Unraveling Variable Relationships in the CoMensha Human Trafficking Dataset through King's EI Model

Applied Data Science Master Thesis

Chih-Yi Hsia
9569243

First Supervisor
Dr. Maarten Cruyff

Second Supervisor
Dr. Kyle Lang

Department of Information and Computing Science
Utrecht University
June 30, 2023

Abstract

Examining the relationships between variables provides extensive information about the data and is beneficial in a variety of applications. However, obtaining the joint distribution is not always feasible considering issues involving confidentiality problems, particularly for sensitive data. The absence of individual-level information has been a challenge in analyzing CoMensha human trafficking dataset in recent years. Providers only supply aggregate data in accordance with GDPR legislation and privacy considerations. The objective of this study is to use King's EI model to transform the marginal distribution into a joint distribution to investigate the correlation between variables in the CoMensha dataset. The four variables sex, age, nationality, and exploitation are matched in six different combinations to form two-by-two contingency tables. The study finds noticeable trends across several groups. It is shown that victims were mainly adults and came from foreign nations, both male and female. Men were exploited in ways other than sexual, but women were predominantly sexually exploited. It also reveals that in different age groups, a greater number of adults were sexually exploited, while a bigger proportion of minors were exploited in other ways. In terms of nationality, it is found that female and sexual exploitation accounted for a higher percentage for almost all of the years in the Dutch or Non-Dutch groups. When analyzing the type of exploitation, females, and adults represents a larger percentage in the sexually exploited group, while men and minors account for a higher proportion of the other types of exploitation. Victims from foreign countries take up the majority of both categories. Validation of the results is difficult to obtain, but it offers potential information for future study, which can utilize multiple models to accomplish the estimation to further examine and compare the results.

Keywords: ecological inference, human trafficking, joint distribution, marginal distribution, contingency table

Introduction

Relationship reasoning between variables is a critical component of any research to be able to obtain essential insights. However, acquiring the joint distribution of variables may be problematic in many circumstances. Data information is only accessible in marginal distributions for numerous settings, including customer surveys, elections, and public datasets from government resources (Dobra et al., 2006). With marginal data, the distribution can only be examined at the aggregate level, with no information at the individual level. For example, with marginal data in a customer survey, a company is limited to targeting women, however, with joint data, the company could concentrate on women between the ages of 26 and 30 for their products. The difficulty emerges as a result of the data collection procedure and data confidentiality. The data in elections is gathered independently under anonymous procedures, and the total ballots voted for each candidate and the persons who vote for them are not connected. To protect consumers' privacy, results of consumer surveys are often provided in marginal frequencies (Chaubey et al., 2003). The CoMensha organization's human trafficking dataset has likewise altered the information it provides in the interest of confidentiality since 2018. In response to the General Data Protection Regulation (GDPR), numerous suppliers have ended providing individual data in instead opting for aggregated data with marginal frequencies, making it more difficult to gain a deeper understanding of the correlations between variables (Van Dijk et al., 2021). Data confidentiality is a means of protecting sensitive information, but it also imposes disclosure constraints (Dobra et al., 2002).

The need to derive an inference of joint distribution has long been a subject of research (Dobra et al., 2006). The technique of getting inferences on joint distributions using marginal distributions is one of the methods to approach the associations among variables and has been studied for years. The approach is known scientifically as ecological inference, which is the act of obtaining details at the individual level from aggregate level data. Numerous models with various assumptions have been developed to address the issue regarding categorical data (Frogner & Poggio, 2019). The simplest approach to this problem is to use an independent model, which posits that two random variables are independent of one another and derives the distributions from fundamental probability theory. Goodman presented the first statistical solution to the problem and formalized it in a simple regression model (Goodman, 1953). The

method investigates variance in the marginal distribution and tries to apply the logic back to the joint distribution. King proposed another EI method for extracting both deterministic and statistical information through the integration of the method of bounds and Goodman's regression (King, 1997). King's EI model was later developed into a hierarchical model based on Bayesian and the Markov Chain Monte Carlo (MCMC) methods (King et al., 1999). The extended model is capable of being used for more complicated problems and is regarded as a good tool for inference with limited information (Karwa & Slavkovic, 2013). Based on King's model, models that replicate the findings using different distributions or sampling strategies are constantly being developed (King et al., 2004).

The goal of this study is to infer the joint distribution of variables with marginal distribution to gain additional details about the data. The study utilizes the use of the CoMensha human trafficking dataset, which has accessible marginal distributions. Marginal and joint distribution is the fundamental concepts in probability. Given a dataset with two variables X and Y, the marginal distribution of X represents the probability distribution of X when Y is ignored and vice versa. The joint distribution, which is referred to as the distribution of X conditional on Y or the distribution of Y on X, considers both variables. The CoMensha dataset contains marginal information for variables, however, the cells of values in the contingency tables are unknown. King's EI model is applied to the dataset since it is more accurate than Goodman's regression and less complicated than other extended models (Rosen et al., 2001). The association between variables may be established by transforming marginal distribution into joint distributions for the dataset employing King's EI model. The additional knowledge about the dataset may be used to make more precise decisions and allocate resources more efficiently.

