Crop Residue Burning in Telangana

Jesse Swinkels
Student number: 2031835

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Business Analytics and Operations Research
Department Econometrics and Operations Research
School of Economics and Management
Tilburg University

Thesis committee:

Joris Wagenaar
Parvathy K Krishnakumari

Tilburg University
School of Economics and Management
Department Econometrics and Operations Research
Tilburg, The Netherlands
December 2022

Acknowledgements

I would like to express my gratitude towards Joris Wagenaar and Parvathy K Krishnakumari for providing supervision of my thesis. Their generous feedback has improved the quality of this research significantly.

Additionally, I am thankful for the opportunity to perform this research project for Zero Hunger Lab and United Nations Development Program. Throughout this process, both teams provided valuable knowledge that has substantially contributed to this research paper. Besides, I greatly appreciated the enjoyable workplace that Zero Hunger Lab offered me.

Lastly, I would like to thank my friends and family for their expertise and moral support. They have kept my spirits and motivation high during this thesis. I wish to extend my special thanks to Mieke Couwenberg. I could not have completed this work without her contribution.

Crop Residue Burning in Telangana

Jesse Swinkels

In this paper, we aim to understand why farmers frequently perform crop residue burning (CRB) despite its negative impact on environmental and human health. We perform data powered positive deviance (DPPD) analysis to obtain insights into which farmers are reducing the practice of CRB and vice versa. We consider the number of agricultural fires, the fire radiative power emissions, NO_2 emissions, and $PM_{2.5}$ emissions. We found that the DPPD allows the government to perform targeted interventions. Additionally, we implement a random forest model that predicts NO_2 and $PM_{2.5}$ emissions per mandal per month in Telangana. The (NO_2) model has acceptable performance ($R^2 \approx 0.68$), high interpretability, and solely uses open data. We conclude that NO_2 and $PM_{2.5}$ are not valid quantifiers for CRB in Telangana because socio-economic and environmental factors are of significantly higher importance than agricultural factors. We conclude that gathering additional data is inevitable to recommend specific interventions to the government of Telangana.

Contents

1. Introduction

People have been paying more attention to living an environmentally friendly life in recent years. They are motivated to bring sustainable habits into their lives, such as using public transportation, recycling sustainable goods, reducing meat consumption, and saving energy. However, as of today, environmental pollution is still one of the most significant problems the world is facing. Environmental pollution causes irreparable damage to the natural world and human society. Primarily India is still dealing with a lot of air pollution. Twenty-one of the world's thirty cities with the worst levels of air pollution is in India, according to data compiled in the 2021 World Air Quality Report (Thiagarajan (2022)). Six Indian cities are in the top ten. Besides, New Delhi has the highest exposure to toxic air.

India's most significant factors for toxic air are industrial pollution and vehicular emissions. However, the pollution grows more intense around May and October due to agricultural fires. It is still being determined how much the annual air pollution peak is due to crop burning. Official figures claim it to be around 10%, while others suggest it could also be higher (Thiagarajan (2022)).

In India, crops are sown during three growing seasons: Rabi, Kharif, and Zaid, of which Rabi and Kharif are the main significant ones. Kharif is from approximately June to October, Rabi from October to June, and Zaid from March to July (Shamin (2022)). Zaid comes between the two seasons and is mainly used to grow fruits. Farmers plant crops during Zaid to earn continuous income, while they have predominantly sowed Rabi and Kharif crops on their lands. Only 2% of the land area reserved for Kharif is required for Zaid (Ram (2021)). Nevertheless, because of the short time between the seasons, difficulties occur. Namely, crop residues remain in the soil after harvesting, which the farmers need to clear before sowing for the next growing season. Usually, burning leftovers is preferred as it is quick and inexpensive (Kala (2021)). We will discuss the growing seasons and farmers' interests in more detail in Section 2.2.

Crop residue burning, abbreviated CRB, emits carbon dioxide (CO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur (SOX), black carbon, particulate matter ($PM_{2.5}/PM_{10}$), and more (Yadav (2019)). These matters cause climate change and increase the chances of human cardiovascular and respiratory diseases. It is hazardous for children because their lungs are less well-developed and have higher respiration rates. A report has found that 66000 deaths in India were due to agricultural burning (Kala (2021)). Another study conducted by a professor of medicine, Vitull K Gupta, revealed that 84.5% of people were suffering from health problems due to the increased smog, including irritation in the eyes, nose, and throat and increased coughing. Furthermore, CRB decreases the quality of the fertile soil because the burning kills critical bacterial and fungal populations (Mohan 2017). These bacterial and fungal populations cause the crops to be more prone to diseases due to the loss of 'friendly' pests (Yadav (2019)). In the future, this can cause problems for the agricultural sector.

Farmers can use the residue in alternative ways, such as cattle feed, compost manure, and biomass energy. However, most of these alternatives are less time efficient than burning. Another method for farmers to process crop residue effectively is by employing agricultural machines. Unfortunately, in general, farmers lack knowledge about the existence of such alternatives. A challenge for mainly small-scale farmers is that they need the facilities to invest in alternative instruments. Many alternative instruments are costly, and the government does not offer (small-scale) farmers enough subsidy such that they can afford these machines (Yadav (2019)). Therefore CRB primarily occurs at small-scale farms. We will discuss existing alternatives to CRB, including advantages and disadvantages, in more depth in Section 2.3.

An action that the government has taken against CRB is introducing policies. For instance, in 2015, the National Green Tribunal (NGT)[1] banned CRB in Rajasthan, Uttar Pradesh, Haryana, and Punjab since these states are at the top of CRB emissions (Yadav (2019)). If farmers continue burning their crop residue, they will be imposed a fine. Since implementing crop residue management policies in certain areas, there has been a slight improvement in reducing the number of crop residue fires in these areas. However, in general, the government's implementations lack strength. It is challenging for the government to perform regular checkups and control whether farmers obey the policy. Due to this, farmers continue to burn residues every season. The government has (unsuccessfully) intervened to curb CRB repeatedly in various ways. We will discuss a more extensive overview of these interventions in Section 2.4.

We have seen that CRB has detrimental effects on the environment and human health. In the past, authorities have attempted to find a balance between environmental, and human health protection and the interests of economically vulnerable farmers by intervening and offering other alternatives to CRB. Despite their efforts, there has not been a breakthrough in reducing crop residue fires. This is mainly due to farmers not receiving enough stimulus to adapt to environmentally and human health-friendly methods. Currently, the United Nations Development Program (UNDP)[2] is working on a platform called Data in Climate Resilient Agriculture (DiCRA) in collaboration with Zero Hunger Lab (ZHL)[3]. UNDP is an organization that focuses, among other things, on sustainable development. ZHL is an organization at Tilburg University that works on finding solutions for the world hunger problem through data science. The DiCRA platform provides insights into the different environmental, socio-economic, and agricultural characteristics of different parts of Telangana, a state in India, through satellite data. During this thesis, we will be contributing to the DiCRA platform to find an answer to the following research question, with Telangana as the region of interest:

How can we utilize satellite data to find factors that influence crop residue burning? We have divided this question into two sub-research questions, which we will cover separately.

Sub-research question 1: How can we identify farming communities that are, over time, reducing or increasing the impact on the environment and health by crop residue burning?
In order to answer this question, we will first focus on identifying specific parts of Telangana that are reducing or increasing the activity of CRB over time to see what parts of Telangana are interesting to analyze.

Sub-research question 2: What changes have occurred for farmers that are increasing or decreasing the activity of crop residue burning in their environmental, socio-economic, and agricultural situations?
After we have answered the first question, we want to gain insights into the causes behind the reduction or increase in the activity of CRB in certain areas.

[1] National Green Tribunal (NGT), https://greentribunal.in/

[2] United Nations Development Program (UNDP), https://www.undp.org/india

[3] Zero Hunger Lab (ZHL),
https://www.tilburguniversity.edu/nl/onderzoek/instituten-en-researchgroepen/zero-hunger-lab

In this research, we will start by explaining some background information about the topic in Section 2. Section 2.1 covers the characteristics of Telangana, Section 2.2 the interests of farmers, Section 2.3 the CRB alternatives, and Section 2.4 the government interventions against CRB. Secondly, we will examine the literature related to CRB prediction in Section 3. Afterward, we will discuss the available (satellite) data that has been used in this research in Section 4. This section is subdivided into Section 4.1, which provides a brief explanation of the data, Section 4.2, which describes the data preparation, Section 4.3, which illustrates the missing data and how we deal with it, and finally Section 4.4, which illustrates some data exploration of the most relevant data. Subsequently, Section 5 discusses the methodology of this research. The methodology is subdivided into the data powered positive deviance in Section 5.1 and the random forest model in Section 5.2. These sections are followed by the results discussed in Section 6. This section is again subdivided into the results of the data powered positive deviance in Section 6.1 and the results of the random forest model in Section 6.2. Finally, based on our findings, we will conclude our thesis with a discussion and limitations in Section 7 and a conclusion and future work suggestion in Section 8. The discussion attempts to answer sub-research question 1, using the methods and results from the data powered positive deviance in Section 5.1 and 6.1. Besides, it attempts to answer sub-research question 2, using the methods and results from the random forest model in Section 5.2 and 6.2. The conclusion finalizes this thesis by answering the main research questions using the findings from the sub-research questions.

2. Background

In this section, we provide context to the problem of CRB in Telangana. It involves Telangana's general characteristics, farmers' interests, current existing CRB alternatives, and implemented interventions by authorities to reduce CRB.

2.1 Characteristics Telangana

Telangana is a state in India containing 33 districts, sub-divided into 594 mandals (of Telangana (2016)). A mandal is a governmental area within a district. We have provided the map of Telangana with its corresponding district names in Appendix A, Figure 1. We have only provided a map of the districts within Telangana because there exist an excessive number of mandals, making it challenging to visualize the mandal names clearly. As we see later, more than 65% of the land in Telangana is used for agricultural purposes. The amount of land that is used for agricultural practices is approximately 49.6 hectares (Ha) (of Telangana (2022)). Of the 55.5 thousand farmers in Telangana, 85.9% are smallholder farmers, meaning that they own less than 2 Ha of land. On the other hand, less than 1% of the farmers are considered large-holder farmers, meaning they own approximately 10 Ha of land (Express (2018)). The average farm holding size in Telangana is approximately 1.12 Ha. Comparing this to the situation in the Netherlands, we see that in 2016 an average farm had approximately 50 Ha. This finding shows that the farmers in Telangana are relatively small compared to those in the Netherlands (CBS (2016)). The most dominant crops shown in Telangana are rice (or paddy) (20.0 Ha), cotton (17.12 Ha), maize (7.52 Ha), soybean (2.42 Ha), chilies (0.79 Ha), and turmeric (0.50 Ha). Most of these crops are sown over the whole of Telangana. However, soybean is primarily sowed in the Northern and Central Zone of Telangana, and turmeric is mainly sown in the Northern part of Telangana (of Telangana (2022)).

2.2 Farmers Interest

In the introduction, we discussed that CRB harms the environment and human health. However, farmers are still performing the activity widely. Farmers have limited resources such as time, knowledge, labor, and finances. We want to analyze the farmers' interests extensively in this section to gain more awareness of their interests.

Time. The first difficulty the farmers face is related to the available time between growing seasons. Recall that we briefly discussed the three growing seasons in India in the introduction. Table 1 provides a detailed overview of these seasons. As previously mentioned, only 2% of the total crop area in Telangana is used for Zaid and the rest for Kharif and Rabi. Due to this division of the crop area devoted to each season, we primarily focus on Kharif and Rabi (Ram (2021)). After harvesting Kharif crops, farmers have approximately twenty to thirty days from October to November to process the crop residue and sow the next season's crop. The most crucial time for Rabi crops to grow is from mid-March to April, and the exact time for total growth is about 140-150 days. This leaves the farmers only a brief period between harvesting, processing crop residue, and sowing the successive crops. Usually, burning the residue seems a simple and time-efficient method due to the lack of time for field preparation and residue decomposition (Thandra et al. (2021)).

Growing Seasons			
Season	**Sow Period**	**Harvest Period**	**Crops**
Zaid (Summer)	March - April	June - July	Fruits, Fodder
Kharif (Autumn)	June - July	September - October	Rice, Soybean, Cotton, Maize, Turmeric, Chillies
Rabi (Spring)	October - November	March-May	Rice, Maize, Chillies

Table 1: Here, the three growing seasons of Telangana are shown, including the sow and harvest period and the most planted crops (Shamin (2022)).

Knowledge. Besides, farmers are limited by knowledge. Many farmers in India believe that burning crop residue benefits soil fertility. In contrast, previous literature states that CRB damages soil health (Yadav (2019)). Moreover, numerous farmers are unaware of all existing alternatives to CRB.

Labor. Additionally, the labor-intensive process of removing crop residues from the soil is a vital factor. Next to that, in 2005, the MGNREGA (Mahatma Gandhi National Rural Employment Guarantee Act) was introduced. This act improved the rights of employees willing to do unskilled manual work by guaranteeing wage employment of at least a hundred days for adult manual workers (Agrawal (2019)). This act caused a rise in wage costs. Due to this increase in costs and the lack of human resources, it is challenging to adapt to labor-intensive CRB alternatives (Thandra et al. (2021)).

Finances. Finally, farmers' finances are limited. Large farmers can invest in costly equipment for the preparation of the fields. However, small-scale farmers cannot do this without suffering a substantial financial risk (Thandra et al. (2021)). In order for these farmers to be able to adapt to investment-rich alternatives, the government needs to ensure there is clear evidence of profit gain or provide subsidies for high-cost alternatives.

Be aware that these interests might also be related. For instance, farmers are limited by time because they do not possess the financial resources to afford time-efficient alternatives, or they are unaware of more profitable solutions.

To conclude, we need to consider that farmers are limited by time, knowledge, labor, and finances when continuing our research. Farmers need to acquire significant benefits or damage in order for them to adapt to alternatives of CRB. Now that we better understand the farmers' interests, we can critically evaluate crop residue management alternatives and government interventions.

2.3 Crop Residue Management Alternatives

As claimed previously, various crop residue management alternatives have been developed over the years. However, none of these methods have successfully reduced CRB. Farmers can alternatively manage crop residue using 1) in-field and 2) off-field management methods. A summary of the advantages and disadvantages of the different crop residue management alternatives is discussed in Table 2.

2.3.1 In-Field Management Methods. In-field management methods convert the crop residue back into the soil. Among the most well-known in-field management methods are compositing, surface retention and mulching, and in-situ with mechanical intensification. We will briefly cover these alternatives here.

Composting. Composting is an example of an ex-situ practice, meaning that we remove the crop residue from its natural location. The crop residue is collected, stored to decay, and returned to the soil (Thandra et al. (2021)). Compost consists of rich organic matter and is vital in sustaining soil fertility, ensuring that the nutrients present in the soil will return to the soil. This sustainable method has a high potential for increased yields as it enhances the soil's structure and nutrient value. A difficulty regarding this method is the decomposability of certain crops. Some crops are hard to decompose due to the hard stubble residues. A solution to this is the low-cost bio-decomposer technique called 'Pusa Decomposer'. It converts crop stubbles into compost by softening the hard stubbles so they can be easily mixed with the soil. Despite the bio-decomposer, this method still has problems in processing large amounts of residues.

Surface Retention and Mulching. In surface retention and mulching, fresh plant material or crop remains are added to the soil (Thandra et al. (2021)). It is an in-situ practice, meaning crop residues remain at their natural location. Mulching is a technique that leaves residues from the preceding crop on the soil without incorporating them into the soil. This method protects the surface soil from erosion by increasing organic carbon and total nitrogen on top of the soil. These methods' advantages are that retaining crop residues on the soil surface decreases weeds' growth and improves the soil's physical, chemical, and biological attributes. On top of that, the decreasing use of chemical fertilizers makes it a low-cost option. Unfortunately, the crops are buried and suffocated when excessively performing these methods. Besides, it can have opposite effects regarding fighting pests and crop diseases when over-performed. Therefore, it seems impossible for this method to take care of all crop residue.

In-Situ Management with Mechanical Intensification. In some regions, the farmer must perform a plowing operation before sowing the next crop due to the significant straw residue height (Thandra et al. (2021)). Machinery can help with this. We will discuss the following machinery: paddy straw chopper, baler, zero till seed drill, and Happy Seeder. The paddy straw chopper cuts paddy stubble into small pieces such that it mixes easily with the soil. Besides, a baler collects all straw from the fields and creates compacted bales. Next, the Zero Till Seed drill prepares the land by making it possible to sow the next crop withstanding the previous crop residue. Finally, the Happy Seeder is a tractor-mounted machine that cuts and lifts rice straw, sows wheat into the soil, and deposits the straw over the sown as mulch (CIMMYT (2019)). The technology is eco-friendly, improves soil fertility, and is time efficient. A paper has evaluated the public and private costs and benefits of ten alternate farming practices to manage rice residue, including burn and non-burning options. Happy Seeder-based systems emerge as the most profitable and scalable residue management practice. Happy Seeders are, on average, 10% to 20% more profitable than burning, with purchasing costs excluded (Shyamsundar et al. (2019)). However, farmers have concerns about investing in Happy Seeders as they can only use them for one season, unlike other machinery such as a tractor (Ram (2020)). Regarding the other methods, there still exist barriers, such as a lack of knowledge and uncertainty about new technologies regarding finances.

2.3.2 Off-Field Management Methods. Off-field management methods are methods in which farmers use crop residue for other purposes. We will discuss well-known crop residue management alternatives: paddy straw for mushroom cultivation, animal fodder, biogas, and biochar.

Paddy Straw for Mushroom Cultivation. Paddy or rice straw mushrooms are a species of fungi mainly cultivated in rice straw. Paddy straw is the main ingredient used as raw material for the mushroom culture (Thandra et al. (2021)). There is a massive availability of paddy straw in India, which is convenient for the mushroom culture. Mushroom cultivation is an environmentally safe disposal practice of the by-product paddy straw. This benefits farmers because they can sell paddy straw for additional income. Despite being an environmentally safe disposal method, farmers do not widely apply this alternative as it involves much more effort than simply burning the residue. Next, the amount of paddy residue is enormous (paddy is the most sowed crop in Telangana), making it impossible to process all residue using this technique.

Fodder for Animals. Another option to manage crop residue is using the crop residues as animal feed (Thandra et al. (2021)). Converting residues to fodder leads to sustainable and safe disposal of crop residue. In addition, farmers can use fodder for their cattle, decreasing their expenses. Unfortunately, this technique only applies to some crop types. For example, it cannot be applied to rice residues, as they contain a higher silica content. Higher silica contents make the decomposition process take much longer because it is more challenging for the microorganisms to act upon. In contrast, the decomposition process of wheat is rapid, making it a suitable crop for animal feed. Unfortunately, wheat is not a dominating crop in Telangana. Therefore, using the crop remains for cattle feed would lead to a minimal gain in Telangana.

Biogas. Biogas is a source of energy than can be produced in various ways. Generating biogas from crop residues is an effective and sustainable alternative to burning. This method assures a way to consume crop residues in an environmentally friendly way, in which we are rewarded with a high-quality fuel gas (Naresh et al. (2021)). Unfortunately, it is only possible to convert some crop residues to biogas, such as rice straw (Thandra et al. (2021)). Nevertheless, there is a high barrier to adapting to this method because farmers must find institutions willing to take their residue for biogas production. Moreover, only a few of these institutions exist in Telangana; most are located near Hyderabad, far away from most croplands.

Biochar. Biochar is a carbon-rich commodity used as a soil amendment to boost soil fertility, carbon conservation, and water infiltration (Naresh et al. (2021)). For the production of biochar, people can use rice straws. We can turn rice straw into biochar when performing a thermal decomposition at certain temperature ranges with a small oxygen supply. Biochar reduces the danger of climate change caused by greenhouse gases (CO_2, CH_4, and NO_2). The carbon footprint of biochar as a soil modification is lower than biochar as fuel. Studies demonstrate that the implementation of this method is not yet feasible. The unfeasibility of this procedure is due to transportation and the necessities for producing biochar. Besides, a clear energy balance and economic benefits are required (Naresh et al. (2021)).

We can conclude that the previously discussed alternatives have advantages and disadvantages and whether they are suitable highly depends on the situation. Reliable additional research on farmers land use, crop growth, and performance of alternatives is necessary such that the alternatives can be evaluated (Naresh et al. (2021)). It is crucial that the situation characteristics, such as the soil, crop type, and location, are appropriately assessed. Suppose we achieve that, then a suitable method that benefits the environment and the farmers can be selected.

2.4 Government Interventions

Over the years, the government of India made several attempts to curb CRB through interventions. In general, these interventions lack specificity. This is because authorities do not specify how to accomplish the goals mentioned in the interventions. In addition, the outcomes were not measured with high quantitative power. Therefore, it is challenging to determine whether an intervention was successful. We will briefly describe some laws that have been operating for many years. To start with, The Air Prevention and Control of Pollution Act (1981) states that the government should prohibit burning material that is not fuel and is likely to cause air pollution. Besides, the Environment Protection Act (1986) claims that any activities that emit environmental pollutants above the prescribed national standard should be prohibited. Finally, Section 144 of the Civil Procedure Code (CPC) states to ban the burning of paddy (Ram (2020)). Next to these older policies, the government implemented more recent interventions. Table 3 provides a summary of the most recent interventions of the government of India. This section discusses these recent interventions in more detail. Be aware that the interventions covered are from India and not specifically Telangana.

Alternatives		
Name	**Advantages**	**Disadvantages**
In Field Management Methods		
Composting	properties soil improve (increased yields), low cost decomposer available	amount of residue
Surface Retention and Mulching	low cost, minimal effort, properties soil improve (increase yields)	risk of overperforming can cause crop suffocation and crop diseases, amount of residue
In-Situ Management with Mechanical Intensification *E.g. paddy straw chopper, baler, zero till seed drill, Happy Seeder*	properties soil improve (increase yields), time efficient, Happy Seeder especially scalable and profitable	high investment costs, lack of knowledge, Happy Seeder can only be used in one season
Off Field Management Methods		
Paddy Straw for Mushroom Cultivation	sustainable disposal method, possibility of selling residue	not suitable for every crop type, higher efforts, amount of residue
Fodder for Animals (wheat)	sustainable disposal method, less costs on fodder for farmers	not suitable for every crop type (wheat is not a primary crop), time-inefficient
Bio Gas (paddy straw)	sustainable disposal method, possibility of selling residue, high-quality fuel gas as product	not suitable for every crop type, higher efforts, amount of residue
Biochar (paddy straw)	properties soil improve (increase yields), biochar as soil modification has a lower carbon footprint than biochar as fuel	not yet feasible

Table 2: This table shows a summary of the most common CRB alternatives, including their advantages and disadvantages.

National Policy for Management of Crop Residue (NPMCR) - 2014. In 2014, the National Policy for Management of Crop Residue (NPMCR) was enforced by the Ministry of Agriculture in India (National Policy for Management of Crop Residue (2014)). Via this law, the government wants to control the burning of the crop remains to prevent environmental degradation and loss of soil fertility by:

1. Promoting in-situ management of crop residue or other purposes such as fodder, mushroom cultivation, and power generation;
2. Bringing more awareness to farmers about the negative effects of CRB and the profitable alternatives;
3. Bringing financial support to farmers, which lowers the threshold to invest in other alternatives;
4. Monitoring CRB using satellite-based remote sensing technologies with the National Remote Sensing Agency (NRSA) and the Central Pollution Control Board (CPCB).

The policy did contribute to realizing objectives. We will see later how the government attempted to accomplish objectives 1, 2, and 3. For objective 4, previous research pointed out that in Punjab, the Punjab Pollution Control Board (PPCB) and the Environmental Prevention and Control Authority (EPCA) used satellite-based remote sensing technologies to define the exact locations of the burning areas around November 2015. Currently, DiCRA shows the exact locations of the fires in Telangana between 2015 and 2021.

National Green Tribunal's ban on Stubble Burning - 2015. Punjab, Rajasthan, and Haryana have shown many crop residue fires over the years. As a result of monitoring the exact locations of the burning areas in November 2015 Punjab, the National Green Tribunal (NGT) banned CRB in Punjab and later also in the states of Rajasthan and Haryana. The National Green Tribunal is a government enterprise focusing on environmental cases. Under this law, farmers still performing CRB can expect a fine between Rs. 2500 to Rs. 15000 (€30-€185) (Bhuvaneshwari (2019)). Be aware that an average person living in India usually earns around Rs. 30000 (€350) per month (Salary (2022)). The introduction of the fine resulted in a reduction of CRB of 38% in Punjab and 25% in Haryana under governance surveillance (Bhuvaneshwari (2019)). In addition, the government offered some state farmers incentives in rewards and subsidies for practicing the control measures. Unfortunately, it is very challenging for the government to perform regular checkups on the farmers, which causes the cash incentive only to reduce the CRB temporarily.

Promotion of Agricultural Mechanization for In-situ Management - 2018. In March 2018, the Cabinet Committee for Economic Affairs started promoting in-situ crop residue management in Punjab, Haryana, Rajasthan, and Uttar Pradesh through subsidies (Ram (2020)). Additionally, farmers were provided farm machinery and technical training through awareness campaigns. In 2018, the government distributed nearly 10000 Happy Seeders in Punjab and nearly 2500 in Haryana. However, it became evident that these machinery types are insufficient to manage the entire paddy crop residue within the given period described in Section 2.2. Moreover, as discussed in Section 2.3.1, the Happy Seeders raise concerns for the farmers regarding period usage limitations.

Promotion of Crop Residue Management with Ex-situ Management - 2012-2018. As previously discussed, in-situ practices cannot manage all crop residue. Consequently, the government has promoted other opportunities to manage crop residue ex-situ. Again, the states of Punjab and Haryana are focused on. One ex-situ practice the government promotes is biogas generation (Bhuvaneshwari (2019)). Ten tons of agricultural residue can be turned into four thousand cubic meters of biogas utilizing the biogas plant that has been certified by multiple academic institutes. Biogas plants offer farmers Rs. 600 (€7) to Rs. 1600 (€20) per ton of straw. One biogas plant can consume 120000 tons of stubble, the residue of approximately fifteen thousand farmers. However, even after perfect implementation of all different ex-situ practice opportunities, there are still large amounts of non-addressed surplus crop residues available (Ram (2020)). Previous literature states that these interventions caused slight reductions in CRB but have certainly not reduced it significantly (Bhuvaneshwari (2019)). Unfortunately, no numerical results are available.

National Clean Air Programme (NCAP) - 2019. The National Clean Air Programme aims to achieve a 20-30% reduction in Particulate Matter concentrations by 2024 by keeping 2017 as a base year (Khaiwal et al. (2022)). The NCAP did not explicitly state how they plan to accomplish this.

Commission for Air Quality Management in National Capital Region and Adjoining Areas Act - 2020. In 2021 the Government of India introduced a new law that handles violations of air pollution laws with heavy penalties and punishments (Khaiwal et al. (2022)). Farmers have reservations about this law since the government cannot provide sustainable alternatives. The government imposes high penalties that severely damage the farmers instead of finding solutions that also consider the farmers' interests. Additionally, this law involves a platform to curb CRB through research development, awareness, and capacity building of farmers.

Despite the numerous interventions, the authorities only make small efforts to make farmers comply with the statements, causing the laws to have little to no effect (Ram (2020)). Additionally, the government lacks specificity in interventions to achieve its goals. Furthermore, be aware there have only been interventions in a few states. Telangana is not covered at all.

Government Interventions			
Intervention	**Description**	**Year**	**Result**
NPMCR	1) Promoting alternatives 2) Bringing more awareness about negative effects of CRB 3) Bringing financial support 4) Monitoring CRB using satellite data	2014	Objective 4 has been achieved in e.g. Punjab, Haryana, Rajesthan and Telangana (DiCRA).
NGT's ban on CRP	In the states Punjab, Rajesthan, and Haryana the government imposed a fine of Rs. 2500 to Rs. 15000 if farmers were still performing CRB.	2015	Under governance surveillance a reduction of 38% in Punjab and 25% in Haryana of CRB has been observed.
Promotion of Agricultural Mechanization for In-situ Management	The government distributed 10000 Happy Seeder in Punjab, and nearly 2500 in Haryana.	2018	It became clear that these machinery types were insufficient in managing the entire paddy crop residue within the given time period.
Promotion of Crop Residue Management with Ex-situ Management	Secundary users of biogas plants offer farmers Rs. 600 to Rs. 1600 per ton of straw by promotion of the government.	2018	A reduction has been observed, but it has become evident that even with this technique there are still large amounts of non-addressed surplus crop residues.
NCAP	Aims to achieve a 20-30% reduction in Particulate Matter Concentrations by 2024 keeping 2017 as base year.	2019	*N/A*
Commission for Air Quality Management in National Capital Region and Adjoining Areas Act	The government introduces a new law that handles violations of air pollution with heavy penalties.	2021	Farmers have reservations against this law as the government offers no sustainable alternatives.

Table 3: A summary of the recent government interventions and their impact is described here.

3. Related Literature

Researchers tried to model CRB to understand its causes and consequences in the past. Based on this knowledge, reliable solutions to reduce CRB can be developed. We will discuss a few useful papers in this paragraph.

One research discusses the prediction of CRB using a Back Propagation Neural Network in Northeastern China (Bai et al. (2021)). It predicts whether a fire will occur in a 3 km x 3 km pixel (not necessarily cropland) with a 0 or 1, where value 0 represents no fire and value 1 represents a fire. It considers air temperature, relative humidity, air pressure, rainfall, wind speed, soil moisture (3 km x 3 km), harvest data (5 km x 5 km), and anthropogenic management and control policies. The model forecasting accuracy under only natural factors (anthropogenic management and control policies excluded) reached approximately 77% during 2013-2017. When considering anthropogenic and control policies, the forecast accuracy dropped to 60% in 2020. This paper is very promising as it is the first paper that includes the effects of government interventions. Despite the drop in accuracy when adding anthropogenic factors, the paper argues that the results are still acceptable because of the relative error between observed fire points and the backpropagation neural network. Although the paper obtains high forecast accuracies, this approach has some difficulties. First, neural networks are known to be black boxes, indicating that they do not feature high interpretability and explainability. Besides, neural networks perform only well in the availability of many data (with high resolutions), which is available in China but not India. Finally, in Telangana, authorities have not (yet) implemented direct control policies as discussed in Section 2.4.

Another paper aims to improve the understanding of CRB through pollution sources by estimating field-level NO_x ($NO_x = NO + NO_2$) emissions from CRB (Shen et al. (2021)). They obtain a new spatial database for Hubei, China, from 2014 to 2016 of 1 km x 1 km resolution. Features they used in their research are crop distribution maps, environmental monitoring stations, fire radiative power, and NO_2. They validated their outcomes by comparing the emission inventory with geostationary satellite observations, previous studies, global fire emission database, NO_2 densities from ozone monitoring instrument satellites, and measurements from environmental monitoring stations. They concluded that NO_x emissions from the global fire emission database were 47% lower than theirs, while other evaluations were significantly higher. These discrepancies were likely to be caused by differences in methodology and data sources. They argue that their method is reliable because of the reasonable correlations with NO_2 and their results in agricultural regions. While this paper shows us that agricultural fires are highly correlated with NO_2 emissions, problems arise in data collection due to the unavailability of high-resolution data in Telangana, India.

