European hake stock structure in the Mediterranean as assessed by otolith shape and microchemistry

Beatriz Morales-Nin, Sílvia Pérez-Mayol, Kirsteen MacKenzie, Ignacio A. Catalán, Miquel Palmer, Thibault Kersaudy, Kélig Mahé

Appendix A: Supplementary material

*S1. Otolith shape: data collection and pre-processing*

After extraction and cleaning, images of the whole left and right sagittal otoliths were scanned under reflected light and saved at high resolution (3,200 dpi). The image acquisition process followed a standardized protocol to minimize user-related biases. Elliptic Fourier Descriptor parameters (EFDs) describing the otolith outline were obtained from binarized images using TNPC 7 software (IFREMER, France). EFDs were normalized with respect to the first harmonic to be invariant to otolith size, rotation and starting point (Kuhl and Giardina, 1982). Provided that fish size and otolith size are well- and linearly correlated (Morales-Nin et al., 1998), we assume that the normalization procedure above also removed any (linear) bias related with fish size (such as those described by Rencher and Christensen, 2012).

To determine the number of harmonics required to safely reconstruct the otolith outline, the cumulated Fourier power (F) was calculated for each otolith $k$ as a measure of the precision of contour reconstruction obtained with $n\_{k} $harmonics (i.e., the proportion of variance in contour coordinates accounted for by the $n\_{k}$ harmonics):

$$F\_{(n\_{k})}= \sum\_{i=1}^{n\_{k}}\frac{A\_{i}²+B\_{i}²+C\_{i}²+D\_{i}²}{2}$$

where $A\_{i}$, $B\_{i}$, $C\_{i}$ and $D\_{i}$ are the coefficients of the $ith$ harmonic. The first 50 harmonics were enough to ensure outline reconstruction with a precision of 99.99% (as suggested by Lestrel, 2008), thus 50 harmonics were used for further analyses.

The resulting matrix containing EFDs (as columns) for each otolith (as rows) was submitted to a principal components analysis (PCA) (Rohlf and Archie, 1984) to reduce matrix dimensionality. PCA was completed using the function *rda* from the *vegan* package (Oksanen et al., 2020) of the R package. The number of Principal Components (PCs) to be retained was determined using the Auer-Gervini algorithm (Auer and Gervini, 2008) as implemented by the function *AuerGervini* from the *PCDimension* package (Coombes and Wang, 2019).

A recent study showed that the strength of directional bilateral asymmetry may vary between different locations within the same species. Thus, *a priori* choosing one otolith (right *versus* left) can introduce some bias when identifying stocks using otolith shape (Mahé et al., 2019). Therefore, the existence of directional asymmetry of otolith shape was tested using the *rda* function from the *vegan* library (Oksanen et al., 2019), according to the model proposed elsewhere (Díaz-Gil et al., 2015; Mahé et al., 2019; Palmer et al., 2010). Given that directional asymmetry was not found, only the left otolith was analysed hereafter. Sexual dimorphism was tested using the same approach, and it was also ignored hereafter as it had no significant effect on otolith shape.

To visualise the differences in otolith shape between different units/locations, the average otolith outline of all otoliths from a given unit was estimated from EFDs, and the pair-wise percentage of non-overlapping surface between the estimated shapes was calculated.

After all the data pre-processing described above, the *large sample* matrix (see main text) used for classification and multivariate analyses of Mediterranean fish using otolith shape was composed of 1,656 otoliths (rows) and 11 PCs (columns) from 40 GSAs.

*S2. Otolith microchemistry: analyses and data pre-processing*

From all the available otoliths above, a subsample of 279 otoliths belonging to females from 10 GSAs (GSA 1b, 9b, 11c, 16b, 18a, 20a, 22b, 25c, 26a, and 27b), and from 2 NE Atlantic divisions (VIIIa and VIIIc) were chosen for microchemical analyses. The number of otoliths analysed per GSA (around 25) was determined according to previous results (Morales-Nin et al., 2005, 2014).

