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

Yiran Wang

Joint work with Dr. Audrey Béliveau and Dr. Martin Lysy

Department of Biostatistics

Yale School of Public Health

April 10, 2025

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

- Goal: Estimate homeless population size from street counts
- Census count: Hard to reach everyone

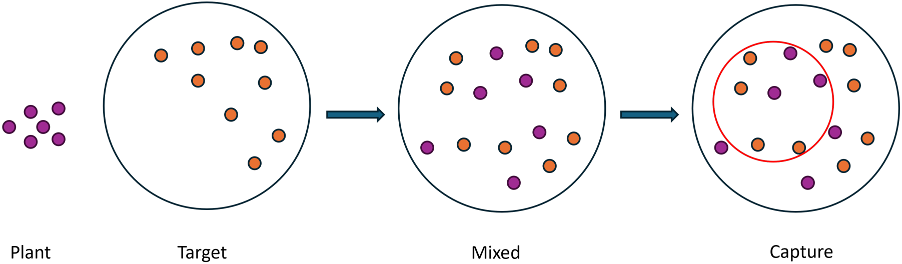
Motivation: Plant-Capture for Point-in-Time Street Surveys

- Goal: Estimate homeless population size from street counts
- Census count: Hard to reach everyone
- Capture-recapture: Mark effects, heterogeneous capture probability ...

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

- Goal: Estimate homeless population size from street counts
- Census count: Hard to reach everyone
- Capture-recapture: Mark effects, heterogeneous capture probability ...
- Plant-Capture:
  - To correct for imperfect detection, homeless decoys ("plants") are planted among the homeless
  - Plants are assumed indistinguishable from homeless individuals.

Plant-Capture Methods [1]



Uncertain Plant Captures

- Assumption: Whether a plant is captured can be assessed with certainty
  - Enumerators: Interview everyone
  - Plants: Answer questionnaire
    - Interviewed (Yes/ No)
    - If not interviewed, seen (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 homeless captured
- $M^{maybe,c}$: Number of plants 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 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 capture
- H: Size of homeless population (target of inference)

Basic Model (Model I)

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

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

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

$$Y - M^{yes} = M^{maybe,c} + H^{c} \tag{4}$$

... but we have more parameters than data points!

Assumptions

1. Plants in M^{yes} were captured and plants in M^{no} were not captured.
2. MAR: **A plant self-assessing 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}

Basic Model (Model I)

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

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

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

$$Y - M^{yes} = M^{maybe,c} + H^{c} \tag{4}$$

Basic Model (Model I)

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

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

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

$$Y - M^{yes} = M^{maybe,c} + H^{c} \tag{4}$$

Basic Model (Model I)

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

$$Y - M^{yes} \mid M^{maybe} \sim \text{Binom}(M^{maybe} + H, p^{c}) \qquad (2^*)$$

where

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

Including Identification in Model

- Enumerators are asked to interview all individuals encountered, but
  - Sleeping
  - Hiding
  - ...
- Therefore, only partial identification information is available

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

Identification Model (Model II)

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

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

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

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

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

...still, we have more parameters than data points!

Assumptions

- Plants in M^{yes} were captured and plants in M^{no} were not captured.
- MAR: Among the plants not identified, **being captured** is *independent* of **self-assessing as "maybe"**

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}

Parametrization

All probabilities can be expressed in terms of p^c, $p^{i|c}$ and $p^{maybe|ni}$:

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

Classification Model (Model III)

So far, we have assumed no heterogeneity in the probability of being captured.

However, in practice:

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

Sites presenting these issues may be classified as “hard”; or otherwise as “easy”. It would be natural for p^c to be larger in easy sites.

Bayesian Normal Approximation (BNA)

Posterior distribution can be written in a general form

$$\Pi(\gamma \mid \boldsymbol{X}) \propto \sum_{\boldsymbol{Z} \in \Omega} L(\boldsymbol{X}, \boldsymbol{Z} \mid \gamma) \Pi(\gamma),$$

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 [2] and NIMBLE [3] 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

Uncertainty Propagation (UP)

- We propose an alternative way to obtain an approximate posterior inference for H.
  - (1) Obtain posterior samples of H^c and p^c.
  - (2) If H^c and p^c are observed, we can construct an approximate representation of the posterior distribution $\pi(H \mid p^c, H^c)$:

$$H|H^c, p^c \sim N\left(\hat{H}_0, \hat{H}_0(1-p^c)/p^c\right),$$

where $\hat{H}_0 = H^c/p^c$ is a variant of the MLE for H [4]. And $\hat{H}_0$ has variance

$$\mathrm{Var}(\hat{H}_0) = \frac{\mathrm{Var}(H^c)}{(p^c)^2} = \frac{H(1-p^c)}{p^c},$$

which is estimated by $\hat{H}_0(1-p^c)/p^c$.

