EMPIRICAL METHODS FOR LOW-QUALITY DATA

by

MICHAEL JERMAN

A DISSERTATION

Presented to the Department of Economics
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2020

DISSERTATION APPROVAL PAGE

Student: Michael Jerman

Title: Empirical Methods for Low-Quality Data

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Economics by:

Alfredo Burlando	Co-chair
Trudy Ann Cameron	Co-chair
Shankha Chakraborty	Core Member
Dejing Dou	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2020

DISSERTATION ABSTRACT

Michael Jerman

Doctor of Philosophy

Department of Economics

September 2020

Title: Empirical Methods for Low-Quality Data

This dissertation presents methods for economic analysis in settings characterized by sparse data. In the first substantive chapter, I show that difference-in-differences estimators can be biased in the presence of treatment externalities. I then develop a model that accounts for these externalities, and estimate the model using data on Indian river pollution. I show that failure to account for treatment externalities can substantially bias estimates toward zero. I find significant reductions in measured pollution levels in the areas downstream of sewage treatment facilities when compared to untreated areas.

Next, I propose a universal method for disaggregating count statistics. The method is able to disaggregate regional statistics such as those collected by censuses or surveys. I demonstrate the algorithm by disaggregating Ugandan census counts of population, tabooda (kerosene lamp) usage, people consuming two or more meals per day, and subsistence farms counted at the subcounty level (the smallest administrative unit reported by the census). Out-of-sample validation suggests that the procedure performs similarly for each statistic and that out-of-sample errors are approximately mean zero throughout the distribution. When combined with nighttime light luminosity data, the disaggregated data can describe within-subcounty distributions of income and poverty. I find that this previously

unobserved within-subcounty inequality accounts for 39.3% of aggregate observed inequality. Next I show that the disaggregated census data can be combined with satellite-derived air pollution data to estimate pixel-level estimates of pollution exposure. I find that 22% of aggregate inequality in air pollution exposure is caused by within-subcounty inequality in exposure.

Finally, I analyze the allocation of environmental resources following India's general election of 1996. Electorally competitive cities in the Ganges Basin during this period were more likely to receive funding for pollution abatement from the federal government of India. These same cities were less likely to receive increased water pollution monitoring. The empirical findings are explained by forward-looking policymakers engaging in clientelism. I emphasize the need for dramatically increased water pollution monitoring along India's rivers and streams.

CURRICULUM VITAE

NAME OF AUTHOR: Michael Jerman

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Central Michigan University, Mount Pleasant, MI
University of Arizona, Tucson, AZ

DEGREES AWARDED:

Doctor of Philosophy, Economics, 2020, University of Oregon
Master of Science, Economics, 2015, University of Oregon
Master of Arts, Economics, 2013, Central Michigan University
Bachelor of Arts, Economics, 2005, University of Arizona

AREAS OF SPECIAL INTEREST:

Income Inequality
Environmental Economics
Development Economics
Political Economy

For Llewelyn

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

In much of the developing world, basic data on markets and economic well-being are unavailable to researchers. In the economics literature, this lack of information is usually addressed in one of two ways. First, much empirical research relies on census, survey, or remotely sensed data that is aggregated over a geographic region. These geographic regions are typically large, and therefore the available data are unable to account for potential heterogeneity that exists within the regions. Second, an emphasis on small-scale randomized control trials (RCTs) has dominated the field of development economics. This method involves costly on-the-ground implementation and monitoring of economic agents over a lengthy period of time.

It is not my objective to disparage these existing methods, as they have provided (and continue to provide) valuable insights. Instead, the purpose of this dissertation is to promote a middle-ground approach to matters of economic development and environmental economics. I present methods for utilizing existing observational data sources that describe economic well-being and the effectiveness of various environmental policies in the developing world.

In Chapter 2, I demonstrate how careful consideration of the data-generating process can yield robust causal estimates in complex environmental settings. Specifically, I analyze the effectiveness of sewage treatment plants constructed in India as part of the National River Conservation Plan on measured pollution levels in India's rivers. The typical difference-in-differences estimator is shown to be biased toward zero in this setting, but a spatial model that explicitly accounts for the downstream persistence of pollution levels provides more robust estimates

in the presence of these spatial spillovers. Using a novel spatial model that takes advantage of the unique data environment, I show that pollution levels are observably lower after the construction of sewage treatment plants. These results contrast with the previously existing literature and with popular perception in India, which both suggest that sewage treatment plants are ineffective at reducing pollution.

Chapter 3 proposes a general algorithm to disaggregate economic and demographic statistics. Most data in the developing world are aggregated over some geographic region, such as census counts reported for census-defined districts or survey data that are representative at the level of some political jurisdiction (such as state or county). I show that freely available satellite data can be utilized to disaggregate count statistics within the geographic region over which the data are originally reported—a method known as dasymetric mapping. The algorithm is a modification of an expectation-maximization algorithm that utilizes a nonlinear random forest model to estimate the relationship between raw satellite imagery and the pixel-level counts. I demonstrate that this nonparametric approach is extremely flexible and is capable of disaggregating a variety of count statistics from the Ugandan census of 2014, including population, households whose primary source of lighting is kerosene lamps, subsistence farmers, and counts of people who consume two or more meals per day. I also show that the disaggregated population data can be combined with other satellite-derived resources to create high-resolution estimates of per-capita economic activity (as proxied by nighttime light luminosity) and per-capita pollution exposure. These estimates suggest a high degree of within-region heterogeneity of important economic variables in Uganda that were previously unobserved across an entire national region.

Chapter 4 returns to the India setting to analyze vote-buying behavior following the parliamentary election of 1996. This election was historically significant in that it ushered in a new era of political instability and was the first of four highly competitive elections that would occur within the next five years. This instability created numerous incentives for vote-buying and other clientelist policies. I find that cities in parliamentary districts that had competitive electoral outcomes (defined as an electoral margin of victory less than 3%) were more likely to receive federal funding for pollution-abatement projects. These findings are robust to a variety of specifications and the inclusion of additional variables that control for the economic, geographic, and political characteristics of this set of cities. These results indicate that policymakers should consider non-environmental factors when allocating environmental resources. However, the environmental and economic cost of this inefficiency is unknown. I conclude that marginal data-collection efforts should emphasize pollution monitoring.

CHAPTER II

MEASURING THE EFFECTS OF POLLUTION ABATEMENT IN THE PRESENCE OF TREATMENT EXTERNALITIES: AN EXAMPLE USING INDIAN DATA

Introduction

The differences-in-differences estimator is commonly used when estimating the average treatment effect of a policy change. However, spillovers (or externalities) occur when treatment affects observations that are not directly targeted by the policy. These spillovers can contaminate the control group in a differences-in-differences setting, therefore leading to biased estimates of the true treatment effect. In this paper, I address this issue in the context of river pollution abatement in India. I find that failing to account for treatment spillovers can substantially bias differences-in-differences estimates in this setting, leading to conclusions that differ from those reported in Greenstone and Hanna (2014) , with different implications for policy.

There are 14,780 sewage treatment plants in the United States (The Center for Sustainable Systems, 2015). India, with four times the population, has just 234 (Mauskar, 2008). In 2011, India's Ministry of Environment and Forests estimated that existing sewage treatment plants have the capacity to treat less than one third of all municipal sewage. In most cases, untreated sewage is discharged directly in to a river or other body of water.

The dangerous pollution levels in India's waterways constitute a public health disaster. Millions of Indian citizens rely on polluted rivers for drinking water, bathing, and cleaning. In addition, many rivers are considered holy in

the Hindu faith and are believed to have purifying effects. Each year, millions of pilgrims ceremonially drink from, and bathe in, polluted rivers.

India's government is aware of the problem and has taken steps to address it. Beginning in 1985 under the Ganga Action Plan, 865 billion USD has been allocated for reducing river pollution across the country. However, it is widely held that the various cleanup schemes implemented by the government have been a failure (The Hindu, 2004). In a 2009 interview, Minister of State for Environment and Forests Jairam Ramesh stated that widescale changes to the government's approach were needed or else resources would "continue to be wasted" (S. Yadav, 2009).

Despite this popular perception, there has been no systematic study of the effectiveness of any specific policy intervention at the national level. This paper addresses this gap by estimating the effect of new sewage treatment plant (STP) construction on pollution levels in India's rivers. Over one hundred new STPs were built as part of the National River Conservation Plan (NRCP) in 1995 at various locations around India. I estimate the effects of these STPs on pollution levels using data from the Central Pollution Control Board's network of river monitoring stations. In a preview of the main results, I find that STP construction is associated with a small but significant decrease in pollution levels, but the effects diminish in the long run.

The findings presented here are directly relevant to Indian policymakers. Despite the perception of waste and ineffectiveness, I find that STP construction can potentially reduce pollution to levels that are considered safe for bathing and recreational activity.

The baseline result is shown using a simple differences-in-differences approach implemented with OLS, as is standard in the development literature (Greenstone & Jack, 2015). However, the existence of treatment externalities confounds identification in unpredictable ways. Aquatic pollutants mix non-uniformly across space; pollution generated at one point along a river will be present downstream, but natural remediation will work to reduce pollution levels as the downstream distance increases. The existence of a sewage treatment plant reduces pollution at the closest downstream monitoring station, but also reduces pollution at all subsequent downstream monitoring stations. Failing to account for treatment externalities can bias differences-in-differences estimators (Miguel & Kremer, 2004).

Error processes may persist downstream as well. Any shock that directly affects pollution levels measured at an upstream monitoring station will affect all downstream monitoring stations. Failing to account for this downstream error process can result in distorted inferences. I therefore develop a spatial econometric model which allows both treatment effects and errors to persist downstream (Anselin, 2013; LeSage & Pace, 2009). In a novel contribution to the spatial literature, I specify an exponential spatial lag to capture the downstream effects of pollution. Under this parameterization, the spatial model is equivalent to nonlinear least squares with exponential decay of the dependent variable. This model is likely to have wide-ranging applications in environmental and development settings, in situations where treatment effects diminish exponentially over distance or time.

In the next section, I review related research in the environmental and development literature. I describe the data in Section II. Section II presents OLS estimation results. As outlined above, the covariance matrix of simple OLS models

is likely misspecified if pollution persists downstream. The novel spatial model that accounts for this error process is presented and estimated in Section II. I discuss the policy significance of the results in Section II before concluding with Section II.

Related Literature

Greenstone and Jack (2015) highlight the importance of identifying (a) the mechanisms that contribute to high pollution levels in the developing world, and (b) which policy measures are effective at reducing exposure to pollution. They emphasize that pollution abatement depends not only on the marginal willingness to pay for abatement for those who are affected by pollution, but also on the institutional environment that builds and maintains the necessary abatement infrastructure. Large utility gains may go unrealized due to weak institutions; resources may be misallocated to the extent that abatement measures are never undertaken.

Water pollution is a huge problem in India, and the ability of the Indian government to implement efficient water pollution abatement measures is analyzed in Greenstone and Hanna (2014). They utilize a differences-in-differences approach to estimate the average effect on pollution levels when certain cities are designated as "problem areas" by the Indian government, relative to cities that receive no such designation. They find that the problem-area designation is not associated with any decrease in pollution levels, a result they attribute to the weak institutional environment in India. The Greenstone and Hanna result serves as the starting point for this paper. Here, I focus on a specific policy intervention involving the construction of new STPs—which allows for a more precise identification of the institutional shortcomings. In particular, I find that while STPs are initially

effective at reducing pollution levels, these effects are short-lived, likely due to the depreciation of physical and human capital.

In a similar study, Lipscomb and Mobarak (2015) use data from a Brazilian water-pollution monitoring network to estimate downstream pollution spillovers. They leverage exogenous variation in shifting county-border locations between upstream-downstream station pairs (as a result of administrative redistricting) to find that pollution generation increases as borders downstream of monitoring stations move upstream; decentralized regulation and abatement authorities do not internalize the downstream effects of pollution and abatement. Each station-pair is treated as a unique observation in the data, and the authors apply OLS with standard heteroscedastic-robust standard errors. As I show in Section II, this assumes that the errors are unrelated across station pairs, an assumption that is unlikely to hold along river systems.

In contrast to Lipscomb and Mobarak, I estimate the effects of a specific policy intervention (i.e. the construction of new sewage treatment plants). However, most of the existing literature on abatement focuses on air pollution, rather than water pollution. For example, Fowlie et al. (2012) analyze the effects of cap-and-trade schemes on pollution levels compared to the effects of command-and-control regulation. They exploit spatial variation in the regulatory environment to identify a significant decrease in pollution levels associated with the cap-and-trade policy.

Using similar methodology, Auffhammer and Kellogg (2011) estimate the effects of gasoline regulation on air pollution. Identification is based on spatial variation in gasoline content laws across US states aimed at reducing ozone created from volatile organic compound emissions. The heterogeneity of implementation

across states allows them to identify which type of abatement measure is more effective. They find that measures aimed at the specific volatile organic compound emissions (those that are more likely to create ozone) are more effective than general quotas on emissions, because general quotas allow polluters to substitute toward other ozone-creating processes.

The effects of heterogeneous abatement compliance identified in these previous studies are particularly acute in developing countries such as India. The estimates in this paper suggest considerable heterogeneity across time in abatement effectiveness. In Section II, I present evidence that a lack of operations oversight, as well as capital depreciation, are likely to blame, echoing the types of findings in the existing environmental literature.

The importance of using quasi-experimental settings when analyzing abatement technologies is highlighted by Dominici et al. (2014). Identification based solely on spatial variation in a policy intervention is generally not sufficient to ensure unbiased estimates of a causal effect, since policies may be non-randomly implemented. In the present context, the locations of new sewage treatment facilities are likely to be non-random, but for downstream beneficiaries, the effects are less likely to be endogenous. For example, the set of municipalities that construct new STPs will likely share unobserved characteristics that impact pollution levels. However, the pollution-generating processes downstream of these cities are plausibly random and independent. If municipalities that discharge large amounts of raw sewage into a river are targeted, other downstream municipalities will derive benefit from the construction of an upstream STP, regardless of the sewage generation and existing abatement technology present in the downstream locations. As long as the unobserved characteristics common to locations with new

STPs are orthogonal to the unobserved characteristics at the downstream location of each monitoring station, Dominici et al.'s exogeneity condition is likely satisfied.

The downstream effects of sewage treatment introduce spillovers into a differences-in-differences estimation environment. In a different context, Miguel and Kremer (2004) estimate the effects of deworming medication for school children Kenya in the presence of analogous spillovers. Children at schools that are randomly selected for treatment influence the health of other children in their neighborhoods—siblings and friends from untreated schools benefit from lower hookworm and roundworm infection rates among their peers. The authors control for this by including the total number of children who live within a certain distance of the schools and the number of treated children within that same distance. This specification accounts for potential bias in point estimates due to treatment spillovers, but the estimated parameter variance-covariance matrix does not account for the same spatial dependence.

The potential for bias when estimating covariance matrices is particularly problematic in an environmental setting. Both pollution generation and abatement have the potential to influence multiple locations (or individuals) simultaneously. Spatial econometric techniques can explicitly account for for this type of simultaneity, but they remain underutilized in both the environmental and development literature.

Data

India's Central Pollution Control Board (under the umbrella of the Ministry of Environment and Forests) maintains a network of 447 river-monitoring stations throughout India. Water samples are taken at specific time intervals (monthly or quarterly), and then sent to regional laboratories for analysis. Data on various

pollutants measured at each station were compiled by Greenstone and Hanna (2014) . Their sample covers the period from 1986 to 2004.

These data also contain information on the geospatial location of each monitoring station. The geospatial information consists of a low-resolution coordinate (accurate to one minute of latitude and longitude—69 miles at the equator) and a brief physical description of the site (for example, the name of the bridge or public beach where the water samples are taken). I use this information in conjunction with Google Maps and OpenStreetMap to pinpoint the exact location of each pollution monitoring station.[1] I then use the coordinates from these locations to calculate the great-circle distance between stations. Each monitoring station can then be plotted using OpenStreetMap, allowing the exact upstream/downstream relationship between stations to be identified.

The 477 monitoring stations in India are located along 60 separate river systems (see Table 1). Each river system therefore has an average of 7.45 stations. For 28 of these river systems, there is only one monitoring station, while the largest river system (Ganga) has 109 stations.

Variable	Mean	St. Dev.	Min	Max
Log of Fecal Coliforms	5.665	2.918	0.000	14.560
Distance to nearest upstream station†	52.410	61.178	0.278	385.700
Stations per river system	7.450	17.476	1	109

†Measured in kilometers. Upstream distance calculated only for stations with upstream neighbors. Of the 447 monitoring stations, 148 have no associated upstream station.

Table 1. Descriptive Statisics

Not all monitoring stations are active during every year of the sample period (Figure 1). New stations are continually added to the network, while some are

[1]Some monitoring stations could not be pinpointed using this method. In this case, the river location nearest to the centroid of the coordinate range was chosen. This strategy was implemented for fewer than 5% of all monitoring stations in the data.

removed. The overall trend is toward more monitoring stations; there are 114 active stations in 1986, and 398 in 2004. As a result, the upstream/downstream relationship between stations is not constant. Distance to the closest upstream station can decrease as new stations are added, and can increase if data for the closest upstream station are not reported in that time period.

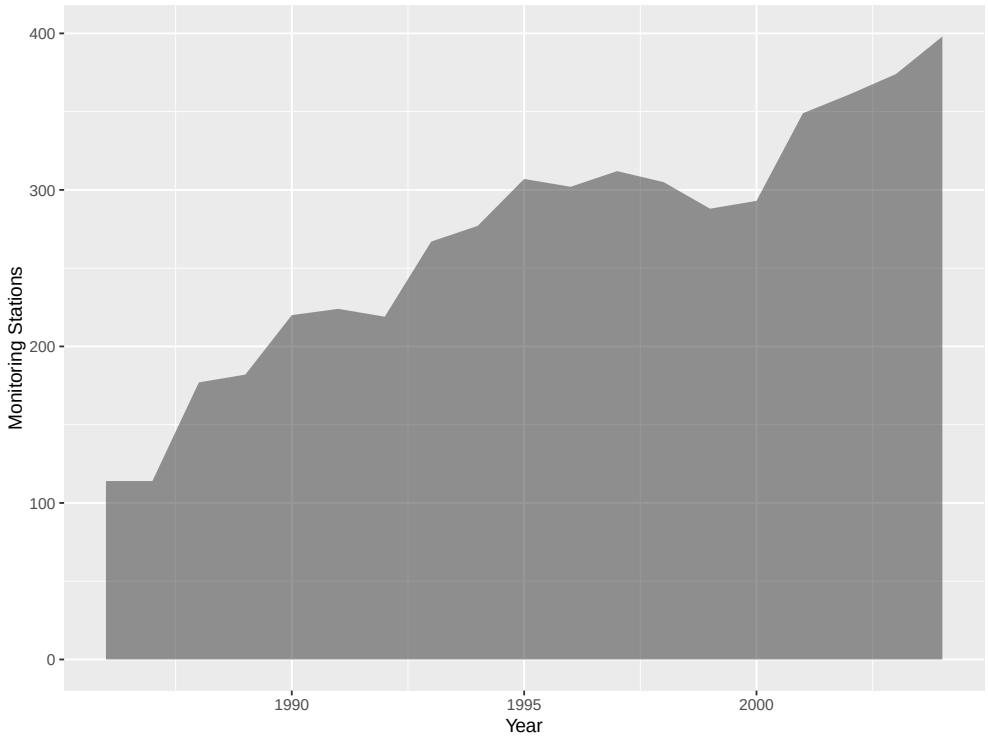


Figure 1. Number of Monitoring Stations by Year

Figure 2 shows the frequency of distances for each station's closest downstream neighbor (excluding those with no downstream neighbors). Roughly 62% of all monitoring stations have another station within 50 kilometers downstream. Any pollution process with even a modest downstream persistence is therefore likely to have its effects be observed at multiple stations in the sample, potentially biasing a differences-in-differences estimator.

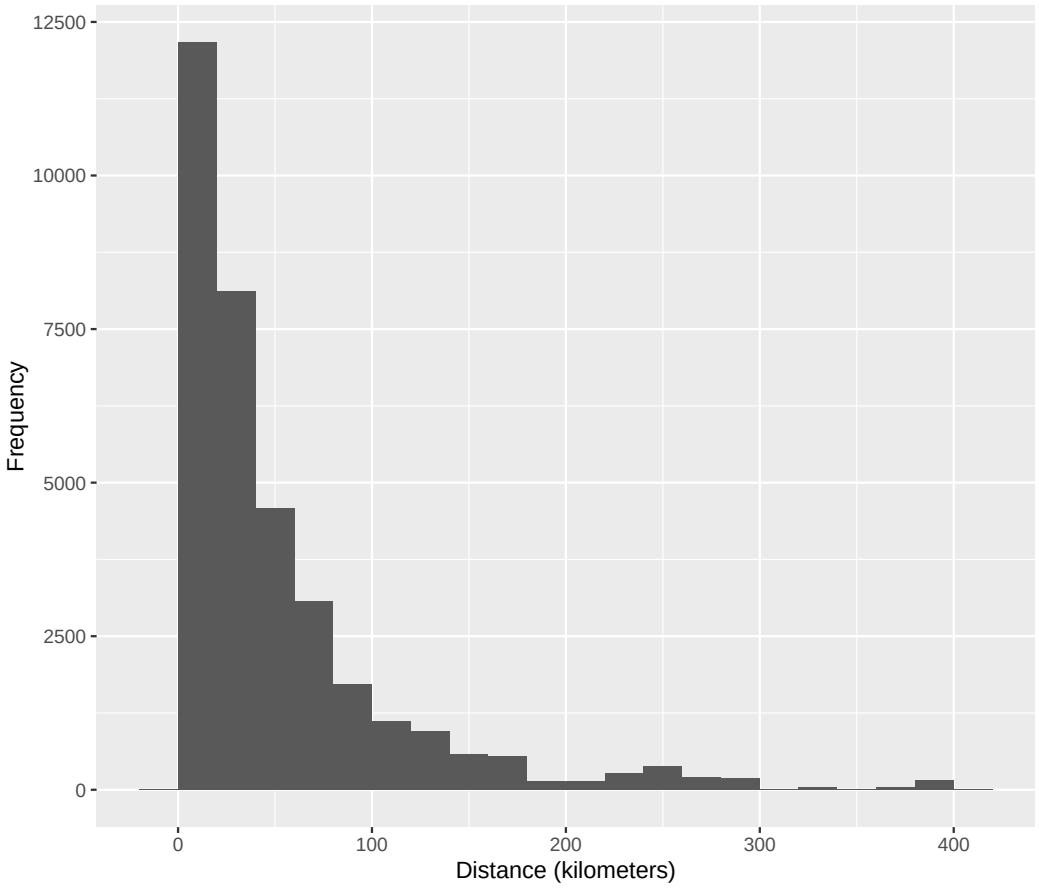


Figure 2. Frequency of Downstream Distances Between Stations

The water samples at each monitoring station are tested for a variety of pollutants and pollution indicators. The indicator utilized in this paper is the count of fecal coliforms (fcoli) in the water. The advantages of this indicator are threefold. First, fcoli is one of the most commonly used water pollution indicators in the environmental science literature, and it is well understood to be closely correlated with the overall health of a water body. Second, fcoli is the most frequently reported pollution indicator in the sample. Lastly, fcoli levels are a direct measurement of the amount of untreated sewage in the water.

Fecal coliforms are a class of coliform bacteria that grow exclusively in the digestive tracts of mammals.[2] Sewage treatment eliminates the food source, so properly treated wastewater therefore has very low levels of fcoli. The levels of fecal coliforms remaining in the water are thus the primary indicator for the effectiveness of sewage treatment.

The counts of fcoli are generally reported in natural logs, a convention I follow in this paper. In the laboratory, water samples are added to an fcoli growth medium and left to incubate for a fixed period, after which the number of colonies can readily be counted. The colony growth over the incubation period is exponential, so taking the natural log allows for more robust comparisons of the actual number of coliforms present in the water source.

Figure 3 shows the distribution of fcoli in the sample, across all measurements at all stations. The dashed, vertical line represents the United States Environmental Protection Agency's "statistical threshold value" (STV) for fcoli levels in recreational water (similar regulatory criteria for India, if they have been established, are not publicly available). At this cutoff, recreational users of the water are likely to fall ill due to fcoli contamination at an estimated rate of 32/1000 (EPA, The Environmental Protection Agency, 2012). Above this level, the EPA deems a body of water to be unsafe for recreational activity.

Nearly half (49%) of all fcoli measurements in the sample are above the EPA's STV. Simply bathing in an Indian river carries a substantial risk of contracting a waterborne illness. Since millions of Indian citizens also rely on river water for cooking, cleaning, and drinking, the fcoli levels in India's rivers represent a significant public health crisis.

[2]Most forms of fecal coliforms are not harmful to humans, with a few notable exceptions. For example, *escherichia coli* (E. coli) is a form of fecal coliform that can cause fatal illness.

