Understanding the Effects of Record Linkage on Estimation of Total when Combining a Big Data Source with a Probability Sample

Benjamin Williams, Statistical Science

World Changers Shaped Here

SMU®

Introduction

- The US National Oceanic and Atmospheric Administration (NOAA) is responsible for:
  - Fishing season lengths
  - Bag limits
  - Keeps fish populations stable, prevents overfishing, and combats effects of natural disasters

Background

- NOAA estimates the total fish caught by recreational anglers

- Catch = *Catch per Unit Effort* X *Effort*

- *CPUE* is estimated from a public dockside intercept survey

- *Effort* is estimated from an address-based mail survey

Problems

- The mail survey suffers from low response rates and lengthy estimation

- The National Research Council has recommended electronic reporting

- Electronic reporting may allow for near real time estimation

Electronic Reporting

- NOAA is experimenting with allowing recreational charter captains to self-report trips

- Captains volunteer to participate – this data can hopefully replace the mail survey and improve estimation

GPS Date 11-15-3916
GPS Time 17:27:31
TID 520036
Iridium Network Available

Lat 39:03.711 N
Long 76:35.009 W
Speed 0 Knots
Course 0 (N)

Electronic Reporting

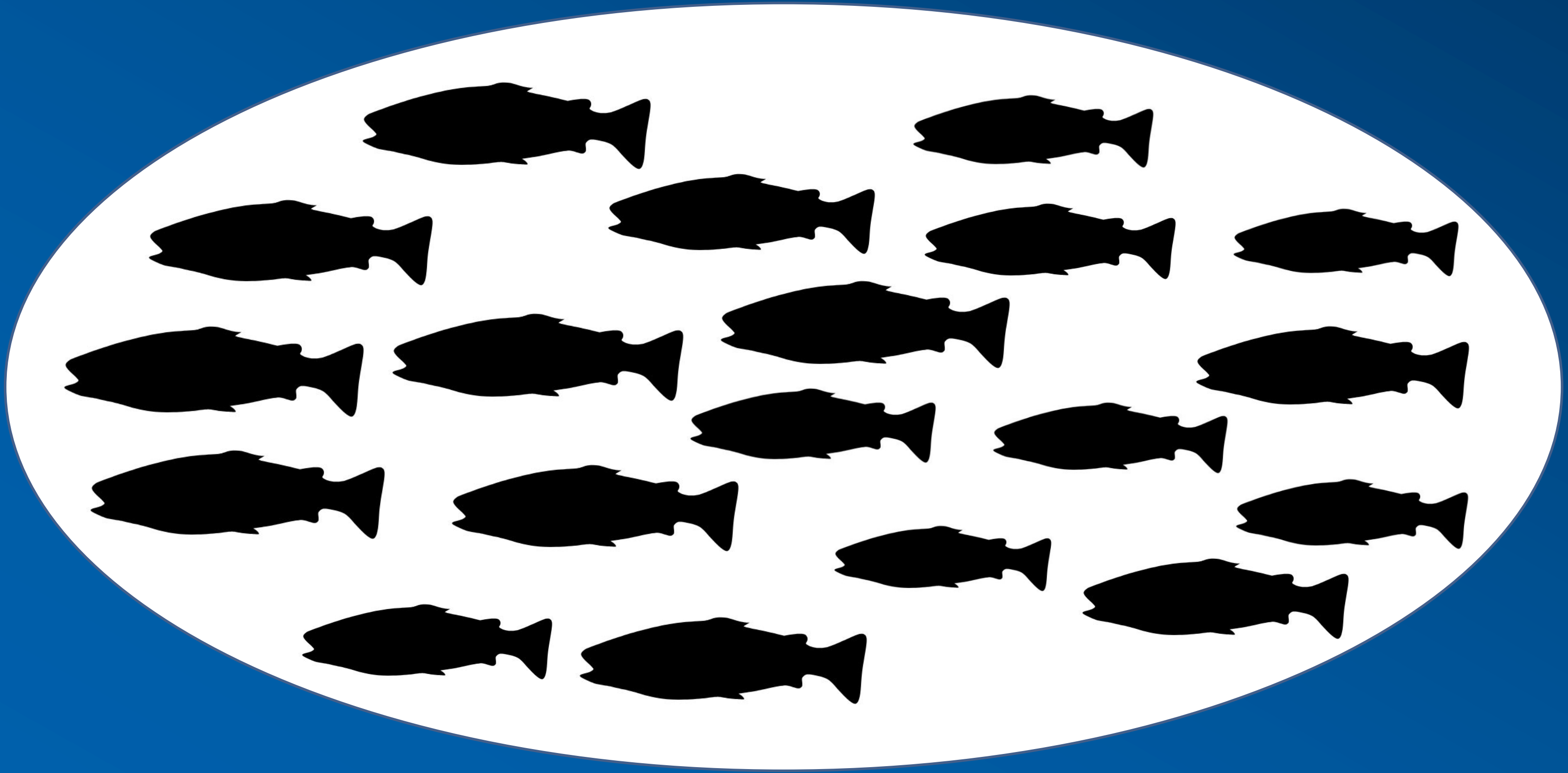
- The self-reports constitute a non-probability sample
- Estimators using data from non-probability samples may suffer
- The self-reports are a large sample containing useful information

