Chapter IV

Erica Craig
Western Ecological Studies, USA

Falk Huettmann
University of Alaska-Fairbanks, USA

ABSTRACT
The use of machine-learning algorithms capable of rapidly completing intensive computations may be an answer to processing the sheer volumes of highly complex data available to researchers in the field of ecology. In spite of this, the continued use of less effective, simple linear, and highly labor intensive techniques such as stepwise multiple regression continue to be widespread in the ecological community. Herein we describe the use of data-mining algorithms such as TreeNet and Random Forests (Salford Systems), which can rapidly and accurately identify meaningful patterns and relationships in subsets of data that carry various degrees of outliers and uncertainty. We use satellite data from a wintering
INTRODUCTION

Individual species and even entire ecosystems are at risk because of climatic changes and destruction of native habitats that are occurring worldwide, simultaneous with increased pressures from the expansion of human populations (Bittner, Oakley, Hannan, Lincer, Muscolino, & Domenech, 2003; Braun, 2005; Knick, Dobkin, Rotenberry, Schroeder, Vander Haegen, & Van Riper, 2003; Millenium Ecosystem Assessment, 2005; Primack, 1998; Zakri, 2003). Knowing and understanding factors that affect species and even that drive entire systems is vital for assessing populations that are at risk, as well as for making land management decisions that promote species sustainability. Advances in Geographic Information System technology (GIS) and digital online data availability coupled with the ability to collect data on animal movements via satellite and GPS have given rise to large, highly complex datasets that have the potential to provide the global community with valuable information for pursuing these goals. However, the sheer volume and complexity of such animal location data can overwhelm biologists charged with making resource management decisions (Huettmann, 2005 for data overview). Further, it can affect the ability to obtain accurate results and to find the best possible solutions for making sustainable decisions. These major obstacles often result in under-utilization of data. Not only is it difficult to accurately filter out erroneous animal locations, it is challenging to identify meaningful patterns from data with multi-dimensional input variables (Braumoeller, 2004). Traditional statistical regression methods are limited by the inability to truly meet the assumptions required for analysis, such as the distribution of variables, model fit, independence of variables and linearity of the data (James & McCulloch, 1990; Nielsen, Boyce, Stenhouse, & Munro, 2002). They also are incapable of explaining the relationships between response and predictor variables (De’ath 2007). However, researchers continue to use the very time consuming, general linear models which use stepwise multiple regression methods (e.g., Manly, McDonald, Thomas, McDonald, & Erickson, 2002; Nielsen et al., 2002) as the predominant analytical approach for such analyses (see Whittingham, Stephens, Bradbury, & Freckleton, 2006 for an assessment of use of these techniques in the ecological literature).

We propose that linking technology with biological applications is not only relevant, but crucial for obtaining sustainability (Chernetsov & Huettmann, 2005; Onyeahialam, Huettmann, & Bertazzon, 2005). Much of ecological data and factors that affect species and their distributions are highly complex and non-linear in nature. Elith et al. (2006) and Olden and Jackson (2002) found that alternative analytical techniques such as classification and regression trees for modeling species distribution widely outperformed traditional modeling methodologies on non-linear data and performed equally as well on linear data (see also Huettmann & Diamond 2006; Yen, Huettmann, & Cooke 2004). While many scientists still continue to propose the use of traditional methods (Kenward, 2001; Millsbaugh & Marzluff, 2001; Braun, 2005), others are seeking alternative solutions because of the non-linearity of most...