Working document to *Transversal Workshop on Fishing Effort Parameters Identification and Measurement Standardization*, Fuengirola (Malaga), Spain 30-31 May 2006

# New improvements in predicting fishing tactics

Joan Moranta[1*], Miquel Palmer[2], Enric Massutí[1] Beatriz Guijarro[1], Francesc Ordines[1] and Antoni Quetglas[1]

[1] IEO- Centre Oceanogràfic de les Balears, P.O. Box 291, 07080 Palma, Spain.

[2] CSIC/UIB- Institut Mediterrani d'Estudis Avançats, Miquel Marquès 21, 07190 Esporles, Spain.

(*) author for correspondence: joan.moranta@ba.ieo.es

## ABSTRACT

Multispecies fisheries in the Mediterranean are versatile.Even during the same fishing trip, which typically comprises a single day, one vessel can use more than one fishing strategy. In the case of the Mallorcan fleet, up to 38% of the fishing trips used more than one tactic. Here, a new method that uses the information gathered in the daily sale bill is proposed to predict the fishing tactics used in a single fishing trip. Four categories were defined a priori depending on the main target species. Daily sales bill and fishing tactics (determined by an on board observed) were available for 224 fishing trips. Abundance data of the commercial categories were submitted to a preliminary Principal Component Analysis and/or Correspondence Analysis and the resulting first few axes were retained in order to define the daily catches. A Generalized Regression neural network (GRNN) was parameterized using 200 fishing trips, and prediction capability was tested using the remaining 24ones. The new approach has proved to be accurate and precise. The percentage of correct classifications of 500 runs (i.e., randomly changing the training and testing data sets) rise up to 93-94% of correct predictions. Trips for which a fully corrected set of predictions were made (i.e., four correct predictions) rise to 80%. This kind of accurate prediction will improve CPUE estimates, and will contribute to understand the complex relationship between the fleet and the resources, which is potentially mediated by either local abundance of the resources or market preferences.

## INTRODUCTION

Mediterranean trawl fishery is multi-specific and exhibit spatial and temporal variability in fishing strategies, mainly depending on the bathymetric range, which determine both target species and communities exploited (e.g. {Alemany, 2003 1622 /id;Goñi, 2004 1363 /id;Massutí, 2005 1321 /id;Moranta, 2000 718 /id}. However, these different tactics are not considered for the management of trawl fishery, which reduce its efficiency to regulate the effective fishing mortality exerted on the specific stocks. In fact, studies aiming to assess fleet dynamics and to identify fishing tactics were developed during the last two decades (e.g. Murrawsky *et al.*, 1983; Rogers and Pikitch, 1992; Pelletier and Ferraris, 2000) and they have not been applied until recent years in the western Mediterranean, both for trawl (Goñi *et al.*, 1999; García-Rodríguez, 2003) and small scale fisheries (García-Rodríguez *et al.*, 2006). These studies applied different multivariate analysis to landings data series from sale bills. Considering that fleets regularly operate daily (e.g. 12 hours at the sea per fishing trip for trawling), the number of fishing days was considered as basic effort.

Trawlers can operate at distinct depth strata during the same fishing day depending on economic factors (e.g. landings price), weather conditions and the season. Therefore, target species and communities exploited may change over short time-scales. This fact increases complexity when predicting fishing tactics from daily sale bills data, although according to expert's opinion this data source still reflects the cases in which more than one fishing tactic has been developed during a single fishing trip. Despite the daily sale bills could be considered too noisy to be used as a reliable image of catch per unit of effort (CPUE), they are the most complete and extensive source of information available because they cover the entire fleet during all the year and often throughout long periods.

The diagnosis (prediction of the fishing tactic, or tactics, used) of thousands of daily sale bills for an expert is subjective and time-consuming. Accordingly, some approaches have been developed in order to increase precision (i.e., objectivity) and to make this task abordable. The most relevant approach is found in multivariate analysis. It consists in pooling the sale bills corresponding to a predefined period (usually by month), which implies to assume negligible within-month variability in the fishing tactics of a specific vessel. The second step consists in filtering these pooled data by using a Principal Components analysis (PCA; or other techniques for reducing dimensionality) and retaining the first few axes , which implies to stablish a cut-off criterion, as those already proposed by {Rohlf 1993 1009 /id} and {Legendre & Legendre 1998 1044 /id}, in order to select how many principal components are retained. . Finally, the monthly-pooled daily sale bills are clustered, using the scores of the selected first few PCA axes, and the major clusters are identified on the resulting dendrogram. However, this step implies a subjective criterion, because the amount of dissimilarity needed to define the cluster is usually data-driven, being problematic to extrapolate them to other data sets.

