**Dillon Hicks**

### 7/20/2018

Engineers for Exploration

**Orfeo Toolbox: K-Means Clustering**

# INTRODUCTION

## PREREQUISITES

Orfeo Toolbox 6.6.0

QGIS 2.18.21

## NOTES

K-means is a clustering algorithm that categorizes pixels in an image into a user specified amount of classes. K-means is unsupervised, meaning that it doesn’t use any labeled data, instead it attempts to split data into different groups that are the most distinct from one another. With clearly different objects and regions in an image, we can possibly classify our areas of interest into the different classes that we want. This is most likely the easiest way to process your data, however K-means’ clustering is arbitrary, and may be to inaccurate for many uses.

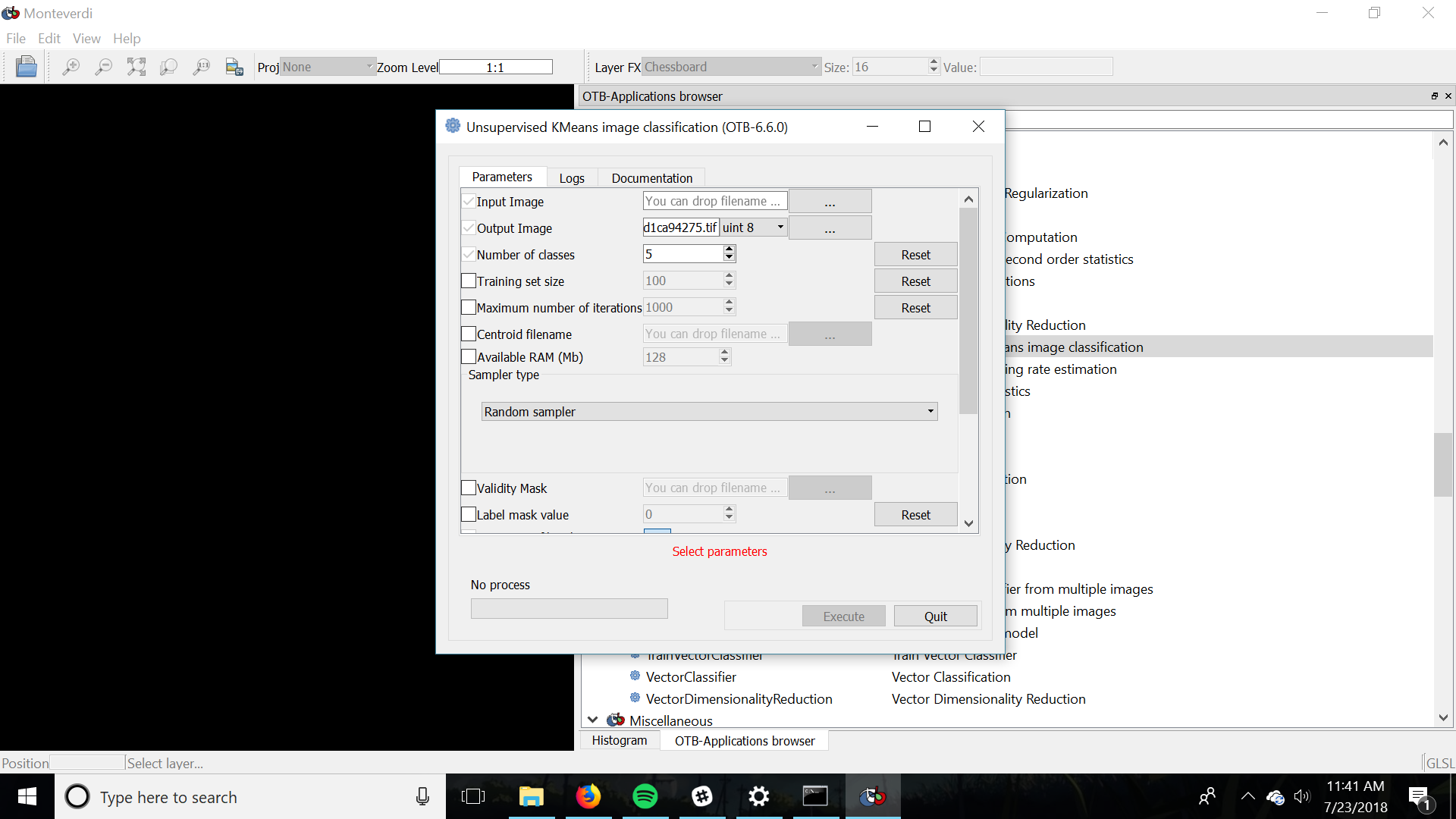
In this tutorial, you will learn how to use K-Means clustering in to classify the data and the QGIS sieve tool to smooth out the imagery of your data.

