REVIEW of GRSC Report by Christman, 5 April 2021

Review of Stunz, G. W., W. F. Patterson III, S. P. Powers, J. H. Cowan, Jr., J. R. Rooker, R. A. Ahrens, K. Boswell, L. Carleton, M. Catalano, J. M. Drymon, J. Hoenig, R. Leaf, V. Lecours, S. Murawski, D. Portnoy, E. Saillant, L. S. Stokes., and R. J. D. Wells. 2021. Estimating the Absolute Abundance of Age-2+ Red Snapper (Lutjanus campechanus) in the U.S. Gulf of Mexico. Mississippi-Alabama Sea Grant Consortium, NOAA Sea Grant. 303 pages.

(File entitled "GRSC Report for GMFMC SSC_filesizereduced_03152021.pdf" received 15 March 2021)

By Mary C. Christman, Courtesy Professor in the Departments of Statistics and Biology at the University of Florida, Gainesville, FL

Date: 5 April 2021

In this review, I address the statistics aspects of the study design and estimation procedures. Generally I do not consider the field methods (sampling gear types) used for collecting, processing or calibrating the data to standard units unless these methods impact the statistical analyses. Nor can I address the assumptions that are made throughout concerning the biology, behavior or phenology of red snapper. Hence, I will not assess whether the non-statistical assumptions are appropriate, the size or direction of the potential biases in the data, and whether the results can be apportioned among age-specific composition.

One point I would like to make before discussing the statistical aspects of the study concerns the structure of the report itself. If possible, the final report should include the additional information provided in the presentations or from the question and answer periods during the review meeting and which was missing in the draft report. For example, there should be a table near the beginning of the report that includes a list of every stratum, and for each stratum: the technologies used for data collection, the sampling design actually used, and the final sample size used in the estimation procedures. Where the numbers of samples collected is larger than the number used in the analyses, an explanation of why some data were not used should be noted. Note that when the design was multi-stage cluster sampling, the stratum sample size should include the sample size for each stage of clustering. In addition, if the number of observations used in the estimation is a subset of the implemented field sample sizes, then that should be noted and explained as well. There should also be short explanations of the reasons for the deviations from the planned design, including explanations for the development of post-strata. The design-based estimators that were used in each stratum should be provided in detail (probably in an appendix) for each sampling design that was implemented. The current report has most of these details, but the information is scattered throughout the document and difficult to infer in some instances. The explanations for the information would inform a reader as to the actual analyses are also sometimes missing. In addition, the information provided about the sampling should be consistent (see Table 1 at end of this report).

One other point is that any estimates of abundance should be accompanied by some measure of their variability (uncertainty). For the lay audience, that is likely the endpoints in a 95% confidence interval; for others, the standard error and sample sizes would be appropriate.

Following are comments that relate to specific TORs where I felt comfortable reviewing.

1. STUDY DESIGN AND SAMPLING APPROACHES

Overall, the study covered a large area in the Gulf of Mexico appropriately. The planned implementation of stratification based on region, depth zone and Random Forest (RF)-generated zones based on probability of occurrence of red snapper sufficiently covered the spatial aspects of the study. The intended design included appropriate stratum sample sizes based on optimal allocation assuming that data previously collected in the Gulf were adequate to describe the distribution of red snapper in the Gulf and under the assumption that the CV of the strata estimates would be approximately 150%. Of course, as expected, implementation was based on a modified set of strata, different sample sizes, and did not always follow a simple random sampling design. Based on the presentations during the review meeting, the modifications appeared to have been done to optimize sampling for some of the technologies that were expensive to deploy. These changes were to be expected given the fiscal and time constraints of the study and appeared to have been performed in such a manner as to allow for appropriate statistical analyses to be done. I discuss in another TOR whether the analyses did use the implemented design correctly. An unfortunate aspect of the changes was that the sample sizes were likely not optimal for minimizing the variance of the estimate as desired and often were too small to obtain sufficient data to adequately characterize a stratum's parameters (mean, variance, total). Some important examples include the strata with a mean density of 0 and which also have very small sample sizes. In those instances, the question of whether these should have been identified as separate strata is an issue. Instead, for FL region at least, perhaps these should have been combined into adjacent strata since the decision to subdivide the FL shelf into regions led to some of these 0 mean strata.

