Journal of Quantitative Economics (2023) 21:641–662
https://doi.org/10.1007/s40953-023-00349-8

ORIGINAL ARTICLE

Model-Selection Inference for Causal Impact of Clusters and Collaboration on MSMEs in India

Samarth Gupta[1]

Accepted: 14 April 2023 / Published online: 4 May 2023
© The Author(s), under exclusive licence to The Indian Econometric Society 2023

Abstract
Do firms benefit more from agglomeration-based spillovers than the technical know-how obtained through inter-firm collaboration? Quantifying the relative value of the industrial policy of cluster development vis-à-vis firm's internal decision of collaboration can be valuable for policy-makers and entrepreneurs. I observe the universe of Indian MSMEs inside an industrial cluster (Treatment Group 1), those in collaboration for technical know-how (Treatment Group 2) and those outside clusters with no collaboration (Control Group). Conventional econometric methods to identify the treatment effects would suffer from selection bias and misspecification of the model. I use two data-driven, model-selection methods, developed by (Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on treatment e ects after selection among high-dimensional controls. Review of Economic Studies, 81(2):608 650.) and (Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post selection and post regulariza- tion inference in linear models with many controls and instruments. American Economic Review, 105(5):486 490.), to estimate the causal impact of the treatments on GVA of firms. The results suggest that ATE of cluster and collaboration is nearly equal at 30%. I conclude by offering policy implications.

Keywords Firm Performance · Entrepreneurship · Formal and Informal Sectors · Shadow Economy · Institutional Arrangements

JEL Classification L25 · L26 · 017

Introduction

The role of clusters and collaboration in valuable knowledge transfer across firms has received substantial attention. However, which of the two provides more gains to the firm? While both can be impactful, they create value through different channels.

✉ Samarth Gupta
samarth.gupta@ahduni.edu.in

[1] Amrut Mody School of Management, Ahmedabad University, Ahmedabad, India

Schmitz and Nadvi (1999) term cluster's benefits as passive, where gains "*fall into the producer's lap*", whereas collaboration requires "*active joint effort*" between consenting parties.

Due to the contrasting mechanisms of knowledge transfer in clusters and collaboration, quantifying the relative benefits of the two can be helpful for many stakeholders. Government agencies in developing countries actively pursue contract-enforcement reforms to promote business collaboration (WorldBank 2020). Policy makers also design place-based policies such as Special Economic Zones (Neumark and Simpson 2014) to increase firm productivity. Estimating the value of these effects, therefore, will aid cost–benefit analysis of the two policies and help prioritize them. From the perspective of a manager or an entrepreneur, collaboration with other firms or relocating to a cluster are both valuable but costly decisions. Estimates of the gains of either treatment may aid in the decision-making process. Finally, estimating the value of these two factors in firm performance will contribute to the literature on TFP spread across firms (Syverson 2011).

I use the MSME Census 2006–07 of India to estimate the impact of clusters and collaboration on firm performance. This dataset enumerates more than 1.2 million registered micro, small, and medium-sized enterprises in India in 2006[1]. It provides detailed information on several firm-level variables such as outcome and productivity metrics such as gross value added (GVA), the value of output and inputs used, labour size, entrepreneur's socio-demographic characteristics and enterprises' location, industrial and operational categories.

I classify firms in this dataset into three mutually exclusive groups. Treatment Group 1 is the set of firms inside a cluster but without collaboration. Firms outside a cluster but collaborating with other firms are classified under Treatment Group 2. Finally, Control group consists of firms which are neither collaborating nor inside a cluster.

Identifying the impact of these treatments is challenging as they are not randomly assigned. More productive firms may be more likely to select into a cluster. Likewise, more productive firms may have more opportunities to collaborate. Higher productivity of firms in a treatment group, then, reflects the selection bias along with the effect of a treatment.

Several conventional methods exist to address the selection bias in a non-experimental observational study (Rosenbaum and Rubin 1983, 1985). However, these methods suffer from model misspecification—either due to the exclusion of a relevant variable or the inclusion of an irrelevant variable. While the former biases the estimates, the latter creates efficiency losses (Rao 1971).

[1] As per the old definition, a firm with initial investments in plants, machinery and equipment less than INR 2.5 millions is defined as micro, with investment between INR 2.5 and INR 50 millions is small, and with investment exceeding 50 millions is defined as medium enterprise. The definition was reclassified on July 1, 2020. Now, an enterprise with investment in plants, machinery and equipment less than Rs. 10 millions and annual turnover less than Rs. 50 millions is defined as micro; with investment in plants, machinery and equipment less than Rs. 100 millions and annual turnover less than Rs. 500 millions is defined as small; and with investment in plants, machinery and equipment less than Rs. 500 millions and annual turnover less than Rs. 2500 millions is defined as medium.

To address selection bias while avoiding model misspecification, I use two recently developed model-selection methods for causal inference. These methods are post-double selection lasso (Belloni et al. 2013) and post-lasso orthogonalized inference (Chernozhukov et al. 2015). Under both techniques, a feature selection method, such as LASSO (Tibshirani 1996), is used to select the variables with the highest predictive power for the outcome variable and the treatment indicator. Belloni et al. (2013) show that Average Treatment Effect (ATE) can be obtained from a regression of the outcome on the treatment indicator if one uses the union of variables selected from the two LASSO regressions as covariates. Under Chernozhukov et al. (2015), the dependent variable and treatment indicator are first orthogonalized using the variables selected from the two LASSO regressions. ATE is then obtained by regressing orthogonalized dependent variable on the orthogonalized treatment indicator.

MSME Census 2006–07 provides a wide array of observable covariates. I include 2500 covariates which absorb many confounders, suggested by prior theoretical and empirical literature, such as firm characteristics, location-industry identifiers and socio-demographic features of the entrepreneurs, which can drive selection into the treatment as well as the outcome. For the outcome variable, I use the log of Gross Value Added (GVA) of the firm, which is the difference between the value of output and inputs of the firm.

By both methodologies, a firm inside a cluster experiences an increase of 30% in its GVA. The effect is significant at the 1% level. Similarly, ATE of collaboration on GVA is nearly 32% using both methods. Thus, the average productivity gains for a firm are marginally higher under collaboration compared to being inside a cluster.

While statistically, the ATE of both treatments is similar, the results have critical implications for policy-makers. There is renewed interest in developing countries, particularly India, for productivity-enhancing industrial policies. For example, in April 2020, the government of India announced the construction of multiple industrial clusters as a long-term relief measure against COVID-19 induced recession. However, these industrial clusters are costly and create negative externalities (de Groot et al. 2009; Busso et al. 2013). On the other hand, improving legal and regulatory standards which promote inter-business partnerships do not carry such costs/risks. In this paper, I show that a firm's productivity gains under collaboration are nearly as high as being inside a cluster. Thus, government agencies can create almost the same productivity effects for firms through low-cost improvements in legal and regulatory environments as from cluster development which carries the risk of negative externalities.

