Mapping Uncharted Waters: Exploratory Analysis, Visualization, and Clustering of Oceanographic Data

Joshua M. Lewis
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093-0515
josh@cogsci.ucsd.edu

Pincelli M. Hull
Scripps Institution of Oceanography
University of California San Diego
San Diego, CA 92093-0208
phull@ucsd.edu

Kilian Q. Weinberger
Yahoo! Research
2821 Mission College Blvd
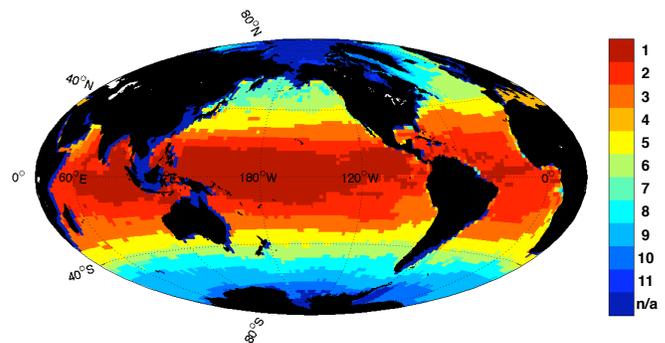Santa Clara, CA 95054
kilian@yahoo-inc.com

Lawrence K. Saul
CSE Department
University of California, San Diego
San Diego, CA 92093-0404
saul@cs.ucsd.edu

Abstract

In this paper we describe an interdisciplinary collaboration between researchers in machine learning and oceanography. The collaboration was formed to study the problem of open ocean biome classification. Biomes are regions on Earth with similar climate (e.g., temperature and rainfall) and vegetation structure (e.g., grasslands, coniferous forests, and deserts). To discover biomes in the open ocean, we apply leading methods in high dimensional data analysis, clustering, and visualization to oceanographic measurements culled from multiple existing databases. We compare traditional approaches, such as k-means clustering and principal component analysis, to newer approaches such as Isomap and maximum variance unfolding. Our work provides the first quantitative classification of open ocean biomes from an automated statistical analysis of multivariate data. It also provides a valuable case study in the use (and misuse) of recently developed algorithms for high dimensional data analysis.

Figure 1. The global projection of oceanic biomes produced by k-means clustering ($k = 11$) after data analysis by maximum variance unfolding; see section 4 for details. The biomes are colored according to cluster membership and sorted by the first principal component of the cluster centroids.

1 Introduction

It is now widely recognized that advances in machine learning are ushering in a new era of computational and experimental science. This era will be characterized by increasingly powerful, automated, and large-scale methods in data analysis and visualization (Mjolsness & DeCoste, 2001). The full potential of machine learning will only be realized, however, by making the field's leading algorithms available and accessible to researchers in other data-driven and computation-intensive areas of science.

As a successful example of this practice, in this paper we describe an interdisciplinary collaboration between researchers in machine learning and oceanography. The collaboration was formed to study the problem of open ocean biome classification. Biomes are regions on Earth with sim-

ilar climate (e.g., temperature and rainfall) and vegetation structure (e.g., grasslands, coniferous forests and deserts). Ecologists are interested in biomes as units of classification because they correspond broadly with the community structure and dynamics of the organisms that live there. In essence, dividing the world into meaningful biomes provides a rough estimate of the number, location, and similarity of the different ecosystems on Earth. Moreover, this problem is ripe for automated analysis because oceanic biomes are large relative to the resolution of available data. In particular, the available data makes it especially feasible to classify surface waters of the open ocean—namely, that part of the ocean that is offshore and restricted to the upper 200 meters of water.

In this paper, we provide the first quantitative classification of open ocean biomes based on a fully automated, statistical analysis of climatological and primary production parameters. To discover these biomes, we applied leading methods in high dimensional data analysis, clustering, and visualization to oceanographic measurements culled from multiple databases and previous ecological studies. Fig. 1 provides an overall visualization of our results.

While our results are interesting in their own right, they also showcase the real-world applicability of several recently developed algorithms for high dimensional data analysis. In addition to standard procedures such as k-means clustering and principal component analysis, we also experimented with spectral clustering (Ng et al., 2002), Isomap (Tenenbaum et al., 2000), and maximum variance unfolding (Weinberger et al., 2004; Sun et al., 2006). Notably, our most interesting results were obtained not from the original implementations of these algorithms, but by adapting the algorithms in various ways and incorporating advances suggested by follow-up work (de Silva & Tenenbaum, 2003; Weinberger et al., 2007).

