Advances in Survey Estimation with Imperfectly Matched Auxiliary Data

Jay Breidt
Colorado State University

6th Italian Conference on Survey Methodology
June 7, 2019

Joint work with Chien-Min Huang and Jean Opsomer
Colorado State University and Westat

Motivation: Recreational fishing in South Carolina

- About 450 charter boats and 15,000 boat trips along the Atlantic Coast of South Carolina each year
- How many black sea bass were caught in 2018?

$$\begin{aligned} U &= \{1, 2, \ldots, N\} \\ &= \{\text{all SC charter fishing boat trips in 2018}\} \end{aligned}$$

  – number of black sea bass caught on kth trip: y_k
  – total black sea bass caught: $T = \sum_{k \in U} y_k$
- Infeasible to obtain data on all $N \simeq 15,000$ boat trips: instead, use a probability sample $s \subset U$

Two sources of information on the charter boat fishery

Sample with angler interviews:

Monthly logbook records:

SOUTH CAROLINA CHARTERBOAT LOGBOOK Revised 4-2012

Vessel (Please Print): ______ Date: ______ Permit No.: ______

Number of Anglers: ______ Trip Start Time: ______ Actual Hours Fished: ______ Location: ______ Example: 32-78-C1 (see map)

Trip Start Location: ______ Artificial Reef Name: ______ Target Species: ______ (Please specifiy)

Locale: ☐ Estuarine ☐ 0 – 3 miles ☐ Offshore

Method: ☐ Troll ☐ Cast / Fly ☐ Bottom ☐ Dive ☐ Gig

Water Depth: Shallowest: ______ feet Deepest: ______ feet

MAIL OR FAX REPORT BY THE 10TH OF THE MONTH TO: SCDNR Fisheries Statistics Section, P.O. Box 12559, Charleston, SC 29422-2559 FAX: (843) 953-9362 Phone: (843) 953-9313

AGENCY USE ONLY

Yr Mo Day Permit # Location Locale Ang# Meth

Target Sp. Hrs. Reef Trip Start Shallowest Deepest

	Species	# Kept	Lbs Kept	# Released Alive	# Released Dead		Species	# Kept	Lbs Kept	# Released Alive	# Released Dead
1050	Dolphin					1423	Gag				
4710	Wahoo					1424	Scamp				
4655	Yellowfin Tuna					1414	Snowy Grouper				
4658	Blackfin Tuna					1416	Red Grouper				
3026	Sailfish					1410	Other Grouper				
2177	White Marlin						(Specify)				
2179	Blue Marlin					3302	Red Porgy (Pinks)				
1940	King Mackeral					3295	Other Porgies				
3840	Spanish Mackeral						(Specify)				
4653	Little Tunny					3764	Red Snapper				
0330	Bonita					3765	Vermillion Snapper				
4654	Skip Jack					3360	Black Sea Bass				
0180	Barracuda					3314	Spottail Pinfish				
3810	Spadefish					1441	White Grunt				
0030	Amberjack					1440	Other Grunts				
0870	Crevalle Jack						(Specify)				
0230	Bluefish					4560	Triggerfish				
0570	Cobia					1082	Red Drum				
4350	Tarpon					1081	Black Drum				
	Other Fish					3447	Spotted Seatrout				
	(Specify)					3446	Weakfish				
						1209	Flounder				
						3560	Sheepshead				
						4410	Ladyfish				
						1970	Whiting				
						2670	Inshore Pinfish				
						3518	Sharpnose Shark				
						3495	Blacktip Shark				
						3483	Bonnethead Shark				
						3521	Spiny Dogfish				
						3511	Smooth Dogfish				
						3508	Other Sharks				
							(Specify)				
						2860	Stingrays				

Captain's Notes:

Signature: ______

Print Name: ______

Goal: combine logbook database with survey data

- Design-based difference estimators
- Extension to multiple frames
- Require matching of sampled elements to auxiliary records
  - most theory and methods assume matching is done without error
- Some results on estimation under imperfect matching
  - properties of difference estimators
  - simulation results based on South Carolina charter boat fishing