The paper also contributes to the literature on factors affecting productivity dispersion across firms. Surveying this literature, Syverson (2011) classifies these factors into two categories—a firm's internal decision-making and its external environment. An important question in this field is how can one reduce productivity dispersion. In this paper, I record the relative advantages of one internal (collaboration) and one external factor (cluster) in generating performance differences. The results have two implications for this literature. First, direct networking and contractual arrangements between firms for technical knowledge may be as valuable as accidental spillovers it obtains from its environment. Second, productivity gains in a

cluster are local to the region, whereas collaboration benefits are more mobile, indicating the potential for performance dispersion.

Most other studies in India which look at the benefits of collaboration or clusters restrict attention to a few successful cases (described later). This is the first study to estimate the impact of these interventions in India on the entire universe of MSMEs. The estimates produced here are the population estimates of ATE.

Machine learning methods such as LASSO are increasingly supplanting conventional econometric techniques. Few recent papers have used machine learning methods to predict the performance of firms instead of estimating causality of treatments on firm performance (Coad and Srhoj 2019; Bargagli-sto et al. 2020). To the best of my knowledge, this is the first paper to use model selection methods for estimating causal impacts of interventions on firm performance.

The rest of the paper is organized as follows. In "Clusters and Collaboration: Literature Review", I provide an overview of the literature on industrial clusters and collaboration between rms. In section "Identification Strategy", I provide a brief description of the identification strategy used in the paper. Those who are well versed the propensity score matching methods may skip this section. Section "Data and Summary Statistics" describes the MSME Census in more details along with some summary statistics. Section "Results" provides the main results and "Conclusion"concludes.

Clusters and Collaboration: Literature Review

Knowledge spillover theory of entrepreneurship identifies human capital, its acquisition and its application as the determinants of an entrepreneur's success (Acs et al. 2013; Ghio et al. 2015). There are several modes of knowledge acquisition. In this paper, I focus on two of these modes—spillovers in clusters and technical know-how from collaboration, both of which can augment the human capital of the entrepreneur[2]. Firms inside a cluster experience positive spillovers from surrounding firms. Collaboration with other firms also benefits firms through technology transfer. Thus, both these modes of knowledge transfer can have a considerable impact on firm performance.

Extensive literature exists on the origins and types of positive externalities of agglomerations. Combes and Gobillon (2014) provide a framework to identify determinants of various agglomeration spillovers. The MSME Census 2006-07 defines a cluster as a special zone, with at least 100 firms of the same industry. As such, knowledge spillovers through local specialization, also known as the

[2] There are other forms of knowledge transfer and human capital acquisition. Rasiah and Govindaraju (2009) record value of research and development for rms in industry-university partnerships. Unger et al. (2013) finds that application of knowledge and skills have a stronger relation with enterprise performance than formal education. Lafontaine and Shaw (2014) and Konstadakopulos (2000) nd evidence for learning-by-doing in performance of an entrepreneur

Marshall-Arrow-Romer spillover, are most likely present in this context (Neffke et al., 2009)[3]. These local specializations may create gains for firms through various channels. For example, Storper and Venables (2004) find how proximity with one's competitors advances skill transfer through imitation. Duranton and Puga (2004) find how local specialization attracts highly skilled labour and reduces matching and search costs for employers.

Given the productivity generation of clusters, they have received substantial policy attention as well (Schmitz 1995; Nadvi 1996; Schmitz and Nadvi 1999; de Groot et al. 2009; Neumark and Simpson 2014; Brakman and van Mar-rewijk 2012; Lucena-Piquero and Vicente 2019). For country-specific reviews, see Palit and Bhattacharjee (2008) for India, Ge (1999) for China and Leong (2013) for a comparison of SEZs in India and China. For India, in particular, evidence on the micro-economic impact of industrial clusters comes mostly from case studies of successful and selective auto-mobile clusters (Venkataramanaiah and Parashar 2007; Okada and Siddharthan 2008). In contrast, this paper conducts a more holistic review of all clusters recorded in the MSME Census 2006–07 to find the causal impact on firms which choose to join a cluster. Thus, the implications of the results in the paper are far wider.

The typology of collaborations has also been well recorded. An interdisciplinary survey suggests that inter-firm coordination overcomes market failures and costly hierarchies (Grandori and Soda 1995; Caglio and Ditillo 2009). In MSME Census 2006–07, enterprises are asked if they collaborate with other firms for technical know-how. As such, the benefits of inter-firm coordination recorded in this paper reflect the economies of scope which reduce costs of technology acquisition (Teece 1980, 1986, Nieto Jesus and Santamaria 2006; Schilling and Phelps 2007)[4].

The literature on collaboration among Indian firms is sparse. Scott-kemmis and Bell (1985) and Levina and Vaast (2008) show that inter-firm boundaries reduce the innovation and performance effects in Indo-British and Indo-American collaborations, respectively. Biswas et al. (2007) conduct a case study of collaboration among textile Indian SMEs. However, this paper contributes by providing a more extensive study of all Indian MSMEs.

Identification Strategy

Consider a firm i with the following production function:

[3] The other two types of agglomeration spillovers are Jacobs' externalities (Jacobs, 1969), which arise due to industrial diversity, and urbanization externalities, which arise when rms are present in or proximate to key infrastructural points such as large cities (de Groot et al. 2009).

[4] Other benefits of collaboration include quasi-integration between firms (Blois 1972), hybrid organizational arrangements which reduce governance costs (Williamson 1981, 1991; Dyer 1998) and overcoming market failures (Powell 1987, 1990; Bradach and Eccles 1989), and, *inter-alia*, complementarity of resources which lead to innovation (Richardson 2003). See also Evan (1966); Whitley (1991); Granovetter (1983, 1985) for views on inter- firm collaboration from outside the economics literature.

$$Yi = e\gamma\ TeX_i\ \beta e\mu$$

where, Y_i is performance of firm i, such as productivity, and X are covariates of the firm/entrepreneur which may influence output. T is a binary variable that takes value 1 if the firm received a treatment, and 0 otherwise. μ is a random error which is i.i.d over firms and $\mu \sim N\ (0,\ \sigma)$.

Applying log transformation of the production function and denoting log Y_i by y_i provides the following linear expression:

$$y_i = \gamma T + \beta X_i + \mu_i \tag{1}$$

If firms receive treatment randomly, then the difference between the mean output of firms with and without treatment provides the treatment effect, i.e. $E(T.X_i) = 0$ implies $E(y_i|T = 1) - E(y_i|T = 0) = \gamma$. However, selecting into a treatment is seldom random and may be a function of observable characteristics of the firm/entrepreneur. For example, entrepreneurs with access to a particular network may be more likely to collaborate with other firms. Similarly, entrepreneurs in some geographic zones may be more likely to join a cluster. In such cases,

$$T = \alpha X_i + \kappa_i \tag{2}$$

where κ_i is the error term in the selection process.

Several methods can account for the selection bias in non-experimental observational studies. One can ideally include all the relevant covariates in a regression of outcome on the treatment indicator. Alternatively, one could structurally model the selection equation (2) and match treated and untreated firms on the propensity of selecting into the treatment. However, under either method, model misspecification poses a risk. The exclusion of some variables can lead to omitted variable bias. In contrast, including unnecessary controls can lead to a higher mean squared error on the treatment effect estimator. Selecting variables by relying on prior theoretical literature is susceptible to charges of p-hacking. Owing to this criticism, I use two data-driven methods which help select the list of covariates to be included as controls in linear regression.

