Why do poor countries make simple products?

Søren Post

Bachelor of Science Programme in Development Studies
SGED10
Supervisor: Karl-Johan Lundquist

LUND UNIVERSITY

Department of Human Geography
Lund University
Sweden
May/2020

Why do poor countries make simple products?

Søren Post

Abstract

The products produced in rich countries are very different from the products produced in poor countries. Why? Unreliable electricity is often cited as one of the main challenges facing industrial production in less developed economies. In this paper, I connect these two observations by linking interruptions in the production enviroment to the sophistication of the production output. Using a data set covering more than 500,000 observations of manufacturing plants in India between 2000 and 2016, I find evidence suggesting that the level of interruption, modeled here using electricity shortages, contained in a plant's input supply is strongly associated with whether or not more complex products have a positive relationship with plant revenues. This suggests a new pathway between the production environment at the local level and the aggregate complexity of the economy.

Keywords: manufacturing, economic complexity, input-output, electricity

Word count (excl. tables, references): 14146

Contents

List of Figures

List of Tables

1 Introduction

The products produced in rich countries are very different from the products produced in poor countries. Why? There is a growing empirical literature showing that the difference between countries' economic sophistication, defined by the kinds of products they produce, explain large variation in GDP per capita (see figure 5). However, very little empirical evidence examines the micro-foundations behind these differences.

From the literature on economic complexity and product upgrading, we know that the productive capabilities in an economy is an important predictor of economic growth (Hausmann and Hidalgo, 2011). However, despite open markets and vast access to information, developing economy uptake to more sophisticated production have been a slow and path dependent process (Hidalgo et al., 2007).

At the same time, the access to reliable electricity is often highlighted as one of the key challenges facing industrial production in less developed economies. The approach taken to quantify this effect is usually based on production stoppages at the unit level (plants, factories or firms). In this paper, I propose a different pathway between a more disruptive production environment and the level of sophistication in the economy. Specifically, I argue that only taking into account the effect of production failures at the plant level underestimates the true effect of a disruptive environment on production networks and incentives.

The mechanism is simple. If more complex products require more intermediate inputs or more steps in production, interruptions or failures in production processes punish them harder than simpler ones. If these disruption-costs are punitive enough, investors will put their money elsewhere and producers will choose less complex products. Given that electricity is an important input in manufacturing - most factories cannot produce anything without running lights, machines, and motors - an unreliably supply could significantly reduce the productivity of a plant. To test for this effect, I examine the ability of electricity interruptions in the production environment to explain the complexity of products made at the plant level of in India between 2000 and 2016.

The approach taken in this paper is primarily related to three strands of literature. First, the literature of economic complexity and the sophistication of an economy's capability base (Frenken et al., 2007; Hausmann et al., 2013; Tacchella et al., 2012). In addition, to develop the relation between production disruptions and product complexity, I draw on O-ring-type effects as modelled in Kremer (1993) and Jones (2011), and the recent literature on volatility in production networks (Acemoglu et al., 2012). While the latter is mainly concerned with aggregate effect of sectoral shocks, the importance of input-output linkages are equally relevant at the plant level.

1.1 Aim and research questions

The objective of this research is to explore the relationship between interruptions in the supply of electricity, here taking the role of the more general production interruptions, and the sophistication of manufacturing production.

More specifically, I'm interested in whether an uncertain production environment can help explain the difference in economic complexity observed at the factory level. To analyse this effect, I put together an extensive set of data on electricity-, state- and plant-level

indicators covering seventeen years across thirty Indian states. Two overall questions guides the research:

1. How is the impact of production interruptions - in the form of electricity shortages - on factory output related to the sophistication of the factory's production?
2. What relationship, if any, does this association between complexity and interruptions have with the complexity of the manufacturing sector in India in the long run?

The first question explores how electricity shortages interact with production sophistication at the plant level. The second question address the impact of this interaction, as it expresses itself in the type of products being made in the economy. In section 3.3 I turn the two research questions into a set of testable hypotheses.

1.2 Structure of thesis

In the following, I first briefly present the background of the widespread electricity shortages in India and its relationship to the manufacturing sector. Second, I develop the theoretical connection between interruptions in production chains, sophistication of economic activity, and electricity shortages. From this model, I construct a set of hypothesis to test against the empirical data. In section 4 I present the data sources used in the analysis. Section 5 presents the construction and operationalisation of my key variables. Here, I also detail my empirical strategy. Section 6 and 7 presents the results of the analysis and discusses the findings. Finally, I conclude.

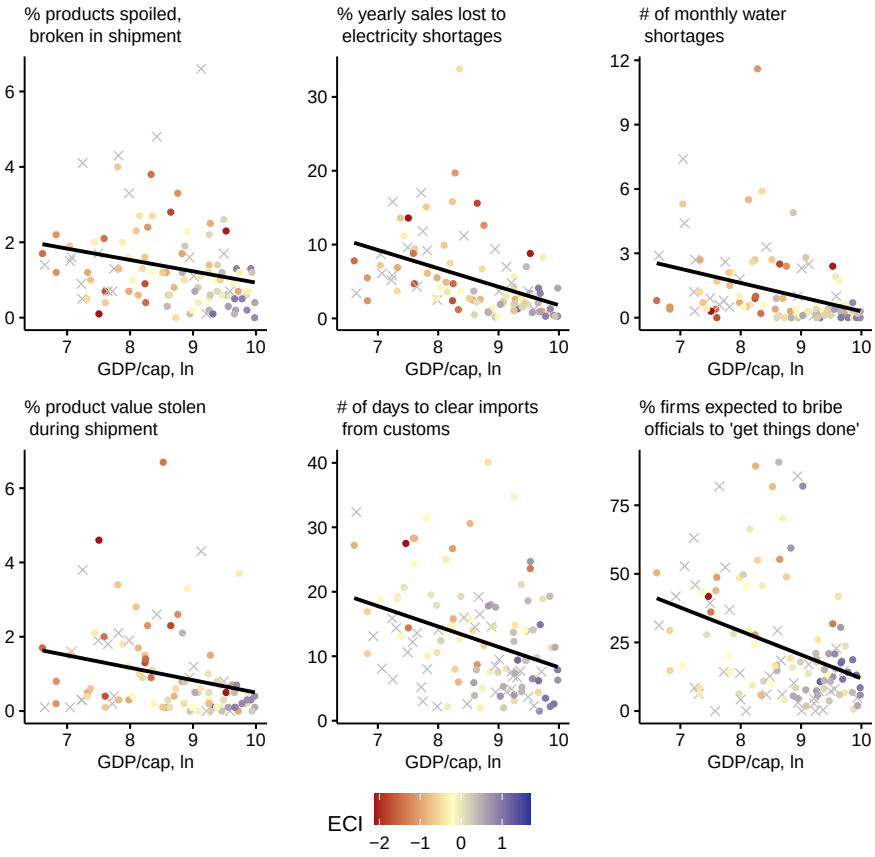


Figure 1: Production disruptions tend to have a higher frequency in poorer, less complex economies. The color of points are country economic complexity index values (ECI), where blue is more complex. Gray crosses has a missing complexity value. Lines are SLR fits to ln(GDP/cap). Disruption data is from World Bank (2020c). ECI data is from Simoes and Hidalgo (2011). Countries are not surveyed the same years, so complexity and GDP/cap values have been matched to the newest survey year for each country before 2017. No country has more than one observation.

2 Background

In the last few years, India has made enormous strides in expanding the quality of the electricity supply. Since 2018, the electricity grid has reached 100% coverage of villages and the distance between the average deficit of required power has been reduced to a less than a percent from an average of 8.5% in 2012. This has in large parts been accomplished by a massive expansion in the generative ability of the power sector: between 2007 and 2017 the capacity more than doubled (Zhang, 2018). Until recently, however, the electricity sector was plagued by poor market incentives, large technical losses and a rapidly increasing demand. This has led to endemic shortages during the past decades.

There are several reasons behind the persistent shortages. Despite the fact the electricity has been open to private enterprises since the 1991 liberalization, state- and government run companies still accounted for 80% of electricity generation in 2010 (split 51% to 29%), as well as most retail distribution companies. For decades, state-run distribution companies have provided fixed-fee electricity provision to agricultural consumers. These un-meetered prices have then been partly paid for by electricity prices for industrial use at almost four times the price (Allcott et al., 2016). For many years, this has led to a "quality-subsidy" trap, where distributors provide poor-quality electricity, consumers accept this because of the low price and the public underwrites the losses of the distribution companies (McRae, 2015).

Despite the government subsidies, distribution companies have run with large yearly deficits: between 1992 and 2010, such companies reported a loss of a 61 billion 2004-USD dollars in total (Allcott et al., 2016). This has led to large underinvestment in the sector. An example of such underinvestment is the "understanding" signed by investors after liberalization in 1991 to build 50 gigawatts of power-generation capacity. Of these 50, 4 was build. Similarly, between 1997 and 2007, only half of the 71 gigawatts capacity planned for construction was realized (CEA, 2013). In addition to the under-capacity comes a large amount of technical losses. For instance, between 1994 and 2009, Indian thermal plants were offline about 28% of the time due to forced outages, planned maintenance, and shortages in coal.

In the same period, the size of the Indian economy more than doubled and the population added some 250 million people. Together, the inability for companies to clear the market, underinvestment, rapid population growth, and massive increases in the size of the economy, this led to wide spread shortages during the study period (2000-2016).

These shortages provide significant barriers to the daily operations of companies. Figure 2 shows the results from two Enterprise Surveys conducted by the World Bank (World Bank, 2020a,b). In 2005, electricity is listed as by far the greatest obstacle. While electricity has been overtaken by corruption as the main concern of companies, more than 30% still perceive that access to power is at least a major obstacle to their operations in 2014.

Such industry-agriculture pricing distortions and lack of reliable power is bound to cause production frictions. The importance of electricity in production inefficiencies have recently been documented in several settings at the level of the individual firm (Grainger and Zhang, 2017; Abeberese et al., 2019; Fisher-Vanden et al., 2015).

Two India-specific patterns are worth highlighting here: the importance of the quality of electricity, rather than just electricity alone, and the evidence on output-costs of electricity shortages. Samad and Zhang (2016) and Chakravorty et al. (2014) both find improvements

in incomes from access to electricity, but much larger income gains from access to quality electricity (from 9.6% to 17% and from 9% to 28.6%)[1]. Abeberese (2017) finds that increases in electricity prices reduces the electricity- and machine-intensity of activities, leading to smaller growth in production output. Finally, Allcott et al. (2016) finds a 5% - 10% reduction in plant-level profits from shortages in electricity, with a lower productivity-penalty due to generator substitution.

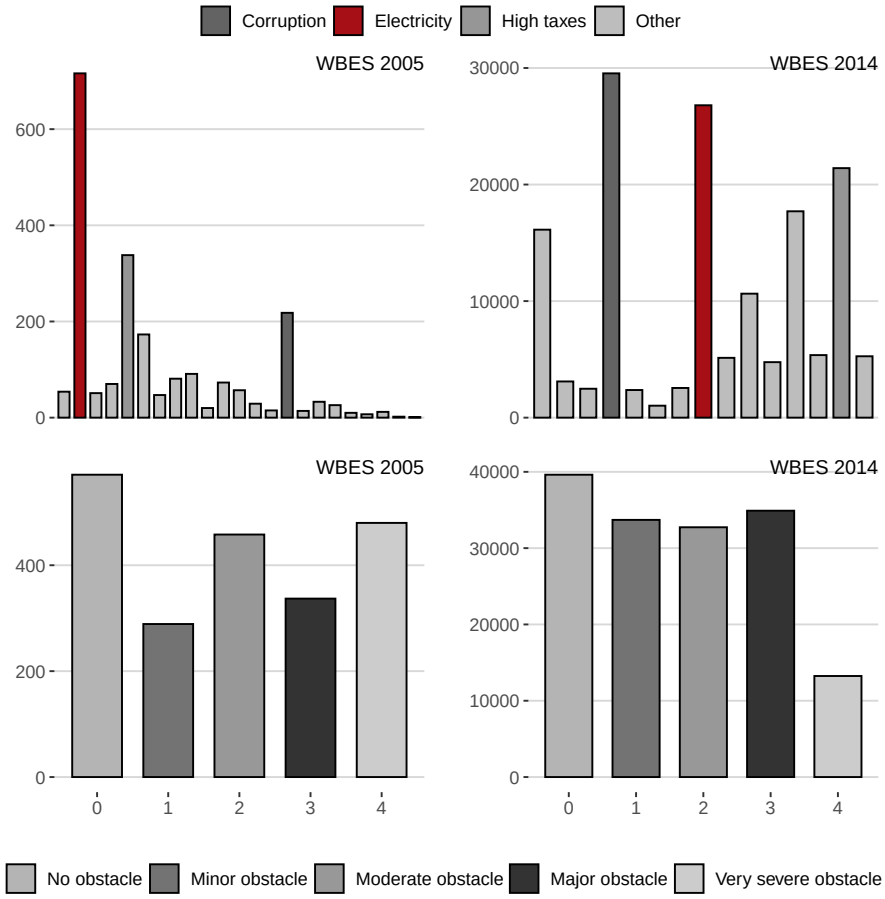


Figure 2: Electricity remains an important concern for firms in India. Top figures: while the share of managers that list electricity quality as the main constraint to their success has fallen between 2005 and 2014, it remains high. Bottom figures: the share of firms that name electricity as a major obstacle, or worse, to their current operations is still around 30 %. Source: World Bank enterprise surveys in India: 2005 and 2014 (World Bank, 2020b,a). Vertical axis is firm counts for 2005 and weighted counts for 2014.

[1]The latter study only non-agricultural incomes.

