COMPARISON OF DATA IMPUTATION METHODS PERFORMANCE FOR MULTIPLE SYSTEM ESTIMATION

(CASE STUDY ON HUMAN TRAFFICKING IN THE NETHERLANDS 2016 - 2019)

Author/Student
Nikolas Anova (2505401)

1st Supervisor
Dr. Maarten Cruyff

2nd Supervisor
Dr. Kyle M. Lang

30 June 2023

MASTER APPLIED DATA SCIENCE

UTRECHT UNIVERSITY

Contents

Acknowledgment

It is a genuine pleasure to express my deepest gratitude to the first supervisor Dr. Maarten Cruyff for mentoring me on the thesis. During this project period, I required patience, experience, and guidance, which the supervisor delivered enthusiastically and professionally. From a student's point of view, I am more than grateful to have such collaboration and support from the supervisor, whose advice helped finalize the thesis. I would like to extend my gratitude to the second supervisor Dr. Kyle M. Lang, for his time, feedback, and interest in my thesis project.

Lastly, I would like to thank Utrecht University and all professors and students from Applied Data Science master studies for one incredible year. As an international student, this journey made me a whole person ready to contribute to the world. I will always reflect on the Utrecht University as remarkable moments that made me a better person and academic citizen.

Abstract

Human trafficking is a problem that still occurs in the modern world, and it is necessary to monitor the number of victims. Since human trafficking is a hidden crime, statistics on identified trafficking victims only reveal a small part of the problem, and the actual number of victims can only be estimated. UNODC recommends using Multiple Systems Estimation (MSE), whereby the size of a hidden population of human trafficking victims is estimated by analyzing the overlap between three or more administrative lists on which persons belonging to that population appear.

In MSE implementation, one of the main problems is missing data. This problem is most likely to occur in the application of MSE due to the use of registration data from several different external sources. The application of the imputation method should be able to solve missing data problems. Since this problem frequently occurs in MSE implementations, however, based on literature reviews, a comparative study of the imputation method performance based on the MSE output has never been conducted. Case in the Netherlands, the missing data problem in human trafficking records also happened in 2016 – 2019. Nevertheless, in previous studies with the same data, multiple imputation was used only with the default method for binary and 2-level categorical data (i.e., logistic regression). The existence of missing data certainly has reduced the quality of population estimates. However, to produce the best MSE output, choosing the suitable imputation method must be done beforehand.

Based on these problems, this study compared the imputation methods performance based on the MSE results in estimating the human trafficking population in the Netherlands from 2016 – 2019. The comparison is seen through the AIC and BIC value of the model. Then the comparison continues between the AIC and BIC version, which is compared based on model complexity, standard error, and reasonableness of estimation. This study focuses on using multiple imputation with seven different methods. These methods are predictive mean matching (PMM), classification and regression trees (CART), random forest, logistic regression, logistic regression with bootstrap, lasso logistic regression, and linear discriminant analysis (LDA).

As a result, different imputation methods produced quite varied MSE model scores and population estimation. The CART method produced the best MSE model compared to other imputation methods. The imputed dataset by CART has the best AIC and BIC scores compared to other imputation methods. The logistic regression method used in previous research produced the rank 6th MSE model in both the AIC and BIC versions. On the other hand, random forest is the imputation method that had the worst MSE model compared to the others. These results show that if there is a problem of missing data in the application of MSE, the choice of the imputation method is proven to affect the quality of the output from MSE.

Keywords: Multiple System Estimation, Human Trafficking, Missing Data, Multiple Imputation, performance comparison

Introduction

1.1 Motivation and Context

Human trafficking is a problem that still occurs in today's modern world. According to the United Nations Trafficking in Persons Protocol, human trafficking is the recruitment, transportation, transfer, harboring, or receipt of people through force, fraud, or deception, intending to exploit them for profit (United Nations, 2000). This problem is fundamental to be resolved for all countries. This urgency is confirmed by the inclusion of the alleviation of Human Trafficking in the Sustainable Development Goals (SDGs) point 16.2 (UNODC, 2015).

Because of the importance of human trafficking, it is necessary to monitor the number/volume of this crime in each country. In SDGs indicator 16.2.2, it is hoped that each country can produce statistics on “the number of victims of human trafficking per 100,000 population, by sex, age, and form of exploitation” (UNSTATS, 2015). This statistic helps provide insight for policymakers to prevent human trafficking crimes.

Since human trafficking is a hidden crime, statistics on identified trafficking victims only reveal a small part of the problem, and the actual number of victims can only be estimated. Moreover, the option to carry out regular population surveys on self-reported victimization by human trafficking is not readily achievable in all countries. Based on these problems, UNODC recommends using Multiple Systems Estimation (MSE), whereby the size of a hidden population of human trafficking victims is estimated by analyzing the overlap between three or more administrative lists on which persons belonging to that population appear (UNODC, 2015).

The implementation of MSE to produce human trafficking indicators is quite successful and has been used in several countries. However, there are some problems in its implementation. One of the main problems in the MSE implementation is missing data (Vincent et al., 2020). This problem is most likely to occur in the application of MSE due to the use of registration data from several different external sources. This missing data problem will undoubtedly affect the quality of the estimation of the MSE method.

In the Netherlands, the missing data problem in human trafficking records also happened in 2016 – 2019 (Van Dijk et al., 2021). CoMensha has been appointed as the official registration organization of all identified possible victims of human trafficking on behalf of the Dutch National Rapporteur. CoMensha obtains data from various sources, such as Police, Inspectorate SZW, and Regional coordinators. From 2016 – 2019, there were missing values in the sense that the joint distribution over all covariates is known, but one or more values are not available. Moreover, since 2018, an application of new European regulations regarding privacy protection has been carried out. As a result, this policy certainly increases the potential for missing data.

The application of the imputation method should be able to solve missing data problems. Van Dijk et al. (2021) implemented multiple imputation to answer these problems. However, the selection of the imputation method was only with the default method for binary and 2-level categorical data (i.e., logistic regression) without looking at other possible methods. The existence of missing data certainly has reduced the quality of population estimates. However, to produce the best MSE output, choosing the suitable imputation method must be done beforehand. It is important because the population estimates produced by the MSE can vary widely with each possible imputed dataset used.

A study is needed to compare the imputation method based on the quality of the output produced by MSE. However, the MSE output in the form of a population estimate is difficult to compare because the actual population size is not known. Comparisons can be made through the AIC or BIC scores generated by the selected MSE model. The complexity of the selected model on MSE can also be used as a comparison. It is based on a study on model assessment and selection in MSE conducted by Cruyff et al. (2020).

As far as the author has carried out the literature study, a comparative study of the imputation method based on the output produced by MSE has never been done before, even though missing data is one of the main problems in the MSE implementation (Vincent et al., 2020). In fact, many comparisons of the imputation method to other analytical tools have been made. For example, Hasan et al. (2017) compared the performance of the imputation method in linear regression analysis. While Olinsky et al. (2003) conducted a similar study on the structural equation model

(SEM) case. Huque et al. (2018) and Yozgatligil et al. (2012) did each in longitudinal and meteorological studies. The last example, Merkle (2011), also conducted a similar study on the case of Bayesian factor analysis.

Based on these problems, this thesis compares the quality of the estimated population of human trafficking victims by MSE from the several imputation methods used. This research focuses on case study datasets from CoMensha. The comparison is seen through the AIC and BIC value of the model, the complexity of the model, standard error, and reasonableness of the population estimation. Meanwhile, the selection of imputation methods is based on the methods that can impute binary or categorical data.

1.2 Literature Review

This sub-section explains in more detail some previous research on the comparison of imputation methods that focus on the application of an analytical tool. The purpose of this literature review is to identify research gaps which are the reasons and references for this study.

Firstly, Hasan et al. (2017) compared the performance of the imputation methods in linear regression analysis. They compared mean imputation, multiple imputation, and maximum likelihood (ML) using simulated data. The performance of the imputation method was assessed using model performance statistics like R-Squared, Adjusted R-Squared, AIC, BIC, MSE, and SSE. This study found that MI and ML always perform better than single imputation (mean imputation). Mean imputation yields very low model performance with a rapid rate of reduction in performance as the percent of missingness increases.

