Data analysis advances in marine science for fisheries management: Supervised classification applications

Jose A. Fernandes

AZTI-Tecnalia
Intelligent Systems Group
The University of the Basque Country

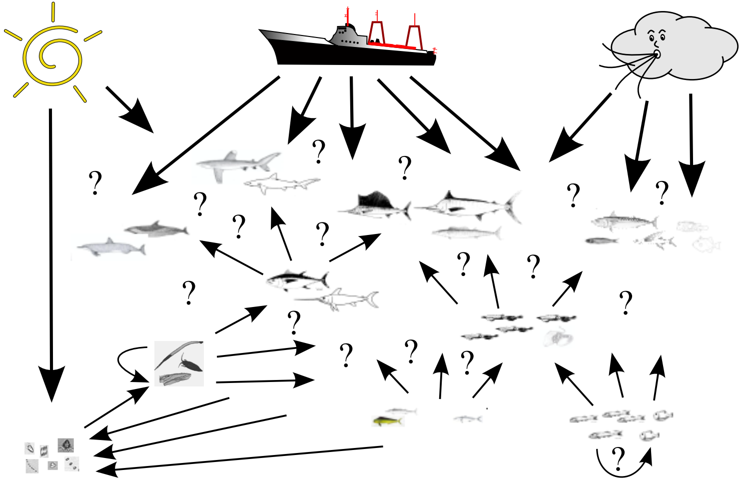
Donostia 6^{th} of May, 2011

Outline

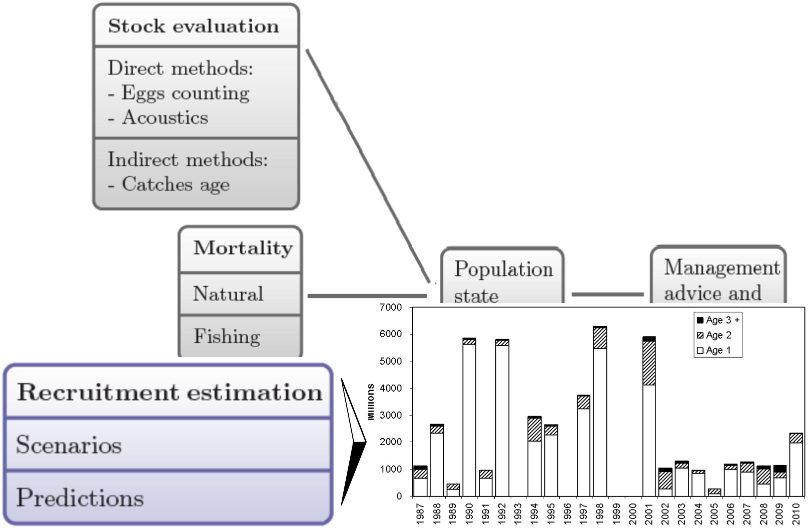
Outline

Knowledge and advice flow in fisheries management

High uncertainty in fisheries research



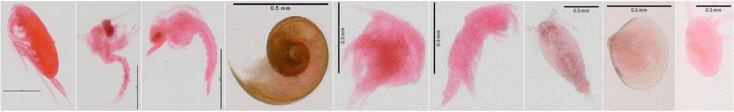
Fish stock estimation for management advice



Data domains in fisheries research

This thesis focuses in:

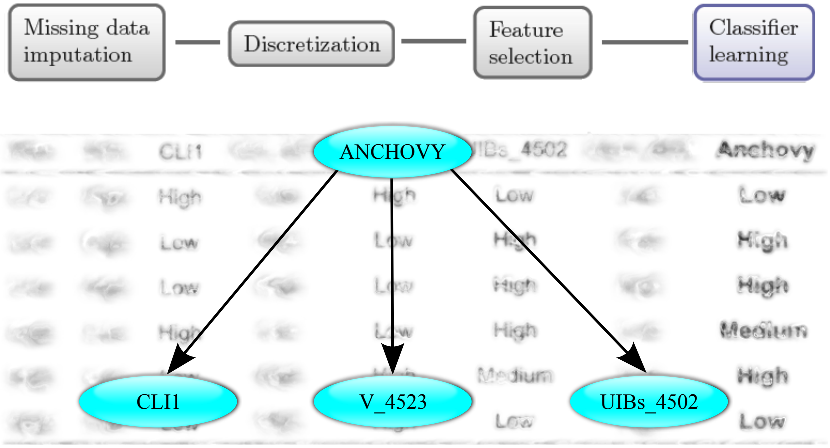
- Samples processing: Zooplankton semi-automatic classification.



