

PHÂN TÍCH KINH TẾ KHÔNG GIAN: HIỆN TƯỢNG NỘI SINH, TỰ CHỌN MẪU VÀ CÁC MÔ HÌNH KHẮC PHỤC

Trần Thị Bích

Khoa Thống kê, Trường Đại học Kinh tế quốc dân

Email: bichtt@neu.edu.vn

Lê Hoàng Minh Nguyệt

Khoa Thống kê, Trường Đại học Kinh tế quốc dân

Email: lehoangminhnguyet@gmail.com

Ngày nhận: 15/9/2016

Ngày nhận bản sửa: 8/10/2016

Ngày duyệt đăng: 18/10/2016

Tóm tắt:

Trong bối cảnh phân bố của các hoạt động kinh tế theo không gian ngày càng mất cân bằng, việc áp dụng các phương pháp phân tích kinh tế không gian để đánh giá hiệu quả phân bố của hoạt động này ngày càng quan trọng. Với mục đích phổ cập hơn nữa việc áp dụng các phương pháp phân tích kinh tế không gian tại Việt Nam, bài viết này tổng hợp lịch sử phát triển của phân tích kinh tế không gian và các vấn đề thường gặp trong phân tích như nội sinh và tự chọn mẫu. Các vấn đề này làm kết quả ước lượng từ các phương pháp thông thường bị chệch và không vững. Từ đó, bài viết đưa ra các mô hình khắc phục và điều kiện áp dụng các mô hình này bao gồm mô hình tác động cố định, mô hình 2 bước của Heckman, biến công cụ và mô hình khác biệt kép.

Từ khóa: Phân tích kinh tế không gian, nội sinh, tự chọn mẫu.

Spatial economics analysis: Endogeneity, selection bias and selected estimation models

Abstract:

While economic activities are highly disproportionately distributed, it is important to apply the spatial economic analysis to evaluate productivity of the concentration of economic activities. With the objective of popularising the application of spatial economic analyses in Vietnam, this paper provides a review of the spatial economic analysis and the issues it often faces, including endogeneity and selection bias. These issues mislead the estimated results when using the Ordinary Least Squared or the General Least Squared methods. The paper then suggests some methods that can solve the problems of endogeneity and selection bias such as the Fixed Effects model, Heckman's Two-Step Model, Instruments Variables and Difference in Differences method.

Keywords: Spatial Economic Analysis, Endogeneity, Selection Bias.

1. Giới thiệu

Phân tích kinh tế không gian (*Spatial Economics Analysis - SEA*) là phương pháp phân tích đã xuất hiện trên thế giới cách đây hơn hai thế kỷ nhằm phân tích hiệu quả của việc phân bố các nguồn lực khan hiếm theo không gian và vị trí của các hoạt động kinh tế. Những thập kỷ gần đây, SEA ngày càng phát triển và được các trường phái kinh tế địa lý mới (*New Economic Geography - NEG*) và kinh tế đô thị mới (*New Urban Economy - NUE*) áp dụng theo hướng phân tích thực nghiệm (*empirical analysis*). Theo Baum-Snow & Ferreira (2015), năm 1990,

49% các ấn phẩm đăng trên Tạp chí Kinh tế đô thị (*The Journal of Urban Economics*) là phân tích thực nghiệm, con số này tăng đến 71% trong năm 2010.

Mục đích của các nghiên cứu SEA là đánh giá tác động của kinh tế cụm (*agglomerations*), thường được đo lường bằng mức độ tập trung (có thể là mật độ dân số, doanh nghiệp, hoặc người lao động), đến năng suất của doanh nghiệp và người lao động (Trần Thị Bích, 2015, 2016). Các nghiên cứu này thường đối mặt với vấn đề nội sinh hay tự chọn mẫu. Điều này là do không phải mật độ tập trung cao của doanh nghiệp hay người lao động tại một cụm nào đó làm

tăng năng suất mà có thể bản thân các cụm đã có sẵn các lợi thế về năng suất cao, do đó thu hút các doanh nghiệp và người lao động (hiện tượng nội sinh). Bên cạnh đó, bản thân doanh nghiệp và người lao động có năng lực cao hơn mới cạnh tranh và tồn tại được ở các cụm có mật độ cao (hiện tượng tự chọn mẫu). Điều này làm cho các nghiên cứu chỉ sử dụng phương pháp bình phương nhỏ nhất với dữ liệu chéo có ước lượng không đáng tin cậy.