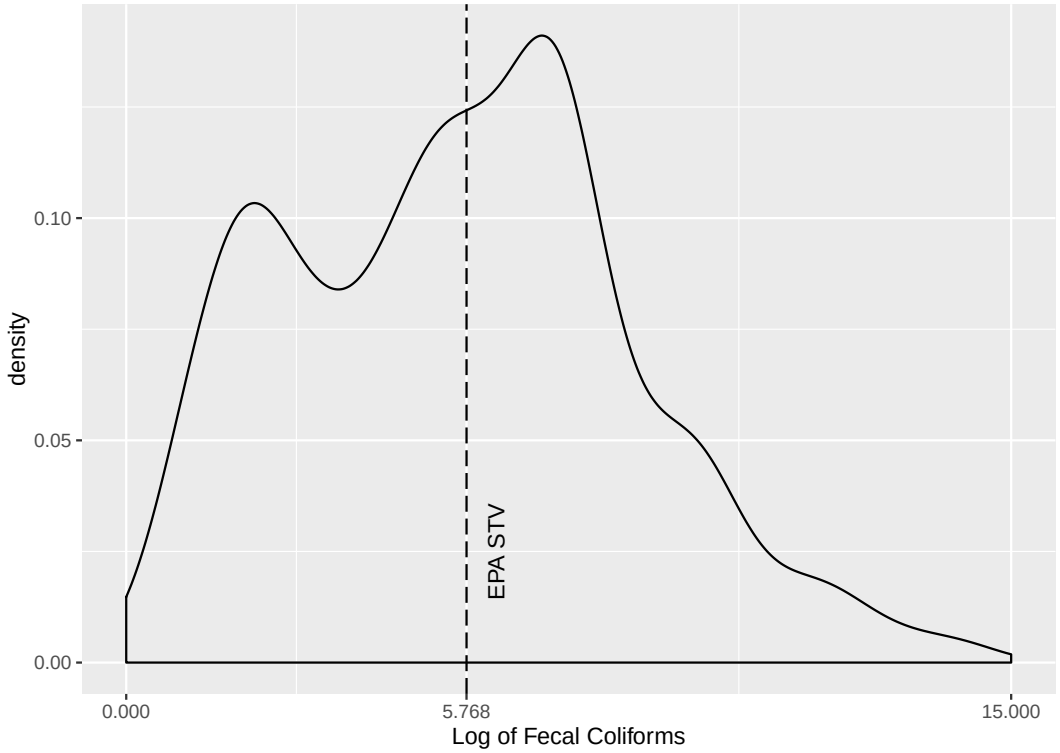


Figure 3. Sample Kernel Density Estimate of Fecal Coliforms

One of the primary objectives of the National River Conservation Plan was the construction of new sewage treatment facilities. In 1995, 107 treatment plants were built along India's rivers (Tyagi, 2013). The location of each of these STPs was obtained from the Ministry of Environment and Forests and satellite photos from Google Earth. The spatial distributions of monitoring stations and STPs are shown in Figure 4.

The effect of STP construction on pollution levels is identified based on the location of the STPs. Fcoli counts at monitoring stations upstream of STPs are compared to counts at monitoring stations downstream from STPs, before and after construction of the STP is completed. The wide geographic dispersion of both monitoring stations and STPs helps ensure that the estimates are representative for India.

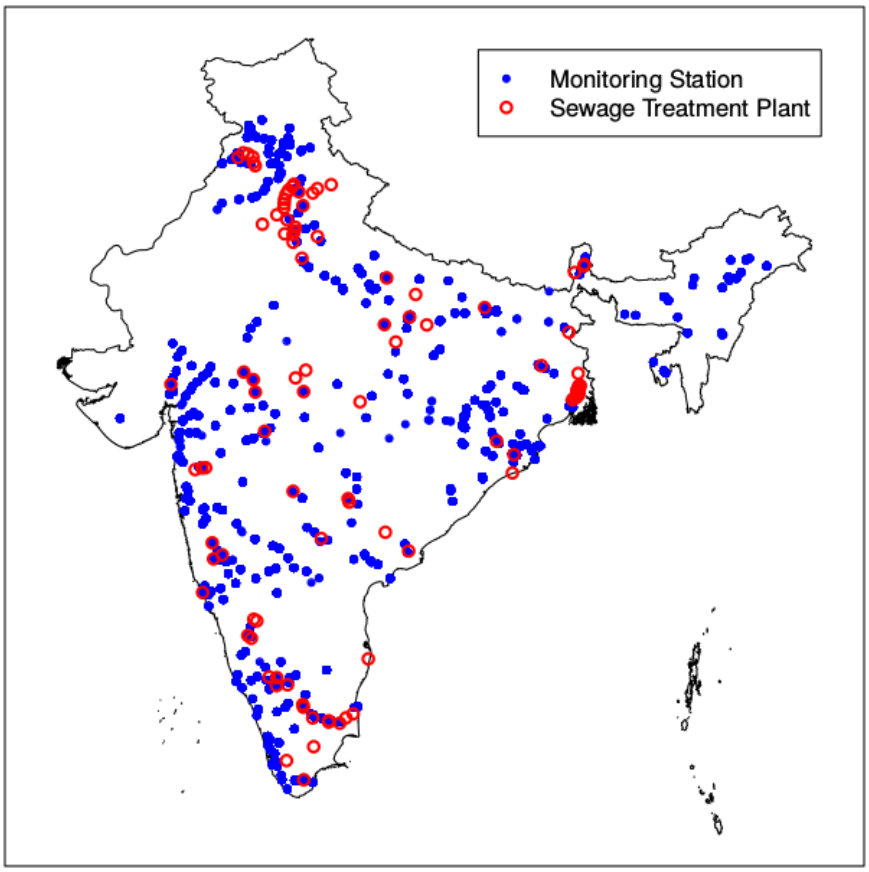


Figure 4. Sewage Treatment Facilities and Monitoring Stations

OLS Results

Least-squares estimation can recover unbiased estimates of treatment externalities if the spatial structure is properly specified. Robust standard-error estimation, however, requires the variance-covariance matrix be properly specified as well. In this section, I ignore the potential problems with estimating the covariance structure and consider a simple, parsimonious model. Despite the potential for biased standard errors, this estimation strategy is common in the literature. The results reported in this section are best viewed as a baseline. In Section II, we will see that the OLS estimates are quantitatively similar to those

obtained from the more robust spatial model, though the OLS estimates of the long-run effect are likely biased toward zero.

Naïve Difference-in-Differences. Following the approach used in many development and environmental settings (Greenstone & Jack, 2015), I specify a simple difference-in-differences model as a baseline. This approach measures the treatment effect on the treated stations relative to untreated stations. This model is written as:

$$P_{it} = \beta_1 T_{it} + \beta_2(t \cdot D_i) + \beta_3(t \cdot T_{it}) + \tau_t + \psi_i + \varepsilon_{it} \tag{2.1}$$
$$\varepsilon_{it} \sim N(0, \sigma_i^2),$$

where P_{it} is the measured pollution level at station i at time t, D_i is a dummy variable indicating whether station i is treated at any time in the sample, and T_{it} is a dummy variable equal to 1 if station i is treated in time t. The parameters τ_t and ψ_i are station- and time-specific fixed effects which control for any unobserved sources of heterogeneity that are constant across time and stations. The term $\beta_2(t \cdot D_i)$ allows for pollution levels at treated stations to evolve on a differential trend from that of the untreated stations, while $\beta_3(t \cdot T_{it})$ allows for a structural break in that trend at the treatment date.

In a "naïve" specification, each station is considered as treated if it is located immediately downstream of an STP, and as untreated otherwise. This specification will capture the first-order effect of sewage treatment on the downstream station relative to all other stations. The indirect effects of treatment on other stations downstream of an STP are assumed to be zero. Later, however, this assumption will be relaxed.

Table 2 shows the results of the naïve differences-in-differences model (equation (2.1)). Column 1 reports the treatment effect with no allowance for

	Dependent variable:					
	Log of Fecal Coliforms					
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Effect (β_1)	-0.388^{***}	-1.266^{***}	-1.165^{***}	-0.873^{***}	-0.759^{***}	-0.773^{***}
	(0.066)	(0.113)	(0.115)	(0.076)	(0.118)	(0.118)
Pre-treatment Trend (β_2)		0.008^{***}	0.004^{**}		-0.002	0.004^{**}
		(0.001)	(0.002)		(0.001)	(0.002)
Post-treatment Trend (β_3)			0.006^{***}			-0.013^{***}
			(0.002)			(0.003)
Post 5 Year Effect (β_1^L)				0.995^{***}	1.085^{***}	1.499^{***}
				(0.073)	(0.105)	(0.129)
Long-run Effect ($\beta_1 + \beta_1^L$)				0.122^{*}	0.326^{*}	0.7256^{***}
				(0.073)	(0.188)	(0.206)
R^2	0.620	0.621	0.621	0.622	0.622	0.622
Adjusted R^2	0.613	0.614	0.614	0.615	0.615	0.615

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedastic-robust standard errors in parentheses. All regressions have 38,625 observations and contain time and station fixed-effects. Estimation results for equation (2.1) are in columns 1-5. Equation (2.3) is reported in column 6.

Table 2. Naïve Differences-in-Differences Estimates

differential trends between the treated and untreated groups ($\beta_2 = \beta_3 = 0$). The estimated treatment effect is small but statistically significant, indicating that the average pollution levels at the stations immediately downstream of an STP are lower, on average, throughout the post-treatment period relative to the pre-treatment period.

Columns 2 and 3 of Table 2 report the estimates for the model with a pre-treatment trend (column 2) and with both pre- and post- treatment trends (column 3). Over the entire sample period, stations immediately downstream of an STP have a positive and significant trend in their pollution levels relative to other stations (β_2, the coefficient on $t \cdot D_i$). This result can be explained by noting that the placement of STPs is nonrandom—it is likely that areas with increasing sewage generation were prioritized for intervention by policymakers.

Specifying linear trends for the treated stations results in a much larger (in magnitude) estimated initial treatment effect—that is, the treatment effect in the first year following treatment. Furthermore, the slope of the post-treatment trend line ($\beta_2 + \beta_3$) is steeper than the pre-treatment trend ($\beta_3 > 0$). Taken as a whole, the parameter estimates in columns 2 and 3 of Table 2 indicate that there is estimated time-heterogeneity in the treatment effect. Large, initial reductions in pollution are offset over time by an increase in the pollution trend of treated stations.

The time-heterogeneity of the treatment effect can be analyzed by interacting the treatment variable with a full set of year dummies:

$$P_{it} = \sum_{\tau} \beta_\tau T_{i\tau} + \delta_t + \psi_i + \varepsilon_{it} \tag{2.2}$$

$$\varepsilon_{it} \sim N(0, \sigma_i^2),$$

where $T_{i\tau}$ is a dummy equal to one if station i is treated in year τ. The sequence $\{\hat{\beta}_\tau\}$ nonparametrically describes the time trend in treated stations relative to untreated stations. These coefficients can be interpreted as treatment-specific year fixed effects.

Figure 5 plots the estimated $\hat{\beta}_\tau$'s along with their 95% confidence intervals. Due to the linear dependence in equation (2.2) (one dummy variable for each year in the sample), the year 1994 (the year before the STPs are built) is specified as the base year. Each $\hat{\beta}_\tau$ therefore measures the average level of pollution relative to the level measured in 1994. The results are striking—stations immediately downstream of sewage treatment facilities show a large, immediate decrease in their measured pollution levels. But the effect is short-lived: pollution levels begin to rise after three years, and are statistically insignificantly different from the pre-treatment levels just six years after the STPs are built. The slight upward trend in pre-treatment means reported in Table 2 can also be observed, although the coefficients on the indicators for each of the years immediately preceding treatment are statistically indistinguishable from one another.

The results presented in Figure 5 motivate the estimation of a specification that distinguishes between short-run and long-run treatment effects:

$$P_{it} = \beta_1 T_{it} + \beta_1^L T_{it}^L + \beta_2(t \cdot D_i) + \beta_3(t \cdot T_{it}) + \tau_t + \psi_i + \varepsilon_{it} \tag{2.3}$$

The new variable T_{it}^L is an indicator equal to one if station i is treated and t is more than five years after the initial treatment date.

The estimation results are reported in Columns 4, 5, and 6 of Table 2. The estimated coefficient on T_{it}^L is positive and significant in all specifications. In fact, the estimated rise in pollution levels at treated stations between 1999 and 2000 is large enough to overcome the initial reduction between 1994 and 1995. The

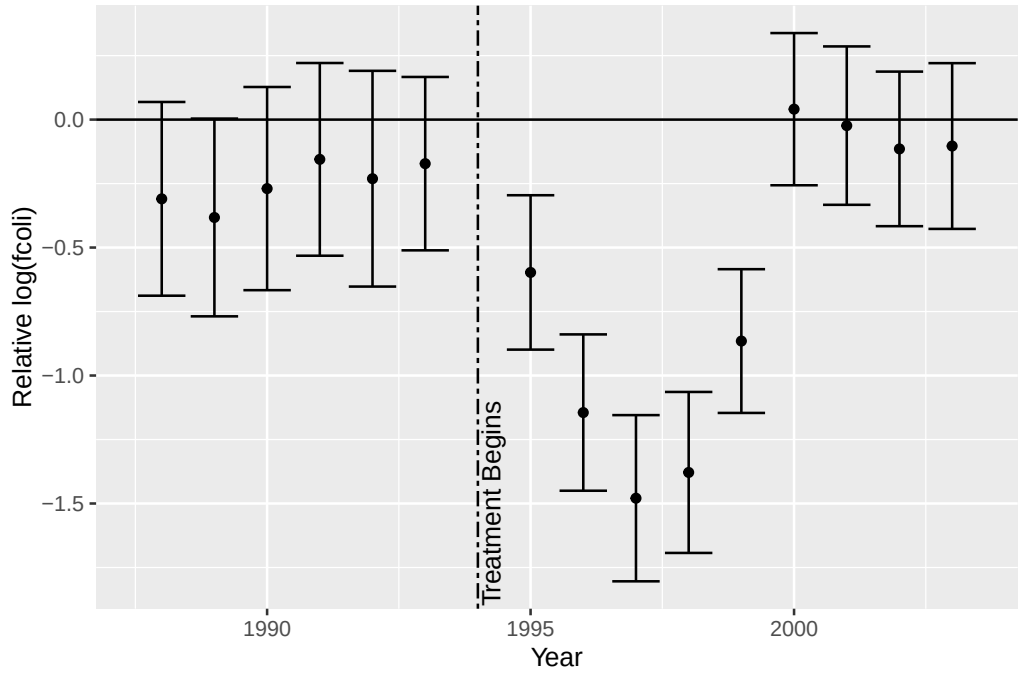


Figure 5. Treatment Effect by Year Relative to 1994

estimated level effect of treatment after five years is given by $\beta_1 + \beta_1^L$, shown in the last row of Table 2. This effect is estimated to be positive and significantly different from zero (at the 10% level) in all three specifications.

These results offer one possible explanation as to why the effects of the NRCP have not been previously observed. Studies comparing pollution levels a few years after the construction of an STP might find little to no effect, despite the substantial decrease immediately following treatment. The potential reasons for this mean-reverting behavior of measured pollution levels will be discussed in Section II.

Treatment Externalities. The identifying assumption underlying the naïve specifications in equations (2.1) and (2.2) is that the construction of an STP impacts pollution levels only at the nearest station immediately downstream. As discussed in Section II, this assumption is unlikely to hold. In particular, other

stations farther downstream from the "treated" stations may also see a reduction in pollution levels, though to a smaller extent than stations closer to the STPs (if there is some recovery or a dilution process that occurs naturally). Estimated treatment effects are likely biased in the presence of these treatment externalities, since this effect amounts to a contamination of the control group by the treatment.

To account for this bias I consider the alternative specification:

$$P_{it} = \sum_{s} \Phi_n(d_{si})\widetilde{T}_{sit} + \tau_t + \psi_i + \varepsilon_{it} \tag{2.4a}$$

$$\Phi_n(d_{si}) = \beta_0 + \beta_1 d_{si} + \beta_2 d_{si}^2 + \cdots + \beta_n d_{si}^n \tag{2.4b}$$

$$\varepsilon_{it} \sim N(0, \sigma_i^2).$$

The variable d_{si} measures the downstream distance of monitoring station i from STP s. $\Phi_n(d)$ is an n^{th}-order polynomial of downstream distance (in kilometers), which allows for treatment effects to vary over distance. $\widetilde{T}_{sit}$ is new treatment indicator equal to one if station i is anywhere downstream of STP s in time t. For each monitoring station i, the distance polynomial is summed over all upstream STPs. Each station can therefore be treated by multiple STPs simultaneously.

The above specification considers all stations downstream of an STP as treated. The intensity of treatment thus varies according to the number of upstream STPs and the downstream distance of each monitoring station from each STP (connected by a river system). In contrast to the model in equation (2.1), this specification explicitly accounts for treatment externalities. Identification is based on the number of upstream STPs and the variation in distances between each station and each upstream STP.

	Dependent variable:				
	Log of Fecal Coliforms				
Variable	(1)	(2)	(3)	(4)	(5)
$\widetilde{T}_{sit}$	-0.227^{***}	-0.454^{***}	-0.553^{***}	-0.338^{***}	-0.418^{***}
	(0.020)	(0.026)	(0.046)	(0.055)	(0.024)
$d \cdot 1000$	0.315^{***}	1.422^{***}	2.376^{***}	-0.981	0
	(0.027)	(0.082)	(0.375)	(0.615)	
$(d \cdot 1000)^2$		-0.808^{***}	-2.607^{***}	9.296^{***}	6.412^{***}
		(0.056)	(0.690)	(1.804)	(0.485)
$(d \cdot 1000)^3$			0.892^{***}	-13.553^{***}	-10.505^{***}
			(0.340)	(1.991)	(0.935)
$(d \cdot 1000)^4$				5.536^{***}	4.484^{***}
				(0.736)	(0.448)
Polynomial Order	1	2	3	4	4
AIC	156214.1	156009.2	156002.5	155949.5	155950.6
BIC	162010.3	161814	161815.8	161771.4	161763.9
R^2	0.621	0.623	0.623	0.624	0.624
Adjusted R^2	0.614	0.616	0.616	0.617	0.617

Notes: *p<0.1; **p<0.05; ***p<0.01. Heteroskedastic-robust standard errors in parentheses. Coefficient estimates for distance measures (d_{si}) are scaled by a factor of 1,000km. All specifications have 38,625 observations and include time and monitoring-station fixed-effects. Estimation results correspond to equation 2.4a.

Table 3. Downstream Abatement Effects

Estimation results are given in Table 3. Column 1 shows the results of a first-order polynomial specification (i.e. where the downstream effects are assumed to be linear in distance). Both estimated polynomial coefficients have the expected sign—initial decreases in pollution immediately downstream of an STP ($d = 0$) that dissipate as distance increases.

It is unlikely that downstream pollution effects are linear in distance. Column 2 corresponds to a polynomial order of two ($\Phi_2(\cdot)$), the most parsimonious specification that allows for nonlinearities. All of the coefficients are statistically significant at the 0.01 level. The "treatment effect" estimate, $\hat{\beta}_0$, is the predicted

change in pollution levels at the location where an STP is constructed (downstream distance is 0 km). The point estimate of -0.454 is more than 10% larger in magnitude than the simple differences-in-differences estimate reported in Table 2. This indicates that failing to account for treatment externalities can substantially bias point-estimates toward zero.

The polynomial specification also allows for the downstream effects to be analyzed. Note that

$$\frac{\partial \widehat{P}^2_{s,t}}{\partial d_{i,s,t} \partial \widetilde{T}} = \Phi'_n(d_{i,s,t}). \tag{2.5}$$

The treatment effects diminish as downstream distance increases only if $\Phi'_n(d_{sit}) > 0$. As expected, this condition is satisfied with the second-order polynomial $\Phi_2(\cdot)$ in column 2—treatment effects appear to be larger when a station is nearer to an STP.

The third-order polynomial of column 3 shows an even larger estimated treatment effect than the second-order specification. In contrast, the fourth-order results in column 4 differ substantially. Of particular interest is the non-monotonic downstream treatment effect. ($\Phi'(d_{sit}) < 0$ when $d < 60.5$). This specification predicts an initial *increase* in the magnitude of the estimated treatment effect over the first 60.5 kilometers downstream. This result is driven entirely by $\widehat{\beta}_1$, the coefficient on d_{sit}. Note that estimated sign is negative, indicating that the minimum of the polynomial function occurs when d is positive. It is unlikely that the true treatment effect reaches its maximum 60.5 km downstream of an STP rather than immediately downstream. This observed behavior in the fourth-order polynomial is probably due to overfitting.

The treatment effects estimated here reflect the average impact of abatement. There is some heterogeneity across rivers and STP locations—

rate of flow and river volume are unfortunately not observed in the data. This heterogeneity may influence estimated downstream effects. Swift-moving water may "carry" the treatment effect downstream more effectively than stagnant rivers. The high-order polynomial specifications may capture some of this heterogeneous effect, thus giving rise to unexpected (but statistically insignificant) non-monotonicity.

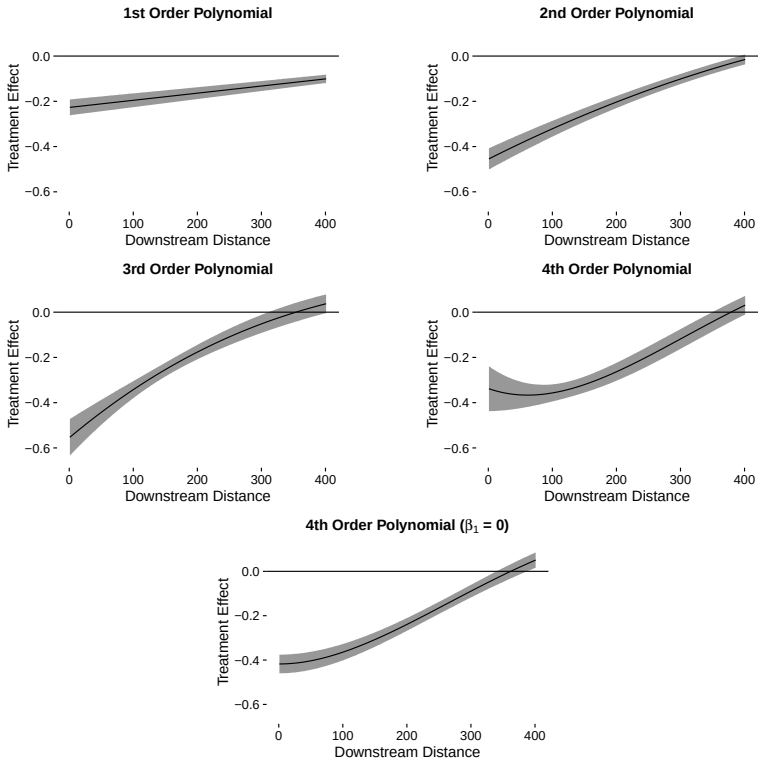
One way to address this non-monotonicity is simply to drop the first-order term d_{sit} from the regression ($\beta_1 = 0$), since the estimated coefficient for this variable is not significantly different from zero anyway. Results of this estimation are reported in column 5 of Table 2. The results are qualitatively similar to the other specifications, and all covariates are significant at the 1% level. Furthermore, the BIC prefers the specification without the d_{sit} term (column 5) to the specification where it is included (column 4). The AIC, which is less sensitive to additional parameters, shows only a marginal preference for the more inclusive model. A monotonic downstream effect can therefore be estimated with little, if any, loss in the explanatory power of the model.

The statistical significance of the downstream effect at any arbitrary distance d can be ascertained by performing a simple F-test on the linear hypothesis

$$H_0 : \Phi_n(d) = 0. \tag{2.6}$$

Figure 6 shows the estimated downstream effects of abatement for each of the specifications in Table 2, along with 95% confidence intervals. The estimates are similar for all of the non-linear specifications (polynomials of order 2 or above). In particular, the downstream effects do not become statistically indistinguishable from zero until 300 or 400 kilometers downstream, indicating a high degree of downstream persistence in the data. The non-monotonic downstream behavior

of the 4th-order polynomial specification is also visible, but the initially wide confidence interval indicates uncertainty about the true shape of the downstream effect over the first 100 kilometers. This initial uncertainty is substantially reduced when the parameter β_1 is set equal to zero, as in the last panel.



Note: Shaded regions represent 95% confidence intervals.

Figure 6. Treatment Effect by Downstream Distance

Spatial Model

The primary shortcoming of the OLS-based approach is its inability to capture the complexity of the error process. Throughout the previous section, it was assumed that errors were independent across monitoring stations, which is unlikely to be the case. This study is the first to find an abatement effect at the national level for STPs in India, so it is imperative to model the errors properly to ensure that the inference is not merely an artifact of a misspecified error structure.

Shocks to local pollution levels that are observed at one station are expected to be observed (to some extent) at all other downstream stations as well. Thus the error processes at stations along the same river are likely to be correlated. Moreover, the extent of this correlation should be smaller for stations separated by larger distances than for stations that are near to each other. This suggests that the commonly used cluster-robust standard errors are unlikely to be adequate, since within-cluster error covariances are not homogeneous for any level of clustering. For instance, error clustering at the river level assumes that monitoring stations located hundreds of kilometers apart along the same river will have the same error correlation as stations separated by just a few kilometers. Similarly, errors clustered at the individual station level assume that errors are independent across stations, even those stations located within the same metropolitan area.

In this section, I assume that the process that describes the downstream treatment effect is identical to the process that describes the downstream error process. This assumption says that the amount of pollution observed at a given distance downstream from a pollution source is the same regardless of the pollution source. Permanent changes in estimated pollution, such as the changes that would be observed after the construction of a new STP, impact downstream stations in the same manner as do transient changes, such as a holiday that decreases industrial production. In other words, the local effects of changes in the upstream pollution generation process are the same regardless of the source of the pollution. This type of spatial relationship can be captured by a "spatial autoregressive model" (Anselin, 2013; LeSage & Pace, 2009).

To illustrate the spatial process, consider two monitoring stations, labeled A and B, located on unconnected bodies of water. Each station has pollution

observations in time period $t = 1, 2$, and an STP is constructed immediately upstream of A at $t = 2$ (while station B remains untreated). The pollution at each monitoring station can be described by

$$P_{A,t} = \beta T_{A,t} + \tau_t + \psi_A + \varepsilon_{A,t}, \quad \varepsilon_{A,t} \sim N(0, \sigma_A^2) \tag{2.7}$$

$$P_{B,t} = \beta(0) + \tau_t + \psi_B + \varepsilon_{B,t}, \quad \varepsilon_{B,t} \sim N(0, \sigma_B^2) \tag{2.8}$$

Now suppose that a river (or canal) is "created" between A and B such that the two monitoring stations are connected by a waterway, so A is now d kilometers upstream of B. The pollution that is measured at A will also be present at some level at station B (now downstream of A). The amount of $P_{A,t}$ that is present at B is $f(d)P_{A,t}$, where $f : \mathbb{R}_+ \to [0, 1]$ measures the proportion of pollution that persists d kilometers downstream. Then we have

$$P_{Bt} = f(d)P_{At} + \tau_t + \psi_B + \varepsilon_{Bt}. \tag{2.9}$$

Note that even though station B is untreated, the decrease in pollution at station A will affect P_{Bt} through the term $f(d)P_{At}$.