Finally, two other similar papers discuss the impact of CRB on $PM_{2.5}$ concentrations. The first paper states that CRB contributed 19% to 29% to $PM_{2.5}$ in Beijing and Tianjin, China (Zhou et al. (2018) and the other paper (Kumar et al. (2020)) examines the accuracy of weather research and forecasting model coupled with chemistry-generated $PM_{2.5}$ forecasts in Delhi, India during the CRB season in 2017. The second paper argues that aerosol-radiation feedback data contributes up to 25% to the prediction of $PM_{2.5}$. Additionally, they claim that including surface temperature forecasts can reduce the bias of up to 30% of $PM_{2.5}$ predictions. They conclude that air quality forecasting can benefit substantially from satellite aerosol radiation and surface temperature forecasts. Both papers give us insights into how to predict $PM_{2.5}$ values more accurately concerning CRB in India. Besides, it provides evidence that $PM_{2.5}$ is related to CRB (in China).

Unfortunately, these papers focus on predicting $PM_{2.5}$ without considering agricultural characteristics.

To conclude, previous research primarily performed case studies in China. In China, high-resolution data is more easily accessible, and control policies have been implemented by the government, in contrast to Telangana. We conclude that CRB is strongly correlated with NO_2 and $PM_{2.5}$ emissions. Altogether, significant features in the prediction of NO_2 and $PM_{2.5}$ are humidity, air pressure, rainfall, wind speed, soil moisture, harvest data, fire radiative power, aerosol radiation, and surface temperature. In addition, we see that previously the aim was primarily to obtain high prediction accuracies. In our specific research, we aim to focus on the interpretability and explainability of the prediction model. Therefore, we will visualize the changes in CRB over time and create a prediction model that can predict CRB or pollution sources from CRB while focusing on the reason behind specific prediction values. This will lead to a higher understanding of CRB and helps to provide the government insights for possible interventions in the future.

4. Data

To proceed with understanding and predicting CRB, gathering reliable data is necessary. The primary data type used in this research is satellite data. Satellite data is data in the form of scans of the earth made by satellites. This type of data allows for depicting geographic characteristics. Satellite images are very straightforward to interpret and contain much information. However, preparing this data for analysis takes time and effort. The reason behind this is that transforming satellite data involves many preprocessing techniques. We must initially prepare and gain general knowledge about the data to apply the analysis techniques to the data later in our research. Therefore, this section covers the general data description, data preparation, missing data, and exploration.

4.1 Data Description

We use different datasets in this research. Therefore, this research will briefly discuss their general description and purpose.

Field Boundaries. Firstly, there is field boundary data of Telangana and its districts and mandals available. This data is retrieved from the open data of the government of Telangana (Telangana (2001)). One GeoJSON file of Telangana contains the polygon shape of Telangana in the column 'geometry'. Another GeoJSON contains the polygon shapes (geometry) of 592 mandals (two less than described in Section 2.1) and the corresponding district name. We have initialized a unique number for each mandal because some mandals within different districts have similar names. We mainly use the field boundaries to convert data to the Telangana level, including the mandal and district levels.

Fire Data. Secondly, we use a GeoJSON file containing the thermal anomalies over 2015-2021 created by NASA, in particular, NASA FIRMS (Fire Information for Research Management Systems) as an indication of the fires (NASA (2022a)). Some of these thermal anomalies are caused by natural circumstances such as lightning strikes, lava, or hot rocks ejected from erupting volcanoes. However, most of these fires are started by humans, by accident, or by purpose. For example, forest fires, fires in buildings, or agricultural fires. The dataset contains information about the location of the fire events (point geometries) and the so-called fire radiative power (FRP) at that location. The FRP in megawatts (MW) provides information on the measured radiant heat of the detected fires. We use FRP to estimate the contribution of biomass burning. The fires represent the center of a one km by one km pixel flagged by the MODIS Fire and Thermal Anomalies algorithm. A pixel is flagged if it contains one or more fires. Note that MODIS stands for Moderate-resolution Imaging Spectroradiometer, an instrument on board a satellite that can measure specific data in the form of, for example, wavelengths of light.

Nitrogen Dioxide (NO_2). NO_2 is a gas that is released at the burning of crop residue. It belongs to the nitrogen oxides (NO_x) family and Greenhouse Gases. NO_x consists of NO and NO_2, which are essential for ozone's chemical formation near the earth's surface. Small ozone concentrations do not endanger human or plant health. However, if high NO_x concentrations are present in the surface atmosphere, NO_x can turn into ozone via a chemical reaction in the presence of sunlight. A higher ozone surface level increases the risk of human and plant health deterioration. Next, burning fossil fuels also forms NO_2. Examples of places where this occurs are in busy streets due to gasoline in car engines, at power plants due to burning coal, or at crop fields due to agricultural fires. We use the dataset of NASA NEO (NASA Earth Observations), which shows the monthly NO_2 values in billion molecules per mm^3 between the years 2004-2022. An Ozone Monitoring Instrument (OMI) available at NASA's Aura satellite measures the NO_2 values. This instrument can distinguish between aerosol types used to derive tropospheric ozone. Therefore, it can successfully measure NO_2 near the surface, resulting in a NO_2 dataset in the form of GeoTiff files at ten kilometers resolution (NASA (2010)). Previously we have seen a strong relationship between NO_2 and CRB. Later we will see that due to this relationship, we select NO_2 as a method to quantify (the emissions of) agricultural fires.

Particulate Matter 2.5 ($PM_{2.5}$). $PM_{2.5}$ refers to tiny particles in the air that are two and one-half microns or less in width. It comes from vehicle emissions or other operations that involve burning fossil fuels, including burning crop residue. Fine particles can be carried in the air over long distances as they are tiny, causing people hundreds of miles away from events to be able to measure the particular matter emission of these events. Exposure to high concentrations of $PM_{2.5}$ causes an increased risk of respiratory and cardiovascular diseases (Kumar et al. (2020)). The dataset provides values of estimated annual and monthly ground-level fine particulate matter ($PM_{2.5}$) for 1998-2020 at ten kilometers resolution in μg/m^3 in the form of NetCDF files (van Donkelaar et al. (2016)). The creators of this dataset have estimated the $PM_{2.5}$ values by combining multiple sources. They combined information from satellite-, simulation-, and motor-based sources. Finally, they used updates based on ground observations to calibrate the $PM_{2.5}$ estimates for the entire time series using a Geographically Weighted Regression (GWR). The $PM_{2.5}$ estimations were highly consistent with the out-of-sample cross-validated $PM_{2.5}$ concentrations from monitors (Kumar et al. (2020)). Previously we have observed that there exists a strong relationship between $PM_{2.5}$ and CRB. Later we will see that due to this relationship, we select $PM_{2.5}$ as a method to quantify (the emissions of) agricultural fires.

Land Cover. In addition, we use a land cover GeoTiff dataset. The Impact Observatory's deep learning AI classification model uses a training dataset of billions of human-labeled image pixels developed by the National Geographic Society to generate yearly data (Karra et al. (2021)). This dataset displays a global map of the land use in Telangana over 2017-2022, divided into ten categories: water, trees, grass, flooded vegetation, crops, scrub, built area, bare ground areas, snow/ice, and clouds (no information due to cloud cover). We derive the land cover data from ESA Sentinel-2 imagery at ten kilometers resolution, resulting in a ten-meter land use yearly time series dataset. This data is helpful as it provides information regarding the land use of a specific location. For example, it can present more background on the cause of a particular fire or illustrate where buildings and croplands are located.

Crop Area. This dataset contains information about crop areas (in Ha) of all crops that are sown in each mandal in Telangana per season (only covering Kharif and Rabi) from 2016 up to and including 2018. Thus, according to Table 1, Kharif harvest data is available from 2016 up to and including 2018, and Rabi harvest data is available from 2017 up to and including 2019. The district Hyderabad is omitted, as there is no crop area. Hyderabad is nearly completely filled with buildings, being Telangana's largest city. The dataset has been obtained from the open data of the government of Telangana (Telangana (2019)). Crop area data is considered because previous research points out that harvest data produces reliable results when included in prediction models for CRB (Bai et al. (2021)) (Shen et al. (2021)). Unfortunately, this dataset is inconsistent in spelling specific mandal names, causing dissimilarities between the mandal names in the crop area data and the field boundary data. We consider the spelling of the field boundary data to be leading. Therefore, based on a similarity score, we attempt to match as many mandal names in the crop area data to the mandal names in the field boundary data. Due to the inconsistencies within the datasets, the matching cannot be performed flawlessly, causing the crop area data to be less accurate.

Soil Type. The soil type dataset contains a global soil map of the world, including more than a hundred different soil types at approximately 300 meters resolution. The soil type of a specific area provides knowledge about its corresponding plant or crop growth and the quality of the soil. Soil types in Telangana are chromic luvisol, chromic vertisol, eutric netosol, lithosol, pellic luvisol, plinthic luvisol, and vertic cambisol. Generally speaking, luvisols are defined as soil types suitable for crop growth if moisture conditions are satisfied. Besides, vertisols are high in nutrients but have high clay content, causing them to be unsuitable for agricultural purposes. Furthermore, lithosols surfaces consist of stone and small rocks, usually on steep slopes. This soil type is unsuitable for cultivation. Finally, cambisols are soil types that show signs of soil formation by a color change or soil structure development. Soil formation is transforming rock material into soil that can support plant growth. The soil type data is a static map provided in a GeoTiff. The map was developed by a collaboration between countless soil scientists and was lastly updated at the beginning of 2022 by the Food and Agriculture Organization of the United States (Food and Organization (2007)). Currently, it is the only global overview of soil types with this range of variety. Incorporating the soil type map in the analysis leads to more knowledge regarding crop growth in specific areas.

Soil Moisture (SSM). Soil moisture is the amount of water content available in the soil. It is a critical measurement in many areas, such as agriculture, meteorology, climate investigations, and natural hazard predictions. This dataset provides soil moisture in millimeters for April 2015 until now on a three-day basis at ten kilometers resolution. The dataset is available through Google Earth Engine[4], a cloud-based geospatial platform that allows users to work with satellite data. However, Google Earth Engine retrieved the data from NASA GFSC, NASA Goddard Space Flight Center (NASA (2022b)). We include soil moisture because previous research mentions that soil moisture is an essential predictor for CRB (Bai et al. (2021)).

Land Surface Temperature day (ST). Land surface temperature gives insights into a particular area's climatic, hydrologic, ecological, and biogeochemical characteristics. This dataset represents the temperature patterns of the top millimeters of the land surface during the day for the years 2000-2022 in Kelvin daily, available at a one-kilometer resolution on Google Earth Engine in GeoTiffs. However, Google Earth Engine retrieved the data from NASA Earth. They gathered the data using MODIS (Wan (2021)). As previously seen, studies claim that land surface temperature is a crucial factor in the prediction of $PM_{2.5}$ (Kumar et al. (2020)).

Burnt Area (BA). The burnt area dataset shows areas sufficiently affected by fire(s). This is determined based on significant changes in vegetation cover, for example, reduction or loss of green material, and in the ground surface, for example, temporarily darker ground due to ash. The burnt area dataset might also give temporal information on the fire season. The dataset has been available 10-daily from 2014 to the present at 300-meter resolution through Copernicus Global Land Service in GeoTiffs (Service (2022)). The dataset shows a one if a 300 by 300-meter grid is marked as a burnt area and a zero if this is not the case.

Precipitation (PRE). Precipitation is a crucial aspect of environmental monitoring. Scientists at the USGS Earth Resources Observation and Science (EROS) Center created this dataset to deliver complete, reliable and up-to-date datasets. They have focused on combining different precipitation estimates from, e.g., NASA, such that high-resolution gridded precipitation datasets are created. These datasets are used for early warning objectives, such as trend analysis or seasonal drought monitoring. The dataset is available daily from 1981 to near-present with a resolution of five kilometers through Google Earth Engine in GeoTiffs (Funk (2022)). It has been included in this research because previous studies claim precipitation to be an essential factor in predicting CRB (Bai et al. (2021)).

Relative Wealth Index (RWI). The relative wealth index predicts the relative standard of living within countries. Researchers at the University of California, Berkeley, and Facebook combine privacy-protecting connectivity data, satellite imagery, and other novel data sources to estimate the relative wealth index of one grid of approximately one km by one km. The stationary RWI map is provided in a GeoTiff. The data was created in 2021 and recently updated in June 2022 (Guanghua Chi (2021)). Previously, we discussed that large-scale farmers feature the financial resources to invest in sustainable CRB alternatives. The RWI data allows us to gain insights into the socio-economic aspects (of farmers) concerning CRB.

[4] Google Earth Engine (GEE), https://earthengine.google.com/

Power Plant Locations. The powerplant dataset contains information about all power plants in India from 1960 until now. It contains information about the powerplant's name, location, and capacity (in MW) for a specific year. Power plants contribute, among other pollutants, significantly, depending on the capacity, to NO_2 and $PM_{2.5}$ emissions. The dataset is provided in an excel sheet format (Portal (2017)).

Population (POP). The population dataset shows the estimated population density per approximately one-by-one-kilometer grids. It is available yearly for the years 2000 up until 2020 in a GeoTiff file format. The units of the population maps are the number of people per squared kilometer (km^2)(WorldPop (2022)). This data provides a clear overview of farmers' social situation regarding whether they are established in densely populated areas or contrariwise. Additionally, it grants knowledge about the contribution of population size to NO_2 and $PM_{2.5}$ emissions.

Data Summary

Data	**Time**	**Period**	**Resolution**[5]	**File**	**Information**	**Units**
Field	N/A	N/A	N/A	GeoJSON	Telangana: geometry (polygon) Mandal: index, district name, mandal name, geometry (polygon), area (km^2)	N/A
Fires	2015-2021	N/A	1 km	GeoJSON	date, geometry (point), FRP (MW)	N/A
NO_2	2004-2022	monthly	10 km	GeoTiff	nitrogen dioxide	*bln. mol.* $/mm^3$
$PM_{2.5}$	1998-2020	monthly	10 km	NETCDF	particulate matter 2.5	$\mu g/m^3$
Crop Area	2016-2018	seasonal	mandal	CSV	year, season, district name, mandal name, sown area (ha)	N/A
Land Use	2017-2022	yearly	10 m	GeoTiff	land use categories	N/A
Soil Type	2022	stationary	300 m	GeoTiff	soil type categories	N/A
SSM	2015- 2022	3-daily	10 km	GeoTiff	soil moisture	mm
ST	2000-2022	daily	1 km	GeoTiff	land surface temperature (day)	K
BA	2014-2022	10-daily	300 m	GeoTiff	burnt area	BIN
PRE	1981-2022	daily	1 km	GeoTiff	precipitation	mm
RWI	2022	stationary	1 km	GeoTiff	relative wealth index	N/A
Power Plants	1960 - 2022	yearly	N/A	Excel	name, capacity (MW), geometry (point), year	N/A
POP	2000-2020	yearly	1 km	GeoTiff	population	$perkm^2$

Table 4: A table of the used datasets is shown here.

[5] Resolution indicates the size of the pixels within the images. Hence, 300 meters means that a pixel is 300 by 300 meters.

4.2 Data Preparation

As mentioned before, this research involves a significant amount of preprocessing due to the usage of satellite imagery data. Therefore, we will shortly discuss the preprocessing of datasets necessary to obtain the dependent and independent variables used further in this research. The dependent variables are defined as the effects of CRB's quantification method(s), and the independent variables, or features, are defined as the causes of changes in the dependent variables. Moreover, the dependent variables are quantifiers of CRB, and the independent variables are the effects of CRB. In this research, we will use the dependent and independent variables to visualize CRB changes over time and model CRB. We have selected the time window from September 2016 up to and until August 2019 because crop type harvest data is only available for 2016 Rabi up to and including 2019 Kharif. Before we explain the preprocessing techniques behind the dependent and independent variables, we will introduce some mathematical notation. A brief description of all introduced sets and variables is shown in Table 5 and Table 6.

We have chosen to perform the analysis on mandal and monthly levels because it is the most precise level that still gives accurate results given the resolution and periodicity of the available datasets (see Table 4). Consequently, we need to aggregate all dataset values per mandal per month (if the dataset is not provided in this format). We will provide more background on these topics and some of the set definitions. We start by explaining how we transform the periodicity to monthly periodicity, and after that, we will discuss the metrics used to aggregate values per mandal.

Transforming Periodicity. The set T'_{t_y} is introduced because some datasets contain multiple observations per month. Consequently, it is necessary to aggregate these values to derive monthly observation values. To achieve this, the set T'_{t_y} is defined as the set of all observations within a month t_y. For instance, if a dataset is available daily, set T'_{t_y} contains all daily observations of a specific month t_y in year y.

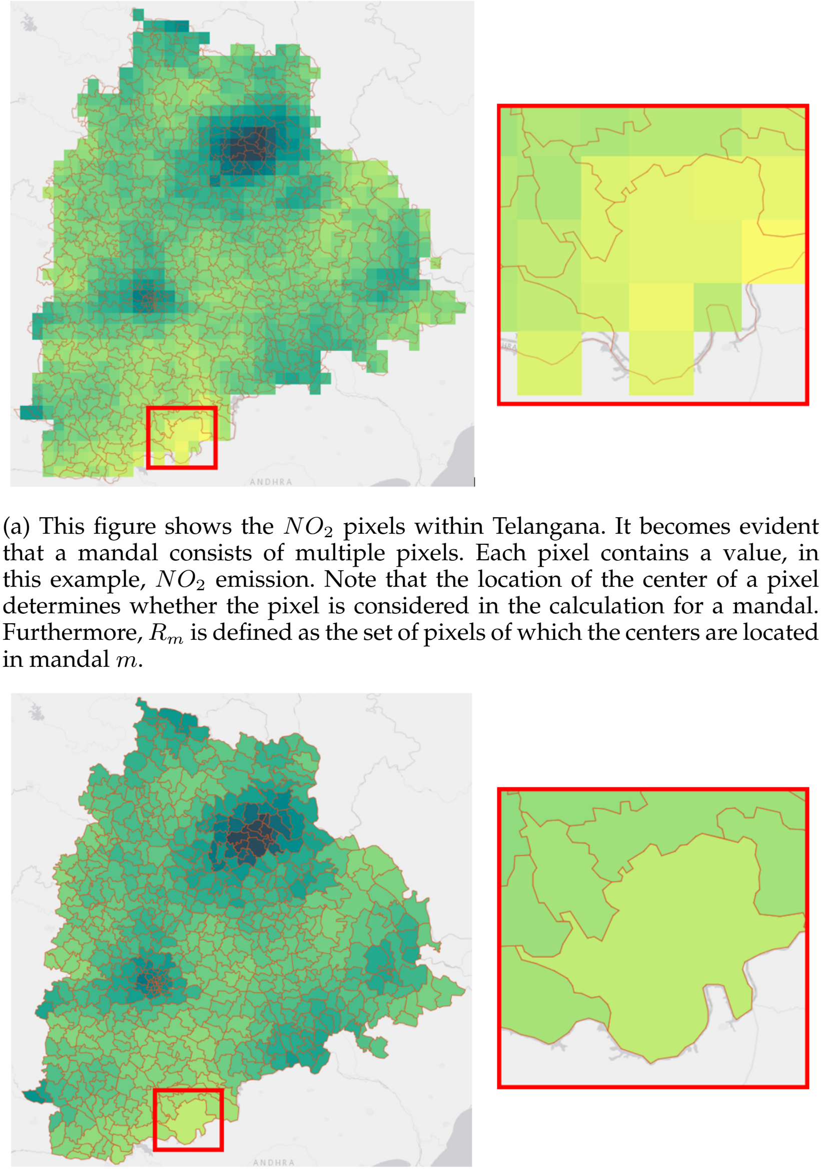
Metrics. The set R_m is introduced because mandals in some data resolution formats contain various pixels. These pixel values need to be aggregated to acquire one value per mandal. This procedure is illustrated in Figure 1. It becomes clear that in Figure 1a, mandals consist of multiple pixels. Set R_m is defined as the set of pixels of which the centers are located in mandal m. Finally, in Figure 1b, the results of averaging the values of the pixels per mandal are shown. Additionally, it is possible to calculate the median, summation, minimum, maximum, and more to gather one value per mandal. Note that we aim to find the metric with the highest explanatory power.

Set Definitions			
Set	**Indices**	**Definition**	**Examples**
M	m	mandals	mandal 1, mandal 2, ...
R_m	r_m	regions (pixels) within a mandal m	1 by 1 km pixel in mandal 1, 10 by 10 km pixel in mandal 2, ...
Y	y	years	2016, ..., 2019
T_y	t_y	months within year y	1, ..., 12
T	t	all months in the data ($T = \bigcup_{y \in Y} T_y$)	1, ..., 36
A	a	categories land use	crops, built area, trees, ...
I	i	fires	fire 1, fire 2, ...
I_{agr}	i_{agr}	agricultural/crop fires ($I_{agr} \subseteq I$)	crop fire 1, crop fire 2, ...
I_{agr,m,t_y}	i_{agr,m,t_y}	crop fires in m at t_y in y ($I_{agr,m,t_y} \subseteq I_{agr}$)	crop fire 1 in mandal 1 in month 1 in year 2016, crop fire 1 in mandal 4 in month 5 in year 2018, ...
S	s	soil types	chromic luviosols, eutric nitosols, ...
W	w	crop types	paddy, cotton, maize, ...
F	f	power plants	Vhadradri Plant, Kakatiya Power Station, ...
T'_{t_y}	t'_{t_y}	set of files belonging to month t_y in y	file of day 1 in month 5 in year 2017, file of week 3 in month 11 in year 2016, ...
P	j	independent variables	PRE, ST, RWI, ...

Table 5: Description of the sets.

Variable Definitions		
Name	**Dimension**	**Definition**
$area_m$	$\forall m \in M$	area (km) of m
β_i	$\forall i \in I$	majority land use within i ($\beta_i \in A$)
$y_{i,a}$	$\forall i \in I \;\; \forall a \in A$	pixels classified as a within i
v_{m,t_y}	$\forall m \in M \;\; \forall t_y \in T_y \;\; \forall y \in Y$	number of crop fires in m within t_y in y
$FRP_{i_{agr,m,t_y}}$	$\forall i_{agr,m,t_y} \in I_{agr,m,t_y}$ $\forall m \in M \quad \forall t_y \in T_y \quad \forall y \in Y$	FRP of i_{agr,m,t_y} in m at t_y in y
FRP_{m,t_y}	$\forall m \in M \;\; \forall t_y \in T_y \;\; \forall y \in Y$	average FRP in m within t_y in y
$NO2_{r_m,t_y}$	$\forall r_m \in R_m \;\; \forall m \in M \quad \forall t_y \in T_y$ $\forall y \in Y$	pixel NO_2 value of r_m at t_y in y
$NO2_{m,t_y}$	$\forall m \in M \;\; \forall t_y \in T_y \;\; \forall y \in Y$	average NO_2 in m within t_y in y
$PM25_{r_m,t_y}$	$\forall r_m \in R_m \;\; \forall m \in M \quad \forall t_y \in T_y$ $\forall y \in Y$	pixel $PM_{2.5}$ value of r_m at t_y in y
$PM25_{m,t_y}$	$\forall m \in M \;\; \forall t_y \in T_y \;\; \forall y \in Y$	average $PM_{2.5}$ in m within t_y in y
x_{j,r_m,t_y}	$\forall j \in P \;\; \forall r_m \in R_m \;\; \forall m \in M$ $\forall t_y \in T_y \;\; \forall y \in Y$	pixel value of j of r_m at t_y in y
$x_{j,r_m,to_{t_y}}$	$\forall j \in P \;\; \forall r_m \in R_m \;\; \forall m \in M$ $\forall to_y \in T'_y \;\; \forall t_y \in T_y \;\; \forall y \in Y$	pixel value of j of r_m at t'_{t_y} within t_y
x_{j,m,t_y}	$\forall j \in P \;\; \forall m \in M \;\; \forall t_y \in T_y$ $\forall y \in Y$	pixel value of j of m at t_y in y
$d_{f,m}$	$\forall f \in F \;\; \forall m \in M$	distance (km) between f and center of m
$c_{f,y}$	$\forall f \in F \;\; \forall y \in Y$	capacity (MW) of f in y
α_m	$\forall m \in M$	majority soil type within m ($\alpha_m \in S$)
$z_{m,s}$	$\forall m \in M \;\; \forall s \in S$	pixels classified as s within m

Table 6: Description of the variables.



(a) This figure shows the NO_2 pixels within Telangana. It becomes evident that a mandal consists of multiple pixels. Each pixel contains a value, in this example, NO_2 emission. Note that the location of the center of a pixel determines whether the pixel is considered in the calculation for a mandal. Furthermore, R_m is defined as the set of pixels of which the centers are located in mandal m.

(b) The pixels in Figure 1a are aggregated per mandal such that we acquire one value for each mandal. This figure visualizes the results of this aggregation. For this example, the mean of all the pixels has been taken to gather one value per mandal.

Figure 1: Illustration of the transformation process of the satellite NO_2 data to mandal level.

Dependent Variables

To determine the dependent variables, we will start by discussing the classification of agricultural fires. Additionally, we will examine multiple quantification methods for agricultural fires.

Classification Fires. The first step in this research is to identify agricultural fires. Recall that the fire dataset contains all kinds of thermal anomalies. Therefore, it is necessary to determine a selection of agricultural fires. A classification approach is to inspect the land cover of a specific fire event. Per fire event i, the number of pixels classified within a category a is calculated, denoted by $y_{i,a}$. These values are calculated using the geometry point columns within the fire dataset and the (categorical) pixel classifications within the land use dataset. Due to the resolution difference between the fire dataset and the land cover dataset (see Table 4), there are different land cover classes within one fire event. Consequently, we classify the land use of a specific fire based on the majority of the land type of that specific fire (Pandey (2022)). The majority of land cover for each fire event is defined as follows:

$$\beta_i = \{\text{a} \in \text{A:} \quad \arg\max y_{ia}\} \quad \forall i \in I \tag{1}$$

Be aware that β_i usually contains only one land cover class. However, when multiple land covers have the majority, β_i contains multiple land cover classes. In practice, this rarely occurs. However, given that it occurs, we select the first land cover class within β_i.

A fire event is classified as an agricultural fire if the majority of the land use of that specific fire event is category crops or flooded vegetation. We also included flooded vegetation because this land cover type usually indicates a rice (paddy) field. This leads to the following definition of set I_{agr}, given that $n = |I|$:

$$I_{agr} = \{i = 1, \ldots, n : (\beta i = crops)|(\beta i = flooded\ vegetation)\} \tag{2}$$

Quantification Methods Agricultural Fires. After determining the agricultural fires, we will focus on quantifying CRB. Given the available data, we can quantify the fires in four ways: 1) the number of agricultural fires per mandal, 2) the average FRP of crop fires per mandal in megawatts (MW), 3) the number of NO_2 emissions per mandal in billion molecules per mm^3, and 4) the number of $PM_{2.5}$ emissions per mandal in μg per m^3. We consider methods 1) and 2) to be direct ways to quantify CRB and 3) and 4) indirect ways to model CRB. Note that 3) and 4) are considered because previous literature (see Section 3) concludes that CRB is strongly correlated with NO_2 and $PM_{2.5}$ emissions.

The formulas for the four quantification methods per mandal m per month t_y of year y are shown below. Note that Equation 3 denotes the number of agricultural fires, Equation 4 denotes the average FRP emissions of agricultural fires, Equation 5 denotes the average NO_2 emissions, and Equation 6 denotes the average $PM_{2.5}$ emissions.

$$v_{m,t_y} = \sum_{i_{agr,m,ty} \in I_{agr,m,ty}} \mathbb{1} \quad \forall m \in M \quad \forall t_y \in T_y \quad \forall y \in Y \tag{3}$$

$$FRP_{m,t_y} = \frac{\sum_{i_{agr,m,t_y} \in I_{agr,m,t_y}} FRP_{i_{agr,m,t_y}}}{|I_{agr,m,t_y}|} \quad \forall m \in M \quad \forall t_y \in T_y \quad \forall y \in Y \tag{4}$$

$$NO2_{m,t_y} = \frac{\sum_{r_m \in R_m} NO2_{r_m,t_y}}{|R_m|} \quad \forall m \in M \quad \forall t_y \in T_y \quad \forall y \in Y \tag{5}$$

$$PM25_{m,t_y} = \frac{\sum_{r_m \in R_m} PM25_{r_m,t_y}}{|R_m|} \quad \forall m \in M \quad \forall t_y \in T_y \quad \forall y \in Y \tag{6}$$

Be aware that agricultural fires are not occurring each month of every mandal. When this is the case we set $v_{m,t_y} = 0$ and $FRP_{m,t_y} = 0$.

We will denote the average among all months within all years at a specific mandal m with $\bar{q}_m$, the average among all mandals at a specific month t_y in year y with $\bar{q}_{t_y}$, and the monthly average among all mandals over all years at month t with $\bar{q}_t$ for $t = 1, \ldots, 12$. Let q denote a quantification method, thus $q = v$, or $q = FRP$, or $q = NO2$, or $q = PM25$. Then $\bar{q}_m$, $\bar{q}_{t_y}$, and $\bar{q}_t$ are defined as follows:

$$\bar{q}_m = \frac{\sum_{y \in Y} \sum_{t \in T_y} q_{m,t_y}}{\sum_{y \in Y} |T_y|} \quad \forall m \in M \tag{7}$$

$$\bar{q}_{t_y} = \frac{\sum_{m \in M} q_{m,t_y}}{|M|} \quad \forall t_y \in T_y \quad \forall y \in Y \tag{8}$$

$$\bar{q}_t = \frac{\sum_{m \in M} \sum_{y \in Y} q_{m,t_y}}{|M| \cdot |Y|} \quad t = 1, \ldots, 12 \tag{9}$$

An advantage of classification methods 1) and 2) is that the outcomes are solely linked to CRB (assuming the classification method is valid). These methods are better for estimating the actual effect of CRB because the outcomes are not linked to other factors than CRB. A disadvantage is that these methods contain many zeros, causing difficulties in prediction modeling. We will discuss this later in more detail. Additionally, we previously discussed that a thermal anomaly in the fire dataset could contain multiple fire events. For simplification, we continue this research assuming that a thermal anomaly consists of only one fire event. However, remember that this causes reduced reliability of methods 1 and 2. For methods 3) and 4), this is contrariwise. NO_2 and $PM_{2.5}$ are strongly correlated with CRB, but NO_2 and $PM_{2.5}$ emissions also come from, e.g., vehicles, power plants, and more. Therefore, we cannot purely estimate the effect of CRB. However, an advantage of these quantification methods is that they do not contain any zeros. This makes these quantification methods less complicated to model. Besides, we do not require significant assumptions for the NO_2 and $PM_{2.5}$ emission values. The quantification methods will be selected considering their strengths and weaknesses. While applying these quantification methods, we consider the number of agricultural fires, the intensity of the agricultural fires, and the impact of agricultural fires on pollution emissions.

