Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 4, pp. 1271 ~1278

A Spatial Regression for Hospital Data[1)]

Yong-Seok Choi[2)] · Changwan Kang[3)] · Seung Bae Choi[4)]

Abstract

Recently, a profit analysis in hospital management is considered as an important marketing concept. When spatial variability is presented, we must analyze the hospital data with spatial statistical methods. In this study, we present a regression model using spatial covariance for adjustment. And we compare the nonspatial model with spatial model.

Keywords : 공간자료, 공간회귀모형, 세미베리오그램, 적합도

1. 서론

최근 국민들의 의료수요가 증가함에 따라, 의료수요를 충족시키기 위해 고급의료를 표방하고 나선 대형 기업병원들이 많이 나타나고 있다. 이와 같은 상황은 병원들 간의 서비스 및 가격경쟁을 심화시켜 기존의 병원 환경에 적지 않은 영향을 주게 되었다. 따라서 병원들은 경쟁 속에서의 생존을 위한 병원 마케팅전략을 수립하는 것은 필수적인 것으로 되었고, 이와 관련된 일반기업의 마케팅이론과 기법을 병원경영에 활용하려는 시도가 증가하고 있는 실정이다.

CRM(Customer Relationship Management)이란 고객에 대한 광범위하고 심층적인 지식을 바탕으로 개개인에게 적합한 차별 서비스를 제공함으로써 고객과의 관계를 지속적으로 강화해 나가는 마케팅 혁신 활동이다. 즉, CRM은 고객에 대한 정보를 수집하고 수집된 정보를 효과적으로 활용하여 신규고객 획득, 우수고객 유지, 고객가치 증진, 잠재고객 활성화, 평생 고객화와 같은 일련의 사이클을 통하여 고객을 적극적으로 관리하고 유지하며, 고객의 가치를 극대화시키기 위한 기업마케팅 전략의 일환이라고 할 수 있다.

---

1) 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.
2) 부산광역시 금정구 장전동 산 30, 부산대학교 통계학과 교수
E-mail: yschoi@pusan.ac.kr
3) 부산광역시 부산진구 가야동 산 24 동의대학교 데이터정보학과 부교수
E-mail : cwkang@deu.ac.kr
4) 부산광역시 부산진구 가야동 산 24, 동의대학교 데이터정보학과 조교수
E-mail : csb4851@deu.ac.kr

최근 경영분야에서 CRM의 마케팅 기법들의 급속적인 발전에 따라 의료분야를 비롯한 많은 분야에서 CRM의 중요성을 인식하여 각 분야에 맞는 경영에 CRM을 위한 마케팅을 적극적으로 도입하고 있다. 이와 더불어 각 분야에 고객 마케팅을 위한 CRM에 대한 연구 역시 다양하게 활발히 진행되고 있다(강창완 외 2인, 2005). 특히, 의료분야에서의 CRM에 대한 연구들로서 병원 선택 동기에 관한 연구(박창균, 1985), 병원에 대한 환자들의 만족도 연구(이태섭, 1993) 등이 있다. 최근의 연구들로서 고객 세분화를 위한 최적 RFM 모형 구축에 관한 연구(이소영 외 3인, 2004), 의료분야에서의 수익성을 예측하는 모형을 개발하는 연구(이소영 외 4인, 2005) 등이 있다. 이러한 연구들은 CRM을 위한 마케팅 전략을 수립함으로써 병원의 수익을 극대화하려는데 목적을 같이하고 있다.

이러한 추세에 따라 본 연구에서는 병원 수익성 모형 개발에 있어서 기존 연구에서는 고려하지 않았던 고객의 위치정보를 포함하는 예측모형을 제안하고자 한다. 즉, 병원 수익성에 영향을 주는 요인으로 고객(환자)의 병원이용정보와 공간적 변인인 고객의 주거 지역정보, 즉 위치정보를 포함한 예측모형을 고려하고자 한다. 그리고 본 연구에서 제안한 공간정보를 도입한 예측모형과 공간정보 없이 통상적으로 기존의 방법으로 구축하는 예측모형과 비교하여 공간정보를 이용한 예측모형이 더 좋은 예측결과를 얻을 수 있음을 실제 예를 통하여 보여 준다.

