Analysis of sea surface temperature maps via topological machine learning

Francesco Conti
Dep. of Mathematics
University of Pisa
Pisa, Italy
name.surname@phd.unipi.it

Oscar Papini
Institute of Information Science and Technologies
National Research Council
Pisa, Italy
orcid: 0000-0003-2069-5068

Davide Moroni
Institute of Information Science and Technologies
National Research Council
Pisa, Italy
orcid: 0000-0002-5175-5126

Gabriele Pieri
Institute of Information Science and Technologies
National Research Council
Pisa, Italy
orcid: 0000-0001-5068-2861

Marco Reggiannini
Institute of Information Science and Technologies
National Research Council
Pisa, Italy
orcid: 0000-0002-4872-9541

Maria Antonietta Pascali
Institute of Information Science and Technologies
National Research Council
Pisa, Italy
orcid: 0000-0001-7742-8126

2023 IX International Conference on Information Technology and Nanotechnology (ITNT) | 979-8-3503-9733-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITNT57377.2023.10139044

***Abstract*—Computational methods to leverage topological features occurring in signals and images are currently one of the most innovative trends in applied mathematics. In this paper a pipeline of topological machine learning is applied to the challenging task of classifying four specific marine mesoscale patterns from remote sensing data, i.e., Sea Surface Temperature maps of the southwestern region of the Iberian Peninsula. Our preliminary study achieves an accuracy of 56% in the 4-label classification. Such results are encouraging, especially considering that the data are affected by noise and that there are low-quality/missing data. Also, the paper devises directions for future improvements.**

Keywords— topological data analysis, machine learning, image classification, temperature map, remote sensing, digital image, 3D point cloud, computational topology, upwelling classification.

I. Introduction

Recent advances in remote sensing are successfully applied to marine observation. Remote sensing provides the experts with a huge number of data acquired by satellite sensors; hence, there is a need for automatic or semiautomatic methods to analyze them. Our case study is the classification of the upwelling regimes of the Iberia/Canary Current System (ICCS), one of the least studied among the upwelling ecosystems [1]. Among all the relevant underlying processes, these mesoscale events (e.g., upwelling, countercurrents and filaments) are of particular interest because they cause the transportation of deeper, colder and nutrient-rich waters to the surface, hence, they affect the biological parameters of the habitat, and enhance the local biodiversity [2]. In other words, there is a connection between the biogeochemical and physical processes occurring in a marine biological system.

Sea surface temperature is the measure of the water's temperature at the ocean's surface. Sea surface temperature is measured by satellite instruments that record the energy emanating from the ocean surface globally, which is emitted at different wavelengths. These data are then validated with temperature readings collected by ships and buoys.

Collecting, processing, and understanding the sea surface data is important because it allows to understand why some regions are warmer or cooler and its impact on the marine environment. Changes in the environment may impact the life of species living there, modifying the access to food, altering the migration patterns, or changing the access to mates.

Globally, the National Oceanic and Atmospheric Administration (NOAA) has collected data describing the global warming: the ocean has warmed approximately 0.13℃ every 10 years over the last century, hence changing the variety of marine organisms and bleaching corals. It also impacts the amount of water vapor available increasing the chance and frequency of more severe and stronger events (storms, floods, and droughts).

To the best of our knowledge, very few solutions have been developed to tackle the automation of the upwelling event classification. One of the most recent has been proposed in [3,4,5]. The method described in [3] aims at the quantitative description of the upwelling events observed in the southwestern Iberian Peninsula; the "spaghetti graphs" are exploited to characterize the upwelling analyzing the temporal evolution of the Sea Surface Temperature (SST) maps. No learning paradigm is applied but, on other hand, the results are preliminary with respect to the automatic classification of large imagery data.

The aim of this work is to develop an automatic method based on topological machine learning to recognize such mesoscale events by analyzing the SST maps acquired by satellite sensing technologies, in the southwestern region of the Iberian Peninsula.

