Optimal Targeting of Interventions to Reduce Air Pollution

Martin Kosík*

3rd November 2022

Abstract

I plan to apply an air pollution transport model (HYSPLIT) to compute the average population exposure to pollution caused by crop residue burning across different locations in northern India. Using these results together with satellite data on cropland fires, I propose how to optimally target interventions to reduce crop residue burning in order to minimize the total harm caused by the created air pollution. My very preliminary results suggest that there could potentially be meaningful efficiency gains from optimal targeting.

*CERGE-EI, Politických vězňů 7, 111 21 Prague, Czech Republic
Email: martin.kosik@cerge-ei.cz

1 Introduction

Biomass burning is responsible for up to 15.4% of infant deaths at the country level due to the air pollution it creates and this share has been on the rise globally from 2004 to 2018 (Pullabhotla et al., 2022). In addition to the heath effects, adverse impacts of air pollution on productivity (Chang et al., 2016) and human capital (Graff Zivin et al., 2020) have been documented.

Many of the biomass fires are intentionally caused by humans. First, fire is often used to destroy forests in order to transform the land for agriculture (Balboni et al., 2021) . Second, crop residue burning (which is the main focus of this paper) is practiced by some farmers to quickly clear-up the land after a harvest and prepare it for the next sowing (Abdurrahman et al., 2020). The problem with crop residue burning is especially severe in northern India where it has been an important contributor to air pollution during the late fall months (Kulkarni et al., 2020). There has been various interventions proposed to incentivize farmers not to burn their land (Abdurrahman et al., 2020). Jayachandran et al. (2019) conducted a RCT to evaluate a simple but potentially highly scalable intervention: offering the farmers a payment conditional on not burning their land.

However, with crop residue burning being fairly widespread (which is the case for for northern India, especially Punjab), a policymaker might not be able to implement a desired intervention in all locations. In this paper, I plan to calculate how the policymaker should target the interventions aimed to reduce crop residue burning across different locations to achieve the greatest reduction in harm with the limited budget. The local weather patterns and unequal distribution of population in space imply that the number of people exposed to the pollution might vary across the source locations. I will use atmospheric models of air pollution transport to compute average exposure to smoke emitted from

different locations in northern India. Combining these results with population density maps enables us to calculate the average number people affected by pollution depending on the source location. Second, I will use a structural model of farmers decisions on crop residue burning using the results of the RCT by Jayachandran et al. (2019), in which the farmers received a payment conditional on not burning their land, for calibration of key parameters. Together with satellite data on fires on agricultural land, this model will give us estimates of the expected reduction of stubble burning achieved by an intervention in a given location (which could be village or other unit of aggregation depending on the context).

2 Literature review

The impact of air pollution on economic, health, psychological, and political outcomes has been studied extensively. Specifically, particulate matter pollution[1] ($PM_{2.5}$ and PM_{10}) have been shown to have adverse effects on health (Deryugina et al., 2019), productivity (Chang et al., 2016), human capital (Graff Zivin et al., 2020), and academic achievement (Gilraine and Zheng, 2022).

The issue of crop residue burning in particular have also received increased attention in recent years (see Abdurrahman et al., 2020 for a review). In atmospheric sciences, the main focus of the work has been on quantifying the contribution of crop residue burning to overall air pollution (Nair et al., 2020) and studying the chemical, physical and optical properties of the aerosols (Mishra and Shibata, 2012; Ram et al., 2016). In economics and public health, the impact of stubble burning on various outcomes has been examined. The air

[1]Particulates (particulate matter) are microscopic particles suspended in the air (Seinfeld and Pandis, 2006, p. 97). PM_{10} to particles that have a diameter smaller or equal to 10 μm whereas $PM_{2.5}$ refers to particles with a diameter smaller or equal to 2.5 μm. These particles are inhalable and can enter the blood stream and brain (especially $PM_{2.5}$ since they tend to be not filtered out in lungs due to their small size) which can cause serious health problems.

pollution caused by agricultural fires has been shown to increase both infant and older-age mortality (He et al., 2020; Pullabhotla et al., 2022) and decrease birthweight, gestational length, and in utero survival (Rangel and Vogl, 2019). Graff Zivin et al. (2020) report that crop residue burning reduces performance at collage entrance exams in China. Finally, economic literature on potential solutions is much more nascent. Most notably, Jayachandran et al. (2019) conducted a RCT in Punjab, India in which payments were made to paddy farmers conditional on them not burning their fields. I describe the RCT in greater detail in subsection 4.1 as its results will be used as an input for my analysis.

In contrast to the studies mentioned above, this paper is focused on optimal targeting of interventions to reduce crop-residue burning, which relate it more to the literature on optimal policies in environmental economics. Assunção et al. (2019) estimate ex-post optimal assignment to "Priority Lists" which were municipalities in Brazil subject to more intense environmental monitoring and enforcement to combat deforestation. The policy of Assunção et al. (2019) aims to minimize total deforestation but in contrast to this paper they do not consider the effect of air pollution and heterogeneity in its dispersion.

3 Modelling air pollution transport

I will use HYSPLIT average dispersion (HyADS) approach introduced by Henneman et al. (2019) for modelling the spatial impact of air pollution. The HyADS approach is based on averaging the pollution concentrations predicted by the HYSPLIT[2] model, which is an air pollution transport model developed and maintained by NOAA (Draxler and Hess, 1998; Stein et al., 2015) that computes air parcel[3] trajectories to determine the dispersion of pollution. In contrast to

[2]The acronym HYSPLIT stands for Hybrid Single-Particle Lagrangian Integrated Trajectory.

[3]An air parcel is an imaginary body of air which can be assigned basic dynamic and thermodynamic properties of atmospheric air.

Chemical transport models (CTMs), HYSPLIT abstracts away from simulating complex chemical processes in the atmosphere. While this might lead to less accurate predictions, it also makes HYSPLIT substantially less computationally demanding. HYSPLIT model has been extensively used in the atmospheric sciences. Its applications include tracking and forecasting the release of wildfire smoke (Rolph et al., 2009), wind-blown dust (Ashrafi et al., 2014), volcanic ash (Stunder et al., 2007), and crop residue burning (Liu et al., 2018). Henneman et al. (2021) shows that HyADS have normalized mean errors between 20 and 28% in comparison with the CTMs predictions of PM2.5 source impacts of coal power plants in the US.

For illustration, figure A1 shows the predicted concentrations by HYSPLIT for an emissions event in which 2500 air parcels are released on October 6, 2019 at 9:00 AM. We can see that the wind direction strongly influences the dispersion of the pollution (in this case dispersing it in the south-east direction).