Here we propose a neural network (ANN) approach for the analysis of daily sale bills, capable to predict those mixing of fishing tactics, which is widely extended in the trawl fleet of the Balearic Islands (and probably in other areas of the Mediterranean). ANN is a data-driven method, in the sense that no functional relationship is assumed to exist between the catch composition (*input layer*) and the fishing tactics (*output layer*). Conversely, a complex net of simple operators (called *neurons*) are interconnected. The weight given at the different connections are optimized for obtaining the currently observed data in the *output layer* from the data in the input layer. To validate the results obtained from this method, information from sampling on board commercial trawlers are also used.

METHODS

1) Description and justification of the *a priori* defined fishing tactics and justification of the "species" included and excluded in the analysis

Previous studies, developed from sampling on board trawlers under commercial conditions, have detected four different fishing strategies in the bottom trawl fishery developed off the Balearic Islands, exploiting different bathymetric ranges (Guijarro and Massutí, 2006; Ordines *et al.*, submitted):

a. On the shallow shelf, between 50 and 80 m depth, in which the most important species are *Spicara smaris*, *Loligo vulgaris*, *Trachurus mediterraneus*, *Mullus surmuletus*, *Octopus vulgaris* and *Pagellus acarne* and *Scorpaena notata*, which are commercialised within a mixed fishes category.

b. On the deep shelf, between 150 and 200 m depth, in which the most important species are *Zeus faber*, *M. surmuletus*, *Merluccius merluccius*, *Lophius budegassa* and *Chelidonichthys cuculus* and *Helicolenus dactylopterus*, which are sold as mixed fishes

c. On the upper slope, between 350 and 600 m depth, in which *Nephrops norvegicus* is the main target species*,* also yielding a valuable bycatch of *Merluccius merluccius*, *Parapenaeus longirostris, Lepidorhombus boscii*, *Micromesistius poutassou*, *Helicolenus dactylopterus* and *Lophius* spp

d. On the middle slope, between 600 and 750 m depth, in which *Aristeus antennatus* is the main target species, also yielding a bycatch of *Galeus melastomus, Phycis blennoides*, *Plesionika martia* and *Geryon longipes*.

All these species characterising the four fishing strategies are also important in landings of the trawl fishery, both in terms of weight and/or economical profit. These fishing strategies, as well as their characteristic bathymetric range and main species, have also been detected by other studies from landings analysis (Alemany and Álvarez, 2003) and bottom trawl surveys (Massutí and Reñones, 2005).

The initial matrix was formed by 81 commercial categories ("species"), mostly corresponding to the above mentioned species,and 224 daily sale bills. These sale bills corresponded to those fishing trips with available information from on-board observers. Species with a frequency of appearance on the sales bills lower than 5% were excluded from the analysis to avoid a lot of zero values in the matrix.

A multivariate analysis of variance (MANOVA) was applied to the sale bills data to test for seasonality and between year effects on species composition (abundance).

2) Description and filtering of the raw data.

Sale bills mainly depend on the fishing tactic adopted. However, there are a number of factors that also affect the catches finally landed. For example, infrequent events might change the characteristics of the daily sale bill. One case is that large catches of rare species might cause that species otherwise considered as by-catch, shift to be considered as valuable catch. This haul-specific variability implies that the raw data might be not suitable as immediate input for classification systems. Therefore, some filtering step is necessary. Here we adopted a filtering strategy that consisted in applying a dimension reduction analysis and retaining the first few resulting axes. A sale bill is fully defined in a space when the number of dimensions taken in account equals the number of species. The position in the original space of a specific sale bill is defined by the abundance of all the species. The techniques for dimension reduction generate a new space with the same dimensionality than the original one, but the new set of axes are . ordered according to the variance explained. Hence,, the first few axes retain the largest amount of information and usually correspond to the main gradients of variability. Principal Components analysis (PCA ) and Correspondence analysis (CA ) were the dimension reduction analyses chosen. Both were applied to abundance data (kg), while CA was also applied to presence/absence data. Because of the distances preserved by each of the methods, Euclidean distances and chi-squared distances for PCA and CA respectively, the resulting axes ordinate sale bills focusing on absolute abundances (PCA) or on (approximated) relative abundances (CA) of the original data ({ter Braak & Smilauer 2002 1126 /id}, {Legendre & Legendre 1998 1044 /id}).

