University of Cape Town

Capturing Transients: An Application of Biostatistics to Astronomy

Author:
Anke van Dyk

Supervisors:
Assoc. Prof. Vanessa McBride
Prof. Paul Groot

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science

in the

Department of Astronomy

November 24, 2021

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Plagiarism Declaration

I, Anke VAN DYK, declare that this thesis titled, 'Capturing Transients:
An Application of Biostatistics to Astronomy' and the work presented in it are my own for the purpose of fulfilling the requirements of the degree of Master of Science whilst in candidature at the University of Cape Town. I know and understand that plagiarism is wrong. I have clearly attributed the work of others where relevant. I have acknowledged the persons and main sources of help.

Signed:

Signed by candidate

Date: 15/12/2021

Acknowledgements

This has been an especially challenging year. There are many things to be grateful for, one of them being the completion of my Master's degree. This is a privilege, no doubt, and it was only due to my immense support system that I could successfully see this to fruition. I would like to start by thanking both of my supervisors:

Vanessa – I thank you for your watchful eye, understanding and words of encouragement throughout this strange time. Your leadership and care was crucial and is deeply appreciated. I look forward to our next collaboration on the PhD in the years to come.

Paul – you continue to challenge me and make me question all the steps along the way, as a seasoned academic should! Thank you for taking me under your wing and directing me through the research process.

Many thanks to the staff at SAAO, particularly *Patricia* and *Valencia*, for overseeing our well-being and the Astronomy department for seeing us through a year of physical distancing. To all at BANG, thank you for your lively discussions and feedback when I presented my work.

To *Dayne* – thank you for your endurance in being married to a hopeful and young academic and seeing me through the ups and downs that accompany that. You keep on supporting me through my own fears and encourage me to tackle my dreams one step at a time.

To my dear mother and father, *Susan* and *Arnold*, who unwittingly keep singing my praises, even though you don't understand what it is that I do – you project love and share in my achievement and allow me to carve my own path. Dankie Moeder vir al die kos, skottelgoed was agter my, en gesels vir die afgelope maande; dit beteken onbeskryflik baie. Aan my Pa, ek blameer jou vir my akademiese entoesiasme, dankie vir jou ewigdurige ondersteuning op alle gebiede. To the rest of my blood and non-blood family: our zoom sessions have been so important to me during the pandemic isolation. That and the pictures and videos of the kids always brightened my day. I miss you all terribly.

Aan *Japie*, vir die daaglikse bederf en humor sê ek groot dankie. Jy was 'n ster die hele pad deur. Ek hoop jy weet dat jou goeie dade nie vergeet is nie.

To my wonderful friends who walked the postgraduate trail with me: *Shilpa*, *Munira*, and *Danté*, thanks for keeping me sane with virtual tea times; it has been my rock and kept me going.

I feel compelled to also mention *Dino*, who has been my constant companion throughout this thesis and routinely summoned me to take breaks on account of needing to go for a walk. You have been both a nuisance and my friend when I needed one.

Lastly, I would like to extend my gratitude to NASSP and NRF SARChI for their funding of this degree, which would otherwise not have been possible.

Abstract

Capture-recapture has been identified as a possible use case for estimating the underlying size of astrophysical transient populations. In this work, we present a series of exploratory analyses using capture-recapture methods from biostatistics.

In the first of three separate analyses, we reproduce results of Laycock (2017). Strategically sampled X-ray lightcurves of simulated populations of high mass X-ray binaries (HMXBs) are used to probe estimator behaviour and efficiency. Overall, these statistically closed population estimators converge to the input population with increasing number of observations, yet estimator efficiency is shown to be significantly be affected by sampling strategy. I then employ non-standard estimator models to account for variations in capture probability of individuals within the population, categorised into 'behavioural', 'temporal', and 'heterogeneous' effects.

In the second analysis, we present a methodology for closed population capture-recapture analysis using real data from the OGLE-IV XROM survey. The data samples consisted of observations that were grouped into epochs. The large variation in quiescent magnitude of the population creates heterogeneity in the capture probability of sources which requires non-standard modelling. Estimation of population size is therefore limited by the choice of observational magnitude threshold. Bias corrected estimation proves to be potentially useful in this context.

In the third and final investigation, we present a 'robust design' approach with a population of Dwarf Nova located towards and in the Galactic Bulge identified from the OGLE-II, -III, and -IV phases. This approach combines closed and open population practices that allows new individuals identified between the survey phases to be added to the study sample for dynamical estimation.

These investigations provide a future course for population size estimation of transients and variable stellar population alongside population synthesis simulations. The generation of capture histories remain non-trivial through the choice of observation grouping, brightness scale, and imposed flux threshold. Further, there remain several unexplored avenues of inquiry and refinement for this methodology pertaining to astronomy using explanatory variables in the modelling. Recommendations are made for further exploration of the topic.

Contents

List of Figures

List of Tables

List of Abbreviations

AGN	Active Galactic Nuclei
AIC	Akaike Information Criterion
BeXRB	Be/X-ray binary
BIC	Bayesian Information Criterion
BH	black hole
CJS	Cormack-Jolly-Seber
CV	Cataclysmic Variable
CRTS	Catalina Real-time Transient Survey
DN	Dwarf Nova
dof	degrees of freedom
GUI	Graphical User Interface
JD	Julian Date
JS	Jolly-Seber
HJD	Heliocentric Julian Date
HMXB	high mass X-ray binary
i.i.d.	independent and identically distributed
IMXB	intermediate mass X-ray binary
LMC	Large Magellanic Cloud
LMXB	low mass X-ray binary
LP	Lincoln-Peterson
LSE	least squares estimate
LSST	Legacy Survey of Time and Space
LLSR	linear least-squares regression

MAGHMXBCAT	Magellanic Clouds High Mass X-ray Binaries Catalog
M_0	homogeneous capture probability (reference) model
M_{bh}	behavioural and heterogeneous capture probability model
M_b	behavioural capture probability model
MCR	multiple capture-recapture
MCMC	Markov Chain Monte Carlo
M_h	heterogeneous capture probability model
ML	maximum likelihood
MLE	maximum likelihood estimation
M_{tbh}	temporal, behavioural, and heterogeneous capture probability model
M_{tb}	temporal and behavioural capture probability model
M_{th}	temporal and heterogeneous capture probability model
M_t	temporal capture probability model
NS	neutron star
OGLE	Optical and Gravitational Lensing Experiment
PTF	Palomar Transient Factory
SCR	spatial capture-recapture
SGXB	supergiant high-mass X-ray binary
SMC	Small Magellanic Cloud
SN	supernova
RSS	residual sum of squares
XRB	X-ray binary
XROM	X-ray variables OGLE Monitoring
ZTF	Zwicky Transient Facility

Chapter 1

Introduction

This project was inspired by the shift in the study of astronomical transient sources from a case-by-case approach to the characterisation of entire populations of transient phenomena. An astronomical transient source has become the misnomer for an astronomical transient *event*, the period when an astronomical object displays detectable variability in its brightness as a function of time. The variability timescale for such objects is short, in relative terms, compared to the evolutionary timescale of astronomical sources, which is typically millions or billions of years. Transient events occur on, broadly ranged, a timescale of milliseconds to decades. Perhaps the most familiar transient event is the *supernova*, the explosive death of a star that emits large amounts of radiation energy across the electromagnetic spectrum, with luminosities on the order of $L \sim 10^{42}$ - 10^{43} erg s^{-1} (Sukhbold et al., 2016). Within our definition, a transient event includes gamma-ray bursts, flare stars, as well as stellar transits and eclipses. Transits and eclipses are not strictly referred to as transients by astronomers, but these time-domain phenomena are applicable under the methodology that we will investigate in this thesis and hence we include them. This examples mentioned is also not exhaustive – the definition of a transient will deliberately be kept general and loose to include sources and events as wildly different as variable stars and gravitational wave mergers.

Astronomy is experiencing a drive for dedicated time-domain surveys covering large amounts of the sky to discover new transients. The drive stems from an interest in the exotic physics that these sources tend to probe, such as the ultra high energy gamma-ray burst phenomena that last for seconds or less but can be seen at other wavelengths for several months. Other transients such as flare star phenomena present sudden, haphazard increases in their brightness by ~ 5 magnitude ($\sim 10^{32}$ erg) in tens of minutes (Fletcher et al., 2011). A proportion of binary stars show accretion during

late-stage evolution, which may similarly give rise to recurrent outbursting phenomena that emit across the electromagnetic spectrum (Eggleton and Pringle, 1985; Eggleton, 2006).

Binary star transients and their evolution have been a hot topic of discussion amongst stellar astronomers since the time it has been shown that $\sim 60 - 70\%$ of stars in our solar neighbourhood are gravitationally bound to a companion in a binary or multiple stellar system (Abt, 1983). Therefore, each star's evolution is not isolated but rather affected by the other (Eggleton and Pringle, 1985; Eggleton, 2006).

The number of transient event alerts identified in past and current programs such as the Palomar Transient Factory (PTF) and the Zwicky Transient Facility (ZTF) demonstrate an eventful and variable sky at up to ~ 600 detections/ deg^2 /night (Law et al., 2009; Bellm, 2014; Bellm et al., 2019; Graham et al., 2019; Patterson et al., 2018). The upcoming Rubin Observatory Legacy Survey of Time and Space (LSST) (Abell et al., 2009; Ivezić et al., 2019) is the next-generation wide-field optical survey telescope and will deliver an unprecedented amount of transient alerts. Current and other future projects include the Catalina Real-time Transient Survey (CRTS) (Drake et al., 2009), Pan-STARRS (Kaiser et al., 2002), SkyMapper (Keller et al., 2007), and MeerLICHT and BlackGEM (Bloemen et al., 2016). Large surveys such as these are bound to increase population sample sizes of a host of different transients. This allows for, and requires, better statistical analysis, study and characterisation of populations as a whole.

We will attempt to address some of the questions: Is it possible to extract robust population size information from a truncated[2] sample in a monitoring campaign? Not only may a subset of known sources be chosen for monitoring, but populations sizes are necessarily biased due to observational flux limits. Thus, can the transient behaviour of sources be exploited to infer population numbers? This project seeks to investigate whether biostatistical capture-recapture methods could aid in transient population size estimation, as circumstantial evidence alongside methods such as population synthesis. The expectation is that we may already be able to investigate the above questions using a combination of simulation and existing data. The most basic and natural way of addressing population size is to keep count through a catalogue of sources; however,

[2]The terms *truncated* and *censored* are often used in statistics to highlight the limitations or incompleteness of a dataset w.r.t. a particular variable. *Truncated* data in this investigation may refer to observed data of members of the population that is bounded by some variable (e.g. apparent magnitude) due to systematic design that excludes observations beyond those bounds. *Censored* data differs from truncated data by including observations beyond the variable bounds by partially describing them in terms of the bound.

the number of sources in a catalogue is not intended to be an estimate of the underlying population size. In this work, we will explore whether capture-recapture statistics can estimate the underlying population size of transient sources. In Chapters 3 to 5, we use simulations and astronomical datasets for application of capture-recapture.

The remainder of this chapter provides an overview of the sources of focus that are later used in the analysis and introduces capture-recapture concepts. A complete outline of the literature and statistics used in capture-recapture is given in Chapter 2. Chapter 3 presents the simulations of lightcurves of high mass X-ray binaries and reproduction of the results of Laycock (2017), along with additional exploration of estimator behaviour and its parameter space as a function of the sampling cadence. Chapters 4 and 5 see the application of methodologies for two different astronomical datasets in the context of closed and open population assumptions. In Chapter 6 I discuss the caveats and limitations of the analysis performed in this study and possibly other approaches to the data reduction based on the results obtained. Chapter 7 concludes and reflects on the future prospects of this work.

1.1 Linking Capture-Recapture to Astronomy

Capture-recapture are methods traditionally used for population size estimation in animal populations. Capture-recapture uses the data of successive sightings or live-trappings of animals to infer population size and/or survival (Otis et al., 1978; Seber, 1982; Amstrup, McDonald, and Manly, 2005). Take, for instance, the example of counting the impala population of the Kruger National Park. The idea is quite simple, but in practice, nearly impossible to do via a conventional tally of impala. Firstly, it would simply take too long. Secondly, one would inevitably encounter the same impala twice or perhaps even multiple times during the survey, and without proper marking of the animals would likely result in over-counting. On the other hand, there are impala that one may never encounter, and simple counting does not allow for the inclusion of animals that are not observed. Lastly, if the study stretches over months or years, the data may become increasingly unrepresentative of the population size due to births and deaths (Chao, 2001).

Capture-recapture involves data collection by sampling the population, usually on multiple occasions, and taking note of *who*, through marking or identification methods, and *how many* are encountered. The simplest implementation of capture-recapture involves a two-step sampling process. A population is sampled on a first occasion, counted and

each member marked with an identifier. The sample is released back into the population and allowed to redistribute. On a second occasion, the same population is sampled where the size and the number individuals that were marked on the first occasion counted. The ratio of marked individuals in the second sample can be equated to the ratio of marked individuals that were released into the population on the first sampling occasion – allowing for an estimate of the population. Capture data are compiled in a capture (a.k.a. encounter) history, typically in either of the following manners (Rivest and Baillargeon, 2019; Chao, 2001):

1. Capture history per individual:
   Using a binary system for marking each individual as a capture '1' or miss '0' at each occasion i for $i \in \mathbb{N}$, $\mathbb{N} := 1, ..., t$, for t number of occasions.

2. Aggregated capture histories:
   A modified version of item 1, where the capture history is pooled (for each individual) across specific observations using a logical OR operation, and followed by an additional column that indicates the frequency of the captures across the entire study.

3. Number of captures per individual:
   The format indicates only the frequency of captures for each individual across the entire study. It is typically used in cases where there are no clearly defined occasions within the longitudinal study (known as a continuous-time model) and is called *repeated counting data*. (This is similar to recurrent event data.)

4. Aggregated numbers of captures:
   This data structure is a two-column dataset that contains the number of captures for each individual in the first column and the frequency information in the second column.

The analogy that Laycock (2017) draws for astronomically *recurring* transients is that their 'movement in brightness', i.e. coming in and out of view of the observer, may serve as the distinction between 'capturing' and 'missing' an individual. The boundary between a capture and a miss is limited by a physical or an imposed flux threshold. Thus, the assumption for transients will mainly be that these sources are spatially fixed but that their lightcurves can be exploited for population size information. Knowledge of the underlying size of transient populations, alongside astrophysical identifying properties, serves as an additional motivator for targeted searches and may further our understanding of star formation and stellar evolution.

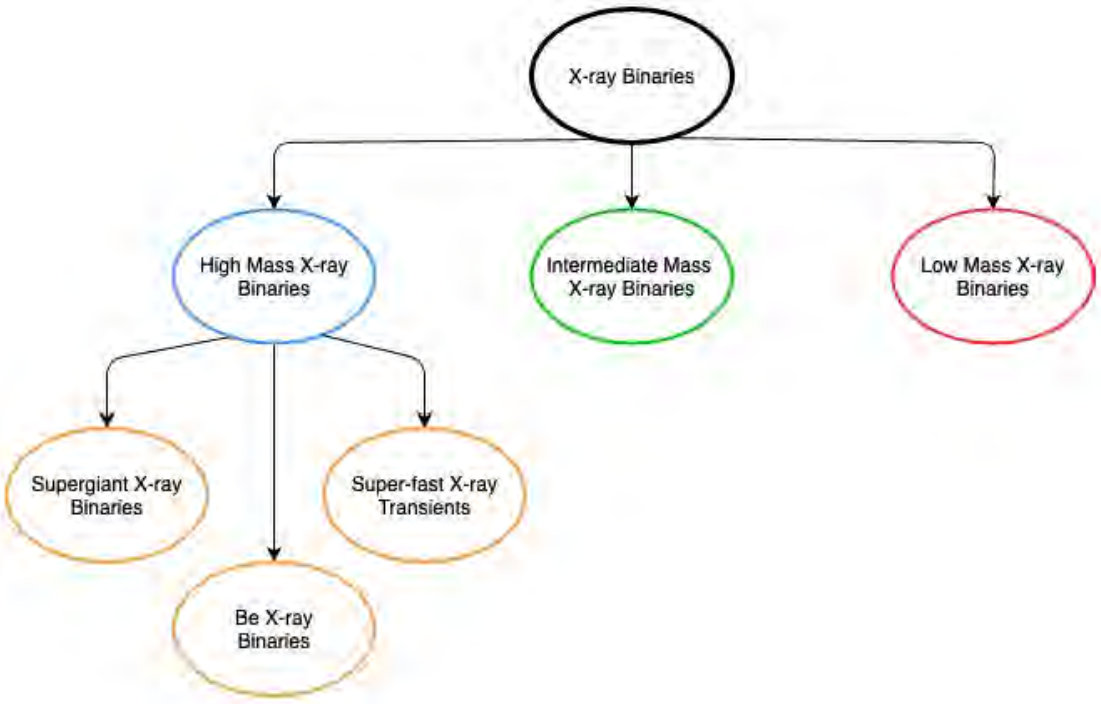


Figure 1.1: Broad classification scheme of X-ray binaries.

1.2 Background on the transients used in the analysis

The datasets used in Chapters 3, 4, and 5 focus on compact binary stars that display recurring outbursts and/or long-term variability in their lightcurves. Below is a short overview of these transient sources and their characteristics. We discuss class characteristics and how capture-recapture applies to these populations.

1.2.1 High Mass X-ray Binaries

An X-ray binary (XRB) is a binary star system that consists of a black hole (BH) or neutron star (NS) compact object and a companion star that donates matter to the compact object (White, Nagase, and Parmar, 1995; Liu, van Paradijs, and Van den Heuvel, 2007). XRBs are broadly categorised according to the mass of the companion (or donor) star, where high mass X-ray binaries (HMXBs) are typically $\geq 10M_{\odot}$, the low mass X-ray binaries (LMXBs) $\leq 1M_{\odot}$, and the intermediate mass X-ray binaries (IMXBs) in-between (Charles and Coe, 2006). Figure 1.1 contains a schematic of the broad classification scheme for XRBs. There are additional sub-classifications of the IMXBs and LMXBs, but they fall out of the scope of this work.

The XRB population in both of the Magellanic Clouds has been well studied because of the favourable low foreground extinction along the line of sight and their reasonable close distance (Liu, van Paradijs, and Van den Heuvel, 2005). The Small Magellanic

Cloud (SMC) is at $d = 62$ kpc (Graczyk et al., 2014) and its line of sight depth estimated at up to 20 kpc (Crowl et al., 2001; Glatt et al., 2008b). The low extinction makes them great test beds for studying their stellar populations. They have as a result been well-investigated and surveyed over the past decades in multi-wavelength surveys such as the Optical and Gravitational Lensing Experiment (OGLE) (Udalski et al., 1992; Udalski et al., 2008), *RXTE* (Corbet, 1999), *XMM-Newton* (Sturm et al., 2011), *Chandra* (Zezas and Antoniou, 2017), *eROSITA* (Merloni et al., 2012), *INTEGRAL* (Kretschmar et al., 2019), *Fermi* (Abdo et al., 2010), *NuSTAR* (Pike et al., 2019), *SWIFT* (Krimm et al., 2013), and S-CUBED (Kennea et al., 2018). The SMC is known to harbour a large population of HMXBs, with the majority of them being classified as the Be/X-ray binary (BeXRB) sub-type with a NS as the compact object (Liu, van Paradijs, and Van den Heuvel, 2005; McBride et al., 2008; Townsend et al., 2011). The study of HMXB populations within the SMC is a component within a broader investigation into massive stellar and stellar binary evolution.

HMXB companions are typically classified as an early-type spectral O/B star. HMXBs with neutron star compact objects are categorised by their X-ray emission as X-ray pulsars or not. Further confirmation for HMXB candidacy is needed by identifying an optical companion that spatially coincidences with the X-ray source emission and classification of the star as early-type (O/B). The optical counterparts of XRBs can be elusive, in part because the positional error in the X-ray tends to be greater than that of optical imaging. However, more recent X-ray surveys such as *XMM-Newton* can reach sub-arcsecond positional accuracy where *RXTE* had ~ 1 degree (Haberl and Sturm, 2016). Besides confirming the X-ray pulsations' positional coincidence to the optical counterpart, further confirmation must be made on the spectral, luminosity, and/or timing attributes for both the X-ray source and the optical counterpart. Only then can a good claim as to the HMXB classification and sub-classification be made (Haberl and Sturm, 2016). Thus, the positional uncertainty coupled with the difficulty of distinguishing X-ray emission between background AGN sources from HMXB sources within the SMC proves a considerable challenge and adds to the uncertainty regarding the underlying population size of HMXBs in the SMC. The application of capture-recapture on X-ray and optical data to compare its population size predictions may provide additional insight from a multi-wavelength perspective. The sub-types of HMXBs differ in terms of their accretion modes. The supergiant high-mass X-ray binary (SGXB) systems predominantly accrete matter from the companion in a stellar wind with a radial outflow, and the BeXRBs, the most prominent sub-type in the SMC, accrete matter onto the compact object from a circumstellar disc (Charles and Coe, 2006; McBride et al., 2008). Two characteristic outburst types are associated with

BeXRBs, namely (Kretschmar et al., 2019; Martin et al., 2014; Charles and Coe, 2006; Stella, White, and Rosner, 1986):

- Type I outburst; an increase in X-ray flux modulated by the orbital period and generally occurring on the binary's periastron passage, which lasts a few days. The X-ray luminosity is on the order of $L_x \sim 10^{37}$ erg s^{-1}. In most cases, the optical and Type I X-ray outbursts coincide, as seen in Alfonso-Garzón et al. (2017).
- Type II outburst; these are giant X-ray outbursts that are longer in duration (weeks or months), not modulated by the orbital period, and with greater luminosity in comparison to Type I, on the order of $L_x \gtrapprox 10^{37}$ erg s^{-1}.

Be stars in isolation are characterised by the presence of a circumstellar disc. BeXRB systems display truncated discs to their isolated cousins. However, the study of the variability and the mass loss mechanisms of Be stars could shed light on BeXRB properties (Kretschmar et al., 2019, and references therein).

1.2.1.1 Orbital period and neutron star spin period

Many HMXBs display X-ray pulsations that are associated with a NS spin period (Haberl and Sturm, 2016). The study of neutron stars and their spin periods (X-ray pulsations) are therefore highly linked with HMXB discovery. BeXRBs display a log-linear correlation between the neutron star spin period and the orbital period of the binary. This is commonly illustrated using a Corbet diagram of $\log(P_s)$ [s] vs. $\log(P_{\rm orbit})$ [d] (Corbet, 1984). The linear relationship has been empirically determined from the up-to-date parameters of 9 Galactic BeXRBs from Corbet 1984 and is given in Laycock (2017, Eq. 13) as:

$$\log_{10}(P_{\rm orbit}) = 1.1329 + 0.4532 \log_{10}(P_{\rm spin}) \tag{1.1}$$

where $P_{\rm spin}$ is the pulsar spin period measured in seconds and $P_{\rm orbit}$ is the binary orbital period measured in days. Figure 1.2 from Kretschmar et al. (2019) shows several systems of HMXB plotted on the Corbet diagram. The BeXRBs plotted in blue suggest a correlation between the binary orbital period (Pb) and spin period (Ps).
The orbital period distribution for the sources in the HMXBCAT (Galactic HMXBs) and MAGHMXBCAT (Magellanic HMXBs) catalogues by Liu, van Paradijs, and Van den Heuvel (2005, 2006) are plotted in Figure 1.3. The catalogued sources in HMXBCAT and MAGHMXBCAT are more abundant at orbital periods of $P_{\rm orbit} < 200$ days. However, not all catalogued systems have known orbital periods. Thus, the true orbital

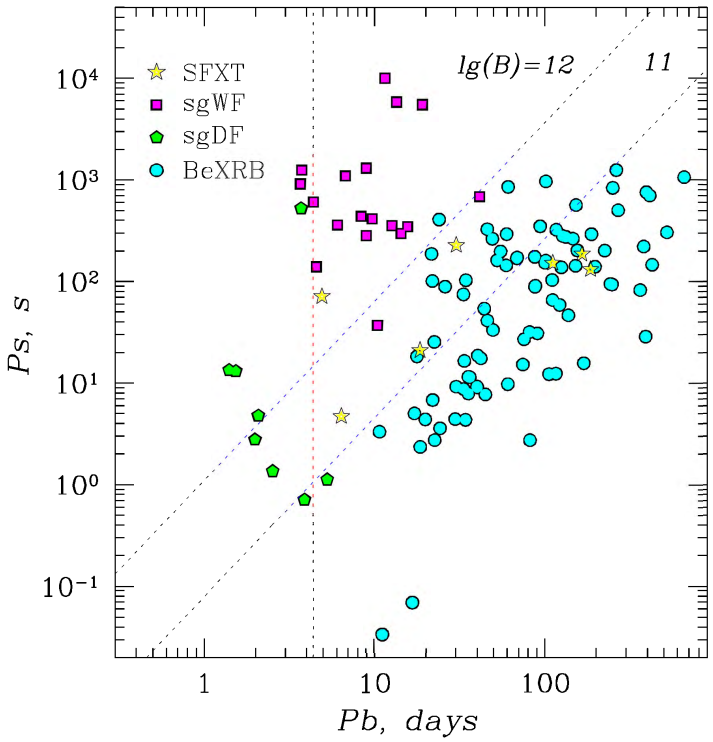


Figure 1.2: Figure from Kretschmar et al. (2019) displays a Corbet diagram for HMXBs with known spin periods. The BeXRBs plotted in cyan suggest a correlation between the neutron star spin period and the binary orbital period.

period distribution remains uncertain. In Chapter 3, we model several orbital period distributions from the spin period-orbital period link (Eq. 1.1).

1.2.1.2 HMXB population in the SMC

The number of known HMXBs have dramatically increased over the past decades. Figure 1.4 shows the number of known HMXB sources as a function of time. The increase in the known HMXBs in the SMC post-1995 is particularly notable as it coincides with the launch of *RXTE* (Galache et al., 2004; Laycock et al., 2003), increasing to more than 80 known sources in 2005. Haberl and Sturm 2016 reviewed 148 candidate HMXBs in the SMC along with a rank in the level of confidence in the classification. Some 27 sources were rejected as misidentifications of normal B stars, whilst 63 of those in the sample contain known pulsars. The current known characteristics leave almost half of the sources with an undetected X-ray pulse period. There may be various reasons for this, including the angle of the neutron star's rotational axis. However, it remains unclear what real fraction of neutron stars in BeXRBs do not exhibit pulsations (Haberl and Sturm, 2016).

[3]HMXBCAT Catalogue: `http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/455/1165`
[4]MAGHMXBCAT Catalogue: `http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/442/1135`

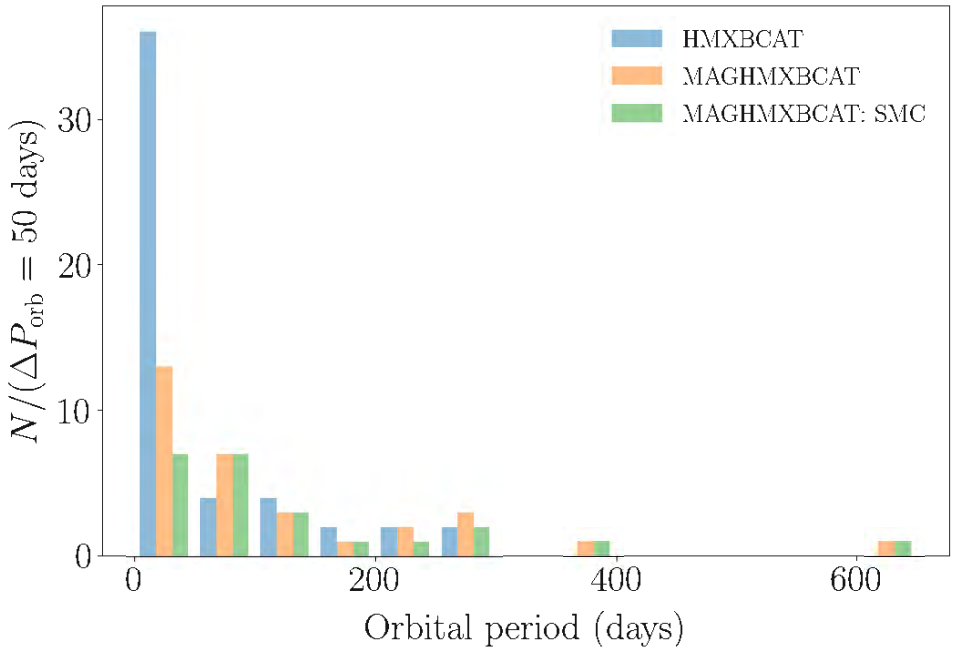


Figure 1.3: Distribution of the known orbital periods of Galactic (HMXBCAT[3]) and Magellanic Cloud (MAGHMXBCAT[4]) HMXBs from Liu et al. (2005; 2006).

1.2.2 Dwarf Nova Cataclysmic Variables

Cataclysmic Variables (CVs) are binary systems that consist of a compact white dwarf type star that is accreting mass from a low-mass companion star, typically a red dwarf of spectral classification M or K, through Roche lobe overflow. Dwarf Novae (DNe) are a subclass of non-magnetic CV binary stars, i.e. where the magnetic field of the white dwarf is negligible or sufficiently weak not to be a factor in the mass accretion flow (Hellier, 2001). The outburst recurrence timescales of DNe range between days and tens of years and typical outburst amplitudes between 2 and 8 magnitude (Warner, 1995). The accepted theory for the cause of DNe outbursts is the *disk-instability* model, a process during which the accretion disc transitions from a cool, low-viscosity to a hot, high-viscosity state (Hellier, 2001, and references therein).

CVs can be identified in many ways – from multiwavelength imaging to spectroscopy. DN subtypes, however, are identified from their optical outburst characteristics. There are several sub-types of DN that are distinguished by their outburst lightcurve morphology. The sub-types of DN are (Hellier, 2001; Warner, 1995):

- *Z Cam*

  Systems that display lengthened standstills at an inflated brightness compared

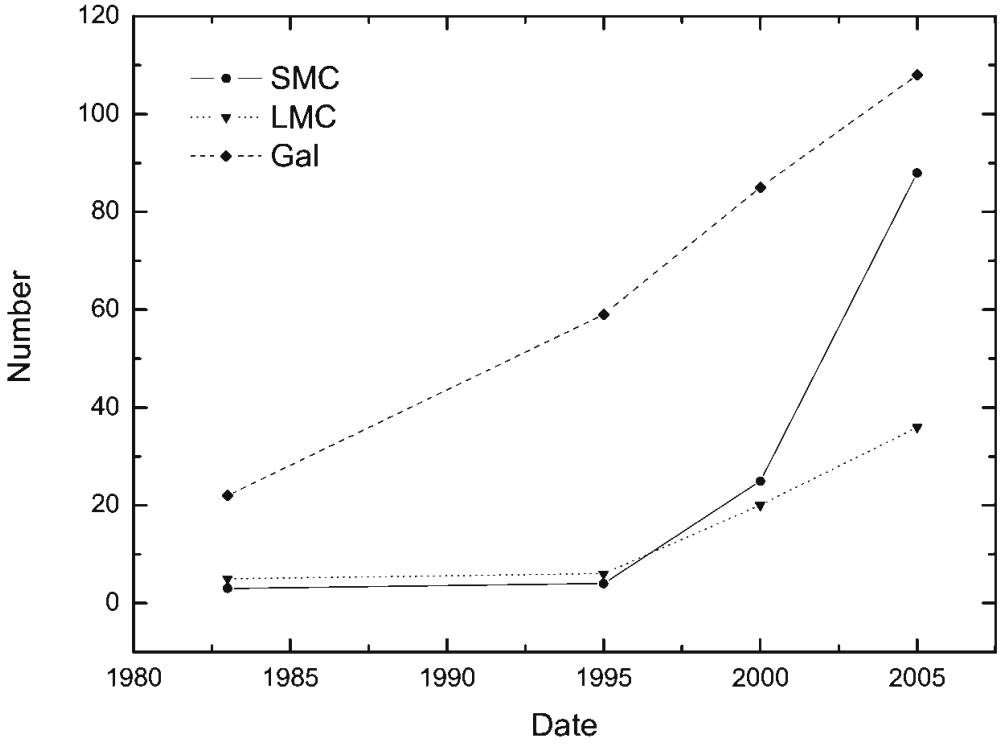


Figure 1.4: Figure 1 from Liu et al. 2005 showing the number of high mass X-ray binaries in the SMC, the Large Magellanic Cloud (LMC) and Galactic Bulge as they have been detected over time.[3, 4]

to pre-outburst, < 1 magnitude from maximum, between epochs of recurrent outbursts. These standstills continue in the range of weeks to years.

- *SU UMa*
  These systems display superoutbursts in addition to outbursts. Superoutbursts are even more energetic and longer-lasting than regular outbursts, which can be on the order of weeks in contrast to days for a regular outburst.

- *U Gem*
  This class has outbursts recurring every few months and lasting up to a few weeks (Hameury, 2020).

In Mróz et al. (2015), the dataset which is used for analysis in Chapter 5, DN sources are identified by the condition of an outburst of at least ≥ 1 magnitude held for a minimum of three consecutive nights. It is therefore expected that the dataset consists of the SU UMa and U Gem sub-types.

1.2.2.1 Galactic CV and DN populations

The evolution of CVs, and subsequently DNe, are of interest for similar reasons to that of HMXBs – to characterise stellar binary evolution and its many facets and the important role that accretion from one star to the other has in the formation of discs and jets.

Coppejans et al. (2016) provide a thorough analysis on the statistical characteristics of a large sample of DNe that have been classified from the CRTS, a large sky-survey searching an area of $\sim 26\,000$ deg^2 (declination range $-30^\circ < \delta < 70^\circ$) for optical transients (Drake et al., 2009, 2014). After having identified more than 1000 CVs in the CRTS, Coppejans et al. (2016) ran a classification algorithm to extract DN-type sources and use their long-period lightcurves (between 8 and 9 years) to constrain the recurrence times and duty cycles of dwarf-nova outbursts.

Similarly, OGLE has provided a large dataset of DNe in the direction of the Galactic Bulge (Mróz et al., 2015), which forms part of the analysis in Chapter 5. However, DNe identified in this survey are situated in a dense stellar field with the possibility of contamination of other sources, such as isolated Be stars, background AGN or optical counterparts of X-ray binaries, at up to 5% of the total dataset. More detail is provided in Chapter 5.

Chapter 2

The Statistical Background

Statistical inference has become an important and necessary tool in modern-day science. It can be defined as the practice of making informed deductions of a parameter based on the random sampling of an unknown underlying distributed parameter space (Roussas, 2003).

Inference has formed the backbone of population studies within the biological sciences. Particularly with relation to the ecological and medical sciences in situations where limited data are available, or the collection of a complete sample would prove far too onerous, or just practically impossible (Chao, 2001, and references therein). Capture-recapture methods have evolved dramatically over the past century, partly because of the 'revolution in statistical thinking' during the 1920s and 1930s, as Fienberg (1992*b*) describes it. The statisticians, R.A. Fisher (1890-1962), J. Neyman (1894-1981), and E.S. Pearson (1895-1980), are but some of the key players that have developed the necessary tools to accommodate scientific study design and data collection through random sampling (Fienberg, 1992*b*; Amorós, 2014). Capture-recapture is fundamentally built on the back of these advances and subsequently has seen its use in ecology for population size and density estimation in animals, but also associated parameters such as mortality (or survival) rates, as well as epidemiological applications for the estimation of the extent of the spread of disease (Fienberg, 1992*b*; Böhning, van der Heijden, and Bunge, 2017). In economics, it is used to infer income level and in various census applications. It is seen in many applications in the social sciences and computer science for software error testing (Chao, 2001; Böhning, van der Heijden, and Bunge, 2017).

There are many overlaps in the statistical methods employed by biologists and (astro-)physicists. The 18th and 19th centuries saw advances made on probability and inferential statistics motivated by biological and sociological research questions (Fienberg, 1992*b*). Pierre-Simon de Laplace (b.1749, d.1827) is sometimes credited with being 'the Father of statistical inference'. He laid the path in his paper on using inference to estimate the size of France's population through data obtained from marriage and birth records in respective parishes (Amorós, 2014; Laplace, 1783; Truscott and Emory, 1902). We also saw the development of the modern clinical trial within the last century, which makes major use of inference for deductions from clinical observations of humans and animals regarding drug testing or disease progression (Fienberg, 1992*b*; Amorós, 2014; Böhning, van der Heijden, and Bunge, 2017). These are standard examples of statistical tools that were developed for the needs of the biological sciences. They have been designed specifically to make reasonable judgements based on samples of the population that are much smaller than the actual size; but assumed to be representative.

What does it mean to record a *representative* sample? It means that the sample should accurately reflect the population under investigation to draw conclusions without biases. Therefore, it typically requires one to randomise the sampling in an appropriate manner suited to the study. Unrepresentativeness is common in astronomical datasets, often due to observational flux limitations and requires post-facto consideration and correction. However, such a randomised sampling technique is often applied with regards to the study of a population – be it to estimate the size or test the effectiveness of a new drug for treating a particular medical condition to infer a result that could apply to a larger population. Inferential statistics has become so ubiquitous, especially in the medical sciences, that we would arguably be the worse for it without it.

Estimating the underlying population size given a sample of observed members has remained an unsolved problem in astronomy. However, biostatistics has successfully addressed population size estimation for more than a century (Laycock, 2017). The extent of the literature that underpins the statistical analysis of the capture-recapture methods displays the continued and active interest in this field to improve the accuracy and precision of estimation. This is reflected in reviews provided in Cormack (1968*a*), Otis et al. (1978), White et al. (1982), Seber (1982, 1986, 1992, 2001), Chao (2001), Schwarz and Seber (1999), Williams, Nichols, and Conroy (2002), and Amstrup, McDonald, and Manly (2005). The methodology includes consideration for sampling technique, experiment design, data collection, estimator bias determination, error analysis, and model selection (Baillargeon and Rivest, 2007).

Modern biostatistics is assisted by hypothesis testing and the assessment of statistical significance and robustness of the results from an analysis of population characteristics. These are inherited concepts and tools in astronomy and physics used to good effect. Examples include variable correlation identification, model fitting, and parameter estimation. Bayesian inference has also become increasingly popular, especially in cosmology and gravitational wave astronomy (see Loredo, 1992; Feigelson and Babu, 2006; Feigelson, 2016; Smith et al., 2020; Sharma, 2017). Methods such as the Markov Chain Monte Carlo (MCMC) are used to recover the underlying distribution of a parameter and are standard practice nowadays. Computational implementations of MCMC algorithms, such as `emcee` or `PyMC3`, are widely available from open source libraries (Foreman-Mackey et al., 2013, 2019; Salvatier, Wieckiâ, and Fonnesbeck, 2016).

In this chapter, I present a general overview of the emergence of capture-recapture theory (§2.1) and the fundamental statistics that accompany the analysis (§2.2), such as error analysis and assumptions, least-square estimation, likelihood theory, and Poisson regression (the latter is directly applicable in this work). Goodness-of-fit and model selection criteria are introduced in §2.2.3 to evaluate models against each other.

Furthermore, a review is given of the main estimators under the relevant closed or open population assumptions, from early twentieth-century to modern use in §2.3 and §2.4. The astronomical capture-recapture framework is over-viewed as presented in Laycock (2017). Lastly, a summary is given in §2.5 on the available software implementations for capture-recapture and presents the **Rcapture** software that is used in this work.

2.1 Capture-Recapture: An evolution

'*How many*' is a fundamental quantitative assessment that provides insight into the population under study. Biostatistics offers various methods for population size estimation[5]. These include complete and incomplete identification, direct and indirect marking (Cattadori et al., 2003; Stephens et al., 2006; Anderson et al., 2013) methods. The topic of interest, the capture-recapture methods (also called mark-recapture or capture-mark-recapture in Otis et al. 1978), generally fall within the complete and direct sampling methods but can be extended to indirect and incomplete measures (Lavallée and Rivest, 2012; McClintock et al., 2014).

By repeatedly sampling a *representative* part of the population, the capture-recapture method allows for estimation, as was said in the example of the Kruger National Park's impala population in Chapter 1. It allows the research problem to become practicable in contrast to *ad infinitum* counting (Chao, 2001).

The first and fundamental distinction that must be made regarding the behaviour or characterisation of the population under investigation is whether the system is isolated or not (Seber and Schofield, 2019). A *closed population* refers to the assumption that no births or deaths can occur within the population i.e. the population remains constant. This implies that a population must be defined by some boundary, spatial or otherwise, and that different population groups cannot interact with one another through the exchange of members. In practice, it restricts the sampling of a population within a short time frame in order for the condition to hold. When this assumption cannot be made, the population is described as *open*. In contrast to being *closed*, it allows for births, deaths, and permanent and temporary migration of members within the population. The sampling structure for open populations, however, differs from closed populations in the sense that sampling epochs are pooled together into sub-epochs where the closure condition holds. For this reason, early uses of open population analysis were geared towards survival estimation rather than size estimation. In contrast, a closed population analysis estimates the size over a limited epoch in order for the condition of closure to hold. (Otis et al., 1978; Seber, 1982; Borchers et al., 2002; Seber and Schofield, 2019)

For any capture-recapture application, one must first decide whether to describe the population as open or closed. In Chapter 1, I mentioned the 'movement in brightness' analogy in the context of astronomically *recurrent*, outbursting stars that are assumed to be static in space. *If* we were to observe a sample of the sky that encompassed

the entire population of interest over a timescale of less than a century, then we could reasonably assume that a population is closed since:

1. the probability of 'birth' or 'death' of stars at a given time is low. The best equivalent that we have for (stellar) death is the supernova (SN), for which Adams et al. 2013 estimate the Galactic SN rate between ~ 1 and ~ 12 per century.

2. the stars are essentially static (we shall assume negligible proper motion for the populations under study here).

3. the timescale in which we observe stellar populations for the study is small in relation to their evolutionary lifespan, making stars within the population unlikely to undergo significant changes in their fundamental characteristics such as spectral and/or luminosity classification. Thus we do not expect 'emigration' in brightness space. Violation of this condition may require resorting to open population analysis.

The above list is conditional on how the population is sampled. Astronomical surveys and/or cross-referenced datasets are necessarily truncated or censored[6] based on a host of selection criteria such as age, mass, orbital separation), distance, or redshift. All pose significant challenges to the data analysis and inference (Efron and Petrosian, 1992; Loredo, 2007). As surveys change in sky footprint over time, there is a need for incorporating open population analysis. Robust estimation, for example, is a sophisticated analysis method for capture-recapture data that fuses the closed and open population sampling structure to estimate size, migration and survival parameters (Pollock, 1981; Kendall, 2012). This sampling setup appears to be a viable option for longitudinal studies extending across decades that may violate classical closed population assumptions. More detail is given in §2.4.

[6]Refer to footnote 2 on P. 2

2.2 Some basic terminology, definitions and theory

To introduce the fundamental statistics that capture-recapture has been developed from, we detail some background on the regression techniques and goodness-of-fit statistics used for model selection in capture-recapture.

2.2.1 Models, data, residual error and measurement error

A model M can be defined as a mathematical formulation used to describe the trend of a data set using explanatory variables[7]. In this work, we will only consider parametric models that depend on a fixed set of parameters, which will be denoted by a parameter vector, θ. A randomly observed and dichotomous (i.e. two category) variable has a realised measurement denoted in lower case by x_{ij} individual $i = 1, \ldots, n$ at sampling occasion $j = 1, \ldots, t$. To fully describe a set of data, the suitability of the model must be evaluated and sources of error taken into consideration. The term "error" may not be consistently attributed or used across the (bio-)statistical and astronomy literature, so I will provide the different types below, described as according to Grace (2016). *Measurement error* may refer to either of the following:

1. *Random error*
   Random error is an unpredictable, uncontrollable, and unreproducible error that is inherent to the measuring procedure.

2. *Systematic error*
   Systematic error (also referred to as *bias*) is a consistent and repeatable type of error resulting from the measuring procedure or related equipment. This error type may be reduced by better procedural planning or using a different or improved measuring device.

3. *Sampling error*
   It is also known as *Estimation error*. It is an uncertainty embedded in the estimate based on the choice of *predictor*[8] of the observed data, by inference from a *sample*[9] of a population, i.e. *limited* data.

Each of the measurement errors mentioned above is ubiquitous in astronomy and a measurement may be affected by a mixture of the three types. It generally describes how the measured datum differs from the ideal and unknown measurement (Grace, 2016;

[7] An explanatory variable is not independent of all the other variables that describe the dataset. Therein lies the subtle difference between independent and explanatory variables.
[8] The estimate used to describe the response of the observed data.
[9] A *sample* refers to a representative subset of the population.

Kelly, 2013). A categorical response variable (which may be dichotomous, or nominal i.e. more than two categories) is denoted by Y_j, for up to n individuals sampled at the jth level (in capture data this translates to being captured exactly j times where $1 \leq j \leq t$ sampling occasions) with realised measurement y_j.

Covariates, denoted $\{X_{ij}\}$, is the term used to describe the set of explanatory variables given individual i at level j. These are underlying measurements (such as age or sex in animal population studies) taken from the *sample* during the experiment that (possibly) *explain* particular effects seen in y_j[10]. The error should be specified as associated with the covariate or response as *covariate measurement error* or *response measurement error*. There is also a distinction between categorical and continuous error variables. Covariates may be either continuous[11] or categorical. An error of a categorical variable may be labelled as *misclassification*, i.e. the probability of an individual being wrongly classified at the jth level. *Continuous measurement error* describes the error around a measurement. Buonaccorsi (2010) and Grace (2016) highlight approaches that correct for measurement error and misclassification, such as bias corrections, moment-based approaches, likelihood techniques, and more. A set of independent and identically distributed (i.i.d.) response variables is denoted as $\mathbb{Y} = \{Y_j\}$ where each Y_j is from the same underlying distribution. Estimates of the parameters $\hat{\theta}_j$ represent a *central* measurement of the realised values of Y_j given by dataset y_j, as in Eq. 2.1. A frequency distribution may be constructed from these estimates that is known as the *sampling distribution*.

$$y_j = \hat{\theta}_j + \epsilon_j \tag{2.1}$$

where $\hat{\theta}_j = E\left[y_j \mid X_{ij}\right]$ is the expected value at level j that is modelled as a function of the set of regression parameters θ_j. The residual term ϵ_j is independent of the covariates, $\{X_{ij}\}$ in the model, distributed normally with a zero mean value and constant variance, σ^2. A *naive* estimator of θ_j would neglect the covariate measurement error and equate the best estimate $\hat{\theta}_j$ of the population sample to the true estimate of the population, θ_0. If we consider the covariate measurement error, i.e. perform a non-naive analysis, a measurement error model must be specified that defines the relationship between the true and observed values. Measurement error may be categorised into two groups (Grace, 2016; Buonaccorsi, 2010). *Classical measurement error* assumes the distribution of the observed values given the true values, whereas *Berkson error* refers to the opposite, i.e. the assumption of the distribution of the true values given the observed values. A model M, which characterises the response variable given a set

[10] These effects will be discussed later in relation to variation in capture probabilities.
[11] A variable characterised by an infinite number of values between any two values.

of parameters θ_j, has covariates $\{X_{ij}\}$. $\{X_{ij}\}$ is linked to *surrogate*[12] covariates $\{X^*_{ij}\}$, either through the classical measurement error model or the Berkson measurement error model assumptions (assuming both with additive error), given by Equations 2.2 and 2.3 (Grace, 2016; Buonaccorsi, 2010).

