Phytoplankton Hotspot Prediction With an Unsupervised Spatial Community Model

Arnold Kalmbach[1], Yogesh Girdhar[2], Heidi M. Sosik[3] and Gregory Dudek[1]

***Abstract*— Many interesting natural phenomena are sparsely distributed and discrete. Locating the hotspots of such sparsely distributed phenomena is often difficult because their density gradient is likely to be very noisy. We present a novel approach to this search problem, where we model the co-occurrence relations between a robot's observations with a Bayesian nonparametric topic model. This approach makes it possible to produce a robust estimate of the spatial distribution of the target, even in the absence of direct target observations. We apply the proposed approach to the problem of finding the spatial locations of the hotspots of a specific phytoplankton taxon in the ocean. We use classified image data from Imaging FlowCytobot (IFCB), which automatically measures individual microscopic cells and colonies of cells. Given these individual taxon-specific observations, we learn a phytoplankton community model that characterizes the co-occurrence relations between taxa. We present experiments with simulated robot missions drawn from real observation data collected during a research cruise traversing the US Atlantic coast. Our results show that the proposed approach outperforms nearest neighbor and k-means based methods for predicting the spatial distribution of hotspots from in-situ observations.**

I. Introduction

This paper addresses the problem of finding spatial density hotspots of a sparsely distributed target phenomenon. We hypothesize that by modeling distributions of co-occurring phenomena, we can predict the presence of the target phenomenon, even in the absence of its direct observation. In particular, we focus on the problem of finding hotspots of target phytoplankton taxa in in-situ observations made by a robotic marine instrument following a fixed survey trajectory.

Phytoplankton are microscopic organisms that form the base of marine food webs. They produce chlorophyll and other pigments to harvest sunlight and fuel photosynthesis, so they can utilize CO_2 and other nutrients to produce O_2 and new organic matter. As such, they play critical roles in global biogeochemical cycles and in structuring marine ecosystems. Marine scientists have long used techniques to measure the amount of chlorophyll in a water sample as a proxy for phytoplankton biomass [2]. These methods are coarse and give only bulk indices, with no information about which species of phytoplankton are present. Phytoplankton are extremely diverse, however, and their community structure plays a major role in shaping ecosystems and their functions. As an extreme example, particular species are known to cause toxic blooms that can threaten wildlife as well as human health.

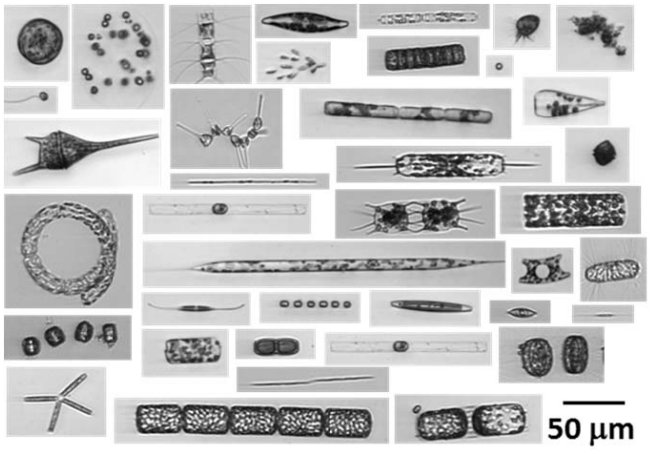


Fig. 1: Example of images captured by the Imaging FlowCytobot (IFCB). These images are classified into 47 classes corresponding to various phytoplankton taxa and other particle types (e.g., detritus) [1]. The proposed topic model automatically discovers community structure from the taxon-specific observational data. We use the model to predict likelihood of observing a target taxon in a given location, without the need for any direct observations of the target.

To meet the gap in observational capability that includes taxonomic resolution, Sosik and Olson have developed the automated, submersible Imaging FlowCytobot (IFCB) [1] and a coupled analysis system [3], [4]. This system can detect and classify phytoplankton automatically in small samples of ocean water collected serially over long periods of time (weeks to years). The images acquired by IFCB have high enough resolution ($\sim$1 µm) that many can be classified to genus or species (Fig. 1). Currently, the IFCB can be routinely moored in the ocean or continuously sample underway on a ship. In addition, prototype deployments have demonstrated its capability to operate on robotic surface vehicles.

In this work we use the detections and detection locations produced by the IFCB as input to the proposed technique, which can enable a marine robot to detect hotpots of sparsely distributed plankton species.

[1]A. Kalmbach and G. Dudek are with the Centre for Intelligent Machines, School of Computer Science, McGill University, 3480 University St. Montréal, QC, Canada {akalmbach,dudek}@cim.mcgill.ca

[2]Y. Girdhar is with the Woods Hole Oceanographic Institution, Applied Ocean Physics and Engineering Department, Woods Hole, MA 02543 yogi@whoi.edu

[3]H.M. Sosik is with the Woods Hole Oceanographic Institution, Biology Department, Woods Hole, MA 02543 hsosik@whoi.edu

Contributions

We present a novel way to robustly estimate the spatial density of a sparsely distributed natural phenomena – phytoplankton taxa – using a probabilistic generative model. The observed distribution of plankton taxa at a location is modelled as a sparse mixture of communities, and the communities are modeled as sparse mixture of plankton taxa. In addition, the model puts smoothness constrains on the spatial distribution of communities. The proposed community model allows us to reason about which plankton taxa we expect to observe together in situations where not all species can be observed.

