WYVERN
BETTER EARTH, FROM SPACE

CxC DATASET

Wyvern is incredibly excited to collaborate with UWaterloo's Data Science Club for CxC! We have provided you resources for tackling a major use case we are digging into here as well - crop classification!

WHO WE ARE

Wyvern is developing unfolding space telescopes to capture high-resolution hyperspectral imagery from space.

Hyperspectral images contain more colors than other types of imagery, meaning these images capture the spectral signature of your crop or forest, for example. With hyperspectral imagery, however, it's hard to get quality images with a high signal-to-noise ratio and high resolution. To mitigate this, we're designing optical telescopes that are compact on launch and *deploy in space*, meaning we pack better performance in a smaller, cheaper-to-launch package. Our increased light collecting area will allow for more light in more bands while maintaining <5 m resolution.

Our first satellites are being launched in spring 2023, which will deliver 5m resolution, VNIR (Visible to Near Infrared) hyperspectral imagery!

USE CASE BACKGROUND

Crop classification is a widely used tool for tracking crop types across large areas. **Many governments produce yearly crop classification maps** including the USA, Canada, Germany, and other EU member states. Crop classification is typically done via machine learning and combining **multiple (typically free) dataset sources**. For example, Canada's AAFC Annual Crop Inventory uses a decision tree model with optical (Landsat, Resourcesat, DMC) and radar (Radarsat-2) imagery.

Collecting and joining different types of datasets is time and resource intensive. Using free sources of imagery also typically have low **spatial and long temporal resolutions**, which makes it unsuitable for commodity trading & strategic decision making.

Our hypothesis is the following:

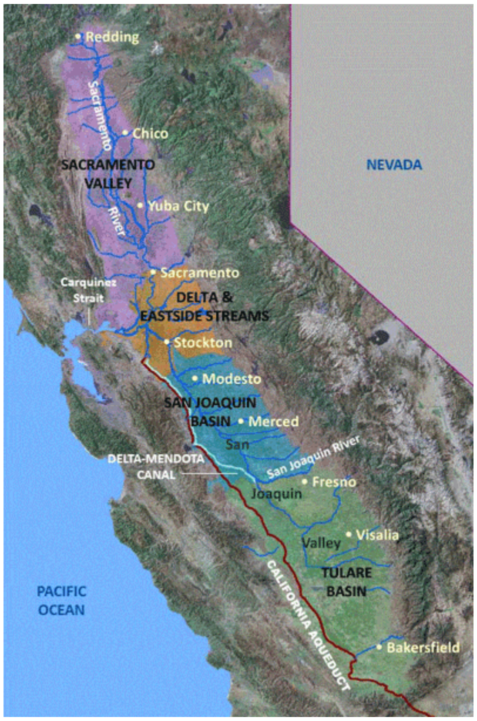
Hyperspectral VNIR imagery can deliver accurate crop classification **without** supplemental datasets using machine learning techniques, due to an increased number of features around the red-edge section of the VNIR spectrum.

Feel free to use any machine learning algorithms, datasets, augmentation, etc that you feel would improve modeling results!

Last update: 2023-01-27

DATASET

California's Central Valley is a large, flat valley that stretches approximately 700 kilometers through the center of the state. It is known for its rich soil and ideal climate for growing a wide variety of crops, making it one of the most productive agricultural regions in the United States. The valley is responsible for **producing a significant portion** of the nation's fruits, vegetables, and nuts, including almonds, grapes, peaches, and tomatoes.



USGS map of Central Valley regions in California

This agricultural diversity, coupled with the California government's excellent record keeping, means that the central valley is the perfect area to train & evaluate machine learning models on a variety of crops. The California Natural Resources Agency produces detailed, vector statewide crop type maps, which makes the data ideal for computer vision approaches.

Since our satellite isn't in space yet, we have to simulate imagery that our satellite will produce! To accomplish this, we've downsampled AVIRIS-NG imagery from 2018 and 2019 to our satellite specification (32 VNIR bands, 5m GSD). The downsampled imagery has been uploaded to a public S3 bucket for use by you!

Imagery is in the form of a **geotiff** file, which adds geospatial reference information on top of a normal tiff file. There are 32 channels (instead of 3 like RGB) within the file, each labeled with the central bandwidth of the channel in nanometers

Labels are in the form of a **shapefile**. These are vector shapes with a geospatial reference that contain details on what crop is growing in what field. They are raw from the California Natural Resources Agency, and will probably require cleanup!

DATASET TABLE

Since this is a ton of imagery, they have been combined into a zip file at the bottom of the page.

