Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 28 - 29, 2006, pp. 99-108

대형 할인점 매출 데이터를 이용한 Semi-Variogram의 추정과 거리에 의한 할인점 이용권 지도 작성에 관한 연구

유성모[1] · 윤연상[2] · 김기환[3]

요 약

대형 할인점 매출 데이터는 G-CRM, 에어리어 마케팅(Area Marketing)에 활용하기 위해 고객의 구매정보와 위치정보를 포함한다. TM중부좌표로 이루어진 고객 위치정보를 이용하여 지점간의 거리를 구할 수 있다. 서로 다른 위치에서 동시에 측정된 자료들이 공간적인 변인에 의하여 영향을 받는다면, 공간적인 변인의 함수식에 의한 예측모형을 설정하는 것이 타당하다. 본 연구에서는 공간적인 변인으로 거리가 주어졌을 때, 대형 할인점 매출 자료에 대한 세미베리오그램(Semi-Variogram)의 모형을 추정하고, 관측되지 않은 지역에 대한 할인점 이용권을 공간예측기법으로 예측하였다. 그리고 공간예측 기법을 통해 예측된 할인점 이용권을 토대로 할인점 이용권 지도를 작성하였다. 또한 매출 데이터의 공간이상치 탐지를 위한 방법을 제시하고 실례로 알아보았다.

주요용어 : Semi-Variogram, 공간예측, 할인점 이용권 지도, 공간이상치, Piecewise linear Model.

1. 서론

상권이란 단일 소매시설 혹은 시설 집단(쇼핑센터)이 고객들을 끌어들이는 지역으로 정의된다(Truman Asa Hartshon, 1992). 최근 유통망의 발달로 대형할인점이 많이 생겨나면서 더욱 상권에 대한 관심이 높아지고 있다. 상권의 유사한 용어로 판매권, 시장지역, 중심지의 보완지역, 배후지역, 시장영역 등이 있다. 이 연구에서는 이 상권을 대형 할인점을 이용하는 고객의 이용권으로 본다.

대형 할인점의 이용권 분포는 할인점이 입지되어 있는 지역의 특성에 따라 다르게 나타난다. 지역의 소매업 매출액, 인구수, 지역 소비정도, 할인점까지의 접근성(거리, 접근 소요시간, 접근 수단), 할인점의 규모, 지역 소비행태 등 여러 요인에 의해 영향을 받게 된다. 하지만 사전 연구사례를 보면 이런 요인 중 접근성 요인이 가장 비중을 크게 차지하는 것으로 나타나고 있다(임명숙, 2004). 따라서 본 연구에서 할인점 이용권에 영향을 주는 요인으로 공간적 변인인 고객과 지점간의 거리

[1]330-841 충청남도 천안시 목천읍 지산리 107, 국제평화대학원대학교 교수. E-mail : syoo@peace.ac.kr
[2]339-700 충청남도 연기군 조치원읍 서창동 208, 고려대학교 정보통계학과 대학원 석사과정.
E-mail : 9887045@korea.ac.kr
[3]339-700 충청남도 연기군 조치원읍 서창동 208, 고려대학교 정보통계학과 교수.
E-mail : korpen@koera.ac.kr

만를 고려해 보고자 한다. 서로 다른 위치에서 측정된 자료들이 실제 접근 거리에 의하여 영향을 받는다면, 공간적인 변인의 함수식에 의한 공간예측모형을 설정하는 것이 타당하다. 본 연구에서는 공간적인 변인으로 거리가 주어졌을 때, 대형 할인점 매출 자료를 이용하여 세미베리오그램(Semi-Variogram)를 추정하고, 여러 세미베리오그램 모형 중 David와 유성모(1993)에 의해 소개된 Piecewise Linear Model을 세미베리오그램(Semi-Variogram)모형으로 한 공간예측 기법으로 적용하고, 이를 통해 예측된 할인점 이용권으로 할인점 이용권 지도를 작성하였다. 또한 매출 데이터의 공간 이상치 탐지를 위한 방법으로 유성모와 엄익현(1997)이 제시한 분포론적 접근방법(Distributional Approach)를 다음 사례에 적용해 보도록 한다.