Design-based inference for the finite population total

- Draw probability sample $s \subset U$ via design with known, positive inclusion probabilities $\Pr[k \in s] = \pi_k > 0$
- Sample membership indicator $I_k = 1$ if $k \in s$, $I_k = 0$ otherwise

$$\mathrm{E}[I_k] = \pi_k, \text{ averaging over all possible samples}$$

- Since $\mathrm{E}[I_k/\pi_k] = 1$ under repeated sampling, unbiased Horvitz-Thompson estimator of T is

$$\widehat{T} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} y_k \frac{I_k}{\pi_k}$$

Now suppose we have the following:

- Auxiliary data $\boldsymbol{x}_\ell$ for all ℓ in some database $\mathcal{A}$
- Perfect, known matching from $\mathcal{A}$ to population U:

$$M_{k\ell} = \begin{cases} 1, & \text{if } \ell \in \mathcal{A} \text{ matches } k \in U, \\ 0, & \text{otherwise} \end{cases}$$

- A "method" $\mu(\cdot)$ for predicting y_k from $\boldsymbol{x}_\ell$:

$$\sum_{\ell \in \mathcal{A}} M_{k\ell}\, \mu(\boldsymbol{x}_\ell) = \widetilde{y}_k \text{ predicts } y_k$$

  – for each element k, look up the correct $\boldsymbol{x}_\ell$

  – apply $\mu(\cdot)$, which does not depend on the sample

Difference estimator combines sample and auxiliary data

- Difference estimator of T is then

$$
\begin{aligned}
\widetilde{T} &= \sum_{k \in U} \sum_{\ell \in \mathcal{A}} M_{k\ell} \mu(\boldsymbol{x}_\ell) + \sum_{k \in s} \frac{y_k - \sum_{\ell \in \mathcal{A}} M_{k\ell} \mu(\boldsymbol{x}_\ell)}{\pi_k} \\
&= \sum_{k \in U} \widetilde{y}_k + \sum_{k \in U} (y_k - \widetilde{y}_k) \frac{I_k}{\pi_k} \\
&= \text{(auxiliary-based prediction)} + \text{(bias adjustment)}
\end{aligned}
$$

where $\widetilde{y}_k$ is not random

- Expectation is

$$
\mathsf{E}\left[\widetilde{T}\right] = \sum_{k \in U} \widetilde{y}_k + \sum_{k \in U} (y_k - \widetilde{y}_k)\, \mathsf{E}\left[\frac{I_k}{\pi_k}\right] = T
$$

Variance of the difference estimator

$$\text{Var}\left(\sum_{k\in U}\widetilde{y}_k + \sum_{k\in U}(y_k - \widetilde{y}_k)\frac{I_k}{\pi_k}\right)$$

$$= \sum_{j,k\in U}\Delta_{jk}\frac{\left(y_j - \widetilde{y}_j\right)}{\pi_j}\frac{(y_k - \widetilde{y}_k)}{\pi_k}$$

- Compare to Horvitz-Thompson estimator:

$$\text{Var}\left(\sum_{k\in U} y_k\frac{I_k}{\pi_k}\right) = \sum_{j,k\in U}\Delta_{jk}\frac{y_j}{\pi_j}\frac{y_k}{\pi_k}$$

Summary so far . . .

