Data Source

Select source

- ○ REDCap API
- ◉ Upload file

Layout

Only Layered (Tidy wave) is available.

- ☑ Reduce overlap with small jitter

Jitter amount (x-axis)

0.15

0.00 0.50

- ☑ Show edges

Edge opacity

0.60

0.05 1.00

Explanation of Cleaning and Estimation Options

Cleaning options (network size adjustments)

- **Fix underreported network size**
  Occasionally participants report a personal network size (degree) smaller than the number of recruits they actually brought into the study.
  For example, if someone recruited 3 peers but reported a network size of 2, this is underreported *Fix underreport* replaces such underreported cases with the observed value from the recruitment tree (out-degree + 1 for non-seeds; out-degree for seeds).
- **Impute median for NA and 0**
  If a participant reports their network size as `NA` (missing) or `0`, the value is invalid for weighting because RDS estimators depend on network size.
  This option replaces `NA` or `0` with a user-specified value, **median of the current network size distribution** (recommended).
- **Set cap**
  Sometimes individuals report extremely large network sizes (e.g., 500), which can disproportionately affect estimates.
  Set cap limits reported values at a user-specified maximum (often the **75th percentile** of the distribution is recommended).
  This reduces the influence of outliers while preserving most of the data.

Estimation methods (weights and hidden size)

- **Gile's Successive Sampling (SS) Weights**
  In Respondent-Driven Sampling (RDS), participants with larger personal networks are more likely to be sampled.
  Gile's SS estimator corrects for this by modeling recruitment as successive draws *without replacement* from a finite population.
  The inclusion probability for each participant depends on:
  - Their reported personal network size
  - The total sample size
  - The assumed prior population size in that region `N` (user-specified)
    Weights are the inverse of these inclusion probabilities, allowing less biased population estimates.
- **SSPSE (Successive Sampling Population Size Estimation)**
  SSPSE builds on the same successive sampling model but aims to estimate the *total hidden population size*.
  It compares the observed distribution of reported network sizes with expected distributions under different candidate population sizes.
  Using Bayesian inference, SSPSE generates a **posterior distribution** for the population size, summarized by mean, median, mode, and credible intervals.
  The output includes both a summary table (prior vs posterior) and a posterior density plot.
  RDS adjustment and SSPSE could only be applied at site-level, ideally with sample sizes >200 for stable estimation (<100 not recommended).

REDCap RDS Tree Automata

Upload CSV/TSV/XLSX

Drag and drop file here
Limit 200MB per file • CSV, TSV, XLSX
Browse files

msmdata app trial.xlsx 1.7MB

Preview

	couponcode	id	image_codecoupon	h0_1	h0_2	h0_3	h0_4	h0_5	h0_5b	h0_6	h0_7	h0_8	h0_8b	h0_9___0	h0_9___1	h0_9___2	h0_9__
0	None	11	None	30	0	30	0	0	None	0	1	0	None	1	0	0	
1	None	12	1710937881198.jpg	18	0	216	0	0	None	0	1	3	None	1	1	0	
2	150100	13	None	35	0	150	0	0	None	0	1	0	None	0	1	1	
3	150101	14	None	35	0	11	0	0	None	0	1	3	None	1	1	1	
4	150102	15	None	35	0	45	0	0	None	0	1	0	None	1	0	0	

Incoming coupon field

couponcode

Seed field

seed.id

Network size field

h3_14

Recruitment out fields

outpon1 × outpon2 × outpon3 ×

Use this uploaded file

Add site-level recruitment

Leading digits length for site code

2

Site prefix digits (e.g., 150):

15

Site name (e.g., Site A):

