USING A LARGE GPS DATASET TO ENHANCE SURVEY MATCHING

Ryan P. A. McShane
Southern Methodist University – Department of Statistical Science
Presented at BigSurv18, October 26th, 2018

Thanks; Contact Info

Thanks to BigSurv18 for funding my travel!

This research was funded by **CLS America** under contract **#PSM00068** and federal funds under award **NA15NMF4540082** from **NOAA Fisheries**, U.S. Department of Commerce.

Me:
Ryan McShane rmcshane@smu.edu

Future project questions:
Dr. Lynne Stokes slstokes@smu.edu

1. INTRODUCTION TO PROBLEM

What is NOAA Fisheries doing?

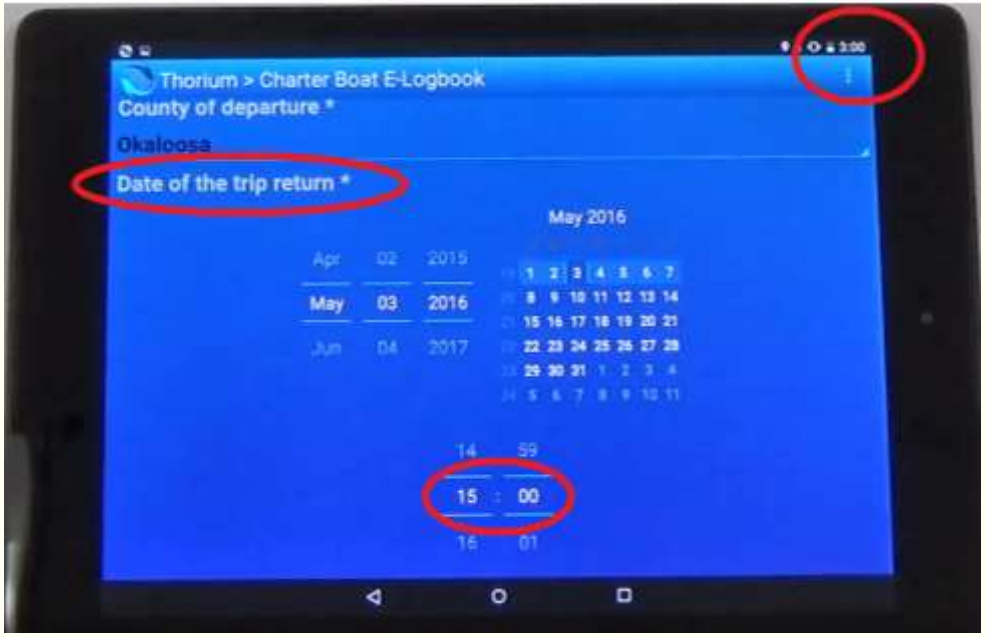
- Currently, NOAA Fisheries collect data on catch and "effort" in the charter boat modality of the recreational fishing sector.
- Catch is estimated with the Access Point Interview Survey (APAIS). The caught fish are inspected, identified, measured, and counted at the passenger level to produce an estimate for catch per unit effort (CPUE).
- "Effort" is estimated via a costly telephone survey with low response rate.
- CPUE and effort are then combined $(CPUE) \times (Effort)$ to produce an estimate of **total catch per species**.

What is the Research Project Proposing?

- An alternative data collection procedure in which charter boat captains report the total catch per species at the end of each trip with a GPS-enabled electronic device.
- The captains' reports (ELBs) are used as auxiliary data to the probability sample of intercepts, resulting in an estimator that has a similar form to a capture recapture estimator.
- This estimation procedure **requires matching intercepted trips to reported ones**. Since intercepted trips are a probability sample, this allows estimation of a reporting rate.

How are the Self-Reports Collected?

- Captains enter their passengers' collective fish catch by species in a tablet app.
- They also report auxiliary information, such as the number of anglers on board, and the time they depart and return.



