Sains Malaysiana 54(10)(2025): 2553-2566
http://doi.org/10.17576/jsm-2025-5410-17

Spatial Cluster Analysis of Human Trafficking in Asia Over a Six-Year Period

(Analisis Kluster Reruang Pemerdagangan Manusia di Asia Sepanjang Tempoh Enam Tahun)

Safat Mohammad Safat, Nuzlinda Abdul Rahman*, Nurul Syafiah Abd Naeeim & Fauhatuz Zahroh Shaik Abdullah

School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia

Received: 24 March 2025/Accepted: 27 August 2025

ABSTRACT

Human trafficking is an important issue which affects many regions of the world. Understanding the spatial distribution of detected trafficking victims is essential to combating modern slavery. The main objective of this study was to determine the hotspot regions of trafficking victims in Asia. The yearly number of detected human trafficking victims for 49 countries, from the year 2016 to 2021, was analysed in this study. The 'hidden nature' of human trafficking leads to a substantial amount of missing data. To perform an in-depth spatial clustering analysis, the missing values were first addressed using several imputation techniques. Spatial clustering techniques are then used to locate the high and low-rate clusters of victims in Asian countries. The results indicated that repeated high-rate clusters consist of countries known for being travel hubs. These clusters also include the bordering and nearby countries which are easily accessible by land transportation. Repeated low-rate clusters do not yield conclusive results yet they infer that the countries located in these clusters may require additional resources to accurately report on the statistics of human trafficking. Spatial clustering analysis was also conducted on the covariates of age, form of exploitation and the sex of the victims. The findings show that the comparison of clusters for different variables can help determine which specific populations are most susceptible to human trafficking with their respective locations. Enforcement agencies and nonprofit organizations can utilize these findings to strengthen their methods of combating human trafficking.

Keywords: Spatial analysis; clustering; human trafficking; SaTScan; scan statistics

ABSTRAK

Pemerdagangan manusia adalah isu penting yang memberi kesan kepada banyak rantau di dunia. Memahami taburan reruang bagi mangsa pemerdagangan manusia yang dikesan adalah penting dalam memerangi perhambaan moden. Objektif utama kajian ini adalah untuk menentukan kawasan panas bagi mangsa pemerdagangan manusia di Asia. Jumlah tahunan mangsa pemerdagangan manusia yang dikesan bagi 49 negara, dari tahun 2016 hingga 2021 telah dianalisis dalam kajian ini. 'Sifat tersembunyi' pemerdagangan manusia menyebabkan sejumlah besar data tidak direkodkan. Bagi menjalankan analisis pengelompokan reruang yang mendalam, nilai yang hilang telah ditangani terlebih dahulu menggunakan beberapa teknik imputasi. Seterusnya, teknik pengelompokan reruang digunakan untuk mengenal pasti pengelompokan kadar tinggi dan rendah mangsa di negara Asia. Keputusan menunjukkan bahawa kelompok kadar tinggi yang berulang terdiri daripada negara yang dikenali sebagai hab perjalanan. Kelompok ini juga merangkumi negara-negara berjiran dan berhampiran yang mudah diakses melalui pengangkutan darat. Kelompok kadar rendah yang berulang tidak memberikan dapatan yang konklusif namun ia memberikan gambaran bahawa negara dalam kelompok ini mungkin memerlukan sumber tambahan untuk melaporkan statistik pemerdagangan manusia dengan lebih tepat. Analisis pengelompokan reruang turut dijalankan ke atas kovariat umur, bentuk eksploitasi dan jantina setiap mangsa. Penemuan menunjukkan bahawa perbandingan antara kelompok bagi pemboleh ubah yang berbeza dapat membantu menentukan populasi tertentu yang paling terdedah kepada pemerdagangan manusia berserta lokasi masing-masing. Agensi penguatkuasaan undang-undang dan organisasi bukan berasaskan keuntungan boleh menggunakan penemuan ini untuk memperkukuhkan kaedah mereka dalam memerangi pemerdagangan manusia.

Kata kunci: Analisis reruang; pengelompokan; pemerdagangan manusia; SaTScan; statistik imbasan

INTRODUCTION

"No one shall be held in slavery or servitude: slavery and the slave trade shall be prohibited in all their forms", is a declaration which was introduced by the United Nations General Assembly in December 1948 (Burke 2022). Thus far, this ideal has yet to be fully attained. 'Modern slavery', a term that is commonly referred to when describing human trafficking, is best defined as the exploitation of a human being for the purpose of profit using force, coercion, fraud or abuse of power. A victim of trafficking cannot simply refuse or escape due to threats, deception and the fear of violence (David et al. 2019). Even if this form of modern slavery has been occurring globally since the beginning of human history, it has not been regarded as a significant worldwide issue until the recent 1990s. Despite the vast amounts of quantitative and qualitative research carried out on human trafficking, creating effective methods to end the trafficking of persons is still far from being reached (Leach 2022).

Victims are trafficked for several reasons including organ removal, illegal adoption, exploitative begging, forced marriage and forced criminal activity. The two most common forms of exploitation are forced labour and sexual exploitation (UNODC 2022). In 2020, about 38.8% of detected trafficking victims around the world were subject to forced labour while around 38.7% were victims of sexual exploitation (UNODC 2022). Men, women, girls and boys are all at risk of trafficking. The distribution of detected human trafficking victims in 2020 was 42%, 23%, 18%, and 17% for women, men, girls and boys, respectively (UNODC 2022). The statistics mentioned are aggregated global results and not a representation of the human trafficking situations in individual regions.

Officials and researchers from governmental, intergovernmental and non-governmental organizations agree there is a scarcity of reliable data involving human trafficking victims (Kangaspunta 2003). Providing an accurate count of the total number of victims currently 'enslaved' is nearly impossible. The present estimated range can be anywhere from twenty to forty million people (Leach 2022). The Global Programme against Trafficking in Human Beings is a global database of information and trends concerning the trafficking of human beings around the world. These statistics are published in the form of yearly reports by the United Nations Office on Drugs and Crime (UNODC). The office collects various open-source information regarding the routes used by traffickers as well as the characteristics of reported victims and convicted traffickers. The main sources include official government reports, information from intergovernmental and non-governmental organizations, research and media reports (Kangaspunta 2003).

An exploratory analysis of the UNODC's data concluded that the majority of the countries of origin and the destination countries of human trafficking victims are located in Asia (Kangaspunta 2003). According to the 'Global Report on Trafficking in Persons 2022', the number of victims detected per one hundred thousand population in East Asia and the Pacific was 0.83 and 0.43 in South Asia, for the year 2019. Crime maps and trafficking flows from the report clearly indicate that the trafficking of human beings is prevalent in the underdeveloped and developing countries in the Asian region.

Crime mapping was first introduced in 1829 by geographer Adriano Balbi and lawyer Andre-Michel Guerry. However, law enforcement agencies did not start using crime maps until the early 1900s. Simple 'push pins' were used on basic wall maps, called 'spot maps', in order to designate the locations where crimes had occurred (Weisbird & McEwan 2015). By definition, crime mapping is a process that uses the geographical information regarding crime events in order to detect spatial patterns of criminal activity (Boba 2001). By the late 1990s, computer software programs, such as Geographic Information System (GIS) were commonly used by law enforcement agencies visualize and analyse the geographical data of crimes along with specific variables of interest (Boba 2001). Virtual point crime maps were utilized for illustrating the specific locations of recorded crime as well as to produce density maps. These types of maps produced quantitative values for representing the areas in which multiple similar crime related events had occurred. These crimes were either distributed over a large area or clustered around specific locations. These clusters of events were termed 'hot spots' (Burgess 2011).