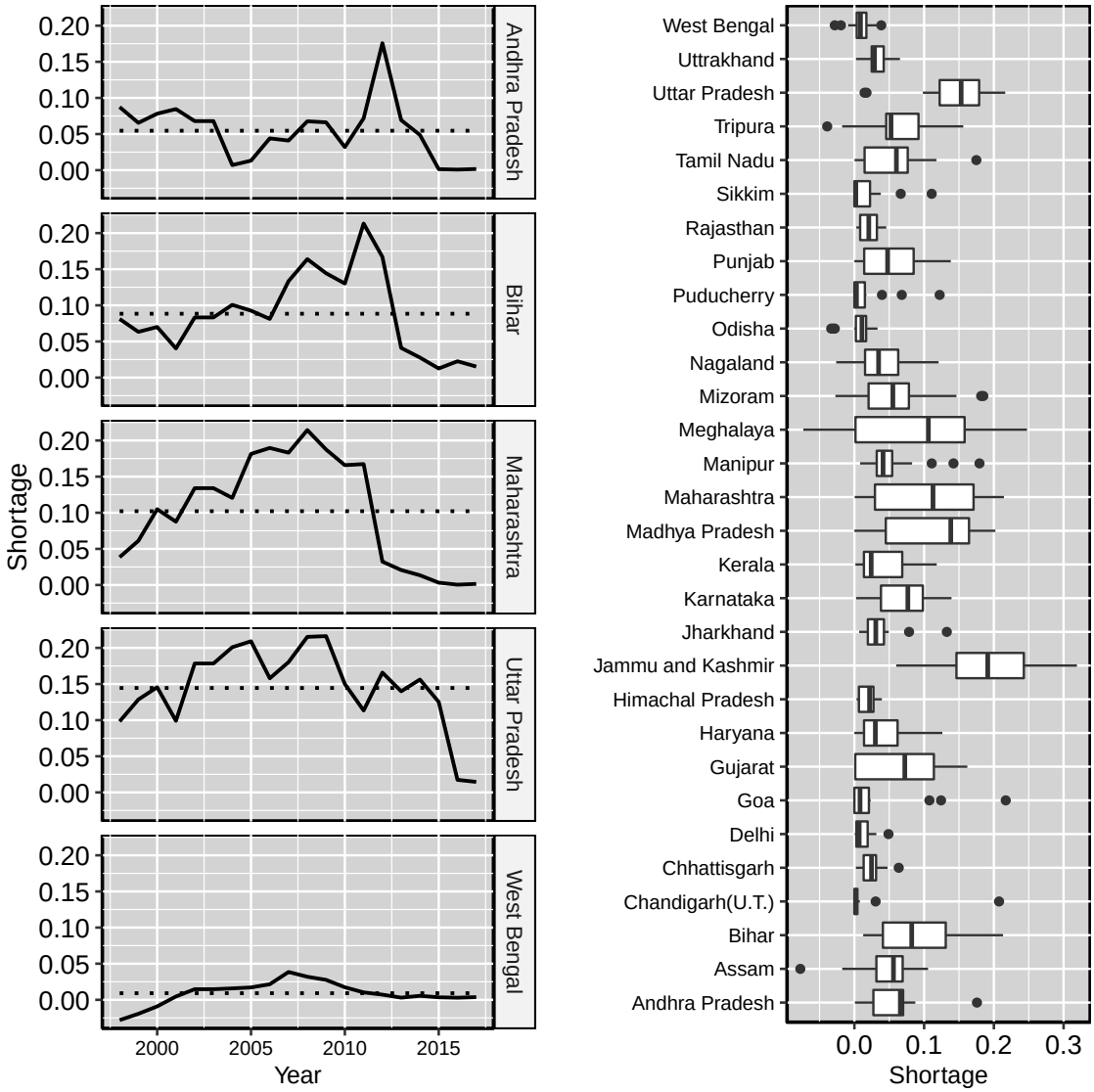


Figure 3: There is a significant variation between the shortage of states in a given year, as well as within states across time. On the left is the yearly average shortage in the five most populous states in 2001. The dotted line is period average for each state. On the right is the distribution of yearly shortages for all states.

3 Literature and theoretical framework

3.1 Product complexity

Since Adam Smith it has been a truism that wealth comes from the economic efficiency of division of labor. The greater the market available, the deeper its participants can specialize and the greater the benefit. This suggests that economic wealth is connected to the increasing number of activities and complexity of interactions in the economy (Romer, 1990).

If the size of the market limits the specialization of firms and workers, the globalization of labor- and input-markets should facilitate broad economic wealth creation. When all countries can exploit the global markets, why then have national differences in the gross domestic product (GDP) per capita skyrocketed during the last two hundred years (Pritchett, 1997)? Despite 50 years of unprecedented international connectivity, international trade, and globalisation (and some notable growth spurts), the data show that developing countries (as a group) are not catching up to more advanced economies (Johnson and Papageorgiou, 2020).

The literature on economic complexity provides one possible answer. If some spill-over effects from the individual activities that arise from specialization - like property rights, tacit know-how, infrastructure, regulation - cannot be imported, they need to be present in the local economy. The productivity of a country then lies in these non-tradable "economic capabilities", and the differences between countries owe (partly) to their number, the complimentarity, and the interactions of these capabilities (Hidalgo et al., 2007; Hausmann et al., 2013).

There are competing methods (Tacchella et al., 2012; Hidalgo and Hausmann, 2009; Inoua, 2016), but approaches to quantifying these capabilities share a common conceptual grounding. Given the difficulties in defining and measuring discrete capabilities, researchers have taken an agnostic approach to specific nature of capabilities. The basic intuition is simple. Say that a set of capabilities are required to effectively produce a product. We can assume that a country that effectively makes the given product posses the necessary capability base. It follows then that products that are produced by many countries requires less rare- or non-tradable capabilities, while rarer products require more complex capabilities. Some products, however, will happen to be present in only a few places for reasons unrelated to the abilities of the economy (diamonds, ostrich eggs). This is solved by implementing an iterative algorithm that repeatedly weighs the complexity of products by the complexity of the countries that export them. See appendix B for a definition of the algorithm used in this paper.

This framework has proven to be a strong predictor of economic performance. Figure 5 shows the robust relationship between country-level economic complexity and GDP per capita (PPP). Since natural resources are a product of geographical luck rather than productive know-how, I separate out economies with more than 10% of resource rents as share of total GDP. Hausmann et al. (2013) shows how the deviations from the observed trend of economic complexity of economies and their GDP/cap is a strong predictor of economic growth, suggesting that they converge to the sophistication of their capabilities (that is, countries below the trend line growth fast, while countries above slows down). Not only does aggregate complexity matter: economies moving into more complex products are more egalitarian (Hartmann et al., 2017), are less carbon-intensive (Can and Gozgor, 2017), and

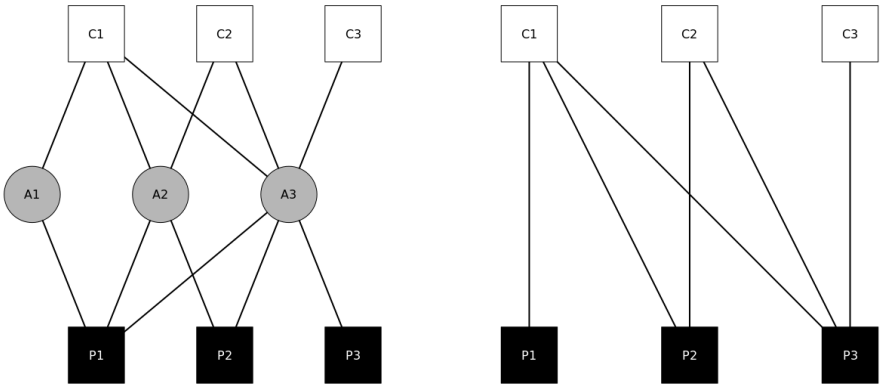


Figure 4: The tripartite graph (left) represents the theoretical model: countries (C) can make the products (P) their capabilities (A) allows them to. The bipartite graph (right) is what we observe in the trade data: countries export a set of products, and from this set of products, we infer their capabilities. For example, every country can produce product three. This suggests that the capabilities required to produce it are ubiquitous. In addition, we can see that the only product country three can make is the one every country produces. This suggest that country three does not have a sophisticated capability-base. In contrast, country one can produce all products including product one, which it is the only one that can produce. Here, country one and product one would the most complex.

have less volatile job-markets (Adam et al., 2019).

The aggregate-level economic complexity is the outcome of a myriad of micro-level decisions, historical conditions, firm decisions. Despite the fact that the economic complexity algorithm explicitly defines aggregate economic complexity as the collection of products' complexity, the factors that drive micro-level economic sophistication are not very well understood, and have seen very little empirical study.

3.2 Interruptions in production networks and economic complexity

I now turn to the relationship between my main variable of interest, the complexity of products at the plant level, and my main explanatory variable, unreliable electricity. Throughout the rest of the paper I use plant, factory, and firm interchangeably. Specifically, this section will argue that interruptions in production networks (such as unreliable electricity) can significantly reduce firms' incentive and ability to produce complex products.

To understand the role of disruptions on products, it is helpful to look at a simple model of production. Disruptions to a plant can happen in two ways: at the level of plant itself, or somewhere in the production chain. I start from the production setup in Acemoglu et al. (2012) used in much of the recent literature on shocks in aggregate production networks. For now, I take one plant to be representative of the production in a given sector, where each plant makes a unique product that can either be sold to consumers or used as intermediate input in the production of a different product. We can then model a multi-sector production by

$$x_i = (z_i l_i)^{\alpha} (\prod_{j=1}^{n} x_{i,j}^{w_{i,j}})^{1-\alpha}$$

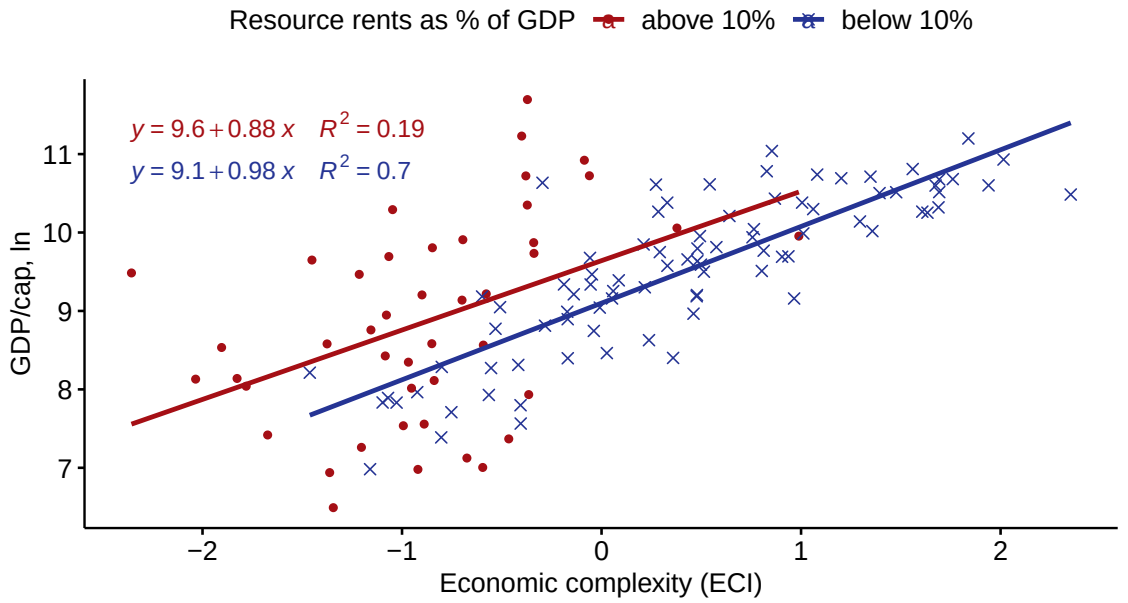


Figure 5: Simple linear best fits on ln(GDP/cap) (PPP, 2011 intl $) by ECI. Data on resource rents is from World Bank (2020e), GDP/cap is from World Bank (2020d) Observations are from 2010.

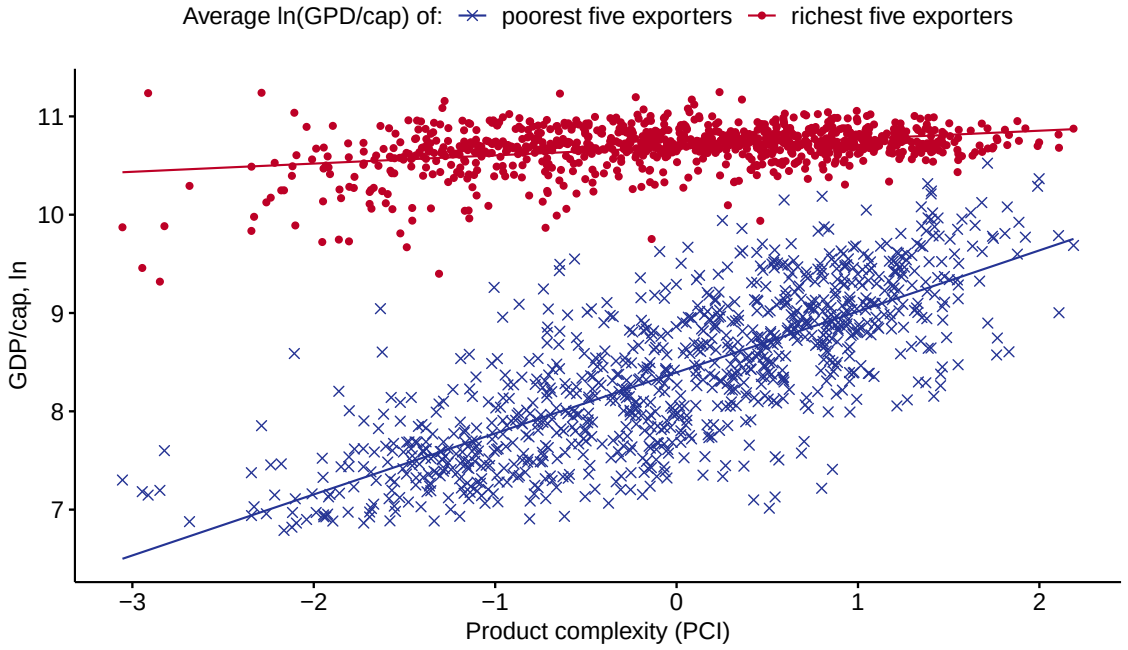


Figure 6: For each product observations are the average ln(GDP/cap) of the five richest (red) and five poorest (blue) significant exporters (countries). The triangular shape suggests an important facet of the distribution of products: while richer countries tend to export all kinds of products, poorer countries seem to face some threshold to compete in more complex products. To be a significant exporter, a country must export the product with a revealed comparative advantage ≥ 1 (Balassa, 1965). GDP/cap is from World Bank (2020d). Observations are from 2010.

where x_i is the output of plant i, l_i is the amount of labor hired by plant i and $\alpha \in (0,1)$ is the share of labor in production. z_i models some risk of production failure (or delay) due to exogenous factors (e.g. fire, theft, power outages, corruption). More accurately, z_i is one minus the risk of failure. In this paper, the interruption I test is unreliable electricity. $x_{i,j}$ is the amount of the output by plant j that is used as intermediate input in the production of x_i. $w_{i,j} \geq 0$ is the amount of good j in the total intermediate input used in the production of good i, and thus represents a sort of production recipe for plant i. I also take $\sum_j w_{i,j} = 1$, i.e. there are constant returns to scale.

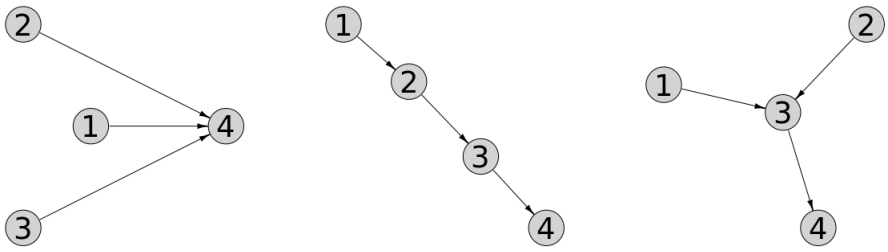


Figure 7: Three stylized input-output configurations in a four sector economy. Each node represents a plant or an economic sector. Arrows show supply relations. The left-most model thus have one plant (4) needing inputs from three other plants (1, 2, 3).

At the individual plant i, z_i is the only source of production failure in this model. However, it is important to note that each individual intermediate input ($x_{i,j}$ for $j = 1, 2, ..., n$) needed for production is also made at a plant, with its own risk of failure. Thus, a potentially large share of disruptions to plant i's production occur through disruptions in it its supplier network. The left-most graph in figure 7 shows this relationship: should plant one fail to deliver, this impacts the productivity of sector four.

The importance of a failure in one of a plant's suppliers depends on how substitutable intermediate inputs are. If the ability of plants to substitute between different types of intermediate inputs are close to 0, the expected output of a factory declines rapidly when n increases. In the simple case that the output of a factory is just the value of the intermediate inputs and the distribution of z is equal across factories, the relationship between expected output and the number of inputs n can be written as $n \cdot (z^n)$ and is identical to the O-ring problem in Kremer (1993)[2] (see figure 11).

