Agglomeration Effects in Cambodia*

Yoshihiro HASHIGUCHI[†]

Institute of Developing Economies

Kiyoyasu TANAKA[‡]

Institute of Developing Economies

March 2013

Abstract

We consider a formal framework to shed light on agglomeration effects in Cambodia. Using the new census dataset on all the Cambodia establishments in 2011, we describe the current patterns of industrial agglomeration in manufacturing and wholesale/retail sectors. As these sectors clearly exhibit a spatial concentration of economic activity with varying degrees across sectors, the descriptive analysis provides a motivation for investigating whether the observed patterns of agglomeration would yield productivity gains. To make a formal assessment, we develop a Bayesian spatial approach and address econometric issues such as spatial autocorrelation between nearby regions and endogeneity of agglomeration. Simulation analysis of our spatial model shows that if strong instruments are available, the framework enables us to identify the impact of agglomeration economies with precision. Thus, our next step is to find a set of plausible instruments to conduct an econometric analysis and derive policy implications for Cambodia.

Keywords: Agglomeration, spillovers, Cambodia

JEL classification: C21, F21, F23, R12, R58

*This paper is written under the project "Multinational Firms and the Globalization of Developing Economies" funded by the Institute of Developing Economies. We would like to thank Fumio Nishi for data assistance. All remaining errors are our own.

[†]Development Studies Center, Institute of Developing Economies; address: 3-2-2 Wakaba, Mihama-ku, Chiba-shi, Chiba 261-8545, Japan. e-mail: yoshihiro_hashiguchi@ide.go.jp

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

1 Introduction

A spatial concentration of economic activity has crucial implications for economic growth. Williamson (1965) argues that agglomeration favors economic growth at an early stage of economic development because limited resources such as capital, human capital, and infrastructure can be most efficiently utilized in an agglomerated area. Fujita and Thisse (2003) illustrates that agglomeration can promote growth in a two-region model of endogenous growth. Indeed, it has long been argued that firms and workers locate in the agglomerated area to benefit from productivity advantages generated by agglomeration economies through 1) more efficient sharing of local suppliers, 2) better matching between employers and workers, and 3) technology or knowledge spillovers among firms and workers (Duranton and Puga, 2004).

However, a spatial concentration of economic activity could grow excessively throughout the process of economic development. Agglomeration starts to bring about heavy congestion and raise factor prices such as wages and land prices. Congestion effects reduce the productivity advantage from agglomeration economies. These offsetting forces make it important to examine agglomeration effects and draw policy implications especially for developing economies. Indeed, it is emphasized that "for policy makers the challenge is to best relax the constraints generated by the congestion and overcrowding of land and resources so that the benefits of agglomeration can be maximized (World Development Report 2009, p. 144)." Thus, a quantitative assessment of agglomeration effects is crucial to evaluate regional policy for maximizing the benefit of agglomeration effects.

In this paper, we develop a Bayesian spatial approach to estimate agglomeration effects in Cambodia. After decades of civil war devastated the Cambodian economy, the reunification of the nation set a stage for rapid progress in economic reconstruction. While the economic growth has averaged 6 percent for the last 10 years, the gross domestic product per capita was merely 931.2 U.S. dollars in 2012 (World Economic Outlook 2012, IMF). As the Cambodian economy is constrained by the limited resources in capita, human capital, and infrastructure, it is a crucial policy issue for its government to investigate the extent to which economic agglomeration should be promoted to maximize economic growth. Moreover, the previous studies on agglomeration economies have focused on high and middle income countries (Melo et al, 2009). Our investigation on low income countries such as Cambodia improves our understanding of the economic magnitude of agglomeration effects across different stages of economic development.

The rest of this paper is as follows. Section 2 provides a description of the data used in this analysis and a simple illustration of a spatial concentration of economic activity. Section 3 explains our empirical framework to estimate agglomeration effects, with an emphasis on spatial autocorrelation and endogeneity issues. Section 4 discusses a current agenda for future research.

2 Data Description

For the analysis, we use *the Economic Census* of Cambodia in 2011(EC2011). The census was conducted to survey economic activities for all the establishments and enterprises over the entire territory of Cambodia in March 2011. The EC2011 was mainly funded by the Japanese ODA and implemented by the National Institute of Statistics, the Cambodian Ministry of Planning, in cooperation with the Japanese government.¹ The survey aims to collect the basic information on firm activities, including area of business place, ownership status, main business activities, employment, establishment year, and so on. The administrative geographic units consist of 1,621 communes in 24 provinces including the Municipality of Phnom Penh. There were total 505,134 establishments in 2011.

