Thai Journal of Mathematics
Volume 22 Number 1 (2024)
Pages 271–284

http://thaijmath.in.cmu.ac.th

ISSN 1686-0209

Annual Meeting in Mathematics 2023

A Comparison of Capture-recapture Modeling to Estimate the Number of Patients with Psoriasis in Trang Province

Orasa Nunkaw* **and Preedaporn Kanjanasamranwong**

Mathematics and Statistics Department, Faculty of Science, Thaksin University, Phattalung, Thailand
e-mail : aorasa@tsu.ac.th (O. Nunkaw); preedaporn@tsu.ac.th (P. Kanjanasamranwong)

Abstract This study aims to estimate the number of patients with psoriasis in Trang Province between 2015 and 2018. The population size estimator based on Poisson distribution called the MLEPoi was used as a basic model for homogeneity population. However, the heterogeneity often occurs in capture-recapture experiment. The population size estimators based on the Poisson mixture model; the MLEGeo, LCMP, TG, Chao, Zelterman and LB estimators were selected for heterogeneous population. By using the ratio plot of the Poisson model for investigating a suitable model for psoriasis data, the results suggested that the Poisson mixture model performs better than the Poisson model. Therefore, the LCMP and MLEGeo estimators provided the best accuracy with excellent goodness-of-fit over competitors. The TG estimator showed promise as an alternative choices to estimate population size in all cases, wile the LB estimator provided the worst result with violated assumption parameters. Using the most appropriate population size estimator determined the average number of hidden patients with psoriasis in Trang Province at 536 persons per year. Combination of reported and unreported cases to the true number of patients, an estimate of patients with psoriasis was 1,104 (95%: 804–1,045) people per year.

MSC: 62P10; 62P25
Keywords: capture-recapture; psoriasis; population size estimator

Submission date: 02.06.2023 / Acceptance date: 31.08.2023

1. Introduction

Psoriasis is considered one of the most common public health issues impacting physical and mental quality of life in many countries around the world. The disease shows as easily visible skin symptoms and negatively impacts personal confidence. Psoriasis afflicts 1-2% of the global population. The underlying cause of the disease has not yet been identified but heredity, the immune system and other internal and external factors have

*Corresponding author.

Published by The Mathematical Association of Thailand.

all been postulated. Patient statistics before the Covid-19 pandemic showed prevalence of 0.0911.4% [1], with disease incidence in Thailand increasing dramatically from 11,015 in 2016 to 16,518 cases by 2018 in the Hospital of Dermatology. The National Institute of Dermatology, Thailand estimated unreported cases and latent patients at around 1.34 million people, with the psoriasis situation now one of the main public health issues. The more humid southern part of the country where most of the population work in agriculture and farming has seen a rapid increase in the disease. Trang Province in the southern part of Thailand has recorded high patient numbers, with hospital medical records listing 469 cases in 2012 and 603 cases in 2017 [2].

Psoriasis is a chronic non-communicable disease that cannot be permanently cured, and patients can become repeatedly infected. Factors that trigger psoriasis are weather and seasonal changes, mental health condition, stress, drinking alcohol, eating certain foods, or a weak immune system. Disease treatment focuses on the reduction of severe infection of the skin. A comprehensive data analysis of the number of infected patients will give more precise statistical information and allow the government to better plan medical treatment and public health budgets and manpower to control the spread of the disease, while also predicting health issue trends in society.

Population size estimation techniques are many and varied, including Bootstrap, Jackknife and Network scale up. However, a powerful tool for estimating elusive target populations involves capture-recapture methods. This technique has been ordinarily used in biological and ecological fields to investigate population dynamics and estimate elusive population size [3], [4], [5]. In recent decades, capture-recapture techniques have been applied to other areas such as epidemiology and surveillance [6], [7], [8], social science [9], [10] and computer system [11], [12] to estimate target population sizes. In this research, the capture-recapture technique was modified to estimate the number of hidden psoriasis-infected patients in Trang Province. These patients were not identified because the symptoms did not clearly show or were very slight. Some patients were treated with alternative medication or at private hospitals. The true number of psoriasis patients in Trang Province is the total of the unobserved and observed counts. Accurate results would benefit national public health development and provide better care for Thai psoriasis patients. A basic model is often used in population studies such as a homogeneous Poisson process. However, in reality, sub-populations caused by covariates such as gender, age, location, size and latent variables render the results inaccurate. This situation involves a heterogeneous population because the capture probability takes on different values. Several estimators have been proposed to estimate population size for capture-recapture data. This research focused on estimators based on both homogeneous and heterogeneous Poisson models.

2. Methods

2.1. Data Sources and Characteristics

Estimation of patients with psoriasis in Trang Province used two data sources. The first was the Trang Provincial Public Health Office and the second was provided by the Southern Regional Hospital of Tropical Dermatology-Trang Province. To avoid a dramatic overestimation of population size due to no overlap between sources, the data were combined into one list. Data matching considered name, surname, gender and age. Identification numbers from the Thai ID card or birth certificate were used to eliminate repeated entries. Every patient who registers in a hospital receives a unique identification code by presenting his/her Thai National ID card or other identification issued by a government office. This code was used to enter patient information including past medical history and current conditions, laboratory tests and test results. Treatment episodes ranged from one day to six months.

