Paper 3259-2019

Capture-Recapture Databases for Data For Good Projects

David J Corliss, Peace-Work

ABSTRACT

Capture-recapture (CRC) is a statistical methodology using multiple independent samples to estimate the size of an entire population. It is typically employed to estimate the size of hard-to-count populations. Originally developed by ecologists to estimate animal populations, capture-recapture has become an important analytic tool for social justice. CRC is used to estimate the number of people affected by a wide variety of problems, including crimes and natural disasters. In this method, careful design, development, and management of the underlying database are critical tasks. This paper demonstrates development and management of databases for capture-recapture analysis, including database organization, integrating additional data sources, addressing privacy issues, and database management and governance.

INTRODUCTION

Capture-Recapture (CRC) is a statistical process for estimating the size of hard-to-count populations using multiple, independent samples. Several different names for the process can be found in the literature, including Capture Recapture, Capture-Mark-Recapture, Mark-Recapture, Multiple System estimation (MSE), the Petersen Method, and the Lincoln Method.

The process was first developed in by C.G. Johannes Petersen in 1896 to estimate the population of a food fish, the European Plaice. Widely used to estimate biological populations, recent decades have seen increasing use in other areas, especially epidemiology, human rights, and other Data for Social Good projects. Following seminal work by Patrick Ball in 2002 on crimes against humanity in the former Yugoslavia, CRC has been used to estimate deaths in police custody, disappearances in Mexico, India, and other places; the growth of hate speech sources in social media, and other human rights applications.

Databases used in Capture-Recapture studies have special structural requirements. Particular attention must be given to the accurate identification of records in successive capture events, a quality known as Record Linkage. When this method is used in Data For Good research, this need to accurately identify individuals across multiple steps creates significant privacy issues which must be addressed. This paper describes the structural requirements of CRC databases, different methods for Record Linkage, data privacy and ethical use, and suggested governance procedures for CRC databases employed in Data For Good research.

THE CAPTURE RECAPTURE PROCESS

Capture-Recapture works by counting multiple samples of the population of interest. When member of the population is identified, it is tagged in a durable manner to assign a unique identifier. Subsequent samples count both the number identified and the number captured in previous samples using the unique identifier. The samples must be independent. A common way to facilitate this to allow sufficient time for those captured in one sampling event to mix with the population before sampling again. For example, the population of a particular species of bird in a given area can be estimated by setting up a net to capture a sample of the birds. The birds are counted, banded, and released. After allowing the captured birds to mix with the local population, a second sample counts the number captured and – using the bands placed at the first sampling – counts the subset captured in

the first sample.

Mathematically, the Lincoln-Petersen Principle states the fraction of marked individuals at time of recapture is equal to the fraction of individuals tagged after the first capture:

$$m / n = M / N$$

Re-arranging, we have

$$N = M / (m / n) = (M \times n) / m$$

In this basic example with two samples, the size of the hard-to-count population is estimated by:

1. The number captured in first sample
2. Multiplied by the number captured in the second sample
3. Divided by the number captured in both samples

It is required that the samples produced by the successive capture events be independent. In practical terms, this means the probability of being captured in a later sample is the same for both those captured and those not captured in an earlier sample. For example, if two capture events are used in the study:

$$p_{2 \in \text{Sample 1}} = p_{2 \notin \text{Sample 1}}$$

A common means of assuring the required independence is by mixing. After the first capture event, those captured and tagged in some manner are allowed to thoroughly mix with the rest of the population before subsequent capture events. While the length of time must allow those captured to mix to the point where they have the same probability of being captured in subsequent samples as those not captured, prolonging the mixing period can lead to loss of the total population through death of individual members or escape from the population being studied.

In some cases, independence can be improved by utilizing different capturing methods. For example, estimating number of fish in a lake can use fishing tackle in one capturing event and nets in another. This could be done to prevent bias toward capturing or evading capture among some individuals by a particular method.

In the case of retrospective capture, a period of mixing may not be necessary. In retrospective capture, a data source is mined after the fact, using independent methods. The mathematical requirement is for independent samples; mixing is merely one means to assure independence. One study (Corliss 2018) estimated the population of hate speech sources in Twitter. Twitter data from previous dates was mined using five different types of filters for identifying hate speech. All of the samples were effectively simultaneous, drawn from the same period of time. The different filters provided the sample independence required by the CRC process.