Independent Variables

We have defined different kinds of independent variable categories concerning CRB. These categories are environmental, socio-economic, and agricultural independent variables. Table 7 provides an overview of all the independent variables, including their appropriate metrics. First, we will talk about transforming each independent variable to monthly periodicity if necessary, and second, we will discuss the calculation of the metrics per independent variable.

Transforming Periodicity. As discussed, the period of the independent variables must correspond with one another. In Table 4, one can see that the period in which the data is available differs frequently. Therefore, we will transform datasets to monthly periodicity if necessary. In Table 8, the definition of monthly periodicity per independent variable is shown.

For the datasets that have a periodicity smaller than monthly, e.g., daily (ST, PRE), 3-daily (SSM), and 10-daily (BA), we can aggregate these values monthly by taking the averages. For $j \in$ {ST, PRE, SSM, BA} the values per region in mandal m at month t_y in year y are defined as follows:

$$x_{j,r_m,t_y} = \frac{\sum_{t'_{t_y} \in T'_{t_y}} x_{j,r_m,t'_{t_y}}}{|T'_{t_y}|} \qquad \forall t_y \in T_y \quad \forall y \in Y \quad \forall r_m \in R_m \quad \forall m \in M \tag{10}$$

For the harvested crop area, we only have seasonal (Kharif and Rabi) data available (see Table 4). Previously, we have defined Rabi crops to be harvested in March, April, and May and Kharif crops to be harvested in September and October (see Table 1). For that reason, we set each harvest month equal to the harvested crop area of the corresponding season. The harvested crop area will be zero if a month is not a harvest month. See the following equation:

$$x_{Harvested\_CropA_w,m,t_y} = \begin{cases} x_{Harvested\_CropA_w,m,Rabi_y}, & \text{if } t_y = 3 | t_y = 4 | t_y = 5 \\ x_{Harvested\_CropA_w,m,Kharif_y}, & \text{if } t_y = 9 | t_y = 10 \\ 0, & \text{if otherwise} \end{cases} \tag{11}$$

$$\forall t_y \in T_y \quad \forall y \in Y \quad \forall m \in M \quad \forall w \in W$$

Note that in Equation 11, we do not consider r_m, but m. This is because the crop area dataset is available at mandal resolution (see Table 4).

Independent Variables			
Name	**Abbreviation**	**Metrics**	**Type**
Environmental			
precipitation	PRE	mean, median, min, max, sum, 90th percentile	numerical
soil moisture	SSM	mean, median, min, max, 90th percentile	numerical
surface temperature	ST	mean, median, min, max, 90th percentile	numerical
burnt area	BA	count	numerical
soil type *chromic luvisols, chromic vertisols, eutric nitosols, lithosols, pellic luvisols, plinthic luvisols, vertic cambisols*	SoilT$_s$ $\forall s \in S$	majority	categorical
Socio-Economic			
relative wealth index	RWI	mean, median, min, max, 90th percentile	numerical
population	POP	mean, median	numerical
power plant score	PowP_score	-	numerical
built area	Built	count, proportion	numerical
Agricultural			
crop area	Crops	count, proportion	numerical
area harvested crops *paddy, maize, soybean, cotton, chillies, turmeric*	Harvested_CropA$_w$ $\forall w \in W$	-	numerical

Table 7: In this table, all preprocessed independent variables are shown. Recall that these variables are available monthly and per mandal.

For $j \in$ {POP, PowP_score, Built, Crops} we have yearly data. Note that Built and Crops are obtained from the land use dataset. We will discuss later how we obtain these values in more detail. We set each monthly value for these independent variables equal to the corresponding yearly value. The formula for $j \in$ {POP, Built, Crops} looks as follows:

$$x_{j,R_m,t_y} = x_{j,R_m,y} \quad \forall t_y \in T_y \quad \forall y \in Y \quad \forall r_m \in R_m \quad \forall m \in M \tag{12}$$

For $j = PowP\_score$, the formula looks slightly different. This is because the power plant score is calculated per mandal m and not per region within a mandal r_m. The exact calculation of the power plant score will be discussed later in this section. For transforming its periodicity, the following holds:

$$x_{PowP\_score,m,t_y} = x_{PowP\_score,m,y} \quad \forall t_y \in T_y \quad \forall y \in Y \quad \forall m \in M \tag{13}$$

Finally, we have two stationary datasets, meaning they only have one value over time. These variables are $j \in$ {SoilT$_s$, RWI}. To transform this to monthly periodicity, we set each month's value in time equal to this stationary value:

$$x_{j,R_m,t_y} = x_{j,R_m} \quad \forall t_y \in T_y \quad \forall y \in Y \quad \forall r_m \in R_m \quad \forall m \in M \tag{14}$$

Recall that $SoilT_s$ is categorical data and that, therefore, $x_{SoilT_s,R_m,y,t}$ will be categorical.

Monthly Periodicity Definitions		
Independent Variables	**Periodicity**	**Definition**
ST	Measured every three days	Average 3-daily soil temperature in pixel r_m in mandal m at month t_y in year y
PRE	Measured every day	Average daily precipitation in pixel r_m in mandal m at month t_y in year y
SSM	Measured every three days	Average 3-daily soil moisture in pixel r_m in mandal m at month t_y in year y
BA	Measured every ten days	Average 10-daily burnt area in pixel r_m in mandal m at month t_y in year y
Harvested_CropA$_w$	Measured every Kharif and Rabi season	Total area harvested of crop w in mandal m at harvest month (Kharif or Rabi) t_y in year y
POP	Measured every year	Population per km^2 in pixel r_m in mandal m in month t_y is equal to the yearly value of year y
PowP_score	Calculated every year	Power plant score per mandal m in month t_y is equal to the yearly value of year y
Built	Measured every year	Built area in pixel r_m in mandal m in month t_y is equal to the yearly value of year y
Crops	Measured every year	Crop area in pixel r_m in mandal m in month t_y is equal to the yearly value of year y
RWI	Measured once	Relative wealth index in pixel r_m in mandal m in month t_y in year y is equal to the stationary value
SoilT$_s$	Measured once	Soil type s of r_m in mandal m in month t_y in year y is equal to the stationary value

Table 8: Definitions of monthly periodicity for each independent variable.

Metrics. After we have transformed all data to monthly periodicity, the different metrics need to be defined. We have selected the following metrics in our research: mean, median, minimum, maximum, summation, 90th percentile, count, proportion, and majority. Not all metrics are used for each independent variable. The suitable metrics have been selected depending on the independent variable's nature. For example, the summation of the precipitation does make sense. However, it does not account for soil moisture. We have chosen both mean and median. The mean has the benefit of considering all values in a mandal, but it is sensitive to outliers. The median, on the other hand, is not sensitive to outliers but does not consider all values. Besides, we cover the 90th percentile and the maximum because the 90th percentile filters out extreme values, whereas the maximum does the opposite. Finally, we consider the proportion next to the count because proportion considers the mandal areas.

We will discuss the appropriate metrics for each independent variable. We will start by showing how to calculate the mean and the median. These metrics are appropriate for independent variables $j \in$ {PRE, SSM, ST, RWI, POP}:

$$mean_{j,m,t_y} = \frac{\sum_{r_m \in R_m} x_{j,r_m,t_y}}{|R_m|} \qquad \forall t_y \in T_y \qquad \forall y \in Y \qquad \forall m \in M \tag{15}$$

Note that for the following equation set R_m and x_{j,r_m,t_y} need to be sorted according to increasing order of x_{j,r_m,t_y} for each j.

$$median_{j,m,t_y} = \begin{cases} x_{j,\frac{|R_m|+1}{2},t_y}, & \text{if } |R_m| \text{ is odd} \\ \frac{x_{j,\frac{|R_m|}{2},t_y} + x_{j,\frac{|R_m|+1}{2},t_y}}{2}, & \text{if } |R_m| \text{ is even} \end{cases} \qquad \forall t_y \in T_y \qquad \forall y \in Y \qquad \forall m \in M \tag{16}$$

Secondly, the minimum, maximum and 90th percentile are defined for $j \in$ {PRE, SSM, ST, RWI}. Note that again, for these equations, set R_m and x_{j,r_m,t_y} need to be sorted according to increasing order of x_{j,r_m,t_y} for each j.

$$min_{j,m,t_y} = x_{j,1,t_y} \qquad \forall t_y \in T_y \qquad \forall y \in Y \qquad \forall m \in M \tag{17}$$

$$max_{j,m,t_y} = x_{j,|R_m|,t_y} \qquad \forall t_y \in T_y \qquad \forall y \in Y \qquad \forall m \in M \tag{18}$$

$$90th\_percentile_{j,m,t_y} = x_{j,\lfloor 0.9 \cdot |R_m| \rfloor,t_y} \qquad \forall t_y \in T_y \qquad \forall y \in Y \qquad \forall m \in M \tag{19}$$

Additionally, the sum is exclusively defined for $j \in$ {PRE}.

$$sum_{j,m,t_y} = \sum_{r_m \in R_m} x_{p,r_m,t_y} \qquad \forall t_y \in T_y \qquad \forall y \in Y \qquad \forall m \in M \tag{20}$$

The determination of the metrics built area (Built), crop area (Crops), and burnt area (BA) are slightly different. This is because Built and Crops are from the land use dataset (see Table 4), and this is a categorical dataset. In addition, BA is a binary dataset, thus $x_{BA,r_m,t_y} \in \{0,1\}$. Here x_{j,R_m,t_y} is defined as follows:

$$x_{Built,R_m,t_y} = \begin{cases} 1, & \text{if landuse is built area} \\ 0, & \text{otherwise} \end{cases} \cdot 0.0001 \quad \forall t_y \in T_y \quad \forall y \in Y \quad \forall r_m \in R_m \quad \forall m \in M \tag{21}$$

$$x_{Crops,R_m,t_y} = \begin{cases} 1, & \text{if landuse is crops} \\ 0, & \text{otherwise} \end{cases} \cdot 0.0001 \quad \forall t_y \in T_y \quad \forall y \in Y \quad \forall r_m \in R_m \quad \forall m \in M \tag{22}$$

$$x_{BA,R_m,t_y} = \begin{cases} 1, & \text{if area is burnt} \\ 0, & \text{otherwise} \end{cases} \cdot 0.09 \quad \forall t_y \in T_y \quad \forall y \in Y \quad \forall r_m \in R_m \quad \forall m \in M \tag{23}$$

We multiply in Equation 21 and 22 by 0.0001 because the land use dataset is available at 0.01 by 0.01 km (see Table 4) pixel sizes. For BA, the pixel size is 0.3 by 0.3 km; hence we multiply in Equation 23 with 0.09. Subsequently, we count all pixels per mandal m in period t_y and obtain the following metric for $j \in$ {Built, Crops, BA}:

$$count_{j,m,t_y} = \sum_{r_m \in R_m} x_{j,R_m,t_y} \quad \forall t_y \in T_y \quad \forall y \in Y \quad \forall m \in M \tag{24}$$

Accordingly, we obtain the proportion for $j \in$ {Built, Crops} as follows:

$$proportion_{j,m,t_y} = \frac{count_{j,m,t_y}}{area_m} \quad \forall t_y \in T_y \quad \forall y \in Y \quad \forall m \in M \tag{25}$$

For $j = PowP\_score$, we have defined a formula ourselves. This is because the amount of power plant emission in a mandal depends on the distance from a mandal to the power plant and the capacity of that specific power plant. Recall that $c_{f,y}$ is defined as the capacity of power plant f in year y and $d_{f,m}$ is defined as the distance between power plant f and the center of mandal m. We define the power plant score as follows:

$$x_{PowP\_score,m,t_y} = \sum_{f \in F} \frac{c_{f,y}}{d_{f,m}} \quad \forall t_y \in T_y \quad \forall y \in Y \quad \forall m \in M \tag{26}$$

We use dummy variables to represent the subgroups of the different soil types in Telangana. Sometimes more subgroups are available within one mandal. To gather one soil type per mandal, we take the majority soil type per mandal, α_m. Recall that $z_{s,m}$ is classified as the number of pixels classified as soil type s within mandal m. α_m is defined as follows:

$$\alpha_m = \{\text{s} \in \text{S}: \quad \arg\max z_{m,s}\} \qquad \forall m = 1, \ldots, |M| \tag{27}$$

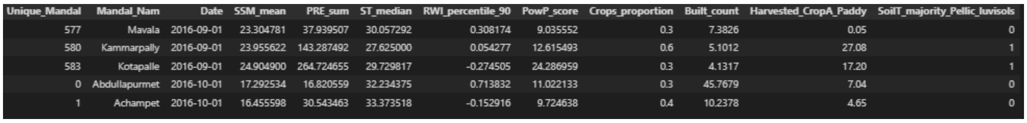
Again be aware that α_m usually contains only one soil type. However, when multiple soil types have the majority, α_m contains multiple soil types. In practice, this rarely occurs. However, given that it occurs, we select the first soil type within α_m. Recall that the set S is defined as the set of all different soil types, with $s \in S$. Then we define the soil type dummy variable majority as follows:

$$majority_{SoilT_s,m,t_y} = \begin{cases} 1 \text{ if } \alpha_m = s, \\ 0 \text{ otherwise} \end{cases} \qquad \forall t_y \in T_y \quad \forall y \in Y \quad \forall s \in S \quad \forall m \in M \tag{28}$$

No metrics calculation of area harvested crops was necessary because the dataset already shows the total production of a crop type for each mandal. We select Telangana's most sown crop types and create a column with the sown area for each crop type. The crop type set is defined as follows, $W =$ {Paddy, Maize, Soybean, Cotton, Chillies, Turmeric}. Then we have columns Harvested_CropsA$_w$ $\forall w \in W$. Recall that a summary of all the independent variables and their metrics are shown in Table 7.

Be aware that the averages of the metrics can be calculated according to Equation 7, Equation 8, and Equation 9 defined in the previous section with q equal to the metric of a variable.

Finally, to give a better overview of the final layout of the independent variables data frame, we have randomly selected five lines of the independent variables and visualized this in Figure 2. The data is now transformed to be available for each mandal and with monthly periodicity.



Unique_Mandal	Mandal_Nam	Date	SSM_mean	PRE_sum	ST_median	RWI_percentile_90	PowP_score	Crops_proportion	Built_count	Harvested_CropA_Paddy	SoilT_majority_Pellic_luvisols
577	Mavala	2016-09-01	23.304781	37.939507	30.057292	0.308174	9.035552	0.3	7.3826	0.05	0
580	Kammarpally	2016-09-01	23.955622	143.287492	27.625000	0.054277	12.615493	0.6	5.1012	27.08	1
583	Kotapalle	2016-09-01	24.904900	264.724655	29.729817	-0.274505	24.286959	0.3	4.1317	17.20	1
0	Abdullapurmet	2016-10-01	17.292534	16.820559	32.234375	0.713832	11.022133	0.3	45.7679	7.04	0
1	Achampet	2016-10-01	16.455598	30.543463	33.373518	-0.152916	9.724638	0.4	10.2378	4.65	0

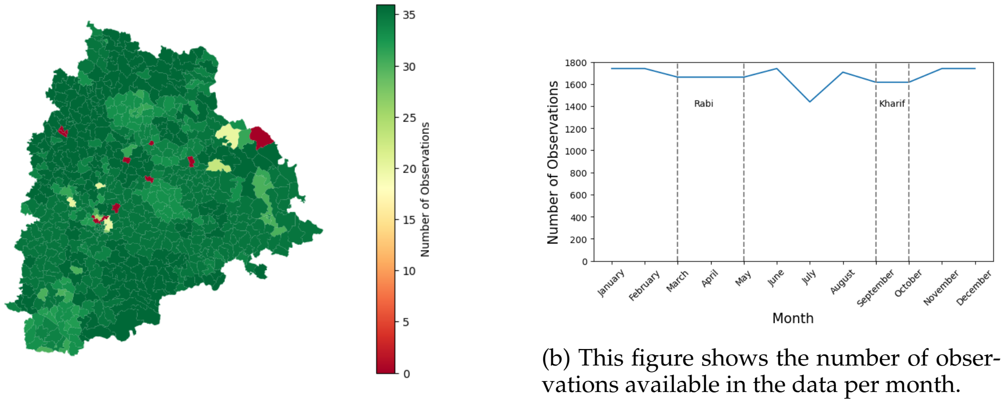
Figure 2: Some randomly selected rows and columns of the final independent variable data frame.

4.3 Missing Data

In the previous section, we discussed how we preprocessed the available data to gather dependent and independent variables necessary for our research. Remember that we have selected the time window from September 2016 until August 2019. Using this time window, we have in total 592 (mandals) x 12 (months) x 3 (years) = 21312 data rows. We observe some missing observations in the data. We will discuss here how we treat them:

- First, there is no crop data, Harvested_CropA$_w$ $\forall w \in W$, available for the following twelve mandals: Alwal, Bachupally, Balanagar, Karimnagar, Kukatpally, Wazeed, Nadikuda, Muduchinthalaphally, Narayanraopet, Mosra, Chandur, and Dhoolumitta. Therefore, we drop all rows for each month for these mandals (12 x 12 x 3 = 432 observations).
- In addition, there are some other missing values within Harvested_CropA$_w$ $\forall w \in W$. These 473 rows have also been dropped.
- Finally, there are some missing values for ST, namely 344 observations. These rows have also been dropped.

After removing the missing data rows, we have a dataset of in total of 20063 rows. In Figure 3, the number of observations per mandal and month is shown. In 3a, we see that the red mandals are the mandals that are removed entirely from the data. In general, most mandals have at least thirty (out of 36) observations in the data. In Figure 3b, we observe for July fewer data points than in other months. This is important to take into account while evaluating the results.



(a) This figure shows the number of observations available in the data per mandal.

(b) This figure shows the number of observations available in the data per month.

Figure 3: This figure shows the amount of data per mandal per month.

4.4 Data Exploration

Previously, we have seen that there are multiple ways in which we can quantify agricultural fires. We will show the results of the general data visualizations of the quantification methods in Section 4.4. After that, we will explore some of the independent variables in Section 4.4. We will give a short overview of both to obtain general knowledge about the data.

Dependent Variables

For some of the quantification methods, we need first to classify fires into agricultural fires. Therefore, we will start by showing the results of the classification.

Classification Fires. As claimed in Section 4.2, we will be looking at the land use of a specific fire event to classify a fire event as agricultural fire or other fire. Figure 4 shows the land cover of Telangana in 2018. What immediately becomes clear is that the majority of the land in Telangana has primarily agricultural purposes (crops). Crops made up more than 65% of all land use in 2018. Unfortunately, the land use dataset is only available for 2017, 2018, and 2019. As we only have observed minor differences per year, we take for 2016 land use 2017 as it is the closest available data year. In Figure 5, the crop area over time is visualized. Subsequently, we argue that it is justified to select the year 2017 from the land use data for the year 2016.

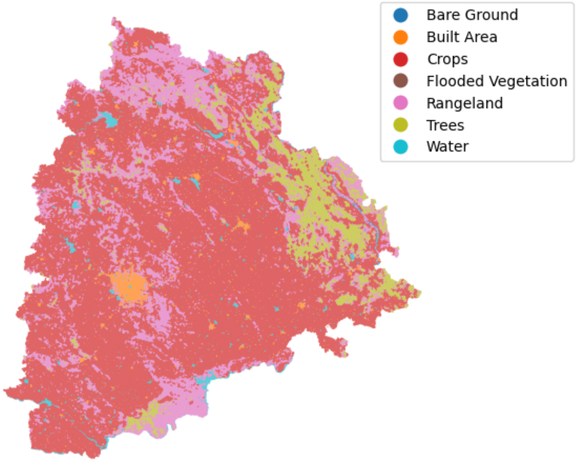
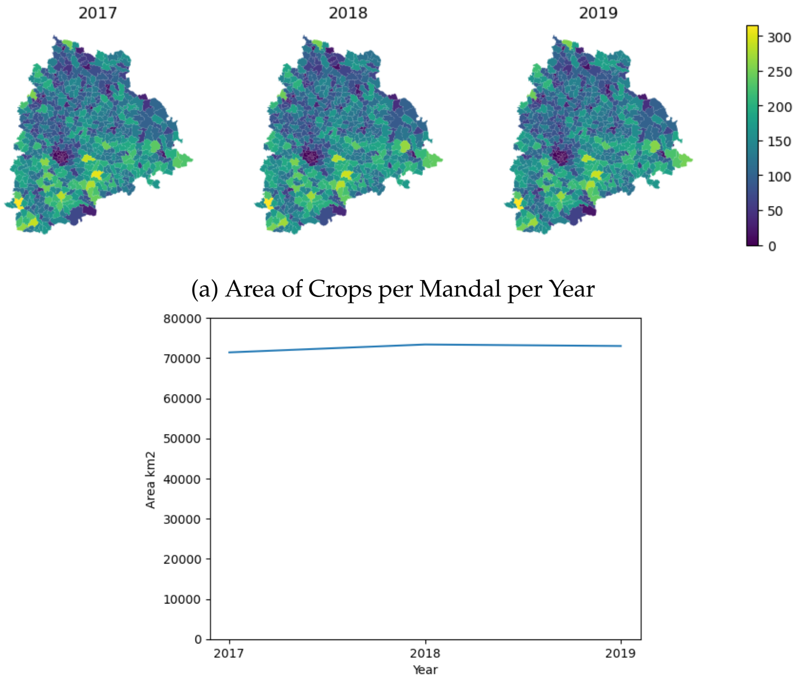


Figure 4: This Figure visualizes the Land Cover of the year 2018.

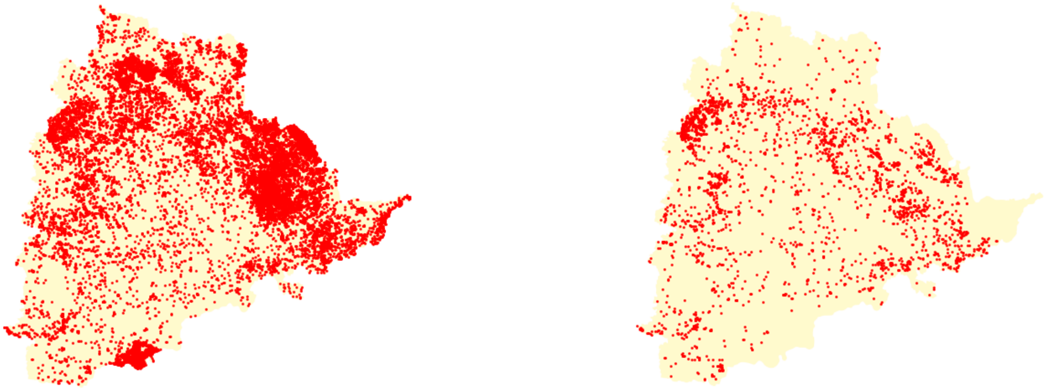
Using the approach described in Section 4.2, we have classified 2956 fires as agricultural fires out of 8794 between September 2016 and August 2019. Recall that set I_{agr} denotes the set of agricultural fires, thus $|I_{agr}| = 2956$. Figure 6 shows the classification results. When comparing Figure 6a with Figure 6b, we see that numerous fires, especially at the top of Telangana, are not classified as agricultural fires. This outcome seems reasonable as we observe in Figure 4 that these locations have land cover rangelands, trees, or water.



(a) Area of Crops per Mandal per Year

(b) Area of Crops per Year

Figure 5: These visualizations show the area of crops in km^2 over time. In Figure 5a, the area of crops is shown on mandal levels, and in Figure 5b, it is shown for the whole of Telangana. Based on these images, we can state that there are only minor differences between the number of crop areas over the years.

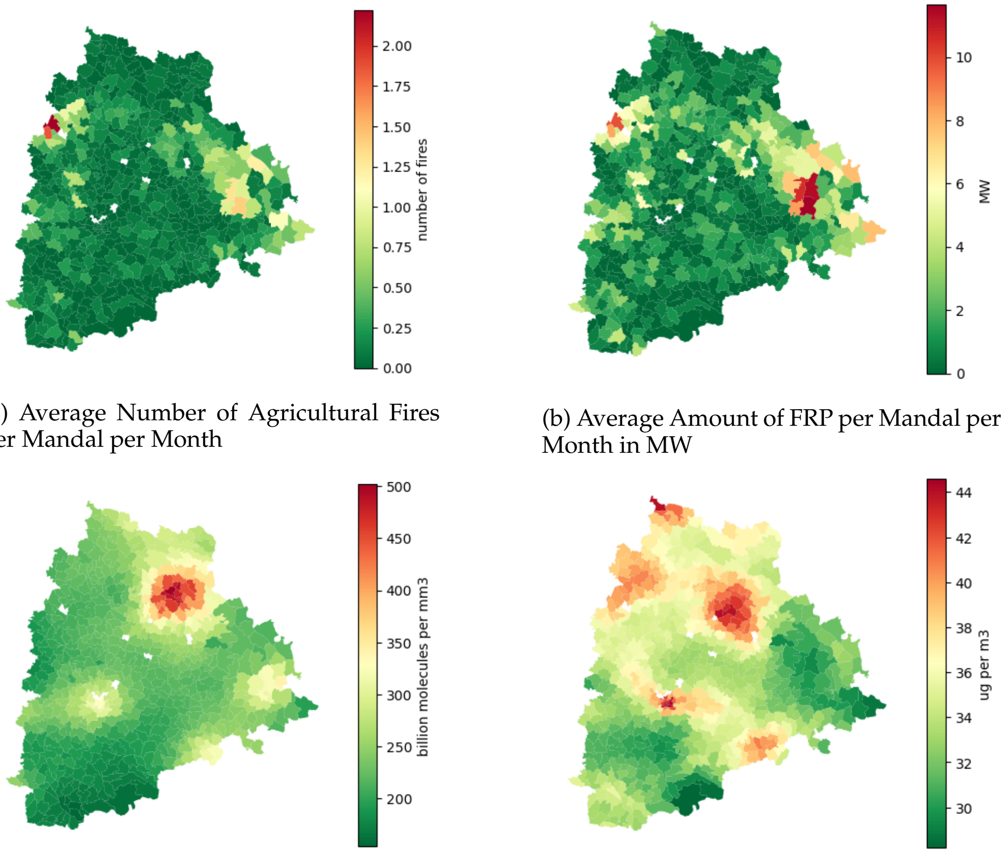


(a) Fires

(b) Agricultural Fires

Figure 6: This Figure shows the locations of (agricultural) fires (red dots) from September 2016 up to and including August 2019.

Quantification Measures CRB. Figure 7 shows the monthly quantification measures per mandal. In Figure 7a, one can see that most mandals have, on average, only a few to no agricultural fires. In districts Nizamabad and Mahabubabad (See Appendix A Figure 1), we observe a higher number of agricultural fires than in other parts of Telangana. Nearly the same holds for the FRP quantification measure. We again observe many values that are zero or close to zero. Besides, we detect that districts Nizamabad and Mahabubabad are also high in FRP emissions. However, the values within district Mahabubabad are higher for FRP than for the number of crop fires. In Figure 7c and 7d, the monthly average NO_2 and $PM_{2.5}$ emissions are shown per mandal. Significant differences exist between these measurements and the number of agricultural fires and FRP measurements. We observe very high values for NO_2, especially in district Peddapali. High $PM_{2.5}$ emissions are located around districts Peddapalli, Hyderabad, Nizamabad, Adilabad, and Nalgonda.



(a) Average Number of Agricultural Fires per Mandal per Month

(b) Average Amount of FRP per Mandal per Month in MW

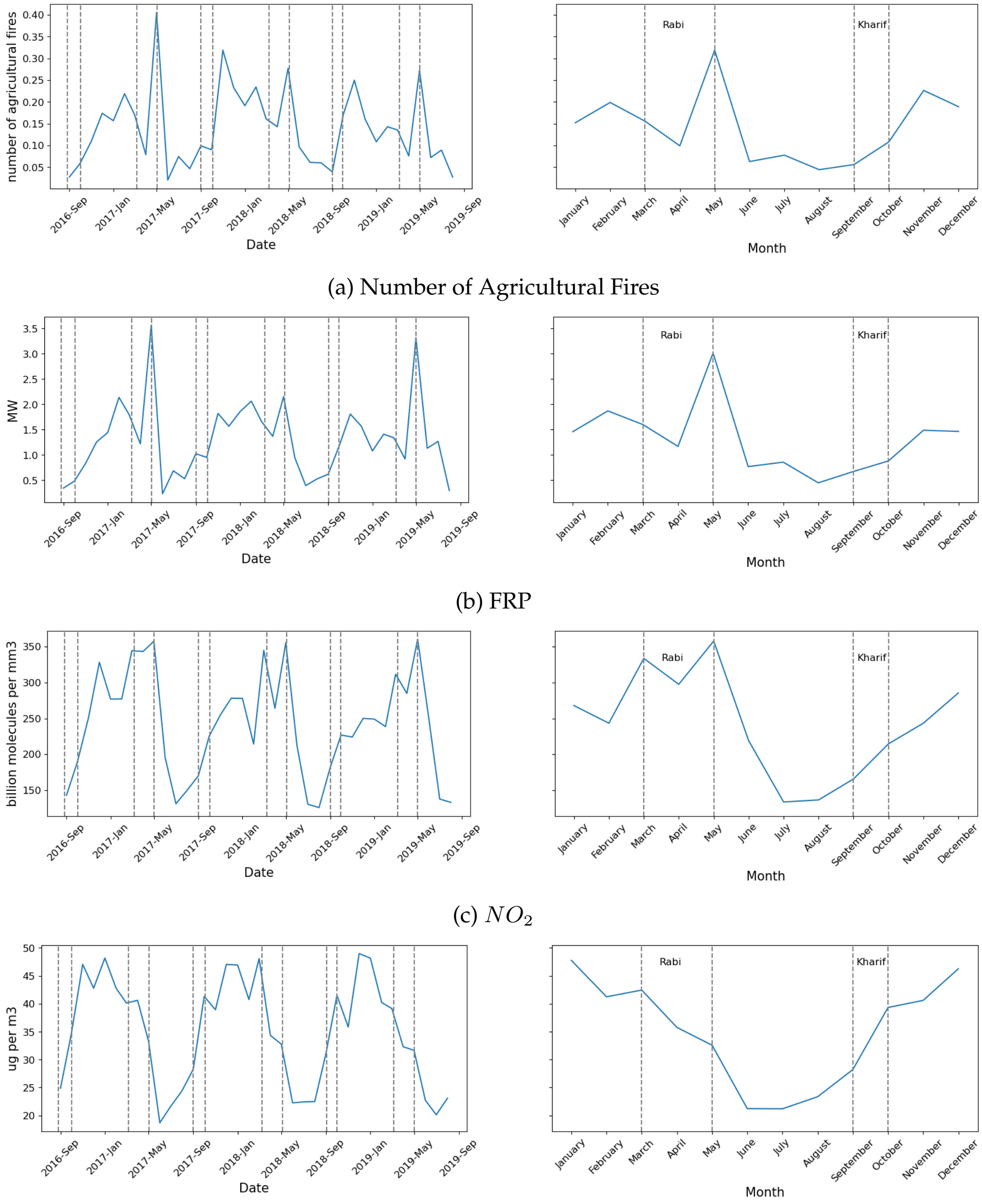
(c) Average NO_2 emission per Mandal per Month in billion molecules per mm^3

(d) Average $PM_{2.5}$ emissions per Mandal per Month in $\mu g/m^3$

Figure 7: These visualizations show the monthly averages per mandal of the different units used to quantify CRB between September 2016 to August 2019.