In the following sections, the study provides the data and methods, results and analysis, conclusion, and discussion. The data and methods section specifies the dataset and the statistical analysis method utilized in the study. The results and analysis provide the estimated percentages for cells in two-by-two tables for each year. The conclusion and discussion section summarizes the findings while also addressing the interpretations, implications, and limitations. The section also includes recommendations for further study and ethical issues.

Data and Methods

The Data

The dataset utilized in this study was contributed by public and private institutions and was managed by an NGO called CoMensha. The dataset includes six registers and four variables from 2016 to 2019. The dataset's details are listed in Tables 1 and 2. Table 3 displays the frequency of each list from 2016 to 2019. The dataset was integrated into the formats provided in Table 4 for analytical preparation. The marginal frequency of each variable was converted into a percentage for each register or list in each year. In List I, for example, there were 68.4 percent males and 31.6 percent females by sex, as well as 94.7 percent adults and 5.3 percent minors by age. Other combinations of every two variables were handled in the same way. The table was stored in R as a data frame, and the model was fitted to get the estimates.

Table 1 Registers in the CoMensha dataset

Code	Registers
I	ISZW: Ministry of Social Affairs and Employment
K	KMar: Border police
O	Opvang: Shelters
P	Politie: Police
R	Regiocoordinator: Regional coordinators
Z	Other

Table 2 Variables in the CoMensha dataset

Code	Variables
S	Sex with levels of Male and Female
L	Age with levels of Adult and Minor
N	Nationality with levels of NL and Other
U	Exploitation with levels of Sexual and Other
J	Year with levels of 2016, 2017, 2018, 2019

Table 3 The frequency of each list

List	2016	2017	2018	2019
I	38	38	75	46
K	109	139	86	48
O	226	206	253	219
P	462	432	530	953
R	240	320	492	266
Z	90	60	61	79

Table 4 Percentage of sex and age distribution in 2016

List	Male	Female	Adult	Minor	N
I	0.684	0.316	0.947	0.053	38
K	0.018	0.982	1.000	0.000	109
O	0.491	0.509	0.686	0.314	220
P	0.113	0.887	0.826	0.174	461
R	0.093	0.907	0.769	0.231	225
Z	0.364	0.636	0.409	0.591	88

The Model

The CoMensha dataset, which comprised six lists from 2016 to 2019, was utilized in this study to investigate four variables: sex, age, nationality, and exploitation. Each list's data includes the total population (N_i), the percentage of females and males by sex, the percentage of adults and minors by age, the percentage of Dutch and other nationalities by nationality, and the percentage of sexual exploitation and other sorts of exploitation by exploitation. Table 5 depicts the simplified notation used to investigate the joint distribution of sex and age in list i. The same interpretation is applied to the combination of additional variables and lists. The parameters of interest are β_i^m and β_i^f, with the observed marginals X_i, Y_i, and N_i. Instead of receiving estimates on a single list, the results are calculated on an average over all lists. The model is established on three assumptions. First, when conditioning on X_i, β_i^m and β_i^f are determined by a truncated bivariate normal distribution. Second, β_i^m and β_i^f are mean independent of X_i. Third, when conditioning on X_i, Y_i in various lists are independent (King, 2013). The model is expressed as follows:

$$Y_i = \beta_i^m X_i + \beta_i^f (1 - X_i) \tag{1}$$

where:
X_i: Proportion of males observed in list i
Y_i: Proportion of adults observed in list i
N_i: Number of victims with the complete observation of sex and age in list i
β_i^m: Proportion of male victims who are adults in list i
β_i^f: Proportion of female victims who are adults in list i
i: 1,2,3,4,5,6

Table 5 Notation for list i of sex and age

	Adult	Minor	Total
Male	β_i^m	$1 - \beta_i^m$	X_i
Female	β_i^f	$1 - \beta_i^f$	$1 - X_i$
Total	Y_i	$1 - Y_i$	N_i

The Likelihood Estimation

The probability density of (β_i^m, β_i^f) is given below. TN means truncated normal distribution, $\mathfrak{B}$ represents the mean vector of (β_i^m, β_i^f), and $\sum$ refers to the variance matrix of (β_i^m, β_i^f).

$$P(\beta_i^m, \beta_i^f) = TN(\beta_i^m, \beta_i^f | \mathfrak{B}, \sum) \tag{2}$$

The truncated bivariate normal distribution with male and female means, standard deviations, and a correction:

$$\psi = \{\mathfrak{B}^m, \mathfrak{B}^f, \sigma_m, \sigma_f, \rho\} = \{\mathfrak{B}, \sum\} \tag{3}$$