Our results also illuminate certain stark differences between these algorithms. For the most part, recently developed algorithms for "manifold learning" yield fairly similar results on the carefully controlled data sets used to benchmark algorithms in this area. The basic assumptions behind these algorithms also appear to be satisfied by most real-world data sets on which they have been tested. The ocean data set in this paper, however, appears to violate the basic assumption of (at least) one popular algorithm for manifold learning, leading to fairly divergent results in a real-world application of interest. We highlight the reasons for this divergence in the discussion of our methods and experimental results. Our work on this particular application thus serves a broader purpose than the application itself. More generally, it provides a valuable case study for researchers in the area of manifold learning.

Our paper is organized as follows. In section 2, we describe the problem of open ocean biome classification in more detail, as well as the data set compiled for this task. In section 3, we briefly survey our methods for data analysis. In section 4, we analyze the results from different methods and compare the different aspects of biome structure that they reveal. Finally, in section 5, we conclude with general lessons from our work.

2 Open ocean biome classification

The endeavor to classify open ocean biomes is not a new one. Oceanic biomes have been identified for more than 100 years based on species range distributions (Giesbrecht, 1892; McGowan, 1971), with the explicit acknowledgment that some set of habitat characteristics were likely driving the patterns. However, it was only in the last decade that oceanic biomes were first determined from physical, chemical and biological habitat characteristics (Longhurst, 1998). This advance was made possible by the recent accumulation of multiple long-term global databases of shipboard measurements and satellite images. This manual classification of oceanic biomes and provinces was based on a visual assessment of a decade of ocean color images corroborated by numerous physical, chemical, and biotic factors. While this biome classification has proven useful, the methodology behind it has limitations. It is difficult to determine the relative importance of biome separations or to assess the similarity of geographically separate regions. Furthermore, visual biome assessments are not readily adapted to oceanographic analyses on different temporal or spatial scales and they offer no systematic method for evaluating the effect of additional variables on biome assignment. In spite of these limitations, there has yet to be a quantitative, multivariate assessment of the identity, location, and similarity of ocean biomes. Automated methods have the potential to (i) provide further insight into the structure, distribution, and interrelationships of surface ocean ecosystems; (ii) enable the identification of biomes on a range of temporal scales; and (iii) support quantitative intercomparisons of open ocean biomes.

In this study we characterize open ocean biomes based on both the long-term annual mean and the average annual range of the following seven global ocean characteristics: temperature (°C), salinity, 10 to 200m density difference, photosynthetically active radiation (PAR, Einsteins/m^2/day), phosphate (μmol/L), dissolved oxygen (ml/L), and net primary productivity (NPP, mg C/m^2/day). These measurements were considered at a 2° by 2° resolution (latitude by longitude), resulting in a $d = 14$ dimensional data set of $n = 9105$ geographic locations. All measurements except PAR and NPP were obtained as objectively analyzed annual and monthly climatologies in a 1° by 1° resolution from the World Ocean Atlas 2005 (WOA05) (Locarnini et al., 2006; Antonov et al., 2006;

Garcia et al., 2006a; Garcia et al., 2006b), with 10 to 200m density difference calculated from WOA05 temperature and salinity climatologies. Monthly NPP and PAR from 1998-2005 at $1/6°$ by $1/6°$ resolution were obtained from the Ocean Productivity website; NPP was estimated based on SeaWiFs ocean color images and the Vertically Generalized Production Model (VGPM) (Behrenfeld & Falkowski, 1997). For all variables, annual range (as a measure of intra-annual variability) was calculated as the range in average monthly conditions for a given $1°$ by $1°$ location. In the final data preparation, we excluded all neritic (i.e. coastal) zones using a 200-meter depth mask from WOA05 and averaged all high resolution variables to a $2°$ resolution. We normalized all oceanographic measures to have zero mean and unit standard deviation. The normalization was applied to attach equal importance to each measurement.

Ocean biomes were identified by searching for clusters in the normalized data described above. Note that by design, latitude and longitude coordinates are not included in this data. Ocean regions that are distant from one another can have very similar features, for instance both polar regions are characterized by low temperatures. Similarly, nearby areas in the ocean can have dramatically different ecological compositions. The geographical coherence in our results is a product of the oceanographic measures alone.

