

Gulf of Mexico Shrimp Effort Estimation Workshop

February 22-23, 2023

Tampa, FL

History

Electronic Logbook (ELB) devices were originally developed by LGL in 2004 as a position logging system for commercial shrimping vessels in the Gulf of Mexico, with the goal of more accurately estimating spatial patterns of trawling effort than those collected by port agents. These devices record vessel location at 10 minute intervals using GPS, and the resulting speeds are used to identify potential vessel trawling activity. From 2004-2013, data from memory chips on these devices were collected and processed by LGL, and total fleet effort was estimated using LGL code. In mid-2013, these responsibilities were transferred to NMFS, where James Primrose was responsible for data management and calculation of effort estimates using a modified version of the original LGL code.

In 2014, cellular Electronic Logbook devices (cELB) were implemented, in which positional data are automatically transmitted back to NMFS servers through the cellular network, as opposed to manual retrieval of memory chips. In early 2014, NMFS selected 500 Gulf of Mexico Shrimp Permit (SPGM) owners using a spatially stratified random sampling method weighted by landings in prior season to participate in the cELB program. Data from these devices are stored in a Galveston Oracle database and were used to generate total fleet effort estimates using the modified LGL code through the 2019 fishing season. In the year 2020, data were received from 450 cELB devices encompassing nearly 13 million pings, 363 of which were identified to have fishing activity. A yearly breakdown of the number of vessels exhibiting ELB fishing activity between 2014 and 2021 is provided below.

Year	2014	2015	2016	2017	2018	2019	2020	2021
Active ELB vessels	444	433	462	452	495	449	363	257

Motivation

Development of a new effort estimation method began in late-2021, after the original code was unable to be executed in a timely manner to generate 2020 estimates. Goals in developing this new method were to produce robust effort estimates with:

- Simplified assumptions
- Increased transparency around input parameters
- More complete use of the cELB data
- Greatly simplified and streamlined R code

This simplified code and logic, resulting from a combination of advances in the R programming language and simplified and unified data sources since the inception of the ELB program, will allow code to be executed more reliably, both in the short-term and by future users.

Summary of Changes

- Effort classification
 - Distances are calculated using the Vincenty ellipsoid method (R *geosphere* package) rather than a Euclidean metric with rough fixed parameters. This results in more accurate distances that take the curvature of the earth into account.
 - A 1-minute resolution NOAA GOM bathymetric grid (R *marmap* package) is used to filter out data at depths too deep for shrimping activity. The default setting filters out any pings occurring at greater than 2,500 feet, but this is an adjustable parameter. The estimated maximum biological depth for royal red shrimp along the continental shelf is 2,395 feet (Perry and Larson 2004).
 - An updated GOM shapefile with higher resolution fathom delineations is now in use. This shapefile now encompasses the entire Gulf EEZ rather than only extending to the shelf edge.
 - The upper fishing speed threshold is calculated using a Gaussian mixture distribution on the observed data rather than using pre-set fixed numbers. This makes the algorithm more robust to changes in fishing dynamics that may occur over time. Starting values that correspond to the expected effort distribution are provided to protect against unrealistic classifications.
 - Towing activity must occur for a one-hour minimum to be classified as effort. This protects against including tri-net tows and other false positives.
- Scaling to total fleet
 - Scaling of effort is now done using landings at aggregate combinations of time (3 quadrimesters) and area (4 zones) in a survey-design framework rather than attempting to match individual trips. This ensures all cELB recorded effort is used to estimate spatial patterns and in the calculation of total effort, rather than only using those trips that are able to be matched to trip ticket landings (~60%). This results in more complete use of the data and fewer assumptions related to the trip matching procedure, given no key exists to truly be able to match trips to landings.

- Code
 - Code for generating annual effort estimates has been greatly simplified and modernized into a single streamlined R script (<500 lines, including all data queries, figures, and comments). The result is a user-friendly product that can be more easily interpreted, executed, and troubleshooted, while requiring the user to know only a single programming language.
 - All numeric decisions (thresholds, etc.) are transparent as function arguments (10) and can be modified as appropriate.
 - There are no randomized components to the code. That is, results are consistent between runs without needing to set a seed.

Assumptions

The following basic assumptions are required to obtain accurate estimates of total effort, given non-universal cELB coverage of the fleet:

1. cELB devices are capturing all fishing activity, and are powered on for the full extent of vessel activity per federal regulations. This assumption is generally supported, but a comparison against observer reported effort suggests cases exist when cELB devices may not be recording.
2. There is no systematic bias in effort classification. That is, there is an equal chance of false-positives and false-negatives. A comparison of cELB classified effort with observer recorded effort generally supports this assumption.
3. The spatial distribution of cELB vessels is representative of the total fleet within strata. There is support for this assumption given the randomized nature of the original cELB selection, however, due to changes to the fleet and vessels dropping from the sample the original selection of vessels may not remain representative.
4. CPUE of vessels with cELBs on board is representative of the total fleet. This is a necessary assumption for using landings to scale up cELB effort.
5. Reporting of landings is similar between vessels with and without cELBs. That is, one group is no more or less likely than the other to completely and accurately report landings. This is a necessary assumption for using landings at an aggregate level to scale up cELB effort.

