이학석사 학위논문

공간 다층모형을 활용한
서울시 골목상권 분석

An Analysis of the Local Businesses in Seoul Using Spatial Multilevel Model

2018년 2월

서울대학교 대학원

통계학과

신 유 진

국문초록

공간 다층모형을 활용한

서울시 골목상권 분석

An Analysis of the Local Businesses in Seoul

Using Spatial Multilevel Model

최근 자영업자들이 늘어나고 개성을 중시하는 소비자들이 늘어나면서 골목상권이 주목을 받고 있다. 골목상권이란 큰 대로변에 있는 상권이 아닌 골목마다 존재하는 상권을 말하는데, 독특한 분위기의 골목상권이 소비자들을 사로잡고 있다. 본 논문에서는 서울시에 있는 여러 골목상권을 분석하고자 한다. 골목상권은 골목 수준, 골목 내 업종 수준으로 나누어 볼 수 있고 이를 분석하기 위해서 다층 모형으로의 접근이 필요하다. 또한 골목의 위치정보가 존재하므로 골목 수준에서 공간적인 상관성을 고려할 필요가 있다. 따라서 공간 다층 모형을 소개하고 이를 활용하여 분석을 진행하려 한다. 공간 다층 모형으로 SAR 모형을 확장시킨 Hierarchical SAR(HSAR) 모형을 활용한다. 이 모형의 경우 모수 추정 시 베이지안 방법인 마르코프 사슬 몬테카를로 방법(MCMC)을 사용한다. 모형 적합 후 최종적으로 어떤 요소들이 골목상권의 매출에 영향을 주는지 알아본다.

주요어 : 공간 자료, 다층 모형, SAR 모형, HSAR 모형, 베이지안 MCMC 방법, 서울시 골목상권

학 번 : 2016-20268

Contents

List of Tables

List of Figures

Chapter 1

서론(Introduction)

최근 위치정보가 있는 다양한 데이터들이 수집되고 이를 효과적으로 분석하기 위해 공간통계(Spatial Statistics)적 방법론이 연구되고 있다. 예를 들면, 위도와 경도에 대한 정보를 포함하고 있는 오존데이터, 강수량 데이터, 미세먼지 데이터 등이 있다. 위치 정보를 활용하여 공간적 상관성을 고려해 분석을 진행할 수 있다.

본 논문에서는 서울시에 존재하는 골목상권을 위치 정보를 고려하여 분석해 보고자 한다. 이 때 골목상권은 큰 대로변에 있는 상권이 아니라 골목마다 존재하는 상권을 의미한다. 더욱이 한 골목 안에는 다양한 업종이 존재하고 있다. 따라서 골목단위, 골목 내의 업종 단위를 나누어 분석을 진행해야 한다. 따라서 다층모형(Multilevel Model)을 활용하여 이러한 골목상권을 분석할 수 있다.

1.1. 연구 소개

다층모형(Multilevel Model)은 사회과학에서 많이 쓰이는 모형으로 수준(level)이 두개 이상인 자료를 분석할 때 쓰인다. 이 모형은 종속변수(y)의 분산을 수준별로 나누어 각 수준에서 구체적인 독립 변수를 사용하여 이 분산을 설명한다. 이를 보통 임의효과(random effect)와 고정효과(fixed effect)라고 부른다. 예를 들어, 여러 그룹별로 실시된 설문조사 자료를 생각해보자. 이는 그룹, 그룹 내 구성원으로 나누어 두 계층 모델(two-level model)로 생각할 수 있다. 임의효과의 의미는 구성원 수준의 회귀 계수가 각 그룹별로 다를 수 있다는 것이다. 혹은 각 그룹별로 오차가 다를 수도 있다. 이는 다층 구조 모형을 어떻게 주느냐에 따라 달라진다.

많은 공간 자료들이 계층적 구조(hierarchical structure)를 가지고 있다. 예를 들어, 지역별 설문조사 자료는 지역별 위치 정보, 지역 정보 등이 존재하고 그 아래에 설문조사를 시행한 개인들에 대한 정보가 계층적으로 존재한다(Elcheroth et al. 2013). 이와 같이 데이터 상의 구조가 서로 포함되는 관계가 있을 때 다층 모형으로 접근하여 좀 더 구체적인 분석이 필요하다.

다층모형(multilevel model)은 기본적으로 상위 수준(higer level)의 개체들 사이에 차이가 존재하는 동시에 상관관계도 존재한다. 즉, 같은 상위 수준 하에 있는 하위 수준(lower level)의 개체들은 상관성이 존재할 것이다. 왜냐하면 같은 상위 수준의 영향을 받고 있기 때문이다. 이러한 다층모형의 구조를 수직적인 구조라고 생각한다면 공간 구조는 수평적으로 볼 수 있다. 따라서 수평적 구조와 수직적 구조를 모두 고려해야 한다. 즉, 다층모형의 상위 수준과 하위 수준간의 연관성을 고려함과 동시에 공간적 구조의 상관성을 고려한다.

SAR(simultaneous autoregressive) 모형을 발전시켜 이를 고려할 수 있다(Dong & Harris. 2015). SAR 모형은 자기 회귀 모수를 통해서 공간적인 상호작용 효과를 반영하는 모형이다(Cressie. 2015). SAR 모형은 공간 가중치 행렬(spatial

weight matrix)을 사용해 공간상의 차이(spatial lag) 혹은 공간적 자기상관성(spatial autocorrelation)을 표현할 수 있다.
SAR 모형에 계층적 구조를 주어 HSAR(Hierarchical SAR) 모형을 만들 수 있다. 공간 가중치 행렬을 정의해 SAR 모형을 만들고 계층적 구조를 반영할 수 있는 항을 포함시킨다. 이를 통해 공간적인 상관성을 고려하면서 계층적인 구조를 모형에 반영할 수 있을 것이다. 앞으로, 분석하고자 하는 서울시 골목상권 데이터에 이 HSAR 모형을 적용시켜 보고자 한다.

1.2. 연구 목표

본 논문에서는 다층(Multilevel) 구조를 가진 골목상권 데이터를 공간 통계적 방법론을 활용하여 분석하고자 한다. 분석 목표는 서울시에 존재하는 골목상권의 매출에 영향을 주는 요소가 무엇인지를 찾고 모델링하는 것이다. 모델링을 통해 각 요소들이 어떤 영향을 주고 있는지 살펴본다. 이 때, 서울시의 골목들은 각각의 고유한 위치 정보를 가지고 있으므로 이 위치 정보를 모델에 반영하여 보다 구체적으로 골목상권의 매출을 분석할 수 있다.

