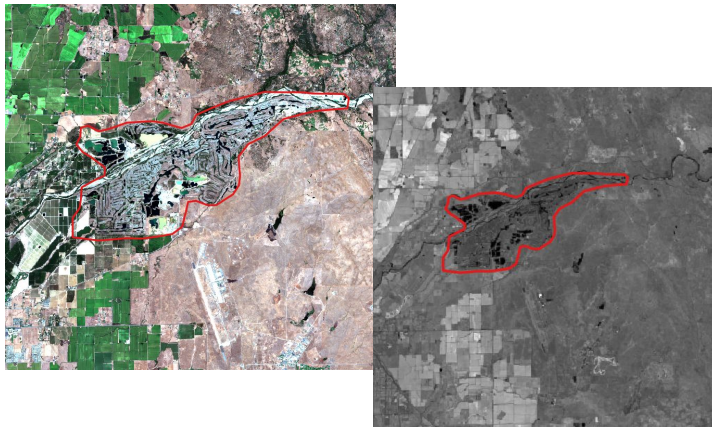
Mining Asset Detection (MAD)

Overview - Goal

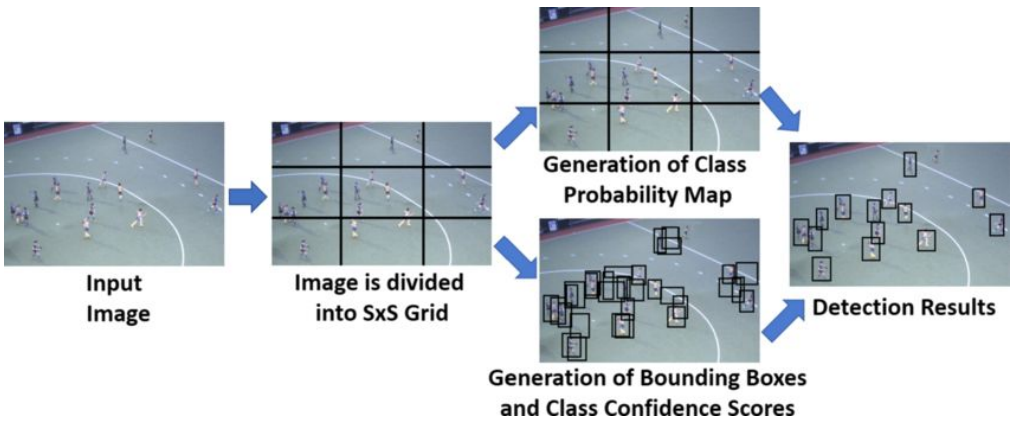


Using **multiband** satellite imagery and a set of **mining polygons**, detect mining sites worldwide.

- Maus mining polygons
- Sentinel L2A Satellite images (12 Bands)

Methodology - YOLO

- Family of visual models
  - Given labelled data (classes and their positions)
  - Produces a classifier that given new images, predicts classes and their position



https://www.researchgate.net/publication/352105926_Review_and_evaluation_of_player_detection_methods_in_field_sports_Comparing_conventional_and_deep_learning_based_methods

Methodology - YOLO

- General idea
  - Train YOLOv11 on satellite images with labelled mines
  - Use this model to predict mine occurrences worldwide

Open questions:

- How/which different bands should be utilised?
- What characteristics of the training data give the best performance?

Methodology - Band selection

Assumption: Other bands than RGB, can help predict mine locations.

In order to test this: Use **Lasso** regression

Methodology - Band selection - Lasso basics

$$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; ||y - Xw||_2^2 + \lambda ||w||_1$$

Regular Least squares:

- Choose coefficients such that the y is optimally reproduces

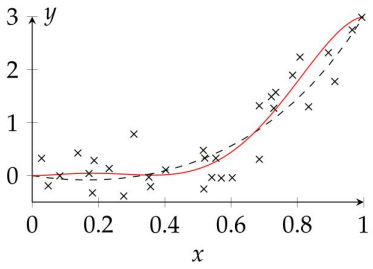
L1 Normalization:

