2020

Statements about the pervasiveness of behavior require data about the pervasiveness of behavior

Craig P. Speelman
Edith Cowan University

Marek McGann

10.3389/fpsyg.2020.594675 Speelman, C. P., & McGann, M. (2020). Statements about the pervasiveness of behavior require data about the pervasiveness of behavior. *Frontiers in Psychology*, *11*, 3117. https://doi.org/10.3389/fpsyg.2020.594675

ICES Journal of Marine Science

ICES Journal of Marine Science (2020), doi:10.1093/icesjms/fsaa180

Imputation of missing data from time-lapse cameras used in recreational fishing surveys

Ebenezer Afrifa-Yamoah [1]*, Stephen M. Taylor [2], Aiden Fisher [1], and Ute Mueller [1]

[1]School of Science, Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Australia
[2]Department of Primary Industries and Regional Development (DPIRD), Western Australian Fisheries and Marine Research Laboratories, PO Box 20, North Beach, WA 6920, Australia

*Corresponding author: e-mail: tel: +61 424559378; e.afrifayamoah@ecu.edu.au.

Afrifa-Yamoah, E., Taylor, S. M., Fisher, A., and Mueller, U. Imputation of missing data from time-lapse cameras used in recreational fishing surveys. – ICES Journal of Marine Science, doi:10.1093/icesjms/fsaa180.

Received 16 June 2020; revised 21 August 2020; accepted 5 September 2020.

While remote camera surveys have the potential to improve the accuracy of recreational fishing estimates, missing data are common and require robust analytical techniques to impute. Time-lapse cameras are being used in Western Australia to monitor recreational boating activities, but outages have occurred. Generalized linear mixed effect models formulated in a fully conditional specification multiple imputation framework were used to reconstruct missing data, with climatic and some temporal classifications as covariates. Using a complete 12-month camera record of hourly counts of recreational powerboat retrievals, data were simulated based on ten observed camera outage patterns, with a missing proportion of between 0.06 and 0.61. Nine models were evaluated, including Poisson and negative binomial models, and their associated zero-inflated variants. The imputed values were cross-validated against actual observations using percent bias, mean absolute error, root mean square error, and skill score as performance measures. In 90% of the cases, 95% confidence intervals for the total imputed estimates from at least one of the models contained the total actual counts. With no systematic trends in performance among the models, zero-inflated Poisson and its bootstrapping variant models consistently ranked among the top 3 models and possessed the narrowest confidence intervals. The robustness and generality of the imputation framework were demonstrated using other camera datasets with distinct characteristics. The results provide reliable estimates of the number of boat retrievals for subsequent estimates of fishing effort and provide time series data on boat-based activity.

Keywords: count data imputation, fully conditional specification, generalized linear mixed effect models, powerboat retrievals, zero-inflated models

Introduction

As many recreational fisheries are of large spatial extent, diverse, and not well defined, it can be challenging and costly to obtain accurate recreational fishing information for sustainable management (Smallwood *et al.*, 2012; Hyder *et al.*, 2018). Remote camera surveys (also referred to as digital camera monitoring) are increasingly being used throughout Europe, North America, and Australasia to monitor recreational fishing effort in marine and freshwater fisheries (Smallwood *et al.*, 2012; van Poorten *et al.*, 2015; Hartill *et al.*, 2016, 2020; Lancaster *et al.*, 2017; Askey *et al.*, 2018). In comparison to onsite surveys (e.g. boat ramp surveys), remote cameras provide a cost-effective method of monitoring the movement of boats (Smallwood *et al.*, 2012; Hartill *et al.*, 2016) or fishers (van Poorten *et al.*, 2015; Askey *et al.*, 2018; Stahr and Knudsen, 2018), where results can then form a basis for subsequent calculations of fishing effort (Hartill *et al.*, 2020). The key strength of this approach is that it can provide wider coverage of boating and fishing effort, including night-time fishing

(Smallwood *et al.*, 2012; Taylor *et al.*, 2018), and activity occurring on the days when onsite survey staff are not present at the ramps. In addition, remote camera data can be used to test the accuracy of other recreational fishing survey methods that have more restricted sampling coverage (Lancaster *et al.*, 2017). The integration of remote camera observations as a complementary technique in recreational fishing surveys can also be used to improve the accuracy and precision of harvest estimates (Steffe *et al.*, 2017).

In theory, remote cameras can provide continuous recordings of boating and recreational fishing activities; however, interruptions of camera operations (herein referred to as "outages") can lead to significant gaps within the data. Camera network outages can occur as a result of technical faults, vandalism, theft, and/or weather conditions, such as temperature and humidity, lightning strikes, flooding, and other environmental factors (Blight and Smallwood, 2015; Hartill *et al.*, 2020). Responding to missing values has been a subject of interest for researchers in many fields, where missing values require proper handling to prevent further loss of precision and reliability of estimates and indices (Kleinke and Reinecke, 2013a; van Poorten *et al.*, 2015; Hartill *et al.*, 2016). In remote camera surveys, missing data can potentially lead to biased estimates. Other problems that could occur include irreproducibility of estimates, and loss of statistical power, with the magnitude determined by the nature and duration of the outage (van Poorten *et al.*, 2015; Hartill *et al.*, 2016). Therefore, it is important to build imputation schemes that are tailored to the pattern and nature of "missingness" and the distributional characteristics of remote camera data. However, despite the rapid emergence of remote camera studies relevant to recreational fishing, relatively few studies have examined analytical approaches for dealing with data outages.

Of the recreational fishing studies that used remote cameras and reported missing data, outages have typically been assumed to be random (Smallwood *et al.* 2012; Taylor *et al.*, 2018). Model-based approaches for imputing remote camera missing data have also been explored (van Poorten *et al.*, 2015; Hartill *et al.*, 2016; Lancaster *et al.*, 2017). A Bayesian hierarchical model was applied to predict total angling effort for 49 lakes in Canada based on remote camera data. Missing camera data were imputed from the average effort from proximate lakes (van Poorten *et al.*, 2015). Hartill *et al.* (2016) used generalized linear models (GLMs) to impute missing values in recreational boats returning to a boat ramp based on recorded remote camera data from neighbouring ramps. In both van Poorten *et al.* (2015) and Hartill *et al.* (2016), neighbouring ramps and viewpoints were used as reference points. However, in instances where outages occur simultaneously across nearby ramps or when ramps with installed remote cameras are widely dispersed with significantly different trends in boating activities, these imputation approaches cannot be applied. The recommendation to incorporate covariates such as climatic and environmental factors was made in both studies. Soykan *et al.* (2014) identified temperature, rainfall, tides, winds (direction, speed, and gust), and sea surface variables as significant predictors of fishing effort, to which boating effort correlates and serves as a good proxy (Johnson *et al.*, 2017). To the best of our knowledge, no study has evaluated the opportunities of using climatic variables to build imputation models to handle missing observations in remote camera data.

The current study sought to formulate and compare several imputation models with climatic and some temporal classifications, as explanatory variables to impute gaps of missing values in the counts of recreational powerboat retrievals from remote camera monitoring along the coastline of Western Australia (WA). Ten real camera outage patterns were applied to a "complete" remote camera monitoring dataset to artificially create missing gaps. Generalized linear mixed effect models built on the fully conditional specification multiple imputation framework were considered to reconstruct missing gaps (van Buuren, 2007; Kleinke and Reinecke, 2013b). Imputed estimates were compared with actual data recorded for the simulated periods of camera outage. The robustness and generality of the modelling scheme was illustrated on two other camera datasets (and covariates) from different locations, to establish the ability of the imputation scheme to handle both short and long outages.