While Olinsky et al. (2003) conducted a similar study on the structural equation model (SEM) case. The five techniques used for comparison are expectation maximization (EM), full information maximum likelihood (FIML), mean substitution (Mean), multiple imputation (MI), and regression imputation (Regression). The study involves two levels of sample size (100 and 500) and seven levels of incomplete data (2%, 4%, 8%, 12%, 16%, 24%, and 32% missing completely at random). The performance of the imputation method was assessed using the variance model. After extensive bootstrapping and simulation, the results indicate that FIML is a superior method in the estimation of most different types of parameters in an SEM format.

Kenyhercz et al. (2016) compared the performance of some data imputation methods with Biodistance Analysis. They selected four data imputation techniques: hot deck, robust iterative model, k-nearest neighbor (kNN = 5), and variable means. Two versions of the data set were then created, wherein values were randomly deleted from each variable so that 25% and 50% of the data were considered missing. Results show that kNN imputation is the most accurate method tested, as it consistently has the lowest difference between actual and imputed values.

Chhabra et al. (2017) did a comparison of multiple imputation methods for data with missing values. However, comparison was made without focusing on a particular analytical tool. They used some multiple imputation methods: predictive mean matching (pmm), classification and regression trees (cart), random forest, linear regression with bootstrap, and Bayesian regression. The performance of the imputation method was assessed using Mean Standard Error, Mean C.I Length.

Based on research on the comparison of imputation methods that have been mentioned and the MSE studies that have been done, the following are the literature gaps identified.

1. One of the main problems in the MSE implementation is missing data (Vincent et al., 2020). This problem is most likely to occur in the application of MSE due to the use of registration data from several different external sources. However, the comparative study of the imputation method that focuses on the application of MSE has never been done before. In fact, several studies on the comparison of the imputation method to other analytical tools have been conducted. As have been conducted by Hasan et al. (2017), Olinsky et al. (2003), and Kenyhercz et al. (2016).
2. A comparison of imputation methods focusing on implementing MSE is possible. Cruyff et al. (2020) conducted an MSE model assessment and selection using the AIC and BIC scores. Hasan et al. (2017) also used the same indicators to compare the imputation method in linear regression cases.

3. Comparison of imputation methods can focus on the multiple imputation. It is based on that Van Dijk et al. (2021) implemented multiple imputation strategy with the default method (logistic regression) to overcome missing data on the same dataset in this thesis. In addition, a comparison of methods on the multiple imputation can be implemented based on research by Chhabra et al. (2017).

1.3 Research Question

Based on the missing data problem in the registration of human trafficking victims in the Netherlands and the motivation to enrich research related to Multiple System Estimation (MSE) mainly related to data imputation, therefore, the research question that this thesis aims to answer is thus as follows:

- How is the performance comparison between the imputation methods based on the results of the Multiple Estimation System (i.e., AIC, BIC, model complexity, standard error, and reasonableness of estimation) in estimating the human trafficking population in the Netherlands 2016 - 2019?

To answer this research question, this study has been conducted and is written in this report. This thesis report explains the data used and the missing data description in the Data section, the imputation methods used and MSE performance indicators comparison details in the Methods section, the results of the comparison in the Result section, then the Conclusion and Discussion section.

Data

2.1 Data Description

This study used data from the Coordination Center against Human Trafficking (CoMensha) from 2016 – 2019. In the Netherlands, CoMensha has been appointed as the official registration organization of all identified possible victims of human trafficking on behalf of the Dutch National Rapporteur. Ideally, all possible victims identified by any organization or person in the Netherlands should be reported to CoMensha. Institutions authorized to carry out criminal investigations into human trafficking are: the National Police (comprising ten regional police districts and one central police unit), coming across all forms of trafficking; Border Police, typically coming across cross-border (sex) trafficking; and the Inspectorate Social Affairs and Employment (Inspectorate SZW) typically coming across cases of labour exploitation. There is no legal obligation for these institutions to report victims they have identified to CoMensha, but they are strongly urged to do so. In addition, designated regional coordinators and other governments, as well as non-governmental institutions – such as organizations providing services to victims/migrants/prostitutes specifically; organizations providing social or legal services; and youth welfare agencies – are invited to report on all cases of presumed victimization. Finally, concerned citizens (or even victims themselves) can identify possible victims and report them to CoMensha directly. Although reporting and registering have steadily improved since this system was implemented several years ago, it cannot be believed that every identified potential victim in the Netherlands is always duly reported.

The data file distinguishes six main categories of registration organizations, namely ISZW (I), KMar (K), Reception (O), Police (P), Regional or care coordination (R), and Other Authorities (Z). The file also contains the covariates of age (adult/minor), sex (male/female), nationality (NL/non-NL), and type of exploitation (sexual and other). Table 1 shows a detailed dataset from coMensha. Meanwhile, the head and tail of this data can be seen in Appendix E.

Table 1: Metadata of Human Trafficking dataset in Netherlands 2016 - 2019 from CoMensha

Variable Name	Data Type	Possible Value
Registrations		
I (ISZW)	Binary	1 (registered); 0 (not registered)
K (KMar)	Binary	1 (registered); 0 (not registered)
O (Reception)	Binary	1 (registered); 0 (not registered)
P (Police)	Binary	1 (registered); 0 (not registered)
R (Regional or care coordination)	Binary	1 (registered); 0 (not registered)
Z (Other Authorities)	Binary	1 (registered); 0 (not registered)
Covariates		
S (sex)	Categorical (2 levels)	M (Male); F (Female)
L (age)	Categorical (2 levels)	A (Adult); M (Minor)
N (Nationality)	Categorical (2 levels)	N (NL); O (Non-NL)
U (Type of Exploitation)	Categorical (2 levels)	S (Sexual); O (Other)
Others		
J (Year)	Categorical (4 levels)	2016 – 2019
Freq	Numeric	The number of observed human trafficking victims
Others Information		
Number of Row	**431**	
Total observed human trafficking victims. (2016 – 2019)	**4742** (2016: **1015**; 2017: **1002**; 2018: **1311**; 2019: **1414**)	

In Ethical and legal considerations, the datasets are fully anonymized, and victims of human trafficking cannot be identified with these data. The datasets are not publicly available and will not be shared with third parties.

2.2 Missing Data Description

One of the main problems in the MSE implementation is missing data (Vincent et al., 2020). This problem is most likely to occur in the application of MSE due to the use of registration data from several different external sources.

In the Netherlands, the missing data problem in human trafficking records happened in 2016 – 2019 (Van Dijk et al., 2021). From 2016 – 2019, there were common missing values in the sense that the joint distribution over all covariates is known, but one or more values are not available. Moreover, since 2018, an application of new European regulations regarding privacy protection has been carried out. As a result, this policy certainly increases the potential for missing data.

Table 2: Frequency of Missing value in each Variable from Human Trafficking dataset

	Variables												Total
	Registrations (12)						Covariates (95)				Others (0)		
	I	K	O	P	R	Z	S	L	N	U	J	Freq	
Frequency of Missing Value	3	3	0	3	2	1	17	25	48	5	0	0	107

Table 2 displays the number of missing values in each variable. Based on Table 2, the total missing value in the human trafficking dataset in the Netherlands 2016 – 2019 was 107. The covariates variables had more missing values than registrations, namely 95 and 12, respectively. Meanwhile, the J and Freq variables had no missing values. Meanwhile, the three covariate variables, namely N, L, and S, had the most significant missing values compared to other variables. The difference in numbers was also quite far compared to other variables.

Methods

3.1 Checking the Missing Data Mechanism

To solve the missing data problem, it is necessary to identify missing data mechanisms that occur in the dataset. Based on this information, it can be determined whether it is necessary to impute the data or just delete the rows containing the missing values. If imputation is required, based on this information helps to determine what imputation method is appropriate. There are three general “missingness mechanisms” moving from the simplest to the most general (Gelman et al., 2017), namely Missingness at completely random (MCAR), Missingness at random (MAR), Missingness Not at Random (MNAR). A common concern when faced with multivariable datasets with missing values is whether the mechanism is MCAR.

In this study, an assessment was carried out to determine whether the missing data in the CoMensha dataset included the MCAR. If classified as MCAR, rows that have missing values are simply deleted. Nevertheless, if not, the imputation is needed on the missing data. The assessment uses Little's (1988) statistical test. The result test statistic is a chi-squared value. The null hypothesis in this test is that the data is MCAR. Little's (1988) statistical test is implemented by using the R libraries "naniar".

