Semi- and fully supervised quantification techniques to improve population estimates from machine classifiers

Eric C. Orenstein,1,* Kasia M. Kenitz,1 Paul L.D. Roberts,2 Peter J.S. Franks,1 Jules S. Jaffe,1 Andrew D. Barton1,3
1Scripps Institution of Oceanography, University of California San Diego, La Jolla, California
2Monterey Bay Aquarium Research Institute, Monterey, California
3Section of Ecology, Behaviour and Evolution, University of California San Diego, La Jolla, California

Abstract

Modern in situ digital imaging systems collect vast numbers of images of marine organisms and suspended particles. Automated methods to classify objects in these images – largely supervised machine learning techniques – are now used to deal with this onslaught of biological data. Though such techniques can minimize the human cost of analyzing the data, they also have important limitations. In training automated classifiers, we implicitly program them with an inflexible understanding of the environment they are observing. When the relationship between the classifier and the population changes, the computer’s performance degrades, potentially decreasing the accuracy of the estimate of community composition. This limitation of automated classifiers is known as “dataset shift.” Here, we describe techniques for addressing dataset shift. We then apply them to the output of a binary deep neural network searching for diatom chains in data generated by the Scripps Plankton Camera System (SPCS) on the Scripps Pier. In particular, we describe a supervised quantification approach to adjust a classifier’s output using a small number of human corrected images to estimate the system error in a time frame of interest. This method yielded an 80% improvement in mean absolute error over the raw classifier output on a set of 41 independent samples from the SPCS. The technique can be extended to adjust the output of multi-category classifiers and other in situ observing systems.

Marine biologists and ecologists increasingly use imaging technologies to study life in the ocean (Culverhouse et al. 2003; Benfield et al. 2007; Lombard et al. 2019). With digital imaging instruments, scientists are able to interrogate the environment at high spatiotemporal resolution, amassing huge amounts of data (Gorsky et al. 2000; Olson and Sosik 2007; Cowen and Guigand 2008; Beijbom et al. 2012; French et al. 2019). Imaging technologies have led to fascinating discoveries, enabled new experimental designs, and facilitated many in situ studies of diverse organisms: parasites of plankton, coral polyps, global-scale protist populations, the role of larvaceans in carbon cycling, and ecological drivers of colony formation to name a few (Peacock et al. 2014; Biard et al. 2016; Mullen et al. 2016; Katiya et al. 2017; Kenitz et al. 2020). But the volume of data is unwieldy – no human, or team of humans, could feasibly taxonomically classify all the data collected by a single instrument. As a result, researchers use automated classification techniques to sort their data (Blaschko et al. 2005; Sosik and Olson 2007; Gorsky et al. 2010; Beijbom et al. 2012; Ellen et al. 2015). Recent advances in image recognition with deep learning algorithms have been particularly encouraging, far surpassing the performance of previous efforts to automatically classify in situ image data (Orenstein and Beijbom 2017; Luo et al. 2018; Ellen et al. 2019).

Most automated classification systems for marine biological image data are supervised: a human expert labels a set of images that are used to train and validate their chosen algorithm. Once the computer is performing at an acceptable level, the classifier is applied to an unlabeled target data set to assess the composition of the community. Regardless of the specific algorithm that is applied, this workflow implicitly assumes that the underlying distributions and relative abundances of the organisms being sampled is constant (Daume III and Marcu 2006; Hand 2006; Pan and Yang 2010). That assumption is difficult to satisfy in oceanographic deployment scenarios, leading to degradation in the quality of annotations returned by a computer classifier. The behavior is broadly described as “dataset shift”: some of the unseen data are
drawn from a different distribution of classes than the original training data.

Dataset shift is particularly challenging in the context of biological oceanography as population fluctuations within a community are often the signals scientists wish to detect. The ocean is a dynamic environment and populations of marine organisms can shift over very short time and space scales (Hauri et al. 1978; Margalef 1978; Franks 2005). In practical terms, this means that the quality of the annotations from a machine classifier will degrade as the composition of the true population deviates from the training set. This could occur due to radical changes in the abundance of a particular organism, the absence of a previously present class, the appearance of a new class, or a change in the appearance of a type of organism.

Dataset shift is a pervasive problem in applied machine learning that has been studied extensively in the context of financial and text analysis (Forman 2008; López et al. 2013; Tasche 2017). Oceanographers have noted dataset shift and proposed solutions such as using a confusion matrix assessed on an independent sample to estimate and correct classifier bias (Solow et al. 2001). A recent study by González et al. (2019) applied three such quantification methods: one relies on an independent estimate of the classifier errors, another casts the original classifier as a probabilistic model, and a third matches the probability distribution of the output and the training set. All of these methods, however, require the knowledge of the community being studied and assume some amount of stability in the presence, relative abundances, and appearance of all the classes.

Here, we compare a simple quantifier to two proposed approaches to explicitly treat dataset shift: adjusted count (AC), prevalence-based adjusted count (pAC), and supervised quantification. AC shifts the raw classifier output with a single estimate of classifier performance made during training and has been previously been applied to the output of plankton image classifiers (Forman 2008; González et al. 2019). pAC applies a similar adjustment using an empirically derived relationship between the classifier output and its performance. Supervised quantification requires a human to annotate a small number of images from the classifier output to assess its time-dependent error, ensuring robustness to a variety of changes in the target population. These approaches are described as “quantification,” as opposed to classification, techniques; rather than considering the quality of individual labels, we discuss how well a computer algorithm describes a population (Moreno-Torres et al. 2012). The methods are illustrated in the case of a binary classifier used to search for diatom chains in data generated by the Scripps Plankton Camera System (SPCS) (Orenstein et al., 2020). Finally, we discuss best practices for applying automated classifiers in the field and propose several avenues for future research.

**Background**

Machine learning techniques attempt to develop a mapping between some input data and a desired output (Duda et al. 2012). In the case of supervised learning, researchers tune their chosen algorithm with a set of expertly labeled training data. Once the system is performing to their liking, it is used to classify new data not seen during the training process. In the ideal case, the automated classifier is effective enough to run through the new data with no human intervention.

The supervised training process implicitly assumes that the training data is an independent identically distributed (i.i.d.) subset of the unseen target data (Daume III and Marcu 2006). In practice, however, it is unlikely that the curated set of images used to train the classifier reflect the real distribution of the classes. Particularly in many observational ecology applications, the target data can be quite different than the expertly labeled training set—so-called dataset shift. Here, we will discuss two common types of shift relevant to oceanographic imaging: prior probability and concept shift.

