INSTITUTE OF DEVELOPING ECONOMIES

IDE Discussion Papers are preliminary materials circulated
to stimulate discussions and critical comments

IDE DISCUSSION PAPER No. 495

Agglomeration Effects of Informal Sector: Evidence from Cambodia

Kiyoyasu TANAKA* and Yoshihiro HASHIGUCHI

May 2015

Abstract
The presence of a large informal sector in developing economies poses the question of whether informal activity produces agglomeration externalities. This paper uses data on all the nonfarm establishments and enterprises in Cambodia to estimate the impact of informal agglomeration on the regional economic performance of formal and informal firms. We develop a Bayesian approach for a spatial autoregressive model with an endogenous explanatory variable to address endogeneity and spatial dependence. We find a significantly positive effect of informal agglomeration, where informal firms gain more strongly than formal firms. Calculating the spatial marginal effects of increased agglomeration, we demonstrate that more accessible regions are more likely than less accessible regions to benefit strongly from informal agglomeration.

Keywords: Agglomeration, Informal Sector, Cambodia, Bayesian
JEL classification: C11, C21, C26, H26, O17, R12

* *Corresponding author*: Inter-Disciplinary Studies Center, Institute of Developing Economies; address: 3-2-2 Wakaba, Mihama-ku, Chiba-shi, Chiba, 261-8545, Japan; e-mail: kiyoyasu_tanaka@ide.go.jp

The Institute of Developing Economies (IDE) is a semigovernmental, nonpartisan, nonprofit research institute, founded in 1958. The Institute merged with the Japan External Trade Organization (JETRO) on July 1, 1998. The Institute conducts basic and comprehensive studies on economic and related affairs in all developing countries and regions, including Asia, the Middle East, Africa, Latin America, Oceania, and Eastern Europe.

INSTITUTE OF DEVELOPING ECONOMIES (IDE), JETRO
3-2-2, WAKABA, MIHAMA-KU, CHIBA-SHI
CHIBA 261-8545, JAPAN

Agglomeration Effects of Informal Sector: Evidence from Cambodia*

Kiyoyasu TANAKA† *Institute of Developing Economies*

Yoshihiro HASHIGUCHI‡ *Institute of Developing Economies*

May 2015

Abstract

The presence of a large informal sector in developing economies poses the question of whether informal activity produces agglomeration externalities. This paper uses data on all the nonfarm establishments and enterprises in Cambodia to estimate the impact of informal agglomeration on the regional economic performance of formal and informal firms. We develop a Bayesian approach for a spatial autoregressive model with an endogenous explanatory variable to address endogeneity and spatial dependence. We find a significantly positive effect of informal agglomeration, where informal firms gain more strongly than formal firms. Calculating the spatial marginal effects of increased agglomeration, we demonstrate that more accessible regions are more likely than less accessible regions to benefit strongly from informal agglomeration.

Keywords: Agglomeration, Informal Sector, Cambodia, Bayesian

JEL classification: C11, C21, C26, H26, O17, R12

1 Introduction

A spatial concentration of economic activity has crucial implications for developing economies. Williamson (1965) argues that agglomeration favors economic growth at an early stage of economic development because limited resources such as capital, human capital, and infrastructure can be most efficiently utilized in an agglomerated area. Fujita and Thisse (2003) demonstrate that agglomeration can promote growth in a two-region model of endogenous growth. Indeed, the importance of agglomeration has been emphasized because firms and workers in

*This paper is written under the project "Multinational Firms and the Globalization of Developing Economies" funded by the Institute of Developing Economies. We acknowledge the financial support of JSPS Grant-in-Aid for Young Scientists (B). We would like to thank Fumihiko Nishi and Souknilanh Keola for data assistance. For useful comments, we thank Kentaro Nakajima and seminar participants at the Japanese Economics Association Conference for 2014 in Fukuoka. All remaining errors are our own.

†Corresponding author: Inter-disciplinary Studies Center, Institute of Developing Economies; Address: 3-2-2 Wakaba, Mihama-ku, Chiba-shi, Chiba, 261-8545, Japan; E-mail: kiyoyasu_tanaka@ide.go.jp

‡Development Studies Center, Institute of Developing Economies; Address: 3-2-2 Wakaba, Mihama-ku, Chiba-shi, Chiba 261-8545, Japan. E-mail: yoshihiro_hashiguchi@ide.go.jp

the agglomerated area benefit from agglomeration externalities through more efficient sharing of local suppliers, better matching between employers and workers, and knowledge spillovers among firms and workers (Duranton and Puga, 2004).

However, it is unsettled whether a spatial concentration of industrial activity would produce similar benefits for low income economies, as has been previously demonstrated for agglomeration economies in high and middle income countries (Rosenthal and Strange, 2004; Melo et al., 2009). Developing economies are substantially different from developed economies in that an informal sector plays a large role in the economy (Schneider and Enste, 2000). A large number of business firms do not register formally to the government and they are different from formally registered firms in economic characteristics, such as productivity, profitability, and size (McKenzie and Seynabou Sakho, 2010; Fajnzylber et al, 2011). For instance, Annez and Buckely (2009, p. 15) state that "some critics argue that informality is unproductive and raises the costs to the formal sector, crowding out agglomeration economies." By contrast, Overman and Venables (2005, p. 20) suggest that "the informal sector also contributes to agglomeration economies."

The controversy over agglomeration economies of informal activity is mainly due to a lack of formal econometric evidence. Although some case studies suggest that the informal sector might contribute to agglomeration effects, Duranton (2009, p. 82) admits that "most of the findings concern the formal sector."[1] Also, Overman and Venables (2005, p. 21) state that "formal evidence on this issue is simply unavailable." Therefore, we know little about the empirical magnitude of agglomeration effects of the informal sector in developing economies.

In this paper, we seek to shed light on the role of an informal sector in agglomeration economies using the Economic Census of Cambodia in 2011 (EC2011). This census covers all the nonfarm establishments across all industrial sectors in all areas of Cambodia and asks whether individual establishments are registered with the Ministry of Commerce. Unregistered economic activities are a commonly used definition of informality and business registration can be used as an objective criterion to classify formal and informal economic activities (Schneider, 2005). We exploit this dataset to address the following questions. What is the size and geographic distribution of the informal sector in Cambodia? Does a spatial concentration of informal activity contribute to the economic performance of informal firms? Do formal firms also benefit from agglomeration effects of informal firms?

To identify the impact of informal agglomeration in Cambodia, we estimate a spatial autoregressive model with an endogenous independent variable. An endogeneity problem of agglomeration economies has been addressed for identification since the seminal work of Ciccone and Hall (1996).[2] However, Artis et al. (2012) point out that the estimate of agglomeration economies remarkably changes when spatial autocorrelation of a dependent variable is controlled. Their findings imply that the spatial autocorrelation may magnify or reduce the local impact of agglomeration economies through a spatial multiplier effect across regions. Because regional economic performance in Cambodia can be spatially dependent for information and knowledge spillovers between nearby regions, we must account for both endogeneity and spatial autocorrelation issues to obtain a precise estimate of informal agglomeration effects. In this respect, our empirical model allows for estimating both direct and indirect impacts of informal agglomeration, the latter of which can be produced by significant spatial autocorrelation in

[1] For instance, Livingstone (1991) discusses agglomeration of informal enterprises in Kenya.

[2] For a detailed discussion of the endogeneity problem, refer to Eberts and McMillen (1999), Rosenthal and Strange (2004), Cohen and Paul (2009), and Puga (2010).

economic performance.

To estimate our model, we develop a Bayesian method by extending the Bayesian instrumental variables (IV) method proposed by Rossi et al. (2005). While their method is developed for a non-spatial linear model with an endogenous independent variable, we extend their approach to estimate a spatial autoregressive model with the endogenous independent variable. We also conduct a Monte Carlo simulation to examine the estimation performance of our model when instruments are weak and strong. The simulation results show that the stronger instruments lead to a smaller dispersion of the posterior distributions of structural parameters, suggesting that the strong instruments for the endogenous independent variable are crucial to identify the structural parameters. It should be emphasized that a Bayesian approach to instrumental variables regression has long been developed since the work of Drèze (1976). On the other hand, the Bayesian approach is increasingly used to estimate a spatial autoregressive model (see, e.g., Banrejee et al., 2004; LeSage and Pace, 2009). To the best of our knowledge, our work is the first to apply the Bayesian method for the spatial autoregressive model with the endogenous independent variable.

A Bayesian approach in this paper is advantageous for mitigating identification problems in estimating a spatial autoregressive model with an endogenous independent variable. Prior work such as Artis et al. (2012) employs the feasible generalized spatial two-stage least squares method proposed by Kelejian and Prucha (1998). In this method, higher-order spatial lags of exogenous independent variables are used as instruments for a spatial lag of a dependent variable under the assumption that the higher-order spatial lags have no direct effect on the dependent variable. Gibbons and Overman (2012) argue that such an exclusion restriction may not be empirically valid. Also, a potentially high correlation among the higher-order spatial lags tends to cause a weak instrument problem. By contrast, our approach does not rely on these identification assumptions by deriving a likelihood function for the dependent variable and applying Bayesian estimation. Because it obviates the need to exploit instruments for estimating the spatial lag of the dependent variable, we can focus on endogeneity problems of informal activity to estimate a causal effect of informal agglomeration economies. Finally, we highlight that our empirical framework can be widely applied in empirical research because endogeneity and spatial autocorrelation problems are common in regional data on economic activity of workers, firms, and governments.

An empirical investigation of agglomeration economies in Cambodia is crucial from both academic and policy perspectives. Schneider et al. (2010) estimate that informal activity accounted for 48.7% of GDP on average for the period 1999-2007, suggesting the substantial contribution of the informal economy. In this respect, Cambodia provides an interesting setting for investigating the role of an informal sector in agglomeration economies. Moreover, the Cambodian economy was devastated by the Pol Pot regime in 1975-79 and the subsequent civil war, before the Paris Conference on Cambodia in 1991 led to agreements on a comprehensive political settlement of the Cambodia conflict. As it set a stage for economic reconstruction, economic growth has averaged 6 percent for the last 10 years. However, per capita GDP reached merely 931.2 U.S. dollars in 2012 (IMF, 2012). Because the Cambodian economy is still constrained by its limited resources in capital, human capital, and infrastructure, it is a crucial policy issue for the country's government to examine the extent to which industrial agglomeration should be promoted to maximize economic growth and whether policy targets should be based on the formal and/or informal sectors.

The main results of our analysis can be summarized as follows. First, we find evidence that

a spatial concentration of informal firms produces a positive impact on the regional economic performance of both formal and informal firms in manufacturing and wholesale/retail industries. Statistical tests of identification problems provide supporting evidence that the estimated coefficients of informal agglomeration can be interpreted to reflect a causal relationship. Second, the positive impact of informal agglomeration on performance tends to be larger for informal firms than for formal firms. This result may reflect that informal firms tend to have weaker backward and forward linkages with formal firms than with the other informal firms. Finally, we calculate spatial multiplier effects of an increase in informal agglomeration and find that more accessible regions are more likely than less accessible regions to benefit from informal agglomeration. The spatial impacts depend crucially on the road infrastructure and geography of each region within the country.

The rest of this paper is organized as follows. Section 2 presents an econometric framework including an empirical model and a Bayesian estimation method, with the details of the Bayesian algorithm described in Appendix A. Variable definitions and instrumental variables are also explained in this section. Section 3 describes the data sources. Section 4 describes the characteristics of an informal sector in Cambodia in terms of the size and geographic distribution of informal firms. Section 5 presents the estimation results and spatial marginal impacts of informal agglomeration. Section 6 presents robustness checks, and Section 7 concludes.

2 Econometric Framework

2.1 Empirical Model

Let us consider the following spatial autoregressive model to estimate the impact of agglomeration on regional economic performance:

$$y_{si} = x_{si}\beta_{s0} + \mathbf{z}_i\boldsymbol{\beta}_{s1} + \rho_s \sum\nolimits_{j=1}^{n} w_{ij}y_{sj} + \varepsilon_{si} \tag{1}$$
$$i = 1, 2, \ldots, n$$
$$s \in \{\text{Manufacturing}, \text{Wholesale/Retail}\}$$

where i and s denote commune-level region and industrial sector and ε_{si} is an error term. The variable x_{si} represents the degree of localized industrial agglomeration, and its coefficient indicates the magnitude of agglomeration effects (localization economies).[3] Estimation is performed separately for the Manufacturing and Wholesale/Retail sectors. As will be mentioned in the following section, we use two measures for y_{si}: sales per worker (*sal*) and wage payment per employee (*wge*). These can be interpreted as a proxy for labor productivity. Since agglomeration economies imply that firms are more productive in more agglomerated areas, regressing the labor productivity (*sal*) on the degree of industrial agglomeration is a natural way to measure the magnitude of agglomeration economies (see, e.g., Ciccone and Hall, 1996; Brülhart and Mathys, 2008; Broersma and Oosterhaven, 2009). Meanwhile, higher wages in agglomerated areas can be seen as evidence of agglomeration economies. The reason is that higher wages increase costs for firms and induce them to relocate elsewhere. However, the presence

[3]Industrial agglomeration is classified into two types: localized and urbanized agglomeration. In line with previous studies, we define localized industrial agglomeration as the economic agglomeration of the same industry and region, and urbanization as the agglomeration of urban population.

of local productivity advantages from agglomeration offsets high labor costs, making the firms stay in high-wage regions. Thus, the wage premium in the agglomerated areas can be also interpreted as evidence of agglomeration economies (see, e.g., Glaeser and Maré, 2001; Wheaton and Lewis, 2002; Combes, et al., 2008).

We consider that $\mathbf{z}_i$ is a vector of exogenous (or predetermined) variables controlling for a variety of commune characteristics. The variable w_{ij} indicates the geographical relationship between regions i and j, which is specified as

$$w_{ij} = \begin{cases} 0 & i = j \\ d_{ij}^{-1} / \sum_j^n d_{ij}^{-1} & i \neq j \end{cases} \tag{2}$$

where d_{ij} is the traveling time between i and j. The parameter ρ_s indicates the magnitude of spatial autocorrelation in y_{si}.

We seek to identify a causal impact of informal agglomeration x_{si} on regional economic performance y_{si} after controlling for a variety of commune characteristics $\mathbf{z}_i$. However, some communes may attract workers in an informal sector by higher wages resulting from unobserved natural advantages such local climate, social infrastructure, and natural resources. Because these natural advantages affects regional economic performance, an omitted-variable bias may arise in the estimated coefficient β_{s0}. Additionally, more productive informal firms may self-select to locate their economic activity in agglomerated regions for positive externality, which introduces an upward bias in the estimated coefficient β_{s0} for reverse causality. In these cases, it is not sensible to assume that the density of informal economic activity is uncorrelated with true differences in the determinants of regional economic performance that are not explicitly controlled in our model.

To deal with such endogenous problems in x_{si}, we employ the Bayesian instrumental variables method proposed by Rossi et al. (2005). We consider that x_{si} is an endogenous variable that is linearly related to a set of instruments $(\mathbf{q}_{si}, \mathbf{z}_i)$ and an idiosyncratic shock η_{si}, where $\mathbf{q}_{si}$ is a vector of variables related to x_{si} but independent of the error terms ε_{si} and η_{si}. Following Rossi et al. (2005), we specify the system of equations as follows:

$$x_{si} = \mathbf{q}_{si}\boldsymbol{\gamma}_{s0} + \mathbf{z}_i\boldsymbol{\gamma}_{s1} + \eta_{si} \tag{3}$$

$$y_{si} = x_{si}\beta_{s0} + \mathbf{z}_i\boldsymbol{\beta}_{s1} + \rho_s \sum\nolimits_{j=1}^{n} w_{ij}y_{sj} + \varepsilon_{si}\,. \tag{4}$$

This system consists of a structural equation (4) with an endogenous independent variable and multiple instruments. If the correlation between η_i and ε_i is positive, there will be a positive endogeneity bias in the estimated coefficient of x_{si} (Rossi et al., 2005). In vector and matrix notation, these equations can be written as

$$\mathbf{x}_s = \mathbf{Q}_s\boldsymbol{\gamma}_{s0} + \mathbf{Z}\boldsymbol{\gamma}_{s1} + \boldsymbol{\eta}_s \tag{5}$$

$$\mathbf{S}_s\mathbf{y}_s = \mathbf{x}_s\beta_{s0} + \mathbf{Z}\boldsymbol{\beta}_{s1} + \boldsymbol{\varepsilon}_s\,, \tag{6}$$

where $\mathbf{S}_s = \mathbf{I}_n - \rho_s\mathbf{W}$, $\mathbf{x}_s = (x_{s1}, \ldots, x_{sn})'$, $\mathbf{y}_s = (y_{s1}, \ldots, y_{sn})'$, $\boldsymbol{\eta}_s = (\eta_{s1}, \ldots, \eta_{sn})'$, $\boldsymbol{\varepsilon}_s = (\varepsilon_{s1}, \ldots, \varepsilon_{sn})'$, $\mathbf{Q}_s = (\mathbf{q}_{s1}, \ldots, \mathbf{q}_{sn})'$, and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$. We use this system to estimate the impact of agglomeration on regional economic performance for each industrial sector.

