Southern Methodist University
SMU Scholar

Statistical Science Theses and Dissertations | Statistical Science

Spring 5-18-2019

Samples, Unite! Understanding the Effects of Matching Errors on Estimation of Total when Combining Data Sources

Benjamin Williams
Southern Methodist University, benjamin@smu.edu

Follow this and additional works at: https://scholar.smu.edu/hum_sci_statisticalscience_etds

Part of the Applied Statistics Commons, and the Statistical Methodology Commons

Recommended Citation
Williams, Benjamin, "Samples, Unite! Understanding the Effects of Matching Errors on Estimation of Total when Combining Data Sources" (2019). *Statistical Science Theses and Dissertations*. 5.
https://scholar.smu.edu/hum_sci_statisticalscience_etds/5

This Dissertation is brought to you for free and open access by the Statistical Science at SMU Scholar. It has been accepted for inclusion in Statistical Science Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.

SAMPLES, UNITE!

UNDERSTANDING THE EFFECTS OF MATCHING ERRORS ON ESTIMATION OF TOTAL WHEN COMBINING DATA SOURCES

Approved by:

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

Dr. Robert Bell
Retired Statistician (Google, AT&T Labs)

Dr. Daniel Heitjan
Professor of Statistical Science (SMU)

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

Dr. Mary Mulry
Principal Researcher (US Census Bureau)

SAMPLES, UNITE!

UNDERSTANDING THE EFFECTS OF MATCHING ERRORS ON ESTIMATION OF TOTAL WHEN COMBINING DATA SOURCES

A Dissertation Presented to the Graduate Faculty of the

Dedman College: School of Humanities and Sciences

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Statistical Science

by

Benjamin M. Williams

B.S. Mathematics, Wheaton College, IL
B.A. Economics, Wheaton College, IL
M.S. Statistics, Southern Methodist University

May 18, 2019

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Stokes, for supplying me with the research topic that became this dissertation. Without her guidance and help, I would not have completed this work. I also would like to thank my committee members for their thoughtful comments throughout the process of writing this dissertation. Thank you to my parents and sister for always encouraging and supporting me. My wife, Alyssa, deserves immense thanks, she has been completely encouraging and supportive throughout my entire time at SMU. She deserves credit for this dissertation right alongside me. Finally, thanks be to God for sustaining me during my 5 years in graduate school. *"For in him all things were created"* (Col 1:16)

Williams , Benjamin M.

B.S. Mathematics, Wheaton College, IL
B.A. Economics, Wheaton College, IL
M.S. Statistics, Southern Methodist University

Samples, Unite!
Understanding the Effects of Matching Errors on Estimation of Total When Combining Data Sources

Advisor: Professor S. Lynne Stokes

Doctor of Philosophy degree conferred May 18, 2019

Dissertation completed April 29, 2019

Much recent research has focused on methods for combining a probability sample with a non-probability sample to improve estimation by making use of information from both sources. If units exist in both samples, it becomes necessary to link the information from the two samples for these units. Record linkage is a technique to link records from two lists that refer to the same unit but lack a unique identifier across both lists. Record linkage assigns a probability to each potential pair of records from the lists so that principled matching decisions can be made. Because record linkage is a probabilistic endeavor it introduces randomness into estimators that use the linked data. The effects of this randomness on regression involving the linked datasets has been examined (for example: Lahiri and Larsen, 2005). However, the effect of matching error has not been considered for the case of estimating the total of a population from a capture-recapture model. In this dissertation we present a general model for matching errors arising from a linkage procedure and examine the effects on bias and variance of some estimators used for such scenarios.

Our work is motivated by the application of estimating fish catch in the Gulf of Mexico. The National Marine Fisheries Service (NMFS) estimates the total number of fish caught by recreational marine anglers. Currently, NMFS arrives at this by estimating from independent surveys the total effort (the number of fishing trips) and the catch per unit effort or *CPUE* (the number of fish caught per species per

trip), and then multiplying them together. Effort data are collected via a mail survey of potential anglers. *CPUE* data are collected via face-to-face intercepts of anglers completing fishing trips at randomly selected times/docks. The interviewers identify the catch totals of intercepted anglers by species.

The effort survey has a high non-response rate. It is also retrospective, which causes the entire estimation process to take more than a month, precluding in-season management. Due to these limitations, the NMFS is experimenting with replacing the effort survey with electronic self-reporting. The anglers report details of their trip via an electronic device and remain eligible to be sampled in the dockside intercept.

Several estimators have been proposed to estimate total catch using these self-reports alongside the dockside intercept using capture-recapture methodology (Liu et al., 2017). For the estimators to be valid, the records from trips that both self-reported and were sampled in the intercept survey must be linked. The self-reported data is a non-probability sample because it is voluntarily submitted and can be considered as a big data source, while the dockside intercept is a smaller probability sample. Liu et al. assumed perfect matching, however this is difficult in practice due to device and measurement error. Currently, the effect of potential matching errors on the estimators is unknown.

In this research, we develop a novel model to investigate the effect matching errors have on the bias and mean square error of the estimators. We describe and implement a record linkage algorithm for our pilot study data following the work of Bell et al. (1994). Then we discuss two other estimators appropriate for scenarios when either there is no undercoverage or angler reporting is completely accurate (Breidt et al., 2018). Finally, we introduce a simulation study and future research plans.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

For Alyssa, who stabilizes my mean though my variance is large.

Chapter 1

Introduction

Statistical sampling plays a vital role in understanding and making inferences for all types of populations. Sampling methods allow for accurate conclusions to be drawn about some population while only observing a subset of the group of interest. Whether examining bacterial spores in the arctic or Internet users in Ecuador, sampling is invaluable. It is especially salient in a world where populations are large and big data is the new standard in both academia and industry. With such large datasets to study, sampling is critical for accurate and rigorous research. More importantly, sampling allows estimates of accuracy and precision. Without a valid probability sample, one cannot know how wrong (or right) they are.

Ideally, samplers have access to a list of all the units in a population, called a sampling frame, from which to draw a sample. This allows them to select individual population units with known probabilities, producing a probability sample (Lohr, 2010). A sample for which probabilities of selection are not known is called a non-probability sample. Probability samples are preferred to non-probability samples because the sampling variance of the estimators calculated from probability samples can be determined using standard sampling theory. Even non-sampling errors, such as non-response bias, may be easier to assess and mitigate for probability samples. The primary disadvantage of non-probability samples is the potential for biased estimation stemming from undercoverage and a lack of representativeness in the samples. Without external sources of information, these sample deficiencies cannot be detected.

It is usually easier to obtain a non-probability sample than a probability sample. For example, a frame may not be available, complicating the selection of a probability sample. A non-probability sample, such as one using volunteer data, can often be

cheaply obtained and have many sampled units. The larger the sample, the more information it may contain. This is an attractive option for analysts working with limited resources studying elusive populations. As a result, statisticians are investigating methods for improving estimation using data from non-probability samples.

1.1. Research Objectives

This research focuses on combining a non-probability sample with a probability sample to improve estimation. The motivating research problem comes from the National Oceanic and Atmospheric Administration (NOAA). NOAA is interested in estimating the total number of fish caught by recreational anglers from the Gulf of Mexico. NOAA uses data from probability samples of anglers to estimate this catch. These estimates are paramount in the process of setting fishing season lengths and bag limits (the number of fish allowed to be caught per trip). Noted problems in the prevailing estimation process, however, include the lengthy time necessary for the estimates to be made, increased costs due to falling response rates, undercoverage due to incomplete sampling frames, and potential nonresponse bias. To combat this, NOAA is experimenting with a new sampling scheme allowing recreational anglers to self-report their trip data via an electronic device. This constitutes a non-probability sample and requires special methods for valid estimation.

The experimental sampling scheme combines the non-probability sample of reports with a probability sample of intercepts of recreational anglers on the dock after they complete their fishing trip. The sample is chosen using a frame of docks and time periods along the Gulf. The intercept sample, called APAIS (Access Point Angler Intercept Survey), uses a frame consisting of only public docks along the Gulf of Mexico, while the self-reports come from trips returning to both public and private docks. The population of interest is all trips completed at a dock in the Gulf by recreational anglers, regardless of whether the trip returned to a public or private dock. The self-reports provide information about trips made to private docks. Because the

intercept sample can be carried out at public docks only, it is insufficient on its own to estimate the total catch.

Recently, Liu et al (2017) developed estimators of total for this scenario. They propose several ways to make use of both the voluntary sample of self-reports and the probability sample, by regarding the data from the self-reports as auxiliary information. Then the sampling properties of their estimators can be evaluated using standard sampling methodology. Their estimators require the identification of trips observed in both samples. Matches occur when a captain both self-reports and is intercepted by a sampler at the dock. It is challenging in practice to identify such trips. Liu et al. did not investigate the effects of matching errors on their estimators.

In this research, we evaluate the effect of matching errors on the estimators proposed by Liu et al., develop a systematic method to match units between the samples, examine additional estimators that are appropriate when there is no undercoverage of the probability sample, and examine the effect of matching error on these additional estimators. The remainder of this chapter describes the motivating example, including how electronic reporting works and why matching trips between the samples is difficult. It also reviews the estimators proposed by Liu et al. (2017). In Chapter 2, we examine the effect matching errors have on the bias and mean square error of the estimators. Chapter 3 includes a literature review of record linkage, a systematic matching method, and an example of how the method has been implemented with our pilot study data. Chapter 4 discusses two other estimators appropriate for scenarios when either there is no undercoverage or angler reporting is completely accurate. Chapter 5 introduces a simulation study, and Chapter 6 presents future research plans.

1.2. Motivating Example

Fisheries in the United States can be divided into two major categories: commercial and recreational. Commercial fisheries are required to report their total catch.

This is not the case for recreational anglers. For some areas and species, the total number of fish caught by recreational anglers can exceed the catch by commercial fisheries (National Research Council, 2017).

Knowing the total fish catch is critical for setting fishing season lengths, bag limits, acceptable biological catch, and the annual catch limit for the various species of fish (National Research Council, 2017). The estimates of catch are inputs to models of fish abundance. Precise estimation of removals (both catch and mortality from discards) improves the performance of these models. Outputs from the models are used for fisheries management to keep population levels stable, prevent overfishing, and combat effects of natural disasters such as oil spills, which can negatively affect fish populations (Tarnecki and Patterson, 2015).

1.2.1. Current Methods

The National Marine Fisheries Service (NMFS), a part of NOAA, estimates the total catch of fish by recreational anglers via the Marine Recreational Information Program (MRIP). MRIP uses two surveys to collect data to estimate total catch. The first is the APAIS, used only for estimating catch per unit effort ($CPUE$) which is the average catch of each species per angler trip. The second is a probability sample of households called the fishing effort survey (FES). It is used to estimate the effort (E) defined as the number of angling trips made. Then the catch is estimated as the product of the two estimates ($\hat{CPUE}$ and $\hat{E}$).

The primary sampling unit (PSU) for the APAIS is a combination of public dock locations and time on a specific day within a two month time period called a wave. The secondary sampling unit (SSU) is a trip made by a recreational angler. The PSUs are selected with a probability proportional to size (number of trips at each PSU) sample design. Many samples are selected and the ones not satisfying certain resource constraints are discarded. One sample is selected at random from these remaining samples for implementation. This constitutes a rejective sampling technique (Fuller,

2009) and the initial probability of inclusion of each PSU is derived using Monte Carlo methods (Breidt and Chromy, 2016; National Research Council, 2017). Interviewers are sent to each selected PSU to interview all recreational anglers returning from a fishing trip. The interviewers record statistics such as the number of fish caught per species per angler, the number of anglers aboard the boat, the number of fish released per species, etc.

The FES is a survey sent by mail to residents living in states that border the marine body of interest. The FES is an address based survey, stratified by closeness to the shore and inclusion of the address on the National Saltwater Registry (NSAR). The NSAR is a registry that most marine recreational anglers are required to join to legally fish in marine waters. Optimal allocation is used to determine the sample sizes within strata. The FES is a sample of all potential recreational anglers, on the NSAR living in a state directly adjacent to the water. Recipients of the FES are asked to provide their effort (the number of times they went fishing) retrospectively for the previous two months for up to five members of the household (National Research Council, 2017). The total catch of fish by recreational anglers is estimated by multiplying $\hat{CPUE}$ and $\hat{E}$ for each species. An adjustment for anglers who live out-of-state is made using an estimated proportion of anglers residing in-state from the dockside intercept sample (APAIS).

This current methodology is the result of years of careful and intentional updates. The National Research Council (NRC) has twice been tasked with reviewing the MRIP and providing recommendations, once in 2006 and again in 2016. The current FES survey is actually an upgrade over its predecessor, the Coastal Household Telephone Survey (CHTS). The CHTS was a telephone based survey that did not stratify based on NSAR, but was a random digit dial survey from the population of residents of coastal counties. The CHTS suffered from declining response rates and undercoverage associated with many residents lacking a traditional landline telephone (Boyle et al., 2009). In the time between the two NRC reviews of the MRIP,

the FES replaced the CHTS, resulting in a three-fold increase in response rate. The response rate from a pilot study of the FES are approximately 35%, indicative of a potential presence of bias due to non-response (Andrews et al., 2014). Weighting class and poststratification adjustments are made to lessen the effect of undercoverage and non-response (National Research Council, 2017). There are still problems in the estimation procedure of the MRIP, many described in the most recent NRC review. We next discuss these issues and then describe the electronic reporting experiment.

1.2.2. Problems with Current Methods

The majority of the issues in the estimation lie with the FES. Though it is an improvement over the CHTS, the FES is far from perfect. The first flaw is non-response. The low response rate of 35% may cause biased estimates (Andrews et al., 2014). Another flaw has to do with measurement error. The FES is conducted at the end of a wave and asks respondents to recall the number of fishing trips made during that wave. This means they are asked to recall events that may have taken place over two months ago. This question may be difficult for respondents to answer accurately. Errors correlated with recall answers include how long ago the event (fishing trip) took place, the importance of the event and the ability to distinguish between events (Eisenhower et al., 1991). Recalling previous trips may be especially difficult for avid anglers, since it may lack salience for them.

The next issue has to do with the efficiency of the estimate production. The NMFS reports it takes approximately 45 days after each wave is completed for estimates of total catch to be made. This results from the time needed for the FES to be mailed out and for enough responses to be returned for valid estimation. Ideally, the estimates would be produced much faster in order to allow the timely setting of fishing limits.

1.2.3. Experiments with Electronic Reporting

In the 2016 NRC review, several recommendations were proposed to fix these flaws and improve estimation. The recommendation which motivates this research advised that electronic data collection methods be used as an alternative to the FES (National Research Council, 2017). The NRC noted these electronic reporting methods might allow for near real time estimation. In the Gulf of Mexico and other areas in the US, fisheries management institutions have begun experimenting with such techniques. The state fish and game agencies in Alabama and Mississippi require anglers to report their catch of the fish species Red Snapper, popular in the Gulf of Mexico (AL Department of Conservation and Natural Resources, 2019; MS Department of Marine Resources, 2019). In South Carolina, self-reporting is done via pen and paper log-books, but the principle of self-reporting remains (Breidt et al., 2018).

In the Gulf, the NMFS is experimenting in several states by asking captains to self-report their trips with an electronic device. These devices allow for many species of fish to be recorded. Our work is motivated by our involvement in this NMFS electronic reporting pilot study conducted in the Gulf of Mexico (AL, MS, and FL) in 2016 and 2017. NMFS has partnered with a private research firm, formerly known as CLS America (CLS) for this experiment. Recreational charter captains can volunteer to participate. Then CLS provides an electronic device to captains allowing them to self-report demographic and fishing data for their recreational fishing trips.

Because the self-reporting is done on an electronic device the data are available for estimation in nearly real time. The captains are instructed to report their data before the boat returns to a landing site at the conclusion of a trip, although this does not always occur. A captain can be selected into the intercept sample and also report her trip with the electronic device, meaning the trip can be present in both samples. The goal of the NMFS experiment is for the voluntary sample of reporting captains to be used in place the FES for estimation. However, unlike the FES, these reports are a non-probability sample and so the current estimation method is not valid. The

APAIS and self-report; however, are actually sufficient to estimate the total catch because they constitute a form of a capture-recapture model.

For valid estimation, sampled trips must be linked to trip reports. In our work as part of the NMFS study, matching the reports to intercepted trips was more difficult than anticipated. Other states, such as Alabama, have encountered difficulty in matching reports to trips as well. One state where matching is not an issue is Mississippi. In Mississippi, reporting is mandatory and without reporting a fishing trip, a recreational angler may not embark on another trip, this is discussed in more depth in Section 3.2.4 (of Marine Resources, 2017). NOAA is currently deciding whether to make reporting mandatory for recreational anglers, because it is not mandatory in all Gulf States.

Of primary concern in deciding whether or not to implement more electronic reporting and whether it should be mandatory is the effect of non-sampling errors on estimators. In this dissertation, we examine how matching errors affect estimators of total catch. We will determine if the effect of matching errors on the estimators is enough to disqualify their use. Other non-sampling errors include undercoverage of the sampling frame, as well as errors arising from a lack of independence between an angler being sampled into the APAIS and self-reporting. Our research group at SMU has examined the magnitude of these non-sampling errors, and matching errors seem to have the largest effect on the estimators we study. Our work is thus critical for NOAA, as they make decisions about the future status of their estimation procedures for recreational angling. We begin the next section with a review of capture-recapture methodology and show how it is adapted for estimating total catch in this application.

1.3. Theory for Motivating Example

1.3.1. Capture-Recapture Model

Capture-recapture methods are powerful ways to estimate total in specific scenarios. In a classic example, suppose a researcher wishes to know the total number of fish (N) in the local fishing hole. On the initial fishing trip, she catches n_1 fish. These fish are tagged so they can be identified later. The next day she returns to the fishing hole and catches n_2 fish. In this second catch, suppose m fish were also caught on the first day, identifiable by their tag. Assuming the second sample is a random sample of the finite population, the Lincoln-Peterson estimator of total (Cren, 1956) gives:

$$\hat{N} = \frac{n_1 n_2}{m}. \tag{1.1}$$

$\hat{N}$ is the maximum likelihood estimator of N under the hypergeometric model.

In the NMFS experiment the self-reporting sample is analogous to the capture portion of a capture-recapture program, while the dockside intercept sample is the recapture component. However, our problem deviates in two ways from the classic capture-recapture setting. First, our goal is to estimate total catch in a population of unknown size N, rather than N itself. Second, the recapture sample is not a simple random sample, but rather a cluster sample with varying selection probabilities. Thus the estimator of catch is in the spirit of (1.1) but has a different form.

1.3.2. Notation for Experimental Estimation Program

Define the universe of interest to be the N recreational fishing trips in the Gulf of Mexico. Define the catch for some species in the i^{th} trip as y_i $(i = 1, ..., N)$. The objective is to estimate $t_y = \sum_{i=1}^{N} y_i$ when N is unknown. In the self-reported data, the reported catch for the i^{th} trip is denoted y_i^*. If the i^{th} trip is not reported, y_i^* is defined to be 0. y_i may differ from y_i^* due to measurement error in the captain's

report. The intercept assessment of catch is taken as the gold standard.

Denote the probability sample (APAIS) by s_2 and the non-probability sample (electronic reports) by s_1. Again, the APAIS is a sample of public docks only, and thus is not a probability sample of the entire universe. There are n_2 trips sampled in s_2 and n_1 reported trips in s_1. We denote by m the number of units present in both s_1 and s_2. Then the number of units in s_1 only is $n_1 - m$, and the number of units in s_2 only is $n_2 - m$. Figure 1 is adapted from Liu et al. (2017) and visualizes the set-up for the two samples.

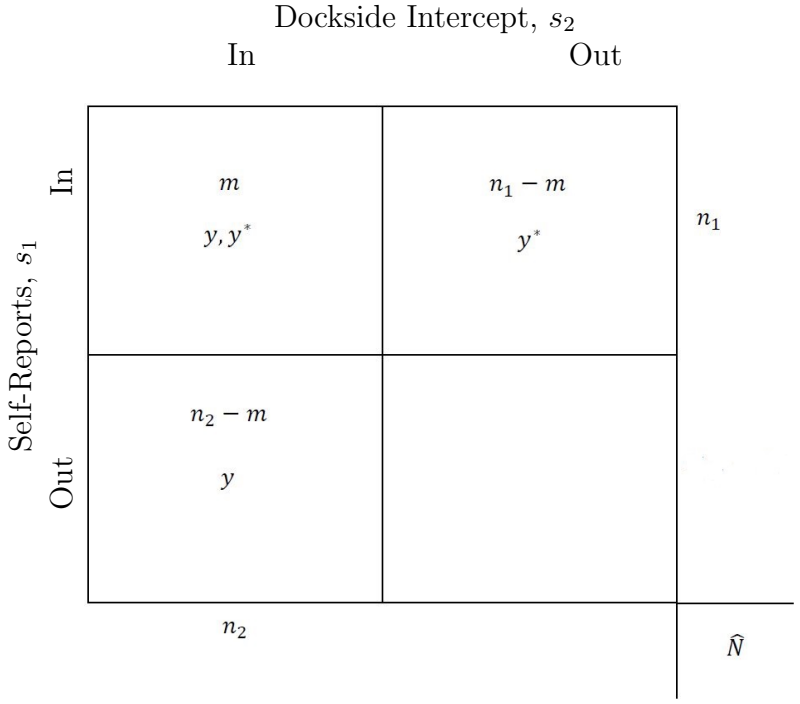


Figure 1.1: Sampling Set-Up

Next, we review current methods for estimating catch from the data of Figure 1.1.

1.3.3. Estimators

Liu et al. (2017) propose several estimators of t_y and examine their performance for various levels of reporting rates and reporting accuracy. They do not investigate

the effect of non-sampling errors, such as matching error. In a later section, we will discuss this effect.

The first estimator is denoted $\hat{t}_{yp}$ and is a generalized version of an estimator developed by Pollock et al. (1994). Their data consists of a probability sample of intercepts and self-reports of trip counts (but not catch). In this method the capture and the recapture samples are used to estimate the total number of trips N, which is multiplied by an estimate of mean catch, determined from the recapture sample only:

$$\hat{t}_{yp} = \hat{N}\hat{\bar{y}}, \tag{1.2}$$

where $\hat{N}$ is defined in (1.1) and $\hat{\bar{y}}$ is the average catch from the intercept sample (Pollock et al., 1994).

Liu et al. (2017) generalizes (1.2) for a complex sample design with sampling weights as

$$\hat{t}_{yp} = \frac{n_1}{\hat{p}_1}\hat{\bar{y}} = n_1\frac{\hat{t}_y}{\hat{n}_1}. \tag{1.3}$$

Here

$$\hat{n}_1 = \sum_{i\in s_2} w_i r_i \tag{1.4}$$

$$\hat{p}_1 = \frac{\sum_{i\in s_2} w_i r_i}{\sum_{i\in s_2} w_i} \tag{1.5}$$

$$\hat{t}_y = \sum_{i\in s_2} w_i y_i. \tag{1.6}$$

where r_i is an indicator of whether a unit is a reporter

$$r_i = \begin{cases} 1 & \text{if } i \in s_1 \\ 0 & \text{otherwise} \end{cases}.$$

$\hat{n}_1, \hat{p}_1$, and $\hat{t}_y$ are Horvitz-Thompson estimators of n_1, $p_1 = \frac{n_1}{N}$ (reporting rate), and t_y, respectively (Horvitz and Thompson, 1952). The $w_i's$ are sampling weights, computed as reciprocals of the selection probabilities. $\hat{t}_{yp}$ is a ratio estimator with ratio $B_p = \frac{t_y}{n_1}$ and auxiliary variable r_i.

The next estimator uses the reported catch, rather than the number of reported trips, as an auxiliary variable:

$$\hat{t}_{yc} = t_{y^*} \frac{\hat{t}_y}{\hat{t}_{y^*}} \tag{1.7}$$

where

$$t_{y^*} = \sum_{i=1}^{n_1} {y_i}^* \tag{1.8}$$

$$\hat{t}_{y^*} = \sum_{i \in s_2} w_i r_i {y_i}^* \tag{1.9}$$

t_{y^*} is the total catch for the reporting domain, and $\hat{t}_{y^*}$ is its estimator. $\hat{t}_{yc}$ is also a ratio estimator with ratio $B_c = \frac{t_y}{t_{y^*}}$ and auxiliary variable $r_i y_i^*$. $\hat{t}_{yc}$ takes the form of a capture-recapture estimator and uses reported catch as auxiliary information.

