Firm Presence, Pollution, and Agglomeration: Evidence from a Randomized Environmental Place-Based Policy

Michael Gechter* and Namrata Kala†

January 18, 2026

Abstract

Firm location decisions impose externalities on other firms due to competitive or agglomerative forces, and on the environment. We study an environmental place-based policy that randomly moved 20,000 firms in New Delhi. Relocation reduces pollution, but firm exit increases. We combine the exogenous assignment of firms to industrial plots with a model to estimate spillovers on neighboring firms, showing that firm survival rates could have been increased by allocating firms to plots accounting for input-output linkages. These results provide causal evidence on how firm presence impacts environmental quality, and how spillovers can be used to minimize costs on regulated firms.

*University of Arkansas; Y-RISE. Email: mgechter@uark.edu

†MIT Sloan School of Management; BREAD; CEPR; NBER; J-PAL. Email: kala@mit.edu.

We are grateful to the International Growth Center (IGC), MIT Sloan School of Management, the National Science Foundation (NSF), Penn State University, Private Enterprise Development in Low-Income Countries (PEDL) and Structural Transformation and Economic Growth (STEG) for financial support. We would like to thank Jonathan Eaton, Glenn Ellison, Robert Gibbons, Michael Greenstone, Paul Grieco, Dominic Leggett, Melanie Morten, Rohini Pande, Claudia Steinwender, Adam Storeygard, Daniel Xu, and seminar audiences at ASU, Berkeley, ISI Delhi, MIT, NBER, Oxford, Penn State, Stanford, UBC, UChicago, University of Pennsylvania, University of Warwick, and Yale for helpful comments, and Christina Antonakakis, David Cai, Zixu Chen, Caio Figueiredo, Valentina Goetz, Shuhang Lou, Hanna Marsho, Esma Ozer, Garima Singal, and Charlotte Wachter for excellent research assistance. All errors are our own.

1 Introduction

Firm location decisions are one of the most important choices managers make, optimizing factors such as proximity to customers, suppliers, and useful information. At the same time, these decisions may have spillover effects on local neighborhoods, by impacting environmental quality and contributing to local economic activity or agglomeration. For this reason, numerous policies attempt to change location choices of firms. For instance, place-based policies incentivize firms to locate in particular areas and zoning regulations restrict their location choices. Location restrictions that seek to limit pollution exposure in particular have a long history, starting with the first zoning laws introduced in the early 20th century in New York, in part to improve environmental quality (Wilson et al., 2008). The inherent endogeneity of firm location decisions renders estimating the causal impact of firm presence on the local economy difficult. Policies that shock firm location decisions can help overcome this but in many cases are bundled with other features. For instance, place-based policies that incentivize firms to locate in certain areas often provide local infrastructure or tax benefits, and may simply displace economic activity rather than creating it.

In this paper we study a policy tool commonly deployed in developing countries to impact firms' location choices: industrial relocation. These policies usually involve the mandated movement of existing firms operating in high-density areas to more remote locations, often, though not always, for environmental reasons. Industrial relocation has been used in several countries, including China, India, Japan, and South Korea, but little is known about the impacts on relocated firms, or whether such policies achieve their environmental goals. Like many place-based policies, they are typically bundled with other policy initiatives, making it difficult to understand their causal impacts. Furthermore, they are usually implemented uniformly over an urban area so there is no clear point of comparison.

We study a large-scale, environmentally-motivated industrial relocation policy in New Delhi, which moved over 20,000 firms to industrial areas outside the city over several years. A unique feature of this policy is that, due to a shortage of industrial plots when relocation began, plot allotment was done via a series of lotteries between 2000 and 2015, with the largest being held prior to 2006. This generates random variation in firm presence over the time period, between neighborhoods with a greater number of firms receiving a plot earlier in the process, and those with a greater number of firms receiving a plot later in the process (conditional on the total number of firms relocated from a neighborhood). It also generates random variation within the industrial areas on who a firm's neighbors are since, conditional on plot size and lottery year, the allocation of firms to plots was also random, something we verify using simulated lotteries.

In the first part of the study, we use the random timing of firm removal to estimate the causal impact of firm presence on neighborhood environmental quality, specifically the policy's target: air pollution. We use a relatively fine definition of neighborhood, a 1km by 1km grid-cell, which is the level at which our air pollution data is measured. This allows us to test whether such policies achieve their environmental objectives. The type of manufacturing firm targeted by the relocation, namely very small to medium-sized, predominates in developing countries so credibly identifying their emission profile is key to informing policymakers on the potential effects of policies targeting them. A reduction in air pollution should not ex ante be taken for granted because most developing countries have limited regulatory capacity. The relocated firms might move back, be replaced by other polluting firms, or pollution may increase due to, for example, the policy's enabling growing vehicular emissions. We combine several data sources with administrative data on the policy.

We use the lottery timing as an instrument for firm removal to the industrial areas across each program year, controlling for the total number of firms that were relocated from a neighborhood. We show that across years, lottery timing has a strong first stage in predicting when a firm took possession of the plot in the industrial area. In neighborhood-level nonlinear instrumental variable (IV) regressions, we find the average neighborhood impacted by the relocation experiences a 1.31 $\mu g/m^3$ drop in particulate matter (PM) levels, about 1.37 % of the mean pre-relocation fine PM concentration for the average neighborhood. Since industrial pollution contributes about 10-12% to Delhi's PM 2.5 (Sharma et al., 2018), relocation reduces industrial pollution in Delhi by about 12.5% for the average neighborhood. The pollution reductions due to industrial relocation are unevenly distributed across space, however, with the 95th percentile neighborhood experiencing a 7.39 $\mu g/m^3$ drop in PM. The most-affected neighborhoods are also among the most populous, in keeping with the goals of the policy.

We conduct several robustness checks for this result, some of which we mention here. First, we show that the latitude and longitude of sending regions with more early vs. late plot assignments via lottery are similar. That is, neighborhoods' location does not predict early vs. later relocation, as we would expect given that the timing is decided randomly. Second, we show consistent results from a complementary two-way fixed effect specification that estimates the effect of the proportion of firms allotted plots early within discrete bins of the total number of firms ever relocated from a neighborhood. Third, we present results using the lottery timing instruments at different levels of temporal aggregation.

In the second part of the paper, we take advantage of the random assignment of firms to industrial area plots (stratified by plot size and year of the lottery) as a lens to understand firm behavior. Specifically, this feature of the policy offers a unique opportunity to understand the causal effects of proximity to competition and input-output linkages on firm outcomes, a type of spillover which has been hypothesized since Marshall (1920), but has been challenging to identify. The largest industrial area, which contained about 15,000 firms had 74 clusters of 150-300 firms, and we use within-cluster variation in the proportion of upstream, downstream and own product firms to study how cluster composition impacts firm survival. This allows us to causally estimate the direction and magnitude of firm spillovers at a fine level of industrial classification, while sidestepping the omitted variables that incentivize firms of certain types to co-locate in observational data. Indeed, the results show that presence of a particular industry can generate in some instances positive and in other instances negative impacts on firms in other industries. We show that a firm's position in the input-output matrix relative to its neighbors, along with net competitive effects, explains the large majority of spillover effects it experiences.[1] Having more upstream firms has the largest positive impacts on firm survival, with a one SD increase in the number of upstream firms within a cluster increasing the likelihood of firm survival by 5.2 percentage points (p.p.), about 20% relative to mean likelihood of firm survival. Downstream firms also increase survival, though the effects are smaller, with a one SD increase in such firms within a cluster increasing firm survival by 2.4 p.p. In contrast, a one SD increase in having firms producing the same product within a cluster reduces the probability of firm survival by 2.2 p.p., indicating that competitive forces are also significant for such firms. We also estimate these spillovers estimates over a range of distance thresholds (rather than within a firm cluster) ranging from 200 meters up to 1km, and show that they operate over small physical distances, and decay with increasing distance, consistent with results in Arzaghi and Henderson (2008), Rosenthal and Strange (2020), and Baum-Snow et al. (2023).

We then model firms' exit decisions as the outcome of the first period of a dynamic game of incomplete information played in each cluster of the largest industrial area, using our causal spillover effect estimates to approximate conditional choice probabilities. We leverage the structure of the model to generate counterfactual exit rates under optimal assignment of firms to plots, using firm spillovers optimally to maximize aggregate firm survival. Compared to the lottery policy's uniform assignment of industries to clusters, the optimal assignment which groups industries with positive spillovers would have

[1] This is in contrast to more aggregated agglomeration analyses which are typically unable to differentiate between positive and negative spillovers (e.g. Ahlfeldt et al. (2015)).

increased aggregate firm survival by 8.2 to 15 percentage points, depending on the stringency of our definition of what constitutes an upstream or downstream linkage (31.5 and 57.2 percent of the average survival rate, respectively).

Finally, we conduct a back of the envelope cost-benefit exercise for the policy. Relocation reduces air pollution in the sending regions, but is costly for the relocated firms. 74% of firms in the largest industrial area comprising about 15,000 firms had exited by 2018. That is, only 26% had survived. To estimate changes in survival probability for relocated firms that were caused by the policy (rather than the "natural" rate of exit), we use the distance firms are relocated. Every kilometer relocated reduces the probability of firm survival by 0.4-1 p.p, illustrating the importance of endogenous location choice and the costs of being moved from that location. Using this estimate, we find that relocation caused an increase in relocated firms' exit rate of about 30 percentage points. We conduct a back of the envelope cost-benefit analysis, converting the reduction in PM levels to the statistical value of lives saved and comparing this to costs associated with firm death and increases in PM in the industrial areas. We find that the benefits outweigh the costs. Notably, optimal assignment of firms to plots taking into account spillover effects would cut the effect of relocation on firm exit by between 31 and 58%, showing the potential of using information on firm spillovers to reduce costs for regulated firms.

The paper builds on several literatures, and lies at the intersection of the impact of environmental policies on firms as well as the returns to agglomeration. First, we contribute to the literature on the impact of environmental regulation targeting firms (Blackman et al., 2010; Chen et al., 2021; De Simone et al., 2024; Do et al., 2018; Fan et al., 2019; Fenske et al., 2023; Foster and Kumar, 2011; Greenstone and Hanna, 2014; Harrison et al., 2015; He et al., 2020; Karplus et al., 2018; Muller and Mendelsohn, 2009; Ryan, 2012; Shapiro and Walker, 2020; Song et al., 2022).[2] We study a unique experiment allowing us to identify long-term impacts of a widely-used type of environmental regulation, one that mandates polluting firms to move out of populated neighborhoods. Such policies are common in low state capacity settings, but relatively little is known about their impacts. A related but distinct literature studies how environmental regulation that does not explicitly target firms' location decisions nevertheless impacts entry and exit decisions (Henderson, 1995; Levinson, 1996; List et al., 2003). These include tests of the pollution haven hypothesis, which posits that firms move to areas with lower environmental standards (Brunnermeier and Levinson, 2004; Choi et al., 2023; Jaffe et al., 1995; Millimet and Roy, 2016; Tanaka et al., 2022). In contrast, we study how environmental policies that target firm

[2]Also related is recent experimental work that randomizes particular aspects of environmental regulation such as auditor discretion Duflo et al. (2018) or emissions trading (Greenstone et al., 2022).

location impact environmental quality and costs for firms.

Second, we are part of a growing literature investigating the nature of agglomeration economies for firms (the classical foundation of this literature being Rosenstein-Rodan (1943), Murphy et al. (1989), Glaeser et al. (1992), Ciccone and Hall (1996), and Henderson et al. (1995). See Combes and Gobillon (2015), Rosenthal and Strange (2020), Duranton and Puga (2020), and Bryan et al. (2024) for excellent reviews of empirical applications in this literature). This empirical literature focuses on measuring the magnitude of such effects (Baum-Snow et al., 2023; Combes et al., 2012; Ellison et al., 2010; Gechter and Tsivanidis, 2020; Greenstone et al., 2010; Kline and Moretti, 2014; Leonardi and Moretti, 2023; Tsivanidis, 2023; Vitali, 2022). This literature is largely (though not exclusively) focused on developed countries, though the nature and magnitude of such effects might vary in developing countries due to differences in firm size, market frictions, and other government policies (Gechter and Tsivanidis, 2020; Vitali, 2022). Furthermore, in this literature, experimental variation is (unsurprisingly) almost entirely absent. Nakajima and Teshima (2017) is an exception, using random assignment of firms within a fish market in Tokyo to investigate the specific channel of shopping externalities. The relocation policy we study affects firms in a variety of manufacturing industries in Delhi, and the random assignment of firms to specific plots in the industrial area allows us to separately identify the effects of these industries on each other's profitability (for instance, upstream vs. downstream relationships), similar to Ellison et al. (2010) but with the benefit of experimental variation. Moreover, we also estimate how firm presence impacts environmental quality.[3]

Before proceeding, we define two key units of analysis used throughout the paper. A *neighborhood* is a 1km $\times$ 1km grid cell, the resolution at which our pollution data is measured. Neighborhoods are our unit of analysis for studying the effects of firm removal on environmental quality. A *cluster* is a block within an industrial area containing 150 - 300 firms. Clusters are our primary unit of analysis when studying spillovers among relocated firms in their new locations.

[3] A second direction in this literature evaluates whether place-based policies such as industrial areas and enterprise zones truly improve firm outcomes, or only displace economic activity from regions not covered by the place-based policy to regions covered by the place-based policy (see Glaeser and Gottlieb (2008) and Neumark and Simpson (2015) for excellent reviews of this literature. Other related work studies place-based policies in developed (Busso et al., 2013; Criscuolo et al., 2019; Greenbaum and Engberg, 2004; Hyman et al., 2022; Neumark and Kolko, 2010; O'keefe, 2004) and developing economies (Betcherman et al., 2010; Chaurey et al., 2023; Hasan et al., 2021; Lu et al., 2019)). Our estimates of firm spillovers do not have this concern of displacement causing an upward bias in any firm spillover estimates, since we make use of a large and known set of firms that were moved and co-located.

2 Context

2.1 Air Pollution in India: Impacts and Related Policies

Air pollution is a leading risk factor for premature death, accounting for over 8 million deaths worldwide in 2021 (HEI, 2021). It is a highly significant public health concern in India, where nearly 100% of the population lives in areas where annual fine PM concentrations exceed WHO recommendations (Greenstone and Fan, 2018). About 1.6 million deaths, 17% of all deaths in India, were attributable to air pollution in 2019 (Pandey et al., 2021), demonstrating the significant health impacts of persistent high pollution exposure.

India has several comprehensive environmental laws, including the Water Act of 1974 and Air Act of 1981. Beginning in the 1980s and 1990s, several pollution reduction policies were initiated, a primary one being the Supreme Court Action Plans (Greenstone and Hanna, 2014) (SCAPs). These plans were comprehensive measures that 17 polluted cities were directed to take by the Supreme Court, and comprised bundles of pollution reduction measures such as fuel switching or firm closures and relocation. Mandates regarding the location of polluting industries were a primary means to reduce industrial pollution mentioned in these action plans. For instance, 13 of 17 Action Plans mention environmentally-driven location mandates like closure or industrial relocation (Harrison et al., 2015).

This focus on regulating firms' location choices to reduce industrial pollution continues in the present. The National Clean Air Programme (NCAP) policy is currently the flagship air pollution reduction policy initiative by the central government, and calls for city-specific abatement plans to reduce air pollution. Relocation of polluting industries continues to be a commonly used tool in these plans.

2.2 Relocation Policy In Delhi

In 1999, the Supreme Court mandated the relocation of manufacturing firms in Delhi that were operating in residential areas, with exemptions for certain types of household industries. The Government started developing three industrial areas on the edges of the city to house these firms (a very small fraction of firms, about 2.5%, were given plots in other industrial areas around Delhi). Firms were asked to apply to the program to be allotted a plot in one of these areas. Only firms that had been registered before 1999, were in a residential neighborhood, and were in an industry that was to be relocated were eligible (Singh, 2007).[4] These firms comprised a large range of small manufacturing firms, including automobile parts, food processing, and rubber and plastics producers. Table 3 presents the most frequent goods produced by these firms.

There was some initial confusion about eligibility for relocation, and 50,000 firms applied to the program because if they were relocation-eligible and did not move, they would have to shut down operations altogether. The government deemed that about 21,000 were actually eligible for relocation and should be assigned plots in the industrial areas. Our analysis focuses on random industrial area plot allotment within the 21,000 firms that applied when the program was announced (pre-2000), and so is internally valid over the relevant sample of relocation-eligible firms. Since the number of industrial plots was limited (more were developed over time), plots were allocated via a series of lotteries over time from this pool of applicant firms. The plot assignment lotteries were simple random draws conditional on the year and plot size category from the pool of eligible firms. For instance, the first lottery assigned plots to about 11,000 firms in the first industrial area in the year 2000 in one of four size categories (100 m^2, 150 m^2, 200 m^2, and 250 m^2). The next large lottery, in 2003, assigned plots to 3,687 firms that had not yet been assigned a plot, and so on. In total, there were 4 large lotteries (with over 2,000 firms each relocated), and many small ones between 2001 and 2015.

Most firms assigned a plot by 2004 were assigned to two industrial areas. In 2005, a third industrial area became operational, and the majority of firms assigned a plot in 2005 or later were assigned to this third area. This difference caused a significant delay in a firm moving, however. The average year in which a firm allocated a plot by 2004 took possession of their plot in the industrial area is in 2005, while the average year in which a firm relocated later (after 2004) received a lease was 2015, a ten-year delay.

Firms that were not allotted an industrial plot in the earlier lotteries could continue operating while they waited for a plot, and once allotted a plot had to move their operations within 3 years. We will show that firms that were assigned plots earlier also were able to take possession of their plot earlier (the first step in moving, which was followed by a lease being issued), generating exogenous variation in the timing of their departure from their original location. The random timing of plot assignment generates random variation in firm presence that we use to estimate impacts of relocation on pollution (more details are provided in Section 4).

The majority of plots had been assigned by 2005, and only about 500-600 firms received an assignment in 2010 or later. Each firm was allotted an industrial plot ranging

[4] Some neighborhoods were deemed to be exempt from relocation, since they were not adequately residential.

from 28 m^2 (which was a spot in a building housing several small firms) to standalone plots of 250 m^2 - the average plot size was between 100 and 150 m^2. Firms were given concessional loans to allow them to build their factories in the allotted plots, and were given leases for these plots. They were not allowed to sell or rent them, and were technically supposed to continue producing the same products they had done while located within Delhi.[5]

Of the three industrial areas, the largest one (Bawana) housed the majority of relocated firms (over 15,000), and included plots of four size categories (100 m^2, 150 m^2, 200 m^2, and 250 m^2, comprising 50.9%, 26.3%, 5.1% and 17.69% of the plots, respectively). We were able to obtain stylized maps of Bawana which we used in combination with Google Earth to get the coordinates of each firm's final allotted plot location (see Figure A4). After additionally geocoding their original address, we estimate how far firms were moved. We find that the average firm relocated to Bawana was moved 20.2 km (12.5 miles) from its original location– the 25th percentile is 16.2km (10.1 miles), while the 75th percentile is 27.3km (17 miles).

