Essays in Development Economics

Martin Kosík

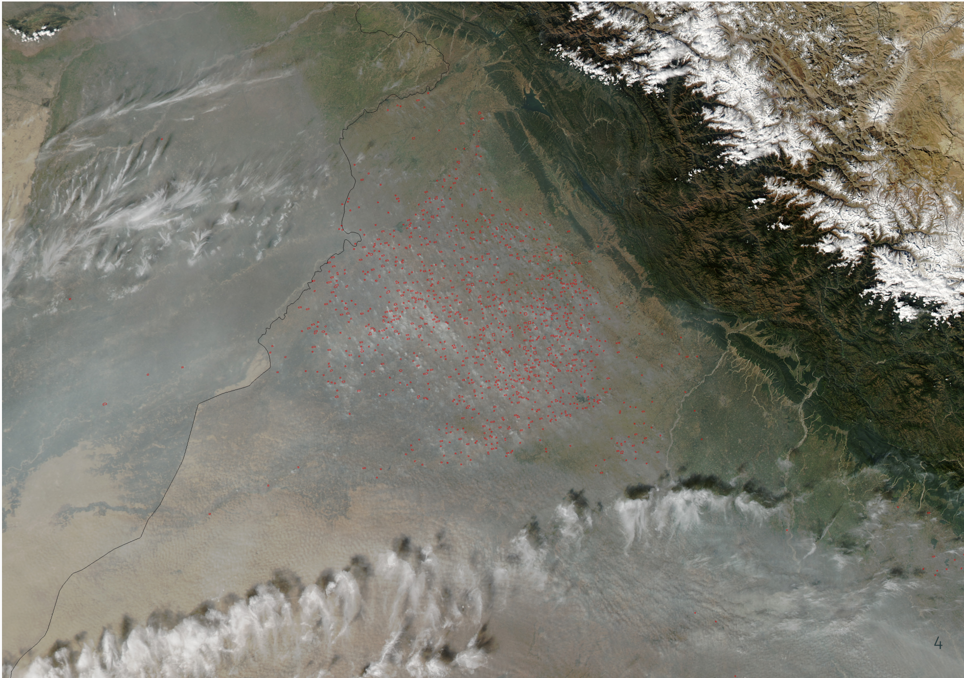
Research projects

- Optimal targeting of interventions to reduce crop residue burning
- Ethnic procurement (with Vasily Korovkin and Pasha Andreyanov)
    - Role of ethnic ties in Russian public procurement auctions ▸ more
- Estimating publication bias in observational studies
    - Use newly available data since the study had been published to re-run the same specifications
      ▸ more

Optimal targeting of interventions to reduce crop residue burning

Motivation

- Air pollution shown to have adverse effects on health (Deryugina et al., 2019), productivity (Chang et al., 2016), and academic achievement (Gilraine and Zheng, 2022)
- Crop residue (stubble) burning is an important contributor to air pollution in some regions (e.g., northwestern India) (Liu et al., 2018)
- Conditional payments to farmers proposed to reduce residue burning (Jack et al., 2023)
- This paper: Given limited resources, which places should be targeted for interventions to reduce air pollution?



Overview

- Goal: Target the interventions into places where the greatest impact can be achieved
- Modeling of two main aspects:
  1. Harm of the air pollution
     - On average, how much harm would additional emissions from a given location cause?
     - Depends on the weather patterns (wind direction, strength, etc.) and spatial distribution of the population
     - I will use an air pollution transport model (HYSPLIT) to estimate the overall impact
  2. Costs
     - How much we would have to spend to reduce the pollution in a given location?
     - Need to model the response of farmers to an intervention

Overview - Contributions

- Targeting: for a given budget, which villages should receive intervention
    - Optimal policy achieves twice the impact of random targeting but only 2% greater impact than targeting locations with most burning
- Scaling up: What will be the average and marginal impacts if we scale up the interventions
    - Increase of budget from 1 to 10 mil. USD decreases the marginal effects by $\approx 30\%$
- Gains from better predictions
    - Predicting where will the burning occur is much more important than weather forecasts

Overview

- Mechanics of the modeling approach
    - interventions $\rightarrow$ burned area $\rightarrow$ emissions $\rightarrow$ air pollution $\rightarrow$ welfare gain
- Inputs into the analysis
    - Share of burned area, historical weather data, RCT results (Jack et al., 2023), population distribution, ...
- Outputs of the analysis
    - Optimal allocation (assignment of interventions to villages)
    - Welfare gain associated with each allocation

Outline

- Background and Literature Review
- Modeling Air Pollution Transport
- Problem Formulation
- Preliminary Results
- Extensions

Background and literature review

Crop residue burning

- Cheap and fast way to clear land after harvest
- Mostly in areas with dual crop systems and mechanized harvesting
  - Dual crop system (e.g., for rice from June to October) might leave little time between harvest and next sowing
  - Mechanized harvesting leaves crop residue on fields, which interferes with the sowing
- Potential solutions
  - Specialized seeding machine (Happy Seeder)
  - Speeding up decomposition of residue
  - Manual removal of crop residue

Literature review

- Effects of air pollution - optimal policy not considered
    - Chang et al. (2016), Heft-Neal et al. (2018), Deryugina et al. (2019), Graff Zivin et al. (2020), Heft-Neal et al. (2020), Gilraine and Zheng (2022), and Pullabhotla et al. (2022)
- Optimal policy in environmental economics - novel context
    - Blundell et al. (2020) and Assunção et al. (2022)
    - Mbakop and Tabord-Meehan (2021) and Kitagawa and Tetenov (2018)
- Crop residue burning - optimal policy not considered
    - RCTs: Pant (2014) and Jack et al. (2023)
    - Atmospheric science: Liu et al. (2018) and Kulkarni et al. (2020)
    - Predictions using satellite data: Liu et al. (2020)
- Gains from more accurate predictions - novel context
    - Rosenzweig and Udry (2019), Anand (2022), and Molina and Rudik (2022)

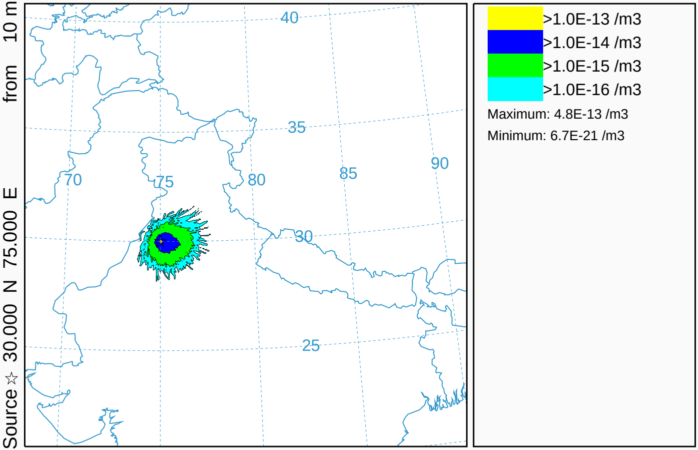
Modeling air pollution transport

Modeling air pollution transport

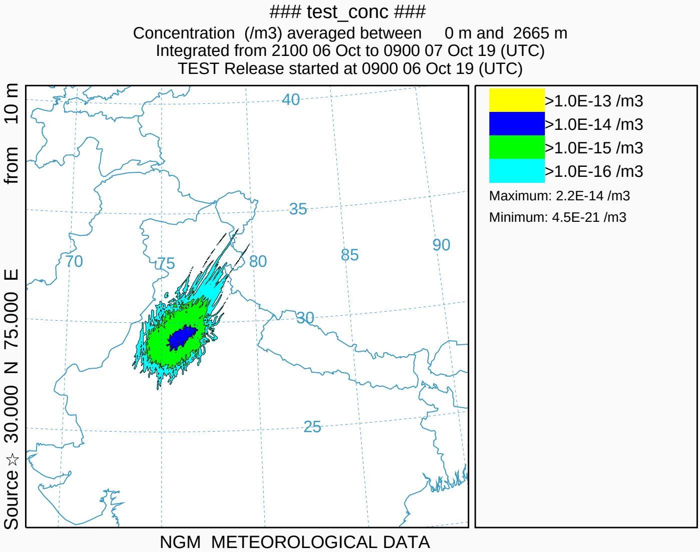
- HYSPLIT dispersion model
  - Hybrid Single-Particle Lagrangian Integrated Trajectory model
  - One of the most extensively used atmospheric transport and dispersion models in the atmospheric sciences
  - Applications include tracking and forecasting the release of wildfire smoke, wind-blown dust, volcanic ash, and crop residue burning (Stein et al., 2015)
- Main output of interest
  - *Source-receptor matrix*: SRM_{ij}
  - Fraction of emissions from source i that are transported into j