All the otoliths were prepared under clean conditions following the protocols described for European hake (Morales-Nin et al., 2014). The otoliths were analysed using a CETAC Teledyne Technologies Laser Ablation System LSX-213 G2+ coupled to a Thermo-Finnigan Element XR Inductively Coupled Plasma-Mass Spectrometer (LA-ICP-MS) working at medium resolution. An area of 200 µm length and 30 µm width was shot on the otolith edge following a straight line, at a scan speed of 5 µm.sec-1 and a fluency of ~7.5 J.cm-2. Up to six control samples of known, certified concentration reference materials (CRMs) were also analysed (NIST612, NIST614, NIST616 (National Institute of Standards and Technology); FEBS-1 (Sturgeon et al., 2005); NIES-22 (Yoshinaga et al., 2000); and MACS-3 (International Association of Geoanalysts)) following a bracketing protocol. A suit of 27 different isotopes was determined: 6Li, 7Li, 11B, 23Na, 24Mg, 25Mg, 27Al, 28Si, 31P, 43Ca, 44Ca, 45Sc, 55Mn, 56Fe, 59Co, 60Ni, 63Cu, 66Zn, 85Rb, 88Sr, 118Sn, 137Ba, 138Ba, 206Pb, 207Pb, 208Pb and 238U; 42Ca was used as internal standard. For each one of the laser-shots and for every measured isotope, the spectrometer provides a temporal profile (intensities over time of counts per second, cps) that is composed of a background interval at which no sample is ablated (*blank*) and a *plateau* interval at which intensity values reach a noisy but steady level after which the signal returns to the blank level again. These raw intensities must be transformed into isotope concentrations (µgisotope · gsample-1 or ppm). This procedure requires several steps: i) the selection of *blank* and *plateau* intervals of every LA-shot (or LA-scan), ii) data transformation and reduction, iii) spectrometer temporal drift correction, iv) normalization by 42Ca as internal standard, and v) quantification of estimated element concentrations in CRMs and otoliths.

The first step (i) was conducted using a Long Short-Term Memory (LSTM) Network built in MatLab and developed by the *Data Processing Group of the University of Vic* and *IMEDEA* (Martí-Puig et al., 2019); the network was trained using previous supervised shots so that it is able to automatically identify and select the *blank* and *plateau* intervals of the raw data for any specific LA-shot. The second part of the data processing (steps ii to v) was implemented in a hierarchical model developed in R and Jags (<https://mcmc-jags.sourceforge.io/>) by the *Fish Ecology Lab* at *IMEDEA*. The procedure is as follows: the distribution of count per second (cps) for a given scan and isotope was determined by subtracting the distribution of cps at the *blank* from the distribution of cps at the *plateau*. When the *plateau-blank* difference was not larger than zero (prob. > 0.05), the scan concentration of the corresponding isotope was considered undetectable (missing data). Next, the *plateau-blank* difference for any given isotope was normalized (ratio) by the *plateau-blank* difference for 42Ca. This normalization (known as *internal standard normalization*) is a widely applied procedure for correcting the so-called matrix effect; that is, disparities related only to structural differences between (synthetic and homogeneous) CRMs and biogenic samples (otoliths). The internal standard used (42Ca) corresponds to a major component of the sample and it is assumed to be homogeneously distributed within it. When a ratio of element:42Ca was not higher than zero (prob > 0.05), the concentration of the corresponding isotope was also considered undetectable (missing value). Finally, the element:42Ca ratio from the CRMs (as dependent variable) was assumed to be a linear combination of the certified concentration for each element (ppm) after correcting by the naturally occurring isotope abundance, the session (categorical variable accounting for session-specific random effects common to all the LA-shots from a given working session), and the time elapsed between the session's start and end (to account for any linear temporal drift of the signal intensity throughout a given session).