Post-Double Selection LASSO

First, I use the post double selection lasso (PDS LASSO from hereon) technique described in Belloni et al. (2013). PDS LASSO obtains the Average Treatment Effect through the following algorithm:

1. First, find the set of covariates which best predict the outcome variable, y_i, using LASSO (Tibshirani 1996). Denote this set as X_1.
2. Second, find the set of covariates which best predict the treatment variable, T, using LASSO. Denote this set as X_2.
3. Regress y_i on T with $X_1 \cup X_2$ as the set of controls. The coefficient on T is the Average Treatment Effect.

Belloni et al. (2013) also provides regularity conditions for inference of the estimator.

Post-LASSO Orthogonalized Regression

Second, I use the post-lasso orthogonalized regression method (Chernozhukov et al. 2015). This process operates through the following algorithm.

1. First, run LASSO of the outcome variable on X to obtain coefficients, $\hat{\beta}$, on covariates.
2. Second, run LASSO of treatment indicator, T, on X to obtain coefficients, covariates[5].
3. Obtain orthogonalized values, $\mu_y = y_i - \hat{\beta}\mathrm{X}$ and $\nu = T - \hat{\theta}\mathrm{X}$.
4. Regress μ_y on ν. The coefficient on ν is the Average Treatment Effect.

$\hat{\kappa}$, on

Asymptotic properties for the estimator are provided in Chernozhukov et al. (2015), which allow inference.

Discussion on the Methodologies

The methodologies in this paper provide two advantages—on covariate selection and on treatment effect estimation.

Advantages in Covariate Selection

The methods in this paper employ LASSO technique to select covariates. The relative advantage of LASSO over OLS in covariate selection are well known—LASSO reduces variance in prediction at the cost of biased estimate. However, LASSO methodology enjoys several advantages over other variable selection methods as well. Tibshirani (1996) compares LASSO, Ridge and subset selection methods and finds that LASSO allows for model interpretability by shrinking some coefficients to zero, unlike Ridge, but also ensures model stability, unlike subset selection. This way, LASSO provides "*advantages of both the methods*". Using simulated data, Hastie et al. (2009) show that coefficient shrinkage in LASSO falls between the coefficient shrinkage in Ridge and other methods like PLC and PRC.

Advantages in Treatment Effect Estimation

The above methodologies provide advantages in treatment effect estimation as well. Like conventional matching methods, these methodologies require the assumption

[5] Since LASSO performs feature selection, some values of $\hat{\beta}$ and $\hat{\theta}$ may be zero. See Hastie et al. (2009) for a more detailed discussion

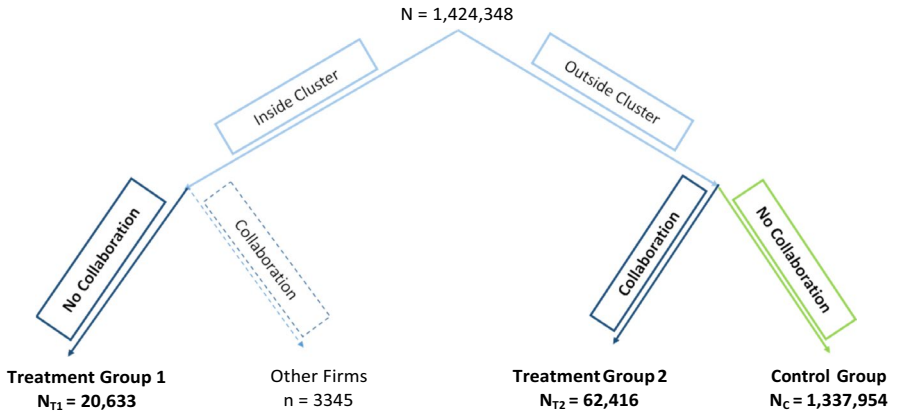


Fig. 1 Selection of Treatment Groups 1 and 2, and Control Group. Figure shows how treatment group 1 and 2, and control group rms were selected from the universe of all firms

of conditional independence (Rosenbaum and Rubin 1983, 1985).; i.e. conditional on observables, treatment is exogenous. However, the conventional methods require the researcher to impose model restrictions (Shortreed and Ertefaie 2017; King and Nielsen 2019), which may increase the coefficient variance (Heckman et al. 1998). Selection of the model in PDS Lasso and Orthogonalized LASSO occurs through data-driven techniques and is independent of the researcher's discretion, which may be susceptible to charges of p-hacking. Alternatively, one could adopt an instrumental variable approach, but satisfying the exclusion restriction assumption is difficult (Angrist and Pischke 2008, see Chapter 4). Thus, model-selection methods allow us to overcome model specification issues without adding more restrictions.

While the conditional independence assumption is a strong requirement, the MSME Census records several high-dimensional covariates. In section "Covariates", I provide a detailed, evidence-backed description of the variables included for selection used in the two methods. Though conditional independence remains untested, including these many controls may help approach this assumption.

Treatment and Control Groups

In this paper, I consider two treatments. The first treatment is whether the firm is inside a cluster developed by the government[6]. The second treatment is whether the entrepreneur obtained technical know-how from a domestic collaboration. To isolate the treatment effects for the two separately, I construct mutually exclusive treatment groups.

Figure 1 shows a schematic to describe treatment groups 1 and 2 and the control group. MSME Census 2006–07 have 1,421,003 observations. Out of this, 23,978

[6] There are also natural clusters, which form due to historical reasons. Since the focus in this paper is on policy implications, I consider only government-developed industrial clusters

are inside some government cluster. 20,633 enterprises inside government clusters do not have a partnership with any other domestic collaborator. These observations constitute Treatment Group 1. Of the firms outside government clusters, 62,416 have obtained technical know-how from collaboration with other domestic firms. These observations constitute Treatment Group 2. The control group is the set of firms outside government clusters which do not have inter-firm collaboration. There are 1,243,149 such enterprises. I identify ATE of treatment groups 1 and 2 against the control group using the model-selection methods, described above.

In addition to the above treatment and control groups, there is another group of 3,345 firms which received both treatments. Comparing this group of firms with any other treatment or control group may be misleading. The concern arises because propensity of one treatment may increase due to the other. In such a situation, it is unclear which treatment is a covariate and which is the outcome. Thus, causal effects for firms receiving both treatments are not presented here. I leave that for future work[7].

Data and Summary Statistics

Several enterprise or factory-level datasets exist in India such as the Unorganized Non-Agricultural Enterprise Survey 2010–11, the Annual Survey of Industries, CMIE-Prowess databases etc. MSME Census 2005–06 provides several advantages which make it suitable for this study.

First, to my knowledge, only MSME Census 2005–06 provides information on whether the enterprise is in a cluster or has a collaboration for technical know-how. Other datasets record the district of the enterprise and not whether the enterprise is in a cluster. Some other datasets provide information on partnerships between firms, whereas MSME Census 2005–06 records collaboration specifically for technical know-how. Second, MSME Census 2005–06 covers a broad information set on a large number of enterprises. An extensive set of covariates for all enterprises allows for model selection inference. Finally, while the database may not be recent, the research question of this paper is still relevant. For e.g., while the definition of MSMEs may have changed over the years, the relative value they incur from these treatments can guide a discussion on this subject.