Do đó, trong những năm gần đây, SEA rất chú trọng đến vấn đề ước lượng để khắc phục hai hiện tượng trên trong đánh giá tác động của kinh tế cụm đến năng suất và lương của doanh nghiệp và người lao động. Thay cho việc chỉ sử dụng phương pháp bình phương nhỏ nhất (*Ordinary Least Square - OLS*) hay hồi quy chéo (*Cross-sectional Regression*), các nghiên cứu có xu hướng chuyển dần sang sử dụng phương pháp biến công cụ (*Instrument Variables - IV*), dữ liệu mảng (*Panel Data*) và mô hình khác biệt kép (*Difference in differences methods*) nhằm loại bỏ các hiện tượng nội sinh (*endogeneity*) và tự chọn mẫu (*selection bias*). Tuy nhiên vẫn còn nhiều nghiên cứu chưa xử lý được vấn đề nội sinh và tự chọn mẫu. Điều này là do chưa có một nghiên cứu nào tổng hợp một cách có hệ thống các mô hình ước lượng khắc phục hiện tượng nội sinh và tự chọn mẫu trong phân tích kinh tế không gian. Trên thực tế mỗi mô hình ước lượng được đề cập một cách vắn tắt và rải rác trong từng nghiên cứu thực nghiệm gây khó khăn cho các nhà nghiên cứu đi sau khi muốn tìm hiểu và ứng dụng các phương pháp ước lượng áp dụng trong SEA.

Mục đích của nghiên cứu này là nhằm hệ thống các vấn đề thường gặp trong SEA và các mô hình khắc phục để có được các ước lượng đáng tin cậy hơn. Điều này càng quan trọng hơn đối với Việt Nam khi phân tích kinh tế không gian vẫn còn là một lĩnh vực khá mới mẻ. Mới chỉ có một số bài báo đề cập đến vấn đề này chẳng hạn như các nghiên cứu của Trần Thị Bích, 2015, 2016. Trong bối cảnh phân bố của các nguồn lực và hoạt động kinh tế theo không gian ngày càng mất cân bằng, việc áp dụng SEA để đánh giá hiệu quả phân bố hoạt động kinh tế ngày càng quan trọng. Nhằm phổ cập hơn việc áp dụng các phương pháp phân tích kinh tế không gian và các mô hình ước lượng tin cậy trong SEA, bài viết này tổng hợp lịch sử phát triển của SEA và hai vấn đề khá phổ biến khi áp dụng phân tích kinh tế không gian là hiện tượng nội sinh và tự chọn mẫu. Bài viết cũng hệ thống hóa các phương pháp đã được sử dụng thành công để khắc phục hiện tượng nội sinh

và vấn đề tự chọn mẫu trong phân tích kinh tế không gian trong thời gian gần đây. Các phương pháp này bao gồm: tác động cố định (*Fixed effects methods*), biến công cụ (*Instruments Variables - IV*), mô hình 2 bước của Heckman (*Heckman's Two - Step Model*) và mô hình khác biệt kép (*Difference in Differences - DID*).

2. Lịch sử phát triển của phân tích kinh tế không gian

Về cơ bản, kinh tế không gian liên quan đến việc phân bổ các nguồn lực khan hiếm theo không gian và vị trí của các hoạt động kinh tế. Trong những nghiên cứu đầu tiên, kinh tế không gian tập trung vào đặc tính tính không lồi (*non-convexities*) để phân tích chi phí vận chuyển trong lĩnh vực thương mại. Nguyên nhân chính là do, mặc dù lợi thế cạnh tranh là một giải thích phù hợp để hiểu về các dòng chảy thương mại, chi phí vận chuyển cũng giải thích một phần cho việc lựa chọn vị trí của các ngành sản xuất trong nước.

Von Thünen (1826) là người đầu tiên áp dụng tính không lồi để thiết lập mô hình sử dụng đất ở nông thôn. Độc lập với những nghiên cứu truyền thống của Thünen, các nghiên cứu của Weber (1909) đã đưa ra cách tiếp cận khác. Cách tiếp cận của Weber được mở rộng theo nhiều hướng để xem xét các hàm sản xuất linh hoạt hơn và các vị trí tối ưu của các cơ sở công cộng. Tiếp theo Weber, Hotelling (1929) cũng khám phá bài toán vị trí mà các nhà sản xuất phải đối mặt nhưng đi theo một hướng rất khác biệt. Nền tảng cơ bản của ông là: các doanh nghiệp phải đối mặt với một sự đánh đổi cơ bản giữa vị trí trung tâm, cho phép họ nắm bắt thị phần lớn hơn và các địa điểm kém trung tâm hơn, cho phép họ giảm nhẹ cường độ cạnh tranh. Việc giải quyết sự đánh đổi này phụ thuộc vào các chi tiết cụ thể của các giả định (đặc biệt là giả định về sự gia tăng tác động của khoảng cách tới sức mạnh định giá của nhà sản xuất). Alonso (1964) đã phát triển một phương pháp dựa trên nguyên tắc tương tự Thünen để tìm hiểu mô hình sử dụng đất ở các thành phố. Nghiên cứu của Alonso đã mở ra xu hướng mở rộng các mô hình cơ bản và sau đó đã hình thành mô hình các thành phố đa trung tâm (Fujita, 1989).