*Note assumptions 3-5 are required due to ELB devices being present on only a subset of the fleet, and could be eliminated with universal VMS.

Data Sources

All tabular data required to generate fleet-wide effort estimates are stored in the following four tables within the SEFSC Oracle database:

1. Raw cELB pings (elb.elb_data@elb_dblk)

Complete data range from 2014-present, encompassing between approximately 12-15 million rows (10-minute pings) annually. With the termination of the 3G network in 2020, data have been uploaded to this table manually from chips beginning in 2021.

Relevant fields include: box number, latitude, longitude, year, month, day, hour, minute, and second.

2. Vessel assignment table (elb_obj.elb_assignments)

This table is used to assign a vessel ID (either US Coast Guard or otherwise state number) to the raw cELB data. The “serial” field in this table is used to join to the “box number” field in the cELB data. Other relevant fields include: vsbn (vessel ID), date installed, date removed, and status (current state of cELB device).

3. Trip ticket landings (scdw.v_fac_landing_prior_year@secdw_dblk)

This is a SEFSC view of complete dealer trip ticket landings for prior years based on data in the GSMFC database. Data are refreshed on a weekly basis to accommodate any updates or corrections to existing data. Relevant fields include: landed_date, fishing_area_sub_area_num (area and subarea of landings on 1-21 GOM trip ticket grid), landing_state_fips, landing_county_fips, species_itis, gear_code, vessel_official_number, and gutted weight (in this case meaning shrimp tail weight). These data are used to summarize landings according to vessel cELB status within time/area blocks.

4. Vessel permit table (scdw.dim_vessel_permit@secdw_dblk)

This table is maintained by SERO and identifies which federal permits are associated with a particular vessel through time. This allows landings to be separated into state and federal categories based on if the vessel ID associated with the landings was in possession of a valid SPGM permit. Relevant fields include: official_number, fishery_code, permit_effective_date, permit_terminated_date, and permit_status.

Additional static inputs include:

1. Gulf of Mexico shapefile with 1-21 trip ticket area grid and 10 and 30 fathom delineations (Courtesy Jo Williams, SEFSC)
2. Gulf of Mexico bathymetric grid (imported from *marmap* package)

Code

All code to produce the PDF report with annual effort estimates, figures and tables is contained within the script 'effort_scaled.R'. This code is executed through the Rmarkdown file 'GOM_shrimp_effort_report.Rmd'. The user simply selects "Knit with Parameters" from the dropdown menu next to the Knit button in the GUI, enters the year for which estimates are desired in the Shiny interface, and clicks Knit to run the code and produce the report.

Required R packages to execute code include: ROracle, dplyr, tidyr, marmap, geosphere, sf, mixtools, survey; ggplot2, scales (to produce figures); knitr, shiny (for report generation)

Function parameters (and defaults) are as follows:

- ❖ ping.hrs.min: minimum time in hours between pings required to attempt to classify vessel state (default 59/3600 hours, i.e., at least one minute)
- ❖ ping.hrs.max: maximum time in hours between pings at which an initial attempt will be made to classify vessel state (default 0.5 hours)
- ❖ knots.min: minimum speed in knots between pings which has the potential to be classified as towing effort. All speeds below this threshold will be considered stopped/idling (default 1.9 knots)
- ❖ knots.max: upper speed bound in knots to be considered a steaming activity rather than an erroneous data point. That is, this value cuts off the right tail of the distribution of steaming speeds (default 11.5 knots)
- ❖ tow.hours.min: sets the minimum amount of consecutive time in hours a vessel must fall in the speed profile of towing activity to be considered a true tow. To reduce false positives, speeds that otherwise fall within the range of towing activity will only be counted as effort if they meet this threshold (default 1 hour)
- ❖ trip.eff.hrs.min: sets the minimum amount of towing time in hours on a "trip" to be considered true effort. This serves as a second pass to reduce false positives that are not filtered out with tow.hours.min (default 2 hours)
- ❖ fill.hrs.max: maximum amount of time in hours between consecutive pings to attempt to fill in missing vessel activity for gaps in ELB data transmission (default 8 hours)
- ❖ trip.hrs.brk: minimum amount of time in hours that must elapse between the end of one tow and the start of the next to be considered a new "trip" (default 24 hours)
- ❖ min.transition.hrs: minimum amount of time in hours that must elapse between the end of one tow and the start of another to be considered a new tow (default $\frac{1}{6}$ hour, i.e., 10 minutes)
- ❖ max.depth.ft: sets maximum depth in feet, according to 1 minute GOM bathymetric grid, for vessel activity to be considered true effort. Any vessel activity occurring in deeper waters will not be considered effort (default 2,500 feet).

The code to estimate effort can be broken down into two primary steps: initial classification of effort from raw cELB data and scaling of cELB effort to the total fleet. These steps are detailed below.

Effort classification

1. Data input

Raw cELB data with timestamp, latitude, and longitude (1) for a specified calendar year are first pulled into R from the Oracle database using the *ROracle* package (version 1.3-1.1). These data are joined to the vessel assignment table (2) by box number, allowing vessel IDs to be associated with the correct boxes at a given time. At this stage, data are also filtered down by box status to remove any demo, development, or test data.

