대한지리학회지 제48권 제5호 2013(768~785)

카운트 데이터 기반 공간 군집 분석 연구의 동향과 방법론적 이슈

조대헌*

Trends and Methodological Issues in Spatial Cluster Analysis for Count Data

Daeheon Cho*

요약 : 행정구역과 같은 공간 단위로 합산된 카운트 데이터는 지리학 연구에 있어 가장 기본적인 데이터이다. 카운트 데이터를 대상으로 하는 공간 군집 분석 연구가 지속적으로 수행되어 왔으나 상대적으로 큰 관심을 받지 못하였을 뿐만 아니라 여러 분야에서 산발적으로 이루어지면서 그 흐름은 물론 주요한 성과와 과제를 제대로 파악하기가 어려운 실정이다. 이 연구에서는 최근 20여 년 동안 이루어진 카운트 데이터 기반의 공간 군집 분석 연구를 대상으로 동향과 방법론적 특성을 살펴본 후 이슈와 과제를 검토함으로써 지리학 연구에 시사하는 바를 파악하고자 한다. 지리학은 물론 보건이나 범죄 등의 영역에서 다양한 방법들이 사용되고 있는데, 이들은 그 목적이나 방법론적 특성이 비교적 뚜렷이 구분될 뿐만 아니라 통계학적 신뢰성과 관련된 이슈 또한 존재한다. 따라서 분석의 실행시 방법론에 대한 면밀한 검토가 필요하며, 향후 방법론과 관련된 실증 연구 및 분석 도구의 개발이 요구된다.

주요어 : 카운트 데이터, 비율 데이터, 공간 패턴, 공간 군집, 분산의 불안정성

Abstract : Count data aggregated into areal units such as administrative boundaries are the most important sources of information for geographic research. Despite of ongoing research on spatial cluster analysis of count data, it has received relatively little attention and besides, it is difficult to comprehend research trends as well as major outcomes and challenges. This study aims to review the research literature conducted during the last two decades, to examine methodological characteristics, and finally to discuss some issues and challenges. Methods for indentifying spatial clusters have been used in various fields including geography, criminology, and epidemiology. However, their methodological features are not only quite distinct from each other, but there are issues related to the statistical reliability. Therefore, these have to be taken into account carefully when particular methods are used, and further empirical research about methodological issues and the development of analysis tools is needed.

Key Words : count data, ratio data, spatial pattern, spatial cluster, instability of variance

이 논문은 2012년 서울대학교 교육종합연구원 연구소 지원금에 의하여 연구되었음.

* 서울대학교 지리교육과 강사(Lecturer, Department of Geography Education, Seoul National University), dhncho@gmail.com

1. 서론

지리적 객체나 사건의 분포는 지리학을 비롯하여 범죄나 보건 등 다양한 연구 분야의 가장 기본적인 관심사가 된다. 이와 관련하여 중요한 관심 주제 중의 하나는 그러한 객체나 사건이 특정 공간 상에 군집되어 분포하는지, 만일 그러하다면 그 영역은 어디인지를 살펴보는 것이다. 인구나 산업의 분포, 범죄나 교통사고의 분포, 질병의 분포 등에서 공간적 군집의 존재 여부는 학술적 측면뿐만 아니라 실용적 측면에서도 매우 중요하다. 예를 들어 산업이나 질병의 군집은 경제지리학이나 역학(epidemiology)의 관심사이기도 하지만 지역 개발이나 방역을 위한 대책 수립과 밀접히 연관된다.

이러한 객체나 사건과 관련된 정보는 특정 위치에서의 존재 유무로 기록되거나 행정구역과 같은 영역 단위로 합산된 카운트로 기록된다. 최근 디지털 공간 데이터 구축의 확산과 함께 개별 객체나 사건에 대한 데이터의 생산이 늘고 있지만 보안이나 사생활침해 등의 이유로 인해 일반적인 연구자가 접근하기는 쉽지 않다. 그래서 학술적인 것이든, 아니면 정책 수립과 같은 실용적인 것이든 인구센서스와 같은 카운트 데이터는 대부분의 관련 연구에서 가장 기본적인 자료원이 된다.

여기서 이슈가 되는 것은 데이터의 유형에 따라 취할 수 있는 공간 군집 분석의 방법이 상이하다는 것이며, 그래서 카운트 데이터에 대한 충분한 이해가 필요하다는 점이다. 예를 들어 어떤 사건이 공간적으로 군집을 이루는지를 평가하기 위해서는 카운트 데이터가 서로 비교 가능한 형태로 제시될 필요가 있다. 이때 한 지역의 카운트(예를 들어 출생아 수)가 그것을 만들어내는 기저 패턴(예를 들어 가임 여성 인구)과 밀접히 관련된다면 이를 고려한 데이터(예를 들어 출생률)를 사용하게 된다.

따라서 연구의 목적에 적절한 분석 방법이 요구되는데, 1990년 이후 지리학 등에서는 다양한 관련 방법론과 그에 기초한 경험 연구들이 본격적으로 수행되어 왔다. 카운트 데이터에 기초한 공간 분석은 1990년대 들어 급성장한 공간통계학(spatial statistics) 혹은 공간데이터분석(spatial data analysis)과 밀접하게 관련되어 있다. 하지만 그러한 발전은 대부분 개별 포인트 데이터나 구역 단위의 비율 척도(ratio scale) 데이터에 집중함으로써(Bailey and Gatrell, 1995) 카운트 데이터는 상대적으로 주목 받지 못하였으며, 발전의 속도 또한 뒤늦었다. 더불어 카운트 데이터에 대한 분석이 상당히 다양한 분야에서 수행되고 있어 주요한 흐름이나 방법론적 특성을 파악하기가 쉽지 않다.

이러한 맥락에서 본 연구는 지난 20여 년간 지리학 등 관련 분야들에서 공간 군집을 파악하기 위해 실제 이루어진 연구들을 되짚어 보고자 한다. 이를 통해 공간 군집의 파악을 위해 카운트 데이터 분석이 어떤 연구에 어떻게 사용되고 있는지 그 동향을 파악함과 동시에 방법론 상의 특성에 초점을 두어 이슈를 검토함으로써 관련 연구에 대한 시사점을 도출할 수 있을 것으로 기대된다. 결과적으로 본 연구의 세부 목적은 다음과 같다. 첫째 지난 20여 년간 수행된 카운트 데이터 기반의 공간 군집 분석 연구의 동향을 정리한다. 둘째, 방법론적인 특성에 기초하여 이들 연구에 사용된 분석 방법들을 유형화하고, 그 특성을 고찰한다. 셋째, 카운트 데이터 기반 공간 군집 분석의 주요 방법론적 이슈를 살펴보고, 관련 연구에 시사하는 바와 향후 과제를 도출한다.

2. 연구의 동향

전술한 바와 같이 지난 20여 년간 지리학 등에서 이루어진 카운트 데이터 기반의 공간 군집 분석 연구를 대상으로 그 동향을 살펴보고자 한다. 따라서 시기적으로는 1990년 무렵부터 최근까지의 연구를 대상으로 하되 전 분야의 전 연구를 다루기보다는 연구자의 판단에 의해 연구 방법 상의 주요한 흐름을 파악할 수 있도록 비교적 자주 사용되는 분석 방법에 초점을 두었다. 조금 더 부연하자면, 객체나 사건의 분포를 다루지만 포인트 패턴 분석과 같은 개별 단위 데이터 분석 방법은 논외로 하며, 카운트 기반이더라도 모델링

과 같이 분석 방법이 패턴 파악에서 벗어난 연구는 제외한 반면 기술적 혹은 탐색적으로 패턴을 파악하는 방법들은 포함하였다. 공간 군집을 분석하는 가장 중요한 목적 중의 하나가 구체적인 지리적 범위를 파악하는 것이므로 분석의 초점은 국지적 분석으로 제한하였다. 최종적으로 약 70편의 연구를 대상으로 결과를 정리하였다(표 1).

이 표로부터 파악할 수 있는 몇 가지 사실을 정리해 보면 다음과 같다. 우선 제한적이기는 하지만 상기의 연구들로부터 공간 군집의 연구 대상이 크게 도시 및 인구(인구 및 가구의 분포, 중심지 식별, 이동 및 통행 패턴), 경제, 범죄, 보건(사망 및 질병) 등으로 나타나고 있음을 알 수 있다. 보건과 관련된 연구가 상당히 활발한 것을 알 수 있는데, 이는 질병과 같은 보건 상의 이슈는 상대적으로 시급한 대책이 요구된다는 점에서 공간적 군집의 파악이 가장 기본적인 관심사가 되기 때문으로 보인다. 보다 전통적인 지리학 연구 중에서는 산업이나 경제 활동의 분포에서 군집을 파악하려는 연구가 많은 비중을 차지하고 있다.