The *Classical measurement error model*, which accommodates experiments where measurements are repeated, is given by:

$$X^*_{ij} = X_{ij} + e_{ij} \tag{2.2}$$

where e_{ij} is the measurement error that is independent of X^*_{ij} and residuals ϵ_j, with a zero mean (or expectation) and constant variance $\sigma^2_{e_{ij}}$. The naive estimator $\hat{\theta}^*_j$, which converges in probability to θ^*_j for samples $j \to \infty$, is subsequently related to the observed value θ_j via $\theta^*_j = \theta_j \left(\sigma^2_{x_{ij}}/(\sigma^2_{x_{ij}} + \sigma^2_{e_j})\right)$ and $\sigma^2_{x_{ij}}$ is the variance in X_{ij}.

The *Berkson measurement error model* is given by:

$$X_{ij} = X^*_{ij} + e_{ij} \tag{2.3}$$

where e_{ij} is the measurement error that is independent of X_{ij} and residuals ϵ_j, with a zero mean (or expectation) and constant variance $\sigma^2_{e_{ij}}$. The naive estimator converges to the the true parameter θ_0 for samples $j \to \infty$. The objective is to find an estimator $\hat{\theta}_j$ to infer the *data generating mechanism* of the true value, θ_0, between a response variable Y_{ij} and corresponding covariates $\{X_{ij}\}$, respectively. However, because the true value, θ_0, is unknown to us in the case of population size estimation and because there are many different ways to construct an estimator of θ_0, it is not possible to determine $\theta_0(\mathbb{Y})$ about the true underlying distribution with model $M(y_{ij} \mid \theta_0)$. Thus, we construct an estimator, such that the residuals, i.e. the difference between the data and the model, associated with the best estimate, is achieved through minimisation under the *least squares algorithm* for the residual sum of squares (RSS): $\sum(\epsilon_j)^2$.

The following general properties are desirable for statistical estimators (Judd, McClelland, and Ryan, 2017):

- *Unbiasedness*
  An unbiased estimator has the expectation value equal to observed peak value of the sampling distribution. If there is large disparity between the expectation

[12] A *surrogate* variable is a measured variable used in the place of an intractable variable. The intractable variable becomes *implicit or unmodelled*. A surrogate represents how the actual measured variable differs from the variable used in the model to describe the relationship between covariates and response. It sometimes called a *proxy* variable.

and observed value, then the estimator is said to be *biased*, i.e. defined as bias $= E\left[\hat{\theta}\right] - \theta$.

- *Efficiency*
  An efficient estimator refers to the narrowness of the sampling distribution spread as it provides higher precision estimation i.e. a lower sample variance.

- *Consistency*
  An estimator is consistent if the spread in the sampling distribution becomes narrower with increasing sample size.

Similarly to estimator properties, the following are desirable for modelling residual error (Judd, McClelland, and Ryan, 2017):

- *Independence*
  Each residual value is independent if it is uncorrelated to any other residual value in the sample. Violation of this condition often means that parameters may need to be modelled together, knows as *within-subject* modelling. This condition's assumption is non-robust, meaning that estimation, which will differ significantly under dependent residual errors, is not comparable to estimation assuming independent residual errors.

- *Identically distributed*
  It refers to the case where all the sampled measurements are acquired from the same distribution. A standard example is the assumption of a normal distribution of residuals with variance σ^2. Violation of this assumption occurs when units within the sample belong to different distributions with residual variances other than σ^2. In certain contexts, later on, we will see cases of behavioural closed population models, which account for the fact that the act of sampling influences the variance of the residual error distribution. Thus, in the capture-recapture context, identically distributed is equivalently called *homogeneity of variance or homoscedasticity*. It describes the sampling as being from the same underlying distribution. If this last condition is not true, each residual is described as being sampled from a distribution with a different variance associated with it and is called the *heterogeneity of variance or heteroscedasticity*. (cf. independent and identically distributed (i.i.d.))

- *Unbiased residuals*
  The assumption would be that the expectation of the residual error distribution is zero. Systematic bias of the residuals would significantly affect the model parameters and would not be situated around zero.

In astronomy, one would typically minimise the residuals using the χ^2 or reduced χ^2_r (χ^2 per degree of freedom) method since it incorporates measurement error into the best estimate and tests for significance of the residuals against the measurement error. However, it implicitly assumes normal probability theory and in a χ^2 contingency analysis, both x and y variables are dichotomous. A multi-way chi-square contingency analysis has an unclear interpretation and logistic regression must instead be used. The capture-recapture measurements are dichotomous variable outcomes of a set of i.i.d. Bernoulli trials (a.k.a. Poisson sampling) where the ith individual at the jth observation, $x_{ij} = 1$ is a capture and $x_{ij} = 0$ is a miss. Their frequency distribution y_j are assumed to be Poissonian and deemed categorical (a fixed number of categories for the number of times that an individual is captured). Thus, any associated 'measured error' in x_{ij} is a misclassification problem rather than an attributed error with variance $\sigma^2_{e_j}$. The Berkson measurement error model can incorporate misclassification into its modelling. However, the models we will apply will generally assume that individuals' measured state x_{ij} is not misclassified. This will be discussed further in §2.3. Extensions of these models for misclassification cases, such as tag-loss models, e.g. for tagged animals, are also available but will not be considered in this work.

2.2.2 Log-linear regression and Maximum Likelihood Estimation

Regression is a statistical method that seeks to describe the relationship between a set of response variables, Y_j, and the explanatory variables, X_{ij} (Hosmer Jr, Lemeshow, and Sturdivant, 2013). The previous section explained how we describe parameters of interest and the relationship between response and covariates. Logistic and log-linear regression are extensions of linear regression and may be employed for situations such as capture-recapture where the measured data is categorical. A multivariate linear regression model, where y_j is the response given the covariates $\{X_{ij}\}$ at every level j, would be described by the following:

$$y_j \mid \{X_{ij}\} = \beta_0 + \beta_1 x_{1,j} + \ldots + \beta_m x_{m,j} + e_j = \sum_{k=0}^{m} \beta_k x_{k,j} + e_j \tag{2.4}$$

with m observed variables ($x_{0,j} \equiv 1$) and m regression coefficients that are slope parameters and β_0 the intercept, and random error term given by e_j. On average, the expected value for the above given the explanatory variables is:

$$E\left[y_j \mid \{X_{ij}\}\right] = \sum_{k=0}^{m} \beta_k x_{k,j} = \boldsymbol{\beta} \cdot \mathbf{X}_j, \quad (j \in \{1, \ldots, t\},\ t = 2)\,; \tag{2.5}$$

where $\cdot$ denotes the dot product and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_m)$ and $\mathbf{X}_j = (1, x_{1,j}, \cdots, x_{m,j})$ are row vectors. We will later come to know $\mathbf{X}_j$ as the *design matrix*. The expectation,

$E[y_j \mid \{X_{ij}\}]$, is sometimes denoted $\hat{\lambda}_j$, and called the linear predictor function. In the case of dichotomous[13] response variables the linear regression form in Equation 2.5 may be used as is, since the mean responses at each level of $\{X_{ij}\}$ fall on a line. Nominal[14] response data in the form of counts *fall on a curve* i.e. the linear form does not hold, and is subsequently described by Poisson distributed frequency responses[15]. Capture data then take on a log-linear format, since the experiment is designed upon the idea of modelling the number of events at a fixed occasion or interval. It follows that the log-expectation is a linear relation for Poisson distributed capture data:

$$\ln\left(E\left[y_j \mid \{X_{ij}\}\right]\right) = \ln\left(\hat{\lambda}_j\right) = \boldsymbol{\beta} \cdot \mathbf{X}_j, \quad (j \in \{1, \ldots, t\},\ t \geq 3)\,; \tag{2.6}$$

Hence, by exponentiation, the expectation is a multiplicative model:

$$\hat{\lambda}_j = e^{\boldsymbol{\beta} \cdot \mathbf{X}_j} \tag{2.7}$$

Assumptions of Poisson regression

1. A response variable, Y, is described by a Poissonian distribution for each observation, therefore
$$\Pr(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$$
2. Each observation must be independent of the other.
This is a strong assumption that will become the topic of discussion for cases that violate this assumption.
3. For a Poisson random variable Y, the expectation (mean) and variance is equal, $\therefore E(Y) = \mathrm{Var}(Y) = \lambda$.
4. The natural logarithm of the expectation is a linear function of explanatory variables x, $\therefore \ln\left(\hat{\lambda}\right) = \boldsymbol{\beta} \cdot \mathbf{X} \iff \hat{\lambda} = e^{\boldsymbol{\beta} \cdot \mathbf{X}}$.

[13] where y_j is the response of only two possible categories e.g. y_j for $j \in \{1, 2\}$.
[14] where y_j is the response of ≥ 3 possible categories e.g. y_j for $j \in \{1, \ldots, t\}$.
[15] $\hat{\lambda}_j$ predicts the number of units captured exactly j times during the study.

The best set of least squares estimates (LSEs) for $\boldsymbol{\beta}$ can be calculated by using a linear least-squares regression (LLSR) algorithm to minimise the error associated with the fitted model under linear regression. The *least squares method* is a thoroughly tried and tested regression technique for linear regression. However, biostatisticians prefer likelihood-based approaches because of greater provision for generality in the modelling and more robust to bias. There are, however, other estimation methods in use, such as the generalised method of moments, profiling method, and functional estimators are employed (Burnham and Anderson, 2002; Grace, 2016). LSE is equivalent to the method of maximum likelihood estimation (MLE) under the conditions of linearity, homogeneity of variance, normally distributed residual error, and independence of residual error. The likelihood function, $\mathcal{L}$, of a parameter θ given the data y, is mathematically defined as:

$$\mathcal{L}(\theta \mid y) = p(y \mid \theta) \tag{2.8}$$

where p is the conditional probability of the data y, given the fixed parameter, θ. The principle of maximum likelihood (ML) seeks to find the best estimate $\hat{\theta}$, by maximising the likelihood, $\hat{\mathcal{L}}$, usually done via first-order differentiation and a second-order derivative confirmation of the maximum. For t sampled observations, the joint likelihood is represented by the product of the individual likelihoods since the events are independent:

$$\mathcal{L}(\theta \mid y_1, \ldots, y_t) = p(y_1 \mid \theta) \cdots p(y_t \mid \theta) = \prod_{j=1}^{t} p(y_j \mid \theta) \tag{2.9}$$

Maximisation of the joint likelihood function, which is a product of t factors, is equivalent to maximising its natural logarithm. The log-likelihood, denoted ℓ, translates to t summands, which is far more convenient to work with. (Hosmer Jr, Lemeshow, and Sturdivant, 2013; Judd, McClelland, and Ryan, 2017; Roback and Legler, 2021)

$$\ell = \ln(\mathcal{L}) = \sum_{j=1}^{t} \ln\left[p(y_j \mid \theta)\right] \tag{2.10}$$

Under log-linear regression, the likelihood function of the linear predictor, $\hat{\lambda}_j$, given the response data of t independent Poissonion observations is (Roback and Legler, 2021):

$$
\begin{aligned}
\mathcal{L}(\lambda_j \mid y_j) &= \prod_{j=1}^{t} p(y_j \mid \lambda_j) \\
\ln\left(\mathcal{L}(\lambda_j \mid y_j)\right) &= \ln\left(\prod_{j=1}^{t} p(y_j \mid \lambda_j)\right) \\
\ell\left(\lambda_j \mid y_j\right) &= \sum_{j=1}^{t} \ln\left(p(y_j \mid \lambda_j)\right) \\
&= \sum_{j=1}^{t} \ln\left(\frac{e^{-\lambda_j}\lambda_j^{y_j}}{y_j!}\right) \\
&= \sum_{j=1}^{t} -\lambda_j + y_j \ln(\lambda_j) - \ln(y_j!) \\
\text{[Substitute } \ln(\lambda) &= \boldsymbol{\beta} \cdot \mathbf{X} \text{ and } \lambda = e^{\boldsymbol{\beta}\cdot\mathbf{X}}\text{]} \\
\therefore \ell\left(\beta\right) &= \sum_{j=1}^{t} -e^{x_j\beta} + y_j x_j \beta - \ln(y_j!)
\end{aligned}
\tag{2.11}
$$

The log-likelihood may then be maximised, and the regression coefficients in β can be solved numerically with an algorithm such as the Newton-Raphson method. The constant term $\ln(y_j!)$ is often neglected when calculating the log-likelihood.

2.2.3 Model Specification

Fitted models require evaluation of their goodness of fit for making valid inferences. Model specification is separated into two components according to Burnham and Anderson (2002):

1. formulation of a set of candidate models;
2. selection of a model, or a number of models.

The most general methods used in the past have been based on model parameter and precision estimation applied on a model of choice. The covariance is a measure of the degree to which a pair of discrete and randomly distributed variables, (X, Y), vary with each other, or in other words, it is an indicator of dependency between variables. For sampled data realisations (x_i, y_i), $i = 1, \ldots, n$, given equal probabilities $p_i = 1/n$, the sampling covariance is determined by:

$$
\operatorname{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - E(X)\right)\left(y_i - E(Y)\right) \tag{2.12}
$$

where $E(X)$ and $E(Y)$ denotes the expectation of X and Y, respectively. As Burnham and Anderson (2002) notes, the likelihood theory developed by Fisher (1922) is a

method for approaching these problems and assumes an already known model structure with only the parameters of that known model requiring estimation. MLE can be used as an objective estimation of the specified model parameters and the sampling covariance matrix – which is conditional on the appropriate model. Models may then be compared via ranking using a likelihood-based test. Below are several tests available for model selection based on a likelihood approach and specified in the context of log-linear (Poisson) models of the analysis in later chapters.

2.2.3.1 Deviance

The deviance measure compares the expected value and the observed value expressed as the likelihood ratio of the fitted model to the *saturated* model. A saturated model refers to when the number of model parameters equals the number of data points. It is a theoretical construct for comparing models based on the same data. Hence, a low score is preferred since a high score represents overfitting. Thus, the total deviance for $Y \sim$ Poisson is given by (Roback and Legler, 2021; Hosmer Jr, Lemeshow, and Sturdivant, 2013):

$$D = 2\sum_{j=1}^{t}\left\{y_j \ln\left(\frac{y_j}{\hat{\lambda}_j}\right) - \left(y_j - \hat{\lambda}_j\right)\right\} \tag{2.13}$$

The deviance is approximately equal to a χ^2 with $t-k$ degrees of freedom as the number of observations become large, where t is the number of observations and k the number of model parameters. In a marginal likelihood (the integrated likelihood function which adds up to 1, which is a property of MLEs in Poisson models) the latter term $\sum_{j=1}^{t}\left(y_j - \hat{\lambda}_j\right) = 0$, since

$$\ell\left(\lambda_j \mid y_j\right) = \sum_{j=1}^{t} -\lambda_j + y_j \ln(\lambda_j) - \ln(y_j!) \implies \frac{d\ell\left(\lambda_j\right)}{d\lambda_j} = \frac{y_j}{\lambda_j} - 1 \implies \hat{\lambda}_j = y_j$$

2.2.3.2 Pearson's χ^2 statistic

The Pearson statistic, χ^2_P, is a standard test for categorical data, which is similar to the deviance, is also asymptotically equivalent to the standard χ^2 statistic as the number of observations, $t \to \infty$ with $t-k$ degrees of freedom. This likelihood-based score evaluates whether the difference between an observed value and the expected value is due to chance. However, the deviance D has the benefit of comparing nested models where χ^2_P does not (Roback and Legler, 2021; Hosmer Jr, Lemeshow, and Sturdivant, 2013).

$$\chi^2_P = \sum_{j=1}^{t} \left(r_j^P\right)^2 \tag{2.14}$$

where r_j^P are the Pearson residuals under Poisson regression (recall $E[Y] = Var[Y] = \lambda$),

$$r_j^P = \frac{y_j - \hat{\lambda}_j}{\sqrt{\hat{\lambda}_j}} \tag{2.15}$$

A rule of thumb is that most residuals should fall between -2 and 2. Residuals larger than that may indicate poorly fitted data (Roback and Legler, 2021; Baillargeon and Rivest, 2007).

2.2.3.3 Information-theoretic approach

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are measures that penalise models based on the number of fitted parameters and the maximised log-likelihood. These are often used for model comparison where the number of model parameters differ and is not intended as a measure of goodness-of-fit for a standalone model. The criteria are defined as:

$$\text{AIC} = 2k - 2\ln(\widehat{\mathcal{L}}) \tag{2.16}$$

$$\text{BIC} = k\ln(t) - 2\ln(\widehat{\mathcal{L}}) \tag{2.17}$$

where k is the number of estimated parameters in the model, t the number of observations, and $\widehat{\mathcal{L}}$ the maximum value of the likelihood function for the fitted model. Lower scores for both measures indicate a preferred model. (Akaike, 1974a,b; Schwarz et al., 1978; Hosmer Jr, Lemeshow, and Sturdivant, 2013)

2.2.4 Confidence interval estimation

Confidence intervals describe the possible values of a parameter to a specified degree of certainty. The point estimate is the 'best guess' amongst those values and the true value being within that interval to a specified degree of probability. Large intervals describe low precision estimation, whereas small intervals indicate high precision. When using a maximum likelihood estimation (MLE) approach for determining $\hat{\theta}$, the log-likelihood profile $\ell(\theta)$ may be used for confidence interval estimation based on the assumption of asymptotic normality as the number of observation $t \to \infty$ (Hosmer Jr, Lemeshow, and Sturdivant, 2013; Judd, McClelland, and Ryan, 2017). Asymptotic normality is a criterion that demands consistency amongst the set of estimators $\hat{\theta}_j$ and approaches

a Gaussian distribution with mean θ_0 (hypothetical true value), and decreasing the variance as t increases. Likelihood confidence interval estimation is often used in non-Gaussian distributed residuals where independence is violated.

2.3 Closed Population Analysis

2.3.1 Early 20th century population size estimators

We return to capture-recapture theory. Now that the fundamentals have been introduced, we can formally lay down the theory in the context of population size estimation. Whilst Laplace used a similar inferential technique in his 1783 study, the first explicit use of capture data was implemented by Petersen (1894), not for estimating size, but rather mortality rates of flat-fish. Some decades later, Lincoln (1930) utilised the method for estimating the size of an American waterfowl population. Lincoln would capture a sample of ducks at their annual breeding ground and mark them with a band. After a shooting season, he would retrieve these bands from the shooters at a seemingly consistent rate. It is an example of a *two capture occasion method* that equate the ratios of marked individuals in a sample to the marked individuals in the entire population, assumed to be fixed:

$$\frac{m}{n} \simeq \frac{M}{N} \implies \widehat{N} = n\frac{M}{m} \tag{2.18}$$

where N is the size of total population, $\widehat{N}$ is the size estimate of the total population, M the number of marked individuals in the total population, n the number of individuals in the sample, and m the number of marked individuals in the sample. To estimate $\widehat{N}$, M individuals of the population are sampled, marked, and released on the first occasion. After sufficient time is given for redistribution of the first sample back into the population, n individuals are sampled on a second occasion and the number of marked individuals, m, that are recaptured are subsequently recorded. There is an implicit assumption that the second occasion is a random sample *without* replacement, meaning that the unmarked individuals in the sample will not be returned to the population with tags to provide information for further sampling (Seber and Schofield, 2019). Dahl (1917) used this method before Lincoln to estimate trout population sizes in the lakes of Norway. However, the estimator is historically assigned the Lincoln index, Lincoln-Peterson (LP) index, or Peterson method (Seber and Schofield, 2019).

The two-sample method still saw much use in the ensuing years, with examples in Sekar and Deming (1949) and Shapiro (1949), however, this was steadily being replaced by the multiple capture-recapture (MCR) method. Negatively biased estimation (or

underestimation) was a general concern under the two-sample method (Seber, 1982; Chao and Huggins, 2005*a*). Further reviews such as Fienberg (1992*a*), Darroch et al. (1993*a*), and Chao and Tsay (1998) also discuss undercounting particularly with respect to using census data. Seber and Schofield (2019) note that undercounting can be estimated with a third sample of the population. Chapman (1951) proposed a different two-sample estimator in attempt to address this bias:

$$\widehat{N} = \frac{(M+1)(n+1)}{(m+1)} - 1 \tag{2.19}$$

where, $\widehat{N}$, M, n, and m are the same parameters as it is defined for the LP index in Equation 2.18. The Chapman 1951 estimator (Eq. 2.19) is derived from a hypergeometric distributed function (i.e. sampling without replacement). Dual-estimators corresponding to the Binomial and Poisson probability distribution may also be used (assumes replacement):

$$\widehat{N}_{Binomial} = \frac{M\,(n+1)}{(m+1)} - 1 \tag{2.20}$$

$$\widehat{N}_{Poisson} = \frac{Mn}{(m+1)} \tag{2.21}$$

Depending on the shape of the sampling distribution, confidence intervals may be determined accordingly. For instances of the two-sample method where random sampling with replacement is more suitable to the experiment then the Bailey 1952 estimator should be used instead of Chapman 1951:

$$\widehat{N} = \frac{M(n+1)}{(m+1)} \tag{2.22}$$

The literature records Schnabel (1938) as the first to pen down an MCR estimator, as a series of LP indices, which she implemented through an example of counting fish in a lake (Krebs, 1999; Seber and Schofield, 2019). Schnabel (1938) assumes that M_j is fixed for each sampling. The estimator, which is based on a binomial approximation to the hypergeometric distribution, is defined as:

$$\widehat{N} = \frac{\sum_{j=1}^{t} (n_j M_j)}{\sum_{j=1}^{t} m_j}, \tag{2.23}$$

where $j = 1, 2, ..., t.$ is the capture occasion, n_j the number of individuals captured in the jth occasion, M_j the number of cumulative marked individuals *at* the jth occasion ($M_1 \equiv 0$), and m_j the number of marked individuals captured in the jth occasion

($m_1 = 0$). The variance was not provided in Schnabel (1938). Chapman (1952) followed this up with a slight modification that adjusts for bias if the ratio of recaptures of the population at each occasion is assumed to be < 0.1:

$$\widehat{N} = \frac{\sum_{j=1}^{t} (n_j M_j)}{\left(\sum_{j=1}^{t} m_j\right) + 1}, \tag{2.24}$$

Equation 2.24 is generally considered to be the 'up-to-date' Schnabel estimator and suggested for use in Seber (1982). A similar method proposed by Schumacher and Eschmeyer (1943) pointed out that since the Schnabel census design is based on a series of LP indices, then parameters $\frac{m_j}{n_j}$ are on average $\frac{M_j}{N}$. Thus, using a linear LSE algorithm on explanatory and response variables, $x = M_j$ and $y = \frac{m_j}{n_j}$ respectively, one could use the slope of $1/N$ passing through the origin to estimate the population size. This leads the following (Seber, 1982):

$$\widehat{N} = \frac{\sum_{j=1}^{t} \left(n_j M_j^2\right)}{\sum_{j=1}^{t} m_j M_j} \tag{2.25}$$

Subsequently, the variance and standard error can be calculated from the slope parameter as:

$$\text{Variance}\left(\frac{1}{\widehat{N}}\right) = \frac{\sum_{j=1}^{t} \left(\frac{m_j^2}{n_j}\right) - \frac{\left(\sum_{j=1}^{t} (m_j M_j)\right)^2}{\sum_{j=1}^{t} (n_j M_j^2)}}{t-2} \tag{2.26}$$

$$\text{Standard Error}\left(\frac{1}{\widehat{N}}\right) = \sqrt{\frac{\text{Variance}\left(\frac{1}{\widehat{N}}\right)}{\sum_{j=1}^{t} (n_j M_j^2)}} \tag{2.27}$$

When the total recaptures are low (< 50), it is advisable to use a Poisson error to estimate the confidence intervals for Schnabel and Schumacher and Eschmeyer methods. If the total recaptures surpass 50 units then the t distribution approximation of the normal distribution is adequate (Seber, 1982): $\frac{1}{\widehat{N}} \pm t_\alpha \, S.E.$, where t_α is the t-distribution given the significance level, α, and $S.E.$ is the standard error.

2.3.2 Modern implementation

The MCR method has evolved since Schnabel (1938) to methods that allow for flexibility on the rigid assumptions of the classical closed population model. Darroch (1958) was the first to correctly derive the probability model that corresponds to the Schnabel (1938) MCR estimator using a multinomial distribution, also allowing for sampling with replacement. The standard model under Darroch (1958) is referred to as the homogeneous capture probability (reference) model (M_0) in Otis et al. (1978). A list of the experimental assumptions (Assumptions 1 to 4) made under the classical closed population model (Otis et al., 1978) is provided on P. 33. These assumptions represented the most restrictive case for closed population estimation and were implicit in most early capture-recapture uses. More often than not, at least one of these assumptions are violated in practice. The notation used in this chapter onward is defined in the comprehensive review of closed population models by Otis et al. (1978). The parameters and their definitions may be found in Table 2.1.

Table 2.1: Mathematical notation and definition of each parameter for closed populations, as described in Otis et al. (1978), and used in §2.3.2.

Parameter	**Definition**
N	The population size (which remains constant).
i	Numbered individual. $i = 1, 2, ..., N$.
j	Capture occasion. $j = 1, 2, ..., t$.
X_{ij}	Capture history matrix.[16]
p_{ij}	The capture probability matrix, where p_{ij} is the probability of capture for the ith individual on the jth capture occasion.
n_j	The number of individuals captured in the jth occasion i.e. $\sum_{i=1}^{N} X_{ij}$.
$n.$	The total number of captures during the study, $\sum_{j=1}^{t} n_j$.
u_j	The number of new (unmarked) individuals captured in the jth capture occasion.
f_j	The capture frequencies i.e. the number of individuals captured exactly j times in t occasions.
M_{t+1}	The number of distinct individuals caught during the experiment (t fixed for a given experiment), $M_{t+1} = \sum_{j=1}^{t} f_j = \sum_{j=1}^{t} u_j$.
M_j	The cumulative number of marked individuals *at* the jth occasion. ($M_1 \equiv 0$)
$M.$	The sum of M_j, $\sum_{j=1}^{t} M_j$.
m_j	The number of marked individuals captured in the jth occasion, $j = 2, \ldots, t$. ($u_j = n_j - m_j$ and $m_1 = 0$)
$m.$	The sum of m_j, $\sum_{j=1}^{t} m_j$.

[16] where $[X_{ij}] = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1t} \\ X_{21} & X_{21} & \ldots & X_{2t} \\ \vdots & \vdots & & \vdots \\ X_{N1} & X_{N2} & \ldots & X_{Nt} \end{bmatrix}$ and $X_{ij} = \begin{cases} 1, & \text{if the } i\text{th individual is caught on the } j\text{th occasion;} \\ 0, & \text{otherwise.} \end{cases}$

Closed population assumptions

$$M_0 \text{ (null model)} : p_{ij} = p$$

Assumption 1. *Population is constant.*
The population remains constant throughout the duration of the experiment.

Assumption 2. *No tag loss.*
Individuals do not lose their tag/mark identification during the study.

Assumption 3. *No marking or recording error.*
Tags are correctly noted and capture states correctly recorded at each occasion.

Assumption 4. *Homogeneity.*
Individuals have equal probability of capture and this remains constant throughout the duration of the study. The act of marking or capturing must also not affect this probability.[17]

The null model (M_0) is generalised by the likelihood function, with parameters N and p, given a set of possible capture histories $\{X_\omega\}$ (ω denotes the sequential capture history of an individual) is given by (Otis et al., 1978):

$$\mathcal{L}(N, p \mid \{X_\omega\}) = \frac{N!}{\left[\prod_\omega X_\omega!\right](N - M_{t+1})!} \cdot p^{n_\cdot}(1-p)^{tN-n_\cdot} \tag{2.28}$$

and the variance described by (Darroch, 1958):

$$\text{Var}(\widehat{N}) = \frac{N}{(1-p)^{-t} + (t-1) - t(1-p)^{-1}} \tag{2.29}$$

Several approaches for estimating N are found in the literature. These include the maximum likelihood method (ML, which is a standard in the field), the jackknife and bootstrap[18] methods, the log-linear method, and a generalised linear method. Furthermore, non-frequentist approaches such as Bayesian methods and mixture models are also employed (Chao, 2001).

When considering an ML approach, there are three different considerations to be made (Amstrup, McDonald, and Manly, 2005):

[17] The homogeneity condition will often not hold in transient astronomy due to factors such as variation in duty cycles and luminosity.
[18] The jackknife and bootstrap are methods of resampling used for parameter and confidence interval estimation.

1. The probability of the observational data of each individual i, assuming independence, i.e. $\mathcal{L}(y_i) = p(Y_i = y_i)$, and constructing the full likelihood as the product of each individual likelihood, $\mathcal{L} = \prod_{i=1}^{n} p_i$.

2. For data which are grouped, the likelihood is constructed from a multinomial distribution that represents the capture data observations fully.

3. For capture data collected from independent groups (e.g. distinct categories by male or female), the likelihood is constructed as the product of the multinomial distribution, which is the product of the likelihood associated with each independent group.

There are three deviations from the standard model that have been identified and have been subject to much discussion. That has expanded the standard model to accommodate violations of some, or all, of these conditions. The three key effects that cause variation in the capture probabilities (which are all violations of Assumption 4) of the population are accommodated under the following frameworks:

(a) temporal capture probability model (M_t),
The capture probability amongst the population is homogeneous but can vary from one capture occasion to the next.

(b) behavioural capture probability model (M_b),
The capture probability is affected by trap-happy, or trap-shy, animal responses after the initial capture. An additional parameter, c, is introduced as the probability that an animal is captured on occasion $j \geq 2$.

(c) heterogeneous capture probability model (M_h),
Each individual in the population has a unique capture probability.

Combinations of (a), (b), and (c), namely: M_{th}, M_{bh}, M_{tb}, and M_{tbh} are also possible frameworks. Borchers et al. (2002) makes the argument that these mentioned above are only 'models in a vague sense' because each is a broad description of the intent in addressing unequal capture probability and provides assumptions within the framework. Each framework contains a myriad of models to implement it. Furthermore, different models under the same framework may behave differently in terms of estimator performance, such as bias, precision (standard error of the estimate) and confidence interval coverage. Hence, the model selection criteria such as those described in §2.2.3 that are used for comparison amongst several model options is paramount to drawing a conclusion on the 'best estimate' of the population.

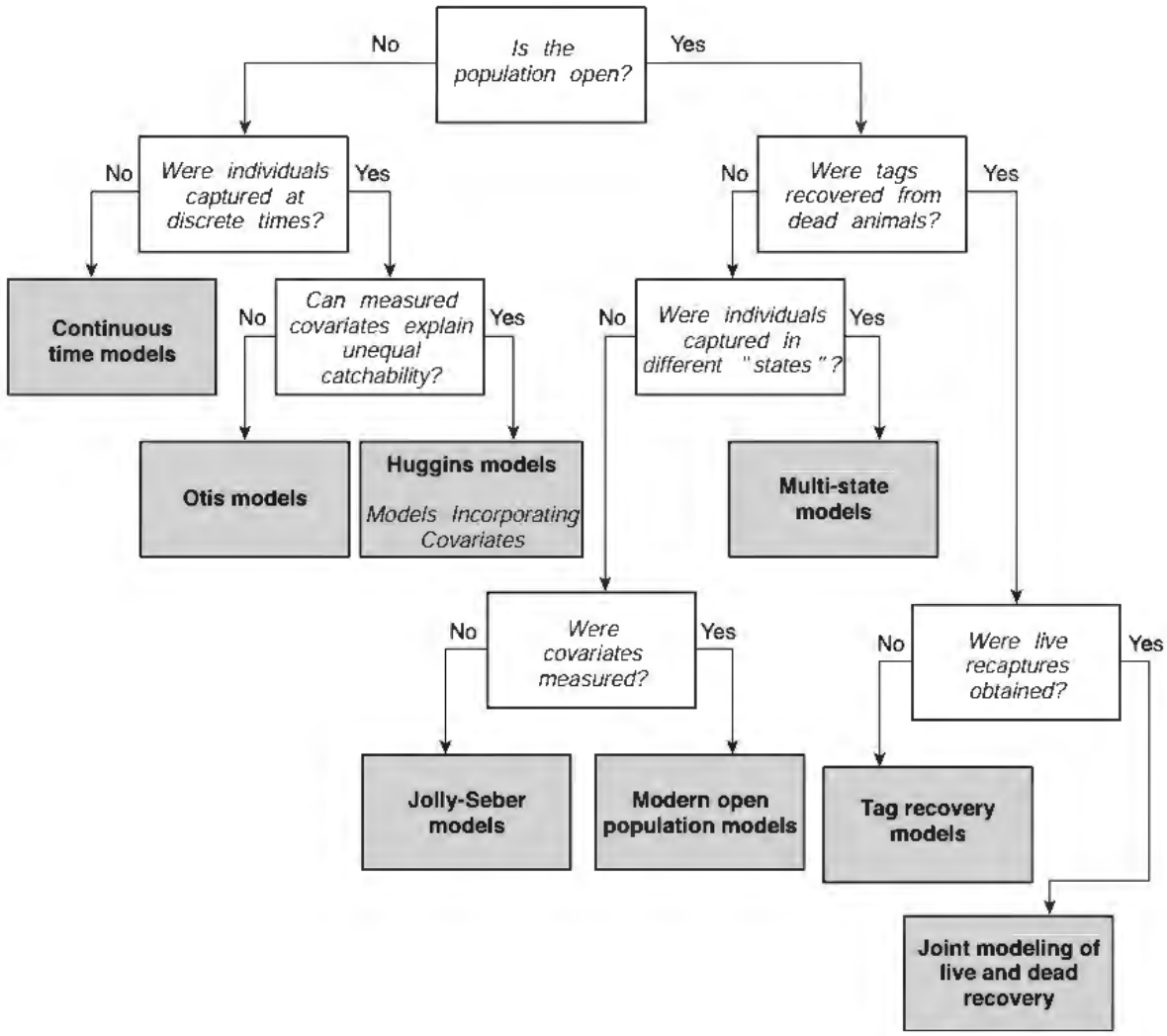


Figure 2.1: Flowchart from Figure 1.2 in Amstrup, McDonald, and Manly (2005) showing the appropriate model based on conditions of population type, sampling design and unequal catchability[19](or heterogeneity). In this work we remain within the bounds of the Otis et al. 1978 models, and a basic implementation of robust design analysis (see §2.4).

An additional distinction is made between discrete-time and continuous-time models that are a result of sampling design. Discrete-time models are conducted according to several distinct occasions. An occasion is typically a day or night in which traps are left, and a group of animals are caught; the exact time of capture is obscured. Discrete-time is the general framework considered in Otis et al. (1978). Continuous-time models have a fixed epoch in which an individual may be caught at any time during data collection. In this case, the capture time is recorded for an individual, and essentially the capture

[19]In many capture experiments, auxiliary data are captured, i.e. the age or sex of the animal. If these variables are found to be correlated with unequal catchability in the population, then they may be incorporated into the population modelling as explanatory variables, as shown in Huggins (1989, 1991).

occasion has a sample size of 1.

Both discrete- and continuous-time frameworks can be accommodated under variations in capture probability and estimators constructed for M_t, M_h, M_b, and combinations thereof. Herewith follows a description of each framework that considers unequal catchability. Figure 2.1 shows the divergent capture-recapture methods. Most methods can be customised to suit a specific need, i.e. sampling design, variation in capture probabilities, or closed vs open population modelling. Following is a review of the effects of variation in capture probabilities.

2.3.2.1 Time effects

Model M_t: $p_{ij} = p_j$

The temporal model accounts for variation in the number of captures made from one occasion to the next. All members of the population remain equally 'catch-able', however, allowance is made for a changing probability according to capture occasion j. The likelihood function under a binomial distribution is described by the following:

$$\mathcal{L}(N, p \mid \{X_\omega\}) = \frac{N!}{\left[\prod_\omega X_\omega!\right](N - M_{t+1})!} \cdot \prod_{j=1}^{t} p_j^{n_j}(1-p_j)^{N-n_j} \tag{2.30}$$

and the approximate variance:

$$\mathrm{Var}(\widehat{N}) = \frac{N}{\prod_{j=1}^{t}(1-p)^{-j} + (t-1) - \sum_{j=1}^{t}(1-p_j)^{-1}} \tag{2.31}$$

Darroch (1958) successfully derived an iterative solution for the multiple capture-recapture scenario, i.e. $t > 2$. When $t = 2$, the maximum likelihood estimation of the above reduces to the famous Lincoln index, shown previously in Eq. 2.18:

$$\widehat{N}_t = \frac{n_1 n_2}{m_2}$$

Otis et al. (1978) note that this model is adequate for capture probabilities around $\bar{p_j} \gtrapprox 0.1$, but that smaller probabilities tend to result in significantly biased estimates.

2.3.2.2 Behaviour effects

Model M_b: $p_{ij} = p$ until first capture, and $p_{ij} = c$ on recapture

This model particularly addresses behavioural responses to capture, with particular emphasis on the first capture. The multinomial likelihood is given by:

$$\mathcal{L}(N, p \mid \{X_\omega\}) = \frac{N!}{\left[\prod_\omega X_\omega!\right](N - M_{t+1})!} \cdot p^{M_{t+1}}(1-p)^{N-M_{t+1}-M_\cdot} \cdot c^{m_\cdot}(1-c)^{M_\cdot - m_\cdot} \quad (2.32)$$

where c is the probability that an animal is captured on any occasion after its first capture occasion. Note that c can be independently estimated from parameters N and p (Otis et al., 1978). It is the so-called 'trap-happy' response (or 'trap-shy' when the response is the opposite) that arises when the initial capture probability of an individual increases ('trap-happy', $p < c$), or decreases ('trap-shy', $p > c$), the subsequent catchability of an individual or multiple individuals (Chao and Huggins, 2005*b*).

The MLEs for this model were derived by Moran (1951) and Zippin (1956, 1958). This model's disadvantage is that it assumes the same behavioural response to capture for all individuals in the population.

2.3.2.3 Individual effects (Heterogeneity)

Model M_h: $p_{ij} = p_i$

Heterogeneity in the capture probability within the population is perhaps the most difficult departure from the classical set of assumptions to account for. Model M_h asserts that each individual j of the population has a capture probability associated with it which is independent of all the others. Tests for heterogeneity is important in order to establish its presence within the population. Tanaka (1951) observed that for a population of red-backed voles, the number of marked animals, M_j, was not linearly related to the proportion of marked and caught. In contrast, a homogeneous sample would be suitably modelled by a linear relation through the origin as interpreted by Schumacher and Eschmeyer (1943). This effect may either be due to heterogeneity or accidental death and removal. Cormack (1968*b*) provides two additional tests for heterogeneity, using the frequency distribution of captures $f(c)$ for a random variable c_i, using a moment estimator about the origin ($\mu'_k = E\left[c^k\right]$, and evaluating $\gamma^2 = (\mu'_2/\mu'_1) - 1$ the coefficient of variation, essentially a measure of heterogeneity. Equal catchability has the null hypothesis of $H_0 : \gamma = 0$. If the null hypothesis does not hold, then heterogeneity is present.

Heterogeneity poses a particular problem because each animal in the population is described by its own explanatory variable; the observed model can only be evaluated once every animal is observed. Without an observational model present for each animal, a likelihood function cannot be constructed. Thus, heterogeneous models have to be tackled through 'design-based'[20] inference or other appropriate methods to estimate the population size, $\widehat{N}$. In this case, the capture history and its frequency distribution become essential to characterise these unequal probabilities.

Former studies have shown that classical homogeneous estimators are negatively biased when heterogeneity is present (Chao and Huggins, 2005*b*). Burnham and Overton (1978) showed, using a non-parametric approach with the assumption the the individual probabilities, p_i, represent a random sample from an unknown distribution, that the capture frequencies, f_j, are sufficient statistics for estimating the population size, $\hat{N}_{\text{Jk}}$, under heterogeneity:

$$\hat{N}_{\text{Jk}} = \sum_{j=1}^{t} a_{jk} f_j \tag{2.33}$$

where k is the k-th order jackknife (generally $k < 5$) selected through sequential testing to find the appropriate order, f_j is the the number of animals captured exactly j times ($j = 1, 2, \ldots, t$) and f_0 represents the number of unobserved individuals, and coefficients a_{jk} as delineated in Burnham and Overton (1978). The jackknife estimator suffers from a usually negative bias; however, it tends to be small with an increasing number of capture occasions (Otis et al., 1978), though for experiments where almost all of the animals are captured, it will overestimate the population size. The variance may also be determined from the jackknife due to the linearity of the estimator. Unfortunately, the jackknife does not offer quantitative measures for describing the heterogeneous capture probability in the population, as is the case for all of the previous models described. Chao (1987, 1989) developed a moment-based estimator, based on Burnham and Overton (1978), which functions as a lower bound of the population size, assuming large t and small p_j. The first-order moment is given by:

$$\hat{N} = M_{t+1} + f_1^2 / (2 f_2) \tag{2.34}$$

with a 95% confidence interval between the bounds:

[20] This refers to *designing* the matrix of the coefficient multipliers (a.k.a. the *design matrix*), $\mathbf{X}$, in the Poisson regression as laid out in Eq. 2.5, to improve the model fit.

$$\left[M_{t+1} + \frac{(\hat{N} - M_{t+1})}{C},\ \ M_{t+1} + (\hat{N} - M_{t+1})C\right] \\ \text{where } C = \exp\left\{1.96\left[\log\left(1 + \hat{\sigma}^2/(\hat{N} - M_{t+1})^2\right)\right]^{1/2}\right\} \tag{2.35}$$

Chao et al. (2005) proposed a bias correction for Eq. 2.34:

$$\hat{N} = M_{t+1} + \frac{(t-1)f_1(f_1 - 1)}{2t(f_2 + 1)} \tag{2.36}$$

Other methods for handling heterogeneity have been proposed by Smith and van Belle (1984) through a bootstrap, Chao and Lee (1992) and Lee and Chao (1994) using sample coverage, Lloyd and Yip (1991) via the Martingale method, to mention a few. Log-linear approaches have been reviewed by Rivest and Lévesque (2001), Rivest and Daigle (2004), Rivest and Baillargeon (2007), and Baillargeon and Rivest (2007) which are applicable in this work. The relevant formats for regression are reviewed in §2.5.1 as implemented by the **Rcapture** software.

2.3.2.4 Time and behaviour effects

Model M_{tb}: $p_{ij} = p_j$ until first capture, and $p_{ij} = c_j$ on recapture

This model encompasses the combination of temporal and behavioural variation in capture probability, given by the likelihood:

$$\mathcal{L}(N, p \mid \{X_\omega\}) = \frac{N!}{\left[\prod_\omega X_\omega!\right](N - M_{t+1})!} \cdot \prod_{j=1}^{t} p_j^{u_j}(1 - p_j)^{N - M_{j+1}} \cdot c^{m_j}(1 - c_j)^{M_j - m_j} \tag{2.37}$$

where c_j is the probability that a previously captured (i.e. marked) individual is captured on the jth occasion ($j = 2, 3, \ldots, t$). However, the model is non-identifiable[21] since the sufficient statistics needed to calculate $\hat{\mathcal{L}}$ $(2k - 1)$ are fewer than the modelling parameters $(2k)$. This model can only be applied under parameter restrictions or assumptions and is also not considered within this work's log-linear framework.

2.3.2.5 Behaviour and individual effects

Model M_{bh}: $p_{ij} = p_i$ until first capture, and $p_{ij} = c_i$ on recapture

[21] There are more parameters than observations; therefore, no functional (one-to-one) mapping exists.

Model M_{bh} is a generalisation of M_h that models dependency of the heterogeneous capture probability on previous capture history. It has similar modelling difficulties to overcome as with M_h. Pollock (1981) characterised the probability distribution as:

$$\begin{aligned} \mathrm{P}\left[\{X_\omega\}\right] = & \frac{N!}{u_1!u_2!\cdots u_t!\,(N-M_{t+1})!} \\ & \cdot \pi_1^{u_1}\pi_2^{u_2}\ldots\pi_t^{u_4}\left(1-\sum_{j=1}^{t}\pi_j\right)^{N-M_{t+1}} \\ & \cdot \mathrm{P}^*\left[\{X_\omega\} \mid u_1, u_2, \ldots, u_t\right] \end{aligned} \tag{2.38}$$

where $j = 1, 2, \ldots, t$, u_j = number of unmarked animals caught at time j, $\pi_j = E\left[(1-p)^{j-1}p\right] = \int_0^1 (1-p)^{j-1}p\, dG_1$, and $\mathrm{P}^*\left[\{X_\omega\} \mid u_1, u_2, \ldots, u_t\right]$ = a conditional probability distribution that does not depend upon the parameter N or the distribution $G_1\left(p;\theta_1\right)$.

According to Otis et al. (1978) and Pollock and Otto (1983), M_{bh} often shows negative bias which is exacerbated by a large heterogeneous effect. Pollock and Otto (1983) developed an improved jackknife estimator in response. It takes on the general form of,

$$\hat{N} = u_1 + u_2 + \cdots + u_{k-1} + ku_k = M_{k+1} + (k-1)u_k. \tag{2.39}$$

The jackknife may similarly overestimate, as with M_h, in cases where nearly all animals are captured (Chao and Huggins, 2005*b*).