Substitutability varies naturally between inputs, and specific elasticities of input and output losses is an ongoing research area (Brummitt et al., 2017; Carvalho, 2014). Much of the recent research on the propagation of shocks through production networks suggest that declines in input availability has a large effect on output for the consuming sector, at least in the short term. For instance, using a 2011 earthquake in Japan as an exogenous shock, Boehm et al. (2019) finds evidence of a near 1-to-1 ratio between input and output losses. This points to elasticities (ability to change one input for another) of intermediate inputs in manufacturing around 0.

This effect is supported by Barrot and Sauvagnat (2016) who find that firm-specific shocks propagate through production-networks for more specific inputs. If more complex production require a larger number of intermediate inputs or more specific (less substitutable) inputs, disruptions will punish the output of complex products more. Intuitively this makes

[2]Here it is assumed that failure in a production ruins the whole production for that period. That is, a fire cannot burn only half the production and electricity shortages cannot last just half the period. These assumptions are only relevant for the example, not the general argument.

sense: more parts goes into medical imaging equipment than baked goods. Similarly, lenses and microchips are highly specific, while cane sugar could feasibly replace beet sugar.

A second effect is modelled in the middle graph in figure 7. Here, supply-relationships depend completely on the output of the node at one stage earlier in the production process. This highlights an important mechanism. If a plant is dependent on the output of another plant, which is again dependent on intermediate inputs, and so on for n stages, then risks multiply in production chains. The effect of losses in the source sector has (decreasing) knock-on effects throughout the production chain[3]. If the outputs at each production stage increase in value, which is a reasonable assumption, failures in later stages are more expensive than in earlier ones.

Taken together, the two points highlighted thus far suggests an important path-way between the complexity of plants and interruptions in the production environment. We would expect more primary production (that is, less complex production) in more disruptive environments if A) more complex products are punished harder by supply-chain unreliability since these products have longer production chains, and B) longer production chains locate their later stages in more reliable environments. This corresponds well with the picture in figure 1.

Furthermore, since risks propagate through production inputs, the marginal returns to increases in primary inputs (like labor) are higher in high-disruption environments. The risks an individual plant faces scales with increases in the share of intermediate inputs $(1-\alpha)$ but not with increases in the share of labor (α). For instance, in the extreme case of $\alpha = 1$ no intermediate inputs are used and the risk to a plant i's production is fully contained in z_i.

The right most configuration in figure 7 depicts a configuration we would expect to meet in a real economy. Here, the output of node four depends directly on sector three, but with sector three's output depending on inputs from two sectors. A shock to sector one, for instance, would in this case be moderated both by sector three's relative dependence on outputs from node one (vs node two), by node three's connection between input and output, and by the degree to which sector four relies on inputs from sector three.

A key result in the seminal paper by Kremer (1993) is that under certain conditions (when quality is not a substitute for quantity and a production function that looks like the one outlined above) an economy will have a larger aggregate output by matching quality in inputs. Quality is Kremer's version of the z used in this paper. To illustrate the idea of matching, say that two firms make one product product each with same simple output function: output is just equal to the number of inputs, n. If production again has z risk of failing (and ruining the output), we can write the expected output as $n(z^n)$. Each product is similar in its use of inputs, and there are four suppliers available, two highly reliable z_h and two less so z_l. These can be matched or mixed in production. We thus have two ways of organizing the production: $2(z_h^2)+2(z_l^2)$ or $2(z_l z_h)+2(z_l z_h)$. If $z_h = 1$ and $z_l = 0.5$, the matched output is 2.5 and the mixed is 2. That is, using the same plants and the same

[3]Decreasing because not all the value of a production process comes from the input.

number of inputs we get a 25% higher outcome by matching[4].

This is an important result for explaining what kinds of products are made in an economy. If production chains match their risk of interruption, even very small differences in the risk of having some kind of production delay can have large effects on where investors and producers choose to place their production infrastructure. The large output penalty to even a few weak links in a production chain suggests that regions may need to reach a certain threshold of production reliability to enter into production of more complex goods. This interpretation is a possible explanation for the pattern seen in figure 6. This effect in turn opens the possibility of an S-curve style effect (see Brummitt et al. (2017) for a dynamic model of such an effect): as long as a certain floor of reliability is not met in the aggregate economy, even a few weak links limits the incentive for investments in more complex productions. This increases the relative marginal returns on producing more primary and simpler goods, which limits the necessity for fixing disruptions.

With regards to electricity-driven interruptions, a potentially important caveat is the ability of plants to invest in buffers against disruptions by purchasing a generator. However, generator electricity is more expensive and they depreciate over time. This means that substituting away from the central electricity supply imposes a kind of input-tax on the production[5]. In this vein, Abeberese (2017) finds that higher electricity prices leads to self-selection into less machine heavy and low-productivity activities. This suggests another pathway between economic complexity and electricity quality.

To summarize, I highlight three important interactions between interruptions and economic complexity at the production-level:

1. Production failures are more costly the longer down the production chain they occur. While some of this effect should be observed in the loss of plant-level revenues due to the higher price of inputs, this value loss will act more in the pre-production choice of where to place production, especially so for firms controlling a larger share of the whole value chain.

2. As the necessary interactions to produce a product increase, small increases in the risk of failure increases the expected cost of output losses exponentially. If the production process is subject to the constraint of very specific inputs, even very small increases in the unreliability of impose quantitatively large output penalties.

3. In terms of production networks, these interactions takes place primarily at the supply-use (or input-output) level. This means that in a typical production function, the effects above does not increase with the share of labor used in making products, but does increase with the reliance on intermediate inputs.

This means that if complex products are typically further down the chain of production, if they rely on a greater number of inputs, or if their inputs are more specific, I would expect

[4]For a proof that this is always true, we can follow Kremer (1993):

$$\begin{aligned}(z_h - z_l)^2 &> 0\\ z_h^2 + z_l^2 - 2z_h z_l &> 0\\ z_h^2 + z_l^2 &> 2z_h z_1\end{aligned}$$

[5]It also requires that suppliers upstream from the self-generating factory invest in generators. I don't test for "self-generating" matching in this paper.

the complexity of production and the impact of interruptions to be positively related (that is, as the complexity of production increases, so does the impact of disruptions).

3.3 Hypotheses

Based on the discussion on the relationship between product complexity and production interruptions in the previous section, I identify five hypotheses to build my analysis around. I specify all hypotheses as null-hypotheses I attempt to reject in the tests.

The first set of hypotheses forms the basis of my further analysis. I expect that a higher marginal product sophistication (plant complexity) is positively associated with higher plant revenues. Given the general importance of electricity and intermediate inputs in manufacturing, I also expect that both state-level and supply-chain electricity interruptions is negatively related to plant revenue. This provides the first set of hypotheses:

> $H1_0$: Plant level complexity is not associated with higher revenues.

> $H2_0$: The level of state-wide electricity shortages is not associated with variation in plant revenues.

> $H3_0$: The level of supply-chain electricity shortages is not associated with variation in plant revenues.

Next, I turn the predictions from section 3.2. Given that state-level shortage should act at plant-level, I expect it to be negatively associated with plant revenues, but not that it interacts with complexity. On the contrary, as outlined in the model above, I expect that shortages in the supply chain will have an increasingly large effect as plants become more complex. This gives me my next hypothesis test:

> $H4_0$: The association between a plant's complexity and revenues does not change across different levels of supply-chain electricity shortages.

Finally, the fifth hypothesis relates to the long-run association between interruptions and the kinds of productive capability that is present in the economy. If there is a matching effect in the entry choices producers make, I would expect that the level of electricity shortages in a state in previous years would dis-incentivize the entry of more complex plants. As before, I specify this as a null hypothesis to reject:

> $H5_0$: The previous level of shortages in a state is not associated with the complexity of new plants.

For all of the following hypotheses tests I use the standard rejection threshold of 95% ($p < 0.05$).

In section 5.1 I present how I operationalize the variables used to test the claims above. I detail my empirical strategy to test hypothesis $H1_0$ through $H4_0$ in section 5.2.1. I present approach to testing the long-run changes of $H5_0$ in section 5.2.2.

4 Data

To perform the tests described in section 5.2 I collect a broad set of data on states, manufacturing plants, and socio-economic indicators. I present it here.

4.1 Economic complexity

I collect data on the complexity of products and countries from the Observatory of Economic Complexity (OEC) (Simoes and Hidalgo, 2011). They provide highly disaggregated data on the economic complexity of products down to the Harmonized System (HS) six-digit level. Here I use products at the four-digit level, classified by the HS 1996-revision. For details on how the complexity of a product is calculated, see appendix B.

4.2 Electricity shortages

My data on electricity shortages comes from India's Central Electricity Authority (CEA). The main feature of the dataset is a measure of shortages based on the difference between the observed consumption of electricity and the estimated counterfactual demand. I extract the energy data from the Power Supply Position of States section of the annual Load Generation Balance Reports published by the CEA (CEA, 2017). Digital versions are only available from 2009-10 at the earliest. For earlier years (1998-2009) I use the dataset constructed by Allcott et al. (2016) who worked in collaboration with the CEA to collect, digitize, and clean earlier reports. I then perform an extra step of cleaning up inconsistencies in the early-years set of observations.

4.3 Manufacturing plants

I use plant level data from the Annual Survey of Industries (ASI) (MOSPI, 2016). The ASI is the primary source of information on industry in India and is collected annually by the Indian Ministry of Statistics and Programme Implementation (MOSPI). The ASI covers the manufacturing units in the registered sector. All registered factories with more than 100 employees (the "census scheme") are surveyed every year. Smaller factories are randomly sampled every year, stratified by industry and state. For the years 2000-2003, the census scheme covered factories with more than 200 workers. Until 2004 the sampling scheme covered around 1/3 of all registered factories. Since then, it has covered around 1/5. For each plant, the ASI provides comprehensive information relating to input, output, value added, employment, and assets. For any analysis using the ASI data, unless otherwise specified, I use the yearly sample weights supplied with the data.

Importantly, the ASI is a unique source of product-level output[6]. In the earlier years of my sample, products are listed according to their 5-digit ASI Commodity Classification (ASICC) codes, whereas later years are listed in NPCMS-2011 codes. This is a non-trivial change in classification system. Given a different classification structure, products can

[6] As an example of the granularity of the original data, the ASICC and NPCMS-2011 codes distinguish between detergent paste, detergent cake, and detergent powder. The HS-96 series groups these together at the four digit level, but distinguishes them from "Organic surface-active agents (not soap); surface-active, washing (including auxiliary washing) and cleaning preparations".

both change in complexity and importance in output-volume. To assign complexity values to plant's production, I convert the product classification into Harmonized System 1996 (HS96) codes through a series of concordances. For a walk-through, see appendix A.

There are a couple of important shortcomings when using ASI data. First, while the census schemes covers all factories with more than 100 (or 200) workers and the sampling scheme is a representative sample of smaller factories, they apply only to registered factories. Nagaraj (2002) shows that only around 48% and 43% of the manufacturing establishments covered in the economic census for 1980 and 1990 appear in the ASI for the given year. Additionally, there is a possibility of under-reporting a plant's value-added for tax-avoidance purposes. However, if the non-included factories or the under-reported value-added are not strongly related to electricity shortages or to product complexity, the results should not be influenced. Even if these issues persist to years covered by my sample, the ASI covers a significant part of the Indian manufacturing sector.

In the raw dataset (2000-2016) there are 908,010 observations of plants. The data set require substantial cleaning. For example, factories can be observed in the survey even if they have closed down. After cleaning the dataset according to the procedure outlined in appendix A, I am left with 565,223 plant-in-year observations between 2000 and 2016 across 30 states[7].

4.4 State-wise variables

I get data on state-wise net domestic product, both total and per capita, from the Reserve Bank of India (RBI). While both series are in constant prices, there are often multiple base-years available. When there are observations for the same year using different base-years, I use the newest. The choice of base is not a completely trivial issue, since there are often substantial differences for estimates for the same year measured against two different base-years. For instance, after rebasing GDP in 1999-2000, the total net domestic gross product increases by almost 65% compared to the same year (1999-2000) measured using 1993-94 as a base. To make sure that rebasing the GDP does not drive results spuriously, I rerun all the main tests with indicator variables for the base year.

From the RBI I also collect data on the total population in each state, the share of population that lives in urban and rural areas and the population density. These variables are mainly used in robustness checks. These values are only available at the 10-year census intervals (1991-2001-2011). For years in between I create a simple imputation of change between observations evenly spaced out on years. For any analysis using population controls, I exclude years after 2011 (last available census). I also collect individual level microdata on from the Indian National Sample Surveys (NSS) (NSSO, 2016). The NSS is conducted by the National Sample Survey Office (NSSO) and provides high frequency representative data on a variety of socio-economic outcomes. Surveys are structured as rounds, with each round typically covering a year. From the NSS I clean and aggregate rounds from 2000 through 2015 to state-wise information on the share of the population that has completed at least a secondary education as well as the share of people between 15-60 years of age (which proxies for the size of the potential work force).

[7] I also exclude a few state (and state-like regions) due to either A) having very few years included in the CEA reports (Lakshadweep, Damordar Valley Corporation, and Telangana), B) because they are not present in the RBI data (Dadra and Nagar Haveli), and C) if they have too few observations in the final base sample (Andaman and Nicobar Islands, Arunachal Pradesh).

5 Methodology

In this section I first develop the variables used to test the hypotheses presented in section 3.3. I outline the empirical strategy in section 5.2. I present the results of the analysis in section 6.

5.1 Key variables

5.1.1 Plant complexity

For each plant, I quantify the complexity of its production output as the weighted average the complexity-values for each product it produces. I assign weights based on the value of the production of each product. That is, the complexity for factory f at time t, $C_{f,t}$, is defined as:

$$C_{f,t} = \sum_{p} PCI_{p,t} \frac{O_{f,p,t}}{\sum_{p} O_{f,p,t}}$$

where $PCI_{p,t}$ is the product complexity of product p at time t and $O_{f,p,t}$ is the output (in current prices) of factory f in product p at time t. The value of the production output is calculated as the net unit sale value of a given product times the amount of units sold.

This definition potentially underestimates the complexity of multi-product factories that produce complex products, but also happens to sell a lot of their low-complexity products. I therefore also include a stricter measure of plant complexity, $C_{f,t}^{\max}$, that uses only the most complex product in a factory's product-portfolio, regardless of the output volume. Alternatively, this measure can be thought of as the top-line or "complexity capacity" of a given plant.

$$C_{f,t}^{\max} = max\{Q_{1,t}I_{1,f,t}, \ \dots \ , Q_{p,t}I_{p,t}\}$$

where

$$I_{p,f,t} = \begin{cases} 1 & \text{if } O_{f,p,t} \geq 0 \\ 0 & \text{if } O_{f,p,t} = 0 \end{cases}$$

5.1.2 Electricity shortage

At the end of each year, the Central Electricity Authority (CEA) and the Regional Power Committees estimate the monthly counterfactual quantity that would have been demanded in each Indian state if there were no shortages. This annual figure, listed in current prices, is the assessed demand of electricity in a state (A). The sum of electricity available from power plants and net imports is the energy available (E). The measure of shortages (S or Shortage) is then defined as the percent of demand in state s in year t that is not met:

$$S_{s,t} = \frac{A_{s,t} - E_{s,t}}{A_{s,t}}$$

In addition, the CEA reports a measure of the power shortages during peak hours (S^p). This "peak shortage" is defined analogously to S but using only peak assessed demand (A^p) and peak energy available (E^p):

$$S^p_{s,t} = \frac{A^p_{s,t} - E^p_{s,t}}{A^p_{s,t}}$$

The final sample consists of state-year observations of 30 states from 1998-2016.