분석의 방법과 관련하여서는 전반적으로 Local Moran's I_i나 Getis-Ord's G_i 및 G_i^*와 같은 국지적 공간통계와 공간스캔통계량(Spatial Scan Statistics) 등

표 1. 카운트 데이터에 기초한 공간 군집 분석 연구

분석 대상	연구자	분석 방법	핵심 변수	변수의 기본 유형
인구 및 가구의 분포	이상일, 2008	표준화상이점수(SSD_i)	학력집단별 전체인구 대비 특정 지역 인구 비중	비율(열비중)
	신상영, 2010	기술적(비율 및 비율의 변동계수)	일반가구 대비 1인가구 비율	비율(행비중)
	Scardaccione *et al.*, 2010	입지계수(LQ)	지역내 특정 종사자 비중과 전지역의 특정 종사자 비중	비율(행비중)
		Local Moran's I_i	인구대비 전입외국인 비율	비율(행비중)
	Lloyd, 2010	Local Moran's I_i	특정 집단 대비 준거 집단 비율의 로그	비율
	최열 등, 2012	기술적(비율)	지역 내 1인가구 비중	비율(행비중)
중심지식별	Giuliano and Small, 1991	기술적(밀도)	고용밀도	비율(밀도)
	Forstall and Greenre, 1997	기술적(비율)	종사자(관심집단) 수와 거주자(준거집단) 수	비율
	Coffey and Shearmur, 2001	기술적(밀도)	고용밀도	비율
		기술적(비율)	종사자(관심집단) 수와 거주자(준거집단) 수	비율
	Anderson and Bogart, 2001	기술적(밀도)	구역 고용밀도	비율(밀도)
	Riguelle *et al.*, 2007	Local Moran's I_i	고용밀도	비율(밀도)
	이주형 · 선권수, 2009	표준점수(z-score)	밀도	비율(밀도)
	이상일 등, 2010	수정 AMOEBA	인구밀도	비율(밀도)
	이상일 · 조대헌, 2012	기술적(밀도: 개별 단위 및 구역)	인구밀도	비율(밀도)
	조대헌 · 이종일, 2013	표준화상이점수(개별단위 및 구역)	전 지역의 고용자 수(및 거주자 수) 대비 대상 지역의 고용자 수(및 거주자 수) 비중	비율(열비중)
		수정 AMOEBA	표준화상이점수	비율(열비중)
이동 및 통행 패턴	Berglund and Karlström, 1999	G_{ij}	이동자수의 잔차(=관찰빈도−추정빈도)	빈도
	Bivand *et al.*, 2009	Local Moran's I_i	순이동률의 잔차	비율(행비중)
	Huang *et al.*, 2009	공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	김영호, 2010	G_{ij}	이동자 수	빈도

분석 대상	연구자	분석 방법	핵심 변수	변수의 기본 유형
경제 활동의 분포	Feser *et al.*, 2005	Getis-Ord's G_i	고용자 수	빈도
		Local Moran's I_i	입지계수(LQ)	비율(행비중)
	Carroll *et al.*, 2008	입지계수(LQ)	지역내 특정 종사자 비중과 전지역의 특정 종사자 비중	비율(행비중)
		Getis-Ord's G_i^*	종사자 수	빈도
	Fernhaber *et al.*, 2008	입지계수(LQ)	지역내 특정 종사자 비중과 전지역의 특정 종사자 비중	비율(행비중)
	신우진 · 신우화, 2009	Getis-Ord's G_i	업체 수	빈도
	김병선 · 김걸, 2009	Local Moran's I_i	업체 수	빈도
	Han and Qin, 2009	Local Moran's I_i	업체 밀도	비율(밀도)
	손승호, 2010	기술적(비율)	지역내 특정 사업체 수 비중과 전지역의 특정 사업체 수	비율(행비중)
		기술적(비율)	총종사자 대비 대상 지역 종사자 비중의 표준점수 (z-score)	비율(열비중)
	Duque *et al.*, 2011	AMOEBA	지역내 특정 종사자 비중	비율(행비중)
	Cromley and Hanink, 2012	입지계수(LQ), 공간적 입지계수(FLQ)	지역(혹은 근린)내 특정 종사자 비중과 전지역 특정 종사자 비중	비율(행비중)
		Getis-Ord's G_i^*	지역내 특정 종사자 비중	비율(행비중)
	이경주 · 권일, 2012a	포아송확률	제조업체 수	빈도
	Liu, 2013	공간적 입지계수(FLQ)	지역(혹은 근린)내 특정 종사자 비중과 전지역 특정 종사자 비중	비율(행비중)
		Local Moran's I_i	입지계수(LQ)	비율(행비중)
범죄 발생의 분포	Ratcliffe and McCullagh, 1999	Getis-Ord's G_i	범죄 발생 밀도	비율(밀도)
	Messner *et al.*, 1999	Local Moran's I_i	범죄율	비율(행비중)
	Craglia *et al.*, 2000	Getis-Ord's G_i	사건 수	빈도
		표준화 범죄율	관찰 사건 수와 기대 사건 수	비율(행비중)
		Local Moran's I_i	범죄율	비율(행비중)
	Zhang and Lin, 2008	$I_{DR,i}$(수정 Local Moran's I_i)	범죄 발생 건수의 편차 잔차(deviance residual)	빈도(기대 빈도와의 잔차)
		$I_{EBI,i}$(수정 Local Moran's I_i)	범죄율의 경험베이즈지수	비율(행비중)(비율의 경험 베이즈 표준 점수)
		$I_{r,i}$ (Local Moran's I_i)	범죄율	비율(행비중)
사망률의 분포	Kulldorff, 1997	공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Hanson and Wieczorek, 2002	공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
		Local Moran's I_i	연령 표준화 사망률의 로그	비율(행비중)
	Bura *et al.*, 2002	Local Moran's I_i	연령 표준화 사망률	비율(행비중)
		Getis-Ord's G_i	연령 표준화 사망률	비율(행비중)
	Boscoe *et al.*, 2003	공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
		RR(relative risk)	관찰 사건 수와 기대 사건 수	비율(행비중)
	Goovaerts and Jacquez, 2004	Local Moran's I_i	표준화사망률의 표준점수(z-score)	비율(행비중)
	Goovaerts and Jacquez, 2005	Local Moran's I_i	사망률	비율(행비중)

분석 대상	연구자	분석 방법	핵심 변수	변수의 기본 유형
사망률의 분포	McLaughlin and Boscoe, 2007	Local Moran's I_i	표준화사망률	비율(행비중)
		Local Moran's I_i	Spatial empirical Bayes smoothed SMR	비율(행비중)
	Zhang and Lin, 2009	수정 공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Rogerson and Yamada, 2009	Fuchs and Kenett's M	관찰 사건 수와 기대 사건 수	빈도 혹은 비율(열비중)
	Cromley and Hanink, 2012	입지계수(LQ, 공간적 입지계수(FLQ)	지역(혹은 근린)내 특정 집단 비중과 전지역 특정 집단 비중	비율(행비중)
		Getis-Ord's G_i^*	지역내 특정 집단 비중	비율(행비중)
질병 발생의 분포	Besag and Newell, 1991	Besag-Newell's R	사건 수	빈도
	Tango, 1995	Tango's index(C_F)	총 사건 수 대비 관찰사건 및 기대사건의 비중	비율(열비중)
	Ord and Getis, 1995	Getis-Ord's G_i	발병률	비율(행비중)
	Rogerson, 1999	Rogerson's R_i	전체 사건 수 대비 관찰 사건 수 비중과 전체 위험 인구 수 대비 지역 인구 수 비중	비율(열비중)
	Tango, 2000	Tango's index(MEET)	총 사건 수 대비 관찰사건 및 기대사건의 비중	비율(열비중)
	Gangnon and Clayton, 2001	공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Kulldorff *et al.*, 2003	공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Jacquez and Greiling, 2003	Local Moran's I_i	표준화사망률	비율(행비중)
	Thomas and Carlin, 2003	Spatially smoothed rates(Baysian method)	진단율	비율(행비중)
		공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Wang, 2004	Rogerson's R_i	전체 시건 수 대비 관찰 사건 수 비중과 전체 위험 인구 수 대비 지역 인구 수 비중	비율(열비중)
	Johnson, 2004	Smoothed SIR(Bayse model)	관찰 사건 수와 기대 사건 수	비율(열비중)
		공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Waller and Gotway, 2004	Local Moran's I_i	사건 수 혹은 발병률	빈도/비율(행비중)
		$I_{cr,i}$(수정 Local Moran's I_i)	관찰 사건 수와 기대 사건 수	비율(열비중)
	Trevelyan *et al.*, 2005	Getis-Ord's G_i	발병률	비율(행비중)
	Rogerson, 2005	U_i	관찰 사건 수와 기대 사건 수	비율(열비중)
	Tango and Takahashi, 2005	공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Hinman *et al.*, 2006	Getis-Ord's G_i^*	사건 수	빈도
	Aldstadt and Getis, 2006	AMOEBA(Getis-Ord's G_i^*)	출산율	비율(행비중)
	Nødtved *et al.*, 2007	Local Moran's I_i	표준화사망률	비율(열비중)
		Smoothed SMR	관찰 사건 수와 기대 사건 수	비율(열비중)
		공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)

