# Forecasting recruitment of multiple fish species using multi-dimensional Bayesian network classifiers

*Jose A. Fernandes[1,2], Jose A. Lozano, Iñaki Inza, Xabier Irigoien, Juan D. Rodríguez, Aritz Pérez*

1 Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth, U.K. PL13 DH. jfs@pml.ac.uk.
2 AZTI-Tecnalia, Marine Research Division, Herrera Kaia z/g. E-20110 Pasaia, Spain

## Summary
A multi-species approach to fisheries management requires taking into account the interactions between species in order to improve recruitment forecasting. Recent advances in Bayesian networks direct the learning of models with several interrelated variables to be forecasted simultaneously. These are known as multi-dimensional Bayesian network classifiers (MDBNs). Pre-processing steps are critical for the posterior learning of the model in these kinds of domains. Therefore, in this study, a set of 'state-of-the-art' uni-dimensional pre-processing methods, within the categories of missing data imputation, feature discretization and subset selection, are adapted to be used with MDBNs. A framework that includes the proposed multi-dimensional supervised pre-processing methods, coupled with a MDBN classifier, is tested for fish recruitment forecasting. The correctly forecasting of three fish species (anchovy, sardine and hake) simultaneously is doubled (from 17.3% to 29.5%) using the multi-dimensional approach in comparison to mono-species models. The probability assessments also show high improvement reducing the average error (Brier score) from 0.35 to 0.27. These differences are superior to the forecasting of species by pairs.

## Introduction
Fisheries management requires forecasting tools to assess the recruitment of commercial fish species in order to control their exploitation. Many studies have addressed this problem by selecting the features and learning a model for each species in isolation (Andonegi et al; 2011) or try to incorporate interactions between species (Fernandes et al., 2013). Classification methods have been also used as mono-species forecasting approaches (Fernandes et al, 2010). However, Multi-dimensional Bayesian networks (MDBNs) permit the learning of classifiers that have multiple class variables in a single model under high uncertainty (van Der Gaag and De Waal, 2006). It would be desirable to be able to combine these MDBNs in a pipeline with pre-processing methods that simultaneously target the forecasting of several species, taking advantage of the relationships between species. Therefore, pre-processing supervised methods adaptation (missing data imputation, discretization and feature subset selection) is needed for the multi-dimensional approach. Within this context, the objectives of this study are: i) to develop pre-processing strategies for multi-dimensional (Mul-D) classifiers based upon uni-dimensional (Uni-D) state-of-the-art methods; ii) to test the proposed pre-processing methods with a real problems of fish recruitment forecasting.

## Material and Methods
A naive Bayes classifier generalized to multiple class variables is selected. Due the novelty of the multi-dimensional classification paradigm several methodological contributions had been proposed before its application: 1) two commonly-used (uni-dimensional) performance measures have been generalized (accuracy and Brier score) to consider multiple response variables; and, 2) pre-processing methods (missing data imputation, feature discretization and feature selection) have been adapted to have multiple objectives. A 10 times repeated 5-fold cross-validation (10x5cv) schema has been selected (Rodríguez et al., 2013). So as to avoid model over-fitting and provide an honest validation, the entire pipeline (data pre-processing and model) is included in the validation scheme (Fernandes et al., 2010); i.e. the data partition, in folds, is performed before the first step of the pipeline. The selected species were Anchovy, Sardine and Hake. Anchovy and hake are species of high commercial interest

in the Bay of Biscay that share the ecosystem with sardine and, consequently, they share the same feature candidates.

