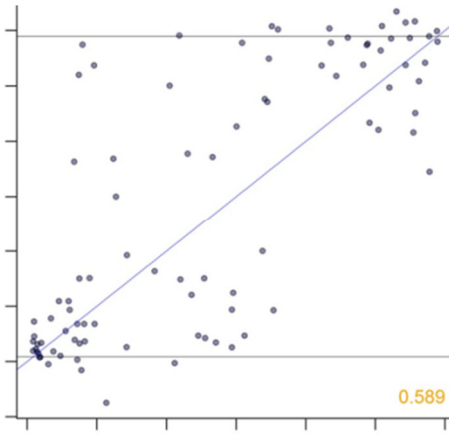
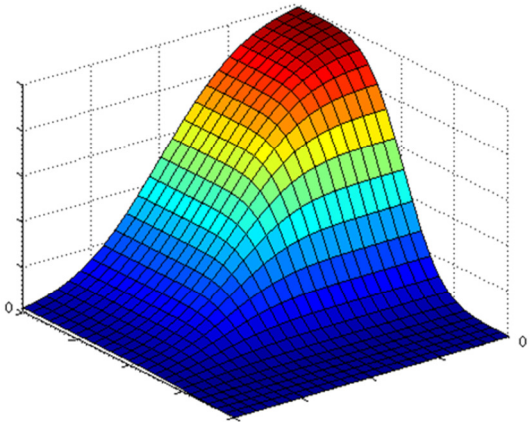
Non-parametric stochastic imputation of length composition data for Atlantic bluefin tuna

Description and cross-validation of imputation methods

Tom Carruthers[1]





January 15th 2013

SHORT-TERM CONTRACT FOR THE MODELLING APPROACHES: SUPPORT TO BLUEFIN TUNA STOCK ASSESSMENT

GBYP 03/2012B

OF THE ATLANTIC-WIDE RESEARCH PROGRAMME ON BLUEFIN TUNA

BETWEEN

THE INTERNATIONAL COMMISSION FOR THE CONSERVATION OF ATLANTIC TUNAS (ICCAT)

AND

TOM CARRUTHERS, UNIVERSITY OF BRITISH COLUMBIA

---

[1] 335 The Fisheries Centre, AERL, 2202 Main Mall, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z4; t.carruthers@fisheries.ubc.ca; (+1) 604 822 6903

1 Executive Summary

- Three non-parametric stochastic imputation approaches that provide a means of imputing Atlantic bluefin tuna length frequency data are described and tested by cross-validation.

- The multivariate normal 'distance' model that randomly imputes data using a multinomial probability function, offered the best predictive capacity by some margin.

- All imputation methods operate most successfully when nearby (in time and space) length observations are imputed preferentially.

- The multivariate normal imputation approach described here offers a basis for quantifying uncertainty from data processing by means of repeated-imputation inference (running multiple stock assessments from multiple imputed data sets).

- By imputing data that are specific to time and region, the approach offers the basis for applying multiple growth curves (for cohort slicing for example) where applicable. Additionally, uncertainty from aging may be simultaneously incorporated into the same repeated-imputation inference framework.

- A number of patterns in the data provide evidence of possible errors in the Task II dataset for Atlantic bluefin tuna, for example, identical length samples replicated in adjacent areas and times.

- While the imputation approach appears to perform reasonably well (assuming the data are reported correctly) future improvements could include the incorporation of greater variability among imputations and further optimization of code to improve computation time (by extending the code to be compatible with parallel processing packages for example).

2 Introduction

Catch composition data are commonly used to support stock assessment approaches that predict the age-structure of a fish population explicitly. Such modelling may be desirable in cases where strong temporal changes have occurred in the strength of cohorts or the vulnerability-at-age of the population. However these data may be patchy in many stock assessment scenarios, particularly large-scale pelagic populations exploited by a number of different fishing fleets. Atlantic bluefin tuna is one such example in which more detailed records of the ICCAT Task II database are incomplete but coarser records of total catch are available in the Task I database.

A version of Virtual Population Analysis (VPA) 'ADAPT' (Gavaris 1988) is currently used as the status quo approach for providing management advice for Atlantic bluefin tuna. A central limitation of VPA analysis is that it requires complete records of catch-at-age (the total landings are apportioned to age classes by year across all fishing fleets). Currently, the complete catch-at-age data are derived by an ad-hoc algorithm that subdivides total reported catches to age classes based on the observed catch composition data. Once derived this total catch-at-age dataset is considered to be known without error for the purpose of the VPA analysis. However two important sources of uncertainty have been removed from the analysis. The first is uncertainty in assigning lengths to the fish that are caught and documented in the Task I dataset:

(1) Length observations (Task II data) + total catches (Task I data) —*Imputation of length data*→ Total catch-at-length

The second source of uncertainty comes from converting these imputed lengths to age classes in order to create the total catch-at-age dataset:

(2) Total catch-at-length —*Length – age conversion (growth model)*→ Total catch-at-age

In the second of two ICCAT contracts '(b) data imputation' of 'modelling approaches: support to Bluefin tuna stock assessment' this research contract aims to address the first of these problems by determining length imputation approaches that allow uncertainty to be expressed in final assessment outputs.

