Fall 12-19-2020

Examining Multiple Imputation for Measurement Error Correction in Count Data with Excess Zeros

Shalima Zalsha
Southern Methodist University, szalsha@smu.edu

EXAMINING MULTIPLE IMPUTATION FOR MEASUREMENT ERROR CORRECTION IN COUNT DATA WITH EXCESS ZEROS

Approved by:

Dr. S. Lynne Stokes
Professor, Department of Statistical Science (SMU)

Dr. Jing Cao
Associate Professor, Department of Statistical Science (SMU)

Dr. Jennifer Dworak
Associate Professor, Department of Electrical and Computer Engineering (SMU)

Dr. Monnie McGee
Associate Professor, Department of Statistical Science (SMU)

EXAMINING MULTIPLE IMPUTATION FOR MEASUREMENT ERROR CORRECTION IN COUNT DATA WITH EXCESS ZEROS

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Statistical Science

by

Shalima Zalsha

B.A., Mathematics, University of Nebraska-Lincoln
B.S.B.A., Management, University of Nebraska-Lincoln
M.S., Statistics, Sam Houston State University

December 19, 2020

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Stokes, for her guidance, support, and training. Dr. Stokes has always inspired and motivated me to be the best version of myself in and outside of class—without her support and encouragement, this dissertation would not have been possible. I would also like thank my committee members, Dr. Cao, Dr. McGee, and Dr. Dworak for their time, encouragement, and suggestions to improve my research. Special thanks to Dr. Potgieter, Dr. Robertson, Dr. South and Sheila for their support during my time at SMU. To my friends and all the wonderful people that I have met at SMU, thank you for making this journey memorable. Finally, I am forever grateful for my family, for their unconditional love—for their support and understanding throughout this journey.

Zalsha, Shalima

B.A., Mathematics, University of Nebraska-Lincoln
B.S.B.A., Management, University of Nebraska-Lincoln
M.S., Statistics, Sam Houston State University

Examining Multiple Imputation
for Measurement Error Correction
in Count Data with Excess Zeros

Advisor: Dr. S. Lynne Stokes

Doctor of Philosophy degree conferred December 19, 2020

Dissertation completed December 2, 2020

Measurement error and missing data are two common problems in wildlife population surveys. These data are collected from the environment and may be missing or measured with error when the observer's ability to see the animal is obscured. Methods such as video transects for estimating red snapper abundance and aerial surveys for estimating moose population sizes are highly affected by these problems since total abundance will be underestimated if missing and mismeasured counts are ignored. We shall refer to this problem as visibility bias; it occurs when the true counts are observed when visibility is high, partially observed when visibility is low (mismeasured), and unobservable when visibility is lost (missing). In addition, data from animal population surveys are often sparse since not all sampled regions are inhabited by the species.

In this dissertation, we examine several multiple imputation techniques which can be used to correct measurement error in sparse count data that are subject to visibility bias. We consider several off-the-shelf imputation models such as normal, Poisson, zero inflated Poisson imputation, and predictive mean matching. In addition, we develop and examine a Hierarchical Bayes Zero Inflated Poisson imputation model which we refer to as HBZIP, and a modified hot deck imputation approach. Each of the method's performance is evaluated in a simulation study for the purpose of estimating total abundance and habitat occupancy rate.

We further assess the robustness of the HBZIP model against visibility model misspecification and incorporate Bayesian model averaging to reduce the impact of visibility model uncertainty. For illustration, we implement the HBZIP imputation model on real data collected from moose population surveys and compare its results to an existing weighting adjustment approach. Finally, we present another simulation study based on the moose data to examine the model's performance when the sampling design is complex and discuss future directions of the research.

TABLE OF CONTENTS

APPENDIX

LIST OF FIGURES

LIST OF TABLES

This thesis is dedicated to my family in the northern and southern hemisphere.

Chapter 1
Introduction

Measurement error and missing data are two common problems in wildlife population surveys. These data are collected from the environment and may be missing or measured with error when the observer's ability to see the animal is obscured. Methods such as video transects for estimating red snapper abundance and aerial surveys for estimating moose population sizes are highly affected by these problems since abundance will be underestimated if missing and mismeasured counts are ignored. We shall refer to this problem as visibility bias; it occurs when the true counts are observed when visibility is high, partially observed when visibility is low (mismeasured), and unobservable when visibility is lost (missing). In addition, data from animal population surveys are often sparse since not all sampled regions are inhabited by the species. In that case, the frequency of zeroes in the data exceeds its expected frequency under the assumed count distribution (zero-inflated).

Methods such as measurement error correction and weighting adjustment have been used to correct visibility bias. In a traditional measurement error correction framework, distributional parameters of the error distribution can be estimated from a validation sample or repeated measurements. These quantities are then used to adjust the estimators so that they are approximately unbiased or consistent estimators of the population or model parameters under study. Therefore, the correction method will vary depending on the parameter of interest. The analysts must select the appropriate correction method and cannot merely implement the same method of analysis as the ones used for complete data. Weighting adjustment, on the other hand, is a design-based adjustment method in which unobserved counts are treated as a non-response problem. It is a routine procedure in survey sampling to compensate for survey nonresponse. This method, however, does not allow adjustments when the response rate is 0; hence, it is unable to adjust the unobserved data when visibility is completely lost.

Another way to deal with mismeasured data is to discard them. Though imperfect, mismeasured data still contain information about the true data; therefore, users may want to avoid discarding data they know are mismeasured and preserve as much information as possible. When data are not just obscured but completely missing, a common approach for preserving information is multiple imputation. Recently, it has also been proposed as a remedy for measurement error (Cole et al., 2006; Reiter and Raghunathan, 2007). The idea behind multiple imputation for mitigating measurement error is similar to the regular imputation for missing data: the true counts are not observed; therefore, plausible values of the true counts are generated stochastically from the assumed model.

Multiple imputation methods are available through most statistical computing software and are relatively easy to implement. They perform well for a broad range of data types so that new models for different types of data have not always been needed. Furthermore, they are also not analysis specific, allowing users of the imputed data to perform the same type of analysis used for complete data. Multiple imputation incorporates not only variability from the sampling design but also the extra uncertainty from imputing the missing and mismeasured data. Therefore, inference about population parameters can be improved by properly adjusting the bias and incorporating the extra variability due to imperfect visibility.

In this study, we examine several multiple imputation techniques which can be used to correct sparse count data that are subject to visibility bias. We consider several multiple imputation models including off-the-shelf imputation methods such as normal, Poisson, zero inflated Poisson imputation, and predictive mean matching. In addition, our specialized imputation models Hierarchical Bayes Zero Inflated Poisson (HBZIP) and a modified hot deck imputation are also presented and examined. We conduct a simulation study to evaluate each method's performance for estimating total abundance and habitat occupancy rate, and its ability to preserve the logical constraints and structural zeros of the data. In addition, we incorporate Bayesian model averaging into the HBZIP model to accommodate multiple visibility models. This approach is then implemented on data from moose population surveys, and its performance is further evaluated in a simulation study.

We provide additional background information about our study in the remainder of this chapter. The rest of the paper is structured as follows: in Chapter 2, we provide and discuss several imputation methods including off-the-shelf and our specialized imputation methods, HBZIP and modified hot deck, which can be used to correct measurement error. In Chapter 3, each method's performance for various count data and visibility mechanism scenarios are examined via a simulation study. In Chapter 4, we introduce Bayesian model averaging and incorporate it into our HBZIP imputation model to overcome visibility model uncertainty. We implement the method on real data, moose population surveys, and conduct an additional simulation study based on the real data in Chapter 5. Finally, conclusion and future direction of the research are presented in Chapter 6.

1.1. Motivating Example

Red snapper is one of the most valuable species in the Gulf of Mexico. Until recently, Red Snapper had been classified as 'overfished' since the first stock assessment in 1988 (Goodyear, 1988). Regulations have been imposed to restrict fishing by implementing fishing quotas in the commercial sector, and season length in the private recreational sector. Since 2007, season length has decreased from 194 days to 2-3 days in 2016 (Stunz et al., 2017). However, there was a widely held perception that the overall stock was more abundant than what the stock assessment model suggested, and a more robust population estimate that covered the diverse red snapper habitats was deemed necessary.

Consequently, researchers determined that multiple sampling strategies were necessary to cover the diverse red snapper habitats. One of the sampling strategies is underwater towed-camera surveys which produce video footage used to enumerate red snappers during the transect. The camera was built on a frame, equipped with a laser device used to measure the width of the visual field, and towed by a vessel along a straight path at a constant speed above the seafloor. In addition, an acoustic sonar device is also available but only for some of the transects. The sonar device can be used to detect fish and adjust the mismeasured or missing counts from the footage when visibility is lost or degraded. During the transect, the surveyors can also obtain the water turbidity level which is a measure of water's clarity or haziness. Therefore, information from these various instruments is valuable and can be used to adjust the mismeasured or missing count data from the video footage.

For transects in which the additional information is unavailable, the analysts are faced with two options: discard the mismeasured data to improve the accuracy of the estimate at the expense of its precision or use the mismeasured data to improve the estimator's precision at the expense of its accuracy. Neither method is optimal since stakeholders of the data, including biologists, conservationists, and policymakers, rely on both accuracy and precision of the estimates. Therefore, alternative methods are needed to preserve the accuracy of these estimates without sacrificing too much of their precision. This can be achieved by incorporating information from the various instruments within a multiple imputation framework.

1.2. Visibility Mechanism and Observation Model

Let N be the true count of fish or local abundance in a transect. Depending on the water clarity as measured by turbidity level (T), the true count may or may not be observed. When T is low, visibility is high, and the true count is observed. However, when T is moderate, only some of the fish are visible, and only the mismeasured count is observed. Finally, when T is high, no fish are visible, and the true count is missing. Hence, T is believed to be highly correlated with the proportion of fish observed in the transect, called the visibility rate (P). Therefore, it is intuitively reasonable to assume that T and P are monotonically related. For the rest of the paper, we shall refer to the relationship between T and P, denoted by

$P = g(T)$, as the visibility mechanism.

Based on the visibility mechanism above, there are multiple ways to model the relationship between the observed and true count. First, the observed count (Y) can be assumed to be the realization of a random process such that given N fish present, and visibility rate P, Y for the i^th transect has the following expectation:

$$E[Y_i|N_i, P_i] = N_i \times P_i, \quad P_i \in [0,1]. \tag{1.1}$$

If N and P are also assumed to be random then the relationship between Y and (N, P) becomes hierarchical and the model can be written in two levels. An alternative model defines the relationship between Y and N in terms of a multiplicative measurement error model. Let P be a random variable between 0 and 1 which has a multiplicative effect on N. The multiplicative measurement error model is:

$$Y_i = N_i \times P_i, \quad E[P_i] = \mu_{p,i} \text{ and } Var[P_i] = \sigma^2_{p,i}. \tag{1.2}$$

Note that based on (1.1) and (1.2), the true count is observed when $P_i = 1$, measured with error when $P_i \in [0,1]$, and missing when $P_i = 0$. Finally, rather than multiplicative measurement error, we may define U, the number of fish unobserved, as a random variable with an additive effect as follows:

$$Y_i = N_i + U_i, \quad E[U_i] = \mu_{u,i} < 0 \text{ and } Var[U_i] = \sigma^2_{u,i}, \tag{1.3}$$

Here, both $\mu_{u,i}$ and $\sigma^2_{u,i}$ depend on the value of P_i. That is, when $P_i = 1$, the true count is observed such that $\mu_{u,i} = \sigma^2_{u,i} = 0$. As P decreases, both $\mu_{u,i}$ and $\sigma^2_{u,i}$ also increase in which case the true count is mismeasured or missing. Therefore, the distribution of U_i depends on the value of P_i.

Incomplete data consists of transects in which Y, T and other covariates (X) are observed but N is missing or measured with error. The validation data are a subset of the transect units in which Y, T, X and N are all observed, hence play the role of gold standard data. Therefore, based on the availability of the variables, we can divide the transect data into incomplete and validation data. Both datasets are used in the estimation; therefore, N must be imputed in the incomplete data. Depending on the method, N can be imputed directly from the conditional distribution of $N|Y,P$ from (1.1). Alternatively, $N|Y,P$ can also be imputed using a standard imputation method such as normal or Poisson regression. It can also be imputed using non-parametric imputation methods such as hot deck imputation and predictive mean matching by first imputing P and U, and then inverting the expression in (1.2) and (1.3) to obtain $N = Y/P$ and $N = Y + U$ respectively.

After imputation, both main and validated data are then used to estimate total abundance $\tau = \sum_{i=1}^{L} N_i$, and habitat occupancy rate $\rho = \sum_{i=1}^{L} I_{[N_i>0]}/L$, where L is the number of transect units in the population. Let I be the number of sampled transects; the parameter estimates are denoted by:

$$\hat{\tau} = \sum_{i=1}^{I} W_i N_i \quad \text{and} \quad \hat{\rho} = \frac{\sum_{i=1}^{I} W_i I_{[N_i>0]}}{\sum_{i=1}^{I} W_i} \tag{1.4}$$

where $W_i = 1/p_{s,i}$ is the inverse probability of selection of the i^{th} transect unit or the sampling weight.

1.3. Missing Data as a Measurement Error Problem

Measurement error can be thought of as a missing data problem. In this framework, the true count is believed to be missing rather than mismeasured when visibility is not perfect. The observed count is then treated as an auxiliary variable which can be used in the imputation model. Missing data can also be thought of as an extreme case of measurement error. That is, depending on the visibility rate, true count can either be missing, measured with error, or observed. For example, recall the measurement error model in (1.2) and (1.3). Depending on P_i, N_i is considered missing when $P_i = 0$, mismeasured when $P_i \in (0, 1)$, and observed when $P_i = 1$. Hence, it seems reasonable to treat P_i as a scale measure which indicates the quality of the observation.

Our initial approach to handle visibility bias was to use the visibility rate (P_i) to decide when the observed data should be retained, adjusted, or imputed. In particular, we thought that it might be reasonable to find a lower cutpoint (γ_l) and upper cutpoint (γ_u) so that the observed data are imputed when $P_i \leq \gamma_l$, adjusted when $\gamma_l < P_i < \gamma_u$, and retained when $P_i \geq \gamma_u$. Note that P_i is a latent variable since it is not directly observed, but rather modeled as a function of turbidity. Hence, this idea led us to develop the HBZIP model presented in Section 2.3, which adjusts the observed count within the model depending on P_i. The cutpoints can then be applied after the HBZIP model is implemented. However, our initial simulations suggest that, for inference purposes, the HBZIP model alone performed better than the HBZIP model with the cutpoints. Therefore, the HBZIP model is implemented in this study without incorporating the cutpoints.

1.4. Literature Review

It is not difficult to see that when visibility bias is ignored, total abundance and occupancy rate are underestimated since $Y \leq N$. Therefore, the observed count must be adjusted appropriately to improve inference about τ and ρ. In this section we discuss several methods which can be used to handle visibility biased problems including measurement error correction methods, weighting adjustments, and hierarchical modeling.

1.4.1. Measurement Error Correction

Within the classical measurement error framework, the true data are measured with errors which are typically assumed to have an additive or multiplicative effect on the true data (Buonaccorsi, 2010). The errors are usually assumed to be homogeneous and follow some distribution whose parameters can be estimated using validation data or replicated measurements. The analyst can then use these quantities to adjust the estimator so that it is approximately unbiased or consistent. In our application, however, measurement errors are unlikely to be homogeneous because visibility and abundance vary among the transects. Replicated measurements are also not likely to be available due to resource constraints and because the true fish count will differ each time the transect is conducted. In addition, many of the measurement error techniques were developed for continuous data or for estimating regression parameters. Hence, not much is available for count data, especially when they are sparse and non-homogeneous.

Alternatively, calibration techniques (Buonaccorsi, 2010) can also be used for adjusting for measurement error. This method assumes a linear regression model $Y|N = \beta_o + \beta_1 N$ which can be inverted to predict the true count $\hat{N} = (Y - \hat{\beta}_o)/\hat{\beta}_1$. Estimates are then made using the predicted values such that $\hat{\tau} = \sum_{i=1}^{l} W_i \hat{N}_i$ and $\hat{\rho} = \sum_{i=1}^{l} W_i I_{[N_i > 0]} / \sum_{i=1}^{l} W_i$. However, this method ignores the extra variability which comes from the parameter estimates. Therefore, estimates of τ and ρ may be unbiased, but their standard errors are too narrow resulting in an undercoverage of the confidence interval.

1.4.2. Weighting/Sightability Adjustment

Weighting adjustment can also be used to correct visibility bias. Within this framework, visibility bias can be viewed as a non-response problem which can be corrected with weighting adjustment, also known as sightability adjustment (Steinhorst and Samuel, 1989). In their application, Steinhorst and Samuel treat sighting probability as response probability, and use standard results from survey sampling to estimate the population size of moose. In the case of single-stage cluster sampling, they define l= number of land units (psu's) sampled, n_k= number of groups in the k^{th} land unit, $m_{i(k)}$ = number of animals in the i^{th} group in the k^{th} land unit, $P_{i(k)}$ = sighting probability for the i^{th} group in the k^{th} land unit, and $p_{s,k}$ = probability selection of the k^{th} land unit. Total abundance is estimated by $\hat{\tau} = \sum_{k=1}^{l} 1/p_{s,k} \sum_{i=1}^{n_k} m_{i(k)}/P_{i(k)}$. The sighting probabilities can either be assumed known from a prior experiment or modeled using logistic regression from validation or experimental data.

Applying this method in our application requires that P_i be defined as the probability of observing a fish rather than the i^{th} group of fish, as in Steinhorst and Samuel (1989). Define I = the number of transects in the population, n_i = the true count of fish in the i^{th} transect, and y_i = number of fish observed for the i^{th} transect. Therefore, the estimate of total abundance in our application is $\hat{\tau} = \sum_{i=1}^{I} W_i Y_i / P_i$. However, this method requires $P_i > 0$; therefore, transects with completely lost visibility must be handled in advance. In addition, the estimates can be very unstable when $P_i \approx 0$, since $1/P_i$ becomes highly sensitive to a slight change in P_i.

1.4.3. Hierarchical Model

The true count and observed data can be formulated as a two-level hierarchical model which specifies the distribution of local abundance in the first level and the observed outcome in the second level. Royle and Dorazio (2006) proposed a flexible hierarchical modeling approach for estimating animal abundance and occurrence when the data are subject to imperfect detection. In their application, the count data are assumed to follow a Poisson distribution with a parameter that can be modeled using Poisson regression. Given the true count, the observed count follows a binomial distribution with a success or detection probability that is modeled using logistic regression. Their study, however, focuses on implementing the hierarchical model, not for imputation but for direct inference about the population parameters. Therefore, estimates are obtained directly using samples from the posterior distribution of the population parameter.

In Royle and Dorazio's (2006) application, the count data are assumed to follow a Poisson distribution; hence, it does not accommodate highly sparse data. Furthermore, they also assume a logistic regression model for the visibility parameter. Therefore, the estimates would be biased if the assumed visibility model is different than the actual visibility mechanism. Regardless, their proposed method is natural because it reflects the hierarchical relationship between the true and observed count. In addition, the method can be used not only to make direct inference about the parameters but also to impute the missing and mismeasured counts.

In this chapter, we presented the motivation behind our research, introduced visibility bias as a measurement error problem and discussed existing methods which can be used to correct visibility bias. Next, we introduce multiple imputation as a method to correct for measurement error and present several imputation models including off-the-shelf and our developed imputation models for correcting measurement error in sparse count data.

Chapter 2

Imputation Methods for Measurement Error Correction

Multiple imputation has been proposed as a method to correct for measurement error (Cole et al., 2006), and to simultaneously handle missing data and measurement error (Ghosh-Dastidar and Schafer, 2003; Blackwell et al., 2017). However, much of the work focuses on continuous data and inference about regression parameters rather than population parameters such as total or occupancy rate defined in (1.4). These parameters are often the main interest in wildlife population surveys such as those used for stock assessment of red snapper.

Within the multiple imputation for measurement error (MIME) framework, the mismeasured count data are assumed to be missing. Therefore, they are imputed multiple (m) times with values randomly generated from $N|Y,X$ where X denotes the covariates in the data set. The resulting file consists of multiple imputed data sets from which estimates and their standard errors can be computed and then combined. Define Q to be the parameter of interest, and denote its estimate from the complete data, as well as its variance denoted by by $\hat{Q}$ and $\hat{U}$, respectively. Though by $\hat{Q}$ and $\hat{U}$, cannot be estimated because the complete data are not available, we can compute the estimators from the m imputed datasets, which we denote by $\hat{Q}_j$ and $\hat{U}_j$ for $j = 1, 2, 3, .., m$. According to Rubin's (1986) rules, the estimates and their variances are pooled as follows:

$$\bar{Q}_m = \sum_{j=1}^{m} \frac{\hat{Q}_j}{m} \quad \text{and} \quad V_m = \bar{U}_m + \frac{m+1}{m} B_m \tag{2.1}$$

where $\bar{U}_m = \sum_{j=1}^{m} \hat{U}_j/m$ and $B_m = \sum_{j=1}^{m} \, (\hat{Q}_j - \bar{Q}_m)^2/(m-1)$. In this section, we discuss several multiple imputation methods which can be used to handle visibility bias for estimating total abundance and occupancy rate. Those considered include imputation methods that are available off-the-shelf and our specialized imputation models: modified hot deck and Bayesian hierarchical model (HBZIP).

2.1. Available Imputation Methods

Imputation is a routine process when dealing with data which have missing values. Various imputation methods are available and generally can be classified into parametric, semi-parametric, and non-parametric methods. Parametric imputation methods make explicit assumptions about the distribution of the response variable. Therefore, the missing response is imputed with random draws from the posterior samples of the assumed distribution. Semi-parametric methods such as predictive mean matching (PMM) utilize normal regression models to find units with similar characteristics. However, it does not make any assumption about the distribution of the response itself. Similarly, non-parametric imputation borrows values from other units with the same characteristics. Unlike semi-parametric methods, these methods do not utilize any model or distribution to find similar units.

In this study, we consider several imputation types including: normal, Poisson (Raghunathan et al., 2001), and zero-inflated Poisson (Kleinke and Reinecke, 2013) regression models for parametric imputation; PMM (Rubin, 1986; Little, 1988) for semi-parametric imputation; and a modified hot deck for nonparametric imputation. In addition, we also consider a modified version of the predictive mean matching method called multiple imputation using distance aided selection (MIDAS) (Siddique and Belin, 2008; Gaffert et al., 2016). These methods are relatively easy to implement and available in R within the *mice* package (van Buuren et al., 2011) for normal and semi-parametric methods as well as *countimp* package (Kleinke and Reinecke, 2013) for Poisson and zero-inflated Poisson.