We demonstrate that our model is able to predict 'hotspots' i.e., locations where a particular taxon obtains high probability of being observed – based on the distribution of the other taxa in a survey. We compare our model's performance in this task to two other strategies: (1) an exhaustive search representing the best any model can be expected to perform if the training and testing data are drawn from the same distribution, but which has a higher computational complexity of than our approach; and (2) a k-means based strategy that has an equivalent complexity to our approach. We show that our model outperforms both other strategies when training and testing data are taken from separate parts of the world, and is competitive when training and testing data are near to one another.

II. Related Work

With recent improvements in in-situ sensing and adaptive sampling algorithms, robots are being used to detect and track many different kinds of natural phenomena underwater. For example, Zhang et al. [5], [6] have demonstrated a technique to autonomously track upwelling fronts in space and time. Ocean upwelling refers to the processes by which nutrient-rich water from the deeper ocean is transported to the surface. Coastal upwelling zones are typically hot-spots for phytoplankton and zooplankton. The authors identify upwelling by detecting vertical temperature gradients.

Much recent work in robotic marine tracking has focused on using visual cues to enumerate species or other phenomena using adaptive sampling techniques [7], [8], [9]. Typical vision systems are much too coarse to provide measurements of phytoplankton populations. The present work makes use of the novel vision capabilities of the IFCB to move towards using similar approaches for phytoplankton tracking applications.

Chlorophyll fluorescence sensors provide a way to detect phytoplankton directly. For example, Das et al. [10] used fluorometers on AUVs and Lagrangian drifters to locate and track phytoplankton patches in the ocean.

Das et al. [11] also developed an approach to predict the abundance of a particular species known to cause harmful algal blooms in the study region. Their objective was to optimize capture of the target species in a small, fixed number of physical samples taken by a robot. Their model is based on a Gaussian Process, with a set of environment variables including fluorescence, temperature, and other chemical properties as inputs, and the results of manual molecular analysis of historical data as training targets. Whereas their method focuses on predicting the abundance of the target species from environmental variables, our method predicts the *relative* abundance of a taxon from the distribution of other taxa. These two perspectives are complementary and both are useful for the problem of automatically choosing the best set of sample locations for extended ex-situ analysis.

Rao et al. [12] proposed the use of a neural network to learn a shared representation over multiple sensor modalities for underwater vehicles (imagery and bathymetry). The learned model is then used to identify information-rich locations given exclusively the bathymetric data. For a small number of classes, this type of multimodal learning framework might capture more of the spatial or temporal complexities of plankton taxon associations. However, as the number of modalities increases this approach is not scalable and therefore it is not suitable for modelling the numerous plankton taxa we consider from this dataset.

Topic modeling [13] offers a natural way to represent highly multimodal data such as the spatio-temporal distribution of plankton taxa. Topic models specify a generative model of the data, where each set of discrete observations is modeled as a mixture of topics (plankton communities) and, in turn, each topic or community is modeled as a mixture of plankton taxa. In topic models, Dirichlet distribution or Dirichlet process [14] priors can be used to control the sparseness of the taxonomic distribution representing a community, and the sparseness of the community distribution at a given location. Girdhar et al. [15], [16] extended the standard topic model to account for spatial and temporal correlation of observations. The plankton community model we propose here is based on the Bayesian nonparametric spatio-temporal topic model (BNP-ROST) [16].

III. Approach

We are interested in identifying areas of the ocean where we are most likely to observe a particular class of plankton. Let w be a plankton observation, such that $w \in [1,V]$, where V is the total number of known plankton taxa, and let x be the spatio-temporal coordinates of this observation, i.e. the vector *[Time since cruise start, Eastings, Northings]*, which we refer to simply as the *location*. The goal is then to estimate the distribution $P(w = v|x,W,X)$, i.e., the distribution of classes v at location x, given all previous observations W and their locations X. We define a hotspot as the set of locations where the probability of observing a class exceeds a class-specific threshold.

Given the high dimensionality of the distribution of classes, we propose the following factorization to approximate the target distribution.

$$P(w = v|x,W,X) = \sum_k P(w = v|z = k)P(z = k|x). \quad (1)$$

Here z is a latent variable, which essentially denotes a plankton community, and the distribution $P(w = v|z = k)$ models the likelihood that the an observation is of taxon

v given that it was drawn from community k. The distribution $P(z=k|x)$ models the spatio-temporal distribution of community k.

We model $P(w|z)$ with a Dirichlet prior. This assumption ensures that our model assigns higher probability to communities represented by sparse taxon distributions. The posterior distribution can be expressed in terms of observation counts:

$$\Phi = P(w=v|z=k,\beta) = \frac{N_k^{(v)}+\beta}{N_k^{(\cdot)}+V\beta}, \tag{2}$$

where β is a symmetric Dirichlet distribution parameter. $\Phi = \{\phi_{v,k}\}$ is a $V \times K$ matrix that represents the community model.

