Reducing Crop Burning to Improve Air Quality in India

Sunayana Ghosh
Cervest and Solve for Good
sunayanag@gmail.com

Carlos Mougan
University of Southampton and Solve for Good
C.Mougan@soton.ac.uk

Robert Berry
Energy Harvest Trust
robert@energyharvesttrust.com

Priyadeep Kaur
Energy Harvest Trust
priyadeep@energyharvesttrust.com

Sukhmeet Singh
Energy Harvest Trust
sukhmeet@energyharvesttrust.com

Niharika Arora
Google
niharikaarora@google.com

Rayid Ghani
Carnegie Mellon University and Solve for Good
rayid@cmu.edu

Abstract

A large proportion of crop residue produced in India is burned every year, leading to environmental and public health issues. This project, a collaboration among Energy Harvest Trust, Carnegie Mellon University, and Solve for Good, aims to tackle crop burning and improve air quality in India. The initial goal of the work was focused on understanding: 1) how much crop waste is burned, 2) where it was burned, and 3) what type of crop was. This paper describes the problem we are tackling, our approach, and early results that show the promise of using simple AI methods that can help achieve our goal.

1. The Problem

India produces $500 - 550$ million tonnes of crop residue every year, most of which is burned by the farmers. This residue burning causes serious air pollution that is responsible for public health emergencies every winter as well as environmental damage. Energy Harvest Charitable Trust (EHS) was started with the goal to tackle the crop residue burning problem in India by educating and training farmers on managing the stubble, helping increase their income, and conducting policy advocacy to change this behavior.

While the scale and impact of the crop burning are massive, the information around the specifics of what crop is being burned, how much of it is being burned, where it is being burned, and at what times is not known to policymakers and advocates. In order to start designing a strategy and associated programs to reduce this crop burning, they need to have this information in an accurate, timely, and granular manner.

2. The Need for AI

One approach to tackle this is to create an on-the-ground effort to manually collect this data from farmers or other individuals in these regions. The effort and coordination required would of course be prohibitive and not make this work scalable. Another approach would be to use satellite imagery data and manually identify what crops are being burned and where on a continuous basis. Besides the scaling issues here as well, we found that the current resolution of available satellite imagery makes it extremely difficult for human experts to accurately identify what type of crops are being burned. Our approach has been to combine these approaches - use on-the-ground data collected by ground staff to create training data and then use satellite imagery to train machine learning models to generalize to new regions and in the future, making this process scalable.

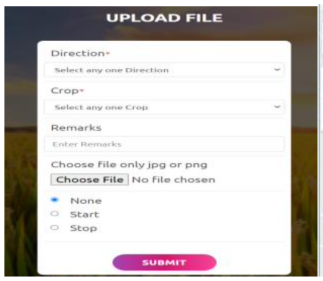
3. Our Work so Far

To answer the issues we described above, we initiated three primary activities as part of this project:

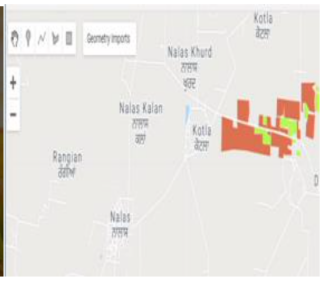
1. Getting better ground truth data
2. Creating a platform for providing information to stakeholders

3. Developing an AI-based crop-type identification model

Figure 1. Components of our system



1 App Developed



2. Use of Google Earth

3. Use of open source application

3.1. Getting better ground truth data

A key issue we faced early on in our efforts to use AI was the lack of reliable ground truth data in order to train and validate the accuracy of our system. While there was anecdotal information available on fires and crops, there was no systematic on-the-ground data collection method that would allow easy and precise data collection. Our initial activities involved a variety of methods and tools to enable the better collection of ground truth data on when and where fires may have occurred as well as precise locations of fields growing specific crops.

Approach 1: Developed a web-app for data collection on the ground

1. Data Type: Mainly for fire points and marking paddy/non-paddy fields. Marked on straight lines or polygons
2. Data Geography: Punjab, India
3. Amount of data collection: 450+ records were made by the data entry team

Challenge: Low-data network issues around farmlands, due to which the application is not able to handle records accurately while a data connection is lost.