Note that in multiple imputation framework, we treat the observed count as a covariate; hence, it is possible that the imputed true count is less than the observed count ($N < Y$).

Therefore, these methods do not preserve the logical constraints of the data. Furthermore, the normal imputation method also does not preserve the discreteness of the imputed data, though the remaining methods considered in this study do. Imputed counts from the normal model may also be negative, which is logically impossible for count data. Intuitively, an easy fix would be to round or truncate the imputed values. However, in their study, Rodwell, Lee, Romaniuk, and Carlin (2014) suggest that post-imputation rounding increases bias in the estimates and inappropriately reduces the variance, while imputation with no rounding or transformation generally performs well.

Alternatively, count imputation models such as Poisson and ZIP may also be used. However, depending on the parameter of interest, it is unclear whether these models would yield preferable results compared to the normal model. Von Hippel (2013) conducted a study and concluded that normal imputation generally works well for estimating means, standard deviation, and regression parameters but performs poorly for estimating parameters that reflect distributional shape. This suggests that normal imputation may perform well for estimating total abundance but poorly when estimating occupancy rate. Hence, choosing the correct imputation model remains important especially if tail probabilities or shape of the distribution are of interest in the analysis.

2.2. Modified Hot Deck

As discussed in the previous section, hot deck imputation is a nonparametric method for handling missing data in survey. It is one of the three imputation methods used by the U.S. Census Bureau to impute item non-response in the Current Population Survey (U.S. Census Bureau, 2016). The hot deck procedure replaces missing values with a value from units with similar characteristics as measured by some distance metric (Andridge and Little, 2010). This allows the imputed data to maintain more realism, since missing data are imputed with real observations. In our data, for example, this would prevent negative or non-integer fish counts from being imputed.

Modified versions of the hot deck method have also been developed. Kim and Fuller 2004 developed a method called “fractional hot deck imputation” in which missing values are replaced with a set of weighted imputed values. Furthermore, McGee and Bergasa (2006) introduced a method called "modified nearest neighbor hot deck", where similar characteristics are chosen from the data and a small bit of noise is added to each observation to reflect sampling variability. In this section, we review the traditional hot deck imputation and then introduce our modified version of the hot deck method.

2.2.1. Hot Deck Imputation Review

Hot deck imputation involves two steps: forming a group of similar units for donor candidates (donor pool) and selecting a donor from the donor pool. The donor pool can be formed using several methods. One of the simplest methods is the adjustment cell method (Brick and Kalton, 1996) in which missing values are replaced with values from units with matching cells or covariates. For example, in population surveys, observations are matched using covariates such as sex, geographical location, and employment status. However, this method can be inefficient if units are sparse and the number of cells is large, requiring the same donor to be used multiple times. Another approach is to form the donor pool based on a distance measure such as Mahalonobis distance to determine similar units. Typically, the size of the donor pool (d) has been pre-determined such that only the d closest units are included in the donor pool. Alternatively, a threshold (δ) value may be used so that only observations with distance below δ are included in the pool. Once the donor pool is formed, a donor is then selected randomly from the pool with equal probability or probability inversely proportional to the distance.

As discussed in Section 2.1, PMM is an example of semi-parametric imputation method. However, PMM can also be thought of as a special case of hot deck imputation (Andridge and Little, 2010). In PMM, the distance is measured using the difference between predictive means of the missing unit and the donor candidates. In other words, they are matched based on the values of their predictive means. Typically, the means are predicted using a regression model. In addition, a modified version of PMM, the MIDAS approach, uses the same distance measure as PMM; however, rather than limiting the size of the donor pool, all units in the data are eligible as donors. For each missing unit, a donor is then selected randomly from the pool with probability inversely proportional to the distance.

Typically, hot deck imputation is implemented to directly impute the missing variable, which is N in our case. However, improvement may result from imputing a function of N rather than N directly (Andridge and Little, 2010). Suppose P is a function of N and an auxiliary variable (Y) highly correlated with N, defined by $P = Y/N$. The analyst may not have much information about N but have more information about P. Thus, hot deck imputation may be performed more effectively on P rather than N. In this case, the imputed values of P and auxiliary variable Y can then be used to obtain $\hat{N} = Y/\hat{P}$. Therefore, better results may yield from indirectly imputing N using $\hat{P}$ rather than directly imputing N. In our case, the variable P here can be thought as the proportion of fish visible to the observer or visibility rate, whereas Y is the number of fish observed.

2.2.2. Adaptation

In our adaptation, hot deck imputation is implemented twice: once to replace P and indirectly impute N where $\hat{N} = Y/\hat{P}$; and a second time to directly impute N when $\hat{P} = 0$ for which $\hat{N} = Y/\hat{P}$ is undefined. The underlying idea behind this method is that, depending on the value of $\hat{P}$, N is either imputed indirectly or directly. When $\hat{P} \in (0,1)$, N is mismeasured; therefore, it is corrected using indirect imputation $\hat{N} = Y/\hat{P}$. However, when N is missing ($\hat{P} = 0$), N must be directly imputed. For each imputation, procedures including forming the donor pool and selecting a donor follow the regular hot deck imputation procedure. Therefore, in its implementation, the analyst can easily implement this adaptation using available imputation packages such as *mice* within R.

Recall that $\hat{N}$ is undefined when $\hat{P} = 0$. In practice, $\hat{N}$ is also very unstable when $\hat{P}$ is near 0. This is because a slight change in $\hat{P}$ can greatly impact $\hat{N}$ when $\hat{P}$ is small. For example, consider a scenario where $Y = 5$, and two possible visibility rates $\hat{P} = 0.05$ or 0.025. The adjusted value is either $\hat{N} = 5/0.05 = 100$ or $5/0.025 = 200$. Therefore, a decrease in visibility rate by 0.025 significantly inflates the adjusted count. The same is not true when $\hat{P}$ is high. Suppose $\hat{P} = 0.95$ or 0.925; the adjusted value is either $\hat{N} = 5/0.95 = 5.26$ or $\hat{N} = 1/0.925 = 5.41$. Hence, the same decrease in $\hat{P}$ only increases $\hat{N}$ by a small amount when $\hat{P}$ is high. For this reason, direct imputation of N is applied not just when $\hat{P} = 0$, but when $\hat{P} < \delta_p$. Here, δ_p denotes the threshold level, near zero, below which $\hat{N}$ is directly imputed, and above which $\hat{N}$ is indirectly imputed.

2.3. The HBZIP Model

Visibility bias can be formulated as a classical measurement error problem in which the observed variable is an additive or multiplicative function of the true count. Given the nature of their relationship, the observed and true count data can also be modeled with hierarchical models (Royle and Dorazio, 2006). Often, a Poisson distribution is assumed for count data. However, in our application, the zero-inflated Poisson distribution is a more sensible choice considering the count data sparsity and the biology of Red Snapper. Some transects are chosen on habitats that cannot support Red Snapper and are therefore not "at risk" for containing any fish, while on others that are feasible habitat they may or may not appear. Based on the true count, the observed count can then be thought of as a realization of a binomial process in which only a fraction of the true count is expected to be observed. Therefore, it is reasonable to model the true count and observed count data with zero-inflated Poisson and binomial distribution respectively.

In the frequentist framework, parameters for the distributions are treated as fixed non-random quantities and must be estimated. In our application, the distributional parameters have a complex relationship that depends on the individual transects' characteristics. In addition, extra variability due to estimating the distributional parameters must also be accounted for in the estimation. Hence, the HBZIP model was developed to incorporate transect specific characteristics and extra variation from the unknown model parameters. The HBZIP is a four-level hierarchical model which includes the observed data model, count data model, priors for the count and observed model parameters, and hyperpriors for the parameters in the priors.

We implement the hierarchical models, not for fully Bayesian analysis of the parameter estimates, but for the purpose of imputation. Therefore, our goal here is not to obtain estimates directly via Bayesian estimation but to generate multiple completed data sets and allow the data users to perform the same type of analysis as they would for complete data. We utilize MCMC sampling techniques to generate random draws from the joint posterior of the hierarchical models. Within the imputation framework, estimates of total abundance, occupancy rate, and their associated standard errors are aggregated using Rubin's (1987) combining rules defined in (2.1). Theoretical development of the model and its Bayesian implementation are discussed in the next two sections.

2.3.1. Theoretical Development

Recall from the previous section that the HBZIP is a four-level hierarchical model which includes the observed data model, true count data model, priors for the distributional parameters, and hyperpriors for the parameters in the priors. The true count is assumed to follow zero-inflated Poisson (Lambert, 1992) with zero-inflation parameter π_i and expected local abundance λ_i for the i^{th} transect, denoted by $N_i \sim ZIP(\pi_i, \lambda_i)$. The probability mass function of N_i is defined by:

$$P(N_i = n) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i}, & n = 0 \\ (1-\pi_i)e^{-\lambda_i}\lambda_i^n(n_i!)^{-1}, & n = 1,2,3,.. \end{cases} \tag{2.2}$$

Given N_i and the visibility rate during the transect (P_i), the observed count (Y_i) is assumed to follow a binomial distribution. Therefore, the true and observed count data distributions are defined by:

$$N_i \sim ZIP(\pi_i, \lambda_i) \tag{2.3}$$

$$Y_i|N_i \sim Binomial(N_i, P_i) \tag{2.4}$$

We expect the model parameters to differ depending on the transect conditions and location. In addition, we assume normalized $\log(\lambda_i)$, $\text{logit}(\pi_i)$, and $\text{logit}(P_i)$ priors in the HBZIP model. However, assuming that the model parameters are known, we can obtain the conditional distribution of $N_i|Y_i$. Hence, our focus in this section is to derive the distribution of $N_i|Y_i$ and understand its characteristics for the different values of π_i, λ_i, and P_i.

Let $N_i \sim ZIP(\pi_i, \lambda_i)$ and $Y_i|N_i \sim Binomial(N_i, P_i)$; then the conditional distribution of $N|Y$ can be obtained following Bayes's theorem:

$$P(N_i|Y_i) = \frac{P(N_i, Y_i)}{P(Y_i)} = \frac{P(N_i, Y_i)P(N_i)}{\sum_{N=y}^{\infty} P(N_i, Y_i)P(N_i)} \tag{2.5}$$

The marginal distribution of Y_i, derived in Appendix A, also follows a zero-inflated distribution $Y_i \sim ZIP(\pi_i, \lambda_i p_i)$ given by:

$$P(Y_i = y) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i p_i}, & y = 0 \\ (1-\pi_i)\frac{e^{-\lambda_i p_i}(\lambda_i p_i)^y}{(y!)}, & y = 1, 2, 3, .. \end{cases} \tag{2.6}$$

Separating the two cases for $N = 0$ and $N > 0$, the joint distribution of N, Y is defined by:

$$P(N_i = n, Y_i = y) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i}, & n = 0; y = 0 \\ \frac{[(1-\pi_i)e^{-\lambda_i}\lambda_i^n]}{y!(n-y)!} p_i^y (1-p_i)^{n-y}, & n > 0; y = 0, 1, 2, .., n \end{cases} \tag{2.7}$$

Using Bayes theorem in (2.5), and the results from (2.6) and (2.7), the conditional distribution of $N|Y$, as shown in Appendix B, is:

$$P(N_i = n|Y_i = y) = \begin{cases} ZIP\left(\frac{\pi_i e^{\lambda_i p_i}}{\pi_i e^{\lambda_i p_i} + (1-\pi_i)}, \lambda_i(1-p_i)\right), & y = 0 \\ y + Poisson(\lambda_i[1-p_i]), & n = y, y+1, ...; y > 0 \end{cases} \tag{2.8}$$

This suggests that when no fish are observed, the conditional distribution $N|Y$ also follows ZIP but with updated parameters $\pi_i' = \frac{\pi_i e^{\lambda_i p_i}}{\pi_i e^{\lambda_i p_i} + (1-\pi_i)}$, and $\lambda' = \lambda[1-p]$. However, if at least one fish is observed, then the distribution of $N|Y$ follows a shifted Poisson distribution with mean λ' and shifting parameter y.

Alternatively, we can define $U_i = N_i - y_i$, the unobserved count, with the following distribution:

$$P(U_i = u|Y_i = y) = \begin{cases} ZIP\left(\frac{\pi_i e^{\lambda_i p_i}}{\pi_i e^{\lambda_i p_i} + (1-\pi_i)}, \lambda_i(1-p_i)\right), & y = 0 \\ Poisson(\lambda_i[1-p_i]), & n = y, y+1, ...; y > 0 \end{cases} \tag{2.9}$$

That is, U_i follows $ZIP(\pi_i', \lambda_i')$ when $y = 0$, and $Poisson(\lambda_i')$ when $y > 0$. Figure 2.1 illustrates the distribution of U for various values of P and λ. Based on the probability mass function in 2.9 and illustration in Figure 2.1, we obtain the following results:

- When visibility is clear, the observed count is the true count. As indicated by the distribution in green in Figure 2.1, $P(U = 0) = 1$ when $p = 1$ regardless of the value of y. Therefore, $N = y$ with probability 1 when $p = 1$.

- When visibility is obscured, the true count is adjusted additively since $N = y + U$. Furthermore, $E[U]$ also increases as visibility rate decreases which suggests that as visibility decreases, the expected unobserved count increases. This is shown by the shift in the distribution of U in each plot, from left to right as P decreases from 1 to 0.01.

- Finally, when visibility is lost ($P = 0$) and no fish are observed ($Y = 0$), as shown by

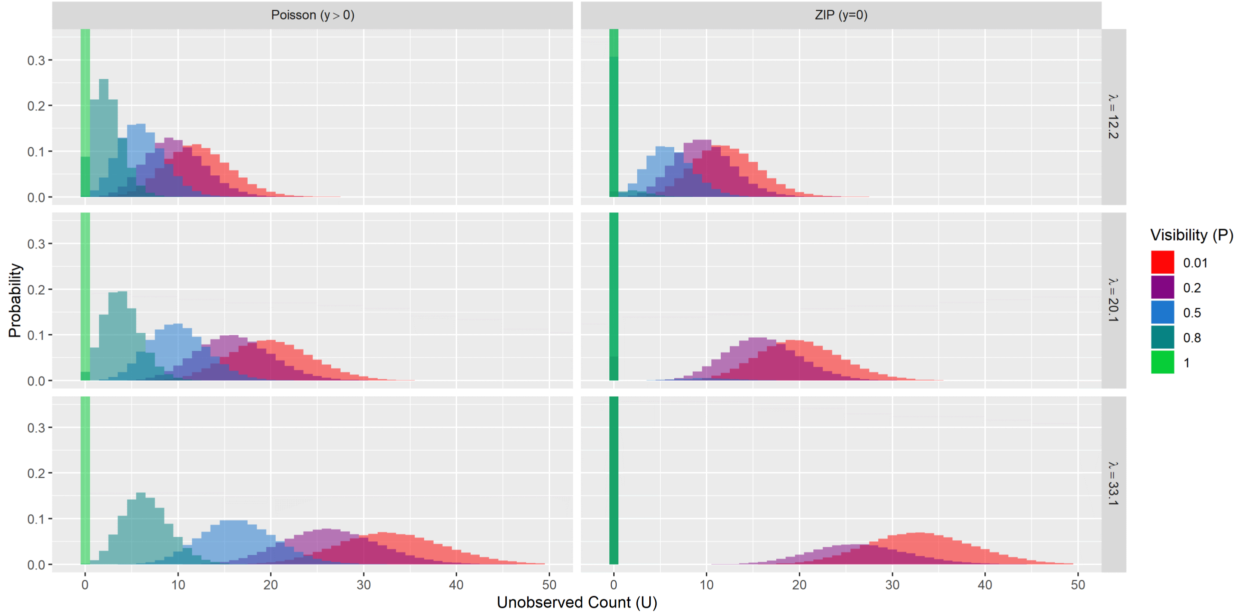


Figure 2.1: Distribution of the unobserved count (U) given the observed y value (column), λ (row), and visibility rate (color key). Each plot illustrates all possible values of U (x-axis) and their associated probability (y-axis).

the plots in the right column, $N|Y$ and N follow the same distribution. This is because $\pi' \rightarrow$ and $\lambda' \rightarrow \lambda$ as $p \rightarrow 0$. This implies that when visibility is lost, missing data are replaced with random draws from the marginal distribution of N, as illustrated by the distribution in red in the right column.

We showed here that the hierarchical models have desirable properties since they preserve the sparsity and discreteness of the data. In addition, the resulting conditional distribution is also consistent with our logical intuition since it allows the observed data to be retained when visibility is high, adjusted when visibility is degraded, and imputed when visibility is lost.

This leaves us with one remaining task: parameters estimation. Distributional parameters can be estimated using a maximum likelihood approach or using regression models with covariates. The maximum likelihood approach requires repeated measures to be available or the transects to be homogenous. This is unlikely in our application due to resources constraint, and since λ_i, π_i, and P_i are likely to vary among transects. Hence, we turn to a Bayesian approach to model the distributional parameters and obtain simulated samples of the true count from the joint posterior for imputation.

2.3.2. Bayesian Implementation

Within the Bayesian framework, normal regression priors are assumed for $\log(\lambda_i)$, $\text{logit}(\pi_i)$, and $\text{logit}(P_i)$ to allow their means to differ based on unit specific characteristics. The normal regression model for $\log(\lambda_i)$ is:

$$\phi_i = log(\lambda_i) \sim Normal(X_i^{(\phi)}\beta, \sigma_\phi^2) \tag{2.10}$$

where $X_i^{(\phi)}$ defines the vector of covariates impacting the mean of local abundance, $\beta = (\beta_o, \beta_1, .., \beta_p)^T$ represents the vector of regression coefficients, and σ_ϕ^2 reflects the extra-Poisson variation in local abundance. The normal regression prior for the $\text{logit}(\phi_i)$ and $\text{logit}(P_i)$ are:

$$\nu_i = logit(\pi_i) \sim Normal(X_i^{(\nu)}\gamma, \sigma_\nu^2), \tag{2.11}$$

$$\eta_i = logit(P_i) \sim Normal(X_i^{(\eta)}\theta, \sigma_\eta^2), \tag{2.12}$$

Similarly, $X_i^{(\nu)}$ and $X_i^{(\eta)}$ are vectors of covariates impacting the local occupancy rate and visibility rate respectively, while $\gamma = (\gamma_o, \gamma_1, .., \gamma_p)^T$ and $\theta = (\theta_o, \theta_1, .., \theta_p)^T$ are vectors representing their regression coefficients. Both σ_ν^2 and σ_η^2 reflect the extra variability which captures the heterogeneity of the local occupancy rate and visibility rate among the transects.

Mutually independent priors are assumed for the hyperparameters which include each regression coefficient in β, γ, and θ, and variance components $\sigma_\phi^2, \sigma_\nu^2, \sigma_\eta^2$. The priors are also set to be vague so that the posterior distribution is dominated by the data likelihood. Priors for each of the regression coefficients in β, γ, and θ are assumed to be mutually independent and follow $\text{Normal}(0, \sigma^2)$ with a large σ^2 value. In addition, uniform priors $U(0, A)$ with a large upper bound A are defined for σ_ϕ, σ_ν, and σ_η. A summary of the hierarchical model and its prior specifications are shown in Appendix C. The joint posterior distribution of the model parameters is proportional to the data likelihood and prior distributions specified above. Therefore, the joint posterior distribution of the hierarchical model's parameters is:

$$[N,\eta,\phi,\nu,\theta,\beta,\gamma,\sigma^2_\eta,\sigma^2_\phi,\sigma^2_\nu|Y,T,X] \propto$$
$$\left(\prod_{i=1}^{l} \underbrace{[Y_i|N_i,\eta_i]}_{\text{Binomial}}\underbrace{[N_i|\nu_i,\phi_i]}_{\text{ZIP}}\underbrace{[\eta_i|X_i^{(\eta)}\theta,\sigma^2_\eta][\phi_i|X_i^{(\phi)}\beta,\sigma^2_\phi][\nu_i|X_i^{(\nu)}\gamma,\sigma^2_\nu]}_{\text{normal regression}}\right)\underbrace{[\theta][\beta][\gamma][\sigma^2_\eta][\sigma^2_\phi][\sigma^2_\nu]}_{\text{hyperpriors}}$$

In its implementation, Bayesian simulation treats missing data similar to an unknown parameter; hence, simulated samples of missing and mismeasured values can be obtained from the joint posterior using Bayesian Markov Chain Monte Carlo (MCMC). This process generates multiple complete data set allowing analysis to proceed as if no data had been missing or mismeasured.

We introduced multiple imputation as a method to correct for measurement error in this chapter. In addition, we presented several imputation techniques including off-the-shelf imputation models such as normal, Poisson, zero-inflated Poisson, PMM, and PMM (MIDAS) and our developed imputation models, modified hot deck and HBZIP. These methods can be used to correct measurement error in zero-inflated count data; however, their performance remain to be evaluated. In the next chapter, we examine the methods' performance for estimating total abundance and habitat occupancy rate through a simulation study.

Chapter 3
Simulation Study

In this chapter, we present the results of a simulation study to evaluate the performance of multiple imputation for measurement error (MIME) methods under various count data model and visibility mechanisms. Note that we distinguish the term visibility mechanism from visibility model here since we use the former to refer to the actual relationship between T and P, while the latter is used to describe the assumed relationship between T and P within the HBZIP model. Therefore, visibility model within the HBZIP model may be inconsistent with the actual visibility mechanism during the transect.

To reflect the characteristics of wildlife abundance, we generated the data from a two-level hierarchical model which specifies the count data in the first level and observed data in the second level. Distributional parameters are modeled according to the regression model in (2.10)-(2.12). The simulation experiment was a three-factor design, with two factors controlling the count data model parameters and one controlling the visibility mechanism. Each of the factors was set at three levels, resulting in 27 parameter combinations. We then implemented imputation procedures including off-the-shelf methods (normal, Poisson, ZIP, PMM, and PMM MIDAS), and our adapted methods (modified hot deck imputation and HBZIP). We evaluated the methods' performance in handling visibility bias for the purpose of estimating total abundance (τ) and occupancy rate (ρ). Performance here was evaluated based on several measures including relative bias, relative standard error, relative mean squared error, and confidence interval coverage defined later in Section 3.2.

