Journal of the Korean Data & Information Science Society
2016, **27**(5), 1193–1201

http://dx.doi.org/10.7465/jkdi.2016.27.5.1193
한국데이터정보과학회지

스캔 통계량을 이용한 암 클러스터 탐색[†]

한준희[1] · 이민정[2]

[1]양산부산대학교병원 의학통계실 · [2]강원대학교 정보통계학과

접수 2016년 8월 28일, 수정 2016년 9월 20일, 게재확정 2016년 9월 21일

요 약

공간 또는 시공간 데이터에서 다른 지역에 비해 유난히 높은 위험률을 보이는 소위 핫 스팟 (hot spot)으로 불리는 클러스터 (cluster)를 찾으려고 하는 경우가 많다. 기존의 많은 방법들은 이러한 클러스터 패턴이 존재하는지에 대한 해답만 주었지만, 최근의 많은 방법들은 클러스터의 위치, 모양, 크기뿐만 아니라 찾아진 클러스터가 통계적으로 유의한지까지 검정해준다. 본 논문에서는 이러한 다양한 방법 중 가장 많이 사용되는 클러스터 탐색 방법 중 하나인 스캔 통계량을 이용한 방법을 소개하고 그 방법이 구현된 무료 소프트웨어 SaTScan을 이용한 결과를 보여주고 장단점을 논하고자 한다. 미국 국립암센터의 SEER 프로그램에서 제공하는 미국의 각 카운티별 암 사망자 자료 중 2006년 여성 폐암 사망자 데이터를 예시 데이터로 사용하여 스캔 통계량을 이용하여 구한 클러스터 탐색 결과를 제시하고 비슷한 연구를 하고자는 연구자에게 도움을 주고자 한다.

주요용어: 공간 또는 시공간 데이터, 상대 위험률, 스캔 통계량, 암 클러스터, SaTScan, SEER 프로그램.

1. 서론

역학 (epidemiology)이나 공공의료 (public health), 질병관리 (disease administration) 등의 의학 관련 분야뿐만 아니라, 사회학, 범죄학 등 다양한 분야에서 지역별 또는 연단위, 월단위의 일정한 시기에 따라 주기적으로 수집이 되는 공간 (spatial), 시간 (temporal), 그리고 시공간 (spatio-temporal) 데이터들을 다루어야 되는 경우가 많다.

이러한 여러 가지 지표들이 공간 또는 시간 데이터인 경우 소지역 추정 (small area estimation) 방법을 이용하여 이웃관계를 고려한 추정값을 산출하는 방식이 많이 사용되고 있다 (Ghosh와 Rao, 1994; Pfeffermann, 2002; Chandra 등, 2007). 주로 베이지안 접근법을 이용하여 유의하다고 판단되는 이웃지역간의 공간상관성 (spatial correlation)을 변량효과 (random effect)로 간주하여 모형을 적합하는 방식이 이용된다 (Banerjee 등, 2004; Lee와 Park, 2015). 종종 지역 간 연관성을 고려한 추정치를 질병지도 (disease mapping)로 나타내어 시각화하기도 한다 (Lawson, 2013; Ahn 등, 2015; Coly 등, 2015).

이런 공간상관성을 고려한 추정량의 계산과 질병지도를 작성하여 데이터의 공간적, 시간적 특성을 이해하려는 접근법 외에 역학적인 요구나 공공정책 수립을 위한 의사결정을 보조하는 수단으로 다른 지역에 비하여 유난히 높은 위험률을 보이는 지역을 찾아내고자하는 경우가 있을 수 있다. 다른 지역에 비하여 높은 위험률을 보이는 지역 (또는 서로 이웃한 지역의 모임)을 클러스터 (cluster) 또는 핫스팟 (hot

† 2016년도 강원대학교 학술연구조성비로 연구하였음.
[1] (50612) 경상남도 양산시 물금읍 금오로 20, 양산부산대학교병원 의학통계실, 조교수.
[2] 교신저자: (24341) 강원도 춘천시 강원대학길 1, 강원대학교 정보통계학과, 조교수.
Email: mlee@kangwon.ac.kr

spot)이라고 부른다. 필요에 의해서는 다른 지역에 비하여 유난히 낮은 위험률을 보이는 지역도 관심의 대상이 되기도 하지만 연구자들은 주로 비교적 높은 위험률을 보이는 지역에 관심을 가지게 된다. 본 논문에서 예시로 사용한 암 데이터의 경우, 다른 지역에 비해 더 높은 발병률이나 사망률을 보이는 암 클러스터 (cancer cluster)를 찾는 것이 주 관심이다.