2. Data filtering

Next, coordinates are converted to decimal degrees and raw pings are filtered to the extent of the Gulf of Mexico EEZ shapefile based on lat/long using the `st_join` function in the *sf* package, such that any data falling outside the region (e.g., South Atlantic) are removed. Additionally, any pings occurring deeper than the maximum specified depth contour (default 2,500 ft.) are removed based on the GOM bathymetric grid (*marmap* package). Any duplicate rows (i.e., identical timestamp, lat/long, vessel) are also removed in this step.

3. Initial speed calculation

Filtered data are next sorted in order of timestamp within each vessel ID, and the `distVincetyEllipsoid` function in the *geosphere* package is used to calculate distances (in nautical miles) between consecutive lat/long points. Speed in knots between consecutive pings are calculated by dividing these distances by the elapsed time in hours. To prevent unrealistic values, only consecutive pings with times between `ping.hrs.min` (~1 minute) and `ping.hrs.max` (30 minutes) are used in this initial calculation to generate a distribution of vessel speeds. The vast majority of pings occur at 10 minute intervals, so this is not an issue in most cases but serves as a safeguard to prevent extreme values from arising. Any speeds still falling above the realistic range of vessel travel (`knots.max`, default 11.5), likely due to bad GPS points, are also removed from the distribution.

4. Classification of vessels speeds

Vessel speeds as calculated in step 3 (above a minimum stopped/idling threshold, default 1.9 knots) tend to clearly follow a bimodal distribution, representing a mixture of the two Normal distributions that result from the distinct fishing and steaming speed footprints. Therefore, a 2-component Normal mixture model (R package *mixtools*) is fit to this distribution of speeds to help identify an optimal cutoff between fishing and steaming activity. This model-based calculation can be useful for identifying subtle changes in fleet behavior over time as opposed to

using a fixed threshold. Starting values of expected means (3, 8), standard deviations (0.5, 1.5), and densities (0.8, 0.2) are provided for the distributions of fishing and steaming speeds, respectively, ensuring that the model converges to a reasonable solution. For 2014-2020 cELB data, the resulting threshold comes out consistently very close to 3.8 knots, with little variation between years (± 0.05 knots).

Upon classification of speeds according to the logic above, values that were unable to be accurately classified initially due to pings occurring below the minimum time threshold are filled in with the most recent known vessel state. In cases where the elapsed time between pings is greater than 30 minutes, classification of vessel state is only attempted for periods of up to a specified maximum (fill.hrs.max, default 8 hours), and only if the calculated vessel speed during that interval falls within the range of possible speeds (< 11.5 knots). Reducing fill.hrs.max below 8 was observed to have very limited impact on the final result, but is a necessary parameter to handle gaps in ELB transmission. For 2014-2020, this logic typically resulted in excess of 99.8% of data points receiving a classification.

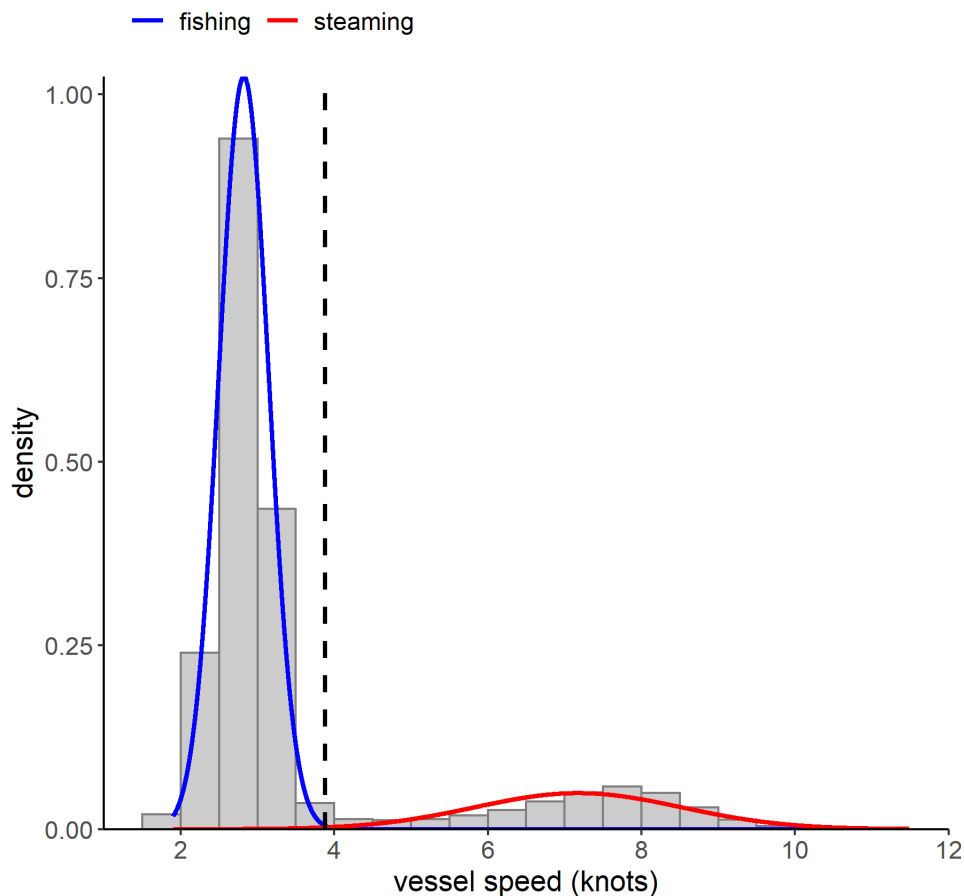


Figure 1: Distribution of cELB vessel speeds (2020) with fitted Gaussian mixture distribution.