Main output of interest for our analysis is the source-receptor matrix (denoted as SRM_{ij}), which describes the concentration of air pollution (in mass units per m^3) in location j for pollutant emitted from source location i under average weather conditions. Average weather conditions in this context means those that are typical when crop residue burning occurs, which in northern India tends to be October and November (for post-monsoon crop residue burning which is the main focus of this paper). To obtain the source-receptor matrix for a typical crop residue burning event, I simply run the simulations for many emission events with different starting dates and then average the results. An emission events consists of releasing the same amount of emissions at the same time from many source locations (for my preliminary results I use 121 source locations) spread over a regular grid. The dispersion of the pollutants is then computed with the weather data for the given time period while keeping track

of the source location of the pollutants so that the source-receptor matrix can be estimated. The technical details of the HYSPLIT simulations used for the preliminary results are described in subsection A.1 in the appendix. Naturally, many simulations need to be run for an accurate estimate of the source-receptor matrix which underscores the computational demands of this project.

Once we have the source-receptor matrix SRM_{ij}, we can estimate the impact of emitting pollution from different locations. Let E_i be the total air pollution emitted from location i and P_i be the total air pollution concentration in i. By definition, we have that $P_j = \sum_i SRM_{ij}E_i$, i.e., the total pollution concentration in j is a sum of the emission from all locations i weighted by the corresponding source-receptor matrix entry. Let us further define $L_j = f(P_j)$ to be loss (harm) to a single person from being exposed to air pollution of concentration P_j. The total loss TL then simply is the sum of the losses across all locations weighted by their population N_j, i.e., $TL = \sum_j L_j \cdot N_j$.[4]

The impact of small change in emissions from i on total loss can be expressed as

$$\frac{\partial TL}{\partial E_i} = \sum_j \frac{\partial L_j}{\partial E_i} N_j = \sum_j \frac{\partial f(P_j)}{\partial P_j} \frac{\partial P_j}{\partial E_i} N_j = \sum_j \frac{\partial f(P_j)}{\partial P_j} SRM_{ij} N_j.$$

Clearly, if a given intervention can achieve the same reduction in emissions at the same costs in all locations then it is optimal to target the locations with the highest $\frac{\partial TL}{\partial E_i}$. Of course, this assumption of uniform effects is not realistic and therefore I will introduce the the full model that takes into account the heterogeneity of costs across locations in section 4.

We can also consider an even simpler case when the loss function is linear

[4]If there are heterogeneous effects of air pollution across individuals (due to e.g., richer households having financial resources to invest into air purifiers), it might be desirable to take these into account when aggregating the loss function. I will consider this in the future work. Nevertheless, the existing studies (Heft-Neal et al., 2018; Heft-Neal et al., 2020) do not show large differences in the effects sizes by wealth levels.

$(f(P) = \psi_0 + \psi \cdot P)$. Then $\frac{\partial TL}{\partial E_i}$ simplifies to

$$\frac{\partial TL}{\partial E_i} = \psi \sum_j SRM_{ij} N_j = \psi \cdot \alpha_i,$$

where the term $\alpha_i \equiv \sum_j SRM_{ij} N_j$ is sometimes referred to as the source impact of i in the literature (Henneman et al., 2021). We can see that under those circumstances, the source-receptor matrix and population in each location become sufficient for determining the optimal allocation. This extreme simplicity allows me to present some very preliminary estimates of the distribution of α_i in section 5.

There is some evidence to support linear effects of $PM_{2.5}$ concentrations at least in the case of infant mortality (at least in the policy relevant ranges). In particular, Heft-Neal et al. (2018) shows that higher order polynomials for the effect of post-birth exposure to $PM_{2.5}$ concentrations on the infant mortality are not statistically significant implying linear response function (nevertheless they do find that the quadratic term for in utero exposure is statistically significant).

4 Model

4.1 Farmers responses to intervention

Now, I proceed to modelling the decision of farmers to understand how they would respond to interventions incentivizing them not to burn their land. The main purpose of this model is to obtain reasonable counterfactual estimates of the share of burned land with and without an intervention for every location. I will use the results of the RCT by Jayachandran et al. (2019) which involved offering farmers a payment conditional on not burning their land to estimate the demand for burning using a discrete choice model (in a spirit of Souza-Rodrigues, 2019).

Consider a farmer with a plot of land p used for winter cropping in a location i (which could be a village or a square on a grid). There is a continuum of such plots in every location and a farmer makes a separate decision whether or not to clear the crop residues by burning them. Let Π^B be the value that farmer attains by clearing the land by burning and and Π^N be the value of the alternative (not burning). I will assume the following functional form

$$\Pi^B_{pi} - \Pi^N_{pi} = \beta s_i + x_i'\gamma + \epsilon_{pi}$$

where s_i is amount of payment offered conditional on not burning and x_i is a vector of location-level characteristics including soil quality, share of farmers with tractors (obtained from 2011 Socioeconomic and caste census), and the share of land burnt in a previous years. Finally, ϵ_{pi} is a idiosyncratic shock that captures unobserved plot-level factors and is assumed to follow the type 1 extreme value distribution. A plot will be cleared by burning (which I will denote by an indicator variable B_{pi}) if the benefits exceed the costs, i.e., $B_{pi} = \mathbb{1}\left(\Pi^B_{pi} > \Pi^N_{pi}\right)$. It follows from the above that the share of land cleared by burning in location in i, denoted as b_i, can be expressed as

$$\log\left(\frac{b_i}{1-b_i}\right) = \beta s_i + x_i'\gamma \tag{1}$$

The coefficients in this equation can be estimated using OLS as has been typically done in the literature (e.g., Pfaff, 1999; Souza-Rodrigues, 2019). For this, we will use the data from Jayachandran et al. (2019) who conducted a RCT in 171 villages in Punjab, India, in which they offered farmers varying amounts of payments conditional on farmers not burning their fields. In addition to conditional payment only, Jayachandran et al. (2019) also included a treatment variant in which the farmers received upfront (unconditional) payment upon

accepting the contract to help alleviate potential liquidity constraints and build trust. In this paper I do not model the difference between these two treatment variants but it is an interesting area for future research. The full data and results from this RCT are not yet available (as of October 2022) but the authors mention that "contracts with partial upfront payment reduce burning by 8-11 percentage points"[5] which suggests that the intervention had a meaningful effect.

Besides the treatment effect, the enrollment into the program and the nature of the self-selection are important determinants of the cost-effectiveness of an intervention (Jack and Jayachandran, 2019). The higher the enrollment of farmers who would never burn their fields, the lower the per enrollee benefits of the program are. In the extreme case, if the vast majority of farmers who enroll in the program are those that would refrain from burning their fields even in the absence of the intervention, then the cost-effectiveness of the program would be very small.