가장 많이 알려져 있고 흔히 사용되는 공간패턴을 찾는 방법 중에는 Moran's I (Moran, 1950)나 Geary's c (Geary, 1954) 통계량을 이용하는 방법이 있다. 그러나, 이런 방법들은 클러스터의 존재유무 (정확히는 클러스터링 패턴의 존재 유무)에 대한 검정은 가능하지만, 그 클러스터의 위치나 크기에 대한 정보를 주기에는 부족하다. 즉, 실용성이 있는 공간 또는 시공간 클러스터를 탐색하는 방법은 클러스터의 위치와 크기를 알려줄 뿐만 아니라 찾아진 클러스터가 통계적으로 유의한지를 판단할 수 있는 기준 (가령 P 값)까지 제공해 주어야 한다.

그러한 클러스터 탐색 방법 중에는 공간 점과정 방법 (spatial point process summary methods), 가장 가까운 이웃 방법 (nearest neighbor method), 그리고 스캔 방법 (local rate scanning method) 등이 있고 찾아진 클러스터의 유의성을 판단하는 방법 또한 다양하다 (Waller와 Jacquez, 1995). 이런 다양한 방법들 중 하나의 최적의 방법이 있다기보다는 주어진 데이터와 연구 목적에 따라서 더 나은 결과를 제공하는 방법이 다를 수 있다고 알려져 있다 (Wheeler, 2007).

본 논문에서는 이 중 Kulldorff (1997)의 스캔 통계량 (scan statistic)을 이용하여 공간 클러스터를 찾는 방법을 소개하고 미국국립암센터 (National Cancer Institute)의 SEER (Surveillance, Epidemiology, and End Results) 프로그램에서 제공하는 2006년 미국의 여성 폐암 사망자 데이터를 사용하여 클러스터를 찾는 예시를 보여주고자 한다. Kulldorff의 스캔 통계량을 이용한 클러스터 탐색 방법은 무료 소프트웨어인 SaTSan (Kulldorff, 2016)으로 구현되어 있고 본 논문에서도 이 프로그램을 사용하여 2006년 미국의 여성 폐암 사망자 데이터를 분석하였다.

논문의 나머지 부분은 다음과 같이 구성되어 있다. 2절에서는 스캔 통계량의 이론적인 원리와 SaTScan 프로그램에서 구현된 스캔 통계량이 어떻게 클러스터를 찾고 또 찾아진 클러스터의 통계적 유의성을 어떻게 검정하는지에 대해 설명한다. 3절에서는 예시로 사용된 SEER 데이터에 대한 소개와 그 분석 결과를 보여주고 마지막으로 4절에서는 결론을 제시하며 마무리한다.

2. 스캔 통계량 (Scan statistic)

2.1. 윈도우의 모양과 크기

앞서 설명한대로 공간적, 시간적, 시공간적 클러스터를 탐색하는 여러 가지 방법 중 하나인 스캔 통계량은 직관적인 원리와 발견된 클러스터에 대한 통계적 유의성을 검정하는 방법을 제공하는 이유로 최근 많은 분야에서 널리 사용이 되고 있다.

스캔 통계량은 이름 그대로 사전에 정의된 모양 (predefined shape)의 윈도우 (window) (Figure 2.1 참고) 를 크기 (size)와 위치 (center)를 계속 바꾸어 가며 관심 지역 전체를 훑어서 모든 가능한 윈도우에 대해, 윈도우 내의 위험도와 윈도우 밖의 위험도의 상대적 비율 (상대위험률; relative risk; RR)을 계산해가는 방식으로 다른 지역에 비해서 비교적 높은 위험률을 보이는 클러스터를 찾아낸다.

스캔 통계량의 접근법은 클러스터로 의심되는 지역의 모양이나 크기에 대한 정보가 없는 상태에서도 사용이 가능하다는 장점이 있지만, 초기 윈도우의 크기와 모양을 어떻게 정의하느냐에 따라 찾아진 클러스터의 위치나 크기가 변할 수 있다는 단점이 있다.