2. 공간변수와 세미베리오그램(Semi-Variogram)

2.1. 공간변수의 정의

공간적 변인인 거리에 의하여 영향을 받는 공간변수들은 공간상으로 가까우면 가까울수록 높은 상관을 보이고 공간상으로 멀어지면 멀어질수록 낮은 상관을 보이게 될 것이다.

따라서, 본 논문에서 다루고 있는 공간변수들 사이의 상관구조는 거리만의 함수식으로 표현된다. 본 논문에서 사용하고 있는 공간변수 $\{Z(s): s\in D\}$는 가우지안(Gaussian) 과정이다. 여기서, D는 d차원 공간 R^d상의 부분집합이다. 또한 공간변수 $Z(\cdot)$는 다음과 같은 가정을 만족한다고 하자.

$$
\begin{aligned}
&① \ E(Z(s)) = \mu, \quad \text{for } all\ s\in D.\\
&② \ Var(Z(s)) \equiv C(0) = \sigma^2 < \infty, \quad \text{for } all\ s\in D.\\
&③ \ Cov(Z(s_i), Z(s_j)) \equiv C(s_i - s_j) < \infty, \quad \text{for } all\ s_i,\ s_j \in D.\\
&④ \ \text{공간변수 } Z(\cdot) \text{는 정규분포를 따른다.}
\end{aligned}
\tag{1}
$$

식(1)의 ③에서 , $s_i - s_j$는 두 지점 s_i와 s_j 사이의 유클리드 거리라 표현한다면, 공간변수들 사이의 상관관계를 나타내는 공분산은 거리만의 함수인 $C(s_i - s_j)$ 형태로 표현된다. $C(s_i - s_j)$의 특별한 경우인 $s_i - s_j = 0$일 때의 $C(0)$는 공간변수의 분산 σ^2이 된다. 식(1)의 가정들에 의하여 두 공간변수의 차 $Z(s_i) - Z(s_j)$는 평균이 0이고 분산이 $2C(0) - 2C(s_i - s_j) \equiv 2\gamma(s_i - s_j)$인 정규분포를 따른다. 여기서 $C(s_i - s_j)$와 $2\gamma(s_i - s_j)$를 각각 코베리오그램(Covariogram)과 베리오그램(Variogram)이라고 부르며, $\gamma(s_i - s_j)$를 세미베리오그램(Semi-Variogram)이라고 부른다.

h를 두 지점 사이의 거리의 차이를 나타내는 값이라고 할때 $s_i = s_j$, $s_j = s + h$라고 놓을 수 있다. 이 때 코베리오그램 $C(\cdot)$와 세미베리오그램 $\gamma(\cdot)$ 사이에는 식(2)와 같은 관계식이 성립한다.

$$\gamma(h) = \frac{1}{2} Var(Z(s) - Z(s+h)$$
$$= \frac{1}{2} Var(Z(s)) + \frac{1}{2} Var(Z(s+h)) - Cov(Z(s), Z(s+h)$$
$$= C(0) - C(h) \quad (2)$$

이제, 세미베리오그램이 가지는 주요 모수들을 정의하고 그것이 가지는 의미를 살펴보기로 하자. 일반적으로 세미베리오그램은 *Still* 과 *Nugget*, *Range* 등 3개의 모수를 가지는 거리 h의 함수식의 형태를 가진다.

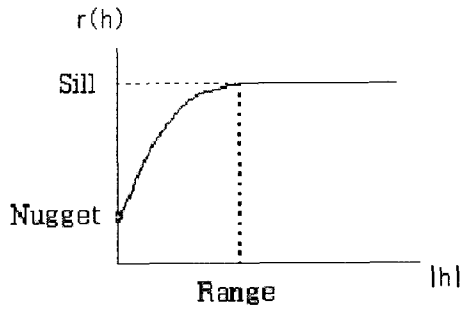


그림 1 : 세미베리오그램의 일반적인 형태와 *Still, Nugget effect* 그리고 *Range*의 관계

2.2. 세미베리오그램의 추정

기존에 제시되어온 모수적인 세미베리오그램 모형에는 *Spherical*모형, Exponential 모형, Rational Quadratic모형, Wave(또는 Hole-effect)모형, Power모형, 그리고 David와 유성모(1993)에 의해서 소개된 Piecewise Linear 모형 등이 있다.

