Evaluating Unlabeled Spatio-Temporal Patterns: A Global Ocean Eddy Monitoring Application

James H. Faghmous
Computer Science and Engineering
The University of Minnesota
Minneapolis, USA
jfagh@cs.umn.edu

Hung Nguyen
Computer Science and Engineering
The University of Minnesota
Minneapolis, USA

Snigdhansu Chaterjee
School of Statistics
The University of Minnesota
Minneapolis, USA

Vipin Kumar
Computer Science and Engineering
The University of Minnesota
Minneapolis, USA

ABSTRACT

Identifying objects in spatio-temporal fields is a growing challenge, especially when the notion of an object is highly uncertain and no ground truth data are available for evaluation. We present a novel spatio-temporal pattern mining method that autonomously extracts spatio-temporal objects from continuous satellite data to study global ocean dynamics. We empirically and objectively evaluate the quality of the features extracted and devise a methodology to balance the tradeoff between admitting false positives and losing false negatives. Our analysis provides data scientists new suggestions on how to analyze uncertain features with no ground truth yet with sufficient information.

1. INTRODUCTION

Data science and knowledge discovery have fully entered the mainstream with popular consumer applications such as Facebook and Netflix relying primarily on KDD to deliver value to customers. However, a new generation of social problems are proving to be harder to tackle using traditional KDD and call for new methods to address both the unique types of data and questions at hand. Global climate change is one of those problems, and it is argued that it is the defining challenge of our era. Despite the urgency, existing methods have limited applicability to climate data that are commonly noisy, heterogeneous, and spatio-temporal. For instance, the most elegant learning algorithms have strong data independence assumptions that are violated in autocorrelated climate data, where values close in space and time tend to be related.

WOODSTOCK '97 El Paso, Texas USA

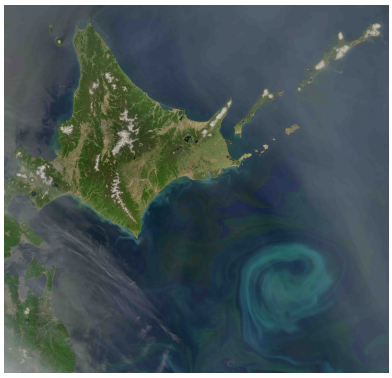

Figure 1: A large eddy off the coast of Japan (bottom right) as observed from a NASA satellite.

One of the biggest challenges when mining climate data is the size of complexity of the datasets at hand. Indeed, the data are so complex and noisy that the search space for patterns is prohibitively large and the likelihood of mistaking spurious patterns as significant is persistent challenge.

One way to cope with this cognitive overload is to use heuristics to reduce the complexity of the space as well as the patterns one is exploring. However, the introduction of any heuristic means that there is a chance of overestimating the significance of spurious patterns (false negatives) as well as discarding real patterns because they do not meet a given heuristic [7]. We are interested in developing novel methods to empirically quantify this trade-off when introducing heuristics to data mining tasks, especially in settings where no ground truth is available but the data convey reasonable information about the task at hand.

This paper focuses on the challenge of identifying unlabeled discrete objects in a continuous spatio-temporal field. Global oceans play a critical role in balancing our planet's heat and energy content. To do so, oceans relies on various phenomena and thanks to global observation systems, such as earth-orbiting satellites, it is possible to monitor them on a global scale. We are interested in monitoring

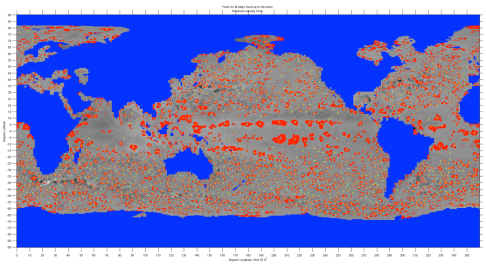

Figure 2: All mesoscale features identified in a single snapshot of SSH. Many of these features do not persist beyond a single week and thus are considered spurious.