Các hướng tiếp cận hiện đại với lý thuyết kinh tế không gian

Tính không lồi trong sản xuất là vấn đề trung tâm của phân tích kinh tế không gian. Các nghiên cứu xem chúng như biến ngoại sinh. Một thời gian không lâu trước đây, các nhà nghiên cứu bắt đầu

quan tâm đến bản chất tính không lỗi, đặc biệt là các nghiên cứu trong lĩnh vực ‘kinh tế đô thị mới’. Để hiểu được khái niệm “khu vực trung tâm” của Alonso, hay tổng quát hơn, tại sao các hoạt động kinh tế lại tập trung theo khu vực hoặc ngành, phân tích kinh tế không gian phải cung cấp nền tảng kinh tế vi mô để giải thích việc doanh thu của địa phương ngày càng tăng. Các nhà kinh tế học không gian đã phát triển phân tích trên cơ sở nghiên cứu của Adam Smith (1776) và Alfred Marshall (1920). Theo Duranton & Puga (2004), các nghiên cứu trong lĩnh vực này đã giải thích hiện tượng doanh thu của địa phương ngày càng tăng có thể do tác động lan tỏa của tri thức (*knowledge spill-over*), giảm chi phí do gần với các nhà cung cấp đầu vào và người mua hàng (*buyer-supplier linkages*), và việc dễ dàng tìm kiếm lao động với các kỹ năng cần thiết trong một thị trường lao động với kỹ năng đặc thù theo ngành (*labour market pooling*).

Các nghiên cứu sử dụng phân tích kinh tế không gian để phân tích hiệu quả của tập trung theo ngành hay tập trung tại các khu vực đô thị có thể được chia ra làm hai hướng chính. Hướng thứ nhất là các nghiên cứu của Henderson (1974) và được biết như là trường phái kinh tế đô thị mới. Hướng thứ hai bắt đầu từ nghiên cứu của Krugman (1991), được biết đến với tên gọi là trường phái địa lý kinh tế mới. Từ sau mô hình nghiên cứu của Krugman, các nghiên cứu trong lĩnh vực phân tích kinh tế không gian đã phát triển theo cấp số nhân, mà đỉnh cao là quyển sách của Fujita & cộng sự (1999).

3. Hiện tượng nội sinh và tự chọn mẫu trong phân tích kinh tế không gian

Nghiên cứu sử dụng phương pháp phân tích kinh tế không gian thường gặp những vấn đề như biến nội sinh và tự chọn mẫu. Những vấn đề này có thể làm cho kết quả ước lượng không chính xác và thiếu tin cậy. Mục này sẽ trình bày chi tiết các vấn đề này.

3.1. Biến nội sinh trong phân tích hồi quy

Nguyễn Thị Minh & Hoàng Bích Phương (2015, 4) nêu rõ, “trong mô hình kinh tế lượng, một biến được gọi là biến nội sinh nếu nó tương quan với sai số ngẫu nhiên trong mô hình, là biến ngoại sinh nếu nó không tương quan với sai số ngẫu nhiên trong mô hình. Chẳng hạn xét một mô hình hồi quy đơn giản sau:

$$Y = \beta_0 + \beta_1 X_i + u \quad (1)$$

Trong đó Y là biến phụ thuộc (biến được giải thích), X_i là biến giải thích và u là sai số ngẫu nhiên.

Mối quan tâm của người ước lượng mô hình là liệu các biến giải thích có phải là biến nội sinh hay không. Chẳng hạn nếu thì X_i là biến nội sinh. Nếu mô hình có ít nhất một biến giải thích là biến nội sinh thì mô hình được gọi là có vấn đề biến nội sinh.”

Khi đó kết quả ước lượng từ phương pháp OLS sẽ là *ước lượng chệch (biased)* và không vững (*inconsistent*).

Để minh họa tại sao biến giải thích nội sinh lại làm cho ước lượng OLS là chệch và không vững, chúng ta trở lại công thức cơ bản sau đây của ước lượng OLS với trường hợp mô hình hai biến số như mô hình (1) ở trên. Khi đó các hệ số của mô hình được ước lượng như sau:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum x_i u_i}{\sum x_i^2} (2)$$

Lấy trung bình hai vế, ta có:

$$E(\hat{\beta}_1) = \beta_1 + \frac{cov(x, u)}{var(x)} (3)$$

Có nghĩa rằng nếu x và u là có tương quan thì $\hat{\beta}_1$ sẽ là ước lượng chệch, $E(\hat{\beta}_1) \neq \beta_1$, và không vững, $\hat{\beta}_1$ không hội tụ về β_1 kể cả khi kích thước mẫu ra rất lớn. Nói tóm lại, biến giải thích nội sinh trong mô hình hồi quy sẽ làm cho các ước lượng không đáng tin cậy, kể cả khi kích thước mẫu là lớn. Do đó nếu mô hình có vấn đề về biến nội sinh thì phương pháp OLS và một số biến thể của nó như phương pháp bình phương tối thiểu tổng quát (*General Least Squares - GLS*) hay phương pháp sai số chuẩn vững sẽ không còn dùng được, và chúng ta cần các phương pháp ước lượng khác.