- Penalizes large coefficients
- Leads to a minimal coefficient set

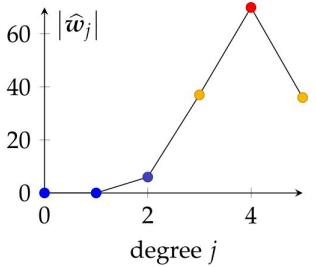
Methodology - Band selection - Lasso characteristics

Lasso helps with finding lower level representations

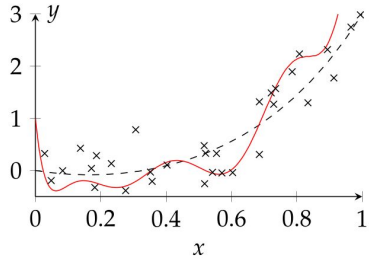
- Simpler model => Higher bias, lower variance



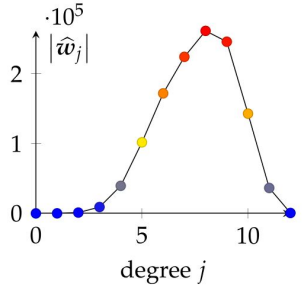
(a) The fitted polynomial.



(b) The coefficients of the fitted polynomial



(a) The fitted polynomial.

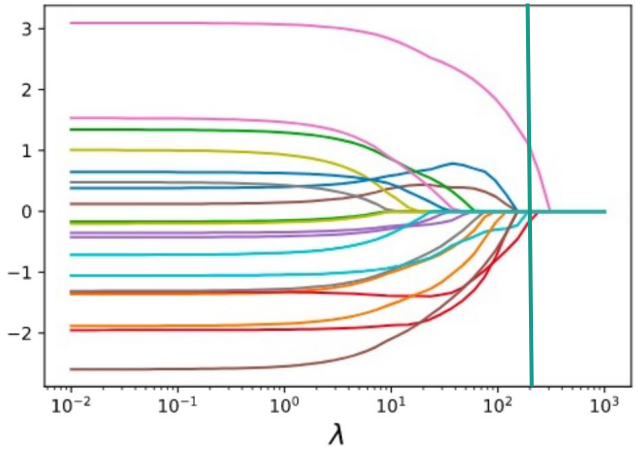


(b) The coefficients of the fitted polynomial.

Methodology - Band selection - Lasso interpretability

Key characteristic for our use case:

- By varying Lambdas, we can determine "*the most essential*" coefficients
- In our example: **the most essential bands**

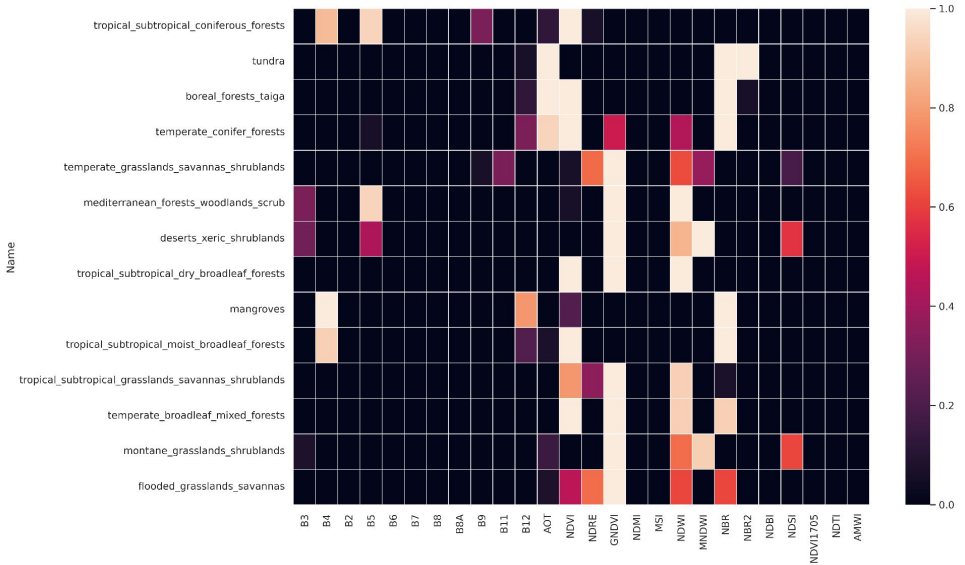


Methodology - Band selection - Lasso results

Do multiple Lasso runs for different biomes

- For the first Lamba where only **3** bands survive, record the bands
- Capture the occurrence of these bands surviving

Result: Which bands are the most important for every biome.



