# Inter-industry Total Factor Productivity Spillovers in India

Bos, Marijke; Goderis, Benedikt; Vannoorenberghe, Gonzague

*Document version:*
Preprint (usually an early version)

*Publication date:*
2014

# Inter-industry Total Factor Productivity Spillovers in India

Marijke Bos, Benedikt Goderis and Gonzague Vannoorenberghe

Tilburg University

June 2014

**Abstract**

It is often argued that the positive effects of policies or institutional reforms on the innovation and productivity of firms are amplified by positive spillovers to other firms. In this paper, we use a comprehensive dataset on Indian manufacturing firms and employ spatial econometric techniques to empirically estimate the strength of inter-firm Total Factor Productivity (TFP) spillovers. To address endogeneity concerns, we use tariff reductions during the Indian period of trade liberalization as an exogenous source of variation in TFP (following Topalova & Khandelwal 2011). We focus on three possible channels, namely observation, labor mobility and intermediate input use. We find no systematic evidence in support of TFP spillovers across Indian manufacturing firms.

---WORKING PAPER JUNE 2014---

This working paper is written in the framework of the research project 'Co-ordinated Country Case Studies: Innovation and Growth Raising Productivity in Developing Countries', funded by the Department for International Development (DFID) of the United Kingdom.

## 1. Introduction

It is often argued that the positive effects of policies or institutional reforms on the TFP of firms are amplified by positive TFP spillovers (externalities) to other firms. Because sustained TFP growth can only come from technological progress, technology spillovers play an important role in new growth theories (Grossman & Helpman 1991, Romer 1990). There exists a vast literature on R&D spillovers (input to technology), patents (the output of technology), and TFP spillovers (the effect of technology), but the latter topic is studied almost exclusively in the context of Foreign Direct Investment (FDI). TFP spillovers are of particular interest to industrial policymakers: if a policy raises the TFP of a particular set of firms, but this increase spills over to other firms in the economy, the benefit of the policy is larger and more widespread than would be expected when ignoring the spillovers.

Although theoretically TFP can spill over between any type of firm, foreign or domestic, the simultaneity in TFP spillovers causes a major concern in empirical work. A growing empirical literature studies productivity spillovers between firms and industries. Paz (2012) studies inter-industry productivity spillovers for 16 Brazilian manufacturing industries. Javorcik (2004) looks at the productivity spillovers from foreign direct investment to domestic firms. Amiti and Konings (2007) study the effects of trade liberalization on plant productivity.

In this paper we use the Indian period of trade liberalization as an exogenous source of variation in firms' total factor productivity and combine this with a comprehensive dataset of Indian manufacturing firms to estimate the strength of inter-firm TFP spillovers. We focus our attention to three possible channels, namely observation, labor mobility and intermediate input use, and test which channels of spillovers are present in the Indian economy. We use spatial econometric techniques and instrumental variables to address the endogeneity problems.

## 2. Theory

### *2.1 Productivity spillovers*

Productivity spillovers take place when one firm's productivity has an effect on the productivity of another firm. Although there are many definitions of productivity, total factor productivity (TFP) is the most widely used measure of productivity and is defined as that part of the output that cannot be explained by the amount of inputs used (such as capital, labor, energy, intermediate inputs). TFP growth comprises of efficiency improvements, technical change and scale economies (Otha, 1974 and Denny et al. 1981). Therefore, productivity spillovers occur when one firm experiences an efficiency improvement, technical change and/or economy of scale, and this has an effect on the efficiency, technology and/or scale of another firm. Each of these three components can have an effect on each one or more of the other three components and raise productivity of another firm. For example, it is possible that a technological improvement in one firm results in an efficiency improvement in the other firm. Technical change incorporates technology improvements, but also increased knowledge that allows for greater production for a given set of inputs. In the remainder I will use the word technology to encompass both.

### *2.2 Technology spillovers*

There is a vast literature on technological spillovers. Similar to productivity, technology is difficult to measure directly. Therefore, empirical studies use indirect measures of which the most common are technology inputs (R&D), technology outputs (patents), and the effect of technology (productivity). Research on R&D spillovers generally models the firm's productivity as a function of own R&D efforts, as well as the R&D efforts of other firms, where the latter is sometimes referred to as the pool of knowledge provided by other firms[1]. Because technology has public good characteristics there is scope for spillovers to other firms. There are two ways to construct this so called spillover stock or knowledge pool. In the symmetric approach, all firms are treated equally and all R&D conducted by these firms is aggregated with equal weight. In the second approach, each pair of firms is treated separately. These studies acknowledge that different firms may draw upon a different pool

[1] Rather than looking at technology spillovers by examining inputs (R&D), some studies have also examined technology outputs, patents, as the source of externalities. (see for example Schankerman, 1979).

of knowledge. In particular, technology spillovers are more likely to occur between firms that are close to each other. In such models, the weight that each firm attaches to the stock of technology of another firm increases when the distance between the pair of firms decreases. The term distance is, however, hard to define empirically. In summary, the literature on technology spillovers identifies four measures of distance: vertical linkages, technological distance, proximity in technological research and geographical distance. Studies using the first measure, the degree of vertical linkages (also referred to as input-output relations), assume that R&D is embodied in purchased inputs and therefore the strength of vertical relations between industries is taken as a proxy for promixity. In particular, this measure of proximity is proportional to a firm's purchases from the other firm. Brown and Conrad (1967) for example used the input-output table and Terleckyi (1974) use the capital and intermediate inputs purchases matrix to proxy for the strength of vertical relations. The second measure, technological distance, is based on cross classification of patents or innovations. Scherer (1982, 1984) examined R&D expenditures which were linked to U.S. invention patents. The sample of more than 15,000 patents was classified by their industry of origin, where the R&D had taken place, and the anticipated industry of use. This resulted in a technology flows matrix which was used to weight the available R&D data, assuming that the flow of knowledge from industry $i$ to industry $j$ is proportional to the fraction of patents of industry $i$ that was anticipated to be used in industry $j$. Griliches and Lichtenberg (1984) interpret a similar estimation on a more detailed data set on TFP growth as measuring improvements in materials and equipment bought from other industries. The third measure of distance is one of proximity in technological research. The assumption underlying this measure is that when two firms are active in the same technological area, as indicated by holding patents in the same patent class, they are more likely to benefit from each other's research. In this view, technology is disembodied and its spillovers are not related to trade in intermediate goods.