3.2. Vấn đề tự chọn mẫu

Tự chọn mẫu được hiểu là sai số hệ thống do mẫu được chọn không ngẫu nhiên từ tổng thể. Điều này là do các nghiên cứu đánh giá hiệu quả của tác động (*treatment effects*) thường không lấy mẫu theo phương pháp thiết kế thực nghiệm. Chẳng hạn trong nghiên cứu đánh giá hiệu quả của chương trình đào tạo đến thu nhập của người lao động, cần so sánh mức thu nhập giữa những người tham gia chương trình đào tạo (nhóm được tác động) và những người không tham gia chương trình (nhóm đối chứng). Để đánh giá hiệu quả của chương trình một cách chính xác thì các cặp nghiên cứu phải tương đồng nhau ở các đặc điểm có thể quan sát được như tuổi, giới tính, tình trạng sức khỏe,... hoặc không quan sát được như năng lực bản thân. Để đáp ứng mẫu theo tiêu chuẩn này cần thực hiện thiết kế thực nghiệm.

Tuy nhiên các nghiên cứu thường không thực hiện theo cách này vì thiết kế thực nghiệm khá tốn kém. Điều này dẫn đến mẫu không được rút ngẫu nhiên từ tổng thể và ảnh hưởng đến độ chính xác của ước lượng.

Phương trình hồi quy sau mô phỏng ví dụ trên:

$$Y = \beta_0 + \beta_i X_i + \gamma Z + u(4)$$

Trong đó Y là thu nhập, X_i là các đặc điểm quan sát được của người lao động, γ đo lường tác động của chương trình đào tạo Z đến Y . Trong phương trình (4), một biến không quan sát được là khả năng có thể tác động đến cả Z và Y . Cụ thể những người có năng lực thường có mức lương cao hơn và họ cũng có xu hướng hay tham gia vào các chương trình đào tạo Z hơn. Tác động của biến không quan sát này được biểu hiện thông qua u , được biết đến như hiện tượng biến bị bỏ sót (*omitted variables*). Như vậy, giữa Z và u có mối liên hệ tương quan dẫn đến giả thiết $cov(Z, u) \neq 0$ và xảy ra ước lượng chệch và không vững như giải thích ở phần biến nội sinh.

3.3. Hiện tượng nội sinh và tự chọn mẫu trong SEA

Như đã đề cập ở trên, phân tích kinh tế không gian đánh giá hiệu quả của việc tập trung (tích tụ - *agglomeration*) theo không gian của các hoạt động kinh tế hay cụ thể là của doanh nghiệp và người lao động. Thông thường, doanh nghiệp và người lao động thường tập trung theo ngành (hay các cụm) hoặc tại các khu vực đô thị đông đúc. Các nghiên cứu trong SEA đánh giá tác động của các cụm hay khu vực đô thị (thông thường được đo lường bằng mật độ tập trung của doanh nghiệp hoặc người lao động) đến năng suất của doanh nghiệp và người lao động. Các nghiên cứu này thường đối mặt với vấn đề nội sinh hay tự chọn mẫu. Điều này là do không phải mật độ tập trung cao của doanh nghiệp hay người lao động tại một cụm nào đó làm tăng năng suất mà có thể bản thân các cụm này có sẵn các lợi thế về năng suất cao và do đó thu hút các doanh nghiệp và người lao động (hiện tượng nội sinh). Bên cạnh đó bản thân doanh nghiệp và người lao động có năng lực cao hơn mới cạnh tranh và tồn tại được ở các cụm có mật độ cao (hiện tượng tự chọn mẫu). Điều này làm cho các nghiên cứu chỉ sử dụng OLS có ước lượng không đáng tin cậy.

Để khắc phục các hiện tượng trên, các nghiên cứu về SEA trong vòng 10 năm trở lại đây có xu hướng sử dụng các phương pháp ước lượng nhằm khắc phục hiện tượng nội sinh và tự chọn mẫu. Các mô hình này được trình bày ở mục 4.