Methods

Study area and camera data description

In WA, an estimated 26% of residents participate in recreational fishing at least once a year (Department of Primary Industries and Regional Development, 2019). Remote cameras have been used since 2006 to monitor trends in boating activities at 30 sites along the coast, including boat ramps, channel entrances and parts of foreshore (Hartill *et al.*, 2020). The type of vessel launched and retrieved is recorded as either commercial, powerboat, jet-ski, kayak or others. Subsequent analysis for this paper was restricted to powerboat retrievals, as this is the common vessel type used for boat-based recreational fishing activities in WA. Counts of the number of powerboat retrievals for each ramp were recorded to the nearest minute. A technical overview of the camera monitoring scheme can be found in Blight and Smallwood (2015).

This study utilized complete data on powerboat retrievals collected between 1 March 2011 and 29 February 2012 at the Leeuwin ramp and ten outage patterns observed at eight boat ramps distributed across the coastline of WA (Figure 1, see also Supplementary Table S1). Outage patterns identified coincided with the state-wide surveys of boat-based recreational fishing in WA (Ryan *et al.*, 2013, 2015, 2017, see Supplementary Table S1). The choice of the ten outage patterns was based on the percentage of missingness, ranging from 0.06 to 0.61 (see Figure 2). The longest outage imputed was 80 days (∼1920 h) and the shortest was 1 h. The ten outage patterns were of variable lengths and uncorrelated among the ramps. The complete record consisted of 8784 hourly entries of count of powerboats retrieved, and 54.4% of all records were zeros. In total, 12 293 powerboat retrievals were recorded.

A simulation scenario was chosen, where observed data of the complete records were turned into missing data based on the ten outage patterns (see Figure 2). This was done to enable cross-validation of the models and to establish the consistency of the models in imputing the various durations of outages. If data were missing for any portion of the hour, the observation for this hour was classified as missing to control for all possible interpretation errors that may have occurred during outages.

In addition, camera records for Mindarie in Ryan *et al.* (2013) and Monkey Mia in Ryan *et al.* (2017) were used. Missing data in these records were imputed for short-term camera outages applying the extrapolation method in Wise and Fletcher (2013). However, missing data for periods of extended outages were not imputed. The traffic intensities at these ramps are very different;

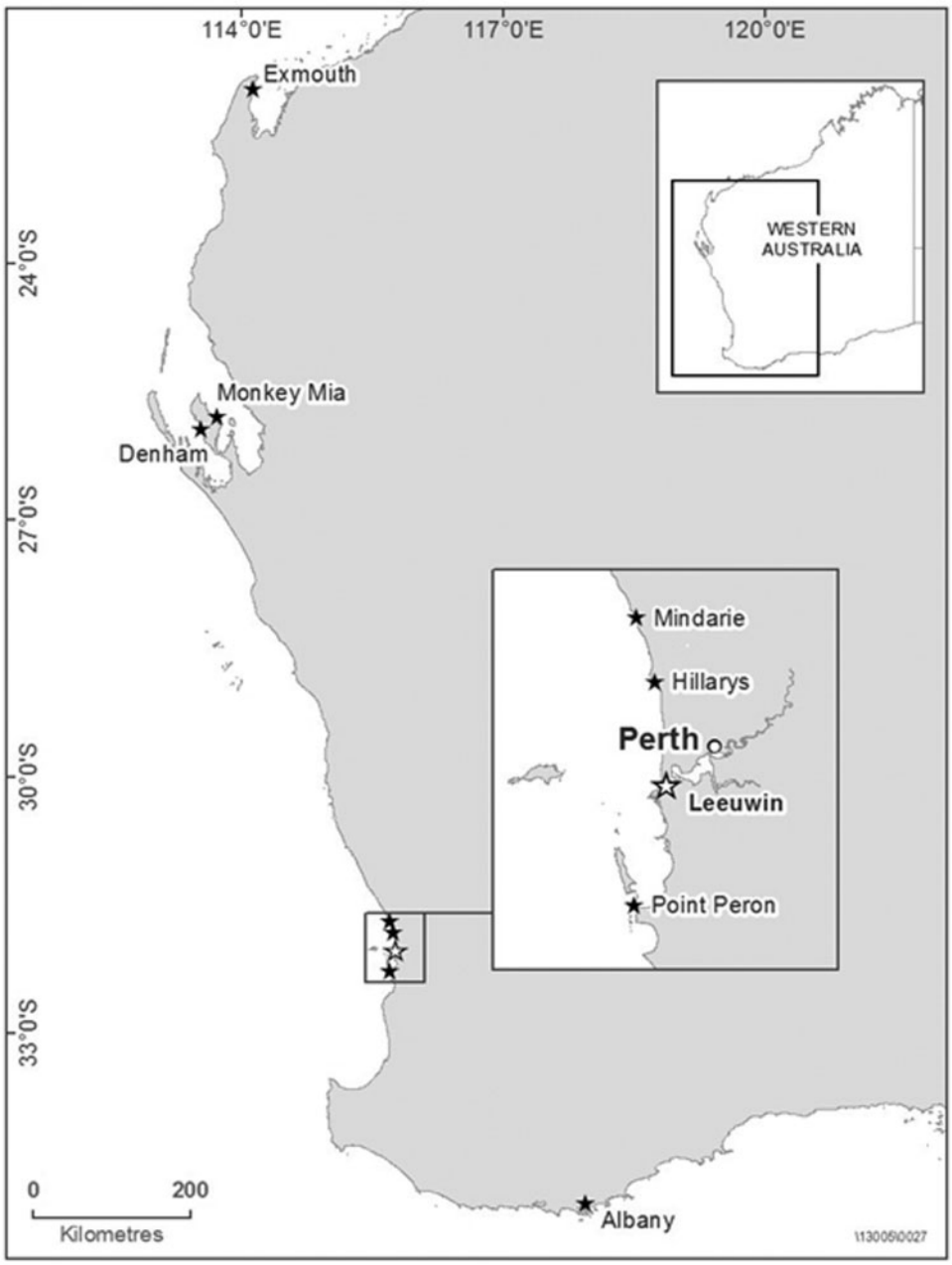

Figure 1. Map of Western Australia showing the remote camera locations from which information on the number of powerboat retrievals was examined in this study. The Leeuwin boat ramp is denoted with a larger white star because no outages occurred in the data from this camera in 2011/12. Real outages that occurred from the other remote cameras (denoted by smaller solid stars) were applied to the complete data set at Leeuwin to examine the various modelling approaches.

Mindarie is a moderately busy ramp with an annual total of ~20 000 powerboat retrievals while Monkey Mia is a less busy ramp with ~6000 powerboat retrievals (Ryan *et al.*, 2013, 2015, 2017). In terms of geographical location and climate type, Mindarie is in the West Coast bioregion with a hot-summer Mediterranean climate while Monkey Mia is in the Gascoyne Coast bioregion with a hot semi-arid climate (see Figure 1). Severe outages were observed at Mindarie as 60% of the fishing year data were missing, with no data for the months of September, April, and July. Similarly, 14% of data at Monkey Mia were missing. No data were available for the month of June.