- Difference estimator is exactly unbiased, regardless of the quality of the method $\mu(\cdot)$
- Has smaller variance than HT provided "residuals"

$$y_k - \widetilde{y}_k$$

  have smaller variation than "raw values" y_k

  – (If $M_{k\ell} \equiv 0$, we get back HT)
- Have an exactly unbiased variance estimator
- Above results assume (1) one frame covers the universe and (2) matching is perfect

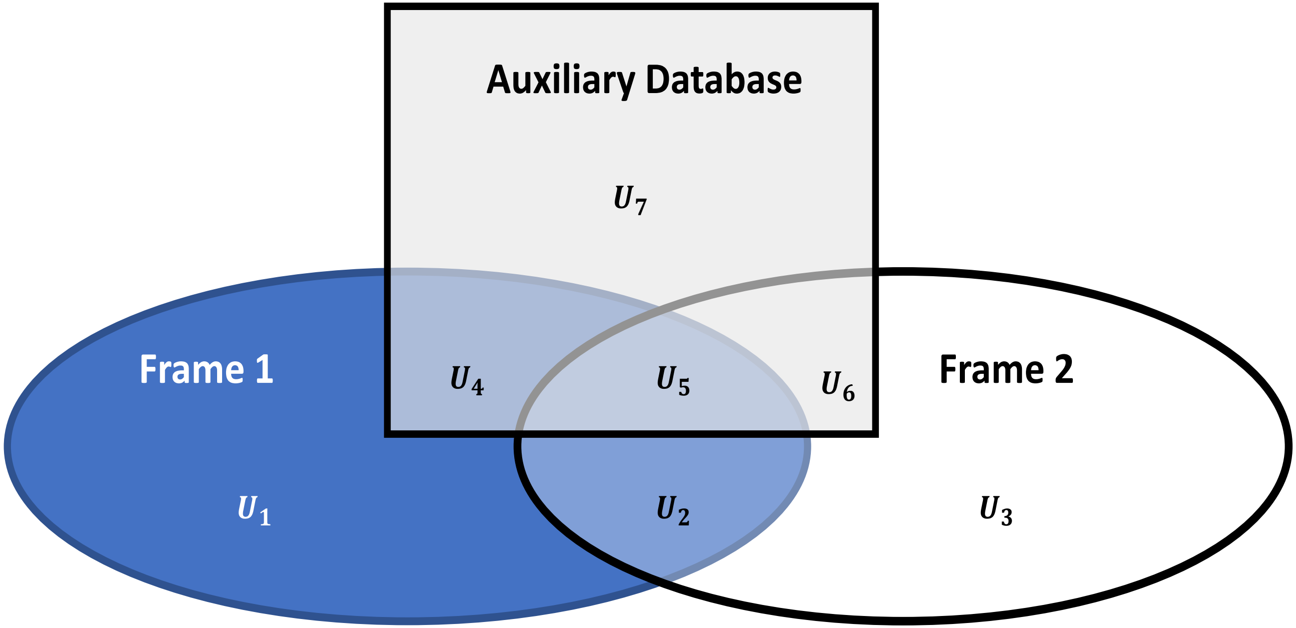
Suppose one frame does not cover the universe

- Assume that the universe U is completely covered by disjoint “overlap domains”:

$$U = \left\{\cup_{g\in G_1} U_g\right\} \cup \left\{\cup_{g\in G_2} U_g\right\} \cup \left\{\cup_{g\in G_3} U_g\right\}$$

- If $g \in G_1$, overlap domain U_g is covered by one or more frames, but not the database
- If $g \in G_2$, overlap domain U_g is covered by one or more frames and the database
- If $g \in G_3$, overlap domain U_g is covered only by the database

$$U = \{U_1 \cup U_2 \cup U_3\} \cup \{U_4 \cup U_5 \cup U_6\} \cup \{U_7\}$$



Overall estimation approach

		In Auxiliary Database?	
		No	Yes
In Sampling Frame(s)?	No	☹	• G_3 • Synthetic predictor • Biased • Zero sampling variance
	Yes	• G_1 • Mecatti estimator • Unbiased • Potentially large variance	• G_2 • Difference estimator • Unbiased • Small variance if auxiliary information is good

Mecatti estimator: adjusting for multiple frames

- From frame f, draw a sample s_{fg} to represent U_g
- Compute Horvitz-Thompson estimator

$$\widehat{T}_{fg} = \sum_{k \in s_{fg}} \frac{y_k}{\pi_k^{(f)}}, \quad \text{where } \mathrm{E}\left[\widehat{T}_{fg}\right] = T_g$$