3.1. Simulation Parameters

Recall that the data were generated from a two-level hierarchical model. In the first level, we generated the count data from zero-inflated Poisson $N \sim ZIP(\pi,\lambda)$. Conditioned on N and P, the observed count (Y) was assumed to follow a distribution, $Y|N,P \sim Binomial(N,P)$. We assumed Poisson regression for λ, similar to (2.10), and allowed it to differ based on two regions, such that for the i^{th} transect:

$$\phi_i = log(\lambda_i = \beta_o + 0.5(area_{2_i}) \tag{3.1}$$

for $i = 1,2,3,..I$. In addition, we specified a logistic regression for P to model the linear relationship between T and logit(P) as follows:

$$\nu_i = logit(P_i) = 5 - 10(T_i) \tag{3.2}$$

Based on the logistic model above, we then modified the linear visibility model to obtain discrete and bounded visibility model as follows:

$$P_{discrete} = \begin{cases} 1, & t_i \leq 0.2 \\ 0.7, & 0.2 < t_i \leq 0.5 \\ 0.3, & 0.5 < t_i \leq 0.8 \\ 0, & t_i > 0.8 \end{cases} \quad \text{and} \quad P_{bounded} = \begin{cases} 1, & t_i \leq 0.2 \\ 0.5, & 0.2 < t_i \leq 0.8 \\ 0, & t_i > 0.8 \end{cases}$$

Figure 3.1 illustrates the three visibility mechanisms simulated in this study. Various mechanisms were considered here since the true relationship between turbidity and visibility in our application is unknown and may vary depending on the data collection strategy.

In the simulation, we varied the Poisson-regression intercept ($\beta_o = 2.5, 3, 3.5$) which indirectly controls the value of λ, the zero-inflation parameter ($\pi = 0.35, 0.5, 0.7$) which affects the occupancy rate (ρ), and the visibility model (linear, discrete, and bounded). Under each

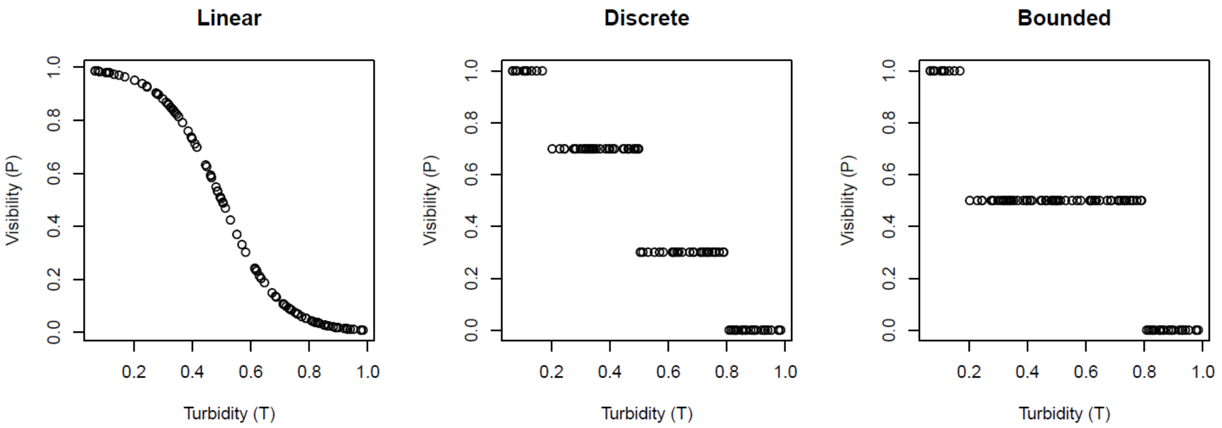


Figure 3.1: Three visibility mechanisms based on turbidity level: linear, discrete, and bounded.

of the 27 parameter combinations, $M = 1000$ sample replicates were simulated. For each replicate, a sample of size 100 samples were generated following a simple random sample. The true counts were observable in 40 of the 100 cases, leaving 60 mismeasured counts to be imputed. Mismeasured counts are then imputed 10 times (m) and used to produce estimates of τ and ρ.

Figure 3.2 illustrates the empirical cumulative distribution function (ECDF) of a randomly generated and multiply imputed sample. The different lines indicate imputed data (colored lines), complete data (black line), and observed data as shown by the naïve method (grey line). As shown in the figure, the HBZIP model, PMM, PMM MIDAS approach appear to reproduce the ECDF best. These distributions closely estimate $P(N = 0)$; hence, we expect them to perform well for estimating habitat occupancy rate. The normal model, on the other hand, underestimates $P(N = 0)$; therefore, we expect the model to overestimate the habitat occupancy rate $P(N > 0)$. We will see that imputed data may not reflect the actual data distribution but still yield valid abundance and occupancy rate estimates. Conversely, imputed data may reflect the actual data distribution well but produce biased or unstable estimates. Therefore, in the simulation, the methods are evaluated based on several criteria for inference purposes discussed in the next section.

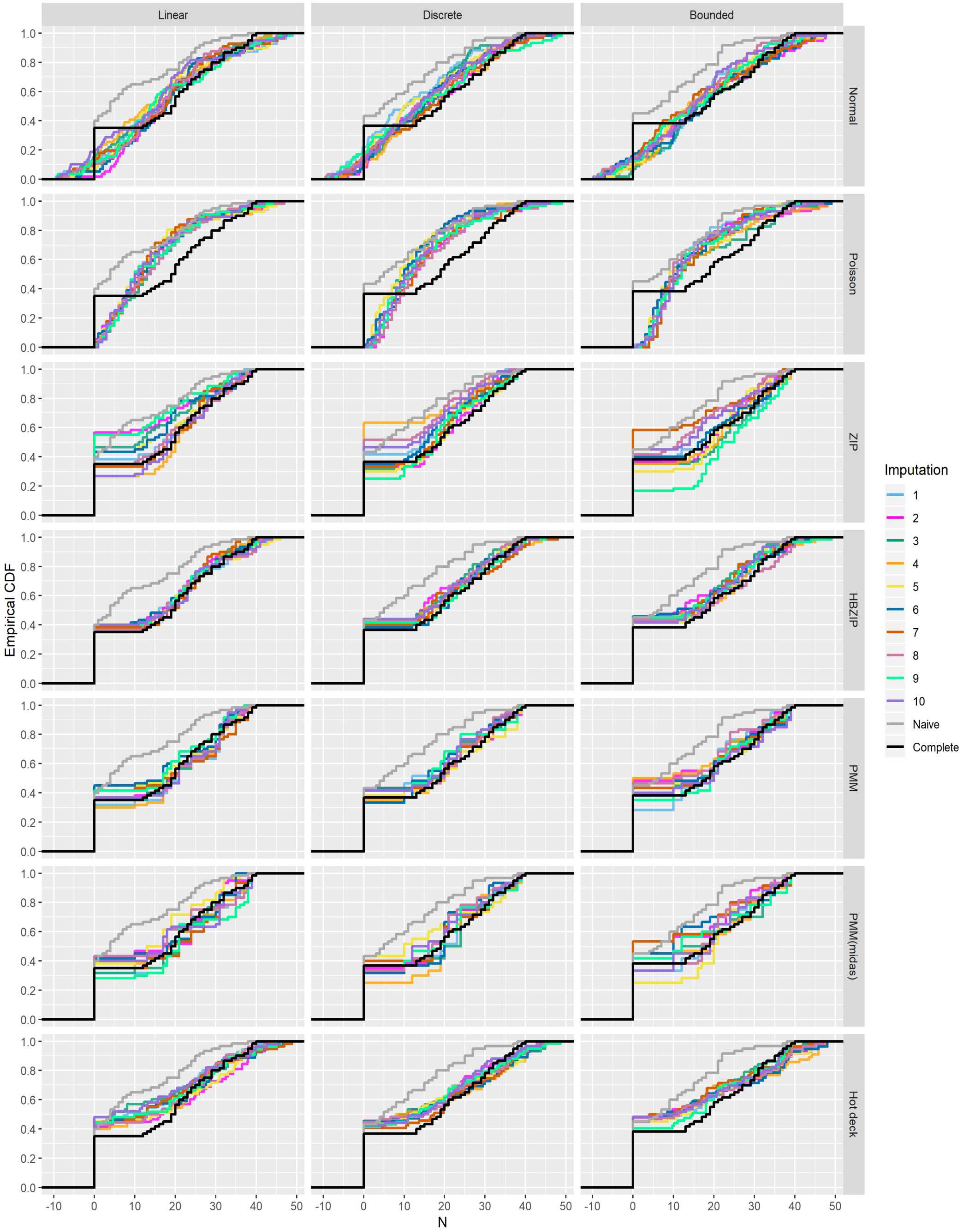


Figure 3.2: Empirical cumulative distribution function (ECDF) of the complete, observed, and multiply imputed data by visibility mechanism (column), and imputation method (row) when $\beta_o = 3$ and $\pi = 0.35$.

3.2. Criteria to Compare Results

We evaluated the performance of various imputation methods for estimating total abundance and occupancy rate defined in (1.4). Let $\theta =$ parameter of interest in the simulation such that $\theta = \{\tau, \rho\}$, and $M =$ simulation replicates. To evaluate the methods' performance, we monitored the estimator's relative bias (RB), relative standard error (RSE), relative root mean squared errors ($RMSE$), and 95% confidence interval calculated as follows:

$$Rel.Bias(\hat{\theta}) = \frac{1}{M}\sum_{j}^{M}(\hat{\theta}_j - \theta)/\theta$$

$$Rel.SE(\hat{\theta}) = \frac{1}{M}\sum_{j}^{M} SE[\hat{\theta}_j]/\theta$$

$$Rel.RMSE(\hat{\theta}) = \frac{1}{M}\sum_{j}^{M}\left[(\hat{\theta}_j - \theta)^2 + Var[\hat{\theta}_j\right]^{1/2}/\theta$$

$$95\%\,CI\,Coverage(\hat{\theta}) = \frac{1}{M}\sum_{j}^{M} I_{[\hat{\theta} - Z_{\alpha/2}SE[\hat{\theta}] < \theta < \hat{\theta} + Z_{\alpha/2}SE[\hat{\theta}]]}$$

Parameter estimates ($\hat{\theta}$) and their associated standard errors ($SE[\hat{\theta}]$) were calculated from the imputed data following Rubin's (1987) rules defined in (2.1).

3.3. Simulation Results

Overall simulation results by visibility mechanism are illustrated in Figure 3.3 for total abundance and Figure 3.4 for habitat occupancy rate. Within the simulation, we also produced estimates from the complete data to illustrate the estimators' performance if no data had been missing or mismeasured. Naive estimates were also produced to demonstrate the estimator's properties if visibility bias is ignored. As shown in both figures, estimates from the complete data analysis suggest that in the absence of visibility bias, the estimates are approximately unbiased ($RB \approx 0$), with confidence interval coverages near their nominal value ($Cov \approx 0.95$). This suggests that sampling distribution of the estimators are approximately normal.

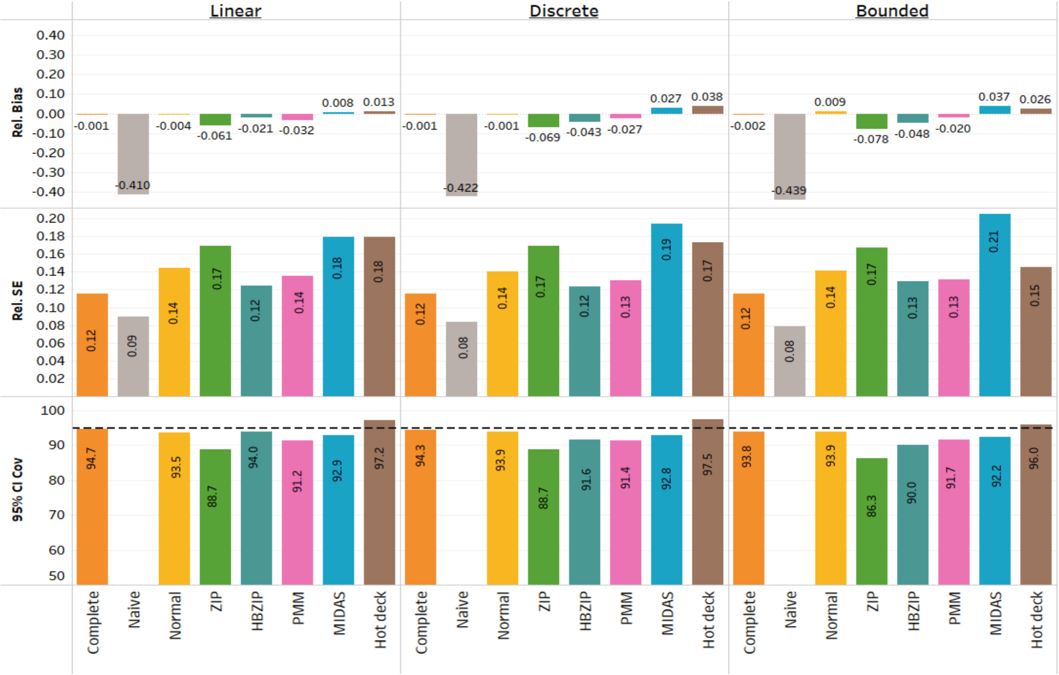


Figure 3.3: Simulation results for total abundance estimates by visibility mechanism for all β_o and π combined. Results from the Poisson model are not presented here due to their highly biased estimates and extreme standard errors.

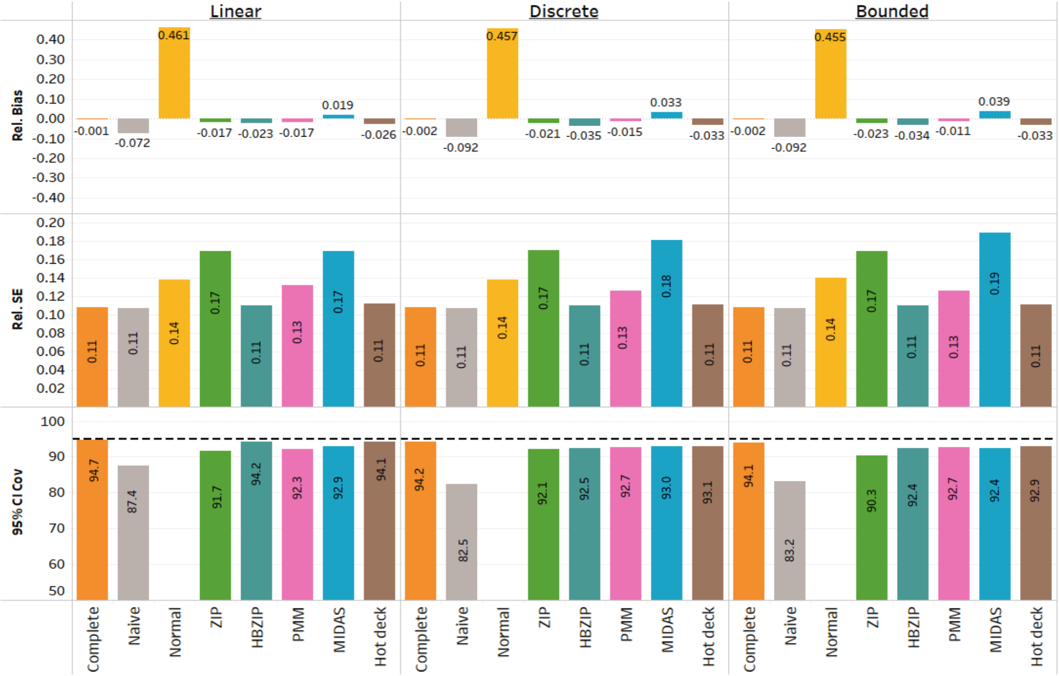


Figure 3.4: Simulation results for occupancy rate estimates by visibility mechanism for all β_o and π combined. Results from the Poisson model are not presented here due to their highly biased estimates and extreme standard errors.

Furthermore, as shown by the naïve estimates, when visibility bias is ignored, total abundance and occupancy rate are underestimated ($RB[\hat{\tau}] \approx -0.41$ to -0.44 and $RB[\hat{\rho}] = -0.07$ to -0.09) with confidence interval coverages that are far below their nominal value. Therefore, we implemented various imputation methods here to improve inference about the parameters and obtain results that are closer to those of complete data. In this section, we present and discuss our simulation results by visibility mechanism.

3.3.1. Linear Visibility

Simulation results for the linear visibility mechanism are summarized in Table 3.1 for abundance estimation and Table 3.2 for occupancy rate estimation. The results show that the methods' performance varies by zero inflation parameter (π) and count data model intercept (β_o). However, when visibility is linear, the HBZIP model outperforms the other methods. Compared to the other methods, estimates from the HBZIP model have a relatively low bias ($RB[\hat{\tau}] = RB[\hat{\rho}] < 0.034$ in magnitude), low standard error, and confidence interval coverage that is near its nominal level ($Cov[\hat{\tau}] = Cov[\hat{\rho}] = 0.925$ to 0.956). The method also yields $RMSEs$ that are the closest to those of complete data. This suggests that when the count data and visibility model in HBZIP are correctly specified, the overall accuracy and precision of the resulting estimates are not significantly lost due to the missing and mismeasured data.

Hot deck imputation also produces estimates with low bias overall ($RB[\hat{\tau}] = RB[\hat{\rho}] < 0.04$ in magnitude). However, it also produces high $RSE[\hat{\tau}]$ which inappropriately increases the confidence interval coverage ($Cov[\hat{\tau}] > 0.965$). Alternatively, PMM MIDAS also shows favorable results for both parameters when the data are not highly sparse ($\pi \leq 0.5$). Under this condition, estimates from PMM MIDAS have low bias ($RB[\hat{\tau}] = RB[\hat{\rho}] < 0.022$), slightly higher RSE compared to HBZIP by about 0.01 to 0.04, and near nominal confidence interval.

π	Method	Rel. Bias			Rel. SE			Rel. RMSE			95% CI Cov		
			β_o			β_o			β_o			β_o	
		2.5	3	3.5	2.5	3	3.5	2.5	3	3.5	2.5	3	3.5
0.35	Complete	0.001	0.001	0.002	0.08	0.08	0.08	0.11	0.11	0.11	94.8	94.9	94.4
	Naive	-0.409	-0.41	-0.411	0.07	0.07	0.07	0.41	0.42	0.42	0	0	0
	Normal	0.008	0.002	-0.001	0.1	0.1	0.1	0.14	0.14	0.14	93	93.3	92.9
	Poisson	0.062	0.036	0.027	0.12	0.1	0.09	0.18	0.15	0.14	87.5	86.7	85
	ZIP	-0.046	-0.042	-0.031	0.12	0.12	0.12	0.17	0.17	0.17	90	92.2	91.5
	HBZIP	-0.027	-0.031	-0.03	0.09	0.08	0.08	0.12	0.12	0.11	94.6	92.7	93.1
	PMM	-0.016	-0.008	-0.008	0.1	0.1	0.1	0.14	0.13	0.13	93	91.6	92.6
	MIDAS	0	0.007	0.014	0.1	0.1	0.1	0.13	0.13	0.13	95.3	94.8	92
	Hot deck	0	-0.001	-0.004	0.13	0.12	0.12	0.16	0.16	0.15	97.2	96.5	97.1
0.5	Complete	0.005	-0.001	-0.002	0.11	0.11	0.11	0.15	0.14	0.14	95.8	94.1	94.6
	Naive	-0.406	-0.411	-0.414	0.08	0.08	0.08	0.42	0.42	0.42	0.7	0.3	0.4
	Normal	-0.003	-0.005	-0.007	0.14	0.13	0.13	0.18	0.18	0.18	94.7	94.5	93.5
	Poisson	0.138	0.074	0.077	0.21	0.16	0.15	0.32	0.24	0.24	91.1	86.8	86.6
	ZIP	-0.066	-0.056	-0.044	0.16	0.16	0.16	0.23	0.23	0.23	88.5	89.8	88.2
	HBZIP	-0.014	-0.03	-0.033	0.11	0.11	0.11	0.15	0.15	0.15	95	93.9	93.4
	PMM	-0.022	-0.022	-0.022	0.13	0.13	0.13	0.18	0.18	0.17	92.2	92.7	92.5
	MIDAS	0.009	0.002	-0.008	0.14	0.14	0.14	0.19	0.19	0.19	94.5	94.9	93.1
	Hot deck	0.01	0.004	0	0.16	0.16	0.16	0.21	0.2	0.2	97.5	96.7	96.6
0.7	Complete	-0.008	0.001	-0.005	0.16	0.16	0.16	0.22	0.22	0.22	95.5	94	94.1
	Naive	-0.414	-0.406	-0.412	0.12	0.12	0.12	0.43	0.43	0.43	10.1	11.5	9.3
	Normal	-0.012	-0.001	-0.016	0.2	0.2	0.2	0.27	0.27	0.27	93.2	93.6	93.2
	Poisson	692.1	12.57	7.587	1,958	19.93	10.78	2,085	23.78	13.31	91.6	90	88.8
	ZIP	-0.101	-0.078	-0.081	0.23	0.23	0.23	0.34	0.34	0.34	85.4	86.7	86.4
	HBZIP	0.016	-0.016	-0.027	0.21	0.17	0.16	0.27	0.23	0.22	95	94.2	93.8
	PMM	-0.067	-0.055	-0.064	0.18	0.18	0.18	0.26	0.26	0.26	89.4	89.5	87.7
	MIDAS	0.064	0.022	-0.039	0.32	0.3	0.28	0.41	0.39	0.39	94.3	90.2	86.8
	Hot deck	0.037	0.039	0.031	0.26	0.25	0.25	0.32	0.31	0.31	98	97.9	97.3

Table 3.1: Simulation results for total abundance estimates for all β_o and ϕ, when visibility is linear.