이론적으로는 어떠한 모양의 윈도우도 가능하고 또한 여러 모양들 중 어떤 모양이 클러스터를 찾는데 더 적절한지에 대한 다양한 연구도 있다 (Tango와 Takahashi, 2005; Patil과 Taillie, 2004). 가장 많이 가정되는 모양은 원형 (circular)과 타원형 (elliptic) (Figure 2.1 참고)이고 타원형의 윈도우를 사용

하는 것이 찾아진 클러스터를 검정할 때 더 높은 통계적 검정력을 보인다고 알려져 있다 (Kulldorff 등, 2006). 따라서 본 논문에서도 타원형의 윈도우를 사용하여 클러스터를 탐색하였다.

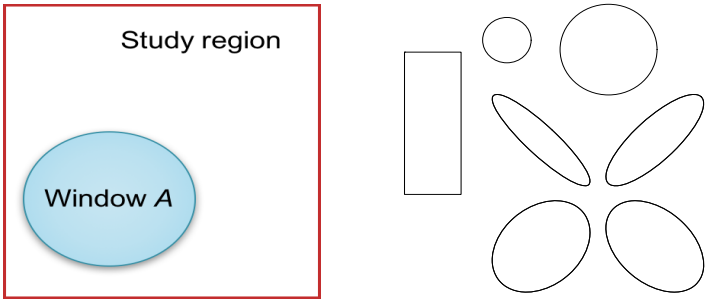


Figure 2.1 Window and variable shapes of windows

크기의 경우는 실제 지리적인 크기 (가령 원형 윈도우의 경우 반경 50km 등)를 이용하기도 하지만, 대부분의 경우는 윈도우 내의 인구수가 연구 영역 전체 인구수에서 차지하는 비율로 정의를 한다. 이렇게 정의하는 이유는 단순히 지리적인 크기로 반경이 고정된 윈도우를 사용할 경우 윈도우 내의 인구수가 많다는 이유로 대도시 근처만 클러스터로 찾아지는 경우를 피할 수 있기 때문이다.

인구수의 비율로 윈도우의 크기를 정하는 경우 최대 허용 윈도우의 크기는 50%보다 작아야 한다. 즉, 가장 큰 클러스터는 전체 연구 영역 인구의 절반까지 담을 수 있고 50%을 넘는다면 그 여집합 (complementary set)인 부분이 클러스터로 간주될 수 있다. 따라서, 이론적으로 후보 클러스터가 될 윈도우의 크기는 최대 50%까지 허용되지만, 실제로 이렇게 큰 윈도우를 가정하는 경우 연구자의 원래 의도와는 달리 대부분의 경우 한 두 개의 큰 클러스터 (global cluster)만 탐색이 되고 연구자가 관심이 있고 의심의 여지가 있는 소지역의 클러스터 (local cluster)는 탐색이 되지 못하는 경우가 종종 있다.

이러한 최대 허용 윈도우 크기의 선택은 통계적인 문제라기보다는 연구목적이나 연구자의 관심에 따라서 의도적으로 정하는 것으로, 연구자가 연구 영역 전체에 걸쳐서 큰 규모의 클러스터 (global cluster)를 찾는데 관심이 있다면 최대 허용 윈도우의 크기를 늘리고, 소규모의 지역 클러스터 (local cluster)를 찾는 것이 더 중요하다면 최대 허용 윈도우의 크기를 줄이면 된다. 스캔 통계량을 이용하여 공간 데이터를 분석한 많은 논문들이 각각의 연구 목적에 따라 다양한 크기의 최대 허용 윈도우 크기를 사용하였다. 최대 허용 윈도우 크기를 어떻게 정하는 가에 따라 찾아진 클러스터 결과들이 달라지며 그 결과들 중 어떤 것이 실제로 더 적절한지에 대해 객관적인 측도를 제공하려는 연구는 계속되고 있다 (Han 등, 2016).

따라서 본 논문에서는 다른 모양에 비해 높은 검정력을 보이는 타원형으로 윈도우 모양을 고정하고 최대 허용 윈도우 크기를 다양하게 변화시켜가며 찾아진 클러스터 결과가 어떻게 다른 지를 보여주고자 한다. 이 결과들 중 어떤 것이 더 적절한지는 암 전문가나 역학 연구 전문가 또는 정책 입안자가 판단하는 것으로 남겨두기로 한다.

2.2. 스캔 통계량을 이용한 클러스터 탐색의 원리

앞 절에서 설명한대로 윈도우는 다양한 모양과 크기를 가질 수 있지만, 스캔 통계량을 이용하여 클러스터를 찾는 기본 원리는 동일하다. 즉, 윈도우 밖의 위험도와 윈도우 안의 위험도를 비교하는 과정을 반복하는 것으로 좀 더 자세한 원리는 아래와 같다.

