Plant-Capture Methods for Estimating Population Size from Uncertain Plant Captures

Yiran Wang, Martin Lysy, Audrey Béliveau

University of Waterloo

Oct 27, 2023

Motivation: Plant-Capture for Point-in-Time Street Surveys

- Goal: Estimate homeless population size
- Plants are instructed to dress and act as if they were homeless, then "mix" with the homeless population.
- Enumerators count how many homeless they see from a distance (Capture without Identification).

Proportion of plants seen $\Longrightarrow$ Probability of being captured

$$\frac{\text{Homeless Count}}{\text{Capture Probability}} \Longrightarrow \text{Homeless population size}$$

Assumptions of Current Methods

1. Closed population
2. The probability of being captured is constant and equal for plants and target population (no heterogeneity)
3. Enumerators' counts are accurate
4. Whether a plant is captured can be told with certainty

Assumptions of Current Methods

1. Closed population
2. The probability of being captured is constant and equal for plants and target population (no heterogeneity)
3. Enumerators' counts are accurate
4. Whether a plant is captured can be told with certainty

- E.g. Plants may be asked to answer whether they were seen by enumerators (Yes / Maybe / No).

Notation

Data

- Y: Total count of captured individuals (including plants)
- M^{yes}, M^{maybe}, M^{no}: Number of plants that are self-assessed as yes/maybe/no to having been captured

Constant

- M: Number of plants

Notation

Latent variables

- H^c: Number of captured individuals from the target population
- $M^{maybe,c}$: Number of plant who are uncertain but were captured

Note:

- $H^c + M^{maybe,c} = Y - M^{yes}$ is known

Notation

Parameters

- p^{yes},p^{maybe}, p^{no}: Probability that a plant was self-assessed as "yes", "maybe" and "no" respectively
- $p^{c|maybe}$: Probability that a plant was captured given self-assessed as "maybe"
- p^{c}: Probability of being captured
- H: Size of target population

Model I

$$(M^{yes}, M^{maybe}, M^{no}) \mid M \sim Multinom(M; \; p^{yes}, p^{maybe}, p^{no}) \quad (1)$$

$$M^{maybe,c} \mid M^{maybe} \sim Binom(M^{maybe}, p^{c|maybe}) \quad (2)$$

$$H^{c} \sim Binom(H, p^{c}) \quad (3)$$

$$Y = M^{yes} + M^{maybe,c} + H^{c} \quad (4)$$

Assumptions

1. Plants in M^{yes} were captured and plants in M^{no} were not captured.
2. **A plant being self-assessed as "maybe"** is *independent* of **being captured by enumerators** :

$$p^{c|maybe} = p^c$$

Table 1: MAR assumption for Model I

	Captured	Not Captured
Maybe	$M^{maybe,c}$	$M^{maybe} - M^{maybe,c}$
Not Maybe	M^{yes}	M^{no}

Model I

Combining equations (2) and (3) into a single Binomial, we can rewrite the model as:

$$(M^{yes}, M^{maybe}, M^{no}) \mid M \sim Multinom(M;\ p^{yes}, p^{maybe}, p^{no}) \quad (1)$$

$$M^{maybe,c} + H^c \mid M^{maybe} \sim Binom(M^{maybe} + H, p^c) \quad (2^*)$$

$$Y = M^{yes} + M^{maybe,c} + H^c \quad (3^*)$$

where

$$p^{yes} = p^c(1 - p^{maybe}) \quad \text{and} \quad p^{no} = (1 - p^c)(1 - p^{maybe})$$

Including Identification in Model

In some surveys, the identity of the captured individuals may be obtained by a direct contact. For example, in the 1990 S-night survey (Martin, 1992; Laska & Meisner, 1993; Martin et al., 1997),

- Enumerators were instructed to interview all individuals encountered in the site, who were not in uniform and were not engaged in obvious money-making activities.
- People found sleeping or covered by sleeping bags or blankets were to be counted but not disturbed or interviewed

Therefore, we propose an alternative model which accounts for this situation.

