Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 30 - May 1, 2004, pp. 183-194

Categorical Data Analysis by Using Spatial Scan Statistics and Echelon Analysis

문승호[1] · 신재경[2]

요 약

본 연구에서는 공간 검색 통계량(spatial scan statistics)과 에셜론 해석법을 이용한 범주형 자료 분석을 다룬다. 이를 위해 우선, 에셜론 덴드로그램을 이용하여 주어진 분할표의 계층적 구조(hierarchical structure)를 결정하고서 이로부터 핫스팟(hotspot)의 후보를 검출한다. 다음으로 우도비(likelihood ratio)를 기초로 유의하게 높거나 낮게 나타나는 지역에 대한 공간 검색 통계량을 산출한다. 마지막으로, 이 통계량을 바탕으로 핫스팟을 검출한다.

Keywords : 범주형 자료분석, 에셜론 해석법, 공간 검색 통계량

1. 서론

분할표로 주어진 데이터를 해석하기 위해서는 독립성 검정, 대수선형모델이나 로짓모델 등을 이용한 접근법이 행해진다. 카이제곱검정 등을 이용해서 독립성 검정을 행할 경우, 2×2 분할표에서의 결과 해석은 명쾌하지만 범주(category) 수가 2이상이 되면 해석하기에 다소 곤란하다. 예를 들어, 아래 표1의 양친의 사회경제상태("high"부터 "low"까지 6단계)와 정신적 건강상태("well"부터 "impared"까지의 4단계)로 분류된 1660명에 대한 6×4 순서 카테고리를 지닌 분할표 데이터를 고려하자.

변수간의 독립성 검정에서 자유도 15인 카이제곱 통계량은 45.985(p-value=0.0001)로 유의하게 되어 이 두 변수는 서로 독립이 아니라는 결론을 내리게 된다. 하지만, 카이제곱 통계량을 이용한 분석에서는 분할표에서의 비독립성이 전체를 통해서 얻어진 것인지 또는, 분할표 내의 특정 셀 또는 셀군이 독립성으로부터 일탈해 있는지에 관한 정보는 얻을 수 없다. 이러한 문제에 관해서는 각 셀 마다 조정된 잔차를 이용한 정규검정 등이 이용된다. 그러나, 표1과 같은 순서 카테고리를 지닌 분할표 데이터의 경우, 인접한 셀 간에는 순서성에 의해 공간적인 연계가 있으므로 셀 마다 개별적 검정을 행하는 것은 적당하지 않으며 인접한 공간적 위치 정보도 고려한 셀 군을 찾아내는 검정을 수행하는 것이 바람직하다. 본 연구에서는 분할표 데이터의 해석에 공간구조 개념을 도입해

[1](608-738) 부산광역시 남구 우암동 55-1, 부산외국어대학교 정보통계학과 부교수
E-mail : shmoon@pufs.ac.kr
[2](641-773) 경남 창원시 사림동 9번지, 창원대학교 통계학과 교수

서 유의하게 이 연관의 핵이 되어있는 셀 군의 검출을 수행하는 방법에 관해 논하고자 한다. 즉, 에셜론 해석에 의해 분할표의 공간적 계층구조를 구하고서, 그 구조를 바탕으로 공간 검색 통계량을 계산함에 따라 유의하게 독립성으로부터 벗어나 있는 셀군을 찾아내는 방법에 관해 연구하고자 한다.

표1. 양친의 사회경제상태와 정신적 건강상태로 분할된 6×4 분할표

Parent's Socio-economic Status	Mental Health Status				
	1. well	2. mild symptom formation	3. moderate symptom formation	4. impared	sum
A(high)	64	94	58	46	262
B	57	94	54	40	245
C	57	105	65	60	287
D	72	141	77	94	384
E	36	97	54	78	265
F(low)	21	71	54	71	217
Sum	307	602	362	389	1660

2. 에셜론 해석2. 신뢰성 시험

합금무계목인발강관은 산업, 선박용 플랜트 사업의 열교환기 튜브로 사용되는데 정격 사용온도는 450℃이고 보증수명은 20년이다. 동 소재의 신뢰성 인증은 "고온고압용 보일에셜론 해석(Myers et al., 1997)은 공간상에 분류된 지도상의 1변량 데이터에 대해 공간적 위치를 표면상의 데이터 고저(高低)를 바탕으로 분할, 공간 데이터의 위상적(topological) 구조 및 변화를 계통적(hierarchical)이고 객관적으로 발견하기 위해서 개발된 해석법이다. 이러한 데이터의 계층적 구조나 원형데이터의 위상적 표면의 변화를 나타내기 위해 에셜론 덴드로그램(echelon dendrogram)을 이용한다. 핫스팟(hotspot)으로 간주할 수 있는 여러 개의 지역들은 에셜론 덴드로그램상에서 피크(peak)로 표현된다.

