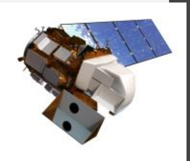
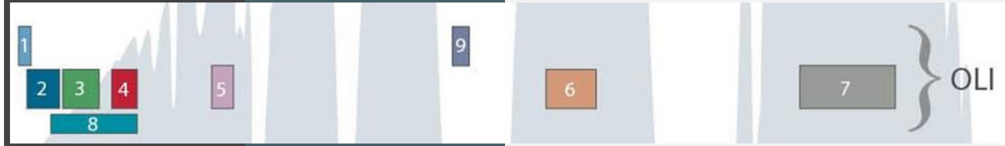
TROUBLE IN PARADISE: Hawai’i in true and false colors

To boldly go where some algorithms have gone before...

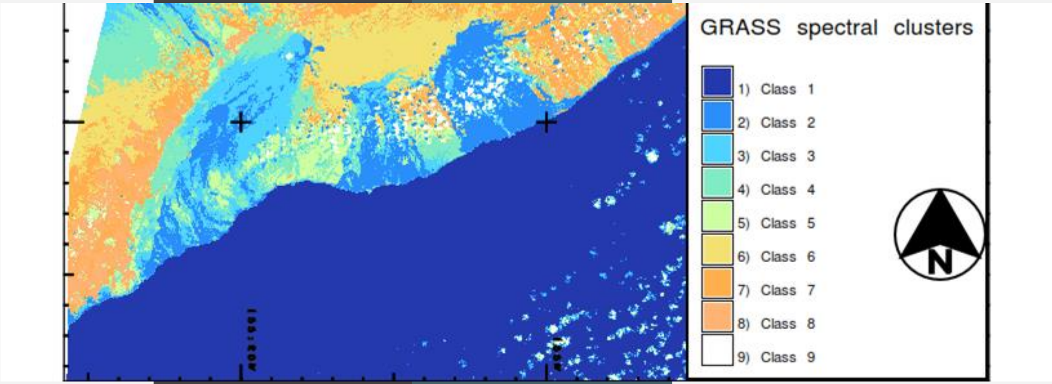
Image by Tommy Beatty from Pixabay

Landsat 8
Operational Land Imager (OLI)





Example of GRASS GIS spectral classification on Landsat 8 data



MOTIVATION

Use multispectral data (visible and infrared) to distinguish lava flow ages.

Infrared spectra are filthy rich with data about vegetation, soil, and rock.

Principal Component Analysis (PCA) to reduce spectral dimensionality.

Clustering to identify

- landscape spectral class (forest, fresh lava…).
- spatial + spectral objects / features (single lava flow).

