Clustering on tropical cyclogenesis over environmental predictors and intertropical oceanic basins.

Jérémie Fragnière

Abstract

Tropical cyclones form at very precise atmospheric and oceanic conditions, therefore it can be hard to apprehend all the mechanisms involved. This paper focus on the unsupervised classification of formed cyclones via their localisation in intertropical basins and their environmental conditions. KMeans and DBSCAN were compared with different hyperparameters. KMeans showed better results for the classification. Three groups were identified showing that cyclones tend to group more according to the environmental predictors, in particular humidity, convection and moisture.

1. Introduction

Knowledge on tropical cyclones has evolved a lot up to know. Knew observations, technologies and discovers in many fields have permitted a better understanding of the dynamics involved in those meteorological phenomena (Emanuel & Center, 2018). Nowadays, a tropical cyclone is defined as an intensified circular storm formed over the inter-tropical seas, and gathering its energy from warm waters below, It is driven by a low atmospheric pressure core generating strong winds and precipitations. The nomenclature depends on the intensity and on the oceanic basins it occurs. In the western North Pacific for example, those storms are called typhoons (Zehnder, 2022). The ensemble of their conditions and mechanisms of formation is called cyclogenesis and is the main topic of this paper.

The conditions for the intensification of a depression into a tropical cyclone are very precise and involve processes from Atmospheric and oceanic dynamics, as well as biogeochemistry (Emanuel & Center, 2018). They include temperature and depth of the oceanic surface layer, already existing atmospheric circulation with conditions favourable to the formation of convective clouds, humidity, geographic position, and wind speed gradient (Zehnder, 2022). Tropical cyclogenesis is a continuous process occurring and evolving at different spatiotemporal scales. Involved dynamics are influenced by external environmental as well as internal factors, generating feedbacks impacting the genesis. Vertical wind shear, oscillation interactions, convective evolution, and friction can be contributing factors for the development and the evolution of tropical cyclones (Tang et al., 2020). Those events are very destructive and deadly. A better understanding of their mechanics could upgrade the way they are apprehended and significantly improve their prevention and risk assessment. There are plenty of methodologies to work with cyclogenesis. In this paper, it was studied using an unsupervised machine learning approach.

Two clustering algorithms, sickit-learn's KMeans and DBSCAN, were used on a dataset containing points of formed cyclones over seven tropical oceanic basins and eleven associated environmental variables. The goal was to compare the algorithms and identify groups in the dataset to analyse how the basins and environmental predictors were put together. The clusters were compared using bar graph for the basins and boxplot for the predictors.

2. Dataset

The dataset consists in a netCDF file containing 11 environmental variables linked to the formation of tropical cyclones. They are climatological tropical cyclone formation probability (CLIM), percent of the area covered by an r=500km circle covered by land (PLND), average weekly Reynold's SST (RSST), average 850-200 hPa vertical shear (VSHD), average 850 hPa relative vorticity (RVOR), average vertical instability parameter (THDV), average 850 hPa horizontal divergence (HDIV), amount of sustained deep convection (PCCD), mid- to upper-level moisture (BTWM), average mean sea level pressure (MSLP), average 850 hPa horizonal temperature advection (TADV), and mid level relative humidity (MLRH). The dataset is linked to the work on cyclogenesis prediction by the regional and mesoscale meteorology branch (RAMMB) at the cooperative institute for research in the atmosphere (CIRA). However, their work isn't addressed in this paper.

The dataset comes in three dimensions: time, lattitude and longitude. The time goes from 1995-01 01T18:00:00 to 2010-12-23T12:00:00, the lattitude from -45° to 45°, and the longitude from 0° to 359°. The values of the 11 variables