To start, ensure that you have the specified QGIS version installed and have Orfeo Toolbox (OTB) setup. If you need help with this, refer back to previous documentation (Orfeo Toolbox: Setting Up). This tutorial will be done in Monteverdi, however everything is found in the processing toolbox of QGIS once you have setup QGIS with OTB

To start, open Monteverdi and click on View and then on OTB-Applications browser (or press Ctrl-A)



Expand the Learning Section and then start the KMeans Classification Application (Unsupervised KMeans image classification).

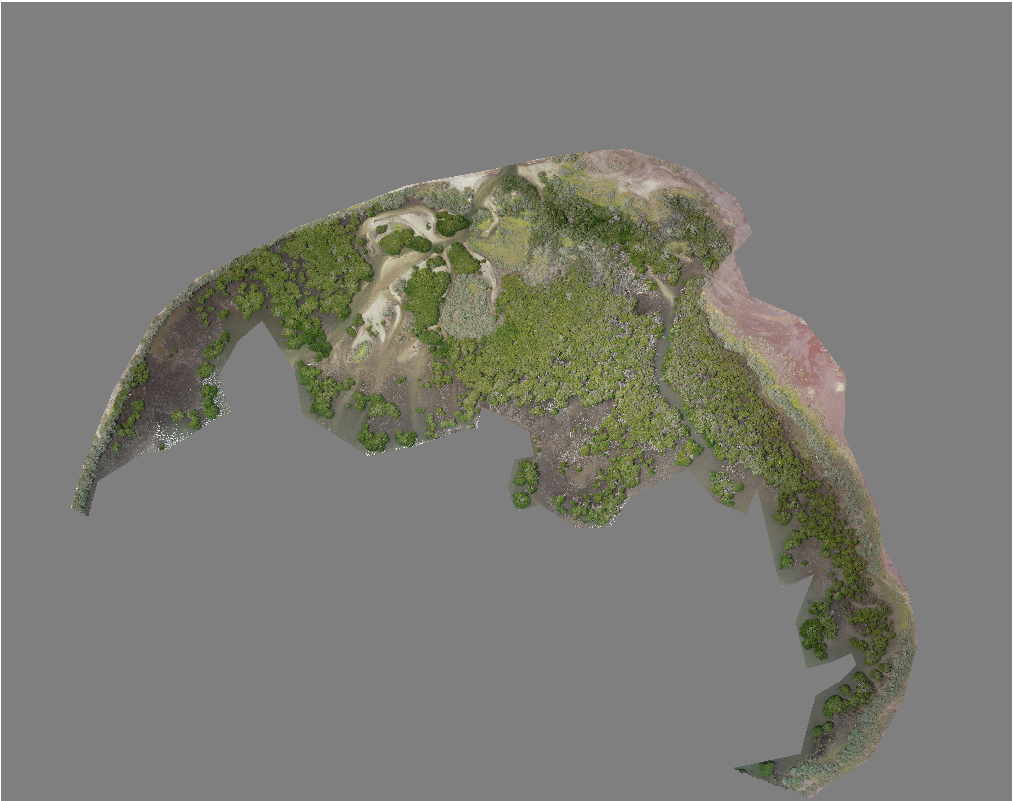
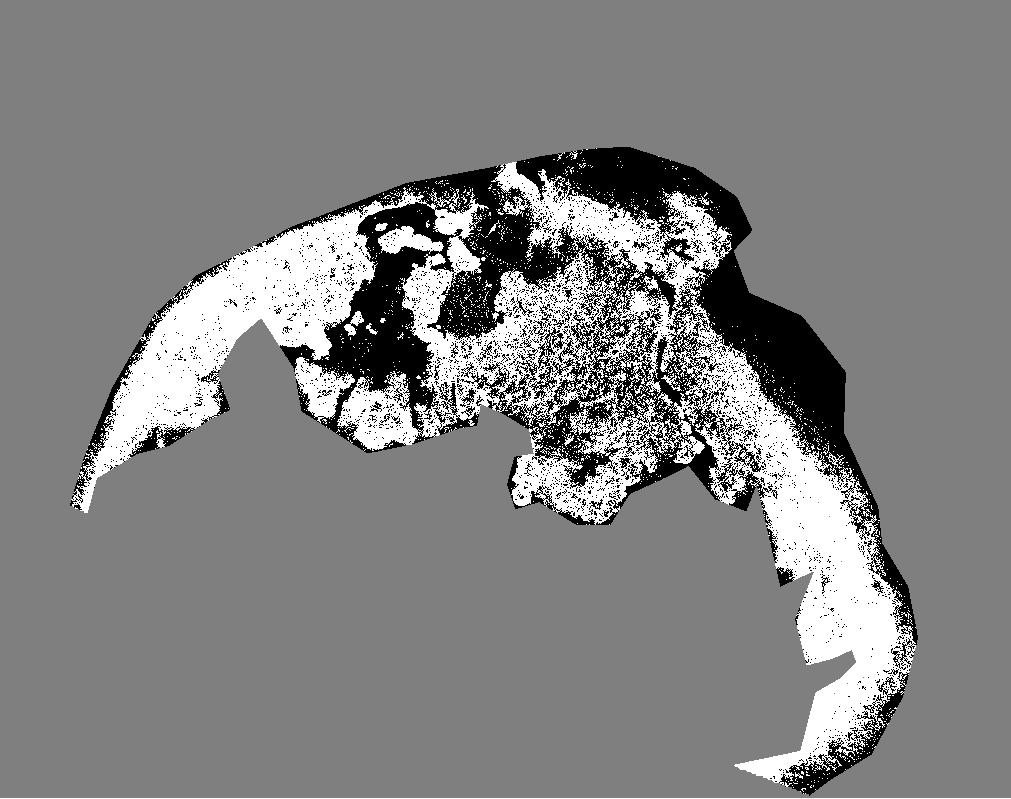


Select your input orthomosaic or other raster data in the input image section, and then choose the location and file name of the classified output image in the output image section.

For the number of classes keep in mind the resolution of your data, the similarity of the colors in the image. If the resolution is low, then the algorithm may struggle to fill the a distinct number of classes that you want. In addition, if you choose too little classes, then it may be possible that K-Means would group together distinct features in your image, leading to lower accuracy. In addition, watch out for anything in your image that is distinct from other parts of the image, but unnecessary. For example, sun glint and shadows from objects. You can increase the amount of classes to account for these distinctions in your image, and be able to remove these in your final view in QGIS or any other program that can display and edit your output raster.

In general, keep all other settings as default, but depending on the computer you are using change the Available Ram (Mb) to speed up the calculation of the clustering algorithm. For example using the BigBoi™ machine in the corner, set the Available ram to 32000, because of the 32GB of RAM on the computer and the abundance of Swap space. However if the processing was done on the computer in which this document is written on, the Available ram was set to 6000 with a measly 8GB of RAM.

