Xi'an Jiaotong-Liverpool University

Research Institute for Economic Integration

Working Papers Series

(RIEI WP series)

# *Agglomeration and Technological Spillovers: Firm-Level Evidence from China's Electric Apparatus Industry*

by

***Ming He, Yang Chen, Ron Schramm***

**Working paper number: 2016-02**

*Notes:*
*Original version: 3 Mar 2016*

# Agglomeration and Technological Spillovers: Firm-Level Evidence from China's Electric Apparatus Industry

Ming He, Yang Chen, and Ron Schramm*
IBSS, Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu Province 215123, China

February 25, 2016

**Abstract**

We use a spatial autoregressive model to study the determinants of firm-level productivity growth using longitudinal data on China's electric apparatus industry over the period of 1999-2007. Factors considered include technological spillover, R&D and export behavior, agglomeration economies, and public expenditure. We propose modifications to Kelejian and Prucha's (1998) FE-2SLS procedure and Mutl and Pfaffermayr's (2011) RE-FG2SLS procedure to cope with the technical difficulties with our unbalanced panel. Statistical evidence strongly favors the fixed effects model over the random effects model. According to our estimates, there are large and significant technological spillovers among firms. Individually, firms benefit from their own R&D and export activities. Market competition and public expenditure in the local and neighboring jurisdictions are found to be important determinants to productivity. Our model also provides direct evidence that the technological spillover effects attenuate rapidly in spatial distance. Finally, the inter-regional spillover effects are found to be more pronounced and more significant on urban districts or jurisdictions with smaller geographical areas. Geographic proximity to neighbors and special administrative role jointly contribute to this observation.

## 1 Introduction

Productivity isn't everything, but in the long run it is almost everything.

—Paul Krugman

The sources of total factor productivity (TFP) growth have been widely debated in the literature from both macro and micro perspectives (see Syverson, 2011 for

*Ming He: ming.he@xjtlu.edu.cn, +86-0512-8816-1712; Yang Chen: yang.chen@xjtlu.edu.cn, +86-0512-8816-1178; Ron Schramm: ron.schramm@xjtlu.edu.cn, +86-0512-8816-7129.

a survey and the references therein). Productivity differences at the micro level are usually attributed to various factors such as access to foreign markets (Clerides, Lach, and Tybout, 1998), firm-level innovation (Griffith, Redding, and Van Reenen, 2004), ownership structure (Aitken and Harrison, 1999), and external market conditions such as competition (Nickell, 1996) and specialization (Glaeser, Kallal, Scheinkman, and Shleifer, 1992), among others. Early literature implicitly assumes the absence of interactions among spatially dispersed firms. A major departure from this tradition was Rosenthal and Strange (2003) who conduct a micro-level analysis of the geographic content of agglomeration economies. Since then the producer practices that may have spillover effects within and across the geographic boundary on the productivity levels of others come into the limelight. These externalities are discussed in the context of classic agglomeration mechanisms such as input sharing, knowledge or technology spillover. Higher productivity correlations among "nearby" producers are usually tested by regressing the productivity level for a specific firm on exogeneous variables such as R&D expenditure of other firms or the presence of foreign investment (Wei and Liu, 2006; Griffith, Redding, and Van Reenen, 2004; Keller and Yeaple, 2009). However, two research questions, although important, are not answered by this strand of literature: (1) to what extent firm-level productivity growth could be explained by that of neighboring firms; and (2) whether the spillovers generated from agglomeration through technology linkages or input sharing attenuate with geographic distance and by how much.

By answering these two questions, this paper contributes to the micro-level productivity analysis in several aspects. First, with county-level geographic information, we are able to conduct a rigorous spatial analysis on firm-level data extracted from the *China Industry Survey* dataset over the period of 1999-2007. Second, we implement a bootstrapped version of Kelejian and Prucha's (1998) FE-2SLS procedure that accounts for heteroskedasticity and spatial autocorrelation in the error term. We also present empirical evidence that our method is superior to the RE-FG2SLS estimator adopted by some recent studies (e.g. Baltagi, Egger, and Kesina, 2015). Third, our empirical specification allows us to study separately the spillovers from different types of neighbors: those in the same region (intra-region) and those in nearby regions (inter-region). Similarly, we are able to gauge inter-regional spillovers of local market conditions.

Our spatial analysis controls for sources of productivity growth identified by the literature, especially Chinese studies. These include R& D activity, export behavior, and local market conditions.

There is a long literature linking productivity with R&D activity (Hall, Mairesse, and Mohnen, 2010). Among others, Doraszelski and Jaumandreu (2013) find evidence of productivity growth as the consequence of R&D expenditures using a panel of Spanish firms. Aw, Roberts, and Xu (2009) find bidirectional causality between R&D and productivity growth among Taiwanese electronics exporters. Chinese studies such as Hu and Jefferson (2004), Hu, Jefferson, and Jinchang (2005), and Boeing, Mueller, and Sandner (2015) report similar findings using Chinese data. Our study shows that a 10% increase in R&D activity, measured by the share of new products in the gross output, contributes to an average of 1.9% increase in productivity in our sample.

The marketization reform (Jefferson and Su, 2006) followed by China's accession to WTO in 2001 results in deep foreign market engagement by Chinese firms. The literature on productivity suggests two mechanisms by export behavior: self-selection and learning by exporting (Syverson, 2011). Even though the theoretical arguments (Melitz, 2003; Grossman and Helpman, 1991) all suggest a positive effect, the empirical evidence based on Chinese data is by far inconclusive. Bao, Huang, and Wang (2015) test the self-selection effect and find it to be positive and significant. But the studies by Lu, Lu, and Tao (2010) and Lu (2010) suggest that Chinese exporters are on average less productive than non-exporters. In this study, we find strong evidence that increase in exports do contribute to firm-level productivity growth.

The seminal work by Glaeser, Kallal, Scheinkman, and Shleifer (1992) studies multiple agglomerative effects on industrial growth: localization, urbanization, and competition. These market conditions are later widely tested in the productivity literature. Firm-level evidence from developed economics suggests that productivity growth benefits from specialization, competition, but not variety (Henderson, 2003; Martin, Mayer, and Mayneris, 2011). Those on Chinese data (e.g. Sheng and Song, 2013; Hu, Xu, and Yashiro, 2015) seem to agree on the role of specialization, but differ in their conclusions regarding competition. This study shows that firm-level productivity benefits strongly from market competition in the same industry, but specialization impairs productivity, which findings are more or less in line with those of Glaeser, Kallal, Scheinkman, and Shleifer (1992).

Public spending as a local market condition is usually overlooked by studies based on firm-level data. But studies on the growth or productivity effects of public capital using aggregate data are abundant. International studies (Aschauer, 1989; Holtz-Eakin, 1994; Hulten and Schwab, 1991; Fernald, 1999) generally find mixed results, but those on China (Vijverberg, Fu, and Vijverberg, 2011; Demurger, 2001) usually suggest an active role of public infrastructure at the province level. We are perhaps among the first to provide micro evidence that public spending contributes to productivity.

By controlling for the five TFP shifters – R&D, export, specialization, competition and local public spending, and constructing two spatial lags, our baseline model reveals strong technological spillovers among firms in China's electronics apparatus industry. The productivity of a firm increases by 3.5% – 4.0% if the intra-regional neighbors experience a 10% productivity growth, much larger in size and more significant than the spillovers from inter-regional neighbors. Given the small spatial scale of the jurisdictions, our findings thus suggest that the productivity transmission process attenuates rapidly in distance. These results are shown to be robust to definitions of spatial neighbors and estimation strategies. Further analysis also shows that urban districts and jurisdictions smaller in geographical area are much more susceptible to inter-regional spillovers. Finally, we find evidence that local market conditions also spillover into neighboring jurisdictions in roughly the same way as they affect local firms, with a smaller magnitude. The findings regarding public spending echo those of Yu, De Jong, Storm, and Mi (2013) and Gomez-Antonio and Fingleton (2012) at a much smaller geographic scale.[1]

The rest of the article is organized as follows: Section 2 presents our econometric methodology and discusses the analytical difficulties brought by the unbalanced panel. Section 3 describes our data source, variables, and their measurement. The empirical results are interpreted in Section 4. Finally, we conclude the study in Section 5.

# 2 Empirical methodology

We plan to gauge the impact of a few socioeconomic factors on China's industrial productivity using firm-level data. The factors considered are either firm-specific or pertinent to a geographic region where multiple firms are co-located. The statistical analysis consists of two independent components: estimating firm-level productivity and estimating the determinants of productivity. Since the panel data is unbalanced, care much be undertaken because spatial regression models based on balanced panels do not extend flawlessly to unbalanced ones. This section overviews the statistical methodologies implemented in our empirical study.

## 2.1 Estimation of total factor productivity

In this study, we adopt Levinsohn and Petrin's (2003) two-step method to estimate TFP. The Levinsohn-Petrin (hereafter LP) method enables us to obtain unbiased estimates of the production function, even if the variable inputs are endogenous to market conditions or other time-varying unobservables that affect productivity.[2]We assume the following Cobb-Douglas production function

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_k k_{it} + \omega_{it} + \eta_{it}, \tag{1}$$

where $y_{it}$, $l_{it}$, and $k_{it}$ are respectively the logarithm of value added, labor input, and capital stock of firm $i$ in year $t$. $\omega_{it}$ is the total factor productivity, and $\eta_{it}$ is the error term. The LP method starts with estimating the 1st-stage least-squares regression

$$y_{it} = \beta_l l_{it} + \phi\left(k_{it}, m_{it}\right) + \eta_{it}, \tag{2}$$

[1] Their studies employ data aggregated at the provincial level.

[2] Our choice of the LP method over that of Olley and Pakes (1996) is largely based on data concerns. The OP estimator proxies productivity by firm's investment decision and state of exit. However, investment is known to have limitations and may not be applicable in general (Levinsohn and Petrin, 2003). More seriously, our data does not provide information on investment, thus it must be derived from capital stock. Data on capital stock prepared by the NBS are infamous for being systematically biased. Although various methods have been developed to estimate the true capital stock (Brandt, Van Biesebroeck, and Zhang, 2012), they all rely on some subjective parameter, thus may introduce substantial noise to the final result. Information on firm's exit is also problematic since it does not truly reflect a change in operational state, thus not a valid proxy. With the LP method, we are able to circumvent all these obstacles.

where $m_{it}$ is the logarithm of intermediate input, and $\phi(\cdot)$ is a polynomial of order three. An estimate of $\phi\left(k_{it}, m_{it}\right)$ is then constructed as

$$\hat{\phi}_{it} = y_{it} - \hat{\beta}_l l_{it}. \tag{3}$$

The 2[nd]-stage of the LP procedure estimates $\beta_k$ through the following nonlinear regression

$$y_{it} - \hat{\beta}_l l_{it} - \beta_k k_{it} = \psi\left(\hat{\phi}_{it-1} - \beta_k k_{it-1}\right) + \eta_{it}, \tag{4}$$

where $\psi(\cdot)$ is another polynomial of order three. In practice, (4) is estimated by minimizing the sum of squared errors. Once $\hat{\beta}_l$ and $\hat{\beta}_k$ are ready, the predicted value of $\omega_{it}$ is given by

$$\hat{\omega}_{it} = y_{it} - \hat{\beta}_l l_{it} - \hat{\beta}_k k_{it}, \tag{5}$$

which is our measure of TFP.

## 2.2 The spatial fixed effects model

We assume the following SAR (spatial autoregressive) model with autoregressive disturbances:[3]

$$\begin{aligned}
\mathbf{y}_t &= \lambda \mathbf{W}_t \mathbf{y}_t + \mathbf{X}_t \boldsymbol{\beta} + \alpha \boldsymbol{\iota}_t + \boldsymbol{\mu}_t + \mathbf{u}_t, \\
\mathbf{u}_t &= \rho \mathbf{M}_t \mathbf{u}_t + \boldsymbol{\epsilon}_t,
\end{aligned} \tag{6}$$

where $t = 1, \cdots, T$ denote the time periods. Let the number of observations in year $t$ be $N_t$, the dependent variable $\mathbf{y}_t$ in (6) is an $N_t \times 1$ vector of firm-level TFP estimates in year $t$. $\mathbf{W}_t$ and $\mathbf{M}_t$ are two $N_t \times N_t$ maximum row-normalized spatial weight matrices with zeros in the main diagonal. Because the number of observations varies with time, both $\mathbf{W}_t$ and $\mathbf{M}_t$ are time-varying. $\mathbf{X}_t$ is an $N_t \times K$ matrix of $K$ exogenous regressors. $\boldsymbol{\iota}_t$ is an $N_t \times 1$ vector of all ones. $\boldsymbol{\mu}_t$ is an $N_t \times 1$ vector of individual fixed effects. If the same firm exists in period $t$ and $t'$, the column entries of $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}_{t'}$ that correspond to the same firm must be identical in value. Finally, the error term $\mathbf{u}_t$ is assumed to be generated by a spatial autoregressive process with i.i.d. disturbances $\boldsymbol{\epsilon}_t$ whose mean is zero and variance is $\sigma_\epsilon^2$.[4]

Stacking the equations over time periods, we can transform (6) into its panel representation

$$\begin{aligned}
\mathbf{y} &= \lambda \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \alpha \boldsymbol{\iota} + \boldsymbol{\mu} + \mathbf{u}, \\
\mathbf{u} &= \rho \mathbf{M} \mathbf{u} + \boldsymbol{\epsilon}.
\end{aligned} \tag{7}$$

Here $N = \sum_{t=1}^{T} N_t$ is the total number of observations. $\mathbf{y} = \left(\mathbf{y}_1', \mathbf{y}_2', \cdots, \mathbf{y}_T'\right)'$, and other vectors (including $\mathbf{X}$) are defined similarly. $\mathbf{W} = \operatorname{diag}\left(\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_T\right)$ is an $N \times N$ block-diagonal matrix with $\mathbf{W}_t$, $t = 1, \cdots, T$ on the diagonal. $\mathbf{M}$ is

[3]The model bears different names in the literature. LeSage and Pace (2009) use the name SARMA (spatial autoregressive-moving average), and Elhorst (2014) calls it the general nested spatial model.

[4]$\boldsymbol{\epsilon}_t$ from different cross sections are also assumed to be independent.

constructed in the same way.