THE LINKAGE PROBLEM

While the mathematics of Capture Recapture are very straight forward, correct execution of the process can face some serious challenges. Perhaps the most important of these is the issue of record linkage: connecting the records in individual capture events to others. This is needed to determine which records were captured only once and which were captured multiple times.

At its simplest, we can imagine a CRC process with a unique ID for each record and two independent capture events (Table 1). In this illustrative example, 82 were captured the first time and 68 captured the second time. A total of 8 were captured both times, giving a population estimate of (82 x 68) / 8 = 697.

While this works very well for birds than can be banded – that is, permanently, physically tagged with a unique ID number – Data For Good applications can be very different. In the case of Data For Good research, Record Linkage usually consists of the determination that two records are the result of multiple contacts with the same person. Tagging meadow voles on the ear with a metal clip inscribed with a serial number is one thing – identifying and tracking human crime victims, as one example, is quite another.

As an illustration of how complex even a simple unique identifier can be, we return to the Twitter hate speech study cited earlier. Twitter adopts a public sharing business model: tweets by any user may be read by others. Further, users are assigned a unique ID. However, this ID uniquely identifies an account, not a person. It so happens that a small number of persons are responsible for the majority of hate speech on Twitter. The worst offenders often use multiple accounts under different names and email addresses. Individuals having multiple IDs can result in an over-count of the population. In the case of social media hate speech, the production of a high volume by a few individuals with multiple accounts will complicate estimating the total number of *incidents*. However, there is no large impact on the number of *sources* because very few hate sources employ multiple Twitter IDs.

As we have seen, Record Linkage can present challenges in Capture Recapture databases even where there appears to be a unique ID. In most Data For Good cases, no such apparently simple identifier exists. In such cases, record linkage can be established by multiple point identification. This can include any information available, from public sources or by informed consent – especially on program intake forms. Date information, where available, can be especially useful.

If a small sample of records are examined in detail to ascertain the quality of the match process, the results can be summarized in a Confusion Matrix:

		Actual	
		Match	Not a Match
Predicted	Match	11	1
	Not a Match	18	70

Figure 1. A Confusion Matrix of 100 Records

In the example shown in Figure 1, the matching process is found to be fairly weak: there are very few false positives and a larger proportion of false negatives. Weak matching processes decrease the denominator in the CRC process, resulting in an over-estimate of the population size. Strong matching processes with many false positives increase the

denominator, resulting in an under-estimate. (In CRC, the numerator is the product of the sample sizes and therefore is unaffected by the quality of the match.)

At a basic level, record linkage can be performed using a series of merges between the samples obtained in each successive capture event. A SET statement combines the matched files followed by the raw, unmatched samples. This produces a combined file containing all the matches first, followed by the raw records before matching. When this combined file is de-duped, first record encountered in a set of duplicates is the one kept. As a result, the de-duped file includes all the linked records, as well as any not found to be linked across multiple samples:

```
 data name_address_match;
    merge sample1 (in=a) sample2 (in=b);
    by last_name address;
    if a=1;
    if b=1;
 run;


 data name bday_match;
    merge sample1 (in=a) sample2 (in=b);
    by last_name bday;
    if a=1;
    if b=1;
 run;

 data combined;
    set name_address_match bday_match sample1 sample2;
    /* Note the order: the first the matches, then the unmatched samples */
 run;

 proc sort data=combined nodupkey;
    by ID
 run;
```

Fuzzy logic used to link records is especially helpful with name and address matches and cases where spelling can vary between samples. Other advanced linkage methods include various machine learning classification methods, and other methods. Ensemble models can be employed, using multiple matching methods to indicate or reject record linkage.