To parsimoniously capture this effect, the model has only two groups that differ in their enrollment rate. First, the plots that would be burned in the absence of the program have enrollment rate $\omega^B = P(R_{ip} = 1|B_{pi} = 1, s_i = 0)$ where R_{ip} is an indicator for enrollment and $s_i = 0$ corresponds to absence of the program in the location i. Second, the plots that would not be burned in the absence of the program have enrollment rate $\omega^N = P(R_{ip} = 1|B_{pi} = 0, s_i = 0)$. Note that $\omega^B > \omega^N$ implies that there is a positive correlation between the cost of enrollment and the cost of reducing burning and hence lower cost-effectiveness. It simply follows from the above that the location-level enrollment rate, r_i, can then be expresses as

$$r_i = \omega^B b(s_i = 0, x_i) + \omega^N \left(1 - b(s_i = 0, x_i)\right)$$

[5] https://jrc.princeton.edu/events/Jayachandran-F22

We can estimate ω^B and ω^N from experimental microdata in Jayachandran et al. (2019) using a method inspired by Jack and Jayachandran (2019). We first use plot-level data from the control group to fit a flexible logit model to obtain estimates of the probability of burning conditional on pre-treatment covariates $\hat{b}_{pi}$ (i.e., the propensity to burn). Due to the random assignment of treatment with respect to location, we should expect the the distribution of the propensities to burn in the control and treatment locations to be the same (ignoring the sampling error). The comparison of the density of $\hat{b}_{pi}$ in the control locations to the corresponding density of those enrolled in the program in the treated locations allows us identify ω^B and ω^N. Notice that for any convex interval $\mathcal{B}$

$$\begin{aligned} P\left(b_{pi} \in \mathcal{B} | R_{pi} = 1, s_i \neq 0\right) =& \omega^B \, \mathbb{E}\left[b_{pi} | b_{pi} \in \mathcal{B}, s_i = 0\right] P\left(b_{pi} \in \mathcal{B} | s_i = 0\right) \\ &+ \omega^N \left(1 - \mathbb{E}\left[b_{pi} | b_{pi} \in \mathcal{B}, s_i = 0\right]\right) P\left(b_{pi} \in \mathcal{B} | s_i = 0\right) \end{aligned}$$

All of the terms above (except ω^B and ω^N) can be estimated from the data. $P\left(b_{pi} \in \mathcal{B} | R_{pi} = 1, s_i \neq 0\right)$ is the density of propensities to burn in a given interval for plots enrolled in the program, $P\left(b_{pi} \in \mathcal{B} | s_i = 0\right)$ is the corresponding density for all plots in the control, and $\mathbb{E}\left[b_{pi} | b_{pi} \in \mathcal{B}, s_i = 0\right]$ is the expected propensity to burn in the given interval in the control. Choosing only two disjoint intervals $\mathcal{B}$ is sufficient for identification of ω^B and ω^N, as it leads to a system of 2 linear equations with 2 unknowns. However, with a finite number of observations, the estimates in the equations are subject to sampling error and therefore it might be preferable to choose a higher number partitions and find the solution that minimizes the sum of least squared errors.

4.2 Problem formulation

We finally proceed to formulating the problem. I will consider a policymaker with a budget M who chooses the level of the conditional payments s_i offered to farmers for not burning their fields in each location i to minimize the total population-weighted loss caused by air pollution. In my formulation mainly for computational reasons, the policymaker is choosing from a finite set of payment levels, $s_i \in \{\bar{s}_1, \ldots, \bar{s}_J\}$ expressed as money units per hectare.[6] Naturally, no intervention (i.e., $\bar{s}_1 = 0$) is always included in this set.

The policymaker minimizes the total loss (TL), which is a sum of losses (L_j) for all locations j weighted by their population N_j

$$\min_{s_i \in \{\bar{s}_1, \ldots, \bar{s}_J\}} \text{TL} = \min_{s_i \in \{\bar{s}_1, \ldots, \bar{s}_J\}} \sum_j L_j N_j, \tag{2}$$

subject to the budget constraint(where l_i is the total area of eligible land in i and r_i is the enrollment rate in the program, and F are the fixed costs of implementing the intervention)

$$\sum_i r_i s_i l_i + \mathbb{1}(s_i > 0) F \leq M, \tag{3}$$

the pollution loss function which specifies the harm by air pollution concentration P_j (which is a sum of pollution concentration due to crop-residue burning p_j^b and other sources p_j^0)

$$L_j = f(P_j) = f(p_j^b + p_j^0), \tag{4}$$

[6] In the RCT of Jayachandran et al. (2019), the levels of the payments were $\{\bar{s}_1, \bar{s}_2, \bar{s}_3\} = \{0, 800, 1600\}$ denominated in INR per acre. Hence if a policymaker does not want to extrapolate the effect of the intervention beyond those actually implemented in the experiment, he or she should only consider this set of values.

the equation for source-receptor matrix decomposition of air pollution

$$p_j^b = \sum_i SRM_{ij} E_i, \tag{5}$$

the equation relating the emissions due to crop residue burning (E_i) to the predicted share of land burned ($b(s_i, x_i)$) and eligible land area l_i

$$E_i = \phi b(s_i, x_i) \cdot l_i, \tag{6}$$

the predicted share of land burned given the conditional payment amount (s_i) and the covariates (x_i)

$$b(s_i, x_i) = \frac{\exp(\beta s_i + x_i'\gamma)}{1 + \exp(\beta s_i + x_i'\gamma)}, \tag{7}$$

and the equation for enrollment rate into the program (discussed in subsection 4.1)

$$r_i = \omega^B b(s = 0, x_i) + \omega^N (1 - b(s = 0, x_i)). \tag{8}$$

There are several key parameters and functional form assumptions that need to be specified. First, the policymaker has to set the maximum budget size M. Second, the loss function for air pollution needs to be defined. One of the main concerns with regard to crop residue burning is its effect on mortality via increasing $PM_{2.5}$ concentrations. In my main analysis, I will therefore focus only on the effects on mortality.[7] Regarding the functional form, linear loss ($f(P) = \psi_0 + \psi \cdot P$) substantially reduces the computational complexity of the optimization and allows greater robustness in specifying some of the parameters (i.e., ψ_0, ψ, p_j^0, ϕ have no influence on the optimal allocation of s_i). Moreover,

[7] Other relevant effects of air pollution have been studied and documented (e.g., on productivity or human capital). Nevertheless, these effects tend to be more context specific and less precisely estimated than the impact on mortality and therefore I only consider mortality.

as discussed at the end of section 3, there is some evidence for a linear effect of $PM_{2.5}$ concentration on infant mortality (Heft-Neal et al., 2018). Nevertheless, alternative parametrizations for the effects on mortality have been proposed in the public health literature. In particular, Burnett et al. (2014) suggests the following function form for the effect of $PM_{2.5}$ concentration C on relative risk of mortality RR:

$$RR(C) = \begin{cases} 1 + \alpha \left[1 - \exp\left(-\gamma \left(C - C_0\right)^{\delta}\right)\right] & \text{for } C > C_0 \\ 1 & \text{for } C \leq C_0 \end{cases} \tag{9}$$

C_0 represents a minimum concentration above which there is evidence indicating health benefits of PM2.5 exposure reductions. This parametrization has been used by several other studies as well (e.g., Apte et al., 2015). Therefore I will consider both of these functional forms to assess the robustness of the final allocations (although I might solve the model with the non-linear loss only for a smaller number of aggregated locations if the computational demands are too high).