Assume that we have an image dataset, a subset of which has been annotated with a label $y$ drawn from a set of classes $Y$. In an effort to develop a classifier, an engineer designs a processing system to extract features from a set of labeled images $X$. Each labeled image is then converted into a vector of features $x$. Ideally, the class label $y$ uniquely determines the set of features $x$ drawn from an image. In the case of a deep learning system, which considers an entire image, $x$ represents all the pixels in a single image drawn from $X$.

The relationship between the features and the labels can be described mathematically as a joint probability distribution $P(x,y) = P(x|y)P(y)$. Here, $P(x|y)$ is the conditional probability of the features based on the class of the image. $P(y)$ is the prior probability of the class—the likelihood of the class showing up in the data. In the case of a binary classifier, where the system must decide between $y = 1$ and $y = 0$, the optimal Bayes decision rule (BDR) between the classes can be written as:

\[
\text{Decide } y = 1 \text{ if } P(x|1)P(1) > P(x|0)P(0); \text{ otherwise } y = 0. \tag{1}
\]

If the priors are equal, $P(0) = P(1)$, then the classification is made purely based on the likelihood of the features conditioned on the class. When the conditional probabilities are equal, $P(x|0) = P(x|1)$, then the observed features are uninformative and the decision is made entirely on the prior probability of the classes.

When training a supervised binary classifier, the system learns the relationships in the BDR based on a set of training data. Once the algorithm is performing at an acceptable level, it is applied to target data set. In the best case for a scientist interested in applying an automated classifier, the conditional and prior probabilities of the training set are equal to those of the target set.

Prior probability shift occurs when the prior of the test set does not equal that of the training set

\[
P_{\text{train}}(y) \neq P_{\text{target}}(y) \tag{2}
\]
where $P_{\text{train}}$ and $P_{\text{target}}$ are the prior distributions of the training set and the target population, respectively. The conditional probability of the features based on the class remains the same between the training and target set. In oceanographic imaging, prior probability shift might happen when the relative abundances of the observed populations change radically. For example, a towed imaging system might encounter very different communities as it moves through a front (Taylor et al. 2012).

Concept shift describes a scenario in which the relationship between the features and the class is changing

$$P_{\text{train}}(x|y) \neq P_{\text{target}}(x|y)$$

In this case, the relative abundances of the organisms being imaged remain the same between the training and target sets, but a new class might appear with features similar to an existing category. Concept shift could also occur if a population were to change appearance, such as fish growing from larval stages to adulthood.

Most assessments of automated classifiers for marine applications have used an independent subset of their training data — images drawn from the original training set that were not used to tune the algorithm — as a test set (Grosjean et al. 2004; Blaschko et al. 2005; Ellen et al. 2015; Faillettaz et al. 2016). The practice is standard for machine learning experiments since the goal is often to fairly compare techniques rather than producing a workflow that operates effectively on a real population. Tuning an algorithm using such a subset will bias the classifier toward the training distribution since the statistics of the test set are identical to the training data — the inequality in Eq. 2 becomes an equality, meaning adjustments made to the classifier based on performance on the target data tunes the algorithm to the training distribution. Training automated classifiers in this manner potentially renders them incapable of detecting meaningful changes in the target population (González et al. 2017). This consequence of the training and evaluation procedure makes it difficult to detect changes in the abundance of constantly fluctuating oceanographic populations from image data using standard machine learning approaches.

The first treatment of dataset shift in the oceanographic literature was Solow et al. (2001). They described an instance of prior probability shift: the population of zooplankton being imaged by the Visual Plankton Recorder changed, leading to a degradation of the classifier performance. To alleviate the issue, they proposed measuring the misclassification rate of the classifier on an independent test set to construct a confusion matrix. That matrix was then inverted and used to correct the estimated class distribution of the target data. The correction improved the automated estimate of the new population distribution. The technique has subsequently been successfully applied to plankton data sets (Sosik and Olson 2007; Lindgren et al. 2013).

Gorsky et al. (2010) described using a semi-supervised validation procedure to correct automated classification of ZooScan data processed with ZooProcess and PkID software. The automated classifier proposes labels for plankton regions that a human operator verifies. The group was reluctant to use a fully automated procedure because they found that the error rate of the classifier varied as a function of season. While highly accurate, the human cost is substantial, necessitating many hours of a trained expert’s time.

González et al. (2017) recently summarized the issue of dataset shift for planktonic imaging and proposed new metrics for tuning and evaluating the efficacy of automated classifiers. Specifically, they advocated for two things: using a training set that mirrors the target distribution as closely as possible (taking the i.i.d. assumption to the logical extreme) and validating the classifier on an independent test distribution as opposed to an independent collection of individual images — the classifier is evaluated on its ability to mimic a known population distribution rather than evaluating it on a per-image basis. This approach explicitly considers dataset shift and mitigates the issue. However, the amount of human labeled data necessary to tune an algorithm in this way is quite large — often unreasonably so. Moreover, there is no guarantee that the test distributions capture all possible variability in the true target domain. This is especially true when considering time series data where, for example, the average annual community composition might be markedly different from the distribution during shorter time frames.

Here, we discuss several approaches to mitigating dataset shift for in situ imaging studies of plankton. Four methods to estimate the population of a sample will be defined and used to analyze time series data from the SPCS. The performance of each approach will be compared by computing the mean absolute error between the output distribution and the ground truth produced by a human annotator. Finally, each approach will be used on a portion of the time series that has not been fully annotated by a human observer to illustrate best practices for systems designed to detect fluctuations in proportions of classes.

**Methods**

**Scripps plankton camera system**

The SPCS is a set of in situ underwater dark field microscopes designed to continuously observe plankton populations (Orenstein et al., 2020). The system deployed on the Scripps Pier consists of two separate cameras, each with a different magnification; one is fitted with a 0.5X and the other with a 5X microscope objective. Together, the two cameras effectively image organisms ranging in size from 10s of microns to several centimeters in size. The system uses no nets, filters, or pumps; it only images objects that pass through the field of view.
Each microscope in the SPCS contains an on-board computer running custom software to capture and segment images in real time. Multiple organisms can be observed in a single frame. Operating at 8 frames per second, the SPCS captures thousands to millions of regions of interest (ROIs) everyday depending on the ambient density of objects. The ROIs are sent via Ethernet to a database for storage and analysis. The system has been running nearly continuously since March of 2015, with occasional lapses for maintenance, so far collecting nearly 1 billion ROIs.

The database hosts a web interface for browsing and annotating the ROIs (spc.ucsd.edu). The system supports sorting by date, object size, and aspect ratio. The ROIs are displayed as a mosaic organized by object size. Users can select multiple objects at once and apply the same label.

Diatom dataset
A human plankton expert examined data from the 5X SPCS microscope in search of diatom chains (Kenitz et al. 2020). They grouped all diatom species into a single class for the sake of simplicity. All ROIs with a major axis length between 50 and 1000 μm in time period of March 2015 and November 2017 were of interest.