We used two criteria to choose the appropriate number of clusters: first, that there should be enough clusters to highlight differences between different methodologies for data analysis; second, that cluster boundaries should suggest visible changes in open ocean biome classification, as assessed by an expert oceanographer. The first criterion sets the lower limit on the number of clusters. In this analysis, positive and negative aspects of the various methodologies were difficult to discern with fewer than 8 clusters. The second criterion acknowledges that for the purpose of classifying open ocean biomes, the data analysis should be performed at a sufficiently coarse resolution to avoid too much subdivision along coastlines. Coastal ocean biomes are known to be geographically much smaller than the open ocean biomes they abut. We began to observe a proliferation of small coastal biomes in our results when the algorithms were asked to compute more than 10 to 12 clusters. This range in cluster numbers is approximately equal to the number of open ocean biomes within an ocean basin as determined by other methods (Longhurst, 1998), further supporting the use of 10 to 12 clusters. Based on these criteria, we used 11 clusters throughout our analyses as a generally useful choice for all algorithms.

As our paper came to press, another group published a largely objective study of surface ocean biomes (Oliver & Irwin, 2008). Our study and this recent study are complementary in the aspects of surface ocean classification that they explore. In the present study we include a wide array of variables and focus on the effect of dimensionality reduction on the classification of biomes. In contrast, Oliver and Irwin focus on the problem of automatically identifying the number of surface ocean biomes. Furthermore, while the present study uses k-means clustering exclusively as a final processing step, Oliver and Irwin combine k-means clustering, Wards linkage agglomerative clustering, and a post-hoc separation of clusters based on geographic continuity. It is notable that in spite of the large differences in variables included and study methodology, the major features of our classifications are similar.

Although oceanographers have identified ocean biomes in the past, there is no standard global set of ocean biomes against which to compare our results, and thus there is no explicit ground truth for evaluating each of our methods. This is hardly a unique problem, however, as there are many application domains lacking a well-defined ground truth. Nevertheless, one may still want to investigate those domains with modern machine learning techniques. In the case of ocean biome classification, our evaluation criteria are defined as (i) the correspondence of cluster boundaries with known currents, faunal breaks, or other biogeochemical boundaries, and (ii) the biogeographic and environmental similarity of clustered regions. While these evaluation criteria lack the succinct appeal of a black-and-white classification set, they reflect the state of knowledge within the domain. To apply these criteria to our results, we rely on the expert analysis of an oceanographer well-versed in the ocean biome literature.

3 Methods

Ocean biomes were discovered by analyzing the $n = 9105$ samples of $d = 14$ normalized measurements described in the previous section. We used a number of different methods for exploratory analysis, visualization, and clustering (Burges, 2005) of the data. These included traditional methods, such as k-means clustering and principal component analysis, as well as newer methods such as Isomap (Tenenbaum et al., 2000) and maximum variance unfolding (Weinberger et al., 2004). The methods in manifold learning are strongly motivated by the three dimensional nature of the ocean itself (characterized by latitude, longitude, and depth). For this reason, we anticipated that the manifold learning algorithms would extract three dimensional representations of the data (which would, in turn, be fed as input to k-means clustering). Manifold learning techniques are natural in this application because they allow us to achieve a manageable (and unbiased) representation of the ocean without losing much information. We also experimented with spectral clustering (Ng et al., 2002), but it performed poorly for reasons that we suggest in section 5.

In the following sections, we provide a brief survey of the algorithms used in our study.

3.1 K-means

K-means is one of the simplest and most popular algorithms for unsupervised clustering of multivariate data. The algorithm assigns each data point $\vec{x}_i \in \Re^d$ to one of k disjoint clusters $\{C_\alpha\}_{\alpha=1}^k$, where $\bigcup_\alpha C_\alpha = \{\vec{x}_i\}_{i=1}^n$. Each cluster C_α has a corresponding centroid $\vec{\mu}_\alpha \in \Re^d$. The centroids and cluster assignments are chosen to minimize the vector quantization error:

$$\mathcal{L}_{k\text{-means}} = \sum_{\alpha=1}^k \sum_{i \in C_\alpha} \|\vec{x}_i - \vec{\mu}_\alpha\|^2. \tag{1}$$

Eq. (1) is minimized by a two-step iterative process. First, each input $\vec{x}_i$ is assigned to the cluster with the closest centroid. Second, each centroid $\vec{\mu}_\alpha$ is re-estimated as the mean of the inputs assigned to it. This process is repeated until convergence. Although this optimization is not convex, in practice it often converges to good local minima with high reliability. We investigated k ranging from 2 to 16. For $k = 2$, we initialized the cluster centroids at random; for $k > 2$, we initialized the centroids by inheriting those from a previous run of the algorithm with $k-1$ clusters, then choosing a new centroid from among the data points in the cluster with the largest variance.

