INDEPENDENT PEER REVIEW OF THE REVISION OF GBYP AERIAL SURVEY DESIGN, IMPLEMENTATION AND STATISTICAL ANALYSES (ICCAT GBYP 12/2020) OF THE ATLANTIC-WIDE RESEARCH PROGRAMME FOR BLUEFIN TUNA (ICCAT GBYP Phase 10)

Stephen T. Buckland[1]

SUMMARY

Aerial surveys of the bluefin tuna stocks of the Mediterranean were carried out in each of seven years between 2010 and 2019 inclusive. The most recent time series of estimates shows large differences from previous estimates, and high interannual variation both within and between regions. I review the survey design, the field methods, and the methods of analysis. I conclude that spatial and temporal coverage of the survey may be insufficient to yield a reliable time series of estimates, especially if spawning locations and spawning times vary across years. Given the difficulties that observers face in recording reliable data for the line transect method, I suggest that the use of high-resolution imagery be explored, possibly in conjunction with long-distance drones. Video or still images taken from higher altitude provide a permanent record, allowing verifiability. I review the methods of analysis used to date, and suggest more advanced model-based methods to complement the design-based methods used to date. I also note the large inconsistencies in some estimates, which point to problems in the computer code.

RÉSUMÉ

Des prospections aériennes des stocks de thon rouge de la Méditerranée ont été réalisées chaque année pendant sept ans entre 2010 et 2019 inclus. Les séries temporelles d'estimations les plus récentes présentent de grandes différences par rapport aux estimations précédentes, ainsi qu'une forte variation interannuelle tant au sein des régions qu'entre elles. Le présent document passe en revue la conception des prospections, les méthodes de terrain et les méthodes d'analyse. Il est conclu que la couverture spatiale et temporelle des prospections peut être insuffisante pour produire une série temporelle fiable d'estimations, en particulier si les lieux et les périodes de frai varient d'une année à l'autre. Compte tenu des difficultés que rencontrent les observateurs pour enregistrer des données fiables pour la méthode de transect linéaire, il est suggéré d'explorer l'utilisation d'images à haute résolution, éventuellement avec des drones à longue distance. Les images vidéo ou fixes prises à haute altitude fournissent un enregistrement permanent, ce qui offre une vérifiabilité. Le document examine les méthodes d'analyse utilisées jusqu'à présent, et suggère des méthodes plus avancées basées sur des modèles pour compléter les méthodes basées sur la conception utilisées jusqu'à présent. De grandes incohérences de certaines estimations sont également notées qui mettent en évidence des problèmes dans le code informatique.

RESUMEN

Se realizaron prospecciones aéreas de los stocks de atún rojo del Mediterráneo en cada uno de los siete años comprendidos entre 2010 y 2019, ambos inclusive. La serie temporal más reciente de estimaciones muestra grandes diferencias con respecto a las estimaciones anteriores, así como una gran variación interanual tanto dentro de las regiones como entre ellas. Revisé el diseño de la prospección, los métodos de campo y los métodos de análisis. Llego a la conclusión de que la cobertura espacial y temporal del estudio puede ser insuficiente para obtener una serie temporal fiable de estimaciones, especialmente si los lugares de desove y las épocas de desove varían a lo largo de los años. Dadas las dificultades a las que se enfrentan los observadores a la hora de registrar datos fiables para el método del transecto lineal, sugiero que se explore el uso de imágenes de alta resolución, posiblemente junto con drones de larga distancia. Las imágenes de vídeo o fijas tomadas desde mayor altura proporcionan un registro permanente que permite

[1]CREEM, The Observatory, St Andrews KY16 9LZ, UK; steve@st-andrews.ac.uk

su verificación. Reviso los métodos de análisis utilizados hasta la fecha, y sugiero métodos más avanzados basados en modelos para complementar los métodos basados en el diseño utilizados hasta la fecha. También observo las grandes incoherencias de algunas estimaciones, que apuntan a problemas en el código informático.

KEYWORDS

Aerial surveys, bluefin tuna, line transect sampling

1. Background

A key part of the Atlantic-wide Research Programme on Bluefin Tuna is aerial surveys of four areas of the Mediterranean Sea that are considered to be the main spawning areas. The latest estimates of abundance, spanning the period 2010-2019, exhibit large variability (Cañadas and Vázquez, 2020), which raises the question of whether the surveys are able to provide reliable data to inform management of the stocks.