Clearly, (7) implies

$$E(\mathbf{u}\mathbf{u}') = \sigma_\epsilon^2 (\mathbf{I}_N - \rho\mathbf{M})^{-1} (\mathbf{I}_N - \rho\mathbf{M}')^{-1} = \sigma_\epsilon^2 \left(\mathbf{I}_N - \rho(\mathbf{M} + \mathbf{M}') + \rho^2\mathbf{M}'\mathbf{M}\right)^{-1}, \quad (8)$$

and

$$E((\mathbf{W}\mathbf{y})\mathbf{u}') = E\left(\mathbf{W}(\mathbf{I}_N - \lambda\mathbf{W})^{-1}\mathbf{u}\mathbf{u}'\right) = \mathbf{W}(\mathbf{I}_N - \lambda\mathbf{W})^{-1} E(\mathbf{u}\mathbf{u}') \neq \mathbf{0}.$$

Therefore, we have both endogeneity and non-spherical disturbances. To obtain consistent estimates of the structural parameters, we can find instruments for the RHS endogenous variable $\mathbf{W}\mathbf{y}$. It remains to circumvent the incidental parameter problem by taking the within transformation of (7) , if the fixed effects themselves are not the interest of the study. For convenience, let's denote the matrices of within and between transformations by $\mathbf{Q}_0$ and $\mathbf{Q}_1$, respectively.[5] Since $\mathbf{Q}_0\boldsymbol{\mu} = \mathbf{0}$, the within transformation eliminates the fixed effects from (7), so that we have

$$\mathbf{Q}_0\mathbf{y} = \lambda\mathbf{Q}_0\mathbf{W}\mathbf{y} + \mathbf{Q}_0\mathbf{X}\boldsymbol{\beta} + \mathbf{Q}_0\mathbf{u}. \quad (10)$$

Despite the autoregressive structure in $\mathbf{u}$, the error term in (10) has zero mean, and the expectation of the RHS endogenous variable $\mathbf{Q}_0\mathbf{W}\mathbf{y}$ turns out to be

$$\begin{aligned} E(\mathbf{Q}_0\mathbf{W}\mathbf{y}) &= \mathbf{Q}_0\mathbf{W}(\mathbf{I}_N - \lambda\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu} + \alpha\boldsymbol{\iota}) \\ &= \mathbf{Q}_0 \sum_{k=0} \lambda^k \mathbf{W}^{k+1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu} + \alpha\boldsymbol{\iota}), \end{aligned} \quad (11)$$

which suggests that $\mathbf{Q}_0\mathbf{W}\mathbf{y}$ in (10) can be instrumented by[6]

$$\mathbf{G}_0 = \left(\mathbf{Q}_0\mathbf{X}, \mathbf{Q}_0\mathbf{W}\mathbf{X}, \mathbf{Q}_0\mathbf{W}^2\mathbf{X}, \cdots\right). \quad (12)$$

In this way, the structural parameters $\boldsymbol{\delta} = (\lambda, \boldsymbol{\beta}', \alpha)'$ can be consistently estimated by 2SLS, which estimator we denote by $\hat{\boldsymbol{\delta}}_W$.

The above is the procedure proposed by Kelejian and Prucha (1998). $\hat{\boldsymbol{\delta}}_W$ is consistent as far as $\mathbf{u}$ is orthogonal to the exogenous regressors $\mathbf{X}$. It remains to find a proper standard error for this estimator. Let's denote $(\mathbf{Q}_0\mathbf{W}\mathbf{y}, \mathbf{Q}_0\mathbf{X})$ by $\mathbf{Z}_0$ and

[5] $\mathbf{Q}_0$ and $\mathbf{Q}_1$ do not have neat matrix representations given the way we order observations in (7). If we permute (7) so that observations are ordered first by individual $i$ then by time $t$, then

$$\begin{aligned} \mathbf{Q}_1' &= \operatorname{diag}\left(\frac{1}{T_1}\boldsymbol{\iota}_{T_1}\boldsymbol{\iota}_{T_1}', \frac{1}{T_2}\boldsymbol{\iota}_{T_2}\boldsymbol{\iota}_{T_2}', \cdots, \frac{1}{T_n}\boldsymbol{\iota}_{T_n}\boldsymbol{\iota}_{T_n}'\right), \\ \mathbf{Q}_0' &= \mathbf{I}_N - \mathbf{Q}_1, \end{aligned} \quad (9)$$

in which $T_i$ denotes the number of time periods in which individual $i$ is observed. Therefore, the matrix representation of $\mathbf{Q}_0$ and $\mathbf{Q}_1$ can be obtained with a proper permutation of (9).

[6] Although a similar construction based on $\boldsymbol{\mu}$ can also be used as instruments, it is infeasible since $\boldsymbol{\mu}$ is unknown at this stage.

the projection matrix onto $\mathbf{G}_0$ by $\mathbf{P}_{\mathbf{G}_0}$, then

$$E\left(\hat{\boldsymbol{\delta}}_W - \boldsymbol{\delta}\middle|\mathbf{X}\right) = \left(\frac{\mathbf{Z}_0'\mathbf{P}_{\mathbf{G}_0}\mathbf{Z}_0}{N}\right)^{-1}\frac{\mathbf{Z}_0'\mathbf{P}_{\mathbf{G}_0}\mathbf{u}}{N}. \tag{13}$$

It is clear from (8) that there are heteroskedasticity and spatial autocorrelations in $\mathbf{u}$. The conventional HAC standard errors are incapable of modeling such correlations. In this study we obtain the standard errors by cluster bootstrapping, where each distinct firm is treated as a cluster. Note that $\hat{\boldsymbol{\delta}}_W$ is simply a one-step GMM estimator using equal weights for the moment conditions. The two-step or the iterated GMM estimator is theoretically more efficient. Again, bootstrapped standard errors are preferred in either case.

## 2.3 The spatial random effects model

A competing model for (7) is the random-effects specification

$$\begin{aligned} \mathbf{y} &= \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \alpha\boldsymbol{\iota} + \mathbf{u}, \\ \mathbf{u} &= \rho\mathbf{M}\mathbf{u} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \end{aligned} \tag{7$'$}$$

where the individual effects $\boldsymbol{\mu}$ is assumed to be uncorrelated with $\mathbf{X}$. The FG2SLS procedure of Kelejian and Prucha (1998) or Mutl and Pfaffermayr (2011) can be easily adapted to unbalanced panels under the random effects assumption. Despite the fact that $\mathbf{Q}_0$ does not commute with $\mathbf{M}$, and that $\boldsymbol{\mu}$ in $\mathbf{u}$ does not vanish after the within transformation, $\mathbf{Q}_0\mathbf{u}$ as a whole is uncorrelated with $\mathbf{Q}_0\mathbf{X}$. Thus, the 2SLS estimator $\hat{\boldsymbol{\delta}}_W$ obtained from (10) remains consistent. Hereby the residuals

$$\hat{\mathbf{u}} = \mathbf{y} - \hat{\lambda}\mathbf{W}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\alpha}\boldsymbol{\iota} \tag{14}$$

are consistent estimates of the error terms in (7$'$).

Although $\hat{\boldsymbol{\delta}}_W$ is consistent, they are not efficient since the error terms $\mathbf{Q}_0\mathbf{u}$ in (10) are non-spherical. A GLS transformation of (10) or (7) followed by another least-squares will restore efficiency. Thus, the next step is to estimate the parameter $\rho$ in the error term $\mathbf{u}$. Since $\boldsymbol{\epsilon}$ is assumed to be spherical and all diagonal elements of $\mathbf{M}$ are zero, we have

$$\begin{aligned} &E\left(\boldsymbol{\epsilon}'\mathbf{Q}_0\boldsymbol{\epsilon}\right) = \operatorname{tr}\left(\mathbf{Q}_0\right)\sigma_\epsilon^2 = (N-n)\sigma_\epsilon^2, \\ &E\left(\boldsymbol{\epsilon}'\mathbf{Q}_0\mathbf{M}\mathbf{Q}_0\boldsymbol{\epsilon}\right) = \operatorname{tr}\left(\mathbf{Q}_0\mathbf{M}\right)\sigma_\epsilon^2 = 0, \\ &E\left(\boldsymbol{\epsilon}'\mathbf{Q}_0\mathbf{M}'\mathbf{M}\mathbf{Q}_0\boldsymbol{\epsilon}\right) = \operatorname{tr}\left(\mathbf{Q}_0\mathbf{M}'\mathbf{M}\right)\sigma_\epsilon^2 = \operatorname{tr}\left(\operatorname{diag}\left(\mathbf{Q}_0\right)\operatorname{diag}\left(\mathbf{M}'\mathbf{M}\right)\right)\sigma_\epsilon^2, \\ &E\left(\mathbf{u}'\mathbf{Q}_1\mathbf{u}\right) = N\sigma_\mu^2 + n\sigma_\epsilon^2. \end{aligned} \tag{15}$$

Here $n$ denotes the number of distinct individuals (firms) in the unbalanced panel, $\operatorname{diag}(\mathbf{Q}_0)$ denotes the diagonal matrix constructed from the main diagonal of $\mathbf{Q}_0$, and $\operatorname{diag}(\mathbf{M}'\mathbf{M})$ is defined similarly.

Since $\mathbf{Q}_0\boldsymbol{\epsilon} = \mathbf{Q}_0 (\mathbf{I}_N - \rho\mathbf{M})\,\mathbf{u}$, (15) lead to the following moment conditions:

$$\begin{aligned}
&\hat{\mathbf{u}}' (\mathbf{I}_N - \rho\mathbf{M}') \mathbf{Q}_0 (\mathbf{I}_N - \rho\mathbf{M}) \hat{\mathbf{u}} = (N-n)\sigma_\epsilon^2, \\
&\hat{\mathbf{u}}' (\mathbf{I}_N - \rho\mathbf{M}') \mathbf{Q}_0\mathbf{M}\mathbf{Q}_0 (\mathbf{I}_N - \rho\mathbf{M}) \hat{\mathbf{u}} = 0, \\
&\hat{\mathbf{u}}' (\mathbf{I}_N - \rho\mathbf{M}') \mathbf{Q}_0\mathbf{M}'\mathbf{M}\mathbf{Q}_0 (\mathbf{I}_N - \rho\mathbf{M}) \hat{\mathbf{u}} = \operatorname{tr}(\operatorname{diag}(\mathbf{Q}_0) \operatorname{diag}(\mathbf{M}'\mathbf{M}))\,\sigma_\epsilon^2, \\
&\hat{\mathbf{u}}' (\mathbf{I}_N - \rho\mathbf{M}') \mathbf{Q}_1 (\mathbf{I}_N - \rho\mathbf{M}) \hat{\mathbf{u}} = N\sigma_\mu^2 + n\sigma_\epsilon^2.
\end{aligned} \tag{16}$$

The three unknowns $\rho$, $\sigma_\epsilon^2$, and $\sigma_\mu^2$ can be estimated from (16) by GMM. Let's denote the estimates by $\left(\hat{\rho}, \hat{\sigma}_\epsilon^2, \hat{\sigma}_\mu^2\right)$, with which we can perform the FGLS transformation on (7):

$$\begin{aligned}
\widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} (\mathbf{I}_N - \hat{\rho}\mathbf{M})\,\mathbf{y} = \lambda \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} (\mathbf{I}_N - \hat{\rho}\mathbf{M})\,\mathbf{W}\mathbf{y} + \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} (\mathbf{I}_N - \hat{\rho}\mathbf{M})\,\mathbf{X}\boldsymbol{\beta} \\
+ \alpha \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} (\mathbf{I}_N - \hat{\rho}\mathbf{M})\,\boldsymbol{\iota} + \boldsymbol{\nu}.
\end{aligned} \tag{17}$$

$\boldsymbol{\Omega}^{-\frac{1}{2}}$ in (17) is the conventional Cochrane-Orcutt transformation for unbalanced panels. For any variable $\xi$,

$$\widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} \xi_{it} = \xi_{it} - \frac{\hat{\sigma}_\epsilon}{\left(T_i \hat{\sigma}_\mu^2 + \hat{\sigma}_\epsilon^2\right)^{\frac{1}{2}}} \bar{\xi}_{i.}, \tag{18}$$

where $T_i$ is the number of observations pertinent to individual $i$ (Baltagi, Egger, and Kesina, 2015).

The last step of the procedure is a 2SLS regression on (17) with the RHS endogenous regressor $\widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} (\mathbf{I}_N - \hat{\rho}\mathbf{M})\,\mathbf{W}\mathbf{y}$ instrumented by

$$\begin{aligned}
\mathbf{G}_1 = \big(&\mathbf{Q}_0\mathbf{X}, \mathbf{Q}_0\mathbf{W}\mathbf{X}, \mathbf{Q}_0\mathbf{W}^2\mathbf{X} \cdots, \mathbf{Q}_0\mathbf{M}\mathbf{X}, \mathbf{Q}_0\mathbf{M}\mathbf{W}\mathbf{X}, \mathbf{Q}_0\mathbf{M}\mathbf{W}^2\mathbf{X} \cdots, \\
&\mathbf{Q}_1\mathbf{X}, \mathbf{Q}_1\mathbf{W}\mathbf{X}, \mathbf{Q}_1\mathbf{W}^2\mathbf{X} \cdots, \mathbf{Q}_1\mathbf{M}\mathbf{X}, \mathbf{Q}_1\mathbf{M}\mathbf{W}\mathbf{X}, \mathbf{Q}_1\mathbf{M}\mathbf{W}^2\mathbf{X} \cdots, \\
&\mathbf{Q}_0\mathbf{W}\boldsymbol{\iota}, \mathbf{Q}_0\mathbf{W}^2\boldsymbol{\iota} \cdots, \mathbf{Q}_0\mathbf{M}\boldsymbol{\iota}, \mathbf{Q}_0\mathbf{M}\mathbf{W}\boldsymbol{\iota}, \mathbf{Q}_0\mathbf{M}\mathbf{W}^2\boldsymbol{\iota} \cdots, \\
&\boldsymbol{\iota}, \mathbf{Q}_1\mathbf{W}\boldsymbol{\iota}, \mathbf{Q}_1\mathbf{W}^2\boldsymbol{\iota} \cdots, \mathbf{Q}_1\mathbf{M}\boldsymbol{\iota}, \mathbf{Q}_1\mathbf{M}\mathbf{W}\boldsymbol{\iota}, \mathbf{Q}_1\mathbf{M}\mathbf{W}^2\boldsymbol{\iota} \cdots, \big),
\end{aligned} \tag{19}$$

which is the optimal set of instruments in the random effects setup.