The use of different technologies in different strata could not be avoided given the environmental variations among strata, for example, the nepheloid layer that caused difficulties with video use in the western gulf regions. This led to several unavoidable problems with the estimation procedure due to the lack of calibration among the different technologies. The report cites only one calibration study that proved inconclusive due to the lack of spatial overlap of the two methods (I base this on the figures provided by Dr. Patterson during the review meeting). If there had been detailed calibration studies in a variety of habitats, I would have been able to evaluate the question of biases and variances due to the variety of technologies used. The only calibration study provided was not sufficient to determine whether the two different methods (video vs hydroacoustics) can be interchanged.

One the other hand, the use of varying technologies does not lead to within-strata problems in that, in most strata, only one field data collection method was used. The exception to this was the overlap of hydroacoustics and C-BASS technologies in an area of the UCB off TX. I address the effect of this in another TOR.

The main question concerning the use of different technologies is whether they essentially provide the same information concerning red snapper densities and abundances such that they can be combined into a single Gulf-wide estimate. I do not believe that is in fact feasible given the different descriptions of the data collected that were provided by the scientists in the review meeting. That is, I am unsure whether the estimates based on a hydroacoustics/video methods are comparable to that obtained by video alone had video along been feasible. This goes back to the issue of the need for additional calibration. On the other hand, if the proportion of sampling

locations yielded an estimate of 0 red snapper at the site, the effect on the stratum mean density of a mismatch in the number of red snapper observed at stations given that red snapper exist at those stations may be rather small and so the estimates from different technologies could be reasonably combined. This must be countered with the determination of whether the detectability differs significantly between two technologies.

Another issue for the implementation of the sampling design was the lack of data collection in a few strata (e.g. LA, parts of the FL shelf) and the requirement to infer mean densities in those missing strata. This introduces additional variability into the estimates for abundance and should be addressed appropriately. I discuss this in another TOR.

So, to answer the questions posed in this TOR:

- **Assess the sufficiency of spatiotemporal sampling by study strata.**

It is sufficient except for the few strata that were not sampled during the study. The use of imputation/substitution of neighboring data potentially introduces bias into the estimates and definitely adds additional variability that cannot be adequately addressed without additional assumptions.

- **Does heterogeneity in sampling by strata affect estimates of absolute abundance and variance around that estimate?**

Yes, but the study could not have been performed otherwise. Both the technologies that were used and the cost/field constraints imposed by those technologies preclude any opportunity to use the same approaches in every stratum. Of more importance is 1) determining whether the two technologies have similar detectability and are comparable for the non-zero fish observations and 2) ensuring that the data are analyzed according to the actual implementation and not based on the planned design.

- **Evaluate assumptions and biases inherent to the design, and the directionality of those biases.**

One sampling issue that likely biases the results is the data collection in the western Gulf that relied on hydroacoustics over a range of depths in the UCB and a separate video survey that was conducted mostly in the deeper depths covered by the hydroacoustics. This was due to the lack of water clarity in the shallower depths but also inherently assumes that the proportion of red snapper of the appropriate size in the species composition is identical over all depths. If that were the case, then the data are unbiased and with appropriate variance but this should be checked by reviewing the distribution of red snapper by depth in other regions if appropriate, i.e. these other regions are a reasonable proxy for the density distribution by depth of red snapper.

For other aspects of the analyses and study design, I do not feel as though I can address the biology inherent in the assumptions.

- **Are sampling approaches collectively appropriate for determining an estimate of absolute abundance for red snapper in the Gulf?**

Yes, given the constraints inherent in any such study.

- **Are different sampling techniques effectively calibrated to each other for generating the absolute abundance estimate?**

Only a single calibration study was done but did not provide appropriate analyses that could be applied to the data collected. As a result it is not possible to evaluate whether the different field methods provide similar estimates and conclusions. If such studies have been done it would be useful to provide them or references to them in the final report.