Data description was completed testing for the existence of differences attributable to year or month (using Canonical correspondence analyses; {Legendre & Legendre 1998 1044 /id}).

3) Predictive system (Artificial neural network).

Conventional classification systems assign each sales bill to one (and only one) category. This type of approach (e.g., the conventional discriminant analysis) is unsuitable here because the same vessel can use two or more fishing tactics during the same trip. Here we propose a generalized regression neural network (GRNN) in order to obtain a multiple prediction (i.e., a prediction for each of the four *a priori* defined fishing tactics). The architecture of a GRNN is defined by an *input layer*, a *hidden layer* and an *output layer* (Fig. 1). The *input layer* consisted in a matrix of q descriptors of the sale bill x p daily sale bills. The foregoing descriptors are the first few axes resulting from filtering the raw data (PCA and CA). The *hidden layer* is defined by p *neurons* that pass the input descriptors through a radial basis function (Fig. 2). The output layer is constituted by four linear neurons (*purelin neurons*), corresponding to the four fishing tactics. The more close to 1 the value for an output neuron is, the larger is the probability that the corresponding fishing tactic have been used.

Any model is able to predict exactly a reponse variable(s) from an explanatory variable(s) simply by increasing the model complexity (i.e., overparameterization). The problem of overparameterized models is that they lose generality (i.e., predictions of the model fail for data different from the ones used for model implementation). This problem is avoided here using a subset of the data to parameterize the ANN (*training data*), and testing the model prediction on a different subset (*testing data*). To minimize the possibility that the estimation of success was biased due to the specific data used as training or testing set, 500 consecutive runs were completed randomly reassigning the cases to one or other set.

CA and PCA were completed using CANOCO and R package. ANN was implemented using the ANN facilities in MATLAB. Multiple runs were completed in MATLAB also.

## RESULTS

### *Data description*

MANOVA analyses showed that species composition (abundance) derived from sales bills showed significant seasonality and between-year differences. However, the amount of variance explained by these variables was small. Presence/absence data did not show any significant effect.

Dimensionality reduction techniques show that daily sales bills are homogenously distributed, suggesting that no outlier points were affecting the analyses (Fig. 3). It is noteworthy the large overlap existing between the samples corresponding to different fishing tactics (Fig. 3). The overlapping between fishing tactics was reflected in the noisy picture depicted by the dendrogram resulting from the first six CA axes (Fig. 4). Therefore, this method seems to be unreliable to identify groups of samples in order to relate them with fishing tactics..

### *ANN capability.*

The values obtained in the four neurons of the output layer show a distribution strongly skewed towards either one or zero (Fig. 5). This fact simplified the selection of a threshold for assuming that a specific fishing tactics have been probably used. Here we select 0.5 as threshold value.

Extensive trials were completed using PCA and CA scores and the number of axes retained was optimized in each case. The best results were obtained when the input data used in the ANN consisted in the scores of the first six axes resulting from applying CA to the species abundance matrix (kg). Alternatively, the CA was applied to the species presence/absence matrix and the use of the resulting scores yielded an error rate of nearly 8%. Similarly, using the PCA scores the percentage of correct predictions was slightly lower. Therefore, in order to simplify it, hereafter we only reported the results corresponding to CA scores resulting from to species abundance matrix.

The predictive power of the ANN must be considered as outstanding. The global percentage of successful predictions rises up to 92-94%. Note that an individual success corresponds to a correct prediction for each fishing tactics. Therefore, a sample showing all four fishing tactics correctly predicted were computed as four correct predictions. Concerning the percentage of samples for which the use of all four fishing tactics were correctly predicted, it rises up 80%.