2.3.2.6 Time and individual effects

Model M_{th}: $p_{ij} = p_i p_j$

Time and heterogeneity effects are combined under this model, assuming independence between samples, hence the multiplicative form for the capture probability. Only more recently have there been reasonable population size estimators constructed for N under this model. Chao and Huggins (2005*b*) notes that this framework does not rely on actual recapture but that resighting of an individual (i.e. non-invasive data capture which implicitly assumes no behavioural capture variability) in the population is sufficient towards data capture. The sample coverage estimator developed by Lee and Chao (1994) can also be used under M_{th}. Logit[22] models (Agresti, 1994; Coull and Agresti, 1999) and log-linear models (Darroch et al., 1993*b*; Agresti, 1994) have also proved fruitful, with the added advantage of being able to compare models because they are

[22]Logit model is short for **Logistic model**, where the regression is of the general form: $\log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$.

under the same regression scheme through model selection criteria, e.g. choosing M_{th} vs. M_h.

2.3.2.7 Time, behaviour, and individual effects

Model M_{tbh}: $p_{ij} = p_{ij}$ until first capture, and $p_{ij} = c_{ij}$ on recapture

M_{tbh} is the most general of all the model frameworks that incorporate all the sources of variation in capture probability. It is often not a practical model if restrictions or assumptions on parameters cannot be made. Some generalised approaches have been explored, such as estimating equations in Chao et al. (2001), and generalised linear models in Huggins (1989, 1991).

2.3.3 Closed population models within the context of astronomy

2.3.3.1 Generalised exponential model

A generalised exponential model for the cumulative count of captured individuals, Nc_k, is derived from first principles in Laycock (2017, Eq. 7) expressed as an exponential function of capture occasion observation k. At each observation k, the expected number of captured sources is n_k. Each source is treated as a Bernoulli random variable with a p probability of capture from a total closed population of N_0 sources. We can therefore approximate that the average number of sources captured at observation k is:

$$n_k = \bar{n} = N_0 p \tag{2.40}$$

For any observation number, k, the sum of the captured and un-encountered sources, Nc_k and Nu_k, respectively, satisfies:

$$N_0 = Nu_k + Nc_k \tag{2.41}$$

Asymptotic behaviour of Eq. 2.41 require that the number of un-encountered sources $Nu_k \to 0$ for $k \to \infty$. The radioactive decay law satisfies this requirement:

$$\begin{aligned} Nu_k &\propto e^{-pk} \\ \text{For } k = 0,\ Nu_0 &= N_0 \\ \therefore\ Nu_k &= N_0 e^{-pk} \quad \text{(Subst. Eq 2.41)} \\ \text{and}\quad Nc_k &= N_0 \left(1 - e^{-pk}\right) \end{aligned} \tag{2.42}$$

The decay law allows for modelling of a homogeneous capture probability p, however, it assumes a continuous variable k and may be classified as a *continuous-time model*

(cf. p. 36). Given a discrete number of observations, Nu_k is better modelled using a geometric distribution $Nu_k \propto (1-p)^{k-1}p$ for $k = 1, 2, 3, \ldots$ with the same mean as in Eq. 2.40. The decay law distribution and the geometric distribution are approximately equal for small p ($p \lessapprox 0.2$). It follows that the un-encountered and the cumulative sources at observation k is:

$$\begin{aligned} Nu_k &\propto (1-p)^{k-1}p \\ Nu_k &= A(1-p)^{k-1}p \\ \text{For } k=0,\ Nu_0 &= N_0 \\ \therefore A &= \frac{N_0(1-p)}{p} \\ \Rightarrow\ Nu_k &= N_0(1-p)^k \\ \text{and}\ \ Nc_k &= N_0\left[1-(1-p)^k\right] \end{aligned} \tag{2.43}$$

2.3.3.2 Sources of variation in capture probability

Laycock (2017) points out that the generalised exponential models derived in Eqns. 2.42 and 2.43 do not hold when the population is subjected to effects of heterogeneity. Heterogeneity (model M_h) will necessarily play a role in the catchability of recurrent astronomical transients in the form of the outburst duty cycle, which is likened to the capture probability p. Recurring transients may show outbursts on widely different time scales, which in the case of a binary transient system, is a function of the orbital period. In such a case, Eq. 2.43 may be extended to (Laycock, 2017, Eq. 9):

$$\{N_{c_k}\}_j = \sum_j N_{0,j}\left[1-(1-p_j)^k\right] \tag{2.44}$$

However, various other physical parameters, such as age or mass, may give rise to heterogeneity and are contingent on the study population. Methods by Huggins (1989, 1991) could be applied using astrophysical explanatory variables to explain underlying unequal catchability in the population. Furthermore, Laycock (2017) shows that the behaviour response model, M_b, is an effective estimator when samples are highly correlated. It becomes relevant when a significant fraction of the individuals in consecutive samples are obtained at a cadence faster than $0.1T$ (T is the typical orbital period) from the simulated X-ray binary population. The high cadence oversamples the population within the same outburst cycle. This result is reproduced in Chapter 3 using simulated X-ray lightcurve data.

2.4 Robust Design Analysis

The open-population model has a 1-capture per epoch design which in practice makes use of pooled samples from capture occasions on multiple closed population epochs (Kendall, 2001). General open population models, such as the Jolly-Seber (JS) and Cormack-Jolly-Seber (CJS) models (Cormack, 1964; Seber, 1965; Jolly, 1965), dealt sufficiently with unequal capture probability in the survival rate estimators but did not incorporate the same level of heterogeneity into the modelling of the population size estimate (Carothers, 1973). Pollock (1982) developed a robust design approach to improve the size estimation for open population models under heterogeneous capture probability by incorporating the sub-structure of closed population analysis during the multiple epochs of the open population analysis. Therefore, the overall scheme was more robust in terms of heterogeneity and considered births and deaths within the population and its dynamics as well, i.e. immigration and emigration (Kendall, 2001).

The JS and CJS models use the non-robust form of the 'return rate' parameter when estimating survival in animal populations. The return rate, in its simplest form, is defined as (Kendall, 2012):

$$R = \varphi \times p \tag{2.45}$$

where R is the measured return rate, φ the apparent survival probability, and p the apparent capture probability – conditional on the individual's survival.

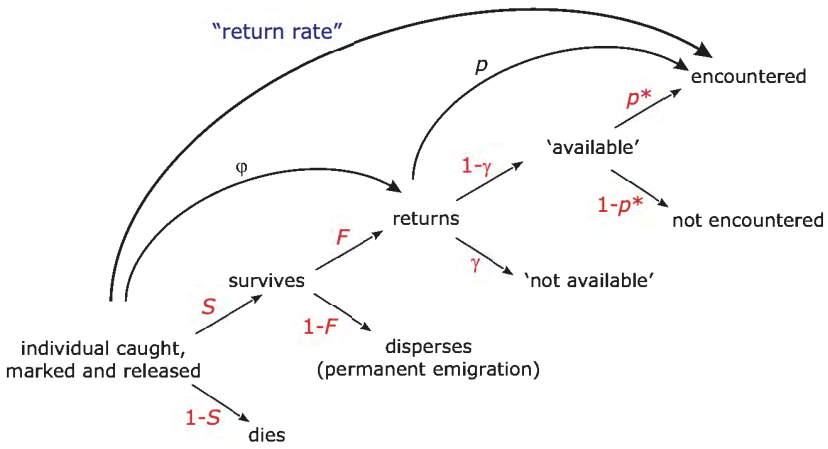


Figure 2.2: 'Fate' diagram from Kendall (2012); Cooch and White (2008) that visualises the makeup of the return rate of individuals in terms of parameters: apparent capture probability p, apparent survival probability φ, fidelity F, and temporary emigration probability γ.

Table 2.2: Breakdown of the return rate parameters.

Parameter	Definition
p^*	true capture probability
γ	temporary emigration probability
p	apparent capture probability ($p = [1 - \gamma]\, p^*$)
S	true survival probability
F	fidelity (opposite to permanent emigration)
φ	apparent survival probability ($\varphi = SF$)
R	return rate of individuals captured, marked and released

Equation 2.45, however, only provides a high-level overview on the matter. Kendall's illustration in Figure 2.2 (a.k.a. the 'fate' diagram) breaks down each component into its most basic features. It makes exceptions in the model for individuals unavailable for capture (but points out that they need to be considered in the overall estimation). The parameters and their definitions are provided in Table 2.2.

In the case of recurring astrophysical transients, we can probably assume that all caught individuals survive so that $S \sim 1$, given that the probability of stellar death is $\ll 1$ — similarly, we permanent emigration in magnitude space over a time-scale of less than a century. Thus, the S and F parameters in the model are not 'real' astronomical effects in this instance. However, when we start to consider observation strategy (sampling design) and the fact that surveys may return for observation to some individual sources more frequently than others, some of these parameters take on a similar meaning for astronomy, in particular where $(1 - \gamma)$ indicates 'availability'.

The overall idea of 'observability' is well-described by Figure 2.3 taken from Kendall (2012). A visual analogy to Figure 2.3 is to imagine that at a given capture occasion, the field of view of the telescope is illustrated by the *observable* study area, whereas

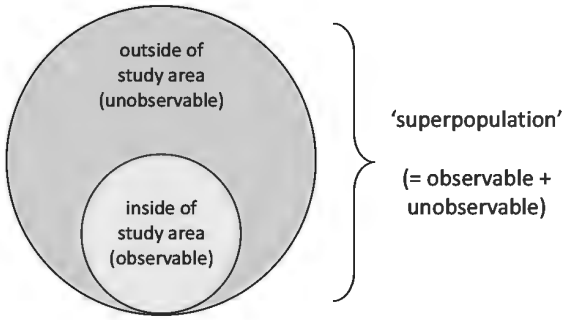


Figure 2.3: Figure from Kendall (2012) illustrates that individuals may either be in an 'observable' or 'unobservable' state which precedes capture probability.

the superpopulation represents the entire field of survey. Not all sources inside the study area will necessarily be captured due to the variable nature of astronomical transients, but, in principle, they are located within the observable (or accessible) area. When the population is open, individuals may 'move' between these states of being 'observable' and 'unobservable' (see Figure 2.4). Parameters γ' and γ'' characterise the state transition as *temporary emigration* parameters.

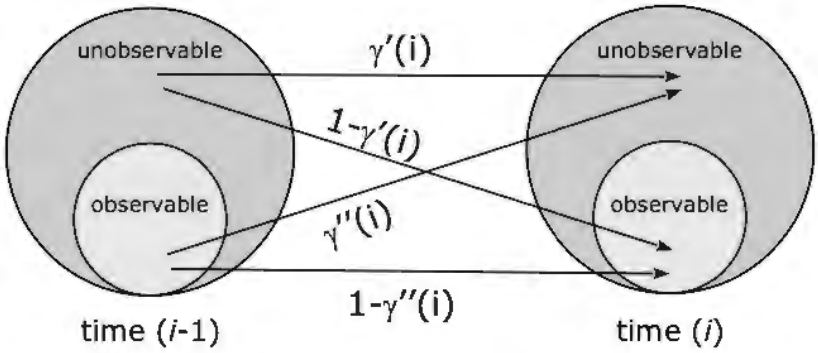


Figure 2.4: Modelling the 'movement' between states of observability in a robust design (Kendall, 2012).

Table 2.3: Temporary emigration parameter definitions.

Parameter	Definition
γ'_i	the probability of being off the study area, unavailable for capture during primary epoch (i) given that the animal was *not* present on the study area during epoch $(i-1)$, and survives to epoch (i).
γ''_i	the probability of being off the study area, unavailable for capture during the primary epoch (i) given that the animal was present during primary epoch $(i-1)$, and survives to epoch (i).

The classical robust design introduced by Pollock assumes a two-level hierarchical sampling structure[23] for which the open population definition holds *between* primary sampling epochs and the closed population definition holds *within* each primary sampling epoch. The primary sampling epochs contain a set of secondary samples, where the number of secondary samples is not necessarily the same within each primary epoch. Population size estimates can be made within each primary epoch from the secondary samples. The dynamical parameters such as survival and immigration are estimated between primary epochs as new individuals are introduced, and others leave.

[23]In capture-recapture literature, the hierarchical structure is referred to as primary and secondary *periods*. To avoid confusion for astronomy, I use the term *epoch* instead of *period*.

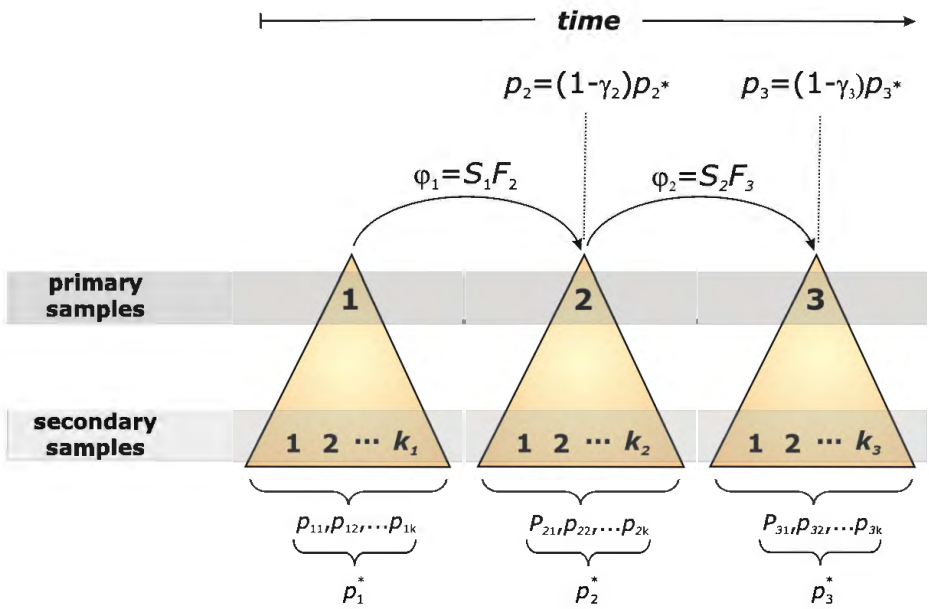


Figure 2.5: Pollock's 1982 robust design model and key parameters within the sampling structure. Figure from Kendall (2012).

In the framework of a closed population, all individuals are by definition 'available' for capture. For each primary epoch, the true probability of capture, following a geometric distribution, may be calculated from the probability of capture from each secondary epoch sample,

$$p^* = 1 - [(1-p_1)(1-p_2)\ldots(1-p_t)] \text{ for } t > 1 \text{ occasions.} \tag{2.46}$$

The apparent capture probability, p, may still be estimated from the standard CJS models, and hence, an ad hoc estimator for $\gamma : \hat{\gamma} = 1-(\hat{p}/\hat{p}^*)$ may be calculated. This is neatly described in Figure 2.5 according to Pollock's 1982 robust design which shows the relevant parameters for the primary and the secondary samples. Between the primary epochs, individuals may then 'move' between the different states of observability, which may be summarised in a *transition matrix* from capture occasion i to $i+1$ as in Table 2.4.

Table 2.4: Transition matrix with probabilities of moving between observable and unobservable states pertaining to capture.

	unobservable, i	**observable**, i
unobservable, $i+1$	γ_i'	γ_i''
observable, $i+1$	$1-\gamma_i'$	$1-\gamma_i''$

Assumptions under classical robust design analysis

Assumption 1. *Population size is constant across secondary sampling occasions.*
The population size remains constant across secondary sampling occasions (within primary epoch) with no additions or losses.

Assumption 2. *Temporary emigration is either random, Markovian*[24]*, or based on a brief first capture response.*

Assumption 3. *Homogeneity of survival probability.*
This survival is not dependent on capture availability, even under Markovian[24] emigration.

The robust design may be extended to characterise multiple states (not only binary) within the scheme; however, that is beyond the scope of this work. The assumptions under robust design incorporate that of closed and open population methods.

2.5 Capture-recapture software

There are various software programs available for processing and analysing capture-recapture data. Perhaps one of the most popular among them is 'Program MARK' (Cooch and White, 2008). Although it is originally written in the `Fortran` programming language, it has only been deployed for Windows machines and is enclosed within a Graphical User Interface (GUI). It offers various models and customisations to the user; however, as Laycock (2017) notes, it is an unlikely candidate for use amongst astronomers because of its unscalability to other programs in the current GUI framework. MARK does have an R environment counterpart called RMARK (Laake, 2013) but it has not been implemented to have the full modelling capabilities of MARK. Whilst `Fortran` is still used in astronomy, nowadays astronomers widely use the Python programming language for computing (e.g. the Astropy project is geared towards computing utilities for astronomers in the Python language (Robitaille et al., 2013; Price-Whelan et al., 2018)), as well as R for advanced statistical purposes. A full wrapper of MARK's capabilities into RMARK, and preferably for Python as well, will be welcomed.

Specialised and proprietary statistical software such as SAS (SAS Institute Inc., 2020), SPSS (IBM, 2020) and Stata (StataCorp LLC, 2020) offer extensive models for general

[24]Refers to a Markov process in which the probability of the current state is dependent on, and only on, the previous state. This is beyond the scope of this work.

linear regression, data imputation and much more. Stata, available on Windows, Mac, and Linux platforms, also offers other biostatistical analysis tools such as survival analysis and epidemiological incidence. Furthermore, Stata allows for Python integration through which one can call it from within the software. SAS is a native Windows platform but can be run on Mac OS and Linux using virtualisation software.

The open-source R programming language remains a popular tool for statisticians and has a large and active community for support (R Development Core Team, 2014, 2010). As explored by Laycock, I also use the **Rcapture** package (written and maintained by Rivest and Baillargeon, 2019) for analysis of capture data, and it is available within the R suite. The **Rcapture** package uses log-linear models with mainly Poissonian regression for its capture-recapture analysis. It supports open and closed population analysis and allows for customisation of model fitting and the design matrix. Several options for capture-probability consideration are available as well as model selection substantiation. **Rcapture** makes use of the native R `glm` function for maximum likelihood estimation. Provided below is a description on using **Rcapture** for analysis that are of relevance to Chapters 3, 4, and 5.

2.5.1 Rcapture software

Rcapture forms the core capture-recapture analysis tool in this work. Herewith follows an outline of the software use and implementation. Full descriptions can be found by the authors in Baillargeon and Rivest (2007) and Rivest and Baillargeon (2007) with regards to the R software.

2.5.1.1 Closed population analysis

The overarching closed population analysis in **Rcapture** consists of four steps:

1. Exploration of the capture data using *descriptive statistics* and identification variations in capture probability through parameters:
   - f_i: the number of units captured exactly i times
   - u_i: the number of units captured for the first time on occasion i
   - v_i: the number of units captured for the last time on occasion i
   - n_i: the number of units captured on occasion i
2. Multiple model fitting.
3. Model selection using comparison criteria like deviance, AIC, and BIC.
4. Final population size estimation, including bias corrections or interactions between individuals in the population.

Capture data are allowed in two input formats. The first is a matrix of t columns as the number capture occasions and one row denoting the capture history, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_t)$, for each captured individual, where $\omega_j = 1$ indicates a capture on the jth occasion and 0 a miss. The second format is an extension of the first. It has an added column at the end of each capture history indicating the number of times the individual was captured, i.e. the matrix has $t + 1$ columns.

The descriptive statistics, f_i, n_i, v_i, and u_i need to be explored for indications of temporal, behavioural and heterogeneous effects on the capture probability of individuals within the population. Significant variations in n_i, for instance, would suggest that there is a temporal effect in the catchability of individuals. Further evaluations should be performed to probe heterogeneity by using the `plot.descriptive` function that plots two parameters as a function of i which is derived and approximated from a geometric distribution. The number of captures in M_0 is modelled by the geometric distribution until the first capture and subsequently follows a binomial distribution. The frequencies, represented as $\log\left(f_i / \binom{t}{i}\right)$, is approximated as:

$$\begin{aligned} \log\left(\frac{f_i}{\binom{t}{i}}\right) &\simeq \log\left(\frac{N \times \Pr(i \text{ captures})}{\binom{t}{i}}\right) \\ \text{where} \quad \log\left(\frac{N \times \Pr(i \text{ captures})}{\binom{t}{i}}\right) &= \log\left(N(1-p)^{t-i}p^i\right) \\ &= \log\left(N(1-p)^t\right) + i\log\left(\frac{p}{1-p}\right) \\ \therefore \log\left(\frac{f_i}{\binom{t}{i}}\right) &\simeq \log\left(N(1-p)^t\right) + i\log\left(\frac{p}{1-p}\right) \end{aligned} \tag{2.47}$$

and the first time capture units, $\log(u_i)$ approximated by:

$$\begin{aligned} \log(u_i) &\simeq \log(N \times \Pr(\text{first capture on occ } i)) \\ \text{where } \log(N \times \Pr(\text{first capture on occ } i)) &= \log\left(N(1-p)^{i-1}p\right) \\ &= \log\left(\frac{Np}{1-p}\right) + i\log(1-p) \\ \therefore \log(u_i) &\simeq \log\left(\frac{Np}{1-p}\right) + i\log(1-p) \end{aligned} \tag{2.48}$$

where p is the capture probability, t the total number of capture occasions, i the given capture occasion, and N the population size to be determined. In **Rcapture**, log refers to the natural logarithm. Equations 2.47 and 2.48 are analytically determined log-linear functions and should display linear relationships under the reference model M_0. Non-linear relationships of either equations may indicate one or a combination of the three capture probability variations. Baillargeon and Rivest (2007) provide the theoretical closed-form shapes of the curves in Table 2.5 for models featuring different capture probability variations.

Table 2.5: Analytic forms of each closed population model as featured in Baillargeon and Rivest (2007, Tab. 1) that indicate variations in capture probability from a capture history based on the graphical forms of parameters $\log\left(f_i/\binom{t}{i}\right)$ and $\log(u_i)$. L refers to linear, L* to almost linear, and C means a concave upward or convex shape. The question mark symbol (?) means that no definitive form exists.

Graph	M_0	M_t	M_h	M_{th}	M_b	M_{bh}
f_i	L	L*	C	L*/C	?	?
u_i	L	?	C	?	L	C

The main workhorse for implementing the log-linear frameworks across multiple capture probability variations is the `closedp` function. Under each of the heterogeneous and heterogeneous-temporal schemes: M_h, M_{th}, and M_{bh}, several estimators may be compared, namely:

- Poisson2 (Sandland and Cormack, 1984; Rivest and Lévesque, 2001)
- Chao (LB) (Chao, 1987, 1989; Rivest and Baillargeon, 2007)
- Darroch (Darroch et al., 1993*b*; Agresti, 1994; Rivest and Baillargeon, 2007)
- Gamma (Rivest and Baillargeon, 2007)

The `closedp` function produces MLEs using the `R glm` function. The MLEs are derived from the log-linear form through an iteratively reweighed least square algorithm. The population size estimate, N, is obtained via the fitted parameters from the MLEs of a Poisson log-likelihood. The probability of an individual with capture history $\boldsymbol{\omega}$ under the reference model M_0 is given by a multinomial distribution:

$$\Pr(\boldsymbol{\omega}) = (1-p)^{t-\sum\omega_j} p^{\sum\omega_j} \tag{2.49}$$

where $\sum\omega_j$ is the number of times the individual is caught. The expected number of individuals, $\mu_{\boldsymbol{\omega}}$, with capture history $\boldsymbol{\omega}$ is therefore:

$$\mu_{\boldsymbol{\omega}} = N(1-p)^{t-\sum\omega_j} p^{\sum\omega_j} \tag{2.50}$$

Manipulating Eq. 2.50 into the log-linear form, $E(\boldsymbol{Y}) = \exp(\boldsymbol{X}\boldsymbol{\beta})$ (cf. Eqns. 2.6 - 2.7), gives:

$$\begin{aligned} \mu_{\boldsymbol{\omega}} &= N(1-p)^{t-\sum\omega_j} p^{\sum\omega_j} \\ \mu_{\boldsymbol{\omega}} &= \exp\left(\log\left(N(1-p)^{t-\sum\omega_j} p^{\sum\omega_j}\right)\right) \\ \mu_{\boldsymbol{\omega}} &= \exp\left(\underbrace{\log\left(N(1-p)^t\right)}_{\gamma} + \sum \omega_j \underbrace{\log\left(\frac{p}{1-p}\right)}_{\beta}\right) \end{aligned} \tag{2.51}$$

where $\boldsymbol{Y}$ is a $(2^t-1)\times 1$ vector of the observed frequencies $n_{\boldsymbol{\omega}}$ (including zero frequencies), and $\boldsymbol{X}$ is a $(2^t-1)\times 2$ design matrix with a first column of ones and a second column defined by $\sum\omega_j$, and $\boldsymbol{\beta} = (\gamma, \beta)^t$. The population size estimate has the general form of:

$$\hat{N} = n + \exp(\hat{\gamma}) \tag{2.52}$$

where n is the total number of individuals caught in the experiment and $\exp(\hat{\gamma})$ is an estimate of the unobserved individuals (much like in Eq. 2.40). We obtain the population size estimate from γ and the capture probability from β using the linear fitted parameters. In fact, one may show that γ is a true estimator through:

$$\begin{aligned} \exp(\gamma) &= \exp\left(\log\left(N(1-p)^t\right)\right) \\ &= \left(N(1-p)^t\right) \\ &= N(1-p)^t \\ &= N \times \Pr\left(\boldsymbol{\omega}_0\right) \\ &= \mu_0 \end{aligned} \tag{2.53}$$

where $\boldsymbol{\omega}_0 = \sum \omega_j = 0$ is the unobservable capture history for individuals never caught and μ_0 is the expected number of the unobserved individuals. Confidence intervals are determined from the profile log-likelihood. The log-linear form of each of the models fitted by `closedp` is given in Table A.2 of Appendix A (Rivest and Lévesque, 2001; Rivest and Baillargeon, 2007). Model selection criteria (deviance, AIC, and BIC) are provided by `closedp` for comparison of the fitted models.

Capture-recapture estimators described in Otis et al. (1978) are biased. Efforts have been directed at characterising and reducing the biases of MLEs by Firth (1992, 1993) and others. Bias-corrections are applied in **Rcapture** using the `closedp.bc` function. The estimator bias-corrections are determined in two ways: via a closed form asymptotic bias approximation or through *frequency modification*. Rivest and Lévesque (2001) and Rivest and Baillargeon (2007) provide generalisations of the Evans and Bonett (1994) log-linear frequency modifications. The corrections are summarised in Table A.3 of Appendix A.

The design-based modelling offered by **Rcapture** is implemented with the `closedp.mX` function. Where the main `closedp` function has all entries in the first column of the design matrix, X, set to 1, `closedp.mX` allows the user to define the first column, which allows for flexibility in the modelling by imposing linear constraints on the underlying structure. Design-based estimation was not attempted in this work since it was focused on exploratory analysis but will be considered in future applications. The `closedp.h` function can be used in conjunction with `closedp.mX` to account for the degree of heterogeneity by varying the Poisson parameter a accordingly.

2.5.1.2 Robust Design Analysis

The robust design analysis is implemented through two functions: `robustd.0` and `robustd.t`. Both are implementations of log-linear parameterisations as featured in Rivest and Daigle (2004). The `robustd.t` function can fit closed population models M_0, M_t, M_h and M_{th} with a response vector of size $2^{\sum_{i=1}^{I} t_i} - 1$, while `robustd.0` has response vector $\prod_{i=1}^{I} (t_i + 1) - 1$ and only accepts models M_0 and M_h. Parameter t_i stands for the number of capture occasions in primary epoch i.

With the statistical background and context now addressed, we will describe these log-linear models' applications to simulated and real datasets in the subsequent chapters.

Chapter 3

Simulated Populations of High Mass X-ray Binaries

To our knowledge, the only instances of capture-recapture applied in astronomy is by Laycock (2017) in a time-domain context and Romine et al. (2016) in a multi-wavelength context to determine the completeness of the population of protostars in the MYStIX survey. However, in a series of five papers on flare stars in the Pleiades, Ambartsumyan et al. (1970, 1971, 1972, 1973) and Mirzoyan et al. (1977) used similar techniques for capture probability and population size estimation without an explicit mention of capture-recapture.

This chapter reproduces results similar to those encountered in the Laycock (2017) paper. This chapter uses simulations to characterise population estimators based on astronomical parameters and constraints such as observational cadence and the rate of detection in time and as a function of the number of observations. [25] The effect of increased brightness threshold is also explored w.r.t. rate of convergence to the true population size as a function of capture occasion observation k. I explain the steps performed and the motivations for doing so to arrive at those results.

3.1 Simulating the outbursts at periastron passage

Laycock (2017) applied the capture-recapture method in the context of simulations of the X-ray lightcurves of HMXBs. Of the subgroups of HMXBs, BeXRB systems are the most abundant and more than half of all known and candidate HMXBs are

[25] Astronomical sampling *cadence* refers to the time between successive observations, akin to sampling period. However, astronomical data have uneven times between samples, thus the term *cadence* can be described as a distribution of the times between samples. *High* cadence implies shorter time between samples whereas *low* cadence implies the opposite.

located within the SMC (Liu, van Paradijs, and Van den Heuvel, 2005, 2007; Haberl and Sturm, 2016, and references therein). For these reasons their prevalence was used as a motivation to simulate a sample of BeXRB systems that is representative of the HMXB population that one may encounter by using the SMC as an example. The features of BeXRB systems include Type I outbursts which are periodic outbursts at periastron (see §1.2.1). The outbursts in the BeXRB systems can thus be modelled using characteristic orbital periods (Charles and Coe, 2006; Laycock, 2017).

Several scenarios of an HMXB population were considered by Laycock to probe estimator performance of the population size by considering the interaction of parameters such as the 'sampling window' (which we will refer to as cadence from hereon) and the recurrence of outbursts of HMXB. Laycock (2017) created six model distributions from pulsar spin periods, described in Table 3.1, which I have also used in the same methodology to simulate X-ray lightcurves of HMXBs. These models represent instances of possible BeXRB spin period distributions, empirically linked to the orbital period through the relation in Eq. 1.1. Each of the models (A through F) were characterised either by a Gaussian (type G) or a uniform (type U) distribution in $P_{\rm pulse}$. A sample of $N = 100$ pulse periods was randomly drawn from each distribution and the periods transformed to a corresponding orbital period using the empirical relationship in Eq. 1.1. The spin period distributions were transformed to skew log-normal orbital period distributions, except for population model B, which was of Type U and subsequently transformed to a triangular distribution. The distributions are shown in Figure 3.1 with peaks ranging between 130 and 300 days.

Table 3.1: Model A to F used for simulating populations of HMXBs with comparison orbital period distributions, as defined in Laycock (2017). Type G refers to a Gaussian distribution with P_1 as the mean and P_2 the standard deviation, where as Type U refers to a Uniform distribution with P_1 as the minimum and P_2 the maximum bounds. T indicates the median orbital period for each model. Negative values for spin periods are excluded.

Model	Type	P_1 s	P_2 s	T d
A	G	150	50	132
B	U	1	1000	< 311
C	G	0	200	16
D	G	200	100	150
E	G	200	50	150
F	G	0	100	16

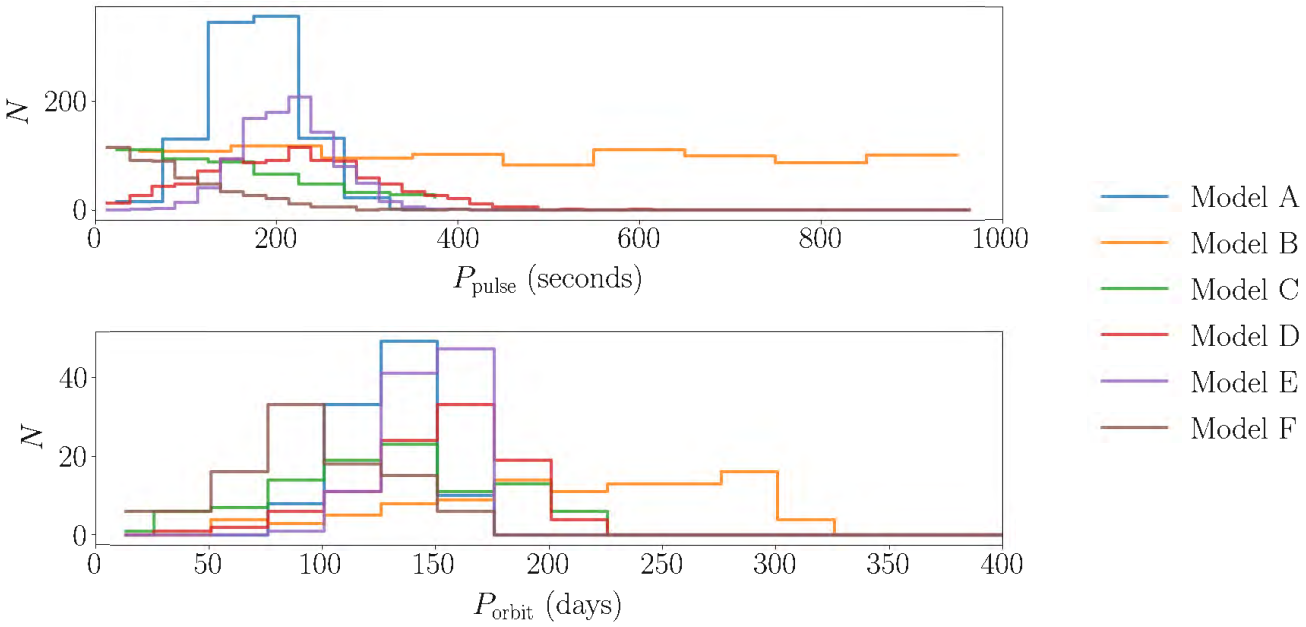


Figure 3.1: Simulated period distribution functions for the pulsar spin period (top) and the $N = 100$ drawn samples that were transformed to BeXRB binary orbital periods (bottom).

Each of the 100 binary orbital periods was used to simulate a BeXRB X-ray lightcurve modelled by a Gaussian-shaped outburst profile which is assumed to reoccur at each periastron passage. The width of the Gaussian-shaped outburst is described by the full-width half-maximum (FWHM) (where FWHM $\approx 2.355\sigma$ where σ standard deviation of the Gaussian). The standard deviation σ was set at a constant 10% of each system's orbital period (which fixes the duty cycle of the sources and hence, the capture probability):

$$\tau_{\rm outburst} = \text{FWHM} \approx 2.355\sigma = 0.2355 P_{\rm orbit} \tag{3.1}$$

A relative flux scale was used for simplicity, with all lightcurves at a zero quiescent flux. This assumes that all systems are located at the same distance and have little variation in luminosity. The amplitude of each outburst was randomly scaled between 0 and 1 from a uniform distribution. The randomness simulates the varying degree to which BeXRB systems outburst based on various factors such as accretion rate (mass transfer rate) and disk size. It allowed for those odd occasions of 'missed' outbursts in BeXRB systems where the geometry of the system varies significantly compared to the regular periastron outbursts (Charles and Coe, 2006) and the outburst is not seen when expected. Each lightcurve signal in the simulation can be described as a sum of Gaussian functions in the time domain by the following equation:

$$\text{Signal form} : \sum_{i}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu_i}{\sigma}\right)^2}; \quad \mu_i \geq t_0$$

where μ_i is the time at the peak flux of the ith outburst, t_0 the simulation start time, and $\sigma = 0.1P_{\text{orbit}} = \tau_{\text{outburst}}/2.355$ the standard deviation related to the outburst cycle in Eq. 3.1.

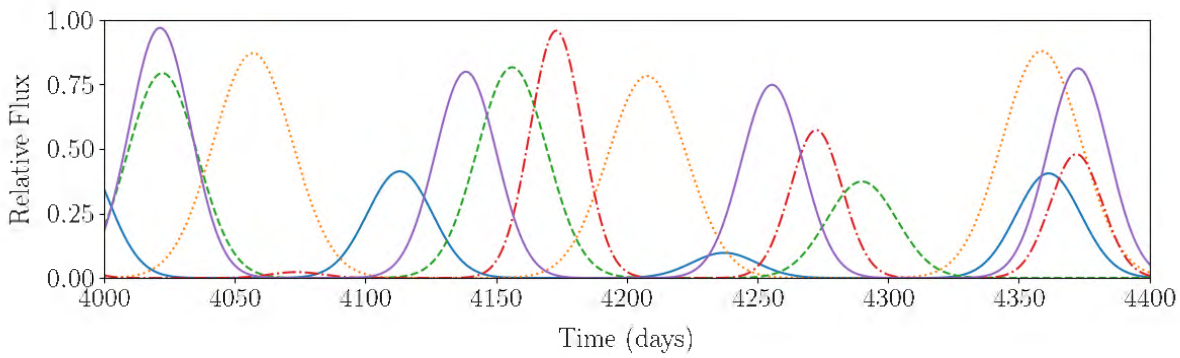


Figure 3.2: Example of a small population of simulated Model A HMXB lightcurves that show periodic outbursts at periastron passage with randomised amplitudes. The lightcurves are plotted for the first 400 days of the simulation.

Models A to F were simulated over a timescale of 4000 days ($\sim$ 11 years). The lightcurves are shown for a few sources in Figure 3.2.

3.2 Sampling the population and observation strategy

The creation of simulated lightcurves was the first step towards creating a dataset for capture-recapture analysis. Astronomical surveys tend to form observing strategies based on several vital scientific requirements, and the strategy is modulated by logistical constraints like spacecraft orbits and/or seasonal visibility. For that reason, the sampling cadence distribution intends to mimic the observation strategy to create a representative capture history. For a set of simulated HMXB populations as denoted in Table 3.1, each lightcurve was sampled within the constraints of a specified cadence distribution. Seven such cadence strategies were implemented for each simulated populations (A to F), and the parameters are provided in Table 3.2.

Laycock (2017) imposed an arbitrary relative flux threshold on the simulated populations (A to F) to emulate the capture methodology for a modelled set of HMXBs. A flux higher than the threshold was defined as a capture, and a flux below defined as a

miss. Capture histories were created by comparing the sampled lightcurves to a chosen relative threshold. This procedure draws the parallel between the capture, tag, and release field experiment for animal population size estimation and its transference to an astronomical context.

Table 3.2: Description of the bounds of the cadence distributions with units in days. The cadence is also given as the fraction of the median orbital period, T, for each model orbital period distribution provided in Table 3.1.

Lower Bound C_{LB} (days)	Upper Bound C_{UB} (days)	Median $\bar{C}$ (days)	Fraction of median model orbital period $\sim \bar{C}/T$					
			A	B	C	D	E	F
7	14	10.5	0.1	≤ 0.1	0.7	0.1	0.1	0.7
15	30	22.5	0.2	0.1	1.4	0.2	0.2	1.4
30	60	45	0.4	0.2	2.8	0.3	0.3	2.8
60	90	75	0.6	0.35	4.7	0.5	0.5	4.7
90	120	105	0.8	0.5	6.6	0.7	0.7	6.6
120	240	180	1.4	0.8	11	1.2	1.2	11

The sampling cadences are defined by uniform distributions with lower and upper bounds, C_{LB} and C_{UB} in days. The time between samples ranges between the bounds and is randomly selected instead of favouring a particular sampling period. The median value, $\bar{C}$ indicates the midpoint of the cadences. It is used to calculate a median sampling period (as a fraction of each simulated population's median orbital period) that will be probed. The parameters are quick identifications of possible aliasing of the estimates w.r.t. observation k.

Table 3.2 probed a range of cadences that allowed the investigation of aliasing effects from the sampling with median orbital periods from each model. Figure 3.3 displays the applied relative flux threshold of 0.2 across the set of simulated lightcurves; chosen arbitrarily with the expectation that with a higher imposed threshold, the units captured per observation will be fewer in number. Increased observation may be needed to recover an estimate to the same accuracy and precision compared to a lower threshold.

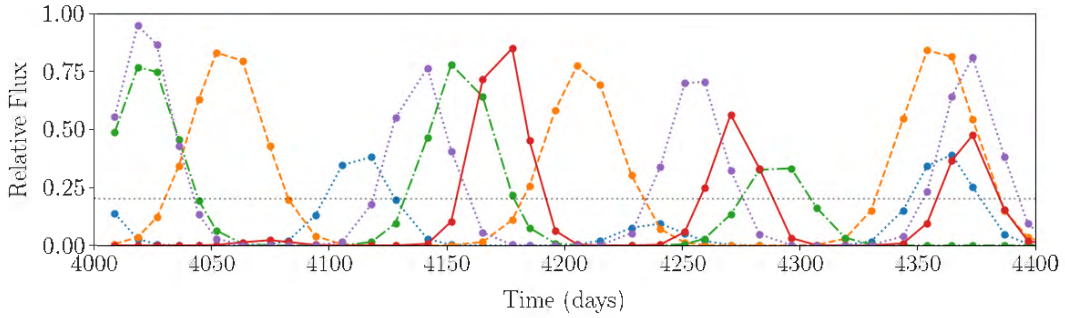


Figure 3.3: Samples (7 to 14 day cadence) of the simulated HMXB lightcurves in shown in Figure 3.2, emulating similar cadencing to *RXTE* (Laycock, 2017). The horizontal line represents a detection threshold that discriminates between a 'capture' or a 'miss' for each observation of a source.

3.3 Implementation of different estimators on Models A to F

3.3.1 Generalised exponential model

The exponential models for continuous and discrete cases were derived in Chapter 2 in Equations 2.42 and 2.43. The cumulative count of identified sources as a function of observation is plotted in Figures 3.4 and 3.5. Each model as specified in Table 3.1 is plotted for capture thresholds at relative fluxes of 0.2 and 0.5 in the respective figures. Seven different cadence strategies are shown for each as defined in Table 3.2.

Each cadence's cumulative captures was fitted with an exponential model as in Equation 2.42 with p as the optimisation parameter. A full table of the model parameters may be found in Table 3.3. The capture probabilities reach $p = 0.33$ (Model A, 30-60 day cadence, $I_{thr} = 0.2$) and hovers between $p = 0.1$ and $p = 0.3$ amongst the models and cadence strategies.

The encounter probability p is consistently smaller with the higher cadences. The 7 to 14-day cadence samples at a small fraction of the median orbital period T of each modelled distribution A to F, and thus encounters fewer outburst occurrences of new sources within the population as a function of observation k compared to the lower cadences. The capture probabilities are strongly affected by an increased relative flux threshold. A flux threshold factor increase of 2.5 (from 0.2 to 0.5) decreases the capture probability by more than a factor of 3 to $p \lessapprox 0.1$ across models and cadences.

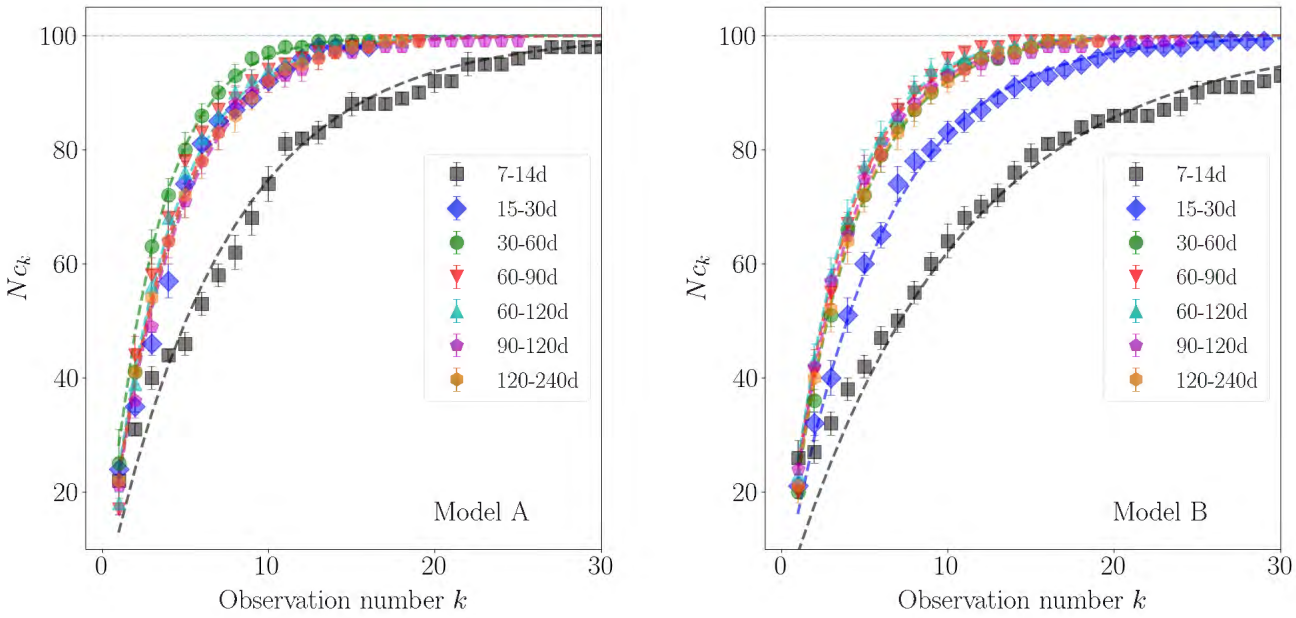


Figure 3.4: Cumulative captures Nc_k plot as a function of observation k (threshold=0.2). They have been modelled by the exponential function in Eq. 2.42. Their fitted parameters are tabulated in Table 3.3. The underlying population size (N=100) is plotted for reference.

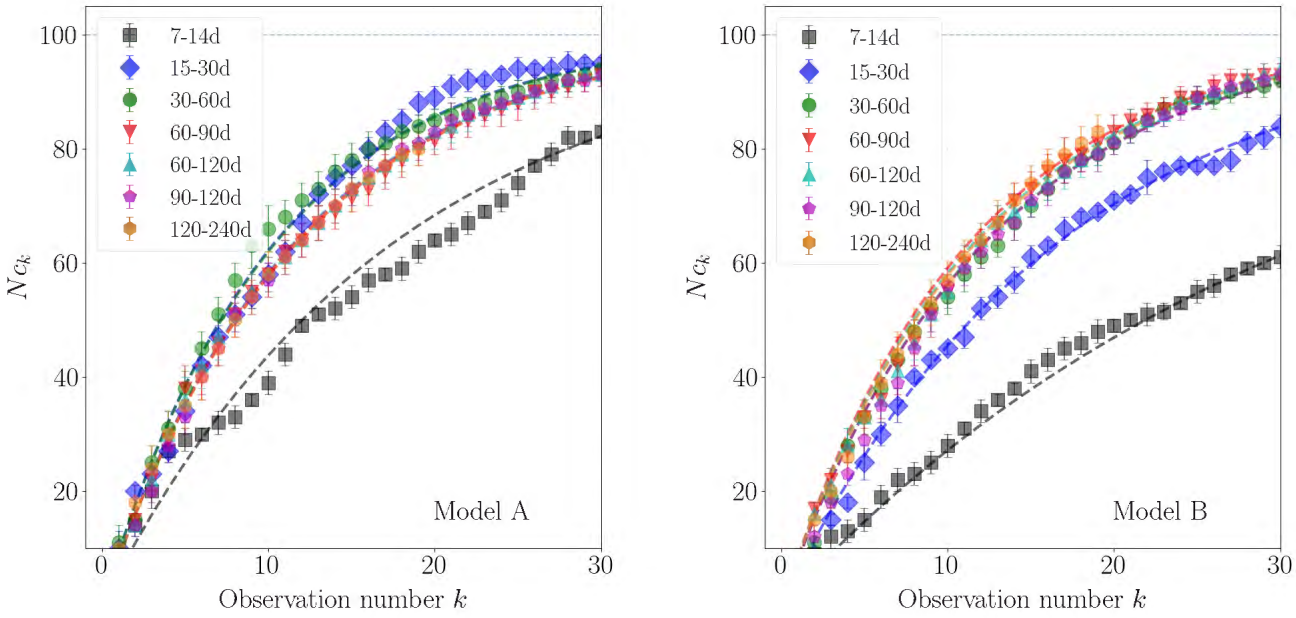


Figure 3.5: Cumulative captures Nc_k plot as a function of observation k (threshold=0.5). They have been modelled by the exponential function in Eq. 2.42. Their fitted parameters are tabulated in Table 3.3. The underlying population size (N=100) is plotted for reference.