URL	Size/Filetype	Type of Data
WYVERN_AVIRISCALI_20180821t191919_v0_1_0.tiff	178.7MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180821t192451_v0_1_0.tiff	178.8MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180821t201135_v0_1_0.tiff	175.0MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180821t201548_v0_1_0.tiff	295.0MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180919t184533_v0_1_0.tiff	172.4MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180919t190324_v0_1_0.tiff	203.1MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180919t202534_v0_1_0.tiff	178.3MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180919t212105_v0_1_0.tiff	179.1MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180922t200321_v0_1_0.tiff	183.9MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180922t201849_v0_1_0.tiff	171.4MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180922t203520_v0_1_0.tiff	306.5MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20181005t230428_v0_1_0.tiff	179.5MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20181005t230859_v0_1_0.tiff	180.4MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20181010t192347_v0_1_0.tiff	278.4MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20181010t195707_v0_1_0.tiff	198.2MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20190623t194727_v0_1_0.tiff	2.0GB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180821t195454_v0_1_0.tiff	274.3MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20190919t200411_v0_1_0.tiff	216.0MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20181005t223802_v0_1_0.tiff	219.0MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20181009t200058_v0_1_0.tiff	173.8MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20191009t183852_v0_1_0.tiff	170.2MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180821t190043_v0_1_0.tiff	170.9MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20181009t181942_v0_1_0.tiff	168.0MB/geotiff	**Imagery**
WYVERN_AVIRISCALI_20180821t184959_v0_1_0.tiff	167.8MB/geotiff	**Imagery**
WYVERN_AVIRIS_CALIFORNIA_IMAGERY_2018-2019.7z	6.0GB/zipped geotiffs	**All Imagery**
2019 Statewide Crop Mapping GIS Shapefile	140MB/shapefile	**Label**
2018 Statewide Crop Mapping GIS Shapefile	123MB/shapefile	**Label**

*Note: datetime information is encoded in the imagery filename: WYVERN_AVIRISCALI_**20181005t230859**_v0_1_0.tiff*

Last update: 2023-01-27

EXPERIMENT DESIGN

We're keeping this section intentionally vague! This use case can be accomplished with a variety of approaches - and to be totally honest we're still trying to identify the best one!

Your approach to modeling will also affect experiment design. Feel free to use your best judgment when tackling this use case.

CROP CLASSES

The raw California Natural Resources Agency shapefiles probably have way too much information in them! One of the first key things to tackle is cleaning and simplifying this dataset in preparation for joining with imagery. This can be accomplished with packages like geopandas.

Additionally, differentiating groups of crops (cereals, legumes, grasses, citrus, vineyards, nuts) may be easier than identifying specific species of crops (wheat, barley, chickpeas vs. soybeans). **Ideally a model would be able to accurately predict as many classes as possible.**

TRAIN/TEST/VALIDATION SPLIT

Deciding on a train/test split is completely up to you! There are a number of different approaches, including holding some images for test/validation or assigning image chips to train/test classes. **Please ensure there is no leakage from train to test.**

ESTABLISHING A BASELINE

Performance can vary significantly across different datasets, models, and species of plants. As a reference, Canada's AAFC Annual Crop Inventory model has a performance of at least 85%. Many current methods for crop classification use decision tree algorithms. Training a basic decision tree algorithm as a baseline would be an excellent way to compare performance.

ASSESSING PERFORMANCE

There are a couple of key areas to dig into when assessing performance:

- Performance across geographic regions
- Performance across different seasons/times/years
- Performance across different crop types

An effective machine learning experiment would assess model performance across all factors to assess if the model is capable of being applied to new imagery.

Your submission will be assessed on both model performance, but also experiment design and validity of results.

RESOURCES

Hyperspectral imagery is a relatively new type of data, especially in machine learning! Below are some potential resources for you to use to tackle this use case. **Additionally, feel free to use more data sources and labels, and whatever you feel would improve your solution!**

WYVERN KNOWLEDGE BASE

https://knowledge.wyvern.space/#/

RASTER VISION ML LIBRARY

https://rastervision.io/

TORCHGEO (PYTORCH)

https://torchgeo.readthedocs.io/en/stable/

MICROSOFT PLANETARY COMPUTER

https://planetarycomputer.microsoft.com/

AVIRIS-NG DATA PORTAL

https://avirisng.jpl.nasa.gov/dataportal/

CALIFORNIA STATEWIDE CROP MAPPING

https://data.cnra.ca.gov/dataset/statewide-crop-mapping

CONTACT

If you have any questions at all, don't hesitate to reach out to Ellie Jones via email or LinkedIn!

https://www.linkedin.com/in/elliejonesyyc/

ellie.jones@wyvern.space

Last update: 2023-01-27