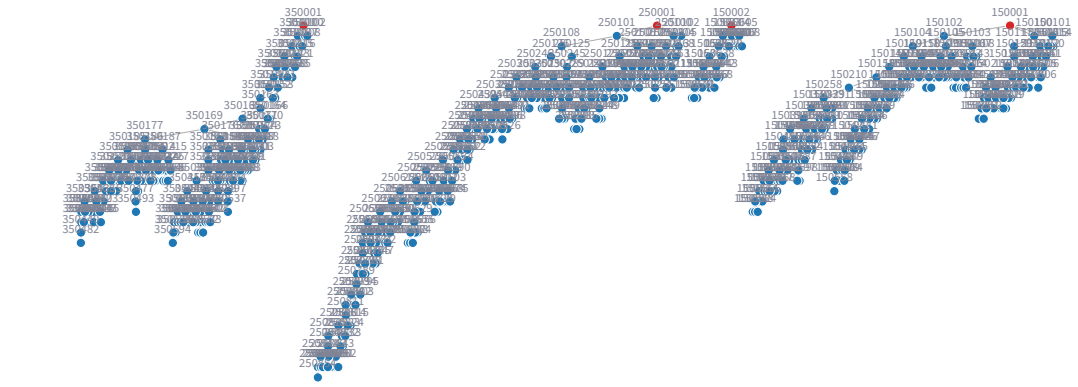
Dakar

Draw

Focus on seed (optional)

<All seeds>

Full Recruitment (Layered · Tidy)



Download full_tree_tidy (HTML)

Underreported count: **79**

Underreported coupon IDs: 150505, 350106, 150155, 250370, 350101, 250325, 250471, 150291, 250100, 350133, 250246, 150611, 150145, 150248, 150121, 250320, 150212, 250192, 250616, 150255, 150104, 350137, 250702, 150157, 150211, 150405, 250380, 150119, 150178, 350205, 150506, 150536, 250129, 250466, 250439, 150118, 150612, 150166, 250478, 150146, 250552, 150394, 250108, 250326, 250216, 150606, 350417, 250589, 250130, 350554, 150421, 250665, 350345, 350263, 250610, 350327, 150163, 350273, 150332, 150338, 350163, 350193, 150218, 150334, 150512, 150158, 350187, 150177, 150142, 250559, 250178, 150651, 150274, 150472, 250804, 250398, 250114, 150167, 250660

Fix underreported networksize

Percentiles of reported_networksize

	reported_networksize
0%	0
25%	3
50%	6
75%	15
100%	450

Impute NA and 0

Imputation value for NA/0

7.00

Median networksize (current): 7.00

Set networksize cap:

1

Apply cap to reported_networksize

Export cleaned networksize (current view)

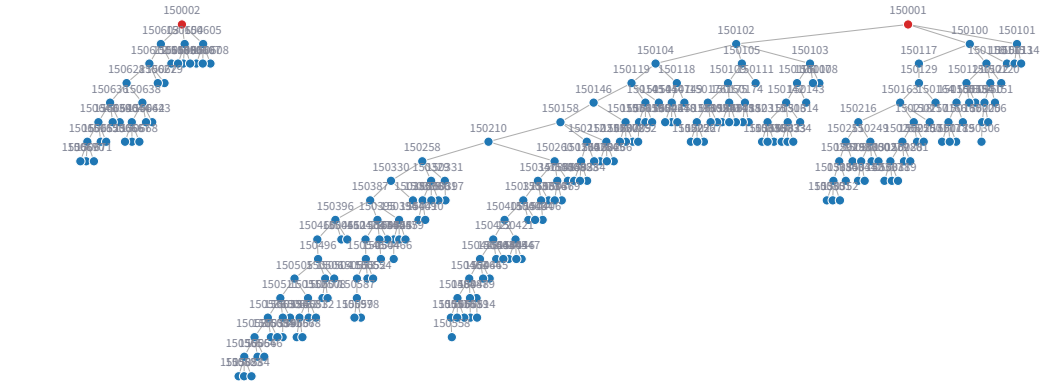
Site-Level Trees

Dakar

[Dakar] Focus on seed

<All seeds>

Tree: Dakar (Layered · Tidy)



Download Dakar_tree_tidy (HTML)