Table 3.3: Capture probabilities from the fitted exponential model to the cumulative captures of Model A to F at different relative flux thresholds.

		Flux Threshold			
		0.2		0.5	
	Cadence (days)	p	σ_p	p	σ_p
Model A	7-14	0.137	0.003	0.058	0.001
	15-30	0.253	0.006	0.097	0.001
	30-60	0.329	0.006	0.097	0.001
	60-90	0.261	0.009	0.086	0.001
	60-120	0.270	0.005	0.087	0.001
	90-120	0.246	0.003	0.087	0.001
	120-240	0.253	0.001	0.086	0.001
Model B	7-14	0.097	0.001	0.032	0.001
	15-30	0.175	0.002	0.061	0.001
	30-60	0.252	0.004	0.083	0.001
	60-90	0.281	0.007	0.089	0.001
	60-120	0.289	0.003	0.087	0.001
	90-120	0.265	0.004	0.082	0.001
	120-240	0.256	0.002	0.086	0.001
Model C	7-14	0.152	0.004	0.058	0.001
	15-30	0.251	0.005	0.089	0.001
	30-60	0.294	0.006	0.096	0.001
	60-90	0.275	0.008	0.088	0.001
	60-120	0.255	0.004	0.089	0.001
	90-120	0.246	0.003	0.090	0.001
	120-240	0.245	0.001	0.085	0.001
Model D	7-14	0.154	0.005	0.047	0.001
	15-30	0.278	0.006	0.076	0.001
	30-60	0.307	0.006	0.087	0.001
	60-90	0.247	0.003	0.082	0.001
	60-120	0.268	0.003	0.084	0.001
	90-120	0.276	0.003	0.085	0.001
	120-240	0.232	0.001	0.079	0.001
Model E	7-14	0.101	0.002	0.047	0.001
	15-30	0.199	0.003	0.089	0.002
	30-60	0.284	0.004	0.099	0.001
	60-90	0.274	0.002	0.093	0.001
	60-120	0.284	0.006	0.094	0.001
	90-120	0.284	0.010	0.091	0.001
	120-240	0.239	0.002	0.085	0.001
Model F	7-14	0.181	0.002	0.061	0.001
	15-30	0.284	0.005	0.089	0.001
	30-60	0.274	0.003	0.091	0.001
	60-90	0.269	0.003	0.089	0.001
	60-120	0.235	0.001	0.083	0.001
	90-120	0.215	0.003	0.082	0.001
	120-240	0.245	0.001	0.083	0.001

As expected, this increase in the brightness threshold for characterising a capture vs a miss also increases the number of observations needed to capture the same cumulative number of sources, Nc_k. In Model A at the 7-14 day cadence, k increases from ~ 11 to ~ 28 corresponding to the flux threshold factor increase of 2.5.

3.3.2 Lincoln-Peterson and Chapman indices

The Lincoln-Peterson (LP) and Chapman (a.k.a. modified LP as referred to in Laycock 2017) indices are basic dual occasion estimators (refer to §2.3, Seber 1982; Lincoln 1930; Petersen 1894; Chapman 1952). Either estimator can, at most, only take into account two consecutive capture occasions and does not distinguish between previously marked or unmarked individuals in follow-up observations. The estimate is calculated based on the current and the previously observed captures and effectively has no 'memory'. Figure 3.6 provides the results of the model A simulation which has been sampled at a 7 to 14-day cadence, roughly at a fraction of 0.1 of the median orbital period of the distribution ($T \sim 132$ days) and using a relative threshold of 0.2. The total individual captures are displayed at every capture occasion in the white squares, which typically range between 20 to 40 individuals at any given time across the 400-day epoch. The black squares indicate the Chapman indices, which estimate the population size based on the current and the previous sample captures.

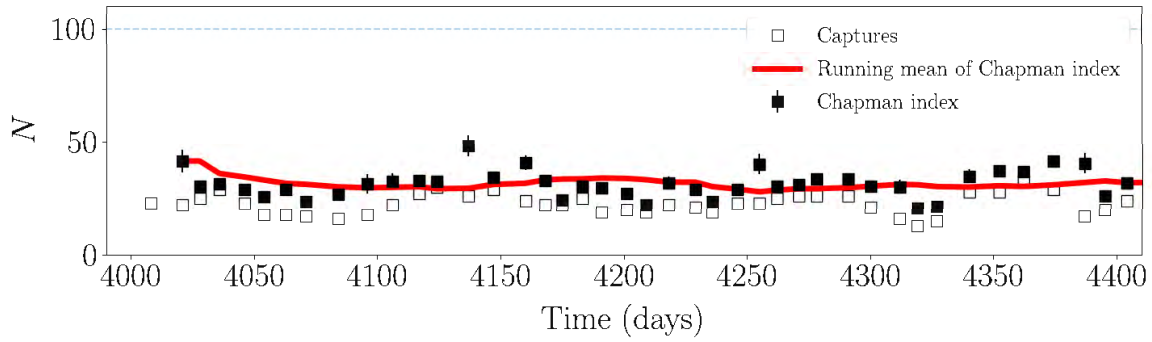


Figure 3.6: Figure of the individual captures and the Chapman estimates from the Model A simulation sampled at a 7 to 14 day cadence. The running mean of the Chapman index across a window of 10 samples is plotted in red.

The Chapman index severely underestimates the underlying population size (known and plotted at $N = 100$ for reference) for the entire duration of sampling as evident in Figure 3.6. The dual estimator lacks a stored history of the already encountered and newly encountered individuals – something that can be improved by implementing a multiple occasion capture-recapture sampling design and estimator as discussed in §2.3.1.

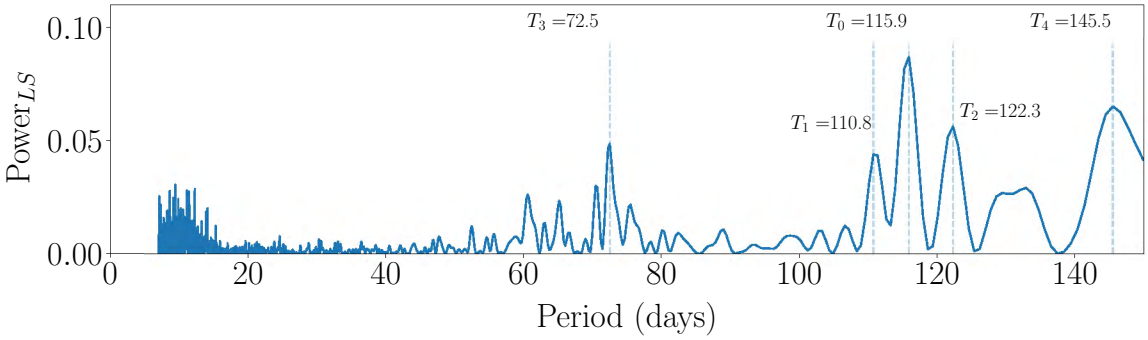


Figure 3.7: Lomb-Scargle periodogram for the captures which vary between 20 and 40 at any given time. Three periodicities are visible at $T_0 = 116$, $T_1 = 146$, and $T_2 = 73$ days.

Figure 3.7 provides a Lomb-Scargle periodogram of the Model A (7 to 14 day cadence) time-series captures that shows an underlying periodic trend present, with peaks at $T_0 = 116$, $T_1 = 146$, and $T_2 = 73$ days. These result from aliasing of the outburst occurrences of the sources in the population with the sampling cadence. Note the sampling distribution profile from 7 to ∼15 days because, within this cadence, it happens more often that a given source is sampled more than once on the same periastron outburst. For this Model A and specified cadence (7 to 14 days), one can also observe the 'periodic' trend in the cumulative captures Nc_k (Fig. 3.4, left) about the fitted model. This periodicity is further reflected in the Chapman estimates as with the captures. A running mean across ten observations smooths out the variability to hover between 28 and 32 across the given timescale, which provides some evidence for the assumption of a binomially distributed average in Eq. 2.40.

Grouping capture occasions can improve underestimates of the Chapman index. Note that this is not equivalent to a lower cadence sampling. Cadenced sampling reflects observations of the *entire* population at known sampling times. The grouping of observation occasions is a binning exercise that allows more captures to accumulate within an allotted time frame and approaches the true population size as more observations are grouped. For example, if we observed an individual across ten occasions and its capture history was noted in time as `{0001000000}`. The merged capture history across every five occasions would be pooled by a logical OR operation to `{10}`. Figure 3.8 illustrates this for the same dataset as in Figure 3.6. The three graphs show every $n = 5$, 10 and 20 occasions with their captures grouped. The Chapman index for these grouped observations already predicts the true size quite well when $n = 10$ captures were merged, whereas, for $n = 5$, there remains a large variation in the estimate as a function of time. For $n = 20$, the captures and recaptures of individuals vary so little

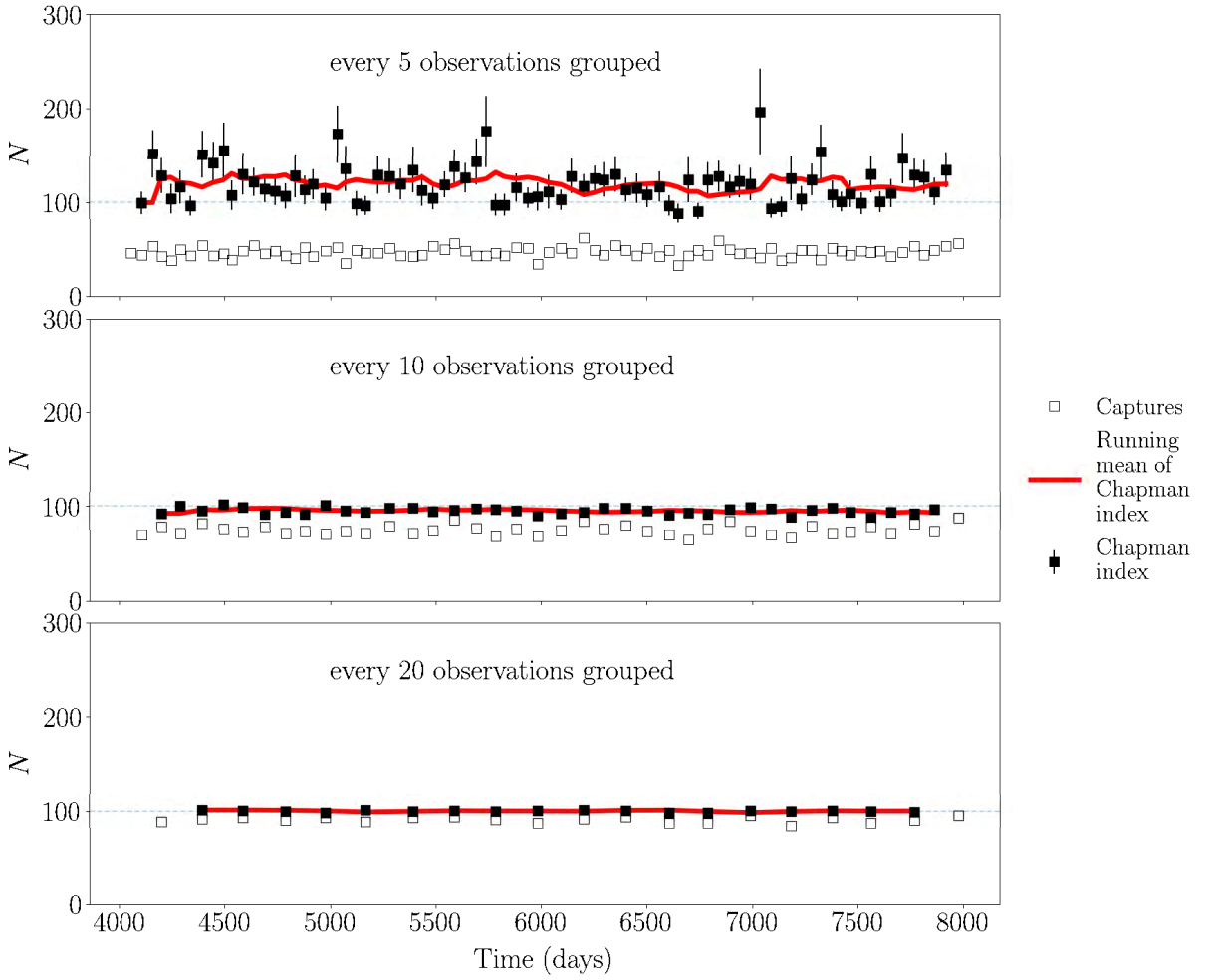


Figure 3.8: Grouped observations of the over-sampled dataset as in Figure 3.6 illustrates less biased estimates of the true population size.

from one occasion to the next that the estimates have high accuracy and precision.

3.3.3 Schnabel and Schumacher-Eschmeyer Estimators

The Schnabel estimate exploits capture information better than its two-sample predecessors. It stores the captures, the number of individuals already encountered, and the individuals re-encountered at each occasion. The Schnabel and Schumacher-Eschmeyer estimators are quite similar in their results, as shown in Figure 3.9 (Schnabel on the left and Schumacher-Eschmeyer on the right) when the population is sampled but once. Both estimators exploit the same underlying assumptions using different regression techniques. The estimates are shown with the solid line and the cumulative count with the dashed line. The colours represent the various cadence strategies applied to the model A simulation.

We observe that the simulated sources ($N = 100$) are captured within 2000 days, and all cadence strategies converge to the true population size from about 500 days ($\pm 20\%$) and

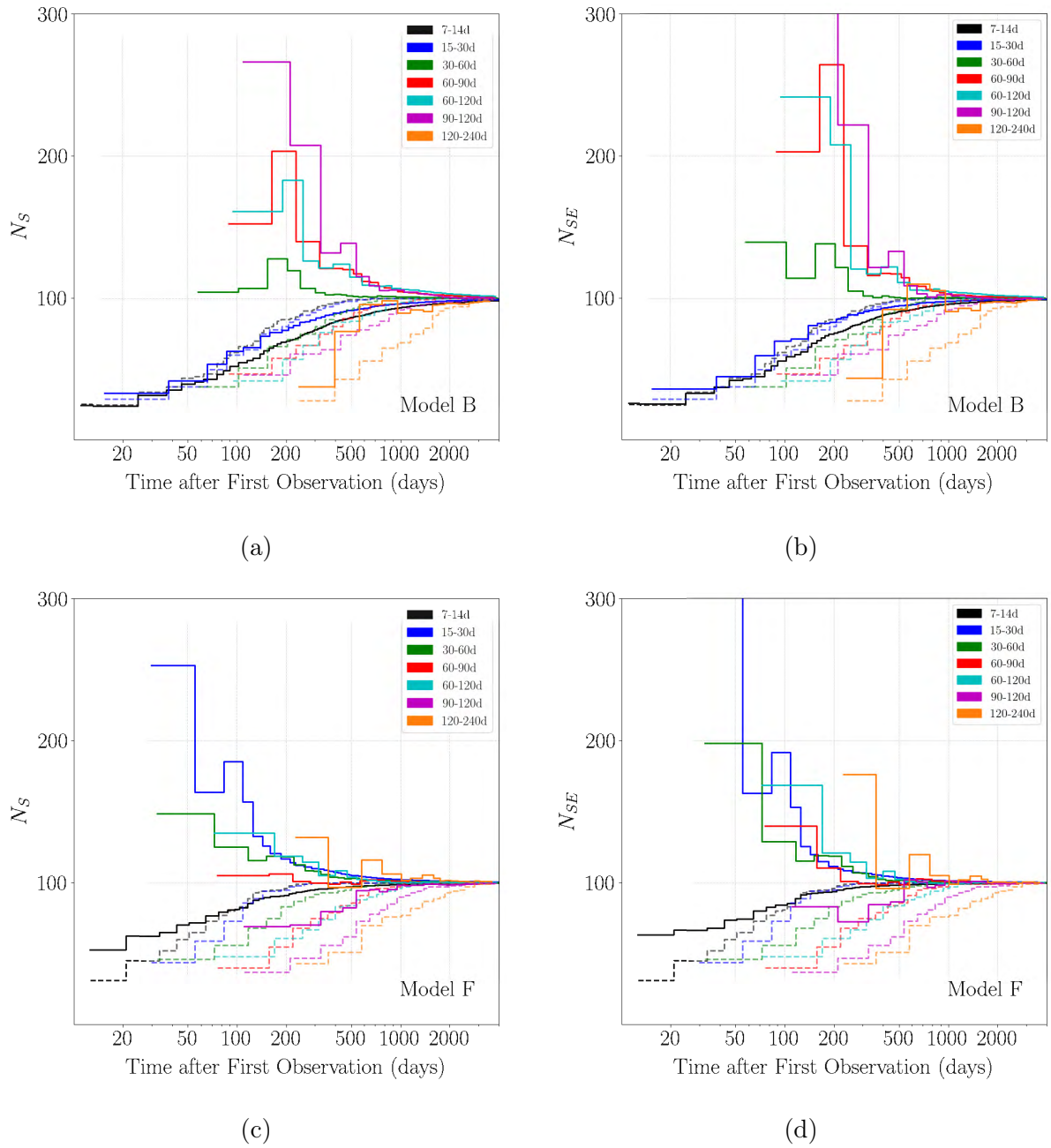


Figure 3.9: Schnabel and Schumacher-Eschmeyer estimates (solid lines), and the cumulative count (dashed lines) when sampled from a simulated population model for various cadences above a 0.2 detection threshold. Both estimators for the various sampling strategies tend to converge to the true population size of $N = 100$ faster than the cumulative count with the exception of the 7 to 14 day cadence. The estimators' convergence is non-monotonic in cases where there is aliasing present between the model median orbital period and the sampling cadence. Convergence to the true population size is reached within 5% after ~ 2000 days. The error on all estimates were consistently $\ll 1$.

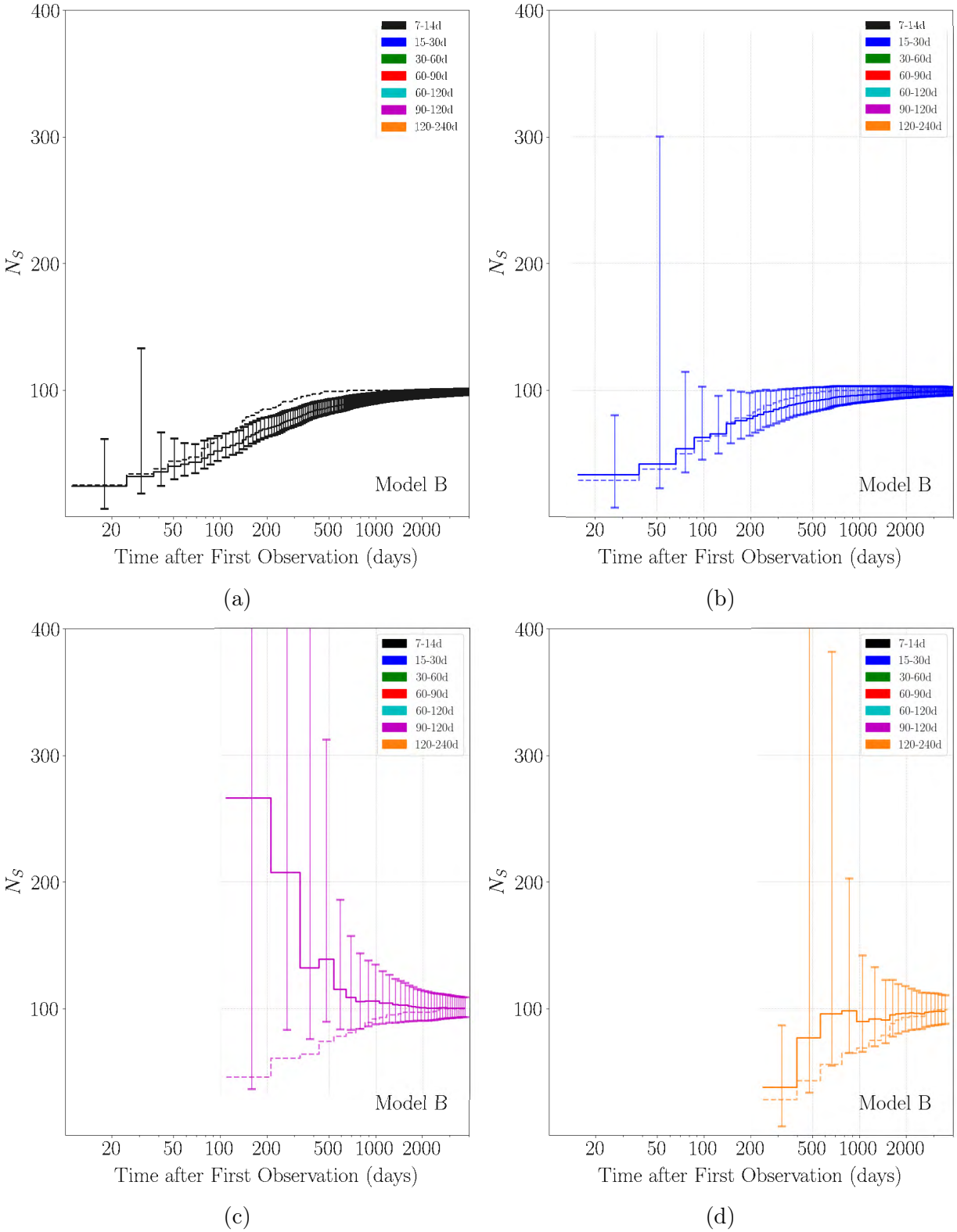


Figure 3.10: Schnabel (solid lines), and the cumulative count (dashed lines), as plotted in Figure 3.9, for Model B along with their 95% confidence intervals. The two highest and two lowest cadences are shown.

beyond. The model A 15 to 30-day cadence strategy converges much faster compared to other cadences. Schnabel and Schumacher-Eschmeyer reach within 5% of the true population size around 200 days because of significant aliasing of the sampling strategy with the common orbital periods. Model A has a median orbital period around 132 days, with the bulk in the range of 100 to 150 days, hence sampled in the $0.1P_{\rm orbit}$ to $0.3P_{\rm orbit}$ range. The periastron outbursts ($\sigma = 0.1P_{\rm orbit}$) are of similar duration to the 15 to 30-day cadence for Model A, which allows for efficient capturing of new sources and periodic recapture of known sources. The high cadences tend to underestimate the population size, whereas the low cadences tend to overestimate early on (though often stillwithin 95% confidence; cf. Model B cadences plotted in in Figure 3.10); since many new individuals are encountered without recapturing a large portion of previous observations. However, they too converge to within 5-10% of the true size between 200 and 500 days.

Each simulated HMXB model was resampled by drawing a new set of cadences from the seven distributions a total of 1000 times. The median of the Schnabel estimates and their 75% confidence intervals, obtained from the resampling of the population models, were plotted as a function of observation number k for models A through F in Figures 3.11 and 3.12 (for brightness thresholds of 0.2 and 0.5). Similar to Figures 3.4 and 3.5, the highest cadence of 7 to 14 days converges slower to the true population size, and even underestimates the cumulative count at later times, because of the much lower associated capture probability. Observations at high cadence also tend to violate independence between population measurements, which the homogeneous Schnabel and Schumacher-Eschmeyer estimators are ill-equipped to correct for. Multiple of the chosen cadences have a spread of 30 days, and these seem to consistently display similar capture probabilities compared to the lower p of the 7-14 and 15-30 day cadence. The high-cadenced sampling relative to the orbital period is ultimately unreliable in estimating the population size when inspected across the various models. It follows that the Schnabel and Schumacher estimators perform best from data with a large spread cadence distribution, and that it optimises estimator efficiency and accuracy.

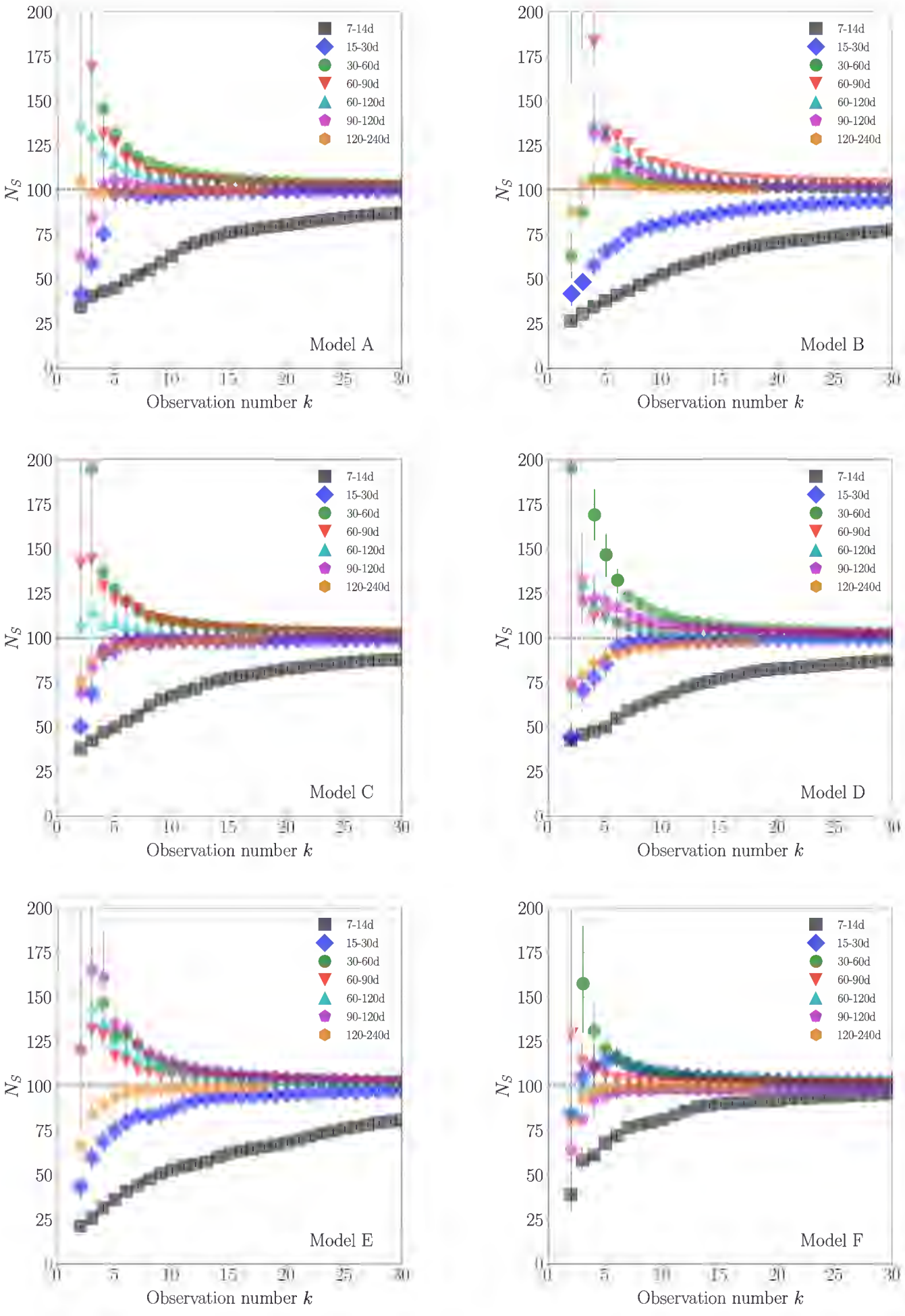


Figure 3.11: Schnabel estimates as a function of observation number k for each simulated HMXB model (threshold=0.2). The cadences are shown in their respective colours for the cumulative count Nc_k.

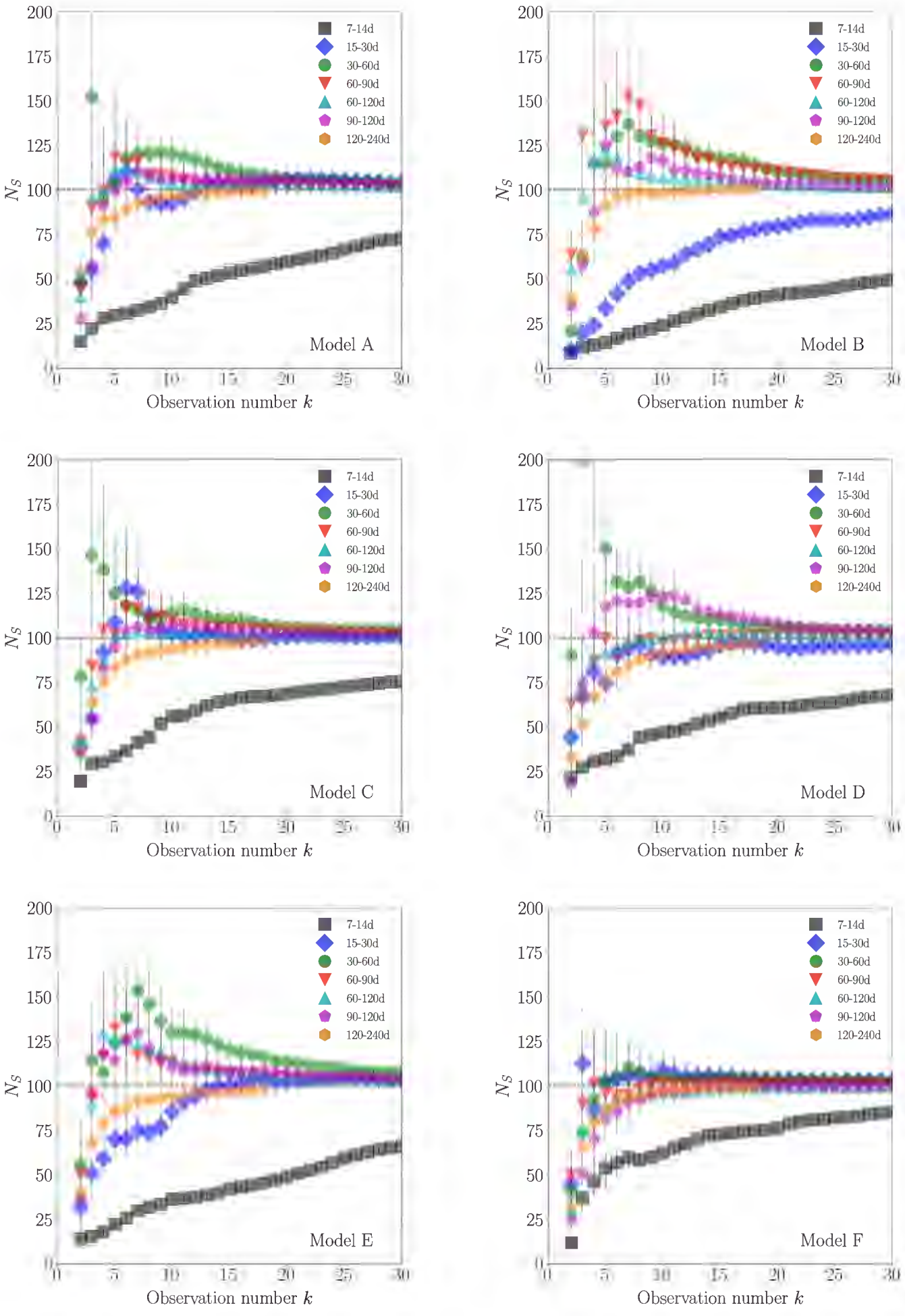


Figure 3.12: Schnabel estimates as a function of observation number k for each simulated HMXB model (threshold=0.5). The cadences are shown in their respective colours for the cumulative count Nc_k.

3.3.4 Models M_0, M_h, M_t, and M_b with Rcapture

Modern capture-recapture software, such as **Rcapture** (Rivest and Baillargeon, 2019), use a combination of estimation approaches such as generalised logistic regression, log-linear and maximum likelihood estimation (MLE) for estimation. The estimators investigated thus far were based on pure regressive methods and assumed equal capture probability amongst individuals. In this section, we have explored a wider variety of models provided by **Rcapture**:

- M_t, which consider capture probability as a function of capture occasion t,
- M_b, which evaluates whether samples are correlated and behave with the so-called 'trap-happy' or 'trap-shy' response, or,
- M_h, the heterogeneous model, accounts for a unique capture probability of each individual.

These concepts have been explained in §2.3 and also refer to the combinations thereof. Models M_{tb} and M_{tbh} are not represented within the standard **Rcapture** `closedp` function since they cannot take a log-linear form, however M_{tb} is available for estimation through non-linear optimisation algorithms. Descriptive statistics (f_i, u_i, v_i, and n_i; cf. §2.5.1) are retrievable through the `descriptive` function from the capture history. The descriptive statistics for the first 15 observations of Model A (sampled at a 7 to 14-day cadence, threshold = 0.2) is provided in Table 3.4. The shape of f_i vs i, shown in Figure 3.13, hints at a Poissonian-type distribution which motivated the use of Poisson regression for estimation.

Rcapture offers inspection of the capture data for signs of heterogeneity based on form of the parameters $\log\left(f_i/\binom{t}{i}\right)$ and $\log(u_i)$ which are either linear, concave, or unspecified (refer to §2.5.1). Figure 3.14 inspects the descriptive parameters for signs of heterogeneous capture probability present for the given dataset (Model A, sampled in a

Table 3.4: Descriptive statistics from Rcapture of Model A at a flux threshold of 0.2 and sampled in a 7-14 day cadence

Individual captured: $n = 88$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
f_i	4	10	21	22	21	9	1	0	0	0	0	0	0	0	0
u_i	21	11	12	0	6	4	6	3	6	6	6	1	1	3	2
v_i	1	0	1	2	3	4	5	7	6	4	6	10	4	12	23
n_i	21	23	30	23	18	17	18	17	17	22	26	29	27	30	23

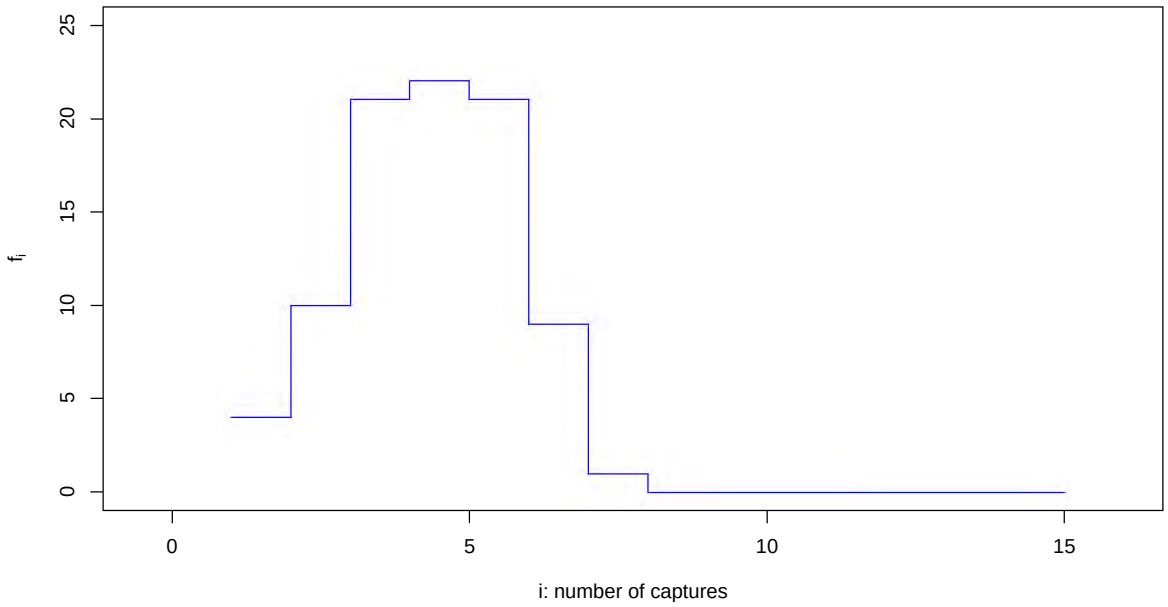


Figure 3.13: Frequency distribution of f_i, the number of individuals captured i times for a total of 15 observations. This is suggestive of a Poissonian distribution.

7 to 14-day cadence, 0.2 threshold) based on the criteria laid out in Table 2.5. In this instance, $\log\left(f_i/\binom{t}{i}\right)$ vs i takes on a concave form whereas $\log(u_i)$ is more linear. This suggests that either M_{th}, M_b, and perhaps M_{bh} may be suitable models for estimation of this particular set. M_b is expected to perform better when compared to Schnabel or M_0 since the higher 7 to 14-day cadence is expected have higher degrees of correlation between samples based on the length of the time of outbursts in Model A ($\sim$ 15d).

Closed population results (application of the `closedp` function for observations $k = 1$ to $k = 15$) shown in Table 3.5 may be compared to the Schnabel estimates (w.r.t. k) in Figures 3.11 and 3.12. The size estimate, $\widehat{N}$, and the standard error of the mean, $\sigma_{\widehat{N}}$, are given in the first two columns. We are also provided with model selection criteria i.e. AIC and BIC, as well as the goodness of fit parameter, deviance, and a mention of the degrees of freedom (dof). For a total of 15 samples with 88 individuals captured, we note that the M_b model has a favourable prediction of $M_b = 98.3 \pm 5.9$ individuals i.e. a $1.7^{+4.2}_{-7.6}\%$ deviation from the true value of $N = 100$. M_{bh} also agrees but has a significantly higher standard error associated with the mean. These two models score lowest on the AIC and BIC model selection criteria which also suggests that they are the most appropriate of the given selection. Where heterogeneity is concerned – there are four different implementations namely: Chao (LB), Poisson2, Darroch, and

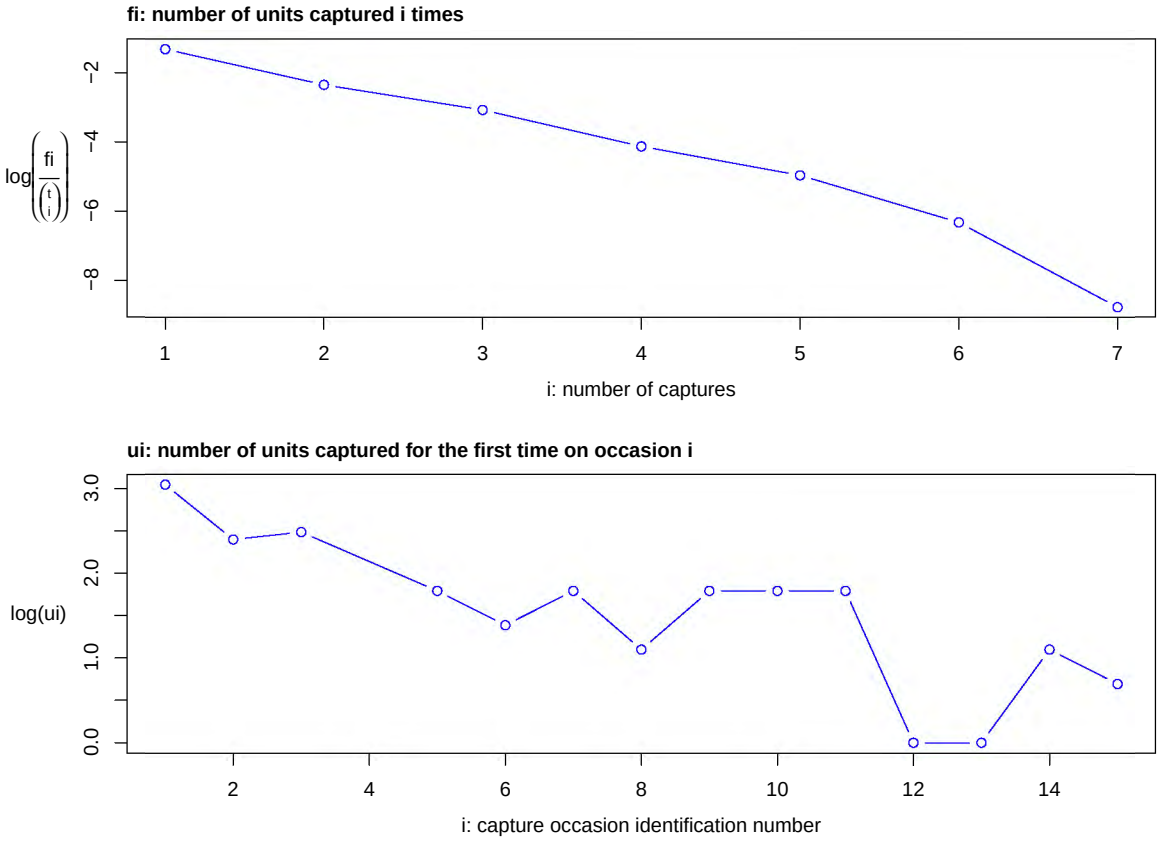


Figure 3.14: Exploratory heterogeneity graph for the capture data of Model A at a 7 to 14 day sampling cadence.

Gamma. In this example, models M_h and M_{th}, amongst those four implementations are very similar in estimation and standard error of the mean. Notably, the M_h and M_{th} models, as well as M_0 and M_t underestimate the true size at $t = 15$.

Another evaluation for model selection was done by comparing each fitted model to the observed u_i parameter as a function of capture occasion i. This is shown in Figure 3.15. The observed newly encountered individuals (u_i) are plotted in the grey triangles. The observed u_i are somewhat erratic. The M_{th} models seem to weigh the initial capture occasions, whereas M_b and M_{bh} gives higher weighting to the new individuals captured on the later occasions – allowing for possible new individuals to be encountered on a future occasion. This also partly explains why M_b and M_{bh} return higher estimates in comparison to the M_{th} models.

Furthermore, fit statistics such as χ^2 and a comparison of the mean ($\bar{u}$) and variance

Table 3.5: Closed population size estimates and model fits with **Rcapture** for Model A 7-14 day cadence for t=15 samples.

Number of captured units: 88

Model	$\widehat{N}$	$\sigma_{\widehat{N}}$	deviance	dof	AIC	BIC
M_0	89.1	1.1	803.132	32765	943.478	948.432
M_t	89.0	1.1	784.202	32751	952.547	992.185
M_h Chao (LB)	89.1	1.1	803.132	32764	945.478	952.910
M_h Poisson2	88.5	0.7	794.187	32764	936.533	943.965
M_h Darroch	88.3	0.5	796.593	32764	938.939	946.371
M_h Gamma3.5	88.2	0.5	797.878	32764	940.224	947.656
M_{th} Chao (LB)	89.0	1.1	784.202	32750	954.547	996.662
M_{th} Poisson2	88.4	0.7	775.686	32750	946.032	988.147
M_{th} Darroch	88.3	0.5	778.060	32750	948.405	990.520
M_{th} Gamma3.5	88.2	0.5	779.303	32750	949.649	991.764
M_b	98.3	5.9	771.472	32764	913.818	921.250
M_{bh}	106.5	29.4	765.824	32763	910.169	920.079

($\sigma^2_{\bar{u}}$) for the model fits to the u_i parameter also favour the M_b and M_{bh} models. What was also considered, because of the small frequency sample sizes (which are only ever as high as $f_i = 22$, refer to Table 3.4), was to apply a bias correction to each size estimate (refer to Table A.3 in Appendix A). These were implemented with the `closedp.bc` function for the models in Table 3.7. M_{bh} has been most affected by this correction through the significant reduction of the standard error $\sigma_{\widehat{N}}$. However, the remaining models showed no significant bias corrections.

If the sampling cadence contained a resonance period of the median/mode of the period distribution of the simulated HMXB model, then fewer than ten observations were typically needed to reach the true population size within 20%. However, it was often not much better than the cumulative individuals' count, with the 7 to 14-day cadence as an example. These models that allow for variation in the capture probability show that with model selection criteria considered, the number of observations for estimation may be reduced consistently to around $10 \leq k \leq 12$ for the 7 to 14-day cadence. In Figures 3.16 and 3.17, the **Rcapture** models are plotted in similar fashion to the Schnabel plots to illustrate convergence of the estimator. However, these figures only consider the estimate and standard error without considering other model selection criteria. Thus, they can inform better or worse models but should not solely direct the choice of model.

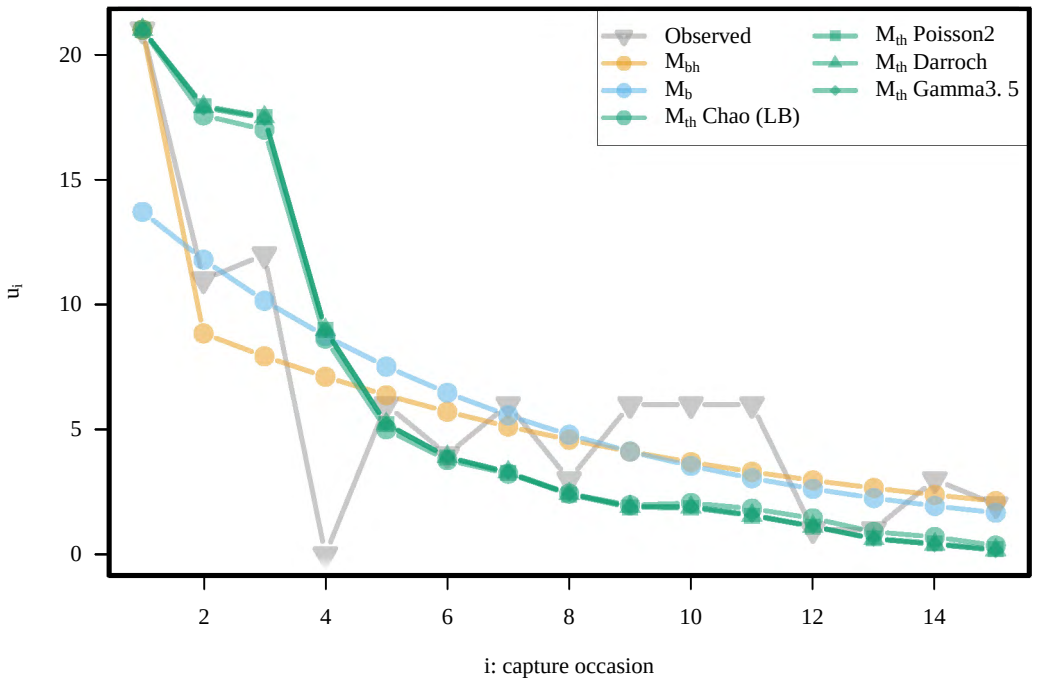


Figure 3.15: Models M_{th}, M_b, and M_{bh} models fitted to the observed u_i parameter as function of capture occasion i.