- **Are the biases and limitations of each approach effectively addressed?**

Yes, they are addressed in each section of the report.

2. STATISTICS AND DATA ANALYSIS

- **Evaluate the statistical methods used to analyze the data, and to construct the absolute abundance estimate and its variance.**

In general, the approaches taken by the two independent analyses are partially correct. In the main analysis that led to Table 6 in the report, the first issue is that all data were treated as though the sampling was simple random sampling (SRS) within *a priori* strata and post-strata. It was not always collected according to SRS. There are several instances where data were collected according to a cluster design. For example, the hydroacoustics survey in the UCB in the TX region was a three-stage cluster design with ship lines the first level of clustering, the hydroacoustic transects at each of the sample locations as the second level of clustering, and if sub-sampling was done along a transect then that is the third level. As was noted in the review meeting, in fact the entire transect was analyzed and so the data for every 15-sec piece is available to be and should be analyzed. In this case, the first level of clustering (the ship lines along which the hydroacoustic transects were conducted) will need to be treated as a random sample of start locations although it is clear from the provided figures that they were not randomly selected. This is also true for the second level (the hydroacoustic transects). It appears that the locations of these along the ship lines were not randomly selected but somewhat systematically. These should probably also be treated as random. One final caution is that the data should not be subdivided into depth strata here since the ship lines that ran perpendicular to the shoreline cross depth strata and so observations between strata are not independent and should not be assumed to be so.

A second example is the C-BASS survey of the pipelines which were incorrectly analyzed in the main analysis but appropriately in the "validation" analyses (Table 7). Note here that the appropriate sample size is the number of pipeline samples not the number of 15-sec videos that were provided. Since the analyses used all of the 15,000 or so 15-sec videos it is not surprising that the CV associated with this estimate is smaller than that from the main analyses which used only a subset of the data and captured only the second-stage variability in the two-stage cluster sampling.

On the other hand, I am unsure whether the data collected in the UCB by the C-BASS surveys where they paired a transect in the UCB with a pipeline transect should be used in the estimation procedures. The sampling frame for these UCB transects is a subset of the sampling frame for the UCB in general and so are potentially sampling a different universe that may or may not have the same characteristics (is the UCB within a km or two of a pipeline representative of the UCB in general?). It should be used with caution and should not be combined with the hydroacoustics surveys in any strata where the two methods overlap. Instead, I believe that the C-BASS should be used only in those areas of the UCB not covered by the hydroacoustics and if it is reasonable to assume that the UCB near pipelines is representative of the other areas of the UCB that could

not be sampled under this approach. Further, if the C-BASS results are to be post-stratified and then combined with the hydroacoustics estimates, the report should include the review of the similarities in the estimates for the overlap region as was discussed in the review meeting.

A third issue is the post-stratification that was performed. In the case of the post-stratification of the pipelines into 3 size classes, that appears to have been appropriate due to the lack of information available before sampling as to the size of the pipelines. The main issue here is whether the post-stratification reduces the standard errors (SEs) sufficiently given the small sample sizes. One point concerning this particular post-stratification is whether it was used for any reason other than reducing variance in a somewhat artificial manner. The usual reasons for post-stratification is domain estimation where the reason for post-stratifying is that the scientist wishes to make estimates for subpopulations that could not be sampled directly as the sampling units' characteristics of interest are not known until the unit is sampled (here that is pipeline diameter). If that is the case for this study then the post-stratification is appropriate; otherwise, it is being used to potentially reduce variability in estimates when it was not *a priori* planned.

On the other hand, any post-stratification that may have been performed due to a review of the data collected should be avoided. Such post-stratification can lead to severe bias and artificially low estimates of SE. It appears from the review meeting that the division of the FL shelf into 3 further substrata (northern, central, and south) was based partly on such a data review. Hence, these should not have been used and could have led partly to the need to perform imputations in the northern areas for differing depth zones.

Another issue is the use of stratification that had been planned but not implemented in the planned manner. An example is the hydroacoustics/video surveys in the TX region where the sampling was over a depth range that was an intersection with two originally planned depth strata. The data analyses should probably treat this a a single stratum, especially since the presentation of the final estimates is not done by depth zone but by habitat and region which implies that depth distribution is not of foremost interest.