Approach 2: Used Google Earth to mark polygons identifying areas of interest.

1. Data Type: Mainly for fire points and marking paddy fields.
2. Data Geography: Punjab, India
3. Amount of data collection: 20+ records were made by the data entry team

Challenge: Scaling is an issue in identifying a large number of fields as it involves marking up fields and fire points manually using Google Earth.

Approach 3: Used an open source application to mark the polygons on the ground

1. Data Type: Mainly for crop identification – used for paddy, cotton stalks, agro-forestry, and open fields. Marked only polygons
2. Data Geography: Punjab, India
3. Amount of data collection: 180+ records were made by the data entry team

Challenge: No specific challenge and can be used for scaling up data collection

3.2. Platform for Reporting Fires

To consolidate the biomass fire identification and reporting, a platform Carbon2Climate was developed which helps in identifying crop waste and reporting intense and frequent farm fires through an online dashboard. It uses an automated program to extract the data from NASA's Visible Infrared Imaging Radiometer Suite (VIIRS). The extracted data is then organized using the latest boundaries of states and UT's of India.

These real-time fire points are displayed on the online dashboard. Users can search and compare fire point-related information for states or districts and can also set preferences for a better user experience. The portal currently shows historical and current data and the ML model is being developed further to predict the farm fires using historical data and integrated into the platform.

Figure 2. Platform for reporting fires

3.3. AI based crop type identification model

In this section, we describe the data extraction process along with the extraction parameters, analyze the extracted field data, and feed the data to an AI system in order to create a classifier that can detect fields and the type of crop planted.

Once we are able to detect crops and fields, we use the Google Earth Engine and FIRMS data to extract satellite images of the fires. Then we run our field and crop detector to identify fields and crop types. By aggregating this information, we are able to estimate the quantity of burned biomass. Details of our approach and source code is available at our github repository[1].

3.3.1 Data Extraction

Using the provided ground truth data and the Google Earth Engine, data were extracted for 505 Paddy fields, 316 Non-harvesting fields and 64 Cotton Fields for 2020. The data is publicly available [link]

3.3.2 Analysis of Extracted Data

After extracting the data, we provide different visualizations of the data prior to the modeling. This is done to provide evidence and support that there is meaningful statistical information that can be learned during the building of the AI models.

During the harvesting months from July to November, the backscatter index shows different trends. SAR indexes, publicly provided by sentinel satellite using Google Earth Engine gave much better results compared to RGB bands which had issues around lower resolution and occurrence of cloud cover in monsoon seasons. Polarized EM values (VH and VV backscatter), soil moisture, land temperature, and NDVI index were studied as features, where VH backscatter plots provided significant differentiation between paddy, non-field, and cotton crops and were primarily used as features for our classification model.

Our preliminary results (Figure 3 and Figure 4) highlight the difference in the statistical distributions and show that it is feasible to distinguish between a paddy field and a non-paddy area, particularly during the harvesting season, leading us to use this information to train an AI model.

3.3.3 Crop identification modelling approach

Here, we describe our approach to building a classification model that can distinguish between the two different types of regions: Non-Field and Paddy Field using the information extracted from Google Earth Engine. The model receives pixel samples of 10m, for the last 3 months of historical data and we perform training and validation on different splits of the data. We use AUC (under the ROC curve) as a macro evaluation metric that illustrates the ability of the model to separate paddy fields from other types of areas. An AUC of 0.5 is equal to doing random predictions, which can be used as a naive baseline for interpreting the model performance results. The model we have currently selected is a generalized linear model with an L1 regularization term, also known as Lasso.