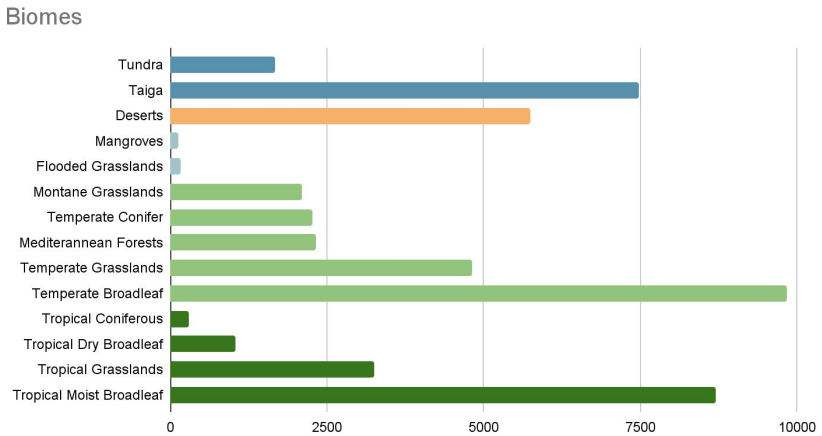
Methodology - Band selection - The caveats

Data is imperfect

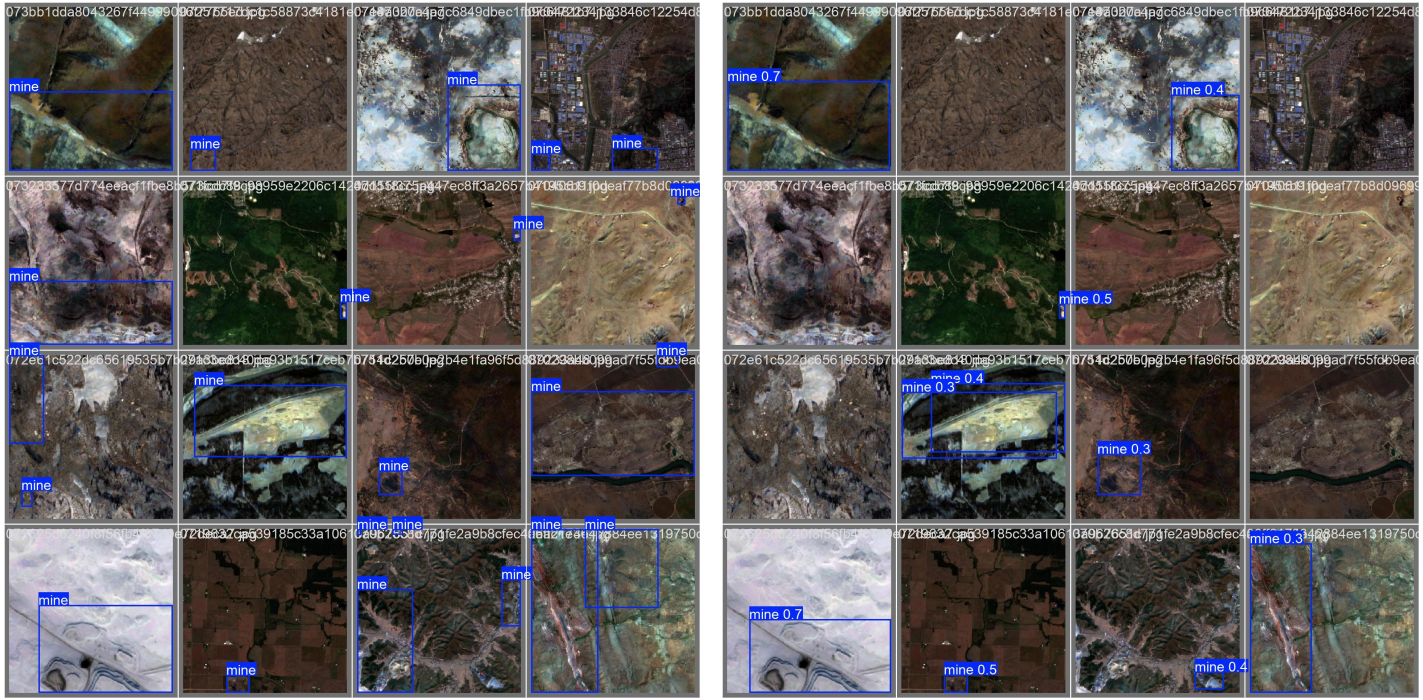
- Not equally distributed for different biomes
- Mining polygons very imprecise