2.2. Capture-recapture Modeling

The capture-recapture technique requires a series of repeated counts from a target population where each count reflects the number of times that a patient has been treated. For a general overview of capture-recapture data see [13]. Let $X_i, i = 1, 2, .., N$ denote the number of treatment episodes that occur during the study period for the N patients with psoriasis in Trang Province. If $X_i = 0$, no episodes are recorded in any of the public treatment service databases. Psoriasis patients are only observed when $X_i \geq 1$. Let $p_x = Pr(X_i = x)$, also let f_x denote the frequency of each patient registered exactly x times, $x = 0, 1, 2, 3, ..., m$ where m is the largest episode count. The total number of population size N can then be written as $N = f_0 + f_1 + f_2 + +... + f_m = f_0 + n$, where $n = \sum_{x=1}^{m} f_x$ is the total number of observed individuals. Since $X_i = 0$ is not observed, the corresponding f_0 is unknown and might be replaced by its expected value, Np_0, where p_0 is the probability of patients not recorded in the sample and has to be estimated. It can also write expected value of population size as $E(N) = Np_0 + N(1 - p_0)$, estimating $N(1-p)$ with n results in $\widehat{N} = \frac{n}{1-p_0}$. Since X_i carries only the non-negative integer values, the Poisson model with parameter λ, $\frac{\exp(\lambda)\lambda^x}{x!}$, may present a basic choice. Obviously, the Poisson model has constrain that the mean and variance are identically then it is rarely occurs in reality. An occurrence of heterogeneous population from covariates (i.e sex, education, location) or latent variables leads to a violation of the Poisson property, which are explained by overdispersion and underdispersion. A classical technique to account for the heterogeneity is a Poisson mixture model. The Poisson parameter is discussed as a latent random variable with arbitrary function $h(\lambda)$. The marginal distribution is given as

$$p_x = \int_{\infty}^{0} \frac{\exp(-\lambda)^x}{\lambda!} h(\lambda) dt, \tag{2.1}$$

where the mixing density $h(\lambda)$ is unknown. For identifying the basic model of count data, the ratio plot of Poisson distribution was used as the graphical approach to investigate an appropriate model. Plotting $(x+1)\frac{f_x+1}{f_x}$ against x can be used to consider the Poisson distribution and the zero-truncated Poisson distribution. If the ratio plot presents a pattern of horizontal lines, this can be taken indication of Poisson and zero-truncated Poisson models. Additionally, [14] introduced a mixed the ratio plot exhibits structured heterogeneity if the ratio plot shows the linear line with positive slope.

Several estimators have been applied to estimate population size using capture-recapture data. Well-known estimators based on homogeneous and heterogeneous Poisson models are described in the next section. The maximum likelihood estimator under a Poisson distribution (MLEPoi) was selected for use in homogeneous cases, while estimators for heterogeneous populations included MLEGeo based on geometric distributions, linear regression based on the Conway-Maxwell-Poisson distribution estimator (LCMP), the

Lanumteng and Bhning estimator (LB), Turings estimator based on the geometric distribution estimator (TG), Chaos lower bound estimator (Chao) and Zeltermans estimator (Zel).

2.2.1. Maximum Likelihood Estimator Based on Poisson Distribution Estimator

Supposed that the capture-recapture count X can be modeled as a Poisson distribution with density $p_x = \frac{\exp(-\lambda)\lambda^x}{x!}$ where λ is a Poisson parameter. The maximum likelihood estimation of zero-truncated counts data has been widely used as it generates small variance. Then, the zero-truncated Poisson distribution is defined as $p_0^+ = \frac{\exp(-\lambda)\lambda^x}{x!\{1-\exp(-\lambda)\}}$, where, $p_0 = \exp(-\lambda)$. Hence, the size of target population can be achieved by

$$\widehat{N}_{MLEPoi} = \frac{n}{1-\exp\left(-\hat{\lambda}_{MLEPoi}\right)}. \tag{2.2}$$

Where $\hat{\lambda}_{MLEPoi}$ is a parameter estimated from the zero-truncated Poisson distribution using the Expectation-Maximization Algorithm or the EM algorithm [15]. The variance of the population estimator under maximum likelihood method of estimation, $\widehat{Var}(\widehat{N}_{MLEPoi})$, can be derived as:

$$\widehat{Var}(\widehat{N}_{MLEPoi}) = \frac{\widehat{N}_{MLEPoi}}{\left\{\exp\left(\frac{\sum\limits_{x=1}^{m} xf_x}{\widehat{N}_{MLEPoi}}\right) - \frac{\sum\limits_{x=1}^{m} xf_x}{\widehat{N}_{MLEPoi}} - 1\right\}}. \tag{2.3}$$

As we mention above, the heterogeneity is often found in nature. The MLEPoi estimator might be not appropriate for capture-recapture data. Alternative estimators have been proposed for more realistic estimation in capture-recapture study.

2.2.2. Maximum Likelihood Based on the Geometric Distribution Estimator

The geometric distribution arises as $p_x = \int_0^\infty g(x;\lambda)h(\lambda;\theta)d\lambda$, where the mixture kernel is the Poisson distribution $g(x;\lambda) = \frac{\exp(-\lambda)\lambda^x}{x!}$, and the mixing density comes from the exponential density $h(\lambda;\theta) = \frac{1}{\theta}\exp(-\frac{\lambda}{\theta})$. Then, the associated marginal density is obtained as $p_x(p) = (1-p)^x p$, where $p = \frac{1}{1+\theta} \in (0,1)$ is the event parameter, and $x = 0, 1, 2, \ldots$. Mean and variance are $E(X) = \frac{1-p}{p}$ and $Var(X) = \frac{1-p}{p^2}$, respectively. Let f_x be the number of individuals identified exactly x time, and m denote the largest observed count. The total number of identifications is achieved from $0f_0 + 1f_1 + 2f_2 + \ldots + mf_m = \sum_{x=0}^m xf_x = S$. As in CR study the number of units identified zero times,f_0, is unknown. It can be written that $S = \sum_{x=0}^m xf_x = \sum_{x=1}^m xf_x$. The maximum likelihood based on the geometric distribution estimator for capture-recapture data was proposed by [16] as