When imputing catch-at-length data it is necessary to understand the missing data problem since it determines the type of imputation method that can be applied legitimately. In this example we have a final dataset of fish lengths that contains missing data points (the Task I dataset indicates that fish were caught but their lengths are not known). If the pattern of the missing length data is unrelated to both the values of the observed and the missing data (the lengths themselves) and all observable variables, the data are considered *missing completely at random* (MCAR) (Gelman *et al.*, 1995). In this case, the probability that data are missing cannot be explained by either the magnitude of the lengths that are missing, a pattern in the observed lengths or any other variable such as fishing gear, or season. Suppose however that small fish may be caught in coastal areas or nearer spawning grounds. If this is the case, the absence of data for small fish is dependent on a variable (sampling location) that is observable. In this example, the missing data mechanism is still independent of the missing values (length) and the data are said to be *missing at random* (MAR) (Rubin, 1976). Where the parameters of the missing data mechanism (that describes the effect of sampling location on probability of capture) are independent of the parameters that describe the probability model of the observed data (*e.g.* given a fish is caught, the chance of obtaining a length measurement is not related to sampling location), the mechanism is considered *'ignorable'* (Rubin, 1976). It may be the case however, that larger fish, of greater power and mobility, are more difficult to measure regardless of where they were caught. Under these conditions the missing data mechanism may be dependent on the missing values themselves. Here the data are not MAR and the mechanism is considered 'non-

ignorable'. In this analysis we will assume that the data are MAR and the mechanism is 'ignorable'; that the probability that a fish is measured is not related to the length.

The simplest model-based method of imputing MAR (ignorable) data is Expectation Maximisation (EM). EM is a model fitting exercise; it estimates the most probable values of parameters in a model that predicts observed data and uses this model to predict replacement data. Expectation maximisation is one of the most widely used of the 'more principled' methods (Little and Rubin 1987). This is exactly the same as fitting a spatio-temporal model to the observed length frequencies. Since such a model has predictive capacity it can impute (fill-in, 'guess') the expected length of a fish caught in a given time or area.

The product of EM is a single imputed data set corresponding to the best fit of the model on the observed data. EM provides no statistically rigorous manner with which to estimate standard errors. It follows that although there is an inevitable uncertainty associated with imputation (*e.g*. a model that fits the observed data poorly), this uncertainty cannot be quantified by EM. With this problem in mind, Rubin (1987) developed a procedure known as Multiple Imputation (MI). In essence, MI is EM with replication. Typically, the replication process occurs between 4 and 10 times creating a different set of imputed values each time. When the mean of the replicate imputations for each data point is used and the standard error incorporated into the final analysis as a measure of uncertainty, the MI method is termed 'repeated-imputation inference'. Under these circumstances, a greater proportion of missing data and/or a model and explanatory variables that do not explain the data well, result in greater uncertainty over the imputed data and the results of any subsequent analyses. Multiple imputation was originally designed to deal with missing data in survey data sets but since its conception in 1987 (Rubin, 1987) it has been successfully applied in a wide range of disciplines (*e.g.* Medicine, Mishra and Dobson 2004; Air quality analysis, Junninen *et al.* 2004). Though MI has received some general criticism on the grounds of inefficiency (Nielsen 2003) it is still thought among many to be the imputation method with the most potential (Olinsky *et al.* 2003).

An alternative to EM is hot-deck imputation which is based on the idea of replacing a missing value with a similar observed value. Though model based, hot-deck imputation differs from EM in that the model is implicit; *i.e*. they are not expressed mathematically (as are the explicit models in EM) and they follow a nonparametric procedure (see Little and Schenker, 1995). Though similarity is the primary condition, hot-deck methods may define similarity by any relevant criterion. For example, 'Hot-deck next case' simply imputes the required data from the next complete case and 'hot-deck nearest neighbour' minimizes a function of the 'distance' between incomplete case and donor case (*e.g*. the number of records, a function of other covariates). Hot-deck methods are most applicable in cases where data are MAR and only the observable variables are those used to decide which case is most similar to that with missing data.

In this research contract I identify a non-parametric approach to imputation that is a hybrid of multiple imputation and hot-deck nearest neighbour. Using a distance model I assign weights to observed data in relation to a set of missing data points and then randomly sample from observations in proportion to their weight in order to fill the missing data points. The method has the ability to create multiple random data sets like multiple imputation but relies on a non-parametric resampling approach and does not require fitting a predictive model to observed data. By generating multiple length data sets and therefor possible catch-at-age datasets, repeated-imputation inference can be obtained by running a VPA stock assessment on each catch-at-age dataset.

3 Methods

3.1 Programming environment

The computer programming for this work was conducted in the statistical environment R (2.14.2, 64bit; R core team 2013). Computations were conducted in Windows 7 using a high-performance laptop (Intel® Core i7-2670QM 2.2Ghz, 16GB RAM). The source code and examples are available online (Carruthers and Kell 2013).

3.2 Data processing

Length data were extracted from the ICCAT Task II dataset 't2sz_20111103' that is available online at www.iccat.int. The data were formatted such that each line of the observed data constitutes an individual length frequency (*e.g.* Table 1).

Table 1. Example formatting of observed length frequency data for the Bluefin tuna data of the Japanese longline fleet. 'ClassFrq' is the fork-jaw length in cm. 'n' is the number of fish sampled at this length. 'Time' is the decimal year. 'DOY' is day of the year to be used in season calculations.

SpeciesCod	FleetCod	GearGrpCod	Lat	Lon	ClassFrq	n	FreqTypeCod	ClassInterval	Time	DOY
BFT	JPN	LL	-7.50043	-22.5139	230	1	FL	1	1956.875	320
BFT	JPN	LL	2.500806	-42.5062	246	1	FL	1	1957.125	46
BFT	JPN	LL	-2.50017	-22.5142	214	3	FL	1	1957.125	46
BFT	JPN	LL	-2.50017	-22.5142	188	1	FL	1	1957.875	320
BFT	JPN	LL	-2.50017	-22.5142	190	12	FL	1	1957.875	320
BFT	JPN	LL	-2.50017	-22.5142	196	1	FL	1	1957.875	320
…										

In cases where lengths are reported in different standards (*e.g.* 'CFL', 'PFFL' or 'CPFFL') these were converted to fork-jaw length using the conversion equations provided on the ICCAT website. Additionally the reporting of geographic location was standardised by taking the gravity of each reported cell using the 'convUL' and 'calcCentroid' functions of the R package 'PBSmapping'.