The Poisson model performs poorly for both estimands. Its performance progressively worsens as π increases which suggests that the Poisson model is inappropriate for imputing sparse data. The ZIP model should also be avoided since its estimates can be highly biased ($RB[\hat{\tau}] > 0.08$ in magnitude) and have interval coverages that are below 0.90. As seen in Table 3.1, the normal imputation model shows favorable results for τ estimation. It produces low bias with $RB[\hat{\tau}] < 0.017$ in magnitude, slightly higher $RSE[\hat{\tau}]$ compared to the HBZIP, and near nominal confidence interval coverage. However, the normal model is highly biased for estimating ρ . Its bias increases as the data becomes more sparse, $RB[\hat{\rho}] \approx 0.2$ when $\pi = 0.35$ to $RB[\hat{\rho}] \approx 0.8$ when $\pi = 0.70$. Hence, the normal model should be avoided for occupancy rate estimation. This suggests that for estimating totals, which is equivalent to estimating means, the normal model is insensitive to misspecification of the data distribution. Multiple studies have shown that the normal imputation model yields consistent estimates for population means and variances even when the data are non-normal (von Hippel and T., 2013; He and Raghunathan, 2006). Therefore, even though distribution of the imputed

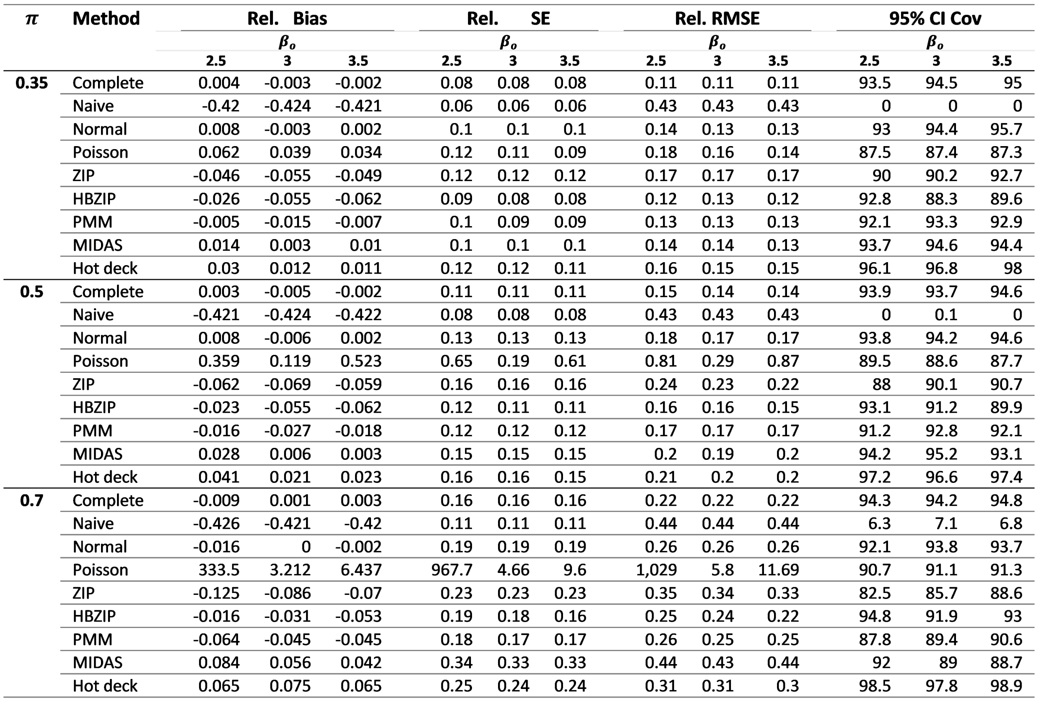
π	Method	Rel. Bias			Rel. SE			Rel. RMSE			95% CI Cov		
			β_o			β_o			β_o			β_o	
		2.5	3	3.5	2.5	3	3.5	2.5	3	3.5	2.5	3	3.5
0.35	Complete	0.001	0.001	0.002	0.07	0.07	0.07	0.1	0.1	0.1	0.001	0.001	0.002
	Naive	-0.091	-0.067	-0.053	0.08	0.07	0.07	0.13	0.12	0.11	-0.091	-0.067	-0.053
	Normal	0.21	0.209	0.211	0.08	0.08	0.08	0.23	0.22	0.23	0.21	0.209	0.211
	Poisson	0.299	0.308	0.319	0.06	0.06	0.05	0.31	0.31	0.32	0.299	0.308	0.319
	ZIP	-0.008	-0.015	-0.01	0.11	0.12	0.11	0.16	0.16	0.16	-0.008	-0.015	-0.01
	HBZIP	-0.024	-0.031	-0.03	0.08	0.07	0.07	0.1	0.1	0.1	-0.024	-0.031	-0.03
	PMM	0	0.003	0.002	0.09	0.09	0.09	0.13	0.12	0.13	0	0.003	0.002
	MIDAS	0.008	0.013	0.021	0.09	0.09	0.09	0.12	0.12	0.12	0.008	0.013	0.021
	Hot deck	-0.027	-0.021	-0.017	0.08	0.08	0.08	0.11	0.11	0.11	-0.027	-0.021	-0.017
0.5	Complete	0.005	-0.001	-0.001	0.1	0.1	0.1	0.13	0.14	0.14	0.005	-0.001	-0.001
	Naive	-0.089	-0.069	-0.056	0.1	0.1	0.1	0.16	0.15	0.14	-0.089	-0.069	-0.056
	Normal	0.365	0.366	0.37	0.12	0.12	0.12	0.38	0.39	0.39	0.365	0.366	0.37
	Poisson	0.486	0.532	0.563	0.1	0.09	0.09	0.5	0.54	0.57	0.486	0.532	0.563
	ZIP	-0.013	-0.017	-0.013	0.16	0.15	0.16	0.21	0.22	0.22	-0.013	-0.017	-0.013
	HBZIP	-0.01	-0.03	-0.033	0.1	0.10	0.1	0.14	0.14	0.14	-0.01	-0.03	-0.033
	PMM	-0.003	-0.009	-0.01	0.12	0.12	0.12	0.17	0.17	0.17	-0.003	-0.009	-0.01
	MIDAS	0.02	0.012	0.003	0.13	0.13	0.13	0.17	0.17	0.17	0.02	0.012	0.003
	Hot deck	-0.029	-0.027	-0.025	0.10	0.10	0.10	0.14	0.14	0.14	-0.029	-0.027	-0.025
0.7	Complete	-0.006	0	-0.006	0.15	0.15	0.15	0.2	0.21	0.21	-0.006	0	-0.006
	Naive	-0.098	-0.07	-0.058	0.15	0.15	0.15	0.22	0.21	0.21	-0.098	-0.07	-0.058
	Normal	0.801	0.809	0.805	0.22	0.22	0.22	0.83	0.84	0.84	0.801	0.809	0.805
	Poisson	0.821	0.978	1.086	0.2	0.19	0.18	0.85	1.00	1.11	0.821	0.978	1.086
	ZIP	-0.024	-0.018	-0.033	0.24	0.24	0.24	0.33	0.33	0.33	-0.024	-0.018	-0.033
	HBZIP	-0.005	-0.018	-0.027	0.16	0.15	0.15	0.21	0.21	0.21	-0.005	-0.018	-0.027
	PMM	-0.043	-0.04	-0.049	0.18	0.18	0.18	0.25	0.25	0.25	-0.043	-0.04	-0.049
	MIDAS	0.081	0.034	-0.026	0.3	0.29	0.28	0.39	0.38	0.38	0.081	0.034	-0.026
	Hot deck	-0.037	-0.024	-0.026	0.16	0.16	0.15	0.21	0.21	0.21	-0.037	-0.024	-0.026

Table 3.2: Simulation results for habitat occupancy rate estimates for all β_o and ϕ, when visibility is linear.

values is inconsistent with the underlying data distribution, the normal imputation model remains effective for inference about the mean. As suggested by Rubin (1996) the goal of imputation is not to recreate individual missing values but to ensure the validity of inference.

3.3.2. Discrete Visibility

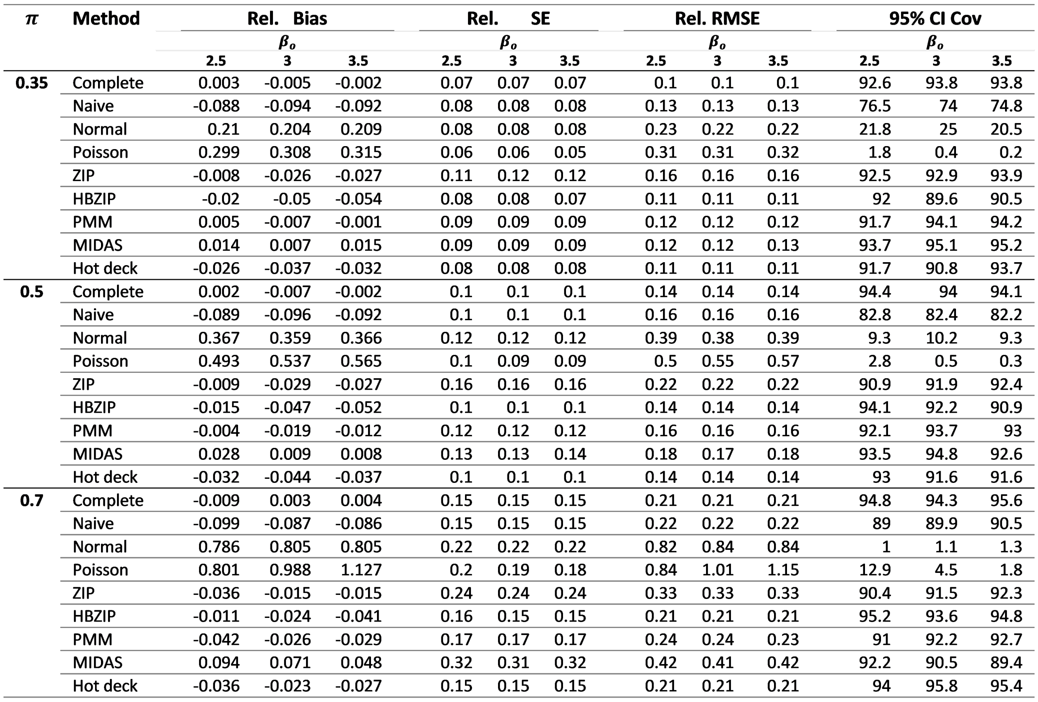
Unlike the linear visibility mechanism, no single method appears to outperform the other methods when visibility mechanism follows the discrete model. The performance of the methods varies significantly depending on the data sparsity as indicated by π. As shown in Table 3.3 and Table 3.4, when the data are not highly sparse ($\pi \leq 0.5$), PMM MIDAS outperforms the other methods for both total abundance and occupancy rate. Though its standard error is slightly higher, PMM MIDAS produces relatively low bias ($RB[\hat{\tau}] < 0.029$) and confidence interval coverage near its nominal value ($Cov[\hat{\tau}] = Cov[\hat{\rho}] = 0.926$ to 0.952). The regular PMM also produces low bias when $\pi < 0.5$. Its RSE, however, is slightly lower which results in a worse confidence interval coverage overall compared to the MIDAS approach. When the data are highly sparse ($\pi = 0.7$), both PMM and PMM MIDAS can have significant bias ($RB[\hat{\tau}] = RB[\hat{\rho}] > 0.06$) and yield confidence interval coverage that are below 0.90.



π	Method	Rel. Bias			Rel. SE			Rel. RMSE			95% CI Cov		
			β_o			β_o			β_o			β_o	
		2.5	3	3.5	2.5	3	3.5	2.5	3	3.5	2.5	3	3.5
0.35	Complete	0.004	-0.003	-0.002	0.08	0.08	0.08	0.11	0.11	0.11	93.5	94.5	95
	Naive	-0.42	-0.424	-0.421	0.06	0.06	0.06	0.43	0.43	0.43	0	0	0
	Normal	0.008	-0.003	0.002	0.1	0.1	0.1	0.14	0.13	0.13	93	94.4	95.7
	Poisson	0.062	0.039	0.034	0.12	0.11	0.09	0.18	0.16	0.14	87.5	87.4	87.3
	ZIP	-0.046	-0.055	-0.049	0.12	0.12	0.12	0.17	0.17	0.17	90	90.2	92.7
	HBZIP	-0.026	-0.055	-0.062	0.09	0.08	0.08	0.12	0.13	0.12	92.8	88.3	89.6
	PMM	-0.005	-0.015	-0.007	0.1	0.09	0.09	0.13	0.13	0.13	92.1	93.3	92.9
	MIDAS	0.014	0.003	0.01	0.1	0.1	0.1	0.14	0.14	0.13	93.7	94.6	94.4
	Hot deck	0.03	0.012	0.011	0.12	0.12	0.11	0.16	0.15	0.15	96.1	96.8	98
0.5	Complete	0.003	-0.005	-0.002	0.11	0.11	0.11	0.15	0.14	0.14	93.9	93.7	94.6
	Naive	-0.421	-0.424	-0.422	0.08	0.08	0.08	0.43	0.43	0.43	0	0.1	0
	Normal	0.008	-0.006	0.002	0.13	0.13	0.13	0.18	0.17	0.17	93.8	94.2	94.6
	Poisson	0.359	0.119	0.523	0.65	0.19	0.61	0.81	0.29	0.87	89.5	88.6	87.7
	ZIP	-0.062	-0.069	-0.059	0.16	0.16	0.16	0.24	0.23	0.22	88	90.1	90.7
	HBZIP	-0.023	-0.055	-0.062	0.12	0.11	0.11	0.16	0.16	0.15	93.1	91.2	89.9
	PMM	-0.016	-0.027	-0.018	0.12	0.12	0.12	0.17	0.17	0.17	91.2	92.8	92.1
	MIDAS	0.028	0.006	0.003	0.15	0.15	0.15	0.2	0.19	0.2	94.2	95.2	93.1
	Hot deck	0.041	0.021	0.023	0.16	0.16	0.15	0.21	0.2	0.2	97.2	96.6	97.4
0.7	Complete	-0.009	0.001	0.003	0.16	0.16	0.16	0.22	0.22	0.22	94.3	94.2	94.8
	Naive	-0.426	-0.421	-0.42	0.11	0.11	0.11	0.44	0.44	0.44	6.3	7.1	6.8
	Normal	-0.016	0	-0.002	0.19	0.19	0.19	0.26	0.26	0.26	92.1	93.8	93.7
	Poisson	333.5	3.212	6.437	967.7	4.66	9.6	1,029	5.8	11.69	90.7	91.1	91.3
	ZIP	-0.125	-0.086	-0.07	0.23	0.23	0.23	0.35	0.34	0.33	82.5	85.7	88.6
	HBZIP	-0.016	-0.031	-0.053	0.19	0.18	0.16	0.25	0.24	0.22	94.8	91.9	93
	PMM	-0.064	-0.045	-0.045	0.18	0.17	0.17	0.26	0.25	0.25	87.8	89.4	90.6
	MIDAS	0.084	0.056	0.042	0.34	0.33	0.33	0.44	0.43	0.44	92	89	88.7
	Hot deck	0.065	0.075	0.065	0.25	0.24	0.24	0.31	0.31	0.3	98.5	97.8	98.9

Table 3.3: Simulation results for total abundance estimates for all β_o and ϕ, when visibility follows the discrete model.

When the data are highly sparse ($\pi = 0.7$), the HBZIP model shows favorable results compared to the other methods. HBZIP estimates have relatively low bias ($RB[\hat{\tau}] = RB[\hat{\rho}] < 0.054$ in magnitude) and the closest $RMSE$ to those of complete data. Under this condition, the highest bias is observed when $\beta_o = 3.5$ for which $RB[\hat{\tau}] = -0.053$. As shown in Table 3.3 and Table 3.4 when $\pi = 0.7$, bias from HBZIP estimates increases progressively as β_o increases— with $RB[\hat{\tau}] = -0.016, -0.031, -0.053$ and $RB[\hat{\rho}] = -0.011, -0.024, -0.041$ when $\beta_o = 2.5, 3, 3.5$ respectively. However, the HBZIP method maintains near 95% confidence interval coverage while the other methods perform worse than the HBZIP under this condition. Finally, both Poisson and ZIP imputation either yield highly biased estimates or poor



π	Method	Rel. Bias			Rel. SE			Rel. RMSE			95% CI Cov		
		β_o			β_o			β_o			β_o		
		2.5	3	3.5	2.5	3	3.5	2.5	3	3.5	2.5	3	3.5
0.35	Complete	0.003	-0.005	-0.002	0.07	0.07	0.07	0.1	0.1	0.1	92.6	93.8	93.8
	Naive	-0.088	-0.094	-0.092	0.08	0.08	0.08	0.13	0.13	0.13	76.5	74	74.8
	Normal	0.21	0.204	0.209	0.08	0.08	0.08	0.23	0.22	0.22	21.8	25	20.5
	Poisson	0.299	0.308	0.315	0.06	0.06	0.05	0.31	0.31	0.32	1.8	0.4	0.2
	ZIP	-0.008	-0.026	-0.027	0.11	0.12	0.12	0.16	0.16	0.16	92.5	92.9	93.9
	HBZIP	-0.02	-0.05	-0.054	0.08	0.08	0.07	0.11	0.11	0.11	92	89.6	90.5
	PMM	0.005	-0.007	-0.001	0.09	0.09	0.09	0.12	0.12	0.12	91.7	94.1	94.2
	MIDAS	0.014	0.007	0.015	0.09	0.09	0.09	0.12	0.12	0.13	93.7	95.1	95.2
	Hot deck	-0.026	-0.037	-0.032	0.08	0.08	0.08	0.11	0.11	0.11	91.7	90.8	93.7
0.5	Complete	0.002	-0.007	-0.002	0.1	0.1	0.1	0.14	0.14	0.14	94.4	94	94.1
	Naive	-0.089	-0.096	-0.092	0.1	0.1	0.1	0.16	0.16	0.16	82.8	82.4	82.2
	Normal	0.367	0.359	0.366	0.12	0.12	0.12	0.39	0.38	0.39	9.3	10.2	9.3
	Poisson	0.493	0.537	0.565	0.1	0.09	0.09	0.5	0.55	0.57	2.8	0.5	0.3
	ZIP	-0.009	-0.029	-0.027	0.16	0.16	0.16	0.22	0.22	0.22	90.9	91.9	92.4
	HBZIP	-0.015	-0.047	-0.052	0.1	0.1	0.1	0.14	0.14	0.14	94.1	92.2	90.9
	PMM	-0.004	-0.019	-0.012	0.12	0.12	0.12	0.16	0.16	0.16	92.1	93.7	93
	MIDAS	0.028	0.009	0.008	0.13	0.13	0.14	0.18	0.17	0.18	93.5	94.8	92.6
	Hot deck	-0.032	-0.044	-0.037	0.1	0.1	0.1	0.14	0.14	0.14	93	91.6	91.6
0.7	Complete	-0.009	0.003	0.004	0.15	0.15	0.15	0.21	0.21	0.21	94.8	94.3	95.6
	Naive	-0.099	-0.087	-0.086	0.15	0.15	0.15	0.22	0.22	0.22	89	89.9	90.5
	Normal	0.786	0.805	0.805	0.22	0.22	0.22	0.82	0.84	0.84	1	1.1	1.3
	Poisson	0.801	0.988	1.127	0.2	0.19	0.18	0.84	1.01	1.15	12.9	4.5	1.8
	ZIP	-0.036	-0.015	-0.015	0.24	0.24	0.24	0.33	0.33	0.33	90.4	91.5	92.3
	HBZIP	-0.011	-0.024	-0.041	0.16	0.15	0.15	0.21	0.21	0.21	95.2	93.6	94.8
	PMM	-0.042	-0.026	-0.029	0.17	0.17	0.17	0.24	0.24	0.23	91	92.2	92.7
	MIDAS	0.094	0.071	0.048	0.32	0.31	0.32	0.42	0.41	0.42	92.2	90.5	89.4
	Hot deck	-0.036	-0.023	-0.027	0.15	0.15	0.15	0.21	0.21	0.21	94	95.8	95.4

Table 3.4: Simulation results for habitat occupancy rate estimates for all β_o and ϕ, when visibility follows the discrete model.

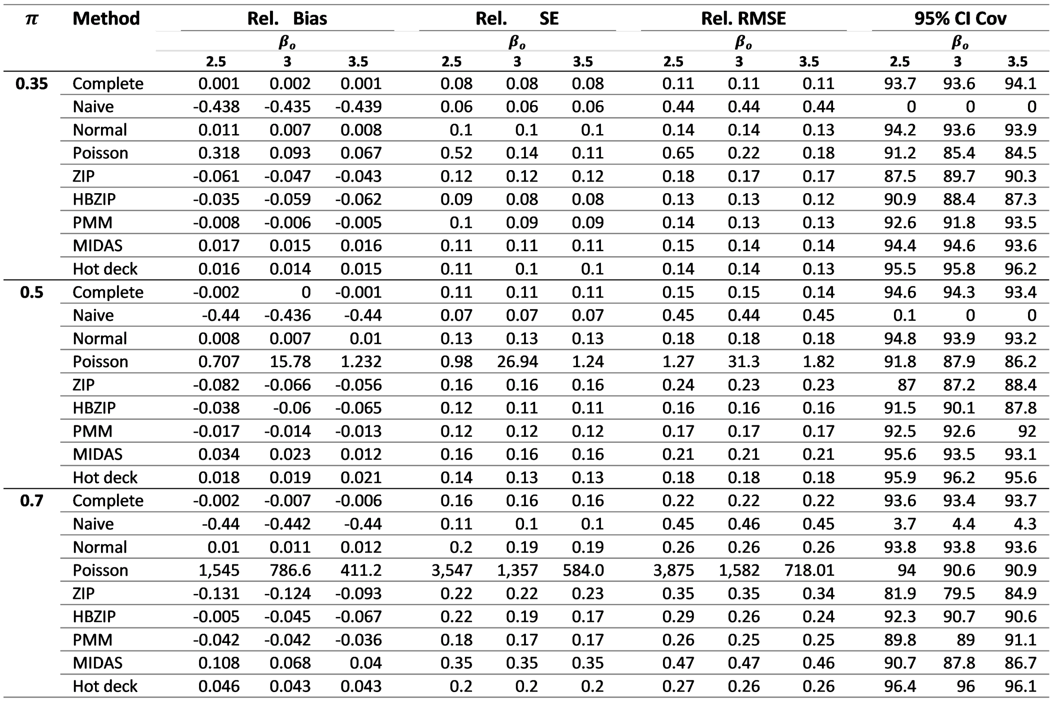
confidence interval coverage for both parameters overall. The normal model, on the other hand, performs well for τ estimation, but again, poorly for ρ estimation.

3.3.3. Bounded Visibility

When visibility mechanism follows the bounded model, PMM MIDAS outperforms the other methods when the data are not highly sparse ($\pi \leq 0.5$). As shown in Table 3.5 and Table 3.6, estimates from PMM MIDAS have a low bias ($RB[\hat{\tau}] = RB[\hat{\rho}] < 0.035$) and near nominal confidence interval coverage ($Cov[\hat{\tau}] = Cov[\hat{\rho}] = 0.931$ to 0.956). Except for hot deck and the normal imputation model, estimates from the other methods have a larger bias and lower confidence interval coverage for both estimands. Hot deck imputation, on the other hand, performs well when estimating τ but its confidence interval coverage falls below 0.93 when estimating ρ.