분석 대상	연구자	분석 방법	핵심 변수	변수의 기본 유형
질병 발생의 분포	Waller *et al.*, 2007	Tango's index	총 사건수 대비 관찰사건 및 기대사건의 비중	비율(열비중)
		공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Yeshiwondim *et al.*, 2009	Local Moran's I_i	사건 수	빈도
		Getis-Ord G_i^*	사건 수	빈도
	Jackson *et al.*, 2009	I_r(Local Moran's I_i)	관찰 사건 수와 기대 사건 수	비율(열비중)
		공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Rogerson and Yamada, 2009	Local Moran's I_i	사건 수 포아송 표준 점수	빈도(포아송 표준 점수)
	Yao *et al.*, 2011	공간스캔통계량(비정형)	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	Torabi and Rosychuk, 2011	Besag-Newell's R	사건 수	빈도
		공간스캔통계량(원형, 비정형)	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
		Tango's index(MEET)	총 사건 수 대비 관찰사건 및 기대사건의 비중	비율(열비중)
		RR(relative risk)(Bayesian disease mapping)	관찰 사건 수와 기대 사건 수	비율(열비중)
	박선일 · 배선학, 2012	공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)
	이경주 · 권일, 2012b	공간스캔통계량	구역내 관찰 사건 수 비중과 기대 사건 수 비중	비율(행비중)

이 가장 빈번히 사용되고 있다. 보다 전통적인 지리학 연구 분야와 그렇지 않은 범죄 및 보건 분야 간에 다소 간의 차이가 파악된다. 전통적인 지리학 연구의 기술 통계나, 입지계수(LQ: location quotient) 등 보다 기술적이고 탐색적인 분석이 이루어지면서도 동시에 최근의 공간통계학 혹은 공간데이터분석 방법을 적용하는 연구가 이루어져 다양한 방법이 사용되고 있다. 반면에 범죄 및 보건 분야의 경우 공간스캔통계량 등 특정 분석 방법이 거의 대부분을 차지하고 있다.

분석에 사용되는 변수의 형태와 관련하여서는 몇 가지 유형을 파악할 수 있다. 카운트 데이터 분석이기 때문에 기본적으로 변수는 빈도이거나 비율 데이터의 형태를 취한다. 하지만 비율의 경우 그 유형은 더욱 세분화되는데, 비율, 비율(밀도), 비율(행비중), 비율(열비중) 등이 그것이다. 일반적인 비율은 성비와 같이 일반적인 준거 집단 대비 관심 집단의 비를 의미하며, 비율(밀도)는 인구밀도나 고용밀도와 같이 비율의 형태이지만 밀도를 의미하는 변수를 나타낸다. 비율에서 행비중 비율은 한 지역에서 관심 집단이 차지하는 비중을, 열비중 비율은 전체 관심 집단 중에서 한 지역의 관심 집단이 차지하는 비중을 의미한다(이상일, 2007). 예를 들어 출생률이나 사망률과 같이 대부분의 인구통계나 보건지표가 행비중 비율에 해당한다면 전체 고학력집단 대비 특정 지역의 고학력 집단 비중은 열비중 비율에 해당한다. 표를 통해 알 수 있듯이 대부분의 분석은 비율 변수에 기초하고 있으며, 그 중에서도 행비중 비율 변수가 다수를 차지하는 것으로 나타났다.

변수가 사용되는 방식과 관련하여 한가지 더 파악할 수 있는 것은 동일한 분석 방법이라 할지라도 사용되는 변수의 유형이 상당히 다양하다는 점이다. 대표적으로 국지적 공간 패턴 분석에서 널리 사용되는 Local Moran's I_i를 들 수 있는데, 빈도(김병선 · 김걸, 2009), 비율(Lloyd, 2010), 비율(밀도)(Han and Qin, 2009), 비율(행비중)(Craglia *et al.*, 2000)과 이를 표준화하는 경우(Liu, 2013) 등 다양한 변수의 형태가 이용되고 있다. 경제활동을 예로 든다면 고용자 수나 고용밀도, 고용률, 특정 산업의 입지계수 등이 모두 Local Moran's I_i 분석에 사용되고 있다는 의미인데 이것과 관련된 이슈에 대해서는 나중에 보다 구체적으

로 다루기로 한다.

3. 분석 방법의 유형화와 특성

1) 선행연구에서의 분류

여러 분석 방법들에 대해 선행연구에서 시도된 분류 체계에 대해 먼저 살펴보고자 한다. 분석의 목적과 관련해서는 두 가지 차원의 분류가 흔히 사용되고 있는데(표 2), 이는 Besag and Newell(1991)의 연구에 힘입은바 크다. 그들은 공간 군집 분석을 그 목적에 따라 "클러스터링 테스트(tests of clustering)"과 "클러스터 탐지 테스트(tests for the detection of clusters)"로, "클러스터링 테스트"는 다시 "일반 테스트(general tests)"와 "초점 테스트(focused tests)"로 구분하고 있다. 클러스터링 테스트는 한 지역 혹은 그 이상의 지역에서 나타나는 군집이 나타나는지에 대한 것이며, 클러스터 탐지 테스트는 전체 지역에서 클러스터로 불릴 국지적 범위를 설정하는 것과 관련된다. 또한 일반 테스트는 전반적인 패턴이 군집을 보이는지에 관심이 있다면 초점 테스트는 특정 지역이 군집에 해당하는지에 관심이 있다. 하지만 일반 테스트와 초점 테스트는 각기 전체 패턴과 구체적인 위치를 다루고, 클러스터 탐지 테스트 또한 구체적인 범위를 다룬다는 점을 고려하면 일반 테스트와 초점 테스트가 하나의 공통된 범주 아래 위치하는 것이 보다 합리적인 것인지에 대해 문제 제기가 가능할 것으로 판단된다.

이런 면에서 Rogerson and Yamada(2009)는 Besag and Newell(1991)의 분류를 수용하면서도 일반 테스트와 초점 테스트를 하나의 범주로 묶지 않고 별개로 제시하고 있다. 이들은 이 범주와 함께 그와 연관된 구체적인 방안을 동시에 제시하는데, 클러스터링 테스트에 해당하는 일반 테스트는 전역적 테스트(global tests)로, 초점 테스트는 국지적 테스트(local tests)로, 클러스터 탐지 테스트는 국지적 테스트를 여러 위치에서 동시에 실행하는 테스트로 개념화하고 있다. 여기서 초점 테스트와 국지적 테스트는 동등한 의미로 사용되고 있다.

Waller and Gotway(2004)는 분석 방법이 크게 두 차원, 즉 클러스터 테스트와 클러스터링 테스트, 그

표 2. 분석의 목적에 따른 공간 군집 분석 방법의 분류

연구자	구분		비고
Besag and Newell (1991)	클러스터링 테스트	일반 테스트	전반적인 패턴이 군집의 경향성을 보이는가?
		초점 테스트	특정 지역(주변)이 군집에 해당하는가?
	클러스터 탐지 테스트		군집에 해당하는 구체적인 범위는 어디인가?
Rogerson and Yamada (2009)	일반 테스트		전반적인 패턴이 군집을 보이는가?: 전역적 테스트
	초점 테스트		특정 지역(주변)이 군집에 해당하는가?: 국지적 테스트
	클러스터 탐지 테스트		군집에 해당하는 구체적인 범위는 어디인가?: 국지적 테스트의 동시 실행
Waller and Gotway (2004)	클러스터링 테스트 – 클러스터 탐지 테스트	클러스터링 테스트	전반적인 패턴이 군집의 경향성을 보이는가?
		클러스터 탐지 테스트	군집에 해당하는 구체적인 범위는 어디인가?
	일반 테스트 – 초점 테스트	일반 테스트	연구 지역 전체 혹은 어디에서든 군집의 경향성 혹은 군집이 존재하는가?
		초점 테스트	특정 지역(주변)에 군집의 경향성 혹은 군집이 존재하는가?
Cromley and McLafferty (2012)	전역적 방법		전반적인 패턴이 군집의 경향성을 보이는가?
	국지적 방법	필드(field) 기반 방법	군집에 해당하는 구체적인 범위는 어디인가? – 스캐닝(scanning) 방법
		객체(object) 기반 방법	군집에 해당하는 구체적인 범위는 어디인가? – 합역(aggregation) 방법
	초점 방법		특정 지역(주변)이 군집에 해당하는가?