Table 3.6: Models M_{th}, M_b, and M_{bh} fitted to the newly observed individual u_i for occasions i up to 15.

Model	Goodness of fit to u_i	Day of first capture	
	χ^2	$\bar{u}$	$\sigma^2_{\bar{u}}$
observed	–	5.41	17.5
M_{th} Chao (LB)	55.5	3.81	9.73
M_{th} Poisson2	76.2	3.65	8.44
M_{th} Darroch	80.9	3.65	8.35
M_{th} Gamma3.5	81.0	3.66	8.40
M_b	22.7	5.41	14.7
M_{bh}	18.0	5.41	17.1

Table 3.7: Bias corrected estimates for selected models of Table 3.5.

Model	$\widehat{N}_{bc}$	$\sigma_{\widehat{N}_{bc}}$
M_{th} Chao (LB)	91.5	3.0
M_{th} Poisson2	89.0	1.1
M_{th} Darroch	89.0	1.1
M_b	97.9	5.7
M_{bh}	106.0	10.6

Model B may be considered the most unusual of the simulated models because of its uniform orbital period distribution and how, together with the sampling cadence, it affects the estimation. The high-cadence (7 to 14-day or $\leq 0.1T$; low capture probability $p \sim 0.1$, cf. Table **??**) in Model M_{bh} proved unstable as several estimates w.r.t. k returned $N < 0$ and others did not converge (these were removed from Figure 3.17). However, after $\sim 6-8$ observations, most of the models and cadence strategies approach within 20% of the true population size within the standard error bounds. The data suggest that a high and narrow sampling cadence is not the best strategy for this application type. Capture-recapture works best under the conditions of homogeneity; in other words, the population sufficiently mixes between each capture occasion to approximate individuals' capture probabilities to be equal. This condition seems to be violated to a larger degree for high in contrast to lower cadences; since captures of new individuals occur at a much slower rate observing rate which leads to underestimates of the population size.

Having gone through all of the above to illustrate the inefficiency of high-cadenced sampling with a small spread, a quick modification to the data with the **Rcapture** `periodhist` function downsamples the capture data that effectively improves the capture probability. The function requires user dictation on how the capture occasions should be grouped. The `periodhist` function is analogous to the merging of capture occasions as was performed manually for the Chapman estimates in §3.3.2. The observation grouping was demonstrated for the 7 to 14-day cadence. Fifteen capture occasions were merged into three capture windows (5 occasions for each window). Table 3.8 presents the results for this merged dataset.

Note that M_0, M_t, M_b, and M_{bh} all estimate the true population size within the standard error. Interestingly, M_h Chao and M_{th} Chao also produce suitable estimates. The other M_h and M_{th} models, however, still underestimate to a similar degree as before, but provide lower (i.e. better) scores for the deviance, AIC, and BIC in comparison to

Table 3.8: Closed population size estimates and model fits for merged capture occasions (3 windows, 5 occasions each) for Model A 7-14 day cadence for $t = 15$ samples.

Number of captured units: 88

Model	$\widehat{N}$	$\sigma_{\widehat{N}}$	deviance	dof	AIC	BIC
M_0	101.0	5.1	14.337	5	48.440	53.394
M_t	100.8	5.1	12.351	3	50.454	60.363
M_h Chao (LB)	101.0	5.1	14.337	5	48.440	53.394
M_h Poisson2	91.3	2.6	5.547	4	41.650	49.082
M_h Darroch	89.6	1.8	5.547	4	41.650	49.082
M_h Gamma3.5	88.7	1.2	5.547	4	41.650	49.082
M_{th} Chao (LB)	100.8	5.1	12.351	3	50.454	60.363
M_{th} Poisson2	91.3	2.6	3.713	2	43.815	56.202
M_{th} Darroch	89.6	1.8	3.713	2	43.815	56.202
M_{th} Gamma3.5	88.8	1.2	3.713	2	43.815	56.202
M_b	101.2	7.7	14.335	4	50.438	57.870
M_{bh}	102.1	16.7	14.329	3	52.432	62.341

the models that accurately estimate the true population size. A confirmation with the exploratory heterogeneity plot solved this matter from the linear plots of $\log\left(f_i/\binom{t}{i}\right)$ and $\log(u_i)$ as functions of i. The test rules out models M_h and M_{th} as suitable candidates. We expected little to no heterogeneous effect in the simulated populations since the duty cycles for all sources have been fixed to the same value as in Equation 3.1. The example of the merging of capture occasions significantly improves the high cadence sampling estimation and can be used as a tool for oversampled datasets.

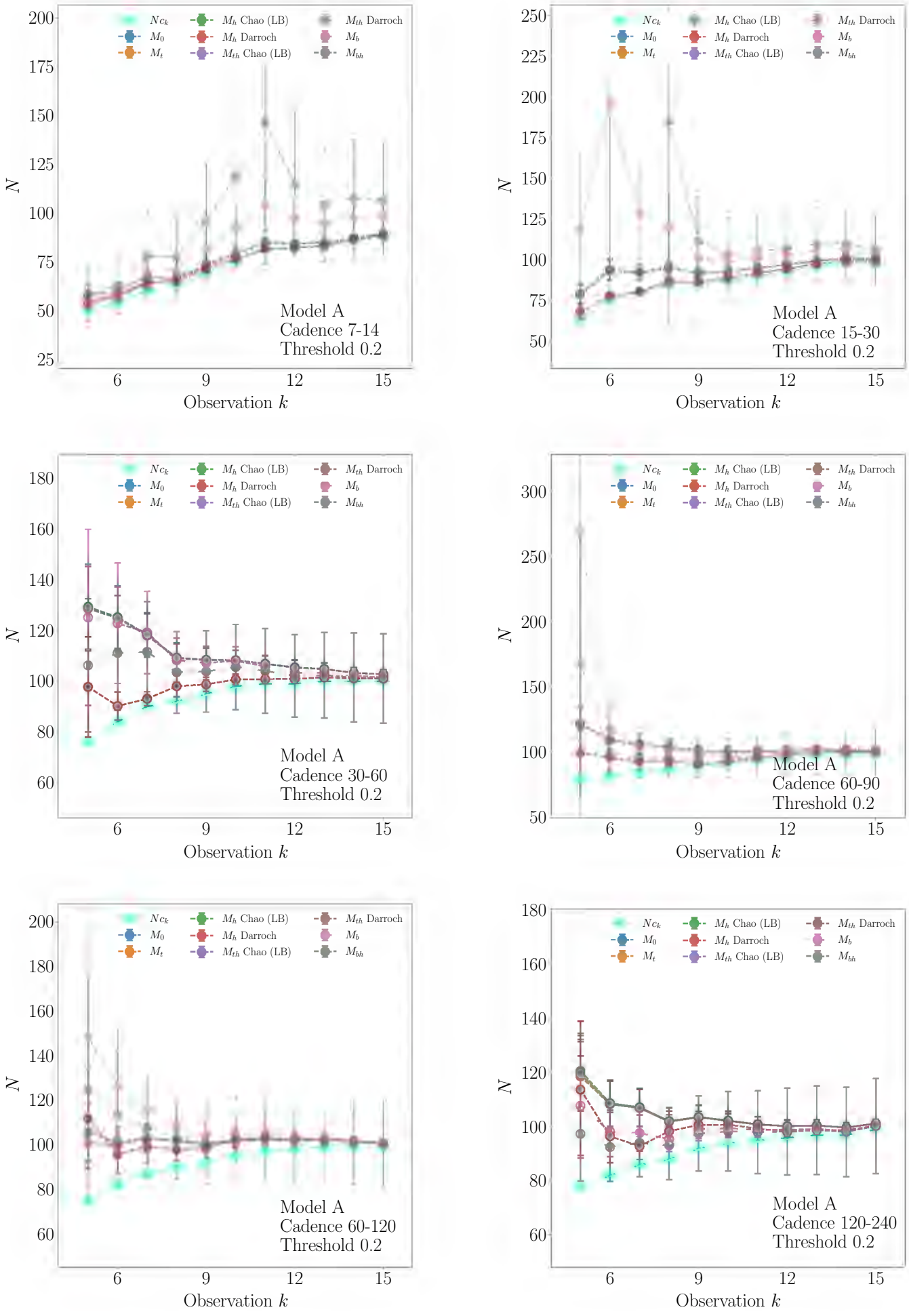


Figure 3.16: Estimates of **Rcapture** `closedp` models fitted for simulated X-ray lightcurves at the same cadences probed for Schnabel and Schumacher-Eschmeyer estimators (Model A, threshold = 0.2).

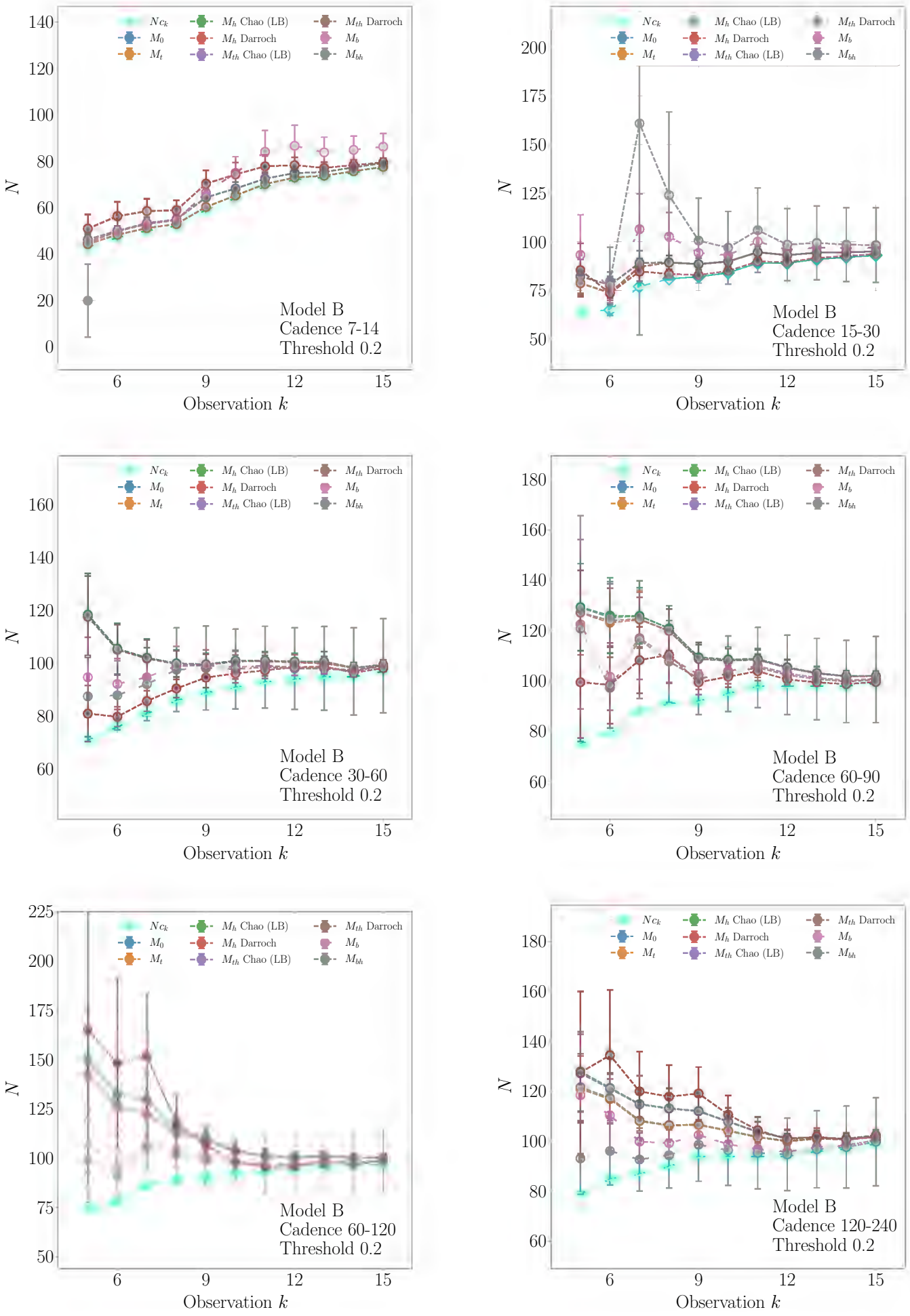


Figure 3.17: Estimates of **Rcapture** `closedp` models fitted for simulated X-ray lightcurves at the same cadences probed for Schnabel and Schumacher-Eschmeyer estimators (Model B, threshold = 0.2).

3.4 Estimates as a function of simulation variables

For most of this chapter, we have assumed key parameters and variables that have been worked into the HMXB simulations, namely:

(a) the duration of the transient outburst (width of the Gaussian profile),

(b) the applied threshold for discriminating between a capture and a miss.

The outburst duration is effectively the duty cycle, the fraction of the orbital period of the binary spent in outburst. Sampling or observation of the set of individuals provides us with the encounter probability, p. As we have seen with the 7 to 14-day cadence, the low encounter probability requires significantly more observations for the Schnabel to recover the true population size. The same may be said of a higher imposed threshold in that more samples will be needed because the probability of encounter is decreased. However, the simulations A to F are rather basic and assume that all the HMXBs have the same minimum relative flux of zero. The simulation is created so that all of the population sources have occurrences where they disappear below the imposed flux threshold. The case is not the same for optical counterparts of HMXBs which will be discussed further in Chapters 4 and 5.

Chapter 4

Application to astronomical population datasets: Part I

In this chapter I present the reduction, analysis and results of the OGLE-IV X-ray variables OGLE Monitoring (XROM) dataset (Udalski, 2008) of the optical counterparts of HMXB sources in the SMC.

Chapter 3 dealt with the exploration of simulations of the X-ray lightcurves of six modelled populations of HMXBs, where each population was constructed from a different binary orbital period distribution. These sources exhibited outbursts on a relative flux scale between 0 and 1, with a zero quiescent flux. The median fluxes of the sources within each modelled simulation were quite similar to each other instead of being broadly distributed, resulting from the pre-and post-outburst zero offset characteristic attributed to all. The individual encounters for each of these models were consequently well described by an equal probability of capture at a given occasion, given sufficient observations ($i : 10 \rightarrow 15$ to within 20% of the true population size), with a standard M_0 model and predecessors like the Schnabel and Schumacher-Eschmeyer estimators. We have seen that high-cadenced sampling suffered from a slow rate of convergence (w.r.t. k) to the true size of the population due to correlation between successive samples with the Schnabel, Schumacher-Eschmeyer and the **Rcapture** M_0 estimators. In certain cases, this could be remedied by merging capture occasions, thus increasing the homogeneity of the capture probability within the population. Exploratory tests of the sampled data, from within the **Rcapture** software, showed that deviations from the homogeneous linear forms of $\log\left(f_i/\binom{t}{i}\right)$ and $\log(u_i)$ vs i, could potentially be addressed with models that allow for heterogeneous capture probability within the population. (Baillargeon and Rivest, 2007; Rivest and Baillargeon, 2019)

In reality we may see broadly distributed quiescent fluxes for sources from the same population within the same spatial vicinity. This stems from inherent variations from the inner physics of stars, thus affecting their stellar luminosities as well as distance and interstellar absorption playing a role. This is evident for the XROM dataset (refer to Figure 4.1). We can confirm this through calculation by using the assumption (see Antoniou et al., 2010; McBride et al., 2008) that the spectral distribution of BeXRB optical counterparts peak at a B0 classification with a luminosity class between III ($M_v = -5.0$ mag) and V ($M_v = -3.3$ mag) (see p.71 in Zombeck, 1990, and references therein) and placing the source at the distance of the SMC ($d = 62$ kpc, Graczyk et al. 2014). Using the distance modulus equation, $M_v - V = 5 - 5\log_{10}(d/[\mathrm{pc}])$, we obtain an apparent V magnitude between ~ 14 and ~ 16. This is roughly in line with what we see in Figure 4.1, even when corrected for the I band ($V - I = -0.42$ from Zombeck 1990, p.68).

The observational cadence is high, peaking at ~ 2 days, but have large gaps resulting from the annual sky visibility of the Magellanic Clouds from the telescope's location. Telescope observation is limited to a proportion of the population at a time, and some fields are observed more often than others. Thus, we do not have equal amounts of on-sky time across the area that hosts the population. The goal has been to develop and understand the constraints of this methodology in an authentic astronomical setting. The dataset in this chapter and the following display different characteristics from the simulations and lead to different population size estimation methodologies. Different assumptions with regards to statistical population regimes have been made for the different datasets. The XROM data in this chapter assume a closed population, and the DNe dataset in Chapter 5 assumes a combination of open and closed population called Robust Design.

4.1 OGLE: The Optical and Gravitational Lensing Experiment and XROM

OGLE has been led by A. Udalski and operated by the University of Warsaw Observatory since 1992. It is a southern hemisphere survey conducted at the Las Campanas Observatory in Chile that aims to find microlensing events in dense stellar fields such as the Magellanic Clouds and the Galactic Bulge (Udalski et al., 1993). The dedicated instrument for the survey since phase-II of operation is the 1.3 m diameter Warsaw Telescope. Although transient and variable star projects have not been the main objective of the survey, the survey's photometric data have proven extremely useful for

longitudinal studies of objects and studying stellar populations because of the large samples and long duration (Udalski, 2008). Each of the survey phases has seen significant improvements, such as the implementation of the 'Early Warning System' (Udalski et al., 1994) for transient monitoring, along with other technical upgrades.

The OGLE-IV phase began its observations on March 4, 2010, and is ongoing (Udalski, 2008). The XROM dataset spans across the OGLE-III and -IV phases. It has intended to provide continuous photometric (I-band) monitoring of variable X-ray sources in the optical regime (Udalski, 2008; Udalski, Szymański, and Szymański, 2015). The XROM sources are surveyed in both of the Magellanic Clouds. I have focused solely on the SMC in this analysis since it is known to harbour a rich population of HMXBs compared to its larger sibling. Many of these are classified as BeXRBs (see Haberl and Sturm 2016 for overview) and this XROM dataset represents a subset of the known and candidate HMXBs population in the SMC when cross-referenced with the X-ray.

OGLE-III contains 48 SMC HMXB sources, whereas OGLE-IV contains 70 sources.[26] A total of 38 sources are encountered in both phases III and IV. In their comprehensive review of X-ray binary sources in the SMC, Haberl and Sturm (2016) report a total of 85 known HMXBs along with another 37 candidates. The Magellanic Clouds High Mass X-ray Binaries Catalog (MAGHMXBCAT)[4] currently records 92 HMXBs in the SMC. This catalogue is an up to date record of the Liu, van Paradijs, and Van den Heuvel (2005) paper. The SMC and its HMXB population has been well studied in the past two to three decades, and the expectation is that the population may be close to fully identified. From that perspective, this is a good test set to see whether it recovers what we assume is more or less the population size since XROM is an optical survey and subset of the currently known and candidate HMXBs of the SMC.

The capture-recapture application to this dataset is pursuant to our interest in HMXBs as revealed in Chapter 3. The optical lightcurve properties of HMXBs can be significantly different to the X-ray. In many cases, the optical outbursts of BeXRBs coincide with the X-ray, but not always. Thus, the results from Chapter 3 will not necessarily translate directly to this dataset. Besides, we did not consider Type II outbursts in the Chapter 3 simulations. The outburst recurrence time in the optical dataset is unknown and not easily constrained, for the reasons mentioned in Chapter 1. However, the optical variability is assumed to be periodic to a lesser degree, *at least* in comparison to the simulations implemented in the previous chapter.

[26] There are 71 sources in the OGLE catalogue, but Haberl and Sturm (2016) notes a misnomer for source AZV285 as the optical counterpart of SXP7.92

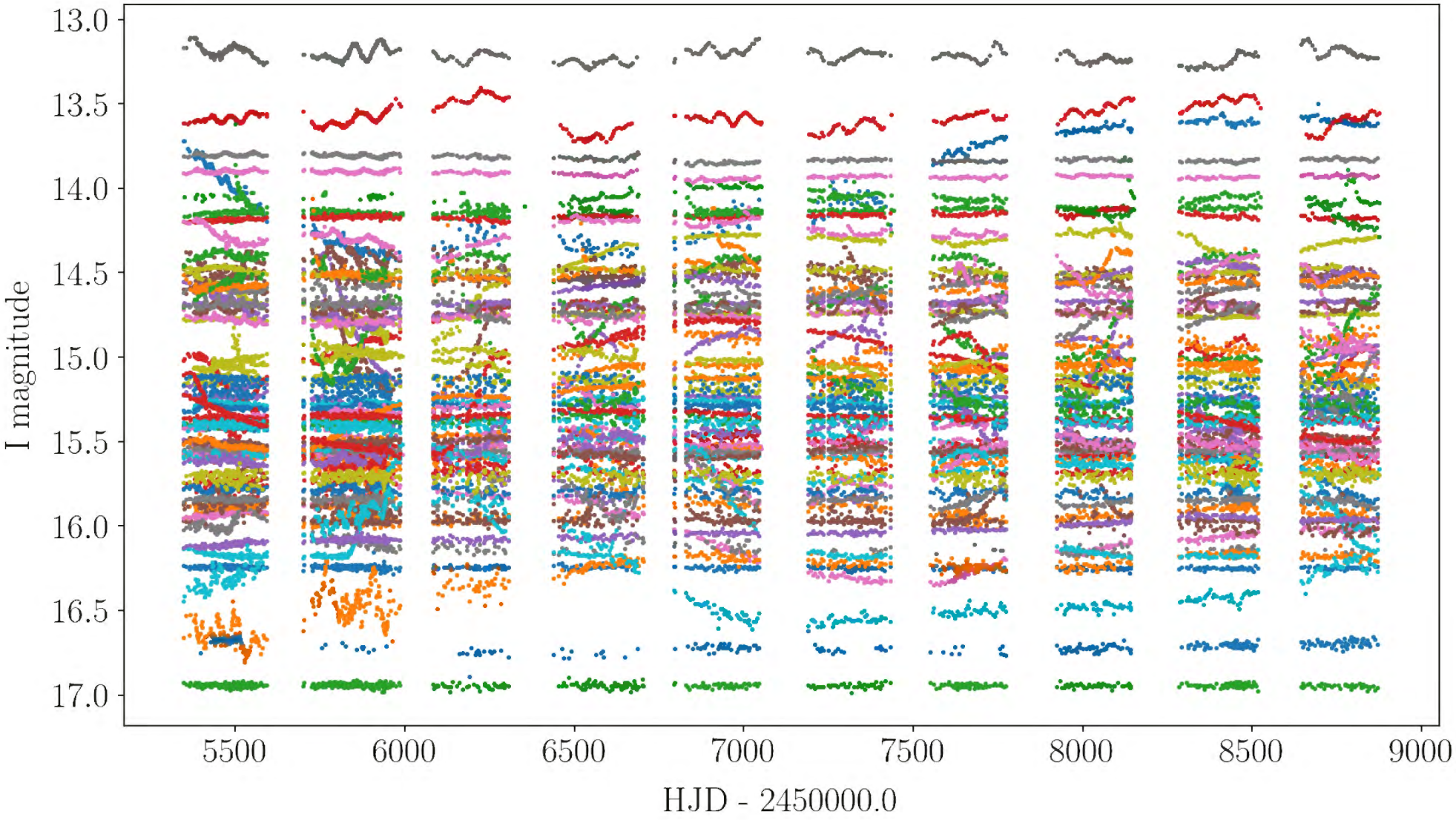


Figure 4.1: OGLE-IV XROM dataset with the lightcurves of 70 sources displayed. Most sources are found between 14^{th} and 16^{th} I mag, tapering off in number at the brighter and fainter ends. The majority of HMXBs in the SMC are classified as BeXRBs, having a young and hot O/B-type star companion in the optical. This is consistent with our previous estimations of B0 spectral type stars' apparent magnitude at that distance, within error constraints and a $V - I$ colour offset.

The **Rcapture** estimators demonstrated robustness (excepting models M_b and M_{bh}) with regards to aliasing of the cadence with the typical orbital periods. Thus, *more random* occurring outbursts (or captures) within the population are expected to be handled at least as robust by the estimators, if not better. The analysis in this chapter continues to use the **Rcapture** software for estimation.

4.2 Characteristics of the data

The working dataset, shown in Figure 4.1, comprises 70 sources. The brightness of these lie between 13th and 17th I mag. Only one of these sources show variability of $\Delta I > 1$ mag over an ~ 11 year timescale. The rest remain within a 1 magnitude variability range. The time axis is given in Heliocentric Julian Date (HJD) which is a correction applied to the Julian Date (JD) based on Earth's motion around the sun.

The distribution of times between consecutive observations for each source from their observed times (seen in the distribution in Figure 4.2) peaks at around 2 days but is

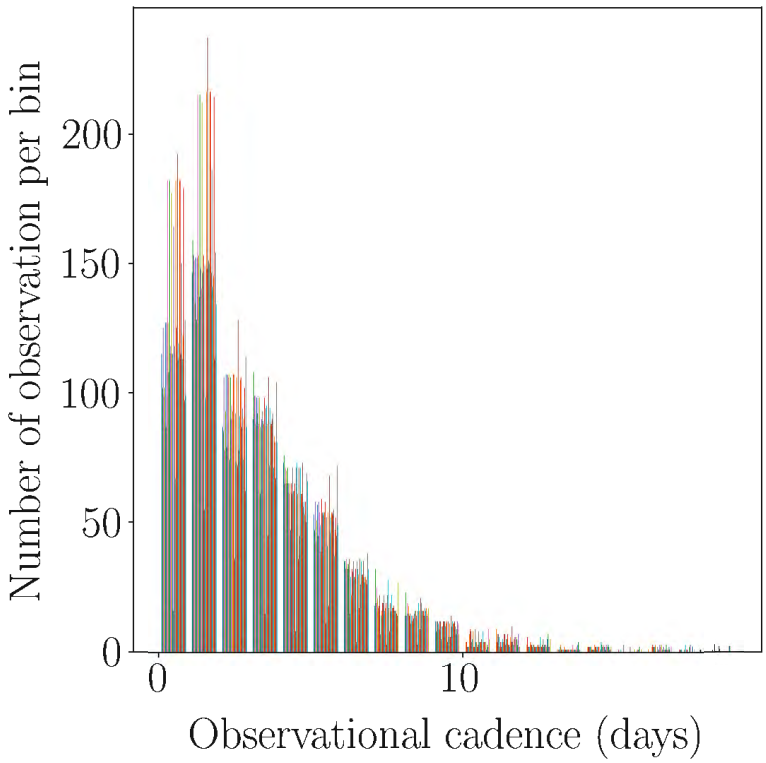


Figure 4.2: Observational cadences for each individual source in the XROM dataset, displayed in different colours. A Poissonian shaped distribution is evident with a peak sampling time at ~ 2 days.

as low as < 1 day or even as high as ~ 280 days. There are also observational gaps of 90 days and longer that are not shown in the diagram that reflect when the SMC was not visible within the pointing constraints of the Warsaw telescope from its location (latitude $\Phi = 29^\circ\ 00'\ 35.8''$ S, longitude $\lambda = 70^\circ\ 42'\ 05.9''$ W, and altitude $a = 2,275$ m). The SMC is annually best visible from August to March in the southern hemisphere.

4.3 Reduction of the data

4.3.1 Organising the data

Our approach to closed population analysis in Chapter 3 assumed a sampling strategy of simultaneous observation of the entire population on successive occasions and assuming a known state for each individual, i.e. no allowance for unknown or missing states of an individual. Only two possible states may be assigned, 1 for a capture and 0 for a miss. A significant problem arises for unobserved sources across specific epochs, which is typical of any astronomical survey. Telescope observations are constrained by its field of view, which, for the OGLE-IV phase spans a large 1.5 deg^2 field of view but still only captures about a tenth or less the sky area of the SMC at any given time.

Concerning Figure 4.2, the set of individual lightcurves had to be organised coherently to emulate simultaneous observations. The data were binned using a 1-day median statistic, starting from the first individual observed at time $t_0 = 2455347.41675$ HJD ($\sim$ May 30 2010).

4.3.2 Creating the capture histories

Capture history datasets were created from the 1-day binned observations. All lightcurves were evaluated against the same threshold, between 13.5 and 17.0 I magnitude, to create a capture history. The threshold variable was arbitrarily chosen to investigate its association to population size estimation. A value brighter than the threshold was logged as a capture (1). A value fainter than the threshold was logged as a miss (0). An absent observation was also logged as a miss, i.e. imputed with a zero. Data imputation is undesirable because it assumes the state of the individual when it has not been observed. Capture occasions were merged using the `periodhist` function (via a logical OR operation) to mitigate data imputation.
Furthermore, we did not require such high cadenced data for estimation and substantiated the use of merged observations. Additionally, Chapter 3 analysis showed that merged occasions reduced (or effectively removed) the 'behavioural' effect caused by high-cadenced oversampling in time. The XROM dataset is conveniently grouped into

annual observations across ten consecutive years. The data were merged into annual capture occasions (i.e. $t = 10$ samples), and a capture translated to being observed at least once during an annual observation run.

4.4 Estimating the population size through closed population analysis

This section presents a closed population analysis from the merged dataset evaluated against a threshold of $I_{thr} = 16$ magnitude. Table 4.1 gives the descriptive statistics of the reduced dataset. The total captures made within each merged sample were fairly constant which means that there is likely no time variation affecting the capture probability. However, $\geq 95\%$ of the individuals identified were already captured in the first sample. Figure 4.1 illustrates that no matter which brightness threshold is imposed, there is a proportion of sources that are always captured simply because their brightnesses never dips below the threshold. These will be referred to as 'always on' sources from here on-wards. The brightness distribution of sources is, as a result, the most significant contributor with regards to heterogeneity in the capture probability in the population. This was substantiated by the concave shape of $\log\left(f_i/\binom{t}{i}\right)$ via the exploratory heterogeneity graph in Figure 4.3. Across the ten merged samples, 60 of the 66 captured individuals were encountered in every sample, one individual captured 8 times, and five individuals captured 3 times or fewer. There is an indication of transient activity for only a small proportion of the population close to the threshold.

Table 4.2 shows the results of the **Rcapture** closed population estimators evaluated at 16^{th} I magnitude. Models M_0 and M_t are excluded as suitable models based on the suggestion of heterogeneity from Figure 4.3 and Table 4.1. The model selection criteria (deviance, AIC, BIC) suggest that both the M_h and M_{th} Chao, Poisson2, and Darroch

Table 4.1: Descriptive statistics for the 10 merged samples from the XROM data, evaluated against a threshold of 16^{th} I magnitude. The definitions of each statistic may be found on P. 49.

N captured: 66

i	1	2	3	4	5	6	7	8	9	10
f_i	2	0	3	0	0	0	0	1	0	60
u_i	63	2	0	0	0	0	0	1	0	0
v_i	1	1	1	0	0	0	1	0	0	62
n_i	63	64	62	61	60	62	61	62	62	62

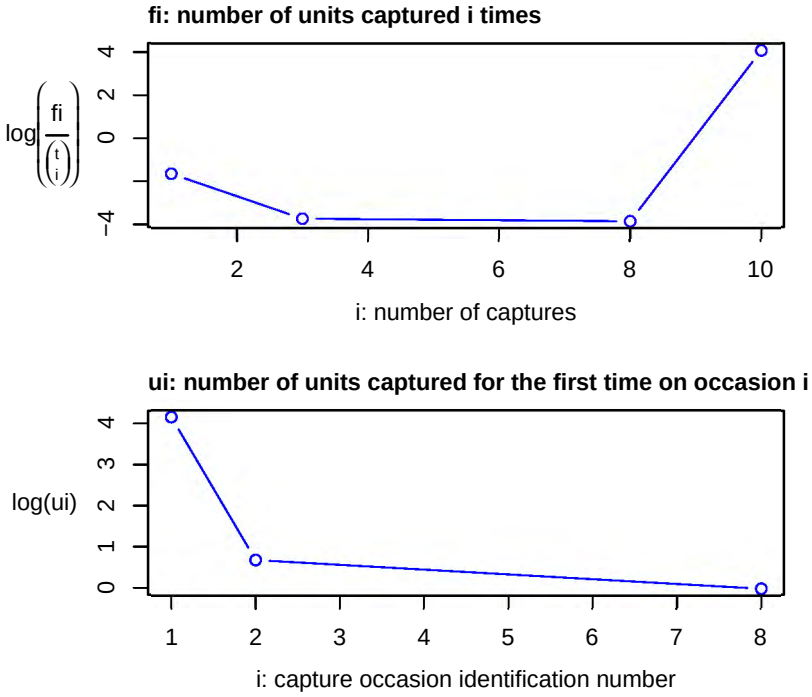


Figure 4.3: Exploratory heterogeneity graph corresponding to the descriptive data in Table 4.1. Heterogeneity is present based on the large fraction of individuals in the sample that are seen at each capture occasion, with few new individuals encountered in follow up occasions.

may be suitable estimators. The Darroch model for both M_h and M_{th} scores the worst on the AIC and BIC selection criteria out of the suitable estimators mentioned. On the other hand, the Chao and Poisson models practically only fit the captures made. This result means that either:

(a) there are no more new HMXBs remain to be found,

(b) that the Chao and Poisson models are too conservative in accounting for the degree of heterogeneity in their fit, or,

(c) that optical brightness is not a good tracer of this group's transient behaviour.

The Chao estimator is recognised as a lower-bound estimate of the population size, and Poisson2 typically yields lower corrections for heterogeneity than Darroch (Baillargeon and Rivest, 2007). Few new sources are identified after capture occasion $i \geq 2$, which intuitively suggests that the population is almost fully identified (for $I_{thr} = 16$). The low count statistics after $i = 2$ and the large degree of heterogeneity present in the data impact the lower bound estimate, so much so that there is very little confidence in predicting new individuals' identification in hypothetical, future capture occasions. Additionally, the bias of Chao's lower bound increases with the degree of heterogeneity

Table 4.2: Closed population size estimates for the merged XROM dataset at a threshold of 16$^{\text{th}}$ I magnitude

Number of captured units: 66

Model	$\widehat{N}$	$\sigma_{\widehat{N}}$	deviance	dof	AIC	BIC
M_0	66.0	0.0	206.184	1021	228.119	232.498
M_t	66.0	0.0	202.241	1012	242.176	266.263
M_h Chao (LB)	66.8	2.0	56.229	1017	86.164	99.302
M_h Poisson2	66.1	0.3	59.628	1020	83.563	90.132
M_h Darroch	78.7	12.3	81.370	1020	105.305	111.874
M_h Gamma3.5	1700.2	2005.3	102.140	1020	126.075	132.644
M_{th} Chao (LB)	66.7	1.8	45.580	1008	93.515	126.360
M_{th} Poisson2	66.1	0.2	49.420	1011	91.355	117.631
M_{th} Darroch	77.2	11.4	70.471	1011	112.406	138.682
M_{th} Gamma3.5	2430.9	2972.5	93.367	1011	135.302	161.578
M_b	66.0	NaN	199.878	1020	223.813	230.382
M_{bh}	66.1	5.6	180.502	1019	206.437	215.196

(size of parameter τ, see Appendix A Table A.2). Based on these facts and the knowledge of the amount of known and candidate HMXBs in the SMC (discussed in §1), the Chao and Poisson2 models do not provide lower bound estimates much better than the number of captures made during the study. The heterogeneous capture probability in the XROM population is arguably better handled by the M_h and M_{th} Darroch estimators.

Poisson2 bounds the capture probabilities from below, which results in the smaller predicted corrections for countering the heterogeneous capture probability compared to the Darroch model. The M_h Chao and M_{th} Chao models present with small deviance scores. The low scores are a by-product of the degree of heterogeneity in the capture probability as indicated by Baillargeon and Rivest (2007). Whilst the Chao models are recognised as being lower bound estimates of the population size, the estimators themselves are not unbiased. These indicators provide further evidence favouring the Darroch estimators as being more reflective of the population size in contrast to the other model options.

The M_h and M_{th} Gamma estimators express large overestimates of the population size compared to the other estimators, which are untenable for this particular scenario and may reasonably be assumed to be unstable based on the standard error of the mean. The Gamma estimators favour very low capture probabilities for modelling,

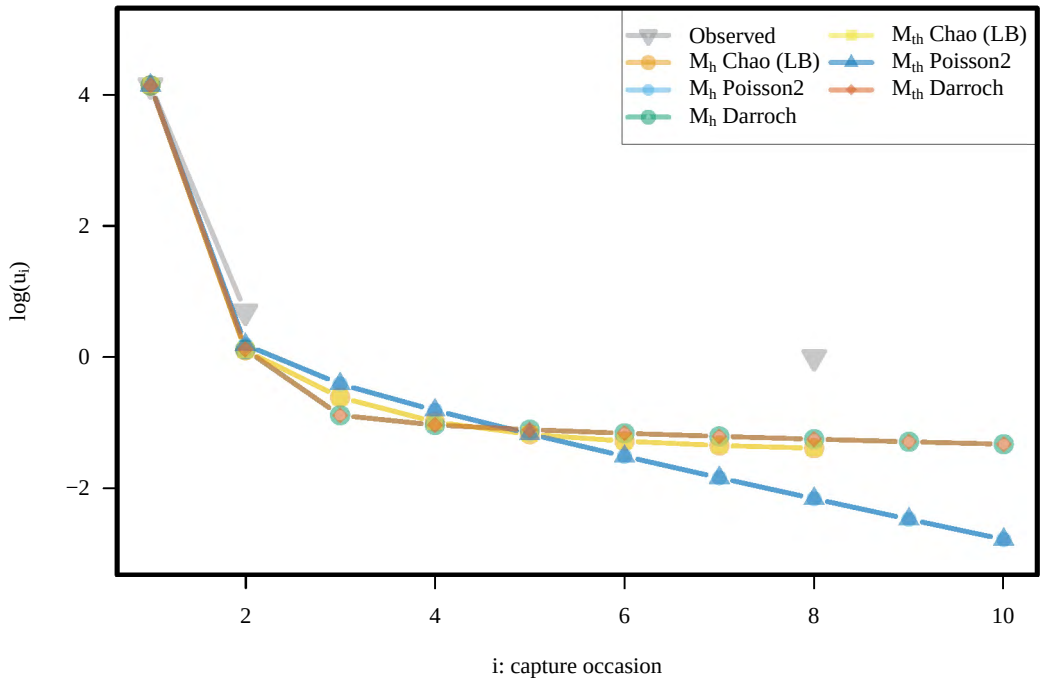


Figure 4.4: Models fitted for the u_i parameters corresponding to Table 4.3.

which results in an over-correction in the size estimate (Rivest and Baillargeon, 2007). The **Rcapture** software also provides a warning for both Gamma estimators that there is a large asymptotic bias present for this dataset. M_b and M_{bh} are also unlikely model contestants since capture occasions have been merged. Hence, we can exclude behavioural (correlative) effects that happen as a result of oversampling.

The M_h and M_{th} models that have been fitted by `closedp` are shown for parameter u_i in Table 4.3 and Figure 4.4 in comparison to the observed new individuals in each sample. The models may not seem superficially different since they are marred by the heterogeneous capture probability which makes it extremely difficult to fit accurately. However, the M_h Darroch estimator has a steadier gradient in its prediction of newly encountered individuals for $i \geq 2$ and fits best in terms of the χ^2 goodness of fit given in Table 4.4. M_h Darroch has the largest variance in u_i, $\sigma^2_{\bar{u}} = 0.716$, compared to any of the other estimators which filters through in the standard error on the estimate provided in Table 4.2: $\widehat{N} \pm \sigma_{\widehat{N}} = 78.7 \pm 12.3$.

A different metric for determining the model goodness-of-fit is the Pearson residuals test (cf. §2.2.3) that can be found as boxplots in Figure 4.5. The residuals are meant to bring out badly fitted data. The interquartile range for each model are small, and

Table 4.3: M_h and M_{th} models fitted to the observed u_i parameter for occasions i.

u_i	observed	M_h Chao (LB)	M_h Poisson2	M_h Darroch	M_{th} Chao (LB)	M_{th} Poisson2	M_{th} Darroch
$u1$	63	62.250	62.250	62.250	63.000	63.000	63.000
$u2$	2	1.403	1.320	1.464	1.951	1.910	1.549
$u3$	0	0.881	0.790	0.504	0.711	0.758	0.947
$u4$	0	0.561	0.555	0.414	0.057	0.052	0.053
$u5$	0	0.364	0.402	0.376	0.050	0.047	0.053
$u6$	1	0.244	0.296	0.350	0.044	0.043	0.052
$u7$	0	0.170	0.221	0.329	0.040	0.039	0.051
$u8$	0	0.125	0.166	0.312	0.146	0.152	0.295

Table 4.4: Goodness of fit parameters for the M_h and M_{th} models fitted to u_i across all capture occasions i.

Model	Goodness of fit to u_i	Day of first capture	
	χ^2	$\bar{u}$	$\sigma^2_{\bar{u}}$
observed	–	1.106	0.398
M_h Chao (LB)	4.703	1.143	0.498
M_h Poisson2	4.162	1.154	0.573
M_h Darroch	3.347	1.169	0.716
M_{th} Chao (LB)	21.614	1.079	0.233
M_{th} Poisson2	22.531	1.080	0.237
M_{th} Darroch	18.880	1.098	0.358

the whiskers of the boxplots extend up to a distance of 1.5 times the interquartile range. The median residuals of each model fit are consistently below zero since the fitted models estimate $u_i > 0$ for all capture occasions i (cf. Table 4.3) where a majority of the observations encounter zero new individuals. Outliers are not shown in this diagram. However, up to 20% of the data for each model have been discarded as outliers. Thus, whilst the residuals are within an acceptable range, the high number of discarded outlier data may indicate poor fitting.

The bias-corrected estimates from the `closedp.bc` function (cf. Table A.3) for the heterogeneous model candidates are located in Table 4.5. M_h Darroch did not show

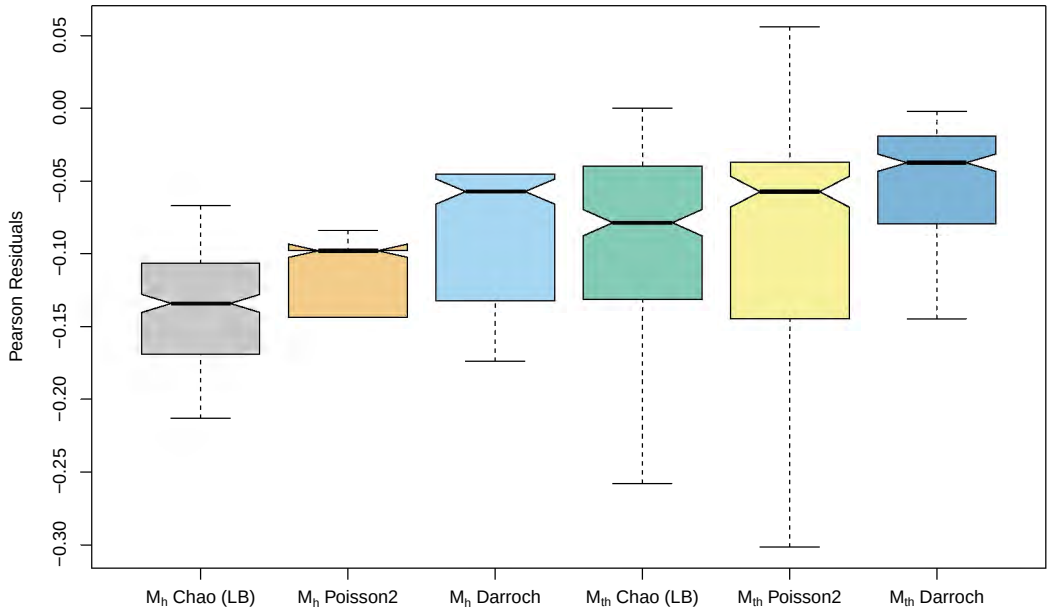


Figure 4.5: Pearson residuals for models in Tables 4.3 and 4.4.

a measurable improvement; however, the M_{th} Chao estimate, $\widehat{N}_{bc}$, increased at the expense of a larger standard error under the bias correction. The correction increases the possibility of a lower bound estimate up to $\widehat{N}_{bc} \sim 83$. M_{th} Darroch shows a measurable increase beyond 3σ and a larger associated standard error when bias-corrected. Bias-corrections may be necessary when confronted with heterogeneity in capture probability – due to the variation in source brightness concerning the imposed threshold.

Table 4.5: Bias corrected closed population size estimates for the XROM dataset evaluated at a threshold of $I_{thr} = 16$ magnitude.

