Satellite Dataset Visual Analysis for Remote Soil Nutrient Estimation

Andrés Isaza-Giraldo[1 4], Manuel Pereira[2 3], Rafael Candeias[2], and Lucas Pereira[2 4]

[1] Faculdade de Belas-Artes, U. Lisboa, Lisbon, Portugal
giraldo@edu.ulisboa.pt
[2] Instituto Superior Técnico, U. Lisboa, Lisbon, Portugal
{manuel.afonso.pereira, rafaelmcandeiras,lucas.pereira}@tecnico.ulisboa.pt
[3] INESC-ID, Lisbon, Portugal
[4] ITI, LARSyS, Funchal, Portugal

Abstract. This paper proposes a methodology for visualizing satellite-based machine learning (ML) datasets to understand the visual components that will be used as inputs for developing ML models. The proposed methodology uses t-Distributed Stochastic Neighbor Embedding (T-SNE) methods to create visualizations of satellite images leveraging models that were pre-trained in ImageNet. T-SNE is a self-supervised learning tool used to transform high-dimensional spaces into two- or three-dimension embeddings, making it easier to visualize a broad dataset in a single image or space. The methodology is demonstrated using the LUCAS Copernicus dataset with satellite images from Sentinel-2. The dataset was constructed using the TerraSense Toolkit (TSTK) and information from the LUCAS Survey, an effort of the European Soil Data Centre. The T-SNE visualization tool aims to improve ML research by providing a clearer understanding of satellite-based datasets.

Keywords: Satellite imagery · t-SNE (t-Distributed Stochastic Neighbor Embedding) · Soil Sensing · Machine-Learning · Agroindustry

1 Introduction

Soil surveying is a process that reveals crucial information about the soil composition, such as nutrients, that could have a great impact towards a more efficient and effective use of land. Traditionally this has been a time consuming and expensive process that requires sampling on the field and intensive testing. Despite the progress done in the agricultural applications of ML using satellite images (e.g.,[1, 2]), datasets are still a challenging matter as there are still few ground-truths available and because of specific limitations of satellite photography. According to Lillesand et al. “When we look at aerial and space images, we see various objects of different sizes, shapes, and colors. (...) The images contain raw image data. These data, when processed by a human interpreter’s brain, become usable information [3].” Following this reasoning, it becomes clear that visually analyzing the dataset, trusting human capacity for interpretation, is one of the paths towards a better dataset that could ultimately improve performance of our toolkit.

Against this background, this paper proposes a methodology to visualize satellite-based ML datasets to better understand the challenges associated with the visual components that utterly affect the performance of ML algorithms. More precisely, the proposed methodology uses t-Distributed Stochastic Neighbor Embedding (T-SNE) [4, 5] methods to create visualizations of the satellite images leveraging models that were pre-trained on ImageNet [6].

T-SNE methods have been used in a wide array of disciplines from data science, medicine, social sciences, media art, etc. This self-supervised learning tool is used to transform high dimensional spaces into two- or three-dimension embeddings, making it easier to visualize a broad dataset in a single image or space. The proposed methodology is demonstrated using the LUCAS Copernicus dataset with satellite images from Sentinel-2 [7–9]. We also analyze the dataset for two specific crops, maize, and common wheat as these are among the most represented crops in the used dataset.

Although t-SNE has commonly been used to visualize the distance between satellite images according to their classification vectors e.g., there are few examples of t-SNE used to plot satellite images themselves for the sake of visually embedding datasets altogether. One prominent example would be the process of creating the interactive platform Land Lines by Zach Lieberman and Google Data Arts Team for which they plotted several satellite images together in a grid before developing their interactive experience [10]. This later type of t-SNE is more commonly used in the domain of digital arts, its application could have a great impact on machine-learning dataset visualization, for example for dataset clearing purposes.

2 Materials and Methods

This section first describes the datasets and tools uses in this research. Then, the proposed methodology is described in detail.

