Applied Data Science Master Thesis

From marginal to joint frequencies in the context of Multiple Systems Estimation

Author : Vasileios Alexiou

Student Number : 2717328

First Supervisor : Dr. Maarten Cruyff

Second Supervisor : Dr. Kyle Lang

Department of Information and Computing Science
Utrecht University

27th of June 2023

Abstract

Population size estimates are important for understanding social phenomena and for informing political decisions. However, privacy restrictions that prevent the disclosure of identities make accurate population estimates difficult. This study addresses these challenges related to victims of trafficking in the Netherlands. I focus on transforming covariate marginal frequencies into joint frequencies in order to understand the relationships within the dataset. Current approaches to transform marginal frequencies into joint are limited in capturing the underlying relationships between covariates. To solve this problem, this work proposes a methodology that uses the Kronecker product to efficiently generate potential combinations of covariate levels. The dataset used consists of marginal frequencies from six population registries associated with Dutch anti-trafficking organizations. This provides insight into the distribution of covariates and highlights differences in gender, age, nationality and type of exploitation. Analysis of joint frequencies reveals patterns and insights, highlighting the vulnerability of specific populations and the prevalence of various forms of exploitation. This result shows the importance of analysis of joint distributions to obtain meaningful results. The proposed methodology bridges the gap between marginal and joint distributions, contributes to the estimation of multiple systems, and improves our understanding of human trafficking in the Netherlands.

Keywords: Multiple Systems Estimation, human trafficking, marginal frequencies, joint frequencies, covariates, population registers, the Netherlands.

Introduction - Literature Review

Estimating the size of a population is a crucial task in social and behavioral sciences, providing valuable insights for policy-making, resource allocation, and program evaluation. Multiple Systems Estimation (MSE) is a widely used technique that leverages information from multiple incomplete population registers to estimate the population size. Simply put, it counts the overlap of victims appearing in different combinations across multiple data sources, (Lyneham, et al., 2019).

The original version of MSE is the mark – recapture estimation for estimating population size. The basic idea is as follows. Suppose that we want to estimate the number of fish in a pond. You catch a large number of fish (eg 100), tag them in some way, and release them. After a while you catch more fish and check how many of the new catches (eg 100 more), were part of the original first catch. For example, if the overlap of both catches is 20, the natural estimate of the total would be 500. MSE extends this idea when there are more than two catches (lists). For example, let's consider a population of individuals who can be part of five different lists. Each individual may appear in one, two, three, four, or all five lists. To estimate the number of individuals who are not present in any of the lists, we can examine the various combinations of lists and count the individuals who appear in one list but not in the others. In total, there are 31 observable combinations representing different overlap patterns among the lists. To apply MSE, we require not only the counts of individuals in each list, but also the knowledge of the sizes of all possible overlaps. By analyzing the patterns of overlaps, MSE allows us to estimate the number of individuals who are not captured in any of the lists. This approach is different from the capture-recapture model, which typically assumes two lists and estimates population size based on the number of individuals appearing in both lists.

MSE offers a complete picture of the population even when some people are absent from particular registers by connecting people across registers. When the identity of the people is excluded owing to privacy concerns, though, this is no easy process. I will have to overcome the challenge of non-identity disclosure when I examine a dataset that contains characteristics (covariates) of victims of human trafficking in the Netherlands. Using numerous datasets, Cruyff, et al., 2021 have thoroughly investigated the prevalence and dynamics of human trafficking in the Netherlands. These studies have specifically looked into the amount of suspected human trafficking victims from 2010 to 2015. Building upon this prior research, my study aims to contribute to the existing knowledge by analyzing a comprehensive dataset covering the period from 2018 to 2019. But first, what is human trafficking?

Human trafficking refers to the illegal and exploitative trade of human beings, involving the recruitment, transportation, transfer, harboring, or receipt of individuals through force, fraud, or coercion for the purpose of (sexual) exploitation. The Netherlands has a complex history and policies regarding prostitution and human trafficking, (Staring, 2009). The Dutch authorities took action in 2000 by removing the prohibition on brothels in the Netherlands, with the goal to establish a system of regulation for prostitution, (Huisman & Kleeman, 2014). Human trafficking though, continues to exist

within the licensed sector and has moved to less regulated areas such as escort services and the internet. Moreover, estimates from the National Rapporteur on Trafficking in Human Beings, which was established in 1997 to monitor and combat human trafficking, suggest that a large number of foreign prostitutes are undocumented. The Dutch government has taken steps to fight human trafficking by conducting financial investigations, adopting a programmatic approach to combat organized crime networks involved in human trafficking, and providing protection and support to victims. Victims can stay in the Netherlands for a longer period to cooperate with law enforcement, and they have the option to apply for a residence permit. However, finding employment after the trial can be challenging for them.

By bridging the gap between marginal distributions and joint distributions, we can uncover the hidden truths and reveal the complete picture, enabling us to infer meaningful results and make a lasting impact in combatting this crime. The integration of data from multiple sources can offer valuable insights (Xie et al., 2009). However, it often falls short of providing the complete joint distributions. Instead, researchers are left with a collection of marginal distributions that only partially reveal the relationships within the combined domain. To illustrate this, let's dive into the details of our human trafficking case. The dataset is composed of six population registers. Due to privacy regulations, however, the registers do not disclose the identity of the victims, so that linkage with other registers is not possible. They did disclose tables with the frequencies of the covariates, like age and gender. For example, they report to have seen 30 males and 40 females, and 15 minors and 55 adults. For the MSE we need to know how the joint distribution, e.g. how many female minors there are, but this information is not available. This highlights the necessity of moving beyond marginal distributions and toward joint distributions to accurately infer meaningful results from the available data.

The above example illustrates the problem of ecological inference. The issue of ecological inference revolves around partial identification, which means that obtaining precise conclusions is often challenging without gathering individual-level data that can uniquely identify each unit, (Glynn & Wakefield, 2010). The problem arises when we try to infer individual-level relationships from aggregate-level observations and it is common in various fields, including political science, sociology, and epidemiology, where researchers often seek to understand the behavior of individuals within a larger group or population. However, due to the aggregation of data, important nuances and heterogeneity at the individual level can be lost, leading to potential biases and inaccuracies in the results. The problem of ecological inference necessitates the development of specialized statistical methods and modeling techniques that carefully account for the limitations and challenges associated with making inferences at the individual level based on aggregate-level data. Researchers strive to find a balance between the available group-level information and the need to accurately capture the underlying individual-level processes, ultimately enhancing our understanding of complex social, political, and epidemiological phenomena.

Research Question:

How do we get joint frequencies that are likely based on the marginal frequencies?

Objective:

Until now, researchers have used a naive approach to convert marginal distributions into joint distributions. The procedure involved identifying the covariate with the highest observed frequency and creating rows in a data frame to represent the missing joint distributions. This process was repeated for each organization (register), considering the frequencies of the covariates.

However, this approach is considered naive for several reasons. First, it is a manual and error-prone method that relies on ad-hoc decision-making. The selection of the covariate with the highest frequency as the starting point may not always yield the optimal results, as it does not take into account the underlying relationships or dependencies among the covariates. Additionally, this procedure leads to a loss of covariate information. By only focusing on a subset of the covariates, other important covariate information is disregarded and not considered in the conversion process. This limitation reduces the accuracy and completeness of the resulting joint distributions. Furthermore, grouping together cases with missing values under a single category (e.g., using NA for all variables except for the register of interest) oversimplifies the representation of the missing data. It fails to capture the potential variations and patterns that might exist within the missing data, which could be important for accurate estimation.