For a given session and a given element, the parameters of a statistical model including all the above variables were estimated using a Bayesian approach. The parameterized model was then used to estimate each LA-scan's concentration (ppm) on otoliths, after accounting for the effects of session and temporal drift. In case of poor convergence of the Bayesian analyses, the corresponding isotope concentration was also considered as missing data. Finally, one of the six CRMs (MACS-3) was treated as a sample of unknown concentration. In case of discrepancy between the certified and the estimated concentration, the LA-shot concentration of the corresponding isotope was also considered as missing data.

Before the classification analysis, the raw microchemical data were tested for normality at the within-group level using the function *mvn* from the *MVN* package (Korkmaz et al., 2014). 24Mg, 55Mn and 138Ba showed non-normal distribution and were transformed by the Box-Cox transformation using the function *boxcox* of the *MASS* package (Venables and Ripley, 2002). This function estimates the likelihood profile for a range of 𝜆 values. The value of 𝜆 showing the maximum likelihood was used for transforming the data using: transformed values = 𝜆−1(raw values𝜆−1).

After a preliminary data exploration, the raw data contained a large number of missing values, thus conventional classification and/or multivariate analyses could not be completed. The conventional approaches for dealing with this problem are deleting variables (elements) and/or samples (otoliths) with missing values. However, this was not an option here because it would reduce the raw data to very few otoliths and elements. Therefore, we adopted a mixed strategy: in a first step, variables (elements) with less than 30% of valid data and otoliths with four or more missing values were deleted. Note that even using these undemanding thresholds, the number of remaining missing values was still high (22%). Therefore, a missing data imputation method was applied in a second step to fill those gaps (Chlioui et al., 2019; Ćwiklińska-Jurkowska et al., 2005). Imputation methods for multivariate data can be based (1) on replacing a given missing value for the corresponding value from the most similar case in the matrix, or (2) on estimating the missing values from the variance/covariance matrix. For the first alternative, we used the *impute.knn* function from the *impute* library of the R package. In that case, a k-nearest neighbours’ algorithm with *k* = 3 was used to find the most similar otoliths in the matrix and the missing data were replaced with the mean values of those three otoliths (Troyanskaya et al., 2001). For the second alternative, the function *imputeData* from the *mclust* and *mix* packages was used (Schafer and Olsen, 1998). Given that similar results were achieved using both methods, hereafteronly the results obtained with the first method (*knn*) were used.

Provided the many classifications methods currently available, some preliminary comparisons were completed. We measured the rate of correct predictions after a leave-one-out cross-validation of 12 computing-intensive methods as they are implemented in the *RWeka* library (Witten et al., 2005; J48, LMT, DecisionStump, Logistic, SMO, IBk, AdaBoostM1, Bagging, LogitBoost, JRip1, OneR and PART). In addition, Linear Discriminant Analyses (LDA) and Quadratic Discriminant Analyses (QDA) were completed using the functions implemented in the *MASS* library (Venables and Ripley, 2002) (Table S2). Since the best classification success results were obtained with LDA (as in Jones et al., 2017), hereafter only LDA was used. Preliminary trials were also completed to compare the missing data imputation procedure at the within-GSA scale *versus* the overall scale (i.e., including all otoliths and ignoring GSA membership). Since the results in terms of classification capability were virtually the same, we ignored GSA membership for missing data imputation in order to avoid any circularity.

After the data pre-processing described above, the *small sample* matrix (see main text) used for classification and multivariate analyses of Mediterranean fish using otolith shape and microchemistry (see Section 2.4) was composed of 154 otoliths (rows) and 7 elements (23Na, 24Mg, 43Ca, 44Ca, 55Mn, 88Sr and 138Ba).

# References

Auer, P., Gervini, D., 2008. Choosing Principal Components: A new graphical method based on Bayesian model selection. Commun. Stat.- Simul. C. 37(5), 962-977. <https://doi.org/10.1080/03610910701855005>.