Simulation Study Setting

- We conducted simulation studies in 2 scenarios:
  - Small Cities: 15 Plants, 150 Homeless
  - Large Cities: 100 Plants, 1500 Homeless

  For each study, we simulated 1,000 datasets

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

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

Priors for Bayesian Inference

- $\log(H) \sim N(0, 100)$
- Model I:
  - $p^c \sim Unif(0,1)$
  - $p^{maybe} \sim Unif(0,1)$
- Model II:
  - $p^c \sim Unif(0,1)$
  - $p^{i|c} \sim Unif(0,1)$
  - $p^{maybe|ni} \sim Unif(0,1)$

Results

Table 3: Simulation Results for H

Model	Method	M	True Value	Estimate	RBias	RRMSE	CP
Model I	BNA	15	150	149	-0.01	0.18	0.91
	MCMC			159	0.06	0.24	0.97
	BNA	100	1,500	1,498	0.00	0.08	0.94
	MCMC			1,513	0.01	0.08	0.94
Model II	BNA	15	150	150	0.00	0.17	0.92
	MCMC			159	0.06	0.24	0.97
	UP			159	0.06	0.22	0.97
	BNA	100	1,500	1,500	0.00	0.07	0.93
	MCMC			1,513	0.01	0.07	0.93
	UP			1,512	0.01	0.07	0.94

Shelter and Street Night Enumeration (S-Night, 1990) [5, 6]

Table 4: 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 II without H^i)

Methods	Metrics	Chicago	New Orleans	Phoenix	New York	Los Angeles
	Estimate	24	69	100	1,688	281
BNA	SD	14	6	10	129	37
	95% CrI	(8, 74)	(59, 82)	(82, 121)	(1,453, 1,962)	(217, 364)
	Estimate	37	70	102	1,709	290
MCMC	SD	68	7	12	142	47
	95% CrI	(11, 156)	(61, 87)	(87, 135)	(1,494, 2,056)	(233, 415)
	Estimate	39	70	102	1,704	291
UP	SD	53	7	12	137	46
	95% CrI	(11, 178)	(61, 86)	(87, 134)	(1,490, 2,026)	(233, 412)

Conclusion

- We proposed a new framework for the plant-capture study, which allows for uncertain assessment of capture, partial identifications and heterogeneity between sites.
- Various inference methods are proposed and evaluated through a 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 MAR assumptions.

Acknowledgements

- Dr. Ruth King (The University of Edinburgh)
- CANSSI Collaborative Research Team Grant: Modern Techniques for Survey Sampling and Complex Data
- The Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants RGPIN-2020-04364, RGPIN-2019-04404

Reference I

[1] EM Laska and M Meisner. "A plant–capture method for estimating the size of a population from a single sample". In: *Biometrics* 49.1 (1993), pp. 209–220.

[2] M Plummer. "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". In: *Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria. 2003, pp. 1–10.

[3] P de Valpine et al. "Programming with models: writing statistical algorithms for general model structures with NIMBLE". In: *Journal of Computational and Graphical Statistics* 26 (2017), pp. 403–417. DOI: 10.1080/10618600.2016.1172487.

[4] AL Rukhin. "Statistical decision about the total number of observable objects". In: *Sankhyā: The Indian Journal of Statistics, Series A* 37.4 (1975), pp. 514–522.

Reference II

[5] E Martin. "Assessment of S-Night street enumeration in the 1990 Census". In: *Evaluation review* 16.4 (1992), pp. 418–438.

[6] E Martin et al. "Issues in the use of a plant–capture method for estimating the size of the street dwelling population". In: *Journal of Official Statistics* 13.1 (1997), p. 59.

Thank You

Model III

Assume we have K classes for sites, and p_k^c could be different among these classes.

$$(M_k^i,\ M_k^{yes},\ M_k^{maybe},\ M_k^{no}) \mid M_k \sim \text{Multinom}(M_k; p_k^i, p_k^{yes}, p_k^{maybe}, p_k^{no}) \tag{10}$$

$$Y_k = M_k^i + M_k^{yes} + M_k^{maybe,c} + H_k^c \tag{11}$$

$$M_k^{maybe,c} \mid M_k^{maybe} \sim \text{Binom}(M_k^{maybe}, p_k^{c|maybe,ni}) \tag{12}$$

$$H_k^c \sim \text{Binom}(H_k, p_k^c) \tag{13}$$

$$H_k^i \sim \text{Binom}(H_k^c, p^{i|c}) \tag{14}$$

Bayesian Normal Approximation (BNA)

For example, in Model II,

$$\gamma = (H, p^c, p^{i|c}, p^{maybe|ni})$$

$$X = Y, M, M^i, M^{yes}, M^{maybe}, H^i$$

$$Z = M^{maybe,c}$$

The posterior is

$$\Pi(Y, M, M^i, M^{yes}, M^{maybe}, H^i \mid H, p^c, p^{i|c}, p^{maybe|ni}) \propto$$
$$\Pi(H, p^c, p^{i|c}, p^{maybe|ni}) \times$$
$$\sum_{M^{maybe,c} = a_{II}}^{b_{II}} L(Y, M, M^i, M^{yes}, M^{maybe}, H^i, M^{maybe,c} \mid H, p^c, p^{i|c}, p^{maybe|ni})$$

Determining a and b for $M^{maybe,c}$

For Model II, the bounds of $M^{maybe,c}$ are based on the following constraints:

1. $$0 \leq M^{maybe,c} \leq M^{maybe}$$

2. $$M^{maybe,c} = Y - M^{i} - M^{yes} - H^{c}$$
$$H^{i} \leq H^{c} \leq H$$
$$\Downarrow$$
$$Y - M^{i} - M^{yes} - H \leq M^{maybe,c} \leq Y - M^{i} - M^{yes} - H^{i}$$

Therefore, we have

$$\max(0, Y - M^{i} - M^{yes} - H) = a_{II} \leq b_{II} = \min(M^{maybe}, Y - M^{i} - M^{yes} - H^{i})$$