Clustering analysis is a widespread method of analysing data by grouping the data into meaningful clusters according to their similarities in characteristics (Maimon & Rokach 2010). When incorporating the spatial context of the data, this becomes known as 'spatial clustering'. Knox (1987) defines a spatial cluster as 'a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance' (Fischer & Getis 2010). The geographic location where the response of interest was observed, otherwise known as spatial information, is vital information. In spatial statistics, one is not only interested in the statistics but as well the specific location of the statistics in question (Schabenberger & Gotway 2005).

The application of spatial clustering analysis can be useful in numerous disciplines. A disease related study explored the spatio-temporal clusters of dengue fever in Peninsular Malaysia using Kulldorff's cylindrical space-time scan statistic, one of the most prevalently used for spatio-temporal clustering analysis, especially with regards to disease clusters (Naeeim & Rahman 2022). SaTScan is a software program which employs Kulldorff's space-time scan statistic to identify the locations of spatial, temporal or spatio-temporal clusters. Another study utilized SaTScan to investigate the spatio-temporal clustering of COVID-19 cases in the state of Johor in Malaysia. The scan statistic determined the virus spread rapidly in populous areas of Johor, especially during the holidays (Ying 2022). Kulldorff's space-time scan statistic is one method which is widely used to determine disease related clusters. Other recent health and disease related studies applying the use of SaTScan to locate clusters include the cluster detection of Legionnaires' disease in the UK (Edens et al. 2019), the clustering analysis of pulmonary tuberculosis in Sichuan, China (Li et al. 2020) and the spatio-temporal cluster analysis of childhood cancer in California (Francis et al. 2020). Spatio-temporal clustering analysis techniques could potentially be invaluable in supporting health care organizations. They could assist with predicting the areas and times with a higher potential risk of disease incidence.

SaTScan and Kulldorff's space-time scan statistic is also employed in various other fields in order to discover significant relationships between the subject matter and space, time or both space and time. Spatio-temporal clustering analysis has been recently employed for the purpose of studying natural disasters and their effects on the environment. Following the tornado that occurred in Joplin, Missouri in 2011, the progress of recovery and the rebuilding of the community was investigated using SatScan (Stimers et al. 2022). SaTScan was used to conduct a spatio-temporal clustering analysis on the environmental factors affecting forest and land fires in Indonesia (Purwaningsih & Cintami 2019). These types of studies help identify the specific locations which are most vulnerable to natural hazards.

Kulldorff first developed SaTScan in 1997 with a group of epidemiologists and statisticians (Kulldorff 2022). Although the main focus of SaTScan was the cluster analysis of diseases, it has also been used for the mapping of crimes. One such crime related study hypothesized that the spatial and temporal movement of homicide in Newark, New Jersey would mimic the spatio-temporal distribution of an infectious disease. The mapping of the homicides and the following spatio-temporal analysis concluded that clusters initially began in the centre of the city but slowly expanded south and westwards throughout the period from 1982 to 2008 (Zeoli et al. 2012). These methods of spatial and temporal clustering analysis on past events and present data can be useful for predicting the potential times and locations of future events.

Analysing the details of victim profiles and utilizing the information in conjunction with the geographical location of past detected trafficking victims can benefit policymakers of programs combating human trafficking. Analysts working for governmental and non-profit organizations can build a victim profile which aids authorities in determining which categories of victims are most at risk of trafficking and in which regions of Asia. This information can also assist organizations in determining which regions need increased awareness of the trafficking situation in order to attempt the prevention of further trafficking. Despite the possible useful inferences which can be concluded from studying the specific geographical distribution of human trafficking victims in Asia, there are no studies conducted on such a large scale. The geographical distribution of human trafficking victims has only been analysed for specific individual countries such as India and China (Dakua, Rahaman & Das 2023; Xia et al. 2022). No recent similar studies have been applied to the entire continent of Asia. Therefore, the purpose of this study is to determine the locations of high-rate and low-rate spatial clusters of human trafficking victims in Asia as well to determine the locations of spatial clusters based on the categorical variables of sex, age and the form of exploitation of human trafficking victims in Asia.

METHODOLOGY

DATA DESCRIPTION

The data used in this study was the yearly counts of detected human trafficking victims of 49 countries from the continent of Asia. The yearly counts for the categorical variables of age, form of exploitation and sex, are also used for this study. The period for this study is from year 2016 to 2021. The United Nations defines the term 'detected trafficking victims' as the number of persons identified by national authorities as victims of human trafficking (UNODC 2022). In this study, the term 'detected trafficking victims' was used when referring directly to the raw dataset obtained from the United Nations.

The data was obtained directly from the United Nations Office on Drugs and Crime (UNODC) database. The data is publicly available on their official website. 'National data are submitted by Member States to UNODC through the United Nations Questionnaire for the Global Report on Trafficking in Persons (GLOTIP) or other means' (UNODC 2022). The UNODC compiled their dataset using data from the national authorities of individual countries which are designated for the identification of human trafficking victims including law enforcement agencies, criminal justice systems and National Referral Mechanisms. It is possible that the member states obtained their data from other sources such as government reports or official websites of national authorities. UNODC (2022) reviews the data for consistency and reliability via comparison with other data sources.

The population estimates for each country in Asia were obtained from the International Database of the United States Census Bureau, an official website of the United States government. The total midyear populations, populations by sex (male and female) and populations by age (single year age groups from 0 to 100+) are provided (United States Census Bureau). The population data is required for calculating the potential risk of becoming a victim of trafficking in a specific country as well as for identifying spatial clusters using SaTScan (Ying 2022).

MISSING VALUES IMPUTATION

The handling of missing data and outliers is one of the most important steps of any data analysis process. The traditional methods of handling missing data include either the deletion of data or a form of imputation (Smith & Climer 2024). Due to a substantial number of missing values, multiple different imputation methods were utilized.

These include cold-deck imputation, hot-deck imputation, mean imputation (MI), nearest neighbour imputation and regression imputation (Spies 2017). The deletion of data is also utilized when necessary. Imputation is applied to both the yearly counts of detected human trafficking victims as well the counts for the individual categorical variables.

Cold-deck imputation is a method which imputes a missing value with a replacement value from a chosen set of other similar sources (Mohd Izham 2018). The only other official similar source which provides publicly accessible information about the victims of human trafficking is U.S. Department of State, a governmental organization which collects and reports data at a scale which is similar to the UNODC. The data is provided in the form of yearly reports on the human trafficking situation for each country.