The system can then be described in matrix form as

$$\begin{bmatrix} P_{A1} \\ P_{A2} \\ P_{B1} \\ P_{B2} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ f(d) & 0 & 0 & 0 \\ 0 & f(d) & 0 & 0 \end{bmatrix} \begin{bmatrix} P_{A1} \\ P_{A2} \\ P_{B1} \\ P_{B2} \end{bmatrix} + \beta \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_1 \\ \tau_2 \end{bmatrix} + \begin{bmatrix} \psi_A \\ \psi_A \\ \psi_B \\ \psi_B \end{bmatrix} + \begin{bmatrix} \varepsilon_{A1} \\ \varepsilon_{A2} \\ \varepsilon_{B1} \\ \varepsilon_{B2} \end{bmatrix} \tag{2.10}$$

This equation can be written more succinctly as:

$$P = WP + \beta T + \tau + \psi + \varepsilon. \tag{2.11}$$

The matrix W is known in the spatial econometrics literature as a "spatial weight matrix." It captures the upstream/downstream relationship between monitoring

stations and allows for measured pollution at upstream stations to affect measured pollution at downstream stations.

Pollution shocks which occur at station A persist downstream and are observed at station B as well, propagating through the matrix W. This specification allows average pollution levels (through the idiosyncratic fixed effects), as well as any pollution shocks, to persist downstream in a similar fashion. Accounting for the error process in this way allows for more-robust inferences concerning the parameter β in equation 2.11, where $\hat{\beta}$ is the estimate of the average treatment effect.

Estimation of (2.11) requires a specific functional form for f. One candidate is the polynomial specification utilized in Section II. However, the inability of the polynomial specification to ensure a monotonic treatment effect over distance is a shortcoming of that specification. In addition, the local nature of polynomial approximations to nonlinear functions means that the error in the predicted downstream effect is likely to increase as downstream distance increases. For all of the polynomial orders estimated in Section II, the overall estimated effect became *positive* at large distances, a result that is unlikely to be consistent with the underlying data generating process.

To preclude these non-monotonic effects, I adopt an exponential form for the pollution process:

$$f(d) = e^{-\rho d} \tag{2.12}$$

In this specification, downstream pollution (and abatement effects) are assumed to decay exponentially as distance increases. In addition to solving the non-monotonicity and positive treatment effects problems that plague the polynomial model, the exponential pollution process is more parsimonious, requiring just

a single additional parameter, namely the exponential rate of decay ρ. This parameter measures the average rate at which pollution decays as it flows downstream. Estimates of ρ that are large in magnitude indicate that pollution decays rapidly downstream, implying also that pollution (and treatment) spillovers have little effect on downstream pollution.

The full spatial model can be expressed as

$$P = W(\rho)P + \beta\widehat{T} + \tau + \psi + \varepsilon \tag{2.13a}$$

$$\varepsilon = M(\lambda)\xi \tag{2.13b}$$

$$\xi \sim N(0, \sigma^2 I\omega)$$

The spatial weight matrix $W(\rho)$ captures the distance between adjacent monitoring stations:

$$W(\rho)_{i,j} = \begin{cases} e^{-\rho d_{ij}}, & \text{if } i \text{ immediately downstream of } j \text{ in time } t \\ 0, & \text{otherwise} \end{cases} \tag{2.14}$$

The treatment vector $\widehat{T}$ is constructed similarly to equation (2.1), but each treated station is weighted by the distance between that station and its nearest upstream STP:

$$\widehat{T} = \begin{cases} e^{-\rho d_{si}} & \text{if station } s \text{ immediately downstream of STP } i, \\ 0 & \text{otherwise} \end{cases} \tag{2.15}$$

The vector ω weights the variance σ^2 for each station, which accounts for station-level heteroskedasticity. This vector is premultiplied by the additional spatial weight matrix $M(\lambda)$, which captures the correlation between stations

located within a certain radius of each other in the geospatial dimension:

$$M(\lambda)_{i,j} = \begin{cases} e^{-\lambda d_{ij}}, & \text{if } d_{ij} < \overline{d} \\ 0, & \text{otherwise} \end{cases} \tag{2.16}$$

These additional spatial weights describe shocks that affect pollution levels at multiple stations which are not (necessarily) located along the same river system. For instance, regional weather patterns (such as rainfall) may impact pollution levels at multiple stations across rivers. The maximum distance parameter $\overline{d}$ gives $M(\lambda)$ some zero entries, which increases computational efficiency. In all specifications, $\overline{d}$ is set to 500 kilometers.[3]

Solving for P in equation (2.13a) yields

$$P = \beta(I - W(\rho))^{-1}\widehat{T} + \widetilde{\tau} + \widetilde{\psi} + (I - W(\rho))^{-1}M(\lambda)\xi \tag{2.17}$$

Expressing the equation as in (2.17) highlights the downstream effects of treatment and the error process. The matrix $(I - W(\rho))^{-1}$ can be expressed as the infinite sum

$$(I - W(\rho))^{-1} = I + W(\rho) + W(\rho)^2 + \dots \tag{2.18}$$

By construction, each nonzero entry of the matrix $W(\rho)$ is an exponential function of the distance between stations i and j, with elements of this matrix given by $w(\rho)_{ij} = e^{-\rho d_{ij}}$. Entries are nonzero if and only if i is immediately upstream of j; $W(\rho)$ captures the first-order relationship between stations. Therefore, the elements of the matrix $W(\rho)^k$ capture the (exponentiated) distance relationship between

[3]The results are not sensitive to the choice of $\overline{d}$. The estimated values of λ in various specifications suggest that the distance-denoominated half-life of nearby pollution shocks is less than one kilometer. The percentage of the shock that persists to the 500km boundary is effectively zero.

stations that are separated by k stations along a river system. For instance, $w(\rho)_{ij}^{[2]}$ (the ij element of matrix $W(\rho)^2$) is nonzero if i and j are separated by one station.

To see this, suppose the intermediate station (between i and j) is indexed by h. Then $w(\rho)_{ih} = e^{-\rho d_{ih}}$ and $w(\rho)_{hj} = e^{-\rho d_{hj}}$. Then, by pre-multiplying $W(\rho)$ by itself, the ij^{th} element of the squared matrix $W(\rho)^2$ is

$$w(\rho)_{ij}^{[2]} = e^{-\rho d_{ih}} e^{-\rho d_{hj}} \tag{2.19}$$

$$= e^{-\rho(d_{ih}+d_{hj})} \tag{2.20}$$

$$= e^{-\rho d_{ij}} \tag{2.21}$$

Rivers are unidirectional; pollution that flows from station i to j must flow through the intermediate station h as well. The river distance that the pollution travels is therefore the sum of the intermediate distances. The novel approach introduced here specifies an exponential function of distance in the weight matrix, which captures this relationship in an intuitive way.

The infinite sum $(I - W(\rho))^{-1}$ then captures the exponential of the distance between each station and all upstream stations.[4] By interacting this matrix with the treatment variable $\widehat{T}$, the spatial lag model collapses to the the treatment externality model in (2.4a), but with an exponential treatment variable, $\widehat{T}_{sit}e^{-\rho d_{i,s}}$:

$$P_{it} = \sum_s \beta \widehat{T}_{sit} e^{-\rho d_{i,s}} + \tau_t + \psi_i + \varepsilon_{it} + \sum_{\text{j upstream of i}} \varepsilon_{jt} e^{-\rho d_{i,j}} \tag{2.22}$$

Note the similarities between equation (2.22) and the OLS model of Section II (equation (2.4a)). The only difference in the non-stochastic portions of these equations is that the polynomial function of distance in equation (2.4a) is replaced with an exponential function in equation (2.22).

[4]There are a finite number of stations along each river in the data. Therefore the matrix $W(\rho)^k$ contains only zeros when k is large, and the sum is finite.

In contrast to the OLS model of equation (2.4a), the upstream error terms are included in the spatial model. The externality matrix $(I - W(\rho))^{-1}$ interacts with the disturbance term, allowing errors to persist downstream in a similar manner to the treatment externalities. The heteroskedastic-robust errors estimated in Section II are unlikely to be accurate in the presence of externalities. In contrast to a simple clustered-error approach, the externality matrix in equation (2.14) allows errors to decay exponentially, rather than specifying common covariances within arbitrarily designated clusters.

Estimating the Spatial Model. The likelihood function for the spatial model in (2.13a)-(2.14) and sufficient conditions for the existence of a maximum are given in Anselin (2013). Although it can be shown that these conditions are met, the presence of the time and station fixed effects and the heterogeneous error weights give rise to a 1,124-dimensional parameter space. Estimating the maximum and Fisher information is not computationally feasible at present.

I instead estimate the spatial model using Bayesian techniques, as outlined by LeSage and Pace (2009). A Bayesian posterior distribution is simply the weighted average of prior beliefs and the likelihood function. Rather than relying on the curvature (the Hessian matrix) at the maximum of the likelihood function to conduct hypothesis tests, Bayesian techniques can be used to estimate the entire posterior distribution, which can be used to express confidence in particular parameter values. Whenever feasible, I employ uniform priors.[5] Furthermore, Bayesian posterior distributions converge to the likelihood function as the sample size approaches infinity. The posterior distributions reported here are therefore

[5]I use the term "uniform" to refer to non-informative priors, where equal prior probability is assigned to each draw.

close approximations to the traditional likelihood function associated with the model.

In general, the posterior distribution cannot be found analytically. Fortunately, various methods exist that allow for samples to be drawn from the posterior. A large number of samples can then be combined in a Monte-Carlo simulation of the true posterior distribution. Most sampling methods require samples to follow sequentially from the previous sample, forming "chains" of samples. Draws are not independent of each other, and it therefore takes a long time to cover a majority of the posterior distribution. For this reason, it is common practice to include a "burn-in period" wherein chains are initiated and samples are created, but the early samples are then discarded before the Monte-Carlo integration takes place. This burn-in allows the sampler to move away from the initial value and toward the mass of the posterior, thereby minimizing the impact of the initial draws (which are chosen by the researcher). Convergence of the chains can be analyzed by initializing various chains with heterogeneous starting values, then comparing the chains to see if they converge to similar regions of the parameter space.

In this paper I utilize the No-U-Turn sampler of Homan and Gelman (2014). This sampler is characterized by minimal path dependence and rapid convergence to the posterior distribution (Neal et al., 2011). Given the large parameter space and the complexity of the likelihood function, these features are necessary to ensure convergence.

The parameter space is minimized by drawing each unique element of ω from an inverse gamma distribution, where the shape and scale hyperparameters of the inverse gamma are drawn from a uniform distribution before each iteration.

Uniform priors are assigned for each of the fixed effects. The exponential decay terms ρ and λ are assigned exponential priors with an inverse-scale parameter of 1. These priors account for the fact that the likelihood function is approximately uniform for large values of ρ and λ, which correspond to very small externalities.[6]

The No-U-Turn sampler is implemented using the Stan language (Carpenter et al., 2015). Column 1 of Table 4 shows the summary statistics for the posterior distributions. The estimated treatment effect is similar in magnitude to the effect estimated in the previous section. Following the sampler's burn-in period, none of the samples were greater than 0 (in fact, none of the samples were larger than -0.2). Recalling that the posterior distribution is quantitatively similar to the likelihood function, it can be concluded with near absolute certainty that the true treatment effect on measured downstream water pollution is negative and nonzero.

Estimation Results. Posterior summary statistics for the two spatial parameters are also reported in Table 4. The posterior mean for the estimate of λ is substantially higher than that of ρ, indicating that regional across-river pollution spillovers are relatively unimportant when compared to downstream pollution spillovers within rivers. The mean draw from the posterior of λ corresponds to a distance-denominated "half-life" for pollution of 0.387 kilometers, which implies that 95% of pollution disappears across rivers located within about 1.6 kilometers of each other. Regional shocks appear to influence only those stations that are very near each other geospatially, for example stations located along different rivers within the same city.

In contrast, the within-river spillover effect is much more substantial. The posterior mean of ρ is 0.006, which corresponds to a downstream distance-

[6]Without specifying these priors, the sampling algorithm tends to get "stuck" in regions with high draws of the exponential parameters, which delays convergence of the posterior draws.

	Dependent variable:	
	Log of Fecal Coliforms	
	(1)	(2)
Treatment (β)	-0.351	-0.474
	[-0.406, -0.298]	[-0.402, -0.551]
Post 5-Year Effect (β^L)		0.216
		[0.144, 0.296]
Long-Run Effect ($\beta + \beta^L$)		-0.258
		[-0.191, -0.324]
Downstream Effect (ρ)	0.006	0.009
	[0.005, 0.007]	[0.006 0.013]
Regional Effect (λ)	1.850	1.105
	[1.097, 2.913]	[0.531, 2.022]
Downstream Half-Life	115.1	75.8
	[93.6, 140.5]	[54.8, 103.7]

Note: Point estimates are posterior means. 95% posterior likelihood in brackets. All specifications include station and time fixed effects and have 38,625 observations. Column 1 corresponds to equation (2.17) and column 2 corresponds to equation (2.23).

Table 4. Bayesian Spatial Estimation Results

denominated half-life for pollution of 115.1 kilometers. Upstream pollution, and therefore upstream pollution abatement, persists downstream for hundreds of kilometers. Ninety-five percent of pollution measured at any given location on a river dissipates after an estimated average of about 500 kilometers downstream.

Figure 7 plots kernel density estimates for some of the of the posterior parameter distributions (β, rho, and λ) presented in Table 4. Of particular interest is the noticeable right-skewness in the posterior distribution for λ, the regional

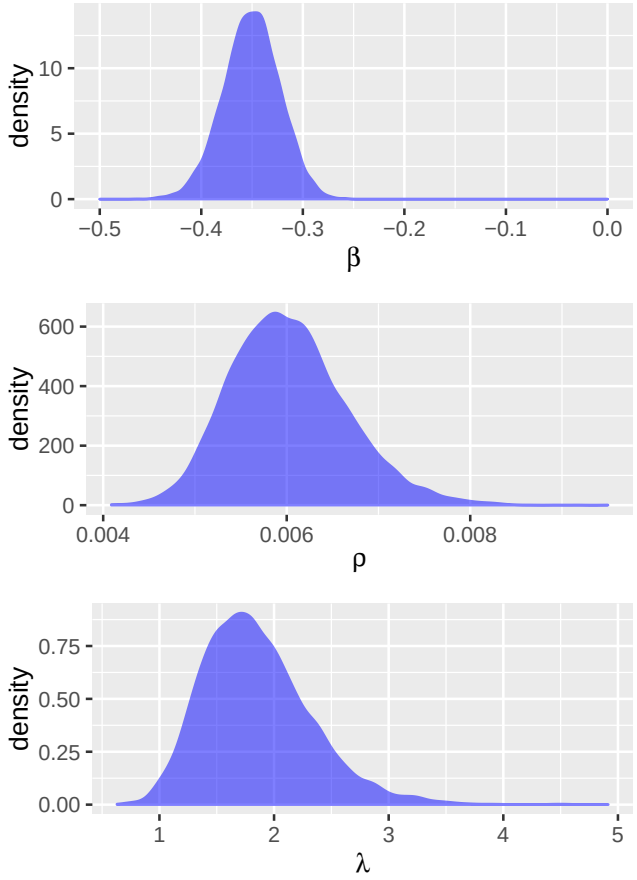


Figure 7. Bayesian Posterior Distributions

effect. The sampler periodically selects large values which correspond to very small regional effects. If assigned a uniform prior, the sampler has difficulty converging to a distribution with a finite mean—the sampler converges only by discounting the large draws of λ. This suggests that regional spillovers are even less significant than the point estimates suggest. Pollution shock processes that have a common effect on stations across rivers appear to have little influence on pollution levels beyond a few kilometers.

Figure 8 shows the trace plot for the posterior draws. Each line represents the posterior draws associated with one of the 32 chains that were initiated (burn-in period included). The rapid convergence of the No-U-Turn sampler is apparent;

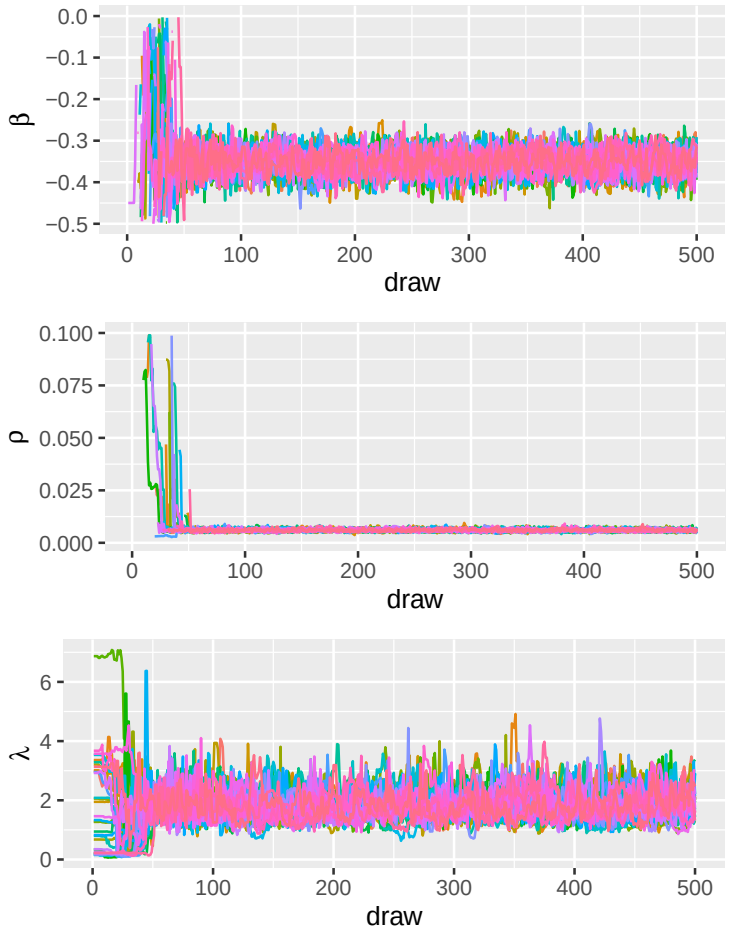


Figure 8. Bayesian Posterior Draws

for all initial values, the sampler converges to the posterior distribution after only 50 draws. Initial values of the treatment effect β range from -1.937 to 2.003, and all chains quickly converge to the posterior centered on -0.351. Initial values of ρ are between 0.150 and 5.992, with all chains converging to the posterior mean of 0.006 within 50 draws. The estimates are therefore robust to a wide range of arbitrary initial values, and the rapid convergence of all chains suggests that the true posterior distribution is being sampled.

The results in Section II suggested that the treatment effect is heterogeneous across time. This heterogeneity is investigated in the spatial setting by estimating

the equation

$$P = W(\rho)P + \beta\widehat{T} + \beta^L\widehat{T}^L + \tau + \psi + \varepsilon, \tag{2.23}$$

which is the spatial analog of equation (2.3). $\widehat{T}$ is again the treatment variable, weighted by the downstream distance from the STP. $\widehat{T}^L$ is similarly defined, but is equal to zero for the first five years after the date when the sewage treatment plant is built, as in equation (2.3). β is therefore the treatment effect for just the first five years after treatment, while $\beta + \beta^L$ is the long-run effect.

Similar to the results in Section II, the treatment effect is found to be larger in the first five years after treatment (column 2 of Table 4). The short-run treatment effect of 0.474 is cut in half after five years, indicating that the sewage treatment plants, on average, lose about half of their effectiveness after five years of operation.

In contrast to Section II, the long-run treatment effect ($\beta + \beta^L$) remains negative and significant. The spatial dependencies in the data appear to bias the long-run OLS estimates toward zero. The richer model that accounts for the spatial lag shows a significant effect on pollution levels downstream of sewage treatment plants even after the initial five-year period. Despite evidence of substantially reduced effectiveness over time, the sewage treatment plants still appear to have a significant long-run beneficial impact on pollution levels.

Joint Posterior Distributions. The estimates of the spillover parameter ρ in column 2 of Table 4 remains similar to that of column 1. While the mean of the posterior shifts substantially, this is mostly due to an increase in the right-hand tail of the posterior distribution (Figure 9). The sampler's propensity to choose larger values of ρ is likely due to the relatively low explanatory power of the additional treatment variable in equation (2.23). The diminished treatment

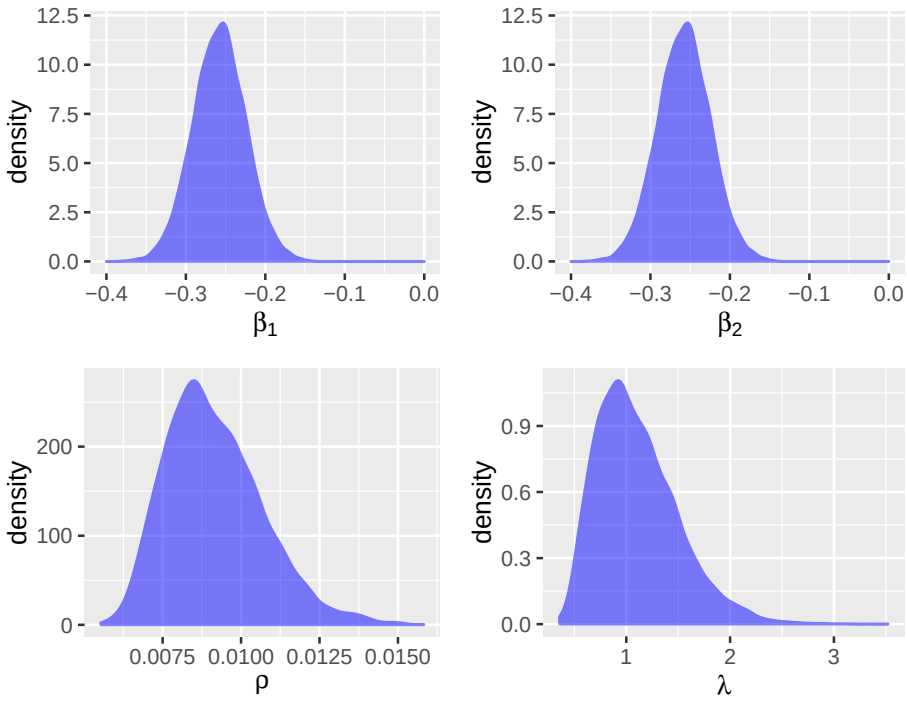
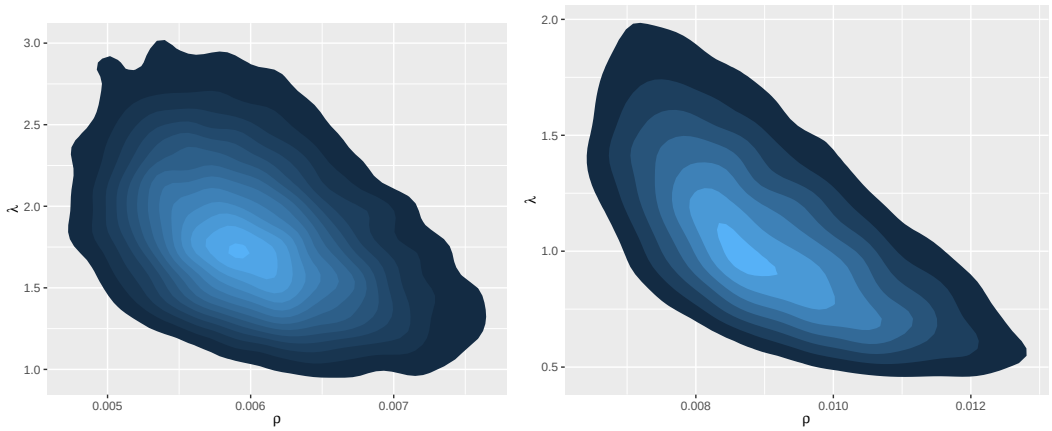


Figure 9. Bayesian Posterior Distributions, Heterogeneous Treatment

effect after five years reduces the amount of systematic right-hand-side variation. The spatial parameters ρ and λ then must explain more of the variation in pollution levels. When this is the case, it becomes more difficult to disentangle the downstream spatial effect (ρ) from the regional spatial effect (λ).

This result can be confirmed visually. Figure 10 shows the joint posterior distributions of ρ and λ for both treatment specifications. With the single treatment variable, the joint posterior is more nearly spherical. In contrast, the additional treatment variable results in a larger negative correlation between the two spatial parameters. Small draws of λ are associated with large draws of ρ, and vice versa.

Downstream Treatment Effects. The parameter β represents the estimated treatment effect measured at a monitoring station immediately



(a) Single Treatment Variable (b) Two Treatment Variables

Figure 10. Joint Posteriors of Spatial Parameters

downstream of an STP ($d_{is} = 0$). The predicted treatment effects for monitoring stations further downstream of an STP are a function of β, the exponential decay parameter ρ, and the downstream distance:

$$T(\beta, \rho, d) = \beta e^{-\rho d} \tag{2.24}$$

This expression can be calculated after each draw of the sampler. Consequently, the posterior distribution of $T(\cdot)$ can also be analyzed.

Figure 11 shows the 95% posterior likelihood plotted against a monitoring station's distance from an upstream STP. The estimated downstream treatment effect is substantial for downstream monitoring stations. The mean estimated reduction in fcoli levels 200 kilometers downstream is larger (in magnitude) than -0.1. Given that nearly all monitoring stations have a another station within 200 kilometers downstream (see Figure 2), failure to account for the downstream effects can potentially introduce substantial bias.

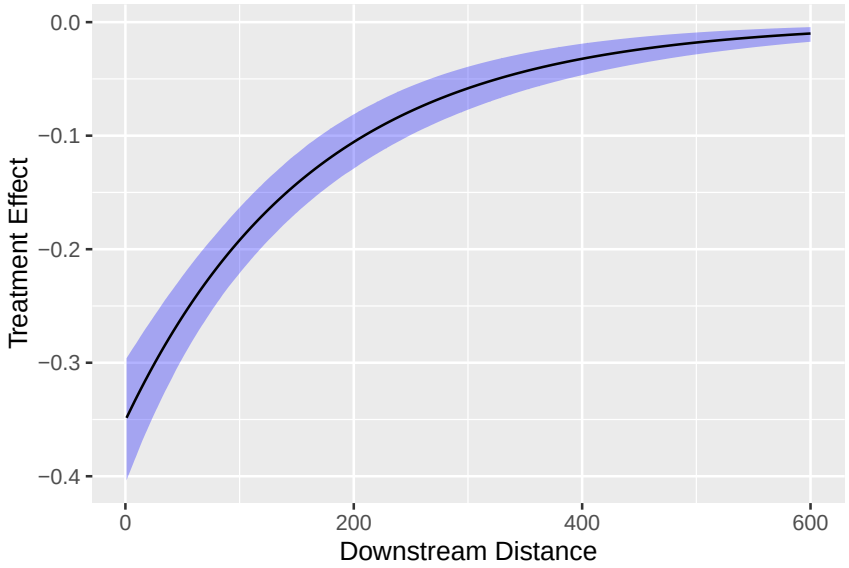


Figure 11. 95% Posterior Likelihood of Downstream Effects

Discussion

The estimated effect of STP construction on measured fcoli counts in the previous sections appears to lie between -0.227 and -1.165. While significantly different from zero in all specifications, this treatment effect is small in magnitude. The sample standard deviation of fcoli levels is 2.92, so estimated treatment effects fall between 13% and 40% of one standard deviation.