The following is a definition of the likelihood function. $N(Y_i|\mu_{i,}\, \sigma_i^2)$ denotes an untruncated normal distribution with mean $\mu_{i,}$ and variance σ_i^2. $R(\breve{\mathfrak{B}}, \breve{\sum})$ represents the normalizing constant for the truncated bivariate normal distribution of $\beta_i^m \; and \; \beta_i^f$. $S(\breve{\mathfrak{B}}, \breve{\sum})$ indicates the normalizing constant from the truncated normal posterior distribution of β_i^m given Y_i and $\breve{\psi}_i$. $\breve{\psi}$ is the corresponding untruncated distribution for ψ to simplify the mathematical computations (King, 1997).

$$L(\breve{\psi} \,|\, Y) \propto \prod_{X_i \in (0,1)} P(Y_i|\breve{\psi})$$

$$= \prod_{X_i \in (0,1)} N(Y_i|\mu_{i,}\, \sigma_i^2) \frac{S(\breve{\mathfrak{B}}, \breve{\sum})}{R(\breve{\mathfrak{B}}, \breve{\sum})} \tag{4}$$

Quantities of Interest

Using a parameterization close to the normal distribution provides a simpler technique to optimize the likelihood function. ϕ a deterministic function of $\breve{\psi}$, is another transformation of $\breve{\psi}$ for convenient estimation.

$$\breve{\psi} = \{\breve{\mathfrak{B}}^m, \breve{\mathfrak{B}}^f, \breve{\sigma}_m^2, \breve{\sigma}_f^2, \breve{\rho}\} \tag{5}$$

$$\phi = \phi_1, \phi_2, \phi_3, \phi_4, \phi_5 \tag{6}$$

$$\phi_1 = \frac{\breve{\mathfrak{B}}^m - 0.5}{\breve{\sigma}_m^2 + 0.25} \tag{7}$$

$$\phi_2 = \frac{\breve{\mathfrak{B}}^f - 0.5}{\breve{\sigma}_f^2 + 0.25} \tag{8}$$

$$\phi_3 = ln(\breve{\sigma}_m) \tag{9}$$

$$\phi_4 = ln(\breve{\sigma}_f) \tag{10}$$

$$\phi_5 = 0.5ln(\frac{1+\breve{\rho}}{1-\breve{\rho}}) \tag{11}$$

Then, using the following processes, a single β_i^m can be extracted from $P(\beta_i^m|Y)$:
First, select one ϕ value from its posterior $P(\phi|T)$ and define it as $\tilde{\phi}$. Then yielding $\tilde{\breve{\psi}}$ by reparametrizing $\tilde{\phi}$ into the untruncated scale. Lastly, add $\tilde{\breve{\psi}}$ to the conditional posterior distribution of β_i^m, known as $P(\beta_i^m|Y, \tilde{\breve{\psi}})$, and randomly select a value of β_i^m from it (King, 1997).

$$P(\beta_i^m|Y) = \int_{-\infty}^{\infty} P(\phi|Y)P(\beta_i^m|Y, \breve{\psi})d\breve{\psi} \tag{12}$$

ei and eiCompare in R

To tackle ecological inference challenges, King and Roberts developed the ei() procedure in R. Using the input data, ei() calculates and optimizes a likelihood distribution, then makes predictions on the quantities of interest based on the potential values or limits within the entire likelihood distribution (King and Roberts, 2012). The eiCompare package generalizes the ei() process created by King and Roberts into a function with several types of table and graph tools (Collingwood et al., 2016). To execute the function, two vectors Y and X containing the known marginal frequencies, and another vector N providing the number of persons of interest are required. The primary function utilized for this study was ei_est_gen(), which is the modified version of the ei() function. The new function optimizes the process of estimating ecological inference outcomes for each independent and dependent variable separately.

Results and Analysis

The CoMensha dataset was used to compute average mean estimates for the values of cells in two-by-two contingency tables across six lists from 2016 to 2019. There are six potential contingency table combinations with four variables: sex, age, nationality, and exploitation, and by exchanging the dependent and independent variables in the model, the combinations can be explored in two ways. This section presents the average mean estimates based on King's EI model. In the appendix, the results of Goodman's regression are provided for comparisons between the two models. The outcomes of the two models can be interpreted similarly.

Age and Sex

The average mean estimates of the distribution of adults and minors for males and females each year are shown in Table 6. According to the model, the average ratio of males who were adults across lists was 51.575 percent in 2016, while the average ratio of males who were minors was 43.941 percent. In 2016, the average proportion of females who were adults throughout the lists was 83.671 percent, while the average percentage of females who were minors was 15.451. Other years' estimates could be explored in the same way as described above. Table 6 reveals that the percentage of adults was higher than minors in both male and female categories every year, implying that the majority of victims were adults regardless of gender. The proportion of adults in the male group was lower than that in the female group, but this trend reversed after 2017. Table 7 shows the average mean estimates of male and female distributions for adults and minors. Males accounted for 17.476 percent of the adult group, while females formed up 82,438 percent. Females have represented a greater proportion of the adult group than males across the years while males exceeded females in the minor group in 2018.