The final estimator is a weighted combination of $\hat{t}_{yp}$ and $\hat{t}_{yc}$ called $\hat{t}_{MR}$.

$$\hat{t}_{MR} = (1 - W_{SRS})\hat{t}_{yp} + W_{SRS}\hat{t}_{yc} \tag{1.10}$$

$\hat{t}_{MR}$ is a multivariate ratio estimator. The optimal weight W_{SRS} if the recapture sample were a simple random sample (each sampling unit has the same probability of being selected, denoted SRS) would be:

$$W_{SRS} = \frac{t_{y^*}}{t_y} \frac{S_{1,yy^*}}{S^2_{1y^*}} = \frac{t_{y^*}}{t_y} \frac{S_{1y}}{S_{1y^*}} R_{1,yy^*} \tag{1.11}$$

where S_{1,yy^*} is the covariance of y and y^* in s_1, $S^2_{1y^*}$ is the variance of y^* in s_1, S_{1y} is the standard deviation of y in s_1, and R_{1,yy^*} is the correlation of y and y^* in s_1 (Olkin, 1958).

In applications, W_{SRS} will be unknown, but it can be consistently estimated (Olkin, 1958) by

$$\hat{W}_{SRS} = \frac{t_{y*}}{\hat{t}_{yc}}, \tag{1.12}$$

only if $\frac{S_{1y}}{S_{1y^*}} R_{1,yy^*} = 1$, which is true if reporting is accurate ($y_i = y_i^*$). When $\hat{W}_{SRS}$ is substituted for W_{SRS} in (1.6), the resulting estimator simplifies to

$$\hat{t}_{y2} = t_{y*} + \frac{n_1}{\hat{n}_1}(\hat{t}_y - \hat{t}_{y*}). \tag{1.13}$$

$\hat{t}_{y2}$ is similar to a difference estimator and is proposed even for complex sample designs. $\hat{t}_{y2}$ adjusts the total reported catch additively rather than multiplicatively, as $\hat{t}_{yc}$ does.

Liu et al. (2017) studied the variances of these estimators for a variety of scenarios, all under simple random sampling. When reporting rates are low, $\hat{t}_{yp}$ and $\hat{t}_{y2}$ outperform $\hat{t}_{yc}$, sometimes substantially. When reporting is fairly accurate and/or reporting rates are high, both $\hat{t}_{yc}$ and $\hat{t}_{y2}$ are better than $\hat{t}_{yp}$. In general, $\hat{t}_{y2}$ performs best or near best under a wide range of conditions when reported catch is not perfectly correlated with actual catch.

In most MRIP applications of self-reporting procedures to date, such as those in Alabama and Mississippi for estimating Red Snapper catch, $\hat{t}_{yc}$ is used. All three estimators could be affected by the undercoverage of the APAIS, which plays the role of s_2. However, it can be shown for each case that if the reporting rate and $CPUE$ are the same for public and private trips, the estimators are approximately unbiased.

Liu et al. (2017) assumed there were no errors in matching trips between the reports and the intercept sample. In the next chapter we discuss the potential matching errors that may occur and their effects on these estimators.

Chapter 2

Matching Errors

For estimating the total catch of fish using a dockside intercept sample (s_1) and reported trips (s_1), for the three estimators discussed in Chapter 1 ($\hat{t}_{yp}, \hat{t}_{yc}$ and $\hat{t}_{y2}$), one must link trips between the samples. In practice, linking trips between the two samples is difficult. For example, if a captain makes two trips in the same day and is intercepted on just one trip, we cannot confidently match the records because the ending time of a trip is not reliably reported. In this case we are unsure which intercept observation matches which report. Other data irregularities making linkage difficult are caused by device errors and reporting errors.

In implementations to date, estimates made using electronically reported data tend to be smaller (and outside the confidence intervals) than the official estimates produced by MRIP's operational procedures. One possible explanation is that bias from non-sampling errors, such as matching errors, is the cause. In this chapter, we will examine the effect of matching errors on the bias and variance on the estimators from Liu et al. (2017) by developing a statistical model to characterize the randomness due to matching.

We have found few papers that assess the impact of residual matching errors after record linkage on estimators calculated from the linked files. Some have addressed the effect on regression coefficients e.g. (Scheuren and Winkler, 1993; Lahiri and Larsen, 2005). Recently, Di Consiglio and Tuoto (2018) performed a sensitivity analysis of the effects that using different linking variables in record linkage, a matching technique described in Chapter 3, has on the bias of resulting regression coefficients. Another examines the impact of matching error on an estimate of undercount for the U.S. Census (Mulry and Spencer, 1991). This is similar to our application because the

undercount estimation can be thought of as a capture-recapture program.

To understand how linkage errors affect estimators we distinguish among the ways trips can be wrongly linked. We use the language of Bell et al. (1994) in our definitions of types of matching error. They define a *match* to occur when two records (one from each data set) refer to the same unit and a *link* to occur when two records are determined to be referring to the same entity (via some matching procedure).

We define three types of matching error for sampled trips. The first occurs when a trip in s_2 (dockside intercept sample) that actually did not report (not in s_1) is incorrectly linked to a report in s_1 (reported trips). That is, a sampled trip is believed to have been reported but it was not. These are called *false positive* links.

The second error happens when a trip in s_2 whose captain submitted a report (thus also in s_1) is linked with a different reported trip. This could happen if a captain reports two trips in a single day but is only sampled once. The one interviewed trip could be linked to the wrong report. Though not a true match, the unit is accurately deemed to have been reported. However, agreement on catch (y_i vs y_i^*) for such links may be poor since each trip refers to a different outing. We call this a *mismatch* link.

The third error occurs when a trip in s_2 whose captain submitted a report (thus also in s_1) is not linked to any reported trip. This is called a *false negative* link.

Every estimator presented uses information from trips which were both reported and intercepted; i.e that should have been matched. If the estimators do not identify these trips correctly, then the estimators will be inaccurate. Because they do not combine information in the same way, the estimators can suffer differently from different linking errors. We now develop a framework to characterize the effects of such errors on the bias and variance of each estimator introduced in Section 1.3.3.

2.1. A Model for Matching Error

To begin we lay out new notation. Recall r_j is the indicator of reporting status:

$$r_j = \begin{cases} 1 & \text{if the } j^{th} \text{ population unit is a reported trip} \\ 0 & \text{otherwise} \end{cases}$$

where $j = 1, ..., N$. r_j is the true reporting status of a trip, not the perception of an analyst. Next, define:

$$m_i(j) = \begin{cases} 1 & \text{if the } i^{th} \text{ population unit is linked to the } j^{th} \text{ trip report} \\ 0 & \text{otherwise.} \end{cases}$$

$m_i(j)$ is an indicator of whether the analyst links the i^{th} trip to the report of the j^{th} trip, for both i and $j = 1, ..., N$.

To model the effect of matching error on the estimators, regard:

$$\underset{\sim}{m}_i = (m_i(1), m_i(2), ..., m_i(n_1), m_i(n_1 + 1), ..., m_i(N), m_i(N + 1))$$

as a random vector, where $m_i(N + 1) = 1 - \sum_{j=1}^{N} m_i(j)$ is an indicator that the i^{th} sample trip is unlinked. If the i^{th} unit is not sampled, it is always true that $m_i(N+1) = 1$, and if the i^{th} trip is sampled, any of the $N+1$ options defined by the above $\underset{\sim}{m}_i$ vector may be 1. That is, an analyst may only link trips selected into the APAIS.

We assume the distribution of $\underset{\sim}{m}_i$, conditional on the reporting units $\mathbf{r} = (r_1, ..., r_N)$, is a single multinomial trial:

$$\underset{\sim}{m}_i \sim Multinomial(1, \underset{\sim}{\pi}_{i|\mathbf{r}})$$
$$\underset{\sim}{\pi}_{i|\mathbf{r}} = (\pi_i(1), \pi_i(2), ..., \pi_i(N), \pi_i(N + 1)),$$

where all $\underset{\sim}{m}_i's$ are mutually independent. $\pi_i(j)$ is the probability of linking the i^{th} population unit with the j^{th} trip report, given the reporting vector $\mathbf{r}$, while $\pi_i(N+1)$ is the probability of failing to link the i^{th} trip. For ease of notation, we suppress the subscripted indicator of conditioning on $\mathbf{r}$ for the individual probabilities, but their value should be understood to be conditional on the reporting status of all trips. Note that a trip cannot link with a non-reported trip, i.e. if $r_j = 0$ then $\pi_i(j) = m_i(j) = 0$.

Our model does allow the same report to be linked to more than one trip; this is a result of the assumption that the multinomial vectors are independent. In practice, one may cull links to eliminate reports linked multiple times and we acknowledge this model does not correctly reflect the effects of such a step. Culling links can only improve the false positive error rate, although at the risk of increasing the false negative error rate. Thus, we believe this generalized model may be slightly pessimistic. In our own linkage implementation for the NMFS data (Chapter 3) we culled links, which reduced our number of links by 11%.

This matching error model is flexible enough to include both deterministic and random linking algorithms. A completely accurate algorithm would have $\pi_i(i) = r_i$, $\pi_i(N+1) = 1 - r_i$, and $\pi_i(j) = 0$ for $j \neq i$. An example of a random one follows.

Suppose an analyst attempts to link trips using the linkage variables boat ID number, date, and time. Suppose the matching algorithm links sampled units to a report that agrees on all three variables. If no links are available, the requirement loosens and the sampled unit is linked to a report if boat ID and at least one other variable agree. If more than one report qualifies as a link at either stage, then the algorithm chooses the report to link at random. Now, suppose a captain accurately reports the boat ID and date for trip i, but reports the time inaccurately. Then $\pi_i(j) = \frac{I[k_{ID,date}>0]}{k_{ID,date}}$, where $k_{ID,date} = \sum_{ID,date} r_j$ is the number of reports on that date by the captain. A mismatch may occur if $k_{ID,date} > 1$. If the captain does not report trip i but does report another trip (i'), a false positive occurs. If the date of trip i is misreported, a false negative occurs for trip i, and a false positive might occur for

a trip j made that day by the same boat. This example illustrates that the $\pi_i(j)'s$ can reflect properties of both the linkage algorithm and the measurement error of variables used for linking.

In the presence of linking errors, the estimators $\hat{t}_{yp}$, $\hat{t}_{yc}$ and $\hat{t}_{y2}$ are affected only through their components $\hat{n}_1$ and $\hat{t}_{y^*}$. These statistics may differ from the values they would have if there were no errors. The other components of the estimators are unaffected by such errors. Note that $\sum_{j=1}^{N} r_j m_i(j)$ is the indicator that a link is made for the i^{th} population unit. If the statistics may be contaminated with linking error, we express their values from (1.4) and (1.10) as:

$$\hat{n}_1 = \sum_{i=1}^{N} z_i w_i \sum_{j=1}^{N} r_j m_i(j) \tag{2.1}$$

$$\hat{t}_{y^*} = \sum_{i=1}^{N} z_i w_i \sum_{j=1}^{N} r_j y_j^* m_i(j). \tag{2.2}$$

z_i is an indicator of whether the i^{th} unit is in the sample s_2. By noting $1 = r_i + (1 - r_i)$, we rewrite $\hat{n}_1$ as

$$\hat{n}_1 = \sum_{i=1}^{N} r_i z_i w_i \sum_{j=1}^{N} r_j m_i(j) + \sum_{i=1}^{N} (1 - r_i) z_i w_i \sum_{j=1}^{N} r_j m_i(j).$$

Then we can show (see Appendix (A.2)) its expectation is

$$\begin{aligned} \mathrm{E}(\hat{n}_1) &= n_1 - \sum_{i=1}^{N} r_i \pi_i(N+1) + \sum_{i=1}^{N} (1 - r_i) \sum_{j \in s_1} \pi_i(j) \\ &= n_1 [1 - \gamma_{\mathbf{r}} + \frac{1 - p_1}{p_1} \eta_{\mathbf{r}}], \end{aligned} \tag{2.3}$$

where

$$\gamma_{\mathbf{r}} = \frac{1}{n_1} \sum_{i=1}^{N} r_i \pi_i(N+1), \tag{2.4}$$

$$\eta_{\mathbf{r}} = \frac{1}{N - n_1} \sum_{i=1}^{N} (1 - r_i) \sum_{j \in s_1} \pi_i(j). \tag{2.5}$$

The parameter $\gamma_{\mathbf{r}}$ is the expected proportion of false negative links of reported trips. $\gamma_{\mathbf{r}}$ is conditioned on the vector of self-reports ($\mathbf{r}$) so the subscript $\mathbf{r}$ is used. $\eta_{\mathbf{r}}$ is defined in the same way and is the expected proportion of false positive links among non-reported trips, also conditioned on $\mathbf{r}$.

We find the expectation of $\hat{t}_{y^*}$ in a similar fashion. Define parameters for the expected mismatch and true positive match proportions of reported trips as

$$\delta_{\mathbf{r}} = \frac{1}{n_1} \sum_{i=1}^{N} r_i \sum_{\substack{j \in s_1 \\ j \neq i}} \pi_i(j) \tag{2.6}$$

and

$$\omega_{\mathbf{r}} = \frac{1}{n_1} \sum_{i=1}^{N} r_i \pi_i(i), \tag{2.7}$$

respectively. We further define three parameters to describe the expected average reported catch obtained among the mismatched, true positive, and false positive links, denoted by $\bar{y}^*_{\delta_{\mathbf{r}}}$, $\bar{y}^*_{\omega_{\mathbf{r}}}$, and $\bar{y}^*_{\eta_{\mathbf{r}}}$, respectively:

$$\bar{y}^*_{\delta_{\mathbf{r}}} = \frac{\sum_{i=1}^{N} r_i \sum_{\substack{j \in s_1 \\ j \neq i}} y^*_j \pi_i(j)}{\sum_{i=1}^{N} r_i \sum_{\substack{j \in s_1 \\ j \neq i}} \pi_i(j)} \tag{2.8}$$

$$\bar{y}^*_{\omega_{\mathbf{r}}} = \frac{\sum_{i=1}^{N} r_i y^*_i \pi_i(i)}{\sum_{i=1}^{N} r_i \pi_i(i)} \tag{2.9}$$

$$\bar{y}^*_{\eta_{\mathbf{r}}} = \frac{\sum_{i=1}^{N} (1 - r_i) \sum_{j \in s_1} y^*_j \pi_i(j)}{\sum_{i=1}^{N} (1 - r_i) \sum_{j \in s_1} \pi_i(j)}. \tag{2.10}$$

The expectation of $\hat{t}_{y^*}$, conditioned on the vector of self-reports $\mathbf{r}$, is

$$\mathrm{E}(\hat{t}_{y^*}) = n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta\bar{y}^*_{\delta} + \eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}(\frac{1-p_1}{p_1})]. \tag{2.11}$$

See appendix (A.5) for the derivation of (2.11).

Note that some relationships must hold among the parameters described above, because each sampled and reported trip is either a true positive, a mismatch, or a false negative link. The expected averages over these categories is equal to the true average reported catch ($\bar{y}^*_1$). This leads to:

$$1 = \omega_{\mathbf{r}} + \delta_{\mathbf{r}} + \gamma_{\mathbf{r}} \tag{2.12}$$

$$\bar{y}^*_1 = \omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \gamma_{\mathbf{r}}\bar{y}^*_{\gamma_{\mathbf{r}}}, \tag{2.13}$$

where $\gamma_{\mathbf{r}}\bar{y}^*_{\gamma_{\mathbf{r}}} = \sum_{i=1}^{N} r_i y_i^* \pi_i (N+1)$.

2.1.1. Effect of Matching Errors on Approximate Relative Bias

Now that a model has been defined, we can examine the bias induced by matching errors in the estimators. To do this we first calculate the expected values of the three estimators defined in (1.3), (1.8), and (1.14) so we can calculate the relative bias of each. Relative bias (RB) is defined as:

$$\mathrm{RB}(\hat{t}_{y,est}) = \frac{E(\hat{t}_{y,est}) - t_y}{t_y}, \tag{2.14}$$

where, $\hat{t}_{y,est}$ represents one of the estimators of t_y. Because all three estimators have components that are ratio estimators, their expected values must be approximated based on the assumption that the intercept sample is large enough that only the first term of its Taylor expansion is required to produce an adequate approximation to the mean. We denote the approximate relative bias by ARB.

To make the ARB expressions more understandable, we define discrepancy parameters:

$$\lambda_{\omega_{\mathbf{r}}} = \frac{\bar{y}^*_{\omega_{\mathbf{r}}}}{\bar{y}^*_1}, \lambda_{\delta_{\mathbf{r}}} = \frac{\bar{y}^*_{\delta_{\mathbf{r}}}}{\bar{y}^*_1}, \text{and } \lambda_{\eta_{\mathbf{r}}} = \frac{\bar{y}^*_{\eta_{\mathbf{r}}}}{\bar{y}^*_1}.$$

These parameters describe the discrepancy between the expected average reported catch among the different types of links and the true average reported catch. If linked trips of each type have expected average reported catch near the true average, then the $\lambda' s$ will be close to 1. This will always occur if reported catch varies little from trip to trip, as seen for some species with small bag limits, such as Red Snapper. Using (2.5) and (2.11), one can show that the approximate relative bias of the three estimators of total catch can be written as:

$$\text{ARB}(\hat{t}_{yp}) = \frac{p_1\gamma_{\mathbf{r}} - (1-p_1)\eta_{\mathbf{r}}}{p_1(1-\gamma_{\mathbf{r}}) + (1-p_1)\eta_{\mathbf{r}}} \tag{2.15}$$

$$\text{ARB}(\hat{t}_{yc}) = \frac{p_1(1-\omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} - \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}}) - (1-p_1)\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}}{p_1(\omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}}) + (1-p_1)\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}} \tag{2.16}$$

$$\text{ARB}(\hat{t}_{y2}) = \text{ARB}(\hat{t}_{yp}) + p_1\frac{\bar{y}^*_1}{\bar{y}}[1 - \frac{p_1(\omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}}) + (1-p_1)\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}}{p_1(\omega_{\mathbf{r}} + \delta_{\mathbf{r}}) + (1-p_1)\eta_{\mathbf{r}}}]. \tag{2.17}$$

See Appendix A for the details of the derivation.

To extract insight from these ARB expressions, we first consider a case where the discrepancy parameters ($\lambda' s$) are all equal to 1, which as noted above, may be reasonable for Red Snapper. If mismatch links are the only observed matching error ($\delta_{\mathbf{r}} > 0, \gamma_{\mathbf{r}} = \eta_{\mathbf{r}} = 0$), then all three estimators are asymptotically unbiased. Thus mismatch links by themselves are not of concern in this setting. Next, if false negative links are the only error type observed ($\gamma_{\mathbf{r}} > 0, \eta_{\mathbf{r}} = \delta_{\mathbf{r}} = 0$) there will be too many links, and the ARB of all the estimators is $\frac{\gamma_{\mathbf{r}}}{1-\gamma_{\mathbf{r}}}$. That is, the estimators are biased upward, without a bound. Conversely, if false positive links are the only error type observed ($\eta_{\mathbf{r}} > 0, \gamma_{\mathbf{r}} = 0$), the ARB of all the estimators is $\frac{-\eta_{\mathbf{r}}}{\frac{p_1}{1-p_1}+\eta_{\mathbf{r}}}$, so the bias is always negative and is a function of the reporting rate p_1. A small reporting rate can produce a large relative bias. A high reporting rate gives a low relative bias because

if few trips are unreported then few trips can be false positive links.

If both false positive and false negative errors occur, the number of links may be closer to the true number of matches than if only one error type occurs because they can offset each other. In this case, the biases above may be mitigated. Figure 2.1 displays the ARB for some such scenarios. The figure displays different settings of false negative and false positive error parameters. The columns pertain to three $\gamma_{\mathbf{r}}$ and $\eta_{\mathbf{r}}$ values, and the rows represent three catch discrepancy values (0.75, 1, 1.33) denoted $\lambda = \lambda_{\omega_{\mathbf{r}}} = \lambda_{\eta_{\mathbf{r}}}$. λ conforms to the constraints seen in the identities of the matching parameters (2.12) and (2.13). In each column $\eta_{\mathbf{r}}$ and $\gamma_{\mathbf{r}}$ average to 0.25, so Figure 2.1 examines how balancing the $\eta_{\mathbf{r}}$ and $\gamma_{\mathbf{r}}$ differently affects ARB.

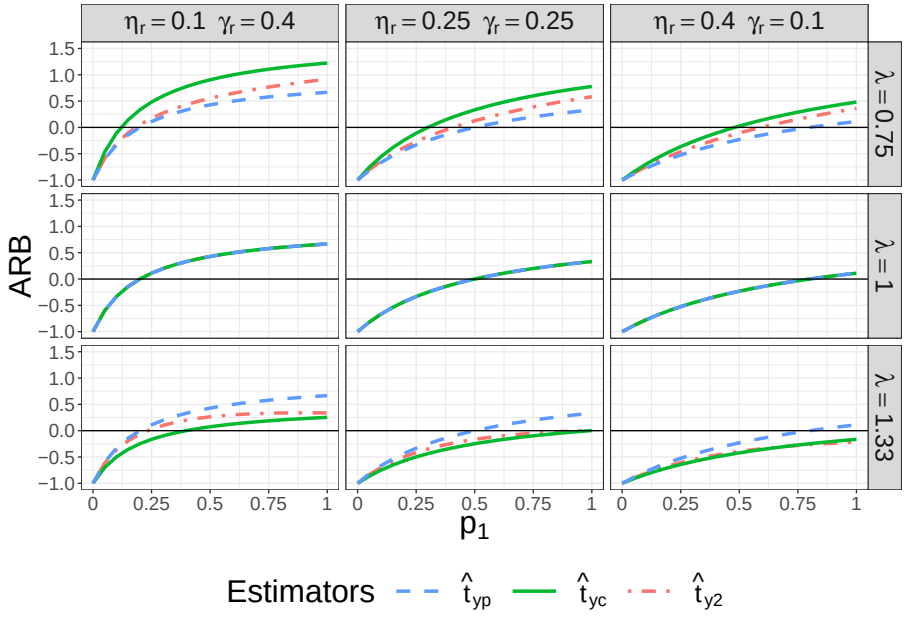


Figure 2.1: ARB of $\hat{t}_{yp}$, $\hat{t}_{yc}$, and $\hat{t}_{y2}$ due to Matching Error as Functions of Reporting Rate for 3 Linkage Scenarios and 3 Reported Catch Discrepancy Values

Figure 2.1 shows that the ARB is an increasing function of the reporting rate p_1 for all three estimators, beginning with a substantially negative bias when p_1 is small. For the chosen settings, each estimator is biased downward by more than 50% when the reporting rate is smaller than 10%. The range of possible ARB values

is largest in the left column ($\eta_{\mathbf{r}} = 0.4, \gamma_{\mathbf{r}} = 0.1$) and smallest in the right column ($\eta_{\mathbf{r}} = 0.1, \gamma_{\mathbf{r}} = 0.4$). The ordering of the estimators by size of bias reverses when λ changes from 0.75 to 1.33, and when $\lambda = 1$, no estimator has an advantage in asymptotic bias.

Another conclusion from Figure 2.1 is that the estimator with the smallest bias depends on the reporting rate. For example, when $\lambda = 0.75$ and $\eta_{\mathbf{r}} = 0.1, \gamma_{\mathbf{r}} = 0.4$, $\hat{t}_{yp}$ has an ARB closest to 0 for reporting rates greater than about 20%, but when $\lambda = 0.75$ and $\eta_{\mathbf{r}} = 0.1, \gamma_{\mathbf{r}} = 0.4$, $\hat{t}_{yp}$ is optimal only for reporting rates over about 70%. We also conclude that $\hat{t}_{y2}$ rarely has the smallest bias, but hardly ever has the worst relative bias either.