All the industrial areas were developed by the government body charged with implementing the relocation policy, and within the industrial area the plot allocation was random as well (conditional on year of plot assignment by lottery and plot size category. e.g. amongst firms to be assigned a plot of 100 sq m in 2000). Figure A1 shows Google Earth images from 2001, 2005, and 2010. Panel (a) shows plot boundaries being delineated but no firm movement yet, while panels (b) and (c) show limited and then significant development of the industrial area in 2005 and 2010, respectively.

Construction was the responsibility of firms so the presence of structures indicates firms actually moving into their plots. We used satellite imagery to identify plots where structures were constructed in one of the five large sectors in Bawana. We found construction on 3048 out of 3392 plots (90%) , demonstrating that a very large fraction of firms to be relocated started operating in the industrial areas. Table 6 presents results from surveyor visits conducted in 2021 (data collection discussed in detail in Section 3), and shows that firm removal from origin locations was persistent, with firms present at less than 10% of the sites they visited.

[5]The 2018 Firm Survival Census in Bawana (described in more detail in the next section) showed that only about 1% of firms changed trade, so this was a relatively rare occurrence.

3 Data

3.1 Air Pollution

To estimate impacts of relocation on pollution, we need high-resolution data on pollution before the relocation began (in 2000) as well as after. This is challenging, since pollution monitors were relatively sparse for Delhi during this time period, with only one air quality monitor reliably reporting data over the entire this period. We use data on fine particulate matter (PM 2.5, or fine PM) initially constructed in van Donkelaar et al. (2016). These data are constructed by combining satellite retrievals of aerosol optical depth, chemical transport modeling, and ground-based measurements (for more details on the latest version of the dataset, refer to Van Donkelaar et al. (2021)). They have been extensively in prior work on the impacts of pollution as well as the impact of regulation on pollution, recently by Greenstone and Fan (2018), De Simone et al. (2024), Barrows et al. (2019), and Behrer et al. (2023).

The data have several advantages for our analysis. First, they are available at the 1km by 1km resolution, providing data at a fine spatial resolution and starting before the policy began to be implemented.[6] Second, they cover fine PM, which is an important air pollutant, responsible for large damages to human health globally (McDuffie et al., 2021). We use the 1km by 1km grid cells as our definition of neighborhood, since this is the finest level for which pollution data are available, and create an annual measure which is our outcome variable for environmental quality.

3.2 Administrative Data on Relocation Policy

The administrative data on the relocation policy is available from the government body that was responsible for the relocation (the Delhi State Industrial and Infrastructure Development Corporation Ltd., DSIIDC). The data include firm name, original address from where they were relocated, details such as applicant name, date of the lottery when they were allotted a plot, a free text description of the firm's products entered by the owner, as well as the ID of the plot they were allocated in the industrial area. It also includes details on the timing of the dates when the firm took possession of the industrial area plot and when the firm's lease began. 21,748 firms received a plot assignment in an industrial area and 21,174 have non-missing information on year of the lottery in which the firm was as-

[6]We validate the Van Donkelaar et al. (2021) data using available pollution monitor readings. Monitor readings are notably noisy and intermittently available, but the rank correlation between the monthly average reading by station and the corresponding Van Donkelaar et al. (2021) data point is nevertheless strong: 0.82.

signed a plot. In Table A1, we show that the availability of lottery year information is not correlated with neighborhood latitude and longitude, i.e. it is uncorrelated with firms' starting geography.[7]

Figure 1 presents the cumulative probability of having been assigned a plot by a given year as well as the probability of having taken possession of the industrial plot by year, showing a positive relationship between the two measures. We present regression results in the next section.

3.2.1 Latitude and Longitude of Sending and Receiving Addresses

Geocoding the original addresses using Google's API, we assign a firm to a neighborhood. Combining this information with the timing of when a firm was given a plot, we create a neighborhood-level dataset of the number of firms that were allotted a plot via lottery each year, as well as the number of firms who took possession of the plot in the industrial area.

We were also able to identify the longitude and latitude of plots within Bawana by digitizing maps we acquired from the industrial associations in this area, and cross-referencing street names in these stylized maps with Google Earth. Figure A4b shows a photograph of the map for one cluster, and Figure A4a a stylized map of the entire industrial area.

3.2.2 Firm Survival Census in 2018: Bawana Industrial Area

In 2018, DSIIDC conducted a census of all plots in the Bawana industrial area, to determine whether the firm assigned to a plot was still operating there. We classify an assigned firm as being present if it was found in the industrial area, and having exited the market if it was found to be closed in this census, or to have (illegally) rented or sold its plot to another firm (a much smaller fraction, about 1%, changed their trade). This definition gives us the figure of about 26% of firms surviving in the industrial area by 2018.

3.2.3 Product Classification

The firm application data include free text information on what firms produce. We determine the 3-digit Annual Survey of Industries Classification Code (ASICC) to assign to each firm a product code through a combination of human judgment and a custom GPT

[7]For another 8.6% of firms, Google geocoding of their addresses places them outside of Delhi. We show that the propensity for this to happen or for the geocodes to be missing altogether is uncorrelated with the timing of the lottery in Table A7, which we discuss further in Section 3.3.1.

pipeline. We discuss the details in Appendix E. From this exercise, we get narrow (3-digit) assignments for 9,200 of about 15,800 firms in Bawana.

For each firm, we create a measure of what proportion of other firms in their assigned cluster in the industrial area are producing the same product. To understand upstream and downstream linkages, we use the 2010 Annual Survey of Industries (ASI) data. From the 2010 ASI, we first retain only firms producing a single product (to avoid the issue of assigning inputs to multiple outputs, which, like in other firm panel datasets, is not available in the ASI), and only keep 3-digit products that have at least five firms producing them. We create product-level expenditure on each input, and generate the proportion of input expenditure for each input.

We use product-level revenues and costs to generate measures of upstreamness and downstreamness. We call product x upstream of product y if the input expenditure producers of y spend on x is above the median amount producers of y spend on any input in the 2010 ASI data. We call product x downstream of product y if producers of y's revenues from selling to producers of x are above the median of producers of y's revenues from selling to any other type of producer To guarantee that the downstream–upstream matrix is symmetric, so that if firm A is downstream of firm B then firm B is upstream of firm A, we combine the two definitions. Specifically, firm A is considered downstream of firm B if either the downstream condition described above is satisfied or if B is upstream of A according to the upstream rule. The upstream definition is completed analogously. For each firm, we create a measure of what proportion of other firms in their assigned cluster in the industrial area are upstream of them.

For each firm, we store counts of the number of other firms in their assigned cluster in the industrial area are downstream and upstream of them, as well as analogous measures for the number of firms within circles of radius of 200 - 1000m. The sum of proportion upstream and proportion downstream give us the proportion of firms in the industrial area that have any input-output linkage with each firm. Our results are robust to alternative definitions of upstream and downstream, including replacing the median with the 75th percentile in these calculations.

3.3 Primary Data Collection

3.3.1 Surveyor Visits to Baseline Addresses of Relocated Firms

To ensure that any geocoding measurement error in assigning firms to neighborhoods is independent of lottery timing, we sent surveyors to the origin addresses of about 15,000 firms (about 71.4% of the total relocated population). [8] Once surveyors located an address, they used GIS-enabled devices to record a manual geocode. The probability that geocodes are missing from either manual or Google geocoding is not correlated with the year a firm was assigned a plot by lottery (see Table A7, and omitting them does not impact the results for pollution.). Moreover, the number of firms assigned a plot by lottery by 2004 according to Google maps and the surveyors are highly correlated, with a correlation of 0.97 at the neighborhood level.

Along with entering a manual geocode, surveyors recorded the current use of the address (see Table 6). It is from this component of the data collection that we infer firm removal was persistent, with any firm present in only 10% of cases.

4 Impact of Relocation on Environmental Quality: Empirical Strategy and Results

4.1 Balance Tests: Firm Original Location

Our use of lottery timing as an exogenous factor determining the level of firm presence in a neighborhood depends on the fact that year of plot assignment by lottery is independent of neighborhood characteristics. To test whether year of plot assignment by lottery is uncorrelated with firms' starting geography, we run the following pair of regressions at the firm level.

$$\begin{aligned} InitialLongitude_j &= \alpha + \kappa YearOfAssignmentByLottery_j + \epsilon_j \\ InitialLatitude_j &= \tilde{\alpha} + \tilde{\kappa} YearOfAssignmentByLottery_j + \tilde{\epsilon}_j. \end{aligned} \tag{1}$$

If year of plot assignment by lottery is unrelated to firm j's original address, κ and $\tilde{\kappa}$ should be estimated as zeros.

The results from estimating equation system (1) are given in Table 1. The coefficient estimates are in degrees, and as expected are very small and statistically insignificant. This indicates that firms assigned plots in specific lotteries were not geographically clustered, as expected. To visualize potential differences in firm origin location by year of plot assignment by lottery beyond mean longitude and latitude, Figure 2 plots the longitudes and latitudes of firm origin addresses by year of plot assignment by lottery. We see no systematic differences in the origin locations of firms by the year they were assigned plots by lottery.

[8] We sent surveyors to 15,811 addresses but they could not find 1,328.

In Appendix B, we further show that small differences in lottery dates can translate into substantial differences in the timing of firms departing their origin locations in part because of a ten-year average difference between in plot possession year between firms assigned plots by lottery between 2000 and 2004, and after (see Table B1). We perform another check of the independence of year of plot assignment by lottery from firm origin location by simulating each of the yearly lotteries, and comparing the actual vs. simulated number of firms relocated early (by 2004) at the neighborhood level. We find the actual and simulated number of firms relocated early are highly correlated (correlation of over 0.95).

4.2 Impacts

In our primary specifications, we use the distribution of firms by year of plot assignment by lottery in a neighborhood to instrument for the number of polluting firms remaining there. This gives us an estimate of the effect of an additional polluting firm on PM 2.5 concentration. We begin by estimating a simple linear specification with full heterogeneity in effects by year using two-stage least squares (2SLS),

$$Y_{it} = \beta_{0,t} + \beta_{1,t} FirmsWithoutPossession_{it} + \beta_{2,t} TotalFirmsRelocated_i + \epsilon_{it}. \tag{2}$$

i denotes a neighborhood, t the year of measurement, and Y_{it} the PM 2.5 concentration in neighborhood i in year t. $FirmsWithoutPossession_{it}$ is the number of firms in origin neighborhood i that had not taken physical possession of their plot in the industrial area by year t. $TotalFirmsRelocated_{it}$ is the number of firms ever relocated from neighborhood i.

Our instruments for $FirmsWithoutPossession_{it}$ are elements of the distribution of firms slated for removal in neighborhood i across different lotteries. Neighborhoods with more firms assigned plots in later lotteries will have more firms remaining (without possession) in any given year t, since firms assigned plots later move later. We control for a neighborhood's total number of firms relocated so that our effects are of lottery-induced variation in the number of active polluting firms among neighborhoods with similar initial numbers of polluting firms. In Appendix C, we experiment with alternative specifications of the total number of firms relocated interacted with proportion relocated early (by 2004), with little change in the results.

Our preferred measure of the distribution of firms in a neighborhood across lotteries aggregates lotteries by year category. Specifically, we define the following 4-element instrument vector Z_i at the neighborhood level: number of firms with a year of plot assignment less than or equal to 2002, number of firms with a year of plot assignment of 2003, number of firms with a year of plot assignment of 2004, and number of firms with a year of plot assignment of 2005. Firms with a draw year greater than 2005 are in the omitted category.

This specification of the distribution of firms in a neighborhood across lotteries is a transparent way of addressing a bias-variance tradeoff. Fully disaggregating the distribution of firms across lotteries so that the instrument vector has counts of the number of firms in the neighborhood assigned plots in each of the 35 lottery dates allows for more accurate predictions of the number of firms having not taken possession of their plot in the industrial area in a given year. This is because, for instance, firms assigned plots in the very last lottery will not have taken possession in earlier years. However, the accuracy comes at the cost of having to estimate more first-stage parameters which increases variance. This cost is particularly substantial because later lotteries only assigned plots to small numbers of firms.

We plot the 2SLS estimates of $\beta_{1,t}$ by year with 90% confidence intervals in Figure 3. Point estimates are around a 0.1 μg/m^3 increase in PM2.5 for each polluting firm remaining in a neighborhood, rising slightly over time. The estimates in most years are statistically significant. In Figure 4 we plot the annual first stage Montiel Olea and Pflueger (2013) effective F-statistics. As discussed in Andrews et al. (2019), standard F-statistics are not appropriate to test for weak instruments in overidentified models in the presence of heteroskedasticity. We can compare the effective F-statistics to Montiel Olea and Pflueger (2013)'s critical values for rejecting the null hypothesis that 2SLS bias exceeds τ% of a of a worst-case benchmark (which coincides with OLS bias under homoskedasticity). We set $\tau = 10\%$ and test at the 5% significance level, as is standard following Stock and Yogo (2005). With these parameters, the average critical value over the years from 2005 to 2015 is 16.26.

Instrument strength as measured by the effective F-statistic increases over time, which is intuitive given that Table B1 shows the longest delay from being assigned a plot to taking possession of it in the industrial area occurs between firms assigned plots by 2004 and after. This contrast starts plays a growing role in explaining the number of polluting firms remaining in a neighborhood as time goes on, peaking around 2013, then declining as more and more neighborhoods have all their polluting firms moved to the industrial area.

The similar point estimates and variance in the year-by-year effects suggest there may be precision benefits from aggregating across years, which we pursue next. Column 1 of

Table 2 presents results from the following pooled linear IV regression across years.

$$Y_{it} = \beta_0 + \beta_1 FirmsWithoutPossession_{it} + \beta_2 TotalFirmsRelocated_i + \omega_t + \epsilon_{it},$$

where ω_t is a year fixed effect. We cluster standard errors at the neighborhood level to allow for arbitrary correlation of errors within neighborhoods over time. Similar to the results from the annual regressions, we find a statistically significant 0.0948 μg/m^3 increase in PM2.5 for each polluting firm remaining in a neighborhood. The effective F-statistic of 56 far exceeds the Montiel Olea and Pflueger (2013) critical value of 17.9, indicating that the instruments are strong.

We next explore nonlinear effects of firm presence on air pollution. In addition to being policy-relevant in their own right (since nonlinearity affects the benefit-cost ratio of the policy), they also help understand distributional effects. This is because relocation of firms displaces air pollution from one part of the metropolitan area to another (more sparsely populated) area further from the city center. If effects are non-linear, relocation may increase or decrease the total amount of air pollution in addition to displacing it.

Our preferred instrumental variable specification therefore includes a quadratic term in the number of firms remaining in a neighborhood in year t.

$$\begin{aligned} Y_{it} = {} & \beta_0 + \beta_1 FirmsWithoutPossession_{it} + \beta_2 FirmsWithoutPossession_{it}^2 \\ & + \beta_3 TotalFirmsRelocated_i + \omega_t + \epsilon_{it} \end{aligned} \tag{3}$$

To identify β_2, we need instruments able to move the square of the number of firms remaining in a neighborhood in year t independently from the level. We therefore include the square of each of the four elements of Z_i in the first stage. In this preferred specification we additionally increase power by incorporating time heterogeneity in the first stage mapping from the distribution of years of plot assignment across firms in a neighborhood to the number of firms remaing in that neighborhood at time t. Time heterogeneity arises because, for example, firms assigned plots in 2005 are unlikely to have taken possession of a plot in the industrial area in that year but, as delays resolve, they become more likely to have taken possession.

We incorporate time heterogeneity into our first stage model by interacting the elements of the instrument vector with the time trend $t - 2005$. Our preferred first stage specification is therefore:

$$FirmsWithoutPossession_{it} = \rho_0 + \rho_1' Z_i + \rho_2'(Z_i \cdot (t - 2005)) + \rho_3' Z_i^2 + \rho_4'(Z_i^2 \cdot (t - 2005)) + \omega_t + \nu_{it}. \tag{4}$$

Column 2 of Table 2 presents 2SLS and first-stage results from estimating Equations (3) and (4). As discussed, having more firms assigned plots in 2005 is initially associated with having more polluting firms remaining in a neighborhood, but this effect reverses over time. For other plot assignment years, the negative association between the number of firms remaining increases in magnitude over time, which is logical. Several of these trend effects are statistically significant, both in predicting the number of firms remaining in a neighborhood and its square. To test for weak instruments in separately identifying the effect of the number of firms remaining in a neighborhood and the effect of the square of that number, we draw on recent theoretical work on the topic and compute Lewis and Mertens (2025)'s g_{min} statistic. The g_{min} statistic of 93.93 exceeds the critical value of 39.72 with $\tau = 10\%$, tested at the 5% significance level, again indicating strong instruments.

Turning to second-stage results, the coefficient on the number of firms remaining in a neighborhood is statistically significant and positive. The relationship is concave, however, since the coefficient on the square of the number of firms remaining is negative and statistically significant. The marginal effect of an additional firm at the mean number of firms relocated from neighborhods with at least one firm relocated is a statistically significant 0.055 μg/m^3 increase in PM2.5.

We conduct several robustness checks in Appendix Table A2. In Column 1, we address the bias-variance tradeoff in selecting instruments more formally by using Belloni et al. (2012)'s post-double-lasso method to select elements from the fully disaggregated distribution of number of firms by individual lottery in each neighborhood. We also use the most predictive of the maximally-aggregated distribution, collapsing the distribution vector to a single element, the number of firms assigned plots by 2004 (Appendix Table A2, Column 2). Results are similar to those in Column 1 of Table 2 across these two specifications. In Column 3 of Appendix Table A2, we run the same quadratic specification as in Column 2 of Table 2, but cluster at a higher level of spatial aggregation, the 2km by 2km grid cell level, to allow for arbitrary spatial correlation at this level. Results are very similar to those in Column 2 of Table 2, with slightly larger second-stage standard errors and a slightly larger critical value for the g_{min} statistic. Results are also robust to restricting the sample to neighborhoods with any relocated firms or to those with surveyor-identified relocated firms.

In Appendix C, we estimate two-way fixed effect (TWFE) specifications using the interaction of the proportion of firms assigned plots by 2004 and an indicator for a year after 2004, as our measure of treatment intensity. These specifications differ conceptually from the reduced form of the specification in Column 2 of Appendix Table A2 in using pre-relocation periods to difference out time-invariant differences in PM 2.5 concentration across neighborhoods. Results are again similar to those in Table 2.

In sum, the results show firm relocation does causally and significantly reduce air pollution, a primary source of mortality and morbidity. In Section 6, we consider a back of the envelope calculation that considers how the benefits of firm removal compare to the costs.