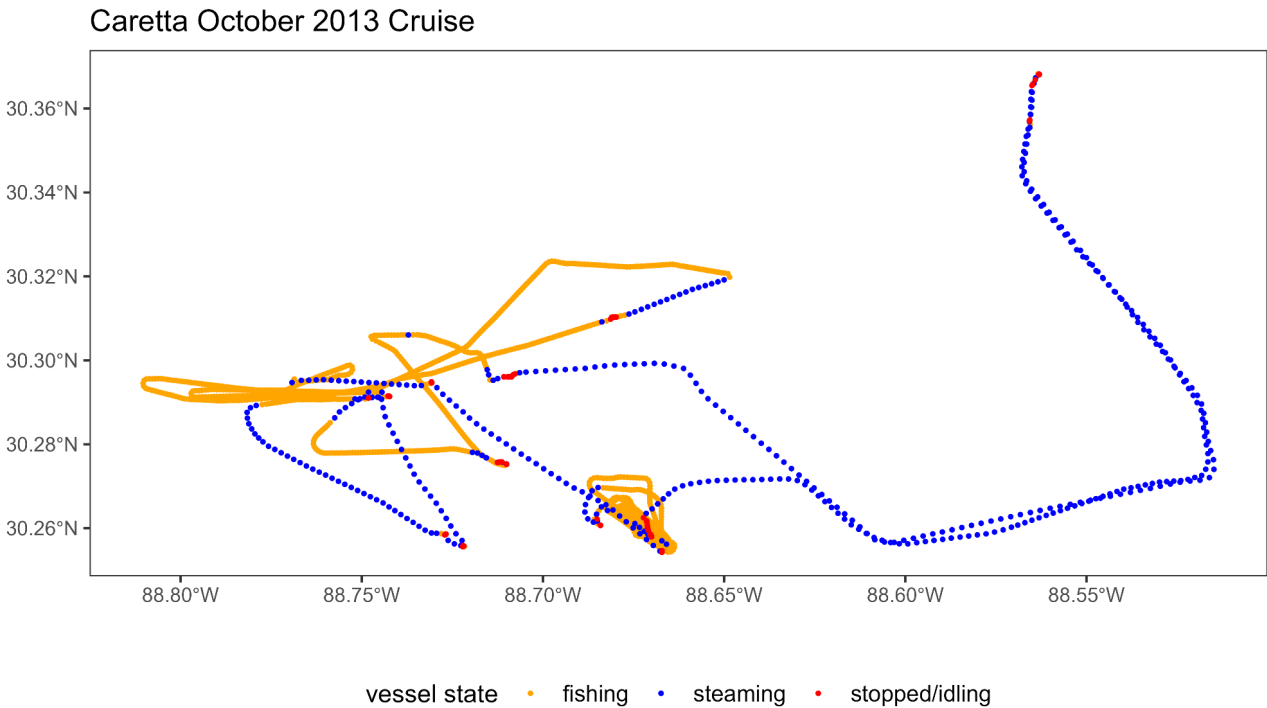


Figure 2: Example vessel track with classified vessel states.

5. Isolation of fishing activity

For the final stage in calculating total cELB effort, a series of steps are taken to remove potential false positive pings based on surrounding vessel activity. The *minimum* break time between tows, or changes between fishing and non-fishing activity, is set to 10 minutes by default (min.transition.hrs). This helps remove changes in vessel state that occur in an unrealistically short interval, likely arising from irregularities in the data (e.g., pings occurring at shorter than 10 minute intervals). A minimum transition time between tows of 10 minutes is also consistent with what is seen in the shrimp observer data. Once tows are defined in this way, only consecutive periods of activity classified as fishing effort falling at or above 1 hour (tow.hrs.min) are kept as true effort. This helps filter much of the false positives arising from vessels that happen to be traveling at towing speeds but are not truly towing. As a final pass at removing false positives, periods of effort that sum to less than a total of two hours (trip.eff.hrs.min) within periods until the duration between consecutive tows on a vessel exceeds 24 hours (trip.hrs.brk) are removed, thought to be too little effort to occur within what likely constitutes a trip. These thresholds are apparent upon examining the resulting distributions of tow times, indicating these periods classified as fishing speeds are likely artifacts of the data rather than true fishing activity.

Once effort data have been isolated from the complete cELB dataset, a spatial join with the shapefile below (Figure 3) is performed to assign a 1-21 statistical zone and depth zone (0-10, 10-30, 30+) to each ping (row).