Prior to implementing cold-deck imputation, a two-sample t-test is used to compare the two datasets to determine whether there is a significant difference between the two groups. If there is no significant difference, then cold-deck imputation is used. The two-sample t-test is used to test the hypothesis of equal means between two independent groups (Curran 2013):

$$H_0 : \mu_{UNODC} = \mu_{U.S.\ Department\ of\ State}$$
$$H_A : \mu_{UNODC} \neq \mu_{U.S.\ Department\ of\ State}$$

μ_{UNODC} and $\mu_{U.S.\ Department\ of\ State}$ are the means of the UNODC and U.S. Department of State datasets, respectively. The null hypothesis indicates the mean count of detected trafficking victims for each country are equal between the two datasets. If the null hypothesis of equal means is rejected, a combination of hot-deck, mean value and nearest neighbour imputation will be used (Spies 2017). Hot-deck imputation is conducted using values from the same source but from a different record, such as a different year (Statistics Canada 2024). The two to three years prior to the year of interest, for which the values are known is used for mean value imputation. The use of these 'nearest neighbours' for imputation is known as deterministic hot-deck imputation (Andridge & Little 2010). If only a single year is available, direct hot-deck imputation is used. If there are no counts available for the country, direct cold-deck imputation from the U.S. Department of State's dataset is used. If no data is available from either source, the country is deleted from the dataset in order to simplify the computation.

For the yearly counts which are cold-deck imputed, the corresponding covariate counts are also imputed using cold-deck imputation. If cold-deck imputation is not possible, a model is used to determine the distribution of the overall counts into the individual categorical variables. The available data of the original and imputed categorical variable counts are used to predict the counts of the missing values. This method, known as regression imputation, preserves the relationship between missing values and other variables (Zhang 2016) by fitting a regression model on the variables of interest.

SCAN STATISTIC USING SaTScan

A scan statistic is a statistical test used to determine if there are clusters in a point process or if it is randomly distributed (Kulldorff 2022). The three main types of scan statistics are spatial, temporal and space-time scan statistics. For the purposes of this study, only the spatial scan statistic is used. A circular scanning window is positioned in the centre of several specified grid points on a map of the study region. The size of each window changes continuously from zero to the upper limit allowed by the SaTScan program. The software program creates a limitless number of distinct circular scanning windows which are composed of the different sets of neighbouring data points within them. Each window is then tested for the presence of a potential cluster (Kulldorff 2022). For this study, the specific location where the human trafficking victim was discovered is not available. Therefore, the latitudes and longitudes of the centre of each country are used as the origin of each circular scanning window.

LIKELIHOOD RATIO TEST

Each individual scanning window is tested for the presence of a potentially significant cluster. The likelihood ratio test is used to test the null hypothesis of equal risk inside and outside the scanning window (Kulldorff 2022). The formal hypothesis statement for this research study is as follows:

H_0: the risk of trafficking inside the scanning window is equal to outside

H_A: the risk of trafficking inside the scanning window is significantly different than outside

With the assumption that the yearly counts of detected human trafficking cases are Poisson distributed, the likelihood function for a specific window is proportional to

$$\left(\frac{c}{E[c]}\right)^{c}\left(\frac{T-c}{T-E[c]}\right)^{T-c} I(\,) \quad , \qquad (1)$$

where T is the total number of cases; c is the observed number of cases within the scanning window and $E[c]$ is the covariate adjusted expected number of cases within the scanning under the null hypothesis; and $T-E[c]$ is the expected number of cases outside the scanning window and I is an indicator function (Kulldorff 2022).

The indicator function is equal to one when scanning for high-rate clusters where the number of cases inside the window is significantly greater than the expected number of cases under the null hypothesis of equal risk inside and outside the window. Otherwise, it will be set to zero. For low-rate clusters, the indicator function is set to one when the number of cases within the window is less than the expected number of cases under the null hypothesis (Ying 2022).

For a purely spatial analysis and without the need for covariate adjustments, the expected number of cases $E[c]$ is computed in Equation (2):

$$E[c] = p \times \frac{T}{P} \quad , \tag{2}$$

where p is the population size in a specific area; T is the total number of cases; and P is the total population size (Ying 2022).

The likelihood function is maximized over all the scanning windows of different locations and sizes. The most probable cluster from this set of candidate clusters is the window with the maximum likelihood (Kulldorff 2022). The p-value is calculated by Monte Carlo hypothesis testing which compares the rank of the maximum likelihoods from the random data sets to the rank of the maximum likelihood from the real data set. If the rank is R then the p-value is calculated using Equation (3):

$$p - value = \frac{R}{1 + number\ of\ simulation} \ . \tag{3}$$

The number of simulations is restricted to either 999 or other numbers that end in 999 such as 1999, 9999, 99999 (Kulldorff 2022). The null hypothesis is rejected when the calculated maximum likelihood value for the most probable cluster from the real dataset is greater than the maximum likelihood value for the most probable cluster calculated from the simulation data (Ying 2022).

SaTScan will continue to run the scan statistics until all potential clusters are discovered. The calculated test statistics for all the clusters are then compared until the cluster with the maximum value of the likelihood test statistic is identified and recognized as the most probable cluster.

The multinomial scan statistic searches and tests clusters for whether the distribution of the cases inside the scanning window is different from the rest of the study area. In this study, the distribution of the variables age, sex and form of exploitation are tested. The null hypothesis states the probabilities for each category are equal:

$$H_0: p_1 = q_1, p_2 = q_2, \ldots, p_{K-1} = q_{K-1}, p_K = q_K$$

The alternative hypothesis claims the probability is not the same for at least one category (Jung et al. 2010). The likelihood function for multinomial model is

$$L(Z, p_1, \ldots, p_k, q_1, \ldots, q_k) \approx \prod_{k=1}^{K} \times \left(\prod_{i \in Z} p_k^{c_{ik}} \times \prod_{i \in Z} q_k^{c_{ik}} \right) \ , \tag{4}$$

where p_k is the probability of being in category k inside window Z and q_k is the probability of being in category k outside window Z (Jung et al. 2010 2010).

The relative risk value is any non-negative number which represents how much more probable it is for the event to occur in a specific area compared to the baseline (Kulldorff 2022). The relative risk is the estimated risk within the cluster divided by the estimated risk outside the cluster (Ying 2022). A value greater than one implies a greater risk while a value of less than one suggests a lower risk. The calculation for relative risk RR is

$$RR = \frac{c/E[c]}{(T-c)/(E[T]-E[c])} = \frac{c/E[c]}{(T-c)/(T-E[c])} \ , \tag{5}$$

where T is the total number of cases in the dataset and c is the observed number of cases within the cluster. Since the analysis is conditioned on the total number of cases observed, $E[T] = T$ (Kulldorff 2022). $c/E[c]$ is the observed number of cases within the cluster divided by the expected number of cases inside the same cluster when the null hypothesis is true. Thus, the equation for RR is a ratio between the observed divided by the expected inside the cluster and the observed divided by the expected outside the cluster (Kulldorff 2022). Figure 1 summarizes the process involved in determining clusters using SaTScan.

RESULTS AND DISCUSSIONS

Prior to determining the location of the spatial clusters, two-sample t-tests of the datasets from the UNODC and U.S. Department of State followed by various forms of imputation were implemented. The results are tabulated in Table 1.