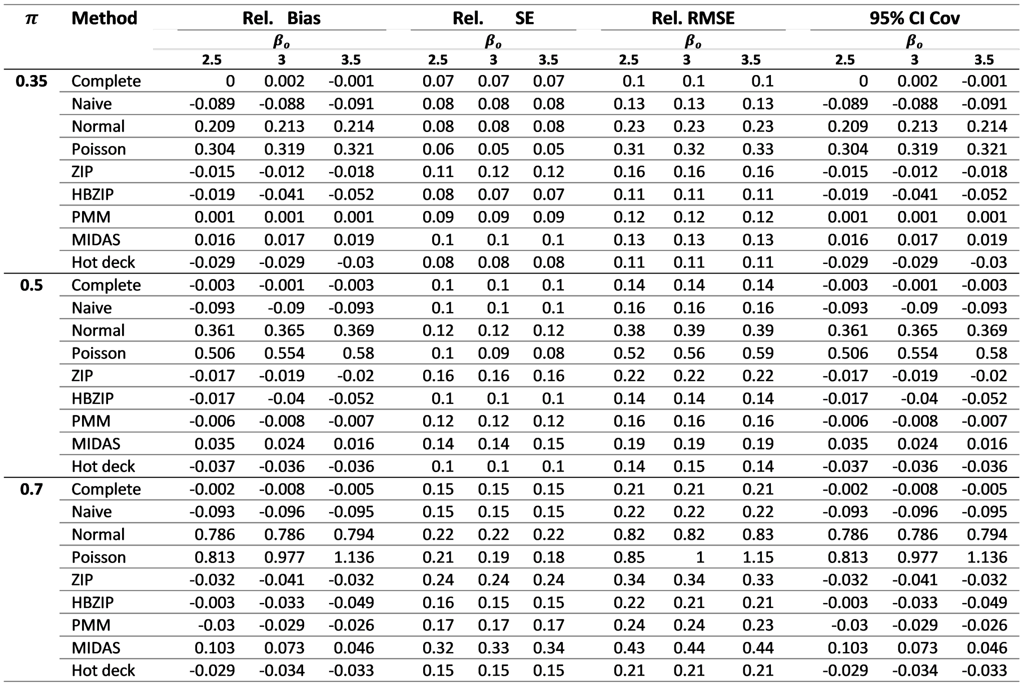
When the data are highly sparse ($\pi = 0.7$), the hot deck method yields the most favorable results for both parameters compared to the other methods. Under this condition, hot deck imputation produces the lowest bias ($RB[\hat{\tau}] < 0.047, RB[\hat{\rho}] < 0.035$ in magnitude), low relative mean squared error ($RMSE[\hat{\tau}] < 0.28, RMSE[\hat{\rho}] = 0.22$), and confidence interval coverage around its nominal value ($Cov[\hat{\tau}] \approx 0.96, Cov[\hat{\rho}] \approx 0.94$).

In addition, the normal model consistently performs well for τ estimation, with $RB[\hat{\tau}] \approx 0$ and $Cov[\hat{\tau}] = 0.932$ to 0.948 for all π values. As seen in the previous cases, the normal imputation model should not be used for estimating ρ. Furthermore, the HBZIP model can be highly biased ($RB[\hat{\tau}] > 0.065$ in magnitude) when visibility is bounded and $\beta_o = 3.5$. The HBZIP also does not outperform the other methods when $\beta_o < 2.5$. Hence, the HBZIP model is not favorable when visibility is bounded.



π	Method	Rel. Bias			Rel. SE			Rel. RMSE			95% CI Cov		
			β_o			β_o			β_o			β_o	
		2.5	3	3.5	2.5	3	3.5	2.5	3	3.5	2.5	3	3.5
0.35	Complete	0.001	0.002	0.001	0.08	0.08	0.08	0.11	0.11	0.11	93.7	93.6	94.1
	Naive	-0.438	-0.435	-0.439	0.06	0.06	0.06	0.44	0.44	0.44	0	0	0
	Normal	0.011	0.007	0.008	0.1	0.1	0.1	0.14	0.14	0.13	94.2	93.6	93.9
	Poisson	0.318	0.093	0.067	0.52	0.14	0.11	0.65	0.22	0.18	91.2	85.4	84.5
	ZIP	-0.061	-0.047	-0.043	0.12	0.12	0.12	0.18	0.17	0.17	87.5	89.7	90.3
	HBZIP	-0.035	-0.059	-0.062	0.09	0.08	0.08	0.13	0.13	0.12	90.9	88.4	87.3
	PMM	-0.008	-0.006	-0.005	0.1	0.09	0.09	0.14	0.13	0.13	92.6	91.8	93.5
	MIDAS	0.017	0.015	0.016	0.11	0.11	0.11	0.15	0.14	0.14	94.4	94.6	93.6
	Hot deck	0.016	0.014	0.015	0.11	0.1	0.1	0.14	0.14	0.13	95.5	95.8	96.2
0.5	Complete	-0.002	0	-0.001	0.11	0.11	0.11	0.15	0.15	0.14	94.6	94.3	93.4
	Naive	-0.44	-0.436	-0.44	0.07	0.07	0.07	0.45	0.44	0.45	0.1	0	0
	Normal	0.008	0.007	0.01	0.13	0.13	0.13	0.18	0.18	0.18	94.8	93.9	93.2
	Poisson	0.707	15.78	1.232	0.98	26.94	1.24	1.27	31.3	1.82	91.8	87.9	86.2
	ZIP	-0.082	-0.066	-0.056	0.16	0.16	0.16	0.24	0.23	0.23	87	87.2	88.4
	HBZIP	-0.038	-0.06	-0.065	0.12	0.11	0.11	0.16	0.16	0.16	91.5	90.1	87.8
	PMM	-0.017	-0.014	-0.013	0.12	0.12	0.12	0.17	0.17	0.17	92.5	92.6	92
	MIDAS	0.034	0.023	0.012	0.16	0.16	0.16	0.21	0.21	0.21	95.6	93.5	93.1
	Hot deck	0.018	0.019	0.021	0.14	0.13	0.13	0.18	0.18	0.18	95.9	96.2	95.6
0.7	Complete	-0.002	-0.007	-0.006	0.16	0.16	0.16	0.22	0.22	0.22	93.6	93.4	93.7
	Naive	-0.44	-0.442	-0.44	0.11	0.1	0.1	0.45	0.46	0.45	3.7	4.4	4.3
	Normal	0.01	0.011	0.012	0.2	0.19	0.19	0.26	0.26	0.26	93.8	93.8	93.6
	Poisson	1,545	786.6	411.2	3,547	1,357	584.0	3,875	1,582	718.01	94	90.6	90.9
	ZIP	-0.131	-0.124	-0.093	0.22	0.22	0.23	0.35	0.35	0.34	81.9	79.5	84.9
	HBZIP	-0.005	-0.045	-0.067	0.22	0.19	0.17	0.29	0.26	0.24	92.3	90.7	90.6
	PMM	-0.042	-0.042	-0.036	0.18	0.17	0.17	0.26	0.25	0.25	89.8	89	91.1
	MIDAS	0.108	0.068	0.04	0.35	0.35	0.35	0.47	0.47	0.46	90.7	87.8	86.7
	Hot deck	0.046	0.043	0.043	0.2	0.2	0.2	0.27	0.26	0.26	96.4	96	96.1

Table 3.5: Simulation results for total abundance estimates for all β_o and ϕ, when visibility follows the bounded model.



π	Method	Rel. Bias			Rel. SE			Rel. RMSE			95% CI Cov		
		β_o			β_o			β_o			β_o		
		2.5	3	3.5	2.5	3	3.5	2.5	3	3.5	2.5	3	3.5
0.35	Complete	0	0.002	-0.001	0.07	0.07	0.07	0.1	0.1	0.1	0	0.002	-0.001
	Naive	-0.089	-0.088	-0.091	0.08	0.08	0.08	0.13	0.13	0.13	-0.089	-0.088	-0.091
	Normal	0.209	0.213	0.214	0.08	0.08	0.08	0.23	0.23	0.23	0.209	0.213	0.214
	Poisson	0.304	0.319	0.321	0.06	0.05	0.05	0.31	0.32	0.33	0.304	0.319	0.321
	ZIP	-0.015	-0.012	-0.018	0.11	0.12	0.12	0.16	0.16	0.16	-0.015	-0.012	-0.018
	HBZIP	-0.019	-0.041	-0.052	0.08	0.07	0.07	0.11	0.11	0.11	-0.019	-0.041	-0.052
	PMM	0.001	0.001	0.001	0.09	0.09	0.09	0.12	0.12	0.12	0.001	0.001	0.001
	MIDAS	0.016	0.017	0.019	0.1	0.1	0.1	0.13	0.13	0.13	0.016	0.017	0.019
	Hot deck	-0.029	-0.029	-0.03	0.08	0.08	0.08	0.11	0.11	0.11	-0.029	-0.029	-0.03
0.5	Complete	-0.003	-0.001	-0.003	0.1	0.1	0.1	0.14	0.14	0.14	-0.003	-0.001	-0.003
	Naive	-0.093	-0.09	-0.093	0.1	0.1	0.1	0.16	0.16	0.16	-0.093	-0.09	-0.093
	Normal	0.361	0.365	0.369	0.12	0.12	0.12	0.38	0.39	0.39	0.361	0.365	0.369
	Poisson	0.506	0.554	0.58	0.1	0.09	0.08	0.52	0.56	0.59	0.506	0.554	0.58
	ZIP	-0.017	-0.019	-0.02	0.16	0.16	0.16	0.22	0.22	0.22	-0.017	-0.019	-0.02
	HBZIP	-0.017	-0.04	-0.052	0.1	0.1	0.1	0.14	0.14	0.14	-0.017	-0.04	-0.052
	PMM	-0.006	-0.008	-0.007	0.12	0.12	0.12	0.16	0.16	0.16	-0.006	-0.008	-0.007
	MIDAS	0.035	0.024	0.016	0.14	0.14	0.15	0.19	0.19	0.19	0.035	0.024	0.016
	Hot deck	-0.037	-0.036	-0.036	0.1	0.1	0.1	0.14	0.15	0.14	-0.037	-0.036	-0.036
0.7	Complete	-0.002	-0.008	-0.005	0.15	0.15	0.15	0.21	0.21	0.21	-0.002	-0.008	-0.005
	Naive	-0.093	-0.096	-0.095	0.15	0.15	0.15	0.22	0.22	0.22	-0.093	-0.096	-0.095
	Normal	0.786	0.786	0.794	0.22	0.22	0.22	0.82	0.82	0.83	0.786	0.786	0.794
	Poisson	0.813	0.977	1.136	0.21	0.19	0.18	0.85	1	1.15	0.813	0.977	1.136
	ZIP	-0.032	-0.041	-0.032	0.24	0.24	0.24	0.34	0.34	0.33	-0.032	-0.041	-0.032
	HBZIP	-0.003	-0.033	-0.049	0.16	0.15	0.15	0.22	0.21	0.21	-0.003	-0.033	-0.049
	PMM	-0.03	-0.029	-0.026	0.17	0.17	0.17	0.24	0.24	0.23	-0.03	-0.029	-0.026
	MIDAS	0.103	0.073	0.046	0.32	0.33	0.34	0.43	0.44	0.44	0.103	0.073	0.046
	Hot deck	-0.029	-0.034	-0.033	0.15	0.15	0.15	0.21	0.21	0.21	-0.029	-0.034	-0.033

Table 3.6: Simulation results for habitat occupancy rate estimates for all β_o and ϕ, when visibility follows the bounded model.

3.4. Conclusion

We examined the performance of various multiple imputation for measurement error correction methods via a simulation study. The results suggest that the methods' performance for estimating total abundance and occupancy rate varies mostly by visibility mechanism and sparsity of the data. The HBZIP model performs well when the actual visibility mechanism is consistent with the assumed visibility model; therefore, it is the best imputation model when the actual visibility mechanism is linear. However, when visibility is not linear and the data are not highly sparse, PMM MIDAS outperforms the other methods with low bias and near nominal confidence interval coverage. Furthermore, when visibility is not linear and the data are highly sparse, the best imputation model is HBZIP when visibility is discrete, and hot deck when visibility is bounded. Finally, the normal imputation model has been shown to be quite effective overall for estimating total abundance, but along with the Poisson model, should be avoided for occupancy rate estimation.

Finally, as discussed in the previous paragraph, estimates from the HBZIP method are shown to be sensitive to visibility model misspecification. Hence, a more flexible method is necessary to improve the robustness of the method against visibility model misspecification. In the next chapter, we present a Bayesian averaging method which can be used to accommodate visibility model uncertainty and mitigate the impact of visibility model misspecification for estimating total abundance and habitat occupancy rate.

Chapter 4
Bayesian Model Averaging Adaptation

The performance of the HBZIP model, based on simulation results in Chapter 3, is sensitive to visibility model specification. Within the HBZIP model, the relationship between turbidity and visibility rate is assumed to be linear, though they may be monotonically related. The analyst may examine the actual visibility mechanism manually using a graphical approach and determine the appropriate visibility model. However, the true visibility model is not always apparent, and the chosen visibility model if determined manually will be subjective and may differ between analysts.

The analyst may also perform a goodness-of-fit test such as Pearson Chi-square and Hosmer–Lemeshow (Hosmer et al., 1997) to assess the appropriateness of the visibility model. However, such preliminary tests can be too stringent; violation of the model assumption does not necessarily invalidate the main objective of the analysis, which in our case is for inference about population parameters. Furthermore, the actual visibility mechanism also varies by survey operation and species. For example, the relationship between visibility rate and turbidity level in video transect surveys of red snapper may be different from the relationship between visibility rate and visual obstruction coverage, a measure of visual clarity in aerial survey of moose. Hence, a more flexible approach which can accommodate the uncertainty around the visibility model is necessary.

To account for the uncertainty in the visibility model and reduce the impact of model misspecification we incorporate Bayesian model averaging (BMA) into the HBZIP model. BMA allows several visibility models to be included in the imputation model and assigns weights to the different models based an assessment of model fit. Within the BMA framework, this is done by averaging the posterior distribution of the parameter of interest under each of the considered models and weighting them by the posterior probability of the model being the correct one. For example, suppose we are interested in the parameter Δ and consider $\mathcal{M}_1, \mathcal{M}_2, .., \mathcal{M}_k$ as possible models for Δ. Following Hoeting, Madigan, Raftery, & Volinsky‘s (1999) notation, the posterior distribution of Δ is

$$P(\Delta|D) = \sum_{k=1}^{K} P(\Delta|\mathcal{M}_k, D)P(\mathcal{M}_k|D). \tag{4.1}$$

Therefore, the posterior probability of $\mathcal{M}_k$ being the right model is

$$P(\mathcal{M}_k|D) = \frac{P(D|\mathcal{M}_k)P(\mathcal{M}_k)}{\sum_{j=1}^{J} P(D|\mathcal{M}_j)P(\mathcal{M}_j)} \tag{4.2}$$

The analyst may then assign the appropriate priors for $P(\mathcal{M}_k)$ based on their knowledge about the true model or assign a vague prior so that $P(\mathcal{M}_k|D)$ is dominated by the likelihood of the data.

4.1. BMA Application on the HBZIP Model

Recall from (2.11), we defined $\eta_i = logit(P_i)$ to be the logit model for visibility rate within the HBZIP model. So far, we have only considered one visibility model and made the assumption that logit(P_i) and turbidity level (T_i) level are linearly related such that logit(P_i) $= \theta_o + \theta_1 T_i$. In other words, this is equivalent to claiming that the pre-selected linear model is the correct model with probability one. In this section we incorporate BMA and consider two visibility models: linear and non-linear. Let $\eta_{ki} = logit(P_{ki})$ be the k^{th} visibility model; we define the linear model ($\mathcal{M}_1$):

$$\mathcal{M}_1 : \eta_{1i} = \theta_{1o} + \theta_{11} log\left(\frac{T_i}{1-T_i}\right) \tag{4.3}$$

where θ_{1o} and θ_{11} denote the intercept and slope in the linear model. Under $\mathcal{M}_1$ we assume that visibility rate and turbidity level are linearly related in the logit scale. Similarly, this means that visibility rate in its raw scale is a monotonic function of the turbidity level. The non-linear model ($\mathcal{M}_2$) is defined by:

$$\mathcal{M}_2 : \eta_{1i} = \theta_{2o} + I_{0<t_i \leq c_1}\theta_{21} + I_{c_1<t_i \leq c_2}\theta_{22} + I_{c_2<t_i \leq c_3}\theta_{23} + I_{c_3<t_i \leq 1}\theta_{24} \tag{4.4}$$

Under $\mathcal{M}_2$, visibility rate is a piecewise function of turbidity level. Depending on their value, turbidity levels are divided into four intervals separated by three cutoff points ($c_k, k = 1, 2, 3$). The effect of turbidity level within each interval is constant and denoted by $\theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}$ respectively. In the model, priors for the cutoff points are assumed to follow a uniform distribution subject to the constraint $0 \leq c_1 \leq c_2 \leq c_3 \leq 1$. Therefore, the cutoff points are not pre-determined by the analyst, rather maximized based on the likelihood of the data.

Figure 4.1 illustrates the various relationships between visibility rate (P) and turbidity level (T), based on different values of θ_{1o} and θ_{11}. As seen in the figure, location of the

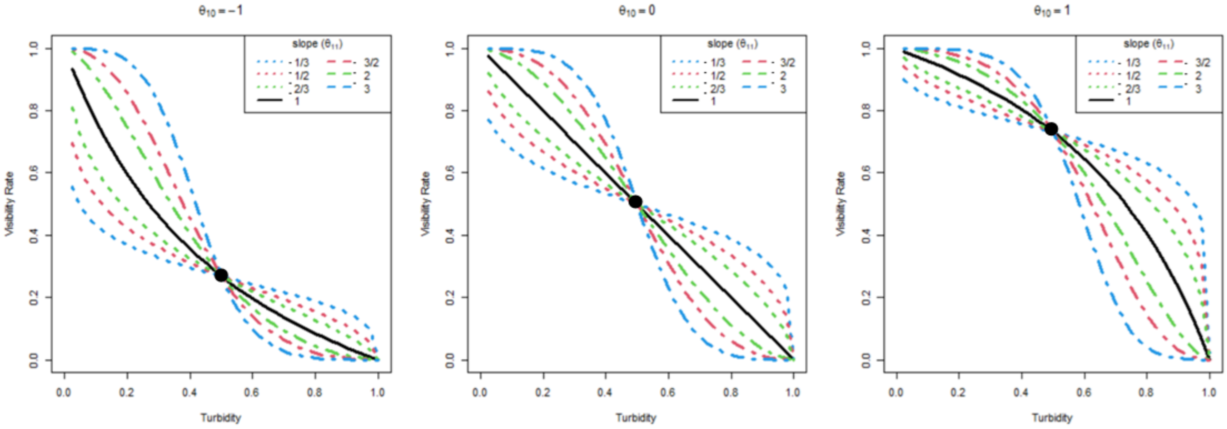


Figure 4.1: Relationship between turbidity and visibility rate under the linear model assumption.

inflection point, the point at which the curve changes direction (i.e. from concave upward to concave downward), indicated by the black dot, is controlled by the intercept. The inflection point is located at $P < 0.5$ when $\theta_{1o} = -1$, $P = 0.5$ when $\theta_{1o} = 0$, and $P > 0.5$ when $\theta_{1o} = 1$. The slope, on the other hand, determines the steepness and concavity of the curve. That is, the larger the slope's magnitude the steeper the curve becomes. In addition, when $\theta_{11} = -1$, the concavity of the graph does not change; hence, the function does not have an inflection point. However, the curve changes from concave downward to concave upward when $\theta_{11} < -1$, and similarly, from concave upward to concave downward when $\theta_{11} > -1$.

Define $P(\mathcal{M}_1) = \delta$, the probability that $\mathcal{M}_1$ is the true model or similarly, a weighting coefficient of $\mathcal{M}_1$. Assuming that the true model is either linear or non-linear, $P(\mathcal{M}_2) = 1 - \delta$ is therefore the probability of $\mathcal{M}_2$ being the true model. Thus, the averaged model is:

$$\eta_i = (\delta)\mathcal{M}_1 + (1-\delta)\mathcal{M}_2 + e_i, \quad e_i \sim N(0, \sigma_\eta^2), \quad \delta \in (0,1).$$

Finally, we assumed a vague beta prior for the weighting coefficient δ. Similar to the HBZIP model in Chapter 2, priors for each of the regression coefficients in $\mathcal{M}_1$ and $\mathcal{M}_2$ are also assumed to be mutually independent and follow vague normal priors. Finally, a vague uniform prior is assumed for σ_η.

4.2. Visibility Model Misspecification

Specifying the appropriate visibility model will not only improve the accuracy of the predicted visibility rates but also improve the accuracy of the imputed values. Recall from (2.9), distribution of the unobserved count (U_i) follows U_i is $ZIP(\pi_i', \lambda_i')$ when $y = 0$, and $Poisson(\lambda_i' = \lambda_i[1 - P_i])$ when $y > 0$, where $\lambda_i' = \lambda_i[1 - P_i]$. Therefore, the count mean (λ_i') decreases as visibility rate (P_i) increases, and vice versa. Consequently, λ_i', the imputed values, and the estimated total abundance will be overestimated if P_i is underestimated and similarly, underestimated if P_i is overestimated.

Visibility model misspecifications are illustrated by the graphs in Figure 4.2. Within each plot, the true visibility rates are indicated by circles and the predicted ones are shown in grey dots. The rows represent the three visibility mechanisms (linear, discrete, and bounded) and the columns indicate the visibility models used within the HBZIP model: linear, non-linear, and BMA. Overall, the predicted visibility rates resemble the pattern of the actual visibility rate most closely when the model is correctly specified: linear model for linear visibility mechanism, and non-linear model for bounded and discrete visibility mechanism. As illustrated by the figures in the third column, though they are not as accurate as the

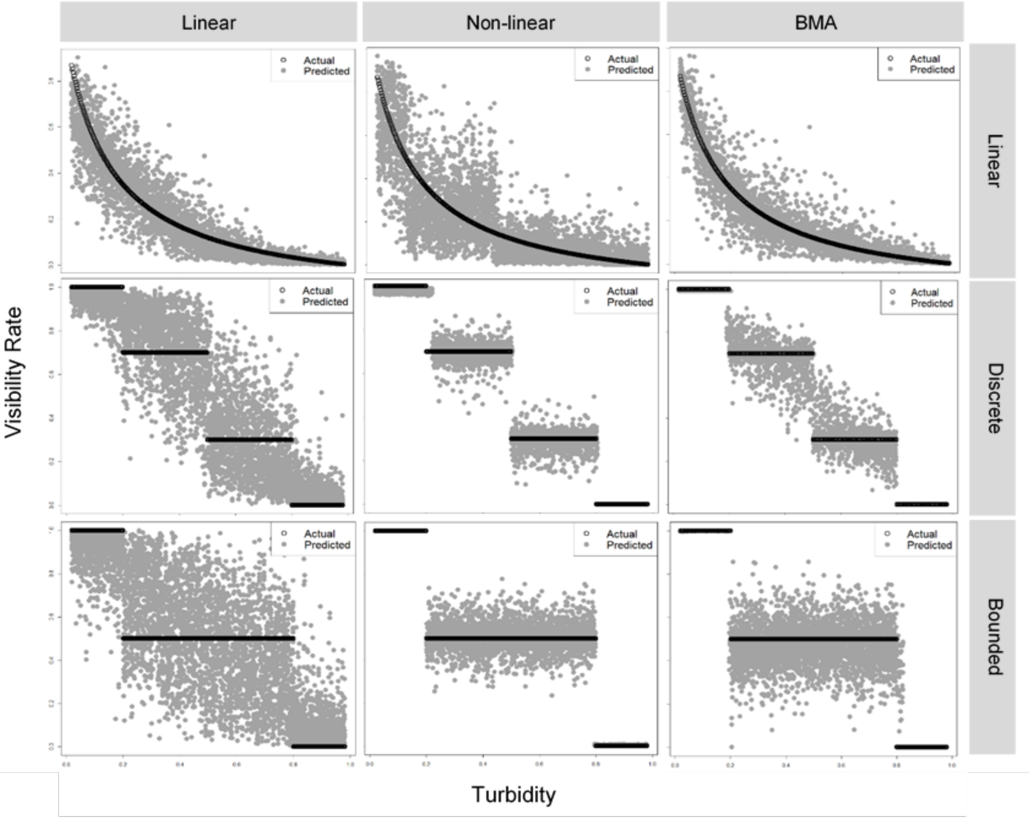


Figure 4.2: Plots of visibility rates based on the its actual visibility mechanism, as shown by the different rows (linear, discrete, and bounded), predicted using HBZIP with linear, non-linear, and BMA visibility models (columns).

correct model, predicted visibility rates from the BMA approach appear to also reflect the actual visibility mechanism.