### *2.3 Agglomeration spillovers*

A fourth measure of proximity is geographical distance between firms. In particular the literature on agglomeration spillovers pays attention to the possibility that technology spillovers are locally appropriated. In other words, there may be geographical boundaries to information flows, especially if the information is tacit (Marshall 1920, Krugman 1991). Moreover, the costs of transmitting knowledge may rise with geographical distance. Marshal (1920) describes three reasons for, or three types of, agglomeration economies: (1) labor pooling, (2) knowledge or technological spillovers, and (3) input-output linkages between vertically related industries. Labor market pooling arises when workers can easily move between firms in the cluster which improves matching between firms and workers. When firms needing the same type of worker cluster together in geographical space it becomes easier for the worker to find a (new) job and for the firm to find suitable employees. Moreover, labor market pooling is one way through which knowledge may spillover. To the extent that knowledge, on for example a firm's products and production processes, is embodied in its workers, flows of workers between firms are in effect flows of knowledge that may raise firm productivity. It is in this way that productivity between firms that share a labor market may spill over from one firm to the other. Several empirical studies have examined the links between worker mobility and productivity spillovers. Stoyanov and Zubanov (2012) for example use matched firm-worker data and find that firm productivity increases when firms hire workers from more productive firms. Without labor market mobility, knowledge and technology may also spillover through observation when firms are located close to another. Proximity in terms of input-output relations (as described above) may also be correlated with proximity of geographical location when firms purposefully locate close to their buyers and suppliers.

### *2.4 FDI spillovers*

A related branch of literature studies technology spillovers through foreign direct investment (FDI). The common assumption in this literature is that spillovers do not arise directly from R&D efforts, but from the premise that for a multinational corporation (MNC) to compete successfully in the local market, it needs to have some superior technology over local firms. This superior technology can spill over to local firms and enhance their productivity. This literature thus looks at one-directional spillovers from the MNC to local firms. The channels through which spillovers take place overlap in part with those identified in the technology and agglomeration spillover literature, and include vertical linkages, demonstration, labor pooling, and effects on market competitiveness (Blomström and Kokko, 1998). Backward linkages, between MNCs and their suppliers, can take the form of assistance and cooperation, but MNCs may also force their suppliers to meet higher quality standards. Contacts between the MNC and its domestic input customers, called forward linkages, could generate spillovers through superior knowledge and technology embodied in their inputs. MNCs may also provide training on how to employ these inputs appropriately (Girma et al., 2008). To the extent that knowledge is embodied in workers, training of employees and demonstration effects may cause spillovers from the MNC to local firms, whether from the same industries or upstream/ downstream industries. Various skills acquired at the MNC may spill over to local firms as the employees switch to other firms. Another way in which spillovers can take place is through observation and copying of technologies. Spillovers through worker mobility and observation are likely to be local, since worker mobility is often geographically limited and observation is easier for firms that located close to the MNC. Another type of spillover arises when the entrance of the MNC leads to more severe competition. This may induce local firms to use their existing technologies more efficiently, by reducing slack or improving X-efficiency. It may also force local firms to search for new and improved technologies, or copy those from the MNC. Technology may spill over unintentionally or intentionally. When costs for the MNC to avoid appropriation are large, or when the costs for the host firm to adopt the technology are small, there may be unintentional spillovers. When, for example, it is costly to prevent labor mobility from the MNC to local firms, technology spills over unintentionally. However, when the MNC expects benefits from making technology available for appropriation, intentional spillovers may occur. For example, the MNC can benefit from providing technical assistance to their suppliers when this increases the quality of the supplied inputs. The channels of productivity spillovers identified by the FDI literature may well apply to a broader context in which one firm, which is not necessarily a foreign or multinational corporation, experiences a productivity boost that may spill over to other firms.

Based on the existing literature, we identify the following channels of productivity spillovers: (1) observation, (2) labor mobility, and (3) intermediate inputs. We focus on inter industry spillovers, and thus exclude potential spillovers from firms in the same industry.

Since competition effects are most likely the result of intra industry pressures, we exclude this channel[2]. The next section describes the empirical analysis.

## 3. Model and Empirical Strategy

### *3.1 Regression model: spatial autoregressive model*

We now turn to the econometric analysis in which we analyze the strength of TFP spillovers across Indian firms. Following the spatial econometrics literature (Anselin 1988, Franzese & Hays 2008), we start from the following spatial lag model:

$$y = \rho W y + X\beta + \varepsilon \tag{1}$$

where $y$ is an $n$ by 1 vector of observations on the dependent variable (Total Factor Productivity), $W$ is an $n$ by $n$ exogenous spatial weights matrix of known constants, $X$ is an $n$ by $k$ matrix of observations on the exogenous variables (including a constant), $\varepsilon$ is a vector of random error terms, $\beta$ is the $k$ by 1 vector of regression coefficients and $\rho$ is the spatial autoregressive parameter, reflecting the average strength of spatial interdependence, or TFP spillovers, across firms.

The matrix $W$ formalizes the relation between observations that are related in some meaningful dimension, such as geographic, technical or economic. Below we will specify, based on the theoretical spillover channels, which dimensions we incorporate in our model. For expository purposes here, the exact dimension is not relevant. The $(i,j)$th element of $W$, denoted by $w_{ij}$, is the spatial weight, which is non-zero if observation $i$ is 'close' to observation $j$. Often, $i$ and $j$ are then said to be neighbors, or neighboring firms. All diagonal elements of the spatial weight matrix are set to zero: $w_{ii} = 0$. The weighting matrix is generally specified to be row standardized in that $\sum_{j=1}^{n} w_{ij} = 1$. Row-standardization is a common feature in spatial econometric models because it addresses potential problems of non-stationarity across space. Moreover, it implies that the spatial

[2] Also, because this research considers inter-industry spillovers within India, we do not discuss international trade as a possible channel for technology diffusion.

lags can be viewed as weighted averages. Given this normalization, weights are generally the same for all neighbors (1 divided by the number of neighbors) and zero for non-neighbors, or declining as a function of some measure of distance between the observations. Taken together, $Wy$ is called the spatial lag and, in our setting, corresponds to a weighted average of neighboring firms' TFP.