In the process of locating the spatial clusters of human trafficking victims, the total counts for the countries in Asia for a specific year are modelled as a retrospective discrete Poisson model. High-rate clusters are scanned in order to determine the hotspot regions of human trafficking activity. Low-rate clusters are scanned to determine the regions which are either potentially safe from trafficking or are considered lacking in reliable data. SaTScan includes the option of simultaneously scanning for both high and low-rate clusters. With the absence of geographical overlap, the contradiction of including low-risk countries in high-rate clusters and vice-versa may be avoided. The cluster with a smaller value for the likelihood ratio test statistic would not be reported and this may result in the absence of significant statistics. To ensure that vital information is not disregarded, the identification of all statistically significant clusters regardless of geographical overlapping combined with the comparison of high-rate and low-rate clusters is considerably more appropriate.

The entire continent of Asia is under consideration in this research study. The maximum possible spatial cluster size is set to 10% of the population at risk to avoid the detection of extremely large circular clusters (AlQadi et al. 2022). If the upper limit of the cluster size is too large then clusters of low rates outside the circle are possibly detected instead of high rates within the circle (Kulldorff 2022). SaTScan does not restrict each individual location within the scanning window to have more cases than expected. Depending on the size of the scanning window, it is possible that low-risk locations are included in high-rate clusters and vice-versa (Abdurrob & Kulldorff 2016). However,

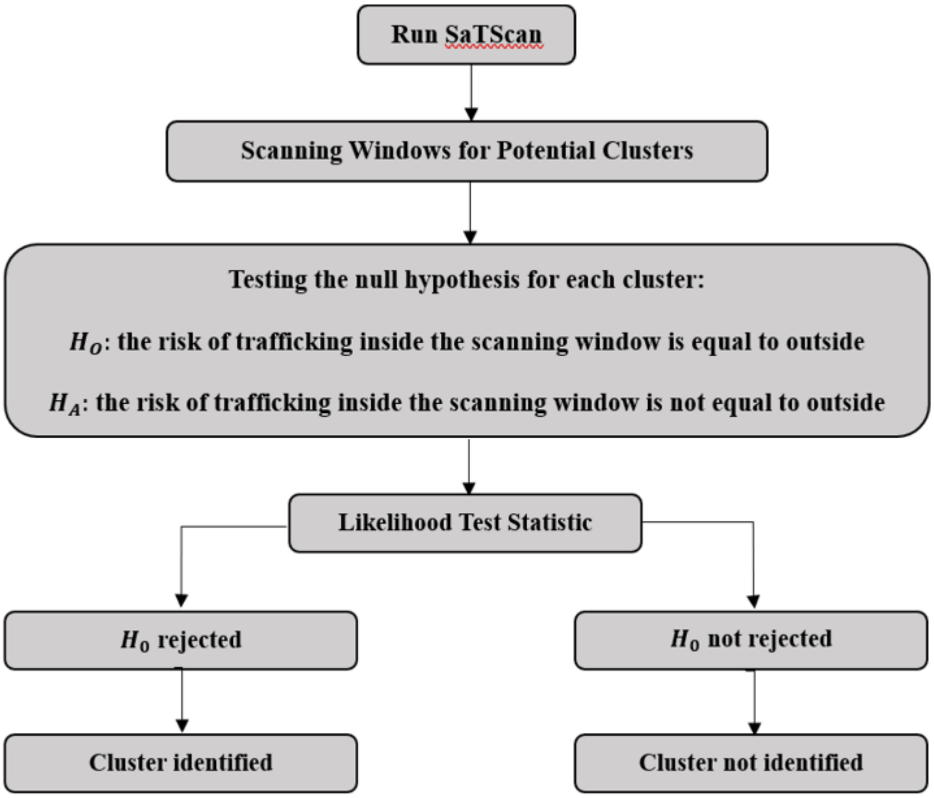

FIGURE 1. A flowchart for the process of determining clusters using SaTScan

TABLE 1. Number of human trafficking victims

Country	2016	2017	2018	2019	2020	2021
Afghanistan	116**	476**	434**	493**	550**	414***
Armenia	26	12	29	8	11	13
Azerbaijan	70	70	98	91	94	95
Bahrain	33	16	11	27	64	84
Bangladesh	355*	733	502*	585*	1683*	1138*
Bhutan	2	3	3	153	181	22
Brunei	5	3	3	3	3	0
Cambodia	326**	986**	222**	262**	417**	364**
China	698	1127	1194	1006***	1006***	1006***
Hong Kong	36**	28**	29**	3**	3**	1**
Macao	4*	3*	1*	0	0	0
Cyprus	30	24	42	36	26	22*
Georgia	16	12	7	31	6	4*
India	23	5898	5788	6616	4709	5934*
Indonesia	1957	1469	297	280	60	813***
Iraq	8	34	90	113	17	81
Israel	48	74	60	75	70	52

continue to next page

continue from previous page

Japan	50	46	28	48	39	48
Jordan	51	48	41	10	47	61
Kazakhstan	606	95	88	44	102	29
Kuwait	38	49	37***	37***	103*	17*
Kyrgyzstan	12	3	8	10	12	5*
Laos	184**	86**	52**	39**	142**	110**
Lebanon	90	57	30	39	87	108
Malaysia	678	1119	132	182	165	348
Maldives	12	3	3	3	3	0
Mongolia	61	46	44	68	39	45*
Myanmar	358	539	528	577	167	38
Nepal	354	312	546	615	2023	187*
Oman	62	46	36	19	12	16*
Philippines	603	1659	2318	2041	1598	1802*
Qatar	2	3	3	9	35*	28
South Korea	82*	77*	80***	80***	80***	80***
Turkey	196	304	135	215	282	402
Saudia Arabia	21	121*	162	585	173	185*
Singapore	33	17	32*	62*	23*	26*
Sri Lanka	23	46*	65	18	9	28*
Syria	10	3	3	7	3	7
Tajikistan	29	79	67	79	108	71
Thailand	850	903	417	1540	146	142
Timor-Leste	21	9***	9***	3	3	9***
Turkmenistan	7	20*	25*	24*	18*	15*
UAE	80	29	74	42	31	1*
Uzbekistan	774	693	494	224	131	243
Vietnam	1128**	670**	490**	300**	121**	126**

*Cold-deck imputation after two-sample t-test
**Cold-deck imputation without two-sample t-test
***Combination of hot-deck, nearest neighbor and mean-value imputation

if the upper limit is too small then statistically significant clusters of interest may be omitted. The minimum number of cases for each scanning window is set to two cases. Adjustments are not made for relative risks. The *p*-value is obtained by using the standard Monte Carlo with 999 simulations to indicate the significance of the cluster. A cluster is statistically significant when its likelihood ratio is greater than the calculated critical value at 1% significance level or the *p*-value is smaller than the significance level of 0.01.

For the discrete Poisson model in SaTScan, high-rate clusters are represented by red circles and low-rate clusters by blue circles. The dots within the circles are the hypothetical specified centres of the countries included within the clusters. It is possible for a country to be included in multiple clusters. Additionally, it is possible for a single dot to be classified as a point cluster, a cluster composed of a single country. Red and blue dots respectively representing high-rate and low-rate point clusters can be located both inside and outside the circles. The cluster, for both high-rate and low-rate clusters, with the highest calculated likelihood ratio test statistic is the most likely cluster and the other clusters are secondary clusters. The spatial cluster maps for the years of 2016, 2017, 2018, 2019, 2020 and 2021 are illustrated as shown in Figure 2.