Furthermore, we have investigated the quantification methods over time. This is shown in Figure 8. The figures on the left show the quantity measures over the years, while the values are aggregated per month on the right side. For all quantity measures, we observe some seasonality on the left side. On the right-hand side, we see that the seasonality of the number of agricultural fires and FRP is mainly caused by the peak around May and slightly around October and November. This is in accordance with the Rabi and Kharif harvest months. Besides, for NO_2, we observe peaks around March-May and slightly December. Finally, for $PM_{2.5}$, we observe peaks mainly around January and December, but high values are also observed around March and October. Generally, we observe a significant decrease around June-July for all quantity measures. The figures show that NO_2 and $PM_{2.5}$ follow approximately the same pattern as the number of agricultural fires and the FRP values.

Appendix B Figure 1 provides additional exploration for the dependent variables in distribution visualizations.



(a) Number of Agricultural Fires

(b) FRP

(c) NO_2

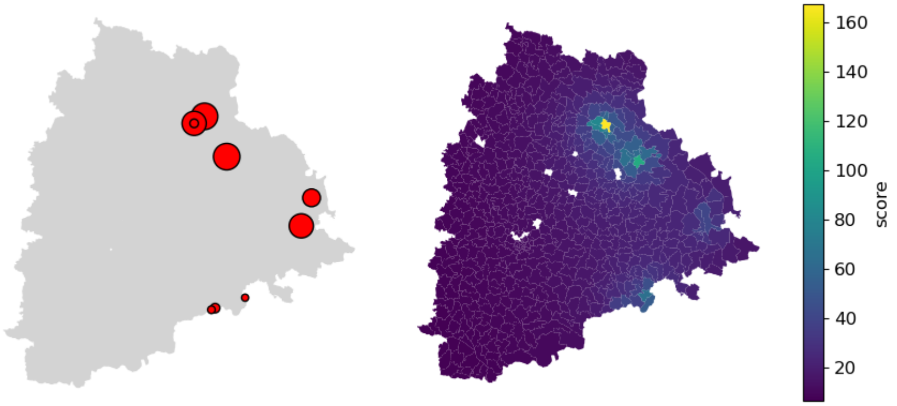
(d) Particulate Matter ($PM_{2.5}$)

Figure 8: This figure shows the different quantification methods over time averaged over mandals. We have visualized it monthly over the years to see a reoccurring seasonal pattern (left side) and averaged it monthly to see the differences per month (right side). Additionally, we have highlighted the harvest periods with grey dotted lines.

Independent Variables

In this section, we will focus on providing some data exploration information about the data of the most relevant independent variables further into our research. The independent variables we will focus on are: the socio-economic variable PowP_score and the environmental variables SSM_percentile_90, PRE_max, and ST_max. Later we will see the reasoning behind choosing these variables for extra discussion.

First, recall that PowP_score is calculated based on power plants and the distances between the center of the mandal and the power plants. In Figure 9a, the locations of the power plants are shown. The marker sizes indicate the capacity of a power plant in 2019. A larger circle indicates a higher capacity for that power plant. In Figure 9b, we observe the average PowP_score over the three years. Note that some of the power plants did not yet exist in 2016 or 2017, causing the average PowP_score to be lower than expected based on Figure 9a. Generally speaking, the average power plant scores are highest in the district Peddapalli and Mancherial (See Appendix A Figure 1). Most mandals have a PowP_score that is equal to or close to zero.

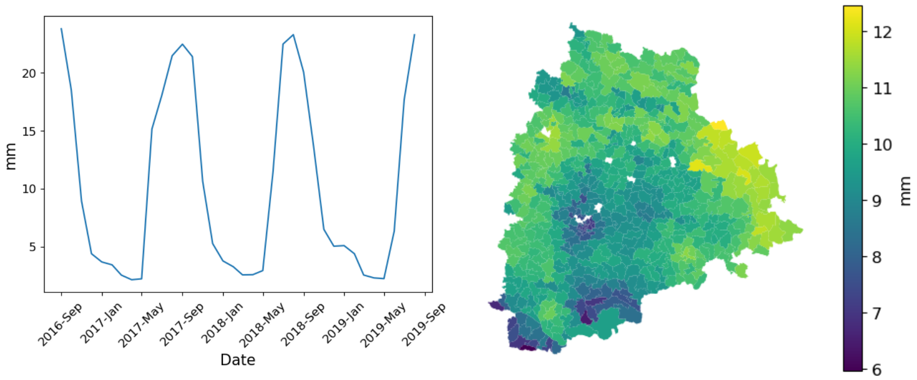


(a) Here, the location of the power plants in 2019 in Telangana are shown. The marker sizes indicate the capacity of the power plant in 2019.

(b) This figure shows the PowP_score of each mandal. This score considers the power plants' capacities and the distance between the power plants and a mandal.

Figure 9: PowP_score

Secondly, we will discuss the SSM_percentile_90. This variable shows the 90-percentile soil moisture of each mandal at a given time. In Figure 10b, we observe the largest average values in the North, East, and West of Telangana, especially in the North-East in districts Mulugu and Jayashankar Bhapalpally. In the South of Telangana, smaller values are perceived. This is particularly in districts Nagarkurnool and Jogulamba Gadwal. In Hyderabad, we also notice slightly lower SSM_percentile_90 values. In Figure 10a, the average SSM_percentile_90 is shown over time. This figure shows that each year during May up to and including November, the SSM_percentile_90 values significantly increase. Besides, it shows a minuscule increase in January.

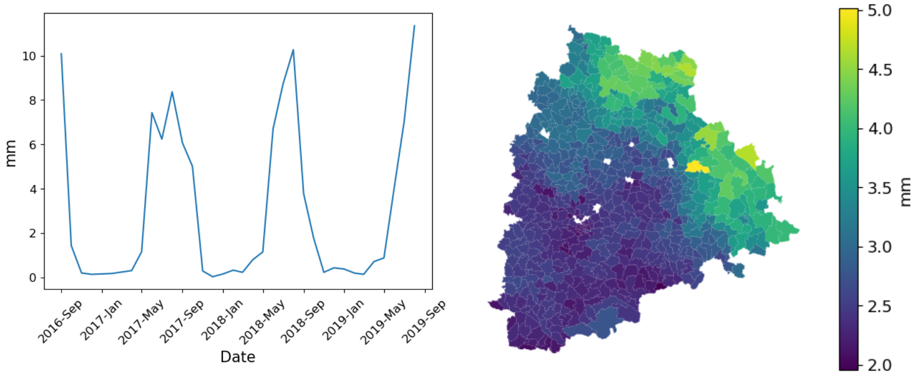


(a) Averaged per Month Observation

(b) Averaged per Mandal

Figure 10: SSM_percentile_90

Thirdly, we will talk through the variable PRE_max. This variable shows the maximum precipitation value within a mandal at a certain time. In Figure 11b, we discover higher levels of the average PRE_max values per mandal in the North and East of Telangana. These rises are primarily in Kamurambheen Asifabad, Jayashankar Bhupalpally, and Mulugu. In the South and Middle of Telangana, smaller levels of average PRE_max values are observed. When investigating the average PRE_max over time in Figure 11b, we notice that the rises and declines are very similar to those of SSM_percentile_90. Thus, high values between May and November. In 2017 a slightly different pattern in the values was detected. Next, 2017 has slightly lower values than the other years.

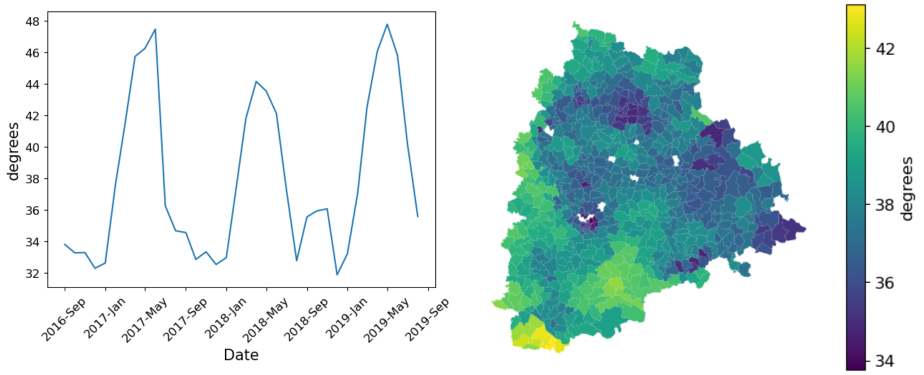


(a) Averaged per Month Observation

(b) Averaged per Mandal

Figure 11: PRE_max

Finally, we will explore the variable ST_max, which shows the maximum soil temperature in a mandal at a certain period in time. Figure 12b depicts that greater maximum temperatures are perceived in the south of Telangana, district Jogulamba Gadwal. The lowest temperatures are observed in Hyderabad, Bhadradri Kothagudem, Jayashankar Bhupalpally, Suryapet, and Jagtial. When analyzing the average ST_max over time in Figure 12a, we conclude that higher temperatures are observed during January up to and including July. Around October and November, minuscule rises are visualized. Especially in the year 2018, this rise is considerably present. Besides, the main peak in 2018 is slightly lower than in 2017 and 2019.



(a) Averaged per Month Observation

(b) Averaged per Mandal

Figure 12: ST_max

The complete overview of all independent variables data exploration is shown in Appendix C Figure 1 up and until including Figure 6.

5. Methodology

This section discusses the methods we applied in our research. Section 5.1 covers the data powered positive deviance analysis, and Section 5.2 the random forest prediction model.

5.1 Data Powered Positive Deviance Analysis

In the data powered positive deviance analysis, abbreviated DPPD, we will identify positive and negative deviant mandals. In this context, positive deviant mandals are mandals that reduce CRB over time, while negative deviant mandals are mandals that increase CRB over time. Both are relevant to analyze as we can gather insights from both parties and better understand each.

Time Series Decomposition. We will use time series decomposition to identify positive and negative deviant mandals. Time series decomposition is a method that reduces time series data into the following three main components: trend or level (g), seasonality (s), and the remainder (r) (Cleveland et al. (1990). The trend shows us the general movement over time, indicating whether it is increasing, decreasing, or constant. We are interested in this because it indicates the change in CRB over time while accounting for the seasonal differences. The seasonal component shows the changes at consistent frequencies. In this research, the seasonal differences are around the end of each harvesting period, October-September and March-May (see Table 1 and Figure 8). These consistent changes occur because farmers burn their crop residue at the end of each harvesting period, and in the rest of the year, only a few to zero agricultural fires are observed. Finally, the remainder is a random fluctuation in the time series. We have chosen to apply time series decomposition because this method analyzes changes over time and considers a seasonality component.

Before applying a time series decomposition model to gather the general trend of the data, we need to remove the consistent seasonal changes. There are two different types of seasonality decomposition methods that can be selected. We can choose between 1) additive seasonality decomposition and 2) multiplicative seasonality decomposition. Additive decomposition is suitable when the time series data is the sum of the three components (g, s, and r), showing that the seasonal variation is constant over time. This model choice is most suitable when the time series has roughly the same variability throughout the series. In contrast, multiplicative decomposition is suitable when the time series data is the product of the three components (Cleveland et al. (1990)). In the multiplicative time series model, the seasonal fluctuations increase or decrease proportionally with the increase and decrease in the level of the series. Therefore, the multiplicative model is suitable when the magnitude of the seasonal pattern increases as the data values increase or the other way around. To better understand the difference between the two methods, we have shown an example of both methods in Figure 13.

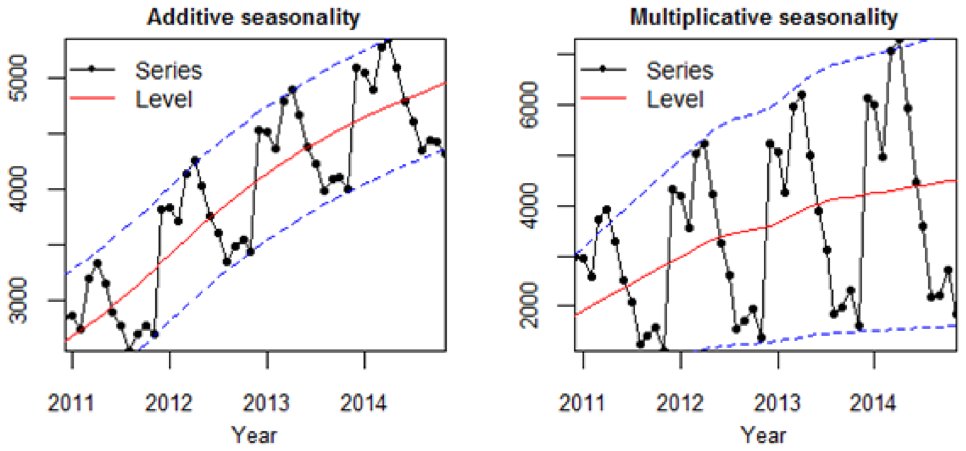


Figure 13: This figure shows an example of additive and multiplicative seasonality. We see the seasonal component remaining roughly the same variation size for the additive model on the left side. In contrast, for the multiplicative model on the right side, the seasonal component's variation increases proportionally over time (Kourentzes (2014)).

When choosing between one of the two decomposition methods, we can look at the data exploration visualizations. In Figure 8 on the left side, the quantity measures, aggregated per mandal, over time are depicted. We conclude from this picture that additive is suitable in this example because there are no proportional increases or decreases in the seasonal component over time. Thus, the time series model formula looks as follows in our context:

$$y_{m,t_y} = g_{m,t_y} + s_{m,t_y} + r_{m,t_y} \qquad \forall t_y \in T_y \qquad \forall y \in Y \qquad \forall m \in M \tag{29}$$

In Equation (29), y_{m,t_y} denotes the time series data observation, representing the amount of CRB. Then g_{m,t_y} is the trend/level component, s_{m,t_y} is the seasonal component, and r_{m,t_y} is the remainder. Recall that m denotes the mandal of interest with $m \in M$ and set T_y denotes the months within a year $y \in Y$. Previous research points out three main time series decomposition models, namely the classical approach using Two-Sided Moving Averages (Hyndman (2010)), X-11 Decomposition (Cleveland and Tiao (1976)), and Seasonal Trend Decomposition using LOESS (Cleveland et al. (1990)). We will only discuss the classical approach and the Seasonal Trend Decomposition in the next part for two reasons. First, Seasonal Trend Decomposition uses the classical approach, and previous research argues that Seasonal Trend Decomposition is currently dominating other time series decomposition methods. Additionally, X-11 Decomposition has not been used frequently recently (Cleveland et al. (1990)) (Plummer (2020))

Classical Approach using Two-Sided Moving Averages (MA). The most well-known decomposition method is the classical approach (Hyndman (2010)). The classical system uses two-sided moving averages (MA), in which the averages of several sequential time series values are taken. MA helps smooth a time series in order to estimate the underlying trend. The idea is that observations close in time are also likely comparable in value. So averaging among those values provides an accurate estimate of the trend while smoothing the data. Let $T = \bigcup_{y \in Y} T_y$ and $t \in T$ indicate all months available in the data over the years. Then the formula of MA looks as follows:

$$\hat{g}_{m,t} = \frac{1}{2 \cdot k + 1} \sum_{j=-k}^{k} y_{m,t+j} \qquad t = k+1, k+2, ..., |T| - k \qquad \forall m \in M \tag{30}$$

Note that in Equation 30 y_m represents the original time series observation value of mandal m at month t and $\hat{g}_{m,t}$ denotes the predicted trend value of mandal m at month t by MA. The number of sequential values before and after an observation at time t is denoted by k. For example, if $k = 2$, then we average among $t-2$, $t-1$, t, $t+1$, and $t+2$ ($2 \cdot k + 1 = 5$ values) to obtain an estimate for the trend at time t. Usually, k is determined based on the number of cycles available in the data. Note that we refer to cycle as the periodic variation in the time series data. In this context, the cycle length is a year because CRB reoccurs yearly within two distinct harvest seasons. Despite MA being very intuitive, it has a few disadvantages. It uses two-sided moving averages to estimate the trend, causing the first k observations and the last k observations to be absent. This leads to a transformed time series with predicted trend values $\hat{g}_{m,k+1}, ..., \hat{g}_{m,|T|-k}$. Next, previous research argues that the trend line usually over-smooths the data, causing a significant remainder component (Plummer (2020).

Seasonal-Trend Decompositions using LOESS (STL). Next to MA, there also exists Seasonal-Trend Decomposition using LOESS (STL). STL is a robust time series decomposition used for additive time series decomposition. The complete explanation of STL is extensive; therefore, we will only briefly cover the procedure. For a more extensive mathematical explanation of STL, we will refer to the paper 'STL: A Seasonal-Trend Decomposition Procedure Based on Loess' Section 5 (Cleveland et al. (1990). LOESS is a smoothing function used a few times in the procedure of STL. MA takes the average of sequential observations ($y_{m,t-k}, \ldots, y_{m,t+k}$) with equal weight. In contrast, LOESS considers an additional weight based on the distance between time series observations. Here the distance is defined as the absolute difference between two time series observations at different points. This absolute difference is defined as follows:

$$|y_{m,t} - y_{m,t+j}| \tag{31}$$

, where $j \in \mathbb{Z}$ ($j \neq 0$) denotes the number of months before ($j < 0$) or after ($j > 0$) time series observation t.

If the value of a time series observation ($y_{m,t+j}$) is similar to the observation, we aim to predict the trend for ($y_{m,t}$), the value of Equation 31 will be small. Then, LOESS applies an additional weight to observation $y_{m,t+j}$ in the calculation for $\hat{g}_{m,t}$. This results in similar observations having a stronger effect on the predicted trend values, causing the line to ease the smoothing of the trend values (in comparison with moving averages). In STL, $\hat{g}_{m,t}$ will be updated by applying LOESS and MA. Additionally, it provides the opportunity for observations with a substantial remainder component ($r_{m,t}$) to diminish their impact on $\hat{g}_{m,t}$. Hence, the method is considered robust (Cleveland et al. (1990). STL does not limit itself to only applying MA and therefore, there are no missing values in the predicted trend values $\hat{g}_{m,t}$. STL uses the number of sequential values q and the number of periods within one cycle p as input. Recall that a cycle is denoted as the periodic variation. In this research, a cycle has a duration of a year; hence the number of periods is the number of observations within the cycle. q is defined as the smoothing parameter because considering more values will significantly smooth the trend. Previous research points out that q has to be an odd number higher than five; otherwise, nearly no smoothing is applied (Cleveland et al. (1990)). The user is free to select a q based on the nature of the research.

To conclude, by applying STL, we gain a few advantages compared to other research methods such as MA. The primary advantage of STL is its robustness, indicating no sensitivity to outliers resulting in an accurate transformed series. Next, the user can control the seasonal component's rate of change (q), indicating the smoothness level of the outcoming trend. Furthermore, it does not lead to missing values in the transformed time series data (Theodosiou (2011)). Consequently, we selected the STL for the trend decomposition of the time series.

Deviant Scores. After decomposing the time series using STL, the predicted trend values, $\hat{g}_{m,t}$ for $t = 1, \ldots, |T|$, for each mandal $m \in M$, are acquired. Our primary focus is the general direction of the line between these values over time $t = 1, \cdots, |T|$ for each mandal m. An increasing predicted trend values line initiates an increasing trend over time in a specific mandal. To compare the mandals, it is necessary to quantify these increases and decreases. To achieve this, we will fit a linear line to the predicted trend values, $\hat{g}_{m,t}$, using simple linear regression (Zou et al. (2003)). We refer to the simple linear regression predicted line trend values as $d_{m,t}$. Note that $d_{m,t}$ is thus a linear approximation of $\hat{g}_{m,t}$. Then, we denote the slope of the line by a_m and the intercept by b_m. The slope measures the line's inclination relative to the period, and the intercept measures the start value of the trend. Consequently, we are mainly concerned with the slope ($d_{m,t}$) because it shows the changes of a mandal m per period t. We further refer to $d_{m,t}$ as the deviant score in this research. The linear regression formula for mandal m is shown in Equation 32.

$$d_{m,t} = b_m + a_m \cdot t \qquad \forall m \in M \tag{32}$$

Note that $a_m > \epsilon$ indicates an increasing trend, $a_m < -\epsilon$ a decreasing trend and $-\epsilon \leq a_m \leq \epsilon$ shows an approximately constant trend. ϵ is defined as the threshold of considering a mandal a positive or negative deviant. This threshold depends on the nature of the research. For example, if $a_m = 0.00001$ when using the number of agricultural fires as a quantifier, it means that per month t, the number of agricultural fires increases with 0.00001. In reality, this rise might be considered negligible, causing mandal m to be an insignificant positive deviant. Be aware that this should always be assessed in combination with expert knowledge. Finally, remark that the simple linear regression models for each mandal might not be suitable for forecasting, only for quantifying the past movement of the trend. It might not be ideal for forecasting because we are not interested in obtaining the best fit to the trend line, possibly a higher degree polynomial, but a simple linear line such that we derive one score for each mandal.

5.2 Random Forest Model

We will create a prediction model focusing on interpretability to understand CRB better and what factors affect it. We have seen in Section 4.2 that we use four ways to quantify CRB, of which two quantify CRB directly. Namely, by counting the number of agricultural fires and averaging over the FRP of the agricultural fires. Remember that a disadvantage of this method is that it is hard to model with known explainable models because the values per month (t) per mandal (m) contain many zeros (see Figure 7a and 7b). Numerous zeros in the prediction values cause the model to predict values relatively close to zero, obtaining low errors but having insufficient explainability power. To resolve this issue, we consider the two indirect quantification methods, NO_2 and $PM_{2.5}$, for the model prediction goal. Unfortunately, this makes it more challenging to measure the causes of CRB. We can exploit modeling CRB using direct quantification methods if additional data is available.

In this research, we will apply a random forest model because it has consistently low generalization errors reported in (similar) previous studies. It is relatively robust to outliers and noise, gives practical internal estimators of error such as strength, correlation, and variable importance, and is a simple and easy-to-interpret method (Breiman (2001)).

This section will discuss the bias and variance tradeoff within decision trees and random forest regression models. It involves evaluation criteria, cross-validation, independent variable selection, hyperparameter turning, variable importance, and the SHAP explainer.

5.2.1 Bias and Variance Tradeoff. The bias and variance tradeoff is a tradeoff within machine learning models that must be considered, such that models produce accurate results (Mehta (2019)). Assume we have independent variables x and dependent variable y. The true relationship between these values is described as follows:

$$y = f(x) + \epsilon \tag{33}$$

, where ϵ is represented as some random noise with $E(\epsilon) = 0$ and $Var(\epsilon) = E(\epsilon^2) - (E(\epsilon))^2 = E(\epsilon^2) = \sigma_\epsilon^2$. Note that ϵ is the irreducible error and the bias and variance are reducible errors. When designing the prediction model, we attempt to find a function $\hat{f}$ that predicts y while finding the true relationship f. Hence, we aim to ensure that $\hat{f}(x)$ is as close as possible to y.

A model is said to be biased if it systematically produces higher or lower predictions than the actual values, also known as underfitting. The bias of a prediction model is defined as follows:

$$bias(\hat{f}(x)) = E(\hat{f}(x)) - f(x) \tag{34}$$

On the other hand, a model is claimed to have high variance if its performance decreases drastically if tested on different data, also known as overfitting. The variance of a prediction model is defined as follows:

$$var(\hat{f}(x)) = E((\hat{f}(x) - \hat{f}(x))^2) \quad \text{(thus } var(\hat{f}(x)) >= 0) \tag{35}$$

The bias captures the error of the model and the variance in the model's generalizability. Ideally, a low bias and low variance are desired. Generally, a more complex model tends to have higher variance but lower bias. In comparison, an oversimplified model tends to have lower variance but will likely have more bias. This tradeoff is visualized in Figure 14.

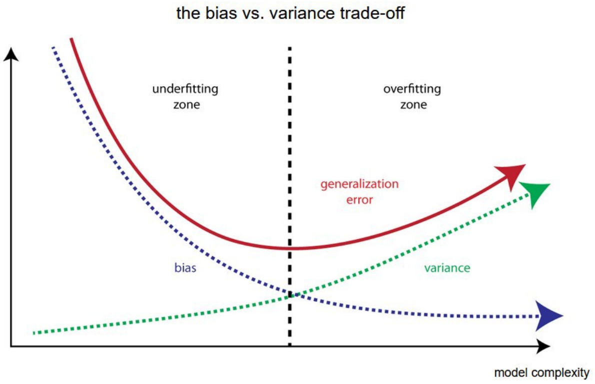


Figure 14: This figure shows the bias and variance tradeoff. Ideally, a low bias and variance are desired. Therefore, we want to minimize the sum of both of them (Mottaghinejad (2021)).

Accordingly, when considering the bias-variance tradeoff, the goal is to choose a model that minimizes both. Usually, models will minimize the mean squared error (MSE). We will show that minimizing this error leads to a minimization of the bias and the variance:

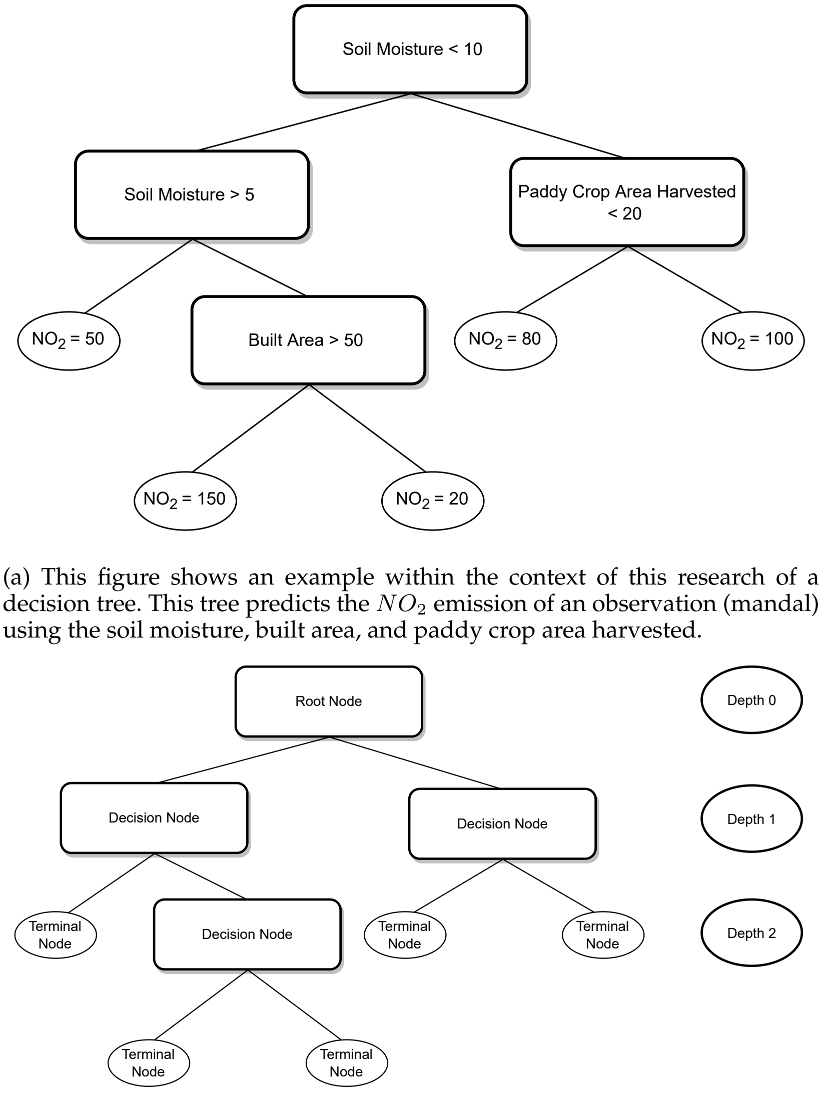
$$\begin{aligned}
E[(y-\hat{f}(x))^2] &= E[(f(x)+\epsilon-\hat{f}(x))^2] \qquad \text{as, } y=f(x)+\epsilon \\
&= E[(f(x)-\hat{f}(x))^2]+E[\epsilon^2]+2E[(f(x)-\hat{f}(x))]E[\epsilon] \\
&= E[(f(x)-\hat{f}(x))^2]+\sigma_\epsilon^2 \qquad \text{as, } E[\epsilon^2]=\sigma_\epsilon^2,\, E[\epsilon]=0 \\
&= E[(f(x)-E[\hat{f}(x)]-(\hat{f}(x)-E[\hat{f}(x)]))^2]+\sigma_\epsilon^2 \quad \text{as, } E[\hat{f}(x)]-E[\hat{f}(x)]=0 \\
&= E[(E[\hat{f}(x)]-f(x))^2]+E[(E(\hat{f}(x)]-\hat{f}(x))^2] \\
&- 2E[(f(x)-E[\hat{f}(x)])(\hat{f}(x)-E[\hat{f}(x)])]+\sigma_\epsilon^2 \\
&= E[\hat{f}(x))-f(x)]^2+E[\hat{f}(x))-\hat{f}(x)]^2 \\
&- 2((f(x)-E[\hat{f}(x)])(E[\hat{f}(x)]-E[\hat{f}(x)]))+\sigma_\epsilon^2 \\
&= bias(\hat{f}(x))^2+var(\hat{f}(x))-0+\sigma_\epsilon^2 \\
&= bias(\hat{f}(x))^2+var(\hat{f}(x))+\sigma_\epsilon^2
\end{aligned} \tag{36}$$

, where $E[(y-\hat{f}(x))^2]$ is defined as the MSE. Additionally, the bias is squared because it can contain negative values in contrast to the variance.

Hence, if we aim to minimize the MSE in a model, we both minimize the bias and the variance. This results in a model that is accurate and generalizable.

5.2.2 Decision Tree. A Decision Tree or Classification And Regression Tree (CART) is a predictive model that uses branching to predict the dependent variable outcome given the independent variable(s). It is a recursive algorithm that partitions the training dataset into subsets, minimizing the MSE in each subset (regression trees). A classification tree is used for categorical dependent variables, and a regression tree is used for numerical dependent variables.

Figure 15a shows an example of a regression decision tree. This tree predicts the number of NO_2 emissions of an observation (mandal) given information at a certain point in time regarding soil moisture, built area, and the number of paddy crop areas harvested. In Figure 15b, the terminology of the nodes is displayed. We will introduce some of this terminology briefly. First, leaf nodes or terminal nodes are the nodes that do not have any successors. In addition, a decision node is a node that contains a condition of which the root node is a special case. The root node is the first node, which is $SoilMoisture < 10$ in Figure 15a. If a condition on a decision node is true, we move in the left direction, and if it is false, we move in the right direction. Besides, the node $BuiltArea > 50$ is considered the parent node of the two child nodes $NO_2 = 150$ and $NO_2 = 20$. The depth of a decision node is the length of the path from the root to the decision node. In addition, the depth of a tree is defined as the maximum depth among all decision nodes in the tree. Accordingly, in the example, the depth is 2. To give an idea of how to read a decision tree, we will shortly go over an observation example of the tree in Figure 15a. Assume a mandal with $SoilMoisture = 3$, $BuiltArea = 80$, and $PaddyCropAreaHarvested = 10$ in March. Starting at the root node, the observation has less than 10 mm of soil moisture. As a consequence, we move to the left decision node. In the following node, we perceive the condition as false because the observation has less than 5 mm soil moisture. Finally, in the last decision node, we move to the left because we observe that the observation has more than 50 km^2 built area. This results in an NO_2 emission of 150 billion molecules per mm^3.