test_conc

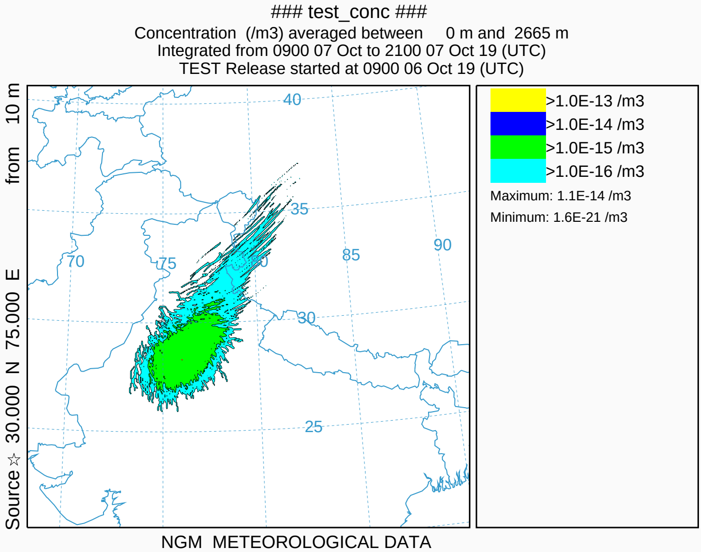
Concentration (/m3) averaged between 0 m and 2665 m
Integrated from 0900 06 Oct to 2100 06 Oct 19 (UTC)
TEST Release started at 0900 06 Oct 19 (UTC)



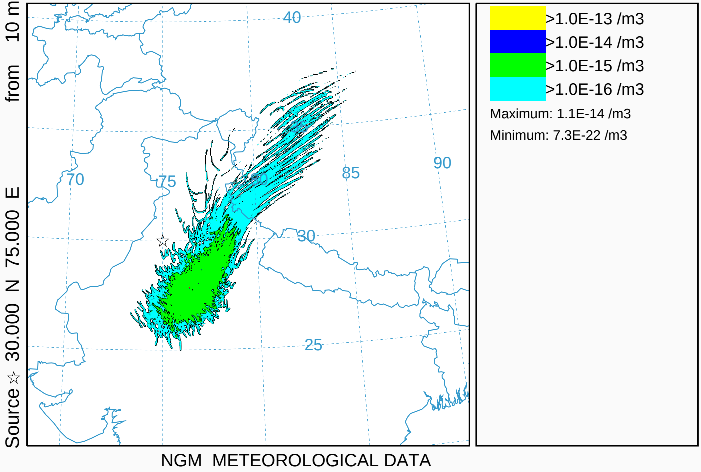
NGM METEOROLOGICAL DATA



test_conc
Concentration (/m3) averaged between 0 m and 2665 m
Integrated from 2100 06 Oct to 0900 07 Oct 19 (UTC)
TEST Release started at 0900 06 Oct 19 (UTC)
Source ☆ 30.000 N 75.000 E from 10 m
40
35
30
25
70
75
80
85
90
>1.0E-13 /m3
>1.0E-14 /m3
>1.0E-15 /m3
>1.0E-16 /m3
Maximum: 2.2E-14 /m3
Minimum: 4.5E-21 /m3
NGM METEOROLOGICAL DATA


test_conc
Concentration (/m3) averaged between 0 m and 2665 m
Integrated from 0900 07 Oct to 2100 07 Oct 19 (UTC)
TEST Release started at 0900 06 Oct 19 (UTC)
Source ☆ 30.000 N 75.000 E from 10 m
40
35
30
25
70
75
85
90
>1.0E-13 /m3
>1.0E-14 /m3
>1.0E-15 /m3
>1.0E-16 /m3
Maximum: 1.1E-14 /m3
Minimum: 1.6E-21 /m3
NGM METEOROLOGICAL DATA

test_conc
Concentration (/m3) averaged between 0 m and 2665 m
Integrated from 2100 07 Oct to 0900 08 Oct 19 (UTC)
TEST Release started at 0900 06 Oct 19 (UTC)



Problem formulation

Definitions

- $SRM_{ij} \dots$ fraction of emissions from source i that are transported into j
- $E_i \dots$ total mass of pollutants emitted from location i
- $P_j = \sum_i SRM_{ij} E_i \dots$ total air pollution concentration in i
- $L_j = f(P_j) \dots$ per capita loss (harm) of exposure to P_j
- $N_j \dots$ total population
- $TL = \sum_j L_j \cdot N_j \dots$ total population-weighted loss caused by air pollution across all locations
- $s_i \dots$ conditional payment amount, $u_i \dots$ share paid upfront, $x_i \dots$ covariates

Problem formulation I

$$\min_{s_i, u_i} \mathsf{TL} = \min_{s_i, u_i} \sum_j L_j N_j, \tag{1}$$

subject to budget constraint (where κ_i is compliance rate)

$$\sum_i \left(u_i r_i + (1 - u_i)\kappa_i\right) s_i l_i \leq M, \tag{2}$$

equation for enrollment rate into the program (▸ details)

$$r_i = \omega^B b(s = 0, x_i, u_i) + \omega^N \left(1 - b(s = 0, x_i, u_i)\right), \tag{3}$$

pollution loss function (▸ details)

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

source-receptor matrix decomposition of air pollution

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

Problem formulation II

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

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

and predicted share of land burned (▸ details)

$$b(s_i, u_i, x_i) = g(s_i, u_i, x_i) \tag{7}$$

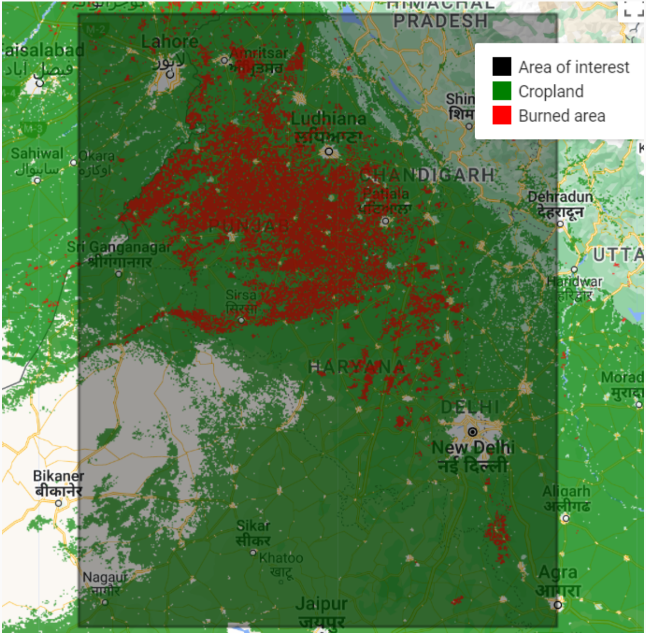
(▸ solving the model)

Preliminary results

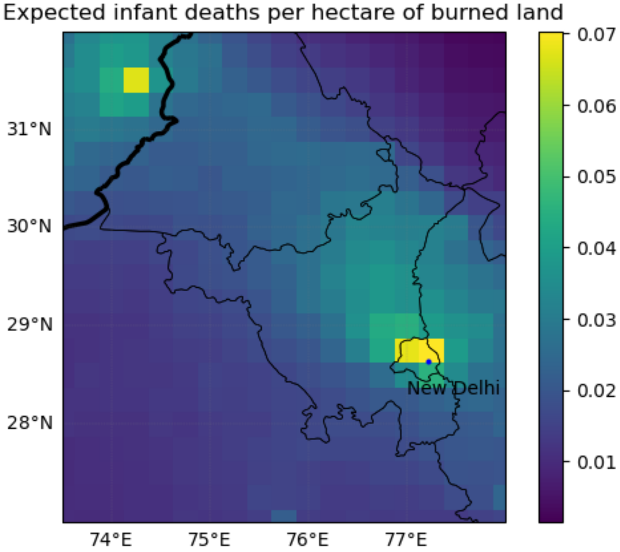
Preliminary results

- I focus on northwestern India where crop residue burning is common
- I run simulations based on October and November weather data for 56 different emission events from 2006 to 2019
- Regular grid of 441 source location
    - Expected infant deaths computed for each location separately, then interpolated on a finer grid
- MODIS satellite images for land cover and burned area estimates on 500m resolution