Table 1 provides summary statistics on socio-demographic characteristics of the owner and industrial characteristics for firms in treatment and control groups. Around 14% of all firms are owned by women. This is slightly higher than in the treatment groups 1 and 2, where nearly 12% and 11% firms are owned by women. Similarly, 10% of all firms are managed by females, as opposed to 9% for firms in treatment groups 1 and 2. Share of manufacturing firms is

[7] Lechner (2005) demonstrates a method where causal effects of multiple, overlapping treatments can be identi ed and estimated. Lopez and Gutman (2017) provides a review of estimation methods for causal effects of multiple treatments. Calia et al. (2016) applied these methods to study the twin effects of innovation and ofshoring on the performance of European firms.

Table 1 Summary statistics on covariates of firms/ entrepreneurs

	Mean	Standard Deviation	Observations
Panel A: all firms			
Female owner	0.140	0.346	1420955
Female manager	0.105	0.306	1421001
Rural	0.468	0.499	1420999
Manufacturing	0.655	0.475	1421001
Panel B: inside cluster (treatment group 1)			
Female owner	0.126	0.331	20633
Female manager	0.090	0.286	20633
Rural	0.457	0.498	20633
Manufacturing	0.791	0.407	20633
Panel C: domestic collaboration(treatment group 2)			
Female owner	0.114	0.318	62394
Female manager	0.092	0.290	62416
Rural	0.448	0.497	62416
Manufacturing	0.719	0.449	62416

Panel A provides summary statistics on major covariates for all firms/entrepreneurs in the MSME Census. Panel B and Panel C provide summary statistics on firms in Treatment Group 1 and Treatment Group 2, respectively

Table 2 Summary statistics on outcomes for firms

	Mean	Standard deviation	Observations
Panel A: all firms			
Gross output (Rs.)	4199185.265	74425988.314	1420220
Total inputs (Rs.)	2621626.759	46370050.446	1418607
Gross value added (Rs.)	1605551.532	48922632.260	1418438
Value of plant and machinery	566184.690	3728200.409	1421001
Panel B: inside cluster (treatment group 1)			
Gross output	6866374.276	63282485.496	20602
Total inputs	4365914.925	31872206.150	20535
Gross value added	2557722.797	46668283.276	20551
Value of plant and machinery	833228.505	3599051.612	20633
Panel C: domestic collaboration(treatment group 2)			
Gross output	11602325.337	1.433e+08	62329
Total inputs	7093805.223	96833354.519	62303
Gross value added	4440024.846	79475803.388	62204
Value of plant and machinery	1132168.485	6577548.263	62416

Panel A provides summary statistics on total output and input of all firms in the MSME Census. Panel B and Panel C provide summary statistics on firms in Treatment Group 1 and Treatment Group 2, respectively

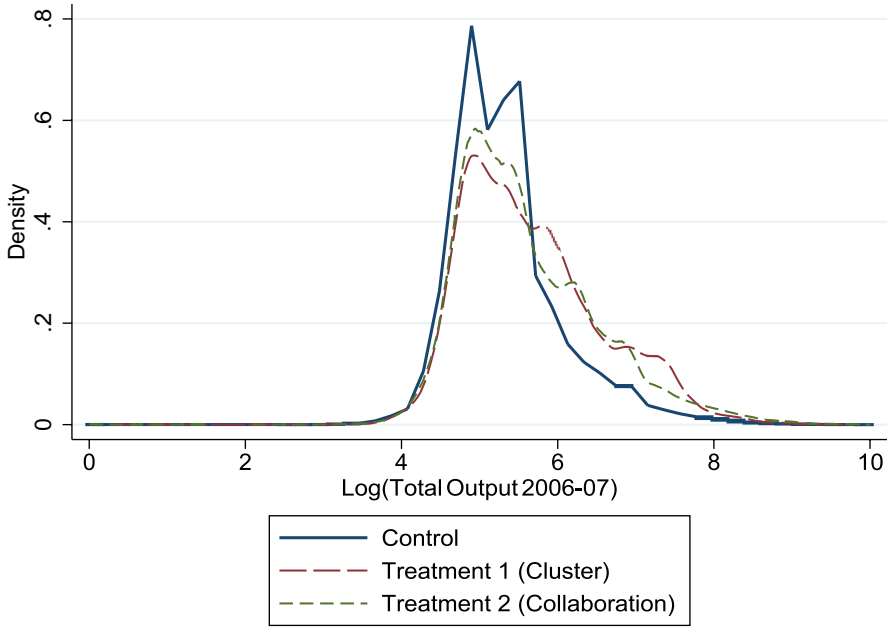


Fig. 2 Distribution of Log Output in 2006–07. Figure shows the distributions of log output of firms in control, treatment group 1 and treatment group 2

substantially higher in clusters (79%) and those with collaboration (71%), compared to the overall ratio (65%).

Table 2 provides summary statistics on the output produced and inputs used by the firms. The average output of all firms is INR 4.2 Millions and average number of workers employed is 5.66 (Panel A). For firms inside a government cluster (Panel B), the average output is INR 6.8 Millions and average number of workers employed is 11.30, both higher compared to all firms. Average output and average workforce for firms with collaboration are INR 11.6 Millions and 9.8, respectively (Panel C). In Figs. 2, 3, I plot the distributions of log output and log inputs, respectively, of firms in the control and treatment groups (output and inputs measured in INR). Log output distributions for firms in either treatment groups stochastically dominate the distribution of log output for control group firms. However, firms in treatment groups also use more inputs, as can be observed from the distributions of their log inputs.

Table 2 also provides mean value of Gross Value Added (GVA), defined as the difference between value of total output and inputs. Average GVA for all firms is INR 1.6 Millions. The corresponding figures for firms in treatment groups 1 and 2 are INR 2.5 Millions and INR 4.4 Millions, respectively. Figure 4 plots the density of log of GVA in control group, treatment group 1 and treatment group 2. For the two treatment groups, distribution of log GVA has a thick right tail.

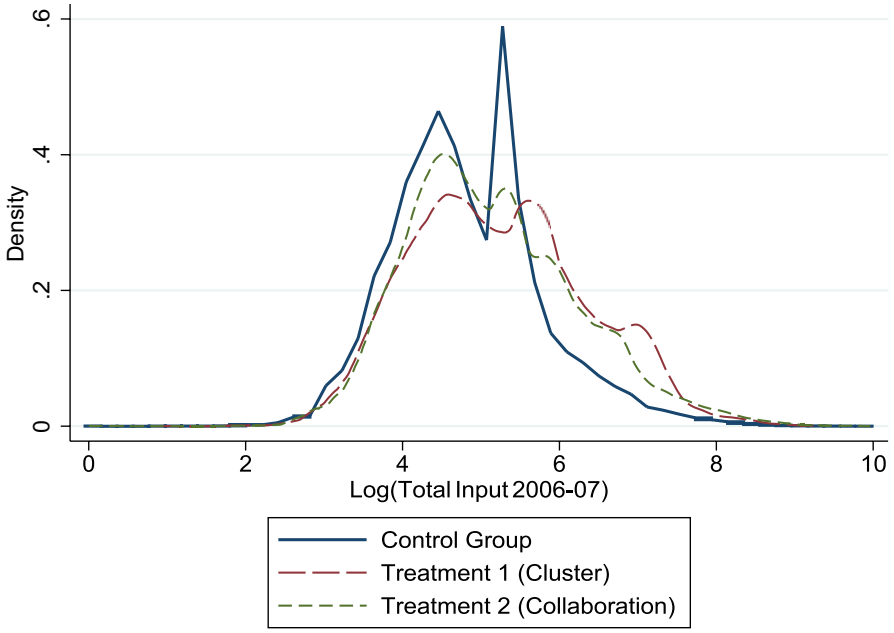


Fig. 3 Distribution of Log Inputs in 2006–07. Figure shows the distributions of log inputs of firms in control, treatment group 1 and treatment group 2