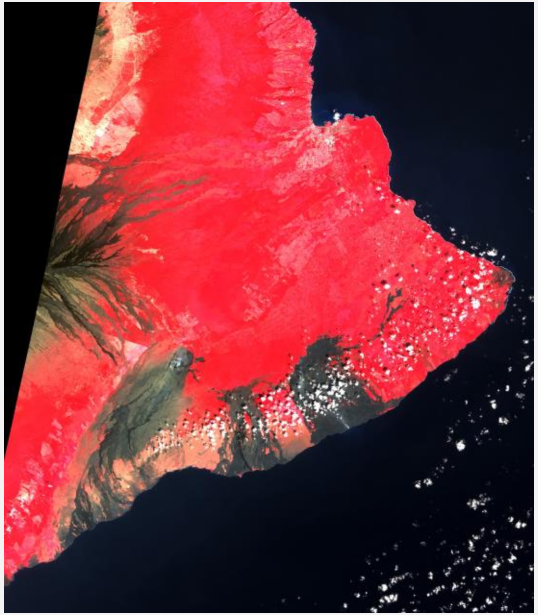
PROBLEM

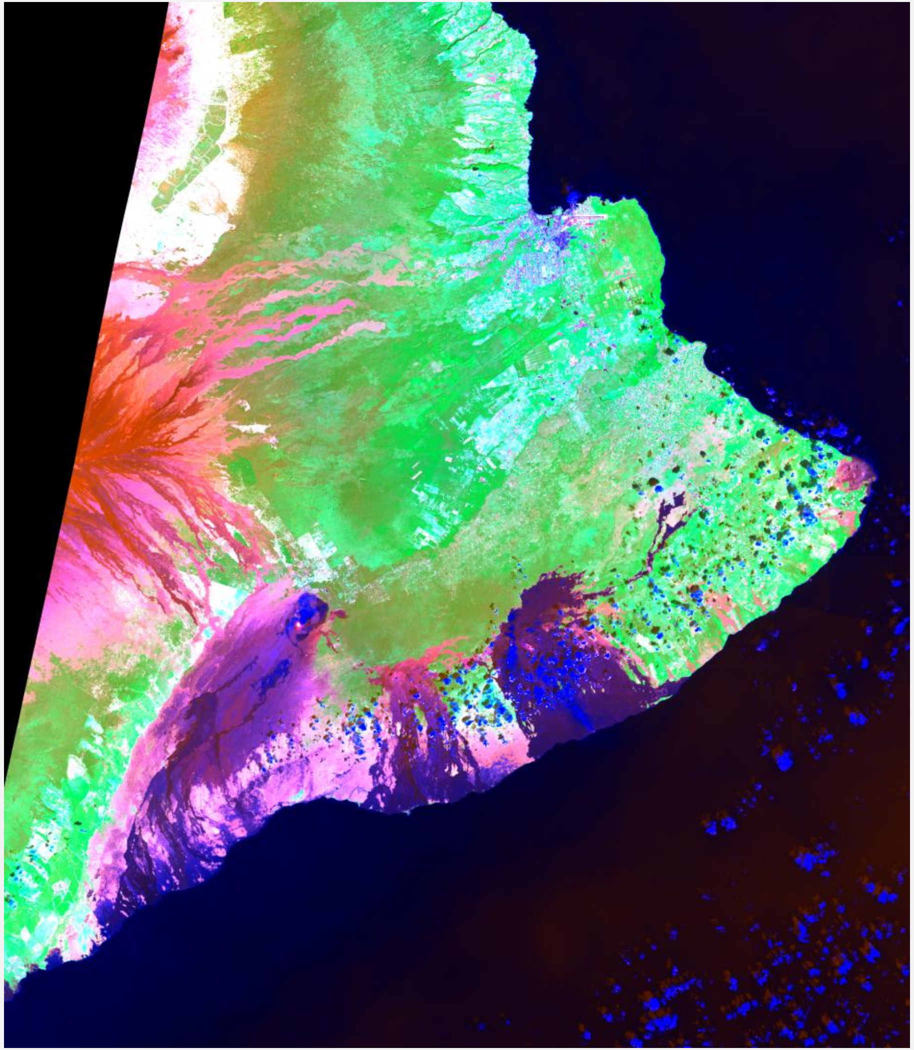
Seven spectral bands cannot, of course, be imaged simultaneously.

Typically this is addressed by creating false color maps where non-visible bands are exchanged with visible colors.