For cross validation purposes a second dataset was formed by simply ignoring the length ('ClassFrq') and aggregating the number of observations 'n' by time and space (*e.g.* Table 2). In this way, for any time and place the known values can be compared with multiple imputed datasets. In this analysis we undertake a cross-validation of the Task II data that are available for the Japanese and Spanish longline fleets

Table 2. Example formatting of the cross-evaluation dataset to be imputed for the Bluefin tuna data of the Japanese longline fleet. 'Time' is the decimal year. 'DOY' is day of the year to be used in season calculations.

SpeciesCod	FleetCod	GearGrpCod	Lon	Lat	Time	DOY	n
BFT	JPN	LL	5.005029	-42.5761	1970.042	16	2
BFT	JPN	LL	-35.0072	32.57651	1973.042	16	1
BFT	JPN	LL	-55.0546	42.57742	1975.042	16	3
BFT	JPN	LL	-24.9674	27.57184	1976.042	16	13
BFT	JPN	LL	-14.9257	27.5744	1976.042	16	2
BFT	JPN	LL	-24.9713	32.57704	1976.042	16	5
…							

3.3 Non-parametric imputation algorithms for catch-at-length data

For any missing data M_i the 'proximity' to an observed data point O_j can be calculated from a multivariate normal distribution based on the distance in terms of season S, year Y and geographic location G. This proximity or weight W can be used in a multinomial sampling model to determine the relative probability of sampling an observation given the location of a missing data point: $P(O_j|M_i)$. Expressed fully this is:

(3a) $$P(O_j|M_i) \propto W_{i,j} = P(S,Y,G) = \frac{\exp(-0.5v/(p_{SY}^2+p_{SG}^2+p_{YG}^2-2p_{SY}p_{SG}p_{YG}-1))}{2\sqrt{2}\pi^{3/2}\sqrt{1-(p_{SY}^2+p_{SG}^2+p_{YG}^2)+2p_{SY}p_{SG}p_{YG}}},$$

where $v = S^2(p_{YG}^2 - 1) + Y^2(p_{SG}^2 - 1) + G^2(p_{SY}^2 - 1) + 2[SY(p_{SY} - p_{SG}p_{YG}) + SG(p_{SG} - p_{SY}p_{YG}) + YG(p_{YG} - p_{SY}p_{YG})]$,

and p is the correlation among variables,

which looks considerably more complicated than the corresponding R code:

'dmnorm(S,Y,G, mean values (S,Y,G),variance-covariance matrix(S,Y,G))'.

When applying this method, I assume there to be no covariance among the three variables leaving only the positive diagonal of the variance-covariance matrix 'sigma' to be specified. A graphical representation of a bivariate normal distribution for differences in time (among years, *Y*) and space (among geographic areas, *G*) is illustrated in Figure 1.

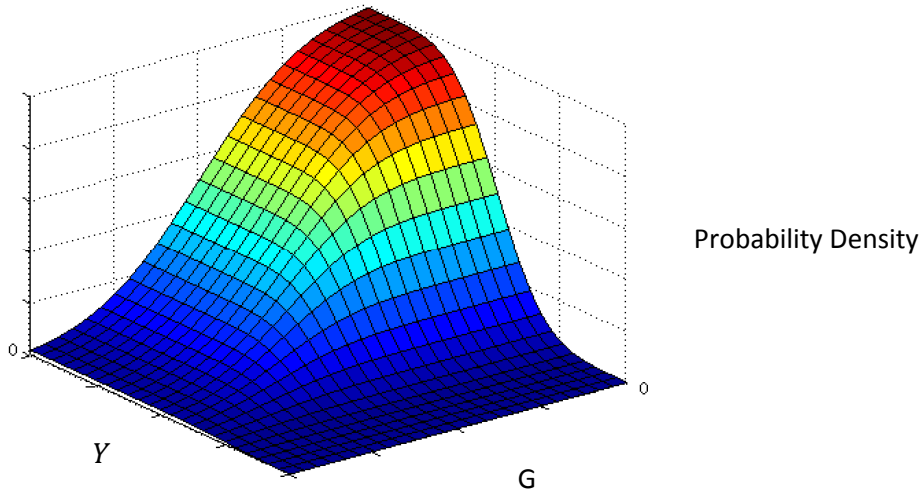


Figure 1. A bivariate normal distribution showing probability density (*W*) with increasing difference in geographic location (*G*, km) and sampling date (*Y*, years). Over this surface, observations O_j are sampled in proportion to their probability density (*W*) according to a multinomial distribution.

Over a large number of observations the multivariate normal approach may be computationally intensive, in which case either an additive univariate normal model:

(3b) $$P(0_j|M_i) \propto W_{i,j} = P(S) + P(Y) + P(G) = \frac{1}{\sigma_S\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{S}{\sigma_S}\right)^2\right) + \frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{Y}{\sigma_Y}\right)^2\right) + \frac{1}{\sigma_G\sqrt{2\pi}}\exp(-\frac{1}{2}\left(\frac{G}{\sigma_G}\right)^2)$$

where the terms σ are the normal standard deviations corresponding to distance in season *S*, year *Y* and among geographic locations *G*.

