Agglomeration economies in the formal and informal sectors : a Bayesian spatial approach

INSTITUTE OF DEVELOPING ECONOMIES

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

IDE DISCUSSION PAPER No. 666

Agglomeration Economies in the Formal and Informal Sectors: A Bayesian Spatial Approach

Kiyoyasu TANAKA* and Yoshihiro HASHIGUCHI

May 2017

Abstract

This paper examines whether localized clusters of similar industries produce agglomeration economies in the formal and informal sectors. We develop a Bayesian method to estimate a spatial autoregressive model with an endogenous independent variable. We use establishment-level census data that cover both formal registered and informal unregistered establishments in Cambodia. We find that the density of local employment has a significantly positive effect on productivity in the informal sector, but little effect in the formal sector. For manufacturing, a doubling of employment density increases productivity in the informal sector by 9% through local linkages and by 17% through spatial multiplier linkages, leading to a 26% increase in total. A spatial network magnifies the local impact of agglomeration economies in the informal sector.

Keywords: Agglomeration economies, Informal sector, Cambodia, Bayes
JEL classification: C11, C21, C26, H26, O17, R12

* Research Fellow, Institute of Developing Economies; Visiting Professor, Royal University of Phnom Penh (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 Economies in the Formal and Informal Sectors: A Bayesian Spatial Approach[†]

Kiyoyasu Tanaka[§]
(Institute of Developing Economies)

Yoshihiro Hashiguchi[ζ]
(Institute of Developing Economies)

April 2017

Abstract
This paper examines whether localized clusters of similar industries produce agglomeration economies in the formal and informal sectors. We develop a Bayesian method to estimate a spatial autoregressive model with an endogenous independent variable. We use establishment-level census data that cover both formal registered and informal unregistered establishments in Cambodia. We find that the density of local employment has a significantly positive effect on productivity in the informal sector, but little effect in the formal sector. For manufacturing, a doubling of employment density increases productivity in the informal sector by 9% through local linkages and by 17% through spatial multiplier linkages, leading to a 26% increase in total. A spatial network magnifies the local impact of agglomeration economies in the informal sector.

Keywords: Agglomeration economies, Informal sector, Cambodia, Bayes
JEL classification: C11, C21, C26, H26, O17, R12

[†] This paper is based on 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) Grant Number 16K17129. We would like to thank Fumihiko Nishi and Souknilanh Keola for data assistance. For useful comments and suggestions, we thank Kentaro Nakajima and seminar participants at 2014 JEA conference, Kyoto University, and 2015 CAED conference. All remaining errors are our own.
[§] Corresponding Author: Research fellow, Institute of Developing Economies, JETRO; address: 3-2-2 Wakaba, Mihama-ku, Chiba-shi, Chiba 261-8545, Japan; e-mail: kiyoyasu_tanaka@ide.go.jp
[ζ] Research fellow, Institute of Developing Economies, JETRO; address: 3-2-2 Wakaba, Mihama-ku, Chiba-shi, Chiba 261-8545, Japan; e-mail: yoshihiro_hashiguchi@ide.go.jp

1. Introduction

A spatial concentration of industrial production has crucial implications for developing economies. Firms and workers in an agglomerated area can benefit from agglomeration externalities through efficient sharing of local suppliers, better matching between employers and workers, and knowledge spillovers (Duranton and Puga, 2004). Limited resources such as capital, human capital, and infrastructure can be most efficiently utilized in an agglomerated area. Consequently, industrial agglomeration can promote economic growth at an early stage of development (Williamson, 1965; Fujita and Thisse, 2003). Thus, agglomeration economies provide a justification for public policies to promote industrial clusters in developing economies.

However, it is an unsettled question as to whether industrial agglomeration produces similar benefits for low income economies, as has been previously demonstrated for high and middle income countries (Rosenthal and Strange, 2004; Melo et al., 2009). In developing economies, self-employment and small enterprises provide major employment opportunities, but often do not formally register with the government. These informal enterprises are less productive and profitable than formally registered enterprises (McKenzie and Seynabou Sakho, 2010; Fajnzylber et al, 2011). Given the coexistence of formal and informal firms, 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." Therefore, whether industrial agglomeration produces productivity gains in both formal and informal sectors is a crucial empirical question.

In this paper, we estimate the long-run magnitude of agglomeration economies produced by a localized concentration of similar industries in Cambodia - a developing economy with the large informal sector. A localized cluster of similar industries can produce both positive externalities and congestion costs for producers. If positive externalities exceed congestion costs, regional productivity should increase with the density of local employment in similar industries, and vice versa. Moreover, industrial clusters form a spatial network among firms and workers through the spatial diffusion of agglomeration externalities. The spatial interdependence of productivity may magnify the local impact of localization economies through a spatial multiplier effect.

We specify a spatial autoregressive model with an endogenous independent variable to identify the causal effect of localization economies in the formal and informal sectors. We address an endogeneity problem in agglomeration by exploiting past data on the density of employment and forest area as instruments for the current density of employment. We account for a spatial network among firms by a spatial lag variable of productivity, which allows us to estimate local and spatial magnitudes of localization economies. To estimate our model, we develop a Bayesian method by extending the Bayesian instrumental variables (IV) method proposed by Rossi et al. (2005).

We exploit a unique dataset based on the Economic Census of Cambodia in 2011 (EC2011). This census covers all 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 activity is a commonly used definition of informality, and business registration is an objective criterion to classify formal and informal economic activities (Schneider and Enste, 2013). The dataset allows us to estimate agglomeration economies in both formal and informal sectors. Additionally,

Cambodia provides an interesting setting for our analysis. The Cambodian economy was devastated by the Pol Pot regime for 1975-79 and the subsequent civil war. The Paris Conference on Cambodia in 1991 led to agreements on the comprehensive political settlement of the Cambodia conflict. While the economy experienced a rapid economic growth thereafter, per capita GDP reached only 931.2 USD in 2012 (IMF, 2012). Schneider et al. (2010) estimate that informal activity accounted for 48.7% of GDP in Cambodia on average for the period 1999-2007.

The main findings can be summarized as follows. The density of local employment has a significantly positive effect on productivity in the informal sector. Since the validity of our instruments is supported, we interpret the positive coefficient of employment density as reflecting a causal effect. By contrast, we find little evidence of localization economies in the formal sector. These contrasting results are found in both manufacturing and wholesale/retail industries. Additionally, a spatial lag of productivity is significantly positive in the informal sector, suggesting that the positive spatial interdependence magnifies the positive localization effects. Using the main result for manufacturing, we find that a doubling of the density of local employment in a region increases productivity in the informal sector by 9% through direct linkages and by 17% through spatial multiplier linkages, leading to a 26% increase in total. Thus, the spatial network is crucial for estimating the precise magnitude of localization economies.

This paper contributes to the related literature in several ways. Livingstone (1991) documents a spatial concentration of manufacturing microenterprises in developing economies, which are interpreted as evidence of externalities perceived by small producers. Drawing on case studies on industrial clusters, Schmitz (1995) and Schmitz and Nadvi (1999) examine how clustering promotes the growth and competitiveness of small manufactures. In addition to local external economies from clustering in Marshall (1890), these papers highlight active joint action among firms in clusters. From a historical perspective, Zeitlin (2008) and Hashino and Otsuka (2013) explore long-term processes of industrial clusters and emphasize cooperative efforts among producers and local governments in clusters. On the other hand, Moreno-Monroy (2012) argues that small enterprises also exist in non-cluster areas and the case studies on *successful* industrial clusters may represent exceptions to the usual experience of small enterprises.

A formal econometric approach has been increasingly taken to estimate agglomeration economies in developing economies, but informal enterprises are largely missed in prior analysis (Combes and Gobillon, 2015).[1] Duranton (2009, p. 82) argues that most of the econometric findings on agglomeration economies concern the formal sector. To fill this gap, Chhair and Newman (2014) use the same dataset in Cambodia as our analysis and show a negative effect of agglomeration on productivity for registered and unregistered firms mainly through a competition channel. Ali and Peerings (2011) address an endogeneity issue in agglomeration by matching firms with similar characteristics between clustered and dispersed areas. In the handloom sector in Ethiopia, they find larger monthly profits for clustered small firms than non-clustered small. By contrast, our paper is distinctive in that we address both the endogeneity of agglomeration and the spatial dependence of productivity to estimate local and spatial impacts of industrial clusters separately for formal and informal firms.

[1] For instance, Lall et al. (2004) estimate agglomeration economies of formal manufacturing industries in India.

Finally, we make a methodological contribution to this line of inquiry. A Bayesian approach to instrumental variables regression has long been developed since the work of Drèze (1976), and it is increasingly used to estimate a spatial autoregressive model (Banrejee et al., 2004; LeSage and Pace, 2009). Extending prior methods, we apply a Bayesian method for a spatial autoregressive model with an endogenous independent variable. To estimate such a model, Kelejian and Prucha (1998) propose higher order spatial lags of exogenous independent variables as instruments for the spatial lag of a dependent variable. However, Gibbons and Overman (2012) argue that this approach suffers from estimation problems for identification because the higher order spatial lags are assumed to have no direct influence on the dependent variable, and these instruments may be weak because of potentially high correlations among the higher order spatial lags. In contrast, our Bayesian approach does not exploit the higher order spatial lags as instruments, thereby relaxing the identification assumptions used in the prior method. Controlling for spatial autocorrelation in the dependent variable, we can focus on the identification of the endogenous independent variable.

The rest of this paper is organized as follows. Section 2 discusses a conceptual framework. Section 3 presents a spatial autoregressive model with an endogenous independent variable. Section 4 explains the Bayesian method for estimating our empirical model. Section 5 describes the data sources and the structure of industrial activity in Cambodia. Section 6 presents the estimation results. Section 7 concludes.

2. Conceptual Framework

The main objective of our analysis is to estimate the long-run magnitude of agglomeration economies produced by a localized concentration of similar industries in a developing economy where formal and informal firms coexist but are largely segregated. As is well known, Marshall (1890) argues that a cluster of firms can be more productive for localization economies through specialized suppliers, labor market pooling, and knowledge spillovers. Firms need specialized equipment and support services to produce and develop goods and services. A localized cluster of firms provides a large market for supporting the specialized suppliers, which in turn produce a wide range of inputs at a low cost for firms in a cluster. Additionally, a cluster of firms supports a large pool of workers with a variety of specialized skills. In a pooled labor market, there is better matching between employers and employees, which reduces labor shortages and unemployment, respectively. Finally, knowledge and technology are crucial inputs of production. An exchange of information and knowledge takes place through personal interactions, and knowledge spillovers benefit producers and workers more effectively in a clustered area. In sum, firms and workers in a cluster benefit from localization externalities through more efficient sharing of local suppliers, better matching between employers and workers, and knowledge spillovers (Duranton and Puga, 2004).

