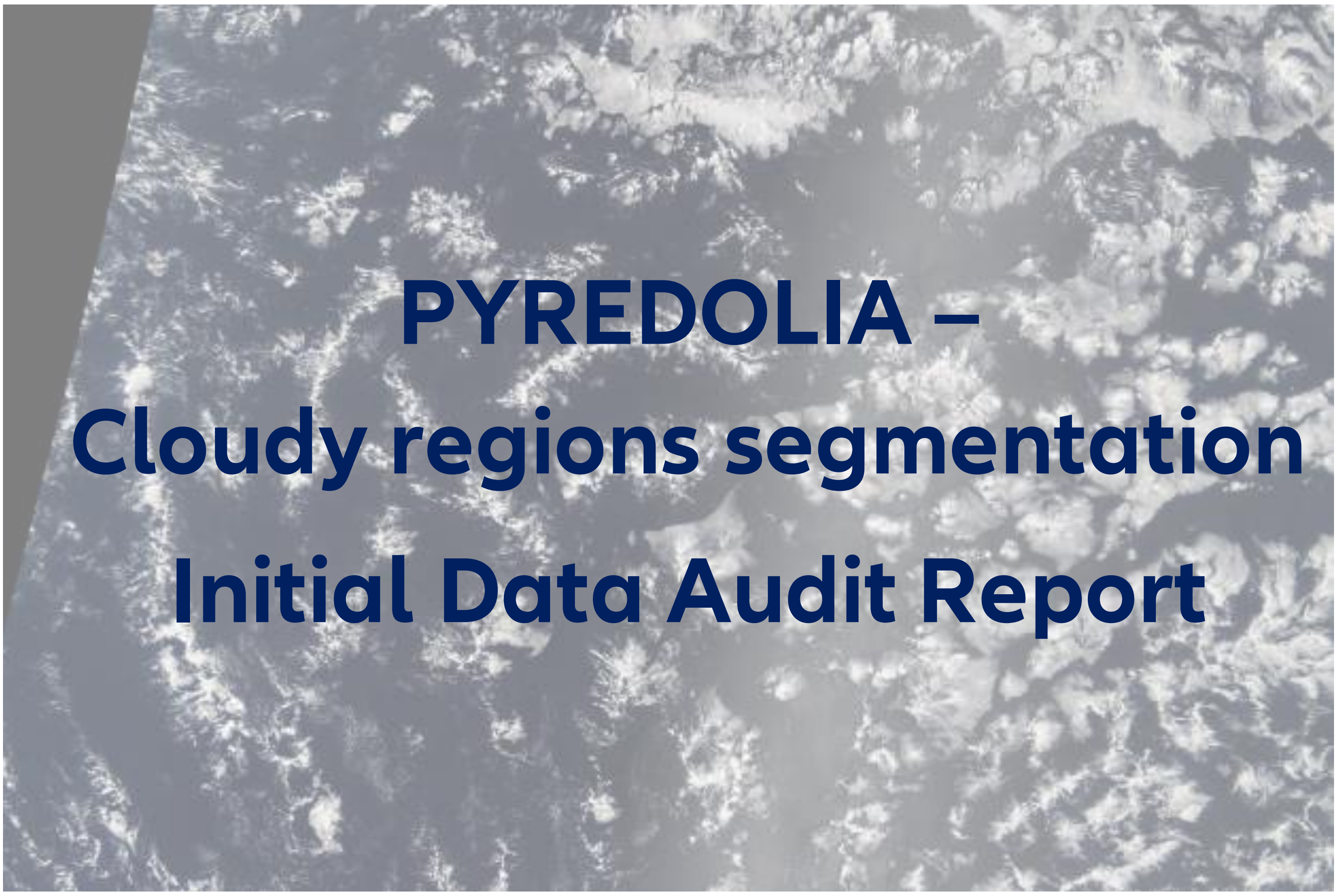


PYREDOLIA – Cloudy regions segmentation Initial Data Audit Report

Authors: Thomas Bury, Afonso Alves, Daniel Staudegger

Date: 16. December 2021

1. Introduction to Problem statement

The objective of this project is to use Deep Learning techniques to develop a model which can classify cloud organisation patterns from satellite images. The project has initially been posted as a Kaggle challenge[1], and was related to a scientific research project[2] of TU Munich, Max Planck Institute for Meteorology in Hamburg, and Sorbonne University in Paris. In particular, the goal of the project is to find a model which recognizes one of four pertinent cloud patterns "Sugar, Flower, Fish and Gravel" (see Appendix 1 for Examples).

2. Data Collection

The scientist downloaded a data set of around 10,000 satellite images from 2 different satellites (Terra and Aqua MODIS) from NASA Worldview over a period from 2007 to 2017 and merged these 2 images into one to cover a wide enough geographic area for studying cloud patterns. Figure 1 shows the regions from which the images were selected, whereby the region east of Barbados ([1] below) was chosen as it is a well-studied region on clouds and climate, and the other regions were chosen due to their climatological similarity to region [1].

