Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Understanding the Effects of Record Linkage on Estimation of Total when Combining a Big Data Source with a Probability Sample

Benjamin Williams[1]

Abstract

The National Marine Fisheries Service (NMFS) estimates the total number of fish caught by recreational anglers in United States waters. NMFS arrives at this by estimating the effort (number of trips) and the catch per unit effort or CPUE (number of fish caught per trip), and then multiplying them together. Effort data are collected via a mail survey. CPUE data are collected via face-to-face intercepts of fishing trips. NMFS is experimenting with replacing the effort survey with voluntary self-reporting. The anglers report trip details via an electronic device and remain eligible for the dockside intercept. Proposed estimators of total use the reports in a capture-recapture setting (Liu et al 2017). For valid estimation, data from reporting and intercepted trips require linkage. Accurate matching is difficult in practice due to non-sampling errors. In this paper, we develop a record linkage algorithm to link trips, and examine the effect that choosing a cut point has on the final estimate.

Key Words: Sampling; Non-Probability Sampling; Record Linkage; Matching Errors.

1. Introduction

1.1 Motivating Example

Fisheries in the United States fall into two categories: commercial and recreational. Commercial fisheries are required to report their total catch. This is not the case for recreational anglers. In some areas, the total number of fish caught by recreational anglers can exceed the total catch of commercial fisheries (National Research Council, 2017). Estimating this total catch is therefore of extreme importance.

Knowing the total fish catch is critical for setting fishing season lengths, bag limits, acceptable biological catch, and annual catch limit for the various species of fish (National Research Council, 2017). The estimates are inputs to models of fish abundance. 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).

The National Marine Fisheries Service (NMFS), of the National Oceanic and Atmospheric Administration (NOAA), estimates the total catch of fish by recreational anglers via the Marine Recreational Information Program (MRIP). The product of two estimates are used to estimate total catch: effort (E), which is the number of trips taken, and catch per unit effort ($CPUE$), which is the average catch of each species per angler trip. The MRIP estimates the effort and catch per unit effort with two surveys. Each survey is a probability sample, meaning every sampling unit has a known probability of being selected into the sample.

The first sample is a dockside intercept sample, known as the Access Point Angler Intercept Survey (APAIS). APAIS is a sample of public docks. The primary sampling unit (PSU) 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. A probability proportional to size (PPS) sample design is used to select the PSU's. Interviewers go to each selected PSU to interview all recreational anglers returning from a fishing trip. The

[1]Benjamin Williams, Southern Methodist University, 3225 Daniel Avenue Dallas, TX, USA, 75205 (benjamin@smu.edu)

interviewers record statistics such as number of fish caught per species per angler, the number of anglers aboard the boat, the number of fish released per species, etc. From the APAIS sample, the *CPUE,* for each species, is estimated.

The second sample is the fishing effort survey (FES). The FES is a survey sent by mail to residents living in states that border the Gulf of Mexico. The FES is an address based survey and is a sample of all potential recreational anglers, on the National Saltwater Registry, living in a state directly adjacent to a marine fishing area (such as the Gulf of Mexico). Recipients of the FES are asked to provide their effort (the number of times they went fishing) retrospectively for the previous wave (two months) for up to five members of the household (National Research Council, 2017). An adjustment for anglers who do not live in the sampled states is made using an estimated proportion of anglers residing in the states from the dockside intercept sample. The total catch of fish by recreational anglers is finally estimated by multiplying *CPUE* and *E* for each species.

This current methodology is the result of years of careful and intentional updates. The National Research Council (NRC) has twice reviewed the MRIP and provided recommendations, once in 2006 and again in 2016. There are still problems in the estimation procedure of the MRIP, many described in the most recent NRC review.

The majority of the issues in the estimation techniques lie with the FES. The first flaw has to do with measurement error. The FES occurs at the end of a wave and asks respondents to recall the number of fishing trips made during that wave. This means they must recall events that may have taken place over two months ago. This question may be difficult for respondents to answer accurately, especially if they are avid anglers. The next issue has to do with the efficiency of the estimate production. The NMFS reports it takes approximately 45 days after each wave for final estimates of total catch. This results from the time needed for delivery and enough responses for valid estimation. Faster estimation could allow the timely setting of fishing limits.