A formation of industrial clusters depends on local natural advantages, including geography, natural resources, economic infrastructure, and social institutions. Some regions with strong natural advantages attract firms and workers, and vice versa. At the same time, the spatial concentration of firms and workers increases congestion costs from congested roads, higher land prices, and environmental pollution. Firms and workers in a cluster benefit from localization economies and incur congestion costs. In addition, the spatial diffusion of knowledge externalities in a cluster shapes the network formation of firms and workers over space. Firms and workers in neighboring areas benefit from

localization economies through spatial interactions with firms and workers in a clustered area. However, such spatial externalities decay as distance increases because the costs of economic and social interactions increase over space and deter the diffusion of knowledge across greater distances. Consequently, industrial clusters shape the spatial network among firms and workers, leading to the spatial interdependence of economic performance.

A distinction between the formal and informal sectors is crucial for understanding localization economies in developing economies. In the formal sector, highly educated entrepreneurs typically manage firms to produce high quality products and services for high income customers using modern production technology with skilled labor and large capital equipment. In the informal sector, uneducated entrepreneurs typically manage firms to produce low quality products and services for low income customers using traditional production technology with unskilled labor and small capital equipment. As discussed in La Porta and Shleifer (2014), the dual view of informality by Lewis (1954) suggests that the formal and informal sectors are largely segregated. Formal firms have more intensive relationships with other formal firms, and informal firms do so with other informal firms. The production and technological linkages between formal and informal firms are weak.[2]

The segregation between the formal and informal sectors suggests that formal and informal firms may benefit differently from a localized concentration of similar industries in which formal and informal firms coexist unevenly in each region. Aggregating formal and informal firms may mask distinctive localization externalities. Additionally, a dense interaction among formal firms leads to the formation of spatial networks in the formal sector, which produces the spatial interdependence of economic performance in the formal sector. On the other hand, a strong interaction among informal firms produces spatial externalities in the informal sector.

3. Econometric Framework

3.1 Structural Equation

To examine our hypotheses, we consider a spatial autoregressive model for region *i* and sector $s \in \{Formal, Informal\}$:

$$y_i^s = \rho \sum_{j=1}^{n} w_{ij} y_j^s + x_i \beta_0 + \boldsymbol{z}_i \boldsymbol{\beta}_1 + \varepsilon_i \qquad (1)$$

where y_i^s and y_j^s are a measure of productivity in regions *i* and *j*, respectively.[3] Regional productivity is measured by the log of total factor productivity (TFP) for region *i* and sector s. The variable w_{ij} represents the degree of geographical connectivity between regions *i* and *j* in the following form:

$$w_{ij} = \begin{cases} 0 & for \;\; i = j \\ \dfrac{d_{ij}^{-1}}{\sum_j^n d_{ij}^{-1}} & for \;\; i \neq j \end{cases}$$

where d_{ij} is travel time between regions *i* and *j*, the parameter ρ is the magnitude of

[2] Mukin (2014) finds that the co-agglomeration of the formal and informal sectors in India is very low whereas Moreno-Monroy et al. (2014) shows that formal sector subcontracting positively correlates with employment growth only in the most modern segments of the informal sector in India.

[3] Estimating a spatial autoregressive model for firm-level productivity is complicated by the difficulty in measuring spatial autocorrelation among individual firms. This issue is addressed by estimating a hierarchical spatial model in Hashiguchi and Tanaka (2014).

spatial autocorrelation in regional productivity, x_i is a variable to capture agglomeration of similar industries in region i, $\boldsymbol{z}_i$ is a vector of control variables, and ε_i is an error term.

We are interested in an estimate of the coefficient β_0, which represent the magnitude of net agglomeration effects arising from both localization economies and congestion effects in a localized cluster of similar industrial activities. Following the prior literature (Ciccone and Hall, 1996; Brülhart and Mathys, 2008; Broersma and Oosterhaven, 2009), we measure industrial agglomeration by using the density of local employment in similar industries: $\ln(Employment_i/Area_i)$. Thus, the coefficient β_0 represents an elasticity of regional productivity with respect to the density of local employment. Since we exploit cross-regional variation in employment density, we interpret the estimated elasticity as indicating the net localization effects that have accumulated during past periods up to the point of the year in our cross-section data.[4]

In examining the formal and informal sectors, we also examine manufacturing and service industries separately. Industrial clusters in these industries may produce agglomeration economies differently. For example, Kolko (2010) finds that manufacturing industries tend to be more agglomerated than service industries in the U.S. Graham (2009) shows that the elasticity of localization economies tends to differ in magnitude across manufacturing and service industries. Thus, we estimate the coefficient β_0 separately for four subsamples of our dataset: formal manufacturing, informal manufacturing, formal services, and informal services.

Potential spatial interdependence in regional productivity is modeled as a spatial lag variable, $\sum_{j=1}^{n} w_{ij} y_j^s$, which can be considered as arising from the network formation of firms and workers over space (Corrado and Fingleton, 2012). Conceptually, our approach is similar to the prior literature in that spatial spillovers in knowledge and technology are captured by the spatial lag variable of total factor productivity in regional production function (Ertur and Koch, 2007; Hashiguchi, 2010). From an econometric point of view, significant spatial autocorrelation may magnify or reduce the impact of local employment density through a spatial multiplier effect on productivity across regions. Thus, the spatial lag variable allows us to estimate the global magnitude of localization economies.

Our control variables are defined as follows. First, the local market size can affect both agglomeration and productivity. We control for the market size by the population size in region *i*. Second, Jacobs urbanization economies can raise productivity in regions with a more diverse range of different industrial activities. To control for cross-industry externalities, we define an industrial diversity index for industry *k* as:

$$Diversity_{ik} = \ln\left(1 \Big/ \sum_{k' \neq k} \left(\frac{Employment_{ik'}}{Employment_i - Employment_{ik}}\right)^2\right)$$

where k' indicates all other industries except for industry *k*. Third, the strength of competition in a local market affects regional productivity. More intensive competition encourages managerial efforts to improve productivity. Meanwhile, stronger competition decreases the market share of individual producers, and a decline in the scale of production reduces productivity. To account for local competition effects, we define a competition index as:

$$Competition_{ik} = \ln(1/(H_{ik}))$$

[4] Martin et al. (2011) use panel data in France to estimate agglomeration effects in the short run.

where $H_{ik} = \sum_{f \in \Omega_{ik}} \left(\frac{Sales_{ikf}}{Sales_{ik}}\right)^2$ is an Herfindahl index of sales concentration within region *i* and industry *k*, and $Sales_{ikf}$ is the volume of sales by firm *f* that belongs to the set of firms in region *i* and industry *k*, Ω_{ik}.

Regional productivity depends in part on industrial infrastructure such as access to electricity. We include an electricity access variable as measured by the proportion of households with access to city power or a generator for the main source of light in region *i*. Additionally, skilled workers may concentrate in denser areas, such as large cities, which contributes to spatial wage disparities (Glaeser and Maré, 2001; Combes et al., 2008). Andersson et al. (2007) show evidence that more productive workers are matched with more productive firms in a denser region through assortative matching and production complementarity. To control for the linkage between skilled workers and agglomeration, we include a high skill variable as measured by the proportion of persons completing technical/vocational diplomas and undergraduate/graduate degrees in region *i*. On the other hand, unskilled workers may reside in dispersed areas whereas regional productivity is lower in such areas. Thus, we include a low skill variable as measured by the proportion of persons completing only primary and secondary school programs in regions *i*.

3.2 Reduced Form Equation

We seek to identify the causal impact of localization economies on regional productivity after controlling for a variety of regional characteristics that may affect regional productivity. The inclusion of control variables $\boldsymbol{z}_i$ helps to reduce an omitted-variables bias in the estimated coefficient of localization economies, β_0. However, some regions may be endowed with unobserved natural advantages such as local climate, social infrastructure, and natural resources. If these natural advantages attract more skilled workers for higher wages, the estimated coefficient, β_0, may contain a bias. Additionally, more productive firms may self-select to locate their economic activity in agglomerated regions to benefit from positive externality. Unobserved heterogeneity in the location decisions by individual firms may cause a reverse-causality bias in the estimated coefficient, β_0. Thus, there is a concern about an endogeneity bias in the estimation of localization economies.[5]

To deal with this endogeneity problem, we employ the Bayesian instrumental variables method proposed by Rossi et al. (2005). We specify the reduced form equation:

$$x_i = \boldsymbol{q}_i \boldsymbol{\gamma}_0 + \boldsymbol{z}_i \boldsymbol{\gamma}_1 + \eta_i \quad (2)$$

where x_i is an endogenous variable and assumed to be linearly related to a set of instruments $(\boldsymbol{q}_i, \boldsymbol{z}_i)$, and $\boldsymbol{q}_i$ is a vector of exogenous variables related to x_i, but independent of the error term η_i. As instrumental variables, we exploit the regional characteristics in the past period including the density of employment in similar industries and the geographic characteristics as measured by the share of forest and shrubland area. The identifying assumption is that the employment density and geographic characteristics in the past period have persistent influences only on the preferences of workers about the location in which they seek employment opportunities. However, these instruments are not correlated with the current differences in regional productivity that are not explicitly

[5] Endogeneity problems in agglomeration economies are discussed in Eberts and McMillen (1999), Rosenthal and Strange (2004), Cohen and Paul (2009), and Puga (2010).

accounted for by our model. Our approach follows Ciccone and Hall (1996) and Ciccone (2002) in that past labor and geography are used as instruments for current agglomeration.

Our instruments can be justified as follows. First, industrial agglomeration is a result of a cumulative process in which individual economic activities are attracted to specific points in geographic areas over time. 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 larger market size and wider availability of intermediate inputs and consumer products (Fujita et al., 1999). Consequently, the past density of labor should positively affect the formation of industrial agglomeration in specific regions. Additionally, we hypothesize that geographic characteristics affect the patterns of locations where workers seek jobs. Forest and shrubland regions are more likely than plain regions to deter the formation of industrial activity for the high cost of leveling the ground, thereby yielding a geographic barrier to worker settlement. Past land characteristics should negatively influence the formation of industrial agglomeration. Consistent with our justifications, Combes et al. (2010) emphasize the usefulness of history and geology as instruments of agglomeration.