One final issue concerning the analyses is the decision to assign all RF high probability of occurrence sampling locations to natural hardbottom. I am unsure of the effect of this decision on the analyses but it could lead to bias. If possible, it could be validated by reviewing the video from the sites that were assigned to this stratum to determine if they do represent hardbottom.

- **Are potential sources of uncertainty effectively incorporated into variance estimates?**

I would argue that the estimated variances, even when corrected by capturing the implemented sampling design, are low due to additional sources of variability currently not included. For example, the expansion factors used to convert densities to abundances are themselves estimates of the true values and so introduce a source of unaccounted variability. In at least one case, this can be accounted for, namely the number of artificial reefs offshore of AL/MS. This estimate of the number of reefs is estimated from a years-long survey and so has an estimate of its variance which could be used to adjust the variance of the estimated abundance of red snapper on artificial reefs in the AL/MS region. I would recommend that this be done to assess the amount of additional variability that was not accounted for in any of the estimates of abundance on artificial reef habitat throughout the Gulf. Note that this is a qualitative assessment since we have no data for the variability of artificial reefs elsewhere in the Gulf and it was intimated at the review meeting that the number of reefs is always changing due to loss and additions over time. I say

qualitative since the exercise would provide some indication of the effect of the uncertainty in habitat size on the estimate of the total abundance.

Another example of a known source of variability that was not accounted for is the use of red snapper proportions from videos applied to the hydroacoustics and C-BASS transects. My understanding is that the proportions were arrived at by combining several videos to obtain the average red snapper proportion and so there is an estimated variance associated with that average that could be used to adjust the variance of the estimated total abundances from those data that use the mean proportion. Again, this provides some qualitative assessment of the variability introduced due to uncertainty in the estimates of habitat size.

There are of course many sources of additional variance in the estimates of total abundance that are not accounted for and which cannot be included due to the lack of information about them. These include the variability introduced in the processing or inclusion of hydroacoustics and videos due to water clarity, the variability due to using imputations/substitutions for missing strata, and the variability in the video estimates of the number of fish on the video images.

Although not specifically related to variances, I wish to comment on the decision to provide two design-based analyses of the same dataset, i.e. the concept that there is a "validation" of the estimated abundances when using design-based inferential procedures. The argument that was given for this validation was that each researcher chose their own approach to post-stratification and choice of estimator. But there should not have been independent choices of post-stratification as they should have been identified before any analyses occurred (e.g. the pipeline size post-stratification) and part of any design-based estimation effort is the review of possible estimators in order to choose the "best" one. So, in design-based inference, the only variability in the statistical analyses should be the choice of estimator, e.g. using a mean of ratios rather than a ratio of means. As a result, one does not need to "validate" the results of design-based inference. The fact that the estimated abundances were similar is not evidence that either approach was correct since this was to be expected. Instead, the only question that arises in design-based estimation should be whether, when the design is correctly followed, one estimator provides a less-biased, less-variable estimate than another estimator. Given the sample sizes used in this study, bias of the estimators is likely not an issue. Hence, there is no need to report two different analyses of the same data as evidence that the main approach is "validated".

- **Are imputations made for unsampled regions appropriate, and what are the potential implications for the direction of biases in the estimates.**

I do not address the biological significance of the choices for which data to use for imputations but instead discuss the statistical aspects. The imputation of the means from other strata does introduce an additional source of potential bias and variance but it also allows for conclusions concerning the estimated total abundance that could not occur otherwise. The problem is that by imputing the variance along with the mean for some of the strata in the FL region, one is artificially assigning a variance based on a sample size that could be completely incorrect. I am unsure how one could correct this without extensive additional research and so recommend that the users of the data be aware that such an issue does exist and could influence the results.