Resulting models are not great at prediction

- They seem to pick up some information
- Questionable if verifiably correct



Methodology - YOLO - First run

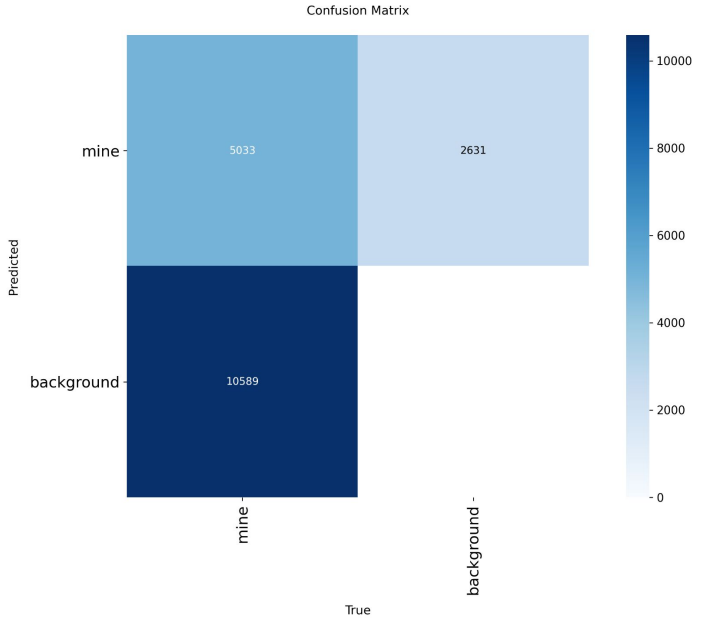


Methodology - YOLO - First run

Results:

- Model seems to learn something
- It has a bad recall (a lot of false negatives)

But why?



Methodology - YOLO - Image Quality

Suspicion on why it does not perform that well:

- Lacking image quality of the training images
  - Mainly due to the fact that we're using averaged out images

Solution:

- Get satellite images from the source
- No accumulation of multiple images, use single snapshots
- Filter based on quality metrics

Quick Demo

Methodology - YOLO - Biomes

Suspicion on why it does not perform that well:

- No distinction between different biomes for now
  - From the Lasso experiments, there is a larger difference
  - So maybe for the visual model too?

Solution:

- Train specialised models based on image metadata or characteristics

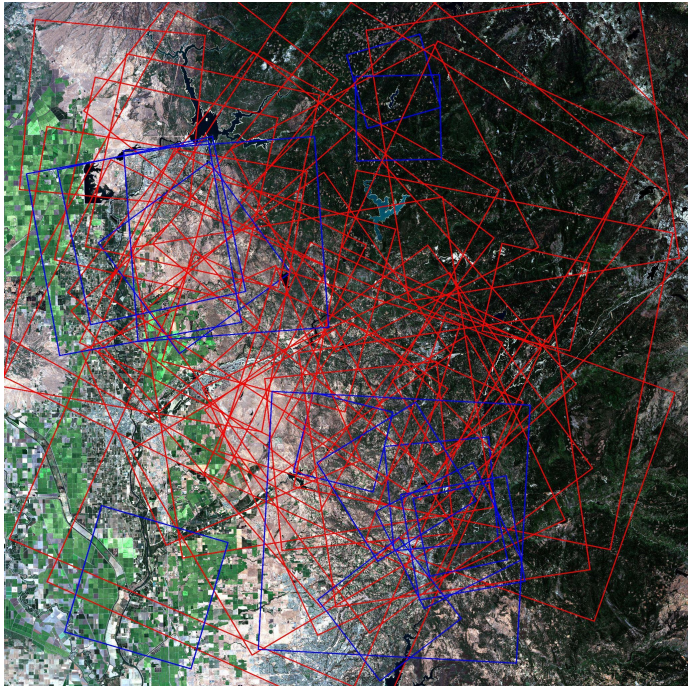
Methodology - YOLO - Variations

Suspicion on why it does not perform that well:

- Relatively small final dataset (~2GB)
  - Adding slight variants of images in the training set should increase the overall size

Solution:

- Increase size of dataset by adding variations



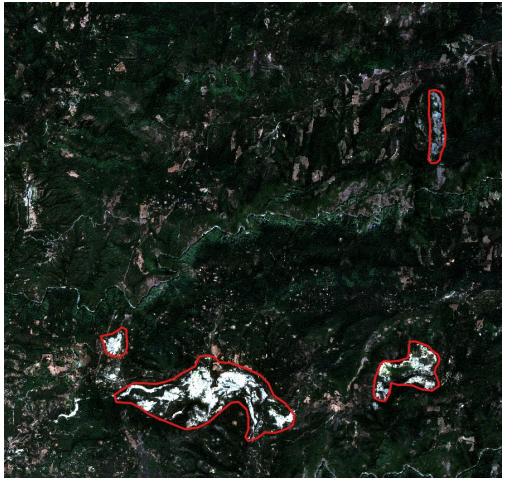
Methodology - YOLO - Bands

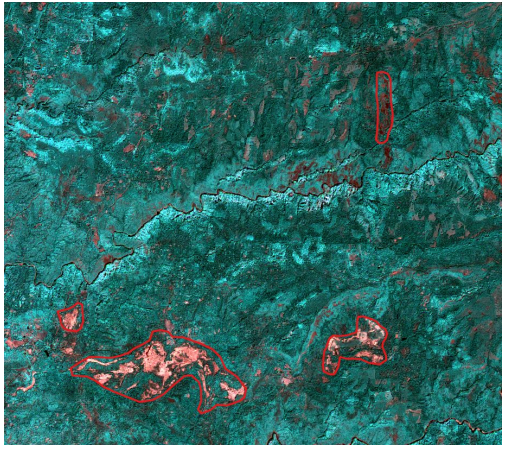
Suspicion on why it does not perform that well:

- We’re only using RGB for the training

Solution:

- Use the more effective band combinations





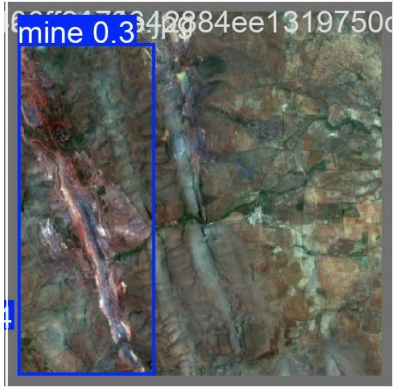
Methodology - YOLO - Labelling

Suspicion on why it does not perform that well:

