Cognome Nome: BELLIN NICOLO’

Affiliazione: Università degli studi di Parma

Ciclo di dottorato: XXXV

Titolo della tesi/ricerca: Applications of machine learning techniques in ecology

Relatori: Pierluigi Viaroli e Valeria Rossi

***ANNO: 2019-2020***

***1 Research***

**1.1 Deep convultional neural network: an artificial eye for species recognition**

A future approach for the species identification by wing venation analysis will be the development of a deep convolutional neural network (DCNN). Park et al. (2020) achived image – based automatic classification of vector mosquiotoes using DCNN with operations of transfern learning and fine tuning. They reach high accurancy rate (97 %) and found that the Convolutional neural network visualized the same morphological features used by an human expert for the classification task. Deep learning algorithms could be a big improvement in suirvellance campaign in absence of human experts able to recognize vector species and to rise the classification accurancy in epidemiological studies. I developed some classifier with different network topologies and different values of the hyperparameters. These operations were combined with regularization methods to prevent overfitting. At present the best DCNN reach a classification accurancy of 65%. Future operations to improve the species classification will be: (1) To increase the number of observations combined with on the fly data augmentation approach. (2) transfer learnig using a DCNN pretrained for other classification task and (3) fine tuning operations. The use of DCNN will be applied also on a dataset of zooplankton images in order to automatizing the classification of freshwater zooplankton species. After trainining the DCNN model we will apply the method of the heatmaps of class activation. This procedure is useful for understanding which parts of an image were identified as belonging to a given class, and thus allows to the human eye to localize particular features in images.

**1.2 A model of egg bank dynamics in ephemeral ponds**

Resting life stages (e.g. dormant seeds and resting eggs) have important implications for ecological and evolutionary processes. In this study, we simulated the impact of different environmental scenarios on the dynamics of resting eggs that make up an “egg bank” of a common fresh water ostracod, *Heterocypris incongruens* (Crustacea). A general model for selection on seed germination in unpredictable environments was used to simulate within pond egg bank dynamics. Metapopulation dynamics were simulated using Levin's and Hanski's models assuming three generalized spatial patterns of pond distribution (random, aggregated along the main wind direction, evenly spaced along the main wind direction) and two dispersion processes (random walk and wind shear). We applied global sensitivity and uncertainty analysis (GSUA) to the models. We estimated the egg bank growth rate based on 30-year simulations under present climatic conditions, and assuming a 2°C rise in winter temperature under global climate change.

**1.3 Egg bank dynamics and extinction rates in ephemeral ponds**

We simulated the egg bank dynamics of six clonal lineages of *Heterocypris incongruens* (Ostracoda) from Northern Italy. Two clonal lineages (W1 and W2) are the most common and “winter ecotypes”. Two clonal lineages (S1 and S2) are allochthonous and “summer ecotypes”. Two clonal lineages (I1 and I2) are relatively rare and generalist in term of seasonality. Fecundity and proportion of resting eggs vary by clonal lineage, temperature and photoperiod. We predicted the clonal extinction risk in present and climate change conditions and we assessed the migration northwards by winter clonal lineages from South Italy. We used the general model of Cohen for simulating egg bank dynamics and we estimated the extinction rate of each clonal lineage with uncertainty analysis. We reported a 30-year simulation in present climate conditions assuming a long-term prediction of climate change under the RCP 8.5 scenario.

***2 Attività formative***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CORSI SPECIFICI E ATTIVITA' MULTIDISCIPLINARI** Corsi su argomenti specifici offerti da Dottorati o da Master / Laurea, Seminari, Scuole, Workshop, Training | | | | | | | | |
| **ATTIVITA'** | | **DATA** | **Disciplina scientifica1** | **N. di ore** | **ECTS2** | **Voto – Giudizio3** | **Teacher** | **Grado**  **(PhD, Master, etc.)** |
| ***WORKSHOPS*** | |  |  |  |  |  |  |  |
| 1 | Il valore della complessità | 31/01/2020 | - | 4.5 | 0.6 | - | Luisa De Cola,  Roberto Ferrari, Roberta Ruotolo, Alessandro Pedrini, Francesco Mezzadri, Monica Mattarozzi, Cristina Sissa, Davide Persico | Master |
| ***SEMINARI*** | |  |  |  |  |  |  |  |
| 1 | Genomic logic of biological diversity | 08/09/2020 | BIO/05 | 2 | 0.3 | - | Riccardo Papa | PhD |
| ***CORSI*** | |  |  |  |  |  |  |  |
| 1 | Machine Learning and data analytics | 12/03/2020 – 28/05/2020 | ING-INF/05 | 48 | 6 | Id | Stefano Cagnoni | Master |
| 2 | Planctologia | 01/03/2020 - 31/05/2020 | BIO/07 | 52 | 6 | Id | Valeria Rossi | Bachelor |
| 3 | Ecological statistic with R | 1/09/2020 - 4/09/2020 | BIO/07 | 23 | 3 | Id | Raul Primicerio | PhD |
| ***SCUOLE*** | |  |  |  |  |  |  |  |
| 1 | Artificial intelligence and machine learning | 08/04/2020 – 08/05/2020 | ING-INF/05 | 60 | 9 | Id | Gardella  Corni  Destri | Master |