Models and missing data assumptions

Let Y be the count of powerboat retrievals data observed from remotely operated camera, with some missing values, such that, $Y = (Y^{obs}, Y^{mis})$, where Y^{obs} represents the observed data and Y^{mis} missing data. Data were assumed to be missing at random. The imputation models were formulated to investigate the conditional distribution:

$$P\left(Y^{mis}, \theta | Y^{obs}, \boldsymbol{X}\right), \quad (1)$$

where θ represents the vector of unknown parameters of the

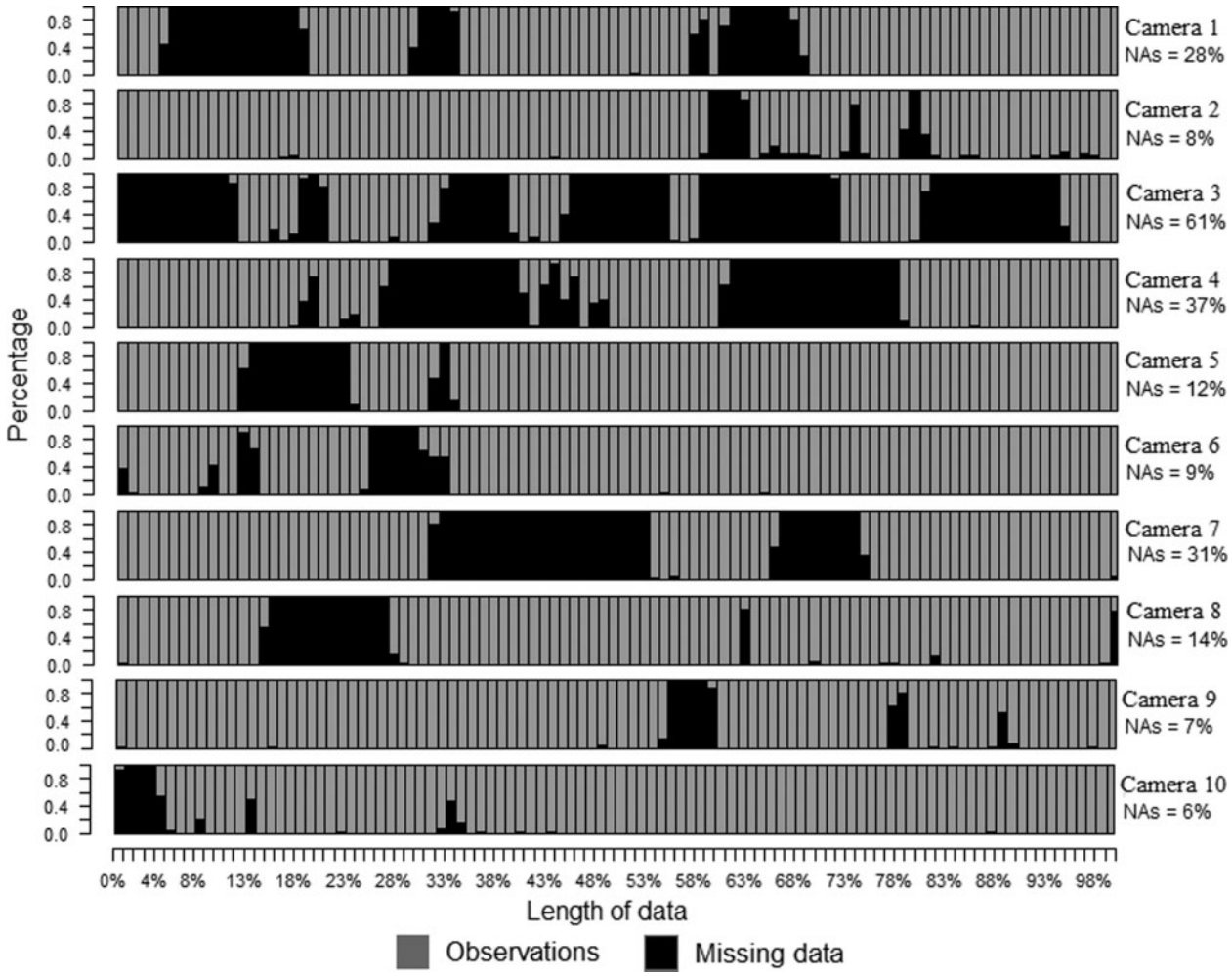

Figure 2. Distribution of the ten outage patterns applied to the Leeuwin dataset and their missing proportion. The horizontal axis represents the length of the camera data partitioned into 100. The vertical axis represents the proportion of missing data in the partitioned block or otherwise. The black bands represent the periods of camera outages, and the grey bands represent the observed data.

model**.** It was further assumed that the data generating process for Y can be derived from generalized linear mixed effect models (Afrifa-Yamoah *et al.*, 2019). Let y be a $n \times 1$ vector of observed outcomes:

$$y = X\beta + Z\upsilon + \varepsilon, \;\; \upsilon \;\sim\; N(0, \;\psi_\upsilon), \;\; \varepsilon \;\sim\; N(0, \Sigma), \qquad (2)$$

where X is a $n \times k$ matrix of fixed effects associated with the outcome y via β, which is a $k \times 1$ vector of coefficients, and Z is an $n \times r$ matrix of random effects associated with y via υ, which is an $r \times 1$ parameter vector. ψ_υ is the $r \times r$ variance–covariance matrix of the random effects and ε is the $n \times 1$ error vector with $\Sigma = \sigma^2 I$, where I is an $n \times n$ identity matrix. Climatic variables were treated as fixed effects, whereas the temporal classifications such as season, type, and time of day were treated as random effects (see Table 1 for variable description). Missing data in the climate data were imputed using the methods in Afrifa-Yamoah *et al.* (2020). It is important to note that these covariates were not directly associated with the missing mechanism and did not explicitly give any information on why the camera records were missing. Although it is common in scientific studies to focus on the relative importance of predictors within statistical models, in this study the focus was to predict boating effort based on the collective contribution of all covariates, irrespective of the statistical significance of their coefficients.

Based on the distributional characteristics of the data, quasi Poisson (denoted as QP), negative binomial (denoted by NB), zero-inflated Poisson (denoted by ZIP), and zero-inflated negative binomial (denoted by ZINB) models were considered. In the two-level models, it was assumed that the two-level processes (i.e. zero and non-zero parts) were influenced by the same set of covariates. In the modelling scheme, random intercept models were fitted, and common slopes were assumed without consideration for interaction effects. This was done to moderate the complexity of the model structure because of the number of predictors used. Predictive mean matching (PMM) as a general purpose method was also investigated. Little is known about the suitability of PMM for count data, because it was developed for imputing missing observations among continuous variables.

Fully conditional specification multiple imputation

The fully conditional specification multiple imputation framework (van Buuren and Groothuis-Oudshoorn, 2011) was used to

Table 1. Study variables and their attributes (NA indicates the number of missing records).

Variable	Type	Description	NA
Retrievals	Count	Hourly aggregated counts of powerboat retrievals.	Variable
Precipitation	Continuous	Average hourly amount of rainfall (mm)	46
Temperature	Continuous	Average hourly air temperature (°C)	4
Humidity	Continuous	Average hourly levels of humidity (%)	4
Wind speed	Continuous	Average hourly wind speed (km h^{-1})	4
Wind direction	Continuous	Average hourly wind direction (true degrees trigonometrically transformed)	4
Wind gust	Continuous	Maximum wind speed ($km(h^{-1})$)	69
Sea level pressure	Continuous	Average sea level pressure (hPa)	4
Day type	Categorical	Weekday or weekend/ public holiday	–
Season	Categorical	Summer (December–February), autumn (March–May), winter (June–August), and spring (September–November).	–
Time of day	Categorical	Dawn (01:00–04:59), early morning (05:00–07:59), morning (08:00–11:59), afternoon (12:00–15:59), late afternoon (16:00–18:59) and evening (19:00–00:59)	–

specify conditional models of the partially observed outcome variable given the covariates, to obtain a posterior predictive distribution defined as:

$$p(y^{mis}|y^{obs},X)=\int p(y^{mis}|y^{obs},X,\ \theta)p(\theta|y^{obs},X)d\theta, \tag{3}$$

where $\theta=(\beta,\psi_{\upsilon},\sigma)$ is the vector of parameters in (1) and $p(\theta|y^{obs},X)$ is the observed data posterior density of θ. From (3), estimate of the model parameters, $\widehat{\theta}=\left(\widehat{\beta},\ \widehat{\psi}_{\upsilon},\ \widehat{\sigma}\right)$, and their variance–covariance, $\widehat{S}_{\widehat{\theta}}$, were obtained.

