Combining Unsupervised and Supervised Learning Techniques for Prediction and Analysis of Rhône's Plume Shape

Daniel Gönczy, Paula Dolores Rescala, María Isabel Ruiz
CS-433: Machine Learning, EPFL, Switzerland

***Abstract*—The plume of the Rhône River plays an important role in the understanding of the environmental processes taken place in Lake Geneva. In this project, we try to understand its shape, extract shape patterns and, finally, conclude with the features that may determine that shape. For this reason, in the first part of the project we will apply several unsupervised clustering algorithms and confirm the results using neural networks and, in the second part, we will perform an image classification based on meteorological and hydrological data.**

I. Introduction

In marine and aquatic coastal environments, river plumes are major transport mechanisms for particulate matter, nutrients and pollutants. Understanding their dynamics is therefore crucial for environment management concepts that ensure the good ecological functioning of coastal systems.

In Lake Geneva, the Rhône River is the principal source of water and sediments for the lake, accounting for 68% of the total water discharge and particulate matter input [1, 2]. After plunging, the Rhône River intrusion into Lake Geneva develops as an interflow [3, 4, 5, 6], i.e. the intrusion follows the depth at which the inflow and lake densities are equal. While the fate of the Rhône intrusion is already documented in the literature, very little is known about the plunge region itself (less than 300 m from the river mouth). The plume region is however of interest since the initial river plume evolution in the nearfield is often important for the subsequent plume spreading in the lake. For instance, detrainment of inflow waters with surface waters may develop during the plume, inducing a change in the inflow density. Based on remote thermal imagery which is a valuable tool to explore cold river plume characteristics, providing information at high temporal and spatial resolution, the present study aims at:

- Determining the main shapes and patterns of the Rhône plunge in Lake Geneva
- Determining the occurrence of overflow and vortices features in the plunge nearfield region

The outcomes of this study are precious to the personnel of the ECOL group as they plan future field work to study the seasonal variability of the Rhône river plunge dynamics.

II. Methods

A. Data exploration

The raw data on which the project is based consists of about $711'000$ grayscale pictures of 640×512 pixels, taken during the last three years. Each image is contained in a folder corresponding to the day at which the picture was taken. Data exploration has shown that each picture of the Rhône's plume is taken with a slightly different angle. Moreover, the number of images taken each day varies as there are some days in which no pictures have been taken and some days in which up to $1'440$ pictures have been taken.

B. Data pre-processing

Prior to beginning our machine learning techniques for detection of the plume shape and features, we pre-processed and refined the data to obtain a better representation of the images.

1) Data binning: The images have been binned to reduce the resolution by a factor 4 in each dimension, down to a resolution of 160×128 pixels. This has been done since the plume's shape is still clearly visible while allowing to reduce the size of the dataset and therefore the number of input parameters of the machine learning algorithms used.

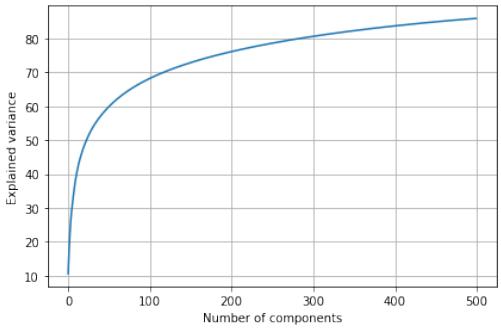
2) Filtering of bad images: Bad images such as images containing clouds covering the Rhône's plune have been removed. The filtering process is done by applying the Roberts edge detection filter to the input image, converting the resulting images into binary images and deciding to discard or to keep it based on the percentage of black pixels.