3.2 Imputation Methods

Imputation is a method to fill in missing data with plausible values to produce a complete data set. The main reason for carrying out imputation is to reduce nonresponse bias, which occurs because the distribution of the missing values, assuming it was known, generally differs from the distribution of the observed items. When imputation is used, it is possible to recreate a balanced design such that procedures used for analyzing complete data can be applied in many situations. Rather than deleting cases that are subject to item nonresponse, the sample size is maintained, resulting in potentially higher efficiency than case deletion.

In general, imputation methods can be divided into two strategies, namely single and multiple imputation methods. In single imputation, the missing data are filled by some means or models, and the resulting completed data set is used for inference. The imputed value is treated as the true value, ignoring the fact that the no imputation method can provide the exact value. While this strategy allows the inclusion of all cases in a standard analysis procedure, replacing missing values with a single value changes the distribution of that variable by decreasing the variance that is likely present. Single imputation does not reflect the uncertainty about the prediction of the missing values (Pigott, 2001).

Meanwhile, multiple imputation, as proposed by Rubin (1987), fills in missing values by generating plausible numbers derived from distributions of and relationships among observed variables in the data set. Multiple imputation differs from single imputation methods because missing data are filled in many times, with many different plausible values estimated for each missing value. Using multiple plausible values provides a quantification of the uncertainty in estimating what the missing values might be, avoiding creating false precision (as can happen with single imputation). Multiple imputation provides accurate estimates of quantities or associations of interest, such as treatment effects in randomized trials, sample means of specific variables, and correlations between 2 variables, as well as the related variances. In doing so, it reduces the chance of false-positive or false-negative conclusions (Li et al., 2015).

Based on the review above, this study implemented a multiple imputation strategy to solve missing data. This selection is also based on Van Dijk et al. (2021), who also used multiple imputation on the human trafficking dataset in the Netherlands. However, they only used the default method (logistic regression) without considering the possibility of other methods.

Several methods can implement the multiple imputation strategy. Based on the missing data type, this study used several methods that can impute binary or categorical (2 levels) data. Table 3 displays a list of the imputation methods used and compared, accompanied by a brief description.

Table 3: Imputation methods used

No.	Imputation Method	Data Type	Short Description
1.	predictive mean matching (pmm)	Any	For each missing entry, the method forms a small set of candidate donors from all complete cases with predicted values closest to the predicted value for the missing entry.

2.	Classification and regression trees (cart)	Any	cart models seek predictors and cut points in the predictors used to split the sample. The cut points divide the sample into more homogeneous subsamples.
3.	Random Forest	Any	Imputes univariate missing data using random forests. Random forests are a way of averaging multiple deep decision trees trained on different parts of the same training set to reduce the variance.
4.	Logistic regression (logreg)	Binary / 2-level categorical	Imputes univariate missing data using logistic regression. Logistic regression is the appropriate regression analysis when the dependent variable is dichotomous (binary).
5.	Logistic regression with bootstrap (logreg.boot)	Binary / 2-level categorical	Imputes univariate missing data using logistic regression by a bootstrapped logistic regression model. The bootstrap method draws a simple bootstrap sample with replacement from the observed data.
6.	Lasso logistic regression (lasso.logreg)	Binary / 2-level categorical	Imputes univariate missing binary data using lasso logistic regression with Bootstrap. LASSO is a penalized regression approach that estimates the regression coefficients by maximizing the log-likelihood function (or the sum of squared residuals) with the constraint of the sum of the absolute values of the regression coefficients.
7.	Linear discriminant analysis (lda)	Categorical (unordered)	As its name suggests, Linear Discriminant Analysis is a linear model for classification and dimensionality reduction.

Implementation of multiple imputation with methods in table 6 will use the "MICE" library in R. Multivariate Imputation by Chained Equations (MICE) has emerged as one of the principled tool of addressing missing data (Chhabra et. al., 2017). In multiple imputation parameter, 25 complete data sets will be generated then averaged. To ensure successful convergence, the number of iterations will be increased to 10 (default is 5). In addition, because it will be compared, all imputation methods used will use the same seed number.

3.3 Multiple Estimation System

3.3.1 Definition

The statistical technique to estimate the volume of hidden populations, known as capture-recapture analysis, was originally developed by biologists to estimate animal populations. Multiple System Estimation (MSE) can be understood as an advanced version of Capture-Recapture, whereby the size of a hidden population of humans is estimated by analyzing the overlap between three or more administrative lists on which persons belonging to that population appear. Persons belonging to the hidden population of trafficking victims can, for example, be registered by several governmental agencies such as police, immigration, labour inspectors, as well as by private providers of legal, medical, or psychological assistance or child or youth care. By modeling the distribution of the recorded victims over these lists, an estimate can be made of those victims who do not appear on any of the lists from either police, NGOs, or other institutions. Figure 1 illustrates how MSE works to estimate populations from several data sources.

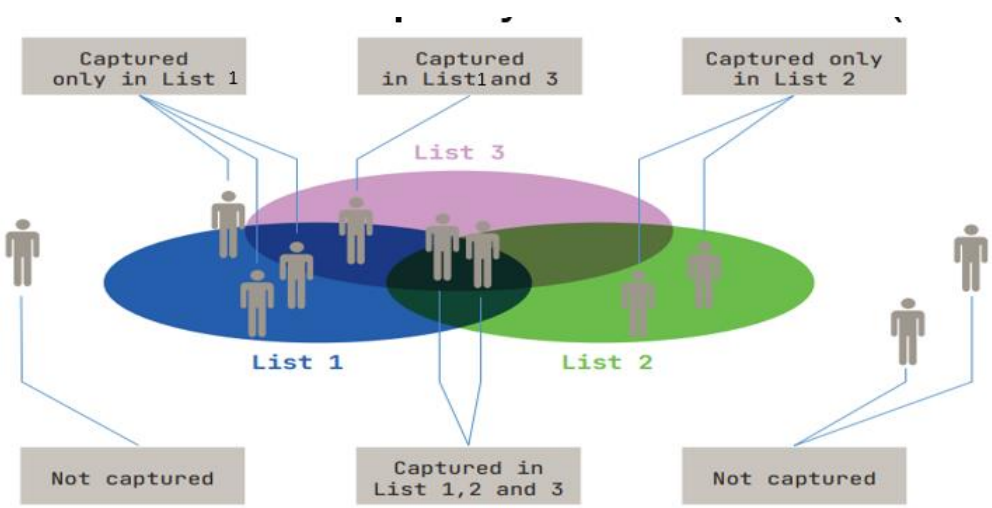

Figure 1: Illustration of Multiple System Estimation (MSE) methods

In its most simple form, the data for MSE consists of a cross-classification of two incomplete population registers, A and B. This results in a two-by-two contingency table with the cells n_{10} representing the number of victims that have been observed in A but not in B, n_{01} representing the number of victims observed in B but not in A, and n_{11} representing victims observed in both

A and B. The cell n_{00} representing the number of victims not observed in A nor in B is to be estimated. The estimation can be performed with the log-linear model (A, B).

$$\text{Log}(u_{ij}) = \lambda_0 + \lambda_i^A + \lambda_j^B$$

For i, j { ∈ 0,1} , where μ_{ij} is the expectation of n_{ij}. The frequencies in the contingency table correspond to the following sets of parameters like in table 4.

Table 4: Example Simple Contingency Table for Log-Linear MSE Model

	B = 0	B = I
A = 0	e^{λ_0}	$e^{\lambda_0+\lambda_I^B}$
A = I	$e^{\lambda_0+\lambda_I^A}$	$e^{\lambda_0+\lambda_I^A+\lambda_I^B}$

The fundamental assumptions of MSE are best explained by discussing the fundamental assumptions of the Dual Systems Estimation (DSE), i.e., MSE with only two lists (or Capture-Recapture). According to an authoritative review (IWGDMF, 1995), data used for DSE must meet four fundamental assumptions:

- the data must relate to persons belonging to a closed population;
- each person must be uniquely identifiable in order to be matched across lists;
- each person must have the same chance to be included on the lists;
- when only two lists are used, the placement of a person on one of the lists must be statistically independent from placement on the other.

The MSE procedure consists of a log-linear analysis of a contingency table of population registers and covariates. As the number of potential models increases exponentially with the number of registers and covariates, it is practically impossible to fit and compare all models. For this reason, model selection is critical. This is essentially a balancing act between model fit (i.e., suitability to observed cell counts) and model complexity (measured by the number of interactions). Typically when performing model selection, associations are assessed using some measure of statistical significance, with only significant associations included in the model.

With a large number of possible models, Cruyff et al. (2020) recommended using the forward selection method. Forward selection is given by completing the following steps.