3.2 PCA and MDS

Principal component analysis (PCA) is a simple linear method for high dimensional data analysis and visualization. PCA computes a linear orthogonal projection $\mathbf{P} \in \Re^{r \times d}$ that maps the original input space $\Re^d$ into a lower dimensional subspace $\Re^r$. For simplicity, assume the data points are centered. Then the projection $\mathbf{P}$ is computed by minimizing the reconstruction error:

$$\mathcal{L}_{\text{PCA}} = \sum_i \|\vec{x}_i - \mathbf{P}^\top \mathbf{P} \vec{x}_i\|^2 \tag{2}$$

subject to $\mathbf{P}\mathbf{P}^\top = \mathbf{I}$, where $\mathbf{I}$ is the $r \times r$ identity matrix. The constraint ensures that the rows of $\mathbf{P}$ are orthonormal. Although the optimization in eq. (2) is not convex, its global minimum can be computed by singular value decomposition. In particular, the rows of $\mathbf{P}$ are given by the top r eigenvectors of the covariance matrix $\mathbf{C} = \sum_i \frac{1}{n} \vec{x}_i \vec{x}_i^\top$. Each eigenvalue of $\mathbf{C}$ reveals the variance of the coordinate obtained from the projection onto its eigenvector. These coordinates represent the data's so-called principal components.

A closely related method for linear dimensionality reduction is metric multidimensional scaling (MDS). This method produces outputs $\vec{y}_i \in \Re^r$ that best preserve the inner product structure of the original data. The outputs are computed by minimizing:

$$\mathcal{L}_{\text{MDS}} = \sum_{ij} (\vec{y}_i \cdot \vec{y}_j - \vec{x}_i \cdot \vec{x}_j)^2 \tag{3}$$

Despite its different motivation, it can be shown that MDS yields the same solution as PCA. Though the loss function explicitly penalizes differences in inner products, MDS is often used to derive outputs that approximately preserve pairwise Euclidean distances. In practice, starting from a pairwise distance matrix, one infers the corresponding inner products which are then fed as input to MDS.
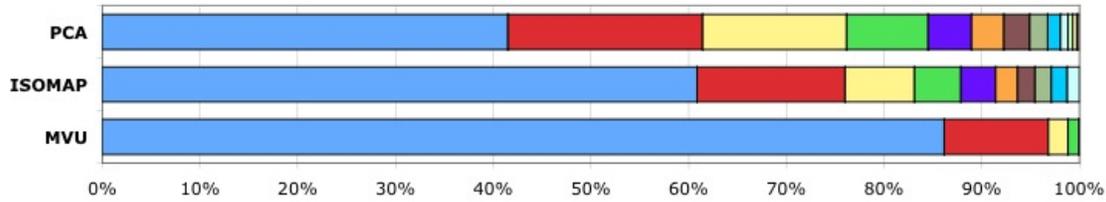
3.3 Isomap

The Isomap algorithm (Tenenbaum et al., 2000) provides a powerful nonlinear extension of multidimensional scaling. It was developed to analyze high dimensional data points sampled from a low dimensional manifold. Whereas MDS focuses on preserving Euclidean distances, Isomap focuses on preserving geodesic distances along the manifold.

The algorithm has three basic steps. The first step computes κ-nearest neighbors of each data point, then uses this information to create an adjacency graph whose nodes represent data points and whose (undirected) edges indicate nearest neighbor relations. The second step estimates the geodesic pairwise distances δ_{ij} along the manifold between points $\vec{x}_i$ and $\vec{j}$. This is done by computing shortest paths through the adjacency graph, with edges weighted by nearest neighbor distances. Finally, the third step uses these pairwise distances as input to MDS. From these distances, MDS outputs low dimensional outputs whose Euclidean distances $\|\vec{y}_i - \vec{y}_j\|$ are approximately equal to the geodesic distances δ_{ij} found in step two.

To analyze our data set of size $n = 9105$ we used a fast approximate implementation of the original Isomap algorithm, known as landmark Isomap (de Silva & Tenenbaum, 2003). Landmark Isomap scales better to large data sets because it only computes the shortest paths between the original data points and some smaller subset of data points designated as landmarks. In our experiments with landmark Isomap, we used 500 landmarks and $\kappa = 15$ nearest neighbors.

The performance guarantees for Isomap depend on an assumption that the data's underlying manifold can be isometrically mapped to a convex subset of Euclidean space. When this assumption does not hold, the algorithm can return spurious results. Later we will examine this assumption in the context of open ocean biome classification.

Figure 2. Normalized eigenvalue spectra from PCA, Isomap, and MVU. Each normalized eigenvalue reveals the relative amount of variance in the corresponding principal component.

