Estimating Missed Detections: A "Two Types" Model

Missed Detections Working Report ECR-003

Paul B Kantor
paul.kantor@rutgers.edu
Vladimir Menkov
vmenkov@gmail.com

June 30, 2018

Abstract

We present a special case of "capture recapture," positing that captured persons are of exactly two types: those who are deterred, once and for all, by the experience of being captured, together with whatever consequences they experience; and those who, having been undeterred by experiencing the consequences once, will not be deterred by repetition of the experience. We may call these "deterables" and "persisters," We show that if a population consists of only those two classes, then data on the number of persisters recaptured two, three, four and more times can provide an estimate of the apprehension rate for that class of persons. We further show how the model can be used to estimate its own validity, using data on multiple recaptures. A simulation model is presented, which supports investigation of the several parameters of the model. Finally, we mention some possible extensions, which merit exploration if this simple model is not validated.

1 The Problem and Model

In estimating the number of missed detections at the border it helps to refer to the diagram of Figure 1. This diagram shows that some number of persons tried to get into the country and some fraction of them were caught. We don't know what fraction it is and we represent that by the letter a. Once we release the ones who have been caught back across the border, or back to their countries of origin, there are three things that each of these people might do. They might decide to never try to enter the country again (in other words they have been deterred); they might decide to enter the country again "soon" (that means within some period of time that we define); or they might decide to try to reentry enter the country again at a time that we will call "later."

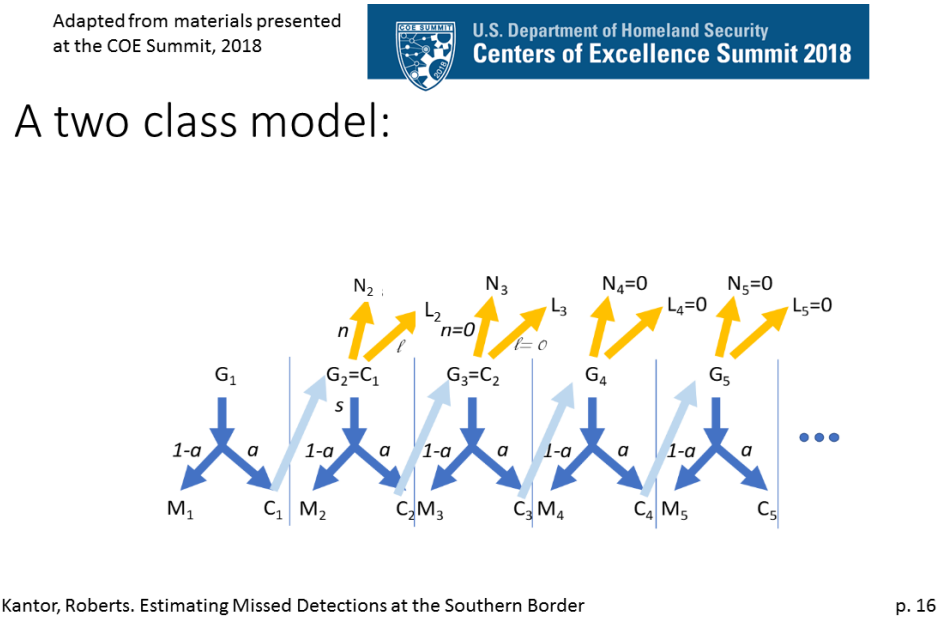

Figure 1: Modeling persistent behavior. See text.

At this point it seems that the situation has gotten worse. We started with one unknown parameter, a, and we now have two more: the fraction who will try again soon, s, and the fraction who will never try again, n. The fraction who will try again later ℓ is also unknown, but it can be computed by subtraction, since $\ell = 1 - s - n$.

However, somewhat like the old joke about turning a cold into pneumonia because there are ways to treat pneumonia, this more difficult situation actually opens new pathways for

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

solution. The key here is to consider the people who have been captured a *second* time, the people described as C_2, and ask what happens to them after they've been returned across the border.

The possibilities are shown in the third part of the diagram. This group of people, C_2, might try again never, or later, or sooner. Now we introduce some specific assumptions.