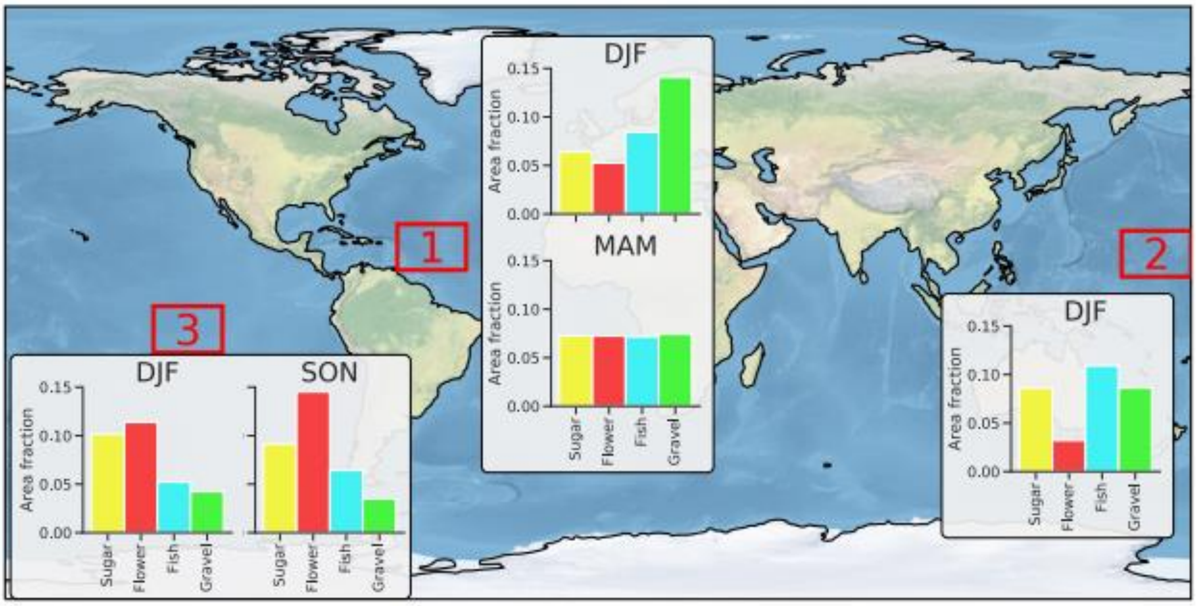


Figure 1: World map showing the three regions selected for the Zooniverse project. Bar charts are showing which fraction of the image area was classified into one of the four regions by the human labelers. Note that the areas do not add up to one. The remaining fraction was not classified.[3]

In order to obtain a training data set for a machine learning model, the scientists then used crowdsourcing and asked 67 scientific experts across the 3 above-mentioned research institutes to manually label the collected satellite images by drawing rectangular, labelled boxes on the images to indicate the respective cloud pattern identified. On average, each image was classified by 3 analysts, and analysing the labelled regions the scientists conclude that "while certainly noisy, clear examples of what was defined as Sugar, Flower, Fish and Gravel could be robustly detected"[4]. Combined with their meteorological expertise and appearing correlation of the labelled regions with data points of prevalent meteorological conditions the scientist conclude that "despite the noise in the labels, there was sufficient consensus between the participants on clear features to warrant further analysis."[4], which gives us comfort that we can use the labelled data set for our model.
For additional details on the data collection and labelling we refer to the scientific paper.

[1] https://www.kaggle.com/c/understanding_cloud_organization/

[2] https://arxiv.org/abs/1906.01906

[3] Figure 2 of Rasp et. al. (https://arxiv.org/pdf/1906.01906.pdf)

[4] https://arxiv.org/pdf/1906.01906.pdf

3. Meta-Data analysis of the (training) data

In the project, the data is already split into a set of 5,546 (labelled) training images and 3,698 test images. We start by a meta-data analysis of the images and perform some checks on the quality.

In a first step, we check the resolution (height and width) of the images as well as check whether they are all indeed colour images. Our analysis confirmed that all the training images are in colour and have the same dimensions (2100 x 1400 pixels).

Next, we want to analyse the contrast of the images. To do this, we first convert the images into a grayscale image, and then compute the contrast by computing the average variances of the pixel values to obtain a measure for the image sharpness. By doing so we obtain the following distribution, which indicates heavy tails that could potentially point to outliers. This is confirmed by looking at the Boxplot of the image sharpness (see Figure 3), which shows outliers on both ends of the spectrum.