This review was commissioned by the Center for Independent Experts (CIE) to provide an independent review of the aerial survey design and statistical analysis used in the development of an index of spawning stock biomass, with an emphasis on the 2019 re-analysis of the time series (Cañadas and Vázquez, 2020). Specific tasks are given in the Terms of Reference (ToR**, Appendix 1**).

2. Description of the individual reviewer's role in the review activities

My expertise is in distance sampling. The bluefin tuna surveys use line transect sampling, which is the most widely-used distance sampling method. My review therefore concentrates on survey design, field methods and analysis methods adopted for the bluefin line transect surveys. I note that Fonteneau *et al*. (2013) proposed that other approaches to estimating adult biomass be investigated. I do not have the expertise to advise on their suggestions, so I have not addressed them. Di Natale and Idrissi (2013) compared the strategy of surveying spawning adults with that of surveying juveniles, and concluded that the former was preferred. MRAG (2016) concluded that aerial surveys and close-kin mark-recapture methods are the leading contenders. Again, I don't have sufficient experience of close-kin mark-recapture to comment on that approach.

3. Review of methods

Cañadas and Vázquez (2020) report annual abundance estimates that vary widely. They also vary widely from one set of analyses to another. Even within a set of analyses, there seem to be large inconsistencies (see **Appendix 2**). It is worth considering possible sources of such large inconsistencies:

1. Survey design.
2. Field methods.
3. Methods of analysis.

I consider each of these in turn.

3.1 Survey design

Survey design varied across years, and the boundaries seem to have been rather arbitrarily drawn. Detections in all areas appear to occur right up to the boundary of areas of overlap across surveys (and where there was search effort, beyond the boundary), suggesting that these overlap areas are too small. Movements in spawning areas across time may therefore compromise the time series of abundance estimates. Areas could be expanded, or a 'buffer zone' created around each area, with a lower level of effort. Given that schools occur through much of the Mediterranean, it may be necessary to develop a stratified design, with good coverage in existing survey areas, and a low level of coverage through all or most of the rest of the Mediterranean. Survey effort might also be spread through the time over which schools might spawn, with a view to developing a spatio-temporal model of spawning school density, from which an annual index might be estimated (see Section 2.3).

The issue of spatial coverage of the survey was raised by the GBYP Steering Committee in 2012, and Cañadas and Vázquez (2013) subsequently developed a proposal that included survey effort with lower coverage through much of the Mediterranean outside the main survey areas. The only areas excluded were ones with no historical spawning, or ones for which flight permits could not be obtained. 'Outside' areas were surveyed in 2013 and 2015.

Because effort in the main survey areas was substantially reduced to keep total cost within budget, Cañadas and Ben Mhamed (2016) recommended that in future, outside areas should not be sampled unless additional resources were allocated. ICCAT (2012) and MRAG (2016) both provide a useful summary of the advantages and disadvantages of surveying a wider area.

3.2 Field methods

Survey methods have varied across surveys, with bubble windows used in some surveys and not others, and with different companies doing the surveys. It is important to keep methods as comparable as possible.

The standard line transect sampling method assumes that a school located on the line is detected with certainty. It is questionable whether this is the case, even when a bubble window is used. If line transect surveys are to be continued, consideration should be given to the use of double-platform methods, so that this assumption can be relaxed. In addition, there is a question of whether all schools are available for detection at the time of the surveys. Double-platform methods can address the problem of schools at the surface and on the trackline that are not detected, but they do not address the problem of schools that are not at the surface (i.e. are unavailable for detection) when the aircraft passes overhead. Cañadas and Ben Mhamed (2016) estimated that only 47% of schools were available for detection at a given time, but presumably did not have adequate data to assess how this proportion varies across years, or through the spawning season within a year.