Based on Figure 2, there are a total 60 high-rate clusters for the six years of the study period. There is one most likely high-rate cluster for each year of the study period from 2016 to 2021. Therefore, there are 6 most likely high-rate clusters and 54 secondary high-rate clusters. Four out of six years of the study period, Cambodia, Malaysia, Laos, Thailand, Myanmar, and Vietnam constitute the most likely high-rate clusters. Thailand, a popular travel hub, is a common transit country with a high frequency of incoming and outgoing traffic via air, sea and land travel. Peninsular Malaysia is easily accessible from south and travelling to Laos, Cambodia and Vietnam towards the east is also relatively simple. Myanmar also shares a land border with the western side of Thailand. Many tourists include these countries in their itinerary when traveling in South East Asia. This leads to a high demand for services from the influx of foreigners in addition to the local population. The ease of accessibility for relocation across land borders combined with a high demand for cheap services produces a hotspot area for traffickers to exploit large numbers of vulnerable victims for forced labour and sexual exploitation.

The characteristics of the most likely high-rate cluster for each year from Figure 2 are tabulated in Table 2. The countries in Table 2 are labelled as: *Cam* for Cambodia, *Viet* for Vietnam, *Tha* for Thailand, *Lao* for Laos, *Mal* for Malaysia, *Mac* for Macau, *Hon* for Hong Kong, *Bhu* for Bhutan, *Phi* for Philippines, *Sin* for Singapore, *Mya* for Myanmar, *Bru* for Brunei, *Nep* for Nepal and *Ban* for Bangladesh. The total population and the total number of cases of all the countries within each cluster observed in Table 2 were used for the calculations of the likelihood ratio and the p-value. From each set of high-rate clusters in Figure 2, the cluster with highest likelihood ratio statistic is determined to be the most likely cluster and the value is

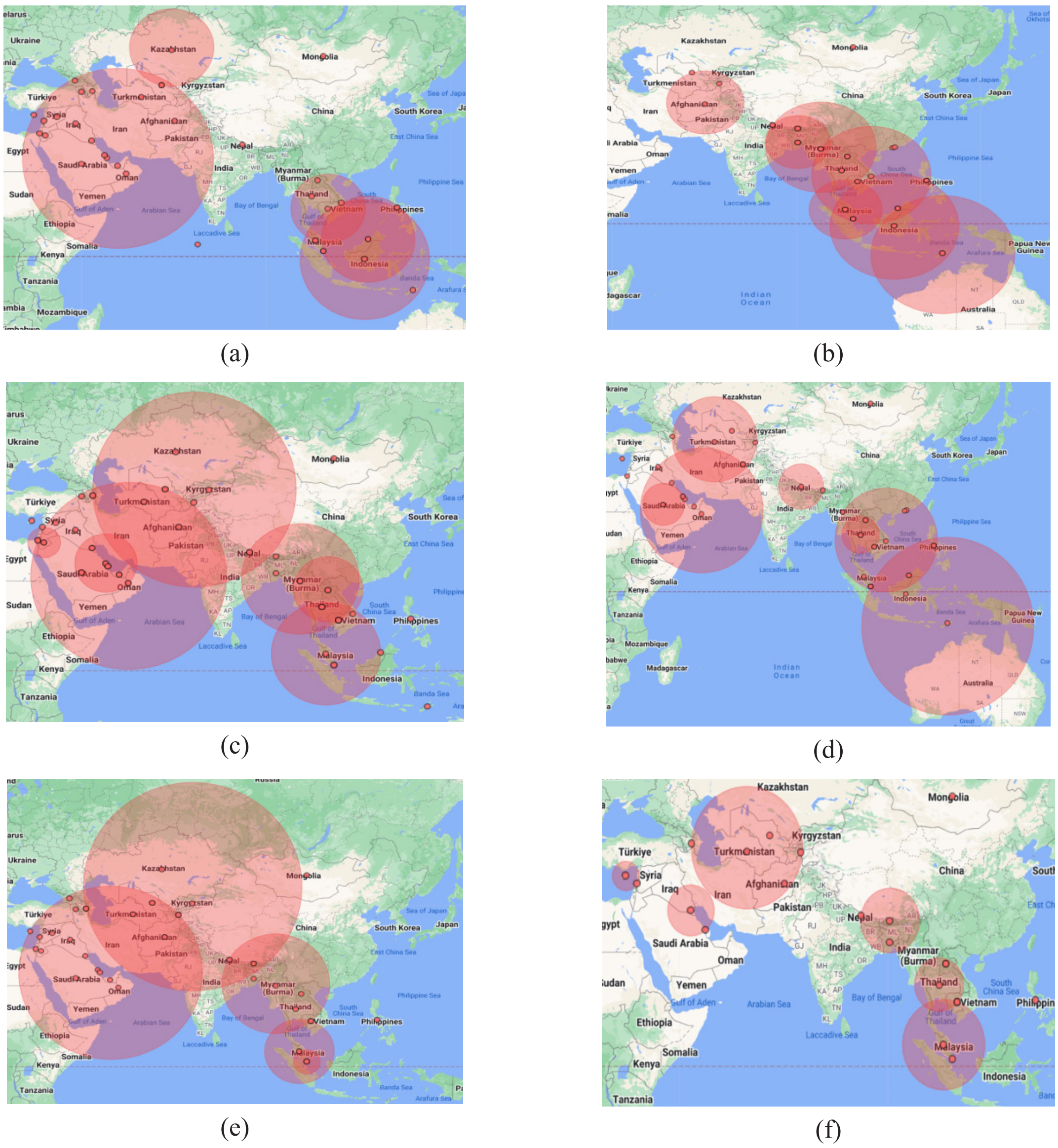

FIGURE 2. High-rate clusters identified for the years: (a) 2016, (b) 2017, (c) 2018, (d) 2019, (e) 2020 and (f) 2021

TABLE 2. Characteristics of most likely high-rate spatial clusters in Asia

Year	Countries	Population	Number of cases	Expected cases	Observed/ Expected	Relative risk	Likelihood ratio	p-value
2016	*Cam, Viet, Tim, Lao, Mal*	219855780	3166	861.62	3.67	4.34	2001.528	0.001
2017	*Viet, Cam, Tim, Lao, Mac, Hon, Bhu, Mal, Phi, Sin, Mya*	395455022	6013	1738.70	3.46	4.69	3803.121	0.001
2018	*Phi*	106513192	2318	379.77	6.10	7.05	2391.763	0.001
2019	*Viet, Cam, Tim, Lao, Mac, Hon, Bru, Mal, Phi, Sing, Mya*	403807747	5009	1616.74	3.10	4.00	2685.764	0.001
2020	*Mya, Ban, Lao, Bhu, Tha, Nep, Cam*	342638561	3139	1058.45	2.97	3.59	1524.964	0.001
2021	*Phi*	112771400	1802	379.05	4.75	5.30	1462.147	0.001

presented in Table 2. The low p-values of 0.001 for each high-rate most likely cluster indicate they are significant.