Under certain assumptions, Mutl and Pfaffermayr (2011) show that the FG2SLS estimator of the random effects model has the expected asymptotic distribution. If we denote $\left(\boldsymbol{\Omega}^{-\frac{1}{2}} (\mathbf{I} - \rho\mathbf{M})\,\mathbf{W}\mathbf{y}, \boldsymbol{\Omega}^{-\frac{1}{2}} (\mathbf{I} - \rho\mathbf{M})\,\mathbf{X}\right)$ by $\widetilde{\mathbf{Z}}$, and the projection matrix onto $\mathbf{G}_1$ by $\mathbf{P}_{\mathbf{G}_1}$, then the FG2SLS random effects estimator $\hat{\boldsymbol{\delta}}_R$ has the asymptotic distribution

$$\hat{\boldsymbol{\delta}}_R \xrightarrow{a} N\left(\boldsymbol{\delta}, \frac{\sigma_\epsilon^2}{N} \left(\frac{\widetilde{\mathbf{Z}}'\mathbf{P}_{\mathbf{G}_1}\widetilde{\mathbf{Z}}}{N}\right)^{-1} \left(\frac{\widetilde{\mathbf{Z}}'\mathbf{P}_{\mathbf{G}_1}\widetilde{\mathbf{Z}}}{N}\right) \left(\frac{\widetilde{\mathbf{Z}}'\mathbf{P}_{\mathbf{G}_1}\widetilde{\mathbf{Z}}}{N}\right)^{-1}\right). \tag{20}$$

This suggests that the variance-covariance matrix of $\hat{\boldsymbol{\delta}}_R$ can be estimated by

$$\hat{\sigma}_\epsilon^2 \left( \widehat{\widehat{\mathbf{Z}}}' \mathbf{P}_{\mathbf{G}_1} \widehat{\widehat{\mathbf{Z}}} \right)^{-1} \widehat{\widehat{\mathbf{Z}}}' \mathbf{P}_{\mathbf{G}_1} \widehat{\widehat{\mathbf{Z}}} \left( \widehat{\widehat{\mathbf{Z}}}' \mathbf{P}_{\mathbf{G}_1} \widehat{\widehat{\mathbf{Z}}} \right)^{-1}, \tag{21}$$

where $\widehat{\widehat{\mathbf{Z}}}$ is given by $\left( \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} (\mathbf{I} - \hat{\rho}\mathbf{M}) \mathbf{W}\mathbf{y}, \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} (\mathbf{I} - \hat{\rho}\mathbf{M}) \mathbf{X} \right)$, while $\hat{\sigma}_\epsilon^2$ is obtained from the second stage of the procedure.

## 2.4 Issues with unbalanced panels

Our specification of the fixed effects model (7) is inconsistent with that of the random effects model (7′). In the FG2SLS literature (e.g. Kelejian, Prucha, and Yuzefovich, 2004; Mutl and Pfaffermayr, 2011), both models are given by (7′), i.e., the individual effects are assumed to be a component of the disturbances. This specification allows a similar feasible GLS transformation on the fixed effects model. The 2SLS estimator obtained from the transformed equation is more efficient than $\hat{\boldsymbol{\delta}}_W$. More importantly, the unified treatment facilitates a subsequent Hausman test. Such advantages, however, rely heavily on the commutativity of the within and the spatial lag transformations, which is automatically satisfied if the panel is balanced and if the spatial weight matrices are time-invariant. Without commutativity, this advantage becomes an analytical burden. To see this, let's note that the within-transformation of (7′) gives

$$\mathbf{Q}_0\mathbf{y} = \lambda \mathbf{Q}_0 \mathbf{W}\mathbf{y} + \mathbf{Q}_0 \mathbf{X}\boldsymbol{\beta} + \mathbf{Q}_0 (\mathbf{I} - \rho\mathbf{M})^{-1} (\boldsymbol{\mu} + \boldsymbol{\epsilon}). \tag{10′}$$

Since $\mathbf{Q}_0 (\mathbf{I}_N - \rho\mathbf{M})^{-1} \neq (\mathbf{I}_N - \rho\mathbf{M})^{-1} \mathbf{Q}_0$, the individual effects $\boldsymbol{\mu}$ do not vanish. Under the fixed effects assumption, the error component $\mathbf{Q}_0 (\mathbf{I} - \rho\mathbf{M})^{-1} (\boldsymbol{\mu} + \boldsymbol{\epsilon})$ is again correlated with $\mathbf{X}$, so the within estimator becomes inconsistent.

By moving the individual effects from the error component to the structural model, (7) ensures the consistency of $\hat{\boldsymbol{\delta}}_W$. With (7), however, the three-stage procedure is no longer feasible. This is because the estimation of $\rho$ is based on the residuals from the first-stage 2SLS, namely

$$\hat{\mathbf{u}}_0 = \widehat{\mathbf{Q}_0\mathbf{u}} = \mathbf{Q}_0 \left( \mathbf{y} - \hat{\lambda}\mathbf{W}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\alpha}\boldsymbol{\iota} \right). \tag{22}$$

If the spatial weights $\mathbf{M}_t$ are time-invariant, then the within transformation $\mathbf{Q}_0$ and the spatial lag operation $\mathbf{M}$ are commutative, so that

$$(\mathbf{I} - \rho\mathbf{M}) \mathbf{Q}_0 \mathbf{u} = \mathbf{Q}_0 (\mathbf{I} - \rho\mathbf{M}) \mathbf{u} = \mathbf{Q}_0 \boldsymbol{\epsilon}. \tag{23}$$

Therefore, $(\mathbf{I} - \rho\mathbf{M}) \widehat{\mathbf{Q}_0\mathbf{u}}$ can be used to construct the moment conditions regarding $\mathbf{Q}_0\boldsymbol{\epsilon}$. Without commutativity, however, (23) is invalid and $(\mathbf{I} - \rho\mathbf{M}) \widehat{\mathbf{Q}_0\mathbf{u}}$ becomes an estimate of $(\mathbf{I} - \rho\mathbf{M}) \mathbf{Q}_0 (\mathbf{I} - \rho\mathbf{M})^{-1} \boldsymbol{\epsilon}$, which contains the unknown parameter $\rho$.

Even if we are given a consistent estimator of $\rho$, the time-varying spatial weights remain an obstacle to GLS estimation. To see this point, let's note that in the transformed equation

$$(I - \rho \mathbf{M})\, \mathbf{y} = \lambda\, (I - \rho \mathbf{M})\, \mathbf{W}\mathbf{y} + (I - \rho \mathbf{M})\, \mathbf{X}\beta + (I - \rho \mathbf{M})\, (\alpha \boldsymbol{\iota} + \boldsymbol{\mu}) + \boldsymbol{\epsilon},$$

$(I - \rho \mathbf{M})\, \boldsymbol{\mu}$ must be eliminated before estimating the structural parameters. In balanced panel models, this is done by the within transformation because $\mathbf{M}$ commutes with $\mathbf{Q}_0$. Without commutativity, the within transformation is bound to fail.

The above discussion reveals the critical constraint imposed by unbalanced panels on the fixed effects model. In order to obtain consistent estimates of the structural parameters from least squares, one must choose (7) over (7′). By doing so, one has to forfeit the efficiency gains from the FG2SLS procedure. Nevertheless, the work by Kelejian, Prucha, and Yuzefovich (2004) shows that such efficiency gains, if any, could be small in magnitude with even a moderate sample size.

Unbalanced panels also introduce minor changes to the FG2SLS procedure for the random effects model. Because $\mathbf{Q}_0$ does not commute with $\mathbf{M}$, the moment conditions in (15) differ from their balanced-panel counterparts. For a similar reason, (19) now consists of more instruments.

Although we use different specifications for the fixed and random effects models, the fixed effects estimator $\hat{\boldsymbol{\delta}}_W$ coincides with the first-stage within estimator of the random effects model. Given specification (7′), both $\hat{\boldsymbol{\delta}}_W$ and $\hat{\boldsymbol{\delta}}_R$ are consistent under the random effects assumption that $E\,(\boldsymbol{\mu}|\mathbf{X}) = \mathbf{0}$, but the latter is more efficient. Therefore, we can design a Hausman test by comparing $\hat{\boldsymbol{\delta}}_W$ and $\hat{\boldsymbol{\delta}}_R$. If the random effects assumption is rejected, we shall estimate the fixed effects model (7) and base our inference on this alternative specification.

# 3 Data description and measurement

We obtain firm-level data on accounting and financial variables from the *China Industry Survey* conducted by the National Bureau of Statistics. The data set was issued on an annual basis from 1996 through 2010, covering over 160,000 industrial establishments annually.[7] Data on accounting and financial variables, including total value added, net value of fixed assets, total employment, total intermediate inputs, etc., enable us to estimate firm-level productivity and construct proxies for firm characteristics. The data set also provides locational information of each firm, with which a serious spatial analysis is possible. Because of its wide coverage and comprehensive information on Chinese firms, this unique data set has been widely used in empirical studies on China's manufacturing sector, especially those on productivity analysis (Brandt, Van Biesebroeck, and Zhang, 2012; Hu, Xu, and Yashiro, 2015; Baltagi, Egger, and Kesina, 2015).

The geo-data are compiled from multiple sources. Our major reference is the official *Codebook of Administrative Divisions* prepared by the National Bureau of

[7]The establishments in the survey are either state-owned enterprises or above-scale (annual revenue from principal business over 5 million CNY) private (potentially with foreign ownership) enterprises.

Statistics. The codebook, published as the National Standard GB/T 2260, assigns six-digit administrative codes to over 3,000 administrative divisions in four levels of the hierarchy: province, prefecture, sub-prefecture, district/county. As of 2007, there were 2866 administrative units in the district/county level. With the codebook, we are able to identify the location (district/county) of each establishment. We acquire the geographic data of the administrative units from a commercial source. These include the coordinates of the administrative centers and a shape file of administrative boundaries. The geo-data enables us to construct spatial neighborhood relations among administrative units either by spatial distance or by contiguity.

Our study also employs data on the budgetary expenditure of local governments in the district/county level. The data is extracted from the *Fiscal Statistics of Cities and Counties* (various issues) compiled by the Ministry of Finance. It is then merged into the main data set by matching administrative codes. Details of our data preparation process can be found in appendix A.

After these steps, we are able to assemble a longitudinal data set of 615,624 industrial firms from 470 four-digit industrial sectors, located in 2,862 districts/counties. Table 1 gives an overview of the data. Judged by the number of firms, total output, and total employment, the data set provides a near-complete coverage of China's above-scale industrial firms over 1999-2007. Remarkably, a large fraction of the sample is replaced by new entrants each year. Sample rotation to such an extent may not truly reflect firms' entry and exit behavior, but is likely to be unintended error in the sampling process.[8] Therefore, exit is no longer a valid proxy for certain unobservable characteristics of a firm, as suggested by (Olley and Pakes, 1996). The sparse nature of the panel data set simply rules out the method of multiple imputation.[9]

In this study, we focus on the sector of electric equipment (industry code 3900) over the period of 1999-2007. The choice of this sector is based on the following concerns: First, the sector is technologically intensive. Influence of socioeconomic factors on productivity may be more pronounced thus easier to detect for this sector. Second, this sector provides a large sample size (over 99,000), with over 27,596 firms located in 1629 districts/counties. The wide spatial distribution of the sector allows us to conduct spatial analysis without a heavy penalty of data loss. The spatial distribution of firms and employment in the sector are shown by figures (1-2).[10] A pattern of agglomeration is evident in these graphs. Firms and employment cluster on the eastern coast and inland industrial centers, while the vast areas in the west are unoccupied. This observation suggests a strong spatial linkage in the locational choice of firms, and likely spatial interactions between firms when they are close.

The estimation of (1) requires firm-level data on value added, inputs in capital, labor, and intermediate goods. In this study, $y_{it}$ is measured by total value added of the firm, $l_{it}$ by annual average number of employed personnels, $k_{it}$ by annual average value of net fixed assets, and $m_{it}$ by value of intermediate inputs. Data on these

[8]There are many gaps in the panel, indicating "exits" followed by "entries."

[9]According to Hughes (2013), multiple imputation may introduce huge bias even if a moderate amount of data is missing. Besides, the unbalancedness may reflect entry/exist behavior of firms. Replacing firms that no longer exist with imputed ones changes the market structure.

[10]The figures use three year averages (log-transformed values) over 2005-2007.

variables are available from the *China Industry Survey* with a few exceptions. Total value added is missing in 2001, 2002, and 2004, thus must be derived from other variables by accounting identities. The 2001 and 2002 values are computed as

$$\begin{aligned}\text{total value added} = {} & \text{gross industrial output - value of intermediate inputs} \\ & + \max\{0, \text{value added tax payable}\},\end{aligned}$$

while those of 2004 are recovered from

$$\begin{aligned}\text{total value added} = {} & \text{revenue from principal business + increase in inventory} \\ & \text{- value of intermediate inputs} + \max\{0, \text{value added tax payable}\}.\end{aligned}$$

We plot the weighted (by employment) average TFP of each district/county in figure (3). Compared to figures (1-2), the spatial pattern of TFP is less clear, partly because individual heterogeneity is smoothed out by taking the average.[11] Nevertheless, in small clusters, such as the metro areas of Chengdu, Guangzhou, and Wuhan, we do observe the spatial gradient of TFP declining from the center to the periphery.