Third, the source-receptor matrix needs to be estimated. In this paper, I use the HyADS approach, which I described in greater details in section 3. Fourth, regarding the relationship between a unit of area burned (in our case hectars) and the mass of pollutant emitted (measured in grams), I will again rely on the existing literature in atmospheric science. In particular, I can apply decompositions used by N. Jain et al. (2014) and Liu et al. (2020) to express ϕ as

$$\phi = CY \times RC \times f_{DM} \times f_{CC} \times EF \tag{10}$$

where CY is the crop yield (produced weight per a unit of area), RC is the residue-to-crop weight ratio, f_{DM} is the dry matter fraction of the crop, f_{CC} is the combustion completeness (fraction of the dry matter burned), and EF is

the emission factor for the pollutant. Using the estimates of these parameters from the literature (provided in table A1) for rice paddy as the crop and $PM_{2.5}$ as the pollutant, we get $\phi \approx 23772$. Finally, I already described above how I would obtain the predictions $\hat{b}_i(s_i, x_i)$ in the previous subsection.

4.3 Solving the model

With potentially large number locations and a non-convex objective function, solving the model might be computationally challenging. However, in the case of linear loss and binary intervention (i.e., each location can either receive or not receive an intervention), we can achieve substantial speed-ups by reformulating the model as a knapsack problem[8] (this is demonstrated in subsection A.2 in the appendix). The knapsack problem is NP-Complete and the existing algorithms may, in the worst case, take exponential time (Kellerer et al., 2004, p. 491). However, there is a fully polynomial-time approximation scheme based on dynamic programming that achieves polynomial time in the number of items while controlling the desired approximation error (Kellerer et al., 2004, p. 37). Moreover, the knapsack solvers are efficiently implemented in OR-tools library by Google.[9] I was able to solve a toy knapsack problem with 10 000 items and rounding of inputs to 2 digits in about one second using the library.

Furthermore, even in the case of linear loss and finite discrete interventions (in which we can assign different levels of payment in each location chosen from a finite set of options), we can still attain significant efficiency gains since the model can be formulated as a multiple-choice knapsack problem. In the multiple-choice knapsack problem, the set of items is partitioned into classes and only one item within a set can be chosen (Kellerer et al., 2004, p. 317).

[8] The knapsack problem is a canonical problem in combinatorial optimization in which the goal is to determine which items (each with its weight and value) to include in order to maximize the value of the included items while not exceeding a given weight limit (the capacity constraint). For more details see Kellerer et al. (2004).

[9] `https://developers.google.com/optimization`

In our case the classes are different types of the interventions (including no intervention) and the constraint enforces that only one type of the intervention will be chosen in a given location. While the multiple-choice knapsack problem is NP-hard, there again exists a fully polynomial-time approximation scheme (Kellerer et al., 2004, p. 338).

Finally, the case of non-linear loss function poses even greater challenge. Nevertheless, there are several heuristic algorithms that have been successfully applied to various optimization problems to find approximate global maximum in a large search space most notably simulated annealing and genetic algorithms (Mitchell, 1998). Simulated annealing methods tend to explore the space very widely in early stages but tend to become more greedy in time (i.e., it tends to select points that are close to currently best solution). Genetic algorithms keep a whole population of solution which are probabilistically mutated, recombined, and discarded. In both cases, the constraint can be included as a penalty into the objective function. Furthermore, regardless of the optimization algorithm selected, the complexity of the problem can always be reduced simply by aggregating some of the locations together and thus reducing the size of the search space.

5 Preliminary results

To demonstrate technical feasibility of the project, I present here some very preliminary results. However, due to high computational demands of the atmospheric simulation, these results are based on only small number of emission events (40) and therefore are likely not very reliable. With more computational resource, I would add more emission events but also increase the number of source locations (which currently is 121) and the increase the resolution of the receptor grid. Furthermore, I could enlarge the region of study from only North

west India to whole South Asia and possible even apply the analysis to other regions where crop residue burning is a serious issue (e.g., China, Brazil, or Sub-Saharan Africa - see Pullabhotla et al., 2022)

First, figure A2 in the appendix plots α_i of every source location on grid according to their longitude and latitude. Note that the number of source location is small due to the necessity of running HYSPLIT simulations for each of them. To increase the resolution of the grid, I applied bilinear interpolation which is shown in figure A3. Finally, only a fraction of land in Northern India is used for growing crops in winter season and crop residue burning is not practiced universally. To account for this, I restrict my analysis to grids that are predicted to have winter cropping as per M. Jain et al. (2016) predictions (based on satellite data) and grids with aggregated total radiating power greater than 5 megawatts (using VIIRS Active Fire data for October 2019). The results for only this subset of locations with likely crop residue burning are plotted in figure A4. A histogram of α_i for only these locations is shown in figure A6. Note that the under linear loss from air pollution (e.g., linear increase in infant mortality with $PM_{2.5}$ concentration), a ratio of any α_i and α_j gives us how much higher reduction in loss is if we reduce a unit of emissions in location i compared to location j. For example, returning to the estimated distribution of α_i in our setting, we have that the ratio of the 9th to 1st deciles is $\frac{0.00192}{0.00058} \approx 3.34$. This implies that, keeping other factors equal, targeting the locations in the 9th decile would be more than 3 times effective in reducing the loss compared to targeting the 1st decile. Naturally, the benefits of targeting would depend on size of the budget. For example, consider a case in which a policy maker can implement the intervention in only 20 % of crop residue burning locations (again the costs of the intervention are uniform across the locations and the loss is linear). Then in our setting, optimal targeting (i.e., locations with highest α_i) would be 84%

more effective than reducing the same amount of emissions in a location with mean α_i.

In the future research, I plan to extend this analysis by relaxing the assumption of uniform costs of the intervention using a structural model of farmers decisions and results of the RCT by Jayachandran et al. (2019).