Chlioui, I., Idri, A., Abnane, I., de Gea, J.M.C.Fernández-Alemán, J.L., 2019. Breast cancer classification with missing data imputation, in: Rocha, Á., Adeli, H., Reis, L., Costanzo, S. (Eds.) New knowledge in information systems technologies. WorldCIST'19 2019. Advances in intelligent systems and computing, vol 932*.*, Springer, pp. 13-23. <https://doi.org/10.1007/978-3-030-16187-3_2>.

Coombes, K.R., Wang, M., 2019. PCDimension: Finding the number of significant Principal Components. R package version 1.1.11. [https://CRAN.R-project.org/package=PCDimension.](https://CRAN.R-project.org/package%3DPCDimension%20.)

Ćwiklińska-Jurkowska, M., Jurkowski, P., Kukulska-Pawluczuk, B., Kołtan, A., Drózdz, W., Hilemann, W., 2005. Different approaches to missing data in discrimination applied to medical problems. Ann. Acad. Med. Stetin*.* 50(S2), 56-58.

Díaz-Gil, C., Palmer, M., Catalán, I.A., Alós, J., Fuiman, L.A., García, E., Gil, M.M., Grau, A., Kang, A., Maneja, R.H., Mohan, J.A., Morro, B., Schaffler, J.J., Buttay, l., Riera-Batle, I., Tolosa, B., Morales-Nin, B., 2015. Otolith fluctuating asymmetry: a misconception of its biological relevance?. ICES J. Mar. Sci. 72(7), 2079-2089. [https://doi.org/ 10.1093/icesjms/fsv067](https://doi.org/%2010.1093/icesjms/fsv067).

Jones, C.M., Palmer, M., Schaffler, J.J., 2017. Beyond Zar, the use and abuse of classification statistics for otolith chemistry. J. Fish Biol. 90(2), 492-504. [https://doi.org/ 10.1111/jfb.13051](https://doi.org/%2010.1111/jfb.13051).

Korkmaz, S., Goksuluk, D., Zararsiz, G., 2014. MVN: An R package for assessing multivariate normality. R J. 6(2), 151-162.

Kuhl, F., Giardina, C., 1982. Elliptic Fourier features of a closed contour. Comput. Graph. Image Proc. 18, 236-258. [https://doi.org/ 10.1016/0146-664X(82)90034-X](https://doi.org/%2010.1016/0146-664X%2882%2990034-X).

Lestrel, P.E., 2008. Fourier descriptors and their applications in Biology, 1st ed. Cambridge University Press. [https://doi.org/ 10.1017/CBO9780511529870](https://doi.org/%2010.1017/CBO9780511529870).

Mahé, K., Ider, D., Massaro, A., Hamed, O., Jurado-Ruzafa, A., Gonçalves, P., Anastasopoulou, A., Jadaud, A., Mytilineou, C.E.R., Ramdane, Z., Bacha, M., Amara, R., de Pontual, H., Ernande, B., 2019. Directional bilateral asymmetry in otolith morphology may affect fish stock discrimination based on otolith shape analysis. ICES J. Mar. Sci. 76(1), 232-243. [https://doi.org/ 10.1093/icesjms/fsy163](https://doi.org/%2010.1093/icesjms/fsy163).

Martí-Puig, P., Pérez-Mayol, S., Serra-Serra, M., Palmer, M., 2019. Automatic interval selection on LA-ICPMS isotopic profiles using an Extreme Learning Machine (ELM) network, in: Sabater-Mir, J., Torra, V.A.I., González-Hidalgo, M. (Eds.), Frontiers in artificial intelligence and applications, IOS Press, pp. 74-83. [https://doi.org/ 10.3233/FAIA190110.](https://doi.org/%2010.3233/FAIA190110.)

Morales-Nin, B., Torres, G.J., Lombarte, A., Recasens, L., 1998. Otolith growth and age estimation of European hake. J. Fish Biol*.* 53, 1155-1168. <https://doi.org/10.1111/j.1095-8649.1998.tb00239.x>.