The primary objective of this thesis is to find an accurate and efficient way of transforming the given marginal frequencies of our “private” dataset to joint frequencies. A more ideal approach allows for a more comprehensive understanding of the relationships between the covariates and the observed frequencies, providing insights into the joint distribution of the variables of interest. Below we can see analytical information about the dataset used for my research.

Data

2018 dataset

	Male	Female	SexNA	Adult	Minor	AgeNA	NL	notNL	NatNA	Sexual	notSex	ExpNA	cases
Veiling Thuis	0	8	0	4	4	0	7	1	0	8	0	0	8
Zoco Limburg	*NA*	*NA*	4	*NA*	*NA*	4	*NA*	*NA*	4	*NA*	*NA*	4	4
Zoco Arnhem	*NA*	1	10	10	1	0	*NA*	*NA*	11	10	1	0	11
Zoco Rotterdam	7	62	0	*NA*	41	28	37	26	16	32	*NA*	37	69
Nidos	2	6	0	2	6	0	0	8	0	*NA*	*NA*	8	8
Zoco Friesland	17	184	0	44	38	119	68	62	71	42	3	156	201
Zoco Gelderland	0	3	0	2	1	0	3	0	0	3	0	0	3
IOM	0	2	0	2	0	0	0	2	0	0	2	0	2
Zoco Groningen	4	36	0	36	4	0	24	16	0	34	2	4	40
SMO Den Bosch	1	5	0	6	0	0	1	5	0	5	1	0	6
Fairwork	146	61	3	207	3	0	0	210	0	0	210	0	210
Scharlaken	0	17	0	17	0	0	7	9	1	17	0	0	17
Zoco Utrecht	1	18	0	12	6	1	9	9	1	18	1	0	19
Terwille	0	8	0	8	0	0	7	1	0	8	0	0	8
Zoco Oost	13	41	0	35	19	0	33	14	7	52	4	0	54

2019 dataset

	Male	Female	SexNA	Adult	Minor	AgeNA	NL	notNL	NatNA	Sexual	notSex	ExpNA	cases
Fairwork	0	1	0	1	0	0	0	1	0	0	1	0	1
Moviera	2	11	1	8	5	1	6	5	3	6	2	6	14
Moviera Gelderland Zuid	*NA*	*NA*	2	2	0	0	2	0	0	2	0	0	2
Moviera Utrecht	0	1	0	1	0	0	0	1	0	1	0	0	1
Nidos	16	17	0	0	33	0	0	33	0	3	*NA*	30	33
Scharlaken koord	0	7	0	6	1	0	2	5	0	7	0	0	7
Terwille	0	3	0	1	*NA*	2	0	3	0	3	0	0	3
Zoorgcoordinatie Eindhoven	0	1	0	1	*NA*	0	1	0	0	1	0	0	1
Zoorgcoordinatie Rotterdam	*NA*	13	0	8	0	1	5	5	3	12	*NA*	1	13

Each row is an organization associated with addressing or providing support to vulnerable populations or social issues in the Netherlands. More details are shown below:

The 3 registers of our dataset are the following:

O : Fairwork, Nidos, SMO

R : Zocos

Z : Veiligthuis, Scharlaken Koord, Terwille, Moviera

- ZOCO Limburg, Arnhem, Rotterdam, Friesland, Gelderland, Groningen, and Utrecht are regional centers for asylum seekers. Asylum seekers are given temporary housing, care, and support at these facilities, also known as "Zorg- en Opvangcentrum" (Care and Reception Centers), while their applications are being reviewed.
- Nidos organization is in charge of looking after unaccompanied minor asylum seekers. Nidos ensures these children's welfare throughout the asylum process and offers them legal and social support.

- IOM is the International Organization for Migration. It is an intergovernmental body that offers services and guidance to governments and immigrants alike regarding migration. The IOM promotes safe, regular, and orderly migration all around the world. They help migrants in a number of ways, including relocation, migration health, and voluntary return and reintegration.
- SMO Den Bosch is located in Den Bosch, the Netherlands. It is an organization that specializes in offering support and care services to those dealing with homelessness, addiction, mental health concerns, or other social obstacles. They provide programs to aid people in reintegrating into society as well as housing, counseling, and other services.
- Fairwork is an organization that strives to end modern slavery and worker exploitation. They offer assistance and resources to employees who are at risk, such as those who have been the victims of forced labor, human trafficking, or other forms of exploitation. These employees' rights are promoted by Fairwork, which also helps them have access to support, protection, and justice.
- Terwille is a Dutch nonprofit organization that specializes in offering aid and support to those who have been touched by prostitution and human trafficking. They provide a range of services, such as safe housing, counseling, and vocational training, to assist people in reestablishing their lives and escaping abusive circumstances.
- The Netherlands-based non-profit Scharlaken Koord focuses on offering assistance and support to those who are involved in or impacted by prostitution. The organization primarily assists sex workers and seeks to advance their wellbeing, security, and empowerment.
- Veiligthuis fights against and prevents elder and child abuse as well as marital violence. The name "Veiligthuis" means "Safe Home" in English, which reflects the organization's dedication to establishing a secure setting for people and families dealing with violence or abuse.
- Moviera is another Netherlands-based non-profit organization committed to helping those impacted by prostitution, human trafficking, and domestic violence. Putting a lot of focus on empowering survivors, with the help of volunteers and donators.

The columns are consisted of the covariates of the victims. More specifically we have the marginal frequency of the victims' :

- Gender (Male, Female)
- Age (Adult, Minor)
- Nationality (NL, not NL)
- Exploitation (Sexual, Other)

Moreover I have created the missing covariate (gender, age, nationality, exploitation) columns, which count the cases where the marginal frequency of each covariate is unknown to us. Finally, the last column counts the total human trafficking cases for each asylum.

The dataset used in this thesis is made up entirely of marginal frequencies and provides details on a number of covariates, including gender, age, nationality, and the type of exploitation experienced by victims of human trafficking. It is important to note that the dataset protects the privacy and identity of

the victims by not revealing their names. As a result, specific ethical considerations regarding the protection of personal information do not apply in this context.

Now, we will delve into the distributions of each covariate (Gender, Age, Nationality, Exploitation) for every organization and each year (2018 and 2019).

2018 covariate distributions :

First, let's examine the gender distribution (Figure 1). The dataset reveals variations in the representation of genders across organizations. It becomes evident that most organizations, for example Zoco Friesland and Zoco Rotterdam exhibit a higher proportion of females compared to males, whereas the organization Fairwork is the only one that demonstrates a larger number of males involved in exploitation cases. It is interesting to note that some Zoco Arnhem and Zoco Limburg are mostly consisted of missing values for the gender covariate (SexNA).

Moving on to the age distribution (Figure 2), the dataset provides valuable insights into the age groups affected by exploitation. Organizations such as Fairwork, Zoco Oost and Zoco Utrecht show a higher incidence of exploitation cases involving adults, while Nidos and Zoco Rotterdam predominantly deal with victims below the age of 18. Additionally, there are instances (Zoco Friesland, Zoco Limburg and Zoco Rotterdam) where there is missing information about the marginal distribution of the age covariate (AgeNA).

The dataset also allows for an exploration of the distribution of victims based on their nationality. Figure 3 shows that organizations like SMO Den Bosch, IOM, Faiwork and Nidos are totally constituted of individuals with non-Dutch nationalities (notNL), indicating the involvement of foreign nationals in exploitation situations. Conversely, organizations like Zoco Rotterdam, Zoco Oost and Zoco Groningen report a larger number of cases involving individuals with Dutch nationality (NL). It is noteworthy that there are instances where the nationality information is missing or unknown (Zoco Arnhem and Zoco Limburg), emphasizing the complexities of identifying and documenting the nationalities of victims.