For each missing observation, independent draws, $\theta^{*}=(\beta^{*},\ \psi_{\upsilon}^{*},\ \sigma^{*})$, were generated from $N(\widehat{\theta},\ \widehat{S}_{\widehat{\theta}})$ using a Gibbs sampler (see van Buuren and Groothuis-Oudshoorn, 2011; Kleinke and Reinecke, 2013a). From θ^{*}, we generated a chain of equations, $y^{*}=x\beta^{*}+z\upsilon^{*}+\varepsilon^{*}$, for the observed data and missing observation. A random draw was made from k Y^{obs} with y^{*} in the closest neighbourhood to that of the missing observation being imputed. This was done to introduce between-imputation variability. The procedure was repeated, generating M plausible complete datasets accounting for the uncertainty in the missing data (Rubin, 1987; Sterne *et al.*, 2009). More detail about the imputation scheme is presented in the Supplementary material.

Conversely, the normality assumption on $\widehat{\theta}$ may sometimes be implausible (Rubin, 1987). A variant implementation of the imputation scheme was carried out, where the draws of θ^{*} were estimated through bootstrapping (Efron, 1994). The parameter vector θ^{*} was estimated by fitting model (1) to a bootstrap sample consisting of $b\leq n-s$ observations (where s is the number of missing observations) drawn from (Y^{obs},X) (van Buuren and Groothuis-Oudshoorn, 2011; Kleinke and Reinecke, 2013b). This scheme resulted in imputations generated as before, from the following models: quasi Poisson (denoted as QP.boot), negative binomial (denoted as NB.boot), zero-inflated Poisson (ZIP.boot), and zero-inflated negative binomial (ZINB.boot). PMM was also implemented in the multiple imputation framework. The scheme used the ordinary multiple linear regression model to formulate the posterior distribution of θ. The imputation schemes replaced missing observations with observed values and thereby preserved the distribution of the observed data (Yu *et al.*, 2007).

The imputation schemes were performed using the *mice* (version 2.14, van Buuren and Groothuis-Oudshoorn, 2011) and *countimp* (version 1.0, Kleinke and Reinecke, 2013b) packages supported by *pscl* (version 1.5.2, Jackman, 2008; Zeileis *et al.*, 2008) and *glmmADMB* (version 0.8.2, Fournier *et al.*, 2012; Skaug *et al.*, 2015) in R (version 3.4.3, R Core Team 2016).

Pooling analysis

For each missing observation, let $\widehat{Y}_{i}^{mis}$ represent the ith imputed value and, then, the mean $\overline{\widehat{Y}_{i}^{mis}}$ and variance $\widehat{Var}\left(\overline{\widehat{Y}_{i}^{mis}}\right)$ of the pooled imputed estimates were obtained as:

$$\overline{\widehat{Y}_{i}^{mis}}=\frac{\sum_{m=1}^{M}\widehat{Y}_{i}^{mis}}{M}, \tag{4}$$

$$\widehat{Var}\left(\overline{\widehat{Y}_{i}^{mis}}\right)=\frac{\sum_{m=1}^{M}\widehat{Var}(\widehat{Y}_{i,m}^{mis})}{M}+\frac{(M+1)}{M(M-1)}\sum_{m=1}^{M}\left(\widehat{Y}_{i,m}^{mis}-\overline{\widehat{Y}_{i}^{mis}}\right)^{2}, \tag{5}$$

where M denotes the total number of estimates for Y^{mis} in the imputation scheme and $\sum_{m=1}^{M}\left(\widehat{Y}_{i,m}^{mis}-\overline{\widehat{Y}_{i}^{mis}}\right)^{2}$ reflects the missing values estimation uncertainties (Rubin, 1987).

Model performance evaluation

Missing values were imputed five times ($M=5$) as this number is considered to provide an appropriate balance of the bias-variance trade-off in the model evaluation (van Buuren and Groothuis-Oudshoorn, 2011; Allison, 2015). The estimation accuracy of the imputed values was assessed with the following performance indicators: percent bias, mean absolute error (MAE), root mean square error (RMSE), and skill score (SS) based on the mean square error:

$$\%\ \text{Bias}=100\times\frac{\sum_{i=1}^{n}(\overline{\widehat{Y}_{i}^{mis}}-Y_{i})}{\sum_{i=1}^{n}Y_{i}}, \tag{6}$$

$$\text{MAE}=\frac{\sum_{i=1}^{n}|Y_{i}-\overline{\widehat{Y}_{i}^{mis}}|}{n}, \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{\widehat{Y}_i^{mis}})^2}{n}}, \quad (8)$$

$$\mathrm{SS} = 1 - \frac{MSE\left(Y,\ \overline{\widehat{Y}^{mis}}\right)}{MSE\left(\overline{Y}, \overline{\widehat{Y}^{mis}}\right)}, \quad (9)$$

where $\mathrm{MSE}\left(Y,\ \overline{\widehat{Y}^{mis}}\right) = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{\widehat{Y}_i^{mis}}\right)^2}{n}$ and $\mathrm{MSE}\left(\overline{Y}, \overline{\widehat{Y}^{mis}}\right) = \frac{\sum_{i=1}^{n}(1-I_i)\left(\overline{Y} - \overline{\widehat{Y}_i^{mis}}\right)^2}{n}$.

Percent bias measures the average tendency of imputed values to be larger or smaller than the associated observed values. A positive score indicates overestimation, whereas a negative score indicates underestimation. The optimal value is 0, with low values indicating plausible imputed values. The MAE and RMSE are widely reported imputation modelling performance indicators. MAE and RMSE have the same units as the variables measured. They are non-negative and unbounded above, with lower values indicating high levels of agreement between observed and estimated values. The SS measures the accuracy of a forecast relative to standard reference. The values of SS are bounded above by 1 and unbounded below. A perfect forecast is observed when a score of 1 is obtained.

Application

To further evaluate the ability of the method to impute plausible values with a distribution comparable to the observed periods, the model with the best overall performance in the cross-validation study was determined. This model (with covariates unique to the locations) was then applied to impute missing data in the camera datasets at Mindarie (Ryan *et al.*, 2017) and Monkey Mia (Ryan *et al.*, 2013). For Mindarie, there were 3 months with complete camera outages, as well as shorter duration outages in the other months; for Monkey Mia, there was a single longer duration outage in addition to intermittent outages.

Results

Case study: ten outage patterns applied to a complete dataset

The percentage of zero counts in the dataset with simulated missing data scenarios ranged from 21.2% to 51.8%. For these simulated periods, the average and standard deviation of observed hourly counts ranged from 1.0 to 1.5 ($2.1 \leq SD \leq 3.0$) while the imputed data had values ranged from 1.0 to 1.8 ($1.8 \leq SD \leq 2.7$). Total imputed estimates obtained from the nine models agreed with the actual totals in most cases (Table 2). For outage patterns 5 and 9, the 95% confidence intervals of the imputed total number of powerboat retrievals from the nine models overlapped with the observed total number of powerboat retrievals. For the ZIP and ZIP.boot models, the 95% confidence intervals of the imputed totals for nine of the outage patterns contained the observed total number of powerboat retrievals. For outage pattern 1, the 95% confidence intervals of the imputed totals did not contain the observed total for any of the nine models (see Figure 3). With respect to temporal strata such as months, the 95% confidence intervals of the imputed counts of powerboat retrievals for most of the models often contained the total observed counts (see Supplementary Table S2).