2.1 Dataset and Tools

The dataset was constructed using information given in the LUCAS Survey, an effort of European Soil Data Centre (ESDAC) for which there have been collected thousands of samples of soil on the European Union describing some of its properties. The ground truths taken were the ones intersecting three different surveys as observed in Fig. 1: the LUCAS Topsoil Analysis that contains information about soil composition, the LUCAS Copernicus that contains information about land use, and LUCAS 2018 that contains GPS coordinates. The LUCAS Copernicus survey is not a complete dataset per se as it does not contain aerial images of the fields. The images were obtained from Sentinel-2 satellite, a hyperspectral satellite orbiting Earth since 2016.

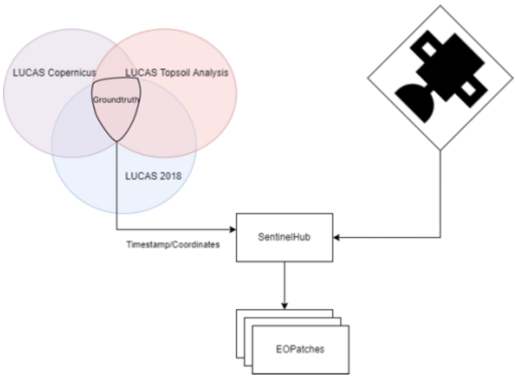

Fig. 1: Dataset construction workflow.

The dataset construction workflow was implemented using the TerraSense Toolkit (TSTK) [11]. For each specific sample of the Lucas Survey several images were taken in a range of 5 days prior and 5 days subsequent of the date in which the ground-truth was taken and as long as the maximum cloud coverage was less than 80%. This is because the satellite captures every point on earth only twice or three times a week due to its orbit around planet Earth. In many cases clouds might block the view to the direct soil rendering the images unusable.

The Sentinel-2 uses a hyperspectral camera with 12 different bands ranging from the visual spectrum to short-wave infrared. The spatial resolution for the visual spectrum is 10m while it varies from 10 to 60 m on other bands. The LUCAS Survey also contains a polygon of up to 51 meters defining an area around the ground-truth where soil uses does not change. The TSTK downloader downloads the polygon and an extra margin in all directions. Resolution of images may vary depending of the size of the polygon. Then images of the visible spectrum were created using red, green and blue bands. Median resolution for the RGB images is 162 x 217px. For the first tSNE presented in this paper, all of the images were exported with a white background and then cropped to 100 x 100px resolution. For later tSNEs images were exported with no background and cropped to 100 x 100px as seen in Fig. 2. Images with smaller resolution were upscaled to match this resolution.

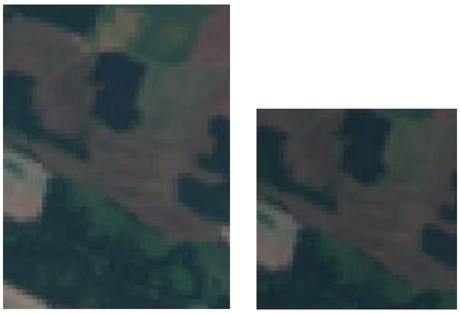

Fig. 2: RGB image with no background and image cropped to 100 x 100 px.

2.2 Feature Extraction and t-SNE

An overview of the proposed methodology is given in Fig. 3. In order to analyze the high amount of data, t-Distributed Stochastic Neighbor Embedding (tSNE) images were created. tSNE is a commonly used clustering method for visualizing high dimensional spaces on a plane or tridimensional space. The VGG16 keras classification model [12], a convolutional neural network (CNN) pretrained on ImageNet, was utilized for high-level feature extraction (see Fig. 4). Although the ImageNet dataset consists of images and labels of specific objects, animals, and people, the current dataset contains a different type of image often with no clear subject and smaller resolution. Whatsoever, the last layer, used for classification on the 1000 categories of ImageNet is ignored, and only the weights for the 4096 connections on the last fully connected layer, whose weights correspond to high-level features. Then a Principal Component Analysis (PCA) extraction takes place, restricting the number of features to only the most relevant.

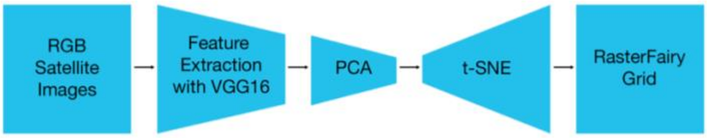

Fig. 3: Workflow diagram of the proposed methodology.