Model	$\widehat{N}_{bc}$	$\sigma_{\widehat{N}_{bc}}$
M_h Chao (LB)	66.0	0.0
M_h Poisson2	66.1	0.4
M_h Darroch	79.2	11.7
M_{th} Chao (LB)	73.6	9.5
M_{th} Poisson2	66.8	1.0
M_{th} Darroch	113.6	29.7

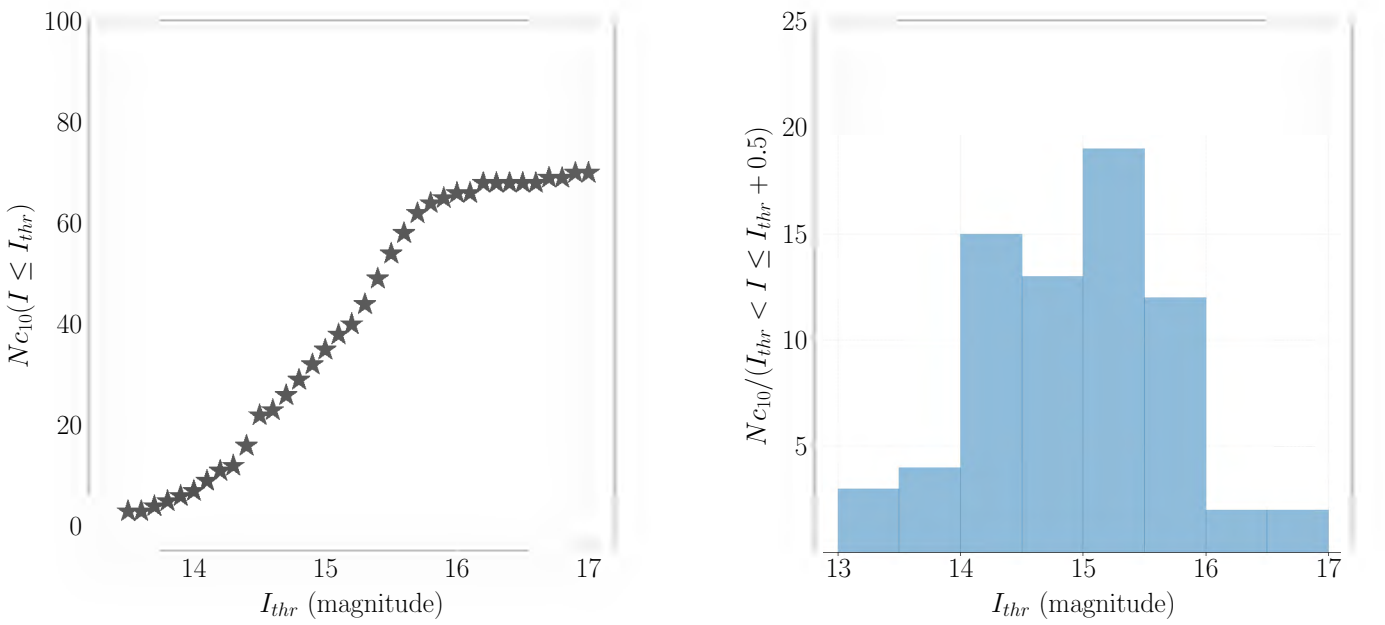
4.5 Population size as a function of magnitude threshold

In the previous section, I have argued points favouring the M_h Darroch estimator with some evidence that bias correction may be needed for the estimates because of the heterogeneous capture probability present among the individuals. The M_h Chao and Poisson2 may be helpful by providing lower bound estimates rather than only relying on the captures made as a lower bound. The previous section's analysis was done for a single threshold and failed to indicate the source brightness distribution. From inspection of Figure 4.1, we can see that the brightness threshold is the *most significant* limiting factor in determining the population size. Hence, it was necessary to evaluate the population size as a function of the imposed threshold.

Figure 4.6 (a) shows the size increasing as a function of threshold. The accumulation of individuals corresponds with each source's maximum brightness. The count increases as the threshold sweeps across the range of expected apparent magnitudes of HMXBs, between 14th and 16th I magnitude, and finally tapering off at the fainter magnitudes. The cumulative distribution may be suggestive of a sigmoid function (or S-curve). Figure 4.6 (b) shows the almost Gaussian distribution, which indicates the maximum brightness bin in which the individuals are captured.

The closed population size estimates, evaluated with a threshold between 13th and 17th I magnitude in increments of 0.1 magnitude, were determined using the same procedure as in the previous section. From the analysis in §4.4, we determined that the most appropriate models for this dataset were the M_h, and M_{th} Chao, Darroch and Poisson2, and thus we restrict the analysis to these. Figure 4.7 (a) indicates the heterogeneous models only, where 4.7 (b) shows the time-heterogeneous models. The Chao, Darroch, and Poisson2 estimates for the M_h and M_{th} models are similar. Model M_{th} struggles with convergence on both the bright and faint ends of the magnitude scale, where the brightness density of sources is significantly lower (notice the standard error that moves off the chart to $\gg 10^6$). Estimates for thresholds beyond the faintest I_{max} in the sample (in this instance $Nc_{t=10}$ for $I_{thr} = 16.9$) should be interpreted with caution.

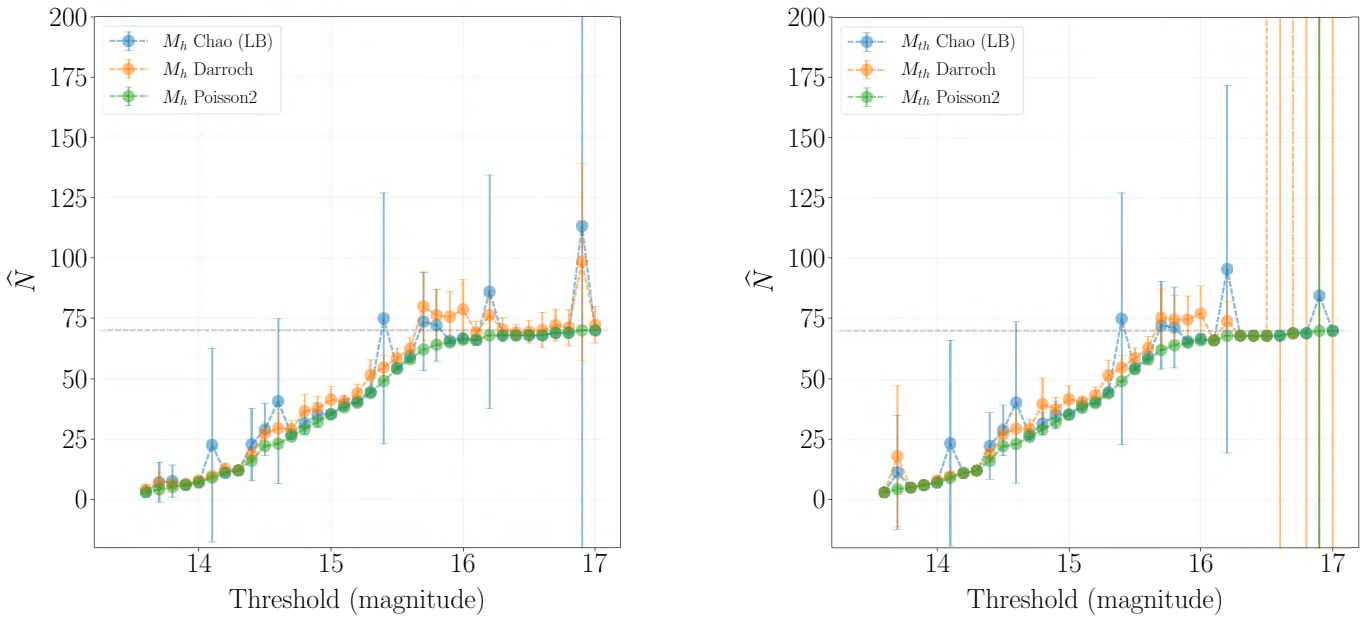
The population size estimation methodology in this dataset differs from the simulations because it is severely restricted by threshold. A proportion of the sources remain obscured, and a proportion is always visible. A certain few individuals hover in brightness in a region of the threshold, crossing it occasionally. The method is therefore only



(a) (b)

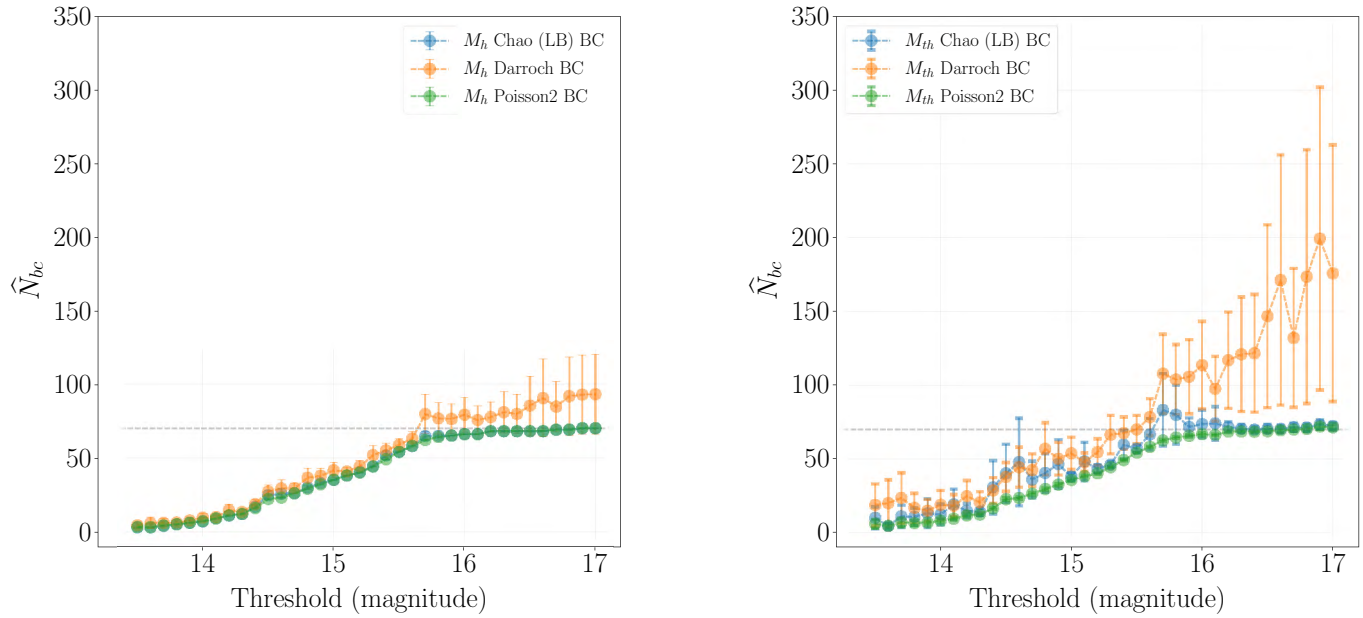
Figure 4.6: (a) Cumulative distribution function of the number of captured individuals as a function of the brightness threshold. (b) Number of individuals captured per magnitude bin. The shape may be indicative of a Gaussian distribution.

sensitive to these few individuals who comprise a small fraction of the dataset. The low count statistics as the threshold in increased to fainter magnitudes is reflected in the larger standard error of the estimate. In Figure 4.7, for both (a) and (b), the Chao estimates are particularly affected in efficiency at thresholds $I_{thr} = 14.1$, 14.6, 15.4, 16.2, and 16.9 magnitude. The thresholds coincide with a sudden rate of increase in captures. At $I_{thr} = 15.7$ and 15.8 magnitude, a similar small spike is seen with both the M_h and M_{th}, Chao and Darroch estimators. These thresholds coincide with a decrease in the rate of captures with respect to the threshold (cf. Figure 4.6). The bias-corrected estimators seem overall more robust in the estimation and correcting for the associated standard error and overcome issues with non-convergence, which can be seen in M_{th} Darroch at $I_{thr} = 16.5$ magnitude. In summary, the heterogeneous set of models, M_h Chao, Darroch, and Poisson, are applicable for handling the unequal capture probability concerning an imposed brightness threshold. Bias-correction plays an important role in offsetting this heterogeneity.



(a) M_h (b) M_{th}

Figure 4.7: Closed population estimates as a function of threshold.



(a) Bias corrected M_h (b) Bias corrected M_{th}

Figure 4.8: Bias corrected closed population estimates as a function of threshold.

By decreasing the OGLE data's time resolution, the approach that we have taken here was a tactic to avoid imputation or interpolation. It did, however, change the underlying variability structure of the sources of the population. Whilst this is, admittedly, far poorer time resolution than we explored in the simulated populations (using X-ray lightcurves), it should be noted that lower time resolution does not affect estimation here. The temporal binning is a minor concern compared to the heterogeneous effect due to the distribution of sources in brightness space. No matter how the capture occasions are binned, u_1 $(i = 1)$ is always tens of times larger than u_i $(i \geq 2)$ and similarly, f_i is capturing large numbers of sources exactly t times (where t is the number of capture occasions in the study). The capture data is no longer Poissonian due to this heterogeneity.

On page 88, I listed several reasons that may explain why the Chao and Poisson models do not estimate any better than the lower bound of the number of units captured during the study. We know that the optical outbursts of HMXBs do not always coincide with the X-ray outbursts and that the optical outburst can reach up to a few magnitude. However, the most extensive variability that is displayed by a source in this dataset is ~ 1 magnitude, whilst most of the sources have $\Delta I_{max} < 0.3$ magnitude, plotted in Figure 4.9. For this reason, we require a relatively deep threshold, i.e. *faint enough*, to estimate within the number range of the currently known size of the HMXB population.

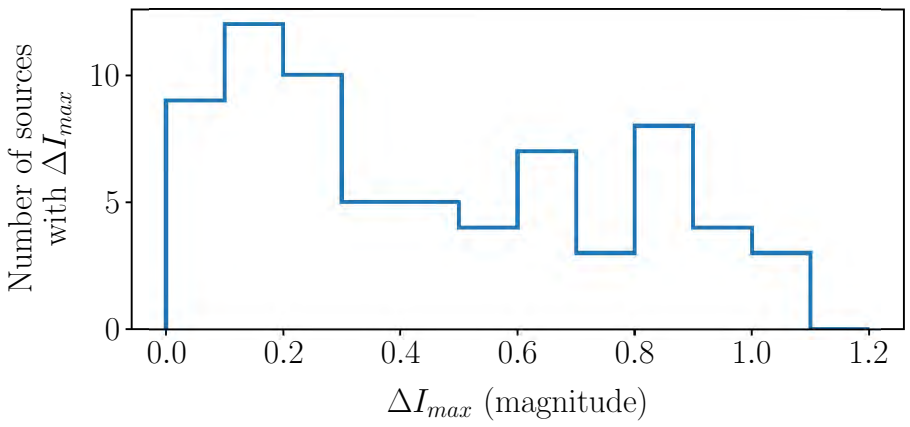


Figure 4.9: Distribution of sources binned according to maximum variability ΔI_{max} displayed in OGLE-IV.

There are numerous other ways that one could attempt the data reduction, for instance:

- One approach could be to 'offset' all of the sources to the same quiescent, mean, or median magnitude. It would significantly alter the heterogeneous effect of the 'always on' sources. The choice of threshold is, in this case, still not arbitrary because a bulk of the sources show variability $\Delta I < 0.2$ magnitude. It would be inadvisable to lower the threshold to such an extent that it captures variability that is not associated with an outburst of an HMXB for some, whilst it is the defining characteristic for others.

- Another option is to augment the dataset. Augmentation may involve interpolating each source's lightcurve on short timescales (e.g. a few weeks at most) using Gaussian processes. Stratified sampling may subsequently be implemented with a Monte Carlo approach, similar to steps applied in §3. The capture histories are created and finally, the population sizes estimated with associated confidence intervals.

One of the benefits of taking the route that we have, i.e. leaving the sources undisturbed in brightness space, is that the imposed threshold attains physical meaning. Design-based methods, which were not implemented here, may also help address the heterogeneity within the sample, regarding the 'Snowshoe hare example' in Baillargeon and Rivest (2007, §2.1). In the example, they design the matrix, X_i, used in the log-linear regression (cf. §2.2.2) to offset the heterogeneity caused by a fraction of hares that were captured at every occasion. Heterogeneous effects and the covariates explaining the unequal catchability require further scrutiny to characterise the underlying population size fully.

A blatant issue is present within this dataset that prevents us from drawing astronomically valid inferences; the sources in this survey have been selected for monitoring and are not encountered at random within the populated survey area. We can therefore not comment on the completeness of the HMXB population within the SMC, even for flux-limited cases. An authentic dataset will need to be generated by returning to the OGLE-IV survey to extract and classify HMXB. However, since the focus of this analysis was on methodology and the interpretation of the statistical results, we opted to treat this dataset as if it were authentic. Future work on this particular population should make attempts at classification and relevant astrophysical parameters that influence present day population size may be incorporated into the problem. Furthermore, the next chapter presents a more realistic approach using survey data by implementing a combination of closed and open population techniques, called 'Robust Design'.

Chapter 5

Application to astronomical population datasets: Part II

5.1 OGLE DNe in and towards the Galactic Bulge

Up until this point, all the population analysis has been in the context of closed populations which assumes that our sampling area remains the same throughout the entire study. Thus all sources in the study area should, in principle, be available for sampling. I mentioned that recurring transients might generally be assumed to be a statistically closed population in the introductory chapters. This definition holds well under the assumptions of negligible proper motion (i.e. spatial movement) and high probability of survival ($S \sim 1$) for astronomical objects across the length of the epochs under study. However, due to the nature of astronomical observation in which we are limited to the telescope field of view, which is typically far smaller than the field of study, the resulting individuals that are 'available' for capture are spatially constrained to us at a given capture occasion. Campaigns such as OGLE define the sky survey area by dividing it into sub-fields of observation. The sky footprint may change between phases of the campaign as the survey undergoes a redesign. Several factors lead to a survey overhaul, with one of them being technical upgrades. Because the population fundamentally changes when the survey area is redesigned, we may draw a parallel with the concept of open populations that incorporate the *sampling unavailability* of individuals between certain phases into its modelling.

An introduction to robust estimation was given in §2.4, which combines the open and closed population methods to yield estimates of both size and survival parameter estimations throughout the survey. This method is argued to apply to the case that will be presented in this chapter. The respective phases of OGLE relevant to this

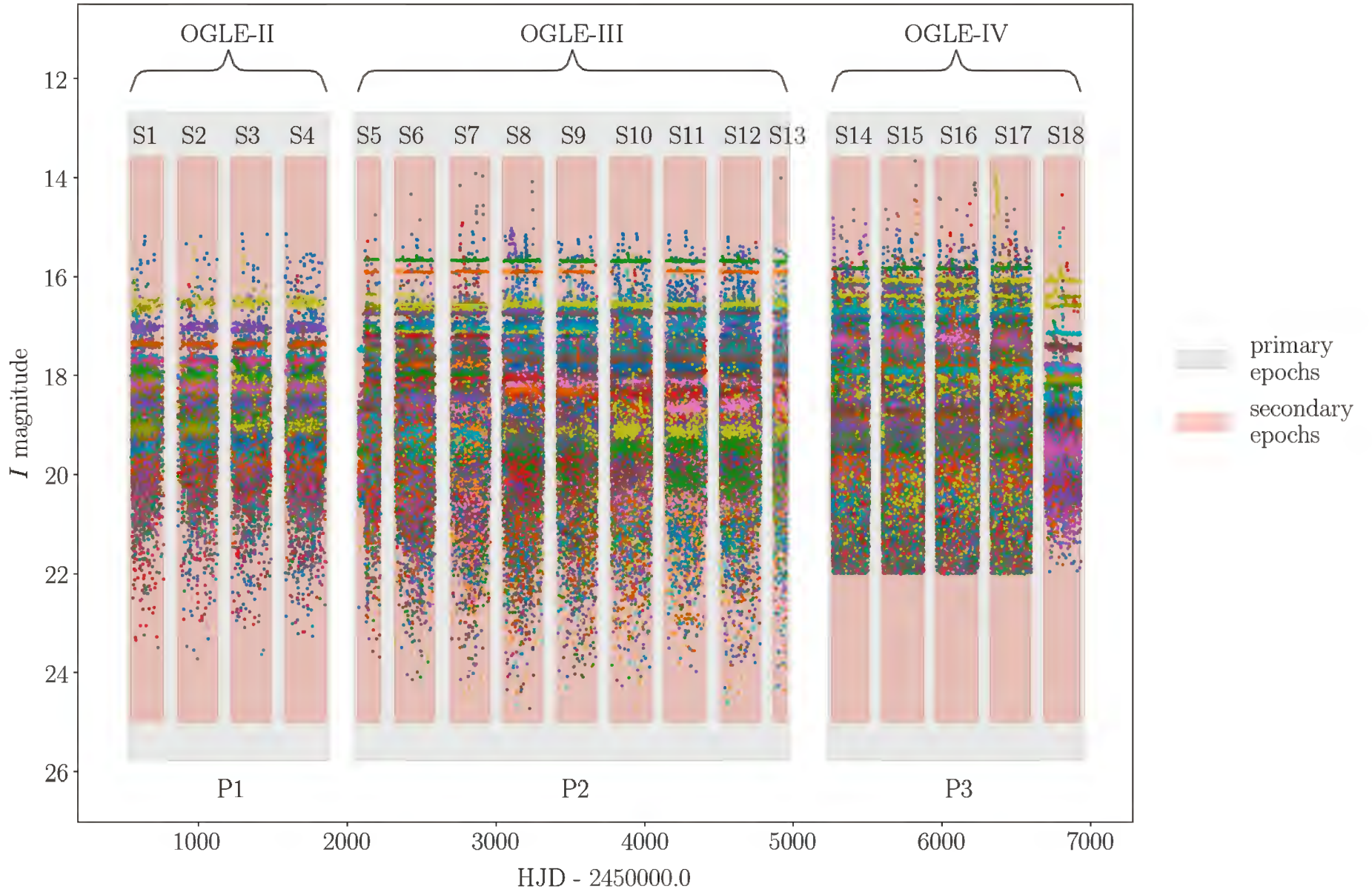


Figure 5.1: OGLE dataset (Mróz et al., 2015; Mroz et al., 2016) containing 1059 DNe towards the Galactic Bulge. The sources display a wide range of magnitudes ($13 < I < 25$ mag) with the highest brightness density located between 18^{th} and 20^{th} I magnitude. The outburst amplitudes vary up to as much as $\Delta I \leq 5$ magnitude for these DNe. The sampling structure of the primary and secondary epochs are defined in §2.4. Colours are arbitrarily assigned to sources.

analysis are OGLE-II (1997-2000, Udalski, Kubiak, and Szymanski 1997), OGLE-III (2001-2009, Udalski et al. 2008), and OGLE-IV (since 2010, Udalski, Szymański, and Szymański 2015)[27].

5.2 Characteristics of the data

The dataset used in this chapter is from the OGLE survey and published in Mróz et al. (2015)[28]. It features the long term lightcurves of the subset of DNe that were observed in OGLE phases II, III, and IV. It contains a total of 1091 DN sources, 1059 of which are located towards the Galactic Bulge, and the remaining 32 located in the direction of the Magellanic Clouds but still belonging to the Milky Way. This dataset contains a significantly larger sized population than the XROM dataset analysed in §4. Therefore we expect better relative precision in the size estimation. I have spatially constrained the analysis to the DN population located towards the Galactic Bulge only.

This dataset is essential in studying a large sample of CV sources for statistical properties. The OGLE survey specifically probes dense stellar regions like the Galactic Disk and Bulge, where other optical surveys such as CRTS, PanSTARRS, PTF and ASAS-SN suffer from higher degrees of confusion in these parts of the sky. The majority of *observable* CVs are thought to be located in the Galactic Bulge, where the stellar density is high and it can be challenging to discriminate and classify stellar objects due to source confusion. The authors point out that the sample is truncated and limited by the following selection criteria:

- technical and seeing limitations that result in source confusion in the densest stellar regions;
- a minimum increase in brightness of 1.0 magnitude that was sustained for at least three consecutive nights;

[27] OGLE-I was a pilot project that did not have a dedicated telescope and was restricted in available telescope time (Udalski, Kubiak, and Szymanski, 1997).
[28] The catalogue may be found at `https://cdsarc.u-strasbg.fr/viz-bin/Cat?J/AcA/65/313`.

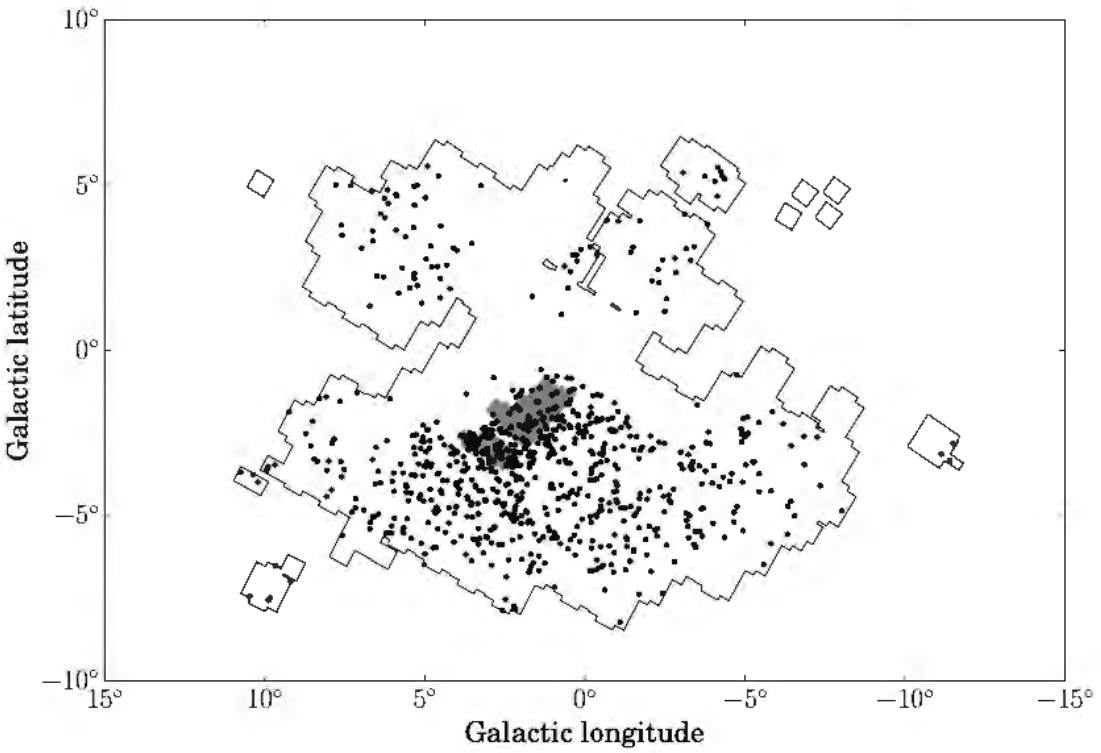


Figure 5.2: Figure 2 from Mróz et al. (2015) showing the spatial distribution of the DN candidate sources in Galactic coordinates, above and below the Galactic plane and centred around the Bulge reflected for OGLE-IV. The density of candidate DNe is lower for sources closest to the Galactic plane due to low cadence sampling, source confusion and extinction. It has limited the robust detection of outbursts based on the selection criteria previously mentioned. The black lines represent the OGLE field survey towards the Galactic Bulge. The shaded area forms part of the K2 Campaign 9 (Howell et al., 2014; Henderson et al., 2016) and is unrelated to this inquiry.

Additionally, Mróz et al. (2015) note that other classes of transients may contaminate this dataset due to misclassification, either because of low-amplitude source variability, faintness or sparse temporal sampling. However, the fraction of misclassified sources is estimated not to exceed 10% of the total sample. Possible contamination sources include:

- extragalactic transient sources i.e. SNe and Active Galactic Nuclei (AGN);
- classical novae which may be very reddened or distant;
- X-ray binaries, which also display outbursts of up to a few magnitudes in the optical, and demonstrated in Figure 4.1;
- young stars that may show a large variety of amplitudes and quasi-periodic outbursts;

- Be stars, which may exhibit more extreme amplitude variability but not known to be larger than ~ 1.2 magnitude.
- microlensing events, where binary lensing events are mistaken for a DNe outburst due to their asymmetric lightcurves compared to single lensing events.

Figure 5.2 displays the spatial location of the identified DN candidates concerning the OGLE fields of the survey in the black lines. The maximum source brightness distribution (Figure 5.3, left) shows a distinct increase in the density towards the faint magnitudes, which are likely DNe contained within the Galactic Bulge. The distribution also reflects the number of captured sources as a function of the applied magnitude threshold. The sources identified with the Mróz et al. (2013) algorithm displayed outburst amplitudes $\Delta I > 1.0$ magnitude, whilst the rest were identified with the OGLE Early Warning System. The sampling cadence of the ~ 11 year survey peaks around 3 days, similar to the XROM dataset (cf. Figure 5.4).

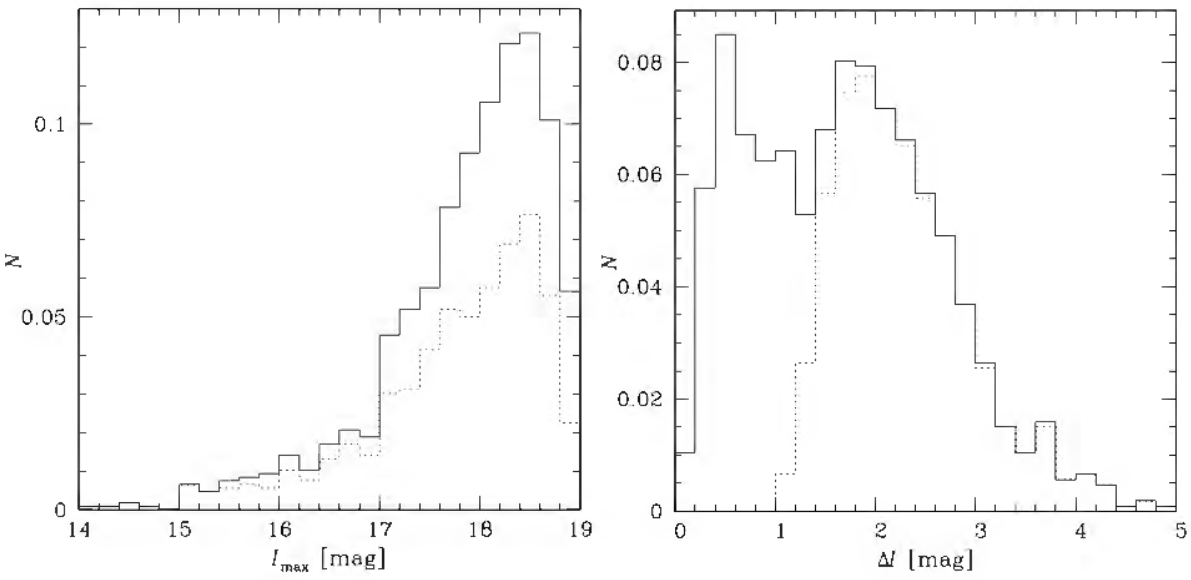


Figure 5.3: Figure 3 from Mróz et al. (2015) showing the normalised distribution of peak magnitudes on the left and the DNe outburst amplitudes on the right for the dataset in Fig. 5.1. The dotted line in both the left and right plots indicate the DNe discovered through the algorithm of Mróz et al. (2015). The bulk of the DNe of the dataset is located at faint magnitudes and low outburst amplitude. Mróz et al. (2015) note that these sources are most likely located within the Galactic Bulge.

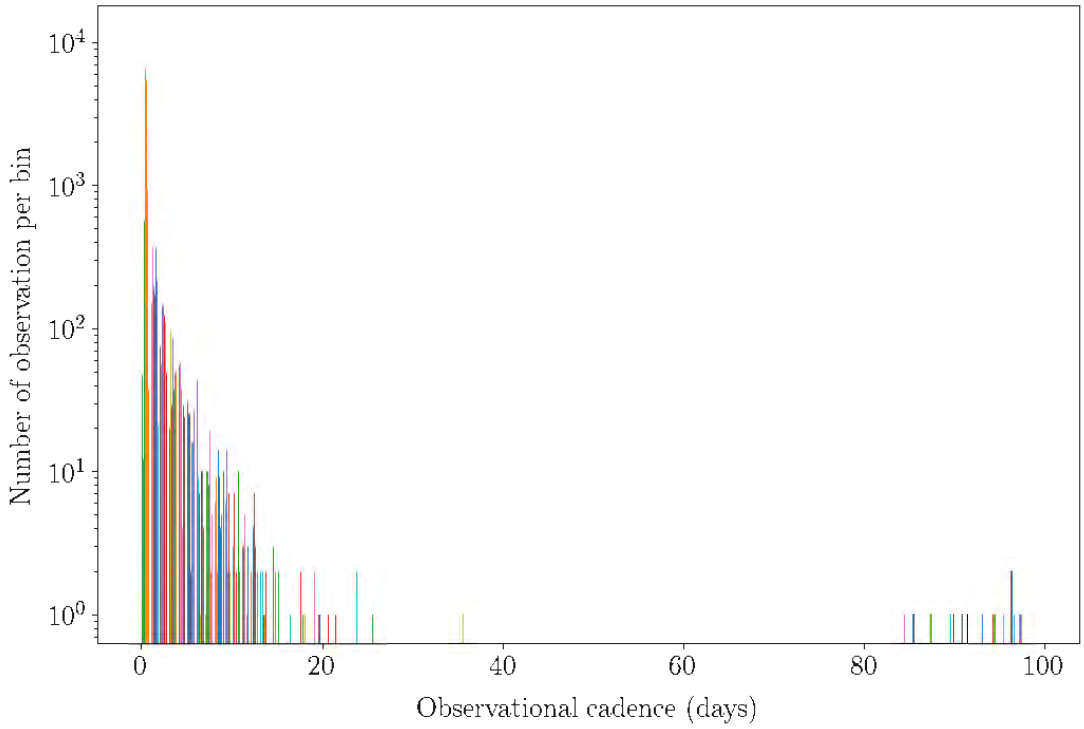


Figure 5.4: Cadence distribution of observations for 1059 DNe located towards the Galactic Bulge. The peak is located between 2 and 3 days.

5.3 Reduction of the data

5.3.1 Organising the data

The dataset was organised into 1-day bins using a median statistic (as was done for the closed population data reduction in §4). Sources that were not observed within a given day were imputed with a value of zero. To use the robust design method, we evaluated whether it would be reasonable to define each temporal phase of the OGLE survey as a closed population system since each phase is allocated defined fields on the sky. The overall size estimate would finally be based on the combined survey area for phases II to IV. This assumption was checked by investigating the number of minimally detected sources (i.e. $I \leq I_{thr}$) across small sub-epochs in time across the $\sim$ 11 year timescale. The observational blocks were separated according to epochs when no sky observation was done for 50 days or longer – these have been labelled as the *secondary epochs*, indicated by 'S#' in Figure 5.1, where '#' is an assigned number. The 50-day limit was chosen based on the cadencing of the observations in Figure 5.4 which is a middle ground between the high and low cadences. The low cadences of 80 days or more are the hiatuses where the Galactic Bulge is not at low enough airmass for observation from the Warsaw telescope.

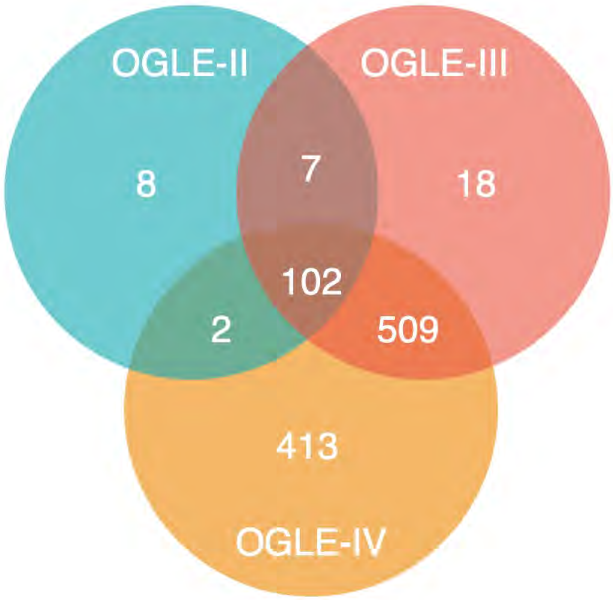


Figure 5.5: Number of sources common to each of the OGLE-II, -III, and -IV survey phases. Each survey contain a significant number of new or different sources which suggests an open population with the surveys as primary epochs in a robust design model.

The number of sources were compared across 18 such secondary epochs (cf. Fig. 5.1) to assess if the condition of closed population holds. Table 5.1 indicates the number of sources, $N_{S\#}$, in each secondary epoch which spans between times t_1 and t_2. Table 5.1 suggests that the closed population condition holds across the following secondary epochs:

$$\text{S1} \rightarrow \text{S4},\ \text{S5} \rightarrow \text{S13},\ \text{and}\ \text{S14} \rightarrow \text{S18}.$$

Notably, between S4 and S5, and S13 and S14, the OGLE surveys were reviewed and where new members join the population and others 'emigrate'. The Venn diagram in Figure 5.5 illustrates the number of sources that belong to each survey and the number of shared sources. It is clear that there are significant increases of 'new' sources with each advancing phase, and thus the assumption that each OGLE phase satisfies the condition for closure is well motivated. The secondary epochs across which the condition held were subsequently grouped into three *primary epochs*, and these correspond to the II, III and IV phases of OGLE. They have been labelled similarly, indicated by 'P#', where '#' is an assigned number. Both the secondary and primary epochs are illustrated in Figure 5.1. The distinct number of sources encountered within a primary epoch is laid out in Table 5.2.

5.3.2 Creating the capture histories

Capture histories were compiled for the 1-day median binned dataset against brightness thresholds of $14.6 \leq I_{thr} \leq 20.1$ magnitude[29] in increments of 0.1 magnitude. Captures were assigned to an individual if $I \leq I_{\rm thr}$, and misses were assigned where $I \geq I_{\rm thr}$ or where sources were unobserved. I have taken a similar approach to the previous chapter by merging the capture histories using the `periodhist` function, in this case, by counting the secondary epoch captures within each primary epoch. A capture translates to being encountered *at least* once within a secondary epoch. To incorporate heterogeneity of individuals into the analysis, the **Rcapture** software required at least three samples per primary epoch. Thus the primary epochs consisted of 4, 9, and 5 samples for P1, P2, and P3.

Table 5.1: Secondary epochs that have been distinguished by at least 50 days of zero observation across the ~ 11 year timescale. The distinct number of sources encountered in each epoch is given by N and the corresponding start and end dates from t_1 to t_2.

Secondary epoch Identifier	$N_{\rm S\#}$	t_1 (HJD − 2450000)	t_2 (HJD − 2450000)
S1	116	541	762
S2	119	857	1128
S3	119	1216	1488
S4	119	1579	1860
S5	620	2072	2225
S6	628	2322	2585
S7	617	2694	2959
S8	559	3041	3316
S9	448	3409	3677
S10	453	3773	4051
S11	365	4132	4412
S12	379	4506	4780
S13	385	4867	4957
S14	1014	5261	5508
S15	1018	5596	5882
S16	1020	5957	6243
S17	1021	6327	6601
S18	60	6688	6933

[29]The choice of thresholds were based on the brightest and the faintest sources. The estimates did not converge outside of the bounds for I_{thr} given above.

Table 5.2: Three primary epochs that correspond to the OGLE-II, -III, and -IV surveys. The distinct number of sources encountered within each epoch is given by N and the corresponding start and end dates from t_1 to t_2.

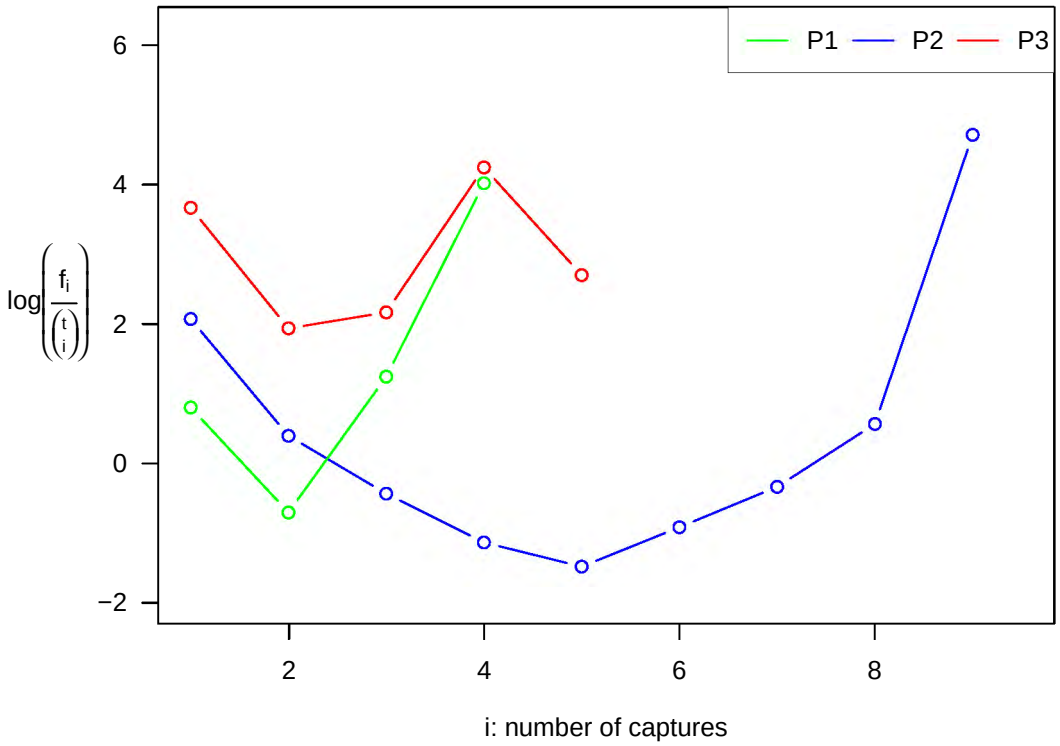
Primary epoch Identifier	$N_{\mathrm{P}\#}$	t_1 (HJD − 2450000)	t_2 (HJD − 2450000)
P1 (OGLE-II)	119	541	1860
P2 (OGLE-III)	636	2072	4957
P3 (OGLE-IV)	1026	5261	6933

5.4 Robust estimation and analysis

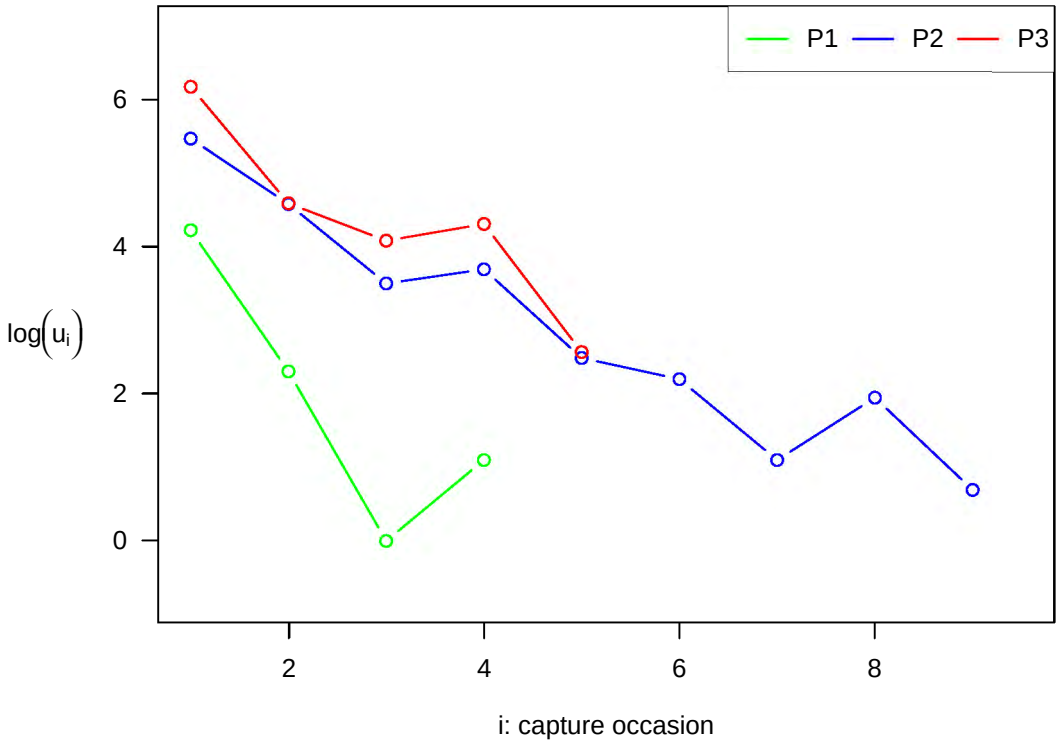
Firstly, a detailed analysis was compiled with the robust design method against a threshold of $I_{thr} = 18.5$ magnitude, where the density of sources in brightness space is reasonably high. The descriptive statistics at each secondary epoch sample is provided in Table 5.3. Parameters $\log\left(f_i/\binom{t}{i}\right)$ and $\log(u_i)$ are plotted in Figure 5.6. In 5.6(a), $\log\left(f_i/\binom{t}{i}\right)$ shows concave-shaped curves within each primary epoch (P3, plotted in red, shows an outlier at $i = 5$) which suggests that heterogeneity is present in the population. Heterogeneity is expected given the analysis in the previous chapter, where we saw it present due to the brightness distribution of sources for the applied threshold. In Figure 5.6(b), $\log(u_i)$ is linear for P2 and P3, whereas P1 is concave. The number of sources captured, n_i, varies noticeably with capture occasion i in P2 (up to $\sim 50\%$ of the mean captures in P2). A temporal effect is present in the capture probability to a greater degree than in P1 or P3. Given these characteristics, we will consider M_h and M_{th} for modelling within the primary epochs of the robust analysis and assess which is the better fit.

Table 5.3: Descriptive statistics for each secondary epoch of the DNe dataset in the direction of the Galactic Bulge ($N_{captured} = 783$, $I_{thr} = 18.5$ magnitude). Parameter definitions can be found in §2.5.1.

	P1 (OGLE-II)				P2 (OGLE-III)									P3 (OGLE-IV)				
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18
i	1	2	3	4	1	2	3	4	5	6	7	8	9	1	2	3	4	5
f_i	9	3	14	56	72	54	55	41	29	34	26	16	112	197	70	88	351	15
u_i	68	10	1	3	236	97	33	40	12	9	3	7	2	477	98	59	74	13
n_i	68	74	65	74	236	301	287	276	244	237	202	212	181	477	504	509	546	44
captured	82				439									721				



(a)



(b)

Figure 5.6: Descriptive statistics $\log(f_i/\binom{t}{i})$ and $\log(u_i)$ plotted for the secondary epoch samples in each primary phase ($I_{thr} = 18.5$ magnitude). Parameter definitions can be found in §2.5.1.