1. Start by fitting the simplest model under consideration. This is the current model and calculates its Information Criterion (IC).
2. Construct a proposal set of models by augmenting the current model one interaction term at a time (while obeying effect hierarchy).
3. Calculate the IC of every model in the proposal set.

If the IC for the current model is less than the smallest IC from the proposal set, then stop. The current model is the final chosen model. Else, set the current model to be the model with the smallest IC from the proposal set and return to step 1.

3.3.2 Implementation

This study implemented Multiple Estimation System (MSE) to estimate the population of human trafficking victims from the imputed dataset. The fact that CoMensha data are collected by some institutions makes multiple systems estimation (e.g., Silverman, 2020) ideally suited for the estimation of the undetected number of victims. MSE was performed with the log-linear model. The advantage of using log-linear models for population size estimation is that they are easily extended to data with more than two lists and covariates. Therefore the log-linear model will suit the data from CoMensha, which is sourced from 6 registrations and consists of 5 covariates.

In MSE, model selection is critical. As the number of potential models increases exponentially with the number of registers and covariates, fitting and comparing all models is practically impossible. For model selection, this study implemented the two most commonly used information criteria (IC), namely the Akaike Information Criterion (AIC; Akaike, 1974) with $c(p) = 2p$ and the Bayesian Information Criterion (BIC; Schwarz, 1978) with $c(p) = \log(n)p$. AIC aims to choose the model which is optimal in terms of prediction. Conversely, BIC approximately chooses the model that is “closest” to the unknown data-generating process. With many possible models, this study implemented forward selection method.

After getting the selected model, this study estimated the population by year and covariates with parametric bootstrap and computes 95% confidence intervals. Bootstrapping method is a numerical approach to generating confidence intervals that use either resampled data or simulated data to estimate the sampling distribution of the maximum likelihood parameter estimates. Parametric bootstrapping methods do not require that the sampling distribution be known. As a result, they provide a robust and straightforward method to estimate confidence intervals (Nelson, 2008).

To summarize, Table 5 displays the implementation of multiple systems estimation on the human trafficking dataset.

Table 5: Implementation of Multiple Systems Estimation

Model Selection		**Estimating Population**	
Input:	7 imputed datasets from different imputation methods	**Input:**	The 2 best models (AIC and BIC versions.)
Criteria:	- Akaike Information Criterion (AIC)	**Type:**	parametric bootstrap
		Iteration:	1000
	- Bayesian Information Criterion (BIC)	**Seed:**	1
		Output:	Population estimation (95% CI) based on the year and covariates (age, sex, nationality, and Exploitation type) for each model.
Method:	Forward Selection		
Output:	Each imputed dataset gets two models with the smallest AIC and BIC values, respectively (14 models).		
Preprocessing imputed datasets: Turn the imputed dataset into a dataset whose rows consist of all possible combinations of registers and covariates variables. • Total number of cell combinations for the registers: 2^6=64 • Total number of cell combinations for the covariates: 2^4×4=64			

So that the total rows are: 64 x 64 = 4096. However, there are 64 rows where all registers have the value of 0 (cannot be observed). The estimation procedure ignores These cells (i.e., not fitted).
Tools: The MSE implementation in this study uses an R package "mse" (https://github.com/MaartenCruyff/mse).

3.4 Comparison

The main objective of this study is to compare the results of several methods in multiple imputations. Chabra et al. (2017) also made a general comparison of several multiple imputation methods using mice. However, the comparison in this study is specific to the output of the Multiple Estimation System. The following are the comparison stages that have been carried out.

1. Through the MSE (model selection) stage, from 7 imputed datasets, 2 of the best models, the AIC and BIC versions, were obtained. So, there are seven best models of the AIC version and seven best models of the BIC version.
2. In each version, it looked for the model that had the smallest value. Therefore, a total of two best models, the AIC and BIC versions, were selected. In this step, the best imputation method for each version of AIC and BIC was also selected.
3. These two best models became the MSE (estimating population) stage input. As a result, each model got the estimation of the population with a confidence interval.
4. From the two models, this study compared the complexity of the model, standard error, and reasonableness of the population estimation. Based on these criteria, the best model and imputation method were selected.
5. After getting the best model, the estimated results of human trafficking victims from this model were generated based on year and covariates (age, sex, nationality, and Exploitation type) with a 95% confidence interval. Estimation results were displayed in tables and graphs, accompanied by descriptions.

To summarize the data and methods applied in this thesis, Figure 2 displays the workflow of this research process. For implementation in R, an html file containing R Code can be downloaded at the link in appendix D.

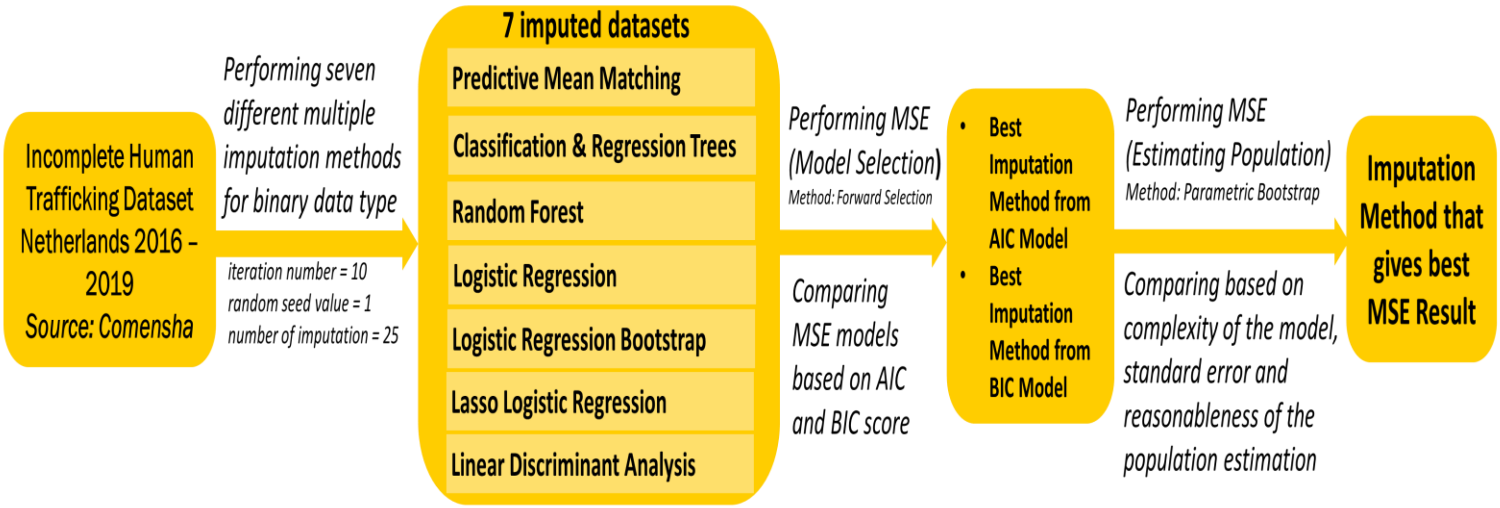

Figure 2: Workflow of Data and Methods

Results

4.1 Missing Data Mechanism

To solve the missing data problem, it is necessary to identify missing data mechanisms that occur in the dataset. A common concern when faced with multivariable datasets with missing values is whether the mechanism is Missingness at completely random (MCAR). If classified as MCAR, rows that have missing values are simply deleted. However, if not, the imputation is needed on the missing data.

Table 6: The results of Little's MCAR Test

	Chi-squared	Df	p.value	Number of missing data patterns
Little's MCAR Test	210	109	0.0000000215	13
H0: Missing Data mechanism is MCAR H1: Missing Data mechanism is not MCAR				

Table 6 displays the results of Little's MCAR Test to determine the missing value in the human trafficking dataset in the Netherlands 2016 - 2019, including the MCAR or not. The p-value was below 0.05, indicating that the null hypothesis (H0) is rejected. Therefore, it can be concluded that the missing value in the human trafficking dataset is not caused by the MCAR mechanism. This conclusion also supports the use of the imputation method in this study.

4.2 Comparison MSE Results from Different Imputation Methods

4.2.1 MSE Model Scores (AIC and BIC)

Seven different imputed datasets were generated from seven different multiple imputation methods. The imputation methods used can impute binary or categorical data. Because they were compared, all imputation methods used the same parameters, such as iteration number (maxit = 10), random seed value (seed = 1), and number of multiple imputation (m = 25).