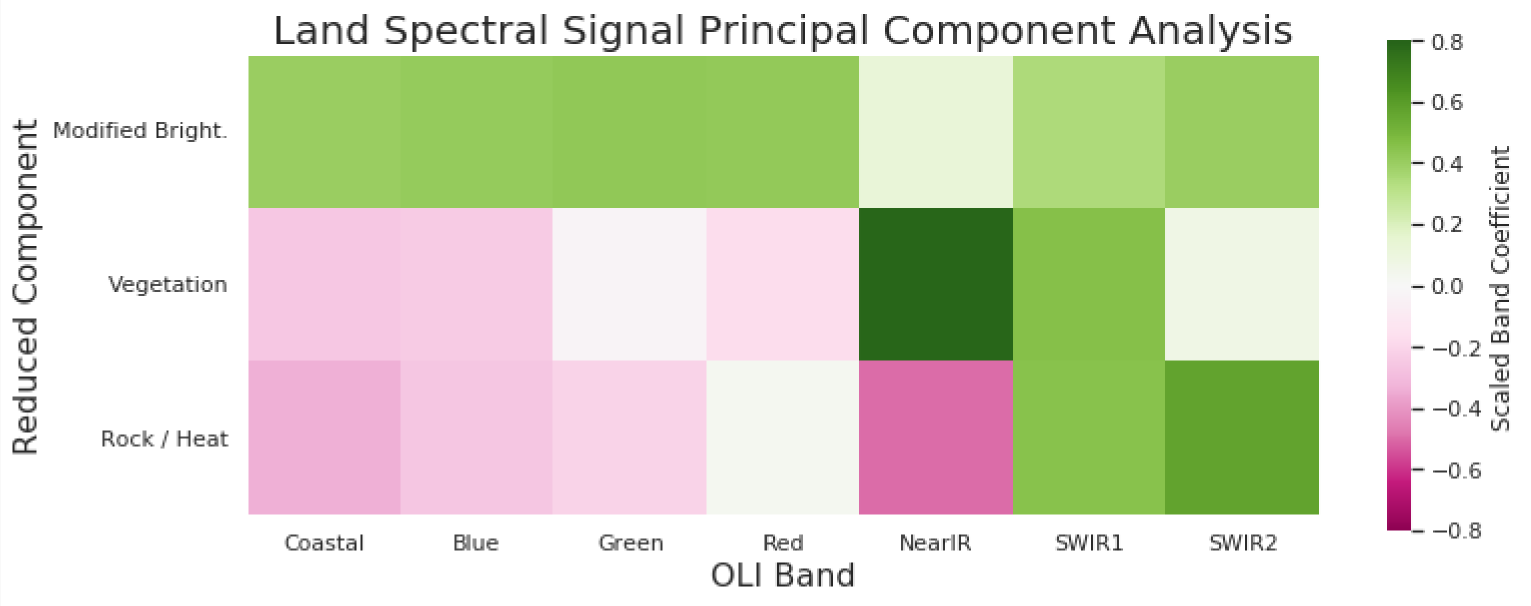
QUESTION

What would we see if we were to conduct a dimensionality reduction via PCA from 7 bands to 3 components and image that?





ENHANCED MAPPING WITH PCA



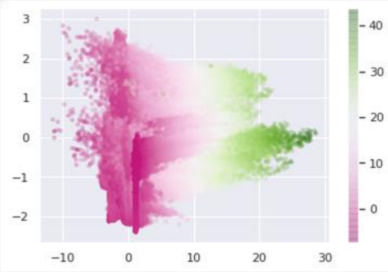
PCA conducted on *land only*
Component 1 - EVR 74%
"Brightness" biased against vegetation.

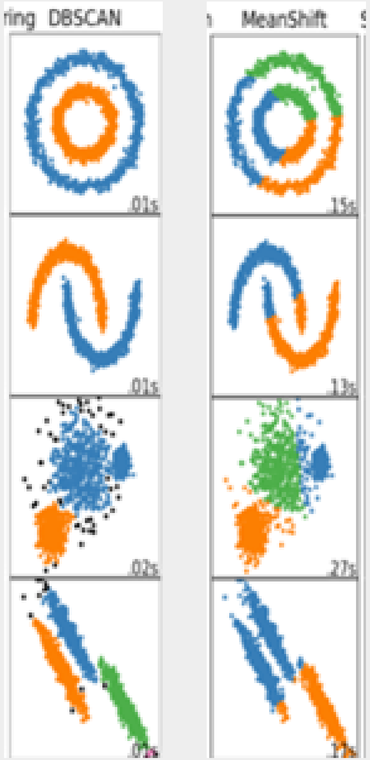
Component 2 - EVR 18%
Strong association with vegetation band 5; suspect soil (hydrated mineral signature) is in band 6.

Component 3 - EVR 6%
"Rockness"? "Heat"? Mid IR band 7 should be associated with non-hydrogen parts of mineral structure.

Red = rock / heat
Green = vegetation
Blue = “brightness”

Greatly enhanced contrast for Hilo streets, structures, airport runways.