- Participants: 241
- Seeds: 2
- Max wave: 18

Underreported count: **37**

Underreported coupon IDs: 150163, 150212, 150505, 150166, 150332, 150146, 150155, 150338, 150255, 150104, 150218, 150334, 150512, 150158, 150157, 150291, 150394, 150177, 150211, 150405, 150142, 150119, 150178, 150651, 150506, 150606, 150274, 150472, 150536, 150421, 150118, 150167, 150611, 150612, 150145, 150248, 150121

[Dakar] Fix underreported networksize

Percentiles of reported_networksize

	reported_networksize
0%	0
25%	2
50%	6
75%	15
100%	165

[Dakar] Impute NA and 0

[Dakar] Imputation value for NA/0

6.00

[Dakar] Median networksize (current): 6.00

[Dakar] Set networksize cap:

15

[Dakar] Apply cap to reported_networksize

[Dakar] Export cleaned networksize

[Dakar] Population estimate:

2000

[Dakar] Run Gile's Weights

[Dakar] Run SSPSE

Delete Dakar

Generate Research Report

- [x] Include Full-Tree section in report
- [x] Include Site-level sections in report (if available)
- [x] Compute Weights & SSPSE during report generation (if not previously saved)

Generate PDF Report

⚠️**Note:** SS-PSE and Gile's SS weights are **not recommended** for full pooled datasets. For valid estimation, these methods should be applied at the **site level**, not the combined dataset.

Model Fit Guide (Good vs Bad)

📘How to Evaluate SS-PSE Model Fit

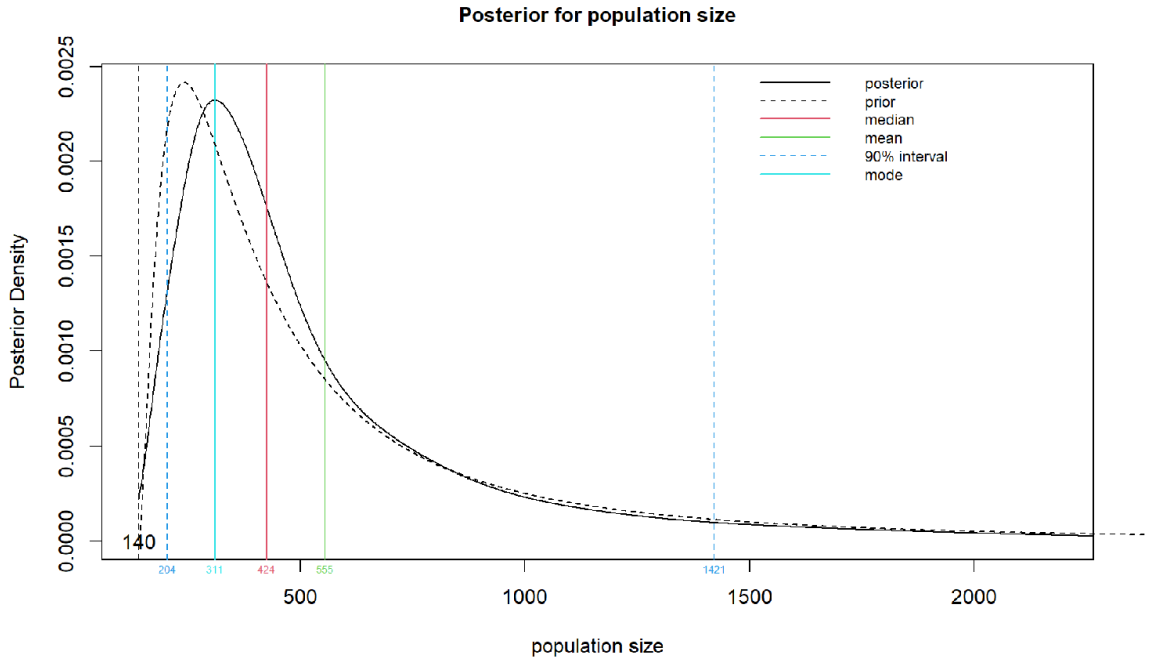
The figures below explain how to evaluate whether the **posterior** (solid curve) and **prior** (dashed curve) indicate a *good* or *bad* SS-PSE model fit.