2.1.2. NMFS Pilot Study Example

For insight into the magnitude of the matching error parameters, we turn to our work on the NMFS electronic reporting pilot study. Our research group was responsible for producing estimates of catch by recreational anglers fishing on charter boats in the Gulf of Mexico over the period of the study. As part of the estimation process, we had to develop a procedure for linking sampled trips with reports. Before developing the record linkage method discussed in Chapter 3, a variety of other linking methods were tried. To gain understanding of the magnitude of the matching error parameters and their effect on bias, we compared the results of two linkage procedures that were implemented independently by two different research team members. In the first, which we called direct matching, trips were linked by comparing the boat ID number, the return date, and return time of the trips. A link was defined if the boat ID was identical and the other two were close, using human judgment. Use of human judgment is untenable if there are a large number of records. The second approach was record linkage, a probabilistic method to link records between data sources with no unique identifier, (Fellegi and Sunter, 1969). Record linkage attempts to replace human judgment with an algorithmic decision. Record linkage

and our implementation of it are discussed in Chapter 3.

Though both methods are sensible, they produced different sets of links. Table 2.1 summarizes the disagreement of the methods for the links from trips sampled in the APAIS in 2017. To examine the potential bias from matching error in our application we used the data from Table 2.1 to estimate the matching error parameters $\omega_{\mathbf{r}}, \delta_{\mathbf{r}}$, and $\eta_{\mathbf{r}}$ twice, by assuming, in turn, that each method is completely accurate. Specifically, we used the off-diagonal cells and the margin totals of the table to estimate the proportion of false negative and false positive links, which we used as estimates of $\gamma_{\mathbf{r}}$ and $\eta_{\mathbf{r}}$, respectively. We assumed a mismatch error parameter of 0 (information about that parameter is not available from the table) since such links have no effect on bias when discrepancy parameters are 1, which we assume.

		Direct Matching		
		Matched	Not Matched	Total
Record Linkage	Linked	62	29	91
	Not Linked	48	1345	1393
	Total	110	1374	1484

Table 2.1: Comparison of Matching Methods in NMFS Electronic Reporting Study, 2017

Table 2.2 shows the estimates of the matching error parameters that would be attained by assuming one of the two linking methods is accurate, along with the estimate of reporting rate each scenario would imply. Estimates of these parameters were obtained as weighted estimates of the proportion of each category made from the intercept sample cases, using one of the set of links as "truth". The weights were sample weights from the intercept sample. We used the parameter values from Table 2.2 (along with an assumption that $\delta_{\mathbf{r}} = 0$ and $\lambda_{\omega_{\mathbf{r}}} = \lambda_{\eta_{\mathbf{r}}} = 1$) to examine the ARB from estimators of Red Snapper catch. From (2.15) - (2.17), we computed the ARB for each set of parameters. Assuming record linkage produces perfect links $\text{ARB}(\hat{t}_{yp}) = \text{ARB}(\hat{t}_{yc}) = \text{ARB}(\hat{t}_{y2}) = -0.14$, while if direct matching produces perfect

Algorithm Assumed Accurate	Parameter	Value
Record Linkage	$\omega_{\mathbf{r}}$	0.59
	$\gamma_{\mathbf{r}}$	0.41
	$\eta_{\mathbf{r}}$	0.03
	p_1	0.05
Direct Matching	$\omega_{\mathbf{r}}$	0.53
	$\gamma_{\mathbf{r}}$	0.47
	$\eta_{\mathbf{r}}$	0.02
	p_1	0.06

Table 2.2: Estimated Parameters for 2 Linking Methods

links, $\text{ARB}(\hat{t}_{yp}) = \text{ARB}(\hat{t}_{yc}) = ARB(\hat{t}_{y2}) = 0.19$. Figure 2.2 shows the ARB of the estimators if the false positive and negative error parameters took the values shown in Table 2.2 when direct matching is considered perfect. Figure 2.2 displays the ARB's as functions of the reporting rate and we assume reporting is representative ($\bar{y}_1^* = \bar{y}$).

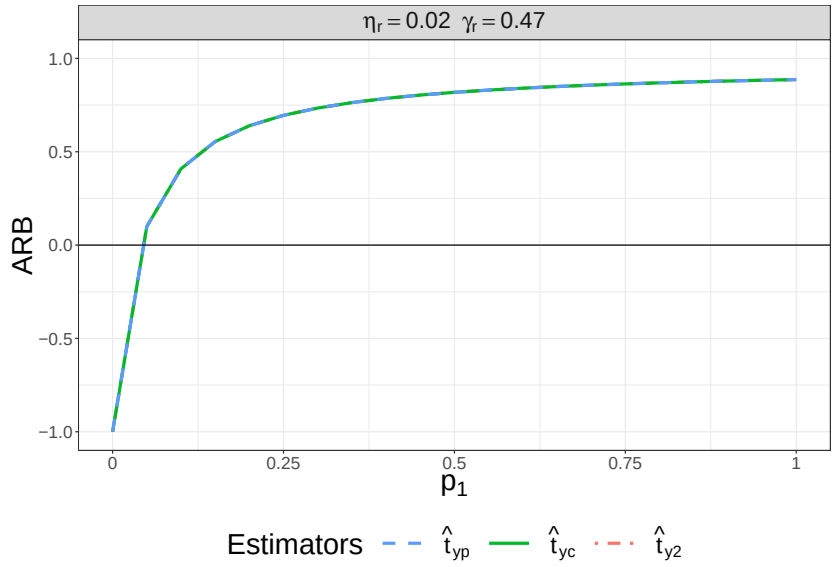


Figure 2.2: ARB of $\hat{t}_{yp}$, $\hat{t}_{yc}$, and $\hat{t}_{y2}$ due to Matching Error as Function of Reporting Rate, 2017 NMFS Electronic Reporting Study; All Lines Overlap

In Figure 2.2, the slope of ARB is steep and crosses from negative to positive for small reporting rates, as seen in the NMFS study. This is why, the calculated ARB assuming record linkage is completely accurate is negative and the calculated ARB assuming direct matching is completely accurate is positive.

2.1.3. Effect of Matching Errors on Approximate Relative Mean Square Error

In Section 2.1.1, we expressed the approximate relative bias of the three estimators as a function of the parameters of the matching error model. This allows an assessment of which estimators' bias may be least affected by various types of matching error and how large the bias may be. However, the three estimators have different variances without matching error, and matching error also affects their variances. So the estimator with the smallest bias in the presence of matching error is not necessarily the best with respect to its mean squared error.

Our goal for this analysis is twofold. First, we would like to determine if the contribution from bias due to matching error to the mean squared error of the estimators is likely to be disqualifying. Second, if it is not, we would like to make a recommendation to NOAA about the best estimation procedures for their fisheries that use electronic reporting data collection, if we assume some matching error is inevitable. Since the relative importance of bias depends on the sample size for all estimators, we compare estimators for sample sizes producing precision in the range achieved by current MRIP estimators.

To do this, we first use the model for matching error to derive the impact on variance of each estimator. For this derivation, we assume a simple random sample design for the intercept sample, with sample size selected to achieve the typical range of precision for MRIP estimators at the wave and state level. Then we make comparisons of the estimator mean squared errors under a variety of scenarios, in order to see if we can gain some insight into which estimator is best to use in which circumstances. Since all three estimators are variations of ratio estimators, their sampling

variance can be approximated in the usual way using the delta method (e.g. Cochran (1963) Section 6.3). The contribution of the matching error variance is obtained by a conditioning argument. See Appendix B for details.

The resulting variance expressions of the three estimators are shown in Appendix (B.8), (B.9), and (B.10). Inspection of these expressions yields little insight because of their complexity. However, in the case of no matching errors, they reduce to the variance formulas presented in Liu et al. (2017), shown in (B.11), (B.12), and (B.13). Besides sample size, the precision of the estimators, when no matching errors are present, in (B.11), (B.12), and (B.13), also depend on the reporting rate p_1, the representativeness of reporting, characterized by the ratio of the average reported catch to the average observed catch ($\bar{y}_1^*/\bar{y}$), and the accuracy of reporting, i.e. the correlation between y and y^* among the reported and sampled trips (R_{1,yy^*}).

When matching errors are possible, the variances also depend on components characterized by matching errors, seen in (B.11)- (B.13). To facilitate further knowledge about the behavior of the variances of the estimators, and mean squared errors, we examine a simplifying scenario, generalizable to a multitude of linkage procedures.

Consider a linking algorithm that links each sampled, reported trip to the correct trip report with probability w, to any other trip report with equal probability, and fails to link to any trip report with probability g. This would result in the following linking model parameters for every reported trip i:

$$\begin{aligned} \pi_i(i) &= w \\ \pi_i(j) &= \frac{1-w-g}{n_1-1} \\ \pi_i(N+1) &= g. \end{aligned}$$

For the unreported trips ($i \notin s_1$), we assume the trip is (correctly) not linked to a trip report with probability m, but is linked to any report with equal probability:

$$\begin{aligned} \pi_i(i) &= 0 \\ \pi_i(j) &= \frac{1-h}{n_1} \\ \pi_i(N+1) &= h. \end{aligned}$$

Under this simplified model:

$$\omega_{\mathbf{r}} = w, \gamma_{\mathbf{r}} = g, \delta_{\mathbf{r}} = 1 - w - g, \eta_{\mathbf{r}} = 1 - h.$$

By varying w, g, and h, we can examine different quality linking algorithms, with the goal of examining the relative mean squared error of $\hat{t}_{yp}$, $\hat{t}_{yc}$, and $\hat{t}_{y2}$, defined as

$$\text{RMSE} = \frac{\text{Bias}^2(\hat{t}_{y,est}) + \text{Var}(\hat{t}_{y,est})}{t_y^2} = \text{RB}^2 + \frac{\text{Var}(\hat{t}_{y,est})}{t_y^2}. \tag{2.18}$$

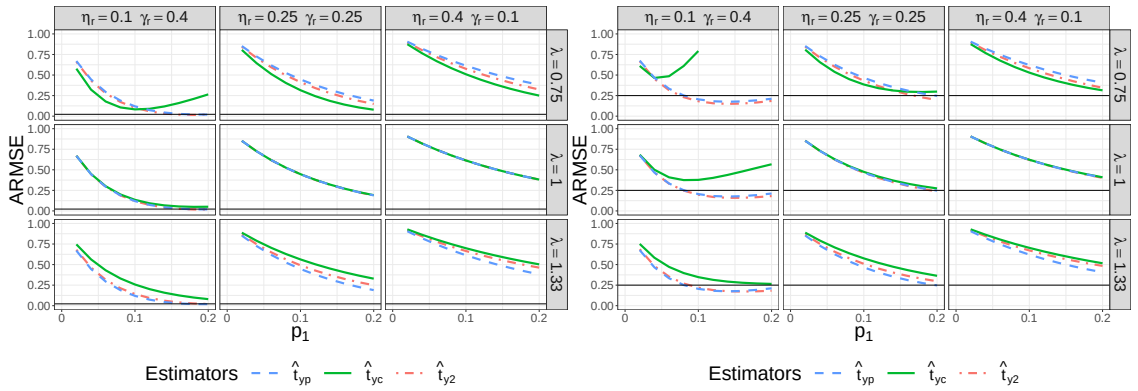
Because we approximate the bias and variance, we actually examine the approximate relative mean square error (ARMSE).

We assume a SRS design for s_2 and examine ARMSE as a function of the reporting rate for several reported catch discrepancy values and linkage error parameters. We set the sample size of the intercept sample, n_2, based on NOAA's standards, because sample size effects the variance of the estimators. NOAA publishes the proportional standard error (PSE) of its estimates as part of its operational estimation program to assist in determining their quality. PSE is the standard error of an estimate divided by the estimate ($^{\text{SE}(\hat{t}_{y,est})}/_{\hat{t}_{y,est}}$). NOAA highlights any estimate with a PSE greater than 0.5, and warns the user of the instability of the estimate. When there are no matching errors, representative reporting ($\bar{y}_1^* = \bar{y}$), and perfect correlation between y^* and y, the relative mean square error is equal to the square of the PSE.

When NOAA makes estimates for a two month period (wave) and/or at the state level rather than for the entire Gulf of Mexico, the estimates have PSE values as large as 0.5. Estimates made for an entire year and/or the entire Gulf of Mexico have PSE values closer to 0.15. Thus, we look at two PSE values, 0.15 and 0.5, for two reporting rates, 0.1 and 0.8. In our NMFS experiment we see a reporting rate of 0.1 and then we examine a reporting rate of 0.8, reasonable for a sampling program where reporting is mandatory, i.e. Mississippi. For each reporting rate we obtain a sample size by setting the RMSE of $\hat{t}_{y2}$ equal to the square of the PSE, assuming no errors in matching, perfect correlation between y^* and y, and representative reporting. We use these scenarios in the figures below.

In Figure 2.3 and 2.4, we assume the CV's (coefficient of variation) for the reported and observed catch are equal, the reported and observed catch are perfectly correlated, and reporting is representative. In Figure 2.3, we choose $w, g,$ and h values to mirror the settings from Figure 2.1, that is, we balance the false negative and false positive error parameters differently in each column of the figure. The rows pertain to three discrepancy values (0.75, 1, and 1.33), where $\lambda = \lambda_{\omega_{\mathbf{r}}} = \lambda_{\eta_{\mathbf{r}}}$. Figure 2.3 shows the approximate relative mean square error for the three estimators, where the sample size has been set to achieve two PSE values (0.15 and 0.5) as described above. We show the ARMSE as a function of the reporting rate for reporting rates near 10%.

From Figure 2.3, the ordering of the estimators' mean squared errors is the same as the ordering from Figure 2.1 (relative bias). Thus, bias due to matching errors dominates the MSE. As in Figure 2.1, here we see the estimator with the smallest ARMSE depends on the discrepancy parameters, linkage parameters, and reporting rate. Additionally, the mean squared error can greatly suffer due to matching errors, especially for low reporting rates. We see that bias is a substantial part of the MSE for smaller PSE values (Figure 2.3a) and less so for larger PSE values (Figure 2.3b). Although $\hat{t}_{y2}$ is never the best estimator, in terms of ARMSE, it is never ever the worst and is always nearly best. We thus recommend $\hat{t}_{y2}$ when one is unsure of



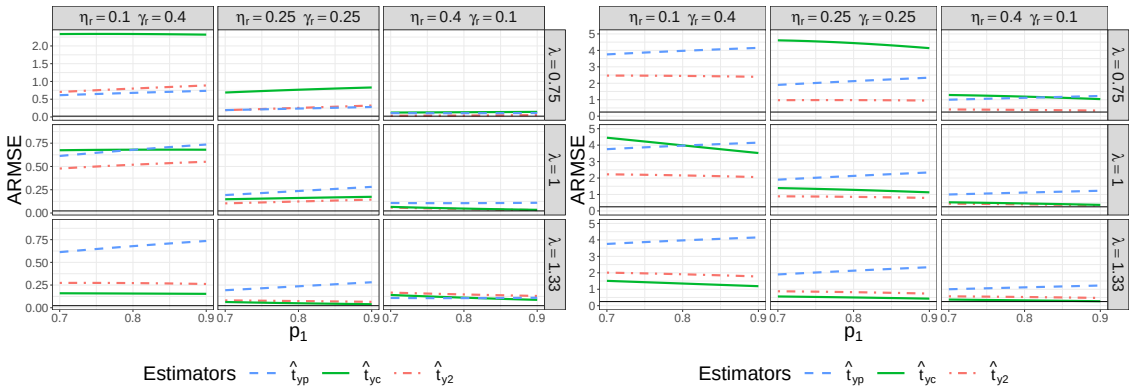
(a) Specified PSE = 0.15; Line at 0.15^2 (b) Specified PSE = 0.50; Line at 0.5^2

Figure 2.3: ARMSE of $\hat{t}_{yp}$, $\hat{t}_{yc}$, and $\hat{t}_{y2}$ as Functions of the Reporting Rate for 3 Linkage Scenarios and 3 Reported Catch Discrepancy Values; Sample Size of s_2 set to Achieve Specified PSE in Scenario of No Matching Error with $p_1 = 0.1$

the parameter values. We also point out that in some cases, $\hat{t}_{yc}$ performs very poorly compared to the other estimators, and we advise NOAA to stop using it in estimation.

In the NMFS electronic reporting pilot study, regardless of the linkage algorithm, we had a very small false positive error rate and it was much smaller than the false negative error parameter (Table 2.2). In the pilot study we were in the setting more similar to the left column of Figures 2.3. For small reporting rates, like those seen in the pilot study, the ARMSE is quite variable.

Figure 2.4 has the same assumptions as those for Figure 2.3, but now, we look at a reporting rate of 80%, and a range of reporting rates near 80%. A high reporting rate such as 80% is achievable for programs with mandatory reporting, like Alabama's. Again, we choose the sample size for two PSE values.



(a) Specified PSE = 0.15; Line at 0.15^2 (b) Specified PSE = 0.50; Line at 0.5^2

Figure 2.4: ARMSE of $\hat{t}_{yp}$, $\hat{t}_{yc}$, and $\hat{t}_{y2}$ as Functions of the Reporting Rate for 3 Linkage Scenarios and 3 Reported Catch Discrepancy Values; Sample Size of s_2 set to Achieve Specified PSE in Scenario of No Matching Error with $p_1 = 0.8$

In Figure 2.4, the patterns from Figure 2.3 and Figure 2.1 no longer hold. For discrepancy parameter (λ) values of 0.75 and 1, $\hat{t}_{y2}$ has the best mean square error. In Figure 2.3, $\hat{t}_{yc}$ was rarely optimal, but here in Figure 2.4, there are times when it is optimal, but other times when it is much worse, in terms of ARMSE, than the other two estimators. Thus again, we recommend $\hat{t}_{y2}$. We also see that in this case, the matching errors contribute substantially to the MSE of the estimators for both large and small PSE values.

Observation of Figures 2.3 and 2.4 shows that choosing the optimal estimator in the presence of matching errors is difficult. In the case of no matching error, Liu et al. (2017) showed $\hat{t}_{y2}$ to have the smallest variance. If matching errors are possible, we still recommend $\hat{t}_{y2}$ when the matching error parameter values are unknown. We also caution NOAA against the use of $\hat{t}_{yc}$ unless the parameters for choosing it can be met.

Finally, note that for Figures 2.3 and 2.4, we assumed representative reporting ($\bar{y}_1^* = \bar{y}$) and accurate reporting ($y_i = y_i^*$). We have looked at cases where each assumption is not held, and note the conclusions from this section remain unchanged.

Chapter 3

Record Linkage

Chapter 2 showed the accuracy of the linking procedure can have a large effect on the quality of the estimators. Therefore, it is important to ensure links are accurately made. For the electronic reporting pilot study, we needed to link trips to perform estimation (see Chapter 1). Originally, we believed the boat ID number, date of the trip, time of the trip, and the location of its return would provide a unique pair of records resulting in a perfect match, since these variables are recorded in both s_1 and s_2. However, this proved far from true.

There are many reasons why linking is difficult. First, in the reports, some information is reported by the captain (e.g., number of passengers) and some is obtained from the electronic device (e.g., location). Both are subject to error, but from different sources. For example, the captains may make errors in recording the number of passengers, while some electronic devices had clocks that were set to incorrect time zones in our pilot study. Furthermore, the time of the dockside intercept will not be identical to the reported return time of a trip, because the interview takes place after the passengers disembark. Additionally, location information from the reports is a series of GPS coordinates recorded at 15 minute intervals. From the intercept survey, we only have the name and identification number of a sampling site. We obtain the coordinates of these sites, but the coordinates for a sampling location are rarely equivalent to the GPS coordinates from a report.

Initially, the linking operation was carried out by hand with a rule filtering trips which were "close" on boat ID number, date and time, and trip ending location. However, few trips were linked. Given the large volume of reports and our knowledge of how many reporting devices were deployed, it seemed unlikely the number

of matches was as small as we found. We decided to loosen the criteria required to identify a match and rely on some of the other variables recorded in both files. To carry out such a method, we needed a principled way to move forward. This led us to the record linkage literature.

3.1. Literature Review

Record linkage is a process to merge two or more data files, using variables present in both data sources. When there is not a unique identifier common to the files, record linkage is valuable. The term *linking variable* is used to identify variables from the separate files which are compared to determine if records match. Record linkage methods were developed after computers became available since it was then feasible to examine every possible pairing between the file records to determine the best one.

Most record linkage development has been for applications when records refer to people. For example, Newcombe et al. (1959) describe an application whose purpose was to follow individuals over time and observe if their health and fertility were affected by exposure to low levels of radiation. Since exposure, marriage, births, and illness information were contained in different files, there was a need to link them with variables common to all the files, such as names and dates. They describe two problems encountered when using linking variables. First, when records actually do refer to the same entity, there may be errors in one or the other linking variable which prevent them from being equivalent. Second, the linking variable values may be equivalent but the records might refer to different entities. They propose a method of evaluating potential matches by creating a score based on aggregating estimates of the odds that each linking variable agrees or not for matching and non-matching pairs (Newcombe et al., 1959).

In NMFS pilot study, we are not linking records referring to people. We are trying to link fishing trips. In our case, we cannot perform accurate clerical review of linked records, which is usually the case when linking files represent people. In

the estimation and calculations for our record linkage implementation, we needed a different way, other than clerical review, to find sets of matches and non-matches in order to estimate parameters of the linking model described below. Bell et al. (1994) presented such a method. Thus, we followed the record linkage implementation laid out in their paper, which was based on the seminal record linkage model developed by Fellegi and Sunter (1969). We first review the Fellegi-Sunter model and then move on to the details of Bell et al. (1994).

Fellegi and Sunter extended and mathematically formalized the ideas of Newcombe et al. (1959) and is the classic method for record linkage. Their algorithm can be described as follows. The two files to be linked are denoted A and B. The set of all potential links between the two files is $A \times B = \{(a, b) : a \in A, b \in B\}$. $A \times B$ is the union of the sets of matched (M) and unmatched (U) pairs,

$$M = \{(a, b) : a = b, a \in A, b \in B\} \text{ and } U = \{(a, b) : a \neq b, a \in A, b \in B\}.$$

Their goal was to produce a rule which declares each member of $A \times B$ as belonging to three possible categories: a match (A_1), a possible match (A_2), or a non-match (A_3).

The linking rule compares the agreement between a and b for a set of linking variables. The result of this comparison is reported as a score. The linking rule is defined by assigning cut-points to the score. Links with high scores are assigned to A_1 while links with low scores are placed in A_3. $L(\mu, \lambda)$ denotes the linking rule achieving specified false positive and false negative rates, denoted as $\mu = P(A_1|U)$ and $\lambda = P(A_3|M)$, respectively. The optimal (μ, λ) rule minimizes the probability that a pair of records cannot be definitively asserted as a match or non-match (i.e. is in A_2) for particular μ and λ values. They prove the optimal linking rule is the one for which the score (S) is defined as:

$$S = \frac{P(\text{agreement status of linking variables of } a \text{ and } b \text{ at observed values} | (a,b) \in M)}{P(\text{agreement status of linking variables of } a \text{ and } b \text{ at observed values} | (a,b) \in U)}. \tag{3.1}$$

Fellegi and Sunter give suggestions for making linkage of large files more practical and also for simplifying the estimation of (3.1). To reduce the number of pairs (a, b) to search through they suggest using blocking variables, variables requiring agreement by both records in order for a comparison to take place. They also suggest selecting linking variables plausibly thought of as independent so (3.1) can be written as the product of ratios of conditional probabilities, for one variable at a time. When one variable is considered at a time, the number of agreement and disagreement patterns to be considered is much smaller. In practice, the log of the product of ratios is taken as the score, so the score is:

$$S = \sum_{k=1}^{K} log \frac{P(\text{ agreement status of a and b for variable } k \text{ at observed value} | (a,b) \in M)}{P(\text{ agreement status of a and b for variable } k \text{ at observed value} | (a,b) \in U)}. \tag{3.2}$$

When there are k linking variables, S is the sum of logged ratios of the k probabilities.

Literature on record linkage has continued to expand beyond the work of Fellegi and Sunter (1969), consisting of new estimation methods for components of the score (e.g., Copas and Hilton (1990), Tancredi and Liseo (2011), Hall and Fienberg (2012), Harron et al. (2014)) and comparisons of various linkage techniques (e.g., Belin (1993)). For our record linkage implementation, we follow the example described in Bell et al. (1994) and adopt their method of estimating the components of (3.2).