본 연구는 2절에서 공간상관을 고려한 회귀모형에 대하여 공간통계학의 기본 개념과 함께 설명한다. 예측을 위해서 일반적으로 자주 사용되는 회귀모형에 대한 내용은 언급을 하지 않는다. 그리고 3절에서는 그리고 3절에서는 부산 소재 D병원의 외래 자료를 이용하여 본 연구에서 제안하는 연구모형과 일반적인 회귀모형에 의한 예측의 우위성 비교를 위하여 실제 적용사례가 다루어진다. 끝으로 4절 맺음말에서 결론을 맺는다.

2. 공간상관을 고려한 회귀모형

2.1 공간자료와 공간상관

공간자료는 확률과정(stochastic process)의 실현치로서 $\{ y(s): \} s \in D \subset R^d$와 같이 표현된다. 여기서 s는 D의 임의의 위치를 나타내고 R^d는 d차원의 유클리드공간으로서 $d = 1, 2, 3$이다. 통상적으로 공간자료는 $(y_i, s_i), i = 1, \cdots, n$으로 표현하며 정상성(stationarity)을 만족한다고 가정한다. 즉 공간상에서 y_i는 위치 s_i에서 얻어진 관측치를 의미하며 보통 공간적으로 서로 멀리 떨어진 경우에는 낮은 상관을, 가까이 있을수록 높은 상관을 갖게 된다. 공간통계학(spatial statistics)에서 공간자료의 분석에서 중요한 목적 중에 하나는 공간 예측인데 이것을 공간통계학에서는 크리깅(kriging)이라고 한다. 크리깅은 (1) 세미베리오그램(semi-variogram) 추정 단계, (2) 추정된 세미베리오그램을 기초로 하여 세미베리오그램을 적합시키는 단계, 그리고 (3) 적합된 세미베리오그램을 이용하여 미지의 위치를 예측하는 단계의 세 가지 단계를 거쳐 수행된다.

세미베리오그램은 공간통계학에서 자주 이용되는 공간종속성의 척도로서 관측치들 간의 거리의 함수로 표현된다. $y(s)$가 본질적인 정상성(intrinsically stationary)을 만

족한다는 가정 하에 세미베리오그램은 다음과 같이 정의된다.

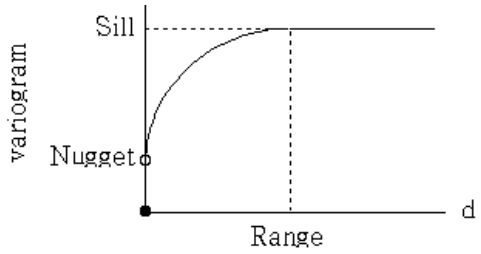
$$\gamma(d) = \frac{1}{2}E(y(s+d)-y(s))^2 \qquad (1)$$
$$= \frac{1}{2}Var(y(s+d)-y(s)) .$$

여기서 $2\gamma(d)$는 베리오그램이라고 하고, $\gamma(d)$는 세미베리오그램이라고 한다. 여기에서는 혼동이 없는 한 $2\gamma(d)$와 $\gamma(d)$을 베리오그램으로 혼용한다.

그리고 자료가 주어진 경우 다음 공식에 의해 표본베리오그램을 계산할 수 있다.

$$\hat{\gamma}(d) = \frac{1}{2N_d}\sum_{N(d)}(y(s_i)-y(s_j))^2 , \qquad (2)$$

단, $N(d)$는 유클리드 거리 $|s_i - s_j| = d$를 갖는 모든 쌍의 집합이고, N_d는 $N(d)$에 속하는 쌍의 수를 의미한다. 일반적으로 베리오그램은 <그림 1>과 같이 nugget, sill, range 등 3개의 모수를 가지는 거리 d에 의존하는 함수식의 형태를 갖는다.