5 Agglomeration and Counterfactual Optimal Industrial Area Design

In this section, we use the policy to estimate the causal impacts of the industrial composition of neighboring firms, as hypothesized by theories of agglomeration going back to (Marshall, 1920). Random assignment of individual plots to firms (within a year and plot size) generates independent variation in the characteristics of a firm's neighbors. We estimate the effect of the industrial composition of firm i's cluster in the industrial area on its long-term survival. We then use these estimates to search for an exit-minimizing alternative to DSIIDC's uniform assignment of firms to plots . We find that 31-58% of the effect of relocation on exit could have been avoided with an optimal industrial area design taking into account neighborhood composition effects.

5.1 Balance Tests for Assignment of Plots in the Industrial Area

We first test whether the plot assignments by product are consistent with random allocation. To simulate the original lotteries, for each lottery year and plot size category, we pick a random set of firms and then assign them randomly to available plots of that size category. For instance, if x firms were assigned 100 m^2 plots in 2000, we randomly pick x of all the 100 m^2 firms and randomly assign them to the 100 m^2 plots that were assigned in 2000. This generates simulated data on the number of firms producing a product in each of the 74 industrial clusters in Bawana. We present the actual and simulated data in Figure 5a. The two distributions are extremely similar, and the Kolmogorov–Smirnov test for equality of distribution functions fails to reject that the two are equal (p-value=0.92). As an additional robustness check, we also present results using the proportion of firms in a cluster producing a product (rather than the number). The results are in Figure 5b and also show that the distributions are very similar (The Kolmogorov–Smirnov test for equality of distribution functions has a p-value of 0.86).

5.2 Neighborhood Composition Effects through Net Competition and Input-Output Linkages

We follow Marshall (1920)'s theory of agglomeration in explaining the effect of the composition of the industrial cluster a firm is assigned to on the firm's survival. Marshall's theory is based on the benefits to firms of minimizing transportation costs in accessing customers, inputs, and ideas. Since the relocated firms largely sell to other firms, we measure the differential access to customers generated by the random allocation of a firm's neighbors by the number of downstream neighbors in the firm's cluster. We measure differential access to inputs by the number of upstream neighbors in the cluster. The number of neighbors producing the same good as the firm could have a negative effect on survival through competition, but also a positive agglomeration effect due to transmission of ideas or a thicker market for the kind of labor required to produce the good. The effects we measure are the net of these two forces, so we refer to them as net competitive effects.

We consider the following partially-linear model of how these cluster composition measures affect the survival of firm i in cluster k, with the parameters of interest being $\alpha_{upstream}$, $\alpha_{downstream}$, and α_{own}:

$$\begin{aligned} Active_{ik} =& \alpha_{upstream} NumberUpstream_{ik} + \alpha_{downstream} NumberDownstream_{ik} + \alpha_{own} NumberOwn_{ik} \\ &+ g\left(\{1\{product_i = m\}\}_{m=1}^{M}, NumberMissing_k, TotalNumberFirms_k\right) + \epsilon_{ik}. \end{aligned} \tag{5}$$

$Active_{ik}$ is an indicator for firm ik's being found active in the 2018 census of the Bawana industrial area. $NumberUpstream_{ik}$ is the number of cluster k's assigned firms who report producing a good upstream of i's product. $NumberDownstream_{ik}$ is the number of cluster k's assigned firms who report producing downstream of i's product. $NumberOwn_{ik}$ is the number of block k's assigned firms producing the same product as i. $\{1\{product_i = m\}\}_{m=1}^{M}$ is a vector of product indicators, with M being the total number of products. $NumberMissing_k$ is the number of firms assigned to cluster k for which a product code could not be assigned, and $TotalNumberFirms_k$ is the total number of plots in cluster k.

Given that we have 180 distinct 3-digit ASICC codes being produced in Bawana, with a skewed distribution of product shares, we take steps to reduce dimension to proceed. We do so using Belloni et al. (2013)'s post-double-LASSO method. We select which fixed effects to include, and whether the other control variables should be included, by running LASSO regressions on a linear approximation to Equation (5) and the treatment variables $NumberUpstream_{ik}$, $NumberDownstream_{ik}$, $NumberOwn_{ik}$. We additionally present a robustness check where we include lottery fixed effects (identified by year and plot size) in this specification in the Appendix (Table A3).

Results are presented in Table 4, where we have divided the composition variables by their standard deviation. Column 1 shows that a one standard deviation increase in the number of firms with any input-output linkage in a cluster increases the probability of firm survival by 5.7 p.p., about 22% relative to the mean survival rate. In contrast, a one standard deviation increase in the number of firms producing the same product has a smaller and negative impact on survival (about 2.2 p.p, or 7% of the mean). The second column shows the impact of input linkages and output linkages separately. Both upstream and downstream linkages matter, but the point estimate on upstream linkages is bigger, about 5.2 p.p, or 20% relative to mean survival rate (however, we cannot reject that these two coefficients are equal). We are able to reject that upstream linkages have the same effect as own-product firms, and the coefficient on own product firms, as in Column 1, negatively impacts firm survival. These results show the significant impact that input-output linkages can have in improving firm survival. Columns 3 and 4 present results using the 75th percentile threshold for defining upstream and downstream rather than the median. Results are similar to those Columns 1 and 2, respectively.[9] We present results adding lottery fixed effects to the LASSO, in Table A3-the results are nearly identical to the main specification. We also present unstandardized estimates in Table A5.

Do these effects decay over distance? To test this, we re-estimate Equation (5) for five different distance thresholds: 200 meters, 400 meters, 600 meters, 800 meters, and 1000 meters. We present these (unstandardized) results in Table A6. Column 1 shows having one additional upstream firm within 200 meters increases the likelihood of firm survival by 0.2 p.p. while one additional downstream firm increases it by 0.06 p.p. One additional firm producing the same product reduces the likelihood of firm survival by 0.3 p.p. The impacts of input-output linkages persist as distance increases up until 600m, after which they are not statistically significant. All point estimates also reduce as distance

[9] We additionally present unstandardized coefficients from this specification in Table A4. One more upstream firm within a cluster increases the likelihood of firm survival by 0.2 p.p., one more downstream firm increases it by 0.07 p.p., and one firm producing the same product reduces the likelihood of firm survival by 0.3 p.p. (Column 2). Results using the 75th percentile cutoff for defining input-output linkages are qualitatively similar.

increases, consistent with agglomeration forces operating over small physical distances and decaying with distance.

Thus, competition and input-output linkages have significant causal effects on firms' long-term survival probability. In addition to being interesting in their own right, these estimates also allow us to generate counterfactual estimates of how much the costs to firms can be reduced by leveraging these spillovers. We describe the construction of these counterfactual estimates and the results in the next subsection.

The specification above assumes that relevant firm spillovers occur through Marshallian forces: input-output linkages and net competitive effects. Is this consistent with the data? In Appendix D, we estimate a much more general model allowing for arbitrary and asymmetric spillovers from the number of producers of each product n on the exit decision of a producer of product m. In this model, each of these spillovers can be determined by the input-output relationship between m and n, by net competition, or by features idiosyncratic to the pair. To address the high dimension of the asymmetric spillover parameter vector, we take a Bayesian approach to inference. We find a posterior mean magnitude of the idiosyncratic effects that makes up only 2.5% of the magnitude of all forces driving spillover effects, with a 90% credible interval of [0.004, 0.142]. This shows that Marshallian forces, as we have modeled them in this section, are indeed driving the large majority of spillover effects across firms.

5.3 Counterfactual Optimal Industrial Area Design

5.3.1 Model of Firm Exit Decisions

We can use the spillover effect estimates from Section 5.2 to evaluate the aggregate effect of alternatives to the uniform assignment of firms to plots in the largest industrial area. To do so, we model firms' decision to remain active in each cluster of the Bawana industrial area as the first period of a dynamic game of incomplete information. Uniquely for the dynamic games literature, we observe the initial condition of each game as the random assignment of firms to the cluster (see e.g. Berry and Compiani (2021) for issues arising from the unobservability of the initial state in dynamic games).

Firms are defined by the same characteristics as in Equation (5): the ASICC code of the product they produce, the plot assignment year, and their plot size allotted. Size allotted proxies for firm size. We denote by $Product_{ik} \in \{1, \ldots, M\}$ the ASICC code of firm i in cluster k, and by $Lottery_{ik} \in \{1, \ldots, L\}$ the lottery in which firm i was assigned a plot, where L is the total number of lotteries. In addition to these observed characteristics, following Benkard et al. (2020) we assume each firm experiences a profit shock ν_{ik} with support $\mathcal{V}$, which is unobserved to us as researchers and to the other firms. We denote the initially-assigned vector of producer shares for cluster k by $s_k \in \Delta^{M-1}$, where Δ^{M-1} is the $(M-1)$-dimensional simplex. This vector of shares constitutes the state variable in the model summarizing conditions following entry of the assigned firms in cluster k. Given s_k, $Product_{ik}$, $Lottery_{ik}$, and ν_{ik}, firm ik chooses whether to remain in operation or exit.

The strategy for firm i is the function $\sigma_{ik} : \Delta^{M-1} \times M \times L \times \mathcal{V} \to \{0, 1\}$, where 0 represents the choice to exit and 1 the choice to stay active. Conditional on the set of firms assigned to each cluster, we assume that a pure strategy Bayes Nash Equilibrium (BNE) exists and that the same BNE is played across all clusters in the data. Under this setup the conditional probability of choosing to stay active is given by the following expression,

$$
\begin{aligned}
&P(Active_{ik} = 1 | s_k, Product_{ik} = m, Lottery_{ik} = l) \\
&= \int \{\nu : \sigma_{ik}(s_k, m, l, \nu) = 1\} dG(\nu | s_k, Product_{ik} = m, Lottery_{ik} = l),
\end{aligned}
$$

the measure of the set of ν values such that firm i chooses to stay active. Uniform random assignment of firms to plots also means regions of the support of the conditioning set of the choice probabilities will be well-represented, which can be an issue in applications of Benkard et al. (2020)'s methodology. More importantly, random assignment of firms to clusters make the conditional choice probabilities causal in the sense that they reflect the expectation of *potential* $Active_{ik}$ with s_k set to s, $Product_{ik}$ set to m, and $Lottery_{ik} = l$. Differences in conditional choice probabilities thus represent treatment effects, not selection.

We can therefore arrive at counterfactual exit rates under the same equilibrium as played in the data by varying the values of the conditioning variables in estimates of $P(Active_{ik} = 1 | s_k, Product_{ik} = m, Lottery_{ik} = l)$. Our estimates from Table 4 yield linear approximations to these conditional probabilities which we will now use to arrive at a set of assignments of firms to clusters which maximize firm survival.

5.3.2 Optimal Counterfactual Assignment

To solve for the optimal survival-maximizing assignment of firms to clusters, we formulate the following optimization program:

$$\begin{aligned} &\max_{s\in[0,1]^{(M-1)\times K}} s'\left(I_K \otimes \delta\right) s \\ &\text{subject to} \\ &\sum_{m=1}^{M-1} s_{mk} \leq p_k \;\forall k \\ &\sum_{k} s_{mk} = S_m \;\forall m. \end{aligned} \tag{6}$$

I_K is the $K \times K$ identity matrix, p_k is the fraction of all plots located in cluster k. S_m is the total share of firms of industry m. s_{mk} is the number of plots assigned to producers of m in cluster k as a share of the total number of plots across the industrial area. δ is the matrix of effects of having an additional producer of each product type on the probability of each type of producer remaining active. We arrive at each product-pairwise effect δ_{mn} by taking the Marshallian effects from Table 4 and applying them to each product pair according to their input-output relationship. For aggregate exit, only the average pairwise spillover effect matters so that we can set $\delta_{nm} = \delta_{mn} = 1/2(\delta_{mn} + \delta_{nm})$ in the δ matrix we use in optimization.

We work with all 180 3-digit product codes with 28 or more firms assigned, which together make up 90% of all firms assigned a 3-digit code, and an "other" category. Figure A2 plots the matrix of net spillover effects across these product codes, and Figure 6 the matrix of spillover effects after symmetrizing. Problem (6) is not convex because δ is unrestricted, and in fact the matrix in Figure 6 is indefinite. To solve the problem, we perform a log barrier reformulation where for each value of the barrier parameter we solve a Karush-Kuhn-Tucker system using Newton's method. We implement this in KNITRO, using its algorithm for selecting initial values and stopping when the optimizer has honed in on a region of the assignment space where starting optimization at different values in the region yields aggregate survival rates that do not differ from the optimum by more than 1 percentage point.

Figure 7 illustrates our optimal assignment. The shares assigned to each cluster by product are indicated by the color of the four squares surrounding each intersection of gray lines, with the horizontal lines representing products and the vertical lines representing industry clusters. The numbers represented by each row add to 1. Comparing with

Figure 6, we see that products with input-output linkages implying a net positive cross-industry effect on the probability of remaining active through 2018 tend to be grouped together in clusters. For instance, product codes 561 and 571 ("printed books, newspaper, periodicals, note books, register etc and other printed matters" and "packing materials made of paper", respectively), whose net spillover implies a strong positive effect on the probability of remaining active, are often assigned to the same cluster.

Relative to uniform assignment, the optimal assignment illustrated in Figure 7 decreases aggregate exit by 15 percentage points. The conclusion that optimally assigning firms to clusters reduces exit is robust to alternative definitions of upstream and downstream. If we call a product x downstream of product y if y's revenues from selling to x are above the 75th percentile of y's revenues from selling to any other type of producer, instead of above the median, and do the same for the definition of upstream, our assignment reduces aggregate exit by 8.2 percentage points.

We also check whether the performance of optimal assignment is reliant on extrapolation beyond the range of producer shares of clusters we observe in the data. If we further constrain the optimal assignment problem so that at most 9% of the producers of a product are assigned to any one cluster (the maximum we observe in the data), aggregate exit falls by 10.8 percentage points relative to uniform assignment. Therefore, leveraging firm spillovers emerges as a robust solution to mitigate a substantial amount of the destructive effect of industrial relocation.

6 Back of the Envelope Cost-Benefit Analysis (CBA)

In this section, we construct a back-of-the-envelope (BOTE) estimate of the health impacts of firm relocation. We note that these are indicative estimates, relying on prior work to estimate how pollution reductions impact mortality.

6.1 Benefits

To estimate the benefits of the program, we consider benefits from avoided mortality as a result of reduced pollution exposure, which is a first-order benefit of pollution reduction. These are a function of three factors: (1) the impact of firm removal on air pollution, (2) the population exposed to pollution, and (3) the mapping between the change in exposure and mortality risk. We discuss each of these in turn.

6.1.1 Pollution-Firm Relationship

For the purposes of the CBA, we use a constrained version of the quadratic instrumental variables model from Equation (3). Specifically, while we do not want to impose any prior shape restriction on the quadratic function in Table 2, for a reasonable CBA we constrain the mapping from number of polluting firms to PM 2.5 to be weakly monotonically increasing.

We do this by restricting the quadratic function estimation to have a turning point at or past the largest observed number of firms having not yet taken possession of their industrial area. Specifically, we take a control function approach (see e.g. Wooldridge (2015)), assuming correct specification of the first-stage equation (Equation (4)). We then reparameterize the quadratic specification from Equation (3) as:

$$\begin{aligned}Y_{it} &= \beta_0 + \beta_1 FirmsWithoutPossession_{it} + \beta_2(\beta_1, \delta) FirmsWithoutPossession_{it}^2 \\ &\quad + \beta_3 TotalFirmsRelocated_i + \lambda\nu_{it} + \omega_t + \epsilon_{it}\end{aligned}$$

where ν_{it} is the residual from the first stage, which in the control function approach captures all sources of endogeneity, and

$$\beta_2(\beta_1, \delta) = -\frac{\beta_1}{2\max\{FirmsWithoutPossession_{it}\}} + \exp(\delta).$$

This parameterization guarantees that the turning point of the quadratic function, $-\beta_1/(2\beta_2)$, lies at or beyond $\max\{FirmsWithoutPossession_{it}\}$, the maximum observed number of firms without possession in a neighborhood. We estimate the parameters by nonlinear least squares and extrapolate, when needed, using the estimated derivative at $\max\{FirmsWithoutPossession_{it}\}$. The function is very close to that estimated in column 2 of Table 2 over the range of the data, with the monotonicity restriction serving primarily to discipline extrapolation.[10] To assign increases in the number of polluting firms to destination neighborhoods that overlap with the industrial areas, we assume uniform firm density in the industrial areas. This means firm increases are distributed proportionally to the intersection area between each neighborhood and the industrial area. Combining firm decreases in the origin neighborhoods and firm increases in destination areas, we compute net firm changes at the neighborhood level, then compute the decrease or increase in pollution using our estimated constrained quadratic polynomial function.

[10]Representing the polluting firm to PM 2.5 relationship using a second degree Bernstein polynomial yields equivalent fitted results. Bernstein polynomials have been used in the econometrics literature to impose monotonicity constraints (see e.g., the supplementary material for Mogstad et al. (2018)).

The second factor impacting policy benefits is the population exposed. We assign each neighborhood the 2011 Population Census population density of the Census ward it lies in, taking an area-weighted average when it intersects multiple wards. The average density for neighborhoods with any relocated firms is 20,225 people per sq km, which reflects the large numbers of people exposed to pollution in Delhi.

To translate pollution changes to mortality impacts, we use estimates from Pope III et al. (2020), who conduct a meta-analysis of the PM2.5-mortality relationship. Studies from Asia in their analysis show that a 10 $\mu\text{g}/\text{m}^3$ reduction in PM2.5 exposure reduces all-cause mortality risk by 5% (hazard ratio 1.05). Scaling linearly, this implies a mortality reduction of 0.5% per $\mu\text{g}/\text{m}^3$. We denote this as $\gamma = 0.005$.

We take the product of neighborhood population density, PM 2.5 change, and γ to get the annual mortality impact in each neighborhood and sum over neighborhoods to get the net mortality impact of 158,401 lives saved. This corresponds to a net mortality benefit of approximately \$111 billion per year. Valued at a Value of Statistical Life (VSL) of \$700,000 (following Majumder et al. (2018)'s estimate for India), this represents approximately \$114 billion in annual mortality benefits.[11]

The distribution of mortality impacts across destination areas is uneven. Bawana and Bawana-II, which received the largest number of allottees, bear a substantial share of the mortality burden, as does the area closest to central Delhi despite having a small number of relocated firms. However, the mortality cost in destination areas is substantially smaller than the benefit in source areas because: (1) fewer people live in the destination industrial areas compared to the dense urban neighborhoods from which firms were removed, and (2) the destination industrial areas are purpose-built zones where population density is lower.

6.2 Costs: Impact on Relocated Firms

In this section, we consider how the relocation policy affected firms that were relocated outside the city. First, for this back-of-the-envelope calculation, we estimate a counterfactual survival probability using distance that a firm was relocated. The probability that firms survived in the industrial area in the long run is strongly decreasing in the distance they were moved, highlighting that location is a consequential endogenously chosen parameter by firms.

