period.

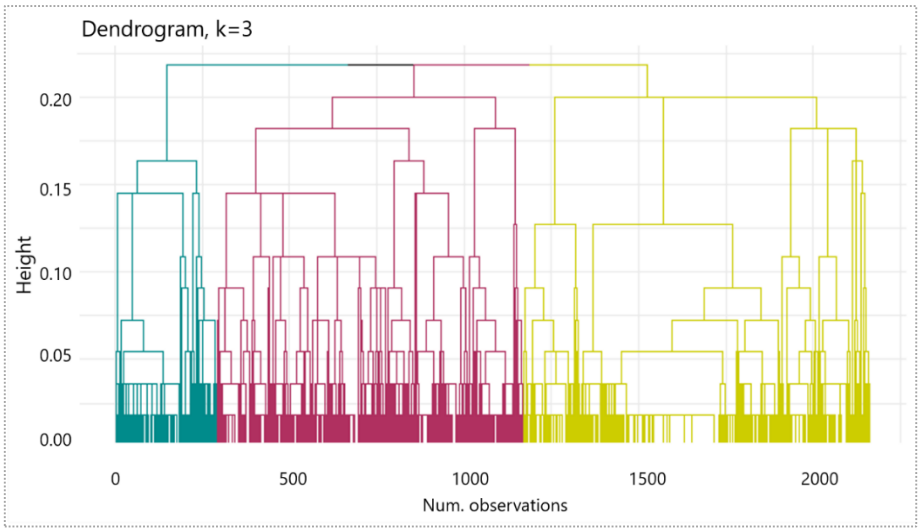

Figure 4.12. Coloured cut dendrogram 2002-2008

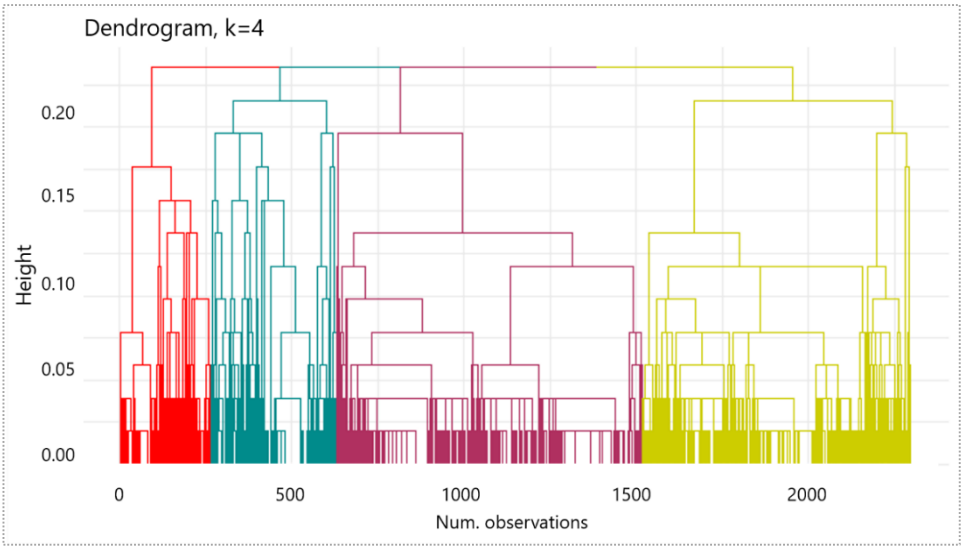

Figure 4.13. Coloured cut dendrogram 2009-2013

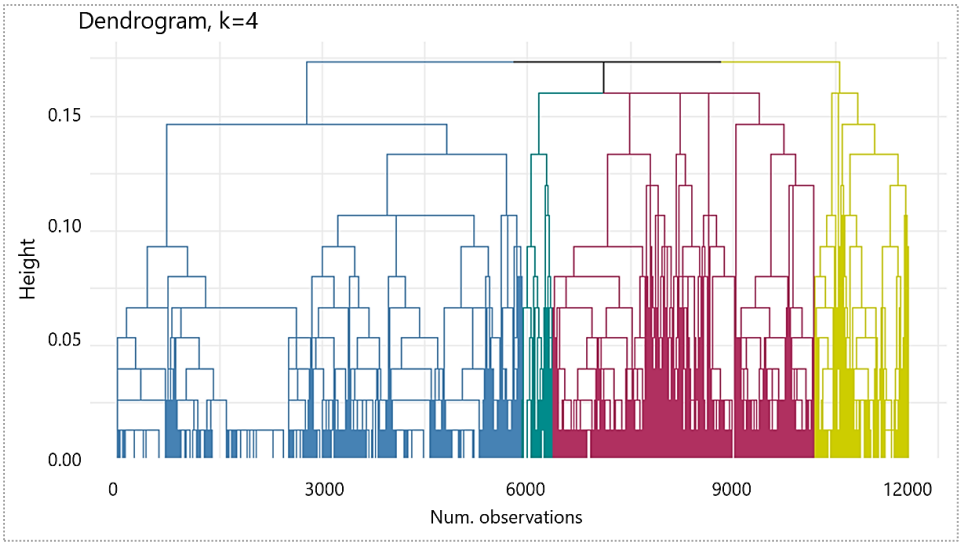

Figure 4.14. Coloured cut dendrogram 2014-2018

4.4. Patterns and Descriptions

For the oldest time period **2002-2008**, three clusters were identified with the first two clusters having almost the same proportion from the sample (Cluster 1 – 41% and Cluster 2 – 46%). Both groups are case managed or reported through a field worker with mostly female cases that are young adults. The first group

Cluster 1.1 are cases that were mostly exploited abroad coming from mostly Eastern European countries and South-eastern Asia. Cases coming from the Eastern European countries were mostly exploited in Eastern and Southern Europe, and Western Asia (countries like Ukraine and Belarus exploited in Russia, Poland, and Turkey). While cases of Indonesians who were exploited in Malaysia were also reported. These middle-income country-origin cases being exploited in middle to high income countries. The first group were exploited for both sexual exploitations for prostitution (67%) and forced labour doing domestic work (31%). This group were controlled by the exploiters through psychological control and physical or sexual abuse.

The second group ***Cluster 1.2*** on the other hand were exploited in their home country in contrary to the first group. These are Eastern Europeans of Moldovan, Ukrainian, Bulgarian, and Romanian origin exploited in their own home countries. This group was exploited also mostly for sexual services purposes (69%) particularly for prostitution (100% of the 69%). Similarly, this group were controlled by the exploiters through psychological control and physical or sexual abuses and restrictions. The mentioned Eastern European countries in this cluster all belong to middle income economy.

Although the third group ***Cluster 1.3*** is only 14% of the sample, it sets itself apart from the first two clusters as most of the cases were minor and male. Also reported through a field case worker, the cases involved are exploitation of Western Africans (85%) exploited mostly in their own home country (Ghana and Sierra Leone), or exploited in their neighbouring Western African country (Guinea-Bissau exploited in Senegal). This indicates exploitation in lower income

