공간 단위의 선정이 회귀분석 결과에 미치는 양상 분석
: 근원적 공간적 자기상관 수준 및 교차 상관 수준을 기준으로 한 무작위 합역 시뮬레이션 연구

▷ 학 생　이 름: 곽지호(지리학과), 강병권(에너지자원공학과)

▷ 지 도　교 수: 이상일

▷ 지도교수 소속: 사범대학 지리교육과

이 보고서는 학생자율교육 프로그램의 취지와 원칙에 따라, 학부생 곽지호, 강병권이 주도적이고 자율적으로 설계하여 시행한 연구를 바탕으로 직접 작성한 결과물임을 확인합니다.	☑

- 목 차 -

I. 서론

공간단위 임의성의 문제(Modifiable Areal Unit Problem, MAUP)는 공간 데이터를 다루는 연구에서 피할 수 없는 전통적인 문제이다. 이는 모든 공간 데이터의 관측치는 특정한 관측지와 대응되며, 관측치와 관측지는 상호 의존적이라는 점에서 기인한다. MAUP은 크게 구획 방식의 영향을 의미하는 구획 효과와 공간 단위 측정 스케일의 영향을 의미하는 스케일 효과로 구분되어진다. 본 연구에서 주목하고자 하는 에어리어 데이터의 맥락에서 스케일 효과와 구획 효과는 작은 공간 단위의 관측지를 합역(spatial aggregation)하는 방식에 따라 해당 데이터는 전혀 다른 정보를 갖게 되는 문제 상황으로 드러난다. 따라서 구득한 공간 데이터를 이용해 분석하고자 할 때에 연구자들은 해당 공간 단위의 수정 가능성을 염두에 두고 통계 분석 결과를 선정된 공간 단위에 대해 상대적으로 해석해야 한다. (이상일, 1999)

Openshaw(1984)가 MAUP을 “공간분석이 직면한 가장 중요한 미제 중 하나”라고 지적한 이래, 일부 연구자들은 그에 대한 대응으로 공간 단위의 선정과 통계 분석 결과의 체계적 연관성을 정량적으로 파악하고자 하였다. (Steel & Holt, 1996; Holt et al., 1996; Duque et al., 2018; Lee et al., 2019) 이를 통해 밝혀진 주요한 사실은 공간적 의존성, 혹은 공간적 자기상관이라는 공간 데이터의 특수성이 MAUP이 통계량에 미치는 영향의 크기 혹은 방향성과 필연적으로 깊은 연관이 있다는 것이다. 그러나, 두 개 변수 이상이 관여되는 회귀 모형에서는 변수의 측정 스케일이 회귀 분석 결과에 미치는 영향이 상대적으로 적게 알려져 있을 뿐 아니라 일관된 경향성이 존재하지 않는다는 결론이 주로 내려져 왔다. (Arbia, 1989; Fotheringham & Wong, 1991; Chen et al., 2011; Jacobs-Crisioni et al., 2014)

본 연구에서는 단순회귀모형에서 MAUP이 발생하는 양상을 보다 체계적으로 살펴보기 위하여 무작위 합역 알고리즘을 이용한 시뮬레이션을 실시한다. 가상의 데이터를 구성할 때 일차적으로는 x (독립변수)와 y (종속변수) 두 변수 간의 상관성에 따라 5가지 수준으로 구분하며, 이차적으로는 x와 y 각각이 가지는 공간적 자기상관의 정도에 따라 구분한다. 구성된 데이터를 10단계의 합역 수준으로 합역하면서 공간 데이터의 측정 스케일이 커짐에 따라 회귀 계수에 발생하는 효율성 손실(efficiency loss)의 양상에 데이터의 공간적 자기상관 및 교차 상관에 의거한 경향성이 관찰되는지 파악하고자 한다. 이때 무작위 합역의 원리에 의해 각 합역 수준별 평균치와 분산은 각각 스케일 효과와 구획 효과로 상정할 수 있다. (Lee et al., 2019) 한편, 실제 공간 데이터를 이용해 회귀 분석을 시행함에 있어서 변수 간 스케일이 일치하지 않는 경우도 적지 않다. 이때 연구자들은 공간 단위가 큰 데이터를 작은 단위로 대입(disaggregation) 혹은 내삽(interpolation)하는 해결책을 취하는 경우가 일반적인데, 각각의 상황에서 MAUP의 양상이 어떻게 달리 나타나는지 또한 회귀 모형에서의 MAUP의 영향을 탐구하고자 할 때 논의되어야 할 부분이다. 따라서 본 연구에서는 (1) x만 합역 후 대입하는 경우 (2) x만 합역 후 (역거리 가중법 등으로) 내삽하는 경우 (3) x, y 모두를 합역하는 경우로 나누어 각각의 경우에서 효율성 손실의 정도가 어떻게 달라지는지 또한 실험하고자 한다.

회귀 계수 추정량의 분산은 해당 변수의 유의성을 파악하는 데 있어서 핵심이 되는 통계량이다. 만일 합역으로 인해 분산이 과대추정될 경우, 해당 변수는 실제(혹은 조밀한

수준에서)보다 유의성이 약한 것으로 판명되게 된다. 특히 단순회귀모형에서 회귀 계수의 분산은 잔차제곱합과 독립변수의 분산과 깊은 관련이 있다. 데이터가 합역될 경우 필연적으로 x와 y의 분산에는 변동이 발생하게 되고, 이는 회귀 계수의 분산에 영향을 미칠 것을 예상할 수 있다. 한편 공간 데이터의 특수성은 이러한 합역이 공간적으로 연접한 개체 간에서 이루어진다는 점이다. 즉, 공간 데이터가 가진 공간적 자기상관의 정도는 합역된 후의 분산에 체계적으로 영향을 미칠 수밖에 없다. 합역과 유사한 평활의 관점에서 Lee(2001)는 Spatial Smoothing Scalar (SSS)를 정의하고 이것이 일종의 '분산 감소 계수'라는 점을 밝히며 (양의) 공간적 자기상관의 정도는 해당 변수의 변동을 체계적으로 감소시킨다는 사실을 지적한 바 있다. 이렇듯 합역으로 인한 x, y의 분산 변화는 회귀모형에 MAUP이 관여하는 핵심 메커니즘이며, x, y의 분산 변화는 공간적 자기상관과 밀접한 관련이 있다. 따라서 공간적 자기상관 수준을 기준으로 분류하여 무작위 합역을 시행한 결과를 제시하는 본 연구의 방법론은 MAUP의 영향을 체계적으로 파악하기 위한 합리적인 실험 설계라고 할 수 있다.