Brueckner (2003) shows that the reduced form corresponding to Equation 1, given that the matrix $(I - \rho W)$ is non-singular with $|\rho| < 1$, is

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} \varepsilon \tag{2}$$

$(I - \rho W)^{-1}$ can be interpreted as the so-called 'spatial multiplier' that links the TFP of one firm to all the independent variables of all neighboring firms in the system. Moreover, Equation 2 illustrates that the TFP of firm $i$ is determined by all error terms in the system. Next to simultaneity, there may also be concern for omitted variable bias. If the error term in equation (1) includes omitted variables that are (i) themselves spatially correlated and (ii) correlated with the spatial lags, then our estimates of interdependence are again biased and inconsistent. As an example, large regional infrastructural improvements, such as the construction of major roads that reduce transports costs, might affect the productivity of many firms in a given region. Alternatively, firms within a given supply chain may all be affected by, for example, (i) regulation aimed at the type of labor that they use intensively, (ii) environmental policies aimed at reducing their carbon dioxide emissions, or (iii) electricity prices that affect their costs of production. All of these omitted factors lead us to wrongfully conclude that there is spatial interdependence in TFP, whereas in fact there is just spatial correlation in the error terms (caused by common shocks to a region or supply-chain). In sum, given $\rho \neq 0$, the spatial lag in Equation 1 is correlated with the error term and can therefore not be consistently estimated using OLS.

### *3.2 Generalized spatial two-stage least squares*

Given the endogeneity of the spatial lag, two main methods have been proposed for consistent estimation. These are Maximum Likelihood Estimation (MLE) (proposed by Ord 1975) and Instrumental Variables Estimation (IVE) (proposed by Kelejian & Prucha 1998). MLE required the assumption of normality of the error terms and is, moreover, computationally very complex especially when dealing with non-symmetric weights (even with moderate sample size), or if the sample size is large, which is our case. Kelejian & Prucha (1998) propose a consistent and computationally simple estimator called the 'Generalized Spatial Two-Stage Least Squares' (S-2SLS) estimator. In particular, they use the spatially lagged exogenous variables, $WX$, as instruments for the endogenous $Wy$. The necessary requirement is that $X$ is a set of exogenous regressors.

In this paper, we use measures of tariffs as exogenous regressors included in the vector $X$. Specifically, following Topalova & Khandelwal (2011), we hypothesize that TFP may depend on two types of tariffs. First of all, a firm's TFP is likely to depend directly on the tariff for goods produced by that firm. We will hereafter refer to this tariff as "output tariff". Secondly, a firm's TFP might depend on the *upstream* tariff for goods that the firm uses as inputs in its production process. We will hereafter refer to this tariff as "input tariff". Using tariffs to construct instruments for the spatial lag may be problematic since, as explained by Topalova & Khandelwal (2011), policymakers may choose to reduce tariffs more in (i) industries that are more productive and hence better able to compete internationally, and (ii) industries that are more politically connected. To assess the importance of these potential sources of endogeneity, Topalova & Khandelwal (2011) perform various tests. First, they show that between 1992 and 1996 (the period of India's Eighth Five-Year Plan) tariff movements were very similar across products, presumably because India had to meet certain externally imposed standards. In contrast, after 1996, tariff movements were less uniform and hence more likely to be endogenous. Second, they show that tariff changes between 1987 and 1997 are not correlated with various politically important industry characteristics in 1987. And finally, they show that regressing tariffs on lagged industry-level productivity yields insignificant results for the period 1989-1996 but not for the period 1997-2001. These results are consistent with an earlier study by Gang & Pandey (1996), who explain their findings in terms of the hysteresis of policy, as trade policy was determined in the Second Five-Year Plan and did not change afterwards. Based on these various tests, Topalova & Khandelwal (2011) conclude that proper identification of the effect of tariffs on productivity is possible only when restricting the analysis to the period immediately before and after the major trade reforms (1989-1996). Following their example, we restrict our analysis to the period 1989-1996. Therefore, given that tariffs are exogenously determined, the Spatial Two-Stage Least Squares estimator will produce a consistent estimate for the autoregressive coefficient, $\rho$.

### *3.3 Bias in non-spatial OLS and spatial OLS*

While the estimation of spatial dependence in itself is interesting, there may also be econometric grounds for including spatial effects. In particular, non-spatial OLS, an OLS regression that ignores potential spatial dependence between observations, may suffer from severe omitted variable bias and efficiency loss. Doreian (1980) is one of the first to compare the results of non-spatial OLS (which suffers from omitted variable bias), spatial OLS (which suffers from simultaneity bias) and spatial MLE using data from two existing non-spatial studies. He finds that non-spatial OLS tends to inflate estimates of non-spatial factors. In other words, non-spatial OLS ignores the spillover effects and attributes the entire effect to a first-order effect. Doreian et al. (1984) perform Monte Carlo experiments and find that non-spatial OLS not only overestimates the estimates of non-spatial regressors, but also underestimates the standard errors, resulting in over-confident conclusions about non-spatial factors. When $\rho$ is larger than 0.1, spatial OLS overestimates $\rho$ and understates its standard error, again resulting in over-confidence but now in the size of the spatial dependence. Land & Deane (1992) compare non-spatial OLS, spatial MLE and spatial 2SLS and find that spatial MLE and spatial 2SLS produce similar results that are superior to non-spatial OLS. Whereas all these studies have used cross-sectional data, Franzese & Hays (2007) compare OLS, spatial OLS, spatial 2SLS and spatial MLE in a panel setting. The results show that any spatial technique significantly improves upon non-spatial OLS, but that a tradeoff must be made between biased but efficient spatial OLS and consistent but less efficient spatial 2SLS. Spatial MLE produces good estimates, but is computationally demanding if not impossible with fairly large sample size. Moreover, spatial 2SLS seems to produce the most accurate standard errors.

The upward bias in non-spatial OLS induced by omitting a spatial lag when spatial interdependence actually exists is simply the omitted variable bias that arises because the omitted variable, the spatial lag, is correlated with the independent variables and a determinant of the dependent variable. The well-known formula for omitted variable bias (OVB) is:

$$OVB = (X'X)^{-1}X'(WY)\rho \tag{3}$$

where $WY$ is a vector of omitted variables and $\rho$ is the true population parameter for the effect of $WY$ on $Y$, given $X$. Thus the size of bias increases with the strength of the spatial interdependence, measured by $\rho$. A priori, it is not possible to determine the magnitude of the bias.

### *3.4 Spatial weights*

Having defined the basic regression model and estimation technique, we will now specify the full model and the different spatial weights that capture the possible spillover channels we are interested in. First, we specify the spatial autoregressive model from Equation 1 for the panel case:

$$TFP_{i,t} = \rho \sum_j \left(w_{ij,t} * TFP_{j,t}\right) + X_{i,t}\beta + \alpha_i + \lambda_t + \varepsilon_{i,t} \tag{4}$$

where we have added firm and time subscripts, firm-fixed effects ($\alpha_i$) and time-fixed effects ($\lambda_t$). Our data set consists of $N$ firms and the subscripts $i$ and $j$ index the firms that 'receive' the spillovers and the firms that 'supply' the spillovers, respectively. We changed notation of the spatial lag to clarify the constructed measure of neighboring TFP. Note that there are no spillovers within firms, $w_{ii} = 0$, so that we effectively sum over all $j \neq i$. From the spillover literature we identified three spillover channels: (1) observation, (2) labor mobility, and (3) intermediate inputs.