Alternatively a Euclidean distance model may offer a viable alternative:

(3c) $$P(0_j|M_i) \propto W_{i,j} = \sqrt{\left(\frac{S}{\sigma_S}\right)^2 + \sqrt{\left(\frac{Y}{\sigma_Y}\right)^2 + \left(\frac{G}{\sigma_G}\right)^2}}^{-1}$$

All three of these implementations are available in an R function 'imputeCAS' (imputation of Catch-At-Size, Carruthers and Kell 2013). The multinomial probability of sampling a length observation O_j given a missing data point M_i is calculated from the weighting values W and is simply:

$$(4) \qquad P(O_j|M_i) = \frac{W_{i,j}}{\sum_j W_{i,j}}$$

4 Results

The predictive ability of the three non-parametric imputation algorithms is illustrated in Figure 2 and also described in Table 3. In terms of predicting the mean length frequency, all three methods show poor predictive capacity when length observations are used for imputation that are from distant years, seasons and areas. For example, the first row of panels in Figure 2 shows the predictive capacity of the additive normal, multivariate normal and Euclidean linear algorithms given the standard deviation (σ) for among years, within year and among geographic areas set to 10 years, 60 days and 5000km, respectively. In this case all three models fail to predict high and low mean lengths. Additionally the R-squared measure (variance in observations that can be explained by imputations) is poor in the case of the additive normal model.

For all imputation algorithms the predictive capacity improves markedly as the standard deviations for each dimension (among year, within year, geographic) are reduced (rows of panels nearer the bottom of Figure 2). This points to spatially and temporally heterogeneous distribution of lengths that are highly specific to season which might be expected given the migratory behaviour of Atlantic bluefin tuna and the general pattern of increasing latitudinal range with size. By far the best performing method is the multivariate normal approach that offers relatively high predictive capacity (R-squared 0.834 for Japanese longline data) when standard deviations are low (middle panel, bottom row Figure 1; σ_y=1.5 years, σ_s=10 days and σ_g=400km).

When extending the same approach to the Spanish longline fleet the method appears to work better at very near geographic scales. For example a standard deviation of 150km offered a significant increase in the predictive capacity (in terms of sample mean length) compared to 400km (Figure 3).

A plot of the cross-evaluation of the best performing method (multivariate normal, σ_y=1.5 years, σ_s=10 days and σ_g=400km) for the 50 largest length samples of the Japanese longline fleet reveals the reliability of the method at approximating sample-specific length frequency characteristics (Figure 4). A standout result is that there is very little variability among imputations when using this approach. All of the imputations (the red, blue and green lines of Figure 4) are similar and when they do not fit the real length frequency data (the black lines) they do not bracket the real data.

In several instances the imputation method fails to recreate clear bimodality in the length frequencies that may be describing important population size structure (for example second panel down, leftmost column of panels). There are a number of instances where the multivariate method appears to impute the data rather too well and it may be due to data reliability problems. For example the third and fourth row of panels from the bottom in Figure 4 show length samples with identical sample size numbers (n) and identical length frequency patterns. It is possible that these are the same data erroneously recorded twice in the Task II dataset with similar but not the same year, season and geographic location. A similar issue can be seen with sample sizes 2917, 2789 and 1274 that show identical frequency patterns and mean observed and predicted lengths.

In some cases the multivariate normal algorithm appears to follow potentially important patterns in the size composition such as bimodality. For example size samples n=3694 and n=2533 of Figure 2 appear to closely approximate the real size sample composition. However in general, the imputation performance of the multivariate method appears to be somewhat worse for the Spanish longline fleet (Figure 5). In several cross-validations the imputations completely miss a clear modality in the observed data (n=1488) or the imputed length frequency runs completely opposite to the pattern in the observed data (n=1218).