countries and the cases were mostly exploited for forced labour purposes for begging, domestic work, and other labour types.

For the four clusters identified the **2009-2013** time period, the first (34%) and second (39%) clusters comprised the big portion of the entire time-period sample. All four clusters have cases reported through a field work. First, ***Cluster 2.1*** is a mix of male and female cases ranging from young to mid adult. This group's cases were exploited in their home country from mostly Eastern European countries, like Ukraine (low income) and Belarus (high income). The cases were mostly trafficked for forced labour (37%) of domestic nature and sexual exploitations (32%) of flesh trade nature or prostitution, and other exploitation types. The means of control employed by the exploiters are significantly through financial and psychological control.

The second group, ***Cluster 2.2*** however are mostly female cases exploited also mostly in their home countries with significant cases in Eastern European countries, including Ukraine and Moldova, and Montenegro, which are low to middle income countries. The second cluster's cases are exploited mostly for sexual services (prostitution) and are psychologically controlled by exploiters.

The third and fourth clusters are of smaller group (16% and 12% respectively) and are also reported through caseworks with a mix of male and female cases. The third group, ***Cluster 2.3***, is consisted of adults who were exploited abroad. These are Ukrainians, Belarusians, and Kyrgyz individuals who were exploited in Russia and Central Asians exploited in Kazakhstan and Uzbekistan. These are cases from lower income countries, exploited in higher income countries, exploited mostly for forced labour purposes (70%) in construction and domestic work. There is a significant financial control and

withholding of documents that exploiters use to control the trafficked individuals in this cluster.

