2021

Imputation, modelling and optimal sampling design for digital camera data in recreational fisheries monitoring

Ebenezer Afrifa-Yamoah
Edith Cowan University

Edith Cowan University

Copyright Warning

Imputation, modelling and optimal sampling design for digital camera data in recreational fisheries monitoring

A Thesis with Publications presented to Edith Cowan University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

Ebenezer Afrifa-Yamoah
BSc (Hons), MEd, MSc, MPhil

2021

Supervisors

Associate Professor Ute Mueller

Dr Stephen M Taylor

School of Science Edith
Cowan University

2021

Use of thesis

[redacted]However, the literary rights of the author must be respected. If any passage from this thesis is quoted or closely paraphrased in a paper or written work prepared by the user, the source of the passage must be acknowledged in the work. If the user desires to publish a paper or written work containing passages copied or closely paraphrased from this thesis, in which the passage copied would in total constitute an infringement for the purpose of the Copyright Act, then the individual must first obtain the written permission of the author to do so.

Declaration

I certify that this thesis does not, to the best of my knowledge and belief:

(i) incorporate without acknowledgement any material previously submitted for a degree or diploma in any institution of higher education;

(ii) contain any material previously published or written by another person except where due reference is made in the text; or

(iii) contain any defamatory material.

(iv) I also grant permission for the Library at Edith Cowan University to make duplicate copies of my thesis as required.

Signed: 

Date: 15/01/2021

Abstract

Digital camera monitoring has evolved as an active application-oriented scheme to help address questions in areas such as fisheries, ecology, computer vision, artificial intelligence, and criminology. In recreational fisheries research, digital camera monitoring has become a viable option for probability-based survey methods, and is also used for corroborative and validation purposes. In comparison to onsite surveys (e.g. boat ramp surveys), digital cameras provide a cost-effective method of monitoring boating activity and fishing effort, including night-time fishing activities. However, there are challenges in the use of digital camera monitoring that need to be resolved. Notably, missing data problems and the cost of data interpretation are among the most pertinent. This study provides relevant statistical support to address these challenges of digital camera monitoring of boating effort, to improve its utility to enhance recreational fisheries management in Western Australia and elsewhere, with capacity to extend to other areas of application.

Digital cameras can provide continuous recordings of boating and other recreational fishing activities; however, interruptions of camera operations can lead to significant gaps within the data. To fill these gaps, some climatic and other temporal classification variables were considered as predictors of boating effort (defined as number of powerboat launches and retrievals). A generalized linear mixed effect model built on fully-conditional specification multiple imputation framework was considered to fill in the gaps in the camera dataset. Specifically, the zero-inflated Poisson model was found to satisfactorily impute plausible values for missing observations for varied durations of outages in the digital camera monitoring data of recreational boating effort.

Additional modelling options were explored to guide both short- and long-term forecasting of boating activity and to support management decisions in monitoring recreational fisheries. Autoregressive conditional Poisson (ACP) and integer-valued autoregressive (INAR) models were identified as useful time series models for predicting short-term behaviour of such data. In Western Australia, digital camera monitoring data that coincide with 12-month state-wide boat-based surveys (now conducted on a triennial basis) have been read but the periods between the surveys have not been read. A Bayesian regression framework was applied to describe the temporal distribution of recreational boating effort using climatic and temporally classified

variables to help construct data for such missing periods. This can potentially provide a useful cost-saving alternative of obtaining continuous time series data on boating effort.

Finally, data from digital camera monitoring are often manually interpreted and the associated cost can be substantial, especially if multiple sites are involved. Empirical support for low-level monitoring schemes for digital camera has been provided. It was found that manual interpretation of camera footage for 40% of the days within a year can be deemed as an adequate level of sampling effort to obtain unbiased, precise and accurate estimates to meet broad management objectives. A well-balanced low-level monitoring scheme will ultimately reduce the cost of manual interpretation and produce unbiased estimates of recreational fishing indexes from digital camera surveys.

Acknowledgements

My heartfelt gratitude goes to Associate Professor Ute Mueller for believing in me, supporting my PhD journey and championing my professional development. I am sincerely grateful to Dr. Stephen M. Taylor for his immense and continuous support. I must say that his enthusiasm and passion for fisheries research was a great source of motivation. To Dr. Aiden Fisher, I thank you for your guidance and contributions in the early to mid-phase of my journey. To Ute, Steve and Aiden, I say your patience, feedback, constructive discussions and criticisms have produced this work which is *a game changer for me*.

I thank the Government of Western Australia Department of Primary Industries and Regional Development (DPIRD) and Edith Cowan University (ECU) for the scholarship to pursue this PhD study. I am thankful to the staff at DPIRD who read and maintained the camera data, I felt pampered when other colleagues were stressed in collecting the data that were analysed in this study. I also thank the lecturers and administrative staff of School of Science, ECU for the diverse support provided to make my studies come to a successful end. The opportunities and trust reposed in me by the school has given me invaluable experience and fond memories.

To dad (Howard Thompson Yamoah) and mum (Agnes Pokua Yamoah), all I can say is *I'VE MADE IT!* and you made it possible. I remember your struggles, prayers and your continuous support for my wellbeing. To Juliet, Victor, Emmanuel and Manuel, thank you for always being there for me. To my wife, Abigail, you've been a pillar and a light house for me. Thank you for trusting my decisions and for supporting this dream. And to my PhD babies, Maame Pokua and Nana Ama, I am sorry for making you cry most mornings that I had to leave for school.

I thank Rev. Dr. Aboagye-Sarfo and family for their kind gestures especially for my early days in Perth. I thank Dr. Hayford Ofori, Emmanuel Aboagye, Dr. Kwaafo Awuah-Mensah, Dr. Adua Eric, Dr. Esther Adama, Dr. VF Nunfam and the Ghanaian community at ECU and Perth for the friendship. I would like to thank my office mates, Anna, Shannon, Eva, Brett, Najmeh, Hira, Caterina, Kan and Josh for the warm conversations and the positive energies transmitted. It has been a long journey for me and would like to thank all teachers and friends who guided and played a role my journey.

I am also grateful to the anonymous reviewers of different internationally recognized journals and the thesis for their precious time and scholarly comments which were helpful in further developing the manuscripts.

"***Finally, Now to Him who is able to do immeasurably more than all that I asked or imagined, according to His power that works in me, to Him be all the glory***".

Source of Funding

This research work was an industry collaboration project between Edith Cowan University and Department of Primary Industries and Regional Development (DPIRD). It addressed an aspect of the grant, Integrated state-wide survey of recreational fishing Phase 2: boat- and shore-based activity.

List of Publications

Published Journal Paper

I. Afrifa-Yamoah, E., Mueller, UA, Taylor, SM and Fisher, AJ (2020). Missing data imputation of high-resolution temporal climate time series data, *Meteorological Applications*, 27(1): e1873. https://doi.org/10.1002/met.1873 (**Chapter Two**)

II. Afrifa-Yamoah E, Taylor SM, Fisher, AJ and Mueller UA. (2020). Imputation of missing data from time-lapse cameras used in recreational fishing surveys. *ICES Journal of Marine Science*, 77(7-8), 2984-2994 https://doi.org/10.1093/icesjms/fsaa180. (**Chapter Four**)

III. Afrifa-Yamoah E, Taylor SM, and Mueller U. (2021). Trade-off assessments between reading cost and accuracy measures for digital camera monitoring of recreational boating effort, *Fisheries Research*, 233, 105757. https://doi.org/10.1016/j.fishres.2020.105757. (**Chapter Seven**)

Conference paper

IV. Afrifa-Yamoah E, Mueller UA, Taylor SM, and Fisher AJ (2019). Fixed versus random effects models: an application in building imputation models for missing data in remote camera surveys. *In the proceedings of the 34th International Workshop on Statistical Modelling (IWSM) (Volume II), Guimarães, Portugal, 7-12 July 2019.* (**Chapter Three**)

Manuscripts

V. Afrifa-Yamoah E, Taylor SM, and Mueller UA. Short-term prediction of recreational boating effort: Evaluation of intermittent demand and count data forecasting methods. *To be submitted to Journal of Time Series Analysis* (**Chapter Five**)

VI. Afrifa-Yamoah E, Taylor SM, Mueller UA Modeling climatic and temporal influences on powerboat launches at two Western Australian boat ramps with relevance to recreational fisheries. *To be submitted to Fisheries Research* (**Chapter Six**)

Copyright statement

I warrant that I have obtained, where necessary, permission from the copyright owners to use any third-party copyright material reproduced in the thesis (e.g. artwork).

Dedication

To:

Abigail (my beloved wife),
Maame Afia Pokua (my first fruit),
Nana Akua Frema (my second fruit),
The Yamoahs.

I respect & appreciate all the investments and sacrifices you've made for me.

List of Abbreviation

ACD	Autoregressive Conditional Duration
ACF	Autocorrelation Function
ACP	Autoregressive Conditional Poisson
AEMO	Australian Energy Market Operator
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
BIC	Bayesian Information Criterion
CI	Credible Interval
CLS	Conditional Least Squares
CV	Coefficient of Variation
DARMA	Discrete autoregressive moving average
DPIRD	Department of Primary Industries and Regional Development
EM	Expectation Maximization
EM-MCMC	Expectation Maximization Markov Chain Monte Carlo
ESS	Effective Sample Size
FCS-MI	Full Conditional Specification Multiple Imputation
FGLS	Feasible Generalized Least Square
GLM	Generalized Linear Models
GLMM	Generalized Linear Mixed Model
GMM	Generalized Method of Moments
HAC	Heteroskedasticity and Autocorrelation
ICES	International Council for the Exploration of the Sea
INAR	Integer-valued Autoregressive
INARMA	Integer-valued Autoregressive Moving Average
INMA	Integer-valued Moving Average
LOO-PIT	Leave-One-Out Cross-Validated Probability Integral Transform
MAE	Mean Absolute Error
MAR	Missing at Random
MASE	Mean Absolute Scaled Error
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo

MLAD	Multiple regression using the Least Absolute Deviation
MNAR	Missing not at Random
MP	Missing Pattern
NA	Missing data
NUTS	No-U-Turn Sampler
PACF	Partial Autocorrelation Function
P-BSHADE	Point Estimation Model of Biased Sentinel Hospitals-based Area Disease Estimation
PMM	Predictive Mean Matching
QP	Quasi Poisson
QQ-plot	Quantile-Quantile plot
RMSE	Root Mean Square Error
RSE	Relative Standard Errors
RSS	Ranked Set Sampling
SARIMA	Seasonal ARIMA
SBA	Syntetos-Boylan Approximation
SBJ	Shale-Boylan-Johnson Approximation
SMAPE	Symmetric Mean Absolute Percentage Error
SRS	Simple Random Sampling
SS	Skill Score
SSRS	Systematic Random Sampling
SSRSP	Stratified Random Sampling with Proportional Allocation
SSRSW	Stratified Random Sampling with Weighted Allocation
WA	Western Australia
ZINB	Zero Inflated Negative Binomial
ZIP	Zero Inflated Poisson

Table of Contents

Use of thesis …… i
Declaration …… ii
Abstract …… iii
Acknowledgements …… v
Source of Funding …… vi
List of Publications ……... vii
Copyright statement …… viii
Dedication …… ix
List of Abbreviation ……... x
List of Tables ……..xvii
List of Figures ……xviii
CHAPTER ONE ……1
General Introduction ……... 1
1.1 Background ……... 1
1.2 Digital camera monitoring in recreational fisheries research: opportunities and challenges 2
1.2.1 Camera outages ……..2
1.2.2 Missing data imputation ……3
1.2.3 Manual data interpretation and sampling ……...5
1.3 Potential drivers of boating activities ……... 7
1.3.1 Climatic variables and the missing data problem ……. 7
1.4 Study area and general data description ……... 8
1.5 Research objectives and questions …… 12
1.6 Thesis structure ……... 13
Chapter References ……... 15
CHAPTER TWO …… 20
Missing data imputation of high-resolution temporal climate time series data …… 20
2.1 Abstract ……... 20
2.2 Introduction …… 21
2.3 Methods ……...24
2.3.1 Study area and data description ……... 24
2.3.2 Autoregressive Integrated Moving Averages (ARIMA) …… 27
2.3.3 Structural Time Series Models …… 27

2.3.4 ARIMA (p, d, q) in state-space forms, Kalman filter and smoothing ……. 28
2.3.5 Multiple Regression Modelling ……... 30
2.3.6 Model performance evaluation …… 32
2.4. Results and Discussion …….. 32
2.5 Limitations and opportunities of the modelling techniques ……42
2.6. Conclusion ……. 43
Chapter Acknowledgements ……43
Chapter References ……... 43
CHAPTER THREE …….. 47
Fixed versus random effects models: an application in building imputation models for missing data in remote camera surveys ……..47
3.1 Abstract ……... 47
3.2 Introduction ……48
3.3 Method …… 49
3.3.1 Data description ……... 49
3.3.2 Missing data and assumptions ……. 51
3.3.3 Modelling framework …….. 51
3.3.4 Full-conditional specification multiple imputation (FCS-MI) ……55
3.3.5 Model evaluation ……. 56
3.4 Results ……57
3.5 Discussion ……... 60
Chapter Acknowledgments ……... 62
Chapter References ……... 62
CHAPTER FOUR ……65
Imputation of missing data from time-lapse cameras used in recreational fishing surveys …. 65
4.1 Abstract ……... 65
4.2 Introduction ……66
4.3 Methods ……...68
4.3.1 Study area and camera data ……..68
4.3.2 Models and missing data assumptions ……70
4.3.3 Fully-conditional specification multiple imputation (FCS-MI) ……...72
4.3.4 Pooling analysis ……... 73
4.3.5 Model performance evaluation …… 73
4.4 Application ……..74

4.5 Results ……74
4.5.1 Case study: Ten outage patterns applied to a complete dataset ……... 74
4.4.2 Application ……...78
4.5 Discussion ……... 80
Supplementary material …… 82
Data Availability Statement ……..82
Chapter Acknowledgements …… 82
Chapter References ……... 83
Chapter Appendix …… 87
CA 1: Imputation model specifications …… 87
CA 2: Imputation algorithms …… 87
Chapter Appendix References …….. 95
CHAPTER FIVE …….. 96
Short term prediction of recreational boating effort: Evaluation of intermittent demand and count data forecasting methods ……96
5.1 Abstract ……...96
5.2 Introduction ……97
5.3 Methods ……...99
5.3.1 Data description ……... 99
5.3.2 Modelling techniques ……102
5.3.3 Point forecast accuracy evaluation ……104
5.4 Results ……...105
5.5 Discussion …… 109
5.6 Conclusion …… 111
Chapter Acknowledgements ……... 112
Chapter References …… 112
CHAPTER SIX ……... 116
Modeling climatic and temporal influences on powerboat launches with relevance to recreational fisheries ……... 116
6.1 Abstract …… 116
6.2 Introduction ……...117
6.3 Methods ……119
6.3.1 Study area and data description …… 119
6.3.2 Bayesian Regression Modeling ……..122

6.3.3 Model specifications and implementation 124
6.3.4 Model evaluation 124
6.4 Results 125
6.4.1 Model fit and cross-validation 125
6.4.2 Reconstructing observed data 129
6.6 Conclusion 134
Declaration of competing interest 134
Chapter Acknowledgement 134
Chapter References 134
Chapter Appendix 138
CHAPTER SEVEN 152
Trade-off assessments between reading cost and accuracy measures for digital camera monitoring of recreational boating effort 152
7.1 Highlights 152
7.2 Abstract 152
7.3 Introduction 153
7.4 Materials and methods 156
7.4.1 Study area 156
7.4.2 Data collection and treatment 156
7.4.3 Sampling units and monitoring design 157
7.4.4 Data analysis 160
7.5 Results 161
7.5.1 Distribution of powerboat retrievals across ramps and strata 161
7.5.2 Estimation of the average number of daily powerboat retrievals 163
7.5.3 Estimation of the annual number of powerboat retrievals and cost 163
7.5.4 Accuracy, precision and coverage rate estimates 165
7.6 Discussion 168
7.7 Conclusion 171
Chapter Acknowledgements 172
Credit authorship contribution statement 172
Chapter References 172
CHAPTER EIGHT 176
General Discussion 176
8.1 Discussion 176

8.2 Limitations and future work ……..180
Chapter References …… 181
Appendix A: Sample R-codes ……183
Appendix B: Statement of co-authors contribution …… 191

List of Tables

Table 2.1: Description of sub-samples and the number of missing values ……. 26

Table 2.2: Multiple linear regression model formulation …..……. 31

Table 2.3: Model ranking based on the pooled average performance indicators' values across the five-fold gaps of missing observations ……..…..…… 35

Table 2.4: The correlation coefficient between the imputed values and the observed values with respect to the estimation techniques for the five-missing folds …….…….... 41

Table 3.1: A) Multiple imputation scheme for the Quasi-Poisson B) Multiple imputation scheme for the Zero-inflated Poisson models …….…... 57

Table 3.2: Models' performance evaluation. The table displays the characteristics of the missing patterns including the minimum (min) and maximum (max) duration of outages, the average total boat counts imputed from the fitted models versus total observed counts with associated standard deviations, and the average performance indicators across the ten imputation runs .. 59

Table 4.1: Study variables and their attributes (NA indicates the number of missing records) 71

Table 4.2: Models' performance evaluation. The table displays the observed total counts and the imputed total counts (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 …… 77

Table CA4.1: Location and survey period corresponding to an outage pattern …….……... 88

Table CA4.2: Total observed and imputed powerboat retrievals (with the 95% confidence bounds) with respect to months ……..……..……..……. (89-90)

Table 5.1: Mean absolute error (MAE), root mean square error (RMSE) and mean absolute scaled error (MASE) values for the performance evaluation of five techniques used to forecast lead times of 12-, 24-, 48- and 168-hours of recreational boating data for three study locations…….. 107

Table 6.1: Study variables and their attributes …….……..……. 120

Table 6.2: Summary statistics for the model fit and posterior distributions of the population level effects and family specific parameter …….…… 128

Table 7.1: Summary of the sampling design and sampling fractions used for digital camera studies on recreational fisheries ..……..……. 155

Table 7.2: Distributional characteristics and attributes of the counts of powerboat retrievals obtained from remote cameras at Denham (low traffic), Leeuwin (moderate traffic) and Hillarys (high traffic) within season and day type strata ……..…… 162

Table 7.3: Average relative standard error (+ standard deviations) and the coverage rate from the 10,000 jackknife draws for the sampling designs across the sampling proportions from the camera records of Denham (low traffic), Leeuwin (moderate traffic) and Hillarys (high traffic) boat ramps (SRS – simple random sampling, SSRS – systematic sampling, SRSP – stratified random sampling with proportional allocation, and SRSW – stratified random sampling with weighted allocation) ……..…… 167

List of Figures

Fig. 1.1: Study area showing the locations of the network of cameras for monitoring boating activities at boat ramps, groynes along sections of the foreshore ……..……. 9

Fig 1.2**:** Distribution of powerboat launches and retrievals at four boat ramps for the integrated survey periods between 01 March 2011 to 29 February 2012 in Western Australia ……11

Figure 2.1: Western Australia's (WA) climate classification ……..……..... 25

Figure 2.2: The distribution of the lengths of gaps measured in hours in the five folds in the four locations ……..……..... 26

Figure 2.3: Distributional characteristics, paired scatter plots and Pearson correlation between study variables with respect to the locations ……..……... 31