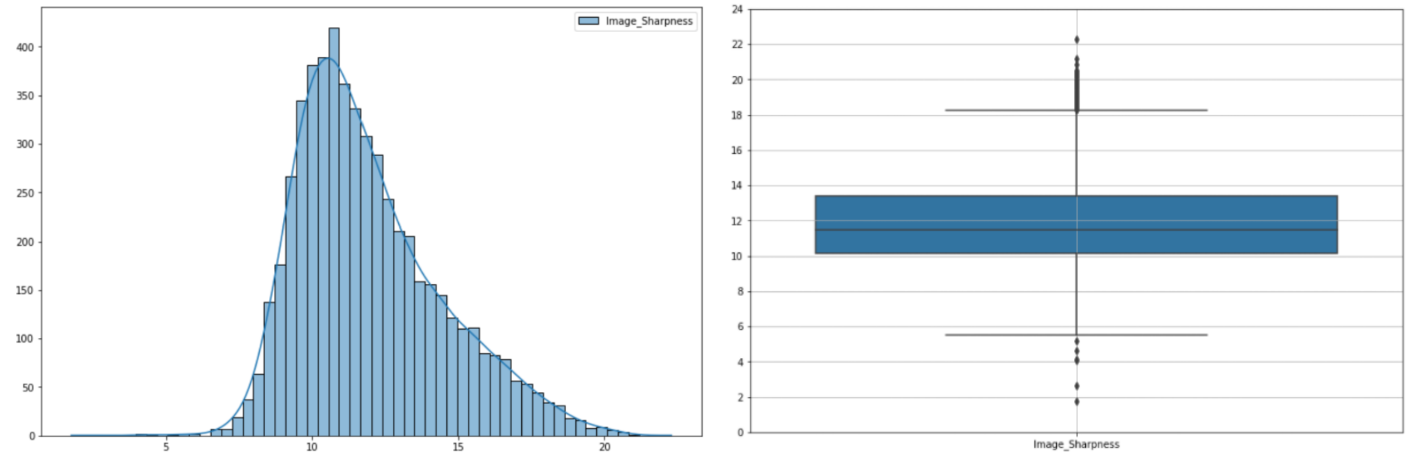


Figure 2: Histogram with density estimate, and Boxplot of the image sharpness.

Looking more closely at these outliers, we observe that those with a low sharpness have a large black stripe in the image (coming from the fact that the pictures were taken from the 2 satellites which were following different trajectories, and not the full geographic region of an image was covered by the merge of the 2 pictures), whereas the outliers with a high sharpness have a very small black stripe and a large area of the image is covered by clouds. See Figure 3 below for examples of these images.

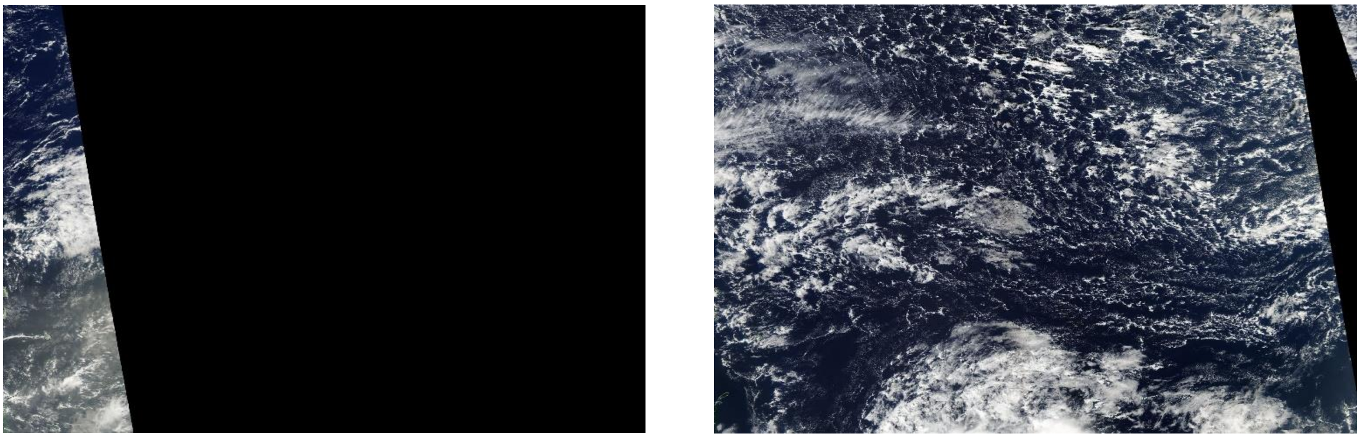


Figure 3: Examples of images with low sharpness (left) and high sharpness (right).

Based on the above meta-data analysis, we conclude that the images correspond in size and general quality, the only point we need to consider is the size of the black stripe in the image. The sharpness indicator we computed can help us identify the pictures with a large area uncovered (i.e., black).

4. Detailed data exploration of the (training) data

Let us now move to studying the training data set in closer detail in order to better understand the data and identify potentially data errors and/or biases we need to deal with. We start by analysing the distribution, how many images we have for each category with at least one labelled segment corresponding to the respective category (see figure 4).