In terms of percent bias, models were ranked differently, but ZIP models were often among the top ranked models. The direction of the estimation of the bias also varied among the outage patterns. For example, the bias was positive for all the models for outage pattern 10, indicating overestimation of the total counts, but for outage pattern 1, all the models recorded negative bias, with underestimated total counts. In terms of MAE and RMSE, the indicators agreed on the top ranked models for all the outage patterns apart from 6. The ZIP models were top ranked most frequently (see Table 2). The relatively low values of MAE and RMSE suggest close agreement between imputed and observed data. The percentage differences in MAE and RMSE values between the two best models ranged from 0.1% to 4.7% and 0.04% to 7.3%, respectively. Nominal differences in MAE and RMSE among the three best models were relatively small and did not appear to be important. SS values, however, revealed some level of disparity in the performance of models and in most cases were distinctive in the choice of the best ranked model. Except for outage patterns 1 and 6, the SS consistently ranked the ZIP and its bootstrap variant as the best models, notably in missing patterns of very long duration (e.g. outage patterns 7 and 8). The percentage difference in the SS values between the two best models (models with larger SS scores) for the ten outage patterns ranged from 0.6% to 35.3%, with the magnitude of errors between −0.158 and 0.312. Although there was no clear systematic trend in the performance of the models with respect to the pattern, the proportion of missing data, and the proportion of zeros in the dataset, ZIP models were generally ranked best.

Application

For Mindarie, distributions across hours of the day adequately depicted the nature of boating activities, particularly for the 3 months where records were missing in their entirety (see Figure 4). This was inferred from the similarities that exist between the distributions for the imputed values and observed months. In addition, for the other months with some missing data, there were some differences in the shapes of the distributions of powerboat retrievals obtained with the imputations from the outlined method compared to the results in Ryan *et al.* (2017). For instance, the distributions for the months of January, May, and June were more regular in shape compared to results in Ryan *et al.* (2017). The variations in the imputations expressed in terms of the monthly total powerboat retrievals were also lower.

Similar patterns were observed in the analysis of Monkey Mia dataset. The distribution for the imputed values for May (with complete outage) adequately reflected the general patterns of the distributions across the observed months (see Supplementary Figure S2). The variation in the estimates of total monthly powerboat retrievals was lower than that for the method applied in Ryan *et al.* (2013).

To further understand the short-term behaviour and the consistency of the imputations, detailed daily distributions of the imputations, particularly for the months with no data have been provided (see Supplementary Figures S1, S3, and S4). The daily distributions of imputed values for April at Mindarie and May at

Table 2. Models' performance evaluation.

Outages	Outage 1					Outage 2					Outage 3					Outage 4				
Models	**Total (95% *CI*)**	**% Bias**	**SS**	**MAE**	**RMSE**	**Total (95% *CI*)**	**% Bias**	**SS**	**MAE**	**RMSE**	**Total (95% *CI*)**	**% Bias**	**SS**	**MAE**	**RMSE**	**Total (95% *CI*)**	**% Bias**	**SS**	**MAE**	**RMSE**
Observed	3 381					819					7 247					3 763				
PMM	3 164 ± 159	**−6.5**	0.19	1.31	2.36	745 ± 74	−15.3	0.25	1.12	1.82	7 446 ± 129	2.7	0.15	1.37	2.31	3 830 ± 165	1.8	0.09	1.22	2.26
QP	3 033 ± 152	−10.3	**0.27**	1.26	2.26	794 ± 71	−6.8	0.20	1.14	1.89	7 273 ± 130	0.4	0.13	1.35	2.33	3 674 ± 154	−2.4	**0.15**	**1.18**	**2.18**
QP.boot	3 124 ± 157	−7.6	0.20	1.31	2.35	764 ± 79	−9.0	0.18	1.19	1.92	7 042 ± 127	−2.8	0.14	1.34	2.32	3 617 ± 153	−3.9	0.14	1.18	2.20
NB	3 046 ± 159	−9.9	0.19	1.31	2.36	725 ± 76	−20.3	0.22	1.11	1.87	7 189 ± 130	−0.8	0.10	1.36	2.37	3 750 ± 159	**−0.3**	0.12	1.20	2.23
NB.boot	3 089 ± 130	−8.6	0.26	**1.24**	**2.26**	753 ± 65	−11.5	0.23	1.08	1.86	7 337 ± 133	1.2	0.05	1.38	2.44	3 494 ± 151	−7.2	0.13	1.18	2.21
ZIP	3 115 ± 131	−7.9	0.26	1.27	2.27	777 ± 55	−7.3	0.26	1.09	1.82	7 279 ± 102	**0.4**	**0.22**	**1.29**	**2.21**	3 712 ± 143	−1.3	0.13	1.81	2.21
ZIP.boot	3 090 ± 130	−8.6	0.25	1.25	2.27	784 ± 55	**−6.5**	**0.28**	**1.05**	**1.79**	7 329 ± 102	1.1	0.22	1.29	2.21	3 780 ± 147	0.4	0.15	1.19	2.19
ZINB	3 059 ± 158	−9.5	0.21	1.29	2.34	753 ± 66	−20.2	0.17	1.13	1.93	6 933 ± 107	−31.9	0.08	1.29	2.41	3 639 ± 168	−3.9	0.05	1.23	2.32
ZINB.boot	2 968 ± 151	−12.2	0.19	1.31	2.37	763 ± 63	−19.0	0.25	1.09	1.84	7 380 ± 162	1.8	−0.16	1.72	2.69	3 818 ± 183	−3.3	0.04	1.32	2.32
Outages	**Outage 5**					**Outage 6**					**Outage 7**					**Outage 8**				
Models	**Total (95% *CI*)**	**% Bias**	**SS**	**MAE**	**RMSE**	**Total (95% *CI*)**	**% Bias**	**SS**	**MAE**	**RMSE**	**Total (95% *CI*)**	**% Bias**	**SS**	**MAE**	**RMSE**	**Total (95% *CI*)**	**% Bias**	**SS**	**MAE**	**RMSE**
Observed	1 642					1 175					2 705					1 966				
PMM	1 749 ± 123	8.3	0.07	1.52	2.49	1 219 ± 102	3.7	0.01	1.62	2.99	3 014 ± 150	11.4	0.02	1.16	2.10	1 897 ± 129	−3.5	0.11	1.51	2.53
QP	1 649 ± 119	6.5	0.16	1.41	2.37	1 185 ± 93	**0.8**	**0.06**	**1.55**	2.90	2 709 ± 141	**0.2**	0.09	1.12	2.02	1 849 ± 123	−5.9	0.17	1.45	2.44
QP.boot	1 626 ± 116	5.1	0.20	1.38	2.32	1 145 ± 91	−1.7	0.12	1.47	**2.82**	2 732 ± 139	1.0	0.09	1.12	2.02	1 856 ± 120	−5.6	0.13	1.46	2.51
NB	1 614 ± 124	−1.7	0.12	1.42	2.43	1 199 ± 103	2.3	0.02	1.57	2.97	2 914 ± 140	7.7	0.06	1.16	2.06	1 807 ± 126	−8.1	0.06	1.46	2.60
NB.boot	1 664 ± 126	1.3	0.16	1.40	2.37	1 220 ± 99	3.8	0.05	1.54	2.92	3 034 ± 148	12.2	0.03	1.18	2.09	1 796 ± 124	−8.7	0.14	1.43	2.49
ZIP	1 625 ± 100	−1.2	0.24	1.39	2.26	1 251 ± 91	6.2	0.06	1.58	2.91	2 716 ± 123	1.9	0.12	1.07	1.99	1 964 ± 107	**−0.1**	**0.21**	**1.40**	**2.38**
ZIP.boot	1 645 ± 100	**0.4**	**0.25**	**1.35**	**2.25**	1 257 ± 90	6.9	0.05	1.57	2.92	2 728 ± 123	2.2	**0.13**	**1.07**	**1.98**	1 972 ± 107	0.3	0.17	1.42	2.44
ZINB	1 732 ± 118	7.9	0.11	1.42	2.44	1 392 ± 117	18.5	−0.09	1.74	3.14	2 653 ± 146	−4.5	0.02	1.11	2.10	2 061 ± 152	4.8	0.05	1.61	2.62
ZINB.boot	1 779 ± 135	8.5	0.14	1.52	2.41	1 797 ± 149	52.9	−0.06	2.08	3.22	2 626 ± 153	−5.1	0.01	1.12	2.11	2 095 ± 151	6.6	0.05	1.61	2.61