Exclusion restrictions of our instruments imply that labor density and geographic characteristics in the past are assumed to affect current regional productivity only through the current level of industrial agglomeration, but should not affect the regional productivity through other channels that are not explicitly accounted for by control variables, $\boldsymbol{z}_i$. If our instruments produce unobserved persistent effects on a local market over time, they affect the contemporaneous determinants of regional productivity. As long as such persistent influences are captured 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 localization economies. Thus, we calculate the Sargan statistic to check the validity of the exclusion restrictions.

Finally, we emphasize that the recent economic growth of the Cambodian economy helps to meet the exclusion restrictions. The economy has experienced the rapid economic growth since the early 1990s, and its industrial structure has been substantially transformed from agriculture to manufacturing and services; for instance, the share of agriculture in GDP declined from 55.6% in 1990 to 33.8% in 2010, while the share of manufacturing increased from 5.2% to 14.9% during this period (Hill and Menon, 2013). These rapid structural changes should isolate our instruments based on past data from unobserved current shocks to regional productivity.

3.3 A System of Equations

Combining equation (2) with a structural equation (1), we specify a system of equations in vector and matrix notation as follows:

$$\boldsymbol{x} = \boldsymbol{Q}\boldsymbol{\gamma}_0 + \boldsymbol{Z}\boldsymbol{\gamma}_1 + \boldsymbol{\eta} \tag{3}$$

$$\boldsymbol{S}\boldsymbol{y} = \boldsymbol{x}\beta_0 + \boldsymbol{Z}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{x} = (x_1, \cdots, x_n)'$, $\mathbf{Q} = (\boldsymbol{q}_1, \cdots, \boldsymbol{q}_n)'$, $\mathbf{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)'$, $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)'$, $\mathbf{y} = (y_1, \cdots, y_n)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)'$, and $\mathbf{S} = \boldsymbol{I}_n - \rho\boldsymbol{W}$. This system of equations consists of a structural equation with an endogenous independent variable and multiple instrumental variables. For brevity, we suppress the superscript s.

After estimating the system of equations, we are interested in calculating a marginal effect of localization economies. If an estimate of spatial correlation, ρ, is not significant,

the marginal effect is simply based on the estimated coefficient of localization economies, β_0. By contrast, if there is significant spatial correlation in regional productivity, a marginal change in agglomeration for each region has an impact not only on its own productivity but also indirectly on the other regions' productivity through the spatial network structure. More specifically, the marginal impact in our model can be written by showing the total derivative of $\boldsymbol{y}$ in equation (4) under the constraints $d\boldsymbol{Z} = \mathbf{0}$ and $d\boldsymbol{\varepsilon} = \mathbf{0}$:

$$d\boldsymbol{y} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}'} d\boldsymbol{x} = \boldsymbol{S}^{-1}\beta_0 d\boldsymbol{x}. \tag{5}$$

The marginal changes in agglomeration for region *i* can affect the productivity in both its own region and other regions. For $\rho = 0$, $\boldsymbol{S}^{-1}$ is reduced to an identity matrix, and spatial effects will be removed. We can measure the marginal effects by the total impact (TI), direct impact (DI), and indirect impact (IDI) of agglomeration on regional productivity in region *i*:

$$TI_i = DI_i + IDI_i,$$

where

$$DI_i = \frac{\partial y_i}{\partial x_i} dx_i,$$
$$IDI_i = \sum_{j=1}^{n} \frac{\partial y_j}{\partial x_i} dx_i \text{ for } j \neq i.$$

In the following analysis, we calculate the averages of TI_i, DI_i, and IDI_i over all regions for $dx_i = 1$. These impacts are referred to as average total impact (ATI), average direct impact (ADI), and average indirect impact (AIDI), following the terminology by LeSage and Pace (2009).

4. Bayesian Estimation

This section describes the Bayesian estimation of our model. The Bayesian estimation requires a *posterior density* to draw an inference regarding unknown parameters in the model. In general, the posterior is proportional to the *likelihood function* times the *prior density*: $\pi(\boldsymbol{\theta}|\ \boldsymbol{y}) \propto f(\boldsymbol{y}|\ \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta}|\ \boldsymbol{y})$ is the posterior; $f(\boldsymbol{y}|\ \boldsymbol{\theta})$ is the likelihood; $\pi(\boldsymbol{\theta})$ is the prior; $\boldsymbol{y}$ is the observed data, and $\boldsymbol{\theta}$ is the unknown parameters. In this section, we describe the likelihood and the prior in our model, and explain the computational scheme for estimating the posterior.

4.1 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.[6] Denoting $\tilde{\boldsymbol{y}} = (\boldsymbol{x}, \boldsymbol{y})'$ and $\mathbf{u}$ as a $(2n\times1)$ vector following a multivariate standard normal distribution $N(\mathbf{0}, \mathbf{I}_{2n})$, we can express the system of equations (3) and (4) as:

$$\mathbf{u} = (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)^{-\frac{1}{2}} \left[\begin{pmatrix} \mathbf{I}_n & \mathbf{0} \\ -\beta_0 \mathbf{I}_n & \boldsymbol{S} \end{pmatrix} \tilde{\boldsymbol{y}} - \begin{pmatrix} \boldsymbol{Q} \\ \mathbf{0} \end{pmatrix} \boldsymbol{\gamma}_0 - \begin{pmatrix} \boldsymbol{Z} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\beta}_1 \end{pmatrix} \right]. \tag{6}$$

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

[6] Alternative forms of error terms would have little influence on the means of posterior distributions, but may affect the standard deviations. It is beyond the scope of this paper to extend our model for alternative assumptions of the error terms.

$$J = \left|\frac{\partial \mathbf{u}}{\partial \tilde{\mathbf{y}}}\right| = |\boldsymbol{\Sigma}|^{-\frac{n}{2}}|\boldsymbol{S}|. \quad (7)$$

The likelihood function is specified by:

$$L = (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}|\boldsymbol{S}| \exp\left\{\begin{bmatrix} \boldsymbol{x} - \boldsymbol{Q}\boldsymbol{\gamma}_0 - \boldsymbol{Z}\boldsymbol{\gamma}_1 \\ \boldsymbol{S}\boldsymbol{y} - \boldsymbol{x}\beta_0 - \boldsymbol{Z}\boldsymbol{\beta}_1 \end{bmatrix}' (\boldsymbol{\Sigma}\otimes\mathbf{I}_n)^{-1} \begin{bmatrix} \boldsymbol{x} - \boldsymbol{Q}\boldsymbol{\gamma}_0 - \boldsymbol{Z}\boldsymbol{\gamma}_1 \\ \boldsymbol{S}\boldsymbol{y} - \boldsymbol{x}\beta_0 - \boldsymbol{Z}\boldsymbol{\beta}_1 \end{bmatrix}\right\}. \quad (8)$$

Independent priors for the unknown parameters are specified as:

$$\begin{gathered} \boldsymbol{\beta}^* \equiv \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} \sim MVN(\boldsymbol{b}_\beta, \boldsymbol{B}_\beta), \quad \boldsymbol{\gamma}^* \equiv \begin{pmatrix} \gamma_0 \\ \boldsymbol{\gamma}_1 \end{pmatrix} \sim MVN(\boldsymbol{b}_\gamma, \boldsymbol{B}_\gamma), \\ \rho \sim U(\lambda_{min}^{-1}, \lambda_{max}^{-1}), \qquad \boldsymbol{\Sigma} \sim IW(b_\Sigma, \boldsymbol{B}_\Sigma), \end{gathered} \quad (9)$$

where $U(\ ,)$ is a uniform distribution, and $IW(\ ,)$ is an inverted Wishart distribution. The prior parameters are denoted by $\boldsymbol{b}_\beta$, $\boldsymbol{B}_\beta$, $\boldsymbol{b}_\gamma$, $\boldsymbol{B}_\gamma$, λ_{min}, λ_{max}, b_Σ, and $\boldsymbol{B}_\Sigma$. The parameters, λ_{min} and λ_{max}, are the minimum and maximum real eigenvalues of **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 **W** contains only real values, this restriction ensures $|\boldsymbol{S}| > 0$. The values of the other prior parameters are assumed as follows: $\boldsymbol{b}_\beta = \mathbf{0}$, $\boldsymbol{B}_\beta = 100\mathbf{I}_k$, $\boldsymbol{b}_\gamma = \mathbf{0}$, $\boldsymbol{B}_\gamma = 100\mathbf{I}_l$, $b_\Sigma = 2$, and $\boldsymbol{B}_\Sigma = 2\mathbf{I}_2$. k and l denote the dimensions of $\boldsymbol{b}_\beta$ and $\boldsymbol{b}_\gamma$, respectively.

Specifying prior parameters is difficult when no information is available for unknown parameters. Following the standard practice in Bayesian estimation, we choose zero values for the location parameters of the prior distributions for the coefficients. This assumption implies that our prior beliefs for the coefficients are centered on zero. Based on these zero values, we investigate the extent to which the posterior distributions move away from the prior distributions. We set these priors to have large variances to ensure that our prior beliefs for the unknown parameters are non-informative.

Before proceeding to sampling procedures, we highlight the advantage of our approach over the prior approach in estimating a spatial autoregressive model with an endogenous independent variable. Kelejian and Prucha (1998) propose higher order spatial lags of exogenous independent variables as instruments for the spatial lag of a dependent variable. This method is applied by Artis et al. (2012) in estimating agglomeration economies. However, Gibbons and Overman (2012) argue that this approach suffers from estimation problems for identification. 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 may be weak because of potentially high correlations among the higher order spatial lags. In this case, it leads to a biased estimate for the spatial lag variable. In contrast, our Bayesian approach does not exploit higher order spatial lags as instruments, thereby relaxing the identification assumptions used in the prior method. By controlling for potential spatial autocorrelation in the dependent variable, we can focus on the identification of the endogenous independent variable in the spatial autoregressive model.

4.2 Markov Chain Monte Carlo Method

Based on the likelihood and the prior for our model, we 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 as follows:

$$\boldsymbol{\beta}^*|\ \boldsymbol{\gamma}^*, \rho, \boldsymbol{\Sigma}, \text{Data} \sim Normal\ \ Distribution,$$
$$\boldsymbol{\gamma}^*|\ \boldsymbol{\beta}^*, \rho, \boldsymbol{\Sigma}, \text{Data} \sim Normal\ \ Distribution,$$
$$\boldsymbol{\Sigma}|\ \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \rho, \text{Data} \sim Inverted\ \ Wishart\ \ Distribution,$$
$$\rho|\ \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\Sigma}, \text{Data} \sim Unknown\ \ Distribution$$

where Data $= \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{Z}\}$. These full conditional distributions are derived from the likelihood function and the specified priors, which are described in Appendix A.