It is possible that observers in different years were searching in different ways. If one is not already in place, an observer training programme should be implemented that includes sufficient detail of the analysis methods that observers understand what is needed if the method is to work. In the absence of such training, observers tend to try to maximise the number of schools detected, which results in detections at large distances which are of little value in the line transect analysis, and potentially a failure to detect small schools near the line. The training should address how to search so that for schools at the surface, probability of detection is at or close to one near the line, and does not fall off rapidly with distance from the line.

Serious consideration should be given to using high-resolution imagery (video or stills) instead of observers. This allows survey aircraft to fly at higher altitude (reducing safety concerns), and strips will be surveyed without the problem of detectability falling off with distance. Also, the images provide a permanent record of what was seen, which can be verified and re-analysed as required. By contrast, the data recorded by the observers are limited by the fact that they are the observers' interpretation of fleeting images. Thus distance estimates and school sizes and weights cannot be verified, and there is no information on missed schools. Using high-resolution imagery also allows the use of long-distance drones rather than piloted aircraft (subject to any restrictions on use of drones in the areas to be surveyed). In the UK, two companies, HiDef (video) and APEM (stills) routinely use high-resolution imagery for marine surveys (see for example Buckland *et al*., 2012). Ideally, the method would be calibrated against current methods necessitating two surveys in at least one year, but given the lack of consistency in the methods used to date, such a calibration exercise would probably have limited value. (Di Natale, 2016, was similarly pessimistic about the potential for calibrating surveys against each other, to try to account for this lack of consistency.) Instead, each estimate might be assumed to be an unbiased estimate of abundance, in which case all estimates would be comparable – at least in principle.

3.3 Methods of analysis

Current estimates should be reviewed and revised, in light of the issues raised in **Appendix 2**. It would appear that the estimates should vary by far less than do some of the estimates tabulated in Cañadas and Vázquez (2020).

The histograms of distance estimates given by Cañadas and Vázquez (2020) are badly spiked; such data are notoriously difficult to model reliably. It appears that many schools are only detected if they are very close to the trackline (within 300m), and there is very little benefit in searching out to 5km or beyond in this circumstance. This alone suggests that survey methods should be revised. This could for example involve bubble windows, with search concentrated within say 1km either side of the line, but a better option would be to use high-resolution imagery from higher altitude, to allow good quality data on a strip of width perhaps 500m or so. Given the extremely rapid fall-off in detectability with distance when using observers, there would be little loss and substantial gain from near-complete detection in a much narrower strip.

Left-truncation of the distance data may affect estimates for 2010-2013. At the start of Section 2.1 of Cañadas and Vázquez (2020), it is stated that *esw* is the reciprocal of the probability (which I will call P_a, and which is referred to as p in Cañadas and Vázquez, 2020) of detecting a school that is within the strip. In fact, $esw = w.P_a$ where w is

the right-truncation distance. The simple (but not unique) way to handle left-truncation is to define $P_a = \int_u^w g(y)dy$ where $g(y)$ is the probability that a group at distance y from the line is detected, for $u \leq y \leq w$, so that u is the left-truncation distance, which is zero when there is no left-truncation. Thus given a fitted detection function $\hat{g}(y)$, an estimate of P_a and hence of *esw* can be obtained. (With covariates in the detection function, this estimation can be carried out for each individual detection, and averaged across detections.) If this is done, I would expect estimates of P_a to be smaller for 2010-2013, for which left-truncation was implemented, than for subsequent years, yet the estimates tabulated are larger. This may be because the effect of left-truncation was slight, but the calculations may need to be checked.

I note here a very large discrepancy of the histograms of perpendicular distances for 2010-2013 of Cañadas and Vázquez (2020) and Figure 1 of ICCAT (2012), showing the perpendicular distance distribution for 2010. For the histograms of Cañadas and Vázquez, left-truncation at 109m appears to eliminate the effect of not being able to detect schools under the aircraft that have no bubble window, there is a large spike of detections between this distance and 300m, and estimated probability of detection for a school 3km from the lines is estimated to be well below 0.2. By contrast, Figure 1 of ICCAT (2012) indicates that no schools were detected within 1km of the line, and the spike of detections occurs at 3km. This discrepancy may have already been investigated. If not, it should be.