Multiple point matches can be employed, often using publicly available information. If enough miscellaneous facts align between two records, they may be considered a match:

Study ID	**398**	**787**	**688**	**398**
Probability Unique	**0.99**	**0.94**	**0.01**	**0.95**
First Name	**David**			**D**
Last Name	**Corliss**			**Corliss**
City	**Plymouth**	**New York**	**Dearborn**	**Ann Arbor**
State	**Michigan**	**New York**	**Michigan**	**Michigan**
College Attended	**Toledo**	**Columbia**		

College Graduation Year	2012	2008		
Mother's Maiden Name			Murphy	
Favorite Sports Team	Detroit Tigers	Apple Lacrosse	Detroit Lions	
Employer	Ford Motor Company	JPMorgan Chase		FCA
Job Class	Scientist	Banking	Manufacturing	Scientist
Previous Employer	Wayne State University			Ford
Previous Job Class	Teacher			Data Scientist
Membership Org 1	American Astronomical Society	Central Park Players	United Auto Workers	American Statistical Association
Membership Org 2	American Statistical Association	Tau Kappa Epsilon		
.				
. (Potentially dozens of fields or more)				
.				

Table 2. An Illustrative Example of Multiple Point Record Linkage

In this example, the first and last individuals are considered similar enough to be considered a match, despite several missing fields and even some conflicts in less important fields. The second record is too different to be a match with the first. The third record is considered too incomplete to assess the likelihood of a match with any person. The researcher may wish to consider dropping the third record here.

ETHICAL CONSIDERATIONS IN THE USE OF DATA

Of course, researchers will always want to employ best practices for the ethical use of data. Data For Good studies

Ethical considerations can include but are not limited to

- Informed consent for the use of privately obtained data
- Use consistent with the purposes stated at the time of collection
- Transfer of data to third parties, including government, law enforcement, and other agencies serving the people impacted
- Data security
- Data retention
- Data ownership

In Data For Good in particular, data ownership can be complex. For example, suppose a news media organization interviews the family of a crime victim and publicly disseminates the report. A Data For Good researcher captures this public report as part of an epidemiological study of that particular crime. In this instance, claims to the ownership of

the data contained in the news report could be made by the victim, the family, the reporter, the media service, the perpetrator, and the general public.

The following practices are recommended as a part of the researcher's data governance practices supporting ethical use of personally identifying data required for Capture Recapture databases created and utilized in Data For Good projects:

- Follow all applicable laws and regulations
- All data with the potential to identify individuals, either alone or in connection with other data, information should be *encrypted at rest*, as well as in motion
- Establish a retention policy and record the date the data were captured
- Avoid making unnecessary copies of the data
- Publicly available sources of data that may be used for record linkage should be treated in the same manner as privately sourced data
- Focus on creating security standards and practices that meet the *reasonable expectations of the persons whose data are collected*. Compliance with applicable laws is necessary, but may not be sufficient to support ethical practice in all cases.

DATABASE DESIGN

Variables contained in CRC databases are one of four types:

- Identification – a single ID identifying across multiple capture events, established using Record Linkage. This available must be unique and populated on every record.

- Sample Indicators – binary field indicating whether a particular record was captured in a specific sample. There must be one of these indicators for each capture event producing a sample of the overall population. Rules for identification must be determined in advance, before the identification of individuals captured in each sample, and clearly stated in the study design.

- Matching Fields – identifying fields used to establish record linkage. Need not be fully populated, illustrated in Table 1.

- Ancillary Modeling Fields – additional fields constituting none of the above, available for analysis and modeling the population.

ID	HateFilter1	HateFilter2	HateFilter3	HateFilter4	HateFilter5	TwitterHandle	Location	LocationMissingSpoof	AccountCreatedDate	AveTweetsPerYear
12793	1	1	0	0	0	@JAGDOB	USA	0	2/26/2017	96
14824	0	0	1	0	1	@ZGFFIZ		1	11/30/2017	4
25947	0	1	1	0	0	@EPBUTJ		1	5/18/2018	78
32329	0	0	1	1	0	@OTXUFL	Brooklyn	0	11/12/2018	72
45115	0	1	0	1	0	@POMTJN	Washington DC	0	7/15/2017	40
62483	0	1	0	1	0	@DIBGLY	Los Angeles	0	12/30/2018	95
68073	1	0	0	0	1	@MYKPPW		1	4/24/2018	330
76112	0	1	1	0	1	@BBCUCT		1	7/3/2018	71
83363	0	0	1	1	1	@CEKTMZ		1	5/23/2018	90
87808	1	0	1	0	0	@FWBKEI	Ultima Thule	1	5/8/2017	26
91591	0	1	0	0	1	@VYXNIP		1	1/15/2018	91
92184	0	1	0	0	1	@JAGDOB	Jupiter	1	9/12/2018	51

94372	1	0	0	1	0	@IRRFVR	Chicago	0	8/8/2017	89
.										
.										
.										