<그림 1> 일반적인 베리오그램 형태

<그림 1>에서 절편항에 해당하는 nugget은 거리 $d=0$일 때 베리오그램의 값을 의미하며 이것은 일종의 측정오차에 해당한다. sill은 거리 d에서 평평하게 되는 곳의 세미베리오그램의 값으로 정의되는데 이는 관측치의 분산에 해당한다. 그리고 베리오그램이 sill에 해당되는 거리 d값을 range라 정의하며 range보다 거리가 작은 경우에 관측치들은 공간적으로 상관되어 있음을 의미한다. 베리오그램의 모형은 그 형태에 따라서 공간통계학에서는 5가지 모형들로 구분한다. 이에 대한 설명은 2.2절에서 소개한다.

2.2 공간자료에서의 회귀모형

공간자료가 주어졌을 때, 예측모형으로 다음과 같은 선형모형으로 표현할 수 있다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \ , \ i = 1, \cdots, n . \qquad (3)$$

일반적으로 오차항들이 서로 독립이고, 동일한 분포 F를 따른다고 가정한다. 그러

나 공간자료인 경우는 오차항들이 공간상관(spatial correlation)을 갖게 되기 때문에 오차항에 대한 다음의 가정들을 고려할 수 있다.

$$Var(e_i) = \sigma^2 + \sigma_I^2, \quad Cov(e_i, e_j) = \sigma^2[f(d_{ij})], \tag{4}$$

여기서 $Cov(e_i, e_j)$는 위치들 s_i와 s_j 사이의 거리 d_{ij}의 함수이다. 거리함수 $f(d_{ij})$는 방향에도 의존할 수 있는데 $f(d_{ij})$는 방향에 의존하지 않는 함수라고 가정한다(isotropic). 거리함수 $f(d_{ij})$는 주어진 형태에 따라서 여러 가지 공분산모형으로 고려해 볼 수 있는데 공간통계학에서 고려하는 대표적 공분산(혹은 베리오그램)모형은 다섯 가지로 다음과 같다.

[모형 1] Spherical 모형 : $f(d_{ij}) = [1 - 1.5(d_{ij}/\rho) + 0.5(d_{ij}/\rho)^3]1(d_{ij} < \rho)$,
[모형 2] Exponential 모형 : $f(d_{ij}) = [\exp(-d_{ij}/\rho)]$,
[모형 3] Linear 모형 : $f(d_{ij}) = [1 - \rho d_{ij}]1(\rho d_{ij} < 2)$,
[모형 4] Guassian 모형 : $f(d_{ij}) = [\exp(-d_{ij}^2/\rho^2)]$,
[모형 5] Power 모형 : $f(d_{ij}) = \rho^{d_{ij}}$.

식 (4)와 [모형 1]에서 [모형 5]까지 5가지 모형들에서 공간통계학에서는 σ_I^2를 nugget, $\sigma^2 + \sigma_I^2$ 부분을 sill, 그리고 ρ를 range라고 부른다. 또한 $Var(e_i) = \sigma^2 + \sigma_I^2$인 경우의 모형을 nugget 효과가 있는 모형이라고 하고, $Var(e_i) = \sigma^2$인 경우를 nugget 효과가 없는 모형이라고 한다. 앞서 설명한 바와 같이 nugget, sill 그리고 range는 베리오그램의 중요 요소이다.

결국 공간 회귀모형은 자료가 주어진 경우 베리오그램의 추정과 회귀모수의 추정을 통해 모형적합이 이루어지게 된다. 여기서 먼저 베리오그램 추정은 베리오그램이 가지고 있는 세 개 모수의 추정문제로 귀결되어질 수 있다. 그러므로 모형적합을 위해서는 첫 번째 단계로, 가장 적절한 공분산(세미베리오그램) 모형을 찾는 일이고 두 번째 단계로는 선정된 세미베리오그램 모수값을 추정하는 일이다. 본 연구에서는 SAS/PROC MIXED 옵션을 이용하여 공분산모수들은 제약최대우도방법(Restricted Maximum Likelihood Method)에 의하여, 그리고 회귀모수들은 혼합모형방정식의 해를 구함으로써 추정치를 구하기로 한다.