Chapter 2

데이터 설명(Data Description)

서울 열린 데이터 광장(http://data.seoul.go.kr/index.jsp)에서 서울시 골목상권 데이터를 제공 받았다. 데이터는 골목의 기본적인 위치 정보, 유동인구, 주변 아파트 단지, 직장인구를 알 수 있다. 또한 골목 내 업종들에 대한 매출 정보, 점포 정보 등을 얻을 수 있다. 홈페이지 내에는 다양한 시점의 정보들이 존재했는데, 시점을 고정시키고 매출에 대한 모델링을 하기 위해서 2016년 12월 데이터만 가져왔다.

2.1. 구성

서울시 골목상권 데이터는 다층모형을 적용할 수 있는 구조를 가진 데이터이다. 상위 수준으로 골목, 하위 수준으로 골목 내의 업종을 두고 골목 단위의 데이터, 업종 단위의 데이터로 나눌 수 있다. 또한 추가적으로 골목별 위치 정보가 제공된다. 각 항목별 자세한 사항은 다음과 같다.

1. 상위 수준(higer level) : 골목

   골목 단위의 데이터로는 우선 상주인구에 대한 정보가 있다. 상주인구는

행정구역별 주민등록 통계 데이터를 건물단위별 가구수 및 성별, 연령대별 인구 수를 추정하여 구해진 값이다. 총 상주인구 수 뿐만 아니라 가구수도 알 수 있다.

서울시를 블록단위로 나눈 뒤 성별, 연령대별 소득 추정액을 구한 월 평균 소득 금액 데이터가 있다.

또한 성, 연령, 요일, 시간대별로 유동인구를 추정한 데이터도 있다. 이 자료는 통신사 데이터, 교통카드 데이터 등을 활용하여 추정하였으며 도로 단위의 유동인구 추정량을 제공한다.

마지막으로 직장인구에 대한 정보가 있다. 직장인구 정보는 50m 셀(cell) 단위의 성, 연령별 직장인구 정보이다.

2. 하위 수준(lower level) : 업종

   골목 내의 업종에 대해서 해당 업종의 점포에 대한 정보가 존재한다. 해당 업종의 점포 수를 제공한다.

   그리고 각 업종별로 매출에 대한 데이터도 존재한다. 이 데이터에는 업종코드, 업종별 평균 영업 개월 수, 당월 매출 금액, 당월 매출 건수, 주중 매출 비율 등이 있다. 이 정보는 BC카드와 신한카드의 신용카드 매출 정보를 가지고 성, 연령대별, 시간대별 거래 패턴을 분석하였고 카드사 점유 비율을 고려하여 매출을 추정한 것이다.

3. 위치 정보

   각 골목별 위치 정보에 대한 데이터도 존재한다. 이 위치의 좌표계는 GRS80TM이다.

Figure 2.1은 각 골목별로 평균 매출을 구해 서울시 지도에 표시한 그림이다. 크기가 크고 색이 진할수록 평균 매출이 크다.

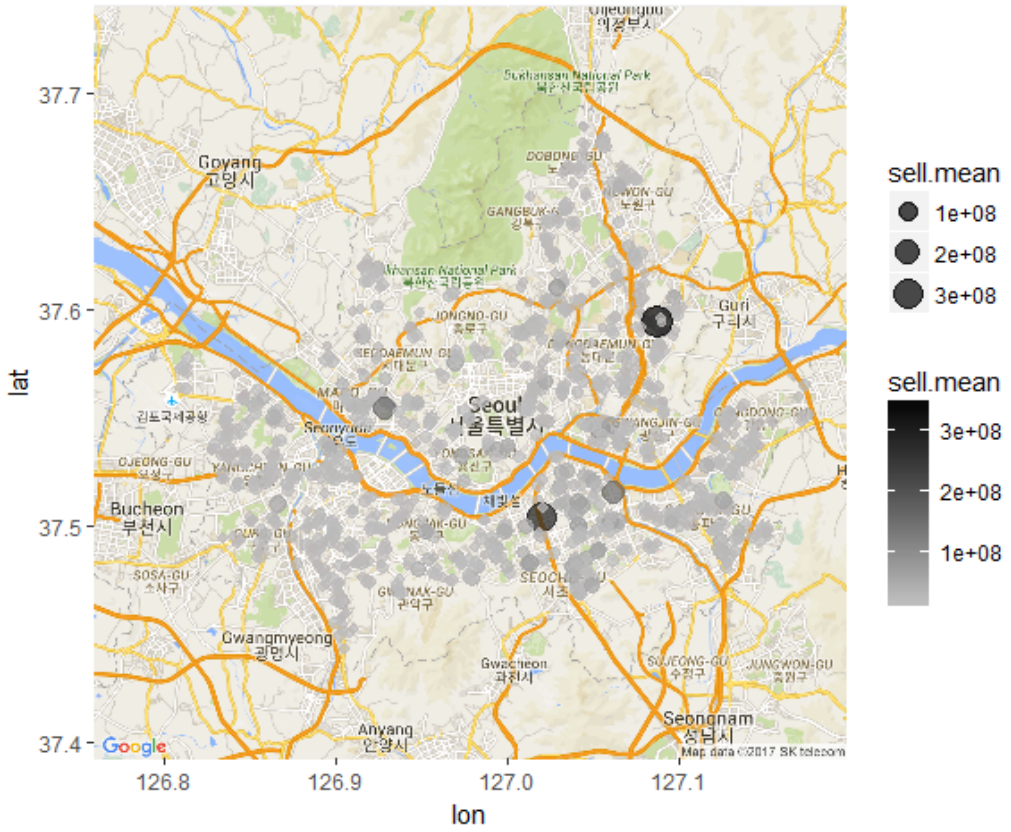

Figure 2.1: 서울시 골목별 평균 매출

한편으로 변수들끼리의 상관관계를 구한 결과 블록단위의 상주인구 수와 가구수가 0.98로 높은 상관관계를 보였다. 또한 그 골목이 속한 블럭의 소비 데이터 역시 0.84로 상주인구 수와 높은 상관관계가 있었다. 따라서 이 세 변수 중 상주인구 수만 사용하기로 하였다.