Ⅱ. 선행 연구 검토

1. 회귀 모형에서 MAUP의 영향

공간 통계 분석 결과에 공간 단위의 설정이 영향을 미친다는 문제는 Gehlke and Biehl(1934)에 의해 최초로 제기되었다. 이 문제가 MAUP이라고 명명(Openshaw and Taylor, 1979; Openshaw, 1984)된 이후, Arbia(1989)는 이변량 상황에서 공간 단위와 통계 분석 결과 간의 체계적 연관성에 주목하였다. 구체적으로 X와 Y간 분산-공분산 행렬의 공분산이 모든 관측치 쌍에 대해 일정한 경우와 국지적으로 상이한 경우로 나눈 뒤 데이터를 합역함에 따라 상관계수의 변화를 관찰하였다. Arbia and Petrarca(2011)은 더 나아가 OLS와 두 가지 공간 회귀 모형, 공간 지체 모형(spatial lag model), SARAR(1,1) 모형에서 합역에 따른 통계 분석 결과의 변화를 관찰하였다. 그 결과 OLS에서는 회귀계수에 무관하게 합역된 공간 단위의 측정 스케일이 커짐에 따라 회귀계수의 분산이 체계적으로 감소하는, 효율성의 손실(efficiency loss)이 관찰되었다. 한편, 공간 회귀 모형에서는 변수의 공간적 자기상관에 따라 다른 양상이 나타났는데, 일반화해서 말하면 양의 공간적 자기상관을 가진 변수는 효율성의 손실을 완화하고, 음의 공간적 자기상관을 가진 변수는 효율성의 손실을 증폭시킨다. 이는 근원적으로 양의 공간적 자기상관을 가진 변수를 합역하여 공간 회귀 모형에 투입하였을 경우 상대적으로 정보량의 손실이 적었기 때문이라고 해석할 수 있다. 그러나, Fotheringham and Wong(1991)과 Green and Flowerdew(1996)에서 지적되었듯이 회귀계수에는 X, Y 각각의 공간적 자기상관 뿐 아니라 X, Y 상호간의 상관이 영향을 미칠 수 있을 것으로도 보인다. 더 나아가 X, Y간의 이변량 공간적 자기상관의 영향을 고려한 연구는 아직 이루어지지 않았다. 따라서 본 연구에서는 대표적 이변량 공간적 자기상관 통계량인 Lee's L(Lee, 2001)을 추가적으로 고려하여 다양한 수준의 Lee's L, Moran's I, Pearson's r을 갖는 데이터를 이용해 실험을 진행하고자 한다.

2. 무작위 합역 절차

공간적 합역 혹은 공간적 애그리게이션(spatial aggregation)은 기본 공간단위를 병합하여 하위지역을 구성하는 과정으로 정의된다. (이상일·이몽현, 2020) 본 연구에서 집중하고자 하는 상황인 에어리어 데이터에서의 공간 단위 불일치의 경우 서로 다른 수준의 합역을 통해 속성값이 재계산되어 있는 상황이다. 이 공간 단위 혹은 합역 수준의 선정은 기본적으로 임의적이기는 하나, 비공간 데이터 분석에서의 애그리게이션과 달리 다음의 두 가지 조건이 고려되어야 한다. 첫째는 상호배제 및 전체포괄 (MECE, mutually exclusive and collectively exhaustive)의 원칙이다. 모든 공간 단위는 중첩되지 않으면서 전체 연구 지역을 완전히 구성할 수 있어야 한다는 것이다. 둘째는 연접성 요구 혹은 연접성 제약이다. 합역은 오로지 연접한 공간 단위들에 대해서만 이루어져야 한다는 것이다. 따라서 이 두 가지 특수한 조건을 고려하면서도 공간 단위 선정의 임의성을 보여줄 수 있는 무작위 합역 절차가 마련되어야 한다. 이때 무작위 합역 절차 역시 다양하게 존재할 수 있으며, 내부적인 알고리즘의 특성에 따라 실험 결과 역시 필연적으로 달라지게 될 것이다. 그러나 무작위 합역 절차 설계의 중요성에도 불구하고, 많은 연구에서 그 내부 알고리즘은 선명하게 드러나 있지 않거나 크게 강조되지 않았다. (이상일·이몽현, 2020)

이에 관련된 선구자적인 연구로 이상일·이몽현(2020)은 무작위 합역 절차를 '무선정 옵션의 적용 방식'과 '후보 공간단위의 선정 확률 결정 방식'에 따라 구분하여 6가지의 무작위 합역 절차를 제시한 뒤 시뮬레이션을 통해 서로 다른 무작위 합역 알고리즘이 만들어 내는 최종 구역의 형태를 연접도와 원형도라는 두 가지 규준에 따라 평가하였다. 이때 연접도와 원형도의 평균은 무작위 합역 알고리즘이 최종 구역의 연접도와 원형도에 어떤 방향으로 영향을 미치는지를 드러내며, 분산은 알고리즘의 상대적 안정성을 평가할 수 있는 지표가 된다. 간략히 결과만을 서술하면 무선정 옵션을 후보 공간단위로 취급하고 후보 공간단위를 선정할 때 연접도에 기반하여 상이한 확률을 적용하는 방식이 가장 현실과의 유사성이 높고 무작위성이 뛰어난 것으로 나타났다. 한편 Lee et al.(2019)는 위의 무작위 합역 절차를 적용하여 MAUP의 스케일 효과와 구획 효과가 근원적 공간적 자기상관 수준에 따라 일변량 통계량 (평균, 분산, Moran's I)에 어떠한 양상으로 나타나는지를 살펴본 연구이다. Moran's I 고유벡터를 이용하여 10가지의 근원적 공간적 자기상관 수준을 모델링한 뒤 무작위 합역을 실행한 결과 모든 통계량에서 근원적 공간적 자기상관 수준이 MAUP이 발생하는 양상과 큰 관련성이 있다는 기존의 가설을 정량적으로 증명해낼 수 있었다. 본 연구는 이변량 이상의 상황을 가정한다는 것을 제외하고 MAUP의 구획 효과나 스케일 효과의 영향을 보고자 하는 목표는 동일하다. 따라서 위의 무작위 합역 절차를 정육각형 테셀레이션에 적용하여 X, Y가 갖는 특성 및 공간 단위의 선정에 따라 회귀모형 결과가 달라지는 양상을 확인하고자 한다.

III. 데이터 및 연구 방법론

1. 가상 데이터 및 공간적 컨피겨레이션의 설정

선행 연구를 통해 회귀 모형에서 공간 단위 선정의 영향은 크게 X와 Y의 (비공간적) 상관관계와 X, Y 각각이 가진 공간적 자기상관 수준에 따라 달라진다는 사실을 파악하였다. 이를 수식적으로 살펴보면, IID 가정 하에서 표준화된 변수의 회귀 계수 분산의 추정량은

$$Var(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{\sum (y_i - \hat{y_i})^2}{(n-1)\sum x_i^2}$$

이고, 분자는 잔차제곱합, 분모는 관측치의 개수 (자유도) 및 독립변수의 분산으로 구성되어 있다. 즉 위 식의 세 부분은 각각은 Pearson's r 및 종속변수의 합역 후 분산, 합역 수준 및 방식, 독립변수의 분산과 관련되어 있으며 수식을 통해서도 Pearson's r과 X, Y의 공간적 자기상관 수준에 따라 가상 데이터를 구성한 뒤 무작위 합역을 통해 분석하는 것이 타당한 것으로 보여진다. 다만 합역에 따른 Pearson's r 및 R^2의 체계적 변화는 수식적으로는 완벽히 도해될 수 없기에 시뮬레이션을 이용해 분석을 할 필요성이 있다.

한편, 유의하여야 할 것은 X와 Y의 상관관계를 나타내는 Pearson's r과 X와 Y 각각의 공간적 자기상관을 나타내는 Moran's I 간에 종속성이 존재한다는 사실이다. 예를 들어 X와 Y가 높은 상관관계를 지닌다면 X와 Y의 Moran's I는 유사할 것이다. 이 사실은 아래 산점도에서 확인할 수 있다.