Morales-Nin, B., Swan, S.C., Gordon, J.D.M., Palmer, M., Geffen, A.J., Shimmield, T., Sawyer, T., 2005. Age-related trends in otolith chemistry of *Merluccius merluccius* from the north-eastern Atlantic Ocean and the western Mediterranean Sea. Mar. Freshwater Res. 56(5), 599-607. [https://doi.org/ 10.1071/MF04151](https://doi.org/%2010.1071/MF04151).

Morales-Nin, B., Pérez-Mayol, S., Palmer, M., Geffen, A.J., 2014. Coping with connectivity between populations of *Merluccius merluccius*: An elusive topic. J. Marine Syst. 138, 211-219. <https://doi.org/10.1016/j.jmarsys.2014.04.009>.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., 2020. vegan: Community Ecology Package. R package version 2.5-7. [https://CRAN.R-project.org/package=vegan](https://CRAN.R-project.org/package%3Dvegan).

Palmer, M., Linde, M., Morales-Nin, B., 2010. Disentangling fluctuating asymmetry from otolith shape. Mar. Ecol. Prog. Ser*.* 399, 261-272. [https://doi.org/ 10.3354/meps08347](https://doi.org/%2010.3354/meps08347).

Rencher, A.C., Christensen, W.F., 2012. Methods of multivariate analysis, 2nd ed. John Willey & Sons.

Rohlf, F.J., Archie, J.W., 1984. A comparison of Fourier methods for the description of wing shape in mosquitoes (Diptera:Culicidae). Syst. Biol. 33(3), 302-317. [https://doi.org/ 10.2307/2413076](https://doi.org/%2010.2307/2413076).

Schafer, J.L., Olsen, M.K., 1998. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivar. Behav. Res*.* 33(4), 545-571. [https://doi.org/ 10.1207/s15327906mbr3304\_5](https://doi.org/%2010.1207/s15327906mbr3304_5).

Sturgeon, R.E., Willie, S.N., Yang, L., Greenberg, R., Spatz, R.O., Chen, Z., Scriver, C., Clancy, V., Lam, J.W., Thorrold, S., 2005. Certification of a fish otolith reference material in support of quality assurance for trace element analysis. J. Anal. Atom. Spectrom. 20(10), 1067-1071. [https://doi.org/ 10.1039/B503655K](https://doi.org/%2010.1039/B503655K).

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17(6), 520-525. [https://doi.org/ 10.1093/bioinformatics/17.6.520](https://doi.org/%2010.1093/bioinformatics/17.6.520).

Venables, W.N., Ripley, B.D., 2002. Modern applied statistics with S, 4th ed.Springer, New York.

Witten, I.A., Frank, E., Hall, M.A., 2005. Data mining: Practical machine learning tools and techniques, 2nd ed., Morgan Kaufmann, San Francisco.

Yoshinaga, J., Nakama, A., Morita, M., Edmons, J., 2000. Fish otolith reference material for quality assurance of chemical analyses., Mar. Chem*.* 69(1-2), 91-97. [https://doi.org/ 10.1016/S0304-4203(99)00098-5](https://doi.org/%2010.1016/S0304-4203%2899%2900098-5).

Table S2. Correct classification rate after leave-one-out cross-validation.

|  |  |  |
| --- | --- | --- |
| Method | R library | Correct classification rate |
| LDA | MASS | 0.30 |
| QDA | MASS | 0.26 |
| J48 | RWeka | 0.21 |
| LMT | RWeka | 0.30 |
| DecisionStump | RWeka | 0.08 |
| Logistic | RWeka | 0.30 |
| SMO | RWeka | 0.27 |
| IBk | RWeka | 0.22 |
| AdaBoostM1 | RWeka | 0.08 |
| Bagging | RWeka | 0.26 |
| LogitBoost | RWeka | 0.26 |
| JRip1 | RWeka | 0.13 |
| OneR | RWeka | 0.18 |
| PART | RWeka | 0.23 |