$$\widehat{N}_{MLEGeo} = \frac{n}{1-\widehat{p}_0} = \frac{n}{1-\frac{n}{S}} = \frac{nS}{S-n}, \tag{2.4}$$

where $\widehat{p}_0$ is estimated by using the maximum likelihood approach based on the zero-truncated geometric distribution.
The variance estimation of the MLEGeo estimator,$\widehat{Var}(\widehat{N}_{MLEGeo})$, was given as

$$\widehat{Var}(\widehat{N}_{MLEGeo}) = \frac{s^2n^2}{(S-n)^3}. \tag{2.5}$$

The MLEGeo estimator often provides a small variance for the true model as an underestimate for contaminated evidence of geometric distribution.

2.2.3. The Turing Estimator Based on the Geometric Distribution

One of the extended the Turing estimator was derived under the geometric distribution by [17]. We have $p_0 = p, p_1 = (1-p)p$ and $E(X) = \frac{1-p}{p}$, therefore, $p_0 = p = \sqrt{p^2} = \sqrt{\frac{(1-p)p^2}{(1-p)}} = \sqrt{\frac{p(1-p)}{(1-p)/p}} = \sqrt{\frac{p_1}{E(X)}}$. In practice, the populations can be estimated by the relative frequencies so that the unobserved probability $\widehat{p}_0$ is achieved by $\widehat{p}_0 = \sqrt{\frac{f_1/N}{S/N}} = \sqrt{\frac{f_1}{S}}$, when $S = \sum_{x=0}^{m} xf_x = \sum_{x=1}^{m} xf_x$. Hence, the extension of Turing's estimator for the geometric distribution (TG) is given as

$$\widehat{N}_{TG} = \frac{n}{1-\sqrt{\frac{f_1}{S}}}. \tag{2.6}$$

The variance of TG estimator, $\widehat{Var}(\widehat{N}_{TG})$, is given as

$$\widehat{Var}(\widehat{N}_{TG}) = \frac{n\sqrt{\frac{f_1}{S}}}{(1-\sqrt{\frac{f_1}{S}})^2} + n^2\left\{\frac{S+f_1}{4S^2\left(1-\sqrt{\frac{f_1}{S}}\right)^4}\right\}. \tag{2.7}$$

The positive points of TG estimator are that it modifies all information of observed counted data (S) in a model, therefor, it seems to be more natural than some estimators. Also, the TG estimator is a straightforward approach to get estimated population size.

2.2.4. The Linear Regression Estimator Based on Conway Maxwell-Poisson Distribution

The Conway Maxwell-Poisson distribution, $CMP(\lambda,\nu)$, is an extension of the Poisson distribution. It generalizes the Poisson distribution by adding an extra parameter ν, which accounts for the cases of over and under-dispersion. The CMP with two parameters has probability mass function as $p_x = \frac{\lambda^x}{(x!)^\nu}\frac{1}{z(\lambda,\nu)}, \lambda > 0, \nu \geq 0$, where $x = 0,1,2,3,...$, and the function $z(\lambda,\nu) = \sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^\nu}$ denotes a normalization constant. A Linear regression estimator based on Conway Maxwell-Poisson distribution was proposed by [18], [19]. An interesting point of this estimator is that it uses the ratio plot of Conway-Maxwell-Poisson distribution with the regression technique to estimate two parameters. Let $(x+1)\frac{p_{x+1}}{p_x} = (x+1)\frac{\frac{\lambda^{x+1}}{(x+1!)^\nu}\frac{1}{z(\lambda,\nu)}}{\frac{\lambda^x}{(x!)^\nu}\frac{1}{z(\lambda,\nu)}} = \frac{\lambda(x+1)!}{(x+1)^\nu}$. Taking logarithm transformation of both sides, it becomes a linear model as $\log\left\{(x+1)\frac{p_{x+1}}{p_x}\right\} = \log\lambda + (1-\nu)\log(x+1)$, where a intercept parameter $\beta_0 = \log\lambda$ and a slop parameter $\beta_1 = 1-\nu$. The LCMP estimator is given as

$$\widehat{N}_{LCMP} \quad = \quad n + \widehat{f_0} = n + f_1\exp(-\widehat{\beta_0}), \tag{2.8}$$

where $\widehat{\beta_0}$ is the intercept point, achieved by plotting the weighted lest square regression $\log\left\{(x+1)\frac{f_{x+1}}{f_x}\right\}$ against $\log(x+1)$, that is $\log\left\{(x+1)\frac{f_{x+1}}{f_x}\right\} = \widehat{\beta_0} + \widehat{\beta_1}\log(x+1)$. A variance estimation of LCMP , $\widehat{Var}(\widehat{N}_{LCMP})$, is given as:

$$\widehat{Var}(\widehat{N}_{LCMP}) = \frac{nf_1e^{-\widehat{\beta}^0}}{n+f_1e^{-\widehat{\beta}^0}} + (e^{-\widehat{\beta}_0})^2 f_1[1+f_1Var(\widehat{\beta}_0)]. \tag{2.9}$$

The benefit of this estimator is that it can use for estimating population size under the Poisson, the geometric distributions.