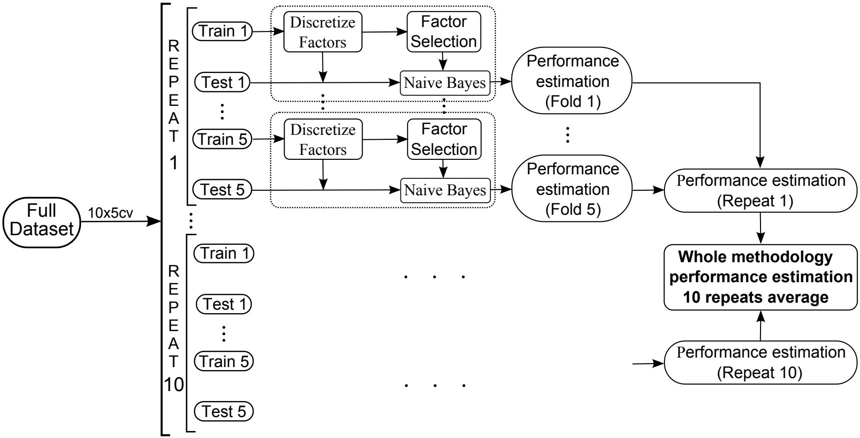
- Robust forecasting: Fish recruitment forecasting.

- Ecosystem-based approach: Simultaneous recruitment forecasting of multiple species.

Supervised classification and the data analysis process



Pipeline validation in filter methods



Performance measures

		Predicted class yes	Predicted class no
Actual class	yes	true positive (TP)	false negative (FN)
	no	false positive (FP)	true negative (TN)

- $Accuracy = \frac{TP+TN}{\#cases}$
- $True\ Positive\ Rate = \frac{TP}{TP+FN}$
- The higher the best for both.

Brier Score

- *Brier Score* $= \frac{1}{\#cases} \sum_{k=1}^{\#cases} \sum_{l=1}^{\#classes} (p_l^k - y_l^k)^2$
- The lower the best (contrary to accuracy & true positive)
- Between 0 & 2, divide by 2 for easier comprehensibility

	$y_l^K = 1$ Actual	$y_l^K = 0$ Otherwise		
	High	Medium	Low	
p^1	0.7	0.2	0.1	$(0.7-1)^2 + (0.2-0)^2 + (0.1-0)^2 = 0.14$
p^2	0.8	0.1	0.1	$(0.8-1)^2 + (0.1-0)^2 + (0.1-0)^2 = 0.06$
p^3	0.1	0.5	0.4	$(0.1-1)^2 + (0.5-0)^2 + (0.4-0)^2 = 1.22$
p^4	0.4	0.5	0.1	$(0.4-1)^2 + (0.5-0)^2 + (0.1-0)^2 = 0.62$

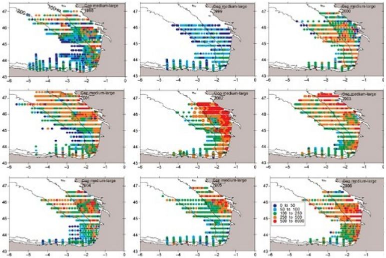
Brier Score: (0.14 + 0.06 +1.22 + 0.62) / 4 = 0.51

Normalized Brier Score: 0.51 / 2 = 0.255

Outline

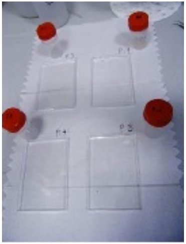
Outline

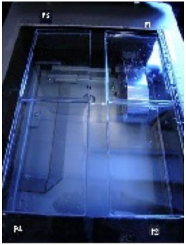
Problem definition

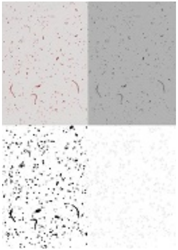


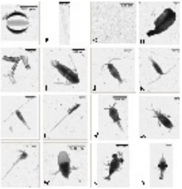
?