Table 6 EI mean estimates of adult and minor distribution by sex (percentage)

	2016		2017		2018		2019	
	Male	Female	Male	Female	Male	Female	Male	Female
Adult	51.575	83.671	99.151	74.745	98.476	88.300	99.460	85.003
SE	0.005	0.125	0.650	0.061	1.559	0.933	0.331	1.794
Minor	43.941	15.451	1.316	25.154	1.311	11.510	0.590	16.329
SE	0.813	0.056	1.766	0.154	1.893	1.218	0.435	0.343
Total	95.516	99.122	100.467	99.899	99.787	99.811	100.051	101.332

Table 7 EI mean estimates of male and female distribution by age (percentage)

	2016		2017		2018		2019	
	Adult	Minor	Adult	Minor	Adult	Minor	Adult	Minor
Male	17.476	40.461	15.826	83.980	26.726	80.717	49.864	20.005
SE	0.523	1.149	0.173	0.470	1.627	16.330	0.594	11.060
Female	82.438	59.406	83.980	83.980	72.439	20.304	50.509	78.561
SE	0.324	1.607	0.470	0.470	1.871	15.324	1.004	13.570
Total	99.914	99.866	99.807	167.960	99.165	101.022	100.373	98.566

Nationality and Sex

Table 8 displays the average mean estimates of the distribution of victims from the Netherlands or other countries by gender for each year. In 2016, the average ratio of males from the Netherlands across lists was 7.075 percent, whereas the average ratio of males from other countries was 93.125 percent. In terms of females, the average proportion of females from the Netherlands was 36.678 percent across all lists, and the average percentage of females from other countries was 63.323 percent in 2016. Estimates for other years could be investigated in the same approach as mentioned above. It was discovered that victims, whether in male or female groups, were mostly from other nations. What can be clearly seen was that the percentage of victims from the Netherlands in the female category was considerably larger than the percentage in the male group. The average mean estimates of male and female distributions for the Dutch nationality and other nationalities are shown in Table 9. Males constituted 11.022 percent of the group with Dutch nationality, while females carried out 88.931 percent. Females surpassed males in both groups of victims with Dutch nationality and victims from other countries from 2016 to 2018, with the trend inverted for victims from other countries in 2019.

Table 8 EI mean estimates of Dutch and other nationalities distribution by sex (percentage)

	2016		2017		2018		2019	
	Male	Female	Male	Female	Male	Female	Male	Female
NL	7.075	36.678	7.318	40.750	3.460	33.570	0.677	31.857
SE	0.728	0.248	1.417	0.187	0.760	1.400	0.023	0.021
Other	93.125	63.323	92.753	59.206	96.510	66.030	99.407	68.082
SE	0.908	0.342	0.507	0.222	5.280	0.640	0.031	0.024
Total	100.199	100.001	100.071	99.956	99.970	99.600	100.083	99.938

Table 9 EI mean estimates of male and female distribution by nationality (percentage)

	2016		2017		2018		2019	
	NL	Other	NL	Other	NL	Other	NL	Other
Male	11.022	27.737	4.700	32.180	31.318	33.378	11.446	53.943
SE	6.076	2.285	4.470	2.860	6.712	2.569	7.151	1.654
Female	88.931	71.216	96.600	67.160	70.296	66.593	86.076	45.950
SE	6.115	1.839	4.460	2.190	6.098	1.715	9.244	1.704
Total	99.953	98.953	101.300	99.340	101.614	99.972	97.522	99.893

Exploitation and Sex

Table 10 provides the average mean estimates of sexual exploitation and other forms of exploitation for males and females each year. In 2016, sexual exploitation in the male category was estimated to be 1.561 percent, while other types of exploitation were 99.385 percent. 79.671 percent of female victims were victims of sexual exploitation in 2016, while 20.502 percent were victims of other types of exploitation. Estimates for other years can be interpreted in the same way suggested above. What stands out in Table 10 is the proportion of different forms of exploitation in the male group varied from the female group. Male victims tended to be exploited in ways other than sexual exploitation, whereas female victims were more likely to be sexually exploited. The average mean estimates of male and female distributions for various forms of exploitation are shown in Table 11. Male victims accounted for 1.630 percent of sexual exploitation victims, while female victims accounted for 98.170 percent. When concentrating on different forms of exploitation, it was clear that the victims of sexual exploitation were mostly female, whereas males composed a larger share of the victims of other types of exploitation.