The goal of Bell et al. (1994) was to identify matches between two large files, one of birth certificates and the other of Medicaid claims. The records were associated with patients and lacked a unique identifier. Demographic information present in both files were used as linking variables. The authors created a score which added or subtracted weight to a matching metric based on the values of the linking variables. Higher scores indicated a higher likelihood that the records constituted a match. Their methods form the basis for the record linkage algorithm created for our NMFS study and are summarized below.

Start by examining one term in (3.2) which applies to the k^{th} linking variable, whose score is denoted by S_k. We denote by x and y the two values observed for that variable for an (a, b) pair. Following Bell et al. (1994), we express S_k as:

$$S_k = log(\frac{P(x, y|M_{(a,b)} = 1)}{P(x, y|M_{(a,b)} = 0)}) = log(\frac{P(x)P(y|x, M_{(a,b)} = 1)}{P(x)P(y)}). \tag{3.3}$$

where $M_{(a,b)}$ is an indicator of a match. The denominator of the last term is justified by assuming non-matching records have characteristics that they would have if paired at random. Then S_k simplifies to

$$S_k = log(P(y|x, M_{(a,b)} = 1)) - log(P(y)). \tag{3.4}$$

S_k has a unique form for each of three potential situations: x and y agree, have similar values, or disagree.

First, consider the case when the values of x and y agree. By assuming the probability that x and y agree for matching records is nearly 1 (i.e. $P(y|x, M_{(a,b)} = 1) = 1$), the score reduces to:

$$S_k = -log(P(y)). \tag{3.5}$$

Next, if x and y do not agree, but are close, we can rewrite

$$P(y|x, M_{(a,b)} = 1) =$$
$$P(x \text{ and } y \text{ are close}|x, M_{(a,b)} = 1)P(y|x \text{ and } y \text{ are close}, x, M_{(a,b)} = 1),$$
$$P(y|x, M_{(a,b)} = 0) =$$
$$P(x \text{ and } y \text{ are close}|x, M_{(a,b)} = 0)P(y|x \text{ and } y \text{ are close}, x, M_{(a,b)} = 0).$$

By assuming the second term in the above two expressions are equivalent and $P(x \text{ and } y \text{ are close}|x, M_{(a,b)} = 1) = P(x \text{ and } y \text{ are close}|M_{(a,b)} = 1)$, the score is:

$$\begin{aligned} S_k = & log(P(x \text{ and } y \text{ are close}|M_{(a,b)} = 1)) - \\ & log(P(\text{y value for a random record is close to } x|M_{(a,b)} = 0)). \end{aligned} \tag{3.6}$$

In our application, we sometimes used "far", and "farther", as additional categories for a potential link. That is, we replaced "close" in (3.6) with "far" or "farther" to obtain a finer categorization of agreement for some linking variable.

Finally, for the cases where x and y disagree for a potential link, we assume the error occurs because, in some subset of the matching records, y behaves as though it is selected at random, independent of x, then
$P(y|x, M_{(a,b)} = 1) = P(y \text{ picked at random } |M_{(a,b)} = 1)P(y)$. Then (3.4) becomes

$$\begin{aligned} S_k &= log(P(y \text{ picked at random } |M_{(a,b)} = 1)) \\ &\approx log(P(x \text{ and } y \text{ disagree } |M_{(a,b)} = 1)). \end{aligned} \tag{3.7}$$

The second line of (3.7) comes from assuming x and y will agree or be close at random, with low probability. So (3.7) is the linking score when x and y disagree.

To estimate the contribution of each variable to S_k, one must know the true match status for a sample of pairs in $A \times B$. Then the distribution of the agreement patterns can be estimated for matches and non-matches. Clerical review is one way to

determine the match status of random samples from $A \times B$. However in cases where clerical review is not possible or too time-consuming, another method is needed. Bell et al. adopted such an approach. For estimating the contribution to S_k for matches, they assumed that links for which the linking variables all agreed were true matches. Then by removing the agreement requirement for one linking variable, they estimated its distribution among this set of near matches. For estimating the contribution to S_k for non-matches, they assumed the agreement status of the entire population of links was the same as that for the non-matching links, since the true matches make up a negligible fraction of all links.

In our NMFS experiment s_1 and s_2 are the files requiring linkage. Because captains record similar variables to those gathered in the intercept sample, we believed record linkage could provide increased linking accuracy. Next, our record linkage algorithm for the NMFS data is described and it closely follows the work of Bell et al. (1994).

3.2. Record Linkage Implementation for NMFS Pilot Study

We are aware of only one paper on the use of linkage methods for matching angler trips. In this paper, the authors matched angler trips from two files using an ad hoc agreement score (Breidt et al., 2018). For each sampled trip, a metric was computed for its pairing with every reported trip, based on a deterministic algorithm, not dependent on the sample. The score, as a measure of evidence, provided by a same or different variable values was not based on the frequency distribution of the values, as it is for record linkage. The scores of the 5 trips with the best scores were normalized to sum to 1 and the score of all other trips were set to 0. Then the normalized scores were treated as a vector of probabilities that the sampled trip matched each reported trips.

3.2.1. NMFS Pilot Study Data

We have three sources of data from the NMFS electronic reporting experiment. The first two are the interview data from the intercept sample (APAIS, s_1) and the electronically reported data from the self-reports (CLS, s_2). In 2016, 1628 trips were sampled in the dockside intercept and 5,976 trips were reported. In 2017, 1484 trips were sampled in the dockside intercept and 6,277 trips were reported. The data quality, especially for the self-reports, was poorer in 2016 than 2017 as the experiment had just begun and flaws in equipment and operations still needed to be resolved. The intercept file was a clean file, as it was prepared and delivered by NOAA in their normal data production cycle. The variables available from the two files are nearly identical, but the method of collecting them differed. Table 3.1 displays the relevant variables available on the two files.

Intercept File (Recorded by Interviewer)	Report File (Reported by Captain or Observed from Device Signal)
Date of Interview	Date of Trip Return (Device)
Time of Interview	Time of Return (Device)
Identification Number of Interview Site	Latitude and Longitude of Last Signal (Device)
Target Species (Angler Response)	Target Species (Captain)
Number of Fish Kept per Species per Angler	Count of Fish Kept per Species for Entire Boat (Captain)
Number of Fish Discarded per Species per Angler	Count of Fish Discarded per Species for Entire Boat (Captain)
Number of Different Species Caught	Number of Different Species Reported (Captain)
Number of Anglers	Number of Anglers (Captain)
Return State	Return State
MRIP Vessel ID Number	CLS Vessel ID Number

Table 3.1: Relevant Variables Available on Intercept and Report Files

Our third data source was location data. Besides the data from the interviews and reports, the GPS location of the device when the report was filed was available,

as well as continuous reports of device location, made at regular intervals (every 15 minutes), around the clock, whether at sea or in port. If the reporting had been done at the landing site, then the location of the device at that time could have been used as a linking variable. However, the timing of reporting varied, so this was not feasible. Instead, the landing site had to be deduced from the trajectory of the location reports, which required much data preprocessing. An algorithm to identify a "trip" from the string of GPS locations, and the locations of its termination point, was developed by members of the research team. These locations were used to determine what site the boat returned to, which was compared to the location (GPS) of each site on the sampling frame, which was also geocoded from addresses and Google Maps aerial views. These complications with determining location made them especially vulnerable to measurement error, therefore the definitions of agreement on location for a potential link for record linkage (i.e. agree, disagree, and close), required human judgment.

We were not provided with the names of boats encountered in the intercept sample due to confidentiality concerns. Instead, NOAA employees attempted to link the names of vessels participating in the experiment to the names of charter boats on the sampling frame. Then they provided an identification number code for each such record from the APAIS. We had no alternative source of information of the identity of the vessels, and no way to detect whether the participating boats (114 boats) had been correctly identified by NOAA.

3.2.2. Algorithm Description

Our linkage algorithm used vessel ID number as the sole blocking variable. There are many options for linking variables, since there are large number of items in common on the intercept questionnaire and the catch reports. These variables range from trip descriptions (date, time, and latitude/longitude of return, number of anglers) to the total number of fish caught, total discards, and number of species caught and re-

leased. Most items were asked in a similar manner on the two questionnaires, thus we expected that if a pair of records constitute a match, their individual variable values should agree or be close. The major exception is location due to the complications described above.

We chose *number of species caught*, *number of species released*, *total catch*, *total release*, *number of anglers*, *date*, *location*, and *vessel ID number* as linking variables. We acknowledge these variables are likely not independent of each other, but we must assume they are for the linking score to be accurately estimated as the sum of the variable components. A violation of this assumption could make the linking score not have the optimality properties ascribed to the Felligi-Sunter linking rule in (3.1). However, Herzog et al. (2007) wrote that the Fellegi-Sunter methods can be used even when there is dependence between the linking variables. They note that in reality, independence may not be mandatory for linking variables (Herzog et al., 2007, pg. 87). However, we hypothesize it may be mandatory for estimation or inferences that come from those linking variables

To use the record linkage score for linking, we estimated the parameters contained in (3.5) - (3.7), for each potential link and each linking variable. These parameters include: $P(y), P(x \text{ and } y \text{ are close} | M_{(a,b)} = 1)$,
$P(y \text{ value for a random record is close to } x | M_{(a,b)} = 0)$, and
$P(x \text{ and } y \text{ disagree } | M_{(a,b)} = 1)$.

First, we produced a set of all potential links, where we required agreement on the blocking variable of each potential link. Call this data set L_{nm}, synonymous to $A \times B$ from Section 3.1. There were 73,313 possible matches. Parameters from (3.5) - (3.7) that are conditioned on records forming a non-match ($M_{(1,b)} = 0$) are estimated using L_{nm} in its entirety, because the proportion of true matches within this dataset is small. After L_{nm} was formed, we filtered out possible matches that agreed on *number of anglers (A)*, *return state (S)*, and *return date(D)*, between s_1 and s_2. There were 208 record combinations meeting this criteria. Denote this set of 208 records by $\widetilde{M}$.

We consider $\widetilde{M}$ to be a set of near matches, and use it to estimate components of (3.4) that are conditional on $M_{(a,b)} = 1$, which follows the methodology of Bell et al. (1994).

We now show how estimation is done for an example linking variable, *number of anglers*. Note that each potential link is assigned a score, so estimation occurs many times. First, for some potential link from L_{nm}, let x denote the number of anglers for the record from s_1 and y the number of anglers for its record from s_2. For *number of anglers*, we define agreement between x and y to mean $x = y$. To estimate (3.5), when x and y agree, let

$$\hat{P}(y) = \frac{\text{\# of occurrences of y value among potential links}}{\text{\# of potential links}}.$$

Next, define $\widetilde{M}_{\{A\}}$ to be the set of potential links that exactly agree on the variables *return state* and *return date*, so it is similar to $\widetilde{M}$, but we loosen the agreement requirement for *number of anglers*. We define x and y to be close when $|x - y| = 1$, i.e. the number of anglers differs by 1 between the two records. Then for (3.6) we estimate

$$\hat{P}(\text{x and y are close}|M_{(a,b)} = 1) = \frac{\text{\# of times x and y are close in } \widetilde{M}_{\{A\}}}{\text{\# of links in } \widetilde{M}_A}.$$

$$\hat{P}(y \text{ value for a random record is close to } x|M_{(a,b)} = 0) = \frac{\text{\# of times y of a potential link is far from specific x}}{\text{\# of potential links}}.$$

Finally, for (3.7), when x and y disagree (defined as $|x - y| > 1$), we estimate:

$$\hat{P}(\text{x and y disagree } |M_{(a,b)} = 1) = \frac{\text{\# of times x and y disagree in } \widetilde{M}_{\{A\}}}{\text{\#number of links in } \widetilde{M}_{\{A\}}}$$

For other linking variables, such as *location*, where y contains the coordinates of the dock where sampling occurred and x contains the coordinates of the predicted end of a reported trip (using GPS information), we use additional comparison categories. For *location*, in addition to agree, close, and disagree, we also have "far" as an agreement pattern. To estimate the score for *location* when x and y are far, we simply replace "close" in (3.6) with "far" to estimate:

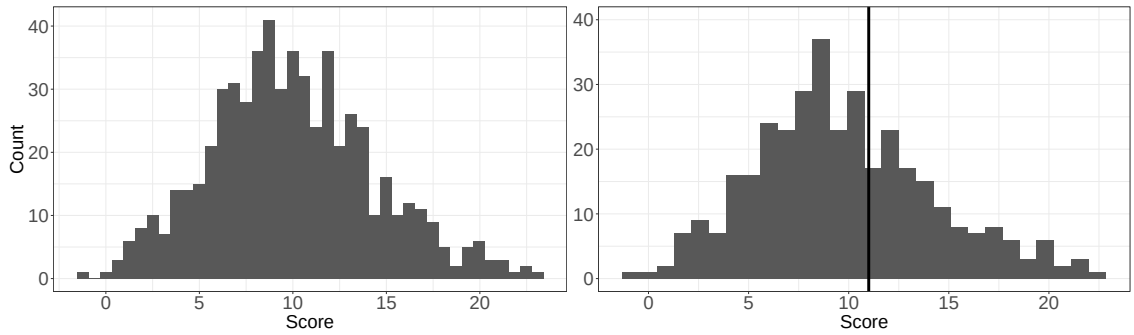
$$\hat{P}(y \text{ value for a random record is far from } x | M_{(a,b)} = 0) = \frac{\#\text{ of times y of a potential link is far from specific x}}{\#\text{ of potential links}}.$$

For *location* we define x and y to be far if the distance between x and y is between 40 and 75 km.

The requirements for the various agreement patterns among the linking variables are different. That is, we define agreement for one linking variable, such as number of anglers, differently than for another variable, such as *location*, due to the differences in the types of variables. The requirements for each agreement category are shown in Table 3.2 at the end of this chapter.

After the scores were calculated for all the potential links for all the linking variables, the scores were summed to obtain a matching score for each potential link. After the score was calculated, for every potentially link between a trip from s_2 with a trip from s_1 we kept only the report with the highest score. In case of a tie between two reports, we kept the link with the smallest distance between the reporting location and sampling location. If more than one trip from s_2 linked to a single report, we kept the link with the best score and for the remaining sampled units, obtained the next best report to link it with. Note, this is more restrictive than the general linkage model described in Chapter 2, but we used it because we felt it is a more realistic algorithmic implementation. There were 591 unique trips with a record linkage score for both years, 243 of those were from 2017. In Figure 3.1, we plot the distribution

of the unique record linkage scores for potential links. Figure 3.1(a) shows trips from both 2016 and 2017 and Figure 3.1(b) shows trips from 2017 only.



(a) All Record Linkage Scores, 2016 & 2017 (b) Record Linkage Scores, 2017 only

Figure 3.1: Distribution of Scores from Record Linkage Implementation

3.2.3. Cut-point Analysis

The next step was to determine a cut-off score. Record pairs with scores below the cutoff are not linked while record pairs with scores above the cutoff are linked. Fellegi and Sunter defined the cut-point such that specified levels of false positive and false negative error rates are met (Fellegi and Sunter (1969)). Fellegi and Sunter provide formulas for μ and λ (false positive and false negative rate, respectively) as:

$$\mu = \sum_{\gamma \in \Gamma} u(\gamma) P(A_1|\gamma) \text{ and } \lambda = \sum_{\gamma \in \Gamma} m(\gamma) P(A_3|\gamma).$$

Here, γ is the comparison vector between two potential links from the two samples to be merged. Γ is the set of all possible agreement patterns between the two data sources, $u(\gamma) = P(\gamma|M = 0)$, and $m(\gamma) = P(\gamma|M = 1)$. Then, a cut-point is chosen such that specified μ and λ are achieved.

Importantly, note μ and λ are not equivalent to $\eta_{\mathbf{r}}$ and $\gamma_{\mathbf{r}}$, respectively. In our case, we suppose the set of reports is fixed and the linking errors from Section 2.1 arise from an analyst incorrectly linking a sampled trip to a report. Fellegi and Sunter

look at each possible pairing of records and attempt to determine whether or not the records match. Thus, they cannot consider mismatches as we do. Fellegi and Sunter assume the randomness in linkage is due to some record pair generation process, while we consider the act of linking a sampled trip to a report as containing the randomness. This is why we use the *Multinomial* distribution in Section 2.1. Because our notion of $\eta_{\mathbf{r}}$ and $\gamma_{\mathbf{r}}$ differs from μ and λ (Fellegi-Sunter method), we cannot use them to choose a cut-point. Even if we could use them, we do not have an informed method of estimating $\eta_{\mathbf{r}}$ and $\gamma_{\mathbf{r}}$. We thus needed a different way to choose a cut-point.

An ideal score distribution is bi-modal and right-skewed. For 2017 we initially chose a score of 11 because there is plausibly a trough at 11, seen in Figure 3.1b. However, the choice of a cutoff score was not obvious. Figure 3.2 shows the estimated [1] value of $\hat{t}_{y2}$ for Red Snapper in Alabama and Florida in 2017 (lower graph) as well its standard error (upper graph) as functions of the cut-point. Red Snapper was chosen as the exemplar for this study because of its importance to the recreational angling community in the Gulf of Mexico and because it was studied in Liu et al. (2017). Figure 3.2 shows the choice of a cut-point has an important effect on the estimates.

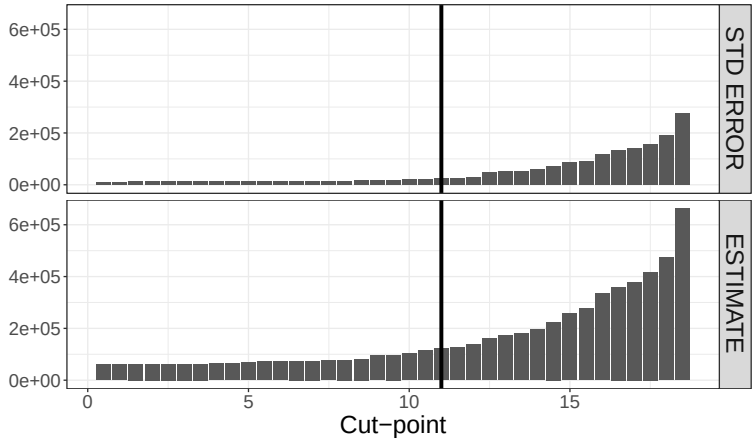


Figure 3.2: Estimate and SE of 2017 Red Snapper Harvest Using $\hat{t}_{y2}$; Cut-point of 11 Shown

[1] Estimation is done with *R's survey* package (Lumley, 2004) so that the complex features of the design were properly accounted for in the standard error.

One approach we used to evaluate our method of cut-point selection was to compare our estimate to external information from NOAA's operational estimation system. NOAA publishes estimates of catch by species, location and time period. If we take their estimates as truth, then we can estimate the bias of our estimator by comparing the two; i.e. bias was estimated for each cut-point as the difference between $\hat{t}_{y2}$ and NOAA's estimate of total Red Snapper catch in Alabama and Florida in 2017.

To compare our estimates with NOAA's, we used NOAA's MRIP query tool[2]. We queried the 2017 Red Snapper total harvest estimates for charter boats in the "all ocean combined" fishing area for the Gulf of Mexico by state (to obtain estimates for Alabama and Mississippi). A screenshot of this query is shown in Figure 3.3.

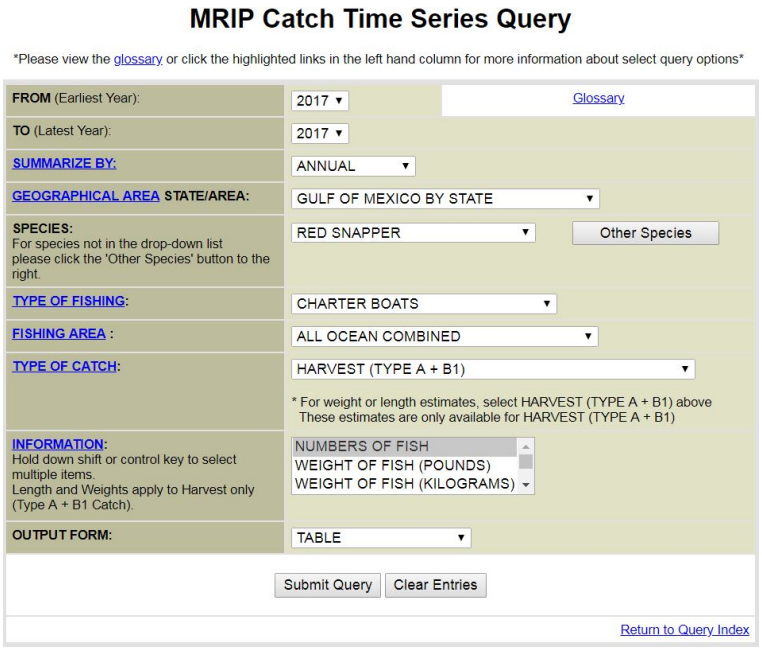


Figure 3.3: MRIP Data Query, 2017 Red Snapper Estimates

Figure 3.4 shows the bias, defined as $\hat{t}_{MRIP} - \hat{t}_{y2}$ where MRIP's estimate is denoted by $\hat{t}_{MRIP}$.

[2]https://www.st.nmfs.noaa.gov/st1/recreational/queries/

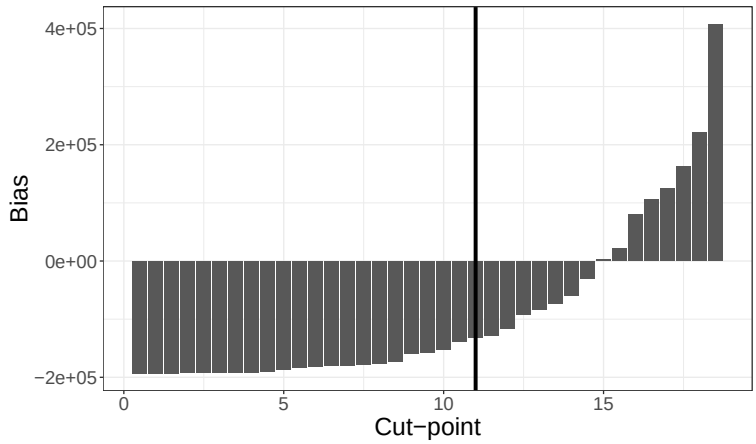


Figure 3.4: Bias of $\hat{t}_{y2}$ 2017 Red Snapper Harvest Estimate; Cut-point of 11 Shown

The cut-point score that makes the bias measure 0 is 15. When the cut-point score is below 15, more links are declared and the graph suggests that our estimator's bias is negative. When the cut-point score is above 15, the bias is positive. This is related to the relative bias expressions from Chapter 2. That is, as the cut-point increases, fewer links are declared, so the number of false negative links increases and the estimator is biased upward. Conversely, as the cut-point decreases, more links are declared, so the number of false positive links increases, and the estimator is biased downward.

By combining this idea of bias with the standard error associated with $\hat{t}_{y2}$, we can obtain a "pseudo-MSE" (pMSE) of $\hat{t}_{y2}$ as a function of the cut-point (Figure 3.5). Define pMSE as $\text{pMSE}(\hat{t}_{y,est}) = \text{Bias}^2(\hat{t}_{y,est}) + \text{SE}^2(\hat{t}_{y,est})$, where the bias is calculated as the difference between our estimate and NOAA's.

As Figures 3.2, 3.4, and 3.5 show, the initial choice of a cutoff score of 11 chosen by simple inspection of the score distribution (Figure 3.1b) does not minimize the standard error, bias, nor "pseudo-MSE" of $\hat{t}_{y2}$. A cut-point of 14.5 minimizes "pseudo-MSE" for Red Snapper.