본 연구에서는 세미베리오그램 모형으로 모형이 간단하면서 자료의 공간적 속성을 잘 설명할 수 있는 Piecewise Linear 모형을 사용하였다.

일반적인 Piecewise Linear 모형의 함수식은 식(3)와 같으며, <그림 2>는 Piecewise Linear 모형 형태를 보여주고 있다.

$$\gamma(h;\theta) = \begin{pmatrix} 0, & h=0 \\ C_0 + \frac{C_s}{a_s}\|h\|, & 0<\|h\|\le a_s \\ C_0 + C_s, & a_s<\|h\|. \end{pmatrix} \quad (3)$$

여기서, 식(3)의 Piecewise Linear 모형은 $\theta=(C_0, C_s, a_s)'$, $C_0\ge 0$, $C_s\ge 0$, $a_s\ge 0$ 를 만족한다.

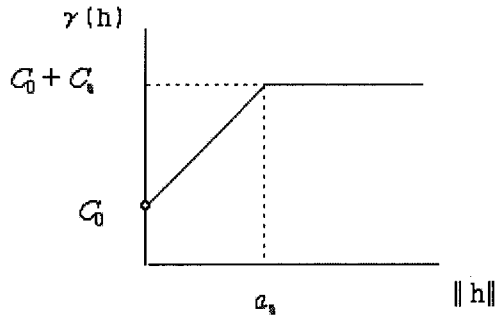


그림 2 : Piecewise Linear 모형의 형태

여기서, C_0는 *Nugget*을, C_0+C_s는 *Still*을, a_s는 *Range*를 나타내며, 특히 C_s를 partial *Still*이라 부른다.

세미베리오그램의 추정은 세미베리오그램 모형이 가지고 있는 모수의 추정문제로 귀결된다. 본 연구에서 고려하고 있는 Piecewise Linear 모형의 경우, 식(3)에서 볼 수 있듯이 세 개의 모수인 C_0 , C_s 그리고 a_s를 추정하는 문제가 된다. 이들 세 개의 모수를 추정하는 방법으로 표본을 이용하여 구한 표본 세미베리오그램과 이론적인 세미베리오그램의 차의 제곱합을 최소화시키는 최소제곱법(Ordinary Least Square estimation)을 채택하였다.

즉, 목적함수

$$Q = \sum_{j=1}^{k}\left\{ \gamma_0(h_j) - C_0 - C_s\frac{h_j}{a_s} \right\}^2 + \sum_{j=k+1}^{N}\{\gamma_0(h_j) - C_s - C_0\}^2 \qquad (4)$$

를 최소화시키는 C_0와 C_s 그리고 a_s를 구한다. 본 연구에서 444개의 지점자료를 이용하여 식(4)를 최소화 시키는 C_0, C_s, a_s를 구하였다.

이 세미베리오그램 모형 적합방법은 시간이 너무 많이 걸린다는 단점이 있으나 원자료를 모두 이용한다는 점에서 모형적합의 정확도면에서는 신뢰성있는 방법이다.

2.3 공간이상치 탐지 방법- 분포론적 접근방법(Distributional Approach)

유성모와 엄익현(1998)은 공간자료에 대한 특이치 탐지문제로 분포론적 접근방법(Distributional Approach)을 소개하였다. 이 분포론적 접근방법에 의한 공간이상치 탐지 방법은 다음과 같다.

공간변수 $Z(\cdot)$는 정규분포를 따른다. 따라서 $Z(s_i)-Z(s_j)$는 평균이 0이고 분산이 $2C(0)-2C(s_i-s_j)\equiv 2\gamma(s_i-s_j)$인 정규분포를 따른다. 여기서, $\frac{[Z(s_i)-Z(s_j)]^2}{2\gamma(s_i-s_j)}$은 자유도가 1인 χ^2분포값과 세미베리오그램을 곱한 형태인 $\gamma(s_i-s_j)\times\chi_1^2$의 분포를 가진다.

또한 이론적인 세미베리오그램은 $\gamma(h)=\frac{1}{2}E(Z(s_i)-Z(s_j))^2$이기 때문에 관측된 지점으로부터 얻은 표본 세미베리오그램인 $\frac{1}{2}(Z(s_i)-Z(s_j))^2$를 이용하여 이론적인 세미베리오그램을 추정할 수 있다.