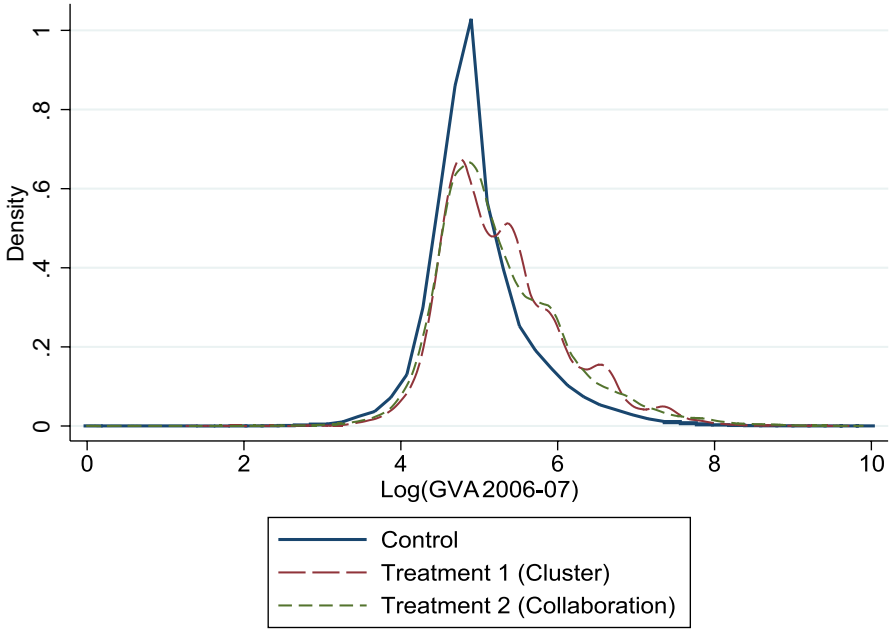


Fig. 4 Distribution of Log GVA in 2006–07. Figure shows the distributions of log GVA of firms in control, treatment group 1 and treatment group 2

Covariates

Although the algorithms used in the paper select the variables, the researcher still has to feed the algorithms with the super-set of covariates. Motivated by prior literature, I use the following set of variables[8].

Firm Characteristics

Firms across and within industries vary substantially in productivity (Gibbons and Henderson 2014). Many papers have shown the impact of trade exposure on performance of firms (Bloom et al. 2016). Organization structure and ownership of the firm can also affect productivity (Hortacsu and Syverson 2009). Following firm characteristics are included in the model selection equations.

Dummy variable for 2-digit industry code of the firm as a covariate.
Dummy for the type of organization. These are proprietary, partnership, private company, public limited company, cooperative and others.
I use original purchase value of plants, machinery and equipment as a covariate for computing propensity score for being in clusters. Cluster developments are usually targeted on the size of the firm. Thus, initial capital of the entrepreneur may determine the choice of entry in a cluster. However, collaboration with other firms may influence initial capital of the firm by direct investment. Thus, instead of being a determinant of treatment group 2, initial capital may be an outcome of being in collaboration. Thus, this variable is not used as a covariate for treatment group 2.

External Characteristics

Local market conditions can have a substantial impact on productivity of firms. Specifically, intra-market competition among firms can induce productivity (Syverson 2004). Similar effects occur due to local level regulations and input markets (Greenstone et al. 2012; Petrin and Sivadasan 2013). For India, in particular, Bhatt et al. (2011) have also found strong correlation in establishment of SEZs and areas with higher inflow of foreign investments in India. Therefore, I include the following indicators

A dummy for district of the enterprise.
A dummy if the enterprise is in rural and urban area.
A dummy for interaction between nature of the enterprise (manufacturing, services and repair) and district of the enterprise.

[8] Ideally, one would want to estimate TFP measures as described in Olley and Pakes (1992) and Levinsohn and Petrin (2003). MSME Census 2006-07 does not provide that detailed information.

Socio-Demographics of the Entrepreneur

I use the following socio-demographic variables as covariates.

Religion: MSME 2006–07 identifies seven categories of religion of the entrepreneur. These are Hindu, Muslim, Sikh, Christian, Jain, Buddhist, and Others. I use a dummy for each religion as a covariate.
Social Category: I include four caste categories: SC, ST, OBC and others.
Gender of the entrepreneur
In addition to this, I also include interactions between gender and social category of the owner.

Overall, there are 2520 covariates.

Results

As described in Sect. "Identification Strategy", under both methods, LASSO regression is run twice—first on the outcome variable and then on the treatment indicator. This helps select the variables which are either used as controls (in post-double selection lasso) or for orthogonalizing the outcome and treatment indicators (in post-lasso orthogonalized regression).

Variable Selection

The model selection criteria follow the LASSO approach where the specification imposes a penalty to the objective function, where the penalty is a function of the coefficient size. Due to this penalty, some coefficients which do not contribute sufficiently enough to the objective function are shrunk to zero. In other words, they are dropped, while others are retained. A more detailed description is available in Hastie et al (2009).

The covariates selected by LASSO are available in the Log files. In this section, I briefly describe the variables which are selected for each treatment indicator and the outcome variable.

Treatment 1: Cluster

LASSO selects 6 covariates for treatment indicator of a cluster. Three of these are interactions between district and nature of enterprise indicators—in some districts, manufacturing industries are more likely to be in a cluster. However, LASSO eliminates most other such interactions indicating that the location and nature of industries are not significant predictors of presence in clusters. LASSO also chooses only one 2-digit NIC indicator. Here again we find that most industry

Table 3 ATE of clusters on log GVA of firms

	(1)	(2)	(3)	(4)
$\hat{\gamma}$	0.701*** (0.010)	0.183*** (0.070)	0.301*** (0.087)	0.306*** (0.087)
Estimation	NA	OLS	PDS LASSO	Post-orthogonalized LASSO

$\hat{\gamma}$ in Column 1 provides the unmatched average difference between log GVA of firms inside and outside cluster. $\hat{\gamma}$ in Column 2 provides regression estimates of the impact of clusters, after including all control variables mentioned in Section "Covariates". $\hat{\gamma}$ in Column 3 provides ATE where controls are selected as per Belloni et al. (2013). $\hat{\gamma}$ in Column 4 provides ATE as suggested by Chernozhukov et al. (2015). *,**, *** denote significance at 10%, 5% and 1% level, respectively. Standard errors are clustered at the district level

level dummy variables are purged in the selection equation. Overall, firm characteristics are not a strong predictor of being present in a cluster.