[11] We note an important caveat: the exposure-mortality relationship may be concave at high pollution levels (Cropper and Park, 2022), in which case marginal reductions in PM2.5 at Delhi's high baseline levels would have smaller health impacts than the linear extrapolation suggests. Under this alternative view, the benefits would be lower.

Only about 26% of firms are still present and functioning in the industrial area by 2018. To estimate a BOTE counterfactual death rate for these firms, we take two approaches - the first is to estimate how the survival rate varies as a function of the distance which the firm was moved. In Figure A3, we illustrate the relationship between distance moved and firm survival. Table 5 presents the corresponding regression results. Column 2 of Table 5 includes neighborhood fixed effects, comparing firms from nearby baseline locations. Each kilometer relocated lowers the probability of firm survival by between 0.4 and 1.4 percentage points. The latter estimate includes origin neighborhood fixed effects, while the former does not. This implies that for the average firm, which was relocated 20.37 kilometers, the probability of firm survival is 28.42 percentage points lower than for a firm that was not relocated (using the second estimate). This is a substantial reduction in firm survival, and similar in magnitude to the 26% survival rate.[12]

Next, we attribute the 28.42 percentage point lower survival probability to the policy to estimate the costs to firms and get a conservative benefit-cost ratio. The greater exit risk is over 12 years (the average firm takes possession of their plot in 2006, and the exit data was collected in 2018), yielding a 2.4% annual decrease in survival probability due to the policy. We use panel data from Center of Monitoring the Indian Economy (CMIE) (Center for Monitoring the Indian Economy, 2021) to impute firm sales and salaries (this is likely an overestimate, since CMIE is more likely to include larger firms). Restricting the sample to the state of Delhi and to the year 2006 when the average firm took possession of their industrial plot, the median firm has \$2.9 million in sales, and pays \$87,700 in wages annually. This implies that the policy caused \$1.46 billion in lost sales and \$44.2 million in lost wages, for a total damage estimate of \$1.51 billion in lost sales and wages per year across the 21,000 relocated firms.

6.3 Comparison of Benefits and Costs

Compared to the benefits of reduced mortality, the benefits are approximately 73 times higher than the costs to firms. Thus, while the relocation policy caused firms to exit at a faster rate and consequently led to lost sales and wages, the benefits of improved air quality were substantially higher.

How do these cost-benefit estimates compare with other pollution reduction possi-

[12] Alternatively, we could use survival rates estimated for Indian firms more broadly - for instance, Sengupta and Singh (2019) find that the probability of firm survival in India over 20 years for registered firms is about 50%, about a reduction of 2.5% each year for survival. Therefore, seventeen years after the policy, about 42.5% of these firms should have survived. As we show below, our cost-benefit estimates do not change significantly if we use this lower counterfactual exit rate.

bilities in India? First, let us consider mitigating damages from coal-fired power plants, which claim 84,650 lives in India each year. Installing scrubbers avoids 72% of deaths in the first year (Cropper et al., 2019), saving 60,948 lives. India has 204 GW of coal capacity (Ministry of Power, 2022); therefore, on average, scrubbers in a 500 MW plant would save 150 lives per year, at a cost of $1.2 billion over 20 years, or 60 million per year. These estimates imply that the cost per life saved is around $400,000. Another pollution reduction possibility is crop residue burning in India, which causes 86,000 premature deaths in India annually (Lan et al., 2022). Jack et al. (2022) find that payment for ecosystem services to incentivize farmers to not burn residue is effective in reducing burning, and applying those treatment effects leads to a cost of life saved of $2,930. Industrial relocation is in the middle of these policies in terms of cost-effectiveness, saving approximately 158,400 lives per year at an economic cost of $1.51 billion, which yields a cost of life saved estimate of approximately $9,530. Taking firm spillovers into account with our exit-reducing assignment cuts the effect of firm relocation on exit roughly in half, bringing the cost per life saved down to approximately $4,765. Thus, leveraging firms' economic spillovers caused by input-output linkages can significantly reduce the costs of such policies on firms.[13]

7 Conclusion

Firm location decisions have important spillovers to the neighborhoods they locate in. These spillovers can be positive, generating employment and knowledge flows, or negative, such as increasing pollution exposure. We find that the presence of the polluting firms studied in this paper negatively impacts neighborhood-level ambient environmental quality in New Delhi. The removal, however, also impacts the relocated firms, substantially decreasing their survival probabilities. These survival probabilities could have been significantly increased by relocating firms taking into account firm spillovers in the relocated industrial areas, indicating that such spillovers can be a powerful force to reduce costs on relocated firms.

The firm removal seems to have been permanent. As described in Section 3.3.1, we sent surveyors to the original addresses of relocated firms to check if they were still present and identify what was present at that address. Results are summarized in Table 6, and show that *any* firm was present in less than 10% of the locations. We also asked the surveyors to collect information on whether the firm could possibly be the same as the

[13] If on the other hand, pollution reductions did not impact health meaningfully due to concavity in the exposure-response relationship, the program's benefit-cost estimate would be less favorable. However, the point about using spillovers to reduce costs on relocated firms still stands.

relocated firm, using firm name and other characteristics to identify this. This happened for less than a third of all firms, or about 3% of all observations. The largest category of land use was residential buildings (in 45% of locations), followed by mixed use residential with retail shops and retail shops alone (21% of locations). The rest were split between construction or empty lots (2.7%), warehouses (4%), or were locked during multiple visits (7.7%).[14]

Removal of firms may have important equity implications, by increasing commuting costs or moving costs for workers, as well as impacting the affordability of a neighborhood. Furthermore, other policy choices may affect the impacts on firms and the environment; for instance, which types of industries should be relocated to provide the maximum environmental benefits and minimize costs to firms and workers, and whether lump sum transfers instead of allocating them land in a fixed place is better for firms' survival. These, and related questions, remain interesting questions for future work.

References

Ahlfeldt, G. M., Sturm, D. M., Redding, S. J., and Wolf, N. (2015). The Economics of Density: Evidence from the Berlin Wall. *Econometrica*, 83(6):2127–2189.

Ahrens, A., Hansen, C. B., and Schaffer, M. E. (2018). PDSLASSO: Stata module for post-selection and post-regularization OLS or IV estimation and inference. Statistical Software Components, Boston College Department of Economics.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics*, 11(Iv):727–753.

Arzaghi, M. and Henderson, J. V. (2008). Networking off Madison Avenue. *Review of Economic Studies*, 75(4):1011–1038.

Barrows, G., Garg, T., and Jha, A. (2019). The health costs of coal-fired power plants in india.

Baum-Snow, N., Pavan, R., and Gendron-Carrier, N. (2023). Local Productivity Spillovers.

Behrer, A. P., Choudhary, R., and Sharma, D. (2023). Air pollution reduces economic activity: Evidence from india. *World Bank Policy Research Working Paper*.

[14]Surveyors were unable to locate about 10% of locations.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Benkard, C., Bodoh-Creed, A., and Lazarev, J. (2020). Simulating the dynamic effects of horizontal mergers: Us airlines. *Working Paper*.

Berry, S. T. and Compiani, G. (2021). Empirical Models of Industry Dynamics with Endogenous Market Structure. *Annual Review of Economics*, 13:309–334.

Betcherman, G., Daysal, N. M., and Pagés, C. (2010). Do employment subsidies work? evidence from regionally targeted subsidies in turkey. *Labour Economics*, 17(4):710–722.

Blackman, A., Lahiri, B., Pizer, W., Planter, M. R., and Piña, C. M. (2010). Voluntary environmental regulation in developing countries: Mexico's clean industry program. *Journal of Environmental Economics and Management*, 60(3):182–192.

Brunnermeier, S. B. and Levinson, A. (2004). Examining the evidence on environmental regulations and industry location. *The Journal of Environment & Development*, 13(1):6–41.

Bryan, G., Frye, K., and Morten, M. (2024). Spatial economics for low-and middle-income countries.

Busso, M., Gregory, J., and Kline, P. (2013). Assessing the incidence and efficiency of a prominent place based policy. *American Economic Review*, 103(2):897–947.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Center for Monitoring the Indian Economy (2021). Prowessdx database.

Chaurey, R., Abeberese, A. B., and Menon, R. (2023). Place-based policies and migration: Evidence from india. *Available at SSRN 4697765*.

Chen, Q., Chen, Z., Liu, Z., Serrato, J. C. S., and Xu, D. Y. (2021). Regulating conglomerates: Evidence from an energy conservation program in china.

Choi, J., Hyun, J., Kim, G., and Park, Z. (2023). Trade policy uncertainty, offshoring, and the environment: Evidence from us manufacturing establishments. *Offshoring, and the Environment: Evidence from US Manufacturing Establishments (February 1, 2023)*.

Ciccone, A. and Hall, R. E. (1996). Productivity and the Density of Economic Activity. *American Economic Review*, 86(1):54–70.

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, 80(6):2543–2594.

Combes, P.-P. and Gobillon, L. (2015). The Empirics of Agglomeration Economies. In *Handbook of Regional and Urban Economics*, chapter 5, pages 247–348. Elsevier.

Criscuolo, C., Martin, R., Overman, H. G., and Van Reenen, J. (2019). Some causal effects of an industrial policy. *American Economic Review*, 109(1):48–85.

Cropper, M. and Park, Y. (2022). The health benefits of air pollution control in india. *Indian Economic Review*, 57(2):245–263.

Cropper, M. L., Guttikunda, S., Jawahar, P., Lazri, Z., Malik, K., Song, X.-P., and Yao, X. (2019). Applying benefit-cost analysis to air pollution control in the indian power sector. *Journal of Benefit-Cost Analysis*, 10(S1):185–205.

De Simone, R., Naaraayanan, S. L., and Sachdeva, K. (2024). Opening the brown box: Production responses to environmental regulation. *Available at SSRN 4707581*.

Do, Q.-T., Joshi, S., and Stolper, S. (2018). Can environmental policy reduce infant mortality? evidence from the ganga pollution cases. *Journal of Development Economics*, 133:306–325.

Duflo, E., Greenstone, M., Pande, R., and Ryan, N. (2018). The value of regulatory discretion: Estimates from environmental inspections in india. *Econometrica*, 86(6):2123–2160.

Duranton, G. and Puga, D. (2020). The Economics of Urban Density. *Journal of Economic Perspectives*, 34(3):3–26.

Ellison, G., Glaeser, E. L., and Kerr, W. R. (2010). What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *American Economic Review*, 100(June):1195–1213.

Fan, H., Zivin, J. S. G., Kou, Z., Liu, X., and Wang, H. (2019). Going green in china: Firms' responses to stricter environmental regulations. Technical report, National Bureau of Economic Research.

Fenske, J., Haseeb, M., and Kala, N. (2023). How rules and compliance impact organizational outcomes: Evidence from delegation in environmental regulation. Technical report, National Bureau of Economic Research.

Foster, A. and Kumar, N. (2011). Health effects of air quality regulations in Delhi, India. *Atmospheric Environment*, 45(9):1675–1683.

Gechter, M. and Tsivanidis, N. (2020). Spatial spillovers from urban renewal: Evidence from the mumbai mills redevelopment. Technical report, Working Paper.

Glaeser, E., Kallal, H., Scheinkman, J., and Shleifer, A. (1992). Growth in Cities. *Journal of Political Economy*, pages 1126–1152.

Glaeser, E. L. and Gottlieb, J. D. (2008). The economics of place-making policies. Technical report, National Bureau of Economic Research.

Greenbaum, R. T. and Engberg, J. B. (2004). The impact of state enterprise zones on urban manufacturing establishments. *Journal of Policy Analysis and Management*, 23(2):315–339.

Greenstone, M. and Fan, C. Q. (2018). Introducing the air quality life index. *AQLI Annual Report*.

Greenstone, M. and Hanna, R. (2014). Environmental regulations, air and water pollution, and infant mortality in india. *American Economic Review*, 104(10):3038–3072.

Greenstone, M., Hornbeck, R., and Moretti, E. (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy*, 118(3):536–598.

Greenstone, M., Pande, R., Sudarshan, A., and Ryan, N. (2022). The benefits and costs of emissions trading: Experimental evidence from a new market for industrial particulate emissions. *Working paper*.

Harrison, A., Hyman, B., Martin, L., and Nataraj, S. (2015). When do firms go green? comparing price incentives with command and control regulations in india. *NBER Working Paper*.

Hasan, R., Jiang, Y., and Rafols, R. M. (2021). Place-based preferential tax policy and industrial development: Evidence from india's program on industrially backward districts. *Journal of Development Economics*, 150:102621.

He, G., Wang, S., and Zhang, B. (2020). Watering down environmental regulation in china. *The Quarterly Journal of Economics*, 135(4):2135–2185.

HEI (2021). State of global air 2024.

Henderson, J. V. (1995). Effects of air quality regulation.

Henderson, V., Kuncoro, A., and Turner, M. (1995). Industrial development in cities. *Journal of Political Economy*, 103(5):1067–1090.

Hyman, B. G., Freedman, M., Khanna, S., and Neumark, D. (2022). Firm responses to state hiring subsidies: Regression discontinuity evidence from a tax credit formula. Technical report, National Bureau of Economic Research.

Jack, B. K., Jayachandran, S., Kala, N., and Pande, R. (2022). Money (not) to burn: payments for ecosystem services to reduce crop residue burning. Technical report, National Bureau of Economic Research.

Jaffe, A. B., Peterson, S. R., Portney, P. R., and Stavins, R. N. (1995). Environmental regulation and the competitiveness of us manufacturing: what does the evidence tell us? *Journal of Economic literature*, 33(1):132–163.

Karplus, V. J., Zhang, S., and Almond, D. (2018). Quantifying coal power plant responses to tighter so2 emissions standards in china. *Proceedings of the National Academy of Sciences*, 115(27):7004–7009.

Kline, P. and Moretti, E. (2014). Local economic development, agglomeration economies, and the big push: 100 years of evidence from the Tennessee Valley Authority. *Quarterly Journal of Economics*, 129(1):275–331.

Lan, R., Eastham, S. D., Liu, T., Norford, L. K., and Barrett, S. R. (2022). Air quality impacts of crop residue burning in india and mitigation alternatives. *Nature communications*, 13(1):6537.

Leonardi, M. and Moretti, E. (2023). The Agglomeration of Urban Amenities: Evidence from Milan Restaurants. *American Economic Review: Insights*, 5(2):141–157.

Levinson, A. (1996). Environmental regulations and manufacturers' location choices: Evidence from the census of manufactures. *Journal of public Economics*, 62(1-2):5–29.

Lewis, D. J. and Mertens, K. (2025). A Robust Test for Weak Instruments for 2SLS with Multiple Endogenous Regressors. *The Review of Economic Studies, forthcoming*.

List, J. A., Millimet, D. L., Fredriksson, P. G., and McHone, W. W. (2003). Effects of environmental regulations on manufacturing plant births: evidence from a propensity score matching estimator. *Review of Economics and Statistics*, 85(4):944–952.

Lu, Y., Wang, J., and Zhu, L. (2019). Place-based policies, creation, and agglomeration economies: Evidence from china's economic zone program. *American Economic Journal: Economic Policy*, 11(3):325–360.

Majumder, A., Madheswaran, S., et al. (2018). *Value of statistical life in India: A hedonic wage approach*. Institute for Social and Economic Change.

Marshall, A. (1920). *Principles of Economics: An Introductory Volume*. Macmillan and Co., London.

McDuffie, E., Martin, R., Yin, H., and Brauer, M. (2021). Global burden of disease from major air pollution sources (gbd maps): a global approach. *Research Reports: Health Effects Institute*, 2021.

Millimet, D. L. and Roy, J. (2016). Empirical tests of the pollution haven hypothesis when environmental regulation is endogenous. *Journal of Applied Econometrics*, 31(4):652–677.

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters. *Econometrica*, 86(5):1589–1619.

Montiel Olea, J. L. and Pflueger, C. (2013). A Robust Test for Weak Instruments. *Journal of Business and Economic Statistics*, 31(3):358–369.

Muller, N. Z. and Mendelsohn, R. (2009). Efficient pollution regulation: getting the prices right. *American Economic Review*, 99(5):1714–1739.

Murphy, K. M., Shleifer, A., and Vishny, R. W. (1989). Industrialization and the big push. *Journal of Political Economy*, 97(5):1003–26.

Nakajima, K. and Teshima, K. (2017). Identifying Neighborhood Effects among Firms: Evidence from Location Lotteries of the Tokyo Tsukiji Fish Market. Technical report, Hitotsubashi University Working Paper.

Neumark, D. and Kolko, J. (2010). Do enterprise zones create jobs? evidence from california's enterprise zone program. *Journal of Urban Economics*, 68(1):1–19.

Neumark, D. and Simpson, H. (2015). Place-based policies. In *Handbook of regional and urban economics*, volume 5, pages 1197–1287. Elsevier.

O'keefe, S. (2004). Job creation in california's enterprise zones: a comparison using a propensity score matching model. *Journal of Urban Economics*, 55(1):131–150.

Pandey, A., Brauer, M., Cropper, M. L., Balakrishnan, K., Mathur, P., Dey, S., Turkgulu, B., Kumar, G. A., Khare, M., Beig, G., et al. (2021). Health and economic impact of air pollution in the states of india: the global burden of disease study 2019. *The Lancet Planetary Health*, 5(1):e25–e38.

Pope III, C. A., Coleman, N., Pond, Z. A., and Burnett, R. T. (2020). Fine particulate air pollution and human mortality: 25+ years of cohort studies. *Environmental research*, 183:108924.

Rosenstein-Rodan, P. N. (1943). Problems of industrialisation of eastern and south-eastern europe. *The economic journal*, 53(210-211):202–211.

Rosenthal, S. S. and Strange, W. (2020). How Close Is Close? The Spatial Reach of Agglomeration Economies. *Journal of Economic Perspectives*, 34(3):27–49.

Ryan, S. P. (2012). The costs of environmental regulation in a concentrated industry. *Econometrica*, 80(3):1019–1061.

Sengupta, R. and Singh, M. (2019). Firm survival in India: Status of firms formed over the last 30 years. *Ideas for India*.

Shapiro, J. S. and Walker, R. (2020). Is air pollution regulation too lenient? evidence from us offset markets. Technical report, National Bureau of Economic Research.

Sharma, S., Bawase, M., Ghosh, P., Saraf, M., Goel, A., Suresh, R., Datta, A., Jhajhjra, A., Kundu, S., Sharma, V., et al. (2018). Source apportionment of pm2. 5 and pm10 of delhi ncr for identification of major sources. *The Energy Resources Institute, Delhi and Automative Research Association of India*.

Singh, V. (2007). *Delhi Master Plan 2021*. Rupa Publications.

Song, R., Xia, F., and Zhang, B. (2022). Pollution taxes as a second-best: Accounting for multidimensional firm heterogeneity in environmental regulations.