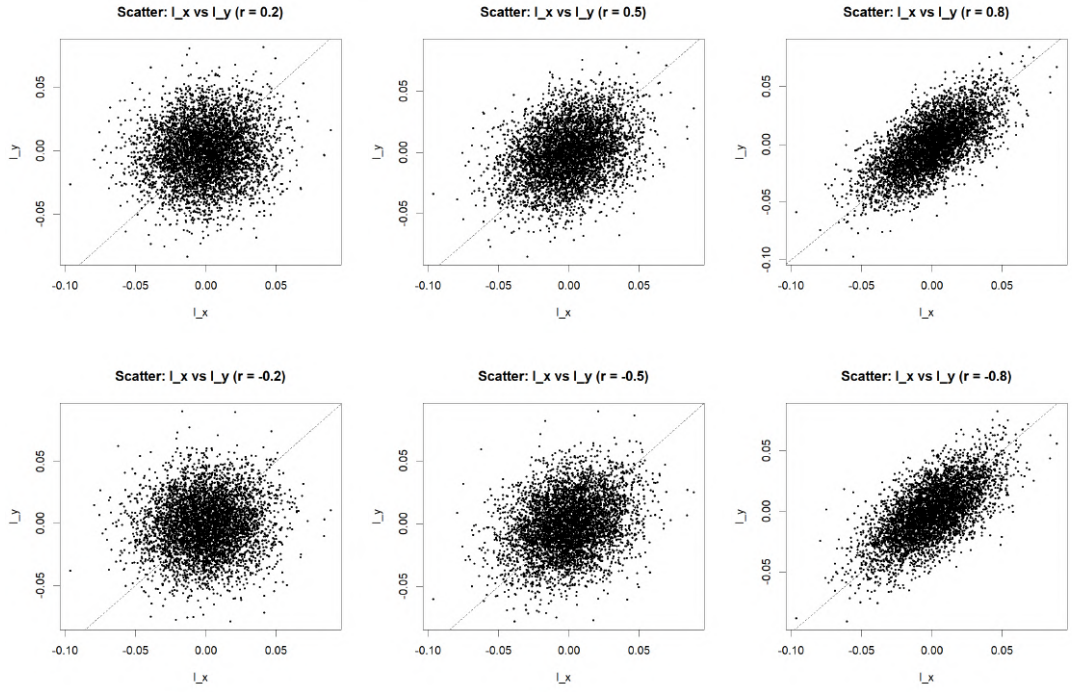

Fig 1. Scatter plot of the spatial autocorrelation between x and y

따라서 예컨대 X의 I가 높고 Y의 I는 낮으면서 X, Y간의 Pearson's r이 높은 상황은 발생할 수 없다. 따라서 본 연구에서는 X, Y 각각이 특정 공간적 자기상관 수준을 지닌 상황으로 모두에 대해 분석을 실시함으로써 모든 경우의 수를 최대한 포괄하고자 하였다. X 혹은 Y의 공간적 자기상관 수준을 설정하기 위하여 본 연구에서는 해당 변수를 모런 고유벡터(Moran eigenvector)로 설정하였다. 모런 고유벡터는 중심화된 공간근접성행렬(SPM, spatial proximity matrix)를 고유분해하여 얻어지는 벡터로, 대응하는 고유치에 해당하는 Moran's I를 나타내는 공간적 패턴을 담고 있다. 고유치를 큰 것부터 나열하면 양의 공간적 자기상관이 가장 큰 1번 고유벡터부터 음의 공간적 자기상관이 가장 큰 n번 고유벡터까지 차례로 얻을 수 있다. (Griffith 1996, 2000, 2003; Tiefelsdorf and Griffith 2007). 본 연구에서는 고유벡터가 가진 Moran's I의 가능치 범위를 등간격으로 나뉘도록 하는 9개의 고유벡터(1, 91, 192, 317, 529, 707, 832, 933, 1024)를 선정하였으며 편의상 각각의 공간적 자기상관 수준을 SA1~SA9으로 명명하고자 한다.

이후 목표하는 Pearson's r에 맞게 데이터가 구성되도록 나머지 변수를 $r\times \mathrm{EV}+\sqrt{(1-r)^2}z$ (단, z와 EV는 독립)으로 구성하였다. 본 연구에서 사용된 테셀레이션은 32×32의 정사각형 테셀레이션이며 가장 높은 공간적 분해 (spatial disaggregation) 수준에서 관측치의 수는 1024이다. ($n=1024$) 무작위 합역을 통해 분석할 10가지 합역 수준에서 각각 관측치의 수는 m=896, 768, 640, 512, 384, 256, 128, 64, 32, 16이며 이를 각각 AG1~AG10으로 명명한다. X와 Y 간의 상관관계는 Pearson's r을 기준으로 0.1, 0.3, 0.5, 0.7, 0.9의 5단계로 나누었다.

또한 본 연구는 X와 Y의 측정 스케일이 불일치하는 경우에 대해서도 분석을 진행하고자 하였다. 서술상의 편의를 위해 우선 각각의 공간적 컨피겨레이션에 대해 용어를 정리하고자 한다. 먼저 xAgg는 Y는 최초 수준 ($n=1024$)으로 고정한 뒤 X에 대해서만 합역을 진행하여 Y에 비해 공간 단위가 큰 경우를 모델링한 것이다. 이때 기본적으로는 평균값을 합역된 구역 내의 관측지들에 일괄적으로 할당하는 방식을 취한다. 즉 전체 관측치의 수는 1024로 동일한 것이다. 한편 IV-3에서는 구역 내 평균값을 동일하게 유지한 역거리가중법(Inverse Distance Weighting, IDW)을 이용해 X를 구성한 경우에 대해서도 분석을 진행한다. 두 번째로 xyAgg는 X와 Y 모두에 대해 동일한 측정 스케일로 합역을 진행한 상황이다. 두 변수의 공간적 컨피겨레이션은 일치하며, 관측치의 수 자체 역시 일괄적으로 줄어들게 된다.

위의 상황을 행렬을 이용한 선형 결합으로 표현하면 다음과 같다.

	초기	xAgg	xyAgg
관측치의 수	$n=1024$	$n=1024$	m
독립변수	$x=(x_1,\cdots,x_n)^\top$	$Px=(\tilde{x}_1,\cdots,\tilde{x}_n)^\top$	$Ax=(\bar{x}_1,\cdots,\bar{x}_n)^\top$
종속변수	$y=(y_1,\cdots,y_n)^\top$	$y=(y_1,\cdots,y_n)^\top$	$Ay=(\bar{y}_1,\cdots,\bar{y}_m)^\top$
적합값	$\hat{y}=(\hat{y}_1,\cdots,\hat{y}_n)^\top$	$\hat{\tilde{y}}=(\hat{\tilde{y}}_1,\cdots,\hat{\tilde{y}}_n)^\top$	$\hat{\bar{y}}=(\hat{\bar{y}}_1,\cdots,\hat{\bar{y}}_n)^\top$

Table 1. Number of Obs., Independent Var., Dependent Var., Fitted Values

이때 $A=RP$이고 P는 평균값을 되뿌리는 대칭·멱등(idempotent)인 $n\times n$ 블록 대각 행렬(block-diagonal matrix), R은 합역된 구역과 초기 셀을 매핑하는 $m\times n$ 행렬이다. 이해를 돕기 위해 다음의 예시를 제시한다. $x=(2,0,4,7,1,-1)^\top$에서 (1, 2), (3, 4, 5), (6)번 구역을 합역하는 상황이라고 하면,

$$A=\begin{bmatrix}\frac{1}{2}&\frac{1}{2}&0&0&0&0\\0&0&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}&0\\0&0&0&0&0&1\end{bmatrix},\ R=\begin{bmatrix}110000\\001110\\000001\end{bmatrix},\ P=\begin{bmatrix}\frac{1}{2}&\frac{1}{2}&0&0&0&0\\\frac{1}{2}&\frac{1}{2}&0&0&0&0\\0&0&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}&0\\0&0&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}&0\\0&0&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}&0\\0&0&0&0&0&1\end{bmatrix}$$

이고, 이때 Ax와 Px는

$$Ax=\begin{bmatrix}(2+0)/2\\(4+7+1)/3\\(-1)/1\end{bmatrix}=\begin{bmatrix}1\\4\\-1\end{bmatrix},\ Px=(1,1,4,4,-1)^{\top}$$

이다.