(1) 데이터 고저가 1차원적 수평위치인 경우

지형도(地形圖)에서의 단면도와 같이 데이터 고저가 1차원적인 수평위치로 주어졌을 경우, 공간(수평) 위치 x에서의 데이터 값(고저를 나타냄)을 h라 하면 데이터는 (x, h)로 표시된다. 그림 2-1에 나타낸 것과 같은 데이터가 있을 경우, 위상적(位相的)으로 동일한 9개의 계급으로 나눌 수가 있다. 이들 계급이 각각 에셜론이며, 그림 2-1에 주어진 번호가 에셜론 번호(echelon number)이다. 에셜론 1부터 5는 피크이며 1차 에셜론이라 부른다. 에셜론 6은 2개 이상의 피크(1차 에셜론 1과 2)의 파운데이션(foundation)이며, 2차 에셜론이라고 한다. 마찬가지로 에셜론 7은 1차 에셜론 3과 4의 파운데이션이며 2차 에셜론이다. 에셜론 8은 2개 이상의 2차 에셜론 6과 7의 파운데이션이며 3차 에셜론이라 부른다. 에셜론 9는 에셜론 5와 8의 파운데이션이지만, 에셜론 8은 3차 에셜론 5는 1차이므로 에셜론 차수는 3차이다. 또한, 에셜론 9는 루트(root)라고도 한다.

그림 2-1에서 보여주는 데이터 구조는 그림 2-2와 같은 에셜론 덴드로그램으로 주어진다.

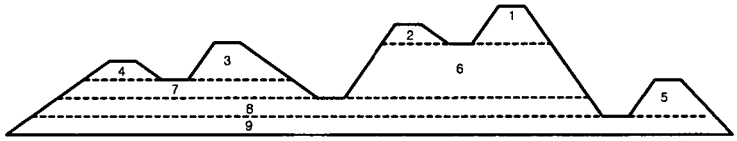

그림2-1. 에셜론 해석에 있어서 동일 위상 영역으로의 분할

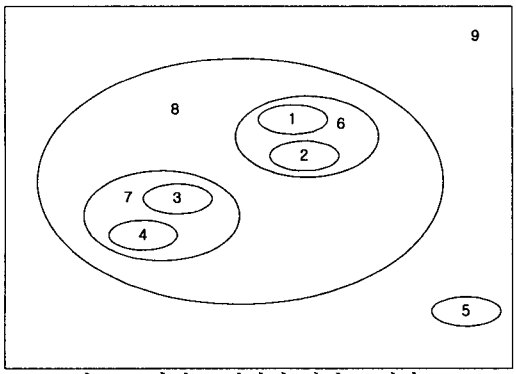

그림2-3. 2차원 공간상의 에셜론 해석

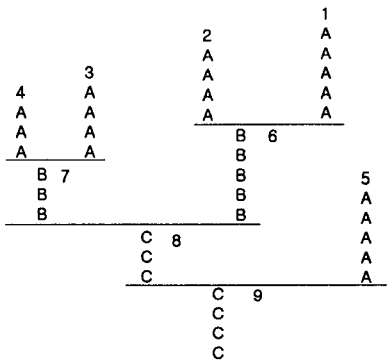

그림2-2. 에셜론 덴드로그램

(2) 데이터 고저가 2차원 공간상의 간단한 구조로 주어진 경우

데이터 고저가 2차원 공간에 주어진 경우, 공간의 위치 (x, y)에 있어서의 데이터 값(고저를 나타냄)을 h라 하면, 데이터는 (x, y, h)로 표현된다. 일반적으로 지형도와 같이 데이터 고저가 비교적 간단한 구조로 주어지는 경우, 이들 세 변수간에는 연속함수 $h=f(x, y)$의 관계가 있다. 그림 2-3에서 나타내는 2차원 공간 데이터의 구조도 그림 2-2와 같은 에셜론 덴드로그램으로 주어진다.