1. First, we assume that people who are going to be discouraged by what happened to them and never try again were already discouraged first time they were caught. This means that the group of people who've been caught twice does not any longer contain people people who will be discouraged. Therefore, these people will try again, either sooner or later. That is $N_3 = 0$.

2. Second, We will assume that people who tried "soon" to enter the country again after the first attempt will also try "soon" after the second. In other words, $L_3 = 0$.

3. With these two assumptions we have effectively posited that after one round, all the people who will try later, or will give up, have already been identified (because they tried again "soon") or have been weeded out. This is our third assumption: that the people who persist will not convert, and become of some other type.

With this set of assumptions, we expect that $L_k = N_k = 0$ for all $k \geq 3$. With these three assumptions about the behavior of the people who were captured twice, (in the time period we call "soon") we can now estimate the probability of apprehension. Since they do not change their behavior with additional recaptures, the same principle can be applied to their further recaptures. In other words, starting with the third part of the diagram, Figure 1 we assume that there are no people left who will never try again, and there also are no people left who will try again "later," then we can estimate the unknown parameter a as C_3 divided by C_2.

By the same argument we can also estimate that a will be given by C_4 divided by C_3, C_5 divided by C_4, and so forth. We are now claiming that several different ratios should be the

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

same. This claim enables us to *test the model* by testing whether these ratios are in fact the same. A standard statistical way to do that is to use chi-squared test and ask whether each of the ratios is equal to the weighted average ratio. The weighted estimate is $\hat{a}$, as shown in Equation 1. In this equation K is the number of time periods for which data are collected.

$$\hat{a} = \frac{\sum_{k\geq 3}^{K} C_k}{\sum_{k\geq 2}^{K-1} C_k} \tag{1}$$

2 Some Possible Extensions of the Model

2.1 Demographics

This discussion averages over all demographics. This makes the estimate of $\hat{a}$ less precise, because there will be differences in behavior, and there may be differences in a across groups. The same is true for n and ℓ. Possibly expert input and a Delphi process could estimate those. In another report we propose methods to elicit and or compute such estimates from human experts, and/or computer simulations. A discussion of this process is presented in another Technical Report, available at http://kantor.comminfo.rutgers.edu/MissedDetections/.

2.2 The meaning of "soon," and other generalizations

To apply the proposed model, we will have to decide what we mean by "soon" and "later." There are several ways to approach this. (a) solicit expert input; (b) use a one-parameter model (exponential, or uniform from zero to some maximum time, r_{wait}) for the distribution of time until the next try. With an exponential model, the probability of not having tried again by time t falls as $e^{-\lambda t}$; (c) use a two parameter model, such as a shifted exponential, or displaced uniform distribution. Alternative (b) admits analytic solutions. Alternative (c) may require non-linear data fitting, but is accessible with current technologies. Note that by varying the definition of the time bin that is counted as "soon," one may probe the observed distribution of time until retry.

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

2.3 Mixtures of persons

As we have discussed elsewhere, another very plausible model for being discouraged is that each experience of being captured[1] results in a non-zero, and constant probability of "giving up." This model does not permit resolution of the product $a(1-n)$ into its factors. Suppose, however, that the population is a mixture (in unknown proportions) of two kinds of persons: those who are deterred once and for all, and those who are incrementally deterred (at a constant rate) by the experience of repeated recapture.

This means is that some of the people (but we do not know how many) will behave as we have described in the model proposed here: once apprehended they will never try again. The other people are such that a fixed fraction of them d, are deterred each time they are captured, not only at first capture.[2] Perhaps surprisingly, because the two populations attenuate differently over time, the mixture can be resolved by numerical methods. Non-linear regression with several parameters, including the allocation of the population between these two types, can reveal how many people are of each type, and also find the apprehension probability $\hat{a}$. In some numerical experiments, not presented here, we have verified that the parameters can be found using the Excel Solver add-in, if the data are not too noisy.