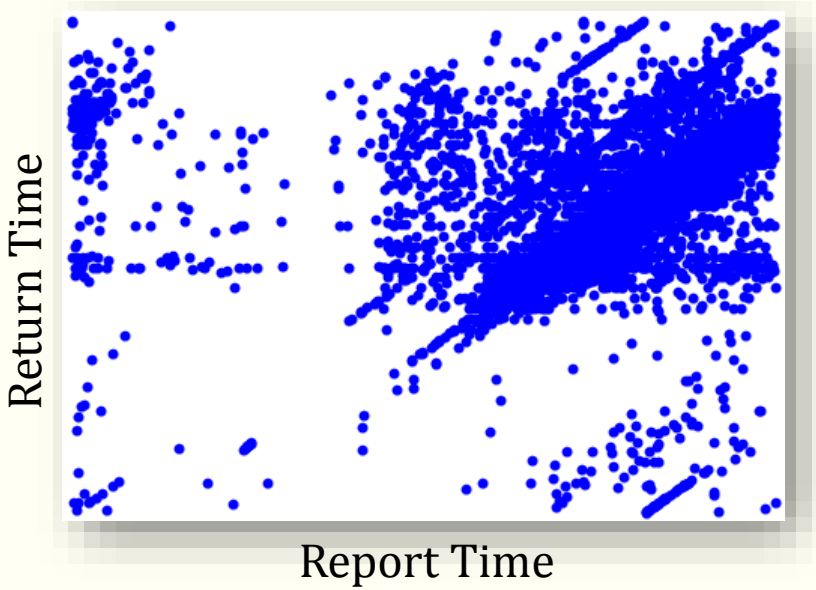
How Reliable are the Self-Reports?

- 16.6% of ELB reports contained a value outside of "reasonable" limits – ex.: reported 320 of one species caught on one trip.
- 31.8% of ELB reports are logically inconsistent, i.e.:
- 56.4% of ELB reports are logically consistent and contain reasonable values.

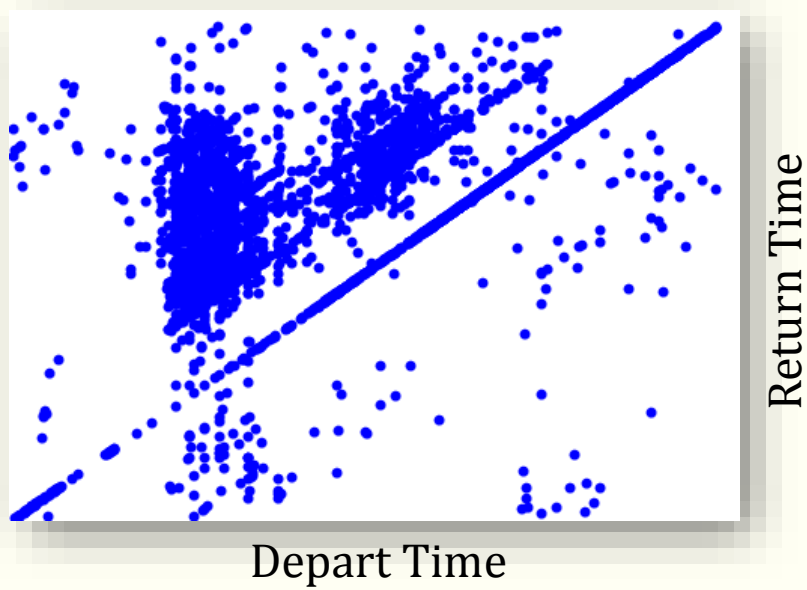
Date-Time Inconsistencies
$depart = return$
$depart > return$
$hours > \text{FLOOR}(return - depart)$
$return_1 > depart_2$

How Reliable are the Self-Reported Times?

Return Time vs. Report Time



Return Time vs. Depart Time

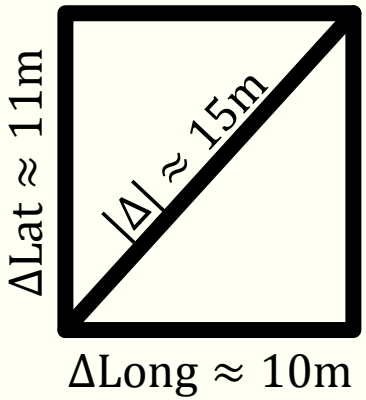


How do we Match Interviews with Self-Reports?

- Captain-reported trips are matched to APAIS interviews by a vessel ID and date.
- However, since charter boats may take multiple trips in a day, the trip report to APAIS interview matching procedure requires a time component.
- The times reported in the ELB are often misleading or erroneous, and thus we sought an alternative source of data to enhance matching – GPS data.
- **Each vessel produces a GPS position report on a periodic basis – to date, there are over 2.5 million such reports.**

What Does the GPS Data Look Like?