(3) 데이터 고저가 2차원 공간의 배열(mesh) 상에 주어진 경우

Remote sensing data와 같이 데이터 고저(高低)가 $n \times m$ 배열상의

$$D_{ij} = \{(x, y) \mid x_{i-1} < x < x_i\ ,\ y_{j-1} < y < y_j\},\ i = 1, 2, \cdots, n,\ \ j = 1, 2, \cdots, m$$

의 계수값 h로 주어지는 공간 데이터인 경우, 데이터는 (i, j, h)로 나타낸다. 이와 같은 공간 데이터의 경우 이들 세 변수의 관계를 나타내는 함수 $h=f(i, j)$는 이산적(離散的)이며 복잡한 함수가 되어 구조도 복잡하게 된다. 그림 2-4와 같은 5×5 배열상의 데이터가 주어질 때, 아래와 같은 방법으로 에셜론 해석이 진행된다. 단, 인접한 데이터란 대상으로 하고 있는 에셜론에 포함되는 데이터에 인접하는 데이터 집합을 뜻한다.

2	24	8	15	3
10	1	14	22	5
4	13	19	23	25
20	21	12	11	17
16	6	9	18	7

그림 2-4. 5 × 5배열상의 데이터

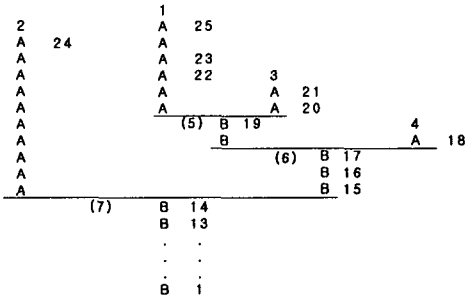

그림 2-5. 에셜론 덴드로그램

Step-1. **피크의 발견**

피크에 속하는 데이터 값은 동일한 피크에 속하는 데이터 이외의 인접하는 데이터 값보다 크다. 그림 4-1의 데이터에 있어서는 4개의 피크가 있다.

① 제1피크

최대값은 25이다. 우선, 25는 제1피크에 포함된다. 25에 인접하는 데이터 중에서 최대값은 23으로, 23은 (25, 23)에 인접하는 데이터보다 크므로 제1피크에 포함된다. (25, 23)에 인접하는 데이터 중에서 최대값은 22로, 22는 (25, 23, 22)에 인접하는 데이터보다 크므로 22도 제1피크에 포함된다. (25, 23, 22)에 인접하는 데이터 중에서 최대값은 19이다. 하지만 19는 (25, 23, 22, 19)에 인접하는 21보다 작으므로 제1피크에 속하지 않는다. 따라서 제1피크는 데이터 25, 23, 22에 의해 구성되며, 에셜론 번호는 1(En 1)이다. 이들 데이터는 동일한 피크 이외의 인접하는 데이터보다 크다.

② 제2피크 이후

제1피크를 제외한 데이터 중에서 최대값은 24이다. 이 24는 제2피크에 포함된다. 24에 인접하는 데이터 중에서 최대값은 14이지만 인접하는 23보다 작기 때문에 제2피크에 속하지 않는다. 그러므로 제2피크 (En 2)는 24에 의해서만 구성된다. 동일한 순서에 의해 제3피크 (En 3)는 21, 20, 제4피크 (En 4)는 18에 의해 구성된다.

Step-2. **파운데이션의 발견**

① 제1파운데이션

4개 피크에 포함되는 데이터를 제외한 최대값은 19이다. 19는 제1피크 (En 1)와 제3피크 (En 3)의 파운데이션이며 에셜론번호는 5가 된다 (En 5). (En 1, 3, 5)에 인접하는 데이터 중에서 최대값은 17이다. 그러나 17은 제4피크의 18보다 작으므로 17은 En 5에 속하지 않는다. En 5는 En 1과 En 3의 부모(parent)이며 이 관계는 에셜론 번호를 이용해서 5(1 3)로 표시된다. 이후 파운데이션을 찾게 되면 En 1과 En 3은 사용되지 않고 대표로 En 5를 이용한다.