We assume that the number of plankton communities is unknown and use a variant of the Chinese restaurant process (CRP) [17], [14] to model the prior for the distribution of communities at a given location. The posterior community distribution at location x is given by:

$$\Theta = P(z=k|x,\alpha,\gamma) \propto \begin{cases} N_{g(x)}^k + \alpha & k \in Z \\ \gamma & k \notin Z \end{cases} \tag{3}$$

Here Z is the current set of all known plankton communities, $N_{g(x)}^k$ is the number of times we have observed a member of community k in the spatio-temporal neighborhood of location x, and α, γ are model hyperparameters. Hence, with high probability proportional to $N_{g(x)}^k + \alpha$, the observation belongs to community k that is common around location x, and with a small probability proportional to γ, the observation belongs to a new, un-modeled community.

We divide the world into spatio-temporal cells such that the cell that contains location x is denoted by $c(x)$, and then define $g(x)$ to be the set of cells in the Von Neumann neighborhood of $c(x)$. The spatio-temporal distribution of communities can then be modeled by $\Theta = \{\theta_{c,k}\}$, which is a $C \times K$ matrix, where C is the total number of spatial cells in the world that have observations.

When the robot explores a new location where the number of observations of the target taxon is zero or too small to be statistically significant, we hypothesize that the plankton topic model can be used to compute a robust estimate of the likelihood of observing the target taxon on the basis of its association with other taxa.

To accomplish this, first we learn the plankton community topic model from previously visited locations. This data could come either from previous missions or locations visited earlier on the same mission. We then compute the maximum likelihood topic assignments for the observations in the neighborhood of the target location. Finally, given the topic assignments and the original model, we can compute the maximum likelihood distribution for all classes, including an unobserved taxon in the target location:

$$P(w=v|x,\Phi) = \sum_k \theta^*_{c(x),k}\phi_{v,k} \tag{4}$$

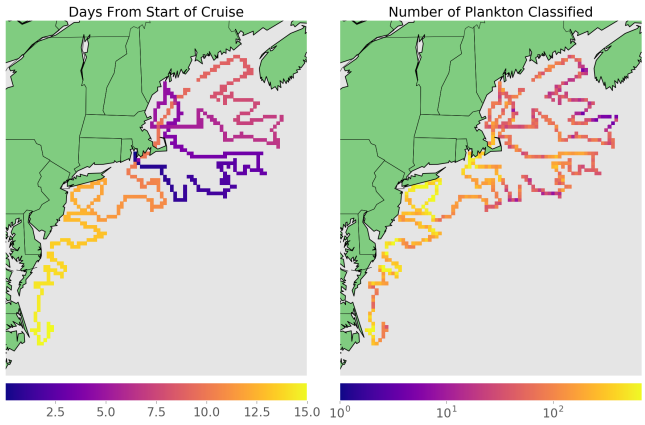


Fig. 2: Summary of data recorded during the Pisces 14-05 cruise. Left, color shows progress in time. Right, color shows the number of plankton observed at each sample location.

Here θ^* is the maximum likelihood topic distribution in the neighborhood of target location x.

To learn the community model, we use an online Gibbs sampler [18], which equally divides the computational resources between computing the posterior topic distribution of the most recent observation, and updating the topic labels and the topic model over the previous observations.

IV. Experiment

To evaluate the hypothesis that the proposed plankton community model can be used to predict hotspots of a target class, we present experiments with simulated missions, drawn from real data, focusing on the worst case scenario where no observations of the target class have been made.

We validate our approach with IFCB classification results from NOAA's Fall 2014 EcoMon Survey aboard the Research Vessel Pisces (Cruise PC 14-05). The IFCB was configured to automatically sample from underway flowing surface seawater (5 ml approximately every 20 min) during the period 4-19 November 2014. The classification system generated over 140,000 individual phytoplankton observations from these water samples. Classification results comprise a dataset with 47 taxa at 852 locations spanning the US Atlantic coast from North Carolina to Maine (See Fig. 2).[1]

We divide the sample locations into equal-sized parts, representing the training and test phases of the simulated mission. The counts of all 47 taxa were kept in the training set and used to learn the topic model. For the test set, we held out each of the 8 most-frequently observed phytoplankton taxa one at a time. We define the hotspots of a taxon to be the top 50 sample locations in the test data, where the relative abundance of the taxon to all other taxa was highest. The 8 tested taxa make up just over 81% of all the observations in

[1]The dataset is available online at `http://ifcb-data.whoi.edu/IFCB102_PiscesNov2014`

the dataset. The most common taxa are miscellaneous centric diatom chains ("mix_elongated"), mixed species of pennate diatoms, *Thalassiosira* spp., *Guinardia delicatula, Guinardia striata, Dictyocha* spp., *Ephemera* spp., and *Phaeocystis* spp.