Training data
To build a labeled training set, the expert annotated ROIs in 29, nonconsecutive 5-h chunks of time. A total of 54,117 ROIs were examined, of which 5043 were diatom chains. The remaining non-diatom-chain ROIs were randomly subdivided to create a set of 5043 non-diatom, or noise, images. The 10,086 ROIs were subdivided into a training and validation set using an 80–20 split; 80% of the images were used for training and 20% for validating the system performance. This labeled image set was used to train several classifiers and will be referred to as the training data (examples of diatom chains and non-diatom chains are shown in Figs. 1 and 2, respectively).

Constructing the training data in this way implicitly biases the classifier to expect more chains than it will generally see in the target data. The decision to artificially even the training distribution was made for two reasons: first, we prefer the classifier not miss any chains and the higher relative abundance of diatom chains in the training distribution leads the classifier to overestimate their presence in target data. Second, we

**Fig. 1.** Examples of diatom chain regions of interest from the Scripps Plankton Camera System. These images were identified by a human annotator.
had a limited amount of training data available and were not confident that the full set of labeled images satisfied the i.i.d. assumption.

By forcing the training distribution to be uniform, we are requiring the classifier to make decisions based only on the features (Eq. 1). When applying a classifier trained this way, one can expect the error rates to change as a function of the relative abundances of the classes. In the case of our binary diatom chain classifier, this would suggest that when there are few chains the false positive rate will be high. We choose to accept this consequence of the training procedure to ensure that the computer would miss few chains regardless of the target distribution.

Test data

The expert observed another set of 41, nonconsecutive 5-h chunks of time, a total of 80,715 images, for independent validation of classifier performance. These blocks of data were selected in an effort develop an independent test including radically different underlying distributions of diatoms chains. Thirty-one of these days contain low relative abundances of chains (< 20% of the total), which we consider to be ambient density. Six days had elevated relative abundances of diatom chains, comprising greater than 50% of the total ROIs. A single day would be described as high if greater than 70% of the observed ROIs were classified as diatom chains.

We note that the annotations in each of the 41 d are not pooled to create the test set. Instead, each individual day is considered an independent sample in order to test how well the classifiers work on data from different underlying distributions of plankton (Orenstein et al. 2015; González et al. 2017).

Target data

The remaining approximately 6.5 million unobserved ROIs within the time and size range will be referred to as the target domain. These images have not been directly observed by a human annotator. Automated classifiers tuned with the training data were applied to the target domain to generate estimates of the proportion of diatom chains. The resulting distribution was corrected using all three proposed methods (AC, pAC, and supervised quantification) to illustrate the output in an entirely new region of time.

The complete resulting time series, corrected with the supervised quantification approach, was used by Kenitz et al. (2020) to examine ecological drivers of chain formation. For the remainder of this paper, we restrict the target domain to data from August to November of 2017 for the sake of clarity and brevity.

Machine classification

The training data were subdivided into a training and validation set – 80% and 20% respectively – to tune and evaluate a support vector machine (SVM) and a deep residual network (ResNet) (Cortes and Vapnik 1995; He et al. 2016). The SVM performed comparably to the ResNet but had less discriminative power (Fig. 3). The ResNet will thus be the focus of the rest of the paper. More information about the SVM can be found in the Supplementary Material.

Fig. 2. Examples of non-diatom chain regions of interest from the Scripps Plankton Camera System. These particular examples may be easily confused with chain-like organisms.
Population estimates via adjusted image classifier

Fig. 3. Estimated and true proportions of diatom chains in the 41 independent test sets. The true proportions are produced by a human annotator. The estimates are computed via classify and count from the trained SVM and ResNet. Each point corresponds to the prevalence of diatom tor. The estimates are computed via classify and count from the trained independent test sets. The true proportions are produced by a human annotator. 

The ResNet architecture is widely used by domain scientists seeking to prototype neural networks because it has a number of attractive qualities for repetitive experiments (Cheng et al. 2019; Schröder et al. 2020). In particular, ResNets allow for efficient training of very deep convolutional neural networks by explicitly encoding shortcuts into the network structure. These operations skip blocks of convolutions of the same dimensions, allowing the network to be deeper at little extra computational cost.

An 18-layer network, ResNet-18, was fine-tuned on the labeled diatom data from a version trained on the ImageNet natural image dataset (Russakovsky et al. 2014; Yosinski et al. 2014). Repurposing a network originally trained from natural images has been shown to be effective for plankton applications when limited training data is available (Orenstein and Beijbom 2017). Input ROIs were resized to 256 \times 256 pixel squares before being randomly flipped, sheared, and resized – a procedure known as data augmentation – to ensure the classifier did not overfit to the training data. No images were replicated, only randomly adjusted each time they were input into the ResNet.

The ResNet-18 was written in Pytorch and run on a server-based NVIDIA GTX 1080. The learning rate was set to 0.002, with a decay rate of 0.1 every 7 epochs, and the momentum to 0.9. These parameters were selected to speed the training procedure and facilitate rapid turnaround for new classifiers. The system was trained over 25 epochs with a batch size of 64. The best classifier yielded \sim 95\% accuracy on the internal validation set.

Analysis

**Classify and count (CC)**

The simplest way to estimate the relative abundance of a class of interest is to count the number of images labeled as a given class by an automated classifier. Assume a classifier \( g \) is trained to operate on the sample \( X \). The estimate of the relative abundance \( p_y \), of class \( y \) directly from the classifier is the sum of all objects \( x_i \) classified as \( y \) normalized to the total number of observed objects \( n \):

\[
p_y = \frac{1}{n} \sum_{i=1}^{n} g(x_i) = y
\]

The subscript \( i \) denotes an individual image in the sample. This is the raw output of an automated classifier.

The CC approach assumes that the automated classifier is operating at a high-enough level to be trusted to run independently. Most automated classifiers applied in ocean sciences rely on some form of CC: the computer estimates the class of the data points in a sample and counts them relative to the total number of points. CC will be used as the baseline against which the other methods are compared.

**Adjusted count**

The CC estimate of the relative class distribution can be corrected for prior probability shift with the measured misclassification rate of the classifier. Assume the classifier is trained to look at a binary problem, where \( y \) can take a value of 0 or 1. The adjusted relative abundance of the binary estimate for class 0 can be expressed as:

\[
q_0 = \frac{p_0 - fpr}{tp - fpr}
\]

where \( q_0 \) is the adjusted class distribution for \( y = 0 \) and \( p_0 \) is the estimated proportion of class 0 from the original classifier output. \( fpr \) is the false positive rate of the system as computed from the test set:

\[
fpr = \frac{\sum fp}{\sum fp + \sum tn}
\]

where \( \sum fp \) is the number of false positives and \( \sum tn \) is the number of true negatives. The \( fpr \) is thus the number of false positives normalized to the total number of objects identified as negatives as evaluated from the test set.