As explained in Section 3.2, to address the endogeneity of the spatial lag, $\sum_j (w_{ij,t} * TFP_{j,t})$, we construct instrumental variables using measures of tariffs. Specifically, the 2SLS estimates we report below are obtained by instrumenting for $\sum_j (w_{ij,t} * TFP_{j,t})$, using two (temporally lagged) instruments: $\sum_j (w_{ij,t-1} * Output\ tariff_{j,t-1})$ and

$\sum_j (w_{ij,t-1} * Input\ tariff_{j,t-1})$.[3] In the regression tables we will refer to these two instruments as "instrument output tariff $(t-1)$" and "instrument input tariff $(t-1)$", respectively. In all our estimations, we control for $Output\ tariff_{i,t-1}$ and $Input\ tariff_{i,t-1}$ (note the subscript $i$ instead of $j$, which was used in the case of the instruments). This is to avoid possible omitted variable bias due to any spatial correlation in tariffs.

### *3.4.1 Channel 1: observation*

To test for spillovers through observation of neighboring firms, we construct the following spatial weights:

$$w_{ij,t}^{D0} = \frac{\delta_{ij,t}^{0}}{\sum_j \delta_{ij,t}^{0}}$$

where $\delta_{ij,t}^{0} = 1$ if $distance_{ij,t} \leq 50km$ and $TFP_{j,t}$ is non-missing, and 0 otherwise.

To test for the presence of spillovers from firms located further than 50km away, we also consider the spatial weight for these firms:

$$w_{ij,t}^{D1} = \frac{\delta_{ij,t}^{1}}{\sum_j \delta_{ij,t}^{1}}$$

where $\delta_{ij,t}^{1} = 1$ if $distance_{ij,t} > 50km$ and $TFP_{j,t}$ is non-missing, and 0 otherwise.

While distance is a non-time varying measure, the time-subscript enters because we have an unbalanced panel dataset such that we do not have TFP observations for all firms for all years. Therefore, we set $\delta_{ij,t}^{0} = \delta_{ij,t}^{1} = 0$ if firm $j$ has no TFP observation in year $t$. The

[3] The variable $Output\ tariff_j$ corresponds to the tariff for goods produced by firm $j$. The variable $Input\ tariff_j$ instead corresponds to the weighted average tariff for goods produced by other firms $k \neq i, j$, where the weights correspond to the importance of firm $k$ as a supplier of inputs to firm $j$.

spatial lag given by weights $w_{ij,t}^{D0}$ ($w_{ij,t}^{D1}$) is now the simple average of the TFP of firms located within (outside) a 50km radius of firm $i$. The regression model to be estimated is the following:

$$TFP_{i,t} = \rho^{D0} \sum_{j} \left( w_{ij,t}^{D0} * TFP_{j,t} \right) + \rho^{D1} \sum_{j} \left( w_{ij,t}^{D1} * TFP_{j,t} \right) + X_{i,t}\beta + \alpha_i + \lambda_t + \varepsilon_{i,t} \qquad (5)$$

We now have two spatial autoregressive coefficients that capture the average strengths of the TFP spillovers from firms located in a 50km radius and located outside the 50km radius, respectively.

### *3.4.2 Channel 2: labor mobility*

To test for the presence of spillovers through labor mobility, we use state-level variation in labor mobility induced by state-level labor regulation. Specifically, we test whether there are more TFP spillovers between firms located in a high-labor mobility state, compared to firms located in states with low labor mobility. We create geographic weights similar to Channel 1, but focus on firms located in the same state (rather than the 50km radius):

$$w_{ij,t}^{S0} = \frac{\sigma_{ij,t}^{0}}{\sum_j \sigma_{ij,t}^{0}}$$

with $\sigma_{ij,t}^{0} = 1$ if $state_i = state_j$ and $TFP_{j,t}$ is non-missing, and 0 otherwise.

The weights given to firms not in the same state are given by:

$$w_{ij,t}^{S1} = \frac{\sigma_{ij,t}^{1}}{\sum_j \sigma_{ij,t}^{1}}$$

with $\sigma_{ij,t}^{1} = 1$ if $state_i \neq state_j$ and $TFP_{j,t}$ is non-missing, and 0 otherwise.

The state-based spatial lag $w_{ij,t}^{S0}$ ($w_{ij,t}^{S1}$) is the simple average of the TFP of firms located in the same (a different) state as firm $i$. The regression model to be estimated is the following:

$$TFP_{i,t} = \rho^{S0} \sum_j \left(w_{ij,t}^{S0} * TFP_{j,t}\right) + \rho^{S1} \sum_j \left(w_{ij,t}^{S1} * TFP_{j,t}\right) + X_{i,t}\beta + \alpha_i + \lambda_t + \varepsilon_{i,t} \qquad (6)$$

We estimate two spatial autoregressive coefficients that capture the average strengths of the TFP spillovers from firms located in the same state and located in a different state, respectively. We estimate equation (6) for two subsamples, based on their level of labor mobility, to test for labor mobility spillovers. We expect no spillovers in the subsample of firms located in states with low labor mobility and positive spillovers in states with high labor mobility.

*3.4.3 Channel 3: intermediate inputs*

To test for spillovers through the use of intermediate inputs from upstream industries, we use industry-level input coefficients from the national Input-Output (IO) table to proxy for firm-firm linkages, as we unfortunately do not have a firm-level IO-table.