3.4 Maximum variance unfolding

Maximum variance unfolding (MVU) (Weinberger et al., 2004; Sun et al., 2006) provides yet another nonlinear extension of multidimensional scaling. Like Isomap, it was developed to analyze data sampled from a low dimensional manifold. While Isomap attempts to preserve geodesic distances, though, MVU focuses on preserving local distances. Let us define the adjacency matrix $\eta_{ij} \in \{0, 1\}$ with $\eta_{ij} = 1$ if $\vec{x}_i$ and $\vec{x}_j$ are κ-nearest neighbors and $\eta_{ij} = 0$ otherwise. MVU attempts to compute the maximum variance configuration of low dimensional outputs $\vec{y}_i \in \Re^r$ that preserve the distances between κ-nearest neighbors. The outputs are computed by minimizing the loss function:

$$\mathcal{L}_{\mathrm{MVU}} = \sum_{ij} \eta_{ij}(\|\vec{y}_i - \vec{y}_j\|^2 - \|\vec{x}_i - \vec{x}_j\|^2)^2 - \nu \sum_i \|\vec{y}_i\|^2,$$

(4)

where the constant ν balances the distance-preserving and variance-maximizing goals of the optimization. The loss function in eq. (4) is not convex. However, its optimization can be reformulated as an instance of semidefinite programming by relaxing the dimensionality of the outputs $\vec{y}_i$. To analyze our data set of size $n = 9105$, we implemented a fast approximation (Weinberger et al., 2007) to the original implementation of MVU. This approximation solves a much smaller semidefinite program to find an approximate minimum of eq. (4). The loss function is then further minimized by conjugate gradient descent.

4 Results

We applied the methods described in the last section to derive low dimensional representations of the data. Fig. 2 shows the normalized eigenvalue spectra from these methods. Note that the top three principal components in PCA account for less than 80% of the variance in the original data. On the other extreme, MVU accounts for nearly all of the variance in three dimensions.

We used the k-means algorithm to derive clusters in the data and to locate likely borders between ocean biomes. The

k-means algorithm was applied to the raw $d = 14$ dimensional data, as well as to the three dimensional representations discovered by Isomap and MVU. Running k-means on top of these representations is appropriate since we are looking for biomes, which conceptually align well with clusters in ocean feature space. To assess the quality of the clustering, k-means cluster assignments were mapped onto global ocean projections and analyzed by an expert oceanographer. The quality of the clustering was measured using two criteria: (i) the correspondence of geographic cluster boundaries with known currents, faunal breaks, or other biogeochemical boundaries, and (ii) the biogeographic and environmental similarity of regions clustered together.

In general, all the methods from section 3 produced reasonable results, exhibiting both geographical continuity and high covariance with major oceanographic and biogeographic regions. However, a detailed review of the results revealed significant differences between the various methods. In what follows, we use two regional comparisons (North Atlantic and Antarctic) to highlight the particular strengths and weaknesses that emerged from this review. For the sake of brevity, we discuss geographic patterns in general terms and in reference to accepted hydrographic (Tomczak & Godfrey, 2003) and biogeographic (Longhurst, 1998) features, with limited examples from species ranges. Also, as shorthand, in the following sections we use Isomap and MVU to refer to the results obtained by k-means clustering of the low dimensional representations discovered by these algorithms.

4.1 North Atlantic comparison

Within the North Atlantic (top panel of Fig. 3), previous expert assessments of biome classification appear to coincide best with the automated results from MVU, followed next by those of raw k-means, then Isomap. Generally speaking, MVU identifies biomes with boundaries and geographic extents that are concordant with oceanographic and biogeographic features, with some evidence for latitudinal over-division. Only MVU has cluster boundaries that lie along a line stretching from northern Canada to northern
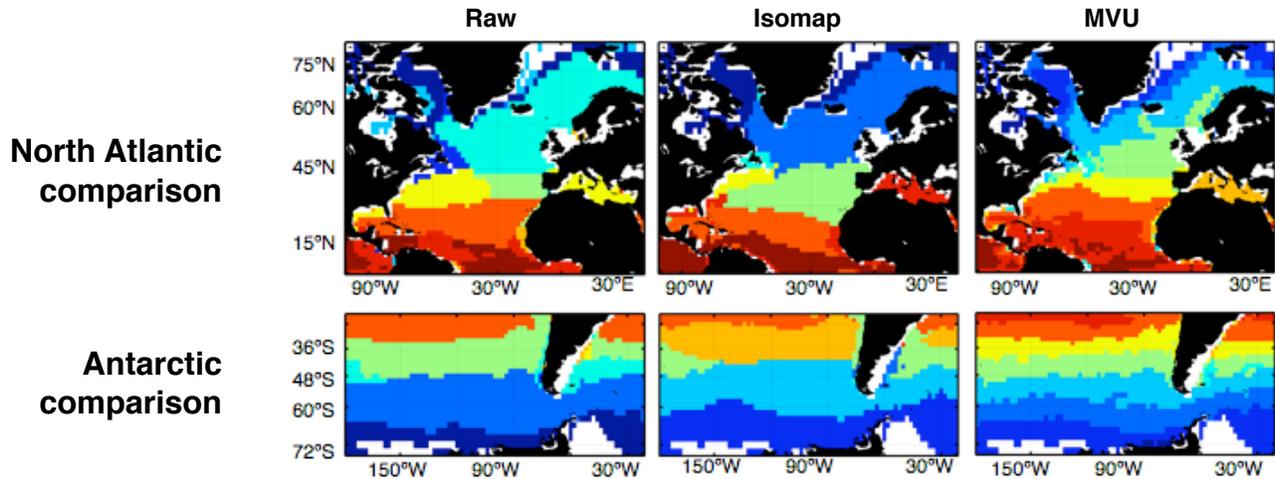
Figure 3. Regional geographic projections of k-means clustering (with $k = 11$) on raw data ($d = 14$) and processed data from the North Atlantic (top) and Antarctic (bottom).