Stock, J. H. and Yogo, M. (2005). Testing for weak instruments in linear iv regression. In Andrews, D. W. K. and Stock, J. H., editors, *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pages 80–108. Cambridge University Press.

Tanaka, S., Teshima, K., and Verhoogen, E. (2022). North-south displacement effects of environmental regulation: The case of battery recycling. *American Economic Review: Insights*, 4(3):271–288.

Tsivanidis, N. (2023). Evaluating the Impact of Urban Transit Infrastructure: Evidence from Bogotá's TransMilenio. *Working Paper*.

Van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C., Kalashnikova, O. V., Kahn, R. A., Lee, C., et al. (2021). Monthly global estimates of fine particulate matter and their uncertainty. *Environmental Science & Technology*, 55(22):15287–15300.

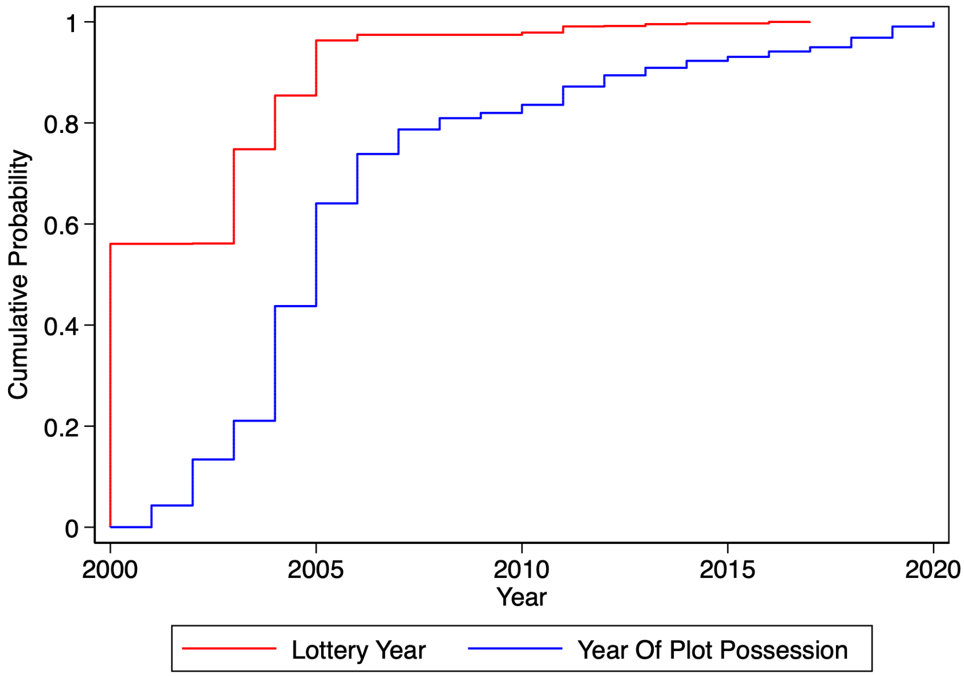
van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M., and Winker, D. M. (2016). Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. *Environmental Science and Technology*, 50:3762–3772.

Vitali, A. (2022). Consumer search and firm location: Theory and evidence from the garment sector in uganda. *Work. Pap*.

Wilson, S., Hutson, M., and Mujahid, M. (2008). How planning and zoning contribute to inequitable development, neighborhood health, and environmental injustice. *Environmental Justice*, 1(4):211–216.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445.

Figure 1: Timing of Plot Lottery and Plot Possession in the Industrial Area



Notes: The blue line depicts the cumulative distribution function of the year in which the firm took possession of the plot in the industrial area. The red line depicts the cumulative distribution function of the year in which the firm was assigned a plot in the industrial area by lottery.

Figure 2: Firm Origin Addresses, Cumulative by Year of Plot Assignment

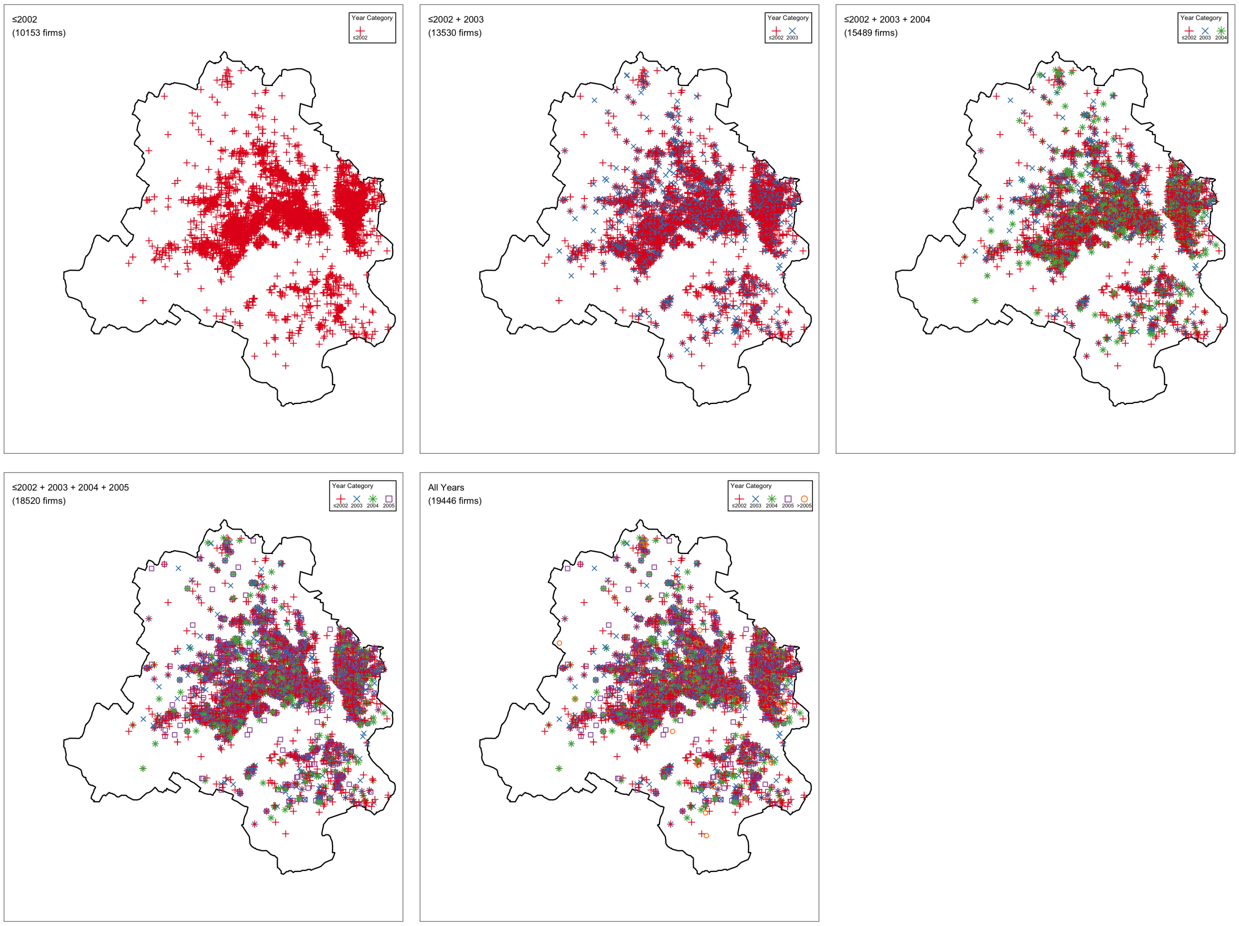
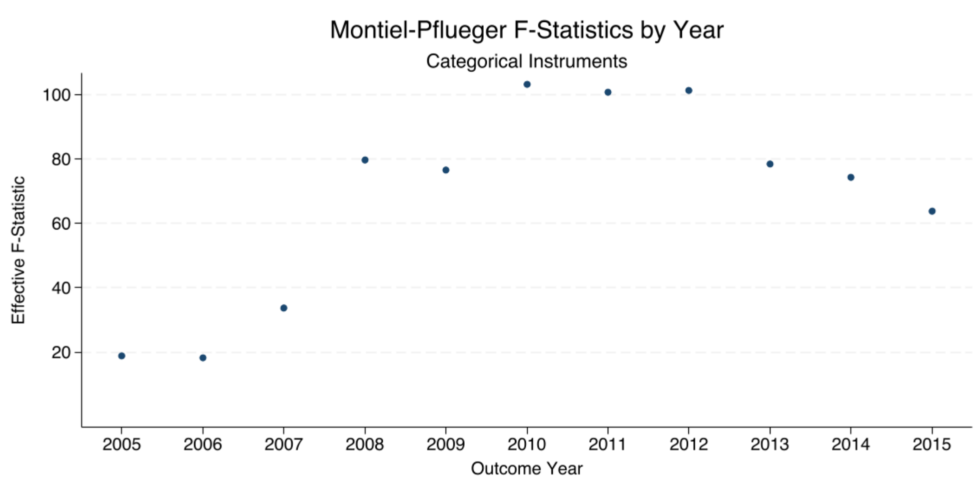


Figure 3: IV Estimates of the Effect of a Remaining Polluting Firm on Mean Neighborhood PM2.5 by Year

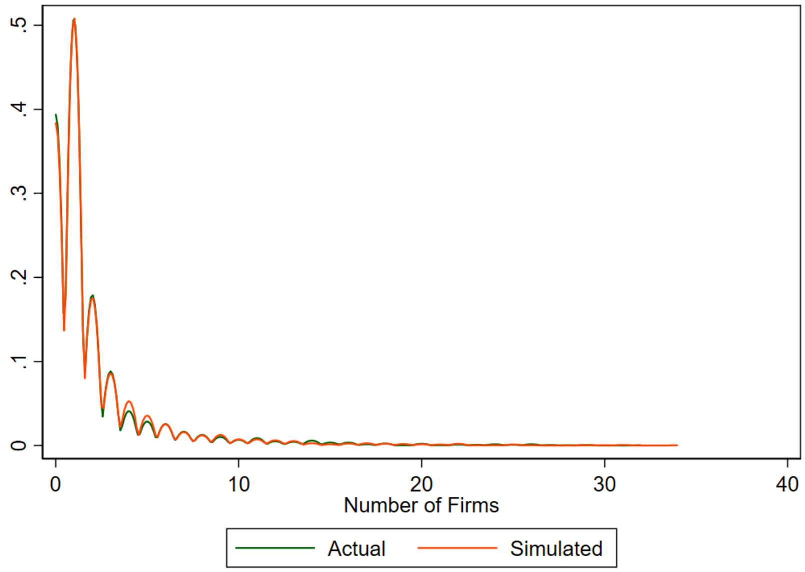
Notes: This figure shows the instrumental variable coefficient estimates (with 90% confidence intervals) for the effect of firms without possession on PM2.5 levels. Each point represents a separate instrumental variables regression, instrumenting the number of firms remaining in a neighborhood in a given year with a vector whose elements are the number of firms with year of plot assignment 2000, 2001, 2002, 2003, 2004, and 2005 (with firms assigned plots after 2005 as the omitted category).

Figure 4: First Stage F-Statistics: Distribution of Firms by Year of Plot Assignment on Number of Remaining Polluting Firms by Year

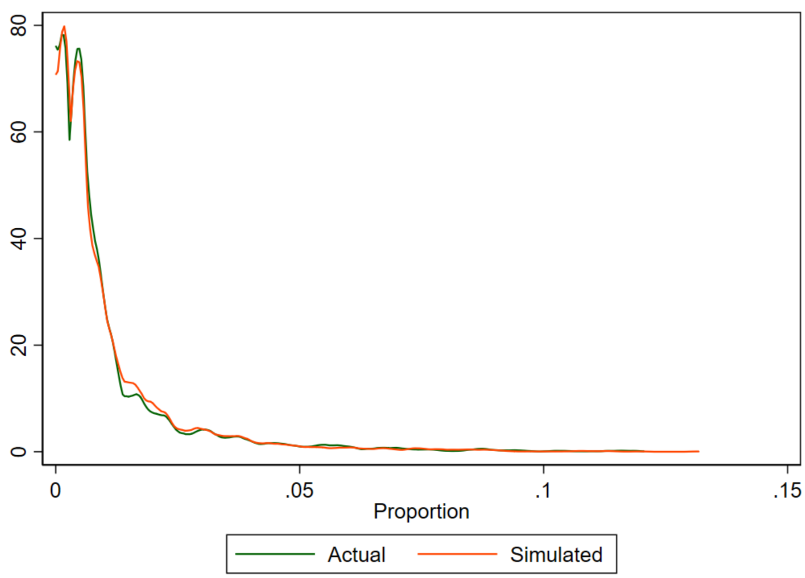


Notes: This figure shows the first stage Montiel Olea and Pflueger (2013) F-statistics testing the joint significance of the vector of number of firms in a neighborhood assigned plots by 2002, in 2003, in 2004, and in 2005 in predicting the number of firms without possession in each year.

Figure 5: Actual vs. Simulated Plot Assignments in the Industrial Area



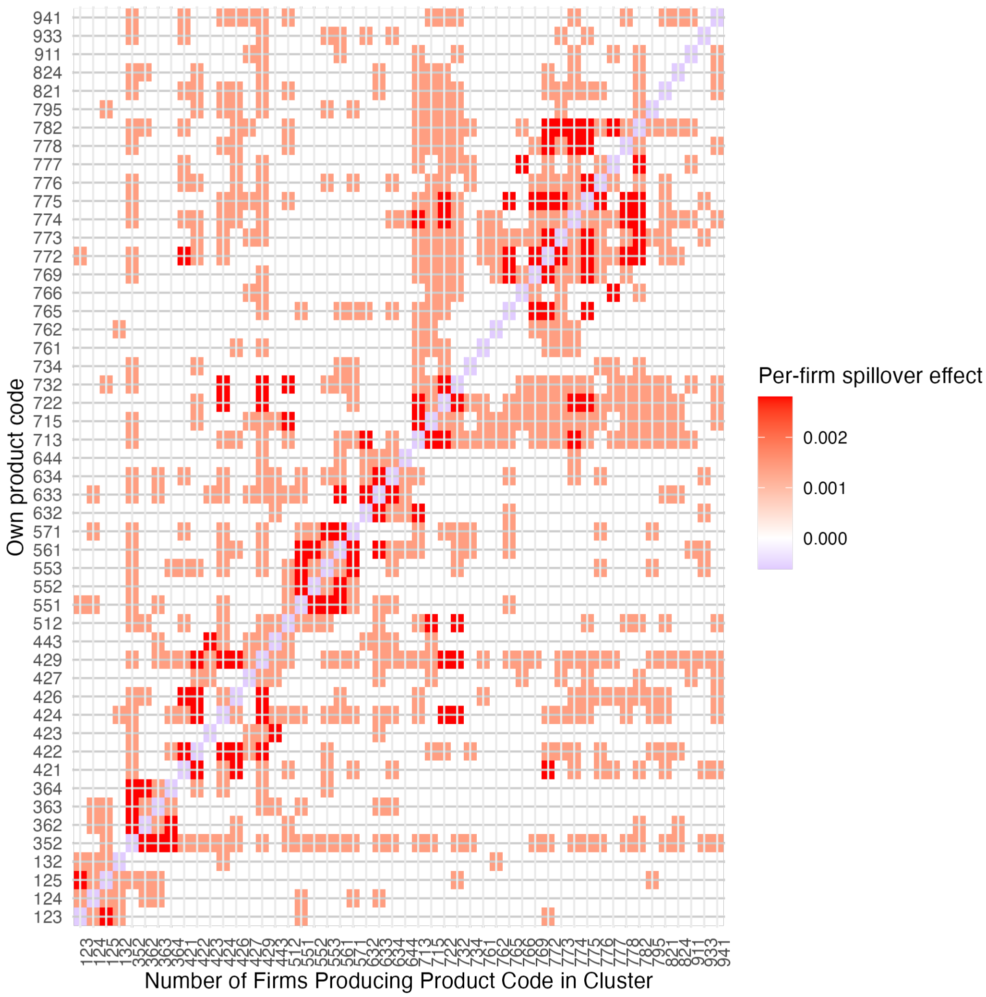
(a) Number of Firms Producing Each Product in a Cluster



(b) Proportion of Firms Producing Each Product in a Cluster

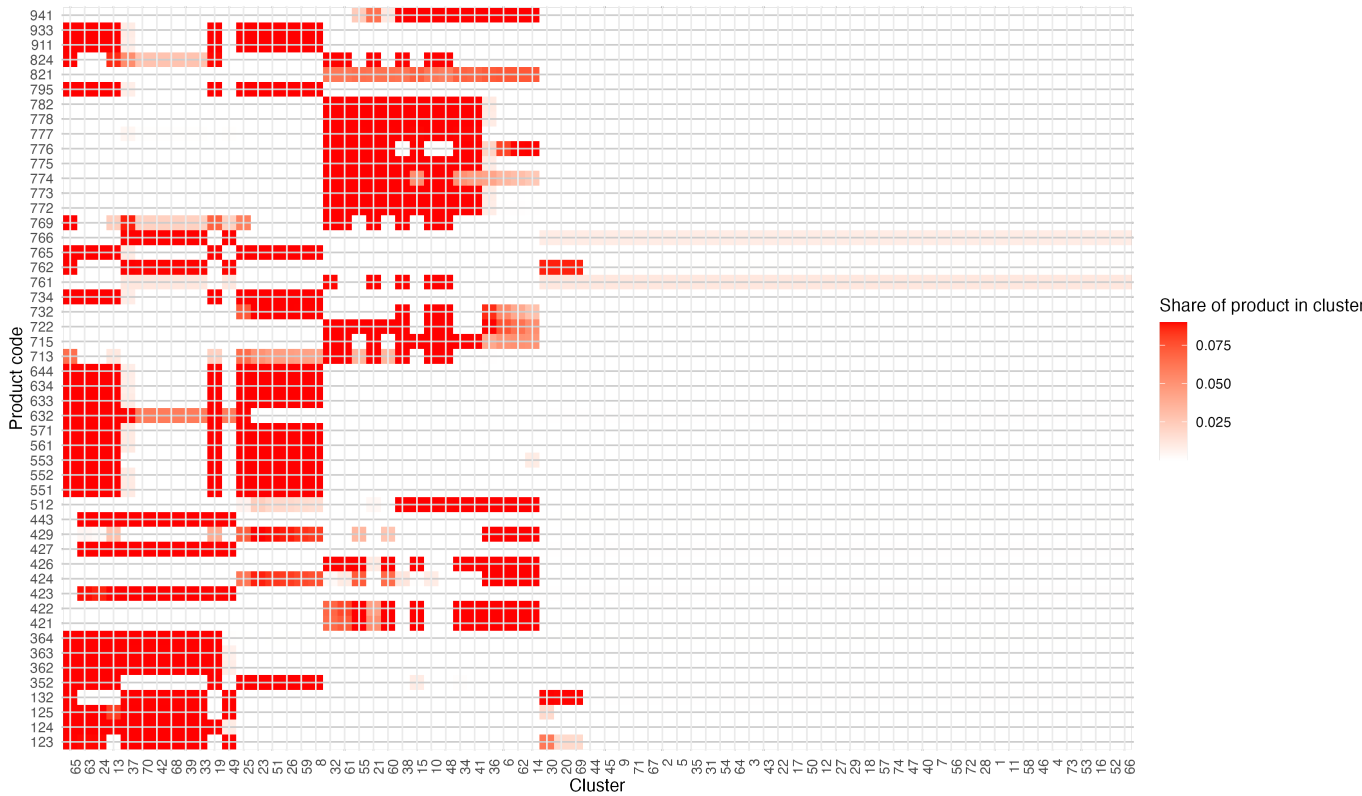
Notes: The orange line in Panel (a) depicts the kernel density of the cluster-product distribution (number of producers of each product in each cluster) generated by the simulated lotteries, while the green line depicts this distribution for actual plot assignments. The orange line in Panel (b) depicts the kernel density of the cluster-product distribution (proportion of each product in each cluster) generated by the simulated lotteries, while the green line depicts this distribution for actual plot assignments. To simulate the original lotteries, for each year of the lottery and plot size category, we randomly pick firms and randomly assign them to available plots of that size category (for instance, if x firms were assigned 100 m^2 plots in 2000, we randomly pick x of all the 100 m^2 firms and randomly assign them to the 100 m^2 plots that were assigned in 2000).

