Computational Statistics and Data Analysis 96 (2016) 74–86

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Estimation of survival and capture probabilities in open population capture–recapture models when covariates are subject to measurement error

Jakub Stoklosa[a,*], Peter Dann[b], Richard M. Huggins[c], Wen-Han Hwang[d,*]

[a] *School of Mathematics and Statistics and Evolution & Ecology Research Centre, The University of New South Wales, New South Wales 2052, Australia*

[b] *Research Department, Phillip Island Nature Parks, Victoria 3922, Australia*

[c] *Department of Mathematics and Statistics, The University of Melbourne, Victoria 3010, Australia*

[d] *Institute of Statistics and Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan*

A R T I C L E   I N F O

Article history:
Received 17 October 2014
Received in revised form 18 October 2015
Accepted 22 October 2015
Available online 30 October 2015

Keywords:
Cormack–Jolly–Seber model
Error-in-variables
Little Penguins
Mark–capture–recapture
Regression calibration
SIMEX

A B S T R A C T

Predictor variables (or covariates) are frequently used in a capture–recapture analysis when estimating demographic quantities such as population size or survival probabilities. If these predictor variables are measured with error and subsequently used in the analysis, then estimates of the model parameters may be biased. Several approaches have been proposed to account for error-in-variables in capture–recapture models, however these methods generally assume the population is closed; hence quantities of interest for open populations such as the survival probabilities do not appear in the likelihood. To account for measurement error in environmental time-varying covariates for open population capture–recapture data, the well-known Cormack–Jolly–Seber model and two statistical methods are considered: (1) simulation–extrapolation; and (2) regression calibration, as well as a new method which accounts for correlation (arising from measurement error) between the survival and capture probabilities. Several simulation studies are conducted to examine the method performances, and a case study is presented which uses capture–recapture data on the Little Penguin *Eudyptula minor* and sea-surface temperature data as an environmental covariate to model their survival and capture probabilities.

1. Introduction

Capture–recapture methods are often used to help estimate and understand important population demographics in many applied disciplines, such as: ecology, conservation biology, epidemiology, medical studies and the social sciences (McCrea and Morgan, 2014). If the population is assumed to be open—*i.e.*, births, deaths, emigration or immigration may occur during sampling, then quantities such as survival probabilities of individuals in the population are usually of interest. The *Cormack–Jolly–Seber (CJS)* model is a commonly used and well developed approach to the analysis of open population capture–recapture experiments (Lebreton et al., 1992; Amstrup et al., 2005; McCrea and Morgan, 2014). In its simplest form, the CJS model consists of two parameters: the apparent survival probability which we denote by S and the capture probability which we denote by P. These parameters may be simply modelled as functions of temporal effects, however they

* Corresponding authors.
E-mail addresses: j.stoklosa@unsw.edu.au (J. Stoklosa), wenhan@nchu.edu.tw (W.-H. Hwang).

http://dx.doi.org/10.1016/j.csda.2015.10.010

can also be modelled as functions of environmental or individual covariates, (such as age or gender) both parametrically or nonparametrically (Pollock, 2002). Typically, the primary interest is in estimating the survival probabilities whereas capture probabilities are treated as nuisance parameters. Importantly, the CJS model falls under a generalized linear model (GLM) framework; this allows for standard practical tools to be incorporated in the analysis, such as: model selection (*e.g.,* using AIC); hypothesis testing (*e.g.,* using a likelihood ratio); and flexible extensions in the linear predictor when considering continuous covariates (*e.g.,* using generalized additive models via penalized splines).

