한국 센서스데이터의 MAUP

강 계 화

Kang gye-hwa

ABSTRACT: Census data are usually provided at an aggregated level. However, the aggregated data are essentially arbitrary geographical areas. The areal units used to report census data have no natural or meaningful geographical identity. Unfortunately, this means that analyses of these area aggregations may be conditional upon the set of zones, which are presented. This effect is known as the modifiable areal unit problem (MAUP) and has two related aspects. First, scale effect is the variation in numerical results that occurs due to the number of zones used in an analysis. Second, results may also differ between different ways of aggregating exactly the same data to the same scale; this may be called the aggregation effect (Openshaw, 1984). This study aims to provide a practical tool for the study of MAUP. I have created a set of 91 areal units based on 280 basic units in Nonhyun-2 dong to solve zoning problem and scale problem. We can easily recognize the importance of areal classification as statistics were different according to areal classification.

Ⅰ. 서론

전통적으로 통계작성자들은 보다 정확하고 의미 있는 통계 작성을 위하여 표본오차를 제대로 측정하고 비표본오차를 최소화하는데 주력해 왔다. 하지만 정말 센서스 자료 작성에 관해 중대한 하자가 없는지 나는 근본적인 문제 하나에 대해 의문을 제기하고자 한다.

그것은 바로 지역구분의 문제이다. 즉, 센서스 자료가 모든 개개인을 조사하고 있음에도 개인정보의 보호라는 명제 하에 자료가 철저히 지역으로 통합되어 발표되고 있는데, 사실 그 지역이라는 것은 명확한 지리적인 실체가 없이 다분히 임의적이라는 사실이다.

하지만 이러한 지역분석에서 지역단위의 설정 여하가 분석결과에 끼치는 영향에 대하여 일찍부터 통계 및 지리학자들은 가변지역단위문제(MAUP : Modifiable Areal Unit Problem)라고 하여 문제의 심각성을 지적한 바 있다(Yule and Kendall, 1950; Robinson, 1950; Openshaw, 1984). 하지만 초기에 문제해결의 곤란 등으로 주목을 받지 못하다가 GIS기술의 발달로 주요 현안으로 부각되고 있다. 비집계데이타와 GIS의 연계에 의해서 자유자재로 지역의 설정이 가능하여 졌기 때문이다. 예를들어 GPS로 위치정보를 취득해가면서 조사현장에서의 입력의 활용은 센서스 필드조사에의 큰 가능성을 내재하고 있으며 기존 센서스데이터의 활용한계를 극복할 수

있는 강력한 수단이 될 수 있으리라 본다(Kang, 2002).

그러나 최근에도 한국에서 개별자료에 관한 통계의 개방은 개인 식별이 가능한 개인정보의 사용을 엄격히 제한하는 통계법의 벽에 막혀 연구가 진척되지 못하고 있는 실정이다. 반면에 구미제국에서 통계는 공공재라는 인식의 확산하에서 비집계데이타를 될 수 있는대로 개방할 방향으로 익명성과 시큐리티의 문제를 명확히 하기 위한 기술적 검토가 적극적으로 진행되고 있다. 이 글에서는 한국 센서스데이터의 MAUP 확인을 통하여 지역구분문제의 중요성을 제고시키고자 한다.

II. 자료 및 분석방법

1. 자료

이 연구에서 사용할 자료는 한국의 2000년 인구센서스 자료인데, 보다 효과적이고 정밀한 지역연구를 위해 서울특별시 강남구의 1개동인 논현2동을 표본으로 선정하였다. 이 지역은 크기가 1.47km²이며, 2000년 인구센서스 당시 인구는 20,415명, 가구는 7,396호 였다. 사실 지리적인 특성 연구를 위해 매우 유용한 대부분의 정보는 10% 지역에서만 조사하는 표본조사 지역에서 작성되기 때문에 모든 지역을 빠짐없이 연구 대상으로 해야하는 본 연구에서는 부득이하게 전수조사항목만을 활용하였다. 이 중 직접 지역단위 연구에 활용할 수 있는 항목은 성, 연령, 교육정도, 혼인상태, 가구구분등의 항목에 불과했다. 주택에 관한 항목도 이들이 주인가구만을 조사함으로써 전체 가구에 대한 정보가 없어 제외시켰다.

2. 분석방법

한국의 센서스 자료에서 MAUP를 확인하는 방법은 지리적으로는 하나의 지역을 다양한 수준의 크기와 통일한 크기내에서 다양한 조합으로 묶는다거나 아니면 개별 자료를 이용한다거나 하는 것이며, 이러한 각각의 다른 수준이나 조합의 지역구분의 문제를 확인하는 방법은 이들 지역에 속하는 통계가 어떻게 달라지는가를 확인하는 것이다. 이 연구에서는 다변량 회귀모델의 적합도와 변수들에 관련된 값들을 비교한다. 가장 먼저 지역단위의 문제를 확인한다. 지역단위의 문제는 지역구분의 수준이 달라지는데 따른 Scale문제와 동일한 지역구분 수준상의 상이한 지역의 통합에서 발생하는 Zoning 문제를 확인하게 된다.