Cañadas and Vázquez (2020) have used the MCDS engine of Distance, with probability of detection varying by school size. In this case, a Horvitz-Thompson-like estimator is used, where school size appears in the numerator inside the summation of the estimator. This only reduces to the first equation of Section 2.1 of Cañadas and Vázquez (2020) when probability of detection is assumed to be a function of distance from the line only. Using the mean school size as in Section 2.1 of Cañadas and Vázquez (2020) will introduce bias when probability of detection is a function of school size as here.

Estimation of animal abundance and total weight requires reliable estimates of school size and weight of school. Vázquez and Cañadas (2019) note that observers' estimates are often poor, and suggest that calibration is needed. They also tested different cameras for obtaining images of schools, potentially allowing calibration to be carried out. Such calibration would seem to be prudent. Alternatively, surveys might switch from line transect surveys with observers to strip transect surveys with high-resolution imagery. Grup Air-Med (2019) successfully implemented a calibration trial, demonstrating the feasibility of calibrating observer estimates.

Consideration should be given to spatial modelling of the data, which may help inform estimation of the extent of spawning areas, and whether these areas move around across years. Density surface modelling (Hedley and Buckland, 2004) should be adequate for this purpose, although more sophisticated point-process models (Yuan *et al.*, 2017) might be considered. Given possible variation in spawning times, it would be advisable to include date in the model, and so develop a spatio-temporal model of spawning density. An index could then be defined in several ways. For example, averaging over space within any given area would give spawning density as a function of date within any given year, and an index could be estimated by integrating over date.

Good progress towards developing a model-based approach was reported by Cañadas and Ben Mhamed (2016), and further development of that approach to show how density varies over space and time might prove useful.

4. Summary of findings for each TOR

In this section, I list each TOR in turn, and summarise my findings relevant to it.

i. Review all relevant information (to be provided by the ICCAT Secretariat - GBYP) on the survey's design, implementation, and statistical approach for the development of the BFT index of abundance. If deemed necessary, discussion over a webinar between CIE reviewers and BFT aerial survey team. Is survey documentation and supporting material adequate to conduct this review?

I found the survey documentation and supporting material to lack detail in places, so that I was unsure of how analyses had been conducted.

ii. Survey design. Evaluate the historical protocols and analytical approaches used in this survey as well as the recommended changes to the design procedures.
    a. Is the current survey design and changes implemented over its history consistent with state-of the art aerial survey design and adequately accounted for in data or statistical treatment?

The survey areas do not fully cover spawning areas, and there is no allowance made for potential changes over time in where schools spawn, or in dates of the main spawning period. While design-based analysis methods potentially yield robust estimates of density within survey areas at the time of the surveys, they offer no insights into how density varies within areas, and whether high density regions extend beyond area boundaries, nor do they indicate whether the surveys captured the main spawning period. Model-based methods applied to past data might help inform design changes, in terms of both spatial extent of survey areas and temporal extent of surveys within a year.

b. Have logistical issues that precluded full attainment of the design been adequately addressed?

If field methods are considered part of survey design, then the introduction of bubble windows has been helpful, although it compromises to some extent the comparability of estimates across years. Improved training of observers would aid comparability. Ideally, the same company and same teams of observers would be used in each survey year. However, now may be the time to consider replacing observer line transect surveys by surveys using high-resolution imagery.

c. Are there further unaccounted for factors?

Changes in aircraft and observer teams are not fully accounted for. Cañadas and Vázquez (2020) use left-truncation for those surveys conducted wholly or partially without bubble windows, which should increase comparability between those estimates and estimates for later years.

iii. Evaluate Statistical treatment and index calculation of the Mediterranean survey time series.

a. Are data treatments (spatial stratification, etc.) appropriate and adequate to account for known factors affecting detection and quantification of spawning biomass?

Model-based analysis methods may help to assess this, and possibly provide improved and more precise estimates of spawning biomass.

b. For issues not addressed by (ii) above, does statistical treatment adequately account for issues affecting detectability, specifically does use of 'school size' in the detection function bias the detection estimates and does the method potentially double count schools detected multiple times?