Norway, delineating the North Atlantic Current (Tomczak & Godfrey, 2003) and many species range limits. For example, copepods like *Calanus hyperboreus* and *Paraeucheata norvegica* range to the north of the current, while copepods like *Neocalanus gracilis* and *Metridia lucens* range to the south (Barnard et al, 2004). On the other hand, only raw k-means identifies a separate coastal biome along the northeastern coast of Africa (in orange); this divides a productive coastal upwelling region from the less productive gyre biome (Longhurst, 1998). Isomap suboptimally clusters this coastal region with the Mediterranian and MVU fails to clearly identify a coastal biome at all. MVU and raw k-means have reasonable albeit different subdivisions of the subtropics, a region characterized by low productivity, a stable pycnocline, and high species richness. MVU subdivides the subtropical gyre into a northern biome including the Gulf Stream Extension and Azores Current (yellow), and a southern biome composed of the Sargasso Sea and a southeastern extension along the Subtropical Convergence (red orange) (Tomczak & Godfrey, 2003). Raw k-means clusters the subtropical gyre along a longitudinal divide with both biomes (in yellow and green) notably truncated in their southern extent relative to the Subtropical Convergence (the conventional southern boundary for the subtropical gyre). MVU's latitudinal subdivision of this region is understandable given the different conditions along the edges of the gyre (e.g., intensified currents and associated features) (Longhurst, 1998). However, MVU latitudinally subdivides the tropical Atlantic as well, suggesting a general tendency for latitudinal over-division. Raw k-means highlights a longitudinal subdivision in the sub-

tropical gyre with scant support from known biogeographic ranges, although the biome division has previously been proposed based on east-west differences in *Sargassum* kelp abundance, among other factors (Longhurst, 1998). Overall in the North Atlantic, MVU best emphasizes the known importance of the North Atlantic Current as both a hydrographic feature and a biogeographic boundary.

4.2 Antarctic comparison

In the Southern Ocean (bottom panel of Fig. 3), the best overall characterization of open ocean biomes is provided by k-means clustering of the raw data. The southernmost biome (south of the Antarctic Divergence at 65°S) appears in the raw k-means and MVU clusters only. This biome classification is consistent with the occurrence of ice adapted fauna (e.g., *Stephos longipes* and *Euphausia crystallorophias*) in this seasonally ice-covered region (Longhurst, 1998). In the waters stretching from the Antarctic Divergence to the Subtropical Front (approximately 65°- 45°S), raw k-means and MVU identify one and two biomes, respectively; both are supported by biogeographic ranges. This broad region spans the marginal ice to ice-free zones and crosses several fronts (Tomczak & Godfrey, 2003); some abundant species span this entire region (e.g., *Salpa thompsoni*, *Calanus propinquus*, *Calanoides acutus*) (Longhurst, 1998), while others characterize waters south or north of the Polar Front (e.g., *Euphausia superba* and *Euphausia frigida* respectively) (Brinton, 1962). The next major hydrographic and biogeographic region is the Subtropical Convergence Zone (north of 65°) (Longhurst, 1998), which is characterized by transitional species like

Figure 4. Color-coded geographic projections of the leading three coordinates from MVU and their combination. To improve the overall visualization, each channel displays the same range of color. The green and blue channels would have much less range if weighted by their actual proportion of the variance.