However, the magnitude of the estimated treatment effect must be considered in the context of the existing pollution abatement infrastructure. In particular, only a small percentage of generated wastewater is treated before it is returned to the rivers. As of 2012, sewage treatment capacity in major Indian cities was 31% of generated sewage (Kaur et al., 2012). The remaining 69% was released untreated into the environment, either in the form of open latrines or municipal sewage systems that empty directly into a river or lake.

The hypothetical effect of treating all generated sewage can be extrapolated using a "back-of-the-envelope" calculation. An estimated 38,554 million liters per day (MLD) of raw sewage is created in India, of which 11,786 MLD are treated, according to the Ministry of Environment and Forests (Kaur et al., 2012). The STPs created under the NRCP had an original capacity of 4,120 MLD. Dividing the untreated sewage generation by the NRCP capacity and multiplying by the estimated treatment effects gives the approximate reduction in fcoli that would result from treating all generated sewage in India. This value is between 2.52 and 7.57. A widespread reduction of 7.57 (the upper estimate) would reduce 98% of the observed samples to below the EPA threshold for recreational usage.

The potential health benefits are large. Wade et al. (2010) surveyed recreational water users one week after bathing in water monitored for coliform bacteria. They find that an increase of fcoli levels of 2.30 doubles the likelihood of experiencing gastrointestinal illness. The estimated reduction of fcoli levels in India due to treatment (-0.351) would therefore decrease the likelihood of contracting a gastrointestinal disease by approximately 15%. Other epidemiological studies find similar health effects of fcoli reductions (Colford et al., 2005; Colford Jr et al., 2007; Lee et al., 1997; Wiedenmann et al., 2006).

The risk of contracting a gastrointestinal disease from recreational water usage with the sample mean fcoli level in the present study is 3.2% (EPA, The Environmental Protection Agency, 2012).[7] The existence of a sewage treatment plant immediately upstream then reduces this risk to 2.72%. In other words, bathing in water immediately downstream of an STP, as opposed to bathing in untreated water, reduces the risk of contracting a gastrointestinal disease by 0.5%.

[7]This risk is calculated at the EPA's statistical threshold value, which is approximately the sample mean in the data.

This percentage may seem small, but to put this change in perspective, an estimated 120 million people bath in the Ganga as part of the annual Maha Kumbh festival (Khaleej Times, 2013). A new sewage treatment plant built upstream of the festival location would eliminate roughly 600,000 cases of gastrointestinal illness among festival attendees alone.

In all of the specifications in this analysis, the estimated treatment effects are larger in the first few years after treatment than they are for later years in the sample. There are two possible explanations for this result. First, sewage generation may be increasing over time. Increasing population and urbanization rates would result in more sewage generation in the areas likely to have acquired a new STP as part of the NRCP. The already meager capacity of the STPs may simply be overwhelmed by these migration and population trends. Unfortunately, no longitudinal data on regional sewage generation in India appear to exist, so this hypothesis cannot be tested directly. However, demographic trends do not tend to fluctuate dramatically in the short run. The rapid decrease in STP effectiveness after the first few years, visible in Figure 5, therefore suggests that migration and population effects are unlikely to be the driving force.

The second possibility is that the STPs themselves diminish in technological effectiveness over the sample period. Facilities need continuous physical and human capital investments to remain fully operational. If these investments are not made, or if the funding is instead diverted to corrupt officials, then the effectiveness of the facilities could be greatly diminished. Unfortunately there is no centralized effort to monitor the effectiveness of the treatment plants. However, episodic government inspections have been undertaken that largely support this explanation by painting a dismal picture of the maintenance status of sewage treatment plants in India.

In 2007, inspectors from the Ministry of Environment and Forests visited 84 of the 236 known STPs. The performance of 46 (55%) of those plants was rated as "poor" or "very poor" by inspectors, with just 8 (9.5%) being rated as "good" (Mauskar, 2008). The report specifically highlights a lack of investment as a serious problem:

> Fund shortage is an important factor in poor operation and maintenance of STPs and has been reported in 26 cases. The problem of fund shortage is mostly reported from States of Bihar, Haryana, U.P., and West Bengal. This trend shows that the root of problem lies in less priority being given to sewage treatment (page 17).

Human capital shortcomings are also highlighted. STP operation is generally carried out by private-sector contractors, who "generally depute unqualified or less qualified staff at site." In some cases, inspectors arrived on site to find the plants entirely abandoned with thick vegetation impeding access to essential sections.

A more comprehensive investigation in 2013 tells a similar story (Tyagi, 2013). At that time, 104 STPs built as part of the NRCP were visited by investigators, and fully 28 (27%) were non-operational or had unsatisfactory performance. The report identifies three areas for improvement: ensuring uninterrupted energy supply, providing more-skilled manpower, and regular maintenance.

Caveats and Directions for Future Research. The estimates presented here are average treatment effects. There is considerable heterogeneity in the technologies and capacities of the sewage treatment facilities which is not captured in any of the specifications. The specific characteristics of each STP may be possible to describe with a combination of interviews with STP managers,

analysis of satellite photographs, and data from the Central Pollution Control Board. This information may allow for heterogeneity analysis with regard to the specific treatment technologies utilized or operational schemes adopted (i.e. public versus private management).

Another important source of heterogeneity that is not accounted for in this study is the hydrological properties of the rivers. The average downstream effect is described by a parsimonious polynomial specification (Section II) or an exponential function (Section II). However, a wide variation in downstream effects likely exists. Broad and slow-moving rivers like the Ganga are likely to carry pollution downstream at different rates than swift-moving mountain streams. These differences are policy relevant, as reductions in pollution along rivers with significant downstream effects will benefit more people than reductions in pollution on rivers with less downstream effect, *ceteris paribus*.

As is common with developing countries, high-resolution panels of demographic and economic variables do not exist for India. This absence may introduce some omitted variable bias into the results. The fixed effects account for any heterogeneity that is constant across time or stations, but some regions may experience demographic or economic changes that are not experienced by other regions. These changes may in turn impact the operation and management of sewage treatment facilities.

This research suggests many avenues for future inquiry. If data on the operation and management of each sewage treatment plant could be obtained, then it may be possible to compare the effectiveness of various management schemes. Given the lack of oversight of the STPs built under the NRCP, for instance, it is possible that STPs run by private contractors may perform poorly when compared

to publicly operated STPs. Another approach is to to follow along similar lines as Lipscomb and Mobarak (2015) and analyze the effectiveness of STPs as a function of distance from the downstream state or district border. Given that the downstream effects are found to be significant, it is possible that local governments are less likely to invest in the operation and maintenance of STPs that are closer to a downstream border and thus will deliver more of their benefits to a different downstream jurisdiction.

The Ministry of Environment and Forests reports that the NRCP sewage treatment plants were built in 1995 (Tyagi, 2013). No comprehensive data on the operational history of these plants is available. It is therefore possible that some STPs began operating after 1995, the treatment date used in this research. In this case, the estimated average treatment effect would be biased toward zero. The operational effectiveness of the STPs can only be observed by their downstream impact on measured pollution levels. Future research will focus on this STP-level heterogeneity.

Conclusion

The estimation results in this paper highlight the importance of controlling for treatment externalities, which are likely to exist in treatment-effect studies in environmental settings. Failing to account for these spillovers can cause considerable bias in point estimates and standard errors for key coefficients. The traditional OLS-based approach can recover unbiased point estimates, but more attention needs to be devoted to the covariance structure if proper inferences are to be made.

The model developed in Section II explicitly accounts for spatial spillovers in both the point estimates and the error structure. The point estimates of this

model are analytically equivalent to the traditional OLS model with treatment externalities if the spillover process is exponential (in distance or any other dimension). The spatial model integrates the same type of spillover process into the error term as well. Spillovers that can be modeled as exponential processes are likely to occur in many environmental settings, whenever the spillovers can be modeled as traveling along distinct paths. Other applications include transportation, communication or social networks, where the size of the treatment effect can be expected to diminish exponentially as it propagates across the network.

The construction of new sewage treatment plants is associated with, on average, a substantial initial decrease in pollution levels. This finding contradicts previous research (Greenstone & Hanna, 2014) and conventional wisdom in India. Treating a larger percentage of generated sewage would have significant health benefits even in the absence of other institutional changes.

However, the India's weak institutional setting appears to compromise the effectiveness of its sewage abatement infrastructure. Additional benefits could be derived simply by utilizing existing infrastructure to its full capacity. The results presented here thus point toward to two obvious policy recommendations. First, greater sewage-treatment capacity should be created along India's rivers. Second, more resources should be devoted to the continuing oversight and maintenance of existing treatment plants.

Both policies would improve environmental quality on their own. But complementarities between the two policies would create greater benefits—new sewage treatment plants would further benefit from improved oversight. If resource constraints are binding, this suggests that short-run policy should

focus on combating the rapid human capital depreciation and physical capital depreciation that occurs at existing treatment plants. Contracts could be written with specific performance goals, with government auditors inspecting facilities at regular intervals. This low-cost approach may have large effects, provided that the independent contractors who operate STPs can be properly incentivized.

CHAPTER III

GEOSPATIAL DISAGGREGATION OF ECONOMIC AND DEMOGRAPHIC DATA

Introduction

In much of the world, census and survey count data are reported for predefined geographic areas. These geographic areas usually define (or at least constrain) the unit of observation in empirical analysis. Heterogeneity within the geographic region is typically hidden from the researcher.

In this paper, I develop a method to spatially disaggregate count data recorded over any arbitrary geographic area. The method uses freely available satellite data to estimate pixel-level counts of the aggregated statistic. The method is *dasymetric* in that aggregate counts are preserved in the pixel-level estimates. Aggregate and mean prediction error is zero by construction.

The method presented here creates estimates of economic and demographic statistics for small (500m or less) geographic areas. This geographic resolution is considerably finer than the resolutions typically used in empirical research. Count statistics aggregated to a administrative boundary might obscure important heterogeneity within the administrative boundary, or may even be a function of the administrative boundary itself, as administrative boundaries are often determined endogenously to many relevant economic outcomes. The method described in this chapter creates geospatial estimates that can be utilized independently of predefined administrative boundaries and can be aggregated to any arbitrary geographic region. For example, a researcher interested in water pollution exposure can estimate the population or poverty rate within any arbitrary distance to a body

of water—or even a function of distance—to a body of water. Such a procedure would allow for robust estimates of treatment intensity.

There is a large literature on the spatial disaggregation of population data. The earliest attempts involved researchers in hot air balloons or light aircraft manual counting housing units. More resent and sophisticated examples use satellite-derived data and statistical techniques to assign population counts to individual pixels within the target geographic area (Mennis, 2009).

The central challenge to producing accurate disaggregated population estimates is the choice of a statistical model. The functional relationship between the satellite-observed spectral qualities of the earth's surface and population density is unknown and likely to be characterized by a high degree of nonlinearity and interaction between observable characteristics. The traditional approach to overcome this problem is to utilize intermediate raster data created by parametric processes. These intermediate data layers generally take the form of land-use classification rasters or open-source geographic databases. Standard parametric statistical models are then used to allocate aggregate population among these various land-use categories.[1] Other approaches involve using non-parametric machine-learning algorithms trained on the aggregate region, then using the estimated model to predict pixel-level population counts (Anderson et al., 2014).

These traditional disaggregation methods create estimates are not constrained to sum to the aggregate population counts. Schroeder and Van Riper (2013) utilize an expectation maximization (EM) algorithm that estimates population level at each pixel, then aggregates the pixel-level estimates to create a set of population weights for the next iteration of the algorithm. The outcome

[1]Examples of this approach include Li and Weng (2005), Gaughan et al. (2013), and Azar et al. (2013)

of the algorithm is a set of pixel-level weights that dasymetrically disaggregate the count data to the pixels within any geographic region.

The approach I propose here combines the EM-based estimation of population weights of Schroeder and Van Riper (2013) with a non-parametric statistical model of the type described by Anderson et al. (2014). Specifically, I utilize the random forest algorithm developed by Breiman (2001). The flexibility of the non-parametric random forest model eliminates the need for intermediate data layers, and this allows the algorithm to run on minimally processed satellite data. Avoiding the need for intermediate layers also means that the algorithm performs similarly on any economic or demographic count data, constrained only by the satellite-observed spectral qualities of the statistic in question. For example, any difference in populations that can be readily observed from above, approximately, by the naked eye—such as population density, construction materials, or urban development boundaries—should be predicted with similar accuracy by the algorithm.

The use of satellite data has become increasingly common in economics.[2] In a landmark study, Henderson et al. (2012) show that satellite-observed nighttime light is highly predictive of aggregate income at the national level. These findings have been used to motivate studies of everything from comparative development (Henderson et al., 2018; Michalopoulos & Papaioannou, 2012) to the local effects of political clientelism (Asher & Novosad, 2017). But despite the well-known relationship between aggregate output and nighttime light, comparatively little is known about the relationship between local economic conditions and nighttime light. For example, Asher and Novosad (2017) use nighttime light data aggregated

[2]See Donaldson and Storeygard (2016) for a comprehensive review of the use of satellite data in economic research

to the electoral district as an outcome variable to show that close electoral outcomes in India lead to a comparative increase in economic growth relative to non-competitive districts. However, they are unable to describe how that growth manifests within the electoral district. Large economic gains accruing to a small minority of high-income earners are empirically indistinguishable to widespread increases in economic activity among poorer residents. In one of the illustrated applications below, I show that combining nighttime light data with disaggregated population data permits pixel-level estimates of economic activity that are highly correlated with other economic outcomes.

This paper is organized as follows. Section III details the dasymetric disaggregation algorithm. The algorithm is demonstrated using data from the 2014 census of Uganda. Next I provide some applications of the algorithm in the Ugandan context. In section III, I combine pixel-level nighttime light data with disaggregated population data to create pixel-level estimates of per-capita economic activity. This strategy allows for assessment of the geographic dispersion of economic activity and highlights the inadequacy of relying on data aggregated at the level of the census district to draw inferences about the distribution of economic activity. Section III demonstrates that the resulting disaggregated count statistics that are ex ante correlated with poverty are highly correlated with pixel-level per-capita nighttime light estimates. In section III, I show that disaggregated census counts can be combined with remotely sensed pollution data to create more-accurate measure of pollution exposure than were previously available to researchers. I show that the previously unobserved variation in pollution exposure within census district has a substantial effect on the overall distribution of pollution

exposure. Finally, I conclude with a discussion of potential future applications and possible ways to improve the predictive performance of the algorithm.

Algorithm

Geospatial disaggregation involves the allocation of an aggregate count statistic, measured over a larger geographic region, to constituent partitions within the region. In the case of satellite data, these partitions are the component pixels (defined as the resolution of the raw satellite data) of a geographic region. Formally, denote Y as the aggregate count statistic measured over some well-defined spatial region, such as a state, county, or census tract. The geographic region is composed of K pixels, indexed $k = 1, 2, \ldots, K$. The pixels partition the geographic space, so the disaggregated contribution of each pixel k to the aggregate count Y is a fraction of the total count $\lambda_k Y$, with $\sum_{k=1}^{K} \lambda_k = 1$. For example, if Y is the aggregate population of some geographic region, $\lambda_k Y$ is the population of pixel k within said geographic region. The geospatial disaggregation problem can therefore be stated in terms of estimating the pixel-level dasymetric weights $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_K)$.

The vector X_k contains remotely sensed data measured at each pixel k. The relationship between the pixel weights and the remotely sensed data is modeled as

$$\lambda_k Y = f(X_k) + \varepsilon_k \tag{3.1}$$

where ε_i is a mean-zero disturbance.

The geospatial disaggregation problem requires joint estimation of f and λ. I proceed using a two-step procedure, similar to an expectation-maximization (EM) algorithm. In the first step (the "maximization" step), $\widehat{f}(X_k)$ is estimated for a given $\widehat{\lambda}$. Next (the "expectation" step), the predicted pixel-level counts $\widehat{y_k} =$

Step 0. Initialization: Randomly select weights $\widehat{\lambda}_k$ for each pixel k, with $0 \leq \widehat{\lambda}_k \leq 1$ and $\left(\sum_{k=1}^{K} \widehat{\lambda}_k = 1\right)$.

Step 1. Estimate $\widehat{\lambda}_k Y = f(X_k)$ with a random forest.

Step 2. Predict pixel-level counts $\widehat{y}_k = \widehat{f}(X_k)$.

Step 3. Create new weights $\widehat{\lambda}_k = \dfrac{\widehat{y}_k}{\sum_{i=1}^{K} \widehat{y}_i}$.

Step 4. Repeat steps 1–3 until convergence.

Figure 12. The Disaggregation Algorithm

$\widehat{f}(X_k)$ are used to construct the predicted pixel-level weights $\widehat{\lambda}_k = \widehat{y}_k / \sum_{i=1}^{K} \widehat{y}_i$. The procedure is seeded using random weights λ, and repeated until some convergence criteria (such as reduction in out-of-sample mean squared error) is satisfied.

The expectation step—the creation of weights $\widehat{\lambda}$—is a necessary step in the algorithm. Because of this, the algorithm is only able disaggregate count data that can be expressed as $Y = \sum_{k=1}^{K} y_k$, where y_k is well-defined on each geographic area k. Importantly, data that represent rates or averages (such as the unemployment rate) cannot be disaggregated by the algorithm, as the aggregate average cannot be expressed as the sum of the constituent pixel averages within the geographic area.

Each iteration of the algorithm produces two relevant values. The estimated dasymetric weights $\widehat{\lambda}$ are used to construct the pixel-level dasymetric allocations $\widehat{\lambda}_k Y$. These are estimates of the pixel-level counts of the aggregated statistic, and automatically have the property $\sum_{k=1}^{K} \widehat{\lambda}_k Y = Y$. That is, the aggregate prediction error is constrained to be zero. Prediction error will therefore attenuate pixel-level

prediction error within the spatial region in question, as the true pixel-level counts are usually a mean-preserving spread of the predicted counts.

The second output of the algorithm is the raw estimate of the pixel-level disaggregation $\widehat{y}_k = \widehat{f}(X_k)$. While lacking the property of a zero aggregate error, these forecasted values can be used for out-of-sample predictions of pixel-level counts. These estimates can be used for various cross validation methods to assess the overall performance of the disaggregation algorithm.

The function f transforms the remotely sensed data X_k into a pixel-level disaggregated count $\widehat{y}_k$. The two primary difficulties in estimating f are variable selection and the functional specification of $f(\cdot)$. In recent years, large quantities of satellite-derived data are available for most of the earth's surface. Traditional dasymetric methods using satellite data rely on intermediate estimates, such as land-cover classifications. The outputs of these intermediate models (pixel-level land classification) are then used as inputs in parametric models to explain pixel-level counts of population.

The parameterizations of the traditional dasymetric methods leave much to be desired. Intermediate data layers are generally constructed independently, without regard to the outcome variable of interest. Binary land-cover classifications (e.g. grassland, developed, or forest) may be useful for disaggregating certain types of count variables like population or farming. However, they are not constructed with their value in predicting population estimates in mind, nor is it likely that they will be useful for disaggregating other types of socioeconomic count variables. Furthermore, there is little ex ante justification for preferring specific parameterizations of f. The commonly used method of ordinary least squares, for

instance, assumes a linear relationship between the variables in the vector X_k and the dasymetric weights λ_k.

For the method introduced in this chapter, these problems are addressed by using a random forest algorithm to estimate $f(X_k)$. The random forest is nonparametric in the sense that it allows for high-order interactions between variables as well as an arbitrary degree of nonlinearity. Further, the random selection of variables at each partition node in the algorithm allows the X_k variables with highest predictive power to drive results, while simultaneously discouraging overfitting by disallowing any single variable (or set of variables) from dominating the construction of each regression tree. The random forest approach therefore allows the researcher to remain agnostic regarding variable selection and parameterization.

I demonstrate this new disaggregation algorithm using data from the 2014 census of Uganda. In addition to the typical counts of people and households, the Ugandan Bureau of Statistics records counts of a variety of socioeconomic statistics, including the number of households who rely on a tabooda (kerosene lamp) for light, counts of people who eat two or more meals per day, and the number of subsistence farms. The smallest geographic area reported by the Bureau of Statistics is the subcounty. Shapefiles produced by the Bureau of Statistics can be matched to these count data for 1,441 identified subcounties having an average area of 168km^2.

Three sources of remotely sensed data are exploited for variables to include as the X_k. The Landsat 8 spacecraft records a variety of spectral radiance wavelengths (see Table 5) on a 16-day interval at a resolution of 15m or 30m. All images over the sample area (Uganda) for the year 2014 are combined using a

"greenest-pixel" composite to create a single raster image of each spectral band for the year. nighttime light data is recorded by the Visible Infrared Imaging Radiometer Suite (VIIRS) aboard the Suomi NPP satellite. Monthly average nighttime light is recorded at a 500m resolution. Annual averages for each pixel are calculated. Finally, elevation data are derived from the Shuttle Radar Topography Mission.

Data	**Spacecraft**	**Spatial Resolution**	**Temporal Resolution**
0.43–0.45 µm reflectance	Landsat 8	30m	16 days
0.45–0.51 µm reflectance	Landsat 8	30m	16 days
0.53–0.59 µm reflectance	Landsat 8	30m	16 days
0.64–0.67 µm reflectance	Landsat 8	30m	16 days
0.85–0.88 µm reflectance	Landsat 8	30m	16 days
1.57–1.65 µm reflectance	Landsat 8	30m	16 days
2.11–2.29 µm reflectance	Landsat 8	30m	16 days
0.52–0.90 µm reflectance	Landsat 8	15m	16 days
1.36–1.38 µm reflectance	Landsat 8	15m	16 days
10.60–11.19 µm reflectance	Landsat 8	30m	16 days
11.50–12.51 µm reflectance	Landsat 8	30m	16 days
nighttime light	Suomi NPP	500m	1 month
Elevation	STS-99	30m	February, 2001

Table 5. Satellite Data Sources

Each of the raster layers are resampled to a 500m resolution, the lowest resolution available in the raw satellite data. Specifically, X_k for each 500m pixel k contains information on the mean and variance of each raster layer within each pixel, for those sources with resolution greater than 500m. All calculations are performed using the Google Earth Engine platform. The final pixel-level data matrix X has 824,272 rows (pixels) and 28 columns (the mean and variance of each explanatory variable described in Table 5).

The algorithm is initialized with a random disaggregation of subcounty counts. To prevent the algorithm from overfitting on the initial (or any other) random disaggregation, the algorithm for each subcounty is run jointly with each

of the bordering subcounties. In other words, $\widehat{f}(X_k)$ is estimated on a collection of neighboring subcounties, but the predicted $\widehat{\lambda}$ is only retained for the central subcounty. In addition to avoiding the problem of overfitting, this process also allows for the assessment of out-of-sample performance of the algorithm. At each iteration, a random forest model is estimated on the surrounding subcounties, excluding the central subcounty. The estimated random forest model is then applied to the X_k variables for the hold-out subcounty, then these constituent pixel-level predictions are aggregated and compared to the census counts. If the algorithm is accurate, the deviation of the predicted counts in the hold-out subcounty to the true census counts for that holdout subcounty should be minimal.

Figure 13 demonstrates the disaggregation algorithm for the city of Kampala and the surrounding area. The top-left panel shows a sample of the raw satellite data—specifically, the radiance associated with the spectral colors red, blue, and green as measured by Landsat 8. Built-up urban areas and (lighter) sparsely populated green areas (darker) are plainly visible. The top right panel shows the raw census count data for the subcounties in the same region. The goal is to disaggregate the choropleth population counts in the top-right panel in a manner consistent with the spectral properties of the satellite data in the top-left panel.

As described above, the algorithm is executed twice on each individual subcounty and its surrounding subcounties. The first execution includes the target subcounty and dasymetrically disaggregates the census population counts. The second execution of the algorithm excludes the target subcounty from the estimation procedure, then uses the estimated random forest based on the surrounding subcounties to predict the pixel-level population in the target subcounty. The resulting estimation is not dasymetric in that aggregate population

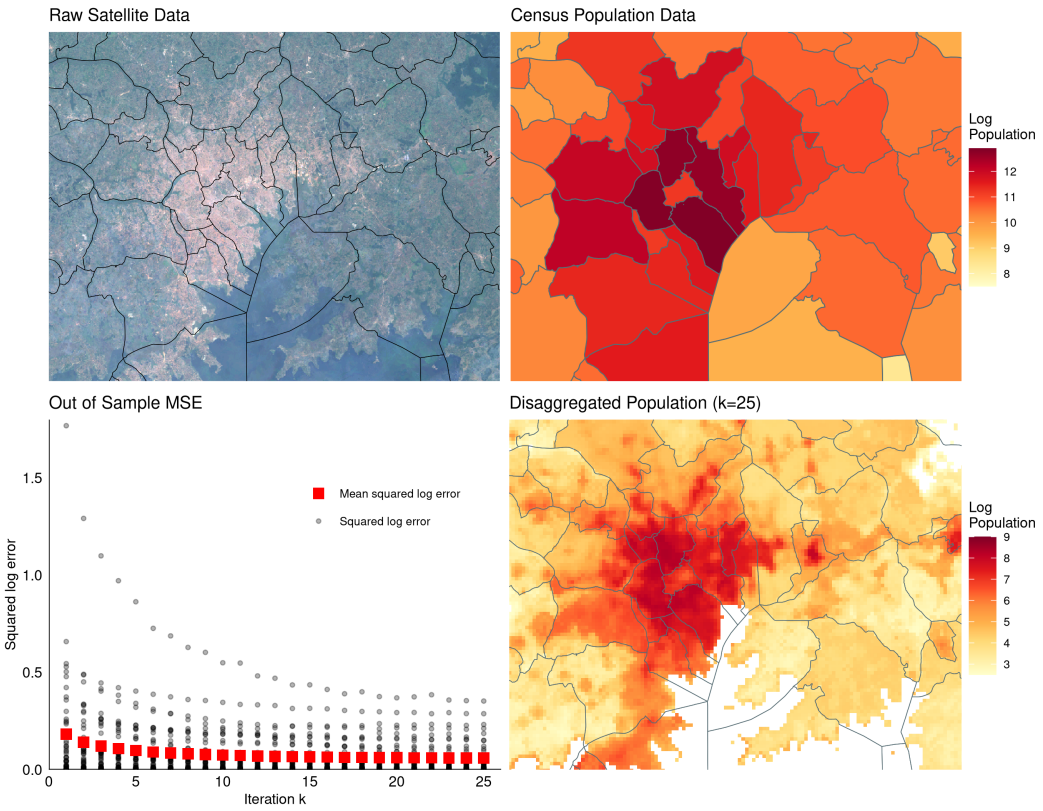


Figure 13. Population Disaggregation of Kampala, Uganda

counts are not preserved, but the (log) error of the predicted populations can be used to assess the overall accuracy of the algorithm.