- Define the coverage indicator

$$F_{fg} = \begin{cases} 1, & \text{if overlap domain } g \text{ is covered by frame } f \\ 0, & \text{otherwise} \end{cases}$$

Mecatti estimator, continued

- Adjust for multiplicity by constructing weights

$$\psi_{fg} = \frac{F_{fg}}{\left(\sum_f F_{fg}\right)}$$

($\psi_{fg} = 1$ if domain covered by only one frame; $1/2$ if two frames, etc.)

- Unbiased Mecatti/multiplicity estimator for $\sum_{g \in G_1} T_g$ is

$$\sum_{g \in G_1} \sum_{f=1}^{F} \psi_{fg} \widehat{T}_{fg}$$

Extending Mecatti to difference estimator

- Multiplicity-adjusted difference estimator for $g \in G_2$:

$$\begin{aligned}
\widetilde{T}^*_g &= \sum_{k \in U_g} \sum_{\ell \in \mathcal{A}} M_{k\ell}\mu(\boldsymbol{x}_\ell) + \sum_{f=1}^{F} \psi_{fg} \sum_{k \in s_{fg}} \frac{y_k - \sum_{\ell \in \mathcal{A}} M_{k\ell}\mu(\boldsymbol{x}_\ell)}{\pi_k^{(f)}} \\
&= \sum_{k \in U_g} \widetilde{y}_k + \sum_{f=1}^{F} \psi_{fg} \sum_{k \in U_g} (y_k - \widetilde{y}_k) \frac{I_k^{(f)}}{\pi_k^{(f)}}
\end{aligned}$$

- Unbiased difference estimator for $\sum_{g \in G_2} T_g$ is then

$$\sum_{g \in G_2} \widetilde{T}^*_g$$

Synthetic predictor for the uncovered population

- G_3 has no sampling frame coverage
- Can only predict with the auxiliary data,

$$\widetilde{T}_g = \sum_{k \in U_g} \sum_{\ell \in \mathcal{A}} M_{k\ell} \mu(\boldsymbol{x}_\ell) = \sum_{k \in U_g} \widetilde{y}_k$$

- Synthetic predictor for $\sum_{g \in G_3} T_g$ is then

$$\sum_{g \in G_3} \widetilde{T}_g = \sum_{g \in G_3} \sum_{k \in U_g} \widetilde{y}_k$$

- Zero sampling variance, unknown bias

Now suppose matching is imperfect

- Replace $M_{k\ell} = 0$ or 1 by match metrics $m_{k\ell} \in [0, 1]$
  - known only for sampled k
- Produced by deterministic algorithm
- Could involve formal probabilistic record linkage (Fellegi and Sunter 1969, Winkler 2009) or other methods
  - conditional probabilities, likelihood ratios, ...
- Whatever their origin, treat $m_{k\ell}$ as fixed in what follows

Modifying the multi-frame estimator

- Under perfect matching, multi-frame estimator is

$$\sum_{g\in G_1}\sum_{f=1}^{F}\psi_{fg}\widehat{T}_{fg}+\sum_{\ell\in\mathcal{A}}\left(\sum_{g\in G_2\cup G_3}\sum_{k\in U_g}M_{k\ell}\right)\mu(\boldsymbol{x}_\ell)$$
$$+\sum_{g\in G_2}\sum_{f=1}^{F}\psi_{fg}\sum_{k\in s_{fg}}\frac{y_k-\sum_{\ell\in\mathcal{A}}M_{k\ell}\mu(\boldsymbol{x}_\ell)}{\pi_k^{(f)}}$$

- Under imperfect, $m_{k\ell}$ is known only for $k\in s_{fg}$
- Cannot just substitute $m_{k\ell}$ for $M_{k\ell}$ in second term, but ok in third