Finally, the fourth cluster, ***Cluster 2.4*** are profiled to have minor individuals as opposed to cluster 3, and are exploited both in their home country and abroad. The cases exploited in their home countries include Haitians and Senegalese. On the other hand, cases of exploitations abroad in this group happen in Laotians and Burmese nationals exploited in Thailand and Guinea-Bissau nationals exploited in Senegal. These are particularly coming from lower-income country origin cases. The individuals in this group were mostly trafficked for forced labour reasons, with exploiters taking advantage of their minor age through psychological control and physical or sexual abuse.

As discussed previously, there were 4 clusters that were identified by the Agglomerative clustering technique for the most recent time period of **2014-2018**. The first group, ***Cluster 3.1***, makes up 37% of the sample for this time period, and described as cases that were handled onsite by a field officer. These cases are mostly female and older adults. The exploitation happened mostly in their home country (Philippines, Ukraine, Moldova, and Cambodia), with instances of South-eastern Asian origins being exploited in Western Asian countries like Kuwait, UAE, Saudi and Qatar. The latter instances are movement from lower income countries to high income ones. This group has been trafficked mostly for labour purposes in the domestic work field (mostly of South-eastern Asian origin), and sexually exploited for prostitution (mostly Eastern Europeans). With a big portion of 37%, this group were also trafficked for other reasons and are controlled through psychological means and physical or sexual

abuse. This group is the quintessential description of the current human trafficking situation globally and can be dubbed as the “classic group”.

The second group, ***Cluster 3.2***, is also comprised of cases that were managed by a field officer but in contrary from the first group are mostly comprised of male cases and are exploited abroad. These individuals are originally from Myanmar and Cambodia exploited in Indonesia (South-eastern Asians exploited in their neighbouring country), or from Ukraine, Belarus and Kyrgyzstan exploited in Russia (Eastern Europeans and Central Asians exploited also in their region). The cases from this group come from varying income countries, but are moving to a higher incomed ones. These mostly male individuals were exploited for the purposes of construction and manufacturing work (forced labour at 85%) and can arguably be described as the “labour group” in a more recent sense. Exploiters control this group through various means including financially, psychologically, and physically.

The third and fourth group have very similar profile or characteristics including movement, means of control, and exploitation type, and where exploitation happened in the United States except for 1 case in the fourth group. Since the fourth group is a small group of 3% proportionally, it was decided on a logical base to provide more distinction that a regrouping is done by putting the observations of non-US case to the second group and the rest (325 observations) were merged to the third group. This now makes the new proportion as Cluster 3.1 – 4,265 (37%), Cluster 3.2 – 1,037 (9%), and Cluster 3.3 – 6,249 (54%). The third group, ***Cluster 3.3***, in description then is different from clusters 3.1 and 3.2 since the cases were handled mostly through a hotline (the use of either phone, text, or chat in connection to a designation helpline for victims). The individuals

in this cluster were mostly female from the younger age group (minor to young adults) who were all exploited in the United States, from the Philippines, Mexico, China, Colombia and the United States itself. The mentioned origin countries other than the United States are coming lower income countries all exploited in the United States which is a high income country. Majority of the cases in this group are exploited for the purpose of sexual exploitation (72%) particularly for prostitution. The exploiters employ mostly physical and sexual abuse to control this group.