2.4 Weighted expert opinion

We have mentioned that experts may be asked to estimate the parameters of specific demographics (that is, the n, ℓ values for that demographic. Of course we cannot know which experts to trust. However, we might also ask each expert (or simulation model) to predict *how soon* people will try again. The resulting numbers are in (unknown) proportion to the observed ratios of apprehensions. Using s_t to represent the number of persons captured in the t-th subinterval, we can compare the proportions to the cor-

[1] and being exposed to essentially the same "consequence" at each capture

[2] We have been told that there is an FOUO report which elaborates on the fact that with constant probabilities of apprehension and quitting, there is only one observable parameter, $z = a(1-n)$. One may of course illustrate that fact using various choices of a, n resulting in the same value of z

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

responding part of the observed data. This is prediction of $s_1 : s_2 : ...$ with a portion $(1-a)S+N+L$ not observed, but inferable from the known number of persons tagged and released. This method is discussed in detail in another report, available at the project web site, http://kantor.comminfo.rutgers.edu/MissedDetections/.

3 The Continuous Case: Additional Possibilities

There are a number of ways this model can be strengthened if the assumptions made here are not confirmed by the data. However, following the principle attributed to Albert Einstein[3], we should test the simplest possible models first. Some possible extensions are mentioned briefly here, but are not recommended unless the data force us to consider them,

3.1 A two-dimensional treatment

Once we recognize that time is continuous, and impose batches on the time between apprehensions, we can do a two-dimensional analysis. One dimension is the number of times the person has been apprehended, and the other is the amount of time between apprehensions.

It is particularly promising that models in which the apprehension rate and the rate of deterrence both depend on *both* the number of recaptures, and on how long a person takes to retry can be made tractable . What matters is the comparison between the number of data points (number of recapture bins)×(number of time bins for re-entry) compared to the number of parameters of the model. Thus, with 3 time bins and 5 recaptures there are 15 data points. In contrast, models of the dependence of apprehension and quitting on those two variables may have only six to eight parameters; that is, three or four parameters to the model of each factor influencing the decision to re-enter.

[3] "A theory should be as simple as possible, but not any simpler"

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

3.2 Does a increase with reapprehension?

We have mentioned that the χ^2 test can be used to check whether the probability of apprehension, a, is constant with repeated apprehensions. Another simple possibility is that a increases with repetitions. This argument may be called a "Darwin" effect. The people who are less likely to be apprehended become an ever smaller portion of the ones who are caught, people are apprehended again. Thus the average probability of being apprehended, across those who are repeatedly apprehended, will gradually rise.

One way to model the Darwinian selection is to assume that the population contains individuals drawn at random from a uniform distribution. The average value of the rate of apprehension will be the average of the upper and lower limits of this distribution. It is convenient to model apprehension rates less than 50% as being drawn from a uniform distribution between 0 and an upper limit $h = 2a$. With each apprehension this distribution is modified by a multiplicative factor representing the apprehension rate. It is easily shown that the resulting mean apprehension rate for the population that is just now captured for the $k - th$ time is given by Equation 2

$$a_k = \frac{\int_0^h x^k dx}{\int_0^h x^{k-1} dx} \tag{2}$$

The integrals are readily computed, yielding .

$$a(k) = \frac{k}{k+1} h \tag{3}$$

It is apparent that repeated recaptures gradually refines the group until it contains only those who are least adept at avoiding apprehension. However, if apprehensions do increase in this way, it will lend some credibility to both the original estimate, and the model itself.

When the average apprehension rate is greater than 50%, this model will not work. Setting $\ell = 2a - 1$ we may write Equation 4.

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

$$a_k = \frac{\int_\ell^1 x^k dx}{\int_\ell^1 x^{k-1} dx} \tag{4}$$

$$= \frac{k}{k+1} \frac{1-\ell^{k+1}}{1-\ell^k} \tag{5}$$

As the number of recaptures increases, the remaining persons have decreasing "fitness" at getting across the border, and the apprehension rate rises. The two models can be patched together in various ways to cover the whole range, or the prior distribution can be taken as a Beta function, which retains its form under the effect of repeated captures.

4 From Concept to Application

The actual computations will be somewhat more complicated. Each person has intervals between his successive apprehensions. In the conceptual picture of Figure 1, we compute as if all are held and released on the same date, at the end of each time period.[4]

In fact, each person will have individual start and capture times, and the "clock" for each person restarts with each release. Thus people who are released months apart can be aggregated for the analysis. The methods of dealing with the data are relatively elementary, but it is difficult to describe them using equations. In Appendix A we provide details of the process.