2. 무작위 합역 실험 설계

본 연구에서는 이상일·이몽현(2020)에서 연접도와 원형도 측면에서 가장 현실과 유관하면서도 적절한 무작위성을 보이는 것으로 밝혀진 무작위 합역 알고리즘을 채택하였다. 앞서 언급하였듯이 이는 무선정 옵션을 후보 공간단위로 취급하고 후보 공간단위를 선정할 때 연접도에 기반하여 상이한 확률을 적용하는 방식이다. 앞서 언급한 10가지 합역 수준에서 관측치 개수인 m은 해당 알고리즘에서 시드의 개수가 된다. 먼저 해당 시드를 무작위로 선정한 뒤, 각 라운드별로 합역할 이웃 공간단위를 선정해 가면서 전체 연구 지역이 m개의 합역된 공간 단위로 겹치지 않으면서 전체 포괄될 때까지 반복적으로 알고리즘이 실행된다.

초기 X, Y 데이터의 1024개 관측치 정사각형 테셀레이션은 고정되고 각 AG마다 xAgg, xyAgg의 공간적 컨피겨레이션은 유지된다. 이때, Pearson's r과 SA 수준을 다르게 하여 각 AG당 200회의 시뮬레이션을 실행한다. 즉, 특정한 SA, AG, Pearson's r, Moran's I 값에 대해 200번의 난수 기반 실험을 반복하여 X, Y 변수의 회귀계수 분산 팽창 계수를 박스 플롯으로 나타낸다. 여기서 분산 팽창 계수는 단순히 각 회귀계수의 분산을 나타내는 것이 아니라 각 합역 방식에 대한 분산을 초기 X, Y 데이터의 회귀계수의 분산으로 나눈 것을 의미한다. 이는 1024개의 X, Y 초기 데이터에 비해 각 합역 방식이 회귀분석의 유의성을 얼마나 감소시켰는지를 설명한다. 또한 Ⅲ-1에서 설명한 것처럼 초기 X, Y의 공간적 자기상관 수준을 각각 조절하며 시뮬레이션을 수행한다.

Ⅳ. 공간 단위 선정이 회귀 분석 결과에 미치는 영향

1. Y의 공간적 자기상관 수준이 다른 경우

먼저 xAgg에 대한 관찰 결과이다. 첫째, Pearson's r과 Moran's I에 무관하게 합역 수준이 증가함에 따라, 즉 공간 데이터의 해상도가 안 좋아질수록 스케일 효과에 의해 평균적인 분산 팽창 계수는 증가한다. xAgg에서 합역 수준이 커진다는 것은 초기 X에

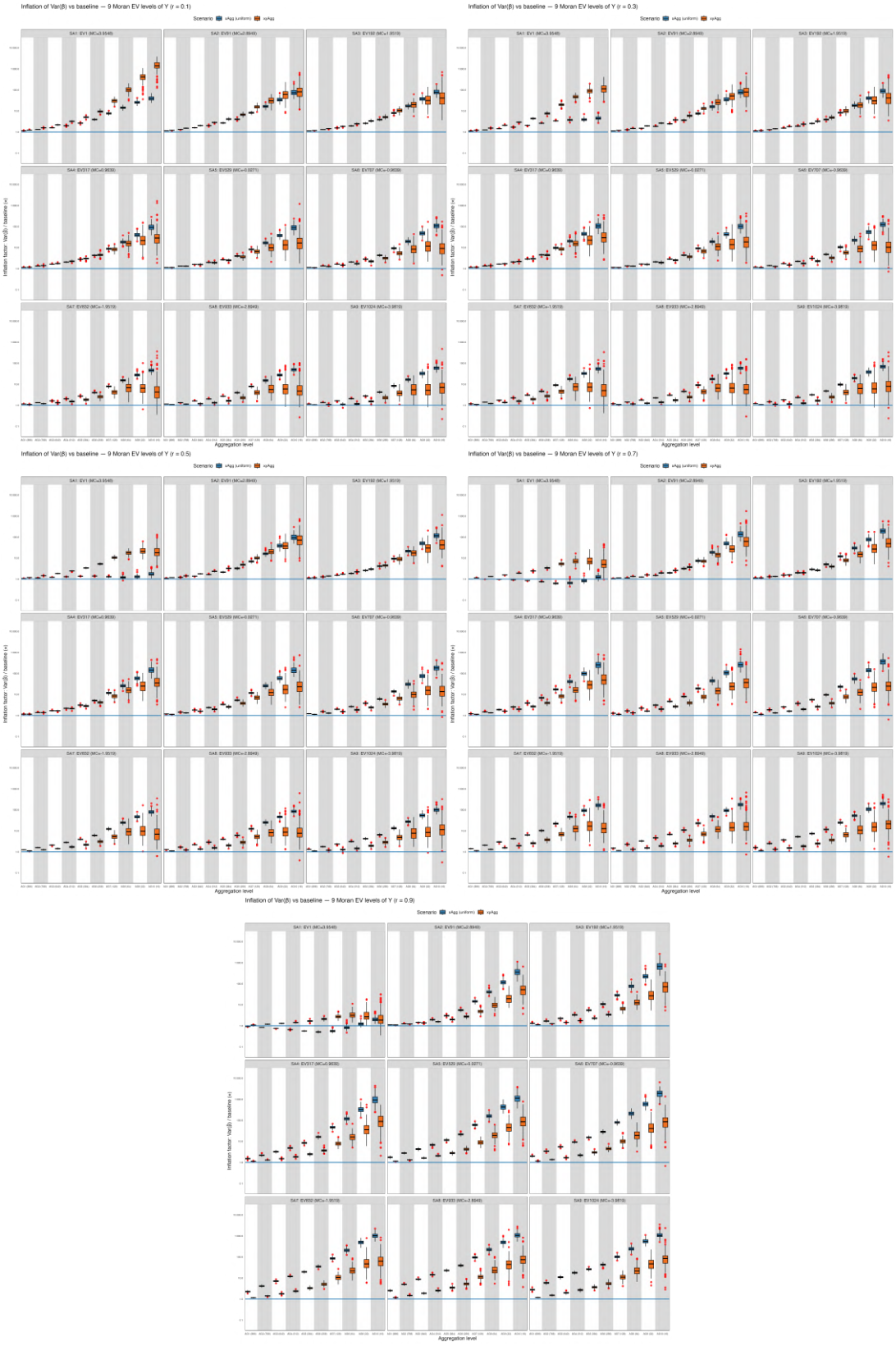

Fig 2. Variance Inflation in 9 Moran EV Levels of Y and 5 levels of Pearson's r

비해 데이터의 다양성이 줄어든다는 것을 의미한다. 따라서 X의 분산이 작아져 Y의 변화에 대한 기울기 추정값이 불안정해지기에 분산 팽창 계수가 증가하게 된다. 둘째, 합역 수준이 증가함에 따라 분산 팽창 계수의 변동 폭이 확대된다. 이는 200가지의 합역 방식에 따라 해당 통계량의 값이 크게 변할 수 있다는 점으로 MAUP의 구획 효과가 합역 수준이 증가할수록 더 확대된다고 볼 수 있다. 그 이유는 자명하다고도 볼 수 있는데, 합역된 구역의 크기가 커질수록 초기 데이터가 가지고 있는 구조를 더 많이 잃어버리기 때문이다. 셋째, Pearson's r이 커짐에 따라 비교하였을 때는 Y가 강한 양의 공간적 자기상관을 가지고 있는 경우(SA1, SA2)와 나머지 경우에서 상이한 패턴이 관찰된다: SA1과 SA2에서는 Pearson's r이 커짐에 따라 분산 팽창 계수가 대부분의 AG 수준에서 감소하는 반면 나머지 SA에서는 Pearson's r이 커짐에 따라 분산 팽창 계수가 증가한다. 그 이유를 추측하기 위해서는 강한 양의 공간적 자기상관을 가진 Y와 높은 Pearson's r을 가진 X를 구성할 경우 X 역시 높은 양의 공간적 자기상관을 가질 수밖에 없다는 점을 주목할 필요가 있다. 따라서 X의 분산이 크게 감소하지 않으며 잔차제곱합의 변화만이 영향을 미친다. (아래 수식에서 분자에 해당)

$$Var(\widehat{\beta_{\mathrm{xAgg}}}) = \frac{\sum(y_i - \hat{\tilde{y}}_i)^2}{(n-1)\sum \tilde{x}_i^2}$$