The paper is arranged as follows: Section II is devoted to the description of the classification pipeline; in Section III it is reported how the experiment has been carried out, i.e., data selection, preprocessing, and the result achieved. A brief discussion of the results is reported in Section IV, along with foreseen improvements of our methods for the classification of upwelling regimes.

II. Methodology

A. Computational Topology

Algebraic Topology is a branch of mathematics dealing with shapes. In a nutshell, you assign an algebraic object to a topological space. Then, you can compute several invariants and descriptors (the Euler characteristic, the Betti numbers, the homology groups). Such invariants and descriptors are used in computational topology to compare shapes, and to define distances between them. One of the most used tools

from computational topology is persistent homology, which is a method for computing topological features of a space across different scales. More persistent features are detected over a wide range of spatial resolutions (have a long "lifespan") and are deemed more likely to represent important, or true features of the underlying space, rather than artifacts of sampling, or noise. Applications of persistent homology span from computer vision and shape analysis to biomedical imaging and complex networks analysis. In this perspective, one of the most promising trends is the merging of persistent homology with machine and deep learning [6].

B. Topological Machine Learning

The core idea in topological machine learning is to read the digital data as an algebraic object, from which topological invariants are computed; hence, use such invariants for training machine learning (ML) classifiers. One of the most effective tools for producing topological descriptors is persistent homology: it computes a persistence diagram (PD) for each dimension to any algebraic representation of the input data (as a simplicial complex with filtration). A wide range of transformations has been devised to exploit the capabilities of PDs in ML algorithms [7,8,9,10,11,12,13]. Each of them has been devised with specific requirements. Among all of them, our solution follows the classification of the MNIST dataset presented in [14]. Firstly, we compute for each input data 18 PDs for each homology dimension considered (0 and 1), generated by the following 18 filtrations:

- the grayscale filtration of both the edge map and its negative version;
- the height filtration (8 directions, applied to the edge map);
- the radial filtration (nine origins);
- the density filtration (with radius = 6 and the Euclidean norm).

Hence, several vectorization methods (persistence image [10] with bandwidth in {0.1, 1, 10} and resolution in {5, 10, 25}, persistence landscape [8] with resolution in {25, 50, 75, 100}, persistence silhouette [7] with resolution in {25, 50, 75, 100}, Betti Curve [9] with resolution in {25, 50, 75, 100}) are used to prepare the input for the classification step. A total of nine classifiers are trained for each representation; such classifiers are: Support Vector Classifiers (SVC) with kernel RBF and C in {1,2,3,5,10,20}; a random forest classifier (#trees = 100); and Lasso (α=1).

In the end, the several vectorizations of the topological descriptors are used to train several ML classifiers. Among all the trained models, the best performing is selected as the final topology-based classification model.

III. The Experiment

In our experiment we collect satellite imagery from two satellite sources: the EUMETSAT's METOP-A and -B [15], and the NASA's Aqua [16]. A visual inspection of SST maps of the southwestern region of the Iberian Peninsula has been performed by experts, looking for a minimal set of patterns. The annotation process identified four typologies of mesoscale events as the most representative. The first mesoscale pattern (E1) is associated with the meander of the southward upwelling jet to the west, near Cape St. Vincent, alongside occurring the development of upwelling filaments. Pattern E2 is depicted by the southwards flow of the upwelling jet overpassing the Cape St. Vincent forming an extended meridional filament. Pattern E3 is characterized by a clear line of cool water throughout the whole southern Iberian coast. Actually, experts distinguish two sub-types in the E3 patterns [3], but we do not take care of this splitting in the present experimentation. Finally, pattern E4 occurs when a warm countercurrent develops near the southern Iberian coast, surrounding Cape S. Vincent, and flowing north near the coast. The present experimentation develops in several steps:

- Definition of the SST dataset;
- Preprocessing of the data;
- Topological Machine Learning pipeline.

This last step produces a topology-based classification model.