Finally, figure 4 provides an important perspective on the prevalence of sexual exploitation cases across organizations. The frequency of sexual exploitation cases does not vary significantly, with exception the organization “Fairwork”, which is reporting a significantly high number of non-sexual exploitation cases. Zoco Limburg is consisted of missing values about the exploitation type and Zoco Arnhem too has many

cases where there are missing values for this covariate. Finally, the main type of abuse for the rest of the organizations is sexual.

2019 covariate distributions :

Similarly to the 2018 dataset procedure, it is wise to start by examining the gender distribution for 2019. The dataset reveals variations in the representation of genders across organizations (Figure 5). It becomes evident that most organizations, such as Fairwork, Moviera, and Terwille, exhibit a higher proportion of females compared to males. However, Nidos show an equal representation of both genders. Interestingly, there are missing values for the gender covariate (SexNA) in some organizations, such as Moviera Gelderland Zuid and Scharlaken koord.

Moving on to the age distribution (Figure 6), the dataset provides valuable insights into the age groups affected by exploitation. Organizations like Nidos predominantly deals with victims below the age of 18. Conversely, Scharlaken Koord, Zoorgcoordinatie Eindhoven, Moviera Utrecht, Moviera Gelderland Zuid and Fairwork primarily involve adults in exploitation cases. Additionally, there are instances (Moviera Gelderland Zuid, Terwille and Zorgcoordinatie Rotterdam) where there is missing information about the marginal distribution of the age covariate (AgeNA).

The dataset also allows for an exploration of the distribution of victims based on their nationality (Figure 7). Organizations like Zorcoordinatie Rotterdam, Moviera, and Scharlaken Koord have a mixture of individuals with Dutch and non-Dutch nationalities, indicating a diverse range of victims. On the other hand, Moviera Utrecht, Fairwork, Nidos and Terwile predominantly involve individuals with non Dutch nationality. It is worth mentioning that there are instances where the nationality information is missing or unknown, such as in Moviera and Zorgcoordinatie Rotterdam.

Regarding the type of abuse, according to Figure 8 the frequency does not significantly vary across organizations, except for Fairwork, which reports only non-sexual exploitation cases. This comes in contrast with the rest of the organizations, which primarily report sexual exploitations. Finally, once again there are missing values for the exploitation type in Moviera Gelderland Zuid and Moviera.

The process of converting marginal frequencies to joint frequencies involves several steps, which are explained below, where I discuss the methodology which I will follow.

Methodology

To address my research question, I will apply the following method to estimate the expected frequencies. In detail I will:

- Calculate the expected frequencies, by converting the observed marginal frequencies into joint frequencies. The expected frequencies represent the frequencies we would expect to see based on the observed marginal frequencies and the assumption of independence between the two covariates.

- To calculate the expected frequencies, we create a list of potential combinations of the levels of the given covariates and compute the number of permutations (i.e., the number of times each pairing of levels can occur). In this context the Kronecker Product can be utilized to efficiently generate the list of potential combinations of the levels of the given covariates. The Kronecker Product is a mathematical operation that combines two matrices to produce a larger matrix, (Van Loan, 2000). It essentially expands the dimensions of the matrices by multiplying each element of one matrix with every element of the other matrix.

- The expected frequency for each combination is obtained by multiplying the number of permutations by the observed marginal frequency and dividing it by the total number of permutations. This calculation assumes that each permutation is equally likely.

The above method is motivated by the need to understand the relationships and associations between different covariates in the dataset. By estimating the joint frequencies, we can gain insights into how the covariates are related and understand whether a specific individual belongs to a certain register or not.

Results

Table 1 shows the function which was used to convert the marginal, to joint frequencies. The function takes three input parameters: **covs**, **freqs**, and register. The covs parameter represents the covariates of interest, such as gender, age, nationality, and type of exploitation. The freqs parameter is a list containing marginal frequency data from our population registers. Finally, the register parameter specifies the name of the population register being analyzed. In detail :

1. The function begins by calculating the total number of cases by summing the frequencies of the first element in the freqs list, (Line 4).
2. Next, the function utilizes the kronecker function to generate the joint frequencies. It iteratively applies the Kronecker product operation to combine the frequencies from all elements in the freqs list. The result is stored in the variable **Freq,** (Line 8).

3. To establish the joint distribution, the function creates a data frame, **d**, using the *expand.grid* function, which represents all possible combinations of covariate levels in a reverse order, (Line 10).
4. The **out** data frame is then initialized with the value 1 in the first column, followed by the covariate combinations stored in d, and the expected frequencies calculated by multiplying the cases by the joint frequencies Freq and dividing it by the sum of Freq, (Line 12).
5. Finally, the name of the population register being analyzed is assigned to the first column of out using the names function, (Line 15).

After constructing the functions, we need to do some data manipulation to extract the covariate combinations in a clear format. This involves the following steps :

1. Add up the frequencies of the organizations for each register using the colSums function. As our dataframe is consisted of all organizations , but not the registers, we need to find the correct indexes and fill each register (R, O, Z), (Lines 21-23).
2. Then, we create lists for covariates (covs) and frequencies (freqs) for each register. Each list contains sub-lists corresponding to different covariates (S, A, N, E) within the register. The resulting lists are stored in variables covs_R, covs_O, covs_Z, freqs_R, freqs_O, and freqs_Z, (Lines 26-31).
3. The marg2joint function is executed for the organization R using the covs_R and freqs_R lists as inputs. The resulting joint distribution is stored in the Rjoint dataframe. The same procedure is repeated for the rest of the registeres, creating 2 more (Ojoint and Zjoint) dataframes, (Lines 39, 52, 66).
4. Finally, several mutations are applied to the all three dataframes . The values in the columns S, A, N, E are replaced based on specific conditions using case_match function, where if the value matches "SexNA", "AgeNA", "NatNA", or "ExpNA", it is replaced with NA, otherwise, it remains the same. The resulting dataframes are stored back after removing unused levels using the droplevels function, (Lines 43-48, 56-61, 70-75).

Below, we can see some statistics for both the 2018 and 2019 datasets, accompanied by the highest frequency – combinations for each register.

<u>2018 dataset:</u>

Table 2. *Summary statistics for 2018 dataset*

Frequency	Register R	Register O	Register Z
Min.	0.02806	0	0
1st Qu.	0.56243	0	0
Median	1.09361	0	0

Mean	4.95062	2.76543	0.3333
3rd Qu.	3.07410	0.01425	0
Max	27.61320	134.11211	12.9630

From the summary of the R register for 2018 (Table 2) we can reveal interesting patterns and insights about the individuals in the dataset. The range of frequencies varies from a minimum value of 0.02806 to a maximum value of 27.61320, with a mean frequency of 4.95062. This indicates that while some combinations are infrequent, there are others that occur more frequently within the dataset. The distribution of frequencies shows a median value of 1.09361 and a third quartile value of 3.07410. These statistics highlight the variation in the occurrence of different combinations and the importance of addressing the higher frequency combinations in understanding and addressing the issue of sexual exploitation.

Based on the summary of Register O's joint frequencies, the majority of those have a value of 0. The maximum frequency in register O is 134.1. As we previously saw, this frequency represents the "Male – Adult – NotNL – notSex" combination of covariates. Finally, the mean frequency for Register O is 2.76543. This indicates that, on average, the joint frequencies for the combinations of covariates in Register O are relatively low. It suggests that there is a wide distribution of frequencies across the different combinations, with many combinations having very low frequency.