The bottom-left panel of Figure 13 records the squared log error for each subcounty in the Kampala region along with the mean squared error for the region. The algorithm converges rapidly as both the mean and the most extreme out-of-sample errors are essentially stable after 25 iterations.

The bottom-right panel of Figure 13 shows the resulting dasymetric population disaggregation after 25 iterations. The disaggregated counts appear to match closely the pattern of development that is observable in the (approximately) true color image in the top-left panel. Furthermore, the lack of substantial

cross-border discontinuities in estimated population is strong evidence that the algorithm is not overfitting, since the estimated $\widehat{f}(\cdot)$ for each subcounty is created independently of the $\widehat{f}(\cdot)$ of each neighboring subcounty.

The non-parametric nature of the disaggregation algorithm allows the method to be applied to any count statistic. I demonstrate this flexibility by disaggregating census counts of population, tabooda (kerosene lamp) usage, population consuming two or more meals per day, and subsistence farms. Each count is disaggregated within all observed subcounties in Uganda.

For each of these four variables, Figure 14 shows the out-of-sample error for each subcounty, across all observed Ugandan subcounties. The 45° dashed line represents zero prediction error. The bulk of the out-of-sample errors are clustered around the 45° line, and none of the census variables indicate systematic out-of-sample error across any part of the subcounty distribution. Further, the distribution of out-of-sample errors are similarly distributed around the 45° line for each of the four variables, indicating that the algorithm performs similarly for each of the four count statistics.

Inequality

Researchers interested in the distribution of economic activity are highly constrained by the availability of economic data, particularly in the developing world. In recent years it has become common to utilize nighttime light data to estimate economic activity in data-poor environments. While the relationship between nighttime light and economic activity is well established at the national level (Henderson et al., 2012), comparatively little is known about the relationship between nighttime light and more-local economic conditions. Additionally, current methods do not distinguish between economic activity reflecting productivity

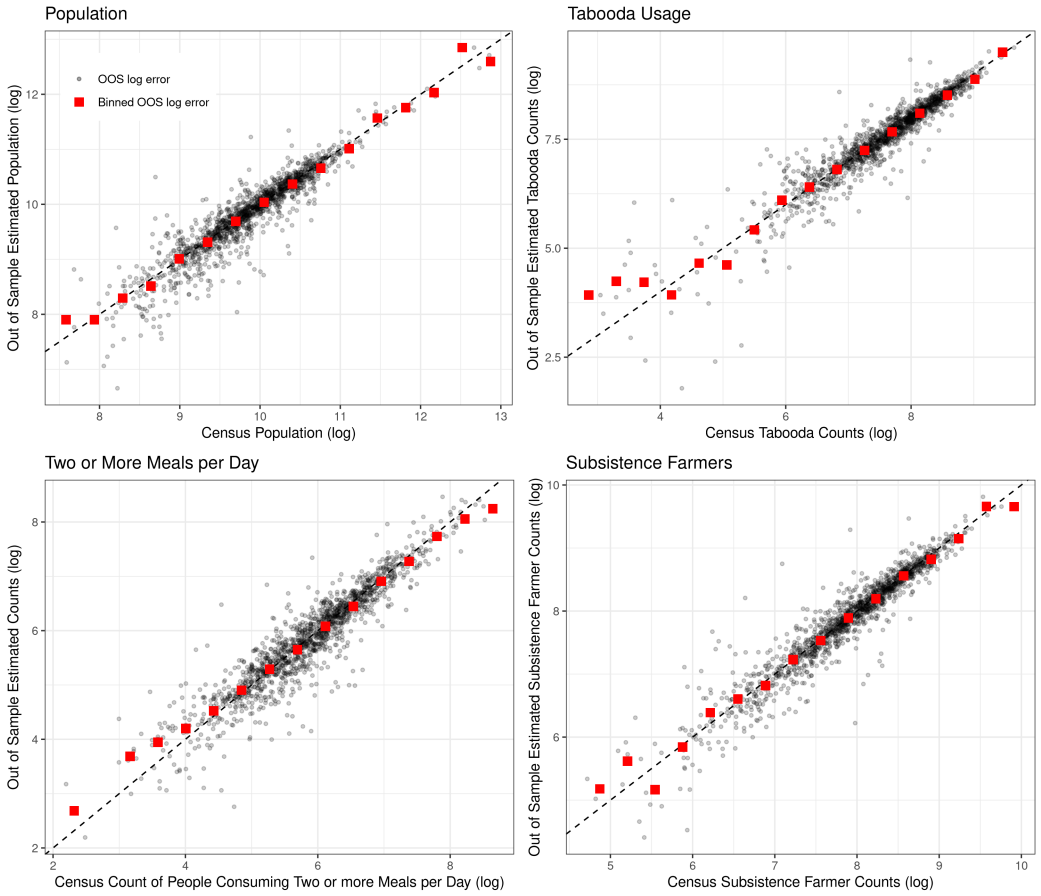


Figure 14. Out-of-Sample Errors

and economic activity reflecting population. If an increase in nighttime light is observed over a given region, it is generally unknown whether the increase reflects an increase in wages or simply an increase in population. In terms of welfare, then, changes in the effects of light activity are ambiguous.

A common approach to address this problem is to aggregate nighttime light data to a known geospatial unit, such as those demarcated by a census. However, this aggregation strategy may mask changes in economic activity that occur within the geospatial unit. In this section, I demonstrate how the dasymetric disaggregation procedure described in the previous section can be combined with

nighttime light data to create higher-resolution estimates of per-capita economic activity.

Denote the set of n economic agents within a geographic area of interest as $N = \{1, 2, \ldots, n\}$. The function $L : \mathbb{P}(N) \to \mathbb{R}$ maps some economic outcome to any subset of economic agents in N, such as income or reliance on subsistence farming. ($\mathbb{P}(N)$ is the power set of N). If economic outcomes are observable for each individual agent—$L(E)$ is known for all $E \subset \mathbb{P}(N)$—then inequality can be calculated in a straightforward manner. For instance, Theil's T index is defined as:

$$T = \sum_{i=1}^{n} \frac{L(\{i\})}{L(N)} \ln \left(\frac{L(\{i\})}{\bar{L}(N)} \right), \tag{3.2}$$

where $\bar{L}(E) = L(E)/|E|$ is the average of the economic outcome of interest across all agents in $E \subseteq N$.

In most settings, $L(\cdot)$ is observed only over a small subset of $\mathbb{P}(N)$. For census and other spatial data, $L(\cdot)$ is observed over geospatial partitions of the universe of individual agents. Formally, let $D = \{D_1, D_2, \ldots, D_J\}$ be a partition of N such that $L(D_j)$ is observable for all $i = 1, 2, \ldots, J$ spatial districts. While $L(\{i\})$ is unobserved, the economic outcome for each i is indirectly observed through i's membership in $D_j \in D$, for some j. The most common way to visualize $L(D_j)$ is with a choropleth map, which shows the geospatial partition of N along with the value of $L(D_j)$ for each partition, typically represented on a color scale. For example, looking ahead, panel A of Figure 15 shows a choropleth map for Uganda, where each D_j represents census-defined subcounties and $L(D_j)$ is the aggregate nighttime light emanating from that subcounty.

A decomposition of equation (3.2) allows aggregate inequality to be described as:

$$T = \sum_{j=1}^{|D|} \frac{L(D_j)}{L(N)} \ln\left(\frac{\bar{L}(D_j)}{\bar{L}(N)}\right) + \sum_{j=1}^{|D|} \frac{L(D_j)}{L(N)} T_j \tag{3.3}$$

$$\text{where } T_j = \sum_{i \in D_j} \frac{L(\{i\})}{L(D_j)} \ln\left(\frac{L(\{i\})}{\bar{L}(D_j)}\right) \tag{3.4}$$

The first term in equation (3.3) is between-district inequality, which can be calculated directly from choropleth values. However, overall inequality is also a function of the unobserved second term, which captures within-district inequality. Inequality calculations based solely only on between-district inequality can dramatically understate—or even overstate—the true level of inequality by omitting within-district inequality (equation (3.4)).

Dasymetric disaggregation of the choropleth using the algorithm described above can provide information about the previously unobserved within-district inequality of nighttime light. First, partition each district D_j into its component pixels. Denote this further partition as $P^j = \{P_1^j, P_2^j, \ldots, P_K^j\}$, and note that $\bigcup_{j=1}^{|D|} P^j$ also partitions N. Equation (3.2) can therefore be further decomposed:

$$T = \sum_{j=1}^{|D|} \sum_{P \in P^j} \frac{L(P)}{L(N)} \ln\left(\frac{\bar{L}(P)}{\bar{L}(D_j)}\right) + \sum_{j=1}^{|D|} \frac{L(D_j)}{L(N)} \ln\left(\frac{\bar{L}(D_j)}{\bar{L}(N)}\right) + \sum_{j=1}^{|D|} \sum_{P \in P^j} \frac{L(P)}{L(N)} T_p \tag{3.5}$$

$$\text{where } T_p = \sum_{i \in P} \frac{L(\{i\})}{L(P)} \ln\left(\frac{L(\{i\})}{\bar{L}(P)}\right) \tag{3.6}$$

The first term in equation (3.5) is the within-district, between-pixel inequality. The second term is the previously described between-district inequality. The last term is the unobserved within-pixel inequality. In essence, the unobserved "residual" component of equation (3.3) is decomposed into the observable between-

pixel inequality term and an unobserved residual (which embeds the term in equation (3.6)). Note that the unobserved residual approaches zero as $|P| \to 1$ for each $P \in \bigcup_{j=1}^{|D|}$ (in the limit, each individual i is uniquely observed), or as $L(\{i\}) \to \bar{L}(P)$ for each $i \in P$ (marginal light output is the same for every agent within a pixel). In other words, unobserved residual inequality approaches zero as the pixel resolution increases.

Monthly nighttime-light luminosity is observed at a 500m resolution beginning in 2012 by the Visible Infrared Imaging Radiometer Suite (VIIRS) aboard the Suomi National Polar-Orbiting Partnership spacecraft. I calculate average nighttime light for 2014 for each pixel and combine these data with the 2014 Ugandan census data on population, algorithmically disaggregated to the same resolution as the VIIRS data. The result is a 500m-resolution estimate of per-capita nighttime light luminosity for the entire country of Uganda. The implied aggregate nighttime light at the subcounty level is displayed in panel A of Figure 15.

Defining $L(\cdot)$ as VIIRS nighttime light allows for direct calculation of the observed components of Theil's T in equation (3.5). Specifically,

$$\sum_{j=1}^{|D|} \sum_{P \in P^j} \frac{L(P)}{L(N)} \ln\left(\frac{\bar{L}(P)}{\bar{L}(D_j)}\right) = 0.2706 \quad \text{(within-subcounty, between-pixel)} \tag{3.7}$$

$$\sum_{j=1}^{|D|} \frac{L(D_j)}{L(N)} \ln\left(\frac{\bar{L}(D_j)}{\bar{L}(N)}\right) = 0.4177 \quad \text{(between-subcounty)} \tag{3.8}$$

Of the total observed inequality, 39.3% is caused by (previously unobserved) within-subcounty inequality, while 60.7% is due to between-subcounty inequality.

A visual representation of between-subcounty and within-subcounty inequality is displayed in Figure 15. Panel A shows the average per-capita light

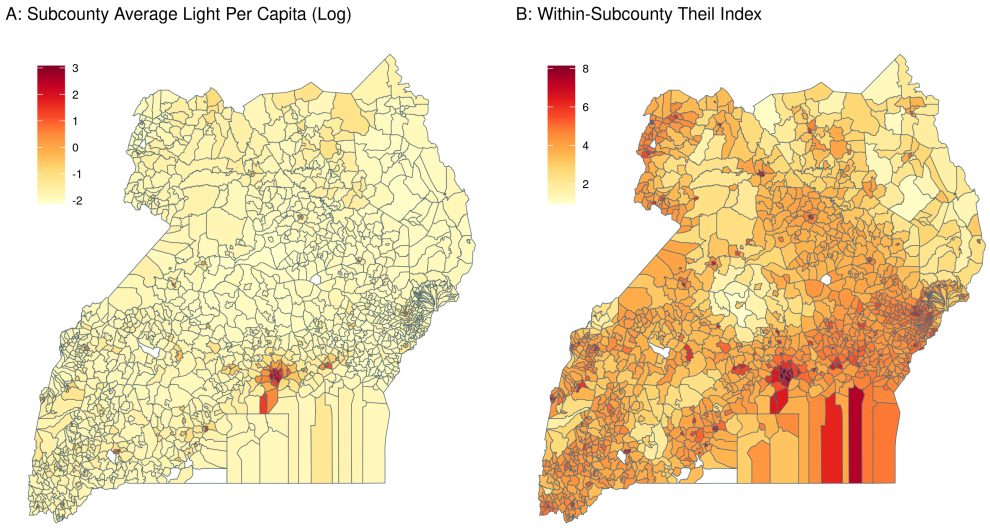


Figure 15. Per-Capita Nighttime Light and Within Subdistrict Inequality Estimates

per subcounty, $\ln(\bar{L}(D_j))$. Extensive inequality is apparent, primarily in the form of small, urban subcounties with high average nighttime light and large, rural subcounties with low average nighttime light.

Panel B of Figure 15 shows the within-subcounty, between-pixel inequality, $\ln\left(\frac{\bar{L}(D_j)}{\bar{L}(N)}\right)$. In comparison to panel A, within-subcounty inequality is more evenly distributed across subcounties, as evidenced by the geographic dispersion of light and dark regions across the country. In comparison, the subcounty-aggregated nighttime lights in panel A show little geographic variation outside of the main urban areas. The stylized fact that emerges from Figure 15 is that between-subcounty inequality is driven by the urban-rural divide, whereas within-subcounty inequality is a nationwide phenomena.

Figure 16 further investigates the distribution of within-subcounty inequality. Panel A shows the subcounty Kuznets curve—inequality plotted against economic activity (in this case, within-subcounty inequality plotted against the

average per-capita nighttime light). Across subcounty areas, there is a strong positive relationship between average per-capita economic activity and inequality, which suggests that Uganda is on the upward-sloping portion of the Kuznets curve.[3] Levels economic development are unevenly distributed within subcounties.

Panel B plots subcounty inequality against the logarithm of population density (average population per inhabited pixel). The overall relationship is negative—subcounties with greater population density tend to have lower between-pixel, within-subcounty inequality. However, caution is required when interpreting this result. Subcounty boundaries are non-random with respect to population density. Specifically, urban subcounties tend to have a smaller geographic area and contain more people. In terms of Theil's T, the unobserved within-pixel inequality (equation (3.6)) is likely to be higher in more dense subcounties. So while between-pixel inequality decreases in population density, the relationship between population density on overall inequality is ambiguous.

The pixel-level estimates of per-capita nighttime light allow for an analysis of the geographic distribution of economic activity. Figure 17 shows the location of each quintile of the per-capita nighttime light distribution within Uganda. For example, panel A shows the geographic location (in red) of the 20% of Ugandans that reside in pixels with the lowest per-capita nighttime light. If nighttime light is taken as a proxy for income, panel A shows the geographic location of the poorest 20% of Ugandans. Broad spatial patterns emerge from the data. First, the 20% of all Ugandans located in the lowest-nighttime light pixels are centrally located, mainly around Lake Kyoga and the south-eastern portion of the country between Lake Victoria and Kenyan border. Few of the poorest 20% of Ugandans

[3]The canonical representation of the Kuznets curve posits an "inverved-U" relationship between inequality and income, with inequality peaking in middle-income regions.

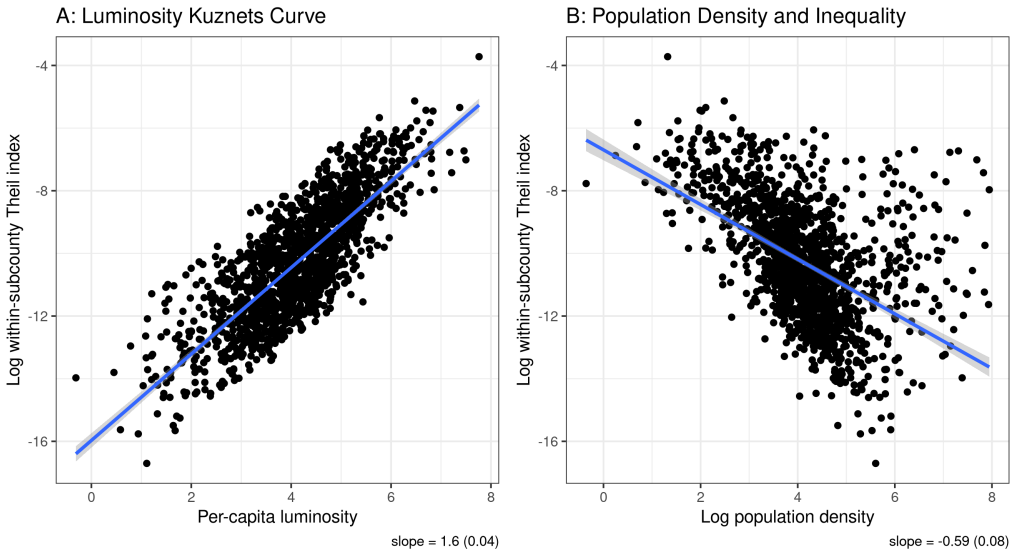


Figure 16. Estimated Relationship Between Nighttime Light, Density, and Inequality

are located in population-dense urban pixels. While possibly reflecting the urban wage premium, it is also possible that within-pixel populations of urban poor cannot be discerned at the pixel-level. This latter possibility will be taken up again in the next section.

The most widely dispersed quintile is the top 20% (panel E of Figure 17). In addition to Kampala (at the bottom-center of the map) and other urban areas the most luminous 20% are heavily represented in rural areas near national parks, perhaps indicating a high return to tourism in those areas. However, a large proportion of the top 20% pixels appear to be in sparsely populated rural areas that do not appear to be regions of significant economic activity. These areas demonstrate a potential shortcoming of using the nighttime light data as a proxy for income in sparsely populated areas, as stray light from transient activity (such as military outposts) can cause the relationship between nighttime light and economic activity to break down. To account for this possibility, panel

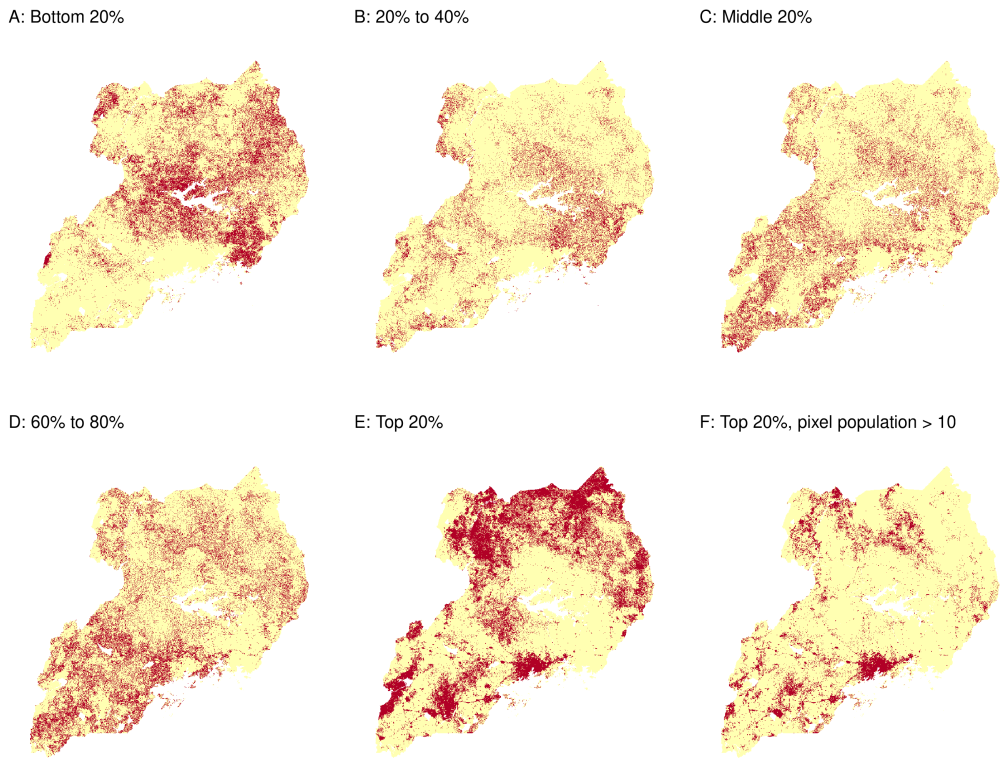


Figure 17. Estimated Geographic Distribution of Population by Quintile of Nighttime Light Distribution

F reproduces panel E but omits pixels that are estimated to have fewer than ten people. The remaining pixels indicate the top 20% of the nighttime light distribution, omitting pixels with population density lower than 40 people per square kilometer. Removing these low-density pixels results in a clearer picture of the geographic distribution of the top 20%, who appear to be located primarily in urban areas, as expected.

The geographic distribution of income quintiles (as measured by nighttime light) reveals information on urbanization patterns and the urban wage premium. Figure 18 reproduces the maps presented in Figure 17 for the Kampala region (located in the bottom center of the maps in Figure 17. Nearly all of the pixels in

Kampala urban area (center) consist of Ugandans in the top 20% of the nighttime light distribution. Pixels in the fourth quintile (60%-80%) are observed in the area immediately surrounding Kampala. Each step backward through the distribution pushes the featured set of pixels farther away from the city center in an almost concentric manner. Regional economic status and the urban wage premium are roughly monotonic in the distance from the urban center.

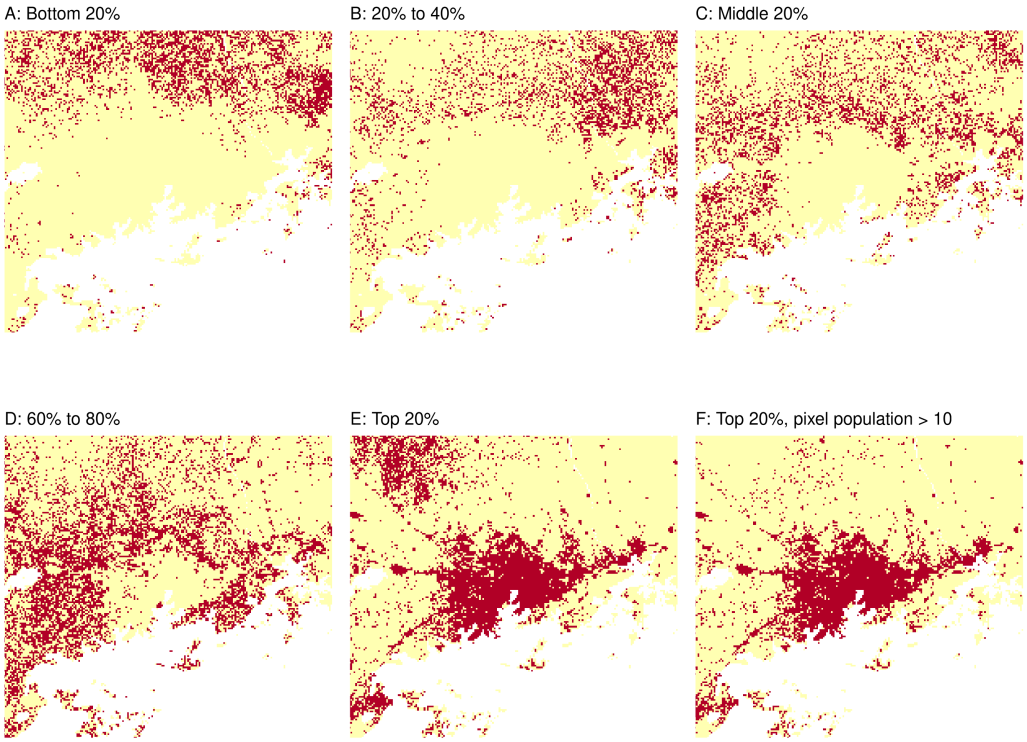


Figure 18. Estimated Geographic Distribution of Population by Quintile of Nighttime Light Distribution, Kampala Region

Poverty

Figure 17 highlights the potential to use the disaggregated per-capita nighttime light estimates to characterize a large portion of the distribution of

economic resources. In particular, binary identification of the poorest pixels can be used to estimate poverty rates across any geographic area.

I identify the geographic distribution of poverty in Uganda by identifying the poorest 41.7% of the population, as defined by the per-capita nighttime light estimates. The 41.7% figure was chosen to correspond to the World Bank's 2016 estimate of the poverty headcount ratio—the proportion of people below the poverty line—for the country. The resulting spatial distribution of the Ugandan poor is therefore approximately the combined distributions of the first two panels of Figure 17. These data can be used to estimate within-subcounty poverty rates by aggregating population counts among poor pixels at the subcounty level.