Current Methods

- Liu et al. (2017) and Breidt, Opsomer, & Huang (2018)
- Use self-reports as auxiliary information, allowing evaluation of the estimators
- The proposed estimators use capture-recapture methodology

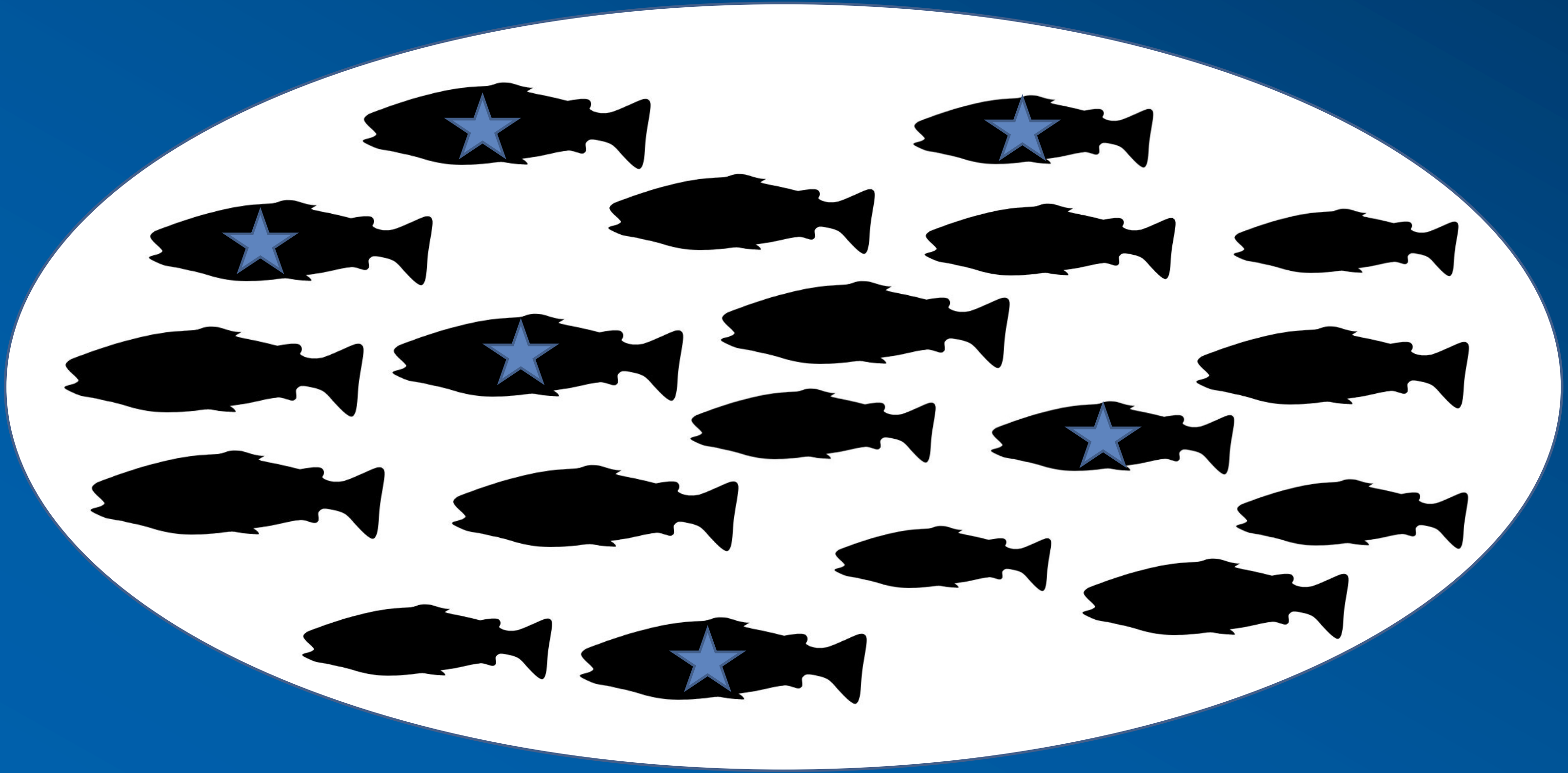
Capture-Recapture

N fish in the pond



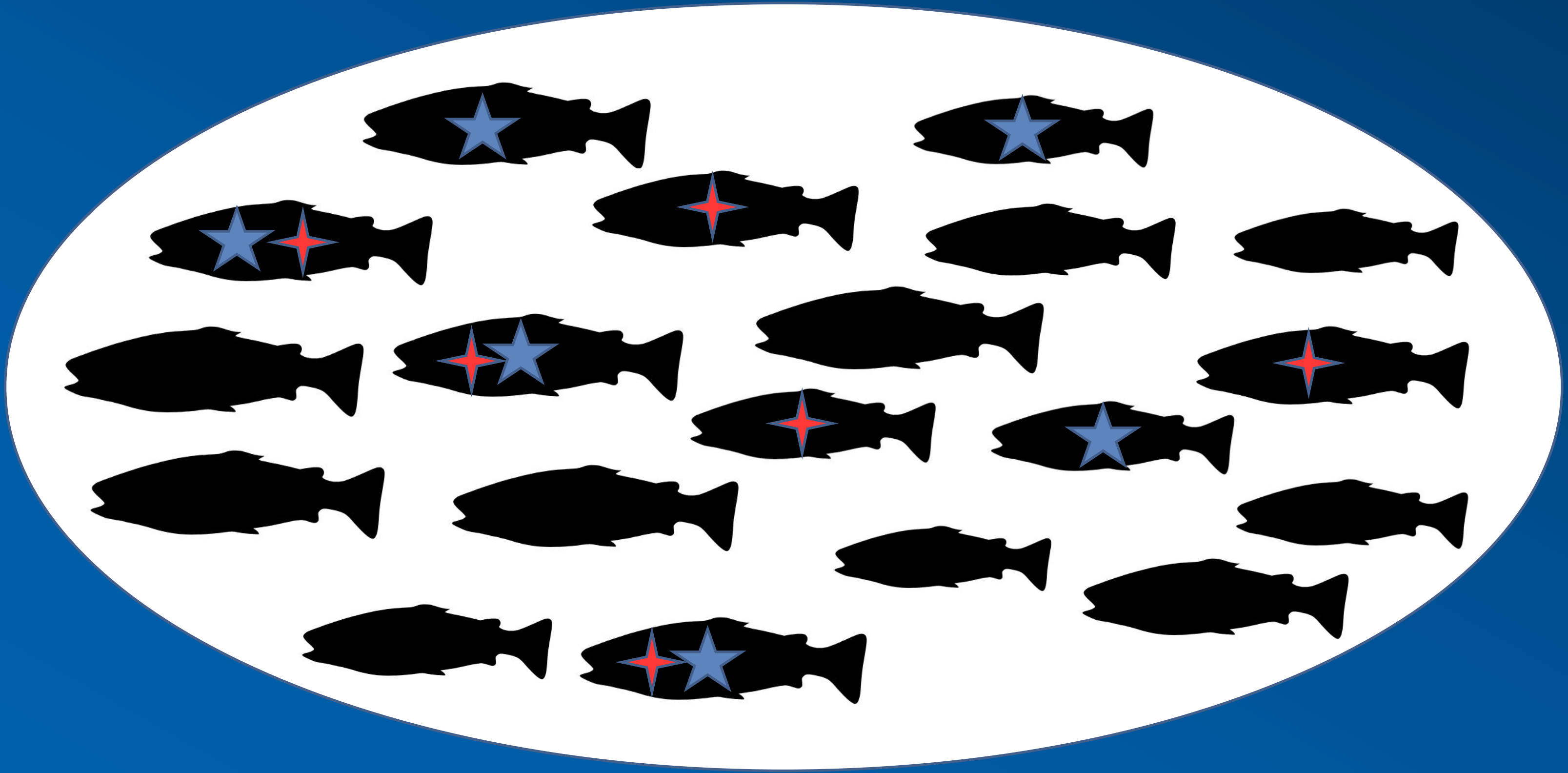