Using the EC2011 data, we construct a measure of regional labor productivity as a dependent variable in our specification. As we focus exclusively on crosssectional variation in our dataset, we simply require a measure of productivity that indicates a relative ranking of one region over another at a point in time. To capture this relative efficiency ranking, regional labor productivity is defined as $\ln(q_i/L_i)$ for region *i*, where q_i is the value added of region *i* and L_i is the amount of employment in that region.

We separately estimate regional labor productivity in manufacturing and wholesale/retail sectors as these sectors provide the largest number of observations with complete information on productivity estimates. To visualize the regional estimates, Figures 1 and 2 show regional variation in labor productivity in the geographical map of the Cambodian land for manufacturing and wholesale/retail

¹Details of EC2011 are found at the Japanese government's website: http://www.stat.go.jp/english/info/meetings/cambodia/census11.htm

sectors, respectively. The darker regions indicate relatively high labor productivity. It is apparent that regional labor productivity has substantial variation over space for these sectors.

[Figures 1 and 2 here]

There are a wide variety of potential determinants of the regional productivity such as local natural advantages. Local endowments such as natural resources, economic infrastructure, geographic environments provide an explanation for accounting for the observed variation in regional labor productivity. In this paper, we attempt to shed light on the role of economic agglomeration in shaping the spatial variation of labor productivity. Thus, we aggregate the number of workers in each commune as a measure of economic agglomeration. Figures 3 and 4 show the results in manufacturing and wholesale/retail sectors, respectively. Both figures point to a large variation in regional employment over space. A possible difference is that employment in manufacturing sectors appears to be more concentrated over space than that in wholesale/retail sectors. Because increasing returns should play a larger role in manufacturing, these observations are sensible from the perspective of economic geography model.

[Figures 3 and 4 here]

3 Econometric Framework

3.1 Literature Review

Over the past thirty years, many scholars have attempted to estimate the magnitude of agglomeration economies in various countries.² Sveikauskas (1975) is one of the early and influential studies, who estimated the impact of city size measured by population on labor productivity to measure the degree of agglomeration economies, and found that a doubling city size increased labor productivity by about 6%. Although he provided a valuable empirical approach, some studies such as Moomaw (1981) have pointed out that the regression analysis may have

²For the details, refer to the comprehensive reviews provided by Eberts and McMillen (1999, section 3), Rosenthal and Strange (2004, section 2), Graham (2008, 65-67), Cohen and Paul (2009), Broersma and Oosterhaven (2009, 487-489), and Puga (2010).

suffered from the endogeneity problem due to the existence of omitted variables and simultaneity bias.³

Ciccone and Hall (1996) and Ciccone (2002), which are seminal works in recent years, offered a model based on regional production function to evaluate agglomeration economies. They applied the two-stage least squares (2SLS) methods to their model to deal with the endogeneity. Ciccone and Hall (1996) used the data of US states in 1988, and Ciccone (2002) applied their model to the cross section data of EU NUTS-3 regions, which include Germany, France, Italy, Spain and the UK. Both of these studies found that a doubling of employment density increased average labor productivity by around 5-6%. Since these seminal works, the model and empirical approach proposed by Ciccone and Hall has been applied by many researchers for many countries or regions. For example, focusing on the studies in recent years, Brülhart and Mathys (2008) extended Ciccone-Hall model to a dynamic panel data model, and estimated the impact of employment density on labor productivity in 245 NUTS-2 regions of 20 European countries from 1980 to 2003, using generalized method of moments (GMM). According to their results, the estimate of the long run elasticity of labor productivity for aggregate industrial sector was 0.13 and significant at the 5%. Broersma and Oosterhaven (2009) applied Ciccone-Hall model to the panel data of 40 Dutch regions from 1990 to 2001, and found that the estimate of the elasticity of labor productivity was 0.033 and significant.

Most of previous studies have observed the existence of agglomeration economies, and have indicated that a doubling employment density increased labor productivity by approximately 3-10%.

Recently, Artis et al. (2012) pointed out that estimate of agglomeration effects is remarkably reduced when spatial autocorrelation is controlled. They used the data of 119 NUTS-3 regions of Great Britain, and adopted a spatial autoregressive model with autoregressive disturbances to deal with the spatial autocorrelation. The spatial model was estimated by feasible generalized spatial 2SLS (FGS2SLS), proposed by Kelejian and Prucha (1998). As a result, the spatial autocorrelation was positive and significant, and the estimates of the elasticity of labor productivity were 0.021-0.024, which were much smaller than their GMM

³For detailed discussion about the endogeneity problem in this literature, refer to Eberts and McMillen (1999), Rosenthal and Strange (2004), Cohen and Paul (2009), and Puga (2010).

estimates (0.039-0.056) which did not take into account the spatial autocorrelation. Their findings imply that estimation of agglomeration effects can lead to spurious conclusions if they ignore the significant degree of spatial autocorrelation.