리고 일반 테스트와 초점 테스트로 구분될 수 있음을 시사하였다. 하지만 이 분류에 따르면 예를 들어 클러스터 탐지 테스트를 일반 테스트나 초점 테스트로 수행할 수 있음을 의미하게 되는데 구체적인 실행 방안의 측면에서 보자면 두 테스트가 어떤 차이를 갖는지 다소 혼동스러운 것으로 판단된다. Cromley and McLafferty(2012)는 공간 군집 분석 방법을 전역적 방법, 국지적 방법, 초점 방법으로 구분하였으며, 그 중에서 국지적 방법은 다시 필드 기반 방법(field-based methods)과 객체 기반 방법(object-based methods)으로 구분하였다. 하지만 여기서 필드 기반 테스트와 객체 기반 테스트는 분석의 목적이라기보다는 분석의 방법과 보다 연관된다. 또한 국지적 방법이라는 범주는 공간통계학적 의미에서는 각 단위와 그 이웃을 대상으로 하는 분석을 의미하고, 따라서 초점 분석의 의미로도 해석 가능하므로 명확한 의미 전달이 어려운 것으로 판단된다.

분석 방법의 특성 관련하여서는 분석의 목적에 의한 분류에 비해 선행연구에서 충분히 다루어지지 않았지만 특정한 연구의 맥락에 적합한 방법을 선택해야 하는 연구자의 입장에서 본다면 상당히 중요한 부분이다. 대부분의 선행연구들은 특별한 분류 없이 각 분석 목적에 해당하는 공간 군집 측정 지수를 그대로 나열하고 있다(Pfeiffer *et al.*, 2008; Rogerson and Yamada, 2009; Fisher and Getis, 2009). 하지만 Waller and Gotway(2004)는 측정 지수의 성격에 따라 스캐닝 방법, 공간적 자기상관(spatial autocorrelation) 지수, 적합도(goodness of fit) 테스트로 구분하고 있다. 스캐닝 방법은 연구지역 전체에서 클러스터로 간주되는 '정확한' 범위를 찾아내기 위해 일정한 영역을 반복적으로 설정하고 그에 대해 테스트를 수행하는 방식으로, 공간스캔통계량(Kulldorff, 1997)이나 GAM(Geographical Analysis Machine)(Openshaw *et al.*, 1988) 등이 이에 해당한다. 공간적 자기상관 지수로는 Moran's I나 Local Moran's I_i, 적합도 테스트로는 카이-스퀘어(χ^2) 테스트나 Tango(1995) 지수 등을 들고 있다.

하지만 이러한 구분은 기준이 명확하지 않아 경우에 따라 동일한 범주로 구분되기도 한다. 예를 들어 Cromley and McLafferty(2012)에 의하면 필드 기반 방법은 위치를 이동해가며 각 지점 상에 일정한 수의 국지적 범위를 설정하고 테스트를 수행하는데, Local Moran's I_i와 같은 LISA(Local Indicators of Spatial Association)나 커널 추정(Kernel Estimation), 공간스캔통계량 등이 이에 해당한다. 이에 반해 객체 지향 방법은 사건이 존재하는 지점들을 출발 '시드(seed)'로 삼아 조건에 맞는 인접 지역들을 반복적인 합역 과정에 의해 확장해감으로써 클러스터를 탐지하는 방법으로 Besag and Newell(1991)의 방법이나 Aldstadt and Getis(2006)의 AMOEBA가 이에 해당한다.

2) 분석의 목적과 방법의 특성에 따른 분류

선행연구에서의 논의를 종합하면서도 연구의 목적에 맞는 분석 방법의 특성을 체계적으로 파악할 수 있도록 분석의 목적과 분석 방법의 특성이라는 두 축을 통해 2차원 분류를 시도하였다. 분석의 목적과 관련하여서는 2단계를 생각할 수 있는데, 우선 기술적 혹은 탐색적인 목적에 주로 사용되는 것인지, 아니면 통계 테스트를 통한 확정적 목적에 주로 사용되는 것인지를 구분하였다. 이어서 통계 테스트를 수행하는 경우 앞서 살펴본 선행연구의 논의를 토대로 전역적 테스트와 초점 테스트, 구역 설정 테스트로 구분하고자 하는데, 초점 테스트와 구역 설정 테스트는 큰 틀에서 국지적 테스트에 해당한다. 전역적 테스트에서는 연구 지역 전체에서 공간 군집의 경향성이 존재하는지를 평가하며, 초점 테스트에서는 주어진 개별 공간 단위가 공간 군집에 해당하는지를 평가한다. 끝으로 구역 설정 테스트에서는 전체 연구 지역에서 군집의 범위가 어디까지인지 그 경계를 파악하기 위한 평가를 수행한다.

본 연구에서의 관심 대상인 국지적 테스트만을 대상으로 앞 장에서 파악한 분석 방법들의 특성을 토대로 각 범주에 해당하는 대표적 측정 지수들을 제시하면 표 3과 같다. 각 측정 지수들에 대한 수식이나 세부적인 내용은 함께 제시된 인용을 참조하도록 하고, 대신 여기에서는 기본적인 특성에 대해서만 살펴보도록 한다. 우선 공간 군집의 파악을 기술적 혹은 탐

색적으로 분석하기 위해 사용되는 방법들에는 공간적 집중도를 요약하는 데 흔히 사용되는 1차적 지표들이 포함된다. 서론에서 잠시 언급한 바와 같이 카운트 데이터는 흔히 공간 단위가 갖는 이질적 특성을 배제하기 위한 과정을 거치며 면적이나 준거집단의 크기를 고려한 비율이 흔히 사용된다. 비율은 그 자체로 사용되기도 하지만 어느 한 지역 값이 다른 지역의 값에 비해 얼마나 더 크고 작은지 파악할 수 있도록 표준화를 거치는 경우가 많다. 상호 간의 비교를 위해 취할 수 있는 방법 중의 하나는 데이터 셋이 가진 통계적 분포 특성을 이용해 값의 스케일을 표준화 하는 방법으로 흔히 사용되는 표준점수(z-score)가 대표적이다(De Smith *et al.*, 2013).

상호 간의 비교를 하는 두 번째 방법은 특정 지역에서의 관찰 값과 기대 값을 비교하는 방법으로 입지계수나 표준화사망률, Krugman 지수 등이 이에 해당한다. 이들 중 입지계수는 경제지리학 등 전통적인 지리학 연구에서, Krugman 지수는 경제학이나 경제지리학 연구에서(Suedekum, 2006), 상대위험도 혹은 표준화사망률의 경우는 인구지리학 혹은 보건지리학 및 공간역학(spatial epidemiology)에서 널리 사용된다. 입지계수의 경우 일반적으로 한 지역에서 특정 집단이 차지하는 비중을 전국에서 그 집단이 차지하는 비중으로 나누어 산출한다. 표준화사망률은 대상 지역에서의 사망률(혹은 사망자 수)을 준거 집단의 기대 사망률(혹은 기대 사망자 수)로 나누어 산출하는데, 이때 기대 사망률은 통상 전역의 연령층별 사망률을 관심 지역의 연령 구조에 반영한 연령 표준화 기대 사망률에 해당하는 값이다(Waller and Gotway, 2004). Krugman 지수는 관찰사건(행비중 비율)과 기대사건(행비중 비율)의 차이에 초점을 둔다.

이어 통계 테스트를 목적으로 하는 방법들을 살펴보고자 하는데, 먼저 초점 테스트의 경우는 주어진 공간 단위가 공간 군집에 해당하는지를 평가하는데 이웃의 단위와는 무관하게 개별 단위만을 대상으로 테스트하는 방법, 주어진 단위와 그 이웃 간의 연관성을 테스트 하는 방법, 그리고 이 둘을 동시에 평가하는 방법으로 구분할 수 있다. 사실 이 구분은 공간 군집과 관련하여 보다 근본적인 질문, 즉 분석 방법이 염두에 두고 있는 귀무가설(혹은 대안가설)의 성격과 관련된다. 즉 목적에 따른 적절한 방법을 사용할 필요가 있는데, 이때 귀무가설이 무엇인지를 살펴보는 것이 유용할 것으로 판단된다(Assunção and Reis 1999; Fisher and Getis, 2009).