Figure 2.4: The mean absolute error (MAE), root mean square error (RMSE) and symmetric mean absolute error (sMAPE) values for the performance evaluation of imputing the five-missing folds of temperature, humidity and wind speed from the different from the different estimation techniques for the four study locations …..… 33

Figure 2.5: Density plots comparing the distribution of the observed and the five-missing folds imputed values for temperature, humidity and wind speed in Perth …. 37

Figure 2.6: Density plots comparing the distribution of the observed and the five-missing folds imputed values for temperature, humidity and wind speed in Broome .. 38

Figure 2.7: Density plots comparing the distribution of the observed and the five-missing folds imputed values for temperature, humidity and wind speed in Exmouth 39

Figure 2.8: Density plots comparing the distribution of the observed and the five-missing folds imputed values for temperature, humidity and wind speed in Esperance …….. 40

Figure 3.1: Distribution of the outage patterns applied to the Leeuwin dataset. 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 brown bands represent the periods of camera outages and the light green shades represent the observed data ……. 50

Figure 3.2: Correlation plot depicting the strength and direction of the correlation among the study variables. (Note: Prec = Precipitation, Temp = Temperature, Hum = Humidity, WinS = Wind speed, Wsin & Wcos = sine and cosine transformation of wind direction, WinG = Wind gust and SLP = sea level pressure) …… 51

Figure 3.3: Error diagnostics for the generalized linear model with serially correlated predictors before the Cholesky transformation ……. 54

Figure 3.4: Error diagnostics for the generalized linear model with serially correlated predictors after the Cholesky transformation …….. 54

Figure 3.5: Model performance based on the percent bias, RMSE and skill scores for the ten replicates of the multiple imputation scheme. Lowest and highest values of RMSE and skill scores respectively indicate best models …….. 60

Figure 4.1: 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 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 ……. 69

Figure 4.2: 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 …….. 70

Figure 4.3: 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 …… 76

Figure 4.4: 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 2015/16. 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 outages are 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 ……. 79

Figure CA4.1: Hourly distribution of the total observed and total imputed powerboat retrievals for the ten outage patterns …… 91

Figure CA4.2: Left: Outage pattern and monthly distributions across hours of the day for the total powerboat retrievals from Monkey Mia (Lat 25.793, Long 113.720) during 2011/12. The distribution of the outage patterns is depicted as follows: the black bars indicate outage periods and white bars indicate observed periods. Right: The distribution of the imputed months with complete outage are represented using the dashed lines. For the other months with missing data, differences can be observed in shapes compared to the results in Ryan *et al*. (2013). Monthly distribution of the total number of powerboat retrievals, with 95% confidence intervals where data imputations were required. The grey bar represents the month with complete camera outage ……..……. 92

Figure CA4.3: Daily distribution of the imputed hourly counts of powerboat retrievals for the month of April at the Mindarie boat ramp (Lat. 31.692, Long 115.702) during 2015/16. Data was sourced from Ryan et al. (2017, p. 181) …….…….. 93

Figure CA4.4: Daily distribution of the imputed hourly counts of powerboat retrievals for the month of May at the Monkey Mia boat ramp (Lat. 25.793, Long 113.702) during 2011/12. Data was sourced from Ryan et al. (2013, p. 152) …….…….…… 94

Figure 5.1: Time series plots, ACF and PACF of the count of powerboat launches from digital camera monitoring observed at Broome between 1 May 2013 and 30 April 2014. The short time series plots reflect the marked areas of the longer time series …… 100

Figure 5.2: Time series plots, ACF and PACF of the count of powerboat launches from digital camera monitoring observed at Denham between 1 May 2013 and 30 April 2014. The short time series plots reflect the marked areas of the longer time series ……101

Figure 5.3: Time series plots, ACF and PACF of the count of powerboat launches from digital camera monitoring observed at Denham between 1 March 2013 and 29 February 2014. The short time series plots reflect the marked areas of the longer time series ...101

Figure 5.4: Reconstruction of observed data based on the five time series models formulated using lead times of 12-, 24-, 48- and 168-hours at A) Broome, B) Denham and C) Leeuwin ...……...……..…. 108

Figure 6.1: Time series and distribution of the number of daily powerboat launches at Hillarys between 1 March 2011 and 29 February 2012 (training set) and 1 May 2013 and 30 April 2014 (test set) …… 121

Figure 6.2: Time series and distribution of the number of daily powerboat launches at Broome between 1 March 2011 and 29 February 2012 (training set) and 1 May 2013 and 30 April 2014 (test set) ……..…….…… 121

Figure 6.3: Time series plots of temperature, humidity and precipitation at Broome and Hillarys between 1 March 2011 and 29 February 2012 (training set) and 1 May 2013 and 30 April 2014 (test set) …… 122

Figure 6.4: Model evaluation – A1 & B1) qq-plot for 300 leave-one-out cross-validated probability integral transform (LOO-PIT) for Hillarys and Broome respectively; A2 & B2) densities overlay from 1000 ensembles of predicted distribution for daily powerboat launches from the fitted models and the observed data for Hillarys and Broome respectively …… 127

Figure 6.5: Model evaluation for predicting test data: (A) pp-plot for observed vs. predicted number of daily powerboat launches at Hillarys (B) densities overlay of the observed and predicted data (C) time series plot of the observed data (black) superimposed by the forecasted data (blue) and the 95% credible interval (CI) (red) 129

Figure 6.6: Model evaluation for predicting test data: (A) pp-plot for observed vs. predicted number of daily powerboat launches at Broome (B) densities overlay of the observed and predicted data (C) time series plot of the observed data (black) superimposed by the forecasted data (blue) and the 95% credible interval (CI) (red) 130

Figure 6.7: Unobserved information on recreational boating effort constructed using Bayesian regression modelling for the periods of 01-March 2012 to 30-April 2013 and 01-May 2014 to 31-July 2015 at A) Hillarys and B) Broome. The red lines represent the constructed data based on models formulated using observed data from preceding year with the dashed lines representing the 95% credible intervals …….. 131

Figure CA6.1: Model evaluation for a single realisation: (A) observed vs. predicted number of daily powerboat launches at Hillarys (B) densities overlay of the observed

and predicted data (C) time series plot of the observed data (black) superimposed by the predicted model (blue) and the 95% credible interval (CI) (red) …….. 138

Figure CA6.2: Marginal effects plots of all population-level predictors (with 95% credible intervals) of the model for predicting the daily count of powerboat launches at Hillarys boat ramp …… 139

Figure CA6.3: Trace and density plots of all relevant parameters of the model for predicting the daily count of powerboat launches at Hillarys boat ramp …… (140-141)

Figure CA6.4: The Gelman-Rubin statistic for the MCMC iterations in estimating relevant parameters of the model for predicting the daily count of powerboat launches at Hillarys boat ramp. The shrinkage factor assesses convergence by comparing the estimated between-chains and within-chain variances for each model parameter. Large differences between these variances indicate non-convergence (Gelman and Rubin, 1992). The model achieved convergence as the potential scale reduction statistics (Gelman-Rubin statistics) were greater than 0.9 and less than 1.05 for all predictors ……..…. (142-144)

Figure CA6.5: Model evaluation for a single realisation: (A) observed vs. predicted number of daily powerboat launches at Broome (B) densities overlay of the observed and predicted data (C) time series plot of the observed data (black) superimposed by the predicted model (blue) and the 95% credible interval (CI) (red) …….. 145

Figure CA6.6: Marginal effects plots of all population-level predictors (with 95% credible intervals) of the model for predicting the daily count of powerboat launches at Broome boat ramp ……..... 146

Figure CA6.7: Trace and density plots of all relevant parameters of the model for predicting the daily count of powerboat launches at Broome boat ramp ……. (147-148)

Figure CA6.8: The Gelman-Rubin statistic for the MCMC iterations in estimating relevant parameters of the model for predicting the daily count of powerboat launches at Broome boat ramp. The model achieved convergence as the potential scale reduction statistics (Gelman-Rubin statistics) were greater than 0.9 and less than 1.05 for all predictors ……. (149-151)

Fig. 7.1: Study area showing the locations of the Hillarys (high-use), Leeuwin (medium-use) and Denham (low-use) boat ramps where remote camera data were analysed ... 157

Fig. 7.2: Average number of powerboat retrievals, coefficient of variation and root mean square error as a function of sample size proportion based on a posteriori data analysis. Sample units were randomly selected without replacement from the camera records of Denham (low traffic), Leeuwin (moderate traffic) and Hillarys (high traffic) boat ramps. Results presented were averaged over 10000 resamples. The error bars are 1 standard error of the average of the estimates from the 10,000 resamples. The horizontal dashed lines represent the true point estimates based on census all counts from observed data sets. (SRS – simple random sampling, SSRS – systematic sampling, SRSP – stratified random sampling with proportional allocation, and SRSW – stratified random sampling with weighted allocation) ……..…… 164

Fig. 7.3: Expanded total number of powerboat retrievals, total cost of manual interpretation and the 95% predicted margin of error as a function of sample size proportion based on a posteriori data analysis. Sample units were randomly selected without replacement using the different sampling techniques from the camera records of Denham (low traffic), Leeuwin (moderate traffic) and Hillarys (high traffic) boat ramps. Results presented were averaged over 10000 resamples. The error bars are 1 standard error of the average of the estimates from the 10000 resamples. The horizontal dashed lines represent the true point estimates based on the observed data sets. (SRS – simple random sampling, SSRS – systematic sampling, SRSP – stratified random sampling with proportional allocation, and SRSW – stratified random sampling with weighted allocation) …….. 166

CHAPTER ONE[1]

General Introduction

1.1 Background

Cameras have been used for surveillance purposes for over 70 years and are widely used in different areas to address a variety of pertinent issues of society (Dornberger, 1954; Zhang, 2017). For instance, in the area of security, digital cameras are being used as a mainstream crime prevention measure across the globe, and for other security purposes such as face detection and recognition (Ashby, 2017; Piza *et al*., 2019; Zhang, 2017). In an empirical analysis, digital camera monitoring was found to significantly increase the chances of solving crimes of different types (Ashby, 2017). In other areas such as tourism, ecology, household management and transportation, digital cameras are being used to help monitor usage of facilities, status of animals, and to promote road and home safety (Baran *et al.*, 2016; Zhang, 2017). Additionally, in research and industries, the application of digital camera monitoring has become a viable option for probability-based survey methods, which often expand sample estimates to population totals and are also used to estimate relevant indexes. There are established measures of effectiveness in the value of intelligence obtained from digital camera surveillance in key decision-making process (Cayford and Pieters, 2018). Importantly, security and data accessibility are key considerations to ensure that the collection of video images does not breach privacy regulations (Bernal, 2016).

In this study, attention is given to the application of digital camera monitoring of boat-based recreational activity in Western Australia. Recreational fishing is a popular outdoor activity worldwide. Approximately 11.5% of the world population are engaged in recreational fishing and it contributes significantly to economies as a source of leisure, job creation and revenue (Cooke and Cowx, 2004; Arlinghaus and Cooke, 2005). Boat-based recreational fishing activity is common and contributes to exploitation of fish populations, with potential sustainability impacts. While the majority of fisheries are managed sustainably, there are concerns for some fish stocks (DPIRD Annual Report, 2019). The participation rate in recreational fishing from this report (~26%) demonstrates the popularity of the activity and the need to ensure it is managed appropriately. Recreational fishing surveys play an integral role in providing information on recreational fishers required for fisheries management. Both on-site (e.g. access point, bus-route, aerial roving, traffic counters and digital camera monitoring) and off-site (e.g. mail, telephone) surveys methods are used (Afrifa-Yamoah *et al.*, 2019, 2020; Hartill *et al.*, 2019; Lai *et al*., 2019;

[1] This thesis is presented and organised as "Thesis with publication" format.

Ryan *et al.*, 2017; Smallwood *et al*., 2012; Taylor *et al.*, 2018; van Poorten and Brydle, 2018). In some parts of the world, recreational fishing activity originates at designated locations (e.g. boat ramps, choke points, estuary channel, groynes) enabling the potential census of boating effort (Hartill *et al*., 2019; Ryan *et al.*, 2017), which serves as a good proxy for fishing effort in some regions (Johnson *et al*., 2017; Taylor *et al*., 2019), and can help in quantifying the recreational catch. In addition, information on boating effort can be useful for validation and corroborative purposes (Steffe *et al.*, 2017), and promotes the understanding of the dynamics of recreational boaters' behaviour to enhance effective management of boat ramps and recreational fisheries.

1.2 Digital camera monitoring in recreational fisheries research: opportunities and challenges

Paucity of information is common in recreational fisheries records as a result of the fact that there are no mandatory requirements for recreational fishers to provide information on their fishing trips. Also due to the large number of recreational fishers in WA and the vast array of fishing spots, the use of digital cameras in fisheries studies is increasing in line with technological advances in recreational fisheries, providing an opportunity for continuous monitoring of boating activities in a field of view, for instance, a boat ramp or choke point (Hartill *et al*., 2019). The use of cameras provides data with wider and better coverage of the temporal sampling frame for fishing effort, although without knowledge of the actual nature of boating activities (Steffe et al., 2008). They can provide a general overview of boating activities and information to complement the sampling challenges from other survey methods in monitoring fishing effort. Cameras can be operated for extended periods of time in remote locations; for example, they can operate 24 hours per day, providing the opportunity to monitor night-time boating activities. They are also effective for capturing daily and seasonal effort trends and are potentially cost-effective (Smallwood *et al*., 2012). Hartill *et al.* (2019) reviewed the literature expanding on the applications and challenges of digital camera monitoring of recreational fishing effort. Fisheries agencies and researchers are interested in building the capacity to fully understand and integrate information obtained from the cameras for management purposes (Bian and Hartill, 2015; Hartill, 2015; Ryan *et al*., 2015).

1.2.1 Camera outages

Although digital camera monitoring of recreational boating activity provides a substantial amount of data, intermittent challenges with cameras' operations can result in the occurrence of significant gaps in the data. Camera outages occur frequently as a result of technical faults,

vandalism, theft, weather conditions such as lightning strikes, and flooding and environmental factors such as extreme temperature and humidity (Blight and Smallwood, 2015). Dealing with missing values has been a subject of interest for researchers in diverse fields. Missing data require proper handling to safeguard precision and reliability of estimates and indexes (van Buuren and Groothuis-Oudshoorn, 2011; van Poorten *et al*., 2015). The types of missing data mechanisms have peculiar patterns and statistical properties that significantly inform the suitable imputation techniques and their implied assumptions. The duration of outages in the remote camera data is also an important consideration in an imputation scheme. The outages in digital camera monitoring may persist for long periods due to technical and logistic inefficiencies or remoteness. The longer the duration of a camera outage the more the overall quality of data is compromised.

1.2.2 Missing data imputation

Several imputation approaches have been proposed and applied in different research areas such as fisheries, meteorology, medicine, neurology, transportation (Amiri and Jensen, 2016; Deb and Liew, 2016; Hartill *et al.*, 2016; Junger and de Leon, 2015; Purwar and Singh, 2015; Sovilj *et al.*, 2016; van Poorten et al., 2015). We can broadly distinguish two types of imputation schemes, single imputations and multiple imputation. Single imputation schemes impute missing data once with the best plausible estimate. The scheme includes simpler methods such as mean substitution and regression-based estimates. However, estimates obtained from such methods usually exhibit greater uncertainties compared to multiple imputation schemes (van Buuren and Groothuis-Oudshoorn, 2011). In multiple imputation schemes, missing observations are imputed m (m>1) times, yielding m plausible complete datasets. Then statistical analyses are performed on the m datasets and the parameter estimates and variances so obtained are pooled to obtain the missing data plausible estimates. These procedures result in improved estimates of uncertainties and are generalizable (van Buuren and Groothuis-Oudshoorn, 2011).

Multiple imputation schemes are generally constructed from two approaches, namely the joint modeling approach (Schafer, 1997) and sequential regression modeling approach (van Buuren and Groothuis-Oudshoorn, 2011). The Bayesian joint modeling approach specifies the joint probability model for the observed and missing data. This approach could possibly run into some analytical problems for large datasets which could result in a non-converging numerical solution (Engel *et al*., 2015). The sequential regression modeling approach, on the other hand, imputes on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable, thus providing a flexible framework for multiple imputation (van Buuren and Groothuis-Oudshoorn, 2011). The focus of multiple imputation is to minimize misclassifications of imputed

values by preserving the original distribution of the dataset during imputation and produces robust and unbiased estimates.

In fisheries research, the imputation work of van Poorten *et al*. (2015) and Hartill *et al*. (2016) on camera data is worth mentioning for the following reasons. van Poorten *et al*. (2015) developed a hierarchical Bayesian model to predict total angling effort (with the model accounting for three typical issues with camera effort) in recreational fishing using a multiple imputation scheme. Missing camera effort data were imputed from the average effort from proximate lakes. In Hartill *et al*. (2016), a high degree of correlation between the number of trailer boats returning at three ramps informed the imputation of missing values for the one ramp (where outage occurred) compared with the observed counts of the other two ramps (which were square root transformed). The number of days with outages represented 7% of the entire survey days. In Hartill's study, generalized linear models (GLMs) were used in the imputation modelling scheme. Both studies recommended incorporating the effects of covariates in the development of imputation scheme for data from remote camera surveys (van Poorten *et al*., 2015; Hartill *et al*., 2016).

Analytical techniques for data involving missing observations make use of various assumptions which are mostly dependent on the underlying missing-data mechanism and the actual physical pattern of missingness (Little and Rubin, 2002). Missingness in data may be completely at random (that is, data are missing independently of both observed and unobserved data), or missingness may be random (implies that given the observed data, data are missing independently of unobserved data) or missingness may be non-random (implies that missing observations are related to values of unobserved data). There is also the case where the missing mechanism is censored (De Jong et al., 2016). In this thesis, the missing mechanism assumed for the remote camera data is missing at random. This was because, camera outages occurred in a random fashion and were independent of the observed data. Therefore, imputation may be performed based on the observed variable, $\boldsymbol{Y}$, and some covariates, $\boldsymbol{X}$ (with complete observations). Modeling the missing mechanism will help remove systematic bias, which poses sampling selection problems and generally makes the process more efficient.

For the model setup, let $\boldsymbol{Y}$ denote an $n \times p$ matrix, where n and p are respectively the number of observations and the number of observed variables, that is, the count of boat launches and retrievals (with some missing observations). The matrix is then divided into elements of observed and missing, that is, $\boldsymbol{Y} = \{\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}\}$ and $P(\boldsymbol{Y}|\theta) = P\left(\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}\middle|\theta\right)$ is the joint

distribution of $\boldsymbol{Y}^{obs}$ and $\boldsymbol{Y}^{mis}$, where θ denotes unknown parameters. Let $\boldsymbol{M}$ denote an $n \times p$ binary matrix for observed and missing data, where m_i is an indicator variable defined as

$$m_i = \begin{cases} 1, & \text{if } y_i \text{ is observed} \\ 0, & \text{if } y_i \text{ is missing} \end{cases} \quad (1.1)$$

with y_i being realizations of the i^{th} $\boldsymbol{Y}$. Assuming that $P(\boldsymbol{M}|\boldsymbol{Y};\theta,\phi)$ is the conditional probability distribution of missingness, where ϕ represents the unknown parameters of $\boldsymbol{M}$ given $\boldsymbol{Y}$, with a joint parameter space (θ,ϕ). The joint distribution of $\boldsymbol{Y}$ and M can be expressed as

$$P(\boldsymbol{Y},\boldsymbol{M}|\theta,\phi) = P(\boldsymbol{Y}|\theta)P(\boldsymbol{M}|\boldsymbol{Y};\theta,\phi) \quad (1.2)$$

For $\boldsymbol{Y}$ and $\boldsymbol{X}$ (an $n \times q$ matrix, where q is the number of covariates considered), the model for a missing at random mechanism of interest is given by

$$P(\boldsymbol{M}|\boldsymbol{Y},\boldsymbol{X};\theta,\phi,\beta) = P(\boldsymbol{M}|\boldsymbol{Y}^{obs},\boldsymbol{X};\theta,\phi,\beta) \quad (1.3)$$

where β represents the unknown parameter(s) of $\boldsymbol{X}$. We are interested in investigating potential models to evaluate

$$P(\boldsymbol{Y}^{mis}|\boldsymbol{Y}^{obs},\boldsymbol{M},\boldsymbol{X};\Omega) \quad (1.4)$$

where Ω is model's parameter space, such that $(\theta,\phi,\beta) \in \Omega$.