With the topic model learned from the training data we compute the maximum likelihood topic assignments for the test data. To simulate the case when there are no observations of the target taxon $v^\star$, we use $\Phi^{(\neg v^\star)}$ instead:

$$\phi_{k,v}^{(\neg v^\star)} \triangleq p\left(w_i = v | z_i = k, \mathbf{z}_{\neg \mathbf{v}^\star}\right) = \frac{N_{-i,k}^{(v)} + \beta}{N_{-i,k}^{(\neg v^\star)} + (V-1)\,\beta} \quad (5)$$

The maximum likelihood taxon distribution for the test data is given by $\Theta\Phi^{(\neg v^\star)}$, but since we do not update the topics given the new data, we can instead estimate $P(w = v^\star|c) = \sum_k \theta_{c,k}\phi_{k,v^\star}$. While our method accounts for the sparsity of taxon distributions, this dataset also features sparsity in terms of the locations of observations. To address this separate issue, we resort to a 2D spatial median filter. Finally, we apply a threshold to identify the hotspot locations.

We compared our method to an exhaustive search strategy and a k-means search strategy. For each sample in the test set, exhaustive search estimates the probability of observing $v^\star$ by looking up the sample in the training set with the most similar distribution to the observed data. This represents the strategy which makes the most use of all the data available for every test sample, at the cost of a linear computational complexity in the number of sample locations in the dataset. In the k-means strategy, we fix a constant test-time complexity by reducing the search space to the K centroids returned by a standard k-means clustering implementation. These centroids are defined such that if each class distribution in the training set were replaced by the nearest of the K centroids, the sum of squared error is approximately minimized, however it does not take into account the sparsity or spatial smoothness of the underlying distributions.

We carried out experiments for two different train/test regimes. First, we used every second sample location for training (see Fig. 3, column 1). This regime simulates a mission where the classifier frequently fails to identify examples of a class, for instance because its acceptance threshold was poorly tuned. Because nearby sample locations tend to have similar distributions, this regime tests the ability of a model to interpolate over small distances. Second, we used the first half of the sample locations as training (Fig. 4, column 1), and the second half for testing. This latter case simulates a mission where the capabilities of the classifier have changed from the first half to the second half. It tests the ability of a model to predict in a new location that is not likely to have any spatially linked correlation with the training data.

V. Results

We ran our model for a range of choices of the hyperparameters $\alpha \in \{0.001, 0.01, 0.1, 0.5, 1\}$, $\beta \in \{0.001, 0.01, 0.1, 0.5, 1\}$, and $\gamma \in \{10^{-6}, 10^{-5}, 10^{-4}\}$ with each of the top 8 taxa held out of the testing data and for both training regimes. We also ran the exhaustive search and k-means strategies for each. The strategies each produce an estimate for $P(w = v^\star|c)$, which we then smooth with a median filter with size parameter σ. For a scalar threshold τ, we predict that cell c is a hotspot if $\Pi_\sigma(P(w = v^\star|c)) > \tau$, where Π_σ is the median function over a square region with side length σ.

To evaluate our results we compare the held-out locations in the test set (Fig. 3 and 4, column 2) to predictions from each of the proposed strategies (Fig. 3 and 4, our model, column 3; exhaustive nearest neighbour search, column 4; and k-means search, column 5). The input to the models is illustrated with the observed values of the held-out class at the training locations (Fig. 3 and 4, column 1). Our findings show that the prediction problem is relatively straightforward for the interleaved experiment (Fig. 3). In contrast, the problem is much more difficult when training and testing locations are in different parts of the world. (Fig. 4). Despite this, for three (Fig. 3 and 4, rows 1, 2, 4) of the four target classes shown here, the spatial location of maxima of our model's predictions are consistently near the maxima in the target distributions.

Varying τ for each strategy and parameter setting we can count the true positive, true negative, false positive, and false negative hotspot predictions compared to the top 50 examples in the held-out data. These counts give the precision and recall for each parameter choice, for each $v^\star$. We also accumulate these counts across all $v^\star$ to compute the overall precision and recall for each choice of parameters. We assign each set of parameters a score given by the area under its aggregated precision-recall curve and select the parameter set with the maximum score for further comparisons. For the interleaved experiment, best performance was achieved with $\alpha = 0.1, \beta = 0.1, \gamma = 10^{-5}, \sigma = 25km$ and for the split experiment, $\alpha = 0.1, \beta = 1.0, \gamma = 10^{-5}, \sigma = 35km$. We chose the number of centroids for the k-means strategy to be the same as the number of topics in the best performing topic model, $K = 9$ for the interleaved experiment, and $K = 6$ for the split experiment.

We compare the aggregated and individual class precision-recall curves for the best parameters for each strategy (Fig. 5). Note that precision refers to the ratio of the number of correctly predicted hotspots to the total number of predicted hotspots, and that recall refers to the ratio of correctly predicted hotspots to the total number of real hotspots. An ideal algorithm will have precision of 1 and recall of 1. From the aggregated precision-recall curves, we find that our model significantly outperforms the exhaustive nearest-neighbor and the k-means strategies on the split-samples regime, especially for low recall requirements. This indicates that the top few predictions of our model were more likely to be true hotspots than those of the other strategies. The exhaustive nearest-neighbor strategy barely performs better than random guessing on the split regime, yet it performs extremely well on the interleaved regime. This result is expected as the exhaustive strategy does not reason at all about the underlying association between plankton