Before proceeding to describe Bayesian estimation, it is useful to highlight the advantage of our approach over the prior approach. In a spatial autoregressive model, Kelejian and Prucha (1998) propose higher order spatial lags of exogenous independent variables as instruments

for the spatial lag of a dependent variable. This approach is likely to suffer from estimation problems for identification (Gibbons and Overman, 2012). Specifically, it is assumed that the higher order spatial lags do not affect the dependent variable directly. If this exclusion restriction is not valid, the instruments are not appropriate. Moreover, these instruments are likely to be weak for a potentially high correlation among the higher order spatial lags; this leads to a biased estimate for the spatial lag variable. In contrast, our approach does not rely on these identification assumptions to estimate the spatial autoregressive model in equations (5) and (6). Controlling for potential spatial autocorrelation in economic performance across regions, we can focus on the endogeneity of x_{si} to identify the coefficient of agglomeration, β_{s0}.

2.2 Bayesian Estimation

Bayesian methodology requires a *posterior density* in order to draw an inference regarding unknown parameters in a model. The posterior is proportional to the *likelihood function* times the *prior density*: $\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})$, where $\mathbf{y}$ represents the observed data, $\boldsymbol{\theta}$ represents the unknown parameters, $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ is the posterior, and $f(\mathbf{y} \mid \boldsymbol{\theta})$ is the likelihood. The following subsections explain the likelihood and the prior for our model, and show the computational scheme for estimating the posterior. It should be noted that sector subscripts s are dropped for ease of notation in the following sections.

2.2.1 Likelihood and priors

To derive the likelihood function, we assume that $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ have a multivariate normal distribution:

$$\begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim MVN(\mathbf{0},\ \boldsymbol{\Sigma} \otimes \mathbf{I}_n),$$

where $\boldsymbol{\Sigma}$ is a 2×2 covariance matrix. Letting $\tilde{\mathbf{y}} = (\mathbf{x}, \mathbf{y})'$ and letting $\mathbf{u}$ denote a $(2n \times 1)$ vector that follows a multivariate standard normal distribution $N(\mathbf{0},\ \mathbf{I}_{2n})$, Equations (5) and (6) can then be rewritten as

$$\mathbf{u} = (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)^{-\frac{1}{2}} \left[\begin{pmatrix} \mathbf{I}_n & \mathbf{0} \\ -\beta_0 \mathbf{I}_n & \mathbf{S} \end{pmatrix} \tilde{\mathbf{y}} - \begin{pmatrix} \mathbf{Q} \\ \mathbf{0} \end{pmatrix} \boldsymbol{\gamma}_0 - \begin{pmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\beta}_1 \end{pmatrix} \right]. \tag{7}$$

The Jacobian for the transformation of $\mathbf{u}$ into $\tilde{\mathbf{y}}$ is

$$\begin{aligned} J = \left| \frac{\partial \mathbf{u}}{\partial \tilde{\mathbf{y}}} \right| &= \left| (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)^{-\frac{1}{2}} \begin{pmatrix} \mathbf{I}_n & \mathbf{0} \\ -\beta_0 \mathbf{I}_n & \mathbf{S} \end{pmatrix} \right| \\ &= |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \, |\mathbf{S}| \left| \mathbf{I}_n - \mathbf{0}\, \mathbf{S}^{-1}(-\beta_0)\mathbf{I}_n \right| \\ &= |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \, |\mathbf{S}| \ . \end{aligned} \tag{8}$$

Then the likelihood function can be obtained:

$$L = (2\pi)^{\frac{n}{2}} \ |\boldsymbol{\Sigma}|^{\frac{n}{2}} \ |\mathbf{S}| \exp \left\{ \begin{bmatrix} \mathbf{x} - \mathbf{Q}\boldsymbol{\gamma}_0 - \mathbf{Z}\boldsymbol{\gamma}_1 \\ \mathbf{S}\mathbf{y} - \mathbf{x}\beta_0 - \mathbf{Z}\boldsymbol{\beta}_1 \end{bmatrix}' [\boldsymbol{\Sigma} \otimes \mathbf{I}_n]^{-1} \begin{bmatrix} \mathbf{x} - \mathbf{Q}\boldsymbol{\gamma}_0 - \mathbf{Z}\boldsymbol{\gamma}_1 \\ \mathbf{S}\mathbf{y} - \mathbf{x}\beta_0 - \mathbf{Z}\boldsymbol{\beta}_1 \end{bmatrix} \right\}, \tag{9}$$

$$\mathbf{S} = \mathbf{I}_n - \rho \mathbf{W}. \tag{10}$$

Independent priors for the unknown parameters are specified as

$$\begin{aligned} &\boldsymbol{\beta}^* \equiv \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} \sim MVN(\mathbf{b}_\beta, \mathbf{B}_\beta), \quad \boldsymbol{\gamma}^* \equiv \begin{pmatrix} \boldsymbol{\gamma}_0 \\ \boldsymbol{\gamma}_1 \end{pmatrix} \sim MVN(\mathbf{b}_\gamma, \mathbf{B}_\gamma), \\ &\rho \sim U(\lambda_{\min}^{-1}, \lambda_{\max}^{-1}), \quad \boldsymbol{\Sigma} \sim IW(b_\Sigma, \mathbf{B}_\Sigma), \end{aligned} \tag{11}$$

where $IW(\,)$ and $U(\,)$ denote an inverted Wishart distribution and a uniform distribution, respectively. The prior parameters are $\mathbf{b}_\beta$, $\mathbf{B}_\beta$, $\mathbf{b}_\gamma$, $\mathbf{B}_\gamma$, $\lambda_{\min}$, $\lambda_{\max}$, b_Σ, and $\mathbf{B}_\Sigma$. The parameters $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum real eigenvalues of $\mathbf{W}$, respectively. We use these values to put a limit on the parameter space of ρ: $\rho \in (\lambda_{\min}^{-1}, \lambda_{\max}^{-1})$. If a vector of the eigenvalues of $\mathbf{W}$ contains only real values, this restriction ensures $|\mathbf{S}| > 0$.[4] The values of the other prior parameters are assumed as $\mathbf{b}_\beta = \mathbf{0}$, $\mathbf{B}_\beta = 100\mathbf{I}_k$, $\mathbf{b}_\gamma = \mathbf{0}$, $\mathbf{B}_\gamma = 100\mathbf{I}_l$, $b_\Sigma = 2$, and $\mathbf{B}_\Sigma = 2\mathbf{I}_2$, where k an l denote the dimensions of $\mathbf{b}_\beta$ and $\mathbf{b}_\gamma$, respectively. Specifying prior parameters is difficult when no information is available for unknown parameters. In line with standard practice, we choose zero values for the location parameters of the prior distributions for the coefficients. This choice means that our prior beliefs for the coefficients are centered on zero. Based on these zero values, we can then investigate the extent to which the posterior distribution moves away from the prior. We set these priors to have large variances in order to ensure that our prior beliefs for the unknown parameters are non-informative.

2.2.2 MCMC algorithm

Having clarified the likelihood and the prior for our model, we now explain the posterior inference procedure. The posterior inference can be carried out by the Markov Chain Monte Carlo (MCMC) method, which allows us to generate samples from the posteriors and to draw a statistical inference using the simulated samples. Bayesian inference is based on the posterior distributions of unknown parameters.

The MCMC sampling requires us to draw samples from the *full* conditional posterior distributions, such as

$$\begin{aligned} \boldsymbol{\beta}^* \mid \boldsymbol{\gamma}^*, \rho, \boldsymbol{\Sigma}, \text{Data} &\sim \textit{Normal Distribution} \\ \boldsymbol{\gamma}^* \mid \boldsymbol{\beta}^*, \rho, \boldsymbol{\Sigma}, \text{Data} &\sim \textit{Normal Distribution} \\ \boldsymbol{\Sigma} \mid \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \rho, \text{Data} &\sim \textit{Inverted Wishart Distribution} \\ \rho \mid \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\Sigma}, \text{Data} &\sim \textit{Unknown Distribution} \end{aligned}$$

where Data $= \{\mathbf{x}, \mathbf{y}, \mathbf{Q}, \mathbf{W}, \mathbf{Z}\}$. Using Equations (9)–(11), these full conditional distributions can be derived. The derivation and the MCMC sampling algorithm are described in Appendix A.

2.3 Variable Definitions

We turn to the definition of variables used in the spatial autoregressive model. For a dependent variable, we use two measures of regional economic performance for formal and informal firms in commune i and industrial sector s: average sales per worker (sal_{si}) and average wage

[4] The inverted minimum and maximum eigenvalues of the spatial weight matrix $\mathbf{W}$ in our sample are $\lambda_{\min}^{-1} = -1.054$ and $\lambda_{\max}^{-1} = 1.000$.

payments per employee (wge_{si}). These variables are defined as follows:

$$sal_{si} = \log\left(Sales_{si}/TW_{si} + 1\right)$$
$$wge_{si} = \log\left(Wage_{si}/TE_{si} + 1\right),$$

where TW_{si} and TE_{si} denote the total numbers of workers and employees, respectively, for commune i and industry s. The workers include self-employed proprietors, unpaid family workers, and regular employees. In measuring wage levels, we use the total number of employees by excluding self-employed proprietors and unpaid family workers from the total number of workers. This definition is crucial for informal firms, which are often operated using family workers only. While prior works have also used alternative performance indicators such as total factor productivity, we do not employ them because a majority of establishments in our dataset do not report a component of total expenses, making it difficult to calculate value added with precision. In fact, it is generally difficult to measure the precise amount of expenses used for business activity in developing economies such as Cambodia, as was demonstrated by de Mel et al. (2009) in the case of Sri Lanka.[5] Also, it is challenging to calculate a reliable measure of profitability, especially for informal small-scale establishments. Therefore, we focus on the relatively reliable economic measures such as sales and wages.

For independent variables, we use the following variables. To estimate localization economies of an informal sector, we define the industrial concentration for commune i and industry s as:

$$agg_{si}^{inf} = \log\left(\frac{Employment_{si} + 1}{Area_i}\right).$$

Employment is defined as the total number of workers only by the informal firms in the same industry and same commune, and *Area* is the geographic area of a commune. Thus, our measure of industrial agglomeration reflects the density of local employment in the same industry. In line with previous studies, we define localized industrial agglomeration as the economic agglomeration of the same industry and commune, and urbanization as the agglomeration of urban population. We account for the effect of urbanization in each commune by total population:

$$pop_i = \log\left(Population_i + 1\right).$$

We control for the role of industrial infrastructure in regional economic performance with electricity access:

$$elec_i = \frac{Household\ with\ electricity\ access_i}{Total\ households_i}.$$

Households have electricity access if they have access to city power or generator for a main source of light. Additionally, we take into account the local labor market by defining a share of high and low skilled workers:

$$hskil_i = \frac{Population\ with\ high\ education_i}{Total\ population_i},$$
$$lskil_i = \frac{Population\ with\ low\ education_i}{Total\ population_i}.$$

[5]For instance, business goods and materials may be used for home consumption, but are not recorded as business expenses.

High education includes persons completing technical/vocational diplomas and undergraduate or graduate degrees. Low education indicates persons completing only primary and secondary school programs. These variables help us to control for potential sorting of heterogeneous workers by quality (Combes et al, 2008).

Each commune has some natural advantages for industrial location, which may confound the agglomeration effect of an informal sector. To account for local advantages, we consider the following variables. Cropland area is included to capture the effect of agricultural industry on economic activity in nonfarm industries:

$$crplnd_i = \log\left(Cropland\ area_i + 1\right).$$

Border regions may have an accessibility advantage for exporting to neighboring economies through a land border and to international markets through marine ports. We include a border variable ($border_i$) to explain these border effects. Additionally, we include the presence of major international airports (air_i) and special economic zones (sez_i) in each commune. Each of these variables is defined as a dummy variable.

2.4 Instrumental Variables

To address the endogeneity of industrial agglomeration of an informal sector, we exploit three instrumental variables:

$$emp_{si}^{1998} = \log\left(\frac{Employment_{si}^{1998} + 1}{Area_i}\right),$$
$$green_i^{2002} = \log\left(Forest\ area_i^{2002} + 1\right),$$

and

$$hskil_i^{1998} = \frac{Population\ with\ high\ education_i^{1998}}{Total\ population_i^{1998}}.$$

The variable emp_{si}^{1998} is the employment density in a commune and industry for the year 1998, and $green_i^{2002}$ is the geographic area of forest and shrublands in each commune for the year 2002. The variable $hskil_i^{1998}$ represents the share of the population completing higher education for the year 1998.

The identifying assumption is that the past level of employment density, presence of high skilled workers, and geographic characteristics have persistent influences only on the preferences of workers about the location in which they seek employment opportunities in an informal sector. However, these variables are not correlated with the current differences in regional economic performance that are not accounted for by our model. The choice to use the past data for eligible instrumental variables is similar to the previous empirical approach of Ciccone and Hall (1996). The geographic characteristics, such as land area, were also used as an instrument in Ciccone (2002).

Our justification for our choice of instruments is as follows. First, industrial agglomeration is a result of a cumulative process in which individual economic activity is attracted to specific points in geographic areas over time. As is predicted by the economic geography model (Fujita et al., 1999), a large concentration of economic activity in one region is more likely than

a small concentration in another region to attract a larger number of workers because of the large market size and wide availability of intermediate inputs and consumer products. As it is reasonable to consider that these persistent effects exist for informal activity, the past density of labor should affect the formation of informal agglomeration in a corresponding industry and region. Additionally, we hypothesize that geographic characteristics affect the patterns of locations where workers seek jobs. Forests and shrublands are more likely than plains to deter the formation of industrial activity as a result of the cost of leveling the ground, thereby yielding a geographic barrier for worker settlement. Thus, the geographic characteristics in the past also affect the informal agglomeration in a region. As is consistent with our explanations, Combes et al. (2010) emphasize the usefulness of history and geology as instruments of agglomeration.

We further exploit the geographic variation in the share of skilled workers in the past period as an instrument. Using data in developed economies, prior studies have shown that more skilled workers tend to concentrate in denser areas, such as large cities, which accounts for a spatial disparity in wages (Glaeser and Maré, 2001; Combes et al., 2008). Andersson et al. (2007) provide the evidence that more productive workers are matched with more productive firms in a denser region through assortative matching and production complementarity. These findings support a strong linkage between skilled workers and industrial agglomeration processes, suggesting that the presence of more productive workers today may attract the entry of new firms and settlement of other workers tomorrow. Although the evidence has been obtained mainly from formal sectors, there is no a priori reason for skilled workers not to congregate in regions with denser informal activity. Thus, we exploit a past skill composition in the local market as a third instrument.

Exclusion restrictions

We now turn to the exclusion restrictions of the instruments used. Specifically, we assume that labor density, geographic characteristics, and skill composition in the past affect current regional economic performance only through the current level of informal agglomeration. However, they should not affect the regional performance through other channels that are not explicitly accounted for by the control variables including population size, electricity access, high and low skilled labor, crop land, and other infrastructures in the current period. Possibly, our instruments may produce unobserved persistent effects on a local market over time, and produce an impact on contemporaneous determinants of regional economic performance. As long as such persistent influences are picked up at least partly by any of the control variables, the exclusion restrictions are satisfied. On the other hand, any remaining correlation between instruments and unobserved current shocks violates the exclusion restrictions, making it difficult to give a causal interpretation for an estimate of informal agglomeration economies.

The validity of exclusion restrictions depends crucially on the hypothesis that unobserved current shocks to regional performance are sufficiently uncorrelated with the labor density, geographic characteristics, and skill endowments in the past. We believe that there is no strong reason to believe that the exclusion restrictions are invalidated in the context of the Cambodian economy.[6] The economy has experienced a remarkably rapid pace of economic growth since the early 1990s, and its industrial structure has been substantially transformed from agricul-

[6]While the lag length of our instruments is much shorter than the duration of historical population used in Combes et al. (2010), we must note that longer lags are likely to cause a weak instrument problem as discussed next.

ture to manufacturing and services; for instance, the share of agriculture in GDP declined from 55.6% in 1990 to 33.8% in 2010, but the share of manufacturing increased from 5.2% to 14.9% during this period (Hill and Menon, 2013). Rapid structural changes are likely to isolate our instruments based on past data from unobserved current shocks to economic performance in industrial sectors.

Checking instrument validity

Up to this point in this section, we have provided theoretical explanations to support the validity of our instruments. However, the credibility of instruments also rests on the empirical ability of our model to satisfy the validity conditions for instruments (Murray, 2006). One condition is a sufficient correlation between an endogenous variable and the instruments, whereas the other is an absence of a correlation between an error term ε_{si} and the instruments. Because we adopt a Bayesian approach for a spatial autoregressive model with an endogenous independent variable, we briefly discuss our statistical approach to an examination of the validity of our instruments.

An identification problem arises when instruments are weakly correlated with an endogenous independent variable (Rossi et al., 2005). According to our Monte Carlo simulation in Appendix B, the use of weaker instruments yields more dispersed posterior distributions for the coefficients, indicating the difficulty of identification under weak instruments. The weakness of instruments can be confirmed by checking the statistical significance of $\boldsymbol{\gamma}_{s0}$ in Equation (5). Additionally, another issue is the validity of exogeneity of instruments. Standard (non-Bayesian) econometrics provides specification tests to check exogeneity. Following the prior literature, we calculate a Sargan's statistic using the posterior mean values of the coefficients, and conduct specification tests for our model using a χ^2 distribution.