1.2.3 Manual data interpretation and sampling

Remote cameras can be operational throughout the year. If the images are interpreted from all 365 days of the year, without any subsampling, it would mean complete monitoring and estimation of boating effort. However, this would result in many images and monitoring traffic at multiple sites could be very demanding. The time intensity and the reading cost (Smallwood *et al*., 2012; Steffe *et al*., 2017) of manual interpretation of images captured by cameras have necessitated the development of strategies that minimize the number of images interpreted (Steffe *et al*., 2008; Hartill, 2015; Hartill *et al*., 2016). With budgetary constraints, managers of remote camera surveys require methods for reducing cost of data interpretation without compromising much of the precision and accuracy levels of estimates obtained (Hartill *et al*., 2016).

Sampling offers a suitable solution to monitoring a part of the whole with generalizable capabilities. It is a means to save time and cost involved in research studies, provides opportunity to study phenomenon with unknown population size and to carry out destructive experiments. There are different sampling protocols for consideration including random sampling schemes, non-random sampling schemes, ranked set sampling and adaptive sampling schemes (Holmes *et al*., 2004; Hartill *et al*., 2016; Thompson *et al*., 2013; Wang *et al*., 2009). The application of traditional sampling designs, such as simple random sampling, systematic sampling and stratified sampling is widespread, because these designs are simple to apply and require minimal *a priori*

information about the population (Holmes *et al*. (2004)). Other sampling schemes that are generally subjective such as purposive and quota sampling are known as non-random. Situations where there is a need to modify the sampling scheme at various stages of survey may require an adaptive sampling scheme. The rank set sampling scheme is non-parametric and may be useful where sample distributional assumption(s) may impede inference.

In fisheries research, many sampling techniques have been investigated (Wang *et al*., 2009; Hartill *et al*., 2016). Wang *et al*. (2009) proposed efficient designs for sampling and subsampling based on ranked sets. Their designs were derived analytically and incorporated highly correlated concomitant variables with variables of interest, such as site selection for a fishery-independent monitoring survey. Ranked set sampling (RSS) for estimating the mean and parameter estimates for simple regression were applied. The relative efficiencies of their designs were compared to the traditional simple random sampling and reported vast improvement in terms of variance and mean squared error. Hartill *et al*. (2016) determined an optimal level of temporal subsampling given a random stratified sampling design using parametric simulations for camera data obtained from monitoring traffic at multiple ramps. The camera data were counts of trailer boats returning daily at three boat ramps over a twelve-month period, which were assigned to respective seasonal/day-type strata. In an iterative simulation scheme for each stratum, an iterative random stratified precision estimator and associated coefficients of variation guided their decision on an optimal subsample size of 60 days per year.

The common tools employed in establishing the optimal size in sampling size planning are power analysis and accuracy in parameter estimation methods. Whereas power analysis is purposely used in hypothesis testing, that is, testing the sample size required for a chosen Type I error rate, the accuracy in parameter estimation method sets a precision level and identifies the required sample that meets the set target (Peterman, 1990; Kelly, 2007). Barrett *et al*. (2017) argued that fisheries studies are geared towards obtaining precise and accurate estimates to guide management decision making processes and in effect the accuracy in parameter estimation is more appropriate. Hartill *et al*. (2016) adopted the accuracy in parameter estimation approach in a related study, where the coefficient of variation was used as a measure of precision. The coefficient of variation and the root mean square error are the most used measures of precision and accuracy respectively in fisheries studies (Yu *et al*., 2012; Hartill *et al*., 2016; Barrett *et al*., 2017).

1.3 Potential drivers of boating activities

Environmental variables can predetermine human behaviour to some degree. According to Soykan *et al.* (2014), environmental variables can inform fishers to maximize their catch and therefore it should be possible to characterize fishing effort and human behaviour as a function of these variables. Environmental, climatic and social factors affect the temporal variability (Maynou and Sardá, 2001; Soykan *et al*., 2014) and can play a key role in survey design and sampling scheme (Steffe *et al*., 2017). For instance, public holidays can greatly influence the variability in sampling boating activities (Desfossess and Beckley, 2015; Steffe *et al*., 2017). The possible correlation between clustered ramps and environmental factors (such as wind speed, rainfall etc.) could also be useful in predicting fishing effort (Soykan *et al.,* 2014). Therefore, an adequate model for effort requires the development of estimation schemes that incorporate a wide range of covariates (van Poorten *et al*., 2015; Hartill *et al.,* 2016). Therefore, the knowledge of the role of the varying climatic, environmental conditions and social events (such as school and public holidays) on boating activities will be useful in the management of recreational fisheries. According to the proposed catchability and effort scheme by Laurec and Le Guen (1981), geographical accessibility of a fishing ground is a major component of catch and effort and is influenced by environmental, climatic and social factors. Therefore, data on boating effort at ramps and other viewpoints, where the effects of these factors have been accounted for, will provide useful accessibility information. However, the effects of these factors are minimally reported in literature.

1.3.1 Climatic variables and the missing data problem

Missing data are common in datasets of climatic variables such as precipitation, temperature, humidity, wind speed, wind gust, and sea level pressure. Notable causes include faulty measuring instruments, routine maintenance and sensor calibration (Yoagatligil *et al*., 2013). Missing observations in climate data are often characterised by occurring consecutively for long periods of time (Simolo *et al.,* 2010). Climate data are typically processed and analysed at low-resolution levels such as daily, weekly, monthly and yearly resolution (Firat *et al.*, 2012, Kanda *et al.*, 2018). It is important to note that the estimation of fundamental statistics such as the means, and covariance is challenging, mostly inaccurate and can be misleading for incomplete data (Schneider, 2001). For instance, the "3/5 omission rule" in the *Guide to Climatological Practices* (3rd edition) stipulates that, when calculating monthly climate normals, any month that is missing more than three consecutive daily values, or more than five daily values in total, should not be included. In a field where missing observations are common, there should be clarity on how low-resolution data are derived from finer resolution data and missing values are handled. The

incomplete state of the data needs to be considered carefully before any meaningful analysis can be carried out. In handling missing data problems, the practice of excluding missing data or censored data from analysis can lead to loss of information, misinterpretation, overestimation or underestimation and introduce bias especially when missingness is not random (Ellington et al, 2015; Maldonado, Aguilera and Salmerón, 2016).

The literature search revealed that several imputation methods have been applied to low-resolution climate data, typically of daily, monthly and yearly resolutions. However, in many instances climatic data at fine resolution are incomplete and this is the case for the data used in this thesis (hourly resolution). Data at lower resolution are commonly based on aggregation from higher resolution data sets. Analysing high-resolution data such as *h*-minutes ($h < 60$) and hourly data, thus would offer greater ability to understand the nature of data variability, behaviours, trends and detection of small changes. In effect, building imputation models to 'fill-in' missing data in high resolution climate data would be a step in the right direction, as there is no significant study on the evaluation of the imputation methods for finer scale missing climate data.

1.4 Study area and general data description

Western Australia (WA) has a population of 2.76 million, with a 1.87% growth rate (Australian Bureau of Statistics, 2020) and an estimated 26% of residents participate in recreational fishing at least once a year (DPRID Annual Report, 2019). WA has a coastal stretch of 12,889 km (Hartill *et al*. 2019). The coastline of WA is divided into four marine bioregions: North, Gascoyne, West and South Coasts (Ryan *et al.*, 2015). In addition to ongoing surveys of boat-based recreational fishing (Ryan *et al*. 2017), digital cameras have been used since 2006 to monitor trends in recreational boating activity at 30 sites along the coast, including boat ramps, channel entrances and parts of the foreshore (Hartill *et al*., 2019). There are 28 cameras monitoring 30 fields of view (see Fig. 1.1).

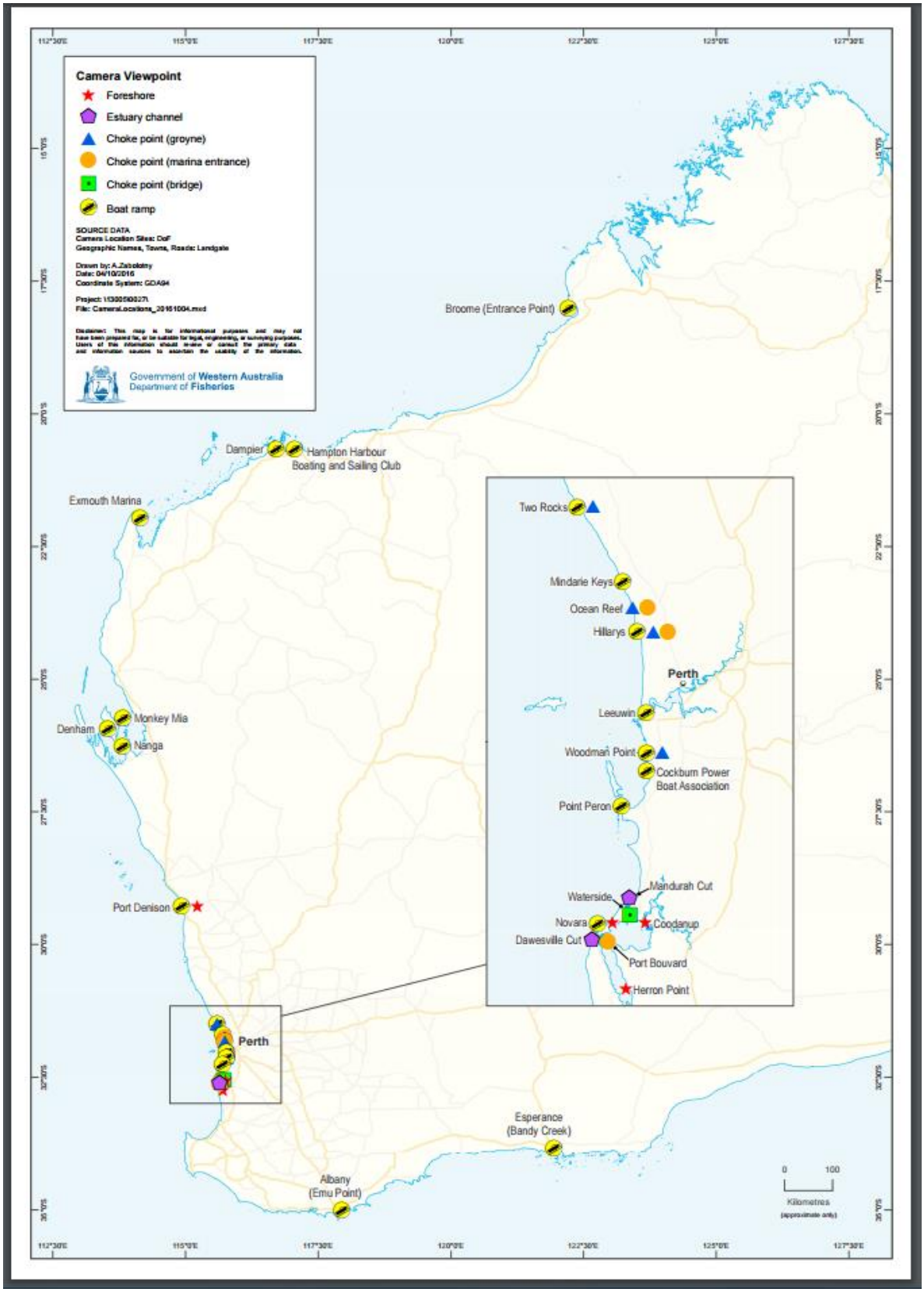

Fig. 1.1: Study area showing the locations of the network of cameras for monitoring boating activities at boat ramps, groynes along sections of the foreshore (Steffe *et al*., 2017)

Recreational boating effort is captured by these cameras operating 24-hours daily. Cameras provide full coverage of traffic activity at viewpoints namely boat ramps, choke-points (estuary channel, marina entrance, groyne, bridge) and foreshore (shallow flats and shore) (Blight and Smallwood, 2015; Hartill *et al.*, 2019). It is important to note that the cameras do not distinguish recreational fishing from other types of recreational activity. Regardless, data obtained from these cameras are an important information source in the estimation of recreational effort and catch estimates for fisheries management purposes. A major drawback in the data is that there are many incidences of outages in the remote camera data (Steffe *et al.*, 2017). In previous work (Ryan *et al*. 2017) short-term camera outages were imputed, however, data for periods of extended outages were not imputed, e.g. an instance of a two months period of outage as mentioned in Ryan *et al*. (2017). Long-term camera outages are common at some locations, due to technical and logistical difficulties or the remoteness of the site (Blight and Smallwood, 2015).

The data for this thesis were obtained from two sources, namely, the Department of Primary Industries and Regional Development (DPIRD) and the Australian Government Bureau of Meteorology. DPIRD provided data of camera monitoring of boat launches and retrievals along the coast of WA. The counts of boat launch and retrieval activities are read for the integrated survey periods between 01 March 2011 to 29 February 2012, from 01 May 2013 to 30 April 2014 and from 01 September 2015 to 31 August 2016. A launch is typically recorded when a boat leaves the shore and a retrieval is recorded when a boat is pulled from the water although these definitions do vary slightly depending on viewpoints at the different ramps (Blight and Smallwood, 2015). Counts of boating traffic (launches and retrievals) for each ramp are recorded to the nearest minute, with time stamps. The type of vessel launched or retrieved is recorded as either commercial, powerboat, jet-ski, kayak and other. In this thesis, analysis focused on data for powerboats, being the most common vessel type used for boat-based recreational activity in WA. In addition, the choice of boat ramps analysed were selected to reflect the vast stretch of the coastline and diverse patterns of traffic intensity. It is reasonable to assume that boat launched are retrieved at the same ramp, since there was high correlation between the number of launches and retrievals at the fields of view. Their distributions across the hours of day were similar (Fig. 1.2). The peak times were between the hours of 0800 - 1000 and 1200 - 1500 for launches and retrievals respectively. In effect, the knowledge of one can help infer the other, therefore the objective of analysis informed the choice of event of boating activity used. For instance, in building imputation models, data on boat retrievals were used, whereas in building forecasting models, data on boat launches were used. Details of the reasoning behind these choices have been expounded in subsequent chapters.

WA has varying environmental features. For instance, according to the Köppen climate classification, the state has ten different climatic zones. From the literature search, the environmental factors considered were air temperature, precipitation, humidity, winds (direction, speed and gust), sea surface pressure. Data were obtained at an hourly resolution for the study duration.

Other temporal variables that are known *a priori* to influence the dynamics of boating effort including months, type of day (categorised as weekday or weekend (include public holidays)), time of day (categorised as dawn, early morning, morning, afternoon, late afternoon and evening) and were also considered as predictors.

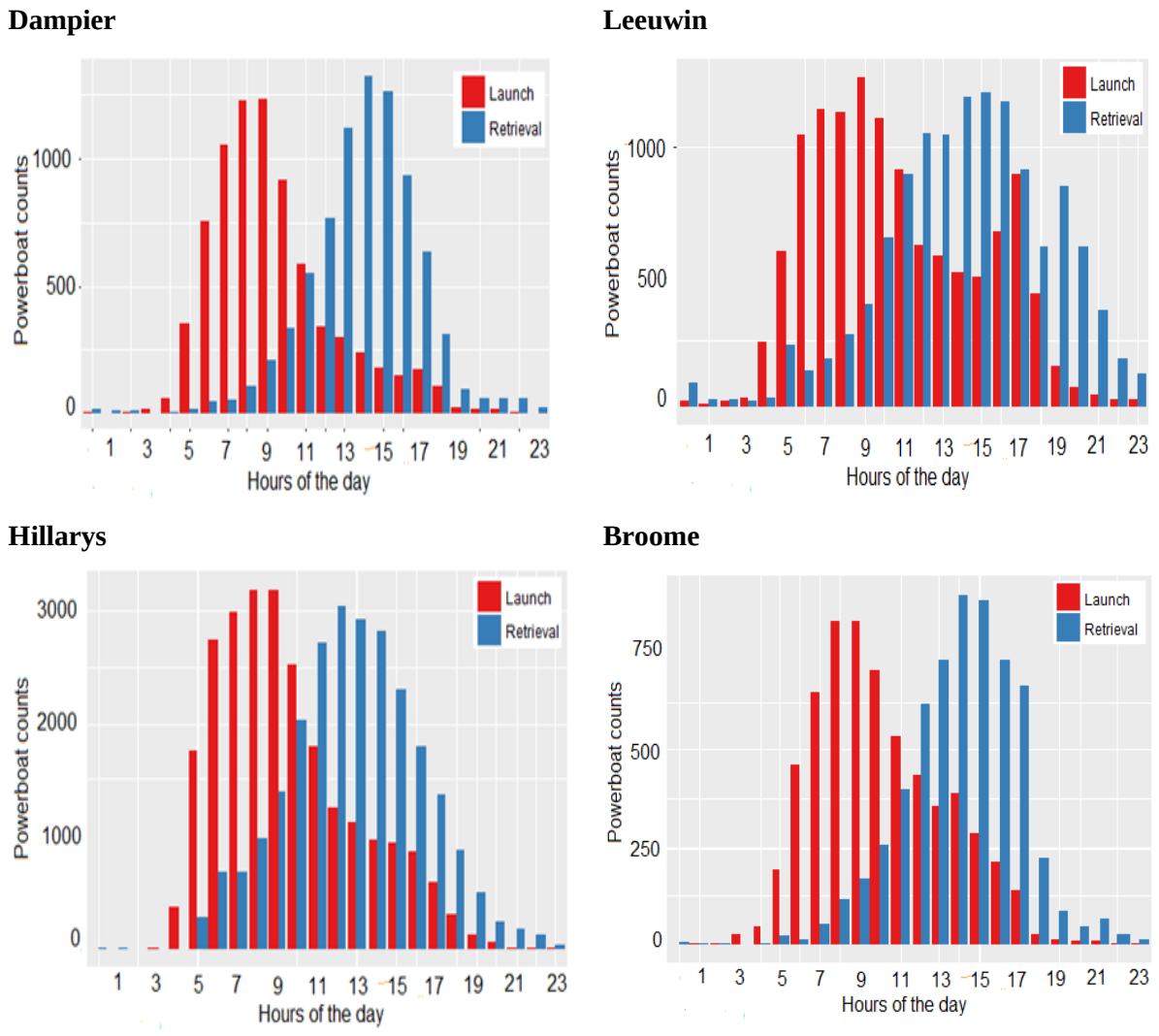

Fig 1.2**:** Distribution of powerboat launches and retrievals at four boat ramps for the integrated survey periods between 01 March 2011 to 29 February 2012 in Western Australia

1.5 Research objectives and questions

This study sought to apply estimation, sampling and modelling techniques to evaluate the trends of boat-based activities from remote camera data using climatic variables as covariates where applicable. Specifically, the objectives (and research questions) were:

1. We investigated imputation methods for finer resolution data to 'fill-in' missing observations in the WA climate data set.