In our application, we cannot perform clerical review, thus taking a sample of the links and non-links to estimate $\eta_{\mathbf{r}}$ and $\gamma_{\mathbf{r}}$ to attempt to balance them or achieve

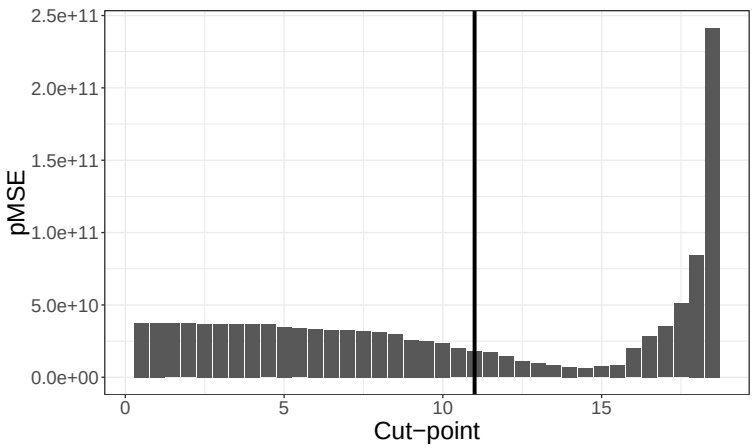


Figure 3.5: "Pseudo-MSE" of $\hat{t}_{y2}$ 2017 Red Snapper Harvest Estimate; Cut-point of 11 Shown

pre-specified levels is not possible. As part of future work, we want to investigate using the GPS data from the trip reports to come up with a more robust method of clerical review of two records. If we could sample the trips and determine the true match status for the sampled trips, we could estimate the false positive and false negative error parameters for a variety of cut-points. We hope that using an improved algorithm to predict the ending site of a trip (via the GPS data) may allow determination of true match status, This would allow estimation of false positive and negative error parameters and determination of the optimal cut-point.

3.2.4. Estimation Using Record Linkage

The purpose of creating the record linkage algorithm was to better match trips between the samples for estimation purposes. To understand the potential benefit of record linkage we estimate the total Red Snapper harvest in Alabama and Florida in 2017 (only reports from AL and FL observed in 2017) using our data matched via record linkage, for 2 different cut-points. We use a cut-point of 11, chosen from inspection of Figure 3.1b and a cut-point of 14.5, chosen from minimizing pMSE of the estimate, anchored in NOAA's publicly available estimates. The table below

presents the estimates and proportional standard error (PSE).

In Table 3.2, the estimates and PSEs are lower for a cut-point of 11 rather than 14.5. The estimates made with a cut-point of 14.5 are much closer to the NOAA estimate (254,525 (14.2)). For a cut-point of 14.5, $\hat{t}_{yp}$ produces the lowest PSE and the estimate closest to NOAA's. For both cut-points, $\hat{t}_{yc}$ performs the worst in terms of estimate and PSE, and $\hat{t}_{y2}$ has an estimate and PSE similar to $\hat{t}_{yp}$. Our initial decision of a cut-point of 11 gives an estimate of $\hat{t}_{y2}$ which are 52% smaller than NOAA's estimate ($\frac{122{,}774-254{,}525}{254{,}525} = -0.52$).

Alabama and Mississippi have implemented angler self-reporting to estimate the total catch of Red Snapper, independently of NOAA. Mississippi requires anglers who wish to catch Red Snapper to report their catch via a smart phone application or on the internet (Tails n' Scales). Anyone wishing to fish for Red Snapper must obtain a identification number before embarking on a fishing trip (MS Department of Marine Resources, 2019) and will not be granted a new identification number for their next trip if they do not report their previous one. Thus, Mississippi does not suffer from matching errors. Mississippi also does not fail an independence assumption, which states that selection in the dockside intercept is independent of reporting. Because reporting is mandatory there is no concern of a lack of independence.

Alabama estimates Red Snapper catch through a smart phone application (SnapperCheck) or a paper report dropped off at the dock (AL Department of Conservation and Natural Resources, 2019). Because the reporting is not done as meticulously as in Mississippi, matching errors are possible. Reporting is also mandatory in Alabama, but from 2014-2016 no disciplinary action was taken against non-reporters (Anson, 2017). The independence assumption does not necessarily hold because a returning trip might be sampled and the presence of the interviewer at the dock may influence the anglers to drop off a report via pen and paper at the dock.

Because NOAA has estimates of Red Snapper catch made independently of both Mississippi and Alabama, we can compare the difference in their estimates (NOAA

vs MS/AL) to the difference between our estimate and NOAA's. In 2014, Mississippi made estimates that were 84% smaller than NOAA's ($\frac{14{,}455-91{,}278}{91{,}278} = -0.84$) (of Marine Resources, 2017). In 2014, Alabama's estimates were 88% smaller than NOAA's ($\frac{43{,}532-350{,}951}{350{,}951} = -0.88$) and in 2015, Alabama's estimates were 85% smaller ($\frac{96{,}937-629{,}849}{629{,}849} = -.85$) (Anson, 2017). As stated above, our estimate (using $\hat{t}_{y2}$ and a cut-point of 11) was 55% smaller than NOAA's. Our estimate has the same trend as those from Alabama and Mississippi (smaller than NOAA's), but is closer to NOAA's estimate than those from Alabama and Mississippi. This is also of note because Mississippi does not suffer from matching errors, as our linkage does.

The estimates made by Mississippi, Alabama, and ourselves point to a possible bias in NOAA's estimates. As mentioned in Chapter 1, the address based FES, which NOAA uses to estimate effort, has a small response rate and may suffer from bias. We believe using record linkage alongside electronic reporting may offer an improvement over current NOAA methodology.

Thus far we have examined estimators from Liu et al. (2017) when undercoverage in the sampling frame and measurement error for the reported catch are both possible. Next, we introduce other scenarios that change these assumptions, and investigate estimators for these situations.

Linking Variable	Agreement Pattern (Proportion of Potential Links)
Number of Species Caught	Agree: $x = y$ (32.2%) Close: $\|x - y\| = 1$ (27.9%) Far: $\|x - y\| = 2$ (18.3%) Disagree: $\|x - y\| > 2$ (21.6%)
Number of Species Released	Agree: $x = y$ (33.2%) Close: $\|x - y\| = 1$ (37.5%) Far: $\|x - y\| = 2$ (18.75%) Disagree: $\|x - y\| > 2$ (10.6%)
Number of Fish Caught	Agree: $\|x - y\| < 1$ (9.6%) Close: $1 \leq \|x - y\| < 6$ (30.3%) Far: $6 \leq \|x - y\| < 16$ (26.9%) Farther $16 \leq \|x - y\| < 26$ (12.0%) Disagree: $\|x - y\| \geq 26$ (21.2%)
Number of Fish Discarded	Agree: $\|x - y\| < 1$ (8.6%) Close: $1 \leq \|x - y\| < 5$ (34.1%) Far: $5 \leq \|x - y\| < 11$ (18.8%) Disagree: $\|x - y\| \geq 11$ (38.5%)
Date	Agree: $\|x - y\| \leq 0.15$ days (4.2%) Close: 0.15 days $< \|x - y\| \leq 1.5$ days (1.7%) Far: 1.5 days $< \|x - y\| \leq 5$ days (4.9%) Farther: 5 days $< \|x - y\| \leq 15$ days (8.8%) Disagree: $\|x - y\| > 15$ days (80.4%)
Location	Agree: Distance between x and $y \leq 15$km (93.3%) Close: Distance between x and y between 15km and 40km (1.4%) Far: Distance between x and y between 40km and 75km (4.3%) Disagree: Distance between x and $y > 75$km (1%)
Number of Anglers	Agree: $x = y$ (75.3%) Close: $\|x - y\| = 1$ (11.4%) Disagree: $\|x - y\| > 1$ (13.3%)

Table 3.2: Agreement Patterns for Linking Variables, Proportion of Potential Links with Given Agreement Pattern in Parenthesis

Year	Method	$\hat{t}_{yp}$	$\hat{t}_{yc}$	$\hat{t}_{y2}$
2017	cut-point = 11	134,136.2 (21.2)	105,157.3 (25.2)	122,774.1 (20.9)
2017	cut-point = 14.5	243,524.8 (30.2)	167,963.1 (43.1)	224,976.6 (32.2)

Table 3.3: Estimates for Red Snapper Harvest in 2017 (AL and FL) Using 2 Different Cut-Points, Proportional Standard Error in Parentheses; NOAA's Estimate (AL and FL): 254,525 (14.2)

Chapter 4

Additional Scenarios

In South Carolina, for-hire recreational fishing captains are required to report their trips via a physical logbook. Because every trip must be reported, the total number of trips taken by recreational anglers is, in theory, known. Here, NOAA still implements a probability sample of dockside intercepts to collect data on catch, and the two sources of data are combined for estimation. Why, then, is an intercept sample necessary if reporting is required? The reason is for reporting compliance enforcement, as well as to ensure accuracy of the reported catch. That is, the sampling operation is presumably a deterrent against intentional non-reporting and the reported data is used to adjust for any non-reporting or inaccurate reporting.

The dockside survey uses a sampling frame consisting of sites accessible by interviewers, and PSUs consist of site and time pairs. As with all dockside surveys, however, some angling sites are not accessible. The time units are shifts of several hours, and may not include overnight hours. Thus, the frame inevitably contains some undercoverage. Nevertheless, it is believed both undercoverage and inaccurate reporting are small in this particular fishery, when compared to others that MRIP monitors. Therefore, estimators that would be too vulnerable to bias from undercoverage or inaccurate reporting for estimating catch by private anglers have been suggested for use in these circumstances.

In this section, we introduce two estimators that have been suggested for use in South Carolina by Breidt et al. (2018). These two estimators are useful when one can assume either (a) undercoverage of the dockside sample frame is small or (b) inaccurate reporting is minimal. Since both of these estimators require matching, they too can suffer from bias due to matching errors. In this section, we examine the

impact of matching errors on bias of these two estimators. The two estimators are

$$\hat{t}_{diff,1} = t_{y^*} + (\hat{t}_y - \hat{t}_{y^*}) \tag{4.1}$$

and

$$\hat{t}_{diff,2} = t_{y^*} + \hat{t}_{y,nr} \tag{4.2}$$

where $\hat{t}_{y,nr}$ is an estimator of the total unreported catch, made from the sample, where

$$\hat{t}_{y,nr} = \sum_{i \in s_2} w_i(1 - r_i)y_i.$$

$\hat{t}_{diff,1}$ is a classical difference estimator. It will be unbiased only if there is no undercoverage of the intercept sampling frame. $\hat{t}_{diff,2}$ also requires no undercoverage, but also requires accurate reported catch, i.e. $y_i^* = y_i$ for all i. Under these conditions these estimators have better performance (smaller variance) than $\hat{t}_{yp}$, $\hat{t}_{yc}$, and $\hat{t}_{y2}$, and are recommended when the required conditions are met.

However, $\hat{t}_{diff,1}$ and $\hat{t}_{diff,2}$ are also vulnerable to bias from matching error. As in Chapter 2, we examine their bias when matching errors are present. Here we assume the conditions for their unbiasedness are met, i.e. there is no undercoverage and catch reporting is exact. Then we determine if one estimator is more susceptible to bias from matching error than the other. Because the scenarios in which an analyst would use $\hat{t}_{diff,1}$ and $\hat{t}_{diff,2}$ are different from the scenario when he would use $\hat{t}_{yp}$, $\hat{t}_{yc}$, or $\hat{t}_{y2}$, we do not compare the sets of estimators. Additionally, due to the presence of undercoverage in the NMFS pilot study, $\hat{t}_{diff,1}$ and $\hat{t}_{diff,2}$ are not appropriate estimators. Thus we do not use these estimators to make estimates of total for our NMFS study.

4.1. Effect of Matching Errors on Relative Bias

Because $\hat{t}_{diff,1}$ and $\hat{t}_{diff,2}$ are not ratio estimators, they are unbiased for any sample size, if they are not subject to non-sampling errors. When matching errors occur, we can also derive the exact expectation of $\hat{t}_{diff,1}$ under the model described in section 2.1.1. To compute the expectation of $\hat{t}_{diff,2}$, we note that since we assume $y_i^* = y_i$, we define $\bar{y}^*_{\omega_\mathbf{r}} = \bar{y}_{\omega_\mathbf{r}}$, $\bar{y}^*_{\delta_\mathbf{r}} = \bar{y}_{\delta_\mathbf{r}}$, and $\bar{y}^*_{\eta_\mathbf{r}} = \bar{y}_{\eta_\mathbf{r}}$ where $\bar{y}^*_{\omega_\mathbf{r}}, \bar{y}^*_{\delta_\mathbf{r}}$, and $\bar{y}^*_{\eta_\mathbf{r}}$ are the expected average catch of the correctly matched, mismatches, and false positive matches, defined in (2.12), (2.13), and (2.14).

Under this model, one can show (See Appendix A)

$$E(\hat{t}_{diff,1}) = t_{y^*} + t_y - n_1[\omega_\mathbf{r}\bar{y}^*_{\omega_\mathbf{r}} + \delta_\mathbf{r}\bar{y}^*_{\delta_\mathbf{r}} + \frac{1-p_1}{p_1}\eta_\mathbf{r}\bar{y}^*_{\eta_\mathbf{r}}] \tag{4.3}$$

$$E(\hat{t}_{diff,2}) = t_{y^*} + t_y - n_1[\omega_\mathbf{r}\bar{y}_{\omega_\mathbf{r}} + \delta_\mathbf{r}\bar{y}_{\delta_\mathbf{r}} + \frac{1-p_1}{p_1}\eta_\mathbf{r}\bar{y}_{\eta_\mathbf{r}}]. \tag{4.4}$$

Then the relative bias of these estimators are

$$\mathrm{RB}(\hat{t}_{diff,1}) = \frac{\bar{y}_1^*}{\bar{y}}[p_1(1 - \omega_\mathbf{r}\lambda_{\omega_\mathbf{r}} - \delta_\mathbf{r}\lambda_{\delta_\mathbf{r}}) - (1-p_1)\eta_\mathbf{r}\lambda_{\eta_\mathbf{r}}], \tag{4.5}$$

$$\mathrm{RB}(\hat{t}_{diff,2}) = p_1(\frac{\bar{y}_1^*}{\bar{y}} - \omega_\mathbf{r}\nu_{\omega_\mathbf{r}} - \delta_\mathbf{r}\nu_{\delta_\mathbf{r}}) - (1-p_1)\eta_\mathbf{r}\nu_{\eta_\mathbf{r}} \tag{4.6}$$

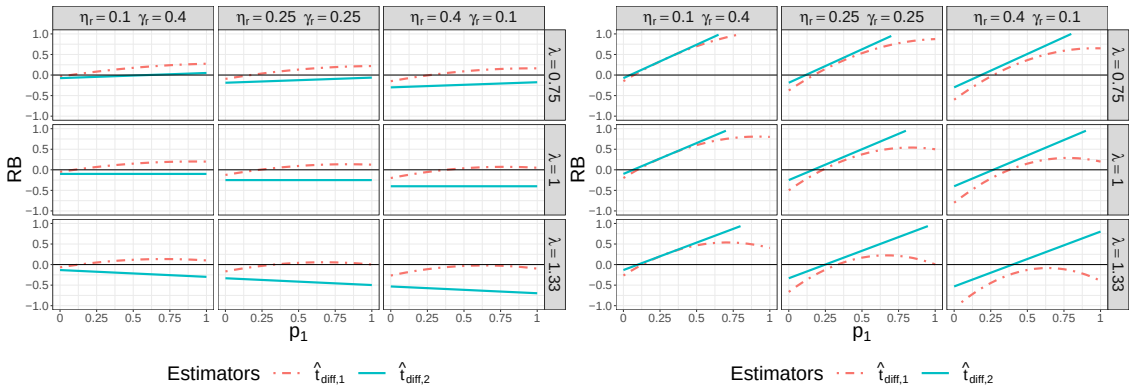
where we define discrepancy parameters (ν) as:

$$\nu_{\omega_\mathbf{r}} = \frac{\bar{y}_{\omega_\mathbf{r}}}{\bar{y}}, \nu_{\delta_\mathbf{r}} = \frac{\bar{y}_{\delta_\mathbf{r}}}{\bar{y}}, \text{and } \nu_{\eta_\mathbf{r}} = \frac{\bar{y}_{\eta_\mathbf{r}}}{\bar{y}}.$$

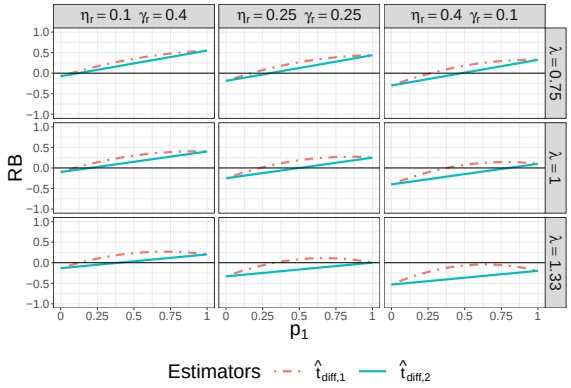
Notice that the RB of both $\hat{t}_{diff,1}$ and $\hat{t}_{diff,2}$ are functions of the reporting rate, p_1. Even though these estimators are proposed for the case when all anglers are required to report, reporting is not consistently 100%, even when mandatory. Varying the reporting rate allows us to see how the estimator fares when there is less than 100% reporting. Inspection of (4.5) and (4.6) gives insight for some special cases.

In the case where all discrepancy parameters ($\lambda's$ and $\nu's$) equal 1, first suppose mismatches to be the only matching error ($\delta_{\mathbf{r}} > 0$). Then $\text{RB}(\hat{t}_{diff,1}) = 0$, while $\text{RB}(\hat{t}_{diff,2}) = p_1(\frac{\bar{y}_1^*}{\bar{y}} - 1)$, where $\frac{\bar{y}_1^*}{\bar{y}}$ is a measure of the representativeness of the reporting. If $\frac{\bar{y}_1^*}{\bar{y}} = 1$, both estimators are unbiased. The bias of $\hat{t}_{diff,2}$ is dependent on both the reporting rate and the representativeness of reporting. Next, if false negative links are the only error allowed ($\gamma_{\mathbf{r}} > 0$), then $\text{RB}(\hat{t}_{diff,1}) = p_1\frac{\bar{y}_1^*}{\bar{y}}\gamma_{\mathbf{r}}$ and $\text{RB}(\hat{t}_{diff,2}) = p_1(\frac{\bar{y}_1^*}{\bar{y}} - 1) + p_1\gamma_{\mathbf{r}}$. If $\frac{\bar{y}_1^*}{\bar{y}} = 1$, the RB of both estimators reduces to $p_1\gamma_{\mathbf{r}}$ and the bias is positive and linear, with a maximum relative bias of 1. Last, supposing the discrepancies are 1, if false positive links are the only allowed error, ($\eta_{\mathbf{r}} > 0$), $\text{RB}(\hat{t}_{diff,1}) = -\frac{\bar{y}_1^*}{\bar{y}}(1 - p_1)\eta_{\mathbf{r}}$ and $\text{RB}(\hat{t}_{diff,2}) = p_1(\frac{\bar{y}_1^*}{\bar{y}} - 1) - (1 - p_1)\eta_{\mathbf{r}}$. If $\frac{\bar{y}_1^*}{\bar{y}} = 1$, the RB of both estimators are equal and not positive.

Now we plot the relative bias of the two new estimators when false positive and false negative links can occur simultaneously, similar to Figure 2.1. The columns of each subplot of Figure 4.1 pertain to several cases of balance in the false negative ($\gamma_{\mathbf{r}}$) and false positive error parameters ($\eta_{\mathbf{r}}$). The rows describe the catch discrepancy parameters, assuming all discrepancies are equivalent ($\lambda = \lambda_{\omega_{\mathbf{r}}} = \lambda_{\eta_{\mathbf{r}}} = \nu_{\omega_{\mathbf{r}}} = \nu_{\eta_{\mathbf{r}}}$). Each subplot considers a case of the ratio of average reported catch to average observed catch $\frac{\bar{y}_1^*}{\bar{y}}$ (0.5, 2, and 1), seen in Figure 4.1a, 4.1b, and 4.1c, respectively. In our NMFS study, Red Snapper has a representativeness measure, $\frac{\bar{y}_1^*}{\bar{y}}$, of 3.22. The value of $\frac{\bar{y}_1^*}{\bar{y}}$ varied for other species, with an $\frac{\bar{y}_1^*}{\bar{y}}$ value of 0.51 for Spanish Mackerel. Though we do not know exactly why Red Snapper gives an $\frac{\bar{y}_1^*}{\bar{y}} = 3.22$, we may surmise that only successful angling trips are reported or that anglers thought only Red Snapper needed to be reported (due to the prevalence of Red Snapper reporting phone applications in AL, MS, and TX), so they only reported when they caught Red Snapper.



(a) $\frac{\bar{y}_1^*}{\bar{y}} = 0.5$ (b) $\frac{\bar{y}_1^*}{\bar{y}} = 2$



(c) $\frac{\bar{y}_1^*}{\bar{y}} = 1$

Figure 4.1: RB of $\hat{t}_{diff,1}$ and $\hat{t}_{diff,2}$ as Functions of Reporting Rate, for 3 Linkage Scenarios, 3 Reported Catch Discrepancy Values, and 3 Settings of the Representativeness of Reporting

From Figure 4.1, if $\frac{\bar{y}_1^*}{\bar{y}}$ is 0.5 or 1, the RB of $\hat{t}_{diff}$ is always larger than $\hat{t}_{diff,2}$. For $\frac{\bar{y}_1^*}{\bar{y}} = 1$ the relative bias is generally smaller than 50% (upward or downward) for the parameter settings examined. The error patterns $(\eta_{\mathbf{r}}, \gamma_{\mathbf{r}})$ shift the RB vertically down as the false positive parameter error increases and the false negative error parameter decreases. When $\frac{\bar{y}_1^*}{\bar{y}} = 0.5$, the relative bias for $\hat{t}_{diff,2}$ is decreasing in p_1, while it is increasing for $R = 1$ and 2. The largest relative bias for the two estimators is seen when $\frac{\bar{y}_1^*}{\bar{y}} = 2$, and can reach 100% in such a case. Here, the relative bias also has the largest range.

In Figure 2.1, we assumed $\frac{\bar{y}_1^*}{\bar{y}} = 1$. There, the ARB of all the estimators had a larger range than Figure 4.1(c) does when $\frac{\bar{y}_1^*}{\bar{y}} = 1$. For $\frac{\bar{y}_1^*}{\bar{y}} = 2$, closest to what we saw for Red Snapper in the NMFS experiment, we recommend $\hat{t}_{diff,2}$ for large reporting rates (always best for reporting rates larger than about 30% regardless of other parameter settings in that case). In South Carolina, reporting rates are large because reporting is mandatory. Regardless of the value of $\frac{\bar{y}_1^*}{\bar{y}}$, bias due to matching error is a serious problem in South Carolina, even with large reporting rates. In this scenario when $\frac{\bar{y}_1^*}{\bar{y}} = 1$ (Figure 4.1c) the relative bias has a smaller range than seen in our case Figure 2.1 (where $\frac{\bar{y}_1^*}{\bar{y}}$ is assumed to be 1). Thus, while matching errors are harmful to the bias in both cases, they hurt the bias of the estimators more in our NMFS scenario.

Chapter 5

Simulation Study

NOAA is motivated to identify and mitigate a variety of non-sampling errors in their samples across the country to estimate fish catch and discards. NOAA's sample designs are complex and the effects of non-sampling errors, such as matching errors, undercoverage, and lack of independence, are nearly impossible to study analytically. An additional complication is that the properties of the estimators can change for different species of fish, due to varying catch distributions as a result of fishing season lengths, bag limits, and the popularity of the fish species. NOAA has become interested in simulating a wide range of scenarios to assess the potential harm of non-sampling errors on the bias and variance of their estimates. We have designed and implemented a simulation approach to examine the effect of matching errors on the estimators presented thus far. Our simulation presents a useful tool to examine the effect of matching errors, and in the future, other non-sampling errors. In our simulation, we recreate the MRIP sample settings and examine the estimators under a few record linkage and measurement error scenarios.

Our goal was to first create a population with features of the actual catch population, and then to simulate sampling from that population with a sample design that mimics the complex features of the APAIS, including clustering and unequal selection probabilities. To do this, we followed the method described in Liu et al. (2017). First we created a population of angler trips in the Gulf of Mexico by replicating each PSU (from the 2017 MRIP) a number of times proportional to its sampling weight. Each PSU was associated with a unique dock location and date. The number of trips per PSU was obtained by randomly selecting, with replacement, from all the possible number of trips per PSU seen in the 2017 APAIS data. Next we simulated the vari-

ables: *number of anglers, Red Snapper harvest, total harvest, total release, number of species harvested*, and *number of species released*. We simulated the variables to be like what we observed in the 2017 APAIS. The specific details of the simulation is described in Appendix C.