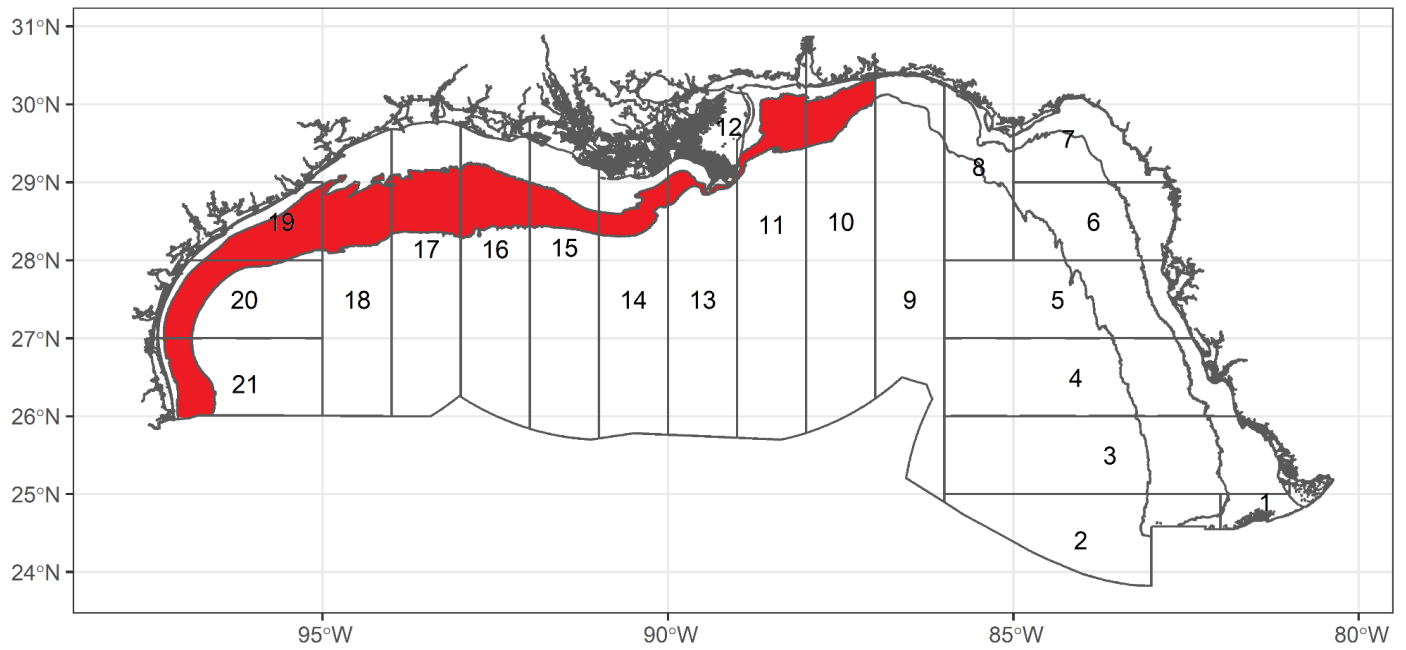


Figure 3: GOM trip ticket statistical grid, with offshore lines delineating 10 and 30 fathom boundaries. Red snapper restricted area is shaded in red.

percentile ■ 50 ■ 95

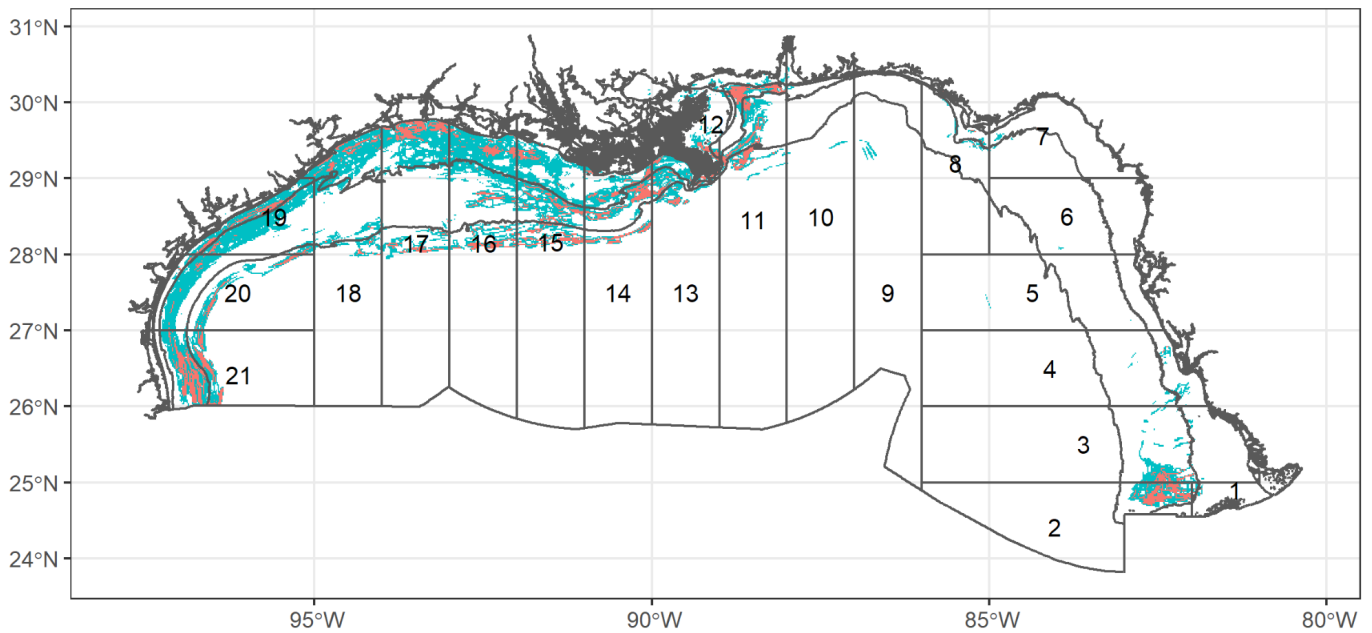


Figure 4: cELB estimated effort (2018-2020) classified by percentile (top 50% of effort falls in red areas, top 95% falls in combination of red and blue).