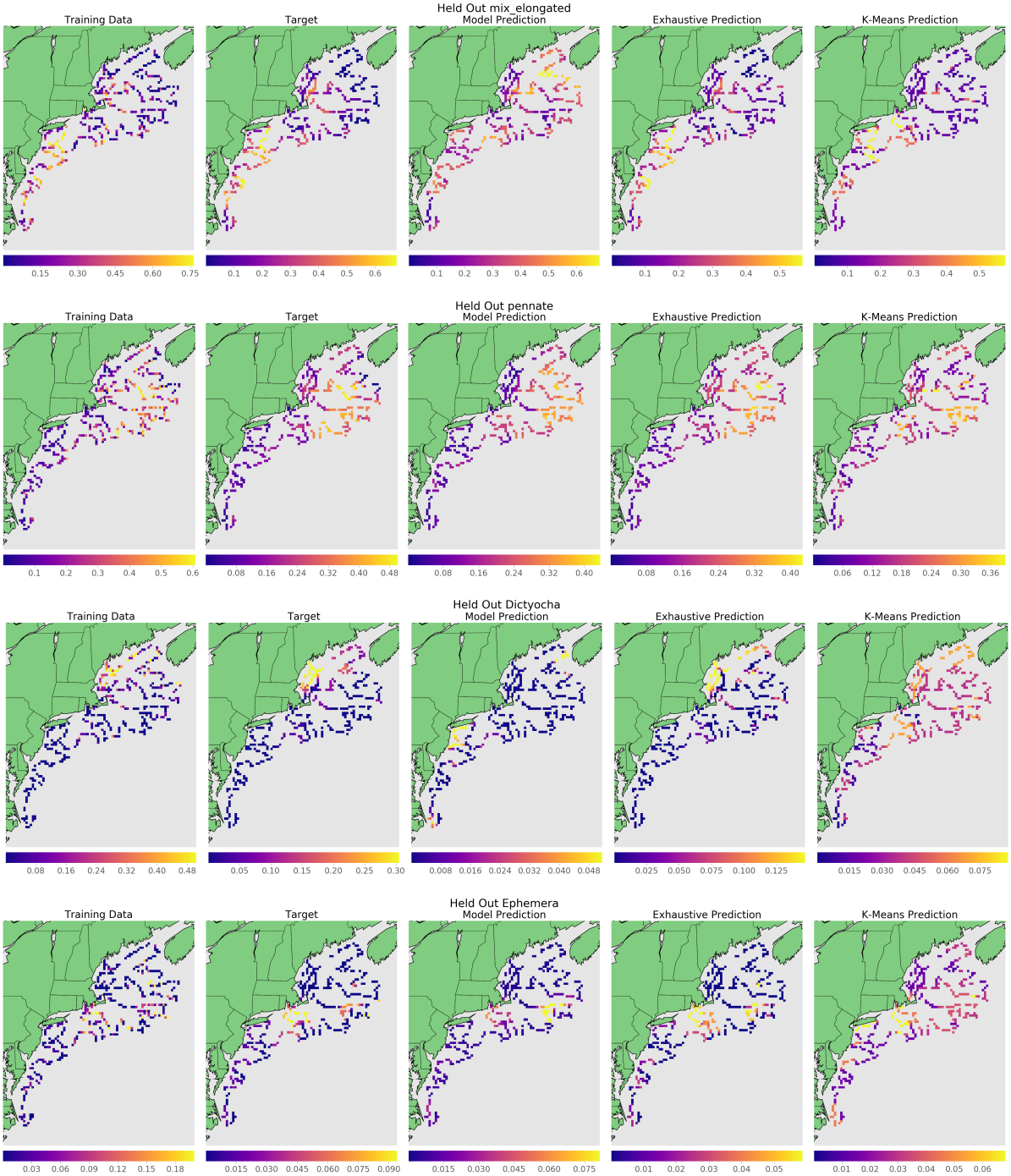


Fig. 3: Spatial distribution for four target classes (rows) in interleaved training/testing samples. The columns correspond to training data (col. 1), held-out target locations (col. 2), and the three models under evaluation (col. 3-5). We find close correspondence between the proposed model and the target data, but exhaustive nearest-neighbor approach has the most similar distribution to held-out target locations. This is because the distribution of plankton is correlated with its spatial neighbors, and hence simple interpolation of the training data is likely to give an accurate plankton distribution at the held-out locations.