We assigned 7% of the population of trips to have been reported (we estimated a reporting rate of approximately 7%). For these trips, variable values of the reports were generated as $x^* = x + \epsilon$, where x^* represents the variable from the reports, x represents the variable from the APAIS, and ϵ was simulated as a zero-mean $Normal$ random variable. Then x^* was rounded to an integer (or to 0 if negative). The variance of ϵ was chosen for each variable to achieve two correlation settings. The first correlation setting closely matched the correlations between the same variables from the linked trips in the 2017 NMFS experiment (via record linkage). The second set the correlations much worse than in the experiment to examine the case of large measurement error. See Table 5.3 at the end of the chapter for the specific correlation settings. The remaining reports (not sampled) were drawn at random from the 2017 self-reported data set and then the variables were given additive error in the form $x'* = x' + \epsilon'$, where $x^{'*}$ is the variable for the reported trip (not sampled), x' is the variable value from the 2017 report data, and ϵ' is a zero-mean $Normal$ random variable. $x^{'*}$ was then rounded to an integer (or to 0 if negative).

Once the population was obtained, a single iteration consisted of drawing a cluster sample of 200 PSUs with probability proportional to size from the population. Next, the self-reports were generated, as described in the previous paragraph. Then we implemented the record linkage algorithm from Section 3.2 to link trips. After record linkage, we estimated the total catch of Red Snapper and the accompanying standard errors[1] for all the estimators discussed in this work. Because the true match status is known, for each iteration we obtained the matching error parameters (false positive, mismatch, and false negative).

[1]Estimation is done with *R's survey* package (Lumley, 2004) so that the complex features of the design were properly accounted for in the standard error.

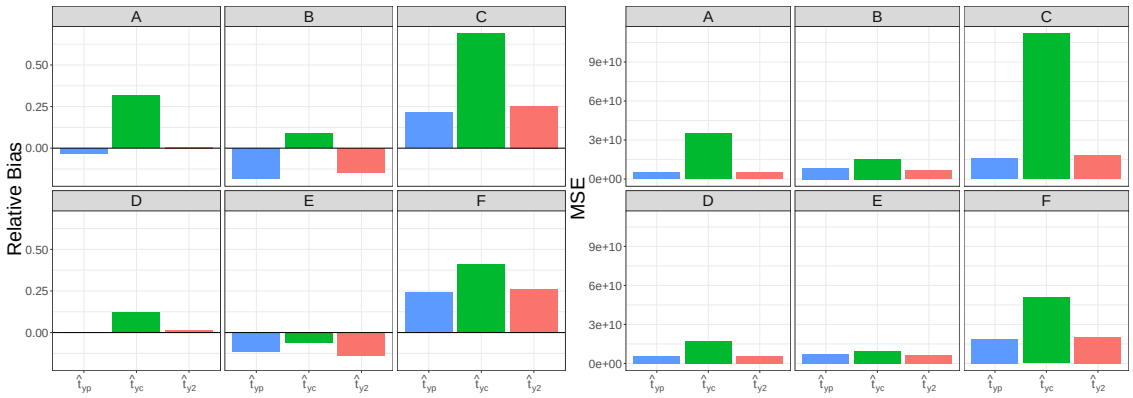
We studied three cut-points for the record linkage algorithm. The first cut-point was made to make the number of identified matches as close as possible to the actual number of matches, which allowed us to examine a best case cut-point. Then we looked at cut-points which made 20% too many links and 20% too few. With three cut-points and two correlation settings, we had 6 simulation settings, seen in Table 5.1. We simulated each scenario 10,000 times. Our simulation was implemented

Simulation	Correlation	Cut-Point
A	Good	Accurate
B	Good	20% More
C	Good	20% Fewer
D	Poor	Accurate
E	Poor	20% More
F	Poor	20% Fewer

Table 5.1: Simulation Settings

on the ManeFrame II high performance computer at Southern Methodist University, and took around 3.75 days. Relative bias and MSE for $\hat{t}_{yp}$, $\hat{t}_{yc}$, and $\hat{t}_{y2}$ are shown in Figure 5.1.

Figure 5.1 shows that in scenario A (good correlation and accurate cut-point), $\hat{t}_{y2}$ and $\hat{t}_{yp}$ have extremely similar, small, MSE's. In scenario A, $\hat{t}_{yc}$ performs much worse than $\hat{t}_{y2}$ and $\hat{t}_{yp}$ in terms of MSE. This confirms what we saw in Chapter 2. Our conclusions from Chapter 2 continue to hold when we look at the effect of the cut-point on the bias of the estimators. For scenarios C and F (cut-point too large) $\hat{t}_{yp}$, $\hat{t}_{yc}$, and $\hat{t}_{y2}$ are biased upward, and when the cut-point is too small (scenarios B and E), $\hat{t}_{yp}$ and $\hat{t}_{y2}$ are biased downward. $\hat{t}_{yc}$ is biased downward in scenario E, but is slightly biased upward in scenario B, however, $\hat{t}_{yc}$ was biased even higher upward



(a) Relative Bias (b) MSE

Figure 5.1: Simulation Results

in scenario A, when the cut-point is chosen to make the correct number of links.

Table 5.2 shows the matching error parameters for each simulation. Table 5.2

Simulation	$\eta_{\mathbf{r}}$	$\gamma_{\mathbf{r}}$	$\delta_{\mathbf{r}}$	$\omega_{\mathbf{r}}$
A	0.044	0.59	0.068	0.34
B	0.052	0.51	0.090	0.40
C	0.034	0.67	0.046	0.28
D	0.044	0.60	0.27	0.13
E	0.053	0.52	0.34	0.15
F	0.035	0.68	0.21	0.12

Table 5.2: Matching Error Parameters Observed in Simulation

shows the false positive error parameter $\eta_{\mathbf{r}}$ is always much smaller than the false negative error parameter $\gamma_{\mathbf{r}}$. This follows the pattern we saw in Section 2.1.2, where we used two linkage methods to examine the matching error rates. We also see the true positive parameter $\omega_{\mathbf{r}}$ is larger for the simulations with good correlation among the linking variables (A, B, C) and smaller for the simulations with poor correlations

(D, E, F). In simulations D, E, and F, the mismatch error parameter $\delta_{\mathbf{r}}$ is larger than it is for simulations A, B, and C, indicating that the poor correlation settings incur more mismatch links and fewer true positive links.

Scenario B is the most similar to the our NMFS scenario, by observing the matching error parameters in Table 2.2. In Table 2.2, when direct matching was assumed accurate, record linkage gave $\omega_{\mathbf{r}} = 0.53$, $\gamma_{\mathbf{r}} = 0.47$, and $\eta_{\mathbf{r}} = 0.02$. In making Table 2.2, we assumed $\delta_{\mathbf{r}} = 0$. From Table 5.2, in simulation B, $\omega_{\mathbf{r}} = 0.40$, $\delta_{\mathbf{r}} = 0.09$, $\gamma_{\mathbf{r}} = 0.51$, and $\eta_{\mathbf{r}} = 0.052$. Summing $\omega_{\mathbf{r}}$ and $\delta_{\mathbf{r}}$ here gives 0.49, which is close to the $\omega_{\mathbf{r}}$ value given in Table 2.2. Simulation B was the case when the cut-point was set to make 20% too many matches and had good correlation between the reported and observed variables. It is encouraging to find that the rates from Table 2.2 match simulation B, and point to the usefulness of this simulation.

The purpose of this simulation is to provide a preliminary framework for understanding the effects of matching error, and other non-sampling errors in the future, under a variety of scenarios. We offer the simulation to NOAA as a tool to examine the performance of their estimators under a multitude of non-sampling errors. This simulation was generated from only 1 year of data. With more data the variable distributions and correlation structure of the reports and sampled units can be better estimated. This will allow us to make further conclusions about the optimal estimator in complex settings under a wide range of errors.

Variable	Good Correlation	Poor Correlation
Number of Anglers	0.896	0.593
Total Harvest	0.734	0.545
Total Release	0.539	0.388
Red Snapper Harvest	0.895	0.626
Number of Species Caught	0.712	0.452
Number of Species Released	0.634	0.348
Date	0.912	0.848
Latitude	0.391	0.233
Longitude	0.976	0.912

Table 5.3: Correlation Settings Between Reported and Observed Variables

Chapter 6

Conclusions and Future Directions

The world we live in is one of ever increasing data creation and gathering. "Big data" is a household phrase, and more and more systems are being built to collect and harness it. Much of this data is not gathered according to rigorous sampling methodologies. One useful way valid inference can happen is by blending the large non-probability data with a probability sample. Our work has derived a framework for rigorous examination of the effects such data augmentation has on the output of an analysis. As a result of this work, we have provided NOAA with information to make better informed decisions about the future of electronic reporting of fishing trips by recreational anglers.

Specifically, we have looked at linking records between two samples, one a non-probability sample, when there is no unique identifier. The two samples constitute a capture-recapture sampling program, and the goal is estimation of the total of a characteristic of the sampling units. We have developed a generalizable model to examine the effect of matching errors on estimators of total in this setting. We have defined parameters to characterize the different types of matching errors and looked at the effects of an array of possible error levels on estimators currently proposed for such capture-recapture methodologies. We believe the contribution of bias, due to matching errors, to the mean squared error can be large enough to disqualify the use of the estimators if matching is poor. Thus, studying matching errors is of utmost importance.

In addition, we built and implemented a record linkage algorithm for our two samples. We used record linkage in a unique setting - one where clerical review of the links is not possible. Our records do not refer to people, but rather fishing trips

containing variables such as time, date, GPS coordinates, etc. Finally, we simulated NOAA's capture-recapture sampling program to examine the estimators in a complex, real setting. This simulation can serve as a guide to NOAA for analyzing the effect of matching errors under a variety of sample settings. NOAA has expressed an interest in using such large scale simulations to examine matching errors as well as other non-sampling errors. This simulation is an important contribution of this dissertation.

In general, we found the estimators proposed for the MRIP in the Gulf of Mexico ($\hat{t}_{yp}, \hat{t}_{yc}$, and $\hat{t}_{y2}$) suffer most when too many or too few records are linked. Additionally, NOAA currently uses $\hat{t}_{yc}$ for estimation. We have shown this is often not the best estimator, in terms of mean squared error. To be conservative, we recommend the use of $\hat{t}_{y2}$ in situations where undercoverage and measurement error (between y and y^*) are possible. When undercoverage and measurement error are of little concern, NOAA should use either $\hat{t}_{diff,1}$ or $\hat{t}_{diff,2}$, with the best estimator depending on the representativeness of the reported catch, reporting rate, linkage error parameters, and the catch discrepancies.

Determining the optimal estimator is difficult because we lack methods to estimate most of the values required to make such a decision. The false positive, mismatch, and false negative parameters, for example, require knowledge of the true match status of a pair of records. In our application clerical review is not possible. However, we are currently looking into use of our auxiliary GPS data as a form of clerical review. We hope to sample links and use the GPS data along with the dock location to determine the true match status of a a pair of linked records. This will allow us to estimate the false negative, mismatch, and false positive error parameters, in addition to the discrepancy parameter values. We hope the use of the GPS data will allow us to improve our record linkage algorithm, because we can find a set of true matches to use in estimating parts of the linkage score (Chapter 3). We also believe this form of clerical review will inform our decision of an optimal cut-point.

There is also more work to be done on the simulation. Primarily, we want to examine the distribution of errors between the observed and reported linking variables. With better understanding of these errors, we could make the simulation more realistic. If we use additional data from a scenario where matching errors are not a problem, like Mississippi, we can learn more about the error structure among matching and non-matching trips. The simulation we have presented here is a preliminary step toward creating a tool to examine various non-sampling errors in NOAA's estimation procedures. Given the emergence of electronic reporting, this simulation provides a crucial tool for making informed decisions about estimation procedures.

Portions of this work have been presented at the conferences *BigSurv*, *Statistics Canada's International Methodology Symposium*, *America Fisheries Society Meetings*, and *The Joint Statistical Meetings*. We have also published an R package called *blendR* in the *Journal of Open Source Software*, which provides code for the estimators given in Liu et al. (2017) (Williams, 2018). The publication of *blendR* is in Appendix D. The package is currently in use by Texas' Department of Parks and Wildlife to make estimates of Red Snapper using data gathered from an intercept sample and a smart phone application for self-reporting. NOAA shows continued interest in our work as they investigate expanding electronic reporting and the consequences of its implementation in a broader setting.

Additionally, we are including part of this work in a separate working paper that assesses the impact of a variety of non-sampling errors in an electronic reporting system setting. In the working paper, we examine undercoverage, independence, and matching errors. Matching errors have the largest potential for error in terms of bias on the estimators examined there. If we can assess the impacts of undercoverage and independence on bias and mean square error, we will be able to give better guidelines for which estimators to use in specific situations. Further work on matching methodology is called for.

This research will not only to be useful for NOAA but also to many other industries. One such avenue is elusive populations. For example, suppose a non-profit organization or government entity wishes to learn about some at-risk demographic in the United States. Because this population is difficult to reach, standard sampling procedures may prove ineffective. One solution to this problem is to recruit volunteers from this demographic to report information from a smart-phone or tablet. These records can be blended with a probability sample, perhaps one of food-stamp users, or the census, to make inferences regarding that population. In the future, we also hope to extend our work to model the effects of matching errors in data linkage on predictions from the data, not just estimation. This might include machine learning models, especially for massive data sources. We also want to investigate blending data sources on the bases of text, i.e. natural language processing, which is an area of interest of the author of this dissertation.

We believe our matching error model and record linkage algorithm can serve as useful tools when seeking to use big data from non-probability samples to make inference. Unfortunately, many firms and individuals believe big data alone will answer all possible questions about a target population or audience. Our research offers a chance to augment such data with a smaller probability sample to harness the valuable information in a theoretically sound manner. Thus, our work has significant potential value going forward.

Appendix A

Matching Error Derivations, Bias

In this Appendix we derive the expected value and relative bias of all the estimators when matching errors are present. Recall from (2.1), $\hat{n}_1 = \sum_{i=1}^{N} z_i w_i \sum_{j=1}^{N} r_j m_i(j)$. Under the model described in Section 2.1, and regarding the $r_j's$ as fixed, we have

$$
\begin{aligned}
\mathrm{E}(\hat{n}_1) &= \mathrm{E}(\sum_{i=1}^{N} z_i w_i \sum_{j=1}^{N} r_j m_i(j)) \\
&= \mathrm{E}(\sum_{i=1}^{N} r_i z_i w_i \sum_{j \in s_1} m_i(j)) + \mathrm{E}(\sum_{i=1}^{N} (1-r_i) z_i w_i \sum_{j \in s_1} m_i(j)) \\
&= \mathrm{E}(\sum_{i=1}^{N} r_i z_i w_i [(1 - \sum_{\substack{j \in s_1 \\ j \neq i}} m_i(j) - m_i(N+1)) + \sum_{\substack{j \in s_1 \\ j \neq i}} m_i(j)]) + \\
&\qquad \mathrm{E}(\sum_{i=1}^{N} (1-r_i) z_i w_i \sum_{j \in s_1} m_i(j)) \\
&= \sum_{i=1}^{N} \mathrm{E}(r_i z_i w_i) - \sum_{i=1}^{N} \mathrm{E}(r_i z_i w_i m_i(N+1)) + \sum_{i=1}^{N} \mathrm{E}((1-r_i) z_i w_i \sum_{j \in s_1} m_i(j)) \\
&= n_1 - \sum_{i=1}^{N} [\mathrm{E}(r_i z_i w_i m_i(N+1) | z_i = 1) P(z_i = 1) + \\
&\qquad \mathrm{E}(r_i z_i w_i m_i(N+1) | z_i = 0) P(z_i = 0)] + \\
&\qquad \sum_{i=1}^{N} [\mathrm{E}((1-r_i) z_i w_i \sum_{j \in s_1} m_i(j) | z_i = 1) P(z_i = 1) + \\
&\qquad \mathrm{E}((1-r_i) z_i w_i \sum_{j \in s_1} m_i(j) | z_i = 0) P(z_i = 0)]
\end{aligned}
$$

$$
\begin{aligned}
&= n_1 - \frac{n_2}{N}\sum_{i=1}^{N}\mathrm{E}(r_i z_i w_i m_i(N+1)|z_i=1)+ \\
&\qquad\qquad \frac{n_2}{N}\sum_{i=1}^{N}\mathrm{E}((1-r_i)z_i w_i \sum_{j\in s_1} m_i(j)|z_i=1) \\
&= n_1 - \sum_{i=1}^{N} r_i\pi_i(N+1) + \sum_{i=1}^{N}(1-r_i)\sum_{j\in s_1}\pi_i(j) \\
&= n_1 - n_1\gamma_{\mathbf{r}} + (N-n_1)\eta_{\mathbf{r}} \\
&= n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]. \qquad\qquad (A.1)
\end{aligned}
$$

Where $\gamma_{\mathbf{r}}$ and $\eta_{\mathbf{r}}$ are defined in (2.4) and (2.5), respectively. $\hat{t}_{y^*}$ has expectation

$$
\begin{aligned}
\mathrm{E}(\hat{t}_{y^*}) &= \mathrm{E}(\sum_{i=1}^{N} r_i z_i w_i \sum_{j\in s_1} y_j^* m_i(j)) + \mathrm{E}(\sum_{i=1}^{N}(1-r_i)z_i w_i \sum_{j\in s_1} y_j^* m_i(j)) \\
&= \mathrm{E}(\sum_{i=1}^{N} r_i z_i w_i y_i^* m_i(i)) + \mathrm{E}(\sum_{i=1}^{N} r_i z_i w_i \sum_{\substack{j\in s_1 \\ j\neq i}} y_j^* m_i(j))+ \\
&\qquad\qquad \mathrm{E}(\sum_{i=1}^{N}(1-r_i)z_i w_i \sum_{j\in s_1} y_j^* m_i(j)) \\
&= \sum_{i=1}^{N}[\mathrm{E}(r_i z_i w_i y_i^* m_i(i)|z_i=1)P(z_i=1) + \mathrm{E}(r_i z_i w_i y_i^* m_i(i)|z_i=0)P(z_i=0)]+ \\
&\qquad\qquad \sum_{i=1}^{N}[\mathrm{E}(r_i z_i w_i \sum_{\substack{j\in s_1 \\ j\neq i}} y_j^* m_i(j)|z_i=1)P(z_i=1)+ \\
&\qquad\qquad\qquad \mathrm{E}(r_i z_i w_i \sum_{\substack{j\in s_1 \\ j\neq i}} y_j^* m_i(j)|z_i=0)P(z_i=0)]+ \\
&\qquad\qquad \sum_{i=1}^{N}[\mathrm{E}((1-r_i)z_i w_i \sum_{j\in s_1} y_j^* m_i(j)|z_i=1)P(z_i=1)+ \\
&\qquad\qquad\qquad \mathrm{E}((1-r_i)z_i w_i \sum_{j\in s_1} y_j^* m_i(j)|z_i=0)P(z_i=0)]
\end{aligned}
$$

$$= \frac{n_2}{N}\sum_{i=1}^{N}\mathrm{E}(r_i z_i w_i y_i^* m_i(i)|z_i=1) + \frac{n_2}{N}\sum_{i=1}^{N}\mathrm{E}(r_i z_i w_i \sum_{\substack{j\in s_1\\ j\neq i}} y_j^* m_i(j)|z_i=1) +$$
$$\frac{n_2}{N}\sum_{i=1}^{N}\mathrm{E}((1-r_i) z_i w_i \sum_{j\in s_1} y_j^* m_i(j)|z_i=1)$$
$$= \sum_{i=1}^{N} r_i \mathrm{E}(z_i y_i^* m_i(i)|z_i=1) + \sum_{i=1}^{N} r_i \mathrm{E}(z_i \sum_{\substack{j\in s_1\\ j\neq i}} y_j^* m_i(j)|z_i=1) +$$
$$\sum_{i=1}^{N}(1-r_i)\mathrm{E}(z_i \sum_{j\in s_1} y_j^* m_i(j)|z_i=1)$$
$$= \sum_{i=1}^{N} r_i y_i^* \pi_i(i) + \sum_{i=1}^{N} r_i \sum_{\substack{j\in s_1\\ j\neq i}} y_j^* \pi_i(j) + \sum_{i=1}^{N}(1-r_i)\sum_{j\in s_1} y_j^* \pi_i(j)$$
$$= n_1 \omega_{\mathbf{r}} \bar{y}_{\omega_{\mathbf{r}}}^* + n_1 \delta_{\mathbf{r}} \bar{y}_{\delta_{\mathbf{r}}}^* + (N-n_1)\eta_{\mathbf{r}} \bar{y}_{\eta_{\mathbf{r}}}^*$$
$$= n_1[\omega_{\mathbf{r}} \bar{y}_{\omega_{\mathbf{r}}}^* + \delta_{\mathbf{r}} \bar{y}_{\delta_{\mathbf{r}}}^* + \frac{1-p_1}{p_1}\eta_{\mathbf{r}} \bar{y}_{\eta_{\mathbf{r}}}^*]. \tag{A.2}$$

Where $\delta_{\mathbf{r}}$ and $\omega_{\mathbf{r}}$ are defined in (2.6) and (2.7), respectively. $\bar{\delta}_{\mathbf{r}}$, $\bar{y}_{\omega_{\mathbf{r}}}^*$, and $\bar{y}_{\eta_{\mathbf{r}}}^*$ are defined in (2.8)-(2.10), respectively.