The true positive rate, \( tpr \) is defined as:

\[
tpr = \frac{\sum tp}{\sum tp + \sum fn}
\]

where \( \sum tp \) and \( \sum fn \) are the total number of true positives and false negatives from the classifier respectively. The \( tpr \) is how often the classifier is correct over all the true labels it returned.

Both the \( tpr \) and \( fpr \) are parameters estimated from an independent test set after the classifier is trained. The AC correction (Eq. 5) is then applied to all target data based on estimates of the rates from the test set.
described in the medical statistics literature where the sensitivity ($tpr$) and specificity ($1 - fpr$) of a diagnostic test are derived from studies where ground truth medical conditions are known (Zhou et al. 2009).

AC is usually applied based on a single estimate of the classifier performance (Forman 2008; Beijbom et al. 2015; González et al. 2019). This suggests that the AC population estimate is only accurate under a prior probability shift. That is, it assumes that the classifier performance and the relationship between the features and the classes remain constant; only the relative abundances of the classes are changing.

The multiclass case can be treated explicitly by using Solow et al.’s (2001) confusion matrix inversion or by applying the binary adjustment in a one-vs-all manner. But the underlying assumptions remain the same: only the underlying distributions of the classes are changing between the training and test set.

**Prevalence adjusted count**

Deriving an empirical relationship between the classifier and the relative abundance of the classes removes the correction’s dependence on a single estimate of performance. In the binary case, a series of randomly selected samples in time or space can be used to estimate how the classifier performance varies as a function of class prevalence. A regression can then be computed between the rates and the relative abundances estimated by the classifier.

Rather than using $tpr$ and $fpr$ defined above, we will compute the false discovery rate ($fdr$) and false omission rate ($for$) of the classifier:

$$fdr = \frac{\sum fp}{\sum fp + \sum tp} \quad (8)$$

and

$$for = \frac{\sum fn}{\sum fn + \sum tn} \quad (9)$$

where $\sum fp$ and $\sum tp$ are the total number of false and true positives as evaluated in the classifier output. Likewise, $\sum tn$ and $\sum fn$ are the sum of true and false negatives in the selected random sample. The $fdr$ and $for$ are related to the familiar $fpr$ and $tpr$, but are subtly different (Genovese and Wasserman 2002).

Consider the $fdr$ vs. the $fpr$ in the context of the diatom chain classifier. The $fdr$ describes the proportion of all ROIs identified as chains that are not chains (the incorrect chain labels over all ROIs assigned by the computer to the chain class). The $fpr$ instead quantifies the proportion of all ROIs that the computer incorrectly identified as chains over the total number of true non-chains (Table 1). Both are useful metrics, but the $fdr$ has the advantage of being related to the output of just one class (Storey 2003). Estimating it from the classifier only requires examining ROIs labeled as the class of interest.

The $fdr$ and $for$ can be treated as function of the classifier-estimated prevalence of a class of interest. The pAC $q_0$ for class 0 at estimated prevalence level $p_0$ then becomes:

$$q_0(p_0) = p_0 [1-fdr(p_0)] + p_0 [for(p_0)] \quad (10)$$

Where the rates are now functions of $p_0$ as estimated by CC. This is a slight modification of AC using $fdr$ and $for$ to account for classifier bias as a function of the CC output. To extend to the multiclass case, the pAC method can again be applied as a series of binary one-vs-all corrections.

For the current experiments with diatom data, we computed an exponential relationship to estimate the dependence of the $fdr$ and $for$ on the classifier output. The exponential was fit to the number of chains found by the machine classifier vs. performance on the human labeled 41 d of test data (Fig. 4). This empirical model was used to select an $fdr$ and $for$ for a target day based on the classifier output.

**Supervised quantification**

Both versions of AC are a form of bias correction that accounts for prior probability shift. When another type of dataset shift is present, such as concept drift, both AC approaches might miss important dynamics. Such a shift could occur due to a change in the appearance of the classes in the target domain, the introduction of a new class or source of noise, or a radical change in the population distribution. To ensure such shifts do not cause the classifier to miss or falsely identify changes in the population, a human annotator can directly estimate the proportion of false positives produced by the classifier in the target domain. This estimate can then be used to adjust the machine classifiers estimate of the relative abundances of the classes.

The supervised quantification sampling design requires a human annotator to directly observe a subset of the target data after it has been classified (Beijbom et al. 2015). Assuming that the machine classifier has a low false negative rate, the observer can look at the machine output to estimate the false discovery rate of the classifier in the target domain:

### Table 1. Summary of notation and equations. All symbols and equations are described in detail in the Methods section.

| Parameter       | Notation | Equation |
|-----------------|----------|----------|
| False positive  | $fp$     | —        |
| False negative  | $fp$     | —        |
| True positive   | $tp$     | —        |
| True negative   | $tn$     | —        |
| False positive rate | $fpr$ | $fpr = \frac{\sum fp}{\sum fp + \sum tp}$ |
| True positive rate | $tpr$ | $tpr = \frac{\sum tp}{\sum fp + \sum tp}$ |
| False discovery rate | $fdr$ | $fdr = \frac{\sum fp}{\sum fn + \sum tn}$ |
| False omission rate | $for$ | $for = \frac{\sum fn}{\sum fn + \sum tn}$ |
where the total number of false positives $\sum fp$ is found by directly observing a subset of $m$ images assigned to class 0 in the target domain $X$ by the machine classifier.

The supervised correction of the classifier output can be written:

$$q_0 = p_0 \times \left(1 - \text{fdr}_{\text{target}}\right)$$

where $p_0$ is the CC estimate of the proportion of class 0 in the target domain (Eq. 4). This correction is robust to most types of dataset shift since the misclassification rate is directly estimated in every new region of space or time where the computer classifier is applied. Note that this formulation explicitly ignores the contribution of false negatives, predicated on the assumption that the classifier does not miss many examples of the class of interest. If the contribution of for is significant, it can be estimated in a similar fashion and used in Eq. 10.

Like the AC methods, a supervised quantification scheme can be extended to the multiclass case by treating each class in a binary one-vs-all manner. This necessitates estimating $\text{fdr}_{\text{target}}$ for each class by selecting $m$ random objects from the classifier output for human verification.

**Assessment**

**Independent test set**

**Classifier performance**

All images in the 41-sample test set were annotated by the domain expert, yielding a true prior distribution of the relative abundance of chains in each sample. The classifier was then applied to each independent day of image data and CC was used to estimate diatom chain prevalence (Fig. 3). The ResNet overestimated the prevalence of chains on each day – by a particularly wide margin on days with low true proportions of diatom chains. The $\text{fdr}$ and for were computed in each of the 41 independent samples and used to estimate the relationship between the rates and the machine estimated chain prevalence used in the application of pAC (Fig. 4).