Consider entry $\alpha_{kh}$ from the IO table which gives the value of inputs supplied by industry $h$ as a share of total value of output of industry $k$. Because it is possible that $\alpha_{kk} \geq 0$, we set $\alpha_{kh} = 0$ for $k = h$.[4] We divide $\alpha_{kh}$ by the number of firms in industry $h$ to proxy for the value of inputs supplied by *those* firms $j \neq i$ that are located in industry $h$, as a share of total value of output of firm $i$ (in industry $k$): $\alpha_{ij,t} = \frac{\alpha_{kh}}{N_{h,t}}$. Thus, the input-output weights are given by:

$$w_{ij,t}^{IO} = \frac{\alpha_{ij,t}}{\sum_j \alpha_{ij,t}}$$

where subscript $j$ corresponds to any firm $j \neq i$ (not just the firms in industry $h$). The intuition behind these spatial weights is that if TFP is incorporated in intermediate inputs and this forms a relevant channel for TFP spillovers, we expect more spillovers from a

[4] Note that $\sum_h \alpha_{kh} + VA_k = 1$, where VA=value added. Our dataset is comprised of manufacturing firms, so we do not consider $\alpha_{kh}$ for industries $h$ that are non-manufacturing, such as Agriculture, Mining and Services.

given upstream firm if its input share in the downstream firm's production is larger. The regression to test this is:

$$TFP_{i,t} = \rho^{IO} \sum_{j} \left( w_{ij,t}^{IO} * TFP_{j,t} \right) + X_{i,t}\beta + \alpha_i + \lambda_t + \varepsilon_{i,t} \tag{7}$$

Unfortunately, the practice of row-standardization has the theoretical implication that the share of total inputs in total production of firm $i$ does not matter for the size of the spillover. Because we expect more TFP spillovers to firms that use relatively more inputs, we estimate equation (7) for subsamples of firms with a *below* median share of inputs in total output and firms with an *above* median share of inputs.

In short, we have 5 spatial weight matrices and three regression equations. The statistical significance (or insignificance) of the spatial autoregressive coefficient estimate for each spatial lag will indicate the presence (or absence) of TFP spillovers working through the respective channels.

## 4. Data and Descriptive Statistics

The firm-level data for this study was collected from the Prowess database, provided by the Centre for Monitoring the Indian Economy (https://prowess.cmie.com). Prowess contains time-series data on the financials of a large number of Indian publicly listed and unlisted companies from 1989/90 and is updated every day. While its coverage is comprehensive, for example the total production of all companies in Prowess is more than 80 percent of India's GDP, the database is not a census of all companies in India, nor is it a random sample. Because Prowess is filled with data from publicly available documents, new firms may enter the database and others may exit the database because of data availability issues rather than because of true entries and exits in the economy. Our sample contains financial and non-financial information on 4000 individual manufacturing companies. Prowess uses the four-digit NIC (2008 revision) classification to categorize companies by industry. There are 116 industries in our sample. The tariff data, reported at the four-digit NIC level, are obtained from Topalova and Khandelwal (2011).[5] Furthermore, we use the Indian Input-Output Transactions Table from 1993-1994 to calculate the spatial weights for each industry pair and approximate the input-output-based spatial weight for each firm pair. The distance-based spatial lag uses headquarter zipcode data from Prowess. We match these with latitude and longitude coordinates to calculate distance between each firm.

Tables 1 and 2 below report summary statistics and cross-correlations for the variables used in estimation. Appendix A contains additional figures and tables that illustrate the geographical distribution and density of the firms in our sample across India. Appendix B provides more detailed information on the data used in our analysis.

Table 1: Summary statistics for the variables used in estimation

| Variable | Observations | Mean | St. Dev. | Minimum | Maximum |
|---|---|---|---|---|---|
| TFP (log) | 17103 | -0.01 | 0.68 | -6.53 | 8.75 |
| Spatial lag (IO weights) | 17103 | -0.03 | 0.11 | -0.64 | 0.38 |
| Output tariff, $t-1$ | 18272 | 0.74 | 0.29 | 0.10 | 3.26 |
| Input tariff, $t-1$ | 18385 | 0.31 | 0.10 | 0.02 | 0.64 |
| Age | 18385 | 21.8 | 19.12 | 0 | 117 |
| Neighbors (50 km) | 17103 | 162.67 | 169.18 | 0 | 608 |
| Distance to port (km) | 18105 | 323.34 | 340.23 | 6.92 | 1180.52 |

Table 2: Pairwise correlations

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1: TFP (log) | 1.00 | | | | | |
| 2: Output tariff, $t-1$ | -0.01 | 1.00 | | | | |
| 3: Input tariff, $t-1$ | -0.09 | 0.51 | 1.00 | | | |
| 4: Spatial lag (IO weights) | 0.08 | -0.26 | -0.34 | 1.00 | | |
| 5: Instrument input tariff | -0.06 | 0.59 | 0.75 | -0.50 | 1.00 | |
| 6: Instrument output tariff | -0.06 | 0.50 | 0.79 | -0.39 | 0.90 | 1.00 |

## 5. Estimation Results

### *5.1 Channel 1: observation*

[5] We are grateful to Petia Topalova and Amit Khandelwal for their generosity in sharing their data and to Petia Topalova for providing a detailed explanation of the data.

Table 3 reports the results of estimating equation (5) and testing whether there are spillovers between firms that are geographically close. Column (1) reports the results from an OLS regression in which we include both the spatial lag that captures average TFP of firms located *within* a 50 km radius and the spatial lag that captures average TFP of firms located *outside* this 50 km radius. In column (2) we run the same OLS regression but

Table 3: Estimation results – spillovers through observation of neighboring firms

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Panel A: Second stage** | | | | |
| Spatial lag $\leq 50km(t)$ | -0.01 (0.03) | 0.02 (0.02) | 0.07 (0.17) | 0.06 (0.17) |
| Spatial lag $> 50km(t)$ | -3.60*** (1.06) | | 1.95 (2.71) | |
| Output tariff $(t-1)$ | 0.06 (0.04) | 0.06 (0.04) | 0.05 (0.04) | 0.06 (0.04) |
| Input tariff $(t-1)$ | -0.44** (0.19) | -0.44** (0.18) | -0.44** (0.18) | -0.44** (0.18) |
| Firm fixed effects | YES | YES | YES | YES |
| Year fixed effects | YES | YES | YES | YES |
| Method | OLS | OLS | 2SLS | 2SLS |
| Number of observations | 16484 | 16484 | 16484 | 16484 |
| $R$-squared within | 0.014 | 0.012 | 0.009 | 0.011 |
| **Panel B: First stage for spatial lag $\leq 50km(t)$** | | | | |
| Output tariff $(t-1)$ | | | -0.01 (0.01) | -0.01 (0.01) |
| Input tariff $(t-1)$ | | | 0.07 (0.05) | 0.05 (0.05) |
| Instrument output tariff $(t-1)$ for spatial lag $\leq 50km(t)$ | | | 0.22*** (0.06) | 0.23*** (0.05) |
| Instrument input tariff $(t-1)$ for spatial lag $\leq 50km(t)$ | | | -0.48*** (0.15) | -0.53*** (0.14) |
| Instrument output tariff $(t-1)$ for spatial lag $> 50km(t)$ | | | -0.15 (1.06) | |
| Instrument input tariff $(t-1)$ for spatial lag $> 50km(t)$ | | | 15.38*** (2.92) | |
| Number of observations | | | 16484 | 16484 |
| $F$-stat | | | 22.94 | 12.33 |
| **Panel C: First stage for spatial lag $> 50km(t)$** | | | | |
| Output tariff $(t-1)$ | | | 0.00 (0.00) | |
| Input tariff $(t-1)$ | | | -0.00* (0.00) | |
| Instrument output tariff $(t-1)$ for spatial lag $\leq 50km(t)$ | | | 0.00*** (0.00) | |
| Instrument input tariff $(t-1)$ for spatial lag $\leq 50km(t)$ | | | -0.01*** (0.00) | |
| Instrument output tariff $(t-1)$ for spatial lag $> 50km(t)$ | | | 0.46*** (0.10) | |
| Instrument input tariff $(t-1)$ for spatial lag $> 50km(t)$ | | | -2.64*** (0.34) | |
| Number of observations | | | 16484 | |
| $F$-stat | | | 16.36 | |