3) Data normalization: We are most interested in the shape and features of the plume in each image. However, with images taken at different times of each day, the brightness differs in each image and can affect our future algorithms when in reality the background colors are not important. Thus, we attempt three different kinds of data normalization, namely histogram flattening, adaptive histogram flattening, and standardization. We analyze our images in grayscale pixels which by construction have values ranging from 0-255. When normalizing data through histogram flattening, images with a narrow range of intensity values are transformed into a wider distribution and ultimately all images reach a similar level of brightness and contrast. With standardization, we simply

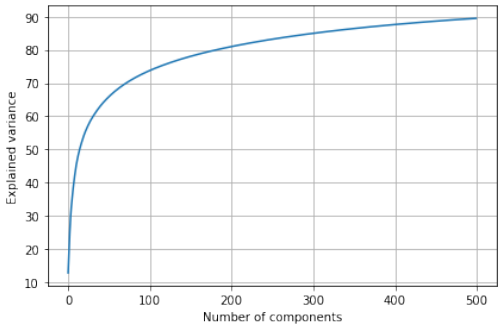
subtract the mean value and divide by the standard deviation in each image so all images ultimately have the same normal distribution.

4) Edge detection: Again, in each image we are most concerned with detecting the shape of the plume. Thus, edge detection in each image is valuable in emphasizing the shape of the plume by detecting lines of high contrast and giving more importance to this than other features such as color and brightness, which still vary to some extent after data normalization. We apply the Roberts cross and Sobel operators separately for edge detection in our images and ultimately try both configurations of the images to see which edge detection algorithm yields better results.

5) Principal Component Analysis (PCA): Our final data preprocessing step is applying PCA for dimensionality reduction of each image. With each image containing 160x128 pixels to begin with, this amounts to 20,480 features per sample, which is far too many features to use in any analysis since it would yield an underdetermined feature matrix. Thus, we apply PCA to reduce the dimensionality of each image to 500. We decide 500 features is a good stopping point after plotting the ratio of explained variance versus number of components in PCA as seen in figure 1 and determining that 500 components is a good trade-off for keeping the dimensions as low as possible without sacrificing too much important information.



(a) Images with filtered with Roberts edge detection filter.



(b) Images filtered with Sobel edge detection filter.

Fig. 1: Explained variances of the PCA.

C. Unsupervised Learning

Since the goal of the project is to determine the possible shapes of the plume from a large dataset of images with no known labels, we decide to apply unsupervised learning techniques. In particular, we use clustering methods to group together images in which the plume displays similar behavior. We attempt three different clustering algorithms to accomplish this task: Kmeans clustering, DBSCAN, and Gaussian Mixture Models.

1) Kmeans: we run the Kmeans clustering with different values of the hyperparameter *number of clusters*. For each execution of the Kmeans algorithm we evaluate its performance using the Silhouette score function.

2) Gaussian Mixture Models: we follow the same procedure described before.

3) DBSCAN: We run the algorithm with several different values of the hyperparameters and allow it to determine the number of clusters.

D. Supervised Learning: Image Classification based on Image Data

After having clustered our images into distinct groups based on plume shape, we now have labels associated with each image and can apply supervised learning techniques for robustness of our results.

In order to easily classify future images, we design a convolutional neural network which accepts a non processed image and classifies it by plume shape.

In particular, we train two neural networks: one using the filtered $5'000$ training subset together with the labels generated by the Kmeans unsupervised clustering algorithm for training and another one using the same images but the labels generated by Gaussian Mixture Models unsupervised clustering algorithm.

Our convolutional neural networks are divided into two parts:

- In the first one, we apply five convolutional layers each of them follow by a max pooling layer.
- In the second part, we just apply several dense layers.

E. Supervised Learning: Image Classification based on meteorological and hydrological data

After having clustered our images into distinct groups based on plume shape, the second part of the project is to unravel the relationship between the plume's shape and the underlying hydrologic parameters of the river and the local weather data. Those parameters were implemented as features to cluster the corresponding image taken at that time into a defined shape. This image clustering was then compared to the clustering of the unsupervised learning of Part I as a metric of classification efficiency.

1) Data gathering: The initial hydrological data given to achieve this part was the hourly flow rate of the Rhône near Vouvry, in Valais, between the 1^{st} of January and the 30^{th} of September. Since more hydrological features might be influencing the river's shape, such as the water level or temperature, it has been tried to gather such data.