A. Satellite SST

A selection of 503 images (381 METOP; 122 Aqua) from the years 2009 to 2016 has been downloaded and manually classified by experts in the mesoscale patterns: {E1, E2, E3, E4}. The resulting dataset is balanced: it is made of 125 images belonging to the E1 class and 126 to each of the E2, E3, and E4 classes (of which respectively 89, 99, 105, 88 from METOP; 36, 27, 21, 38 from Aqua). The satellite sources declare a spatial resolution at nadir of 1 km, and a temperature accuracy of 0.01 °C (METOP) or 0.005 °C (Aqua); the SST maps collected show values ranging in [-2°C, 36°C]. The files were provided in either NetCDF-4 or HDF format (the latter only for pre-2014 Aqua files) and subsequently converted into 8-bit grayscale PNG images; in particular, for each image, the following steps were performed:

- information about the latitude, longitude and temperature value was extracted from the NetCDF/HDF file and stored in three NumPy arrays;
- a Cartopy GeoAxis was prepared with a Plate Carrée projection and an extent of [36° N, 39.5° N] × [10.5° W, 7° W];
- a grayscale colormap was defined such that a temperature of 5 °C corresponds to gray 95%, a temperature of 25 °C to gray 0% (black) and the in-between values are linearly interpolated; moreover, the white color has been assigned to missing or low-quality data;
- the temperature data was plotted in the GeoAxis using MatPlotLib's pcolormesh method (normalized between 5 °C and 25 °C) and saved using MatPlotLib's savefig method, with a 0.2-inch white padding, resulting in a 409×409 PNG image (370×370 without the white frame).

Note that the thermal resolution of the raw data is 0.01°C, at least, but when the temperature map is converted into a PNG file, such a resolution may be lower.

B. Preprocessing

The dataset considered is made of images which, even if they could be correctly classified by experts, very frequently are affected by noise (clouds), or may contain vast areas of missing data (e.g., large white stripe, leaving only 50% of sea surface visible). To enhance the signal content of each image

and, consequently, the mesoscale pattern of the sea surface, a preprocessing pipeline has been applied (see Fig. 1):

- The Iberian Peninsula is filtered out, using a black mask;
- Multi-threshold Otsu 5-class segmentation, Scikit-Image implementation [17]
- Median filtering (kernel size: 7) followed by Gaussian filtering (kernel size: 3), implemented in OpenCV
- Fat Edge extraction using the Python PIL inbuilt function ImageFilter.FIND_EDGE (based on a Laplacian kernel of size 3).

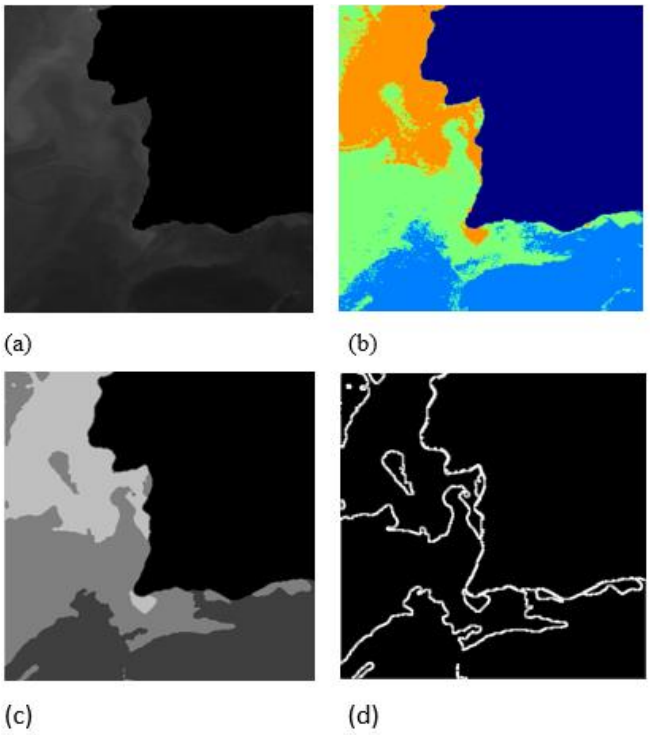

Fig. 1 Image preprocessing. (a) Original greyscale image; (b) segmented image obtained through the Otsu multi-thresholding; (c) smoothing and blurring through median and Gaussian filtering; (d) fat edges. This last image is the input of the topological pipeline of classification.