Finally, the maximum frequency in Register Z is 12.9630. The mean frequency for Register Z is 0.3333. This indicates that, on average, the joint frequencies for the combinations of covariates in Register Z are relatively low. It suggests that there is a sparse distribution of frequencies across the different combinations, with many combinations having very low or even zero frequencies.

Table 3. *5 highest-frequency combinations for register R, 2018*

S	A	N	E	Freq
Female	<NA>	NL	<NA>	27.61320
Female	<NA>	NL	Sexual	26.23941
Female	Adult	NL	<NA>	25.25155
Female	Adult	NL	Sexual	23.99525
Female	<NA>	notNL	<NA>	20.15446

Table 3 reveals a considerable presence of females from the Netherlands in register R. The top 5 combinations with the highest frequencies reveal significant patterns in the dataset. The most common combination consists of females from the Netherlands, where specific information about age and exploitation is not available. Another prevalent combination includes women from the Netherlands who have experienced sexual exploitation. Additionally, there is a notable frequency of adult, Dutch females, both with and without information about sexual exploitation. Finally, there is a relatively high frequency

of females from countries other than the Netherlands, indicating the presence of a diverse group in the same register.

Table 4. *5 highest-frequency combinations for register O, 2018*

S	A	N	E	Freq
Male	Adult	notNL	notSex	134.112109
Female	Adult	notNL	notSex	64.805851
Male	Minor	notNL	notSex	5.613995
Male	Adult	notNL	<NA>	5.084819
Male	Adult	notNL	Sexual	3.178012

In Table 4, the most prevalent combination involves adult males from countries other than the Netherlands who have not experienced sexual exploitation. This suggests a significant representation of adult males from non-Dutch backgrounds in register O, potentially indicating specific migration or labor patterns. Following closely is a combination of adult females from non Dutch countries who have not experienced sexual exploitation. This highlights the presence of adult females in similar demographic categories, potentially facing similar challenges. Additionally, there is a relatively low frequency of minor males from countries other than the Netherlands who have not experienced sexual exploitation. Furthermore, there are combinations of adult males from countries outside of the Netherlands, both with and without information about sexual exploitation.

Table 5. *5 highest-frequency combinations for register Z, 2018*

S	A	N	E	Freq
Female	Adult	NL	Sexual	12.9629630
Female	Adult	notNL	Sexual	11.1111111
Female	Adult	NL	notSex	1.0370370
Female	Adult	<NA>	Sexual	0.9259259
Female	Adult	notNL	notSex	0.8888889

Finally, table 5's top 5 combinations with the highest frequency for register Z in 2018 show that adult females, primarily from the Netherlands, are most frequently the victims of sexual exploitation, with a frequency of 12.963. Adult females of non-Dutch nationalities who have been sexually exploited constitute another statistically significant combination, with a frequency of 11.111. The three last combinations are again constituted by adult females and their frequencies are significantly lower.

<u>2019 dataset:</u>

Table 6. *Summary statistics for 2019 dataset*

Frequency	Register R	Register O	Register Z
Min.	0	0	0.001829
1st Qu.	0	0	0.025606
Median	0	0	0.076818
Mean	0.1728	0.4198	0.333333
3rd Qu.	0	0	0.243256
Max	5.0143	15.4152	5.351623

Based on the summary of register R's joint frequencies (Table 6), we can gain valuable information about the characteristics of individuals within the dataset. The range of frequencies in register R varies from a minimum value of 0 to a maximum value of 5.0143. This indicates that the range of frequencies for the different combinations of covariates in register R for 2019 is not as wide as that for 2018. The mean frequency for register R is 0.1728, suggesting that, on average, the joint frequencies for the combinations of covariates in register R are relatively low. It indicates that some combinations occur more frequently, while others are less common within the dataset. The distribution of frequencies in register R shows a median value of 0, indicating that there is a central tendency around this value. The third quartile value of 0 suggests that a significant proportion of the combinations have frequencies below this threshold.

In register O, the frequencies exhibit a wide range, from a minimum value of 0 to a maximum value of 15.4152. This indicates that there is considerable variability in the frequencies of the different combinations of covariates in O, even higher than that of register R. The mean frequency for O is 0.4198, suggesting that, on average, the joint frequencies for the combinations of covariates in this register are relatively low. It indicates that there is a wide distribution of frequencies across the different combinations, with many combinations having very low or even zero frequencies. The distribution of frequencies shows a median value and a third quartile value of 0. This solidifies the aforementioned statement, that a large number of combinations have very low or negligible frequencies within the dataset.

Finally, upon examining the summary of register Z's joint frequencies we notice that they range from a minimum value of 0.001829 to a maximum value of 5.351623. This suggests that, once again, there is considerable variation in the frequencies of the different combinations of covariates within Z, lower though than that of register O. The mean frequency for register Z is 0.3333, so on average, the joint frequencies for its combinations of covariates are relatively low. It suggests that there is a sparse distribution of frequencies across the different combinations, with many combinations having very low or even zero frequencies. The distribution of frequencies shows a median value of 0.076818 and a third quartile value of 0.243256 too, which signify as well that a significant portion of the combinations have very low or negligible frequencies within the dataset.

Table 7. *5 highest-frequency combinations for register R, 2019*

S	A	N	E	Freq
Female	Adult	NL	Sexual	5.0142857
Female	Adult	notNL	Sexual	4.1785714
Female	Adult	<NA>	Sexual	2.5071429
Female	<NA>	NL	Sexual	0.5571429
Female	<NA>	notNL	Sexual	0.4642857

The top 5 register R frequency combinations for 2019 are shown in Table 7. It is clear from the analysis of the data that most victims in these combinations are adult females. Their nationality covariate varies, but their sexual exploitation covariate is constant. With a frequency of 5.014, the most frequent combination involves Dutch victims who have been the victims of sexual exploitation. Moreover, a frequency of 4.178 proves that a group of victims of human trafficking from nations other than the Netherlands who have also been sexually exploited follows closely. Additionally, there are combinations in register R where specific information about nationality is missing (NA), but sexual exploitation is reported, indicating the presence of unidentified victims. Lastly, the fifth highest frequency combination includes citizens from countries outside the Netherlands who have not experienced sexual exploitation.

Table 8. *5 highest-frequency combinations for register O, 2019*

S	A	N	E	Freq
Female	Minor	notNL	<NA>	15.4152249
Male	Minor	notNL	<NA>	13.7024221
Female	Minor	notNL	Sexual	1.5415225
Male	Minor	notNL	Sexual	1.3702422
Female	Minor	notNL	notSex	0.5138408

Table 8 explores the top 5 register O combination frequencies for 2019. With a frequency of 15.415, the combination of female adolescents from nations other than the Netherlands and no information concerning sexual exploitation (NA) appears to be the most common. Similar to female minors, male minors from non-Dutch nationalities also have a high incidence of 13.702. It is noteworthy that the frequency of sexual exploitation is lower in the combinations, indicating that it could not be the main issue for minors inside register O.

Table 9. *5 highest-frequency combinations for register Z, 2019*

S	A	N	E	Freq
Female	Adult	notNL	Sexual	5.351623
Female	Adult	NL	Sexual	3.822588
Female	Minor	notNL	Sexual	1.783874
Female	Adult	notNL	<NA>	1.689986

Female	Minor	NL	Sexual	1.274196

The 5 combinations with the highest frequencies for register Z in the same year are examined in Table 9. Combinations in this record mostly include female adults of non-Dutch nationalities who have been the victims of sexual exploitation. Adult females from non-Dutch backgrounds who have experienced sexual exploitation make up the most frequent combination, with a frequency of 5.351. Dutch adult females who have also been sexually exploited, albeit at a much lower frequency of 3.822, constitute another interesting combination. Once again, it is important to note that combinations without sexual exploitation and combinations with missing nationality information (NA) also exist.