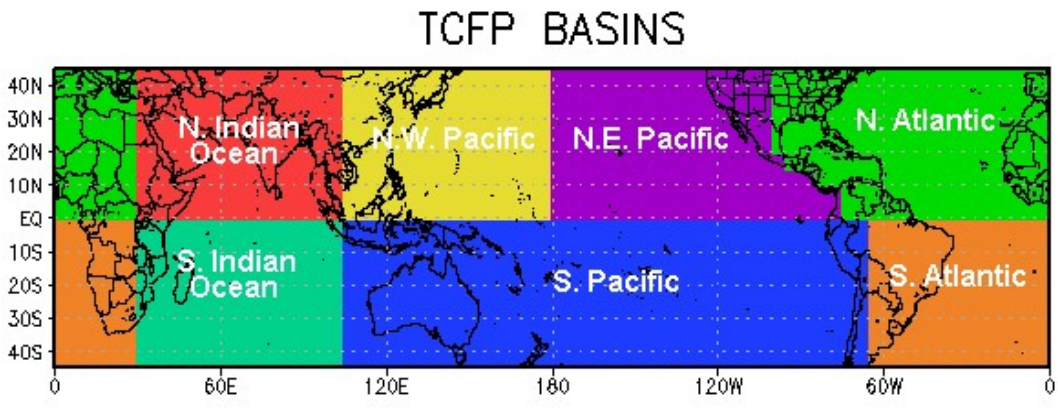

Figure 1. Map of tropical cyclone formation probability oceanic basins.

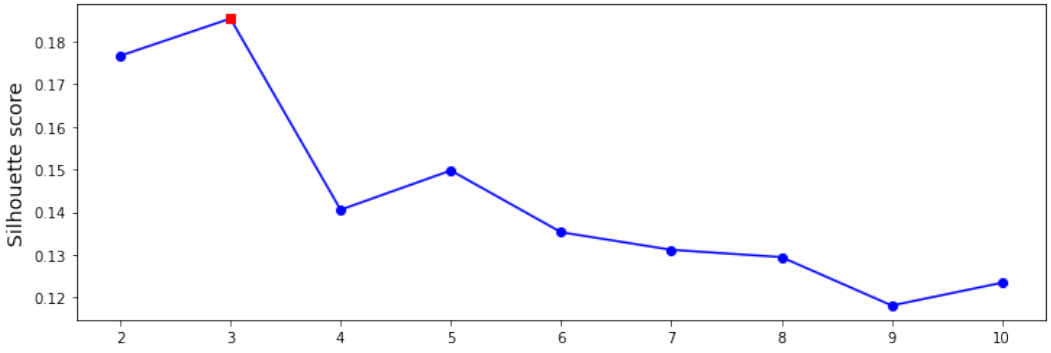

Figure 2. KMeans silhoutte scores in function of the number of clusters k.

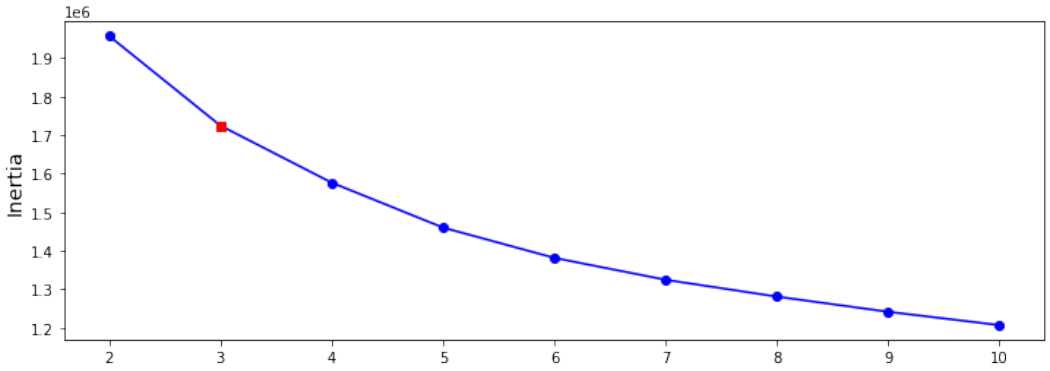

Figure 3. KMeans inertias in function of the number of clusters k.

for each set of coordinates represents the conditions at which a tropical cyclone actually formed. Those are called the true positive datapoints and represents the data used in this project. A sample of the conditions for which a cyclone didn't form, the true negative datapoints, is also available but it wasn't used in this paper because the focus was on the effective formation of cyclones.