3. 의료자료를 이용한 사례분석

3.1 자료

분석 자료는 부산 지역에 위치하는 D의료원의 2002년 1월 1일부터 2003년 12월 31일까지 2년간 퇴원환자 19841명 중 무작위추출을 이용하여 얻어진 1000명의 자료이다. 이들 중에서 위치정보인 횡단 메르카토르 좌표계인 TM(Transverse Mercator) 좌표가 생성 가능한 629명을 최종 분석대상으로 선정하여 분석하였다. 참고로 TM좌표는 임의의 지역에 대한 기준 지점을 좌표 원점으로 정하고 원점을 중심으로 투영한 평면상에서 원점을 지나는 자오선을 X축, 동서방향의 위도선을 Y축으로 하여 각 지점의 위치를 m단위의 평면 직각 좌표계로 표시한 것을 말한다. 우리나라에서는 TM 좌표 또는 평면 직각 좌표계에서의 좌표 기준점으로 서부 원점(125° E, 38° N), 중부 원점(127° E, 38° N), 동부 원점(129° E, 38° N)의 3개 원점을 사용하며 여기서 X축은 북쪽 방향이 양의 값을 나타내고, Y축은 동쪽 방향이 양의 값을 나타난다. 본 연구에서는 환자들의 주소를 이용한 TM 좌표 생성하였으며 분석에 사용된 변수와 자료 형태는 [표 1]에 나타나있다.

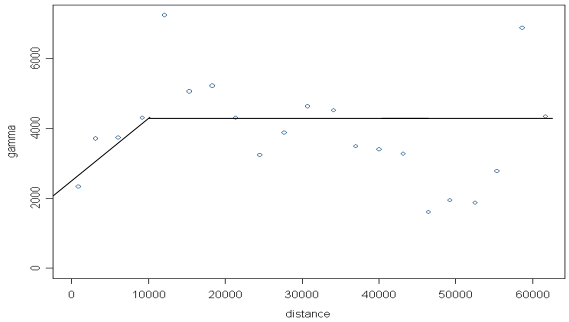
<표 1> 분석변수 및 자료

등록번호	총진료비	입원횟수	TM_x	TM_y
23**	24743148	3	389360.9	189296.6
34**	1826964	1	386677.0	187353.8
48**	3019556	1	388979.5	187676.0
76**	7362658	2	388986.6	187821.1
44**	2520264	2	389000.1	187901.5
.	.	.	.	.
63**	2624668	2	388669.3	187376.6

<표 1>에서 총진료비는 입원환자들의 분석기간 동안의 진료비의 총합을 의미하며 위치정보를 의미하는 TM 좌표 값은 각각 TM_x 와 TM_y 변수로 나타내고 있다.

3.2 베리오그램(variogram) 추정

먼저 공간회귀분석에 앞서 분석에 필요한 세미베리오그램 모형을 추정한 결과가 <그림 2>에 제시되어 있다. <그림 2>를 통하여 적절한 모형으로 exponential 혹은 spherical 모형을 선택할 수 있으며 해당모수의 초기 추정치로 sill=4000, nugget=2000, 그리고 range =10000 선택하였다. 한편 마케팅 관점에서 range 추정값은 흥미 있는 해석이 가능한데(유성모 외 2인, 2006) 즉 모수 range는 일정거리 이상이 되면 수익성(총진료비)에 영향을 주지 않는 거리로 해석이 되며 병원입장에서 보면 병원이용고객의 최대 마케팅경계로 볼 수 있다. 이러한 관점에서 D의료원의 유효마케팅 경계는 D의료원에서 반경 10 km 이내로 볼 수 있다.



<그림 2> 총진료비를 이용한 세미베리오그램

3.3 공간상관을 고려한 회귀분석

<표 1>의 자료를 이용하여 병원 수익성 예측모형으로 총진료비를 종속변수로, 그리고 환자의 입원횟수를 설명변수로 하여 다음의 회귀모형을 고하였다.

$$\text{총진료비} = \alpha + \beta \cdot \text{입원횟수} + e_i . \tag{5}$$

이때 공간상관을 고려한 공간회귀분석 프로그램은 아래와 같다(SAS Inc, 1992).

```
PROC MIXED scoring=50;
        MODEL totcost=freq1/solution;
        REPEATED/subject=intercept local
                        type=sp(sph)(tm_x1 tm_y1);
```