(a) This figure shows an example within the context of this research of a decision tree. This tree predicts the NO_2 emission of an observation (mandal) using the soil moisture, built area, and paddy crop area harvested.

(b) This figure shows the terminology of a decision tree.

Figure 15: Figure 15a visualises a decision tree example and Figure 15b shows its terminology.

Now that we intuitively have an adequate feeling of the process behind a decision tree, we will look at a decision tree from a mathematical point of view. In Figure 16, the tree from Figure 15a is visualized with mathematical notation.

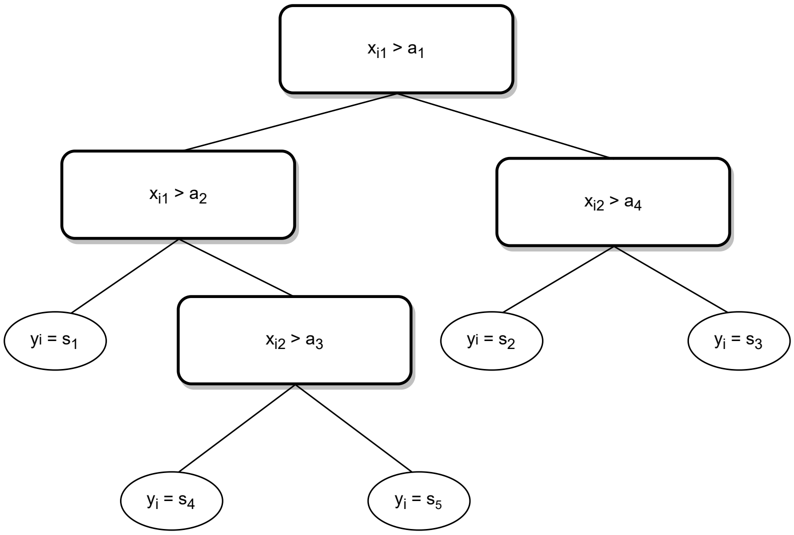


Figure 16: Here, the tree from Figure 15a is visualized with the mathematical formulation.

Let X denote the independent variable values and Y the dependent variable values described in Section 4.2. P' is the subset of all independent variables in the final model ($P' \subseteq P$), with $j' \in P'$. Subsequently, we define the number of independent variables as $p = |P'|$ and the total number of observations as n. Therefore, X is an n x p matrix and Y is an n x 1 matrix. Consequently, $x_{i,j'}$ indicates the value of the independent variable j' of observation i. The data, D, is denoted $D = \{X_i, y_i | i = 1, \ldots, n\}$, where $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$. Thus X and Y look as follows:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & \ddots & \cdots & x_{2,p} \\ \vdots & \cdots & \ddots & \vdots \\ x_{n,1} & \cdots & \cdots & x_{n,p} \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{37}$$

Finally, we define $\hat{y}_i$ to be the prediction of the regression tree model given X_i.

CART consists of three main components: a rule for assigning a value to each terminal node, the splitting rule, and the termination criterion (Fernando Rainho Alvest Torgo (1999)). First, we will discuss the rule for assigning a value to each terminal node, l, with l' the total number of terminal nodes. Define D_l as the set of observations of D within terminal node l and $n_l = |D_l|$. Using this mathematical notation, the prediction value for a terminal node, s_l, is calculated as follows:

$$s_l = \frac{1}{n_l} \sum_{i \in D_l} y_i \tag{38}$$

Hence $\hat{y}_i \in \{s_l | l = 1, \dots l'\}$. This means that if observation i ends at terminal node $l = 2$, $\hat{y}_i = s_2$. In Equation 38, we notice that s_l is the mean of the observations within each terminal node. This value is chosen because it minimizes the mean squared error. Recall that we previously defined the MSE as $E[(y - \hat{f}(x))^2]$, with y the dependent variable, x the independent variable(s), and $\hat{f}$ the prediction function. Considering the newly introduced notation, this is equal to:

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{39}$$

To prove that minimizing Equation 39 leads to Equation 38, we assume Y is a continuous random variable with probability density function $f(y)$. First, we will rewrite Equation 39:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = E[(Y - s_l)^2] \quad &\text{, as } E[Y] = \frac{1}{n} \sum_{i=1}^{n} y_i \text{ , and } \hat{y}_i \in \{s_l | l = 1, \dots l'\} \\
= \int_{-\infty}^{\infty} (y - s_l)^2 f(y) dy \quad &\text{, as } E[Y] = \int_{-\infty}^{\infty} y f(y) dy \\
= \int_{-\infty}^{\infty} (y^2 - 2y \cdot y_{s_l} + s_l^2) f(y) dy & \\
= \int_{-\infty}^{\infty} y^2 f(y) d(y) - 2s_l \int_{-\infty}^{\infty} y f(y) dy + s_l^2 \quad &\text{, as } \int_{-\infty}^{\infty} y f(y) = 1
\end{aligned} \tag{40}$$

Now we can minimize Equation 40 with respect to s_l:

$$\begin{aligned}
0 &= \frac{\partial}{\partial s_l} E[(Y - s_l)^2] \\
0 &= \frac{\partial}{\partial s_l} \int_{-\infty}^{\infty} y^2 f(y) d(y) - 2 s_l \int_{-\infty}^{\infty} y f(y) dy + s_l^2 \\
0 &= -2 \int_{-\infty}^{\infty} y f(y) dy + 2 s_l \\
s_l &= \int_{-\infty}^{\infty} y f(y) dy \\
s_l &= \frac{1}{n_l} \sum_{i \in D_l} y_i \qquad \text{, as} \int_{-\infty}^{\infty} y f(y) dy = E[Y] = \frac{1}{n} \sum_{i=1}^{n} y_i
\end{aligned} \tag{41}$$

Secondly, we will examine the splitting rule. We refer to splitting as branching a node into multiple nodes. The goal of a splitting rule is to choose a split that maximizes the decrease in the tree error resulting from this division. We will restrict the discussion of the splitting rule to the case of trees in which decision nodes have at most two children. Let o be a node corresponding to a partition D_o. D_o is defined as the data observations that reach node o, and $n_o = |D_o|$. Remember that l is a terminal node, and o can be any node. Using this notation, we define the fitting error of a node o, with s_o defined in Equation 41, as follows:

$$Err(o) = \frac{1}{n_o} \sum_{D_o} (y_i - s_o)^2 \tag{42}$$

Let a be the split value of node o, o_L the left child of node o, and o_R the right child node of o. Then D_{o_L} is denoted as the set of data observations that meet the condition $x_{i,j'} > a$. Hence, $D_{o_L} = \{< X_i, y_i >\in D_o : X_i \to a\}$, with $n_{o_L} = |D_{o_L}|$. Additionally, $|D_{o_R}|$ is defined as the set of the data observations that meet the condition $x_{i,j'} < a$. Therefore, $D_{o_R} = \{< X_i, y_i >\in D_o : X_i \nrightarrow a\}$, with $n_{o_R} = |D_{o_R}|$. We define the error of a split a as the weighted average of the errors of the resulting sub-nodes:

$$Err(a, o) = \frac{n_{o_L}}{n_o} Err(o_L) + \frac{n_{o_R}}{n_o} Err(o_R) \tag{43}$$

The best split a^* at decision node o maximizes the following equation:

$$\Delta Err(a, o) = Err(o) - Err(a, o) \tag{44}$$

Finally, the termination rule denotes when to stop growing the tree. This termination rule is essential for the reliability of the tree. If the tree grows unlimited, it will perform perfectly on the training data. However, it will perform poorly on the test data. In this case, the model's generalization is unacceptable due to overfitting. Therefore, pruning the tree is necessary. Note that pruning is defined as preventing the tree from growing unlimited. Pruning the tree can be done in various ways. We use a combination of setting a maximum tree depth (max_depth), a minimum number of observations within a node before a split is created (min_samples_split), and a minimum number of observations in a terminal node (min_samples_leaf). Defining a max_depth prevents the tree from growing unlimitedly large, min_samples_split causes the number of decision nodes to decrease, and min_samples_leaf increases the number of observations within a terminal node. Hence, these three termination criteria cause the tree to be more simplistic. In Section 5.2.7, we will discuss how to determine the values of these parameters.

Growing the complete tree will be done in a recursive matter. Algorithm 1 describes the recursive procedure using $< X_i, y_i >, i = 1, \ldots, n$ as input. The best split value is chosen at each node according to the existing tree. Thus, it does not consider each possible tree, causing the algorithm to be greedy.

Algorithm 1 Recursive Partitioning Algorithm

```
if Termination criteria reached then
    Create a leaf node, $l$ and assign it a constant value $s_l$.
    Return leaf node.
else
    Find the best split value, $a^*$.
    Create a node $o$ with $a^*$ as split value.
    The left branch of node $o$ = RecursivePartitioning($< X_i, y_i >$: $X_i \to a$)
    The right branch of node $o$ = RecursivePartitioning($< X_i, y_i >$: $X_i \not\to a$)
    Return node $o$
end if
```

5.2.3 Random Forest Regression. Random forest is a non-parametric, supervised machine learning algorithm that can be applied for classification and regression purposes. It consists of a large number of individual decision trees that operate together. The random forest algorithm is briefly outlined in Algorithm 2 (Breiman (2001)). Note that in the formula in step 3 of Algorithm 2, B denotes the number of bootstrap samples (or the number of trees) in the random forest, and $\hat{y}_{i,b}$ the prediction for observation i of the tree b. We refer to bootstrapping as a statistical procedure using random and replacement sampling to simulate multiple samples. In Section 5.2.7, we will discuss how to determine the number of (bootstrap) samples. Additionally, note that B is defined as the number of trees in a random forest model. Finally, $\hat{y}_i$ defines the final prediction of the random forest for observation i.

Algorithm 2 Random Forest Algorithm

```
1: Draw $B$ (bootstrap) samples from the original data.
2: Grow a tree for each sample with $p$ number of independent variables. This causes a
   reduction in the correlation between the trees.
3: Aggregate the predictions of the $T$ trees and get the prediction values. For a regres-
   sion tree $\hat{y}_i \leftarrow \frac{1}{B}\sum_{b=1}^{B} \hat{y}_{i,b}$
```

Unlike single decision tree models (CART), a random forest is unlikely to overfit the data. This is because the prediction is based on the average forecast across low-correlated decision trees, causing a decrease in variance. We will prove that this holds. Assume we draw B bootstrap samples from the original data, constructing $T_1, T_2, \dots, T_B$ trees. The trees are identically distributed trees, however not independent. Therefore, there exists a positive correlation ρ_{T_b,T_j} if $b \in B$ and $j \in B$ for $b \neq j$ between these trees. We will show that the variance reduces if various trees $T_1, T_2, \dots, T_b$ are used in the random forest model.

We know that $Var(T_1) = Var(T_2) = \cdots = Var(T_b) = \sigma^2$. Besides, we assume $Cov(T_b, T_j) > 0$ if $b \in B$ and $j \in B$ for $b \neq j$. Then the following holds:

$$\begin{aligned}
Var(\frac{\sum_{b=1}^{B} T_b}{B}) &= \frac{1}{B^2} Var(\sum_{b=1}^{B} T_b) \\
&= \frac{1}{B^2}(\sum_{b=1}^{B} Var(T_b) + \sum_{j=1}^{B} \sum_{b=1, b\neq j}^{B} Cov(T_b, T_j)) \\
&= \frac{1}{B^2}(\sum_{b=1}^{B} Var(T_b) + \sum_{j=1}^{B} \sum_{b=1, b\neq j}^{B} \rho_{T_b,T_j} \sqrt{Var(T_b)Var(T_j)}) \\
&\leq \frac{1}{B^2}(\sum_{b=1}^{B} Var(T_b) + \sum_{j=1}^{B} \sum_{b=1, b\neq j}^{B} \sqrt{Var(T_b)Var(T_j)}) \qquad \text{as, } 0 < \rho_{T_b,T_j} \leq 1 \\
&\leq \frac{1}{B^2}(\sum_{b=1}^{B} Var(T_b) + \sum_{j=1}^{B} \sum_{b=1, b\neq j}^{B} \frac{Var(T_b) + Var(T_j)}{2}) \qquad \text{as, } \sqrt{ab} \leq \frac{a+b}{2} \\
&= \frac{1}{B^2}(\sum_{b=1}^{B} \sigma^2 + \sum_{j=1}^{B} \sum_{b=1, b\neq j}^{B} \frac{2\sigma^2}{2}) \\
&= \frac{1}{B^2}(B\sigma^2 + (B^2 - B)\sigma^2) \\
&= \sigma^2
\end{aligned} \tag{45}$$

Besides, we can show that the random forest model prediction remains unchanged. Let $E(T_1) = E(T_2) = \cdots = E(T_b) = \mu$. Then the following holds:

$$E(\frac{\sum_{b=1}^{B} T_b}{B}) = \frac{1}{B} E(\sum_{b=1}^{B} T_b) = \frac{1}{B} \sum_{b=1}^{B} E(T_b) = \frac{1}{B} \sum_{b=1}^{B} \mu = \frac{1}{B} \cdot B \cdot \mu = \mu \tag{46}$$

To conclude, a random forest reduces the variance compared to a decision tree because $Var(\frac{\sum_{b=1}^{B} T_b}{B}) \leq \sigma^2$, while the random forest predictions remain unchanged because $E(\frac{\sum_{b=1}^{B} T_b}{B}) = \mu$.

5.2.4 Evaluation Metrics. In previous research (Masih (2019)) (Kumar et al. (2020)) (Kamińska (2019)) of random forests in time series applications, three evaluation metrics for regression-based random forests stand out, namely R-squared R^2, Root Mean Squared Error $RMSE$, and Mean Absolute Error MAE.

The R^2 shows how much of the variation of the dependent variable is explained by the independent variables in the model. If $R^2 < 0$, it means the model is predicting worse than the mean of the dependent variable, if $R^2 = 0$, none of the variations of the dependent variable is explained by the independent variables (model predicts the mean), and if $R^2 = 1$ all variance is explained by the independent variables (model predicts without error). The R^2 is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{47}$$

In formula 47 y_i refers to the actual value of observation i, $\bar{y}$ indicates the mean of all actual values, $\hat{y}_i$ denotes the predicted value of observation i, and $\bar{\hat{y}}$ is the mean of all predictions.

The $RMSE$ is a measurement of the spread of the residuals of a model. In which residuals are defined as $\hat{y}_i - y_i$. The MAE measures how far away, on average, without considering direction, the predicted value is from the actual value. Note that for $RMSE$, higher residuals are punished more significantly than MAE due to the square. The formulas of these evaluation metrics are described here:

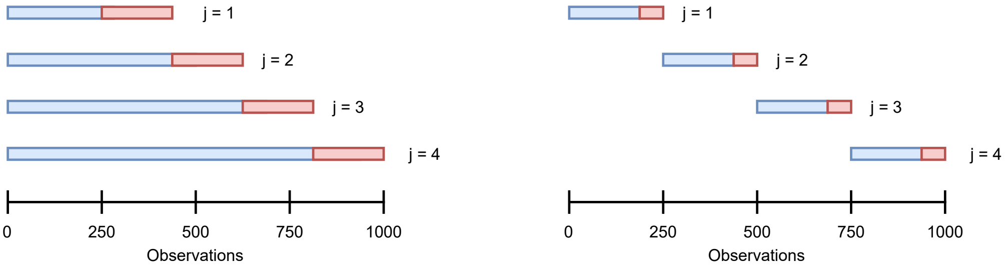
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{48}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{49}$$

Note that $RMSE$ and MAE are absolute evaluation metrics, meaning that one cannot compare different dependent variables model outcomes across one another. In contrast, this is possible for R^2.

In this research, we create a random forest model that predicts monthly values per mandal. Consequently, we can calculate the average evaluation metrics per mandal and month over the years. We can calculate these averages using the formulas in Equation 8 and Equation 7 for $q = R^2$, $q = MAE$, $q = RMSE$.

5.2.5 Cross-Validation in Time Series. A technique that is primarily used to evaluate the performance of a machine learning model on unseen data is called cross-validation. It usually leads to lower biases because the model is evaluated on distinct data. Cross-validation splits the data into k number (≥ 2) of subsets, at least one for training and one for evaluating the model. This method improves the model's generalizability because the model is less prone to overfitting as we test the model on unseen data. For time series data, cross-validation needs to be performed in a specific way such that data from the future is not used to forecast values in the past. Therefore, classical cross-validation techniques are unsuitable and can cause overfitting if applied. Fortunately, there are two methods of applying cross-validation for time series data: Rolling cross-validation and Blocked cross-validation (Bergmeir and Benítez (2012)). These methods are visualized in Figure 17. In Figure 17, the blue boxes denote the training data, and the red boxes denote the test data. At each fold, the model will be trained and tested on different dataset parts, and at the end, the errors will be averaged among all folds. To demonstrate this mathematically, let e_j be the error measurement in fold j then the average error is calculated as follows:



(a) Rolling Cross-Validation

(b) Blocking Cross-Validation

Figure 17: This figure shows the differences between rolling and blocked cross-validation. The number of folds is defined by $j = 1, j = 2 \ldots, j = k$, with $k = 4$. Note that blue boxes denote the training data, and red boxes the test data.

$$e = \frac{1}{k} \sum_{j=1}^{k} e_j \tag{50}$$

For rolling cross-validation (see Figure 17a), one starts with a small training dataset, predicting future values and calculating the accuracy of the test set. This training dataset is continuously increased in each fold. Unfortunately, this technique might lead to leakage from future data to the model, because the model can observe future patterns and will attempt to memorize them.

Therefore, blocking cross-validation is introduced; see Figure 17b. In blocked cross-validation, margins are added at two positions such that the model is prevented from memorizing patterns from one iteration to the next. In Algorithm 3, the calculation of the train and test observations within each fold is shown. The user can choose the number of folds, depending on the data characteristics. For example, if the dataset of interest includes three years of data, selecting $k = 3$ is not highly reliable because the data will be evaluated and trained on a similar part of the year in each fold. Be aware that selecting a higher fold number, in general, decreases the number of train and test observations in each fold. Another parameter the user can choose is the train/test ratio within each fold. Previous research justifies empirically that between 70-80% training data and 20-30% test data leads to the most accurate results (Gholamy et al. (2018)). Consequently, we have selected a train data size of 75% and a test data size of 25% in each fold. Unfortunately, using blocking cross-validation leads to a reduced number of data observations in each fold compared to rolling and regular cross-validation.

Algorithm 3 Blocking Time Series Cross-Validation

```
Define the number of $k$ folds
$train\_test\_split \leftarrow 0.75$
$kfold\_size \leftarrow \lfloor \frac{n}{k} \rfloor$
$indices \leftarrow [0, \ldots, n-1]$
$j \leftarrow 0$
while $j \leq (k-1)$ do
    $start \leftarrow j \cdot kfold\_size$
    $stop \leftarrow start + kfold\_size$
    $split \leftarrow \lfloor (train\_test\_split \cdot (stop - start)) \rfloor + start$
    $train_j \leftarrow indices[start : stop]$
    $test_j \leftarrow indices[split : stop]$
    $j \leftarrow j + 1$
end while
```

To conclude, we select blocking cross-validation over rolling cross-validation in this research as it prevents leakage from future data to the model. However, while applying this method, we must consider that the model in each fold is trained and evaluated on fewer data points than rolling cross-validation (and other regular cross-validation methods).

5.2.6 Independent Variable Selection. To obtain the best independent variable subset (P') for the model, we only need to select the ones with strong explainable power. Later, we will see that highly correlated variables will be given an overall reduced feature importance in a random forest compared to the same tree without correlated counterparts. In addition, random forests have a general preference for variables with high cardinality. Therefore, the outcomes from the model depend on the methods we choose to select the independent variables. In order to account for this, we will calculate the Spearman correlation coefficient for each independent variable combination and create variable subsets based on these correlation scores. The Spearman correlation coefficient measures the monotonic relationship between two variables. A monotonic relationship is described as follows: the value of one variable increases, and so does the other variable value, or the value of the variable increases and the other variable decreases (Schober (2018)). In contrast with a linear relationship, monotonic relationship values do not change at a constant rate. Spearman works with rank-ordered variables. Rank-ordered variables denote the ranks assigned to the variable in ascending order. Hence, if we have data points 5, 3, 9, and 1, the ranks of these points would be 3, 2, 4, 1. Note that x_i and x_j denote the values of independent variable i and j, with $i, j \in P$. Recall that P denotes the set of all independent variables. In addition, $R(x_i)$ denotes the ranking of independent variable x_i. Then the Spearman correlation coefficient, r_{x_i,x_j}, between variable x_i and x_j is calculated as follows (Zar (2005)):

$$rs_{x_i,x_j} = \frac{cov(R(x_i), R(x_j))}{\sigma_{R(x_i)}\sigma_{R(x_j)}} \qquad \forall i \in P \qquad \forall j \in P \tag{51}$$

, where Equation 51, $cov(R(x_i), R(x_j))$ is defined as the covariance of variable x_i and x_j. Note that if $i = j$, $rs_{x_i,x_j} = 1$.

Let $\bar{R(x_i)}$ denote the mean of the rank variables of x_i, then $cov(R(x_i), R(x_j)$ is calculated as follows:

$$cov(R(x_i), R(x_j)) = \frac{\sum_{k=1}^{n}(R(x_{i_k}) - \bar{R(x_i)})(R(x_{j_k}) - \bar{R(x_j)})}{n} \tag{52}$$

In Equation 51, $\sigma_{R(x_i)}$ denote the standard deviation of $R(x_i)$. This value is calculated as follows:

$$\sigma_{R(x_i)} = \sqrt{\frac{\sum_{k=1}^{n}(R(x_{i_k}) - \bar{R(x_{i_k})})^2}{n}} \tag{53}$$

In Equation 51 $rs_{x_i,x_j} = 0$ means no correlation, $rs_{x_i,x_j} = -1$ mean perfect negative correlation between ranks, and $rs_{x_i,x_j} = 1$ means perfect positive correlation between the ranks of the variables. It is challenging to set thresholds, e.g., weak, moderate, or strong relationship correlations. Most researchers agree that a coefficient of $rs_{x_i,x_j} < |0.1|$ indicates negligible correlation, $rs_{x_i,x_j} > |0.8|$ a strong relationship, and $rs_{x_i,x_j} > |0.9|$ a very strong relationship. However, rather than using oversimplified rules, the coefficients should be interpreted as related to the research context (Schober (2018)). As briefly discussed, we will create independent variable subsets of variables strongly correlated with one another using the Spearman correlation coefficients. Subsequently, we will train the random forest model on these strongly correlated independent variable subsets and calculate the variable importance on the test data. Finally, we will only select one independent variable from the strongly correlated subset. If this variable is of predictive power (e.g., high R^2 and high variable importance), it will be added to the model. The other variables within the subset will not be added to the model such that the variable importance of the independent variables is reliable. We will explain the variable importances more extensively in Section 5.2.8. We refer to P' as the subset of the final model's independent variables ($P' \subseteq P$). This subset contains all independent variables that are significant in the prediction of the dependent variable.

5.2.7 Tuning Hyperparameters. Hyperparameters are parameters in a machine learning model whose value controls the learning process. The primary hyperparameters of a random forest and their description and default values are shown in Table 9 (Pedregosa et al. (2011)) (Probst et al. (2019)). Hyperparameter tuning is used to customize the model to the dataset. If the hyperparameters are untuned, inaccurate results are probably to be achieved. This is because distinct datasets can vary significantly in, for example, size, the number of independent variables, variable types (categorical or numerical), and more. The hyperparameters depend on these specific dataset characteristics and must be tuned.

Often the general effects of hyperparameters are known. However, selecting the optimal hyperparameters can be challenging due to the interaction effects. We will discuss the general effects of each hyperparameter briefly.

Hyperparameter Tuning		
Hyperparameter	**Description**	**Default**
Bootstrap	Boolean indicating bootstrap observations are used	$TRUE$
n_estimators	Number of trees in the forest, or the number of (bootstrap) samples	500, 1000
max_features	Number of features in each tree	$\frac{p}{3}$, p [6]
min_samples_leaf	Minimum number of observations in terminal node	5
max_depth	Maximum number of levels in each decision tree	$None$
min_samples_split	Minimum number of observations required to split an internal node	2

Table 9: In this table, the hyperparameters and their default values for the random forest regression model are shown (Probst et al. (2019)), (Pedregosa et al. (2011)).

- **bootstrap:** First, we will examine the $bootstrap$ parameter. We have previously discussed that a random forest consists of multiple trees. These trees are created using bootstrap samples or the entire dataset. Recall that bootstrapping is defined as a statistical procedure using random sampling with replacement to simulate multiple samples. We select $TRUE$ if $bootstrap$ samples are used when building trees and $FALSE$ when the whole dataset is used when building each tree. Here, the bias and variance tradeoff needs to be considered because the first option results in lower variance but higher bias, and the second option results in lower bias but higher variance.

- **n_estimators:** Secondly, $n\_estimators$ represents the number of trees or the number of bootstrap samples in the random forest model. Here generally holds that more trees increase the model reliability. However, adding more trees is computationally expensive; beyond a certain point, the tradeoff might not be worth it.

- **max_features:** Additionally, $max\_features$ describes the number of features or independent variables from the final independent variable subset each split in a tree can select. Generally speaking, a higher number of $max\_features$ leads to higher correlated trees because more overlapping independent variables are present in each tree. On the other hand, an insufficient number of $max\_features$ leads to increased bias because we do not have enough explainability in the subset of selected independent variables.

[6] Note that the first value was supported by (Breiman (2001)) in 2001, and the second value was more recently justified empirically in (Geurts et al. (2006))

- **min_samples_leaf:** Besides, we should consider the $min\_samples\_leaf$ number. Recall that $min\_samples\_leaf$ is the minimum number of observations within one terminal node. Again, the bias and variance tradeoff is essential for this parameter. A lower $min\_samples\_leaf$ value means fewer observations in each terminal node, causing the model to be able to get more complex. On the other hand, a higher $min\_samples\_leaf$ can cause the model to be very simplistic. A tradeoff must be found so that the model does not overfit or underfit.
- **max_depth:** Another hyperparameter is the max_depth. Remember that max_depth denotes the number of splits a tree can make before a prediction is derived. If a higher max_depth value is chosen, the tree is allowed to grow more prominent, decreasing bias. However, the model is likely to overfit the training data. Overfitting is not desirable because we want the model to have low variance and thus be generalizable. Therefore, we once more must find a tradeoff between bias and variance.
- **min_samples_split:** Finally, remember that $min\_samples\_split$ tells us the minimum number of observations within a node before a split is calculated. The bias and variance tradeoff is also crucial in the last hyperparameter. This is because a lower $min\_samples\_split$ value will cause the model to be more complex. In general, this leads to more decision nodes within the trees. In contrast, a higher $min\_samples\_split$ value causes fewer decision nodes in the trees. Note that $min\_samples\_split \geq 2$, otherwise we cannot create a split between the observations.

Now we have a general idea of the effects of each hyperparameter. However, as previously mentioned, hyperparameters have interacting effects. Testing each hyperparameter setting combination is impossible because there are infinitely many combinations. Additionally, testing a large number of combinations is very computationally expensive. To solve this, we will use hyperparameter tuning to objectively search for different model hyperparameter values. For this, we will use random search and grid search.

In random search, we define a search space for each hyperparameter and randomly sample points in that domain. It uses random combinations of hyperparameters to find the best solution for the built model. Then we use the output of random search to have a more specific search space and evaluate every position in the grid using grid search (Brownlee (2020)). Grid search can be applied multiple times to search more precisely for optimal hyperparameters. Let $f : \mathbb{R}^h \rightarrow \mathbb{R}$ be the function that calculates the R^2 (with k-fold blocked cross-validation) using the actual values of the data y_i and the predicted values $\hat{y}_i$ with the hyperparameter input vector $x \in \mathbb{R}^n$, h the number of hyperparameters, and n_{iter} the number of iterations of the algorithm. The algorithm of random search is described in Algorithm 4. It is desirable to have n_{iter} as high as possible because more combinations of hyperparameters are tested. However, this is computationally expensive. Therefore, we must find the right balance between the number of combinations tested (which also depend on the search space size) and the computational costs. We define the number of all hyperparameter value combinations for grid search as n_{com}. The algorithm of Grid Search is described in Algorithm 5.

Algorithm 4 Random Search

```
Randomly sample $x \in \mathbb{R}^h$ without replacement from the search space.
$j \leftarrow 0$
while $j \leq (n_{iter} - 2)$ do
    Randomly sample $y \in \mathbb{R}^h$ without replacement from the search space.
    if $f(y) > f(x)$ then
        $x = y$
    end if
    $j \leftarrow j + 1$
    Return $x$ as the best hyperparameter value combination
end while
```

Algorithm 5 Grid Search

```
Run the first hyperparameter value combination called $x \in \mathbb{R}^h$ from the search space.
$j \leftarrow 0$
while $j \leq (n_{com} - 2)$ do
    Run the $(j + 2)$-th hyperparameter value combination called $y \in \mathbb{R}^h$ from the search space.
    if $f(y) > f(x)$ then
        $x = y$
    end if
    $j \leftarrow j + 1$
end while
```

Note that for a combination of random and grid search to reduce the computational costs efficiently, the total search space of grid search needs to be smaller than the total search space of random search. More specifically, the hyperparameter values tested in a random search will be of greater range, with larger step sizes than those in a grid search. The exact search spaces will be defined based on experimentation with the data. Be aware that these search methods do not result in the optimal outcome. This is because our search area is limited, and it is impossible to examine every combination.