4. Một số mô hình khắc phục hiện tượng nội sinh và tự chọn mẫu

Mục này sẽ liệt kê những mô hình và ý tưởng cơ bản của các mô hình trong việc giải quyết hiện tượng nội sinh và tự chọn mẫu. Mô hình khắc phục đầu tiên là mô hình tác động cố định (*Fixed effects models*). Mô hình này là lựa chọn tốt nhất để loại bỏ hiện tượng nội sinh và tự chọn mẫu bằng cách khai thác bản chất của dữ liệu mảng. Tuy nhiên, hạn chế của mô hình khi áp dụng vào trường hợp của Việt Nam là thiếu dữ liệu mảng. Mô hình thứ hai là mô hình hai bước của Heckman (*Heckman's Two-step method*), nó có thể giải thích hiện tượng nội sinh do tự chọn mẫu. Nhược điểm của phương pháp này là có thể làm giảm quy mô của mẫu. Mô hình thứ ba là biến công cụ (*Instrument Variables*), được sử dụng để giải thích cho mỗi nguyên nhân của nội sinh và tự chọn mẫu liệt kê trong mục 3 và áp dụng được cho cả dữ liệu chéo. Mặc dù vậy, mô hình có nhược điểm là rất khó tìm được biến công cụ thỏa mãn các giả thiết đặt ra, đặc biệt trong điều kiện của Việt Nam do thiếu lưu trữ dữ liệu trong quá khứ. Ngoài ra, mục này còn đề cập thêm mô hình khác biệt kép (DID) như là một ứng dụng cụ thể của tác động cố định.

4.1. Mô hình tác động cố định (*Fixed effects models*)

Dữ liệu mảng có thể cung cấp các giải pháp hữu ích cho vấn đề nội sinh hay tự chọn mẫu đã đề cập ở trên bằng cách sử dụng các mô hình tác động cố định (*Fixed Effects*). Mục đích của việc đưa thêm các tác động cố định c_i là nhằm loại bỏ các tác động không quan sát được và cố định theo thời gian. Khi đánh giá tác động của quy mô (hay mật độ của doanh nghiệp hoặc người lao động) của các cụm, SEA xử lý vấn đề nội sinh như lợi thế của các cụm hay năng lực không quan sát được của doanh nghiệp và người lao động bằng cách đưa các tác động cố định của cụm, doanh nghiệp, và người lao động vào mô hình hồi quy. Mô hình này được viết như sau:

$$y_{it} = X_{it} \beta + c_i + u_{it} \quad (5)$$

Với X_{it} chứa các biến giải thích chỉ thay đổi theo i hoặc t , hoặc cả hai. Các u_{it} là những nhiễu (sai số) có thể thay đổi theo i và t . Lưu ý rằng tác động không quan sát được được giả định không thay đổi theo thời gian và thường là phản ánh lợi thế của các cụm hay năng lực không quan sát được của doanh nghiệp hoặc người lao động. Vấn đề then chốt là liệu c_i có đảm bảo không tương quan với các biến giải thích quan sát được X_{it} với $t = 1, 2, \dots, T$ hay không.

Nếu không có sự tương quan giữa X_{it} và c_i , nghĩa là $Cov(X_{it}, c_i) = 0$ với mọi X_{it} , $t = 1, 2, \dots, T$, thì c_i được coi là một “tác động ngẫu nhiên riêng lẻ - *individual random effect*”. Ngược lại, nếu tác động không quan sát được được phép tương quan với các biến giải thích, c_i sẽ được coi là một “tác động cố định riêng lẻ - *individual fixed effect*”.

Ước lượng tác động cố định (*Fixed Effects*) khai thác bản chất của dữ liệu mảng bằng cách biến đổi phương trình ước lượng (5) nhằm loại trừ ảnh hưởng của các tác động không quan sát được c_i . Một cách để thực hiện điều đó là chuyển đổi các tác động cố định bằng cách tính giá trị trung bình cho các biến ở phương trình (5):

$$\bar{y}_i = \bar{X}_i\beta + c_i + \bar{u}_i(6)$$

và sau đó lấy (5) trừ (6) để loại bỏ các c_i không đổi:

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)\beta + u_{it} - \bar{u}_i(7)$$

Khi c_i đã được gỡ bỏ, phương trình (7) cho kết quả ước lượng tin cậy. Lưu ý việc loại bỏ c_i đồng nghĩa với loại bỏ các tác động cố định bao gồm lợi thế về năng suất của cụm và khả năng không quan sát được của doanh nghiệp và người lao động. Do vậy, kết quả ước lượng phản ánh trung thực tác động của các cụm đến năng suất của doanh nghiệp và người lao động và do đó khắc phục được hiện tượng nội sinh và tự chọn mẫu trong SEA.

4.2. Mô hình 2 bước của Heckman (*Heckman's Two - Step Model*)

Mô hình tác động cố định đòi hỏi phải có dữ liệu mảng (ở cấp doanh nghiệp hoặc người lao động) và điều này nhiều khi khó khăn khi sử dụng dữ liệu của Việt Nam. Khi chỉ có dữ liệu chéo và như vậy thường gặp phải vấn đề tự chọn mẫu (chẳng hạn doanh nghiệp và người lao động có năng lực cao hơn thường tập trung vào các cụm có nhiều lợi thế về tiền lương và năng suất). Do vậy, ước lượng tác động của quy mô hoặc mật độ của các cụm đến lương hay năng suất bị sai lệch. Để khắc phục, các nhà phân tích dữ liệu có thể sử dụng mô hình hai bước của Heckman. Ý tưởng của mô hình này là thực hiện ước lượng *probit* hoặc *logit* ở bước 1 nhằm tìm ra các quan sát có điểm số (*scores*) tương đồng nhau trên cơ sở các biến quan sát được. Nói một cách khác, mô hình ở bước 1 sẽ tìm ra và kết nối các quan sát (doanh nghiệp hoặc người lao động) có đặc điểm tương tự nhau. Bước 2 thực hiện đánh giá tác động của quy mô hoặc mật độ của các cụm đến năng suất hay lương của các nhóm doanh nghiệp hoặc người lao động có đặc điểm tương tự nhau đã được