The basic geographic unit in our data is an (urban) district or a county. The location of a firm is identified by the district/county in which it operates, but we have no further locational information within the district/county. We assume that firms in the same district/county are all located at the administrative center. In this regard, a firm has two types of neighbors: those in the same district/county and those in neighboring district/counties. Thus, we introduce two spatial weight matrices to the SAR model (6):

$$\mathbf{W}_{1t} y_{ikt} = \sum_{\substack{j \in I_t(k) \\ j \neq i}} l_{jkt} y_{jkt} \Big/ \sum_{\substack{j \in I_t(k) \\ j \neq i}} l_{jkt}, \tag{24}$$

$$\mathbf{W}_{2t} y_{ikt} = \sum_{\substack{j \in I_t(k') \\ k' \in N(k)}} l_{jk't} y_{jk't} \Big/ \sum_{\substack{j \in I_t(k') \\ k' \in N(k)}} l_{jk't}. \tag{25}$$

Here the subscripts $i$ and $j$ denote firms, $k$ and $k'$ denote districts/counties, and $t$ denotes time. $I_t(k)$ is the set of all firms located in district/county $k$ in year $t$, and $N(k)$ is the set of all neighboring districts/counties of district/county $k$. $y$ is any variable to be weighted, and $l_{ikt}$ is the employment of firm $i$ in district/county $k$ and year $t$. The construction is based on the premise that larger firms (measured by employment) exert stronger influence on their neighbors than smaller ones. Contiguous (Rook style) districts/counties are treated as neighbors. We also use an alternative definition based on distance. Two districts/counties are regarded as neighbors if their administrative centers are within 50 kilometers in great circle distance.[12]

[11] A few studies (e.g. Hu, Xu, and Yashiro, 2015) show that firm type, including ownership structure and size, is correlated with productivity. If a firm type is inproportionally high or low in a district/county, the average TFP will be biased compared to those of neighbors.

[12] The choice of the 50-kilometer cutoff value is based on the observation that the mean distance

Both $\mathbf{W}_{1t}$ and $\mathbf{W}_{2t}$ are maximum row normalized, and they have zeros in the main diagonal. It is easy to see $\mathbf{W}_{1t}\mathbf{W}_{2t} = \mathbf{W}_{2t}$. This property helps to alleviate the computation burden in the regression stage. For simplicity, we use a single AR term in the error component, i.e., we assume $\mathbf{u} = \rho \mathbf{W}_1 \mathbf{u} + \boldsymbol{\epsilon}$ in (7) and $\mathbf{u} = \rho \mathbf{W}_1 \mathbf{u} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$ in (7′).

The SAR model (6) considers two types of exogenous variables: firm idiosyncrasies that have no effect on other firms, and market conditions that impact not only local firms, but also likely firms in neighboring districts/counties. The literature (Sheng and Song, 2013; Hu, Xu, and Yashiro, 2015; Baltagi, Egger, and Kesina, 2015) has identified multiple firm-level characteristics that are correlated with productivity, including ownership structure, size of the firm, years of operation, R&D activity, and participation in the international market. Our current study specifies individual effects. Thus, only time-varying factors can be properly estimated by the model. We use two variables to proxy R&D ($rd$) and export ($ex$) activities. They are respectively measured by the share of new merchandise in gross output and the fraction of gross sales that are exported. We consider three aspects of the local market environment: specialization ($spec$), competition ($comp$), and public spending ($pub$). According to Marshall's (1890) hypothesis, a city benefits from specialization because of spillovers between firms in the same industry. Our measure follows that of Glaeser, Kallal, Scheinkman, and Shleifer (1992), i.e.,

$$spec_{kt} = \frac{\frac{\text{sectoral employment in area } k \text{ and year } t}{\text{total industrial employment in area } k \text{ and year } t}}{\frac{\text{sectoral employment in China and year } t}{\text{total industrial employment in China and year } t}}$$

Porter (1990) argues that competition among local firms boosts productivity. Instead of measuring the average firm size (Glaeser, Kallal, Scheinkman, and Shleifer, 1992; Rosenthal and Strange, 2003), we use the Herfindahl-Hirschman index of sectoral employment in the district/county. Using the notations from (24), we define

$$comp_{kt} = \sum_{i \in I_t(k)} \left( \frac{l_{ikt}}{\sum_{i \in I_t(k)} l_{ikt}} \right)^2 .$$

The HHI take value between zero and unity. A smaller value indicates stronger competition. Because the HHI decreases in the number of firms even if the distribution of employment among firms remains unchanged, it also proxies the number of firms, which has also been proposed as a measure of agglomeration (Hu, Xu, and Yashiro, 2015). Although the relationship between public investment and productivity has been studied for long (Aschauer, 1989; Fernald, 1999; Vijverberg, Fu, and Vijverberg, 2011), it remains a missing link in the empirical studies using micro data. In this study, we use total budgetary expenditure by the local government ($pub$).[13] The

between contiguous neighbors is 64 kilometers, while the median is 47 kilometers. In this way, the estimate of $\lambda_2$ will have a similar interpretation as in the contiguity case.

[13]The expenditure categories in different issues of the *Fiscal Statistics of Cities and Counties* are

variable is log-transformed so that the coefficient measures the effect of a percentage change in public spending.

In figures (4-8), these exogenous variables are plotted on the map of administrative divisions. There is a clear spatial pattern in *ex*, *comp*, and *pub*. Evidently firms in the metro centers along the eastern coastal are export-oriented. Competition, measured by HHI, is more fierce in regional centers, including those located in central and western China. The spatial variation in public spending is less pronounced, but the urban cores in the Yangtze Delta and the Pearl River Delta receive far more public spending than the rest of the nation.

Our model allows local market conditions (*spec*, *comp*, and *pub*) to influence firms in neighboring districts/counties. Therefore, their spatial lags $\mathbf{W}_{2t}spec_{kt}$, $\mathbf{W}_{2t}comp_{kt}$, and $\mathbf{W}_{2t}pub_{kt}$ are also included as regressors. Finally, we end up with the following empirical model

$$\begin{aligned} tfp_{ikt} = \lambda_1 \mathbf{W}_{1t} tfp_{ikt} + \lambda_2 \mathbf{W}_{2t} tfp_{ikt} + \beta_1 rd_{kt} + \beta_2 ex_{kt} \\ + \beta_3 spec_{kt} + \beta_4 comp_{kt} + \beta_5 pub_{kt} \\ + \beta_6 \mathbf{W}_{2t} spec_{kt} + \beta_7 \mathbf{W}_{2t} comp_{kt} + \beta_8 \mathbf{W}_{2t} pub_{kt} \\ + \alpha + \text{error term}, \end{aligned} \tag{26}$$

where the error term is either $\boldsymbol{\mu} + (I - \rho \mathbf{W}_1)^{-1} \boldsymbol{\epsilon}$ in the fixed effects model or similarly $(I - \rho \mathbf{W}_1)^{-1} (\boldsymbol{\mu} + \boldsymbol{\epsilon})$ in the random effects model.

In practice, we retain the observations that have both types of neighbors, then those with complete observations. This results in an effective sample size of 84727 if the neighbor relationship among district/counties is defined by contiguity and 81331 if the neighbor relationship is defined by the 50 km criterion.

# 4 Empirical results

## 4.1 The baseline model

Table 4 summarizes the estimates of (26) by different model specifications. The conventional fixed effects estimates are reported in column (FE) as benchmark. Column (FE-IV) reports the the instrument variable 2SLS estimator discussed in section (2.2). Since we have included two spatial lags in (26), the list of instruments (12) has to be expanded. In practice, the instruments considered are *rd*, *ex*, *spec*, *comp*, *pub*, and their spatial lags by $\mathbf{W}_1$, $\mathbf{W}_2$, or their interactions up to the second power. Since the error terms in the within transformed model are potentially autocorrelated in space, the conventional HA and HAC type standard errors are questionable. The standard errors reported here are obtained through 50 bootstrap sample of firms. In column (FE-GMM), we implement the conventional two-step GMM using the same set of instruments. The same bootstrapping procedure is used to obtain the standard errors. Finally, we report the FG2SLS estimates of the random effects model

not compatible.

in column (RE-FG2SLS). Note that we use $\mathbf{M} = \mathbf{W}_1$ in the error component, so the instruments suggested by (19) are built on (1) *rd*, *ex*, *spec*, *comp*, *pub*, the vector of ones, and (2) their spatial transformations by $\mathbf{W}_1$, $\mathbf{W}_2$, or their interactions up to the second power. These variables are within-transformed (except for the vector of ones) and between-transformed into the instruments. Since the error terms in the GLS transformed structural equation (17) are spherical, we report the conventional standard errors in column (RE-FG2SLS).

The estimates in column (FE-IV) are notably different from those in column (FE), indicating substantial endogeneity bias in the latter. We note that the clustered standard errors for the 2SLS within estimator are sizably smaller (not reported) than the bootstrapped values. This observation justifies our earlier concerns. According to these estimates, the productivity of a firm increases by four percent if the productivity of neighboring firms in the same district/county increases uniformly by ten percent.[14] Judged by the magnitude and significance, the spatial spillovers within the same district/county are strong. The coefficient on $\mathbf{W}2\_tfp$ is much smaller in size and insignificant. The result echoes the findings made by other researchers that spatial interactions attenuate rapidly in distance (Rosenthal and Strange, 2003, among others). Recently Baltagi, Egger, and Kesina (2015) analyze the spillover effects among Chinese firms using a statistical framework similar to ours. Interestingly, they uses the same data source as ours. There, they made an unusual observation that the strength of spillover effects, measured by the size of the spatial AR coefficient, does not change much as they extend the geographic scope from districts/counties to prefecture units, then to provinces. By incorporating two different spatial lags, our model is able to address this issue in an explicit way.[15] It is worthwhile to note that the within model sweeps off idiosyncratic effects, thus it estimates how fluctuations in productivity propagate over space. It does not reflect the selection and sorting effects suggested by the recent literature (e.g. Behrens, Duranton, and Robert-Nicoud, 2014).

The coefficients on *rd* and *ex* are both highly significant. According to these estimates, firms are more productive as they increase their development of new products or export more. These results are in line with the empirical evidence in the literature, especially those on China. The result suggests that the level of specialization in the local area or neighboring districts/counties has little effect on firm productivity. In contrast, productivity benefits significantly from competition.[16] Both results are similar to those of Glaeser, Kallal, Scheinkman, and Shleifer (1992). Finally, we find strong evidence that firms benefit from public expenditures. A ten percent increase in local public expenditure increases productivity by 1.4 percent. The marginal effects of *comp* and *pub* in neighboring districts/counties are much smaller in size, but still highly significant. These show firms also benefit from favorable market conditions in neighboring areas.

[14] Our spatial weight matrices are row-normalized unless the firm is an island without neighbors.

[15] Their spatial model uses a single spatial AR term which corresponds to $\mathbf{W}_1$ in ours, thus is less flexible than ours.

[16] The HHI is inversely related to the number of firms, which is a commonly used measure for the agglomeration. Thus, we also find evidence that agglomeration boosts productivity.

The two-stage GMM estimates reported in column (FE-GMM) are very close to the 2SLS estimates. Despite its theoretical advantage in efficiency, the GMM procedure yields virtually identical standard errors. Therefore, we prefer the 2SLS estimator because it is much easier to implement.[17] It is noteworthy that the estimated idiosyncrasies $\hat{\boldsymbol{\mu}}$ in (7) is strongly correlated with the regressors $\mathbf{Z} = (\mathbf{Wy}, \mathbf{X})$, namely, $\text{cor}\left(\hat{\boldsymbol{\mu}}, \mathbf{Z}\hat{\boldsymbol{\delta}}_W\right) = -0.33$. Furthermore, the standard error of $\boldsymbol{\mu}$ (1.137) is large relative to that of $\mathbf{Q}_0\mathbf{u}$ (0.671). These observations invalidates the random effects assumption. The random effect FG2SLS estimates are reported in the last column. They are in sharp contrast to the 2SLS estimates, while the standard errors are notably smaller in size. The evidence thus rejects the random effects assumption.[18]

We rerun the regressions using the distance based definition of neighborhood. The estimates are summarized in table 5. We observe a similar patter as in table 4: the 2SLS and GMM routines produce similar estimates for the within model, which differ from the conventional within estimates or the random effects FG2SLS estimates. The difference is more pronounced in the spatial AR coefficients. Using the 2SLS estimates, we find $\text{cor}\left(\hat{\boldsymbol{\mu}}, \mathbf{Z}\hat{\boldsymbol{\delta}}_W\right) = -0.32$, and a large standard error for $\boldsymbol{\mu}$ (1.133) compared to that of $\mathbf{Q}_0\mathbf{u}$ (0.667). Thus we favor the within model over the random effects model. Again, the GMM estimator does not show a clear advantage in terms of efficiency, so we base our inference on the 2SLS estimates. The estimates based on the alternative neighborhood concept are comparable to the previous ones. the spillovers from firms in the same district/county remain significant, but slightly weaker. The spillovers from neighboring areas remain insignificant. Among the exogenous regressors, the specialization index is again insignificant, while all other regressors are highly significant with expected signs.

## 4.2 Urban districts and distance

The baseline model shows strong and significant technological spillovers among firms in the same district/county. It also shows that the spillovers become much weaker and insignificant when the spatial linkage is extended to include firms in neighboring areas. This section further investigates the key socioeconomic or geographic factors behind these spillover effects. The first factor that comes to our mind is China's administrative division. For historical reasons, urban districts and counties are very different in socioeconomic characteristics. The traditional urban districts serve as the administrative and economic centers of prefecture-level cities. They are small in area, but equipped with high quality public infrastructure. Starting from the early '90s, a new type of urban districts emerged. They are upgraded from counties to host the

[17] The two-stage GMM procedure is computationally burdensome. For every bootstrap sample, the GMM criterion function has to be minimized twice. It took roughly 100 minutes to finish 50 repetitions on modern hardware.

[18] A formal Hausman test on the random effects specification demands theoretical development on the joint distribution of $\hat{\boldsymbol{\delta}}_W$ and $\hat{\boldsymbol{\delta}}_R$, which hasn't been accomplished at this stage. Given the huge difference between the two sets of estimates, the random effects specification is not likely to survive such a test.

growing body of manufacturing firms. Since their birth, these urban districts have experienced rapid growth in industrial output, employment, and infrastructure. Some of them have grown into new urban centers. Compared to counties or county-level cities, both types of urban districts have a much higher concentration of firms and employment, but smaller geographic areas (table 2).[19] The distinction suggests that the spillover effects studied in the baseline model may behave differently in urban districts and counties.

We divide the sample into two sub-samples by administrative type, and run the baseline regression on them. The estimates are reported in table 6. The numbers in the first column are taken from table 5 (column FE-IV). The second and third columns report estimates from the sub-samples. There is a sharp contrast in the estimated AR coefficients. $\lambda_1$ is highly significant in both samples, but the estimate is smaller in the urban sample. $\lambda_2$ estimated from the urban sample is much more significant and larger in size than that of the county sample. To formally test whether the AR coefficients are different between the sub-samples, we introduce a dummy variable (*county*) for counties and make it interact with $\mathbf{W}_1 tfp$ and $\mathbf{W}_2 tfp$.[20] The estimates are reported in column 4. Clearly, the difference in $\lambda_1$ is highly significant. We also perform the analysis using the contiguity based neighbor relationship (table 7). There we observe the same pattern.