한편 Y가 강한 양의 공간적 자기상관을 갖고 있지 않다면 합역했을 때 Pearson's r이 높은 경우 X 또한 기존의 정보량을 크게 잃어버리고, 이것이 잔차 변동의 효과를 압도하여 분산 팽창 계수가 체계적으로 커지는 효과로 나타난다. MAUP이 일변량의 분산에 미치는 영향이 양의 공간적 자기상관의 정도가 큰 데이터에서만 특징적으로 나타난다는 사실은 Lee et al.(2019)에서도 밝혀진 결과로, 위의 Figure는 그 사실과 상당 부분 상응한다.

다음으로는 xyAgg에 대한 관찰 결과이다. 첫째, Pearson's r과 I에 무관하게 합역 수준이 증가함에 따라 스케일 효과에 따른 분산의 팽창이 드러난다. xyAgg에서 합역 수준이 커진다는 것은 데이터의 다양성이 줄어드는 것과 함께 샘플 수 자체가 작아지는 것이다. 이에 마찬가지로 분산 팽창 계수가 증가한다. 둘째, 구획 효과 역시 마찬가지로 AG가 커짐에 따라 확대되는 경향을 보인다. 셋째, Pearson's r이 증가함에 따라 SA1, SA2를 제외하고는 분산 팽창 계수가 증가하는 경향성을 보인다. 이에 대한 해석은 xAgg에서와 크게 다르지 않다.

한편 xAgg와 xyAgg를 비교하면 다음의 사실이 관찰된다. 첫째, 같은 조건에서 xAgg의 분산 팽창 계수가 대체로 크게 나타난다. 하지만 SA1처럼 Y의 공간적 자기상관이 큰 경우에는 xyAgg의 분산 팽창 계수가 xAgg보다 더 크게 나타난다. xyAgg의 경우 Y도 초기의 데이터를 잃게 되는데, Y가 강한 공간적 자기상관을 가질 경우에는 변동이 크게 감소하지 않는다. 이 경우 오히려 더 많은 수의 관측치를 가진 xAgg 상황에서의 분산이 더 작을 수 있는 것이다. 둘째, Pearson's r이 커짐에 따라 xAgg와 xyAgg의 분산 팽창 계수 간의 차이가 더 커짐을 확인할 수 있다. 셋째, xAgg에 비해 xyAgg에서의 구획 효과가 더 크게 나타난다. 대부분의 AG 수준에서 xyAgg Boxplot의 길이가 더 긴 것을 확인할 수 있다. xyAgg에서는 X와 Y 모두 합역되므로 구획 효과가 증폭되어 나타나게 되기 때문이다.