귀무가설 [H_0: 모든 연구 영역에서 위험도는 같다 (즉, 상대위험도 (RR) = 1)]과 대립가설 [H_1: 윈도우 A내의 위험도가 더 높다 (또는 낮다)]에 대해 스캔 통계량을 이용하여 가설검정을 하는 것이 기본 원리이다. 즉, 귀무가설과 대립가설 하에서 우도 (likelihood)를 계산하고 이 둘을 비교하는 우도비 (likelihood ratio; LR) 검정을 이용하여 통계적으로 유의한 클러스터를 찾아낸다.

이론적으로는 어떤 유형의 데이터에 대해서도 가능하지만 (Huang 등, 2007; Huang 등, 2009), 본 논문에서는 예시에서 사용되는 암 사망률 데이터와 같이 각 지역의 인구수와 암 사망자수가 주어진 카운트 (count) 데이터의 경우와 같이 포아송 모형을 가정하여 우도비를 구하는 경우만 설명하고자 한다.

모든 가능한 윈도우의 집합을 Z라고 하고 이 집합의 임의의 원소는 윈도우가 될 수 있고 z 라고 표기하자. 이때, z는 지리적으로 이웃한 (공통의 경계를 가진) 지역 단위의 모임으로 다양한 모양과 크기를 가질 수 있다. 지역 단위의 경우 미국은 카운티, 우리나라는 시군구 등의 행정구역 단위가 주로 사용된다.

이때, c_z와 n_z를 윈도우 z내의 실제 관측된 케이스 수 (가령 암 사망자수)와 기대되는 케이스 수 (또는 인구수)라 하고 연구 영역내의 모든 관측된 케이스의 합과 모든 기대되는 케이스 합을 $C = \sum_z c_z$, $N = \sum_z n_z$이라 하면, 포아송 모형을 가정할 경우 윈도우 z 에 대한 우도비 (LR)는 아래와 같이 주어진다.

$$LR(z) = \left\{ \left(\frac{c_z}{n_z} \right)^{c_z} \left(\frac{C - c_z}{N - n_z} \right)^{C - c_z} \right\} I(c_z > n_z).$$

이때, n_z가 윈도우 z내의 기대되는 케이스 수라면, C=N 임을 쉽게 알 수 있다. 대부분의 경우 윈도우 내의 관측된 케이스 수보다 기대되는 케이스 수가 더 많은 즉, 위험도가 더 높은 경우의 클러스터를 찾는 것이 주 관심이고 이는 식의 뒷부분에 있는 지표 함수 (indicator function), $I(c_z > n_z)$로 표현되고 있다. 즉, 다른 지역보다 낮은 위험도를 보이는 클러스터를 찾는 것이 관심이라면 이 지표 함수를 $I(c_z < n_z)$로 대체하거나, 또는 위험도가 높고 낮음에 상관없이 모든 클러스터를 찾고 싶다면, 지표 함수 부분을 제거한 우도비를 이용하면 된다. 많은 경우 우도비에 로그 취한 로그 우도비 (log likelihood ratio; LLR)를 이용하듯이 스캔 통계량의 경우도 마찬가지로 $LLR(z) = \log(LR(z))$를 흔히 이용한다.

이제, Z내의 모든 윈도우 z에 대해 우도비의 최대값 $T = \max_z LLR(z)$를 구하면 그 때의 윈도우 z가 가장 유의한 클러스터 (primary cluster)가 되고 이 클러스터를 제외하고 즉, 이 클러스터와 겹치지 않는 (not overlapping) 클러스터들 중 다음으로 유의한 클러스터들 (secondary clusters)을 비슷한 방법으로 찾을 수 있다.

2.3. SaTScan에서의 검정방법

앞 절에서 설명한 스캔 통계량의 기본 원리를 이용하여 찾아진 클러스터의 통계적 유의성을 검정하는 방법에는 여러 가지가 있다 (Waller와 Jacquez, 1995).

이 중 Kulldorff의 SaTScan은 각 지역 단위를 중심으로 점점 크기를 증가시켜 서로 겹치는 윈도우를 계층적으로 만들어가며 모든 윈도우에 대해 로그 우도비 $LLR(z)$를 계산하고 그 최대값인 검정통계량 $T = \max_z LLR(z)$를 구한다. 스캔 통계량은 closed form이 존재하지 않기 때문에 몬테카를로 (Monte Carlo) 접근법을 사용하여 유의성 검정을 한다.