5.2.8 Variable Importances. A random forest can provide additional information, such as permutation feature importance. The permutation feature importance measures the increase in the prediction error after permuting independent variable (or feature) values. A feature is unimportant if permuting its values does not influence the model error. Breiman introduced the idea behind the method in 2001 for random forests specifically (Breiman (2001)). Based on this idea, a generalized method was introduced. The algorithm is shown in Algorithm 6 (Molnar (2022)). Recall that X is the feature matrix (n x p) and Y (n x 1) is the target variable. Besides, recall that $j' \in P'$, with P' the set selected independent variables in the final model, $P' \subseteq P$, and $p = |P'|$. We run this algorithm on the validation set because it gives us an idea of the generalizability of the selected features. Necessary for the permutation feature importance is that the independent variables are not strongly correlated with one another. As previously discussed, strongly correlated features can lead to a lower importance value for both

Algorithm 6 Permutation Variable Importances

```
Estimate the original model error, $e_{orig}$ using $R^2$ for regression purposes, on the validation dataset.
while $j' \in P'$ do
    Generate a feature matrix $X_{perm}$ in which feature $j'$ is permuted in the data X (breaking the association between feature $j'$ and the true outcome)
    while $i = 1, \ldots, n$ do
        Estimate the model error, $e_{i,perm}$ on observation $i$ based on the predictions of the permuted feature matrix $X_{i,perm}$
        Calculate the permutation feature importance $FI_{j'} \leftarrow e_{orig} - \frac{1}{n}\sum_{i=1}^{n} e_{i,perm}$
        Calculate the permutation feature importance standard deviation $FISD_{j'} \leftarrow \sqrt{\frac{\sum_{i=1}^{n} |e_{i,perm} - \frac{1}{n}\sum_{i=1}^{n} e_{i,perm}|^2}{n}}$
    end while
end while
Sort the feature importances by descending $FI_j$
```

features, although they might be very relevant. One can handle this issue by clustering highly correlated features and only keeping at most one independent variable from each cluster. We have described the procedure in this in Section 5.2.6, Independent Variable Selection

5.2.9 Shapely Additive Explanations (SHAP). SHAP (SHapely Additive exPlanations) is a method to interpret individual predictions. It calculates the contribution of each feature to the prediction such that the impact of a feature on the model outcome is better understood. The idea behind SHAP is based on the Shapley values, a method from coalitional game theory. Shapley values tell us how to distribute payout fairly among players in a game. In this context, the game is the prediction value, and each player is a different feature or feature combination. Let f be the predictive model that needs to be explained. Recall that P' is the set of all selected independent variables, and S is a feature subset from P', $S \subseteq P'$. Then the SHAP values are defined as follows (Lundberg and Lee (2017)):

$$\phi_i = \sum_{S \subset P' \setminus \{i\}} \frac{|S|!(|P'| - |S| - 1)!}{|P'|!} (f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)) \tag{54}$$

More intuitively, the effect of a model $f_{S\cap\{i\}}$ is trained with feature i present, and another model f_s is trained with feature i absent. The comparison of the two feature sets is shown in the $(f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S))$ part of the equation. This is performed for every possible feature subset $S \subseteq P' \setminus \{i\}$. Then the Shapley values shown in Equation 54 are weighted averages of all possible differences. This section will discuss TreeSHAP, an alternative to KernelSHAP specifically designed for tree-based machine learning models such as decision trees, random forests, and gradient-boosted trees. A disadvantage of KernelSHAP is that it is prolonged because it has an exponential running time, whereas TreeSHAP only has a polynomial running time. TreeSHAP uses conditional expectations (Lundberg and Lee (2017)):

$$f_S(x) = E[\hat{f}(x)|X_s] \tag{55}$$

One property of SHAP that is specifically interesting in this research is the SHAP summary plot. This property is relevant because it combines feature importance with feature effects to achieve better model understanding and, thus, explainability. For each feature, the SHAP value is shown (the impact on the model output) in combination with the feature value (high or low). This indicates relationships between the independent variables and the effect on the predictions.

6. Results

This section will discuss our results from the DPPD analysis and the random forest model. In the DPPD analysis, we will look at changes over time of the previously defined CRB quantification methods, namely the number of agricultural fires, average FRP, average NO_2, and average $PM_{2.5}$. Additionally, we will highlight the areas that show significant change. Finally, we will examine the results of training and testing the random forest model for the prediction of NO_2 and $PM_{2.5}$. Using these results, we attempt to clarify changes in CRB.

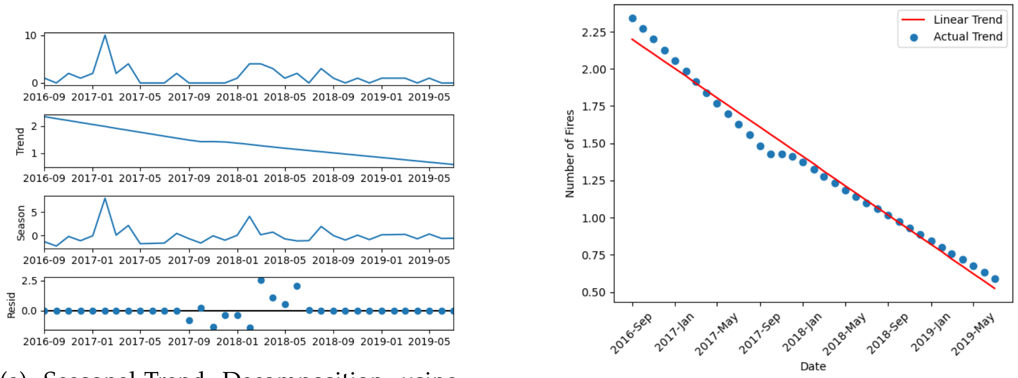
6.1 Data Powered Positive Deviance Analysis

As previously discussed, Seasonal-Trend Decomposition using LOESS (STL) will be used for the DPPD analysis to obtain the trend values of a specific mandal m at month t, with $m \in M$ and $t \in T$. Here the set M is defined as the 592 mandals in Telangana. Set T denotes all months available in the data. The data contains observations between September 2016 and August 2019. Hence, $t = 1$ denotes September 2016, $t = 2$ October 2016, $\ldots$, and $t = 36$ August 2019. In Section 5.1, $y_{m,t}$ is denoted as the quantity of interest in mandal m at month t. In this section we will perform STL for $y_{m,t} = v_{m,t}$ (number of agricultural fires), $y_{m,t} = FRP_{m,t}$ (average FRP), $y_{m,t} = NO2_{m,t}$ (average NO_2), and $y_{m,t} = PM25_m$ (average $PM_{2.5}$). Before applying STL, we need to choose the number of periods within one cycle, p, and a suitable smoothing parameter, q. We have yearly cycles with twelve months, thus $p = 12$. For selecting q, we look at the nature of the data and the goal of this research. In this research, we apply STL on three-year data, a relatively small amount, to measure changes over time. We want significant differences to be visible so we can still draw conclusions. Therefore, we set q as low as possible (and odd) such that it is still smoothing, thus $q = 7$.

We apply STL for each quantification method using $p = 12$ and $q = 7$, resulting in the predicted trend values, denoted by $\hat{g}_{m,t}$. Secondly, we will fit a linear regression line to $\hat{g}_{m,t}$ $\forall t \in T$ separately for each mandal. This results in the slope, a_m, and the intercept, b_m. We will refer to the deviant score as a_m.

A negative deviant score is a positive deviant, indicating a decreasing trend in agricultural fires, FRP, NO_2, or $PM_{2.5}$. In contrast, a positive score is a negative deviant, which means an increase in agricultural fires, FRP, NO_2, or $PM_{2.5}$. Finally, zero (or very close to zero) shows no change in the values. Figure 18 shows the specific procedure of the mandal Yellandu (located in district Bhadradri Kothagudem) when

$y_{m,t} = v_{m,t}$. We observe in Figure 18a the results of STL. STL divides the line at the top in Figure 18a into the trend, the seasonal component, and the residual. Figure 18b shows the trend values from STL results and the fitted Linear Regression line to these trend values. The slope of the red line is the deviant score, a_m. For Yellandu, this is $a_{Yellandu} = -0.05$, meaning that per month the number of crop fires in Yellandu decreases by 0.05.



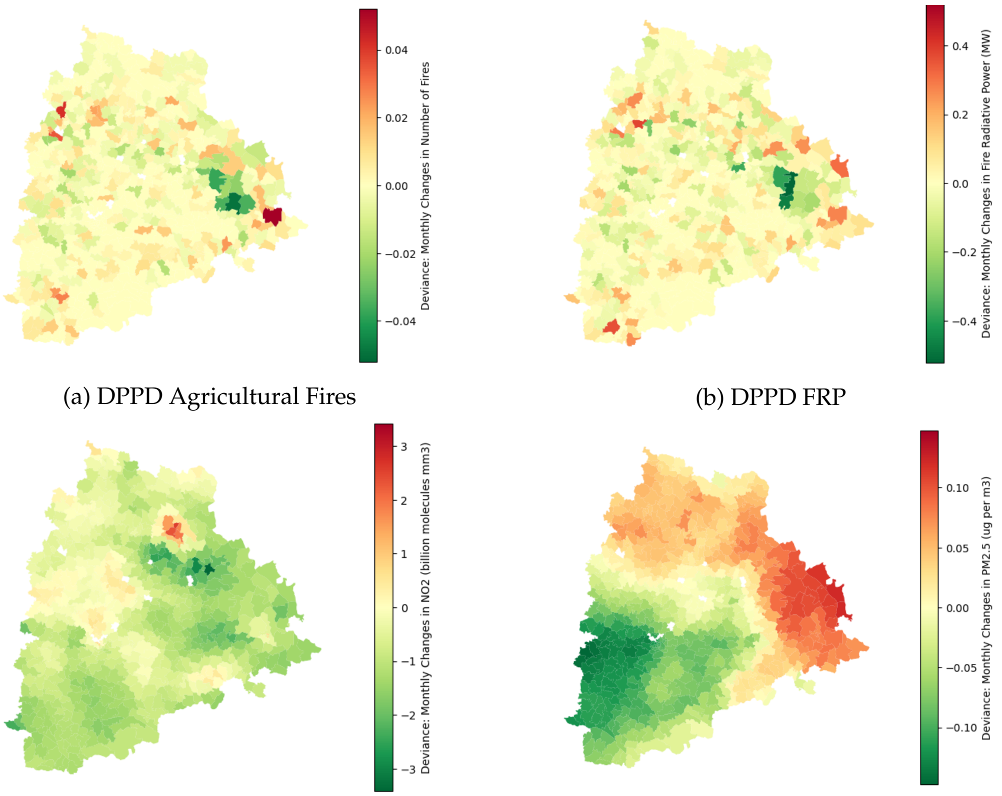
(a) Seasonal-Trend Decomposition using LOESS (STL) Mandal Yellandu Number of Crop Fires

(b) Linear Regression on the Trend of the Mandal Yellandu Number of Agricultural Fires

Figure 18: These visualizations show the procedure of the DPPD analysis for one single mandal, Yellandu. In this example, we have selected the number of agricultural fires as the CRB quantifier. In Figure 18a, we observe the results of STL splitting the line at the top into the trend, the seasonal component, and the residual. Figure 18b shows the trend values from STL results and the fitted Linear Regression line to these trend values. The slope of this line, $a_{Yellandu} = -0.05$, indicates the deviance score. This means that per month the number of crop fires in Yellandu decreases by -0.05.

Figure 19 shows the deviant scores of all mandals per CRB quantification method. We notice that for $y_{m,t} = v_{m,t}$ (number of agricultural fires Figure 19a) and $y_{m,t} = FRP_{m,t}$ (FRP Figure 19b) relatively high negative deviant scores are around the Mahabubabad district. Besides, the relatively high positive deviant scores are around districts Nizamabad, Bhadradri Kothagudem, and Wanaparthy. For Figure 19a, the highest deviant score, $a_m = 0.05$, shows an increase of 0.05 crop fires per month, which is nearly a decrease of two agricultural fires over three years. The same holds for the lowest deviant score, $a_m = -0.05$, showing a decrease of 0.05 crop fires per month, which is a reduction of nearly two agricultural fires over three years. These changes seem considerable when comparing these values with the average mandal values $\bar{v}_m$, visualized in Figure 7a. However, in reality, the differences might not be very significant. We observe the same for the deviant scores of the FRP quantification method in Figure 19b in comparison with Figure 7a. It is challenging to set the threshold for defining when a mandal shows a significant deviant score, $|a_m| > \epsilon$. There is no set-in-stone value because the yearly agricultural fires can vary considerably. It is up to Telangana's government to interpret and respond to these numbers using expert knowledge and statistics.

When examining the outcomes of the deviant scores of NO_2 and $PM_{2.5}$ ($y_{m,t} = NO2_{m,t}$ and $y_{m,t} = PM25_{m,t}$) in Figure 19c and Figure 19d, we observe that for NO_2 the highest rises are around district Peddapalli. The highest NO_2 drops are around Karimnagar and Jayashankar Bhupalpally. For $PM_{2.5}$, the highest inclines are at the North-East of Telangana, and the highest declines are at the South-West of Telangana. However, when comparing the value sizes with 7c and Figure 7d, sizes of increases and decreases are relatively insignificant. Again there exists no set-in-stone value for those thresholds. It is up to the government of Telangana how they want to interpret and respond to these numbers.


0.04
0.02
0.00
−0.02
−0.04
Deviance: Monthly Changes in Number of Fires
0.4
0.2
0.0
−0.2
−0.4
Deviance: Monthly Changes in Fire Radiative Power (MW)
3
2
1
0
−1
−2
−3
Deviance: Monthly Changes in NO2 (billion molecules mm3)
0.10
0.05
0.00
−0.05
−0.10
Deviance: Monthly Changes in PM2.5 (ug per m3)

(a) DPPD Agricultural Fires

(b) DPPD FRP

(c) DPPD NO_2

(d) DPPD $PM_{2.5}$

Figure 19: These visualizations show the results of the DPPD analysis for all mandals for all quantifiers: the number of agricultural fires, FRP, NO_2, and $PM_{2.5}$. Note that a red mandal indicates an increase over time, and a green mandal a decrease over time. The scores per mandal are called the deviance scores a_m. The time is in months.

6.2 Random Forest Model

This section will provide the results from the NO_2 and $PM_{2.5}$ random forest model. This model is trained and tested on monthly available data for each mandal between September 2016 and October 2019. First, we will discuss the results from the independent variable selection. Secondly, we will disclose the final hyperparameter settings we have selected. Additionally, we cover the evaluation metrics of both emission values. Finally, we provide the random forest explainability, including permutation importances and SHAP values.

6.2.1 Independent Variable Selection. Previously, we have discussed that adding multiple highly correlated independent variables in a random forest model causes unreliable permutation importance values of a predictor j, FI_j, with $j \in P$. To resolve this, we will create independent variable subsets, based on the Spearman correlations, rs_{x_i,x_j} with $i, j \in P$. Highly correlated independent variables will be added together in a subset. The model will be trained and evaluated on each subset, and permutation importances are calculated to see what independent variable of a subset has the strongest predictive power. Only this variable will be added to the final model. Sometimes the variables seem to be of low predictive power. This is the case when $R^2 < 0$ or $R^2 \approx 0$ and when $FI_j < 0$ or $FI_j \approx 0$. In this case, none of the variables will be added to the final model. Note that for calculating the R^2, we have selected 4-fold ($k = 4$) blocked cross-validation. This is because the dataset consists of three years of data. Hence, to ensure that the model is trained and tested on different months in each fold, we have selected 4-fold blocked cross-validation. Additionally, we did not select $k > 4$ since this decreases the number of data observations within each fold.

In Figure 20, the Spearman correlations between all independent variables are shown in a heatmap. In Section 4.4, we already discussed that setting thresholds for highly correlated independent variables is challenging and depends on the nature of the research. In this research, we aim to have highly reliable model explainability results. Accordingly, strongly correlated independent variables should not both be included in the final model. Considering this, we choose to set the threshold correlation to be >$|0.8|$. All variables with higher Spearman correlations will be added together in a subset. Using this approach, we obtain the following subsets:

- {SSM_mean, SSM_median, SSM_max, SSM_min, SSM_percentile_90}
- {PRE_mean, PRE_median, PRE_max, PRE_min, PRE_percentile_90, PRE_sum}
- {ST_mean, ST_median, ST_min, ST_max, ST_percentile_90}
- {POP_mean, POP_median}
- {RWI_mean, RWI_median, RWI_percentile_90}

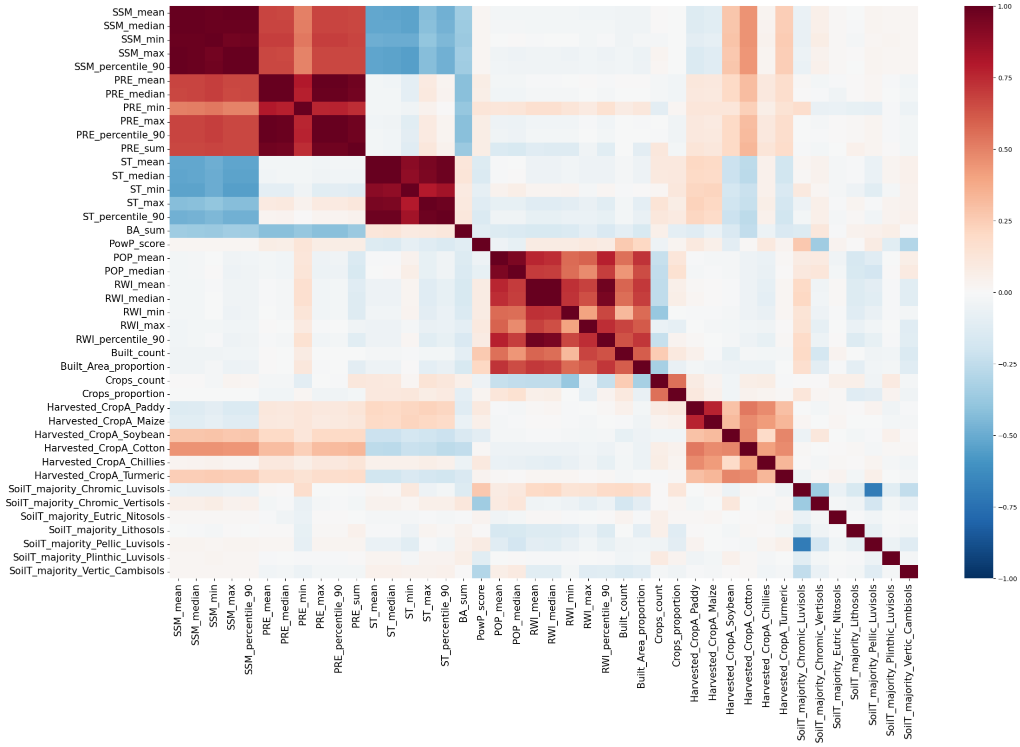


Figure 20: Spearman Correlation Coefficients

Note that not mentioned independent variables can freely be added to the model. We will train a random forest model for each independent variable subset and the single independent variables, predicting NO_2 and $PM_{2.5}$. Using the R^2 and $FI_j \; \forall j \in P$ from the model results, the final independent variables are chosen. We refer to the final independent variables as the set P', with $j' \in P'$, with $P' \subseteq P$. These independent variables are added to the final model. The selected independent variables for both NO_2 and $PM_{2.5}$ are presented per category in Table 10.

Final Independent Variables		
	NO_2	$PM_{2.5}$
Environmental	SSM_percentile_90, PRE_max, ST_max, SoilT_majority_Pellic_Luvisols	SSM_median, PRE_max, ST_max
Socio-Economic	RWI_mean, PowP_score, POP_median, Built_count	RWI_median, PowP_score
Agricultural	Harvested_CropA_Paddy, Harvested_CropA_Maize,	Harvested_CropA_Paddy, Harvested_CropA_Maize

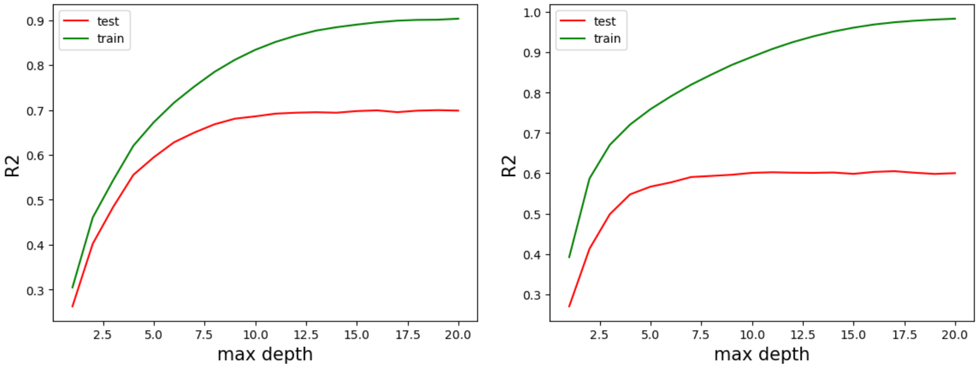
Table 10: Final Independent Variables selected for NO_2 and $PM_{2.5}$

6.2.2 Hyperparameters. In Table 11, the final tuned hyperparameters are shown. Previously, we discussed applying random and grid searches to tune the hyperparameters. Remember that these methods do not necessarily result in optimal hyperparameter settings. We have determined our search area based on knowledge of the data attained through experimentation. For the determination of the hyperparameters, we use 4-fold blocked cross-validation for the same reasons previously mentioned.

In Figure 21, we have visualized the procedure of selecting the ranges for max_depth. For both Figure 21a and 21b, we observe little to no increase in the R^2 if the max_depth increases in the test dataset (red line). Additionally, the increase in R^2 for the training dataset (green line) remains when increasing max_depth, causing the model to be trained precisely to the training data. Hence, at this point, the model starts overfitting. The model will perform accurately on the training data. However, it will not be very generalizable. We observe that the choices of $max\_depth = 9$ and $max\_depth = 7$ for NO_2 and $PM_{2.5}$ seem to be reliable because around these points, the most substantial increase in R^2 has ended. Note that similar figures can be created for other hyperparameters.

Hyperparameters			
	Search Ranges	NO_2	$PM_{2.5}$
$max\_features$	[auto, $\lceil\frac{\lvert P'\rvert}{3}\rceil$]	$\lceil\frac{\lvert P'\rvert}{3}\rceil = 2$	$\lceil\frac{\lvert P'\rvert}{3}\rceil = 2$
max_depth	[2, 3, $\ldots$, 9, 10]	9	7
$n\_estimators$	[300, 400, $\ldots$, 1900, 2000]	400	1400
$bootstrap$	[True, False]	True	True
$min\_samples\_leaf$	[1, 2, 3, 4]	3	1
$min\_samples\_split$	[5, 6, $\ldots$, 19, 20]	13	5

Table 11: Tuned Hyperparameters for NO_2 and $PM_{2.5}$

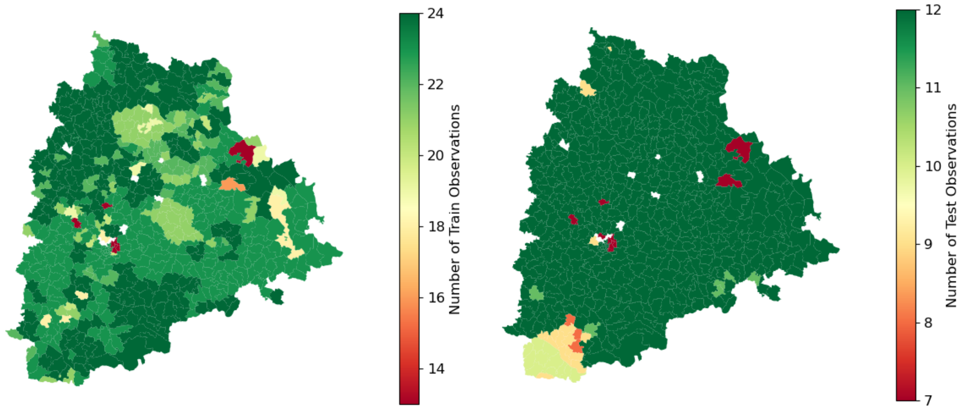


(a) NO_2

(b) $PM_{2.5}$

Figure 21: Model Performances versus Maximum Tree Depth

6.2.3 Evaluation Metrics. After we have selected the independent variable subsets and determined the final hyperparameters for the NO_2 and $PM_{2.5}$ prediction models, we implement the random forest regression models and evaluate the results. For this, we use September 2016 up to and including October 2018 as training data and September 2018 up to and including October 2019 as test data. This results in 13238 training data observations and 6825 test data observations. We have made this selection such that we can evaluate results on a year over time. Additionally, we did this split for the final evaluation to account for the number of data observations issue of blocked cross-validation. Recall that selecting blocked cross-validation reduces the number of data observations within a fold significantly compared to other cross-validation methods. Figure 22 shows the number of train and test observations per mandal. Note that for both train (Figure 22a) and test (Figure 22b), most mandals are relatively close to the maximum number of observations per mandal. This especially holds for the test data. A few mandals have considerably lower train and test data observations (marked dark red). Additionally, in Figure 23, we observe that for the training data, there are significantly fewer observations available in July. We will take both findings into account while assessing the model.



(a) Train Data

(b) Test Data

Figure 22: Number of Observations per Mandal in Train and Test Dataset.

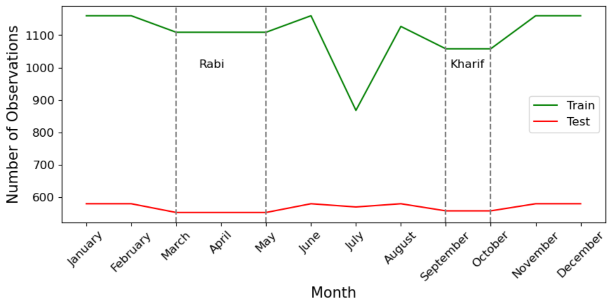


Figure 23: Number of Observations per Month in Train and Test Dataset. The harvest periods are marked with dotted grey lines.

The metrics results are shown in Table 12. We observe a higher R^2 for NO_2 than for $PM_{2.5}$. This means that for NO_2, more variation is explained through the independent variables than for $PM_{2.5}$. In addition to the general performance of the model, we will show the model performance per mandal and over time.

Evaluation Metrics Results		
	NO_2	$PM_{2.5}$
RMSE	52.58	6.80
MAE	41.51	5.15
R^2	0.68	0.59

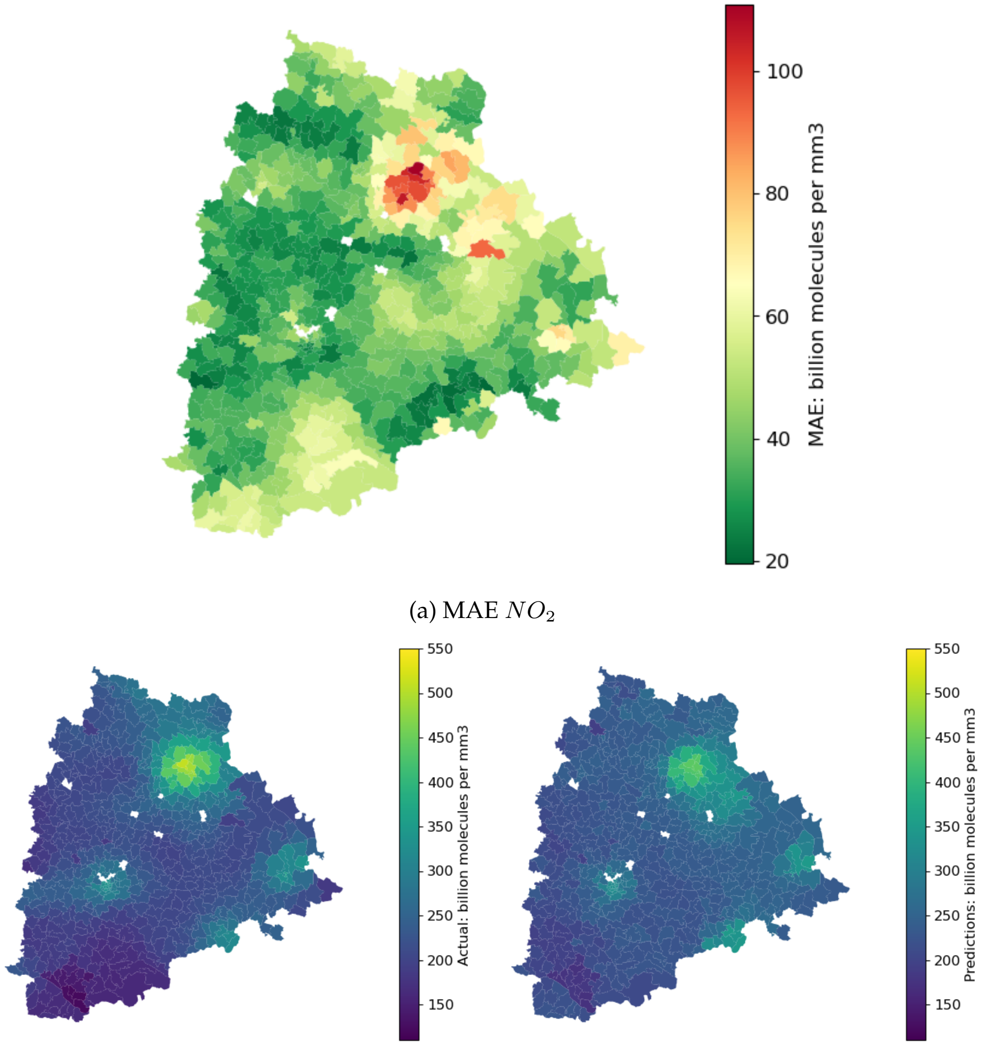
Table 12: Evaluation metrics results for NO_2 and $PM_{2.5}$ on the test data. Recall that the RMSE and MAE of NO_2 and $PM_{2.5}$ are incomparable because they are in different units.

Performance per Mandel. First, we will look at the performances of the prediction models on the test data for NO_2 and $PM_{2.5}$ for each individual mandal. We use September 2018 and August 2019 to evaluate the model. This ensures one full year of test data instances on all months and mandals.

For NO_2, we observe in Figure 24 the performances averaged per mandal. Figure 24a shows the MAE per mandal. Moreover, Figure 24b and 24c show the predicted and actual NO_2 values. Remark that Figure 24c, and Figure 24b contain similar legend ranges, and the legend range of Figure 24a differs. We observe that, in general, the model predicts lower than the actual values. However, the distribution of prediction seems to be approximately correct. This means that the NO_2 values are predicted close to the real values, but generally, the predicted values are smaller than the real NO_2 emission values. The highest errors seem to occur in the district Peddapalli, the region with the highest NO_2 emissions in general. We suspect this is due to the large power plants located in Peddapalli. However, additional research is necessary to confirm this. Recall that an overview of the district names is illustrated in Appendix A, Figure 1.

For $PM_{2.5}$, the performances averaged per mandal are shown in Figure 25. Again remark that Figure 25c and Figure 25b contain similar legend ranges, and the legend range of Figure 25a differs. In Figure 25a, we see that the model performs overall worse around the districts Adilabad, Nizamabad, Nirmal, Peddapalli, Hyderabad, Rangareddy, and Suryapet. For these districts, Rangareddy excluded, we observe higher actual $PM_{2.5}$ values in Figure 25b. In Figure 25c, we notice that the model predicts slightly higher values for the mandals within these districts (in comparison with the average district $PM_{2.5}$ predictions). However, the prediction values are still lower than the actual values. The predictions are too high for mandals around the districts Rangareddy, Mahabubnagar, Nagarkumool, and Bhadradri Kothagude. These districts have generally lower actual $PM_{2.5}$ values. Consequently, we generally conclude that mandals with extreme $PM_{2.5}$ values are predicted worse compared to mandals with average $PM_{2.5}$ values. This indicates that explainable variable(s) are currently missing in the data.

We observe no pattern between the number of data observations per mandal and the MAE of both NO_2 and $PM_{2.5}$. Therefore, we assume that differences in the number of data instances per mandal do not cause the errors.