***3 Altre attività***

|  |
| --- |
| 1. Attività di tutoraggio Ecologia e etologia per la conservazione della natura |
| 1. Attività di tutoraggio Scienze e tecnologie per l’ ambiente e le risorse |
| 1. Correlatore tesi triennale biologia |

***ANNO: 2020-2021***

***1 Research***

**1.1 Geometric morphometrics and machine learning as tools for the identification of sibling mosquito species of the Maculipennis complex (*Anopheles*).** Geometric morphometrics allows researchers to use the specific software to quantify and to visualize morphological differences between taxa from insect wings. Our objective was to assess wing geometry to distinguish four Anopheles sibling species of the Maculipennis complex, *An. maculipennis s. s.*, *An. daciae sp. inq.*, *An. Atroparvus* and *An. melanoon*, found in Northern Italy. We combined the geometric morphometric approach with different machine learning alghorithms: support vector machine (SVM), random forest (RF), artificial neural network (ANN) and an ensemble model (EN). Centroid size was smaller in *An. atroparvus* than in *An. maculipennis s. s.* and *An. daciae sp. inq.* Principal component analysis (PCA) explained only 33% of the total variance and appeared not very useful to discriminate among species, and in particular between *An. maculipennis s. s.* and *An. daciae sp. inq.* The performance of four different machine learning alghorithms using procrustes coordinates of wing shape as predictors was evaluated. All models showed ROC-AUC and PRC-AUC values that were higher than the random classifier but the SVM algorithm maximized the most metrics on the test set. The SVM algorithm with radial basis function allowed the correct classification of 83% of *An. maculipennis s. s.* and 79% of *An. daciae sp. inq.* ROCAUC analysis showed that three landmarks, 11, 16 and 15, were the most important procrustes coordinates in mean wing shape comparison between *An. maculipennis s. s.* and *An. daciae sp. inq.* The pattern in the threedimensional space of the most important procrustes coordinates showed a clearer differentiation between the two species than the PCA. Our study demonstrated that machine learning algorithms could be a useful tool combined with the wing geometric morphometric approach.

**1.2 Unsupervised Machine Learning and Data Mining Procedures Reveal Short Term, Climate Driven Patterns Linking Physico-Chemical Features and Zooplankton Diversity in Small Ponds.** Machine Learning (ML) is an increasingly accessible discipline in computer science that develops dynamic algorithms capable of data-driven decisions and whose use in ecology is growing. Fuzzy sets are suitable descriptors of ecological communities as compared to other standard algorithms and allow the description of decisions that include elements of uncertainty and vagueness. However, fuzzy sets are scarcely applied in ecology. In this work, an unsupervised machine learning algorithm, fuzzy c-means and association rules mining were applied to assess the factors influencing the assemblage composition and distribution patterns of 12 zooplankton taxa in 24 shallow ponds in northern Italy. The fuzzy c-means algorithm was implemented to classify the ponds in terms of taxa they support, and to identify the influence of chemical and physical environmental features on the assemblage patterns. Data retrieved during 2014 and 2015 were compared, taking into account that 2014 late spring and summer air temperatures were much lower than historical records, whereas 2015 mean monthly air temperatures were much warmer than historical averages. In both years, fuzzy c-means show a strong clustering of ponds in two groups, contrasting sites characterized by different physico-chemical and biological features. Climatic anomalies, affecting the temperature regime, together with the main water supply to shallow ponds (e.g., surface runoff vs. groundwater) represent disturbance factors producing large interannual differences in the chemistry, biology and short-term dynamic of small aquatic ecosystems. Unsupervised machine learning algorithms and fuzzy sets may help in catching such apparently erratic differences.