Each imputed dataset was applied to multiple estimation systems. At this stage, model selection was carried out based on AIC and BIC with the forward method. As a result, in each imputed dataset, the two best MSE models, the AIC and BIC versions, were obtained. The smaller the AIC and BIC scores, the better the quality of the model.

Table 7: Information Criterion (AIC and BIC) score for the Imputation Method used

Imputed Dataset no.	Imputation Method Used	Information Criteria of Model			
		AIC		BIC	
		Score	Rank	Score	Rank
(1)	(2)	(3)	(4)	(5)	(6)
1	predictive mean matching (pmm)	3738	4th	4221.4	4th
2	Classification and regression trees (cart)	3609.9	1st	4092.9	1st
3	Random Forest	4006.8	7th	4489.8	7th
4	Logistic regression (logreg)	3835.1	6th	4308.8	6th
5	Logistic regression with bootstrap (logreg.boot)	3685.1	3rd	4158.2	3rd
6	Lasso logistic regression (lasso.logreg)	3765.3	5th	4248.9	5th
7	Linear discriminant analysis (lda)	3680.5	2nd	4155.5	2nd

First, a comparison was made on the AIC score. Table 7 column 3 displays the AIC score of the best MSE model (AIC version) from the seven imputation methods. It shows that the classification and regression trees (cart) imputation method produces the best AIC version model compared to other imputation methods. The cart had the lowest AIC score compared to other imputation methods.

Meanwhile, the logistic regression (logreg) method, used in previous research (Van Dijk et al., 2021), produced the rank 6th AIC version model. The score distance was also quite far (around 300) by cart. Meanwhile, the modified logistic regression methods, namely bootstrap logreg (logreg.boot) and lasso logreg, were better than the original method and ranked third and fifth. On the other hand, random forest is the imputation method that produced the worst AIC version model compared to the others. Only random forests had AIC scores above 4000.

The next is the comparison for the BIC version model. Table 7 column 5 displays the BIC score of the best MSE model (BIC version) from the seven different imputation methods. It shows that the classification and regression trees (cart) imputation method also produced the best BIC version model compared to other imputation methods. The cart had the lowest BIC score compared to other imputation methods.

Meanwhile, linear discriminant analysis (lda) was the second-best imputation method, but the score differed very slightly from logistic regression with bootstrap (logreg.boot) in the third place. The logistic regression (logreg) method used in previous research (Van Dijk et al., 2021) was the ranked 6th BIC version model. On the other hand, random forest is the imputation method that produced the worst BIC version model compared to the others. Overall, the imputation methods on the BIC version have the same rank order as the AIC version.

To summarize, different imputation methods generated quite varied MSE model scores. At AIC, the seven multiple imputation methods had an average score of 3760.1 with a standard deviation of 120.4. Whereas in BIC, the seven multiple imputation methods had an average score of 4239.3 with a standard deviation of 121.3. Variations in the score of this model are, of course, followed by variations in the resulting population estimation results (Appendix A). This shows that if there is a problem of missing data in the application of MSE, the choice of the imputation method is proven to affect the quality of the output from MSE.

Overall, at this comparison stage, the best AIC model is from the imputed dataset by the CART method. Meanwhile, the best BIC model is also from the CART method. At this stage, it has been concluded that CART is the best imputation method. However, the next stage is still being carried out to compare the 2 best MSE models, the AIC and BIC versions of the CART imputation method.

4.2.2 Complexity of MSE Model

The next stage of comparison is to look at the complexity of the model. The aim is to find parsimonious models that are neither too simple nor too complex and thus trade-off between bias and variance. As a model becomes more complex, the fit to observed counts improves, reducing bias in the population size estimate. However, at the same time, the chance of including spurious interactions increases which can lead to high variability in the estimates. Conversely, a too-simple model does not fit the observed counts, providing estimates with low variability but with potentially high bias.

Model complexity is measured by the number of parameters and interaction variables used in the formula. Table 8 displays a summary of the number of parameters and variable interactions of the best AIC and BIC models from the results of the previous stages (both of which come from CART imputation).

Table 8: The Number of Parameters and Variable Interactions from Best AIC and BIC Models

	Number of Parameters	**Number of Variable Interactions**
Best AIC Model (Appendix C1)	62	51
Best BIC Model (Appendix C2)	50	39

In the best model version of AIC, overall, the total number of parameters was 62. While inside the formula, it had 51 variable interactions. In the best model version of BIC, overall, the total number of parameters was 50. While inside the formula, it had 39 variable interactions.

Based on the number of parameters and interaction variables, both models can be classified as complex models. As a complex model, the fit to observed counts improves, reducing bias in the population size estimate. However, at the same time, the chance of including spurious interactions increases which can lead to high variability in the estimates.

When compared, the AIC model is a more complex model than the BIC model. This can be seen from the number of parameters and variable interactions of the AIC model, which is more than the BIC model. With the principle of looking for a model that is neither too simple nor too complex, the BIC model is better than the AIC model. However, the assessment model continues at the next stage of comparison, which focuses directly on population estimation.

4.2.3 MSE Population Estimation: Standard Error

The following comparison stage focuses on population estimation as the main output of the MSE. Firstly, the population estimation comparison uses the standard error. It assesses how far a sample statistic likely falls from a population parameter. A good MSE model is a model that produces a small standard error.

For the comparison to be valid, the standard error is converted to the normalized standard error, calculated by dividing the standard error by the mean value. The quality of the estimate should be better if the normalized standard error is minor. So the actual population is in the less probable range. Table 9 displays the results of the absolute length of the confidence interval and normalized standard error on the AIC and BIC models.

Table 9: Comparison of Normalized Standard Error between AIC and BIC Model

Year	Best AIC Model					Best BIC Model				
	Nhat	min95	max95	length CI	Normalized Standard Error	Nhat	min95	max95	length CI	Normalized Standard Error
2016	4512	3513	6258	2745	0.16	6504	5279	8173	2894	0.11
2017	3208	2550	4181	1631	0.13	5133	4213	6465	2252	0.11
2018	6023	4735	7884	3149	0.13	7751	6409	9698	3289	0.11
2019	5812	4647	7814	3167	0.14	6063	5039	7524	2485	0.10

From Table 9, in 2016 – 2019, the AIC model produced population estimations with higher normalized standard error than the BIC model. Although in absolute value, the AIC model's confidence interval length was shorter than the BIC model.

The difference in the normalized standard error between the two models was not that big, but it still significantly affected the quality of the population estimate. The most significant normalized standard error gap between AIC and BIC occurred in 2016, with a value of 0.05. The gap narrowed in 2017 and 2018, with a value of 0.02. However, the gap increased again in 2019, with a value of 0.04.

In both models, in absolute value, the confidence interval length from 2016 to 2019 fluctuated quite a bit. After being standardized, the normalized standard error value in the AIC model still tends to fluctuate from 2016 to 2019. On the other hand, the BIC model tends to stagnate.

Overall, based on the normalized standard error, the BIC model has a better estimation of the human trafficking population than the AIC model.

4.2.4 MSE Population Estimation: Reasonableness

Comparison of population estimates is continued by looking at the reasonableness. It can be seen by comparing it with the observed human trafficking population. If the estimated number is too large compared to the observed population, it can be concluded that the estimation results are not good enough. Table 10 presents the ratio between estimated and observed populations in both models (AIC and BIC).

Table 10: Comparison between Estimated and Observed Human Trafficking Population

Year	Observed Population	Best AIC Model			Best BIC Model		
		Estimated Population	Difference (Est– Obs)	Ratio (Est / Obs)	Estimated Population	Difference (Est– Obs)	Ratio (Est / Obs)
2016	1015	4512	3497	4.4	6504	5489	6.4
2017	1002	3208	2206	3.2	5133	4131	5.1
2018	1311	6023	4712	4.6	7751	6440	5.9
2019	1414	5812	4398	4.1	6063	4649	4.3

Table 10 shows that the AIC model has a larger estimated population than the observed population. The largest ratio of differences occurred in 2016 and 2018, when the estimated populations were 4.4 and 4.6 times larger than the observed population, respectively. In 2017 and 2019, each period had an estimated population of around 3 and 4 times larger than the observed population.

Table 10 also shows that the BIC model has a larger estimated population than the observed population. The most significant ratio of differences occurred in 2016, with an estimated population 6.4 times larger than the observed population. The second largest difference ratio occurred in 2018, nearly six times. In 2017 and 2019, each period had an estimated population of 5.1 and 4.3 times larger than the observed population.