이렇게 서로 다른 방법들이 사용되는 이유는 공간 군집의 개념과 관련되어 있는 것으로 판단되는데, 일반적으로 공간적 군집 혹은 클러스터는 "지리적 혹은 시간적으로 우연히 발생했다고 보기 어려울 정도로

표 3. 카운트 데이터 기반의 공간 군집 분석 방법의 분류

분석의 목적		분석 방법	측정 지수
기술 및 탐색적		개별 단위 분석	• 비율(행비중, 열비중) 및 밀도, 이들의 표준점수(z-score) • 입지계수 / Krugman 지수(Suedekum, 2006) / SMR 혹은 SIR(Waller and Gotway, 2004)
통계테스트	초점 테스트	개별 단위 분석	• Fuchs and Kenett's M (Rogerson and Yamada, 2009) • 표준화상이점수(SSD_i) (이상일, 2007; 2008)
		근린 분석	• 국지적 공간통계: Local Moran's I_i(Anselin, 1995) / Getis-Ord's G_i & G_i^*(Ord and Getis, 1995) • 수정 국지적 공간통계(I): $I_{cr,i}$ (Waller and Gotway, 2004) / I_{r_i} (Jackson *et al.*, 2009) • 수정 국지적 공간통계(II): $I_{DR,i}$ (Zhang and Lin, 2008) • 수정 국지적 공간통계(III): $I_{EBI,i}$ (Zhang and Lin, 2008) • 공간적 입지계수(focal location quotient) (Cromley and Hanink, 2012)
		개별 단위 및 근린 분석의 결합	• 공간적 카이-스퀘어 통계량: Tango's C_F (Tango, 1995) / Rogerson's R_i(Rogerson, 1999)
	구역 설정 테스트	초점 테스트의 병렬 수행	상기 참조
		스캐닝	• 공간스캔통계량(Kulldorff, 1997)
		합역	• (수정) AMOEBA(국지적 공간통계 기반)(Aldstadt and Getis, 2006; 이상일 등, 2010)

충분히 크고 밀집된 사건들의 집합체"(Knox, 1989)로 인식된다. 널리 사용되는 Local Moran's I_i와 같은 공간적 자기상관 분석의 경우 인접한 공간 단위 간에 값이 유사한지를 분석한다. 주어진 지역의 값이 높고, 그 이웃에도 높은 값들이 위치한다면 공간 군집으로 파악하게 되는데, 높은 값을 갖는 공간 단위들이 밀집해 있는지가 초점이 된다. 문제는 대부분의 카운트 데이터가 구역 단위로 합산된 데이터이기 때문에 한 구역 자체로 기대 이상의 높은 빈도를 나타내면 이웃 구역과는 무관하게 군집에 해당하는지를 평가하는 것이 가능할 수 있다는 점인데, 결과적으로 이는 개별 단위 분석의 근거가 된다. 세 번째 방법은 이 둘을 결합하는 것으로 주어진 단위 자체의 빈도가 아주 높거나, 꼭 그렇지는 않더라도 주변에 높은 빈도가 집중되어 있다면 군집으로 파악할 수 있다는 논리를 반영하고 있다.

개별 단위 분석에 언급된 방법들은 주어진 단위별로 관찰사건의 크기가 기대되는 것과 얼마나 차이를 나타내는지를 테스트한다. Fuchs and Kenett's M (Rogerson and Yamada, 2009)이 전역적인 적합도 테스트인 카이-스퀘어 테스트의 국지적 버전과 유사하다면, 이상일(2007; 2008)이 제안한 표준화상이점수(standardized score of dissimilarity)는 거주지 분리 등의 연구에서 널리 사용되는 상이지수(index of dissimilarity)의 국지적 버전과 유사하다고 할 수 있다. 이들은 관찰 빈도와 기대 빈도의 차이 혹은 관찰 열비중 비율과 기대 열비중 비율 간의 차이가 큰 경우 군집으로 파악하게 된다.

근린 분석에 언급된 대표적 방법들은 흔히 공간통계학에서 공간적 연관성을 측정하는 국지적 지수들이다. 이들 지수는 대상지역과 그 이웃(즉, 근린)에 대해 투입되는 변수의 유사성(예를 들어 Local Moran's I_i(Anselin, 1995))이나 집중도(예를 들어 Getis-Ord's G_i 및 G_i^*(Ord and Getis, 1995))를 측정하는데, 그 과정은 대부분 주어진 근린 내의 값들과 평균, 혹은 이웃 값과의 비교를 통해 이루어진다. 이들 지수는 기본적으로 대상지역과 이웃 간의 관계에 초점을 두기 때문에 앞 장에서 전술한 것처럼 입력 변수의 형태는 상당히 다양할 수 있다. 국지적 공간통계와는 달리 공간적 입지계수(Cromley and Hanink, 2012)는 입지계수와 동일한 논리를 가지되 측정 시 이웃에 해당하는 지역을 함께 고려하며 통계적 유의성 평가를 지원하고 있다.

그런데 국지적 공간통계 지수들은 공간통계학에서 비율척도로 측정되는 연속형의 변수들을 대상으로 개발되어 왔기 때문에 카운트 데이터, 특히 카운트 기반의 비율 데이터에 적용하기 위해서는 주의가 필요하다는 지적이 있어 왔다(Besag and Newell, 1991; Walter, 1992). 원 측도들이 분산의 불안정성(instability of variance) 혹은 작은 수의 문제(small number problem)(Waller and Gotway, 2004)에 취약할 수 있다는 지적(이에 대해서는 4장 2절 참조)과 함께 이에 대응하기 위한 방안이 요구되었고, 수정 국지적 측도들은 그 대응 방안의 일환이라고 할 수 있다. 열비중 방식의 비율과 국지적 분산을 사용하는 방안(수정 국지적 공간통계(I)), 관찰 빈도와 기대 빈도의 잔차를 사용하는 방안(수정 국지적 공간통계(II)), 행비중 비율을 사용하되 분산의 불안정성을 개선하기 위해 표준화를 하는 방안(수정 국지적 공간통계(III)) 등이 제안되었다. 카운트를 그대로 사용하는 분석 역시 표준적인 측도들이 기초로 하고 있는 정규분포의 가정으로부터 자유롭기 어렵다. 이 이슈에 대해서는 다음 장에서 다시 논의한다.

초점 테스트의 세 번째 방법들은 기본적으로 개별 단위 분석과 근린 분석의 결합이다. 이 방법을 추구하는 연구자들(Tango, 1995; Rogerson, 1999)에 따르면 개별 단위 통계량은 주변 지역과의 관련성을 충분히 드러내지 못 하는 반면 공간적 자기상관 통계량은 각 지역과 이웃 간의 '유사성'에만 초점을 두고 있어 관심 지역 자체에 대한 평가가 충분히 이루어지지 못한다. 즉, 앞의 둘은 공간 군집과 관련된 서로 다른 두 측면을 다루고 있으므로 서로 결합될 필요가 있음을 주장한다. 여기서 더 나아가 이 방법들은 바로 위에서 지적한 카운트 데이터가 가진 문제점에 대응하기 위한 요소를 포함하는 지수들이라고 할 수 있다.

지금까지 살펴본 초점 테스트와는 달리 구역 설정 테스트의 경우 구역의 범위를 미리 지정하지 않고, 분석을 통해 그 영역을 파악하는 것으로 역시 세 가지

방법이 가능하다. 가장 흔히 사용되는 방법으로는 초점 테스트의 병렬 수행을 들 수 있다. 국지적 공간통계를 예로 들자면, 모든 각 단위 지역에서 통계량을 산출하게 되고, 그 결과 핫스팟(hot spot)이나 콜드스팟(cold spot)의 위치를 파악할 수 있다. 이때 통계적으로 유의미한 지역들의 공간적 연합을 공간 군집으로 간주하게 되는데, 그 지도는 흔히 '클러스터 지도' 혹은 '유의성 지도'라 불린다(Anselin, 2003). 하지만 이 결과가 곧 클러스터의 경계와 같은지에 대해서는 명확히 알 수 없으며(이상일 등, 2010), 따라서 구역 자체를 평가하는 다른 두 가지 방식의 대안이 개발되어 왔다.