6 Conclusion

In this paper, I proposed how to optimally target interventions to reduce crop residue burning in order to minimize the total harm caused by the created air pollution. The very preliminary results I presented suggest that there could potentially be meaningful efficiency gains from targeting the interventions optimally, especially for relatively small budgets. Nevertheless, there remains much work to be done in future research. Firstly and most importantly, the full analysis proposed in this could not be executed since the results of the RCT of Jayachandran et al. (2019) were not yet published. Secondly, there are rather purely technical improvements that could strengthen the credibility of the results. This includes increasing the number of source locations, simulating more emission events and using meteorological data with greater resolution. Finally, there are possible new directions in which to extend this project could be extended. There are other sources of biomass burning such as forest fires and slash-and-burn agriculture, which are at least partially caused by human activity (Balboni et al., 2021). The modelling framework I developed could be applied with some modifications to these problems as well to better understand the costs and benefits of possible interventions in different regions of the world.

References

Abdurrahman, Muhammad Isa, Chaki, Sukalpaa and Saini, Gaurav (2020), 'Stubble burning: Effects on health & environment, regulations and management practices', *Environmental Advances* 2, p. 100011, DOI: 10.1016/j.envadv.2020.100011.

Akagi, S. K., Yokelson, R. J., Wiedinmyer, C., Alvarado, M. J., Reid, J. S., Karl, T., Crounse, J. D. and Wennberg, P. O. (2011), 'Emission factors for open and domestic biomass burning for use in atmospheric models', *Atmospheric Chemistry and Physics* 11 (9), pp. 4039–4072, DOI: 10.5194/acp-11-4039-2011.

Apte, Joshua S., Marshall, Julian D., Cohen, Aaron J. and Brauer, Michael (2015), 'Addressing Global Mortality from Ambient PM2.5', *Environmental Science & Technology* 49 (13), pp. 8057–8066, DOI: 10.1021/acs.est.5b01236.

Ashrafi, Khosro, Shafiepour-Motlagh, Majid, Aslemand, Alireza and Ghader, Sarmad (2014), 'Dust storm simulation over Iran using HYSPLIT', *Journal of Environmental Health Science and Engineering* 12 (1), p. 9, DOI: 10.1186/2052-336X-12-9.

Assunção, Juliano, McMillan, Robert, Murphy, Joshua and Souza-Rodrigues, Eduardo (2019), *Optimal environmental targeting in the amazon rainforest*, tech. rep., National Bureau of Economic Research.

Balboni, Clare, Burgess, Robin and Olken, Benjamin A. (2021), *The Origins and Control of Forest Fires in the Tropics*, tech. rep., Tech. Rep.

Burnett, Richard T., Pope, C. Arden, Ezzati, Majid, Olives, Casey, Lim, Stephen S., Mehta, Sumi, Shin, Hwashin H., Singh, Gitanjali, Hubbell, Bryan, Brauer,

Michael, Anderson, H. Ross, Smith, Kirk R., Balmes, John R., Bruce, Nigel G., Kan, Haidong, Laden, Francine, Pr, üss-Ustün Annette, Turner, Michelle C., Gapstur, Susan M., Diver, W. Ryan and Cohen, Aaron (2014), 'An Integrated Risk Function for Estimating the Global Burden of Disease Attributable to Ambient Fine Particulate Matter Exposure', *Environmental Health Perspectives* 122 (4), pp. 397–403, DOI: 10.1289/ehp.1307049.

Chang, Tom, Graff Zivin, Joshua, Gross, Tal and Neidell, Matthew (2016), *The Effect of Pollution on Worker Productivity: Evidence from Call-Center Workers in China*, Working Paper 22328, Series: Working Paper Series, National Bureau of Economic Research, DOI: 10.3386/w22328.

Deryugina, Tatyana, Heutel, Garth, Miller, Nolan H., Molitor, David and Reif, Julian (2019), 'The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction', *American Economic Review* 109 (12), pp. 4178–4219, DOI: 10.1257/aer.20180279.

Draxler, Roland R. and Hess, G. D. (1998), 'An overview of the HYSPLIT_4 modelling system for trajectories', *Australian meteorological magazine* 47 (4), pp. 295–308.

Gilraine, Michael and Zheng, Angela (2022), *Air Pollution and Student Performance in the U.S.* Working Paper 30061, Series: Working Paper Series, National Bureau of Economic Research, DOI: 10.3386/w30061.

GOI, DAC (2018), 'Agricultural Statistics at a glance 2018', *Government of India, Ministry of Agriculture & Farmers Welfare, Department of Agriculture, Cooperation & Farmers Welfare, Directorate of Economics and Statistics.*

Graff Zivin, Joshua, Liu, Tong, Song, Yingquan, Tang, Qu and Zhang, Peng (2020), 'The unintended impacts of agricultural fires: Human capital in

China', *Journal of Development Economics* 147, p. 102560, DOI: 10.1016/j.jdeveco.2020.102560.

He, Guojun, Liu, Tong and Zhou, Maigeng (2020), 'Straw burning, PM2.5, and death: Evidence from China', *Journal of Development Economics* 145, p. 102468, DOI: 10.1016/j.jdeveco.2020.102468.

Heft-Neal, Sam, Burney, Jennifer, Bendavid, Eran and Burke, Marshall (2018), 'Robust relationship between air quality and infant mortality in Africa', *Nature* 559 (7713), pp. 254–258, DOI: 10.1038/s41586-018-0263-3.

Heft-Neal, Sam, Burney, Jennifer, Bendavid, Eran, Voss, Kara K. and Burke, Marshall (2020), 'Dust pollution from the Sahara and African infant mortality', *Nature Sustainability* 3 (10), pp. 863–871, DOI: 10.1038/s41893-020-0562-1.

Henneman, Lucas R. F., Choirat, Christine, Ivey, Cesunica, Cummiskey, Kevin and Zigler, Corwin M. (2019), 'Characterizing population exposure to coal emissions sources in the United States using the HyADS model', *Atmospheric Environment* 203, pp. 271–280, DOI: 10.1016/j.atmosenv.2019.01.043.

Henneman, Lucas R. F., Dedoussi, Irene C., Casey, Joan A., Choirat, Christine, Barrett, Steven R. H. and Zigler, Corwin M. (2021), 'Comparisons of simple and complex methods for quantifying exposure to individual point source air pollution emissions', *Journal of Exposure Science & Environmental Epidemiology* 31 (4), pp. 654–663, DOI: 10.1038/s41370-020-0219-1.

Jack, B. Kelsey and Jayachandran, Seema (2019), 'Self-selection into payments for ecosystem services programs', *Proceedings of the National Academy of Sciences* 116 (12), pp. 5326–5333, DOI: 10.1073/pnas.1802868115.

Jain, M., Mondal, P., Galford, G.L., Fiske, G. and DeFries, R.S. (2016), *India Annual Winter Cropped Area, 2001-2016*, Type: dataset, DOI: 10.7927/H47D2S3W.

Jain, Niveta, Bhatia, Arti and Pathak, Himanshu (2014), 'Emission of air pollutants from crop residue burning in India', *Aerosol and Air Quality Research* 14 (1), pp. 422–430.