The low-rate clusters of yearly human trafficking victims in Asian countries are displayed in Figure 3. There are a total 52 low-rate clusters for the six years of the study period. There is one most likely low-rate cluster for each year of the study period from 2016 to 2021. Therefore, there are 6 most likely low-rate clusters and 46 secondary low-rate clusters.

The characteristics of the most likely low-rate cluster for each year are tabulated in Table 3. The countries in Table 3 are labelled as: *Jap* for Japan, *Sou* for South Korea, *Tim* for Timor-Leste, *Indo* for Indonesia, *Hon* for Hong Kong and *Mac* for Macau. Japan and South Korea are included in the most likely low-rate clusters for three out of six years of the study period. The low relative risk of human trafficking in these clusters does not indicate that the population is safe from the threat of trafficking. In the UNODC dataset, South Korea has many missing values which suggests the authorities and the system for reporting victims in the country may need improvements in the future. Japan and Indonesia are also both point clusters for several most likely and secondary significant low-rate clusters. The reason they are point clusters instead of clusters consisting of multiple countries may potentially be due to the fact that they are islands. It would be difficult to transport victims by sea or air. These results support the theory of land accessibility between bordering countries leading to significant clusters of multiple countries.

For the multinomial model in SaTScan, high-rate and low-rate clusters are not considered and thus the application of a red circle or red dot does not indicate high-rate clusters. The clusters are regions for which the distribution of the categorical variables are significantly different than the rest of the area (Jung, Kulldorff & Richard 2010). For this study, the multinomial model is used to search for clusters of the different levels of each categorical variable. This method considered all possible groupings of the categories and searches for clusters where the distribution of the categories is significantly different in comparison to the distribution of the categories outside the clusters. The p-value is used to determine the presence of a potential cluster at a significance level of 0.01. The values of the relative risks for each category are important for understanding the dangers of human trafficking with respect to different variables. A spatial clustering analysis is conducted for each year of the study period. Due to the substantial amount of information, only the year of 2019 is directly analysed in this paper.

The clusters for the categorical variables of age, form of exploitation and sex are illustrated in Figure 4. Clusters were identified for each variable for each year of the study period. A single most likely cluster was discovered for each categorical variable for each year, which produces a total of 18 most likely clusters. In addition, 152 secondary clusters were showed during the analysis.

The characteristics of the most likely cluster for each categorical variable is tabulated in Table 4. The number of cases, expected cases, the observed over expected value, the relative risk and the percent of the cases in the region for each level of the categorical variables are included. The likelihood ratio test statistic is the highest calculated for each variable respectively amongst all the possible clusters. Each cluster has a p-value of 0.001, indicating its significance.

The most likely spatial clusters in 2019 for the variables of age and sex are point clusters in Myanmar. The relative risk values for these point clusters indicate that the ages and sexes of detected trafficking victims were exceedingly likely to be unknown compared to other countries in Asia. Recent conflicts and internal strife within the country may have caused the population of Myanmar to become a hotspot for human trafficking. This might be due to insufficiency in the organization and reporting of accurate information pertaining to the personal details of trafficking victims within a conflict zone. The secondary clusters provide useful inferences due to the availability of more information on the details of the victims. Many useful inferences are summarized in Table 5.

The comparison of spatial clusters across different categorical variables such as age, form of exploitation and sex can assist local authorities and policy makers in building a victim profile. The results can be useful in determining the specific populations in certain countries who have a higher risk of becoming victims of trafficking. For example, the results in Table 5 indicate that a victim from Nepal is likely to be an adult male.

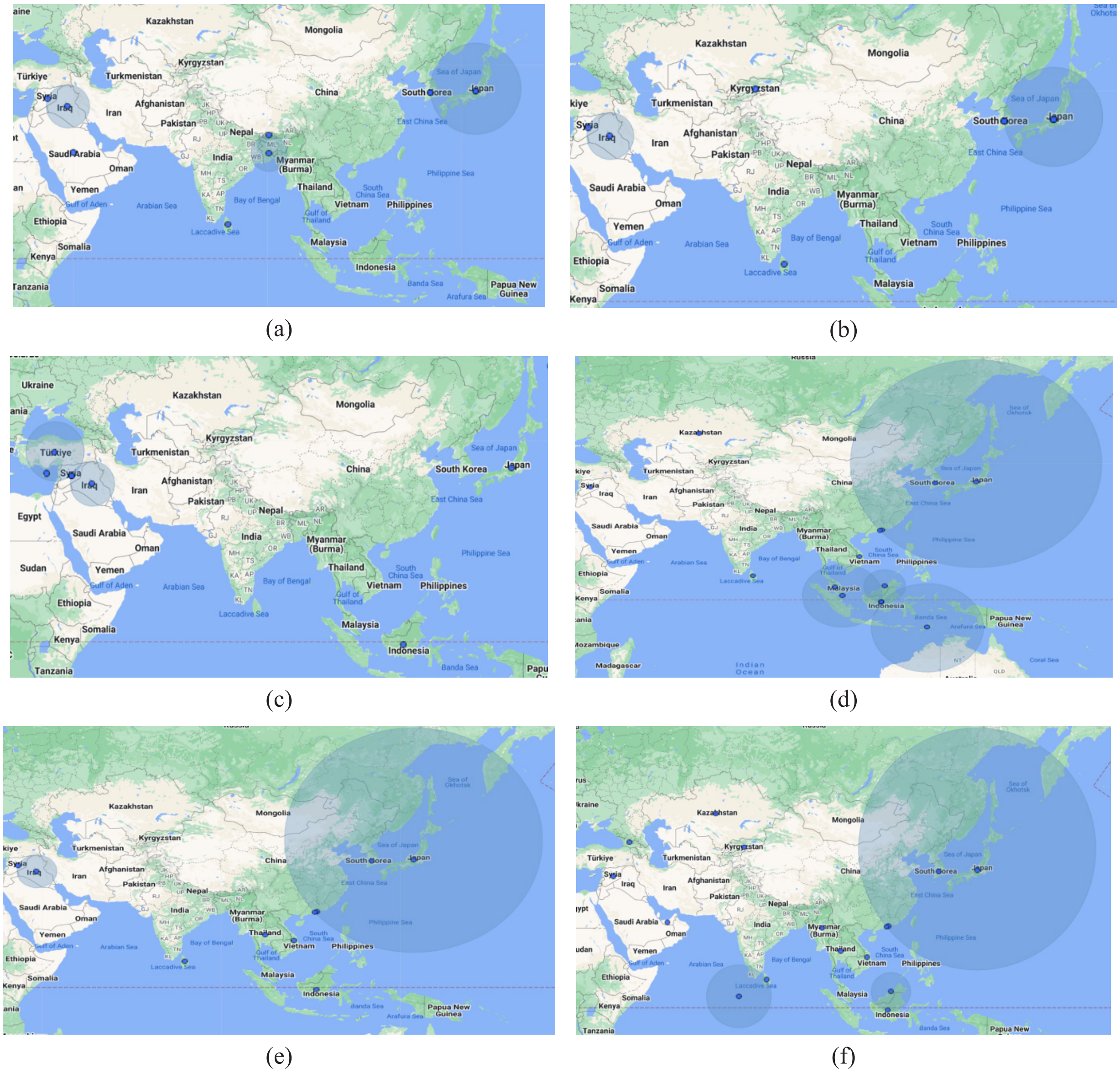

FIGURE 3. Low-rate clusters identified for the years: (a) 2016, (b) 2017, (c) 2018, (d) 2019, (e) 2020 and (f) 2021