먼저 공간상관이 존재하는지에 대한 검정 즉, $H_o : \rho = 0$은 독립모형(공간상관이 없는 모형)과 공간상관을 고려한 모형(exponential 모형인 경우)과의 -2 REML의 차이값을 우도비 검정통계량으로 계산할 수 있으며 그 결과가 [표 2]에 나타나 있다. <표 2>를 보면 유의수준 1%에서 귀무가설 H_o를 기각할 수 있으며 이는 분석 자료에 공간상관이 존재한다는 것을 의미한다. 또한 적절한 공간 공분산모형의 선택을 위한 적합도 통계량을 살펴보면 exponential 모형이 spherical 모형보다 그 값이 작아 다소 더 좋은 모형으로 나타나 최종적으로 공간회귀모형으로 exponential 모형을 이용한 회귀 모형을 적합시켰다. 따라서 <표 2>에 의한 결과에서 기존의 회귀모형(독립모형) 보다 공간정보를 이용한 공간회귀모형이 예측의 측면에서 보다 더 우수한 모형임을 알 수 있다.

<표 2> 적합도를 이용한 모형 비교

적합도 통계량	독립 모형 (no spatial)	exponential 모형	spherical 모형
-2 REML log likelihood	9861.4	9790.9	9842.0
AIC	9863.4	9796.9	9848.0
BIC	9867.9	9810.4	9861.3
Null model LR Test : H_o: $\rho = 0$ χ^2=70.5 , p-value = 0.0001			

결론적으로, exponential 공간상관을 고려여 얻어진 최종 회귀추정식은 다음과 같다.

총진료비 = 69.72 + 218.7×입원횟수

4. 맺음말

본 연구에서는 병원마케팅에서 활용할 수 있는 수익성 예측모형으로 공간상관 구조를 고려한 회귀모형을 제시하였다. 특히, 서로 다른 위치에서 관찰된 자료들이 공간적 변인인 거리에 의해서 서로 상관되었을 때 이들 공간변수에 대한 예측 모형을 병원마케팅 관점에서 살펴보았다. 즉, 베리오그램 모형 추정을 통해 수익성에 대한 유효마케팅 거리를 추정할 수 있었고, 공간상관을 고려한 총 진료비에 대한 예측식을 구하였다.

그러나 본 연구는 공간상관을 고려한 가장 단순한 예측모형을 제시함으로써 실제 병원 수익성에 영향을 줄 수 있는 입원기간과 같은 많은 요인들을 고려하지 못 하였다. 실질적 병원마케팅을 위해서는 좀 더 많은 설명요인을 고려한 예측모형을 고려해야 할 것이다.

참고문헌

1. 강창완, 최용석, 임승범(2005). 프론티어 모델을 이용한 수익성분석, *Journal of the Korean Data Analysis Society,* Vol. 7, No. 5, 1695-1703.
2. 박창균(1985). 병원마케팅 전략수립을 위한 환자들의 병원선택요인에 관한 연구. 연세대학교 석사논문.
3. 유성모, 윤연상, 김기환(2006). 대형 할인점 매출 데이터를 이용한 Semi-Variogram의 추정과 거리에 의한 할인점 이용권 지도 작성에 관한 연구, *Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society*, April 28-29, 99-108.

4. 이소영, 최승배, 김규곤, 강창완(2004). 고객세분화를 위한 최적 RFM 모형 구축에 관한 연구, *Journal of the Korean Data Analysis Society,* Vol. 6, No. 6, 1829-1840.
5. 이소영, 최승배, 김규곤, 김형도, 강창완(2005). 고객세분화를 위한 최적 RFM 모형 구축에 관한 연구, *Journal of the Korean Data Analysis Society*, Vol. 6, No. 6, 1829-1840.
6. 이태섭(1993). K대학병원 서비스에 대한 소비자 만족도, 계명의대논문집.
7. Cressie, N.A.C.(1991). *Statistics for Spatial Data*, New York: John Wiley & Sons, Inc.
8. SAS Institute(1992). *SAS Technical Report P-229 SAS/STAT Software : Changes and Enhancements.* Cary, NC.

[2006년 10월 접수, 2006년 11월 채택]