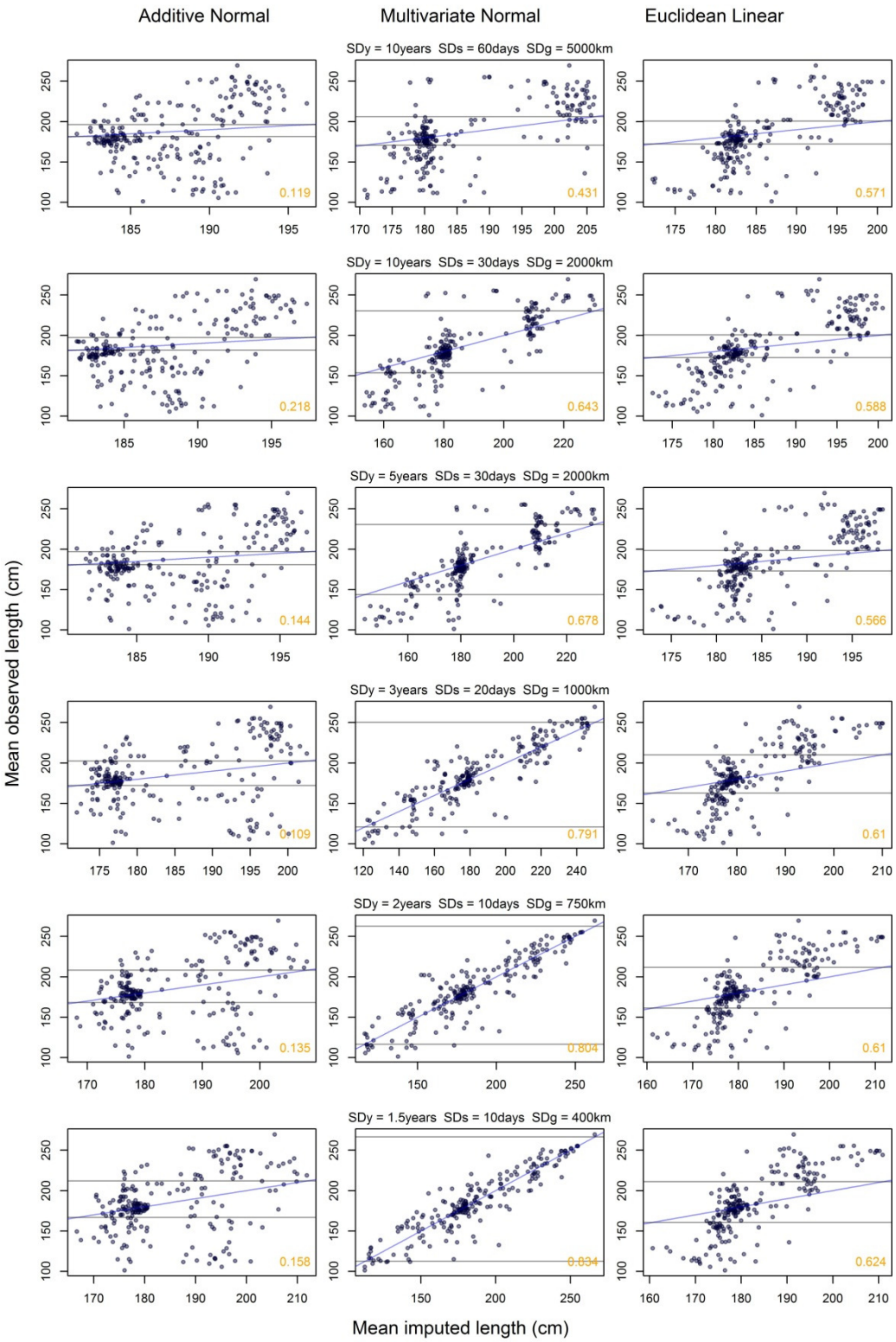


Figure 2. The predictive ability of the imputation model for the Japanese longline fleet. The mean of the observed length frequency samples plotted against mean imputed lengths for sample sizes over 200 (225 points). Each plot contains a 1:1 line and two horizontal lines representing the range of the predicted mean data. Each row represents different imputation conditions in terms of the standard deviation among years (σ_y=SDy), the seasonal standard deviation within year (σ_s=SDs) and the standard deviation in geographic space (σ_g=SDg). The orange number in the bottom right of each panel is the R-squared statistic representing the fraction of variance in observed mean length explained by the imputed mean length (assuming a linear model).

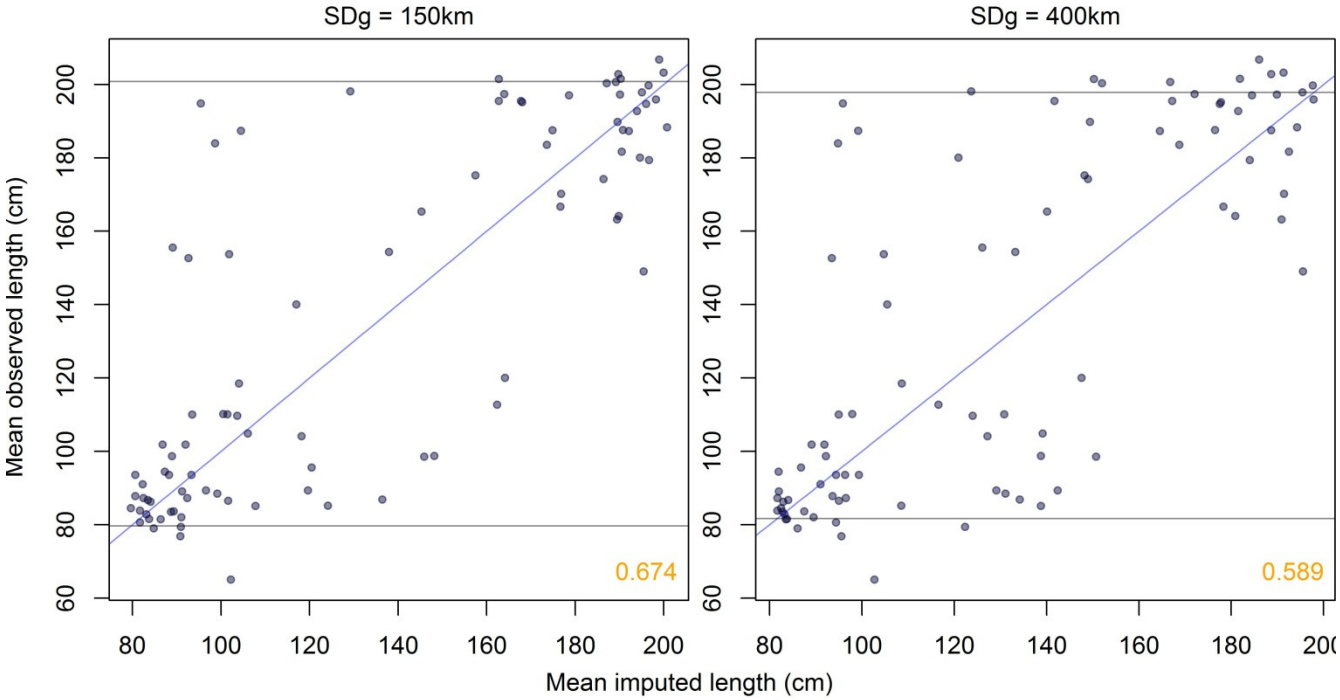


Figure 3. The predictive ability of the imputation model for the Spanish longline fleet. The mean of the observed length frequency samples plotted against mean imputed lengths for sample sizes over 200 (90 points). Each plot contains a 1:1 line and two horizontal lines representing the range of the predicted mean data. In both panels the standard deviation among years and the standard deviation within year are kept the same (SDy=1.5years, SDs=10days). The left-hand panel however includes a lower standard deviation and is more likely to sample lengths that are closer geographically (SDg=150km and SDg=400km respectively). The orange number in the bottom right of each panel is the R-squared statistic representing the fraction of variance in observed mean length explained by the imputed mean length (assuming a linear model).

Table 3. The performance of the different imputation algorithms given six different parameterizations. These are plotted graphically in Figure 2. The parameters σ_Y, σ_S and σ_G refer to the multivariate normal standard deviations for among years, within years (seasonally) and among geographic areas. Good performance is highlighted by three criteria that may be satisfied simultaneously, a positive regression slope of unity (between 0.9 and 1.1) that is statistically significant (p-value < 0.05) and explains over 50% of the variability in the observed lengths. Results fitting these criteria are highlighted in light grey.