3. Methodology

First, to determine in which oceanic basin each sample belongs, a categorical variable representing the basins was added to the dataset. The basins were defined using the tropical cyclone formation probability map made by the RAMMB (Fig.1). A black and white map with the same dimensions was loaded. The pixels were used as geographical coordinates and assigned a variable between 1 to 7 in function of the basin they represented as follow: N.Atlantic (1), N.Indian (2), N.-W.Pacific (3), N.E.Pacific (4), S.Atlantic (5), S.Indian (6), and S.Pacific (7). The result was merged with the dataset to assign a basin to each true positive point. After some NaN cleaning, the data were reshaped in (n_samples, n_features) format as required for the clustering. The data were also scaled using StandardScaler class from sickit-learn.

3.1. KMeans

A KMeans model was trained with a number of clusters k going from 2 to 10, to find which one fitted the data the best. The method of initialization was "k-means++" and the K-means algorithms tried were "lloyd" and "elkan". The number of initial centroïds was set to 100 in order to have a stronger model without taking too much computation time. The performance metrics used were only silhouette score and inertia, since there were no true labels for this dataset. Those metrics were extracted and plotted in function of k to find the best value. The clusters were then predicted with the best models. The distribution of the basin variable was represented using bar graph, and the environmental predictors using boxplots.

3.2. DBSCAN

Unlike KMeans, DBSCAN is a clustering algorithm based on density (Shubert et al., 2017). It doesn't need an pre-initialization of the number of clusters and allows to get rid of the noise points of the dataset. It has two major hyperparameters: "eps" and "min_samples". The latter was determined with different epoch of the model and a range from the default one 5, to 100. The biggest numbers were expected to work better because they seem to be more appropriate for large and noisy datasets (Shubert et al., 2017). The second hyperparameter was determined using the sickit learn's k-nearest neighbors method with the min_samples

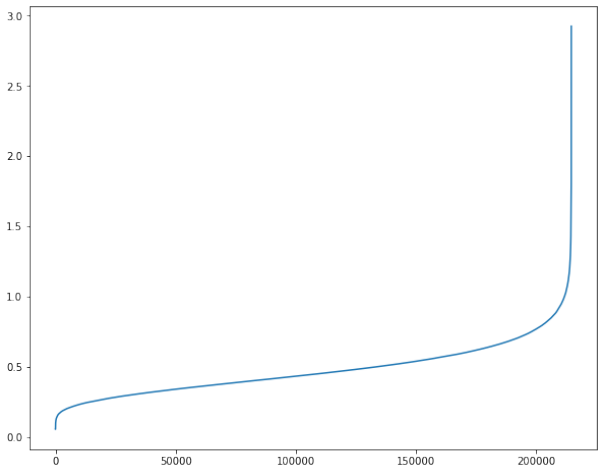

Figure 4. K-nearest neighbors method for DBSCAN.

value as the n_neighbors parameter. The most appropriate values to use should be in the elbow of the graph were the slope is the bigger (Mane, 2020). The metrics used to calculate the distance between the instance were 'euclidian' and 'manhattan'. The resulting DBSCAN model was then fitted to the data. The distribution was also represented as bar graph and boxplots, but the noise points were excluded.

4. Results

4.1. KMeans

For the "lloyd" algorithm. The silhouette score (Fig. 2) and inertia (Fig. 3) graphics show that the best number of clusters is 3, with respective values of 0.185 and 1.72 x 10^6. The computation with the "elkan" variation give the same results, so only the 'lloyd' algorithm is considered for the graphic vizualisation. The low silhouette score shows that the model don't do very well with the dataset.

The bar graph (Fig.5) shows that the first cluster groups a big part of the points from all the basins, with the most frequents being N.W. and N.E. Pacific. The second cluster have a majority of N.Atlantic points, but both N.Pacific basins are also well represented. The third cluster is dominated by N.Atlantic and N.E. Pacific.

On the three boxplot graphs (Fig.8), the variable TADV is ditributed around 0 and have outliers covering the whole range from negative to positive. It seems however that the values are higher in cluster 3. Every cluster shows a relatively even distribution of the values between negative and positive. It could be explain by the fact that points from every basins are present in all three clusters. However, the first cluster shows positive dominating values for MLRH and PCCD, and negative dominating values for BTWM and PLND. In the cluster 2, the variable PLND isn't much