5.1.3 Supply chain shortage

In addition to the CEA estimated electricity shortage, I construct a separate measure of supply-shortages. As discussed in section 3.2, using only the plant-state shortage relationship underestimates the importance of connections between different plants in production networks. In order to construct my measure of supply-chain quality, I need to connect three different sets of information. First, I have the interruption-variable, Shortage, on the state-level. Next, I know which plants produce which products, how much of them each contributes, and in which states they are located. Finally, I also have information on what kinds of inputs plants use. I connect them in two steps.

First, for each year, I find how much of each product is produced in each state. I then assign a weighted "production shortage"-average to each product. Formally, I find the product-shortage value for product p S_p in year t by weighting the shortage of each state by the share of the product that is produced there:

$$S_{p,t} = \sum_s w_{p,s,t} S_{s,t}$$

where $w_{p,s,t}$ is the share of product p's total yearly output that is accounted for by state s in year t. As before $S_{s,t}$ is the average shortage for state s in year t. It is important to note that this metric does not take into account any spatial relationships. If the supply-linkages is strongly conditioned on geographical closeness this approach will mis-assign the importance of shortages in industries. That is, I make the implicit assumption that industry-wide output is distributed evenly geographically. Since the ASI does not carry information of where plants source their inputs from, I can't account for this effect.

Second, for each plant I find the importance of each input. I define this as the purchase-value of the input-product as a share of the plants total revenue. Then, for each plant in each year, I attach the "product-shortage" value of their inputs. I multiply this value by the product's share of the individual plant's revenues and sum the result for all products for each plant. This has two advantages: it means that products are given importance relative to their importance to the plant, and that the supply-shortage is more important for plants that use more intermediate inputs (because the supply uncertainty is less important for plants that hardly use any inputs) to create their revenue. Formally, let $m_{f,p,t}$ be the ratio between the amount of product p plant f lists as input (expressed in current prices)

divided by the total revenue of plant f. The supply shortage value for plant f in a given year t is then found in $D_{f,t}$ (D for disruption, to distinguish it from S):

$$D_{f,t} = \sum_{p} m_{f,p,t} S_{p,t}$$

Note that the m values does not necessarily sum to one, since it is weighted by the share of revenues, not by share of all inputs. This means that plants with a smaller share of inputs in their production process will be assigned a smaller supply shortage, even if the product-shortage of the input is high.[8]

5.2 Empirical strategy

In this section I outline the empirical strategy I use to test hypotheses presented above. As a general approach, I use variations over several different specifications of least squares regressions with fixed effects. I present the tests for hypotheses $H1_0$, $H2_0$, $H3_0$, and $H4_0$ in 5.2.1 and the test for hypothesis $H5_0$ in 5.2.2, respectively.

5.2.1 Plant revenues, complexity, and electricity shortages

In this section, I describe how I test hypotheses 1 to 4 through a regression analysis.

Let $O^g_{f,t}$ be the (natural log of) revenue (total gross sales) of a plant f in year t expressed in 2004 constant Rs. This my dependent variable. I add controls for the number of workers employed by each plant (averaged over a year) $E_{f,t}$, the prosperity in the state (net domestic product per capita, constant prices) $N_{s,t}$, and the share of revenue paid as wages $W_{s,t}$. To make sure revenues are not just driven by large turnovers, I also control for the plant's position in the distribution of total production costs in the year I observe the plant. I use the distribution instead of the actual monetary value because I don't have a reliable deflator for the total production costs[9]. I take the Z-score of the total production costs of the plant by by subtracting the average value of observed plants in the same year, and dividing by the standard deviation. I denote this value as $X^z_{f,t,i}$. I then add fixed effect controls for state R_s, year T_t, two-digit industry I_i. Finally, I add my variables of interest: plant complexity $C_{f,t}$ and the two disruption variables, Shortage ($S_{s,t}$) and Supply shortage ($D_{f,t}$). I also include an interaction term between complexity and each of the two electricity shortage variables.

The interaction term is added to address hypothesis 4 on how supply chain shortages affect the impact of complexity on plant output. One of the key predictions from the discussion on the relationship between more complex production and disruptions is that electricity shortages at the plant level should affect the plant-revenue more or less equally across complexity level. Disruptions in the supply-chain, however, should not. As the importance of intermediate products in a plant's production increases, especially if they

[8] This measure only takes into account downstream effects. Should there also be an effect where factories that has to stop their work because shortages result in customers buying less input it would have an effect on supplying industries. However, unless more complex factories are more likely to be situated as a supplier to other factories, it is hard to see how this downstream link would effect plant-complexity.

[9] In contrary to only adjusting prices of plant output, this was unfeasible because it would involve deflating wages, energy inputs, intermediate inputs, etc, separately.

themselves have a long supply line, the lost value of input-supply disruptions increases. In other words, for more complex plants, we would expect that shortages in the supply-chain increases in importance. We can test this relationship by modelling the average revenues of plants, but adding an interaction term between the two different kinds of shortages and complexity. If it is true that supply chain interruptions is more important for more complex plants, I would expect that the interaction between complexity and supply-chain shortage is significant and negative, but the interaction between the state shortage and complexity is less so[10][11].

To keep at least some legibility, let all the plant-level controls (E, W, X^z) be contained in $PLANT$ and the fixed effect indicators (R, T, I) be contained in FE. I also display only the three coefficients I'm interested in: the coefficient of complexity on revenues, the coefficient of electricity shortages on revenues, and their interaction.

Indexing plants by f, years by t, and states by s this gives me the full form of my main two main equations:

$$O^g_{f,t,i} = \beta_{comp} C_{f,t} + \beta_{short} S_{s,t} + \beta_{S \times C} C_{f,t} \times S_{s,t} + PLANT + FE \tag{1}$$

$$O^g_{f,t,i} = \beta_{comp} C_{f,t} + \beta_{supply} D_{s,t} + \beta_{D \times C} C_{f,t} \times S_{s,t} + PLANT + FE \tag{2}$$

I test for hypothesis $H1_0$, $H2_0$, and $H3_0$ using the coefficient significance test by stepwise adding the variables for complexity (C), state-level shortage (S), and supply chain shortages (D). To be precise, I test the significance of coefficients β_{comp}, β_{short}, and β_{supple} in explaining the revenues of plants.

For hypothesis $H4_0$, on how supply chain shortages alter the effect of complexity on plant revenue, I turn to the $\beta_{D \times C}$ coefficient. If $\beta_{D \times C}$ is statistically significant and negative, it suggests as plants supply shortages increase, the added value of a plant being more complex declines.

Notice that I conduct my tests with clustered standard errors. Since my shortage variable is applied at the level of states in different years, plants that are observed in the same year and state have potentially correlated errors. To adjust for this I cluster my standard errors by state-years (so that "Jammu and Kashmir in 2009" is one cluster and "Jammu and Kashmir in 2010" is another).

I present the results from the regressions above, as well as some alternative models, in section 6.2.

5.2.2 Electricity shortages and long-run changes

I now turn to the effect of longer-run shortages on the distribution of factories. Are plants with a more complex production output less likely to be constructed in states that have

[10] This assumes that plants do not have the majority of their supply chain in the same state.

[11] I only have a good industry-specific (3-digit) deflator of gross sales up to 2011, meaning that I limit my sample to the plants observed between 2000 and 2011 for the main analysis. I further restrict it by removing the observations that are flagged by the procedure described in appendix A.1. This leaves me with some 285,000 observations of plants from 2000 to 2011.

unreliable electricity? The analysis described here constitutes the test for hypothesis $H5_0$: if the rolling two-year average state level shortage, $\bar{S}$, is significantly associated with the complexity of new plants in the state, I reject the null hypothesis.

The ASI asks plants when they had first year of commercial production. This makes it possible to date the entry of the plant, and match it with the state shortage variable S. While the productive capital in the manufacturing sector, like factory machinery, is presumably relatively static and product specific, there is no guarantee that a plant produces the same product at the time of survey and the time of the inaugural production. This means that the plant complexity observed is not necessarily the plant complexity at entry.

Since I do not have across-year plant indicators, lumping all plants that is introduced during my sample means that there is a risk of observing the same factories again and again. This entails that it is possible that any results is being driven by factories correlating with themselves. I therefore limit the analysis to observing factories exactly 2 years after their reported inaugural production (no overlap).[12] This allows plants to have a year getting started, but is not so far removed at that observations of plant-characteristics at entry are unlikely to hold.

Essentially, this analysis asks: how is the complexity of plants observed two years after their initial year of production (reported, but not observed) related to the average electricity shortage in the year of their initial production and the year before? For instance, if a plant observed in Assam in 2005 reports that it had its first commercial production in 2003, I assign the average shortage of 2001-2003 in as Assam to this observation.

As before, I use robust standard errors and cluster by state-years. To account for the concerns about downward biased standard errors due to serial correlation discussed in Bertrand et al. (2004) (that is, if some effect "contained" the shortage outcomes carries over to other years), I also cluster by state as a robustness check.

I test hypothesis $H5_0$ in two configurations: a minimal and an expanded model. As before, let $C_{f,i,s,t}$ be the complexity of a plant f with its first year of production in year t indexed by 2-digit NIC industry i and state s. I repeat all of the following analyses using C^{max} as well. I then define $\bar{S}_{s,t}$ as the average shortage in the year of entry and the year prior in state s.[13] R_s and T_t represents state and year specific indicator variables. Finally, I_i represents industry indicators. While I would prefer use industry-year dummies to control for specific temporal industry trends (like an event or demand - policy, climate, etc - that influences the entry of plants within an industry, but is not related to shortages), the (relatively) few observations means that some state-industry-entry year groups becomes very small.

Again, I only display the coefficient I'm interested in. The regression equation then takes the form:

$$C_{f,i,s,t} = \beta_{\bar{S}} \bar{S}_{s,t} + R_s + T_t + I_{i,t} \tag{3}$$

Here, the $\bar{S}_{s,t}$ takes on the role of possible producers' knowledge or expectation of z in the

[12] Estimates were slightly smaller, but essentially the same, when using plants observed three years after entry instead.

[13] It is worth noting that I define the lagged shortage as the average value of the $S_{s,t}$ variable across the years, not the ratio between the average availability and demand over the period (since, presumably, producers won't be interested in overall power capacity, only the amount they are short).

theoretical model discussed earlier. That is, are they willing to start up a factory, if they believe the electrical infrastructure is poor?

I also test using a more expansive model that controls for various factors that could have an "attraction" effect on the entry of plants at the state-level. While most of the differences between states should be caught in the fixed effects, I now allow for variance in the share of the population that are working age (15-60), the share of people with at least a secondary education completed, the growth of the state economy and the total size of the state economy.

If the coefficient for the two-year average state electricity shortage $\bar{S}$, $\beta_{\bar{S}}$ is significantly associated with the complexity values of new plants in both the minimal and the expanded model, I reject hypothesis $H5_0$.

Table 1: World Bank Enterprise Surveys and the Shortage variable

	Self-gen share	Obstacle	Power quality	Self-gen share	Obstacle
	(1)	(2)	(3)	(4)	(5)
Shortage	66.787**	4.672***	−8.509*	45.567***	10.823**
	(21.129)	(0.470)	(3.367)	(12.081)	(4.038)
Constant	15.835***	2.118***	6.290***	4.827***	1.463***
	(2.843)	(0.092)	(0.306)	(1.002)	(0.258)
Industry FE	Yes	Yes	Yes	Yes	Yes
Obervations:	1126	2278	2270	4712	7365
WBES:	2005	2005	2005	2014	2014

Notes: Column one, two, and three are based on the WBES in India in 2005. Column four and five are from the WBES in India in 2014. Dependent variable in column (1) and (4) is the reported share of self-generated electricity. Column (2) and (5) are the degree to which electricity is an obstacle to the firms operation (from 0: "No obstacle" to 4: "Very severe obstacle"). Column (3) only exists for the 2005-survey and is the reported quality of the power grid (from 1: "Extremely bad" to 10: "Excellent"). All columns use industry-fixed effects. Standard errors are robust and clustered by state.

6 Analysis and results

In this section I present the findings from the analysis detailed in section 5.2. I first discuss some initial results on the validity of the electricity shortage variable. I then turn to the results from testing hypotheses $H1_0$, $H2_0$, $H3_0$ (6.2.1), $H4_0$ (6.2.2), and $H5_0$ (6.3).

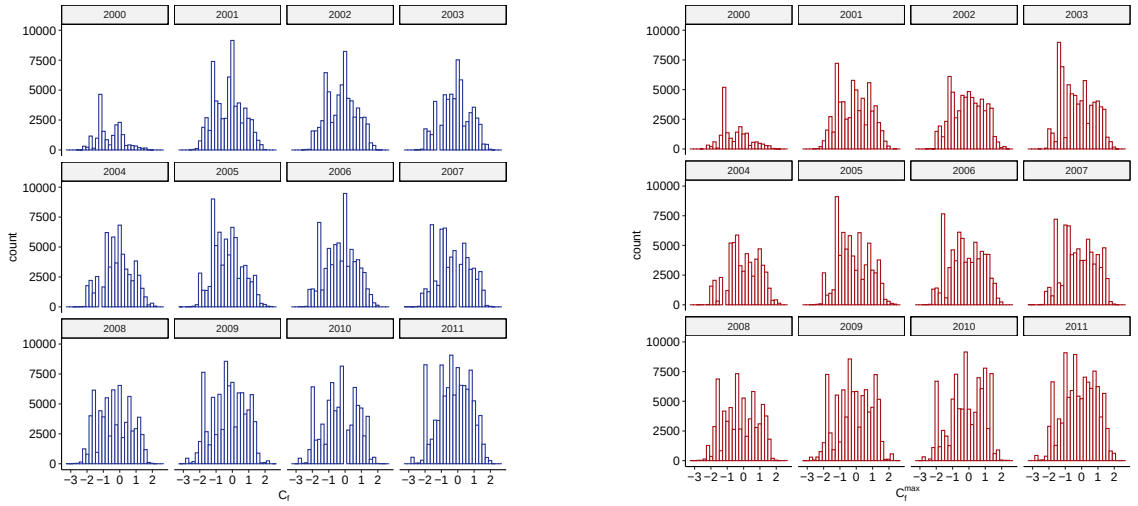
6.1 Validity of electricity variable

The shortage observations depends on an official estimation of the non-shortage demand and is likely to be affected by measurement error. To confirm that the energy data is meaningful, I run a series of simple regression tests to "ground truth" state shortages to microdata on the experience of manufacturing firms. The World Bank periodically conducts enterprise surveys in a range of countries. These surveys cover a broad variety of indicators on the production environment of firms, including perceived challenges and quality of public service provision. Survey data is available for India from 2005 and 2014 (World Bank, 2020a,b)[14]. As shown in Table 1, the shortage variable is a significant predictor (at the $p < 0.05$ level) of the reported quality of the electricity supply (2005), the severity of electricity quality as a barrier to doing business (2005, 2014), the share of electricity generated by firms' own generator (2005, 2014)[15].

While there is likely to be some degree of attenuation bias in the shortage variable, it is a significant predictor of electricity quality across firms and time. To add, Alam (2013) use a measure based on night light composites to identify blackouts and shows that it is highly correlated with the peak version of the shortage variable. This suggests that the shortage measurements carry meaningful information on the electricity reliability at the state level.

[14]The 2005 and 2014 sample covers 2,286 and 7,365 firms, respectively. The 2005 survey does not employ sample weights. For the 2014 survey, I use the strict weighting scheme.

[15]The reason the power quality variable is only listed for 2005 is that the questionnaire have been updated.



(a) Distribution of C_f

(b) Distribution of C_f^{max}

Figure 8: Distribution of factories by complexity for the sample used in the main analysis on the interaction between Shortage and complexity.