Based on the full conditional posteriors described in Appendix A, we conduct the following 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 $MVN\left(\widehat{\boldsymbol{b}}_\beta, \widehat{\boldsymbol{B}}_\beta\right)$, given $\boldsymbol{\gamma}^{*(r-1)}$, $\boldsymbol{\Sigma}^{(r-1)}$, $\rho^{(r-1)}$, Data.

Draw $\boldsymbol{\gamma}^{*(r)}$ from $MVN\left(\widehat{\boldsymbol{b}}_\gamma, \widehat{\boldsymbol{B}}_\gamma\right)$, given $\boldsymbol{\beta}^{*(r)}$, $\boldsymbol{\Sigma}^{(r-1)}$, $\rho^{(r-1)}$, Data.

Draw $\boldsymbol{\Sigma}^{(r)}$ from $IW\left(\widehat{\boldsymbol{b}}_\Sigma, \widehat{\boldsymbol{B}}_\Sigma\right)$, given $\boldsymbol{\beta}^{*(r)}$, $\boldsymbol{\gamma}^{*(r)}$, $\rho^{(r-1)}$, Data.

Draw ρ', a candidate of $\rho^{(r)}$, from $TN\left(\hat{\rho}, \hat{\sigma}_\rho^2\right)$, given $\boldsymbol{\beta}^{*(r)}$, $\boldsymbol{\gamma}^{*(r)}$, $\boldsymbol{\Sigma}^{(r)}$, Data.

Calculate an acceptance probability:

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

Set $\rho^{(r-1)} = \rho'$ with a probability $\alpha\left(\rho', \rho^{(r-1)}\right)$, and set $\rho^{(r)} = \rho^{(r-1)}$ with a probability $1 - \alpha\left(\rho', \rho^{(r-1)}\right)$.

If $r <$ M, set $r = r + 1$ and return to step (ii). Otherwise, proceed to step (iii).[7]

(iii) Discard the samples with the superscript $r = 1,2, \cdots, M_0$, and keep the samples with $r = M_0 + 1, M_0 + 2, \cdots, M$.

In this paper, we set $M = 30{,}000$ and $M_0 = 10{,}000$, implying that we retain 20,000 replications for the posterior inference.[8]

4.3 Monte Carlo Simulation

There is an identification problem in an instrumental-variable regression model when instruments are weakly correlated with an endogenous independent variable. The weak instruments affect the identification of structural parameters (Rossi et al., 2005). To investigate the estimation performance of our model, we conduct Monte Carlo simulation in two cases when instruments are strong and weak. Details of the simulation setups are provided in Appendix B.

In both cases, we find that the median of posterior distributions for the structural parameters is close to their true value and their 95% credible intervals contain the true value. These findings support the credibility of our model for inference. In the weak-instruments case, the posterior distributions have a larger dispersion for the structural parameters and the MCMC sampler has a higher autocorrelation. In this case, additional MCMC draws may be required to obtain more precise posterior distributions.

---

[7] Because the large size of the spatial matrix $\mathbf{W}$ makes it computationally difficult to compute the determinant $|\mathbf{I}_n - \rho\mathbf{W}|$, we follow the approximation by Ord (1975, p.121): $|\mathbf{I}_n - \rho\mathbf{W}| = \prod_{i=1}^{n}(1 - \rho\lambda_i)$, where $\lambda_1, \cdots, \lambda_n$ are real eigenvalues of $\mathbf{W}$.

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

5. Data Description

5.1 The Economic Census of Cambodia in 2011

Our main dataset is based on the Economic Census of Cambodia in 2011 (EC2011). The purpose of the EC2011 is to survey economic activities of all nonfarm establishments and enterprises over the entire territory of Cambodia. The administrative geographic units surveyed include 1,621 communes in 24 provinces. The survey was mainly funded by Japanese Official Development Assistance and implemented by the National Institute of Statistics (NIS) in the Cambodian Ministry of Planning, in cooperation with the Japan International Cooperation Agency.[9] The census enumeration was conducted in March 2011 to survey all the establishments and enterprises, including the street vendors that operate at a fixed location but can move.[10] To collect the data, census enumerators visited each establishment to interview its representative and/or owner. Through face-to-face interviews, the enumerators filled out a questionnaire for each establishment. The NIS collected all the questionnaires for data input and checked data consistency by comparing two data files that were created separately by two data-input operators.

In the EC2011 questionnaire, each establishment was asked 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 the activities of unregistered firms. In the registration process, formal firms must provide the registrar with the specific location of their office and the name of their agent.[11] The registration procedures require a firm (1) to deposit the legally required initial capital in a bank and obtain deposit evidence, (2) conduct an initial check of the 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. In the literature, Schneider and Enste (2013, Chapter 2) define the national economy as the dual economy of official sector and underground sector, the latter of which consists of the shadow economy based on market transactions and the household sector based on non-market transactions for self-sufficient individual needs. Our definition of the formal and informal sectors corresponds to the official sector and the shadow economy, respectively.

To gauge the relative importance of the shadow economy in Cambodia, Table 1 presents aggregate data on formal and informal establishments, with the financial figures

[9] In preparing for the EC2011, the NIS created the establishment listing in Phnom Penh for 2006 and conducted the establishment survey in Phnom Penh for 2007. The nation-wide establishment listing was created for 2009.

[10] The survey does not cover the establishments classified into (1) agriculture, forestry, and fishing, (2) public administration and defense, (3) activities of households as employers, (4) activities of extraterritorial organizations and bodies, and (5) mobile establishments such as a bike taxi and a street peddler.

[11] A regulatory framework for commercial enterprises in Cambodia 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. In 2005, the National Assembly in Cambodia adopted the ”Law on Commercial Enterprise”, which applies to partnerships, private limited companies, public limited companies, and foreign businesses.

measured for one month in February 2011. Across all industries, there were 17,378 formal establishments and 487,756 informal establishments.[12] The informal sector represents 96.6% of the total number of establishments, 23.4% of total sales, 40.8% of total wages, and 66.4% of total employment. For manufacturing industries, there were 1,723 formal establishments and 69,693 informal establishments, with the informal sector representing 97.6% of firms, 60.8% of sales, 35.8% of wages, and 32.5% of employment. On the other hand, there were 6,245 formal establishments and 286,101 informal establishments in wholesale and retail trade industries, with the informal sector representing 97.9% of total firms, 87.8% of sales, 60.5% of wages, and 93.1% of employment.

---Table 1 here---

Our data show that the informal sector is remarkably large in Cambodia's industrial activities. While the formal sector provides large employment opportunities in manufacturing, the informal sector is prominent in both manufacturing and wholesale/retail industries. Additionally, these industries account for the large share of total industrial activity. The wholesale and retail industries accounted for 33.1% of total employment in nonfarm industries, and the manufacturing industries constituted 31.7%. Thus, the analysis of these industries is critical to understanding agglomeration effects in Cambodia.

Using the EC2011, we construct the variables on regional productivity, the density of local employment, an industrial diversity index, and a competition index. More specifically, we estimate regional productivity y_i^s in two steps, with details of the data construction provided in Appendix C1. First, we assume that production function for firm f follows a Cobb-Douglas form, $q_f = A_f {K_f}^{\beta_k} {L_f}^{\beta_l}$, where q_f is value added; K_f is capital input; L_f is labor input, and A_f is the efficiency level that is observed by a producer, but not by an econometrician. We estimate a log-linear form of the above production function by OLS for a sample on individual firms in each industry. Based on the OLS estimates, we calculate the log of the residuals as a proxy for total factor productivity, $\ln(\mathrm{TFP}_f)$. Second, we compute the regional productivity of firms in region i and sector $s \in \{Formal, Informal\}$ separately for each industry:

$$y_i^s = \sum_{f \in \Omega_{is}} \frac{q_{fis}}{\sum_{f \in \Omega_{is}} q_{fis}} \ln(\mathrm{TFP}_{fis})$$

where Ω_{is} is the set of firms that belong to region i and sector s. The average productivity across region i and sector s for each industry is approximated by the average of firm-level TFP weighted by the share of its own value added in the total value added. An increase in regional productivity results from an increase in the firm-level TFP values and an expansion of the market share of more productive firms, i.e., allocative efficiency.

5.2 Data on Spatial Weight Matrix

A spatial-weighting matrix is designed to capture the degree of connectivity between all pairs of distinct regions, which should approximate the spatial dependence in regional productivity that has been formed in past periods up to the year of our cross-sectional data. To approximate such connectivity, we construct a dataset on the shortest travel time between communes in Cambodia. Shorter travel time from one commune to another

[12] We exclude NGOs from the data.

commune indicates greater connectivity between these communes, whereas longer travel time suggests weaker connectivity.

We compute the travel time for the shortest path among all possible routes 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 travel distance and travel speed. These data are constructed from (1) the Geographic Information System (GIS) shape file on 1,621 administrative units at the commune level in Cambodia, (2) the map of the Cambodian road network published by the Cambodian Ministry of Public Works and Transport, (3) the Japan External Trade Organization (JETRO) survey on ASEAN logistics network map (JETRO, 2009), and (4) satellite-image data of geographic conditions in each commune. Details of the data construction are given in Appendix C2.

5.3 Other Data Sources

Data on population and households are taken from the Population Census of Cambodia in 1998 and 2008. These datasets contain information on the characteristics of the population and households, including residential location, education, and electricity access. Additionally, geographic data on forest and shrubland are constructed from the satellite land cover images provided by the Land Processes Distributed Active Archive Center, a component of NASA's Earth Observing System Data and Information System. Land cover maps in Cambodia are matched with the GIS shape file of communes to calculate the proportion of land areas classified as forest and shrubland for year 2002.

6. Estimation Results

6.1 Main Results

Table 2 presents the summary statistics of the sample used. The mean and standard deviation of variables are shown separately for the formal and informal sectors in manufacturing and wholesale/retail industries, respectively. Table 3 reports the main estimation results. In manufacturing, the posterior mean coefficient of employment density is 0.001 and insignificant for the formal sector in column (1). By contrast, the posterior mean coefficient is 0.088 for the informal sector in column (2) and statistically significant at the 1% level because the 99% credible interval of the posterior distribution does not contain zero. In wholesale/retail, the posterior mean coefficient of employment density is 0.050 and insignificant for the formal sector in column (3). The posterior mean coefficient is 0.052 and significant at the 1% level for the informal sector in column (4). These results indicate that the density of local employment has a significantly positive effect on the regional productivity only in the informal sector.