여기에서 소개하고자 하는 공간 특이치 탐지 방법은 관측된 지점으로부터 얻은 표본 세미베리오그램 $\frac{1}{2}(Z(s_i)-Z(s_j))^2$이 $\gamma(s_i-s_j)\times\chi^2_{1;1-\alpha}$보다 큰 경우를 유의수준 α하에서의 탐지하여 그러한 표본 세미베리오그램들에 대하여 가장 영향을 많이 미친 공간지점을 특이치로 판정하는 방법이다.

3. 대형할인마트 매출데이터를 이용한 실증적 사례분석

3.1. 자료의 제시

실증적 분석을 위하여 2005년 5월 4주간 H 대형할인점(영등포점) 카드이용 고객 매출액을 이용해 각 동별 이용 고객수가 10명 이하인 동을 임의로 제외한 444개 행정동의 자료를 산출하였다.

각 동은 행정동으로 구분하였으며, 도로상거리는 네이버 전자지도에서 제공하는 도로상의 최단 운행거리를 사용하였다. 각 동의 TM좌표는 해당 동의 고객 위치 TM좌표의 평균값을 이용하였으며, 매출평균도 해당동의 4주간의 고객 매출평균값을 이용하였다. 매출합계는 4주간 해당 동 고객에게 판 매출 총 합계를 이용하였다.(단위는 만원)

표 1 : 분석에 이용된 H 대형할인점(영등포점)의 서울시 동별 매출데이터 요약

지역 (구/동)			공간적 영향변인	TM좌표		동별 매출액 자료	
ID	gu	dong	도로상거리 (Km)	x (동서방향)	y (남북방향)	매출평균 (단위:만원)	매출합계 (단위:만원)
1.	강남구	개포1동	19.21	205000.35	442027.39	81580	244740
2.	강남구	개포2동	22.23	206297.92	442383.04	150377	451130
3.	강남구	개포3동	22.05	205988.89	442911.66	29220	116880
4.	강남구	개포4동	18.38	204539.72	441364.72	12803	51210
5.	강남구	논현1동	15.24	202923.66	445844.28	77833	1634490
6.	강남구	논현2동	17.37	202555.35	445377.31	61357	552210
7.	강남구	대치1동	19.02	205118.99	443500.64	243635	974540
8.	강남구	대치2동	20.70	205724.67	443917.59	34563	138250
9.	강남구	대치3동	20.75	205732.80	444610.70	99199	892790
10.	강남구	대치4동	21.54	204924.16	444417.09	32570	162850
	...	...	...	...	...	...	...
441.	중랑구	상봉1동	26.22	207742.75	455436.51	71285	142570
442.	중랑구	신내1동	29.08	208758.48	455589.78	65573	262290
443.	중랑구	신내2동	29.31	208127.24	456913.04	59350	118700
444.	중랑구	중화3동	24.44	206634.45	455502.54	20660	61980