Semi-automatic classification

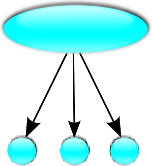






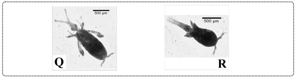


Minor	ECD	Class
0.63	0.79	Copepod
0.46	0.79	Decapoda
0.16	0.79	Artifact
1.39	0.79	Copepod

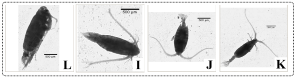


Need for trade-off between performance and taxa detail

Corycaeidae
Oncaeidae
Calanoida Lateral
Calanoida Dorsal I
Calanoida Dorsal II
Calanoida Dorsal III
Temoridae
Sapphirinidae
Marine Snow
Cladocera
Poicilo Lateral
Eucalanidae
Oithonidae
Miraciidae
Cirripeda
Appendicularia
Gastropoda
Chaetognatha
⋮



Cyclopoida



Calanoida

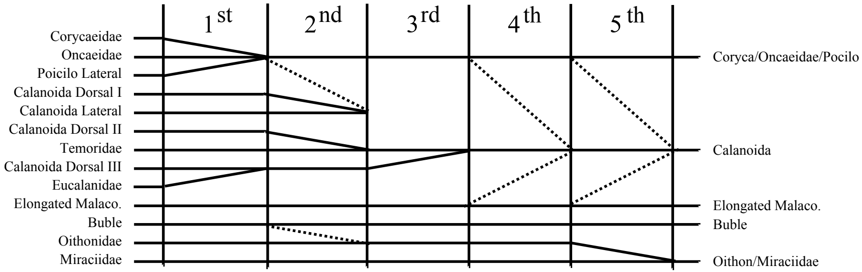
- Maximun taxa detail (# classes) is desired
- 75% times +1 class decreases performance
- Distribution reliability decreases with performance
- A class with a 0.2 true posive rate is useless
- Maximum performance is desired
- In general lower # classes = higher performance
- But, 25% times +1 class increases performance
- How to find good trade-off? Trial and error?
- Need of a tool to guide and help the expert!!!

The experimentation

Datasets	# indiv.	# classes	Avg. indiv. per class
Tulear 2004	1839	37	46
Bioman 98-06	17803	24	1232
Bioman 2007	6694	30	632

Method for trade-off between performance and taxa

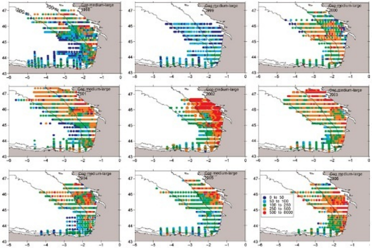
- Expert specifies the most detailed training-set possible.
- The performance of the training-set is evaluated.
- Class mergers that improve the performance are proposed.
- Expert selects those that have biological meaning.

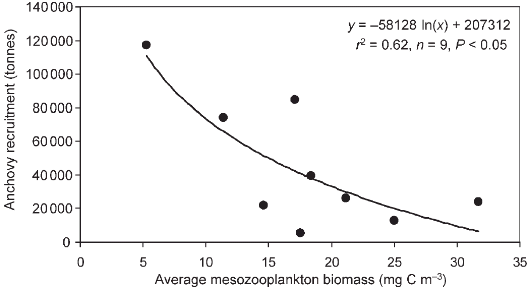


Method results

	Tulear 2004	Bioman 98-06	Bioman 2007
Initial # classes	37	24	30
Final # classes	25	19	26
Initial accuracy (%)	64.7	85.7	82
Final accuracy (%)	74	88.8	82.1
Initial TP # < 0.5	10	9	12
Final TP # < 0.5	3	2	3

Zooplankton biomass, a limiting factor to recruitment?



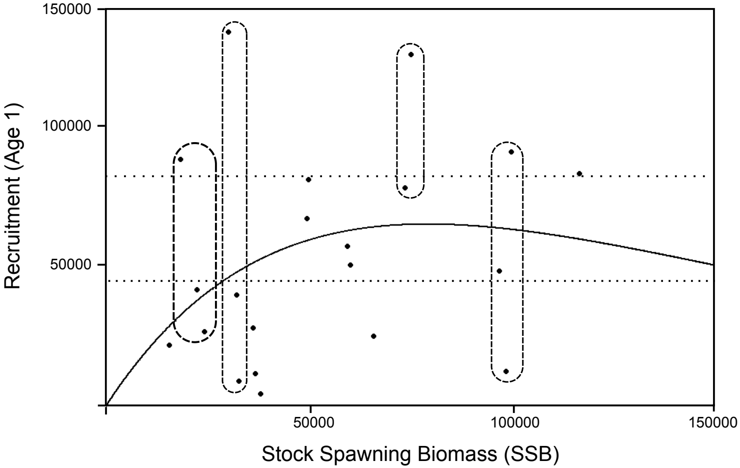


?

