Machine Learning for Cloud Masking in Sentinel-3 SLSTR Data

Samuel Jackson (samuel.jackson@stfc.ac.uk), Caroline Cox (caroline.cox@stfc.ac.uk),
Jeyan Thiyagalingam (t.jeyan@stfc.ac.uk)

Introduction

Clouds appear ubiquitously in the Earth's atmosphere, and thus present a persistent problem for the accurate retrieval of remotely sensed information. The task of cloud masking can be represented as a binary classification problem where we aim to assign each pixel in an image “cloudy” or “clear”.

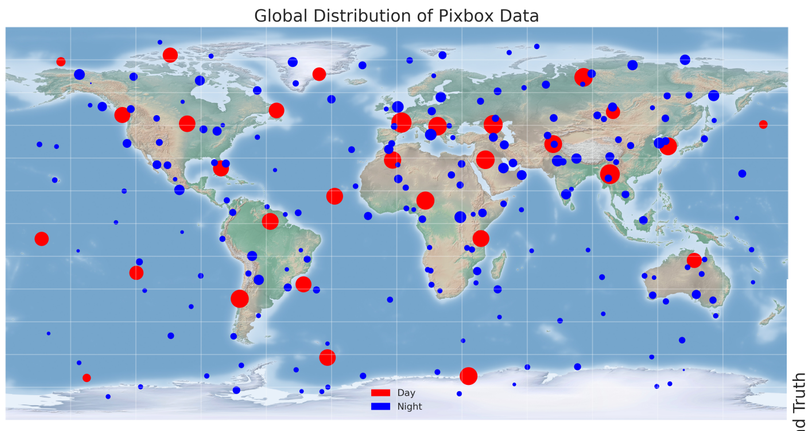
With the deluge of data now being collected on a daily basis from Earth observation platforms, machine learning algorithm are becoming an increasingly attractive option to tackle highly complex problems using a data driven approaches where traditional rule based systems can fail.

A major barrier to applying machine learning to EO data is the lack of labelled data from which to train models. In this work we make use of a relatively small hand-labelled dataset of single pixels to train a model which can distinguish between cloud and clear sky.

Data

The Sentinel-3 satellites carry the Sea and Land Surface Temperature Radiometer (SLSTR) [1]. The SLSTR records in two resolutions and in two viewing angles: nadir and oblique. It measures in 11 different channels from Near infrared (NIR) through short wave infrared (SWIR) to thermal infrared (TIR) bands. The product also includes 2 existing cloud masks: A mask which uses thresholding tests (Empirical) and one which uses a Bayesian model (Bayesian).

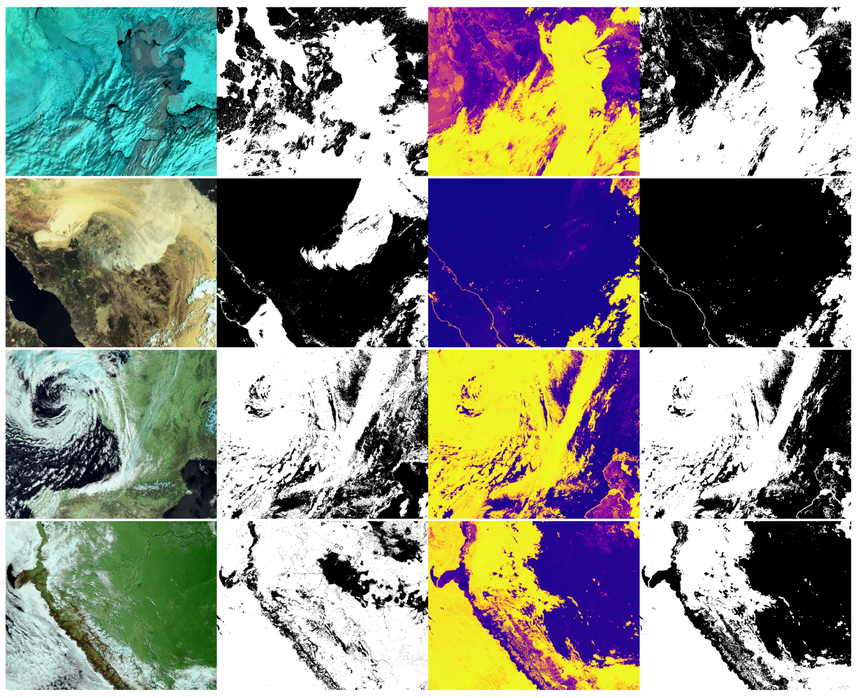
For training data we use a hand-labelled dataset consisting of ~18,000 pixels from 413 different level 1 products originally produced for validating the Bayesian mask.



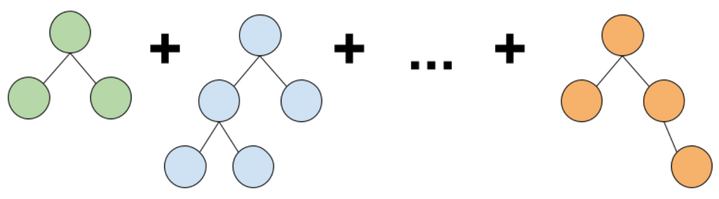
Above: Global distribution of labelled pixels. Size indicates the number of labelled pixels in each product. Colour indicates day vs. night.

Methods

The input features we use the 9 NIR, SWIR, and TIR channels. Additionally, we compute spatial features over 5x5 windows centred on each labelled pixel such as the min, max, mean, and standard deviation, similar to [2]. We also include auxiliary information including variables from meteorological prediction (such as TCWV), as well as binary flags for day/night and ocean/land.



Above: Example masks generated using GBDTs from a random sample of day time products. Columns (left to right) are false colour scene, Bayesian cloud mask, probability of cloud from GBDTs, and GBDTs cloud mask.