Judged by these estimates, the spatial autoregressive structure is very different in urban districts and counties. Firms in urban districts are subject to both types of spillover effects (intra-regional and inter-regional). The estimates of $\lambda_1$ and $\lambda_2$ are comparable in size because the two effects are equally potent. For firms in counties, the inter-regional spillover effect is very weak in size and significance. Therefore, the intra-regional effect plays the dominant role, and $\lambda_1$ is large in size. An easy explanation to this observation can be based on the special role of urban districts in China's administrative hierarchy. They are designated regional hubs, and they have tight economic linkages with the rest of the prefecture, including their neighbors. Counties are stand-alone administrative units under the prefecture. Consequently such inter-regional linkages are far weaker for counties.

The NEG theory suggests another factor that may also explain the urban-county difference: distance. Urban districts are on average much smaller in area than counties. Consequently, they are closer to their neighbors in space. Firms located in urban districts are subject to both types of spillovers, but the those located in counties are hardly affected by inter-regional spillovers because of greater distance. This argument also suggests significant $\lambda_1$ and $\lambda_2$ for urban districts (smaller area); a large and significant $\lambda_1$ for counties (larger area). In order to ascertain the true mechanism behind the urban-county difference, we need to conduct a similar investigation into the second argument.

We sort all 2866 administrative units by area. The lower 50% are marked as

[19]The 2005-2007 sub-sample consists of 538 urban districts and 483 counties or county-level cities. They host 9539 and 5815 firms (annual average) in sector 3900, respectively.

[20]At the same time, the set of instruments are expanded to include their interactions with the dummy variable.

small, and the rest are marked as large.[21] We then construct a dummy variable *large* to identify the large administrative units. It can be seen from table 3 that 75% of the administrative units in the sample are small ones, hosting roughly 73% of the firms. Evidently the majority of urban districts are small, but over 50% of the counties are also small. The sample correlation between *county* and *large* is 0.269. Despite the overlap, they actually measure different concepts.

The following analysis is similar to what we have done previously. We run the baseline model on the sample of small administrative units, then on the sample of large ones. Then we let *large* interact with $\mathbf{W}_1 tfp$ and $\mathbf{W}_2 tfp$ and run the regression on the full sample. The results are reported in tables 8 and 9. The pattern is strikingly similar to that of tables 6 and 7. $\lambda_1$ is highly significant in all specifications, and significantly smaller in size for small administrative units. $\lambda_2$ is significant in the sample of small administrative units but insignificant in the other sub-sample. The difference in $\lambda_2$ is again significant. The evidence thus strongly supports the distance-based argument.

We thus find independent evidence that administration type and spatial area are factors that influence the spillover effects on districts/counties. They are related, though not identical concepts. A regression model that accounts for both factors helps to reveal the true causal effect, or at least, which factor is relatively more important. We thus include the interactions of $\mathbf{W}_1 tfp$ and $\mathbf{W}_2 tfp$ with both *county* and *large*. The estimates are reported in the last column of tables 8 and 9. The results are mixed. In table 8, where neighborhood is defined by the 50 km criterion, the interactions with $\mathbf{W}_1 tfp$ are significant but those interacting with $\mathbf{W}_2 tfp$ are not. It indicates that both administration type and spatial size matters for $\lambda_1$, which is smaller in size in urban districts and/or smaller administrative units. When we switch to contiguity based neighborhood, $\mathbf{W}_1 tfp$ and $\mathbf{W}_2 tfp$ interacting with *large* are significant. Surprisingly, the interaction of $\mathbf{W}_2 tfp$ and *county* is significant but the sign is hard to explain: It seems that counties benefit more from inter-regional spillovers than urban districts, controlling for spatial area. Except for this parameter, other estimates are all consistent with the previous ones. We conclude that both administration type and size jointly determines the strength of productivity spillovers.

# 5 Concluding remarks

In this article, we analyze the determinants of firm-level productivity in China's electric apparatus industry. The geolocational information provided by the *China Industrial Survey* data set allows us to perform a joint estimation of intra-regional and inter-regional effects with a spatial autoregressive model. Because of theoretical limitations, not all spatial econometric methods can be applied to the unbalanced panel data set. We show that the Kelejian and Prucha (1998) fixed effect 2SLS estimator and the Mutl and Pfaffermayr (2011) random effect FG2SLS estimator can be modified to use an unbalanced panel. The empirical estimates of our baseline model reveal strong correlation between the individual effects and the exogenous

[21] The 50% quantile is 1552 square kilometers.

regressors. The individual effects are also found to be large in size compared to the error terms. Based on these observations, we choose the fixed effects model for subsequent analyses.

Estimates of the baseline model shows strong spillovers among firms in the same district/county, while the spillover effects among neighboring administrative units are found to be small in size and insignificant. R&D and exports are found to contribute to higher productivity. Firms also benefit from favorable local market conditions, including competitiveness and public expenditure. However, specialization has little impact on the productivity of local firms. The market conditions in neighboring districts/counties have similar effects on productivity, but to a less extent.

The analyses on different types of administrative units reveal more information on the spillover effects. The special administrative and economic role of urban districts allows firms to interact more with peers in neighboring regions; while firms in small administrative units also have more inter-regional interactions because of shorter spatial distance. The empirical findings support both views. Although the spillover effects diminish in space, firms in urban districts or small administrative units benefit more from inter-regional spillovers because of their advantageous location.

The current research can be extended to provide more methodological rigor or empirical evidence. As we explained in section 2.4, the Kelejian-Prucha type within estimator is in capable of addressing the autoregressive structure in the error component if the panel is unbalanced. Consequently, our fixed effects model (10) has non-spherical error terms, which make the 2SLS estimator inefficient. A GMM procedure that estimates all parameters in (7) will restore efficiency, regardless of data type. On the other hand, Mutl and Pfaffermayr's (2011) Hausman test could be extended to unbalanced panels. Such a test can be constructed from the 1st stage within estimator and the random effects FG2SLS estimator.

The *China Industry Survey* data set provides a broad range of opportunities for empirical study. It would be interesting to see how the factors identified in the current study work on other industrial sectors. Another interesting topic is ownership structure. A large fraction of China's industrial firms are SOEs, which have gained increasing control over the market in recent years. Private firms and foreign-owned firms may differ substantially from SOEs in their capacities in generating or absorbing technological spillovers. A study on this topic could have important policy implications. Finally, the current study identifies public expenditure as a source of productivity growth, but it remains unclear how different types of public expenditure contribute to productivity. Further investigations is needed to cast some light on this question.

# References


Aitken, B. J., and A. E. Harrison (1999): "Do Domestic Firms Benefit from Direct Foreign Investment? Evidence from Venezuela," *American economic review*, 89(3), 605–618.

Aschauer, D. A. (1989): "Is Public Expenditure Productive?," *Journal of Monetary Economics*, 23(2), 177–200.

Aw, B. Y., M. J. Roberts, and D. Y. Xu (2009): "R&D Investment, Exporting, and Productivity Dynamics," Discussion paper, National Bureau of Economic Research.

Baltagi, B. H., P. H. Egger, and M. Kesina (2015): "Sources of Productivity Spillovers: Panel Data Evidence from China," *Journal of Productivity Analysis*, 43(3), 389–402.

Bao, Q., J. Huang, and Y. Wang (2015): "Productivity and Firms' Sales Destination: Chinese Characteristics," *Review of International Economics*.

Behrens, K., G. Duranton, and F. Robert-Nicoud (2014): "Productive Cities: Sorting, Selection, and Agglomeration," *Journal of Political Economy*, 122(3), 507–553.

Boeing, P., E. Mueller, and P. G. Sandner (2015): "China's R&D Explosion–Analyzing Productivity Effects Across Ownership Types and Over Time," *ZEW-Centre for European Economic Research Discussion Paper*, (15-006).

Brandt, L., J. Van Biesebroeck, and Y. Zhang (2012): "Creative Accounting or Creative Destruction? Firm-Level Productivity Growth in Chinese Manufacturing," *Journal of Development Economics*, 97(2), 339–351.

Clerides, S. K., S. Lach, and J. R. Tybout (1998): "Is Learning by Exporting Important? Micro-dynamic Evidence from Colombia, Mexico, and Morocco," *Quarterly Journal of Economics*, 113(3), 903–947.

Demurger, S. (2001): "Infrastructure Development and Economic Growth: an Explanation for Regional Disparities in China?," *Journal of Comparative Economics*, 29(1), 95–117.

Doraszelski, U., and J. Jaumandreu (2013): "R&D and Productivity: Estimating Endogenous Productivity," *The Review of Economic Studies*, 80(4), 1338–1383.

Elhorst, J. P. (2014): *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer, Heidelberg.

Fernald, J. G. (1999): "Roads to Prosperity? Assessing the Link between Public Capital and Productivity," *American Economic Review*, 89(3), 619–638.

Glaeser, E. L., H. D. Kallal, J. A. Scheinkman, and A. Shleifer (1992): "Growth in Cities," *Journal of Political Economy*, 100(6), 1126–1152.

Gomez-Antonio, M., and B. Fingleton (2012): "Analyzing the Impact of Public Capital Stock using the NEG Wage Equation: A Spatial Panel Data Approach," *Journal of Regional Science*, 52(3), 486–502.

Griffith, R., S. Redding, and J. Van Reenen (2004): "Mapping the Two Faces of R&D: Productivity Growth in a Panel of OECD Industries," *Review of Economics and Statistics*, 86(4), 883–895.

Grossman, G., and E. Helpman (1991): *Innovation and Growth in the World Economy*. The MIT Press, Cambridge, MA.

Hall, B. H., J. Mairesse, and P. Mohnen (2010): "Measuring the Returns to R&D," vol. 2, pp. 1033–1082. Elsevier.

Henderson, J. V. (2003): "Marshall's Scale Economies," *Journal of urban economics*, 53(1), 1–28.

Holtz-Eakin, D. (1994): "Public-Sector Capital and the Productivity Puzzle," *The Review of Economics and Statistics*, 76(1), 12–21.

Hu, A. G., and G. H. Jefferson (2004): "Returns to Research and Development in Chinese Industry: Evidence from State-Owned Enterprises in Beijing," *China Economic Review*, 15(1), 86–107.

Hu, A. G., G. H. Jefferson, and Q. Jinchang (2005): "R&D and Technology Transfer: Firm-level Evidence from Chinese Industry," *Review of Economics and Statistics*, 87(4), 780–786.

Hu, C., Z. Xu, and N. Yashiro (2015): "Agglomeration and Productivity in China: Firm Level Evidence," *China Economic Review*, 33(4), 50–66.

Hughes, G. (2013): "Estimating Spatial Panel Models Using Unbalanced Panels," United kingdom stata users' group meetings 2013, Stata Users Group.

Hulten, C. R., and R. M. Schwab (1991): "Public Capital Formation and the Growth of Regional Manufacturing Industries," *National Tax Journal*, 44(4), 121–134.

Jefferson, G. H., and J. Su (2006): "Privatization and Restructuring in China: Evidence from Shareholding Ownership, 1995–2001," *Journal of Comparative Economics*, 34(1), 146–166.

Kelejian, H. H., and I. R. Prucha (1998): "A Generalized Spatial Two-Stage Least Squares Procedure of Estimating a Spatial Autoregressive Model with Autoregressive Disturbances," *Journal of Real Estate Finance and Economics*, 17(1), 99–121.

Kelejian, H. H., I. R. Prucha, and Y. Yuzefovich (2004): "Instrumental Variable Estimation of A Spatial Autoregressive Model with Autoregressive Disturbances: Large and Small Sample Results," in *Advances in Econometrics, Volume 18*, ed. by J. P. Lesage, and R. K. Pace, pp. 163–198. Emerald Group Publishing Limited, United Kingdom.

Keller, W., and S. R. Yeaple (2009): "Multinational Enterprises, International Trade, and Productivity Growth: Firm-level Evidence from the United States," *The Review of Economics and Statistics*, 91(4), 821–831.

LeSage, J., and R. K. Pace (2009): *Introduction to Spatial Econometrics*. CRC Press, Boca Raton, FL.

Levinsohn, J., and A. Petrin (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 70(2), 317–341.

Lu, D. (2010): "Exceptional Exporter Performance? Evidence from Chinese Manufacturing Firms," Working paper, University of Chicago.

Lu, J., Y. Lu, and Z. Tao (2010): "Exporting Behavior of Foreign Affiliates: Theory and Evidence," *Journal of International Economics*, 81(2), 197–205.

Marshall, A. (1890): *Principles of Economics*. Macmillan, London.

Martin, P., T. Mayer, and F. Mayneris (2011): "Spatial Concentration and Plant-Level Productivity in France," *Journal of Urban Economics*, 2(69), 182–195.

Melitz, M. J. (2003): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71(6), 1695–1725.

Mutl, J., and M. Pfaffermayr (2011): "The Hausman Test in a Cliff and Ord Panel Model," *The Econometrics Journal*, 14(1), 48–76.

Nickell, S. J. (1996): "Competition and Corporate Performance," *Journal of Political Economy*, 104(4), 724–746.

Nie, H., T. Jiang, and R. Yang (2012): "Chinese Industrial Enterprise Database: Current Use and Potential Problems," *Journal of World Economy*, 143(5), 142–158.

Olley, G. S., and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64(6), 1263–1297.

Porter, M. E. (1990): *The Competitive Advantage of Nations*. Free Press, New York.

Rosenthal, S. S., and W. C. Strange (2003): "Geography, Industrial Organization, and Agglomeration," *Review of Economics and Statistics*, 85(2), 377–393.

Sheng, Y., and L. Song (2013): "Re-estimation of Firms' Total Factor Productivity in China's Iron and Steel Industry," *China Economic Review*, 24(3), 177–188.

Syverson, C. (2011): "What Determines Productivity?," *Journal of Economics Literature*, 49(2), 326–365.

Vijverberg, W. P., F.-C. Fu, and C.-P. C. Vijverberg (2011): "Public Infrastructure as a Determinant of Productive Performance in China," *Journal of Productivity Analysis*, 36(1), 91–111.