- Each vessel is equipped with a Thorium VMS device which reports the vessel's GPS position periodically.
- Time-stamped GPS position reports are produced hourly, every fifteen minutes, or a blend of the two.
- Latitude and Longitude are reported with four decimal places, i.e. $(24.9367, -80.6124)$.
- Even when vessels sit still overnight, coordinates fluctuate ± 0.0001 from time to time (translated to meters in figure).



2. IDENTIFYING LANDING SITES

What are Landing Sites?

- Trips are identified using the GPS data, but we chose to define trips as starting and stopping at the pick-up and drop-off locations, or **landing sites**.
- NOAA provides a list of **APAIS Interview Sites** with GPS coordinates in a ***Site Register***. Since we can only match ELBs with APAIS interviews, the landing sites "can only" be APAIS interview sites.
- This also importantly allows us to estimate when APAIS Interviews would have taken place.

How are Landing Sites Identified?

- We start by seeing how often each vessel is within 400m of each site on the ***Site Register***: The most frequent sites are selected as candidates.
- This process took ~6hrs of compute time with >1000 interview sites and 2.5M observations.
- Then, the distribution of time lags between each observation *at the interview site* is examined – a sufficient number of 4 to 16 hour lags should exist, representing trips.

0 "trips"

Hour_Dif	Frequency	Percent
1	144	100.00

251 "trips"

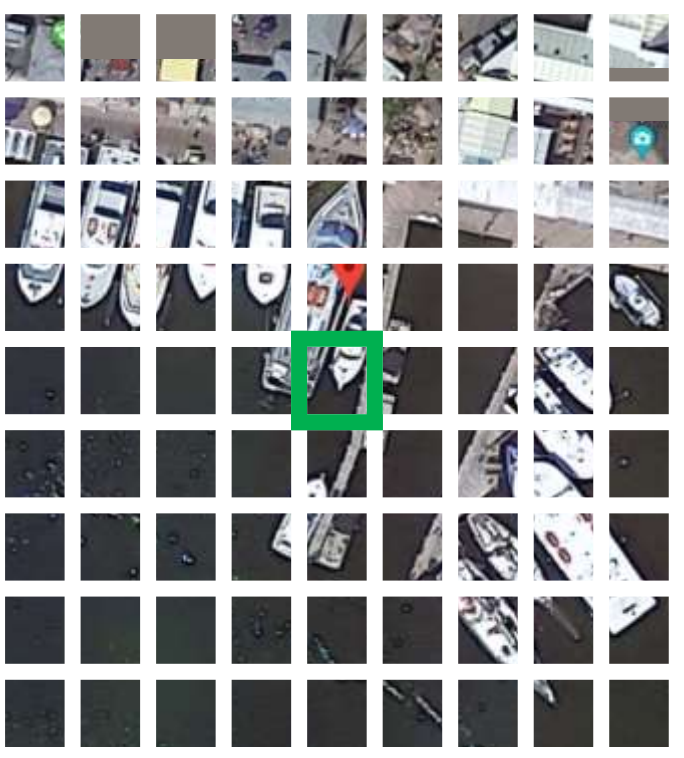
Hour_Dif	Frequency	Percent
0.25	12411	54.85
1	9760	43.13
2	15	0.07
3	5	0.02
4	29	0.13
5	28	0.12
6	42	0.19
7	42	0.19
8	17	0.08
9	27	0.12
10	16	0.07
11	20	0.09
12	16	0.07
13	8	0.04
14	2	0.01
15	4	0.02

How are Landing Sites Identified?

- We also check Google Maps to see if this is a plausible pick-up and drop-off location.
- We often found that these were repair shops, hotels, restaurants, backyards, driveways, et cetera, that just happened to be close to an interview site.