CLUSTERING METHODOLOGY

<u>Reduce dimensionality with PCA.</u>

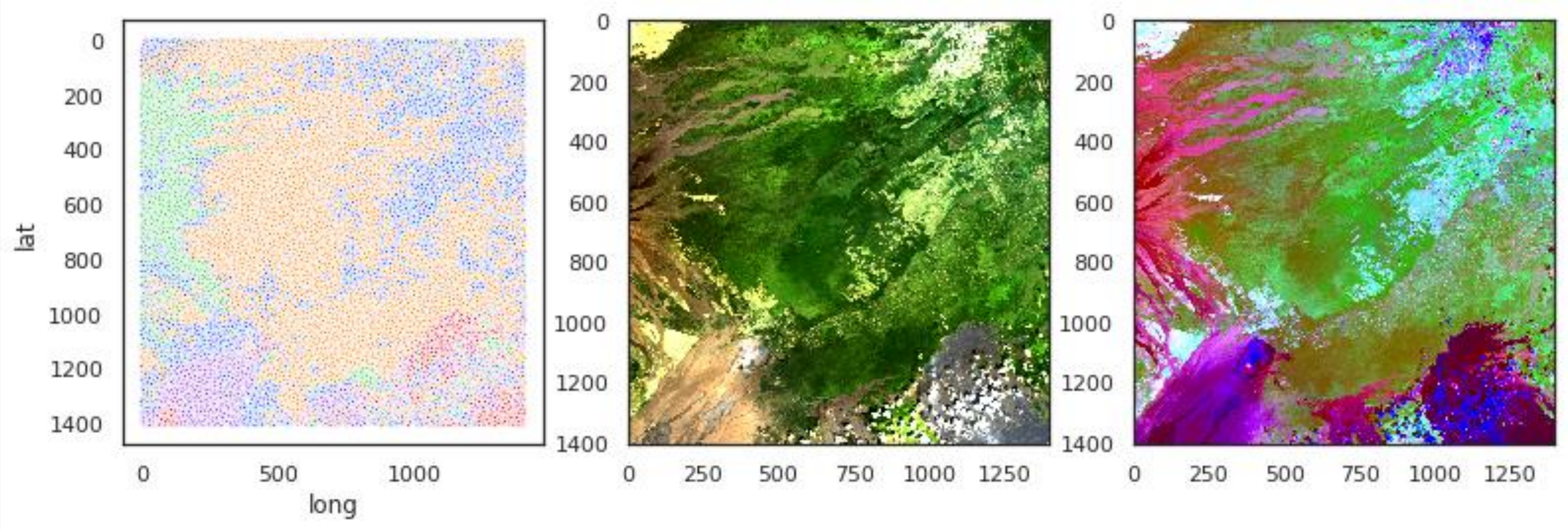
<u>DBSCAN</u> and <u>Mean Shift</u> algorithms:

- Avoid imposing a preset *k*.
- Cluster (leaving data points as noise) rather than partition.

<u>Gridsearch across hyperparameter space:</u>

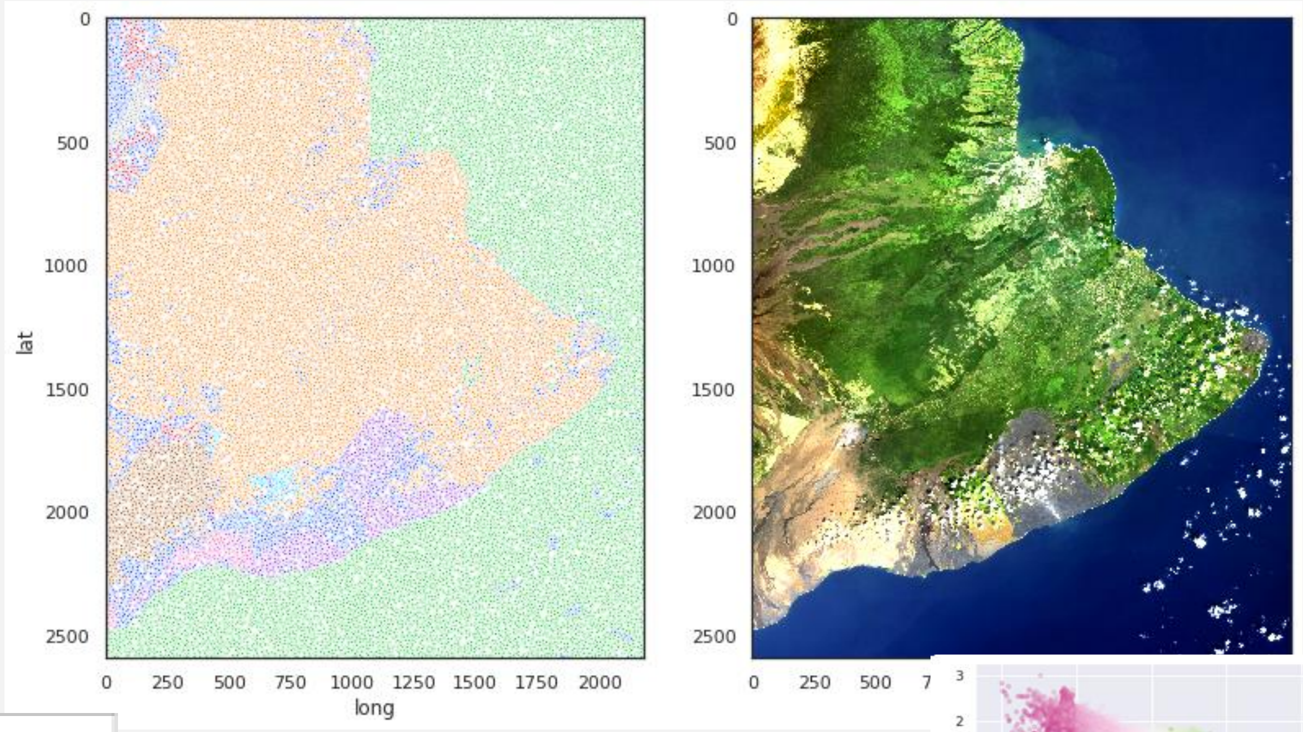
- DBSCAN:
  - core point distance
  - core number
- Mean Shift:
  - window size