Outline

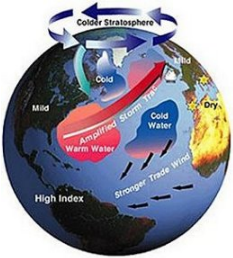
Problem definition

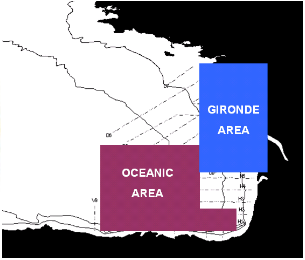
- Stock spawning biomass - recruitment relationship?

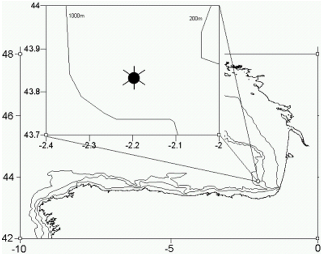


Relationship between recruitment and climate

- Identified relationships with global climatic patterns.
- Identified relationships with regional factors.
- Identified relationships with local factors.
- Previous attempts of forecasting based on climatic and environmental factors not very succesfull.

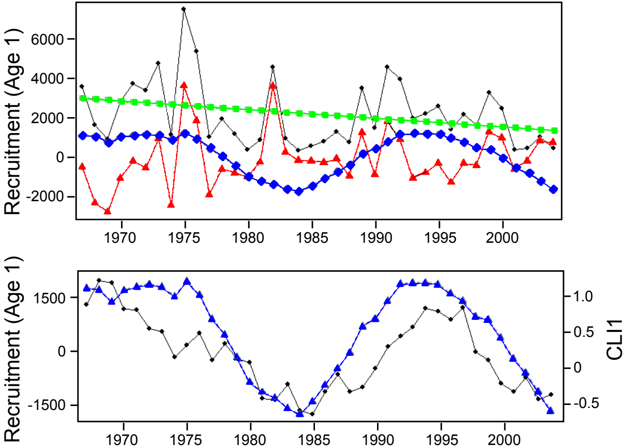




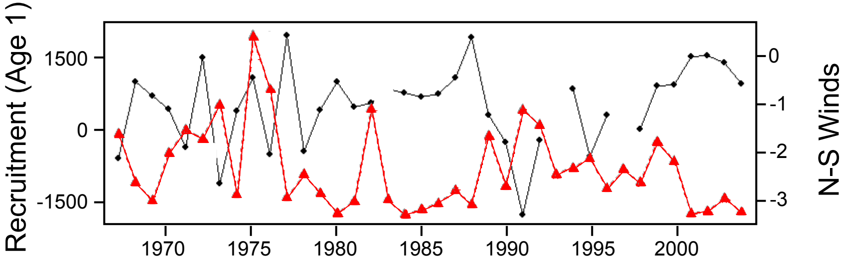


The time-series

- Cyclical behaviour of anchovy recruitment?
- Driven by cyclical behaviour of climate patterns?



Other factors



The data

- Workshop on Long-term Variability in SW Europe (2007).
- Data extended from other sources such as NOAA.
- From 100 to 200 factor candidates (columns).
- From 30 to 50 years of data (rows).
- Noisy data.
- Need of probabilistic forecast.
- Need of robust results.

Methodological pipeline

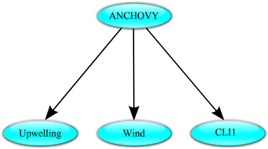
- A novel class discretization method balancing error.
- Discretization: Fayyad and Irani's MDL method.
- Multivariate feature selection LOO CFS.
- Probabilistic model: naive Bayes classifier.
- Honest model validation and comparison.

Recruitment levels:
- Low: < 1200
- Medium
- High: > 3250

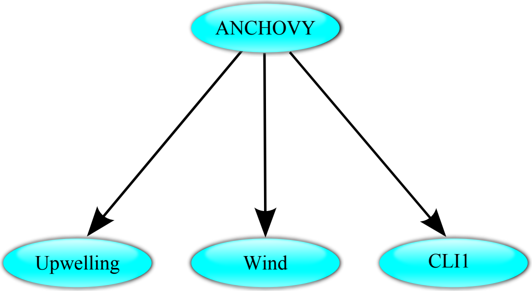
Factors:
- No level (SSB)
- Two levels
- Three levels