Capture-Recapture

Day 1: Catch n_1 fish and tag them



Capture-Recapture

Day 2: Catch n_2 fish, m of which are tagged

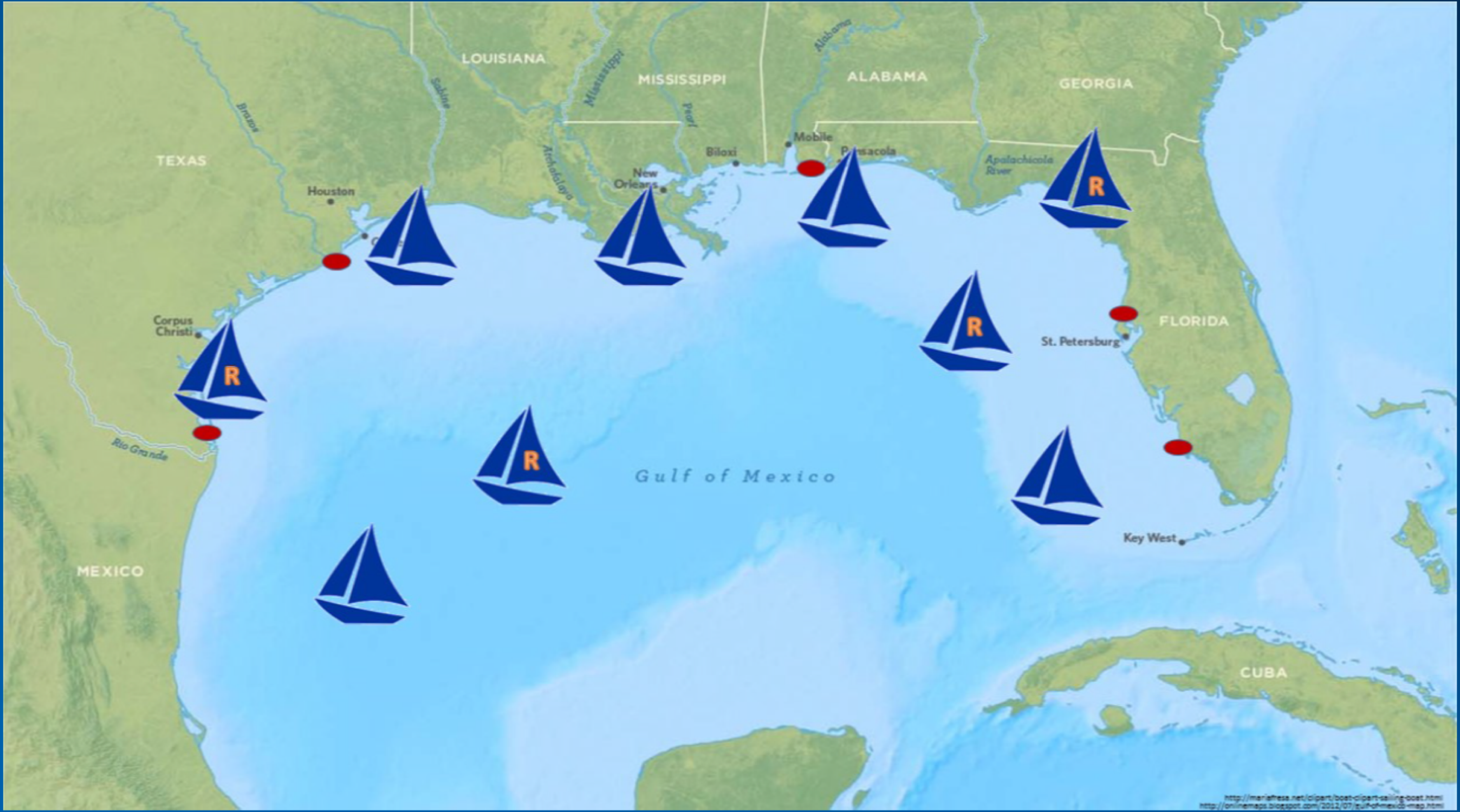


Capture-Recapture

- Estimate the total catch of fish with:

$$\hat{N} = \frac{n_1 n_2}{m}$$

- Assume: recapture sample is a probability sample
- Self-reports are analogous to the capture sample
- Dockside intercept is analogous to recapture sample
- We want to estimate total of a characteristic, not the total units in the population



http://mariafresa.net/clipart/boat-clipart-sailing-boat.html
http://onlinemaps.blogspot.com/2012/07/gulf-of-mexico-map.html