New Notation

Data

- M^i: Number of identified plants
- H^i: Number of identified individuals from the target population

Parameters

- $p^{maybe|ni}$: Probability that a plant was self-assessed as "maybe" given not identified
- p^i: Probability that a plant was identified
- $p^{i|c}$: Probability that a plant was identified given captured by an enumerator

Model II

$$(M^{i},\ M^{yes},\ M^{maybe},\ M^{no}) \mid M \sim Multinom(M;\ \ p^{i},\ p^{yes},\ p^{maybe},\ p^{no}) \tag{5}$$

$$M^{maybe,c} \mid M^{maybe} \sim Binom(M^{maybe}, p^{c|maybe,ni}) \tag{6}$$

$$H^{c} \sim Binom(H, p^{c}) \tag{7}$$

$$H^{i} \sim Binom(H^{c}, p^{i|c}) \tag{8}$$

$$Y = M^{i} + M^{yes} + M^{maybe,c} + H^{c} \tag{9}$$

Assumptions

- Plants in M^{yes} were captured and plants in M^{no} were not captured. $p^{i|c}$ is constant for plants and target population.
- **Being captured by an enumerator** is *independent* of **being self-assessed as "maybe" for a plant** among the plants not identified

Table 2: MAR assumption for Model II

	Captured	Not Captured
Interviewed	M^i	
Maybe	$M^{maybe,c}$	$M^{maybe} - M^{maybe,c}$
Not Maybe	M^{yes}	M^{no}

Parameters

- $p^{i} = p^{c} p^{i|c}$
- $p^{yes} = p^{c}(1 - p^{i|c})(1 - p^{maybe|ni})$
- $p^{maybe} = p^{c}(1 - p^{i|c})p^{maybe|ni} + (1 - p^{c})p^{maybe|ni}$
- $p^{no} = (1 - p^{c})(1 - p^{maybe|ni})$
- $p^{c|maybe,ni} = \frac{p^{c}(1-p^{i|c})}{p^{c}(1-p^{i|c})+(1-p^{c})}$

Model III

So far, we have assumed no heterogeneity in the probability of being captured (Assumption 2).

However, in practice:

- visual barriers
- drug activities
- heard gunshots
- enumerators did not approach everyone or did not have enough time to complete the enumeration

If more than 50 percent of enumerators mentioned any of these problems, the site was classified as "hard"; otherwise, it was classified as "easy". It would be natural for p^c to be larger in easy sites.

Maximum Likelihood Estimation

All the proposed models in our study can be written in a general form

$$L(\boldsymbol{X}|\gamma) = \sum_{\boldsymbol{Z}\in\Omega} L(\boldsymbol{X},\boldsymbol{Z}|\gamma),$$

where

- γ: Parameters
- $\boldsymbol{X}$: Data
- $\boldsymbol{Z}$: Latent Variables

By summing over Z, we can remove the latent variable Z from the likelihood.

MCMC Algorithm by JAGS

- When the model complexity increases (e.g. large number of latent variables), marginalization could be infeasible.
- Probabilistic programming languages such as JAGS and NIMBLE have grown in popularity among practitioners, therefore it is desirable to implement our models using these languages.
- A simple solution for two latent variables with known sum is to use *dsum()* function in JAGS.
  - Alternative methods: Custom function and distribution in NIMBLE; Zeros/Ones trick

Simulation Study Setting

- We conducted simulation studies for 3 models in 2 scenarios:
  - Small Cities:
    15 Plants, 150 Homeless (Model I & II)
    30 Plants, 300 Homeless (Model III)
  - Large Cities: 100 Plants, 1,500 Homeless

  For each study, we simulated 1,000 datasets

- MCMC: 3 chains, 30,000 iterations and 15,000 burn-ins for each chain

- MLE: log transformation for H and logit transformation for p^c,$p^{i|c}$ $p^{maybe|ni}$