C. *Best model and Classification results*

The methodology described in the previous Section produced 588 topology-based models, which have been trained to classify the upwelling patterns {E1, E2, E3, E4} on the balanced dataset of 403 b/w edge maps, which are the output of the preprocessing steps described in the previous subsection. An external test set is used to perform the model selection, in order to select the best-performing classification model: the best-performing model exploits the Betti curves in both dimensions 0 and 1 (as topological descriptors), and the Ridge classifier (as ML method). The results of classification of such model are shown in the confusion matrix of Fig. 2. In the end, our method is fully automatic and achieved a 56% of overall accuracy in the task of 4-label classification. Also, Fig. 2 shows that the classification is excellent for E3 and E4 patterns, while unsatisfactory for E1 and E2 patterns. In more detail, the misclassification between E2 and E3 caught our attention: there are 14 (13+1) images out of 50 which are classified incorrectly. Actually, in both E2 and E3 there is a cold region flowing; while in E2 the cold water flows in the south, in E3 the cold water flows closer to the coast, bending to east. Moreover, often such an event disappears before it has taken a definitive and clearly visible turn.

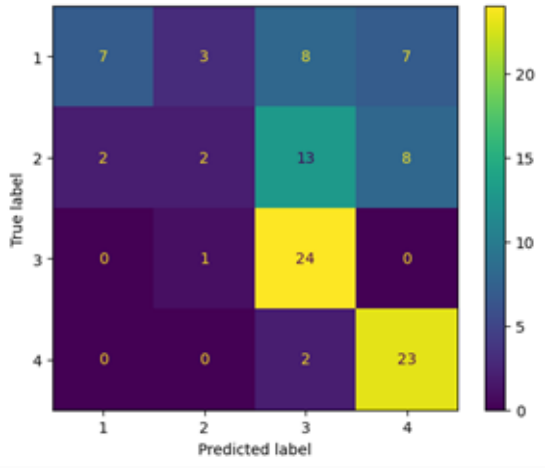

Fig. 2. Confusion matrix of classification. Note: E3 and E4 events are classified very well, while the classification of E1 and E2 requires additional work to be improved.

IV. Discussion and Conclusions

The results described in the previous section are encouraging because they show that topological descriptors extracted from SST maps could provide an excellent support for the detection of E3 and E4 patterns, as in both cases, the classification accuracy is very high (respectively 24/25 and 23/25), showing remarkable robustness against noise and missing signal. On the contrary, the classification of E1 and E2 patterns requires to be improved. There are many approaches we plan to adopt in order to improve our results:

(i) consider a temporal sequence of SST maps, in order to appreciate the evolution of a specific pattern. This approach would help to better discriminate E2 from E3, for example, and also to develop an automatic tool able not only to recognize a specific upwelling event, but also to predict it;

(ii) improve the image preprocessing by exploiting convolutional neural networks to tackle the problem of noisy signal, or missing signal in SST maps, e.g. using an autoencoder with Bayesian optimization, as done in [18,19].

Acknowledgment

The authors express gratitude to Prof. Flávio Martins and Dr João Janeiro from the University of Algarve, Centre for Marine and Environmental Research, for their support.

Funding

This work has been partially funded by the European Union's Horizon 2020 research and innovation programme, under grant agreement No. 101000825 (NAUTILOS, https://www.nautilos-h2020.eu/).

References

[1] J. Kämpf and P. Chapman, "Upwelling systems of the world: A scientific journey to the most productive Marine ecosystems", Springer International Publishing, pp. 31-42, 2016. doi: 10.1007/978-3-319-42524-5_4.

[2] R. Varela, F. P. Lima, R. Seabra, C. Meneghesso and M. Gómez-Gesteira, “Coastal warming and wind-driven upwelling: A global analysis,” Science of The Total Environment, vol. 639, pp. 1501-1511, 2018. doi: 10.1016/j.scitotenv.2018.05.273.