In this section, we propose a different approach to dealing with the problems of both the endogeneity and the spatial autocorrelation. Our approach is based on a Bayesian Instrumental Variables (IV) method proposed by Rossi, at el. (2005). We apply Rossi's method to the spatial autoregressive model (Anselin 1988; 2001).

3.2 Bayesian Approach to Spatial Autoregressive Model with an Endogenous Regressor

3.2.1 Model

Let us consider a spatial autoregressive model to estimate the impact of agglomeration on regional productivity, such as

$$y_i = x_i \beta_0 + \mathbf{z}_i \beta_1 + \rho \sum_{j=1}^n w_{ij} y_j + \varepsilon_i, \quad i = 1, 2, \dots, n$$
(1)

where *i* denotes a region, y_i and x_i represent regional productivity level and the degree of industrial agglomeration, respectively, and ε_i is an error term. w_{ij} is a variable which shows the geographical relationship between regions *i* and *j*, and specified as

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

where d_{ij} is the traveling time between *i* and *j*. The parameters β_0 and ρ indicate the magnitude of agglomeration effects and spatial autocorrelation in y_i . We consider that \mathbf{z}_i is a vector of exogenous (or predetermined) variables, and x_i is an endogenous variable and has a linear relationship with a set of instruments ($\mathbf{q}_i, \mathbf{z}_i$) and an idiosyncratic shock η_i , where \mathbf{q}_i is a vector of variables related to x_i but independent of the error terms ε_i and η_i . Following Rossi et al. (2005), we specify the system of equations as follows:

$$x_i = \mathbf{q}_i \boldsymbol{\gamma}_0 + \mathbf{z}_i \boldsymbol{\gamma}_1 + \eta_i \tag{3}$$

$$y_i = x_i \beta_0 + \mathbf{z}_i \beta_1 + \rho \sum_{j=1}^n w_{ij} y_j + \varepsilon_i \,. \tag{4}$$

This system indicates a structural equation (4) with an endogenous regressor and multiple instruments. If the correlation between η_i and ε_i is positive, there will be a positive endogeneity bias (Rossi et al, 2005). In vector and matrix notation, these equations can be written as

$$\mathbf{x} = \mathbf{Q} \,\boldsymbol{\gamma}_0 + \mathbf{Z} \,\boldsymbol{\gamma}_1 + \boldsymbol{\eta} \tag{5}$$

$$\mathbf{S}\mathbf{y} = \mathbf{x}\beta_0 + \mathbf{Z}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}\,,\tag{6}$$

where $\mathbf{S} = \mathbf{I}_n - \rho \mathbf{W}$, $\mathbf{x} = (x_1, \dots, x_n)'$, $\mathbf{y} = (y_1, \dots, y_n)'$, $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$, $\mathbf{Q} = (\mathbf{q}_1, \dots, \mathbf{q}_n)'$, and $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)'$. We use this system to estimate an impact of agglomeration on regional productivity.

3.2.2 The Likelihood and Priors

To derive the likelihood function, we assume that η and ε have a multivariate normal distribution:

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

Let $\tilde{\mathbf{y}} = (\mathbf{x}, \mathbf{y})'$ and \mathbf{u} denote a $(2n \times 1)$ vector which follows a multivariate standard normal distribution $N(\mathbf{0}, \mathbf{I}_{2n})$. Equations (5) and (6) can be rewritten as

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

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

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

And then the likelihood function can be obtained as follows:

$$L = (2\pi)^{\frac{n}{2}} |\mathbf{\Sigma}|^{\frac{n}{2}} |\mathbf{S}| \exp\left\{ \begin{bmatrix} \mathbf{x} - \mathbf{Q} \,\boldsymbol{\gamma}_0 - \mathbf{Z} \,\boldsymbol{\gamma}_1 \\ \mathbf{S} \mathbf{y} - \mathbf{x} \,\boldsymbol{\beta}_0 - \mathbf{Z} \,\boldsymbol{\beta}_1 \end{bmatrix}' [\mathbf{\Sigma} \otimes \mathbf{I}_n]^{-1} \begin{bmatrix} \mathbf{x} - \mathbf{Q} \,\boldsymbol{\gamma}_0 - \mathbf{Z} \,\boldsymbol{\gamma}_1 \\ \mathbf{S} \mathbf{y} - \mathbf{x} \,\boldsymbol{\beta}_0 - \mathbf{Z} \,\boldsymbol{\beta}_1 \end{bmatrix} \right\}, \quad (9)$$
$$\mathbf{S} = \mathbf{I}_n - \rho \mathbf{W}. \tag{10}$$

