Continental Shelf Research 36 (2012) 19–28

Contents lists available at SciVerse ScienceDirect

Continental Shelf Research

journal homepage: www.elsevier.com/locate/csr

Research papers

Methods of training set construction: Towards improving performance for automated mesozooplankton image classification systems

Chun-Yi Chang[a], Pei-Chi Ho[a], Akash R. Sastri[a,b], Yu-Ching Lee[a], Gwo-Ching Gong[c,e], Chih-hao Hsieh[a,d,*]

[a] Institute of Oceanography, National Taiwan University, 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan
[b] Department of Biological Sciences, Université du Québec à Montréal, Montreal, Canada H3C 3P8
[c] Institute of Marine Environmental Chemistry and Ecology, National Taiwan Ocean University, 2, Pei-Ning Rd., Keelung 20224, Taiwan
[d] Institute of Ecology and Evolutionary Biology, National Taiwan University, 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan
[e] Center of Excellence for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, 2, Pei-Ning Rd., Keelung 20224, Taiwan

ARTICLE INFO

Article history:
Received 25 October 2010
Received in revised form
3 November 2011
Accepted 9 January 2012
Available online 18 January 2012

Keywords:
Balanced training
Category-specific accuracy
Hydrographic heterogeneity
Global training set
Water mass-specific training set
ZooScan

ABSTRACT

The correspondence between variation in the physico-chemical properties of the water column and the taxonomic composition of zooplankton communities represents an important indicator of long-term and broad-scale change in marine systems. Evaluating and relating compositional change to various forms of perturbation demand routine taxonomic identification methods that can be applied rapidly and accurately. Traditional identification by human experts is accurate but very time-consuming. The application of automated image classification systems for plankton communities has emerged as a potential resolution to this limitation. The objective of this study is to evaluate how specific aspects of training set construction for the ZooScan system influenced our ability to relate variation in zooplankton taxonomic composition to variation of hydrographic properties in the East China Sea. Specifically, we compared the relative utility of zooplankton classifiers trained with the following: (i) water mass-specific and global training sets; (ii) balanced versus imbalanced training sets. The classification performance (accuracy and precision) of water-mass specific classifiers tended to decline with environmental dissimilarity, suggesting water-mass specificity However, similar classification performance was also achieved by training our system with samples representing all hydrographic sub-regions (i.e. a global classifier). After examining category-specific accuracy, we found that equal performance arises because the accuracy was mainly determined by dominant taxa. This apparently high classification accuracy was at the expense of accurate classification of rare taxa. To explore the basis for such biased classification, we trained our global classifier with an equal amount of training data for each category (balanced training). We found that balanced training had higher accuracy at recognizing rare taxa but low accuracy at abundant taxa. The errors introduced in recognition still pose a major challenge for automatic classification systems. In order to fully automate analyses of zooplankton communities and relate variation in composition to hydrographic properties, the recognition power of the system requires further improvements.

1. Introduction

A variety of approaches have been employed in monitoring projects that investigate how aquatic ecosystems respond to anthropogenic disturbances and climate change (Harley et al., 2006; Hsieh et al., 2008, 2009, 2010). In particular, monitoring the response of zooplankton communities is especially useful and important, because zooplankton (i) are sensitive indicators to environmental changes (Beaugrand et al., 2002), (ii) constitute an important biogeochemical component of aquatic systems (Wohlers et al., 2009), and (iii) represent the dominant link between phytoplankton communities and higher trophic groups such as fish and seabirds (Hsieh et al., 2005; Beaugrand et al., 2010).

Of particular interest is the ability to rapidly gather accurate information about how the taxonomic composition and the size distribution of planktonic communities varies over large spatial and temporal scales (Benfield et al., 2007). Unfortunately, the time, expertise, and effort required to manually enumerate and microscopically identify zooplankton represents a major obstacle in routine application of zooplankton studies to monitoring programs

[*] Corresponding author at: Institute of Oceanography, National Taiwan University, 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan. Tel.: +886 2 33669745; fax: +886 2 33669746.
E-mail address: chsieh@ntu.edu.tw (
.-h. Hsieh).

0278-4343/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csr.2012.01.005

(MacLeod et al., 2010). Recognizing this problem, zooplankton taxonomists have devoted considerable attention to the development and application of automated methods that may be routinely used as part of large-scale and long-term plankton monitoring programs (Wiebe and Benfield, 2003).

Rapid developments in image capturing and analyses, machine learning techniques and computing power have led to the recent emergence of automated mesozooplankton image classification methods (see review in Benfield et al., 2007). Preliminary applications of artificial neural networks to pattern recognition of planktonic organisms have gained significant attention because they have clearly demonstrated the feasibility of the methodology (Simpson et al., 1992; Culverhouse et al., 1996; MacLeod et al., 2010). At its extreme, some studies have proposed the use of automated in situ underwater observation tools that can potentially collect real-time information on plankton distribution patterns; however, resolving accuracy issues associated with both machine learning and identification currently limits the widespread adoption of this approach (Tang et al., 1998; Davis et al., 2004). A compromise between manual microscopy and automated *in situ* profilers is scanning preserved samples with bench-top optical devices. These devices are attractive because collections of preserved zooplankton samples can be analyzed (identification and individual size measurement) in a short time. To date, the methods employing commercially available scanners have generally demonstrated a good correspondence between automatic (scanned images) and manual abundance counts (Bell and Hopcroft, 2008; Plourde et al., 2008; Gislason and Silva, 2009).