The analysis of the erroneous predictions indicates the possible biases and the behaviour of the ANN. The inspection of the confusion matrix (Table 1) indicates that errors mainly concentrate in consecutive depth strata. Confirming this result, the percentage of correct assignments arrived up to 98% when only two fishing tactics were considered. A second pattern observed is that mixed samples tend to accumulate more errors (in relative terms) than the samples corresponding to a single fishing tactics. Finally, the absence of observed bills of some combination of fishing tactics should deteriorate the power of the method.

## DISCUSSION

The results confirmed that data-driven methods are a reliable alternative to the multivariate approach described above. It is remarkable that the power of the method was here tested for first time with an independent data set. The specific case of ANN used here (GRNN) shows some interesting advantages. First, GRNN makes independent predictions for each tactics, and therefore allows mixed predictions. Second, the algorithm is very fast, and easily to implement. Third, predictions for large data set are possible. This permit to obtain reliable predictions of the most probable fishing tactics for the thousands of sales bills commonly availables.

The success of ANN and other data-drive methods rely on the type of input provided. Consequently, extensive trials have been done, and the best input matrix was build up by the scores on the first six axes extracted by a correspondence analysis on the raw abundance matrix. The existence of a significant effect of seasonality (month) and between-year differences suggest that the scores resulting after removing these effects should give more accurate results. Contrasting, the error rate is slightly larger (7-8%) in comparison with the error corresponding to a crude CA (6-7%). We suggest that both, the small amount of variance explained by seasonality (10%) and between-year differences (3%), and the fact that the selection of the fishing tactics changes also with seasonality smooth the a priori advantages of removing the effects of seasonality and between-year differences.

The approach proposed here will improve CPUE estimates, and will contribute to understand the complex relationships between the fleet and the resources, which is potentially mediated by local abundance of the resources or market preferences.

REFERENCES

Alemany, F., Alvarez, F., 2003. Determination of effective fishing effort on hake Merluccius merluccius in a Mediterranean trawl fishery. Scientia Marina 67, 491-499

García-Rodriguez, M., 2003. Characterisation and standardisation of a red shrimp, *Aristeus antennatus* (Risso, 1816), fishery off the Alicante gulf (SE Spain). Scientia Marina 67, 63-74

García-Rodriguez, M., Fernandez, A.M., Esteban, A., 2006. Characterisation, analysis and catch rates of the small-scale fisheries of the Alicante Gulf (SE Spain) over a 10 years time series. Fisheries Research 77, 226-238

Goñi, R., Adlerstein, S., Álvarez, F., Garcia, M., Sanchez, P., Sbrana, M., Maynou, F., Viva, C., 2004. Recruitment indices of European hake, *Merluccius merluccius* (Linnaeus 1758), in the Northwest Mediterranean based on landings from bottom-trawl multispecies fisheries. ICES Journal of Marine Science 61, 760-773

Goñi, R., Álvarez, F., Adlerstein, S., 1999. Application of generalized linear modeling to catch rate analysis of Western Mediterranean fisheries: the Castellón trawl fleet as a case study. Fisheries Research 42, 291-302

Guijarro, B., Massutí, E., 2006. Selectivity of diamond- and square-mesh codends in the deepwater crustacean trawl fishery off the Balearic Islands (western Mediterranean). ICES Journal of Marine Science 63, 52-67

Hidalgo, J.M., Oliver, P., Massutí, E., Guijarro, B., Morsanta, J., Cartes, J., Lloret, J., Morales-Nin, B. (submitted). Seasonal and short spatial patterns in European hake recruitment process at the Balearic Sea (NW Mediterranean): the role of the environment on distribution and condition. Journal of Marine Systems.