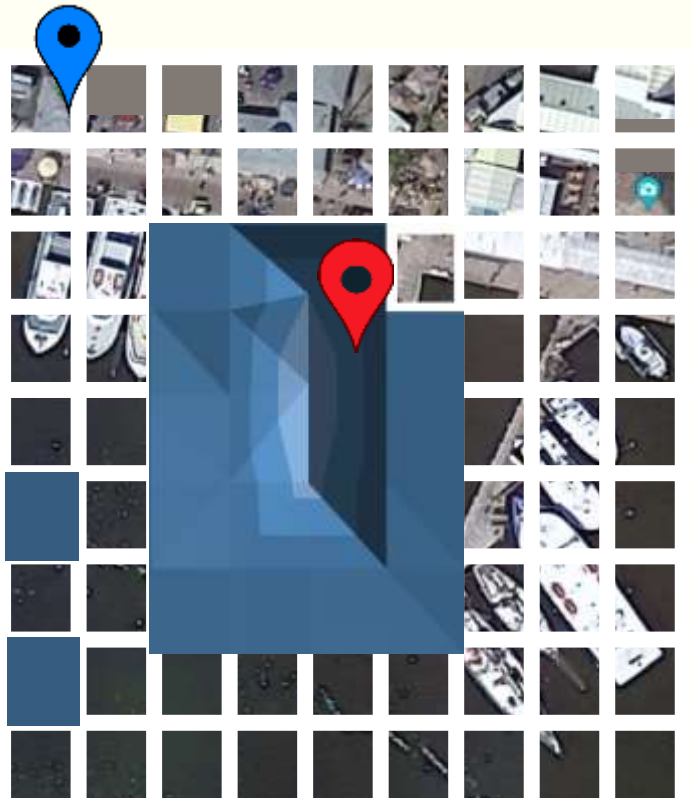
Google Mapping Made Efficient

- These locations, however, are too granular to examine individually. We grouped them by $\pm.0001 \cdot n, n \in \mathbb{N}$ in both LAT and LONG.
- Ex.: $n = 4$ would group 81 individual locations in a 9x9 LAT/LONG grid.



Google Mapping Made Efficient

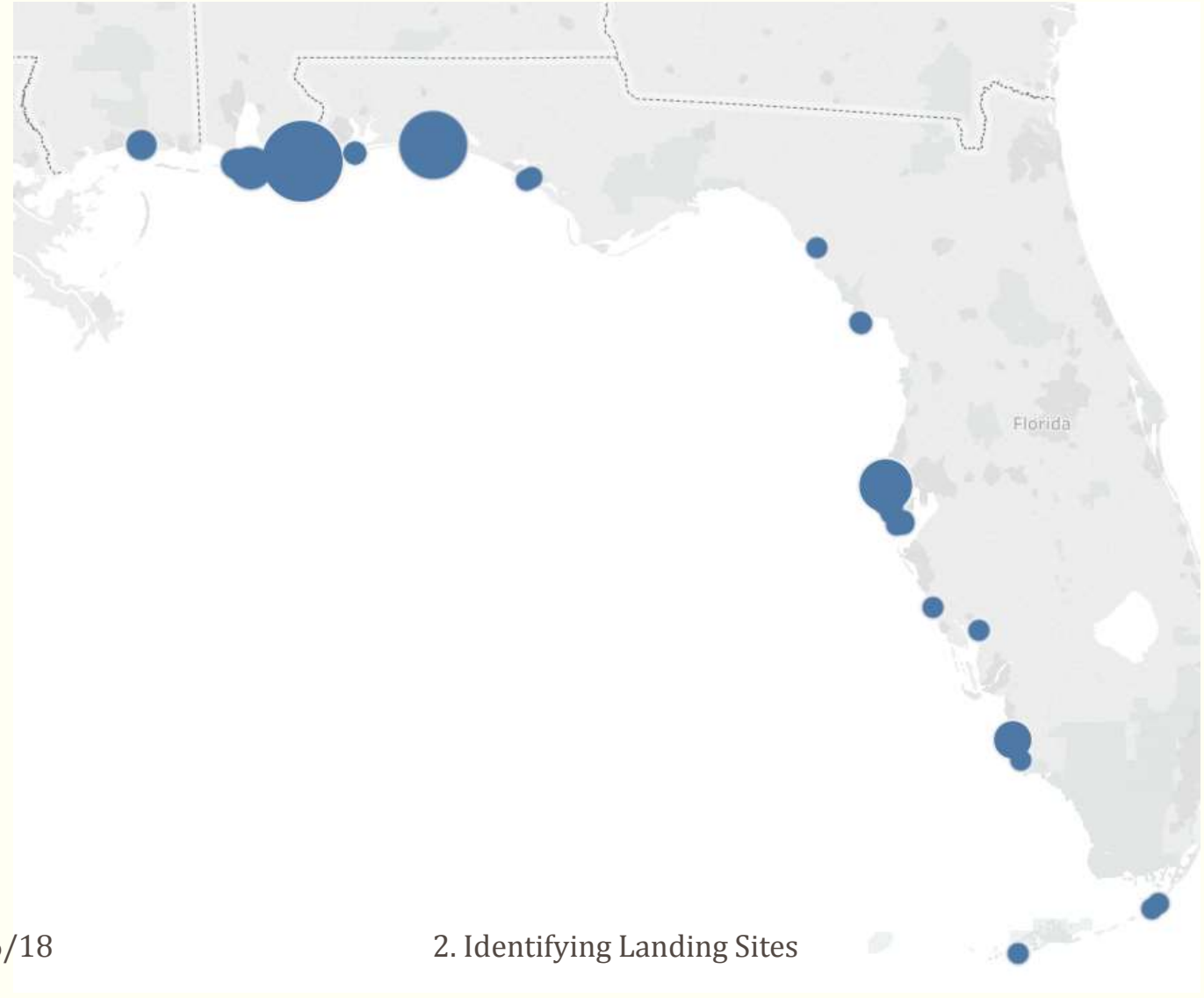
- We found a weighted average of these locations.
- This also let us check how far the vessel is from the interview site.