Notes: Table 3 reports the results from estimating equation (5). The dependent variable is logged TFP. In all columns, robust standard errors (clustered by firm and industry-year) are reported in parentheses. The reported $F$-

stat corresponds to the first stage $F$ test of excluded instruments. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

including the first spatial lag ($\leq 50km$) only. Columns (3) and (4) report the results from estimating the same specifications as in columns (1) and (2) but now, instead of OLS, applying a 2SLS estimator. For each spatial lag, we use the two variables "instrument output tariff ($t-1$)" and "instrument input tariff ($t-1$)" as instrumental variables. The second-stage results of the 2SLS procedure are reported in Panel A, while the first-stage results are reported in Panels B and C.

The results seem to indicate that TFP does not significantly diffuse across geographic space, as all but one of the spatial lag coefficients are statistically insignificant. The only exception is the coefficient for the second spatial lag ($> 50km$), which is negative and significant at the 1 percent level, but this result disappears once we address endogeneity concerns in column (3). The relevant instrumental variables for the spatial lags enter with coefficients that are statistically significant and have the expected sign. The $F$-statistic is in all three cases above 10, suggesting that the instruments are not weak.

### *5.2 Channel 2: labor mobility*

Table 4 reports the results of estimating equation (6) and testing whether there are spillovers between firms that are located in the same state (or in different states) and whether any such spillovers are stronger for states with high labor mobility. Columns (1) and (2) report the results from OLS regressions on subsamples of states with high labor mobility and states with low labor mobility. Columns (3) and (4) report the results from estimating the same specification for the same two subsamples as in columns (1) and (2) but now, instead of OLS, applying a 2SLS estimator. As before, for each spatial lag, we use the two variables instrument output tariff ($t-1$) and instrument input tariff ($t-1$) as instrumental variables. The second-stage results of the 2SLS procedure are reported in Panel A, while the first-stage results are reported in Panels B and C.

The only spatial lag that enters with a coefficient that is significant at the 5 percent level is the spatial lag for firms in the same state in the 2SLS regression (column 3). The coefficient is positive which suggests the presence of positive TFP spillovers within Indian states. However, we cannot be very confident in this result, given the first-stage insignificance of

Table 4: Estimation results – spillovers through labor mobility

| | (1) high mob. | (2) low mob. | (3) high mob. | (4) low mob. |
|---|---|---|---|---|
| **Panel A: Second stage** | | | | |
| Spatial lag same state ($t$) | -0.08 (0.24) | -0.10 (0.09) | 2.26** (0.98) | 0.07 (0.21) |
| Spatial lag other state ($t$) | -4.87* (2.49) | -0.65 (1.12) | 0.75 (8.11) | -0.88 (1.38) |
| Output tariff ($t-1$) | 0.08 (0.08) | 0.05 (0.04) | 0.10 (0.08) | 0.05 (0.04) |
| Input tariff ($t-1$) | -0.72** (0.29) | -0.25 (0.21) | -0.79*** (0.30) | -0.25 (0.21) |
| Firm fixed effects | YES | YES | YES | YES |
| Year fixed effects | YES | YES | YES | YES |
| Method | OLS | OLS | 2SLS | 2SLS |
| Number of observations | 4808 | 10040 | 4808 | 10040 |
| $R$-squared within | 0.020 | 0.011 | 0.06 | 0.01 |
| **Panel B: First stage for spatial lag same state** | | | | |
| Output tariff ($t-1$) | | | -0.01 (0.01) | -0.01 (0.01) |
| Input tariff ($t-1$) | | | 0.04 (0.04) | 0.04 (0.03) |
| Instrument output tariff ($t-1$) for spatial lag same state | | | 0.04 (0.17) | 0.49 (0.34) |
| Instrument input tariff ($t-1$) for spatial lag same state | | | -0.10 (0.43) | -2.22*** (0.75) |
| Instrument output tariff ($t-1$) for spatial lag other state | | | -3.69* (2.13) | -5.18*** (1.36) |
| Instrument input tariff ($t-1$) for spatial lag other state | | | -0.52 (8.44) | 4.74 (3.44) |
| Number of observations | | | 4808 | 10040 |
| $F$-stat | | | 4.06 | 62.92 |
| **Panel C: First stage for spatial lag other state** | | | | |
| Output tariff ($t-1$) | | | -0.00 (0.00) | 0.00 (0.00) |
| Input tariff ($t-1$) | | | -0.01 (0.00) | -0.00 (0.00) |
| Instrument output tariff ($t-1$) for spatial lag same state | | | -0.02 (0.01) | 0.08*** (0.01) |
| Instrument input tariff ($t-1$) for spatial lag same state | | | -0.11** (0.04) | -0.16*** (0.02) |
| Instrument output tariff ($t-1$) for spatial lag other state | | | -0.04 (0.20) | 1.89*** (0.07) |
| Instrument input tariff ($t-1$) for spatial lag other state | | | 0.41 (0.86) | -4.35*** (0.14) |
| Number of observations | | | 4808 | 10040 |
| $F$-stat | | | 8.31 | 353.14 |

Notes: Table 4 reports the results from estimating equation (6). The dependent variable is logged TFP. In all columns, robust standard errors (clustered by firm and industry-year) are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

the relevant instrumental variables. Also, the $F$-statistic is 4.06, suggesting that the instruments are indeed weak.