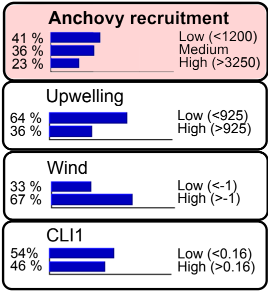
Factors from:
- 7 global
- 13 regional
- 19 local

To:
- CLI1
- Upwelling
- Winds

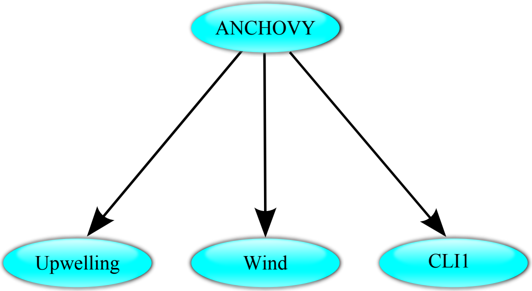


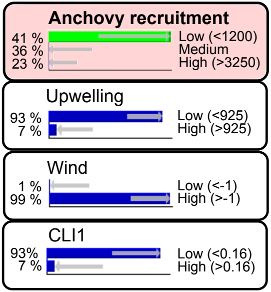
Anchovy final model



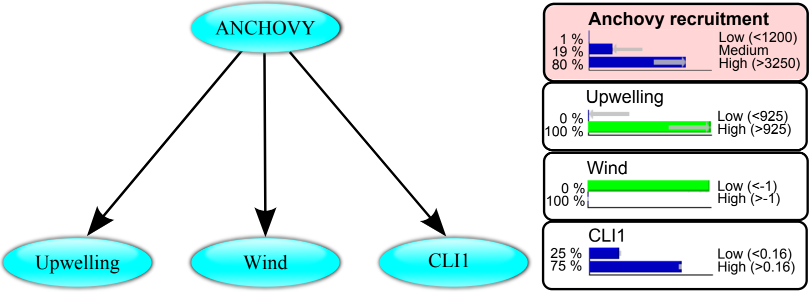


Anchovy final model





Anchovy final model

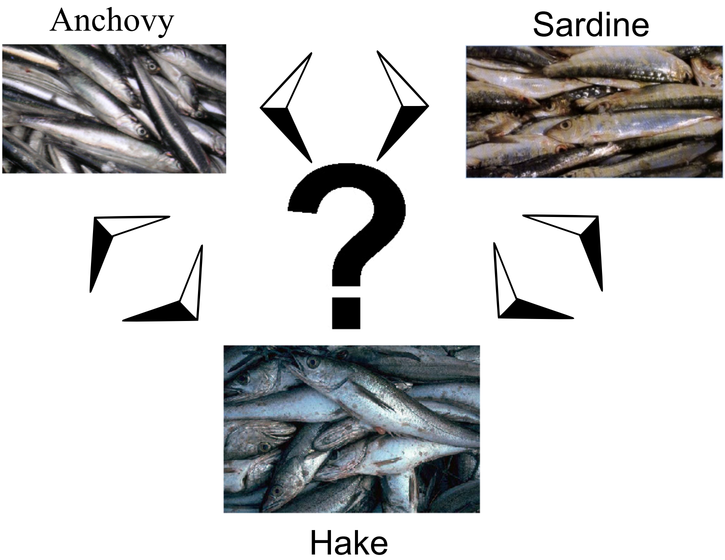


Classifiers comparison

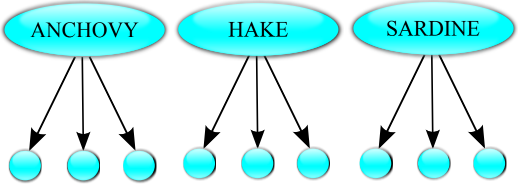
Metrics	**NB**	TAN	J48DT	MPNN	**SVM**
10 x 5cv Acc. (%)	$\mathbf{44.9 \pm 5.0}$	38.4 ± 9.1	46.3 ± 7.3	46.3 ± 7.7	$\mathbf{45.8 \pm 5.1}$
Brier score	$\mathbf{0.24 \pm 0.05}$	0.26 ± 0.06	0.27 ± 0.05	0.29 ± 0.05	$\mathbf{0.22 \pm 0.05}$
TP low	**0.473**	0.393	0.488	0.474	**0.454**
TP medium	**0.270**	0.276	0.313	0.29	**0.376**
TP high	**0.394**	0.323	0.348	0.356	**0.325**
CPU-time (min)	**29.0**	29.8	29.7	82.3	**33.4**