We further assessed whether BMA model can correctly identify the true visibility model by evaluating the estimated posterior probability of the true model. Ideally, given the data, the correct model (linear model for linear visibility mechanism and non-linear model for bounded or discrete visibility) should have a larger posterior probability than the incorrect model; hence, it should be weighted more in the averaged model. In Figure 4.2, the estimated posterior probability of the correct model based on BMA is 0.999 for linear visibility mechanism, 0.774 for discrete, and 0.954 for bounded visibility mechanism. This suggests that BMA can identify the correct model though it appears to detect the linear model better than the non-linear models. The non-linear model is more complex since it involves more parameters to be estimated than the linear model. Hence, a larger sample size may be necessary for BMA to correctly choose the non-linear model with equally high posterior probability, compared to the linear model.

Figure 4.3 illustrates the predicted visibility rates for varying sample size and their estimated posterior probability of the correct model, denoted by $\hat{P}(\mathcal{M}_c|D)$. The figure illustrates that as sample size increases, $\hat{P}(\mathcal{M}_c|D)$ also increases, and the true visibility mechanism is better reflected by the predicted visibility rates. In addition, given the same sample size, $\hat{P}(\mathcal{M}_c|D)$ is also higher when visibility mechanism follows the linear model than when it follows one of the non-linear models. Therefore, while the correct model is weighted more overall in the averaged model, the estimated posterior probability of the correct model appears to be impacted by the sample size and the visibility model.

While assessing the posterior probability is important to check whether BMA properly selects the right model, the purpose of BMA implementation here is not to predict the true visibility model, but rather to improve the imputation model's performance for estimating total abundance and habitat occupancy rate. Even when the sample size is small and BMA

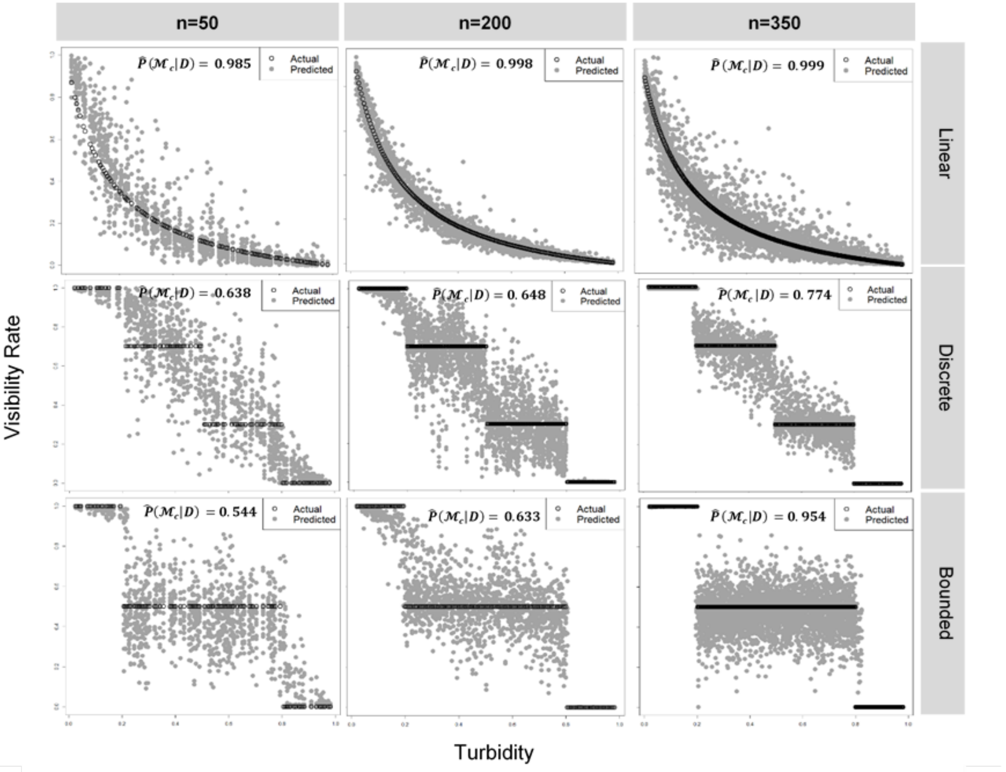


Figure 4.3: Plots of visibility rates by sample size (column) and visibility mechanism (row).

is not choosing the right model with high posterior probability, BMA still yields visibility rates that better reflect the actual visibility mechanism than the misspecified model. Hence, imputation performance of the HBZIP model with BMA implementation can still perform relatively well even if it is not choosing the right model based on the posterior probability of the model.

The model's imputation performance for estimating total abundance and habitat occupancy rate is investigated in the next section. We evaluated the performance of the HBZIP method with BMA implementation, which we refer to here as HBZIP-BMA, in a simulation study similar to the one presented in Chapter 3.

4.3. HBZIP-BMA's Model Performance

Simulation results in Chapter 3 suggest that the HBZIP model yields favorable outcomes when the visibility mechanism follows the linear model. However, its performance worsened when the visibility mechanism follows the non-linear model, as its bias increases and confidence interval coverage fall significantly below 95%. We included HBZIP-BMA in the simulation presented in Chapter 3 and evaluated its performance; the results are reported in Table 4.1 and Table 4.2. Previous simulation results from the HBZIP model without BMA implementation are also presented in the table for comparison. The complete simulation results for HBZIP-BMA are reported in Appendix D.

π	Method	Linear			Discrete			Bounded		
		Rel. Bias	Rel. SE	95% CI Cov	Rel. Bias	Rel. SE	95% CI Cov	Rel. Bias	Rel. SE	95% CI Cov
0.35	HBZIP	-0.029	0.08	93.5	-0.048	0.08	90.2	-0.052	0.09	88.9
	HBZIP-BMA	0.014	0.08	94.5	0.012	0.08	93.8	0.012	0.08	93.5
0.5	HBZIP	-0.025	0.11	94.1	-0.047	0.11	91.4	-0.054	0.11	89.8
	HBZIP-BMA	0.018	0.11	94.4	0.013	0.11	93.9	0.01	0.11	94.2
0.7	HBZIP	-0.009	0.18	94.3	-0.034	0.18	93.2	-0.039	0.19	91.2
	HBZIP-BMA	0.025	0.17	94.7	0.014	0.17	94.1	0.008	0.17	93.8

Table 4.1: Simulation results for total abundance estimates by visibility mechanism and zero-inflation rate π.

As shown in Table 4.1, the largest improvement from BMA implementation is observed when the visibility model follows the non-linear model. The relative bias decreases in magnitude, on average, from -0.043 to 0.013 when the visibility follows the discrete model, from -0.048 to 0.01 when visibility follows the bounded model. However, only a slight improvement is observed, from -0.021 to 0.019 on average, when visibility follows the linear model. Similarly, the largest improvement in confidence interval coverage is observed when visibility follows the discrete and bounded model. On average, confidence interval coverage increased from 91.6% to 93.9% for discrete visibility mechanism and from 89.9% to 93.8% for bounded visibility, and from 94.0% to 94.5% for linear visibility mechanism.

Similar results are observed for habitat occupancy rate estimates. As shown in Table 4.2, little improvement is seen when visibility follows the linear model. However, when the visibility mechanism follows the discrete model, the magnitude of the relative bias decreases from -0.035 to 0.015 on average; the confidence interval coverage also improved from 92.5% to 93.7%. When the visibility mechanism follows the bounded model, the overall magnitude of the bias decreased from -0.034 to 0.008 and the confidence interval coverages also improves from 92.3% to 93.5%. Finally, the two models produce similar results when visibility follows the linear model; confidence interval coverage is 94.2% on average for HBZIP compared to 94.4% for HBZIP-BMA.

π	Method	Linear			Discrete			Bounded		
		Rel. Bias	Rel. SE	95% CI Cov	Rel. Bias	Rel. SE	95% CI Cov	Rel. Bias	Rel. SE	95% CI Cov
0.35	HBZIP	-0.024	0.1	94.7	-0.038	0.1	92.4	-0.036	0.1	92.7
	HBZIP-BMA	0.02	0.1	94.5	0.013	0.1	93.8	0.008	0.1	93.8
0.5	HBZIP	-0.028	0.07	93.2	-0.041	0.08	90.7	-0.037	0.08	90.7
	HBZIP-BMA	0.016	0.07	93.8	0.012	0.07	92.7	0.01	0.07	93
0.7	HBZIP	-0.017	0.15	94.7	-0.025	0.15	94.5	-0.028	0.15	93.7
	HBZIP-BMA	0.024	0.16	94.8	0.019	0.16	94.7	0.006	0.16	93.7

Table 4.2: Simulation results for habitat occupancy rate by visibility mechanism and zero-inflation rate π

4.4. Conclusion

In this chapter, we applied BMA to improve the robustness of the HBZIP imputation model against visibility model misspecification. We incorporated both linear and non-linear visibility models within the HBZIP imputation model to account for visibility model uncertainty and to reduce the impact of model misspecification. We demonstrated via a simulation study that BMA application on the HBZIP imputation model improves the method's performance for estimating total abundance and habitat occupancy rate estimation, especially when the visibility mechanism follows the non-linear models. While the two visibility models presented in this chapter accommodate various shapes for the visibility mechanism, they do not cover all of them. Therefore, more research can be done to explore other potential visibility models or simpler yet flexible models that can cover a broad range of visibility mechanism, especially when visibility mechanism follows the non-linear model.

Thus far, we have shown through a simulation study that Bayesian model averaging is effective to improve the robustness of HBZIP imputation model. In the next chapter, we illustrate the application of HBZIP model on real data collected from moose population surveys and compare its performance to an existing weighting adjustment approach.

Chapter 5

Real Data Example

In this chapter, we provide an example of HBZIP-BMA implementation using data from an aerial survey of moose in northeastern Minnesota. The data can be accessed through the *SightabilityModel* package in R (Fieberg, 2012) and has been used to demonstrate other visibility bias correction methods for estimating moose population size (Giudice et al., 2012; ArchMiller et al., 2018).

We implemented HBZIP-BMA to estimate moose population abundance from 2005 to 2007 and compared the results with total abundance estimates obtained from normal imputation, naive approach, and the sighability adjustment method (Steinhorst and Samuel, 1989). Occupancy rate is not estimated here since the moose data are not sparse. In addition, we performed an additional simulation study and generated hypothetical data which reflected the sampling design and characteristics of the moose data. In the simulation, we imposed sparsity on the data and evaluated each of the method's performance for total abundance and habitat occupancy rate estimation.

The moose data included data from two surveys: sightability trial and operational. In both surveys, the moose were visually surveyed from a helicopter and the amount of screening cover or percent of visual obstruction coverage (voc) was recorded. The actual presence of the moose was only known in the sightability trial data. During the sightability trial, the researchers took advantage of moose that were radio-collared for a different, moose survival study (Fieberg et al., 2013). The sightability data contain 124 sightability trials used to estimate the visibility model parameters. The operational surveys comprised of 805 sightings of moose groups from 124 plots, the primary sampling units, sampled from three strata between 2004 to 2007. Therefore, multiple plots were selected from each of the strata and multiple sightings were possible within each plot. The strata were determined prior to sampling based on the moose population density.

5.1. Sightability Adjustment Method Overview

Recall from Section 1.4.2 that sightability adjustment is a correction method in which visibility bias is viewed as a non-response problem; hence, it can be corrected using weighting adjustment (Steinhorst and Samuel, 1989). In its application, Steinhorst and Samuel (1989) treat visibility rate analogously to response probability. They developed a modified Horvitz-Thompson estimator of population size which incorporates both the survey design selection probability and sighting probability, as if it is an additional stage of sampling. To estimate the unknown sighting probability, they proposed using a logistic regression model to predict visibility rates using other covariates measured during the survey. To account for the extra variation due to visibility bias, Steinhorst and Samuel (1989) derived a complex expression for the estimator's variance which includes three components of variability accounting for sampling error, sightability error, and errors for estimating the visibility model parameters.

The sightability adjustment method has been implemented by Fieberg (2012) on the moose data. Abundance estimates and their standard errors can be reproduced using the *SightabilityModel* package in R (Fieberg, 2012). The package also allows the user to specify the logistic regression model used to predict the visibility rate. Fieberg (2012) demonstrated fitting a logistic regression model in which the relationship between voc and visibility is not linear; this was done by fitting a logistic regression model with natural cubic splines (Chambers and Hastie, 1992; Perperoglou et al., 2019) of the voc as the predictor. In our analysis, we use both the linear and spline logistic regression model to estimate abundance using the sightability adjustment method.

HBZIP-BMA Comparison to the Sightability Adjustment Method

The sightability adjustment method differs from the HBZIP-BMA method in three ways. First, the sightability adjustment method is analysis or parameter specific. That is, the variance expression for the sightability estimator is specific to the parameter of interest. Therefore, a new variance expression must be derived if the analyst wishes to use the sightability adjustment method to estimate population parameters other than total abundance. The HBZIP-BMA on the other hand is an imputation-based approach. Therefore, population parameters can be estimated from the imputed data using the same estimation method used if no data had been missing or mismeasured; point and variance estimates from the multiply imputed data can be pooled using the standard imputation combining rule.

Second, the sightability adjustment method is a design-based correction approach, in contrast with HBZIP-BMA, which is model-based. That is, variance in the sightability estimator is viewed as the resulting variability from the sampling design and non-response error. In addition, correction in the sightability approach is performed via weighting adjustment, by dividing the observed count with the visibility rate. Therefore, this method requires visibility rate to be greater than zero. In a model-based approach such as the HBZIP, variability in the estimate comes not only from the sampling design but also from the stochastic variation of the variables in the data and estimated parameters in the imputation model. Therefore, correction is performed stochastically based on the imputation model, rather than deterministically by weighting adjustment. Hence, the sightability correction method is non-parametric, in contrast with HBZIP which makes assumptions about the underlying statistical distribution of the data.

Last, the sightability approach separates sources of variance into three components (sampling error, sightability, and visibility model) while the HBZIP method separates variability into two components (within and between imputation variance). The within imputation variance is obtained by averaging the sampling variance from each of the imputed data set; hence, it is equivalent to the sampling error in the sightability adjustment method. The between imputation variance is analogous to the two latter variance components: sightability and visibility model. Together they account for the variance arising from the mismeasured data and the extra variability from estimating the model parameters.

5.2. Adaptation and Total Abundance Estimation

To implement the HBZIP-BMA method, adaptations were necessary for two reasons. First, similar to the regular imputation approach, the HBZIP-BMA requires that the missing/mismeasured data (operational survey) and complete data (sightability trial) are subsets of the same data. Therefore, the same set of covariates would be available in the mismeasured and complete data. The sightability trial, however, was lacking stratum-specific information necessary in the imputation model. Thus, to implement HBZIP-BMA, we randomly assigned each sightability trial to one of the three strata. Since the average size of moose groups varies among strata, we calibrated the group size so that the average size of moose groups is similar in both the sightability and operational data.

Second, visibility rate here is defined as the probability of sighting an animal group rather than sighting an individual animal. Therefore, the count data model is also defined for moose groups rather than for individual moose. Hence, N within the imputation model denotes the true count of moose groups; therefore, moose abundance within each plot (primary sampling unit) is estimated by adding the imputed count of moose groups ($\hat{N}$) multiplied by their respective group size (m).

Let S_h= number of sampling units (plots) in the h^{th} stratum, $N_{h,i}$= number of moose groups in the i^{th} unit of the h^{th} stratum, and $m_{h,i,j}$ = the number of moose in the j^{th} group, in the i^{th} unit of the h^{th} stratum; total abundance for the population is therefore defined by $\tau = \sum_{h=1}^{3} \sum_{i=1}^{S_h} \sum_{j=1}^{N_{h,i}} m_{h,i,j}$. Given s_h sampled units in the h^{th} stratum, and $p_{h,i}$ selection probability of the i^{th} unit in the h^{th} stratum, the estimator of τ in the absence of visibility bias is $\hat{\tau} = \sum_{h=1}^{3} \sum_{i=1}^{s_h} \frac{1}{p_{h,i}} \sum_{j=1}^{N_{h,1}} m_{h,i,j}$. Due to obscured visibility, not all groups are observed within each sampling unit. Therefore, the total abundance estimate based on the sightability adjustment method is:

$$\hat{\tau} = \sum_{h=1}^{3} \sum_{i=1}^{s_h} \frac{1}{p_{h,i}} \sum_{j=1}^{n_{h,i}} \frac{m_{h,i,j}}{\pi_{h,i,j}} \tag{5.1}$$

where $\pi_{h,i,j}$ = probability of observing the j^{th} animal group in the i^{th} unit of the h^{th} stratum and $n_{h,i}$ = number of moose groups sighted in the i^{th} unit of the h^{th} stratum. Alternatively, let $\hat{N}_{h,i,j}$ be the imputed count of moose group with size $m_{h,i,j}$. The estimated total abundance based on the HBZIP approach is

$$\hat{\tau} = \sum_{h=1}^{3} \sum_{i=1}^{s_h} \frac{1}{p_{h,1}} \sum_{j=1}^{n_{h,i}} \hat{N}_{h,i,j} m_{h,i,j} \tag{5.2}$$

Variance of the estimate in (5.1) and (5.2) are computed following the variance formula for the stratified sampling total estimator (Lohr, 2009). Total abundance estimates and their standard errors from the multiply imputed data are then pooled using Rubin's combining rule shown in (2.1). The missing counts are imputed with posterior samples of N drawn from the HBZIP-BMA model. The HBZIP-BMA model was fit using JAGS (Plummer, 2003), a program for analysis of Bayesian hierarchical models which utilizes Markov Chain Monte Carlo simulation. The program was run through *R2jags* package (Su and Yajima, 2020). The JAGS code used in the analysis can be found in Appendix E.

5.3. Results

Estimates of total abundance and their 95% confidence intervals are illustrated in Figure 5.1. Except for the naïve approach, the lowest abundance in 2005 was estimated by the linear sightability approach (8,158), followed by HBZIP BMA (8,216), normal imputation (8,349), and spline sightability (8,459). Overall, except for the naïve approach, estimates from the other methods are within each other's margin of error as illustrated in Figure

5.1. Total abundance estimates, their standard errors, and relative standard errors are also reported in Appendix F. The resulting standard errors and therefore margin of errors from both sightability methods are wider compared to those from normal and HBZIP approach.

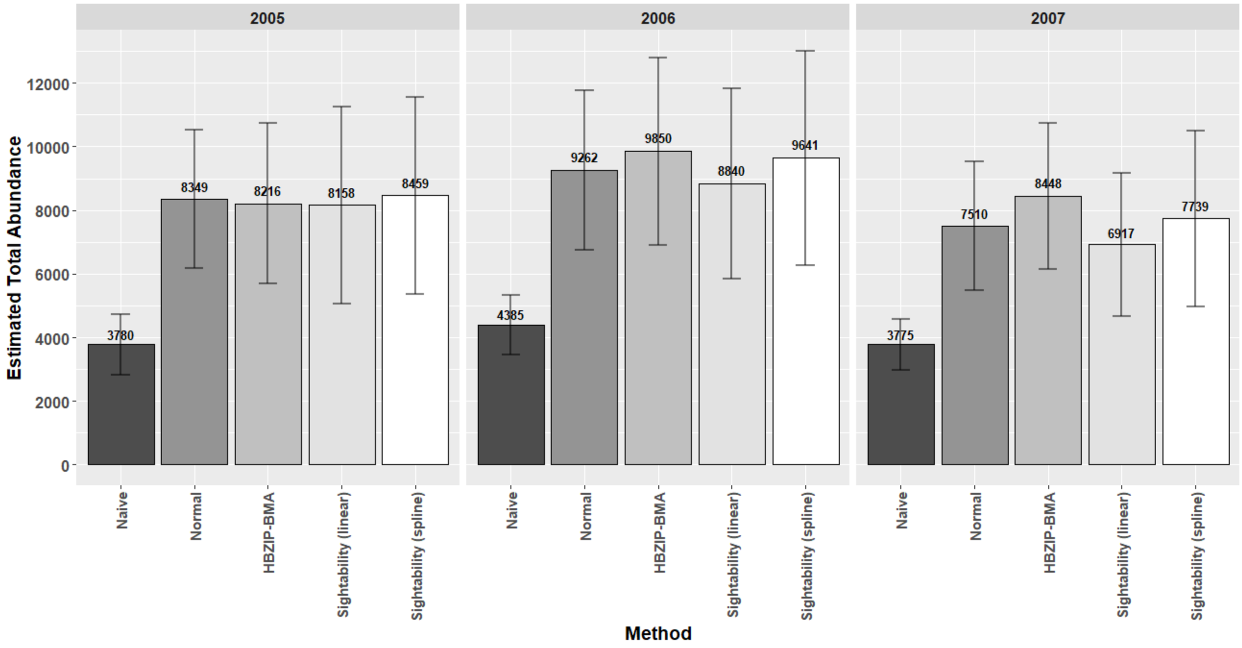


Figure 5.1: Estimates of total moose abundance and their 95% confidence interval.

Though the estimates are not significantly different from each other except for the naïve approach, the linear sightability method consistently produces the lowest estimates. As shown in the figures, discrepancies between the different methods are especially apparent in 2006 and 2007. However, it remains unclear which method yields the best estimate since the true total abundance is unknown.

To explain these discrepancies, we examine visibility rates predicted by each of the methods except for the normal model. Note that visibility rate is not used or modeled within the normal imputation method; hence, it is not predicted. The predicted rates are then compared to the actual visibility rates estimated from the data. Figure 5.2 illustrates the actual and predicted visibility rates from the different correction methods.