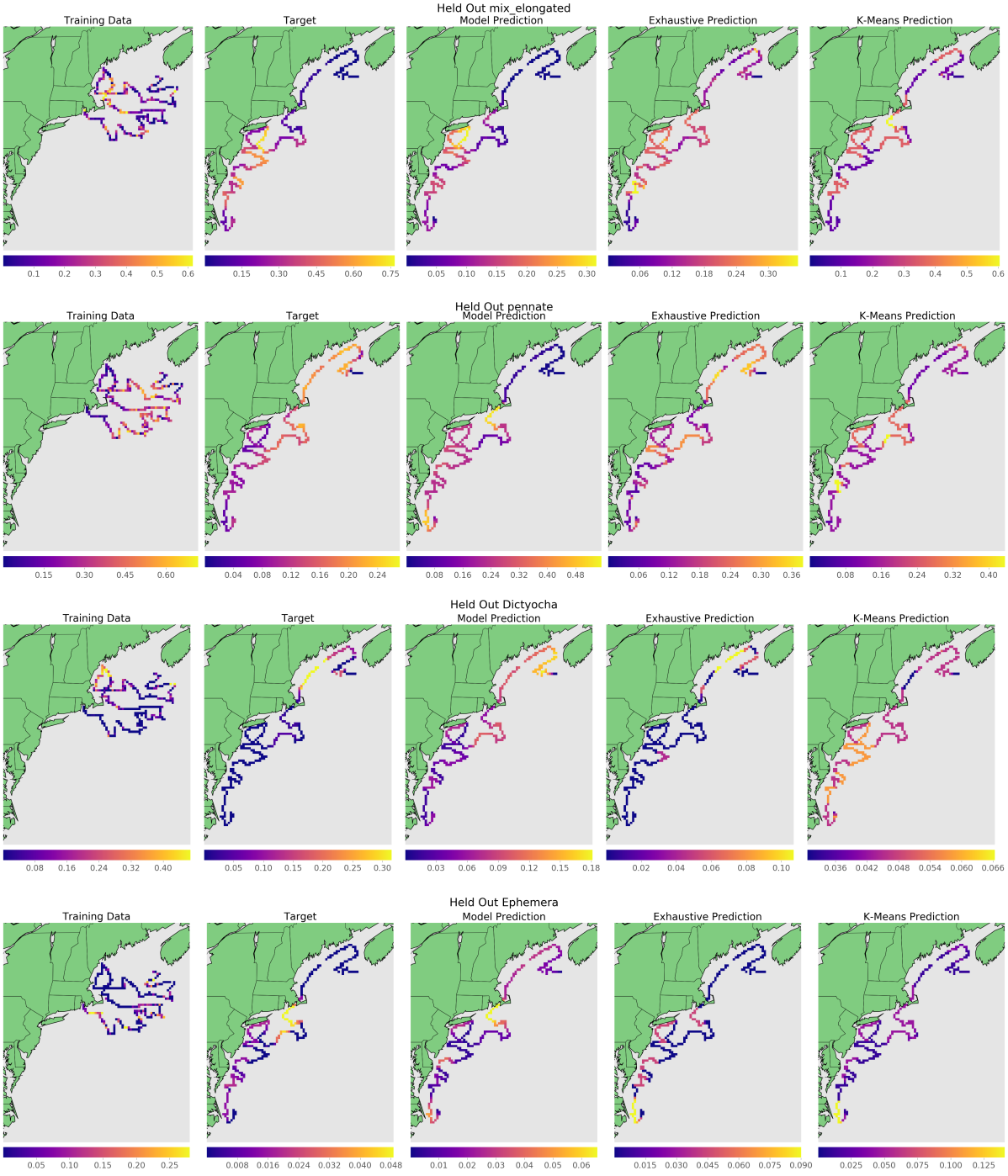
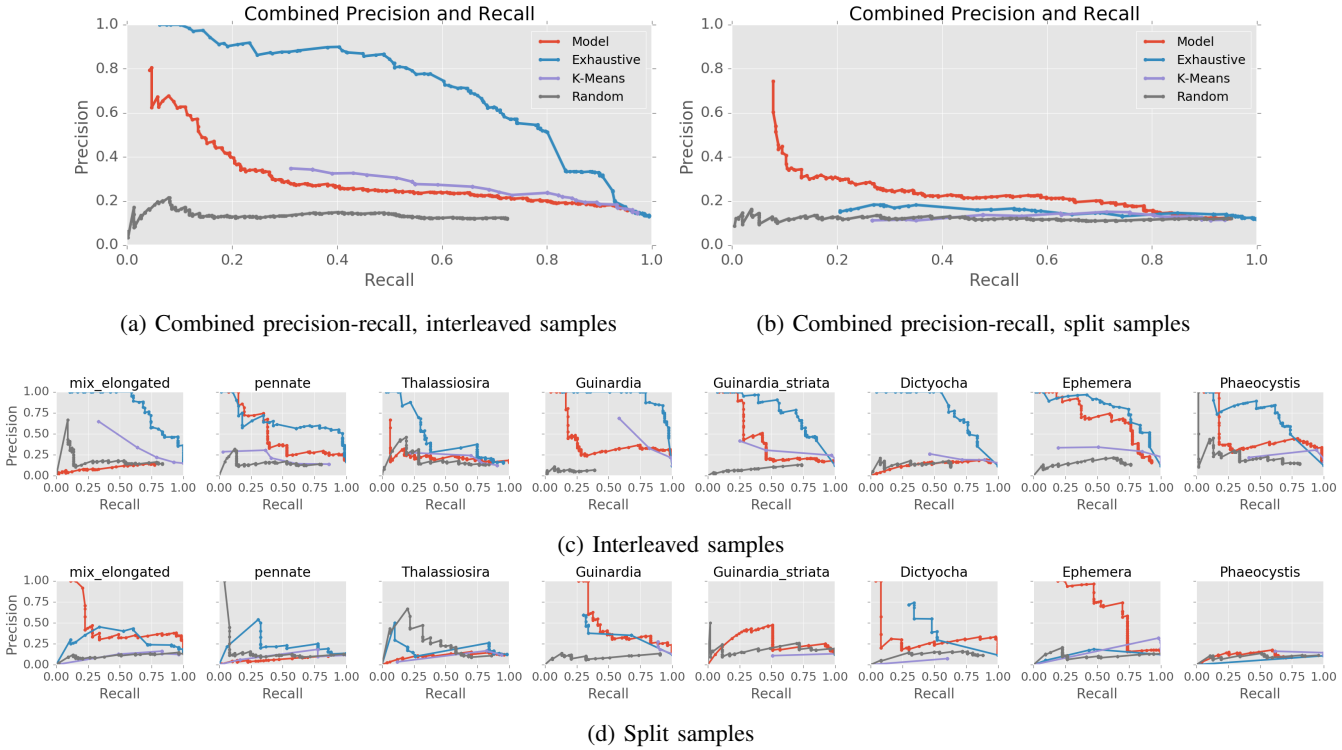