② 제2파운데이션

En 1부터 5에 포함되는 데이터를 제외한 최대값은 17이다. 17은 En 5와 En 4의 파운데이션이며 En 6이다. En 6은 En 5와 En 4의 부모가 되며 6(5(1 3)4)이다. 이후, En 1, En 3, En 4, En 5는 대표로 En 6을 이용한다. En 1부터 6에 포함되는 데이터를 제외한 최대값은 16이다. 이 16은 (En 1, 3, 4, 5, 6)에 인접하는 데이터보다 크므로 En 6에 속한다. En 1부터 6(단, 16은 En 6에 포함된다)에 포함되는 데이터를 제외한 최대값은 15이다. 이 15도 En 6에 속한다. En 1부터 6(단, 16, 15는 En 6에 포함된다)에 포함되는 데이터를 제외한 최대값은 14이다. 그렇지만 14는 (En 1, 3, 4, 5, 6)에 인접하는 데이터의 제2피크의 24보다 작으므로 14는 En 6에는 속하지 않는다.

③ 제3파운데이션(root)

En 1부터 6에 포함되는 데이터를 제외한 최대값은 14이다. 14는 En 2와 En 6의 파운데이션이며 En 7이다. En 7은 En 2와 En 6의 부모가 되며 7(2 6(5(1 3)4))이다. 13이하의 데이터는 루트 En 7에 속한다. 이상에 의해서 이 5×5 배열 데이터의 구조는 그림 2-5와 같은 에셜론 덴드로그램에 의해 주어진다.

(4) 지역 데이터의 경우

인구 데이터나 질병의 발생률 데이터와 같이 대상으로 하는 데이터가 거의 그 지역에 널리 퍼진 영역 데이터는, 지역 D_i $(i = 1, 2, \cdots, n)$ 안의 데이터 값을 h라 하면, 데이터는 (i, h)로 표시된다. 그림 2-6의 데이터는 2001년도 주민등록인구에 따른 전국 16개 시·도의 인접정보와 총인구 데이터를 나타낸다(단위는 10만 명). 이와 같은 지리 데이터에 대해서도 피크, 파운데이션과 루트를 발견할 수 있다. 이들 시·도의 총인구에 대한 구조는 그림 2-7과 같은 에셜론 덴드로그램에 의해 표현되어진다. 이 그림으로부터 경기도를 포함한 경인지역과 다음으로 부산·경남 지역의 두 개의 큰 피크가 있음을 알 수 있다(문승호, 2003).

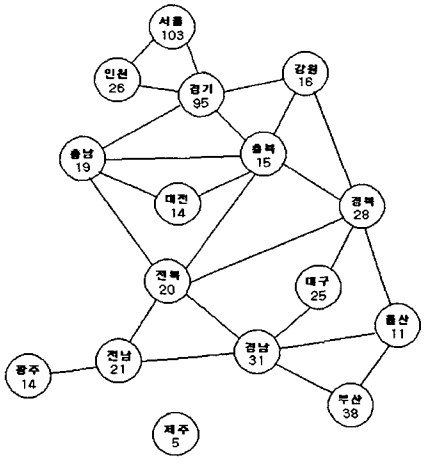

그림2-6. 지역의 인접정보

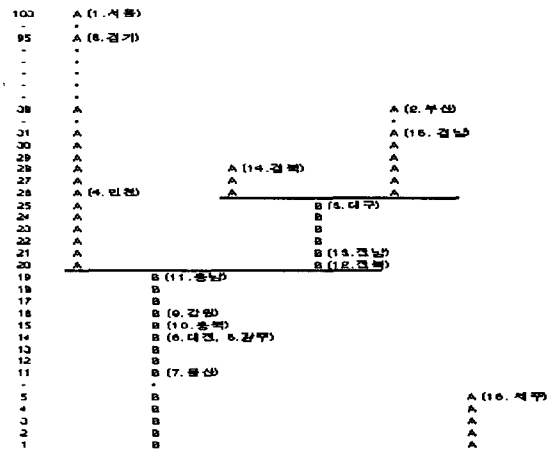

그림2-7. 총인구 데이터의 에셜론 덴드로그램

3. 공간 검색 통계량(spatial scan statistic)

공간통계학에서의 일반적인 관심사항은 일련의 점(point)들이 랜덤(random)하게 분포되어 있느냐, 아니면 일정한 군집(cluster)성을 가지고 있느냐는 것이다. 이러한 군집의 위치를 찾아내고자 하는 경우에 사용할 수 있는 방법이 바로 공간 검색 통계량(Kulldorff, 1977)이다. 공간 검색 통계량은 질병의 발생률과 같이 지역마다 얻어지는 데이터에 있어서 유의하게 높은 비율로 나타나는 핫스팟(hotspot) 지역을 발견하는데 이용된다.