The approach that was taken for the LA region appears to have been more appropriate use of the imputation in that the data for those parts of the TX region most similar to LA was combined with the red snapper proportions collected in LA in order to obtain estimates of abundance more likely for the LA region. The problem is that once again, the sample sizes are not valid and so

any estimate of variance is likely not correct since it is based in part on sample size. It might be worth exploring calculating the variance of the estimate imputed for the missing stratum by assuming that the missing stratum sample size is the value that was planned for that stratum before data collection occurred, i.e. estimate a standard error for the estimate used in that stratum by dividing the standard deviation of the dataset used in the imputation by the square root of the sample size intended for the stratum being imputed. An alternative might be to randomly select with replacement observations from the dataset used for the imputation (e.g. the FL, northern, shallow depth, low occurrence dataset) and assign those randomly selected values to the stratum to be imputed (e.g. the FL, northern, mid-depth, low occurrence) where the sample size would be the planned sample size of the missing stratum. This is similar to the imputation method known as "hot-decking". The bias introduced by such methods depends on the adequacy of the selected replacement information for describing the missing information.

3. RESULTS

- **Is the estimate and its variance reliable, consistent with input data and population biological characteristics, and useful as an estimate of absolute abundance of age 2+ red snapper?**

I am not sure whether these results are reliable and consistent with population biological characteristics as I am not a red snapper expert. Instead, I would state that the results if corrected for the above noted statistical issues where possible and reasonable are consistent with the data and can be useful in at least a regional context. I am not as comfortable with combining the western and eastern regions into a single value that is an estimate of the "absolute abundance" of red snapper in the Gulf since the technologies and abilities to obtain data are so different between the two regions that they may not be describing the same quantity. Further, I am concerned with combining the UCB estimated abundance with the other habitats since the variance of the estimated abundance for the UCB when recalculated based on the actual design implemented is likely an underestimate. This is partly due to the small sample sizes in that stratum and so it is likely that the full range of possible density values or variety of habitats in the UCB were not observed. Hence, confidence in the estimates for that habitat is lower than for the other habitats. So, my main conclusions are that the estimated totals are appropriate for their regions and habitats but they are more variable than indicated in the current estimates of CV and further that they may not truly be measures of absolute abundance Gulf-wide but can be considered in a regional context.

- **Assumptions and biases inherent to the methods: Are assumptions made appropriate, given study design considerations? Describe the magnitude and directionality of any biases.**

Yes, the lists of assumptions appear to be appropriate and I do not believe there is any systematic bias due to the sampling design or statistical analyses, except where noted above related to variance estimation.

- **Do you think the data presented can be combined with age-specific composition information for generating an age-specific estimate of abundance?**

I cannot answer this question as it relates to biological considerations.

Table 1. Sample sizes reported in text or Tables 6-7 or appendices

Region	Habitat	Sub class	Sample Size in Text	Page in Text	Table 6 Values	Table 7 values	Table 12 Ahrens et al.[g]
FL	Natural		749	32	505	295	180
	Artificial		65	32	84	84	180
	Natural+ Artificial		927	Fig 5			
	UCB[a]				530	453	
AL/MS	Natural[c]		32		32	32	114
	Artificial		130[f]		198	198	90
	Artificial[b]	Shallow AL	68	Table 3			
		Mid AL	45	Table 3			
		Deep AL	4	Table 3			
		MS	13	Table 3			
	UCB[a]				931	628	
TX	Natural		40	55	36	30	1071
	Artificial		18	55		22	90
	Artificial	Large			45		
		Small			4		
	UCB	10-100 m	140	65	6435	3538	
	UCB[d]		8	55			
LA	Natural		22[e]	68		656	603
	Artificial		42[e]	68		42	90
	UCB		1540[e]	68		3745	
Pipelines					27	15618	

[a] no explanation is provided in text in Regional Sections for FL, AL/MS
[b] number of reefs sampled, not number of selected grid cells
[c] number of "natural features"
[d] considered to be UCB due to habitat features noted during the survey
[e] unclear whether these are the number of samples from TX used for imputation or are the final sample sizes within the LA region
[f] sum of the 4 classses within artificial for AL/MS
[g] numbers are for number of artificial reefs or number of 3 arc-second square areas (~90 m x 90 m) for natural hardbottom