3 Data Description

3.1 Economic Census of Cambodia

Our main dataset is based on the *Economic Census* of Cambodia in 2011 (EC2011). The census was conducted in March 2011 to survey economic activities of all the nonfarm establishments and enterprises over the entire territory of Cambodia. The EC2011 was mainly funded by the Japanese ODA and implemented by the National Institute of Statistics, the Cambodian Ministry of Planning, in cooperation with the Japanese government.[7] The survey aimed to collect information about a firm's activities, including financial statement, the persons engaged in the business, and main line of business. The administrative geographic units consist of 1,621 communes in 24 provinces including the Municipality of Phnom Penh. Establishment-level data are aggregated across communes to construct regional data. In the analysis, we use establishment and firm interchangeably.

A regulatory framework for commercial enterprises was first established by the "Law Bearing upon Commercial Regulations and the Commercial Register," which was enacted in 1995 and modified in 1999. This law defines the meaning of commercial enterprise and commercial activity, stipulates the obligation of companies to register, and details the formal procedures of commercial registration. Moreover, the National Assembly in Cambodia adopted the "Law on

[7] Details of EC2011 can be found at the Japanese government's website:
`http://www.stat.go.jp/english/info/meetings/cambodia/census11.htm`

Commercial Enterprise" in 2005, which is applied to partnerships, private limited companies, public limited companies, and foreign businesses. A partnership or company must register with the registrar the specific location of the office and the name of the agent.

In the questionnaire of EC2011, each establishment is asked about its registration to administrative agencies and the names of the ministries licensing or approving its operation. Specifically, each establishment must answer whether or not they have registered with the Ministry of Commerce or the Provincial Department of Commerce. We exploit this question to define the formal sector as the business activities of registered firms and the informal sector as those of unregistered firms.[8] In Cambodia, this definition implies that the formal firms have followed several procedures for registration: (1) to deposit the legally required initial capital in a bank and obtain deposit evidence, (2) conduct an initial check for uniqueness of the company name at the Intellectual Property Department and the Business Registration Office, and (3) publish an abstract of the company organization documents and incorporate the company with the Business Registration Department in the Ministry of Commerce (World Bank, 2014). These procedures are estimated to cost at least 400 USD and to take one month.

3.2 Other Data Sources

Data on population and households are taken from the *Population Census* of Cambodia in 1998 and 2008. These datasets provide the characteristics of the population and households, including residential location, education, and living environments. Data on special economic zones and major airports are taken from the Cambodia Investment Guidebook in 2010. We also use this data source to define the border regions. Additionally, geographic data on cropland, forest, and shrublands are taken from the satellite images. Specifically, the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's spacecraft Terra and Aqua produces image of land use with global coverage and high spatial and temporal resolution. MODIS land products are received, distributed, and archived at the Land Processes Distributed Active Archive Center (LP DAAC), a component of NASA's Earth Observing System (EOS) Data and Information System (EOSDIS).

The spatial-weighting matrix is based on the shortest traveling time between commune pairs, which consists of a 1,621×1,621 matrix with each element estimated in hours. We employ the Floyd-Warshall algorithm to compute the shortest time between communes using data on traveling distance and traveling speed.[9] Data for this variable come from the Geographic Information System (GIS) shape file on 1,621 administrative units at the commune level in Cambodia. We obtain data on the latter variable from the map of the Cambodian road network published by the Cambodian Ministry of Public Works and Transport, which documents the location of national roads at the one- and two-digit level. Additionally, we use the JETRO survey on ASEAN logistics network map (JETRO, 2009) and satellite-image data of geographic conditions in each commune. These datasets justify our assumptions about travelling speeds between a given pair of neighboring communes. Our approach to measuring connectivity is similar to the empirical investigation of agglomeration economies in Indian formal manufacturing industries by Lall et al. (2004).

[8] See Williams and Lansky (2013) for discussions on the definition and measurement of informal employment.
[9] See Appendix C for more details.

4 Characteristics of the Informal Sector

4.1 Size of the Informal Sector

Table 1 presents aggregate figures of formal and informal activity in Cambodia, where financial figures are measured for one month of February 2011. Over all industries, there were 17,378 formal establishments and 487,756 informal establishments.[10] In terms of numbers, 96.6% of establishments can be classified as informal. This figure about informality is larger than those from surveys in other economies as reported by de Paula and Scheinkman (2011) and Rand and Torm (2012). Despite slightly different definitions of informality, our data suggest that the sample surveys are likely to miss a large number of informal firms.

[– Table 1 –]

Turning to financial aspects, we find that informal firms accounted for 76.6% of total sales and formal firms were responsible for 23.4% of them. We obtain similar percentage shares when measuring with the expenses that include purchases of products, costs for providing services, rents and employees' wages. Additionally, informal firms were responsible for 59.2% of total wages and formal firms accounted for 40.8% of them. Although the total wage payments point to the dominant role of informal activity, the importance of the informal sector declines slightly. Finally, 66.4% of all workers worked for informal firms and 33.6% of them worked for formal firms. These figures indicate that a majority of industrial workers belong to the informal sector.

Table 1 also presents the figures for manufacturing and wholesale/retail industries. In the following analysis, we focus on these industries for two major reasons. First, these industries play a substantial role in the industrial activity of Cambodia. Wholesale and retail industries accounted for 33.1% of total employment in nonfarm industries, and manufacturing industries constituted 31.7%. Thus, the analysis of these industries is critical to understanding agglomeration effects in Cambodia. Second, it has been demonstrated that manufacturing and service industries tend to exhibit different patterns of industrial concentration and agglomeration economies. For example, Kolko (2007) finds that manufacturing industries tend to be more agglomerated than service industries in the U.S. Moreover, Graham (2009) shows that the elasticity of localization economies tends to differ in magnitude across manufacturing and service industries.[11] Following the previous literature, we shed light on possible differences between manufacturing and service industries. Compared to the prior papers, our study is distinctive in that we focus on *informal* activity in these industries.

In the manufacturing industry, there were 69,693 informal establishments, which represent 97.6% of all manufacturing establishments. As 60.8% of total manufacturing sales came from informal firms, the dominant role of informal activity in manufacturing is apparent. However, formal firms accounted for 64.2% of total wage payments and 67.5% of total employment, implying that the formal sector provides large employment opportunities for local workers. Additionally, there were 286,101 informal establishments in wholesale/retail industries, which accounted for 97.9% of all establishments. Informal activity constituted 87.8% of all sales. By

[10] We exclude NGOs from the data.

[11] Based on a meta-analysis, Melo et al. (2009) show that service industries are likely to exhibit larger urban agglomeration. Dekle (2002) and Morikawa (2011) present evidence of positive agglomeration economies for service industries in Japan.

contrast with the manufacturing industry, the informal firms played a larger role in generating total wage payments and total employment.

In sum, our data suggest that the informal sector plays a substantial role in the industrial activity of Cambodia. The prominence of informality is also observed in individual industries, such as manufacturing and wholesale/retail. However, the size of informality depends on specific industries and measurement. We find that formal firms accounted for the larger shares of total wages and employment in manufacturing than in wholesale/retail. Also, the share of the informal sector appears to be remarkably large in terms of the number of establishments. Although the importance of informality tends to decline in terms of sales and wages, these patterns imply potentially different impacts of informal agglomeration across industries.

4.2 Geographic Distribution of the Informal Sector

We now turn to an examination of the geographic distribution of informal activity in manufacturing and wholesale/retail industries. We first aggregate employment of informal firms across provinces for these industries. Provincial shares of informal employment in percentage terms are shown for manufacturing industry in Figure 1 and for wholesale/retail industry in Figure 2. Note that Phnom Penh is the capital of Cambodia, which shares national borders with three countries: the western border with Thailand from Koh Kong to Preah Vihear, the northern border with Laos from Preah Vihear to Ratanak Kiri, and the eastern border with Vietnam from Ratanak Kiri to Kampot. The southern provinces, Koh Kong, Preah Sihanouk, Kep and Kampot, face the Gulf of Thailand.

[– Figures 1 and 2 –]

Figure 1 shows that the regional share of informal employment is relatively high in the central provinces such as Phnom Penh, Kandal, Kampong Cham, and Takeo. On the other hand, it is relatively small in the peripheral provinces such as Stung Treng, Ratanak Kiri, and Mondul Kiri. Also, the southern provinces such as Koh Kong and Preah Sihanouk have a lower share of informal employment. These observations imply that informal activity in manufacturing tends to concentrate toward the central area of Cambodia. Looking at wholesale and retail industries in Figure 2, we see a similar geographic distribution of informal employment. A noticeable difference is that the spatial concentration of informal activity appears to be more pronounced in Phnom Penh.

To further analyze a spatial distribution of informal activity, we compute an index of agglomeration for these industries.[12] Specifically, we follow the standard approach proposed by Ellison and Glaeser (1997). For region i, a simple measure of the geographic concentration of informal firms in industry s is defined as:

$$G_s \equiv \sum_{i=1}^{n} (g_{is} - z_i)^2 ,$$

where g_{is} is the share of informal employment by firms belonging to industry s in region i, and z_i is the share of aggregate employment by all firms in region i. As is pointed out by Ellison and Glaeser (1997), this measure does not take into account the size distribution of firms in the

[12] Hatsukano et al. (2012) used the nation-wide establishment listing in 2009 to compute several indexes of geographic concentration at the district level for 22 manufacturing industries.

industry and the level of geographic aggregation of regions. An alternative measure is the EG agglomeration index:

$$\gamma_s \equiv \frac{G_s - \left(1 - \sum_r z_i^2\right) H_s}{\left(1 - \sum_i z_i^2\right)(1 - H_s)},$$

where H_s is the firm-level Herfindahl index in industry s, which is computed using only informal firms in the corresponding industry. This index is comparable across industries and robust to the fineness of spatial aggregation of regions, with the value of zero indicating a no-agglomeration benchmark.[13]

Table 2 presents the results of these indexes computed for informal firms in manufacturing and wholesale/retail industries. At the commune-level, we find $G = 0.003$ and $\gamma = 0.002$ for manufacturing. The corresponding figures for wholesale and retail industries are fairly similar. Table 2 also shows the results of the indexes computed at the district- and province-level. Comparing these estimates, we find more concentration of informal activity at the more aggregated level of regions for both manufacturing and wholesale/retail industries. As a reference point, Ellison and Glaeser (1997) show that less concentrated industries such as newspapers, bottled soft drinks, and concrete products in the U.S. take on a value of γ between 0.002 and 0.012. On the other hand, more concentrated industries such as automobiles, automobile parts, and photographic equipment give a value between 0.089 and 0.174. As compared with the results in the U.S., it appears reasonable to conclude that the geographic concentration is not strong for either manufacturing or wholesale/retail industries in Cambodia, at least at the commune- and district-level.

[– Table 2 –]

5 Estimation Results

5.1 Benchmark Results

The summary statistics of the sample used are presented in Table 3. As we estimate separate empirical models for manufacturing and wholesale/retail industries, we show the summary statistics for these industries separately. Table 4 reports the estimation results using sal_{si}^{fml} and sal_{si}^{inf} as an dependent variable.[14] In the manufacturing industry, the posterior mean coefficients of agg_{si}^{inf} are 0.089 for the formal sector and 0.135 for the informal sector. As the 99% credible interval of these posterior distributions does not contain zero, these coefficients are statistically significant at the 1% level. In the case of wholesale/retail industries, the posterior mean coefficients of agg_{si}^{inf} are 0.136 for the formal sector and 0.153 for the informal sector. Both of these coefficients are also statistically significant at the 1% level. Therefore, these results indicate that the density of informal activity has a significantly positive effect on sales per worker for both

[13] Although Cassey and Smith (2014) suggest simulating confidence intervals for the EG index, it is beyond the scope of this paper.

[14] The estimation is implemented with *Ox* version 6.20 (Doornik, 2006).

formal and informal firms across different industries in Cambodia.

[– Tables 3 and 4 –]

We now address the validity of our specification in Table 4. First, the posterior mean of ρ for manufacturing is significantly positive in both formal and informal sectors. This result points to the spatial interdependence of economic performance in manufacturing activity, regardless of formality in commercial registration. Although the posterior mean of ρ for wholesale/retail industries is not significant in either the formal or informal sectors, it is crucial to control for the spatial lag of a dependent variable in estimating precise effects of informal agglomeration economies. Second, the posterior mean coefficients of emp_{si}^{1998} are significantly positive across specifications whereas those of $green_i^{2002}$ are significantly negative. These results imply that the past level of employment density and geographic characteristics should significantly affect the spatial pattern of informal agglomeration. Although the posterior mean coefficients of $hskil_i^{1998}$ are not significant across specifications, we conclude that our instrumental variables are not likely to suffer seriously from a weak instrument problem.

Finally, we further examine the validity of our instruments by calculating a Sargan statistic and conducting a statistical test of exclusion restrictions. Based on a χ^2 distribution, the p-values of the Sargan statistic are fairly large across specifications, implying that there is no strong evidence to reject the null hypothesis that our instrumental variables do not correlate with an error term in Equation (6). Taken together, these statistical tests lend support for the empirical specification of our spatial autoregressive model and for the identification assumption of our instruments. While we find a significantly positive correlation between informal agglomeration and regional economic performance in the formal and informal sectors, these results allow us to conclude that the density of informal activity has a causal impact on sal_{si}^{fml} and sal_{si}^{inf} in both manufacturing and wholesale/retail industries.

Comparing the posterior mean coefficients of agg_{si}^{inf} across specifications, we find two distinctive patterns. First, the density of informal activity has a larger positive impact on informal firms than on formal firms. A plausible explanation is that informal firms tend to have weaker backward and forward linkages with formal firms than with the other informal firms for various reasons. Informal firms do not obtain formal commercial registration, thereby lacking a tax registration code from the government. As formal firms may not be able to report transactions with informal firms for tax deductions, these transactions should increase the tax burden for the formal firms. In addition, informal firms are often incapable of meeting the quality standards for products requested by formal firms and their low productivity is a barrier to linkage with formal firms. Thus, we can interpret that informal agglomeration has a weaker externality for formal firms partly because they are reluctant to engage in economic transactions with informal firms. Indeed, this explanation is consistent with those in the literature. Arimah (2001) finds weaker linkages between formal and informal sectors in Nigeria. Also, Mukin (2013) shows that formal and informal manufacturing firms in India do not agglomerate strongly for low linkages.

Second, we find that the density of informal activity tends to have a stronger impact on wholesale/retail industries than on manufacturing industry. Such a difference is observed for economic performance in both formal and informal sectors. These results are consistent with the meta-analysis of prior estimates in Melo et al. (2009); the average of agglomeration effects is greater for service industry than for manufacturing industry. While the previous findings are based primarily on the formal sector, our analysis further shows that the stronger effects of

agglomeration in the service industry are also observed for a spatial concentration of informal activity.

Looking at the other independent variables, we find that the posterior mean coefficients of pop_i and $elec_i$ are significantly positive across specifications. Consistent with our intuition, urbanization and electricity infrastructure contribute to an increase in regional economic performance in both formal and informal sectors. On the other hand, the posterior mean coefficients of $hskil_i$ have a positive correlation with sal_{si}^{fml}, but a negative association with sal_{si}^{inf} across industries. In wholesale/retail industries, these coefficients are significant for both sal_{si}^{fml} and sal_{si}^{inf}. These results suggest that informal firms tend to have lower sales per worker in communes endowed with more skilled labor, but formal firms in wholesale/retail industries are likely to exhibit larger sales per worker in such communes. Finally, we find that cropland area is significantly and positively associated with sal_{si}^{inf} in manufacturing.

Before turning to discuss the marginal effects of agg_{si}^{inf} in Table 4, we present in Table 5 the estimation results using wge_{si}^{fml} and wge_{si}^{inf} as a dependent variable. We find that the posterior mean coefficients of agg_{si}^{inf} are significantly positive in manufacturing, implying that wages per employee are significantly higher in communes with more dense informal activity. As statistical tests of instruments, such as the Sargan statistic support the validity of the empirical specification, we can interpret these results as indicating a causal relationship. In the case of wholesale/retail industries, we find that the posterior mean coefficients of agg_{si}^{inf} are not significant for wge_{si}^{fml} and significantly positive only for wge_{si}^{inf}. As the validity of instruments is supported by the statistical tests, we conclude that the density of informal activity in wholesale/retail industries has a causal positive impact on wages per employee in informal firms. Consistent with the results in Table 4, we find that wage premiums from informal agglomeration economies tend to be larger for the informal sector than the formal sector across different industries.

[– Table 5 –]

5.2 Spatial Marginal Effects

We proceed to estimate the marginal effects of informal agglomeration. In order to compute marginal effects in a spatial autoregressive model, we must take into account spatial dependence in economic performance across regions. Without significant spatial correlation, the estimation is reduced to a standard elasticity interpretation based on a non-spatial regression model. Our estimation results show that the posterior mean coefficients of agg_{si}^{inf} should simply pick up the magnitude of the elasticity of localization economies in wholesale/retail industries because their posterior means of ρ are not significant in both formal and informal sectors. However, we find significantly positive posterior means for ρ in manufacturing formal and informal firms. Thus, a marginal change in agg_{si}^{inf} for each region has an impact not only on its own performance but also indirectly on the other regions' performance through the spatial structure. In the following, we focus on the spatial marginal effects in manufacturing.