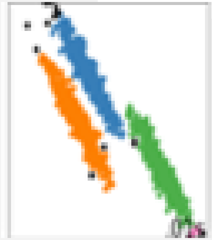
LANDSCAPE CATEGORIES



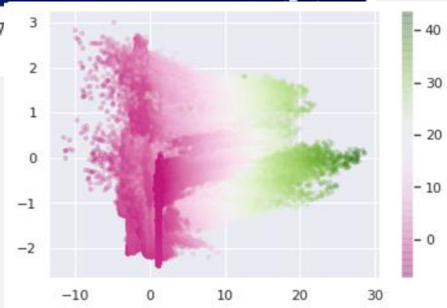
Mean Shift: window size 0.676 scaled units,
40% of estimate_bandwidth
Number of clusters: 186 (??!!)
Number of noise points: 22142
out of 39367 sampled pixels

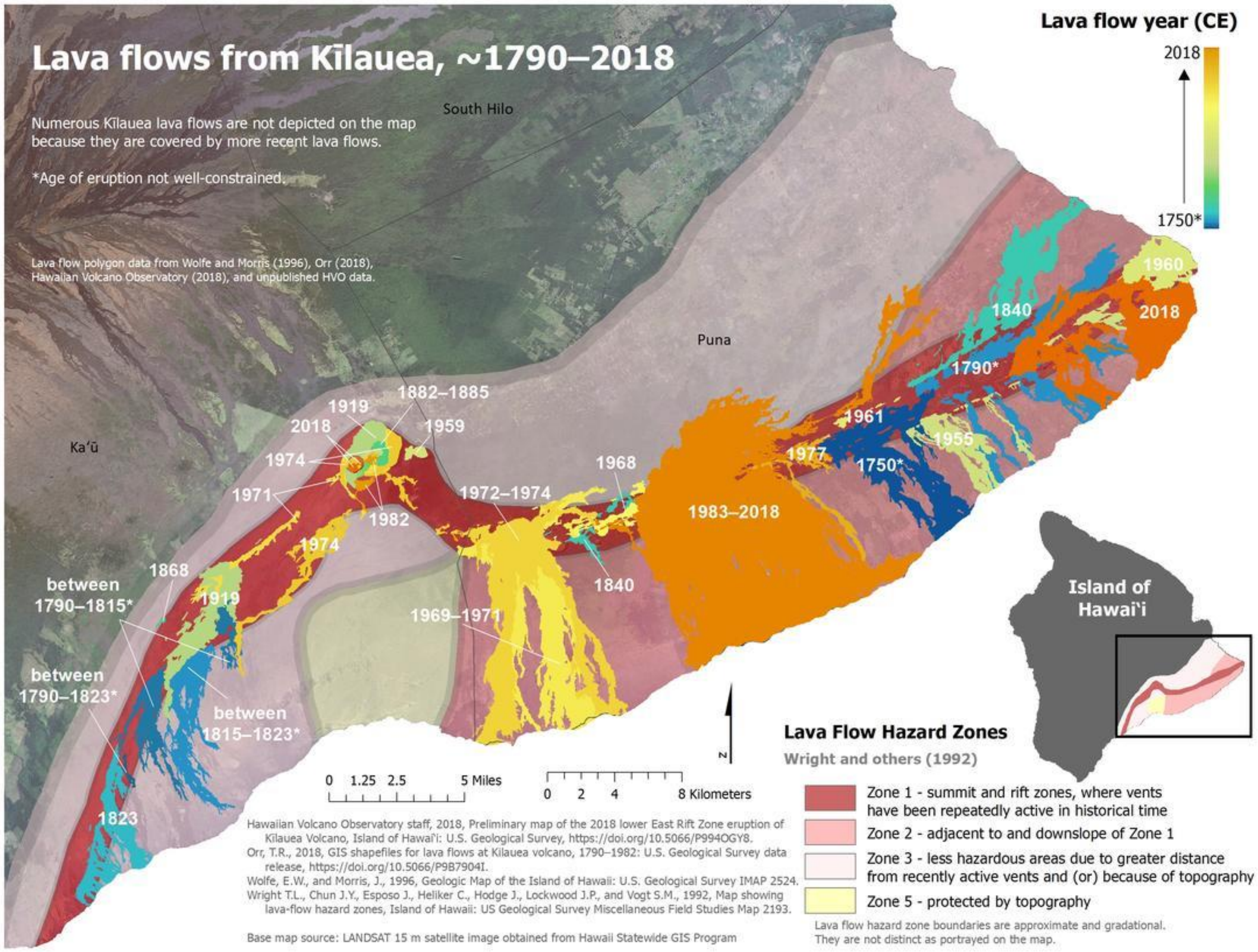
LANDSCAPE OBJECTS





DBSCAN: Core distance and core number:
0.20 scaled units, 25 points for core
Number of clusters: 16
Number of noise points: 8360
out of 56718 sampled pixels





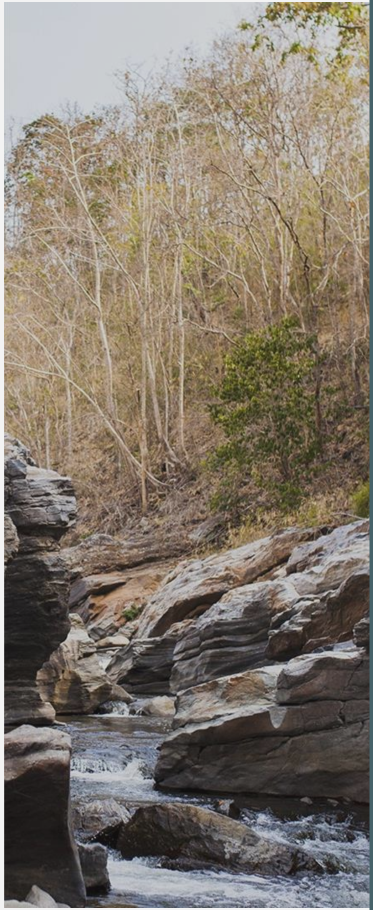
APPLICATIONS

<u>Volcanic hazard mapping:</u>

Not every ocean island or other basaltic province has been mapped as carefully as Hawai'i.

<u>Transferable methodology example:</u>

Mapping contaminant distribution via stressed vegetation.