2.2.5. Lanumteng and Böhning Estimator

Lanumteng and Böhning [20] introduced an alternative estimator of population based on the negative binomial distribution. Suppose that, $g(x) = Gam(\lambda;\theta,k) = \frac{\theta^{-k}\lambda^{k-1}exp(\frac{-\lambda}{\theta})}{\Gamma(k))}$ with parameters θ and $k > 0$. The mixed Poisson-Gamma with shape parameter k and scale parameter θ can be rewritten as $p_x = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}p^k(1-p)^x$, indicating the probability of the negative binomial with shape parameter k and scale parameter $\theta = \frac{1-p}{p}$. The ratio of neighboring negative binomial probability is given as: $\log\frac{xp_x}{p_{x-1}} = \log(k-x-1) + \log(1-p) \approx \log(1-p) + \log(k-1) + \frac{1}{k-1}x$. This pattern indicates the linear regression with a intercept $\beta_0 = \log(1-p) + \log(k-1)$ and a slope $\beta_1 = \frac{1}{k-1}$. Then, for $x=1$, $x=2$ and $x=3$, the ratios are given as $\log\left(\frac{f_1}{f_0}\right) = \beta_0 + \beta_1$, $\log(\frac{2f_2}{f_1}) = \log\frac{2f_2}{f_1} = \beta_0 + 2\beta_1$ and $\log\left(\frac{3f_3}{f_2}\right) = \beta_0 + 3\beta_1$, respectively. Solving these equations, hence, the LB population size estimator for capture-recapture ultimately provided as

$$\widehat{N}_{LB} = n + \frac{3f_1^3f_3}{4f_2^3}. \tag{2.10}$$

An estimation of the variance of the LB estimator, $\widehat{Var}(\widehat{N}_{LB})$, is given by

$$\widehat{Var}(\widehat{N}_{LB}) = \left(\frac{9}{4}\right)^2\left\{\frac{f_1^5f_3^2}{f_2^6}\right\}\left\{\frac{f_1}{f_2}+1\right\} + \left(\frac{3}{4}\right)^2\left\{\frac{\{f_1^6f_3\}}{f_2^6}\right\}\left\{1-\frac{f_3}{n}\right\} + \frac{\frac{3n}{4}f_1^3f_3}{nf_2^3+\frac{3}{4}f_1^3f_3}. \tag{2.11}$$

2.2.6. Chao's Estimator

The Chao estimator was introduced by [21], [22], a lower bound for estimating the population size. Chao's lower bound estimator of the population size N is

$$\widehat{N}_{Chao} = n + \frac{f_1^2}{2f_2}. \tag{2.12}$$

Note that only f_1 and f_2 are used in Chao's lower bound estimator. Chao's estimator probably represents lower bound estimates. The Variance of Chao's estimator , $\widehat{Var}(\widehat{N}_{Caho})$, is given as

$$\widehat{Var}(\widehat{N}_{Chao}) = \left(\frac{1}{4}\right)^2 \frac{f_1^4}{f_2^3} + \frac{f_1^3}{f_2^2} + \left(\frac{1}{2}\right)\frac{f_1^2}{f_2}, \tag{2.13}$$

see [23] and [24] for more detail. Extended versions of Chao's estimator has been recently developed based on the log-normal distribution in [25].

2.2.7. The Zelterman Estimator

The Zelterman estimator was modified by using the truncated Poisson distribution. The highlight of this estimator is well-known as a robust estimator since the first and the second of frequencies are used in the model [26]. However, it has the limitation of a long tail count data set [20]. The Zelterman estimator to estimate population size is provided as

$$\widehat{N}_{Zel} = \frac{n}{1 - \exp\left(-\frac{2f_2}{f_1}\right)}. \tag{2.14}$$

The Variance of the Zelterman's Estimator,$\widehat{Var}(\widehat{N}_{Zel})$, was given as

$$\widehat{Var}(\widehat{N}_{Zel}) = nG(\widehat{\lambda})\left[1 + G(\widehat{\lambda})\widehat{\lambda}^2\left(\frac{1}{f_1 + f_2}\right)\right], \tag{2.15}$$

where $G(\widehat{\lambda}) = \frac{\exp(-\widehat{\lambda})}{\{1-\exp(-\widehat{\lambda})\}^2}$ and $\widehat{\lambda} = \frac{2f_2}{f_1}$. More details can be found in [23] and [24] .

A confidence interval of population size estimation can be constructed by using the normal approximation as $\widehat{N} \pm z_{0.975}\widehat{S.E}(\widehat{N})$, where $\widehat{S.E}(\widehat{N}) = \sqrt{\widehat{Var}(\widehat{N})}$ for all estimators.

2.3. Model Evaluation

Model selection criteria comprise rules used to select the best statistical model among a set of candidate models. Let $\hat{f}_1 = \hat{N}\hat{p}_1$, $\hat{f}_2 = \hat{N}\hat{p}_2$,$\hat{f}_3 = \hat{N}\hat{p}_3$,..., $\hat{f}_m = \hat{N}\hat{p}_m$ denote the fitted frequencies of psoriasis patient with $1, 2, 3, ...m$ treatment episodes and probabilities $\hat{p}_1, \hat{p}_2, \hat{p}_3, ..., \hat{p}_m$. These values can be estimated under the studied model. The frequencies fitted under the model can be compared with the observed, presented by a graphical evaluation fitted. Additionally, the root mean square error ($RMSE$) is computed under each model as:

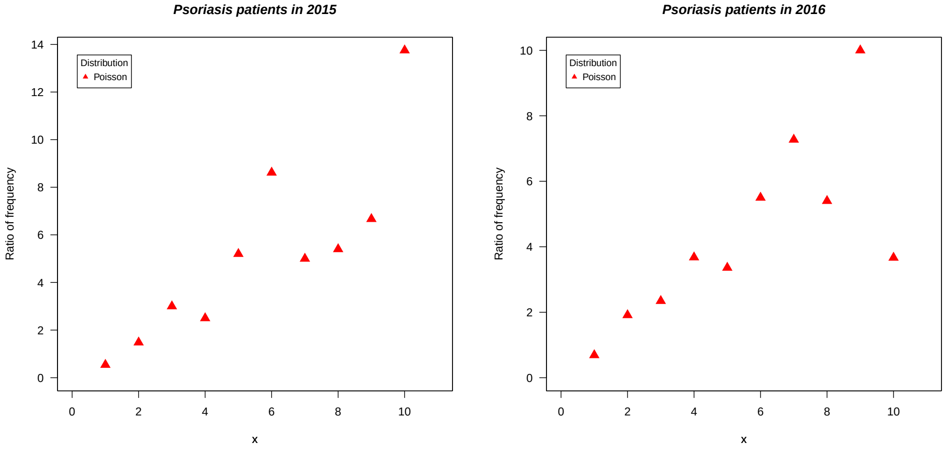
$$RMSE = \sqrt{\frac{1}{m}\sum_{x=1}^{m}\left(f_x - \hat{f}_x\right)^2}. \tag{2.16}$$

The $RMSE$ provides the overall quality of estimation, lower values of values indicate better fit [27].

3. Results

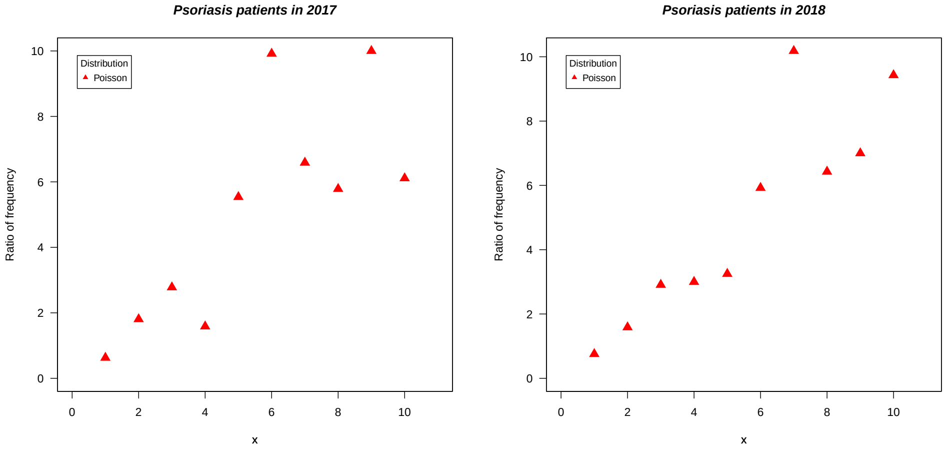
Different estimators were applied to psoriasis data sets in Trang Province between 2015 and 2018, and the ratio plot of the Poisson distribution was used to investigate the basic model. According to Figure 1 and Figure 2. It is clear that all ratio plots display linear trends with positive slopes. Thus, it is reasonable to assume that a heterogeneous model would be appropriate to estimate population sizes. Seven estimators were presented in

Table 1 and Table 2 to determine the population size of psoriasis patients. The MLEPoi population size estimator gave the lowest number of psoriasis patients and the lowest standard errors, resulting in a narrow length of 95% confidence intervals in Trang Province every year. These results were not surprising because the MLEPoi always shows an underestimate for a heterogeneous population [17]. Population size estimators based on a homogeneous Poisson-based model should, therefore, be avoided. Alternative estimators based on the parametric models were selected to deal with this problem such as the LCMP, MLEGeo and LB estimators. The non-parametric Chao lower bound, TG and Zelterman estimators were also studied as interesting choices.



(A) (B)

Figure 1. Ratio plot of Poisson distribution in 2015 and 2016



(C) (D)