Detection of the same school from different transects, either because the school is visible from more than one transect or because the school moves from one strip to another, does not systematically bias estimates. See Buckland *et al*. (2001:36) for discussion of the first issue, and Buckland *et al.* (2001:32) for discussion of the second issue. Provided a Horvitz-Thompson-like estimator (Buckland *et al*., 2004:38-45) is used, modelling probability of detection as a function of school size does not bias estimation of abundance or biomass. However, if density is estimated simply by multiplying school density by mean school size, as stated in the first equation of Section 2.1 of Cañadas and Vázquez (2020), then bias does occur; abundance or biomass would be over-estimated, because mean school size in the sample overestimated mean school size in the population, as larger schools have a higher probability of detection than smaller schools.

c. Does the most recent (2019) index construction represent the most effective treatment?

Model-based methods may prove more effective. See recommendations.

d. Does the high inter-treatment variability of the index due to poorly estimated or highly variable detection functions render the index unreliable as a time series?

There are problems with the estimates as presented. See **Appendix 2** for examples. If these issues can be resolved, a useful index should be achievable.

e. Are better statistical (spatial/temporal) treatments possible?

I believe that a model-based approach, developing a spatio-temporal model, might give a more informative index. See recommendations.

iv. Suitability of GBYP aerial survey

a. Does it achieve full objective (all Mediterranean spawning grounds) or partially (on specific spawning areas)

Partial only. Spatial coverage is not sufficient, and temporal coverage within a year may also be problematic.

b. Are known logistical/biological/unaccountable factors adequately addressed?

See recommendations.

c. Are unknown factors (availability of fish, timing of spawning, behavioural changes) too substantial, rendering the survey unable to achieve its full or partial goals?

I believe a model-based approach may help address these issues. See recommendations.

d. Provide general recommendations for potential improvements

See recommendations.

e. Determine if the current approach meets the established criteria for an index of abundance. If not provide an explanation of why and whether or not the data can be re-evaluated to meet these criteria.

See recommendations.

v. Provide recommendations on the future of this survey, as well as potential design modifications, standardization and/or research to improve the survey

See recommendations.

5. Discussion and recommendations

Surveys by aircraft or long-distance drones at relatively high altitude and using high-resolution imagery offer several advantages over line transect surveys using observers. First, the raw data are images, which provide a permanent record of the survey and for which estimates can be updated if analysis methods are improved. By contrast, for line transect surveys, only the observers' interpretations of what was seen are available to the analyst. Second, detectability of schools should not be a function of distance from the transect line, and so the difficulty of modelling the rapid fall-off in probability of detection with distance from the line in the line transect surveys is avoided, as is the difficulty in detecting schools directly below the aircraft. Third, estimates of school size and weight can be estimated objectively from the images, rather than subjectively by observers. Fourth, once suitable image analysis software has been acquired or developed, analysis should be relatively straightforward and rapid.

Recommendation 1. The feasibility and cost of conducting surveys using high-resolution video or stills instead of observers, and from either piloted aircraft or long-distance drones, should be assessed. This should include a pilot study carried out say in Area A, with aircraft that have bubble windows and observers as well as high-resolution cameras, so that data from the two approaches can be compared.

Many spawning schools occur outside of the areas of overlap across years, and outside the dates over which the surveys are conducted.

Recommendation 2. Survey designs that offer improved spatial and temporal cover at acceptable cost should be developed.

The surveys can detect only spawning schools that are at the surface when the aircraft passes over. Quílez-Badía *et al.* (2016) proposed a combined aerial survey and tagging programme to allow estimation of the proportion of schools available to be detected in the survey.

Recommendation 3. Information on behaviour of spawning schools and on timing of spawning should be reviewed, to assess whether there might be time trends in the proportion of schools that are at the surface and hence detectable at a given time point during the surveys, and if so, whether an approach along the lines proposed by Quílez-Badía *et al.* (2016) is feasible.

Standard line transect methods assume that all schools on the line are detected. 'Double-platform' surveys allow estimation of the probability of detection, without assuming that detection of surface schools on the line is certain. Having front observers as one 'platform' and rear observers as the other is unlikely to be effective if only one platform has bubble windows. A better option is to have the observers as one platform and a high-resolution camera (or multiple cameras, to increase field of view) as the second. Another option of having two aircraft flying in tandem might be effective, but would double field costs.