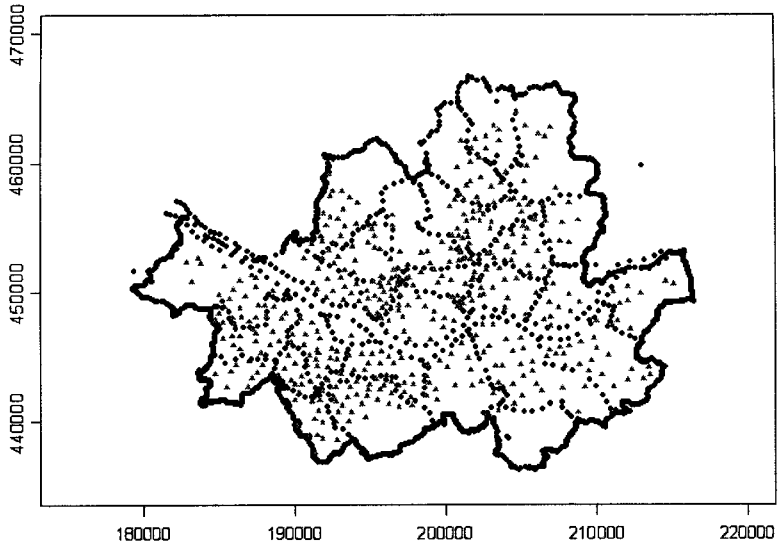
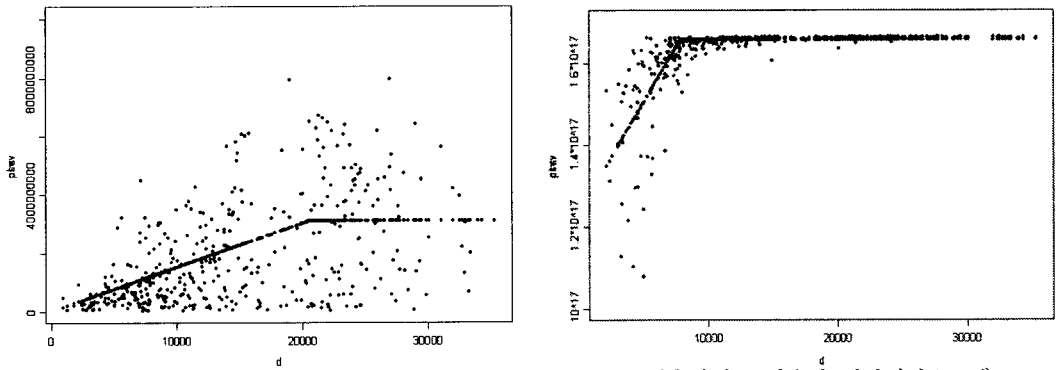


그림 3 : 분석에 이용된 서울시 444개 행정동 분포

3.2. 세미베리오그램의 추정

이론적인 세미베리오그램은 $\gamma(h)=\frac{1}{2}E(Z(s_i)-Z(s_j))^2$이기 때문에 관측된 지점으로부터 얻은 444개의 표본 세미베리오그램 $\frac{1}{2}(Z(s_i)-Z(s_j))^2$를 이용하여 이론적인 세미베리오그램을 추정할 수 있다. <그림 4>는 444개 행정동(동) 매출평균과 매출합계 자료로부터 444개의 표본 세미베리오그램 $\frac{1}{2}(Z(s_i)-Z(s_j))^2$를 점으로 나타낸 후 Piecewise Linear 모형을 적합시킨 결과를 보여주고 있다.



<매출평균을 이용한 세미베리오그램> <매출합계를 이용한 세미베리오그램>

그림 4 : 표본 세미베리오그램과 Piecewise Linear 모형 적합결과

표 2 : 매출평균과 매출합계에 대한 세미베리오그램 모수

세미베리오그램 모수	매출평균 데이터	매출합계 데이터
C_0	0	1.22×10^{17}
C_s	3.11×10^{9}	0.42×10^{17}
a_s	20671.49 (20.6Km)	7872.18 (7.8Km)
Q	2.30×10^{21}	4.73×10^{34}

여기에서 의미있는 세미베리오그램 모수로 a_s를 볼 수 있다. 이 모수 a_s는 일정거리 이상이 되면 거리가 매출에 영향을 주지 않는 즉, 거리와 매출이 서로 독립이 되는 거리를 말한다. 이 a_s를 할인점의 이용권의 최대 상권경계로 볼 수 있으며, 할인점 입장에서는 유효 마케팅 경계로 볼 수 있다. 매출평균 데이터는 a_s가 20671.49(m)로 매우 넓게 설정되었으나, 매출합계 데이터는 a_s가 7872.18(m)로 거리 상대적으로 짧으면서 의미있게 나왔다.

이승한(2003)이 할인점의 매출액에 따른 거리 분포를 구한 것을 토대로 살펴보면 이 7.8Km는 표 3에서 나타났듯이 대도시 할인점의 이용률에 따른 거리분포 중 매출의 95%에 해당하는 거리이다. 즉, 대도시 대형할인점의 매출의 95%를 포함하는 거리가 세미베리오그램에 의해 추정된 할인점 이용권 최대 상권경계와 거의 일치한다고 볼 수 있다.