Jayachandran, Seema, Kala, Namrata, Pande, Rohini, Jack, Kelsey and Rowe, Caitlin (2019), *Paying Farmers Not to Burn: A Randomized Trial of Payments for Ecosystem Services in India*, tech. rep., DOI: 10.1257/rct.4508.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, Roy and Joseph, Dennis (1996), 'The NCEP/NCAR 40-Year Reanalysis Project', *Bulletin of the American Meteorological Society* 77 (3), Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, pp. 437–472, DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kellerer, Hans, Pferschy, Ulrich and Pisinger, David (2004), *Knapsack Problems*, Berlin: Springer.

Kulkarni, Santosh H., Ghude, Sachin D., Jena, Chinmay, Karumuri, Rama K., Sinha, Baerbel, Sinha, V., Kumar, Rajesh, Soni, V. K. and Khare, Manoj (2020), 'How Much Does Large-Scale Crop Residue Burning Affect the Air Quality in Delhi?', *Environmental Science & Technology* 54 (8), pp. 4790–4799, DOI: 10.1021/acs.est.0c00329.

Lasko, Kristofer and Vadrevu, Krishna (2018), 'Improved rice residue burning emissions estimates: Accounting for practice-specific emission factors in air

pollution assessments of Vietnam', *Environmental Pollution* 236, pp. 795–806, DOI: 10.1016/j.envpol.2018.01.098.

Liu, Tianjia, Marlier, Miriam E., DeFries, Ruth S., Westervelt, Daniel M., Xia, Karen R., Fiore, Arlene M., Mickley, Loretta J., Cusworth, Daniel H. and Milly, George (2018), 'Seasonal impact of regional outdoor biomass burning on air pollution in three Indian cities: Delhi, Bengaluru, and Pune', *Atmospheric Environment* 172, pp. 83–92, DOI: 10.1016/j.atmosenv.2017.10.024.

Liu, Tianjia, Mickley, Loretta J., Singh, Sukhwinder, Jain, Meha, DeFries, Ruth S. and Marlier, Miriam E. (2020), 'Crop residue burning practices across north India inferred from household survey data: Bridging gaps in satellite observations', *Atmospheric Environment: X* 8, p. 100091, DOI: 10.1016/j.aeaoa.2020.100091.

Mishra, Amit Kumar and Shibata, Takashi (2012), 'Synergistic analyses of optical and microphysical properties of agricultural crop residue burning aerosols over the Indo-Gangetic Basin (IGB)', *Atmospheric Environment* 57, pp. 205–218, DOI: 10.1016/j.atmosenv.2012.04.025.

Mitchell, Melanie (1998), *An Introduction to Genetic Algorithms*, Reprint edition, Cambridge, Mass.: MIT Press.

Nair, Moorthy, Bherwani, Hemant, Kumar, Suman, Gulia, Sunil, Goyal, Sanjeev and Kumar, Rakesh (2020), 'Assessment of contribution of agricultural residue burning on air quality of Delhi using remote sensing and modelling tools', *Atmospheric Environment* 230, p. 117504, DOI: 10.1016/j.atmosenv.2020.117504.

Pan, X. L., Kanaya, Y., Wang, Z. F., Komazaki, Y., Taketani, F., Akimoto, H. and Pochanart, P. (2013), 'Variations of carbonaceous aerosols from open

crop residue burning with transport and its implication to estimate their lifetimes', *Atmospheric Environment* 74, pp. 301–310, DOI: 10.1016/j.atmosenv.2013.03.048.

Pfaff, Alexander S. P. (1999), 'What Drives Deforestation in the Brazilian Amazon?: Evidence from Satellite and Socioeconomic Data', *Journal of Environmental Economics and Management* 37 (1), pp. 26–43, DOI: 10.1006/jeem.1998.1056.

Pullabhotla, Hemant Kumar, Zahid, Mustafa, Heft-Neal, Sam, Rathi, Vaibhav and Burke, Marshall (2022), 'Global biomass fires and infant mortality', Publisher: EarthArXiv.

Ram, Kirpa, Singh, Sunita, Sarin, M. M., Srivastava, A. K. and Tripathi, S. N. (2016), 'Variability in aerosol optical properties over an urban site, Kanpur, in the Indo-Gangetic Plain: A case study of haze and dust events', *Atmospheric Research* 174-175, pp. 52–61, DOI: 10.1016/j.atmosres.2016.01.014.

Rangel, Marcos A. and Vogl, Tom S. (2019), 'Agricultural Fires and Health at Birth', *The Review of Economics and Statistics* 101 (4), pp. 616–630, DOI: 10.1162/rest_a_00806.

Rolph, Glenn D., Draxler, Roland R., Stein, Ariel F., Taylor, Albion, Ruminski, Mark G., Kondragunta, Shobha, Zeng, Jian, Huang, Ho-Chun, Manikin, Geoffrey, McQueen, Jeffery T. and Davidson, Paula M. (2009), 'Description and Verification of the NOAA Smoke Forecasting System: The 2007 Fire Season', *Weather and Forecasting* 24 (2), pp. 361–378, DOI: 10.1175/2008WAF2222165.1.

Seinfeld, John H. and Pandis, Spyros N. (2006), *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, 2nd edition, Hoboken, N.J: Wiley-Interscience.

Souza-Rodrigues, Eduardo (2019), 'Deforestation in the Amazon: A Unified Framework for Estimation and Policy Analysis', *The Review of Economic Studies* 86 (6), pp. 2713–2744, DOI: 10.1093/restud/rdy070.

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D. and Ngan, F. (2015), 'NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System', *Bulletin of the American Meteorological Society* 96 (12), Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, pp. 2059–2077, DOI: 10.1175/BAMS-D-14-00110.1.

Stunder, Barbara J. B., Heffter, Jerome L. and Draxler, Roland R. (2007), 'Airborne Volcanic Ash Forecast Area Reliability', *Weather and Forecasting* 22 (5), pp. 1132–1139, DOI: 10.1175/WAF1042.1.

A Appendix

A.1 Description of HYSPLIT simulations

In general, I tend to follow the setup and the parameters of Henneman et al. (2019), nevertheless I deviate from their approach in certain aspects, which I will describe below, due the differences in goals (Henneman et al., 2019 aim to estimate exposure to actual historical pollution from coal power plants whereas I am interested in counterfactual exposure) and context (coal power plants in the US vs. crop residue burning in India).

As mentioned in section 3, I estimate the source-receptor matrix by averaging the source-receptor concentration estimates across emission events. An emission event proceed as follows: An unit mass of pollutant is released from each source at the height 7.5 meters. The 121 source locations were spread over a regular rectangular grid with latitude ranging from 40° to 45° and longitude from 78° to 70°. The HYSPLIT model then tracks the dispersion of 2500 air parcels for each sources (this number was chosen to balance the fidelity and computational demands of the simulations) for 4 full days (96 hours). The emission event length of 4 days was chosen based on the results of Pan et al. (2013) who measured that the approximate atmospheric lifetime of carbonaceous aerosols from crop residue burning is 1 to 6 days. The source-specific concentrations were calculated for each location on a grid with a resolution 0.05° of latitude and longitude. The concentration grids were averaged over the emission events to produce a single grid of the mean concentrations for each source location. These resulting grids formed our final estimate of the source-receptor matrix.