Figure 6: Effect of the Number of Producers of Each Good in the Cluster on Firm Survival, Symmetric



Notes: the color of the four squares around each intersection of gray lines represents the effect of the share in an industrial cluster of the product code corresponding to the vertical gray line on the survival of a firm producing the good corresponding to the horizontal line. Effects δ_{mn}, where m denotes the own product represented by the horizontal gray line and n the number of producers of the product represented by the vertical gray line in the cluster from Figure A2 after making $\delta_{nm} = \delta_{mn} = 1/2(\delta_{mn} + \delta_{nm})$ so that the matrix is symmetric. Spillover effects are on survival, which is a dummy variable that takes the value 1 if the firm was operating in the largest industrial area, Bawana, in 2018, and 0 otherwise.

Figure 7: Optimal Product Composition of Clusters in the Industrial Area



Notes: the color of the four squares around each intersection of gray lines represent the share of the product represented by the horizontal gray line which is optimally assigned to the industrial cluster represented by the vertical gray line. The shares add up to 1 across rows. There are 74 industrial clusters in the largest industrial area.

Table 1: Balance by Year of Lottery: Longitude and Latitude

	(1) Latitude	(2) Longitude
Lottery Year	0.0000932	-0.000906
	(0.000326)	(0.00148)
Constant	28.48****	79.00****
	(0.652)	(2.960)
Mean of Dependent Variable	28.66	77.19
R-Squared	0.0000	0.0000
N	19,508	19,508

Notes: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Latitude and longitude are the latitude and longitude of the firm's origin address as provided by Google Maps. Lottery Year is the year in which a firm received their plot assignment.

Table 2: Pooled Instrumental Variables Estimates: Effect of Firm Presence on Air Pollution

	Fine PM (μg/m^3)		
	Linear	Quadratic	
	(1)	(2)	
2SLS			
# of Firms Without Possession	0.0948*	0.0736****	
	(0.0573)	(0.0144)	
# of Firms Without Possession^2		-0.0003****	
		(0.0001)	
Total # of Firms Relocated	-0.0153	-0.0027	
	(0.0135)	(0.0034)	
Constant	101.1630****	101.2559****	
	(0.2263)	(0.1730)	
Marginal Effect at Mean (32.66 firms)		0.0548****	
		(0.0136)	
	# of Firms	# of Firms Without Poss.	
	Without Poss.	Linear	Squared
First Stage			
# of Firms with Lottery Year $\leq$ 2002	-1.2860****	-1.1048****	-138.6016****
	(0.1197)	(0.1008)	(28.4922)
# of Firms with Lottery Year $\leq$ 2002, Squared		0.0001	0.0401
		(0.0003)	(0.1230)
# of Firms with Lottery Year $\leq$ 2002 $\times$ Trend		-0.0158***	-1.4996
		(0.0060)	(1.8305)
# of Firms with Lottery Year $\leq$ 2002, Squared $\times$ Trend		-0.0000	-0.0057
		(0.0000)	(0.0157)
# of Firms with Lottery Year 2003	-1.1766****	-0.7695****	-76.4027
	(0.1261)	(0.1647)	(52.2515)
# of Firms with Lottery Year 2003, Squared		-0.0048	-1.2969
		(0.0057)	(1.8878)
# of Firms with Lottery Year 2003 $\times$ Trend		-0.0482**	-8.7588
		(0.0203)	(6.4943)
# of Firms with Lottery Year 2003, Squared $\times$ Trend		0.0006	0.1690
		(0.0008)	(0.2476)
# of Firms with Lottery Year 2004	-0.9735****	-1.1777****	-466.8534****
	(0.1060)	(0.1839)	(80.9374)
# of Firms with Lottery Year 2004, Squared		0.0250***	18.9934****

(continued on next page)

(continued from previous page)

	Fine PM (μg/m^3)		
	Linear	Quadratic	
		(0.0084)	(3.9360)
# of Firms with Lottery Year 2004 × Trend		0.0183	42.7673****
		(0.0196)	(9.1523)
# of Firms with Lottery Year 2004, Squared × Trend		-0.0037****	-2.4607****
		(0.0010)	(0.5100)
# of Firms with Lottery Year 2005	-0.3450***	0.0096	-44.4800
	(0.1288)	(0.1346)	(34.3298)
# of Firms with Lottery Year 2005, Squared		-0.0046	-1.3044
		(0.0031)	(1.0811)
# of Firms with Lottery Year 2005 × Trend		-0.0582****	-12.2785***
		(0.0148)	(4.0662)
# of Firms with Lottery Year 2005, Squared × Trend		0.0010**	0.3173**
		(0.0004)	(0.1424)
Total # of Firms Relocated	1.2590****	1.1754****	138.8614****
	(0.1100)	(0.0770)	(22.8728)
Constant	2.5105****	0.8446****	150.8515***
	(0.2682)	(0.1084)	(48.5740)
N	15235	15235	
R-Squared	0.7730	0.7748	
Mean of Dependent Variable	112.6	112.6	
First Stage Test Statistic			
(Montiel-Pflueger F / Lewis-Mertens g_{min})	56.00	93.93	

Notes: Standard errors clustered at the neighborhood (1km × 1km grid cell) level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Fine PM is the annual concentration of fine particulate matter in μg/m^3 at the neighborhood-year level. Firms Without Possession is the number of firms in a neighborhood that have not yet taken possession of their plots. Column 1 (Linear): instruments the single endogenous variable (Firms Without Possession) using 4 base lottery year category instruments. Column 2 (Quadratic): instruments both linear and quadratic terms (Firms Without Poss. and Firms Without Poss. Squared) using 16 instruments (base draw year categories, their squares, and interactions with a time trend). Base instruments are number of firms with plot assignment years ≤2002, 2003, 2004, 2005 (firms with plot assignment years > 2005 are omitted). Total Firms Relocated is the total number of firms ever relocated from the neighborhood. First-stage regressions show effect of instruments on endogenous variables. All specifications include year fixed effects. Sample restricted to years 2005-2015. Column 1 reports the Montiel-Pflueger (2013) effective F-statistic with $\tau = 10\%$, tested at the 5% significance level (critical value: 17.90). Column 2 reports the Lewis-Mertens (2025) gmin statistic for multiple endogenous variables with $\tau = 10\%$, tested at the 5% significance level (critical value: 39.72).

Table 3: Most Common Products in the Largest Industrial Area

Product Name	Number of Firms
Motor vehicles (Passengers/goods transportation & special purpose vehicles)	1239
Printed books, newspaper, periodicals, notebooks, register etc. & other printed matters	770
Packing materials made of paper	614
Lamp, filament, electrodes, anodes/connectors, fittings & parts	487
Bags/boxes/panels/containers of plastic/pvc	401
tubes/pipes/basin & sanitary fittings of plastic/pvc	321
Film (non-sensitive/photographic)/foil/rolls/tape/rope of plastic/pvc & related materials	314
Articles, parts of plastic/pvc n.e.c	294
Copper and copper alloy, worked	266
Finished products of iron/steel	257
Wooden (incl plywood) furniture, boxes (incl packing box) and other wooden articles	248
Electrical motors, generators, transformer, power pack (this incl pump set fitted with electric motor)	226
Footwear plastic/pvc	210
Audio/video/sound apparatus & parts	182
Aluminium and aluminium alloys, worked	180

Notes: The product names are three digit ASICC codes (from 2010 product codes for the ASI). The second column shows the number of firms that produce products with the closest match with that ASICC code according to GPT (please see Appendix E for details on the matching procedure).

Table 4: Reduced Form Impacts of Firm Spillovers

	Firm Survival			
	(1)	(2)	(3)	(4)
No. Firms Upstream or Downstream(Std)	0.0567****			
	(0.0150)			
No. Firms Own Type (Std)	-0.0217**	-0.0244**	-0.0196*	-0.0217*
	(0.0109)	(0.0109)	(0.0117)	(0.0117)
No Firms Upstream (Std)		0.0517****		
		(0.0128)		
No Firms Downstream (Std)		0.0244*		
		(0.0146)		
No. Firms Upstream or Downstream (Std)			0.0384**	
			(0.0152)	
No Firms Upstream (Std)				0.0334***
				(0.0128)
No Firms Downstream (Std)				0.0159
				(0.0145)
P-Val: Any=Own	0.000***		0.014**	
P-Val: Upstream=Own		0.000***		0.008***
P-Val: Downstream=Own		0.017**		0.079*
P-Val: Upstream=Downstream		0.153		0.368
Mean of Dependent Variable	.26	.26	.26	.26
Linkage Measurement	Median	Median	75th Pctile	75th Pctile
N	9111	9111	9111	9111

Notes: Robust standard errors clustered at the cluster-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Firm Survival is a dummy variable that takes the value 1 if the firm was operating in the industrial area in 2018, and 0 otherwise. All displayed independent variables are standardized by dividing by their respective standard deviation. Post double LASSO (Belloni et al., 2013) estimates, including product fixed effects, number of plots with missing plot assignments and total number of plots in the cluster included as controls.

Table 5: Impacts of Distance Relocated on Firm Survival

	(1)	(2)
	$\mathbb{1}$(Firm Survival)	
Distance Relocated	-0.00402***	-0.0141***
	(0.000643)	(0.00479)
Constant	0.349***	0.554***
	(0.0137)	(0.0974)
Mean of the Dependent Variable	0.267	0.268
N	13580	13431
R2	0.00287	0.0365

Notes: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Firm survival is a dummy variable that takes the value 1 if the firm is operating in the industrial area (Bawana) in 2018, and 0 otherwise. Distance relocated is the distance in kilometers between the firm's original address and the plot it is assigned in the industrial area.

Table 6: Descriptive Evidence From Surveyor Visits in 2021/2022

What is Located at Original Address	Proportion
Residential	0.452
Commercial and Residential/Market	0.21
Firm	0.099
Locked building	0.077
Warehouse	0.04
Empty Plot/Under Construction	0.027

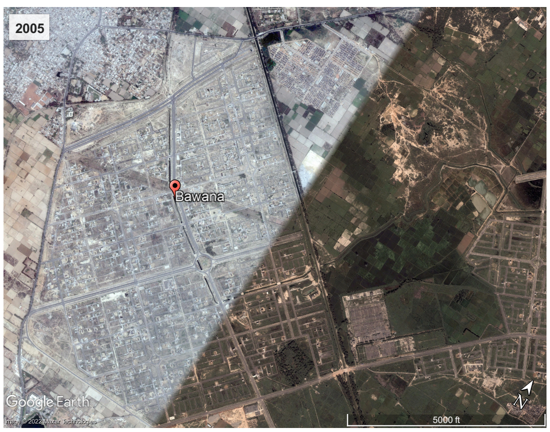
Notes: 15,756 addresses were visited in total by surveyors. Surveyors were asked to reach the origin address, and record what they found there e.g. a firm, residence etc. If there was a firm, they administered a short survey to elicit the firm's name, size, and products sold. 10% of addresses were not found by surveyors. See Section 3.3.1 for details of the data collection.

Appendices

A Appendix Figures and Tables

Figure A1: Bawana Industrial Area: 2001, 2005, and 2010

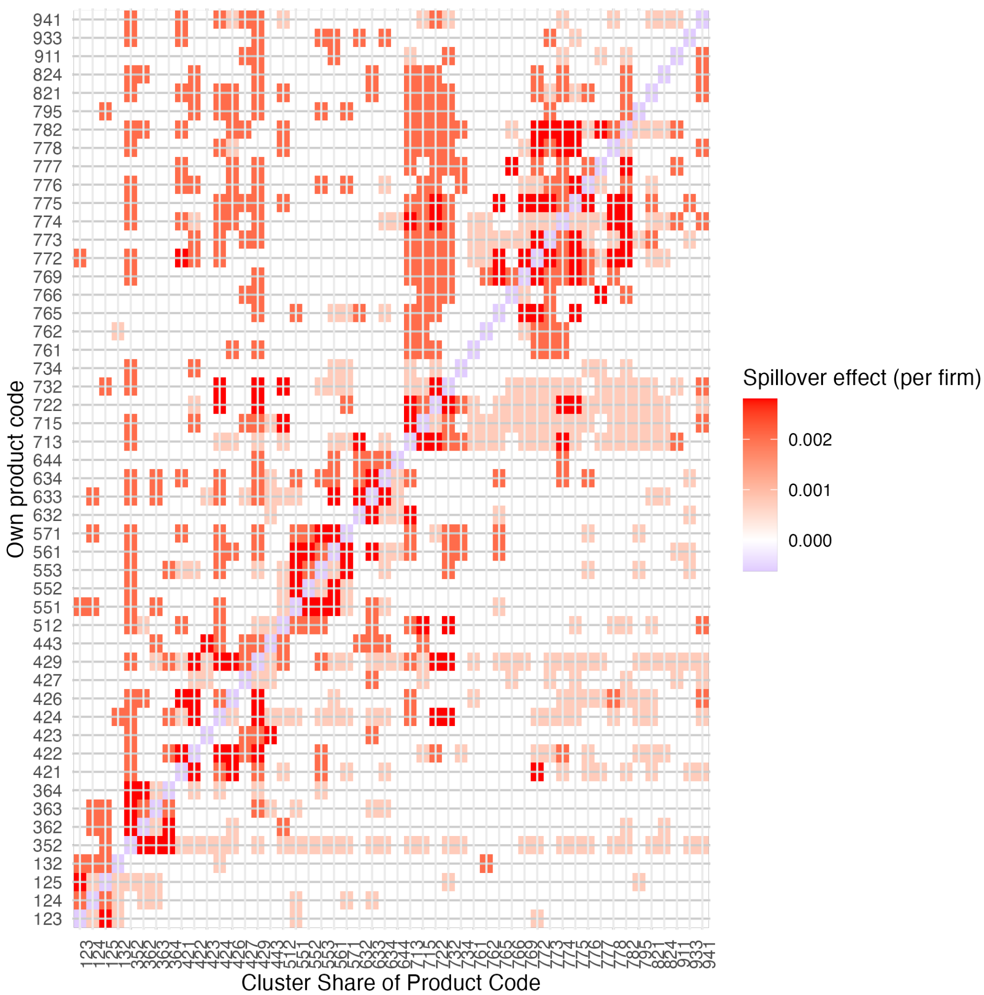
(a) 2001



(b) 2005

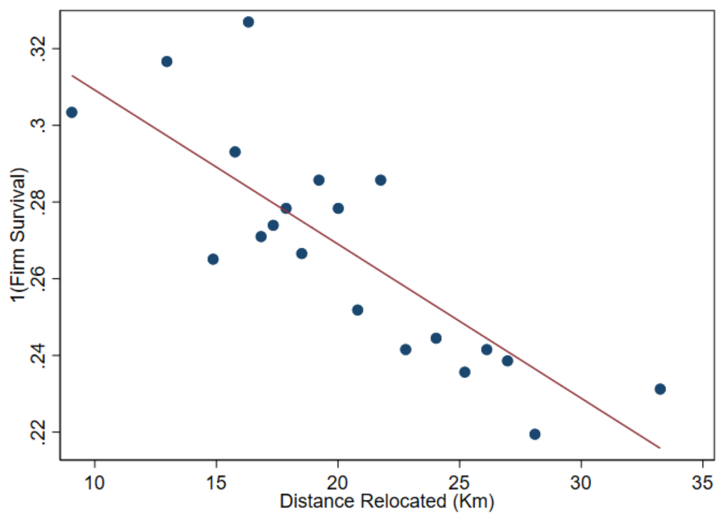
(c) 2010

Figure A2: Effect of the Number of Producers of Each Good in the Cluster on Firm Survival, by Own Good



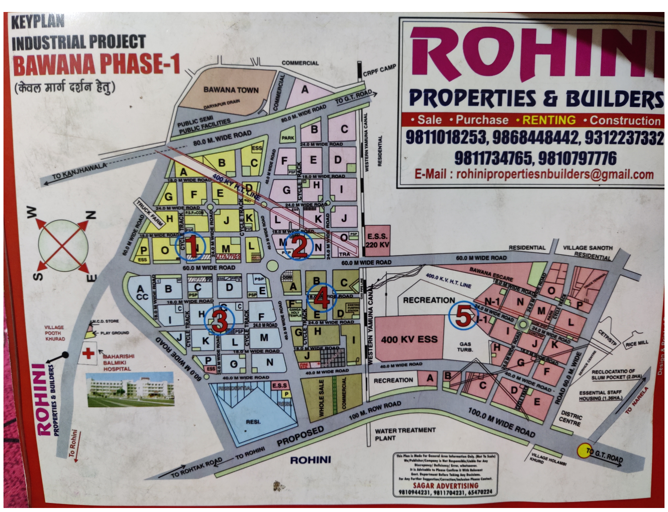
Notes: the color of the four squares around each intersection of gray lines represents the effect of the number of firms in an industrial cluster with the product code corresponding to the vertical gray line on the survival of a firm producing the good corresponding to the horizontal line. Survival is a dummy variable that takes the value 1 if the firm was operating in the largest industrial area, Bawana, in 2018, and 0 otherwise. Effects combine estimates from Table 4 with input-output relationships between individual product codes as described in the text.