표 3 : 할인점의 매출액에 따른 거리 분포 (각 점포 평균적용) : (이승한, 2003)

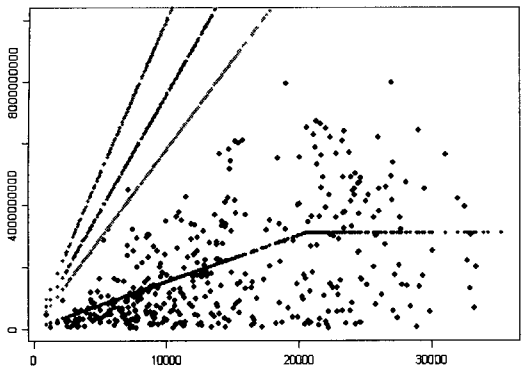
구분	80%	85%	90%	95%	100%
대도시	3.96	4.76	6.01	**7.70** **	17.67
중소도시	4.40	4.66	5.97	**8.33** *	16.65
주거지역	3.63	4.22	5.50	**7.45** *	16.88
상업지역	4.47	5.41	6.67	**8.55** *	17.86

3.3. 공간이상치 탐지 - Distributional Approach의 적용

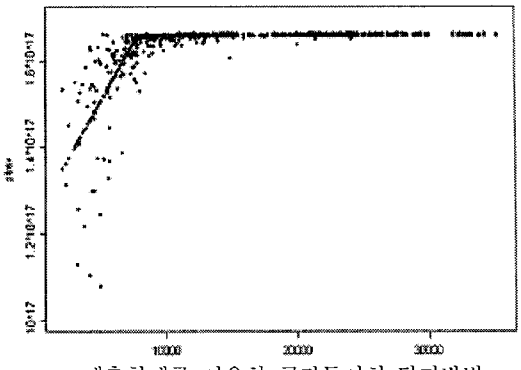
본 연구에서 실증분석을 위한 자료인 대형할인점 매출액 데이터를 이용하여 2.3절에서 소개한 공간이상치 탐지를 위한 Distribution Approach를 $\alpha=0.01/0.05/0.1$하에서 그려보았다(유성모, 엄익현, 1998).

그림 5에서 그래프상의 점은 444개의 표본 세미베리오그램 $\frac{1}{2}(Z(s_i)-Z(s_j))^2$을 나타내고 있고, 표본 세미베리오그램 토대로 추정한 Piecewise Linear 세미베리오그램 $\hat{\gamma}(s_i-s_j)$를 나타내고 있으며, 좌측 3개의 선은 유의수준 $\alpha=0.01/\ 0.05/\ 0.1$에서의 예측상한을 나타내고 있다.

그림 5의 <매출평균을 이용한 공간특이치 탐지방법>에 의한 그래프를 자세히 보면 $\alpha=0.1$의 경우는 총 444개의 표본 세미베리오그램 중 1개가 90% 예측상한 보다 크다. 그러나 유의수준을 5%로 설정하였으므로 총 444개 표본 세미베리오그램 중 1개가 예측상한을 벗어났다고 해서 특이치로 보기는 어렵다.



<매출평균을 이용한 공간특이치 탐지방법>



<매출합계를 이용한 공간특이치 탐지방법>

그림 5 : 대형할인마트 매출평균과 매출합계 자료에서 Distributional Approach에 의한 공간 특이 치 탐지 방법 (유의수준 $\alpha = 0.01/0.05/0.1$)

<매출합계를 이용한 공간특이치 탐지방법>에 의한 그래프를 보면 $\gamma(s_i - s_j) \times \chi^2_{1;1-\alpha}$값이 충분히 커서 공간특이치에 해당하는 것이 없었다. 그래프상에서도 $\gamma(s_i - s_j) \times \chi^2_{1;1-\alpha}$의 예측상한을 표시하기 어려울 정도로 값이 커서 공간특이치에 대한 문제는 없다고 볼 수 있다. 이런 결과는 분석 사전에 표본 세미베리오그램 중 값이 너무 커서 세미베리오그램의 추정모수의 값을 왜곡시키는 표본 세미베리오그램 값을 분석에서 제외하기 위해 표본 세미베리오의 최소값에서 부터 2.5%와 최대값에서 부터의 2.5%에 해당하는 전체 5%(22개 자료)를 제외하였기 때문이기도 하다.