---Tables 2 and 3 here---

We examine the validity of our instruments to check whether positive localization economies in the informal sector can be interpreted as suggesting a causal relationship. First, the past density of employment has the significantly positive posterior mean coefficients at the 1% level across specifications. The past forest variable has significantly negative posterior mean coefficients at the 1% level for the informal sectors in columns (2) and (4). Thus, the past level of employment density and geographic characteristics should significantly affect the spatial pattern of agglomeration. Our specifications are not likely to suffer from a weak instrument problem. Second, we compute the Sargan statistic and p-values based on a χ^2 distribution to check the empirical validity of the exclusion

restrictions. The p-values of the Sargan statistic are large across specifications. For instance, the p-value is 0.204 in column (2) and 0.483 in column (4). We find 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 identification assumption of our instruments, thereby suggesting a causal impact of the localized concentration of similar industries for the informal sector.

We turn to discuss the spatial lag in our model. In manufacturing, the posterior mean of the spatial lag of productivity is not significant for the formal sector in column (1), but significantly positive at the 1% level for the informal sector in column (2). In wholesale/retail, the posterior mean coefficients for the spatial lag are insignificant for the formal sector in column (3), but significantly positive at the 1% level for the informal sector in column (4). These results suggest that there is not significant spatial interdependence of economic performance among formal firms. Meanwhile, there exists the positive spatial interdependence among informal firms, implying that the local impact of positive localization economies should be magnified by the positive spatial network within the informal sector. Thus, we find it crucial to control for the spatial lag of regional productivity in estimating agglomeration economies in the informal sector.

To examine the magnitude of localization economies, we compute marginal effects of employment density on regional productivity in the informal sector for manufacturing and wholesale/retail industries. Table 4 presents a summary of average impacts on productivity, which are computed from the results in Table 3. In manufacturing industry, the average direct and indirect impacts are 0.09 and 0.17, respectively. Combining these impacts, we find the average total impact of 0.26. Intuitively, these results indicate that a doubling of the density of local employment increases informal firms' productivity by 9% through direct linkages in their own region and by 17% through spatial multiplier linkages in other regions. These linkages combine to produce a 26% increase in productivity. Additionally, the average direct and indirect impacts in wholesale/retail are 0.05 and 0.22, respectively. The average total impact is 0.22. Intuitively, a doubling of the density of local employment increases informal firms' productivity by 5% through direct linkages in their own region and by 22% through spatial multiplier linkages, which combine to produce a 28% increase in productivity. Thus, the average indirect impact is larger than the average direct impact in both industries, implying that the spatial network structure is crucial for estimating the precise magnitude of localization economies.[13]

---Table 4 here---

6.2 Robustness Checks

In the benchmark analysis, we measure industrial agglomeration by the density of workers. Henderson (2003) and Martin et al. (2011) argue that localization economies can be examined by using the number of firms in the same industry and region. If agglomeration externalities occur mainly through an interaction within firms rather than within workers, we might underestimate the localization economies in the benchmark results. To address this concern, we extend the main specification by measuring the agglomeration as the density of firms in the same industry and commune.

Table 5 presents the estimation results. In manufacturing, the posterior mean

[13] The regional variation in total impacts is determined by the structure of the spatial weighting matrix. In general, more accessible communes tend to have larger total impacts whereas less accessible communes have smaller total impacts.

coefficient of firm density is -0.001 and insignificant for the formal sector in column (1). The coefficient is 0.098 and significant at the 1% level for the informal sector in column (2). In wholesale/retail, the posterior mean coefficient of firm density is 0.050 and insignificant for the formal sector in column (3). The coefficient is 0.052 and significant at the 1% level for the informal sector in column (4). In columns (2) and (4), the posterior mean coefficients of past employment density and past forest are significantly positive at the 1% level. The Sargan statistic has a large p-value in these specifications. These results support the validity of our instrumental variables in the specification using the density of firms. Thus, the density of firms also has a positive impact on regional productivity in the informal sector across industries, and the magnitude of the posterior mean coefficient is similar to the results using employment density. The benchmark results are robust to the alternative definition of industrial agglomeration.

---Table 5 here---

For further robustness checks, we estimate the benchmark specification with additional control variables.[14] In the first specification, we include cropland area to capture the effect of agriculture on industrial location and a dummy variable for border regions to account for an accessibility advantage in exporting to foreign markets through land borders. We also include a dummy variable for major international airports and special economic zones. In the second specification, we include the spatial lag of the density of local employment to address a possible confounding factor due to a spatial concentration of similar industries in other regions. We find that the estimation results in these alternative specifications do not differ substantially from the main estimation results, supporting the positive localization economies in the informal sector.

6.3 Further Issues

Previous discussions up to this point suggest that the density of local employment in similar industries has a positive impact on regional productivity in the informal sector, but has little influence in the formal sector. These findings raise further questions. First, we have focused on the density of total employment in both formal and informal sectors to investigate the average cluster benefit. It is plausible that formal firms benefit strongly from a localized cluster of formal firms and weakly from that of informal firms, suggesting that our measure might lead to insignificant localization economies for the formal sector. On the other hand, informal firms may also benefit strongly (weakly) from a localized cluster of informal (formal) firms. The previous results may mask heterogeneous localization economies in terms of the proportion of formal firms in a cluster.

To address this issue, we extend the benchmark specification with an interaction term between the density of employment and the proportion of formal-firm employment in each region. Table 6 presents the estimation results. In manufacturing, the posterior mean coefficient of employment density remains insignificant for the formal sector in column (1). The posterior mean coefficient of the interaction term is -0.152 and the 95% credible interval of the posterior distribution does not contain zero. A large proportion of formal firms reduces positive localization impacts for the formal sector, possibly suggesting that a localized concentration of formal manufacturing firms significantly raises congestion costs. Additionally, the posterior mean coefficients of employment density and the

[14] These estimation results are available upon request.

interaction are 0.109 and -0.157 for the informal sector in column (2), respectively. The 99% credible intervals of their posterior distributions do not contain zero. Consistent with the result in column (1), a large share of formal firms decreases positive localization effects for the informal sector. In other words, a large proportion of informal firms in a cluster magnify positive localization effects for the informal sector. Finally, the results for wholesale/retail industry show that the posterior mean coefficients of the interaction are insignificant for both formal and informal sectors in columns (3) and (4).

---Table 6 here---

Up to this point, we have estimated a linear relationship between industrial clustering and regional productivity. This approach raises a question of whether more dense employment always increases positive localization effects. Because more clustering would increase not only localization benefits but congestion costs, congestion costs may exceed positive localization externalities in high clustering regions. Thus, there may be a nonlinear relationship between clustering and productivity.

To examine this issue, we extend the benchmark specification by including a quadratic term of the density of local employment. Table 7 presents the estimation results. In manufacturing, the posterior mean coefficient of employment density is 0.184 and insignificant for the formal sector in column (1), whereas it is 0.102 and significant for the informal sector in column (2). In column (2), the posterior mean coefficient of the quadratic term is -0.012 and the 99% credible interval of the posterior distribution does not contain zero. Thus, the density of local employment increases regional productivity in the informal sector at a decreasing rate. This result can be interpreted that more clustering not only generates positive localization economies for informal manufacturing firms, but increases congestion costs at an increasing rate. Industrial clustering produces a nonlinear effect on productivity. Additionally, the results for wholesale/retail industry show that the posterior mean coefficients of employment density are insignificant for the formal sector in column (3) and significant for the informal sector in column (4). The posterior mean coefficients of the quadratic term are not significant for both sectors. We find little evidence of nonlinear localization effects in wholesale/retail industry.

---Table 7 here---

7. Concluding Remarks

It is widely held that industrial clusters produce agglomeration economies and promote industrialization in developing economies, as has been previously demonstrated for high and middle income countries. However, the large informal sector in developing economies poses a question as to whether both formal and informal firms benefit from industrial clusters. Using a comprehensive dataset on formally registered and unregistered business establishments in Cambodia, this paper estimates the long-run magnitude of agglomeration economies produced by a localized concentration of similar industries. We develop a Bayesian spatial approach to address the endogeneity of industrial agglomeration and spatial dependence in economic performance.

We find that localization economies are significantly positive in the informal sector, but have little effect on the formal sector. In manufacturing, a doubling of employment density increases regional productivity in the informal sector by 9% through local linkages and by 17% through spatial multiplier linkages, leading to a 26% increase in total. Similar results are found for wholesale/retail industry. A spatial network serves to magnify the positive impact of localization economies in the informal sector.

These results provide implications for developing economies. The low level of cluster benefits in the formal sector may indicate the weak linkages between formal and informal firms. The distinction between the formal and informal sectors is crucial for understanding agglomeration economies in developing economies. Additionally, the spatial network among firms can magnify agglomeration economies through geographic connectivity. This finding highlights that an improvement in transportation infrastructure can enhance agglomeration economies.

References

Ali, M, and Peerlings, J., 2011. Value added of cluster membership for micro enterprises of the handloom sector in Ethiopia. *World Development*, 39 (3), 363-374.

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., pp. 1-45.

Artis, M.J., Miguelez, 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.

Chhair, S., and Newman, C., 2014. Clustering, productivity and spillover effects: evidence from Cambodia. *UNU-WIDER Working Paper*, No. WP2014-065.

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., 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, Chicago, pp. 15-66.

Combes, P., and Gobillon, L., 2015. The empirics of agglomeration economies. In: G. Duranton, J. V. Henderson, and W. C. Strange, (ed) *Handbook of Regional and Urban Economics*, vol. 5, Elsevier, Amsterdam, pp. 247-348.

Corrado, L, and Fingleton, B., 2012. Where is the economics in spatial econometrics? *Journal of Regional Science*, 52 (2), 210-239.

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., pp. 67-113.

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, Elsevier, Amsterdam, pp. 2063-2117.

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

Ertur, C., and Koch, W., 2007. Growth, technological interdependence and spatial externalities: theory and evidence. *Journal of Applied Econometrics*, 22 (6), 1033-1062.

Fajnzylber, P., Maloney, W.F., and Montes-Rojas, G.V., 2011. Does formality improve micro-firm 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.