the krill *Thysanoessa gregaria* and *Nematoscelis megalops* (Brinton, 1962). The Subtropical Convergence appears in raw k-means and Isomap (green and orange respectively), but is over-split in MVU (green and yellow bands). Unlike the extra subdivision in the waters stretching from the Antarctic Divergence to the Subtropical Front, the additional MVU division in the Subtropical Convergence is not supported by species ranges or by hydrographic features. Therefore, given the concordance with known faunal and physical features in the Southern Ocean, raw k-means is preferred in this region over both MVU and Isomap.
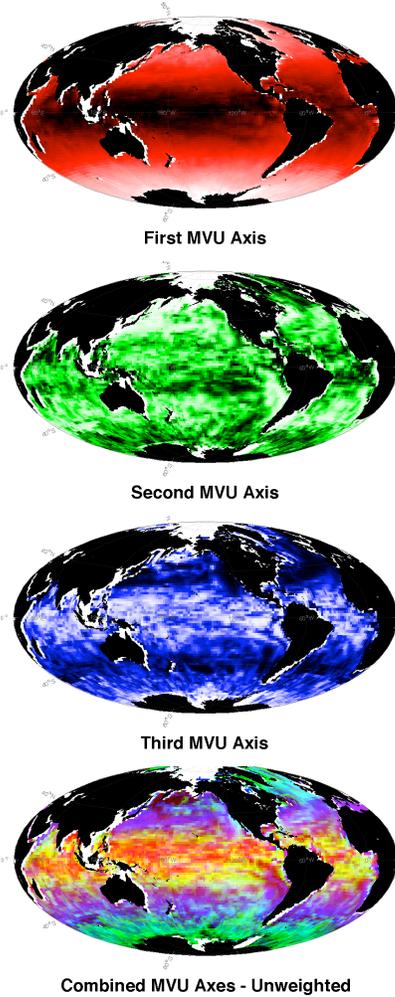
5 Discussion

In this paper, we have applied leading methods in high dimensional data analysis to the visualization and clustering of oceanographic data. We conclude by highlighting the general lessons that emerged from this particular application.

Exploratory analysis and visualization are greatly facilitated by the ability to discover faithful two or three dimensional representations of multivariate data. As shown in Fig. 2, on the $d = 14$ dimensional data set in this paper, both PCA and Isomap fail to discover such representations. PCA fails presumably because it focuses only on linear structure, and our data does not lie principally in a two or three dimensional subspace. We speculate that Isomap's failure stems from its underlying assumption that the data can be isometrically mapped to a convex subset of Euclidean space. Oceanic regions are bounded by currents, across which sharp discontinuities in environmental conditions occur. This characteristic of the ocean may lead to gaps in the continuity of the environmental data (i.e., "holes") and thus violate the convexity assumptions behind the Isomap algorithm, leading to spurious results.

Where PCA and Isomap fail in our application, however, MVU essentially succeeds. Fig. 4 maps the three dimensional representation discovered by MVU onto the globe. The figure enables the three coordinates of MVU to be interpreted (by an expert) in terms of actual environmental variables. In particular, the first MVU coordinate has a clear latitudinal gradient which is highly correlated to the mean annual temperature, PAR, and dissolved oxygen. Likewise, the second and third MVU coordinates each show a weak correlation to five other environmental variables (annual variability in NPP, PAR, temperature, and water column density difference as well as mean annual NPP) that are poorly captured by the first MVU coordinate. Finally, the composite image of these coordinates in the bottom panel of the figure provides a highly graphic and interpretable visualization of the entire data set.

Our results in clustering also present an opportunity to compare traditional versus more recently proposed methods in high dimensional data analysis. For the purpose of open ocean biome classification, the k-means algorithm on the raw data provides a surprisingly successful clustering. However, raw k-means appears to be somewhat insensitive to fine gradients in conditions, as evidenced by the coarse clusters identified in the subpolar North Atlantic. In the subtropical to polar North Atlantic, the results from MVU produce clusterings most consistent with known oceanographic biomes. Unfortunately, this attention to detail by MVU also leads to an emphasis of latitudinal variations in the Antarctic, unlike the results from raw k-means. These distinctions are certainly present in the data, but they might not be useful

if they are too fine for organisms to exploit or if the ecological importance of an environmental change is not linearly related to the magnitude of the change itself.

Based on the results from all of our methods, we recommend MVU for oceanographic analyses within specific subregions (e.g., North Atlantic, North Pacific) or for those seeking to measure fine spatial or temporal gradients. For analyses that are concerned more with general trends or take place on larger scales, however, simple k-means may be the most appropriate method. In addition to the above methods, we tried several other methods not reported here. Spectral clustering (Ng et al., 2002) produced results that lacked geographical continuity, perhaps because, like kernel PCA with a Gaussian kernel, it projects distant points into orthogonal vectors, making it ill-equipped to discover low dimensional manifolds (Weinberger et al., 2004). We investigated several other techniques (e.g., locally linear embedding and additional variants of MVU (Song et al., 2008)) but they were too similar to our main methods to warrant separate discussion.