Table 10 EI mean estimates of sexual exploitation and other types of exploitation distribution by sex (percentage)

	2016		2017		2018		2019	
	Male	Female	Male	Female	Male	Female	Male	Female
Sexual	1.561	79.671	8.088	82.150	25.114	92.945	44.124	84.222
SE	2.006	0.047	0.968	0.296	0.085	0.030	2.809	1.978
Other	99.385	20.502	91.970	17.753	74.950	7.026	56.075	15.413
SE	0.549	0.265	1.123	0.255	0.027	0.012	2.074	1.610
Total	100.946	100.173	100.058	99.902	100.064	99.970	100.199	99.635

Table 11 EI mean estimates of male and female distribution by exploitation (percentage)

	2016		2017		2018		2019	
	Sexual	Other	Sexual	Other	Sexual	Other	Sexual	Other
Male	1.630	53.860	4.434	57.263	10.392	85.406	30.768	76.231
SE	1.110	1.630	1.395	2.852	0.536	4.382	1.317	3.152
Female	98.170	47.600	95.574	42.642	89.824	13.398	69.237	23.996
SE	1.170	2.490	1.130	2.575	0.319	2.274	2.613	2.882
Total	99.800	101.460	100.008	99.905	100.215	98.804	100.005	100.227

Age and Nationality

Table 12 shows the annual average mean estimates of the distribution of adults and minors by nationality. According to the model, there were 64.719 percent of adults and 35.442 percent of minors among the victims from the Netherlands in 2018. The majority of victims from other countries were adults, accounting for 99.939 percent of all victims from foreign countries in 2018. A similar process could be used for analyzing estimates for different years. The average mean estimations of different nationalities for adults and minors are displayed in Table 13. In terms of adult victims, 82.806 percent were from foreign nations, while 17.515 percent were from the Netherlands in 2018. In 2018, 97.777 percent of minor victims were from the Netherlands, while 3.170 percent were from other countries. The description above uses the year 2018 rather than another year since the sums of the other years were not close to 100. Further discussion will be provided in the next section.

Table 12 EI mean estimates of adult and minor distribution by nationality (percentage)

	2016		2017		2018		2019	
	NL	Other	NL	Other	NL	Other	NL	Other
Adult	98.850	70.945	91.541	73.672	64.719	99.939	58.170	94.884
SE	1.142	1.172	8.961	2.522	0.496	0.005	1.193	0.487
Minor	98.850	28.081	91.541	22.670	35.442	0.086	58.170	5.096
SE	1.142	3.105	8.961	5.275	0.145	0.020	1.193	0.458
Total	197.701	99.026	183.082	96.342	100.161	100.025	116.339	99.980

Table 13 EI mean estimates of Dutch and other nationalities distribution by age (percentage)

	2016		2017		2018		2019	
	Adult	Minor	Adult	Minor	Adult	Minor	Adult	Minor
NL	39.711	13.171	32.590	35.743	17.515	97.777	17.397	20.772
SE	0.675	3.121	0.765	3.309	0.659	3.759	0.498	1.671
Other	60.224	86.717	32.590	63.963	82.806	3.710	82.833	79.119
SE	0.743	3.541	0.765	2.298	0.228	5.712	0.600	0.607
Total	99.935	99.889	65.180	99.706	100.321	101.488	100.230	99.891

Age and Exploitation

Table 14 displays the average mean estimates of the distribution of adults and minors for sexual exploitation and other types of exploitation each year. As indicated by the model, 98.205 percent of victims of sexual exploitation were adults, while 1.791 percent were minors in 2016. In terms of other types of exploitation, 37.505 percent were adults, and 62.515 were minors in 2016. Estimates for other years could be researched in a similar way stated above. Victims of sexual exploitation were mostly adults throughout the years, whereas minors accounted for greater percentages in other forms of exploitation except for the year 2019. Table 15 presents the mean average estimates of each type of exploitation for adults and minors. When only adult victims were considered, 76.771 percent were victims of sexual exploitation and 23.183 were victims of other exploitation in 2016. Adult victims were primarily exploited sexually, while youngsters were mostly exploited in other ways.