클러스터가 존재하지 않는다 (H_0: 모든 연구 영역에서 위험도는 같다)는 가정하에 많은 수의 데이터 셋을 임의로 생성한 뒤, 각 데이터 셋에 대한 검정통계량 T를 계산하여 실제 데이터로부터 계산된 검정통계량의 값의 순위를 확인하여 유의성 여부를 판단한다. 예를 들어, m번의 시뮬레이션을 했을 경우 실제 데이터로부터 계산된 검정통계량의 순위가 r이라면, 몬테카를로 P값 (Monte Carlo P-value)은 $\frac{r}{1+m}$로 계산된다. 유의수준을 5%로 정한다면 몬테카를로 시뮬레이션으로 생성된 데이터 셋에서 계산

된 검정통계량의 값과 비교하여 실제 데이터에서 계산된 값이 상위 5%내에 든다면, 찾아진 클러스터는 유의하다고 판단한다. SaTScan은 기본적으로 999번의 시뮬레이션을 시행하므로 발견된 클러스터로부터 계산된 검정통계량의 값이 상위 50%내로 든다면 이 클러스터가 유의하다고 판단한다.

3. 데이터 분석

3.1. 미국 국립암센터의 SEER 프로그램

미국의 국립암센터의 SEER 프로그램은 미국 내의 암 관련 통계 정보를 제공한다 (NCI, 2016). SEER 프로그램은 암 관련 핵심적인 요약정보를 얻고자 하는 일반대중 뿐만 아니라 암 생존율 (cancer survival rates)이나 경쟁 위험 (competing risks) 연구 등 암 관련 역학에 대해 좀 더 과학적인 이해를 하고자는 연구들에게도 아주 중요한 데이터를 제공하고 있다. 일정한 기간을 거친 뒤에는 일반대중에게 공개되어 누구나 데이터를 얻을 수 있다.

SEER 데이터는 Table 3.1에서 보는 것과 같이 폐암 (lung cancer), 유방암 (breast cancer) 등 소위 말하는 주요 암 (common cancer)을 비롯하여 방광암 (bladder cancer), 자궁경부암 (cervical cancer)와 같은 드문 암 (rare cancer)에 이르기까지 현재까지 보고되는 대부분의 암종에 대한 사망자와 발생자 정보를 제공하고 있다.

Table 3.1 Categorization of cancer based on total number of cases in 2006, USA

Cancer Site	Total number of deaths in 2006	Class
Lung, Male	88,791	Common
Lung, Female	69,037	Common
Breast	40,600	Common
Prostate	28,256	Medium
Ovary	14,781	Medium
Bladder, Male	9,368	Rare
Bladder, Female	4,049	Rare
Cervical	3,953	Rare

특히, 암 사망자와 같은 인구수와 관련된 (population based) 카운트 데이터의 경우 실제 매년 관측되는 사망자수에 따라 common, medium, rare의 세 가지 클래스로 분류해서 관리하는 것으로 알려져 있다. 이는 자료 분석 시 너무 적은 케이스 수로 모형을 적합할 경우 생길 수 있는 불안정한 추정 등의 문제들에 대한 기준을 제시하는 것으로 알려져 있다. 가령, 드문 암의 경우 3∼5년치의 케이스 수를 합쳐서 추정값을 구하는 것을 권장하고 있다 (NCI, 2016).

3.2. 암 클러스터 탐색 예시

이 절에서는 SEER 프로그램에서 제공하는 미국의 각 카운티 (county)별 암 사망자 데이터 중 2006년 미국의 여성 폐암 사망자 데이터에 스캔 통계량을 이용하여 통계적으로 유의한 클러스터를 찾은 결과를 예시로 보여주고 결과의 해석과 주의점에 대해 논의하고자 한다.

먼저, Figure 3.1에서는 미국에서 각 카운티별로 2006년에 여성의 폐암 사망률의 상대위험률 (relative risk)을 색지도 (choropleth map)로 나타낸 것이다. 여기서, 상대위험률은 그 카운티내의 폐암 사망위험률과 그 카운티를 제외한 모든 지역의 폐암 사망위험률의 상대적인 비율을 의미한다. 상대위험률이 1이라는 것은 그 카운티의 폐암사망률이 미국 전체의 폐암사망률과 다르지 않다는 의미이고, 클러스터가 아니라는 뜻이다. 상대위험률이 1보다 크다는 것은 일단 그 지역이 상대적으로 더 위험하다는 의미이긴 하나, 통상적으로는 상대위험률이 1.2보다 큰 경우에만 주의를 기울이도록 한다 (NCI, 2006).