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.
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.
Hashiguchi, Y., 2010. Bayesian estimation of spatial externalities using regional production function: the case of China and Japan. *Economics Bulletin*, 30 (1), 751-764.
Hashiguchi, Y., and Tanaka, K., 2014. Agglomeration and firm-level productivity: a Bayesian spatial approach. *Papers in Regional Science*, 94 (S1), 95-114.
Hashino, T., and Otsuka, K., 2013. Cluster-based industrial development in contemporary developing countries and modern Japanese economic history. *Journal of the Japanese and International Economies*, 30, 19-32.
Henderson, J. V., 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.
International Monetary Fund (IMF), 2012. *World Economic Outlook 2012*, IMF, Washington, D.C.
Japan External Trade Organization (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., 2010. Urbanization, agglomeration, and coagglomeration of service industries. In: E.L. Glaeser, (ed) *Agglomeration Economics*, The University of Chicago Press, Chicago, pp. 151-180.
Lall, S.V., Shalizi, Z., and Deichmann, U., 2004. Agglomeration economies and productivity in Indian industry. *Journal of Development Economies*, 73 (2), 643-673.
La Porta, R., and Shleifer, A., 2014. Informality and development. *Journal of Economic Perspectives*, 28 (3), 109-126.
LeSage, J., and Pace, P.K., 2009. *Introduction to Spatial Econometrics*. CRC press, Boca Raton.
Lewis, W. A., 1954. Economic development with unlimited suppliers of labor, *Manchester School of Economic and Social Studies*, 22 (2), 139-191.
Livingstone, I., 1991. A reassessment of Kenya's rural and urban informal sector. *World Development*, 19 (6), 651-670.
Martin, P., Mayer, T., and Mayneris, F., 2011. Spatial concentration and plant-level productivity in France. *Journal of Urban Economics*, 69 (2), 182-195.
Marshall, A., 1890. *Principles of Economics*. Macmillan, London.
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), 5-24.
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.
Moreno-Monroy, A., 2012. Critical commentary. Informality in space: understanding agglomeration economies during economic development. *Urban Studies*, 49 (10), 2019-2030.
Moreno-Monroy, A., Pieters, J., and Erumban, A. A., 2014. Formal sector subcontracting and informal sector employment in Indian manufacturing. *IZA Journal of Labor & Development*, 3: 22.
Mukin, M., 2014. Coagglomeration of formal and informal industry: evidence from India. *Journal of Economic Geography*, 15 (2), 329-351.
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.

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., 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*, No. 5356, World Bank.

Schneider, F., and Enste, D. H., 2013. *The Shadow Economy: An International Survey*. Cambridge University Press, Cambridge.

Schmitz, H., 1995. Collective efficiency: growth path for small-scale industry. *Journal of Development Studies*, 31 (4), 529-566.

Schmitz, H., and Nadvi, K., 1999. Clustering and industrialization: introduction. *World Development*, 27 (9), 1503-1514.

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

World Bank, 2014. *Doing Business 2014: Understanding Regulations for Small and Medium-size Enterprises*, World Bank, Washington D.C.

Zeitlin, J., 2008. Industrial districts and regional clusters. In: G. Jones and J. Zeitlin, (ed) *The Oxford Handbook of Business History*, Oxford University Press, Oxford, pp. 219-243.

Table 1. The Size of the 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; figures in parentheses show the percentage share of registered or unregistered firms for the corresponding variable.

Table 2. Summary Statistics

Sample	Manufacturing				Wholesale/Retail			
	Formal		Informal		Formal		Informal	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Productivity	1.125	0.945	0.853	0.858	1.218	0.951	0.621	0.654
Density of employment	2.154	2.426	0.298	2.289	3.156	2.167	1.410	1.909
Formal employment share	0.298	0.289	0.092	0.230	0.087	0.117	0.022	0.070
Population size	9.281	0.581	8.867	0.594	9.275	0.624	8.820	0.643
Industrial diversity	1.124	0.273	0.977	0.256	1.357	0.410	1.224	0.393
Local competition	2.189	0.936	2.048	1.050	3.853	1.104	3.421	0.892
Electricity access	0.487	0.382	0.192	0.281	0.506	0.362	0.188	0.277
High skilled labor	0.062	0.092	0.020	0.047	0.059	0.085	0.019	0.046
Low skilled labor	0.764	0.192	0.605	0.234	0.777	0.187	0.594	0.240
Past density of employment	1.587	2.745	-0.789	2.531	2.577	2.650	-0.043	2.662
Past forest	0.098	0.597	0.125	0.697	0.095	0.626	0.128	0.715
No. of obs.	281		1,558		353		1,621	

Table 3. Main Estimation Results

Dependent variable: regional productivity

	(1)	(2)	(3)	(4)
	Manufacturing		Wholesale/Retail	
Variable	Formal	Informal	Formal	Informal
Density of employment	0.001	0.088**	0.050	0.052**
	(0.055)	(0.022)	(0.044)	(0.017)
Population size	0.286**	0.368**	0.208*	0.137**
	(0.103)	(0.038)	(0.091)	(0.031)
Industrial diversity	0.288	0.096	0.090	0.188**
	(0.237)	(0.091)	(0.121)	(0.039)
Local competition	-0.093	-0.177**	-0.075	-0.068**
	(0.059)	(0.024)	(0.051)	(0.021)
Electricity access	0.382	0.586**	0.438*	0.836*
	(0.264)	(0.112)	(0.217)	(0.082)
High skilled labor	0.998	-0.763	-0.531	-0.440
	(0.882)	(0.603)	(0.915)	(0.487)
Low skilled labor	0.151	-0.091	0.901**	-0.352**
	(0.438)	(0.127)	(0.346)	(0.086)
Spatial lag of productivity	-0.442	0.660**	0.131	0.812**
	(0.369)	(0.162)	(0.324)	(0.111)
Instrumental variable				
Past density of employment	0.663**	0.555**	0.711**	0.541**
	(0.051)	(0.021)	(0.026)	(0.013)
Past forest	-0.151	-0.154**	-0.107	-0.121**
	(0.125)	(0.046)	(0.063)	(0.029)
Sargan	0.431	1.615	0.001	0.491
(p-value)	(0.512)	(0.204)	(0.980)	(0.483)
No. of obs.	281	1,558	353	1,621

Notes: Figures and those in parentheses indicate the mean and standard deviation of posterior distribution for each variable, respectively; constant is not reported; asterisks * and ** indicate that the 95% and 99% credible intervals of the posterior distribution do not contain zero, respectively.

Table 4. Marginal Effects of Employment Density on Productivity

	Manufacturing Informal	Wholesale/Retail Informal
Average Total Impact	0.26	0.28
Average Direct Impact	0.09	0.05
Average Indirect Impact	0.17	0.22

Notes: Total, direct, and indirect impacts are computed for an increase of one unit in the log of employment density, implying that these impacts show a percentage change in productivity resulting from a one percentage increase in employment density; the calculation is based on the estimation results in Table 3.

Table 5. Estimation Results for the Density of Firms

Dependent variable: regional productivity

	(1)	(2)	(3)	(4)
	Manufacturing		Wholesale/Retail	
Independent variable	Formal	Informal	Formal	Informal
Density of firms	-0.001	0.098**	0.050	0.052**
	(0.060)	(0.025)	(0.045)	(0.017)
Population size	0.286**	0.391**	0.209*	0.140**
	(0.099)	(0.037)	(0.091)	(0.030)
Industrial diversity	0.288	0.094	0.090	0.188**
	(0.234)	(0.090)	(0.121)	(0.039)
Local competition	-0.093	-0.196**	-0.079	-0.071**
	(0.065)	(0.027)	(0.051)	(0.021)
Electricity access	0.386	0.685**	0.443*	0.835**
	(0.248)	(0.105)	(0.216)	(0.082)
High skilled labor	1.015	-1.021	-0.455	-0.360
	(0.933)	(0.616)	(0.883)	(0.476)
Low skilled labor	0.154	-0.088	0.896**	-0.356**
	(0.448)	(0.125)	(0.348)	(0.086)
Spatial lag of productivity	-0.476	0.680**	0.131	0.816**
	(0.369)	(0.157)	(0.324)	(0.109)
Instrumental variable				
Past density of employment	0.609**	0.502**	0.709**	0.539**
	(0.039)	(0.017)	(0.025)	(0.012)
Past forest	-0.102	-0.125**	-0.111	-0.124**
	(0.096)	(0.037)	(0.062)	(0.028)
Sargan	0.424	1.479	0.001	0.506
(p-value)	(0.515)	(0.224)	(0.979)	(0.477)
No. of obs.	281	1,558	353	1,621

Notes: Figures and those in parentheses indicate the mean and standard deviation of posterior distribution for each variable, respectively; constant is not reported; asterisks * and ** indicate that the 95% and 99% credible intervals of the posterior distribution do not contain zero, respectively.

Table 6. Estimation Results for the Share of Formal Sector

Dependent variable: regional productivity

	(1)	(2)	(3)	(4)
	Manufacturing		Wholesale/Retail	
Independent variable	Formal	Informal	Formal	Informal
Density of employment	0.047	0.109**	0.053	0.055**
	(0.071)	(0.025)	(0.049)	(0.017)
Density of employment × Formal share	-0.152*	-0.157**	-0.024	-0.098
	(0.069)	(0.031)	(0.097)	0.056
Population size	0.387**	0.406**	0.211*	0.142**
	(0.103)	(0.038)	(0.092)	(0.031)
Industrial diversity	0.386	0.131	0.090	0.188**
	(0.233)	(0.090)	(0.121)	(0.039)
Local competition	-0.145*	-0.207**	-0.080	-0.076**
	(0.065)	(0.027)	(0.054)	0.021
Electricity access	0.480	0.767**	0.436*	0.841**
	(0.254)	(0.110)	(0.219)	(0.082)
High skilled labor	0.484	-0.897	-0.427	0.164
	(0.948)	(0.605)	(0.985)	(0.581)
Low skilled labor	0.179	-0.110	0.898*	-0.371**
	(0.433)	(0.126)	(0.350)	(0.087)
Spatial lag of productivity	-0.412	0.643**	0.157	0.802**
	(0.368)	(0.174)	(0.315)	(0.123)
Instrumental variable				
Past density of employment	0.524**	0.500**	0.674**	0.531**
	(0.037)	(0.018)	(0.026)	(0.013)
Past forest	-0.110	-0.126**	-0.114	-0.122**
	(0.087)	(0.040)	(0.061)	(0.029)
Sargan	0.461	1.359	0.001	0.535
(p-value)	(0.497)	(0.244)	(0.977)	(0.464)
No. of obs.	281	1,558	353	1,621

Notes: Figures and those in parentheses indicate the mean and standard deviation of posterior distribution for each variable, respectively; constant is not reported; asterisks * and ** indicate that the 95% and 99% credible intervals of the posterior distribution do not contain zero, respectively; formal share shows the share of employment by registered firms in each region.