By comparing the ratio between the AIC and BIC models from 2016 – 2019, the AIC model had a smaller ratio than the BIC model. However, both models still produce reasonable populations.

To be even more convincing, the reasonableness of the population is also seen from a comparison with the results of previous research by Van Dijk et al. (2021), which uses the same dataset. The reasonableness of the population is seen by looking at the similarity of the ratio between the estimated and the observed population. The more similar the ratio value, the population can be said to be reasonable. Table 11 presents a comparison between estimated and observed populations by Van Dijk et al. (2021).

Table 11: Comparison between Estimated and Observed Human Trafficking Population from Previous Research

Year	Observed Population	Van Dijk et. al (2021) Model			AIC Model Ratio	BIC Model Ratio
		Estimated Population	Difference (Est– Obs)	Ratio (Est / Obs)		
2016	1015	4196	3181	4.1	4.4	6.4
2017	1002	2947	1945	2.9	3.2	5.1
2018	1311	5435	4124	4.1	4.6	5.9
2019	1414	5139	3725	3.6	4.1	4.3

Table 11 shows that the AIC and BIC models had larger ratios than the Van Dijk et al. model from 2016 to 2019. In other words, the population estimations from the AIC and BIC models are greater than those from the Van Dijk et al. model. However, the two models did not produce estimates that differed significantly from Van Dijk et al. model. The ratio value of the AIC model was only slightly more similar to Van Dijk et al. model than the BIC model.

Overall, based on the two assessment models above, it can be concluded that both models (AIC and BIC) have reasonable population estimation.

4.2.5 Recapitulation of Comparison MSE Results

The best models of the AIC and BIC versions were compared based on the complexity of the models, standard error, and reasonableness of the population estimation. Table 12 displays the recapitulation of the comparison results.

Table 12: Recapitulation of Comparison Results

	Comparison				
	First Stage		**Second Stage (Comparison between 2 best models from first stage, AIC and BIC version)**		
	AIC score	**BIC score**	**Model Complexity**	**Standard error of Population Estimation**	**Reasonableness of Population estimation**
Result	AIC best model is from CART imputation method.	BIC best model is from CART imputation method.	BIC model is better than AIC	BIC model is better than AIC	Both models' estimation are reasonable

Based on the recapitulation in Table 12, it can be concluded that overall, the BIC version of the model produces a better MSE output (estimated human trafficking population) than the AIC. In the case of imputation methods, with the human trafficking dataset in the Netherlands, the classification and regression trees (cart) imputation method has the best MSE model compared to other imputation methods. The selected MSE model from the cart imputation method (BIC version) can be seen from appendix C3.

4.3 Human Trafficking Population based on Selected Model

In this section, the results of estimating the human trafficking population based on the selected model, namely the BIC version of the CART imputation method, will be discussed. Note that these figures are not official statistics.

Table 13 and Figure 3 show the overall estimation results by year. In Table 13, the ratio between the observed victims and the total estimate from 2016 – 2019 ranges from 4 to 6 times. These results indicate that the number of unobserved human trafficking victims in the Netherlands is quite a lot compared to those recorded. The population had decreased from 2016 to 2017. However, it increased quite drastically in 2018 (the highest in this period), then decreased again in 2019. This pattern is similar to the estimation results by Van Dijk et al. (2021).

Table 13: Population Estimation of Human Trafficking Victims in Netherlands 2016 – 2019

Year	Observed Population	Estimated Total Population	Ratio (Est / Obs)
2016	1015	6504	6.4
2017	1002	5133	5.1
2018	1311	7751	5.9
2019	1414	6063	4.3

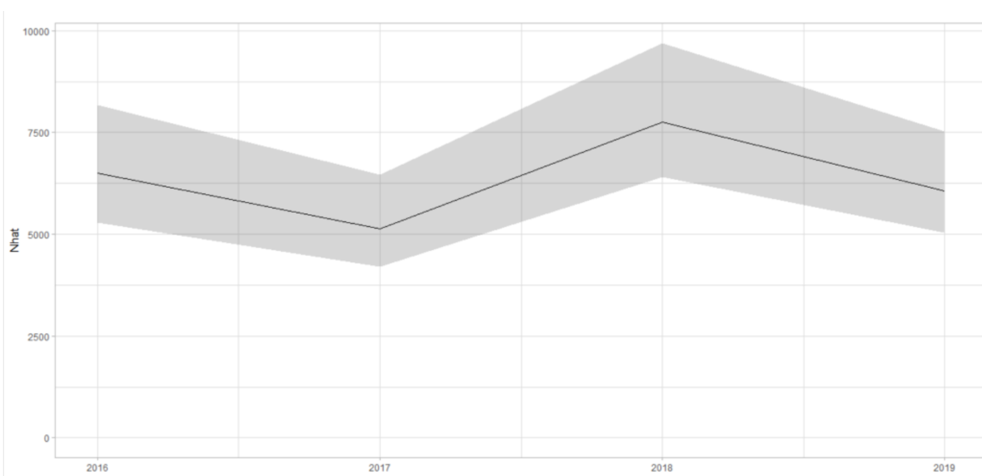

Figure 3: Population Estimation of Human Trafficking Victims in Netherlands 2016 – 2019

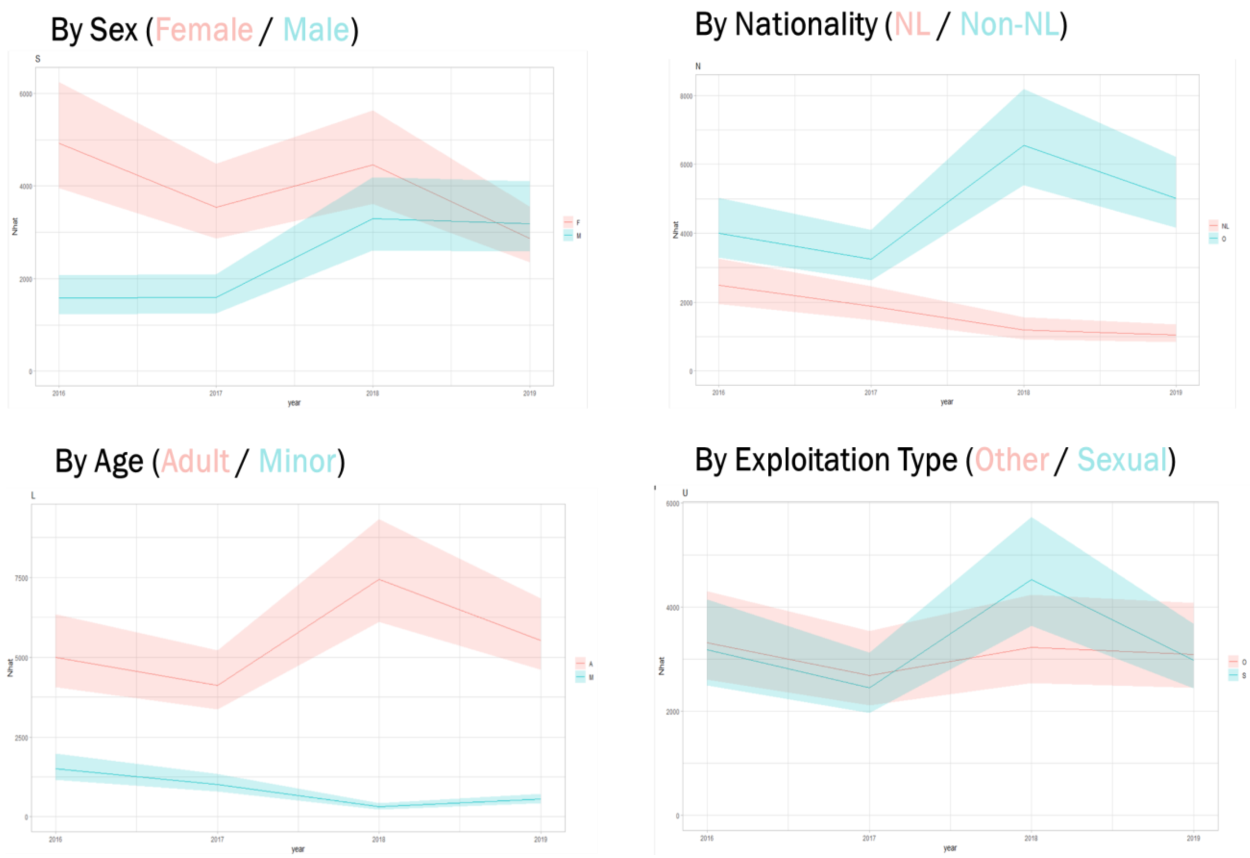

Figure 4: Population Estimation of Human Trafficking Victims in Netherlands 2016 – 2019 by Covariates (Sex, Age, Nationality, and Exploitation Type)

Figure 4 on the top left displays the population estimation of human trafficking victims by Sex (Male / Female) in the Netherlands from 2016 – 2019. In 2016 and 2017, the number of female victims was more than male; non-intersecting confidence intervals show this. However, in 2018 and 2019, the number of male victims tended to increase, and the confidence interval between the number of male and female victims intersected. Therefore, it cannot be concluded which sex was bigger or smaller in number.