Figure 2. Ratio plot of Poisson distribution in 2017 and 2018

Table 1. Observed and estimated capture frequencies

	Model	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	$f_{\geq 12}$	$RMSE$
2015	Observed	299	81	40	30	15	13	16	10	6	4	5	31	
$\hat{\lambda} = 1.84$	fitted (Poi)	190	174	107	49	18	6	1	0	0	0	0	-	60.55
$\hat{\nu} = 0, \hat{\lambda} = 0.44$	fitted (CMP)	304	135	60	27	12	5	2	1	0	0	0	-	29.31
$\hat{p} = 0.46$	fitted (Geo)	249	135	74	40	22	12	6	3	2	1	1	-	20.37
$\hat{k} = 1.99, \hat{p} = 0.007$	fitted (NB)	0	0	0	0	0	0	0	0	0	0	0	-	-
$\hat{\beta}_0 = -0.013, \hat{\beta}_1 = 1.00$														
2016	Observed	264	91	58	34	25	14	11	10	6	6	2	25	
$\hat{\lambda} = 1.99$	fitted (Poi)	188	174	107	49	18	6	1	0	0	0	0	-	48.5
$\hat{\nu} = 0, \hat{\lambda} = 0.50$	fitted (CMP)	270	136	69	35	17	9	4	2	1	1	0	-	15.61
$\hat{p} = 0.43$	fitted (Geo)	236	134	76	43	24	14	8	4	3	1	1	-	19.63
$\hat{k} = 1.98, \hat{p} = 0.008$	fitted (NB)	6	8	10	12	13	14	15	15	16	16	16	-	97.74
$\hat{\beta}_0 = -0.28, \hat{\beta}_1 = 1.02$														
2017	Observed	312	98	59	41	13	12	17	14	9	9	5	14	
$\hat{\lambda} = 2.1$	fitted (Poi)	177	186	130	68	29	10	3	1	0	0	0	-	63.60
$\hat{\nu} = 0, \hat{\lambda} = 0.49$	fitted (CMP)	309	151	73	36	17	9	4	2	1	0	0	-	19.55
$\hat{p} = 0.42$	fitted (Geo)	252	147	85	50	29	17	10	6	3	2	1	-	29.90
$\hat{k} = 1.95, \hat{p} = 0.006$	fitted (NB)	5	6	8	9	10	11	12	13	13	13	14	-	53.09
$\hat{\beta}_0 = -0.26, \hat{\beta}_1 = 1.06$														
2018	Observed	275	104	55	40	24	13	11	14	10	7	6	15	
$\hat{\lambda} = 2.2$	fitted (Poi)	157	173	127	70	31	11	4	1	0	0	0	-	55.95
$\hat{\nu} = 0, \hat{\lambda} = 0.52$	fitted (CMP)	276	143	74	39	20	10	5	3	1	1	0	-	14.07
$\hat{p} = 0.42$	fitted (Geo)	244	140	81	46	27	15	9	5	3	2	1	-	19.58
$\hat{k} = 2.35, \hat{p} = 0.11$	fitted (NB)	1	1	2	2	3	3	4	4	5	5	6	-	106.44
$\hat{\beta}_0 = 0.18, \hat{\beta}_1 = 0.74$														

Table 2. Estimation of unobserved and total cases of patients with psoriasis in Trang Province

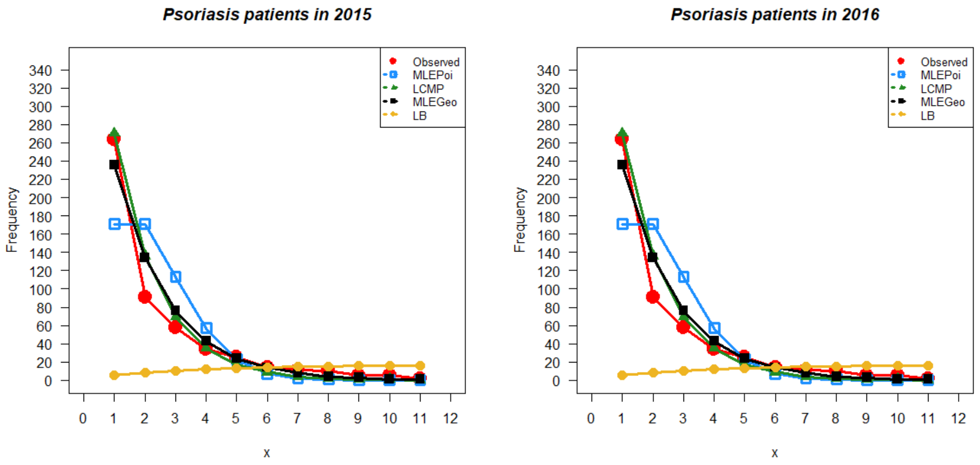
Year	estimator	unobserved	total	total/observed	$\widehat{S.E}(\widehat{N})$	95% CI
2015	MLEPoi	103	649	1.18	13.69	622–676
$n = 545$	LCMP	674	1,220	2.23	290.31	652–1,789
	MLEGeo	459	1,005	1.84	39.46	928–1,082
	LB	1,509	2,055	3.76	594.61	890–3220
	Chao	552	1,098	2.01	91.57	919–1,277
	TG	549	1,098	2.00	33.14	1,027–1,157
	Zel	760	1,306	2.39	130.35	1,051–1,562
2016	MLEPoi	86	631	1.16	12.05	607–655
$n = 546$	LCMP	524	1, 064	1.94	165.14	746–1,393
	MLEGeo	417	962	1.77	36.01	891–1,033
	LB	1,062	1,607	2.95	401	821–2,393
	Chao	383	928	1.70	64.93	801–1,055
	TG	461	1,006	1.85	29.26	949–1,063
	Zel	550	1,095	2.01	98.17	903-1,288
2017	MLEPoi	85	688	1.14	11.68	665–711
$n = 603$	LCMP	640	1,243	2.06	254.18	745–1,742
	MLEGeo	433	1,036	1.71	35.74	966–1,106
	LB	1,428	2,031	3.37	515.267	1,021–3,041
	Chao	497	1,100	1.82	78.588	946–1,254
	TG	524	1,127	1.87	31.40	1,065–1,189
	Zel	690	1,293	2.144	114.23	1,070–1,517
2018	MLEPoi	72	646	1.13	10.53	625–667
$n = 574$	LCMP	530	1,104	1.92	153.21	804–1,405
	MLEGeo	445	1,048	1.83	36.63	976–1,1202
	LB	763	1,337	2.33	274.44	799–1,875
	Chao	364	938	1.63	59.64	821–1,055
	TG	451	1,025	1.79	28.49	969–1,081
	Zel	508	1,082	1.82	88.87	908–1,257