***2 Attività formative***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CORSI SPECIFICI E ATTIVITA' MULTIDISCIPLINARI** Corsi su argomenti specifici offerti da Dottorati o da Master / Laurea, Seminari, Scuole, Workshop, Training | | | | | | | | | |
| **ATTIVITA'** | | **DATA** | **Disciplina scientifica1** | **N. di ore** | **ECTS2** | **Voto – Giudizio3** | | **Teacher** | **Grado**  **(PhD, Master, etc.)** |
| ***SEMINARI*** | |  |  |  |  | |  |  |  |
| 1 | Darwin Day: Spillover ecologia ed evoluzione di una pandemia | Maggio | BIO/07 | 3.30 | 1 | | Id | Docenti vari | PhD |
| ***CORSI*** | |  |  |  |  | |  |  |  |
| 1 | Species distribution Modeling with Remote Sensing | 12,17,19 Agosto | BIO/07 | 6 | 1 | | Voto | Docenti Vari  (NASA) | PhD |
| 2 | Introduzione ai modelli misti, maximum likelihood e statistica bayesiana con R | Giugno | BIO/07 | 24 | 3 | | Id | Stefano Leonardi | PhD |
| 3 | A primer in ecological networks. Data and theory  Summer school | Settembre | BIO/07 | 24 | 3 | | Id | Antonio Bodini | PhD |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ATTIVITA' INTERDISCIPLINARI (SOFT SKILLS)** Corsi interdisciplinari dedicati a competenze generali, ad esempio comunicazione, competenze informatiche, gestione della ricerca o reti e proprietà intellettuale ecc. | | | | | | | | |
| **ATTIVITA'** | | **DATA** | **Disciplina scientifica1** | **N. di ore** | **ECTS2** | **Voto – giudizio3** | **Teacher** | **Grado**  **(PhD, Master, etc.)** |
| ***WORKSHOPS*** | |  |  |  |  |  |  |  |
| 1 | Modelling hail probability over Italy with a machine Learning approach | Giugno | ING-INF/05, GEO/12 | 1 | 0.1 | Id | Centro euro-Mediterraneo sui Cambiamenti Climatici (CMCC) | PhD |
| ***SEMINARI*** | |  |  |  |  |  |  |  |
| 1 | Seminari di Ecologia e Sostenibilità S.It.E | Febbraio - Ottobre | BIO/07 | 10 | 1 | Id | Docenti Vari | PhD |

***3 ALTRE ATTIVITA’***

|  |
| --- |
| 1. Tutor corso di Studio: Ecologia ed Etologia per la conservazione della Natura (Unipr) |
| 2. ATTIVITA’ DIDATTICHE INTEGRATIVE: Analisi dei Dati ecologici (Unipr) |
| 3. ATTIVITÀ DI TUTORATO E ATTIVITÀ DIDATTICO-INTEGRATIVE, PROPEDEUTICE E DI RECUPERO nell’ambito del progetto PLS - Progetto Nazionale Biologia Biotecnologie (Unipr) |

***ANNO: 2021-2022***

***1 Research***

**1.1 Unsupervised machine learning and geometric morphometrics as tools for the identification of inter and intraspecific variations in the Anopheles Maculipennis complex**

Geometric morphometric analysis was combined with two different unsupervised machine learning algorithms, UMAP and HDBSCAN, to visualize morphological differences in wing shape among and within four Anopheles sibling species (*An. atroparvus*, *An. melanoon, An. maculipennis* s.s. and *An. daciae* sp. inq.) of the Maculipennis complex in Northern Italy. Specifically, we evaluated: (1) wing shape variation among and within species; (2) the consistencies between groups of *An. maculipennis* s.s. and *An. daciae* sp. inq. identified based on COI sequences and wing shape variability; and (3) the spatial and temporal distribution of different morphotypes. UMAP detected at least 13 main patterns of variation in wing shape among the four analyzed species and mapped intraspecific morphological variations. The relationship between the most abundant COI haplotypes of *An. daciae* sp. inq. and shape ordination/variation was not significant. However, morphological variation within haplotypes was reported. HDBSCAN also recognized different clusters of morphotypes within *An. daciae* sp. inq. (12) and *An. maculipennis* s.s. (4). All morphotypes shared a similar pattern of variation in the subcostal vein, in the anal vein and in the radio-medial cross-vein of the wing. On the contrary, the marginal part of the wings remained unchanged in all clusters of both species. Any spatial-temporal significant difference was observed in the frequency of the identified morphotypes. Our study demonstrated that machine learning algorithms are a useful tool combined with geometric morphometrics and suggest to deepen the analysis of inter and intra specific shape variability to evaluate evolutionary constrains related to wing functionality.