Modifying the multi-frame estimator, continued

- Second term under perfect matching is

$$\sum_{\ell \in \mathcal{A}} \left(\sum_{g \in G_2 \cup G_3} \sum_{k \in U_g} M_{k\ell} \right) \mu(\boldsymbol{x}_\ell)$$

- If ℓth record matches some element in $\cup_{g \in G_2 \cup G_3} U_g$, then (parenthetical term) $= 1$
- Under imperfect matching, estimate parenthetical term as equal to 1
  - (or construct a complicated, and biased, estimator)

Modifying the multi-frame estimator, final

- Analogue of perfect-match multi-frame estimator

$$\sum_{g\in G_1}\sum_{f=1}^{F}\psi_{fg}\widehat{T}_{fg}+\sum_{\ell\in\mathcal{A}}\left(\sum_{g\in G_2\cup G_3}\sum_{k\in U_g}M_{k\ell}\right)\mu(\boldsymbol{x}_\ell)$$
$$+\sum_{g\in G_2}\sum_{f=1}^{F}\psi_{fg}\sum_{k\in s_{fg}}\frac{y_k-\sum_{\ell\in\mathcal{A}}M_{k\ell}\mu(\boldsymbol{x}_\ell)}{\pi_k^{(f)}}$$

  is then

$$\begin{aligned}\widetilde{T}_{diff} &= \sum_{g\in G_1}\sum_{f=1}^{F}\psi_{fg}\widehat{T}_{fg}+\sum_{\ell\in\mathcal{A}}(1)\mu(\boldsymbol{x}_\ell)\\ &\quad+\sum_{g\in G_2}\sum_{f=1}^{F}\psi_{fg}\sum_{k\in s_{fg}}\frac{y_k-\sum_{\ell\in\mathcal{A}}m_{k\ell}\mu(\boldsymbol{x}_\ell)}{\pi_k^{(f)}}\end{aligned}$$

Bias of the estimator

- Bias depends on matching and prediction error:

$$\mathsf{E}\left[\widetilde{T}_{diff}\right] - T = -\sum_{g \in G_3} T_g + \sum_{\ell \in \mathcal{A}} \mu(\boldsymbol{x}_\ell) - \sum_{\ell \in \mathcal{A}} \left(\sum_{g \in G_2} \sum_{k \in U_g} m_{k\ell} \right) \mu(\boldsymbol{x}_\ell)$$
$$= -(\text{total uncovered}) + (\text{database}) - (\text{overlap})$$

- Sufficient conditions for unbiased estimation are

$$G_3 = \emptyset \text{ and } \sum_{g \in G_2} \sum_{k \in U_g} m_{k\ell} = 1 \text{ for all } \ell \in \mathcal{A}$$

- Asymptotic unbiasedness and mean square consistency requires some combination of "not too much" matching error or undercoverage, and "good" prediction of the uncovered population

Variance of the estimator

- Variance of the estimator is (setting $m_{k\ell} \equiv 0$ for $k \in \cup_{g\in G_1} U_g$):

$$\sum_{f=1}^{F} \sum_{g\in G_1\cup G_2} \sum_{g'\in G_1\cup G_2} \psi_{fg}\psi_{fg'} \sum_{j\in U_g} \sum_{k\in U_{g'}} \Delta_{jk}^{(f)} \frac{d_j}{\pi_j^{(f)}} \frac{d_k}{\pi_k^{(f)}}$$

with $d_j = \begin{cases} y_j - \sum_{\ell\in\mathcal{A}} M_{j\ell}\mu(\boldsymbol{x}_\ell), & \text{perfect matching} \\ \text{prediction error} & \\ y_j - \sum_{\ell\in\mathcal{A}} m_{j\ell}\mu(\boldsymbol{x}_\ell), & \text{imperfect matching} \\ \text{matching and/or prediction error} & \end{cases}$