2.2. 데이터 전처리

다음 Figure 2.2는 전체 매출 정보를 히스토그램으로 그려본 것이다. 히스토그

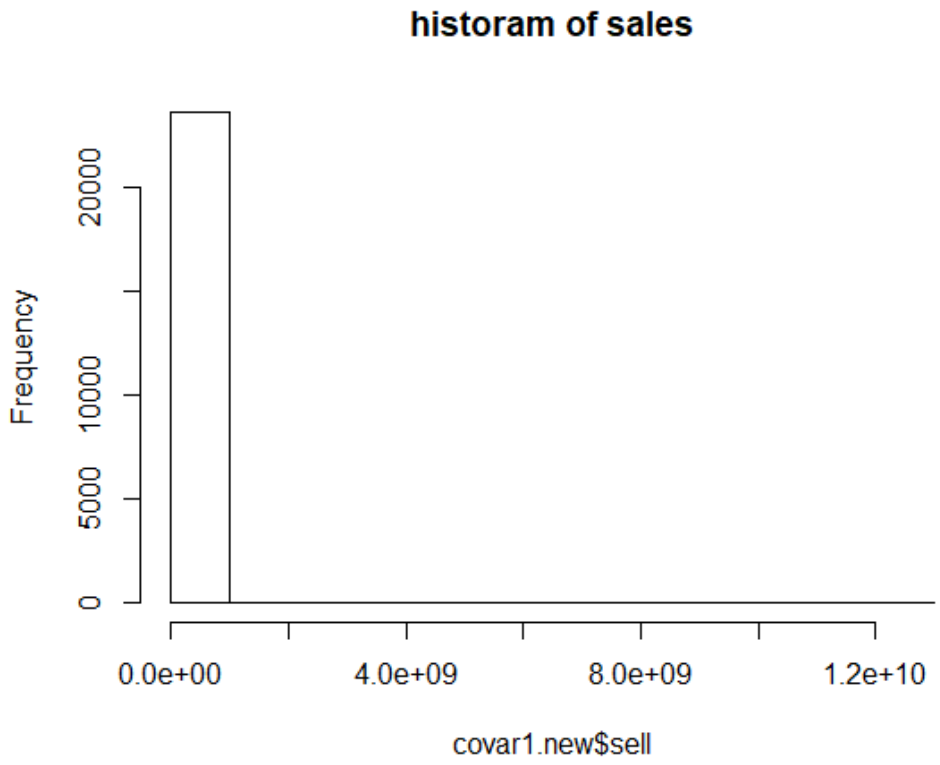

Figure 2.2: 매출 히스토그램

램이 왼쪽으로 지나치게 치우친 것을 볼 수 있다. 따라서 로그변환(log transformation)을 통해서 데이터를 정제해주었다. 결과는 아래 Figure 2.3과 같다.

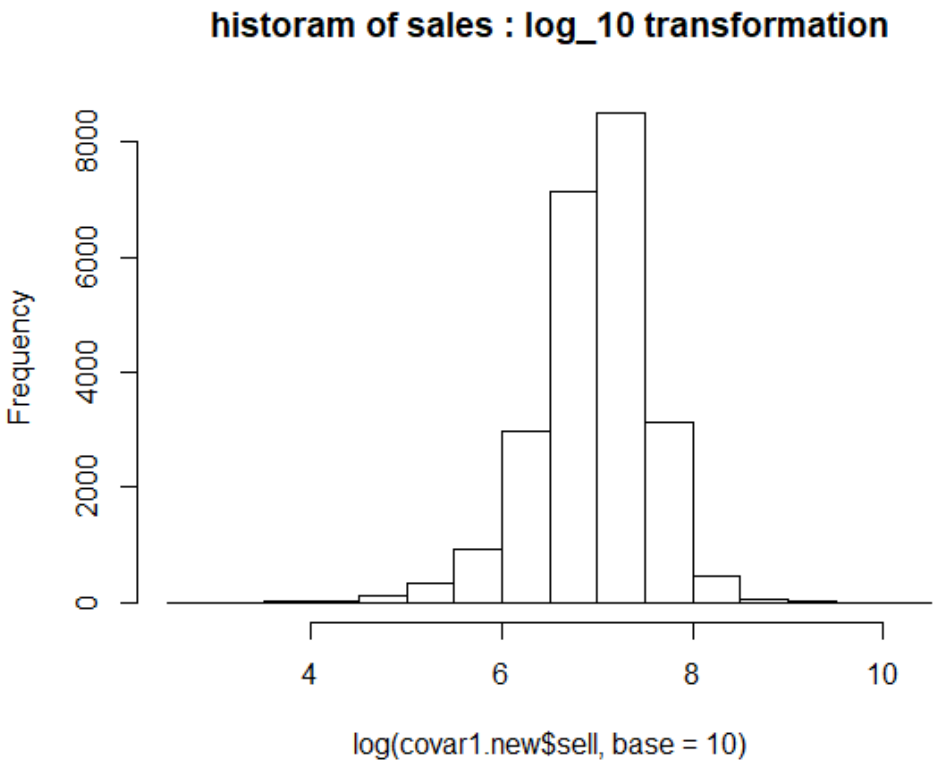

Figure 2.3: 매출 히스토그램 : 로그 변환

이와 같이 한쪽으로 치우친 데이터들에 대해서 모두 로그변환을 해주었다. 또한 측정이 되어 있지 않은 데이터가 있었다. 상주인구, 가구수, 수입, 소비데이터에서 총 451개의 결측값(missing value)이 존재해 이를 빼고 분석을 진행하였다. 이에 해당하는 데이터는 총 23112개의 관측치이고 골목의 수는 총 985개이다.

Chapter 3

방법론(Method)

공간 다층모형(multilevel model)을 활용해 골목상권을 분석하기 위해서 첫 번째로 기본적인 다층모형에 대해서 알아본다. 그리고 다음으로 많이 쓰이는 공간 모형인 SAR(Simultaneous Autoregressive) 모형을 다층 구조로 확장시킨 HSAR(Hierarchical Simultaneous Autoregressive) 모형에 대해서 알아보도록 한다.

3.1. 다층모형(Multilevel Model)

다층모형은 회귀 모형의 일반화된 모형으로 수준(level)이 여러개인 모형을 말한다. 예를 들어, 여러 구에 있는 학교, 그리고 학교 안에 여러 학급에 존재하는 학생들에 대한 모형은 다층모형으로 표현할 수 있다. 다양한 구가 가장 높은 수준이고 그 다음으로 학교, 학교 내의 학급들, 학급 내의 학생들 순서대로 수준을 나눌 수 있다. 이와같이 상위, 하위 수준을 가지고 있는 데이터의 경우에 다층모형을 적합시키면 보다 정확한 분석을 할 수 있다. 수준별로 공변량을 고려할 수 있고 같은 수준 내에 있는 데이터들의 상관관계를 반영해

줄 수 있기 때문이다.

수준이 두개인 다층모형을 생각해보자. 총 N개의 데이터가 존재하고 J개의 상위 수준이 존재한다고 하자. 각 수준에는 n_j개의 하위 수준들이 존재한다. 이를 식으로 나타내면 다음과 같다.

$$\begin{gathered}
y_{ij} = \beta_{0j} + x_{ij}^T \boldsymbol{\beta}_j + \epsilon_{ij} \text{ (하위 수준)} \\
\beta_{kj} = \gamma_{k0} + x_j^T \boldsymbol{\gamma}_k + u_{kj} \text{ (상위 수준)} \\
var(\epsilon_{ij}) = \sigma_\epsilon^2 \ ; \ var(u_{kj}) = \sigma_{uk}^2 \\
cov(\epsilon_{ij}, u_{kj}) = 0 \ ; \ cov(\epsilon_{ij}, \epsilon_{i'j'}) = 0 \ if \ j \neq j' \\
i = 1, \ldots, n_j, \ j = 1, \ldots, J, \ k = 0, 1, \ldots, K \\
n_1 + \cdots + n_J = N
\end{gathered} \tag{3.1}$$

식 (3.1)에서 i는 하위 수준의 인덱스, j는 상위 수준의 인덱스이다. x_{ij}^T는 하위 수준의 공변량(covariate)이고 x_j^T는 상위 수준의 공변량이다. 따라서 $\boldsymbol{\beta}$와 $\boldsymbol{\gamma}$는 각각 하위, 상위 수준의 계수(coefficient)이다. 또한 u_{kj}는 상위 수준의 오차(error), ϵ_{ij}는 하위 수준의 오차이다.