In general, the marginal impact differs across all the regions. This can be shown by the following reduced form of Equation (6):

$$\mathbf{x}_s = \mathbf{Q}_s \boldsymbol{\gamma}_{s0} + \mathbf{Z}_s \boldsymbol{\gamma}_{s1} + \boldsymbol{\eta}_s \tag{12}$$

$$\mathbf{y}_s = \mathbf{S}_s^{-1} \beta_{s0} \mathbf{x}_s + \mathbf{S}_s^{-1} \mathbf{Z}_s \boldsymbol{\beta}_{s1} + \mathbf{S}_s^{-1} \boldsymbol{\varepsilon}_s \,, \tag{13}$$

where $\mathbf{S}_s = \mathbf{I}_n - \rho_s \mathbf{W}$. The total derivative of $\mathbf{y}_s$ under the constraints $d\mathbf{Z}_s = \mathbf{0}$ and $d\boldsymbol{\varepsilon}_s = d\boldsymbol{\eta}_s = \mathbf{0}$ is

$$\begin{bmatrix} dy_{s1} \\ dy_{s2} \\ \vdots \\ dy_{sn} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_{s1}}{\partial x_{s1}} & \frac{\partial y_{s1}}{\partial x_{s2}} & \cdots & \frac{\partial y_{s1}}{\partial x_{sn}} \\ \frac{\partial y_{s2}}{\partial x_{s1}} & \frac{\partial y_{s2}}{\partial x_{s2}} & \cdots & \frac{\partial y_{s2}}{\partial x_{sn}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_{sn}}{\partial x_{s1}} & \frac{\partial y_{sn}}{\partial x_{s2}} & \cdots & \frac{\partial y_{sn}}{\partial x_{sn}} \end{bmatrix} \begin{bmatrix} dx_{s1} \\ dx_{s2} \\ \vdots \\ dx_{sn} \end{bmatrix} \tag{14}$$

$$d\mathbf{y}_s = \tilde{\mathbf{S}}_s \, d\mathbf{x}_s$$

where $\tilde{\mathbf{S}}_s = \partial \mathbf{y}_s / \partial \mathbf{x}_s' = \mathbf{S}_s^{-1} \beta_{s0}$ is an $n \times n$ matrix. Equation (14) indicates that

$$\text{Total impact from } x_{si} \text{ to } \mathbf{y}_s: \quad TI_i = \sum\nolimits_{j=1}^{n} \frac{\partial y_{sj}}{\partial x_{si}} dx_{si}, \ (i = 1, 2, \ldots, n)$$

$$\text{Direct impact from } \mathbf{x}_s \text{ to } \mathbf{y}_s: \quad DI = \begin{bmatrix} \frac{\partial y_{s1}}{\partial x_{s1}} & 0 & \cdots & 0 \\ 0 & \frac{\partial y_{s2}}{\partial x_{s2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\partial y_{sn}}{\partial x_{sn}} \end{bmatrix} \begin{bmatrix} dx_{s1} \\ dx_{s2} \\ \vdots \\ dx_{sn} \end{bmatrix}$$

$$\text{Indirect impact from } \mathbf{x}_s \text{ to } \mathbf{y}_s: \quad IDI = \begin{bmatrix} 0 & \frac{\partial y_{s1}}{\partial x_{s2}} & \cdots & \frac{\partial y_{s1}}{\partial x_{sn}} \\ \frac{\partial y_{s2}}{\partial x_{s1}} & 0 & \cdots & \frac{\partial y_{s2}}{\partial x_{sn}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_{sn}}{\partial x_{s1}} & \frac{\partial y_{sn}}{\partial x_{s2}} & \cdots & 0 \end{bmatrix} \begin{bmatrix} dx_{s1} \\ dx_{s2} \\ \vdots \\ dx_{sn} \end{bmatrix}.$$

Thus, the marginal changes in agglomeration for region i can influence the performance of not only that region but also other regions. As $\tilde{\mathbf{S}}_s$ reduces to an identity matrix of size n under $\rho = 0$, such indirect effects disappear if $\rho = 0$.

To summarize the marginal effects of agg_{si}^{inf}, we calculate the averages of TI_i, DI and IDI over all regions, based on the posterior means of parameters and $dx_s = 1$. These are referred to as Average Total Impact (ATI), Average Direct Impact (ADI), and Average Indirect Impact ($AIDI = ATI - ADI$), using the terminology proposed by LeSage and Pace (2009). Table 6 presents a summary of estimated impacts on the economic performance of formal and informal firms. In terms of sales per worker, ADI is 0.089 for formal sector and 0.135 for informal sector. These figures are equivalent to the coefficients of agg_{si}^{inf} in Table 4. $AIDI$ is 0.019 for the formal sector and 0.025 for the informal sector. Combining these values we find ATI of 0.108 and 0.160, respectively, for the formal and informal sectors. Intuitively, these results indicate that a doubling of the density of informal activity in a region increases the sales per worker of formal firms by 8.9% in that region and by 1.9% in other regions through spatial multiplier linkages, which result in a total 10.8% increase. On the other hand, a doubling of the density of informal activity in a region increases the sales per worker of informal firms by 13.5% in that region and by 2.5% in other regions through the spatial multiplier linkages, which result in a total 16.0% increase. Additionally, we find that the ATI of wages per employee is 0.042 for the formal sector and 0.090 for the informal sector. Although the ADI is much larger than the $AIDI$, it is evidently important to account for both local and spatial impacts in the

presence of significant spatial dependence of economic performance.

[– Table 6 –]

The analysis up to this point demonstrates that the magnitudes of TI_i in the informal sector are globally larger than in the formal sector. However, the geographic pattern of informal activity is not uniform across regions, and it is useful to investigate the geographic distribution of estimated total impacts. We show TI_i on sales per worker for the formal sector in Figure 3 and for the informal sector in Figure 4. These figures show that the total impacts are not uniformly distributed among regions for either sector. These differences depend on the structure of $\mathbf{S}_s^{-1}\beta_{s0} = (\mathbf{I}_n - \rho_s\mathbf{W})^{-1}\beta_{s0}$, which can be expanded to the geometric series $(\beta_{s0}\mathbf{I}_n + \rho_s\mathbf{W} + \rho_s^2\mathbf{W}^2 + \cdots)$. Since the spatial matrix $\mathbf{W}$ is common to both sectors, the differences between Figures 3 and 4 is produced by the difference of ρ_s and β_{0s}, which are estimated separately for the formal and informal sectors. Our estimates of these parameters imply that the difference between these figures is attributed to the larger estimate of β_{0s} for the informal sector than that for the formal sector.

[– Figures 3 and 4 –]

The regional variation in TI_i is determined by the structure of $\mathbf{W}$, as the parameters ρ_s and β_{0s} are constant for all communes. Since we construct $\mathbf{W}$ as the degree of connectivity among communes measured by the traveling time, more accessible communes tend to have larger TI_i. On the other hand, less accessible communes are likely to have smaller TI_i. To investigate the determinants of these spatial impacts more specifically, we regress TI_i on commune characteristics, including road infrastructure, geographic conditions, and relative location within the country.[15] The regression results for the formal and informal sectors are presented in Table 7. These results suggest that TI_i is relatively large in the communes near to one- and two-digit national roads, but relatively small in the tree-covered and regularly flooded communes. Additionally, we find that both formal and informal firms in Phnom Penh province, the capital of Cambodia, tend to benefit from increased agglomeration of informal sectors. Finally, the spatial impacts are smaller in the border communes.

[– Table 7 –]

6 Robustness Checks

6.1 Alternative Definition of Informal Agglomeration

In the benchmark analysis, we measure informal agglomeration by the density of workers in the informal sector. As discussed by Henderson (2003) and Martin et al. (2011), localization economies can be measured by the number of firms in the same industry and region. If agglomeration externalities occurred mainly through the interactions of firms rather than workers, we might underestimate the impact of informal agglomeration in the benchmark results. To

[15] See Appendix C for the variable definitions.

address this concern, we redefine the informal agglomeration as the density of firms in the same industry and commune, and re-estimate all the benchmark specifications using the firm-density variable.

Table 8 presents the estimation results using sal_{si}^{fml} and sal_{si}^{inf} as a dependent variable. In manufacturing industries, the posterior mean coefficients of agg_{si}^{inf} are 0.095 for the formal sector and 0.144 for the informal sector. Compared to the benchmark estimates in Table 4, these estimates are slightly larger. In wholesale/retail industries, the posterior mean coefficients of agg_{si}^{inf} are 0.136 for the formal sector and 0.154 for the informal sector. These estimates are similar to the benchmark results in Table 4. Across specifications, the posterior mean coefficients of emp_{si}^{1998} and $green_{i}^{2002}$ are significantly positive and negative, respectively. Also, a Sargan statistic has the large p-values across specifications. Thus, these results support that the density of informal firms has a positive impact on sales per worker for both formal and informal firms. In the case of manufacturing, the existence of informal firms appears to play a slightly larger role in localization economies than the presence of their workers.

[– Table 8 –]

Table 9 shows the estimation results using wge_{si}^{fml} and wge_{si}^{inf} as a dependent variable. In manufacturing industries, we find that the posterior mean coefficients of agg_{si}^{inf} are significantly positive for both formal and informal sectors. On the other hand, agg_{si}^{inf} has a significantly positive coefficient only for the informal sector. When comparing with the benchmark results in Table 5, we see that the significance and magnitude of the estimates are similar across specifications. As statistical information, such as a Sargan statistic, supports the validity of instrumental variables, the density of informal firms has a positive effect on wages per employee in the informal sector. In the case of manufacturing, the formal sector also benefits from the density of informal firms.

[– Table 9 –]

6.2 Spatial Dimension of Informal Agglomeration

In the main analysis, we focused on the effects of informal agglomeration within the same commune because the interaction of formal and informal firms should be the most intensive in proximity and would weaken with increasing distance. On the other hand, it is plausible that informal activity in nearby communes also generates some agglomeration externality, implying that localization economies of informal activity may be better captured by taking into account the density of informal activity in other communes, with a larger weight for the closer communes. To investigate this issue, we redefine the informal agglomeration as $(\mathbf{I} + \mathbf{W})agg_{si}^{inf}$, where $\mathbf{W}$ is a spatial-weighting matrix. Using the spatial variable of the density of workers in informal firms, we estimate all the benchmark specifications. Note that we also redefine the instrumental variables as $(\mathbf{I} + \mathbf{W})q_{si}$, which accounts for the spatial dimension with the same weighting matrix.

Table 10 presents the estimation results using sal_{si}^{fml} and sal_{si}^{inf} as a dependent variable. In manufacturing industries, the posterior mean coefficient of $(\mathbf{I} + \mathbf{W})agg_{si}^{inf}$ is 0.047 for the informal sector with high significance. In wholesale/retail industries, the posterior mean coefficient

is 0.038 for the informal sector with high significance. Because the statistical tests support the validity of instrumental variables, the spatially defined measure of informal agglomeration also yields positive causal impacts on sales per worker for informal firms. Compared with the benchmark estimates in Table 4, the estimates of informal agglomeration are substantially smaller. Moreover, the estimates are not significant for the formal sector in these industries. Additional, unreported results show that the posterior mean coefficient of $(\mathbf{I} + \mathbf{W})agg_{si}^{inf}$ is significantly positive when wge_{si}^{inf} in manufacturing industry is used as a dependent variable. However, the estimates of informal agglomeration across specifications tend to become smaller than the benchmark estimates in Table 5.

[– Table 10 –]

6.3 Alternative Weighting Matrix

The benchmark specification is estimated with a spatial weight matrix constructed from the minimum travelling time across communes in Cambodia. We believe that this spatial structure represents connectivity across communes sufficiently well. However, as discussed by LeSage and Pace (2009, chapter 6), there is a question of which spatial weight matrix is most appropriate to account for spatial dependence of regional economic performance. Alternative spatial weight matrices could produce different results for informal agglomeration effects. To deal with this concern, we exploit a nearest-neighbor weight matrix based on 5 or 10 nearest-neighboring communes. With these alternative weight matrices, we re-estimate all the benchmark specifications.

Table 11 presents the estimation results for manufacturing industry in which a weight matrix is based on 5 or 10 nearest-neighboring communes. The dependent variable is sal_{si}^{fml} or sal_{si}^{inf}. With the 5-nearest neighbors, the posterior mean coefficients of agg_{si}^{inf} are 0.088 for the formal sector and 0.133 for the informal sector, respectively. The posterior mean coefficients are similar for 10 nearest neighbors. As the validity of instrumental variables is supported by the statistical tests, the significantly positive impact of informal agglomeration on regional economic performance is robust to the use of an alternative spatial-weighting matrix. Additionally, in unreported results using wge_{si}^{fml} and wge_{si}^{inf} as a dependent variable, we find similar posterior mean coefficients of agg_{si}^{inf} to those given in Table 5. Finally, the results are also similar for wholesale/retail industries.

[– Table 11 –]

It should be emphasized that an alternative weight matrix has a large influence on the magnitude of the spatial autocorrelation in y_{si}, denoted by ρ_s. In the case of sal_{si}^{fml} for manufacturing industry, the benchmark estimate of ρ_s is 0.180. In Table 11, ρ_s is 0.132 for 5 nearest neighbors and 0.211 for 10 nearest neighbors. When sal_{si}^{inf} is used, the estimates of ρ_s are 0.156 for the main analysis, 0.107 for 5 nearest neighbors, and 0.166 for 10 nearest neighbors. Compared with the benchmark spatial weight matrix, the 5-nearest neighbors matrix tends to produce a lower estimate for the spatial dependence in regional economic performance whereas the 10-nearest neighbors matrix yields a higher estimate. These patterns are also found in the results using wge_{si}^{fml} and wge_{si}^{inf} as a dependent variable. Therefore, the implication is that alternative

spatial-weighting matrices should have an influence on the indirect effects of informal agglomeration over space. It is, therefore, crucial to construct an appropriate weight matrix in order to obtain precise estimates of spatial marginal effects in the presence of significant spatial dependence.

6.4 Additional Control Variable

Finally, we check the robustness of estimation results to additional control variable. A plausible concern is that the presence of formal sector may also affect regional economic performance, but is not controlled in benchmark specifications. To address such a potential confounding factor, we include a dummy variable that takes on unity for communes in which at least one formal firm in similar industry exists, and zero otherwise. In the data, formal firms exist in 379 communes for manufacturing and in 425 communes for wholesale/retail industries, respectively. These communes accounted for 23.4% and 26.2% of 1,621 communes in total, respectively. Because the dummy variable of formal sector explains whether we observe economic performance of formal firms, it may not be sensible to examine the robustness for formal firms' performance. Thus, we focus to examine whether the presence of formal sector affects the positive impact of informal agglomeration on regional economic performance of informal firms.

Table 12 presents the estimation results of benchmark specifications with the additional control variable dum_{si}^{fml}. In the case of manufacturing, the posterior mean coefficients of agg_{si}^{inf} are 0.130 for sal_{si}^{inf} and 0.069 for wge_{si}^{inf}, respectively. For the sample of wholesale/retail industries, the posterior mean coefficients are 0.141 for sal_{si}^{inf} and 0.062 for wge_{si}^{inf}, respectively. These posterior mean coefficients are slightly smaller in magnitude than those in the corresponding benchmark specifications, but they are statistically significant across specifications. Also, the inclusion of dum_{si}^{fml} has little influence on the validity of instrumental variables, implying that statistical tests lend support for the causality interpretation. Taken together, these results indicate the robustness of benchmark results to the additional control variable dum_{si}^{fml}.

[– Table 12 –]

7 Concluding Remarks

An informal sector plays a significant role in developing economies and a spatial concentration of informal activity poses a crucial question about whether and to what extent informal agglomeration generates positive or negative externality. This paper employs the economic census on all the establishments in Cambodia to estimate the impact of informal agglomeration on the economic performance of formal and informal firms at the regional level. To estimate a causal impact, we develop a Bayesian spatial approach to take into account both the endogeneity of informal agglomeration economies and spatial dependence in economic performance. The results show that the spatial concentration of informal firms yields a positive impact on the regional economic performance of both formal and informal firms in manufacturing and wholesale/retail industries. The positive impact is larger for informal firms than for formal firms, implying that informal firms may have weaker linkages with formal firms than with other informal firms. Additionally, informal agglomeration generates spatial multiplier effects across regions, where

more accessible regions are more likely than less accessible regions to benefit strongly from informal agglomeration.

These results have important implications for policy makers trying to formulate an effective industrial policy. In terms of maximizing agglomeration economies, governments in developing economies should consider both formal and informal activity, because supporting informal firms may also contribute effectively to improving the performance of domestic industry. On the other hand, policy instruments should be carefully designed by taking into account the geography of industrial activity because informal firms in the more accessible regions tend to yield larger externality than those in the less accessible regions. Moreover, transportation infrastructure should be carefully constructed and improved in urban and rural regions because congestion effects of industrial concentration may weaken the positive externality and the peripheral regions may not produce a large positive externality.

Finally, we mention some remaining issues for future research on informal agglomeration. While our study focuses on the agglomeration effects of informal activity, selection effects of individual firms are not separately identified. As prior studies, such as Combes et al. (2012) and Arimoto et al. (2014), have investigated sources of productivity improvements in cities and industrial clusters, respectively, an examination of the distinctive channels is promising. A related question is how informal firms choose to become formal through official registration and whether industrial clusters would promote the formalization of informal activity through agglomeration externality.

References

Andersson, F., Burgess, S. and Lane, J.I. (2007) Cities, matching and the productivity gains of agglomeration. *Journal of Urban Economics*, 61(1), 112–128.

Annez, P.C. and Buckely, R.M. (2009) Urbanization and growth: setting the context. In: M. Spence, P. C. Annez, and R. M. Buckely, (ed) *Urbanization and Growth*, the Commission on Growth and Development, Washington, D.C.

Arimah, B.C. (2001) Nature and determinants of the linkages between informal and formal sector enterprises in Nigeria. *African Development Review*, 13(1), 114–144.