Outages	Outage 9				Outage 10			
Models	**Total (95% *CI*)**	**%Bias**	**SS**	**MAE**	**RMSE**	**Total (95% *CI*)**	**% Bias**	**SS**	**MAE**	**RMSE**
Observed	724					746				
PMM	758 ± 80	18.5	0.20	1.23	2.02	804 ± 88	21.1	−0.08	1.47	2.44
QP	695 ± 79	−9.8	0.23	1.18	1.98	836 ± 93	25.5	0.05	1.37	2.29
QP.boot	684 ± 74	−8.3	0.31	1.13	1.88	834 ± 88	25.3	−0.05	1.44	2.40
NB	732 ± 86	14.9	0.22	1.18	2.00	855 ± 103	28.0	0.01	1.42	2.35
NB.boot	793 ± 87	23.3	0.10	1.30	2.15	820 ± 97	23.3	−0.03	1.38	2.39
ZIP	710 ± 64	**−7.4**	0.25	1.19	1.97	757 ± 71	**14.9**	**0.14**	**1.32**	**2.18**
ZIP.boot	708 ± 63	−7.9	**0.31**	**1.15**	**1.88**	783 ± 74	18.3	0.09	1.36	2.24
ZINB	741 ± 87	16.2	0.10	1.31	2.15	805 ± 99	21.3	−0.06	1.51	2.42
ZINB.boot	735 ± 85	15.3	0.13	1.31	2.12	814 ± 101	22.5	−0.15	1.60	2.52

The table displays the observed total counts and the imputed total powerboat retrievals (with 95% confidence intervals), the percentage bias, the skill score, the mean absolute error, and the root mean error from the fitted models in relation to the ten missing patterns. The best models have bold scores with respect to the performance indicators.

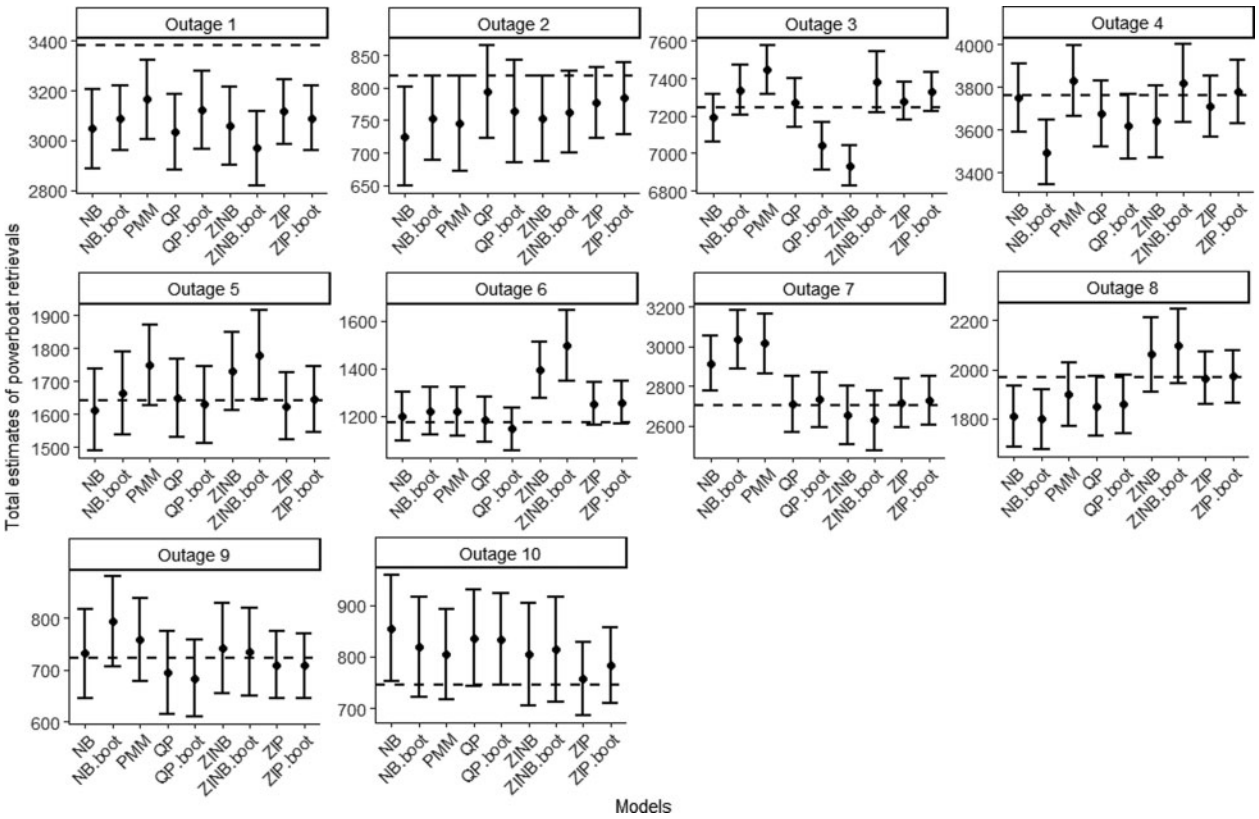

Figure 3. Total estimates of powerboat retrievals (with 95% confidence intervals) obtained from the nine fitted models for the ten missing patterns studied. The horizontal dashed lines represent the true observed total counts of powerboat retrievals at the Leeuwin boat ramps from the missing periods.

Monkey Mia adequately depicted the nature of traffic intensities at the two boat ramps.

Discussion

Generalized linear mixed models built on a fully conditional specification multiple imputation framework were found to reconstruct plausible values of counts of powerboat retrievals for the durations of outages studied. The modelling framework has demonstrated suitability for the imputation of missing data in count data sets. Generally, the choice and type of model will depend on the nature and characteristics of the data set and the missing patterns. However, the ZIP model in the multiple imputation scheme (with its "self-correcting properties") is likely to perform well for count data with many zeros and possible overdispersion. This is because, for such data, zero-inflated models provide a rigorous analytical approach. In addition, the overdispersed nature of such data will not impact on the results, since it uses the dependencies within the dataset to hierarchically model the variance structure. We recommend further simulation studies to assess varying modelling conditions and missing mechanisms of various types including missing completely at random (MCAR), where it is assumed that there is no relationship between missingness of the data and any values, observed or missing.

Robust imputation models with the ability to uncover the relationship between variables to "fill-in" gaps of missing data with values that will fit the distribution of the powerboat retrievals are ideal. In the framework outlined, the predicted values from the chain of equations formulated with the covariates were used to guide the random draws from the observations to impute missing data. To obtain plausible estimates using regression modelling requires the use of as much information as available (Kaiser and Tracy, 1988; van Buuren and Groothuis-Oudshoorn, 2011). Conceptually, it is difficult to determine all variables related to boating activity, as many other factors not considered in this study may be important. The covariates in the imputation modelling phase do not completely capture all the variability in the powerboat retrievals data. However, inclusion of more variables more variables might lead to collinearity with the response and among control variables. Perturbation analysis (see Hendrickx, 2018) can be applied to mitigate the impact of collinearity on the response. If collinear control variables do not covary with the response variable(s) (which was the case in this study), there will be no effect on coefficient estimation or model performance (Allison, 2012).