다층모형의 경우 상위수준을 어떻게 모델링하느냐에 따라서 임의효과(random effect), 고정효과(fixed effect)로 나누어질 수 있다. 따라서 다층모형을 임의효과 모형(random effect model)으로 정의하면 선형혼합모형(linear mixed model)으로 보고 추정할 수 있다. 만약 y가 이산형이거나 가산 자료인 경우 일반화 선형혼합모형(generalized linear mixed model)을 적용할 수 있다.

3.2. SAR 모형 (Simultaneous Autoregressive Model;SAR Model)

SAR 모형은 공간 통계학에서 많이 쓰이는 모형으로 계량경제학이나 지리학 연구 등에 널리 쓰이는 모형이다. 가장 간단한 모형부터 살펴보자.
종속 변수인 y에 공간상의 차이(spatial lag)를 고려한 모형이고 공간시차모형(spatial lag model)이라고 부른다. 식으로 표현하면 다음과 같다.

$$y = \rho W y + \epsilon \tag{3.2}$$

식 (3.2)에서 W는 $n \times n$ 행렬로, 공간 가중치 행렬(spatial weight matrix)이다. 그리고 ρ는 자기회귀 모수(autoregression coefficient)로 데이터로 부터 추정한다. 이 모형은 주어진 위치의 y값이 가까운 위치에 있는 값들과 연관이 있다는 것을 반영한다. W는 항상 행의 합이 1이 되도록 표준화 한다. 따라서 주변 값들의 가중평균을 회귀식에 반영하게 된다. 식 (3.2)를 다음과 같이 나타낼 수도 있다.

$$y = (I - \rho W)^{-1}\epsilon$$

추가적인 설명변수 X가 존재한다면 혼합 공간 자기회귀 모형(mixed spatial autoregressive model)이라 부르고 다음과 같은 식으로 표현할 수 있다.

$$y = X\beta + \rho W y + \epsilon$$

보통의 회귀식에 자기회귀(autoregressive) 항을 넣은 것과 같다. 종속 변수 y에 대해서 정리하면 아래와 같다.

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\epsilon$$

이 모형은 공간상의 자기상관성(spatial autocorrelation)과 다른 설명 변수들의 영향을 함께 포함한다.
SAR 모델의 다른 접근으로 공간 오차 모형(spatial error model)이 있다. 이 모형은 공간상의 차이(spatial lag)가 아니라 오차항의 공간적 자기상관성(spatial autocorrelation)을 반영한다. 식으로 표현하면 다음과 같다.

$$y = X\beta + \epsilon,\ where\ \epsilon = \lambda W\epsilon + u$$

기본 모형은 일반적인 회귀 모형이나 오차항이 공간인 상관성을 띈다. 이 때, u는 서로 독립이고 동일한 분포를 따르는 오차이다.
SAR 모형을 정리하면 다음과 같다.

$$y = X\beta + \rho W_1 y + \epsilon\ where\ \epsilon = \lambda W_2\epsilon + u,\ u \sim N(0, \sigma^2 I) \qquad (3.3)$$

이 때, W_1과 W_2는 $n \times n$ 공간 가중치 행렬(spatial weight matrix)이다. $\rho = 0,\ \lambda = 0$이면 일반 선형회귀(linear regression) 모형이다. $\lambda = 0$이면 혼합 공간 자기회귀 모형(mixed spatial autoregressive model; spatial lag model)이다. 마지막으로 $\rho = 0$이면 공간 오차 모형(spatial error model)이 된다.

3.3. 공간 다층모형(Spatial Multilevel Model)

본 논문에서는 다층모형에 공간적인 구조를 넣은 모형으로 HSAR(Hierarchical SAR) 모형을 소개하도록 하겠다. HSAR 모형은 SAR 모형을 확장시켜서 반응 변수와 상위 수준(higher level)의 잔차(residuals)에 기본적인 다층모형을 적용시켰다. 이 때, 수준(level)이 2개인 다층모형을 기본으로 하였다.
일반적인 다층모형의 경우 상위 수준에 따라서 계수가 달라질 수 있지만 본 논문에서는 상위수준이 다르더라도 계수가 달라지지 않는 모형을 고려하였다.

상위 수준의 변수가 1개일 때 식 (3.4)와 같은 예를 들 수 있다.

$$\begin{gathered} y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij} \text{ (하위 수준)} \\ \beta_{0j} = \gamma_0 + u_{0j} \; and \; \beta_{1j} = \gamma_1 + u_{1j} \text{ (상위 수준)} \\ where \; \epsilon_{ij} \sim N(0, \sigma_e^2) \; and \; u_{kj} \sim N(0, \sigma_{uk}^2) \; (k = 0, 1) \end{gathered} \tag{3.4}$$

두 식을 합치면 아래와 같다.

$$\begin{gathered} y_{ij} = \gamma_0 + \gamma_1 X_{ij} + u_{0j} + u_{1j} X_{ij} + \epsilon_{ij} \\ where \; \epsilon_{ij} \sim N(0, \sigma_e^2) \; and \; u_{kj} \sim N(0, \sigma_{uk}^2) \; (k = 0, 1) \end{gathered}$$

항 $u_{0j} + u_{1j} X_{ij}$이 존재하여 상위 수준(j)마다 달라지는 임의 효과(random effect)가 존재하는 모형임을 알 수 있다. 하지만 본 논문에서는 이 항을 축소하여 u_{0j}만 존재한다고 가정한다. 즉, 식 (3.4)에서 $\beta_{1j} = \gamma_1 + u_{1j}$가 아니라 $\beta_{1j} = \gamma_1$으로 놓는다.