Legendre, P., Legendre, L., 1998. Numerical Ecology. Elsevier Science B.V., Amsterdam

Massutí, E., Renones, O., 2005. Demersal resource assemblages in the trawl fishing grounds off the Balearic Islands (western Mediterranean). Scientia Marina 69, 167-181

Moranta, J., Massutí, E., Morales-Nin, B., 2000. Fish catch composition of the deep-sea decapod crustacean fisheries in the Balearic Islands (western Mediterranean). Fisheries Research 45, 253-264

Murawski, S.A., Lange, A.M., Sissenwine, M.P., Mayo, R.K., 1983. Definition and analysis of multispecies otter-trawl fisheries off the northeast coast of the United States. Journal du Conseil 41, 13-27

Ordines, F., Massutí, E., Guijarro, B., Mas, R., (submitted). The effect of mesh geometry on the selectivity of a multi-species bottom trwal fishery on the continental shelf off the Balearic Islands (western Mediterranean): diamond *vs.* square mesh in the codend. Aquatic Living Resources

Pelletier, D., Ferraris, J., 2000. A multivariate approach for defining fishing tactics from commercial catch and effort data. Canadian Journal of Fisheries and Aquatic Sciences 57, 51-65

Rogers, J.B., Pikitch, E.K., 1992. Numerical definition of groundfish assemblages caught off the coasts of Oregon and Washington using commercial fishing strategies. Canadian Journal of Fisheries and Aquatic Sciences 49, 2648-2656 ter Braak C.J.F., Smilauer P., 2002. CANOCO reference manual and user's guide to Canoco for Windows: Software for Canonical Community ordination (version 4.5). Microcomputer Power, Ithaca, NY, USA

Table 1: Confusion matrix (observed versus predicted classes) corresponding to 500 runs (the testing data set was randomly selected in each run). The number of cases are indicated in the upper panel, and the percentages (by columns) in the lower panel.

| | Obs. | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Predicted | SS | DS | US | MS | SS+DS | SS+US | SS+MS | DS+US | DS+MS | US+MS |
| SS | 1064 | | | | 315 | | | | 1 | |
| DS | | 755 | | | 193 | | | 138 | 3 | |
| US | | | 361 | 47 | | 108 | | 46 | | |
| MS | | | 10 | 4351 | | | 2 | | | 54 |
| SS+DS | 362 | 287 | | | 478 | 60 | 9 | | 64 | |
| SS+US | | | 113 | | 44 | 977 | | 16 | | |
| SS+MS | | | | | 58 | | 793 | | 120 | |
| DS+US | | 54 | 15 | | | 9 | | 161 | | |
| DS+MS | | | | | | | 100 | | 158 | |
| US+MS | | | | 52 | | | | | | 114 |
| TOTAL | 1426 | 1096 | 499 | 4450 | 1088 | 1154 | 904 | 361 | 346 | 168 |

| | Obs. | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Predicted | SS | DS | US | MS | SS+DS | SS+US | SS+MS | DS+US | SS+MS | US+MS |
| SS | 74.6 | | | | 29 | | | | 0.3 | |
| DS | | 68.9 | | | 17.7 | | | 38.2 | 0.9 | |
| US | | | 72.3 | 1.1 | | 9.4 | | 12.7 | | |
| MS | | | 2 | 97.8 | | | 0.2 | | | 32.1 |
| SS+DS | 25.4 | 26.2 | | | 43.9 | 5.2 | 1 | | 18.5 | |
| SS+US | | | 22.6 | | 4 | 84.7 | | 4.4 | | |
| SS+MS | | | | | 5.3 | | 87.7 | | 34.7 | |
| DS+US | | 4.9 | 3 | | | 0.8 | | 44.6 | | |
| DS+MS | | | | | | | 11.1 | | 45.7 | |
| US+MS | | | | 1.2 | | | | | | 67.9 |

FIGURE LEGENDS

Fig. 1: ANN architecture. The number of neurons in each layer is indicated. The function operating in each layer is indicated also (unimodal response in the intermediate layer and linear response in the output layer).

Fig. 2: An example of the output of a radial basis function. Note that the original descriptors of the species composition of one specific sale bill (x-axis) are rescaled between 0 an 1.

Fig. 3. Biplots corresponding to CA on “species” composition of 224 daily sales bills. The term “Species” refers here to the commercial categories considered. Scaling focuses on preserving (chi-squared) distances between samples. The top panel shows the polygons enclosing all samples belonging to each of the considered categories. Species composition of these categories can be approximated from the bottom panel. The distance between a sample point and species points approximates the predicted relative frequency of this species in the sample (for example, *Aristeus antennatus* and *Galeus melastiomus* tend to be related with TM). Moreover, species points in proximity correspond to species often occurring together. The species more closely related to the main gradients have been included only.