전체지역을 G라하고, 그 중 하나의 지역을 Z라 하자. 지역 Z안에 일련의 점들이 포함될 확률을 p, 영역 Z외부의 점들이 포함될 확률을 q라 하자. 이때 이 점들은 서로 독립이다. 따라서, 가설은 다음과 같다.

$$\text{귀무가설 } H_0: p=q \ v.s. \text{ 대립가설 } H_1: p>q.$$

이 때,

$n(G)$: 전체영역 G에서의 모집단 수, $n(Z)$: 영역 Z 내부의 모집단 수,

$c(G)$: 전체영역 G에서 속성을 지닌 개체 수, $c(Z)$: 영역 Z 내부에서 속성을 지닌 개체 수라 하자.

두 모형 베르누이 모형(Bernoulli model)과 포아송 모형(Poisson model)을 고려하자.

(1) 베르누이 모형(Bernoulli model)

베르누이 모형의 우도함수(likelihood function)는 다음과 같이 주어진다.

$$L(Z,p,q)=p^{c(Z)}(1-p)^{n(Z)-c(Z)}q^{n(G)-n(Z)}(1-q)^{(n(G)-n(Z))-(c(G)-c(Z))}$$

$x(A)$를 영역 A에서 일련의 속성을 가지는 개체 수를 나타내는 확률변수라 하면, 귀무가설이 참이라는 가정 하에서는

$$x(Z)\sim B(n(Z),p), \quad x(Z^c)\sim B(n(G)-n(Z),p)$$

대립가설 하에서는

$$x(Z)\sim B(n(Z),q), \quad x(Z^c)\sim B(n(G)-n(Z),q)$$

이 성립한다.

군집이 될 가능성이 높은 지역을 찾기 위해서, 영역 Z가 주어진 조건에서 이 우도함수를 최대로 하는 값을 구해야 한다.

$$L(Z)=\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)}\left(1-\frac{c(Z)}{n(Z)}\right)^{n(Z)-c(Z)}$$
$$\times\left(\frac{(c(G)-c(Z))}{(n(G)-n(Z))}\right)^{(c(G)-c(Z))}\left(1-\frac{(c(G)-c(Z))}{(n(G)-n(Z))}\right)^{((n(G)-n(Z))-(c(G)-c(Z)))}$$
$$\text{if}\quad \frac{c(Z)}{n(Z)}>\frac{c(G)-c(Z)}{n(G)-n(Z)}\ \frac{c(G)-c(Z)}{n(G)-n(Z)}\ \text{and otherwise}$$
$$L(Z)=\left(\frac{c(G)}{n(G)}\right)^{c(G)}\left(1-\frac{c(G)}{n(G)}\right)^{n(G)-c(G)}.$$

우도비 λ는 핫스팟을 발견하기 위해서 전체영역의 부분집합인 영역 Z에서 최대인 것으로 한다.

$$\lambda=\frac{Max_Z L(Z)}{L_0}$$

단, L_0는 귀무가설 하에서의 우도함수 값이다.

$$L_0=\left(\frac{c(G)}{n(G)}\right)^{c(G)}\left(1-\frac{c(G)}{n(G)}\right)^{n(G)-c(G)}$$

대립가설 하에서 속성을 지닐 확률은 영역 밖보다 안쪽이 보다 높은 핫스팟이 있지만, 그 후보를 발견하는 검정통계량으로서 로그 우도비 통계량 $\log\lambda$를 계산한다. Kulldorff(1997)는 검색할 원의 중심을 시·도의 중심으로 하고 반지름은 인구의 절반이 될 때까지 변화시켜, 최대의 공간검색 통계량을 지니는 원을 most likely cluster로 정의하고 있다. 또한, 귀무가설 상의 $\log\lambda$를 몬테칼로(Monte Carlo)법에 의해 계산, p-값(p-value)도 계산하고 있다.

(2) **포아송 모형**(Poisson model)

전체영역 G에서 속성을 지닐 개체 수가 $c(G)$가 될 확률은

$$\frac{\exp[-pn(Z)-q(n(G)-n(Z))][pn(Z)+q(n(G)-n(Z))]^{c(G)}}{c(G)!}$$

모든 영역 안에서의 지점 x에서의 밀도는

$$\frac{pn(x)}{pn(Z)+q(n(G)-n(Z))}\quad \text{if } x\in Z,\quad \frac{qn(x)}{pn(Z)+q(n(G)-n(Z))}\quad \text{if } x\in Z^c.$$

이 때, 포아송 모형에 대한 우도함수는 아래와 같이 주어진다.