The results varied for lower-level temporal stratification and there were instances of under- and overestimation, notably in the PMM and the negative binomial models for time of day. The shortfalls of the PMM in imputing non-continuous variables are apparent (Allison, 2015). The approach performed well in imputing estimates for the large scale (e.g. the 12-month total number of powerboat retrievals at the Leeuwin ramp) but struggled at a finer-scale (e.g. time of the day). This was because PMM used an ordinary linear regression model in the estimation process and did not capture the clustering effects especially for temporal variables with several levels. Conversely, the estimation processes for the negative binomial models were more cumbersome and

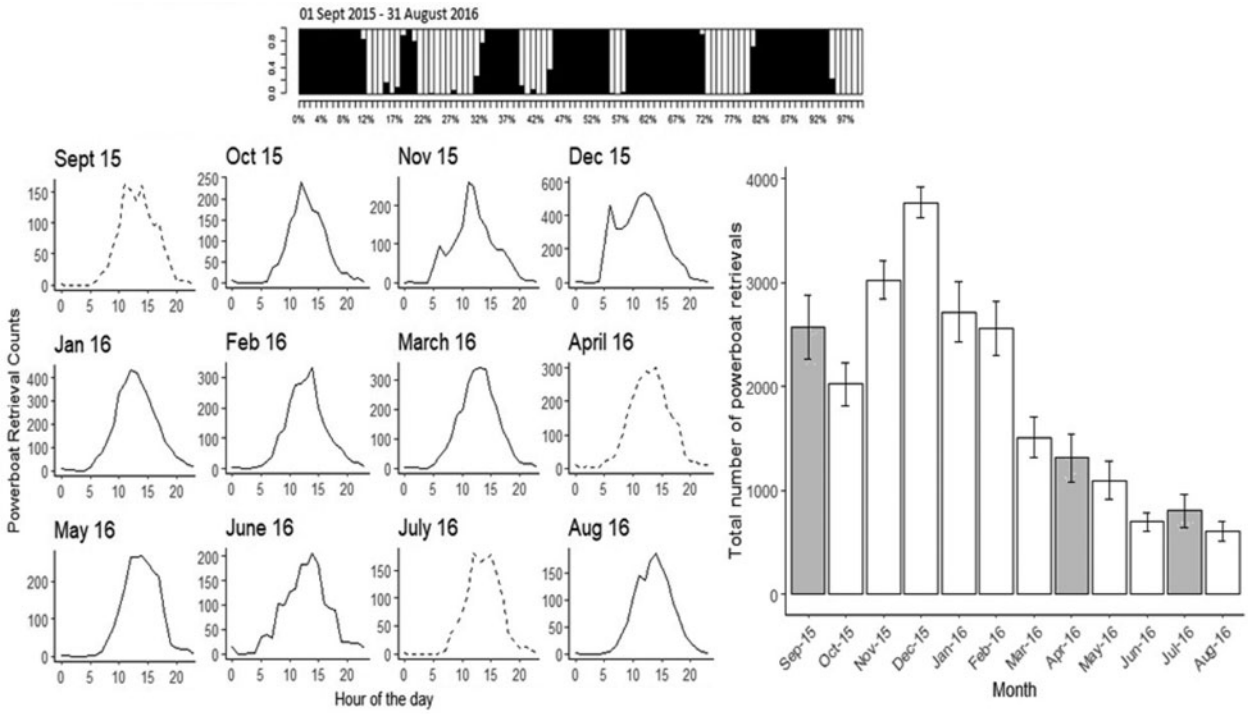

Figure 4. Left: outage pattern and monthly distributions across hours of the day for the total powerboat retrievals from Mindarie (lat 31.692, long 115.702) during 2015/2016. The distribution of the outage patterns is depicted as follows: the black bars indicate outage periods and white bars indicate observed periods. The distribution of the imputed months with complete outage is represented using dashed lines. For the other months with missing data, differences can be observed in shapes compared to the results in Ryan *et al.* (2017). Right: monthly distribution of the total number of powerboat retrievals, with 95% confidence intervals where data imputations were required. The grey bars represent the months with complete camera outage.

sometimes models had to be run for long periods of time before convergence. This was mostly a consequence of the variance being a quadratic function of the mean, which affected the iteratively weighted least squares algorithm. Ver Hoef and Boveng (2007) found the quasi-Poisson regression to be superior to the negative binomial regression in estimating the overall abundance of harbour seals (*Phoca vitulina*) with overdispersed count data, as the negative binomial regression tends to assign more weight in the parameter estimation process.

The opportunities that remote camera surveys provide for complementary and corroborative purposes in recreational fishing research are evident (Smallwood *et al.*, 2012; Hartill *et al.*, 2016; Lancaster *et al.*, 2017; Steffe *et al.*, 2017; Askey *et al*, 2018). Although some of the challenges of missing data in remote camera studies can be mitigated with measures such as regular maintenance schedules, back-up power supplies for cameras, and installing cameras in proximate locations to assist data sharing (van Poorten *et al.*, 2015; Hartill *et al.*, 2016), missing data cannot be completely eliminated. Our study has demonstrated that there is a need to explore, using known response data for the variable of interest, the extent to which imputation models successfully describe the distribution of that variable and adequately impute plausible values for missing periods. The current method is suitable for imputing reasonably long outage periods, but in an instance where an entire season is missing, more assumptions would be required. For instance, major outages (e.g. 9-month in a year) will compromise the quality of imputed estimates and the level of acceptance of the results. Alternatively, the nearest neighbourhood concept used in Hartill *et al.* (2016) could be applied. In addition, we propose that for ramps where continuous camera data have been collected, inferences could be made from estimates from the preceding years.

The outcomes of the imputation modelling can assist in dealing with outages in remote camera studies elsewhere. Within WA, the detailed analysis undertaken at the boat ramp at Leeuwin will form the basis of imputing missing counts of powerboat retrievals for camera outages for the other locations where remote cameras have been installed. This will enable the monitoring of long-term trends in boating and recreational fishing activity. The current study has a wide area of application in fisheries, ecological, and related studies involving remotely operated cameras and automatic traffic counters. For instance, remote camera monitoring data has been used to estimate nocturnal shore-based recreational fishing effort (Taylor *et al.*, 2018), to estimate angling effort (van Poorten *et al.*, 2015; Askey *et al*, 2018; Stahr and Knudsen, 2018), and to monitor human use of artificial reefs and areas of the coast (Wood *et al.*, 2016; Flynn *et al.*, 2018). As the number of fisheries and ecological studies using remote cameras is likely to increase, the need to consider accurate approaches for imputing missing data resulting from outages will become increasingly important to guide key management decisions.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Data availability statement

The data underlying this article were provided by the Department of Government of Western Australia Primary Industries and Regional Development under licence. Data will be shared on request to the corresponding author with permission of Government of Western Australia Primary Industries and Regional Development.

Funding

This study was funded and supported by Government of Western Australia Department of Primary Industries and Regional Development (DPIRD) and Edith This study was funded and supported by Government of Western Australia Department of Primary Industries and Regional Development (DPIRD) and Edith Cowan University (ECU) (G1002222).

Acknowledgements

The authors express their sincere gratitude to the staff of DPIRD who spent much time reading the camera data. The authors thank Agata Zabolotny for producing the map and Eva Lai for producing some of the figures in the Supplementary material. The authors would like to thank Stuart Blight and Karina Ryan for maintaining the network of cameras. A special thanks to Dr Norman Hall, Mr Cameron Desfosses, and Mr Mark Pagano for their time and input during the internal review process by DPIRD. The authors are also grateful to the Australian Government Bureau of Meteorology for providing the climate data for the study. The authors also thank the three anonymous reviewers for their constructive criticisms and the useful suggestions provided.

References

Afrifa-Yamoah, E., Mueller, U. A., Fisher, A. J., and Taylor, S. M. 2019. Fixed versus Random effects models: an application in building imputation models for missing data in remote camera surveys. *In* Proceedings of the 34th International Workshop on Statistical Modelling (IWSM), Guimarães, Portugal, 7–12 July.

Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., and Fisher, A. J. 2020. Missing data imputation of high-resolution temporal climate time series data. Meteorological Applications, 27: 1–18.