Scaling to total fleet

A survey-weighting approach based on aggregated trip-ticket reported landings (see Data Sources, #3) is used to scale up estimated effort from vessels with cELB devices to that of the total offshore fleet. Landings have been observed to correlate particularly strongly with effort in the GOM shrimp fishery ($r = 0.94$), especially in relation to reef fisheries (Chollett, unpubl.).

1. Stratum definitions

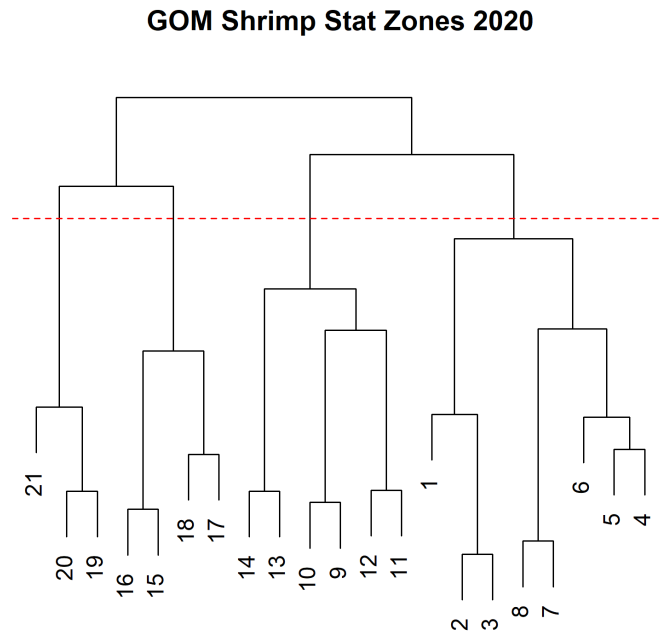


Figure 5: Dendrogram from hierarchical cluster analysis of spatial extent of trips (2020).

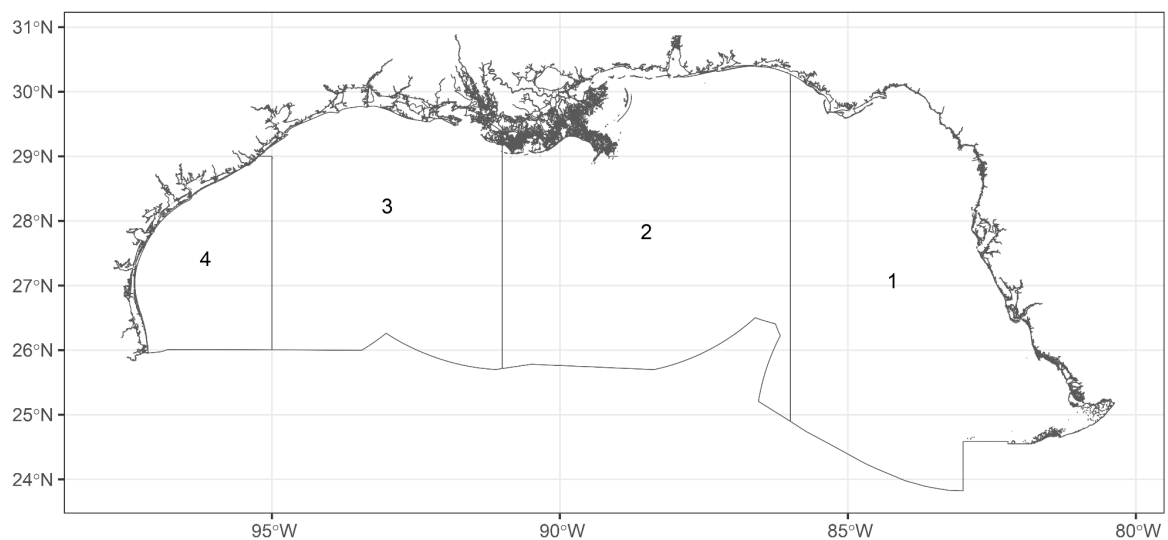


Figure 6: Map of resulting 4 aggregate areas (stat zones: 1: 1-8, 2: 9-14, 3: 15-18, 4: 19-21).

Strata are defined by the combination of quadrimester (months 1-4, 5-8, and 9-12) and area. As shrimp trips tend to be spatially extensive and span multiple statistical zones, a suitable aggregation of zones is needed to define broader areas used for scaling. To help identify these areas quantitatively, a hierarchical cluster analysis using Ward's method was performed to identify patterns in the spatial extent of trips based on an indicator matrix with trips as rows and statistical zones as columns based on cELB activity. The result for 2020 is presented in Figure 5. When the tree is cut to obtain four branches (dashed red line), the aggregations result in area 1: 1-8, area 2: 9-14, area 3: 15-18, area 4: 19-21 (Figure 6). This results in a total of 3 quadrimester x 4 area = 12 time/area strata.