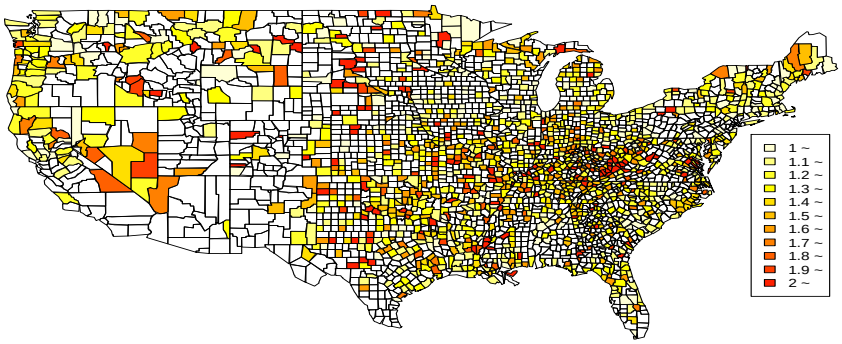
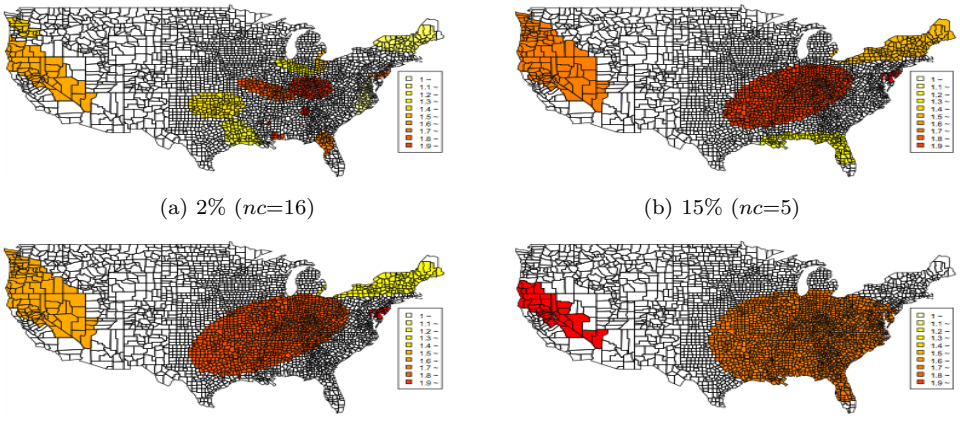


Figure 3.1 Relative risks of U.S. female lung cancer mortalities by counties, 2006

Figure 3.1과 같이 상대위험률을 색지도로 표현하였을 때, 인접한 지역이 비슷한 색, 즉, 비슷한 상대위험률을 가지는 듯 보인다. 그러나, 그러한 패턴이 실제로 클러스터를 이루는지, 또한 그 클러스터의 정확한 위치와 모양, 크기를 단지 색지도로 알아내기는 힘들다. 더구나, 그러한 클러스터에 대한 통계적 유의성을 검정하는 것은 더 힘든 일이다.

따라서, Figure 3.1에서와 같은 패턴을 보이는 2006년 미국 여성의 폐암 사망률이 실제로 어떤 공간적인 클러스터를 형성하는지를 탐색하고자 2절에서 설명한 스캔 통계량을 이용한 결과가 Figure 3.2에 제시되었다. SaTScan의 기본 설정을 이용하여 포아송 모형을 적합하였다. 즉, 각 카운티별 여성 인구수와 폐암 사망자수가 데이터 입력값이 되고 공간정보는 각 카운티의 중앙점의 좌표가 제공이 되었다. 이렇게 주어진 중앙 좌표를 기준으로 각 점들에 대해 미리 정한 최대 허용 윈도우 크기까지 윈도우의 크기를 점진적으로 키워가며 모든 윈도우에 대해 우도비를 계산하고 그 최대값인 최대우도비값을 구하여 몬테카를로 시뮬레이션을 통해 몬테카를로 P값을 구한다. 기본적으로 SaTScan은 가능한 모든 클러스터를 찾아서 보여주므로 결과 문서를 확인하여 P값이 예를 들어 5%보다 작은 클러스터만 유의한 것으로 판단해야한다.