Last, because *CPUE* is estimated from APAIS, it is made from trips returning to public sites only. An implicit assumption made in multiplying *CPUE* and *E* is the *CPUE* is the same for trips returning to public and private docks. This may be a reasonable assumption, but it also may be the case that private trips catch more or less fish per angler per trip. This is another reason to turn to electronic reporting instead of the FES, as the self-reporting allows for total catch to be reported for trips returning to private docks.

In the 2016 NRC review, the council proposed several recommendations to fix these flaws and improve estimation. The recommendation, which motivates this research, advised an evaluation of electronic data collection methods to possible replace the FES (National Research Council, 2017). The NRC noted these electronic reporting methods might allow for near real time estimation. In some areas in the United States, fisheries management institutions have begun experimenting with such techniques. In the Gulf of Mexico, the NMFS is experimenting in several states, by asking captains to self-report their trips with an electronic device.

The NMFS has collaborated with a private research firm (referred to as CLS) for this experiment. Recreational charter captains can volunteer to participate. CLS provides an electronic device to volunteers allowing them to self-report demographic and fishing data for their recreational fishing trips. Because the self-reporting occurs on an electronic device, the data is available for estimation in nearly real time. A captain can be selected into the intercept sample and report her trip with the electronic device, meaning the trip can be present in both samples. The goal of the experiment is for the voluntary sample of captains who self-report to replace the address based FES. However, this voluntary sample is a non-probability sample and so the current estimation method is not valid. These two samples are sufficient to estimate the total catch because they constitute a form of a capture-recapture model.

1.2 Current Methodology

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 given a tag to 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 caught on the first day as well, identifiable by their tag. If the proportion of tagged fish in the second sample is approximately equal to the proportion of tagged fish in the population of fish, then $E\left(\frac{n_1}{N}\right) \approx \frac{m}{n_2}$. This leads to the Lincoln-Peterson estimator of total (Cren, 1956):

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

$\widehat{N}$ is the maximum likelihood estimator under the hypergeometric model, which assumes the recapture sample is a simple random sample (SRS). This SRS assumption does not necessarily extend to the initial sample.

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. Because the APAIS is a probability sample, estimators can be similar to the Lincoln-Peterson estimator from equation 1.1. However, some of the capture units, the self-reported trips, which return to private docks have no chance at being in the recapture sample. Thus, using 1.1 requires the reporting rates for trips returning to public and private docks to be equivalent. Liu et al (2017) investigated this situation in Texas and produced consistent estimators of total catch by adapting $\widehat{N}$ from 1.1.

First, 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 $(i = 1, \dots, N)$. The goal is to estimate $t_y = \sum_{i=1}^{N} y_i$. 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^* is distinguished from y_i as measurement error resulting from inconsistencies between the captain's report and the interviewer's data.

Denote the probability sample (APAIS) by s_2 and the non-probability sample (electronic self-reports) by d_1, since it is helpful to think of reporters as a domain. There are n_2 trips sampled in s_2 and n_1 reported trips in d_1. We examine one estimator recommended in this scenario from Liu et al. (2017). This estimator is called $\hat{t}_{y2}$ and is a multivariate ratio estimator (Olkin, 1958). It has form:

$$\hat{t}_{y2} = t_{y^*} + \frac{n_1}{\hat{n}_1}\left(\hat{t}_y - \hat{t}_{y^*}\right) \tag{1.2}$$

where $t_{y^*} = \sum_{i \in d_1} y_i^*$, $\hat{n}_1 = \sum_{i \in s_2} r_i w_i$, $\hat{t}_y = \sum_{i \in s_2} w_i y_i$, $\hat{t}_{y^*} = \sum_{i \in s_2} r_i y_i^*$. Here, $r_i = 1$ if the i^{th} sampled unit was reported, and is 0 otherwise; w_i is the sampling weight, defined as the inverse of the sampling probability for the i^{th} unit. Liu et al. (2017) point out that $\hat{t}_{y2}$ is an adjusted average underreport of catch added to the total reported catch. Estimation of $\hat{n}_1$ and $\hat{t}_{y^*}$ require linking of trips which were both reported and intercepted.