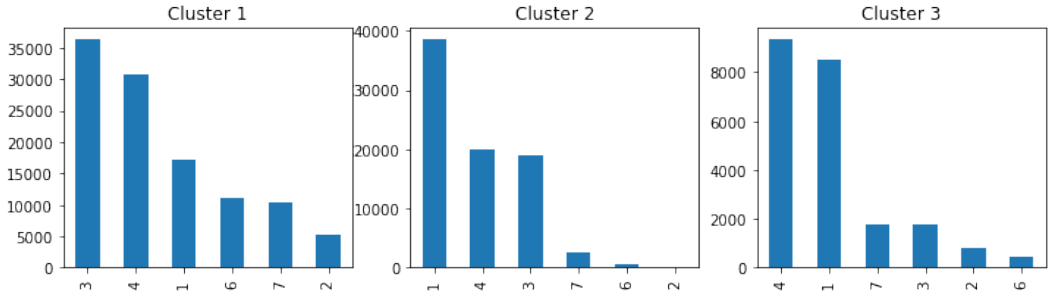

Figure 5. Bar graphs of the basins frequency for KMeans clusters.

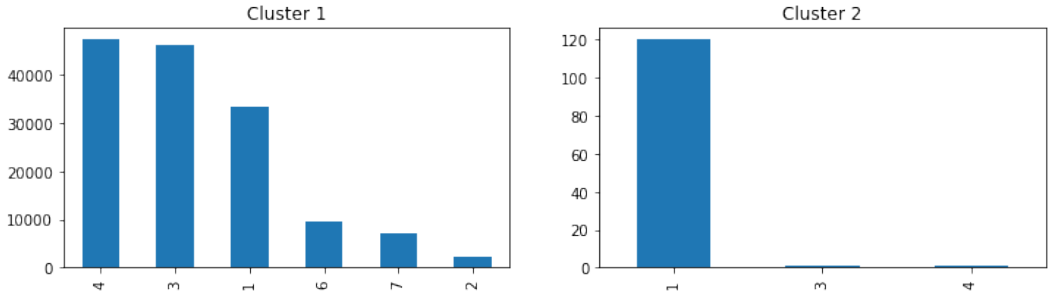

Figure 6. Bar graphs of the basins frequency for DBSCAN clusters.

represented, but shows a dominance of positive values. The boxplot shows that PCCD and MLRH are mostly spread in the negative, while BTWM and HDIV dominated mostly the positive values. The cluster 3 shows the clearest tendency, with a PLND variable largely dominated by positive values, and THDV being mostly spread in negative domain.

4.2. DBSCAN

The graphic analysis from the graph of the k-nearest neighbors (Fig.4) helped choosing 1.1 as value for the ”eps” hyperparameter. The metric kept for the distance is ”euclidian” because ”manhattan metric generated too much clusters and noise points. After many tries, the used value for the min_samples parameter was 60, who gives 2 clusters, 68665 noise points and 0.067 as silhouette score. This model seems to do even worse than the KMeans on the dataset, the results could therefore not be very reliable.

The algorithm put almost all the points in the first cluster, and around 120 points from the N.Atlantic basin in a second cluster (Fig.6). The boxplot (Fig.7) shows that the first cluster have an even distribution of all the variables. The second cluster shows high values for MSLP and THDV, spread around zero for HDIV and MLRH, and negative dominating values for all the other variables.

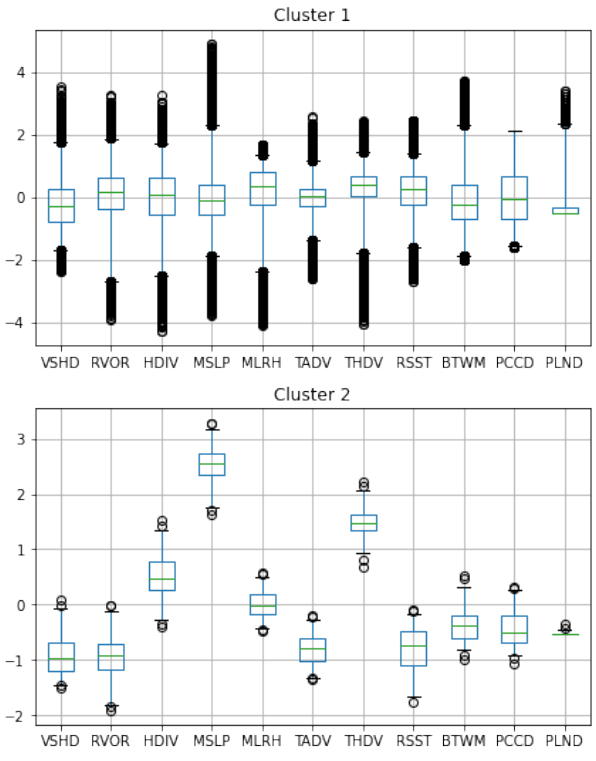

Figure 7. Boxplots for the environmental predictors of DBSCAN clusters.