6.2 Interaction between shortages and complexity

I now present the results from the analysis outlined in section 5.2.1 on the relationship between plant revenues, electricity shortages, and interruptions in the supply chain. I first address hypotheses $H1_0$, $H2_0$, and $H3_0$. Lastly I present the findings on the interaction effects on plant complexity and electricity shortages in the supply chain ($H4_0$).

6.2.1 Revenues, complexity and shortages ($H1_0$, $H2_0$, and $H3_0$)

Table 2 and 3 displays the results from the regression in equation 1 and 2.

As for hypothesis $H1_0$, column (1) through (7) shows that the complexity of a plant explains a statistically significant part of the variation on plant revenues, under a range of controls and using both measures of electricity interruption (with and without interactions). Since the adjusted revenues are expressed in their natural log, the coefficients are readily interpreted: without the interaction effects, a one unit increase in PCI of a plant's most complex product is associated with a marginal increase in revenues of 15% (see Table 3). The equivalent number is more more pedestrian 4% for the average complexity of the plant (see Table 2). Since PCI-values are standardized for each year, a one unit increase means producing products that are one standard deviation more complex. However, given that I test for interaction effects in the next section, one should be careful about interpreting the association between complexity and revenues separately.

While the explanatory power a plant's complexity adds to the model is very small, the association is robust across the different specifications. I reject hypothesis $H1_0$: "Plant level complexity is not associated with higher revenues."

Surprisingly, the state-level Shortage is not found to be significantly associated with changes in plant revenues. Although the coefficient on Shortage is negative in both Table 2 and 3, it is statistically insignificant in both cases. I therefore cannot reject hypothesis $H2_0$: "The level of state-wide electricity shortages is not associated with variation in plant

revenues."

Next I turn to hypothesis $H3_0$. The strongly significant Supply shortage coefficient in column (5) in Table 2 and 3 suggests that marginal increases in interruptions in a plant's suppliers is associated with a loss of plant revenues. I reject hypothesis $H3_0$: "The level of supply-chain electricity shortages is not associated with variation in plant revenues."

While they are not reported here, I also run a number of different variations using intermediate input share of revenues, electricity intensity (revenue/kWh), and log-forms of the predictors. None of the results reported above is sensitive to any of these alternatives.

In conclusion, I find significant positive associations between marginal increases in the complexity of a plant's production and significant negative associations between marginal increases in the amount of shortages contained in a plants supply-network and its revenues. I don't find a significant relationship between marginal changes in state-level electricity shortage and variations in plant revenue. I therefore reject hypotheses $H1_0$ and $H3_0$, but not $H2_0$.

6.2.2 Interaction between supply-chain shortages and plant complexity ($H4_0$)

I now turn to the main effect studied in this paper: the relationship between a plant's complexity and supply chain shortages. Again, Table 2 and 3 displays the main results. Column (5), (6), and (7) shows the result from using Supply shortage as control, adding an interaction with complexity, and further adding state-level Shortage as control. The coefficient of the Supply shortage is significant and negative at each step.

The most important result is in the interaction with either of the complexity measures. The negative complexity $\times$ Supply shortage coefficient means that as the plants get more complex, the negative association between electricity shortages on plant revenues becomes increasingly strong. Analogously, it could be interpreted as when Supply shortages increase, it becomes increasingly less profitable to produce more complex products.

To make sure that the result is not just driven by within-state shortages, I run the same analysis using the state-level Shortage. This version of the model is reported in column (4) (with Shortage alone) and column (7) (using Shortage as a control). While the Shortage interaction on C_f is mildly significant in column (4), the fact that it, contrary to the Supply shortage coefficient, is insignificant alone, interacted with C_f^{max} and as a control for the Supply shortage interaction with C_f, suggests that the result is less robust.

If complex production processes are punished more severely by interruptions, we would expect that electricity shortages in their supply chain would be more costly to the revenue of plants as you move up in the complexity distribution. This is precisely the picture that emerges in Table 2 and 3. I further explore the relationship in two ways.

The theoretical model predicts that while reliance on inputs (namely specific inputs) have multiplying effects on the level of shortage-induced losses, a larger share of labour in production should not have this effect. I would therefore expect that the interaction effect between the wage-share of revenues and Supply shortages to be in the opposite direction: as the level in supply-chain interruptions increase, higher shares of wages to revenue (proxying for labor inputs) will be increasingly positive for plant revenue. Column (1) in Table 4 shows that while the coefficient of wage shares alone remain negatively related to revenues,

in the face of supply chain interruptions, the effect is completely reversed.[16] I also tests for the relationship between the amount a plant spends on intermediate inputs (as share of revenue) and the Supply shortage. There is no significant relationship. Given the noisy model, this could be due to a weak effect being lost. It could also, however, highlight the fact that inputs are not alike: some inputs are highly specific and vulnerable to supply-interruptions, while others are not. Based on the data presented here, it is not possible to make judgement either way.

In addition, I run a test with a slightly adjusted sample. The Supply shortage variable is highly skewed for a few observations (around 30). In column (1) and (2) of Table 5 I remove all the observation that have a Supply shortage value above one. A supply shortage value of more than one would indicate that plants use more than all of their revenue on inputs, and that they source all of these inputs from a state that has a Shortage value of 1 - that is, no electricity available at all. I run the two main regressions again. Column (3) and (4) in the same table further limits the sample to only include plants that list at least one input, no matter its share of their revenues. Just removing the 32 observations increases the importance of both the Supply shortage variable and its mitigating effect on plant complexity even further. Indeed, from the effect on revenues being stronger than the dampening effect on complexity, the reverse is now observed[17].

Similarly to the previous section, I repeat the tests with different variations in controls, including electricity intensity (kWh/revenue). I also test the interaction using centered versions of C_f, C_f^{max} and Supply shortage. None of the results change significantly.

Together, these results are strongly in line with the theoretical predictions. Three effects are worth highlighting.

- First, the analysis showed a weak- or non-existing relationship between the amount of state-level Shortage and the association of complexity and revenues.
- Second, the analysis found a persistently strong association between higher levels of supply-network shortages and a less positive relationship between plant complexity and revenues.
- Finally, the association between the share of revenues paid in wages and plant revenues is reversed as the level of Supply shortages increase. This suggest that plants with a high share of wages paid relative to revenue are relatively less adversely impacted by supply interruptions.

[16] I only report the relationship to C^{max}, but the outcome is statistically identical for C.

[17] One should be careful making too strong claims about the numerical impact of the shortages, though. Depending on the sample used, the mean value of the Supply shortage variable ranges around 0.04-0.045, with a standard deviation of about the same, meaning that it is very easy to make conclusions on out-of-sample values.

Table 2: Association between complexity (C_f) of plants, shortages, and revenues.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(Intercept)	16.53***	16.58***	16.58***	16.59***	16.62***	16.62***	16.62***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Wage/rev share	−4.38***	−4.38***	−4.38***	−4.39***	−4.37***	−4.38***	−4.38***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Self-gen (1)	1.23***	1.23***	1.23***	1.23***	1.23***	1.23***	1.23***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Number of employees	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Production costs (z)	0.32***	0.32***	0.32***	0.32***	0.32***	0.32***	0.32***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
State NDP/cap	0.00***	0.00**	0.00**	0.00**	0.00**	0.00***	0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
C_f		0.04**	0.03**	0.06***	0.03**	0.06***	0.06***
		(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Shortage			−0.11	−0.17			−0.07
			(0.17)	(0.17)			(0.17)
C_f × Shortage				−0.29*			
				(0.14)			
Supply shortage					−1.34***	−1.50***	−1.49***
					(0.22)	(0.27)	(0.27)
C_f × Supply shortage						−0.52*	−0.52*
						(0.23)	(0.23)
R^2	0.41	0.41	0.41	0.41	0.41	0.41	0.41
Adj. R^2	0.41	0.41	0.41	0.41	0.41	0.41	0.41
Num. obs.	285126	285126	285126	285126	285126	285126	285126
RMSE	2.71	2.71	2.71	2.71	2.71	2.71	2.71
N Clusters	331	331	331	331	331	331	331

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

[1] This table presents the result of the regression analysis in section 6.2 on the interaction between plant complexity, energy shortages, and plant revenue. The dependent variable in all columns is the natural log of yearly total plant revenue (constant 2004 Rs). Independent variables are in order: wages paid as a share of revenue, an indicator for whether or not the plant owns a generator, the total cost of production standardized per year, the plants' total number of yearly employees (average), the net domestic product in each state per capita, the weighted average complexity the plant, shortage in the state, interaction between complexity and shortage, the supply shortage value of the plant, and interaction between complexity and supply shortage. Each column includes state-, two-digit industry-, and year fixed effects. Standard errors are robust and clustered by state-year.

Table 3: Association between the most complex product in plants (C_f^{max}), shortages, and revenues.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(Intercept)	16.53***	16.73***	16.73***	16.73***	16.77***	16.77***	16.77***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)
Wage/rev share	−4.38***	−4.40***	−4.40***	−4.40***	−4.39***	−4.39***	−4.39***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Self-gen (1)	1.23***	1.23***	1.23***	1.23***	1.22***	1.22***	1.22***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Number of employees	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Production costs (z)	0.32***	0.31***	0.31***	0.31***	0.31***	0.31***	0.31***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
State NDP/cap	0.00***	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
C_f^{max}		0.15***	0.15***	0.16***	0.15***	0.17***	0.17***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Shortage			−0.09	−0.10			−0.04
			(0.18)	(0.18)			(0.18)
C_f^{max} × Shortage				−0.14			
				(0.13)			
Supply shortage					−1.33***	−1.42***	−1.42***
					(0.21)	(0.24)	(0.24)
C_f^{max} × Supply shortage						−0.43*	−0.42*
						(0.21)	(0.21)
R^2	0.41	0.41	0.41	0.41	0.41	0.41	0.41
Adj. R^2	0.41	0.41	0.41	0.41	0.41	0.41	0.41
Num. obs.	285126	285126	285126	285126	285126	285126	285126
RMSE	2.71	2.71	2.71	2.71	2.70	2.70	2.70
N Clusters	331	331	331	331	331	331	331

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

[1] This table presents the result of the regression analysis in section 6.2 on the interaction between plant complexity, energy shortages, and plant revenue. The dependent variable in all columns is the natural log of yearly total plant revenue (constant 2004 Rs). Independent variables are in order: wages paid as a share of revenue, an indicator for whether or not the plant owns a generator, the plants' total number of yearly employees (average), the total cost of production standardized per year, the net domestic product in each state per capita, complexity of the most complex product a plant produces, shortage in the state, interaction between complexity and shortage, the supply shortage value of the plant, and interaction between complexity and supply shortage. Each column includes state-, two-digit industry-, and year fixed effects. Standard errors are robust and clustered by state-year.

Table 4: Association between Supply shortages, wage-share, intermediate input share, and revenues.

	(1)	(2)	(3)	(4)
(Intercept)	16.82***	16.77***	16.73***	16.73***
	(0.07)	(0.06)	(0.07)	(0.06)
C_f^{max}	0.15***	0.15***	0.15***	0.15***
	(0.01)	(0.01)	(0.01)	(0.01)
Supply shortage	−2.62***	−1.33***		
	(0.33)	(0.21)		
Wage/rev share	−4.60***	−4.39***	−4.41***	−4.40***
	(0.07)	(0.06)	(0.09)	(0.06)
Self-gen (1)	1.22***	1.22***	1.23***	1.23***
	(0.04)	(0.04)	(0.04)	(0.04)
Number of employees	0.00***	0.00***	0.00***	0.00***
	(0.00)	(0.00)	(0.00)	(0.00)
Production costs (z)	0.31***	0.31***	0.31***	0.31***
	(0.05)	(0.05)	(0.05)	(0.05)
State NDP/cap	0.00**	0.00**	0.00**	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)
Supply shortage × wage/rev share	3.49***			
	(0.79)			
Int. input/rev share		0.00*		−0.00
		(0.00)		(0.00)
Int. input share × Supply shortage		−0.00		
		(0.00)		
Shortage			−0.10	−0.09
			(0.20)	(0.18)
Shortage × wage/rev share			0.10	
			(0.93)	
Shortage × int. input share				0.00
				(0.00)
R^2	0.41	0.41	0.41	0.41
Adj. R^2	0.41	0.41	0.41	0.41
Num. obs.	285126	285126	285126	285126
RMSE	2.70	2.70	2.71	2.71
N Clusters	331	331	331	331

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

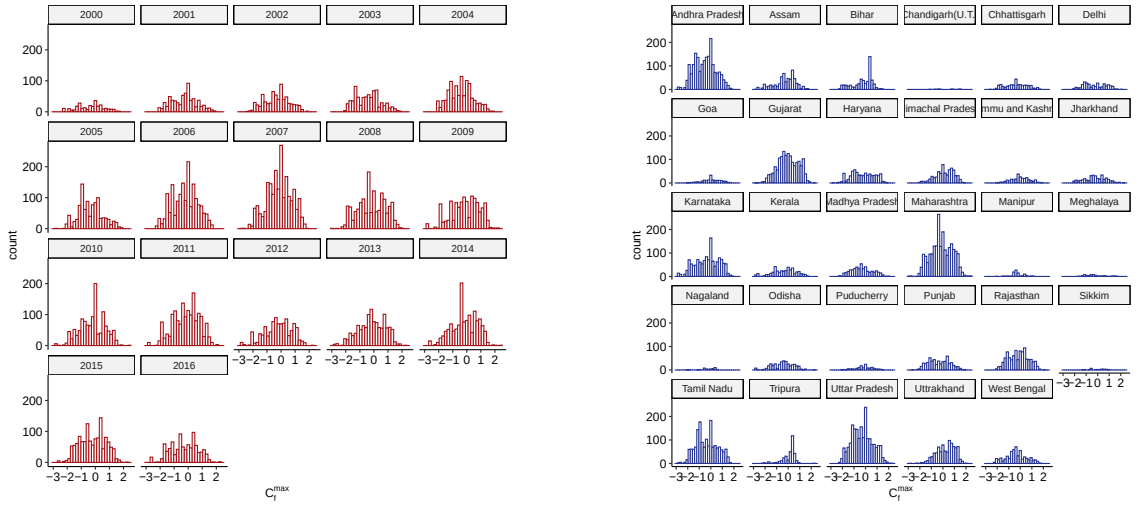
[1] This table presents a modified version of the equation in 6.2, here with focus on the interaction between wage/revenue share, intermediate input/share (products), shortages and plant revenue. The dependent variable in all columns is the natural log of yearly total plant revenue (constant 2004 Rs). Independent variables are the same as in Table 3 except for the inclusion of Int. input/rev share, which is the purchase value of intermediate input products as a share of revenues (the given plant in the given year). Each column includes state-, two-digit industry-, and year fixed effects. Standard errors are robust and clustered by state-year.

Table 5: Association between the complexity of plants and Supply shortage: adjusted sample.