However, the water level and temperature of the river is only relevant compared to the lake's temperature and level, which

is measured only once every two weeks. Moreover, the lake's data for 2020 is not available yet and Swiss hydrological data is not freely accessible either.

On the contrary, the weather data of Vouvry can easily be freely accessed online [7]. The relevant parameters that were retrieved are the air temperature, the wind speed and the wind direction.

2) Data pre-processing: A feature expansion was made to get the wind speed squared, since it seemed quite straightforward that the latter could correlate with the kinetic energy of the wind and thus give more accurate results. Also, since the data gathered spans throughout a year and a half, it has been decided to normalize the data with a mean and standard deviation corresponding to a year's worth of data. This allows to better capture the distribution of these parameters over a yearly period. Each image had in its filename the Unix time, which was used to link the latter with the closest time at which the parameters were measured. This means that the features are taken from the closest round hour at which the picture was taken.

3) Semi-supervised learning: Since the image classification algorithm used previously rely on both Gaussian mixture models and Kmeans, two sets of labels were used for the semi-supervised learning on this part as a metric to validate the clustering. Even though it might seem more evident to develop a convolution neural network to do supervised learning, it was desired not to rely on the previously found labels as ground truths since no absolute metric was available to ensure the clustering was optimal. Therefore, it was chosen to do semi-supervised learning to rely only on the features to classify the images and then to check the coherence of the results with the previously found image cluster labels. In addition, since this part was more of an extension of the project guidelines than part of the project itself, a choice had to be made between trying Kmeans and a convolution neural network on this dataset.

III. Results

A. Unsupervised Clustering Results

As previously mentioned in Section II-C, clustering performance was evaluated with the Silhouette Score and sum of squared errors for a range of values of k, the hyperparameter setting the number of clusters for Kmeans and Gaussian Mixture Models. Although DBSCAN was attempted as well, the algorithm produced clearly incorrect clusters by either putting all images into one cluster or all into separate clusters. Even after attempting different values of the hyperparameters, we could not obtain better results. Thus, we discarded this algorithm as a valid option for our project and believe it is because of the uneven densities of our clusters.

Although we measured performance with the silhouette score, however, we were unable to get good results or determine the best number of clusters from this test. We tested a range of k values from 2 to 20, since with some domain knowledge we expected around 4-6 clusters. However, no number of clusters gave a good silhouette score (with all below 0.15 whereas close to 1 is ideal) no matter how we preprocessed our data or which edge detection operator was used. Figure 2 shows the plot of silhouette scores and sums of squared errors. We expected to observe a clear global maximum somewhere around the expected number of clusters which would depict the best hyperparameter value, but we observed the opposite in fact. Thus, instead of using this test, we selected $k = 4$ from domain knowledge of expected clusters to see what that performance would look like. To visualize the final clusters, we reduced the dimensions again of the images using T-distributed Stochastic Neighbor Embedding (tSNE) to 2. Figure 3 shows the visualization of the clusters in this two-dimensional space.

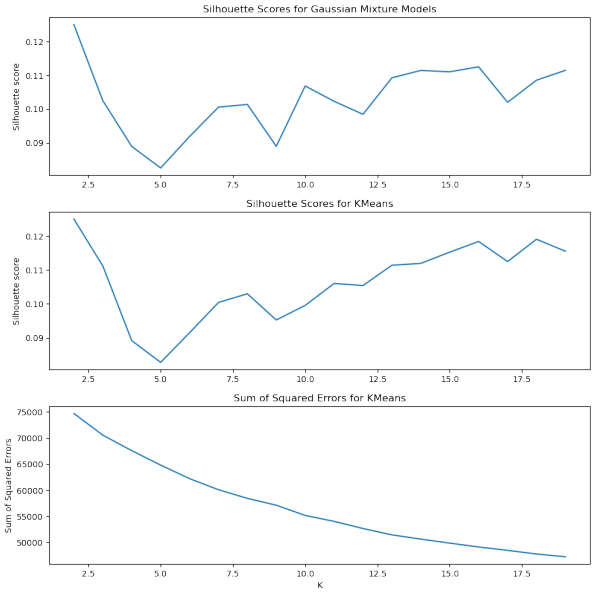


Fig. 2: Plots of evaluation metrics for clustering techniques. Silhouette score versus number of clusters in the top two plots and Sum of Squared Errors in the last plot. First plot is for Gaussian Mixture Models whereas the second and third are for Kmeans.