The start dates for the emission events were either 9 AM on October 1 or 9 AM on October 20 both for years 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2015, 2016, and 2017 (hence 20 emission events in total). I used the NCEP reanalysis (Kalnay et al., 1996) for historical meteorological data mainly due to

its global coverage, long temporal coverage, and decent resolution.

A.2 Reformulation to knapsack problem

In case of linear loss function and binary intervention, we can reformulate the model as a canonical knapsack problem. First, let us first denote

$$b_i^0 = \hat{b}_i(s_i = 0, x_i)$$

and

$$b_i^1 = \hat{b}_i(s_i = \bar{s}^T, x_i)$$

therefore we can express $b_i(s_i, x_i)$ as

$$b_i(s_i, x_i) = b_i^1 \cdot \mathbb{1}\left(s_i = s^T\right) + b_i^0.$$

By plugging this into the objective function and dropping the linear terms that do not depend on s_i and all the scaling terms, we get

$$\underset{s_i \in \{0, \bar{s}^T\}}{\arg\min} = \sum_j \sum_i N_j SRM_{ij} l_i b_i^1 \cdot \mathbb{1}\left(s_i = s^T\right)$$

Since maximization of the negative of a function is equivalent to minimzation of the original function, we can write

$$\underset{s_i \in \{0, \bar{s}^T\}}{\arg\max} = \sum_i \sum_j (-1) N_j SRM_{ij} l_i b_i^1 \cdot \mathbb{1}\left(s_i = s^T\right)$$

If we denote the values to be $v_i = \sum_j (-1) N_j SRM_{ij} l_i b_i^1$, the weights to be $w_i = \bar{s}^T l_i$ and binary treatment to be $t_i = \mathbb{1}\left(s_i = s^T\right)$, we arrive at the canonical

formulation of the knapsack problem:

$$\underset{s_i \in \{0, \bar{s}^T\}}{\arg\max} = \sum_i v_i \cdot t_i$$

subject to

$$\sum_i w_i \cdot t_i \leq M$$

For linear loss and finite discrete interventions, it can be easily shown that the multiple knapsack formulation can be obtained by simply including the additional interventions as new items and by imposing new constraints requiring only up to one intervention to be chosen in each location.

Table A1: Emission conversion parameters for rice paddy and $PM_{2.5}$

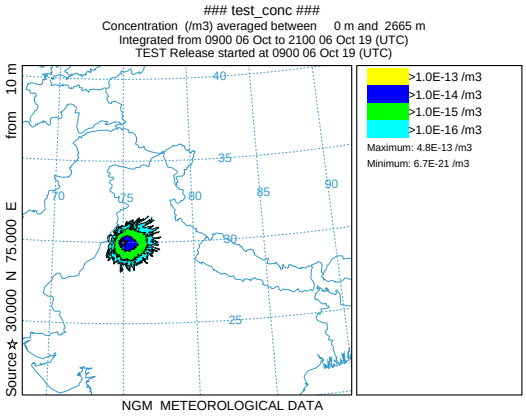
Parameter	Description	Units	Value	Source
CY	Crop yield (production per area)	$\frac{kg}{ha}$	3774	GOI (2018, p. 150) [10]
RC	Residue-to-crop ratio	unitless	1.5	N. Jain et al. (2014)
f_{DM}	Dry matter(DM) fraction of the crop	unitless	0.86	N. Jain et al. (2014)
f_{CC}	Combustion completeness	unitless	0.78	Lasko and Vadrevu (2018) [11]
EF	Emission factor for $PM_{2.5}$	$\frac{g}{kg}$	6.26	Akagi et al. (2011)

[10] The average of crop yield for Punjab (4366) and Haryana (3181) in 2017-18 season was used since two states cover the most of our area of interest.

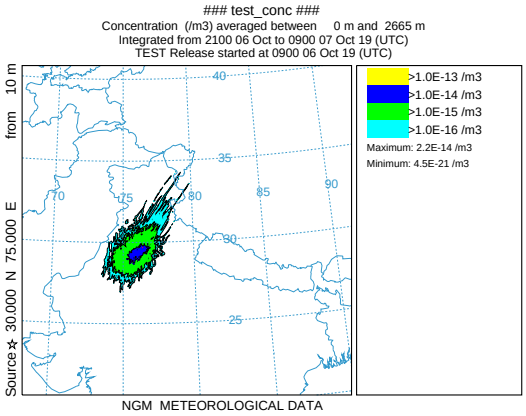
[11] The average of values for complete burn (0.89) and partial burn (0.67) was used.