[1] https://github.com/ProdigyNumbers/energy_harvest_trust

Figure 3. Yearly evolution of VH Backscatter

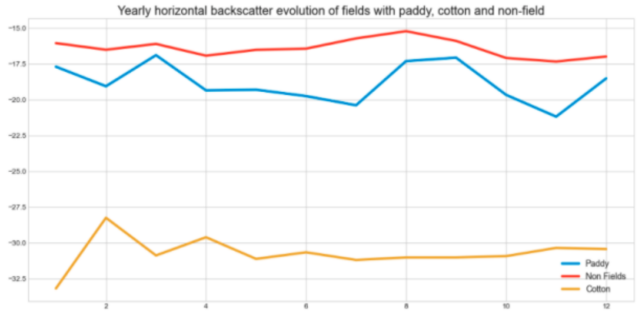
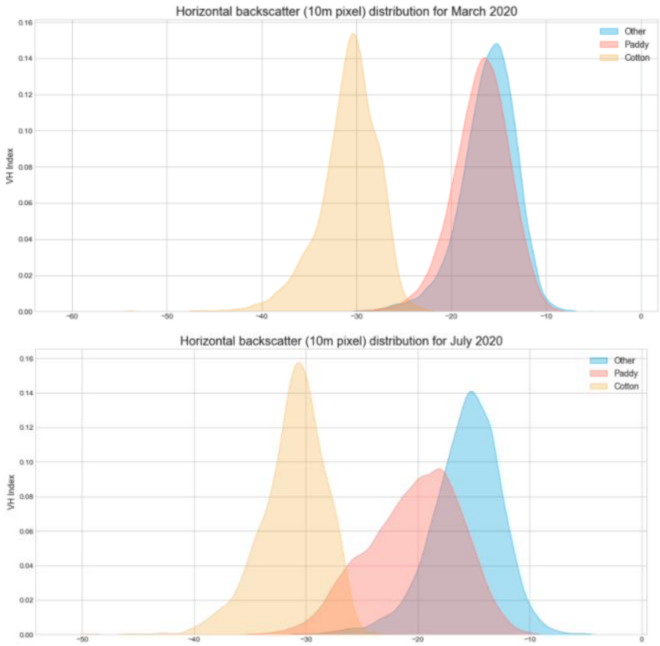


Figure 4. Differences in distributions between paddy and a non-paddy area

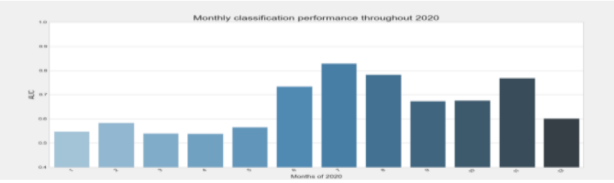


4. Limitations

While our work shows a lot of promise, it has several limitations including:

- The fire detection is based on the data provided by

Figure 5. Classifier used to distinguish between Paddy and Non-Paddy fields



FIRMS.

- Our model relies on ground truth label collection that can suffer from potential selection biases.
- The evaluation of the models is based on the state of Punjab, the model performance on out-of-state data remains unknown and is part of future work.

5. Reproducibility Statement

To ensure reproducibility, we make the data, data preparation code, code repositories, and methods publicly available at our GitHub repository (http://github.com/ProdigyNumbers/energy_harvest_trust)

6. Impact

Biomass Fire Reporting: Although the initial goal of the project was to map straw-burning fires in the state of Punjab, we were able to extend the platform for mapping the fires from the states of Punjab, Haryana, Madhya Pradesh, Uttar Pradesh, Andhra Pradesh, and Jharkhand. More than 900,000 fire points were mapped for the years 2020 and 2021.

On-the-ground data collection: 700+ on-the-ground data collection points were mapped for fires and crop types, allowing us as well as others interested in supporting these efforts to build and validate more scalable solutions.

Straw Collection: During the process of on-the-ground data collection, Energy Harvest also collected more than 50 tonnes of straw from farmers resulting in their 7% income generation and prevention of air pollution.

7. Next steps

After getting a better understanding of 1) How much crop waste is burned, 2) where it is getting burned, and 3) What type of crop waste is getting burned, we plan to focus on the following tasks in the next phase:

- **Intervention**: By creating a marketplace for crop residue and connecting farmers, collectors, and buyers of crop residue.
- **Policy advocacy**: Inform government about areas where crop waste is getting burned and also do advocacy for policy changes required for effective use of crop waste.