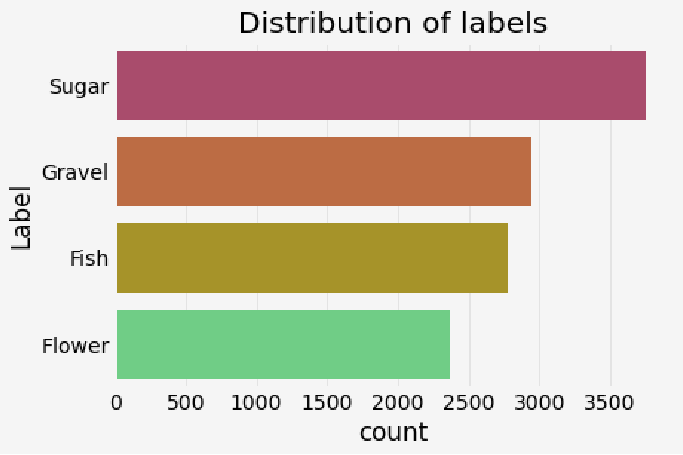


Figure 4: Number of images containing the respective category.

Figure 4 shows a slightly non-uniform distribution with "Sugar" being the category contained in images the most. We will keep this small imbalance in mind, as it might produce a small bias in our model.

In the second step, we analyse the distribution of the number of labels per image (ranging from 0 to 4).

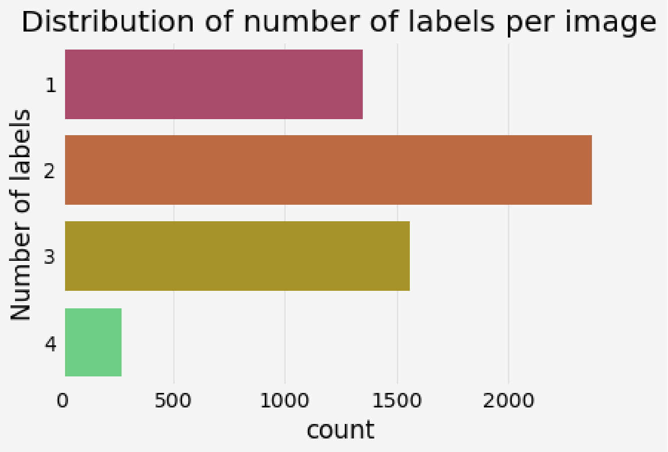


Figure 5: Number of labels per image.

We observe that each picture contains at least one label, and that most pictures contain 2 categories while having all 4 categories in one picture is quite rare. Looking at these combinations in more detail, we observe that "Sugar" is prevalent in 7 of the 8 most frequent patterns, which is in line with our above observation that Sugar is the most prevalent category.

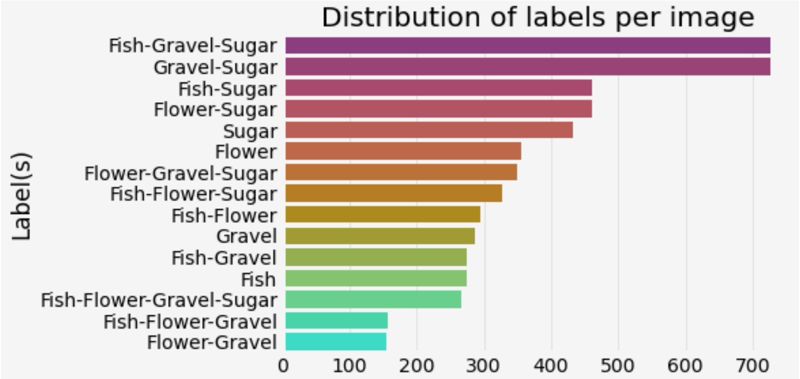


Figure 6: Distribution of label combinations per image.

We next move to better understand and the masks, which correspond to the segments of the images labelled by the scientists as containing one of the cloud patterns. These segments are encoded via a 1-d array of pixels corresponding to the respective cloud pattern segment. The masks in general are polygons and can have different shapes, as shown in an example image below, where we show only the segments that were labelled (here corresponding to a Sugar pattern) and blacken out the rest.

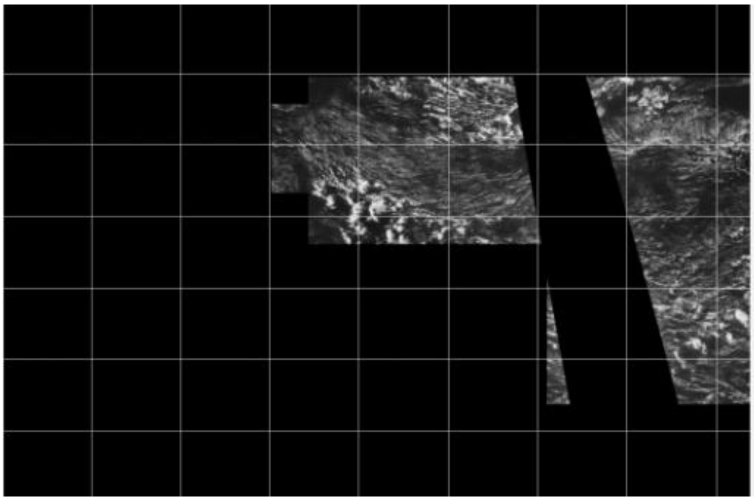


Figure 7: Example of a mask containing all labelled pixels of the Sugar pattern.