All the images are then plotted with tSNE to a two-dimensional plane, where neighboring relationships indicate similarity of features. This is done through an iterative random process that ultimately optimizes probability of distribution through a stochastic process. On a last step the images are replotted into a grid using RasterFairy, tool created by Mario Klingemann [13]. This tool raster the images into a regular structure whilst trying to preserve neighboring relationship.

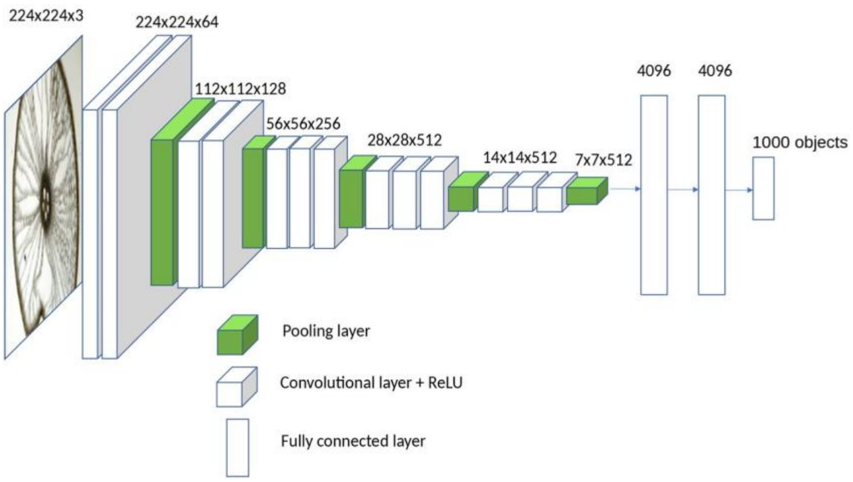

Fig. 4: VGG16 architecture.

3 Results and Discussion

This section presents and discusses the obtained results. First, the results are presented for the entire dataset. Then, two specific crops, common wheat and maize, are explored in more detail.

3.1 Entire Dataset tSNE

The whole dataset consists of over 50.000 captures of soil for over 15.809 different points or soil ground-truths. Meaning there might be more than 1 image for any point. The images were created using only the red, green and blue channels of the patches. Then they were automatically cropped to 100 x 100px due to the size of the capture.

The t-SNE shown in Fig. 5 was made with a random sample of 10.000 images. Primary Component Analysis (PCA) was restricted to 300 features.

Some obvious conclusions can be drawn from the t-SNE. The first being the persistence of the clouds on the dataset even though the clouds were already maxed out they still take about 20% of the dataset. In the blue shaded section of the tSNE at the bottom, it could be observed the presence of heavy moisture due to atmospheric conditions that blur the view to the soil. All of these images add a lot of random noise to the dataset as no specific features can be extracted from there.

On the lower right side there are a set of images that have white lines on their margins. These images have a smaller resolution than 100px wide because the polygon provided by the LUCAS Copernicus Survey was smaller than average, thus when applied the center crop an extra white margin was left. In future experiments images with less than 100px wide were upscaled.

There is also some clustering of black images or partially black images on the bottom of the tSNE, towards the center and center right. The absence of visual information is due to the satellite camera not covering the whole of the requested area during its orbit.

There are also some reddish images clustered on the top left. Whatsoever some other reddish images were grouped alongside clouds on the lower center right. Further analysis indicates that some of these images are also brighter than some of the clouds. It is possible that VGG16 assigned similar features to dim clouds and bright reddish soil, which ended up grouping them together after the stochastic clustering.

The rest of the dataset seems to be clear and distinguishable. A wide variety of greens and browns is observed. But also, there is a high presence of distinct forms due to cultivation area; presence of roads, houses and other structures; and topographical features.

3.2 Common Wheat and Maize Case Study

The TSTK was used to train individual models for each crop present on the LUCAS Survey, being that each crop has different visual properties related to soil nutrient. From previous analysis it was stablished that the model trained on Common Wheat was the most accurate one to predict soil nutrients while the one trained on Maize was the least accurate [11]. The biggest difference for each crop is the size of the dataset. As seen in Fig. 5 the number of samples for common wheat is 2.5 times larger than that for maize (4.133 and 1.796, respectively [7]).