The negative exponential relationship between the classifier output and the $\text{fdr}$ reflects our intuition for dataset shift (Fig. 4a). When the actual proportion of chains is far below the proportion in the training set, the $\text{fdr}$ goes up. Likewise, as the abundance of chains increases, the $\text{fdr}$ drops. This behavior is to be expected when the training distribution is forced to be uniform and is exhibited by the ResNet. The for remains below 0.2 for all samples, with most below 0.1, indicating that the classifier does not miss many chains (Fig. 4b).

When the estimated prevalence was below 0.2, the $\text{fdr}$ was generally close to 1, meaning most ROIs labeled as chain were incorrect. This is a function of both the classifier’s error and the state of the system being observed. When the actual abundance of chains is small, few images will be captured in the sample volume of the SPCS. Since the computer is trained to expect an even proportion of chains, it only considers the likelihood of the features independent of the class distribution (Eq. 1).

**AC and pAC**

The AC and pAC methods were evaluated by selecting 30 of the 41 fully annotated days to estimate the parameters for each correction. For AC, the $\text{tpr}$ and $\text{fpr}$ used in Eq. 5 were computed as the mean rates from the 30 d. For pAC, the 30 samples were used to derive the empirically relationship between the CC relative abundance estimates and $\text{fdr}$ and for. The remaining 11 d were held out for testing (Fig. 5).

Both corrections shifted the classifier output closer to the true population value when the proportion of chains was close to the training distribution. When the target distribution of chains varied radically from that of the training set, both AC quantifiers had similar tendencies: if the true prevalence was
low, both methods shifted the output below the true population. Likewise, when the true prevalence was high, the AC approaches shifted the diatom numbers above the original classifier output (Fig. 5b).

The over- or underestimation behavior is to be expected in cases when the classifier is trained on a uniform distribution. Doing so makes the classifier unbiased to one class or the other, effectively causing the computer to ignore the prior in the BDR (Eq. 1). The classifier then only considers the likelihood of the features, causing it to put most long, skinny ROIs into the chain class. That in turn makes the denominator of Eq. 5 small when considering the relatively rare chain class, yielding a quantifier highly sensitive to errors in the estimate of \( tpr \) or \( fpr \).

The pAC dampened the AC’s sensitivity by relying on \( fdr \) and \( fpr \) estimates from the exponential relationships with the CC prevalence. The output from pAC was typically closer to the true prior than either the CC or AC prevalence estimates (Fig. 5).

**Supervised quantification**

The fully annotated 41-sample test set was used to model and assess the performance of a supervised quantification scheme. In all, 1000 random subsets of \( m \) ROIs were drawn from the ResNet output in each of the 41 test samples. The \( fdr \) from each random subset was computed based on the human labels. Each \( fdr \) was then used to adjust the prevalence according to Eq. 12. The \( q_0 \) from each random subset was then used to compute a mean prevalence and standard error of the supervised quantification procedure in each of the 41 samples. The procedure estimates the sampling distribution of a supervised quantifier at different settings of \( m \) (Fig. 6).

Supervised quantification was assessed for \( m = 10, 20, \) and 50 ROIs. The mean estimated prevalence for all values of \( m \) was very close to the actual prior. As \( m \) increased, the width of the distribution around the mean estimated prevalence got narrower. This intuitively makes sense: as \( m \) approaches the size of the whole sample, the estimate of the \( fdr \), and thus the prevalence, become progressively closer to the true values.

A larger number of human-annotated ROIs from the classifier output will ensure a more accurate estimate of the \( fdr \) and the relative abundance of the class. The number of machinesorted ROIs checked by a human can be adjusted according to the project goals and constraints. We note, however, that estimating the \( fdr \) as few as 10 images per target space or time period and applying the correction of Eq. 12 can yield a remarkable improvement in abundance estimates.

**Performance metric**

We use the mean absolute error, or MAE, to compare the performance of each quantifier in the test set:

\[
\text{MAE} = \frac{1}{k} \sum_{i=1}^{k} |p_{i,\text{true}} - q_{i,0}| \quad (13)
\]

where \( k \) is the number of independent samples, \( p_{i,\text{true}} \) is the true prevalence of chains in each sample as evaluated by the human annotator, and \( q_{i,0} \) is the corresponding estimated prevalence of chains from the quantifier. If the MAE is 0, the quantifier perfectly returned the true prevalence of the population.

AC, pAC, and all supervised quantifiers returned a better MAE than the CC output from the ResNet (Table 2). We produced standard error bounds from the modeled performance of the supervised quantifier to illustrate the possible breadth of prevalence estimates. Even with only 10 human annotated random samples, the upper bound of the MAE is lower than either of the AC approaches.

**Target domain**

The ResNet classifier was run over all unannotated SPCS data in the desired time and size range (Fig. 7). CC was used to estimate the prior distribution of relative chain abundance for each day in the target domain based on the ResNet labels. The
AC, pAC, and supervised quantification corrections were then applied to the CC generated prior distribution. The parameters for the AC and pAC were computed from all 41 samples in the test for application to the target domain. Note that the true prior is unknown over this data set – all curves are estimates of the prevalence as measured by the classifiers and subsequent adjustments.

The AC correction was applied to each day’s worth of data by computing the average fpr and tpr from the independent test set. The AC method shifted the time series downward when the classifier estimated a very low relative abundance of chains and upward when the raw classifier output was high (Fig. 7). The pAC approach used the empirically derived exponential functions to estimate the fdr and for based on the 41 independent test samples (Fig. 4). The pAC output is qualitatively similar to the AC, with less exaggerated upward shifts when the original classifier estimates a high prevalence.

The domain expert labeled 30 random ROIs from the classifier output to directly estimate the fdr on a daily basis for supervised quantification. ROIs not labeled as chains were ignored as the for was assumed to be uniformly low based on the 41 independent test days. The daily estimate of the fdr was used in Eq. 12 to adjust the relative chain abundance. The supervised quantification correction was consistently downward, regardless of the original classifier’s prevalence estimate. The noisy fluctuations in the shift are a consequence of the

**Table 2.** Mean absolute error (MAE) computed for each method applied to the independent test dataset. The MAE is computed by measuring the difference between the true prevalence of diatom chains and the prevalence estimated by the quantifier. MAE is computed based on all 41 test samples for classify and count (CC) and supervised quantification methods based on a subset of 11 samples for adjusted count (AC) and prevalence adjusted count (pAC), as indicated by *. Since CC, AC, and pAC only produced a single estimate of the distribution, standard error was not computed. The data column summarizes which sets of labeled ROIs are used for each method. The amount of data necessary is a proxy for human hours needed; the annotator time increases with each additional set of labeled data.

| Method | MAE  | SE   | Data     |
|--------|------|------|----------|
| CC     | 0.062| —    | Training |
| AC*    | 0.078| —    | Training+test |
| pAC*   | 0.056| —    | Training+test |
| quant10| 0.012| 0.017| Training+10/sample |
| quant20| 0.012| 0.012| Training+20/sample |
| quant50| 0.012| 0.008| Training+50/sample |

**Fig. 6.** Supervised quantification applied to the ResNet output on all 41 samples in the test dataset. The supervised quantifiers were modeled using 1000 random samples of the size indicated by the associated number in the legend. The error bars represent the standard error of the mean prevalence estimate computed from these 1000 random samples. (a) Estimated prevalence as a function of time. (b) Estimated prevalence as a function of the true prior. The black line is a visual aid with a slope of 1.