Priors for Bayesian Inference

- $H \sim TN_{[0,\infty]}(100, 200^2)$ (rounded, small cities)
  $H \sim TN_{[0,\infty]}(1000, 10000^2)$ (rounded, large cities)
- Model I:
  - $p^c \sim Unif(0,1)$
  - $p^{maybe} \sim Unif(0,1)$
- Model II & III:
  - $p^c \sim Unif(0,1)$
  - $p^{i|c} \sim Unif(0,1)$
  - $p^{maybe|ni} \sim Unif(0,1)$

Results

Table 3: Model I Results

Method	M	Parameter	True Value	Estimate	SD	RBias	RRMSE	CP
Bayesian	15	H	150	162	40	0.09	0.23	0.97
		p^c	0.7	0.67	0.12	-0.05	0.16	0.97
		p^{maybe}	0.2	0.23	0.10	0.14	0.49	0.98
MLE	15	H	150	149	31	-0.01	0.24	0.85
		p^c	0.7	0.73	0.12	0.04	0.19	0.98
		p^{maybe}	0.2	0.20	0.10	0.01	0.51	0.98
Bayesian	100	H	1,500	1,523	122	0.02	0.08	0.94
		p^c	0.7	0.70	0.05	-0.01	0.07	0.95
		p^{maybe}	0.2	0.20	0.04	0.02	0.20	0.94
MLE	100	H	1,500	1,497	114	-0.00	0.08	0.93
		p^c	0.7	0.70	0.05	0.01	0.07	0.95
		p^{maybe}	0.2	0.20	0.04	0.00	0.20	0.94

Results

Table 4: Model II Results

Method	M	Parameter	True Value	Estimate	SD	RBias	RRMSE	CP
Bayesian	15	H	150	162	36	0.08	0.22	0.96
		p^c	0.7	0.67	0.11	-0.05	0.16	0.97
		$p^{maybe\|ni}$	0.2	0.26	0.14	0.30	0.73	0.96
		$p^{i\|c}$	0.8	0.80	0.04	-0.01	0.05	0.95
MLE	15	H	150	150	29	0.00	0.22	0.88
		p^c	0.7	0.72	0.12	0.03	0.18	0.98
		$p^{maybe\|ni}$	0.2	0.21	0.13	0.04	0.83	0.96
		$p^{i\|c}$	0.8	0.80	0.04	-0.00	0.05	0.95
Bayesian	100	H	1,500	1,520	113	0.01	0.08	0.94
		p^c	0.7	0.69	0.05	-0.01	0.07	0.94
		$p^{maybe\|ni}$	0.2	0.21	0.06	0.04	0.29	0.96
		$p^{i\|c}$	0.8	0.80	0.01	0.00	0.02	0.96
MLE	100	H	1,500	1,498	107	-0.00	0.07	0.93
		p^c	0.7	0.70	0.05	0.01	0.07	0.94
		$p^{maybe\|ni}$	0.2	0.20	0.06	-0.00	0.30	0.97
		$p^{i\|c}$	0.8	0.80	0.01	0.00	0.02	0.96

Results

Table 5: Model III Results

Method	M	Parameter	True Value	Estimate	SD	RBias	RRMSE	CP
Bayesian	30	H	300	336	64	0.12	0.20	0.95
		p^c_{easy}	0.9	0.84	0.08	-0.06	0.09	0.94
		p^c_{hard}	0.4	0.39	0.12	-0.03	0.30	0.97
		$p^{maybe\|ni}$	0.2	0.22	0.10	0.09	0.49	0.97
		$p^{i\|c}$	0.8	0.80	0.03	-0.00	0.03	0.94
MLE	30	H	300	313	65	0.04	0.25	0.97
		p^c_{easy}	0.9	0.90	0.07	0.00	0.08	0.96
		p^c_{hard}	0.4	0.42	0.14	0.06	0.38	0.97
		$p^{maybe\|ni}$	0.2	0.19	0.10	-0.07	0.54	0.98
		$p^{i\|c}$	0.8	0.80	0.03	0.00	0.03	0.95
Bayesian	100	H	1,500	1,571	173	0.05	0.12	0.94
		p^c_{easy}	0.9	0.89	0.04	-0.01	0.05	0.95
		p^c_{hard}	0.4	0.39	0.08	-0.03	0.20	0.95
		$p^{maybe\|ni}$	0.2	0.21	0.06	0.04	0.29	0.95
		$p^{i\|c}$	0.8	0.80	0.01	-0.00	0.02	0.95
MLE	100	H	1,500	1,510	142	0.01	0.10	0.97
		p^c_{easy}	0.9	0.91	0.04	0.01	0.05	0.93
		p^c_{hard}	0.4	0.40	0.08	0.01	0.20	0.97
		$p^{maybe\|ni}$	0.2	0.20	0.06	-0.01	0.30	0.96
		$p^{i\|c}$	0.8	0.80	0.01	-0.00	0.02	0.95