What are Some Errors we Encountered?

- Our default landing site for a vessel was one that was already identified from APAIS interviews.
- However, from time to time, APAIS interview sites were inaccurate, or “mislabeled” in APAIS data.

Where were the Landing Sites?



3. IDENTIFYING TRIPS

How do we Identify Basic Trips?

- Simply put, a "trip" is when a vessel leaves their designated landing site and returns between 3 and 16 hours.
- A trip "starts" at t_0 when the location at t_1 is sufficiently far from t_0's location. It "ends" at t_n when the vessel is back at the landing site.
- Additional information about the trip is also kept and used – for example, the maximum distance from the landing site and the corresponding location and time, the average distance from shore, and the length of the trip.

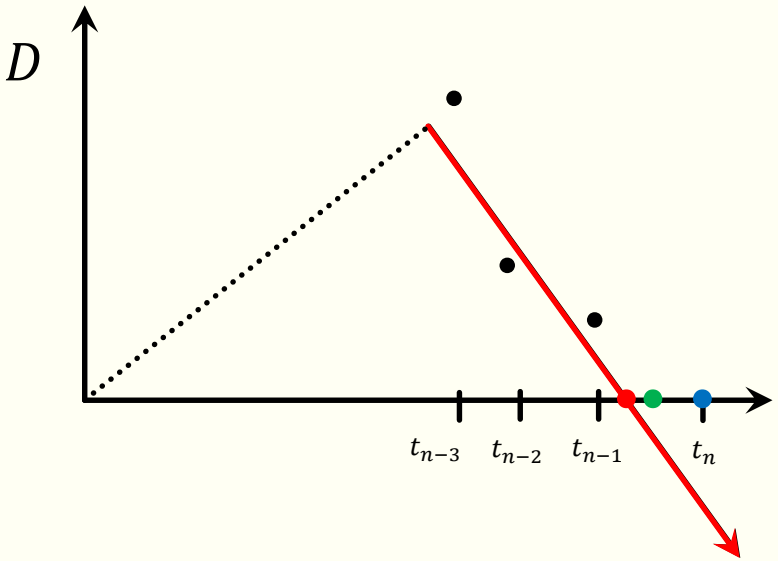

Tallahassee
Jacksonville
Orlando
Florida
Tampa

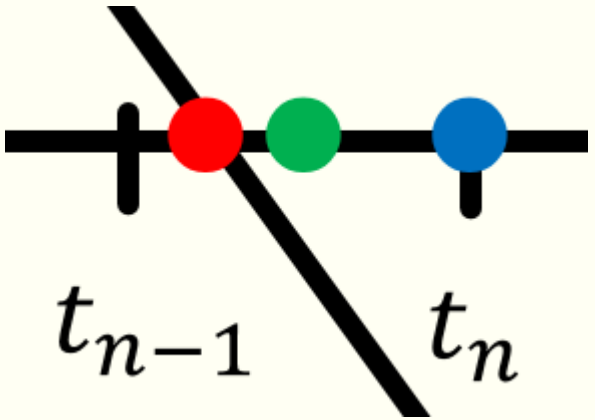
How do we Re-Estimate Arrival Time?