Arimoto, Y., Nakajima, K., Okazaki, T. (2014) Sources of productivity improvement in industrial clusters: the case of the prewar Japanese silk-reeling industry. *Regional Science and Urban Economics* 46, 27–41.

Artis, M.J., Miguélez, E. and Moreno, R. (2012) Agglomeration economies and regional intangible assets: an empirical investigation. *Journal of Economic Geography*, 12(6), 1167–1189.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton.

Broersma, L. and Oosterhaven, J. (2009) Regional labor productivity in the Netherlands: evidence of agglomeration and congestion effects. *Journal of Regional Science* 49(3), 483–511.

Brülhart, M. and Mathys, N.A. (2008) Sectoral agglomeration economies in a panel of European regions. *Regional Science and Urban Economics* 38(4), 348–362.

Cassey, A.J., Smith, B.O. (2014) Simulating confidence for the Ellison-Glaeser index, *Journal of Urban Economics* 81, 85–103.

Ciccone, A. (2002) Agglomeration effects in Europe. *European Economic Review* 46(2), 213–227.

Ciccone, A. and Hall, R.E. (1996) Productivity and the density of economic activity. *American Economic Review* 86 (1), 54–70.

Cohen, J.P. and Paul, C.J.M. (2009) Agglomeration, productivity and regional growth: production theory approaches. In: Capello R, Nijkamp P (ed) *Handbook of regional growth and development theories*. Edward Elgar, Cheltenham, UK, pp. 101–117.

Combes, P., Duranton, G. and Gobillon, L. (2008) Spatial wage disparities: sorting matters! *Journal of Urban Economics* 63(2), 723–742.

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

Combes, P., Duranton, G., Gobillon, L. and Roux, S. (2010) Estimating agglomeration economies with history, geology, and worker effects. In: E.L. Glaeser, (ed) *Agglomeration Economics*, The University of Chicago Press.

Dekle, R., 2002. Industrial concentration and regional growth: evidence from the prefectures. *Review of Economics and Statistics* 84 (2), 310-315.

de Mel, S., McKenzie, D.J. and Woodruff, C. (2009) Measuring microenterprise profits: must we ask how the sausage is made? *Journal of Development Economics* 88(1), 19–31.

de Paula, Á. and Scheinkman, J.A. (2011) The informal sector: an equilibrium model and some empirical evidence from Brazil. *Review of Income and Wealth*, 57(S1), S8–S26.

Doornik, J.A. (2006) *Ox: An objected-oriented matrix programming language*. Timberlake Consultants, London.

Drèze, J. H. (1976) Bayesian limited information analysis of the simultaneous equations model, *Econometrica*, 44(5), 1045–1075.

Duranton, G. (2009) Are cities engines of growth and prosperity for developing countries? In: M. Spence, P. C. Annez, and R. M. Buckely, (ed) *Urbanization and Growth*, The Commission on Growth and Development, Washington, D.C.

Duranton, G. and Puga, D. (2004) Micro-foundations of Urban Agglomeration Economies. In J.V. Henderson and J.-F. Thisse, (ed) *Handbook of regional and urban economics*, vol. 4, North-Holland, New York; pp. 2063–2117.

Eberts, R.W. and McMillen, D.P. (1999) Agglomeration economies and urban public infrastructure. In: Cheshire, P., Mills, E.S. (eds) *Handbook of regional and urban economics*, volume 3. North-Holland, New York, pp. 1455–1495.

Ellison, G. and Glaeser, E.L. (1997) Geographic concentration in U.S. manufacturing industries: a dartboard approach. *Journal of Political Economy*, 105(5), 889–927.

Fajnzylber, P., Maloney, W.F., and Montes-Rojas, G.V. (2011) Does formality improve microfirm performance? Evidence from the Brazilian SIMPLES program. *Journal of Development Economics*, 94(2), 262–276.

Fujita, M., Krugman, P. and Venables, A.J. (1999) *The spatial economy: Cities, regions, and international trade*. MIT Press, Cambridge and London.

Fujita, M. and Thisse, J.F. (2003) Does Geographical Agglomeration Foster Economic Growth? And Who Gains and Loses From It? *Japanese Economic Review* 54(2), 121–145.

Gamerman, D. and Lopes, H.F. (2006) *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference*, second edition. Chapman and Hall/CRC, Boca Raton. Gibbons, S. and Overman, H.G. (2012) Mostly Pointless Spatial Econometrics? *Journal of Regional*

Science, 52(2), 172-191.
Gibbons, S., Overman, H.G. (2012) Mostly pointless spatial econometrics? *Journal of Regional Science*, 52(2), 172-191.
Glaeser, E.L. and Maré, D.C. (2001) Cities and skills, *Journal of Labor Economics*, 19(2), 316–342.
Graham, D.J. (2009) Identifying urbanization and localisation externalities in manufacturing and service industries. *Papers in Regional Science*, 88(1), 63–84.
Hatsukano, N, Kuroiwa, I. and Tsubota, K. (2012) Economic integration and industry location in Cambodia. In: I. Kuroiwa (ed) *Economic Integration and the Location of Industries*, Palgrave Macmillan, New York.
Henderson, J. Vernon (2003) Marshall's scale economies. *Journal of Urban Economics* 53(1), 1–28.
Hill, H. and Menon, J. (2013) Cambodia: rapid growth with weak institutions. *Asian Economic Policy Review*, 8(1), 46–65.
IMF. (2012) *World Economic Outlook 2012*, Washington, D.C.
JETRO. (2009) *ASEAN Logistics Network Map*, 2nd ed. JETRO, Tokyo.
Kelejian, H.H. and Prucha, I.R. (1998) A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* 17, 99–121.
Kolko, J. (2007) Agglomeration and co-agglomeration of services industries. *MPRA Paper* No. 3362, University Library of Munich, Munich.
Lall, S.V., Shalizi, Z. and Deichmann, U. (2004) Agglomeration economies and productivity in Indian industry. *Journal of Development Economies*, 73(2), 643–673.
LeSage, J. and Pace, P.K. (2009) *Introduction to spatial econometrics*. CRC press, Boca Raton.
Livingstone, I. (1991) A reassessment of Kenya's rural and urban informal sector. *World Development*, 19(6), 651–670.
Martin, P., Mayer, T., Mayneris, F. (2011) Spatial concentration and plant-level productivity in France. *Journal of Urban Economics* 69(2), 182–195.
McKenzie, D. and Seynabou Sakho, Y. (2010) Does it pay firms to register for taxes? The impact of formality on firm profitability. *Journal of Development Economics*, 91(1), 15–24.
Melo, P.C, Graham, D.J. and Noland, R.B. (2009) A meta-analysis of estimates of urban agglomeration economies. *Regional Science and Urban Economics*, 39(3), 332–342.
Morikawa, M. (2011) Economies of density and productivity in service industries: an analysis of personal service industries based on establishment-level data. *Review of Economics and Statistics* 93(1), 179–192.
Mukin, M. (2014) Coagglomeration of formal and informal industry: evidence from India. *Journal of Economic Geography*, forthcoming.
Murray, M.P. (2006) Avoiding Invalid Instruments and Coping with Weak Instruments. *Journal of Economic Perspectives*, 20(4), 111–132.
Ord, K. (1975) Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349), 120–126.
Overman, H.G. and Venables, A.J. (2005) Cities in the Developing World. *CEP Discussion Paper* No. 695, Department for International Development, London.
Puga, D. (2010) The magnitude and causes of agglomeration economies. *Journal of Regional Science* 50(1), 203–219.

Rand, J. and Torm, N. (2012) The benefits of formalization: evidence from Vietnamese manufacturing SMEs. *World Development*, 40(5), 983–998.

Rosenthal, S., and Strange, W.C. (2004) Evidence on the nature and sources of agglomeration economies. In: J.V. Henderson and J.F. Thisse, (ed) *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam. pp. 2119–2171.

Rossi, P.E., Allenby, G.M. and McCulloch, R. (2005) *Bayesian statistics and marketing*. John Wiley and Sons Lts, West Sussex.

Schneider, F. (2005) Shadow economies around the world: what do we really know? *European Journal of Political Economy*, 21(3), 598–642.

Schneider, F., Buehn A. and Montenegro, C.E. (2010) Shadow economies all over the world: new estimates for 162 countries from 1999 to 2007. *Policy Research Working Paper* 5356, World Bank.

Schneider, F. and Enste, D.H. (2000) Shadow economies: size, causes, and consequences. *Journal of Economic literature*, 38(1): 77–114.

Wheaton, W.C. and Lewis, M.J. (2002) Urban wages and labor market agglomeration. *Journal of Urban Economics* 51(3), 542–562.

Williamson, J.G. (1965) Regional inequality and the process of national development, *Economic Development and Cultural Change*, 13(4), 3–45.

William, C.C. and Lansky, M.A. (2013) Informal employment in developed and developing economies: perspectives and policy responses. *International Labour Review*, 152(3-4), 355–380.

World Bank. (2014) *Doing business 2014: understanding regulations for small and medium-size enterprises*, Washington D.C.

–Tables and Figures–

Table 1: Formal and Informal Sectors in Cambodia

	All		Manufacturing		Wholesale/Retail	
	Formal	Informal	Formal	Informal	Formal	Informal
Number of establishment	17,374	487,719	1,723	69,693	6,245	286,101
	(3.4)	(96.6)	(2.4)	(97.6)	(2.1)	(97.9)
Sales (mil. USD)	140.31	459.75	20.91	32.49	42.01	301.82
	(23.4)	(76.6)	(39.2)	(60.8)	(12.2)	(87.8)
Wages (mil. USD)	14.46	20.97	4.18	2.33	1.52	2.33
	(40.8)	(59.2)	(64.2)	(35.8)	(39.5)	(60.5)
Employment (mil. people)	0.561	1.112	0.358	0.172	0.038	0.515
	(33.6)	(66.4)	(67.5)	(32.5)	(6.9)	(93.1)

Notes: Formal and Informal indicate registered and unregistered firms, respectively; values in parentheses are the percentage share of each sector in the corresponding variable.

Table 2: Agglomeration Indexes of Informal Sectors

	Manufacturing	Wholesale/Retail
A. Commune-level Employment		
Geographic concentration	0.003	0.002
EG agglomeration index	0.002	0.002
B. District-level Employment		
Geographic concentration	0.009	0.005
EG agglomeration index	0.009	0.005
C. Province-level Employment		
Geographic concentration	0.036	0.018
EG agglomeration index	0.041	0.021

Table 3: Summary Statistics

	Obs.	Mean	Std. Dev.	Min	Max
Manufacturing					
sal_{si}^{fml}	1621	0.562	1.267	0.000	6.976
sal_{si}^{inf}	1621	2.548	1.133	0.000	8.286
wge_{si}^{fml}	1621	0.217	0.596	0.000	5.017
wge_{si}^{inf}	1621	0.644	0.744	0.000	5.314
agg_{si}^{inf}	1621	−0.028	2.190	−7.191	8.949
emp_{si}^{1998}	1621	−0.866	2.555	−7.582	9.820
Wholesale/Retail					
sal_{si}^{fml}	1621	0.954	1.840	0.000	8.249
sal_{si}^{inf}	1621	3.625	0.954	0.878	7.825
wge_{si}^{fml}	1621	0.237	0.590	0.000	3.604
wge_{si}^{inf}	1621	0.550	0.707	0.000	5.127
agg_{si}^{inf}	1621	1.393	1.866	−3.898	11.188
emp_{si}^{1998}	1621	−0.043	2.662	−7.582	10.701
Other variables					
pop_i	1621	8.820	0.643	5.700	11.403
$elec_i$	1621	0.188	0.277	0.003	1.000
$hskil_i$	1621	0.019	0.046	0.000	0.512
$lskil_i$	1621	0.594	0.240	0.025	1.000
$crplnd_i$	1621	2.981	1.481	0.000	6.111
$border_i$	1621	0.062	0.242	0.000	1.000
air_i	1621	0.026	0.159	0.000	1.000
sez_i	1621	0.079	0.270	0.000	1.000
$green_i^{2002}$	1621	0.128	0.715	0.000	7.145
$hskil_i^{1998}$	1621	0.064	0.098	0.000	1.000

Table 4: Estimation Results for Sales per Worker

	Manufacturing		Wholesale/Retail	
	sal_{si}^{fml}	sal_{si}^{inf}	sal_{si}^{fml}	sal_{si}^{inf}
Explanatory Variables (x_{si}, $\mathbf{z}_{si}$)				
agg_{si}^{inf}	0.089**	0.135**	0.136**	0.153**
	(0.033)	(0.032)	(0.041)	(0.025)
pop_i	0.262**	0.443**	0.425**	0.236**
	(0.062)	(0.059)	(0.076)	(0.047)
$elec_i$	1.524**	0.531**	3.023**	0.797**
	(0.150)	(0.140)	(0.202)	(0.124)
$hskil_i$	1.164	−2.677**	3.122**	−2.206**
	(0.901)	(0.824)	(1.204)	(0.711)
$lskil_i$	−0.021	0.121	0.152	−0.057
	(0.172)	(0.163)	(0.202)	(0.125)
$crplnd_i$	−0.008	0.057**	−0.014	0.023
	(0.023)	(0.022)	(0.030)	(0.019)
$border_i$	0.037	−0.099	0.285	0.034
	(0.112)	(0.106)	(0.146)	(0.091)
air_i	−0.340	−0.022	−0.171	−0.010
	(0.180)	(0.170)	(0.236)	(0.147)
sez_i	0.022	0.068	−0.233	0.006
	(0.109)	(0.103)	(0.142)	(0.088)
$const$	−2.143**	−2.060**	−3.556**	0.783
	(0.538)	(0.515)	(0.611)	(0.410)
ρ	0.180**	0.156**	−0.081	0.108
	(0.048)	(0.057)	(0.052)	(0.058)
Instrumental Variables ($\mathbf{q}_{si}$)				
emp_{si}^{1998}	0.500**	0.500**	0.523**	0.524**
	(0.022)	(0.022)	(0.013)	(0.013)
$green_i^{2002}$	−0.180**	−0.175**	−0.116**	−0.108**
	(0.048)	(0.048)	(0.030)	(0.029)
$hskil_i^{1998}$	0.283	0.329	0.147	0.147
	(0.366)	(0.368)	(0.227)	(0.220)
Sargan	1.003	0.070	1.195	0.470
(p-value)	(0.605)	(0.966)	(0.550)	(0.791)
No. of obs.	1621	1621	1621	1621

Notes: Asterisks * (**) indicate that the 95% (99%) credible interval of the posterior distribution does not contain zero. Values in parentheses are standard deviations for each posterior distribution.

Table 5: Estimation Results for Wage per Employee

	Manufacturing		Wholesale/Retail	
	wge_{si}^{fml}	wge_{si}^{inf}	wge_{si}^{fml}	wge_{si}^{inf}
Explanatory Variables (x_{si}, $\mathbf{z}_{si}$)				
agg_{si}^{inf}	0.034*	0.078**	0.021	0.075**
	(0.016)	(0.020)	(0.013)	(0.018)
pop_i	0.136**	0.182**	0.112**	0.171**
	(0.030)	(0.037)	(0.023)	(0.033)
$elec_i$	0.634**	0.961**	0.904**	0.772**
	(0.071)	(0.089)	(0.062)	(0.086)
$hskil_i$	1.086*	−1.116*	2.876**	−0.452
	(0.438)	(0.523)	(0.389)	(0.502)
$lskil_i$	−0.060	−0.049	−0.041	0.077
	(0.082)	(0.103)	(0.062)	(0.086)
$crplnd_i$	−0.004	0.040**	−0.010	0.005
	(0.011)	(0.014)	(0.009)	(0.013)
$border_i$	0.130*	−0.012	0.071	−0.148*
	(0.053)	(0.067)	(0.045)	(0.063)
air_i	0.005	0.022	0.079	0.027
	(0.085)	(0.107)	(0.073)	(0.102)
sez_i	0.026	−0.026	−0.014	−0.017
	(0.052)	(0.065)	(0.043)	(0.061)
$const$	−1.131**	−1.301**	−0.949**	−1.312**
	(0.257)	(0.322)	(0.187)	(0.263)
ρ	0.187**	0.129*	−0.009	0.096
	(0.051)	(0.054)	(0.047)	(0.057)
Instrumental Variables ($\mathbf{q}_{si}$)				
emp_{si}^{1998}	0.500**	0.501**	0.523**	0.523**
	(0.022)	(0.022)	(0.013)	(0.013)
$green_i^{2002}$	−0.183**	−0.172**	−0.113**	−0.112**
	(0.048)	(0.048)	(0.030)	(0.030)
$hskil_i^{1998}$	0.295	0.292	0.149	0.145
	(0.367)	(0.370)	(0.228)	(0.227)
Sargan	1.946	4.521	0.477	0.309
(p-value)	(0.378)	(0.104)	(0.788)	(0.857)
No. of obs.	1621	1621	1621	1621

Notes: Asterisks * (**) indicate that the 95% (99%) credible interval of the posterior distribution does not contain zero. Values in parentheses are standard deviations for each posterior distribution.