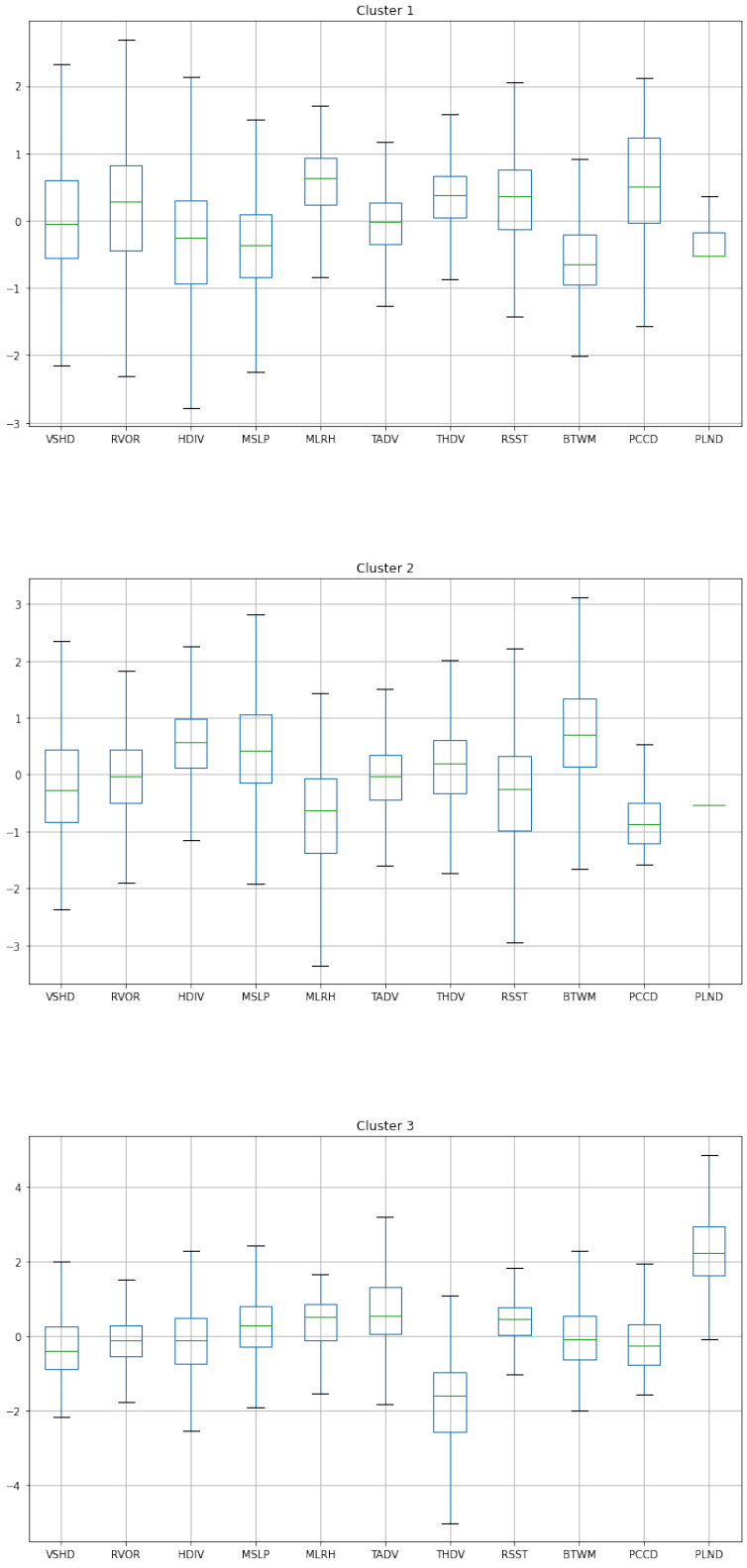

Figure 8. Boxplots for the environmental predictors of KMEans clusters.