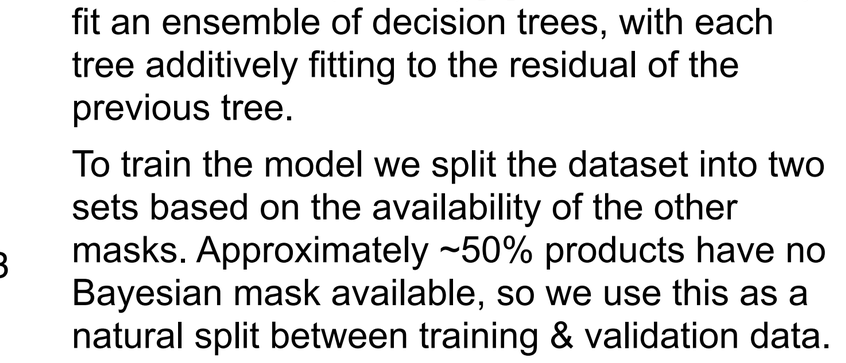
Above: Schematic illustration of fitting GBDTs. At each step the algorithm fits a new decision tree to the residuals of the previous tree ensemble.

For the model we use Gradient Boosting Decision Trees (GBDTs) [3]. GBDTs iteratively fit an ensemble of decision trees, with each tree additively fitting to the residual of the previous tree.

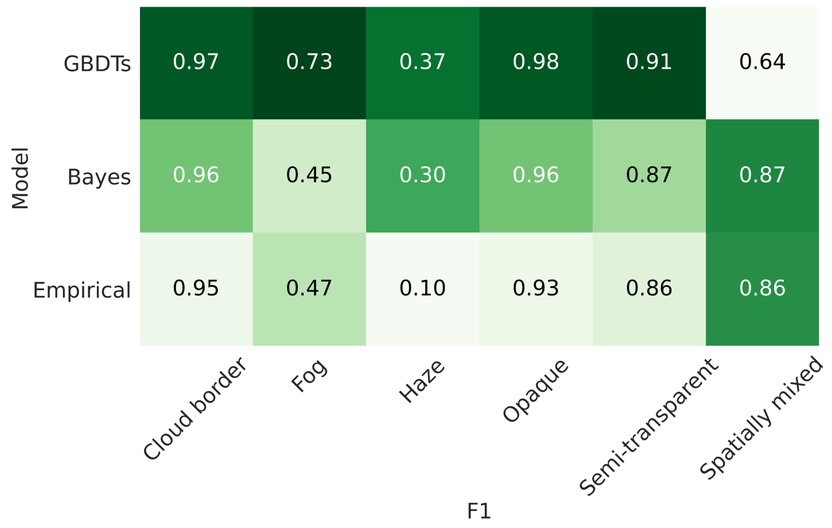
To train the model we split the dataset into two sets based on the availability of the other masks. Approximately ~50% products have no Bayesian mask available, so we use this as a natural split between training & validation data.

Results

The results of show increased performance compared to the two existing masks provided with the product across a range of different cloud types, regions, and terrain.

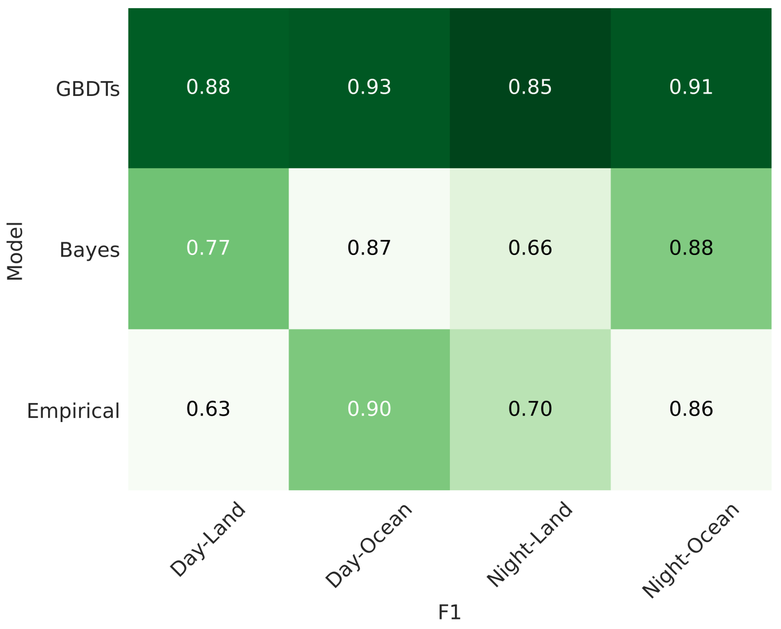


Above: Normalised confusion matrix for GBDTs, Bayesian, and Empirical model on the test set.



Above: F1 scores for GBDTs, Bayesian/Probabilistic and Empirical cloud classifiers across different types of cloud.

Below: F1 scores for GBDTs, Bayesian/Probabilistic and Empirical cloud classifiers between day & night and ocean & land.



Future Directions

Future work will aim to focus in several directions:

- Validation of model on sea and land surface temperature retrievals.
- Co-alignment & validation against CALIOP data
- Investigate Semi-, Self-supervised, and transfer learning methods.

Acknowledgements

We would like to thank Brockmann Consult for providing the hand labelled data used in this work. This work was supported by AI for Science theme from the Alan Turing Institute.

References

1. Coppo P *et al*. *Journal of Modern Optics* 2010;57(18):1815-30.
2. Kilpatrick K.A. *et al. Journal of Atmospheric and Oceanic Technology*. 2019;36(3): 387-407.
3. Chen T. *ACM International Conference on Knowledge Discovery and Data Mining* 2016;785-794.