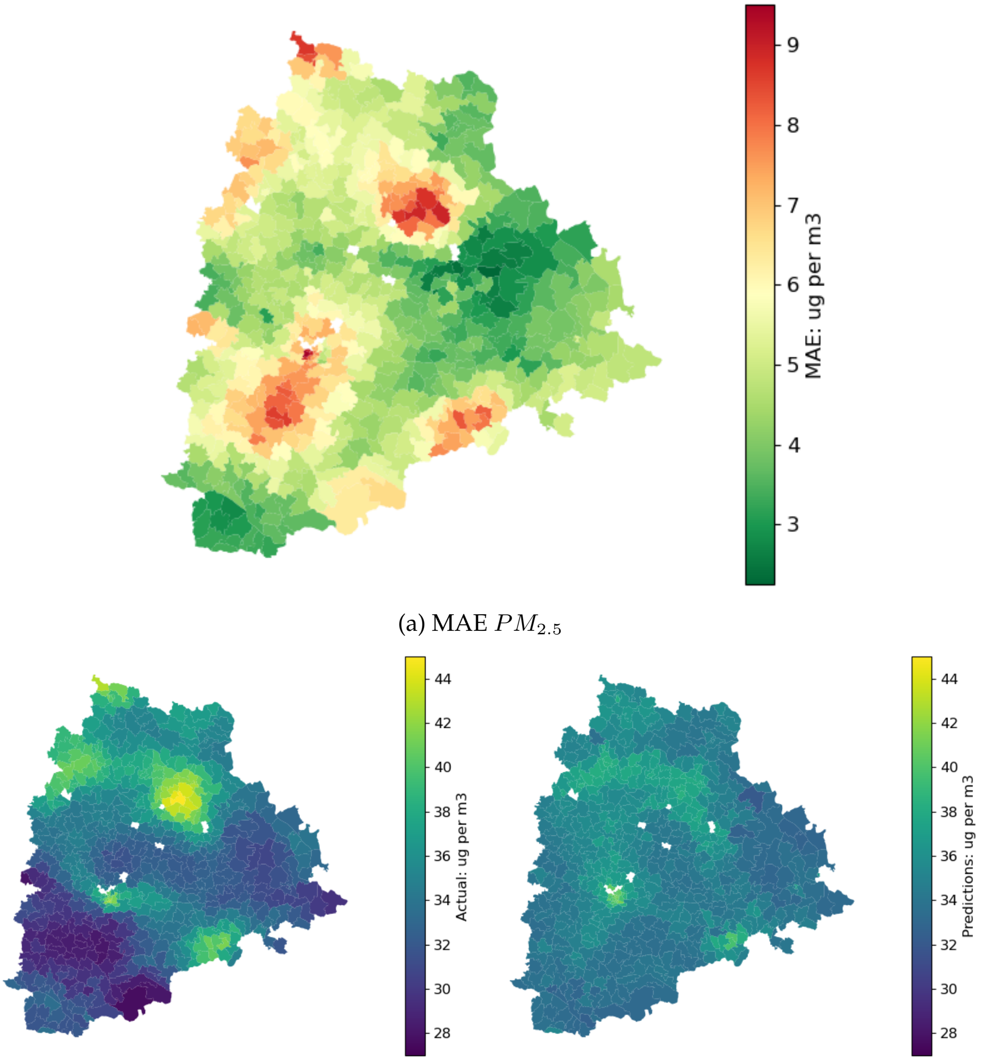

100
80
60
40
20
MAE: billion molecules per mm3
550
500
450
400
350
300
250
200
150
Actual: billion molecules per mm3
550
500
450
400
350
300
250
200
150
Predictions: billion molecules per mm3

(a) MAE NO_2

(b) Actual Values NO_2

(c) Predicted Values NO_2

Figure 24: Model outcomes visualised per mandal for NO_2. Be aware that Figure 24b and 24c have the same legend ranges and that Figure 24a has a different legend range.



(a) MAE $PM_{2.5}$

(b) Actual Values $PM_{2.5}$

(c) Predicted Values $PM_{2.5}$

Figure 25: Model outcomes visualised per mandal for $PM_{2.5}$. Be aware that Figure 25b and 25c have the same legend ranges and that Figure 25a has a different legend range.

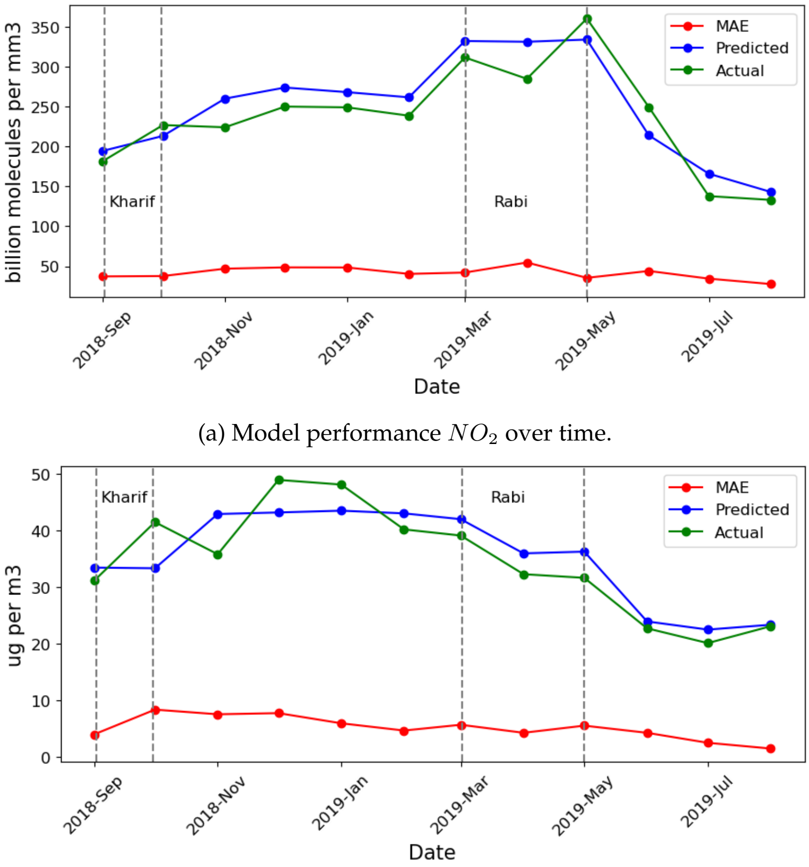
Performance over Time. Secondly, we will look at the model performances of the test data over time for NO_2 and $PM_{2.5}$ visualized in Figure 26. Previously, we examined the performances of the model for each mandal. Subsequently, we will explore whether the model performs worse or better in certain months. Again be aware that we use September 2018 and August 2019 to evaluate the model. This ensures one full year of test data instances on all months and mandals.

For NO_2 these results are shown in 26a. We identify that the model predicts the highest NO_2 values in March, April, and May and the lowest NO_2 values in July and August. High values in March up to and including May is plausible as this is during the harvest period of Rabi crops. Additionally, we observe an increase in actual and predicted values at the end and beginning of the year. It is unknown to us why this increase is occurring. Future research is required to obtain an answer. The MAE is approximately equal over time. However, we observe a slight increase in error in April.

For $PM_{2.5}$ the results are shown in Figure 26b. We perceive that the model predicts the highest $PM_{2.5}$ values in November, December, January, and February, while the lowest values are predicted in June and July. Again, we are unsure about the cause of the increase in $PM_{2.5}$ during November up to including February. The model predicts very accurately during August and less accurately during October, November, and December. Again this indicates that explainable variables(s) are currently missing in the data.

Again, we observe no pattern between the number of data observations over time and the MAE of both NO_2 and $PM_{2.5}$. We previously detected that July has fewer data observations in the training data. However, this month does perform averagely accurately for both NO_2 and $PM_{2.5}$. Hence, we assume that the differences in the number of data instances over time do not cause the errors.

Generally speaking, we conclude that the NO_2 prediction model has a better performance than $PM_{2.5}$ in terms of R^2 values (0.69>0.59). For NO_2, we primarily notice that the prediction distribution seems close to the actual distribution. However, the model mainly provides safe predictions. As a consequence, extreme values are not predicted accurately. The MAE over time is approximately the same. However, a minor rise is displayed around April during the harvest of Kharif. For $PM_{2.5}$, we obtain the same conclusion: primarily extreme $PM_{2.5}$ values are not predicted accurately. Here are the predictions even more off than for NO_2. We suspect this occurs because $PM_{2.5}$ contains more regions with extreme values. However, additional research is necessary to confirm this. The MAE over time shows that the predictions for August are closest to the actual values. During harvest season, the MAE of the predictions in October and November show a higher error value. We refer to Appendix D Figure 1 for a more extensive model performance review. In this figure, for each prediction model (NO_2 or $PM_{2.5}$), the best, average, and worst performing mandal are visualized.



(a) Model performance NO_2 over time.

(b) Model performance $PM_{2.5}$ over time.

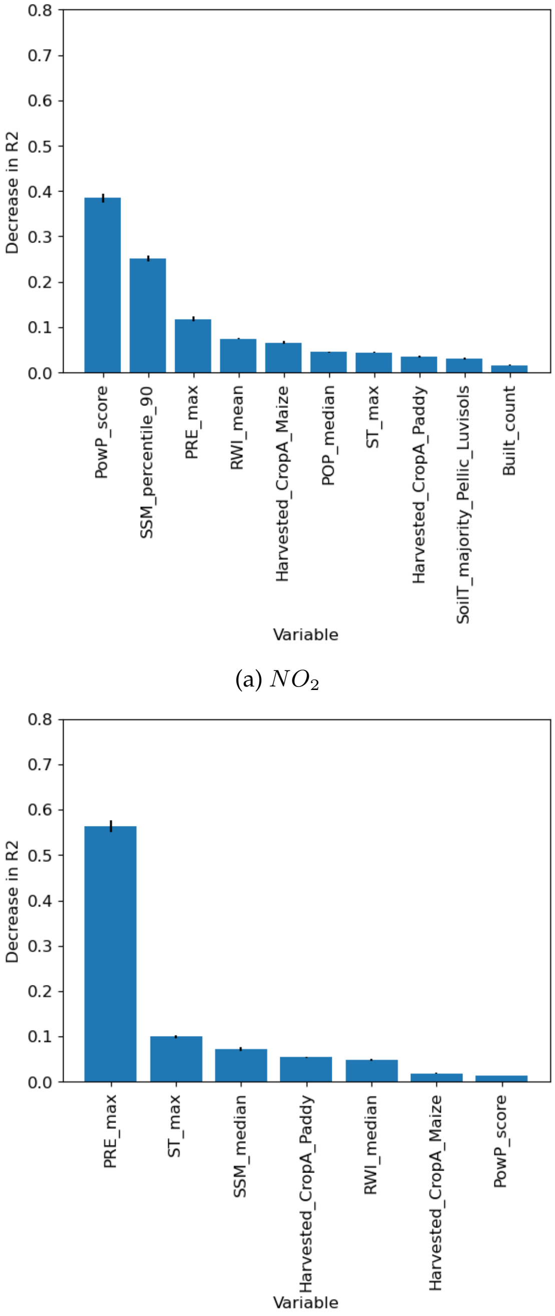
Figure 26: These visualizations show the model outcomes of the test data for NO_2 and $PM_{2.5}$ over time. In Figure 26a the values for NO_2 are shown and in Figure 26b the values for $PM_{2.5}$ are shown. Note that the harvest periods are marked with grey dotted lines.

6.2.4 Random Forest Explainability. To understand why the model gives the previously discussed predictions, we will analyze the permutation importances and the SHAP values for each predictor j'. Recall that $j' \in P'$ is the set of the selected independent variables in the final model. The permutation importances give us insight into how important the independent variable is for the prediction, and the SHAP values show how the independent variable influences the model output. We will discuss the results for NO_2 and $PM_{2.5}$. In Table 13, we summarize the independent variable permutation importances and their model relation according to the SHAP values. The model relation describes the relationship between changes in the independent variable and the model outcomes, SHAP values. If it is positive, an increase in the independent variable value leads to an increase in the model prediction value and vice versa. In addition, if it is negative, an increase in the independent variable value leads to a decrease in the model prediction value and vice versa. Here we discuss the permutation importances and the SHAP values of NO_2 and $PM_{2.5}$ in more detail.

Model Explainability Summary			
Independent Variable	**Category**	FI	**Model Relation** [7]
	NO_2		
PowP_score	Socio-Economic	0.39	+
SSM_percentile_90	Environmental	0.25	-
PRE_max	Environmental	0.11	-
RWI_median	Socio-Economic	0.07	+
Harvested_CropA_Maize	Agricultural	0.06	N/A
POP_median	Socio-Economic	0.04	+
ST_max	Environmental	0.04	+
Harvested_CropA_Paddy	Agricultural	0.04	+
SoilT_majority_Pellic_Luvisols	Environmental	0.03	+
Built_count	Socio-Economic	0.02	+
	$PM_{2.5}$		
PRE_max	Environmental	0.56	-
ST_max	Environmental	0.10	-
SSM_median	Environmental	0.07	-
Harvested_CropA_Paddy	Agricultural	0.05	+
RWI_median	Socio-Economic	0.05	+
Harvested_CropA_Maize	Agricultural	0.02	N/A
PowP_score	Socio-Economic	0.01	+

Table 13: A summary of the overall effect of the independent variable on the model predictions is shown. Note that the independent variables are sorted according to their permutation importance scores. We start with the independent variable with the highest permutation score (FI) and end with the lowest permutation score for NO_2 and $PM_{2.5}$.

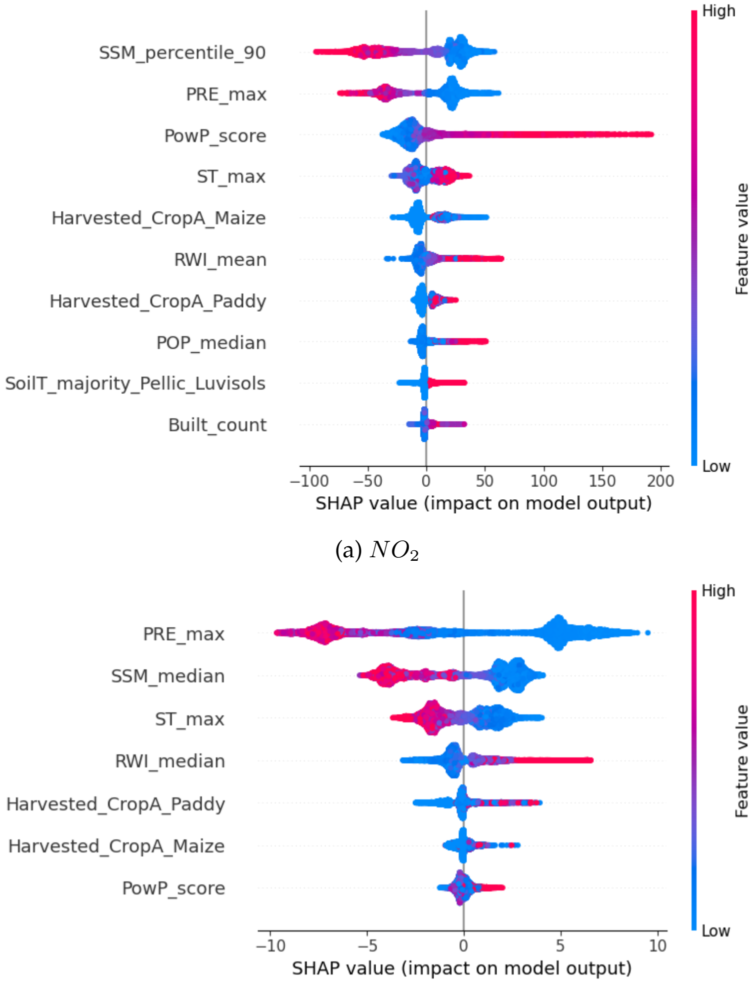
[7] We define an independent variable to have a positive model relation (+) if an increase in independent variable values leads to an increase in SHAP values and an independent variable to have a negative model impact (-) if an increase in independent variable values leads to a decrease in SHAP values. We define the model impact to be N/A if there is no real pattern visible in the SHAP values.



(a) NO_2

(b) $PM_{2.5}$

Figure 27: This figure shows the permutation importances, $FI_{p'}$ of the final selected independent variables for the final model. In Figure 27a the results are shown for the NO_2 prediction model and in Figure 27b the results are shown for the $PM_{2.5}$ prediction model.



(a) NO_2

(b) $PM_{2.5}$

Figure 28: This figure shows the SHAP values, ϕ_i for each observation i. In Figure 28a the results are shown for the NO_2 prediction model and in Figure 28b the results are shown for the $PM_{2.5}$ prediction model.

NO_2. We will analyze Figure 27a and 28a to draw conclusions from the independent variable permutation scores and the SHAP values. First, we conclude that the most relevant socio-economic independent variable is PowP_score. It has a permutation score of nearly 0.39, indicating that if this variable is permuted, the R^2 will decrease by 0.39. A higher PowP_score leads to higher model predictions. In other words, if a mandal is located close to a power plant, it will obtain higher NO_2 emission predictions. If we proceed to look at all independent variables accordingly, this will be their impact (sorted according to the permutation importances):

- PowP_score ($FI \approx 0.39$): If a mandal is located close to a power plant, it will obtain higher NO_2 emission predictions and vice versa.
- SSM_percentile_90 ($FI \approx 0.25$): If a mandal's 90th percentile soil moisture is high (usually during May - November), it will obtain lower NO_2 predictions and vice versa.
- PRE_max ($FI \approx 0.11$): If the maximum rainfall measured in a mandal is higher (usually during May - November), it will obtain lower NO_2 predictions and vice versa.
- RWI_mean ($FI \approx 0.07$): If the average relative wealth index of a mandal is high, it will result in higher NO_2 predictions for that mandal and vice versa.
- Harvested_CropA_Maize ($FI \approx 0.06$): If the area of maize crops harvested in a mandal is small (usually off-harvest season, thus January - February, June - August, November - December), it can result in both higher and lower NO_2 predictions. On the other hand, if the area of maize crops harvested in a mandal is large (usually during harvest season, thus March-May and September - October), it can predict higher NO_2 values.
- POP_median ($FI \approx 0.04$): If the population's median in a mandal is high, it will obtain higher NO_2 predictions and vice versa.
- ST_max ($FI \approx 0.04$): If the maximum soil temperature in a mandal is high (usually during January - July), it will obtain higher NO_2 predictions and vice versa.
- Harvested_CropA_Paddy ($FI \approx 0.04$): If the area of paddy harvested in a mandal is large (usually during harvest season, thus March-May and September - October), it will obtain higher NO_2 predictions and vice versa.
- SoilT_majority_Pellic_Luvisols ($FI \approx 0.03$): If the majority of the soil type within a mandal is pellic luvisols, it will obtain higher NO_2 predictions and vice versa. Recall that this soil type is suitable for crop growth.
- Built_count ($FI \approx 0.02$): If the area of built area in a mandal is large, it will obtain higher NO_2 predictions and vice versa.

$\boldsymbol{PM_{2.5}}$. While describing the independent variable influences on the $PM_{2.5}$ model, we will analyze Figure 27b and 28b. We will proceed to look at all independent variables for $PM_{2.5}$ just like we did for NO_2 and discuss their impact (sorted according to the permutation importance):

- PRE_max ($FI \approx 0.56$): If the maximum rainfall measured in a mandal is higher (usually during May - November), it will obtain lower $PM_{2.5}$ predictions and vice versa.
- ST_max ($FI \approx 0.10$): If the maximum soil temperature in a mandal is high (usually during January - July), it will obtain higher $PM_{2.5}$ predictions and vice versa.
- SSM_mean ($FI \approx 0.07$): If the average soil moisture in a mandal is high (usually during May - November), it will obtain lower $PM_{2.5}$ predictions and vice versa.
- Harvested_CropA_Paddy ($FI \approx 0.05$): If the area of paddy harvested in a mandal is large (usually during harvest season, thus March-May and September - October), it will obtain higher $PM_{2.5}$ predictions and vice versa.
- RWI_median ($FI \approx 0.05$): If the median of relative wealth index in a mandal is high, it will result in higher $PM_{2.5}$ predictions for that mandal and vice versa.
- Harvested_CropA_Maize ($FI \approx 0.02$): If the area of maize crops harvested in a mandal is minor (usually off-harvest season, thus January - February, June - August, November - December), it can result in both higher and lower $PM_{2.5}$ predictions. On the other hand, if the area of maize crops harvested in a mandal is large (usually during harvest season, thus March-May and September - October), it can predict higher $PM_{2.5}$ values.
- PowP_Score ($FI \approx 0.01$): If a mandal is located close to a power plant, it will obtain higher $PM_{2.5}$ emission predictions and vice versa.

We conclude that most of Telangana's NO_2 emissions are due to the independent socio-economic and environmental variables and not necessarily from agricultural emissions. Especially, the distance from a mandal to a powerplant (with specific capacity) and the 90th percentile soil moisture of a mandal have much impact on the NO_2 predictions. Additionally, for $PM_{2.5}$, we conclude that the $PM_{2.5}$ emissions in Telangana are mainly influenced by the independent environmental variables and not necessarily the agricultural and socio-economic independent variables. Especially, the maximum rainfall value within a mandal has much impact on the $PM_{2.5}$ predictions. Both models conclude that the crops that contribute most are maize and paddy. The fact that we observe paddy is logical because it is the most dominant crop in Telangana. However, it is remarkable that maize is selected because it has a much lower sown crop area than paddy in Telangana. This indicates that maize could be a particularly polluting crop in terms of NO_2 and $PM_{2.5}$ emissions.

7. Discussion and Limitations

In the discussion in Section 7.1, we focus on explaining and evaluating the results of our research. Additionally, in the limitations in Section 7.2, we will discuss weaknesses in the design and analyses of our research.

7.1 Discussion

Given the discovered results, we revisit the research questions presented earlier in this paper.

1. How can we identify farming communities that are, over time, reducing or increasing the impact on the environment and health by CRB?

To answer the first sub-research question, we explore the DPPD analysis results in Section 6.1. In the DPPD analysis, we estimate the general course of the trend of CRB in a mandal while accounting for seasonal differences. We use multiple CRB quantifiers, such as the number of agricultural fires, the FRP emissions of agricultural fires, the NO_2 emissions, and the $PM_{2.5}$ emissions. After performing the DPPD, we obtain a deviant score for each mandal for each quantifier. This score represents a specific quantifier's general increase or decrease per month. These scores are relevant for the government to determine targeted interventions. However, further research is necessary to set appropriate thresholds for the deviance scores such that the government only responds to significant deviant scores, called positive and negative deviants. Here positive deviants are defined as farmers reducing the practice of CRB, and negative deviants as farmers increasing the practice of CRB. The Telangana government can establish thresholds by applying expert knowledge and statistics.

2. What changes have occurred for farmers that are increasing or decreasing the activity of crop residue burning in their environmental, socio-economic, and agricultural situations compared to other farmers?

Additionally, to answer the second sub-research question, we analyze the findings from the random forest model. The number of agricultural fires and FRP emissions of agricultural fires contain many zero values causing increased difficulty in using common predictive modeling techniques. Therefore, we decide to build a prediction model for monthly emissions of NO_2 and $PM_{2.5}$. We find for both NO_2 and $PM_{2.5}$ that the most significant factors for the prediction are socio-economic factors, such as power plant capacities, and environmental factors, such as soil moisture, rainfall, and soil temperature. Consequently, we acknowledge that agricultural factors seem less relevant in the prediction of NO_2 and $PM_{2.5}$ in Telangana, as initially presented in previous research. We find maize and paddy the most significant crops contributing to NO_2 and $PM_{2.5}$. This is expected for paddy because it is the most dominant crop in Telangana. However, for maize, this is remarkable. This indicates that maize could be a particularly polluting crop in terms of NO_2 and $PM_{2.5}$ emissions. These findings emphasize the need for additional data to explore direct CRB effects. Additionally, we find that the NO_2 predictions are more accurate than those for $PM_{2.5}$. For both quantifiers, we perceive approximately correct prediction distributions between mandals. However, both models fail to predict extreme values accurately. Therefore, we conclude that the random forest model mainly results in predictions close to the mean. Especially for

$PM_{2.5}$, the prediction values of extreme values seem to be far off. Accordingly, we have sufficient evidence that explainable variables are currently missing in the data. The model performance evaluation over time in Section 6.2.3 shows no significant changes in performance in specific months or during the harvest season. In general, we conclude that the performance of the prediction model for NO_2 is acceptable and that additional data can improve the prediction model for $PM_{2.5}$.

How can we utilize satellite data to find factors that influence crop residue burning?

Finally, to answer the main research question, we combine the answers to the previous sub-research questions. This approach results in possibilities for the government regarding targeted interventions and a random forest model with acceptable performance that is of great explainable power. Be aware that this is one of the first research papers that attempts to predict CRB, focusing primarily on explainable modeling techniques. In addition, the analysis is completely developed with open data, making it widely accessible and reproducible. Consequently, future data can be easily added to the model. We conclude that agricultural factors do not contribute to NO_2 and $PM_{2.5}$ as much as initially expected. Besides, there is evidence that maize could be a particularly polluting crop in terms of NO_2 and $PM_{2.5}$ emissions. Moreover, this research presents the importance and possibilities of gathering additional data. The government of Telangana can exploit the DPPD analysis to identify positive and negative deviant farmers and find causes for these positive and negative deviant farmers using the random forest model. Unfortunately, due to the lack of data, we cannot suggest direct interventions to the government of Telangana to curb CRB. In the limitations in section 7.2, we will further examine this matter. Concludingly, we can gain more insights into CRB factors in Telangana using a combination of the DPPD analysis and the random forest model. However, to obtain concrete solutions, additional data is essential.

7.2 Limitations

This research involves limitations regarding the available data, DPPD analysis, and the random forest model.

As previously discussed, we are primarily limited by the available data in this research. The first limitation is regarding the quantification of CRB. We select four methods for this procedure. Whereas quantifiers, such as the number of fires and FRP emissions, are not entirely reliable or challenging to predict due to the number of zeros, others are not solely directed to CRB. We observed that the number of agricultural fires and FRP are solely linked to CRB. However, the algorithm that flags fire events marks an event if one or more fires occur. Therefore, the exact number of agricultural fires still needs to be discovered. Besides, we cannot perform evaluation techniques on the classification of agricultural fires due to the unavailability of fire labels. We have found that agricultural factors do not contribute significantly to NO_2 and $PM_{2.5}$. Hence, the validity of these quantification methods declines. We assume that adding data about the amount of crop residue that is burnt or the emission that is solely emitted through CRB to the model would increase the relevance of this research.

Furthermore, the main reason we cannot extract direct interventions from our research is the unavailability of data regarding CRB alternatives and government interventions. Consequently, we are incapable of evaluating CRB alternatives or government interventions. Accordingly, it remains to be seen what alternatives or government interventions are most effective under which scenario.

Additionally, we were forced to make various assumptions concerning the sown and harvested crop area data. First, there is the inconsistent spelling of the district, mandal, and crop names in the crop area data. Hence, it is probable that mistakes were made in the process of matching these names. Subsequently, some mandals are absent in the crop area data despite the land use primarily being crops. To add to that, this data only contains seasonal data. As a result, it is unclear when certain crops are specifically sown and harvested. We have decided to set the harvested crop area of each month equal to the total crop area. Altogether this is likely to decrease the data validity. Adding higher reliable crop data will improve the model's reliability.

The last limitation concerning the available data is the data resolution. Currently, only low-resolution data is available. This causes developing interventions on the farmer level to be impossible. We have focused on mandal-level data because it is still the most accurate data at which important data is available. It can provide insights about farmers within a specific mandal. However, the primary intention was to focus on farmer-level interventions or alternatives.

Next, we discuss the limitations we encounter regarding the DPPD analysis. First, we stated that in the DPPD, we determine the linear line movement of CRB using the four quantification methods. We average this movement over a certain period as a deviant score. This score summarizes the general trend while not showing sudden strong inclines and declines. For example, if we interpret a score of approximately zero, this indicates a roughly constant movement of CRB over the years. However, a substantial increase followed by a strong decrease results in a similar score. Misinterpreting these scores might result in falsely selected interventions.

Additionally, we have yet to perform extensive threshold-setting research to select a suitable threshold. We know that a threshold highly depends on the quantification method and nature of the research. However, additional research on statistically justifying thresholds would increase the relevance of this research.

Finally, we have chosen to apply a random forest model because it is reliable and straightforwardly interpretative. Nevertheless, we are unsure whether this choice results in the most accurate findings and performances. We stated that the NO_2 provides acceptable results. However, if a wider variety of models had been evaluated, this could have been verified more confidently. Another interesting approach may be to study less common modeling techniques for predicting quantifiers containing many zeros could have led to more direct CRB causes.

Furthermore, we cannot conclude causal effects with high certainty. This is because the independent variables might not directly affect NO_2 or $PM_{2.5}$. The independent variables might be correlated with NO_2 and $PM_{2.5}$ but do not necessarily contain causal effects.

8. Conclusion and Future Work

In Section 8.1, we discuss the main takeaways of this research. Subsequently, we suggest additional future work in Section 8.2.

8.1 Conclusion

Farmers are still performing CRB frequently due to its efficiency and cost-effectiveness, despite its negative impact on environmental and human health. Previous government interventions and existing CRB alternatives provide insufficient effects to curb CRB. Therefore, in this research, we aim to find methods to utilize satellite data to find factors that influence CRB.

We succeeded in marking positive and negative deviant areas in Telangana through the DPPD analysis that can be used for targeted interventions. To add to that, we discovered that in Telangana, environmental factors (precipitation, soil moisture, and soil temperature) and socio-economic factors (power plant capacity) contribute significantly more to the NO_2 and $PM_{2.5}$ emissions than agricultural factors (sown crop area). Therefore, we claim that NO_2 and $PM_{2.5}$ are not valid quantifiers of CRB in Telangana. Nevertheless, we found that the maize crop area significantly influences the model performance next to the paddy crop area. This is remarkable since the total area size of maize is considerably low compared to the area size of paddy in Telangana. Despite stating that NO_2 and $PM_{2.5}$ are not valid methods to quantify CRB, we developed a prediction model with open data, acceptable performance, and high explainable power. Unfortunately, gathering additional data is inevitable to come up with direct interventions. To conclude, this research has set the first step in identifying CRB areas and causes. However, additional data is essential to obtain concrete CRB indicators and solutions.

8.2 Future Work

For future work regarding the DPPD, it might be beneficial to add additional information to the deviant scores for completeness. For example, DiCRA can show the deviant scores per mandal and add an interactivity element that shows the trend when the user clicks on a certain mandal. This approach leads to a higher understanding of the deviant scores and changes during specific times. In addition, it makes determining a significant threshold for deviance less complicated because there is more knowledge about the changes. In the future, the significance thresholds can be additionally explored through t-tests, ANOVA tests, and confidence intervals (Cox (1982)).

Furthermore, for the modeling of CRB, it can be of interest to explore models that perform well when the dependent variable contains an excess of zero-valued data. An example is the Zero Inflated Poisson Regression Model (Loeys et al. (2012)). This model focuses on count-based datasets in which the dependent variable contains many zeros. This approach enables modeling the number of agricultural fires and FRP emissions as direct quantifiers for CRB.