	(1)	(2)	(3)	(4)
(Intercept)	16.79***	16.64***	16.84***	16.70***
	(0.06)	(0.06)	(0.07)	(0.07)
C_f^{max}	0.18***		0.21***	
	(0.01)		(0.02)	
Supply shortage	−1.99***	−2.15***	−2.34***	−2.61***
	(0.24)	(0.24)	(0.23)	(0.23)
Wage/rev share	−4.40***	−4.38***	−4.40***	−4.38***
	(0.06)	(0.06)	(0.06)	(0.06)
Self-gen (1)	1.22***	1.23***	1.22***	1.22***
	(0.04)	(0.04)	(0.04)	(0.04)
Number of employees	0.00***	0.00***	0.00***	0.00***
	(0.00)	(0.00)	(0.00)	(0.00)
Production costs (z)	0.31***	0.32***	0.29***	0.30***
	(0.05)	(0.05)	(0.05)	(0.05)
State NDP/cap	0.00**	0.00***	0.00*	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)
C_f^{max} × Supply shortage	−0.56**		−1.15***	
	(0.21)		(0.22)	
C_f		0.07***		0.11***
		(0.01)		(0.02)
C_f × Supply shortage		−0.76***		−1.43***
		(0.22)		(0.23)
R^2	0.41	0.41	0.41	0.41
Adj. R^2	0.41	0.41	0.41	0.41
Num. obs.	285094	285094	266539	266539
RMSE	2.70	2.71	2.68	2.69
N Clusters	331	331	331	331

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

[1] This table presents the results of the same analysis as in Table 2 and 3 (interaction between Supply shortage and plant complexity), but with a slightly adjusted dataset. Column (1) and (2) restricts the sample to filter very few strong outlier-values (values above 1 is removed). Column (3) and (4) also restricts the sample to plants that have at least 1 intermediate input. Each column still includes state-, two-digit industry-, and year fixed effects. Standard errors are robust and clustered by state-year.



(a) Factories by year

(b) Factories by state

Figure 9: Distribution of factories in the plant-entry sample (2-year after entry). Some states only have a few factories that are observed after their initial production. However, all results hold after rerunning the tests without the least represented states. The vertical axis is the (unweighted) count of factories with the given C_f^{max} in the state/year.

6.3 Do shortages discourage the entry of complex plants?

I now present the results from the regression model highlighted in section 5.2.2.

First, a note about the sample size. Limiting the analysis to only plants observed in a set amount of time of their initial production naturally reduces the sample size substantially - from around 500,000 observations to around 20,000[18]. Figure 9 show that some states are reduced to very few observations. For my main analysis, I run the regression excluding the four least represented states (Chandigarh, Manipur, Nagaland, and Sikkim). All the findings were repeated when using the full sample. It is also worth noting that in figure 9, there is no evidence of increasing left-skew in the distribution of complexity as time moves forward. This suggests that the effect is not just driven by all factories getting more complex with time (so that plant entries in later years carry the significance). To make sure of this, I rerun the analysis with a variable that indicates the median and mean complexity of all plants entering in the given year (not reported). This control does not change any of the results.

6.3.1 Entry of new plants: minimal model ($H5_0$)

Do higher average shortages in a state dissuade the entry of factories with a complex production? Table 6 offers some tentative evidence that the previous quality of the electricity supply is in fact associated with which kinds of plants starts producing in a state. Column (1) and (2) shows the results from the minimal model. Both the most complex products produced in a plant and the plant's average complexity is positively associated with a smaller degree of electricity disruptions. The size of the effect, however, is very small: a one percentage point increase in the average shortage over the previous years is associated with a 0.006 decrease in the PCI of the products a plant produces.

[18]Since I don't use the adjusted revenues in this analysis, I can include my full sample of factories.

In column (3) and (4), I change the plant-level complexity value with the median complexity of all plants within their two-digit industry code. A negative coefficient would mean here that plants in more complex industries would enter into states with lower average shortages. There is no evidence that this is the case. These columns exclude the industry-fixed effects used in the other columns. Column (5) shows the linear probability that a plant entering into a state with higher average electricity shortages will possess a generator. The result is positive and strongly significant. Finally, column (6) shows that the electricity/revenue share of new plants is not significantly associated with the state's past shortage.

It is worth highlighting that the two-digit industry of plants accounts for almost all of the variance explained in the model. This is perhaps unsurprising as we would expect the variation in production complexity to be much greater between different industries than within them. Given the large effect of industry-indicators, I also test an expanded model where I exclude indicators for some of the configurations. I now move to this model.

6.3.2 Entry of new plants: expanded model ($H5_0$)

Table 7 and 8 shows the marginal association between $\bar{S}$ and C_f and C_f^{max} in the model with an expanded group of controls. Column (1)-(3) excludes industry-effects. Again the average electricity shortage retains a robust, but weak, association with the kinds of plants that begin production. The result is strongest when adding all controls, including the industry-indicator: controlling for the share of the population with a secondary education, the share of people in working age, the prosperity of the state, the economic growth in the state, and the size of the total economy, a one percentage point increase in the electricity shortage of the past two years is associated with a 0.009 decrease in the PCI of the most complex products new plants produce (and a slightly smaller decrease in the average). Column (5) highlights the importance of the industry the plant belongs to: the model's explanatory power barely changes using only the entry year-, state- and industry-fixed effects.

In conclusion, while the quantitative effect is very small, the data show a weak association between the two-year average electricity shortage in a state and the type of factories that enters into operation. Given that this result is robust to a variety of controls, including changing the error-clustering to states (rather than state-year), adding a control for overall change in production complexity in India, and a range of different combinations of controls, and using both the maximum and the average complexity of the products plants produce, I reject hypothesis $H5_0$: "The association between a plant's complexity and revenues does not change across different levels of supply-chain electricity shortages."

Table 6: Association between complexity of new plants, electricity use, and shortages

	C_f^{max}	C_f	C_{ind}^{max}	C_{ind}	Self-gen (1)	Electricity rev share
(Intercept)	-1.24^{***}	-1.31^{***}	-0.41^{***}	-0.50^{***}		180.09
	(0.07)	(0.07)	(0.05)	(0.05)		(434.39)
$\bar{S}_{s,t}$	-0.60^{*}	-0.61^{*}	0.33	0.32	0.53^{***}	-9433.58
	(0.30)	(0.29)	(0.23)	(0.22)	(0.13)	(7225.74)
R^2	0.50	0.50	0.08	0.07	0.26	0.02
Adj. R^2	0.50	0.50	0.07	0.07	0.26	0.01
Num. obs.	18974	18974	18974	18974	17680	12329
RMSE	1.18	1.14	1.27	1.19	0.67	16237.57
N Clusters	420	420	420	420	418	280

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

[1] This table presents the results from equation 3. $\bar{S}$ is the two-year average state shortage at the time of entry. Column (1) is the complexity of the most complex product a plant produces. Column (2) is the weighted average complexity of the products it produces. Column (3) and (4) are the median C^{max} and C of plants in an industry (2-digit level). In column (5) and (6) the dependent variable has been changed to the linear probability of a factory self-generating electricity and the adjusted revenue per kWh electricity used. All columns use fixed effects on state and year, but only column (1), (2), (5), and (6) uses 2-digit industry fixed effects. Standard errors are robust and clustered by state-year .

Table 7: Association between the most complex product produced in new plants (C_f^{max}) and electricity shortages: more controls

	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.29	−1.52	1.38	2.99*	−1.24***
	(0.67)	(1.87)	(1.88)	(1.39)	(0.07)
$\bar{S}_{s,t}$	−0.80*	−0.76*	−0.78*	−0.92***	−0.61*
	(0.33)	(0.33)	(0.32)	(0.25)	(0.30)
Share with sec. education	0.08	−0.32	−0.16	−0.14	
	(0.74)	(0.81)	(0.78)	(0.58)	
Share between 15 and 60	−1.19	−0.78	−0.40	−0.14	
	(1.10)	(1.14)	(1.09)	(0.82)	
ln(State NDP/cap)		0.18	−0.24	−0.53***	
		(0.17)	(0.19)	(0.14)	
State NDP/cap growth		−0.21	−0.25	−0.10	
		(0.17)	(0.17)	(0.14)	
ln(State NDP)			0.45***	0.51***	
			(0.11)	(0.09)	
R^2	0.09	0.09	0.09	0.51	0.51
Adj. R^2	0.09	0.09	0.09	0.51	0.50
Num. obs.	15450	15450	15450	15450	18839
RMSE	1.58	1.58	1.58	1.16	1.18
N Clusters	313	313	313	313	383

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

[1] This table presents the results from the model of new plants with expanded controls. The dependent variables in all five columns in the complexity of the most complex product produced by a new plant ($C_{f,t}^{max}$). The independent variables in order: the two-year average state shortage at the time of entry, the share of population that has completed (at minimum) a secondary education, the share of the population between 15 and 60 years old, the natural log of state net domestic product by capita (constant prices), the yearly change in net domestic product by capita (constant prices), the natural log of total state net domestic product (constant prices, divided by 1 mio). Columns (1) through (3) use state- and entry-year fixed effects. Column (4) and (5) also use two-digit industry fixed effects. All standard errors are robust and clustered by state-year.

Table 8: Association between the complexity of new plants (C_f) and electricity shortages: more controls

	(1)	(2)	(3)	(4)	(5)
(Intercept)	−0.06	−1.47	1.36	3.01*	−1.31***
	(0.65)	(1.75)	(1.75)	(1.28)	(0.07)
$\bar{S}_{s,t}$	−0.69*	−0.66*	−0.69*	−0.85**	−0.61*
	(0.32)	(0.32)	(0.31)	(0.26)	(0.29)
Share with sec. education	−0.18	−0.50	−0.34	−0.35	
	(0.69)	(0.74)	(0.71)	(0.53)	
Share between 15 and 60	−0.68	−0.34	0.02	0.27	
	(1.07)	(1.11)	(1.07)	(0.80)	
ln(State NDP/cap)		0.14	−0.27	−0.57***	
		(0.16)	(0.18)	(0.13)	
State NDP/cap growth		−0.18	−0.22	−0.07	
		(0.16)	(0.15)	(0.13)	
ln(State NDP)			0.44***	0.54***	
			(0.11)	(0.09)	
R^2	0.08	0.08	0.08	0.50	0.50
Adj. R^2	0.08	0.08	0.08	0.50	0.50
Num. obs.	15450	15450	15450	15450	18839
RMSE	1.52	1.52	1.52	1.12	1.14
N Clusters	313	313	313	313	383

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

[1] This Table presents the results from the model on new plants with expanded controls. The dependent variables in all five columns in weighted average complexity of new plants ($C_{f,t}$). The independent variables in order: the two-year average state shortage at the time of entry, the share of population that has completed (at minimum) a secondary education, the share of the population between 15 and 60 years old, the natural log of state net domestic product by capita (constant prices), the yearly change in net domestic product by capita (constant prices), the natural log of total state net domestic product (constant prices, divided by 1 mio). Columns (1) through (3) use state- and entry-year fixed effects. Column (4) and (5) also use two-digit industry fixed effects. All standard errors are robust and clustered by state-year.

7 Discussion

In this section I discuss the most important findings from the analyses presented in the previous section, how they relate to the theoretical model from the framework, as well as some important limitations.

7.1 Findings

Based on the framework developed in section 3.2 I made several theoretical predictions on the relationship between interruptions at the plant- and supply-chain level and the complexity of the products that a plant makes. In section 3.3 I translated them into a set of hypotheses to test my data against. How well did these predictions stand up to the findings from the empirical analysis?

The first set of hypotheses (1 to 3) were the "basic" predictions: all else equal, a higher product complexity is associated with higher revenues and a higher level of electricity interruptions is associated with lower revenues. While both measures of plant complexity and the supply shortage variable matched expectations (although with a small effect size), it is surprising that the state-level shortage was not associated with lower average revenues. For example, Allcott et al. (2016) estimates a substantial loss in revenues based directly on the same measure. Theoretically, I would expect there to be a relationship. A higher amount of time without electricity would more production down-time. Even for plants using generators, the higher prices of electricity inputs should be observed in the data.

Two reasons could plausibly explain the different result. First, they use an IV-approach to instruments for changes in the state-level shortages. This reduces a lot of ambiguous relationship between factors that impact both revenues and shortages (like economic growth). Second, they have access to individual plant-level indicators across different sample years. Again, this reduces the noise in the data significantly: by using plant-level indicators, they are able to control all the intrinsic between-plant heterogeneity. In my model, the differences between plants becomes added statistical noise. Finally, since their sample (1992-2010) is different from mine (2000-2010 for the specific analysis), it is possible (but unlikely) that the relationship does not hold for only a subset of the years.

Next, an important implication from the model discussed in section 3.1 is that while electricity shortages acting at the plant-level - such as outages on the general supply network - will act on plants in a similar way (depending on their electricity intensity), disruptions in the supply chain will not. Testing for this effect, I find a significant relationship between a decreasing association between increases in marginal product complexity and higher plant revenues as the level of supply shortages increase. In other words, the data suggest that if there is a high degree of supply-chain uncertainty, the more complex factories gets less profitable (compared to less complex factories). Such an interaction effect is not observed between the state-level shortages and the association between plant complexity and revenues[19].

So what does such a finding mean? Essentially, this suggests that in higher uncertainty environments, such as many developing economies, it becomes less attractive to invest in a more complex production setup. Given the relatively strong interaction coefficient of

[19]While one term, $C_f \times S_s$, was weakly significant under a specific set of controls, all the other combinations were not. I therefore disregard that result here.

supply-shortages on the benefit of complexity, this suggest that for at given level of disruption in the production environment, producers can "overshoot" the complexity of their production. Should there also exist a higher price of entry into producing more complex products, this could have a significant impact on the incentives to invest in higher complexity production technology. This point is highlighted by the result that while a higher wage/revenue share is strongly negatively associated with plant revenues, the interaction between wage-share and shortages in the input supply is strongly positive. This suggests that, again, the presence of a more disruptive production environment perverts incentives away from what usually considered preferable (less primary industry, more complex production).

Finally, I find a weak but significant association between the two-year average state wide electricity shortage and the complexity of new factories. In this test, we can think of the two-year average electricity shortages as a proxy for potential entrants knowledge or belief about how reliable the electricity supply is. This result suggests that producers and investors are less willing to begin production of higher complexity products if they believe that the state-wide electricity is more unreliable.

Given that I find no complexity specific interaction between state-level shortages and revenues, there is no clear reason for why the complexity of plant should condition whether or not it is likely to enter into specific states based on their previous shortage-level[20]. Instead, a possible interpretation is through the matching effect discussed in the theoretical model. If we assume that plants that produce a more complex product are situated longer down the value-chain, failures or production stops are more expensive here (at the supply-chain level). Since the aggregate expected value of the production chain is substantially reduced by adding even a few more easily interrupted links, we would expect that markets organize their high and low interruption industries together (Kremer, 1993). At the aggregate level, this would mean that high value production chains would be less likely to enter into previous high-shortage states. Should more complex products be more likely to appear in such high-value chains, this would explain the pattern found in the data.

As a final point, the results presented in this paper suggests that much of the literature that estimates the costs to low-quality electricity underestimates two effects. First, there is the unaccounted for electricity disruption in supply chain network. If this effect is not included, a supply-driven loss in revenues would go unnoticed in the "treatment"-group and cause downward-biased estimates of the true output penalty to the electricity outages. To my knowledge, this supply chain-effect has not been introduced in literature on electricity costs on manufacturing output. Second, while a few papers explicitly find some change in firm behaviour, like substituting local for imported inputs (Fisher-Vanden et al., 2015) or changing the relative share of production inputs (Abeberese, 2017), there is typically no account of the structural effects on what kinds of new production is introduced into the economy (except for explicitly electricity-related outcomes).

7.2 Limitations

I now outline a few of the limitations should be kept in mind when interpreting the results discussed above.

[20] As with any of the results, this should be taken with the caveat that the statistical power of the test might not be strong enough to detect a result, should there be one.

7.2.1 Research design

First, and perhaps most importantly, the effects shown in this paper are purely associative. I do not randomize any assignment of shortages, and I don't instrument for changes in the electricity availability. This means that any causal claims has to be purely conjectural. In addition, the main difficulty in testing the predictions from the framework is in controlling from the other factors that co-move with power quality and plant complexity. Without causal assignment, it is difficult to assess if this research cleared that bar.