Figure A1: $PM_{2.5}$ dispersion from illustrative emission event

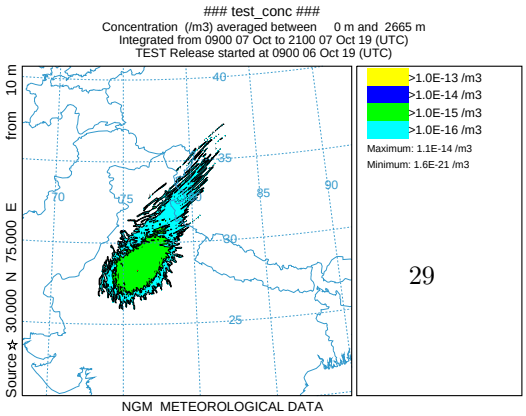
(a) 0 to 12 hours after release



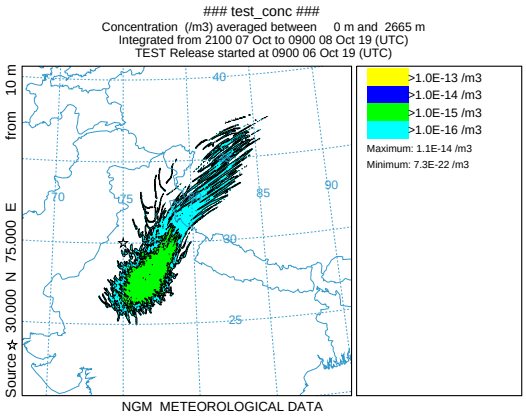
(b) 12 to 24 hours after release



(c) 24 to 36 hours after release



(d) 36 to 48 hours after release



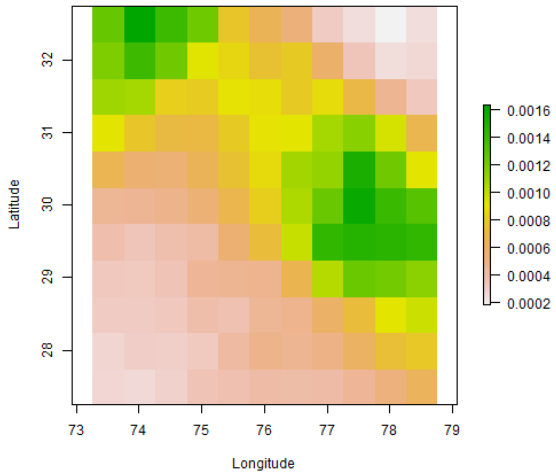


Figure A2: α_i by source location

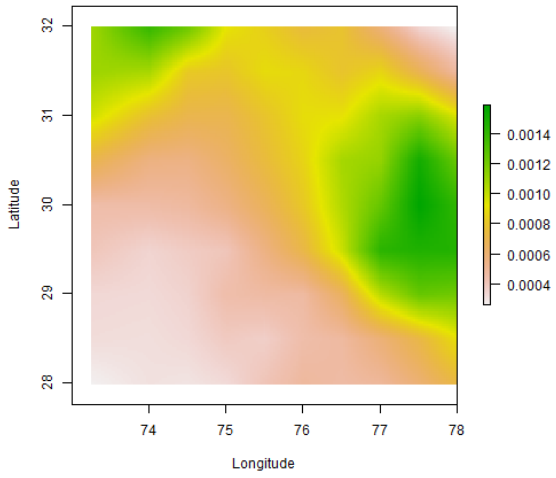


Figure A3: α_i by source location - interpolated

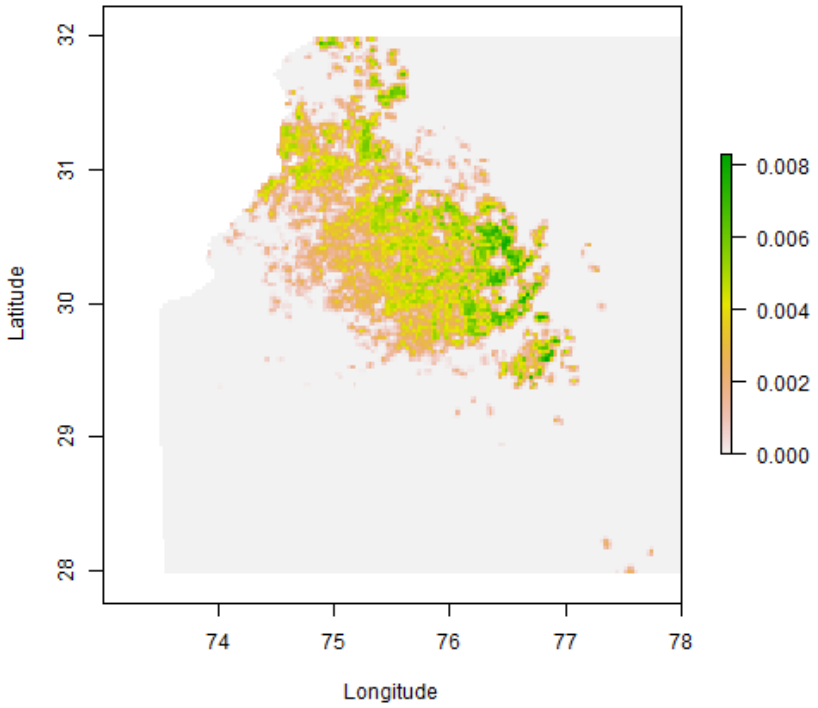


Figure A4: Linear loss contributions ($\alpha_i \cdot b_i \cdot l_i$) by pixel

Note: Pixel resolution is approximately 2.7 km

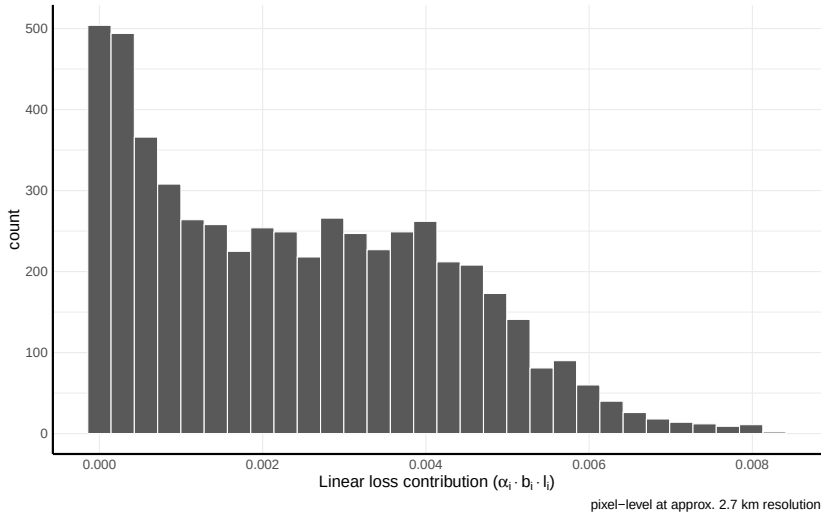


Figure A5: Histogram of linear loss contributions ($\alpha_i \cdot b_i \cdot l_i$) by pixel

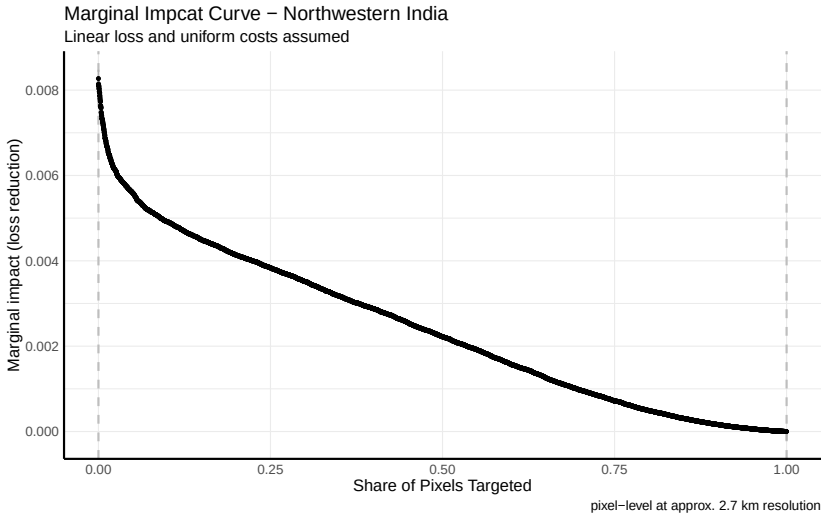


Figure A6: Marginal impact curve under linear loss (pixel-level)