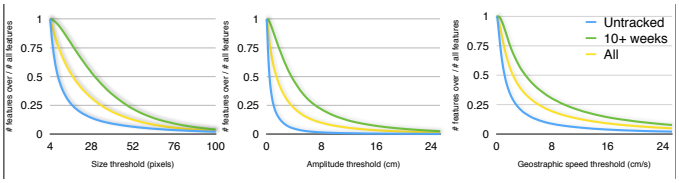

Figure 3: Curves showing the fraction of features that exceed a given threshold for: feature size (left), amplitude (middle), and geostrophic speed (right). For any given cutoff we can see how many features would be omitted from the data. Also, this figure shows that lifetime is a reasonable metric for eddy quality, as the 10+ week curve (green) is consistently above the entire population (yellow) as well as the untracked features (blue).

global mesoscale ocean eddy activity from sea surface height (SSH) satellite data. Ocean eddies are coherent rotating water structures that dominate the ocean's kinetic energy and play a fundamental role in the transport of heat, energy, and nutrients throughout the global oceans. As such, understanding global eddy activity is critical for our planet's sustainability.

Traditionally, eddies are identified by thresholding the continuous SSH field. However, given the large variability and noise in the data the number of candidate objects is artificially high (Figure 2). To reduce the risk of spurious discoveries, domain scientists resort to heuristics to define the physical characteristics that make an eddy. For example, [1] imposed a minimum and maximum feature size of 9 and 1000 pixels respectively. While these necessary, yet arbitrary, thresholds do curb false positives, they also lead to false negatives when real features are discarded because they fail a strict threshold. Furthermore, different studies employ different heuristics. For example, two prominent eddy studies, one published in *Science* [2] and the other in *Nature* [6] relied on significantly different sets of heuristics, yet it unclear how those choices impacted their results.

We contribute to the data science community by providing a new objective approach to evaluating features and the trade-offs involved in applying strict heuristics. More specifically:

- In the case of unlabeled data, what can we learn from the information in the data despite no ground truth?
- When it comes to expert heuristics, can we quantify what we lose/gain by introducing a certain threshold?
- Is there an objective way to measure the importance of such heuristics and how they interact with one another?

Although we focus on a climate application, we believe that any field that is faced with a lack of ground truth yet sufficient feature quality information could use this work as a road-map on how to navigate this uncharted territory.

2. APPROACH AND RESULTS

We identified and tracked eddies for the 19 years of data available using methods presented in [3, 4]. Due to space limitations we will present and discuss results from a single year: 2009. In the 52 weeks spanning 2009, we identified 195,967 eddy-features, 39,601 (20%) of which didn't survive past a week while 53,872 (27%) lasted ten weeks of longer. We noticed that for small features, the proportion of untracked features is around $40-50\%$ of all features identified. However as eddies become bigger that proportion goes down to 10%. Thus size seems to be a reasonable heuristic when looking to remove uncertain features.

Yet by pruning the features based on their physical characteristics, we have introduced spurious features into the system. Similar to other studies, we would like to use some heuristics to clean the dataset of the lowest quality features. However, instead of choosing an arbitrary heuristic, we propose a more systematic approach to selecting a threshold and quantifying the trade-offs associated with such a choice.

3. EVALUATION

This evaluation is performed under several assumptions: First, we assume that the objective of introducing heuristics or thresholds is to remove any low quality features, while keeping as many high quality features as possible. Second, we are dealing with unlabeled and uncertain data. This is because, the SSH data from which the features are derived is subject to noise and errors. Furthermore, the eddy identification as well as the tracking approaches have limitations that may introduce erroneous features or trajectories [5, 4]. Thus, even though a feature might pass all heuristics, due to the absence of ground truth there is still a possibility that such a feature is spurious. Finally, we will use eddy lifetime as the indicator a feature's quality. Therefore, the untracked eddies are the poorest quality eddies, while the longer lived ones are considered high quality features. In this evaluation we consider eddies that lived 10+ weeks as the "high quality" eddies. We tested with various lifetimes and other quality metrics and the results remained the same and due to space constraints we will only report the results for 10+ weeks.