데이터에 총 J개의 상위 수준이 존재하고 각 상위 수준 $j,\ j = 1, \ldots, J$에 대해 n_j개의 하위 수준이 존재한다고 하자. 그리고 총 N개의 자료가 있다고 하자. 그러면 아래와 같은 다층모형으로 다시 쓸 수 있다.

$$\begin{gathered} y_{ij} = \beta_0 + x_{ij}^T \boldsymbol{\beta} + x_j^T \boldsymbol{\gamma} + u_j + \epsilon_{ij} \\ var(\epsilon_{ij}) = \sigma_\epsilon^2 \; ; \; var(u_j) = \sigma_u^2 \\ cov(\epsilon_{ij}, u_j) = 0 \; ; \; cov(\epsilon_{ij}, \epsilon_{i'j'}) = 0 \; if \; j \neq j' \\ i = 1, \ldots, n_j, \; j = 1, \ldots, J \\ n_1 + \cdots + n_J = N \end{gathered} \tag{3.5}$$

이제 식 (3.5)의 다층모형에 공간 구조를 반영한 HSAR 모형을 알아보자. HSAR 모형의 식을 살펴보면 다음과 같다.

$$
\begin{aligned}
y &= \rho W y + X\beta + Z\gamma + \triangle\theta + \epsilon \\
\theta &= \lambda M\theta + u \\
\triangle &= \begin{bmatrix} l_1 & 0 & \dots & 0 \\ 0 & l_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & l_J \end{bmatrix} \\
where\ l_j &= \left.\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right\} n_j \ (\ j = 1, \dots, J\ )
\end{aligned}
\tag{3.6}
$$

식 (3.6)에서 y는 반응 변수로 크기가 N인 벡터(vector)이다. X는 $N \times K$ 행렬이고, 하위 수준(lower level)의 공변량(covariates)이다. β는 크기가 K인 회귀계수(regression coefficients)이다. Z는 $N \times P$ 행렬로 상위 수준(higher level)의 공변량이다. γ는 크기가 P인 회귀 계수이다.

W는 하위 수준의 공간 가중치 행렬(spatial weight matrix)로 크기는 $N \times N$이고 행의 합이 1이 되도록 표준화된 행렬이다. ρ는 하위 수준에서의 공간 자기회귀 모수(spatial autoregressive parameter)이다. 이 모수는 해당 수준에서의 공간상의 영향을 나타낸다. 따라서 하위 수준의 공간정보가 있는 경우 이 항을 통해서 공간 구조를 고려할 수 있다. 단, 2장에서 설명한 골목상권 데이터의 경우 하위수준에서의 위치 정보가 존재하지 않는다. 따라서 실제 4장에서 데이터 분석을 진행할 때 $\rho W y$ 항이 없는 모형을 적용할 것이다.

θ는 상위 수준(higher level)의 오차로서 임의효과(random effect)를 나타낸다. W와 비슷하게 M은 표준화된 공간 가중치 행렬인데 상위 수준을 대상으로 한다. 크기는 $J \times J$이다. λ는 상위 수준의 공간 자기회귀 모수이다.

$\triangle$은 $N \times J$ 블록 행렬(block matrix)이고 각 $j = 1, \dots, J$에 대해 l_j는 j번째

상위 수준에 속한 하위 수준의 개수(n_j)를 크기로 하는 1 벡터이다.
ϵ과 u는 오차이다. 각각 분포는 $\epsilon \sim N(0, \sigma_e^2)$, $u \sim N(0, \sigma_u^2)$로 가정한다. ϵ과 u는 독립(independent)이라고 가정하자.
θ의 공분산 행렬(covariance matrix)는 $cov(\theta) = \sigma_u^2(B'B)^{-1}$으로 $B = I_J - \lambda M$이다. 따라서 $\theta \sim N(0, \sigma_u^2(B'B)^{-1})$이다.
식 (3.6)으로 부터 y의 조건부 기댓값(conditional expectation)을 구할 수 있다.

$$E(y|X\beta, Z\gamma) = (I_N - \rho W)^{-1}(X\beta + Z\gamma) \tag{3.7}$$

식 (3.7)로 부터 한 위치에서 공변량 값들의 변화가 해당 위치에서의 결과뿐만 아니라 다룬 위치들의 결과에도 영향을 준다는 것을 알 수 있다. 그 이유는 $(I_N - \rho W)^{-1}$가 앞에 곱해져 있기 때문이다.
식 (3.6)에서 소개한 모형은 두 개의 수준이 있는 SAR 모형으로 두 수준에서의 공간상의 효과를 추정해야 한다. 따라서 모수들에 대해 추가로 사전분포(prior distribution)을 가정한 베이지안 모형을 생각한다. 그리고 마코프체인 몬테카를로(Markov Chain Monte Carlo;MCMC) 방법을 이용해 모수를 추론한다. 이를 위해서 각 모수에 대한 사전분포(prior distibution)의 정의가 필요하다.

3.3.1. 사전분포(prior distribution)

HSAR 모형은 MCMC 방법을 통해서 추정할 수 있는데, 이 MCMC 방법은 각 모수의 사후분포(posterior distribution)로부터 연속적으로 표본(sample)을 추출한다. 모수 η에 대한 사후분포는 식 (3.8) 같이 표현할 수 있다. 이 때 $P(\eta)$는 사전분포(prior distribution)이다.

$$P(\eta|\text{Data}) \propto P(\text{Data}|\eta) \times P(\eta) \tag{3.8}$$

표현의 편리함을 위해서 $\boldsymbol{\beta} = [\beta\ \gamma]$, $\boldsymbol{X} = [X\ Z]$라고 놓자. 3.3장에서 설명한 식 (3.6)의 모수 $\eta = \{\rho,\ \lambda,\ \boldsymbol{\beta},\ \sigma_e^2,\ \sigma_u^2\}$이고 Data=$\{y,\ \boldsymbol{X},\ W,\ M\}$이다. 사전

분포는 사전에 알고 있는 정보와 모수에 대한 불확실성을 반영한다. 따라서 식 (3.8)에 의하면 데이터를 통해 사전 정보를 업데이트한 후의 η에 대한 정보가 사후분포에 반영된다(LeSage & Pace. 2009).

본 논문에서는 $\boldsymbol{\beta}$의 사전분포를 다변량 정규분포(multivariate normal distribution)로 사용하였다. 이 때, 평균은 $M_b = (0, \ldots, 0)$이고 분산은 $V_b = 100 \times I$이다(I는 단위행렬(identity matrix)). 따라서 $P(\boldsymbol{\beta}) \sim N(M_b, V_b)$이다.

그리고 ρ와 λ에 대해서는 균등분포(uniform distribution)을 가정하였다. $P(\rho) \sim U[1/\nu_{min}, 1/\nu_{max}]$이고 ν_{min}과 ν_{max}는 각각 하위 수준 가중치 행렬 W의 고유값(eigenvalue)의 최솟값, 최댓값이다. 그런데 W는 행별로 표준화되었기 때문에 $\nu_{max} = 1$이다(LeSage & Pace. 2009). 따라서 $P(\rho) \sim U[1/\nu_{min}, 1]$이다. 마찬가지로 $P(\lambda) \sim U[1/\nu^*_{min}, 1]$이고 ν^*_{min}는 상위 수준 가중치 행렬 M의 최솟값이다.