It should be noted that a different closed population size estimator can model each primary epoch and that M_{th} is likely to be a better fit for P2. I first applied the same model at each primary epoch for consistency and analysed the results. In order to compare like with like, the `robustd.t` function was used to evaluate both models M_h and M_{th} (cf. §2.5; M_h evaluated with `robustd.0` arrives at the same estimates as M_h with `robustd.t`, but with far fewer degrees of freedom, whereas M_{th} can only be implemented with `robustd.t`). Table 5.4 presents the model fit parameters and key estimates for the Gamma estimator. The Gamma estimator has previously shown to give large weight to small capture probabilities. It tends to overestimate the population size as a result, especially for small sample sizes (cf. §4.2). The estimated capture probabilities in P1 are low for both M_h and M_{th} Gamma, with inefficient estimation given the large associated standard errors. The capture probability impacts the size estimation in P1, resulting in standard errors that are larger than the size estimates. However, the overall size estimates do not appear greatly affected in terms of efficiency and accommodate between $\sim$ 50 and $\sim$ 130 sources unaccounted for at $I_{thr} = 18.5$ (roughly 6% - 17% of the 783 captured sources). Overall, the M_{th} Gamma delivers far lower deviance and AIC scores than M_h Gamma, which suggests that it may be the more appropriate model.

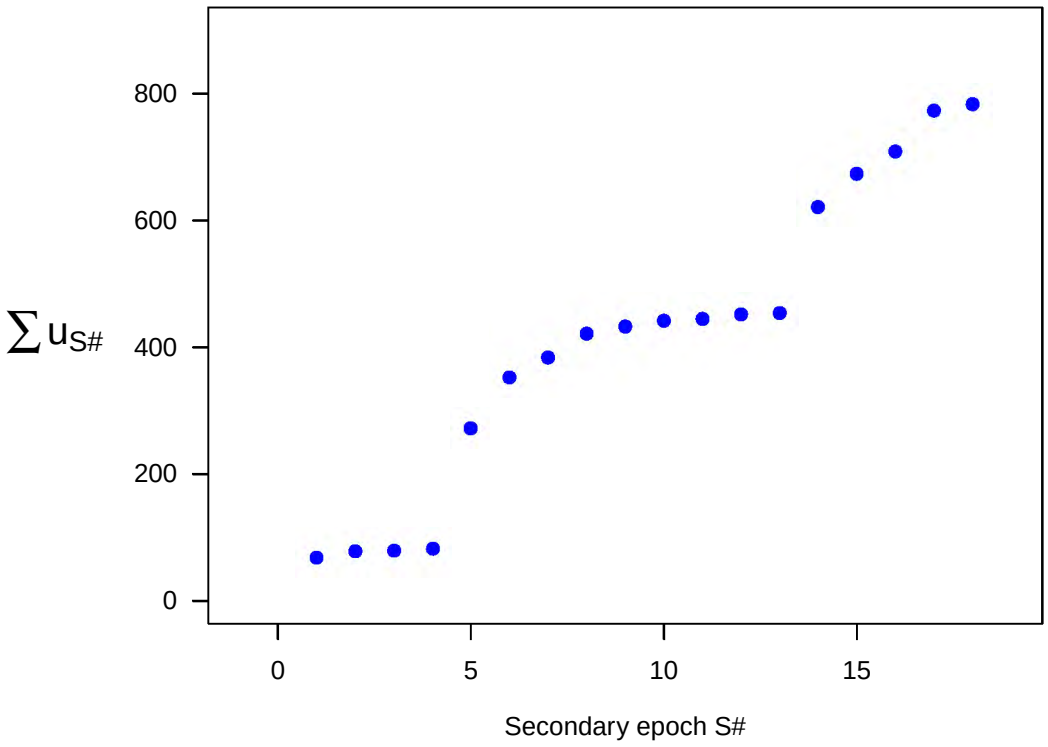


Figure 5.7: Cumulative individuals as a function of each secondary epoch sample S# ($I_{thr} = 18.5$ magnitude).

Table 5.4: Robust analysis result with a comparison of the heterogeneous models, with and without temporal effect (M_h and M_{th}), across the primary OGLE phases II, III, and IV ($I_{thr} = 18.5$ magnitude.)

Closed population model for every epoch:	M_h Gamma			M_{th} Gamma		
Function:	`robustd.t`			`robustd.t`		
	deviance	dof	AIC	deviance	dof	AIC
Model fit:						
fitted model	4589.775	262133	5222.909	3085.686	262118	3748.821
Test for temporary emigration:						
model with temporary emigration:	4487.364	262132	5122.498	2976.665	262117	3641.8
	estimate		std. error	estimate		std. error
Capture probabilities:						
P1	0.1398		0.1539	0.1147		0.1311
P2	0.5124		0.0444	0.4843		0.0446
P3	0.8327		0.0166	0.7799		0.0193
Survival probabilities:						
P1 $\rightarrow$ P2	0.956		0.0411	0.9905		0.0439
P2 $\rightarrow$ P3	1.000		0.0000	1.000		0.0000
Size:						
P1	586.8		648.8	714.7		820.3
P2	802.4		68.7	847.3		77.1
P3	845.7		12.7	875.3		16.2
Number of new arrivals:						
P1 $\rightarrow$ P2	241.4		624.1	139.4		815.9
P2 $\rightarrow$ P3	43.3		69.0	28.0		77.1
Total number of units that ever inhabited survey area:						
all epochs	871.5		36.5	882.1		30.1
Total number of captured units:	783			783		

Recall from §2.4 that the capture probabilities and population sizes are calculated within each (closed) primary epoch. In contrast, the survival probabilities and number of new arrivals are calculated from the transition between primary epochs – to account for the open population dynamics. Figure 5.7 shows the cumulative count of sources encountered as a function of the secondary epoch. There is a characteristic exponential shape present that reappears at the start of every new primary epoch (at S5 and S14), which shows the change in the population's scope at specific instances. This exponential shape was seen in Figure 3.11 and 3.12 for the simulated X-ray binary populations. The rate of increase of newly detected individuals further motivates the use of open population modelling for this application.

Other combinations of heterogeneous fits were tested. Two combinations are presented in Table 5.5 that score very similar on the deviance and AIC selection criteria. The difference is only in the choice of the heterogeneous fit for P1. In Combination 1, P1 is fitted by M_h Chao, which attributes a high capture probability to the epoch and subsequently produces a population size estimate of $\widehat{N}_{\mathrm{P1}} = 92.1 \pm 9.5$, which is more in line with the sample size of P1 (cf. Table 5.1) than the Gamma estimate of $\widehat{N}_{\mathrm{P1}} = 534.8 \pm 22.3$. However, we know that P1 is a truncated sample of the 'superpopulation', and P2 and P3 adds new individuals to the population because they have become 'available' for capture.

The capture probability and population size estimates filter through to the estimated 'new arrivals'. In Combination 1, because the population size estimate is low in P1, the number of new arrivals are considered 'real', whereas Combination 2 suggests that they are included in P1 but were just not captured. The overall size estimates of Combinations 1 and 2 are marginally different. The P1 Gamma fit of Combination 2 leads to a higher estimate, as expected.

When looking back at M_{th} Gamma in Table 5.4, the estimates in P1, P2, and P3 all agree within the standard error bounds, which suggests that the population is closed. This agreement is a fascinating result since we know that the underlying population does not *really* physically change. A cross-check with a closed population analysis is presented in the next section.

Table 5.5: Robust analysis result for different heterogeneous models across the primary OGLE phases: II, III, and IV ($I_{thr} = 18.5$ magnitude.)

Closed population model for each epoch:	Combination 1			Combination 2		
P1	M_h Chao			M_h Gamma		
P2	M_{th} Chao			M_{th} Chao		
P3	M_{th} Darroch			M_{th} Darroch		
	deviance	dof	AIC	deviance	dof	AIC
Model fit:						
fitted model	2622.505	262114	3293.639	2623.405	262116	3290.539
Test for temporary emigration:						
model with temporary emigration:	2622.307	262113	3295.441	2623.208	262115	3292.343
	estimate		std. error	estimate		std. error
Capture probabilities:						
P1	0.8901		0.0856	0.1532		0.0167
P2	0.9192		0.0174	0.9191		0.0174
P3	0.8124		0.0164	0.8124		0.0164
Survival probabilities:						
P1 $\rightarrow$ P2	0.8916		0.0368	0.8915		0.0368
P2 $\rightarrow$ P3	1.0000		0.0000	1.0000		0.0000
Size:						
P1	92.1		9.5	534.8		22.3
P2	476.7		9.9	476.8		9.9
P3	850.3		12.8	850.3		12.8
Number of new arrivals:						
P1 $\rightarrow$ P2	394.6		23.7	0.0		0.0
P2 $\rightarrow$ P3	373.6		25.0	373.6		25.0
Total number of units that ever inhabited survey area:						
all epochs	860.3		12.8	908.4		23.6
Total number of captured units:	783			783		

5.5 Is the population open or is it closed?

In the example in §2.1 and §3.3 of Baillargeon and Rivest (2007), capture data of a snowshoe hare population is investigated in both closed and open population approaches. There is little difference between the two approaches in terms of the model selection criteria. Hence, the possibility of a closed population of hares is not fully discarded in the example. For completeness, I present the closed population size estimates of the Galactic Bulge DNe data, given by `closedp` across the secondary epochs as samples. We already know that there is both heterogeneity and temporal effects in the capture probability amongst the population; therefore, model M_{th} is most likely to provide the preferred fit. Table 5.6 provides the results.

Table 5.6: Closed population size estimates across secondary epochs for $I_{thr} =$ 18.5 magnitude.

Number of captured units: 783

Model	$\widehat{N}$	$\sigma_{\widehat{N}}$	deviance	AIC	BIC
M_0	785.7	1.7	6548.152	6954.932	6964.258
M_t	784.0	1.0	4121.516	4550.296	4610.917
M_h Chao (LB)	995.7	44.0	5262.089	5680.869	5718.174
M_h Poisson2	789.7	2.7	5788.030	6196.811	6210.800
M_h Darroch	867.2	12.8	5448.951	5857.731	5871.721
M_h Gamma3.5	1244.5	66.4	5478.043	5886.823	5900.812
M_{th} Chao (LB)	969.0	38.8	1822.934	2265.714	2358.977
M_{th} Poisson2	785.5	1.6	2875.298	3306.078	3371.362
M_{th} Darroch	937.2	21.4	2021.429	2452.209	2517.493
M_{th} Gamma3.5	2366.0	231.9	2247.882	2678.662	2743.946

Table 5.7: Bias corrected closed population size estimates for the DNe dataset evaluated at a threshold of $I_{thr} = 18.5$ magnitude.

Model	$\widehat{N}_{bc}$	$\sigma_{\widehat{N}_{bc}}$
M_{th} Chao (LB)	977.2	40.0
M_{th} Poisson2	785.7	1.7
M_{th} Darroch	941.6	21.7

The model selection criteria confirm that M_{th} is the preferred model for this dataset. Based on this brief closed population analysis we can see that the M_{th} Chao and Darroch fits (both suitable) are not far off compared to the overall estimates of the robust design analysis (Tables 5.4 & 5.5). The bias-corrected estimates in Table 5.7 do not show significant difference to the M_{th} `closedp` estimates.

We cannot fully discard the validity of a closed population analysis for this dataset. It raises a valid question as to why we should concern ourselves with a robust design analysis and whether our open population claim has merit. However, the robust method performed in this chapter has provided a path forward by presenting an astronomical case study that could aid in the analysis of transient survey populations that do not satisfy the closed population conditions. The robust method has also provided flexibility in the modelling by allowing different models and fits at different epochs and compiling the estimates together into a final population size estimate across the study. In contrast, the closed population method restricts the analysis to the same model across all samples.

5.6 A final note on the analysis

As faced in the closed population analysis in Chapter 4, the same observational limitations for population size estimation occur in this open population analysis. A brightness threshold truncates the sample, which creates a representation bias of the population size to an unknown degree. Similar suggestions could be followed as presented in §4.5. Furthermore, multi-state approaches may help distinguish whether a source has been missed (pointed at but not observed) and whether the source was not pointed at (hence not observed).

5.7 The Galactic DNe population from the literature

Space density, denoted by ρ_0, is a key parameter for modelling and constraining the Galactic CV population. Binary population synthesis studies in the literature have consistently predicted higher space densities for the Galactic CV population compared to observations, noting the seminal works of de Kool (1992) and Politano (1996) with predictions $\rho_0 \simeq 2 \times 10^{-5}$ - 2.0×10^{-4} pc^{-3}.

In attempts to reconcile the observable population with the intrinsic population, studies such as Gänsicke et al. (2009) have shown that there is potentially an abundance of intrinsically faint CVs close to the minimum orbital period which only adds to problem of collecting volume-limited samples (Pretorius, Knigge, and Kolb, 2007; Rau et al.,

2007). This poses a problem for direct comparison of our population size estimates the literature. Another unknown is the fractional composition of sub-populations, such as DNe, of the total Galactic CV population. Sub-populations also present according to different scale heights that stem from binary evolution within the galaxy and thus the Mróz et al. (2015) dataset probes the thin disk population ($-10° < \ell < 10°$) up to a maximum vertical distance of ~ 1.5 kpc (Patterson, 1984). Higher space densities are expected for the long orbital-period type CVs at these low Galactic latitudes but the duration of the OGLE survey (~ 10 years) means that it is insensitive to long orbital-period and low mass transfer rate ($\dot{M}$) systems with long outburst recurrence times (Patterson, 1984; Pretorius, Knigge, and Kolb, 2007). The instrinsic faintness together with extinction and source confusion in the crowded field hinders representative sampling of the population. Recent efforts have been made to compile volume-limited samples and to map the distribution of the Galactic CV population

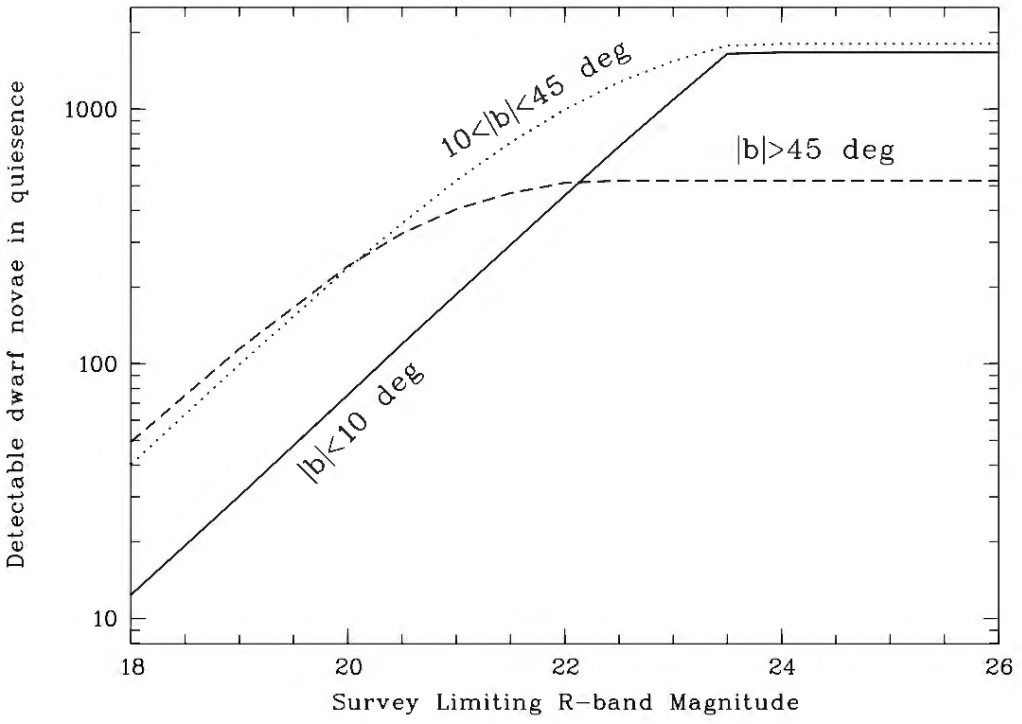


Figure 5.8: Figure 8 taken from Rau et al. (2007) showing predictions of the detectable population of quiescent DN according to the limiting R-band magnitude of a given survey.

Figure 5.8 taken from Rau et al. (2007) shows the expected number of quiescent DN systems that should be detectable brighter than a given limiting R-band magnitude. According to the graph, for latitudes $|b| < 10^{\circ}$ and $R < 21$ magnitude, around 200 quiescent DN should be detectable, whilst only 20 are expected for $R < 18$ magnitude. If we correct for the colour difference between the I and R bands assuming a spectral classification of a K0→M8 type donor star then $I - R \sim -0.42$ to -2.2 (colour corrections calculated from Zombeck (1990, P.69)). Thus, for a limiting $I \sim 21$ magnitude (i.e. that of the OGLE survey), this number increases to a range between ~ 300 and $\simeq 1700$ detectable quiescent sources (Rau et al. (2007) assumes a space density $\rho = 3 \times 10^{-5}$ pc−3). This is in agreement with the detected sample size of 1059 DNe from the Mróz et al. (2015) dataset.

For $I < 18.5$ magnitude we might expect between 30 and 100 detectable quiescent sources according to Figure 5.8. If we define the quiescent state of each source as a $\Delta I \leq 0.5$ magnitude bin that it spends the majority of time over the course of the survey, then $\sim$ 152 DNe from the Mróz et al. (2015) sample are detectable in quiescence for $I < 18.5$ magnitude. This number is slightly larger than predictions in Rau et al. (2007). We note the unusually large fraction of the sample identified with outbursting magnitudes of less than 1 which could be argued to be misclassifications. However, we will assume correct source classification for this discussion. Referring back to Table 5.3, the detected sample by the end of OGLE-IV is around 5 times larger ($N = 783, I < 18.5$)) than those defined detectable at quiescence. The open population size estimates shown in Tables 5.4 and 5.5 range between 860.3 ± 12.8 and 908.4 ± 23.6. This implies that up to $\sim 17\%$ of the population remains undetected in this flux-limited sample. A naive estimate of the space density of DN contained within the solid angle of $|b| < 10^{\circ}, |\ell| < 10^{\circ}$ out to a distance of 1 kpc (based on CV distances measured by *Gaia*-DR1 presented in Ramsay et al. (2017)) results in $\rho \sim 4 \times 10^{-6}$ pc^{-3}, neglecting scaleheight considerations. This estimate is an order of magnitude smaller than the assumed space density used for calculation in Rau et al. (2007). Although, our crude estimate is in agreement with the lower bound of $1.1^{+2.3}_{-0.7} \times 10^{-5}$ pc^{-3} obtained by Pretorius, Knigge, and Kolb (2007).

Unfortunately, the observational data for the CV population at Galactic latitudes of $|b| < 10^{\circ}$ is very incomplete. A large CV sample (722 DN-type and 309 other-type) presented in Coppejans et al. (2016) from the CRTS focuses on latitudes above $|b| > 15^{\circ}$. Pretorius, Knigge, and Kolb (2007) uses the Palomar-Green (survey limited to latitudes $|b| > 30^{\circ}$, Green et al. 1982) sample of CVs to simulate a potentially unseen population of DNe for magnitude limited cases of $V < 20, 16$, and 14 magnitude. The

population size increases by a factor of more than 100 from the limit of $V < 16$ to $V < 20$, most notably at the short orbital-periods, as they constitute the overwhelming majority of the total population. It is not simple to make meaningful comparisons with our results with regards to the intrinsic population located in the thin disk. Given the various caveats of the flux-limited ($I < 18.5$ magnitude) sample, we present our results as lower bound estimates of the intrinsic thin disk population of DNe.

Chapter 6

Discussion

Animal capture-recapture experiments exploit the fact that animals move spatially around an area and redistribute themselves meaningfully between samples. The analogy for capture-recapture in astronomy is cemented in 'brightness variability'. The principle of independent Bernoulli trials[30] is the foundation for multiple capture sampling that leads to statistical inference of the population size. Spatial coordinates (RA and DEC) are the most widely used identifiers in astronomy. They have acted as the "tag" identifier, which remains unchanged for an astronomical source on the assumption that the sources are static. This analogy has been the basis for transferring the capture-recapture application to astronomy as established by Laycock (2017) and similarly applied in this work. The analysis was centred around two capture-recapture approaches, namely closed population and robust design analysis.

6.1 Summary of analyses

The Chapter 3 X-ray lightcurve simulation and exploration of closed population estimator behaviour w.r.t. efficiency, accuracy, and precision has provided us with the following insights:

(Ch3) 1. How to create capture histories from sampled lightcurves, estimate the population size from the capture history using likelihood theory and Poisson regression and apply selection criteria to choose a preferred fit from the model candidates.

(Ch3) 2. The rate of estimation as a function of observation k is improved for low cadences with a large spread in the cadence distribution. The more 'random' the sampling cadence distribution is, the better since it increases the probability of capturing individuals p within the population.

[30]Recall that the standard M_0 model has this condition

(Ch3) 3. Aliasing of the outburst period, orbital period and sampling cadence demonstrate a 'burn-in period' for estimation (an analogy to MCMC sampling where the initial iterations have large bias until the chains 'settle down') for the Schnabel and Schumacher-Eschmeyer estimators. The population estimate appears to converge quickly, but often-times overshoots until a steady slope of convergence is reached. A minimum of 5 sample observations is therefore recommended for a relative threshold of 0.2× maximum outburst magnitude, but preferably up to 10 observations, to overcome biased estimation (cf. Figures 3.11).

   The **Rcapture** MLEs shown in Figures 3.16 and 3.17 show overall that convergence is achieved in fewer observations compared to the Schnabel and Schumacher-Eschmeyer estimators, most crucially for the 7 to -14 day cadence. Wherever possible, sampling should be done randomly, as mentioned in point 2 above.

(Ch3) 4. Assessment of variation in the capture probability is crucial for correct inference of the population size, as estimators under different capture probability frameworks can produce significantly different results. The **Rcapture** software provides tests for distinguishing between the types of capture probability variations, i.e. temporal, behavioural, or heterogeneous cases.

(Ch3) 5. An increased relative brightness threshold decreases the rate of estimation as a function of observation k due to a significantly decreased probability of capture p (cf. Table **??** and **??**).

(Ch3) 6. Model M_b ('behavioural model') is appropriate for high-cadence sampled data (relative to the outburst period/duty cycle of the binary) due to the correlation of captures when in the same outburst cycle.

It was clear that the more modern MLE approaches, such as the estimators implemented in **Rcapture** (§3.3.4), are superior in terms of efficiency and rate of convergence to the true population. The number of observations required using **Rcapture** was nearly half that of the Schnabel and Schumacher-Eschmeyer estimators for the same degree of convergence across the cadences regimes.

Chapters 4 and 5 assumed a fixed number of observations from Chapter 3 w.r.t observation strategy and the number of observations needed for estimating convergence within specific bounds of the underlying population. For both of the real-data applications, the main difference in the analysis from Chapter 3 was the use of an apparent magnitude scale rather than a relative magnitude (or flux) scale to generate the capture histories. Generating a capture history from astronomical data is non-trivial. Scaling to the same quiescent magnitude may be needed but will ultimately depend on the desired

outcome, i.e. whether the aim is to estimate the total population or the population as a function of magnitude (brightness) density. In Chapter 4, there was some focus on the cumulative sources captured and the population estimations as a function of the applied detection threshold. In contrast to a relative flux (or magnitude) scaling, this has now forced consideration for handling sources that are always captured when the detection threshold is below their minimum brightness and the bias in estimation that it incurs. On a point of modelling, no consideration was given in Chapter 4 for two- (or more) factor interactions (as referred to in §2.5.1) that uses *design-based* modelling to counter heterogeneity. Bias corrections significantly improved convergence and estimation in the XROM data. Further insights gained in the Chapter 4 XROM closed population analysis included:

(Ch4) 1. Reduction of lightcurve data into corresponding capture histories via pooling, thereby eliminating issues with data imputation for sources that were not sampled at a specific point in time.

(Ch4) 2. Assessment of the source of variation in capture probability, particularly heterogeneity, within the population using methodology from **Rcapture**.

(Ch4) 3. Observing the limitations of the heterogeneous estimators such as M_h Poisson and Chao, which often only produced lower bound estimates of the population, in line with the captured amount of individuals.

(Ch4) 4. Bias corrected estimates were imperative in this context due to large biases incurred from the degree of heterogeneity.

In Chapter 5, the robust analysis method was proposed for estimating population size to allow for continuity in estimation across survey stages of OGLE Galactic DN lightcurves. Surveys that stretch across decades may be affected by multiple changes, as previously discussed. Adding new individuals to a survey phase (or dropping some) may influence estimation, the reason being that in some instances, implicit changes may have been made to the sampling design or where closed population conditions were violated. Such sampling design changes may motivate towards assuming an open population study to do a coherent population analysis. The dataset of DNe located in the direction of the Galactic Bulge chronicled survey changes, technical and in OGLE across its phase II to III, and III to IV. Several issues cropped up with the robust population analysis in Chapter 5:

(Ch5) 1. The capture histories created for the robust analysis was akin to the reduction of the XROM data. Captures were grouped via a logical `OR` function across a specified epoch, ignoring the occasions of 'missing data' (so-called because the

source was not pointed at and therefore we cannot make a claim about its state of capture). Multi-state models may partially address this issue; however, this function is not available within **Rcapture**.

(Ch5) 2. Seber and Schofield (2019) comment that the **Rcapture** open population models lack the flexibility for heterogeneity modelling and interpreting the results when there is departure from the underlying assumptions. The robust design method, however, has merit for use in transient population estimation by joining the population modelling across otherwise distinct epochs. Future changes can be made for the inclusion of multi-state captures, i.e. a ternary description may be: 'observed but not captured', 'observed and captured', 'not observed and not captured'. It will require the use of modified software to assist in heterogeneity modelling.

(Ch5) 3. **Rcapture** did not provide bias-corrections in the open population modelling as for the closed population, which has been shown necessary for analysis using a magnitude flux scale instead of a relative flux scale.

6.2 General considerations

Several questions remain that will need continuous consideration. The first is:

C1. Biases in estimation from truncated and censored survey samples.

The simulated set of HMXB X-ray lightcurves in Chapter 3 provided us with control over a 'truth value' for the population size. The sampling procedure of the simulated populations was representative of the underlying population due to the simulation setup. We similarly expect the OGLE-IV sub-sample of HMXBs in the SMC to be representative because the population has been well-studied across the electromagnetic spectrum. For the DN dataset located towards the Galactic Bulge, representativeness is less certain. Distances to each Dwarf Nova (DN) source may differ greatly, and foreground extinction due to gas and dust may obscure a large number of DN sources that creates this uncertainty in population representativeness. The selection criteria for DNe were laid out in §5.2 that creates an inherently truncated dataset.

C2. Identification of classes of transient sources is not done at the time of observation.

In contrast to an ecological capture-recapture experiment where an individual can be marked upon capture because it is recognisable based on certain characteristic features, astronomical transients can often not be classified using a single photometric datum. X-ray binary transients, for example, are classified and sub-classified through a multi-step process. It starts with detecting the X-ray source, with or without X-ray pulsations and follows with optical counterpart identification and long-term photometric monitoring. Also, spectra are analysed to identify defining features such as the presence of a disc and the type of compact object for confirmation on the sub-type (Prestwich et al., 2003; Soria, Cropper, and Motch, 2005; Negueruela and Schurch, 2007; Gopalan, Vrtilek, and Bornn, 2015).

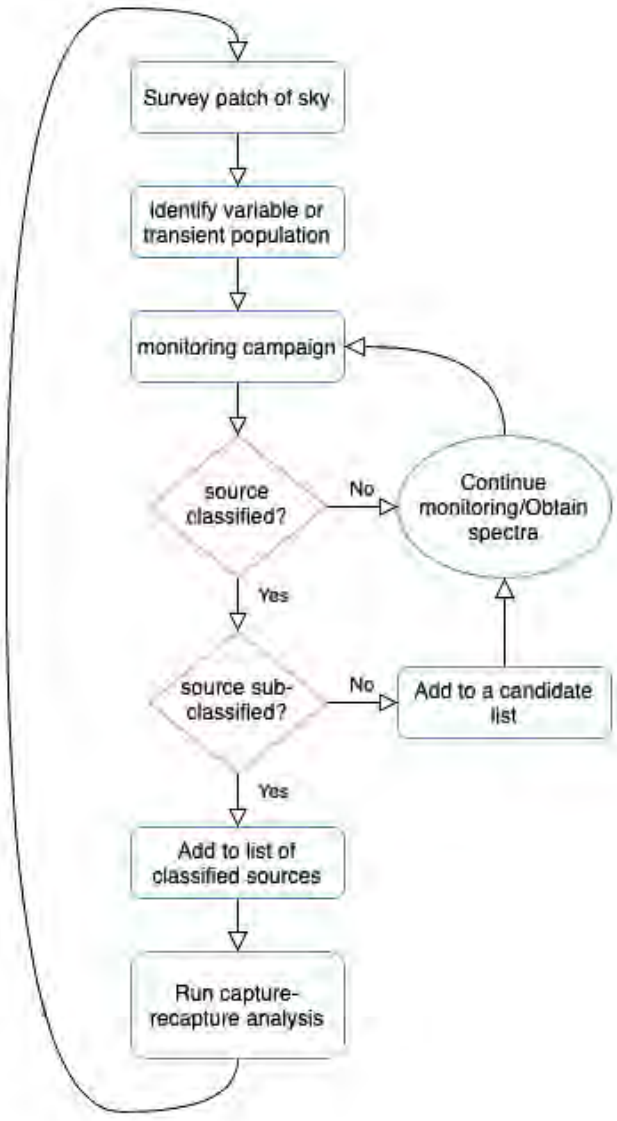


Figure 6.1: Flowchart of the pipeline from observation to population size estimation.

We recognise that for both points made above, population size estimation in astronomy will be an iterative process (cf. Figure 6.1) involving recurrent use of old data, with new data incorporated into the analysis once it becomes available (Laycock, 2017). Since it is not typical to classify sources from a single observation, iterative estimates will need to be computed as sources are categorised. The need for classification also invites the use of fused data from various sources (catalogues, surveys, multi-wavelength regimes and more) that may provide valuable information that could directly or indirectly contribute to population size estimation because of better representativeness.

6.3 The many different approaches to capture-recapture

Chapter 2 gave a general overview of the introductory closed and open population approaches. However, regarding Figure 2.1, it is evident that we have thus far only covered the theory of, and have explored, a small fraction of what capture-recapture methods have to offer. First of all, model frameworks, M_{tb} and M_{tbh} were not considered in this work merely because of the focus on log-linear regression for estimation and the choice of software, and the fact that neither model can be represented in log-linear form. Although this is a minor concern, it should be followed up with other available capture-recapture software, particularly for the high-cadence sampling (relative to orbital period) where temporal and behavioural effects were present, but not heterogeneity.

Of particular interest may be the logit models[31] by Huggins (1989, 1991) that incorporate covariates that explain unequal "catchability" in the population. These models could add physical astronomical variables, whether categorical or continuous, to the framework as statistical weights to determine capture probabilities. Parameters relevant to this work are orbital period and duty cycle (outburst time as a fraction of the orbital period) (Chao and Huggins, 2005*a*).

Another practical approach to incorporate with capture-recapture is recurrent event analysis which is, as Chao and Huggins (2005*b*) mention, two sides of the same coin. Recurrent event analysis is a standard technique used by astronomers (see Oppermann, Yu, and Pen, 2018; Connor, Pen, and Oppermann, 2016; Bird et al., 2009). In this case, the captures are described via a Poisson process (or deviation from the Poisson process) across the epoch $(0, T)$ with a rate of λ. The capture probability may then

[31] See footnote 22 on p. 40

be described as approximately $p \approx \lambda\, dt$ across interval $(t, t + dt)$. Captures can only be described as independent for non-overlapping intervals. The former describes the continuous model's basic framework under the assumption of homogeneity (Chao and Huggins, 2005*b*). Likelihood can be determined henceforth. A general overview of the topic pertaining to the biological and health sciences may be found in Cook and Lawless (2007).

Lastly, an emerging field called spatial capture-recapture (SCR) could potentially be applied to astronomical populations w.r.t. magnitude (brightness) density estimation where the latter can be analogous to spatial density estimation for ecological populations (Royle et al., 2013). Numerous ad-hoc methods exist to deal with animals' spatial density estimation but are often problematic according to Royle et al. (2013). The analysis in Chapters 4 and 5 showed clear signs of heterogeneity, not only because the sources in the population demonstrate different duty cycles but also because they are *location*-bound within magnitude space (i.e. they have a well-defined magnitude distribution). SCR could provide a formal solution to this issue by constructing a well-defined sample area in magnitude space (Royle et al., 2013).

Chapter 7

Conclusion

In this work, we set out to explore uses and their implementation of capture-recapture techniques for population size estimation in the context of time-domain astronomy. This exploratory work was undertaken with a view of developing a methodology for estimating the intrinsic population of stellar transients in addition to population synthesis modelling. We developed an understanding of estimator model and their behaviours in several astronomical contexts as well as in the interpretation of the resulting estimates to select the best model.

Chapter 3 reproduced the results of Laycock (2017) through the strategic sampling of simulated X-ray lightcurves of six different HMXB population models; each defined by its corresponding orbital period distribution. We presented detailed methodology on the epoch downsampling of high-cadence data to improve population size estimates along with non-standard modelling of the capture probability given tests for the presence of heterogeneous, behavioural, or temporal effects.

In Chapter 4, we developed a strategy for the reduction and analysis of survey data in the context of a closed population as represented by a monitored sample population of HMXBs situated in the SMC. We investigated several models, of which M_h appears to be most applicable to populations distributed widely in quiescent magnitudes. The estimators availalbe under the M_h model-framework were compared, noting that the Chao and Poisson estimators represent lower bounds. The Darroch estimate was discussed at length as a choice model. Bias-corrected estimates were investigated and shown to be potentially important in determining the intrinsic population.

Chapter 5 saw the use of a 'Robust-Design' approach to account for a changing population of Galactic DNe between phases of a survey; due to redesign of the sky survey area

or improved capacity to detecting new sources. A magnitude limited analysis suggested that up to $\sim$ 17% of the population had gone undetected for $I < 18.5$ magnitude. Comparisons were drawn with the literature and the space density of the Galactic DNe population estimates using crude assumptions. Our estimate corresponds to a lower bound of that obtained by Pretorius, Knigge, and Kolb (2007).

In these three contrasting datasets and their corresponding analysis we have shown the potential for capture-recapture to be used within time-domain astronomy for several population size estimation applications.

Appendix A

Log-linear representations of Rcapture models

Table A.1: Parameter definitions for log-linear relations and bias corrections.[32]

Parameter	**Definition**
t	total number of capture occasions
t_0	specified capture occasion ($t_0 > 1$)
j	jth capture occasion ($j = 1, 2, \ldots, t$ unless otherwise specified)
i	ith individual
N	population size
n	the number of individuals captured at least once
p	capture probability
p_j	capture probability at occasion j
ω	capture history
ω_j	capture history at jth occasion
$\sum \omega_j$	number of times an individual is caught
μ_ω	expected number of individuals with capture history ω
μ_0	expected number of individuals that were unobserved
γ, β	intercept and slope parameters estimated through log-linear regression
γ_C, β_C	intercept and slope parameters estimated through log-linear regression for the Chao estimator
ψ_m	model used for determining heterogeneity, e.g. Darroch, Poisson, or Gamma.
τ_m	degree (or size) or the heterogeneity parameter
a	base exponent value of a Poisson model, $\psi_m = (a^m - 1)\,,\ a > 0$
$\phi(t)$	log-Laplace transform of $\log(a)X$ where a is a positive number and X is a random variable with a Poisson distribution
$F(x)$	A combination of distributions that characterises the capture probability of individuals.
$G(x)$	distribution dependent on $F(x)$ and used to determine $\phi(t)$
λ	base exponent value of a Gamma model, $\psi_m = -\log(\lambda + k) + \log(\lambda)$
b	bias
$E(Y)$	expectation of variable Y
f_j	the number of individuals caught exactly j times
$p*$	the probability of being caught at least once ($N = \mu_0/(1 - p^*)$ for Chao estimator)
c_λ	Scale parameter used in the Gamma model: $c_\lambda = \log\{1 - 1/(\lambda + 1)^2\} / \log\{1 - 1/(\lambda + 2)^2\} > 1$
u_j	number of individuals caught for the first time at jth occasion

[32] All models are characterised by the form $\hat{N} = n + \exp\hat{\gamma}$, and variance estimator $v(\hat{N}) = \exp\hat{\gamma} + (\exp 2\hat{\gamma}) v_P(\hat{\gamma})$, where v_P is the Poisson variance.

Table A.2: Log-linear form of each model fitted within `closedp`. See Rivest and Lévesque (2001); Rivest and Baillargeon (2007) for full derivations.

Model	**Log-linear format**
M_0 (See p. 52)	$$\log(\mu_{\boldsymbol{\omega}}) = \gamma + \beta \sum_j^t \omega_j,$$ $$\gamma = \log\left\{N(1-p)^t\right\}, \quad \beta = \log\left\{p/(1-p)\right\}.$$
M_t	$$\log(\mu_\omega) = \gamma + \sum_j^t \omega_j \beta_j$$ $$\gamma = \log\left\{N \prod_j^t (1-p_j)\right\}, \quad \beta_j = \log\left\{p_j/(1-p_j)\right\}.$$
M_h(Chao (LB)) (See Eqns. 2.34 and 2.36.)	$$\log \mu_\omega = \gamma_C + \beta_C \sum_{j=1}^t \omega_j + \sum_{m=3}^t \psi_m \left(\sum_{j=1}^t \omega_j\right) \tau_m,$$ $$\gamma_C = \gamma + 2\varphi(1) - \varphi(2), \quad \beta_C = \beta + \varphi(2) - \varphi(1),$$ $$\tau_m = \varphi(m) - (m-1)\varphi(2) + (m-2)\varphi(1),$$ $$\psi_m(x) = \begin{cases} 1, \text{ if } x = m; \\ 0, \text{ otherwise. (for } m = 3, \ldots, t) \end{cases}$$ $$\varphi(t) = \log\left\{\int_R \exp(tx) dG(x)\right\},$$ $$dG(x) \sim \frac{dF(x)}{\prod_{j=1}^t \{1 + \exp(\beta + x)\}}.$$

M_h(Poisson2)	$\mu_\omega = \exp\left\{\gamma + \beta\sum_j^t \omega_j + \tau\psi\left(\sum_j^t \omega_j\right)\right\}$, $\psi(k) = (a^k - 1)$, Condition: $\psi(3) - 2\psi(2) + \psi(1) = 1$, so that $\tau = \tau_3$ as for M_h(Chao).
M_h(Darroch)	$\mu_\omega = \exp\left\{\gamma + \beta\sum_j^t \omega_j + \tau\psi\left(\sum_j^t \omega_j\right)\right\}$, $\psi(k) = k^2/2$, Condition: $\psi(3) - 2\psi(2) + \psi(1) = 1$, so that $\tau = \tau_3$ as for M_h(Chao).
M_h(Gamma3.5)	$\mu_\omega = \exp\left\{\gamma + \beta\sum_j^t \omega_j + \tau\psi\left(\sum_j^t \omega_j\right)\right\}$, $\psi(k) = -\log(\lambda + k) + \log(\lambda), \quad (k = 1, \ldots, t-1)$, Condition: $\psi(3) - 2\psi(2) + \psi(1) = 1$, so that $\tau = \tau_3$ as for M_h(Chao).

M_{th}(Chao (LB))	$$\log \mu_\omega = \gamma_C + \sum_{j=1}^{t} \omega_j \beta_{Cj} + \sum_{m=3}^{t} \psi_m \left(\sum \omega_j \right) \tau_m,$$ $$\gamma_C = \gamma + 2\varphi(1) - \varphi(2), \quad \beta_{Cj} = \beta_j + \varphi(2) - \varphi(1),$$ $$\tau_m = \varphi(m) - (m-1)\varphi(2) + (m-2)\varphi(1),$$ $$\psi_m(x) = \begin{cases} 1, \text{ if } x = m; \\ 0, \text{ otherwise. (for } m = 3, \ldots, t) \end{cases}$$ $$\varphi(t) = \log \left\{ \int_R \exp(tx) dG(x) \right\},$$ $$dG(x) \sim \frac{dF(x)}{\prod_{j=1}^{t} \{1 + \exp(\beta_j + x)\}}.$$
M_{th}(Poisson2)	$$\mu_\omega = \exp \left\{ \gamma + \sum_{j}^{t} \omega_j \beta_j + \tau \psi \left(\sum_{j}^{t} \omega_j \right) \right\},$$ $$\psi(k) = (a^k - 1),$$ Condition: $\psi(3) - 2\psi(2) + \psi(1) = 1$, so that $\tau = \tau_3$ as for M_{th}(Chao).
M_{th}(Darroch)	$$\mu_\omega = \exp \left\{ \gamma + \sum_{j}^{t} \omega_j \beta_j + \tau \psi \left(\sum_{j}^{t} \omega_j \right) \right\},$$ $$\psi(k) = k^2/2,$$ Condition: $\psi(3) - 2\psi(2) + \psi(1) = 1$, so that $\tau = \tau_3$ as for M_{th}(Chao).

M_{th}(Gamma3.5)	$\mu_\omega = \exp\left\{\gamma + \sum_j^t \omega_j\beta_j + \tau\psi\left(\sum_j^t \omega_j\right)\right\}$, $\psi(k) = -\log(\lambda + k) + \log(\lambda), \quad (k = 1, \ldots, t-1)$, Condition: $\psi(3) - 2\psi(2) + \psi(1) = 1$, so that $\tau = \tau_3$ as for M_{th}(Chao).
M_b	$\log(\mu_j) = \gamma + (j-1)\beta, \quad j = 1, \ldots, t - t_0$; $\gamma = \log(Np), \quad \beta = \log(1-p)$.
M_{bh}	$\log(\mu_{j+t_0}) = \gamma + (j-1)\beta, \quad j = 1, \ldots, t - t_0$; $\gamma = \log(N_1 p), \ \beta = \log(1-p)$, where p is capture probability on occasions $t_0 + 1, \ldots, t$.

Table A.3: Estimator bias corrections using the Rcapture `closedp.bc` function for each model produced by `closedp`. See Rivest and Lévesque (2001); Rivest and Baillargeon (2007) for full derivations.