3.1 Impact of threshold choice

We evaluate the effect of pruning features based on their physical characteristics. For any given threshold th, k features would be removed because they fell below th, while $N-k$ would remain, with N being the total number of features. In our case, we would like to select a threshold that would remove as many untracked features as possible while preserving the highest quality features – those that lived 10 weeks or more.

For each characteristic (size, amplitude, and geostrophic

speed), we can set th to take every possible value of that variable and see how many features are removed or remain. Figure 3 shows the fraction of features that exceed a threshold for size, amplitude, and geostrophic speed. We segmented the data into two groups, eddies that were untracked and those that lived for ten or more weeks. We also plot the data for all eddies to show how each population is performing relative to the entire dataset.

Figure 3 allows us to visualize how many features are above or below a given threshold, effectively allowing us to assess how harsh a threshold might be. More importantly, we can compute the fraction of high and low quality features we would throw away for any given threshold th. For all panels in Figure 3, the threshold starts out as too low such that no eddy has a characteristic that falls below that threshold, thus all curves begin at 1 (*i.e.* 100% of the observations exceed the threshold). Similarly, as the threshold becomes harsher there will be a point where all eddies would fail to meet the threshold and all curves will collapse to 0. Let's use Figure 3 to assess the impact a threshold has on the results. For instance, in the left panel of Figure 3, if we apply the commonly used 9 pixel size threshold, we would remove $1 - 74.83\% = 25.17\%$ of features discovered including $1 - 94.03\% = 5.97\%$ of the best features (10+ weeks) and $1 - 47.48\% = 52.52\%$ of the worst features (untracked).

From this figure we can see that there is clear separation between the untracked and 10+ week curves. However for certain extreme thresholds the separation becomes less stark.

3.2 How to select the "best" heuristic?

While Figure 3 give us an idea of how each threshold is performing, we cannot easily compare across thresholds to identify a single best threshold (if it exists). One way to compare the importance of different variables to pruning, we can group the three curves in Figure 3 by normalizing each threshold's performance relative to how many features are discarded from the general population. This is achieved by going through every fraction in the "All" curve (yellow) and recording the corresponding fraction for either tracked and untracked. For example, at threshold $th = 8cm$ in the amplitude panel (Figure 3 center) 12.5% of all features have an amplitude greater than $8cm$. Now we can compute the corresponding fraction of 10+ week and untracked features that remain when the fraction of all remaining features is 12.5%. In this case, there are 30% 10+ week features that remain based on the amplitude threshold, 20% according to the geostrophic speed threshold, and 15% in the size threshold. We repeat this for each feature type and each threshold resulting in 6 curves (3 thresholds $\times$ 2 feature types). Figure 4 shows the normalized plot. Now at any fraction of total features (x-axis), we can know the fraction of good or bad features removed based on every threshold. This allows for a 1-to-1 comparison between thresholds. For example, when 50% of features remain the amplitude threshold removes 20% of good features and 85% of poor features. In contrast, the geostrophic threshold removes 32% and 70% of good and poor features respectively. Figure 4 suggests that for most fractions of total features amplitude does significantly better than the two other metrics, since it removes the largest fraction of low quality eddies while minimizing the fraction of high quality eddies it discards.