2. We developed and investigated a suitable imputation technique which accounts for environmental conditions to "fill in" missing observations in remote camera data.
   - What are the properties of remote camera failures in WA?
   - What reasonable assumptions can be made for the imputation of remote camera missing data?
   - How can the effect of covariates be incorporated in imputing missing data from digital camera monitoring?

3. We built time series and regression models for predicting the temporal distribution of boat launches and retrievals activities. We further explored the predictive abilities of environmental (climatic and oceanographic) and other factors to describe boating activity distribution in WA recreational fisheries.
   - How well can statistical models predict boating activities in WA?
   - How does the length of the boating data time series affect model performance in WA?

4. We developed sampling schemes for low-level monitoring of remote cameras to meet broad monitoring objectives.
   - What sampling scheme will ensure adequate representative monitoring of boat-based activities in WA?
   - Given that the amount of traffic differs among ramps, what is the optimal sample size (number of days) for monitoring in a remote camera survey for each of the ramps in WA?

1.6 Thesis structure

This thesis has been presented and organised as "Thesis with publication" format[2]; and structured in chapters as follows:

[2] "Thesis with publication" format is an acceptable format of thesis for postgraduate research at ECU policy. The current thesis has been written based on the guideline provided at

Chapter 1 presented the background of this PhD research and brief literature overview. The objectives of the research and the structure of the thesis were discussed.

Chapter 2 addressed objective 1. The contextual literature search on imputation schemes for high-resolution climatic data revealed that the development and evaluation of imputation schemes for such data are at the early phase. There was the need to investigate imputation schemes that would help address the missing data problem in our high-resolution climate data. Multiple approaches to the imputation of missing values were investigated including structural time series models with Kalman smoothing, an ARIMA models with Kalman smoothing and multiple linear regression models. Results for chapter 2 have been published in the journal *Meteorological Application* (**Study I**).

Chapter 3 partly addressed objective 2. Establishing the modelling framework that best describes the relationship between our study response variable (count of boating activities) and covariates (climatic and temporally classified variables) was key in the objective of building imputation models to 'fill-in' missing observations in digital camera monitoring data. There is no clear-cut decision on treating temporally classified variables as fixed or mixed effects, especially if the variable has more levels (Harrison *et al.*, 2018). We compared the accuracy of the fits of treating temporally classified variables, including time of day, and type of day as fixed or mixed effects respectively in a generalized linear modelling setup to impute missing observation in digital camera monitoring data. Results for chapter 3 have been published *in the proceedings of the 34th International Workshop on Statistical Modelling (IWSM) (Volume II), Guimarães, Portugal* (**Study II**).

Chapter 4 partly addressed objective 2. Missing data are common in digital camera monitoring because of camera outages. We have presented a robust imputation technique that incorporated climatic and some temporal classification variables to impute missing data. We compared several generalized linear mixed effect models formulated in the fully-conditional specification multiple imputation framework to impute missing data, with climatic and some temporal classifications as covariates. An article based on the results in chapter 4 has been published in the journal *ICES Journal of Marine Sciences* (**Study III**).

---

http://www.ecu.edu.au/GPPS/policies_db/policies_view.php?rec_id=0000000434. In this format, the submitted thesis can consist of publications that have already been published, are in the process of being published, or a combination of these.

Having established how to impute missing data, we turned our attention to forecasting and distinguished between short-term and long-term forecasts. For the short-term forecasts, data were treated as time series and no covariates were considered, for the long-term forecasting, covariates were used.

The results from Chapters 5 and 6 addressed objective 3. Data generated from recreational boating activity are characterised by infrequent counts, often of variable size and sparse periods of zero counts and such data are difficult to predict. In Chapter 5, we explored the short-term forecasting capabilities of intermittent demand and some count data time series methods for recreational boating effort data observed from digital camera monitoring (**Study IV**). This may guide decisions, such as, filling in short duration gaps, and scheduling routine maintenance of boat ramp.

Survey methods used in recreational fisheries management do not ensure that data collection is continuous, due to budgetary constraints and logistical restrictions. Based on the success achieved in using the study covariates in the imputation search in **Studies II & III**, we explored their potential to describe the temporal distribution of recreational boating effort for longer periods. In Chapter 6 (**Study V**), a Bayesian regression modelling technique was considered as a long-term forecasting tool to formulate predictive models to determine the temporal distribution of boating traffic at two ramps in Western Australia.

As a final topic, the sampling of camera data for estimating boating effort was considered to address objective 4. Manual interpretation of data from digital camera monitoring can be expensive, especially across multiple sites. In improving the utility of digital camera monitoring, the cost of reading data must be managed.

In Chapter 7 an *a posteriori* analysis study design was used to investigate the trade-offs between the reading cost and accuracy measures of estimates of boat retrievals obtained at various sampling proportions for low, moderate and high traffic boat ramps, thereby informing decisions on approaches to be used for future reading of camera data. The article based on the results for chapter 7 has been published in the journal *Fisheries Research* (**Study VI**).

A general discussion of the findings from the various studies has been presented in Chapter 8.

Chapter 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 the proceedings of the 34th International Workshop on Statistical Modelling (IWSM)*, Guimarães, Portugal, 7-12 July.

Afrifa-Yamoah, E., Taylor, S. M. and Fisher, A. J., Mueller, U. (2020). Imputation of missing data from time-lapse cameras used in recreational fishing surveys, *ICES Journal of Marine Science*, 10.1093/icesjms/fsaa180.

Amiri, M. and Jensen, R. (2016). Missing data imputation using fuzzy-rough methods. *Neurocomputing*, 205, 152-164.

Arlinghaus, R. and Cooke, S. J. (2005). Global impact of recreational fisheries. *Science*, 307, 1561-1562.

Arlinghaus, R., Tillner, R., and Bork, M. (2015). Explaining participation rates in recreational fishing across industrialised countries. *Fisheries Management and Ecology*, 22, 45-55.

Ashby, M. P. J. (2017). The Value of CCTV Surveillance Cameras as an Investigative Tool: An Empirical Analysis. *Eur J Crim Policy Res.*, 23, 441–459.

Australian Bureau of Statistics (2020). Australian Demographic Statistics. Retrieved from https://www.abs.gov.au/ on 3/09/20.

Baran, R., Rusc, T. and Fornalski, P. A. (2016). Smart camera for the surveillance of vehicles in intelligent transportation systems. *Multimed Tools Appl*., 75, 10471–10493.

Barrett, B. N., van Poorten, B., Copper, A. B., and Haider, W. (2017). Concurrently assessing survey mode and sample size in off-site angler survey. *North American Journalof Fisheries Management*, 37; 756-767.

Bernal, P. (2016) Data gathering, surveillance and human rights: recasting the debate, *Journal of Cyber Policy*, 1:2, 243-264.

Bian, R. and Hartill, B. (2015). Modelling of recreational fishing effort in QMA 1. *New Zealand Fisheries Assessment Report 2015/26*, Ministry of Primary Industries, Welling-ton, New Zealand, 50p.

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. 121*, Department of Fisheries, Western Australia.

Cayford, M. and Pieters, W. (2018). The effectiveness of surveillance technology: What intelligence officials are saying, *The Information Society*, 34(2), 88-103.

Cooke, S. J. and Cowx, I. G. (2004). The role of recreational fishing in global fish crises. *Bioscience*, 54(9), 857-859.

Deb, R. and Liew, A. W.-C. (2016). Missing value imputation for the analysis of incompletetraffic accident data. *Information Science*, 339, 274-289.

Department of Primary Industries and Regional Development Annual Report (2019). Retrieved on 5.11.2019 from *https://dpird.wa.gov.au/annual-report.*

De Jong, R., van Buuren, S., and Spiess, M. (2016). Multiple Imputation of Predictor Variables Using Generalized Additive Models. *Communications in Statistics-Simulation and Computation*, 45, 968-985.

Desfossess, C. and Beckley, L. E. (2015). Temporal and environmental factors affecting thelaunching of recreational boats at entrance point boat ramp, broome, Western Australia. In Beckley, L. E., editor, *Final Report of Project 2.1.1 of the Kimberley Marine Research Program Node of the Western Australian Marine Science Institution*, Chapter 5, pages 77–91. WAMSI, Perth, Western Australia.

Dornberger, W. (1954). "V-2, Ballantine books," in *ASIN: B000P6L1ES*, pp. 14 – 15.

Ellington, E. H., Bastille-Rousseau, G., Austin, C., Landolt, K. N., Pond, B. A., Rees, E. E., Rober, N., Murray, D. L. (2015). Using multiple imputation to estimate missing data in meta-regression. *Methods in Ecology and Evolution* **6:** 153-163.

Engel, U., Jann, B., Scherpenzeel, A., and Sturgis, P. (2015). *Improving Survey Methods*. Routledge, Taylor & Francis Group, New York.

Firat M, Dikbas F, Cem Koc A, Gungor M. 2012. Analysis of temperature series: estimation of missing data and homogeneity test. *Meteorological Applications* **19**: 397-406.

Hartill, B. (2015). Evaluation of web camera-based monitoring of levels of recreational fishing effort in FMA 1. *New Zealand Fisheries Assessment Report 2015/22*, Ministryof Primary Industries, Wellington, New Zealand.

Hartill, B. W., Payneb, G. W., Rusha, 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, Weltersbach, M. S. (2019). Digital camera monitoring of recreational fishing effort: Applications and challenges. *Fish and Fisheries*, DOI: 10.1111/faf.12413.

Harrison, X. A., Donaldson, L., Correa-Cano, M. E. *et al*. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 1-32.

Holmes, K. W., Van Niel, K., Baxter, K., and Kendrick, G. (2004). Designs for marine remote sampling: a review and discussion of sampling methods, layout, and scaling issues. *CRC for Coastal Zone Estuary and Waterway Management, Technical Report 87*, Project CB3: Benthic Biology and Habitat Mapping Task 2.1 Milestone Report, 37p.

Johnson, A. F., Moreno-Báez, M., Giron-Nava, A., Corominas, J., Erisman, B., E., E., and Aburto-Oropeza, O. (2017). A spatial method to calculate small-scale fisheries effort in data poor scenarios. *PLoS ONE*, 12(4), e0174064.

Junger, W. L. and de Leon, A. P. (2015). Imputation of missing data in time series for airpollutants. *Atmospheric Environment*, 102, 96-104.

Kanda, N., Negi, H. S., Rishi, M. S., Shekhar, M. S. (2018). Performance of various techniques in estimating missing climatological data over snowbound mountainous areas of Karakoram Himalaya. *Meteorological Applications*, 25, 337-349.

Kelly, K. (2007). Sample size planning for the coefficient of variation from the accuracy inparameter estimation approach. *Behavior Research Methods*, 39; 755-766.

Lai, E. K. M., Mueller, U., Hyndes, G. A. and Ryan, K. L. (2019). Comparing estimates of catch and effort for boat-based recreational fishing from aperiodic access-point surveys. *Fisheries Research*, 219, 1-12. https://doi.org/10.1016/j.fishres.2019.06.003

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wileyand Sons Inc., Hoboken, New Jersey.

Laurec, A. and Le Guen, J.-C. (1981). Dynamique des populations marines exploitées. Tome 1. Concepts et modèles. Publications du C.N.E.X.O. Série. *Rapports scientifiques et techniques*, 45; 1-120.

Maldonado, A. D., Aguilera, P. A. and Salmerón, A. (2016). An experimental comparison of methods to handle missing values in environmental datasets. *International Congress on Environmental Modelling and Software* ***3***, https://scholararchive.byu.edu/iemssconference/2016/Stream-C/3

Maynou, F. and Sardá, F. (2001). Influence of environmental factors on commercial trawl catches of Nephrops norvegicus (L). *ICES Journal of Marine Science*, 58; 1318-1325.

Peterman, R. M. (1990). Statistical power analysis can improve fisheries research manage-ment. *Canadian Journal of Fisheries and Aquatic Sciences*, 47(1); 2-15.
Piza, E. L., Welsh, B. C, Farrington, D. P and Thomas, A. L. (2019). CCTV surveillance for crime prevention: A 40-year systematic review with meta-analysis. *Criminology and Public Policy*, 18, 135-159.

Purwar, A. and Singh, S. K. (2015). Hybrid prediction model with missing value imputationfor medical data. *Expert Systems with Applications*, 42, 5621-5631.

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, No. 268*, Department of Primary Industries and Regional Development, 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, No. 287*, Department of Primary Industries and Regional Development, Western Australia.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall, London.

Schneider, T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* **14**, 853-871.

Simolo C, Brunetti M, Maugeri M, Nanni T. 2010. Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *International Journal of Climatology* **30:** 1564-1576

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(6): 1265-1276.

Soykan, C. U., Equchi, T., Kohin, S., and Dewar, H. (2014). Prediction of fishing effort distributions using boosted regression trees. *Ecological Applications,* 24(1), 71-83.

Sovilj, D., Eirola, E., Miche, Y., Björk, K.-M., Nian, R., and Akusok, A. (2016). Extreme learning machine for missing data using multiple imputations. *Neurocomputing*, 174,220-231.

Steffe, A. S., Murphy, J. J., and Reid, D. D. (2008). Supplemented Access Point Sampling Designs: A Cost-Effective Way of Improving the Accuracy and Precision of Fishing Effort and Harvest Estimates Derived from Recreational Fishing Surveys. *North American Journal of Fisheries Management*, 28(4), 1001-1008.

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* No. 286, Department of Primary Industries and Regional Development, WA.

Taylor, S. M., Blight, S. J., Desfosses, C. J., Steffe, A. S., Ryan, K. L., Denham, A. M., & 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(6), 2107–2116.

Taylor, S. M., Smallwood, C. B., Desfosses, C. J., Ryan, K. L., and Jackson, G. (2019). Integrated survey of boat-based recreational fishing in inner Shark Bay 2018/19. *Fisheries Research Report No. 298*. Department of Primary Industries and Regional Development.

Thompson, D. R., Cabrol, N. A., Furlong, M., Hardgrove, C., Low, B. K. H., Moersch, J., and Wettergreen, D. (2013). Adaptive sensing of time series with application to remote exploration. In Robotics and Automation (ICRA), *2013 IEEE International Conference, IEEE*, pp. 3463-3468.

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. and Brydle, S. (2018). Estimating fishing effort from remote traffic counters: Opportunities and challenges. *Fisheries Research*, 204, 231-238.

van Poorten, B. T., Carrutters, 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.

Wang, Y.-G., Ye, Y., and Milton, D. A. (2009). Efficient designs for sampling and sub-sampling in fisheries research based on ranked sets. *ICES Journal of Marine Science*, 66: 928–934.

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 fin-fish fisheries (Final Report for FRDC Project 2005/034 and WAMSI Subproject 4.4.3). *Fisheries Research Report*.

Yozgatligil C, Aslan S, Iyigun C, Batmaz I. 2013. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and Applied Climatology* **112**: 143-167.

Yu, H., Jiao, Y., Su, Z., and Reid, K. (2012). Performance comparison of tradition sampling designs and adaptive sampling designs for fishery-independent surveys: A simulation study. *Fisheries Research*, 113, 173-181.

Zhang, T. (2017). *Exploring the Frontier of Smart Video Surveillance: Novel Domains and Fine-Grain Event Understanding*. Doctoral Thesis, The University of Queensland, Australia.

Chapter 2 is not included in this version of the thesis.

Chapter 2 has been published as:

Afrifa-Yamoah, E., Mueller, UA, Taylor, SM and Fisher, AJ (2020). Missing data imputation of high-resolution temporal climate time series data, *Meteorological Applications*, 27(1): e1873. https://doi.org/10.1002/met.1873

CHAPTER THREE

Fixed versus random effects models: an application in building imputation models for missing data in remote camera surveys[3]

3.1 Abstract

The decision to specify model predictors as fixed or random effects is not always clear cut and at times different interpretations regarding their nature might be possible. This study investigated modelling frameworks for the imputation of missing counts of powerboat retrievals from camera data using climatic and other temporal classification variables as predictors. The temporal classification variables could be treated as fixed or random effects. To evaluate the impact of the treatment of these predictors, patterns of observed outages were applied to a set of complete 12-month hourly camera data. The proportion of missing data ranged from 0.06 to 0.31. A variety of generalized linear and mixed models built on the full-conditional specification multiple imputation framework were formulated to impute the missing values. The models were assessed using the percentage bias, root-mean-square error and skill score. Results from ten replicated multiple imputation schemes showed that the mixed effect models obtained plausible mean estimates of the total number of powerboat retrievals with less variability than those from fixed effect models. A comparison with predictive mean matching was also performed which showed that the popular predictive mean matching performed worse.

Keywords: imputation of count data, generalized linear mixed models, Bayesian sequential regression, Cholesky transformation

[3] A summarized version of the study has been published in volume II of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM). It can be cited as: Afrifa-Yamoah E**,** Mueller UA, Taylor SM, and Fisher AJ (2019). Fixed versus random effects models: an application in building imputation models for missing data in remote camera surveys. *In the proceedings of the 34th International Workshop on Statistical Modelling (IWSM) (Volume II), Guimarães, Portugal, 7-12 July 2019.*

3.2 Introduction

Complex data structures often require more sophisticated statistical modelling techniques. Generalized additive models, Bayesian regression, linear and generalized mixed effect models are among the modelling techniques that could be applied to such data. These models can handle a mixture of variables. The distinction between specifying covariates as fixed or random effect is not always obvious and the multiple definitions in literature add to the dilemma (Gelman and Hill, 2007). The treatment of the temporal variables, however, should be motivated by the goals of the analysis (Gelman and Hill, 2007). For example, if the variable "time of day" is considered as a categorical predictor for the count of recreational boating effort; there are several possible levels that can be chosen for time aggregation and regardless of the choice there is potentially the problem of heterogeneity within the levels. In addition, if the sampling space is uneven across the levels of a temporal variable, then this will be a crucial consideration in deciding on how to treat the variable in model building. The estimation process for the model involving such covariates must have the potential to reduce the probability of false positives (Type I error rates) and false negatives (Type II error rates). Moreover, the process must have the ability to appropriately infer the magnitude of variation within and among clusters or hierarchical levels (Crawley, 2013; Harrison *et al.*, 2018). These modelling considerations must be tailored to the area of application (e.g. dealing with recreational boating effort data obtained from digital camera monitoring).

Digital camera monitoring provides continuous recordings of recreational boating activities; however, interruptions of cameras' operations can lead to significant gaps in the data (referred to as 'outages'). Despite the rapid emergence of camera-based studies relevant to recreational fishing, relatively few studies have examined analytical approaches for dealing with the modelling challenges to address outages in this type of data (van Poorten *et al*., 2015; Hartill *et al.*, 2016). The modelling challenges sought to be addressed include the formulation of models: 1) with the ability to capture the grouping effect of key temporal variables such as season, time of day etc. on the number of powerboat retrievals; 2) to allow the variance-covariance structures to be explicitly modelled, typical for correlated data which characterise the counts of boat retrievals; 3) to sufficiently address any issues relating to over- and under- dispersion of the count data; and 4) to account for any zero inflation in the data.