Most of the aforementioned studies agree that careful evaluations of the types of data used in training set construction are needed because training set structure can significantly affect the performance and utility of classifiers for use in ecological studies. For instance, strong hydrographic heterogeneity exists among subregions in our monitoring area, the East China Sea (ECS). This heterogeneity should be reflected in zooplankton community composition (Mackas, 1984; Beaugrand et al., 2009) and suggests that a water mass-specific classifier (WSC) would out-perform a global classifier (GC, i.e. a training set consisting of data pooled from throughout the entire sampling area) within similar water masses where similar zooplankton communities are present. However, constructing a WSC for different water masses limits the broad-scale applicability of automated imaging systems and is far more time consuming than constructing a GC for an entire region.

In addition, the hollow curve that typically describes the species abundance distribution of natural populations (McGill et al., 2007) presents another issue surrounding training set construction. This distribution imposes difficulties for training classifiers because taxa abundances within a sample are never equal (i.e. overall abundance is dominated by only a few groups). As a consequence, a classifier trained with natural proportions of taxonomic groups will be imbalanced. Both balanced and imbalanced training have been used in previous studies (Grosjean et al., 2004; Bell and Hopcroft, 2008; Fernandes et al., 2009; Gislason and Silva, 2009; Gorsky et al., 2010) without addressing the potential impact on either training set type on category-specific performance. From a statistical perspective, the accuracy of an imbalanced training set is usually biased toward the dominant categories at the expense of rarer categories (Chen et al., 2004; Thomas et al., 2006). Such outcomes are undesired if our purpose is to obtain useful community-level data.

Here, we compared (i) the performances of WSC and GC, and (ii) the category-specific performance of classifiers constructed from imbalanced (naturally occurring) and balanced training sets. Our comparisons were carried out using zooplankton samples collected during several research cruises in the ECS (2006–2009). Samples were processed using the ZooScan system (Grosjean et al., 2004; Gorsky et al., 2010). We evaluated our results in an ecologically relevant context by considering how the taxonomic composition of data obtained from various classification schemes varied with hydrographic properties of our study area.

2. Materials and methods

2.1. Study area

The East China Sea (ECS, Fig. 1) is the largest marginal sea of the western Pacific and is characterized by considerable hydrographic heterogeneity. The heterogeneous nature of the ECS is due to the complicated circulation of several currents, such as the China Coastal Current, Yangtze River Diluted Water, Taiwan Warm Current, Kuroshio Branch Current, and Yellow Sea Mixed Water (Liu et al., 2003). In addition, circulation patterns are modulated by the East Asia monsoon (Lee and Chao, 2003). As a result, the ECS displays a strong spatial gradient and seasonality in its environmental characteristics, e.g. temperature, salinity, and nutrients (Gong et al., 1996, 2003).

2.2. Mesozooplankton sampling and digitizing

Mesozooplankton samples were collected in the ECS from 2006 to 2009 (Fig. 1). Sampling stations in each research cruise varied depending on weather conditions. Station-specific environmental properties during each cruise are provided in Appendix A. Plankton samples were collected using a 330 μm ORI net (Ocean Research Institute plankton net) with a 160 cm diameter mouth. The net was hauled obliquely from 10 m above the sea floor to the surface at a speed of 0.3 m s^{-1}. When the water depth was greater than 200 m, the net was hauled from 200 m to the surface. Mesozooplankton samples were fixed with 10% buffered formalin and brought back to the lab for digitizing.

Prior to scanning, samples were split into approximately 1500–2000 objects per aliquot (Grosjean et al., 2004) using a

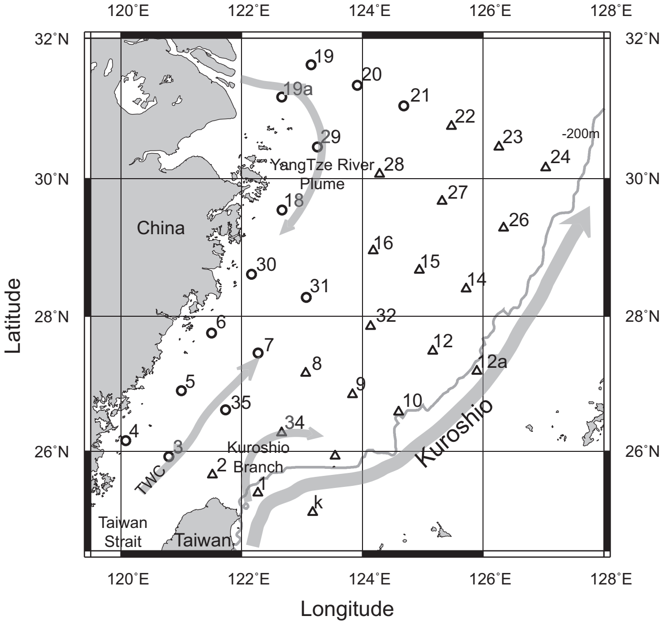

Fig. 1. Map of sampling stations in the East China Sea. Circles represent nearshore stations (distance to coastline within 100 nautical miles); triangles represent offshore stations (distance to coastline beyond 100 nautical miles). TWC, Taiwan Warm Current.