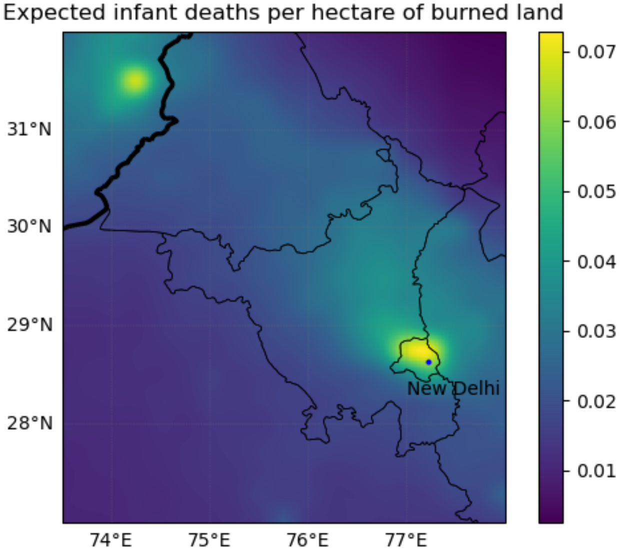
Cropland and burned area



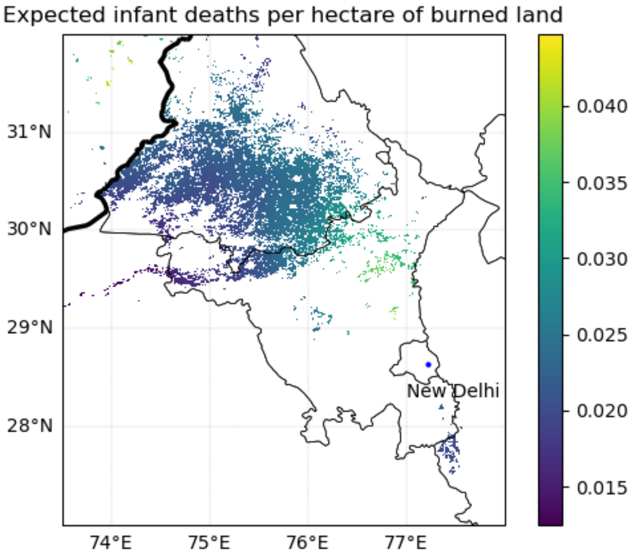
Infant deaths per ha of burned land - no interpolation



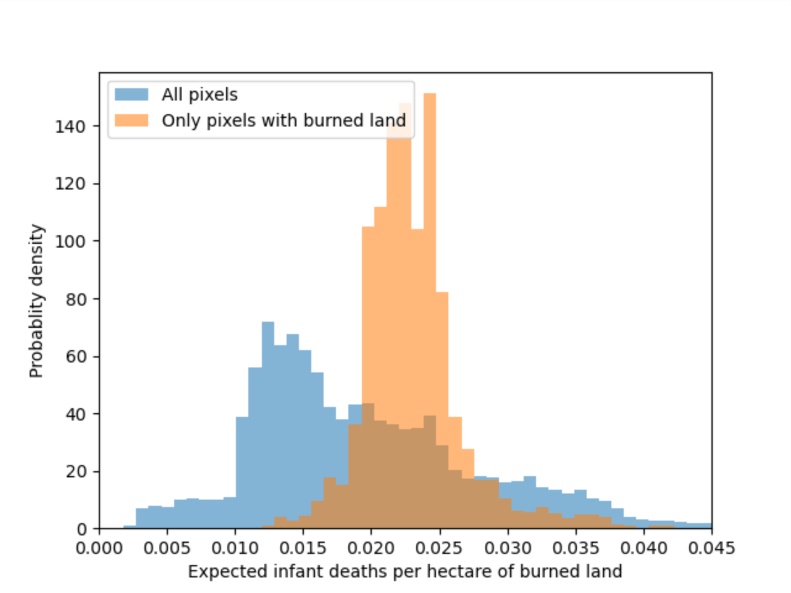
Infant deaths per ha of burned land - interpolated



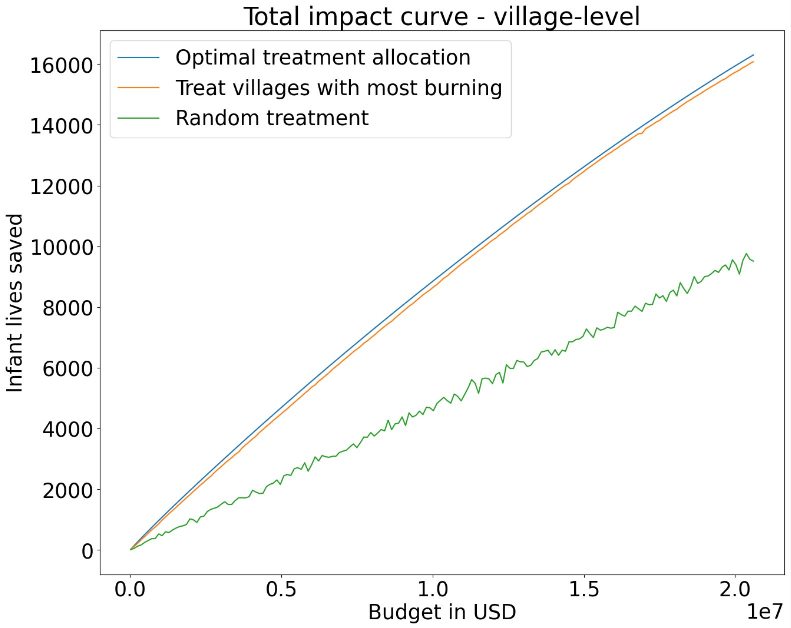
Infant deaths per ha of burned land - only burned land



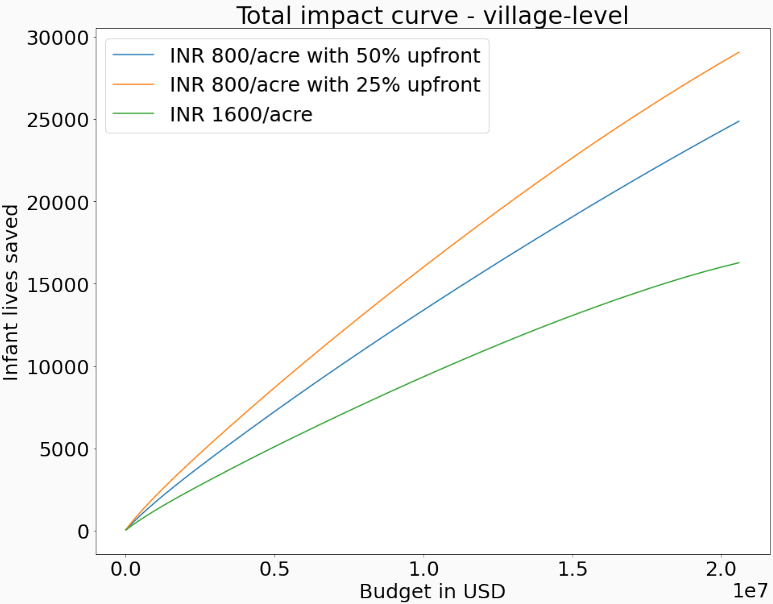
Substantial differences in harm depending on location



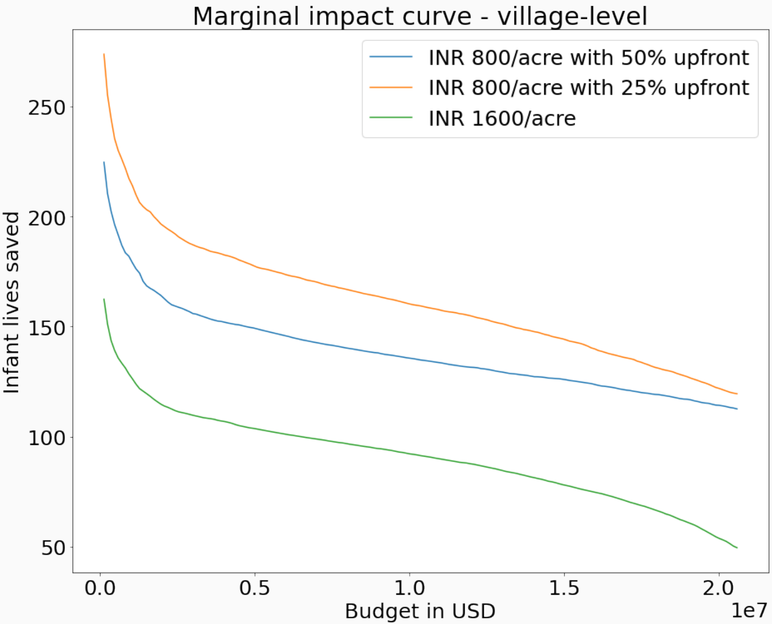
Targeting locations with most burning is nearly optimal



How do conditional payments scale up?