$$L(Z,p,q)=\frac{\exp[-pn(Z)-q(n(G)-n(Z))]}{c(G)!}p^{c(Z)}q^{c(G)-c(Z)}\prod_{x_i}n(x_i)$$

$x(Z)$를 영역 Z내에서 속성을 지니는 개체 수를 나타내는 확률변수라 할 때, 임의의 영역 A에 있어서

$$x(A) \sim Poisson(pn(A \cap Z) + qn(A \cap Z^c))$$

귀무가설 하에서는 임의의 영역 A에 있어서

$$x(A) \sim Poisson(pn(A))$$

우도함수를 최대로 하기 위해서 영역 Z가 주어진 조건에서 최대우도함수를 계산한다.

$$L(Z) = \begin{cases} \frac{e^{-n(G)}}{c(G)!}\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)}\left(\frac{c(G)-c(Z)}{n(G)-n(Z)}\right)^{c(G)-c(Z)}\prod_{x_i} n(x_i), & \text{if } \frac{c(Z)}{n(Z)} > \frac{c(G)-c(Z)}{n(G)-n(Z)} \\ \frac{e^{-n(G)}}{c(G)!}\left(\frac{c(G)}{n(G)}\right)^{c(G)}\prod_{x_i} n(x_i) = L_0, & otherwise \end{cases}$$

우도비 λ는 핫스팟을 발견하기 위해서 전체 영역의 부분집합인 영역 Z에서 최대인 것으로 한다.

$$\lambda = \begin{cases} \frac{Max_Z L(Z)}{L_0} = \frac{\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)}\left(\frac{c(G)-c(Z)}{n(G)-n(Z)}\right)^{c(G)-c(Z)}}{\left(\frac{c(G)}{n(G)}\right)^{c(G)}}, & \text{if } \frac{c(Z)}{n(Z)} > \frac{c(G)-c(Z)}{n(G)-n(Z)} \\ 1, & otherwise \end{cases}$$

검정통계량 λ는 아래와 같이 나타낼 수도 있다.

$$\lambda = \begin{cases} \left(\frac{c(Z)}{e(Z)}\right)^{c(Z)}\left(\frac{c(G)-c(Z)}{e(G)-e(Z)}\right)^{c(G)-c(Z)}, & \text{if } c(Z) > e(Z) \\ 1, & otherwise \end{cases}$$

이 때, $e(Z)$는 영역 Z내에서 속성을 지니는 개체 수의 기댓값이며 $e(G) = c(G)$이다.

4. $r \times c$ 순서범주형 분할표에 대한 핫스팟의 검출

어느 모집단으로부터 추출된 표본이 두 변수 x와 y에 관해 각각 r개와 c개의 순서범주에 의해 분류된 $r \times c$ 분할표를 고려하자. 이 때, 이 분할표 데이터는 원격탐사(remote sensing) 데이터와 같이 $r \times c$의 2차원 배열(mesh) 상에 공간구조를 가지는 데이터로 간주할 수 있다. p_{ij}를 i행 j열 범주에 대한 모집단 확률($i = 1, 2, \cdots, r$ $j = 1, 2, \cdots, c$), p_i와 p_j를 각각 행과 열의 주변확률을 나타내는 것으로 다음과 같이 정의하자.

$$p_{i.} = \sum_j p_{ij}, \quad p_{.j} = \sum_i p_{ij}, \quad p_{..} = \sum_i p_{i.} = \sum_j p_{.j} = \sum_i \sum_j p_{ij} = 1$$

이때, 귀무가설로 두 변수가 독립이라는 가설을 고려하자.

$$H_0 : p_{ij} = p_{i.} \cdot p_{.j} \quad \text{for } i = 1, 2, \cdots, r \text{ and } j = 1, 2, \cdots, c$$

$n(ij)$와 $c(ij)$를 각각 i행 j열 범주에 대한 모집단의 크기 및 관측도수, $n(i.), c(i.)$과 $n(.j), c(.j)$를 각각 행과 열의 모집단 및 관측도수에 대한 주변도수라 하자. 또, $n(..)$과 $c(..)$는 각각 모집단 크기 및 총 관측도수로 한다. 귀무가설 하에서, i행 j열 범주의 모집단 확률

$$p_{ij} = \frac{n(i.)n(.j)}{n(..)^2}$$

의 최우추정량 및 기대도수는 다음 식으로 주어진다.