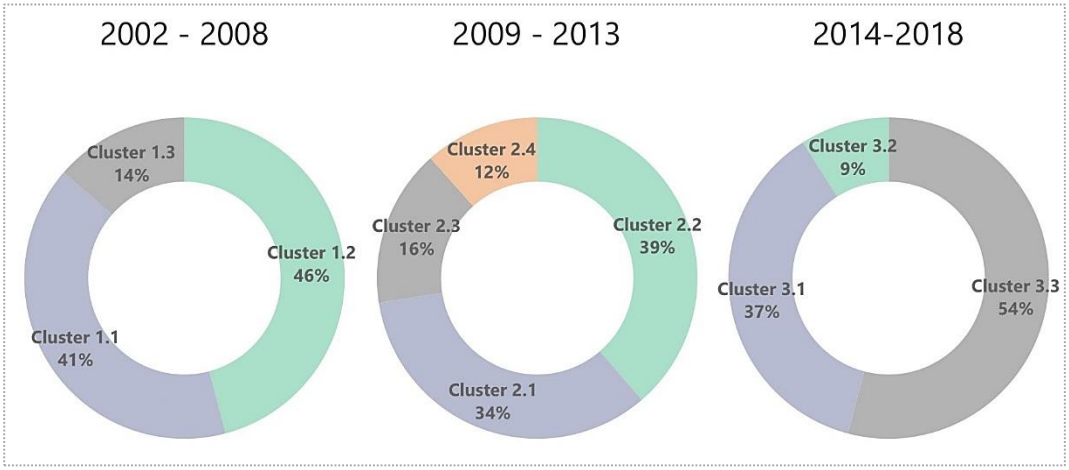

Figure 4.15. Cluster sizes proportion per time period from agglomerative clustering

Cluster	2002-2008	2009-2013	2014-2018
1	877	777	4,265
2	991	888	1,037
3	293	364	6,249
4	-	266	-

Table 4.5. Cluster sizes proportion per time period from agglomerative clustering

4.5. Fuzzy k-mode clustering results

Fuzzy k-mode clustering, another clustering technique for categorical data type was performed after the agglomerative clustering. As mentioned, this technique was performed to compare non-hierarchical clustering for categorical data be compared with the descriptive results of agglomerative hierarchical clustering. The partitioning decision was similar to the agglomerative clustering for profile matching.

Although the 2014-2015 Dunn coefficient provided the highest indicator clustering performance, overall results show mediocre indices indicating indistinct clustering. In addition, the cluster sizes tend to be more even compared to the agglomerative hierarchical clustering results. After relating the clusters back to the dataset for description, the grouping of made little discrepancy to one another. For example, the Fuzzy k-mode clustering failed to distinctly group all US-based cases in one group and where distributed on all clusters in the 2014-2018 time period. Another example for the same time period is that all clusters were predominantly female group, whereas agglomerative hierarchical cluster was able to distinguish the male cases in South-eastern and Central Asia.

Time Period	Dunn Index
2002-2008	0.3989682
2009-2013	0.3487838
2014-2015	0.4498390

Table 4.6. Dunn index coefficients result from Fuzzy k-mode clustering

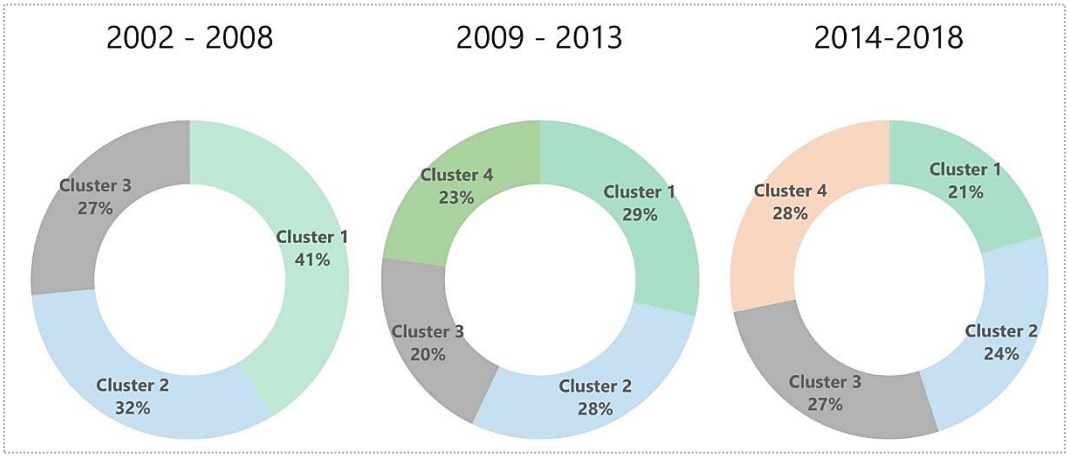

Figure 4.16. Cluster sizes proportion per time period from Fuzzy k-mode clustering

Cluster	2002-2008	2009-2013	2014-2018
1	894	658	2,377
2	694	651	2,816
3	574	464	3,095
4	-	522	3,263

Table 4.7. Cluster sizes proportion per time period from Fuzzy k-mode clustering

5. Conclusion

This paper illustrates how machine learning is highly functional and valuable in social science research of real-world dataset, confirming its potential to the field. Distinctive groups or clusters were identified in the human trafficking dataset from CTDC.