Independent priors for the unknown parameters are specified as

$$\boldsymbol{\beta}^* \equiv \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} \sim MVN(\mathbf{b}_{\boldsymbol{\beta}}, \, \mathbf{B}_{\boldsymbol{\beta}}) \,, \quad \boldsymbol{\gamma}^* \equiv \begin{pmatrix} \boldsymbol{\gamma}_0 \\ \boldsymbol{\gamma}_1 \end{pmatrix} \sim MVN(\mathbf{b}_{\boldsymbol{\gamma}}, \, \mathbf{B}_{\boldsymbol{\gamma}}) \,,$$

$$\rho \sim U(\lambda_{\min}^{-1}, \lambda_{\max}^{-1}) \,, \quad \boldsymbol{\Sigma} \sim IW(b_{\boldsymbol{\Sigma}}, \mathbf{B}_{\boldsymbol{\Sigma}}) \,,$$
(11)

where IW() and U() denote the inverted Wishart distribution and the uniform distribution. The prior parameters are \mathbf{b}_{β} , \mathbf{B}_{β} , \mathbf{b}_{γ} , \mathbf{B}_{γ} , λ_{\min} , λ_{\max} , b_{Σ} , and \mathbf{B}_{Σ} . The λ_{\min} and λ_{\max} are the minimum and maximum eigenvalue of \mathbf{W} , and we put a limit on the parameter space of ρ such as $\rho \in (\lambda_{\min}^{-1}, \lambda_{\max}^{-1})$. If a vector of eigenvalues of \mathbf{W} contains only real values, this restriction ensures $|\mathbf{S}| > 0$.

3.2.3 MCMC algorithm

Bayesian inference is based on the posterior distributions of unknown parameters. The Markov Chain Monte Carlo (MCMC) methods enable us to generate samples from the posteriors and to draw statistical inference using the simulated samples. Our MCMC sampling is based on the following *full* conditional posterior distributions:

$$\beta^* \mid \gamma^*, \rho, \Sigma, \text{Data}$$
$$\gamma^* \mid \beta^*, \rho, \Sigma, \text{Data}$$
$$\Sigma \mid \beta^*, \gamma^*, \rho, \text{Data}$$
$$\rho \mid \beta^*, \gamma^*, \Sigma, \text{Data}$$

where Data = { $\mathbf{x}, \mathbf{y}, \mathbf{Q}, \mathbf{W}, \mathbf{Z}$ }. Using Equations (9)–(11), we can derive these full conditionals as follows.

Full conditional posterior of β^*

Given the parameter γ^* , η can be observed, and consequently, Equation (6) conditional on η is

$$\mathbf{S}\mathbf{y} = \mathbf{x}\beta_0 + \mathbf{Z}\beta_1 + \boldsymbol{\varepsilon} \mid \boldsymbol{\eta}, \tag{12}$$

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

$$E(\boldsymbol{\varepsilon} \mid \boldsymbol{\eta}) \equiv \boldsymbol{\mu}_{\varepsilon \mid \eta} = E(\boldsymbol{\varepsilon}) + (\sigma_{12}\mathbf{I}_n)(\sigma_{11}\mathbf{I}_n)^{-1}(\boldsymbol{\eta} - E(\boldsymbol{\eta}))$$

$$= (\sigma_{12}\mathbf{I}_n)(\sigma_{11}\mathbf{I}_n)^{-1}\boldsymbol{\eta}$$
(13)
$$V(\boldsymbol{\varepsilon} \mid \boldsymbol{\eta}) \equiv \boldsymbol{\Sigma}_{\varepsilon \mid \eta} = (\sigma_{22}\mathbf{I}_n) - (\sigma_{12}^2\mathbf{I}_n)(\sigma_{11}\mathbf{I}_n)^{-1},$$

where σ_{ij} is a (i, j)th element of Σ . Equation (12) can be rewritten as

$$\Sigma_{\varepsilon|\eta}^{-1/2} \left(\mathbf{S} \mathbf{y} - \boldsymbol{\mu}_{\varepsilon|\eta} \right) = \Sigma_{\varepsilon|\eta}^{-1/2} \begin{bmatrix} \mathbf{x} & \mathbf{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{bmatrix} + \boldsymbol{\xi},$$