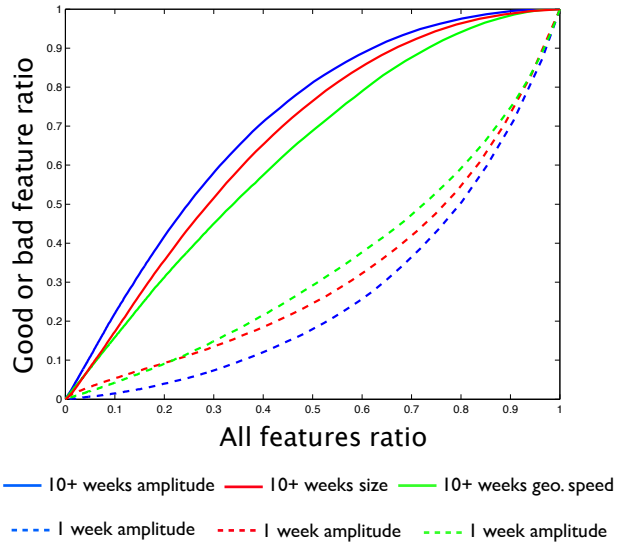

Figure 4: Normalized curve that shows the tradeoffs between selecting a threshold. For any given proportion of the general population that remains, we can tell how many good/bad features were discarded by the threshold that resulted in the general population fraction of interest.

3.3 Maximizing multiple objectives

What become apparent from the previous analysis is that we are interested in optimizing two objectives: Keep as many good features as possible and removing as many poor features along the way. This is a traditional case of Pareto optimality, where improving one objective would result in taking away from the other. Figure 5 shows the Pareto optimality curve for maximizing the fraction of good features kept and the fraction of bad features removed. To generate this figure, we simply map the fraction of good features kept given a threshold from Figure 4 (the solid curve) to the x-axis, and we map 1 minus the value of the corresponding dashed curve in Figure 4 to the y-axis. Similarly to Figure 4, amplitude tends to dominate the other variables for most of the ranges.

Using Figure 5 it is now possible to compute a *gain metric* of how well any threshold value is optimizing these two objectives relative to random chance. At any point in one of the Pareto optimality curves, we can extract how well a threshold optimizes the two objectives. Assume that the fraction of good features kept by a threshold is x and the fraction of bad features discarded is y, thus we are trying to optimize $x + y$. If we introduced a random quantity to use as a threshold, its optimality curve would be along the diagonal (from 1 to 1). In the random case, there is no gain of applying a threshold since the proportion of good features kept is equal to the proportion of bad features kept or $1 - y$. Thus we can compute a gain score relative to random chance by projecting the x and y values for a given threshold unto the diagonal, which results in $gain = x - (1 - y)$.

3.4 Interactions between variables

Now that we are able to quantify the gain of using a threshold for any given variable, we can compute an inter-

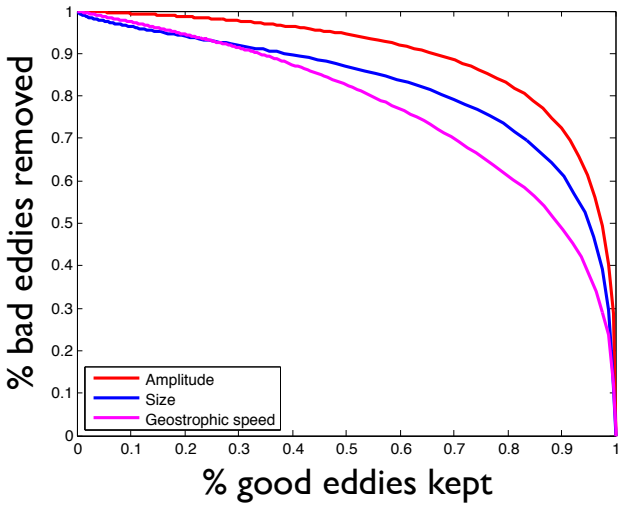

Figure 5: Pareto optimality curve for maximizing the fraction of good features kept and the fraction of bad features removed. This curve is similar to an ROC curve in the sense that the diagonal (random chance) provides no gain.

actions gain score where we compute the gain of applying two thresholds from different variables. Let's assume that we wish to quantify the gain of applying a threshold th_i for amplitude and another threshold th_j for size, where th_i needs not to be equal to th_j. To compute the gain of applying these two thresholds, we first apply both thresholds on the data and remove any feature that fails either threshold. Now, we can reconstruct a Pareto optimality curve like the one in Figure 5 but using the fraction of good features kept and bad features removed using the dataset that meets th_i and th_j. We can then compute the score for that pair of threshold $gain_{ij} = x - (1 - y)$. We repeat this procedure for every pair of threshold values and report the interactions gain matrix in Figure 6.