Scaling up - marginal effects ($\approx \Delta$ USD 120,000)

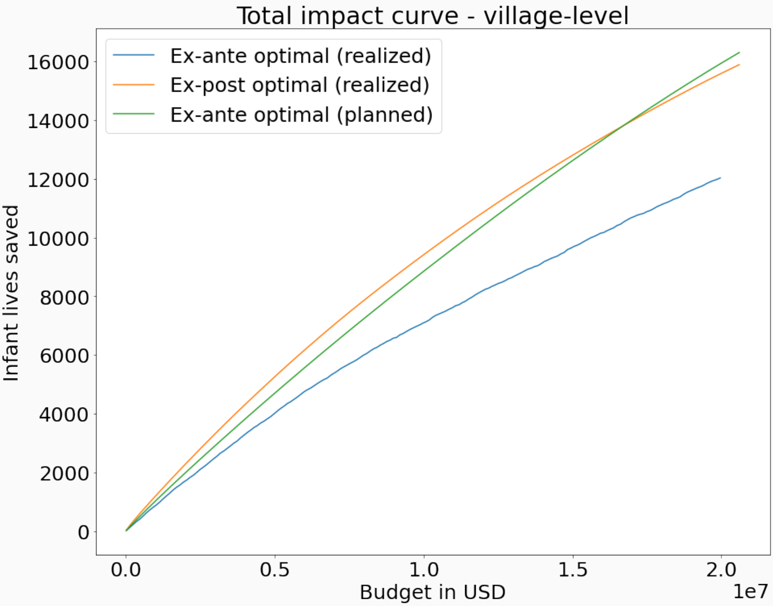


Extensions

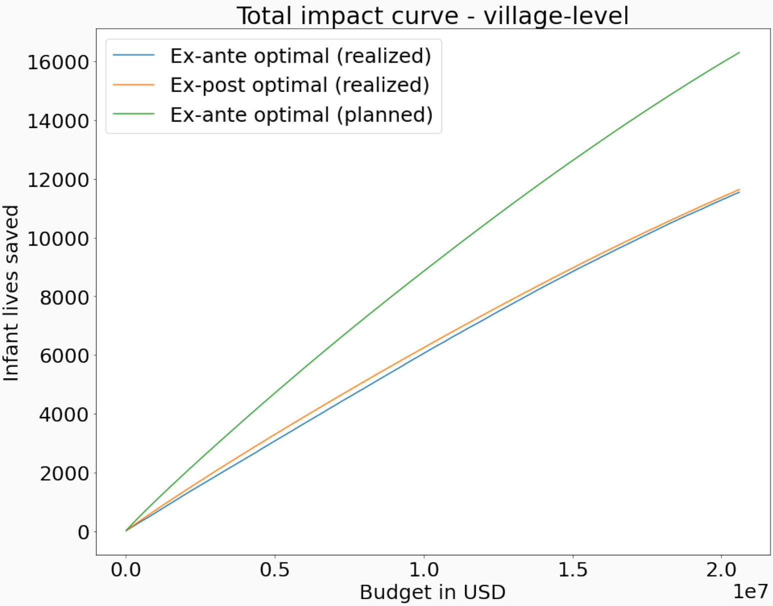
Gains from better predictions

- Policymaker faces uncertainty with respect to location of burning in the absence of intervention and weather shocks (among other sources of uncertainty)
    - $\rightarrow$ realized impact of the program can be different from what was planned
- Better predictions can reduce this uncertainty
- We can upper-bound these gains by computing ex-post optimal allocations
- Results for 2017 suggest that the predictions of burned area are more important than those of weather shocks

Gains from better predictions - burned area



Gains from better predictions - weather shocks



Alternative (reduced form) approach

- Based on the results of Pullabhotla et al. (2022) who estimate the effect of an additional km^2 burned area within a 30 km radius on infant mortality in the downwind direction
- HYSPLIT: Area burned $\rightarrow$ mass emitted $\rightarrow$ $PM_{2.5}$ conc. $\rightarrow$ infant mortality
- Reduced-form: Area burned $\rightarrow$ infant mortality (within 30 km)
- Results on the same order of magnitude as HYSPLIT approach, although with much higher variance

details

Conclusion

- Targeting
  - Optimal policy achieves twice the impact of random targeting but only 2% greater impact than targeting locations with most burning
- Scaling up
  - Increase of budget from 1 to 10 mil. USD decreases the marginal effects by $\approx 30\%$
- Gains from better predictions
  - Predicting where will the burning occur is much more important than weather forecasts

Other projects

- Ethnic procurement (with Vasily Korovkin and Pasha Andreyanov)
  - Role of ethnic ties in Russian public procurement auctions
  - Large database of 20 million purchases with information on prices, detailed product categories, bureaucrats etc.
  - Ethnicity of bureaucrats and firm managers is predicted from surnames using a specialized neural network
  - Preliminary results show more evidence in favor of statistical rather than taste-based discrimination ▸ more
- Estimating publication bias in observational studies
  - Dependence of publication probability on the results can lead to systemic bias
  - Use newly available data since the study had been published to re-run the same specifications

▸ more

References

- Acemoglu, Daron, Tristan Reed, and James A. Robinson (2014). "Chiefs: Economic Development and Elite Control of Civil Society in Sierra Leone". In: *Journal of Political Economy* 122.2. Publisher: The University of Chicago Press, pp. 319–368. DOI: *10.1086/674988*.
- Anand, Vaibhav (2022). *The Value of Forecast Improvements: Evidence from Advisory Lead Times and Vehicle Crashes*. en. SSRN Scholarly Paper. Rochester, NY. DOI: *10.2139/ssrn.4206910*.
- Andrews, Isaiah and Maximilian Kasy (2019). "Identification of and Correction for Publication Bias". en. In: *American Economic Review* 109.8, pp. 2766–2794. DOI: *10.1257/aer.20180310*.

Assunção, Juliano, Robert McMillan, Joshua Murphy, and Eduardo Souza-Rodrigues (2022). “Optimal Environmental Targeting in the Amazon Rainforest”. In: *The Review of Economic Studies*, rdac064. DOI: 10.1093/restud/rdac064.

Blundell, Wesley, Gautam Gowrisankaran, and Ashley Langer (2020). “Escalation of Scrutiny: The Gains from Dynamic Enforcement of Environmental Regulations”. en. In: *American Economic Review* 110.8, pp. 2558–2585. DOI: 10.1257/aer.20181012.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020). “Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics”. en. In: *American Economic Review* 110.11, pp. 3634–3660. DOI: 10.1257/aer.20190687.

Chang, Tom, Joshua Graff Zivin, Tal Gross, and Matthew Neidell (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, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif (2019). “The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction”. en. In: *American Economic Review* 109.12, pp. 4178–4219. DOI: *10.1257/aer.20180279*.

Gilraine, Michael and Angela Zheng (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*.

Graff Zivin, Joshua, Tong Liu, Yingquan Song, Qu Tang, and Peng Zhang (2020). “The unintended impacts of agricultural fires: Human capital in China”. en. In: *Journal of Development Economics* 147, p. 102560. DOI: 10.1016/j.jdeveco.2020.102560.

Heft-Neal, Sam, Jennifer Burney, Eran Bendavid, and Marshall Burke (2018). “Robust relationship between air quality and infant mortality in Africa”. en. In: *Nature* 559.7713, pp. 254–258. DOI: 10.1038/s41586-018-0263-3.

Heft-Neal, Sam, Jennifer Burney, Eran Bendavid, Kara K. Voss, and Marshall Burke (2020). “Dust pollution from the Sahara and African infant mortality”. en. In: *Nature Sustainability* 3.10, pp. 863–871. DOI: 10.1038/s41893-020-0562-1.

Jack, B. Kelsey, Seema Jayachandran, Kala Namrata, and Pande Rohini (2023). *Money (Not) to Burn: Payments for Ecosystem Services to Reduce Crop Residue Burning*. Working Paper.