[3] M. Reggiannini, J. Janeiro, F. Martins, O. Papini, and G. Pieri, “Mesoscale patterns identification through SST image processing” in Proc. of the 2nd Int. Conf. on Robotics, Computer Vision and Intelligent Systems, pp. 165–172, 2021, doi: 10.5220/0010714600003061

[4] M. Reggiannini, M., O. Papini, G. Pieri, “An Automated Analysis Tool for the Classification of Sea Surface Temperature Imagery” Pattern Recognit. Image Anal., vol. 32, pp. 631–635, 2022. doi: 10.1134/S1054661822030336

[5] G. Pieri, J. Janeiro, F. Martins, O. Papini, and M. Reggiannini, “MEC: A Mesoscale Events Classifier for Oceanographic Imagery,” Applied Sciences, vol. 13, no. 3, p. 1565, Jan. 2023, doi: 10.3390/app13031565.

[6] D. Moroni, M.A. Pascali, “Learning Topology: Bridging Computational Topology and Machine Learning”, Pattern Recognit. Image Anal. 31, 443–453, 2021, doi: 10.1134/S1054661821030184.

[7] F. Chazal, B.T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, “Stochastic convergence of persistence landscapes and silhouettes” Proceedings of the Thirtieth Annual Symposium on Computational Geometry, Kyoto, pp. 474–483, 2014.

[8] P. Bubenik, et al. “Statistical topological data analysis using persistence landscapes” J. Mach. Learn. Res. 16, 77–102, 2015.

[9] Y. Umeda, “Time series classification via topological data analysis”, Inf. Media Technol., vol. 12, pp. 228–239, 2017.

[10] H. Adams, et al.” Persistence images: A stable vector representation of persistent homology” J. Mach. Learn. Res., art. no. 18, 2017.

[11] C. Chen; X. Ni, Q. Bai, Y. Wang , “A topological regularizer for classifiers via persistent homology”, Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, pp. 2573–2582, 2019.

[12] C.S. Pun, S. X. Lee, K. Xia. “Persistent-homology-based machine learning: a survey and a comparative study.” Artif. Intell. Rev., vol. 55, no. 7, pp. 5169–5213, 2022.

[13] R. Corbet, U. Fugacci, M. Kerber, C. Landi, B. Wang, “A kernel for multi-parameter persistent homology”, Comput. Graph. X, vol. 2, art. no. 100005, 2019.

[14] F. Conti, D. Moroni, and M. A. Pascali, "A Topological Machine Learning Pipeline for Classification" Mathematics 10, no. 17, art. no. 3086, doi: 10.3390/math10173086, 2022.

[15] OSI SAF. Full resolution L2P AVHRR Sea Surface Temperature MetaGRanules (GHRSST) - Metop. 2011. doi: 10.15770/EUM_SAF_OSI_NRT_2013.

[16] NASA/JPL. GHRSST Level 2P Global Sea Surface Skin Temperature from the Moderate Resolution Imaging Spectroradiometer (MODIS) on the NASA Aqua satellite (GDS2), 2020. doi: 10.5067/GHMDA-2PJ19.

[17] P.-S. Liao, T.-S. Chen, and P.-C. Chung, “A fast algorithm for multilevel thresholding”, Journal of Information Science and Engineering, vol. 17, no. 5, pp. 713-727, 2001.

[18] A. Barth, A. Alvera-Azcárate, M. Licer, J.-M. and Beckers, “DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations”, Geosci. Model Dev., vol. 13, pp. 1609–1622, 2020. https://doi.org/10.5194/gmd-13-1609-2020

[19] A. Barth, A. Alvera-Azcárate, C. Troupin, J.-M., and Beckers, “DINCAE 2.0: multivariate convolutional neural network with error estimates to reconstruct sea surface temperature satellite and altimetry observations”, Geosci. Model Dev., vol. 15, pp. 2183–2196, 2022. doi: 10.5194/gmd-15-2183-2022