Second, I don't have any across-time plant identifiers. While these do exist in the original data, since the 2018 release of the ASI on MOSPI's data platform this variable have been scrubbed. This reduces the power of the research substantially. As noted earlier, this also strongly reduces the explanatory power of the plant-entry analysis, given the limitations it puts on the sample size. Finally, all observations becomes significantly more noisy without specific plant identifiers.

Third, given the rather small effect-sizes (which is, at least in some part, a function of the lack of plant ID), strictly testing any hypotheses on the base of coefficient p-values potentially overstate the significance of findings. That said, all my main results are robust to a range of different specifications and controls. As an additional point, some of the key results, like the interaction between input shortages and plant complexity, is driven by variables with a relative narrow interval of values. Hence, one should take care not to draw conclusions on out-of-sample range values.

Finally, the result found here is only based on the Indian experience during the sample-years, in particular that of the manufacturing sector. There is, however, no reason why the effect of interruptions on the production choices could not exists in many places. The only barrier to a similar study being replicated on developing economies elsewhere is the scarcity of data (where the ASI is of particularly high quality, especially given the time span). However, efforts like the UNU-WIDER and the University of Copenhagen's Myanmar Enterprise and Monitoring System (MEMS) would be a straight-forward candidate for a similar study.

7.2.2 Endogeneity and attenuation bias

There are a couple of reasons the effect of electricity disruptions on economic activities are difficult to study empirically. First, the relationship is likely to have a significant endogenous component, but it is unclear in which direction. More complex production could be related to a more intensely developed economy, which could also be related to more stable electricity supply. On the other hand, a more developed economy could have a more complex production, but would also have a higher electricity demand which could lead to shortages, especially given the disconnect between market demand and the electricity supply-sector in India (as outlined in section 2). Second, electricity reliability is difficult to measure without error. My electricity variable are estimations and on a quite broad unit of measurement (states). This means that any uncertainty in the estimation of counterfactual demand is added to the natural variation between electricity in different areas in each state and is carried into my analysis. This very likely introduces some measurement error in my independent variable.

One way to address this issue would be to repeat the analysis using one the IV-approaches in the literature. For instance, Allcott et al. (2016) uses the marginal extra available

energy from hydroplants introduced by rainfall at higher elevations to instrument for yearly variation in electricity supply. This would be valuable as an alternative to S in the analysis of supply-shortage interaction with complexity in section 6.2.

7.2.3 Modifiable Area Unit Problem

My tests rely on "ground down" state-wide variables on electricity reliability to individual plants. As with much of research across spatial units, this runs into the issue of artificial boundaries. It is not uncommon that results that are based on aggregate units disappear at more fine-grained analysis. For instance, states might not be appropriate scale of measurement, or be homogeneous in its distribution of reliable electricity. Min et al. (2017) construct a power supply irregularity index of 600,000 villages in India based on high frequency night-light photos. They show that there is substantial variance within states. Whether or not this village variance would impact the variance on plants within states is unclear. The authors did not respond to multiple requests, and their concrete methodology is unclear from their work. It would be valuable to test the results from this paper against a more fine-grained assignment of electricity reliability (also, one that is not estimated from official side).

8 Conclusion

In this paper I have explored the relationship between the level of electricity shortages contained in a plant's production environment, both locally and in the supply chain, and the economic complexity of its output.

I first develop a theoretical model of the connection between economic complexity and interruptions. From this model, I then create a set of hypotheses that I bring to data. Using a variety of different regression specifications on a large, unit-level data set (spanning from 2000-2016), I find robust evidence on the connection between electricity shortages and production complexity in two important ways.

First, I find evidence suggesting that while the local electricity supply is not related to the importance of complexity for plant revenues, interruptions in the input supply of plant strongly conditions the association between complexity of a plant's output and revenues.

Second, I find a small, but robust, association between the past two-year average electricity shortages of states and the entry choices of new plants. A higher two-year level of average electricity shortages is associated with less complex plants entering into production in the state.

Taken together, these results suggests an important relationship between the complexity of an economy and its level of production disruptions. The association between interruptions in the supply-chain, complexity, and revenues suggests that higher levels of risk of production-failure (or stoppage) can lead to perverse incentives in the manufacturing sector by increasing the returns to primary production relative to more sophisticated activities.

While these results are purely correlations, they suggest two reasons that much of the previous literature have underestimated the costs of a poor quality power supply. The first reason concerns the revenue effects of interruptions in the input supply. Many studies exploit some version of an instrumental variable assignment of electricity shortages to study the loss of revenue. However, if these studies do not take into account the revenue impacts of interruptions in the input supply, some of the variation in revenues that is caused by electricity shortages will be mis-assigned. This creates a downward bias in the estimate of the true cost of poor electricity. Secondly, the results presented in this paper suggests that as interruptions contained in the supply-chains increase, the value of more complex production decreases. This suggests that studies on the aggregate loss to electricity shortages also need to take into account possible self-selection into less productive industries. In light of strong empirical evidence linking country-level economic complexity and growth, this is especially important.

References

Abeberese, A. B. (2017). Electricity Cost and Firm Performance: Evidence from India. *The Review of Economics and Statistics*, 99(5):839–852.

Abeberese, A. B., Ackah, C. G., and Asuming, P. O. (2019). Productivity Losses and Firm Responses to Electricity Shortages: Evidence from Ghana. *The World Bank Economic Review*, (lhz027).

Acemoglu, D., Carvalho, V. M., Ozdaglar, A., and Tahbaz-Salehi, A. (2012). The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016.

Adam, A., Garas, A., and Lapatinas, A. (2019). Economic complexity and jobs: An empirical analysis.

Allcott, H., Collard-Wexler, A., and O'Connell, S. D. (2016). How do electricity shortages affect industry? Evidence from India. *American Economic Review*, 106(3):587–624.

Balassa, B. (1965). Trade liberalisation and "revealed" comparative advantage. *The manchester school*, 33(2):99–123.

Barrot, J.-N. and Sauvagnat, J. (2016). Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks. *The Quarterly Journal of Economics*, 131(3):1543–1592.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275.

Boehm, C. E., Flaaen, A., and Pandalai-Nayar, N. (2019). Input linkages and the transmission of shocks: Firm-level evidence from the 2011 Tōhoku earthquake. *Review of Economics and Statistics*, 101(1):60–75.

Brummitt, C. D., Huremovic, K., Pin, P., Bonds, M. H., and Vega-Redondo, F. (2017). Contagious disruptions and complexity traps in economic development. *Nature Human Behaviour*, 1(9):665–672.

Can, M. and Gozgor, G. (2017). The impact of economic complexity on carbon emissions: Evidence from France. *Environmental Science and Pollution Research*, 24(19):16364–16370.

Carvalho, V. M. (2014). From Micro to Macro via Production Networks. *Journal of Economic Perspectives*, 28(4):23–48.

CEA (2009:2017). Load Generation Balance Report. Technical report, Central Electricity Authority, New Delhi.

CEA (2013). Planwise Capacity Addition. Technical report, Central Electricity Authority, New Delhi.

Chakravorty, U., Pelli, M., and Ural Marchand, B. (2014). Does the quality of electricity matter? Evidence from rural India. *Journal of Economic Behavior & Organization*, 107:228–247.

Fisher-Vanden, K., Mansur, E. T., and Wang, Q. J. (2015). Electricity shortages and firm productivity: Evidence from China's industrial firms. *Journal of Development Economics*, 114:172–188.

Frenken, K., Van Oort, F., and Verburg, T. (2007). Related Variety, Unrelated Variety and Regional Economic Growth. *Regional Studies*, 41(5):685–697.

Grainger, C. A. and Zhang, F. (2017). *The Impact of Electricity Shortages on Firm Productivity: Evidence from Pakistan*. Policy Research Working Papers. The World Bank.

Hartmann, D., Guevara, M. R., Jara-Figueroa, C., Aristaran, M., and Hidalgo, C. A. (2017). Linking Economic Complexity, Institutions and Income Inequality. *World Development*, 93:75–93.

Hausmann, R. and Hidalgo, C. A. (2011). The network structure of economic output. *Journal of Economic Growth*, 16(4):309–342.

Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., Simoes, A., and Yildrim, M. A., editors (2013). *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. The MIT Press, Cambridge, MA.

Hidalgo, C. and Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575.

Hidalgo, C. A., Klinger, B., Barabasi, A.-L., and Hausmann, R. (2007). The Product Space Conditions the Development of Nations. *Science*, 317(5837):482–487.

Inoua, S. (2016). A simple measure of economic complexity. *arXiv preprint arXiv:1601.05012*.

Johnson, P. and Papageorgiou, C. (2020). What remains of cross-country convergence? *Journal of Economic Literature*, 58(1):129–175.

Jones, C. I. (2011). Intermediate Goods and Weak Links in the Theory of Economic Development. *American Economic Journal: Macroeconomics*, 3(2):1–28.

Kremer, M. (1993). The O-Ring Theory of Economic Development. *The Quarterly Journal of Economics*, 108(3):551–575.

McRae, S. (2015). Infrastructure quality and the subsidy trap. *American Economic Review*, 105(1):35–66.

Min, B., O'Keeffe, Z., and Zhang, F. (2017). *Whose Power Gets Cut? Using High-Frequency Satellite Images to Measure Power Supply Irregularity*. Policy Research Working Papers. The World Bank.

MOSPI (2000:2016). Annual Survey of Industries. http://microdata.gov.in/nada43/index.php/home.

Nagaraj, R. (2002). How to improve India's industrial statistics. *Economic and Political Weekly*, pages 966–970.

NSSO (2000:2016). National Sample Survey (NSS) data (unit level): Survey round: 55, 56, 57, 58, 60, 61, 62, 63, 64, 66, 68, 70, 71, 72. http://microdata.gov.in/nada43/index.php/catalog.

Pritchett, L. (1997). Divergence, big time. *Journal of Economic perspectives*, 11(3):3–17.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, 98(5, Part 2):S71–S102.

Samad, H. and Zhang, F. (2016). *Benefits of Electrification and the Role of Reliability: Evidence from India*. Policy Research Working Papers. The World Bank.

Simoes, A. J. G. and Hidalgo, C. A. (2011). The economic complexity observatory: An analytical tool for understanding the dynamics of economic development. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., and Pietronero, L. (2012). A new metrics for countries' fitness and products' complexity. *Scientific reports*, 2:723.

World Bank (2020a). Enterprise Surveys: Firm-level data on India (2005). https://www.enterprisesurveys.org/en/data.

World Bank (2020b). Enterprise Surveys: Firm-level data on India (2014). https://www.enterprisesurveys.org/en/data.

World Bank (2020c). Enterprise Surveys Indicators Data. https://www.enterprisesurveys.org/en/data.

World Bank (2020d). World Development Indicators: GDP, PPP (constant 2011 international $). https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD.

World Bank (2020e). World Development Indicators: Total natural resources rents (% of GDP). https://data.worldbank.org/indicator/NY.GDP.TOTL.RT.ZS.

Zhang, F. (2018). In the Dark: How Much Do Power Sector Distortions Cost South Asia? page 261.

A Appendix: Data cleaning

A.1 Cleaning Annual Survey of Industries (ASI)

The ASI is distributed by the Ministry of Statistics and Programme Implementation, Government of India, (MOSPI) as ten blocks for every year. These blocks require substantial cleaning and harmonization of variables. Here I outline the filtering procedure.

I first create the base sample. Plants can be included in the survey, even if they are reported to be closed or are missing. I drop all observations not listed as open. I also drop all factories that are not listed as in a manufacturing sector and observations that don't report revenues (defined as the total gross sale value of all production output). Finally, I drop all observations that are exact copies of other observations.

After the initial filtering process, there can still be observations that have misreported values of specific variables. When analysing these variables, I further limit the sample using a "flagging"-system[21].

I assign observations an "input-revenue" flag if their labour or material-costs is more than two times their revenues or if their fuel and electricity costs are greater than revenues. Similarly, I can also observe the quantity of electricity consumed. I multiply the amount of electricity consumed by the state-year median price paid (current Rs/kWh). If the amount is higher than the revenue, I assign a flag. For every time I run an analysis involving any of these variables as an outcome, I exclude observations that are flagged. I drop all the observations that have two or more flags completely. If an observation reports 0 electricity consumption, I set all electricity variables as missing for the observation. I further drop observations on an ad-hoc basis during the analysis. When I do, this is explicitly written in the section.

A.2 Product concordance for ASI

As mentioned in the data-section, the ASI lists products according to two different classification methods. In earlier years (before 2010) the ASICC classification is used, whereas later years lists product by their NPCMS-2011 code. The standard nomenclature for international trade, however, is the Harmonized System classification (HS). Since I assign complexity to plants by their the products they produce, and since I calculate the complexity of products by their position in the international trade network, I need to map the HS system to the codes used in the ASI.

This is rather round-about process. The reason behind the shift from ASICC to NPCMS-2011 is that the early scheme was severely flawed in the grouping-classification and was poorly suited to international comparison. This means that the mapping between ASICC and NPCMS-2011 is imperfect. The NPCMS-2011 mapping is based directly on the international standard Central Product Classification, which again is different from the Harmonized System used in trade-accounting. I first match all products from the ASICC years to the NPCMS-2011 classification with the concordance Table provided by MOSPI [22]. I then turn the NPCMS-2011 codes into the CPC-2 classification by removing the last

[21] This method was inspired by Allcott et al. (2016).

[22] http://www.csoisw.gov.in/CMS/En/1027-npcms-national-product-classification-for-manufacturing-sector.aspx

two digits (which are India specific). I use the concordance Table supplied by UNSD to map the CPC-2 codes to HS-2007. Finally, I use turn the HS-2007 codes into HS-1996 to match the trade data.

Often, one product code from the source classification maps to two different codes in the destination classifications. There is no way to solve this issue completely. Instead, I create two mappings: a "strict" and a "lenient" match. The "strict" match uses only products that have a non-partial match and leaves other products as missing. The "lenient" approach assigns the first of the partial mappings as a match. Since the difference is usually quite small between partially mapped products, is is usually feasible to purposely "misassign" the products to a mapping that exists, rather than drop it altogether. For instance, the ASICC listings of "Lobsters, processed/frozen" (11329), "Prawns, processed/frozen" (11331), "Shrimps, processed/frozen" (11332) all map to two different NPCMS-2011 codes: "Crustaceans, frozen" (212500) and "Crustaceans, otherwise prepared" (212700). Similarly, "Butter" (11411) maps to three different kinds of butter (based on cattle-milk, buffalo-milk, or other milk) in the NPCMS-2011 system. While not particularly rigorous, very little information should be lost on the complexity of the production output between these three mappings. Indeed, many such categories will be clubbed together anyhow when converting NPCMS-2011 to Harmonized System codes. It is worth noting that I use the "strict"/"lenient" approach throughout the concordance chain. This means a substantial product loss in the "strict" approach: products from the ASICC classification (five digits) that might be together in the final Harmonized System code can be dropped because they map to two different NPCMS-2011 codes (that are seven digits vs the four I use in the HS-code). At any rate, while the observations are substantially reduced in some states, the distribution of plant complexity changes very little (see figure 10). I therefore use the lenient approach in my main analysis.