S-Night Data Analysis

Table 6: S-Night Data from Literature

	Chicago	New Orleans	Phoenix	New York	Los Angeles
Plants	13	58	26	94	25
Interviewed	2	41	18	40	16
Yes	0	6	3	19	1
Maybe	5	5	1	13	2
No	6	6	4	22	6
Census	11	109	104	1240	217

Real Data Analysis (Model 2 without H^i)

Table 7: Real Data Results

	Bayesian			MLE		
Parameter	Estimate	SD	95% CrI	Estimate	SD	95% CI
Chicago						
H	63	68	(17, 270)	54	38	(13, 217)
p^c	0.15	0.10	(0.04, 0.41)	0.16	0.10	(0.04, 0.46)
$p^{maybe\|ni}$	0.46	0.13	(0.21, 0.72)	0.45	0.15	(0.20, 0.73)
$p^{i\|c}$	0.75	0.21	(0.25, 0.99)	1.00	0.00	(0.00, 1.00)
New Orleans						
H	71	7	(61, 87)	69	6	(58, 82)
p^c	0.84	0.05	(0.72, 0.93)	0.86	0.05	(0.73, 0.94)
$p^{maybe\|ni}$	0.31	0.10	(0.13, 0.54)	0.29	0.11	(0.13, 0.54)
$p^{i\|c}$	0.82	0.06	(0.70, 0.92)	0.83	0.06	(0.68, 0.91)
Phoenix						
H	103	12	(87, 135)	98	10	(80, 120)
p^c	0.80	0.08	(0.63, 0.92)	0.84	0.08	(0.64, 0.94)
$p^{maybe\|ni}$	00.18	0.12	(0.03, 0.48)	0.12	0.12	(0.02, 0.54)
$p^{i\|c}$	0.82	0.08	(0.63, 0.94)	0.84	0.08	(0.61, 0.94)
New York						
H	1715	137	(1500, 2039)	1688	131	(1450, 1964)
p^c	0.68	0.05	(0.58, 0.78)	0.70	0.05	(0.59, 0.79)
$p^{maybe\|ni}$	0.25	0.06	(0.15, 0.37)	0.24	0.06	(0.14, 0.37)
$p^{i\|c}$	0.61	0.06	(0.49, 0.73)	0.61	0.06	(0.48, 0.73)
Los Angeles						
H	287	40	(232, 388)	282	40	(215, 372)
p^c	0.70	0.09	(0.52, 0.85)	0.71	0.09	(0.50, 0.86)
$p^{maybe\|ni}$	0.26	0.13	(0.07, 0.56)	0.22	0.14	(0.06, 0.58)
$p^{i\|c}$	0.89	0.08	(0.69, 0.98)	0.92	0.07	(0.63, 0.99)

Conclusion

- We proposed a new framework for the plant-capture study, which allows for uncertain assessment of being captured and/or identified in the model and also considers heterogeneity.
- Two inference methods are proposed and evaluated using simulation study.
- Further investigation should be conducted for the applicability of our models in real-world scenarios, with a particular focus on assessing the validity of the independence assumption.