Sampling Setup

		Dockside Intercept		
		Sampled	Not sampled	
Self-Reports	Did report	m y^*, y	$n_1 - m$ y^*	n_1
	Did not report	$n_2 - m$ y		
		n_2		$\hat{N}$

Estimator

Liu et al (2017):

$$\hat{t}_{y2} = t_{y^*} + \frac{n_1}{\hat{n}_1}\left(\hat{t}_y - \hat{t}_{y^*}\right)$$

$t_{y^*} = \textit{total reported catch}$

$\hat{n}_1 = \textit{estimated number of reports}$

$\hat{t}_y = \textit{estimated total catch from intercept sample}$

$\hat{t}_{y^*} = \textit{estimated total reported catch from intercept sample}$

Matching Errors

Estimation requires linking trips between the samples, but this is difficult due to:

- Captains may report well after a trip ends
- Device/Measurement error
- Timing of end of a trip and time of interview will be different in both samples, cannot identify multiple trips in same day
- Self-reports consist of device reported data and captain reported data

Matching Error Types

Type 1: *False-positive* (biases estimators downward)

- Link a sampled trip to a reported trip
- That trip did not actually report

Type 2: *Mismatch* (not much of a concern)

- Link a sampled trip to a reported trip
- That trip did actually report, but linked to wrong reporting trip

Type 3: *False-negative* (biases estimators upward)

- Fail to link a sampled trip to a reported trip
- That trip did actually report

Record Linkage

- The quality of the estimators depends on accurate linking
- Due to non-sampling errors, matching trips is difficult
- Implement Record Linkage
    - Fellegi and Sunter (1969), Bell et al (1994)

Record Linkage

- Call x and y the two values observed for the k^{th} linking variable
- The linking score is:

$$S_k = \log\left(\frac{P(x, y|M = 1)}{P(x, y|M = 0)}\right)$$

$$= \log\big(\mathrm{P}(\mathrm{y}|\mathrm{x}, \mathrm{M} = 1)\big) - \log(\mathrm{P}(\mathrm{y}))$$

where M is an indicator of a match

Record Linkage

- The linking score is simplified based on the amount of agreement or disagreement between x and y (Bell et al 1994)

- Estimate pieces of linking score conditional on a match by holding other linking variables constant

- Assuming independence, sum the scores for each linking variable to obtain a score for each potential link

Example

- Data from 2 years of an electronic reporting experiment in the Gulf of Mexico (2016 – 2017)

- In 2016: 1,628 intercepts, 5,976 self-reports
- In 2017: 1,484 intercepts, 6,277 self-reports