Employment of multiple imputation. Multiple imputation by chained equations (MICE) has been effective in handling the missing data in this study, due to its impressive strength in the preservation of the proportions. MICE or MI overall is a technique that has yet to become familiar to social researchers according to a few authors due to its proven more unbiased results than deletion or single imputation.

Clustering with comparison of techniques. Moreover, agglomerative hierarchical clustering, aided with its heuristics, has been effective in distinctively identifying clusters over fuzzy k-mode on the three time periods the imputed dataset has been segmented.

Inference from the clusters. The segmentation of the imputed dataset into three time periods was for the purpose of identifying trends that related to pattern discovery. The agglomerative clustering was instrumental in the segmented groups and provided a better characteristics and profiling of human trafficking victims over the last 17 years. In more recent years (2014-2018), the descriptive coining of the three groups as “classic”, “labour”, and “US” was possible due to the distinctive clustering results.

With presumably enough insights, research, and enforcement done on human trafficking by subject matter and authorities, this study is a confirmation that machine learning is an advantage and viable aid in the crack-down of human

trafficking. The CTDC dataset has been contributory in proving that machine learning techniques are highly applicable in the knowledge discovery of social problems.

6. Prospective Research

As the first to use the CTDC dataset applied with machine learning techniques per researcher's extent of knowledge, it also sets itself apart from other related studies, able to use case-based record datasets. Prospectively, it is still an adequate motivation to explore other techniques of unsupervised learning for the betterment of clustering or other multiple imputation techniques using supervised learning. Against multiple imputation of missing data, there are literatures that support clustering with missing data and is a possible option to be explored.

As the clusters were identified particularly in the more recent time period, to move further is to use these as labels and to implement profiling checks on individuals through an accomplished learner using the labels.

Social science researchers often have studies on policy effects and post-implementation analysis. With the massive efforts and resources being utilized in as part of counter-trafficking initiatives, it is a considerable angle to look at how policy-makers and implementers can benefit from the machine learning techniques. This study, in relation to resource and effort movement, is already a step on how subject matter experts can review the policies and where efforts should be placed in mitigating the social problem.

References

Aggarwal, C. (2015). *Data mining:the textbook*. New York:Springer.

Ahmad, A. N. (2008). The labour market consequences of human smuggling: 'Illegal' employment in London's migrant economy. *Journal of Ethnic and Migration Studies*, vol. 34 (6): 853-74.

Akay, Ö. and Yüksel, G. (2018). Clustering the mixed panel dataset using Gower's distance and k-prototypes algorithms. *Communications in Statistics-Simulation and Computation*, vol. 47(10), pp.3031-3041.

Alvari, H., Shakarian, P. and Snyder, J.K. (2017). Semi-supervised learning for detecting human trafficking. *Security Informatics*, vol. 6(1), p.1.

Asis, M.M. (2008). Human trafficking in East and South-East Asia: searching for structural factors. *Trafficking in humans: Social, cultural and political dimensions*, pp.181-205.

Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, vol. 20(1), pp.40-49.

Bholowalia, P. and Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*, vol. 105(9).

Buuren, S.V. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, pp.1-68.

Caraway, N. (2006). Human rights and existing contradictions in Asia-Pacific human trafficking politics and discourse. *Tulane Journal of International & Comparative Law*, vol. 14 (2), pp. 295-316.

Cismondi, F., Fialho, A.S., Vieira, S.M., Reti, S.R., Sousa, J.M. and Finkelstein, S.N. (2013). Missing data in medical databases: Impute, delete or classify?. *Artificial intelligence in medicine*, vol. 58(1), pp.63-72.