Jain, Niveta, Arti Bhatia, and Himanshu Pathak (2014). “Emission of air pollutants from crop residue burning in India”. In: *Aerosol and Air Quality Research* 14.1. Publisher: Taiwan Association for Aerosol Research, pp. 422–430.

Jensen, Paul A. and Jonathan F. Bard (2002). *Operations Research Models and Methods*. English. Hoboken, NJ: Wiley.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, Roy Jenne, and Dennis Joseph (1996). “The NCEP/NCAR 40-Year Reanalysis Project”. en. In: *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.

Kitagawa, Toru and Aleksey Tetenov (2018). “Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice”. en. In: *Econometrica* 86.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA13288, pp. 591–616. DOI: *10.3982/ECTA13288*.

Kulkarni, Santosh H., Sachin D. Ghude, Chinmay Jena, Rama K. Karumuri, Baerbel Sinha, V. Sinha, Rajesh Kumar, V. K. Soni, and Manoj Khare (2020). “How Much Does Large-Scale Crop Residue Burning Affect the Air Quality in Delhi?” en. In: *Environmental Science & Technology*. Publisher: American Chemical Society. DOI: *10.1021/acs.est.0c00329*.

Kvarven, Amanda, Eirik Strømland, and Magnus Johannesson (2020). “Comparing meta-analyses and preregistered multiple-laboratory replication projects”. In: *Nature Human Behaviour* 4.4. Publisher: Nature Publishing Group UK London, pp. 423–434.

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

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

Mbakop, Eric and Max Tabord-Meehan (2021). "Model Selection for Treatment Choice: Penalized Welfare Maximization". en. In: *Econometrica* 89.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16437, pp. 825–848. DOI: *10.3982/ECTA16437*.

Michalopoulos, Stelios and Elias Papaioannou (2016). "The Long-Run Effects of the Scramble for Africa". en. In: *American Economic Review* 106.7, pp. 1802–1848. DOI: *10.1257/aer.20131311*.

Molina, Renato and Ivan Rudik (2022). *The Social Value of Predicting Hurricanes*. en. SSRN Scholarly Paper. Rochester, NY. DOI: *10.2139/ssrn.4266614*.

Pant, Krishna Prasad (2014). *Monetary Incentives to Reduce Open-Field Rice-Straw Burning in the Plains of Nepal*. Tech. rep.

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

Rosenzweig, Mark R. and Christopher R. Udry (2019). *Assessing the Benefits of Long-Run Weather Forecasting for the Rural Poor: Farmer Investments and Worker Migration in a Dynamic Equilibrium Model*. Working Paper. DOI: *10.3386/w25894*.

Stein, A. F., R. R. Draxler, G. D. Rolph, B. J. B. Stunder, M. D. Cohen, and F. Ngan (2015). "NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System". EN. In: *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*.

Thank you for your attention

Program enrollment rate I

- Program enrollment given by

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

  where $\omega^B = P(R_{ip} = 1 | B_{pi} = 1, s_i = 0)$, $\omega^N = P(R_{ip} = 1 | B_{pi} = 0, s_i = 0)$, R_{ip} is enrollment indicator, and B_{pi} is burning indicator
- This allows for self-selection into the program based on demand for burning
  - Important for cost effectiveness of the program
  - The higher the enrollment of farmers who would never burn their fields, the lower the per enrollee benefits of the program

go back

Program enrollment rate II

- ω^B and ω^N can be estimated from experimental micro data
- For any convex interval $\mathcal{B}$

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

where b_{pi} pre-treatment probability of burning

- All of the terms above (except ω^B and ω^N) can be estimated from the data
- 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

go back

Pollution loss function

- I will only consider the effects on infant mortality
- I will use linear loss function as a default
  - There is evidence supporting linear effect of $PM_{2.5}$ concentration on infant mortality (Heft-Neal et al., 2018)
  - Substantial reductions in 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)
  - I will use estimates from Pullabhotla et al. (2022)
- Future work: alternative parametrizations of the loss function

go back

Burned area $\rightarrow$ emissions

- We can apply decomposition used by Jain et al. (2014) and Liu et al. (2020) to express ϕ as

$$\phi = CY \times RC \times f_{DM} \times f_{CC} \times EF \tag{8}$$

  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 from the literature for rice paddy as the crop and $PM_{2.5}$ as the pollutant, we get $\phi \approx 23772$

go back

Modeling the treatment effect I

- Challenge: Extrapolate the treatment effect from the sample of villages in the RCT (all of them had high pre-treat. burned land share) to all villages in Punjab and Haryana
- My approach: Impose functional form assumptions and then test sensitivity of the results
- Microfoundation of this approach more

go back

Modeling the treatment effect II

1. Get estimates of $b(s_i = 0, x_i)$ from ridge regression wit pre-treatment data
2. Compute village specific treatment effect

$$\hat{\beta}_i = h(\hat{b}(s_i = 0, x_i), \hat{\beta}, \bar{b}_{\text{RCT-control}})$$

3. Compute the counterfactual burned cropland share under treatment

$$\hat{b}(s_i = 1, x_i) = \hat{b}(s_i = 0, x_i) - \hat{\beta}_i$$

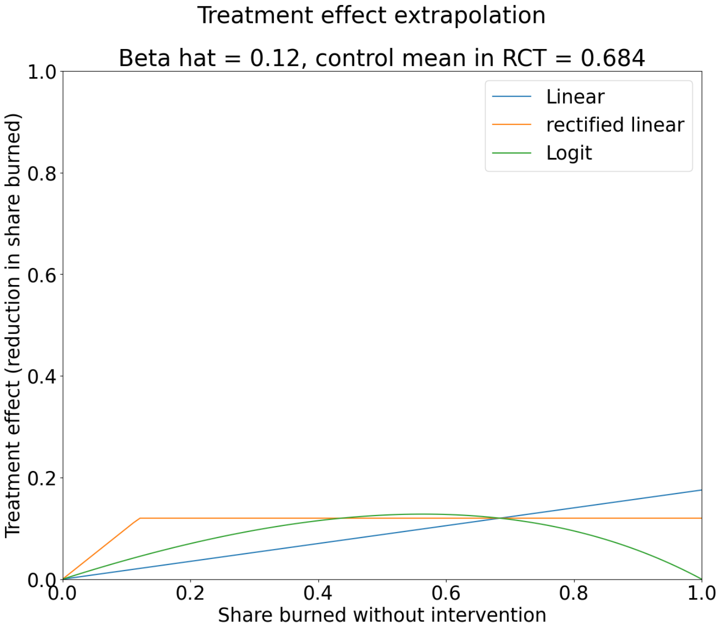
go back

Modeling the treatment effect III

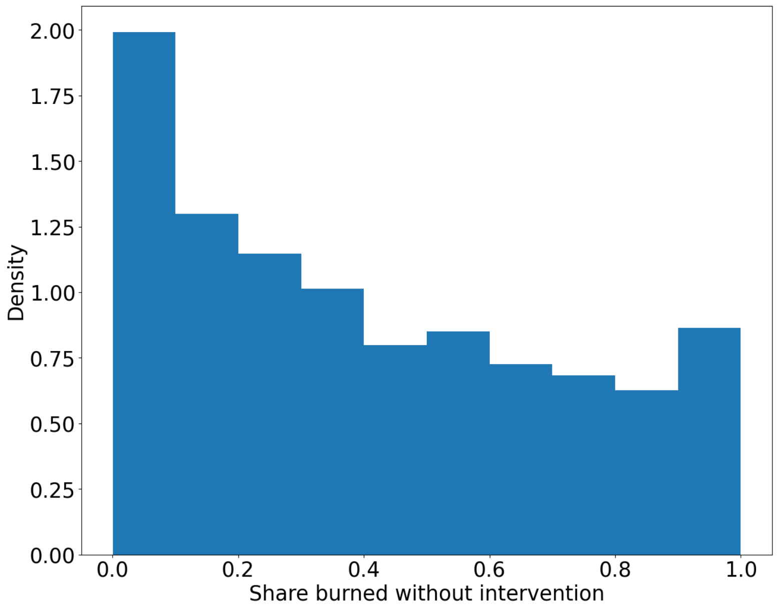
- Possible forms of $h(.)$:
  1. Linear:
     $$\hat{\beta}_i = \hat{b}(s_i = 0, x_i) \cdot \frac{\hat{\beta}}{\bar{b}_{\text{RCT-control}}}$$
  2. Rectified linear:
     $$\hat{\beta}_i = \begin{cases} \hat{\beta} & \text{for } \hat{b}(s_i = 0, x_i) > \hat{\beta} \\ \hat{b}(s_i = 0, x_i) & \text{for } \hat{b}(s_i = 0, x_i) \leq \hat{\beta} \end{cases}$$
  3. Logit:
     $$\hat{\beta}_i = \hat{b}(s_i = 0, x_i) - \sigma\left(\sigma^{-1}(\hat{b}(s_i = 0, x_i)) + \hat{\beta}_i^{\text{logit}}\right)$$
     where
     $$\hat{\beta}_i^{\text{logit}} = \sigma^{-1}\left(\bar{b}_{\text{RCT-control}}) + \hat{\beta}\right) - \sigma^{-1}\left(\bar{b}_{\text{RCT-control}})\right)$$
     and $\sigma(x) = 1/(1 + \exp -x)$ is the standard logistic function