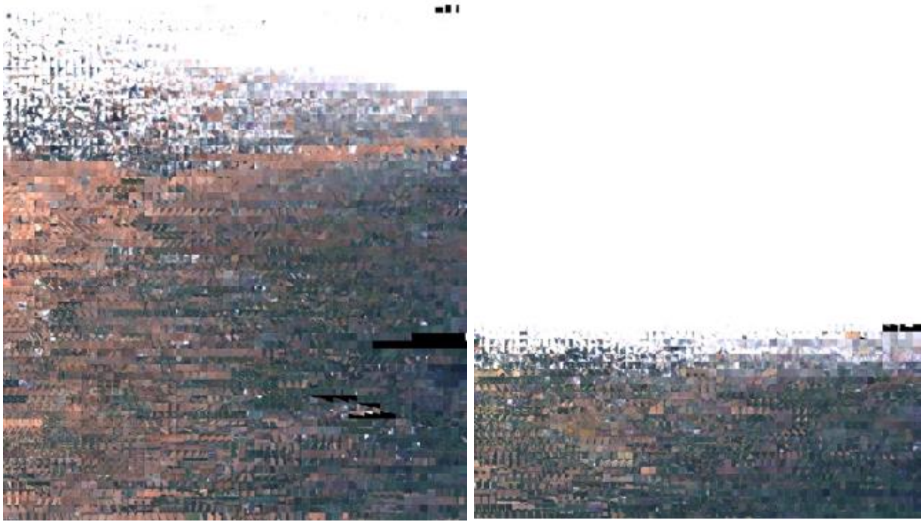

Fig. 5: tNSE representation for individual crops: common wheat (left), maize (right).

Furthermore, the t-SNE differentiates two different types of images for each crop, possibly for when the plant is fully grown, and for when it is harvested or recently planted as observed on the detail shown in Fig. 7. This distinction is much clearer on wheat than maize. It could be hypothesized that the TSTK model is able to understand these differences and make separate calculus for when there is leaf presence and when there are no leaves.

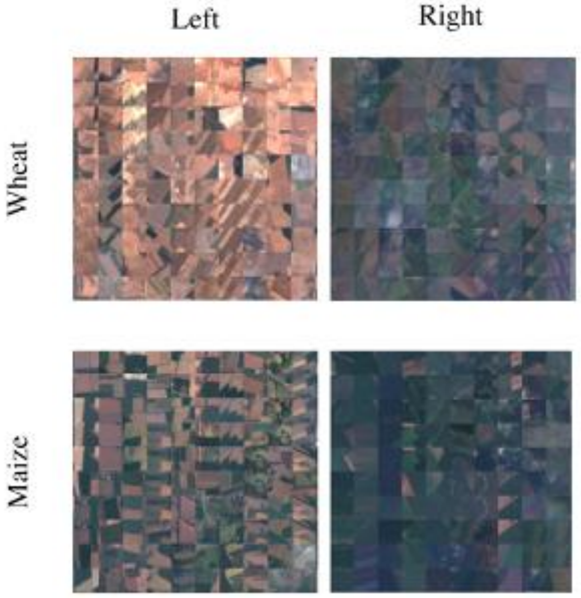

Fig. 6: Details of different clusters generated for each crop showing the different stages of plant growth.

4 Conclusions and Future Works

Satellite images, regardless of the spectrum they captured, are images whatsoever and they contain visual information. To really understand what is happening inside a satellite images dataset it is important not only to look at the visual information but to find ways to organize it to make it easier to comprehend by a human user. For such reason, using a t-SNE is a useful solution to better look at visual information altogether and so to explore relationship within images but also to understand the possible challenges driven out of the dataset creation, composition, etc.

Satellite images, such as the provided by the Sentinel-2, are a cost-efficient way to obtain visual information from the soil. Whatsoever information regarding soil composition is still limited and more robust dataset are needed to actually make a big picture of soil nutrients all across the European Union and the whole of the world.

4.1 Future Work

The current dataset is still subject to different types of visual analysis. On one hand it is needed to study the visual spectrum for each of the crops that are not cited on this article to understand the how much visual information is actually present on the dataset.