5 Discussion and Conclusions

We have discussed a particular model of behavior that supports both estimation of the apprehension rate and rigorous test of the validity of the model itself. The model is extremely simple, and perhaps one cannot reasonably expect it to work for very many combinations

[4]We have been told that current analysis aggregates people into fiscal years. If true, this means that persons apprehended "in the next year" might be 12 months or a few days from their previous apprehensions. Such data are of little use for the rigorous analysis proposed here.

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

of the demographic and the consequences to which people are exposed. However, it is not difficult to sketch a path towards more realistic models, which can be systematically explored. The existence of a simulation model and data generator makes it easy to formulate a model, generate data, and then test whether the analytic methods proposed here yield results sufficiently close to the selected parameters of the model.

Most immediately, simulated data may be used as a tool to assess the distribution of the random variable "time until trying to cross the border again." This can be done by returning to the original data, and analyzing separately the cases for which the interval between successive apprehensions falls into each of several bins. In the example data set described in A, reasonable bins would be $[10-19]; [20-29]; [30-39]; [40-\infty)$. The analysis should to show that the product of the apprehension rate and the fraction returning in each such bin is constant across successive reapprehensions.

To support this kind of exploration, we have developed a package for simulating data B. This contains an option to write the data out to a file, so that the data can be analyzed in various ways, to look for deviation from the assumptions of the basic model presented here.

Acknowledgments. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. We thank Dennis Egan for ongoing leadership of this research project, and Dr. Marc Rosenblum and his colleagues at the Office of Immigration Statistics for helpful conversations.

Appendices

A Generating Simulated Data

To generate simulated data, we can take a large time span (say one year) and generate "dates of first release" uniformly across that range. Each "first release" will have a unique identifying number NID, corresponding to the biometric unique identification of persons. At the time of release, each person will, with probability n be deterred forever, and removed from the sample. The program accepts two more parameters *rewait*, and probability of apprehension, a. The former is the upper bound of a uniform distribution of the time until next attempt. We add[5] a parameter τ characterizing the time between apprehension and release. In this example we take $\tau = 10$ days.

For analysis the data must be processed through a set of steps. Data will typically be first presented in time order. The events of interest are, for each individual: the time of each successive release, and the time of the next recapture. [We cannot exclude the possibility that the individual has entered and left the country between those two events, but in this initial analysis we ignore that possibility. It can, in principle, also be modeled.] After some preprocessing, the data may be brought to a format such as shown in Table 1

Recapture	Time	Person ID	Delta time
9	255	3	13
10	280	3	25
1	105	4	38
2	128	4	23
...	...	...	...

Table 1: Selected from a file of simulated data for 100 persons with specified random parameters. See text.

The data in Table 1 were generated using the following parameters:

```
tau=10;   # time between apprehension and release; set at 10 days. could be randomized
app=0.8;  # apprehension probability
```

[5]This parameter is only needed if the information on apprehensions does not include the time of release. If that data is available, the time until recapture can be computed directly from the data.

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

```
pquit=.4  # probability to be permanently deterred
tmax=365; # data collection extends over 365 days
rewait=30; # max time from release to next attempt: Uniform [1 to rewait]
nPersons=100;  #
```

The analysis is straightforward. The results may be summarized in a Table such as Table 2. The analysis is for 100 simulated persons. In each row the data are:

1. Recapture: the number of times a person has been recaptured;
2. The number of persons recaptured this number of times;
3. The apprehension rate estimated by comparing this number of recaptures to the number of persons recaptured one less time.
4. The predicted number of persons who would be recaptured this number of times, based on the aggregated estimate of the apprehension rate, Equation 1.
5. The contribution of this data element to the total calculation of chi-squared.

Recapture	100	est-app	predicted	part-chi-sq
1	54	54.0 %		
2	45	83.3 %	44.7	0.0019
3	37	82.2 %	37.3	0.0018
4	33	89.2 %	30.6	0.1829
5	32	97.0 %	27.3	0.8012
6	27	84.4 %	26.5	0.0097
7	22	81.5 %	22.4	0.0056
8	16	72.7 %	18.2	0.2692
9	13	81.3 %	13.2	0.0046
10	11	84.6 %	10.8	0.0052
11	8	72.7 %	9.1	0.1346
12	7	87.5 %	6.6	0.0214
13	3	42.9 %	5.8	1.3484
14	1	33.3 %	2.5	0.8864
Denom	308		deg freed	12
Numer	255	82.8 %	chi-sqd	3.6728

Table 2: Example computation of χ^2 for simulated data. See text.