Figure A3: Binscatter Plot: Distance Relocated and Firm Exit

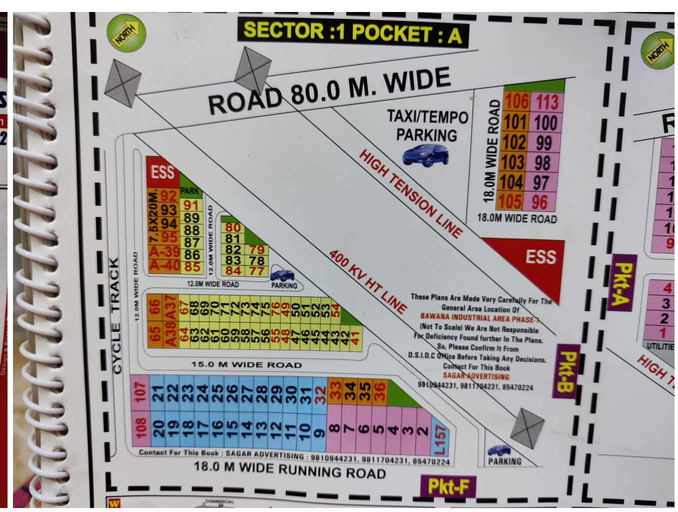


Notes: Distance Relocated is the distance between the firm's original address geocoded using Google Maps and the assigned plot in the industrial area. 1(Firm Survival) is a dummy variable that takes the value 1 if the firm was operating in the industrial area in 2018, and 0 otherwise.

Figure A4: Map of the Largest Industrial Area Cluster (Bawana)



(a) Map of Largest Industrial Area: Bawana



(b) Map of an Industrial Area Cluster

Table A1: Balance by Whether Lottery Year is Missing: Longitude and Latitude

	(1) Latitude	(2) Longitude
1(Lottery Year Missing)	-0.000228	-0.00567
	(0.00546)	(0.0150)
Constant	28.66****	77.19****
	(0.000926)	(0.00405)
Mean of Dependent Variable	28.66	77.18
R-Squared	0.0000	0.0000
N	24373	24373

Notes: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Latitude and longitude are the latitude and longitude of the firm's origin address as provided by Google Maps. 1(Lottery Year Missing) is a binary variable that takes the value 1 if the year of the firm's assignment by lottery is missing, and 0 otherwise.

Table A2: Instrumental Variables Estimates Robustness: Alternative Instrument Specifications

	Fine PM (μg/m^3)		
	IV LASSO	Firms Relocated by 2004	Quadratic (2km Clusters)
	(1)	(2)	(3)
2SLS			
# of Firms Without Possession	0.1157**	0.0965*	0.0736****
	(0.0551)	(0.0576)	(0.0151)
# of Firms Without Possession^2			-0.0003****
			(0.0001)
Total # of Firms Relocated	-0.0201	-0.0157	-0.0027
	(0.0128)	(0.0136)	(0.0036)
Constant	101.1085****	101.1586****	101.2559****
	(0.2224)	(0.2274)	(0.3293)
		# of Firms Without Poss.	# of Firms Without Poss.

(continued on next page)

(continued from previous page)

	Fine PM (μg/m^3)			
	IV LASSO	Firms Relocated by 2004	Quadratic (2km Clusters)	
			Linear	Squared
First Stage				
# of Firms Relocated by 2004		-0.9549****		
		(0.0723)		
# of Firms with Lottery Year $\leq$ 2002			-1.1048****	-138.6016****
			(0.0888)	(27.2552)
# of Firms with Lottery Year $\leq$ 2002, Squared			0.0001	0.0401
			(0.0004)	(0.1299)
# of Firms with Lottery Year 2003			-0.7695****	-76.4027
			(0.1401)	(50.7950)
# of Firms with Lottery Year 2003, Squared			-0.0048	-1.2969
			(0.0056)	(1.8700)
# of Firms with Lottery Year 2004			-1.1777****	-466.8534****
			(0.2012)	(94.1166)
# of Firms with Lottery Year 2004, Squared			0.0250***	18.9934****
			(0.0078)	(3.6425)
# of Firms with Lottery Year 2005			0.0096	-44.4800
			(0.1186)	(34.1749)
# of Firms with Lottery Year 2005, Squared			-0.0046	-1.3044
			(0.0031)	(1.1709)
# of Firms with Lottery Year $\leq$ 2002 $\times$ Trend			-0.0158**	-1.4996
			(0.0065)	(2.0274)
# of Firms with Lottery Year $\leq$ 2002, Squared $\times$ Trend			-0.0000	-0.0057
			(0.0000)	(0.0167)
# of Firms with Lottery Year 2003 $\times$ Trend			-0.0482**	-8.7588
			(0.0197)	(6.3072)
# of Firms with Lottery Year 2003, Squared $\times$ Trend			0.0006	0.1690
			(0.0008)	(0.2464)
# of Firms with Lottery Year 2004 $\times$ Trend			0.0183	42.7673****
			(0.0234)	(10.7897)
# of Firms with Lottery Year 2004, Squared $\times$ Trend			-0.0037****	-2.4607****
			(0.0009)	(0.4728)
# of Firms with Lottery Year 2005 $\times$ Trend			-0.0582****	-12.2785***
			(0.0138)	(4.4364)

(continued on next page)

(continued from previous page)

	Fine PM (μg/m^3)			
	IV LASSO	Firms Relocated by 2004	Quadratic (2km Clusters)	
# of Firms with Lottery Year 2005, Squared × Trend			0.0010**	0.3173**
			(0.0004)	(0.1577)
Total # of Firms Relocated		0.9894****	1.1754****	138.8614****
		(0.0555)	(0.0628)	(22.8133)
Constant		2.5465****	0.8446****	150.8515***
		(0.2725)	(0.1519)	(53.6170)
N	15235	15235	15235	
R-Squared		0.7730	0.7748	
Mean of Dependent Variable	112.6	112.6	112.6	
First Stage Test Statistic	N/A	174.62	95.26	

Notes: Standard errors clustered at the neighborhood level in parentheses. Columns 1-2 use 1km × 1km grid cells; Column 3 uses 2km × 2km grid cells. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Fine PM is the annual concentration of fine particulate matter in μg/m^3 at the neighborhood-year level. Firms Without Possession is the number of firms in a neighborhood that have not yet taken possession of their plots. Column 1 uses IV LASSO (Ahrens et al., 2018; Belloni et al., 2012) with cross-validation to select among all available lotteries as instruments. Candidate instruments for IV LASSO included lottery draws on the following dates, with number of winners in parentheses: Sep 21, 2000 (172); Sep 22, 2000 (7); Oct 1, 2000 (94); Oct 3, 2000 (2851); Oct 8, 2000 (558); Oct 12, 2000 (5928); Oct 18, 2000 (109); Oct 23, 2000 (2); Aug 12, 2002 (24); Jun 6, 2003 (3363); Aug 26, 2004 (1955); Oct 20, 2005 (3024); Nov 27, 2006 (286); Jan 9, 2010 (20); Jun 18, 2010 (68); Aug 4, 2010 (5); Mar 3, 2011 (8); Mar 18, 2011 (249); Aug 19, 2011 (4); Aug 26, 2011 (49); Mar 2, 2012 (15); Oct 26, 2012 (7); Mar 28, 2013 (98); Nov 12, 2013 (1); Jan 28, 2014 (38); May 1, 2014 (1); Nov 18, 2014 (1); Mar 15, 2016 (1); Mar 22, 2016 (69); Apr 19, 2016 (1); Aug 8, 2017 (2); Nov 20, 2017 (1); Mar 20, 2020 (1). The procedure selected the following instruments: Oct 12, 2000; Jun 6, 2003; Aug 26, 2004; Oct 20, 2005; Nov 27, 2006; Jun 18, 2010; Mar 18, 2011; Aug 26, 2011; Mar 28, 2013; Jan 28, 2014; Nov 18, 2014; Mar 22, 2016. Column 2 instruments Firms Without Possession using the number of firms relocated by 2004 (i.e., firms with lottery dates before 2005). Column 3 uses a quadratic specification, instrumenting both Firms Without Possession and its square using 16 instruments (4 base draw year categories, their squares, and interactions with time trend). Total Firms Relocated is the total number of firms relocated from the neighborhood. First-stage regressions for Columns 2 and 3 show effects of instruments on endogenous variables. All specifications include year fixed effects. Sample restricted to years 2005-2015. Column 1 does not report an F-statistic because the theory for weak instrument testing with IV LASSO has not been developed for specifications with time-invariant instruments and cluster-robust inference. Column 2 reports the Montiel-Pflueger (2013) effective F-statistic with $\tau = 10\%$, tested at the 5% significance level (critical value: 23.11). Column 3 reports the Lewis-Mertens (2025) gmin statistic for multiple endogenous variables with $\tau = 10\%$, tested at the 5% significance level (critical value: 46.68).

Table A3: Reduced Form Impacts of Firm Spillovers Adding Lottery Fixed Effects in LASSO Controls

	Firm Survival			
	(1)	(2)	(3)	(4)
No. Firms Upstream or Downstream(Std)	0.0568****			
	(0.0150)			
No. Firms Own Type (Std)	-0.0225**	-0.0242**	-0.0198*	-0.0222*
	(0.0110)	(0.0110)	(0.0117)	(0.0118)
No Firms Upstream (Std)		0.0525****		
		(0.0129)		
No Firms Downstream (Std)		0.0240		
		(0.0146)		
No. Firms Upstream or Downstream (Std)			0.0399***	
			(0.0152)	
No Firms Upstream (Std)				0.0349***
				(0.0128)
No Firms Downstream (Std)				0.0170
				(0.0146)
P-Val: Any=Own	0.000***		0.012**	
P-Val: Upstream=Own		0.000***		0.006***
P-Val: Downstream=Own		0.018**		0.068*
P-Val: Upstream=Downstream		0.138		0.358
Mean of Dependent Variable	.26	.26	.26	.26
Linkage Measurement	Median	Median	75th Pctile	75th Pctile
N	8964	8964	8964	8964

Notes: Robust standard errors clustered at the block-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Firm Survival is a dummy variable that takes the value 1 if the firm was operating in the industrial area in 2018, and 0 otherwise. All displayed independent variables are standardized by dividing by their respective standard deviation. Post double LASSO (Belloni et al., 2013) estimates, including product fixed effects, a dummy variable for each lottery (lottery-year and plot-size combination), number of plots with missing plot assignments and total number of plots in the block included as controls.

Table A4: Reduced Form Impacts of Firm Spillovers: Unstandardized

	Firm Survival			
	(1)	(2)	(3)	(4)
No. Firms Upstream or Downstream(Std)	0.00132****			
	(0.000348)			
No. Firms Own Type	-0.00304**	-0.00342**	-0.00274*	-0.00303*
	(0.00153)	(0.00153)	(0.00163)	(0.00164)
No Firms Upstream		0.00205****		
		(0.000508)		
No Firms Downstream		0.000755*		
		(0.000451)		
No. Firms Upstream or Downstream			0.00105**	
			(0.000416)	
No Firms Upstream				0.00162***
				(0.000619)
No Firms Downstream				0.000612
				(0.000559)
P-Val: Any=Own	0.011**		0.045**	
P-Val: Upstream=Own		0.002***		0.020**
P-Val: Downstream=Own		0.014**		0.054*
P-Val: Upstream=Downstream		0.053*		0.230
Mean of Dependent Variable	.26	.26	.26	.26
Linkage Measurement	Median	Median	75th Pctile	75th Pctile
N	9111	9111	9111	9111

Notes: Robust standard errors clustered at the block-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Firm Survival is a dummy variable that takes the value 1 if the firm was operating in the industrial area in 2018, and 0 otherwise. Post double LASSO (Belloni et al., 2013) estimates, including product fixed effects, number of plots with missing plot assignments and total number of plots in the block included as controls.

Table A5: Reduced Form Impacts of Firm Spillovers Adding Lottery Fixed Effects in LASSO: Unstandardized

	Firm Survival			
	(1)	(2)	(3)	(4)
No. Firms Upstream or Downstream(Std)	0.00132****			
	(0.000350)			
No. Firms Own Type	-0.00316**	-0.00339**	-0.00277*	-0.00310*
	(0.00154)	(0.00154)	(0.00164)	(0.00165)
No Firms Upstream		0.00208****		
		(0.000511)		
No Firms Downstream		0.000743		
		(0.000452)		
No. Firms Upstream or Downstream			0.00110***	
			(0.000417)	
No Firms Upstream				0.00169***
				(0.000619)
No Firms Downstream				0.000652
				(0.000561)
P-Val: Any=Own	0.009***		0.042**	
P-Val: Upstream=Own		0.002***		0.017**
P-Val: Downstream=Own		0.016**		0.049**
P-Val: Upstream=Downstream		0.047**		0.216
Mean of Dependent Variable	.26	.26	.26	.26
Linkage Measurement	Median	Median	75th Pctile	75th Pctile
N	8964	8964	8964	8964

Notes: Robust standard errors clustered at the block-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Firm Survival is a dummy variable that takes the value 1 if the firm was operating in the industrial area in 2018, and 0 otherwise. Post double LASSO (Belloni et al., 2013) estimates, including product fixed effects, a dummy variable for each lottery (lottery-year and plot-size combination), number of plots with missing plot assignments and total number of plots in the block included as controls.

Table A6: Distance Gradient of Reduced Form Impacts of Firm Spillovers: Unstandardized

	Firm Survival				
	(1)	(2)	(3)	(4)	(5)
No. Firms-US (200m)	0.00215****				
	(0.000403)				
No. Firms-DS (200m)	0.000631*				
	(0.000350)				
No. Firms-Same (200m)	-0.00324***				
	(0.00121)				
No. Firms-US (400m)		0.000471****			
		(0.000132)			
No. Firms-DS (400m)		0.0000631			
		(0.0000989)			
No. Firms-Same (400m)		-0.00105**			
		(0.000452)			
No. Firms-US (600m)			0.000192***		
			(0.0000742)		
No. Firms-DS (600m)			0.0000364		
			(0.0000519)		
No. Firms-Same (600m)			-0.000686***		
			(0.000263)		
No. Firms-US (800m)				0.0000875	
				(0.0000534)	
No. Firms-DS (800m)				-0.00000517	
				(0.0000323)	
No. Firms-Same (800m)				-0.000481***	
				(0.000181)	
No. Firms-US (1000m)					0.0000356
					(0.0000426)
No. Firms-DS (1000m)					-0.0000131
					(0.0000219)
No. Firms-Same (1000m)					-0.000281**
					(0.000140)
P-Val: US=Own	0.000***	0.004***	0.004***	0.009***	0.066*
P-Val: DS=Own	0.004***	0.020**	0.009***	0.010**	0.057*
P-Val: US=DS	0.004***	0.011**	0.075*	0.125	0.299
Mean of Dependent Variable	.26	.26	.26	.26	.26
Threshold (meters)	200	400	600	800	1000
N	9,111	9,111	9,111	9,111	9,111

Notes: Robust standard errors clustered at the block-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Firm Survival is a dummy variable that takes the value 1 if the firm was operating in the industrial area in 2018, and 0 otherwise. Post double LASSO (Belloni et al., 2013) estimates, including product fixed effects, number of plots with missing plot assignments and total number of plots in the distance threshold included as controls.

Table A7: Correlation Between Missing Geocoding and Timing of Firm Lottery

	(1) Surveyor Geocode Missing	(2) Google Geocode Missing
Lottery Year	-0.00125	-0.000170
	(0.00115)	(0.000930)
Constant	2.601	0.421
	(2.303)	(1.862)
Mean of the Dependent Variable	0.0980	0.0800
N	15010	20186
R2	0.0000749	0.00000167

Notes: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Surveyor Geocode Missing is a dummy variable that takes the value 1 if the surveyor was unable to find the address, and 0 otherwise. Google Geocode missing is a dummy variable that takes the value 1 if Google Maps was unable to find a geocode for the firm within the greater Delhi area, and 0 otherwise.

B The Impact of Lottery Timing on Movement to the Industrial Area

In this section, we test that the timing of receiving a plot in the industrial area in the lottery process drove the timing of moving to the industrial area. We begin by estimating the following equation for outcomes related to movement to the industrial area, analogous to the balance equation (1):

$$Y_j = \alpha + \mu \text{Year of Assignment by Lottery}_j + \epsilon_j \tag{7}$$

where Y_j is the year in which the firm took possession of the industrial plot. As mentioned in the main text, this action is the first that firms undertake to begin construction on their assigned plots, and move to the industrial area. The coefficient of interest is μ, which measures the marginal impact of a one-year delay in "winning" a lottery for firm j.

Results from estimating Equation (7) are presented in Column 1 of Table B1. A one year delay in getting a plot assignment in a lottery causes just over a year's delay in the firm taking possession of the plot. Column 2 replaces Year of Assignment by Lottery$_j$ with an indicator for having been assigned a plot by lottery by 2004. We see that if a firm did not receive a plot assignment by 2004 this delayed it taking possession of the plot by 10 years, indicating a large gap between firm removal for firms that received a plot in an early vs. late lottery. This was because a third industrial area had to be constructed to accommodate these remaining firms. Thus, these results indicate that the lottery timing generated variation in the timing of firm removal from a neighborhood, allowing us to estimate the causal impacts of firm presence on environmental quality.[15] In our main analyses, we exploit the large delay in movement caused by the relatively small difference in year of plot assignment by lottery between firms assigned plots by 2004 and those assigned after.

Thus, these results indicate that the lottery timing generated variation in the timing of firm removal, allowing us to estimate the causal impacts of firm presence.

[15] We also show results at the neighborhood level i.e. we estimate

$$\begin{aligned}\text{Log(Firms Taken Plot Possession)}_i = &\alpha + \beta \text{Log(Firms Assigned Plots by 2004)}_i \\ &+ \nu \text{Log(Total Number of Firms Relocated)} + \epsilon_i\end{aligned}$$

where i denotes the neighborhood. We estimate this equation for the year 2006, as well as nine years later when the third industrial area is open, in 2015. Results are presented in Table B2. It shows that a one percent increase in the number of firms assigned plots early increases the number of firms who took plot possession by nearly one percent in 2006, so an almost one-for-one increase. Nine years later, in 2015, the first stage estimate is about half that, as firms continue to be moved to the industrial areas over time.

Table B1: First Stage of Lottery Timing Impacting Firm Movement to Industrial Area

	Year of Plot Possession	
	(1)	(2)
Lottery Year	1.153****	
	(0.0110)	
Lottery Year > 2004		9.968****
		(0.0691)
Constant	-301.4****	2004.7****
	(22.11)	(0.0186)
Lottery Year	All	
Mean of Dependent Variable	2006.16	2006.16
R-Squared	0.4578	0.6454
N	19553	19553

Notes: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Year of Plot Possession is the year in which the firm took possession of the plot in the industrial area. Lottery Year is the year in which a firm received their plot assignment.