마지막으로 $P(\sigma_e^2)$,$P(\sigma_u^2)$는 역감마 분포(inverse gamma distribution)로 가정하였고 구체적으로는 $P(\sigma_e^2) \sim IG(0.01, 0.01)$과 $P(\sigma_u^2) \sim IG(0.01, 0.01)$로 가정하였다. 이때 $IG(\alpha, \beta)$ 분포에서 α와 β는 각각 형상모수(shape parameter)와 척도모수(scale parameter)이다. 이 분포의 구체적인 식은 다음과 같다.

$$p(x) \propto x^{-\alpha-1} exp(-\beta/x)$$

Chapter 4

데이터 분석(Data Analysis)

서울시의 골목상권 분석을 위해 앞서 3장에서 설명한 모형을 적용해 보자. 우선, 총 985개의 골목(상위 수준;higher level)이 존재하고 23112개의 골목상권(하위 수준;lower level)이 존재한다. 주어진 데이터 상의 위치 정보는 골목 단위의 위도, 경도 정보가 있다. 3.3장에서 소개된 HSAR 모형은 상위 수준의 위치 정보, 하위 수준의 위치 정보 모두 존재하는 경우였다. 따라서 이 데이터의 경우 식 (3.6)을 적용시킬 때 하위 수준의 위치정보가 없다는 것을 고려해야 한다.

2장에서 언급한 바와 같이 변수들 중 한 쪽으로 치우친 것들은 로그변환을 해주었다. 공변량으로 쓰일 변수들은 아래와 같다.

상위 수준(higer level) : 골목

- 유동인구 수(LogPFlow)
- 직장인구 수(LogPWork)
- 상주인구 수(LogPStay)
- 월 평균 소득(LogIncome)

- 주변 버스정류장, 지하철역 수(LogNTrans수

하위 수준(lower level) : 업종

- 평균 영업 개월 수(LogBMon)

- 매출 건수(LogNSale)

- 주말 매출 비율(WeRatio)

- 골목 내 해당 업종 점포 수(LogNStore)

식 (3.6)에 서울시 골목상권 데이터를 적용시키면 다음과 같다.

$$y = X\beta + Z\gamma + \triangle\theta + \epsilon,\ \theta = \lambda M\theta + u$$

$$\triangle = \begin{bmatrix} l_1 & 0 & \dots & 0 \\ 0 & l_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & l_{985} \end{bmatrix} \ where\ l_j = \left\{ \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\} n_j\ (\ j = 1, \dots, 985\ )$$

반응 변수 y, 공변량 X와 Z, 블록 행렬 $\triangle$은 모두 데이터 상 주어지고 회귀계수 β와 γ, 자기 회귀 모수 λ를 추정해야 한다. 이 때 공간 가중치 행렬인 M을 정의해야 할 필요가 있다. 공간 가중치 행렬은 한 위치에서 거리가 가까운 값에 가중치를 높게 주고 멀면 가중치를 낮게 준다. 이러한 특성을 반영하여 다음과 같이 공간 가중치 행렬을 정의한다.

$$M_{ij} = \begin{cases} exp\left(-10 \times (d_{ij}^2)/(d^2)\right) & if\ d_{ij} \le d \\ 0 & ,otherwise \end{cases} \tag{4.1}$$

식 (4.1)은 가중치 행렬 M의 i행 j열의 원소를 나타낸 것이다. d_{ij}는 i골목과 j골목 사이의 거리를 나타내고 d는 가중치를 부여할 최소의 한계점(threshold)

를 나타낸다. 한계점은 약 5km로 두었다. 골목간의 거리에 따라 W_{ij}를 계산한 후 각 행에 대해서 표준화를 해주었다.
MCMC 방법으로 추정할 때 총 3개의 체인(chain)을 사용하였고 각 체인별로 10,000번의 반복을 진행하였다. 또한 모수를 추정할 때 마지막 5,000개의 표본만을 사용하였다.
초기값은 3가지 체인에 대해 다르게 주었다. 첫 번째 체인의 경우 β는 일반 선형 회귀 모형의 계수를 사용하였다. 그리고 λ는 0.5, ϵ과 u의 분산인 σ_e^2과 σ_u^2은 모두 2로 주었다. 두 번째 체인은 λ, σ_e^2과 σ_u^2을 모두 1로 주고 β는 모두 0으로 주었다. 세 번째 체인은 λ를 2, σ_e^2과 σ_u^2을 0.5로 주고 β는 모두 1로 주었다.
모형 적합시에 R 프로그램의 'HSAR' 패키지를 사용하였다.

4.1. 분석 결과(Results)

MCMC 방법으로 추정한 결과는 다음과 같다. 이 때 2.5%와 97.5%는 95% credible interval이다. Credible interval이란 사후 표본(posterior samples)의 2.5%와 97.5%분위(quantile)를 말하며 이 구간에 0이 포함되면 해당 변수는 유의하지 않다고 해석할 수 있다.
Table 4.1의 $\hat{R}$ 값은 수렴성을 보여주는 수치로 1에 가까울수록 MCMC 체인이 수렴했다고 볼 수 있다(Gelman & Rubin. 1992).

Table 4.1: HSAR 모형 적합 결과(모델 1)

	Posterior Mean	Posterior Std. Error	2.5%	97.5%	$\hat{R}$
Intercept	10.738	0.160	10.416	11.060	1.00
영업개월 수	0.102	0.007	0.089	0.115	1.00
판매 건수	0.532	0.004	0.524	0.538	1.00
주말 비율	0.012	0.0004	0.010	0.011	1.00
점포 수	0.045	0.011	0.027	0.068	1.00
유동인구 수	0.007	0.013	-0.020	0.030	1.00
직장인구 수	0.172	0.013	0.149	0.199	1.00
상주인구 수	-0.086	0.011	-0.106	-0.066	1.00
수입	0.075	0.011	0.054	0.098	1.00
대중교통 수	-0.007	0.005	-0.016	0.002	1.00
$\boldsymbol{\lambda}$	0.786	0.055	0.567	0.859	1.00
$\boldsymbol{\sigma_e^2}$	0.831	0.008	0.813	0.844	1.00
$\boldsymbol{\sigma_u^2}$	0.014	0.002	0.014	0.024	1.02

유의 수준 5%에서 유의한 변수는 Table 4.1에서 굵게 표시하였다. 유의하지 않은 변수는 유동인구 수(LogPFlow), 골목 주변 버스와 지하철 수(LogNTrans)이다. 버스와 지하철 수는 골목 주변의 접근성이 좋은지에 대한 지표로 유동인구 수와 유사한 성격을 띄고 있는 것으로 보이는데, 두 변수가 유의하지 않다는 의외의 결과가 나왔다.