Click Execute and wait for the processing to complete. Depending on how powerful the processing computer is and how large the image is, it can take from 15 minutes to a few hours.



Shown above on the left is the resultant K-means processed image, with the original file on the left. As you can see, with the wide variety of colors and, K means struggles to cluster the shape of the regions accurately. Part of this can be attributed to a very noisy resultant image, and we can help this by applying the sieve function in QGIS to get rid of small polygons that cause a noisy image.

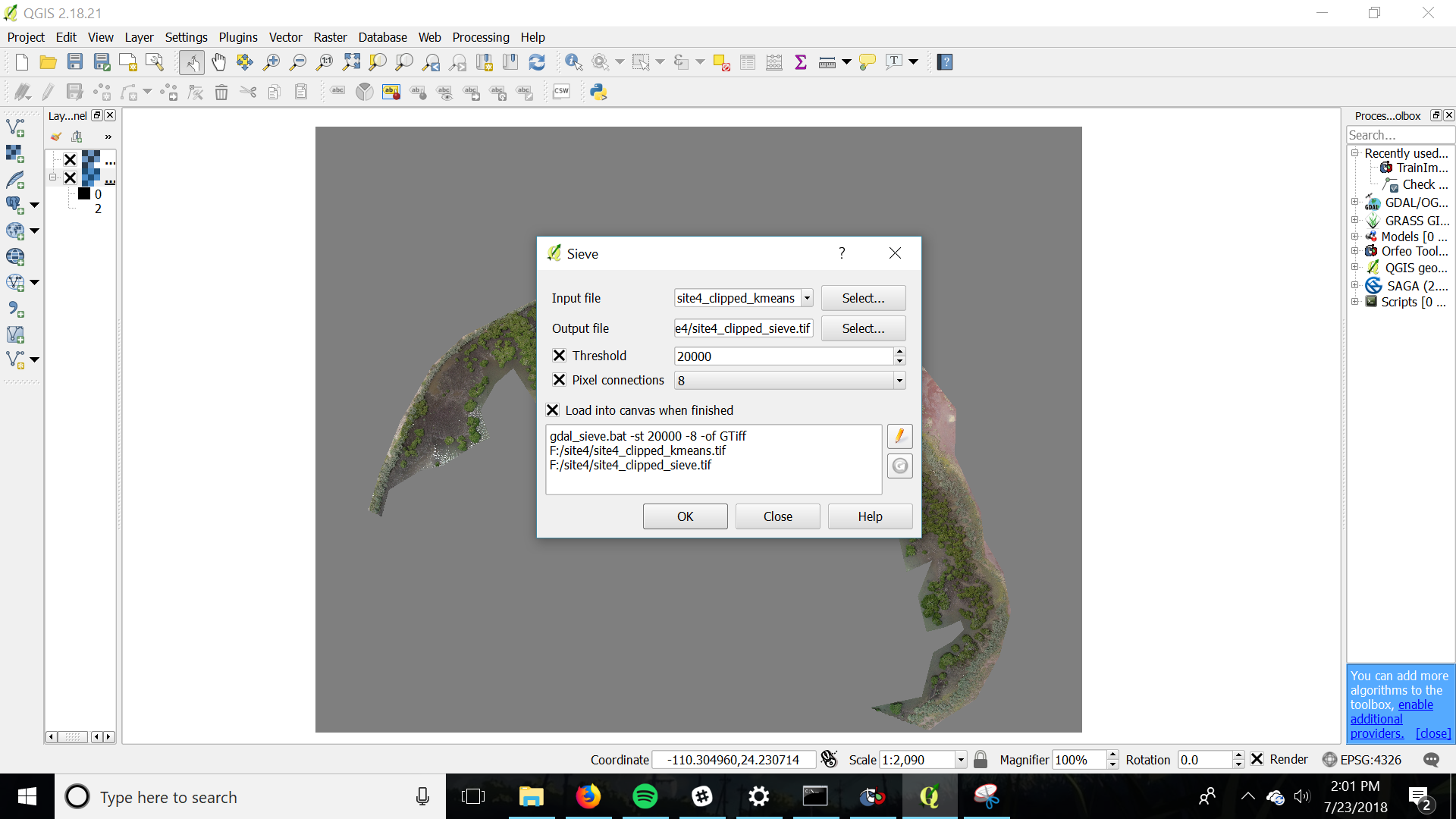
To use the sive on your already processed K-means raster, go to Raster and then Analysis and then click Sieve...