Wei, Y., and X. Liu (2006): "Productivity Spillovers from R&D, Exports and FDI in China's Manufacturing Sector," *Journal of International Business Studies*, 37(4), 544–557.

Yu, N., M. De Jong, S. Storm, and J. Mi (2013): "Spatial Spillover Effects of Transport Infrastructure: Evidence from Chinese Regions," *Journal of Transport Geography*, 28, 56–66.

# A Data preparation

## A.1 The China Industry Survey Data Set

The China Industry Survey data set is known (Nie, Jiang, and Yang, 2012) to contain numerous errors and internal inconsistencies. A major challenge we face is the identification problem: Even though the raw data provides detailed information on firm ID, geolocation, and industrial sector of each establishment, in many cases their values are non-unique in the survey period. The obscurity is caused by recording errors and revisions in the coding system. These key identifiers must be cleaned before we can build a longitudinal data set. Our data-cleaning procedure is described below.

Create the unique ID

Each observation (establishment) is jointly identified by the organization code and the name of the business. The organization code is the official identification issued by the registration office, which should remain unchanged throughout the life cycle of the establishment. On the other hand, the name of the business is supposed to be unique but is subject to change over time. To resolve the indeterminacy in organization codes, we adopt the matching procedure proposed by Nie, Jiang, and Yang (2012) as follows.

1. Pool the observations from all annual data sets, and group them by organization code. The observations in each group share the same organization code. If observations from the same group have more than one business names, it indicates that the establishment changed its business name in the corresponding year.

2. For each group $G$ constructed in the previous step, find other groups that has at least one observation sharing the same business name with an observation from group $G$. Once these

groups are identified, append them to $G$, and remove duplicated observations from the resulting group $G'$.[22] The updated group $G'$ may contain more than one organization codes.

3. Repeat the previous step until the group structure no longer changes, then remove all duplicated groups. Each of the remaining group represents a unique establishment in the longitudinal data, to which an unique ID is assigned. We use the organization code associated with the group as the unique ID.

The raw data consists of 658213 unique organization codes over 1996-2010. The algorithm converges in three iterations, retaining 620020 groups. The numbers indicate that 6% of full sample might have been misidentified as independent establishments without this treatment.[23]

Among the 620020 unique establishments identified in the previous step, 8825 have multiple observations in one or more years. About half of them (4429) have more than two observations but only one multiplicity, in which case we compare the values of accounting and financial variables with those in adjacent years. The observation with a closer match is chosen and the other one is discarded. The observations associated with the remaining 4329 establishments are discarded.[24]

Identify the geolocation

The geolocation of each establishment is identified by a six-digit administrative code. The data sets provide the six-digit (1996-2003) or twelve-digit (2004-2010) administrative code plus the six-digit zip code for each observation. The first six digits of the administration code define an administrative unit in the district/county level, while the last six digits define the township and community. In this study, we use the first six digits to identify the location of each establishment. We clean the data on administrative codes by the following procedure:

1. The administrative codes in the data are formatted with different versions of the GB/T 2260 standard.[25] We construct a junction table that maps earlier versions of the map code into the 2007 revision. All valid administrative codes are then converted into the values specified by the 2007 revision of the national standard (GB/T 2260-2007).

2. For the 16771 observations that do not have valid administrative codes, we constructed a table that maps 37249 zip codes to six-digit administrative codes (2007 version). Their administrative codes are then recovered from zip codes if the latter can be found in our mapping table.

3. If the previous steps result in a unique administrative code for an establishment in all years, it is used to identify the geolocation. 30811 establishments are found to have multiple values, in which case we choose the most frequent one.

Identify the industrial division

The industrial division of each establishment is identified by a four-digit industry code. The data set uses two different coding systems: GB/T 4754-1994 until 2002 and GB/T 4754-2002 afterwards. Old industry codes are mapped to their new values specified by GB/T 4754-2002. Similar to what we did to administrative codes, the industry codes are further cleaned so that all observations associated with an establishment are assigned a unique industry code.

[22] Note that a group may be simultaneously appended to multiple groups, and that the group-combining operation may result duplicated groups that have identical set of observations.

[23] As far as the organization code and business name are changed one at a time, the iterated matching procedure is able to track down the same establishment over time, even if multiple changes take place in succession. However, the algorithm fails whenever an establishment changes the organization code and business name at the same time.

[24] 2047 establishments have only two observations, both of which are recorded in the same year.

[25] The administrative divisions have undergone four major revisions in 1995, 1999, 2002, and 2007. Each time, a large number of administrative divisions were either renamed, merged, split, or (newly) created. In most cases, the administrative division affected was assigned a new six-digit code. The old administrative code was abolished and won't be used in the future. The 1995 revision of the code map, GB/T 2260-1995 consists of 3404 distinct administrative codes, 1200 of which were revoked by 2007. The changes being made are tabulated in the appendix of the publication.

## A.2 The Budgetary Expenditure Data Set

There are two major issues with the fiscal data. First, it uses a different administrative division than the one specified by the official code map. Data loss is inevitable when we merge data. Nevertheless, we are able to retain most of the administrative units in the district/county level.[26] The city of Shenzhen was treated as one piece by the yearbook until 2007. The city is known to be a major manufacturing hub in the Pearl River Delta, hosting a large number of firms. In order to retain these observations in our sample, we treat the five urban districts of Shenzhen as one administrative unit. Second, the expenditure categories were revised twice during the study period, first in 2003 and then in 2007. As a result, the variables on expenditure types are not compatible in different issue of the yearbook, except for total budgetary expenditure.

[26]The number of dropouts ranges from 124 in 1999 to 34 in 2006.

Table 1: Summary Statistics of the China Industry Survey Dataset, 1996-2010

| year | full sample | | | above-scale firms[1] | | | percent of national aggregate[2] | | | change in sample[3] | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | count | output | employees | count | output | employees | count | output | employees | in | out |
| 1996 | 23480 | 3383.75 | 375.16 | 21921 | 3349.19 | 365.76 | – | – | – | – | – |
| 1997 | 23889 | 3653.88 | 369.57 | 21671 | 3578.76 | 356.59 | – | – | – | 2374 | 1965 |
| 1998 | 163810 | 6690.22 | 612.80 | 123351 | 6525.97 | 563.84 | 74.7% | 96.3% | 59.0% | 143162 | 3241 |
| 1999 | 160623 | 7174.11 | 573.57 | 120709 | 7001.87 | 531.34 | 74.5% | 96.3% | 91.5% | 23801 | 26988 |
| 2000 | 161401 | 8441.10 | 549.41 | 125140 | 8274.83 | 512.93 | 76.8% | 96.6% | 92.2% | 26088 | 25310 |
| 2001 | 159703 | 8772.10 | 503.62 | 129458 | 8616.13 | 473.15 | 75.6% | 90.3% | 87.0% | 39519 | 41217 |
| 2002 | 179871 | 10915.00 | 545.11 | 150171 | 10757.41 | 517.95 | 82.7% | 97.1% | 93.8% | 41862 | 21694 |
| 2003 | 194452 | 14018.17 | 567.64 | 170950 | 13886.01 | 547.60 | 87.1% | 97.6% | 95.3% | 41265 | 26684 |
| 2004 | 276826 | 15636.75 | 654.10 | 259219 | 15627.63 | 645.00 | 93.8% | 77.5% | 97.4% | 126147 | 43773 |
| 2005 | 269752 | 24834.29 | 681.88 | 256486 | 24652.75 | 672.47 | 94.4% | 98.0% | 97.5% | 38885 | 45959 |
| 2006 | 299599 | 31198.10 | 727.40 | 287773 | 30906.18 | 720.55 | 95.3% | 97.6% | 97.9% | 55030 | 25183 |
| 2007 | 334151 | 39943.59 | 778.38 | 327118 | 39730.99 | 774.30 | 97.1% | 98.1% | 98.3% | 61981 | 27429 |
| 2008 | 260106 | 37652.23 | 639.00 | 257080 | 37548.71 | 637.28 | 60.3% | 74.0% | 72.1% | 7787 | 81832 |
| 2009 | 202724 | 32049.91 | 505.83 | 200863 | 32032.94 | 505.06 | 46.2% | 58.4% | 57.2% | 0 | 57382 |
| 2010 | 342077 | 85057.47 | 1290.07 | 335441 | 84824.55 | 1286.28 | 74.1% | 121.4% | 134.76% | 160572 | 21219 |

[1]Above-scale firms are those with an annual revenue from principal business over 5 million CNY.
[2]The sub-total of above-scale firms in the sample as a percentage of the national aggregate reported by the *China Statistical Yearbook* (2012 issue).
[3]Numbers of entrants and exists in the current year.

Table 2: Summary statistics of urban districts and counties, sample average in 2005-2007

| | urban districts | | | | | counties | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | min | 1st quartile | median | 3rd quartile | max | min | 1st quartile | median | 3rd quartile | max |
| area[1] | 7.3 | 105.4 | 324.8 | 810.0 | 6653.0 | 167.5 | 1004.0 | 1449.0 | 2103.0 | 8928.0 |
| number of firms | 0.7 | 3.0 | 6.0 | 15.3 | 755.3 | 0.7 | 2.0 | 3.7 | 9.3 | 466.7 |
| employment | 10.0 | 369.2 | 1160.0 | 3472.0 | 357200.0 | 17.0 | 178.3 | 497.0 | 1682.0 | 90050.0 |
| public expenses[1] | 70.3 | 299.9 | 577.6 | 1016.0 | 63280.0 | 164.4 | 488.4 | 677.8 | 975.4 | 6064.0 |

[1] Units: square kilometer (area) and RMB 1 million (public expenditure).

Table 3: Sample division by administration type and size, 2005-2007 average

| | small | | large | |
|---|---|---|---|---|
| | number of admin units | number of firms | number of admin units | number of firms |
| urban districts | 492 | 7766.3 | 46 | 1772.7 |
| counties | 270 | 3369.7 | 213 | 2445.3 |

Table 4: The base model: contiguous districts/counties treated as neighbors, 84727 observations on 26174 distinct firms

Dependent variable: *tfp*

| regressor | FE[1] | FE-IV[2] | FE-GMM[2] | RE-FG2SLS |
|---|---|---|---|---|
| *W1_tfp* | 0.185** | 0.403** | 0.410** | 0.673** |
| | (0.011) | (0.074) | (0.075) | (0.008) |
| *W2_tfp* | 0.130** | 0.111 | 0.118 | 0.134** |
| | (0.013) | (0.096) | (0.092) | (0.008) |
| *rd* | 0.181** | 0.192** | 0.185** | 0.389** |
| | (0.027) | (0.030) | (0.029) | (0.016) |
| *ex* | 0.083** | 0.088** | 0.087** | 0.070** |
| | (0.025) | (0.028) | (0.028) | (0.009) |
| *spec* | 0.019* | -0.006 | -0.006 | -0.011** |
| | (0.009) | (0.014) | (0.014) | (0.002) |
| *comp* | -0.634** | -0.832** | -0.863** | 0.132** |
| | (0.075) | (0.123) | (0.123) | (0.015) |
| *pub* | 0.234** | 0.137** | 0.132** | -0.009** |
| | (0.013) | (0.036) | (0.036) | (0.003) |
| *W2_spec* | -0.030** | -0.021 | -0.019 | -0.035** |
| | (0.009) | (0.013) | (0.013) | (0.002) |
| *W2_comp* | -0.573** | -0.421** | -0.400** | 0.031* |
| | (0.077) | (0.135) | (0.133) | (0.017) |
| *W2_pub* | 0.030** | 0.019** | 0.018** | -0.002 |
| | (0.005) | (0.006) | (0.006) | (0.002) |
| *Intercept* | 1.477** | 1.279** | 1.234** | 1.261** |
| | (0.139) | (0.159) | (0.158) | (0.040) |
| $R^2$ | 0.13 | 0.12 | — | 0.73 |
| $\hat{\sigma}_\mu$ | 1.166 | 1.137 | — | 1.399 |
| $\hat{\sigma}_u$ $(\hat{\sigma}_\epsilon)^3$ | 0.666 | 0.671 | — | 0.560 |
| $\hat{\rho}$ | — | — | — | -0.962 |

Significance codes: '**' 0.05, '*' 0.10

[1] Clustered standard errors in parentheses

[2] Bootstrapped standard errors in parentheses

[3] In columns (FE) and (FE-IV), the numbers are standard errors of $\mathbf{Q}_0\mathbf{u}$ in (10); in column (RE-FG2SLS), the number is $\hat{\sigma}_\epsilon$ estimated by (16).

Table 5: The base model: districts/counties within 50 kilometers treated as neighbors, 81331 observations on 24903 distinct firms

Dependent variable: *tfp*

| regressor | FE[1] | FE-IV[2] | FE-GMM[2] | RE-FG2SLS |
|---|---|---|---|---|
| *W1_tfp* | 0.163** | 0.354** | 0.360** | 0.592** |
| | (0.011) | (0.066) | (0.070) | (0.010) |
| *W2_tfp* | 0.181** | 0.088 | 0.087 | 0.188** |
| | (0.015) | (0.081) | (0.078) | (0.009) |
| *rd* | 0.179** | 0.186** | 0.178** | 0.465** |
| | (0.027) | (0.027) | (0.027) | (0.017) |
| *ex* | 0.083** | 0.087** | 0.086** | 0.055** |
| | (0.025) | (0.026) | (0.026) | (0.009) |
| *spec* | 0.020** | -0.004 | -0.004 | -0.002 |
| | (0.009) | (0.011) | (0.011) | (0.002) |
| *comp* | -0.590** | -0.776** | -0.808** | 0.118** |
| | (0.074) | (0.104) | (0.111) | (0.017) |
| *pub* | 0.210** | 0.168** | 0.168** | -0.020** |
| | (0.013) | (0.036) | (0.034) | (0.003) |
| *W2_spec* | -0.042** | -0.022 | -0.019 | -0.050** |
| | (0.009) | (0.016) | (0.016) | (0.002) |
| *W2_comp* | -0.743** | -0.524** | -0.495** | 0.055** |
| | (0.081) | (0.122) | (0.127) | (0.020) |
| *W2_pub* | 0.023** | 0.019** | 0.019** | -0.003 |
| | (0.004) | (0.007) | (0.006) | (0.002) |
| *Intercept* | 1.689** | 1.498** | 1.468** | 1.524** |
| | (0.141) | (0.191) | (0.198) | (0.042) |
| $R^2$ | 0.14 | 0.13 | — | 0.70 |
| $\hat{\sigma}_\mu$ | 1.140 | 1.133 | — | 1.325 |
| $\hat{\sigma}_u$ $(\hat{\sigma}_\epsilon)^3$ | 0.663 | 0.667 | — | 0.588 |
| $\hat{\rho}$ | — | — | — | -0.824 |

Significance codes: '**' 0.05, '*' 0.10

[1]Clustered standard errors in parentheses

[2]Bootstrapped standard errors in parentheses

[3]In columns (FE) and (FE-IV), the numbers are standard errors of $\mathbf{Q}_0\mathbf{u}$ in (10); in column (RE-FG2SLS), the number is $\hat{\sigma}_\epsilon$ estimated by (16).