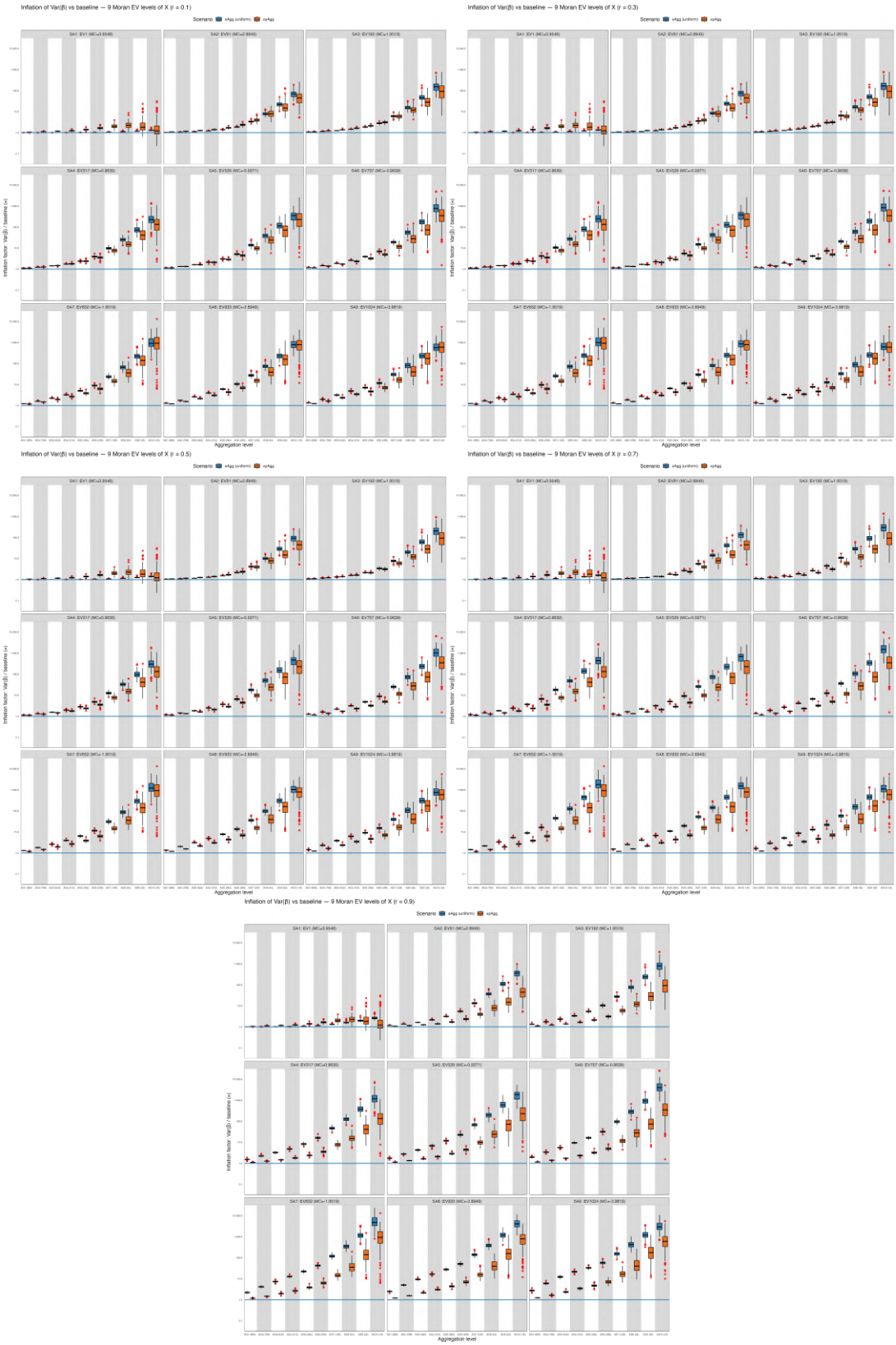

Fig 3. Variance Inflation in 9 Moran EV Levels of X and 5 levels of Pearson's r

2. X의 공간적 자기상관 수준이 다른 경우

먼저 xAgg에 대한 관찰 결과이다. 첫째, IV-1에서와 마찬가지로 스케일 효과가 뚜렷하게 관찰된다. 이는 설명력의 변동 혹은 근원적 공간적 자기상관 정도와 무관하게 합역으로 인해 X의 분산이 감소하는 경향성(Lee et al., 2019)에서 기인한 결과라 할 수 있다. 둘째, IV-1에서와 마찬가지로 구획 효과가 합역 수준이 증가함에 따라 확대되는 패턴을 보인다. 셋째, IV-1에서와 달리 Pearson's r이 증가함에 따라 모든 X의 SA에서 평균적인 분산 팽창 계수는 증가하였다. 특히 IV-1에서 나타났던, SA1 또는 SA2에서 Pearson's r이 증가함에 따라 분산 팽창 계수가 큰 폭으로 감소하였던 패턴이 관찰되지 않았다. 이는 xAgg 상황에서는 Y가 강한 양의 공간적 자기상관을 가지고 있는지가 회귀 계수 분산을 변화시키는 데 보다 결정적인 역할을 한다는 사실을 알려준다. 특히 r이 동일하더라도 X의 I가 높은 경우와 Y의 I가 높은 경우 간에는 큰 차이가 있었다는 사실도 Y의 영향이 더욱 결정적일 수 있다는 점을 시사한다. 다만 이는 r이 동일한 상황에서 X와 Y를 다양한 방식으로 재구성하며 민감도 분석을 실시한 결과를 통해 검증될 필요가 있다.

한편, xyAgg에서의 분산 팽창 계수는 모든 Pearson's r에서 동일한 결과를 얻었다. 그 이유는 분산 팽창 계수를 수식으로 도해할 경우 자명한데, 잔차제곱합을 $(1-R^2)\times SST$로 나타내면 $R^2_{\text{init}} \simeq \mathrm{R}^2_{\text{xyAgg}}$로 r이 관여되는 부분이 소거되어 Pearson's r과 무관하게 같은 값을 지니게 된다. (현재 세팅에서는 X를 EV로 고정한 반면 IV-1에서는 Y를 EV로 고정하고 원하는 r에 맞게 X를 구성하였기 때문에 xyAgg가 r에 따라 다르게 나타났던 것이다.)

$$Var(\widehat{\beta_{\text{xyAgg}}}) = \frac{\sum(\overline{y_i}-\widehat{\overline{y_i}})^2}{(m-1)\sum x_i^{-2}} = \frac{(1-R^2_{\text{xyAGG}})\times \text{SST}_{\text{xyAgg}}}{(m-1)\sum x_i^{-2}}$$

$$Var(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{\sum(y_i-\hat{y_i})^2}{(n-1)\sum x_i^2} = \frac{(1-R^2_{\text{init}})\times \text{SST}_{\text{init}}}{(n-1)\sum x_i^2}$$

따라서 해당 경향성은 스케일 효과의 발현, 합역 수준 증가에 따른 구획 효과 증가라는 두 가지로 요약될 수 있다.

마지막으로 xAgg와 xyAgg를 비교하면 다음의 사실을 관찰할 수 있었다. 대부분 Y의 공간적 자기상관 수준을 조절한 경우와 비슷한 양상을 보였는데, xyAgg가 고정되어 있어 다음의 두 가지 결론만이 도출된다. 첫째, SA1을 제외하면 대부분 xAgg의 분산 팽창 계수가 xyAgg에 비해 높았다. 둘째, xAgg에 비해 xyAgg에서 구획 효과가 증폭되어 나타났다.

3. 내삽 방식을 달리하였을 경우

X를 무작위 합역한 이후 초기 데이터의 해상도로 리샘플링할 때 IDW 내삽을 활용한 이후 Y와 회귀분석을 진행한 경우가 xAgg(IDW)에 해당한다. xAgg(IDW)도 r과 I와는 무관하게 합역 수준이 증가함에 따라 분산 팽창 계수가 증가하는 것으로 나타나고, 분산 팽창 계수의 변동 폭도 확대되는 것으로 나타난다. 이외에도 IV-1의 결과와 유사하

다. 다만 모든 조건에서 xAgg(IDW)가 xAgg보다 회귀분석 신뢰도가 더 높다. 이는 공간 데이터의 보간 방법 중에서 단순한 내삽보다 공간적 구조를 이용하는 내삽 방법이 더 효과적일 것이라는 기대를 만족한다. 하지만 그 차이가 크게 드러나지는 않는다. 그나마 SA 1 즉, Y의 공간적 자기상관 수준이 매우 큰 경우에는 xAgg(IDW)가 xAgg의 분산 팽창 계수 차이가 큰 편이다. 그 이유는 다음과 같이 설명할 수 있다. Y의 공간적 자기상관 수준이 매우 크면 인접한 셀간의 값이 유사하고 이는 부드러운 패턴이라고 할 수 있다. 하지만 X를 무작위 합역하여 구역의 평균을 단순 복사한 방법은 구역 간 경계부에서 값이 불연속적이게 된다. 특히 AG 수준이 큰 경우에 경계부에서 값이 급격하게 변한다. 반면 IDW는 이를 완만하게 보간함으로써 Y의 부드러운 패턴을 잘 설명한다.

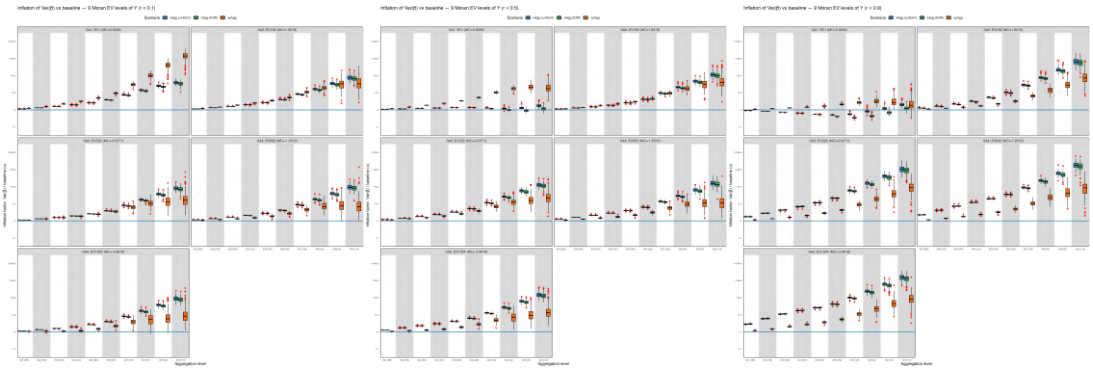

Fig 3. Variance Inflation of xAgg, xAgg(IDW), and xyAgg (EV = Y)