By taking the minimum and maximum pixel value per mask in each dimension, we can draw bounding boxes. An example for the respective masks and corresponding bounding boxes is shown in figure 8 below.

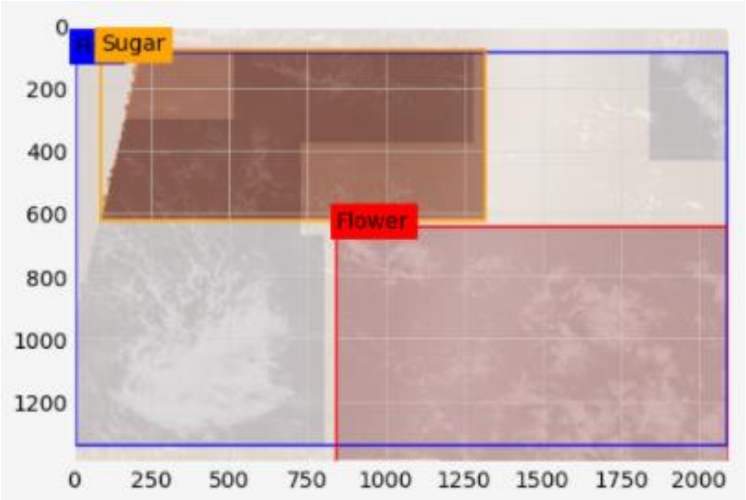


Figure 8: Masks (polygons) with the corresponding rectangular bounding boxes.

This example highlights a potential issue of using the bounding box as a target feature for modelling: since the box is built with the minimum and maximum x and y values, two or more cloud masks of the same label can create a box that encapsulates a high area not belonging to the actual mask. For example, in the case where there are 2 small cloud patterns of a specific category in each of the corners of the picture, then the masks can separate and show these 2 smaller segments, while the bounding boxes will cover almost the entire image.

The above suggest that masks are more accurate in spotting and distinguishing the regions in a picture of a given cloud pattern, while the bounding boxes would provide a simpler input to the model and are less segmented. We might try both options in our modelling, but we hypothesis that the masks will provide more accurate results for image segmentation.

A second observation we note from the above image is that an image can also have 2 segments per category. In order to take this into account, we add 2 additional analyses: First, we compare the average size of the picture covered by the respective masks by counting the pixels. Figure 9 below shows the distribution of the sum of the pixels of a given cloud category in the image.

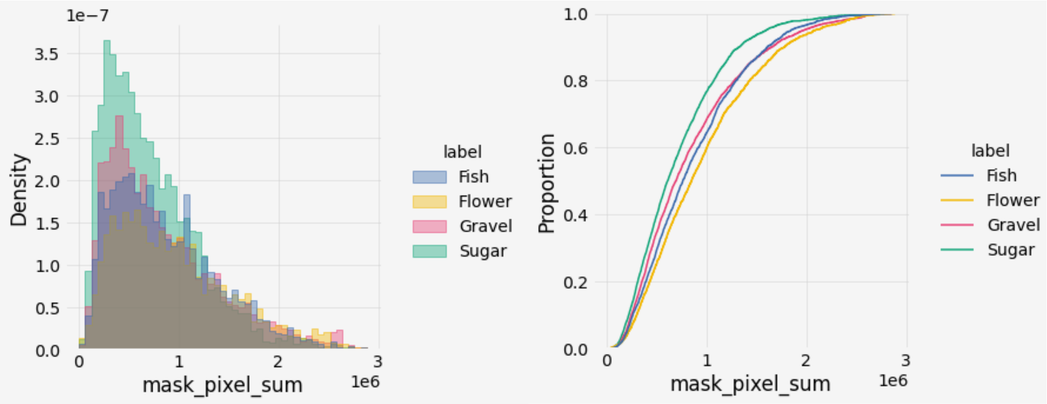


Figure 9: Distribution of the pixel sum (area covered per image) per cloud type.

It seems that the surface distributions are spanning the same range. For the “Sugar” it seems to be more likely that this mask in the image is spanning a smaller area compared to the other categories.

Second, a function was written to count how many different segments of a given category we have within a labelled image. The function utilizes two strategies to find and count different cloud structures: horizontal search - cloud structures that can be separated by a horizontal line, and vertical search - cloud structures that can be separated by a vertical line. For vertical search, the basic logic is to look for jumps in the starting pixels of each segment (high pixel numbers in the 1-d segment space correspond to high x values in the 2-d matrix space). For horizontal search, the function scans the y values for each x value of the mask, and infers the presence of one or more clouds. More concretely, it analyses changes in the y distribution from x_{i-1} to x_i. Some other elements were taken into account to increase the accuracy of the algorithm, namely: consider mask segments that take up the entire y range (mask that "wraps" around the figure); discard segments that have height lower than 35 pixels (corresponding to irregularities in the edges of the cloud mask and isolated pixels frequently found in the black stripe); up to 3 overlapping cloud structures (with the same x value).