먼저 스캐닝 방식의 경우, Kulldorff의 공간스캔통계량이 대표적인데, 기본적으로 대상 지점을 중심으로 원을 계속 확장해가며 우도비 통계량을 산출하게 된다. 이때 분모의 경우는 관찰 데이터 상에서 구역 내부 및 외부에서 빈도가 실현될 가능성(우도)을, 분자의 경우는 귀무가설 하에서 구역 내부 및 외부의 빈도가 실현될 가능성(우도)을 산출해 그 값이 최대인 원의 영역을 군집으로 파악하게 된다. AMOEBA 알고리즘과 같은 합역 방식의 경우, 주어진 각 공간 단위를 출발 시드로 삼아 인접해 있는 공간 단위들을 계속 합역해 가며 구역에 대한 국지적 공간통계 테스트를 수행하게 되고, 조건을 충족하면서 그 유의성이 최대인 범위를 군집의 경계로 설정하게 된다. 이 두 방식에 있어 통계량의 속성 상 어느 공간 단위를 먼저 합역해가는 것이 합당한지를 판단할 수 있다면 합역 방식을 택하겠지만 그렇지 않은 경우는 스캐닝 방식을 사용하여야 할 것으로 판단된다.

4. 방법론적 이슈와 과제

1) 분석 방법의 적용과 관련된 이슈

이상의 논의를 토대로 카운트 데이터 기반의 공간 군집 분석과 관련한 주요 방법론적 이슈 및 향후 과제를 정리해보자면 다음과 같다. 이는 크게 두 가지 유형의 측면에서 검토될 수 있는데, 첫 번째는 성격이 비교적 뚜렷이 구분되는 분석 방법이 다양하게 존재하고 있고, 따라서 방법론의 특성에 따라 분석 결과 및 그 해석 또한 영향을 받을 수 있다는 점이다. 예를 들어 행비중 비율 간의 비를 사용하는 입지계수형의 지수는 특정 집단이 한 지역에서 차지하는 비중이 기대되는 비중에 비해 몇 배나 더 높은지(혹은 낮은지)를 측정한다. 즉 대상 지역에서 특정 사건이 차지하는 비중에 초점이 있어 전국의 총 사건 중 어느 지역에 얼마나 많은 사건이 몰려있느냐와는 다른 개념이며, 보다 정확히는 집중도(concentration)라기보다는 특화도(specialization)에 해당한다는 것이다(Hildebrand and Mace, 1950; Waller and Gotway, 2004; Suedekum, 2006). 이와 같은 특성으로 인해 카운트 데이터 기반의 공간 군집 분석에는 어떤 유형의 변수를 사용할 것인지에 대해 신중한 검토가 필요하다. 원칙적으로 이야기하자면 관찰 빈도를 설명할 기저 변수를 상정할 수 있다면 비율을 사용하는 것이 합당하며, 이 경우 행비중 비율의 경우는 집중도 보다는 특화도에 초점이 있는 반면, 열비중 비율의 경우는 집중도에 보다 초점이 있음을 인식할 필요가 있다(Ceapraz, 2008).

이와 연관하여 더 지적할 것은 각 방법들에서 한 지역의 측정 값을 다른 지역이나 평균적인 경향과 비교하기 위해 기본적으로 사용하는 방식이 비율이냐 차이이냐의 문제를 들 수 있다. 예를 들어 입지계수는 기대 값과의 비율을 사용하고 있는 반면 표준화상이점수는 기대 값과의 차이를 사용하고 있는데, 전자가 상대 값이라면 후자는 절대 값에 해당한다. 일반적으로 상대 값에 대한 해석이 보다 직관적이지만 결국 이 역시 비율에 기반하고 있음을 유의할 필요가 있다. 다시 입지계수를 예로 들자면 분모에 해당하는 비율(전국에서 특정 집단이 차지하는 비중)은 모든 지역에서 값이 동일한 상수이므로 위에서 살펴본 행비중 비율이 갖는 특성을 그대로 가지게 될 뿐만 아니라 비율 데이터의 통계학적 신뢰성과 관련된 이슈에서도 자유로울 수 없다(이에 대해서는 다음 절에서 다시 상술한다).

이를 앞에서 언급한 변수의 기본 유형과 함께 살펴

본다면 카운트 데이터의 공간 분석은 변수의 형태(행비중과 열비중), 기대 값(혹은 평균)과의 비교 방법(비율과 차)의 두 축으로 방법들 간의 특성을 비교할 수 있다. 이들 중 행비중 비율 변수이면서 기대 값과의 비에 초점을 두는 지수(입지계수나 표준화사망률, 공간스캔통계량 등), 열비중 비율 변수이면서 기대 값(열비중 비율)과의 차이에 초점을 두는 지수(표준화상이점수, 공간적 카이-스퀘어 통계량 등), 주어진 변수에 대해 평균이나 이웃 값과의 차이에 초점을 두는 지수(국지적 공간통계 및 AMOEBA) 등이 대표적이다. 이러한 방식의 차이에 대한 연구가 필요한데, 관련하여 이상일(2007; 2008)은 특정 집단의 공간적 집중을 파악하기 위해 행비중 비율 간의 비를 사용하기 보다는 열비중 간의 차이를 사용하는 방식이 보다 합리적인 결과를 도출할 수 있음을 제시한 바 있다. 하지만 이러한 접근법의 차이가 공간 군집 분석 결과에 어떠한 결과를 미치는지 그 구체적인 과정과 영향에 대해서는 심층적인 실증 연구가 필요할 것으로 판단된다.

이와 함께 어떻게 측정하느냐의 문제도 중요한데, 구체적으로 초점 테스트와 구역 설정 테스트를 위해 사용되는 각 방법들의 간의 차이에 대한 비교 검토가 요구된다는 점을 지적할 수 있다. 초점 테스트의 경우 개별 분석, 근린 분석, 그리고 이 둘의 결합 등 세 가지 유형으로 구분하였는데, 이러한 차이가 미치는 영향은 물론 공간 군집의 개념이 이러한 서로 다른 접근에 모두 동등하게 합당한지를 면밀히 분석할 필요가 있다. 구역 설정 테스트 역시 초점 테스트를 병렬적으로 실행하는 방법과 그렇지 않은 테스트 간에 본질적인 차이를 규명할 필요가 있을 것으로 판단된다. 이와 관련하여서는 초점 테스트의 병행 시 각 분석 지역 간에 이웃이 중복되어 통계량 또한 상관성을 가지는 소위 다중 비교의 문제(multiple testing problem)(Anselin, 1995; Rogerson and Yamada, 2009)나 구역 자체에 대한 평가를 하지 않음으로 인해 발생하는 결과 상의 문제(Bogart and Ferry, 1999; 이상일 · 조대헌, 2012; 조대헌 · 이종일, 2013) 등을 포함한 검토가 필요하다. 그와 동시에 구역 설정을 위한 새로운 접근 방법의 개발이 필요할 것으로 생각되는데, 예를 들어 구역에 대해 초점 테스트를 하는 방안이 대안이 될 수 있을 것으로 판단된다.

2) 카운트 데이터의 통계학적 특성과 관련된 이슈

카운트 데이터 기반의 공간 군집 분석과 관련된 두 번째 유형의 이슈는 카운트 데이터 자체의 통계학적 특성과 관련된다. 일반적으로 카운트 데이터, 특히 비율 데이터와 관련하여 서로 연관된 두 가지 이슈가 큰 주목을 받고 있다. 먼저 비율 데이터는 통계적 분포에서 상당한 비대칭을 나타내고 결과적으로 비정규성이 커진다는 점이 오래 전부터 지적되어 왔다(Atchley *et al.*, 1975; Barnes, 1982; Kane and Meade, 1998). 이는 곧 정규분포를 기본 가정으로 하여 패턴을 파악하는 대부분의 통계학적 분석에서 비율 데이터가 문제를 일으킬 수 있음을 의미한다. 예를 들어 한 비율 변수에 대해 표준점수를 산출하는 경우 그 값은 기계적으로 산출되겠지만 그것이 가진 통계학적 해석의 근거는 흔들리게 된다. 이 문제는 앞서 살펴본 바와 같이 원 비율은 물론 그것을 표준화는 입지계수형의 비율에도 동일하게 적용된다.