4. 기타 통계량에 미치는 영향 및 민감도 분석

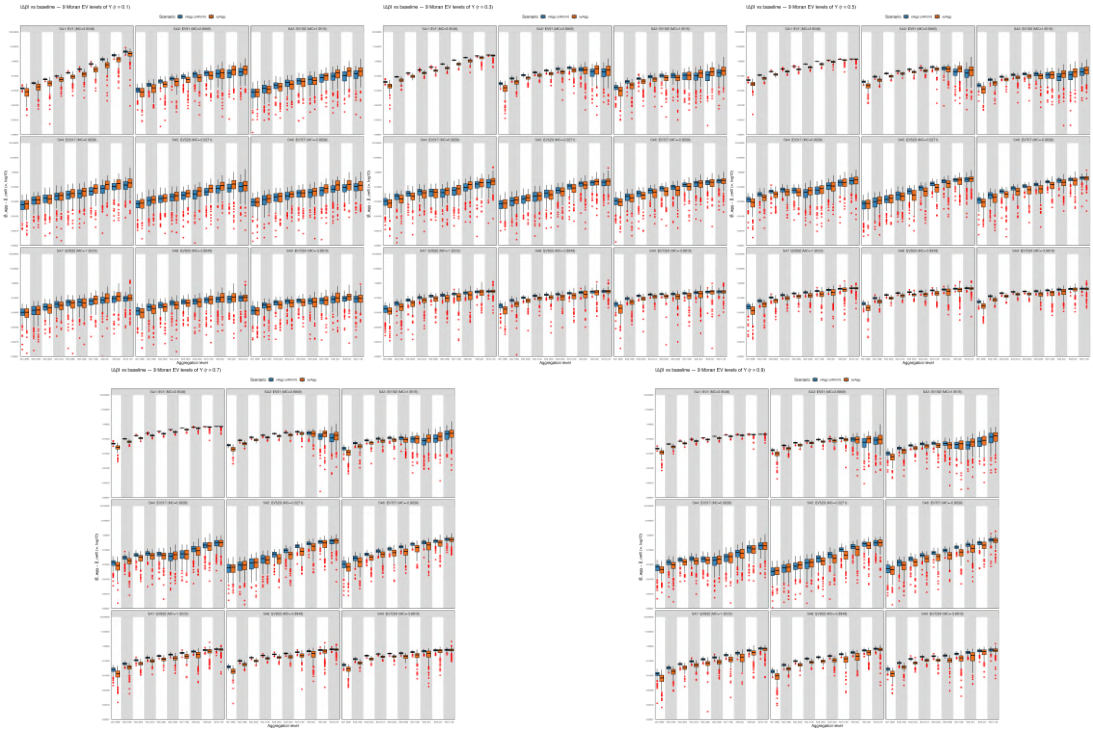

Fig 4. Log-scaled boxplots of the absolute change in slope estimates

회귀 계수 추정량의 t-value는 귀무가설 하에서 $t = \frac{\hat{\beta}}{se(\hat{\beta})}$이다. 따라서 분모에 해당하는 추정량의 분산뿐 아니라 분자에 해당하는 회귀 계수 추정량 자체가 크게 변동할 경우 초기 세팅에서의 회귀 분석 결과와 다른 결론이 내려질 수 있다. 따라서 동일한 방식으로 무작위 합역을 진행하되 위의 식을 바탕으로 최초 X, Y로 추정된 회귀계수에 비해 편향이 얼마나 발생하는지에 대한 boxplot을 작성하였다. 관찰 결과 합역에 따라 편향이 대체로 발생하였고 낮은 초기 r에서 구획 효과가 더 크게 나타나는 경향성은 관찰되었으나, 분산의 팽창 정도에 비하면 편향의 영향은 상대적으로 작았다.

다음으로는 민감도 분석(sensitivity analysis)를 실시하였다. 앞서 데이터를 구성할 때, 하나의 변수를 모런 고유벡터로 고정한 뒤 원하는 r 수준에 맞추어 나머지 변수를 구성하는 방식을 채택하였다. 위의 분석 결과가 일반성을 잃지 않는지 확인하기 위해 시드를 변화시키면서 나머지 변수가 가진 특성(분산, 공간적 자기상관의 정도)이 크게 달라지거나 그에 따라 전체적인 패턴의 변화가 발생하는지에 대한 민감도 분석을 실시할 필요가 있다. 비록 단순히 시드를 변화시키는 것이기 때문에 미처 파악하지 못한 경우가 있을 수는 있으나, 분석 결과 대부분 기존에 서술한 결과에서 크게 벗어나지 않는 경향성을 나타내었다.