go back

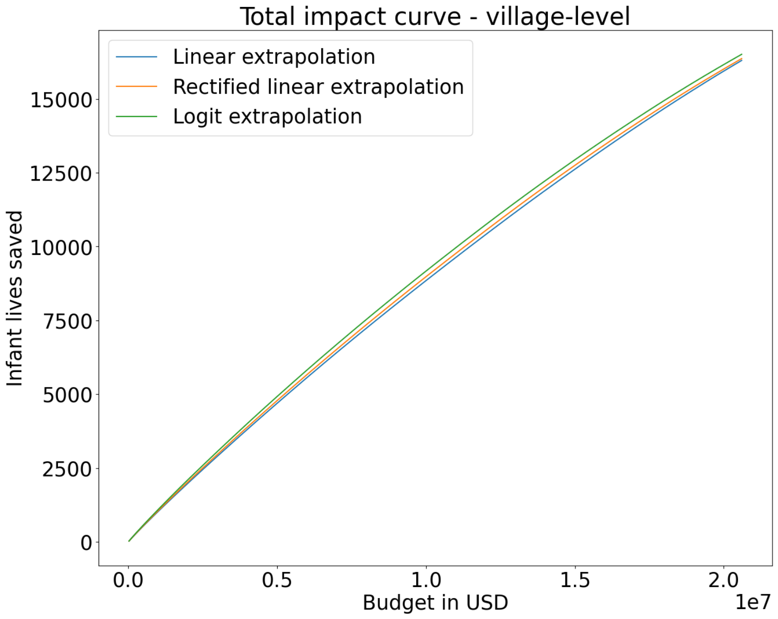
Treatment effect extrapolation



Histogram of share of burned cropland



Extrapolation and total loss



Microfounding the model

- Farmer with plot p in village i compares the profits from burning (Π^B_{pi}) and not burning (Π^N_{pi})

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

- ϵ_{pi}
  - idiosyncratic shock that captures unobserved plot-level factors
  - If of the type 1 extreme value distribution then the b_i, can be expressed as

$$b(s_i, x_i) = \frac{\exp\left(\beta s_i + x_i'\gamma\right)}{1 + \exp\left(\beta s_i + x_i'\gamma\right)}, \tag{9}$$

▸ go back

Solving the model

- Linear loss and finite discrete interventions
  - We can formulate the optimization as a multiple-choice knapsack problem
  - Fast algorithms for approximate solutions of multiple-choice knapsack problem (OR-tools library in Python)
- Linear loss and continuous interventions
  - Separable programming can be applied to obtain approximate solutions by piecewise linearization of the nonlinear objective and then using linear programming (Jensen and Bard, 2002, chapter 10.4)

go back

Reduced-form approach - details

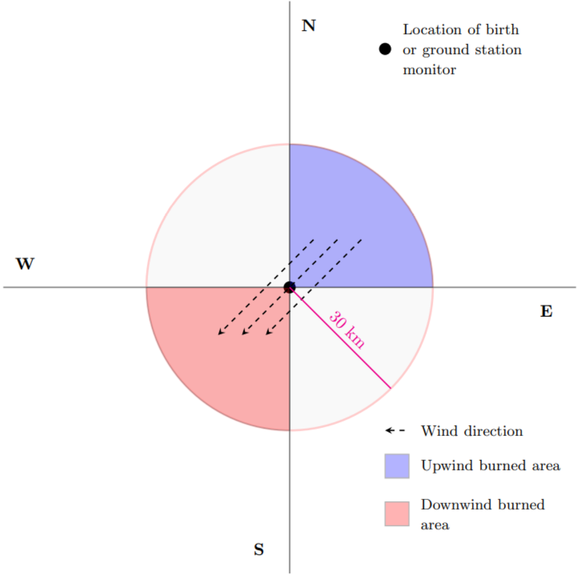
- Wind direction data: NCEP reanalysis (Kalnay et al., 1996)
  - October and November from 1948 to 2022
- Expected infant deaths are calculated as

$$L_i = \sum_d N_{id} \cdot c \cdot f_{id} \cdot \beta_{\text{BA}}$$

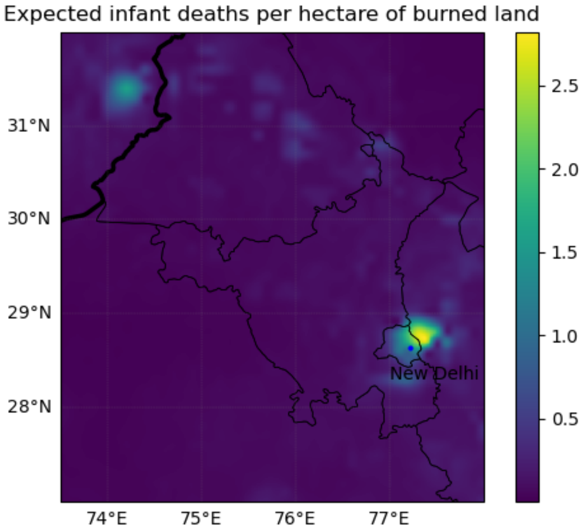
where L_i is the expected increase in infant deaths per an additional burned hectare in location i, f_{id} is the relative frequency of historical monthly wind directions, N_{id} is population living in a quarter in the direction d within the 30 km radius of location i, c is the population share of infants (defined as individuals under the age of one), and β_{BA} is the effect of an additional hectare burned on infant mortality.

go back

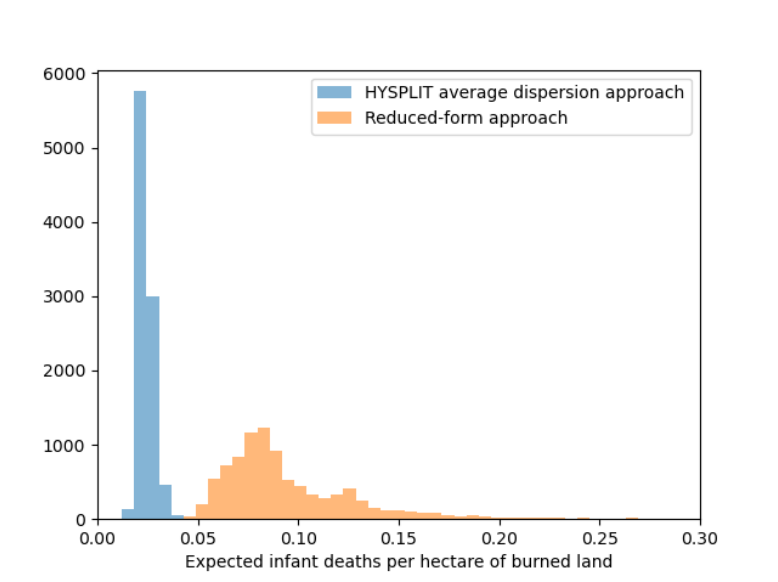
Reduced form approach



Reduced form approach - interpolated



Comparison - histogram (only locations with burned land)



Setting: Public Procurement, Bureaucrats, and Firms

- **Public body**—e.g., a hospital wants to buy antibiotics
  1. This hospital needs to run an auction
  2. **Bureaucrats** run auctions for this hospital
  3. Hospital submits documentation including reserve price
  4. Bureaucrat manages all paperwork & runs bid review
  5. Firms submit their sealed bids
  6. Hospital signs the contract with the winning **firm**
- Observe ethnic markers of bureaucrats & firm decision-makers
- **Result 1:** firms tend to win contracts with coethnic bureaucrats