Table 7. Estimation Results for Nonlinear Agglomeration Effects

Dependent variable: regional productivity

	(1)	(2)	(3)	(4)
	Manufacturing		Wholesale/Retail	
Independent variable	Formal	Informal	Formal	Informal
Density of employment	0.184	0.102**	0.073	0.049*
	(0.135)	(0.023)	(0.092)	(0.024)
Density of employment squared	-0.028*	-0.012**	-0.003	0.001
	(0.014)	(0.003)	(0.009)	(0.003)
Population size	0.344**	0.361**	0.202*	0.139**
	(0.103)	(0.038)	(0.094)	(0.032)
Industrial diversity	0.298	0.077	0.089	0.189**
	(0.240)	(0.091)	(0.121)	(0.039)
Local competition	-0.137*	-0.196**	-0.074	-0.068**
	(0.066)	(0.025)	(0.051)	(0.021)
Electricity access	0.251	0.679**	0.430	0.833**
	(0.295)	(0.112)	(0.222)	(0.082)
High skilled labor	1.829*	0.032	-0.380	-0.500
	(0.917)	(0.619)	(0.936)	(0.541)
Low skilled labor	-0.315	-0.178	0.865*	-0.346**
	(0.561)	(0.131)	(0.379)	(0.099)
Spatial lag of productivity	-0.409	0.639**	0.159	0.790**
	(0.373)	(0.174)	(0.315)	(0.128)
Instrumental variable				
Past density of employment	0.311**	0.534**	0.449**	0.426**
	(0.044)	(0.021)	(0.026)	(0.013)
Past forest	-0.173	-0.147**	-0.084	-0.090**
	(0.089)	(0.045)	(0.048)	(0.026)
Sargan	0.065	1.579	0.001	0.491
(p-value)	(0.798)	(0.209)	(0.976)	(0.484)
No. of obs.	281	1,558	353	1,621

Notes: Figures and those in parentheses indicate the mean and standard deviation of posterior distribution for each variable, respectively; constant is not reported; asterisks * and ** indicate that the 95% and 99% credible intervals of the posterior distribution do not contain zero, respectively.

Appendix A

In Appendix A, we describe the derivation of full conditional posterior distributions for the parameters $\boldsymbol{\beta}^*$, $\boldsymbol{\gamma}^*$, $\boldsymbol{\Sigma}$, and ρ.

A1. Full Conditional Posterior of $\boldsymbol{\beta}^*$

Given the parameter $\boldsymbol{\gamma}^*$, we can observe $\boldsymbol{\eta}$. A structural equation conditional on $\boldsymbol{\eta}$ is written as:

$$\boldsymbol{S}\boldsymbol{y} = \boldsymbol{x}\beta_0 + \boldsymbol{Z}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \mid \boldsymbol{\eta}. \tag{A1}$$

The expectation and variance of $\boldsymbol{\varepsilon} \mid \boldsymbol{\eta}$ are given by:

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

where σ_{ij} is the (i,j)th element of $\boldsymbol{\Sigma}$. Equation (A1) can be rewritten as:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\beta^* + \boldsymbol{\xi}. \tag{A2}$$

where $\boldsymbol{y}^* \equiv \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}\big(\boldsymbol{S}\boldsymbol{y} - \boldsymbol{\mu}_{\varepsilon|\eta}\big)$, $\boldsymbol{X}^* \equiv \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2}[\boldsymbol{x}, \boldsymbol{Z}]$, $\beta^* \equiv [\beta_0, \boldsymbol{\beta}_1]'$, and $\boldsymbol{\xi} \sim MVN(\mathbf{0}, \mathbf{I}_n)$. Using equation (A2) and the prior of $\boldsymbol{\beta}^*$, we obtain the full conditional multivariate normal distribution of $\boldsymbol{\beta}^*$:

$$\boldsymbol{\beta}^* \mid \boldsymbol{\gamma}^*, \rho, \boldsymbol{\Sigma}, \text{Data} \sim MVN\big(\widehat{\boldsymbol{b}}_\beta, \widehat{\boldsymbol{B}}_\beta\big)$$

where $\widehat{\boldsymbol{B}}_\beta = \big[\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{B}_\beta^{-1}\big]^{-1}$ and $\widehat{\boldsymbol{b}}_\beta = \widehat{\boldsymbol{B}}_\beta\big[\boldsymbol{X}^{*\prime}\boldsymbol{y}^* + \boldsymbol{B}_\beta^{-1}\boldsymbol{b}_\beta\big]$.

A2. Full Conditional Posterior of $\boldsymbol{\gamma}^*$

Substituting and rearranging the system of equations gives us the following:

$$\begin{pmatrix} \boldsymbol{X} \\ \frac{\boldsymbol{S}\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}_1}{\beta_0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{Z} \\ \boldsymbol{Q} & \boldsymbol{Z} \end{pmatrix}\boldsymbol{\gamma}^* + \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\eta} + \frac{\boldsymbol{\varepsilon}}{\beta_0} \end{pmatrix}. \tag{A3}$$

The covariance matrix of $\boldsymbol{\gamma}^*$ is given by:

$$V\begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\eta} + \dfrac{\boldsymbol{\varepsilon}}{\beta_0} \end{pmatrix} \equiv \boldsymbol{\Omega} = [\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'] \otimes \mathbf{I}_n$$

where $\mathbf{A} \equiv \begin{bmatrix} 1 & 0 \\ 1 & 1/\beta_0 \end{bmatrix}$. Multiplying both sides of equation (A3) by $\boldsymbol{\Omega}^{-1/2}$ yields:

$$\boldsymbol{\Omega}^{-1/2}\begin{pmatrix} \boldsymbol{X} \\ \frac{\boldsymbol{S}\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}_1}{\beta_0} \end{pmatrix} = \boldsymbol{\Omega}^{-1/2}\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{Z} \\ \boldsymbol{Q} & \boldsymbol{Z} \end{pmatrix}\boldsymbol{\gamma}^* + \boldsymbol{\zeta} \tag{A4}$$

where $\boldsymbol{\zeta} \sim MVN(\mathbf{0}, \mathbf{I}_{2n})$. Using this equation and the prior of $\boldsymbol{\gamma}^*$, we obtain the full conditional multivariate normal distribution:

$$\boldsymbol{\gamma}^* \mid \boldsymbol{\beta}^*, \rho, \boldsymbol{\Sigma}, \text{Data} \sim MVN\big(\widehat{\boldsymbol{b}}_\gamma, \widehat{\boldsymbol{B}}_\gamma\big)$$

where $\widehat{\boldsymbol{B}}_\gamma = \big[\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^* + \boldsymbol{B}_\gamma^{-1}\big]^{-1}$ and $\widehat{\boldsymbol{b}}_\gamma = \widehat{\boldsymbol{B}}_\gamma\big[\boldsymbol{Z}^{*\prime}\boldsymbol{y}^+ + \boldsymbol{B}_\gamma^{-1}\boldsymbol{b}_\gamma\big]$. Note that $\boldsymbol{y}^+ \equiv \boldsymbol{\Omega}^{-1/2}\begin{pmatrix} \boldsymbol{X} \\ \frac{\boldsymbol{S}\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}_1}{\beta_0} \end{pmatrix}$ and $\boldsymbol{Z}^* \equiv \boldsymbol{\Omega}^{-1/2}\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{Z} \\ \boldsymbol{Q} & \boldsymbol{Z} \end{pmatrix}$.

A3. Full Conditional Posterior of $\boldsymbol{\Sigma}$

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

$$\boldsymbol{\Sigma} \mid \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \rho, \text{Data} \sim IW\big(\hat{b}_\Sigma, \widehat{\boldsymbol{B}}_\Sigma\big)$$

where $\hat{b}_\Sigma = n + b_\Sigma$ and $\widehat{\boldsymbol{B}}_\Sigma = [\mathbf{E} + \mathbf{B}_\Sigma^{-1}]^{-1}$. Note that $\mathbf{E} \equiv \begin{bmatrix}\boldsymbol{\eta}' \\ \boldsymbol{\varepsilon}'\end{bmatrix}[\boldsymbol{\eta} \quad \boldsymbol{\varepsilon}]$, where $\boldsymbol{\eta} = \boldsymbol{x} - \boldsymbol{Q}\boldsymbol{\gamma}_0 - \boldsymbol{Z}\boldsymbol{\gamma}_1$ and $\boldsymbol{\varepsilon} = \boldsymbol{S}\boldsymbol{y} - \boldsymbol{x}\beta_0 - \boldsymbol{Z}\boldsymbol{\beta}_1$.

A4. Full Conditional Posterior of ρ

We reformulate equation (A2) as follows:

$$\widetilde{\boldsymbol{y}} = \rho\widetilde{\boldsymbol{X}} + \boldsymbol{\xi}. \tag{A5}$$

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

$$P(\ \rho|\ \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\Sigma}, \text{Data}) \propto |\boldsymbol{I}_n - \rho\boldsymbol{W}| \exp\left\{-\frac{1}{2}\left[\widetilde{\boldsymbol{y}} - \rho\widetilde{\boldsymbol{X}}\right]'\left[\widetilde{\boldsymbol{y}} - \rho\widetilde{\boldsymbol{X}}\right]\right\} I[\rho \in (\lambda_{min}^{-1}, \lambda_{max}^{-1})]$$
$$\propto |\boldsymbol{I}_n - \rho\boldsymbol{W}| \exp\left\{-\frac{1}{2\hat{\sigma}_\rho^2}(\rho - \hat{\rho})^2\right\} I[\rho \in (\lambda_{min}^{-1}, \lambda_{max}^{-1})]$$

where $\hat{\sigma}_\rho^2 = \left[\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right]^{-1}$ and $\hat{\rho} = \hat{\sigma}_\rho^2\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{y}}$. Note that $I[\rho \in (\lambda_{min}^{-1}, \lambda_{max}^{-1})]$ is an indicator function that takes on a value of 1 for $\rho \in (\lambda_{min}^{-1}, \lambda_{max}^{-1})$. Because this density function is not standard, we use the Metropolis-Hastings (MH) technique (Gamerman and Lopes, 2006, Chapters 5 and 6). The candidate generating function used in the *MH* algorithm is a normal distribution truncated on the interval $(\lambda_{min}^{-1}, \lambda_{max}^{-1})$ with the mean of $\hat{\rho}$ and the variance of $\hat{\sigma}_\rho^2$, which is denoted as $TN(\hat{\rho}, \hat{\sigma}_\rho^2)$.

