UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Dottorato di Ricerca in Ingegneria dell'Informazione

Doctoral Course

"Big Data Analytics: Marine Data as a Case Study"

Dr. Gianpaolo Coro

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI) - CNR- Italy

gianpaolo.coro@cnr.it

Short Abstract:

In this course, practical methodologies for **marine data analysis** and **modelling** will be presented.

The course will cover specific classes of problems in marine science and their corresponding solutions, adopting state-of-the-art computer science technologies and methodologies. The explained techniques will include:

1) Unsupervised approaches to discover patterns of habitat change and predict fishing vessel activity patterns: Principal Component Analysis and Maximum Entropy for feature selection; KMeans, XMeans, DBScan, and Local Outlier Factor cluster analysis; Singular Spectrum Analysis for time series forecasting;

2) Supervised approaches for species distribution prediction and invasive species monitoring: Feed-Forward Artificial Neural Networks, Support Vector Machines, AquaMaps, Maximum Entropy;

3) Bayesian models to predict fish stock availability in specific fishing areas;

These methods will be applied to marine data such as vessel transmitted data, species observation records, and catch and vessel time series that fall into the Big Data category. These data are crucial to safeguard food availability and economic welfare, which are fundamental to human life. For example, predicting the impact of climate change on species habitat distribution contributes to avoiding economic and biodiversity collapse due to sudden ecosystem change. Likewise, monitoring the effect of overfishing on fish stocks and marine biodiversity prevents ecosystem and economic collapse.

The explained techniques will address real use cases of the United Nations (FAO, UNESCO, UNEP, and others) for marine food and ecosystem safety and illustrate the new lines of research in this context. They are also general enough to be applied to Big Data of other domains. The analysed data have indeed general characteristics of Big Data such as constantly incrementing volume, vast heterogeneity and complexity, and unreliable content. For this reason, the methodologies will be illustrated in the context of the Open Science paradigm, characterized by the repeatability, reproducibility, and cross-domain reuse of all experimental phases.

The course will be interactive and made up of practical exercises. Attendees will use online environments to parametrize the models, run the experiments, and potentially modify the models.

Course Contents in brief:

- Big data and marine data
- Geospatial data
- Parameter selection techniques for environmental variables
- Distance and density-based cluster analysis for habitat and vessel pattern recognition
- Artificial Neural Networks, Support Vector Machines, and Maximum Entropy models for species distribution modelling
- Techniques for time series forecasting applied to marine data
- Open Science approaches

Total # of hours of lecture: 16

References:

Coro, G. (2020). OPEN SCIENCE AND ARTIFICIAL INTELLIGENCE SUPPORTING BLUE GROWTH. Environmental Engineering & Management Journal (EEMJ), 19(10).

Coro, G., Sana, L., & Bove, P. (2024). An open science automatic workflow for multi-model species distribution estimation. International Journal of Data Science and Analytics, 1-20.

Coro, G., Bove, P., & Kesner-Reyes, K. (2023). Global-scale parameters for ecological models. Scientific Data, 10(1), 7.

Coro, G., Tassetti, A. N., Armelloni, E. N., Pulcinella, J., Ferrà, C., Sprovieri, M., ... & Scarcella, G. (2022). COVID-19 lockdowns reveal the resilience of Adriatic Sea fisheries to forced fishing effort reduction. Scientific Reports, 12(1), 1052.

Coro, G., Vilas, L. G., Magliozzi, C., Ellenbroek, A., Scarponi, P., & Pagano, P. (2018). Forecasting the ongoing invasion of Lagocephalus sceleratus in the Mediterranean Sea. *Ecological Modelling*, *371*, 37-49.

Coro, G., Panichi, G., Scarponi, P., & Pagano, P. (2017). Cloud computing in a distributed e-infrastructure using the web processing service standard. *Concurrency and Computation: Practice and Experience*, *29*(18), e4219.

Coro, G., Candela, L., Pagano, P., Italiano, A., & Liccardo, L. (2015). Parallelizing the execution of native data mining algorithms for computational biology. *Concurrency and Computation: Practice and Experience*, *27*(17), 4630-4644.

Coro, G., Bove, P., Armelloni, E. N., Masnadi, F., Scanu, M., & Scarcella, G. (2022). Filling gaps in trawl surveys at sea through spatiotemporal and environmental modelling. Frontiers in Marine Science, 9, 919339.

CV of the Teacher

Gianpaolo Coro is a Physicist with a PhD in Computer Science. His research focuses on Artificial Intelligence and Data Mining. He has been working for 20 years on Machine Learning and Signal Processing with applications to Computational Biology, Brain Computer Interfaces, Language Technologies and Cognitive Sciences. The aim of his research is the study and experimentation of models and methodologies to process biological data, and the application of the results to ecological problems management. His approach relies on distributed e-Infrastructures and uses parallel and distributed computing via Grid and Cloud based technologies through an Open Science approach.

Final Exam: Re-application of a selection of methods practiced during the course.

Room and Schedule

Room: *Aula Riunioni del Dipartimento di Ingegneria dell'Informazione, Via G. Caruso 16, Pisa – Ground Floor*

Schedule:

Day1 – May 6, 2025 – h. 9.00-13.00 Introduction to marine data and Open Science methodologies

Day2 – May 7, 2025 – h. 9.00-13.00 Data selection techniques and pattern recognition

Day3 – May 8, 2025 – h. 9.00-13.00 Supervised modelling of species distributions and invasions

Day4 – May 9, 2025 – h. 9.00-13.00 Data mining techniques for extracting knowledge from biodiversity and vessel data