In Western Australia (WA), digital cameras have been used since 2006 to monitor trends in recreational boating activity at up to 28 sites along the coast, including boat ramps, channel entrances and parts of the foreshore (Steffe *et al.*, 2017), in addition to ongoing surveys of boat-

based and shore-based recreational fishing (Ryan *et al*., 2017). The resulting data have outages and patterns of groupings in the number of counts of powerboat retrievals with respect to key temporally classified predictors, such as the time of the day, day type and, to a larger extent, seasons. For instance, boating traffic is busier in summer than in winter. Likewise, more boat retrievals are observed in the afternoon compared to the early morning. The clustered structure of the response variable would result in correlated observations and violate the independence assumption of ordinary least squares modelling. Additionally, the nature of boating retrieval data requires models with the ability to estimate the variance hierarchically, to ensure that the data generating process adequately estimates between-group variations in means, as well as the variations within groups (Harrison *et al*., 2018). In this study, the treatment of temporal predictors as either random effects and fixed effects was investigated in a generalized linear modelling framework. The model fit was evaluated based on its ability to reconstruct gaps of missing values in data on the number of powerboat retrievals observed at a ramp in WA.

3.3 Method

3.3.1 Data description

Among the digital camera records, one was complete and was used to validate models considered. The complete record consisted of 8,784 entries, and 54.4% of all records were zeros. A total of 12,293 powerboat retrievals was recorded. Four distinct patterns of outages were applied to the complete record. The choice of the 4 outage patterns was based on the percentage of missingness, ranging from 0.06 to 0.31, reflecting both incidents of short and long outages. The longest outage was 80 days (~1,920 hours) and the shortest was one hour. Outage patterns were of variable lengths and uncorrelated among the ramps. The four distinct observed outage patterns applied to a set of complete 12-month hourly camera data of the count of the number of powerboat retrievals were as presented in Figure 3.1.

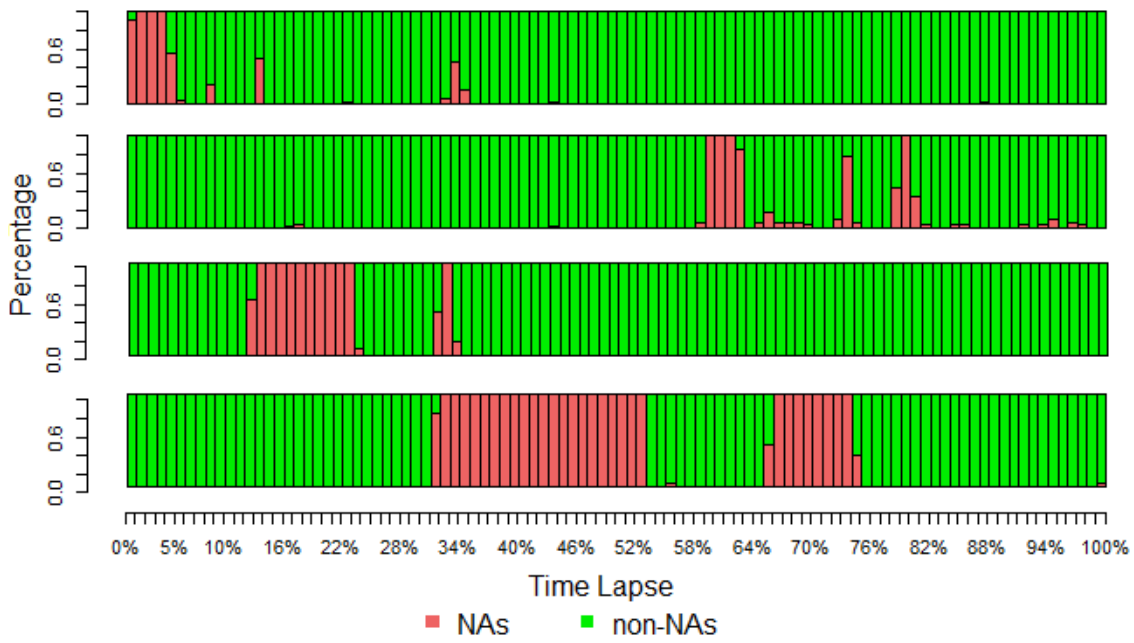

Figure 3.1: Distribution of the outage patterns applied to the Leeuwin dataset. The horizontal axis represents the length of the camera data partitioned into 100. The vertical axis represents the proportion of missing data in the partitioned block or otherwise. The brown bands represent the periods of camera outages and the light green shades represent the observed data.

Hourly data on precipitation, temperature, humidity, wind speed and direction, and sea level air pressure for the Perth Metro station (009105) were obtained from the Australian Bureau of Meteorology. The correlation structure among the study variables is presented in Figure 3.2. The challenge associated with these covariates is that missing observations are inevitable. Advanced time series models with state-space representation amenable to Kalman filter and smoothing algorithms and multiple regression modelling techniques were applied to impute missing observations (Afrifa-Yamoah *et al.*, 2020).

The temporal variables were hours of the day (dawn, early morning, morning, afternoon, late afternoon and evening), the type of day (weekday or weekend/public holidays) and austral seasons (winter, summer, autumn and spring).

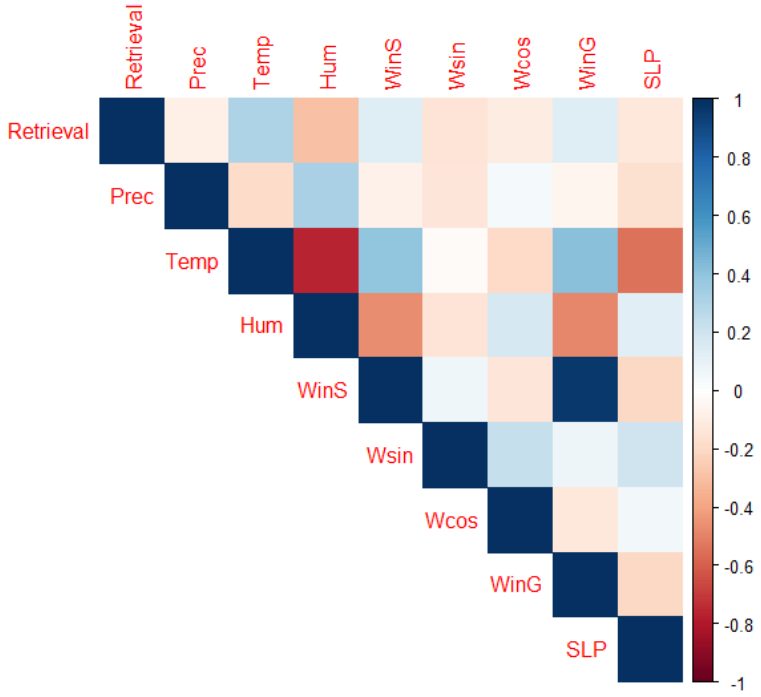

Figure 3.2: Correlation plot depicting the strength and direction of the correlation among the study variables. (Note: Prec = Precipitation, Temp = Temperature, Hum = Humidity, WinS = Wind speed, Wsin & Wcos = sine and cosine transformation of wind direction, WinG = Wind gust and SLP = sea level pressure)

3.3.2 Missing data and assumptions

For a given Y with some missing values, and some covariates, X, the imputation models were formulated to investigate the conditional distribution

$$P(Y^{mis}, \Omega | Y^{obs}, X) \tag{3.1}$$

where Ω represents the vector of unknown model parameters.

Data were assumed to be missing at random (MAR) and generating process for Y was assumed to be a generalized linear model. Two generalized linear model types were considered, the first treated all variables as fixed effects, and the second treated the temporal variables as random effects. The temporal variables were considered as random effects as there were several levels that could have been chosen for time aggregation and to account for potential heterogeneity due to this decision. Quasi-Poisson and zero-inflated Poisson models were considered based on the relationship between mean and variance of the counting process for the camera data.

3.3.3 Modelling framework

Generalized linear models (GLM) are popular modeling extensions for ordinary linear models. In this modelling framework, it is assumed that that the distribution of Y belongs to the

exponential family of distributions. This enables the modelling of real-life scenarios that follow distributions such as Poisson, gamma, binomial and normal (see Dobson and Barnett (2008)).

3.3.3.1 Fixed effect model

The generalized linear model is given by

$$g(\mu) = \boldsymbol{X}\beta, \tag{3.2}$$

where $\mu = \mathbb{E}(Y)$, g is a link function which is monotonic and smooth, $\boldsymbol{X}$ is the model matrix and β is a vector of unknown parameters.

Additionally, the general-purpose predictive mean matching (PMM) was also applied. The approach is generally applicable for the imputation of numeric, non-normal, heteroscedastic residuals and non-linear association between variables (Rubin, 1996; Morris *et al.*, 2014). It has been used to impute missing observations for continuous (van Buuren and Groothuis-Oudshoorn, 2011) and semi-continuous variables (Vink *et al.*, 2014). Little is known about how PMM compares to models that are specifically designed to handle count data. To the best of our knowledge, the technique has received minimal attention in count data imputation problem. PMM was implemented using the *mice* algorithm (van Buuren and Groothius-Oudshoorn, 2011). It uses the ordinary multiple linear regression model to formulate the posterior distribution of the model parameters and imputes missing observations with observed values and thus could preserve the distribution of the observed data (Yu *et al*., 2007).

3.3.3.2 Mixed effect model

The generalized linear mixed model (GLMM) is given by

$$g(\mu) = \boldsymbol{X}\beta + \mathbf{Z}\mathrm{b}, \quad \mathrm{b} \sim N(\mathbf{0}, \psi), \tag{3.3}$$

where b is a random vector containing random effects, with zero expected value and covariance matrix ψ_Ω , with unknown parameters in Ω; $\mathbf{Z}$ is a model matrix for the random effects. (following from the presentation in Wood, 2017).

Random effects are useful when a categorical variable has many levels, uneven sampling across those levels and some observations are correlated (Bolker, 2015). Within the boat retrieval data, there were several levels that could have been chosen for time aggregation and thereby account for potential heterogeneity. Additionally, the sampling across the levels of our temporal variables were uneven, for instance, weekdays were sampled more often than weekends. The estimation of random effects is done with partial pooling, which ensures that a level's effect estimate will be based partially on the more abundant data from the other levels. The nature of the boating retrieval data required a model with the ability to estimate the variance hierarchically, to ensure

that the data generating process adequately estimates the between-group variations in means, as well as, the variations within groups (Harrison *et al*., 2018). In the modelling scheme, random intercept models were fitted, interaction effects were not considered. This was done to moderate the complexity of the model structure because of the large number of predictors involved.

Serial correlation is common among climatic time series variables such as temperature. The presence of serial correlation in the covariates leads to a violation of the assumption of independence among the errors (see Figure 3.3) and will result in the misspecification of the error covariance structure, leading to biased model parameter estimates (Jahng and Wood, 2017).

A Cholesky transformation was applied to the variables (see Jahng and Wood, 2017 for the mathematical details), leading to heterogeneous error covariance structure (see Figure 3.4). Suppose we have a linear relationship given by

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + u_t \tag{3.4}$$

for $t = 1, \cdots, n$ where n is the number of data points and u_t is generated by the Markoff scheme

$$u_t = \alpha u_{t-1} + \varepsilon_t \tag{3.5}$$

with random error ε_t and a known autoregression coefficient α. The resultant equation from substituting equation (3.5) into equation (3.2) is given by

$$y_t' = \beta_0' + \beta_1 x_{1,t}' + \cdots + \beta_k x_{k,t}' + \varepsilon_t \tag{3.6}$$

where $y_t{}' = y_t - \alpha y_{t-1}$, $x_{1,t}' = x_{1,t} - \alpha x_{1,t-1}$, $\cdots$ and $x_{k,t}' = x_{k,t} - \alpha x_{k,t-1}$.

To estimate y_t from given $x_{t_1}, \cdots, x_{t_k}$, equation (3.6) could be improved by

$$y_t = \beta_0{}' + \beta_1 (x_{t_1} - \alpha x_{t_1-1}) + \cdots + \beta_k (x_{t_k} - \alpha x_{t_k-1}) + \alpha y_{t-1} \tag{3.7}$$

where β_0', β_1, $\cdots$, β_k are estimated from equation (3.4).

For each of the climatic variables, the error structure was assessed to determine the autoregressive parameters for the transformation. Then, each of the climate variables was transformed independently via regression-with-autoregressive-error models with the remaining variables as predictors for each time point. The transformed variables were then used to build the imputation model.

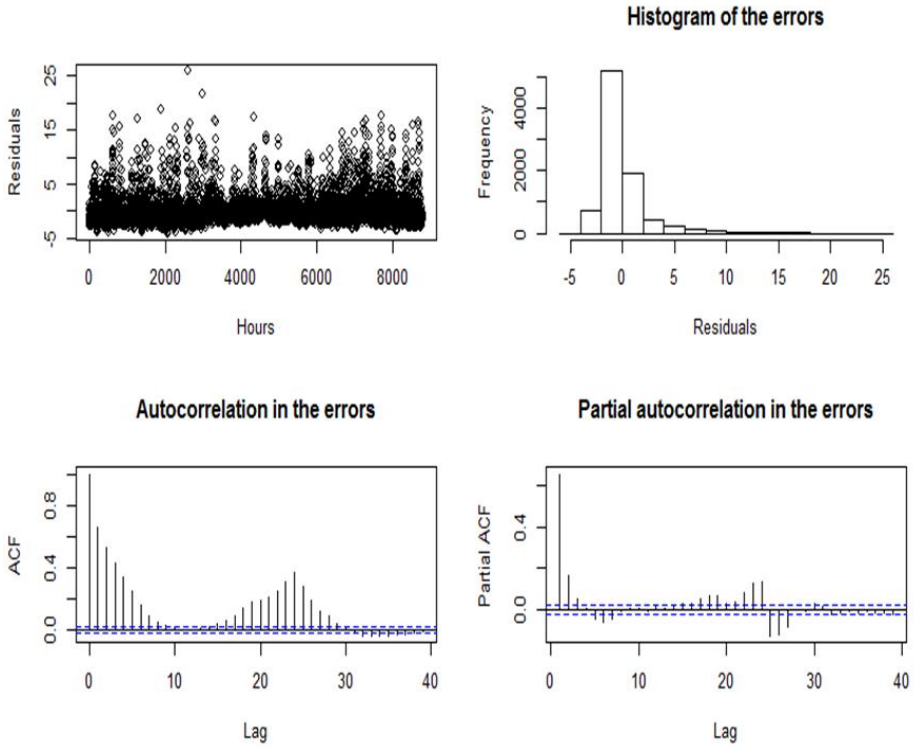

Figure 3.3: Error diagnostics for the Gaussian model output with serially correlated predictors before the Cholesky transformation.

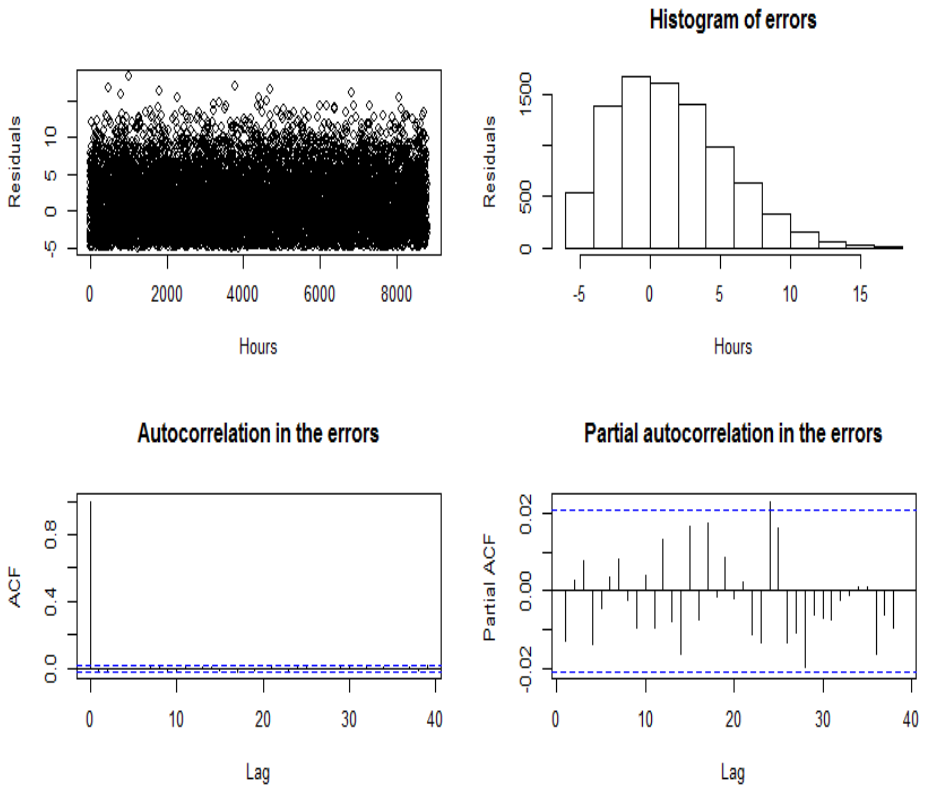

Figure 3.4: Error diagnostics for Gaussian model output with serially correlated predictors after the Cholesky transformation.

3.3.4 Full-conditional specification multiple imputation (FCS-MI)

In FCS-MI (also known as Bayesian sequential regression (van Buuren, 2007), imputed values for missing observations of a variable are randomly drawn conditioned on the observed outcomes and possibly some covariates (van Buuren and Groothuis-Oudshoorn, 2011). These draws are guided through the formulation of models from the distribution that best approximates the association between the variables. For a univariate missing data imputation, the FCS-MI framework involves specifying a conditional model of partially observed variable given some covariates, to obtain a predictive distribution. In this framework, independent draws are generated from the posterior predictive distribution for the missing data. The posterior predictive distribution was obtained by

$$p\left(y^{mis}\middle|y^{obs},X\right)=\int p\left(y^{mis}\middle|y^{obs},X,\Omega\right)p(\Omega|y^{obs},X)d\Omega \tag{3.8}$$

where $\Omega=(\beta,\psi,\sigma)$ is the vector of parameters in equation (3.1) and $p(\theta|y^{obs},X)$ is the observed data posterior density of Ω.

Let $\widehat{\Omega}$ be the vector of parameter estimates with covariance matrix, $\widehat{\mathrm{Var}}\left(\widehat{\Omega}\right)$, where Ω represents the vector of parameters obtained from the fitted model. For a missing observation (Y^{mis}), the law of iterated expectation was used to find a consistent estimator of Ω by solving

$$E_{f\left(Y^{mis}\middle|Y^{obs},\boldsymbol{X}\right)}[u\left(Y^{obs},Y^{mis},\boldsymbol{X},\Omega\right)]=0 \tag{3.9}$$

where $f(.)$ is the conditional predictive distribution of the missing data obtained from the fitted model and $u(.)$ is the score function, which is the gradient of the log-likelihood function, $\ell(\Omega|Y,X)$. In this scheme, missing observations were imputed with values sampled from the predictive distribution of the observed data. The between-imputation variability was introduced using a regression-type approach of fitting specified models to different samples for each of the *M* imputations, where *M* is the number of multiple imputations (see Klienke and Reinecke, 2013). The scheme repeatedly draws estimates of Y^{mis} from Y^{obs} based on $f\left(Y^{mis}\middle|Y^{obs},\boldsymbol{X}\right)$ and then combines the results for inference (Salfrán, 2018). The multiple imputation scheme accounts for the uncertainty in the missing data, since the exact true values cannot be determined (Rubin, 1987; Sterne *et al.*, 2009; van Buuren and Groothius-Oudshoorn, 2011; Klienke and Reinecke, 2013). The imputation algorithms for the models are presented in Table 3.1.