- Use Boat ID number as blocking variable

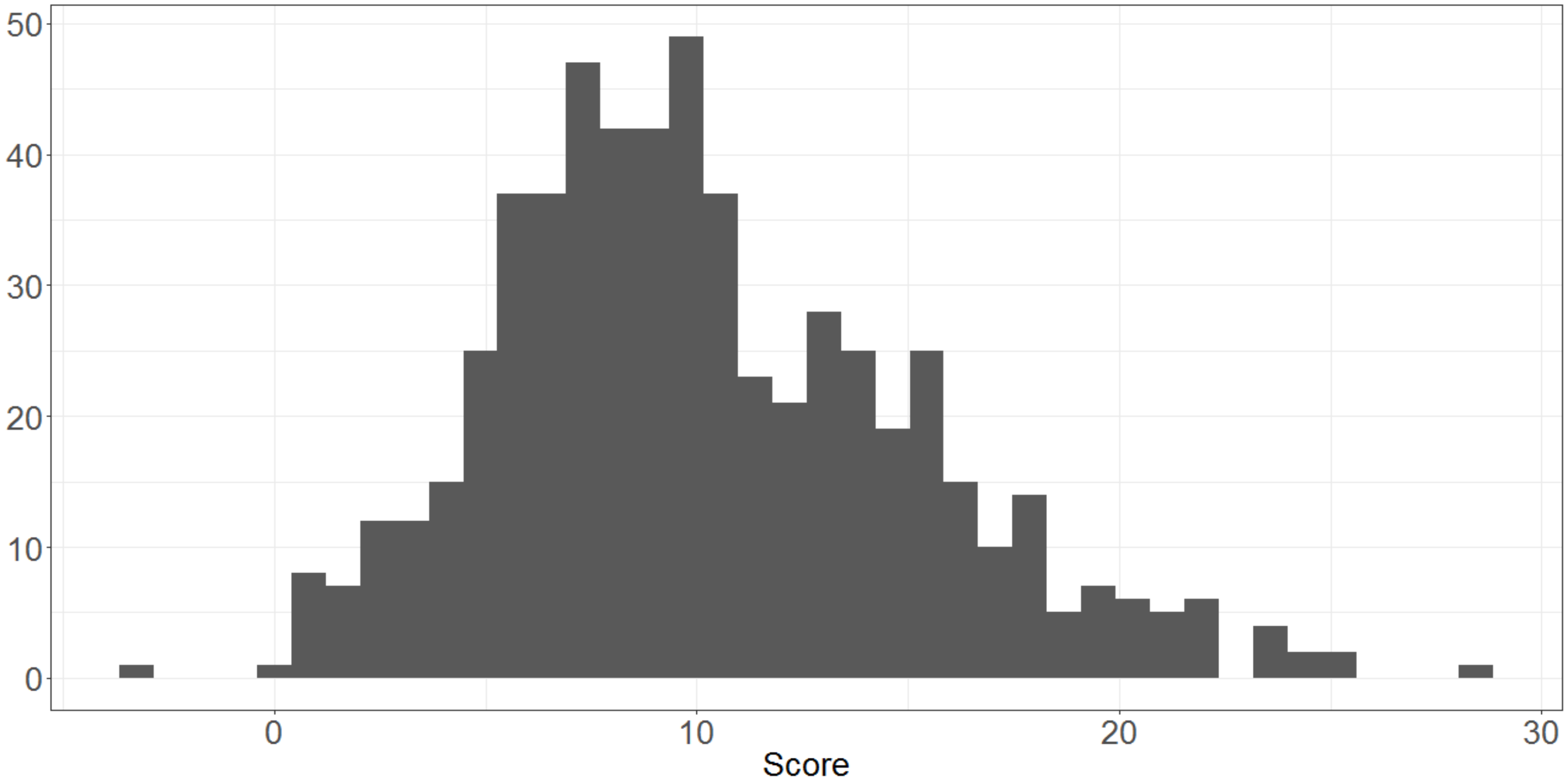
Augmenting GPS Data

- GPS position reported for each self-reported trip
- Over 2.5 million GPS reports to date
- Predict *Return Time*, *Return Location*
- See Ryan McShane in 40.105 at 11:30 for specifics