phân loại ở bước 1. Do vậy, tác động của cụm đã được bóc tách khỏi ảnh hưởng năng lực của doanh nghiệp và người lao động. Mô hình này, còn được gọi là Heckit (Heckman, 1979) được diễn giải như sau. Lấy lại mô hình (4) ở mục 3 như sau.

$$Y = \beta_0 + \beta_i X_i + \gamma Z + u$$

Bước 1: Thực hiện ước lượng (thường là bằng mô hình *probit*):

$$z_i = \omega W_i + \epsilon_i(8)$$

Với z_i là một biến giả nhận giá trị bằng 1 nếu cá thể i tham gia vào tác động (tham gia chương trình đào tạo chẳng hạn) và bằng 0 nếu không tham gia. W_i là một tập hợp các biến giải thích tác động đến z_i và quan sát được. Sau khi ước lượng mô hình (8), mỗi quan sát sẽ có một giá trị (*score*- z_i) và các cá thể tham gia vào tác động hay không tham gia được so sánh với nhau trên cơ sở có các giá trị z_i tương đồng và do vậy loại bỏ được vấn đề tự chọn mẫu. Sau đó, các giá trị phù hợp z_i từ mô hình được sử dụng để tính toán nghịch đảo tỷ lệ Mills, $\lambda_i = \phi(z_i)/\Phi(z_i)$, với $\phi(\cdot)$ và $\Phi(\cdot)$ tương ứng là hàm mật độ và phân phối xác suất cho một biến chuẩn.

Bước 2: Cộng nghịch đảo tỷ lệ Mills như một biến giải thích thêm vào mô hình hồi quy ban đầu:

$$y_i = X_i\beta + \alpha\lambda_i + v_i(9)$$

Theo các giả định u_i và v_i tuân theo phân phối chuẩn hai biến, mô hình này cung cấp các ước lượng phù hợp của các hệ số hồi quy β và α , hoặc bằng OLS hoặc bằng MLE trong trường hợp của các mô hình phản ứng nhị phân (*binary response models*).

Mặc dù có lợi thế là áp dụng được cả cho dữ liệu chéo, mô hình hai bước của Heckman có nhược điểm là làm giảm quy mô của mẫu do nhiều quan sát bị loại bỏ trong quá trình kết nối tương đồng. Mẫu nhỏ cũng ảnh hưởng ít/nhiều đến kết quả ước lượng và vấn đề này cần cân nhắc khi phân tích dữ liệu của Việt Nam do bản thân quy mô mẫu vốn đã không lớn.

4.3. Biến công cụ (*Instrument Variables - IV*)

Như đã đề cập ở trên, khi ước lượng tác động của các cụm hay khu vực đô thị đến năng suất của doanh nghiệp và người lao động có thể gặp vấn đề là bản thân các cụm có sẵn các lợi thế về năng suất cao và do đó thu hút các doanh nghiệp và người lao động. Do vậy, cần tìm các biến ngoại sinh (IV) phù hợp để thay thế cho biến nội sinh (trong trường hợp này là

quy mô của các cụm) rồi thực hiện ước lượng.

Để minh họa cách IV thực hiện trong ước lượng bình phương nhỏ nhất hai giai đoạn (2SLS), xem xét mô hình tiêu chuẩn sau:

$$y_i = X_i\beta + u_i(10)$$

Giả sử $X_i = [X_{1i}, X_{2i}]$, với X_{1i} là các biến giải thích ngoại sinh và X_{2i} là các biến giải thích nội sinh, ví dụ:

$$y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + u_{ii} \quad (11)$$

Theo Wooldridge (2010), để ước lượng $\beta = [\beta_1, \beta_2]$ một cách tin cậy thông qua ước lượng IV, trước hết cần có một tập hợp các biến công cụ Z_i đáp ứng hai điều kiện sau đây:

(i) Tương quan với các biến giải thích nội sinh: $\text{Cov}[Z_i, X_{2i}] \neq 0$

(ii) Không có mối tương quan với phần sai số: $\text{Cov}[Z_i, u_i] = 0$

Điều kiện (i) gọi là điều kiện hạng (*rank condition*), trong khi điều kiện (ii) được gọi là điều kiện ngoại sinh. Điều kiện hạng yêu cầu phải có số biến công cụ lớn hơn hoặc bằng biến giải thích nội sinh. Nếu không, các phương trình không được xác định.