Figure 4 on the bottom left displays the population estimation of human trafficking victims by age (Adult / Minor) in the Netherlands from 2016 – 2019. For four years, the number of adult victims was always greater than minor victims. It was shown that the confidence intervals did not overlap and were quite far apart. Apart from that, information was also obtained that the number of minor victims continued to decrease from 2016 to 2019. Meanwhile, the number of adult victims fluctuated quite a bit.

Figure 4 on the top right displays the population estimation of human trafficking victims by nationality (NL / Other) in the Netherlands from 2016 – 2019. For four years, the number of victims of non-Netherlands nationality was always greater than victims of Netherlands nationality. It can be concluded that the confidence intervals were not intersecting. In addition, information was also obtained that the number of NL victims continued to decline from 2016 to 2019. Meanwhile, the number of non-NL victims fluctuated quite a bit and tended to increase.

Figure 4 on the bottom right displays the population estimation of human trafficking victims by exploitation type (Other / Sexual) in the Netherlands from 2016 – 2019. In four years, the number of victims of sexual and other exploitation tends to be the same. This was concluded because the confidence intervals continued to overlap, especially in 2016 and 2017.

Appendix B shows the complete results of population estimation for human trafficking victims in the Netherlands based on year (2016 – 2019) and covariates (age, sex, nationality, and Exploitation type) from the selected model.

Conclusion and Discussion

5.1 Conclusion

Based on the results, the following is the conclusion to answer the research question.

- This study found that different imputation methods produce quite varied MSE model scores. Moreover, AIC and BIC score versions produced the same imputation method performance ranking. From the seven multiple imputation methods used to impute data on Human Trafficking Victims in the Netherlands 2016 - 2019, the classification and regression trees (CART) method produced the best Multiple System Estimation (MSE) models compared to other imputation methods. The imputed dataset by CART has the best AIC and BIC scores compared to other imputation methods. The logistic regression method used in previous research produced the rank 6th MSE model. Meanwhile, the modified logistic regression methods, namely bootstrap logreg and lasso reglog, were better than the original method and ranked third and fifth. On the other hand, random forest is the imputation method that produced the worst MSE model compared to the others. Compared by version, the BIC model from the CART method is better than the AIC version. This conclusion can be seen from the comparison of the complexity of the model, standard error, and reasonableness of the population estimation. The BIC model is better on model complexity and standard error criteria. Meanwhile, both models have reasonableness in their population estimates.

5.2 Discussion

These results show that if there is a problem of missing data in the application of MSE, the choice of the imputation method is proven to affect the quality of the output from MSE. In this case, Classification and Regression Trees is the best imputation method. Its performance exceeds the Logistic Regression (logreg) method, used in previous research (Van Dijk et al., 2021). However, the results of this study do not strictly recommend the use of any imputation method.

For future research, this comparison can be continued by using dummy data. That will certainly make it easier to do simulations and can indicate a variety of conditions. For example, the performance of the imputation method can be seen in more detail at different percentage levels or distributions of missing data. Moreover, this research still uses the default MSE hyperparameter. It will be interesting if future studies carry out more detailed performance on different MSE hyperparameters. Of course, a simulation like this will be able to provide more detailed recommendations for suitable imputation methods.

Moreover, the results of the estimation of human trafficking victims in the Netherlands in this study are not official statistics. This research focuses more on comparing the imputation method's performance.

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control.

Chhabra G., Vashisht V., Ranjan J. (2017). A Comparison of Multiple Imputation Methods for Data with Missing Values. Indian Journal of Science and Technology, Vol 10. https://sciresol.s3.us-east-2.amazonaws.com/IJST/Articles/2017/Issue-19/Article14.pdf

Cruyff M, Overstall A., McRea R. (2020). Multiple System Estimation of Victims of Human Trafficking: Model Assessment and Selection. https://doi.org/10.1177/0011128720981908

Fienberg, S. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. Biometrika, vol. 59.

Gelman, A. and Hill, J. (2007) Ch 25: Missing-data imputation in Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York.

Hasan H., Ahmad S. (2017). A comparison of model-based imputation methods for handling missing predictor values in a linear regression model: A simulation study. https://doi.org/10.1063/1.4995930

Huque M. D., Carlin J. B., Simpson J. A., Lee K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0615-6

IWGDMF (International Working Group for Disease Monitoring and Forecasting). (1995). Capture-recapture and multiple record systems estimation. Part 1. History and theoretical development. American Journal of Epidemiology

Kenyhercz, M., & Passalacqua, N. (2016). Missing Data Imputation Methods and Their Performance With Biodistance Analyses. *Biological Distance Analysis*, 181-194. https://doi.org/10.1016/B978-0-12-801966-5.00009-3

Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple Imputation: A Flexible Tool for Handling Missing Data. JAMA, 314(18), 1966. https://doi.org/10.1001/jama.2015.15281

Little R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association Volume 83, 1988. https://www.tandfonline.com/doi/ref/10.1080/01621459.1988.10478722?scroll=top&role=tab&aria-labelledby=refs

Merkle E. C. (2011). A Comparison of Imputation Methods for Bayesian Factor Analysis Models. https://doi.org/10.3102/1076998610375833

Nelson, W. (2008). Statistical Methods. *Encyclopedia of Ecology*, 3350-3362. https://doi.org/10.1016/B978-008045405-4.00661-3

Olinsky A., Chen S., Harlow S. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. https://doi.org/10.1016/S0377-2217(02)00578-7

Pigott T. D. (2001). A Review of Methods for Missing Data. Educational Research and Evaluation vol. 7 no. 4 pp. 353-383.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, New York, Chichester.

Schwarz, G. E. (1978). Estimating the dimension of a model. Annals of Statistics.

Silverman W. B. (2020). Multiple-systems analysis for the quantification of modern slavery: classical and Bayesian approaches. Journal of the Royal Statistical Society. https://www.nottingham.ac.uk/research/beacons-of-excellence/rights-lab/resources/academic-publications/2020/silverman-multiple-systems-analysis-for-the-quantification-of-modern-slavery.pdf

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

UNODC. (2015). Sustainable Development Goals. https://www.unodc.org/roseap/en/sustainable-development-goals.html

UNODC. (2015). Research Brief: Multiple Systems Estimation for estimating the number of victims of human trafficking across the world.

UNSTATS. (2015). SDG indicator 16.2.2 metadata. https://unstats.un.org/sdgs/metadata/files/Metadata-16-02-02.pdf

Van Dijk, J.J.M., Cruyff, M. & Van der Heijden, P.G.M. (2021). A Multiple Systems Estimation of presumed victims of human trafficking in the Netherlands during 2016-2019. https://open.overheid.nl/documenten/ronl-e19ddd79-02ad-424b-bccc-b2767f794a98/pdf

Vincent K. S., Far S. S., Papathomas M. (2020). Common Methodological Challenges Encountered With Multiple Systems Estimation Studies. https://doi.org/10.1177/001112872098190

Yozgatligil C., Aslan S. (2012). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. https://link.springer.com/article/10.1007/s00704-012-0723-x

Appendix

Appendix A:

Complete Results of Implementation of Multiple System Estimation (i.e., model selection) on Seven Imputed Datasets from Seven Different Imputation Methods

Imputation Method Used	**Information Criteria of Model**					
	AIC			**BIC**		
	Score	**Rank Score**	**Population (2016 – 2019)**	**Score**	**Rank Score**	**Population (2016 – 2019)**
predictive mean matching (pmm)	3738.0	4th	20210	4221.4	4th	22768
Classification and regression trees (cart)	3609.9	1st	19555	4092.9	1st	25451
Random Forest	4006.8	7th	15221	4489.8	7th	16353
Logistic regression (logreg)	3835.1	6th	19145	4308.8	6th	26388
Logistic regression with bootstrap (logreg.boot)	3685.1	3rd	19944	4158.2	3rd	17298
Lasso logistic regression (lasso.logreg)	3765.3	5th	19051	4248.9	5th	17812
Linear discriminant analysis (lda)	3680.5	2nd	25172	4155.5	2nd	31869