In the imputation process, the observed data and covariates were used to fit a model, to obtain $\widehat{\Omega}$ and $\widehat{Var}(\widehat{\Omega})$. For each missing datum, chains of equations were formulated with parameter estimates drawn from the $\Omega\sim N(\widehat{\Omega},\widehat{Var}(\widehat{\Omega}))$. Predicted values were obtained and corresponding

observed datapoints with the closest predicted values to that of the missing observation were sampled. The imputed values were subsequently drawn from the observed data. The process was repeated $M = 5$ times, so that for each imputed value $\hat{Y}^{mis}$ there were M replicates $\hat{Y}_{i,m}^{mis}$ for $m = 1, \cdots, M$. For a missing observation, the combined imputed estimates of the mean $\overline{\hat{Y}_i^{mis}}$ and variance ($\widehat{Var}(\overline{\hat{Y}_i^{mis}})$) were obtained as

$$\overline{\hat{Y}_i^{mis}} = \sum_{m=1}^{M} \hat{Y}_{i,m}^{mis} \tag{3.10}$$

$$\widehat{Var}\left(\overline{\hat{Y}_i^{mis}}\right) = \frac{\sum_{m=1}^{M} \widehat{Var}(\hat{Y}_{i,m}^{mis})}{M} + \frac{M+1}{M(M-1)} \sum_{m=1}^{M} \left(\hat{Y}_{i,m}^{mis} - \overline{\hat{Y}_i^{mis}}\right)^2 \tag{3.11}$$

where $\sum_{m=1}^{M}\left(\hat{Y}_{i,m}^{mis} - \overline{\hat{Y}_i^{mis}}\right)^2$ reflects the missing values estimation uncertainties (Rubin, 1987). The multiple imputation scheme was repeated ten times for each model to establish consistency or otherwise of the missing value estimates obtained from the models. All imputation modelling approaches were carried out using *mice* (van Buuren and Groothius-Oudshoorn, 2011) and *countimp* (Klienke and Reinecke, 2013) packages in R (R Core Team, 2016).

3.3.5 Model evaluation

The estimation accuracy of the imputed values was assessed via the percent bias, mean absolute error (MAE), root mean square error (RMSE) and skill score (SS) based on the mean square error

$$\%Bias = 100 \times \frac{\sum_{i=1}^{n}(\overline{\hat{Y}_i^{mis}} - Y_i)}{\sum_{i=1}^{n} Y_i} \tag{3.12}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{\hat{Y}_i^{mis}})^2}{n}} \tag{3.13}$$

$$SS = 1 - \frac{MSE(Y, \overline{\hat{Y}^{mis}})}{MSE(\overline{Y}, \overline{\hat{Y}^{mis}})} \tag{3.14}$$

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

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-magnitude values indicating plausible imputed values. RMSE is widely reported imputation modelling performance indicators. For RMSE, the range is 0 to +∞, and lower values indicate high levels of agreement between observed and estimated values and have the same units as the variables measured. The

skill score measures the accuracy of a forecast relative to standard reference. The values of SS range between $-\infty$ and 1. A perfect forecast is observed when a score of 1 is obtained.

Table 3.1: A) Multiple imputation scheme for the Quasi-Poisson B) Multiple imputation scheme for the Zero-inflated Poisson models.

Algorithm A
1. Obtain estimates of $\boldsymbol{\Omega}, \widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{Var}}(\widehat{\boldsymbol{\Omega}})$ from the fitted model
2. For $\boldsymbol{I_i = 0}$, draw $\boldsymbol{\Omega}^*$ from $\boldsymbol{N}(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{Var}}(\widehat{\boldsymbol{\Omega}})$
3. Formulate chained equations $\boldsymbol{f}(\boldsymbol{Y}\|\boldsymbol{X}, \boldsymbol{\Omega}^*)$
4. Compute $\widehat{\boldsymbol{Y}}$ from the chained equations
5. Randomly draw one of the $\boldsymbol{Y^{obs}}$ with $\widehat{\boldsymbol{Y}}$ closet to that of $\boldsymbol{Y^{mis}}$
6. Repeat steps 2-5 $\boldsymbol{M}$ times
Algorithm B
1. Obtain estimates of $\boldsymbol{\Omega} = \{\boldsymbol{\Omega_z}, \boldsymbol{\Omega_c}\}$, $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{z}}$ and $\widehat{\boldsymbol{Var}}(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{z}})$ from the zero model, and $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{c}}$ and $\widehat{\boldsymbol{Var}}(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{c}})$ from the count model.
2. For $\boldsymbol{I_i = 0}$, draw $\boldsymbol{\Omega}^*_{\mathbf{z}}$ from $\boldsymbol{N}\left(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{z}}, \widehat{\mathbf{Var}}\left(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{z}}\right)\right)$.
3. From $\boldsymbol{\Omega}^*_{\boldsymbol{z}}$ compute predicted probabilities for having a zero vrs non-zero count.
4. $\boldsymbol{Y^{mis}}$ for the zero part are imputed with zeros, remembering cases for the non-zero part.
5. For $\boldsymbol{Y^{mis}}$ for the count part, draw $\boldsymbol{\Omega}^*_{\mathbf{c}}$ from $\boldsymbol{N}\left(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{c}}, \widehat{\mathbf{Var}}\left(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{c}}\right)\right)$.
6. Formulate chained equations $\boldsymbol{f}(\boldsymbol{Y}\|\boldsymbol{X}, \boldsymbol{\Omega}^*_{\mathbf{c}})$
7. Compute $\widehat{\boldsymbol{Y}}$ from the chained equations
8. Randomly draw one of the $\boldsymbol{Y^{obs}}$ with $\widehat{\boldsymbol{Y}}$ closet to that of $\boldsymbol{Y^{mis}}$
9. Repeat steps 2-8 $\boldsymbol{M}$ times

3.4 Results

The percentage of zero counts in the dataset with simulated missing data scenarios ranged from 35.1% to 51.8% and missing proportion of missing observations were between 0.06 and 0.31 (Table 3.2). The 95% confidence interval of the average total imputed estimates obtained from the five models contained the actual totals in most cases. However, for outage pattern 1, all models underestimated the observed number of powerboat retrievals. In terms of percent bias, models were ranked differently with four different models ranked as the best for the 4 outage patterns and PMM was ranked worst each time. The direction of the estimation of the bias also varied among the outage patterns. For example, the bias was negative for all the models for outage pattern 1, indicating underestimation of the total counts, but for outage pattern 4, four of the models recorded positive bias, with overestimated total counts.

In terms RMSE, the zero-inflated models were ranked the best apart from outage 3 (Table 3.2). The percentage differences in RMSE values between the two best models ranged from 1.53% to 3.44%. In terms of SS, the zero-inflated models were ranked the best, with the fixed models often ranked as the best (Figure 3.5). The percentage difference in the SS values between the two best models (models with larger SS scores) for the ten outage patterns ranged from 0.06% to 10.3%, with the magnitude of errors between 0.01 and 0.03. 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. PMM typically showed the worst performance, with comparatively large magnitude of bias over- and under-estimation, because it generally fits the ordinary linear regression model in the parameter estimation process. From the ten replications, the random effect models provided more consistent estimates of the total powerboat retrievals, evidence the narrower confidence intervals compared to the fixed models.

Table 3.2: Models' performance evaluation. The table displays the characteristics of the missing patterns including the minimum (min) and maximum (max) duration of outages, the average total boat counts imputed from the fitted models versus total observed counts with associated standard deviations, and the average performance indicators across the ten imputation runs.

Outage	Missing prop	Min	Max	Number of zeros	Model	Average Total Estimate (SD)	% Bias (SD)	SS	RMSE
Outage 1	0.06	1	393	4549 (51.8%)	Observed	746			
					PMM	**734 (22.5)**	-19.8 (17.6)	-0.13	3.25
					QP.fixed	795 (20.7)	23.1 (9.2)	-0.09	2.89
					ZIP.fixed	778 (15.6)	15.1 (8.7)	0.18	2.54
					QP.mixed	800 (20.5)	25.5 (9.3)	0.02	2.86
					ZIP.mixed	766 (14.0)	**14.9 (7.8)**	**0.21**	**2.51**
Outage 2	0.08	1	345	4434 (50.5%)	Observed	819			
					PMM	739 (22.5)	-34 (5.9)	0.14	3.24
					QP.fixed	782 (15.7)	-12.0 (3.3)	0.19	2.86
					ZIP.fixed	**795 (13.2)**	**-7.7 (4.0)**	**0.26**	**2.45**
					QP.mixed	768 (10.6)	-15.3 (3.2)	0.20	2.95
					ZIP.mixed	787 (7.8)	-8.9 (2.6)	0.25	2.72
Outage 3	0.12	144	940	4243 (48.3%)	Observed	1642			
					PMM	1783 (22.9)	17.1 (3.2)	0.09	5.38
					QP.fixed	1627 (26.6)	**-6.1 (5.6)**	0.15	3.45
					ZIP.fixed	1671 (13.5)	6.7 (1.5)	**0.30**	**2.08**
					QP.mixed	1634 (10.9)	-6.5 (4.2)	0.16	3.29
					ZIP.mixed	**1638 (19.0)**	-8.3 (7.7)	0.27	2.24
Outage 4	0.31	3	1920	3080 (35.1%)	Observed	2705			
					PMM	2956 (66.2)	18.7 (2.5)	0.08	4.31
					QP.fixed	2741 (26.4)	1.1 (1.9)	0.20	2.63
					ZIP.fixed	2693 (31.4)	-6.8 (5.2)	**0.23**	2.74
					QP.mixed	2725 (21.7)	**0.7 (3.9**)	0.21	**2.59**
					ZIP.mixed	**2694 (17.0)**	-1.9 (4.7)	0.22	2.67

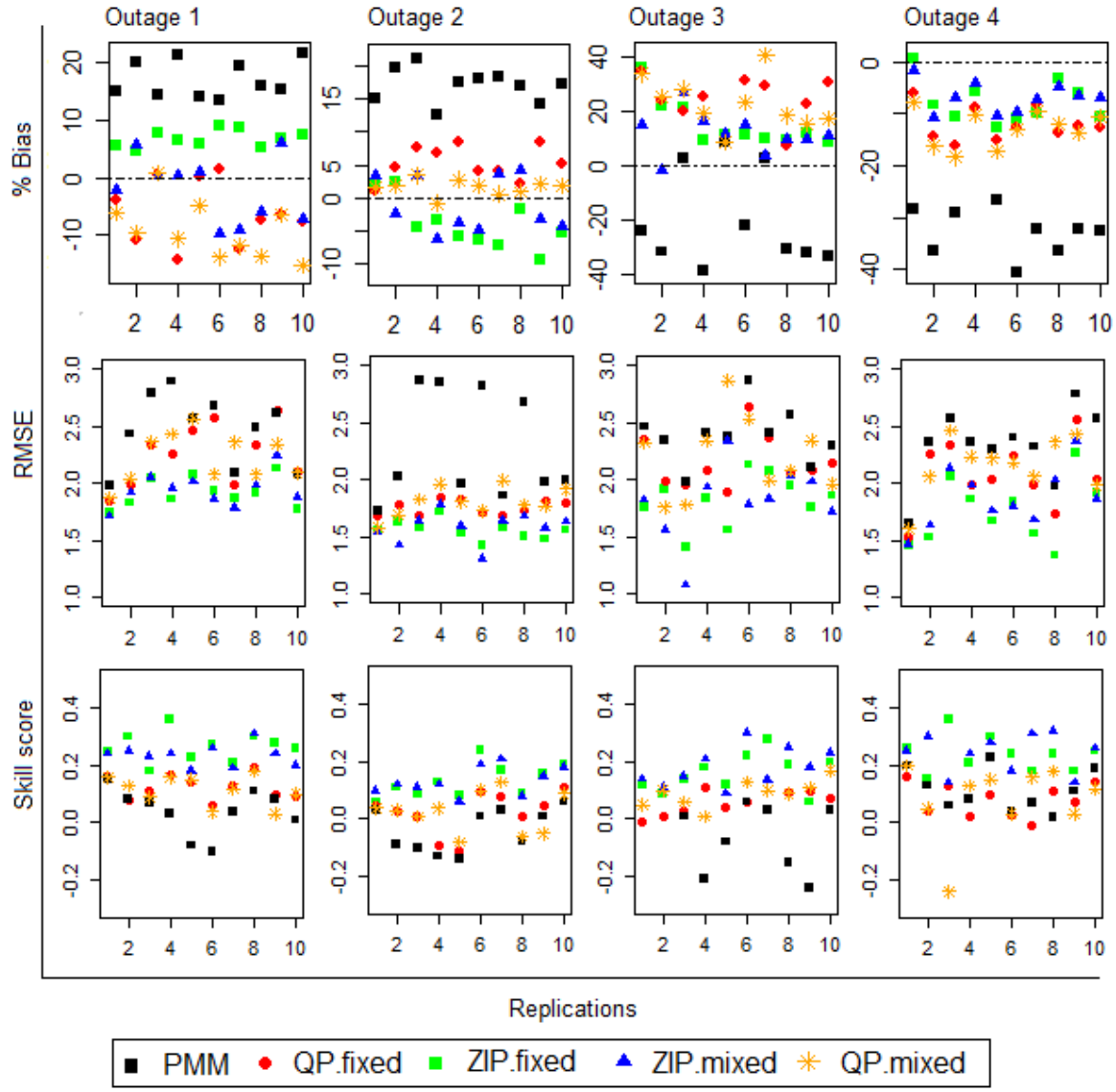

Figure 3.5: Model performance based on the percent bias, RMSE and skill scores for the ten replicates of the multiple imputation scheme. Lowest and highest values of RMSE and skill scores respectively indicate best models.

3.5 Discussion

In this study, although the groupings of the temporal predictors were collectively exhaustive, treating them as random effects resulted in relatively more stable outcomes compared to treating them as fixed effects. Controlling for non-independence within the levels of temporal variables improved the accuracy of the parameter estimation process. In the random effect models, we fitted only random intercepts which allowed only the group means to vary, for simplicity and fast convergence of the imputation scheme. Although, treating random slope to clustered data to help control Type I errors (Aarts *et al.*, 2015), the mixed models were generally provided more

consistent estimates of the total powerboat retrievals with narrower confidence intervals than their fixed effect counterparts. The difference between the model choices was more apparent in terms of the variability of the estimates around the mean imputed totals from the ten replications. The mixed effect models, notably the zero-inflated mixed models, were found to report the least variability, implying that more stable estimates were obtained.

The covariates used in this study are indirectly related to boat retrievals and the imputation modelling scheme required enough data points to train the models to recognise the general patterns to impute plausible values for the missing data, which was made possible by the level of data resolution. The decision to make the imputation models more dependent on the observed data was grounded in the assumption that data were missing at random, with the premise that some information about the missingness in the data could be inferred from the observed data and some covariates. The significance of climatic and temporal strata such as time of day, type of day and season has been established for recreational and commercial fishing activity (Desfosses and Beckley, 2015; Maynou and Sardá, 2001; Soykan *et al*., 2014). In recreational fishing surveys, these temporal strata are often used for sampling scheme development, and in some instances for monitoring and evaluation purposes (Ryan *et al*., 2017; Taylor *et al.*, 2018). These temporal strata to a large extent control the clustering effects of boating activities over time. The choice of the mixed effect modelling approach enabled the estimation process to be dependent on the groupings within these temporal variables. This ensured that the within-group variations were adequately captured to obtain estimates that were representative of the group.

If there is weak association between outcome variable and predictors and substantial missingness in the outcome variable (see Figure 3.1), the full-conditional specification multiple imputation (FCS-MI) has been found to be more robust to model misspecification than the joint model multiple imputation in restricted general location modelling settings (Seaman and Hughes, 2018). Also, formulating joint models in multilevel setting may mathematically be unachievable or could require high level computational skills. Resche-Rigon and White (2018) illustrated the mathematical difficulty in formulating a simple joint model in specifying conditional models in the multilevel setting. Within the FCS-MI framework, this complexity can be easily dealt with by fitting a mixed effect model.

In Hartill *et al*. (2016), GLM were applied to impute the number of trailer boats that were retrieved when camera outage was experienced. GLM generally fails to reflect possible groupings in outcome variables in its estimation process (Faraway, 2010), and would perform

poorly in capturing the clustering effects within the boating activity data, especially for finer-scale datasets. The application of GLMMs in modelling data in ecological studied has been reviewed by Bolker *et al.* (2008), in which the challenges in the estimation and inferential procedures and the opportunities have been outlined. This modelling approach is more applicable to non-normal data involving random effects. The approach allows flexibility in specifying the desired distribution, the appropriate link function and the structure of the random effect. Lancaster *et al.* (2017) used a GLMM to study the significance of ecological and geographical variables, including rugosity, bottom type, depth and the presence or otherwise of bullkelp bioband (*Nereocystis luetkeana*) in a location to predict shore-based recreational fishing effort using counts of shore-based fishers observed from digital cameras as response variable. Shore-based monitoring of recreational fishing effort generally involves a relatively smaller area and it is possible to obtain consistent measurements of the covariates used. In the case of boat-based surveys, consistent measurements for the variables considered in Lancaster *et al.* (2017) may not be feasible and different covariates would have to be considered. The current study has the potential for predicting recreational activity using camera data as the response, and climatic and temporally classified covariates in both shore- and boat-based surveys. Comparatively, on the cost of obtaining data on predictors of recreational boating effort, the climatic and temporal variables would be a huge cost-saving option.

Chapter Acknowledgments

This study was funded and supported by the Government of Western Australia Department of Primary Industries and Regional Development (DPIRD). The authors acknowledge the Australian Government Bureau of Meteorology for providing the climate data.

Chapter References

Aarts, E., Dolan, C. V., Verhage, M. and Sluis, S. (2015). Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neuroscience*, 16:94, 1-15.

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*, 1-18. https://doi.org/10.1002/met.1873

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. and White, J-S. S. (2008). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, 24 (3), 127 -135.

Bolker, B. M. (2015). Linear and generalized linear mixed models. In Fox, Negrete-Yankelevich and Sosa (Eds.), *Ecological Statistics: Contemporary Theory and Application.* Oxford University Press, United Kingdom.

Crawley, M. (2013). *The R Book* (Second Edition). Chichester: Wiley.

Desfosses, C., and Beckley, L. E. (2015). Temporal and environmental factors affecting the launching of recreational boats at entrance point boat ramp, Broome, Western Australia. In Beckley, L. E. (Ed), *Final Report of Project 2.1.1 of the Kimberley Marine Research Program Node of the Western Australian Marine Science Institution,* Chapter 5, 77-91.

Dobson, A. J. and Barnett, A. (2008). *An Introduction to Generalized Linear Models.* CPC Press.

Faraway J. (2010). Generalized Linear Models. *International Encyclopedia of Education*: 178-183.

Gelman A. and Hill J. (2007). Data Analysis using Regression and Hierarchical/Multilevel Models. New York: Cambridge University Press.

Harrison, X. A., Donaldson, L., Correa-Cano, M. E. *et al*. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 1-32.

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.

Jahng, S., and Wood, P. K. (2017). Multilevel models for intensive longitudinal data with heterogeneous autoregressive errors: the effect of misspecification and correction with Cholesky transformation. *Frontiers in Psychology*, 8:262, 1-15.

Klienke, K. and Reincke, J. (2013). Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica*, 67(3): 311-336.

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 Conserv. Mar Freshw Ecosyst*. 27: 804-813.

Maynou, F. and Sardá, F. (2001). Influence of environmental factors on commercial trawl catches of *Nephrops norvegicus* (L). *ICES Journal of Marine Science*, 58: 1318-1325.