SaTScan은 탐색된 클러스터의 상대위험률, 중심점의 위치, 크기 (전체 인구수 대비 클러스터 내의 인구수 비율)를 비롯하여 클러스터를 구성하는 지역 (이 경우 카운티)의 아이디를 텍스트 문서의 형태로 결과를 제공한다. 즉, Figure 3.2와 같은 색지도를 만들기 위해서는 외부 프로그램을 이용해야만 한다. 본 논문에서는 R 프로그램을 이용하였다.



(a) 2% (nc=16) (b) 15% (nc=5)

(c) 25% (nc=4) (d) 50% (nc=2)

Figure 3.2 Clusters detected with varying maximum window size allowed from 2% in (a) to 50% in (d). The numbers in the parenthesis are number of significant clusters detected for each case.

앞서 2절에서 스캔통계량을 사용할 때 유의점 내지 한계점으로 지적이 되었던 문제도 Figure 3.2를 통해 확인할 수 있다. 즉, (a) 2%에서 (d) 50%에 이르기까지 최대 허용 윈도우의 크기에 따라 최종적으로 찾아진 유의한 클러스터들의 크기, 모양, 위치 등이 달라짐을 알 수 있다. 이와 같은 경우 (a)∼(d) 중 어느 결과가 데이터의 실제 공간 클러스터 패턴을 잘 설명하는지에 대한 판단을 하기는 힘들다. 다만, 여러 가지 다른 관점들, 가령 역학적 의미라든지, 정책적인 의미 등을 고려하여 연구자의 목적에 부합하는 결과를 제공할 수 있는 최대 허용 윈도우 크기를 선택하는 것이 옳을 것이다. 예를 들면, 소규모 지역 단위로 상대위험률이 높은 클러스터를 찾는 것이 목적이라면 2%와 같이 작은 윈도우를 선택하고, 반대로 연구 영역 전체에 걸친 어떤 패턴을 찾고자 하는 것이 목적이라면 이론적으로 허용 가능한 50%의 큰 윈도우로 스캔통계량을 계산하면 될 것이다. 한 가지 주의할 점은 대체적으로 클러스터 크기가 클수록 클러스터 자체의 상대위험도는 낮고, 클러스터 크기가 작을수록 그 클러스터의 상대위험도가 높음을 알 수 있는데, 이는 Figure 3.2에서 같은 중심을 가진 클러스터들이 크기가 클수록 색이 옅어지는 것으로 확인 가능하다. 다만, 아주 높은 상대위험도를 가지는 작은 크기의 클러스터들이 합쳐지면서 그 중심이 달라지는 경우가 발생할 수 있으며, 이러한 경우 클러스터의 크기가 커짐에도 불구하고 색이 옅어지지 않는 결과가 생길 수도 있으니 결과 해석시 주의가 필요하다.

위의 결과는 미국 국립암센터의 암 전문가들이 북부 캘리포니아와 네바다주를 비롯하여 몇몇 지역들은 여성 히스패닉 인구가 급격히 늘어난 지역들이고 이들 히스패닉 여성들의 흡연율이 비교적 높은 편이라는 점을 알아내면서, 찾아진 클러스터들이 실제로도 역학적으로 의미가 있는 결과임을 보여준다.

4. 결론 및 제언

본 연구에서는 공간 또는 시공간 데이터에서 다른 지역이나 시간에 비해 위험률이 유난히 높은 지역이나 시간인 클러스터를 찾는 방법 중 하나인 스캔 통계량을 소개하였다. 또한, 비교적 덜 알려진 무료 소프트웨어 SaTScan을 국내에 소개하여 미국국립암센터의 SEER 프로그램을 통해 제공되는 미국 폐암 사망자 데이터를 분석한 결과를 예시로 보여주고 해석에서의 유의점에 대해서도 언급을 하였다.

스캔 통계량은 그 기본적인 원리와 이론적 배경이 매우 직관적이라서 많은 분야에서 그 활용성이 점점 높아지고 있다. 다만, 그 원리상 많은 컴퓨팅 자원 (computing resources)을 요구하지만, SaTScan은 이를 효과적으로 다룰 수 있는 알고리즘을 구현하여 이러한 단점을 많이 극복하였다 (Kulldorff, 2016).

스캔 통계량을 구하는 과정에서 선택하는 윈도우의 크기에 따라 최종적인 결과가 달라질 수 있는 점도 많은 연구를 통해서 해법이 찾아지고 있는 중이며 많은 경우에 따라서는 연구 목적에 따라 임의로 그 크기를 한정할 수도 있다는 점을 2006년 미국 여성 폐암 사망자 데이터를 분석한 결과를 통해 논의해보았다.