Table B2: Impact of Lottery Timing on Firm Movement to Industrial Area: Neighborhood-Level

	(1) Log(Firms Given Plot Possession By 2006)	(2) Log(Firms Given Plot Possession) By 2015)
Log(Firms Assigned Plot in Lottery by 2004)	0.997****	0.502****
	(0.0437)	(0.0467)
Constant	-0.112****	-0.00445
	(0.0169)	(0.00409)
Year	2006	2015
Mean of Dependent Variable	23.88	28.93
R-Squared	0.9818	0.9955
N	555	579

Notes: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Firms Given Plot Possession by 2006 is the total number of firms in a neighborhood who have taken possession of their industrial plot by 2006. Firms Given Plot Possession by 2015 is the total number of firms in a neighborhood who have taken possession of their industrial plot by 2015. Firms Relocated by 2004 is the total number of firms in a neighborhood who have been assigned an industrial plot in a lottery by 2004. All regressions include controls for the log of total number of firms relocated in a neighborhood.

C Two-Way Fixed Effect Specifications for the Impact of Firm Removal on Environmental Quality

In this appendix, we estimate two two-way fixed effect (TWFE) specifications for the effect of firm removal on air pollution as measured by PM 2.5 concentrations. This specification, complementary to the main specification, also includes sending neighborhood fixed effects to difference out time-invariant components of air pollution levels across neighborhoods.

The TWFE specification conditions nonlinearly on the total number of firms ever relocated from a neighborhood. We divide neighborhoods into groups based on quartiles of the number of firms ever relocated from them, then estimate the impact of the proportion of firms assigned a plot by 2004 in each group on air pollution. We also present a specification which uses deciles of the number of firms relocated ever relocated from a neighborhood instead of quartiles.

In years following the beginning of firm movement to the industrial areas, neighborhoods with a higher proportion of early plot assignments should see reductions in pollution relative to those with a lower proportion of early plot assignments. Based on the average plot possession years of firms assigned plots by 2004 and after, we hypothesize that the reduced form effect of getting a plot assignment early should be small prior to 2005 since neighborhoods with early vs later plot assignments should have similar pre-trends in their outcomes. Therefore, the interactions terms are omitted for the year 2004.

We begin with an event-study specification:

$$\begin{aligned}
\mathrm{Y}_{it} =& \alpha + \beta_t \sum_{t=1998}^{t=2015, t\neq 2004} \left[\mathrm{Bin}_i \times \text{Prop Firms Lottery by } 2004_i \times \mathbb{1}(\text{Year} = t)\right] \\
&+ \xi_t \sum_{t=1998}^{t=2015, t\neq 2004} \left[\text{Prop Firms Lottery by } 2004_i \times \mathbb{1}(\text{Year} = t)\right] \\
&+ \pi_t \sum_{t=1998}^{t=2015, t\neq 2004} \left[\mathrm{Bin}_i \times \mathbb{1}(\text{Year} = t)\right] + \psi_i + \tau_t + \epsilon_{it},
\end{aligned}$$

where Y_{it} is average annual PM 2.5 concentration for neighborhood i in year t. The main coefficients of interest are β_t, which measure the marginal impact of a greater proportion of firms winning a lottery earlier in the process (i.e. by 2003) in each year. Prop Firms Lottery by 2004 is the proportion of firms in neighborhood i that received a plot assignment by 2004. Bin takes on 1 of 4 values (and in the decile specification, one of 10 values) depending on the number of firms relocated. It is 1 if the number of relocation-eligible firms is in the first quartile, 2 if it is in the second, and so on (in the decile specification, it is 1 if the number of relocation-eligible firms is in the first decile, 2 if it is in the second, and so on). The analogous homogeneous-effect-by-year specification replaces the year dummies implicit in the definition of the β_t parameters with a post-dummy 2004 dummy. As in the main text, we cluster standard errors at the 1 km^2 grid cell level.

Event study estimates for this specification are presented in Figure C1, with Figure C1a presenting the results using the specification with four bins, and Figure C1b using the specification with ten bins. Aggregate results are presented in Table C3 (Column 1 presenting the results using the specification with four bins and Column 2 presenting the results using the specification with ten bins). The event study estimates again do not show evidence of a pre-trend before 2004, and show a negative impact after. The aggregate effect estimate shows that conditional on a baseline number of relocation-eligible firms, a marginal increase in the proportion of firms relocated early causes a 1.02 μg/m^3 reduction in fine PM. Normalizing by the standard deviation of proportion relocated early, this would imply that a one standard deviation increase in the proportion of firms relocated early causes a 0.45 μg/m^3 reduction in fine PM. Results using the specification with 10 bins are consistent with the specification using 4 bins, showing a negative coefficient on the triple interaction that is precisely estimated.

Figure C2 presents the triple interaction estimates converted to per-firm effects of firm removal, allowing for direct comparison with the reduced form of the instrumental variables specification estimated in Column 2 of Table A2. The total effect of the proportion of firms removed on PM is $\beta \cdot \text{Bin}_i + \xi$, where β is the triple interaction coefficient and ξ is the main effect, the coefficient on the proportion-by-post interaction. The per-firm effect of removal is $(\beta \cdot \text{Bin}_i + \xi)/\text{Total Number of Firms Relocated}_i$. To approximate the average of this effect within each bin in a manner that uses information from grid cells that relocated zero firms, we divide by the mean of Total Number of Firms Relocated$_i$ within each bin to obtain: $(\beta \cdot b + \xi)/\bar{N}_b$. Standard errors are computed using the delta method, accounting for the covariance between β and ξ.

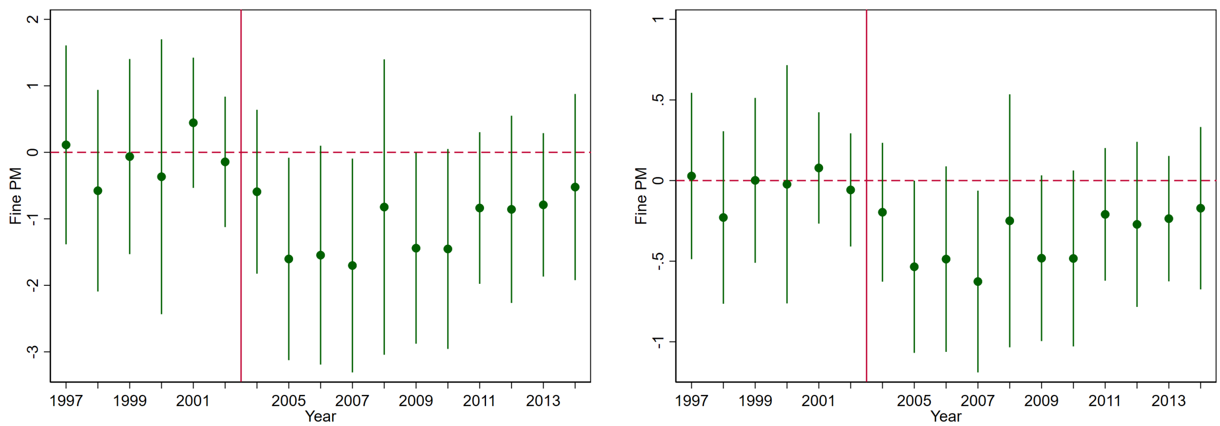
For comparison we plot the pooled reduced form coefficient associated with Column 2 of Table A2, which is obtained by regressing PM on the count of firms assigned a plot by 2004, controlling for total firms relocated and year fixed effects. The triple interaction per-firm effects vary across bins, with larger-magnitude effects in neighborhoods with fewer firms ever relocated (lower bins). The reduced form coefficient is constant across bins and approximately appears as a precision-weighted average. The comparison of these two reduced form approaches reassures us that our main results are robust to alternative ways of conditioning on the number of firms relocated, and to including neighborhood fixed effects. This pattern of heterogeneous effects is also consistent with the concave relationship between firm presence and pollution documented in Column 2 of Table 2, where the coefficient on firms squared is negative.

Table C3: Impact of Random Firm Removal on Air Pollution (Fine PM, $\mu g/m^3$): Triple Interaction Between Baseline Number of Firms and Proportion Lotteried Early

	Fine PM ($\mu g/m^3$)	
	(1)	(2)
Triple Interaction(Bins=4)	-1.021***	
	(0.376)	
4 Bins X Post	0.894***	
	(0.300)	
Proportion 2004 × Post	1.214**	0.545
	(0.600)	(0.425)
Triple Interaction(Bins=10)		-0.330**
		(0.135)
10 Bins X Post		0.293***
		(0.108)
Constant	106.2****	106.5****
	(0.298)	(0.212)
Mean of Dependent Variable	106.89	106.89
R-Squared	0.9767	0.9767
N	10,692	10,692

Notes: Standard errors clustered at the neighborhood (1km by 1km) grid cell level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$. Fine PM is the annual concentration of fine particular matter in $\mu g/m^3$ at the neighborhood-year level. Triple Interaction (Bins=4) is the triple interaction between three variables. The first is a variable (labeled "4 Bins" in the table) that takes the value 1 if the baseline total number of firms relocated is in the first quartile, 2 if it is the second, and so on. The second is the proportion of firms in the neighborhood that had received an industrial plot by 2004 (labeled "Proportion 2004" in the table). The third is a dummy variable that takes the value 1 if the year is 2005 or later (labeled "Post" in the table). Triple Interaction (Bins=10) is the triple interaction between three variables. The first is a variable (labeled "10 Bins" in the table) that takes the value 1 if the baseline total number of firms relocated is in the first decile, 2 if it is the second, and so on. The second is the proportion of firms in the neighborhood that had received an industrial plot by 2004 (labeled "Proportion 2004" in the table). The third is a dummy variable that takes the value 1 if the year is 2005 or later (labeled "Post" in the table).

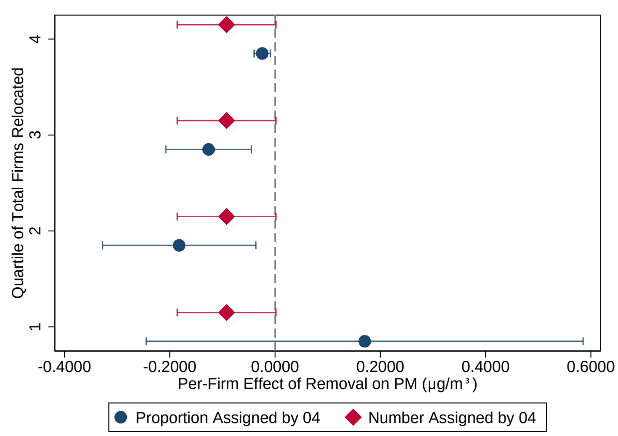
Figure C1: Impact of Random Firm Removal on Air Pollution (Fine PM, $\mu g/m^3$): Triple Interaction Between Baseline Number of Firms and Proportion Lotteried Early, 4 and 10 Bins



(a) Four Bins (b) Ten Bins

Notes: Triple interactions presented of three variables. The first in Panel (a) is a variable that takes the value 1 if the baseline total number of firms relocated is in the first quartile, 2 if it is the second, and so on. In Panel (b), it is a variable that takes the value 1 if the baseline total number of firms relocated is in the first decile, 2 if it is the second, and so on. The second is the proportion of firms in the neighborhood that had received an industrial plot by 2004. The third are year dummy variables (2004 is omitted). Grid ID and year fixed effects, as well as all double interactions included. Bars represent 90% confidence intervals.

Figure C2: Reduced Form Effect of Firms Assigned a Plot by 2004 on PM 2.5



Notes: The figure shows per-firm effects of the number of firms assigned a plot by 2004 on PM 2.5 by quartile of total firms relocated. Proportion Assigned by 04: we convert the effect of the proportion of firms assigned a plot by 2004 from Table C3 to an approximate per-firm effect according to the procedure described in Appendix C. Reduced Form: coefficient on number of firms assigned a plot by 2004, constant across bins. This is the reduced form associated with the instrumental variable specification from Column 2 of Table A2. Horizontal bars represent 90% confidence intervals.

D Fully Heterogeneous Spillover Specification

In this section, we develop an asymmetric, fully heterogeneous counterpart of Equation (5). Our aim here is to investigate the extent to which spillovers as whole can be explained by the Marshallian heterogeneity we allow for in Equation (5). We find that the large majority of heterogeneity in spillovers is indeed driven by Marshallian relationships between the goods firms produce.

Our fully-heterogeneous model estimates the following equation.

$$Active_{ik} = 1\Big\{ \sum_{m=1}^{M} \kappa_m 1\{product_i = m\} + \sum_{m,n \neq (M,M)} \delta_{mn} 1\{product_i = m\} NumberOfFirms_{nk} \\ + \theta TotalNumberOfFirms_k + \sum_{l} \phi_l 1\{Lottery_i = l\} + \varepsilon_{ik} \geq 0 \Big\} \tag{8}$$

where ε_{ik} follows a logistic distribution. The spillovers here are asymmetric because the effect of the count of n-producers in industrial cluster k depends on what i produces.

As in Section 5.3, product indices $\{m : m < M\}$ correspond to 3-digit product codes with 28 or more firms assigned, and M to all other products. We could estimate Equation (8), and then run a meta-regression of the estimated $\hat{\delta}_{mn}$ parameters on indicators for input-output relationships between each mn pair. However, since the dimension of the δ vector is the square of the number of products we consider (180 distinct products) minus one, the individual parameters would be very imprecisely-estimated.

We instead take a Bayesian approach to combine the two steps and regularize our estimates. We specify

$$\delta_{mn} = \alpha^{mn}_{upstream} 1\{n \to m\} + \alpha^{mn}_{downstream} 1\{n \leftarrow m\} + \alpha^{m}_{own} 1\{n = m\} + \xi_{nm}$$

where $1\{n \to m\}$ indicates that n is upstream of m and $1\{n \leftarrow m\}$ that n is downstream of m. We use the definitions of upstream and downstream described in Section 3.2.3. We assume κ_m, $\alpha^{mn}_{upstream}$, $\alpha^{mn}_{downstream}$, α^{m}_{own} are independently normally distributed with means μ_κ, $\mu_{upstream}$, $\mu_{downstream}$, μ_{own}, and standard deviations σ_κ, $\sigma_{upstream}$, $\sigma_{downstream}$, σ_{own}. The idiosyncratic effects ξ are independently normally distributed with mean 0 and standard deviation σ_ξ. We assign μ_κ a Normal(0, 10) prior and σ_κ a LogNormal(0, 0.75) prior. For numerical stability we normalize the firm count variables to have mean zero and standard deviation one. For the corresponding spillover effect parameters (upstream, downstream, own-industry, and idiosyncratic effects), we use priors appropriate for these standardized counts: Normal(0, 0.7) for means and LogNormal(-2.6, 0.75) for standard deviations. These priors are uninformative for shares of variation in δ_{mn} across product pairs due to Marshallian vs. idiosyncratic features.

We estimate the model using Hamiltonian Monte Carlo in Stan (Carpenter et al., 2017) and calculate the posterior share of the variance in heterogeneous spillover effects δ_{nm} attributable to features other than input-output linkages and net competitive effects between n and m. Table D4 shows the results, comparing the posterior distribution of the share to its prior. We find that Marshallian forces explain the large majority of variation in the δ_{mn} parameters. The posterior mean for the share of the variance in δ_{mn} explained by factors other than input-output relationships is 0.088. The 90% credible interval for the share of variation in spillovers due to non-Marshallian factors ranges only from 0.013 to 0.234. This is as compared with the 0.564 mean share due to non-Marshallian features over 10,000 draws from the prior, with a 5% to 95% interquantile range of 0.066 to 0.970. Therefore, a large fraction of spillovers are indeed mediated through Marshallian forces, which motivates our main spillover specifications.

Table D4: Share of Variation in Firm Spillovers Due to Non-Input-Output Effects

		Quantiles			
	Mean	5%	25%	75%	95%
Posterior	0.088	0.013	0.037	0.117	0.234
Prior	0.564	0.066	0.306	0.838	0.970

Notes: Quantiles of the share of variation in the effect of the share of producers of product n in a product-m-producer's block in the industrial area which is unexplained by the input-output relationship between n and m. 3000 burn-in Hamiltonian Monte Carlo draws and 2000 samples from the posterior per chain. 4 chains. R-hat criterion indicating good chain mixing lies below 1.005 for all parameters.

The results in Table D4 depend on how frequently we classify product pairs as upstream and downstream from one another. An alternative way of evaluating the importance of non-input-output-based spillovers is to compute the posterior distribution of $\frac{\sigma_{\xi}}{\sigma_{upstream}+\sigma_{downstream}+\sigma_{own}+\sigma_{\xi}}$. We do this in Table D5. The 90% credible interval for this quantity ranges only from 0.002 to 0.081. The posterior mean of the share is 0.025. This is as compared to the prior mean of 0.246, with the 5% to 95% interquantile range of the prior distribution ranging from 0.009 to 0.772.

Table D5: Relative Magnitude of Variation in Firm Spillovers Due to Non-Input-Output Effects

		Quantiles			
	Mean	5%	25%	75%	95%
Posterior	0.025	0.002	0.007	0.031	0.081
Prior	0.246	0.009	0.051	0.376	0.772

Notes: Quantiles of the share of variation in the effect of the share of producers of product n in a product-m-producer's block in the industrial area which is unexplained by the input-output relationship between n and m. 3000 burn-in Hamiltonian Monte Carlo draws and 2000 samples from the posterior per chain. 4 chains. R-hat criterion indicating good chain mixing lies below 1.005 for all parameters.

E Matching Firm Product Information to Annual Survey of Industries Commodity Classification Codes (ASICCs)

E.1 Procedure

Before any analysis we created a custom pipeline using the GPT-4-turbo model to assign Annual Survey of Industries Commodity Classification (ASSIC) codes to firms based on the goods they report producing. We performed prompt engineering to guide the model using a set of detailed instructions that included accounting for potential typos and nuanced elements like distinguishing between manufacturing a particular product vs. manufacturing the machinery that produces that product. The output consists of a single product code assignment or a "missing" label for cases where GPT could not find a good match. The details of the prompts used are included in the next section. These requests are submitted to the GPT-4-turbo model using OpenAI's API, and the model's responses were parsed and integrated back into the output data. GPT models are inherently non-deterministic. To minimize randomness and enhance reproducibility, we set the temperature parameter, which measures response entropy, to 0.

E.2 Human Evaluation

To evaluate the results, we randomly selected samples of unique product descriptions and compared them to human assignments to ASICC codes done by research assistants. We compared the human assignments with the results of different versions of the GPT model concluding that GPT model assignment produced a reasonable and efficient fit for nearly all cases and was more consistent than a team of research assistants. We also used the human evaluation to extensively test different input prompts in search of best matching. This resulted in an extensive initial prompt that covers the edges cases in which GPT-4 performed poorly.

E.3 Output Processing

We process the assignment requests using the Batch method of OpenAI's API. We download the resulting output as JSON files and extract the assignment answer for each firm. GPT provides up to 5 product codes in order from most to least preferred. We pick the most preferred code. The full list of industry codes and detailed prompt is available upon request.