A good model fit should show:

- ✓ Prior and posterior curves overlap **moderately**
- ✓ Posterior is **not wildly different** from prior
- ✓ Posterior median stays within the high-density posterior region
- ✓ Posterior mode shifts *somewhat*, but not drastically
- ❌ Posterior should **NOT** completely ignore the prior
- ❌ Posterior should **NOT** be many orders of magnitude away

If the curves diverge strongly → **your prior assumption is likely incorrect**,
and you should adjust the median prior and rerun SS-PSE.

✅Example of GOOD fit (GUA_MT)

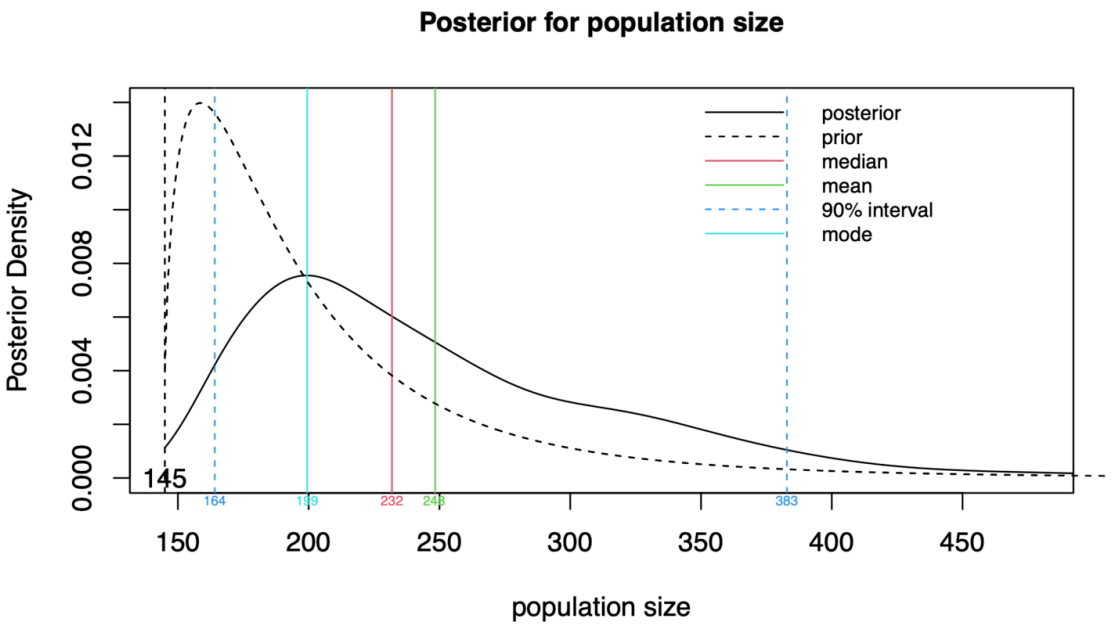


Good fit example: Prior and posterior curves overlap well.

Why this is a GOOD fit?

- Posterior curve shifts but **still overlaps** with the prior.
- Prior information meaningfully contributes to the posterior.
- Sample visibility and recruitment patterns are consistent with the prior belief.

❌Example of POOR fit (GUA_UDI)



Bad fit example: Prior and posterior curves do NOT overlap.

Why this is a BAD fit?

- Prior (dashed) and posterior (solid) almost **do not overlap**.
- Posterior strongly diverges → **the median prior is likely incorrect**.
- The sample's visibility and recruitment dynamics contradict prior belief.

📌**In this case, adjust your prior and rerun SS-PSE**
Aim for a curve that moderately overlaps with the prior
(similar to the GOOD example above).