Fig. 4: Spatial distribution for four target classes (rows) in split training/testing samples. The columns correspond to training data (col. 1), held-out target locations (col. 2) , and the three models under evaluation (col. 3-5). The proposed plankton topic model provides predictions that agree better with the held-out observations than do the simpler k-means based plankton community model or the exhaustive nearest neighbor search.



(a) Combined precision-recall, interleaved samples

(b) Combined precision-recall, split samples

(c) Interleaved samples

(d) Split samples

Fig. 5: Plots showing precision-recall curves that indicate the performance of the proposed technique at predicting hotspots of the target plankton species. (a,c) When the training data is interleaved with the target locations, the exhaustive nearest-neighbor has the best average performance. (b,d) The proposed model has the best average performance in cases where observations from nearby locations are not available (split samples), and hence a robust plankton community model is required.

types. Instead, it depends on having observed a training point whose distribution is similar to every test point. In contrast, our model performs nearly as well on the split regime as the interleaved regime.

Our model also outperforms the k-means strategy on the split-samples regime. Note that the k-means strategy is exactly equivalent to the exhaustive search strategy in the limit where K is the number of training points. Both these strategies rely on a distance metric over the class distributions. The high dimensionality of the distributions acts to the detriment of the distance metric. As the dimensionality of a space increases, the discriminating power of distance metrics within that space decreases. The amount of data needed to find meaningful clusters grows exponentially with the number of dimensions, a phenomenon sometimes called *the curse of dimensionality*. As a result, the two search-based strategies perform well when test points are very near to training points in taxon distribution space, but when test points are further away, a distance metric is less informative and performance is negatively impacted. Our model mitigates this problem with additional constraints in the form of a hierarchical generative model, the CRP prior on the spatial distribution of communities, and sparse Dirichlet priors on the plankton class distribution that describes each community.

VI. Discussion

Our ongoing efforts are focused on using this work to improve on autonomous sampling techniques for sparsely distributed class counts, including phytoplankton taxa. In this work we have considered the case where the classifier fails to make any predictions whatsoever about some class, however the results are also relevant to scenarios where unexpectedly low or high numbers of a class are observed. Recall that IFCB samples only 5 mL of water at a time, yet researchers would like to characterize the plankton distribution in a wide area of ocean. As a result, the measured distributions in individual cells are extremely noisy. This was not taken into account while collecting the present dataset, and as a result we need to use spatial smoothing on the order of a 15km radius to achieve meaningful predictions. An interesting future direction for this work is to compare our model's prediction to real-time measurements, and use this comparison to decide whether more data is needed. We are currently developing a variant of IFCB that can be deployed on autonomous surface vehicles such as the WHOI JetYak which will allow dynamic planning with respect to the plankton observations and our model.

We also plan to explore further models which build on the one proposed in this work. A natural research direction is to develop models of the relationships between taxa and environment variables. However, the high-dimensionality,

sparsity of observations, and noisiness of the distribution of individual taxa make learning these relationships difficult. Our initial explorations have suggested that the relationships between environment variables and communities are easier to characterize with simple models than those with individual taxa. We are particularly interested in such models which also incorporate temporal aspects, as they could enable learning causal relationships involved in phytoplankton lifecycles and the changing ecosystems. Finally, we plan to explore deeper generative models of the observations, which we expect will discover more complex community structures.

VII. Conclusion

We have presented a novel technique for finding hotspots of discrete targets that are sparsely distributed in the world. The proposed method utilizes a probabilistic generative model to describe spatial co-occurrence relationships between the target and other kinds of observations. Our technique uses these relationships to estimate the target's spatial distribution in locations where robust measurements are not available. We apply our approach to the problem of finding hotspots of phytoplankton taxa in observations made by a robotic marine instrument.

The proposed technique utilizes a mixture model with spatial smoothness and sparsity constraints on phytoplankton distributions to enable accurate predictions, even when the observed plankton distribution is very different from training data. We validated our approach with real data collected on a two week fixed-trajectory survey mission. Results from experiments show that our model produces a better community representation that can more accurately predict hotspot locations than either exhaustive nearest-neighbour search or a k-means based plankton community model.