Morris, T. P., White, I. R., and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(75), 1-13.

R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Resche-Rigon, M. and White, I. R. (2018). Multiple imputation by chained equation for systematically and sporadically missing multilevel data, *Statistical Methods in Medical Research*, 27(6), 1634-1649.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434); 473-489.

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. (2017). State-wide survey of boat-based recreational fishing in Western Australia 2015/16. *Fisheries Research Report, No. 287,* Department of Primary Industries and Regional Development, Western Australia.

Salfrán, D. V. (2018). *Multiple imputation for complex data sets*. Doctoral Dissertation. Universität Hamburg, Hamburg, Germany.

Seaman, S. R. and Hughes, R. A. (2018). Relative efficiency of joint-model and full-conditional-specification multiple imputation when conditional models are compatible: the general location model. *Statistical Methods in Medical Research*, 27(6), 1603-1614.

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* No. 286, Department of Primary Industries and Regional Development, WA.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009; 338 (b: 2393):157–160.

Soykan, C. U., Equchi, T., Kohin, S., and Dewar, H. (2014). Prediction of fishing effort distributions using boosted regression trees. *Ecological Applications,* 24(1), 71-83.

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*, doi:10.1093/icesjms/fsy066.

van Buuren S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res.*; 16 (3):219–42.

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., Carrutters, 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.

Vink, G., Frank, L. E., Pannekoek, J., and van Buuren, S. (2014). Predictive mean matching imputation of semi-continuous variables. *Statistica Neerlandica*, 68 (1): 61-90.

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 (3): 233-243.

Chapter 4 is not included in this version of the thesis.

Chapter 4 has been published as:

Afrifa-Yamoah E, Taylor SM, Fisher, AJ and Mueller UA. (2020). Imputation of missing data from time-lapse cameras used in recreational fishing surveys. *ICES Journal of Marine Science*, 77(7-8), 2984-2994 https://doi.org/10.1093/icesjms/fsaa180.

CHAPTER FIVE

Short term prediction of recreational boating effort: Evaluation of intermittent demand and count data forecasting methods [4]

5.1 Abstract

Aspects of recreational fisheries management rely on the analysis of count data of boating activity obtained from digital camera monitoring. Data are often highly variable, and are characterized by sparse periods of zero counts, dominated by seasonality which result in a repetitive cycle making modelling a challenge and forecasting difficult. In this study, five forecasting methods were evaluated and accuracies of their point estimates of forecasts for lead times of 12, 24, 48 and 168 hours were assessed using cross-validation techniques. Specifically, intermittent demand forecasting techniques, including Croston's method and Syntetos-Boylan Approximation (SBA) models, and count data forecasting methods including autoregressive conditional Poisson (ACP) models, integer-valued moving average (INMA) models, and integer-valued autoregressive (INAR) models were evaluated. Digital camera monitoring data of hourly counts of powerboat launches at a boat ramp in Western Australia were used. The length of this time series was one year. ACP and INAR models performed better than intermittent demand forecasting techniques for short forecast horizons and provided some evidence of their sufficiency in forecasting the dynamics in recreational boating activities. This result established that, in as much as intermittency may be a key feature for a given dataset, it should not override the systemic characteristics of data in the application of forecasting techniques. Our results provide plausible estimates for short-term outages in such data and promote pragmatic management decision-making.

Keywords: digital camera monitoring, time series modelling, intermittent data, integer-valued models

[4] This chapter will be submitted to *Journal of Time Series Analysis* for publication. The full text has been removed from this version of the thesis.

CHAPTER SIX

Modeling climatic and temporal influences on powerboat launches with relevance to recreational fisheries [5]

6.1 Abstract

Digital camera monitoring data on recreational boating activity are often manually interpreted and the reading cost can be expensive for multiple sites. Typically, this scheme is used along with other periodic boat-based surveys and it is common practice that camera data between survey periods are not read, creating significant gaps in the time series. We predicted boating behaviour during these periods of non-observation using historical data and secondary variables to complete the time series data. Predictive models, built in a Bayesian regression modelling framework, were formulated to determine the temporal distribution of daily boating traffic at two ramps in Western Australia based on climatic variables (including temperature, humidity, wind speed and gust, sea level pressure and wind direction) and temporal classifications (including months, and day type). Two observed year-long datasets from digital camera monitoring of powerboat launches were used, with a yearlong gap between them. One set was used to build models, and the other set was used for validation purposes. Models were cross-validated using leave-one-out sample, ensemble prediction and reconstruction of observed datasets. Fitted models explained 50% [95% CI of R^2: 0.40 – 0.58] and 62% [95% CI of R^2: 0.58 – 0.66] of the variabilities in the daily number of powerboat launches at the two locations, respectively. Subsequently, using the data for the preceding period where camera data were read, we constructed plausible data for the period between the readings. Constructed and reconstructed data generally aligned well with the observed data, with some temporal biases at the bulk and upper tail of the distributions. The 95% credible intervals of the reconstructed periods adequately captured the observed data at both locations. Data for the constructed periods depicted the general trends for the observed periods. Our results provide useful insights into using environmental factors to predict boating activity to 'fill in the gaps' between survey years. This could assist in the ongoing monitoring and sustainable management of recreational fisheries.

Keywords: temporal analysis, digital camera monitoring data, distributional regression, Bayesian regression modelling, recreational fisheries management

[5] This chapter will be submitted to *Fisheries Research* for publication. The full text has been removed from this version of the thesis.

Chapter 7 is not included in this version of the thesis.

Chapter 7 has been published as:

Afrifa-Yamoah E, Taylor SM, and Mueller U. (2021). Trade-off assessments between reading cost and accuracy measures for digital camera monitoring of recreational boating effort, *Fisheries Research*, 233, 105757. https://doi.org/10.1016/j.fishres.2020.105757.

CHAPTER EIGHT

General Discussion

8.1 Discussion

The wealth of information obtained from digital camera monitoring has established its wide application in different fields of study. The amount of data generated from digital camera monitoring is enormous and statistical knowledge is required to fully unravel patterns, trends and derive meaningful summaries that translate raw data into problem solving tools. This study has provided relevant statistical support to improve the utility of digital camera monitoring of boating effort. Specifically, methods have been developed and investigated for dealing with missing observations in high-resolution climate data (**Study I**) and digital camera monitoring data (**Study II & III**), modelling and describing the temporal distribution of recreational boating activity (**Study IV & V**) and designing an appropriate low level and cost-saving monitoring scheme for digital camera usage in recreational fisheries research (**Study VI**) . The main findings of this thesis are: 1) climatic and temporal variables are useful predictors for describing the distribution of recreational boating effort and are suitable for model building, 2) structural time series models with Kalman smoothing, ARIMA models with Kalman smoothing and multiple linear regression are potentially useful methods for imputing missing observations in high-resolution climate data, 3) generalized linear mixed models built on the full conditional specification multiple imputation (FCS-MI) framework are suitable for imputing missing observations in digital camera monitoring data, 4) autoregressive conditional Poisson (ACP) models and integer-valued autoregressive (INAR) models were appropriate for short horizons forecasting of count data that are highly variable, 5) manual interpretation of camera footage for 40% of the days within a year can be deemed as an adequate level of sampling effort to obtain unbiased, precise and accurate estimates to meet broad management objectives.

In digital camera monitoring outages are expected and, in many instances, some form of imputation will be required. An aspect of this thesis investigated missing observations in data generated from monitoring of recreational boating effort. Recreational fisheries studies that have used digital camera monitoring have dealt with missing observations by borrowing information from cameras in proximate locations (Hartill *et al.*, 2016; van Poorten *et al.*, 2015) or by using methods that were of limited use for imputing long outages (Ryan *et al.*, 2013, 2015, 2017). However, long outages are common in such data, especially in remote locations and so an imputation scheme was required that would deal with imputing plausible values for varied durations of outages (**Study II & III**) In the proposed scheme, climatic and some

temporal variables were used to describe the distribution of boating effort to guide the imputation process. Climatic variables including temperature, precipitation, humidity, wind speed and gust, and wind direction are commonly observed meteorological measurements. The climatic data set had missing observations so it was first necessary to impute them before they were subsequently used to model the camera data. **Study I** investigated and proposed suitable imputation methods for addressing relatively short duration missing observations in high-resolution temperature, humidity and wind speed data. Measurements on climatic variables are universally available for use, thus promoting the practicality of the proposed technique elsewhere, albeit with contextual variations.

Some considerations were made in using these covariates in the imputation model building for the digital camera monitoring data. It is typical in scientific studies to focus on the relative importance of predictors within statistical models. However, it is important to note that the covariates used were not directly associated with the missing mechanism and did not explicitly give any information on why the camera records were missing. Thus, the focus was on assessing the predictive information on boating effort that the covariates collectively contributed. Part of the objective was to establish the means to effectively combine these covariates to extract the signals from the noises in the digital camera monitoring data. The climatic variables were treated as fixed effects in the model building. However, the treatment of the temporal variables could be motivated by the goals of the analysis (Gelman and Hill, 2007). For example, time of day was considered as a categorical variable, and there were several levels that could have be chosen for time aggregation and regardless of the choice there existed heterogeneity within the levels. Additionally, the uneven sampling across the levels of such classifications was a crucial consideration in terms of how they should be treated in model building. Particularly, type of day was a categorical variable with levels weekday and weekend; weekday was sampled more often in the model building. In effect, the estimation process must have the potential to reduce the probability of false positives (Type I error rates) and false negatives (Type II error rates). In additional, the process must have the ability to appropriately infer the magnitude of variation within and among clusters or hierarchical levels (Crawley, 2013; Harrison *et al.*, 2018).

The distinction between specifying covariates as fixed or random effect is not always obvious and the multiple definitions in the literature add to the dilemma (Gelman and Hill, 2007). **Study II** used a generalized linear modelling framework to explore two ways of treating the temporal variables, either as fixed or random effects. It was found that treating these variables as random

effects produced consistent estimates with narrower confidence intervals compared to the fixed effect counterparts in ten replicated runs. It was noted that treating the variables as random effects enabled the explicit modelling of the random structures in the boating effort data. This aided the correct inference about fixed effects, depending on which level of the data's hierarchy was being manipulated. For example, when fixed effects varied or were manipulated at the level of time of day, then treating number of boat counts from a time block as independent represented pseudo replication, which was controlled carefully by using random effects. Similarly, if fixed effects varied at the levels of time of day, then the non-independence of the hours within the time blocks was also accounted for. The estimation of random effects was done with partial pooling, which ensured that level's effect estimate was based partially on more abundant data across levels, thus addressing the sampling disparity issues.

In **Study III**, generalized linear mixed effect models with climatic and temporal variables were considered to build imputation models to "fill-in" missing observations in the digital camera monitoring data. The study design used was a simulation scenario, where observed data of complete records were turned into missing data based on 10 observed outage patterns, with missing proportion ranging from 0.06 to 0.61. Nine models were built on the full conditional specification multiple imputation (FCS-MI) framework (van Buuren and Groothius-Oudshoorn, 2011; Kleinke and Reinecke, 2013). The FCS-MI framework was used to specify conditional models of the partially observed outcome variable given the covariates, to obtain a posterior predictive distribution. Two approaches were investigated to obtain independent draws of the parameters for the partially observed outcome variable. The first approach used a Gibbs sampler to make independent draws from an assumed normally distributed pool of model parameters. The second approach estimated the parameters using bootstrapping. The models were found to reconstruct plausible values of counts of powerboat retrievals for the durations of outages studied. The longest outage imputed was 80 days (~1,920 hours) and the shortest was one hour. There were no systematic trends in performance among the models, however, zero-inflated Poisson (ZIP) and its bootstrap variant models consistently ranked amongst the top three models and possessed the narrowest confidence intervals. The outlined framework has adaptable properties, as the choice and type of model will generally be dependent on the nature and characteristics of the data set and the missing patterns being investigated. However, the ZIP models are likely to perform well for count data with many zeros. This was established in the satisfactory results obtained when the ZIP models were applied to impute missing observations in digital camera monitoring data observed at two different ramps in WA. The

additional advantage of the multiple imputation scheme is that it has self-correcting properties which makes it robust even in cases where imputation models are slightly mis-specified (Salfran and Spiess, 2015). Admittedly, in the outlined framework there is the possibility of specifying models where the conditional distributions will not correspond to valid joint distributions, however, it has been established that in practice this will have little impact on the results (Raghunathan *et al.*, 2001; van Buuren, 2007).

Understanding accessibility patterns at ramps by recreational boaters will help managers to have greater anticipation of the current and future management needs to inform regulatory policies. In recreational fisheries management practices, short-term forecasts may assist in guiding decisions that will help to draw the right balance between sustainable recreational fisheries management practices and high-quality fishing experience for recreational fishers. Digital camera monitoring of boating effort may observe data that are highly variable, of fine granularity and can be characterized by sparse periods of zero counts, dominated by seasonality that results in cyclicity, making forecasting difficult. **Study IV** evaluated and compared point estimates of short-term forecasts of boating effort using intermittent demand forecasting techniques, including Croston's method and Syntetos-Boylan Approximation (SBA) models, as well as count data forecasting methods including autoregressive conditional Poisson (ACP) models, integer-valued moving average (INMA) models, and integer-valued autoregressive (INAR) models. Integer-valued autoregressive (INAR) and autoregressive conditional Poisson (ACP) models were identified as useful for predicting short-term behaviour of recreational boating effort. Based on the success achieved in using the study covariates to impute plausible values for the missing observations in the boating effort data, **Study V** explored the opportunities to use them to predict the temporal distribution of recreational boating effort. Using a Bayesian regression framework, it was found that the covariates contained adequate predictive abilities to reveal patterns and trends in recreational boating effort. Using the No-U-Turn Sampler (NUTS) proposed by Hoffman and Gelman (2014), the modelling framework provided greater flexibility and power to uncover complex relationship structures within the datasets. This modelling scheme would provide continuous time series data on boating effort and provide additional support for dealing with missing observation issues in digital camera monitoring data.

One major challenge of digital camera monitoring of recreational boating effort is the cost of manual interpretation of video imagery, especially if multiple sites are involved. In the final study, **Study VI,** an *a posteriori* analysis was undertaken to investigate the trade-offs between

the reading cost and accuracy measures of estimates of boat retrievals obtained at various sampling proportions for low, moderate and high traffic boat ramps, thereby informing decisions on approaches to be used for future reading of camera data. Classical random sampling techniques including simple random sampling, systematic sampling and stratified sampling designs with proportional and weighted allocation were found to produce unbiased estimates of the total number of powerboat retrievals in 10,000 jackknife resampling draws. It was concluded that manual interpretation of camera footage for 40% of the days within a year can be deemed as an adequate level of sampling effort to obtain unbiased, precise and accurate estimates to meet broad management objectives. The relative standard error (RSE ± standard deviations) obtained for sampling proportions from 0.4 onwards were below the 20% threshold adopted in some fisheries research practices (Vølstad *et al*., 2014) for three of the sampling designs across the three boat ramps. Coverage rates of over 90% were observed for the confidence intervals for the estimated annual number of powerboat retrievals, with low relative standard errors (RSE < 20%). While the automation of the monitoring system would ultimately provide a cost-efficient means of data interpretation (Buch *et al.*, 2011), advances in this technology are in an early phase for monitoring recreational fishing effort (Hartill *et al.*, 2019). Thus, in the interim and beyond, the current study can improve the utility of digital camera monitoring by reducing the cost of manual data interpretation. The consistency in the trends of the relationships between the performance indicators, cost across ramps and sampling proportion from the sampling designs are indicative of the significant gains achieved and their reliability in practice.

8.2 Limitations and future work

In this thesis, the study objectives were achieved by applying diverse concepts of statistical techniques, specifically imputation, modelling, time series analysis and sampling. In this thesis the FCS-MI framework was used to impute plausible values for the periods of missing data, however other frameworks such as a Bayesian joint modelling approach or machine learning algorithms were not explored. A comparative study of different imputation modelling frameworks would be useful to fully understand the strengths and weaknesses that may exist in the approach chosen. This would provide opportunities of applications and help contextualize analytical techniques to meet specific imputation objectives. Secondly, the length of data used in this thesis is relatively short for a time series analysis, and there was the additional challenge of the paucity in the series. An attempt has made to construct the gaps between surveys, however, there were no rigorous validation of the modelled series. It would

be useful to explore adaptive schemes where sparse readings between survey years are matched to the modelled series. In addition, longer continuous time series of digital camera monitoring data would enhance the opportunities of using statistical modelling techniques to improve the utility of digital camera monitoring in recreational fisheries. Thirdly, the study predictors, notably the climatic variables are not exhaustive and leave a wide area for exploration as recreational boaters' exhibit greater stochasticity in their boating behaviours. To gain greater understanding of the processes underlying the relationships of climatic variables and recreational boating activity, a further exploration of varying modelling setups and adaptation of the climatic conditions in a simulation study would be useful. Finally, automation of digital camera monitoring system will improve efficiency and enhance application.

Chapter References

Allison, P. (2012). When can you safely ignore multicollinearity? Statistical Horizons. Retrieved on 31.10.2018 from *http://statisticalhorizons.com/multicollinearity.*

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *Preprint arXiv:1701.02434v2*. Columbia University, New York.

Buch, N., Valastin, S. A., and Orwell, J. (2011). A review of computer vision techniques for the analysis of urban traffic. *IEEE Transactions on Intelligent Transportation Systems*, 2(3), 920-939.

Crawley, M. (2013). *The R Book* (Second Edition). Chichester: Wiley.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Hierarchical/Multilevel Models*. New York: Cambridge University Press.

Hartill, B. W., Payneb, G. W., Rusha, 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. (2019). Digital camera monitoring of recreational fishing effort: applications and challenges, *Fish and Fisheries*, DOI: 10.1111/faf.12413

Harrison, X. A., Donaldson, L., Correa-Cano, M. E. *et al*. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 1-32.

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

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1593–1623.

Jahng, S., and Wood, P. K. (2017). Multilevel models for intensive longitudinal data with heterogeneous autoregressive errors: the effect of misspecification and correction with Cholesky transformation. *Frontiers in Psychology*, 8:262, 1-15.

Klienke, K. and Reincke, J. (2013). Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica,* 67(3): 311-336.

Raghunathan, T., Lepkowski J., Van Hoewyk, J., Solenberger P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology,* 27(1):85–95.

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*, No. 249, Department of Fisheries, Western Australia.

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*, No. 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*, No. 287, Department of Primary Industries and Regional Development, Western Australia.

Salfrán, D. and Spiess, M. (2015). A Comparison of Multiple Imputation Techniques. Technical Report, Discussion Paper. Universität Hamburg.

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(6): 1265-1276.

van Buuren S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res.*;16(3):219–42.

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., Carrutters, 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.

Vølstad, J. H., Afonso, P. S, Baloi, A. P., de Premegi, N., Meisfjord, J. and Cardinale, M. (2014). Probability-based survey to monitor catch and effort in coastal small-scale fisheries. *Fisheries Research*, 151, 39-46.