Conclusion and Discussion

In conclusion, this thesis has addressed the research question of how to obtain joint frequencies that are likely based on marginal frequencies in the context of Multiple Systems Estimation (MSE). The research question focused on finding an accurate and efficient method to transform marginal frequencies into conditional (joint) frequencies.

Through the methodology employed, which involved calculating expected frequencies and utilizing the Kronecker product to generate potential combinations of covariate levels, this thesis successfully estimated joint frequencies based on observed marginal frequencies. The approach considered the relationships and dependencies among the covariates, providing a more comprehensive understanding of the dataset and enabling insights into the joint distribution of variables of interest.

The implications for the proper domain setting are significant. By moving beyond marginal distributions and delving into joint distributions, researchers can gain a holistic understanding of the relationships within the domain. This knowledge enhances the accuracy and completeness of population size estimations, contributing to effective policy-making, resource allocation, and program evaluation in social and behavioral sciences. Moreover, in the specific domain of human trafficking, understanding the joint frequencies of covariates can aid in identifying vulnerable populations, informing targeted interventions, and improving victim support and protection services.

This research also needs to take ethical implications and issues into account. Since the dataset used in this thesis only included marginal frequencies without revealing individual identities, it safeguarded the privacy and anonymity of each victim. This method made sure that moral standards were followed and protected private data.

In summary, by creating a precise and effective approach to estimate joint frequencies based on marginal frequencies in the setting of MSE, this thesis has successfully addressed the research question. The results give a thorough understanding of relationships within the domain, guiding interventions and policy-making to combat human trafficking. The study also acknowledges and addresses the ethical implications, highlighting the significance of ethical considerations when researching delicate subjects like human trafficking.

REFERENCES

Glynn, A. N., & Wakefield, J. (2010). Ecological inference in the social sciences. Statistical Methodology, 7(3), 307-322. https://doi.org/10.1016/j.stamet.2009.09.003 von Eye, A., Mun, E.-Y., & Mair, P. (2012). Log-linear modeling. WIREs Computational Statistics, 4(2), 218-223. https://doi.org/10.1002/wics.203

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. Journal of the American Statistical Association, 100(469), 332–346. http://www.jstor.org/stable/27590542

Huisman, W., Kleemans, E.R. The challenges of fighting sex trafficking in the legalized prostitution market of the Netherlands. Crime Law Soc Change 61, 215–228 (2014). https://doi.org/10.1007/s10611-013-9512-4

International Working Group for Disease Monitoring and Forecasting. (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. American Journal of Epidemiology, 142(10), 1047-1058.

Lyneham, S., Dowling, C., & Bricknell, S. (2019). Estimating the dark figure of human trafficking and slavery victimisation in Australia. Statistical Bulletin No. 16 Canberra: Australian Institute of Criminology. https://www.aic.gov.au/publications/sb/sb16

Silverman, B. (2013). Modern slavery: An application of multiple systems estimation. Home Office.

Staring, R. H. J. M. (2012). Human trafficking in the Netherlands: Trends and recent developments. International Review of Law, Computers & Technology, 26(1), 59-72. https://doi.org/10.1080/13600869.2012.646797

Van Dijk, J. G., Cruyff, M., & Van Der Heijden, P. (2021). Multiple Systems Estimation Slachtoffers Mensenhandel Nederland 2016-2019. https://repository.wodc.nl/handle/20.500.12832/3123?show=full

Van Loan, C. F. (2000). The ubiquitous Kronecker product. Journal of Computational and Applied Mathematics, 123(1–2), 85-100. ISSN 0377-0427. https://doi.org/10.1016/S0377-0427(00)00393-9.

Xie, C., Zhong ,W. & Mueller, K. (2017). A Visual Analytics Approach for Categorical Joint Distribution Reconstruction from Marginal Projections. IEEE Transactions on Visualization and Computer Graphics, 23(1), 51-60. doi: 10.1109/TVCG.2016.2598479