To validate the feature Cloud_Count, some graphs with the masks where built and visually compared with the new feature value (5 for each Cloud_Count value). While in most examples the value and visual observation matched, some exeptions existed, namely in images with many clouds of a given type / irregularites in the shape. Some of the images used to validate the function can be found in the Appendix.

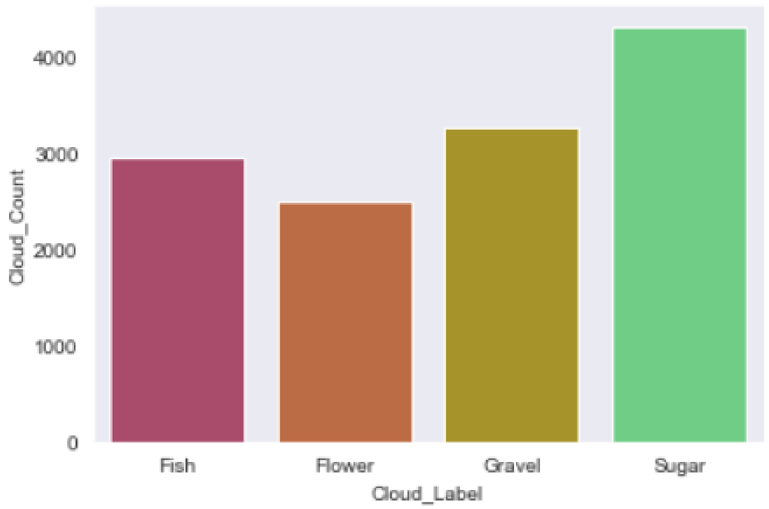


Figure 10: Distribution of the Cloud_Count total per cloud label.

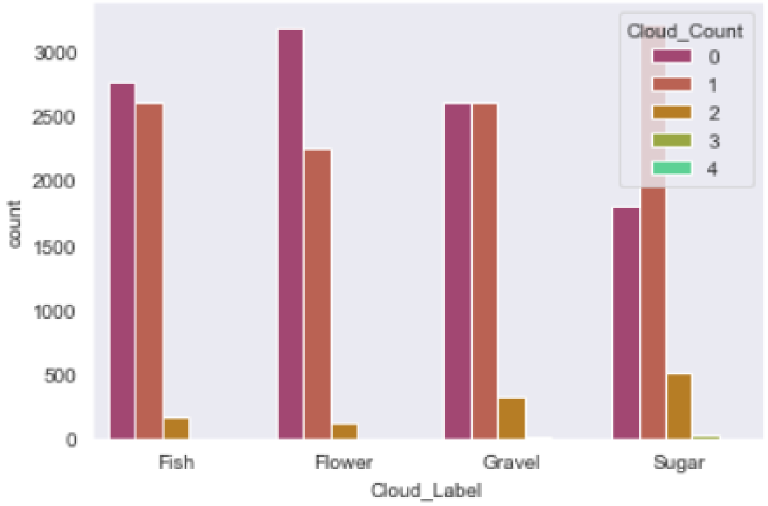
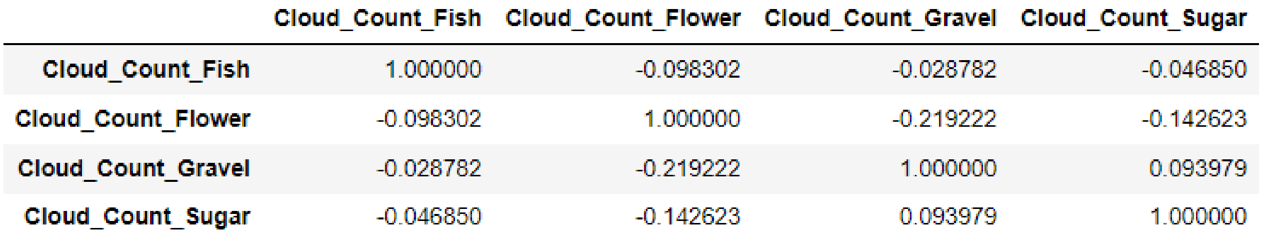


Figure 10: Distribution of each Cloud_Count value grouped by cloud label.

The above plots show similar results to previous visualizations: the distribution of Flower is more skewed to the left (lowest average value), while the Sugar distribution is more skewed to the right (highest average value).