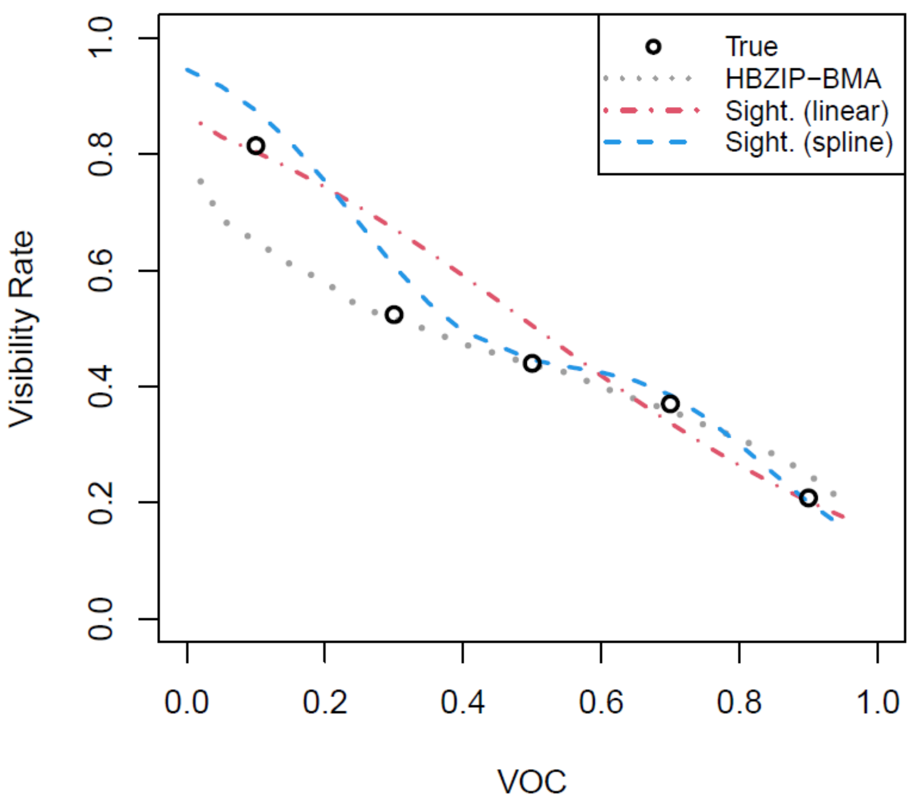


Figure 5.2: True and predicted visibility rates from the HBZIP, linear sightability, and spline sightability adjustment method.

The actual visibility rate is estimated by dividing the voc into five equal-width intervals and calculating the proportion of observed moose groups within each interval. The actual visibility rates in the moose survey, indicated by the black circles, do not appear to follow the linear model. Visibility rates appear to be overestimated by the linear sightability approach; therefore, the observed counts are not sufficiently inflated which would result in an underestimation of the total abundance. On the other hand, the spline sightability approach closely predicts the true visibility rate, though it is overestimated when voc is below 0.5. Similarly, the HBZIP-BMA also closely predicts the true visibility rate but slightly underestimates it when voc is below 0.3. Therefore, based on these predicted visibility rates, the true abundance is likely to be between the estimated total abundance from the HBZIP-BMA method and the sightability approach.

5.4. Simulation Study Based on the Moose Data

Moose abundance estimates in the previous section were estimated based on real and not simulated data. Therefore, the true value of the total abundance parameter was unknown and evaluating the methods' performance for inference purposes was not possible. In this section, we generate hypothetical data which reflect the characteristics and sampling design of the moose survey. We impose sparsity on the count data and estimate total abundance as well as habitat occupancy rate.

For total abundance estimation, we evaluated the following correction methods: normal imputation, HBZIP-BMA, linear sightability, and spline sightability approaches. The sightability estimator of habitat occupancy rate has not yet been developed and is only available for population size and population ratios (e.g. calf to cow ratio) (Samuel et al., 1992). Therefore, only normal and HBZIP-BMA imputation method were evaluated for habitat occupancy rate estimation.

5.4.1. Simulation Design

The moose data consist of 121 sampled plots and 805 sightings of moose groups across the three strata between 2004-2007. These data were then treated as the target population from which a total of 35 plots were randomly sampled. The plots were sampled with an equal sampling rate within each stratum, which resulted in 11, 18, and 6 sampled plots from the first, second, and third stratum respectively. Within each plot, the group size (N) was generated from a zero-inflated Poisson with a constant zero-inflation rate and differing Poisson means (λ) in each stratum. The observed counts (Y) were generated following Binomial(N, P) where P is the probability of observing an individual moose. Of the sampled units, the true counts were observed in n_v of the units within each stratum; hence, only the observed counts (Y) were available in the remaining samples.

The simulation experiment was a three-factor factorial design with four levels of validation sample size (n_v =2, 3, 4, 5), three levels of zero-inflation rate (π =0, 0.3, 0.5) and two visibility mechanisms (linear and non-linear). Validation sample size here was defined as the number of sampled plots for which the true counts were observed. The average number of sightings per sampling plot was 6.6; therefore, approximately 13, 20, 27, and 33 sightings from 2, 3, 4, and 5 sampling plots were validated respectively. Multiple zero-inflation rates were explored to evaluate the effect of data sparsity on the estimates. The two visibility mechanisms, linear and non-linear, were also simulated to assess the robustness of the different correction methods to visibility model mispecification. In summary, the simulation study was conducted with 24 parameter combinations total, each replicated 500 times.

In the simulation, visibility rate was defined as the probability of observing an individual moose rather than a group of moose. Hence, total abundance was estimated by adjusting for the number of moose observed rather than the number of moose groups observed. Occupancy rate was estimated by the proportion of non-zero moose counts in the secondary sampling units. Estimates of total abundance and occupancy rate were evaluated based on their relative bias, relative standard error, and 95% confidence interval coverage.

5.4.2. Results

Simulation results for total abundance estimation are illustrated in Figure 5.3. Estimates from the complete data are also included in the figure to illustrate the estimator's performance in the absence of visibility bias. As shown in Figure 5.3, both normal and HBZIP methods yielded relatively low bias estimates regardless of the visibility mechanism, zero-inflation rate (π), and validation sample size (n_v). The sightability adjustment methods yielded relatively low bias ($|RB[\hat{\tau}]| < 0.03$) when visibility mechanism follows the linear model. However, the bias intensifies when the visibility mechanism follows the non-linear model, especially when the validation sample size is small ($n_v = 2$) as shown by the dashed lines.

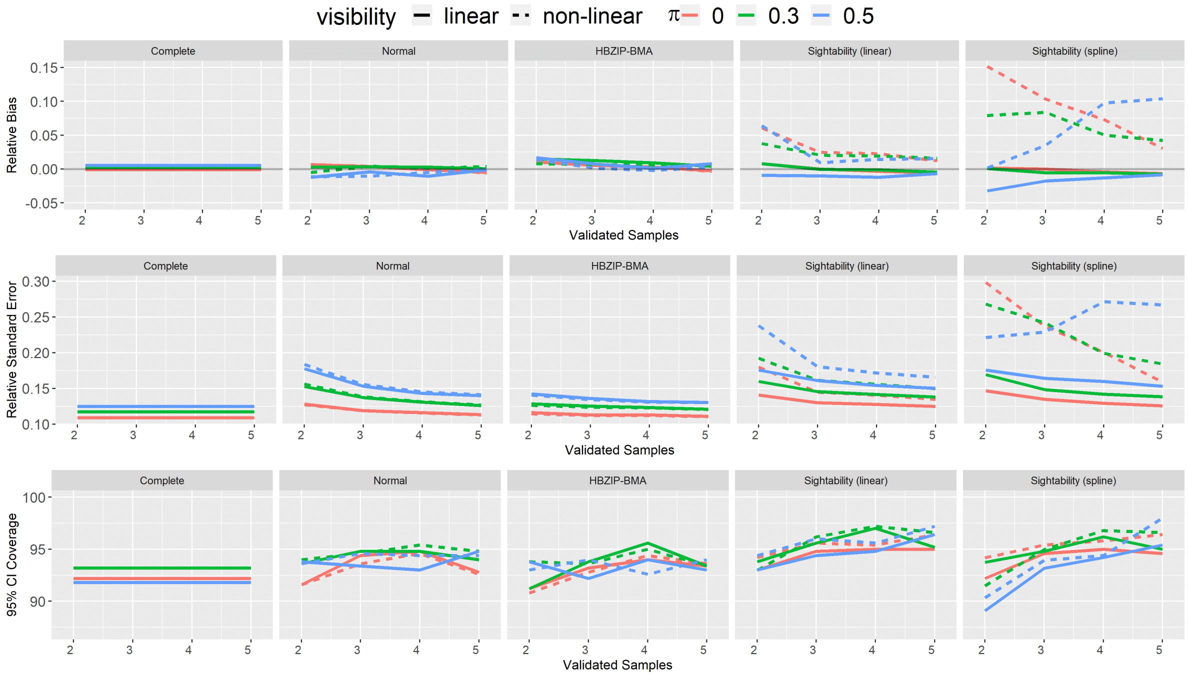


Figure 5.3: Simulation results for total abundance estimates by correction method and measuring criteria: relative bias (first row), relative standard error (second row), and 95% confidence interval coverage (third row). Within each plot, visibility mechanism is indicated by the line type (dashed or solid), while zero-inflation is indicated by the different color.

The same pattern is observed for the relative standard error. Both sightability approaches yield standard error that is sensitive to the visibility mechanism. For all correction methods, the standard error decreases as n_v increases; however, the rate at which the standard error decreases also decreases as n_v increases. Furthermore, compared to the other methods, the resulting standard error from the HBZIP-BMA method is lower and closer to the standard errors obtained from the complete data. The lower standard error for the HBZIP-BMA model suggests that distribution of the true count data is better reflected by the HBZIP-BMA model than the other imputation models.

Both linear sightability models yield approximately 95% confidence interval coverage when $n_v > 2$, the actual visibility follows the linear model, and the data are not zero-inflated. However, when visibility follows the non-linear model and n_v is large, overcoverage is observed for both sightability models. The HBZIP-BMA method yielded confidence interval coverages that are only slightly below 95% though they are relatively close to the resulting confidence interval coverage from the complete data. Except for the spline sightability model, all methods yield comparable confidence intervals, especially when n_v is large.

Recall that the sightability estimator of habitat occupancy rate is not available; therefore, only simulation results from the complete data, normal and HBZIP imputation are reported. Simulation results for habitat occupancy rate estimates are illustrated in Figure 5.4. The results suggest that estimates from the complete data are unbiased. In addition, the normal-based confidence interval should be avoided when $\pi \approx 0$ as its confidence interval coverage falls significantly below 95%. This is expected since the regular normal-based confidence interval coverage becomes highly unstable and converges to 0 as the proportion approaches the boundaries, 0 or 1 (Brown et al., 2001).

As shown in Figure 5.4, the normal model yields highly biased estimates when the data are sparse ($\pi > 0$), though the relative bias decreases significantly as n_v increases. Consistent with the simulation results presented in Chapter 3, normal imputation method should be avoided overall for habitat occupancy rate estimation. Furthermore, Figure 5.4 illustrates that the HBZIP-BMA yields approximately unbiased occupancy rates regardless of the sample size, visibility mechanism, and sparsity of the data. As illustrated in the figure, overall, the confidence interval coverage from the HBZIP method converges to its nominal value as n_v increases.

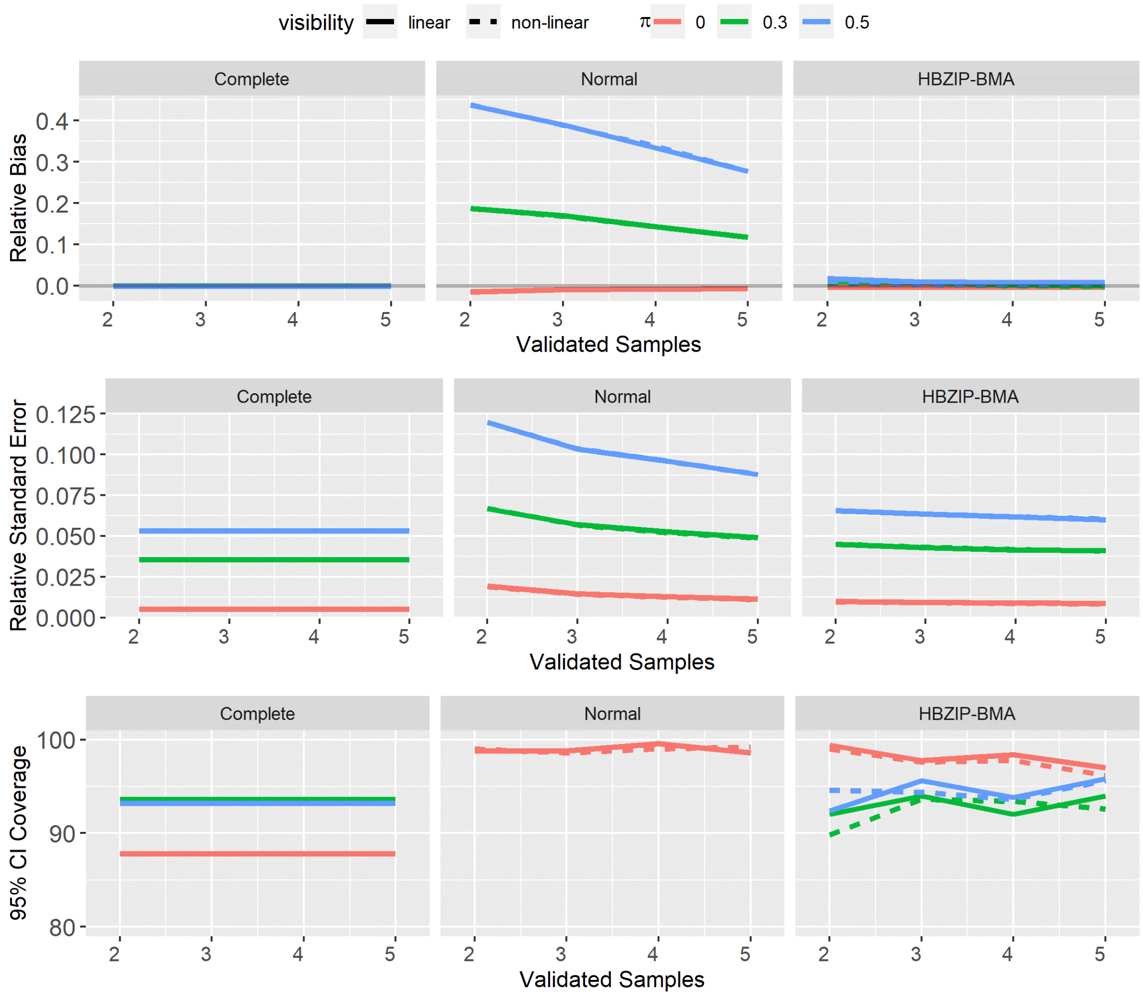


Figure 5.4: Simulation results for habitat occupancy rate estimates by correction method and measuring criteria: relative bias (first row), relative standard error (second row), and 95% confidence interval coverage (third row). Within each plot, visibility mechanism is indicated by the line type (dashed or solid), while zero-inflation is indicated by the different color.

5.5. Conclusion

In this chapter we applied HBZIP with BMA implementation on real moose population survey data. We conducted a simulation study to understand the properties of the different measurement error correction methods for total abundance and habitat occupancy rate estimation. The real data analysis suggests that estimates obtained using HBZIP-BMA imputed data are consistent with estimates obtained using the sightability adjustment methods. Furthermore, results from the simulation study suggest that HBZIP-BMA performs relatively well regardless of the underlying visibility mechanism and sparsity of the data for estimating both total abundance and habitat occupancy rate. The normal imputation model is also robust for estimating total abundance but is highly biased for estimating habitat occupancy rate. Thus, we have shown through both real data analysis and simulation study that multiple imputation utilizing both Bayesian hierarchical models and Bayesian model averaging together can be a viable approach for mitigating measurement error in zero-inflated count data.

Chapter 6
Conclusions and Future Direction

The way in which we gather and collect our data has become increasingly more complex; therefore, a more flexible and advanced measurement error correction method is necessary so that more relevant information can be added into the process. This is often the case with count data from animal population surveys—they are error-prone and collected through complex survey operations involving multiple instruments and observers. Therefore, a more advanced and flexible correction method is necessary to combine data, augment information, and model the complex relationship among the variables.

In this dissertation, we examined a flexible multiple imputation approach for mitigating measurement error in count data collected from animal population surveys. We explored off-the-shelf imputation methods and developed specialized imputation models to handle zero-inflated count data subject to visibility bias. We compared the methods' performance for estimating total abundance and habitat occupancy rate in a simulation study. We found that the off-the-shelf normal imputation model is robust and consistently produced favorable results for total abundance estimation. Though its standard error is slightly larger than our specialized imputation model (HBZIP), abundance estimates from the normal imputation model are approximately unbiased and have confidence interval coverage that is close to its nominal value. However, the normal imputation model should be avoided when the parameter of interest in the estimation is habitat occupancy rate.

In this research, we have shown that the HBZIP imputation model is effective for handling visibility bias, though it is sensitive to visibility model misspecification. To reduce the impact of visibility model misspecification, we incorporated Bayesian model averaging into the HBZIP model. We implemented the method on real data collected from moose population surveys and conducted a simulation study based on these data. The results showed that HBZIP with BMA implementation produced estimates that have desirable properties—approximately unbiased with low standard error regardless of the visibility model and sparsity of the data. In addition, total abundance estimates from HBZIP-BMA were also consistent with estimates from the sightability adjustment method. Furthermore, unlike the sightability adjustment method, HBZIP-BMA also allows analysis of the data to be performed using the same statistical methods as the ones used for complete data. This suggests that HBZIP-BMA offers more flexibility and can be a viable alternative for handling visibility bias in animal population surveys.

More work can be done to generalize the use of method. The correction methods presented in this research assume that validation data are a subset of the main survey. That is, validation data are collected internally, rather than externally solely for the purpose of estimating the visibility model. This assumption is inherent in the regular imputation framework; both complete and missing data are part of the same survey and used in the estimation. Thus, more research can be done to allow external validation data such that only the imputed data are used for inference. In this case, the modeling process remains unchanged. However, the standard method for combining the estimates may no longer be appropriate (Reiter and Raghunathan, 2007) and other combining rules may be explored.

While the non-linear visibility model has been shown to perform well in our simulation study, other simpler and more efficient models may also be considered, one of which is the natural cubic spline function, similar to the one used within the sightability adjustment method. Other non-linear models from the class of generalized additive models (Hastie and Tibshirani, 1986) may also work. In generalized additive models, the linear predictor is replaced with additive smoothing functions which allow for non-linear relationship between the predictor and the response variable. Therefore, more research can be done to explore alternative non-linear visibility models to be included within the HBZIP model.

Finally, correcting visibility bias is not a trivial task due to the complexity of the survey operation and animal behavior that varies by species and habitat. Though our method was initially developed to handle visibility bias in underwater video transects survey of red snapper, we demonstrated that it can also be adapted for aerial surveys of moose populations. We believe that, given similar data and measurement error structure, the method can also be adapted for other wildlife population surveys such as aerial survey of beluga whale in Alaska (Shelden and Wade, 2020) or elk in Oregon (Biederbeck et al., 2020). Finally, our research highlights the utility provided by Bayesian hierarchical models and flexibility of multiple imputation approach for measurement error correction in wildlife population surveys. We hope that techniques and results presented in this research will contribute to and motivate the development of other multiple imputation for measurement error correction techniques for other types of surveys.