Table 6: Spatial Marginal Effects

	Manufacturing			
	sal_{si}^{fml}	sal_{si}^{inf}	wge_{si}^{fml}	wge_{si}^{inf}
Average Total Impact	0.108	0.160	0.042	0.090
Average Direct Impact	0.089	0.135	0.034	0.078
Average Indirect Impact	0.019	0.025	0.008	0.012

Table 7: Determinants of Spatial Impacts

Dependent variable: total impact of increased informal agglomeration

	(1)	(2)
Variable	Formal	Informal
One digit national road	0.014*	0.017*
	(0.00044)	(0.00056)
Two digit national road	0.0035*	0.0045*
	(0.00049)	(0.00062)
Tree/flooded area	−0.0088*	−0.011*
	(0.00053)	(0.00066)
Phnom Penh province	−0.0074*	−0.0094*
	(0.00087)	(0.0011)
National border	−0.0042*	−0.0053*
	(0.00079)	(0.00099)
R-squared	0.46	0.46
No. of obs.	1,621	1,621

Notes: Values in parentheses are standard errors; * indicates significance at the 1% level.

Table 8: Robustness to Firm Density for Sales per Worker

	Manufacturing		Wholesale/Retail	
	sal_{si}^{fml}	sal_{si}^{inf}	sal_{si}^{fml}	sal_{si}^{inf}
Explanatory Variables (x_i, $\mathbf{z}_i$)				
agg_{si}^{inf}	0.095**	0.144**	0.136**	0.154**
	(0.036)	(0.034)	(0.041)	(0.025)
pop_i	0.271**	0.457**	0.426**	0.237**
	(0.060)	(0.058)	(0.076)	(0.047)
$elec_i$	1.561**	0.584**	3.015**	0.788**
	(0.151)	(0.141)	(0.202)	(0.122)
$hskil_i$	1.159	−2.709**	3.236**	−2.082**
	(0.905)	(0.829)	(1.193)	(0.697)
$lskil_i$	−0.025	0.114	0.142	−0.067
	(0.173)	(0.164)	(0.203)	(0.124)
$crplnd_i$	−0.005	0.060**	−0.014	0.023
	(0.023)	(0.022)	(0.030)	(0.018)
$border_i$	0.037	−0.098	0.290*	0.039
	(0.112)	(0.107)	(0.146)	(0.090)
air_i	−0.346	−0.030	−0.186	−0.029
	(0.180)	(0.170)	(0.236)	(0.145)
sez_i	0.018	0.063	−0.234	0.004
	(0.109)	(0.103)	(0.142)	(0.087)
$const$	−2.165**	−2.093**	−3.489**	0.833*
	(0.534)	(0.511)	(0.620)	(0.413)
ρ	0.177**	0.158**	−0.080	0.118*
	(0.048)	(0.057)	(0.052)	(0.057)
Instrumental Variables ($\mathbf{q}_{si}$)				
emp_{si}^{1998}	0.469**	0.469**	0.522**	0.523**
	(0.020)	(0.020)	(0.012)	(0.012)
$green_i^{2002}$	−0.164**	−0.160**	−0.118**	−0.112**
	(0.045)	(0.045)	(0.028)	(0.028)
$hskil_i^{1998}$	0.281	0.327	0.126	0.128
	(0.341)	(0.343)	(0.216)	(0.212)
Sargan	1.029	0.066	1.175	0.499
(p-value)	(0.598)	(0.968)	(0.556)	(0.779)
No. of obs.	1621	1621	1621	1621

Notes: Asterisks * (**) indicate that the 95% (99%) credible interval of the posterior distribution does not contain zero. Values in parentheses are standard deviations for each posterior distribution.

Table 9: Robustness to Firm Density for Wage per Employee

	Manufacturing		Wholesale/Retail	
	wge_{si}^{fml}	wge_{si}^{inf}	wge_{si}^{fml}	wge_{si}^{inf}
Explanatory Variables (x_i, $\mathbf{z}_i$)				
agg_{si}^{inf}	0.036*	0.084**	0.021	0.075**
	(0.017)	(0.022)	(0.013)	(0.018)
pop_i	0.139**	0.190**	0.112**	0.173**
	(0.029)	(0.036)	(0.023)	(0.033)
$elec_i$	0.649**	0.993**	0.903**	0.768**
	(0.072)	(0.091)	(0.062)	(0.086)
$hskil_i$	1.092*	−1.129*	2.894**	−0.382
	(0.440)	(0.530)	(0.386)	(0.498)
$lskil_i$	−0.062	−0.054	−0.043	0.072
	(0.083)	(0.104)	(0.062)	(0.087)
$crplnd_i$	−0.003	0.043**	−0.010	0.005
	(0.011)	(0.014)	(0.009)	(0.013)
$border_i$	0.130*	−0.012	0.072	−0.146*
	(0.053)	(0.068)	(0.045)	(0.063)
air_i	0.003	0.018	0.077	0.019
	(0.086)	(0.108)	(0.073)	(0.102)
sez_i	0.025	−0.029	−0.014	−0.018
	(0.052)	(0.065)	(0.043)	(0.061)
$const$	−1.140**	−1.315**	−0.938**	−1.276**
	(0.255)	(0.322)	(0.190)	(0.267)
ρ	0.182**	0.124*	−0.008	0.095
	(0.051)	(0.054)	(0.047)	(0.057)
Instrumental Variables ($\mathbf{q}_{si}$)				
emp_{si}^{1998}	0.469**	0.470**	0.522**	0.522**
	(0.020)	(0.020)	(0.012)	(0.012)
$green_i^{2002}$	−0.168**	−0.153**	−0.116**	−0.115**
	(0.045)	(0.045)	(0.028)	(0.028)
$hskil_i^{1998}$	0.291	0.260	0.128	0.124
	(0.341)	(0.344)	(0.217)	(0.216)
Sargan	1.964	4.437	0.471	0.309
(p-value)	(0.375)	(0.109)	(0.790)	(0.857)
No. of obs.	1621	1621	1621	1621

Notes: Asterisks * (**) indicate that the 95% (99%) credible interval of the posterior distribution does not contain zero. Values in parenthesis are standard deviation for each posterior distribution.

Table 10: Robustness to Neighboring Informal Agglomeration

	Manufacturing		Wholesale/Retail	
	sal_{si}^{fml}	sal_{si}^{inf}	sal_{si}^{fml}	sal_{si}^{inf}
Explanatory Variables (x_i, $\mathbf{z}_i$)				
$(\mathbf{I}+\mathbf{W})agg_{si}^{inf}$	0.033	0.047**	0.017	0.038**
	(0.019)	(0.018)	(0.024)	(0.015)
pop_i	0.310**	0.521**	0.523**	0.327**
	(0.057)	(0.055)	(0.072)	(0.045)
$elec_i$	1.537**	0.542**	3.124**	0.889**
	(0.150)	(0.141)	(0.200)	(0.121)
$hskil_i$	1.303	−2.498**	3.942**	−1.689*
	(0.924)	(0.867)	(1.228)	(0.732)
$lskil_i$	0.118	0.345*	0.431*	0.189
	(0.160)	(0.153)	(0.193)	(0.119)
$crplnd_i$	−0.017	0.043	−0.039	0.001
	(0.023)	(0.022)	(0.030)	(0.018)
$border_i$	0.026	−0.119	0.252	0.012
	(0.113)	(0.108)	(0.147)	(0.090)
air_i	−0.314	0.023	−0.124	0.039
	(0.178)	(0.170)	(0.236)	(0.145)
sez_i	−0.007	0.024	−0.258	−0.021
	(0.107)	(0.103)	(0.142)	(0.087)
$const$	−2.637**	−2.822**	−4.411**	−0.004
	(0.482)	(0.469)	(0.566)	(0.386)
ρ	0.168**	0.138*	−0.085	0.100
	(0.049)	(0.059)	(0.053)	(0.059)
Instrumental Variables ($\mathbf{q}_{si}$)				
$(\mathbf{I}+\mathbf{W})emp_{si}^{1998}$	0.598**	0.598**	0.607**	0.607**
	(0.018)	(0.018)	(0.012)	(0.011)
$(\mathbf{I}+\mathbf{W})green_i^{2002}$	−0.249**	−0.247**	−0.203**	−0.198**
	(0.055)	(0.055)	(0.036)	(0.036)
$(\mathbf{I}+\mathbf{W})hskil_i^{1998}$	0.218	0.276	0.181	0.147
	(0.432)	(0.433)	(0.284)	(0.280)
Sargan	2.083	0.169	1.888	1.079
(p-value)	(0.353)	(0.919)	(0.389)	(0.583)
No. of obs.	1621	1621	1621	1621

Notes: Asterisks * (**) indicate that the 95% (99%) credible interval of the posterior distribution does not contain zero. Values in parentheses are standard deviations for each posterior distribution.

Table 11: Robustness to Alternative Spatial Weight Matrix

	Manufacturing			
	5 nearest neighbors		10 nearest neighbors	
	sal_{fml}	sal_{inf}	sal_{fml}	sal_{inf}
Explanatory Variables (x_i, $\mathbf{z}_i$)				
agg_{si}^{inf}	0.088**	0.133**	0.084*	0.131**
	(0.033)	(0.032)	(0.033)	(0.032)
pop_i	0.256**	0.441**	0.246**	0.421**
	(0.062)	(0.059)	(0.062)	(0.059)
$elec_i$	1.542**	0.543**	1.518**	0.553**
	(0.149)	(0.140)	(0.148)	(0.140)
$hskil_i$	1.387	−2.563**	0.988	−2.604**
	(0.883)	(0.824)	(0.886)	(0.822)
$lskil_i$	−0.040	0.118	−0.041	0.102
	(0.172)	(0.163)	(0.171)	(0.162)
$crplnd_i$	−0.004	0.059**	−0.004	0.056**
	(0.023)	(0.022)	(0.023)	(0.022)
$border_i$	0.037	−0.107	0.059	−0.095
	(0.112)	(0.106)	(0.112)	(0.106)
air_i	−0.384*	−0.024	−0.367*	−0.025
	(0.179)	(0.170)	(0.178)	(0.169)
sez_i	0.007	0.069	0.004	0.066
	(0.109)	(0.103)	(0.108)	(0.103)
$const$	−2.060**	−1.920**	−2.007**	−1.883**
	(0.538)	(0.509)	(0.535)	(0.508)
ρ	0.132**	0.107**	0.211**	0.166**
	(0.032)	(0.035)	(0.038)	(0.044)
Instrumental Variables ($\mathbf{q}_{si}$)				
emp_{si}^{1998}	0.500**	0.500**	0.500**	0.500**
	(0.022)	(0.022)	(0.022)	(0.022)
$green_i^{2002}$	−0.181**	−0.176**	−0.181**	−0.176**
	(0.048)	(0.048)	(0.048)	(0.048)
$hskil_i^{1998}$	0.281	0.330	0.287	0.328
	(0.366)	(0.368)	(0.367)	(0.369)
Sargan	1.252	0.094	1.425	0.071
(p-value)	(0.535)	(0.954)	(0.491)	(0.965)
No. of obs.	1621	1621	1621	1621

Notes: Asterisks * (**) indicate that the 95% (99%) credible interval of the posterior distribution does not contain zero. Values in parentheses are standard deviations for each posterior distribution.

Table 12: Robustness to Additional Control Variable

	Manufacturing		Wholesale/Retail	
	sal_{si}^{inf}	wge_{si}^{inf}	sal_{si}^{inf}	wge_{si}^{inf}
Explanation Variables (x_i, $\mathbf{z}_i$)				
agg_{si}^{inf}	0.130**	0.069**	0.141**	0.062**
	(0.031)	(0.020)	(0.025)	(0.017)
pop_i	0.434**	0.165**	0.208**	0.143**
	(0.059)	(0.037)	(0.047)	(0.032)
$elec_i$	0.473**	0.834**	0.560**	0.524**
	(0.147)	(0.093)	(0.132)	(0.091)
$hskil_i$	-2.673**	-1.070*	-2.085**	-0.308
	(0.823)	(0.520)	(0.702)	(0.495)
$lskil_i$	0.126	-0.035	-0.072	0.058
	(0.162)	(0.102)	(0.124)	(0.085)
$crplnd_i$	0.056*	0.039**	0.023	0.005
	(0.022)	(0.014)	(0.019)	(0.013)
$border_i$	-0.097	-0.013	0.020	-0.162**
	(0.107)	(0.067)	(0.090)	(0.062)
air_i	-0.015	0.040	-0.003	0.038
	(0.172)	(0.107)	(0.147)	(0.101)
sez_i	0.064	-0.034	0.013	-0.009
	(0.104)	(0.065)	(0.088)	(0.060)
dum_{si}^{fml}	0.088	0.206**	0.291**	0.308**
	(0.073)	(0.046)	(0.063)	(0.044)
$const$	-2.009**	-1.166**	0.921*	-1.082**
	(0.506)	(0.316)	(0.394)	(0.256)
ρ	0.164**	0.106*	0.137**	0.111*
	(0.054)	(0.054)	(0.051)	(0.057)
Instrumental Variables ($\mathbf{q}_{si}$)				
emp_{si}^{1998}	0.505**	0.506**	0.526**	0.526**
	(0.022)	(0.022)	(0.013)	(0.013)
$green_i^{2002}$	-0.177**	-0.175**	-0.109**	-0.113**
	(0.048)	(0.048)	(0.029)	(0.030)
$hskil_i^{1998}$	0.335	0.312	0.144	0.140
	(0.368)	(0.370)	(0.220)	(0.227)
Sargan	0.067	4.795	0.463	0.197
(p-value)	(0.967)	(0.091)	(0.793)	(0.906)
No. of obs.	1621	1621	1621	1621

Notes: Asterisks * (**) indicate that the 95% (99%) credible interval of the posterior distribution does not contain zero. Values in parenthesis are standard deviation for each posterior distribution.

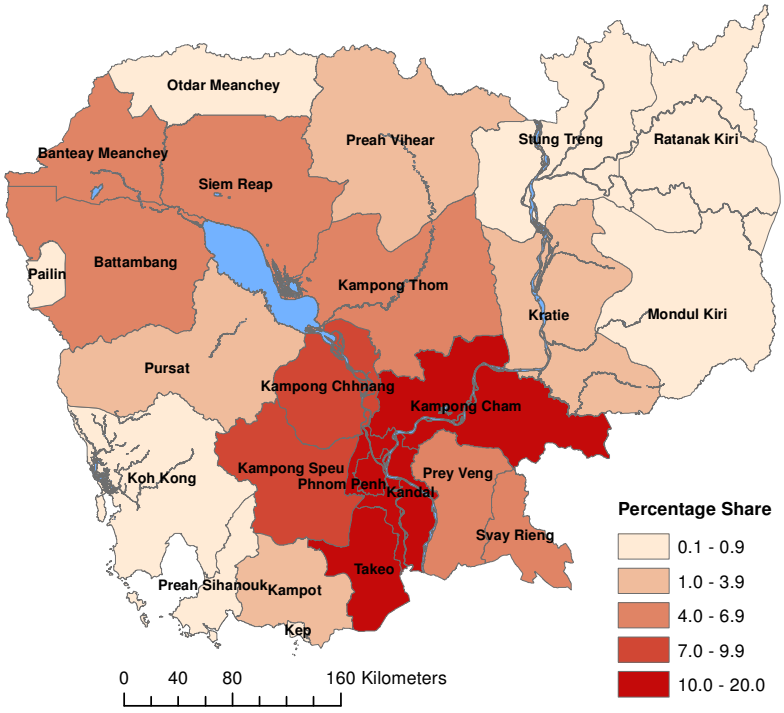

Figure 1: Provincial Shares of Informal Employment (Manufacturing)

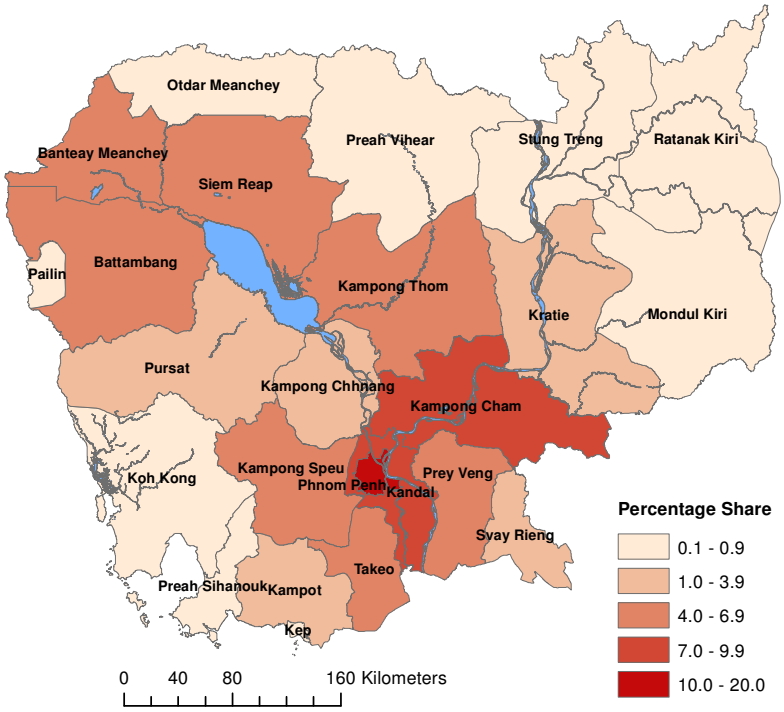

Figure 2: Provincial Shares of Informal Employment (Wholesale/Retail)