Finally, if additional data is available, possibilities for building an optimization model can be explored. For instance, designing a model that optimizes farmer yield while considering CRB emissions as a penalty term. This optimization model supports the farmers in deciding what crops to sow and harvest in which period. Additionally, the model could offer CRB alternatives, give insights into effective government interventions, provide optimal penalty costs, or provide optimal subsidies for curbing CRB. Be aware that implementing this suggestion requires a large amount of additional data.

References

Agrawal, G. K. (2019). Mahatma gandhi national rural employment guarantee act: Design failure, implementation failure or both? *Management and Labour Studies*, 44(4):349–368.

Bai, B., Zhao, H., Zhang, S., Zhang, X., and Du, Y. (2021). Can neural networks forecast open field burning of crop residue in regions with anthropogenic management and control? a case study in northeastern china. *Remote Sensing*, 13(19).

Bergmeir, C. and Benítez, J. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

Bhuvaneshwari, S., H. H. . M. J. N. (2019). Crop residue burning in india: Policy challenges and potential solutions. *International journal of environmental research and public health*, 832.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brownlee, J. (2020). Hyperparameter optimization with random search and grid search. *Machine Learning Mastery*, 12.

CBS (2016). Gemiddelde oppervlakte grond per bedrijf op land- en tuinbouwbedrijven, 2016.

CIMMYT (2019). Happy seeder can reduce air pollution and greenhouse gas emissions while making profits for farmers.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess (with discussion). *Journal of Official Statistics*, 6:3–73.

Cleveland, W. P. and Tiao, G. C. (1976). Decomposition of seasonal time series: A model for the census x-11 program. *Journal of the American Statistical Association*, 71(355):581–587.

Cox, D. R. (1982). Statistical significance tests. *British journal of clinical pharmacology*, 14(3):325.

Express, T. N. I. (2018). 85.9 per cent of farmers in telangana are marginal.

Fernando Rainho Alvest Torgo, L. (1999). Inductive learning of tree-based regression models. pages 57–6.

Food and Organization, A. (2007). Digital soil map of the world. data retrieved from, https://data.apps.fao.org/map/catalog/srv/eng/catalog.search#/metadata/446ed430-8383-11db-b9b2-000d939bc5d8.

Funk, Chris, P. P. M. L. D. P. J. V. S. S. G. H. J. R. L. H. A. H. . J. M. (2022). Climate hazards group infrared precipitation with station data. data retrieved from University of California, https://chc.ucsb.edu/data/chirps.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Gholamy, A., Kreinovich, V., and Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation.

Guanghua Chi, Han Fang, S. C. J. E. B. (2021). Relative wealth index. data retrieved from OCHA Services, https://data.humdata.org/dataset/relative-wealth-index.

Hyndman, R. (2010). *Moving Averages*, pages 866–869.

Kala, N. (2021). Crop residue burning in india.

Kamińska, J. A. (2019). A random forest partition model for predicting no2 concentrations from traffic flow and meteorological conditions. *Science of The Total Environment*, 651:475–483.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., and Brumby, S. P. (2021). Global land use / land cover with sentinel 2 and deep learning. pages 4704–4707. data retrieved from, https://env1.arcgis.com/arcgis/rest/services/Sentinel2_10m_LandCover/ImageServer.

Khaiwal, R., Singh, T., and Mor, S. (2022). Covid-19 pandemic and sudden rise in crop residue burning in india: issues and prospects for sustainable crop residue management. *Environmental Science and Pollution Research*, 29.

Kourentzes, N. (2014). Additive and multiplicative seasonality.

Kumar, R., Ghude, S. D., Biswas, M., Jena, C., Alessandrini, S., Debnath, S., Kulkarni, S., Sperati, S., Soni, V. K., Nanjundiah, R. S., and Rajeevan, M. (2020). Enhancing accuracy of air quality and temperature forecasts during paddy crop residue burning season in delhi via chemical data assimilation. *Journal of Geophysical Research: Atmospheres*, 125(17):e2020JD033019. e2020JD033019 2020JD033019.

Loeys, T., Moerkerke, B., De Smet, O., and Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65(1):163–180.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Masih, A. (2019). Application of random forest algorithm to predict the atmospheric concentration of no2. In *2019 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, pages 252–255.

Mehta, P., W. C. H. D. A. R. C. B. M. F. C. K. . S. D. J. (2019). A high-bias, low-variance introduction to machine learning for physicists. 810:1–124.

Mohan, V. (2017). Stubble burning hit health of 84% people in punjab: Survey.

Molnar, C. (2022). Interpretable machine learning. 191:192–213.

Mottaghinejad, S. (2021). The intuition behind bias and variance.

Naresh, R., Dhaliwal, S., Chandra, M., Malhotra, S., Jana, H., Singh, P., Kumar, V., Baliyan, A., and Gawdiya, S. (2021). Alternative uses, in and off-field managements to reduce adverse impact of crop residue burning on environment: A review. *International Journal of Environment and Climate Change*.

NASA, F. (2022a). Modis active fire detections. data retrieved from NASA FIRMS, `https://earthdata.nasa.gov/firms`.

NASA, G. (2022b). Soil moisture. data retrieved from NASA GSFC, `https://developers.google.com/earth-engine/datasets/catalog/NASA_USDA_HSL_SMAP10KM_soil_moisture#description`.

NASA, N. (2010). Nitrogen dioxide. data retrieved from NASA NEO, `https://avdc.gsfc.nasa.gov/pub/data/satellite/Aura/OMI/V03/L3/OMNO2d_HR/`.

National Policy for Management of Crop Residue, N. (2014).

of Telangana, G. (2016). Telangana open data portal 2016.

of Telangana, G. (2022). Directorate of economics and statistics.

Pandey, S. (2022). Crop residue burning in telangana — undp dicra.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Randomforestregressor. *Journal of Machine Learning Research*, 12:2825–2830.

Plummer, A. (2020). *Different Types of Time Series Decomposition*.

Portal, K. D. (2017). Power plants. data retrieved from NASA GSFC, `https://datasource.kapsarc.org/explore/dataset/coal-power-plants-database_india_final/information/?disjunctive.region&disjunctive.state&disjunctive.district&disjunctive.location&disjunctive.plant_name&disjunctive.coal_type&disjunctive.owner&disjunctive.sector&disjunctive.status&refine.state=Telangana&basemap=a92047&location=12,13.27787,79.34275`.

Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3):e1301.

Ram, N. (2020). Crop residue management: Solution to achieve better air quality.

Ram, S. (2021). Season of harvest - what is harvested when where in india?

Salary, A. (2022).

Schober, Patrick MD, P. M. B. C. P. M. S. L. A. M. P. M. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia Analgesia*, 126.

Service, C. G. L. (2022). Burnt area. data retrieved from Copernicus, `https://land.copernicus.eu/global/products/ba`.

Shamin, J. (2022). Trends and pattern of foodgrains production in indias. *Natural and Social Sciences*, 10:7–24.

Shen, Y., Jiang, C., Chan, K. L., Hu, C., and Yao, L. (2021). Estimation of field-level nox emissions from crop residue burning using remote sensing data: A case study in hubei, china. *Remote Sensing*, 13(3).

Shyamsundar, P., Springer, N. P., Tallis, H., Polasky, S., Jat, M. L., Sidhu, H. S., Krishnapriya, P. P., Skiba, N., Ginn, W., Ahuja, V., Cummins, J., Datta, I., Dholakia, H. H., Dixon, J., Gerard, B., Gupta, R., Hellmann, J., Jadhav, A., Jat, H. S., Keil, A., Ladha, J. K., Lopez-Ridaura, S., Nandrajog, S. P., Paul, S., Ritter, A., Sharma, P. C., Singh, R., Singh, D., and Somanathan, R. (2019). Fields on fire: Alternatives to crop residue burning in india. *Science*, 365(6453):536–538.

Telangana, G. (2001). District and mandal shapefiles of telangana. data retrieved from the Government of Telangana, `https://data.telangana.gov.in/story/telangana-district-and-mandal-shape-files`.

Telangana, G. (2019). Crop sown and harvested per district and mandal. data retrieved from the Government of Telangana, `https://data.telangana.gov.in/dataset/kamareddy-district-mandal-wise-crop-areas`.

Thandra, B., Sairam, M., Shankar, T., Maitra, S., and Praharaj, S. (2021). Crop residue burning in india: Causes, impacts and solutions. 12:37430–37437.

Theodosiou, M. (2011). Forecasting monthly and quarterly time series using stl decomposition. *International Journal of Forecasting*, 27(4):1178–1195.

Thiagarajan, K. (2022). The world's most polluted capital city.

van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M., and Winker, D. M. (2016). Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology*, 50(7):3762–3772. data retrieved from, https://sites.wustl.edu/acag/datasets/surface-pm2-5/.

Wan, Z., H. S. H. G. (2021). Modis/terra land surface temperature/emissivity daily l3 global 1km sin grid v061 [data set]. data retrieved from NASA Earth, https://doi.org/10.5067/MODIS/MOD11A1.061.

WorldPop (2022). Population. data retrieved from WorldPop, https://hub.worldpop.org/geodata/listing?id=76.

Yadav, R. S. (2019). Stubble burning: A problem for the environment, agriculture and humans.

Zar, J. H. (2005). *Spearman Rank Correlation*. John Wiley Sons, Ltd.

Zhou, Y., Han, Z., Liu, R., Zhu, B., Li, J., and Zhang, R. (2018). A modeling study of the impact of crop residue burning on PM2.5 concentration in beijing and tianjin during a severe autumn haze event. *Aerosol and Air Quality Research*, 18(7):1558–1572.

Zou, K. H., Tuncali, K., and Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, 227(3):617–628.

Appendix A: Visualisation Telangana

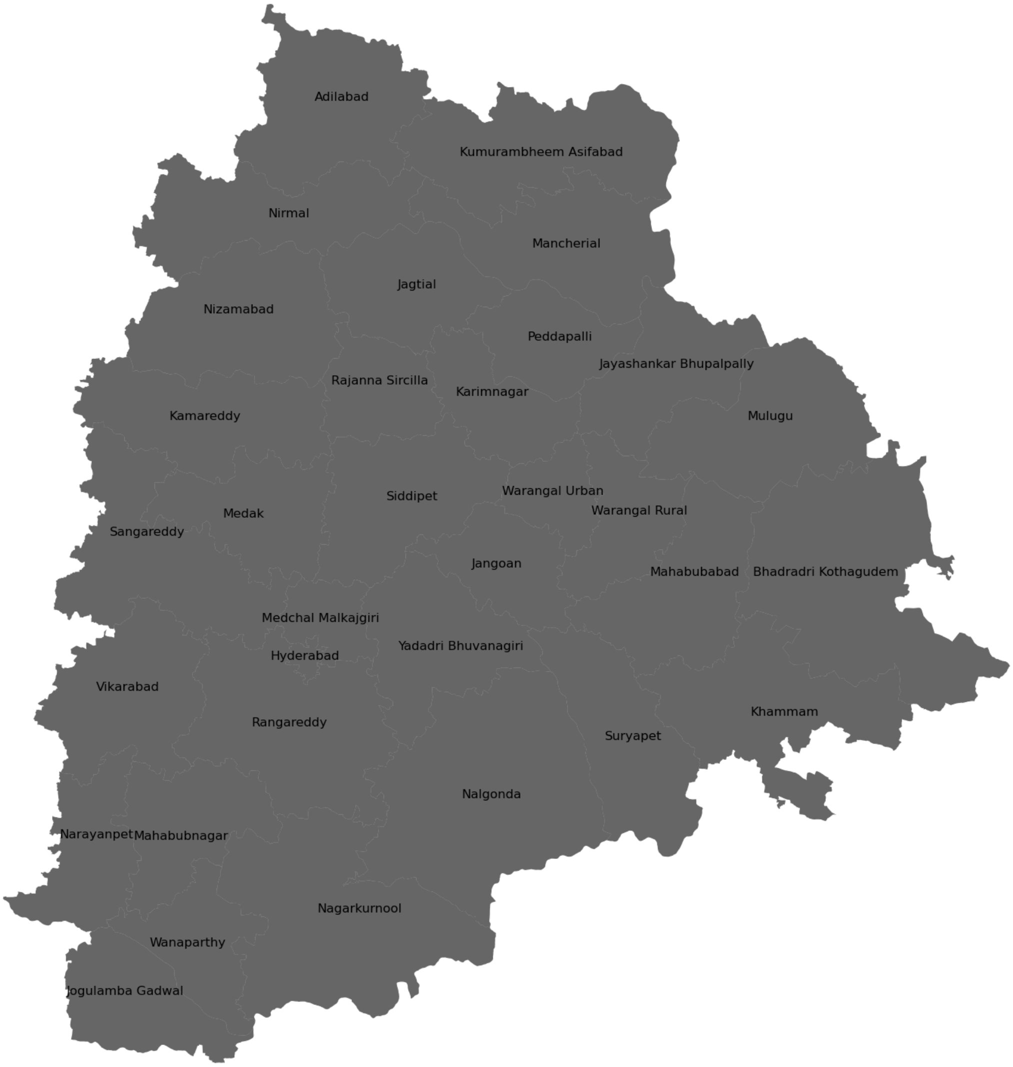
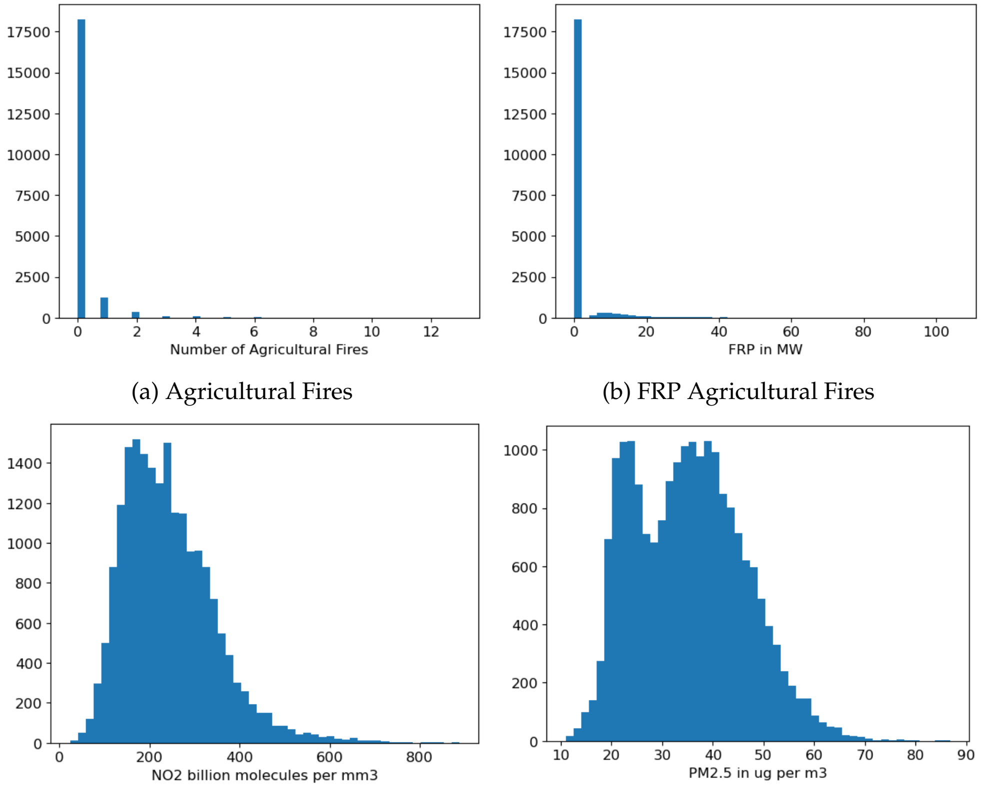


Figure 1: District Names in Telangana

Appendix B: Dependent Variable Exploration



(a) Agricultural Fires

(b) FRP Agricultural Fires

(c) NO_2

(d) $PM_{2.5}$

Figure 1: Distributions of the Dependent Variables

Appendix C: Independent Variables Exploration

Environmental Independent Variables

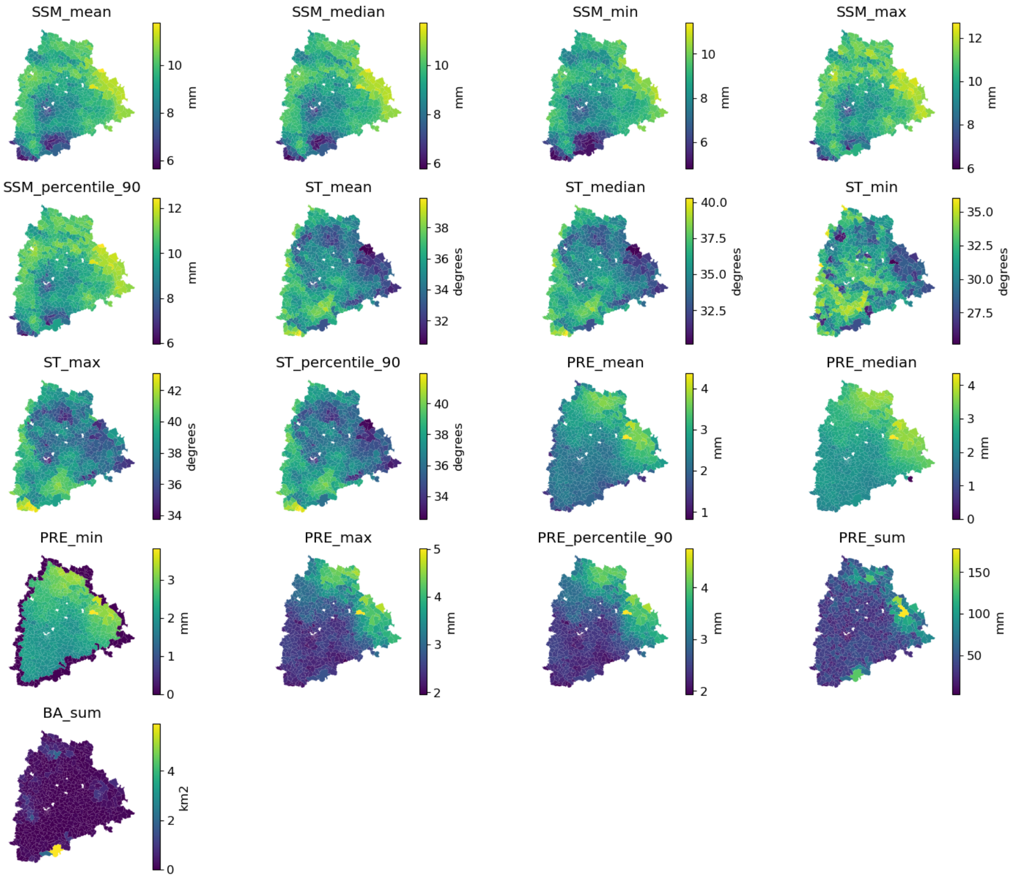


Figure 1: Numerical Environmental Variables Geoplots Telangana [8]

[8] Note that in the PRE_min geoplot we observe zero values on the edges of Telangana. This seems to be an error in the data. Fortunately, this error is not included in the prediction model because we determine that only PRE_max is a dominant predictor.

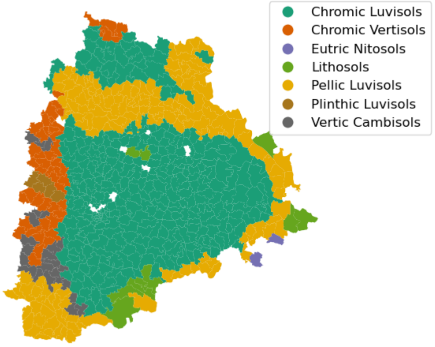


Figure 2: Soil Types Geoplot Telangana

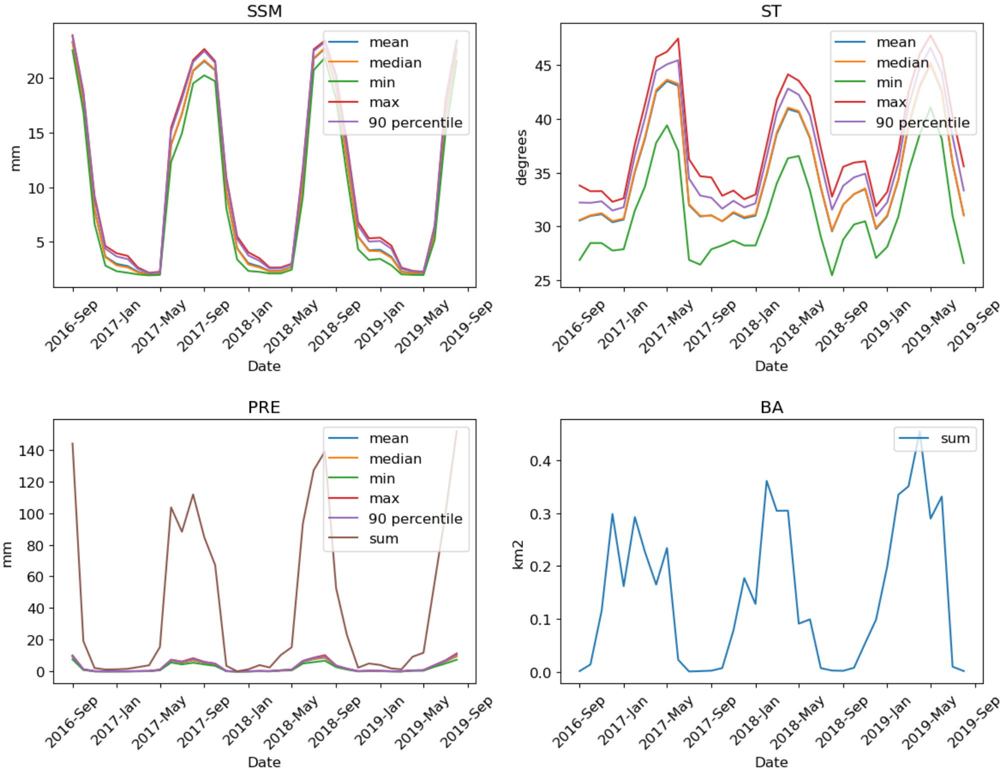


Figure 3: Numerical Environmental Variables over Time Telangana

Socio-Economic Independent Variables

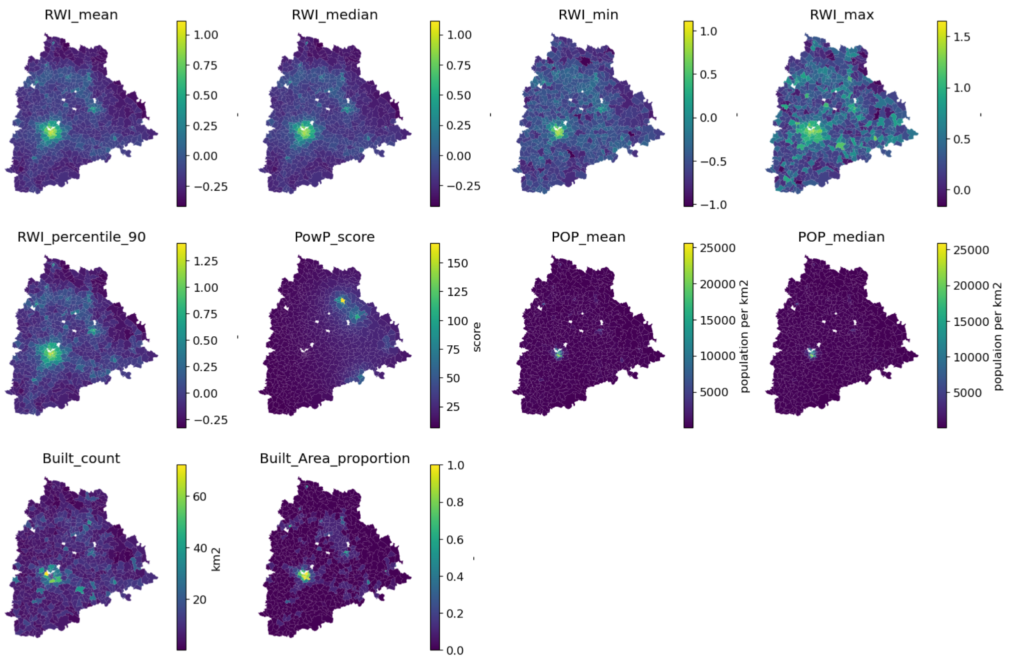


Figure 4: Numerical Socio-Economic Variables Geoplots Telangana[9]

[9] Note that we do not show socio-economic variables over time because these values do not vary per month. E.g. RWI is static, PowP_score is yearly, POP is yearly, and Built is yearly available.

Agricultural Independent Variables

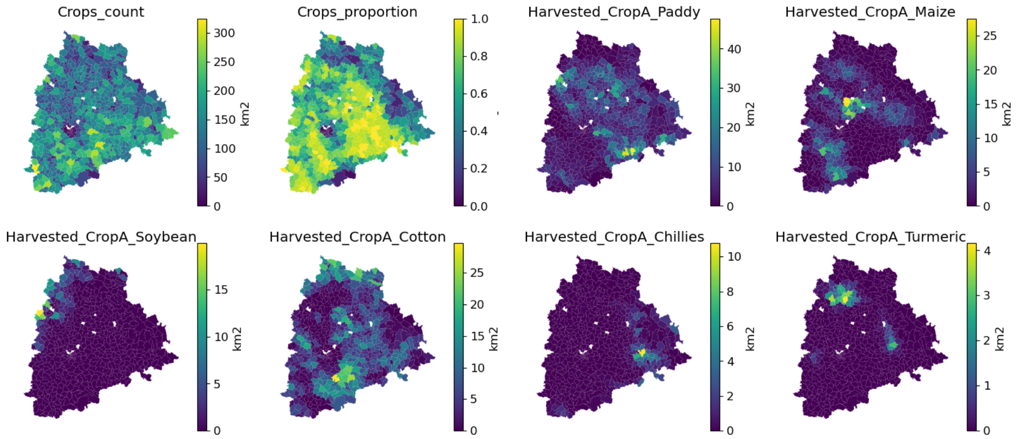


Figure 5: Numerical Agricultural Variables Geoplots Telangana

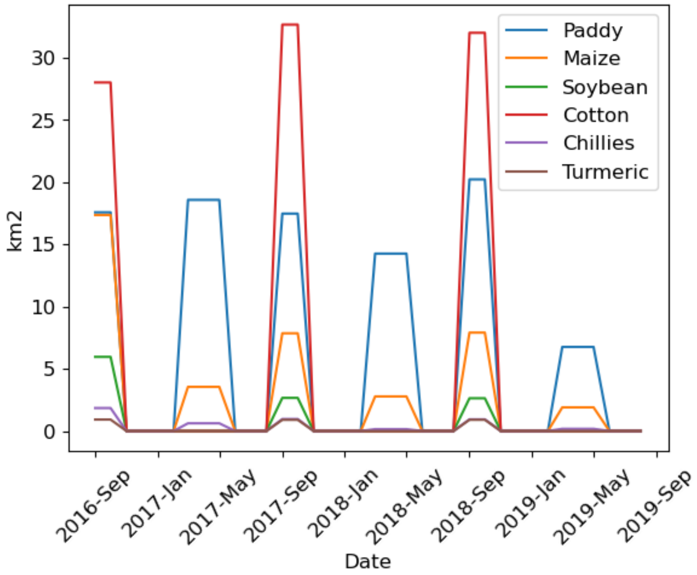
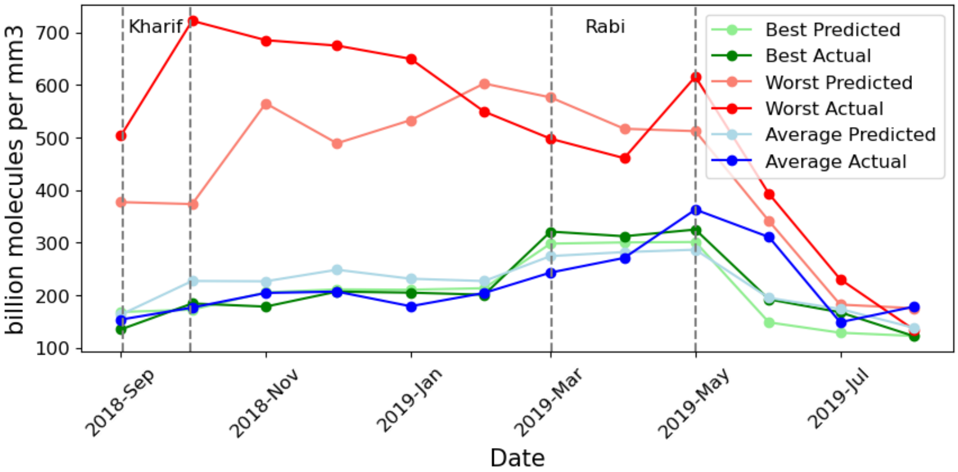
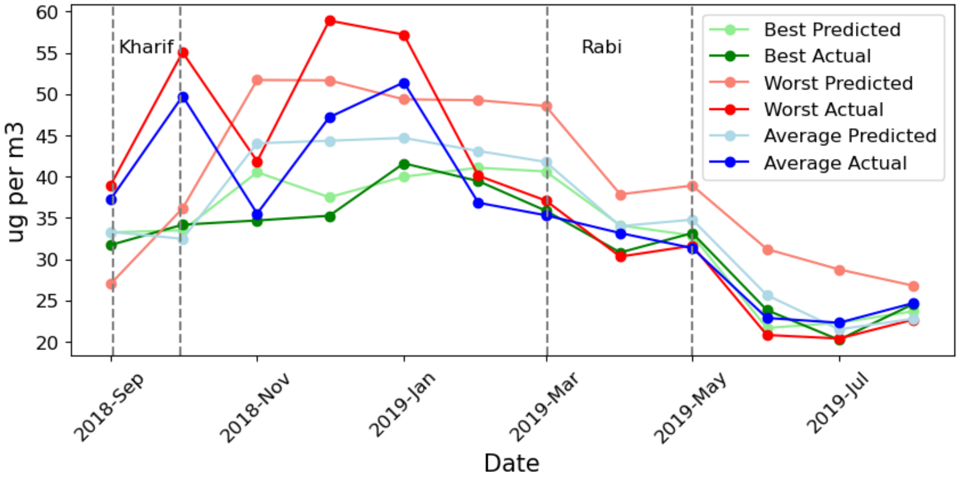


Figure 6: Average Crops Harvested over Time Telangana

Appendix D: Model Performances, Best, Average, and Worst Mandal



(a) Model outcomes per mandal NO_2: *best mandal*: Kodangal (MAE: 19.6), *worst mandal*: Anthergaon (MAE: 110.9), and *average mandal*: Valigonda (MAE: 41.5)



(b) Model outcomes per mandal $PM_{2.5}$: *best mandal*: Nallabelli (MAE: 2.3), *worst mandal*: Shaikpet (MAE: 9.5), and *average mandal*: Choutuppal (MAE: 5.2)

Figure 1: This figure shows the values of the best, worst, and average mandal. In Figure 1a this is shown for the NO_2 prediction model and in Figure 1b this is shown for the $PM_{2.5}$ prediction model. Note that the harvest periods are marked with grey dotted lines.