2. Record Linkage

2.1 Matching Errors

The accuracy of the linking operation can have a large effect on the quality of the estimators of catch. Therefore, it is important to ensure links are as accurate as possible. Originally, we believed the boat identification number, the date and time of the trip, and the location of its return would provide a unique link which would result in a perfect match, since information about these variables are recorded in both the intercept survey data and the self-reported data. However, this was not the case. For the reports, the captain reports some of the information (e.g., number of passengers) and some comes directly from the electronic device (e.g., location). Both are subject to error, but from different causes. The location information from the reports is actually a series of GPS locations the device reports at 15 minute intervals. From the intercept survey, we have an identification number and name of a marina or other location on the frame from which the intercept PSU's are chosen. Thus, the same trip reported and intercepted will not have identical GPS locations, but rather should simply be close. Other non-sampling errors that make linkage difficult include device errors and captains reporting well after the ending of their fishing trip.

Initially, we performed a linkage operation by hand, with a rule that filtered trips that were "close" on boat ID number, date and time of each trip, and the ending location of a trip. However, we identified few matching trips using this approach. Because of the large number of reports and our knowledge of the number of electronic devices deployed, it seemed unlikely the number of reported trips encountered was as small as we found. We decided to loosen the criteria required to identify a match, or to rely on some of the many other variables reported 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.

2.2 Record Linkage

Record linkage is a process to merge two or more data files based upon variables present in both data source. When there is not a unique identifier common to the data files, record linkage is used to link them. For linking trips, we follow the record linkage techniques of Fellegi and Sunter (1959) and Bell et al. (1994).

Fellegi and Sunter (1959) formalized record linkage and their work now summarized. The two files to be linked are denoted by A and B. The set of all possible ordered pairs of links between the two files is denoted by: $A \times B \{(a, b): a \in A, b \in B\}$. This set is the union of the set of matched and nonmatched pairs; i.e.,[$M = \{(a, b): a = b, a \in A, b \in B\}$] and $U = \{(a, b): a \neq b, a \in A, b \in B\}$]. Their goal is to produce a linking rule that sorts each member of $A \times B$ into one of three possible categories: either it is declared a match (A_1), a possible match (A_2), or not a match (A_3). The linking rule is based on comparing a and b on a set of linking variables and determining for whether they agree. The result of this comparison is a score. Cut-points for the score define the linking rule. The links with high scores are assigned to A_1 while those with low scores are assigned to A_3 (Fellegi and Sunter, 1959).

Bell, Keesey, and Richards (1994), define a *match* to occur when two records (one from each data set) refer to the same unit. A *link* occurs when two records are determined to be referring to the same entity in both files (via some matching procedure). They say two records *agree* when the records display the same values on the linking variables, but are not necessarily a match. They created a score like that from Fellegi and Sunter (1959) which adds or subtracts weight based on the amount of agreement between values of the linking variables.

Denote by x and y the two values observed for that variable for a (a, b) pair. By assuming records that do not constitute matches are paired at random, the score for the k^{th} linking variable can be written as:

$$S_k = log\big(P(y|x, M = 1)\big) - log\big(P(y)\big) \tag{2.1}$$

where M is an indicator of a match. The matching score of 2.1 has a unique form for each of three potential situations: x and y *agree*, are *close*, or *disagree*. The scores for each situation, respectively, are:

$$S_k = -log\big(P(y)\big) \tag{2.2}$$

$$S_k = log\big(P(\text{x and y are close}|M = 1)\big) - log\big(P(\text{a random record is close to x}|M = 0)\big) \tag{2.3}$$