Parameters			**Additive normal**			**Multivariate normal**			**Euclidean linear**		
σ_Y (years)	**σ_S (days)**	**σ_G (km)**	**Slope**	**p-value**	**R^2**	**Slope**	**p-value**	**R^2**	**Slope**	**p-value**	**R^2**
10	**60**	**5000**	3.496	1.03E-07	0.119	2.276	3.67E-29	0.431	4.207	7.02E-43	0.571
10	**30**	**2000**	4.06	1.44E-13	0.218	1.587	9.40E-52	0.643	3.858	8.98E-45	0.588
5	**30**	**2000**	3.041	3.81E-09	0.144	1.543	8.59E-57	0.678	4.363	2.91E-42	0.566
3	**20**	**1000**	1.282	4.00E-07	0.109	1.045	1.15E-77	0.791	3.071	1.85E-47	0.61
2	**10**	**750**	1.313	1.32E-08	0.135	0.984	8.39E-81	0.804	2.833	1.63E-47	0.61
1.5	**10**	**400**	1.361	5.91E-10	0.158	0.974	7.02E-89	0.834	2.987	2.94E-49	0.624

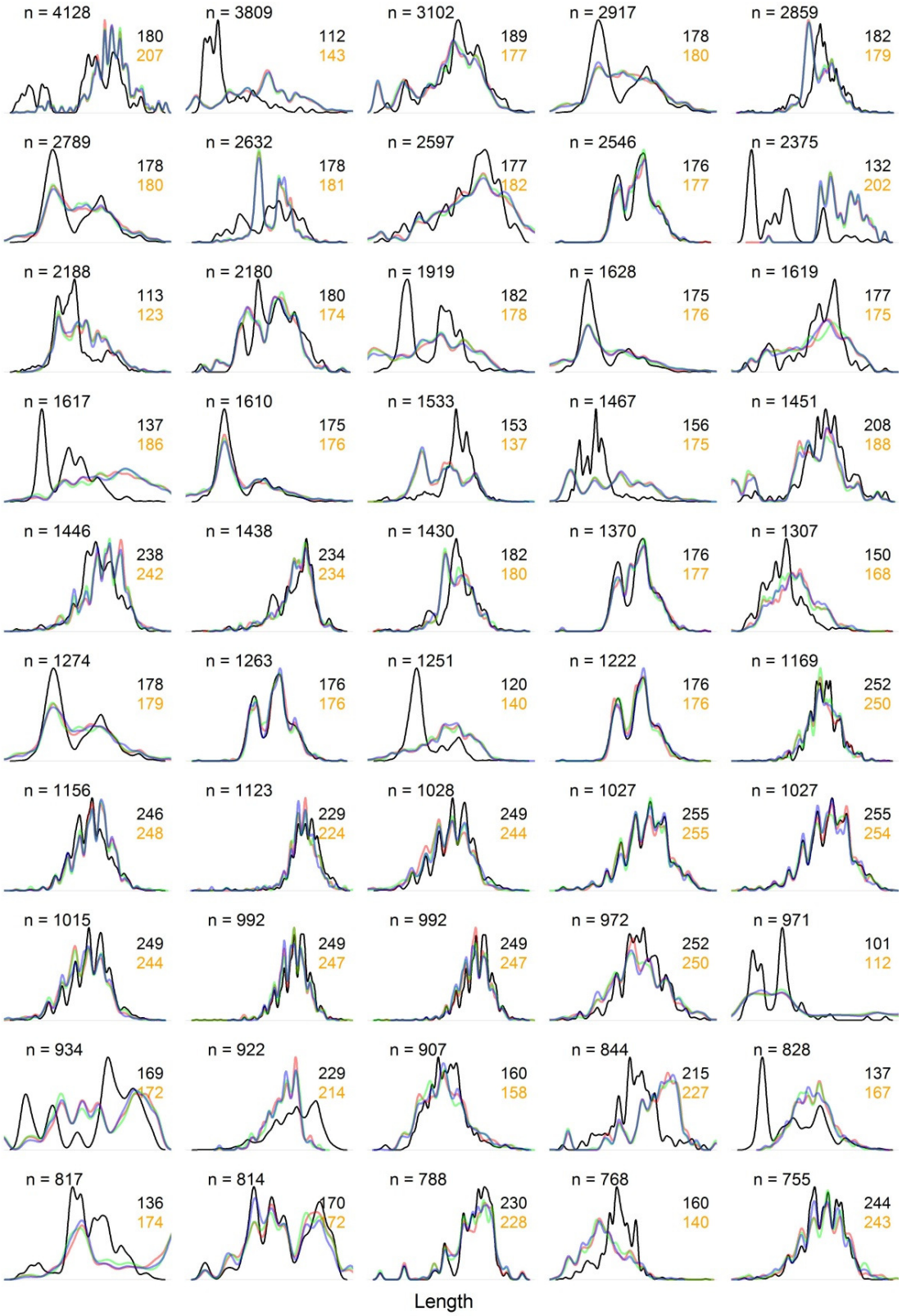


Figure 4. The observed and imputed length samples for the 50 largest samples of the Japanese longline fleet. The black line is the observed length frequency. The red, blue and green lines are three imputed length frequencies using the multivariate normal method (σ_Y =1.5years, σ_S =10days, σ_G =400km). The number in the top-left of each panel is the number length observations plotted. The two numbers in the top-right are the mean observed length (black) and the mean imputed length (orange).