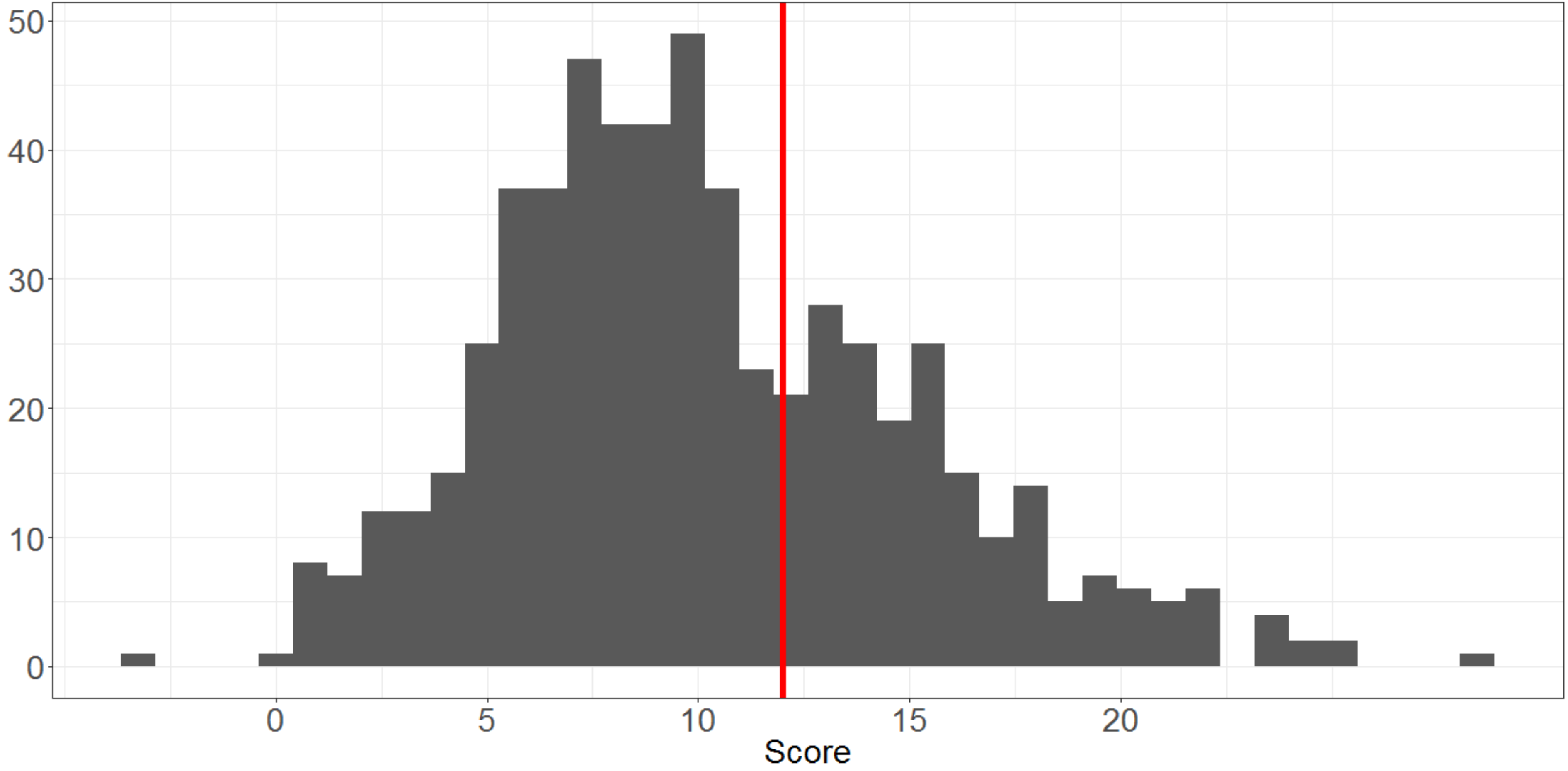
Additional Linking Variables

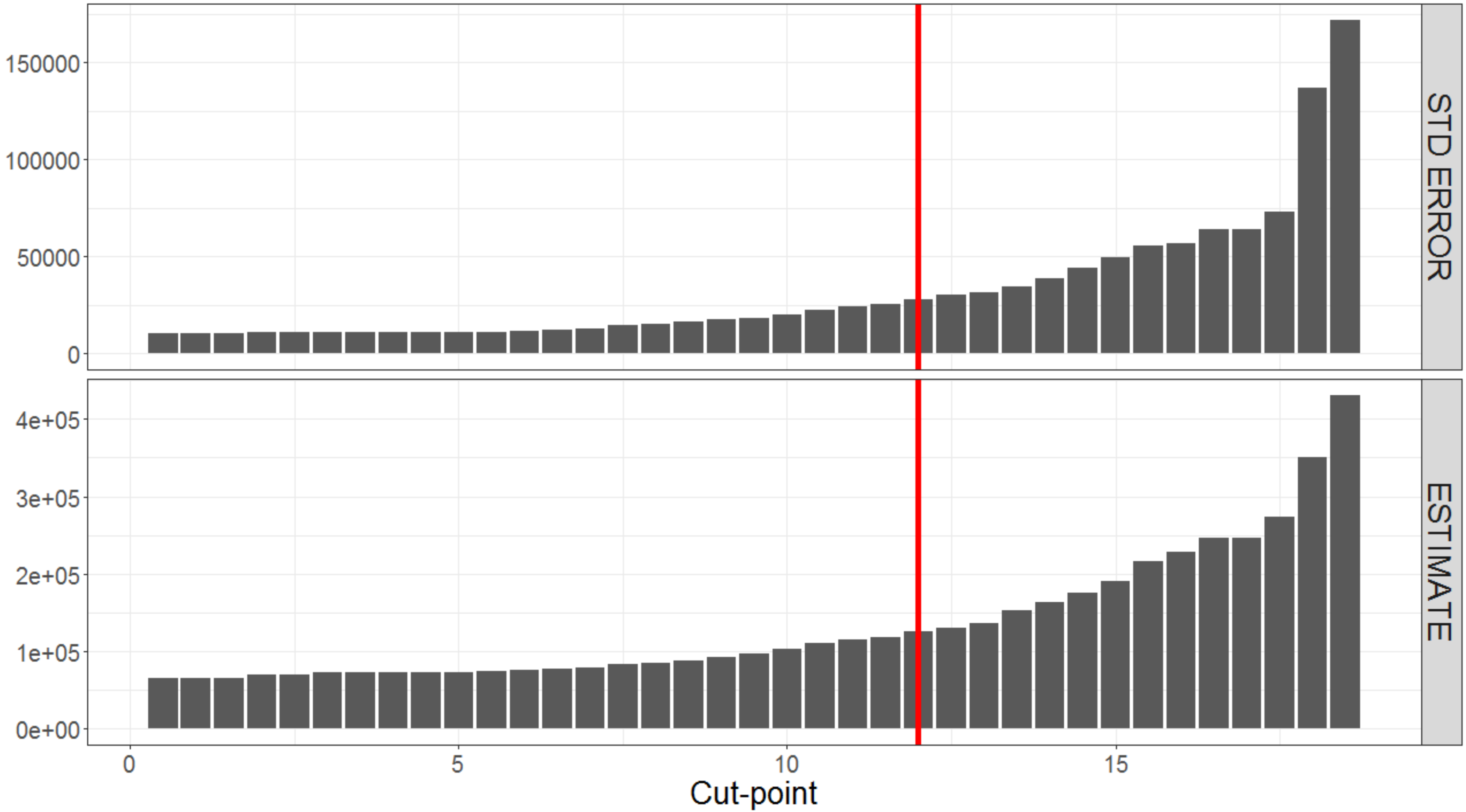
Intercept File (Recorded by Interviewer)	Report File (Reported by Captain of Observed from Device Signal)
Date of Interview	Date of Trip Return (Device)
Interview Site Number	Predicted Return Site (Device)
Number of Fish Harvested per Angler	Number of Fish Harvested for Entire Boat (Captain)
Number of Fish Discarded per Angler	Number of Fish Discarded for Entire Boat (Captain)
Number of Different Species Caught	Number of Different Species Reported (Captain)
Number of Anglers	Number of Anglers (Captain)

Score Distribution



Score Distribution




150000
100000
50000
0
STD ERROR
4e+05
3e+05
2e+05
1e+05
0e+00
ESTIMATE
0
5
10
15
Cut-point

Estimates - Red Snapper Harvest 2017

For AL and FL

NOAA : 13.9% PSE

Record Linkage : 22.2% PSE (cut-point = 12, 86 matches)

Current and Future Work

- Compare other metrics to determine cut-point
- Estimate matching error
- R package for estimation: *blendR* (Williams, 2018)
- Many extensions

References

Bell, R. M., Keesey, J., & Richards, T. (1994). The Urge to Merge: Linking Vital Statistics Records and Medicaid Claims. *Medical Care, 32*(10), 1004-1018.