Using these expectations, we see $\hat{t}_{yp}$ has an expectation and ARB of

$$\mathrm{E}(\hat{t}_{yp}) \approx n_1 \frac{t_y}{n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]} = \frac{t_y}{1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}}, \tag{A.3}$$

$$\mathrm{ARB}(\hat{t}_{yp}) = \frac{\frac{t_y}{1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}} - t_y}{t_y} = \frac{\gamma_{\mathbf{r}} - \frac{1-p_1}{p_1}\eta_{\mathbf{r}}}{1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}}$$
$$= \frac{p_1\gamma_{\mathbf{r}} - (1-p_1)\eta_{\mathbf{r}}}{p_1(1-\gamma_{\mathbf{r}}) + (1-p_1)\eta_{\mathbf{r}}}. \tag{A.4}$$

Next, $\hat{t}_{yc}$ has an expected value and ARB of:

$$\mathrm{E}(\hat{t}_{yc}) \approx t_{y*}\frac{t_y}{n_1[\omega_{\mathbf{r}} \bar{y}_{\omega_{\mathbf{r}}}^* + \delta_{\mathbf{r}} \bar{y}_{\delta_{\mathbf{r}}}^* + \frac{1-p_1}{p_1}\eta_{\mathbf{r}} \bar{y}_{\eta_{\mathbf{r}}}^*]}$$
$$= \frac{\bar{y}_1^* t_y}{\omega_{\mathbf{r}} \bar{y}_{\omega_{\mathbf{r}}}^* + \delta_{\mathbf{r}} \bar{y}_{\delta_{\mathbf{r}}}^* + \frac{1-p_1}{p_1}\eta_{\mathbf{r}} \bar{y}_{\eta_{\mathbf{r}}}^*}, \tag{A.5}$$

$$\begin{aligned}
\text{ARB}(\hat{t}_{yc}) &= \frac{\frac{\bar{y}_1^* t_y}{\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}} - t_y}{t_y} \\
&= \frac{\bar{y}_1^* - (\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}})}{\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}} \\
&= \frac{1 - (\omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}})}{\omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}} \\
&= \frac{p_1(1 - \omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} - \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}}) - (1-p_1)\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}}{p_1(\omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}}) + (1-p_1)\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}}. \qquad \text{(A.6)}
\end{aligned}$$

where $\lambda_{\omega_{\mathbf{r}}} = \frac{\bar{y}^*_{\omega_{\mathbf{r}}}}{\bar{y}_1^*}, \lambda_{\delta_{\mathbf{r}}} = \frac{\bar{y}^*_{\delta_{\mathbf{r}}}}{\bar{y}_1^*}$, and $\lambda_{\eta_{\mathbf{r}}} = \frac{\bar{y}^*_{\eta_{\mathbf{r}}}}{\bar{y}_1^*}$, as defined in Chapter 2. Next $\text{E}(\hat{t}_{y2})$ has an expectation and ARB of:

$$\begin{aligned}
\text{E}(\hat{t}_{y2}) &\approx t_{y^*} + \frac{n_1}{n_1[1 - \gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}(t_y - n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]) \\
&= t_{y^*} + \frac{t_y}{1 - \gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}} - \frac{n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]}{1 - \gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}}, \qquad \text{(A.7)} \\
\text{ARB}(\hat{t}_{y2}) &= \frac{t_{y^*} + \frac{t_y}{1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}} - \frac{n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]}{1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}} - t_y}{t_y} \\
&= p_1\frac{\bar{y}_1^*}{\bar{y}} + \frac{\gamma_{\mathbf{r}} - \frac{1-p_1}{p_1}\eta_{\mathbf{r}}}{1 - \gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}} - \frac{p_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]}{\bar{y}(1 - \gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}})} \\
&= p_1\frac{\bar{y}_1^*}{\bar{y}} + \frac{\gamma_{\mathbf{r}} - \frac{1-p_1}{p_1}\eta_{\mathbf{r}}}{1 - \gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}} - \frac{p_1[\omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}]}{\frac{\bar{y}}{\bar{y}_1^*}(1 - \gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}})} \\
&= p_1\frac{\bar{y}_1^*}{\bar{y}}[1 - \frac{\omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}}{1 - \gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}}] + \frac{\gamma_{\mathbf{r}} - \frac{1-p_1}{p_1}\eta_{\mathbf{r}}}{1 - \gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}} \\
&= \text{ARB}(\hat{t}_{yp}) + p_1\frac{\bar{y}_1^*}{\bar{y}}[1 - \frac{p_1(\omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}}) + (1-p_1)\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}}{p_1(\omega_{\mathbf{r}} + \delta_{\mathbf{r}}) + (1-p_1)\eta_{\mathbf{r}}}]. \qquad \text{(A.8)}
\end{aligned}$$

Next $\hat{t}_{diff,1}$ has an expectation and ARB of:

$$\begin{aligned}
\text{E}(\hat{t}_{diff,1}) &= \text{E}(t_{y^*} + (\hat{t}_y - \hat{t}_{y^*})) \\
&= t_{y^*} + t_y - n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]
\end{aligned}$$

$$\begin{aligned}
\text{ARB}(\hat{t}_{diff,1}) &= \frac{t_{y^*} + t_y - n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}] - t_y}{t_y} \\
&= \frac{t_{y^*} - n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]}{t_y} \\
&= \frac{p_1(\bar{y}^*_1 - \omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} - \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} - \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}})}{\bar{y}} \\
&= p_1\frac{\bar{y}^*_1}{\bar{y}}(1 - \omega_{\mathbf{r}}\lambda_{\omega_{\mathbf{r}}} - \delta_{\mathbf{r}}\lambda_{\delta_{\mathbf{r}}}) - \frac{\bar{y}^*_1}{\bar{y}}(1-p_1)\eta_{\mathbf{r}}\lambda_{\eta_{\mathbf{r}}}. \qquad \text{(A.9)}
\end{aligned}$$

$\hat{t}_{diff,2}$ has an expectation and ARB of:

$$\begin{aligned}
\text{E}(\hat{t}_{diff,2}) &= \text{E}(t_{y^*} + \sum_{i=1}^{N} z_i w_i y_i (1 - \sum_{j=1}^{N} r_j m_i(j))) \\
&= t_{y^*} + \text{E}[\sum_{i=1}^{N} z_i w_i y_i - (\sum_{i=1}^{N} r_i z_i w_i y_i \sum_{j\in s_1} m_i(j) + \\
&\qquad \sum_{i=1}^{N}(1-r_i) z_i w_i y_i \sum_{j\in s_1} m_i(j))] \\
&= t_{y^*} + \sum_{i=1}^{N}[\text{E}(z_i w_i y_i | z_i = 1)P(z_i = 1) + \text{E}(z_i w_i y_i | z_i = 0)P(z_i = 0)] - \\
&\qquad \sum_{i=1}^{N}[\text{E}(r_i z_i w_i y_i m_i(i) | z_i = 1)P(z_i = 1) + \\
&\qquad \text{E}\sum_{i=1}^{N}\text{E}(r_i z_i w_i y_i m_i(i) | z_i = 0)P(z_i = 0)] - \\
&\qquad \sum_{i=1}^{N}[\text{E}(r_i z_i w_i y_i \sum_{\substack{j\in s_1 \\ j\neq i}} m_i(j) | z_i = 1)P(z_i = 1) + \\
&\qquad \text{E}(r_i z_i w_i y_i \sum_{\substack{j\in s_1 \\ j\neq i}} m_i(j) | z_i = 0)P(z_i = 0)] - \\
&\qquad \sum_{i=1}^{N}[\text{E}((1-r_i) z_i w_i y_i \sum_{j\in s_1} m_i(j) | z_i = 1)P(z_i = 1) + \\
&\qquad \text{E}((1-r_i) z_i w_i y_i \sum_{j\in s_1} m_i(j) | z_i = 0)P(z_i = 0)]
\end{aligned}$$

$$
\begin{aligned}
&= t_{y^*} + \sum_{i=1}^{N} \mathrm{E}(z_i y_i | z_i = 1) - \sum_{i=1}^{N} r_i \mathrm{E}(z_i y_i m_i(i) | z_i = 1) - \\
&\quad \sum_{i=1}^{N} r_i \mathrm{E}(z_i y_i \sum_{\substack{j \in s_1 \\ j \neq i}} m_i(j) | z_i = 1) - \sum_{i=1}^{N} (1 - r_i) \mathrm{E}(z_i y_i \sum_{j \in s_1} m_i(j) | z_i = 1) \\
&= t_{y^*} + \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} r_i y_i \pi_i(i) - \sum_{i=1}^{N} r_i y_i \sum_{\substack{j \in s_1 \\ j \neq i}} \pi_i(j) - \sum_{i=1}^{N} (1 - r_i) y_i \sum_{j \in s_1} \pi_i(j) \\
&= t_{y^*} + t_y - n_1 \omega_{\mathbf{r}} \bar{y}_{\omega_{\mathbf{r}}} - n_1 \delta_{\mathbf{r}} \bar{y}_{\delta_{\mathbf{r}}} - (N - n_1) \eta_{\mathbf{r}} \bar{y}_{\eta_{\mathbf{r}}} \\
&= t_{y^*} + t_y - n_1 [\omega_{\mathbf{r}} \bar{y}_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}} \bar{y}_{\delta_{\mathbf{r}}} + \frac{1 - p_1}{p_1} \eta_{\mathbf{r}} \bar{y}_{\eta_{\mathbf{r}}}], \qquad \text{(A.10)} \\
\mathrm{ARB}(\hat{t}_{diff,2}) &= \frac{t_{y^*} + t_y - n_1 \omega_{\mathbf{r}} \bar{y}_{\omega_{\mathbf{r}}} - n_1 \delta_{\mathbf{r}} \bar{y}_{\delta_{\mathbf{r}}} - (N - n_1) \eta_{\mathbf{r}} \bar{y}_{\eta_{\mathbf{r}}} - t_y}{t_y} \\
&= \frac{p_1 (\bar{y}_1^* - \omega_{\mathbf{r}} \bar{y}_{\omega_{\mathbf{r}}} - \delta_{\mathbf{r}} \bar{y}_{\delta_{\mathbf{r}}} - \frac{1 - p_1}{p_1} \eta_{\mathbf{r}} \bar{y}_{\eta_{\mathbf{r}}})}{\bar{y}} \\
&= p_1 \frac{\bar{y}_1^*}{\bar{y}} - p_1 \omega_{\mathbf{r}} \nu_{\omega_{\mathbf{r}}} - p_1 \delta_{\mathbf{r}} \nu_{\delta_{\mathbf{r}}} - (1 - p_1) \eta_{\mathbf{r}} \nu_{\eta_{\mathbf{r}}}. \qquad \text{(A.11)}
\end{aligned}
$$

Where $\bar{y}_{\omega_{\mathbf{r}}}, \bar{y}_{\delta_{\mathbf{r}}}, \bar{y}_{\eta_{\mathbf{r}}}$ are all defined as in equations (2.8), (2.9), and (2.10) by replacing y_j^* with y_i and $\nu_{\omega_{\mathbf{r}}} = \frac{\bar{y}_{\omega_{\mathbf{r}}}}{\bar{y}}, \nu_{\delta_{\mathbf{r}}} = \frac{\bar{y}_{\delta_{\mathbf{r}}}}{\bar{y}}$, and $\nu_{\eta_{\mathbf{r}}} = \frac{\bar{y}_{\eta_{\mathbf{r}}}}{\bar{y}}$.

Appendix B

Matching Error Derivations, Variance

In this Appendix, we derive the variances of each estimator when matching errors are present. Assume a SRS design in these derivations, and that the self-reports $\mathbf{r}$ are fixed. We begin by deriving $\mathrm{V}(\hat{n}_1)$. First, denote $\sum_{j=1}^{N} r_j m_i(j)$ as $\tilde{r}_i$, then call $\mathrm{E}(\tilde{r}_i) = \sum_{j=1}^{N} r_j \pi_i(j) = \pi_i(\cdot)$ and $\overline{\pi(\cdot)} = \frac{1}{N}\sum_{i=1}^{N} \pi_i(\cdot)$.

$$\begin{aligned}
\mathrm{V}(\hat{n}_1) &= \mathrm{E}(\mathrm{V}(\hat{n}_1|z_i's)) + \mathrm{V}(\mathrm{E}(\hat{n}_1|z_i's)) \\
&= \mathrm{E}(\mathrm{V}(\sum_{i=1}^{N} z_i\tilde{r}_i|z_i's)) + \mathrm{V}(\mathrm{E}(\sum_{i=1}^{N} z_i\tilde{r}_i|z_i's)) \\
&= \mathrm{E}(\frac{N^2}{n_2^2}\sum_{i=1}^{N} z_i^2\pi_i(\cdot)(1-\pi_i(\cdot))) + \mathrm{V}(\frac{N}{n_2}\sum_{i=1}^{N} z_i\pi_i(\cdot)) \\
&\approx \frac{N^2}{n_2}p_1\overline{V}_{1M} + \frac{N^2(1-\frac{n_2}{N})}{n_2}S_\pi^2 \qquad \text{(B.1)}
\end{aligned}$$

where $\overline{V}_{1M} = \frac{1}{n_1}\sum_{i=1}^{n_1} \pi_i(\cdot)(1-\pi_i(\cdot))$ and $S_\pi^2 = \frac{1}{N}\sum_{i=1}^{N}(\pi_i(\cdot) - \overline{\pi(\cdot)})^2$. Note that if there are no errors in matching, $S_\pi^2 = p_1(1-p_1)$. Next, see:

$$\begin{aligned}
\mathrm{Cov}(\hat{t}_y, \hat{n}_1) &= \mathrm{E}[\mathrm{Cov}(\hat{t}_y, \hat{n}_1|z_i's)] + \mathrm{Cov}[\mathrm{E}(\hat{t}_y|z_i's), \mathrm{E}(\hat{n}_1|z_i's)] \\
&= \mathrm{Cov}[\mathrm{E}(\hat{t}_y|z_i's), \mathrm{E}(\hat{n}_1|z_i's)] \\
&= \mathrm{Cov}[\mathrm{E}(\frac{N}{n_2}\sum_{i=1}^{N} z_iy_i|z_i's), \mathrm{E}(\frac{N}{n_2}\sum_{i=1}^{N} z_i\tilde{r}_i|z_i's)] \\
&= \mathrm{Cov}[\frac{N}{n_2}\sum_{i=1}^{N} z_iy_i, \frac{N}{n_2}\sum_{i=1}^{N} z_i\pi_i(\cdot)] \\
&= \frac{N^2(1-\frac{n_2}{N})}{n_2}S_{y,\pi(\cdot)} \qquad \text{(B.2)}
\end{aligned}$$

where $S_{y,\pi(\cdot)} = \frac{1}{N}\sum_{i=1}^{N} y_i\pi_i(\cdot) - \bar{y}\overline{\pi(\cdot)}$. Then the variance of $\hat{t}_{yp}$ is

$$
\begin{aligned}
\mathrm{V}(\hat{t}_{yp}) = \mathrm{V}(n_1\frac{\hat{t}_y}{\hat{n}_1}) &\approx n_1^2\mathrm{V}[\frac{\mathrm{E}(\hat{t}_y)}{\mathrm{E}(\hat{n}_1)} - \frac{\mathrm{E}(\hat{t}_y)}{\mathrm{E}^2(\hat{n}_1)}(\hat{n}_1 - \mathrm{E}(\hat{n}_1)) + \frac{1}{\mathrm{E}(\hat{n}_1)}(\hat{t}_y - \mathrm{E}(\hat{t}_y))] \\
&\approx \frac{n_1^2}{\mathrm{E}^2(\hat{n}_1)}\mathrm{V}[\hat{t}_y - \frac{\mathrm{E}(\hat{t}_y)\hat{n}_1}{\mathrm{E}(\hat{n}_1)}] \\
&\approx \frac{n_1^2}{\mathrm{E}^2(\hat{n}_1)}[V(\hat{t}_y) + \frac{\mathrm{E}^2(\hat{t}_y)}{\mathrm{E}^2(\hat{n}_1)}\mathrm{V}(\hat{n}_1) - 2\frac{\mathrm{E}(\hat{t}_y)}{\mathrm{E}(\hat{n}_1)}\mathrm{Cov}(\hat{t}_y, \hat{n}_1)] \\
&\approx \frac{n_1^2}{(n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}])^2}[\frac{N^2(1-\frac{n_2}{N})}{n_2}S_y^2 + \\
&\qquad \frac{t_y^2}{(n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}])^2}[\frac{N^2}{n_2}p_1\overline{V}_{1M} + \frac{N^2(1-\frac{n_2}{N})}{n_2}S_\pi^2] - \\
&\qquad\qquad 2\frac{t_y}{n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}\frac{N^2(1-\frac{n_2}{N})}{n_2}S_{y,\pi(\cdot)}] \\
&\approx \frac{N^2(1-\frac{n_2}{N})}{n_2}\frac{n_1}{[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}[S_y^2 + \frac{t_y^2[p_1\overline{V}_{1M} + S_\pi^2]}{(n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}])^2} - \\
&\qquad\qquad 2\frac{t_y S_{y,\pi(\cdot)}}{n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}] \\
&\approx \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\frac{1}{[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}[CV_y^2 + \frac{N^2[p_1\overline{V}_{1M} + S_\pi^2]}{(n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}])^2} - \\
&\qquad\qquad \frac{2NS_{y,\pi(\cdot)}}{n_1\bar{y}[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}] \\
&\approx \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\frac{1}{[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}[CV_y^2 + \frac{p_1\overline{V}_{1M} + S_\pi^2}{(p_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}])^2} - \\
&\qquad\qquad \frac{2S_{y,\pi(\cdot)}}{p_1\bar{y}[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}].
\end{aligned}
\tag{B.3}
$$

Next, examine $V(\hat{t}_{yc})$. First, let $\tilde{y}_i^* = \sum_{j=1}^{N} y_j^* r_j m_i(j)$, $\mathrm{E}(\tilde{y}_i^*) = \sum_{j\in s_1} y_j^*\pi_i(j) = \mu_i^*(\cdot)$, $\frac{1}{N}\sum_{i=1}^{N}\mu_i^*(\cdot) = \overline{\mu^*(\cdot)}$, and $\mathrm{V}(\tilde{y}_i^*) = \sum_{i=1}^{N} y_j^{*2}\pi_i(j) - \mu_i^{*2}(\cdot) = S^2_{y^*\underset{\sim}{\pi}_i}$. Begin with:

$$
\begin{aligned}
\mathrm{V}(\hat{t}_{y^*}) &= \mathrm{E}(\mathrm{V}(\hat{t}_{y^*}|z_i's)) + \mathrm{V}(\mathrm{E}(\hat{t}_{y^*}|z_i's)) \\
&= \mathrm{E}(\mathrm{V}(\frac{N}{n_2}\sum_{i=1}^{N} z_i\tilde{y}_i^*|z_i's)) + \mathrm{V}(\mathrm{E}(\frac{N}{n_2}\sum_{i=1}^{N} z_i\tilde{y}_i^*|z_i's))
\end{aligned}
$$

$$= \mathrm{E}(\frac{N^2}{n_2^2}\sum_{i=1}^{N} z_i^2 S^2_{y^*\underset{\sim}{\pi}_i}) + \mathrm{V}(\frac{N}{n_2}\sum_{i=1}^{N} z_i \mu_i^*(\cdot))$$
$$\approx \frac{N^2}{n_2} p_1 \overline{V}_{1y^{*2}M} + \frac{N^2(1-\frac{n_2}{N})}{n_2} S^2_{\mu^*(\cdot)} \tag{B.4}$$

where $\overline{V}_{1y^{*2}M} = \frac{1}{n_1}\sum_{i\in s_1} S^2_{y^*\underset{\sim}{\pi}_i}$ and $S^2_{\mu^*(\cdot)} = \frac{1}{N-1}\sum_{i=1}^{N}(\mu_i^*(\cdot) - \overline{\mu^*(\cdot)})^2$. Next,

$$\begin{aligned}
Cov(\hat{t}_y, \hat{t}_{y^*}) &= \mathrm{E}[\mathrm{Cov}(\hat{t}_y, \hat{t}_{y^*}|z_i's)] + \mathrm{Cov}[\mathrm{E}(\hat{t}_y|z_i's), \mathrm{E}(\hat{t}_{y^*}|z_i's)] \\
&= \mathrm{Cov}[\mathrm{E}(\hat{t}_y|z_i's), \mathrm{E}(\hat{t}_{y^*}|z_i's)] \\
&= \mathrm{Cov}[\mathrm{E}(\frac{N}{n_2} z_i y_i|z_i's), \mathrm{E}(\frac{N}{n_2} z_i \tilde{y}_i^*|z_i's)] \\
&= \mathrm{Cov}[\frac{N}{n_2}\sum_{i=1}^{N} z_i y_i, \frac{N}{n_2}\sum_{i=1}^{N} z_i \mu_i^*(\cdot)] \\
&= \frac{N^2(1-\frac{n_2}{N})}{n_2} S_{y,\mu^*(\cdot)},
\end{aligned}$$

where $S_{y,\mu^*(\cdot)} = \frac{1}{N}\sum_{i=1}^{N} y_i\mu_i^*(\cdot) - \bar{y}\overline{\mu^*(\cdot)}$. This gives:

$$\begin{aligned}
V(\hat{t}_{yc}) = V(t_{y^*}\frac{\hat{t}_y}{\hat{t}_{y^*}}) &\approx t_{y^*}^2 V[\frac{\mathrm{E}(\hat{t}_y)}{\mathrm{E}(\hat{t}_{y^*})} - \frac{\mathrm{E}(\hat{t}_y)}{\mathrm{E}^2(\hat{t}_{y^*})}(\hat{t}_{y^*} - \mathrm{E}(\hat{t}_{y^*})) + \frac{1}{\mathrm{E}(\hat{t}_{y^*})}(\hat{t}_y - \mathrm{E}(\hat{t}_y))] \\
&\approx \frac{t_{y^*}^2}{\mathrm{E}^2(\hat{t}_{y^*})} V[\hat{t}_y - \frac{\mathrm{E}(\hat{t}_y)\hat{t}_{y^*}}{\mathrm{E}(\hat{t}_{y^*})}] \\
&\approx \frac{t_{y^*}^2}{\mathrm{E}^2(\hat{t}_{y^*})}[V(\hat{t}_y) + \frac{\mathrm{E}^2(\hat{t}_y)}{\mathrm{E}^2(\hat{t}_{y^*})} V(\hat{t}_{y^*}) - 2\frac{\mathrm{E}(\hat{t}_y)}{\mathrm{E}(\hat{t}_{y^*})} Cov(\hat{t}_y, \hat{t}_{y^*})] \\
&\approx \frac{t_{y^*}^2}{(n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}])^2}[\frac{N^2(1-\frac{n_2}{N})}{n_2} S_y^2 + \\
&\quad \frac{t_y^2}{(n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}])^2}[\frac{N^2}{n_2} p_1 \overline{V}_{1y^{*2}M} + \frac{N^2(1-\frac{n_2}{N})}{n_2} S^2_{\mu^*(\cdot)}] - \\
&\qquad\qquad \frac{2t_y}{n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]} \frac{N^2(1-\frac{n_2}{N})}{n_2} S_{y,\mu^*(\cdot)}]
\end{aligned}$$

$$\approx \frac{N^2(1-\frac{n_2}{N})}{n_2}\frac{\bar{y}_1^{*2}}{[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]^2}[S_y^2+ \frac{\bar{y}^2[p_1\overline{V}_{1y^{*2}M}+S^2_{\mu^*(\cdot)}]}{(p_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}])^2}- \frac{2\bar{y}}{p_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]}S_{y,\mu^*(\cdot)}]$$

$$\approx \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\frac{\bar{y}_1^{*2}}{[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]^2}[CV_y^2+ \frac{p_1\overline{V}_{1y^{*2}M}+S^2_{\mu^*(\cdot)}}{(p_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}])^2}-\frac{2S_{y,\mu^*(\cdot)}}{p_1\bar{y}[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]}]. \tag{B.5}$$

Next, move on to $V(\hat{t}_{y2})$ and note $\hat{t}_{y2}$ can be written as a ratio estimator, $\hat{t}_{y2}=t_{y^*}+n_1\frac{\hat{\bar{\delta}}}{\hat{n}_1}$, where $\hat{\bar{\delta}}_{\mathbf{r}}=\hat{t}_y-\hat{t}_{y^*}$. Then the variance can be written as:

$$\begin{aligned}
V(\hat{t}_{y2}) &= V(n_1\frac{\hat{\bar{\delta}}}{\hat{n}_1}) \approx n_1^2V(\frac{\mathrm{E}(\hat{\bar{\delta}})}{\mathrm{E}(\hat{n}_1)}-\frac{\mathrm{E}(\hat{\bar{\delta}})}{\mathrm{E}^2(\hat{n}_1)}(\hat{n}_1-\mathrm{E}(\hat{n}_1))+\frac{1}{\mathrm{E}(\hat{n}_1)}(\hat{\bar{\delta}}-\mathrm{E}(\hat{\bar{\delta}})))\\
&\approx \frac{n_1^2}{\mathrm{E}^2(\hat{n}_1)}V(\frac{\mathrm{E}(\hat{\bar{\delta}})\hat{n}_1}{\mathrm{E}(\hat{n}_1)}+\hat{\bar{\delta}})\\
&\approx \frac{n_1^2}{\mathrm{E}^2(\hat{n}_1)}\{V(\frac{\mathrm{E}(\hat{\bar{\delta}})\hat{n}_1}{\mathrm{E}(\hat{n}_1)})+V(\hat{\bar{\delta}})-2\frac{\mathrm{E}(\hat{\bar{\delta}})}{\mathrm{E}(\hat{n}_1)}\mathrm{Cov}(\hat{n}_1,\hat{\bar{\delta}})\}\\
&\approx \frac{n_1^2}{\mathrm{E}^2(\hat{n}_1)}\{V(\hat{\bar{\delta}})+\frac{\mathrm{E}^2(\hat{\bar{\delta}})}{\mathrm{E}^2(\hat{n}_1)}V(\hat{n}_1)-2\frac{\mathrm{E}(\hat{\bar{\delta}})}{\mathrm{E}(\hat{n}_1)}\mathrm{Cov}(\hat{n}_1,\hat{\bar{\delta}})\}.
\end{aligned}$$