In conclusion, perhaps the most important lesson of our work is that modern methods in machine learning provide new avenues for exploratory analysis of scientific data. In general, we have found that an interdisciplinary approach is needed to combine the statistical expertise of machine learning researchers with the domain knowledge of natural scientists. This paper provides one example of such a fruitful collaboration.

References

Antonov, J. I., Locarnini, R. A., Boyer, T. P., Mishonov, A. V., & Garcia, H. E. (2006). World Ocean Atlas 2005. Volume 2: Salinity. In S. Levitus (Ed.), *NOAA Atlas NESDIS*, vol. 62, 1–182. Washington, D.C.: U.S. Government Printing Office.

Barnard et al, R. (2004). Continuous plankton records: Plankton atlas of the North Atlantic Ocean (1958-1999). ii. biogeographical charts. *Marine Ecology-Progress Series*, *Supplement*, 11–75.

Behrenfeld, M. J., & Falkowski, P. G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and Oceanography*, *42 (1)*, 1–20.

Brinton, E. (1962). The distribution of pacific euphausiids. *Bulletin of the Scripps Institution of Oceanography*, *8 (2)*, 46–269.

Burges, C. J. C. (2005). Geometric methods for feature extraction and dimensional reduction. In L. Rokach and O. Maimon (Eds.), *Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers*. Kluwer Academic Publishers.

de Silva, V., & Tenenbaum, J. B. (2003). Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems 15* (pp. 721–728). Cambridge, MA: MIT Press.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., & Antonov, J. I. (2006a). World Ocean Atlas 2005. Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen saturation. In S. Levitus (Ed.), *NOAA Atlas NESDIS*, vol. 63, 1–342. Washington, D.C.: U.S. Government Printing Office.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., & Antonov, J. I. (2006b). World Ocean Atlas 2005. Volume 4: Nutrients (phosphate, nitrate, silicate). In S. Levitus (Ed.), *NOAA Atlas NESDIS*, vol. 64, 1–396. Washington, D.C.: U.S. Government Printing Office.

Giesbrecht, W. (1892). *Systematik und Faunistik der pelagischen Copepoden des Golfes von Neapel und der angrenzenden Meeres-Abschnitte*. Berlin: R. Friedlaender.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., & Garcia, H. E. (2006). World Ocean Atlas 2005. Volume 1: Temperature. In S. Levitus (Ed.), *NOAA Atlas NESDIS*, vol. 61, 1–182. Washington, D.C.: U.S. Government Printing Office.

Longhurst, A. (1998). *Ecological Geography of the Sea*. San Diego: Academic Press.

McGowan, J. A. (1971). Oceanic biogeography of the pacific. In B. Funnell and W. R. Riedel (Eds.), *The Micropalaeontology of the Oceans*, 2–74. Cambridge: Cambridge University Press.

Mjolsness, E., & DeCoste, D. (2001). Machine Learning for Science: State of the Art and Future Prospects. *Science*, *293*, 2051–2055.

Ng, A. Y., Jordan, M., & Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems 14* (pp. 849–856). Cambridge, MA: MIT Press.

Oliver, M. J., & Irwin, A. J. (2008). Objective global ocean biogeographic provinces. *Geophys. Res. Lett.*, *35*, L15601.

Song, L., Smola, A., Borgwardt, K., & Gretton, A. (2008). Colored maximum variance unfolding. In J. Platt, D. Koller, Y. Singer and S. Roweis (Eds.), *Advances in neural information processing systems 20*. Cambridge, MA: MIT Press.

Sun, J., Boyd, S., Xiao, L., & Diaconis, P. (2006). The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Review*, *48(4)*, 681–699.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2323.

Tomczak, M., & Godfrey, J. S. (2003). *Regional Oceanography: An Introduction*. Delhi: Daya Publishing House.

Weinberger, K., Sha, F., Zhu, Q., & Saul, L. K. (2007). Graph Laplacian regularization for large-scale semidefinite programming. In B. Schölkopf, J. Platt and T. Hofmann (Eds.), *Advances in neural information processing systems 19*. Cambridge, MA: MIT Press.

Weinberger, K. Q., Sha, F., & Saul, L. K. (2004). Learning a kernel matrix for nonlinear dimensionality reduction. *Proceedings of the Twenty First International Conference on Machine Learning (ICML-04)* (pp. 839–846). Banff, Canada.