In Table 2 the expressions "Denom." and "Numer" refer to the fraction of Equation 1.

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

The probability of a chi-squared statistic exceeding this value (3.7) for 12 degrees of freedom is *very* large (98.86%); while this value does not have an interpretation as a confidence, it indicates that the data truly give no reason to suppose that the apprehension rate is not constant, and has the estimated value near 82.8%. One may also compute a nominal confidence interval, for example, based on the variance of the individual estimates, but it may be simpler[6] to impute the binomial standard error $\sqrt{a(1-a)/N}$.

B The MRA Codes

We have developed a package called mra, to assist investigation of these ideas. We may think of this as an acronym for "Multiple Recapture Analysis." The code can be used in several ways.

1. As a tool to generate simulated data files with specific delay before release, probability of deterrence at first apprehension, and probability of apprehension.

2. As a tool to analyze an existing data file, using the model described here

3. As a package for exploring the impact of randomness and/or specific variations of the parameters of the model

In the third mode, the program generates data, and analyzes it on the fly, which avoids input/output delays and does not use very much storage. The program options include a variable that controls the random seed. When the program is run repeatedly with the same seed, the results should be the same.

Discussion. The parameters of the MRA code are shown in Table 3. Dtau is now set to be a fixed time. It would not be difficult to modify the source code so that it is generated according to some specified random distribution.

The number of persons, DnPersons=, should be set fairly large, as in the default setting of $1,000$, to provide good statistics on the simulated data. For any particular border station

[6] This issue merits for further statistical examination, as the successive observations are not independent. Nonlinear fit of a geometric model may be a good approach.

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

Parameter	Meaning	Default Value
-DinFile=	Name of the input file	nul
-DoutFile=	Name of the ouptut file	nul
-Dtau=	Fixed days until release	10
-Dapp=	Probability of apprehension	0.8
-Drewait=	Maximum number of days in uniform distribution of time until next attempt to enter	30 (days)
-Dpquit=	Probability of quitting	0.4
-DnPersons=	Number of persons to simulate	1000
-Dseed	Seed for the random number generator	1
-Dtmax=	Maximum number of days during which data are simulated.	365 (days)

Table 3: Parameters of the MRA code. See text.

it might take quite some time to accumulate such a large number of persons. We note that little is known about whether people try the same route again. As a first cut, it may make sense to apply this model to data on captures at any port of entry, or point between POEs. This will capture information on people whether they persist via the same path, or change their mode of attempted entry.

The random variable controlling the time until a person tries to enter the country again is set to be uniform, with the default value ranging from 1 to 30 days. As with the interval between capture and release, it would not be difficult to modify the source code to use another distribution, such as the exponential distribution.

The length of the simulation, Dtmax=, turns out to have a subtle relation to the analysis of the simulated data. If the data collection period is too short compared to the delay before a renewed attempt to enter, the data gathering will be incomplete at larger numbers of recaptures, and that will be reflected in variation of the estimated apprehension rate, and a large value of the χ^2 statistic.

Some general notes on usage are given in Exhibit 1.

Exhibit 1

```
# Usage:
#  run.sh [options]
# Options:
# Random number generator seed:
```

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

```
#   -Dseed=1
# Simulation parameters:
#   -Dtau=10 -Dapp=0.8 -Dpquit=0.4 -Dtmax=365 -Drewait=30 -DnPersons=1000
# Output file (optional):
#   -DoutFile=events.dat
# Input file (instead of built-in simulation), same format as input file
#   -DinFile=events.dat
#------------------------------------------------------------
# For example
#   ./run.sh -Dtau=50 -Dapp=0.9 -Dseed=3
#   ./run.sh -Dtau=50 -Dapp=0.9 -Dseed=3 -DoutFile=tmp.dat
#   ./run.sh -DinFile=tmp.dat
#------------------------------------------------------------

java -classpath lib/commons-math3-3.6.1.jar:classes $@edu.rutgers.mra.MakeReApprehension
```

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH Subcontract: R-17-0050 The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.