First, we derive an expression for $\mathrm{Cov}(\hat{n}_1,\hat{t}_{y^*})$ as:

$$\begin{aligned}
\mathrm{Cov}(\hat{n}_1,\hat{t}_{y^*}) &= \mathrm{Cov}(\mathrm{E}(\hat{n}_1|z_i's),\mathrm{E}(\hat{t}_{y^*}|z_i's))+\mathrm{E}(\mathrm{Cov}(\hat{n}_1,\hat{t}_{y^*}|z_i's))\\
&= \mathrm{Cov}(\mathrm{E}(\frac{N}{n_2}\sum_{i=1}^N z_i\tilde{r}_i|z_i's),\mathrm{E}(\frac{N}{n_2}\sum_{i=1}^N z_i\tilde{y}_i^*|z_i's))+\\
&\qquad \mathrm{E}(\mathrm{Cov}(\frac{N}{n_2}\sum_{i=1}^N z_i\tilde{r}_i,\frac{N}{n_2}\sum_{i=1}^N z_i\tilde{y}_i^*|z_i's))\\
&= \mathrm{Cov}(\frac{N}{n_2}\sum_{i=1}^N z_i\pi_i(\cdot),\frac{N}{n_2}\sum_{i=1}^N z_i\mu_i^*(\cdot))+\mathrm{E}(\frac{N^2}{n_2^2}\sum_{i=1}^N\mathrm{Cov}(\tilde{r}_i,\tilde{y}_i^*))
\end{aligned}$$

$$
\begin{aligned}
&= \frac{N^2(1-\frac{n_2}{N})}{n_2} S_{\pi(\cdot),\mu^*(\cdot)} + \mathrm{E}(\frac{N^2}{n_2^2}\sum_{i=1}^{N} z_i[\mathrm{E}(\tilde{r}_i\tilde{y}_i^*) - \mathrm{E}(\tilde{r}_i)\mathrm{E}(\tilde{y}_i^*)]) \\
&= \frac{N^2(1-\frac{n_2}{N})}{n_2} S_{\pi(\cdot),\mu^*(\cdot)} + \frac{N}{n_2}\sum_{i=1}^{N}[\mu_i^*(\cdot)(1-\pi_i(\cdot))] \\
&= \frac{N^2(1-\frac{n_2}{N})}{n_2} S_{\pi(\cdot),\mu^*(\cdot)} + \frac{N^2}{n_2} p_1 \overline{V}_{1\mu^*(1-\pi(\cdot))}, \qquad \text{(B.6)}
\end{aligned}
$$

where $\overline{V}_{1\mu^*(1-\pi(\cdot))} = \frac{1}{n_1}\sum_{i\in s_1}\mu_i^*(\cdot)(1-\pi_i(\cdot))$ Then see

$$
\begin{aligned}
\mathrm{E}(\hat{\bar{\delta}}) &= \mathrm{E}(\hat{t}_y - \hat{t}_{y^*}) = t_y - n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}], \\
\mathrm{V}(\hat{\bar{\delta}}) &= V(\hat{t}_y - \hat{t}_{y^*}) = \mathrm{V}(\hat{t}_y) + \mathrm{V}(\hat{t}_{y^*}) - 2\mathrm{Cov}(\hat{t}_y, \hat{t}_{y^*}) \\
&\approx \frac{N^2(1-\frac{n_2}{N})}{n_2}[S_y^2 + p_1\overline{V}_{1y^{*2}M} + S^2_{\mu^*(\cdot)} - 2S_{y,\mu^*(\cdot)}], \\
\mathrm{Cov}(\hat{n}_1, \hat{\bar{\delta}}) &= \mathrm{Cov}(\hat{n}_1, \hat{t}_y - \hat{t}_{y^*}) = \mathrm{Cov}(\hat{n}_1, \hat{t}_y) - \mathrm{Cov}(\hat{n}_1, \hat{t}_{y^*}) \\
&\approx \frac{N^2(1-\frac{n_2}{N})}{n_2}(S_{y,\pi(\cdot)} - S_{\pi(\cdot),\mu^*(\cdot)} - p_1\overline{V}_{1\mu^*(1-\pi(\cdot))}).
\end{aligned}
$$

Now, putting pieces together,

$$
\begin{aligned}
V(\hat{t}_{y2}) &\approx \frac{n_1^2}{\mathrm{E}^2(\hat{n}_1)}\{V(\hat{\bar{\delta}}) + \frac{\mathrm{E}^2(\hat{\bar{\delta}})}{\mathrm{E}^2(\hat{n}_1)}V(\hat{n}_1) - 2\frac{\mathrm{E}(\hat{\bar{\delta}})}{\mathrm{E}(\hat{n}_1)}\mathrm{Cov}(\hat{n}_1, \hat{\bar{\delta}})\} \\
&\approx \frac{n_1^2}{(n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}])^2}\{\frac{N^2(1-\frac{n_2}{N})}{n_2}[S_y^2 + p_1\overline{V}_{1y^{*2}M} + S^2_{\mu^*(\cdot)} - 2S_{y,\mu^*(\cdot)}] \\
&\quad + \frac{(t_y - n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}])^2}{(n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}])^2}[\frac{N^2}{n_2}p_1\overline{V}_{1M} + \frac{N^2(1-\frac{n_2}{N})}{n_2}S_\pi^2] - \\
&\quad \frac{2(t_y - n_1[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}} + \delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}])}{n_1[1-\gamma_{\mathbf{r}} + \frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}\frac{N^2(1-\frac{n_2}{N})}{n_2}(S_{y,\pi(\cdot)} - \\
&\quad S_{\pi(\cdot),\mu^*(\cdot)} - p_1\overline{V}_{1\mu^*(1-\pi(\cdot))})\}
\end{aligned}
$$

$$
\begin{aligned}
&\approx \frac{N^2(1-\frac{n_2}{N})}{n_2}\frac{1}{[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}\{S_y^2+p_1\overline{V}_{1y^{*2}M}+S^2_{\mu^*(\cdot)}-2S_{y,\mu^*(\cdot)}+\\
&\qquad\frac{(\frac{1}{p_1}\bar{y}-[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}])^2}{[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}[p_1\overline{V}_{1M}+S^2_{\pi}]-\\
&\quad 2\frac{\frac{1}{p_1}\bar{y}-[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]}{[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}(S_{y,\pi(\cdot)}-S_{\pi(\cdot),\mu^*(\cdot)}-p_1\overline{V}_{1\mu^*(1-\pi(\cdot))})\}\\
&\approx \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\frac{1}{[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}\{CV_y^2+\frac{p_1\overline{V}_{1y^{*2}M}+S^2_{\mu^*(\cdot)}-2S_{y,\mu^*(\cdot)}}{\bar{y}^2}+\\
&\qquad\frac{(\frac{1}{p_1}\bar{y}-[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}])^2}{\bar{y}^2[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}[p_1\overline{V}_{1M}+S^2_{\pi}]-\\
&\quad 2\frac{\frac{1}{p_1}\bar{y}-[\omega_{\mathbf{r}}\bar{y}^*_{\omega_{\mathbf{r}}}+\delta_{\mathbf{r}}\bar{y}^*_{\delta_{\mathbf{r}}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}\bar{y}^*_{\eta_{\mathbf{r}}}]}{\bar{y}^2[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}(S_{y,\pi(\cdot)}-S_{\pi(\cdot),\mu^*(\cdot)}-p_1\overline{V}_{1\mu^*(1-\pi(\cdot))})\}
\end{aligned}
\tag{B.7}
$$

To recap, when matching errors may be present, and assuming $\lambda=\lambda_{\omega_{\mathbf{r}}}=\lambda_{\delta_{\mathbf{r}}}=\lambda_{\eta_{\mathbf{r}}}$, the variances of the estimators are:

$$
\begin{aligned}
V(\hat{t}_{yp}) &\approx \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\frac{1}{[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}[CV_y^2+\frac{p_1\overline{V}_{1M}+S^2_{\pi}}{(p_1[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}])^2}-\\
&\qquad\qquad\frac{2S_{y,\pi(\cdot)}}{p_1\bar{y}[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}]
\end{aligned}
\tag{B.8}
$$

$$
\begin{aligned}
V(\hat{t}_{yc}) &\approx \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\frac{1}{\lambda^2(1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}})^2}[CV_y^2+\\
&\qquad\frac{p_1\overline{V}_{1y^{*2}M}+S^2_{\mu^*(\cdot)}}{p_1^2\bar{y}_1^{*2}\lambda^2(1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}})^2}-\frac{2S_{y,\mu^*(\cdot)}}{p_1\bar{y}\bar{y}_1^*\lambda[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}]
\end{aligned}
\tag{B.9}
$$

$$
\begin{aligned}
V(\hat{t}_{y2}) &\approx \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\frac{1}{[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}\{CV_y^2+\frac{p_1\overline{V}_{1y^{*2}M}+S^2_{\mu^*(\cdot)}-2S_{y,\mu^*(\cdot)}}{\bar{y}^2}+\\
&\qquad\frac{(\frac{1}{p_1}\bar{y}-\bar{y}_1^*\lambda[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}])^2}{\bar{y}^2[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]^2}[p_1\overline{V}_{1M}+S^2_{\pi}]-\\
&\quad 2\frac{\frac{1}{p_1}\bar{y}-\bar{y}_1^*\lambda[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}{\bar{y}^2[1-\gamma_{\mathbf{r}}+\frac{1-p_1}{p_1}\eta_{\mathbf{r}}]}(S_{y,\pi(\cdot)}-S_{\pi(\cdot),\mu^*(\cdot)}-p_1\overline{V}_{1\mu^*(1-\pi(\cdot))})\}
\end{aligned}
\tag{B.10}
$$

When no matching errors are present, Liu et al. (2017) showed the estimators to have the following variances. Note, R_{1,yy^*} is the correlation of y and y^* in the domain of reported trips.

$$V(\hat{t}_{yp}) = \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\{CV_y^2 + (1+\frac{1}{p_1}) - 2(\frac{\bar{y}_1}{\bar{y}})\} \tag{B.11}$$

$$V(\hat{t}_{yc}) \approx V(\hat{t}_{yp}) + \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\{\frac{CV_{1y^*}^2}{p_1} - 2(\frac{\bar{y}_1}{\bar{y}})R_{1,yy^*}CV_{1y}CV_{1y^*}\} \tag{B.12}$$

$$V(\hat{t}_{y2}) \approx V(\hat{t}_{yp}) + \frac{t_y^2(1-\frac{n_2}{N})}{n_2}\{p_1(\frac{\bar{y}_1^*}{\bar{y}})CV_{1y*}[(\frac{\bar{y}_1}{\bar{y}})CV_{1y^*} - 2(\frac{\bar{y}_1}{\bar{y}})R_{1,yy^*}CV_{1y}]\}. \tag{B.13}$$

Appendix C

Simulation Details

In this appendix, we describe how we simulated the population variable values for the simulation in Chapter 5. We will examine one variable at a time. We begin with *number of anglers*. We estimated the distribution of *number of anglers* empirically from the 2017 APAIS sample, then we made draws from the estimated frequency distribution for each simulated angling trip. The empirical distribution is given in Table C.1.

For the variable *total harvest*, we saw that in the 2017 APAIS, 13.6% of trips did not harvest any fish. So we randomly assigned 13.6% of the simulated population to have a *total harvest* value of 0. The remaining population trips were first given a *total harvest* per angler value (think *CPUE*) generated from a zero-truncated $Gamma(1.8, 0.4)$ distribution. Then this harvest per angler value was multiplied by the Number of Anglers for that specific population trip, and rounded to the nearest integer, to obtain the *total harvest* for each simulated population trip.

For the variable *total release*, we followed the same steps as for *total harvest*, but we used a zero-truncated $Gamma(2, 0.35)$ random variable instead of a zero-truncated $Gamma(1.8, 0.4)$.

For *Red Snapper Harvest*, we saw that in the 2017 APAIS, 88% of trips did not catch Red Snapper. We randomly assigned 88% of the population trips to catch no Red Snapper. We randomly assigned 51.7% of the remaining trips to have a *Red Snapper harvest* per angler value of 2 (empirically estimated from the 2017 APAIS). The other 49.3% of these trips were given a *Red Snapper Harvest* per angler value drawn from a $Uniform(0.1, 1.95)$ (estimated from the 2017 APAIS). Then, this *Red Snapper Harvest* per angler value was multiplied by the *Number of Anglers* and

Value	Proportion
1	0.0481
2	0.251
3	0.169
4	0.204
5	0.114
6	0.146
7	0.0156
8	0.0176
9	0.0122
10	0.0108
11	0.00474
12	0.00609
13	0.00135

Table C.1: Empirical Distribution of *Number of Anglers* from the 2017 APAIS

rounded to the nearest integer for each population trip to obtain *Red Snapper harvest*.

For the variable *number of species caught* we estimated its disribution empirically from the 2017 APAIS. Then we made draws from the estimated frequency distribution for each simulated angling trip. The empirical distribution is given in Table C.2.

number of species released was estimated in the same way as *number of species caught* was. Its empirical distribution is in Table C.3.

Value	Proportion
0	0.137
1	0.284
2	0.28
3	0.156
4	0.0748
5	0.0404
6	0.0182
7	0.00606
8	0.00202
9	0.00135
10	0.000674

Table C.2: Empirical Distribution of *Number of Species Caught* from the 2017 APAIS

Value	Proportion
0	0.247
1	0.338
2	0.253
3	0.113
4	0.0371
5	0.0115
6	0.00135

Table C.3: Empirical Distribution of *Number of Species Released* from the 2017 APAIS

Appendix D

R Package Publication

The Journal of Open Source Software

Combining a Probability and a Non-Probability Sample in a Capture-Recapture Setting

Benjamin Williams[1]

1 Department of Statistical Science, Southern Methodist University

DOI: 10.21105/joss.00886

Software

- Review
- Repository
- Archive

Submitted: 06 August 2018
Published: 14 August 2018

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

Statistical sampling plays a vital role in understanding and making inferences with respect to all types of populations and is especially salient in a world where populations are large and big data is the new standard. In an ideal situation, samplers have access to a list of all the units in a population, called a sampling frame, from which to draw a sample. This allows them to select individual population units with known probabilities, producing a probability sample (Lohr, 2010). A sample for which probabilities of selection are not known is called a non-probability sample. Probability samples are preferred to non-probability samples because the sampling variance of the estimators calculated from probability samples can be determined using standard sampling theory. The primary disadvantage of non-probability samples is the potential for biased estimation stemming from undercoverage and a lack of representativeness in the samples. Without external sources of information, these sample deficiencies cannot be detected.

It is usually easier to obtain a non-probability sample than a probability sample. For example, a frame may not be available, which complicates the selection of a probability sample. A non-probability sample, such as one using data from volunteers, does not require an expensive nonresponse follow-up. The larger the sample, the more information it may contain. This is an attractive option for analysts working with limited resources while studying elusive populations. As a result, statisticians have begun to investigate methods for improving estimation using data from non-probability samples (for example, Elliott & Haviland (2007)). One way to improve such estimation is to combine a non-probability sample with a probability sample.

The blendR package (available on GitHub) provides four statistically valid estimators of total when combining a non-probability sample with a probability sample. These estimators have applications in many areas, such as: the internet of things, where a non-probability sample could be taken from devices connected to the internet; insurance claims, where claims could be voluntarily reported; and estimation of the death toll due to a natural disaster, where survivors could self-report deaths in a family. In each of these situations, the estimators from blendR can combine the information from the respective non-probability samples with a probability sample to make more accurate estimates. The prevalence of non-probability samples continues to grow in both academia and industry, due in large part to technological advances and the availability of big data. blendR is needed to allow analysts from a variety of disciplines to use non-probability samples to improve estimation.

The estimators are taken from Liu, Stokes, Topping, & Stunz (2017) and Breidt, Opsomer, & Huang (2018). The sampling program considers the non-probability sample as a capture sample and the probability sample as a recapture sample, meaning units selected into the non-probability sample can be again sampled into the probability sample. Capture-recapture methodology provides powerful tools to estimate the total number of units in a population (Le Cren, 1965). The goal of the four estimators presented is to make valid

estimates of the total of some variable of interest gathered in both samples. The values may disagree for units which are part of both samples (due to measurement error, for example).

The estimators from Liu et al. (2017) are ratio estimators and the one from Breidt et al. (2018) is a difference estimator. One ratio estimator uses whether or not the unit was a part of the non-probability sample as auxiliary information, one uses the value of the variable of interest gathered in the non-probability sample as auxiliary information, and the third is a weighted combination of the first two estimators. The difference estimator adds the total value of the variable of interest gathered in the non-probability sample to the estimated difference between the value of the variable in the probability sample and the value of the variable in the non-probability sample. These estimators can be used in any situation of combining samples via a capture-recapture sampling program and have many exciting possible extensions.

The estimators are currently used to estimate the total catch of the fish in several settings, including the fish Red Snapper by Texas Parks and Wildlife (TPWD). TPWD and other entities, including the National Oceanic and Atmospheric Administration (NOAA) estimate the total fish catch in the Gulf of Mexico. The blendR package provides data from a 2016 TPWD capture-recapture sampling program in which the capture sample was a non-probability sample of captains who reported the number of Red Snapper they caught via a smartphone app. The recapture sample was a dockside intercept sample in which boats were boarded and interviewers collected data about the number of Red Snapper caught (a probability sample).

The National Research Council has advised NOAA to continue experiments with electronic reporting to better estimate the total fish caught in marine waters by recreational anglers (National Research Council, 2017). Accurate estimation is critical to setting approriate fishing seasons and bag limits. As such, this is an important research field.

This work is part of dissertation research by the author (Benjamin Williams). It is also being used in working papers regarding non-sampling errors and sample size calculations for electronic reporting experiments by a fisheries research team at Southern Methodist University led by Lynne Stokes. Bug reports, contributions, and other useful comments are welcomed as issue tickets on Github and will be attended to in a timely manner.

References

Breidt, J. F., Opsomer, J. D., & Huang, C.-M. (2018). Model-assisted survey estimation with imperfectly matched auxiliary data. In *Predictive econometrics and big data*, Studies in computational intelligence (Vol. 753, pp. 21–35). Springer.

Elliott, M. N., & Haviland, A. (2007). Use of a Web-Based Convenience Sample to Supplement a Probability Sample. *Survey methodology*, *33*(2), 211–5. Retrieved from http://www.thewitnessbox.com/10498-en.pdf

Le Cren, E. D. (1965). A Note on the History of Mark-Recapture Population Estimates. *The Journal of Animal Ecology*, *34*(2), 453. doi:10.2307/2661

Liu, B., Stokes, L., Topping, T., & Stunz, G. (2017). Estimation of a Total from a Population of Unknown Size and Application to Estimating Recreational Red Snapper Catch in Texas. *Journal of Survey Statistics and Methodology*, *5*(3), 350–371. doi:10.1093/jssam/smx006

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Brooks/Cole.

National Research Council. (2017). *Review of the marine recreational information program*. Washington, D.C.: National Academies Press. doi:10.17226/24640

Bibliography

AL Department of Conservation and Natural Resources (2019). Snapper Check. https://research.dcnr.alabama.gov/Snapper/. Accessed: 2019-04-08.

Andrews, R., Brick, J. M., and Mathiowetz, W. N. A. (2014). Development and Testing of Recreational Fishing Effort Surveys. Technical report.

Anson, K. (2017). Private Recreational Electronic Census Reporting of Red Snapper Catch in Alabama: 2014-2015. Technical report, Alabama Department of Conservation and Natural Resources: Marine Resources Division.

Baker, R., Brick, M. J., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., and Tourangeau, R. (2013). Report of the AAPOR Task Force on Non-Probability Sampling.

Belin, T. R. (1993). Evaluation of Sources of Variation in Record Linkage Through a Factorial Experiment. *Survey Methodology*, 19(1):13–29.

Bell, R. M. (2017). Diverse Applications of Probabilistic Record Linkage. Schucany Lecture Series.

Bell, R. M., Keesey, J., and Richards, T. (1994). The Urge to Merge: Linking Vital Statistics Records and Medicaid Claims. *Medical Care*, 32(10):1004–1018.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, 78(2):161–188.

Boyle, J. M., Lewis, F., and Tefft, B. (2009). Cell Phone Mainly Households: Coverage and Reach for Telephone Surveys Using RDD Landline Samples. *Survey Practice*, 2(9):1–10.

Breidt, J. F. and Chromy, J. (2016). Marine recreational information program access point angler intercept survey. Support Statement Marine Recreational Information Program 0648-0659, OMD Control.

Breidt, J. F., Opsomer, J. D., and Huang, C.-M. (2018). Model-assisted survey estimation with imperfectly matched auxiliary data. In Kreinovuch, V., Sriboonchitta, S., and Chakpitak, N., editors, *Predictive Econometrics and Big Data*, volume 753 of *Studies in Computational Intelligence*, pages 21–35. Springer.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurment Error in Nonlinear Models A Modern Perspective*. Chapman and Hall/CRC, Boca Raton, FL, 2 edition.

Cochran, G., W. (1963). *Sampling Techniques*. Wiley Publication in Applied Statistics. John Wiley & Sons, 2 edition.

Copas, J. B. and Hilton, F. J. (1990). Record Linkage: Statistical Models for Matching Computer Records. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3):287.

Cren, E. D. L. (1956). A Note on the History of Mark-Recapture Population Estimates. *The Journal of Animal Ecology*, 34(2):453.

Di Consiglio, L. and Tuoto, T. (2018). When adjusting for the bias due to linkage errors: A sensitivity analysis. *Statistical Journal of the IAOS*, 34(4):589–597.

Eisenhower, D., Mathiowetz, N. A., and Morganstein, D. (1991). *Recall Error: Sources and Bias Reduction Techniques*. John Wiley and Sons, Inc.

Elliott, M. N. and Haviland, A. (2007). Use of a Web-Based Convenience Sample to Supplement a Probability Sample. *Survey methodology*, 33(2):211–5.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183.

Fuller, W. A. (2009). Some Design Properties of a Rejective Sampling Procedure. *Biometrika*, 96(4):933–944.

Hall, R. and Fienberg, S. E. (2012). Valid Statistical Inference on Automatically Matched Files. In *Privacy in Statistical Databases*, pages 131–142. Springer.

Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., and Goldstein, H. (2014). Evaluating Bias Due to Data Linkage Error in Electronic Healthcare Records. *BMC medical research methodology*, 14(1):36.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer, NY, New York.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the Americal Statistical Association*, 47:663–685.

Lahiri, P. and Larsen, M. D. (2005). Regression Analysis with Linked Data. *Journal of the American Statistical Association*, 100(469):222–230.

Liu, B., Stokes, L., Topping, T., and Stunz, G. (2017). Estimation of a Total from a Population of Unknown Size and Application to Estimating Recreational Red Snapper Catch in Texas. *Journal of Survey Statistics and Methodology*, 5(3):350–371.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. Brooks/Cole, 2 edition.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19.

MS Department of Marine Resources (2019). Mississppi Red Snapper. http://dmr.ms.gov/index.php/component/content/category/159-snapper-tails-n-scales. Accessed: 2019-04-08.

Mulry, M. H. and Spencer, B. D. (1991). Total Error in PES Estimates of Population. *Journal of the American Statistical Association*, 86(416):839.

National Research Council (2017). *Review of the Marine Recreational Information Program*. National Academies Press.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic Linkage of Vital Records. *Science*, 130(3381):954–959.

of Marine Resources, M. D. (2017). Mandatory red snapper reporting program 2016 methods and results.

Olkin, I. (1958). Multivariate Ratio Estimation for Finite Populations. *Biometrika*, 45(1):154.

Opsomer, J. (2017). South Carolina Charter Boat Validation Estimation Progress Report. Joint Statistical Meetings 2017.

Pollock, K., Turner, S., and Brown, C. (1994). Use of Capture-Recapture Techniques to Estimate Population Size and Totals when a Complete Frame is Unavailable. *Survey Methodology*, 20(2):117–124.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387):516.

Scheuren, F. and Winkler, W. E. (1993). Regression Analysis of Data Files that are Computer Matched - Part 1. *Survey Methodology*, pages 39–58.

Tancredi, A. and Liseo, B. (2011). A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems. *The Annals of Applied Statistics*, 5(2):1553–1585.

Tarnecki, J. H. and Patterson, W. F. (2015). Changes in Red Snapper Diet and Trophic Ecology Following the Deepwater Horizon Oil Spill. *Marine and Coastal Fisheries*, 7(1):135–147.

Williams, B. (2018). Combining a Probability and a Non-Probability Sample in a Capture-Recapture Setting. *Journal of Open Source Software*, 3(28):886.