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta}^* + \boldsymbol{\xi},$$
 (14)

where $\boldsymbol{\xi} \sim MVN(\mathbf{0}, \mathbf{I}_n), \, \mathbf{y}^* \equiv \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2} (\mathbf{S}\mathbf{y} - \boldsymbol{\mu}_{\varepsilon}), \text{ and } \mathbf{X}^* \equiv \boldsymbol{\Sigma}_{\varepsilon|\eta}^{-1/2} [\mathbf{x}, \mathbf{Z}].$ Using Equation (14) and the prior of $\boldsymbol{\beta}^*$ yields the *full* conditional distribution of $\boldsymbol{\beta}^*$:

$$\boldsymbol{\beta}^{*} \mid \boldsymbol{\gamma}^{*}, \boldsymbol{\rho}, \boldsymbol{\Sigma}, \text{Data} \sim MVN(\hat{\mathbf{b}}_{\beta}, \hat{\mathbf{B}}_{\beta})$$
$$\hat{\mathbf{B}}_{\beta} = \left[\mathbf{X}^{*'} \mathbf{X}^{*} + \mathbf{B}_{\beta}^{-1}\right]^{-1}$$
$$\hat{\mathbf{b}}_{\beta} = \hat{\mathbf{B}}_{\beta} \left[\mathbf{X}^{*'} \mathbf{y}^{*} + \mathbf{B}_{\beta}^{-1} \mathbf{b}_{\beta}\right].$$
(15)

Full conditional posterior of γ^*

Substituting Equation (5) for (6) and reformulating these equations, we can obtain the following equation:

$$\begin{pmatrix} \mathbf{x} \\ \underline{\mathbf{s}}_{\mathbf{y}-\mathbf{Z}\boldsymbol{\beta}_1} \\ \beta_0 \end{pmatrix} = \begin{pmatrix} \mathbf{Q} & \mathbf{Z} \\ \mathbf{Q} & \mathbf{Z} \end{pmatrix} \boldsymbol{\gamma}^* + \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\eta} + \frac{\boldsymbol{\varepsilon}}{\beta_0} \end{pmatrix}.$$
(16)

The covariance matrix of η^* is

$$V\begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\eta} + \boldsymbol{\varepsilon}/\beta_0 \end{pmatrix} \equiv \boldsymbol{\Omega} = [\mathbf{A} \boldsymbol{\Sigma} \mathbf{A}'] \otimes \mathbf{I}_n, \qquad (17)$$
$$\mathbf{A} \equiv \begin{bmatrix} 1 & 0 \\ 1 & 1/\beta_0 \end{bmatrix}.$$

Multiplying both sides of Equation (16) by $\Omega^{-1/2}$ yields

$$\mathbf{\Omega}^{-\frac{1}{2}}\begin{pmatrix}\mathbf{x}\\\underline{\mathbf{s}\,\mathbf{y}-\mathbf{Z}\,\boldsymbol{\beta}_{1}}\\\beta_{0}\end{pmatrix} = \mathbf{\Omega}^{-\frac{1}{2}}\begin{pmatrix}\mathbf{Q}&\mathbf{Z}\\\mathbf{Q}&\mathbf{Z}\end{pmatrix}\boldsymbol{\gamma}^{*} + \boldsymbol{\zeta}\,,\quad \boldsymbol{\zeta}\sim MVN(\mathbf{0},\,\mathbf{I}_{2n}).$$
(18)

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

$$\boldsymbol{\gamma}^{*} \mid \boldsymbol{\beta}^{*}, \boldsymbol{\rho}, \boldsymbol{\Sigma}, \text{Data} \sim MVN(\hat{\mathbf{b}}_{\gamma}, \hat{\mathbf{B}}_{\gamma}),$$
$$\hat{\mathbf{B}}_{\gamma} = \left[\mathbf{Z}^{*'} \mathbf{Z}^{*} + \mathbf{B}_{\gamma}^{-1} \right]^{-1}$$
(19)
$$\hat{\mathbf{b}}_{\gamma} = \hat{\mathbf{B}}_{\gamma} \left[\mathbf{Z}^{*'} \mathbf{y}^{+} + \mathbf{B}_{\gamma}^{-1} \mathbf{b}_{\gamma} \right],$$

.

where

$$\mathbf{y}^{+} \equiv \mathbf{\Omega}^{-\frac{1}{2}} \begin{pmatrix} \mathbf{x} \\ \underline{\mathbf{s}} \, \mathbf{y} - \mathbf{Z} \, \boldsymbol{\beta}_1 \\ \overline{\boldsymbol{\beta}_0} \end{pmatrix}, \quad \mathbf{Z}^{*} \equiv \mathbf{\Omega}^{-\frac{1}{2}} \begin{pmatrix} \mathbf{Q} & \mathbf{Z} \\ \mathbf{Q} & \mathbf{Z} \end{pmatrix}.$$