B. Neural Network Results

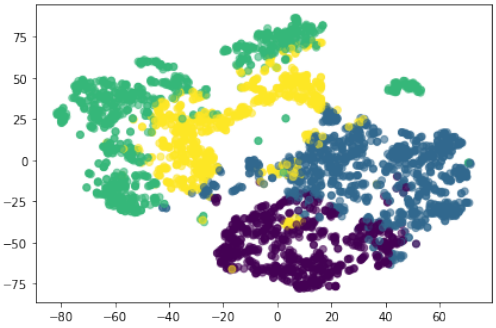
In Figure 4 we can see the accuracy obtained for the training examples when predicting with our neural networks in both cases. As we can see, the results are better using the labels obtained using the Gaussian Mixture Models unsupervised clustering algorithm.

The accuracy levels show that we may improve the clustering or the neural network architecture so that the Neural Network predicts the same when applying it to the clustered images (training images of the neural network).

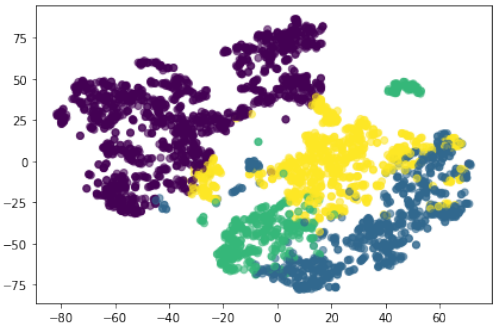
This issue may be solved by adding more convolutional layers or by training with more images (i.e. using a bigger dataset for the training process).

C. Semi-supervised learning Results

The Kmeans algorithm was both applied on the whole feature dataset and then solely with the flow rate, since it



(a) Unsupervised clustering with Kmeans (4 clusters).



(b) Unsupervised clustering with Gaussian Mixture Models (4 clusters).

Fig. 3: Unsupervised clustering results.

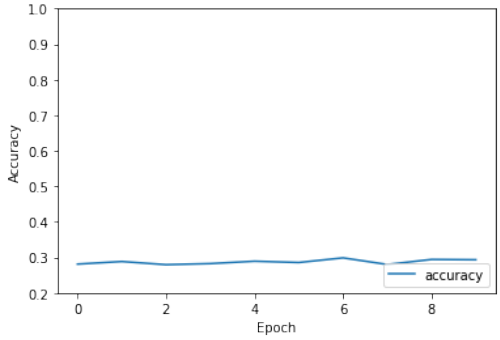
is expected that it is by far the leading parameter in the classification of the plume's shape. The adjusted rand index, ARI, provides a good metric of the coherence between the labeled clusters from image classification and the determined clusters based on the physical data. On the other hand, the silhouette score gives a good insight at the clustering efficiency of the kmeans algorithm in this context. We can see that only using the flow rate allows for better separated clusters at the cost of bad ARI. This means that the flow rate is a dominant feature but that it is clearly not sufficient to cluster the data well due to underfitting. Also, the kmeans with all the features has a slightly better ARI but it is still quite bad. This is partly due to the fact that the labels themselves do not depict well the physical underlying clusters on the images, as we have seen that results are not stellar.

TABLE I: Result metrics of the Kmeans clustering with respectively all the features and solely the flow rate of the river, as it was expected that this is the most relevant feature.