**Fig. 7.** The output of the ResNet on the target data from August 2017 to November of 2017. The Adjusted Count (AC) and prevalence Adjusted Count (pAC) adjustments introduce a downward or upward shift depending on the estimate prior from the computer. The supervised quantifier detects substantially different signals, notably during the first week of August.
random sampling of the classifier output and are consistent with the error bars in Fig. 6.

The shapes of all the corrections roughly followed that of the original CC estimate with a downward shift (Fig. 7). A notable exception occurred in early August 2017 when the ResNet detected an elevated abundance of chains, estimating chain prevalence at nearly 0.8. The supervised quantifier reduced this estimate down to around 0.2 – a reduction of more than 50%. Both AC methods maintained the elevated abundance estimate of CC.

**Discussion**

Shifts in the abundance or types of organisms in a particular study area can confound attempts at automated classification. In many cases, fully autonomous computer classifiers are not yet a perfect solution to sorting big, biological image datasets. Marine ecosystems exhibit ecological and environmental variability on a range of timescales; species invasions, extinctions, and range shifts are likely. As a consequence, skilled and flexible corrections for dataset shift are essential. Here, we presented both automated and supervised methods to correct the output of an automated classifier.

In our diatom time series, most target time periods experienced a shift in the underlying distributions of diatom chains relative to the training set. By subsampling the noise images for training data, we implicitly taught the classifiers to expect an equal ratio of chains to other objects. In effect, we biased the classifiers to expect more chains than generally are present on a given day. This is reflected in the $fdr$ and $for$ computed in the 41 independent samples – the measured $for$ is generally low while the $fdr$ is often quite high – indicating that the classifier expects a higher fraction of chains than is typically present (Fig. 4).

The bias is most obvious on days when chains were rare enough to not be effectively sampled by the SPCS – when chains are present at low abundances, the number of chains per unit volume is small enough that there is a low likelihood that one drifts into the camera sample volume. Because the computer is trained to expect an even proportion of chains, it overestimated their relative abundance when the actual value was low.

In the binary case, it is possible to empirically choose a threshold to ignore data from the classifier based on its estimates of the population. Such a threshold could be computed using the measured computer performance on independent test data to determine an expected signal-to-noise ratio (SNR). As in acoustics, SNR describes the level at which a signal is indistinguishable from noise. In the diatom chain case, a low SNR would indicate days where the measured number of chains is too low to be separated from other noise images. Such analysis, however, is also prone to missing important dynamics by relying on a static estimate of the classifier performance.

The AC and pAC corrections were in close agreement. This is expected because the pAC is a modification of the static approach. When the estimated prevalence of chains was low, the AC methods yielded a downward shift of the relative chain abundance compared to the raw classifier output. Both methods, however, assume that the only difference between the target and training set is the relative abundances of the classes and that the measured error rates are acceptable for every sample in the target domain. If either condition is broken, the AC methods will not yield appropriate estimates of the distribution of classes.

The AC methods are most effective when the test data are deliberately selected to contain a range of prevalences of the class of interest. pAC is particularly sensitive to the choice of test domain – the empirical relationship must be developed with a representative range to generate reasonable corrections. This can be achieved by choosing to annotate independent periods of time or space with a spread of estimated prevalences via CC (the raw classifier output).

The ResNet and both AC approaches misrepresented the diatom chain prevalence from images captured in the first week of August 2017. CC identified a peak in diatom chain prevalence, with proportions above 60%. The AC corrections made small adjustments but also indicated that chains were present in an elevated abundance. The supervised quantifier suggested that a better estimate of the chain prevalence was around 20%. During this time frame, the human annotator observed an abundance of long, non-chain filaments that appeared to confuse the ResNet (Fig. 2). These objects appear very similar to diatom chains, but do not have distinct cells. This introduction of new, confounding noise is an example of concept drift – when the relationship between the features and the label changes. The long, thin objects were not present in sufficient numbers in the original training set to ensure that the machine properly interpreted such input. The new noise source artificially boosted the estimated relative abundance of chains, potentially distorting subsequent ecological analyses.

Caution is necessary when interpreting the output of an automated classifier on biological image data. Even when corrections are made based on the machine’s performance on independent test data, abundance estimates can still be misleading or incorrect. Supervised quantification is an effective, efficient way to ensure relative abundance estimates are as close to the true ecosystem state as possible.

While supervised quantification entails additional human annotation effort, examining a small number of random images per sampling period can yield substantial improvements in abundance estimates (Fig. 6; Table 2). It is, however, substantially less effort than reviewing all the classifier output by hand. The extra human effort will help improve the relative abundance estimates regardless of possible changes between the machine’s training set and the target data. Note, however, that the adjustment of the supervised quantifier will
be noisy when small numbers of images are observed to estimate the \( fdr \).

For all the proposed corrections, the computational and human costs increase substantially for multiclass classifiers. The simplest way to extend these methods for multiple categories is to treat each class in a binary, one-vs-all manner. Doing so necessitates estimating a \( fdr \) for each class individually. For supervised quantification, this means the number of images per sampling period that must be annotated by a human increases linearly.

In practice, the specific scientific goal will dictate the exact procedure for training, evaluating, and deploying a given automated classification system. For time series data, we are confident that automated classifiers – properly evaluated on independent data exhibiting true distributions of the classes of interest – can be deployed as detectors of population changes without modification. An automated classifier could be used in this way if the measured \( for \) is uniformly low and the behavior of the \( fdr \) is well understood. All peaks, however, should be evaluated by a human labeler to ensure the fidelity of the population estimates from the machine classifier.

To make the process of supervised quantification more efficient one could select the number of images for human observation dynamically: for each sample period, select \( n \) data points without replacement until a desired number of true positives are observed. Such a process might ensure better statistics and limit the number of ROIs viewed by a human. Similarly, an empirical noise floor could be defined to prevent a human annotator from considering days unlikely to contain the organism of interest. The exact parameters of such approaches would need to be calibrated depending on the specific context of the project.