Table 14 EI mean estimates of adult and minor distribution by exploitation (percentage)

	2016		2017		2018		2019	
	Sexual	Other	Sexual	Other	Sexual	Other	Sexual	Other
Adult	98.205	37.505	97.208	48.473	97.208	48.473	86.096	96.773
SE	0.008	0.116	3.671	4.095	3.671	4.095	0.031	0.073
Minor	1.791	62.515	0.746	54.151	0.746	54.151	13.921	3.197
SE	0.016	0.009	0.500	0.464	0.500	0.464	0.036	0.054
Total	99.995	100.020	97.953	102.624	97.953	102.624	100.017	99.970

Table 15 EI mean estimates of sexual exploitation and other types of exploitation distribution by age (percentage)

	2016		2017		2018		2019	
	Adult	Minor	Adult	Minor	Adult	Minor	Adult	Minor
Sexual	76.771	8.152	77.997	21.186	78.374	4.084	70.253	29.778
SE	0.030	0.007	1.309	5.709	0.053	7.396	0.250	0.319
Other	23.183	91.692	21.612	80.088	21.947	98.608	29.778	29.778
SE	0.015	0.083	0.462	3.248	0.646	1.600	0.319	0.319
Total	99.955	99.843	99.609	101.274	100.321	102.692	100.032	59.557

Nationality and Exploitation

In Table 16, the average mean estimates of the distribution of various nationalities for the two types of exploitation are displayed annually. The model estimated that the average percentage of sexual exploitation victims who were Dutch was 44.375 percent, while the average percentage of victims who were from other countries was 55.642 in 2016. Other forms of exploitation victims included 4.778 from the Netherlands and 95.224 from other countries in 2016. Other years' estimates can be examined in the same way as mentioned above. It was revealed that, whether in sexual exploitation or other exploitation groups, victims of foreign nationalities contributed a larger percentage than victims of Dutch nationality. The average mean estimates of sexual exploitation and other types of exploitation for Dutch and other nationalities are shown in Table 17. Sexual exploitation accounted for 59.040 percent of victims in the Netherlands, while other kinds of exploitation accounted for 41.020 percent. Except for 2018, a larger proportion of Dutch victims were victims of sexual exploitation. Victims from foreign nations were mostly sexually exploited similarly.

Table 16 EI mean estimates of Dutch and other nationalities distribution by exploitation (percentage)

	2016		2017		2018		2019	
	Sexual	Other	Sexual	Other	Sexual	Other	Sexual	Other
NL	44.375	4.778	46.054	7.243	31.974	1.638	25.892	1.514
SE	0.218	0.277	0.909	1.222	0.269	0.470	0.375	2.039
Other	55.642	95.224	53.700	92.472	67.987	98.257	74.512	98.586
SE	0.168	0.311	0.867	1.617	0.337	0.568	0.682	1.395
Total	100.018	100.002	99.753	99.715	99.961	99.895	100.403	100.100

Table 17 EI mean estimates of sexual exploitation and other types of exploitation distribution by nationality (percentage)

	2016		2017		2018		2019	
	NL	Other	NL	Other	NL	Other	NL	Other
Sexual	59.040	67.046	57.085	68.928	36.008	63.025	53.375	69.344
SE	0.046	0.139	5.789	1.523	0.014	0.005	10.948	1.831
Other	41.020	32.906	44.019	30.618	63.994	36.977	44.275	31.507
SE	0.004	0.008	3.653	0.525	0.054	0.004	12.799	2.619
Total	100.060	99.952	101.104	99.546	100.002	100.002	97.650	100.851

Conclusion and Discussion

This research aimed to discover the joint distribution of variables within the CoMensha human trafficking dataset by applying King's EI model. Based on the estimation of the EI model, it is conceivable to infer that there were detectable trends between different groups. When the statistics were broken down by gender, the majority of the victims were adults and came from other countries. However, most of the males were exploited in ways other than sexual, whereas many of the females were sexually exploited. In different age groups, it can be demonstrated that a larger percentage of adults were sexually exploited, while a greater proportion of minors were exploited in other ways. In terms of nationality, Dutch victims were no different from victims of other nations when compared in gender and type of exploitation. In both groups, female and sexual exploitation accounted for a larger percentage for nearly all the years. When comparing victims of sexually exploited and other ways of exploitation, females, and adults accounted for a higher proportion in the sexually exploited group, whereas males and minors represented a greater share in the other types of exploitation group. However, victims from other nations made up a bigger fraction of both groups.

Despite the true percentages of the joint distribution could not be verified, this method gives some insight into the dataset. Part of the findings from the joint distribution of sex and exploitation, as well as exploitation and nationality, were consistent with the trend in the prior study (UNODC, 2022). The outcome of the potential distribution offers the relationship and patterns between variables for future use in policy-making, resource management, or further research. This study, however, has some limitations that should be considered. As previously stated, it is difficult to verify the actual distribution of data, and some cases overlapped across several lists, which did not occur in the election data used by King in his model. The complete observation of marginal frequency has an influence on the accuracy of the results, which implies that missing values in marginal data will produce estimate challenges. When examining the study's findings, it is discovered that some of the summations of expected percentages are not near 100, and it is unclear why this is the case.