Appendix A: Sample R-codes

```
 Chapter 2
# Set working directory
# Use read.table to load data

# Multiple regression modelling
# Create indicator variable for variable with missing
# observations
 Var.name <- function(aug){
               x<-dim(length(aug))
                x[which(!is.na(aug))]=1
                x[which(is.na(aug))]=0
             return(x)
                        }

## Imputation function
 for (i in 1:nrow(Data))
 {
   if(Data$Var.name[i]==0)
   {
     Data$var_1[i]= c_0+c_ii*Data$vars[-var_1]
   }
 }
# where vars are the set of predictors and var_1 is the
# response

# Structural models & State-space ARIMA(with Kalman filtering)
# library(imputeTS) (version 2.7)
# Transform data as ts object
 na.kalman(ts.object, model=c("StructTS","auto.arima")

# accuracy measures: mean absolute error (MAE), root mean
# square error (RMSE) and symmetric mean absolute percentage
# error (SMAPE)
 MAE <- (sum(abs(Data$Observed-Data$Estimate)))/nrow(Data)
 MSE <- (sum((Data$Observed-Data$Estimate)^2))/nrow(Data)
 RMSE <- sqrt(MSE)
 SMAPE <- ((sum(abs((Data$Observed-Data$Estimate)/
          Data$Observed)))/nrow(Data))*100
 cbind(MAE, MSE, RMSE, SMAPE)
```

```
 Chapter 3 & 4
# Set working directory
# Use read.table to load data

#library(mice) (version 2.14)
#library(VIM)  (version 4.8.0)
#library(pscl) (version 1.5.2)
#library(countimp) (version 1.0)
#library(lattice) (version 0.20-38)
#library(ggplot2)

# Initialize the mice algorithm
   ini<-mice(Data, m=5, maxit = 0, print = FALSE)

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

# Assign a new predictor matrix to meet modelling specification
   pred<-ini$predictorMatrix
   pred[var[i],] <- c(); c() may contain 0, 1, 2, and/or 3
# Note that number of entries in c() is dependent on number of
# predictors

# Specify the model
   meth<-ini$method
   meth[var[i]]<-"model"
# For example, model=='2l.zip' imputes missing data based on a
# generalized linear mixed effects Zero-inflated Poisson model.
# See Klienke, K. and Reincke, J. (2013). Multiple imputation
# of incomplete zero-inflated count data. Statistica
# Neerlandica, 67(3): 311-336.

# Run imputation algorithm
# Specify the number of multiple imputations (m) and iterations
# (maxit)
# Set seed for reproducibility
   imp <- mice(Data, m = num, method = meth,maxit = num,
              predictorMatrix = pred, seed = num, print = TRUE)
# This algorithm runs on 'countimp' using 'mice' and 'pscl' as
portable interfaces

# For illustration, suppose we are imputing missing data with
# zero-inflated Poisson model where underlying process for the
# zero and count parts are influenced by the same set of
# predictors assuming there are 3 predictors). The modelling
# configuration will be as follows;
   ini<-mice(Camera, m=5, maxit = 0, print = FALSE)
   pred<-ini$predictorMatrix
   pred[1,] <- c(0,1,1,1)
```

```
   meth<-ini$method
   meth[2]<-"2l.zip"
   imp <- mice(Data, m = num, method = meth, maxit = num,
               predictorMatrix = pred, seed = num, print = TRUE)

# Evaluation of imputation
# generate the convergence properties of the multiple
# imputations
   plot(imp)

# generate plot where the distributions of the observed and
# imputed data are compared
   Long <- complete(imp,"long")
   levels(long$.imp) <- paste("Imputation",1:5)
   long <- cbind(long, La.na =is.na(imp$data$Var.nam))

   densityplot(~Var.name|.imp, data=long, group = La.na,
               plot.point = FALSE, ref=TRUE, xlab="Counts",
               scales = list(y=list(draw=F)), par.settings =
               simpleTheme(col.line = rep(c("blue","red"))),
               auto.key = list(columns=2, text = c("Observed",
               "Imputed")))
# accuracy measures: mean absolute error (MAE), root mean
# square error (RMSE) and skill score (SS)
   Bias <- 100*(sum(Imputed - Observed))/sum(Observed)
   MAE <- (sum(abs(Observed- Imputed)))/nrow(Data)
   MSE <- (sum((Observed- Imputed)^2))/nrow(Data)
   RMSE <- sqrt(MSE)
   MSE_1 <- (sum((Observed- Imputed)^2))/nrow(Data)
   MSE_2 <- (sum((mean(Observed)-Imputed)^2))/nrow(Data)
   SS <- 1 - (MSE_1/MSE_2)
   cbind(Bias, MAE, RMSE, SS)
```

```
 Chapter 5
# Set working directory
# Use read.table to load data

# library(forecast) (version 8.7)
# library(lubridate) (version 1.7.4)
# library(tsintermittent) (version 1.9)
# library(acp) (version 2.1)
# library(dplyr) (version 0.8.3)
# library(tscount) (version 1.4.2)
# Specify the format of date
 Data$Date <- dmy_hm(Data$Date)

# Split data into training and test set
 Train_data <- ts(Data$var[i], start = num, end = num)
 Test_data <- ts(Data$var[i],start = num, end = num)

# Fit Croston & SBA model to the dataset
 Model <- crost(Data, h = time horizon, type = c("croston",
                "SBA"))

# Fit Autoregressive Conditional Poisson (ACP) model
 Model <- acp(var[i]~-1,data = Data,p = num, q = num,
              family="acp")
# See Heinen, A. (2003). Modeling Time Series Count Data: An
# Autoregressive Conditional Poisson Model, MPRA Paper No.8113

# Fit Integer-valued Autoregressive (INAR) model
 Model <-tsglm (Data, model = list(past_obs = c(num,num)),
                distr="distribution")

# Fit Integer-valued Moving Average (INMA) model
 Model <-tsglm (Data, model = list(past_mean=c(num)),
                distr="poisson")

# measure accuracy of model for in-sample and out-of-sample
# prediction
 accuracy(Model$components$c.in[,1],Data)

# Mean Absolute Scaled Error
 mase <- function(Train_ts, Test_ts, outsample_forecast){
            naive_insample_forecast <- stats::lag(Train_ts)
            insample_mae <- mean(abs(Train_ts -
            naive_insample_forecast), na.rm = TRUE)
            error_outsample <- Test_ts - outsample_forecast
            ase <- error_outsample / insample_mae
            mean(abs(ase), na.rm = TRUE)
            }
 mase(Train_data,Test_data,h_ahead)
```

```
 Chapter 6
# Set working directory
# Use read.table to load data

# library(sjmisc) (version 2.8.2)
# library(rstan) (version 2.19.2)
# library(rstanarm) (version 2.19.2)
# library(sjstats) (version 0.17.7)
# library(rstantools) (version 2.0.0)
# library(brms) (version 2.10.0)
# library(mgcv) (version 1.8-28)
# library(coda) (version 0.19-3)
# library(ggplot2) (version 3.2.1)
# Set priors on the predictors
 priori <- get_prior(Response ~ Predictors, family="model",
                     data   = Data)

 prior <- c(set_prior("dist. specification", class = "", coef
            = "Predictor"),)

# Fit Bayesian regression model
 Modeltest <- brm(Response ~ Predictors,
                  data   = Data,
                  family = "distribution",
                  warmup = num,
                  iter   = num,
                  chains = num,
                  prior = prior,
                  control = list(adapt_delta = num),
                  inits  = "random",
                  cores  = num)
# This model runs in 'brms' using 'rstanarm' as a portable
# interface for running the model in Stan for full Bayesian
# inference
# See Bürkner, P.-C. (2017). Brms: An R package for Bayesian
# Multilevel Models using Stan. Journal of Statistical
# Software, 80(1): 1-28, doi:10.18637/jss.v080.i01
# See also Bürkner, P.-C. (2018). Advanced Bayesian Multilevel
# Modelling with the R package brms. R Journal, 10(1), 395-411.
# https://doi.org/10.32614/RJ-2018-017

# Evaluation - display densities overlay from ensembles of
# predicted distribution for daily powerboat launches from the
# fitted models and the observed data
    par(mfrow=c(1,2))
    pp_check(Modeltest,nsamples = num)
    pp_check(Modeltest,nsamples = num,type = "loo_pit_overlay")

# To estimate Bayes R^2
    plot(bayes_R2(Modeltest))
```

```
# To obtain the marginal_smooths for the predictors
    modelposterior <-as.mcmc(Modeltest)

# To obtain the Geweke diagnostic use – to assess convergence
# by comparing the estimated between-chains and within-chain
# variances for each model parameter
    geweke.diag(modelposterior[,1:23],frac1 = 0.1,frac2 = 0.9)
    geweke.plot(modelposterior[,1:23],frac1 = 0.1,frac2 = 0.9)

# Plot fitted means against actual response
    dat1 <- as.data.frame(cbind(Y = standata(Modeltest)$Y,
                          fitted_values1))
    ggplot(dat1) + geom_point(aes(x = Estimate, y = Y))
#
# Predict for new dataset of predictors
    future <- predict(Modeltest, newdata = newdata)
```

```
 Chapter 7
# library(TeachingSampling) (version 3.4.2)
# library(sampling) (version 2.8)
# library(SamplingStrata) (version 1.4-1)
# Simple random sampling (SRS)
  Popsize =366
  sample_stats <- function(df, n=n){
# randomly sample size of n without replacement from the
# dataframe
       df1 <- df[sample(1:nrow(df), n, replace=F),]
# post-stratification
      m_1<- filter(df1, Interact == "Autumn/Weekday") %>%
      summarise(Awd_n=n(),Awd_Tot=sum(var.nam),
             Awd_M=mean(var.nam),Awd_SE=sd(var.nam),
             Awd_E=strata_1*Awd_M, Awd_ESE=
            (sqrt((strata_1-Awd_n)/(strata_1-1)))
    *(sqrt(strata_1^2/Awd_n)) * Awd_SE, Awd_PME = 1.96*Awd_ESE)
# Do for all 8 strata
# estimate the mean of retrievals from the sample
  mx <- mean(df1$var.nam)
# estimate the standard deviation with fpc factor
  sdx <- sd(df1$var.nam)
  sumx<- sum(df1$var.nam)
# expanded total estimate
  Total <- Popsize*mx
  Total_se <- sqrt((Popsize^2)/n)*sdx
  CI_U <- Total+1.96*Total_se
  CI_L <- Total-1.96*Total_se
  PME <- 1.96*Total_se
# Check if actual value is found within confidence interval
  Count <- between(Actual, CI_L,CI_U)
# coeeficient of variation
  cv <- sdx/mx
# estimate the root mean square error.
  RMSE <- sqrt((sum((df1$var.nam -mx)^2)/n))
  return(c(Mean = mx, SD = sdx, S_total = sumx, E_total = Total,
      E_Tse = Total_se, CL = CI_L, CU = CI_U, PME = PME,
#    Cov = Count, CV = cv, RMSE = RMSE, m_1, m_2, m_3, m_4, m_5,
     m_6, m_7, m_8))
     }
# 10000 Jacknife draws
# results <- replicate(10000, sample_stats(Data, n = num))
# Systematic sampling (SSRS)
 sample_stats <- function(df){
       sam<-S.SY(Popsize, int)
       df1<-df$var[sam]
       m_1 <-summary(df1)
       df2<-df$Ret[sam]
       mx <- mean(df2)
       sdx <- sd(df2)*sqrt((Popsize-length(df2))/(Popsize-1))
       sumx<-sum(df2)
```

```
       Total <- Popsize*mx
       Total_se <- sqrt((Popsize^2)/length(df2))*sdx
       CI_U <- Total+1.96*Total_se
       CI_L <- Total-1.96*Total_se
       PME <- 1.96*Total_se
       Count <- between(Actual, CI_L,CI_U)
       cv <- sdx/mx
       RMSE <- sqrt((sum((df2-mx)^2)/length(df2)))
  return(c(Mean=mx, SD=sdx, S_total=sumx, E_total=Total,
           E_Tse=Total_se, CL=CI_L, CU=CI_U, PME=PME,
           Cov= Count, CV= cv, RMSE=RMSE,m_1))
    }
 int = num
 results <- replicate(10000, sample_stats(Camera))
# Stratified sampling (SRSP)
  sample_stats <- function(df){
   df <- df %>% group_by(Interact) %>% sample_frac(sp)
   m_1<- filter(df, Interact == "Autumn/Weekday") %>%
    summarise(Awd_n=n(),Awd_Tot=sum(var.nam),
              Awd_M=mean(var.nam),Awd_SE=sd(var.nam),
              Awd_E=strata_1*Awd_M, Awd_ESE=
              (sqrt((strata_1-Awd_n)/(strata_1-1)))
      *(sqrt(strata_1^2/Awd_n)) * Awd_SE,Awd_PME=1.96*Awd_ESE)
# Do for all 8 strata
# Estimate the all relevant statistics from the sample
  return(c(all relevant estimates))
  }
 sp = num
 nsize = num
 results <- replicate(10000, sample_stats(Data))
# Optimal stratified sampling design (SRSW)
 sample_stats <- function(df){
      df1<-strata(df, stratanames ="Interact", size= size,
                  method="srswor")
      df2<-getdata(df,df1)
      m_1<- filter(df2, Interact == "Autumn/Weekday") %>%
           summarise(Awd_n=n(),Awd_Tot=sum(var.nam),
              Awd_M=mean(var.nam),
              Awd_Ma=strata_1*mean(var.nam),
              Awd_SE=(sqrt((strata_1-Awd_n)/(strata_1-1)
              *(sum(df2$ var.nam -Awd_M)^2)/(Awd_n-1)
              *(nsize/Awd_n)*(strata_1/Popsize)^2)),
              Awd_s=Awd_SE/(Awd_n)^2,Awd_PME=1.96*Awd_SE)
# Do for all 8 strata
# estimate the all relevant statistics from the sample
  return(c(all relevant estimates))
  }
 nsize = num
 size = c(num,num,num,num,num,num,num,num)
 results <- replicate(10000, sample_stats(Data))
```

Appendix B: Statement of co-authors contribution

Co-authorship statement for publications with the PhD

With reference to ECU thesis with publication policy, statements from co-author(s) attesting to the PhD candidate's contribution to the joint publications must be included in the appendix.

Paper title: Missing data imputation of high-resolution temporal climate time series data

Journal: Meteorological Applications

Paper status: Published

List of authors: Ebenezer Afrifa-Yamoah, Ute A. Mueller, Stephen M. Taylor, Aiden J. Fisher.

PhD candidate: Ebenezer Afrifa-Yamoah

Scientific contributions to the paper:

The PhD candidate contributed to conceptualization, data analysis and validation, manuscript writing, review and editing constituting 80% of the work.

Stephen M. Taylor contributed to the development of the idea, interpretation of the results and critical revision of the manuscript (5%).

Aiden J. Fisher contributed to the development of the idea and critical revision of the manuscript (5%).

Ute A. Mueller contributed to the development of the idea, interpretation of the results and critical revision of the manuscript (10%).

Signature, PhD candidate

I, as a co-author, endorse that this level of contribution by the candidate indicated above is appropriate.

Signatures, co-authors

Associate Professor Ute A. Mueller:  Date: 16/10/2020

Dr Stephen M. Taylor:  Date: 16/10/2020

Dr Aiden J. Fisher:  Date: 20/10/2020

Co-authorship statement for publications with the PhD

With reference to ECU thesis with publication policy, statements from co-author(s) attesting to the PhD candidate's contribution to the joint publications must be included in the appendix.

Paper title: Imputation of missing data from time-lapse cameras used in recreational fishing surveys.

Journal: ICES Journal of Marine Science

Paper status: Accepted

List of authors: Ebenezer Afrifa-Yamoah, Stephen M. Taylor, Aiden J. Fisher, Ute A. Mueller.

PhD candidate: Ebenezer Afrifa-Yamoah

Scientific contributions to the paper:

The PhD candidate contributed to conceptualization, data analysis and validation, manuscript writing, review and editing constituting 70% of the work.

Stephen M. Taylor contributed to the development of the idea, interpretation of the results and critical revision of the manuscript (10%).

Aiden J. Fisher contributed to the development of the idea and critical revision of the manuscript (5%).

Ute A. Mueller contributed to the development of the idea, interpretation of the results and critical revision of the manuscript (15%).

Signature, PhD candidate

I, as a co-author, endorse that this level of contribution by the candidate indicated above is appropriate.

Signatures, co-authors

Associate Professor Ute A. Mueller:  Date: 16/10/2020

Dr Stephen M. Taylor:  Date: 16/10/2020

Dr Aiden J. Fisher:  Date: 20/10/2020

Co-authorship statement for publications with the PhD

With reference to ECU thesis with publication policy, statements from co-author(s) attesting to the PhD candidate's contribution to the joint publications must be included in the appendix.

Paper title: Trade-off assessments between reading cost and accuracy measures for digital camera monitoring of recreational boating effort.

Journal: Fisheries Research

Paper status: Accepted

List of authors: Ebenezer Afrifa-Yamoah; Stephen M. Taylor; Ute Mueller

PhD candidate: Ebenezer Afrifa-Yamoah

Scientific contributions to the paper:

The PhD candidate contributed to conceptualization, data analysis and validation, manuscript writing, review and editing constituting 75% of the work.

Stephen M. Taylor contributed to the development of the idea, interpretation of the results and critical revision of the manuscript (10%).

Ute A. Mueller contributed to the development of the idea, interpretation of the results and critical revision of the manuscript (15%).

Signature, PhD candidate

[redacted]

I, as a co-author, endorse that this level of contribution by the candidate indicated above is appropriate.

Signatures, co-authors

Dr Stephen M. Taylor: [redacted] Date: 16/10/2020.

Associate Professor Ute A. Mueller: [redacted] Date: 16/10/2020

Co-authorship statement for publications with the PhD

With reference to ECU thesis with publication policy, statements from co-author(s) attesting to the PhD candidate's contribution to the joint publications must be included in the appendix.

Paper title: Modeling climatic and temporal influences on powerboat launches with relevance to recreational fisheries

Journal: To be submitted to Fisheries Research

Paper status: Manuscript

List of authors: Ebenezer Afrifa-Yamoah; Stephen M. Taylor; Ute Mueller

PhD candidate: Ebenezer Afrifa-Yamoah

Scientific contributions to the paper:

The PhD candidate contributed to conceptualization, data analysis and validation, manuscript writing, review and editing constituting 75% of the work.

Stephen M. Taylor contributed to the development of the idea and critical revision of the manuscript (10%).

Ute A. Mueller contributed to the development of the idea, interpretation of the results and critical revision of the manuscript (15%).

Signature, PhD candidate

I, as a co-author, endorse that this level of contribution by the candidate indicated above is appropriate.

Signatures, co-authors

Dr Stephen M. Taylor:  Date: 16/10/2020

Associate Professor Ute A. Mueller:  Date: 16/10/2020

Co-authorship statement for publications with the PhD

With reference to ECU thesis with publication policy, statements from co-author(s) attesting to the PhD candidate's contribution to the joint publications must be included in the appendix.

Paper title: Short-term prediction of recreational boating effort: Evaluation of intermittent demand and count data forecasting methods.

Journal: To be submitted to Journal of Time Series Analysis

Paper status: Manuscript

List of authors: Ebenezer Afrifa-Yamoah; Stephen M. Taylor; Ute Mueller

PhD candidate: Ebenezer Afrifa-Yamoah

Scientific contributions to the paper:

The PhD candidate contributed to conceptualization, data analysis and validation, manuscript writing, review and editing constituting 75% of the work.

Stephen M. Taylor contributed to the development of the idea and critical revision of the manuscript (10%).

Ute A. Mueller contributed to the development of the idea, interpretation of the results and critical revision of the manuscript (15%).

Signature, PhD candidate

I, as a co-author, endorse that this level of contribution by the candidate indicated above is appropriate.

Signatures, co-authors

Dr Stephen M. Taylor:  Date: 16/10/2020

Associate Professor Ute A. Mueller:  Date: 16/10/2020