1) 이를 위해 먼저 대상지역을 280개의 기초단위구으로 나누게 된다. 각 기초단위구는 2000년 센서스에서 대충 30가구 정도의 크기가 되도록 한다.
2) Scale 문제를 확인하기 위해서는 이 280개 기초단위구를 seed unit으로 하여 2개씩 묶어 140개 수준, 4개씩 묶어 70개 수준, 8개씩 묶어 35개 수준으로 지역의 scale 크기를 달리했다. 각 수준별로는 각각 20개의 다른 세트로 지역통합을 실시했는데, 이는 20개의 세트의 자료를 구함으로써 적은 자료를 사용할 때 범할 지도 모르는 극단적인 자료 사용의

가망성을 사전에 배제시켰다. Scale 문제 확인을 위해서 중회귀 모델을 만들어 이 모델에 대한 정보가 Scale을 달리 했을 때 어떻게 변하는지를 규명함으로써 Scale 문제를 도출하였다.

3) Zoning 문제를 위해서는 35개 기초단위구 수준의 자료를 50개 세트를 생성하여 각 자료별로 각종 변수들의 기술통계량(평균, 표준편차) 및 위 모델의 R2 값, 매개변수들의 Beta 값 등의 변동수준을 파악한다.

III. 한국 센서스데이터의 MAUP 확인

MAUP확인을 위하여 개별자료를 통합한 지역수준의 자료에 대해서는 다양한 통계적인 방법의 적용이 가능하도록 다음과 같이 양적인 척도로 변수를 조정하였다.

- 주인가구 비율 : 주인가구/전체*100
- 남자 가구주 비율 : 남자/전체*100
- 가구주의 평균 연령
- 가구주 학력의 대학이상 비율 : (대학+ 대학교+ 석사+ 박사)/전체*100
- 가구주중의 유배우 비율 : 유배우/전체*100

위의 변수들은 기술통계량에서는 평균과 표준편차를 계산하여 비교하였고 다변량 분석의 중회귀모델에서는 모델의 설명력을 나타내는 R2, 통계적으로 유의한 변수의 종류, 변수의 Beta 값 및 이들의 표준오차를 계산하여 정리하였다. 한편 다변량 회귀분석에 사용한 모델의 가설 및 변수들은 다음과 같다.

① 가설

- 가구주의 연령이 높으면 높을수록 자가비율이 높음
- 가구주의 학력이 높으면 높을수록 자가비율이 높음
- 가구주가 유배우이면 자가비율이 높음
- 가구주가 남성인 경우가 자가비율이 높음

② 모델

Y= B0 + B1 Page + B3 Pedu + B4Pmarried + B5 Pmale + e

여기서 종속변수 Y(Pown) : 가구주의 자가비율

독립변수 : Page(가구주의 평균연령)
Pedu(가구주의 대졸이상 비율)
Pmarried(가구주의 유배우비율)
Pmale(가구주의 남성비율)

1. Scale 문제

쥰별 분석결과의 차이를 보기 위해서는 기초단위구를 280개 수준, 140개 수준, 70개 수준

및 35개 수준으로 인접성을 조건으로 각기 다른 20개 세트별로 자료를 작성하였다. 검토한 변수는 주인가구 비율, 남자 가구주 비율, 가구주의 평균 연령, 가구주 학력의 대학이상 비율, 가구주의 유배우 비율의 4가지였다.

표1에 나타난 것처럼 회귀모델의 경우에 각 scale별로 통계적으로 유의한 변수가 다소 달리 나타났다. 먼저 유의한 변수의 수에 있어서는 280개 존은 3개(유배우 비율, 대졸이상 교육정도, 연령 등)의 변수가 유의한 것으로 나타났으나 140개 존에서는 전체 20개 세트중 8개만이, 70개 존에서는 6개, 35개 존에서는 5개 만이 동일한 모델을 만들어 냈다. 140개 존의 경우에는 유배우 비율과 대졸이상 교육정도 등의 2개 변수를 모델로 하는 세트가 12개로 가장 많았고 70개 존의 경우에도 같은 모델이 14개 세트로 가장 많았다. 반면에 35개 존의 경우에는 다른 종류의 2개 변수(유배우 비율과 연령 등)를 모델로 하는 세트(9개 세트)가 가장 많았다. 35개 존에서는 그 밖에도 대졸이상 교육정도와 연령을 모델로 하는 세트가 2개, 대졸이상 교육정도만을 모델로 하는 세트가 2개, 유배우 비율만 모델로 하는 경우가 2개 세트 등 다양하게 나왔다. 이로써 동일한 모델을 연구하더라도 스케일은 물론 존닝 방법을 달리 했을 경우에 자료 이용자들이 상이한 결론에 도달한 것은 너무나도 자명하다고 하겠다.