Dataset shift is not restricted to the observation of planktonic ecosystems; the assumption of a stationary relationship between a computer classifier and a population is difficult to satisfy for most marine organisms. Experiments have been conducted to evaluate supervised quantification on coral datasets demonstrating the utility of such methods on spatial variable data types (Beijbom et al. 2015). The problem of dataset shift has also been identified in the bioacoustics literature in the context of using whale calls to monitor population changes (Širović 2016). The issue is pervasive in biological oceanography and should be treated carefully when drawing ecological conclusions from automated data analysis.

This work demonstrates the efficacy and necessity of keeping a human annotator involved in an automated classification scheme for biological oceanographic imagery. On the 41 independent samples of diatom chain data, supervised quantification yielded improved population estimates over the raw classifier output by an average of 80% in terms of MAE. When applied to real target data, the supervised quantifier identified a false spike in diatom chains identified by the other methods. The procedure entails additional human effort, but not substantially more than that necessary to train the machine learning algorithm in the first place. We believe keeping a human in the loop in this manner will ensure the fidelity of population estimates from a machine classifier.

Comments and recommendations

In this work, we randomly sampled our noise class to artificially create a uniform training distribution. This procedure discarded approximately 40,000 non-diatom chain images. In the binary case illustrated in this paper the noise class is broad, and we almost certainly lost information regarding what defines “not a diatom chain.” We chose to sacrifice this information in the interest of highlighting the quantification procedures and making the results easy to interpret.

While many of the ROIs in the non-diatom chain category could be parsed into other groups for a multiclass study, the noise class would still be dominant. Such radically imbalanced training distributions are a common feature of labeled plankton image datasets (González et al. 2017; Luo et al. 2018). How an operator deals with the problem is very much a function of the project goals, human annotator constraints, the desired algorithm, and the final deployment strategy. Here, we offer a few suggestions for future work.

In practice, there are many ways of accounting for imbalanced training sets that might improve the accuracy of the baseline classifier (He and Garcia 2009). If training an ensemble or margin classifier such as the SVM detailed in the supplement, one could use a weighting scheme to account for the minority class (Huang and Du 2005). This training procedure allows one to use all the labeled images while enforcing that the classes are considered equally. The approach is demonstrably effective but requires estimating the necessary weights – a process that becomes more complex as the number of classes increases.

When building a neural network, an operator could consider using data augmentation procedures rather than subsampling (Razavian et al. 2014). To account for the imbalance, the procedure oversamples a minority class and applies random transformations to alter an individual ROIs appearance. This must be done judiciously in order to avoid overfitting the classifier to the particular set of images available for training.

New research in out-of-domain detection approaches suggests that intelligent subsampling of dominant classes could improve deep neural network classification results (Li and Vasconcelos 2020). Instead of randomly subsampling an overpopulated class, a computer is trained to select the most informative examples to preserve as much variability as possible. Such approaches could also improve a computer’s ability to detect outliers and examples of novel classes.

References

Beijbom, O., P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman. 2012. Automated annotation of coral reef survey images. Proceedings of the 2012 IEEE Conference on
Ellen, J. S., C. A. Graff, and M. D. Ohman. 2019. Improving
Ellen, J., H. Li, and M. D. Ohman. 2015. Quantifying Califor-
Duda, R. O., P. E. Hart, and D. G. Stork. 2012. Pattern classi-
Forman, G. 2008. Quantifying counts and costs via classi-
Franks, P. J. 2005. Plankton patchiness, turbulent transport
González-Gil. 2003. Do experts make mistakes? A compari-
Gorsky, G., M. Picheral, and L. Stemmann. 2000. Use of the
Gorsky, G., and others. 2010. Digital zooplankton image anal-
Gorospe, M., A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik. 2019. Automatic plankton quantification using deep features. J. Plankton Res. 41: 449–463. doi: 10.1093/plankt/fbz023
Gorsky, G., and others. 2010. Digital zooplankton image analysis using the ZooScan integrated system. J. Plankton Res. 32: 285–303. doi: 10.1093/plankt/fbp124
Gorsky, G., M. Picheral, and L. stemmann. 2000. Use of the underwater video profiler for the study of aggregate dynamics in the north Mediterranean. Estuar. Coast. Shelf Sci. 50: 121–128. doi: 10.1006.ecss.1999.0539
Grosjean, P., M. Picheral, C. Warembourg, and G. Gorsky. 2004. Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. ICES J. Mar. Sci. 61: 518–525. doi: 10.1016/j.icesjms.2004.03.012
Hand, D. J. 2006. Classifier technology and the illusion of progress. Stat. Sci. 21: 1–14. doi: 10.1214/088342306000000060
Haury, L. R., J. A. McGowan, and P. H. Wiebe. 1978. Patterns and processes in the time-space scales of plankton distributions, p. 277–327. In Spatial pattern in plankton communities. Springer. doi: 10.1007/978-1-4899-2195-6_12
He, H., and E. A. Garcia. 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21: 1263–1284. doi: 10.1109/TKDE.2008.239
He, K., X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 770–778. doi: 10.1109/CVPR.2016.90
Huang, Y.-M., and S.-X. Du. 2005. Weighted support vector machine for classification with uneven training class sizes. 2005 International Conference on Machine Learning and Cybernetics. Proceedings of the 2005 International Conference on Machine Learning and Cybernetics. Vol. 7. 4365–4369.
Katiya, K., R. E. Sherlock, A. D. Sherman, and B. H. Robison. 2017. New technology reveals the role of giant larvaceans in oceanic carbon cycling. Sci. Adv. 3: e1602374. doi: 10.1126/sciadv.1602374