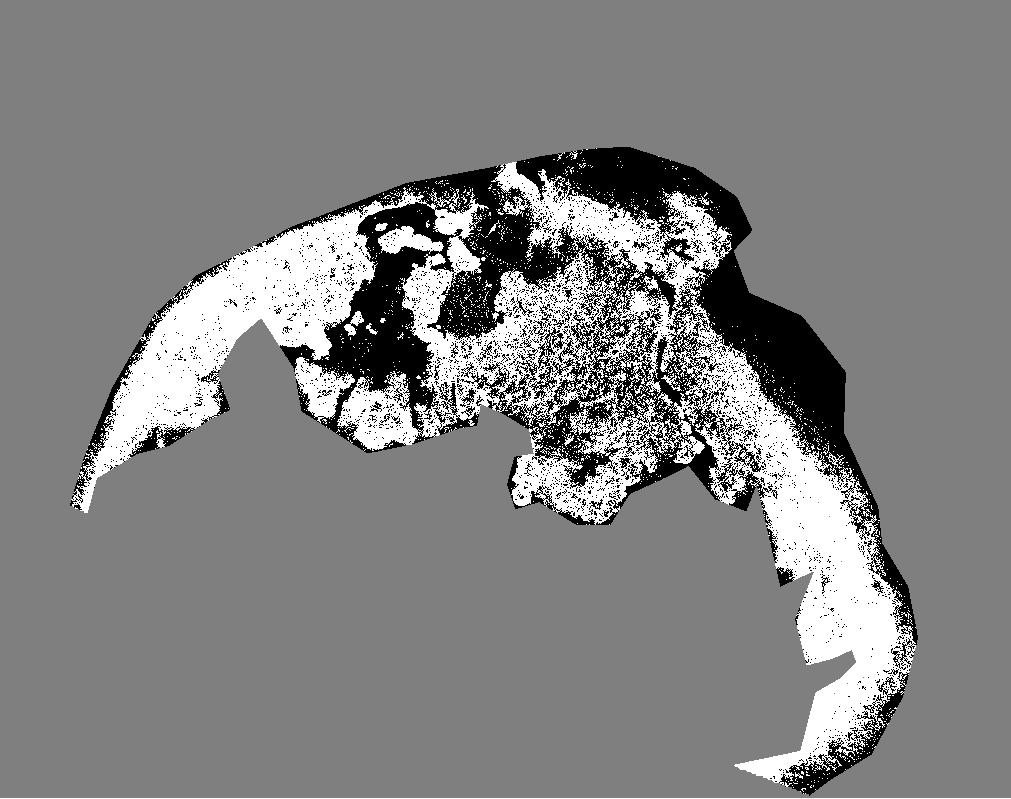


Select the input processed file and then select the location and name of the output file.

Set the threshold to a number in order to set the limit at which you want polygons under that size absorbed by the largest polygon that is adjacent.

Set the pixel connections to 4 if you want the pixels directly above or to the the sides of the pixel, or 8 if you want pixels that are diagonal in addition to pixels above and to the sides to be accounted for in the sieve.





As you can see, the picture on the right is far less noisy and is a smoother representation of labels in the original image after using the sieve.

# 

# REFERENCES

QGIS Documentation

https://docs.qgis.org/2.8/en/docs/index.html#

GDAL Documentation

https://trac.osgeo.org/gdal/

Orfeo Toolbox Documentation

https://www.orfeo-toolbox.org/documentation/