Model	**Bias** b
M_0	Asymptotic bias: $b = \frac{(t-1)p(1-p)^{t-1}\{1-(1-p)^t\}}{2\{1-(1-p)^t - tp(1-p)^{t-1}\}^2}$ OR Frequency modifications: $\sum_j \omega_j = 1, \quad b = -1/2t$ $\sum_j \omega_j = 2, \quad b = 2/\{t(t-1)\}$ $\sum_j \omega_j > 2, \quad b = 0$
M_t	Asymptotic bias: $b = \frac{1}{2}\frac{\left(\sum_{j=1}^t \frac{p_j}{1-p_j}\right)^2 - \sum_{j=1}^t \left(\frac{p_j}{1-p_j}\right)^2}{\left(\frac{1}{\prod_{j=1}^t (1-p_j)} - 1 - \sum_{j=1}^t \frac{p_j}{1-p_j}\right)^2}$ OR Frequency modifications: $\sum_j \omega_j = 1, \quad b = (t-2)/2t$ $\sum_j \omega_j = 2, \quad b = 2/\{t(t-1)\}$ $\sum_j \omega_j > 2, \quad b = 0.$
M_h(Chao (LB))	$b = E(f_0) - \frac{(t-1)E(f_1)^2}{2tE(f_2)}$

M_h(Poisson2)	Bias relative to Chao (LB) estimate: $(1-p^*)\{\exp(-\tau/a)-1\}$ where $\tau=\tau_3$ as for M_h(Chao)
M_h(Darroch)	Bias relative to Chao (LB) estimate: $(1-p^*)\{\exp(-\tau)-1\}$ where $\tau=\tau_3$ as for M_h(Chao)
M_h(Gamma3.5)	Bias relative to Chao (LB) estimate: $(1-p^*)\{\exp(-c_\lambda\tau)-1\}$ where $\tau=\tau_3$ as for M_h(Chao)
M_{th}(Chao (LB))	Frequency modifications: $\sum_j \omega_j = 1, \quad b=(t-2)/2t$ $\sum_j \omega_j = 2, \quad b=2/\{t(t-1)\}$ $\sum_j \omega_j > 2, \quad b=0$
M_{th}(Poisson2)	Bias relative to Chao (LB) estimate: $(1-p^*)\{\exp(-\tau/a)-1\}$ where $\tau=\tau_3$ as for M_h(Chao) OR

	Frequency modifications: $\sum_j \omega_j = 1, \quad b = (t-2)/2t$ $\sum_j \omega_j = 2, \quad b = 2/\{t(t-1)\}$ $\sum_j \omega_j > 2, \quad b = 0$
M_{th}(Darroch)	Bias relative to Chao (LB) estimate: $(1-p^*)\{\exp(-\tau)-1\}$ where $\tau = \tau_3$ as for M_h(Chao)
M_{th}(Gamma3.5)	Bias relative to Chao (LB) estimate: $(1-p^*)\{\exp(-c_\lambda \tau)-1\}$ where $\tau = \tau_3$ as for M_h(Chao)
M_b	Asymptotic bias: $b = \frac{t(1-p)^{t+1}\left[\frac{t\{1+(1-p)^t\}}{1-(1-p)^t} - \frac{2-p}{p}\right]}{2p^2\{1-(1-p)^t\}^2\left[\frac{1-p}{p^2} - \frac{t^2(1-p)^t}{\{1-(1-p)^t\}^2}\right]^2}$ OR Frequency modifications: $t - t_0 < 6, \quad b = u_j + (2,-1,-1)$ $t - t_0 \geq 6, \quad b = u_j + (3,-1,-1,-1)$
M_{bh}	Frequency modifications: $t - t_0 < 6, \quad b = u_j + (2,-1,-1)$ $t - t_0 \geq 6, \quad b = u_j + (3,-1,-1,-1)$

Appendix B

OGLE XROM Sources

Table B.1: OGLE-III XROM identification and location information

	Name	**Field**	**Star No**	**RA (J2000)**	**Dec (J2000)**
1	SXP0.09	SMC128.4	118	0:42:35.40	-73:40:34.2
2	SXP0.92	SMC125.3	19461	0:45:35.28	-73:19:02.9
3	SXP31.0	SMC116.6	33	1:11:08.53	-73:16:46.0
4	SXP22.07	SMC116.1	3322	1:17:40.18	-73:30:50.5
5	SXP967	SMC114.7	39	1:02:06.66	-71:41:15.9
6	SXP455B	SMC113.7	18374	1:01:20.64	-72:11:18.6
7	SXP1323	SMC113.6	27699	1:03:37.50	-72:01:32.9
8	SXP726	SMC113.3	10946	1:05:55.22	-72:03:50.3
9	AZV285	SMC110.5	5	1:01:55.78	-72:32:36.0
10	SXP2.763	SMC109.2	5	0:59:12.74	-71:38:44.8
11	SXP46.6	SMC108.8	30	0:53:55.28	-72:26:45.1
12	SXP140	SMC108.8	35838	0:56:05.53	-72:21:59.2
13	SXP342	SMC108.8	33	0:54:03.84	-72:26:32.9
14	MA93-798	SMC108.8	6254	0:54:46.37	-72:25:22.6
15	SXP34.08	SMC108.7	24657	0:55:28.46	-72:10:56.5
16	SXP152.1	SMC108.3	36	0:57:50.38	-72:07:55.9
17	SXP280.4	SMC108.3	16091	0:57:49.58	-72:02:35.7
18	SXP304	SMC108.3	12190	1:01:02.88	-72:06:58.7
19	SXP455A	SMC108.2	34801	1:01:20.65	-72:11:18.6
20	SXP202	SMC108.1	4929	0:59:21.03	-72:23:17.0
21	SXP565	SMC108.1	19293	0:57:35.99	-72:19:33.7
22	SXP2.37	SMC107.5	25	0:54:33.47	-73:41:01.1
23	SXP101	SMC106.7	15343	0:57:27.03	-73:25:19.0
24	SXP504	SMC105.7	36877	0:54:55.87	-72:45:10.7
25	SXP701	SMC105.6	36015	0:55:18.43	-72:38:51.8
26	SXP59.0	SMC105.5	35420	0:54:56.16	-72:26:47.6
27	SXP645	SMC105.5	35415	0:55:35.14	-72:29:06.5
28	SXP293	SMC105.4	22335	0:58:12.58	-72:30:48.4
29	SXP202B	SMC105.3	29894	0:59:28.66	-72:37:03.9
30	SXP169.3	SMC102.2	13410	0:52:55.28	-71:58:06.0
31	SXP91.1	SMC102.1	32	0:50:56.99	-72:13:34.2
32	SXP18.3	SMC101.8	19552	0:49:11.45	-72:49:37.1
33	SXP327	SMC101.4	25097	0:52:52.21	-72:17:14.8
34	SXP7.78	SMC101.3	33277	0:52:05.61	-72:26:03.9
35	SXP8.80	SMC101.3	4737	0:51:53.12	-72:31:48.3

36	SXP138	SMC101.3	38781	0:53:23.85	-72:27:15.1
37	SXP82.4	SMC101.2	12997	0:52:08.94	-72:38:02.9
38	SXP756	SMC100.8	22903	0:49:42.02	-73:23:14.2
39	SXP9.13	SMC100.7	63223	0:49:13.63	-73:11:37.4
40	SXP25.5	SMC100.7	54315	0:48:14.14	-73:10:03.4
41	SXP264	SMC100.7	45007	0:47:23.36	-73:12:26.9
42	SXP893	SMC100.7	63151	0:49:29.81	-73:10:58.0
43	SXP74.7	SMC100.5	55158	0:49:03.36	-72:50:52.0
44	SXP172	SMC100.2	44100	0:51:52.01	-73:10:33.7
45	SXP323	SMC100.2	48	0:50:44.71	-73:16:05.0
46	SXP15.3	SMC100.1	48026	0:52:13.99	-73:19:18.3
47	RX-J0052.1-7319	SMC100.1	48068	0:52:15.39	-73:19:15.0
48	OGLE-SMC-SC5-180008	SMC100.1	30389	0:50:48.00	-73:18:17.6
49	RX-J0544.1-7100	LMC186.6	38	5:44:05.14	-71:00:49.7
50	XMMU-J054134.7-682550	LMC182.5	6	5:41:34.36	-68:25:48.2
51	CAL-83	LMC182.5	11204	5:43:34.22	-68:22:22.0
52	RX-J0513.9-6951	LMC112.2	48890	5:13:50.88	-69:51:47.6
53	RX-J0516.0-6916	LMC100.7	52359	5:16:00.03	-69:16:08.4
54	RX-J0520.5-6932	LMC100.1	14805	5:20:29.97	-69:31:55.8

Table B.2: OGLE-IV XROM identification and location information

	Name	**Field**	**Star No**	**RA (J2000)**	**Dec (J2000)**
1	XMMJ013250.6-742544	SMC739.11	1265	1:32:51.43	-74:25:45.2
2	SXP1062	SMC738.07	1966	1:27:45.96	-73:32:56.4
3	SXP31.0B	SMC733.32	102	1:11:08.59	-73:16:46.3
4	RX_J0123.4-7321	SMC733.26	24	1:23:27.42	-73:21:22.3
5	SXP22.07	SMC733.21	3978	1:17:40.16	-73:30:50.6
6	IGR_J01217-7257	SMC732.03	3540	1:21:40.61	-72:57:30.9
7	LIN_358	SMC728.27	1904	0:59:12.21	-75:05:17.6
8	SXP202BB	SMC726.31	28921	0:59:28.67	-72:37:04.2
9	AZV285	SMC726.30	25187	1:01:55.80	-72:32:36.4
10	SXP6.85	SMC726.29	18	1:02:53.31	-72:44:35.1
11	SXP164	SMC726.28	23178	1:04:29.12	-72:31:37.1
12	SXP65.8	SMC726.27	20665	1:07:12.60	-72:35:33.9
13	SXP11.5	SMC726.20	19470	1:04:42.30	-72:54:04.2
14	SXP101	SMC726.15	77	0:57:27.06	-73:25:19.4
15	SXP31.0A	SMC726.08	69	1:11:08.58	-73:16:46.3
16	SXP2.763B	SMC725.25	11632	0:59:12.74	-71:38:44.9
17	SXP967	SMC725.23	10589	1:02:06.67	-71:41:16.2
18	SXP152.1B	SMC725.16	363	0:57:50.39	-72:07:56.3
19	SXP304	SMC725.15	85	1:01:02.89	-72:06:59.1
20	SXP455	SMC725.15	42	1:01:20.66	-72:11:19.1
21	SXP348	SMC725.14	251	1:03:13.91	-72:09:14.3
22	SXP1323	SMC725.14	19468	1:03:37.54	-72:01:33.2
23	SXP523	SMC725.14	375	1:02:47.58	-72:04:51.4
24	SXP3.34	SMC725.13	54	1:05:02.53	-72:10:56.2
25	SXP726	SMC725.12	86	1:05:55.25	-72:03:50.6
26	XMMUJ010743.1-715953	SMC725.12	12425	1:07:43.42	-71:59:53.9
27	SXP175	SMC725.06	151	1:01:52.29	-72:23:34.1
28	SWIFT_J01077-7228	SMC725.04	92	1:07:44.25	-72:27:40.9
29	SXP0.92	SMC720.29	467	0:45:35.26	-73:19:03.4
30	SXP9.13	SMC720.28	40399	0:49:13.61	-73:11:37.8
31	SXP25.5	SMC720.28	40482	0:48:14.10	-73:10:03.9
32	SXP11.87	SMC720.28	190	0:48:14.04	-73:22:04.0
33	SXP323	SMC720.27	43155	0:50:44.71	-73:16:05.4
34	SXP756	SMC720.27	17	0:49:42.00	-73:23:14.6
35	OGLE-SMC-SC5-180008	SMC720.27	303	0:50:47.99	-73:18:18.0

36	SXP15.3	SMC720.26	47	0:52:13.99	-73:19:18.8
37	SXP172	SMC720.26	36274	0:51:52.02	-73:10:34.0
38	RX-J0052.1-7319	SMC720.26	531	0:52:15.39	-73:19:15.4
39	SXP0.09	SMC720.23	178	0:42:35.37	-73:40:34.8
40	SMC_3	SMC720.20	22115	0:48:20.01	-73:31:52.1
41	SXP2.37	SMC720.17	50	0:54:33.44	-73:41:01.3
42	SXP146.6	SMC720.12	13583	0:45:17.77	-73:47:05.7
43	XMM_J004855-73494	SMC720.11	13342	0:48:55.36	-73:49:45.7
44	SXP169.3	SMC719.28	16997	0:52:55.28	-71:58:06.3
45	SXP34.08	SMC719.27	338	0:55:28.44	-72:10:56.8
46	SXP152.1A	SMC719.26	179	0:57:50.38	-72:07:56.3
47	SXP280.4	SMC719.26	19885	0:57:49.57	-72:02:36.1
48	SXP91.1	SMC719.21	21951	0:50:56.99	-72:13:34.4
49	SXP4.78	SMC719.21	22049	0:51:38.79	-72:17:04.8
50	SXP7.78	SMC719.20	144	0:52:05.63	-72:26:04.1
51	SXP46.6	SMC719.20	34	0:53:55.32	-72:26:45.3
52	SXP138	SMC719.20	325	0:53:23.86	-72:27:15.4
53	SXP327	SMC719.20	27652	0:52:52.23	-72:17:15.1
54	SXP6.88	SMC719.19	181	0:54:46.37	-72:25:22.8
55	SXP59.0	SMC719.19	162	0:54:56.18	-72:26:47.8
56	SXP342	SMC719.19	166	0:54:03.86	-72:26:32.8
57	SXP645	SMC719.19	125	0:55:35.15	-72:29:06.6
58	SXP7.92	SMC719.18	465	0:57:58.51	-72:22:29.2
59	SXP140	SMC719.18	27492	0:56:05.54	-72:21:59.6
60	SXP565	SMC719.18	27563	0:57:36.01	-72:19:34.1
61	IGR_J00569-7226	SMC719.18	378	0:57:02.19	-72:25:55.4
62	SXP202	SMC719.17	49	0:59:21.03	-72:23:17.4
63	SXP293	SMC719.17	2	0:58:12.59	-72:30:48.8
64	SXP8.80	SMC719.12	35646	0:51:53.16	-72:31:48.6
65	SXP82.4	SMC719.11	36482	0:52:08.95	-72:38:03.2
66	SXP504	SMC719.10	139	0:54:55.88	-72:45:10.8
67	SXP701	SMC719.10	33543	0:55:18.44	-72:38:51.9
68	SXP202BA	SMC719.08	27301	0:59:28.67	-72:37:04.2
69	SXP74.7	SMC719.05	41432	0:49:03.33	-72:50:52.5
70	SXP214	SMC719.04	754	0:50:11.26	-73:00:26.0
71	SXP2.763A	SMC718.01	10792	0:59:12.74	-71:38:44.9
72	MAXI_J0158-744	MBR109.24	18	1:59:25.87	-74:15:28.0
73	SWIFT_J0208.4-7428	MBR109.12	1333	2:06:45.16	-74:27:47.6

74	IGR_J015712-7259B	MBR108.25	11	1:57:16.19	-72:58:32.4
75	IGR_J015712-7259A	MBR102.01	984	1:57:16.17	-72:58:32.5
76	CAL83	LMC554.12	404	5:43:34.17	-68:22:22.1
77	LXP61.6	LMC554.06	18693	5:41:34.32	-68:25:48.5
78	LXP4.40	LMC553.23	35075	5:41:26.66	-69:01:23.7
79	LXP91.1	LMC552.06	6898	5:44:05.22	-71:00:50.8
80	LXP187	LMC531.05	4251	4:51:06.86	-69:48:03.2
81	SWIFT_J0456.0-7020	LMC530.18	23149	4:55:58.83	-70:20:00.0
82	LXP272	LMC519.29	9750	5:30:11.38	-65:51:23.9
83	LXP69.5	LMC519.22	10153	5:29:47.84	-65:56:43.7
84	1A0538-66	LMC518.26	21298	5:35:41.00	-66:51:53.7
85	LXP4.10	LMC512.15	25	5:02:51.74	-66:26:26.9
86	LXP4.4B	LMC511.17	17	5:13:42.61	-67:24:10.0
87	LXP169	LMC510.20	50751	5:07:55.47	-68:25:05.3
88	LXP956	LMC510.17	45733	5:12:59.98	-68:26:38.0
89	IGRJ05007-7047	LMC508.31	62	5:00:46.06	-70:44:36.1
90	SWIFT_J0513.4-6547	LMC506.16	16	5:13:28.26	-65:47:18.7
91	LXP4.4A	LMC505.07	38	5:13:42.61	-67:24:10.1
92	RX-J0520.5-6932	LMC503.11	88150	5:20:29.86	-69:31:55.7
93	RX_J0513.9-6951	LMC503.07	319	5:13:50.77	-69:51:47.5
94	ERASST_J100130.9-614021	GD2060.13	17175	10:01:30.88	-61:40:21.0

Bibliography

Abdo, A. et al. (2010). 'Detection of the Small Magellanic Cloud in gamma-rays with Fermi/LAT'. *Astronomy & Astrophysics*, vol. 523:p. A46.

Abell, P. A. et al. (2009). 'LSST Science Book, Version 2.0'. URL `https://www.openaire.eu/search/publication?articleId=od_______165::90d6b6900c3d5b4748d45b42a3d5c2f8`.

Abt, H. A. (1983). 'Normal and abnormal binary frequencies'. *Annual review of astronomy and astrophysics*, vol. 21:pp. 343–372.

Adams, S. M., Kochanek, C., Beacom, J. F., Vagins, M. R., and Stanek, K. (2013). 'Observing the next galactic supernova'. *The Astrophysical Journal*, vol. 778(2):p. 164.

Agresti, A. (1994). 'Simple capture-recapture models permitting unequal catchability and variable sampling effort'. *Biometrics*:pp. 494–500.

Akaike, H. (1974a). 'Stochastic theory of minimal realization'. *IEEE Transactions on Automatic Control*, vol. 19(6):pp. 667–674.

Akaike, H. (1974b). 'A new look at the statistical model identification'. *IEEE Transactions on Automatic Control*, vol. 19(6):pp. 716–723.

Alfonso-Garzón, J. et al. (2017). 'Long-term optical and X-ray variability of the Be/X-ray binary H 1145-619: Discovery of an ongoing retrograde density wave'. *A&A*, vol. 607:p. A52. URL `https://doi.org/10.1051/0004-6361/201630211`.

Ambartsumyan, V., Mirzoyan, L., Parsamyan, E. S., Chavushyan, O., and Erastova, L. (1970). 'Flare stars in the Pleiades'. *Astrophysics*, vol. 6(1):pp. 1–10.

Ambartsumyan, V., Mirzoyan, A., Parsamyan, E., Chavushyan, O., and Erastova, L. (1971). 'Flare stars in the Pleiades. II'. *Astrophysics*, vol. 7(3):pp. 189–196.

Ambartsumyan, V. et al. (1972). 'Flare stars in the pleiades. III'. *Astrophysics*, vol. 8(4):pp. 287–299.

Ambartsumyan, V. et al. (1973). 'Flare stars in the Pleiades. IV'. *Astrophysics*, vol. 9(4):pp. 267–277.

Amorós, J. (2014). 'Recapturing Laplace'. *Significance*, vol. 11(3):pp. 38–39.

Amstrup, S. C., McDonald, T. L., and Manly, B. F. (Editors) (2005). *Handbook of capture-recapture analysis* (Princeton University Press).

Anderson, C. W. et al. (2013). 'Comparison of indirect and direct methods of distance sampling for estimating density of white-tailed deer'. *Wildlife Society Bulletin*, vol. 37(1):pp. 146–154.

Antoniou, V., Zezas, A., Hatzidimitriou, D., and Kalogera, V. (2010). 'Star formation history and X-ray binary populations: the case of the Small Magellanic Cloud'. *The Astrophysical Journal Letters*, vol. 716(2):p. L140.

Bailey, N. T. (1952). 'Improvements in the interpretation of recapture data'. *The Journal of Animal Ecology*:pp. 120–127.

Baillargeon, S. and Rivest, L.-P. (2007). 'Rcapture: Loglinear Models for Capture-Recapture in R'. *Journal of Statistical Software*, vol. 19(5). URL `http://econpapers.repec.org/article/jssjstsof/19_3ai05.htm`.

Bellm, E. (2014). 'The Zwicky Transient Facility'. In *The Third Hot-wiring the Transient Universe Workshop*. pp. 27–33.

Bellm, E. et al. (2019). 'The Zwicky Transient Facility: System Overview, Performance, and First Results'. *Publications of the Astronomical Society of the Pacific*, vol. 131(995):p. 18 002. URL `https://arxiv.org/abs/1902.01932`.

Bird, A. J. et al. (2009). 'Discovery of a 30-d period in the supergiant fast X-ray transient SAX J1818. 6–1703'. *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 393(1):pp. L11–L15.

Bloemen, S. et al. (2016). 'MeerLICHT and BlackGEM: custom-built telescopes to detect faint optical transients'. In *Ground-based and airborne telescopes VI*, vol. 9906 (International Society for Optics and Photonics), p. 990664.

Böhning, D., van der Heijden, P. G., and Bunge, J. (2017). *Capture-recapture methods for the social and medical sciences* (CRC Press).

Borchers, D. L., Buckland, S. T., Stephens, W., Zucchini, W. et al. (2002). *Estimating animal abundance: closed populations*, vol. 13 (Springer Science & Business Media).

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications* (CRC press).

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (Springer New York), 2nd ed.

Burnham, K. P. and Overton, W. S. (1978). 'Estimation of the size of a closed population when capture probabilities vary among animals'. *Biometrika*, vol. 65(3):pp. 625–633.

Carothers, A. (1973). 'The effects of unequal catchability on Jolly-Seber estimates'. *Biometrics*:pp. 79–100.

Cattadori, I. M., Haydon, D. T., Thirgood, S. J., and Hudson, P. J. (2003). 'Are indirect measures of abundance a useful index of population density? The case of red grouse harvesting'. *Oikos*, vol. 100(3):pp. 439–446.

Chao, A. (1987). 'Estimating the population size for capture-recapture data with unequal catchability'. *Biometrics*:pp. 783–791.

Chao, A. (1989). 'Estimating population size for sparse data in capture-recapture experiments'. *Biometrics*:pp. 427–438.

Chao, A. (2001). 'An overview of closed capture-recapture models'. *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 6(2):pp. 158–175.

Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T.-J. (2005). 'A new statistical approach for assessing similarity of species composition with incidence and abundance data'. *Ecology letters*, vol. 8(2):pp. 148–159.

Chao, A. and Huggins, R. (2005*a*). *Classical Closed-population Capture–Recapture Models*, chap. 2 (Princeton University Press, Princeton, New Jersey), pp. 22–35.

Chao, A. and Huggins, R. (2005*b*). *Modern Closed-population Capture–Recapture Models*, chap. 4 (Princeton University Press, Princeton, New Jersey), pp. 58 – 87.

Chao, A. and Lee, S.-M. (1992). 'Estimating the number of classes via sample coverage'. *Journal of the American statistical Association*, vol. 87(417):pp. 210–217.

Chao, A. and Tsay, P. (1998). 'A sample coverage approach to multiple-system estimation with application to census undercount'. *Journal of the American Statistical Association*, vol. 93(441):pp. 283–293.

Chao, A., Yip, P. S., Lee, S.-M., and Chu, W. (2001). 'Population size estimation based on estimating functions for closed capture–recapture models'. *Journal of Statistical Planning and Inference*, vol. 92(1-2):pp. 213–232.

Chapman, D. G. (1951). 'Some properties of hypergeometric distribution with application to zoological census'. *University of California Publications Statistics*, vol. 1:pp. 131–160.

Chapman, D. G. (1952). 'Inverse, multiple and sequential sample censuses'. *Biometrics*, vol. 8(4):pp. 286–306.

Charles, P. A. and Coe, M. J. (2006). *Optical, ultraviolet and infrared observations of X-ray binaries*. Cambridge Astrophysics (Cambridge University Press), pp. 215–266.

Connor, L., Pen, U.-L., and Oppermann, N. (2016). 'FRB repetition and non-Poissonian statistics'. *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 458(1):pp. L89–L93.

Cooch, E. and White, G. C. (Editors) (2008). *Program MARK," A gentle introduction"* (phidot). URL `http://www.phidot.org/software/mark/docs/book/`.

Cook, R. J. and Lawless, J. (2007). *The statistical analysis of recurrent events* (Springer Science & Business Media).

Coppejans, D. L. et al. (2016). 'Statistical properties of dwarf novae-type cataclysmic variables: the outburst catalogue'. *MNRAS*, vol. 456(4):pp. 4441–4454. `1512.03821`.

Corbet, R. (1984). 'Be/neutron star binaries-A relationship between orbital period and neutron star spin period'. *A&A*, vol. 141:pp. 91–93.

Corbet, R. (1999). 'Monitoring and Discovering X-Ray Pulsars in the SMC'. *RXTE Proposal*:p. 40 089.

Cormack, R. (1964). 'Estimates of survival from the sighting of marked animals'. *Biometrika*, vol. 51(3/4):pp. 429–438.

Cormack, R. (1968*a*). 'Statistics in Biology: Statistical Methods for Research in the Natural Sciences, Volume I'.

Cormack, R. (1968*b*). 'The statistics of capture-recapture methods'. *Oceanogr. Mar. Biol. Ann. Rev*, vol. 6:pp. 455–506.

Coull, B. A. and Agresti, A. (1999). 'The use of mixed logit models to reflect heterogeneity in capture-recapture studies'. *Biometrics*, vol. 55(1):pp. 294–301.

Crowl, H. H. et al. (2001). 'The line-of-sight depth of populous clusters in the Small Magellanic Cloud'. *The Astronomical Journal*, vol. 122(1):p. 220.

Dahl, K. (1917). 'Studier og forsøk over ørret og ørretvand'. *Kristiana: Centraltrykkeriet.*, (15).

Darroch, J. N. (1958). 'The multiple-recapture census: I. Estimation of a closed population'. *Biometrika*, vol. 45(3/4):pp. 343–359.

Darroch, J. N., Fienberg, S. E., Glonek, G. F., and Junker, B. W. (1993*a*). 'A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability'. *Journal of the American Statistical Association*, vol. 88(423):pp. 1137–1148.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (1993*b*). 'A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability'. *Journal of the American Statistical Association*, vol. 88(423):pp. 1137–1148. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476387`.

de Kool, M. (1992). 'Statistics of cataclysmic variable formation'. *A&A*, vol. 261(1):pp. 188–202.

Drake, A. et al. (2009). 'First results from the catalina real-time transient survey'. *The Astrophysical Journal*, vol. 696(1):p. 870.

Drake, A. et al. (2014). 'Cataclysmic variables from the catalina real-time transient survey'. *Monthly Notices of the Royal Astronomical Society*, vol. 441(2):pp. 1186–1200.

Efron, B. and Petrosian, V. (1992). 'A simple test of independence for truncated data with applications to redshift surveys'. *The Astrophysical Journal*, vol. 399:pp. 345–352.

Eggleton, P. (2006). *Evolutionary processes in binary and multiple stars*, vol. 40 (Cambridge University Press).

Eggleton, P. and Pringle, J. (Editors) (1985). *Interacting Binaries*, vol. 150 of *Series C* (NATO Advanced Science Institutes (ASI), Cambridge, UK). Held from 31 July-13 August 1983.

Evans, M. A. and Bonett, D. G. (1994). 'Bias reduction for multiple-recapture estimators of closed population size'. *Biometrics*:pp. 388–395.

Feigelson, E. D. (2016). 'The changing landscape of astrostatistics and astroinformatics'. *Proceedings of the International Astronomical Union*, vol. 12(S325):pp. 3–9.

Feigelson, E. D. and Babu, J. (2006). *Statistical challenges in astronomy* (Springer Science & Business Media).

Fienberg, S. E. (1992*a*). 'Bibliography on capture-recapture modelling with application to census undercount adjustment'. *Survey Methodology*, vol. 18(1):pp. 143–154.

Fienberg, S. E. (1992*b*). 'A brief history of statistics in three and one-half chapters: A review essay'. *Statistical Science*, vol. 7(2):pp. 208–225.

Firth, D. (1992). 'Bias reduction, the Jeffreys prior and GLIM'. In *Advances in GLIM and Statistical Modelling* (Springer), pp. 91–100.

Firth, D. (1993). 'Bias reduction of maximum likelihood estimates'. *Biometrika*:pp. 27–38.

Fisher, R. A. (1922). 'On the mathematical foundations of theoretical statistics'. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222(594-604):pp. 309–368.

Fletcher, L. et al. (2011). 'An observational overview of solar flares'. *Space science reviews*, vol. 159(1-4):p. 19.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2013). 'emcee: the MCMC hammer'. *Publications of the Astronomical Society of the Pacific*, vol. 125(925):p. 306.

Foreman-Mackey, D. et al. (2019). 'emcee v3: A Python ensemble sampling toolkit for affine-invariant MCMC'. *arXiv preprint arXiv:1911.07688*.

Galache, J. et al. (2004). 'Long-Term Behaviour of High Mass X-Ray Binaries in the SMC'. In *American Astronomical Society Meeting Abstracts*, vol. 205. pp. 102–11.

Gänsicke, B. T. et al. (2009). 'SDSS unveils a population of intrinsically faint cataclysmic variables at the minimum orbital period'. *MNRAS*, vol. 397(4):pp. 2170–2188. 0905.3476.

Glatt, K. et al. (2008b). 'Age determination of six intermediate-age Small Magellanic Cloud star clusters with HST/ACS'. *The Astronomical Journal*, vol. 136(4):p. 1703.

Gopalan, G., Vrtilek, S. D., and Bornn, L. (2015). 'Classifying X-ray Binaries: A Probabilistic Approach'. *The Astrophysical Journal*, vol. 809(1):p. 40.

Grace, Y. Y. (2016). *Statistical analysis with measurement error or misclassification* (Springer).

Graczyk, D. et al. (2014). 'The Araucaria Project. The Distance to the Small Magellanic Cloud from Late-type Eclipsing Binaries'. *ApJ*, vol. 780(1):59. 1311.2340.

Graham, M. J. et al. (2019). 'The Zwicky transient facility: science objectives'. *Publications of the Astronomical Society of the Pacific*, vol. 131(1001):p. 078 001.

Green, R. F., Ferguson, D. H., Liebert, J., and Schmidt, M. (1982). 'Cataclysmic variable candidates from the Palomar Green Survey'. *PASP*, vol. 94:pp. 560–564.

Haberl, F. and Sturm, R. (2016). 'High-mass X-ray binaries in the Small Magellanic Cloud'. *Astronomy & Astrophysics*, vol. 586:p. A81.

Hameury, J. M. (2020). 'A review of the disc instability model for dwarf novae, soft X-ray transients and related objects'. *Advances in Space Research*, vol. 66(5):pp. 1004–1024. 1910.01852.

Hellier, C. (2001). *Cataclysmic Variable Stars-how and why they vary* (Springer Science & Business Media).

Henderson, C. B. et al. (2016). 'Campaign 9 of the K2 mission: observational parameters, scientific drivers, and community involvement for a simultaneous space-and ground-based microlensing survey'. *Publications of the Astronomical Society of the Pacific*, vol. 128(970):p. 124 401.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, vol. 398 of *Wiley Series in Probability and Statistics* (John Wiley & Sons, Inc., Hoboken, New Jersey.), 3rd ed.

Howell, S. B. et al. (2014). 'The K2 mission: characterization and early results'. *Publications of the Astronomical Society of the Pacific*, vol. 126(938):p. 398.

Huggins, R. (1989). 'On the statistical analysis of capture experiments'. *Biometrika*, vol. 76(1):pp. 133–140.

Huggins, R. (1991). 'Some practical aspects of a conditional likelihood approach to capture experiments'. *Biometrics*:pp. 725–732.

IBM (2020). 'IBM SPSS Statistics'. URL `https://www.ibm.com/za-en/products/spss-statistics`, accessed 22-10-2020.

Ivezić, Ž. et al. (2019). 'LSST: from science drivers to reference design and anticipated data products'. *The Astrophysical Journal*, vol. 873(2):p. 111.

Jolly, G. M. (1965). 'Explicit estimates from capture-recapture data with both death and immigration-stochastic model'. *Biometrika*, vol. 52(1/2):pp. 225–247.

Judd, C. M., McClelland, G. H., and Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond* (Routledge).

Kaiser, N. et al. (2002). 'Pan-STARRS: a large synoptic survey telescope array'. In *Survey and Other Telescope Technologies and Discoveries*, vol. 4836 (International Society for Optics and Photonics), pp. 154–164.

Keller, S. C. et al. (2007). 'The SkyMapper telescope and the southern sky survey'. *Publications of the Astronomical Society of Australia*, vol. 24(1):pp. 1–12.

Kelly, B. C. (2013). 'Measurement Error Models in Astronomy'. In Feigelson, E. D. and Babu, G. J. (Editors), *Statistical Challenges in Modern Astronomy V*.

Kendall, W. L. (2001). 'The robust design for capture-recapture studies: analysis using program MARK'.

Kendall, W. L. (2012). *Program MARK: a gentle intoduction*, chap. 15: The 'robust design' (MARK), 19th ed., pp. 1–48.

Kennea, J., Coe, M., Evans, P., Waters, J., and Jasko, R. (2018). 'The First Year of S-CUBED: The Swift Small Magellanic Cloud Survey'. *The Astrophysical Journal*, vol. 868(1):p. 47.

Krebs, C. J. (1999). *Ecological methodology*. 574.5072 K7.

Kretschmar, P. et al. (2019). 'Advances in Understanding High-Mass X-ray Binaries with INTEGRALand Future Directions'. *New Astronomy Reviews*, vol. 86:p. 101 546.

Krimm, H. A. et al. (2013). 'The swift/bat hard x-ray transient monitor'. *The Astrophysical Journal Supplement Series*, vol. 209(1):p. 14.

Laake, J. L. (2013). 'RMark: an R interface for analysis of capture-recapture data with MARK'.

Laplace, P.-S. (1783). 'Sur les naissances, les mariages et les morts'. *Histaire de I'Academie Royale des Sciences*:p. 693.

Lavallée, P. and Rivest, L.-P. (2012). 'Capture-recapture sampling and indirect sampling'. *Journal of Official Statistics*, vol. 28(1):p. 1.

Law, N. M. et al. (2009). 'The Palomar Transient Factory: System Overview, Performance, and First Results'. *Publications of the Astronomical Society of the Pacific*, vol. 121(886):pp. 1395–1408. URL https://www.jstor.org/stable/10.1086/648598.

Laycock, S. G. T. (2017). 'From blackbirds to black holes: Investigating capture-recapture methods for time domain astronomy'. *New A*, vol. 54:pp. 91–102. URL http://dx.doi.org/10.1016/j.newast.2017.01.003.

Laycock, S. G. T. et al. (2003). 'X-Ray and optical observations of XTE J0052—723: a transient Be/X-ray pulsar in the Small Magellanic Cloud'. *Monthly Notices of the Royal Astronomical Society*, vol. 339(2):pp. 435–441.

Lee, S.-M. and Chao, A. (1994). 'Estimating population size via sample coverage for closed capture-recapture models'. *Biometrics*:pp. 88–97.

Lincoln, F. C. (1930). *Calculating waterfowl abundance on the basis of banding returns.* 118 (US Department of Agriculture).

Liu, Q. Z., van Paradijs, J., and Van den Heuvel, E. (2005). 'High-mass X-ray binaries in the Magellanic Clouds'. *A&A*, vol. 442(3):pp. 1135–1138. URL https://heasarc.gsfc.nasa.gov/W3Browse/all/maghmxbcat.html.

Liu, Q., van Paradijs, J., and Van den Heuvel, E. (2006). 'Catalogue of high-mass X-ray binaries in the Galaxy'. *Astronomy & Astrophysics*, vol. 455(3):pp. 1165–1168.

Liu, Q., van Paradijs, J., and Van den Heuvel, E. (2007). 'A catalogue of low-mass X-ray binaries in the Galaxy, LMC, and SMC'. *Astronomy & Astrophysics*, vol. 469(2):pp. 807–810.

Lloyd, C. J. and Yip, P. (1991). 'A unification of inference from capture-recapture studies through martingale estimating functions'. *Estimating Equations*:pp. 65–88.

Loredo, T. J. (1992). *Promise of Bayesian Inference for Astrophysics.* Statistical Challenges in Modern Astronomy (New York: Springer), pp. 275–297.

Loredo, T. J. (2007). 'Analyzing data from astronomical surveys: Issues and directions'. In *Statistical Challenges in Modern Astronomy IV*, vol. 371. p. 121.

Martin, R. G., Nixon, C., Armitage, P. J., Lubow, S. H., and Price, D. J. (2014). 'Giant Outbursts in Be/X-Ray Binaries'. *The Astrophysical Journal Letters*, vol. 790(2):p. L34.

McBride, V. A., Coe, M. J., Negueruela, I., Schurch, M. P. E., and McGowan, K. E. (2008). 'Spectral distribution of Be/X-ray binaries in the Small Magellanic Cloud'. *MNRAS*, vol. 388(3):pp. 1198–1204. 0805.0008.

McClintock, B. T. et al. (2014). 'Mark-resight abundance estimation under incomplete identification of marked individuals'. *Methods in Ecology and Evolution*, vol. 5(12):pp. 1294–1304.

Merloni, A. et al. (2012). 'eROSITA science book: mapping the structure of the energetic universe'. *arXiv preprint arXiv:1209.3114*.

Mirzoyan, L. V. et al. (1977). 'Flare stars in the pleiades. V'. *Astrophysics*, vol. 13(2):pp. 105–116.

Moran, P. A. P. (1951). 'A mathematical theory of animal trapping'. *Biometrika*, vol. 38(3/4):pp. 307–311.

Mróz, P. et al. (2013). 'Dwarf Novae in the OGLE Data. II. Forty New Dwarf Novae in the OGLE-III Galactic Disk Fields'. *arXiv preprint arXiv:1307.1238*.

Mróz, P. et al. (2015). 'One Thousand New Dwarf Novae from the OGLE Survey'. *Acta Astronomica*, vol. 65(4):pp. 313–328. 1601.02617.

Mroz, P. et al. (2016). 'VizieR Online Data Catalog: One thousand new dwarf novae from the OGLE Survey (Mroz+, 2015)'. *VizieR Online Data Catalog*:J/AcA/65/313.

Negueruela, I. and Schurch, M. (2007). 'A search for counterparts to massive X-ray binaries using photometric catalogues'. *Astronomy & Astrophysics*, vol. 461(2):pp. 631–639.

Oppermann, N., Yu, H.-R., and Pen, U.-L. (2018). 'On the non-Poissonian repetition pattern of FRB121102'. *Monthly Notices of the Royal Astronomical Society*, vol. 475(4):pp. 5109–5115.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). 'Statistical Inference from Capture Data on Closed Animal Populations'. *Wildlife Monographs*, vol. 62:pp. 3–135.

Patterson, J. (1984). 'The evolution of cataclysmic and low-mass X-ray binaries.' *ApJS*, vol. 54:pp. 443–493.

Patterson, M. T. et al. (2018). 'The Zwicky Transient Facility Alert Distribution System'. *Publications of the Astronomical Society of the Pacific*, vol. 131(995):p. 018 001.

Petersen, C. G. J. (1894). *On the biology of our flat-fishes and on the decrease of our flat-fish fisheries.*, vol. 4 (Report of the Danish Biological Station).

Pike, S. N. et al. (2019). 'Observing the transient pulsations of SMC X-1 with NuSTAR'. *The Astrophysical Journal*, vol. 875(2):p. 144.

Politano, M. (1996). 'Theoretical Statistics of Zero-Age Cataclysmic Variables'. *ApJ*, vol. 465:p. 338.

Pollock, K. H. (1981). *A capture-recapture sampling design robust to unequal catchability.* Ph.D. thesis, North Carolina State University.

Pollock, K. H. (1982). 'A Capture-Recapture Design Robust to Unequal Probability of Capture'. *The Journal of Wildlife Management*, vol. 46(3):pp. 752–757. URL https://www.jstor.org/stable/3808568.

Pollock, K. H. and Otto, M. C. (1983). 'Robust estimation of population size in closed animal populations from capture-recapture experiments'. *Biometrics*:pp. 1035–1049.

Prestwich, A. et al. (2003). 'Classifying X-ray Sources in External Galaxies from X-ray Colors'. *The Astrophysical Journal*, vol. 595(2):p. 719.

Pretorius, M. L., Knigge, C., and Kolb, U. (2007). 'Understanding selection effects in observed samples of cataclysmic variables'. In di Salvo, T. et al. (Editors), *The Multicolored Landscape of Compact Objects and Their Explosive Origins*, vol. 924 of *American Institute of Physics Conference Series*. pp. 546–554. `astro-ph/0610648`.

Price-Whelan, A. M. et al. (2018). 'The Astropy project: Building an open-science project and status of the v2. 0 core package'. *The Astronomical Journal*, vol. 156(3):p. 123.

R Development Core Team, 2014 (2010). 'R: a language and environment for statistical computing'.

Ramsay, G., Schreiber, M. R., Gänsicke, B. T., and Wheatley, P. J. (2017). 'Distances of cataclysmic variables and related objects derived from Gaia Data Release 1'. *A&A*, vol. 604:A107. `1704.00496`.

Rau, A. et al. (2007). 'The Incidence of Dwarf Novae in Large Area Transient Searches'. *ApJ*, vol. 664(1):pp. 474–480. `astro-ph/0611933`.

Rivest, L.-P. and Baillargeon, S. (2007). 'Applications and Extensions of Chao's Moment Estimator for the Size of a Closed Population'. *Biometrics*, vol. 63(4):pp. 999–1006. URL `http://www.ingentaconnect.com/content/bpl/biom/2007/00000063/00000004/art00002`.

Rivest, L.-P. and Baillargeon, S. (2019). *Package 'Rcapture'*. URL `https://cran.r-project.org/package=Rcapture`.

Rivest, L.-P. and Daigle, G. (2004). 'Loglinear models for the robust design in mark–recapture experiments'. *Biometrics*, vol. 60(1):pp. 100–107.

Rivest, L.-P. and Lévesque, T. (2001). 'Improved log-linear model estimators of abundance in capture-recapture experiments'. *Canadian Journal of Statistics*, vol. 29(4):pp. 555–572.

Roback, P. and Legler, J. (2021). *Beyond Multiple Linear Regression: Applied Generalized Linear Models And Multilevel Models in R* (CRC Press). URL `https://bookdown.org/roback/bookdown-BeyondMLR/`.

Robitaille, T. P. et al. (2013). 'Astropy: A community Python package for astronomy'. *Astronomy & Astrophysics*, vol. 558:p. A33.

Romine, G., Feigelson, E. D., Getman, K. V., Kuhn, M. A., and Povich, M. S. (2016). 'YOUNG STELLAR POPULATIONS IN MYStIX STAR-FORMING REGIONS: CANDIDATE PROTOSTARS'. *The Astrophysical Journal*, vol. 833(2):p. 193. URL `https://doi.org/10.3847%2F1538-4357%2F833%2F2%2F193`.

Roussas, G. G. (2003). *An introduction to probability and statistical inference* (Elsevier).

Royle, J. A., Chandler, R. B., Sollmann, R., and Gardner, B. (2013). *Spatial capture-recapture* (Academic Press).

Salvatier, J., Wieckiâ, T. V., and Fonnesbeck, C. (2016). 'PyMC3: Python probabilistic programming framework'. *Astrophysics Source Code Library*:pp. ascl–1610.

Sandland, R. L. and Cormack, R. M. (1984). 'Statistical Inference for Poisson and Multinomial Models for Capture- Recapture Experiments'. *Biometrika*, vol. 71(1):pp. 27–33. URL `https://www.jstor.org/stable/2336393`.

SAS Institute Inc. (2020). 'SAS Analytics Software & Solutions'. URL `https://www.sas.com/`, accessed 22-10-2020.

Schnabel, Z. E. (1938). 'The Estimation of Total Fish Population of a Lake'. *The American mathematical monthly*, vol. 45(6):pp. 348–352. URL `https://www.jstor.org/stable/2304025`.

Schumacher, F. X. and Eschmeyer, R. W. (1943). 'The estimation of fish populations in lakes and ponds'. *Journal of the Tennessee Academy of Science*, vol. 18(228).

Schwarz, C. J. and Seber, G. A. (1999). 'A review of estimating animal abundance III'. *Statistical Science*, vol. 14(4):pp. 427–456.

Schwarz, G. et al. (1978). 'Estimating the dimension of a model'. *The annals of statistics*, vol. 6(2):pp. 461–464.

Seber, G. A. (1965). 'A note on the multiple-recapture census'. *Biometrika*, vol. 52(1/2):pp. 249–259.

Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters* (Charles Griffin), 2nd ed. URL `https://natlib-primo.hosted.exlibrisgroup.com/primo-explore/search?query=any,contains,9912810353502836&tab=catalogue&search_scope=NLNZ&vid=NLNZ&offset=0`.

Seber, G. A. F. (1986). 'A Review of Estimating Animal Abundance'. *Biometrics*, vol. 42(2):pp. 267–292. URL `https://www.jstor.org/stable/2531049`.

Seber, G. A. (1992). 'A review of estimating animal abundance II'. *International Statistical Review/Revue Internationale de Statistique*:pp. 129–166.

Seber, G. A. (2001). 'Some new directions in estimating animal population parameters'. *Journal of agricultural, biological, and environmental statistics*:pp. 140–151.

Seber, G. A. and Schofield, M. R. (2019). *Capture-recapture: Parameter Estimation for Open Animal Populations* (Springer).

Sekar, C. C. and Deming, W. E. (1949). 'On a method of estimating birth and death rates and the extent of registration'. *Journal of the American Statistical Association*, vol. 44(245):pp. 101–115.

Shapiro, J. (1949). 'Ecological and life history notes on the porcupine in the Adirondacks'. *Journal of Mammalogy*, vol. 30(3):pp. 247–257.

Sharma, S. (2017). 'Markov Chain Monte Carlo methods for Bayesian data analysis in astronomy'. *ARA&A*, vol. 55:pp. 213–259.

Smith, R. J., Ashton, G., Vajpeyi, A., and Talbot, C. (2020). 'Massively parallel Bayesian inference for transient gravitational-wave astronomy'. *Monthly Notices of the Royal Astronomical Society*, vol. 498(3):pp. 4492–4502.

Smith, E. P. and van Belle, G. (1984). 'Nonparametric estimation of species richness'. *Biometrics*:pp. 119–129.

Soria, R., Cropper, M., and Motch, C. (2005). 'Classifying the zoo of ultraluminous X-ray sources'. *Chinese Journal of Astronomy and Astrophysics*, vol. 5(S1):p. 153.

StataCorp LLC (2020). 'Stata software'. URL `https://www.stata.com/`, accessed 22-10-2020.

Stella, L., White, N., and Rosner, R. (1986). 'Intermittent stellar wind accretion and the long-term activity of Population I binary systems containing an X-ray pulsar'. *ApJ*, vol. 308:pp. 669–679.

Stephens, P., Zaumyslova, O. Y., Miquelle, D., Myslenkov, A., and Hayward, G. (2006). 'Estimating population density from indirect sign: track counts and the Formozov–Malyshev–Pereleshin formula'. *Animal Conservation*, vol. 9(3):pp. 339–348.

Sturm, R. et al. (2011). 'The XMM-Newton survey of the Small Magellanic Cloud: discovery of the 11.866 s Be/X-ray binary pulsar XMMU J004814. 0-732204 (SXP11. 87)'. *A&A*, vol. 527:p. A131.

Sukhbold, T., Ertl, T., Woosley, S., Brown, J. M., and Janka, H.-T. (2016). 'Core-collapse supernovae from 9 to 120 solar masses based on neutrino-powered explosions'. *The Astrophysical Journal*, vol. 821(1):p. 38.

Tanaka, R. (1951). 'Estimation of vole and mouse populations on Mt. Ishizuchi and on the uplands of southern Shikoku'. *Journal of Mammalogy*, vol. 32(4):pp. 450–458.

Townsend, L., Coe, M., Corbet, R., and Hill, A. (2011). 'On the orbital parameters of Be/X-ray binaries in the Small Magellanic Cloud'. *Monthly Notices of the Royal Astronomical Society*, vol. 416(2):pp. 1556–1565.

Truscott, F. W. and Emory, F. W. (Editors) (1902). *Pierre-Simon Laplace: A Philosophical Essay on Probabilities* (Chapman & Hall, Limited, London.), translated from sixth french ed.

Udalski, A. (2008). 'XROM and RCOM: Two New OGLE-III Real Time Data Analysis Systems'. *Acta Astronomica*, vol. 58:pp. 187–192. 0810.2244.

Udalski, A., Kubiak, M., and Szymanski, M. (1997). 'Optical Gravitational Lensing Experiment. OGLE-2–the Second Phase of the OGLE Project'. *arXiv preprint astro-ph/9710091*.

Udalski, A., Szymański, M., Kaluzny, J., Kubiak, M., and Mateo, M. (1992). 'The Optical Gravitational Lensing Experiment'. *Acta Astronomica*, vol. 42:pp. 253–284.

Udalski, A., Szymanski, M., Soszynski, I., and Poleski, R. (2008). 'The optical gravitational lensing experiment. final reductions of the OGLE-III Data'. *arXiv preprint arXiv:0807.3884*.

Udalski, A., Szymański, M. K., and Szymański, G. (2015). 'OGLE-IV: Fourth Phase of the Optical Gravitational Lensing Experiment'. *Acta Astronomica*, vol. 65(1):pp. 1–38. 1504.05966.

Udalski, A. et al. (1993). 'The optical gravitational lensing experiment. Discovery of the first candidate microlensing event in the direction of the Galactic Bulge'. *Acta astronomica*, vol. 43:pp. 289–294.

Udalski, A. et al. (1994). 'The Optical gravitational lensing experiment. The Early warning system'. *arXiv preprint astro-ph/9408026*.

Warner, B. (1995). *Cataclysmic variable stars*, vol. 28.

White, G. C., Anderson, D. R., Burnham, K. P., and Otis, D. L. (1982). *Capture-recapture and removal methods for sampling closed populations* (Los Alamos National Laboratory).

White, N., Nagase, F., and Parmar, A. (1995). *The properties of X-ray binaries*, chap. 1. Cambridge Astrophysics Series (Cambridge University Press), pp. 1–57.

Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002). *Analysis and management of animal populations* (Academic press).

Zezas, A. and Antoniou, V. (2017). 'A deep survey of the X-ray binary populations in the SMC'. In Ness, J.-U. and Migliari, S. (Editors), *The X-ray Universe 2017*. p. 244.

Zippin, C. (1956). 'An evaluation of the removal method of estimating animal populations'. *Biometrics*, vol. 12(2):pp. 163–189.

Zippin, C. (1958). 'The removal method of population estimation'. *The Journal of Wildlife Management*, vol. 22(1):pp. 82–90.

Zombeck, M. V. (1990). *Handbook of Space Astronomy and Astrophysics* (Cambridge University Press), 2nd ed. URL `http://ads.harvard.edu/books/hsaa/toc.html`.