Fig. 4: Tree plot corresponding to a cluster analysis (UPGMA) on the Chi-squared distances within the space defined by the six first axes extracted by a CA analysis on the “species” abundances. It is not possible to objectively define clusters of samples assignable to specific fishing tactics.

Fig. 5: Histogram of the distribution of the values at the output layer. Almost all values are larger than 0.9 or smaller than 0.1.

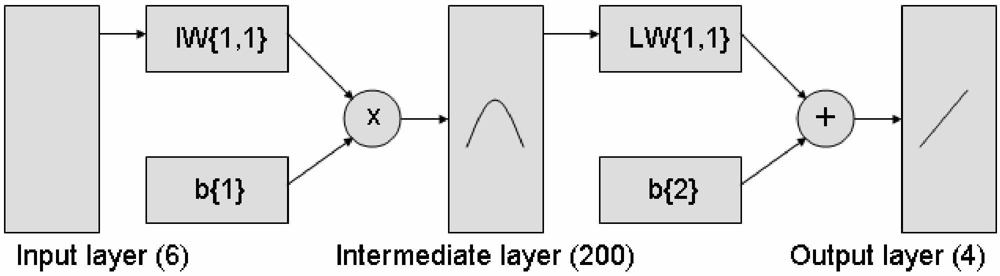


Figure 1. Moranta et al.

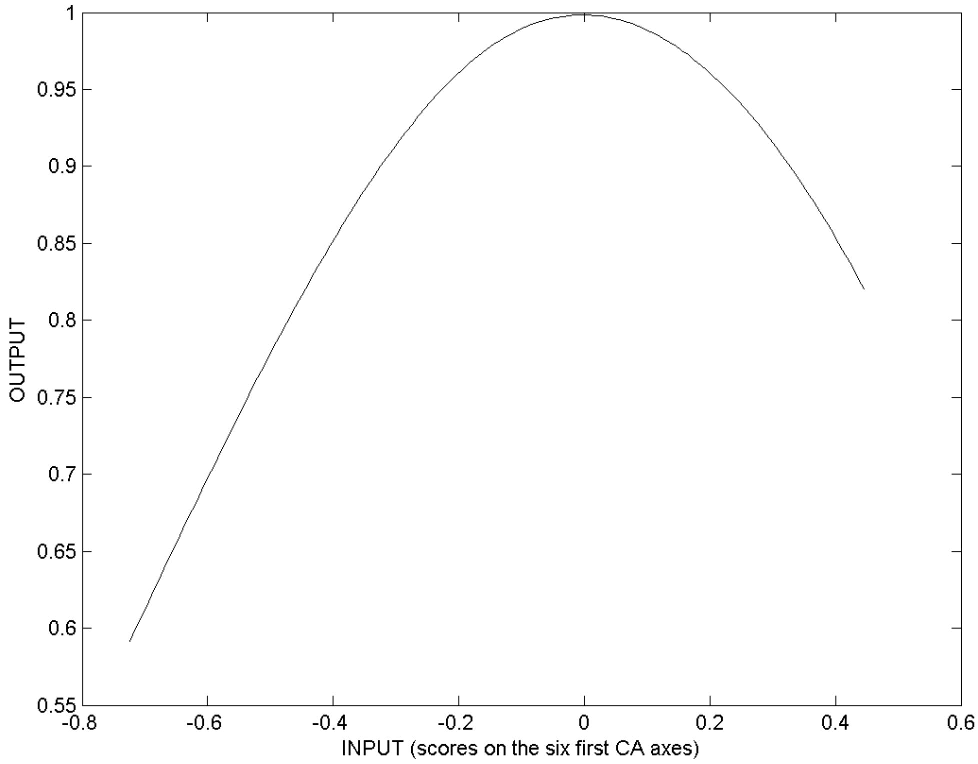


Figure 2. Moranta et al.