TABLE 3. Characteristics of most likely low-rate spatial clusters in Asia

Year	Countries	Population	Number of cases	Expected cases	Observed/ Expected	Relative risk	Likelihood ratio	*p*-value
2016	*Jap, Sou*	177378296	132	695.15	0.19	0.18	354.125	0.001
2017	*Jap, Sou*	177349900	123	779.76	0.16	0.15	441.933	0.001
2018	*Jap*	125918128	28	448.95	0.062	0.061	349.393	0.001
2019	*Tim, Indo*	271884737	283	1088.56	0.26	0.25	444.742	0.001
2020	*Indo*	272856384	60	842.88	0.071	0.067	649.053	0.001
2021	*Jap, Sou, Hon, Mac*	184296086	129	619.45	0.21	0.20	296.787	0.001

(a) (b) (c)

FIGURE 4. Spatial clusters in 2019 for the categorical variables of (a) age, (b) form of exploitation and (c) sex

TABLE 4. Characteristics of most likely clusters of each categorical variable

Characteristics	Cluster
Variable: Age	
Location IDs included	Myanmar
Total Cases	577
Category	[0 to 17 years], [18 years or over], [Unknown]
Number of cases	84, 274, 219
Expected cases	173.000, 386.120, 17.880
Observed / expected	0.490, 0.710, 12.250
Relative risk	0.480, 0.700, 22.510
Percent cases in area	14.600, 47.500, 38.000
Likelihood ratio	450.617
p-value	0.001

continue to next page

continue from previous page

Variable: Form of Exploitation	
Location IDs included	Bhutan
Total Cases	153
Category	[Forced labour], [Sexual exploitation], [Other forms of exploitation]
Number of cases	153, 0, 0
Expected cases	43.510, 54.940, 54.550
Observed / expected	3.520, 0, 0
Relative risk	3.620, 0, 0
Percent cases in area	100.000, 0, 0
Likelihood ratio	194.496
p-value	0.001
Variable: Sex	
Location IDs included	Myanmar
Coordinates / radius	(21.914 N, 95.956 E) / 0 km
Total Cases	577
Category	[Male], [Female], [Unknown]
Number of cases	61, 297, 219
Expected cases	213.890, 328.840, 34.270
Observed / expected	0.290, 0.900, 6.390
Relative risk	0.280, 0.900, 8.190
Percent cases in area	10.600, 51.500, 38.000
Likelihood ratio	323.619
p-value	0.001

TABLE 5. Summary of relative risk inferences based on age, form of exploitation and sex

Cluster	Relative risk inference
Secondary cluster 1 (age)	The victim is 25% more likely to be an adult if they are from the region which includes Bhutan, Bangladesh and Nepal
Secondary cluster 2 (age)	The victim is 25% more likely to be an adult if they are from the region which includes Timor-Leste and Indonesia
Secondary cluster 3 (age)	The victim is 26% more likely to be a child if they are from the region which includes Hong Kong, Macao, Vietnam
Secondary cluster 1 (form)	The individual is 108% more likely to be a victim of forced labour if they are from Saudia Arabia
Secondary cluster 2 (form)	The individual is 105% more likely to be a victim of forced labour if they are from Cambodia
Secondary cluster 3 (form)	The individual is 137% more likely to be a victim of sexual exploitation if they are from the region which includes Japan, Hong Kong, Macao and Mongolia
Secondary cluster 1 (sex)	The victim is 34% more likely to be a female if they are from the region which includes Malaysia, Singapore, Cambodia and Vietnam
Secondary cluster 2 (sex)	The victim is 46% more likely to be a male if they are from the region which includes Saudia Arabia, Bahrain and Kuwait
Secondary cluster 3 (sex)	The victim is 73% more likely to be a male if they are from Nepal

CONCLUSIONS

To conduct a useful analysis in this study, missing values are first dealt with by determining appropriate estimations using a mixture of cold-deck, hot-deck, mean value, nearest neighbour and regression imputation techniques. As a result, these imputations allowed for most of the 49 countries to retain a sufficient amount of data to be effectively analysed. Purely spatial clustering techniques were used on the imputed dataset in order to determine the areas of high and low rates of human trafficking in Asia. Results indicated there are repeated high-rate clusters of trafficking victims in the regions which included countries recognized as popular transit destinations, such as Thailand. Countries easily accessible by land travel from Thailand, such as Myanmar, Malaysia, Cambodia and Laos were also part of these clusters. These clusters of bordering and nearby countries support the theory that many victims are regularly transported across land borders. Also, high and low-rate point clusters consisted of islands with no land borders. The lack of information from other countries within these clusters suggests that traffickers tend to avoid sea and air routes when transporting victims across borders. Japan and South Korea were repeatedly included in low-rate clusters. The presence of repeated low-rate clusters does not necessarily mean that these countries are safe from the risk of trafficking. Additional reliable data collection and reporting should be taken into consideration for the implementation of a more comprehensive analysis of the human trafficking situation in any country.

Purely spatial clusters were also used to analyse the spatial distribution of the age, form of exploitation and sex of the victims for the year 2019. The results varied from country to country but provided useful inferences. For example, the risk of sex trafficking was higher in the region of Hong Kong and Macau than in other areas of Asia. Also, compared to other regions, the victims in Hong Kong and Macau were more likely to be children. Therefore, the comparison of the location of clusters across different variables can be used to construct a potential victim profile which assists in identifying the specific individuals at risk of certain forms of trafficking and which regions are most dangerous for them.

The findings of this research study are vital in determining the hotspot regions of human trafficking in Asia as well as the regions for which specific populations are at a greater risk of trafficking. The hotspot locations of human trafficking victims determined by this study allow for the appropriate allocation of resources. The authorities can use this information to identify and rescue victims. By means of sharing more in-depth information and the cooperation between law enforcement agencies of multiple countries can result in the prevention of cross-border trafficking as well as the apprehension of traffickers at specific determined locations. Further study of low-risk regions can determine whether the region is either safe from trafficking or if the countries in the region require additional support to efficiently collect and organize data on victims. National, international and nonprofit organizations can provide this support by using these results to update policies and techniques for detecting human trafficking victims. These organizations can share the study findings with the general population to spread information and raise awareness on the risks of trafficking in order to prevent future trafficking.

The analysis in this study was conducted using several imputed values and therefore the results are not representative of the complete human trafficking situation. The most restricting limitation in this study is the presence of a large number of missing values. The UNODC and the U.S Department of State are the two most reliable organizations responsible for providing the publicly available data on detected trafficking victims. However, additional detailed data is required in order to conduct a more in-depth analysis. The proper collection, organization and representation of data is required to complete an accurate and comprehensive study. The cooperation of governmental organizations from several countries is also required for the construction of efficient statistical methods for gathering and presenting information on the demographics of victims. This data is essential for conducting comparable studies to obtain more appropriate and improved results. The results of this study imply that despite the varying factors among different countries which affect the risk of trafficking, there are groupings of countries which share similar characteristics. Note that this paper is not a commentary on the policies and enforcement capabilities of the countries mentioned in this paper.