Panel A of Figure 19 plots subcounty per-capita nighttime light against the estimated headcount ratio, defined as the proportion of the subcounty population living in pixels in the poorest 41.7% of the nighttime light distribution. The apparent negative relationship suggests that poverty rates are lower in comparatively high-income subcounties. This contrasts with the inequality results presented in Figure 15, which indicated that high-income subcounties have lower inequality. These two findings need not be contradictory—lower variance of the income distribution is likely to be observed in low-income areas, particularly in rural areas with high levels of subsistence farms. Areas with higher productivity may have more opportunities for job specialization, therefore increasing the variance of the income distribution.

Of particular interest in panel A of Figure 19 is the bunching of subcounties at each extrema of the headcount range (zero and one). The pattern suggests censoring, despite the bounded range of the headcount ratio. However, the headcount ratio is estimated at the pixel-level, whereas the within-pixel headcount

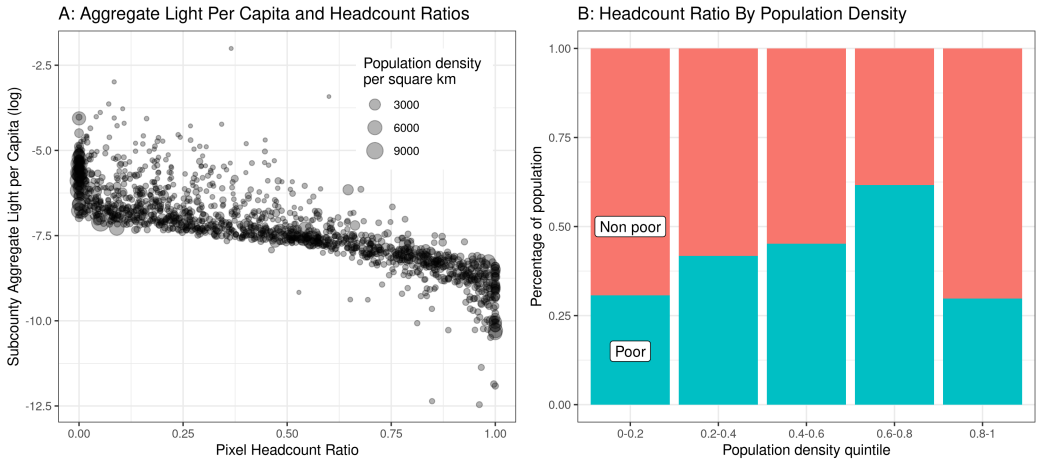


Figure 19. Subcounty Per-Capita Lights, Estimated Headcount Ratios, and Population Density

ratio may be heterogeneous. Evidence of systematic heterogeneity in the within-pixel headcount ratio can be observed by noting that the "censored" observations at the left-hand extremum (subcounties with a pixel headcount ratio of zero on the left-hand side of panel A) tend to be high density, urban subcounties. Within-pixel heterogeneity is likely to be higher in dense urban subcounties where socioeconomic conditions may vary considerably over short distances and poor inhabitants may be "free riding" on nighttime light from their wealthier within-pixel neighbors.

A similar pattern emerges at the other extremum of the headcount range (the right-hand side of panel A). The apparent censoring of subcounties with a pixel headcount ratio of one also appears to be dominated by subcounties with high population density. These subcounties are urban areas, but with with low per-capita nighttime light, and therefore are expected to have greater within-pixel poverty rates.

Further evidence of the systematic variation in within-pixel heterogeneity is presented in panel B of Figure 19, which groups subcounties by quintile of

the population density distribution. The proportion of poor pixels within the subcounties (the estimated headcount ratio) also increases as population density increases. But this trend suddenly reverses in the highest-density quintile, which has the lowest proportion of poor pixels. Taken together, the two panels of Figure 19 suggest that economic outcomes based on pixel-level per-capita nighttime light estimates are likely to be less accurate in pixels with higher population densities.

The Ugandan census does not directly measure poverty rates. However, many different count statistics reported by the census are likely to be correlated with poverty and income. For example, the census counts (1) the number of people in each subcounty that rely on a tabooda—a type of kerosene lamp—for their primary source of light. Other recorded counts that are likely correlated with income are (2) the number of people who consume two or more meals per day and (3) the number of people who rely on subsistence farming as their primary source of calories. As described previously, these counts can be dasymetrically disaggregated using the same algorithm that is used to disaggregate the population counts.

Pixel-level estimates of these three alternative poverty proxies are regressed against nighttime light to assess the ability of nighttime light to explain differences in the other economic outcomes across pixels. Regression results are reported in Table 6. While nighttime light alone is a significant predictor for each of the alternative poverty proxies in the cross-section of pixels (columns 1, 3, and 5), very little of the estimated pixel-level variation is explained by nighttime light alone. Furthermore, the sign on the nighttime light variable is the counter to the expected sign in models (1) and (5), for the cases of tabooda usage and subsistence farms. This indicates that nighttime light, by itself, is a poor measure of economic development at the pixel-level. In much applied research, however,

	Dependent variable:					
	Tabooda		$\geq$ Two Meals		Subsistance Farms	
	(1)	(2)	(3)	(4)	(5)	(6)
Nighttime light	1.364***	−0.339***	1.123***	0.326***	0.637***	−0.430***
	(0.009)	(0.006)	(0.005)	(0.003)	(0.005)	(0.001)
Population		1.561***		0.731***		0.978***
		(0.001)		(0.001)		(0.0003)
Observations	818,937	818,937	818,937	818,937	818,937	818,937
R^2	0.025	0.664	0.070	0.652	0.019	0.925

Notes: *p<0.1; **p<0.05; ***p<0.01. All variables are in natural logs.

Table 6. Luminosity as a Pixel-Level Predictor of Other Census Counts

pixel-level nighttime light is the only outcome variable available. By including pixel-level population estimates (models 2, 4, and 6), the explained variation in these alternative poverty measures increases dramatically (note also that the coefficients on nighttime light have the intuitively correct signs). Nighttime light and estimated population jointly explain 66% of the variation in tabooda usage, 65% of the variation in population consuming two or more meals per day, and 92% of the variation in counts of subsistence farms.

Pollution Exposure

As an additional illustrative application, note that advances in atmospheric science and satellite data analysis have resulted in high-resolution estimates of air pollution over much of the earth's surface. These estimates have been used in a variety of contexts and are particularly valuable in the developing world where on-the-ground pollution monitoring is scarce.

As is the case with nighttime light data, the traditional approach is to aggregate the pollution data to some known spatial unit, such as a district or municipality. Information about the finer geographic distribution of air pollution

is lost during this aggregation process. Unlike nighttime light, pollution exposure is not additive—average pollution levels within a given area are not equivalent to the average pollution exposure experienced by each individual within the area (per-capita nighttime light, in contrast, is reduced by adding an economically idle person to a geographic area). But in general, pollution exposure is the variable of interest.

Estimates of average pollution exposure can be recovered from the dasymetric disaggregation procedure. Estimated pixel-level ambient pollution levels can be weighted by pixel population estimates to produce estimate average pollution exposure at the pixel level. For small pixel sizes and pollutants are uniformly mixing at the local level, these pixel-level estimates are approximately equivalent to pollution levels experienced by each person within a pixel.

I demonstrate this process by utilizing satellite-derived $PM_{2.5}$ (airborne particulate matter smaller than 2.5 µm) estimates from NASA's Socioeconomic Data and Applications Center (van Donkelaar et al., 2018). The data consist of annual average $PM_{2.5}$ estimates that cover most of the earth's surface. Each pixel represents an estimate of the annual average $PM_{2.5}$ level at each pixel.

Aggregating these data to the Ugandan subcounty, average pollution levels for the year 2014 are shown in panel A of Figure 20. Nationwide patterns of pollution exposure are evident, with higher pollution levels in the south and southwest, and lower pollution levels in the northeast.

These subcounty averages do not contain information on within-subcounty differences in exposure. To assess the relative importance of within-subcounty exposure, I again turn to equation (3.5) and define the function $L(\cdot)$ as pollution exposure. The estimated contributions to overall inequality in pollution exposure

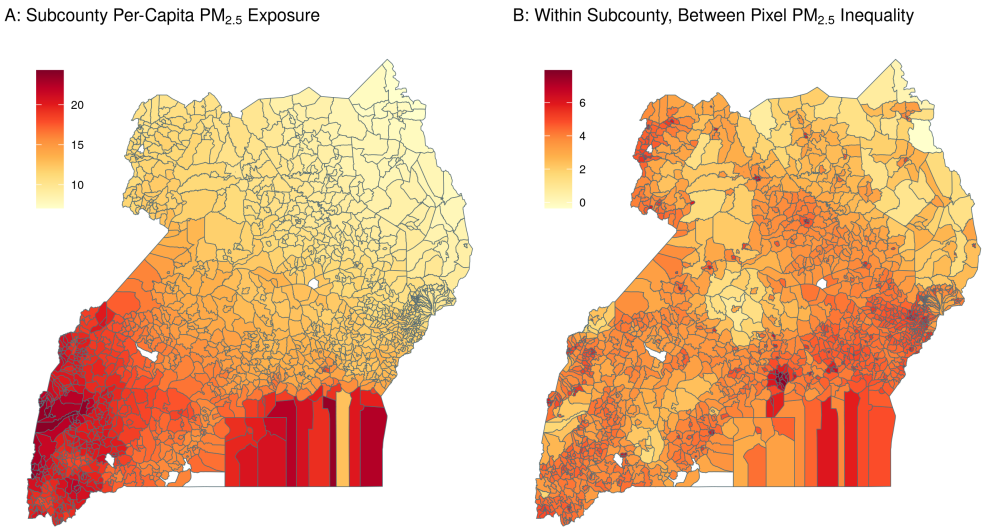


Figure 20. Estimated Distribution of $PM_{2.5}$ Exposure

are:

$$\sum_{j=1}^{|D|} \sum_{P \in P^j} \frac{L(P)}{L(N)} \ln\left(\frac{\bar{L}(P)}{\bar{L}(D_j)}\right) = 0.1493 \quad \text{(within-subcounty, between-pixel)} \tag{3.9}$$

$$\sum_{j=1}^{|D|} \frac{L(D_j)}{L(N)} \ln\left(\frac{\bar{L}(D_j)}{\bar{L}(N)}\right) = 0.5201 \quad \text{(between-subcounty)} \tag{3.10}$$

22% (equation (3.9) divided by the sum of equations (3.9) and (3.10)) of the observed pixel-level $PM_{2.5}$ exposure is due to within-subcounty variation in exposure rates across pixels.

Within-subcounty pollution inequality varies considerably across subcounties. Returning to Figure 20, panel B plots this variation. Importantly, within-subcounty inequality in pollution exposure is not highly correlated with average subcounty exposure (panel A). Applied research that uses subcounty-level exposure estimates and fails to take into account this variation could lead to attenuated or otherwise biased causal estimates in models that include per-person pollution exposure.

The disaggregated pollution exposure estimates allow for pixel-level comparisons of environmental quality and economic outcomes. One commonly discussed comparison is the environmental Kuznets curve, which describes the relationship between income and environmental quality. Figure 21 shows this relationship in Uganda, using nighttime light as an income proxy. For clarity, binned averages of $PM_{2.5}$—constructed by subdividing the range of pixel-level nighttime light estimates into 200 equally sized bins—are presented (the raw data contains 818,937 observations). The classic "inverted-U" Kuznets shape is apparent in this cross section of data. Average pollution exposure peaks at about the 25th percentile of the per-capita nighttime light distribution, the income proxy.

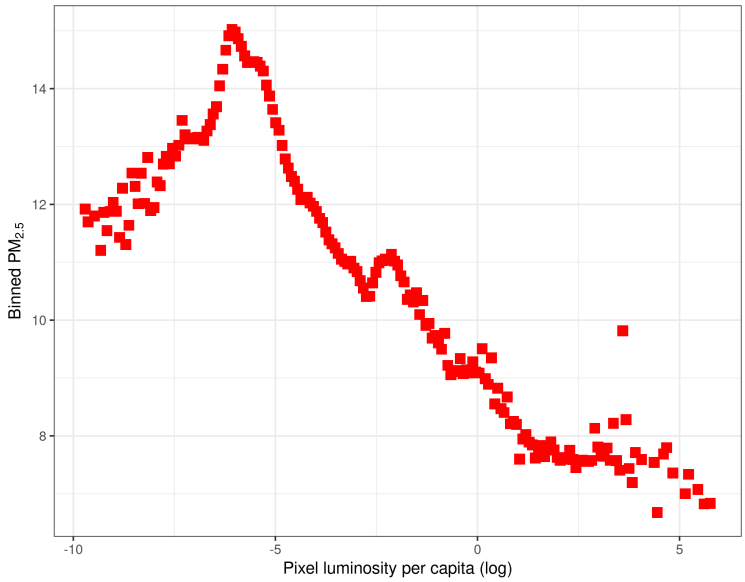


Figure 21. $PM_{2.5}$ Exposure by Per-Capita Nighttime Light

Conclusion

This paper proposes a general method for geospatial disaggregation of administrative socioeconomic data. The algorithm produces dasymetric maps that preserve aggregate counts, which eliminates errors in the aggregate. The non-

parametric nature of the procedure provides generality that can be leveraged to disaggregate a wide variety of count statistics in addition to traditional dasymetric population maps.

I demonstrate the algorithm by disaggregating four separate count statistics recorded by the Ugandan census of 2014 at the subcounty level. These variables include population and three alternative proxies for income and poverty—tabooda usage, people consuming two or more meals per day, and subsistence farms. The algorithm performs similarly well with each variable, yielding small out-of-sample errors centered on zero throughout the distribution. The algorithm is estimated using freely available, high-resolution data with minimal parametric adjustment beyond atmospheric correction of the raw satellite data.

The methodology proposed here has numerous applications. By combining the disaggregated population estimates with nighttime light data, estimates of inequality and poverty rates can be calculated at the pixel-level—a resolution smaller than what is available through official statistics. Similarly, satellite-derived pollution levels can be used to estimate pollution exposure at the pixel-level, which provides a more accurate estimation of annual average exposure at the individual level.

There are many other possible applications. In environments where economic data are scarce, census bureaus are often able to count variables that are correlated with economic outcomes. For example, estimated treatment effects of rural electrification programs may be expected to vary with respect to pre-existing electrification rates. If census takers count households with electrification in place, then disaggregated estimates of the counts may be used to calculate treatment intensity.

Finally, the flexibility of the disaggregation algorithm allows for a variety of remotely sensed data to be utilized. All of the examples presented here were calculated using three satellite sources: Landsat 8, SRTM, and Suomi NPP. But many terabytes of satellite data from countless sources are made available each day in publicly available data sets. It is likely that higher-quality data sources, or data sources that are expected to have ex ante theoretical relationships with the outcome variable of interest, can be leveraged to improve the disaggregated estimates. The flexibility and universality of the estimation procedure are its greatest strengths.

CHAPTER IV

THE EFFECT OF INDIA'S 1996 LOK SABHA ELECTION ON POLLUTION ABATEMENT AND MONITORING

Introduction

Efficient allocation of scarce environmental resources is particularly important in regions with low environmental quality. In this paper, I highlight the inefficient allocation of funds from a large-scale government program in India aimed at reducing water pollution levels in the Ganges Basin (the Ganga Action Plan, Phase II). Using the Lok Sabha (parliament) election of 1996, I show that cities where the margin of victory is small are more likely to receive government funding for pollution abatement and less likely to have routine water quality measurements after the election. These findings are consistent with a model of policymakers exchanging public resources for electoral support in subsequent elections.

River pollution is a serious concern in India. The Ministry of Environment and Forests estimates that more than two-thirds of the wastewater generated in India's cities and towns is discharged directly into waterways without treatment. As of 2011, the entire nation of India had just 234 sewage treatment plants. By comparison, the United States has 14,780 plants, despite having one fourth the population (Mauskar, 2008; The Center for Sustainable Systems, 2015). The dangerous levels of pollution in India's waterways constitute a public health disaster. Hundreds of millions of Indian citizens rely on polluted rivers for drinking water, bathing, and cleaning. In addition, many rivers are considered holy in the Hindu faith and are believed to have purifying effects. Each year, millions of pilgrims ceremonially drink from—and bathe in—dangerously polluted rivers.

Policymakers who wish to improve environmental quality in India and across the developing world are faced with high marginal costs and political economy constraints (Greenstone & Jack, 2015). Despite billions of dollars in expenditures since the mid 1980s, the various abatement schemes undertaken by the federal government of India have had little measurable affect on water pollution levels (Greenstone & Hanna, 2014). Large-scale wastewater treatment facilities are costly, but are only capable of treating a small percentage of the total wastewater created in metropolitan areas. On the political economy front, corruption and earmarking may lead to inefficient allocations of resources or a lack of information regarding regulatory compliance and pollution levels (Duflo et al., 2013).

Clientelism, defined as "the proffering of material goods in return for electoral support" (Stokes, 2011), is commonly observed in India. The social and economic structure of Indian society is particularly susceptible to clientelist politics. Dense social networks organized around small groups of individuals (*jatis*, or castes), along with a highly hierarchical social structure, create numerous opportunities for vote-buying (Anderson et al., 2014). Somanathan and Banerjee (2007) find that politicians with a similar ethnicity as the majority of an electoral district are more corrupt than "outsider" politicians because they are better positioned to exploit local social and economic networks. Using a regression discontinuity design, Lehne et al. (2018) exploit the first-past-the-post nature of Indian elections to show that public contracts for rural road development are disproportionately granted to contractors from the same *jati* as the state representative for the district.

While most of the literature on clientelism in India focuses on state and local governments, comparatively little is known about clientelism at the federal

level. Using a similar regression discontinuity design as Lehne et al. (2018), Asher and Novosad (2017) find that representation by a member of parliament (MP) from the ruling coalition is associated with improved economic outcomes, such as firm profits, employment, and output. The authors suggest that the primary channel by which politicians are able to influence business outcomes is through regulation.

This paper identifies an alternative mechanism by which clientelist politicians can influence economic outcomes: the allocation of federal government expenditures for large-scale public works projects. In the first set of results, I estimate the effects of electoral outcomes on federal pollution abatement expenditures as part of the GAP II program. Next, I estimate the effects of electoral outcomes on the expansion of the nationwide network of water pollution monitors that is administered by the federal government. I first show that abatement expenditure is partially determined by the vote-buying behavior of policymakers during a time of electoral uncertainty. Specifically, I find that cities in the Ganges Basin are more likely to receive federal funding for pollution abatement if they are in electorally competitive districts. Second, I show that policymakers are less likely to monitor water pollution levels in electorally competitive cities. Given that access to environmental information increases demand for environmental quality (Jalan & Somanathan, 2008), policymakers avoid reporting on the effectiveness of abatement schemes because they know that the high marginal cost of environmental quality means marginal investments will have little measurable effect on actual pollution levels.

The existing research most closely related to the present work is that of Cole (2009). Using a similar identification strategy, they find that agricultural credit offered by state banks increases during election years in districts that are electorally

competitive. They do not find evidence that similar increases occur in non-election years nor among private banks, indicating that direct manipulation of financial services by politicians is likely to be responsible for the increase. Importantly, they find no simultaneous increase in agricultural output among the electorally competitive districts. The clientelist targeting of electorally important districts is therefore (weakly) inefficient.

To preview the main results, I find that cities in contested electoral districts, where the difference in vote share between the ruling coalition and the opposition coalition is less than 3%, are on average between 7.0% (5.4%) and 17% (7.6%) more likely to receive funding as part of Phase II of the Ganga Action Plan (GAP II), a federal program aimed at reducing water pollution in the Ganges and its tributaries. The same cities are 2.5% (0.7%) less likely to be monitored by downstream water pollution monitoring when compared to cities that are not electorally competitive.

The paper is organized as follows. In section IV I describe the institutional setting in which this study takes place. Section IV I describe the data used in the empirical analysis. The main results are separated into two sections. I first describe the effect of close electoral outcomes on the receipt of federal abatement funding (section IV). Then in section IV I address the effects of close electoral outcomes on water pollution monitoring. The final section summarizes the main results and discusses the policy implications.

Background

Federal administration of Phase II of the Ganga Action Plan (GAP II) began in 1996 when a variety of state and federal programs were consolidated under the auspices of the National River Conservation Directorate. States apply to the

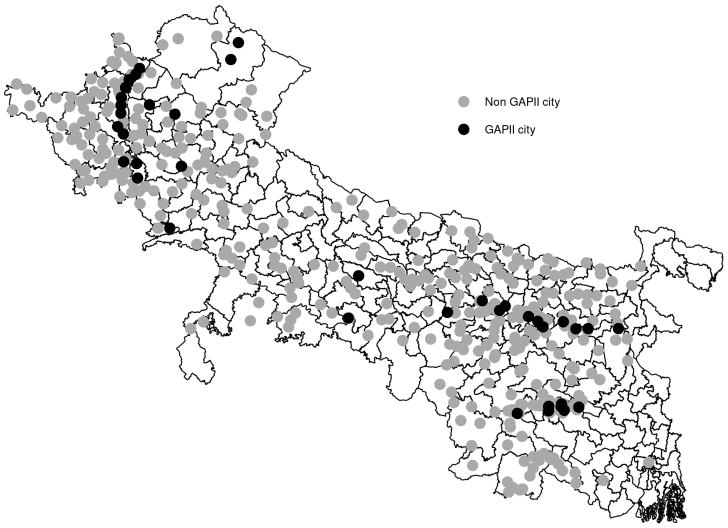
federal government for abatement projects, which may include the construction of wastewater treatment facilities, riverfront development, or the provisioning of low-cost toilets to prevent open defecation along river banks. No official guidelines for project approval appear to exist. Projects and funding allocations are highly discretionary.

GAP II abatement projects were approved in 99 cities and towns in seven states within the Ganges basin.[1] Figure 22 shows the geographic distribution of cities receiving GAP II funding, among all cities and towns in the Ganges Basin identified by the 2001 Census of India.

Little oversight is provided after projects are approved and funds are allocated. A 2007 survey of sewage treatment facilities constructed with GAP II funding by the Ministry of Environment and Forests found that more than half of the plants were rated as "poor" or "very poor" by inspectors, with only 10% of these plants being rated in "good" condition. Anecdotal reports from inspectors document many instances where local vegetation had begun to reclaim the treatment facilities (Mauskar, 2008). Kapur (2020) attributes such examples to India's "precocious democracy." Social cleavages and the relative political strength of state governments make it difficult for the federal government to provide sustained funding for small, ongoing projects. As a result, federal policymakers tend to favor large scale, highly visible programs, such as the one-time infrastructure improvements provided by GAP II.

The timing of the GAP II program coincides with a period of political upheaval in India. The historical dominance of the Indian National Congress party (often referred to simply as "Congress") had begun to diminish in the late

[1]The seven states with approved GAP II projects in the Ganges Basin are Bihar, Delhi, Haryana, Jharkhand, Uttarakhand, Uttar Pradesh, and West Bengal.



Notes: All cities and towns (defined as population greater than 5,000) from the 1991 Census of India included. Location is derived from the Wikipedia entry for each city. Ganga Action Plan, Phase II (GAP II) funding status obtained from India's Minestry of Environment and Forests.

Figure 22. Electoral Districts and Cities of the Ganges Basin

1980s, giving rise to the "third electoral system" of India (Y. Yadav, 1999). While individual *jatis* have always voted more or less monolithically, the third electoral system is characterized by voting patterns organized around large groups of *jatis*, often centered around social status. Caste identity became the dominant feature of Indian politics during this time (Munshi, 2019). From this system, the Bharatiya Janata Party (BJP)—a Hindu nationalist party to the political right of Congress—began to challenge Congress's dominance at the federal level. The general election of 1996 was the culmination of these trends. The BJP won the most seats in the Lok Sabha (parliament), but no party was able to gain a clear majority. A short-

lived government was formed by the BJP, but was quickly replaced, just 13 days later, by a ruling coalition consisting of Congress and a collection of left-wing parties (see Table 7 the post-election organization of the major political parties). The new electoral reality and the hung parliament created an expectation of imminent new elections and political uncertainty.

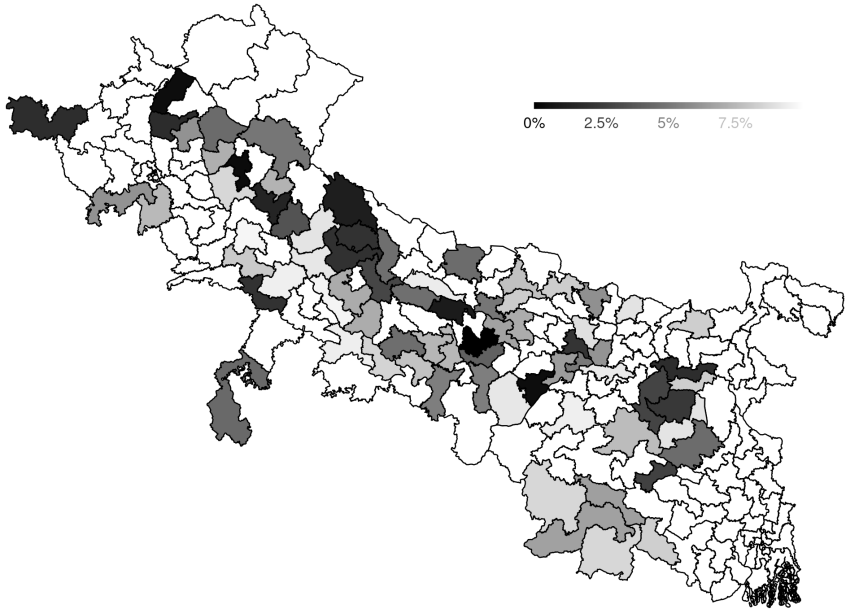
Ruling coalition	**Opposition coalition**	**Unaffiliated parties**
India National Congress	Samata Party	Bahujan Samaj Party
Janata Dal	Bharatiya Janata Party	All India Indira Congress (Tiwari)
Samajwadi Party	Haryana Vikas Party	Jharkhand Mukti Morcha
Communist Party of India		
Revolutionary Socialist Party		
All India Forward Bloc		

Table 7. Coalitions of the 11th Lok Sabha

These expectations manifested in the form of three subsequent elections and the formation of five separate governments formed within the following three years. The "third electoral system", which featured expanded voter participation among the lower castes (Y. Yadav, 1999), created a high degree of electoral volatility, with Lok Sabha seats changing hands with a frequency never before seen in India (Y. Yadav, 1999). This setting created many opportunities and incentives for vote-buying, as the ruling coalition (see table 7) could reasonably expect to lose closely-won seats and win closely-lost seats in subsequent elections.