CREDITS

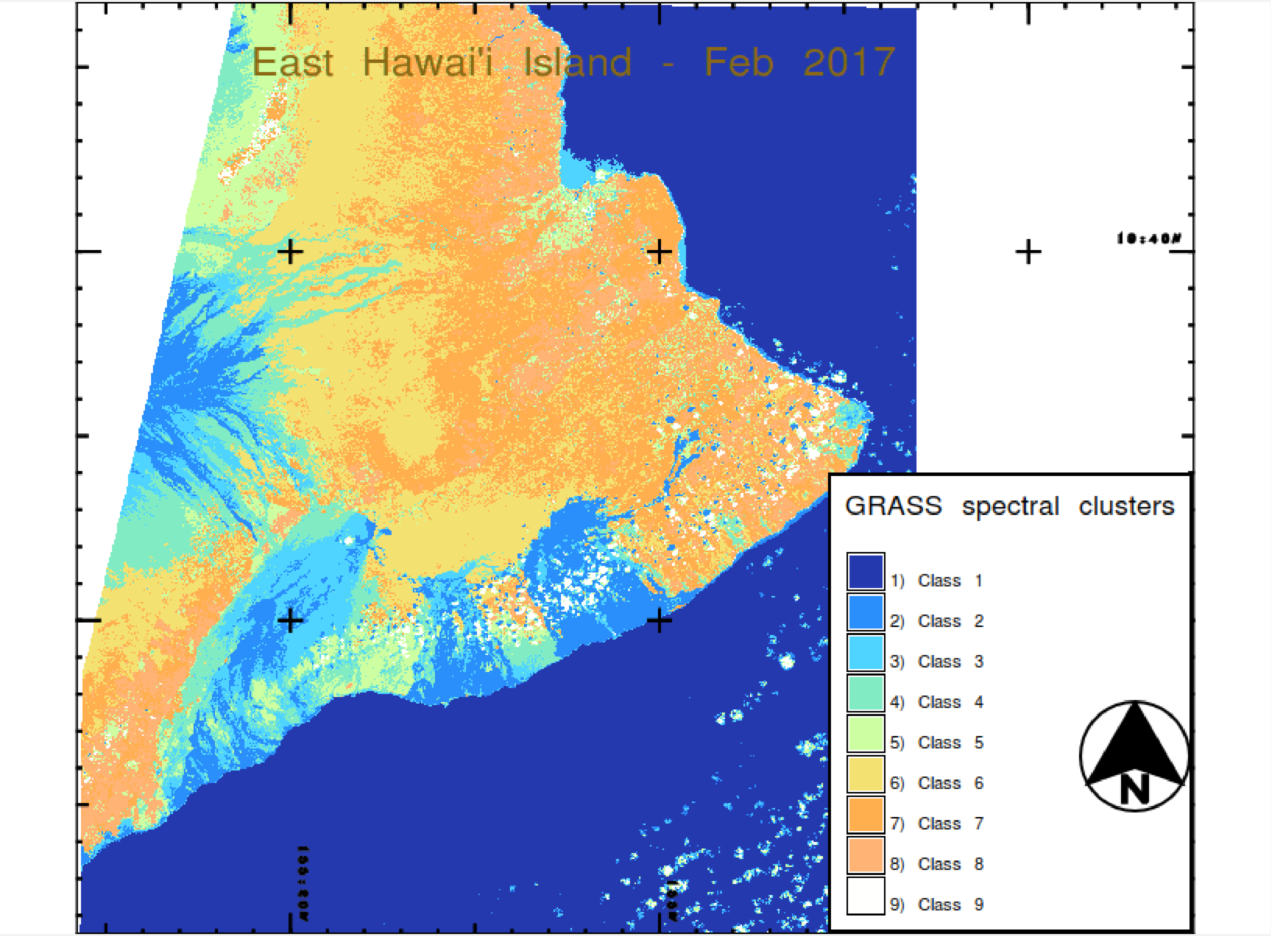
- Landsat 8 data collected by NASA and served by USGS.
- GRASS GIS from U.S. ACE *et al.*
- Particularly helpful tutorial on Medium by Robert Simmon dealing with the GDAL library and geospatial image data in general.
- Many other online sources of help!

- Presentation template by Slidesgo
- Icons by Flaticon
- Infographics by Freepik
- Author introduction slide photo created by Freepik
- Text & Image slide photo created by Freepik.com

Existing algorithms and software

Future work

BUT HASN'T THIS ALL BEEN DONE BEFORE?



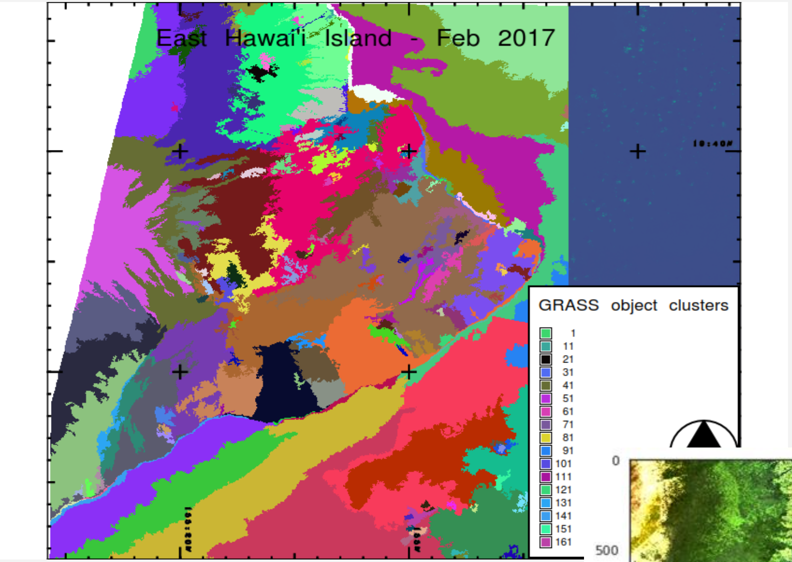
Well… mostly.