Outline

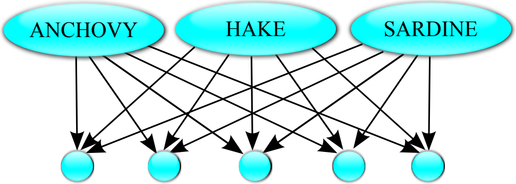
Problem definition



Multi-dimensional classification approach to fisheries

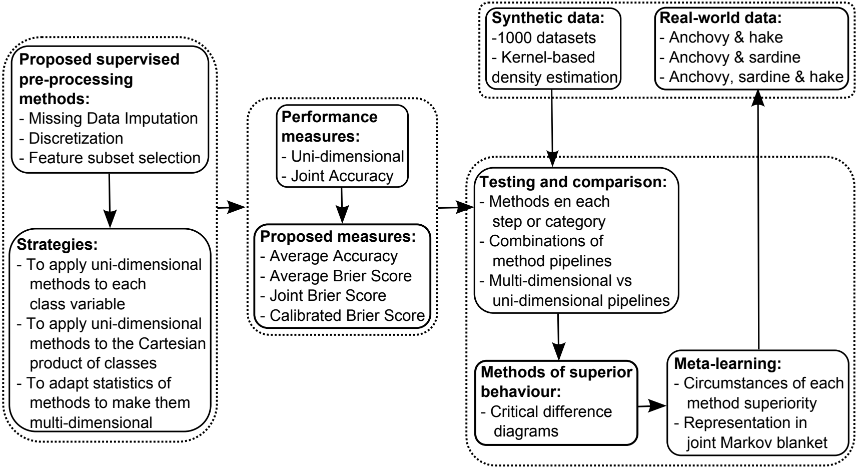


	X_1	...	X_n	C
Instance 1	x_1^1	...	x_n^1	c^1
Instance 2	x_1^2	...	x_n^2	c^2
Instance 3	x_1^3	...	x_n^3	c^3
...		...		
Instance N	x_1^N	...	x_n^N	c^N



	X_1	...	X_n	C_1	...	C_m
Instance 1	x_1^1	...	x_n^1	c_1^1	...	c_m^1
Instance 2	x_1^2	...	x_n^2	c_1^2	...	c_m^2
Instance 3	x_1^3	...	x_n^3	c_1^3	...	c_m^3
...		...			...	
Instance N	x_1^N	...	x_n^N	c_1^N	...	c_m^N

Experimental design



Multi-dimensional supervised pre-processing strategies

- Target each class variable **separately** and merge results.

$X_1, X_2, ..., Xn \rightarrow$ **Anchovy** $\rightarrow X_1, X_2$

$X_1, X_2, ..., Xn \rightarrow$ **Hake** $\rightarrow X_2, X_7$

$X_1, X_2, ..., Xn \rightarrow$ **Sardine** $\rightarrow X_1, X_9$

$\rightarrow X_1, X_2, X_7, X_9$

- Targeting the **Cartesian product** of classes.

Anchovy	**Hake**	**Sardine**		**A x H x S**
Low	Low	Medium	→	LLM
Low	Medium	Low	→	LML
Medium	High	Medium	→	MHM
Medium	Low	High	→	MLH
High	High	Low	→	HHL
High	Medium	High	→	HMH

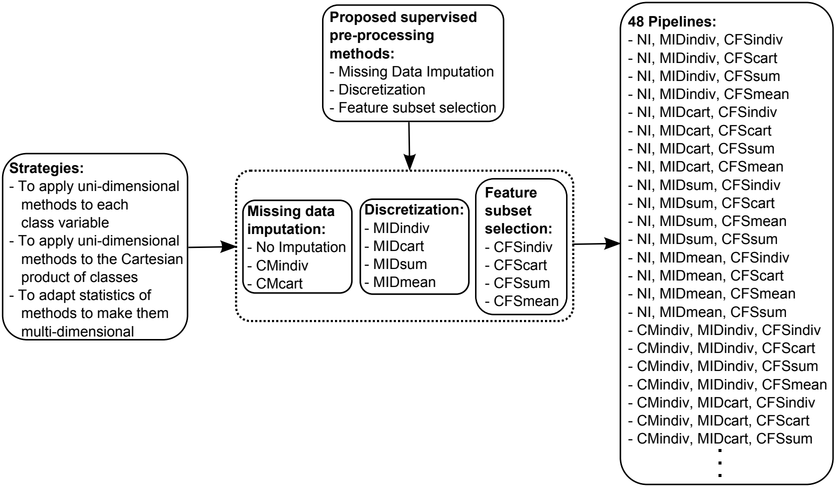
$X_1, X_2, ..., Xn \rightarrow$ **A x H x S**