Recommendation 4. If line transect surveys are continued, the feasibility of carrying out a 'double-platform' survey should be assessed, to allow the estimation of a detection function without having to assume that detection of schools at the surface and on the line is certain.

Observers must seek to detect all or nearly all schools near the line, ensure that probability of detection does not fall sharply with distance from the line, estimate the size and weight of detected schools, and generate comparable data over time. (The need to train observers on how to search was recognised by ICCAT, 2012.)

Recommendation 5. If line transect surveys are continued, and if such a programme does not already exist, an observer training programme should be developed, to ensure that observers understand what is required of them, and to gain experience in detecting schools, and estimating their size and weight.

Trials suggest that observers' estimates of school size and weight are very variable, and some observers' estimates may have large bias.

Recommendation 6. If line transect surveys are continued, it would seem essential to calibrate observers' school size and weight estimates, using photographic images of at least a sample of detected schools.

While design-based methods generally yield robust estimates of density, model-based methods offer several advantages: spatial models will indicate whether high densities potentially occur outside of surveyed areas; spatio-temporal models will indicate whether temporal coverage of the surveys is adequate; a degree of extrapolation to unsurveyed areas may be possible; precision may be improved.

Recommendation 7. A spatial (e.g. density surface) model should be fitted to the data for each survey year. This should be extended to a spatio-temporal model, with date within a year being added to the spatial model, to explore the options for estimating an annual index from the fitted model, which will allow for spatial and temporal variation in spawning densities across years.

Estimates of Cañadas and Vázquez (2020) show some inconsistencies that are probably a result of problems in the computer code.

Recommendation 8. If these inconsistencies cannot be resolved, an experienced user of the R distance software should be hired, either to review the code used to generate the estimates, or to write independent code to do the same analyses, allowing differences between the new estimates and the estimates of Cañadas and Vázquez (2020) to be identified and resolved.

The methods of analysis used by Cañadas and Vázquez (2020) are not fully described, and there appear to be inaccuracies in the description. A full description of the methods would allow verification of estimates.

Recommendation 9. The methods of analysis should be described in sufficient detail that an independent analyst could repeat the analysis, and the computer code used should be available.

Acknowledgements

This work has been carried out under the ICCAT Atlantic-Wide Research Programme for Bluefin Tuna (GBYP), which is funded by the European Union, several ICCAT CPCs, the ICCAT Secretariat, and other entities (see https://www.iccat.int/gbyp/en/overview.asp). The content of this paper does not necessarily reflect ICCAT's point of view or that of any of the other sponsors, who carry no responsibility. In addition, it does not indicate the Commission's future policy in this area.

References

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. 2001. Introduction to Distance Sampling. Oxford University Press, Oxford.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. (eds) 2004. Advanced Distance Sampling. Oxford University Press, Oxford.

Buckland, S.T., Burt, M.L., Rexstad, E.A., Mellor, M., Williams, A.E. and Woodward, R. 2012. Aerial surveys of seabirds: the advent of digital methods. J. App. Ecol. 49: 960-967.

Cañadas, A. and Ben Mhamed, A. 2016. Power analysis and cost-benefit analysis for the ICCAT GBYP aerial survey on bluefin tuna spawning aggregations. Atlantic-wide Research Programme on Bluefin Tuna Phase 5.

Cañadas, A. and Vázquez, J.A. 2013. Short-term contract for assessing the feasibility of a large-scale survey on bluefin tuna spawning aggregations in all the Mediterranean Sea for obtaining useful data for operating modeling purposes. Atlantic-wide Research Programme on Bluefin Tuna Phase 3.

Cañadas, A. and Vázquez, J.A. 2020. 2010-2019 aerial survey data re-analysis. Atlantic-wide Research Programme on Bluefin Tuna Phase 9.

Di Natale, A. 2016. Tentative SWOT analysis for the calibration of ICCAT GBYP aerial survey for bluefin tuna spawning aggregations. Collect. Vol. Sci. Pap. ICCAT, 72(6): 1463-1476.

Di Natale, A. and Idrissi, M'H. 2013. ICCAT GBYP aerial survey: spawners vs. juveniles. A SWOT analysis of both perspectives. Collect. Vol. Sci. Pap. ICCAT, 69(2): 803-815.