	Cloud_Count_Fish	Cloud_Count_Flower	Cloud_Count_Gravel	Cloud_Count_Sugar
Cloud_Count_Fish	1.000000	-0.098302	-0.028782	-0.046850
Cloud_Count_Flower	-0.098302	1.000000	-0.219222	-0.142623
Cloud_Count_Gravel	-0.028782	-0.219222	1.000000	0.093979
Cloud_Count_Sugar	-0.046850	-0.142623	0.093979	1.000000

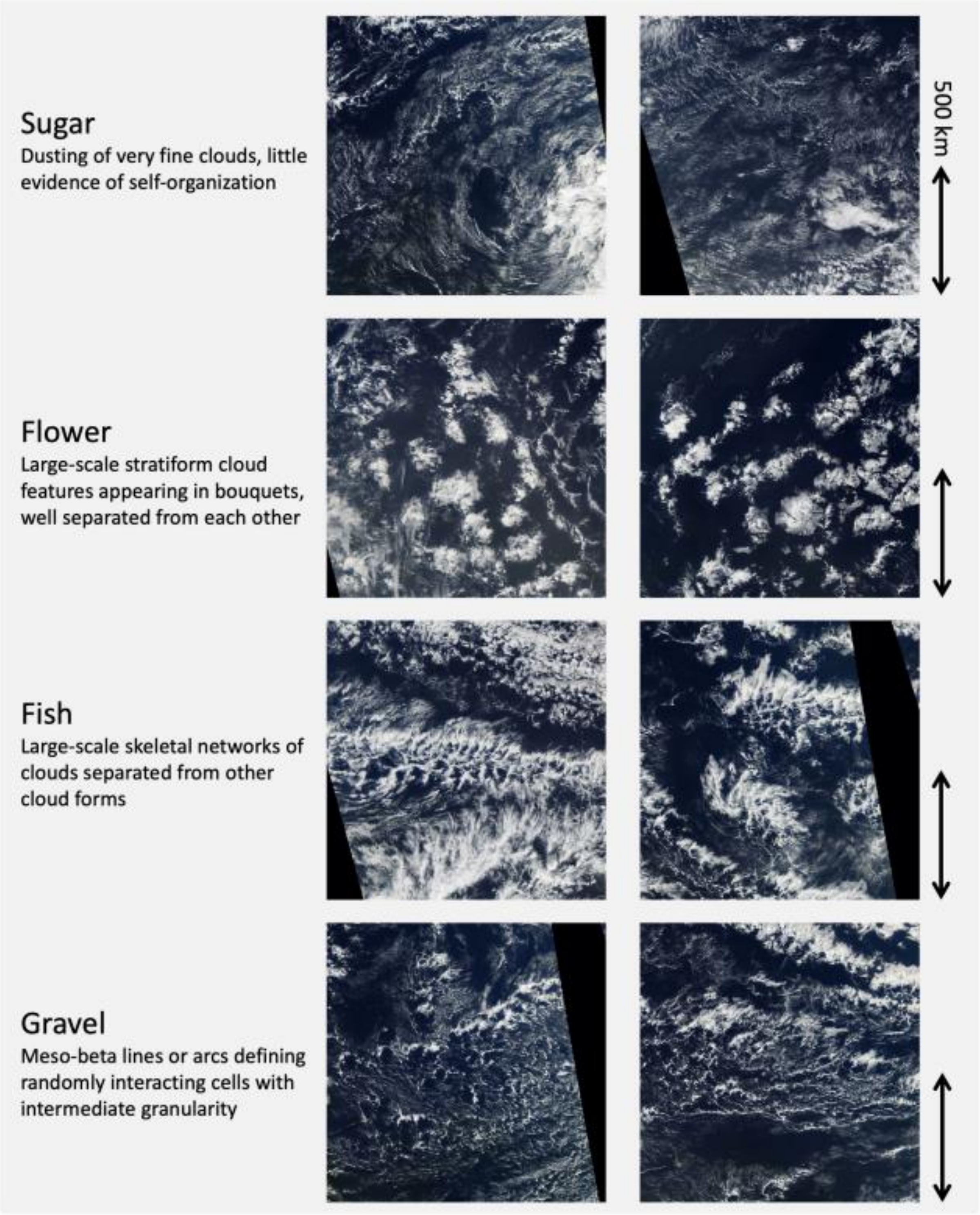
Table 1: Correlations between the different cloud types.

Table 1 presents mostly negative and near zero correlation values. The exception is the Sugar-Gravel correlation, which is low but positive. This might be explained by Figure 6: the Sugar-Gravel combinations are the most frequent ones.

A. Appendix

A.1 Cloud Patterns

See below canonical examples for images of the four cloud organisation patterns studied in the paper.


Sugar
Dusting of very fine clouds, little evidence of self-organization
Flower
Large-scale stratiform cloud features appearing in bouquets, well separated from each other
Fish
Large-scale skeletal networks of clouds separated from other cloud forms
Gravel
Meso-beta lines or arcs defining randomly interacting cells with intermediate granularity
500 km

Some masks used to validate the Cloud_Count feature:

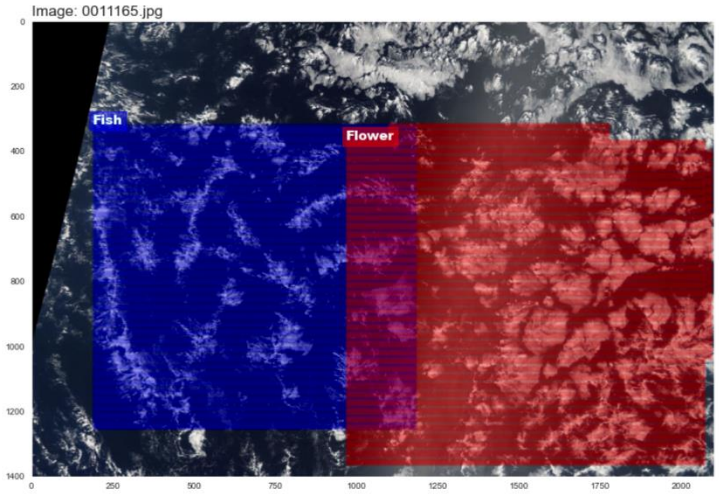


Figure 12: Image with expected Cloud_Count = 1 for Fish label.

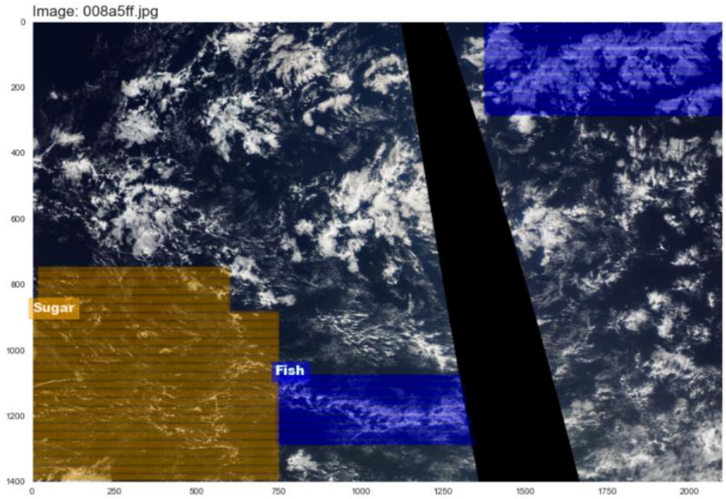


Figure 13: Image with expected Cloud_Count = 2 for Fish label.

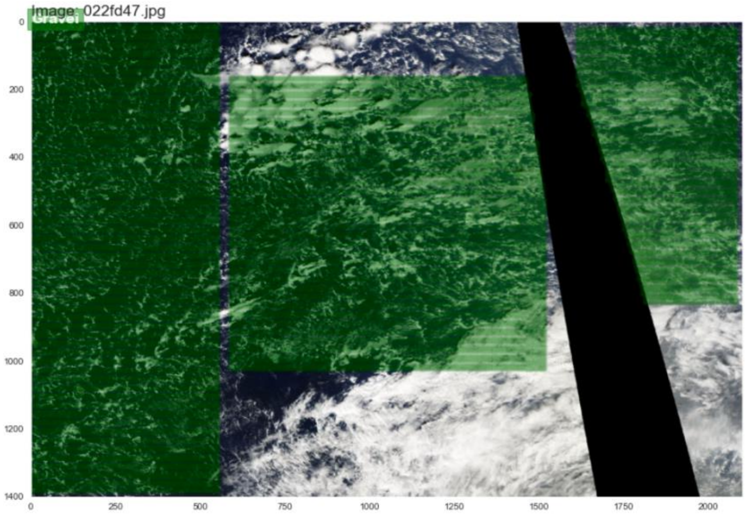


Figure 14: Image with expected Cloud_Count = 3 for Gravel label.

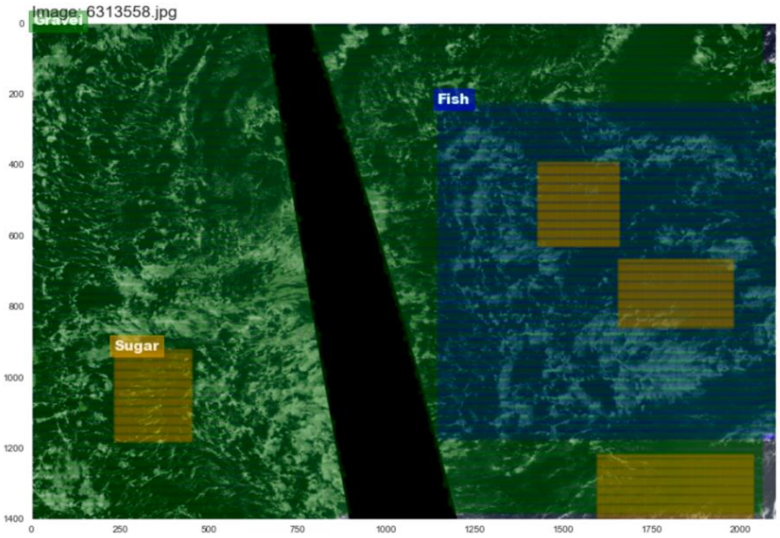


Figure 15: Image with expected Cloud_Count = 4 for Sugar label.