표 3 표1. 존별 회귀모델의 차이

회 귀 모 델	280존	140존	70존	35존
Y = a1*married + a2*edu + a3*age + b	1	8	6	5
Y = a1*married + a2*edu + b		12	14	
Y = a1*married + a2*age + b				9
Y = a1*married + b				2
Y = a1*edu + a2*age + b				2
Y = a1*edu + b				2
합 계	1	20	20	20

Key : marrried는 가구주의 유배우 비율
edu은 대졸이상 가구주 비율
age는 가구주의 연령

한편 scale 별로 각 변수의 Beta 값 및 표준오차를 보면 scale 수준별로 다소 상이한 패턴을 보이는데, 280개 존의 경우 대부분의 변수에서 다른 수준의 스케일에 비해서 아주 높거나 낮은 극단적인 값을 보였으며 scale이 작을 수록 Beta 및 이들의 표준오차의 분포가 더욱 다양해지는 패턴을 보여주고 있다. 이는 스케일이 커질수록 평균효과의 특성으로 자료들이 통합되고 있음을 보여 주고 있다. 따라서 스케일이 큰 지역집계단위는 개인정보들의 특성을 상실하게 되어 소지역 연구나 미크로데이터의 연구에서 큰 스케일의 집계단위로의 연구가 불합리하다는 것을 알 수 있었다.

2. Zoning 문제

전체 50개 세트별로 회귀 모델의 분석결과 표2에 나타난바와 같이 회귀식에서 통계적으로 유의한 변수들이 달리 나타나기도 하였다. 우선 전체 50개 세트중 17개 세트가 대졸자 가구주 비율(edu)과 가구주 연령(age)를 통계적으로 유의한 변수로 나타낸 모델을 제시하였으며, 15개 세트는 유배우 가구주(marriage)와 가구주 연령(age)을 설명변수로 제시하였다. 반면에 유배우 가주주 비율(marriage) 만을 포함한 모델은 3개 세트, 대졸 가구주 비율(edu)만을 포함한 모델은 1개 세트에서만 나타났다. 이로써 이 모델은 50개 쥰을 어떻게 선택하느냐에 따라 설명변수의 종류가 달라지는 모습을 보여 주고 있다.

표 4 표 2. 50개 쥰별 회귀모델의 차이

회 귀 모 델	세트 수	전체 세트 수 중 구성비(%)
$Y = a_1$*marriage + a_2*edu + a_3*age + b	12	24.0
$Y = a_1$*marriage + a_2*age + b	15	30.0
$Y = a_1$*marriage + b	3	6.0
$Y = a_1$*edu + a_2*age + b	17	34.0
$Y = a_1$*edu + b	2	4.0
$Y = a_1$*age + a_2*male + b	1	2.0

Key: marriage는 유배우 가구주 비율, edu은 대졸이상 가구주 비율,
age는 가구주의 연령, male는 남자 가구주 비율임

Ⅳ. 결론

전술한 바와 같이 스케일과 죠닝시스템에서의 지역수준의 크기와 방법을 달리하여 다양한 각각의 지역단위에 스텝와이즈 방법을 사용한 중회귀모델을 적용하여 보았다. 우리가 잘 아는 바와 같이 중회귀분석은 사회현상이나 자연현상을 설명하는데 있어 관련된 변수들 간의 제조건중, 어떤 요인이 가장 작용하고 있는가를 추출하는 현실사회의 변화에 더 큰 설득력을 가진다고 말하여진다. 그러나 스케일과 죠닝시스템에서 변동에 대한 모델은 지역단위에서 추출한 자료의 중회귀분석에서 매우 비현실적인 결과를 만들어 내는 것이 보여졌다.

지역분석에서 센서스자료를 중회귀분석에 활용하는 것은 사회학자들에게 매우 일반적인 관행이지만 여기에서 MAUP에 대하여는 거의 언급되고 있지 않는 실정이다. 현재의 GIS의 기술발전은 센서스자료의 지역정의나 지역특성을 시뮬레이션할 수 있는 여건과 환경을 제공하고 있지만 정치적·행정적 벽에 막혀있다. 그러나 앞으로 MAUP의 적극적인 연구와 관심의 고조, 또는 그 중요성과 심각성에 대한 인식의 저변확대로 개별자료에 대한 이용이 더욱 강조될 것이다. 물론 개별자료에 대한 공공연한 사용은 개인정보의 보호가 엄격히 고려되어야만 한다. 이제는 센서스 담당자들이 자료의 통합에 대한 일상적인 관행이 잘못된 방향으로 정책입안자들을 오도할 수 있다는 것을 깨달아야 하며, 따라서 센서스 자료가 제표되는 지역단위나 이용자 정

의의 지역단위의 타당성을 검토하기 시작해야만 할 때이다.

<참고문헌>

Kang, G. 2002. Availability of a mobile GIS loading GPS for census operations. -A case study of the development and application in Korea- Theory and Applications of GIS, 10-1. 103-109.

Openshaws, S. 1984. The modifiable areal unit problem. Concept and Technique in Modern Geography 38. Norwich : Geo Books.

Robinson, W.S. 1950. Ecological correlations and the behavior of individuals. American Sociological Review,15. 351-357.

Yule, G. U. and Kendall, M.G. 1950. An introduction to the theory of statistics. Griffin: London.