Fonteneau, A., Suzuki, Z. and Payne, A.I.L. 2013. Mid-term review of the ICCAT Atlantic-wide Research Programme on Bluefin Tuna Phase 4.

Grup Air-Med. 2019. Implementacion del muestreo aereo para la calibracion y validacion. Atlantic-wide Research Programme on Bluefin Tuna Phase 9.

Hedley, S.L. and Buckland, S.T. 2004. Spatial models for line transect sampling. Journal of Agricultural, Biological and Environmental Statistics 9: 181-199.

ICCAT. 2012. 2011 GBYP workshops on aerial surveys, and operational meetings on biological sampling and on tagging of bluefin tuna. Collect. Vol. Sci. Pap. ICCAT, 68(1): 1-65.

MRAG. 2016. Second review of the ICCAT Atlantic-wide Research Programme on Bluefin Tuna Phase 6.

Quílez-Badía, G, Tensek, S., Di Natale, A., Pagá García, A., Cañadas, A., Kitakado, T. and Kell, L.T. 2016. A consideration of additional variance in the Mediterranean aerial survey based on electronic satellite tags. Collect. Vol. Sci. Pap. ICCAT, 72(6): 1544-1552.

Vázquez, J.A. and Cañadas, A. 2019. Design of an aerial survey calibration exercise and an updated aerial surveys sighting protocol. Atlantic-wide Research Programme on Bluefin Tuna Phase 8.

Yuan, Y., Bachl, F.E., Lindgren, F., Borchers, D.L., Illian, J.B., Buckland, S.T., Rue, H. and Gerrodette, T. 2017. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. Annals of Applied Statistics 11: 2270-2297.

Appendix 1

Statement of Work

1. Background and Objectives

The BFT aerial survey is one of the major activities of the Atlantic Wide Research Programme for Bluefin Tuna (GBYP). It was launched in 2010 with the purpose of obtaining a relative abundance index of spawning biomass for the Mediterranean Sea. The index is obtained from aerial transects conducted during June in the four main spawning areas using a combination of scientific and professional spotters deployed on airplanes. Since its start, the survey has faced numerous logistical challenges and has had to alter its design and data processing protocols multiple times.

Currently, the most recent (2019) iteration of the index exhibits substantial differences from prior time series and the index exhibits high interannual variability both within and between regions. The magnitude of the difference between prior time series and the high variability has raised concerns regarding the estimation procedures and the overall efficacy of the survey to reflect annual spawner abundance in the Mediterranean Sea. Given the need to evaluate the survey and to soon take decisions regarding the nature of its continuation, ICCAT requests an independent desk review of the survey design, statistical treatments and analytical procedures and of its general capacity to achieve its objectives.

Expertise required to conduct this review will include two independent and highly qualified experts with a combined background and experience in aerial survey design, statistical time series evaluation, and a strong understanding of population modeling and stock assessment. Reviewers will have no financial or perceived conflicts of interest related to the subject matter to be reviewed. Finally, reviewers are to be approved by ICCAT upon selection but only as approval related to reviewer expertise to conduct the review and/or any conflicts of interest not discovered over the reviewer identification and selection process. The CIE will however make the final decision on the eligibility and effectiveness of all selections in such cases.

2. Reviewer Tasks

To provide an independent review of the Mediterranean Sea Bluefin tuna aerial survey design and statistical analysis used in the development of an index of spawning stock biomass, with an emphasis on the 2019 re-analysis of the time series. Specific tasks will include, but not be limited to, the following Terms of Reference (ToR):

i. Review all relevant information (to be provided by the ICCAT Secretariat - GBYP) on the survey's design, implementation, and statistical approach for the development of the BFT index of abundance. If deemed necessary, discussion over a webinar between CIE reviewers and BFT aerial survey team. Is survey documentation and supporting material adequate to conduct this review?

ii. Survey design. Evaluate the historical protocols and analytical approaches used in this survey as well as the recommended changes to the design procedures.
    a. Is the current survey design and changes implemented over its history consistent with state-of the art aerial survey design and adequately accounted for in data or statistical treatment?
    b. Have logistical issues that precluded full attainment of the design been adequately addressed?
    c. Are there further unaccounted for factors?