Conceptual Framework: Taste-based Bias or Information

- Two competing stories: taste-based bias vs. screening
- First-price sealed-bid auction, firm payoff:

$$\mathbb{E}[\pi(b)] = (b - c)Pr(win|b)$$

  - Assumption: firms ex-ante symmetric
  - Assumption: costs for project are i.i.d. draws from the same distribution

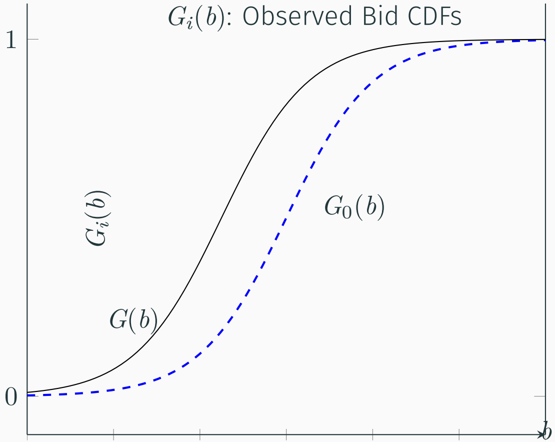
- **Tasted-based:** add perturbation to probability of winning

$$\mathbb{E}[\pi_0(b)] = (b - c)(Pr(win|b) + \epsilon) \Rightarrow \frac{\partial^2 \mathbb{E}[\pi_0(b)]}{\partial b \partial \epsilon} > 0 \Rightarrow b_0(\epsilon, c) \uparrow$$

  - I.e., bureaucrats select inefficient firms of same type

Illustration: Tasted-based Bid Distribution

- For generic distributions should lead to **higher prices**



Conceptual Framework: Taste-based bias or Information

- First-price sealed-bid auction, firm payoff:

$$\mathbb{E}[\pi(b)] = (b - c)Pr(win|b)$$

  - Assumption: firms are asymmetric—either different costs or screening
  - Formally: assume different cost distributions or add information frictions

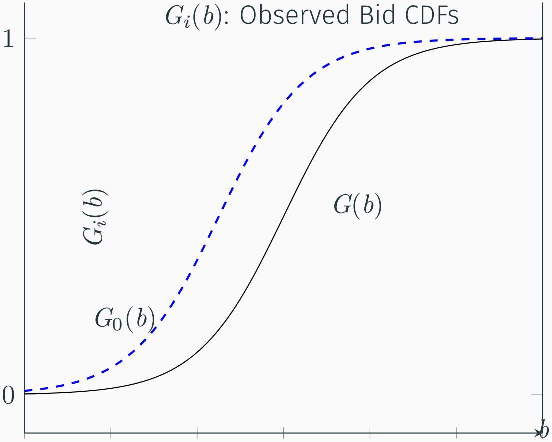
- Screening: informally, add some perturbation to realized costs

$$\mathbb{E}[\pi_0] = (b - (c - \epsilon))(Pr(win|b)) \Rightarrow \frac{\partial^2 \mathbb{E}[\pi_0]}{\partial b \partial \epsilon} \Rightarrow b_0(\epsilon, c) \downarrow$$

  - Fewer information frictions, or on average lower costs

Illustration: Information or Cost-Based Bid Distribution

- For generic distributions should lead to **lower prices**
- What is happening in the data?



- **Result 2:** lower prices in coethnic pairs

Setting

- Russian procurement over the period 2014 through 2018:
    - High ethnic diversity between regions
    - More importantly, exploit *within-organization* variation in bureaucrats
- Micro-level data on
    - Procurement outcomes
    - Firm decision-makers ethnicities
    - Other bidding and firm outcomes
- Main findings
    - **Result 1:** firms tend to win contracts with coethnic bureaucrats
    - **Result 2:** lower prices in coethnic pairs—second-best?

Classification Algorithm

- Ethnicity predicted from surnames
    - Memorial data—over 1 million individuals with ethnicity labels
    - Pool ethnicities into five groups
        - Not always the largest but the most salient features of last names
        - E.g., Belorussian spelling “Russianized”—almost impossible to differentiate
        - Same for Yakuts, and some other groups—very hard
    - Character-level bidirectional LSTM neural network
    - Around 97% out-of-sample accuracy
    - Calibration of predicted probabilities using isotonic regression

Main results

- Cross-sectional OLS results show
    - Firms with managers co-ethnic with bureaucrats are more likely to win
    - The winning firms with co-ethnic managers receive lower prices
- Use switches of bureaucrats for identification:

$$\text{Share of Minority Firms}_{it} = \alpha_i + \mu_t + \beta_0 \cdot \text{Share of Min. Bureaucrats}_{it} + \epsilon_{it},$$

  where the data is aggregated on organization i and time t level and estimate
- Augment with lags and leads
- back

Baseline DiD

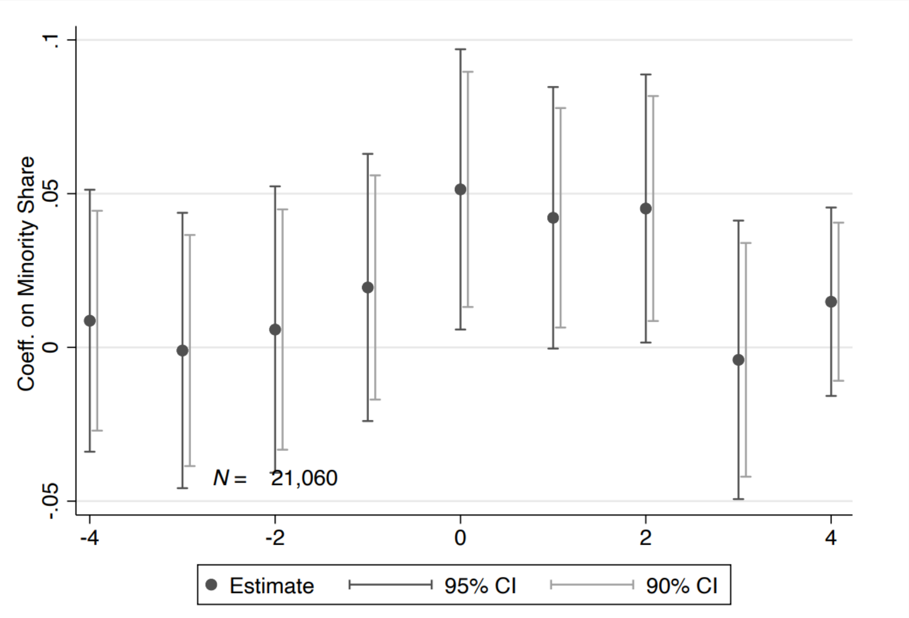
Dependent Variable: Share of Co-Ethnics among Contract Winners

	(1)	(2)	(3)	(4)	(5)
Share of Minority Bureaucrats	0.010**	0.051**			0.158***
	(0.005)	(0.023)			(0.017)
Share of Slavic Bureaucrats			0.010**	0.037*	
			(0.005)	(0.021)	
# of Purchases by Org. per period > 0					0.024***
					(0.002)
Lags & Leads		✓			✓
Lags & Leads, Slavic				✓	
$\hat{\beta}_0 + \sum_{k=1}^{K}(\hat{\beta}_k - \hat{\beta}_{-k})$		0.117		0.065	0.172
Mean Dep. Var.†	0.105	0.100	0.895	0.900	0.093
Observations	143,513	21,060	143,513	21,060	82,720

Notes: Standard errors are clustered on organization level. * p<0.1, ** p<0.05, *** p<0.01.

back

Two-way FE coefficients



Baseline DiD - prices

Dependent Variable: Log-Price per Contract

	(1)	(2)	(3)	(4)
	Log-Price		Log-Max-Price	
Coethnic Share	-0.028***	-0.110*	-0.024**	-0.103**
	(0.011)	(0.059)	(0.009)	(0.046)
Lags & Leads		✓		✓
$\hat{\beta}_0 + \sum_{k=1}^{K}(\hat{\beta}_k - \hat{\beta}_{-k})$		-0.264		-0.260
Observations	142,962	20,988	142,962	20,988

Notes: Standard errors are clustered on organization level. * p<0.1, ** p<0.05, *** p<0.01.