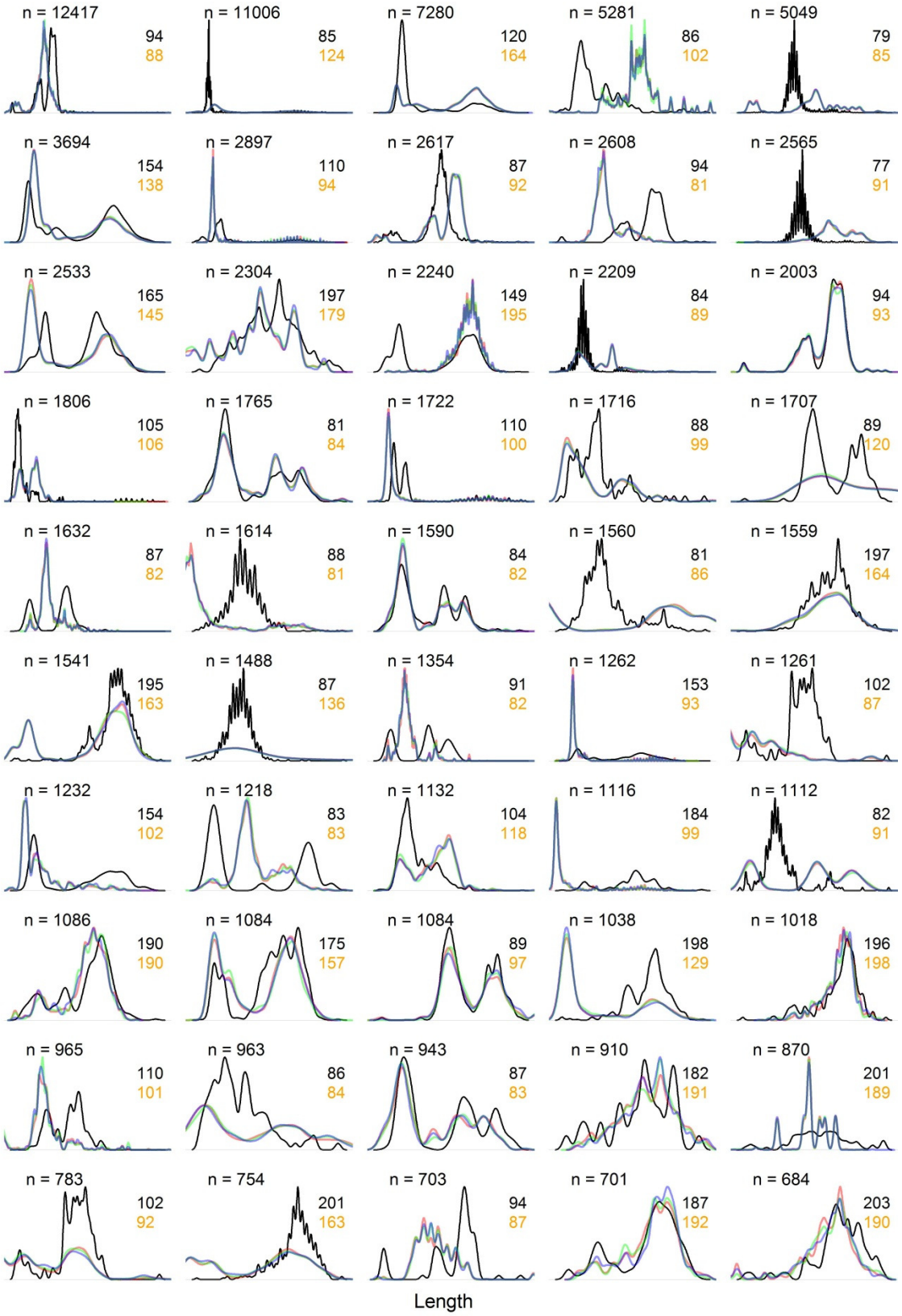


Figure 5. The observed and imputed length samples for the 50 largest samples of the Spanish longline fleet. The black line is the observed length frequency. The red, blue and green lines are three imputed length frequencies using the multivariate normal method (σ_Y =1.5years, σ_S =10days, σ_G =150km). The number in the top-left of each panel is the number length observations plotted. The two numbers in the top-right are the mean observed length (black) and the mean imputed length (orange).

5 Discussion

This brief report describes a relatively simple approach for the imputation of length frequency data for pelagic species such as Atlantic bluefin tuna. While the predictive capability of the method is strongly determined by how the 'distance' between observed and missing data are determined these may be tailored to a specific case using cross-validation. A more rigorous approach to tuning the imputation method using cross- validation may be optimization. For example it may be possible to numerically solve for the year, season and geographic standard deviation inputs that maximize the R-squared statistic for imputed versus observed means.

A core advantage of imputing length frequency data that is specific to time and location is that appropriate growth models may be applied to convert the lengths into ages. For example different growth rates for fish caught in Mediterranean and Gulf of Mexico areas. When combined with growth models, the multivariate normal imputation model assumed here may be used to investigate how uncertainty over imputing lengths affects the management recommendations of an assessment. This can be achieved by repeated-imputation inference. The analysis starts by imputing a number of length datasets that are all different, say 20. For each length dataset generated, a length-age conversion takes place that creates an individual total catch-at-age matrix for an assessment. For each imputation there is a corresponding assessment and a corresponding management recommendation. In this way the relative contribution of uncertainty from data processing can be quantified. Additionally it may be possible to integrate uncertainty in the length-age conversion by simply adding this to each imputed dataset (for example by sampling from distributions for values for growth parameters).