$$S_k = log\big(P(\text{x and y disagree}|M = 1)\big). \tag{2.4}$$

Equation 2.2 assumes the probability x and y agree for a match is close to 1, equation 2.3 assumes x $P(\text{x and y are close}|M = 1, x)$ are independent, and equation 2.4 assumes the y value randomly disagree with the x value for some proportion of the matches (Bell, Keesey, and Richards, 1994). Then, each piece of equations 2.2 – 2.4 must be estimates. The score in 2.1 is estimated empirically from one of the files (in our case we used intercept data). To estimate the component of 2.2 conditioned on nonmatches, we compute empirically from $A \times B$, under the approximation that almost all pairs are nonmatches. To estimate the pieces of 2.3 and 2.4 conditional on the records constituting a match, a subset of both data files that agree on a few key linking variables is assumed to be essentially a set of matches. The probability for the omitted linking variable is estimated using this set. This methodology follows from Bell, Keesey, and Richards (1994).

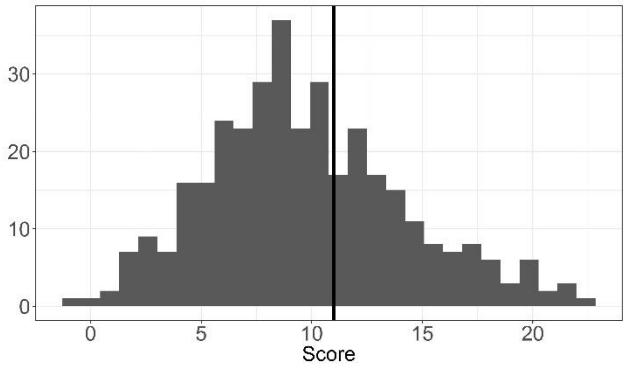
Our two files are the file of angler intercepts and the file of reported trips. We have data from two years of the NMFS experiment, 2016 and 2017. In 2016, 1569 trips were in the dockside intercept and 6514 trips were reported. In 2017, there were 1380 trips sampled in the dockside intercept and 9132 self-reported trips. The data quality, especially for the self-reported trip file, was poorer in 2016 than in 2017 as the experiment had just begun and flaws still needed to be resolved, including issues with vessel identification numbers. NOAA prepared and delivered the intercept file in their normal data production cycle. The variables available from the two files are nearly identical, but the method of collecting them differs.

For the linkage, we used vessel identification number as the sole blocking variable. We expect some true matches to disagree at least slightly on date of returning to the dock after a trip, so date was not a blocking variable. There are many options for linking variables, including date of return to dock, state to which trip returned, number of anglers, total catch of fish by all anglers, total number of species caught, latitude/longitude of dock or return location, individual catch and release numbers for over 40 species of fish, and others. We chose vessel identification number, total catch,

total discards, number of anglers on the trip, latitude/longitude of the return location recorded by the electronic device of self-reporters and the dock location specified in the dockside sample, the date on which the trip occurred, the number of species caught, and the number of species released.

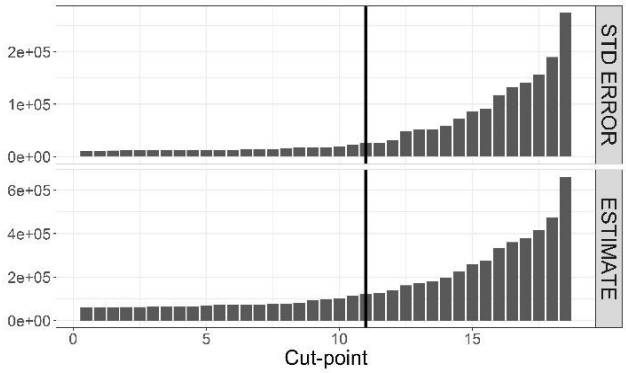
After the record linkage, a matching score was calculated for every unique potentially self-reporting vessel and date/time recorded in s_2 we kept only the self-report with the highest score. In case of a tie between two self-reported trips the link with the smallest distance between the self-report and intercept site location was kept. If any self-reports were repeated in the data, we dropped them and for those trips from s_2, obtained the next closest match from the self-reports. After the matching procedure, there were 591 unique trips with a score for both years.