Future research should consider the constraints noted above when making predictions. Other statistical estimating methods might be used and evaluated in future research to delve further

into the data. This work focuses on two-by-two contingency tables; approaches such as Bayesian inference may be used to infer R by C tables. It is ideal to learn more about the data, but there are some aspects to consider when performing research. With limited information on ecological inference, it is difficult to link it back to specific persons, maintaining confidentiality. Since the results' reliability and correctness cannot be validated, they should be taken with caution. The potential impacts of the research on society and the battle against human trafficking should be evaluated and stereotypes of certain groups should be avoided. It is achievable to enhance societal fairness and equality through the proper use of the data.

Acknowledgments

I would like to express my deepest appreciation to Maarten for guiding me with his knowledge and experience.

Bibliography

Chaubey, Y. P., Nebebe, F., & Sen, D. (2003, November). Estimation of joint distribution from marginal distributions. In Proceedings of the American Statistical Association, Section on Survey Research Methods (pp. 883-888).

Collingwood, L., Oskooii, K. a. R., Garcia-Rios, S., & Barreto, M. A. (2016). eiCompare: Comparing Ecological Inference Estimates across EI and EI: RC. *R Journal*, *8*(2), 92. https://doi.org/10.32614/rj-2016-035

Dobra, A., Fienberg, S., E., & Trottini, M. (2002). Assessing the Risk of Disclosure of Confidential Categorical Data. *Bayesian Statistics*, *7*, 125–144.

Dobra, A., Tebaldi, C., & West, M. (2006). Data augmentation in multi-way contingency tables with fixed marginal totals. Journal of Statistical Planning and Inference, 136(2), 355–372. https://doi.org/10.1016/j.jspi.2004.07.002

Frogner, C., & Poggio, T. (2019). Fast and Flexible Inference of Joint Distributions from their Marginals. In *International Conference on Machine Learning* (pp. 2002–2011). http://proceedings.mlr.press/v97/frogner19a/frogner19a.pdf

Goodman, L. A. (1953). Ecological Regressions and Behavior of Individuals. *American Sociological Review*, *18*(6), 663. https://doi.org/10.2307/2088121

Karwa, V., & Slavkovic, A. B. (2013). Conditional inference given partial information in contingency tables using Markov bases. *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.1256

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press.

King, G., & Roberts, M. (2012). EI: A (n R) Program for Ecological Inference. *Harvard University*, 76.

King, G., Rosen, O., & Tanner, M. A. (1999). Binomial-Beta Hierarchical Models for Ecological Inference. *Sociological Methods & Research*, *28*(1), 61–90. https://doi.org/10.1177/0049124199028001004

King, G., Rosen, O., & Tanner, M. A. (2004). *Ecological Inference: New Methodological Strategies*. http://ci.nii.ac.jp/ncid/BA69497001

Rosen, O., Jiang, W., King, G., & Tanner, M. A. (2001). Bayesian and Frequentist Inference for Ecological Inference: The RxC Case. *Statistica Neerlandica*, *55*(2), 134–156. https://doi.org/10.1111/1467-9574.00162

United Nations Office on Drugs and Crime (UNODC). (2022). Monitoring human trafficking prevalence through multiple system estimation 2022.

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

Appendix

Estimation of Goodman's Regression

Table 18 Goodman's mean estimates of adult and minor distribution by sex (percentage)

	2016		2017		2018		2019	
	Male	Female	Male	Female	Male	Female	Male	Female
Adult	67.902	83.777	94.140	66.164	100.019	93.085	93.620	93.256
SE	81.952	86.264	82.130	85.678	16.059	17.790	25.138	25.974
Minor	32.098	16.223	5.860	33.836	-0.019	6.915	6.380	6.744
SE	81.952	86.264	82.130	85.678	16.059	17.790	25.138	25.974
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 19 Goodman's mean estimates of male and female distribution by age (percentage)

	2016		2017		2018		2019	
	Adult	Minor	Adult	Minor	Adult	Minor	Adult	Minor
Male	40.473	59.419	52.443	21.605	47.651	-132.961	47.414	42.882
SE	113.908	87.134	107.100	79.562	540.389	296.568	394.963	223.983
Female	59.527	40.581	47.557	78.395	52.349	232.961	52.586	57.118
SE	113.908	87.134	107.100	79.562	540.389	296.568	394.963	223.983
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 20 Goodman's mean estimates of Dutch and other nationalities distribution by sex (percentage)

	2016		2017		2018		2019	
	Male	Female	Male	Female	Male	Female	Male	Female
NL	-6.984	14.714	-4.893	21.168	-2.504	15.072	-0.072	19.611
SE	64.866	68.388	77.057	81.181	39.939	44.132	61.775	64.055
Other	106.984	85.286	104.893	78.832	102.504	84.928	100.072	80.389
SE	64.866	68.388	77.057	81.181	39.939	44.132	61.775	64.055
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 21 Goodman's mean estimates of male and female distribution by nationality (percentage)