Bước đầu tiên của 2SLS là ước lượng phương trình dạng khử, ví dụ hồi quy các biến nội sinh theo tất cả các biến ngoại sinh, bao gồm cả IV:

$$X_{2i} = Z_i\alpha + X_{1i}\gamma + \epsilon_i(12)$$

với thống kê F có thể được sử dụng để kiểm định mức ý nghĩa của α về mức độ tương quan giữa các biến nội sinh và các biến công cụ.

Trong bước thứ hai các giá trị phù hợp của phương trình dạng khử được sử dụng thay cho X_{2i} để ước lượng phương trình (12) ở trên. Kết quả ước lượng được viết như sau:

$$\hat{\beta}_{2SLS} = ([\hat{X}_{i2}, X_{i1}]' [\hat{X}_{i2}, X_{i1}])^{-1} [\hat{X}_{i2}, X_{i1}]' y_i(13)$$

Kiểm định Hausman (1978) có thể sử dụng để kiểm tra hiện tượng nội sinh bằng cách so sánh thông số (*parameters*) của 2SLS với những thông số có thể thu được bằng phương pháp OLS khi không có IV. Do 2SLS và OLS đều phù hợp khi tất cả các biến giải thích là ngoại sinh, có thể kết luận rằng có hiện tượng nội sinh nếu hai thông số khác biệt đáng kể.

Ciccone & Hall (1996) là một trong những nhà nghiên cứu tiên phong sử dụng biến công cụ để phân tích tác động của kinh tế cụm đến năng suất tại Mỹ. Hai ông gợi ý các IV phải phản ánh được sức hút của các cụm đối với doanh nghiệp và người lao động trong quá khứ nhưng lại không có mối liên hệ với năng suất ở thời điểm phân tích. Theo đó, các biến

như mật độ dân số của các cụm hay điều kiện hạ tầng giao thông từ xa xưa trong lịch sử có thể sử dụng được làm IV. Do vậy, khó khăn trong sử dụng phương pháp IV đối với số liệu của Việt Nam là rất khó tìm được một biến thay thế cho mật độ tập trung của các cụm mà không liên quan đến năng suất của doanh nghiệp và người lao động do việc lưu trữ dữ liệu quá khứ của Việt Nam rất yếu.

4.4. Mô hình khác biệt kép (Difference in differences methods - DID)

Phương pháp khác biệt kép DID là một ứng dụng phổ biến cụ thể của tác động cố định. Về cơ bản, DID so sánh các nhóm tác động và đối chứng dựa trên những khác biệt trong kết quả ở từng thời kỳ quan sát. Cụ thể, thực hiện điều tra cả nhóm tham gia vào tác động và nhóm đối chứng trước khi có tác động. Sau khi thực hiện tác động, thực hiện một điều tra tiếp theo về cả hai nhóm này. Từ thông tin này, ta tính toán được sai biệt giữa các kết quả trung vị được quan sát trên nhóm tác động và đối chứng trước và sau tác động.

Theo Bertrand & cộng sự (2004), khi có dữ liệu ban đầu, ta có thể tính toán được mức tác động bằng cách giả định tính không đồng nhất không được quan sát không đổi theo thời gian và không có liên hệ với can thiệp trong thời kỳ. Giả định này mở hơn yếu tố ngoại sinh có điều kiện và thể hiện những thay đổi trong kết quả nhóm đối chiếu gồm đối tượng không tham gia. Cũng có thể xác định sự khác biệt theo cơ chế hồi quy, trong đó phép hồi quy được gán quyền số để tính toán sai số có thể có.

Ưu điểm của DID là giảm thiểu giả định về yếu tố ngoại suy có điều kiện hay chỉ lựa chọn các đặc tính được quan sát, đồng thời cung cấp một phương pháp trực giác, dễ sử dụng để tính toán việc lựa chọn các đặc tính không được quan sát. Tuy vậy, phương pháp đòi hỏi thông tin phải được quan sát trên cả nhóm tác động và đối chứng ở ít nhất hai mốc thời gian nên khó áp dụng ở Việt Nam do thiếu số liệu.

5. Kết luận

Tại các nước đang phát triển nói chung và tại Việt Nam nói riêng, các hoạt động kinh tế ngày càng có xu hướng tập trung cao tại một số khu vực, đặc biệt là các khu vực đô thị gây nên áp lực rất lớn cho các thành phố đã quá tải. Do vậy, việc đánh giá hiệu quả của phân bố theo khu vực nhằm cung cấp bằng chứng khoa học cho hoạch định chính sách phát triển kinh tế không gian ngày càng quan trọng.