APPENDIX

Figure 1. *Gender distribution 2018*

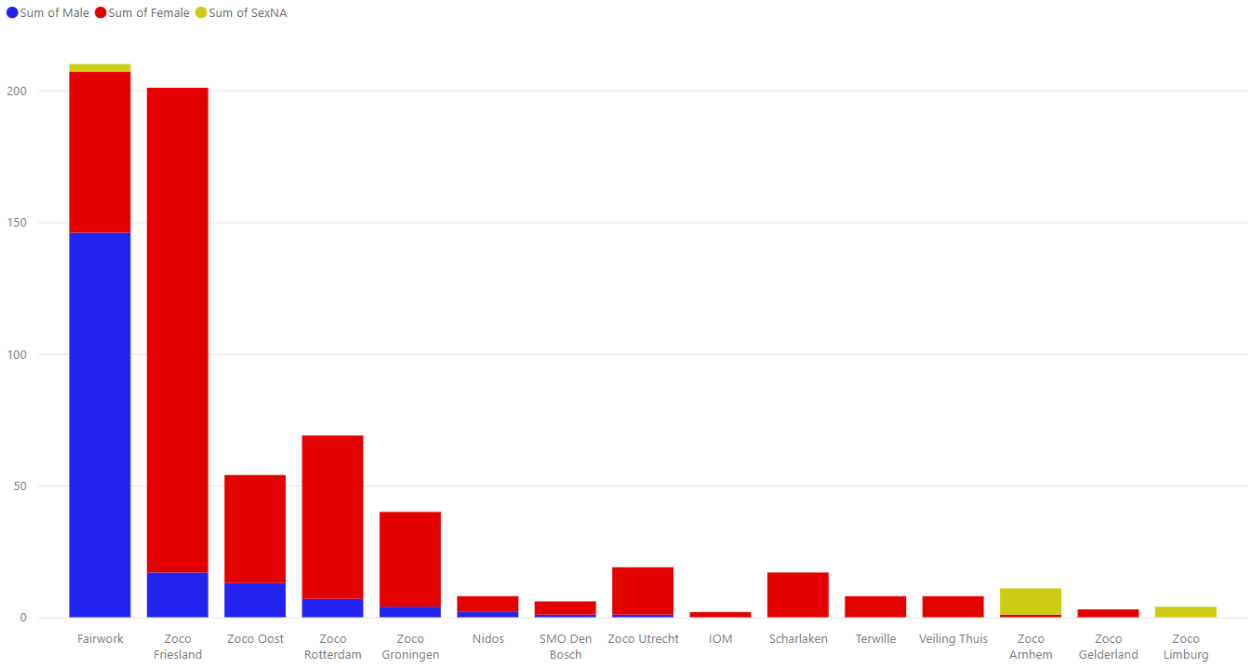

Figure 2. *Age distribution 2018*

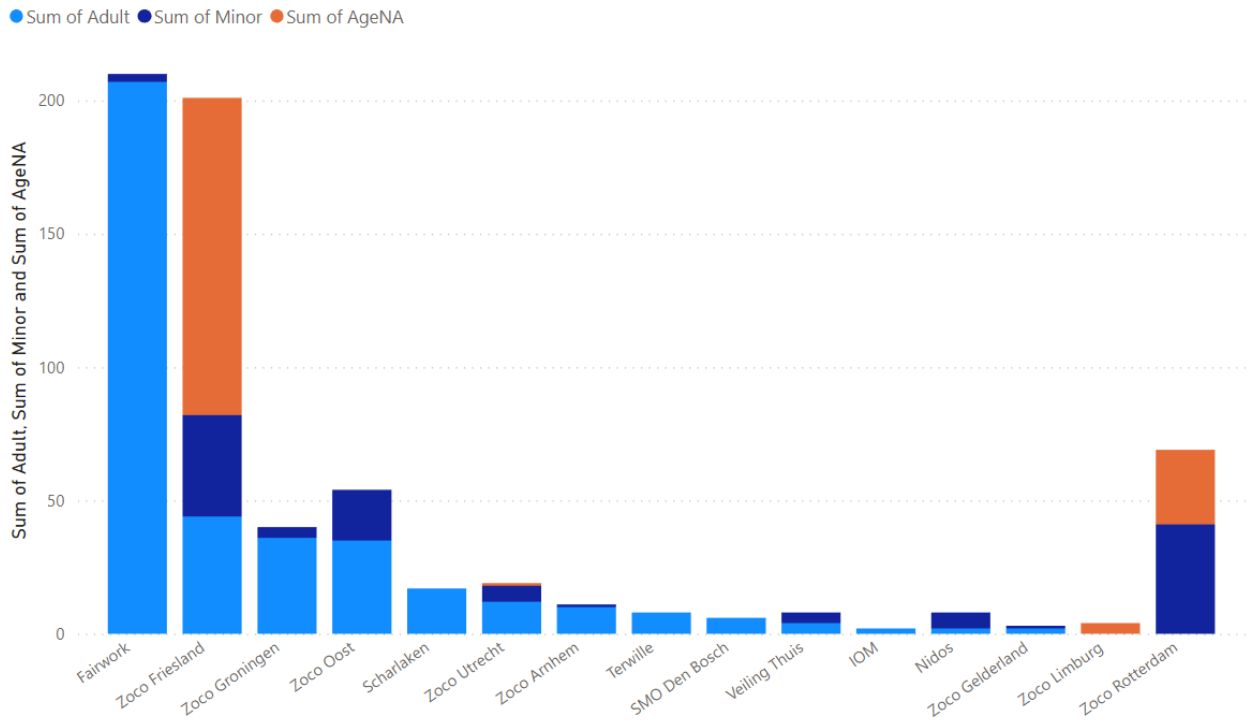

Figure 3. *Nationality distribution 2018*

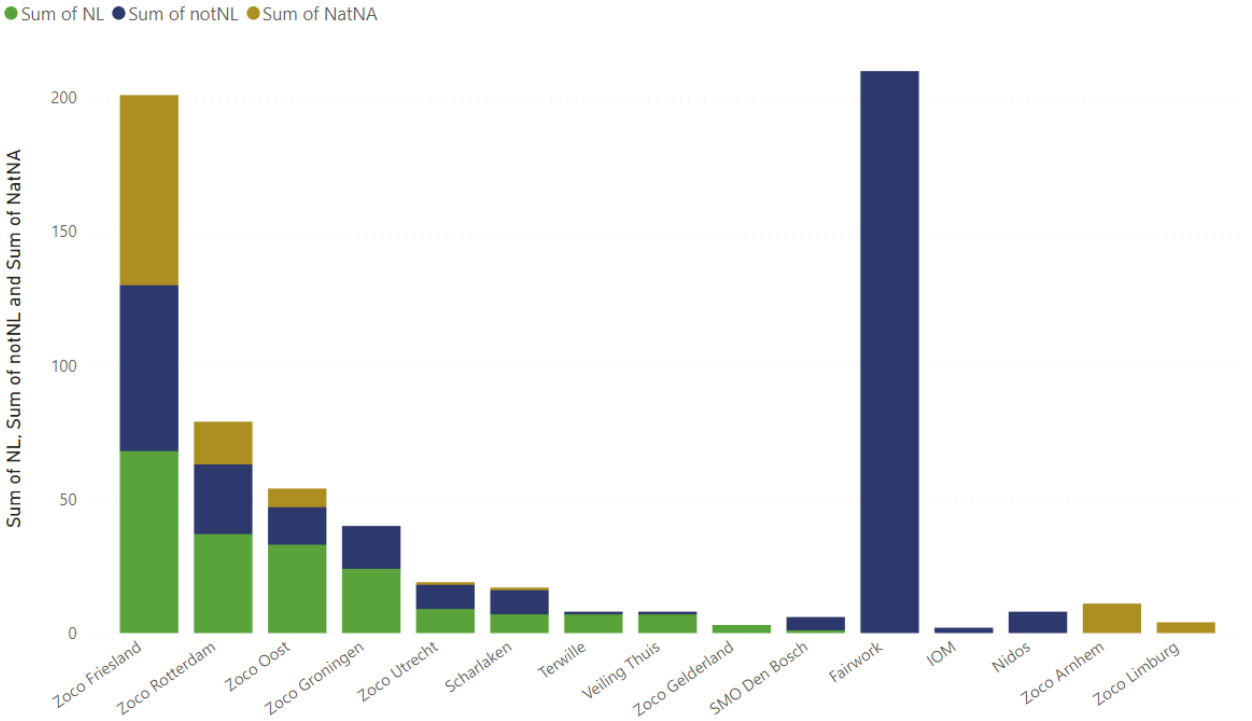

Figure 4. *Exploitation distribution 2018*

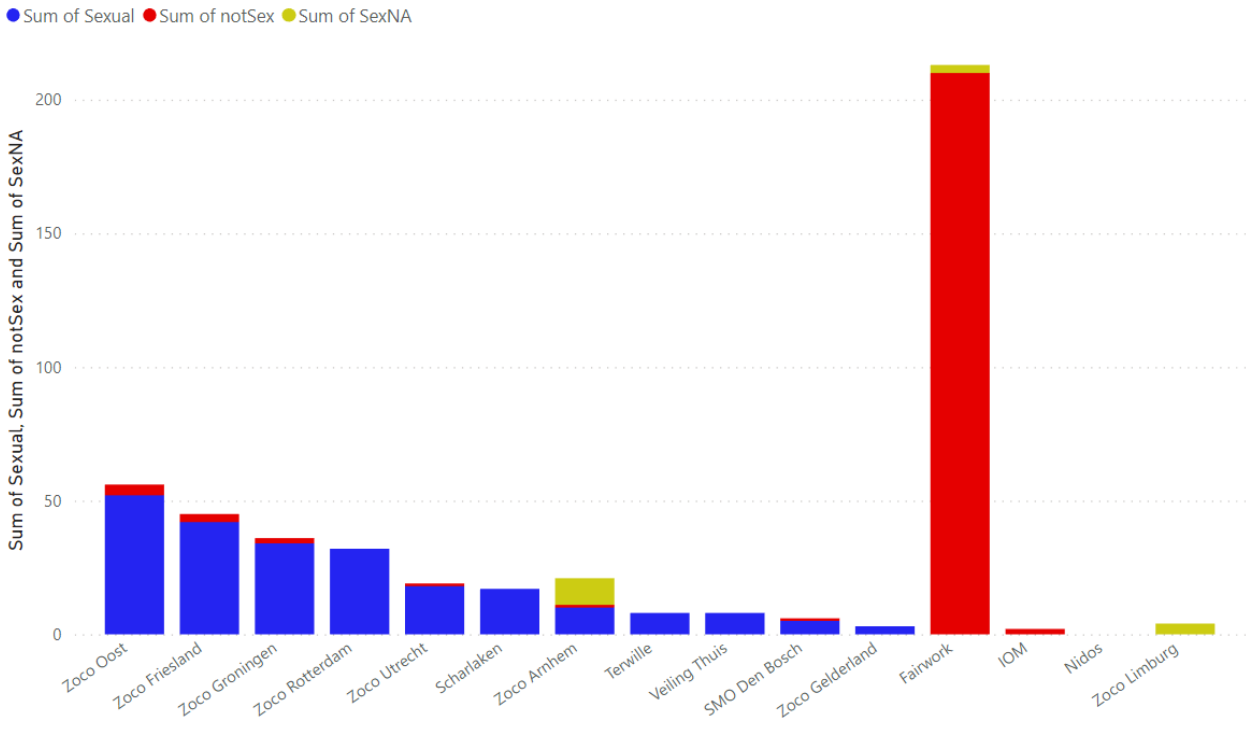

Figure 5. *Gender distribution 2019*

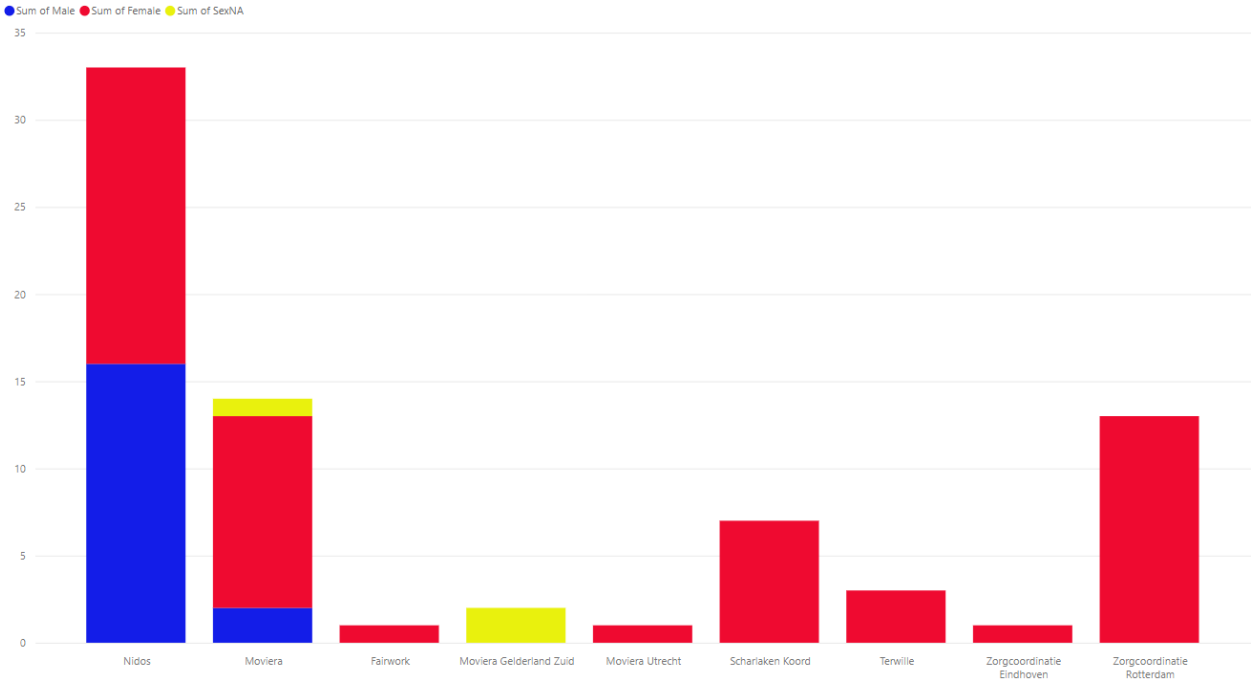

Figure 6. *Age distribution 2019*

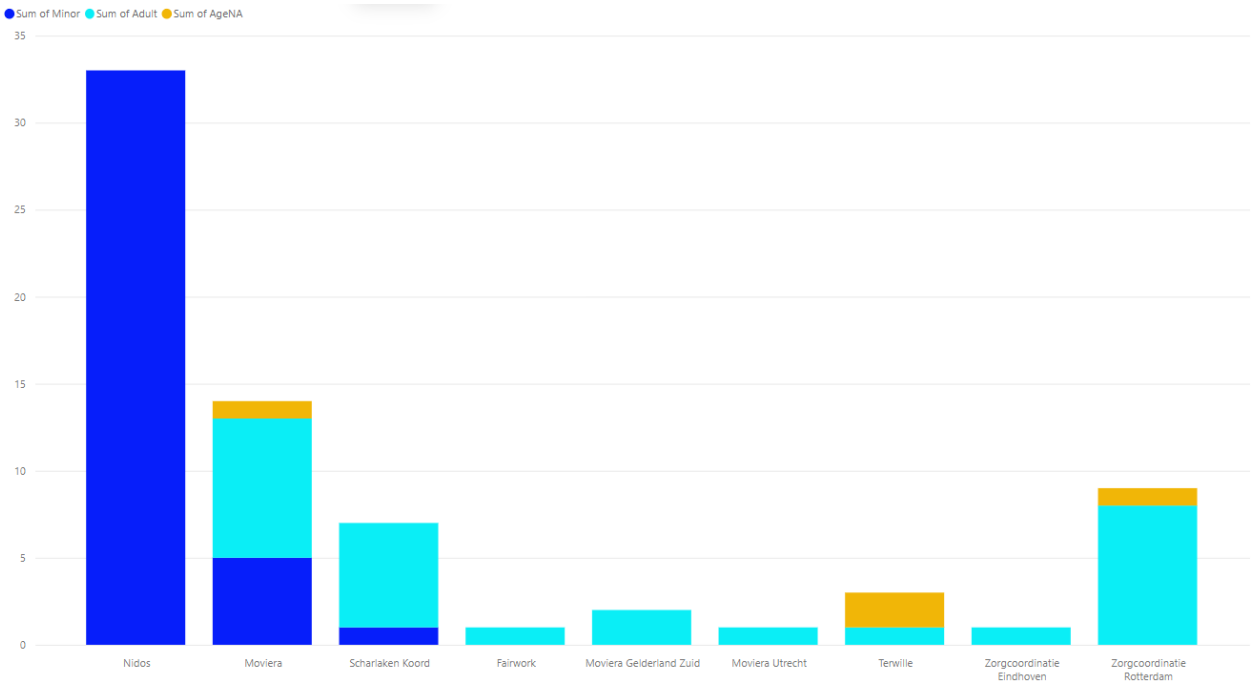

Figure 7. *Nationality distribution 2019*

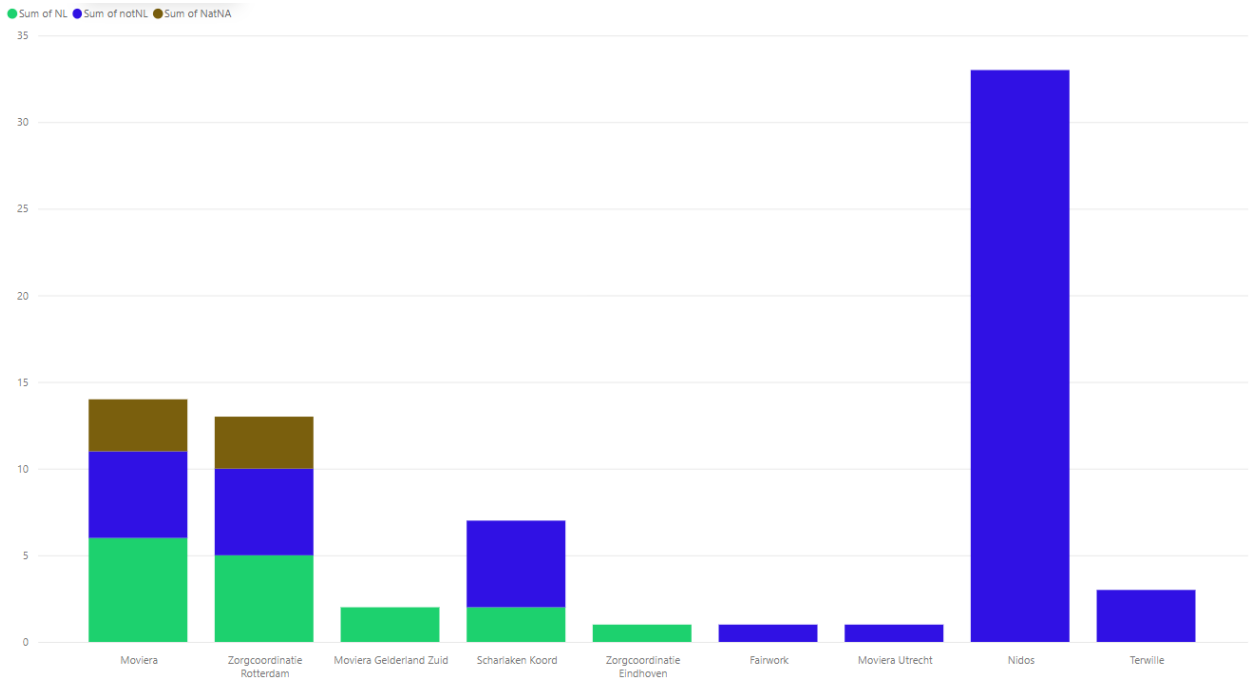

Figure 8. *Exploitation distribution 2019*

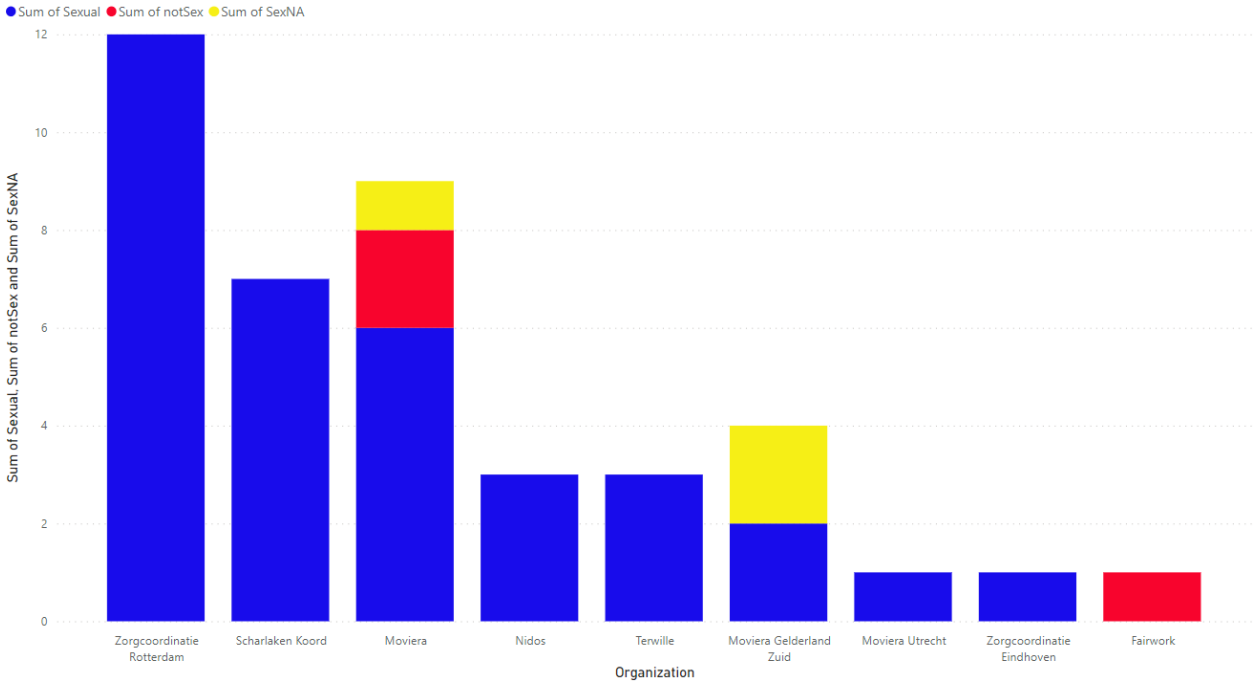

Table 1. *R code for marginal to joint frequency conversion*

```r
#Make the marginal to joint function
marg2joint <- function(covs, freqs, register) {

 cases <- sum(freqs[[1]])     # the number of cases

 Freq <- freqs[[1]]

 for(i in 2:length(freqs)) Freq <- kronecker(Freq, freqs[[i]])

  d <- expand.grid(rev(covs) )

  out <- data.frame(1, rev(d), Freq = cases * Freq / sum(Freq)) # Freq is