- $\mathrm{Var}\left(\widetilde{T}_{diff}\right) = O\left(\frac{N^2}{\min_f n_f}\right)$ and $N^{-1}\widetilde{T}_{diff} \overset{\text{m.s.}}{\to} N^{-1}\mathsf{E}\left[\widetilde{T}_{diff}\right]$
- Unbiased variance estimation provided all $\pi_{jk}^{(f)} > 0$ in each frame

Use SC recreational fishery to devise a simulation study

- About 450 charter boats and 15,000 boat trips along the Atlantic Coast each year
- Survey data from sampled angler on boat trip on the actual date
  - coverage error: not all sites and times are in-frame
  - lots of sampling error
- Logbook data from captain's report, later that month
  - nonresponse
  - measurement error
- Lots of matching error!

Characterizing the match metrics

- Perfect match: $m_{k\ell} = 1$ for $\ell = \ell_1$ and 0 otherwise
- High-quality match: $m_{k\bullet} = \sum_{\ell \in \mathcal{A}} m_{k\ell} = 1$

$$m_{k\ell} = \begin{cases} 1/3, & \text{if } \ell = \ell_1,\ \ell = \ell_2 \text{ or } \ell = \ell_3, \\ 0, & \text{otherwise} \end{cases}$$

- Low-quality match: $m_{k\bullet} = \sum_{\ell \in \mathcal{A}} m_{k\ell} < 1$

$$m_{k\ell} = \begin{cases} 1/6, & \text{if } \ell = \ell_1,\ \ell = \ell_2 \text{ or } \ell = \ell_3, \\ 0, & \text{otherwise} \end{cases}$$

- No match: $m_{k\ell} = 0$ for all $\ell \in \mathcal{A}$

Custom matching algorithm

- Match metrics $\{m_{k\ell}\}_{k\in s,\ell\in\mathcal{A}}$ developed by South Carolina Department of Natural Resources staff

Interview variables	Logbook variables
Date of interview	Date of reported trip
Time of interview	Estimated trip end time
License number of vessel	License number of vessel
Name of vessel given	Name of vessel reporting
Interview site	Reported start site

- Large fraction of unmatched trips and low-quality matches

	No Match	LQ	HQ	Perfect
Empirical	11.0%	52.5%	36.5%	0.0%

Simulation experiment, I

- Use real logbook data to create artificial population with $|U| = 10{,}647$ boat trips, sorted in space and time
- Use Markov chain to assign (unobservable) states to groups of population boat trips:

	state from Markov chain			
	no match	LQ	HQ	perfect
size of group of elements:	1	10	5	1
logbook records created:	0	5	5	1
metric sum:	0	$1/2$	1	1

- If an LQ element is selected, metrics (correctly) indicate it might match one of five records, or none of them

Simulation experiment, II

- Set Markov chain parameters to simulate match metrics $\{m_{k\ell}\}$ and logbook database $\mathcal{A}$ under two scenarios:

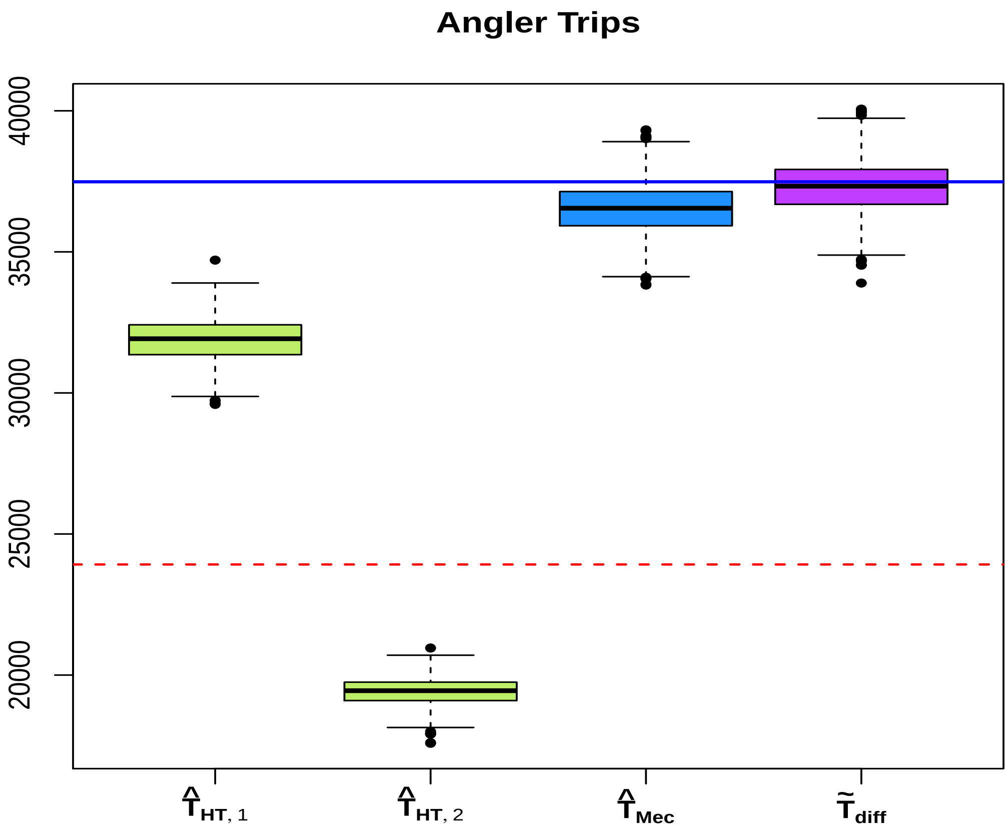
	$\lvert\mathcal{A}\rvert$	No Match	LQ	HQ	Perfect
Poor Match	6,836	8.6%	54.4%	31.7%	5.3%
Better Match	9,031	2.3%	23.3%	69.8%	4.7%
Empirical		11.0%	52.5%	36.5%	0.0%

- Population of boat-trips and database of logbook records is then fixed
- Create two incomplete frames, partially overlapping
- Sample repeatedly from this finite population

Simulation experiment, III

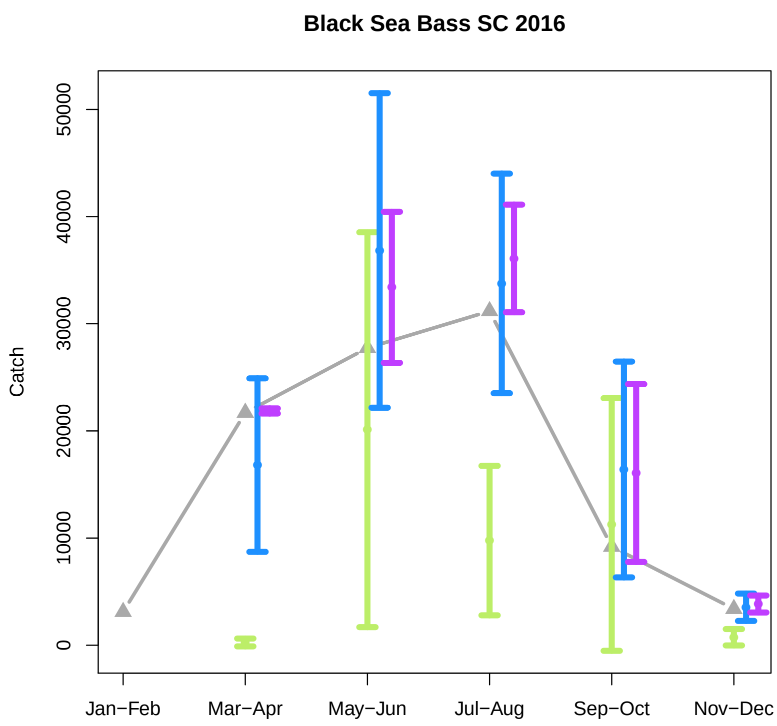
- Simplifications:
  - no "differential matching": quality of $m_{k\ell}$ does not depend on y_k
  - no measurement error: $\mu(\boldsymbol{x}_\ell) = y_k$ for perfect match
- Draw 1000 repeated samples from simulated population
  - stratified, two-stage, unequal-probability selection
- Compute $\widehat{T}_{HT,1}$, $\widehat{T}_{HT,2}$, $\widehat{T}_{Mec}$, $\widetilde{T}_{diff}$ for number of angler trips and several species in each simulated sample
- Assess bias, variance, and MSE for each estimator