	2016		2017		2018		2019	
	NL	Other	NL	Other	NL	Other	NL	Other
Male	8.086	46.746	15.530	48.063	-27.711	46.286	13.223	50.710
SE	93.290	146.421	80.920	120.518	130.346	213.717	93.066	146.868
Female	91.914	53.254	84.470	51.937	127.711	53.714	86.777	49.290
SE	93.290	146.421	80.920	120.518	130.346	213.717	93.066	146.868
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 22 Goodman's mean estimates of sexual exploitation and other types of exploitation distribution by sex (percentage)

	2016		2017		2018		2019	
	Male	Female	Male	Female	Male	Female	Male	Female
Sexual	-50.802	91.813	-44.800	80.561	-23.281	97.108	-9.898	98.106
SE	26.892	28.153	37.345	38.951	22.099	26.590	78.628	80.801
Other	150.802	8.187	144.800	19.439	123.281	2.892	109.898	1.894
SE	26.892	28.153	37.345	38.951	22.099	26.590	78.628	80.801
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 23 Goodman's mean estimates of male and female distribution by exploitation (percentage)

	2016		2017		2018		2019	
	Sexual	Other	Sexual	Other	Sexual	Other	Sexual	Other
Male	-3.636	63.600	-9.459	62.467	-0.857	78.459	18.030	73.506
SE	12.735	15.235	20.365	25.802	17.108	15.766	41.653	43.057
Female	103.636	36.400	109.459	37.533	100.857	21.541	81.970	26.494
SE	12.735	15.235	20.365	25.802	17.108	15.766	41.653	43.057
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 24 Goodman's mean estimates of adult and minor distribution by nationality (percentage)

	2016		2017		2018		2019	
	NL	Other	NL	Other	NL	Other	NL	Other
Adult	50.505	78.378	21.707	84.322	61.561	97.854	63.821	88.210
SE	88.281	138.469	63.423	94.173	8.132	13.003	39.942	63.283
Minor	49.495	21.622	78.293	15.678	38.439	2.146	36.179	11.790
SE	88.281	138.469	63.423	94.173	8.132	13.003	39.942	63.283
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 25 Goodman's mean estimates of Dutch and other nationalities distribution by age (percentage)

	2016		2017		2018		2019	
	Adult	Minor	Adult	Minor	Adult	Minor	Adult	Minor
NL	0.964	19.066	-2.939	54.560	-4.666	245.438	0.722	64.162
SE	83.158	63.296	81.637	61.149	89.980	50.268	159.791	101.386
Other	99.036	80.934	102.939	45.440	104.666	-145.438	99.278	35.838
SE	83.158	63.296	81.637	61.149	89.980	50.268	159.791	101.386
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 26 Goodman's mean estimates of adult and minor distribution by exploitation (percentage)

	2016		2017		2018		2019	
	Sexual	Other	Sexual	Other	Sexual	Other	Sexual	Other
Adult	88.462	72.275	77.016	79.891	91.418	98.469	82.591	91.303
SE	55.938	66.659	60.721	76.416	12.625	12.936	27.519	27.725
Minor	11.538	27.725	22.984	20.109	8.582	1.531	17.409	8.697
SE	55.938	66.659	60.721	76.416	12.625	12.936	27.519	27.725
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 27 Goodman's mean estimates of sexual exploitation and other types of exploitation distribution by age (percentage)

	2016		2017		2018		2019	
	Adult	Minor	Adult	Minor	Adult	Minor	Adult	Minor
Sexual	36.929	-3.696	22.007	27.829	31.063	359.467	37.327	129.138
SE	162.179	123.978	151.932	112.711	675.140	376.685	337.506	210.669
Other	63.071	103.696	77.993	72.171	68.937	-259.467	62.673	-29.138
SE	162.179	123.978	151.932	112.711	675.140	376.685	337.506	210.669
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 28 Goodman's mean estimates of Dutch and other nationalities distribution by exploitation (percentage)

	2016		2017		2018		2019	
	Sexual	Other	Sexual	Other	Sexual	Other	Sexual	Other
NL	14.312	1.371	15.995	7.379	18.073	-1.653	14.405	5.905
SE	45.316	54.248	58.499	73.429	33.024	33.843	49.456	49.803
Other	85.688	98.629	84.005	92.621	81.927	101.653	85.595	94.095
SE	45.316	54.248	58.499	73.429	33.024	33.843	49.456	49.803
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 29 Goodman's mean estimates of sexual exploitation and other types of exploitation distribution by nationality (percentage)

	2016		2017		2018		2019	
	NL	Other	NL	Other	NL	Other	NL	Other
Sexual	73.732	25.037	40.359	21.776	166.925	37.069	75.700	46.336
SE	137.015	215.241	110.591	164.387	155.479	248.616	129.274	203.454
Other	26.268	74.963	59.641	78.224	-66.925	62.931	24.300	53.664
SE	137.015	215.241	110.591	164.387	155.479	248.616	129.274	203.454
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000