Trong những thập kỷ gần đây, phân tích kinh tế không gian rất phát triển. Các phương pháp phân

tích kinh tế không gian thường gặp phải hiện tượng nội sinh và tự chọn mẫu làm cho kết quả ước lượng không chính xác nếu chỉ sử dụng ước lượng bình phương nhỏ nhất thông thường với dữ liệu chéo. Trong vòng 10 năm trở lại, các nghiên cứu có xu hướng sử dụng các mô hình khắc phục ngày càng tăng làm cho kết quả ước lượng đáng tin cậy và chính xác hơn.

Nhằm phổ cập hơn các phương pháp ước lượng đáng tin cậy trong phân tích kinh tế không gian tại Việt Nam, bài viết này đã cung cấp tổng quan về lịch sử phát triển của phân tích kinh tế không gian, phân tích sâu một số vấn đề về hiện tượng nội sinh và tự chọn mẫu. Trên cơ sở đó, bài viết đã đưa ra một số

mô hình để khắc phục các hiện tượng này.

Mô hình đầu tiên là Mô hình tác động cố định. Trong một số điều kiện nhất định mô hình này có thể loại bỏ các tác động của nội sinh và biến bị bỏ sót bằng cách khai thác bản chất của dữ liệu mảng. Mô hình thứ hai là mô hình hai bước của Heckman. Mô hình này có thể khắc phục hiện tượng nội sinh do tự chọn mẫu. Thứ ba là mô hình với biến công cụ được sử dụng để giải thích cho mỗi nguyên nhân của nội sinh. Cuối cùng là mô hình khác biệt kép, là một ứng dụng cụ thể của tác động cố định. Bài viết đã đưa ra những ý tưởng cơ bản, lý thuyết về cách thực hiện và một số điều kiện đi kèm của các mô hình trên.

Tài liệu tham khảo:

- Alonso, W. (1964). *Location and Land Use*. Cambridge, MA: Cambridge University Press.
- Baum-Snow, N. & Ferreira, F. (2015), 'Causal Inference in Urban and Regional Economics' in *Handbook of Regional and Urban Economics*, edited by Gilles Duranton, J. Vernon Henderson & William C. Strange. Amsterdam, The Netherlands: Nikki Levy Publisher.
- Bertrand, M., Duflo, E. & Mulainathan, S. (2004), 'How much should we trust Differences in Differences Estimates', *Quarterly Journal of Economics* 119 (1): 249-75.
- Ciccone, A. & Hall, R.E. (1996), Productivity and the Density of Economic Activity. *The American Economic Review*, 86(1), 54-70.
- Duranton, G. & Puga, D. (2004). Micro-foundations of urban agglomeration economies. In *Handbook of Regional and Urban Economics*, vol. 4, ed. V. Henderson and J.-F. Thisse. Amsterdam: North-Holland.
- Fujita, M. (1989), *Urban Economic Theory: Land Use and City Size*, Cambridge: Cambridge University Press.
- Fujita, M., Krugman, P. & Venables, A. (1999), *The Spatial Economy: Cities, Regions, and International Trade*, Cambridge, MA: MIT Press.
- Hausman, J. A. (1978), Specification tests in econometrics, *Econometrica*, 46 (6), 1251-1271.
- Heckman, J. J. (1979), Sample selection bias as a specification error, *Econometrica*, 47 (1), 153-161.
- Henderson, J. (1974), The sizes and types of cities, *American Economic Review* 64, 640-56.
- Hotelling H. (1929), Stability in competition, *Economic Journal* 39, 41-57.
- Krugman, P. (1991), Increasing Returns and Economic Geography. *Journal of Political Economy*, 99, 483-499.
- Marshall, A. (1920), *Principles of Economics*, 8th Ed. London: MacMillan.
- Nguyễn Thị Minh & Hoàng Bích Phương (2015), "Tàn mạn về vấn đề biến nội sinh trong phân tích kinh tế", truy cập ngày 11 tháng 8 năm 2016, từ <<http://www.mfe.edu.vn/bai-nghien-cuu/no-02-2015-tan-man-ve-van-de-bien-noi-sinh-trong-phan-tich-kinh-te.html>>.
- Smith, A. (1776), *An Inquiry into the Nature and Causes of the Wealth of Nations*, London: Printed for W. Strahan and T. Cadell.
- Trần Thị Bích (2015), "Sản xuất công nghiệp phi chính thức ở Việt Nam: tập trung theo ngành hay đa dạng hoá", *Tạp chí Kinh tế và Phát triển*, Số 223, 21-29, 2016.
- Trần Thị Bích (2016), "Chênh lệch về lương theo địa phương: nghiên cứu từ lao động phi chính thức tại đô thị ở Việt Nam", *Tạp chí Kinh tế và Phát triển*, Số 227, 98-106, 2016.
- Von Thünen, J. (1826), *Isolated State: An English edition of Der isolierte Staat*, Trans. by C. Wartenberg and P. Hall, Oxford: Pergamon Press, 1966.
- Weber, A. (1909), *Alfred Weber's Theory of the Location of Industries*, Trans. by C. Friedrich, Chicago: University of Chicago Press, 1929.
- Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press, 2nd edn.