Table 2: Structure of a Capture-Recapture Database

In this example, ID is a unique identifier serving as the Identification field. Five samples were taken and merged into a single SAS dataset using the unique ID for record linkage. Each of the five binary Filter fields in this example serves as a Sample Indicator, indicating whether the record was captured in a particular sample. TwitterHandle is a Matching Field. While these records were matched in a SAS merge using the Identification field ID, TwitterHandle can be used to find additional matches. In the example, the records for ID=12793 and ID=92184 both have the same TwitterHandle @JAGDOB, despite having a different Twitter ID. Thee should really be regarded as a single record, with the data combined following rules determined before the study is conducted and documented in the study design. The remaining variables are Ancillary Modeling Fields – useful for subsequent analysis but not employed in the CRC process for estimating the total population.

While many CRC studies use two capture events, this example captures five separate samples of the Twitter universe. Each using a different filter to capture the records of interest – in this case, Tweets containing different forms of hate speech. The five independent samples are captured simultaneously and retrospectively - extracted from Twitter's archive of previous tweets – with the different filters providing the required independence of the samples.

POPULATION ESTIMATION USING A CAPTURE RECAPTURE DATABASE

While the database in this example was created using SAS, the population calculation was performed using the R package multimark, supporting Bayesian Capture Recapture applying an MCMC-based algorithm to a large number of samples. This type of analysis is made possible by the structural design of database, specifically,

- Keyed by a Unique ID
- One column for each independent capture sample
- Sample Indicators fields are Boolean
- No duplicate IDs

This code is used to estimate the size of the total population, including the uncertainty:

```
library(multimark)


twitter <- read.csv("C:/SASUniversityEdition/myfolders/Peace-
Work/twitter_hate_data_crc.csv", header=FALSE)


twitterm <- as.matrix(twitter)

twitterm <- twitterm[-1,] #delete column names in first row

twitterm <- twitterm[,-1]    #delete Twitter ID - Not needed for CRC
algorithm
```

```
Twittersetup <- processdata(Enc.Mat=twitterm,data.type="never")
twitter.dot <- multimarkClosed(mms=twittersetup, mod.p= ~1)
summary(twitter.dot$mcmc)
```

Algorithm Output:

Iterations = 2001:12000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
pbeta[(Intercept)]	-4.342e+00	3.956e-02	3.956e-04	7.185e-03
N	3.760e+05	1.446e+04	1.446e+02	2.520e+03
delta_1	9.999e-01	5.843e-05	5.843e-07	5.843e-07
delta_2	4.159e-05	4.154e-05	4.154e-07	4.270e-07

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
pbeta[(Intercept)]	-4.415e+00	-4.369e+00	-4.344e+00	-4.316e+00	-4.264e+00
N	3.485e+05	3.661e+05	3.761e+05	3.858e+05	4.039e+05
delta_1	9.998e-01	9.999e-01	9.999e-01	1.000e+00	1.000e+00
delta_2	9.546e-07	1.207e-05	2.869e-05	5.781e-05	1.561e-04

In addition to estimating the total population, the study used in this example performed a time series analysis by estimating the population in a series of years:

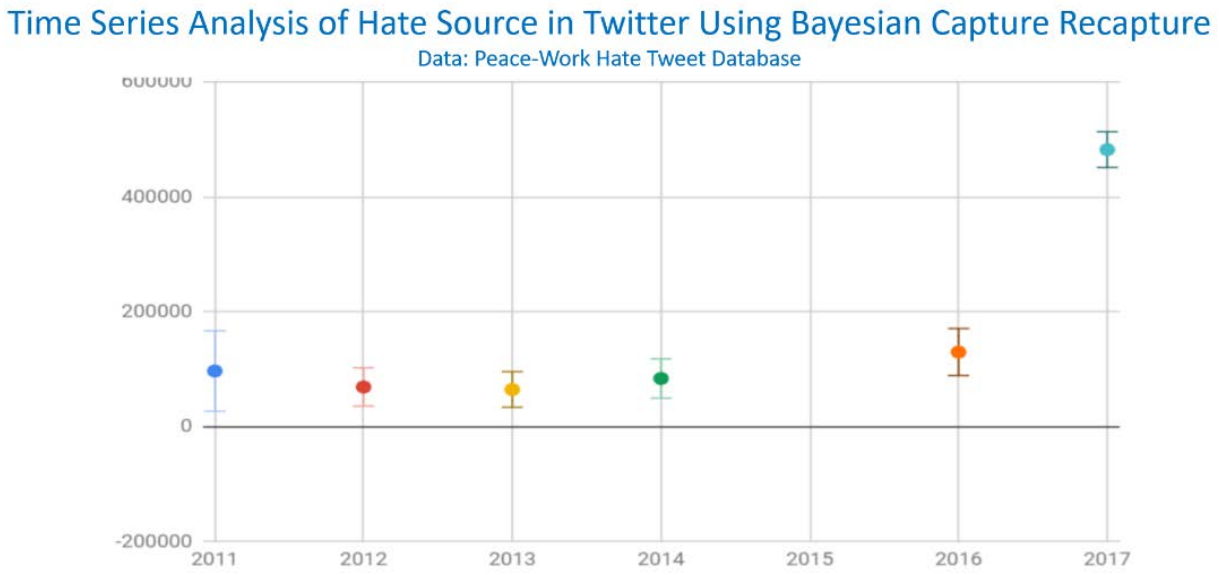

Figure 2. Bayesian Capture Recapture Estimation of Hate Speech Sources in Twitter

This analysis indicates a fairly steady level of hate speech sources in the Twitter universe between 2011 and 2016, with a sharp increase in 2017.

CONCLUSION

Capture-recapture (CRC) is an effective method for estimating the size of hard-to-count populations using multiple independent samples. In recent years, a growing body of research applies CRC in epidemiology, human rights, social justice, and other critical Data For Good applications. Accurate estimations using this method impose restrictions on the design, structure, and development of the analytic database and its variables. Database structural requirements include a unique identifier, matching variables, and a separate Boolean variable for each sample populated for every record indicating whether the record is contained in the particular sample. Record linkage – matching records across the multiple, independent samples - is required to combine data. Addressing the critical problem of record linkage is the focus of much of the research on Capture-Recapture methodology.

When CRC is employed in Data For Good studies, the requirement of record linkage results in significant issues in data privacy, governance, and the ethical uses of data. Ownership is often complicated by the presence of multiple claims and stakeholders. Storage and use of personal, highly sensitive data require adherence to all applicable laws. Above and beyond legal requirements, the focus in ethical use of data should be on security of personally identifying data and use consistent with the informed reasonable expectations of data owners. Governance recommendations for ethical use further include informed consent including research purposes and access, use consistent with the purposes stated to stakeholders when the data are obtains, strong data security, encryption in motion and at rest, and adherence to a clearly stated data retention policy.

REFERENCES

Corliss, D. J. 2018. “Bayesian Capture - Recapture in Social Justice Research”; *Young Bayesians and Big Data for Social Good,* Marseilles, France: International Centre Meetings Mathématiques

Modlin, D. 2018. Getting Started With Bayesian Analysis. Cary, NC: SAS Institute.

Smith, P. J. 1988. “Bayesian Methods for Multiple Capture-Recapture Surveys.” *Biometrics,* December 1988.

Yan, L. and Lix, L. 2016. “Use Capture-Recapture Method To Estimate Prevalence Of Disease In SAS”; *Proceedings SAS Global Forum 2016,* Las Vegas, NV: SAS Global Users Group

ACKNOWLEDGMENTS

SAS programming and results used in this paper were produced using SAS University Edition.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David J Corliss
Peace-Work
734.837.9323
davidjcorliss@peace-work.org
www.peace-work.org