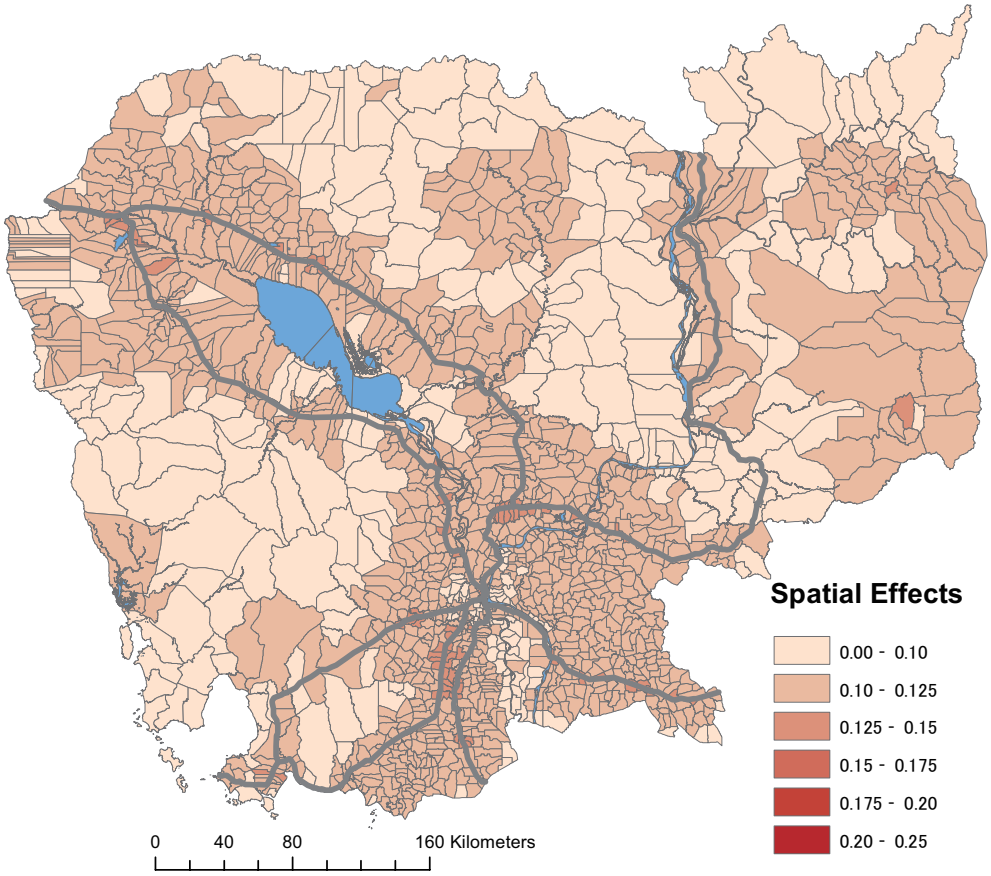

Figure 3: Spatial Effects on Manufacturing Formal Firms

Notes: The map shows the total impacts on sales per worker in the formal sector in each commune resulting from a uniform increase in the density of informal activity across communes; one-digit national roads are shown in the map as gray lines.

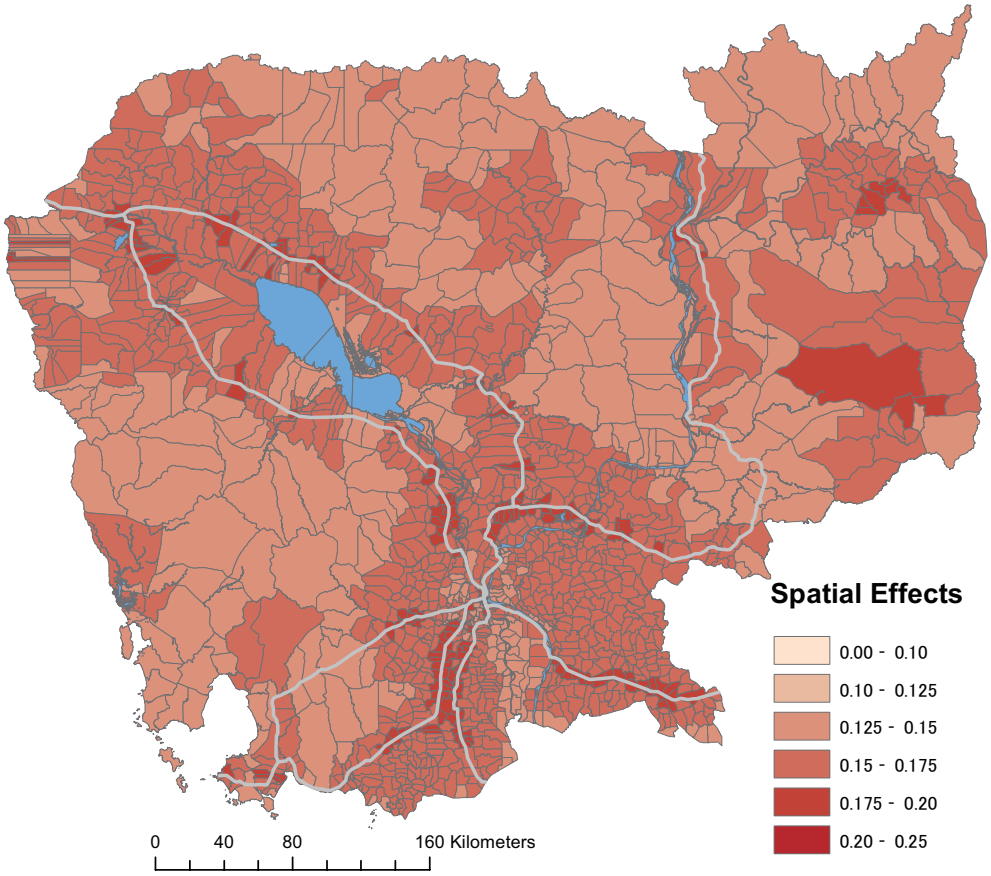

Figure 4: Spatial Effects on Manufacturing Informal Firms

Notes: The map shows the total impacts on sales per worker in the informal sector in each commune resulting from a uniform increase in the density of informal activity across communes; one-digit national roads are shown in the map as light gray lines.

Appendix A: Full Conditional Posterior Density

This Appendix describes how to generate samples from the full conditional posterior distribution.

Full conditional posterior of $\boldsymbol{\beta}^*$

Given the parameter $\boldsymbol{\gamma}^*$, $\boldsymbol{\eta}$ can be observed, and consequently, Equation (6) conditional on $\boldsymbol{\eta}$ is

$$\mathbf{Sy} = \mathbf{x}\beta_0 + \mathbf{Z}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \mid \boldsymbol{\eta}, \tag{15}$$

and the expectation and variance of $\boldsymbol{\varepsilon} \mid \boldsymbol{\eta}$ are

$$\begin{aligned} \mathrm{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{\eta}) \equiv \boldsymbol{\mu}_{\varepsilon|\eta} &= \mathrm{E}(\boldsymbol{\varepsilon}) + (\sigma_{12}\mathbf{I}_n)(\sigma_{11}\mathbf{I}_n)^{-1}(\boldsymbol{\eta} - \mathrm{E}(\boldsymbol{\eta})) \\ &= (\sigma_{12}\mathbf{I}_n)(\sigma_{11}\mathbf{I}_n)^{-1}\boldsymbol{\eta} \\ \mathrm{V}(\boldsymbol{\varepsilon} \mid \boldsymbol{\eta}) \equiv \boldsymbol{\Sigma}_{\varepsilon|\eta} &= (\sigma_{22}\mathbf{I}_n) - (\sigma_{12}^2\mathbf{I}_n)(\sigma_{11}\mathbf{I}_n)^{-1}, \end{aligned} \tag{16}$$

where σ_{ij} is the (i, j)th element of $\boldsymbol{\Sigma}$. Equation (15) can be rewritten as

$$\begin{aligned} \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}\left(\mathbf{Sy} - \boldsymbol{\mu}_{\varepsilon|\eta}\right) &= \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}\begin{bmatrix}\mathbf{x} & \mathbf{Z}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \boldsymbol{\beta}_1\end{bmatrix} + \boldsymbol{\xi}, \\ \mathbf{y}^* &= \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\xi}, \end{aligned} \tag{17}$$

where $\boldsymbol{\xi} \sim MVN(\mathbf{0}, \mathbf{I}_n)$, $\mathbf{y}^* \equiv \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}(\mathbf{Sy} - \boldsymbol{\mu}_{\boldsymbol{\varepsilon}})$, and $\mathbf{X}^* \equiv \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}[\mathbf{x}, \mathbf{Z}]$. Using Equation (17) and the prior of $\boldsymbol{\beta}^*$ yields the *full* conditional multivariate normal distribution of $\boldsymbol{\beta}^*$:

$$\begin{aligned} \boldsymbol{\beta}^* \mid \boldsymbol{\gamma}^*, \rho, \boldsymbol{\Sigma}, \text{Data} &\sim MVN(\hat{\mathbf{b}}_\beta, \hat{\mathbf{B}}_\beta) \\ \hat{\mathbf{B}}_\beta &= \left[\mathbf{X}^{*\prime}\mathbf{X}^* + \mathbf{B}_\beta^{-1}\right]^{-1} \\ \hat{\mathbf{b}}_\beta &= \hat{\mathbf{B}}_\beta\left[\mathbf{X}^{*\prime}\mathbf{y}^* + \mathbf{B}_\beta^{-1}\mathbf{b}_\beta\right]. \end{aligned} \tag{18}$$

Full conditional posterior of $\boldsymbol{\gamma}^*$

Substituting Equation (5) for (6) and reformulating, we can obtain:

$$\begin{pmatrix}\mathbf{x} \\ \frac{\mathbf{Sy} - \mathbf{Z}\boldsymbol{\beta}_1}{\beta_0}\end{pmatrix} = \begin{pmatrix}\mathbf{Q} & \mathbf{Z} \\ \mathbf{Q} & \mathbf{Z}\end{pmatrix}\boldsymbol{\gamma}^* + \begin{pmatrix}\boldsymbol{\eta} \\ \boldsymbol{\eta} + \frac{\boldsymbol{\varepsilon}}{\beta_0}\end{pmatrix}. \tag{19}$$

The covariance matrix of $\boldsymbol{\eta}^*$ is

$$\begin{aligned} \mathrm{V}\begin{pmatrix}\boldsymbol{\eta} \\ \boldsymbol{\eta} + \boldsymbol{\varepsilon}/\beta_0\end{pmatrix} \equiv \boldsymbol{\Omega} &= [\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'] \otimes \mathbf{I}_n, \\ \mathbf{A} &\equiv \begin{bmatrix}1 & 0 \\ 1 & 1/\beta_0\end{bmatrix}. \end{aligned} \tag{20}$$

Multiplying both sides of Equation (19) by $\boldsymbol{\Omega}^{-1/2}$ yields

$$\boldsymbol{\Omega}^{-\frac{1}{2}}\begin{pmatrix}\mathbf{x} \\ \frac{\mathbf{Sy} - \mathbf{Z}\boldsymbol{\beta}_1}{\beta_0}\end{pmatrix} = \boldsymbol{\Omega}^{-\frac{1}{2}}\begin{pmatrix}\mathbf{Q} & \mathbf{Z} \\ \mathbf{Q} & \mathbf{Z}\end{pmatrix}\boldsymbol{\gamma}^* + \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim MVN(\mathbf{0}, \mathbf{I}_{2n}). \tag{21}$$

Using this equation and the prior of γ^*, we can obtain the *full* conditional multivariate normal distribution:

$$\begin{aligned} \boldsymbol{\gamma}^* \mid \boldsymbol{\beta}^*, \rho, \boldsymbol{\Sigma}, \text{Data} &\sim MVN(\hat{\mathbf{b}}_\gamma, \hat{\mathbf{B}}_\gamma), \\ \hat{\mathbf{B}}_\gamma &= \left[\mathbf{Z}^{*\prime}\mathbf{Z}^* + \mathbf{B}_\gamma^{-1}\right]^{-1} \\ \hat{\mathbf{b}}_\gamma &= \hat{\mathbf{B}}_\gamma \left[\mathbf{Z}^{*\prime}\mathbf{y}^+ + \mathbf{B}_\gamma^{-1}\mathbf{b}_\gamma\right], \end{aligned} \tag{22}$$

where

$$\mathbf{y}^+ \equiv \boldsymbol{\Omega}^{-\frac{1}{2}} \begin{pmatrix} \mathbf{x} \\ \frac{\mathbf{S}\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}_1}{\beta_0} \end{pmatrix}, \quad \mathbf{Z}^* \equiv \boldsymbol{\Omega}^{-\frac{1}{2}} \begin{pmatrix} \mathbf{Q} & \mathbf{Z} \\ \mathbf{Q} & \mathbf{Z} \end{pmatrix}.$$

Full conditional posterior of $\boldsymbol{\Sigma}$

The full conditional posterior of $\boldsymbol{\Sigma}$ has an inverted Wishart distribution:

$$\begin{aligned} \boldsymbol{\Sigma} \mid \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \rho, \text{Data} &\sim IW(\hat{b}_\Sigma, \hat{\mathbf{B}}_\Sigma) \\ \hat{b}_\Sigma &= n + b_\Sigma \\ \hat{\mathbf{B}}_\Sigma &= \left[\mathbf{E} + \mathbf{B}_\Sigma^{-1}\right]^{-1}, \end{aligned} \tag{23}$$

where

$$\mathbf{E} \equiv \begin{bmatrix} \boldsymbol{\eta}' \\ \boldsymbol{\varepsilon}' \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta} & \boldsymbol{\varepsilon} \end{bmatrix},$$

$\boldsymbol{\eta} = \mathbf{x} - \mathbf{Q}\boldsymbol{\gamma}_0 - \mathbf{Z}\boldsymbol{\gamma}_1$, and $\boldsymbol{\varepsilon} = \mathbf{S}\mathbf{y} - \mathbf{x}\beta_0 - \mathbf{Z}\boldsymbol{\beta}_1$.

Full conditional posterior of ρ

We can reformulate Equation (17) as

$$\begin{aligned} \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}\left(\mathbf{y} - \begin{bmatrix} \mathbf{x} & \mathbf{Z} \end{bmatrix}\boldsymbol{\gamma}^* - \boldsymbol{\mu}_{\varepsilon|\eta}\right) &= \rho\,\boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}\mathbf{W}\,\mathbf{y} + \boldsymbol{\xi}, \\ \tilde{\mathbf{y}} &= \rho\,\tilde{\mathbf{X}} + \boldsymbol{\xi}, \end{aligned} \tag{24}$$

where $\tilde{\mathbf{y}} \equiv \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}(\mathbf{y} - [\mathbf{x}, \mathbf{Z}]\,\boldsymbol{\gamma}^* - \boldsymbol{\mu}_{\varepsilon|\eta})$, and $\tilde{\mathbf{X}} \equiv \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}\mathbf{W}\,\mathbf{y}$. Then, the full conditional posterior density function of ρ can be obtained as

$$\begin{aligned} P(\rho \mid \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\Sigma}, \text{Data}) &\propto \mid \mathbf{I}_n - \rho\,\mathbf{W} \mid \exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}} - \rho\tilde{\mathbf{X}}\right]'\left[\tilde{\mathbf{y}} - \rho\tilde{\mathbf{X}}\right]\right\} I[\rho \in (\lambda_{\min}^{-1}, \lambda_{\max}^{-1})] \\ &\propto \mid \mathbf{I}_n - \rho\,\mathbf{W} \mid \exp\left\{-\frac{1}{2\,\hat{\sigma}_\rho^2}(\rho - \hat{\rho})^2\right\} I[\rho \in (\lambda_{\min}^{-1}, \lambda_{\max}^{-1})], \end{aligned} \tag{25}$$

where $\hat{\sigma}_\rho^2 = \left[\tilde{\mathbf{X}}'\,\tilde{\mathbf{X}}\right]^{-1}$ and $\hat{\rho} = \hat{\sigma}_\rho^2\,\tilde{\mathbf{X}}'\,\tilde{\mathbf{y}}$. Here, $I[\rho \in (\lambda_{\min}^{-1}, \lambda_{\max}^{-1})]$ is an indicator function equal to 1 exactly when $\rho \in (\lambda_{\min}^{-1}, \lambda_{\max}^{-1})$. Since this density function is not standard, we use the Metropolis–Hastings (MH) technique.[16] The candidate generating function used in the MH algorithm is $TN(\hat{\rho}, \hat{\sigma}_\rho^2)$, which is a normal distribution truncated on the interval $(\lambda_{\min}^{-1}, \lambda_{\max}^{-1})$, for which the mean is $\hat{\rho}$ and the variance is $\hat{\sigma}_\rho^2$.

[16] For more details about the Metropolis–Hastings and Gibbs sampling techniques, refer to Gamerman and Lopes (2006, Chapters 5 and 6).

MCMC sampling algorithm

Now we describe the MCMC sampling algorithm for our model.