Full conditional posterior of Σ

The full conditional posterior of Σ takes a form such as

$$\Sigma \mid \boldsymbol{\beta}^{*}, \boldsymbol{\gamma}^{*}, \boldsymbol{\rho}, \text{Data} \sim IW(\hat{b}_{\Sigma}, \hat{\mathbf{B}}_{\Sigma})$$
$$\hat{b}_{\Sigma} = n + b_{\Sigma}$$
$$\hat{\mathbf{B}}_{\Sigma} = \left[\mathbf{E} + \mathbf{B}_{\Sigma}^{-1}\right]^{-1},$$
(20)

where

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

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

Full conditional posterior of ρ

Reformulate Equation (14) such as

$$\Sigma_{\varepsilon|\eta}^{-1/2} \begin{pmatrix} \mathbf{y} - \begin{bmatrix} \mathbf{x} & \mathbf{Z} \end{bmatrix} \boldsymbol{\gamma}^* - \boldsymbol{\mu}_{\varepsilon|\eta} \end{pmatrix} = \rho \Sigma_{\varepsilon|\eta}^{-1/2} \mathbf{W} \mathbf{y} + \boldsymbol{\xi}, \\ \tilde{\mathbf{y}} = \rho \, \tilde{\mathbf{X}} + \boldsymbol{\xi}, \end{cases}$$
(21)

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

$$P(\rho \mid \boldsymbol{\beta}^{*}, \boldsymbol{\gamma}^{*}, \boldsymbol{\Sigma}, \text{Data}) \propto |\mathbf{I}_{n} - \rho \mathbf{W}| \exp\left\{-\frac{1}{2}\left[\tilde{\mathbf{y}} - \rho \tilde{\mathbf{X}}\right]' \left[\tilde{\mathbf{y}} - \rho \tilde{\mathbf{X}}\right]\right\} I[\rho \in (\lambda_{\min}^{-1}, \lambda_{\min}^{-1})]$$
$$\propto |\mathbf{I}_{n} - \rho \mathbf{W}| \exp\left\{-\frac{1}{2\hat{\sigma}_{\rho}^{2}}(\rho - \hat{\rho})^{2}\right\} I[\rho \in (\lambda_{\min}^{-1}, \lambda_{\min}^{-1})],$$
(22)

where $\hat{\sigma}_{\rho}^2 = \left[\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right]^{-1}$ and $\hat{\rho} = \hat{\sigma}_{\rho}^2 \tilde{\mathbf{X}}'\tilde{\mathbf{y}}$. $I[\rho \in (\lambda_{\min}^{-1}, \lambda_{\min}^{-1})]$ is an indicator function equal to 1 if $\rho \in (\lambda_{\min}^{-1}, \lambda_{\min}^{-1})$. Since this density function is not standard, we use the Metropolis–Hastings (MH) technique.⁴ The candidate generating function used

⁴For more details about the Metropolis–Hastings and Gibbs sampling techniques, refer to Gamerman and Lopes (2006, chapters 5 and 6).

in the MH algorism is $TN(\hat{\rho}, \hat{\sigma}_{\rho}^2)$, which is a normal distribution truncated on the interval $(\lambda_{\min}^{-1}, \lambda_{\max}^{-1})$, in which the mean is $\hat{\rho}$ and the variance is $\hat{\sigma}_{\rho}^2$.

MCMC sampling algorithm

Now we describe the MCMC sampling algorithm for our model.

MCMC sampling algorithm

- (i) Choose the arbitrary initial value for all parameters and set up r = 1, where r is the number of times of MCMC sampling.
- (ii) Repeat the following sampling:

Draw $\boldsymbol{\beta}^{*(r)}$ from $MVN(\hat{\mathbf{b}}_{\beta}, \hat{\mathbf{B}}_{\beta})$, given $\boldsymbol{\gamma}^{*(r-1)}, \boldsymbol{\Sigma}^{(r-1)}, \rho^{(r-1)}$, Data. Draw $\boldsymbol{\gamma}^{*(r)}$ from $MVN(\hat{\mathbf{b}}_{\gamma}, \hat{\mathbf{B}}_{\gamma})$, given $\boldsymbol{\beta}^{*(r)}, \boldsymbol{\Sigma}^{(r-1)}, \rho^{(r-1)}$, Data. Draw $\boldsymbol{\Sigma}^{(r)}$ from $IW(\hat{b}_{\Sigma}, \hat{\mathbf{B}}_{\Sigma})$, given $\boldsymbol{\beta}^{*(r)}, \boldsymbol{\gamma}^{*(r)}, \rho^{(r-1)}$, Data. Draw ρ' (a candidate of $\rho^{(r)}$) from $TN(\hat{\rho}, \hat{\sigma}_{\rho}^{2})$, given $\boldsymbol{\beta}^{*(r)}, \boldsymbol{\gamma}^{*(r)}, \boldsymbol{\Sigma}^{(r)}$, Data. Calculate an acceptance probability:

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

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

If r < M, set r = r + 1 and return to (ii). Otherwise, go to (iii).