CTDC (2019). "Telling Stories Through Open Data". [online]. [Accessed 3 February 2019]. Available at: https://www.ctdatacollaborative.org/about-us

DARPA (2019). "About DARPA". [online].[Accessed 19 July 2019]. Available at: https://www.darpa.mil/about-us/about-darpa

data.worldbank.org (2018). "GNI per capita, Atlas method (current US$) Data". [online]. [Accessed 15 March 2019]. Available at: https://data.worldbank.org/indicator/ny.gnp.pcap.cd?year_high_desc=true

Dhar, V. (2012). Data science and prediction. *United States of America: New York University.*

Dubrawski, A., Miller, K., Barnes, M., Boecking, B. and Kennedy, E. (2015). Leveraging publicly available data to discern patterns of human-trafficking activity. J*ournal of Human Trafficking*, vol. 1(1), pp.65-85.

Ford, M., Lyons, L. and van Schendel, W. (2012). Labour migration and human trafficking: an introduction. *In Labour Migration and Human Trafficking in Southeast Asia* (pp. 17-38). Routledge.

Fowler, J.H., Heaney, M.T., Nickerson, D.W., Padgett, J.F. and Sinclair, B. (2011). Causality in political networks. *American Politics Research*, vol. 39(2), pp.437-480.

Franco-Arcega, A., Franco-Sánchez, K.D., Castro-Espinoza, F.A. and García-Islas, L.H. (2014), November. Data Mining for Discovering Patterns in Migration. *In Mexican International Conference on Artificial Intelligence* (pp. 285-295). Springer, Cham.

Górecki, J., Hofert, M. and Holeňa, M. (2017). Kendall's tau and agglomerative clustering for structure determination of hierarchical Archimedean copulas. *Dependence Modeling*, vol. 5(1), pp.75-87.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, vol. 27(4), p. 859.

Grimmer, J. (2014). We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science & Politics*, vol. 48 (01), pp. 80-83.

Haji-Maghsoudi, S., Haghdoost, A.A., Rastegari, A. and Baneshi, M.R. (2013). Influence of pattern of missing data on performance of imputation methods: an example using national data on drug injection in prisons. *International journal of health policy and management*, vol. 1(1), pp.69–77.

Haken, J. (2011). *Transnational crime in the developing world*. Global financial integrity, vol. 12(11).

Han, J., Pei, J. and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. T*he ANNALS of the American Academy of Political and Social Science*, vol. 659(1), pp.48-62.

Hundman, K., Gowda, T., Kejriwal, M. and Boecking, B. (2018). Always Lurking: Understanding and Mitigating Bias in Online Human Trafficking Detection. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 137-143). ACM.

IOM.int (2016). "*G*lobal Trafficking Trends in Focus; IOM Victim of Trafficking Data, 2006-2016". [online]. [Accessed 5 March 2019]. Available at: https://www.iom.int/sites/default/files/our_work/DMM/MAD/A4-Trafficking-External-Brief.pdf

IOM.int (2019). "About IOM". [online]. [Accessed 16 June 2019]. Available at: https://www.iom.int/about-iom

IOM X (2019). "IOM X Story*".* [online]. [Accessed 16 June 2019]. Available at: https://iomx.iom.int/about

Ji, J., Pang, W., Zhou, C., Han, X. and Wang, Z. (2012). A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowledge-Based Systems*, vol. 30, pp.129-135.

Jordan, M.I. and Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. *Science,* vol. 349(6245), pp.255-260.

Kamande, S.W., Miriti, E.A.K., Ahishakiye E. (2018). Consumer Segmentation and Profiling using Demographic Data and Spending Habits Obtained through Daily Mobile Conversations*. International Journal of Computer Applications*, vol. 181(9), pp.33-42.

Kejriwal, M. and Szekely, P. (2018), April. Technology-assisted investigative search: A case study from an illicit domain. *In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (p. CS17). ACM.

Kejriwal, M., Ding, J., Shao, R., Kumar, A. and Szekely, P. (2017). Flagit: A system for minimally supervised human trafficking indicator mining. *arXiv preprint arXiv*:1712.03086.

Konrad, R., Trapp, A., Palmbach, T. & Blom, J. (2017). Overcoming human trafficking via operations research and analytics: Opportunities for methods, models, and applications. *European Journal of Operational Research*, vol. 259 (2), pp. 733-745.