## Results and Discussion

The application of the propose multi-dimensional/species methodology is compared with equivalent methodology using single species methodology (Table I). In general, the multi-species approach improves the single-species approach in the accuracy of each species (often without statistically significant differences). However, a key issue in this domain is that the multi-dimensional approach improves in terms of the single species probabilities estimation (Brier score) of each species with differences that are statistically significant. In addition, the most important result is that there are notable improvements of the joint accuracy, when the multi-species approach is used, i.e. the chance of being correct, at the same time, in all of the species is higher than using a single-species classifier for each species. In fact, the increase of joint accuracy is higher than average accuracy of single species, or the accuracy of each species. This simultaneous improvement in all species is crucial in terms of the ecosystem-based fisheries management approach. The principal objective of proposing a set of multi-dimensional pre-processing approaches, appropriate for fisheries management, is achieved. The application to a real oceanographic problem reveals benefits, improving the forecasting of fish recruitment. Firstly, improvement in the forecasting of each species is achieved. Secondly, a significant improvement in the chance of simultaneous correct forecast for all of the species, which is a key issue for the ecosystems management approach, can be highlighted. Finally, significant improvement in the a posteriori estimated class probabilities, which leads to better informed decisions, can be shown (single, average and joint Brier score). This is a key objective in knowledge-based fisheries management.

| Pre-processing pipeline | ARI Acc. | ARI BS | SR Acc. | SR BS | HR Acc. | HR BS | Joint Acc. |
|---|---|---|---|---|---|---|---|
| CM-MID-CFS (Uni-D) | $52.7 \pm 6.7$ | 0.36 | $55.7 \pm 6.7$ | 0.34 | $67.6 \pm 3.3$ | 0.27 | $17.3 \pm 4.8$ |
| CMcart-MIDmean-CFSsum | $54.6 \pm 7.3$ | 0.35 | **$65.4 \pm 5$** | 0.27 | $72.9 \pm 5.5$ | 0.21 | $28.9 \pm 4.5$ |
| CMcart-MIDindiv-CFScart | $46.5 \pm 4.3$ | 0.32 | $59.8 \pm 6.3$ | **0.24** | $71.4 \pm 5.8$ | 0.19 | $22.6 \pm 4.3$ |
| CMcart-MIDmean-CFSmean | $45 \pm 7.6$ | 0.32 | $58.4 \pm 6$ | 0.25 | **$75 \pm 4.6$** | **0.18** | $19.7 \pm 5.5$ |
| CMcart-MIDmean-CFScart | **$57.9 \pm 5$** | **0.30** | $60.6 \pm 4.8$ | 0.27 | $68.9 \pm 7.3$ | 0.21 | **$29.5 \pm 4$** |
| CMcart-MIDmean-CFSindiv | $53.8 \pm 4.8$ | 0.32 | $63.4 \pm 2.9$ | 0.27 | $71.6 \pm 6.1$ | **0.18** | $28.5 \pm 4.7$ |

Table 1. Comparison of multi-dimensional *vs* uni-dimensional approach for anchovy, sardine and hake. Best results are emphasized in bold. The first line represents the use of single species methods. The rest of lines represent different combinations of multi-dimensional methods. Brier score (BS) is the probabilistic error. Joint accuracy (Acc.) is the chance of being right in all the species simultaneously.

## References

Andonegi, E., Fernandes, J.A., Quincoces, I, Uriarte, A, Pérez, A., Howell, D., Stefansson, G., 2011. The potential use of a Gadget model to predict stock responses to climate change in combination with Bayesian Networks: the case of the Bay of Biscay anchovy. ICES. J. Mar. Sci., 68: 1257-1269.

Fernandes, J.A.; Irigoien, X.; Goikoetxea, N.; Lozano, J.A.; Inza, I.; Pérez, A.; Bode, A., 2010. Fish recruitment prediction, using robust supervised classification methods. Ecol. Model., 221 (2): 338-352.

Fernandes, J.A., Cheung, W.W.L., Jennings, S., Butenschön, M., Mora, L., Frölicher, T.L., Barange, M., *eta al.*, 2013. Modelling the effects of climate change on the distribution and production of marine fishes: accounting for trophic interactions in a dynamic bioclimate envelope model. Global Change Biol., 19(8): 2596-2607.

Rodríguez, J.D., Pérez, A., Lozano, J.A., 2013. A General Framework for the Statistical Analysis of the Sources of Variance for Classification Error Estimators. Pattern Recognition, 46(3): 855-864.

Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol. Model., 203 (3-4): 312-318.

Van Der Gaag, L.C., De Waal, P.R., 2006. Multi-dimensional Bayesian network classifiers. Technical report CS-UU 2006-056.