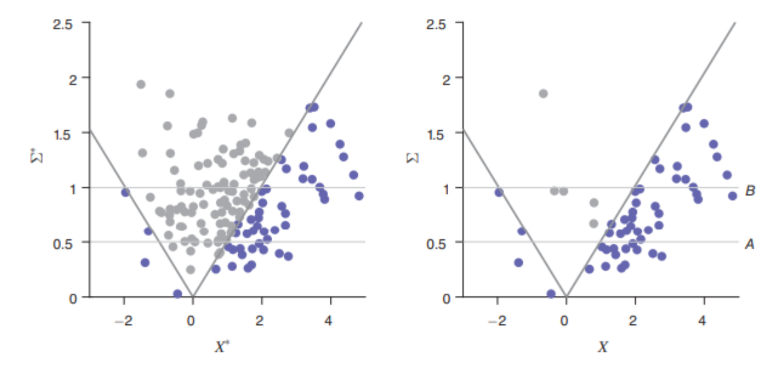
back

Motivation

- Publication of empirical studies depends on their results (effect size, significance, ...)
- This can lead to bias in the published estimates
- In experimental research, systematic replication studies were conducted to identify
- However, no systematic replication of observational studies done in economics
- This paper: Use newly available data since the study had been published to re-run the same specifications

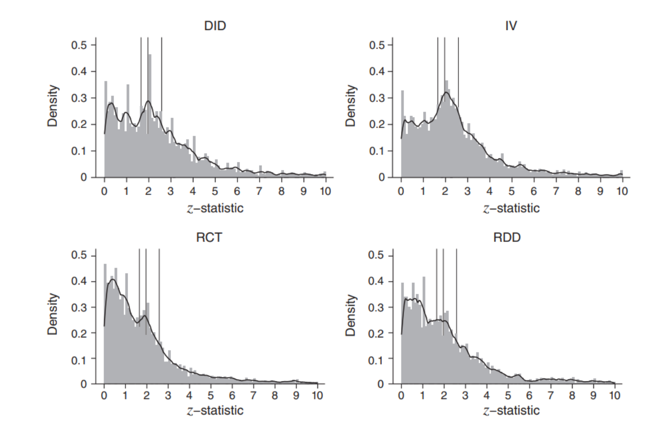
Existing methods for observational research

- Meta-analytic approach
  - Assume that effect sizes and standard errors are independent across all studies (very strong assumption)
  - Kvarven et al. (2020) compared the bias-adjusted effect sizes obtained using these methods are almost three times as large as those from the systematic replication studies



Existing methods for observational research

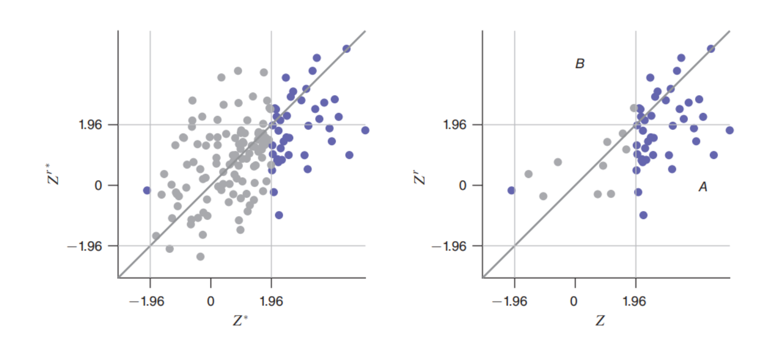
- Using the distribution of z-statistics (Brodeur et al., 2020)
  - Based on comparing the density of z-statistics around the significance threshold
  - Cannot detect p-hacking that would have large impact on the z-statistics



Andrews and Kasy (2019) approach

- Uses systematic replication studies
- Assumes the true effects for the original study and replication are draws from the same distribution
- Selection on publication identified up to scale from

$$\frac{f_{Z,Z^r}(b,a)}{f_{Z,Z^r}(a,b)} = \frac{p(b)}{p(a)}$$



Systemic replication of observational studies - Economics

- One can also use newly available datasets (e.g., DHS surveys)
- Potential issue: the true effects might decline in time
  - Estimate the decline in the true effects using multiple time periods of the new data
  - Focus on effects where the decline is likely to be small
    - E.g., effects that according to the published research should persist over 100 years
- Examples: Acemoglu et al. (2014) and Michalopoulos and Papaioannou (2016)

back

Michalopoulos and Papaioannou (2016) table 2 - original

	All Ethnicity-Country Homelands						Ethnicity-Country H	
	All Observations				Excl. Outliers	Excl. Capitals	All Obser	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
					Panel A. Negative Binomial ML Estimates			
SPLIT (Partitioning)	0.4513***	0.3329**	0.4495***	0.4626***	0.4494***	0.4565***	0.9247***	0.8050***
	(0.1611)	(0.1851)	(0.1254)	(0.1201)	(0.1172)	(0.1236)	(0.1704)	(0.2372)
SPIL (Adjacent Split)	0.0481	0.3910	0.4619*	0.4920*	0.4834*	0.4256*	0.0879	0.5679
	(0.2789)	(0.3430)	(0.2626)	(0.2628)	(0.2686)	(0.2760)	(0.5748)	(0.4733)
Log Likelihood	-4506.794	-4280.172	-4119.95	-4108.723	-3993.148	-3781.286	-1697.469	-1561.61
R-square	0.203	0.528	0.645	0.633	0.168	0.182	0.148	0.343
					Panel B. Linear Probability Model (LPM) Estimate			
SPLIT (Partitioning)	0.0562**	0.0660***	0.0783***	0.0819***	0.0839***	0.0789***	0.0874**	0.0835*
	(0.0241)	(0.0238)	(0.0258)	(0.0266)	(0.0266)	(0.0266)	(0.0399)	(0.0484)
SPIL (Adjacent Split)	0.0571	0.1146***	0.1284***	0.1443***	0.1487***	0.1468***	0.1787***	0.2246***
	(0.0486)	(0.0394)	(0.0397)	(0.0408)	(0.0402)	(0.0408)	(0.0594)	(0.0604)
adjusted R-square	0.304	0.430	0.44	0.445	0.446	0.446	0.315	0.463
Simple Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location Controls	No	No	Yes	Yes	Yes	Yes	No	No
Geographic Controls	No	No	No	Yes	Yes	Yes	No	No
Country Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	No	Yes
Observations	1212	1212	1212	1212	1199	1165	579	579

Michalopoulos and Papaioannou (2016) table 2 - replication

	Model 1	Model 2	Model 3	Model 4
split10pc	-0.012	0.086	0.053	0.047
	(0.177)	(0.115)	(0.106)	(0.105)
spil	0.322	0.514*	0.313	0.370+
	(0.284)	(0.226)	(0.215)	(0.213)
Num.Obs.	1212	1212	1212	1212
Std.Errors	by: cluster	by: cluster	by: cluster	by: cluster
FE: wbcode		X	X	X
split10pc	0.018	0.014	-0.010	-0.008
split10pc	(0.023)	(0.022)	(0.025)	(0.024)
spil	0.038	0.044	0.015	0.021
spil	(0.056)	(0.047)	(0.050)	(0.052)
Num.Obs.	1212	1212	1212	1212
Std.Errors	by: wbcode & cluster	by: wbcode & cluster	by: wbcode & cluster	by: wbcode & cluster
FE: wbcode		X	X	X

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Acemoglu et al. (2014) table 5 - original

	Dependent Variable			
	Weight for Height Z-Score		Moderate to Severe Anemia	
	(1)	(2)	(3)	(4)
	A. Baseline Specification			
ln(number of ruling families)	.212	.211	−.099	−.091
	(.117)	(.117)	(.041)	(.040)
R^2	.045	.052	.055	.066
	B. Baseline Specification with Additional Geographic Controls			
ln(number of ruling families)	.189	.167	−.136	−.129
	(.127)	(.132)	(.039)	(.039)
R^2	.052	.059	.067	.077
Observations	1,521	1,519	1,423	1,421
Number of chiefdoms	116	116	114	114
District fixed effects	Yes	Yes	Yes	Yes
Mother controls	No	Yes	No	Yes

Acemoglu et al. (2014) table 5 - replication

	(1)	(2)	(3)	(4)
fam_num_ln	0.109	0.138	0.025	0.005
	(0.103)	(0.101)	(0.043)	(0.038)
Num.Obs.	2264	2264	2322	2322
R2	0.012	0.020	0.016	0.050
R2 Adj.	0.006	0.010	0.010	0.040
Std.Errors	by: CODE	by: CODE	by: CODE	by: CODE
fam_num_ln	0.083	0.108	0.025	0.005
	(0.113)	(0.112)	(0.043)	(0.038)
Num.Obs.	2264	2264	2322	2322
R2	0.016	0.023	0.016	0.050
R2 Adj.	0.007	0.010	0.010	0.040