$X_1, X_2, ..., Xn \rightarrow X_1, X_3, X_7, X_9$

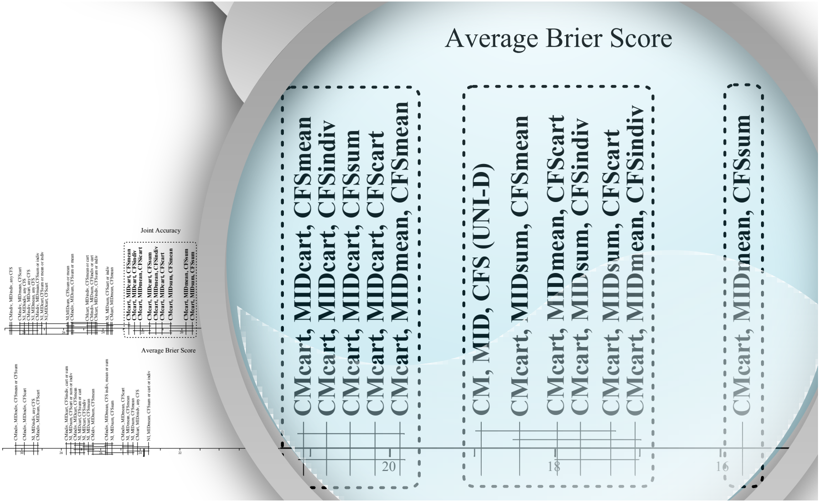
- **Adapt** the statistics of the methods based on *mean* or *sum*.

$X_1, X_2, ..., Xn \rightarrow$ **Anchovy, Hake, Sardine** $\rightarrow X_2, X_3, X_8, X_9$

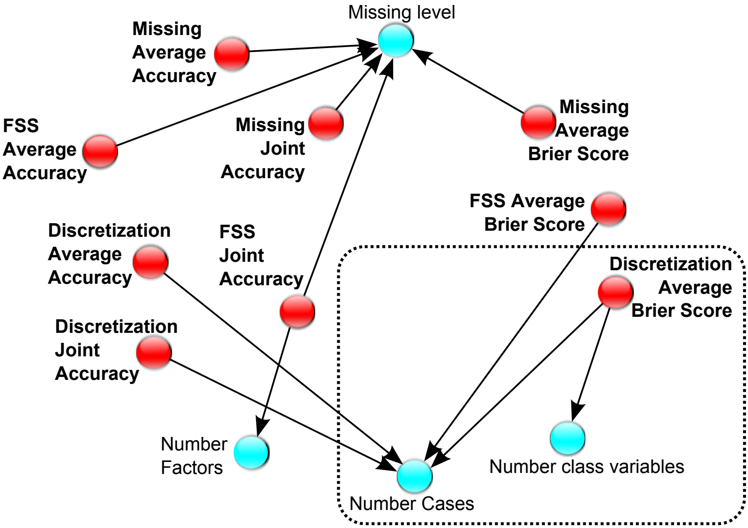
Pipelines of pre-processing methods for multi-dimensional classification



Pre-processing methods of superior behaviour



Metalearning: Circumstances of each method superiority



Simultaneous forecasting of 3 fish species recruitment

- Doubled the chance of being right in all species simultaneously.
- Superior Brier score for each species.
- The advantage of a single model suiting the ecosystem-based approach.

Pre-processing pipeline	Anchovy BS	Sardine BS	Hake BS	Joint Acc.
CM-MID-CFS (Uni-D)	0.36	0.34	0.27	17.3 ± 4.8
CMcart-MIDmean-CFSsum	0.35	0.27	0.21	28.9 ± 4.5
CMcart-MIDindiv-CFScart	0.32	**0.24**	0.19	22.6 ± 4.3
CMcart-MIDmean-CFSmean	0.32	0.25	**0.18**	19.7 ± 5.5
CMcart-MIDmean-CFScart	**0.30**	0.27	0.21	$\mathbf{29.5 \pm 4}$
CMcart-MIDmean-CFSindiv	0.32	0.27	**0.18**	28.5 ± 4.7

Outline

Optimizing number of classes in Zooplankton classification

Conclusions

- A method for experts to define and evaluate training-sets.
- 9 years of data with more than 4,000 samples processed.
- No relation found between anchovy recruitment and zooplankton biomass.