Treatment 2: Collaboration

For this treatment indicator as well, LASSO purges most of the variables. Further, the algorithm selects only three district level dummy variables, which are mostly the metropolitan districts. This indicates higher propensity for collaboration for MSMEs in richer areas.

Outcome Variable: Log of GVA

Given that the outcome variable is indicative of size of the firm, LASSO selects several close correlates from the superset of the covariates. These variables reflect the geographic, industrial, organizational and entrepreneur-specific heterogeneity in firm size. Among industrial classification, we find the following ones chosen—mining, manufacture of coke and refined petroleum, manufacturing of chemical products, manufacturing of rubber and plastic products, other non-metallic products, basic metals, motor vehicle, wholesale trade and land transport.

I also find that the interactions of districts dummy and nature of the enterprise (manufacturing, service and trade) survive the covariate selection exercise. Many of these interaction dummy variables belong to bigger states such as Rajasthan and Uttar Pradesh and manufacturing industries.

ATE of Clusters

Table 3 provides the estimate of Average Treatment Effect for treatment group 1. Column 1 reports the uncontrolled differences of means between firms in treatment group 1 and control group. The average GVA for a firm inside a cluster is nearly 70% higher. Column 2 regresses log GVA of all firms on the indicator for cluster along with all the controls described in Sect. "Covariates". Now, the coefficient on treatment indicator reduces to 0.18 with a standard error of 0.070. Thus, conditional on all the observable covariates, firms inside a cluster are nearly 18% more productive.

Table 4 ATE of collaboration on log GVA of firms

	(1)	(2)	(3)	(4)
$\gamma\hat{}$	0.570*** (0.006)	0.146*** (0.039)	0.319*** (0.059)	0.324*** (0.059)
Estimation	NA	OLS	PDS LASSO	Post-orthogonalized LASSO

$\gamma\hat{}$ in Column 1 provides the unmatched average difference between log GVA of firms with and without collaboration. $\gamma\hat{}$ in Column 2 provides regression estimates of the impact of clusters, after including all control variables mentioned in Section "Covariates". $\gamma\hat{}$ in Column 3 provides ATE where controls are selected as per Belloni et al. (2013). $\gamma\hat{}$ in Column 4 provides ATE as suggested by Chernozhukov et al. (2015). *,**, *** denote significance at 10%, 5% and 1% level, respectively. Standard errors are clustered at the district level

To address the bias in column 1 and overfitting in column 2, I use the data driven techniques described in Belloni et al. (2013) and Chernozhukov et al. (2015). In column 3, I report the post-double selection LASSO estimator. The coefficient on treatment indicator now is 0.306 with a standard error of 0.087. Post-LASSO orthogonalized estimator, reported in column 4, is 0.301 with a standard error of 0.087. The Average Treatment Effect on GVA of a firm locating in a cluster is nearly 30%. The full list of selected covariates and their statistical significance is available in Table 5.

ATE of Collaboration

Table 4 reports the estimates of ATE for treatment 2. Uncontrolled mean difference between log GVA for firms in treatment group 2 and control group is 0.57; firms with collaboration have a 57% higher GVA. Column 2 provides the results of a regression of log GVA on indicator for treatment group 2 along with all the controls. Now, the coefficient on treatment group indicator reduces to 0.14 with a standard error of 0.039. Thus, controlling for all the covariates, a firm with collaboration for technical know-how has nearly 14% higher GVA.

Estimates in column 2 of Table 4 stem out of a naive model selection method where all the observable covariates are included in the regression. Columns 3 and 4 report ATEs after employing model selection methods proposed under Belloni et al. (2013) and Chernozhukov et al. (2015), respectively. The ATE in column 3 is 0.324 and is significant at 1% level. Similarly, the ATE in column 4 is 0.319 with a standard error of 0.059. Thus, a data- driven model-selection exercise suggests that a firm observes a 32% increase in its GVA after entering in a collaboration Table 5.

Conclusion

In this paper, I provide estimates of the GVA gains for a firm if it moves to a cluster or enters a collaboration for technical know-how. A firm inside an industrial cluster observes an increase in GVA of nearly 30%. Similarly, a firm with collaboration experiences a growth of almost 32% in its GVA. These estimates are higher

than the conventional OLS estimates from a regression controlling for all observable covariates.

Overall assessment of treatment would require not just the value but the costs of the treatment. Building a cluster can be expensive. Clusters may also impose negative externalities on surrounding areas such as congestion or pollution (de Groot et al. 2009; Busso et al. 2013). Sampat (2008) records the displacement and violations of democratic procedures in establishing SEZs in India. On the other hand, inter-firm collaboration can have several advantages apart from private knowledge between firms. For eg; Nieto Jesus and Santamaria (2006) show how technical collaboration can push the innovation frontier for smaller firms. Thus, the overall benefit of clusters, accounting for negative externalities, may be lower compared to inter-firm collaboration.

The paper also addresses two important questions in the literature on persistent performance differences across firms (Gibbons and Henderson 2014)—what determines this productivity spread, and how can it dissipate across firms? Both the treatments considered here create an impact on GVA of firms. However, the benefits of clusters are local to the region and cannot be transmitted out to other firms. The advantages of being in collaboration, on the other hand, need not require such location specificity.

How can policy-makers encourage collaboration between firms? A beneficial partnership requires an incentive-compatible contract to ensure the needs of both parties. In environments with poor contract enforcement mechanisms and slow judicial processes, such as India, contracting imposes a high cost of uncertainty. Evidence for this comes from Rao (2019), which finds the detrimental impact of the slow judicial process on the formal sector firm growth in India. Scott-kemmis and Bell (1985) also demonstrate this in Indo-British collaborations where suppliers were unwilling to provide complete technology transfer. Thus, the benefits of collaboration can be potentially higher with effective legal and regulatory safeguards in place, with credible punishment on contract reneging. Legal reforms will assure that larger firms to collaborate with smaller partners.

Appendix

(See Table 5).