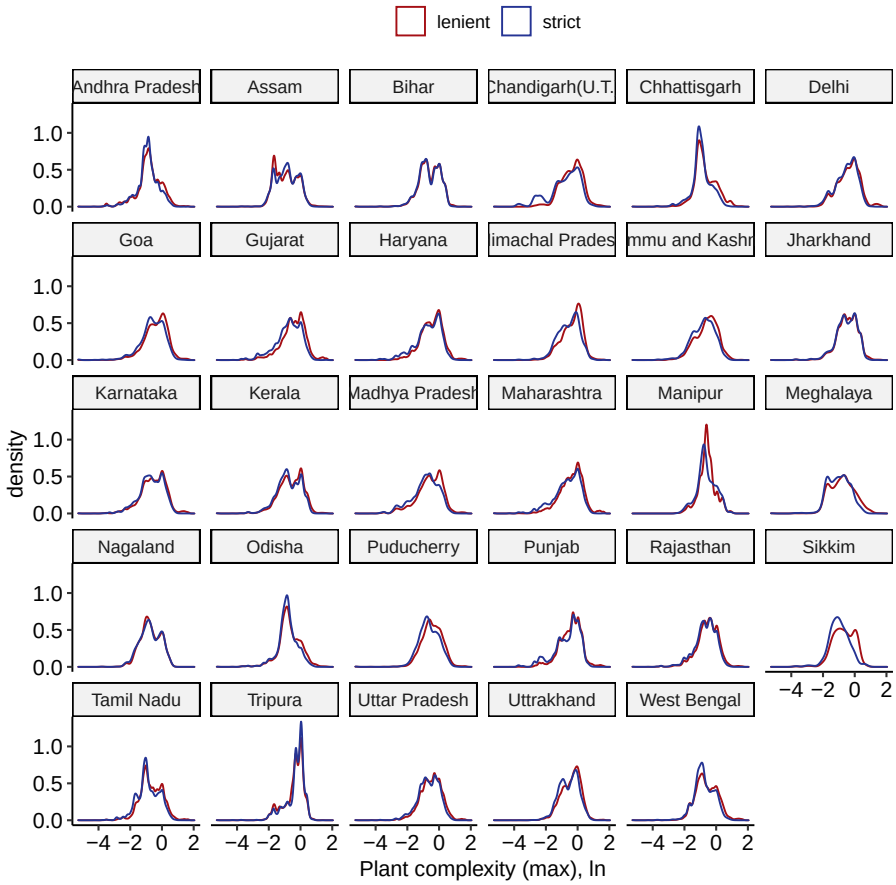


Figure 10: State-wise density of plant complexity by using only strict or lenient matches. All years are pooled. Gaussian kernel.

B Appendix: Calculating product complexity

I use product complexity-values extracted using the Hausmann-Hidalgo (HH) algorithm (Hidalgo and Hausmann, 2009). Here I present the algorithm.

The HH algorithm is essentially an iterative calculation that repeatedly weighs products based on the sophistication of countries that export them and countries based on the products they export. Given the vast international differences in economy-sizes, the export data is first binarized using the revealed comparative advantage (RCA) (Balassa, 1965). The RCA is taken for each country in each of the around 1200 products the HS 96 series. RCA normalizes the export share in a country's total export with the share of a product's global export value in the value of all global exports together. Hence, RCA of country c in product p:

$$RCA_{cp} = \frac{X_{cp}}{\sum_p X_{cp}} \Big/ \frac{\sum_c X_{cp}}{\sum_c \sum_p X_{cp}}$$

where X_{cp} is the export value of country c in product p. I then define an RCA matrix M_{cp} with countries in rows as products in columns as:

$$M_{cp} = \begin{cases} 1 & \text{if } RCA_{cp} \geq 1 \\ 0 & \text{if } RCA_{cp} < 1 \end{cases}$$

As mentioned, the economic complexity of countries and products was originally calculated by repeatedly discounting products by their ubiquity (how many countries exports them with RCA ≥ 1) and weighting countries by the products they export. However, following Hausmann et al. (2013) the end-values of the algorithm can be found by finding the eigenvector that correspondents to the second largest eigenvalue. The product complexity index PCI is then defined as:

$$PCI = \frac{\vec{Q} - < \vec{Q} >}{\text{stdev}(\vec{Q})}$$

where $\vec{Q}$ = eigenvector of $\hat{M}_{pp'}$ associated with the second largest eigenvalue and

$$\hat{M}_{pp'} = \sum_p \frac{M_{cp} M_{cp'}}{k_{c,0} k_{p,0}}$$

And finally, $k_{c,0} = \sum_p M_{cp}$ and $k_{p,0} = \sum_c M_{cp}$. The end result is a product-specific value of economic sophistication that is completely based on whether or not products tend to co-occur in countries' export baskets. No assumptions are made on the intrinsic complexity of different classes of products, only that the most complex products are those that are hardest to make, and they are made by those countries that have the best production knowledge.

C Appendix: Figures

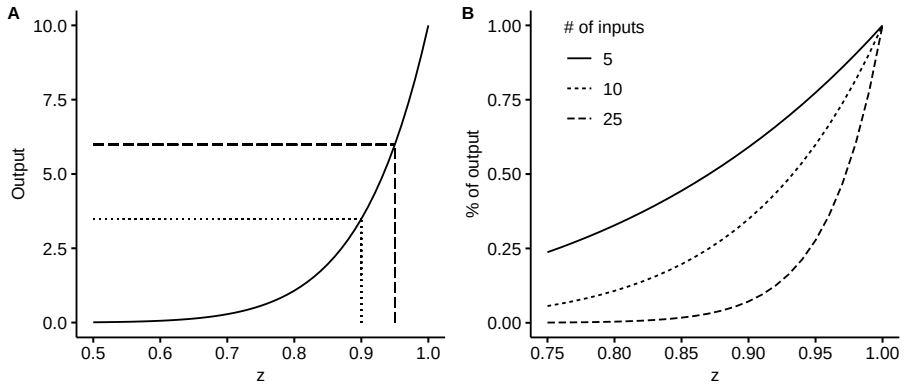


Figure 11: Expected plant output by risk and number of inputs: (A) for very small increases in risk we see massive drops in expected output (in the example $n = 10$): when going from $z = .95$ (dashed) to $z = .90$ (dotted) - a decrease of 5.26% - expected output falls more than 40 %; (B) as complexity of production increases, the drop from potential output driven by marginal decreases in quality skyrockets. In terms of production, this suggest that small changes in the risk of failure disproportionately punishes higher complexity producers.

D Appendix: Tables

Table 9: 'Lenient' vs 'strict' matching to HS96: observations by year

year	unmatched	lenient	strict	**lenient change**	**strict change**
1999	9687	8780	5376	**-0.09**	**-0.45**
2000	24676	21870	11454	**-0.11**	**-0.54**
2001	57113	50797	25762	**-0.11**	**-0.55**
2002	63136	55954	28503	**-0.11**	**-0.55**
2003	65558	57901	29455	**-0.12**	**-0.55**
2004	86605	75863	38476	**-0.12**	**-0.56**
2005	74444	65430	33053	**-0.12**	**-0.56**
2006	79707	71335	35902	**-0.11**	**-0.55**
2007	78660	70255	36000	**-0.11**	**-0.54**
2008	69988	61854	31438	**-0.12**	**-0.55**
2009	68060	66980	34139	**-0.02**	**-0.50**
2010	71788	70548	37054	**-0.02**	**-0.48**
2011	75273	75273	43178	**0.00**	**-0.43**
2012	78302	78302	44208	**0.00**	**-0.44**
2013	75156	75156	43829	**0.00**	**-0.42**
2014	78809	78809	46868	**0.00**	**-0.41**
2015	81359	81359	49344	**0.00**	**-0.39**
2016	81366	81366	48376	**0.00**	**-0.41**

Table 10: 'Lenient' vs 'strict' matching to HS96: observations by state

state	unmatched	lenient	strict	**lenient change**	**strict change**
A and N Islands	479	462	202	**-0.04**	**-0.58**
Andhra Pradesh	90398	86038	50030	**-0.05**	**-0.45**
Arunachal Pradesh	251	251	176	**0.00**	**-0.30**
Assam	20795	20314	11480	**-0.02**	**-0.45**
Bihar	13647	13270	9452	**-0.03**	**-0.31**
Chandigarh(U.T.)	5261	4755	2433	**-0.10**	**-0.54**
Chhattisgarh	19981	19343	11102	**-0.03**	**-0.44**
Dadra and Nagar Haveli	12564	11715	5546	**-0.07**	**-0.56**
Daman and Diu	14219	13161	6249	**-0.07**	**-0.56**
Delhi	25517	23626	11033	**-0.07**	**-0.57**
Goa	14441	12852	5715	**-0.11**	**-0.60**
Gujarat	109480	102425	52153	**-0.06**	**-0.52**
Haryana	49166	45289	26453	**-0.08**	**-0.46**
Himachal Pradesh	26985	25577	8942	**-0.05**	**-0.67**
Jammu and Kashmir	10171	9598	5209	**-0.06**	**-0.49**
Jharkhand	14220	13563	8483	**-0.05**	**-0.40**
Karnataka	70911	65227	34660	**-0.08**	**-0.51**
Kerala	37497	35639	20910	**-0.05**	**-0.44**
Madhya Pradesh	39901	37857	20943	**-0.05**	**-0.48**
Maharashtra	177551	162988	86218	**-0.08**	**-0.51**
Manipur	1535	1533	970	**0.00**	**-0.37**
Meghalaya	2145	2111	1394	**-0.02**	**-0.35**
Nagaland	2301	2244	1446	**-0.02**	**-0.37**
Odisha	20783	20159	11856	**-0.03**	**-0.43**
Puducherry	8447	7757	3761	**-0.08**	**-0.55**
Punjab	66850	62618	37874	**-0.06**	**-0.43**
Rajasthan	43642	41938	22892	**-0.04**	**-0.48**
Sikkim	1202	1201	377	**0.00**	**-0.69**
Tamil Nadu	126272	120292	66847	**-0.05**	**-0.47**
Telangana	9656	9656	5738	**0.00**	**-0.41**
Tripura	5802	5735	4512	**-0.01**	**-0.22**
Uttar Pradesh	95337	90631	49092	**-0.05**	**-0.49**
Uttrakhand	29273	27934	12084	**-0.05**	**-0.59**
West Bengal	53007	50073	26183	**-0.06**	**-0.51**

Table 11: 'Lenient' vs 'strict' matching to HS96: output by year (current R)

year	unmatched	lenient	strict	**lenient change**	**strict change**
1999	8.404983e+13	1.743172e+13	7.433884e+12	**-0.79**	**-0.91**
2000	4.281514e+12	4.113282e+12	2.512859e+12	**-0.04**	**-0.41**
2001	8.919334e+12	8.664177e+12	4.079980e+12	**-0.03**	**-0.54**
2002	7.570050e+12	7.294395e+12	4.059058e+12	**-0.04**	**-0.46**
2003	8.882428e+12	8.570030e+12	4.907379e+12	**-0.04**	**-0.45**
2004	9.928743e+12	9.602304e+12	5.615234e+12	**-0.03**	**-0.43**
2005	1.364389e+13	1.310630e+13	7.605898e+12	**-0.04**	**-0.44**
2006	2.343411e+14	2.321173e+14	9.896938e+13	**-0.01**	**-0.58**
2007	8.492717e+15	8.450439e+15	3.480926e+14	**0.00**	**-0.96**
2008	1.918965e+15	1.896303e+15	1.066407e+15	**-0.01**	**-0.44**
2009	3.068860e+13	3.040062e+13	1.868528e+13	**-0.01**	**-0.39**
2010	2.843052e+13	2.825666e+13	1.656423e+13	**-0.01**	**-0.42**
2011	4.113326e+13	4.113326e+13	2.691098e+13	**0.00**	**-0.35**
2012	5.222862e+13	5.222862e+13	3.460210e+13	**0.00**	**-0.34**
2013	5.441068e+13	5.441068e+13	3.383971e+13	**0.00**	**-0.38**
2014	5.901109e+13	5.901109e+13	3.601650e+13	**0.00**	**-0.39**
2015	6.000773e+13	6.000773e+13	3.883916e+13	**0.00**	**-0.35**
2016	9.163110e+15	9.163110e+15	6.761738e+15	**0.00**	**-0.26**

Table 12: 'Lenient' vs 'strict' matching to HS96: output by state (current R)

state	unmatched	lenient	strict	**lenient change**	**strict change**
A and N Islands	3.730948e+10	3.727433e+10	3.639415e+10	**0.00**	**-0.02**
Andhra Pradesh	1.141583e+14	1.091365e+14	5.515054e+13	**-0.04**	**-0.52**
Arunachal Pradesh	2.618636e+10	2.618636e+10	2.120664e+10	**0.00**	**-0.19**
Assam	1.861085e+13	1.859523e+13	1.142270e+13	**0.00**	**-0.39**
Bihar	2.959615e+13	2.958399e+13	2.451601e+13	**0.00**	**-0.17**
Chandigarh(U.T.)	2.857063e+13	2.789530e+13	2.613528e+13	**-0.02**	**-0.09**
Chhattisgarh	2.817048e+13	2.815152e+13	1.042075e+13	**0.00**	**-0.63**
Dadra and Nagar Haveli	2.262976e+14	2.258103e+14	9.721252e+13	**0.00**	**-0.57**
Daman and Diu	2.259624e+13	2.240136e+13	1.126080e+13	**-0.01**	**-0.50**
Delhi	1.989231e+14	1.976959e+14	1.754432e+14	**-0.01**	**-0.12**
Goa	3.615391e+14	3.604067e+14	2.487690e+13	**0.00**	**-0.93**
Gujarat	8.810380e+15	8.798361e+15	3.018780e+14	**0.00**	**-0.97**
Haryana	2.495809e+14	2.482845e+14	2.003876e+14	**-0.01**	**-0.20**
Himachal Pradesh	9.147398e+14	9.120361e+14	7.526687e+14	**0.00**	**-0.18**
Jammu and Kashmir	2.116757e+13	2.095423e+13	9.981348e+12	**-0.01**	**-0.53**
Jharkhand	2.413089e+14	2.409273e+14	1.457840e+13	**0.00**	**-0.94**
Karnataka	3.462383e+15	3.461098e+15	3.334097e+15	**0.00**	**-0.04**
Kerala	8.018367e+13	7.812379e+13	4.529216e+13	**-0.03**	**-0.44**
Madhya Pradesh	9.148944e+13	8.974632e+13	5.105227e+13	**-0.02**	**-0.44**
Maharashtra	1.398309e+15	1.389269e+15	5.757764e+14	**-0.01**	**-0.59**
Manipur	2.198879e+10	2.198578e+10	1.132198e+10	**0.00**	**-0.49**
Meghalaya	4.882104e+11	4.871523e+11	4.334702e+11	**0.00**	**-0.11**
Nagaland	1.405110e+11	1.404993e+11	1.353239e+11	**0.00**	**-0.04**
Odisha	1.799426e+13	1.794603e+13	1.227580e+13	**0.00**	**-0.32**
Puducherry	1.869095e+13	1.856310e+13	1.142226e+13	**-0.01**	**-0.39**
Punjab	3.392139e+14	3.341858e+14	4.290774e+13	**-0.01**	**-0.87**
Rajasthan	7.302604e+14	6.648560e+14	5.729336e+14	**-0.09**	**-0.22**
Sikkim	4.520094e+12	4.520092e+12	2.083536e+12	**0.00**	**-0.54**
Tamil Nadu	6.537198e+14	6.496537e+14	4.049141e+14	**-0.01**	**-0.38**
Telangana	2.000670e+14	2.000670e+14	1.373189e+14	**0.00**	**-0.31**
Tripura	1.642075e+11	1.641171e+11	8.823975e+10	**0.00**	**-0.46**
Uttar Pradesh	1.530646e+15	1.527540e+15	1.405733e+15	**0.00**	**-0.08**
Uttrakhand	2.797105e+14	2.788987e+14	1.394167e+14	**0.00**	**-0.50**
West Bengal	1.986139e+14	1.806156e+14	6.499627e+13	**-0.09**	**-0.67**