REFERENCES

Abdurrob, A. & Kulldorff, M. 2016. *SaTScan Tutorial #3 Advanced Options.* Accessed 2nd September 2024. https://www.satscan.org/tutorials/advancedoptions/SaTScanTutorialAdvancedOptions.pdf

AlQadi, H., Bani Yaghoub, M., Wu, S., Balakumar, S. & Francisco, A. 2022. Prospective spatial-temporal clusters of COVID-19 in local communities: Case study of Kansas City, Missouri, United States. *Epidemiology and Infection* 151: e178.

Andridge, R.R. & Little, R.J.A. 2010. A review of hot deck imputation for survey non-response. *International Statistical Review* 78(1): 40-64.

Boba, R. 2001. *Introductory Guide to Crime Analysis and Mapping*. Washington: U.S. Dept. of Justice.

Burgess, M. 2011. Understanding crime hotspot maps. *Crime and Justice Statistics*. Issue paper No. 60.

Burke, M.C. 2022. *Human Trafficking Interdisciplinary Perspectives*. 3rd ed. New York: Routledge.

Curran, J.M. 2013. The frequentist approach to forensic evidence interpretation. In *Encyclopedia of Forensic Sciences*, edited by Siegel, J., Saukko, P. & Houck, M. London: Elsevier, Academic Press. pp. 286-291.

Dakua, T., Rahaman, M. & Das, K.C. 2023. An analysis of the spatial and temporal variations of human trafficking in India. *Cogent Social Sciences* 10(1): 2293315.

David, F., Latham-Sprinkle, J., Bryant, K. & Larsen, J. 2019. *Migrants and Their Vulnerability to Human Trafficking, Modern Slavery and Forced Labour*. London: International Organization for Migration.

Edens, C., Alden, N.B., Danila, R.N., Fill, M.-M.A., Gacek, P., Muse, A., Parker, E., Poissant, T., Ryan, P.A., Smelser, C., Angelo, T. & Schrag, S.J. 2019. Multistate analysis of prospective Legionnaires' disease cluster detection using SaTScan, 2011-2015. *PLoS ONE* 14(5): e0217632.

Fischer, M.M. & Getis, A. 2010. *Handbook of Applied Spatial Analysis Software Tools, Methods and Applications*. 1st ed. Berlin: Springer.

Francis, S.S., Enders, C., Hyde, R., Gao, X., Wang, R., Ma, X., Wiemels, J.L., Selvin, S. & Metayer, C. 2020. Spatial-temporal cluster analysis of childhood cancer in California. *Epidemiology* 31(2): 214-223.

Global Report on Trafficking in Persons. 2022. Accessed 16th August 2024. https://www.unodc.org/documents/data-and-analysis/glotip/2022/GLOTiP_2022_web.pdf

Jung, I., Kulldorff, M. & Richard, O.J. 2010. A spatial scan statistic for multinomial data. *Statistics in Medicine* 29(18): 1910-1918.

Kangaspunta, K. 2003. Mapping the inhuman trade: Preliminary findings of the database on trafficking in human beings. *Forum on Crime and Society* 3(1&2): 81-103.

Knox, E.G. 1989. Detection of clusters. In *Methodology of Enquiries into Disease Clustering,* edited by Elliot, P. London: Small Area Health Statistics Unit 17(1): 20.

Kulldorff, M. 2022. *SaTScan™ User Guide for version 10.1*. (2022). Accessed 16th August 2024.
https://www.satscan.org/cgi-bin/satscan/register.pl/SaTScan_Users_Guide.pdf?todo=process_userguide_download

Leach, R. 2022. Ending demand for modern-day slavery: An analysis of human trafficking in the global marketplace. *Indiana Journal of Global Legal Studies* 29(1): 232-259.

Li, T., He, J.G., Yang, C.H., Li, J., Xiao, Y., Li, Y.K., Chen, C. & Wu, J.L. 2020. A comparative study on SaTScan and FleXScan software for spatial clustering analysis regarding the incidence of pulmonary tuberculosis. *Zhonghua Liu Xing Bing Xue Za Zhi* 41(2): 207-212.

Maimon, O. & Rokach, L. 2010. *Data Mining and Knowledge Discovery Handbook.* 2nd ed. Springer.

Mohd Izham bin Mohd Jaya. 2018. Cold deck missing value imputation with a trust-based selection method of multiple web donors. PhD Thesis. Serdang: Universiti Putra Malaysia (Unpublished).

Naeeim, N.S.A. & Rahman, N.A. 2022. Spatio-temporal clustering analysis using two different scanning windows: A case study of dengue fever in Peninsular Malaysia. *Spatial and Spatio-Temporal Epidemiology* 41(4): 100496.

Purwaningsih, T. & Cintami, A. 2019. Analysis of factors affecting the area of forest and land fires in Indonesia uses spatial regression Geoda and SaTScan. *Jurnal Informatika* 12(2): 58.

Schabenberger, O. & Gotway, C.A. 2005. *Statistical Methods for Spatial Data Analysis*. 1st ed. New York: Chapman and Hall/CRC.

Smith, K. & Climer, S. 2024. *Improving Data Cleaning Using Discrete Optimization.* https://arxiv.org/abs/2405.00764

Spies, L. 2017. Possible imputation procedures for the Census 2021. In *Work Session on Statistical Data Editing*, 24-26 April 2017, Netherlands. The Hague: United Nations Economic Commission for Europe.

Statistics Canada. 2024. *3.4.4 Imputation*. Accessed 16th August 2024. https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch3/imputation/5214784- eng.htm

Stimers, M., Lenagala, S., Haddock, B., Paul, B.K. & Mohler, R. 2022. Space-time clustering with the space-time permutation model in SaTScanTM applied to building permit data following the 2011 Joplin, Missouri Tornado. *International Journal of Disaster Risk Science* 13(6): 962-973.

UNODC. 2022. *UNODC Research - Data Portal – Trafficking in Person.* Accessed 16th August 2024. https://dataunodc.un.org/sites/dataunodc.un.org/files/metadata_trafficking_in_ persons.pdf

Weisbird, D.L. & McEwan, T. 2015. *Introduction: Crime Mapping and Crime Prevention*. PhD Thesis. Hebrew University of Jerusalem (Unpublished).

Xia, H., Li, G., Zhou, J., Yu, Y., Du, M. & Feng, C. 2022. Spatial and temporal patterns, influencing factors, and prevention and control measures of human trafficking in southwest China. *Tropical Geography* 42 (9): 1534-1546.

Ying, F.F. 2022. Spatio-temporal clustering analysis of COVID-19 cases in Johor. Master's thesis. Universiti Sains Malaysia (Unpublished).

Zeoli, A.M., Pizarro, J.M., Grady, S.C. & Melde, C. 2012. Homicide as infectious disease: Using public health methods to investigate the diffusion of homicide. *Justice Quarterly* 31(3): 609-632.

Zhang, Z. 2016. Missing data imputation: Focusing on single imputation. *Ann. Transl. Med.* 4(1): 9.

*Corresponding author; email: nuzlinda@usm.my