Figure 1.2-1
Record Linkage Scores 2017, Cut-point of 11 shown



The next step was to determine a cut-off score. Record pairs with scores below the cut-off are not linked while record pairs with scores above the cut-off are linked. For 2017 we chose a cut-off score of 11 because there is a trough at a score of 11. An ideal score distribution is bi-modal and right-skewed. While the distribution in Figure 2.2-1 is slightly skewed, a cut-off score is not obvious. Figure 2.2-2 shows the estimated value total for the fish species Red Snapper in 2017 as well its standard error as functions of the cut-point. The cut-point certainly has an effect.

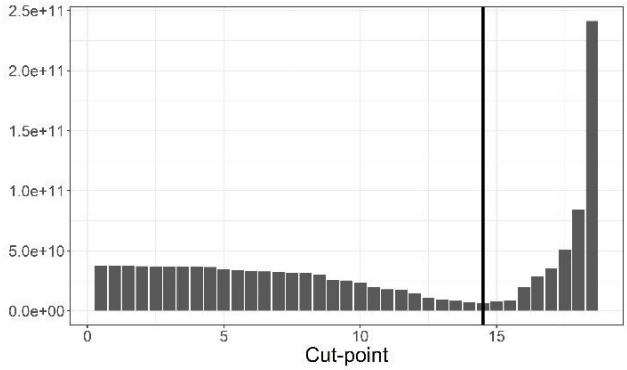
Figure 2.2-2
Estimate and SE of Total Red Snapper harvest (2017), Cut-point of 11 shown



To choose the correct cut-point, Fellegi and Sunter defined the cut-off score such that specified levels of false positive and false negative error rates are met (Fellegi and Sunter, 1959). This is certainly reasonable, but in our case, we cannot determine with certainty which trips are true matches or true non-matches. Unlike usual record linkage problems in which names are involved and clerical review can lead to correct match status determination, we are considering two fishing trips and comparing their numeric linking variables. There may be sizable measurement error in these variables, at least from the reported trips, adding to the difficulty of clerical review.

A first idea to determine the cut-point is to compare these estimates to the publicly posted estimates on the NOAA website to get an idea of the bias of $\hat{t}_{y2}$. Though this is an unsatisfactory method, due to the possible bias in NOAA's estimates themselves, it is still useful to understand the effects of the cut-point. By combining this idea of bias with the standard error associated with $\hat{t}_{y2}$, we can obtain a "pseudo-mse" of $\hat{t}_{y2}$ as a function of the cut-point (Figure 2.2-3).

Figure 2.2-3
"Pseudo-MSE" as Function of Cut-point (2017), Cut-point of 11 shown



As figures 2.2-2 and 2.2-3 show, the initial choice of a cut-off score of 11 chosen by simple inspection of the score distribution (figure 2.2-1) does not minimize the standard error or "pseudo-mse" of the $\hat{t}_{y2}$. The choice of a cut-point leads to many questions, including: how can one determine a cut-point which will affect the estimates of different species of fish in different ways. Perhaps a smaller cut-point gives better estimates for Red Snapper but not for White Grunt. We hesitate to choose a cut-point solely based on estimates made for a single species.

3. Future Work

This work is far from finished. We are currently working on estimating false positive and false negative error rates to attempt to choose a cut-point in the vein of Fellegi and Sunter (1959). We have also outlined a model for matching error to introduce randomness into the matching procedure. This will allow us to examine the effect of a variety of matching errors on the bias and variance of $\hat{t}_{y2}$ and other estimators proposed in the literature. We are also finishing a simulation study to examine the complex nature of the intercept sample. The simulation will allow us to investigate the effects of matching error and record linkage on the estimates of total. We believe this work of blending samples has incredible value in the current age of big data and in harnessing the potential of non-probability samples.

References

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

Cren, E. D. L. (1956). "A Note on the History of Mark-Recapture Population Estimates", *The Journal of Animal Ecology*, 34, pp. 453 – 454.

Fellegi, I. P., and A. B. Sunter (1969), "A Theory of Record Linkage", *Journal of the American Statistical Association*, 64, pp. 1183 – 1210.

Liu, B., L. Stokes, T. Topping, and G. Stunz (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*, 100, pp. 222 – 230.

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

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