Figure 6 shows that for any combination of variables there is a considerable high gain areas and the thresholds need not to be precise. Furthermore, we find that the best threshold combination is that of 1cm amplitude and 9 pixels size, which is the heuristics used in [5]. However, now we are able to empirically quantify how much better is that choice of heuristics compared to others and it seems that there are several other reasonable threshold combinations.

4. CONCLUSIONS

In this paper we presented an objective method to both select a heuristic to reduce the risk of false positives and false negatives.Furthermore, we are able to empirically quantify the tradeoffs between selecting a certain heuristic. This is useful since, despite the fact the previous studies had identified the optimal heuristic combination using expert knowledge, we were able to recover the same pair of heuristic but with the added insight of the exact value and tradeoffs of choosing such a combination. One future avenue could look at different optimization objectives. In our case, we equally valued keep good features and removing bad ones. But other might have different preferences. One could re-write the gain equation as $gain = x - \beta(1 - y)$; where β is one's tolerance

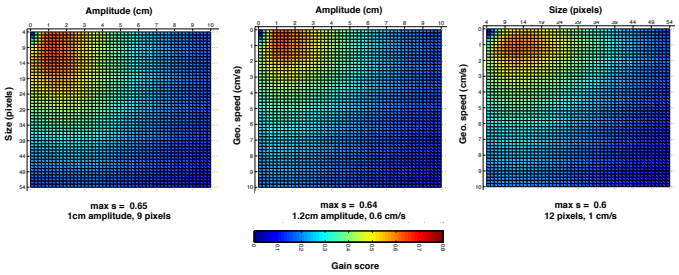

Figure 6: Variable interactions gain matrix which shows the gain relative to random chance to applying to thresholds th_i and th_j to the data. This shows that best gain is attained by using a dual threshold of 1cm amplitude and 9 pixels (left panel). However, there is a significant range of high gain scores in that neighborhood.

for including poor features. So if one is interested in removing all poor features, then β would be high. Such an optimization can be used across other applications such as e-commerce where one might prefer to not lose an existing customer versus gaining a new one.

5. ACKNOWLEDGMENTS

This research was funded by a University of Minnesota Doctoral Dissertation Fellowship and NSF Expeditions in Computing Grant (IIS-1029711).

6. REFERENCES

[1] D. Chelton, M. Schlax, and R. Samelson. Global observations of nonlinear mesoscale eddies. *Progress in Oceanography*, 2011.

[2] D. B. Chelton, P. Gaube, M. G. Schlax, J. J. Early, and R. M. Samelson. The influence of nonlinear mesoscale eddies on near-surface oceanic chlorophyll. *Science*, 334(6054):328–332, 2011.

[3] J. H. Faghmous, M. Le, M. Uluyol, S. Chaterjee, and V. Kumar. Parameter-free spatio-temporal data mining to catalogue global ocean dynamic. In *Thirteenth International Conference on Data Mining (ICDM-13)*, 2013.

[4] J. H. Faghmous, V. Mithal, M. Uluyol, M. Le, L. Styles, S. Boriah, and V. Kumar. Multiple hypothesis object tracking for unsupervised self-learning: An ocean eddy tracking application. In *Twenty-Sventh Conference on Artificial Intelligence (AAAI-13)*, 2013.

[5] J. H. Faghmous, L. Styles, V. Mithal, S. Boriah, S. Liess, F. Vikebo, M. d. S. Mesquita, and V. Kumar. Eddyscan: A physically consistent ocean eddy monitoring application. In *Intelligent Data Understanding (CIDU), 2012 Conference on*, pages 96 –103, oct. 2012.

[6] I. Frenger, N. Gruber, R. Knutti, and M. Münnich. Imprint of southern ocean eddies on winds, clouds and rainfall. *Nature Geoscience*, 6(8):608–612, 2013.

[7] E. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards parameter-free data mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 206–215. ACM, 2004.