유의한 변수들에서 상주인구 수(LogPStay)의 추정치만 음수이다. 상주인구가 많을수록 매출이 줄어든다는 결과가 나왔다. 나머지 변수들을 보면 평균 영업 개월수(LogBMon)가 클수록, 판매 건수(LogNSale)가 많을수록, 주말 매

출비율(WeRatio)이 높을수록, 점포 수(LogNStore)가 많을수록, 직장인구 수(LogPWork)가 많을수록 그리고 소득(LogIncome)이 많을수록 매출이 올라간다는 결과가 나왔다. 특히 판매 건수가 가장 큰 추정치를 보인다.

모형의 평가 기준으로 DIC(Deviance Information Criterion)와 로그 우도(log likelihood)값을 살펴보았다. DIC는 베이지안 모형 선택(model selection) 문제에서 많이 쓰이는 기준으로, 낮을수록 좋고 로그 우도값은 클 수록 좋다. 이 모형의 DIC는 62116.17이고 로그 우도 값은 -30672.37이다. $\hat{R}$은 거의 1.00으로 모두 수렴했다고 볼 수 있다.

모든 변수를 사용한 모델 1(Table 4.1)에서 유의하지 않은 변수들을 뺀 후 같은 방법으로 다시 적합을 해보았다. 그 결과는 다음과 같다.

Table 4.2: HSAR 모형 축소 모델(reduced model) 적합 결과(모델 2)

	Posterior Mean	Posterior Std. Error	2.5%	97.5%	$\hat{R}$
Intercept	10.821	0.108	10.566	11.008	1.00
영업개월 수	0.101	0.006	0.089	0.115	1.00
판매 건수	0.532	0.004	0.524	0.539	1.00
주말 비율	0.011	0.0004	0.010	0.011	1.00
점포 수	0.045	0.011	0.023	0.068	1.00
직장인구 수	0.169	0.012	0.143	0.194	1.00
상주인구 수	-0.088	0.010	-0.107	-0.067	1.00
수입	0.076	0.011	0.055	0.097	1.00
$\boldsymbol{\lambda}$	0.787	0.053	0.585	0.859	1.00
$\boldsymbol{\sigma_e^2}$	0.831	0.008	0.815	0.846	1.00
$\boldsymbol{\sigma_u^2}$	0.015	0.002	0.014	0.023	1.01

이 모델의 경우 DIC 값은 62112.73으로 모델 1(Table 4.1)보다 작다. 또한 로그 우도 값이 -30669으로 모델 1(Table 4.1)보다 크다. 추정된 값은 거의 비슷하다. 두 기준으로 보아 모델 1(Table 4.1)보다 모델 2(4.2)가 선호되는 모델이다.

마지막으로 공간 구조를 반영하지 않은 기본적인 다층모형(Multilevel model)을 적합시켜 모델 2(Table 4.2)와 비교해 보았다.

Table 4.3: 기본 다층모형 적합 결과(모델 3)

	Posterior Mean	Posterior Std. Error	2.5%	97.5%	$\hat{R}$
Intercept	10.773	0.109	10.540	10.983	1.00
영업개월 수	0.102	0.007	0.089	0.115	1.00
판매 건수	0.531	0.004	0.524	0.539	1.00
주말 비율	0.011	0.0004	0.010	0.011	1.00
점포 수	0.045	0.011	0.023	0.068	1.00
직장인구 수	0.170	0.012	0.146	0.196	1.00
상주인구 수	-0.086	0.010	-0.106	-0.067	1.00
수입	0.079	0.011	0.057	0.100	1.00
$\boldsymbol{\sigma_e^2}$	0.829	0.009	0.815	0.845	1.00
$\boldsymbol{\sigma_u^2}$	0.019	0.004	0.015	0.024	1.01

추정된 값들은 크게 차이를 보이지 않았으나 DIC 값은 62119.87로 가장 컸다.

DIC가 작을수록 좋으므로 HSAR를 적합시키고 유의한 변수들만 사용한 모델 2(Table 4.2)가 가장 좋은 모델이라고 볼 수 있다.

Chapter 5

결론(Conclusion)

지금까지 공간 다층모형을 통해 서울시의 골목상권을 분석해보았다. 공간 다층모형은 HSAR(Hierarchical Simultaneous Autoregressive) 모형을 활용하는데, 이는 SAR(Simultaneous Autoregressive) 모형에 계층적 구조를 반영한 것이다.

논문의 1장과 2장에서는 연구의 목표와 내용을 설명하고 서울시 골목상권 데이터를 살펴보았다. 골목상권 데이터는 골목 수준과 골목 내의 상권 수준으로 두 가지의 수준으로 나눌 수 있고 각 수준마다 공변량이 존재한다. 3장에서는 분석을 위한 방법론들을 소개하며 구체적인 식과 의미를 알아보았다. 기본적인 다층모형, SAR 모형, 그리고 HSAR 모형에 대해서 살펴보았다. 마지막으로 4장에서는 2장에서 설명한 데이터를 3장의 방법론들을 토대로 실제로 분석해보고 결과를 도출해 보았다. 결과적으로 유의한 변수 7개를 선정하여 모델링(모델 2)을 할 수 있었다. 해당 변수는 하위 수준(lower level)에서 평균 영업개월 수, 판매 건수, 주말 매출 비율, 점포 수가 있고 상위 수준(higher leve)에서 직장인구 수, 상주인구 수, 소득이 있다.

참고문헌

[1] Ballo, J. G. (2013). Determinants of active labor market policy: A spatial and multilevel analysis of ALMP expenditures in 29 OECD countries between 1985 and 2010 (Master's thesis, The University of Bergen).

[2] Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). Hierarchical modeling and analysis for spatial data. Crc Press.

[3] Corrado, L., & Fingleton, B. (2011). Multilevel modelling with spatial effects.

[4] Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.

[5] Dong, G., & Harris, R. (2015). Spatial autoregressive models for geographically hierarchical data structures. Geographical Analysis, 47(2), 173-191.

[6] Dong, G., Harris, R., Jones, K., & Yu, J. (2015). Multilevel modelling with spatial interaction effects with application to an emerging land market in Beijing, China. PloS one, 10(6), e0130761.

[7] Elcheroth, G., Penic, S., Fasel, R., Giudici, F., Glaeser, S., Joye, D., ... &

Spini, D. (2013). Spatially weighted context data and their application to collective war experiences. Sociological Methodology, 43(1), 364-411.

[8] Gelfand, A. E., Banerjee, S., Sirmans, C. F., Tu, Y., & Ong, S. E. (2007). Multilevel modeling using spatial processes: Application to the Singapore housing market. Computational Statistics & Data Analysis, 51(7), 3567-3579.

[9] Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevelhierarchical models (Vol. 1). New York, NY, USA: Cambridge University Press.

[10] Gelman, A., & Rubin, D.B. (1992). "Inference from Iterative Simulation using Multiple Sequences". Statistical Science, 7, 457-511.

[11] LeSage, J. P., & R. K. Pace. (2009). Introduction to Spatial Econometrics. Boca Raton, FL: CRC Press/Taylor & Francis.