Allison, P. 2012. When Can You Safely Ignore Multi-Collinearity? Statistical Horizons. http://statisticalhorizons.com/multicollinearity (last accessed 31 October 2019).

Allison, P. 2015. Imputation by Predictive Mean Matching: Promise and Peril. Statistical Horizons. https://statisticalhorizons.com/predictive-mean-matching (last accessed 23 September 2019).

Askey, P. J., Ward, H., Godin, T., Boucher, M., and Northrup, S. 2018. Angler effort estimates from instantaneous aerial counts: use of high-frequency time-lapse camera data to inform model-based estimators. North American Journal of Fisheries Management, 38: 194–209.

Blight, S., and Smallwood, C. 2015. Technical Manual for Camera Survey of Boat- and Shore-Based Recreational Fishing in Western Australia. Fisheries Occasional Publication, No. 124, Department of Fisheries, Western Australia.

Department of Primary Industries and Regional Development. 2019. Department of Primary Industries and Regional Development Annual Report. https://dpird.wa.gov.au/annual-report (last accessed 5 September 2019).

Efron, B. 1994. Missing data, imputation and bootstrap. Journal of the American Statistical Association, 89: 463–475.

Flynn, D. J. H., Lynch, T. P., Barrett, N. S., Wong, L. S. C., Devine, C., and Hughes, D. 2018. Gigapixel big data movies provide cost-effective seascape scale direct measurements of open-access coastal human use such as recreational fisheries. Ecology and Evolution, 8: 9372–9383.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M., Nielsen, A., *et al.* 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software, 27: 233–249.

Hartill, B. W., Payne, G. W., Rush, N., and Bian, R. 2016. Bridging the temporal gap: continuous and cost-effective monitoring of dynamic recreational fisheries by web cameras and creel surveys. Fisheries Research, 183: 488–497.

Hartill, B. W., Taylor, S. M., Keller, K., and Weltersbach, M. S. 2020. Digital camera monitoring of recreational fishing effort: applications and challenges. Fish and Fisheries, 21: 204–215.

Hendrickx, J. (2018). Collinearity in Mixed Models. Paper AS03. PhUSE EU Connect.

Hyder, K., Weltersbach, M. S., Armstrong, M., Ferter, K., Townhill, B., Ahvonen, A., Arlinghaus, R., *et al.* 2018. Recreational sea fishing in Europe in a global context—participation rates, fishing effort, expenditure and implication for monitoring and assessment. Fish and Fisheries, 19: 225–243.

Jackman, S. 2008. pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory. Stanford University. Department of Political Science, Stanford University, Stanford, California.

Johnson, A. F., Moreno-Báez, M., Giron-Nava, A., Corominas, J., Erisman, B., Ezcurra, E., and Aburto-Oropeza, O. 2017. A spatial method to calculate small-scale fisheries effort in data poor scenarios. PLoS One, 12: e0174064.

Kaiser, J., and Tracy, D. B. 1988. Estimation of missing values by predicted score. Proceedings of the Section on Survey Research, American Statistical Association 1988, 631–635.

Kleinke, K., and Reinecke, J. 2013a. Multiple imputation of incomplete zero-inflated count data. Statistica Neerlandica, 67: 311–336.

Kleinke, K., and Reinecke, J. 2013b. Countimp 1.0—A Multiple Imputation Package for Incomplete Count Data [Tech. Rep.]. University of Bielefeld, Bielefeld, Germany. http://www.uni-bielefeld.de/soz/kds/pdf/countimp.pdf.

Lancaster, D., Dearden, P., Haggarty, D. R., Volpe, J. P., and Ban, N. C. 2017. Effectiveness of shore-based remote camera monitoring for quantifying recreational fisher compliance in marine conservation areas. Aquatic Conservation: Marine and Freshwater Ecosystems, 27: 804–813.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D. B. 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Ryan, K. L., Hall, N. G., Lai, E. K., Smallwood, C. B., Taylor, S. M., and Wise, B. S. 2015. State-wide survey of boat-based recreational fishing in Western Australia 2013/14. Fisheries Research Report 268, Department of Fisheries, Western Australia.

Ryan, K. L., Hall, N. G., Lai, E. K., Smallwood, C. B., Taylor, S. M., and Wise, B. S. 2017. State-wide survey of boat-based recreational fishing in Western Australia 2015/16. Fisheries Research Report 287, Department of Primary Industries and Regional Development, Western Australia.

Ryan, K. L., Wise, B. S., Hill, N. G., Pollock, K. H., Sulin, E. H., and Gaughan, D. J. 2013. An integrated system to survey boat-based recreational fishing in Western Australia 2011/12. Fisheries Research Report 249, Department of Fisheries, Western Australia.

Skaug, H., Fournier, D., Bolker, B., Magnusson, A., and Nielsen, A. 2015. Generalized Linear Mixed Models using 'AD Model Builder'. R package version 0.8.2.

Smallwood, C. B., Pollock, K. H., Wise, B. S., Hall, N. G., and Gaughan, D. J. 2012. Expanding Aerial-Roving Surveys to include counts of shore-based recreational fishers from remotely operated

cameras: benefits, limitations and cost-effectiveness. North American Journal of Fisheries Management, 32: 1265–1276.

Soykan, C. U., Eguchi, T., Kohin, S., and Dewar, H. 2014. Prediction of fishing effort distributions using boosted regression trees. Ecological Applications, 24: 71–83.

Stahr, K. J., and Knudsen, R. L. 2018. Evaluating the efficacy of using time-lapse cameras to assess angling use: an example from a high-use metropolitan reservoir in Arizona. North American Journal of Fisheries Management, 38: 327–333.

Steffe, A. S., Taylor, S. M., Blight, S. J., Ryan, K. L., Desfossess, C., Tate, A., Smallwood, C. B., Lai, E. K., Trinnie, F. I., and Wise, B. S. 2017. Framework for integration of data from remotely operated cameras into recreational fishery assessments in Western Australia. Fisheries Research Report 286, Department of Primary Industries and Regional Development (DPIRD), WA.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., *et al.* 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. British Medical Journal, 338: b2393–160.

Taylor, S. M., Blight, S. J., Desfosses, C. J., Steffe, A. S., Ryan, K. L., Denham, A. M., and Wise, B. S. 2018. Thermographic cameras reveal high levels of crepuscular and nocturnal shore-based recreational fishing effort in an Australian estuary. ICES Journal of Marine Science, 75: 2107–2116.

van Buuren, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 16: 219–242.

van Buuren, S., and Groothuis-Oudshoorn, K. 2011. MICE: multivariate imputation by chained equations in R. Journal of Statistical Software, 45: 1–67.

van Poorten, B. T., Carruthers, T. R., Ward, H. G. M., and Varkey, D. A. 2015. Imputing recreational angling effort from time-lapse cameras using an hierarchical Bayesian model. Fisheries Research, 172: 265–273.

Ver Hoef, J. M., and Boveng, P. L. 2007. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology, 88: 2766–2772.

Wise, B. S., and Fletcher, W. J. 2013. Determination and development of cost-effective techniques to monitor recreational catch and effort in Western Australian demersal finfish fisheries. Final Report for FRDC Project 2005/034 and WAMSI Subproject 4.4.3. Fisheries Research Report 245, Department of Fisheries, Western Australia.

Wood, G., Lynch, T. P., Devine, C., Keller, K., and Figueira, W. 2016. High-resolution photo-mosaic time-series imagery for monitoring human use of an artificial reef. Ecology and Evolution, 6: 6963–6968.

Yu, L. M., Burton, A., and Rivero-Arias, O. 2007. Evaluation of software for multiple imputation of semi-continuous data. Statistical Methods in Medical Research, 16: 243–243.

Zeileis, A., Kleiber, C., and Jackman, S. 2008. Regression models for count data in R. Journal of Statistical Software, 27: 1–25.

Handling editor: Kieran Hyder