French, G., M. Mackiewicz, M. Fisher, H. Holah, R. Kilburn, N. Campbell, and C. Needle. 2019. Deep neural networks for analysis of fisheries surveillance video and automated monitoring of fish discards. ICES J. Mar. Sci. 77: 1340–1353. doi: 10.1002/icesjms/fsz149
Genovese, C., and L. Wasserman. 2002. Operating characteristics and extensions of the false discovery rate procedure. J. R. Stat. Soc. Series B Stat. Methodol. 64: 499–517. doi: 10.1111/1467-9868.00347
González, P., E. Álvarez, J. Díez, Á. López-Urrutia, and J. J. del Coz. 2017. Validation methods for plankton image classification systems. Limnol. Oceanogr.: Methods 15: 221–237. doi: 10.1002/lom3.10151.
González, P., A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik. 2019. Automatic plankton quantification using deep features. J. Plankton Res. 41: 449–463. doi: 10.1093/plankt/fbz023
Gorsky, G., and others. 2010. Digital zooplankton image analysis using the ZooScan integrated system. J. Plankton Res. 32: 285–303. doi: 10.1093/plankt/fbp124
Gorsky, G., M. Picheral, and L. stemmann. 2000. Use of the underwater video profiler for the study of aggregate dynamics in the north Mediterranean. Estuar. Coast. Shelf Sci. 50: 121–128. doi: 10.1006.ecss.1999.0539
Grosjean, P., M. Picheral, C. Warembourg, and G. Gorsky. 2004. Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. ICES J. Mar. Sci. 61: 518–525. doi: 10.1016/j.icesjms.2004.03.012
Hand, D. J. 2006. Classifier technology and the illusion of progress. Stat. Sci. 21: 1–14. doi: 10.1214/088342306000000060
Haury, L. R., J. A. McGowan, and P. H. Wiebe. 1978. Patterns and processes in the time-space scales of plankton distributions, p. 277–327. In Spatial pattern in plankton communities. Springer. doi: 10.1007/978-1-4899-2195-6_12
He, H., and E. A. Garcia. 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21: 1263–1284. doi: 10.1109/TKDE.2008.239
He, K., X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 770–778. doi: 10.1109/CVPR.2016.90
Huang, Y.-M., and S.-X. Du. 2005. Weighted support vector machine for classification with uneven training class sizes. 2005 International Conference on Machine Learning and Cybernetics. Proceedings of the 2005 International Conference on Machine Learning and Cybernetics. Vol. 7. 4365–4369.
Kenitz, K. M., E. C. Orenstein, P. L. D. Roberts, P. J. S. Franks, J. S. Jaffe, M. L. Carter, and A. D. Barton. 2020. Environmental drivers of population variability in colony-forming marine diatoms. Limnology and Oceanography. 65: 2515–2528. doi:10.1002/lom3.11468.

Li, Y., and N. Vasconcellos. 2020. Background data resampling for outlier-aware classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13218–13227.

Lindgren, J. F., I.-M. Hassellöv, and I. Dahllöf. 2013. Analyzing changes in sediment meiofauna communities using the image analysis software ZoonImage. J. Exp. Mar. Biol. Ecol. 440: 74–80.

Lombard, F., et al. 2019. Globally consistent quantitative observations of planktonic ecosystems. Front. Mar. Sci. 6: 196. doi:10.3389/fmars.2019.00196

López, V., A. Fernández, S. García, V. Palade, and F. Herrera. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inform. Sci. 250: 113–141. doi:10.1016/j.ins.2013.07.007

Luo, J. Y., J.-O. Irisson, B. Graham, C. Guigand, A. Sarafraz, C. Mader, and R. K. Cowen. 2018. Automated plankton image analysis using convolutional neural networks. Limnol. Oceanogr.: Methods 16: 814–827. doi:10.1002/lom3.10285

Margalef, R. 1978. Life-forms of phytoplankton as survival alternatives in an unstable environment. Oceanol. Acta 134: 493–509. https://archimer.ifremer.fr/doc/00123/23403/.

Moreno-Torres, J. G., T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. 2012. A unified view on dataset shift in classification. Pattern Recognit. 45: 521–530. doi:10.1016/j.patcog.2011.06.019.

Mullen, A. D., T. Treibitz, P. L. Roberts, E. L. Kelly, R. Horwitz, J. E. Smith, and J. S. Jaffe. 2016. Underwater microscopy for in situ studies of benthic ecosystems. Nat. Commun. 7: 12093.

Olson, R. J., and H. M. Sosik. 2007. Automated taxonomic classification of plankton sampled with imaging-in-flow cytometry. Limnol. Oceanogr.: Methods 5: 204–216. doi:10.4319/lom.2007.5.204.

Orenstein, E., O. Beijbom, E. E. Peacock, and H. M. Sosik. 2015. Whoi-plankton-a large scale fine grained visual recognition benchmark dataset for plankton classification. arXiv preprint arXiv:1510.00745.

Pan, S. J., and Q. Yang. 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22: 1345–1359. doi:10.1109/TKDE.2009.191

Peacock, E. E., R. J. Olson, and H. M. Sosik. 2014. Parasitic infection of the diatom Guinardia delicatula, a recurrent and ecologically important phenomenon on the New England shelf. Mar. Ecol. Prog. Ser. 503: 1–10. doi:10.3354/meps10784.

Razavian, A. S., H. Azizpour, J. Sullivan, and S. Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 512–519.

Russakovsky, O., and others. 2014. ImageNet large scale visual recognition challenge. arXiv:1409.0575 [cs].

Schröder, S.-M., R. Kiko, and R. Koch. 2020. MorphoCluster: Efficient annotation of plankton images by clustering. Sensors 20: 3060. doi:10.3390/s20113060.

Širović, A. 2016. Variability in the performance of the spectrogram correlation detector for north-East Pacific blue whale calls. Bioacoustics 25: 145–160. doi:10.1080/09524622.2015.1124248.

Solow, A., C. Davis, and Q. Hu. 2001. Estimating the taxonomic composition of a sample when individuals are classified with error. Mar. Ecol. Prog. Ser. 216: 309–311.

Sosik, H. M., and R. J. Olson. 2007. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. Limnol. Oceanogr.: Methods 5: 204–216. doi:10.4319/lom.2007.5.204.

Storey, J. D. 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. Ann. Statist. 31: 2013–2035. doi:10.1214/aos/1074290335

Tasche, D. 2017. Fisher consistency for prior probability shift. J. Mach. Learn. Res. 18: 3338–3369.

Taylor, A. G., R. Goericke, M. R. Landry, K. E. Selph, D. A. Wick, and M. J. Roadman. 2012. Sharp gradients in phytoplankton community structure across a frontal zone in the California Current Ecosystem. J. Plankton Res. 34: 778–789. doi:10.1093/plankt/fbs036

Tok, E. C., D. R. Ratele, C. Briseño-Avenda, M. L. Carter, P. J. S. Franks, J. S. Jaffe, and P. L. D. Roberts. 2020. The Scripps Plankton Camera system: A framework and platform for in situ microscopy. Limnol. Oceanogr.: Methods 18: 681–695. doi:10.1002/lom3.10394.

Orenstein, E. C., and O. Beijbom. 2017. Transfer learning and deep feature extraction for planktonic image data sets. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). 1082–1088.
Acknowledgments

E.C.O., P.L.D.R., J.S.J., and P.J.S.F. were funded by the National Science Foundation BIGDATA Initiative (NSF IIS 15-46351). A.D.B. and K.M.K. were supported by the Simons Foundation.

Conflict of interest

None declared.

Submitted 25 November 2019
Revised 28 August 2020
Accepted 29 September 2020

Associate editor: Ivona Cetinic