Appendix B

In Appendix B, we describe the setup of Monte Carlo simulation. We generate a dataset for the following model:

$$x_i = q_{1i}\gamma_{01} + q_{2i}\gamma_{02} + q_{3i}\gamma_{03} + z_{1i}\gamma_{11} + z_{2i}\gamma_{12} + z_{3i}\gamma_{13} + \eta_i \tag{B1}$$
$$y_i = \rho\sum_{j=1}^{n} w_{ij}y_j + x_i\beta_0 + z_{1i}\beta_{11} + z_{2i}\beta_{12} + z_{3i}\beta_{13} + \varepsilon_i \tag{B2}$$

where $\begin{pmatrix}\eta_i \\ \varepsilon_i\end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ and $z_{1i} = 1$. We assume that z_{2i} and z_{3i} follow a standard normal distribution $N(0,1)$. The element w_{ij} is constructed using a dataset on the minimum travel time between 1,621 communes in Cambodia. For the case of strong instruments, we construct a dataset with the parameters as $\gamma_{01} = \gamma_{02} = \gamma_{03} = 4$. For weak instruments, a dataset is constructed for $\gamma_{01} = \gamma_{02} = \gamma_{03} = 0.1$. In both datasets, we set all other parameters as follows; $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$, $\gamma_{11} = \gamma_{12} = \gamma_{13} = 1$, $\rho = 0.5$, and $\boldsymbol{\Sigma} = \begin{bmatrix}1 & 0.8 \\ 0.8 & 1\end{bmatrix}$. We use relatively diffuse priors $N(0,100)$ for coefficients and $\boldsymbol{\Sigma} \sim IW(2, \mathbf{I}_2)$ for the covariance matrix. The prior of ρ is $U(\lambda_{min}^{-1}, \lambda_{max}^{-1})$. The simulation results are summarized in Table B1. Figure B1 shows the MCMC sampling path for the parameters while Figure B2 presents histograms of posterior distributions of the MCMC samples.

[Table B1, Figures B1 and B2]

Appendix C

C1. Data on Regional Productivity

Value added is computed as sales minus intermediate input for one month in February 2011. Sales include all income gained from operating activities such as selling goods and providing services. Intermediate input is computed from expenses minus wages because the expenses already include every expense being paid for operating activities such as the

purchase of materials for sale, instruments for providing services, rent, and employees' salaries and wages. To address outliers in the value added data, we exclude the observations in the top and bottom 1% of value added for each industry.[15] Since a number of firms do not report fixed assets, capital input is proxied by the area of business place in square meters where the firm operates. Additionally, labor input is the total number of persons engaged in business activities, including self-employed proprietors, unpaid family workers, regular employees who are employed on a continuous basis for more than a one-month period, and other employees.

C2. Travel Time between Commune Pairs

The shortest travel time is calculated by the Floyd-Warshall algorithm. To prepare a dataset on the geographic distances between all commune pairs, we use the GIS shape file of administrative units in Cambodia and obtain the shortest distance between these communes. We construct a dataset on travel speeds between *neighboring* communes by using the JETRO surveys that are based on a field survey of the actual status of logistics infrastructures within Cambodia.

Our assumptions about travel speeds are summarized in Table C1. For instance, if a 1-digit national road connects both communes in contiguity, we assume a speed of 90 kilometers per hour. If a 1-digit national road connects one commune and a 2-digit national road connects the neighboring commune, we assume a speed of 70 kilometers per hour. If neighboring communes are characterized by tree covered or regularly flooded areas, we assume a speed of 4 kilometers per hour. Finally, we assume a speed of 30 kilometers per hour for all the other pairs of neighboring communes.

Given these assumptions, we calculate travel time for all the neighboring communes. Because each commune must be connected to all the other communes at least indirectly through its neighboring communes, we can compute travel time for a large number of possible routes. For this task, we execute the Floyd-Warshall algorithm. For a given pair of communes i and j, we compute travel 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 travel time from commune i to commune j via commune k, $T_{ik} + T_{kj}$. 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 travel time to be 10,000 hours for missing observations of commune pairs in the dataset. We repeat this recursive algorithm for $k = 1,2,\cdots,N$, where N is the total number of communes.

The dataset shows that the average geographic distance among the 1,621 communes is 189.8 kilometers with a standard deviation of 109.5, and the average travel time is 6.72 hours with a standard deviation of 4.8. The correlation coefficient of time and distance is 0.61, implying that the relationship between travel time and geographic distance are not perfectly linear.

[Table C1]

Reference

Gamerman, D., and Lopes, H. F., 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd Edition. Chapman and Hall/CRC, Boca Raton.

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.

[15] For measurement issues about business activity in developing economies, see de Mel et al. (2009).

Table B1. Results of Monte Carlo Simulation

Panel A: Strong Instruments					
Parameter	True value	95%L	Median	95%U	Std. Dev.
β_0	1	0.987	0.994	1.002	0.004
β_1	1	0.737	0.956	1.157	0.106
β_2	1	0.965	1.016	1.066	0.026
β_3	1	0.980	1.032	1.085	0.027
ρ	0.5	0.459	0.507	0.559	0.026
σ_{11}	1	0.981	1.050	1.128	0.038
σ_{12}	0.8	0.779	0.843	0.913	0.034
σ_{22}	1	0.963	1.033	1.110	0.037
Panel B: Weak Instruments					
Parameter	True value	95%L	Median	95%U	Std. Dev.
β_0	1	0.515	0.836	1.091	0.148
β_1	1	0.485	1.061	1.727	0.319
β_2	1	0.909	1.180	1.518	0.154
β_3	1	0.929	1.191	1.519	0.152
ρ	0.5	0.370	0.520	0.679	0.075
σ_{11}	1	0.980	1.050	1.127	0.038
σ_{12}	0.8	0.738	1.011	1.356	0.159
σ_{22}	1	0.881	1.331	2.085	0.311

Note: 95%L and 95%U indicate the lower and upper bounds of the 95% credible interval, respectively.

Table C1. Assumptions on Travel Speed in Kilometers per Hour between Commune Pairs

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

Note: Figures show travel speed in kilometers per hour for travel between neighboring communes with corresponding characteristics.

Figure B1. MCMC Sampling Path

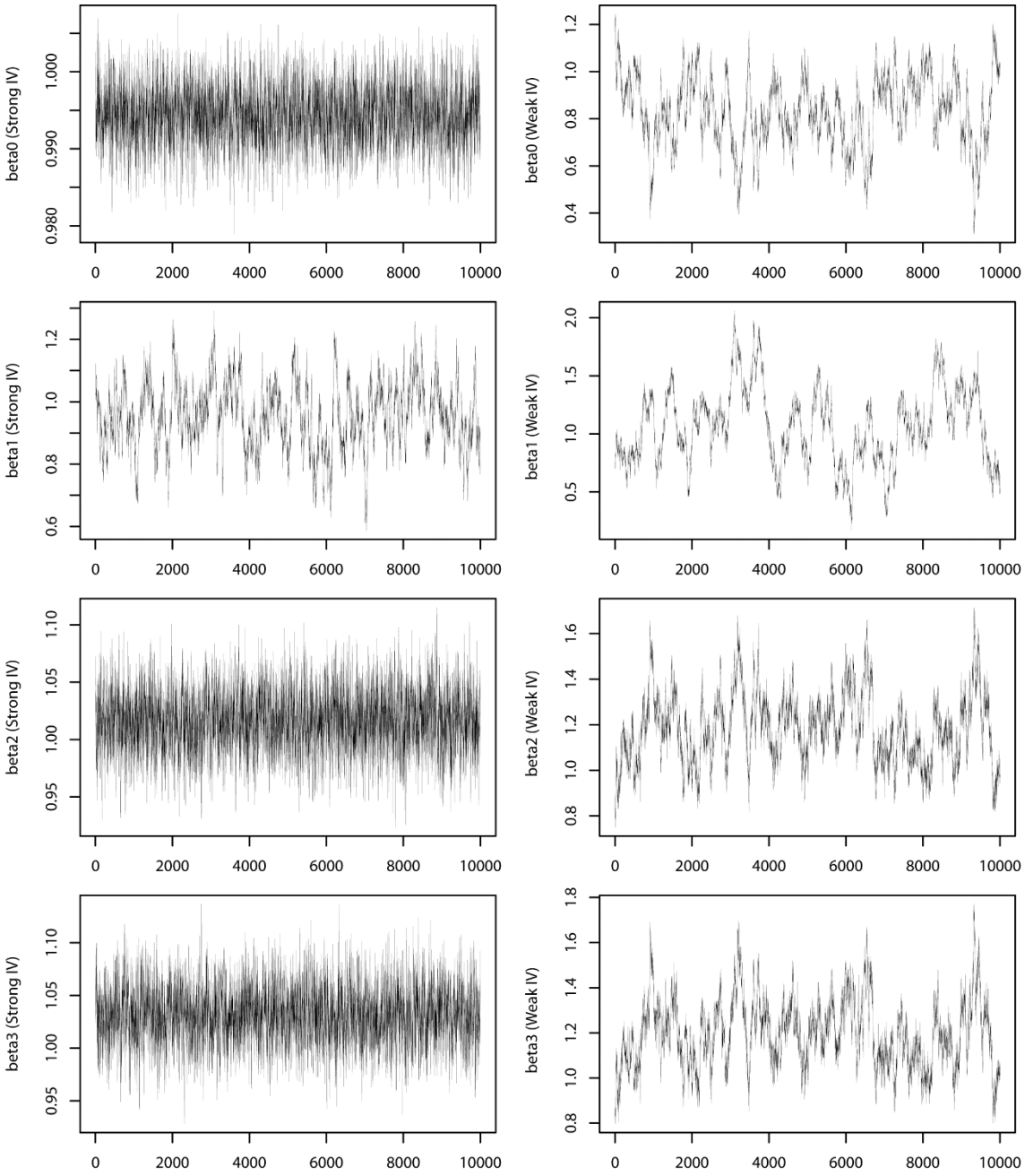

Figure B2. Histograms of MCMC Samples