Lee, K.J. & Carlin, J.B. (2012). Recovery of information from multiple imputation: a simulation study. Emerging themes in epidemiology, vol. 9(1), p.3.

Lipton, Z.C. and Steinhardt, J. (2018). Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*.

Lison, P. (2015). An introduction to machine learning. *Norway: University of Oslo.*

Mahmoud, T.O. and Trebesch, C. (2010). The economics of human trafficking and labour migration: Micro-evidence from Eastern Europe. *Journal of comparative economics*, vol. 38(2), pp.173-188.

Misuraca, M., Spano, M. and Balbi, S. (2019). BMS: An improved Dunn index for Document Clustering validation. *Communications in Statistics-Theory and Methods*, vol. 48(20), pp. 5036-5049.

Newman, D.A. (2009). Missing data techniques and low response rates. *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*, vol. 7.

Newman, D.A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, vol. 17(4), pp.372-411.

Ng, M.K. and Jing, L., 2009. A new fuzzy k-modes clustering algorithm for categorical data. *IJGCRSIS,* 1(1), pp.105-119.

Pavoine, S., Vallet, J., Dufour, A.B., Gachet, S. and Daniel, H. (2009). On the challenge of treating various types of variables: application for improving the measurement of functional diversity. *Oikos*, vol. 118(3), pp.391-402.

Poelmans, J., Elzinga, P., Ignatov, D.I. and Kuznetsov, S.O. (2012). Semi-automated knowledge discovery: identifying and profiling human trafficking. *International Journal of General Systems*, vol. 41(8), pp.774-804.

Rdocumentation.org (2019). "mice function | R Documentation". (2019). [online]. [Accessed 18 March 2019]. Available at: https://www.rdocumentation.org/packages/mice/versions/3.6.0/topics/mice

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B. and Rand, D.G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, vol. 58(4), pp.1064-1082.

Ryazantsev, S.V., Karabulatova, I.S., Mashin, R.V., Pismennaya, E.E. and Sivoplyasova, S.Y. (2015). Actual problems of human trafficking and illegal migration in the Russian federation. *Mediterranean Journal of Social Sciences*, vol. 6(3 S1), p.621.

Sabitha, A.S., Mehrotra, D. and Bansal, A. (2014). Similarity based convergence of learning knowledge objects and delivery using agglomerative clustering. *Journal of Information Technology and Application in Education*, vol. 3(1), pp.9-17.

Saha, I., Sarkar, J.P. and Maulik, U. (2019). Integrated Rough Fuzzy Clustering for Categorical data Analysis. *Fuzzy Sets and Systems*, vol. 361, pp.1-32.

Sharma, N. and Gaud, N. (2015). K-modes Clustering Algorithm for Categorical Data. *International Journal of Computer Applications*, vol. 127(1), p.46.

SSRN.com (2019). "social-sciences::SSRN". [online]. [Accessed 1 July 2019]. Available at: https://www.ssrn.com/index.cfm/en/social-sciences/

Sterne, J.A., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M. and Carpenter, J.R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, vol. 338, p.b2393.

UN.org (2019). "United Nations DGACM". (2019). [online]. [Accessed 16 March 2019]. Available at: https://www.un.org/depts/DGACM/RegionalGroups.shtml

UNODC.org (2004). "United Nations Convention Against Transnational Organized Crime And The Protocols Thereto". [online][Accessed 19 March 2019]. Available at: https://www.unodc.org/documents/treaties/UNTOC/Publications/TOC%20Convention/TOCebook-e.pdf

Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton:Chapman and Hall/CRC.

Van den Hoven, J. (2015). Clustering with optimised weights for Gower's metric. *Netherlands: University of Amsterdam*.

White, I.R., Royston, P. & Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, vol. 30(4), pp. 377–399.

Wulff, J.N. and Ejlskov, L. (2017). Multiple Imputation by Chained Equations in Praxis: Guidelines and Review. *Electronic Journal of Business Research Methods*, vol. 15(1).

Zambelli, A.E. (2016). A data-driven approach to estimating the number of clusters in hierarchical clustering. *F1000Research*, vol. 5.

Zhang, Z. (2016.) Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of translational medicine*, vol. 4(2).