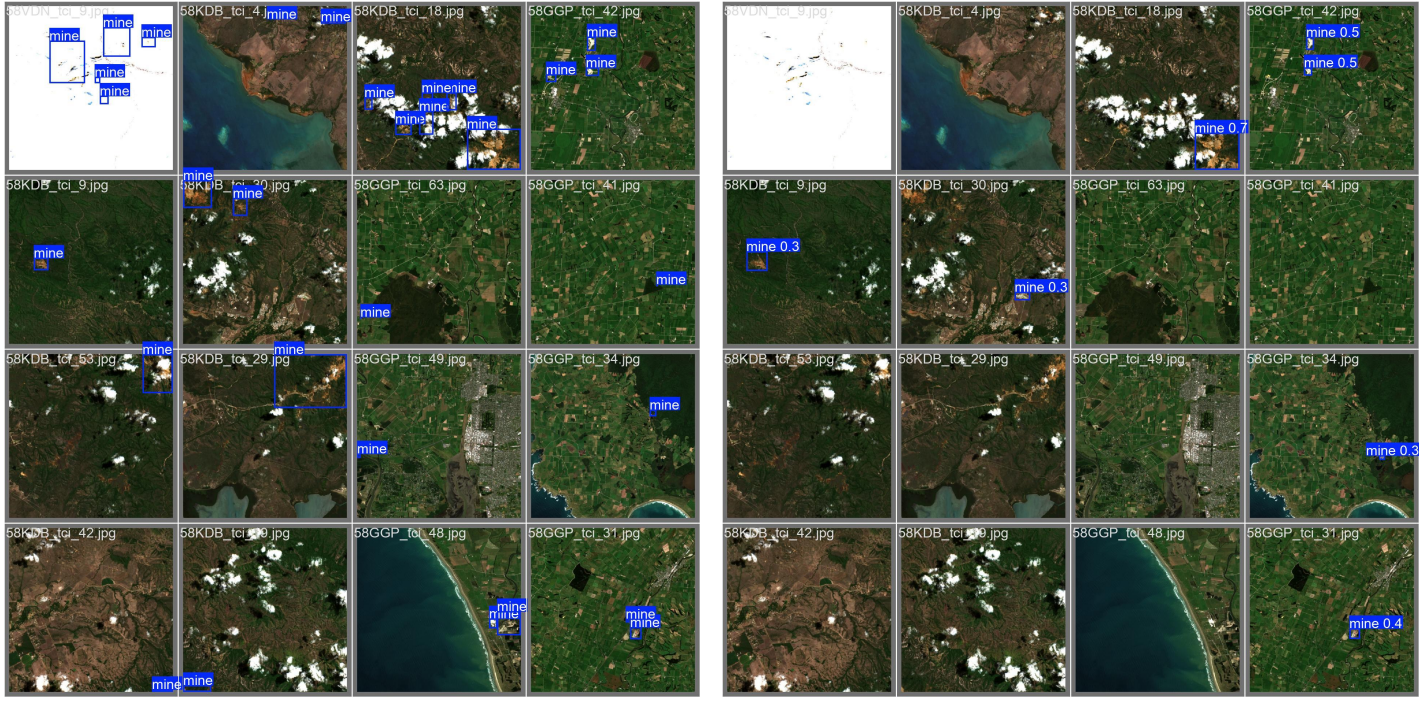
- Unsatisfactory labeling
  - Overlapping labels might cause issues with the learning
  - Labelling is too imprecise (stems from the maus polygons)