Table 5 Impact of clusters on Log GVA with selected covariates (post double selection method)

	Coefficient	Standard Error
I(Cluster)	0.306441	0.087502
Mining	0.768344	0.075309
Manufacture of paper products	0.554136	0.049535
Manufacture of coke and refined petroleum products	0.599035	0.083926
Manufacture of chemical products	0.228998	0.054214
Manufacture of rubber and plastic products	0.208249	0.046294
Manufacture of other non-metallic products	0.525929	0.048304
Manufacture of basic metals	0.851873	0.070218
Manufacture of motor vehicles	0.432815	0.081184
Wholesale trade and commission trade	− 0.13467	0.044529
Land transport	1.552294	0.105672
Electricity	− 0.16511	0.10094
Proprietary	− 0.18819	0.080888
Partnership	1.451599	0.091225
Private company	2.082758	0.155082
Public limited	1.913049	0.153612
Cooperatives	0.711551	0.135016
I(Dibang valley)	0.196893	0.043535
I(Purba midnapur)	− 0.24666	0.056261
I(Ratnagiri)	1.12383	0.048263
DelhiNorthWest*manufacturing	− 0.10028	0.082542
Dungarpur*manufacturing	0.005933	0.062494
Chittaurgarh*manufacturing	− 0.02752	0.058366
Kheri*manufacturing	− 0.18363	0.052259
Hardoi*manufacturing	− 0.1905	0.049947
Ghazipur*manufacturing	− 0.02536	0.03771
PurbaChamparan*trade	0.004969	0.034642
Sheikpura*manufacturing	− 0.11683	0.077019
Gaya*manufacturing	− 0.05969	0.06441
Haora*manufacturing	− 0.34281	0.045663
Garhwa*manufacturing	0.00434	0.042469
Ranchi*service	− 0.06456	0.105833
Gajapati*service	− 0.09287	0.055096
Janjgir-Chapma*service	− 0.07268	0.060716
Ranandgaon*service	− 0.00869	0.063564
Gwalior*service	− 0.01414	0.061161
Narsimhapur*service	− 0.01518	0.066414
Dindori*service	− 0.03096	0.056868
Amravati*trade	− 0.26571	0.032381
Osmanabad*manufacturing	− 0.66382	0.058385
Bagalkot*trade	− 0.07566	0.074349
Bagalkot*manufacturing	− 0.0494	0.072756

Table 5 (continued)

	Coefficient	Standard Error
Chikmaglur*trade	0.02132	0.065421
Thiruvananthpuram*service	0.009041	0.080726
Tiruvanmallai*trade	1.587233	0.056395
Karur*manufacturing	0.011563	0.049477
Karur*trade	− 0.0313	0.074527
Ariyalur*mnufacturing	− 0.01244	0.043991
Ariyalur*trade	− 0.08672	0.081144

The table shows coefficients selected from the Post Double Selection LASSO method. The corresponding table for the impact on collaboration is available on request

Data availability The data is publicly available. The data processed by the author is available on request.

References

Acs, Z.J., D.B. Audretsch, and E.E. Lehmann. 2013. The knowledge spillover theory of entrepreneurship. *Small Business Economics* 41 (4): 757–774.

Aghion, P., R. Burgess, S.J. Redding, and F. Zilibotti. 2008. The unequal e ects of liberalization: Evidence from dismantling the license raj in India. *American Economic Review* 98 (4): 1397–1412.

Angrist, J.D., and J.S. Pischke. 2008. *Mostly harmless econometrics: an empiricist's companion. Mostly harmless econometrics an empiricist's companion*. Princeton University Press,.

Bargagli-Stoffi, F.J., J. Niederreiter, and M. Riccaboni. 2020. Supervised Learning for the Prediction of Firm Dynamics. In *Data Science for Economics and Finance: Methodologies and Applications, number September*. Springer.

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

Besley, T., and R. Burgess. 2004. "Can labor regulation hinder economic performance? Evidence from India." *The Quarterly journal of economics* 119.1: 91–134.

Bhatt, A., M. Puri, and A. Appolloni. 2012. "SEZ proliferation in India: are the objectives being realized?." *Transition Studies Review* 19: 79–88.

Biswas, S., Roy, S., and Seshagiri, S. (2007). Collaboration in Indian SME Clusters: A Case Study. In: The 3rd International Conference on Communities and Technologies, pages 28 30.

Blois, K. 1972. Vertical quasi-integration. *The Journal of Industrial Economics* 20 (3): 253–272.

Bloom, N., M. Draca, and J. Van Reenen. 2016. Trade induced technical change? The impact of Chinese imports on innovation, IT and productivity. *The Review of Economic Studies* 83 (1): 87–117.

Bradach, J. and Eccles, R. (1989). Markets versus hierarchies: from ideal types to plural forms. *Annual Review of Sociology*, 15(1).

Brakman, S., and C. van Marrewijk. 2013. "Reflections on cluster policies." *Cambridge Journal of Regions, Economy and Society* 6.2: 217–231.

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

Ditillo, A., and A. Caglio. 2011. *Controlling Collaboration between Firms: How to build and Maintain Successful Relationships with External Partners*. Butterworth-Heinemann.

Calia, P., I. D'Attoma, and S. Pacei. 2016. Do Innovation and Offshoring Make a Differ- ence? An Empirical Exploration of the Effects on the Performance of European Firms. *International Business Research* 9 (7): 98.

Chernozhukov, V., C. Hansen, and M. Spindler. 2015. Post selection and post regulariza- tion inference in linear models with many controls and instruments. *American Economic Review* 105 (5): 486–490.

Coad, A., and S. Srhoj. 2019. Catching Gazelles with a Lasso: Big data techniques for the prediction of high-growth firms. *Small Business Economics* 55 (541): 565.
Combes, P.P., and L. Gobillon. 2014. The empirics of agglomeration economies. *SSRN Journal*. https://doi.org/10.2139/ssrn.2505370.
de Groot, H.L., J. Poot, and J.M. Smit. 2009. New Zealand Agglomeration externalities, innovation and regional growth : theoretical perspectives and meta-analysis. *Handbook of Regional Growth and Development Theories, Pages* 256: 281.
Dehejia, R. H., and S. Wahba. 2002. "Propensity score-matching methods for nonexperimental causal studies." *Review of Economics and statistics* 84.1: 151–161.
Duranton, G., and D. Puga. 2004. "Micro-foundations of urban agglomeration economies." *Handbook of regional and urban economics*. Vol. 4. Elsevier. 2063–2117.
Dyer, J.H. 1998. E ective Inter rm collaboration: how firms minimize transaction costs and maximize transaction value. *Strategic Management Journal* 18 (May): 1998.
Evan, W. M. (1966). The organization-set: Toward a theory of interorganizational relations. *Approaches to organizational design*, pages 173–191.
Ge, W. 1999. *Special economic zones and the economic transition in China*. World Scientific.
Ghio, N., M. Guerini, E.E. Lehmann, and C. Rossi-Lamastra. 2015. The emergence of the knowledge spillover theory of entrepreneurship. *Small Business Economics* 44 (1): 8.
Gibbons, R., and R. Henderson. 2012. "Relational contracts and organizational capabilities." *Organization science* 23.5: 1350–1364.
Grandori, A., and G. Soda. 1995. Inter- rm networks: antecedents, mechanisms and forms. *Organization Studies* 16 (2): 183–214.
Granovetter, M. 1983. The strength of weak ties: A network theory revisited. *Socia- Logical Theory* 1 (1983): 201–233.
Granovetter, M. 1985. Economic Action and Social Structure: the Problem of Embeddedness. *Journal of Economic Sociology* 3 (3): 44–58.
Greenstone, M., J. A. List., and C. Syverson. 2012. The effects of environmental regulation on the competitiveness of US manufacturing. No. w18392. National Bureau of Economic Research.
Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning*, vol. 27. New York: Springer.
Heckman, J.J., H. Ichimura, and P.E. Todd. 1997. Matching as an econometric eval- uation estimator: evidence from evaluating a job training programme. *The Review of Economic Studies* 64 (4): 605–654.
Heckman, J.J., H. Ichimura, and P. Todd. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65 (2): 261–294.
Hortacsu, A., and C. Syverson. 2009. Why do firms own production chains? *US Census Bureau Center for Economic Studies, Pages* 9: 31.
Jacobs, J. (1969).. The Economy of Cities. Vintage, New York
Hsieh, C.-T., and P.J. Klenow. 2009. Misallocation and Manufacturing TFP in China and India. *Quarterly Journal of Economics* 12 (4): 1403–1448.
Hsieh, C.-T., and B.A. Olken. 2014. The Missing Missing Middle. *Journal of Economic Perspectives* 28 (3): 89–108.
Johnson, J. (1999). O sets in the international marketplace: an aerospace industry view. Trends and challenges in aerospace o sets.
King, G., and R. Nielsen. 2019. Why propensity scores should not be used for matching. *Political Analysis* 27 (4): 435–454.
Konstadakopulos, D. 2000. Learning behaviour and co-operation of small high technology firms in the ASEAN region: some evidence from the Singapore-Johor agglomeration. *Asean Economic Bulletin* 17 (1): 48–59.
Lafontaine, F., and K. Shaw. 2016. "Serial entrepreneurship: Learning by doing?." *Journal of Labor Economics* 34.S2: S217–S254.
Lechner, M. 2005. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. *SSRN Electronic Journal, Pages* 43: 58.
Leong, C.K. 2013. Special economic zones and growth in China and India: an empirical investigation. *International Economics and Economic Policy* 10 (4): 549–567.
Levina, N., and E. Vaast. 2008. Innovating or doing as Told? Status differences and overlapping boundaries in offshore collaboration. *MIS Quarterly* 32 (2): 307–332.
Levinsohn, J., and A. Petrin. 2003. Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*. https://doi.org/10.1111/1467-937X.00246.