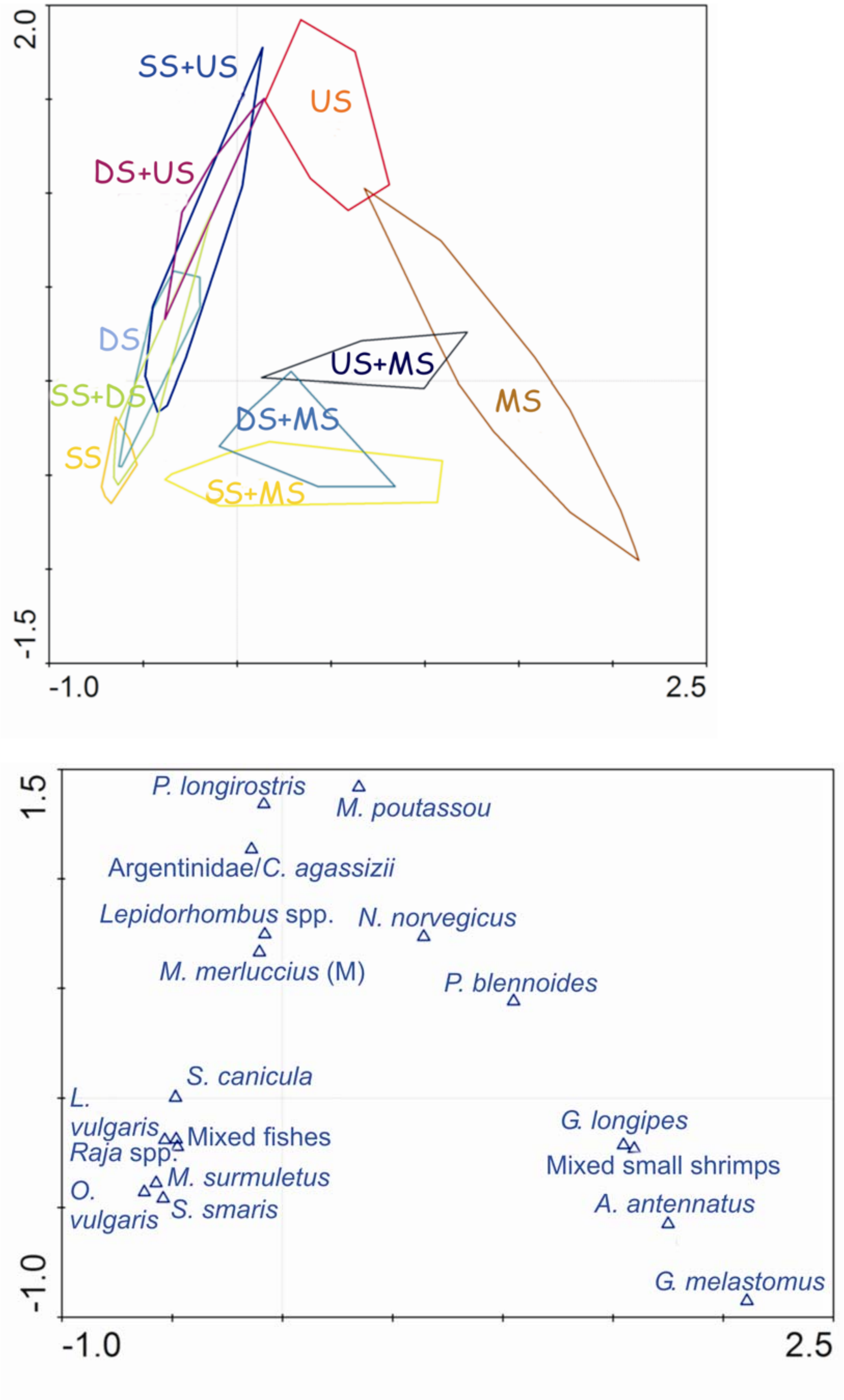


Figure 3. Moranta el al.