MCMC sampling algorithm
(i) Choose arbitrary initial values for all parameters and initialize a counter $r = 1$.
(ii) Repeat the following steps:
Draw $\boldsymbol{\beta}^{*(r)}$ from $MVN(\hat{\mathbf{b}}_\beta, \hat{\mathbf{B}}_\beta)$, given $\boldsymbol{\gamma}^{*(r-1)}, \boldsymbol{\Sigma}^{(r-1)}, \rho^{(r-1)}$, Data.
Draw $\boldsymbol{\gamma}^{*(r)}$ from $MVN(\hat{\mathbf{b}}_\gamma, \hat{\mathbf{B}}_\gamma)$, given $\boldsymbol{\beta}^{*(r)}, \boldsymbol{\Sigma}^{(r-1)}, \rho^{(r-1)}$, Data.
Draw $\boldsymbol{\Sigma}^{(r)}$ from $IW(\hat{b}_\Sigma, \hat{\mathbf{B}}_\Sigma)$, given $\boldsymbol{\beta}^{*(r)}, \boldsymbol{\gamma}^{*(r)}, \rho^{(r-1)}$, Data.
Draw ρ' (a candidate of $\rho^{(r)}$) from $TN(\hat{\rho}, \hat{\sigma}^2_\rho)$, given $\boldsymbol{\beta}^{*(r)}, \boldsymbol{\gamma}^{*(r)}, \boldsymbol{\Sigma}^{(r)}$, Data.
Calculate an acceptance probability:

$$\alpha(\rho', \rho^{(r-1)}) = \min\left\{1, \frac{|\mathbf{I}_n - \rho' \mathbf{W}|}{|\mathbf{I}_n - \rho^{(r-1)} \mathbf{W}|}\right\}.$$

Set $\rho^{(r)} = \rho'$ with probability $\alpha(\rho', \rho^{(r-1)})$, and set $\rho^{(r)} = \rho^{(r-1)}$ with probability $1-\alpha(\rho', \rho^{(r-1)})$.
If $r < M$, set $r = r + 1$ and return to (ii). Otherwise, go to (iii).[17]
(iii) Discard the samples with the superscript $r = 1, 2, \ldots, M_0$, and save the samples with $r = M_0 + 1, M_0 + 2, \ldots, M$.

In this paper, we set $M = 30,000$ and $M_0 = 10,000$, so that 20,000 replications are retained and used for the posterior inference.

Appendix B: Monte Carlo Simulation

An identification problem arises in IV methods when instruments are 'weak.' As instruments become weaker, we approach an unidentified case (Rossi et al., 2005). To investigate the estimation performance of our model with weak instruments, we consider two cases: *strong instruments* and *weak instruments*.

The data for the two cases are all generated from

$$\begin{aligned} x_i &= q_{1i}\gamma_{01} + q_{2i}\gamma_{02} + q_{2i}\gamma_{03} + z_{1i}\gamma_{11} + z_{2i}\gamma_{12} + z_{3i}\gamma_{13} + \eta_i \\ y_i &= x_i\beta_0 + z_{1i}\beta_{11} + z_{2i}\beta_{12} + z_{3i}\beta_{13} + \rho\sum_{j=1}^{n} w_{ij}y_j + \varepsilon_i \\ \begin{pmatrix}\eta_i \\ \varepsilon_i\end{pmatrix} &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{26}$$

and $z_{1i} = 1$ while z_{2i} and z_{3i} follow a standard normal distribution N(0, 1). The element w_{ij} is constructed for Cambodian commune-level regions, so the sample size of the artificial dataset is the same as the number of Cambodian communes. The parameter settings of the datasets are as follows.

[17] Since it is infeasible to directly calculate the determinant $|\mathbf{I}_n - \rho\mathbf{W}|$ given the large size of our spatial matrix $\mathbf{W}$ (1621×1621), we use the approximation given by Ord (1975, p. 121):

$$|\mathbf{I}_n - \rho\mathbf{W}| = \prod_{i=1}^{n}(1 - \rho\lambda_i),$$

where $\lambda_1, \ldots, \lambda_n$ are real eigenvalues of $\mathbf{W}$.

Data set 1: *with 'strong instruments'*

$$\boldsymbol{\gamma}_0 = \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \gamma_{03} \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}.$$

Data set 2: *with 'weak instruments'*

$$\boldsymbol{\gamma}_0 = \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \gamma_{03} \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}.$$

All other parameters are the same between datasets 1 and 2:

$$\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1\,,$$
$$\gamma_{11} = \gamma_{12} = \gamma_{13} = 1\,,$$
$$\rho = 0.5\,,$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

We analyze the difference of estimation performance between the two cases. Relatively diffuse priors, such as $N(0,\ 100)$, are used for coefficients and for the covariance matrix, $\boldsymbol{\Sigma} \sim IW(2, \mathbf{I}_2)$. The prior of ρ is $U(1/\lambda_{\min}, 1/\lambda_{\max})$.

Table B1 shows the results of simulation analysis. For both cases, the median is close to the true value, and the 95% credible intervals of the structural parameters contain their true value. However, the posterior distributions of the weak instruments cases have a higher dispersion for the coefficients of the structural parameters, indicating the difficulty of identification. Figures B1 and B2 also show the large dispersion of the posteriors in the weak instruments case. In addition, Figure B1 shows that MCMC sampler for weak instruments has a higher autocorrelation, implying that more MCMC draws are needed to obtain a precise posterior distribution.

[– Table B1–]
[– Figures B1 and B2–]

Appendix C: Traveling Time between Commune Pairs

A spatial-weighting matrix is supposed to capture the degree of connectivity, which is relevant for economic activity, between all pairs of distinct regions. For a given pair of regions, greater connectivity should be associated with stronger influences of economic activity on each other. Conversely, weaker connectivity should be associated with smaller influences. In this paper, we measure connectivity by the shortest traveling time between communes. A shorter time to travel from one commune to another commune points to the greater connectivity between these communes. In contrast, a longer time to move between communes suggests weaker connectivity. Thus, we need precise estimates of the traveling time for the shortest path among all possible routes to construct the spatial-weighting matrix.

The shortest traveling time is calculated by the Floyd–Warshall algorithm. To execute this method, we first prepare a dataset consisting of the geographic distances between all commune

pairs. The GIS shape file of administrative units in Cambodia enables us to obtain the shortest distance between these communes. Next, we prepare a dataset consisting of the traveling speeds between *neighboring* communes. We exploit the data sources cited in the text to make reasonable assumptions about how fast we can move between neighboring communes. Specifically, the JETRO survey documents, which are based on a field survey of the actual status of logistics infrastructures, detail traveling speeds for a number of transportation routes within Cambodia. Table A1 summarizes our assumptions about traveling speed. For instance, if a 1-digit national road crosses both communes in contiguity, we assume a crossing speed of 90 kilometers per hour.[18] If the 1-digit national road crosses one commune with the 2-digit national road crossing neighboring commune, we assume a speed of 70 kilometers per hour. Additionally, if neighboring communes are characterized by tree covered or regularly flooded areas, we assume a speed of 4 kilometers per hour for crossing. Finally, we assume 30 kilometers per hour for all the other pairs of neighboring communes.

Given the above assumptions, we can calculate the traveling time for all the neighboring communes. Because each commune must be connected to all the other communes at least indirectly through its neighboring communes, this dataset allows us to compute traveling times for a significant number of possible routes. For this task, we employ the Floyd-Warshall algorithm, which is executed in the following steps. First, for a given pair of communes i and j, we compute traveling time from commune i to intermediate route commune k, T_{ik}, and that from commune k to commune j, T_{kj}. This gives us the traveling time from commune i to commune j via commune k, $T_{ik} + T_{kj}$. We then compare traveling times between direct and indirect routes. If $T_{ik} + T_{kj} < T_{ij}$, we replace the shortest path $T_{ij} = T_{ik} + T_{kj}$. Otherwise, we keep the original path T_{ij}. To enable computation, we initially set the traveling time to 10,000 hours for missing observations of commune pairs in the dataset on traveling time. We repeat this recursive algorithm for $k = 1, 2, \ldots, N$, where N is the total number of communes.

The above procedure gives us a sample with more than 1 million observations of the shortest traveling time between communes. The average geographic distance among the 1,621 communes is 189.8 kilometers with a standard deviation of 109.5, and the average traveling time is 6.72 hours with a standard deviation of 4.8. The correlation coefficient of time and distance is 0.61, implying that traveling time and geographic distance are positively correlated, but are not necessarily in a perfectly linear relationship. Thus, it is important to take into account traveling speed in constructing a spatial-weighting matrix based on geographic distance.

[– Table C1 –]

[18] Roughly speaking, the 1-digit national roads radiate in all directions from Phnom Penh, the capital of Cambodia, which is located at the geographic center of the territory, reaching out to national borders and the ocean. The 2-digit national roads cross provinces.

–Appendix Tables and Figures–

Table B1: Results of Simulation Analysis

Strong instruments: $\gamma_0 = [4, 4, 4]'$

	True value	95%L	Median	95%U	Std. Dev.
β_0	1	0.9869	0.9944	1.0017	0.0037
β_1	1	0.7367	0.9563	1.1573	0.1062
β_2	1	0.9652	1.0155	1.0661	0.0258
β_3	1	0.9805	1.0317	1.0849	0.0266
ρ	0.5	0.4586	0.5070	0.5595	0.0259
σ_{11}	1	0.9806	1.0502	1.1281	0.0377
σ_{12}	0.8	0.7794	0.8432	0.9127	0.0341
σ_{22}	1	0.9632	1.0334	1.1101	0.0375

Weak instruments: $\gamma_0 = [0.1, 0.1, 0.1]'$

	True value	95%L	Median	95%U	Std. Dev.
β_0	1	0.5147	0.8358	1.0914	0.1476
β_1	1	0.4850	1.0611	1.7271	0.3195
β_2	1	0.9087	1.1804	1.5179	0.1536
β_3	1	0.9294	1.1908	1.5193	0.1516
ρ	0.5	0.3705	0.5195	0.6791	0.0749
σ_{11}	1	0.9804	1.0497	1.1272	0.0376
σ_{12}	0.8	0.7376	1.0114	1.3557	0.1588
σ_{22}	1	0.8806	1.3305	2.0852	0.3109

Note: 95%L and 95%U are the lower and upper bounds of the 95% credible interval.

Table C1: Assumptions on Traveling Speed in Kilometers per Hour

Commune Characteristics	One digit national road	Two digit national road	Congested hub (i.e., Phnom Penh area)	Tree covered, regularly flooded area
One digit national road	90			
Two digit national road	70	50		
Congested hub (i.e., Phnom Penh area)	62.5	42.5	35	
Tree covered, regularly flooded area	47	27	19.5	4

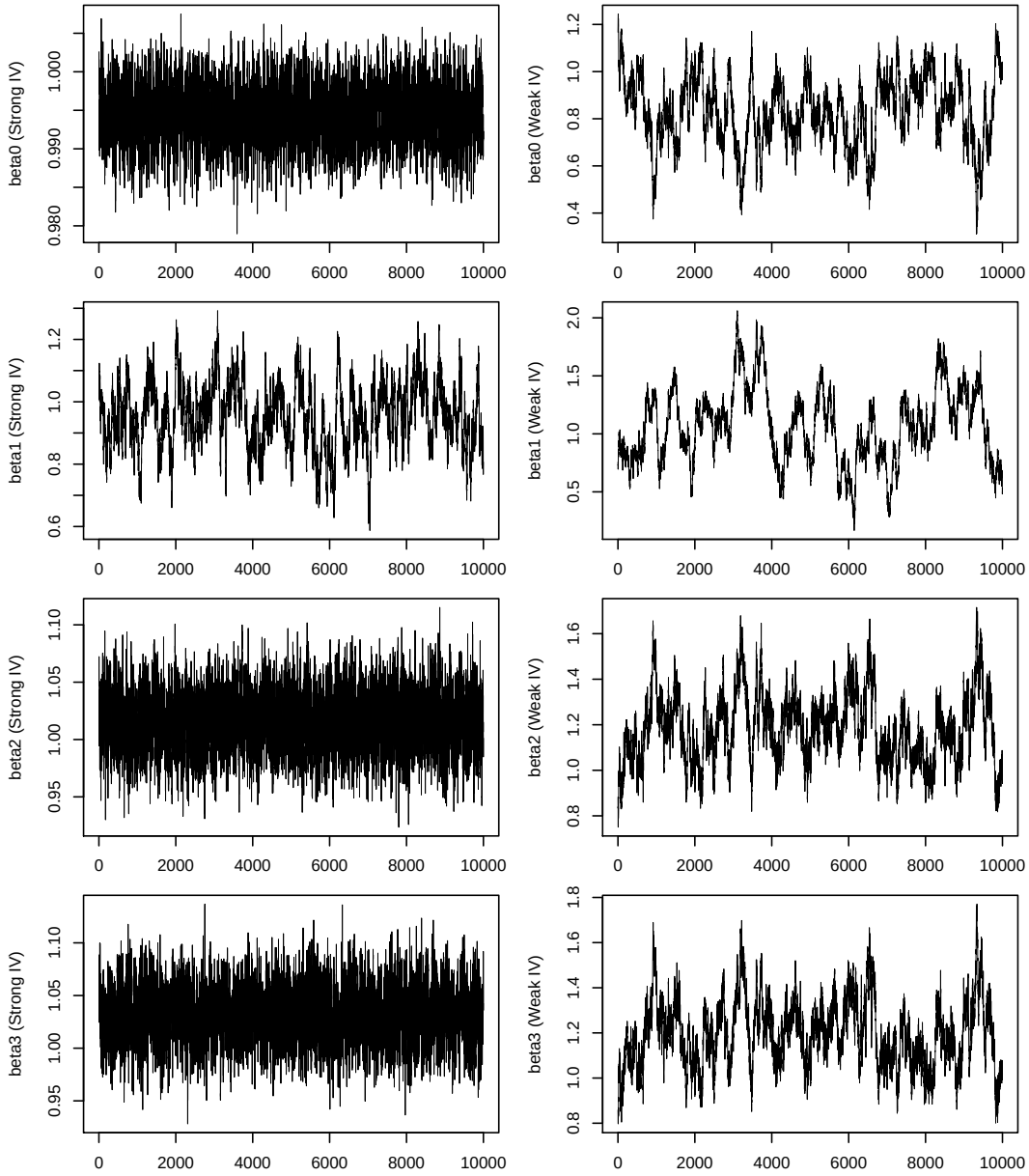

Figure B1: MCMC Sampling Path

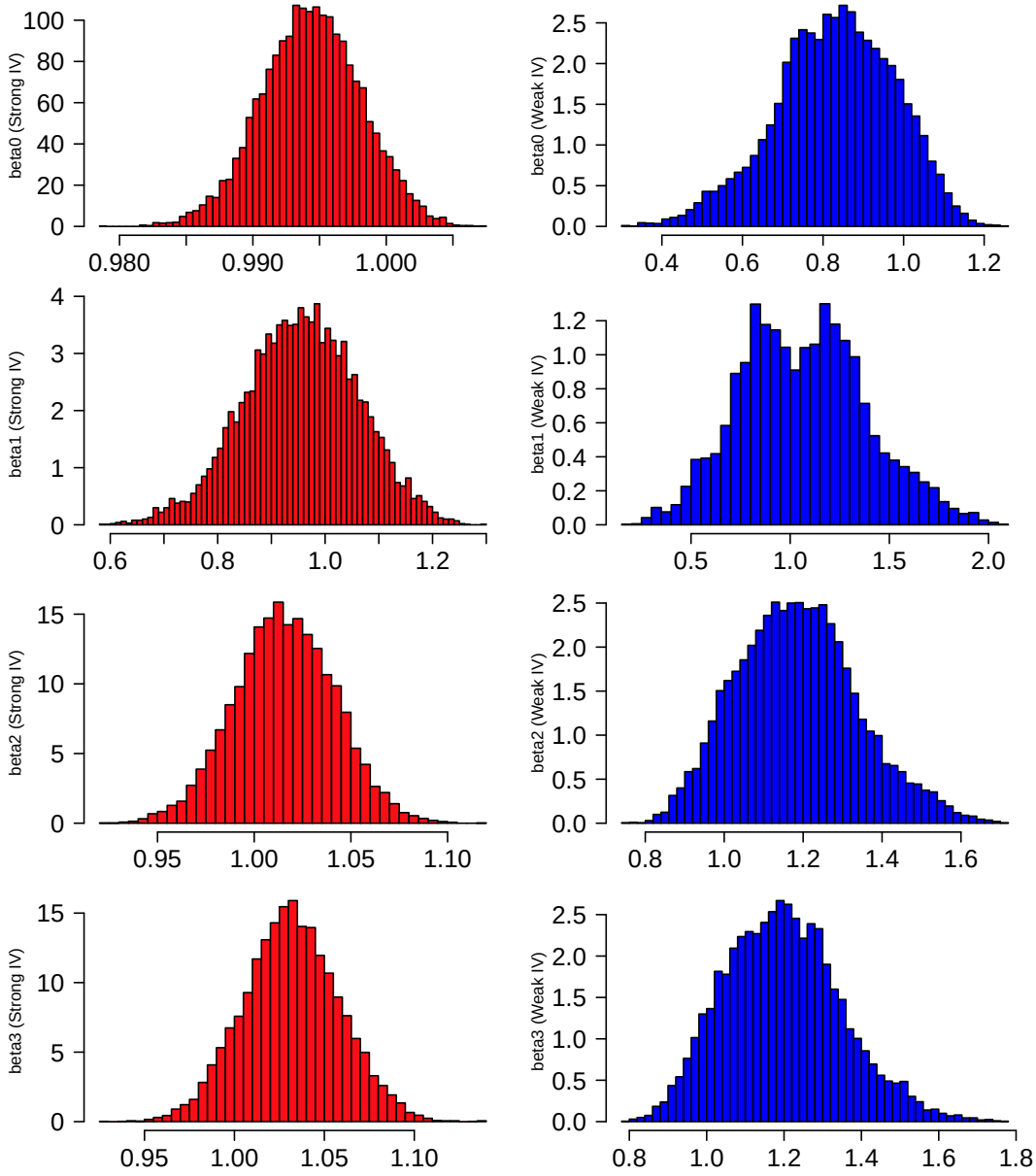

Figure B2: Histograms of MCMC Samples