Figure 3 and Figure 4 compared estimated frequencies under the parametric based estimators. The MLEGeo estimator showed the best accuracy with the lowest number of $RMSE$ in 2015. Since the MLEGeo estimator was constructed based on the geometric distribution, it might use the TG estimator as an alternative. The parameters of the zero-truncated Conway-Maxwell-Poisson model are $\hat{\lambda} = 0.44$, and $\hat{\nu} = 0$ i.e. the geometric distribution was obtained as a special case of Conway-Maxwell-Poisson distribution. Then, the LCMP estimator was used for estimating the number of psoriasis patients in 2015. It is even more clear from the graph that the truncated Poisson and the truncated negative binomial distributions were not suitable for this data set. To estimate the number of psoriasis patients in Trang Province in 2016, 2017 and 2018, the LCMP estimator showed the best accuracy since they provide the lowest $RMSE$. The graphs seem to fit well and provide estimated values of frequencies in line with observed values. As the dispersion parameters of the zero-truncated Conway-Maxwell-Poisson distribution, $\hat{\nu} = 0$, the MLEGeo and the TG estimators based on the zero-truncated geometric model can be used to estimate population size. However, they often give a lower value of population size than the LCMP [17]. The MLEPoi estimator showed a very low estimate for $\hat{N}$ with a small standard error in all data sets, resulting in the corresponding 95% confidence intervals did not overlap with those selected estimators.

The use of the best model of capture-recapture approaches to estimate the total of patients with psoriasis in Trang Province show in Table 3. The number of the psoriasis patients in Trang Province in 2015 is provided by the MLEGeo estimator as 1,005 with 95% CI (928 – 1,082). Also, the LCMP estimator give the number of patients in 2016, 2017 and 2018 as 1,064 (95% CI: 746–1,393), 1,243 (95% CI: 745–1,742) and 1,104 (95% CI: 804–1,405) persons, respectively. Overall, the average of hidden of psoriasis patients in Trang Province was 536 persons per yaer, and the true number of psoriasis patients was approximately 1,104 (95% CI: 806–1,405). The ratio of total estimated case to the observed cases is interesting, the average ratio of 1.94 would mean that for every a hundred treatment patients there are ninety-four patients untreated. The reason for the unseen cases might be mild signs of psoriasis.

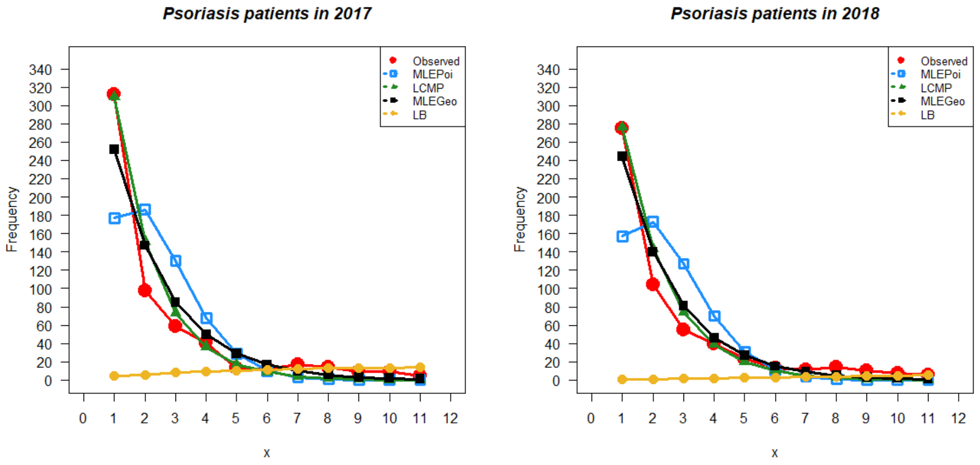
Table 3. Estimation of unobserved and total of psoriasis patients in Trang Province between 2015 and 2018

Year	estimator	unobserved	total	total/observed	$\widehat{S.E}(\hat{N})$	95% CI
2015	MLEGeo	459	1,005	1.84	39.46	928–1,082
2016	LCMP	524	1,064	1.94	165.14	746–1,393
2017	LCMP	640	1,243	2.06	254.18	745–1,742
2018	LCMP	530	1,104	1.92	153.21	804–1,405
Average		536	1,104	1.94		806–1,405



(E) (F)

FIGURE 3. Observed frequencies with fitted frequencies based on the zero-truncated Poisson, the zero-truncted geometric, the zero-truncated Conway-Maxwell-Poisson and the zero-truncated negative binomial distributions in 2015 and 2016



(G) (H)

FIGURE 4. Observed frequencies with fitted frequencies based on the zero-truncated Poisson, the zero-truncated geometric, the zero-truncated Conway-Maxwell-Poisson and the zero-truncated negative binomial distributions in 2017 and 2018

4. Conclusion

A variety of estimators in the capture-recapture field have been proposed and applied in diverse areas of interest to estimate elusive target population sizes. This paper compared the capabilities of seven estimators to estimate the number of patients with psoriasis in Trang Province between 2015 and 2018. The Poisson ratio plot was used as the graphical device to investigate a suitable model. Results suggested that all data sets displayed the occurrence of a heterogeneous population. The LCMP estimator under the Conway-Maxwell-Poisson distribution and the MLEGeo estimator showed optimal accuracy, while the LCMP estimator best captured different levels of heterogeneity flexibly. The Conway-Maxwell-Poisson distribution showed two parameters as a generalized form of the Poisson distribution that included the geometric distribution as a sub-model when $\nu = 0, 0 < \lambda < 1$ and a Poisson distribution when $\nu = 1$. The LCMP estimator was selected to estimate the number of patients with psoriasis in Trang Province in 2016, 2017 and 2018, while the MLEGeo showed excellent estimation in 2015.