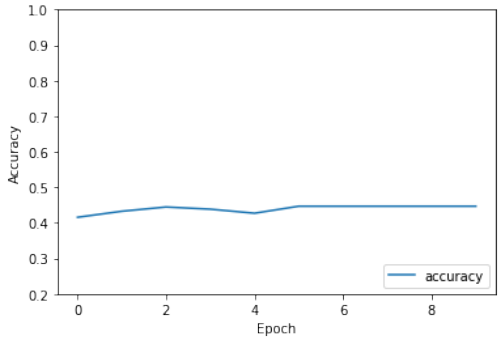
	ARI, Gaussian Mixture Models	ARI, Kmeans	Lowest SSE	Silhouette Score
Kmeans, all features	0.1182	0.1166	4929.64	0.2675
Kmeans, only flow rate	0.0478	0.0807	183.60	0.5611

IV. Discussion

Our objective in this project is to predict the Rhône's plume shape. With a wide variety of machine learning clustering methods to try, it takes time to decide which unsupervised method use. We ultimately chose a Gaussian Mixture Model



(a) Accuracy obtained with the Convolutional Neural Network trained with the Kmeans labels.



(b) Accuracy obtained with the Convolutional Neural Network trained with the Gaussian Mixture Models labels.

Fig. 4: Accuracy obtained by our Convolutional Neural Networks.

run on purposefully preprocessed data and tuned hyperparameter (number of clusters) to achieve a good clustering of the images.

Furthermore, we have used this clustered images to train a convolutional network with the aim of testing the consistance of our results. Unfortunately, the results obtained were not as good as expected as the accuracy level obtained is low. However, this accuracy levels may be improved by changing the architecture of the neural network or by improving the clustering process using new machine learning techniques.

Ultimately we were able to predict the shape of the plume by only using meteorological and hydrological data establishing a relationship between the underlying hydrologic parameters of the river and the local weather data and the plume's shape.

V. Conclusion

In conclusion, we achieved to filter efficiently the taken images in order to get rid of images of bad quality or with a consequent cloud cover. Then, the fact that the unsupervised learning gave subpar results was partly due to the lack of a ground truth metric to evaluate the clustering. Future work could try to improve this by labeling the data and run supervised or semi-supervised learning on that aspect. Finally, the physical data semi-supervised learning could be improved by adding more features such as hydrological data and try a

fully supervised learning algorithm to see if the results could improve.

REFERENCES

[1] P. Burkard. "Hydrologie - Bilan hydrologique". In: *Commission Internationale pour la Protection des Eaux du Léman (CIPEL), Lausanne: Le Léman Synthèse 1957–1982* (1984), pp. 43–48.

[2] P. Zahner and J.-P. Vernet. "Dynamique du système lacustre". In: *Commission Internationale pour la Protection des Eaux du Léman (CIPEL), Lausanne: Le Léman Synthèse 1957-1982* (1984), pp. 55–63.

[3] F.-A. Forel. "Les ravins sous-lacustres des fleuves glaciaires". In: *Comptes Rendus de l'Académie des Sciences de Paris* 101 (1885), pp. 725–728.

[4] J. Dominik, D. Burrus, and J. P. Vernet. "A preliminary investigation of the Rhône River plume in eastern Lake Geneva". In: *J. Sediment. Petrol.* 53 (1983), pp. 159–163. DOI: 10.1306/212F817A-2B24-11D7-8648000102C1865D.

[5] F. Giovanoli and A. Lambert. "Die Einschichtung der Rhône im Genfersee: Ergebnisse von Strömungsmessungen im August 1983, Schweiz". In: *Z. Hydrol* 47 (1985), pp. 159–178. DOI: 10.1007/BF02551939.

[6] J. Hadler et al. "Application of δ^{18}O, $\delta^{18}C_{DIC}$, and major ions to evaluate micropollutant sources in the Bay of Vidy, Lake Geneva". In: *Isotopes Environ. Health Stud.* 52 (2014), pp. 94–111. DOI: 10.1080/10256016.2014.971786.

[7] *World weather online for developers*. URL: https://www.worldweatheronline.com/developer/premium-api-explorer.aspx (visited on 12/12/2021).