Studi Trent. Sci. Nat., 86 (2009): 47-48 ISSN 2035-7699

Atti XVIII Convegno Gadio 2008: *Un mondo che cambia: successioni ecologiche, invasioni biologiche ed alterazioni antropiche*
Sessione 4 - Poster

Short note - Nota breve

Decision tree approach to model *Salmo marmoratus* presence in Piedmont (North-Western Italy)

Tina TIRELLI* & Daniela PESSANI

Dipartimento di Biologia Animale e dell'Uomo, Università degli studi di Torino, Via Accademia Albertina 13, 10123 Torino, Italy
* Corresponding author e-mail: *santina.tirelli@unito.it*

RIASSUNTO - *Realizzazione di alberi decisionali per predire le caratteristiche dell'habitat occupato da* Salmo marmoratus *in Piemonte* - In Piemonte, il considerevole impatto delle attività antropiche sui fiumi ha avuto conseguenze pesanti sulla flora e sulla fauna. E' quindi indispensabile sviluppare strumenti che forniscano valutazioni ecologiche accurate dei corsi d'acqua e dei loro abitanti. Nella presente ricerca, sono stati costruiti alberi decisionali (DTs) per predire le caratteristiche dell'habitat occupato da Salmo marmoratus(Couvier 1817), salmonide a rischio di estinzione, utilizzando un dataset di 198 siti piemontesi. Il database è costituito da 7 parametri ambientali (inputs) e da dati di presenza/assenza (output). I modelli ottenuti mostrano performance ed attendibilità piuttosto basse. Si suggerisce pertanto di utilizzare diverse tecniche per costruire modelli predittivi di presenza/assenza al fine di evitare l'uso di tecniche improprie per affrontare specifici problemi di management.

Key words: decision trees, ecological modelling, species prediction, Piedmont (Italy)
Parole chiave: alberi decisionali, modelli ecologici, predizione di specie, Piemonte

1. INTRODUCTION

There are several well known machine-learning techniques highly suitable for habitat modeling (Goethals & De Pauw 2001; Dakou *et al*. 2007). We applied DT to model presence/absence of *Salmo marmoratuss* (Couvier 1817), an endangered salmonid listed in the Annex II of the European Union Habitats Directive 92/43/CEE and in the International Union for Conservation of Nature and Natural Resources Red List. The present study aims at evaluating the reliability and applicability of DT to model presence/absence of *S. marmoratus*.

2. STUDY AREA AND METHODS

The study system consisted of 198 sites located along Piedmont rivers (North-Western Italy). *Salmo marmoratus* was recorded at 67 of the sampling sites, corresponding to the 33.38% of them.

We considered a data set of 20 variables and presence/absence fish data, obtained from the "Monitoraggio della fauna ittica in Piemonte" (Regione Piemonte 2006).

River and habitat data were proportionally normalized between 0 and 1 in the range of the maximum and minimum values, the attribute selection - best-first search (Witten & Frank 2005) and Goldberg's genethic algorithm (Goldberg 1989) - allowed to obtain a subset of 7 inputs to buid the trees.

The "Top-Down Induction of Decision Trees" (Quinlan 1986) - J48 algorithm with binary split (Dakou *et al*. 2007) - was used and tree post-pruning optimization (confidence factor 0.15-0.25) applied. Ten fold cross-validation was repeated 10 times. We evaluated the five performance parameters, usually reported in literature, on the basis of matrices of confusion (Fielding & Bell 1997). Unpruned and pruned trees performance were compared (Mann-Whitney tests).

3. RESULTS

The performance of the models were not really satisfactory (CCI= 71.3%, sensitivity= 78.6%, specificity= 59.6%, Cohen's k = 0.35, area under the ROC curve= 0.7); the optimal confidence level of pruning was 0.17. No significant differences in the predictive performance between pruned and unpruned trees were detected.

4. DISCUSSION AND CONCLUSIONS

The obtained models showed predictions based principally on chance (Cohen's k value). The low performance obtained are probably due to the low frequency of occurrence of *S. marmoratus* (33.38%). In fact, the

performance of DTs is strongly related to the frequency of occurrence of the predicted taxa (Goethals *et al*. 2001; Manel *et al*. 2001). Moreover, DTs may not be the best performing technique to be applied in certain cases (see Dakou *et al*. 2007). Therefore, we strongly recommend the use of different techniques to build presence/absence models to avoid using improper techniques to help managing species.

REFERENCES

Dakou E., D'heygere T., Dedecker A.P., Goethals P.L.M., Lazaridou-Dimitriadou M. & De Pauw N., 2007 - Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). *Aquatic Ecol*., 41: 399-411.

Fielding A.H. & Bell J.F., 1997 - A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv*., 24: 38-49.

Goethals P. & De Pauw N., 2001 - Development of a concept for integrated ecological assessment in Flanders, Belgium. *J. Limnol*., 60: 7-16.

Goethals P.L.M., Džeroski S., Vanrolleghem P. & De Pauw N., 2001 - Prediction of benthic macro-invertebrate taxa (Asellidae and Tubificidae) in watercourses of Flanders by means of classification trees. In: IWA 2nd World water congress, Berlin: 5-6.

Goldberg D.E., 1989 - *Genetic algorithm in search, optimization and machine learning*. Addison-Winsley Publishing Company, Reading, 412 pp.

Manel S., Williams H.C. & Ormerod S.J., 2001 - Evaluating presence/absence models in ecology: the need to account for prevalence. *J. Appl. Ecol*., 38: 921-931.

Quinlan JR., 1986 - Induction of decision trees. *Mach. Learn*., 1: 81-106.

Regione Piemonte, 2006 - *Monitoraggio della fauna ittica in Piemonte*. Direzione Pianificazione delle Risorse Idriche. Torino, 149 pp.

Witten I.H. & Frank F., 2005 - *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA, 525 pp.