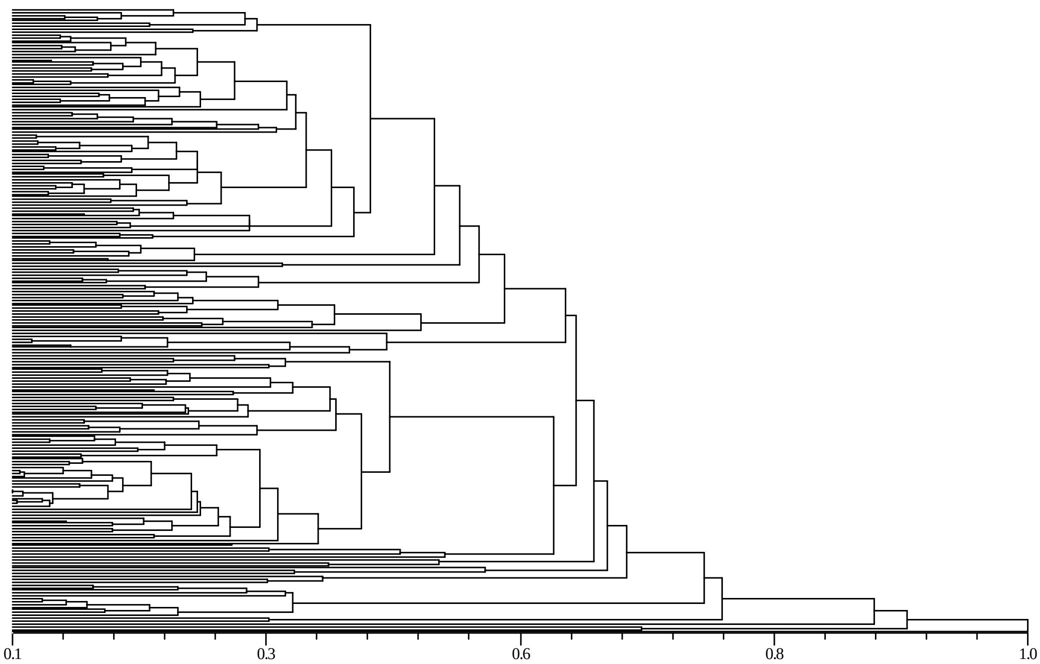


Figure 4. Motranta et al.

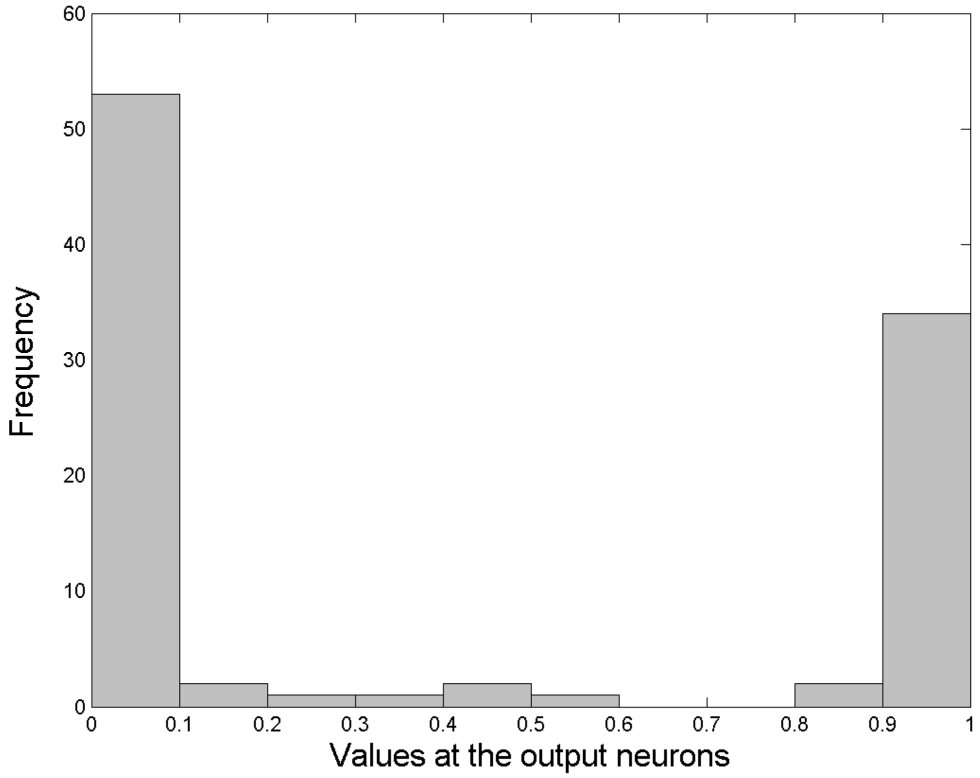


Figure 5. Moranta el al