3.5. 할인점 이용권 지도의 작성

그림6에서 매출평균에 의한 3D 표면도(Surface)는 동별 평균매출이기에 각 동의 소득수준을 반영하게 나타났다. 위치상 기준이 영등포임에도 강남권의 평균매출이 거리상 가까운 영등포의 매출평균과 비슷하거나 높게 나타난다. 반면에 매출합계에 의한 3D 표면도는 거리의 특성을 잘 반영해 영등포를 기준으로 거리가 가까운 곳의 총매출의 합계가 크게 나타나고 있다.

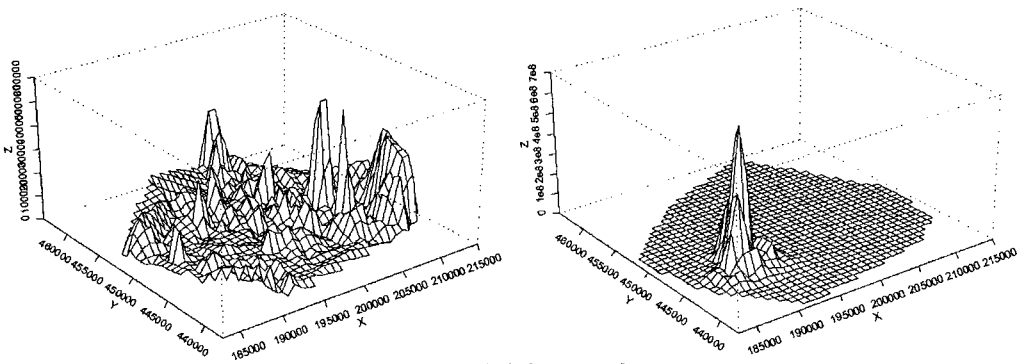


그림 6 : 매출평균/합계의 3D 표면도(Surface)

그림 7은 서울시 전체에서 영등포점의 매출에 관해 알 수 있는데, 매출평균에 의한 등고선도는 그림6의 표면도와 같이 각 동의 소득수준에 영향으로 매출평균이 전체적으로 고르게 퍼져 분포하였으며, 반면 매출합계에 의한 등고선도는 표2의 a_s가 7872.18(m)인 것을 보여준다. 이처럼 대형 할인점의 이용권을 구체적으로 표현한 등고선도는 그림8의 매출합계에 의한 등고선도라고 할 수 있다.