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

Thus $M - M_0$ replications are retained and used for the posterior inference.

3.2.4 Monte Carlo Simulation

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

Strong instruments: $\gamma_0 = [4, 4, 4]'$								
	True value	95%L	Median	95%U	StDev.			
β_0	1	0.9869	0.9944	1.0017	0.0037			
β_1	1	0.7367	0.9563	1.1573	0.1062			
β_2	1	0.9652	1.0155	1.0661	0.0258			
β_3	1	0.9805	1.0317	1.0849	0.0266			
ho	0.5	0.4586	0.5070	0.5595	0.0259			
σ_{11}	1	0.9806	1.0502	1.1281	0.0377			
σ_{12}	0.8	0.7794	0.8432	0.9127	0.0341			
σ_{22}	1	0.9632	1.0334	1.1101	0.0375			

Table 1: Results of Simulation Analysis

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

Weak instruments: $y_0 = [0.1, 0.1, 0.1]$									
	True value	95%L	Median	95%U	StDev.				
β_0	1	0.5147	0.8358	1.0914	0.1476				
β_1	1	0.4850	1.0611	1.7271	0.3195				
β_2	1	0.9087	1.1804	1.5179	0.1536				
β_3	1	0.9294	1.1908	1.5193	0.1516				
ho	0.5	0.3705	0.5195	0.6791	0.0749				
σ_{11}	1	0.9804	1.0497	1.1272	0.0376				
σ_{12}	0.8	0.7376	1.0114	1.3557	0.1588				
σ_{22}	1	0.8806	1.3305	2.0852	0.3109				

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

The data for the two cases are all generated from

$$\begin{aligned} x_{i} &= q_{1i} \gamma_{01} + q_{2i} \gamma_{02} + q_{2i} \gamma_{03} + z_{1i} \gamma_{11} + z_{2i} \gamma_{12} + z_{3i} \gamma_{13} + \eta_{i} \\ y_{i} &= x_{i} \beta_{0} + z_{1i} \beta_{11} + z_{2i} \beta_{12} + z_{3i} \beta_{13} + \rho \sum_{j=1}^{n} w_{ij} y_{j} + \varepsilon_{i} \\ \begin{pmatrix} \eta_{i} \\ \varepsilon_{i} \end{pmatrix} \sim MVN(\mathbf{0}, \mathbf{\Sigma}) \,, \end{aligned}$$

$$(23)$$

and $z_{1i} = 1$; z_{2i} and z_{3i} follow a standard normal distribution N(0, 1). The w_{ij} is constructed by Cambodian commune-level region, and then the sample size of the artificial data set is the same as the number of Cambodian communes. The parameter settings of the data set are given as follows:

Data set 1: The case of 'strong instruments'

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

Data set 2: The case of 'weak instruments'

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

All other parameters are set up the same among the data sets 1 and 2, such as

$$\beta_{0} = \beta_{1} = \beta_{2} = \beta_{3} = 1,$$

$$\gamma_{11} = \gamma_{12} = \gamma_{13} = 1,$$

$$\rho = 0.5,$$

$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

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

Table 1 shows the results of simulation analysis. For both cases, the median is close to the true value, and the 95% credible intervals of the structural parameters contain their true value. However, the posterior distributions of the weak instruments cases have a higher dispersion for the coefficients of the structural parameters, indicating the difficulty of identification. Figures 5 and 6 also show the large dispersion of the posteriors under the weak instruments situation. In addition, Figure 5 shows that MCMC sampler under the weak instruments has higher autocorrelation, implying that more MCMC draws are needed to obtain a precise posterior distribution.