Appendix B:

Complete Results of Implementation of Multiple System Estimation (i.e., estimation population) on the Selected Model (from the CART imputation method with the BIC version) for Human Trafficking Victims in the Netherlands based on year (2016 – 2019) and covariates (age, sex, nationality, and Exploitation type)

Year	Observed Victims	Estimated Victims	Confidence Interval (95)		Ratio (est /obs)
2016	1015	6504	5279	8173	6.4
2017	1002	5133	4213	6465	5.1
2018	1311	7751	6409	9698	5.9
2019	1414	6063	5039	7524	4.3

Age	Year	Observed Victims	Estimated Victims	Confidence Interval (95)		Ratio (est /obs)
Adult (>=18)	2016	788	4995	4056	6336	6.3
	2017	813	4118	3366	5214	5.1
	2018	1248	7439	6116	9328	6.0
	2019	1277	5517	4612	6833	4.3
Minor (< 18)	2016	227	1509	1163	1984	6.6
	2017	189	1015	793	1350	5.4
	2018	63	312	218	428	5.0
	2019	137	547	421	714	4.0

Sex	Year	Observed Victims	Estimated Victims	Confidence Interval (95)		Ratio (est /obs)
Female	2016	806	4921	3960	6248	6.1
	2017	768	3538	2873	4485	4.6
	2018	866	4455	3610	5634	5.1
	2019	768	2873	2350	3557	3.7
Male	2016	209	1583	1229	2069	7.6
	2017	234	1595	1256	2092	6.8
	2018	445	3296	2609	4192	7.4
	2019	646	3191	2588	4097	4.9

Nationality	Year	Observed Victims	Estimated Victims	Confidence Interval (95)		Ratio (est /obs)
Netherlands	2016	326	2498	1944	3270	7.7
	2017	333	1885	1487	2464	5.7
	2018	223	1198	922	1561	5.4
	2019	251	1057	844	1350	4.2
Others	2016	689	4006	3288	5029	5.8
	2017	669	3248	2643	4090	4.9
	2018	1088	6553	5397	8194	6.0
	2019	1163	5007	4168	6216	4.3

Form Of Exploitation	Year	Observed Victims	Estimated Victims	Confidence Interval (95)		Ratio (est /obs)
Others	2016	383	3320	2610	4300	8.7
	2017	361	2683	2112	3543	7.4
	2018	408	3227	2540	4235	7.9
	2019	496	3088	2449	4084	6.2
Sexual	2016	632	3184	2498	4143	5.0
	2017	641	2450	1964	3129	3.8
	2018	903	4523	3641	5732	5.0
	2019	918	2976	2445	3669	3.2

Appendix C: MSE Models Formula and Parameter

C1. Best AIC Model from CART imputation Method

```
Freq ~ I + K + O + P + R + Z + S + L + N + U + J + O:U + S:U +
    S:N + I:U + L:N + R:N + P:J + S:J + P:R + O:Z + L:J + Z:L +
    K:N + K:S + O:N + R:S + L:U + R:J + P:U + K:P + N:J + Z:N +
    P:Z + I:L + K:L + K:J + U:J + I:Z + K:U + O:L + I:P + I:N +
    I:R + O:P + Z:U + R:U + S:L + P:S + K:O + O:J + I:J + K:R +
    I:O + Z:S + N:U + I:S + P:L + I:K + Z:J + K:Z + O:S
```

C2. Best BIC Model from CART imputation Method

```
Freq ~ I + K + O + P + R + Z + S + L + N + U + J + O:U + S:U +
    S:N + I:U + L:N + R:N + P:J + S:J + P:R + O:Z + L:J + Z:L +
    K:N + K:S + O:N + R:S + L:U + R:J + P:U + K:P + N:J + Z:N +
    P:Z + I:L + K:L + K:J + U:J + I:Z + K:U + O:L + I:P + I:N +
    I:R + O:P + Z:U + R:U + S:L + P:S + K:O
```

C3. The Selected/Best MSE model from the CART imputation method with BIC version (with parameter values)

```
(Intercept)        I1        K1        O1        P1        R1        Z1
       6.02     -5.49     -7.47     -4.76     -3.83     -3.40     -7.02
         SM        LM        NO        US     J2017     J2018     J2019
      -1.41      0.20      0.54      0.54     -0.32     -1.05     -1.26
      O1:US     SM:US     SM:NO     I1:US     LM:NO     R1:NO  P1:J2017
      -1.65     -2.33      1.67     -4.76     -1.87     -0.63      0.14
   P1:J2018  P1:J2019  SM:J2017  SM:J2018  SM:J2019     P1:R1     O1:Z1
      -0.31      0.82      0.40      0.99      1.32      1.39      2.42
   LM:J2017  LM:J2018  LM:J2019     Z1:LM     K1:NO     K1:SM     O1:NO
      -0.17     -1.46     -0.64      1.93      3.30     -1.71      2.07
      R1:SM     LM:US  R1:J2017  R1:J2018  R1:J2019     P1:US     K1:P1
      -1.06     -0.85      0.59      0.84      0.32      1.64     -2.58
   NO:J2017  NO:J2018  NO:J2019     Z1:NO     P1:Z1     I1:LM     K1:LM
      -0.09      0.79      0.59      1.23      0.88     -1.52     -2.24
   K1:J2017  K1:J2018  K1:J2019  US:J2017  US:J2018  US:J2019     I1:Z1
       0.58     -0.81     -0.72      0.02      0.84      0.52      2.50
      K1:US     O1:LM     I1:P1     I1:NO     I1:R1     O1:P1     Z1:US
       1.51      0.67      1.22      1.37      1.66      0.56      1.12
      R1:US     SM:LM     P1:SM     K1:O1
       0.65     -0.53      0.39     -2.08
```

Appendix D:

Complete R Code and Output

The HTML document containing the R Code and the output of this thesis can be downloaded at the link below:

https://drive.google.com/file/d/1-URFrdqtvy2Jp4dgCPbcAgiXGZVqbvKt/view

Appendix E: Human Trafficking datasets in Netherlands 2016 – 2019 from CoMensha

Head

	I	K	O	P	R	Z	S	L	N	U	J	Freq
1	0	1	0	0	0	0	F	A	NL	O	2016	1
2	0	0	0	1	0	0	F	A	NL	O	2016	8
3	0	0	0	0	1	0	F	A	NL	O	2016	11
4	0	0	0	0	0	1	F	A	NL	O	2016	3
5	1	0	0	0	0	0	M	A	NL	O	2016	1
6	0	0	0	0	1	0	M	A	NL	O	2016	1
7	0	0	1	0	1	0	M	A	NL	O	2016	1
8	0	0	0	1	1	0	M	A	NL	O	2016	1
9	1	0	0	1	1	0	M	A	NL	O	2016	1
10	0	0	0	1	0	0	F	M	NL	O	2016	10
11	0	0	0	0	1	0	F	M	NL	O	2016	16
12	0	0	0	0	0	1	F	M	NL	O	2016	1
13	0	0	0	0	1	0	M	M	NL	O	2016	1
14	0	0	0	0	1	0	F	NA	NL	O	2016	1

Tail

	I	K	O	P	R	Z	S	L	N	U	J	Freq
419	0	0	0	1	0	0	M	M	O	S	2019	2
420	0	0	0	0	1	0	M	M	O	S	2019	1
421	0	0	0	1	1	0	M	M	O	S	2019	1
422	NA	NA	1	NA	NA	1	NA	M	O	S	2019	17
423	0	0	0	1	0	0	F	A	NA	S	2019	2
424	0	0	0	1	0	0	M	A	NA	S	2019	1
425	0	0	0	0	1	0	M	A	NA	S	2019	1
426	0	0	0	1	0	0	F	M	NA	S	2019	1
427	0	0	0	0	1	0	NA	M	NA	S	2019	1
428	0	0	0	0	0	1	NA	NA	NA	S	2019	10
429	NA	NA	1	NA	1	NA	F	M	O	NA	2019	7
430	0	0	1	0	0	0	NA	M	O	NA	2019	3
431	0	0	0	0	1	0	F	NA	NA	NA	2019	7