5. Discussion

The results from the basins classification doesn't seem to show any clear tendency. Only the N.Atlantic, N.W. and N.E Pacific appears to be grouped more often with the KMeans. But they have also more points in the dataset, so it could be why they are more likely to be present in the clusters. The small group from N.Atlantic basin separated by the DBSCAN have differents environmental conditions than the other points, and it can explain why it is separated. It has high values of mean sea level pressure and vertical instability parameter. It could represent a small group of cyclones formed with those particular conditions. The environmental conditions for cyclogenesis are very precise and all the points from the dataset come from the intertropical zone. Therefore, the global conditions of formation are very similar and it could be why the two clustering algorithms tend to group all the basins together. Also, the structure of the dataset could not be very favorable for clustering algorithms.

Environmental predictors analysis showed some differences between the KMeans clusters. The first groupe have more positives values for mid relative humidity and sustained deep convection, and more negative values for mid- to upper-level moisture. By contrast, the second group have more negative value for mid relative humidity and sustained deep convection, and more positive values for mid- to upper-level moisture. This could shows two group of tropical cyclones formed with opposite conditions of humidity, convection and moisture. The variable representing the percentage of land coverage close to the area, who wasn't well represented in the two first group, shows a large domination in the positive values from the third group. There is also a domination of the negative values for the average vertical instability parameter, and higher value for the average 850 hPa horizonal temperature advection. This group could put together cyclones formed closer to land covered area with a tendency for negatives vertical instability parameter and high horizontal temperature advection.

6. Conclusion

The KMeans algorithm did better than the DBSCAN algorithm to the classification of the cyclogenesis true positive datapoints, although either having very high performance scores on this dataset. The similarities in the formation of the cyclones seem to depend more on their environmental conditions rather than on the location of their formation. Further analysis with other algorithms or parameter could bring different methodologies and point of views to develop the analysis. A comparison with the true negative points could also be instructive to understand the differences with the environmental conditions at which tropical cyclones don't form.

Links

- **Dataset:** https://unils-my.sharepoint.com/personal/milton_gomez_unil_ch/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fmilton%5Fgomez%5Funil%5Fch%2FDocuments%2FJeremie&ga=1
- **Code:** https://github.com/jejefragniere/MLEE_final_project/blob/main/Dev/MLEE_final_project.ipynb
- **TCFP image:** https://rammb.cira.colostate.edu/projects/gparm/images/TCFP_basins.gif

References

[1] Brian H. Tang, Juan Fang, Alicia Bentley, Gerard Kilroy, Masuo Nakano, Myung-Sook Park, V.P.M. Rajasree, Zhuo Wang, Allison A. Wing, Liguang Wu, Recent advances in research on tropical cyclogenesis, Tropical Cyclone Research and Review, Volume 9, Issue 2, 2020, Pages 87-105, ISSN 2225-6032, https://doi.org/10.1016/j.tcrr.2020.04.004.

[2] Emanuel, Kerry & Center, Lorenz. (2018). 100 Years of Progress in Tropical Cyclone Research. Meteorological Monographs. 59. 10.1175/AMSMONOGRAPHS-D-18-0016.1.

[3] Schubert, Erich & Sander, Jörg & Ester, Martin & Kriegel, Hans & Xu, Xiaowei. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. ACM Transactions on Database Systems. 42. 1-21. 10.1145/3068335.

[4] Schumacher, A. B., M. DeMaria, J. A. Knaff, 2009: Objective Estimation of 24-h Probability of Tropical Cyclone Formation. Wea. Forecasting, 24, 456-471.

[5] Tanmay Mane. (2020). Jupyter Notebooks - NearestNeighbors to find optimal 'eps' in DBSCAN. https://www.kaggle.com/code/tanmaymane18/nearestneighbors-to-find-optimal-ep -in-dbscan Accessed 22 December 2022.

[6] Zehnder, Joseph A.. "tropical cyclone". Encyclopedia Britannica, 4 Oct. 2022, https://www.britannica.com/science/tropical-cyclone. Accessed 20 December 2022.