카운트 데이터에 기초한 비율 데이터의 또 다른 통계학적 이슈는 분산의 불안정성 혹은 작은 수의 문제로 알려져 있다(Wallter and Gotway, 2004; Anselin *et al.*, 2006; Rogerson and Yamada, 2009; Fisher and Getis, 2009; Cromley and McLafferty, 2012). 흔히 사용되는 비율 데이터는 연구 지역에서 특정 사건의 빈도(예: 사망자 수)가 차지하는 비율, 즉 행비중 비율로 나타내며 이를 통해 사건의 발생 확률을 추정하게 된다. 그런데 문제는 발생률의 분포에서 분산의 값이 분모에 해당하는 준거 집단, 즉 위험 인구의 크기에 의존한다는 점이다(Anselin *et al.*, 2006). 보다 정확히 말하자면 발생률의 분산은 위험 인구의 규모에 반비례하고, 따라서 분모의 크기가 작은 경우 분산은 커지는 경향을 보이게 된다. 다시 말해 분모가 작은 지역에서 사건 수가 조금만 달라지더라도 그 비율 값의 차이는 매우 커지게 되는데, 실제 관찰 데이터에서도 이를 쉽게 발견할 수 있다. 이와 같이 비율 값의 변동성이 서로 이질적이라면 이를 고려하지 않는 경우

통계학적 신뢰성이 저하될 수 있다.

비율 데이터가 갖는 분산의 불안정성의 문제는 곧 분모에 해당하는 위험 인구의 분포가 나타내는 이질성의 문제인데, 이는 개별 공간 단위별 분석은 물론 공간적 연관성 측도들에 기반한 분석에도 영향을 미쳐 매우 복잡한 양상을 보인다(Fisher and Getis, 2009). 사실 이질적인 인구의 분포를 고려하지 않고 공간 군집을 분석하는 경우 통계학적인 신뢰성에 문제를 유발한다는 지적은 1990년대부터 이루어져왔다(Besag and Newell, 1991; Walter, 1992; Assunção and Reis, 1999). 구체적으로 이들의 연구는 카운트 혹은 비율 데이터인 경우 Moran's *I*가 기초하는 통계학적 가정이 모두 비현실적이며, 인구의 이질적 분포를 고려하지 않는 경우 측정 결과에 1종 오류가 증가하는 반면 통계적 파워는 감소함을 파악하였다.

이러한 문제들에 대한 인식이 높아지면서 이에 대응하기 위한 다양한 방안들이 연구되어 왔다. 그 방향은 공간 단위의 조정, 데이터의 변환(transformation), 평활(smoothing), 측정 지수의 개선, 대안적인 통계학적 평가 방법의 사용 등으로 정리될 수 있다. 우선 공간 단위의 조정은 비율 값의 분모에 해당하는 변수의 크기가 공간 단위들 간에 서로 유사하도록 단위를 재설정 한 후 비율 값을 산출하는 방법을 말한다(Wallter and Gotway, 2004). 하지만 일반적으로 인구의 규모가 이질적인 경우는 재조정된 공간 단위 간의 면적에 불균형이 발생하게 된다.

데이터 변환은 주어진 원 비율 데이터에 적절한 함수나 체계를 적용해 변환하는 과정인데, 가장 간단하게는 비율 값에 제곱근이나 로그를 취하는 방법이 있으며(Kane and Meade, 1998), 보다 복잡하게는 Freeman-Tukey 변환이나 ArcSin 변환, 경험적 베이즈(Empirical Bayes) 표준화 등을 들 수 있다(Anselin *et al.*, 2006). 전자의 방법들이 비율 데이터가 갖는 비정규성의 문제를 개선하는데 초점을 둔다면 후자는 분산의 불안정성의 문제를 개선하는데 초점을 두고 있지만 어떤 경우이든 원 변수의 값이 달라지기 때문에 그 의미를 해석하기가 어렵다는 문제점을 갖는다. 그럼에도 불구하고 경험적 베이즈 표준화 기법은 최근 관련 연구에서 그 활용도가 커지는 추세에 있다(Assunção and Reis, 1999; Zhang and Lin, 2008). 이 기법은 소위 사전 분포(prior distribution)로 알려진 부가적인 정보(사전 분포의 평균과 분산)를 이용해 비율 값들이 평균적 경향을 향해 '축소(shrinkage)' 되도록 하는 일종의 표준 점수를 산출한다(Anselin *et al.*, 2006).

비율 데이터의 문제를 개선하는 세 번째 방안은 원 데이터 상의 변수 값들을 이용해 통계적인 평활을 수행하는 것이다. Anselin *et al.*(2006)은 평활이 적용되는 방식에 따라 평균 및 중앙값 기반 평활, 비모수적 평활, 경험적 베이즈 기반 평활, 모델 기반의 평활 등으로 구분하고 있는데, 본질적으로 원 자료가 국지적인 경향성을 반영하여 추정되어 변환된다는 특징을 가진다. 그런데 평활 방식에서 문제가 되는 요소는 공간적 평활 기법의 경우 대상 지역의 결과 값이 자신 및 주변지역의 값에 의존적이기 때문에 평활 과정 그 자체에서 공간적 자기상관의 요소가 부가된다는 것으로, 여기에 공간적 자기상관의 측정을 시도하는 것은 주의가 필요하다 (McLaughlin and Boscoe, 2007).

비율 데이터의 문제와 관련하여 취할 수 있는 또 다른 방안은 통계량에 비율 데이터의 특성을 반영하는 것이다. 특히 Moran's *I* 및 Local Moran's I_i를 중심으로 한 공간적 연관성 측도에 대한 개선이 집중적으로 시도되었는데, 3장에서 살펴본 것처럼 그 방향은 크게 3가지 측면에서 진행되어 왔다(Oden, 1995; Waldhör, 1996; Assunção and Reis, 1999; Waller and Gotway, 2004; Zhang and Lin, 2008; Jackson *et al.*, 2009). 개별 단위에 대한 테스트와 공간적 연관성 테스트를 동시에 수행하는 공간적 카이-스퀘어 통계량(Tango, 1995; Rogerson, 1999) 역시 열비중 비율에 초점을 두면서 국지적 분산을 고려함으로써 위험인구의 이질성을 고려하는 맥락 속에서 개발되어 왔다. 이러한 측정 지수의 개선은 행비중 비율을 사용하기보다는 열비중 비율에 초점을 두고 있으나 제시된 방안들 상호 간에 비율 데이터의 특성이 어떤 영향을 미치는지에 대한 추가적인 실증 연구가 요구된다.

비율 데이터의 공간 패턴 분석과 관련하여 마지막으로 고려할 수 있는 대응 방안은 패턴에 대한 통계학적 검증 방법과 관련된 것이다. 대부분의 통계학적 가

설 검증은 그것이 가정하는 확률 분포에 근거하는데 예를 들어 Moran's I 및 Local Moran's I_i는 검정 통계량의 유의성을 파악하기 위해 정규성 가정에 의한 방법과 랜덤화 가정에 의한 방법을 사용한다. 정규성 가정의 경우 관찰 변수가 서로 독립적이며 발생확률이 동일한 IID(independent and identically distributed) 정규 분포를 가정하지만, 비율 데이터에서 이는 보증되기 어려운 가정이며, 더욱이 Local Moran's I_i의 경우는 통계량의 분포 특성 자체가 잘 알려져 있지 않다(Anselin, 1995; Rogerson and Yamada, 2009). 따라서 모집단에 대한 정규분포 가정을 하지 않아도 되는 랜덤화 가정이 보다 선호되는 경향이 있다(Anselin, 1995; 이상일, 2008).

그러나 랜덤화 가정 역시 비율 데이터의 분모, 즉 위험 인구의 분포가 갖는 공간적 구조로 인해 영향 받으며, 특히 전역적 Moran's I에서는 유지되기 어려운 가정이다(Besag and Newell, 1991). 따라서 이에 대한 대응책이 요구되는데, 랜덤화 적용시 비율 자체를 랜덤화하기 보다는 분모의 값은 고정하고 비율 데이터의 분자에 해당하는 카운트만을 랜덤화하거나(McLaughlin and Boscoe, 2007), 랜덤화 대신 몬테카를로(Monte Carlo) 시뮬레이션을 통해 이질적 포아송(heteroneneous Poisson) 분포로부터 카운트를 모의함으로써 귀무 분포를 형성하는 방법(Waller and Gotway, 2004) 등이 제안되었다. 특히 카운트 데이터의 공간 분석에서 몬테카를로 시뮬레이션은 일찍부터 유용한 평가 방법으로 주목받아 왔다(Besag and Newell, 1991; Fisher and Getis, 2009). 하지만 현재 널리 사용되고 있는 방법들 중에서 몬테카를로 시뮬레이션을 주된 방법으로 사용하고 있는 경우(Kulldorff, 1997)는 드문데 그것은 실용 가능한 도구의 부재가 주요 요인인 것으로 판단되지만 다양한 경험 연구를 통해 그 타당성을 평가할 필요가 있다.