Research Ethics

Ethical approval was obtained from Thaksin University Research Ethics Committee (Reference Number: 0106/66).

Acknowledgements

Authors would like to thank the reviewers for the very helpful comments which lead to considerable improvement of the paper as well as to interesting and promising modifications of the proposed estimator. Authors are grateful to Thaksin University for providing facilities support our research.

References

[1] Who, 2016 [Online]: Avalible at: https://www.who.int/publications/i/item/9789241565189 (Jan 5, 2023).

[2] Southern Regional Hospital of Tropical Dermatology Trang Province [Online]: Avalible at: https://sis.trangskin.go.th/ (April 26, 2022).

[3] K.H. Pollock, Capture-recapture design robust to unequal probability of capture, The Journal of Wildlife Management 46 (3) (1982) 752–757.

[4] D. Zelterman, Robust estimation in truncated discrete distributions with application to capture-recapture experiments, Journal of Statistical Planning and Inference 18 (2) (1988) 225–237.

[5] A. Chao, Estimating the population size for capture-recapture data with unequal catchability, Biometrics (1987) 783–791.

[6] A.D. Kiakalayeh, M.R. Taramsari, R. Mohammadi, S.D. Kiakalayeh, H. Kavakpour, Comparison of the capture-recapture method and seroprevalence survey for estimation of COVID-19 prevalence in the Islamic Republic of Iran, Eastern Mediterranean Health Journal 29 (2) (2023) 126–131.

[7] D. Böhning, I. Rocchetti, A. Maruotti, H. Holling, Estimating the undetected infections in the Covid-19 outbreak by harnessing capture–recapture methods, International Journal of Infectious Diseases 97 (2020) 197–201.

[8] K. Wang, L. Ding, Y. Yan, C. Dai, M. Qu, D. Jiayi, X. Hao, Modelling the initial epidemic trends of COVID-19 in Italy, Spain, Germany, and France, PloS One 15 (11) (2020) e0241743.

[9] M. Mwale, K. Mwangilwa, E. Kakoma, K. Iaych, Estimation of the completeness of road traffic mortality data in Zambia using a three source capture recapture method, Accident Analysis & Prevention 186 (2023) 107048.

[10] A.M. Coumans, M.J.L.F. Cruyff, P.G.Van der Heijden, J.R.L.M. Wolf, H.J.S.I.R. Schmeets, Estimating homelessness in the Netherlands using a capture-recapture approach, Social Indicators Research 130 (2017) 189–212.

[11] N. Accettura, G. Neglia, L.A. Grieco, The Capture-Recapture approach for population estimation in computer networks, Computer Networks 89 (2015) 107–122.

[12] S. Zander, L.L. Andrew, G. Armitage, G. Huston, Estimating IPv4 address space usage with capture-recapture, In 38th Annual IEEE Conference on Local Computer Networks-Workshops, IEEE (2013) 1010–1017.

[13] S.C. Amstrup, T.L. McDonald, B.F. Manly (Eds), Handbook of Capture-Recapture Analysis, Princeton University Press, 2005.

[14] D. Bhning, M.F Baksh, R. Leardsuwansri, J. Gallagher, Use of the ratio plot in capture–recapture estimation, Journal of Computational and Graphical Statistics 22 (1) (2013) 135–155.

[15] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society: Series B (Methodological) 39 (1) (1977) 1–22.

[16] S.A. Niwitpong, D. Böhning, P.G. van der Heijden, H. Holling, Capture–recapture estimation based upon the geometric distribution allowing for heterogeneity, Metrika 76 (2013) 495–519.

[17] O. Anan, D. Böhning, A. Maruotti, On the Turing estimator in capture–recapture count data under the geometric distribution, Metrika 82 (2019) 149–172.

[18] O. Anan, Capture-Recapture Modelling for Zero-Truncated Count Data Allowing for Heterogeneity, Doctoral Dissertation, University of Southampton, 2016.

[19] O. Anan, D. Böhning, A. Maruotti, Population size estimation and heterogeneity in capture–recapture data: A linear regression estimator based on the Conway–Maxwell–Poisson distribution, Statistical Methods & Applications 26 (2017) 49–79.

[20] K. Lanumteang, D. Böhning, An extension of Chao's estimator of population size based on the first three capture frequency counts, Computational Statistics & Data Analysis 55 (7) (2011) 2302–2311.

[21] A. Chao, Estimating the population size for capture-recapture data with unequal catchability, Biometrics (1987) 783–791.

[22] A. Chao, Estimating population size for sparse data in capture-recapture experiments, Biometrics (1989) 427–438.

[23] D. Böhning, A simple variance formula for population size estimators by conditioning, Statistical Methodology 5 (5) (2008) 410–423.

[24] D. Böhning, Some general comparative points on Chao's and Zelterman's estimators of the population size, Scandinavian Journal of Statistics 37 (2) (2010) 221–236.

[25] C.H. Chiu, Y.T. Wang, A. Bruno, B.A Walther, A. Chao, An improved nonparametric lower bound of species richness via a modified good–turing frequency formula, Biometrics 70 (3) (2014) 671–682.

[26] D. Zelterman, Robust estimation in truncated discrete distributions with application to capture-recapture experiments, Journal of Statistical Planning and Inference 18 (2) (1988) 225–237.

[27] C.A. Mushagalusa, A.B. Fandohan, R. Glèlè Kakaï, Random forests in count data modelling: An analysis of the influence of data features and overdispersion on regression performance, Journal of Probability and Statistics 2022 (2022) Article ID 2833537.