Acknowledgement

This work was supported in part by awards to YG from NOAA through its Cooperative Institute for the North Atlantic Region (CINAR) program, and from WHOI; and to HMS from NASA's Ocean Biology and Biogeochemistry Program, and from NOAA through CINAR. We are indebted to Emily Brownlee for expert assistance with IFCB data collection and Joe Futrelle for facilitating IFCB data access and analysis workflows. We also thank the captain and crew of the Research Vessel Pisces and scientists from NOAA's Northeast Fisheries Science Center for enabling our participation in EcoMon surveys. We gratefully acknowledge the support via grant to GD of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] R. J. Olson and H. M. Sosik, "A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot," *Limnology and Oceanography: Methods*, vol. 5, no. 6, pp. 195–203, jun 2007. [Online]. Available: http://doi.wiley.com/10.4319/lom.2007.5.195

[2] C. J. Lorenzen, "A method for the continuous measurement of in vivo chlorophyll concentration," *Deep-Sea Research*, vol. 13, pp. 223–227, 1966.

[3] H. M. Sosik and R. J. Olson, "Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry," *Limnology and Oceanography: Methods*, vol. 5, no. 6, pp. 204–216, jun 2007. [Online]. Available: http://doi.wiley.com/10.4319/lom.2007.5.204

[4] H. M. Sosik, J. Futrelle, E. F. Brownlee, E. Peacock, T. Crockford, and R. J. Olson, "IFCB-Analysis software system." [Online]. Available: https://doi.org/10.5281/zenodo.153978

[5] Y. Zhang, J. G. Bellingham, J. P. Ryan, B. Kieft, and M. J. Stanway, "Autonomous Four-Dimensional Mapping and Tracking of a Coastal Upwelling Front by an Autonomous Underwater Vehicle," *Journal of Field Robotics*, vol. 33, no. 1, pp. 67–81, jan 2016. [Online]. Available: http://doi.wiley.com/10.1002/rob.21617

[6] Y. Zhang, J. P. Ryan, J. G. Bellingham, J. B. J. Harvey, and R. S. McEwen, "Autonomous detection and sampling of water types and fronts in a coastal upwelling system by an autonomous underwater vehicle," *Limnology and Oceanography: Methods*, vol. 10, no. 11, pp. 934–951, nov 2012. [Online]. Available: http://doi.wiley.com/10.4319/lom.2012.10.934

[7] F. Shkurti, A. Xu, M. Meghjani, J. C. G. Higuera, Y. Girdhar, P. Gigure, B. B. Dey, J. Li, A. Kalmbach, C. Prahacs, K. Turgeon, I. Rekleitis, and G. Dudek, "Multi-domain monitoring of marine environments using a heterogeneous robot team," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, pp. 1747–1753.

[8] S. Manjanna, N. Kakodkar, M. Meghjani, and G. Dudek, "Efficient terrain driven coral coverage using gaussian processes for mosaic synthesis," in *2016 13th Conference on Computer and Robot Vision (CRV)*, June 2016, pp. 448–455.

[9] O. Pizarro, A. Friedman, M. Bryson, S. B. Williams, and J. Madin, "A simple, fast, and repeatable survey method for underwater visual 3d benthic mapping and monitoring," *Ecology and Evolution*, pp. n/a–n/a, 2017. [Online]. Available: http://dx.doi.org/10.1002/ece3.2701

[10] J. Das, F. Py, T. Maughan, T. O'Reilly, M. Messie, J. Ryan, G. S. Sukhatme, and K. Rajan, "Coordinated sampling of dynamic oceanographic features with underwater vehicles and drifters," *The International Journal of Robotics Research*, vol. 31, no. 5, pp. 626–646, apr 2012. [Online]. Available: http://ijr.sagepub.com/cgi/doi/10.1177/0278364912440736

[11] J. Das, F. Py, J. B. J. Harvey, J. P. Ryan, A. Gellene, R. Graham, D. A. Caron, K. Rajan, and G. S. Sukhatme, "Data-driven robotic sampling for marine ecosystem monitoring," *The International Journal of Robotics Research*, vol. 34, no. 12, pp. 1435–1452, oct 2015. [Online]. Available: http://ijr.sagepub.com/cgi/doi/10.1177/0278364915587723

[12] D. Rao, A. Bender, S. B. Williams, and O. Pizarro, "Multimodal information-theoretic measures for autonomous exploration," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2016, pp. 4230–4237. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7487618

[13] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, p. 77, apr 2012. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2133806.2133826

[14] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian Nonparametric Models with Applications," *Bayesian nonparametrics*, pp. 158—-207, 2010.

[15] Y. Girdhar, P. Giguere, and G. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," *The International Journal of Robotics Research*, vol. 33, no. 4, pp. 645–657, nov 2013. [Online]. Available: http://ijr.sagepub.com/cgi/doi/10.1177/0278364913507325

[16] Y. Girdhar and H. Singh, "Unsupervised Lifelong Learning for a Curious Underwater Exploration Robot," in *ICRA 2016 Workshop: AI for Long-term Autonomy*, 2016, p. 4.

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, dec 2006. [Online]. Available: http://pubs.amstat.org/doi/abs/10.1198/016214506000000302

[18] Y. Girdhar and G. Dudek, "Gibbs Sampling Strategies for Semantic Perception of Streaming Video Data," *ArXiv e-prints*, p. 7, 2015. [Online]. Available: http://arxiv.org/abs/1509.03242