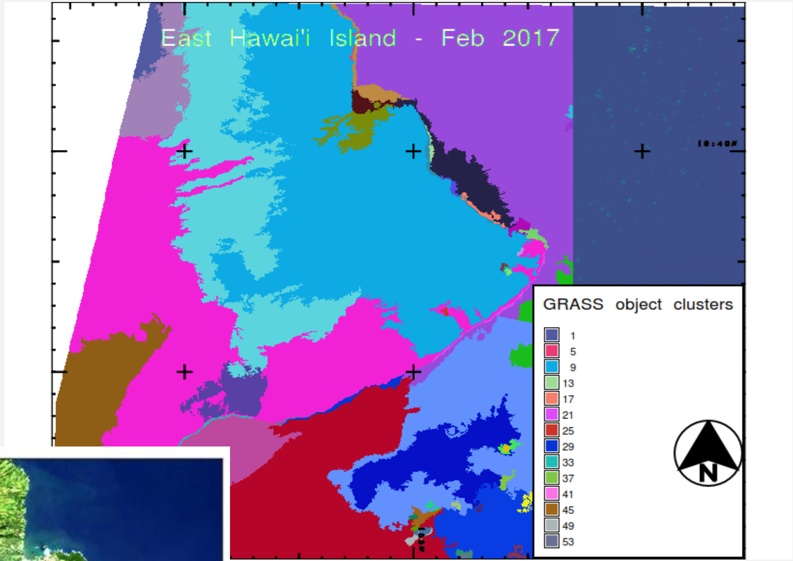
GRASS GIS i.cluster + i.maxlik routines: their own special sauce clustering algorithm, not quite *k*-means.

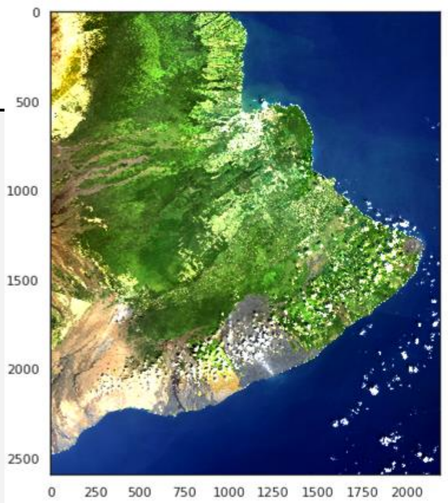
10 classes selected as input parameter, but some form of regularization removed one.

I don't feel bad about being outdone by a project with this many contributor-hours.

BUT HASN'T THIS ALL BEEN DONE BEFORE?







GRASS GIS i.segment routine used recursively / hierarchically (iterations 5 and 6 shown, thresholds 0.75 and 0.90).

Overall this is still superior, but there are a few advantages already apparent in using my alternative implementation. (I don't split the ocean, for example.)

Where to go from here?

Used advanced computing resources (Dask, Spark) on a remote instance to cluster all data.
Autoencoder to develop alternative reduced dimension spaces in which to cluster.

The original plan was to do initial exploration with clustering before moving to neural network classifiers.
There are several landscape classification training sets on TensorFlow alone to explore.
Train or find a pretrained model to then specialize to Hawai'i.

Tileserver to upgrade app so that large images can be served with short lag time.

Eventually attack the original motivating question: regress lava flow ages.
Find or create shapefiles for dated lava flows to use as training data.
Carefully construct & evaluate network to train it to recognize the difference between windward and leeward landscapes.
A solid model for lava flow age regression could be used to study volcanic hazards.
Similar methodology could attack questions like imaging stressed vegetation to infer contaminant distributions.