$$\hat{p}_{ij} = \frac{c(i.)c(.j)}{c(..)^2}, \quad \hat{e}_{ij} = c(..)\,\hat{p}_{ij} = \frac{c(i.)c(.j)}{c(..)}$$

대립가설 하에서 두 변수는 독립이 아니라 어떠한 연관이 있다. 어떤 범주가 관련하고 있는지를 살펴보기 위해서는 아래와 같은 통계량에 기인한 방식이 적당하다.

$$d1(ij) = c(ij) - \hat{e}(ij), \quad d2(ij) = \frac{c(ij)}{\hat{e}(ij)}, \quad d3ij = \frac{(c(ij) - \hat{e}(ij))^2}{\hat{e}(ij)}, \quad d4ij = \frac{c(ij) - \hat{e}(ij)}{\sqrt{\hat{e}(ij)}}$$

이들 통계량이 큰 값을 취하는 범주 영역이 핫스팟 후보가 된다. 이처럼 유의하게 높은 연관이 있는 범주 영역을 조사하기 위해서 공간 검색 통계량을 이용한다.

분할표에서 셀군(영역)을 Z, 분할표 전체를 G로 하자. 이때, 포아송 모형에 대한 우도비는 다음과 같이 주어진다.

$$\lambda = \frac{\left(\frac{c(Z)}{n(Z)}\right)^{c(Z)} \left(\frac{c(G)-c(Z)}{n(G)-n(Z)}\right)^{c(G)-c(Z)}}{\left(\frac{c(G)}{n(G)}\right)^{c(G)}} = \left(\frac{c(Z)}{e(Z)}\right)^{c(Z)} \left(\frac{c(G)-c(Z)}{e(G)-e(Z)}\right)^{c(G)-c(Z)}$$

포아송 모형에서의 로그 우도비 통계량을 이용하여 표1의 조기퇴직 시스템과 연금제도에 관한 분할표 데이터에 대해서 유의하게 독립성에서 벗어나는 셀을 발견한다. 여기서는 관련의 근원을 재는 데이터로서 Kulldorff(2003) 등이 개발한 소프트웨어 SaTScan에 따라 분할표의 셀 데이터로서 상대 위험율(relative risk)을 이용한다.

5. 수치예

분할표의 각 셀에서 상대 위험율을 계산한 결과가 표2이다. 이 데이터를 2절 (3)에서 설명한 바와 같이 고저가 2차원 공간 배열 상의 데이터로서 다루면, 그림2-8과 같은 에셜론 덴드로그램을 작성할 수 있다.

에셜론 해석에 의한 분할표의 공간적 계층구조로서 그림2-8이 얻어졌는데, 그 구조를 바탕으로 해서 공간검색 통계량을 계산, 유의하게 독립성에서 벗어난 셀을 발견한다. 에셜론 덴드로그램을 바탕으로 로그 우도비 통계량을 계산하면, 제1피크로서 impared-F(low)을 정점으로 impared-D, E, F와 moderate-F가, 제2피크로서는 well-A(high)를 정점으로 well-A, B, C와 mild_B, 그리고, 이들의 파운데이션으로서 moderate-C에 의해 구성되며, ((F4, E4, F3, D4), (A1, B1, C1, B2), C3)가 로그 우도비 통계량 9.71로 핫스팟의 후보로 뽑혀진다. 또한, SaTScan을 이용한 핫스팟의 후보로 첫 번째는 (F4, F3, E4)로 통계량은 5.86(p-value=0.036), 두 번째는 (B1, A1)으로 통계량은 3.13(p-value=0.343)이다. 이로부터 우리는 에셜론 덴드로그램에 의해 구조를 파악하고 그 구조에서 피크로 나타나는 셀들을 핫스팟으로 간주하는 에셜론 해석법과 검색 전용 소프트웨어인 SaTScan의 결과(표3 참조)가 유사함을 알 수 있다.