### *5.3 Channel 3: intermediate inputs*

Table 5 reports the results of estimating equation (7) and testing whether there are spillovers from upstream (intermediate-input supplying) firms to downstream (intermediate-input buying) firms and whether any such spillovers are stronger for firms for which inputs are important relative to output. Columns (1) and (2) report the results from OLS regressions on subsamples of firms with low and high levels of inputs. Columns (3) and (4) report the results from estimating the same specification for the same two subsamples as in columns (1) and (2) but now, instead of OLS, applying a 2SLS estimator. As before, for the spatial lag, we use the two variables instrument output tariff $(t-1)$ and instrument input tariff $(t-1)$ as instrumental variables. The second-stage results of the 2SLS procedure are reported in Panel A, while the first-stage results are reported in Panels B and C.

The spatial lag enters with a statistically significant coefficient only in the OLS regression of column (2) for firms with high levels of inputs. The coefficient is positive which suggests the presence of positive TFP spillovers. However, we cannot be very confident in this result, as it disappears entirely when we address endogeneity in column (4). The relevant instrumental variables for the spatial lags enter with coefficients that are statistically significant and have the expected sign. The $F$-statistic is in both cases above 10, suggesting that the instruments are not weak.

Table 5: Estimation results – spillovers through intermediate inputs

| | (1) low inputs | (2) high inputs | (3) low inputs | (4) high inputs |
|---|---|---|---|---|
| **Panel A: Second stage** | | | | |
| Spatial lag input suppliers ($t$) | 0.00 (0.10) | 0.26*** (0.07) | -0.10 (0.20) | -0.00 (0.123) |
| Output tariff ($t-1$) | 0.08 (0.06) | 0.09** (0.04) | 0.07 (0.06) | 0.09** (0.04) |
| Input tariff ($t-1$) | -0.33 (0.25) | -0.92*** (0.24) | -0.31 (0.25) | -0.80*** (0.22) |
| Firm fixed effects | YES | YES | YES | YES |
| Year fixed effects | YES | YES | YES | YES |
| Method | OLS | OLS | 2SLS | 2SLS |
| Number of observations | 8147 | 7845 | 8147 | 7845 |
| $R$-squared within | 0.009 | 0.035 | 0.008 | 0.028 |
| **Panel B: First stage for spatial lag input suppliers** | | | | |
| Output tariff ($t-1$) | | | -0.00 (0.02) | 0.07** (0.03) |
| Input tariff ($t-1$) | | | 0.04 (0.07) | -0.31 (0.19) |
| Instrument output tariff ($t-1$) for spatial lag input suppliers | | | 1.07*** (0.17) | 1.30*** (0.24) |
| Instrument input tariff ($t-1$) for spatial lag input suppliers | | | -2.04*** (0.41) | -3.06*** (0.53) |
| Number of observations | | | 8147 | 7845 |
| $F$-stat | | | 19.17 | 17.02 |

Notes: Table 5 reports the results from estimating equation (7). The dependent variable is logged TFP. In all columns, robust standard errors (clustered by firm and industry-year) are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

## 6. Conclusions

It is often argued that the positive effects of policies or institutional reforms on the TFP of firms are amplified by positive TFP spillovers to other firms. In this paper we have empirically estimated the strength of such inter-firm TFP spillovers, using the Indian period of trade liberalization as an exogenous source of variation in TFP. We have focused our attention to three possible channels, namely observation, labor mobility and intermediate input use. We find no systematic evidence in support of TFP spillovers across firms.

## References


Amiti, M., and J. Konings, 2007. “Trade Liberalization, Intermediate Inputs, and Productivity: Evidence from Indonesia.” *American Economic Review* 97: 1611-38.

Anselin, L. (1988). Spatial Econometrics: Methods and Models. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Balakrishnan, P., K. Pushpangadan, and M. Babu (2000). Trade liberalization and productivity growth in manufacturing: Evidence from firm-level panel data. Economic and Political Weekly 35 (41), 3679-3682.

Blalock, G., and P. Gertler. 2008. “Welfare Gains from Foreign Direct Investment through Technology Transfer to Local Suppliers.” *Journal of International Economics* 74: 402-21.

Brueckner, J. K. (2003). Strategic interaction among governments: An overview of empirical studies. International Regional Science Review 26 (2), 175-188.

Doreian, P. (1980). Linear models with spatially distributed data: Spatial disturbances or spatial effects. Sociological Methods & Research 9 (1), 29-60.

Doreian, P., K. Teuter, and C.-H. Wang (1984). Network autocorrelation models: Some monte carlo results. Sociological Methods & Research 13, 155-200.

Franzese, R. and J. Hays (2007). Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. Political Analysis 15, 140-164.

Franzese, R.J., and J.C. Hays. 2008. “Empirical Models of Spatial Interdependence.” In *Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford: Oxford University Press.

Gang, I. N. and M. Pandey (1996). Trade protection in india: Economics vs. politics? Working Paper 9616.

Goerg, H., and E. Strobl. 2005. “Spillovers From Foreign Firms through Worker Mobility: An Empirical Investigation.” *Scandinavian Journal of Economics* 107: 693-709.

Javorcik, Beata S. 2004. “Does Foreign Direct Investment Increase the Productivity of Domestic Firms? In Search of Spillovers through Backward Linkages.” *American Economic Review* 94: 605–27.

Kelejian, H. and I. Prucha (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. Journal of Real Estate Finance and Economics 17 (1), 99-121.

Keller, W. 2004. "International Technology Diffusion." *Journal of Economic Literature* 42: 752-82.

Levinsohn, J. and A. Petrin (2003). Estimating production functions using inputs to control for unobservables. Review of Economic Studies 70 (2), 317-342.

Paz, Lourenco. 2012. "Trade Liberalization and Inter-Industry Productivity Spillovers: An Analysis of the 1989-1998 Brazilian Trade Liberalization Episode." MPRA Paper No. 38859.

Topalova, Petia, and Amit Khandelwal. 2011. "Trade Liberalization and Firm Productivity: The Case of India." *Review of Economics and Statistics* 93: 995-1009.

Xu, B. 2000. "Multinational Enterprises, Technology Diffusion, and Host Country Productivity Growth." *Journal of Development Economics* 62: 477-93.