- $t^* = (t_{n-1} + t_n)/2$
- $s_{n-k} = \frac{Geodist(G_{n-k-1}, G_{n-k})}{t_{n-k-1} + t_{n-k}}$
- $\hat{s} = \frac{1}{m}\sum_{\{i \in M\}} s_i$

  $M = \{\text{returning obs}\}$
- $t' = \frac{s_n}{\hat{s}}(t_n - t_{n-1}) + t_{n-1}$
- So, $t_n > t^*, t'$
- If $t' < t^*$, then $\hat{t} = \frac{t' + t^*}{2}$,

  else $\hat{t} = \frac{t' + t_n}{2}$





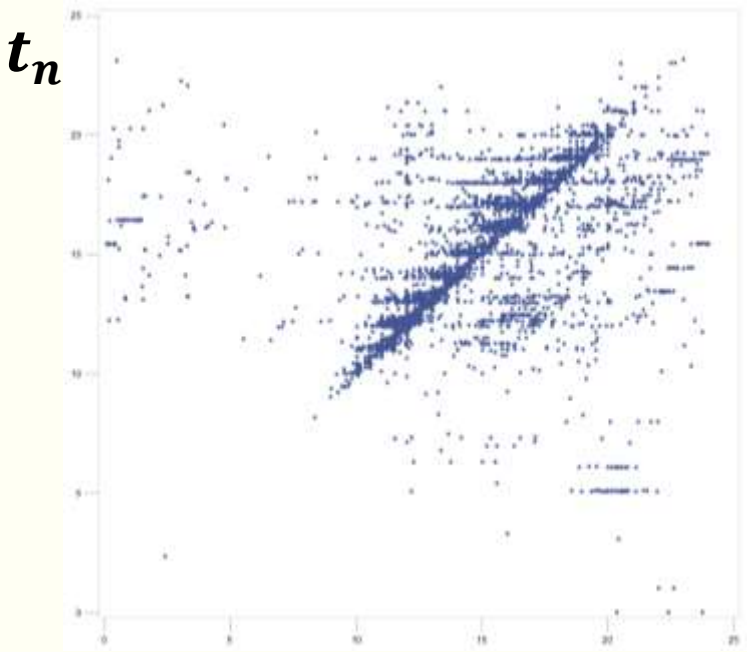
4. BENEFITS OF THE LANDING SITE AND TRIP DATA

Matching Trips, ELB Reports, & APAIS Interviews

- First, ELB Reports are sorted by Vessel ID, Date, and Return Time and assigned a trip # for the day. GPS trips and APAIS interviews, similarly.
- Then, GPS Trips are matched to ELB Reports by Vessel ID, Date, and Trip # lining up to minimize (Return Time $- t_n$).
- Finally, ELB Reports are matched to APAIS interviews by matching on Vessel ID, Date, and trip #, again minimizing Estimated return time $-$ Interview time.

Estimated Arrival Time Impact

t_n vs Arrival Times



- Estimating arrival time increased the number of matches by 15%, mostly in reported "$depart = return$" cases.

Were there Captains that Didn't Report?

- About 12,000 GPS trips were produced while only ~6,000 ELB Reports were produced – as few as half of trips went unreported.

Assumption Validation (Public vs Private)

- The Total Catch estimator is a capture-recapture estimator of the form $\frac{n_1 n_2}{m}$. It assumes that
  1. $RR_{public} = RR_{private}$
  2. $CPUE_{public} = CPUE_{private}$
  3. $CPUE^*_{public} = CPUE^*_{private}$
- 1 & 2 can be validated using the GPS data! (and otherwise could not have been validated).