표2. 표1의 6×4 분할표 데이터의 상대 위험율

Parent's Socio-economic Status	Mental Health Status			
	1. well	2. mild symptom formation	3. moderate symptom formation	4. impared
A(high)	1.321	0.989	1.015	0.749
B	1.258	1.058	1.011	0.697
C	1.074	1.009	1.039	0.892
D	1.014	1.013	0.920	1.045
E	0.735	1.009	0.934	1.256
F(low)	0.523	0.902	1.141	1.396

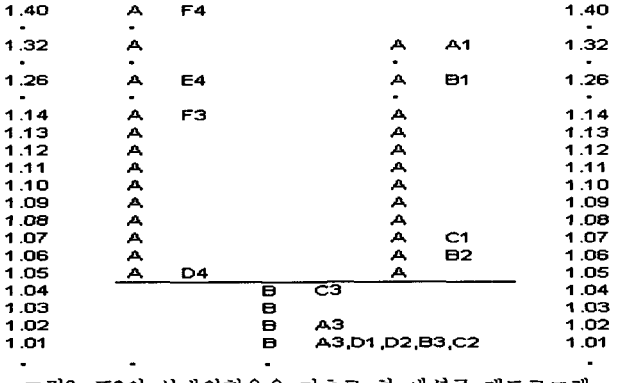

그림8. 표2의 상대위험율을 기초로 한 에셜론 덴드로그램

표3. 표1의 6×4 분할표 데이터에 대한 SaTScan 결과(일부)

MOST LIKELY CLUSTER

1.Location IDs included.: 24, 23, 20
Coordinates / radius..: (5.999989 N, 4.000000 E) / 111.12 km
Population.....: 266052
Number of cases.....: 203 (160.27 expected)
Annual cases / 100000.: 76.4
Overall relative risk.: 1.267
Log likelihood ratio..: 5.862172
Monte Carlo rank.....: 36/1000
P-value.....: 0.036

SECONDARY CLUSTERS

2.Location IDs included.: 5, 6, 9, 1
Coordinates / radius..: (1.999978 N, 1.000000 E) / 111.12 km
Population.....: 391248
Number of cases.....: 272 (235.69 expected)
Annual cases / 100000.: 69.6
Overall relative risk.: 1.154
Log likelihood ratio..: 3.129963
Monte Carlo rank.....: 343/1000
P-value.....: 0.343

6. 마무리

본 연구에서 우리는 분할표로 주어진 데이터에 대해 공간 검색 통계량과 에셜론 해석법을 적용해 보았다. 데이터의 공간적 계층구조를 이용하는 에셜론 해석법의 결과가 객관적으로도 충분한 해석이 가능하다는 장점을 수치 예에서 알 수 있었다. 이처럼 공간 탐색 통계량을 이용한 에셜론 해석법에 의한 핫스팟의 발견은 여러 분야에의 적용이 가능하다. Kurihara and Hong(2002)에서는 이를 이용하여 유아 돌연사 증후군(sudden infant death syndrome)에 대한 핫스팟 지역의 발견을 다루고 있다. 또한, 공간 군집화 방법(spatialclu stering methods)과 에셜론 분석을 이용하여 국내에서 발생한 업무상 질병자에 대한 질병지도를 작성, 핫스팟 지역을 찾아내어 그 지역의 예방활동의 기초자료로 활용할 수 있는 연구도 있다(홍한표, 2003). 이와 같이 에셜론 분석법을 이용해서 핫스팟 지역을 탐색, 이 결과와 기존의 공간 검색 통계량의 결과를 비교, 분석한다면 여러 가지 질병 등에 대한 원인 규명과 함께 예방에도 많은 도움이 될 것이다. 국내의 관련 자료에 대한 연구가 차후의 연구과제로 남아있다.

참고문헌

1. 문승호 (2003). Echelon Analysis, *Journal of the Korean Data Analysis Society,* 5(2), 273-281.

2. 홍한표 (2003). 공간 군집화 방법과 에셜론 분석을 이용한 핫스팟 지역의 발견, 통계청 「통계연구」, 8(2), 131-153.

3. Anderson, E.B. (1994). *The Statistical Analysis of Categorical Data.* Springer-Verlag.

4. Kulldorff, M. (1997). A spatial scan statistics, *Communications in Statistics, Theory and Methods,* 26, 1481-1496.

5. Kulldorff, M. (2003). SaTScanTM User Guide for version 4.0. (http://www.satscan.org).

6. Kurihara, K. and Hong, H. (2002). Detection of Hotspots for Geospatial Data with Echelon Analysis Based on Spatial Scan Statistics, *Proceedings of the 4th Conference of the IASC,* 189-192.

7. Myers, W.L., Patil, G.P. and Joly, K. (1997). Echelon approach to areas of concern in synoptic regional monitoring, *Environmental and Ecological Statistics,* 4, 131-152.

8. Strole, L., Langner, S.T., Michael, S.T., Opler, M.K. and Rennie, T.A.C. (1978). Mental health in the Metropolis : The midtown Manhattan study, NYU Press.