## Appendix A: Additional Information (separate document)

## Appendix B: Data

### *B1 Variables used in TFP calculation*

Firm-level TFP is estimated using the following firm-level variables from the Prowess database: sales, change in inventory, gross fixed assets, compensation to employees, power and fuel expenses, and raw material expenses. These variables are used to construct measures of output, capital, labor, power and fuel input, and raw material inputs. We use the period 1988-2002 to estimate the coefficients of the production function for each industry using the methodology proposed by Levinsohn and Petrin (2003). Once we have obtained the coefficients, we can calculate TFP as the residual. While our regression estimates in this paper are based on the period 1998-1996, we use a longer time span to obtain consistent estimates of the productivity function. Once we have estimated firm-level TFP, we calculate a TFP index to make the firm-level TFP estimates comparable across industries. This is done by substracting the average industry's productivity in 1997 from the estimated firm-level TFP.

#### Output

The value of output is calculated as the difference between total sales and change in inventory and deflates this using two-digit NIC98 industry-specific wholesale price indices (WPI). The wholesale price index is obtained from published by the Office of the Economic Advisor to the Ministry of Commerce and Industry (http://www.eaindustry.nic.in/) and were matched to the two-digit NIC98.

#### Capital

The measure of capital input is constructed from the data on gross fixed assets and depreciation using a modified perpetual inventory method (Balakrishnan et al. 2000). This measure uses the perpetual inventory method (PIM) while taking into account that capital is recorded at historic, and not replacement cost. Specifically, consider $GFA_t^H$, gross fixed assets at time $t$ recorded in the company's books at historic costs:

$$GFA_t^H = P_t I_t + P_{t-1} I_{t-1} + \cdots + P_{t-n} I_{t-n}$$

Where $P_t$ denotes the price index at time $t$, $I_t$ is the investment made at time $t$ and $n$ is the age of the firm. For a firm, it's current gross fixed assets is the sum of previous investments (all recorded at historic prices) plus this year's investment (recorded at current costs). In other words, while this year's investment is recorded at replacement costs, previous years' investments are not. For productivity estimates, it would be more informative to have a measure of gross fixed assets at *replacement* costs:

$$GFA_t^R = P_t I_t + P_t I_{t-1} + \cdots + P_t I_{t-n}$$

Assuming a fixed inflation and growth rate given by $\pi = \frac{P_t - P_{t-1}}{P_{t-1}}$ and $g = \frac{I - I_{t-1}}{I_{t-1}}$, the revaluation factor, defined as the ratio of the value of assets at replacement costs to the value of assets at historic costs, is given by:

$$R_t^R = \frac{GFA_t^R}{GFA_t^H} = \frac{P_t I_t (1 + \frac{1}{(1+g)} + \frac{1}{(1+g)^2} + \cdots + \frac{1}{(1+g)^n}}{P_t I_t (1 + \frac{1}{(1+g)(1+\pi)} + \frac{1}{[(1+g)(1+\pi)]^2} + \cdots + \frac{1}{[(1+g)(1+\pi)]^n}}$$

which can be rewritten to:

$$\frac{[(1+g)^{n+1} - 1](1+\pi)^n [(1+g)(1+\pi) - 1]}{g([(1+g)(1+\pi)]^{n+1} - 1)}$$

Using the revaluation factor and price index for base year $t$, we convert capital (recorded at historic costs) at current price in the base year to capital at replacement costs at constant prices. Subsequent year's investment, $I_{t+1} = \frac{GFA_{t+1}^H - GFA_t^H}{P_{t+1}}$, is added to obtain capital at replacement costs at constant prices in the next years.

Assuming the capital stock has finite economic life (20 years), the lifetime of capital stock, $n$, is set to $\min(20, age)$ where $age$ is the age of the firm (capital stock) in the base year. Data on gross rate of capital formation was obtained from the World Bank Development Indicator and the inflation rate of capital was constructed from India's National Account Statistics publications on Gross Capital Formation at current and constant prices. To maximize the number of observations in the base year, we take 1997 as the base year.

Because we deal with an unbalanced panel, we interpolate $GFA_t = \frac{GFA_{t-1} + GFA_{t+1}}{2}$ if $GFA_t$ is missing. This ensures we can use the PIM also if we have a missing observation in the middle of a company's time series. Once we have calculated capital at replacement costs for all years (original and interpolated), we use only the capital at replacement costs for year observations in which the capital (at historic cost) was non-missing originally.

Finally, we are left with the problem that some firms have observations only prior to the base year 1997, or only after the base year 1997. This is about 15 percent of the firms. To nevertheless calculate revalued capital for these firms for the entire period 1988-2002, we take base year 1996 for those firms with latest observation in 1996, and 1995 when latest year of observation is 1995 etc.

**Labor**

Because of a lack of data on the number of employees per firm, labor input is proxied by compensation to employees. This measure includes salaries and wages, but also bonuses and staff training expenditures. The advantage of using this measure compared to the number of workers is that it reflects both worker quality and quantity. The deflator used is constructed from the WPI on all commodities.

**Raw material inputs**

Due to a lack of information on physical quantities, we use deflated raw material expenses as proxy for raw material inputs. The deflator used is constructed from the WPI series on primary articles.

**Power and fuel input**

Deflated raw material expenses are used as a proxy power and fuel input. The deflator used is constructed from the WPI series on fuel, power, light and lubricants.

### *B2 Details on the construction of output and input tariffs*

We use tariff data from Topalova & Khandelwal (2011). The effect of trade liberalization on firm's total factor productivity may come through two types of tariffs. First, the output tariff is the tariff foreign firms in a particular industry face when exporting their products to India. A lower tariff thus means lower trade protection of domestic firms or conversely, a higher level of trade liberalization. Trade protection may also come in the form on input tariffs, which is defined as the tariff on the intermediate inputs used in the production of the industry's final goods.

Topalova & Khandelwal (2011) constructed a database of annual tariff data for 1987 to 2001 at the six-digit level of the Indian Trade Classification Harmonized System (HS) code. These 5,045 HS6 product lines were matched to 116 industries at the four-digit NIC1998, to calculate industry-level tariffs. Since then, Prowess has updated its industry classification to NIC08, so in order to combine the tariff data with firm-level data, we further matched the industry-level tariffs from NIC1998 to NIC2008 to obtain output tariffs. The input tariff for a given industry $j$ is calculated using:

$$Inputtariff_{j,t} = \sum_{s} \alpha_{js} Outputtariff_{s,t}$$

which corresponds to Eq.4 in Topalova and Khandelwal (2011), and where $\alpha_{js}$ is the share of input of product $s$ in the value of output of industry $j$. These weights are obtained from the product-industry Input-Output table as explained in the data description section. The industry classification of the Input-Output table is, however, different from the NIC classification. Since there exist no correspondence tables between IO and NIC. Note that we do not use exactly the same matching as \cite{Topalova2011} due to the updated National Industry Classification used in Prowess.