On the other hand, it is needed to be understood which of the bands of the Sentinel-2 satellite are being useful for the study of each crop. As suggested by Wand and Wei [14, 15] certain bands might be more useful than others to estimate the presence of chlorophyl in leaves which can be an indication of certain nutrients on the soil, specifically nitrogen for the study. The bands depend on the specific crop but some common bands might also be determinant. Lu et al. experiments propose that for the estimation on potassium from plant canopy certain near infrared and shortwave infrared are more adequate for such task [16]. It would be interesting to see t-SNE experiments with the different bands provided by the satellite to actually understand how it clusters differently to the visual spectrum cluster.

According to Mandrake study to detect sulfur on the soil [17] although some sulfur might be easy to differentiate though aerial imagery, some other might have similar visual properties to non-sulfur components which might lead to false positive results. For such reason it is needed to study which visual characteristic might lead further algorithms to provide false positive results.

Visualizing the results of the t-SNE for the present study, it has been suggested several times that this kind of image clustering could be applied through an interactive interface for both dataset analysis and dataset cleaning. This kind of tool would be extremely useful for cleaning larger datasets avoiding going through images and files.

However, what could have the biggest impact in the success of the TSTK would be an improved dataset of soil ground-truths alongside crop type and GPS coordinates. Although the effort of the LUCAS Survey has been extremely useful, the dataset is not big enough for machine-learning algorithms result to be fully trusted. Further improvement and refinement of our dataset is going to be performed, but help of agricultural institutions might play a big role in the further development on this cost-efficient technology.