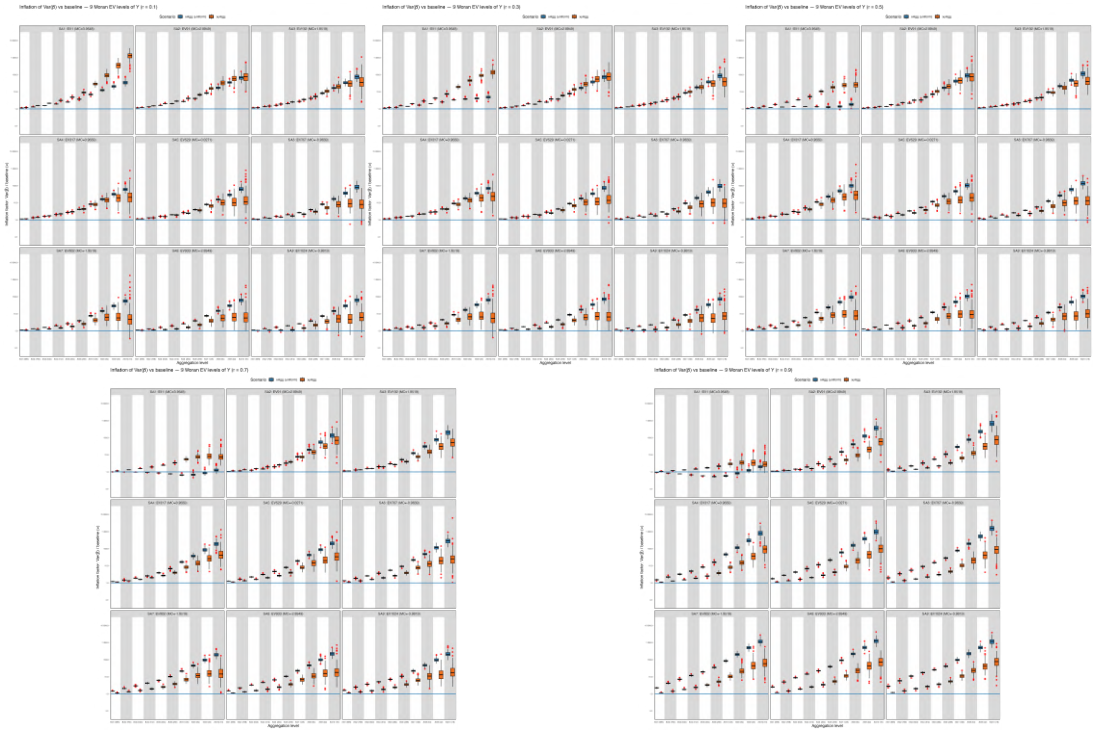

Fig 5. Variance Inflation in 9 Moran EV Levels of Y with Different X

V. 결론

본 연구는 공간 단위의 선정이 회귀분석에 미치는 영향을 파악하고자 하였다. 특히 근원적 공간적 자기상관 정도와 교차 상관 정도에 따라 MAUP의 영향에 차이가 있을 것이라는 가설을 바탕으로 r과 Moran's I를 조절하면서 가상의 데이터를 구성하였다. MAUP의 스케일 효과 및 구획 효과를 파악하기 위한 방법으로는 무작위 합역을 통한 시뮬레이션 실험을 택하였다. 또한 X와 Y의 공간 단위가 불일치하는 상황, 특히 X의 공간 해상도가 더 낮은 상황을 가정하여 해당 상황에서 회귀 분석 결과가 어떻게 달라지는지를 살펴보았다. 실험을 통해 밝혀진 결과는 크게 다음 네 가지로 요약할 수 있다. 첫째, 공간적 컨피겨레이션 (xAgg와 xyAgg), r, Moran's I 등에 무관하게 스케일 효과는 뚜렷하게 나타났다. 공간 단위의 크기가 증가할수록 분산이 더욱 증가하는 평균적 경향성을 보였다. 둘째, 공간적 컨피겨레이션, r, Moran's I 등에 무관하게 구획 효과는 공간 단위 크기가 증가함에 따라 더 강하게 나타났다. 한편 X, Y 모두가 합역되는 xyAgg 상황에서 증폭되어 구획 효과의 크기는 더욱 커졌다. 셋째, 대부분의 경우 xAgg에 비해 xyAgg에서 분산 팽창 계수가 더 낮은 경향성이 나타났다. 심지어 내삽을 통해 최대한 원래 데이터의 변동을 복원하고자 했음에도 불구하고 xyAgg의 분산 팽창 계수가 더 낮았는데, 이는 xAgg의 경우 양의 공간적 자기상관이 강하지 않은 데이터에서 정보량의 중복으로 인한 효율성 손실을 피하기 어렵다는 점을 함의한다. 넷째, xAgg에 비해 xyAgg에서 분산 팽창 계수가 더 크게 나타나는 현상은 Y의 양의 공간적 자기상관의 정도가 매우 강한 경우에만 관찰되었다. 양의 공간적 자기상관이 강할 때만 특이한 패턴을 보이는 것은 일변량 상황에서 자기상관의 정도가 강할 때 분산 변동이 달라지는 양상이 상이했던 결과와 상응한다. '무 공간적 자기상관'과 '음의 공간적 자기상관'의 차이가 분산 변동의 맥락에서는 큰 차이를 발생하지 못하는 것이 다시 확인된 것이다. 기존 연구에서 회귀분석 상황에서 특정한 경향성이 발견하지 못했다고 보고한 이유 중 하나로 추측되며, 본 연구에서는 근원적 공간적 자기상관 수준을 모런 고유벡터를 통해 모델링함으로써 이러한 결과를 발견해낼 수 있었다.

본 연구는 수식적으로 밝혀내기 어려운 공간 단위 설정의 영향을 시뮬레이션을 통해 파악하였고, 일정 수준의 경향성을 파악하였다는 의의가 있다. 특히 본 연구의 결과는 특별히 해상도가 다른 두 개 이상의 공간 데이터를 분석하고자 할 때 결과 해석에 주의해야 함을 시사한다. 회귀분석에서 독립변수만을 합역하는 것은 두 변수 간 공간적인 불일치를 유발하여 회귀분석의 신뢰도를 낮출 수 있음에 주의해야 한다. 연구 목적에 어긋나지 않는다면 오히려 두 변수 모두에 대해 알려진 상위의 체계로 분석하거나, 다변량 회귀 모형의 경우 해당 독립변수를 제거하는 것을 검토해볼 필요가 있다. 특히 종속변수의 공간적 자기상관 수준이 매우 큰 경우에 합역 수준에 따라 회귀 계수의 분산이 크게 변동하는 것으로 나타났기에 주의해야 한다. 한편, X의 공간적 자기상관 수준을 조절하였을 경우보다 Y의 공간적 자기상관 수준을 조절하였을 경우에 분산 팽창 계수의 변동폭이 더 큰 것으로 나타났다. 이는 종속변수의 공간적 자기상관 수준이 독립변수의 공간적 자기상관 수준에 비해 회귀분석에 있어 더 중요한 요인이 되었기 때문으로 보인다. 하지만 특정 r에서 X와 Y의 I값을 조절하며 비교한 것이 아니기에 해석하는데 한계를 지니며, 데이터의 초기 조건을 변경해보며 더욱 정련화된 상황을 실험해 볼 필요가 있다.

합역 이후 리샘플링 과정에 IDW의 내삽을 추가한 분석에서는 공간적 특성을 고려한 보간법이 단순히 평균을 대입한 방법보다 회귀분석 신뢰도를 높인다는 기대를 만족한다. 하지만 Y의 공간적 자기상관 수준이 매우 큰 경우를 제외하고는 그 차이가 뚜렷하게 나타나지 않는다. 본 연구에서는 IDW 한 가지 방법으로만 내삽을 진행하였으나, 실제로 공간 분석에 있어서 다양한 내삽 기법이 사용되고 있기에 이를 더 확인해 볼 필요가 있다. 또한 해상도가 다른 공간 데이터 간 Pearson's r값과 Moran's I값을 이용해, 각 조건에 해당하는 데이터가 회귀분석의 신뢰성을 가장 높일 수 있는 내삽 방법을 찾는 연구로 발전할 가능성이 있다. 그러나 본 연구에서처럼 두 변수간 Pearson's r, 각 변수의 Moran's I, 합역 수준 등 다양한 요인을 통제하며 결과를 해석하는 것은 직관적이지 않으며, 각 효과들이 명확하지 않은 부분이 있다. 따라서 회귀분석을 왜곡하지 않는 공간단위 선정 문제에 어울리는 새로운 통계량을 고안하는 후속 연구가 진행되어야 할 것이다.

참고문헌

이상일, 1999, "기능지역의 설정과 '공간단위 수정가능성의 문제 (MAUP)'," 한국지리환경교육학회지, 7(2), 757-783.

이상일 · 이몽현, 2020, "무작위합역 절차의 다양성에 대한 시뮬레이션 연구," 한국지도학회지, 20(3), 93-107.

Arbia, G., 1989, Spatial Data Configuration in the Statistical Analysis of Regional Economics and Related Problems, Kluwer, Dordrecht.

Arbia, G. and Petrarca, F., 2011, "Effects of MAUP on Spatial Econometric Models," Letters in Spatial and Resource Sciences, 4(3), 173-185.

Chen, J., Zhang, Y., & Yu, Y., 2011, "Effect of MAUP in Spatial Autocorrelation," Acta Geographica Sinica, 66(12), 1597-1606.

Duque, J. C., Laniado, H., & Polo, A., 2018, "S-maup: Statistical Test to Measure the Sensitivity to the Modifiable Areal Unit Problem," PLOS ONE, 13(11), e0207377.

Fotheringham, A. S. and Wong, D. W. S., 1991, "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis," Environment and Planning A, 23(7), 1025-1044.

Gehlke, C. E. and Biehl, K., 1934, "Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census-Tract Material," Journal of the American Statistical Association, Supplement, 29, 169-170.

Green, M. and Flowerdew, R., 1996, "New Evidence on the Modifiable Areal Unit Problem," in Longley, P. & Batty, M. (eds.), Spatial Analysis: Modelling in a GIS Environment, GeoInformation International, Cambridge, 41-54.

Holt, D., Steel, D. G., Tranmer, M., & Wrigley, N., 1996, "Aggregation and Ecological Effects in Geographically Based Data," Geographical Analysis, 28(3), 244-261.

Jacobs-Crisioni, C., Rietveld, P., & Koomen, E., 2014, "The Impact of Spatial Aggregation on Urban Development Analyses," Applied Geography, 47, 46-56.

Lee, S.-I., 2001, "Developing a Bivariate Spatial Association Measure: An Integration of

Pearson's r and Moran's I," Journal of Geographical Systems, 3, 369-385.

Lee, S.-I., Lee, M., Chun, Y., & Griffith, D. A., 2019, "Uncertainty in the Effects of the Modifiable Areal Unit Problem under Different Levels of Spatial Autocorrelation: A Simulation Study," International Journal of Geographical Information Science, 33(6), 1135-1154.

Openshaw, S., 1984, The Modifiable Areal Unit Problem (CATMOG 38), Geo Books, Norwich.

Openshaw, S. and Taylor, P. J., 1979, "A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem," in Wrigley, N. (ed.), Statistical Applications in the Spatial Sciences, Pion, London, 127-144.

Steel, D. G. and Holt, D., 1996, "Rules for Random Aggregation," Environment and Planning A, 28(6), 957-978.