Appendix A

Marginal Distribution of Y Derivation

Given $Y|N \sim Binomial(N,P)$ and $N \sim Poisson(\lambda)$, the marginal distribution of Y in (2.4) is derived as follows:

$$\begin{aligned}
P(Y) &= \sum_{n=y}^{\infty} P(Y|N)P(N) \\
&= \sum_{n=y}^{\infty} \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} [\pi + (1-\pi)e^{-\lambda}]^{n=0} [(1-\pi)e^{-\lambda}\lambda^n (n!)^{-1}]^{n>0} \\
&= \begin{cases} \pi + (1-\pi)e^{-\lambda}, & n = 0, y = 0 \\ \sum_{n=y}^{\infty} \frac{[(1-\pi)e^{-\lambda}\lambda^n]}{y!(n-y)!} p^y (1-p)^{n-y}, & n > 0, y = 0, 1, .., n \end{cases} \\
&= \begin{cases} \pi + (1-\pi)e^{-\lambda}, & n = 0, y = 0 \\ \frac{[e^{-\lambda}(1-\pi)p^y]}{y!} \sum_{n=y}^{\infty} \frac{\lambda^n}{(n-y)!} (1-p)^{n-y}, & n > 0, y = 0 \\ \frac{[e^{-\lambda}(1-\pi)p^y]}{y!} \sum_{n=y}^{\infty} \frac{\lambda^n}{(n-y)!} (1-p)^{n-y}, & n > 0, y = 1, ..n \end{cases}
\end{aligned}$$

Let $d = n - y$

$$= \begin{cases} \pi + (1-\pi)e^{-\lambda}, & y = 0, d = 0 \\ \frac{[e^{-\lambda}(1-\pi)(\lambda p)^y]}{y!} \sum_{d=1}^{\infty} \frac{\lambda^d (1-p)^d}{d!}, & y = 0, d = 1,2,3,.. \\ \frac{[e^{-\lambda}(1-\pi)(\lambda p)^y]}{y!} \sum_{d=0}^{\infty} \frac{\lambda^d (1-p)^d}{d!}, & y > 0, d = 0,1,2,3,.. \end{cases}$$

$$= \begin{cases} \pi + (1-\pi)e^{-\lambda}, & y = 0 \\ [e^{-\lambda}(1-\pi)](e^{\lambda(1-p)} - 1), & y = 0 \\ \frac{[e^{-\lambda}(1-\pi)(\lambda p)^y]}{y!}(e^{\lambda(1-p)}), & y > 0 \end{cases}$$

$$= \begin{cases} \pi + (1-\pi)e^{-\lambda p}, & y = 0 \\ (1-\pi)\frac{(\lambda p)^y e^{-\lambda p}}{y!}, & y > 0 \end{cases}$$

Hence, $Y \sim ZIP(\pi, \lambda p)$ ■

Appendix B

Derivation of the Conditional Distribution $N|Y$

In (2.8), we defined the conditional distribution of $N|Y$ which can be derived using the Bayes theorem in (2.5) and the results from (2.6) and (2.7). Recall from (2.5), $P(N|Y) = \frac{P(Y,N)}{P(N)}$. Suppose we separate $P(N|Y)$ in three cases for $(n=0, y=0)$, $(n>0, y=0)$, and $(n>0, y>0)$, then the conditional distribution of $N|Y$ is derived as follows:

$$\begin{aligned}
P(N|Y) &= \frac{P(Y,N)}{P(N)} \\
&= \begin{cases} \frac{\pi+(1-\pi)e^{-\lambda}}{\pi+(1-\pi)e^{-\lambda p}}, & n=0, y=0 \\ \frac{[(1-\pi)e^{-\lambda}\lambda^n]}{y!(n-y)!} p^y (1-p)^{n-y} \times \frac{1}{\pi+(1-\pi)e^{-\lambda p}}, & n>0, y=0 \\ \frac{[(1-\pi)e^{-\lambda}\lambda^n]}{y!(n-y)!} p^y (1-p)^{n-y} \times \frac{1}{\pi+(1-\pi)e^{-\lambda p}}, & n>0, y>0 \end{cases} \\
&= \begin{cases} \frac{\pi e^{\lambda p}}{\pi e^{\lambda p}+(1-\pi)} + \frac{1-\pi}{\pi e^{\lambda p}+(1-\pi)} e^{-\lambda(1-p)}, & n=0, y=0 \\ \frac{1-\pi}{\pi e^{\lambda p}+(1-\pi)} e^{-\lambda(1-p)} \frac{[\lambda(1-p)]^n}{n!}, & n>0, y=0 \\ \frac{e^{-\lambda(1-p)}}{(n-y)!} [\lambda(1-p)]^{n-y}, & n>0, y>0 \end{cases}
\end{aligned}$$

Therefore, $(N|Y=y) \sim ZIP\left(\frac{\pi e^{\lambda p}}{\pi e^{\lambda p}+(1-\pi)}, \lambda[1-p]\right)$ if $y=0$ and
$(N|Y=y) \sim y + Poisson(\lambda[1-p])$ if $y>0$. ■

Appendix C
HBZIP Model and Prior Specifications

Models within the HBZIP are defined as follows:

True Count:

$$N_i \sim Poisson(\lambda_i, \pi_i)$$

Observed Count:

$$Y_i | N_i \sim Binomial(N_i, P_i)$$

Priors for Local Parameters:

$$\phi_i = log(\lambda_i) \sim N(X^{(\phi)}\beta, \sigma^2_\phi)$$

$$\nu_i = log(\pi_i) \sim N(X^{(\nu)}\gamma, \sigma^2_\nu)$$

$$\eta_i = log(\pi_i) \sim N(X^{(\eta)}\theta, \sigma^2_\eta)$$

Priors for Hyperparameters:

$$\beta_0, \beta_1, .., \beta_{p1} \sim N(0, 10000)$$

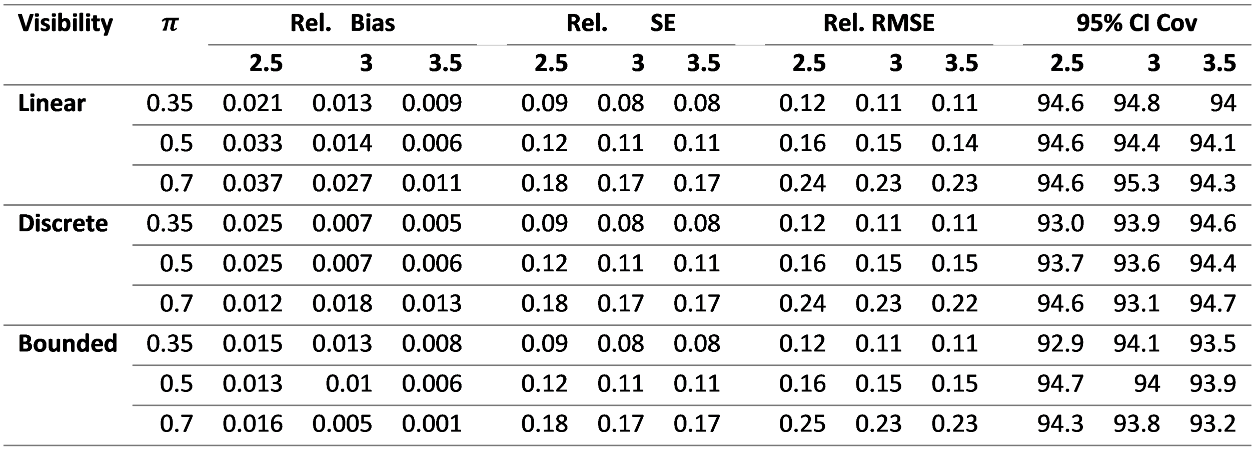
$$\gamma_0, \gamma_1, .., \gamma_{p2} \sim N(0, 10000)$$

$$\theta_0, \theta_1, .., \theta_{p3} \sim N(0, 10000)$$

$$\sigma^2_\phi, \sigma^2_\nu, \sigma^2_\eta \sim Uniform(0, 100)$$

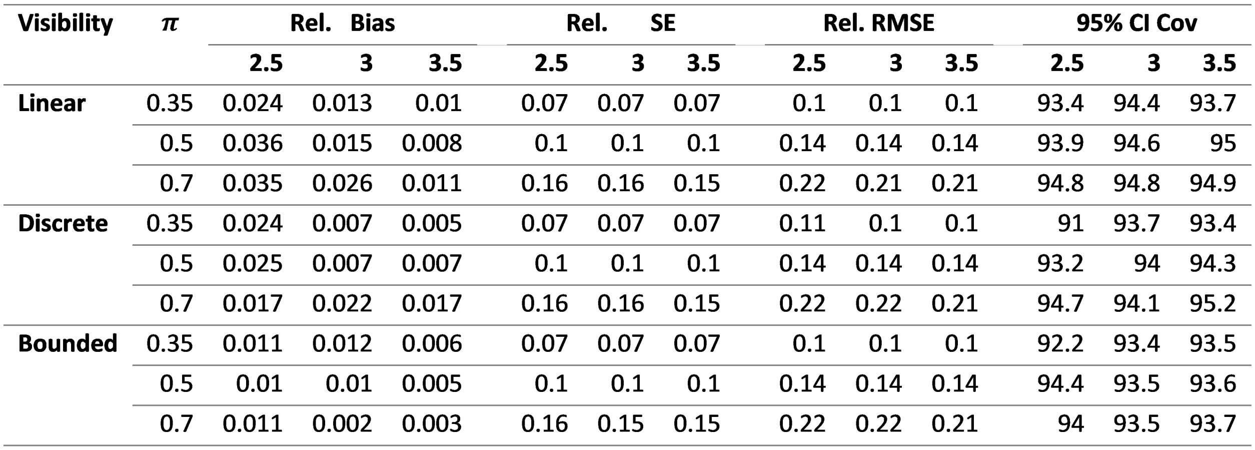
Appendix D

Simulation Results for HBZIP-BMA



Visibility	π	Rel. Bias			Rel. SE			Rel. RMSE			95% CI Cov		
		2.5	3	3.5	2.5	3	3.5	2.5	3	3.5	2.5	3	3.5
Linear	0.35	0.021	0.013	0.009	0.09	0.08	0.08	0.12	0.11	0.11	94.6	94.8	94
	0.5	0.033	0.014	0.006	0.12	0.11	0.11	0.16	0.15	0.14	94.6	94.4	94.1
	0.7	0.037	0.027	0.011	0.18	0.17	0.17	0.24	0.23	0.23	94.6	95.3	94.3
Discrete	0.35	0.025	0.007	0.005	0.09	0.08	0.08	0.12	0.11	0.11	93.0	93.9	94.6
	0.5	0.025	0.007	0.006	0.12	0.11	0.11	0.16	0.15	0.15	93.7	93.6	94.4
	0.7	0.012	0.018	0.013	0.18	0.17	0.17	0.24	0.23	0.22	94.6	93.1	94.7
Bounded	0.35	0.015	0.013	0.008	0.09	0.08	0.08	0.12	0.11	0.11	92.9	94.1	93.5
	0.5	0.013	0.01	0.006	0.12	0.11	0.11	0.16	0.15	0.15	94.7	94	93.9
	0.7	0.016	0.005	0.001	0.18	0.17	0.17	0.25	0.23	0.23	94.3	93.8	93.2

Table D.1: HBZIP-BMA simulation results for total abundance estimates for all β_o, π, and all visibility models.

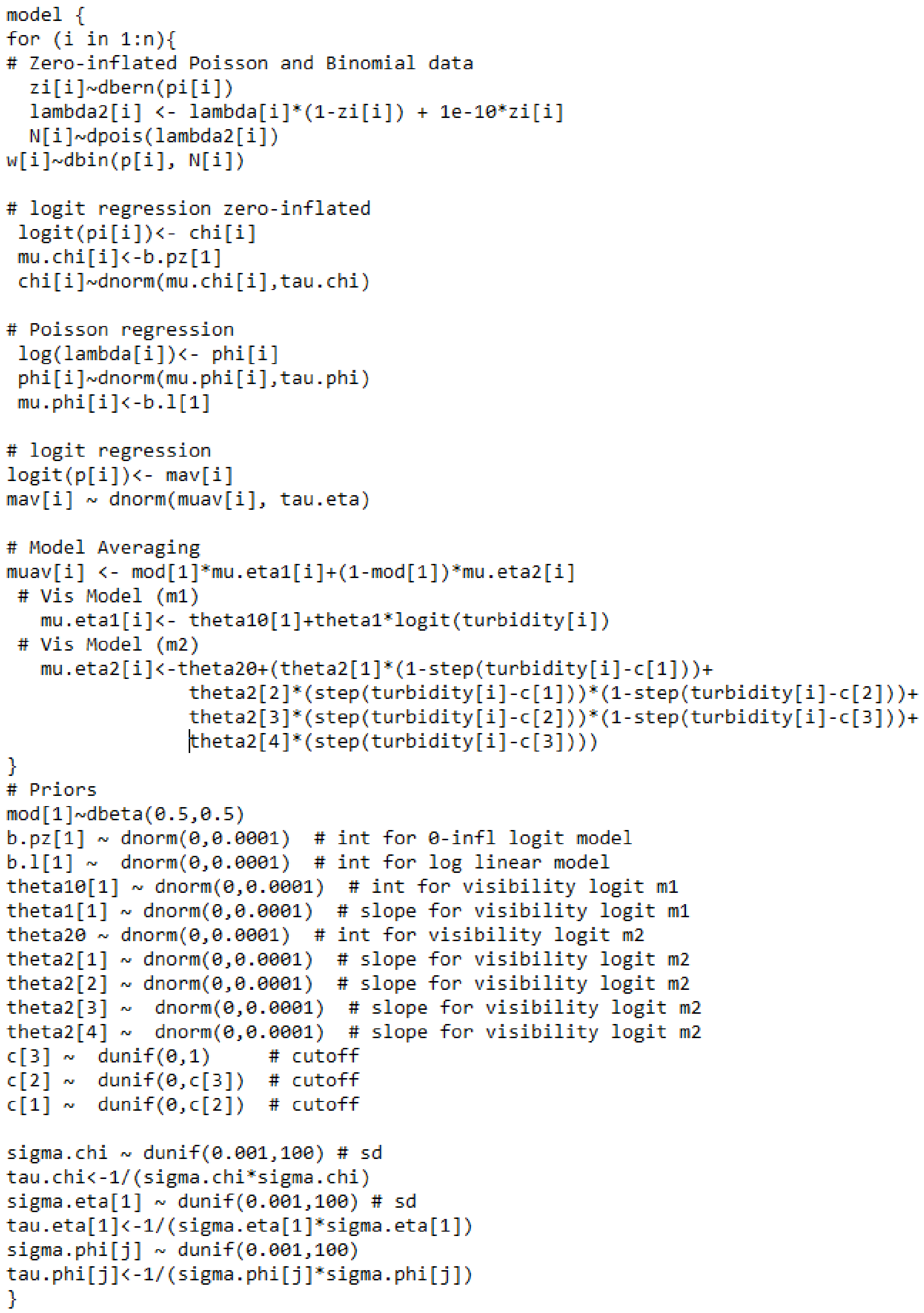


Visibility	π	Rel. Bias			Rel. SE			Rel. RMSE			95% CI Cov		
		2.5	3	3.5	2.5	3	3.5	2.5	3	3.5	2.5	3	3.5
Linear	0.35	0.024	0.013	0.01	0.07	0.07	0.07	0.1	0.1	0.1	93.4	94.4	93.7
	0.5	0.036	0.015	0.008	0.1	0.1	0.1	0.14	0.14	0.14	93.9	94.6	95
	0.7	0.035	0.026	0.011	0.16	0.16	0.15	0.22	0.21	0.21	94.8	94.8	94.9
Discrete	0.35	0.024	0.007	0.005	0.07	0.07	0.07	0.11	0.1	0.1	91	93.7	93.4
	0.5	0.025	0.007	0.007	0.1	0.1	0.1	0.14	0.14	0.14	93.2	94	94.3
	0.7	0.017	0.022	0.017	0.16	0.16	0.15	0.22	0.22	0.21	94.7	94.1	95.2
Bounded	0.35	0.011	0.012	0.006	0.07	0.07	0.07	0.1	0.1	0.1	92.2	93.4	93.5
	0.5	0.01	0.01	0.005	0.1	0.1	0.1	0.14	0.14	0.14	94.4	93.5	93.6
	0.7	0.011	0.002	0.003	0.16	0.15	0.15	0.22	0.22	0.21	94	93.5	93.7

Table D.2: HBZIP-BMA simulation results for occupancy rate estimates for all β_o, π, and all visibility models.

Appendix E

Jags Code for HBZIP-BMA



```
model {
for (i in 1:n){
# Zero-inflated Poisson and Binomial data
  zi[i]~dbern(pi[i])
  lambda2[i] <- lambda[i]*(1-zi[i]) + 1e-10*zi[i]
  N[i]~dpois(lambda2[i])
w[i]~dbin(p[i], N[i])

# logit regression zero-inflated
 logit(pi[i])<- chi[i]
 mu.chi[i]<-b.pz[1]
 chi[i]~dnorm(mu.chi[i],tau.chi)

# Poisson regression
 log(lambda[i])<- phi[i]
 phi[i]~dnorm(mu.phi[i],tau.phi)
 mu.phi[i]<-b.l[1]

# logit regression
logit(p[i])<- mav[i]
mav[i] ~ dnorm(muav[i], tau.eta)

# Model Averaging
muav[i] <- mod[1]*mu.eta1[i]+(1-mod[1])*mu.eta2[i]
 # Vis Model (m1)
   mu.eta1[i]<- theta10[1]+theta1*logit(turbidity[i])
 # Vis Model (m2)
   mu.eta2[i]<-theta20+(theta2[1]*(1-step(turbidity[i]-c[1]))+
                theta2[2]*(step(turbidity[i]-c[1]))*(1-step(turbidity[i]-c[2]))+
                theta2[3]*(step(turbidity[i]-c[2]))*(1-step(turbidity[i]-c[3]))+
                theta2[4]*(step(turbidity[i]-c[3])))
}
# Priors
mod[1]~dbeta(0.5,0.5)
b.pz[1] ~ dnorm(0,0.0001)  # int for 0-infl logit model
b.l[1] ~  dnorm(0,0.0001)  # int for log linear model
theta10[1] ~ dnorm(0,0.0001)  # int for visibility logit m1
theta1[1] ~ dnorm(0,0.0001)  # slope for visibility logit m1
theta20 ~ dnorm(0,0.0001)  # int for visibility logit m2
theta2[1] ~ dnorm(0,0.0001)  # slope for visibility logit m2
theta2[2] ~ dnorm(0,0.0001)  # slope for visibility logit m2
theta2[3] ~  dnorm(0,0.0001)  # slope for visibility logit m2
theta2[4] ~  dnorm(0,0.0001)  # slope for visibility logit m2
c[3] ~  dunif(0,1)     # cutoff
c[2] ~  dunif(0,c[3])  # cutoff
c[1] ~  dunif(0,c[2])  # cutoff

sigma.chi ~ dunif(0.001,100) # sd
tau.chi<-1/(sigma.chi*sigma.chi)
sigma.eta[1] ~ dunif(0.001,100) # sd
tau.eta[1]<-1/(sigma.eta[1]*sigma.eta[1])
sigma.phi[j] ~ dunif(0.001,100)
tau.phi[j]<-1/(sigma.phi[j]*sigma.phi[j])
}
```

Appendix F

Estimates and standard errors of moose abundance in northeastern Minnesota

Method	**Parameter**	**Year**		
		2005	**2006**	**2007**
Naive approach	$\hat{\tau}$	3,780	4,385	3,775
	SE[$\hat{\tau}$]	487	481	405
	RSE[$\hat{\tau}$]	(12.9%)	(11%)	(10.7%)
Sightability (linear)	$\hat{\tau}$	8158	8840	6917
	SE[$\hat{\tau}$]	1574	1523	1154
	RSE[$\hat{\tau}$]	(19.3%)	(17.2%)	(16.7%)
Sightability (splines)	$\hat{\tau}$	8459	9641	7739
	SE[$\hat{\tau}$]	1584	1716	1406
	RSE[$\hat{\tau}$]	(18.7%)	(17.8%)	(18.2%)
HBZIP-BMA	$\hat{\tau}$	8216	9850	8448
	SE[$\hat{\tau}$]	1295	1498	1176
	RSE[$\hat{\tau}$]	(15.8%)	(15.2%)	(13.9%)
Normal	$\hat{\tau}$	8349	9262	7509
	SE[$\hat{\tau}$]	1104	1282	1029
	RSE[$\hat{\tau}$]	(13.2%)	(13.8%)	(13.7%)

Table F.1: Estimates of total moose abundance, their standard errors and relative standard error represented in %. Sightability estimates were obtained using the *SightabilityModel* package (Fieberg, 2012).

Bibliography

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey nonresponse. *International Statistical Review*, 78(1):40–64.

ArchMiller, A. A., Dorazio, R. M., and Fieberg, J. R. (2018). Time series sightability modeling of animal populations.

Biederbeck, H. H., DeWaine, H. J., and VandeBergh, D. J. (2020). Aerial high resolution digital imagery elk survey. *Wildlife Technical Report 006-2016, Oregon Department of Fish and Wildlife. Retrieved*, 11.

Blackwell, M., Honaker, J., and King, G. (2017). Multiple overimputation: A unified approach to measurement error and missing data. *Sociological Methods & Research*, 46(3):342–369.

Brick, J. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3).

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for binomial proportion.

Buonaccorsi, J. P. (2010). Measurement error: Models, methods, and applications. page 300.

Chambers, J. M. and Hastie, T. J. (1992). Chapter 7 of statistical models in s. In *T. J.*

Cole, S. R., Chu, H., and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, 35(4):1074–1081.

Fieberg, J. (2012). Estimating population abundance using sightability models: R SightabilityModel package. *Journal of Statistical Software*, 51(9):1–20.

Fieberg, J., Alexander, M., Tse, S., and Clair, K. S. (2013). Abundance estimation with sightability data: a Bayesian data augmentation approach.

Gaffert, P., Meinfelder, F., and Bosch, V. (2016). Towards an MI-proper Predictive Mean Matching (Discussion paper).

Ghosh-Dastidar, B. and Schafer, J. L. (2003). Multiple edit/multiple imputation for multivariate continuous data. *Journal of the American Statistical Association*, 98(464):807–817.

Giudice, J., Fieberg, J. R., and Lenarz, M. S. (2012). *Spending Degrees of Freedom in a Poor Economy: A Case Study of Building a Sightability Model for Moose in Northeastern Minnesota.* The Journal of Wildlife Management 76(1):75-87; 2012; DOI: 10.1002/jwmg.213.

Goodyear, C. P. (1988). Recent trends in the red snapper fishery of the gulf of mexico. *National Marine Fisheries Service, Southeast Fisheries Science Center, Miami, Florida. Technical Report CRD*, 87/88-16.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statist. Sci., Volume 1, Number 3*, 1:297–310.

He, Y. and Raghunathan, T. E. (2006). Tukey's gh Distribution for Multiple Imputation. *The American Statistician*, 60:3.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14.

Hosmer, D., Hosmer, T., Cessie, S., and Lemeshow, S. (1997). Hosmer, d., hosmer, t., cessie. *S. L., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model.. Statistics in medicine*, 16:965–80.

Kim, J. and Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, 91:559–578.

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

Lambert, D. (1992). Zero-inflated poisson regression. *with an Application to Defects in Manufacturing. Technometrics*, 34(1):2307/1269547.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics. 6 (3), 287-296. doi:10*, 6:2307/1391878.

Lohr, S. (2009). Sampling: Design and analysis. *Cengage Learning*, 79::.

McGee, M. and Bergasa, N. (2006). Analysis of a pilot study for amelioration of itching in liver disease: When is a failed trial not a failure? *The American Statistician, Nov*, 60(4):303–308.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). R. *A review of spline function procedures in. BMC Med Res Methodol*, 19(46.).

Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. *v3. 2. 0. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing. vol. 124. Vienna;*, 2.(0.).

Raghunathan, T. E., Lepkowski, J., Hoewyk, V., H., J., and Solenberger, P. W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(2).

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–471.

Rodwell, L., Lee, K. J., Romaniuk, H., and Carlin, J. B. (2014). Comparison of methods for imputing limited-range variables: a simulation study. *BMC Medical Research Methodology*, 14(1).

Royle, J. A. and Dorazio, R. M. (2006). Hierarchical models of animal abundance and occurrence. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):249–263.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley and Sons.

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

Samuel, M. D., Steinhorst, R. K., Garton, E. O., and Unsworth, J. W. (1992). Estimation of wildlife population ratios incorporating survey design and visibility bias. *The Journal of Wildlife Management*, 56(4):2307/3809465.

Shelden, K. E. and Wade, P. R. (2020). Aerial surveys, distribution, abundance, and trend of belugas (Delphinapterus leucas) in Cook Inlet, Alaska. 11.

Siddique, J. and Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. pages 83–102.

Steinhorst, R. K. and Samuel, M. D. (1989). Sightability adjustment methods for aerial surveys of wildlife populations. *Biometrics*, 45(2):2307/2531486.

Stunz et al. (2017). *Estimating the absolute abundance of Red Snapper in the U.S Gulf of Mexico.* (Research proposal). Texas A&M University-Corpus Christi, Harte Research Institute for the Gulf of Mexico.

Su, Y. and Yajima, M. (2020). *R2jags: Using R to Run 'JAGS'.* https://CRAN.R-project.org/package=R2jags'.

U.S. Census Bureau (2016). Imputation of unreported data items. Accessed: 02-3-2020, https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/imputation-of-unreported-data-items.html.

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3).

von Hippel and T., P. (2013). Should a normal imputation model be modified to impute skewed variables.