Figure 25 shows the margin of victory between the two main coalitions that were created following the 1996 elections. Competitive electoral districts (the darker-shaded regions) are widely distributed across the Ganges Basin, with the exception of the Ganges Delta (at the bottom right of Figure 25), which is dominated by Congress and affiliated parties. The central thesis of this paper is that policymakers targeted the darker regions of Figure 25 (regions with close

elections) for GAP II funding, primarily for political reasons, rather than for reasons directly related to pollution levels.



Notes: Darker areas correspond to closer elections. Margin of victory defined as the absolute value of the difference in vote share between the ruling coalition (post election) and the main opposition coalition.

Figure 23. Winning Coalition Margin of Victory by Electoral District, 1996 Lok Sabha Elections

Data

Summary statistics for all variables used in the analysis are reported in Table 8. The main explanatory variable is the margin of victory between the two major coalitions. This is the (absolute value of the) difference in the percentage of votes earned by the leading candidate from the ruling coalition and the percentage

of votes earned by the leading candidate from the opposition coalition within each parliamentary district. The main outcome variables are the GAP II indicator (equal to one if a city received funding from Phase II of the Ganga action plan) and the monitoring station indicator(s) (which are equal one if a new pollution monitoring station is opened downstream of a city within three years of the election). Control variables include city population, distance from the city to a major river, and city nighttime light luminosity (an income proxy).

Statistic	Mean	St. Dev.	Min	Max
$\mathbb{1}$[GAP II]	0.182	0.386	0	1
$\mathbb{1}$[GAP I]	0.010	0.097	0	1
Ruling coalition share	31.229	10.711	8.229	65.981
Opposition coalition share	34.985	10.254	6.215	57.001
Vote margin	12.184	9.963	0.296	49.522
$\mathbb{1}$[New monitoring station within 10km, 1996-1998]	0.019	0.137]	0	1
$\mathbb{1}$[New monitoring station within 20km, 1996-1998]	0.033	0.180	0	1
$\mathbb{1}$[Existing monitoring station within 10km, 1995]	0.136	0.343	0	1
$\mathbb{1}$[Existing monitoring station within 20km, 1995]	0.336	0.473	0	1
Population (log)	9.965	0.978	6.886	13.771
Distance to Ganges/major tributary	0.663	0.644	0.001	2.854
Per capita luminosity	3.010	1.711	−9.460	6.810
Luminosity growth rate, 1992–1999	0.428	1.156	−1.514	11.325

Notes: Each variable has 631 observations. Unit of observation is a city within the Ganges Basin. GAP variables are indicators equal to one if a city received funding as part of Phase I or Phase II. Ruling (opposition) share is the percentage of the vote earned by highest vote-getting party in the ruling (opposition) Lok Sabha coalition. Electoral margin is the difference in vote share between a party in the ruling coalition of the 1996 Lok Sabha and the vote share of a party in the main opposition coalition. The major tributaries to the Ganges are the Alaknanda, Atrai, Ghaghara, Gomti, Koshi, Mandakini, Punpun, Ramganga, Son, Varuna, Yamuna, and Hooghly. Luminosity is the sum total of all recorded luminosity within a circle with a 10km radius from the city. Per-capita luminosity is defined as the inverse hyperbolic sine of raw luminosity divided by city population. Luminosity growth is the inverse hyperbolic sine of the ratio of luminosity in 1999 to luminosity in 1992.

Table 8. Summary Statistics

The unit of observation for all empirical specifications is a city or town in the Ganges Basin. The boundaries of the Ganges watershed were determined from a drainage direction raster calculated from a digital elevation model obtained from

the HydroSHEDS database (Lehner, Verdin, and Jarvis 2008). Cities and towns within India (and their populations) were obtained from the 1991 Census of India. The geographical coordinates of each city and town were derived algorithmically by scraping the Wikipedia article for each city. These coordinates were then matched to the Ganges watershed boundaries, creating a dataset of cities and towns within the Ganges Basin.

Centerlines for the Ganges and its major tributaries (see the Table 8 notes) were obtained from OpenStreetMaps (OpenStreetMap contributors, 2017). The distance from each city to the Ganges or one of its major tributaries was calculated as the great circle distance from the city coordinates to the closest river centerline.

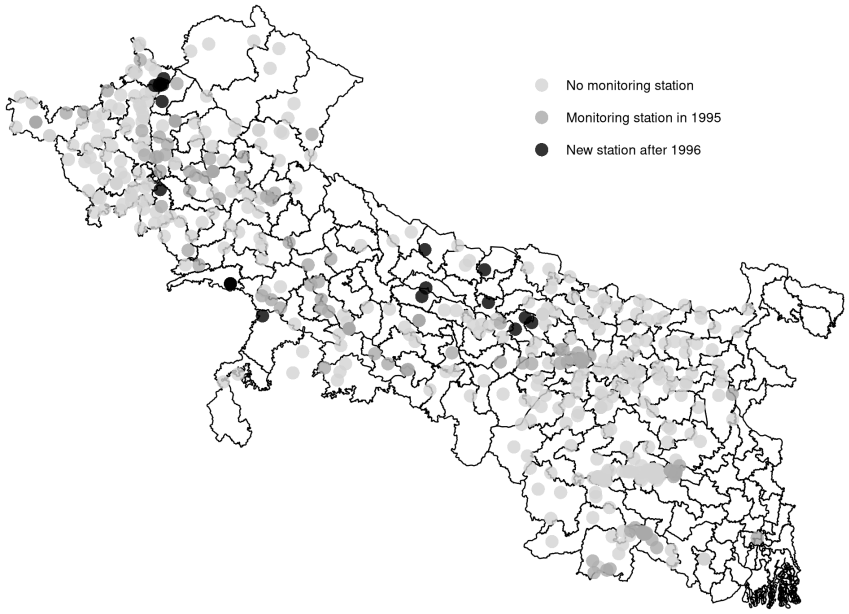
Nighttime light luminosity is used as a proxy for city income (Henderson et al., 2012). Luminosity is measured by the Operational Linescan System instruments aboard the spacecraft in the Defense Meteorological Satellite Program. For income levels, I calculate the inverse hyperbolic sine of the aggregate annual nighttime light activity within a 10km radius of the city coordinates for the year prior to the election (1995). As a proxy for income growth I calculate the inverse hyperbolic sine of the ratio of luminosity in the year 1999 to the year 1992.[2]

A list of cities that received funding from Phase I and Phase II of the Ganga Action Plan was obtained from the Ministry of Environment and Forests (see Figure 22). Electoral results and parliamentary district boundaries were obtained from the Electoral Commission of India.

Finally, the locations and operational dates for the water pollution monitoring network were obtained from India's Central Water Commission and the Central Pollution Control Board. Figure 24 shows the locations of the pollution

[2]Inverse hyperbolic sine is used instead of the natural log in order to include cities with no measured nighttime light luminosity.

monitoring stations within the Ganges Basin. The drainage direction raster was then used to match each city with all the associated downstream monitoring stations.



Notes: Includes all cities and towns (defined as population greater than 5,000) from the 1991 Census of India included. Location is derived from the Wikipedia entry for each city. Monitoring station location provided by India's Central Pollution Control Board.

Figure 24. Cities with Pollution Monitors

Close Elections and GAP II Funding

The period of political uncertainty following the Lok Sabha election of 1996 created unique incentives for policymakers which coincided with the roll-out of the Ganga Action Plan, Phase II (GAP II) program. Government expenditures

on the program exceeded Rs. 693 billion ($9 billion USD). In this section, I assess the relationship between these expenditures and the electoral results of the 1996 election. Specifically, I test the hypothesis that cities in electoral districts with a high degree of electoral competition have a higher likelihood of receiving GAP II funding than cities in non-competitive districts, ceteris paribus.

The empirical relationship between close electoral outcomes in the Lok Sabha election of 1996 and a city's inclusion in the Phase II of the Ganga Action Plan (GAP II) is modeled as:

$$Y^*_{ips} = \beta \cdot \mathbb{1}[\text{margin}_i < \bar{M}] + X_i'\gamma + \pi_p + \sigma_s + \varepsilon_{ips} \tag{4.1}$$

On the righthand side of equation (4.1), the primary variable of interest is $\mathbb{1}[\text{margin}_i < \bar{M}]$, which is a indicator equal to one if the electoral margin of victory in city i is less than $\bar{M}$. The electoral margin is defined as the difference between the proportion of ballots cast for a candidate from the ruling coalition and those cast for a candidate in the main opposition coalition, conditional on a candidate in either of the two coalitions winning the election in that district. Districts where neither of the two main coalitions are included in all regressions to help identify γ, the coefficients associated with the control variables). Intuitively, the indicator is equal to one if the district in which the city lies is electorally competitive from the perspective of both of the major coalitions. City-level controls are captured in the vector X_i. The parameters π_p and σ_s are political party and state level fixed effects, respectively. The city-specific disturbance ε_{ips} is mean zero.

The latent outcome variable Y^* is manifested in the form of the observable outcome variable Y_{ips}:

$$Y_{ips} = \begin{cases} 1 & \text{if } Y^*_{ips} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

In this section, the outcome Y_{ips} is an indicator equal to one if city i in state s represented by a member of parliament from party p received GAP II funding. For the main results, the model is estimated with a probit link function and by setting $\bar{M} = 3\%$ (elections are considered "close" if the margin of victory is less than 3%).[3] The value $\bar{M} = 3$ is selected to balance the definition of electoral closeness against the number of "treated" cities.

Estimates of the average marginal effects are reported in Table 9. The baseline Model 1 is the most parsimonious, formed by restricting to zero all right-hand side parameters other than β in equation (4.1). Cities in parliamentary districts in where the 1996 election was close (having less than a 3% margin of victory) are estimated to be 9.8% more likely, on average, to receive GAP II funding.

The primary identification assumption is the zero mean of the error ε_{ips}, conditional on the explanatory variables. The probit estimate of β is biased if the distribution of districts with close elections is nonrandom with respect to unobserved city-level characteristics that affect a city's GAP II status. For instance, it is reasonable to expect that the electoral composition of a city may be correlated with income, environmental preferences, and existing abatement infrastructure.

Potential sources of this omitted variable bias are addressed in Models 2–5. City-level controls (X_i) include population, an indicator variable equal to one if

[3]Logit and linear probability model specifications are reported in Table A.15 on page 109.

	Dependent variable:				
	𝟙[GAP II city]				
	(1)	(2)	(3)	(4)	(5)
𝟙[Electoral margin < 3%]	0.098**	0.103**	0.130***	0.112***	0.132***
	(0.044)	(0.043)	(0.047)	(0.043)	(0.049)
Population (log)		0.066*	0.064**	0.065**	0.060**
		(0.038)	(0.025)	(0.033)	(0.028)
𝟙[Distance from major river < 20km]		0.066**	0.082**	0.073**	0.090***
		(0.033)	(0.033)	(0.032)	(0.034)
𝟙[Distance from major river < 50km]		0.100***	0.097***	0.081**	0.125***
		(0.030)	(0.027)	(0.032)	(0.034)
Per capita luminosity, 1995		0.025	0.016	0.020	0.015
		(0.033)	(0.023)	(0.031)	(0.026)
Luminosity growth rate		0.023***	0.019***	0.023**	0.016**
		(0.009)	(0.007)	(0.009)	(0.008)
𝟙[GAP I city]		0.893***	0.870***	0.881***	0.866***
		(0.013)	(0.018)	(0.017)	(0.017)
Winning Party FE	No	No	Yes	No	Yes
State FE	No	No	No	Yes	Yes

Notes: *p<0.1; **p<0.05; ***p<0.01. Point estimates correspond to probit average marginal effects. Standard error of average marginal effect in parentheses. Dependent variable indicates if a city received funding from the GAP II program. Independent variables are described in the Table 8 notes on page 88. All regressions have $n = 631$ observations.

Table 9. The Effect of Close Elections on GAP II Funding

a city is within 20km of the Ganges or a major tributary, the level and growth of income (using nighttime light proxies), and GAP I funding status. The inclusion of these variables (Model 2) results in a larger and more significant estimate of the marginal effect of close elections. For the estimated positive coefficient on the indicator for close elections to be the result of unobserved omitted variable bias, the omitted variable(s) must be substantially more correlated with electoral outcomes than the observed control variables included in Model 2. This appears

unlikely given the relatively small effect that the observed variables in X_i have on the estimate of β.

Model 3 adds winning-party fixed effects, π_p. These fixed effects account for any unobserved heterogeneity across cities that vote for the same party. This accounts for the possibility of systematic differences based on electoral preferences. For instance, while the BJP garners considerable support in Uttar Pradesh, West Bengal is dominated by their political opponents from the Congress party. The cultural and religious foundations of these electoral preferences may also impact demand for abatement expenditures, but the fixed effects may account for these differences.

Model 4 adds state fixed effects to the specification (σ_s). These account for any state and federal complementarities (or rivalries) that are common in India due to the political ascendance of state governments. Once again, the estimated coefficient for the within-state effect of close elections is qualitatively similar to the other estimates, as is the estimate reported in Model 5 which includes both state and winning-party fixed effects.

Next, I assess whether GAP II status is contingent on electoral results. The clientelism exhibited by the Lok Sabha may be a reward (or punishment) for previous voting behavior, or it may be a process akin to vote-buying wherein the ruling coalition targets electoral districts in order to change future voting behavior. Empirically, this distinction can be addressed by examining the symmetry of the close election variable using a modification of equation (4.1):

$$\begin{aligned} Y^*_{ips} =& \beta \cdot \mathbb{1}[\text{margin}_i < \bar{M}] + \tilde{\beta} \cdot \mathbb{1}[\text{margin}_i < \bar{M}] \cdot \mathbb{1}[\text{ruling coalition victory}_i] + \\ & X_i'\gamma + \pi_p + \sigma_s + \varepsilon_{ips} \end{aligned} \tag{4.3}$$

The additional parameter $\tilde{\beta}$ captures the difference between GAP II funding probabilities in districts where the ruling coalition narrowly lost the election versus the districts where the ruling coalition narrowly won. A value $\tilde{\beta} > 0$ is consistent with districts being rewarded for their previous voting behavior, whereas $\tilde{\beta} = 0$ indicates that policymakers are clientelistically forward-looking.

Table 10 shows the probit estimation results for equation (4.3). The estimate for the added coefficient $\tilde{\beta}$ (second row) is statistically indistinguishable from zero across all specifications. Furthermore, the estimates for β (first row) are similar to their analogs in Table 9. These estimates are therefore consistent with forward-looking policymakers who are anticipating frequent elections and the resulting high level of government uncertainty following the 1996 election. In other words, the estimates suggest that it is electoral competition, rather than specific electoral outcomes, that are more directly responsible for the systematic differences in GAP II allocations across cities.

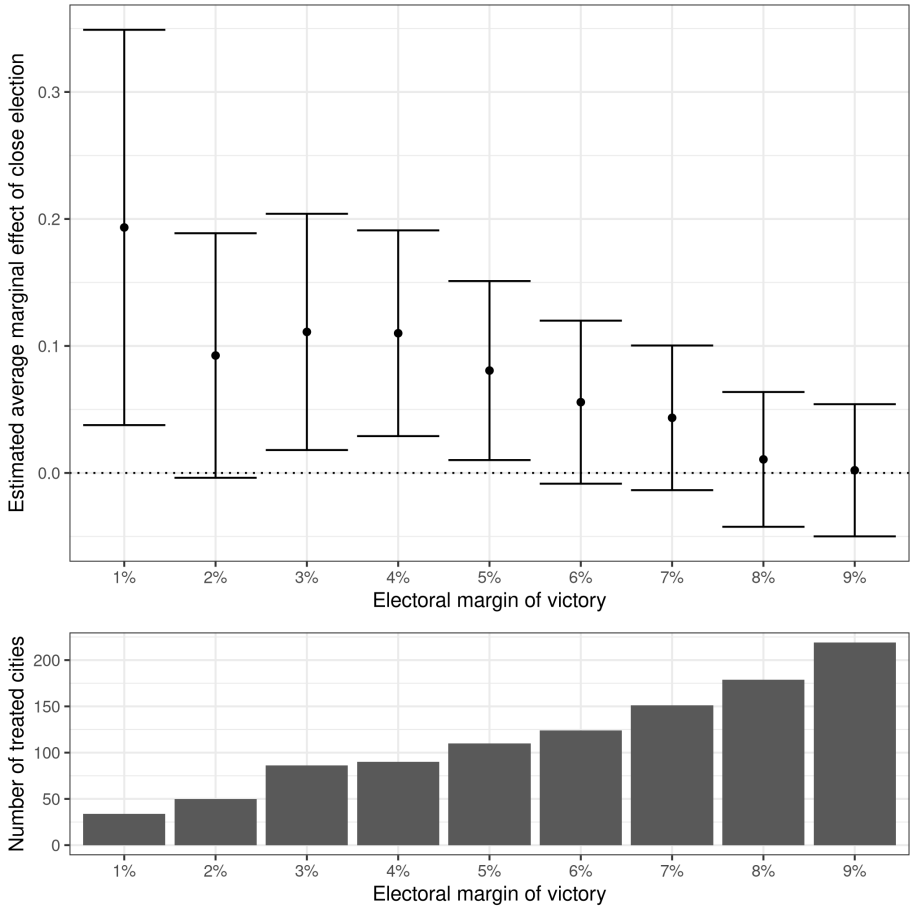
The parameter $\bar{M}$ in equations (4.1) and (4.3) determine the closeness of the election. Identification of β (the main parameter of interest) requires $\bar{M}$ to be chosen such that error term ε_{ips} is uncorrelated with the binary variable $\mathbb{1}[\text{margin}_i < \bar{M}]$, which summarizes the closeness of the election. This assumption is unlikely to hold. To assess the importance of this assumption to the empirical results, estimation of the regression equation (4.1) is repeated for $\bar{M} \in \{1\%, 2\%, \ldots, 9\%\}$. Ex ante, the point estimate of β is expected to decreases monotonically toward zero as the indicator for close electoral outcomes becomes less restrictive. Figure 25 confirms this hypothesis. Cities in electoral districts where the margin between major coalitions is less than 9% are no more likely to receive GAP II funding on average. But decreasing $\bar{M}$ to 1% increases the estimated effect

	Dependent variable:				
	𝟙[GAP II city]				
	(1)	(2)	(3)	(4)	(5)
𝟙[Electoral margin < 3%]	0.070	0.119*	0.112*	0.149*	0.098
	(0.054)	(0.064)	(0.067)	(0.079)	(0.072)
𝟙[Margin < 3%] ×	0.037	−0.017	0.034	−0.034	0.066
𝟙[Ruling coalition victory]	(0.066)	(0.048)	(0.073)	(0.048)	(0.091)
Population (log)		0.066***	0.064**	0.062***	0.061**
		(0.017)	(0.025)	(0.022)	(0.030)
𝟙[Distance from major river < 20km]		0.065*	0.084**	0.071**	0.094***
		(0.034)	(0.033)	(0.031)	(0.035)
𝟙[Distance from major river < 50km]		0.103***	0.093***	0.085***	0.126***
		(0.031)	(0.027)	(0.029)	(0.036)
Per capita luminosity, 1995		0.025*	0.016	0.017	0.016
		(0.013)	(0.023)	(0.021)	(0.029)
Luminosity growth rate		0.023**	0.019***	0.022**	0.017**
		(0.011)	(0.007)	(0.009)	(0.008)
𝟙[GAP I city]		0.893	0.871***	0.886***	0.868***
		(1.431)	(0.018)	(0.014)	(0.017)
Winning Party FE	No	No	Yes	No	Yes
State FE	No	No	No	Yes	Yes

Notes: *p<0.1; **p<0.05; ***p<0.01. Point estimates correspond to probit average marginal effects. Standard error of average marginal effect in parentheses. Dependent variable indicates if a city received funding from the GAP II program. Ruling coalition victory is an indicator equal to one if the winning party is a member of the ruling party coalition. Other independent variables are described in Table 8 notes on page 88. All regressions have $n = 631$ observations.

Table 10. The Effect of Close Elections and Electoral Victory on GAP II Funding

to 22.4%. However, decreasing $\bar{M}$ is associated with an increase in the standard error of the estimate as the number of cities coded as electorally competitive shrinks. The $\bar{M} = 3$ specification offers a good balance between standard errors and electoral competition, and therefore is the preferred specification.



Notes: Each point in the top panel represents the coefficient estimates for the indicator variable $\mathbb{1}$[Electoral margin $<$ $\bar{M}$%]. The number of treated cities (bottom panel) is the total number of observations for which the electoral margin is less than $\bar{M}$% for each electoral margin. There are 631 total observations. All other variables are the same as in Model 4 from Table 9.

Figure 25. Estimated Marginal Effects for Various Electoral Margins of Victory

The coefficient estimates in Figure 25 also mitigate concerns regarding omitted variable bias. Identification of β requires that the conditional probability of receiving GAP II funding is the same (on average) for all cities in the sample, except for differences in electoral outcomes. Any unobserved omitted variable that is correlated with both electoral outcomes and GAP II funding would threaten this identification. For instance, it is perhaps unreasonable to believe that cities in electoral districts that are highly supportive of the ruling coalition are conditionally

similar to those with more pluralistic politics. But the unobserved differences between highly competitive electoral districts and slightly less competitive electoral districts is unlikely to be as much of a concern. Figure 25 indicates that substantial differences in likelihood of receiving GAP II funding exist between cities with electoral margins less than 3% and cities with electoral margins greater than 3%. Both groups of cities contain diverse political preferences and are unlikely to differ systematically across unobserved characteristics.

Despite this evidence, it is empirically impossible to rule out the possibility that cities with competitive elections are more likely to receive GAP II funding for reasons unrelated to clientelism. However, if clientelism were not the cause of the increased likelihood of GAP II funding in electorally competitive cities, this result should hold in all cities that are electorally competitive regardless of whether a party in the ruling coalition is competitive in those districts.

In Table 11, the treatment variable is modified to estimate the effect of close elections on the probability of receiving GAP II funding among cities in districts that are not contested by a party in the ruling coalition. These are cities with a high degree of political pluralism, but they lack the same incentive for vote-buying by the ruling coalition. Models 1 and 2 (specifications with and without winning party fixed effects, respectively) show that cities with close electoral outcomes which are not contested by the ruling coalition are statistically indistinguishable from cities without close elections. Increased probability of GAP II funding is not the result of close elections per se, but rather of close elections that are contested by the ruling party.

Among cities that are contested by the ruling coalition, the mean vote share for the ruling coalition is 33%. If the composition of the electorate, rather

	Dependent variable:			
	𝟙[GAP II city]			
	(1)	(2)	(3)	(4)
𝟙[Electoral margin < 3%]×𝟙[non-competitive]	−0.104	0.033		
	(0.091)	(6.953)		
Ruling coalition share between 30%–36%			−0.063**	−0.067**
			(0.031)	(0.030)
Opposition coalition share between 30%–36%			−0.067***	−0.077***
			(0.026)	(0.026)
Population (log)	0.068***	0.065***	0.062***	0.057***
	(0.019)	(0.017)	(0.019)	(0.017)
𝟙[Distance from major river < 50km]	0.050	0.082**	0.062*	0.089**
	(0.035)	(0.038)	(0.035)	(0.039)
𝟙[Distance from major river < 20km]	0.089**	0.122***	0.083**	0.113***
	(0.036)	(0.042)	(0.036)	(0.043)
Per capita luminosity, 1995	0.026*	0.024	0.020	0.015
	(0.016)	(0.015)	(0.016)	(0.014)
Luminosity growth rate	0.025**	0.022*	0.026**	0.022*
	(0.011)	(0.012)	(0.011)	(0.011)
𝟙[GAP I city]	0.880***	0.859	0.882***	0.867*
	(0.317)	(0.667)	(0.321)	(0.490)
State FE	Yes	Yes	Yes	Yes
Winning Party FE	No	Yes	No	Yes

Notes: *p<0.1; **p<0.05; ***p<0.01. Estimates correspond to probit average marginal effects. Standard error of average marginal effect in parentheses. Dependent variable indicates if a city received funding from the GAP II program. Electoral margin is the difference in vote share between a party in the ruling coalition of the 1996 Lok Sabha and the vote share of the main opposition coalition. Cities are "non-competitive" if neither the ruling nor opposition coalition wins a seat in the election. The sample mean of coalition vote share among competitive cities is approximately 33%. Other independent variables are described in the Table 8 notes on page 88. All regressions have $n = 631$ observations.

Table 11. The Effect of Vote Share on GAP II Allocations in Non-Competitive Cities

than clientelism, is driving the differences in funding, one would expect to observe an increased probability of GAP II funding in cities that have similar shares of voters supporting one of the two major coalitions, regardless of whether the election is close. This hypothesis is tested in Models 3 and 4 of Table 11. The main explanatory variables are indicators equal to one if the vote shares of the ruling and opposition coalition are within 3% of the mean vote share among

electorally competitive cities (30%–36%). In terms of the electorate, these cities are compositionally similar to the cities with close elections, but they are not themselves electorally competitive. The estimates of the average marginal effect are negative for both indicator variables across both specifications. If anything, these results suggest that the differences in electoral composition across competitive and non-competitive cities is biasing the results toward zero.

The GAP II program was preceded by Phase I of the Ganges Action Plan (GAP I), in 1985. Though smaller in scope, the objectives and implementation of GAP I were similar in most respects to GAP II. However, GAP I was implemented during a time that was not characterized by political instability. GAP I allocations therefore offer a placebo test for the main results. The 1996 electoral outcomes should not have an effect on GAP I funding as there was less of an incentive for vote buying during this earlier period of relative stability. If, instead, the electoral results of 1996 are predictive of GAP I funding, this would be strong evidence against the clientelism explanation as political instability should not be predictive of clientelist policies in the time before the instability occurs.

Table 12 estimates equation (4.1) but replaces the outcome variable Y_{ips} with an indicator equal to one if city i received GAP I funding in 1985, eleven years prior to the political uncertainty following the 1996 election. The estimate of the average marginal effect of a close election is statistically indistinguishable from zero across all specifications. If there are any unobserved characteristics that determine both the electoral composition (the relative fraction of voters supporting each major coalition) of a city and the probability of receiving federal funds for pollution abatement, those characteristics were not present ten years prior to the sample period.