2. Sampling weight estimation

Sampling weights (i.e., scaling factors) for effort in each stratum are estimated using the ratio of total offshore landings (defined by reported subarea and bottom otter trawl gear type) among vessels holding a federal SPGM permit to offshore landings among the subset of vessels with fishing activity identified by a cELB device in a given calendar year (Eq. 1). Finite population correction factors for variance calculations are computed similarly as the number of cELB vessels with landings over the number of total vessels with landings (Eq. 2).

$$w_{area/quad} = total\ landings_{area/quad} / ELB\ landings_{area/quad} \quad (1)$$

$$fpc_{area/quad} = ELB\ vessels_{area/quad} / total\ vessels_{area/quad} \quad (2)$$

For 2020, these values were as follows:

<u>quad</u>	<u>area</u>	<u>w</u>	<u>fpc</u>
1	1	2.01	0.465
1	2	2.24	0.689
1	3	2.10	0.631
1	4	1.81	0.641
2	1	2.68	0.438
2	2	2.14	0.573
2	3	1.89	0.605
2	4	1.89	0.638
3	1	2.50	0.376
3	2	2.36	0.543
3	3	1.96	0.608
3	4	1.80	0.621

3. Total effort computation

A stratified 1-stage cluster sampling design is used to obtain estimates of effort totals and variances using the sampling weights as calculated in Step 2. Estimated cELB effort for each

vessel is aggregated at the quadrimester/area level and the sampling weights are applied accordingly to obtain totals (Eq. 3).

$$\text{total federally fleet effort} = \sum cELB \text{ effort}_{\text{area/quad}} \times W_{\text{area/quad}} \quad (3)$$

Variance of the estimated total is calculated using the standard formula for a stratified one-stage cluster sample (Woodruff 1971), with vessels representing primary sampling units. In the following formula, y represents vessel effort, w represents the sampling weight (Eq. 1), f represents the finite population correction (Eq. 2), n represents the number of PSUs (vessels) in a strata, and h denotes a given quadrimester/area stratum.

$$\hat{V}(\hat{Y}) = \sum_{h=1}^H \hat{V}_h(\hat{Y})$$

where, if $n_h > 1$, then

$$\hat{V}_h(\hat{Y}) = \frac{n_h(1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi\cdot} - \bar{y}_{h\cdot\cdot})^2$$

$$y_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$$

$$\bar{y}_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} y_{hi\cdot} \right) / n_h$$

This design is implemented in the R *survey* package (Lumley 2020) with the following code, where “cell” is a concatenation of quadrimester and area:

```
design = box_tow_effort_grp %>%
  group_by(quadrimester, area,
    GOM_region = ifelse(StatZone %in% 10:21, "West", "East"), DepZone, VSBN) %>%
  summarise(days = sum(days)) %>% mutate(cell = paste(quadrimester, area)) %>%
  inner_join(sample_wgts, by = c("quadrimester" = "quadrimester_land", "area")) %>%
  svydesign(~VSBN, strata = ~cell, weights = ~w, fpc = ~fpc, data = ., check.strata = FALSE)
```

The following code can be used to produce estimates of totals and variances within specified domains (here quadrimester/area):

```
svyby(~days, ~quadrimester + area, design = ., svytotal, vartype = "var", keep.names = FALSE)
```

Further details of statistical computations can be found here:

https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_surveymeans_details06.htm#statug_surveymeans_vartotaldetails

Confidence intervals around estimates (in units of 24 hour days) are calculated from the resulting standard errors (i.e., square root of the variances) multiplied by the desired quantile of a t-distribution with n-h degrees of freedom (number of vessels minus number of strata).

4. Offshore state landings

Landings reported in offshore waters (beyond the COLREG line) from vessels not in possession of a federal SPGM permit are assumed to occur in state waters (i.e, 0-3 or 0-9 nm offshore). These landings typically constitute a small percentage of the offshore total (~3.5% in 2020). For purposes of effort scaling, these landings are allocated among depths in proportion to the area of offshore state water encompassed by each depth zone within each of the four aggregate areas (with the majority falling in the 0-10 fathom zone). For example, to allocate within a given area:

$$state\ landings_{area/depth} = state\ landings_{area} * state\ water\ area_{area/depth} / \sum state\ water\ area_{area} \quad (4)$$

As the cELB sampling frame consists only of vessels with federal SPGM permits, the total federal fleet effort and variance estimates as calculated in the previous section are adjusted upward as follows:

$$state\ correction = (federal\ landings + state\ landings) / federal\ landings \quad (5)$$

$$total\ effort = total\ federal\ effort * state\ correction \quad (6)$$

$$var(total\ effort) = var(total\ federal\ effort) * state\ correction^2 \quad (7)$$

5. Allocation of total landings and effort among depth zones

Since the original distribution of cELB effort data is preserved, scaled effort and landings can be allocated among depth zones and statistical zones (or other custom groupings) proportionally to the observed distribution of cELB effort. For example, to allocate landings by depth zone within each area and quadrimester:

$$landings_{area/quad/depth} = landings_{area/quad} * effort_{area/quad/depth} / \sum effort_{area/quad} \quad (8)$$

For allocation of landings, this assumes constant offshore CPUE among the depth zones, as depth information is not reported on trip tickets, though this assumption seems to be reasonably well met based on analyses conducted by SSRG.