[12] Pierewan, A. C., & Tampubolon, G. (2014). Spatial dependence multilevel model of well-being across regions in Europe. Applied Geography, 47, 168-176. ISO 690

[13] Ren, Z., Wang, J., Liao, Y., & Zheng, X. (2013). Using spatial multilevel regression analysis to assess soil type contextual effects on neural tube defects. Stochastic environmental research and risk assessment, 27(7), 1695-1708.

[14] Snijders, T. A. (2011). Multilevel analysis. In International Encyclopedia of Statistical Science (pp. 879-882). Springer Berlin Heidelberg.

[15] Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. Tutorials in Quantitative Methods for Psychology, 8(1), 52-69.

Appendix A

R 코드

```
rm(list=ls())
library(MCMCpack)
library(mvtnorm)
library(spdep)
library(spatialprobit)
library(Matrix)
library(foreign)
library(rgdal)
library(RColorBrewer)
library(HSAR)

#read final data
data<-read.csv("data.csv")
data<-data[,-c(1,3,4,5)]

store<-data.frame(data,c(data$num.sub+data$num.bus))
```

```
colnames(store)<-c(colnames(data),"num.bussub")

# the number of total business for each alley
AB <- as.data.frame(table(store$id))
# total number of alley
total <- dim(AB)[1]
num <- AB[,2]
id <- rep(c(1:total),num)
n <- nrow(store)
#random effect
Delta <- matrix(0,nrow=n,ncol=total)
for(i in 1:total) {
Delta[id==i,i] <- 1
}
rm(i)

Delta <- as(Delta,"dgCMatrix")

#higher level weight matrix M
loc<-data.frame(unique(store$lon),unique(store$lat))
d<-as.matrix(dist(loc))
for(i in 1:total){
tmp<-which(d[,i]>0.1)
tmp2<-which(d[,i]<0.1)
d[tmp,i]<-0
d[tmp2,i]<-exp(-10*(d[tmp2,i] / 0.1)^2)
}
```

```
M<-d
for(i in 1:nrow(as.matrix(M))){
M[i,]<-d[i,]/sum(d[i,])
}
M<-as(M,"dgCMatrix")

## fitting the full model
res.formula.f <- sell ~ busi.month+sell.count+weekend+num.store+pop
    .fl+pop.work+pop.stay+income+num.bussub
betas.f= coef(lm(formula=res.formula.f,data=store))

#chain1
pars.f1=list( rho = 0.5,lambda = 0.5, sigma2e = 2.0, sigma2u = 2.0,
     betas = betas.f )
#chain2
pars.f2=list(rho=1,lambda=1, sigma2e=1,sigma2u=1, betas=rep(0,10))
#chain3
pars.f3=list(rho=2,lambda=2, sigma2e=0.5,sigma2u=0.5, betas=rep
    (1,10))

#fitting for chain1
res.f1 <- hsar(res.formula.f,data=store,M=M,Delta=Delta,
burnin=5000, Nsim=10000, thinning = 1, parameters.start=pars.f1)
summary(res.f1)

#fitting for chain2
res.f2 <- hsar(res.formula.f,data=store,M=M,Delta=Delta,
```

```
burnin=5000, Nsim=10000, thinning = 1, parameters.start=pars.f2)
summary(res.f2)

#fitting for chain3
res.f3 <- hsar(res.formula.f,data=store,M=M,Delta=Delta,
burnin=5000, Nsim=10000, thinning = 1, parameters.start=pars.f3)
summary(res.f3)

##fitting the reduced model
res.formula0 <- sell ~ busi.month+sell.count+weekend+num.store+pop.
    work+pop.stay+income
betas0= coef(lm(formula=res.formula0,data=store))

#chian1
pars01=list( rho = 0.5,lambda = 0.5, sigma2e = 2.0, sigma2u = 2.0,
    betas = betas0 )
#chain2
pars02=list(rho=1,lambda=1, sigma2e=1,sigma2u=1, betas=rep(0,8))
#chain3
pars03=list(rho=2,lambda=2, sigma2e=0.5,sigma2u=0.5, betas=rep(1,8)
    )

#fitting for chain1
res01 <- hsar(res.formula0,data=store,M=M,Delta=Delta,
burnin=5000, Nsim=10000, thinning = 1, parameters.start=pars01)
summary(res01)
```

```
#fitting for chain2
res02 <- hsar(res.formula0,data=store,M=M,Delta=Delta,
burnin=5000, Nsim=10000, thinning = 1, parameters.start=pars02)
summary(res02)

#fitting for chain3
res03 <- hsar(res.formula0,data=store,M=M,Delta=Delta,
burnin=5000, Nsim=10000, thinning = 1, parameters.start=pars03)
summary(res03)

##summary
res<-res.f1 #model that wants to summary
# The regression coefficients
tvalue=as.vector(res$Mbetas/res$SDbetas)
coef <- data.frame(Mbetas=as.vector(res$Mbetas),MSDbetas=as.vector(
    res$SDbetas),t=tvalue)

#when res is full model
row.names(coef) <- c("Intercept", "busi.month","sell.count","
    weekend","num.store","pop.fl","pop.work","pop.stay","income","
    num.bussub")

#when res is reduced model
row.names(coef) <- c("Intercept", "busi.month","sell.count","week",
    "num.store","pop.work","pop.stay","income")

# lambda, higher level spatial autoregressive parameter
```

```
lambda<-data.frame(res$Mlambda,res$SDlambda,t=res$Mlambda/res$
    SDlambda)

# lower level variance sigma2e
sigmae<-data.frame(res$Msigma2e,res$SDsigma2e,t=res$Msigma2e/res$
    SDsigma2e)

# higher level variance sigma2u
sigmau<-data.frame(res$Msigma2u,res$SDsigma2u,t=res$Msigma2u/res$
    SDsigma2u)
```

Abstract

Yujin Shin
The Department of Statistics
Graduate School
Seoul National University

Recently, the number of self-employed people and people who value marked individuality is increasing. Therefore local businesses are receiving attention. The meaning of local businesses is not to the business on the main street but to the business in every alley. The unique atmosphere of these businesses is dominating consumers. In this article, I want to analyze various local businesses in Seoul.

Local businesses can be divided into the level of alley and business in alley, so to analyze them, access to a multilevel model is needed. Also as location information of the alley exists, it is necessary to consider spatial correlations at the alley level. Therefore the multilevel model will be introduced and analysis will proceed using the model. In this case, use Hierarchical SAR model that extends a SAR model with multilevel model. For this model, the Bayesian approach Markov Chain Monte Carlo(MCMC) is used to estimate parameters. After fit of model, finally, let's figure out which factors affect sales at the local businesses.

Keyword :*Spatial data, Multilevel model, Simultaneous Autoregressive model, Hierarchical SAR Model, Local Businesses in Seoul*

Student Number : 2016-20268