Even with poor match, difference dominates Mecatti

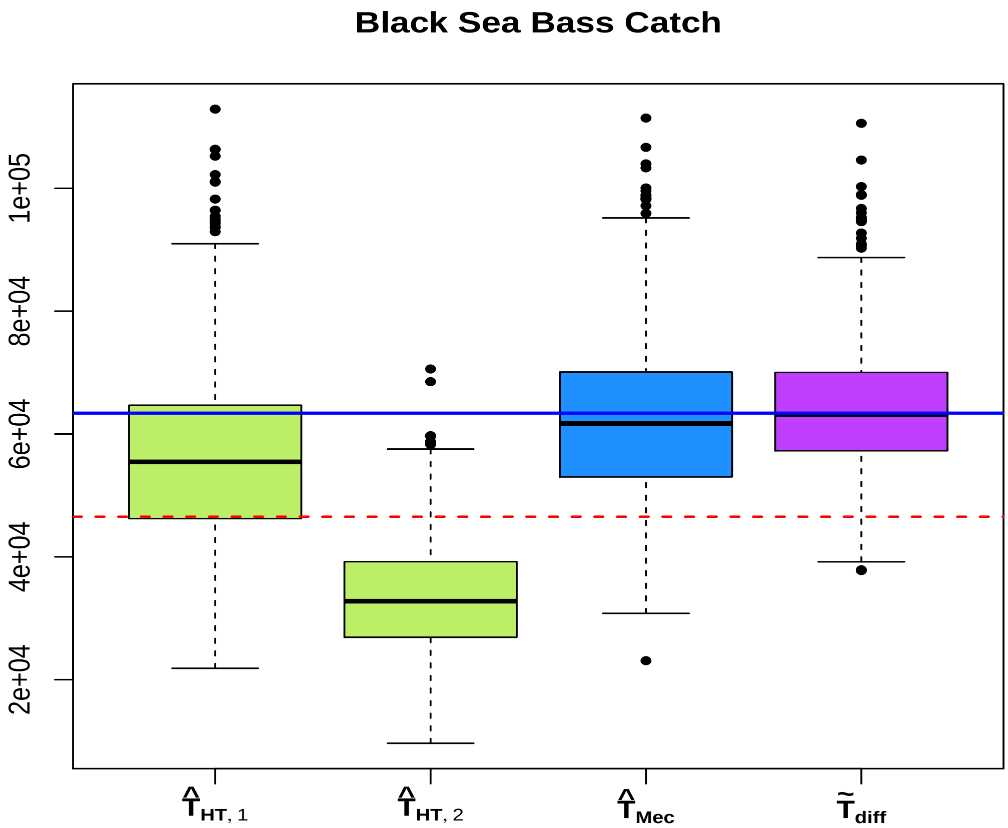


Real Black Sea Bass logbook, HT, and combinations

- Frequently targeted and caught; appears regularly in both sources



Difference estimator dominates Mecatti



Summary

- Auxiliary information is useful even with imperfect matching
  - naive difference estimator improves accuracy and precision of multiplicity estimator
  - variance estimators and confidence intervals (not shown) work well
- Matching across frames or matching across auxiliary databases adds challenges
- Grazie mille!
  Contact info: `FJay.Breidt@colostate.edu`