Our motivation is capture–recapture data collected on adult Little Penguins *Eudyptula minor* from a colony located on Phillip Island, Australia. Within this colony, the Little Penguin will nest in either natural or artificial burrows, and their numbers are expected to be highest on the island during breeding and moulting—which usually occurs between late spring and early autumn (Reilly and Cullen, 1981, 1983). However, a large proportion of their lifetime is spent at sea; they may spend a day to more than several weeks at sea at a time foraging for food (Sutherland and Dann, 2012). This has motivated several studies to examine the population size distribution, breeding success and survival as functions of sea-surface temperature (Cullen et al., 2009; Sidhu et al., 2012; Dann and Chambers, 2013; Huggins and Stoklosa, 2013). When modelling capture–recapture data using the CJS model, both apparent survival and capture probabilities can be modelled as functions of sea-surface temperature which are treated as covariates (*e.g.,* see Stoklosa and Huggins, 2012). Measurements of the sea-surface temperatures may of course be subject to uncertainty in the way they were measured; so using these contaminated covariates may influence the results in the capture–recapture analysis.

It is well-known that *measurement error* (or error-in-variables) in predictor variables causes biased estimates when the measurement error is ignored. There is extensive literature on dealing with measurement error in covariates (see Carroll et al., 2006); two common approaches that we consider are simulation–extrapolation (SIMEX, Cook and Stefanski, 1994) and regression calibration (Armstrong, 1985; Carroll and Stefanski, 1990; Hardin et al., 2003). For further details on both methods see the respective Chapters 4 and 5 of Carroll et al. (2006). In these methods, only the variance and the distribution of the measurement error are assumed known and are specified prior to the analysis. Notably, both SIMEX and regression calibration are applicable to general estimation methods such as GLMs (Carroll et al., 2006).

In the capture–recapture context, Gould et al. (1999) have proposed SIMEX to correct for the effect of measurement error in a catch–effort model, where the measurement error was associated with the catch and effort recordings. Measurement error may also arise from imprecise measurements collected on trait characteristics from individuals. For example, uncertainty or inaccuracy associated with a measuring device that is used for measuring body weights or head-to-tail lengths of captured individuals. Several approaches using individual covariates for closed populations have been proposed: both Hwang and Huang (2003) and Huggins and Hwang (2010) used a modified regression calibration approach; Xi et al. (2009) used the EM algorithm for the case of both missing data and measurement error, Hwang et al. (2007) used a conditional score approach, and Hwang and Huang (2007) used measurement error methods for continuous time capture–recapture models. In closed populations, since there is no birth, death, emigration or immigration, the survival and birth probabilities are one and zero, respectively.

For open populations, Oliver (2012) examined the effects of measurement error in covariates collected on individuals on survival rates using a refinement of regression calibration within a CJS framework, and Barker et al. (2002) used measurement error methods to quantify density dependence between survival and abundance. As in Oliver (2012), we relax the assumption of closure but consider the simultaneous estimation of survival and capture probabilities, and model these as functions of environmental time-varying covariates with measurement error. We also adopt the CJS framework but use both SIMEX and refined regression calibration, along with a new approach to account for possible correlation (caused by the measurement error) between survival and capture probabilities.

The structure of the paper is as follows: we review the CJS model for the error-free covariate case in Section 2, and then discuss how to account for measurement error using SIMEX and regression calibration followed by the new proposed approach in Section 3. In Section 4 we conduct several simulation studies to investigate the sensitivity and robustness of the estimators in the presence of measurement error, and in Section 5 we fit all models to the Little Penguin data. Some discussion is given in Section 6.

2. The Cormack–Jolly–Seber model

We first give the likelihood function for the CJS model when incorporating environmental covariates in the error-free case, further details are given in Lebreton et al. (1992) or McCrea and Morgan (2014). We use maximum likelihood to carry-out the estimation, although Bayesian methods (Gimenez et al., 2006) or the EM algorithm (Van Duesen, 2001) can also be used.

A capture–recapture experiment consists of animals/individuals being captured from some population at a fixed number of τ capture occasions. If an individual is captured for the first time, they are marked and their presence is noted. Individuals which have been previously marked from past occasions are also noted. The marks allow the identification of each captured individual. These observed data (commonly referred to as capture histories) usually consists of a sequence of zeros and ones for each individual captured at least once in the experiment—*i.e.,* a matrix consisting of captured (or not captured) indicators across each capture occasion. We assume that the population is open and that individuals do not lose their tags, and tags are recorded correctly.