Solution: Improve the labelling process as well as possible

- Try to omit intersection labels
- Filter badly intersecting polygons

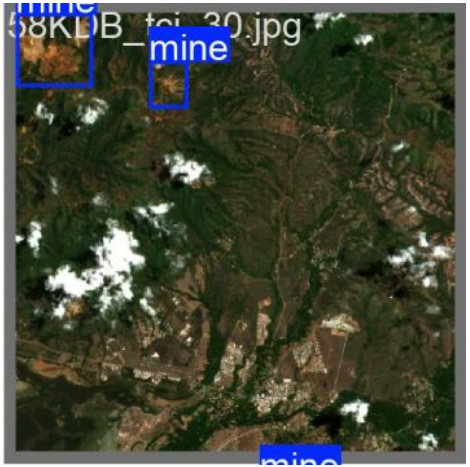


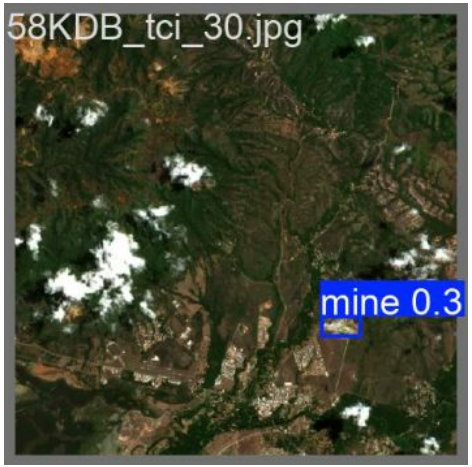
Labelling investigation



Labelling investigation

False negatives in the dataset…

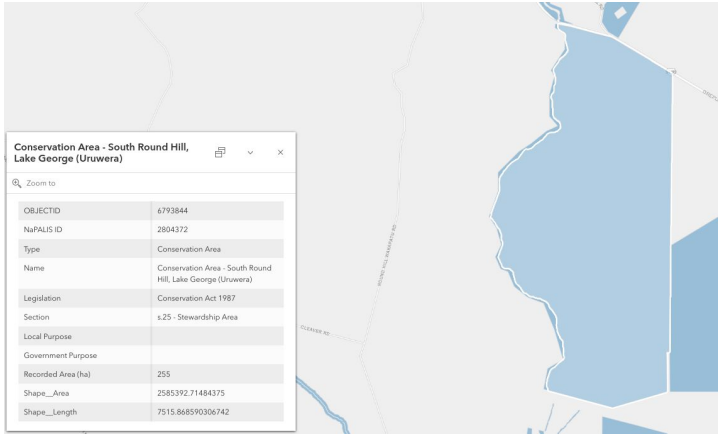




Labelling investigation

False negatives + false positives in the dataset…

Labelling investigation



This Protected Area Layer contains land and marine areas, most of which are administered by the Department of Conservation Te Papa Atawhai (DOC) and are protected by the Conservation, Reserves, National Parks, Marine Mammal and Marine Reserves Acts. All of the areas have been identified spatially. The attributes in this dataset are derived from the National Property and Land Information System (NaPALIS), which is a centralised database for all DOC and LINZ administered land.

The boundaries for most protected areas are derived from the Landonline Primary Parcel(s). In some cases, the boundaries may have been based on unsurveyed parcels defined to varying degrees of accuracy. As such please note that the boundaries are indicative only.

The **Longwood Range** is a range of hills to the west of the Southland Plains, Southland, New Zealand.[1] From the 1860s until the 1950s gold mining was prevalent in the Longwood Ranges.[2] There are many small towns and localities situated around the periphery of these hills: clockwise from the south-east, these include Riverton, Pourakino Valley, Colac Bay, Pahia, Orepuki, Tuatapere, Otautau and Thornbury.[3]

The Te Araroa Trail runs through the forest.

Labelling investigation

2.1 Locality

The existing WWTP is located on a site of approximately 3 ha, situated 1.1 km northwest of the Edendale – Wyndham Road bridge over the Mataura River. The WWTP has an existing pipe conveying the treated wastewater to the Mataura River outfall following the alignment of Edendale - Wyndham Road, within the existing road reserve.

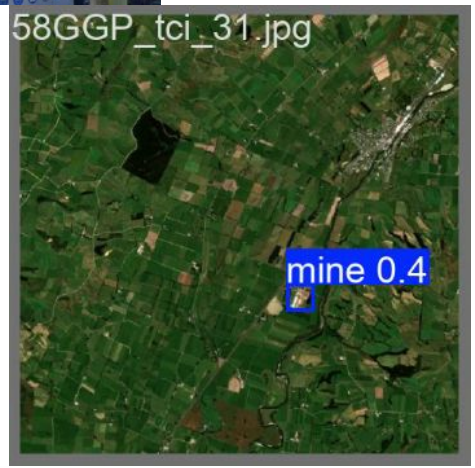
2.2 Land use

2.2.1 Existing Site

The site is currently used for the Edendale – Wyndham WWTP. The plant is based on a vermiculture treatment system and comprises the following elements:

- Inlet screens (2 units).
- Filter belt press.
- Vermiculture treatment beds (5 beds), “worm beds”.
- Phosphorus removal system.
- UV disinfection.

Labelling investigation



Labelling investigation

This is purely anecdotal! We can't claim the model to be great at its job.

But…

Maybe the image quality is not really the bottleneck but the underlying truth is?

- It's hard for the model to learn a substantial amount with false negatives
- At the least, the final test set should not contain any false values

What now?

1. Some manual labelling is needed at least for the final test set.
    a. With ~ 4900 images ~ 10-15% should be labelled by hand

2. Similar to the image quality issue, the ground truth should be augmented iteratively
    a. Detect false negatives in the system, correct them where possible
    b. Maybe in an automated way?
    c. With which rules in place? What even is a mine?

3. If possible, a well labelled, geographically diverse country with little to no false labellings would be great.
    a. Could also be a diverse collection.