본 논문은 이미 많은 연구가 진행된 미국의 사례를 참고하여 우리나라의 암 등록 자료에 대해 클러스터 분석을 할 때 시행착오를 줄이고 어떤 점들에 유의하여 분석을 해야 하는지에 대한 도움을 주고자 하였다.

References

Ahn, D. S., Han, J. H., Yoon, T. H., Kim, C. H. and Noh, M. S. (2015). Small area estimations for disease mapping by using spatial model. *Journal of the Korean Data & Information Science Society*, **26**, 101-109.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*, CRC Press, New York.

Chandra, H., Salvati, N. and Chambers, R. (2007). Small area estimation for spatially correlated populations-a comparison of direct and indirect model-based methods. *Statistics in Transition*, **8**, 887-906.

Coly, S., Charras-Garrido, M., Abrial, D. and Yao-Lafourcade, A. (2015). Spatiotemporal disease mapping applied to infectious diseases. *Procedia Environmental Sciences*, **26**, 32-37.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, **5**, 115-145.

Ghosh, M. and Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, **9**, 55-76.

Han, J., Zhu L, Kulldorff, M., Hostovich, S., Stinchcomb, D., Tatalovich, Z., Lewis D. and Feuer, E. (2016). Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics. *International Journal of Health Geographics*, 15-27.

Huang, L., Kulldorff, M. and Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics*, **63**, 109-118.

Huang, L., Tiwari, R. C., Zhaohui, Z., Kulldorff, M. and Feuer, E. J. (2009). Weighted normal spatial scan statistic for heterogeneous population data. *Journal of the American Statistical Association*, **104**, 886-898.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1487-1496.

Kulldorff, M. (2016). *SaTScan user guide v9.4.4*, `http://www.satscan.org/`.

Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, **25**, 3929-3943.

Lawson, A. B. (2013). *Bayesian disease mapping: Hierarchical modeling in spatial epidemiology*, 2nd Ed., Chapman and Hall/CRC, New York.

Lee, W. and Park, C. (2015). Prediction of apartment prices per unit in Daegu-Gyeongbuk areas by spatial regression models. *Journal of the Korean Data & Information Science Society*, **26**, 561-568.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17-23.

NCI. (2016). Surveillance, Epidemiology, and End Results (SEER) Program, `www.seer.cancer.org`.

Patil, G. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, **11**, 183-197.

Pfeffermann, D. (2002). Small area estimation: New developments and directions. *International Statistical Review/Revue Internationale De Statistique*, **70**, 125-143.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4-11.

Waller, L. A. and Jacquez, G. M. (1995) Disease models implicit in statistical tests of disease clustering. *Epidemiology*, **6**, 584-590.

Wheeler, D. C. (2007). A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003, *International Journal of Health Geographics*. 6-13.

Journal of the Korean Data & Information Science Society
2016, **27**(5), 1193–1201

http://dx.doi.org/10.7465/jkdi.2016.27.5.1193
한국데이터정보과학회지

Cancer cluster detection using scan statistic[†]

Junhee Han[1] · Minjung Lee[2]

[1]Division of Biostatistics, Pusan National University Yangsan Hospital
[2]Department of Statistics, Kangwon National University

Received 28 August 2016, revised 20 September 2016, accepted 21 September 2016

Abstract

In epidemiology or etiology, we are often interested in identifying areas of elevated risk, so called, hot spot or cluster. Many existing clustering methods only tend to a result if there exists any clustering pattern in study area. Recently, however, lots of newly introduced clustering methods can identify the location, size, and shape of clusters and test if the clusters are statistically significant as well. In this paper, one of most commonly used clustering methods, scan statistic, and its implementation SaTScan software, which is freely available, will be introduced. To exemplify the usage of SaTScan software, we used cancer data from the SEER program of National Cancer Institute of U.S.A. We aimed to help researchers and practitioners, who are interested in spatial cluster detection, using female lung cancer mortality data of the SEER program.

Keywords: Cancer cluster, relative risks, SaTScan, scan statistic, spatial and spatio-temporal data.

[†] This study was supported by 2016 Research Grant from Kangwon National University.
[1] Assistant professor, Division of Biostatistics, Pusan National University Yangsan Hospital, Gyeongsangnam-do 50612, Korea.
[2] Corresponding author: Assistant professor, Department of Statistics, Kangwon National University, Gangwon-do 24341, Korea. Email: mlee@kangwon.ac.kr