[Figures 5 and 6 here]

4 Concluding Remarks

Industrial agglomeration has an important implication for economic growth as it is likely to promote economic development through more efficient utilization of limited resources in an agglomerated area. In particular, agglomeration effects enable us to derive a policy implication for developing economies in which agglomeration economies might not be fully realized at an early stage of economic growth. In this paper, we consider a formal framework to shed light on agglomeration effects in Cambodia. To make a formal assessment, we employ the new census dataset on all the Cambodia establishments in 2011. We describe the current patterns of industrial agglomeration in manufacturing and wholesale/retail sectors. These sectors clearly exhibit a spatial concentration of economic activity with varying degrees across sectors. A descriptive analysis provides a motivation for conducting a formal econometric analysis to examine whether the observed patterns of agglomeration yield productivity gains. Furthermore, we develop a Bayesian spatial approach to address econometric issues such as spatial autocorrelation between nearby regions and endogeneity of agglomeration. Simulation analysis of our spatial model shows that if strong instruments are available, the framework enables us to identify the impact of agglomeration economies with precision. Thus, our next step is to find a set of plausible instruments to conduct an econometric analysis and derive policy implications for Cambodia.

References

- Anselin, L. (1988) *Spatial econometrics: methods and models*. Kluwer Academic Publishers, Dordrecht.
- Anselin, L. (2001) Spatial econometrics. In: Baltagi, B. (eds) A companion to theoretical econometrics. Blackwell Publishers, Malden, Mass, pp. 310– 330.
- Artis, M.J., Miguélez, E., Moreno, R. (2012) Agglomeration economies and regional intangible assets: an empirical investigation. Journal of Economic Geography 12(6), 1167–1189.
- Broersma, L., Oosterhaven, J. (2009) Regional labor productivity in the Netherlands: evidence of agglomeration and congestion effects. *Journal of Regional Science* 49, 483–511.
- Brülhart, M., Sbergami, F. (2009) Agglomeration and growth: cross-country evidence *Journal of Urban Economics* 65(1), 48–63.
- Brülhart, M., Mathys, N.A. (2008) Sectoral agglomeration economies in a panel

of European regions. Regional Science and Urban Economics 38, 348–362.

- Duranton, G., D. Puga (2004) Micro-foundations of Urban Agglomeration Economies.
 In: J.V. Henderson and J.-F. Thisse (eds) *Handbook of regional and urban economics*, vol. 4, North-Holland, New York, pp. 2063–2117.
- Ciccone, A. (2002) Agglomeration effects in Europe. *European Economic Review* 46, 213-227.
- Ciccone, A., Hall, R.E. (1996) Productivity and the density of economic activity. *American Economic Review* 86, 54–70.
- Cohen, J.P., 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.
- Doornik, J.A. (2006) *Ox: an objected-oriented matrix programming language*. Timberlake Consultants, London.
- Eberts, R.W., McMillen, D.P. (1999) Agglomeration economies and urban public infrastructure. In: Cheshire, P., Mills, E.S. (eds) *Handbook of regional and urban economics*, volume 3. North-Holland, New York, pp. 1455–1495.
- Fujita, M., Thisse, J.-F. (2003) Does geographical agglomeration foster economic growth? and who gains and loses from it? *Japanese Economic Review* 54(2), 121–145.
- Graham, D.J. (2008) Identify urbanisation and localisation externalities in manufacturing and service industries. *Papers in Regional Science* 88, 63–84.
- Kelejian, H.H., 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.
- Melo, P. C, Graham, D.J., Noland, R.B. (2009) A meta-analysis of estimates of urban agglomeration economies *Regional Science and Urban Economics* 39(3), 332–342.
- Moomaw, R.L. (1981) Productivity and city size: a critique of the evidence. *Quarterly Journal of Economics* 96, 675–688.
- Puga, D. (2010) The magnitude and causes of agglomeration economies. *Journal of Regional Science* 50, 203–219.
- R Development Core Team (2009) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.

- Rosenthal, S.S., Strange, W.C. (2004) Evidence on the nature and sources of agglomeration economies. In: Henderson JV, Thisse JF (eds) *Handbook of regional and urban economics*, volume 4. Elsevier, Amsterdam, pp. 2119-2171.
- Rossi, P.E., Allenby, G.M., McCulloch, R. (2005) *Bayesian statistics and marketing*. John Wiley and Sons Lts, West Sussex.
- Sveikauskas, L. (1975) The productivity of cities. *Quarterly Journal of Economics* 89, 393–413.
- Williamson, J. G. (1965) Regional inequality and the process of national development *Economic Development and Cultural Change* 13(4), 3–45.
- World Bank (2008) World development report 2009: reshaping economic geography, Washington DC.



Figure 1: Productivity in Manufacturing Sector



Figure 2: Productivity in Wholesale and Retail Sectors



Figure 3: Spatial Concentration in Manufacturing Sectors



Figure 4: Spatial Concentration in Wholesale and Retail Sectors



Figure 5: MCMC Sampling Path



Figure 6: Histograms of MCMC Samples