CPUE calculations can then proceed accordingly for given times/areas/depths:

$$CPUE = \sum landings_{region/year/depth} / \sum effort_{region/year/depth} \quad (9)$$

Results

Table 1: 2020 Gulf-wide Offshore Estimates

subarea	tow days	lower	upper	landings	CPUE
offshore	56,918	55,414	58,422	67,513,636	1,186

Table 2: 2020 Stratified Offshore Effort Estimates

GOM region	depth zone	tow days	lower	upper	landings	CPUE
East	Offshore 0-10 fm	695	564	826	892,354	1,284
East	Offshore 10-30 fm	5,225	4,759	5,691	7,250,556	1,388
East	Offshore 30+ fm	123	66	179	104,504	852
West	Offshore 0-10 fm	19,497	18,471	20,523	23,258,890	1,193
West	Offshore 10-30 fm	18,898	18,127	19,669	21,715,169	1,149
West	Offshore 30+ fm	12,480	11,757	13,204	14,292,163	1,145

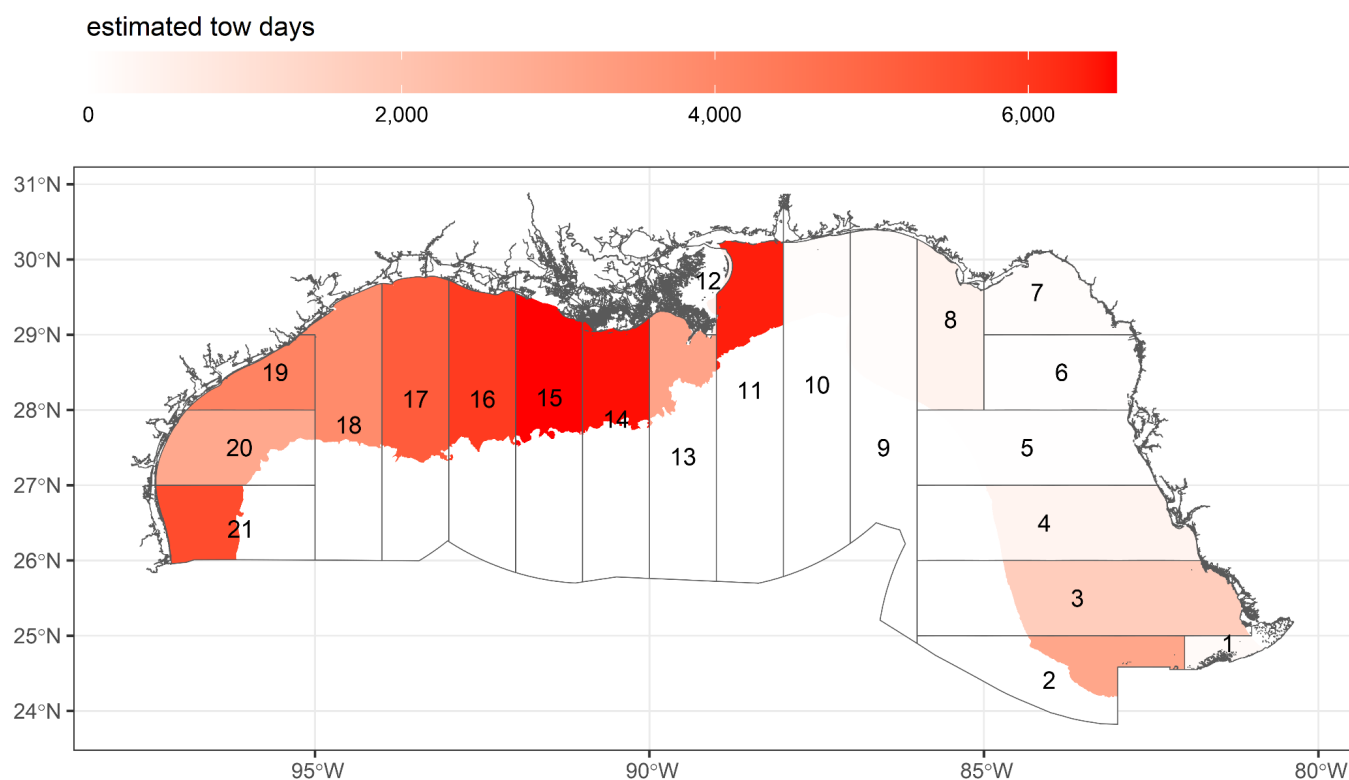


Figure 7: Total estimated 24 hour tow days by statistical zone (2020).



Figure 8: Comparisons of LGL and SEFSC algorithm results for total offshore effort (A) and effort within the red snapper restricted area (defined as depths between 10 and 30 fathoms in the Western GOM, areas 10-21). (B).

Limitations

The methods described here are only as good as the data they depend on. Incomplete ELB data (e.g., boxes not functioning or turned on for complete trips) or missing/inaccurate landings reports have the potential to create bias in estimates. Additionally, while data collection depends on the receipt of physical memory chips, we equally depend on all vessels with ELB units to submit complete records in order to produce accurate estimates.