Table 6: Spillover effects on urban districts and counties: districts/counties within 50 kilometers treated as neighbors.

Dependent variable: *tfp*[1]

| regressor | baseline | urban districts | counties | full model |
|---|---|---|---|---|
| *W1_tfp* | 0.354** | 0.204** | 0.587** | 0.246** |
| | (0.066) | (0.079) | (0.097) | (0.065) |
| *W2_tfp* | 0.088 | 0.203** | 0.092 | 0.135** |
| | (0.081) | (0.101) | (0.099) | (0.068) |
| interactions: | | | | |
| *W1_tfp×county* | — | — | — | 0.343** |
| | | | | (0.066) |
| *W2_tfp×county* | — | — | — | -0.088 |
| | | | | 0.057 |
| *rd* | 0.186** | 0.157** | 0.206** | 0.176** |
| | (0.027) | (0.039) | (0.037) | (0.028) |
| *ex* | 0.087** | 0.053* | 0.154** | 0.084** |
| | (0.026) | (0.031) | (0.040) | (0.025) |
| *spec* | -0.004 | 0.000 | 0.004 | -0.002 |
| | (0.011) | (0.018) | (0.015) | (0.010) |
| *comp* | -0.776** | -0.511** | -0.987** | -0.733** |
| | (0.104) | (0.126) | (0.152) | (0.101) |
| *pub* | 0.168** | 0.136** | 0.162** | 0.157** |
| | (0.036) | (0.039) | (0.034) | (0.030) |
| *W2_spec* | -0.022 | -0.023 | -0.035* | -0.020 |
| | (0.016) | (0.017) | (0.019) | (0.014) |
| *W2_comp* | -0.524** | -0.480** | -0.720** | -0.503** |
| | (0.122) | (0.134) | (0.217) | (0.107) |
| *W2_pub* | 0.019** | 0.026** | -0.021** | 0.021** |
| | (0.007) | (0.006) | (0.010) | (0.006) |
| *Intercept* | 1.498** | 1.979** | 0.316 | 1.358** |
| | (0.191) | (0.199) | (0.315) | (0.179) |
| $N$ | 81331 | 51480 | 29851 | 81331 |
| $n$ | 24903 | 15575 | 9328 | 24903 |
| $R^2$ | 0.13 | 0.10 | 0.18 | 0.13 |
| $\hat{\sigma}_\mu$ | 1.133 | 1.144 | 1.114 | 1.404 |
| $\hat{\sigma}_u$ | 0.667 | 0.699 | 0.605 | 0.667 |

Significance codes: '**' 0.05, '*' 0.10

[1]All regressions estimated by fixed effects 2SLS, bootstrapped standard errors in parentheses.

Table 7: Spillover effects on urban districts and counties: contiguous units treated as neighbors.

Dependent variable: *tfp*[1]

| regressor | baseline | urban districts | counties | full model |
|---|---|---|---|---|
| *W1_tfp* | 0.403** | 0.320** | 0.542** | 0.314** |
| | (0.074) | (0.071) | (0.112) | (0.071) |
| *W2_tfp* | 0.111 | 0.137* | 0.175 | 0.135 |
| | (0.096) | (0.082) | (0.140) | (0.084) |
| interactions: | | | | |
| *W1_tfp*× *county* | — | — | — | 0.300** |
| | | | | (0.067) |
| *W2_tfp*× *county* | — | — | — | -0.037 |
| | | | | (0.070) |
| *rd* | 0.192** | 0.167** | 0.213** | 0.183** |
| | (0.030) | (0.030) | (0.044) | (0.029) |
| *ex* | 0.088** | 0.057** | 0.149** | 0.085** |
| | (0.028) | (0.026) | (0.029) | (0.027) |
| *spec* | -0.006 | -0.016 | 0.014 | -0.004 |
| | (0.014) | (0.015) | (0.019) | (0.014) |
| *comp* | -0.832** | -0.643** | -0.944** | -0.791** |
| | (0.123) | (0.099) | (0.153) | (0.117) |
| *pub* | 0.137** | 0.121** | 0.123 | 0.124** |
| | (0.036) | (0.036) | (0.088) | (0.032) |
| *W2_spec* | -0.021 | -0.014 | -0.046** | -0.023** |
| | (0.013) | (0.014) | (0.022) | (0.012) |
| *W2_comp* | -0.421** | -0.359** | -0.588** | -0.405** |
| | (0.135) | (0.123) | (0.198) | (0.123) |
| *W2_pub* | 0.019** | 0.026** | -0.012 | 0.023** |
| | (0.006) | (0.008) | (0.012) | (0.006) |
| *Intercept* | 1.279** | 1.793** | 0.045 | 1.168** |
| | (0.159) | (0.198) | (0.257) | (0.153) |
| $N$ | 84727 | 54513 | 30214 | 84727 |
| $n$ | 26174 | 16679 | 9495 | 26174 |
| $R^2$ | 0.12 | 0.12 | 0.19 | 0.12 |
| $\hat{\sigma}_\mu$ | 1.137 | 1.147 | 1.100 | 1.398 |
| $\hat{\sigma}_u$ | 0.671 | 0.705 | 0.603 | 0.672 |

Significance codes: '**' 0.05, '*' 0.10

[1] All regressions estimated by fixed effects 2SLS, bootstrapped standard errors in parentheses.

Table 8: Spillover effects on small and large administrative units: units within 50 kilometers treated as neighbors.

Dependent variable: *tfp*[1]

| regressor | baseline | small admins | large admins | model 1 | model 2 |
|---|---|---|---|---|---|
| *W1_tfp* | 0.354** | 0.286** | 0.584** | 0.277** | 0.219** |
| | (0.066) | (0.087) | (0.104) | (0.056) | (0.058) |
| *W2_tfp* | 0.088 | 0.196** | -0.065 | 0.141* | 0.117* |
| | (0.081) | (0.096) | (0.087) | (0.080) | (0.070) |
| interactions: | | | | | |
| *W1_tfp*×*large* | — | — | — | 0.400** | 0.317** |
| | | | | (0.070) | (0.083) |
| *W1_tfp*×*county* | — | — | — | — | 0.163** |
| | | | | | (0.075) |
| *W2_tfp*×*large* | — | — | — | -0.162** | -0.137 |
| | | | | (0.073) | (0.091) |
| *W2_tfp*×*county* | — | — | — | — | 0.035 |
| | | | | | (0.073) |
| *rd* | 0.186** | 0.206** | 0.099** | 0.176** | 0.170** |
| | (0.027) | (0.029) | (0.038) | (0.027) | (0.028) |
| *ex* | 0.087** | 0.079* | 0.116** | 0.092** | 0.088** |
| | (0.026) | (0.033) | (0.040) | (0.025) | (0.025) |
| *spec* | -0.004 | -0.020 | 0.004 | -0.016* | -0.009 |
| | (0.011) | (0.013) | (0.023) | (0.009) | (0.010) |
| *comp* | -0.776** | -0.725** | -0.869** | -0.759** | -0.709** |
| | (0.104) | (0.133) | (0.190) | (0.090) | (0.091) |
| *pub* | 0.168** | 0.121** | 0.209** | 0.147** | 0.164** |
| | (0.036) | (0.042) | (0.062) | (0.038) | (0.032) |
| *W2_spec* | -0.022 | -0.018 | -0.050** | -0.024* | -0.025* |
| | (0.016) | (0.016) | (0.020) | (0.014) | (0.013) |
| *W2_comp* | -0.524** | -0.574** | -0.403* | -0.511** | -0.526** |
| | (0.122) | (0.131) | (0.222) | (0.098) | (0.098) |
| *W2_pub* | 0.019** | 0.004 | 0.023** | 0.012* | 0.017** |
| | (0.007) | (0.008) | (0.009) | (0.006) | (0.006) |
| *Intercept* | 1.498** | 1.923** | 0.474* | 1.577** | 1.493** |
| | (0.191) | (0.180) | (0.282) | (0.176) | (0.166) |
| $N$ | 81331 | 60463 | 20868 | 81331 | 81331 |
| $n$ | 24903 | 18358 | 6545 | 24903 | 24903 |
| $R^2$ | 0.13 | 0.11 | 0.20 | 0.13 | 0.13 |
| $\hat{\sigma}_\mu$ | 1.133 | 1.118 | 1.136 | 1.384 | 1.507 |
| $\hat{\sigma}_u$ | 0.667 | 0.681 | 0.619 | 0.667 | 0.666 |

Significance codes: '**' 0.05, '*' 0.10

[1]All regressions estimated by fixed effects 2SLS, bootstrapped standard errors in parentheses.

Table 9: Spillover effects on small and large administrative units: contiguous units treated as neighbors.

Dependent variable: *tfp*[1]

| regressor | baseline | small admins | large admins | model 1 | model 2 |
|---|---|---|---|---|---|
| *W1_tfp* | 0.403** | 0.278** | 0.679** | 0.298** | 0.278** |
| | (0.074) | (0.083) | (0.161) | (0.075) | (0.071) |
| *W2_tfp* | 0.111 | 0.261** | -0.023 | 0.218** | 0.178** |
| | (0.096) | (0.096) | (0.108) | (0.098) | (0.084) |
| interactions: | | | | | |
| *W1_tfp*×*large* | — | — | — | 0.438** | 0.520** |
| | | | | (0.077) | (0.086) |
| *W1_tfp*×*county* | — | — | — | — | -0.034 |
| | | | | | (0.078) |
| *W2_tfp*×*large* | — | — | — | -0.227** | -0.373** |
| | | | | (0.085) | (0.098) |
| *W2_tfp*×*county* | — | — | — | — | 0.247** |
| | | | | | (0.085) |
| *rd* | 0.192** | 0.202** | 0.142** | 0.184** | 0.179** |
| | (0.030) | (0.036) | (0.034) | (0.030) | (0.030) |
| *ex* | 0.088** | 0.075** | 0.122** | 0.093** | 0.093** |
| | (0.028) | (0.034) | (0.037) | (0.027) | (0.027) |
| *spec* | -0.006 | -0.013 | 0.001 | -0.015 | -0.007 |
| | (0.014) | (0.014) | (0.032) | (0.013) | (0.014) |
| *comp* | -0.832** | -0.707** | -1.006** | -0.796** | -0.754** |
| | (0.123) | (0.127) | (0.196) | (0.120) | (0.116) |
| *pub* | 0.137** | 0.097** | 0.127 | 0.103** | 0.108** |
| | (0.036) | (0.036) | (0.082) | (0.033) | (0.028) |
| *W2_spec* | -0.021 | -0.026* | -0.042** | -0.029** | -0.032** |
| | (0.013) | (0.014) | (0.021) | (0.012) | (0.012) |
| *W2_comp* | -0.421** | -0.481** | -0.418** | -0.436** | -0.447** |
| | (0.135) | (0.118) | (0.203) | (0.124) | (0.115) |
| *W2_pub* | 0.019** | 0.008 | 0.020 | 0.011* | 0.014** |
| | (0.006) | (0.007) | (0.013) | (0.006) | (0.006) |
| *Intercept* | 1.279** | 1.692** | 0.385* | 1.316** | 1.258** |
| | (0.159) | (0.168) | (0.233) | (0.142) | (0.146) |
| $N$ | 84727 | 60218 | 24509 | 84727 | 84727 |
| $n$ | 26174 | 18324 | 7850 | 26174 | 26174 |
| $R^2$ | 0.12 | 0.10 | 0.17 | 0.12 | 0.12 |
| $\hat{\sigma}_\mu$ | 1.137 | 1.114 | 1.172 | 1.373 | 1.465 |
| $\hat{\sigma}_u$ | 0.671 | 0.683 | 0.638 | 0.671 | 0.672 |

Significance codes: '**' 0.05, '*' 0.10

[1] All regressions estimated by fixed effects 2SLS, bootstrapped standard errors in parentheses.

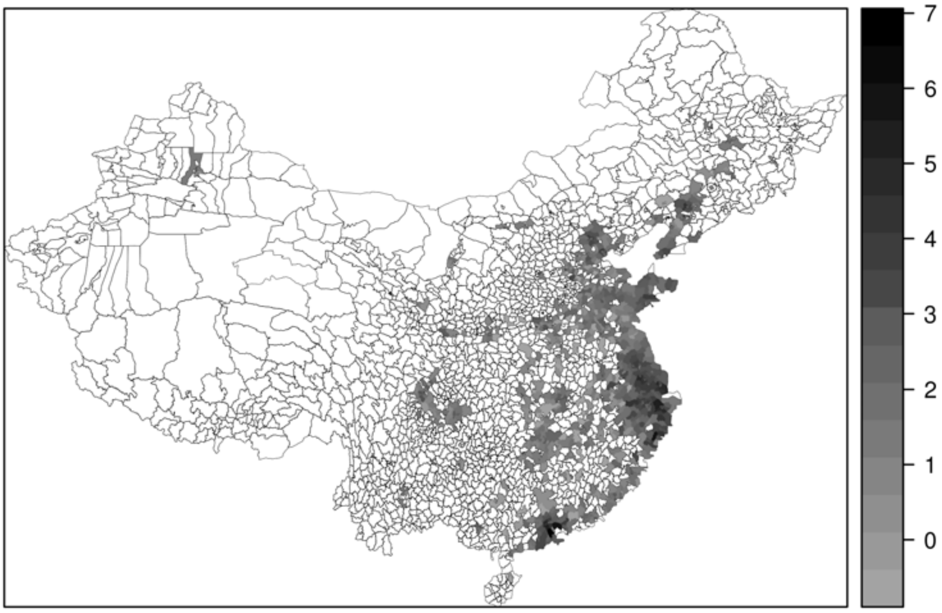


Figure 1: District/county subtotal: number of firms in sector 3900, 2005-2007 average.