Assumption Validation (Public vs Private)

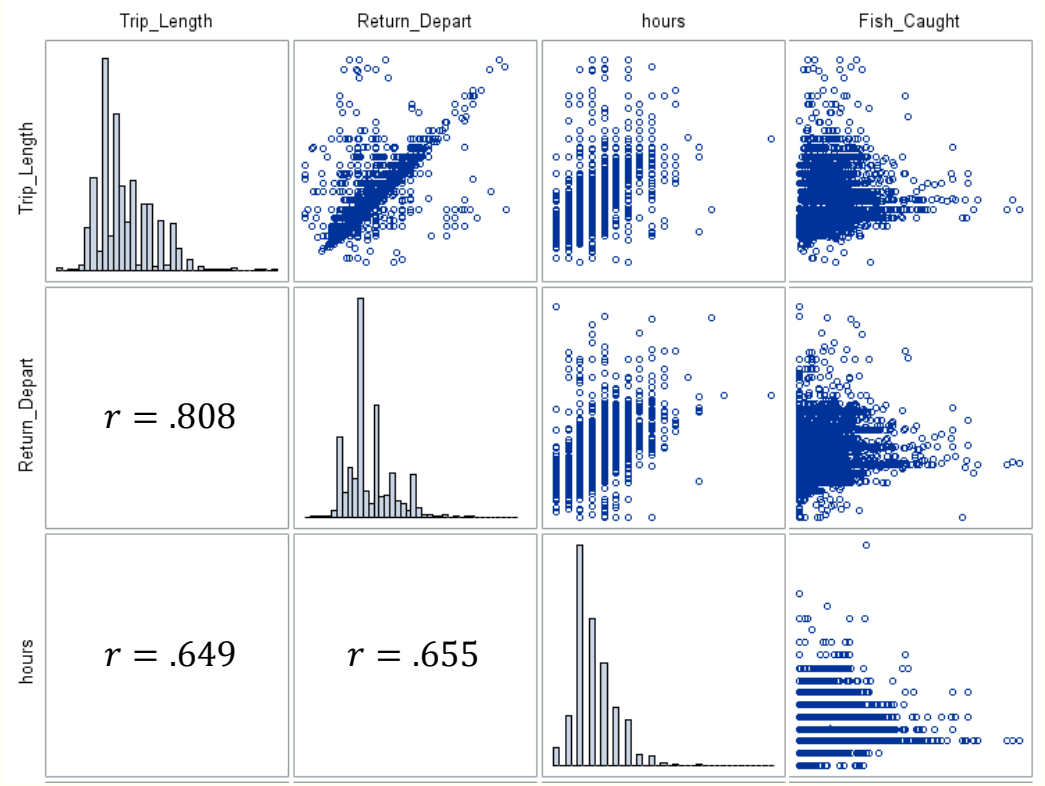
- The Total Catch estimator is a capture-recapture estimator of the form $\frac{n_1 n_2}{m}$. It assumes that
  1. $RR_{public} = RR_{private}$
  2. $CPUE_{public} = CPUE_{private}$
- Let $\lambda_X = \frac{X_{private}}{X_{public}}$ and $\pi_{\bar{F}} = \frac{\sum_i \mathbb{I}[i \in private]}{n}$
- $RB(\hat{C}) = \frac{\pi_{\bar{F}}(\lambda_{RR}\lambda_{CPUE^*} - \lambda_{CPUE})}{1 + \pi_{\bar{F}}(\lambda_{CPUE} - 1)}$

Suggestions for Improvement

- Captains asked for a list of landing sites with exact GPS coordinates, or that they have the Thorium device produce a special position report indicating a landing site whenever a captain goes to a "new" landing site.
- More specific captain training on what a "Return time" is and **how to report** it.
- **Real-time edit check** to prevent unreasonable values or inconsistent values.

How does Trip Duration Relate to Fish Caught?

- Dropped bad values.
- GPS: Trip_Length
- ELB Δ: Return−Depart
- ELB: Hours [Spent 🐠]



- Fish caught correlation $\boldsymbol{r} = (.152, .197, .121)$

5. APPENDIX

Thanks; Contact Info

Thanks to BigSurv18 for funding my travel!

This research was funded by **CLS America** under contract **#PSM00068** and federal funds under award **NA15NMF4540082** from **NOAA Fisheries**, U.S. Department of Commerce.

Me:
Ryan McShane rmcshane@smu.edu

Future project questions:
Dr. Lynne Stokes slstokes@smu.edu

Other SMURF members: Ben Williams, Shalima Zalsha, Alan Elliott, Mo Chen, Bingchen Liu