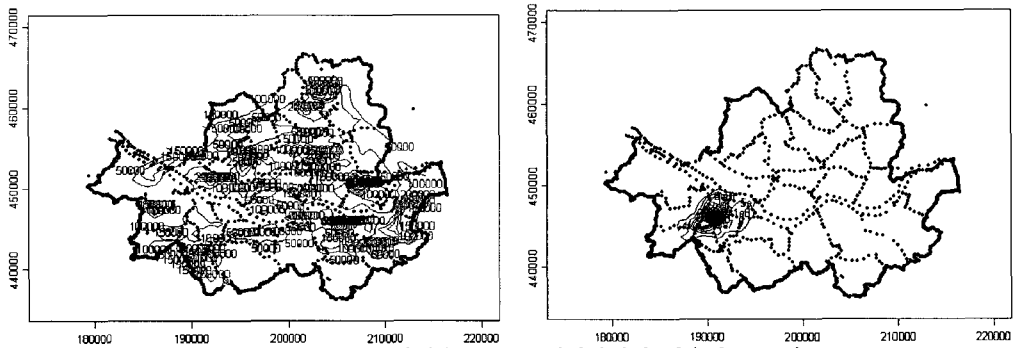


그림 7 : 매출평균/합계을 이용한 대형할인점 이용권 등고선도

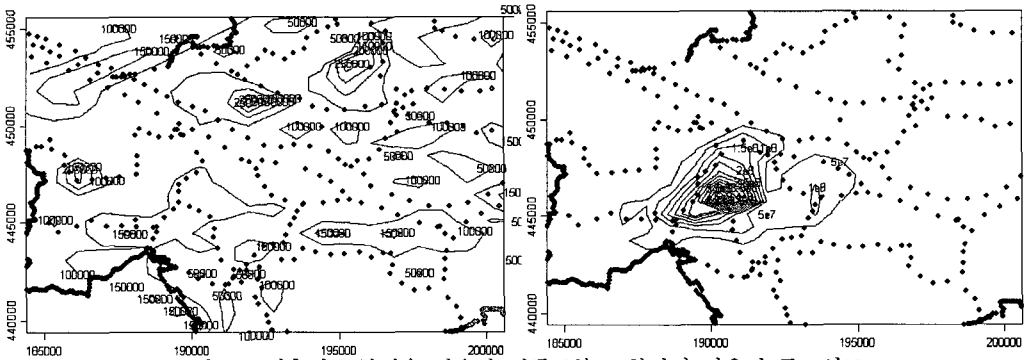


그림 8 : 매출평균/합계을 이용한 영등포부근 할인점 이용권 등고선도

4. 결론

본 논문에서는 서로 다른 위치에서 관찰된 자료들이 공간적인 변인인 거리에 의해서 서로 상관되어 있을 때, 이들 공간 변수에 대한 세미베리오그램의 추정, 공간 특이치 탐지 등의 방법들을 연구하였으며, 대형할인점 매출 자료를 이용한 사례분석을 통하여 실증적으로 검토해 보았다. 이러한 방법들은 공간상에 산재되어 있는 자료들의 공간적 상관관계를 모형화시킴으로서 기존의 독립성에 대한 가정의 경우보다는 보다 현실적이고 신뢰성있는 방법론이다.

하지만 본 논문에서는 가장 간단한 모형을 제시하고자 공간 변인 중 거리만을 고려한 모형을 제시하였다. 할인점 이용권, 구매력을 나타내기 위한 영향 요인으로 서론에서 거론했듯이 많은 요인들이 존재한다. 이런 거리만을 고려한 모형의 단점을 보완하기위해 현재 BPI(Buying Power

Index)를 이용한 세미베리오그램 모형을 연구중에 있다. 이 BPI는 서울시에 대한 지역의 인구수 비율, 소득액 비율에 점포까지의 실제 거리를 고려한 지수로써 이 지수를 이용하여 세미베리오그램을 추정하고 모형제시를 한다. 이는 좀 더 현실적인 접근으로 상권의 영향 요인을 포함한 모형으로 현재 거리만 고려한 모형보다 결과가 많이 개선될 것이다. 또한 세미베리오그램모형은 Piecewise Linear 모형 만에 국한하여 분석을 시도하였는데 거리외에 다른 요인이 공간적 변인으로 작용할 경우, 세미베리오그램의 종류를 보다 다양하게 적용하는 문제는 추후의 연구로 남겨둔다.

참고문헌

[1] 나종화, 김정숙 (2002). *SPLUS 입문과 활용*, 자유아카데미, 서울.

[2] 성균관대학교 측지정보학 연구실 홈페이지, 좌표변환서비스, http://gro.skku.ac.kr

[3] 신정신 (1997). 구매력의 지표와 측정에 관한 연구, *광고연구*, 1997년 여름호, 163-176.

[4] 유성모, 엄익현 (1999). 강우강도 데이터를 이용한 세미베리오그램의 추정과 공간이상치에 관한 연구, *응용통계연구*, 제12권 1호, 125-141.

[5] 이승한 (2003). *유통시설의 지역경제 파급효과 및 이용권 분석 - 할인점 중심으로 -*, 박사학위논문, 한양대학교 대학원, pp. 84-113.

[6] 임명숙 (2004). *대형쇼핑시설의 유형별 입지 특성 및 소비자행태에 관한 연구*, 박사학위논문, 단국대학교 대학원.

[7] Cressie, N. (1993). *Statistics for Spatial Data*, New York, John Wiley & Sons, INC.

[8] Mando Map &Soft. 지도정보서비스, http://www.speednavi.co.kr

[9] Matheron, G. (1963). Principles of Geostatistics, *Economic Geology*, 58, pp. 1246-1266.

[10] Mathsoft Inc. (1996). *S+SPATIALSTATS User's Manual*, Mathsoft Inc, Seattle, Washington.

[11] Insightful Corporation (2001). *S-plus for Window Programmer's Guide*, Insightful Corporation, Seattle, Washington.

[12] Truman Asa Hartshon (1992). *Interpreting The City: An Urban Geography*, New York: John Willey & Sons, Inc., p. 387.