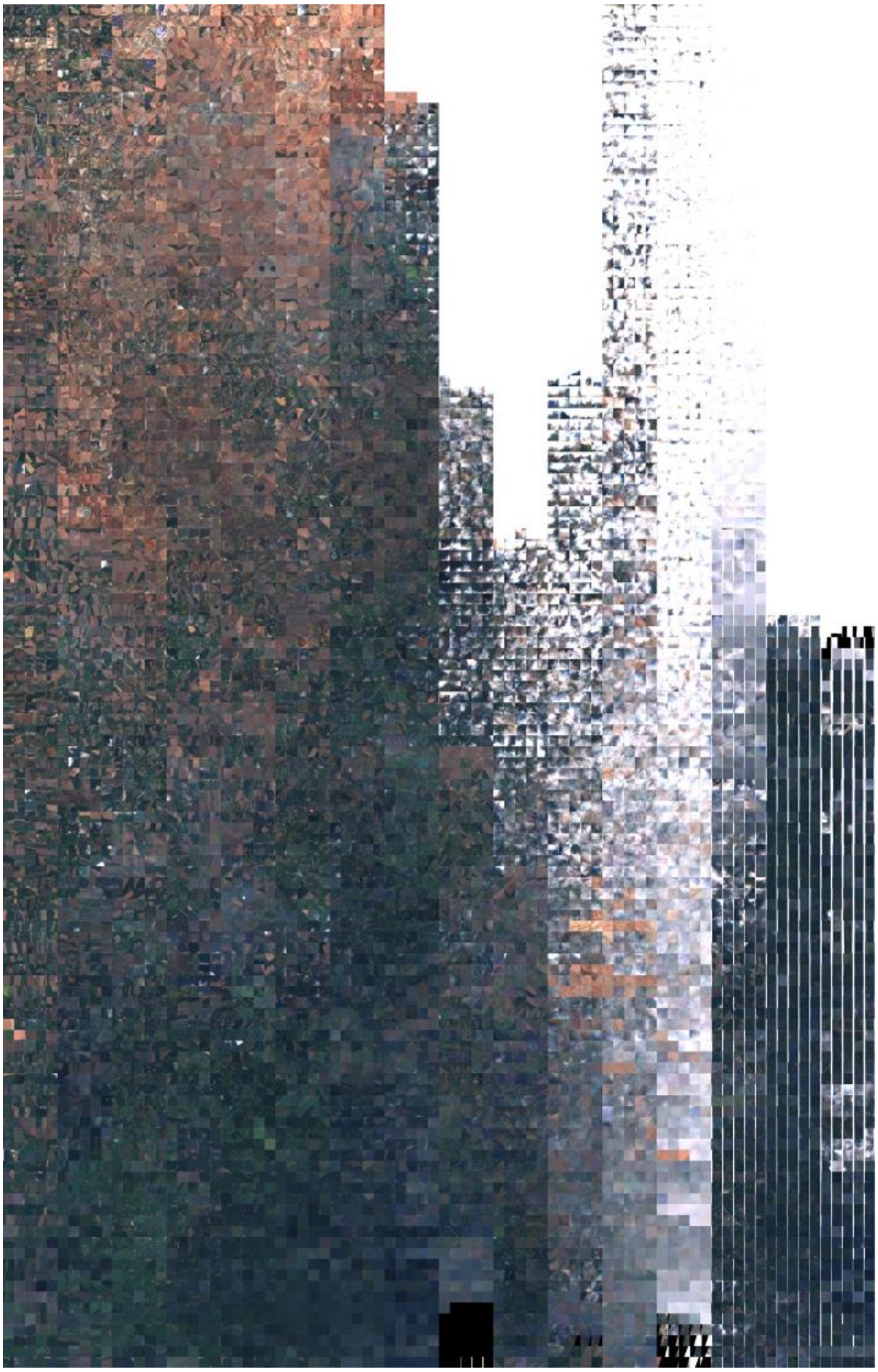

Fig. 7: tNSE representation for the entire dataset.

References

1. Karthikeyan N, Shashikkumar M, Ramanamurthy J (2010) A study on vegetation vigour as affected by soil properties using remote sensing approach. https://doi.org/10.1109/RSTSCC.2010.5712811
2. Walshe D, McInerney D, De Kerchove RV, Goyens C, Balaji P, Byrne KA (2020) Detecting nutrient deficiency in spruce forests using multispectral satellite imagery. Int J Appl Earth Obs Geoinformation 86:101975. https://doi.org/10.1016/j.jag.2019.101975
3. Liu JG, Mason PJ (2016) Image Processing and GIS for Remote Sensing: Techniques and Applications, 2nd edition. Wiley-Blackwell
4. Cai TT, Ma R (2022) Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data
5. Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov):2579–2605, 2008.
6. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Miami, FL, pp 248–255
7. d'Andrimont R, Verhegghen A, Meroni M, Lemoine G, Strobl P, Eiselt B, Yordanov M, Martinez-Sanchez L, van der Velde M (2021) LUCAS Copernicus 2018: Earth-observation-relevant in situ data on land cover and use throughout the European Union. Earth Syst Sci Data 13:1119–1133. https://doi.org/10.5194/essd-13-1119-2021
8. Tóth G, Jones A, Montanarella L (2013) The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environ Monit Assess 185:7409–7425. https://doi.org/10.1007/s10661-013-3109-3
9. European Commission. Joint Research Centre. (2020) Assessment of changes in topsoil properties in LUCAS samples between 2009/2012 and 2015 surveys. Publications Office, LU
10. Liebermann, Zach (2016). Land Lines. https://zachlieberman.medium.com/land-lines-e1f88c745847
11. 1Pereira MAS (2022) TerraSenseTK: a toolkit for remote soil nutrient estimation. MasterThesis
12. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition
13. Klingemann M (2022) About RasterFairy-Py3
14. Wang W, Yao X, Tian Y, Liu X, Ni J, Cao W, Zhu Y (2012) Common Spectral Bands and Optimum Vegetation Indices for Monitoring Leaf Nitrogen Accumulation in Rice and Wheat. J Integr Agric 11:2001–2012. https://doi.org/10.1016/S2095-3119(12)60457-2
15. Wang L, Wei Y (2016) Revised normalized difference nitrogen index (NDNI) for estimating canopy nitrogen concentration in wetlands. Optik 127:7676–7688. https://doi.org/10.1016/j.ijleo.2016.05.115
16. Lu J, Eitel JUH, Jennewein JS, Zhu J, Zheng H, Yao X, Cheng T, Zhu Y, Cao W, Tian Y (2021) Combining Remote Sensing and Meteorological Data for Improved Rice Plant Potassium Content Estimation. Remote Sens 13:3502. https://doi.org/10.3390/rs13173502
17. Mandrake L, Wagstaff KL, Gleeson D, Rebbapragada U, Tran D, Castaño R, Chien S, Pappalardo RT Hyperspectral Sulfur Detection Using An Svm With Extreme Minority Positive Examples Onboard EO-. 12