**1.2 Species distribution modeling and machine learning in assessing the potential distribution of freshwater zooplankton in Northern Italy**

Here we explore a species distribution model framework (SDM) combined with machine learning algorithms to describe the distribution of two freshwater zooplankton species *Daphnia longispina* (Cladocera) and *Eucyclops serrulatus* (Copepods) in a system of 283 shallow and ephemeral freshwater habitats in the Northern Italian Appennines. We model the habitat suitability by comparing one regression-based model, one generalized linear model (GLM) and two machine learning algorithms: random forest (RF) and artificial neural network (ANN) with one hidden layer. We used a total of 27 predictor variables. The modeling framework was used considering a scenario of future climate change in order to evaluate potential shifts in spatial distribution of the zooplankton species. For both species, the supervised machine learning algorthn (ANN) produced the highest mean values for all the performance metrics. Both species, in a future climatic change scenario, are expected to shift their distribution mainly toward lower northern altitudes with an overall expansion of 7% with respect to the past/present climatic conditions.

**1.3 Make the CPUs do the hard work - Automated acoustic feature extraction and visualization for marine ecoacoustics applications illustrated using marine mammal Passive Acoustic Monitoring datasets**

We present an alternative to the use of ecoacoustic indices and describe the application of multiple machine learning techniques to the analysis of a large PAM dataset. We combine pre-trained acoustic classification models, dimensionality reduction, and random forest algorithms to demonstrate how machine-learned acoustic features capture different aspects of the marine environment. We processed two PAM databases and conducted 13 trials showing how acoustic features can be used to: i) discriminate between the vocalizations of marine mammals, beginning with high-level taxonomic groups, and extending to detecting differences between conspecifics belonging to distinct populations; ii) discriminating amongst different marine environments; and iii) detecting and monitoring anthropogenic and biological sound sources. We argue that acoustic feature extraction, visualization, and analysis allows the retention of most of the environmental information contained in PAM recordings, overcoming the limitations encountered when using ecoacoustics indices.

***2 Attività formative***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CORSI SPECIFICI E ATTIVITA' MULTIDISCIPLINARI** Corsi su argomenti specifici offerti da Dottorati o da Master / Laurea, Seminari, Scuole, Workshop, Training | | | | | | | | | |
| **ATTIVITA'** | | **DATA** | **Disciplina scientifica1** | **N. di ore** | **ECTS2** | **Voto – Giudizio3** | | **Docente** | **Grado**  **(PhD, Master, etc.)** |
| ***WORKSHOPS*** | |  |  |  |  | |  |  |  |
| 1 | International Workshop on the Integration of Genomic and Geographic Information System data for wildlife conservation (WIGGIS) | 15/09/2022-16/09/2022 | Bio/07 | 16 | 2 | | Id | Stéphane Joost  Elia Vajana  Oliver Selmoni  Marco Andrello | PhD |
| ***CORSI*** | |  |  |  |  | |  |  |  |
| 1 | Ecosystem science | 19-20-21 Ottobre | Bio/07 | 24 | 3 | | Id | Prof. Viaroli  Prof. Rossetti  Prof. Nizzoli | PhD |
| ***ALTRO*** | |  |  |  |  | |  |  |  |
| 1 | PhD DAY | 7-8 Luglio 2022 | Bio/07 | 16 | 2 | | Id | Università di Firenze, Ferrara e Parma | phD |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ATTIVITA' INTERDISCIPLINARI (SOFT SKILLS)** Corsi interdisciplinari dedicati a competenze generali, ad esempio comunicazione, competenze linguistiche, competenze informatiche, gestione della ricerca o reti e proprietà intellettuale ecc. | | | | | | | | |
| **ATTIVITA'** | | **DATA** | **Disciplina scientifica1** | **N. di ore** | **ECTS2** | **Voto – giudizio3** | **Teacher** | **Grado**  **(PhD, Master, etc.)** |
| ***SEMINARI*** | |  |  |  |  |  |  |  |
| 1 | Evaluating Ecosystem Services with Remote Sensing | 23-25-30 Agosto 2022 | Bio/07 | 7 | 1 | Id | Nasa Arset | PhD |
| 2 | Using Earth Observations for Pre- and Post- Fire Monitoring | 18-20 Gennaio 2022 | ING-INF/05,  GEO/12 | 7 | 1 | Id | Nasa Arset | PhD |

***3 Altre attività***

|  |
| --- |
| 1. TUTOR CORSO DI STUDIO: Biologia Laurea Triennale (Università di Parma) |
| 2. ATTIVITA’ DIDATTICHE INTEGRATIVE: Analisi dei Dati ecologici (Università di Parma) |
| 3. Attività di orientamento per studenti della scuola secondaria di secondo grado attraverso attività laboratoriali |