Lopez, M.J., and R. Gutman. 2017. Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science* 32 (3): 432–454.
Lucena-Piquero, D., and J. Vicente. 2019. The visible hand of cluster policy makers: An analysis of Aerospace Valley (2006–2015) using a place-based network methodology. *Research Policy* 48 (3): 830–842.
Nadvi, K. (1996). Small firm industrial districts in Pakistan. PhD thesis, University of Sussex.
Neffke, F., M. Svensson Henning, R. Boschma, K.-J. Lundquist, and L.-O. Olander. 2009. The dynamics of agglomeration externalities along the life cycle of industries. *Regional Studies* 12: 1537–1546.
Neumark, D., and H. Simpson. 2015. "Place-based policies." *Handbook of regional and urban economics*. Vol. 5. Elsevier, 1197–1287.
Nieto, M. J., and L. Santamaría. 2010. "Technological collaboration: Bridging the innovation gap between small and large firms." *Journal of small business management* 48.1: 44–69.
Okada, A., and N.S. Siddharthan. 2008. Automobile clusters in India: evidence from chennai and the national capital region. *The Flowchart Approach to Industrial Cluster Policy, Pages* 109: 144.
Olley, S., Pakes, A., 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64:1263–1298
Palit, A., and S. Bhattacharjee. 2008. *Special economic zones in India: myths and realities*. Anthem Press.
Petrin, A., and J. Sivadasan. 2013. Estimating lost output from allocative inefficiency, with an application to chile and firing costs. *Review of Economics and Statistics* 95 (1): 286–301.
Powell, W.W. 1987. Hybrid organizational arrangements: new form or transitional development. *California Management Review* 30 (1): 67–87.
Powell, W.W. 1990. Neither market nor hiearchy: network forms of organization. *Reseach in Organizational Behavior* 12 (1990): 295–336.
Qian, H., and Z.J. Acs. 2013. An absorptive capacity theory of knowledge spillover en- trepreneurship. *Small Business Economics* 40 (2): 185–197.
Rao, P. 1971. Some notes on misspeci cation in multiple regressions. *The American Statistician* 25 (5): 37–39.
Rao, M. 2019. Judges, lenders, and the bottom line: Court-ing firm growth in india. working paper.
Rasiah, R. and Govindaraju, C. (2009). University-Industry R&D Collaboration in the Automotive, Biotechnology and Electronics Firm in Malaysia. In; 7th Globelics Conference, (October):6 8.
Richardson, G. B. 2003. "The organization of industry re-visited." *DRUID Summer Conference*.
Rosenbaum, P.R., and D.B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (41): 55.
Rosenbaum, P.R., and D.B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. https://doi.org/10.2307/2683903.
Sampat, P. 2008. Special economic zones in India. *Economic and Political Weekly* 43 (28): 25–27.
Schilling, M.A., and C.C. Phelps. 2007. Inter rm collaboration networks: the impact of large-scale network structure on firm innovation. *Management Science* 53 (7): 1113–1126.
Schmitz, H. 1995. Collective e ciency: growth path for small-scale industry. *The Journal of Development Studies* 31 (4): 529–566.
Schmitz, H., and K. Nadvi. 1999. Clustering and industrialization: introduction. *World Development* 27 (9): 1503–1514.
Scott-kemmis, D. and Bell, M. (1985). Technological Dynamism and Technological Content of Collaboration: Are Indian Firms and TechnolAre Indian Firms Missing Opportunities? *Economic and Political Weekly*, 20(45).
Shortreed, S.M., and A. Ertefaie. 2017. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* 73 (4): 1111–1122.
Storper, M., and A.J. Venables. 2004. Buzz : face-to-face contact and the urban economy. *Journal of Economic Geography* 4 (4): 351–370.
Syverson, C. 2004. Market structure and productivity: A concrete example. *Journal of Political Economy* 112 (6): 1181–1222.
Syverson, C. 2011. What determines productivity? *Journal of Economic Literature* 49 (2): 326–365.
Taylor, T. (2004). Using procurement o sets as an economic development strategy. In: Brauer, J. and Dunne, P., editors, Arms Trade and Economic Development: Theory, Policy and Cases in Arms Trade O sets, chapter 2, pages 34 47. Routledge.

Teece, D. 1980. Economies of scope and the scope of the enterprise. *Journal of Economic Behavior and Organization* 1 (3): 223–247.
Teece, D.J. 1986. Pro ting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy* 15 (6): 11–46.
Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical. Society Series B* 58 (1): 267–288.
Tybout, J. 2000. Firms manufacturing in developing countries : how well they do, and why. *Journal of Economic Literature* 38 (1): 11–44.
Unger, J.M., A. Rauch, M. Frese, and N. Rosenbusch. 2013. Human capital and En- trepreneurial success: a meta-analytical review. *Journal of Business Venturing* 26: 341–358.
Venkataramanaiah, S., and S. Parashar. 2007. Enhancing the competitiveness of SMEs through industrial clusters: The Indian experience. *International Journal of Technology Management and Sustainable Development* 6 (3): 164–174.
Whitley, R.D. 1991. The social construction of business systems in East Asia. *Organization Studies* 12 (1): 1–28.
Williamson, O. 1981. The economics of organization: the transaction cost approach. *The American Journal of Sociology* 87 (3): 548–577.
Williamson, O.E. 1991. Comparative economic organization : the analysis of discrete structural alternatives. *Administrative Science Quarterly* 36 (2): 269–296.
WorldBank. 2020. *Doing Business 2020 Technical report*. Washington DC: World Bank.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.