Future work

- To consider the non-random spatial distribution of plankton in samples.
- To apply novel approaches such as semi-supervised classification.
- To design a system that can be used at sea (real-time).

Robust machine learning methods for fish recruitment forecasting

Conclusions

- A methodology for fish recruitment forecasting based on state-of-art machine learning has been proposed.
- The method has been used on real-world advice last 2 years.

Future work

- To extend the methodology to deal with continuous variables.
- To develop long-term forecasting mixing with mechanistic models.
- To improve the analysis of factors stability (help to detect mechanisms).
- To incorporate cost-sensitive modelling.

Pre-processing for multi-dimensional fish recruitment forecasting

Conclusions

- A set of pre-processing methods has been proposed.
- Tested with synthetic and real data.
- Methods of superior behaviour and circumstances of each method superiority identified.
- Suitability of multi-dimensional classification for ecosystem based approach.

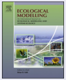
Future work

- To explore different model structures (from probabilistic to mechanistic relationships).
- To propose methods with continuous data.
- To apply to other data domains of multi-dimensional nature.

Thesis publications

International Journals: first author

J.A. Fernandes, X. Irigoien, G. Boyra, J.A. Lozano, I. Inza (2009) Optimizing the number of classes in automated zooplankton classification. *Journal of Plankton Research*, 31(1): 19-29.



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

J.A. Fernandes, J.A. Lozano, I. Inza, X. Irigoien, J.D. Rodríguez, A. Pérez, A. (2011) Supervised pre-processing approaches in multiple class-variables classification for fish recruitment forecasting. *Applied Soft Computing*, submitted.

International Journals: collaborations

L. Zarauz, X. Irigoien, J.A. Fernandes (2008) Modelling the influence of abiotic and biotic factors on plankton distribution in the Bay of Biscay, during three consecutive years (2004-06). *Journal of Plankton Research*, 30(8): 857-872.

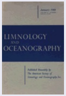
L. Zarauz, X. Irigoien, J.A. Fernandes (2009) Changes in plankton size structure and composition, during the generation of a phytoplankton bloom, in the central Cantabrian sea. *Journal of Plankton Research*, 31(2): 193-207.

X. Irigoien, J.A. Fernandes, P. Grosjean, K. Denis, A. Albaina, M. Santos (2009) Spring zooplankton distribution in the Bay of Biscay from 1998 to 2006 in relation with anchovy recruitment. *Journal of Plankton Research*, 31(1): 1-17. Featured article.

International Journals: collaborations

X. Irigoien, G. Chust, J.A. Fernandes, A. Albaina, L. Zarauz (2011) Factors determining the distribution and betadiversity of mesozooplankton species in shelf and coastal waters of the Bay of Biscay. *Journal of Plankton Research*, in press.

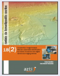
E. Andonegi, J.A. Fernandes, I. Quincoces, A. Uriarte, A. Pérez, D. Howell and G. Stefansson (2011) Improving semi-automated zooplankton classification using an internal control and different imaging devices. *ICES Journal of Marine Science*, in press.



E. Bachiller, J.A. Fernandes, X. Irigoien (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. *Limnology and Oceanography: Methods*, submitted.

Other publications

Technical reports



E. Bachiller, J.A. Fernandes (2011) Zooplankton Image Analysis Manual: automated identification by means of scanner and digital camera. *Revista Investigación Marina*, 18(2): 16-37.

L. Ibaibarriaga, A. Uriarte, S. Sanchez, J.A. Fernandes, X. Irigoien (2010) Use of juvenile abundance indices for the management of the Bay of Biscay. *Working document ICES WGANSA*, Lisbon, Portugal.

E. Andonegi, I. Quincoces, H. Murua, J.A. Fernandes, A. Uriarte, S. Sanchez, S. Cerviño, et al. (2010) Fish stock recovery strategies - Report from the Bay of Biscay. *UNCOVER*.

J.A. Fernandes, X. Irigoien, A. Uriarte, J.A. Lozano, I. Inza (2009) Anchovy Recruitment Mixed Long Series prediction using supervised classification. *Working document ICES WKSHORT*, Bergen, Norway.

Data analysis advances in marine science for fisheries management: Supervised classification applications

Jose A. Fernandes

AZTI-Tecnalia
Intelligent Systems Group
The University of the Basque Country

Donostia 6^{th} of May, 2011