	Dependent variable:		
	𝟙[GAP I city]		
	(1)	(2)	(3)
𝟙[Electoral margin < 3%]	0.014	0.013	0.006
	(0.014)	(0.009)	(0.004)
Population (log)		0.012***	0.023***
		(0.002)	(0.004)
𝟙[Distance from major river < 50km]		0.020***	0.111
		(0.006)	(0.535)
𝟙[Distance from major river < 20km]		0.003	−0.053
		(0.002)	(0.292)
Per capita luminosity, 1995		−0.003***	0.001
		(0.001)	(0.004)
Luminosity growth rate		0.002***	0.010
		(0.001)	(0.013)
State FE	No	No	Yes

Notes: *p<0.1; **p<0.05; ***p<0.01. Point estimates correspond to probit average marginal effects. Standard error of average marginal effect in parentheses. Dependent variable indicates if a city received funding from the GAP I program. Phase I allocations were completed before the election of 1996. Independent variables are described in the Table 8 notes on page 88. All regressions have $n = 631$ observations.

Table 12. Placebo Test: The Effect of Close Elections Allocations Prior to the Election (GAP I)

Monitoring Stations

India's Central Pollution Control Board maintains a network of water quality monitoring stations (WQMS) across the country. Those monitoring stations are located along rivers and streams where periodic (monthly or quarterly) water samples are drawn and sent to laboratories for analysis. Most of the available

information regarding water quality in the Ganges Basin is derived from these monitoring stations.[4]

The WQMS network is geographically sparse. At the time of the 1996 Lok Sabha election, the average distance between each city in the Ganges Basin to its nearest downstream WQMS was 49.9 kilometers. As a result, both the level and trend in water quality in most cities is undetectable, as are the effects of abatement measures implemented at the city level.

Water pollution monitoring is relatively inexpensive, compared to abatement technologies. A rowboat and access to a river is typically all that is required to add a marginal monitoring location to the network. The marginal value of adding a specific location to the network is much harder to assess, however, and could therefore be subject to political manipulation. A ruling coalition concerned about its tenuous hold on the reins of government may be reluctant to increase pollution monitoring when pollution levels and marginal abatement costs are high, or when the marginal decrease in pollution levels that results from abatement is low. The high pollution levels in the Ganges were highly publicized in the press during the sample period, so policymakers may have been reluctant to add information pertaining to the scale of the environmental catastrophe given their inability to combat rising pollution levels.

I test the hypothesis that policymakers are unlikely to increase pollution monitoring in electorally competitive districts by estimating equation (4.1) with the outcome variable Y_{ips} set equal to one if a new WQMS opens downstream of city i within three years of the 1996 election. GAP II status—the outcome variable from

[4]See Chapter 2 for more information regarding the pollution monitoring network.

the previous section—can now be included as a (potentially endogenous) control variable.

The estimated average marginal effects are reported in Table 13. Panel A defines the outcome variable as equal to one if a new WQMS is constructed within 10km downstream, though this distance is discretionary. For robustness, panel B considers new monitoring stations withing 20km downstream.

The estimated average marginal effect of close elections on the addition of a WQMS downstream of a city is negative and statistically significant across all specifications. Water pollution monitoring is less likely to occur downstream of cities that are likely to be electorally competitive in subsequent elections. The evidence in favor of politically endogenous placement of new monitoring stations is strengthened by the negative and significant estimated average marginal effect of a city receiving GAP II funding. New abatement technology is associated with a decrease in pollution monitoring, suggesting that policymakers recognize the low expected marginal benefit (in terms of documented pollution reductions) of abatement expenditures. Large public investments in abatement that do not produce measurable reductions in pollution levels may be considered a political embarrassment to the ruling coalition.

The robustness of this result can be tested against in the context of a placebo treatment that specifies the dependent variable as new monitoring stations opened prior to the 1996 election. If WQMS location is determined by political considerations in the post election period, those same considerations should be absent before the election. Table 14 shows the estimated average marginal effect of close electoral outcomes in the 1996 general election on WQMS location in the three years prior to the election. Model 1 shows that future electoral considerations

	Dependent variable:				
	𝟙[New monitoring station, 3 years post 1996 election]				
	(1)	(2)	(3)	(4)	(5)
Panel A: New monitoring station within 10km					
𝟙[Electoral margin < 3%]	−0.026***	−0.024***	−0.024***	−0.026***	−0.022***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)
Pre-existing monitoring station, 1995		0.012	0.022	0.023	0.028
		(0.021)	(0.026)	(0.025)	(0.029)
Population (log)		0.005	0.005	0.006	0.006
		(0.007)	(0.006)	(0.006)	(0.006)
𝟙[Distance from major river < 20km]		0.032	0.029	0.031	0.021
		(0.027)	(0.027)	(0.020)	(0.018)
𝟙[Distance from major river < 50km]		−0.001	0.005	−0.046	−0.032
		(0.021)	(0.020)	(0.040)	(0.036)
Per capita luminosity, 1995		0.001	−0.0003	−0.001	−0.001
		(0.005)	(0.004)	(0.004)	(0.004)
Luminosity growth rate		−0.007	−0.001	0.002	0.002
		(0.010)	(0.007)	(0.004)	(0.005)
𝟙[GAP II city]		−0.022**	−0.026**	−0.026**	−0.029**
		(0.010)	(0.011)	(0.011)	(0.012)
Panel B: New monitoring station within 20km					
𝟙[Electoral margin < 3%]	−0.036***	−0.036***	−0.033**	−0.038***	−0.024
	(0.012)	(0.012)	(0.014)	(0.012)	(0.018)
Pre-existing monitoring station, 1995		0.017	0.016	0.023	0.030
		(0.019)	(0.019)	(0.020)	(0.022)
Population (log)		−0.0001	0.001	−0.005	−0.003
		(0.011)	(0.010)	(0.010)	(0.010)
𝟙[Distance from major river < 20km]		0.041	0.036	0.040	0.031
		(0.031)	(0.030)	(0.029)	(0.026)
𝟙[Distance from major river < 50km]		−0.004	0.001	−0.034	−0.051
		(0.026)	(0.025)	(0.034)	(0.041)
Per capita luminosity, 1995		0.006	0.006	−0.0001	0.001
		(0.007)	(0.007)	(0.006)	(0.006)
Luminosity growth rate		−0.022	−0.022	−0.009	−0.007
		(0.015)	(0.016)	(0.014)	(0.012)
𝟙[GAP II city]		−0.030**	−0.034**	−0.030**	−0.035**
		(0.015)	(0.014)	(0.014)	(0.014)
Winning Party FE	No	No	Yes	No	Yes
State FE	No	No	No	Yes	Yes

Notes: *p<0.1; **p<0.05; ***p<0.01. Estimates correspond to probit average marginal effects. Standard error of average marginal effect in parentheses. Dependent variable indicates if a water pollution monitoring station was constructed within ten (Panel A) or twenty (Panel B) kilometers downstream of a city within three years after the 1996 election. Other independent variables are described in the Table 8 notes on page 88. All regressions have $n = 631$ observations.

Table 13. The Effect of Close Elections on the Introduction of Pollution Monitoring Stations

seem to be predictive of WQMS location, but this estimate is not robust to the inclusion of other control variables (Models 2–5). Taken as a whole, Table 14 suggests that policymakers were not considering electoral politics when placing

monitoring stations prior to the period of electoral instability beginning with the 1996 election.

	Dependent variable:				
	𝟙[New monitoring station, 3 years prior to the 1996 election]				
	(1)	(2)	(3)	(4)	(5)
𝟙[Electoral margin < 3%]	−0.036*	0.005	0.002	−0.016	0.003
	(0.020)	(0.024)	(0.026)	(0.023)	(0.025)
Pre-existing monitoring station, 1992		−0.075***	−0.074***	−0.080***	−0.080
		(0.010)	(0.009)	(0.010)	(0.209)
Population (log)		0.087***	0.082***	0.126***	0.106***
		(0.019)	(0.018)	(0.028)	(0.023)
𝟙[Distance from major river < 20km]		0.241***	0.185***	0.244***	0.191
		(0.048)	(0.058)	(0.047)	(0.288)
𝟙[Distance from major river < 50km]		−0.070*	−0.069***	−0.103**	−0.089***
		(0.037)	(0.024)	(0.047)	(0.032)
Per capita luminosity, 1995		0.075***	0.070***	0.115***	0.094***
		(0.017)	(0.017)	(0.026)	(0.023)
Luminosity growth rate		−0.134***	−0.225***	−0.092***	−0.138***
		(0.027)	(0.041)	(0.031)	(0.040)
𝟙[GAP II city]		−0.064***	−0.063***	−0.065***	−0.064
		(0.008)	(0.007)	(0.008)	(1.738)
Winning Party FE	No	No	Yes	No	Yes
State FE	No	No	No	Yes	Yes

Notes: *p<0.1; **p<0.05; ***p<0.01. Estimates correspond to probit average marginal effects. Standard error of average marginal effect in parentheses. Dependent variable indicates if a water pollution monitoring station was constructed within ten (Panel A) or twenty (Panel B) kilometers downstream of a city within three years after the 1996 election. Other independent variables are described in the Table 8 notes on page 88. All regressions have $n = 631$ observations.

Table 14. The Effect of Close Elections on Pollution Monitoring Stations Prior to Election (Placebo Test)

Discussion

The unprecedented political uncertainty in India that began in the late 1990s has had long-lasting economic and political consequences for the nation. The preceding analysis identifies one such consequence: the inefficient allocation of environmental services and monitoring. However, the source of the inefficiency is all that has been identified. The magnitude of the inefficiency remains unknown.

The policy implications of this inefficiency are substantial, and therefore represent an enticing topic for future research.

Baseline pollution levels in India are high and abatement infrastructure is inadequate. Given these facts, it is possible that the marginal social benefit of increased abatement expenditure is universally high throughout the Ganges Basin. In this case, the inefficiencies created by a clientelist allocation of expenditures may be low relative to the social benefit.

Alternatively, political pluralism in India is not evenly distributed. Political affiliation is often predicted by socioeconomic status. Large, diverse cities may attract more political attention than poorer areas, and these large and diverse cities might benefit disproportionately from increased abatement expenditure.

A social-welfare-maximizing planner would allocate marginal abatement expenditures to locations where they would have the largest marginal benefit. But the planner's problem is characterized by a high degree of uncertainty about many relevant factors. The sparsity of the water pollution monitoring network means that planners have very little basic information on the location of both high-pollution areas and point sources of pollution. Similarly, the effectiveness of any marginal expenditure is unknowable in most locations. The paucity of information regarding water pollution in India leads to one prevailing and inescapable policy conclusion: the scale of India's WQMS network must be dramatically increased before any meaningful discussion of policy effectiveness can take place. Adequate monitoring is a precondition for efficient allocation of abatement expenditures.

CHAPTER V

CONCLUSION

The preceding chapters demonstrate various methods for accounting for economic settings with low quality data. The specific approaches can be summarized as follows:

1. Careful consideration of the data generating process.
2. The creation of novel datasets.
3. The identification of the areas in which marginal data collection would have the largest benefit.

As demonstrated in Chapter 2, careful consideration of the data generating process can yield significant and policy-relevant results in settings where traditional econometric methods may be biased or misleading. The geographic relationship between sewage treatment plants and water quality monitoring stations can be exploited in a spatial econometric model to reduce the bias of treatment effect estimates. Moreover, the unidirectional nature of rivers allows for specific parameterizations that are able to produce significant estimates of water pollution reduction that can be attributed to the construction of sewage treatment plants. I also find that the effectiveness of sewage treatment plants tends to diminish over time, a pattern that is in agreement with anecdotal evidence of poor management and a lack of oversight of the treatment facilities.

The second approach to accounting for low-quality data is conceptually straight-forward: create data that is of higher quality. Many terabytes of data that span the entire globe are recorded by satellites each day, and much of it is freely available to researchers. Chapter 3 introduces a method for transforming that

data into high-resolution statistics that closely track a variety demographic and economic outcomes. The resulting data offers a middle ground between aggregated statistics, as recorded by censuses or surveys, and direct data collection as part of a costly randomized control trial. The algorithm I propose is demonstrated using census data from Uganda, and I show that commonly used aggregate measures of economic welfare mask a considerable amount of heterogeneity at the pixel level.

Finally, I show that existing data sources can be used to illuminate domains in which increased data collection will have the most marginal impact. Chapter 4 identifies the inefficient allocation of pollution abatement expenditures and pollution monitoring that results from political considerations following the 1996 general election of India. I find that electorally competitive cities in the Ganges basin in the aftermath of the election were more likely to receive funding for pollution abatement from the federal government of India, and that these same cities were less likely to receive increased water pollution monitoring after the 1996 election. However, pollution monitoring is inconsistent and inadequate, which precludes estimation of the consequences of the inefficient allocations.

APPENDIX

ADDITIONAL TABLES

	Dependent variable:				
	𝟙[GAP II city]				
	(1)	(2)	(3)	(4)	(5)
Panel A: Linear Probability Model					
𝟙[Electoral margin < 3%]	0.091**	0.087**	0.105**	0.090**	0.105**
	(0.038)	(0.043)	(0.042)	(0.042)	(0.043)
Population (log)		0.066***	0.069***	0.066***	0.065***
		(0.021)	(0.020)	(0.020)	(0.019)
𝟙[Distance from major river < 20km]		0.073*	0.091**	0.078*	0.097**
		(0.043)	(0.043)	(0.044)	(0.043)
𝟙[Distance from major river < 50km]		0.087**	0.079**	0.067*	0.085**
		(0.034)	(0.032)	(0.036)	(0.036)
Per capita luminosity, 1995		0.016	0.010	0.012	0.006
		(0.010)	(0.009)	(0.010)	(0.009)
Luminosity growth rate		0.015	0.016	0.016	0.013
		(0.016)	(0.013)	(0.016)	(0.013)
𝟙[GAP I city]		0.652***	0.653***	0.655***	0.646***
		(0.067)	(0.075)	(0.070)	(0.076)
Panel B: Logit					
𝟙[Electoral margin < 3%]	0.091**	0.092**	0.130***	0.101***	0.131**
	(0.045)	(0.038)	(0.047)	(0.039)	(0.051)
Population (log)		0.082**	0.070**	0.078*	0.065*
		(0.038)	(0.034)	(0.044)	(0.038)
𝟙[Distance from major river < 20km]		0.057*	0.070**	0.064**	0.078**
		(0.030)	(0.031)	(0.030)	(0.033)
𝟙[Distance from major river < 50km]		0.106***	0.104***	0.092***	0.130***
		(0.030)	(0.028)	(0.032)	(0.039)
Per capita luminosity, 1995		0.039	0.022	0.033	0.021
		(0.030)	(0.030)	(0.040)	(0.036)
Luminosity growth rate		0.024**	0.020**	0.024*	0.017*
		(0.012)	(0.009)	(0.012)	(0.010)
𝟙[GAP I city]		0.892***	0.869***	0.884***	0.866***
		(0.013)	(0.019)	(0.015)	(0.017)
Winning Party FE	No	No	Yes	No	Yes
State FE	No	No	No	Yes	Yes

Notes: *p<0.1; **p<0.05; ***p<0.01. Estimates for the logit model correspond to average marginal effects. Standard errors in Panel A are heteroskedastic-robust. Standard errors in Panel B are estimated for the average marginal effects. Dependent variable indicates if a city received funding from the GAP II program. Other independent variables are described in the Table 8 notes on page 88. All regressions have $n = 631$ observations.

Table A.15. The Effect of Close Elections on GAP II Funding, LPM and Logit Models

	Dependent variable:				
	𝟙[New monitoring station, 3 years post election]				
	(1)	(2)	(3)	(4)	(5)
Panel A: Linear Probability Model					
𝟙[Electoral margin < 3%]	−0.026*	−0.020**	−0.021	−0.023***	−0.014
	(0.014)	(0.008)	(0.015)	(0.009)	(0.015)
Population (log)		0.003	0.005	0.006	0.006
		(0.006)	(0.006)	(0.006)	(0.006)
𝟙[Distance from major river < 20km]		0.028	0.032	0.032	0.025
		(0.019)	(0.020)	(0.020)	(0.020)
𝟙[Distance from major river < 50km]		0.006	0.007	−0.028	−0.022
		(0.012)	(0.021)	(0.019)	(0.023)
Per capita luminosity, 1995		0.002	0.001	0.001	−0.0003
		(0.003)	(0.003)	(0.003)	(0.003)
Luminosity growth rate		−0.003	−0.0001	0.002	0.003
		(0.006)	(0.005)	(0.006)	(0.006)
𝟙[GAP I city]		−0.034	−0.018	−0.011	−0.008
		(0.026)	(0.028)	(0.031)	(0.032)
𝟙[GAP II city]		−0.017	−0.028	−0.025	−0.032
		(0.016)	(0.018)	(0.019)	(0.020)
Panel B: Logit					
𝟙[Electoral margin < 3%]	−0.026***	−0.024***	−0.024***	−0.025***	−0.023***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)
Population (log)		0.005	0.005	0.006	0.006
		(0.005)	(0.005)	(0.004)	(0.005)
𝟙[Distance from major river < 20km]		0.032	0.029	0.033	0.021
		(0.030)	(0.028)	(0.024)	(0.020)
𝟙[Distance from major river < 50km]		−0.001	0.006	−0.050	−0.035
		(0.023)	(0.020)	(0.052)	(0.046)
Per capita luminosity, 1995		0.002	0.0005	0.0003	0.001
		(0.005)	(0.002)	(0.002)	(0.002)
Luminosity growth rate		−0.005	−0.0003	0.002	0.002
		(0.017)	(0.007)	(0.003)	(0.003)
𝟙[GAP I city]		−0.020***	−0.020***	−0.020***	−0.020***
		(0.006)	(0.006)	(0.006)	(0.006)
𝟙[GAP II city]		−0.018*	−0.020**	−0.020**	−0.021**
		(0.009)	(0.010)	(0.009)	(0.009)
Winning Party FE	No	No	Yes	No	Yes
State FE	No	No	No	Yes	Yes

Notes: *p<0.1; **p<0.05; ***p<0.01. Estimates for the logit model correspond to average marginal effects. Standard errors in Panel A are heteroskedastic-robust. Standard errors in Panel B are estimated for the average marginal effects. Dependent variable indicates if a water pollution monitoring station was constructed within ten kilometers downstream of a city within three years of the election. Other independent variables are described in the Table 8 notes on page 88. All regressions have $n = 631$ observations.

Table A.16. The Effect of Close Elections on the Introduction of Pollution Monitoring Stations, Linear Probability Model and Logit Specifications

REFERENCES CITED

Anderson, W., Guikema, S., Zaitchik, B. & Pan, W. (2014). Methods for Estimating Population Density in Data-Limited Areas: Evaluating Regression and Tree-Based Models in Peru. *PLOS ONE*, *9*(7), e100037.

Anselin, L. (2013). *Spatial econometrics: Methods and models*. Springer Netherlands.

Asher, S. & Novosad, P. (2017). Politics and Local Economic Growth: Evidence from India. *American Economic Journal: Applied Economics*, *9*(1), 229–273.

Auffhammer, M. & Kellogg, R. (2011). Clearing the air? the effects of gasoline content regulation on air quality. *The American Economic Review*, 2687–2722.

Azar, D., Engstrom, R., Graesser, J. & Comenetz, J. (2013). Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sensing of Environment*, *130*, 219–232.

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

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P. & Riddell, A. (2015). Stan: A probabilistic programming language. *Journal of Statistical Software*.

Colford, J. M., Wade, T. J., Schiff, K. C., Wright, C., Griffith, J. F., Sandhu, S. K. & Weisberg, S. B. (2005). Recreational water contact and illness in mission bay, california. *Westminster, CA: Southern California Coastal Water Research Project*.

Colford Jr, J. M., Wade, T. J., Schiff, K. C., Wright, C. C., Griffith, J. F., Sandhu, S. K., Burns, S., Sobsey, M., Lovelace, G. & Weisberg, S. B. (2007). Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. *Epidemiology*, *18*(1), 27–35.

Dominici, F., Greenstone, M. & Sunstein, C. R. (2014). Particulate matter matters. *Science (New York, NY)*, *344*(6181), 257.

Donaldson, D. & Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives*, *30*(4), 171–198.

Duflo, E., Greenstone, M., Pande, R. & Ryan, N. (2013). Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from india. *The Quarterly Journal of Economics*, *128*(4), 1499–1545.

EPA, The Environmental Protection Agency. (2012). 2012 recreational water quality criteria [https://www.epa.gov/sites/production/files/2015-10/documents/rec-factsheet-2012.pdf Accessed: 8/29/2016].

Fowlie, M., Holland, S. P. & Mansur, E. T. (2012). What do emissions markets deliver and to whom? evidence from southern california's no x trading program. *The American Economic Review*, 965–993.

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P. & Tatem, A. J. (2013). High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLOS ONE*, *8*(2), e55882.

Greenstone, M. & Hanna, R. (2014). Environmental regulations, air and water pollution, and infant mortality in india. *American Economic Review*, *104*(10), 3038–72.

Greenstone, M. & Jack, B. K. (2015). Envirodevonomics: A research agenda for an emerging field. *Journal of Economic Literature*, *53*(1), 5–42.

Henderson, J. V., Squires, T., Storeygard, A. & Weil, D. (2018). The Global Distribution of Economic Activity: Nature, History, and the Role of Trade. *The Quarterly Journal of Economics*, *133*(1), 357–406.

Henderson, J. V., Storeygard, A. & Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, *102*(2), 994–1028.

Homan, M. D. & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research*, *15*(1), 1593–1623.

Jalan, J. & Somanathan, E. (2008). The importance of being informed: Experimental evidence on demand for environmental quality. *Journal of development Economics*, *87*(1), 14–28.

Kapur, D. (2020). Why Does the Indian State Both Fail and Succeed? *Journal of Economic Perspectives*, *34*(1), 31–54.

Kaur, R., Wani, S., Singh, A. & Lal, K. (2012). Wastewater production, treatment and use in india. *National Report presented at the 2nd regional workshop on Safe Use of Wastewater in Agriculture*.

Khaleej Times. (2013). Record 120 million take dip as maha kumbh fest ends [http://www.khaleejtimes.com/article/20130311/ARTICLE/303119878/1028 Accessed: 8/29/2016].

Lee, J., Dawson, S., Ward, S., Surman, S. & Neal, K. (1997). Bacteriophages are a better indicator of illness rates than bacteria amongst users of a white water course fed by a lowland river. *Water Science and Technology*, *35*(11-12), 165–170.

Lehne, J., Shapiro, J. N. & Vanden Eynde, O. (2018). Building connections: Political corruption and road construction in India. *Journal of Development Economics*, *131*, 62–78.

LeSage, J. & Pace, R. (2009). *Introduction to spatial econometrics*. CRC Press.

Li, G. & Weng, Q. (2005). Using Landsat ETM + Imagery to Measure Population Density in Indianapolis, Indiana, USA. *Photogrammetric Engineering & Remote Sensing*, *71*, 947–958.

Lipscomb, M. & Mobarak, A. M. (2015). Decentralization and water pollution spillovers: Evidence from the re-drawing of county boundaries in brazil. *Forthcoming, Review of Economic Studies*.

Mauskar, J. M. (2008). *Evaluation of operation and maintenance of sewage treatment plants in india, 2007* (tech. rep.). Ministry of Environment and Forests. Delhi, India.

Mennis, J. (2009). Dasymetric Mapping for Estimating Population in Small Areas. *Geography Compass*, *3*(2), 727–745.

Michalopoulos, S. & Papaioannou, E. (2012). National institutions and african development: Evidence from partitioned ethnicities.

Miguel, E. & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), 159–217.

Munshi, K. (2019). Caste and the Indian Economy. *Journal of Economic Literature*, *57*(4), 781–834.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*, 113–162.

OpenStreetMap contributors. (2017). Planet dump retrieved from https://planet.osm.org.

Schroeder, J. P. & Van Riper, D. C. (2013). Because muncie's densities are not manhattan's: Using geographical weighting in the expectation–maximization algorithm for areal interpolation. *Geographical analysis*, *45*(3), 216–237.

Somanathan, R. & Banerjee, A. (2007). The Political Economy of Public Goods: Some Evidence from India. *Journal of Development Economics*, *82*, 287–314.

Stokes, S. C. (2011). Political Clientelism. *The Oxford Handbook of Political Science*.

The Center for Sustainable Systems. (2015). U.s. wastewater treatment factsheet. *University of Michigan*.

The Hindu. (2004). Ganga action plan bears no fruit.

Tyagi, A. (2013). *Performance evaluation of sewage treatment plants under nrcd* (tech. rep.). Central Pollution Control Board, Ministry of Environment and Forests. Delhi, India.

van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M. & Winker, D. M. (2018). Global annual pm2.5 grids from modis, misr and seawifs aerosol optical depth (aod) with gwr, 1998-2016. *NASA Socioeconomic Data and Applications Center (SEDAC)*.

Wade, T. J., Sams, E., Brenner, K. P., Haugland, R., Chern, E., Beach, M., Wymer, L., Rankin, C. C., Love, D., Li, Q. et al. (2010). Rapidly measured indicators of recreational water quality and swimming-associated illness at marine beaches: A prospective cohort study. *Environmental Health*, *9*(1), 1.

Wiedenmann, A., Krüger, P., Dietz, K., López-Pila, J. M., Szewzyk, R. & Botzenhart, K. (2006). A randomized controlled trial assessing infectious disease risks from bathing in fresh recreational waters in relation to the concentration of escherichia coli, intestinal enterococci, clostridium perfringens, and somatic coliphages. *Environmental health perspectives*, 228–236.

Yadav, S. (2009). If we do not change the national river conservation plan, the money will continue to get wasted [http://indiatoday.intoday.in/story/'If+we+do+not+change+the+National+River+Conservation+Plan,+the+money+will+continue+to+get+wasted'/1/76482.html Accessed: 8/19/2016]. *India Today*.

Yadav, Y. (1999). Electoral Politics in the Time of Change: India's Third Electoral System, 1989-99. *Economic and Political Weekly*, *34*(34/35), 2393–2399.