Breidt, J. F., Opsomer, J. D., & Huang, C. (2018). Model-Assisted Survey Estimation with Imperfectly Matched Auxiliary Data. In V. Kreinovich, S. Sriboonchitta, & N. Chakpitak (Eds.), *Predictive Econometrics and Big Data* (Vol. 753, pp. 21-35). Springer, Cham.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association, 64*(328), 1183-1210.

Liu, B., Stokes, L., Topping, T., & Stunz, G. (2017). Estimation of a Total from a Population of Unknown Size and Application to Estimating Recreational Red Snapper Catch in Texas. *Journal of Survey Statistics and Methodology, 5*(3), 350-317.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic Linkage of Vital Records. Science, 130(3381):954–959.

Tarnecki, J. H. and Patterson, W. F. (2015). Changes in Red Snapper Diet and Trophic Ecology Following the Deepwater Horizon Oil Spill. Marine and Coastal Fisheries, 7(1):135–147.

The National Academies of Sciences, Engineering, and Medicine. 2016. *Review of the Marine Recreational Information Program (MRIP)*. Washington, DC: The National Academies Press. doi: 10.17226/24640

Williams, B. (2018). Combining a Probability and a Non-Probability Sample in a Capture-Recapture Setting. *Journal of Open Source Software*, 3(28), 886, https://doi.org/10.21105/joss.00886.

Thank you!