  the expected frequency

  names(out)[1] <- register

  out
 }

#Add up the frequencies of the organizations for each register (R, O, Z)
R <- colSums(marg[c(2, 3, 4, 6, 7, 9, 13, 15), ], na.rm = T)
O <- colSums(marg[c(5, 10, 11), ], na.rm = T)
Z <- colSums(marg[c(8, 12, 14), ], na.rm = T)

```

```r
#Now the lists
covs_R <- list(S = names(R)[1:3], A = names(R)[4:6], N = names(R)[7:9],
E = names(R)[10:12])
covs_O <- list(S = names(O)[1:3], A = names(O)[4:6], N = names(O)[7:9],
E = names(O)[10:12])
covs_Z <- list(S = names(Z)[1:3], A = names(Z)[4:6], N = names(Z)[7:9],
E = names(Z)[10:12])


freqs_R <- list(S = R[1:3], A = R[4:6], N = R[7:9], E = R[10:12])
freqs_O <- list(S = O[1:3], A = O[4:6], N = O[7:9], E = O[10:12])
freqs_Z <- list(S = Z[1:3], A = Z[4:6], N = Z[7:9], E = Z[10:12])

#Run the function for organization R
Rjoint <- marg2joint(covs = covs_R, freqs = freqs_R, register = "R")

head(Rjoint)

Rjoint <- mutate(Rjoint,
                 S = case_match(S, "SexNA" ~  NA, .default = S),
                 A = case_match(A, "AgeNA" ~  NA, .default = A),
                 N = case_match(N, "NatNA" ~  NA, .default = N),
                 E = case_match(E, "ExpNA" ~  NA, .default = E)) %>%
 droplevels()

summary(Rjoint)

#Run the function for organization O
Ojoint <- marg2joint(covs = covs_O, freqs = freqs_O, register = "O")

head(Ojoint)

Ojoint <- mutate(Ojoint,
                S = case_match(S, "SexNA" ~  NA, .default = S),
                 A = case_match(A, "AgeNA" ~  NA, .default = A),
                 N = case_match(N, "NatNA" ~  NA, .default = N),
                 E = case_match(E, "ExpNA" ~  NA, .default = E)) %>%
  droplevels()

summary(Ojoint)

#Run the function for organization Z
Zjoint <- marg2joint(covs = covs_Z, freqs = freqs_Z, register = "Z")

head(Zjoint)

Zjoint <- mutate(Zjoint,
                 S = case_match(S, "SexNA" ~  NA, .default = S),
                 A = case_match(A, "AgeNA" ~  NA, .default = A),
                 N = case_match(N, "NatNA" ~  NA, .default = N),
                 E = case_match(E, "ExpNA" ~  NA, .default = E)) %>%
  droplevels()

```

```
77 summary(Zjoint)
```