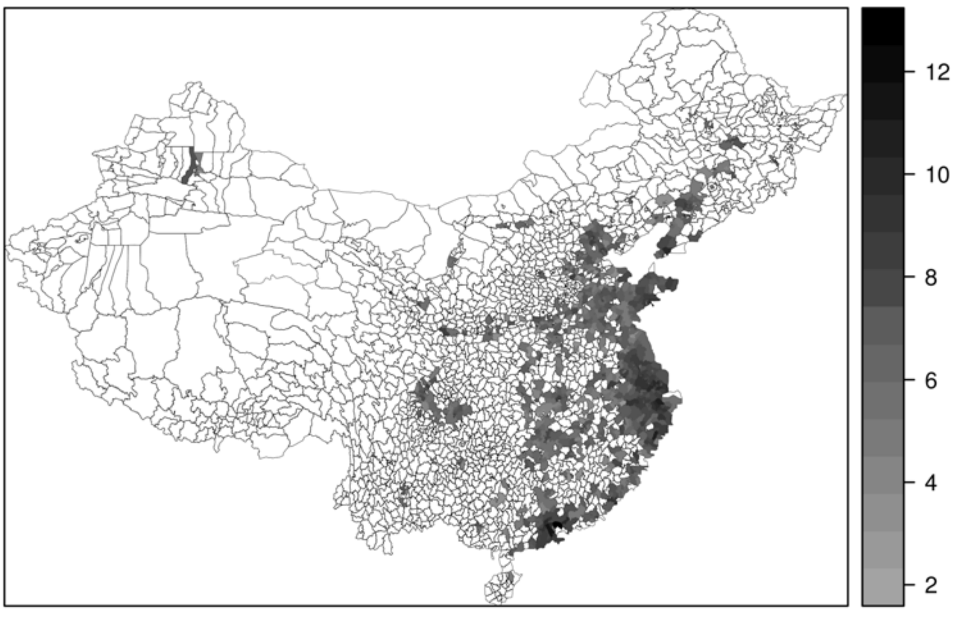


Figure 2: District/county subtotal: employment in sector 3900, 2005-2007 average.

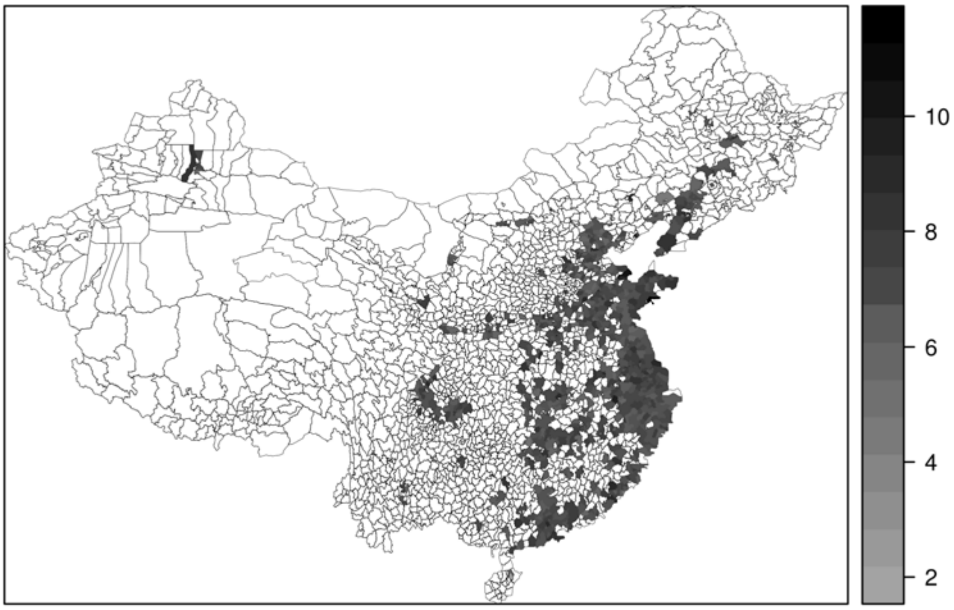


Figure 3: District/county average TFP in sector 3900, futher averaged over 2005-2007.

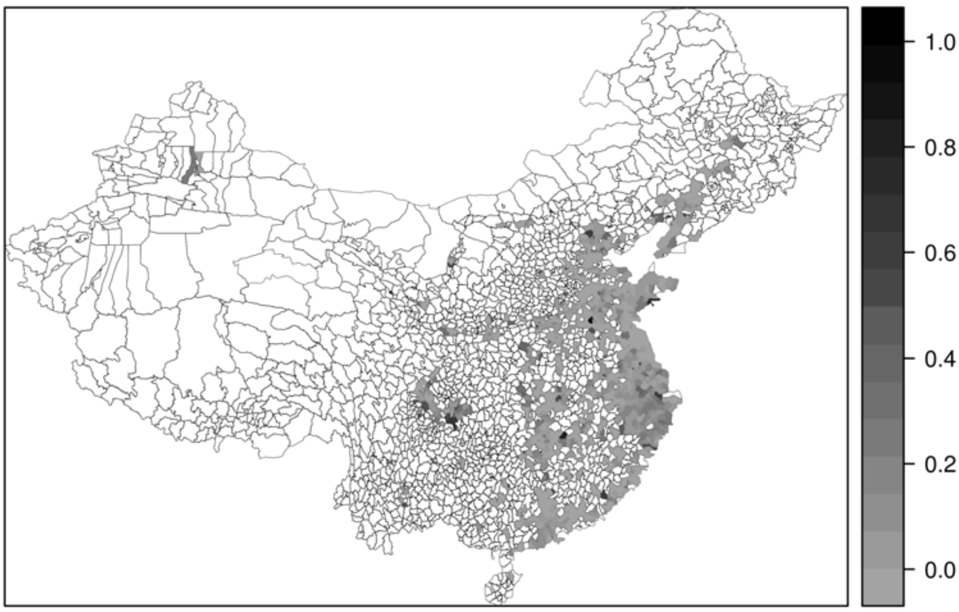


Figure 4: District/county average new product-total ouput ratio in sector 3900, futher averaged over 2005-2007.

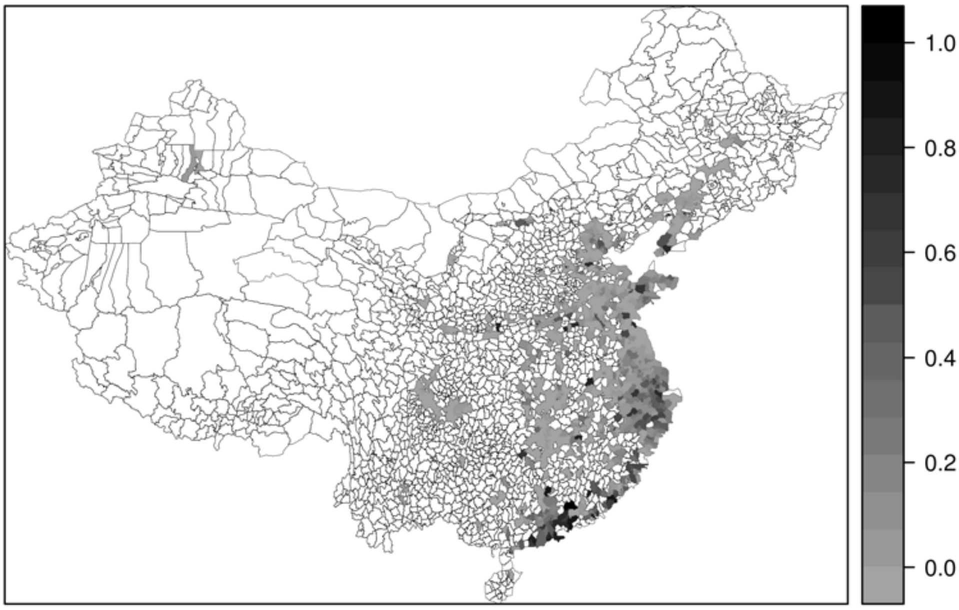


Figure 5: District/county average export-sales ratio in sector 3900, futher averaged over 2005-2007.

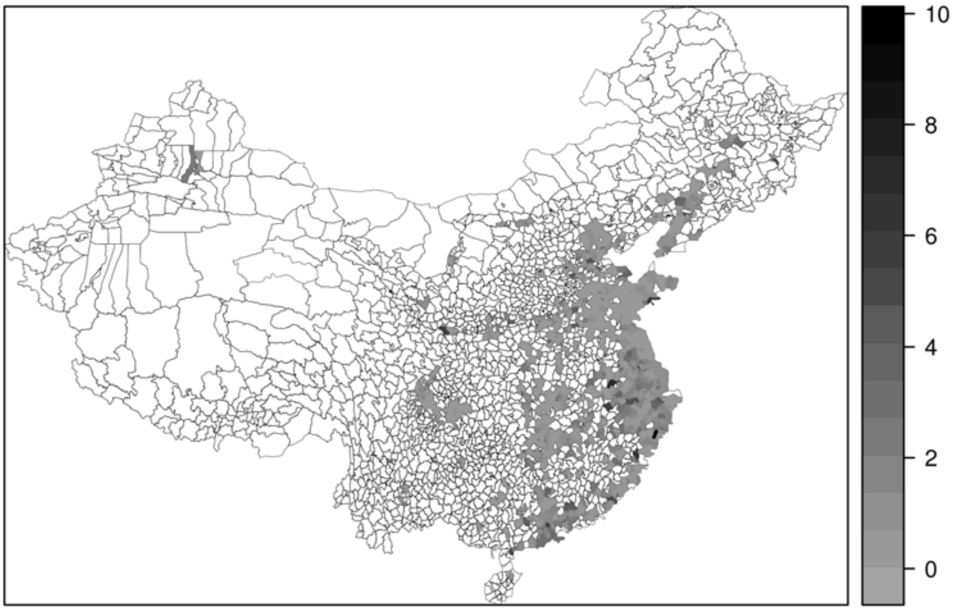


Figure 6: District/county specialization index in sector 3900, futher averaged over 2005-2007.

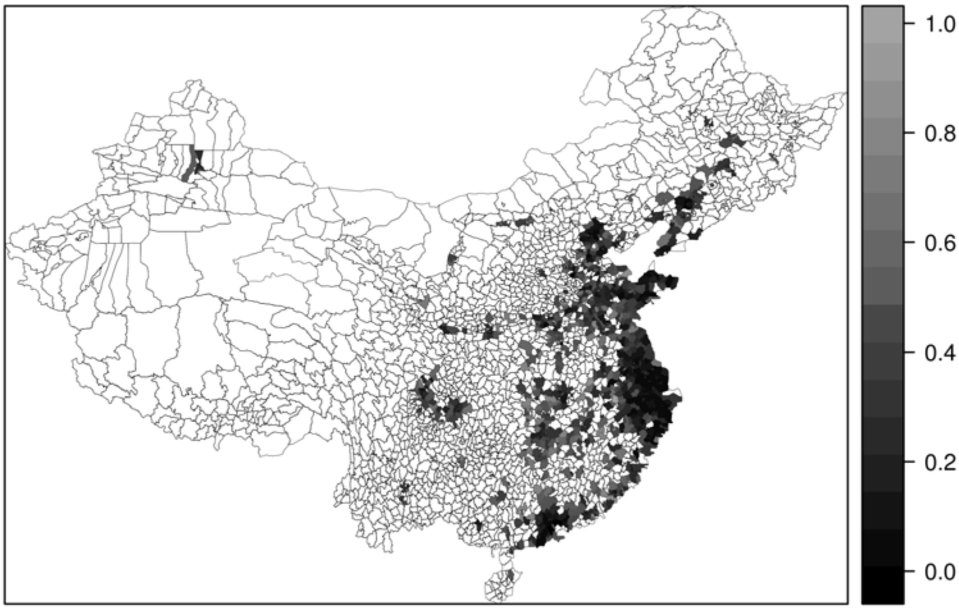


Figure 7: District/county Herfindahl-Hirschman index in sector 3900, futher averaged over 2005-2007.

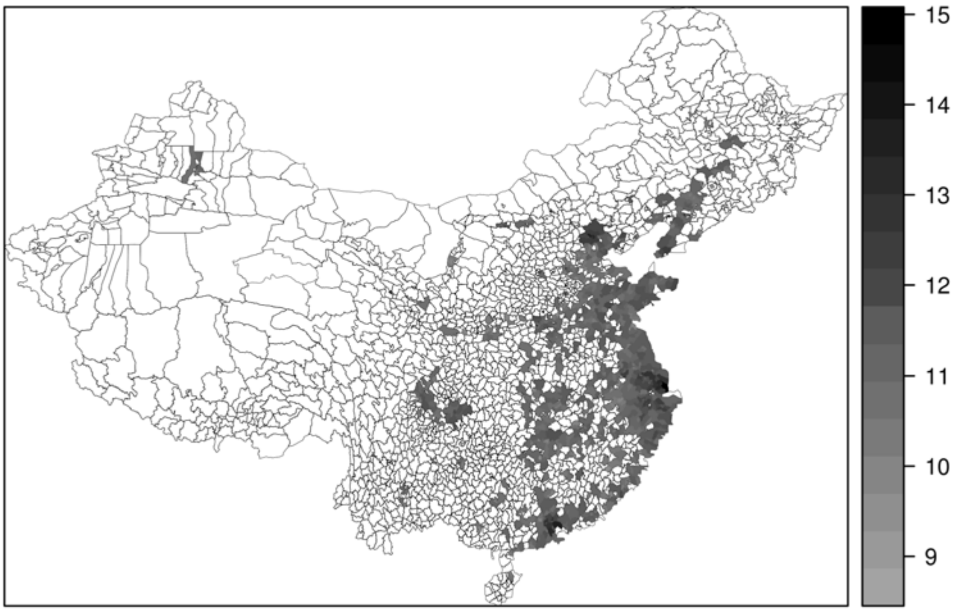


Figure 8: District/county total budgetary expenditure (log-transformed) in sector 3900, futher averaged over 2005-2007.