There are several reasons to be concerned about the length data that were used for cross-validation of the imputation algorithms. It is possible that such errors might provide an artificially favourable basis for performance evaluation. For example if identical data were erroneously replicated in the dataset in a similar, time or location, these data would provide a numerically perfect (albeit logically flawed) basis for imputing replicate data. This may explain three phenomena observed in the imputation of bluefin tuna longline data: (1) the strong tendency for performance to increase as 'distances' are made very small, (2) as distances are reduced the imputed data get very similar in composition (there is much less variability among imputations), (3) the very high R-squared number for the prediction of data that are inherently stochastic observations. The third point is perhaps the most compelling: R-squared values of almost 85% are very high considering that only time and space are used to predict a very complex sampling process. It seems likely that this is an artefact of erroneous replication of size samples. Possible mechanisms for this might be the replication of data from a wider area (*e.g.* 20 x 20 degree ocean square) to a number of smaller areas (the four 5 x 5 degree sub areas of the 20 x 20 degree square) or replication from course temporal block (quarter) to finer temporal block (month). A comprehensive data-vetting procedure should be undertaken prior to cross-validation to avoid incorrectly tuning the approach. Alternatively, a spatially explicit population dynamics model could be used to simulate data to test the imputation algorithms.

Assuming that the Japanese longline data are correctly reported, it proved considerably harder to impute the Spanish longline length compositions. It is likely that covariate data other than time, season and location could provide a better basis for predicting missing lengths. Such covariates could include specific gear type (rather than the coarser longline 'gear group code') and set depth information (or a proxy such as hooks-per-basket). The imputation for the Spanish longline data helped to illustrate the variable computation demand of the multivariate normal imputation approach that is squared in nature. While the Japanese imputation (1250 length frequency samples imputed from 41 000 observations) took 45 seconds to complete, the Spanish longline imputation (260 length frequency samples imputed from 6700 observations) took just 3 seconds. Due to the design of the algorithm replicate imputations incur virtually no computation cost (for example less than a 1% increase in computation time per 20 replicates for the Japanese longline imputation). Since the multivariate normal probability density function operates on each line (each set of length frequencies to be imputed) the process can be re-coded to make use of parallel processing using R packages such a snowfall.

Development priorities for the imputation algorithm should focus on improving the bracketing behaviour of the model: where the method fails to predict length frequency patterns the imputations should at the very least

overlap with the known length composition. Other possible avenues for development could include the separation of longitudinal and latitudinal distance to form a four-dimensional multivariate distribution and the addition of artificial error to increase the variability among imputations.

It should be noted that imputing lengths according to spatio-temporal model may generate artificial patterns in the data that are inconsistent with biological or ecological theory that have important consequences for subsequent management implications. An alternative to imputation is to test alternative approaches that do not require complete catch composition data. For example both Multifan-CL (Fournier *et al.* 1990) and CASAL (Bull *et al.* 2012) allow length frequency data to inform size structure without complete records of catch at length or cohort slicing. A possible downside to such an approach may be the inability to fit a spatial assessment model and therefor account for regional growth characteristics.

The non-parametric approaches of this preliminary results paper were chosen after investigating other modelling options. Initial investigation of Conditional Auto-Regressive (CAR) models and GLM modelling of parametric length compositions revealed a relatively poor predictive capacity. Both approaches rely on the formulation of a flexible parametric model of length compositions. The principal problem was the inability of such methods to approximate multimodal length frequency distributions and the computational demand of creating multiple imputations.

In this paper I describe a computationally efficient approach to the imputation of length data for Atlantic bluefin tuna. While the approach requires fine tuning and further development it can generate stochastic length samples that through repeated-imputation inference can quantify the uncertainty from data-processing. The method offers a reproducible alternative to the current status quo that is statistically rigorous and can be easily applied to stocks of other pelagic species such as tropical tunas.

6 Acknowledgements

Many thanks to both Carlos Palma and Laurie Kell for their helpful comments and advice on methods and data.

7 References

Carruthers, T.R., Kell, L. 2013. Imputation of Catch-At-Length. ICAL. GitHub Repository. https://github.com/tcarruth/ICAL.git

Bull, B., Francis, R.I.C.C., Dunn, A., McKenzie, A., Gilbert, D.J., Smith, M.H., Bain, R., Fu, D. (2012). CASAL (C++ algorithmic stock assessment laboratory): CASAL user manual v2.30-2012/03/21 . NIWA Technical Report 135. 280 p.

Fournier, D.A., Sibert, J.R., Majkowski, J., Hampton, J. (1990). MULTIFAN: a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (Thunnus maccoyii). Canadian Journal of Fisheries and Aquatic Sciences 47, 301-317.

Gavaris, S. 1988. An Adaptive framework for the estimation of population size. CAFSAC Res. Doc. 88/29. 12pp.Gelman, A., J. B. Carlin, H. Stern, and D. Rubin 1995. Bayesian Data Analysis. Chapman and Hall, London.

Junninen, H., H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen. 2004. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38:2895- 2907.

Little, R. J. A., and D. B. Rubin 1987. Statistical Analysis with Missing Data. Wiley, New York.

Little, R. J. A., and N. Schenker. 1995. Missing data in G. Arminger, C. C. Clifford, and M. E. Sobel, editors. Handbook of Statistical Modeling for the Social and Behavioral Sciences. Plenum Press, New York.

Mishra, G. D., and A. J. Dobson. 2004. Multiple imputation for body mass index: lessons from the Australian Longitudinal Study on Women's Health. Statistics in Medicine 23:3077-3087.

Nielsen, S. F. 2003. Proper and improper multiple imputation. International Statistical Review 71:593-607.

Olinsky, A., S. Chen, and L. Harlow. 2003. The comparative efficacy of imputation methods for missing data in structural equation modeling. European Journal of Operational Research. 151:53-79.

R Core Team. 2013. R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-9. http://www.R-project.org.

Rubin, D. B. 1976. Inference and Missing Data. Biometrika 63:581-592.

Rubin, D. B. 1987. Multiple Imputations for Nonresponse in Surveys. John Wiley and Sons, New York.