iii. Evaluate Statistical treatment and index calculation of the Mediterranean survey time series.
    a. Are data treatments (spatial stratification, etc.) appropriate and adequate to account for known factors affecting detection and quantification of spawning biomass.?
    b. For issues not addressed by (ii) above, does statistical treatment adequately account for issues affecting detectability, specifically does use of 'school size' in the detection function bias the detection estimates and does the method potentially double count schools detected multiple times?
    c. Does the most recent (2019) index construction represent the most effective treatment?
    d. Does the high inter-treatment variability of the index due to poorly estimated or highly variable detection functions render the index unreliable as a time series?
    e. Are better statistical (spatial/temporal) treatments possible?

iv. Suitability of GBYP aerial survey
    a. Does it achieve full objective (all Mediterranean spawning grounds) or partially (on specific spawning areas)
    b. Are known logistical/biological/unaccountable factors adequately addressed?
    c. Are unknown factors (availability of fish, timing of spawning, behavioural changes) too substantial, rendering the survey unable to achieve its full or partial goals?
    d. Provide general recommendations for potential improvements

v. Determine if the current approach meets the established criteria for an index of abundance. If not provide an explanation of why and whether or not the data can be re-evaluated to meet these criteria.

vi. Provide recommendations on the future of this survey, as well as potential design modifications, standardization and/or research to improve the survey

3. Deliverables

Deliverable #1- CIE reviewer shall submit a draft review report (formatted as an SCRS document) providing complete documentation of the review and recommendations (late September-early October 20200.

Deliverable #2 – CIE reviewer will present the draft review report findings to the Bluefin Tuna Working Group (BFTWG) at its next available meeting (early October 2020) (virtual presentation).

Deliverable #3- CIE reviewer will submit a final review report (formatted as an SCRS document), revised as based on comments provided by the BFTWG (first week in November 2020).

Appendix 2

Details on estimates' inconsistencies

Here, I give details of some of the inconsistencies in the estimates of Cañadas and Vázquez (2020).

1. Table 3.2.1. Estimates of 'average *p*' are consistently higher for 2010-13 than for later years, but as noted in the report, as there was left-truncation for these early years, I would expect them to be lower. However, the amount of left-truncation was relatively modest at 109m, so this may be just a chance effect.

2. At the start of section 2.1, the authors provide an equation for estimating density. This equation includes mean school size, which will give a biased estimate of density because probability of detection of a school was found to be a function of school size. However, if mean school size is dropped from this equation, it should give an estimate of school density. Applying this equation to the estimates of Table 4.1.1, school density is estimated as number of schools detected on effort divided by (twice the esw times the transect length). This gives the estimated densities of row 2 below, while the authors' estimates are in row 3:

```
 2010   2011   2013   2015   2017   2018   2019
0.00049 0.00048 0.00078 0.00093 0.00221 0.00229 0.00179
0.00016 0.00022 0.00035 0.00058 0.00523 0.00402 0.00363
```

   There is clearly no correspondence in these estimates. While the discrepancies for 2010-13 may have something to do with left-truncation of the data, this is not the case for later years. Many of the other estimates in this table show the same discrepancies.

3. Expected cluster size in Table 4.1.1 for 2013 and 2015 are similar (415 vs 445), yet expected weight for 2013 is 0.505 and for 2015 is 79.361, indicating that average fish weight was more than 70 times greater in 2015 than in 2013. I do not have the data to check this.

4. As noted by the authors, total is estimated as density multiplied by area. Thus for area A in Table 4.1.1, total weight should be density of weight multiplied by 61,837. Assuming weight densities are correct, my estimates of total weight are in row 2 below, while the authors' estimates are in row 3.

```
2010   2011   2013   2015   2017   2018   2019
30300   8657  12367 2832135 9402934 10865998 6663555
2119   963   1946   2832   9403  10866   6664
```

   The authors do not specify their units, but for 2015-19, presumably the difference is kg vs tonnes. However, the estimates for 2010-13 bear no correspondence at all, and the density estimates of Table 4.1.1 appear to be orders of magnitude out.

Similar issues appear to be present in other tables, but I haven't checked these.