카운트 데이터의 자체적인 특성과 관련된 이러한 이슈들은 결국 통계학적 테스트의 신뢰성을 어떻게 확보할 것인가라는 문제로 귀결된다. 요약하자면 데이터 혹은 변수를 가공하거나, 대안적인 측정 지수를 개발하거나, 대안적인 통계 테스트를 적용하는 대응이 가능할 것으로 판단된다. 공간 단위를 조정하거나 변수를 변화하는 방법, 통계학적 평활을 시도하는 방법은 원 데이터가 가진 상세함이나 의미가 사라져 결과적으로 해석 상의 어려움을 초래한다는 면에서 나머지 두 방식이 보다 합리적일 것으로 판단된다. 측도를 개선하거나 개발하는 경우는 열비중 비율이 통계 테스트 과정에서의 문제를 개선하는데 보다 유효할 것으로 기대되지만 이에 대해서는 추가 연구가 필요하다. 통계학적인 유의성을 평가하는 방안과 관련하여서는 몬테카를로 시뮬레이션의 가능성을 보다 적극적으로 평가해 볼 필요가 있는데 GIS나 공간 분석 S/W에서 이를 용이하기 수행할 수 있는 도구의 개발이 추가적으로 이루어질 필요가 있다.

5. 결론

지리적 현상의 공간적 분포를 파악하는데 있어 카운트는 가장 기본적이고 중요한 데이터원이다. 카운트 데이터를 이용해 공간 군집을 파악하려는 시도는 지리학의 오랜 관심사 중의 하나이지만 다양한 분야에서 관련 연구가 빠르게 늘어나고 있다. 본 연구에서는 이러한 최근의 성과를 살펴보기 위해 경험 연구들을 대상으로 연구의 동향을 정리함과 동시에 다양하게 개발되고 있는 분석 방법들을 유형화하고, 그 특성을 파악하고자 하였다.

먼저 경험 연구들을 살펴본 결과 카운트 데이터에 기초한 군집 분석은 전통적인 지리학의 영역뿐만 아니라 보건이나 범죄 등의 분야에서 활발히 수행되고 있었다. 방법론적인 측면에서 볼 때 기술적 혹은 탐색적 요약을 위한 방법을 일부 포함하여, 최근의 공간통계학 혹은 공간데이터 분석 분야를 중심으로 한 분석 방법들이 다양하게 사용되고 있었다. 빈도를 그 자체로 분석하기보다는 비율 데이터에 기초한 분석이 대부분이며, 특히 대상 지역 내에서 특정 집단이나 사건이 차지하는 비중을 의미하는 행비중 비율에 기초한 분석이 가장 널리 사용되었다. 이어 분석 방법들에 대한 유형화를 시도하였으며 동시에 그 특성을 살펴보았다. 유형화를 위해 분석의 목

적과 분석 방법의 특성이라는 두 가지 축을 설정하고, 국지적 분석 방법만을 대상으로 분류를 시도하였다. 분석의 목적과 관련해서는 기술 및 탐색적 분석과 통계테스트에 의한 분석으로 구분 후 다시 초점 테스트와 구역 설정 테스트로 구분하였으며, 각 목적을 달성하기 위해 사용하는 접근 방법의 특성에 따라 방법론적인 구분이 가능하였다.

이상의 과정을 통해 살펴본 결과 카운트 데이터 기반의 공간 군집 분석 방법은 그 목적이나 방법에 따라 상호 간에 비교적 명확한 차이를 보였다. 즉, 분석 방법에 따라 다소 간에 다른 측면을 측정하고 평가하는 것으로 나타났다. 이는 방법에 따라 분석하는 내용이 달라질 수 있음을 의미하고, 따라서 구체적인 특정 연구에서 분석 방법의 선택이 상당히 신중해야 함을 의미한다. 이런 면에서 특정 분석 방법이 염두에 두고 있는 귀무가설(혹은 연구가설)이 무엇인지를 면밀히 살펴보는 것은 매우 중요한 과정이 될 것으로 판단된다. 분석 방법을 적용하는 과정 상의 문제와는 별개로 카운트 데이터 자체의 특성과 관련된 이슈들과 그에 대한 대응책에 대해서도 살펴보았다. 카운트 데이터 혹은 비율 데이터의 경우 분산의 불안정성으로 인해 통계적 신뢰성의 저하 가능성이 있음을 인식할 필요가 있으며, 이에 대한 다양한 대응 방안들이 존재하므로 가장 효과적인 전략에 대한 검토가 요구된다.

요컨대, 카운트 데이터 기반의 공간 군집 연구는 여러 분야에서 비교적 활발히 이루어지면서 다양한 방법론들이 제안되고 개발되어 왔다. 하지만 궁극적으로 카운트 데이터 기반의 공간 군집이라는 것이 어떤 것을 의미하는지 그 개념을 명확히 하고, 그에 적합한 방법이 무엇인지를 보다 체계적이고 실증적인 연구를 통해 밝힐 필요가 있으며, 이를 지원할 수 있는 분석 도구의 개발 또한 중요한 과제가 된다.

참고문헌

김병선 · 김걸, 2009, “서울시 생산자서비스 산업의 공간적 분포 패턴 변화분석: FIRE 산업을 중심으로,” 국토지리학회지, 43(3), 399-408.

김영호, 2010, “서울시 자전거 이용의 공간 네트워크 패턴 연구: 공간적 네트워크 자기상관을 중심으로,” 국토지리학회지, 44(3), 339-352.

박선일 · 배선학, 2012, “구제역의 시 · 공간 군집 분석: 2010~2011 한국에서 발생한 구제역을 사례로,” 한국지역지리학회지, 18(4), 464-472.

손승호, 2010, “사회 · 경제적 속성을 통해 본 인천의 도시 구조,” 한국도시지리학회지, 13(3), 27-38.

신상영, 2010, “1인가구 주거지의 공간적 분포에 관한 연구: 서울시를 사례로,” 국토계획, 45(4), 81-95.

신우진 · 신우화, 2009, “서울시 소매업종 공간분포패턴에 관한 연구,” 부동산연구, 19(2), 279-296.

이경주 · 권일, 2012a, “비도시 지역의 공장 개별입지 난개발에 관한 실증적 분석,” 한국지역개발학회지, 24(5), 145-159.

이경주 · 권일, 2012b, “공간현상 분석을 위한 GIS 기반의 공간통계적 접근방법에 관한 고찰: 공간 군집지역 탐색을 위한 공간검색통계량의 실증적 사례분석,” 한국공간정보학회지, 20(1), 81-90.

이상일, 2007, “거주지 분화에 대한 공간통계학적 접근(I): 공간 분리성 측도의 개발,” 대한지리학회지, 42(4), 616-631.

이상일, 2008, “거주지 분화에 대한 공간통계학적 접근(II): 국지적 공간 분리성 측도를 이용한 탐색적 공간데이터 분석,” 대한지리학회지, 43(1), 134-153.

이상일 · 조대헌, 2012, “비선호 시설의 인구분포 관련 입지기준 평가를 위한 GIS-기반 방법론 연구: 원자력 발전소의 경우,” 대한지리학회지, 47(5), 755-774.

이상일 · 조대헌 · 손학기 · 채미옥, 2010, “공간 클러스터의 범역 설정을 위한 GIS-기반 방법론 연구: AMOEBA 기법,” 대한지리학회지, 45(4), 502-520.

이주형 · 선권수, 2009, “토지이용밀도 및 주거유형별 분포에 따른 서울시 중심지 변화에 관한 연구,” 한국지역개발학회지, 21(2), 253-280.

조대헌 · 이종일, 2013, “고용 중심지의 범역 설정을 위한 GIS 기반 접근법의 적용,” 2013년 한국도시지리학회 하계학술대회 자료집, 충북대학교, 60-66.

최열 · 신종훈 · 박원전, 2012, “1인가구 분포 및 밀집지역 유형 분석: 부산광역시 사례,” 대한토목학회논문집, 32(6D), 655-662.

Aldstadt, J. and Getis, A., 2006, Using AMOEBA to create a spatial weights matrix and identify spatial clusters, *Geographical Analysis,* 38(4), 327-343.

Anderson, N. B. and Bogart, W. T., 2001, The structure of sprawl: identifying and characterizing employment centers in polycentric metropolitan areas, *American Journal of Economics and Sociology,* 60, 147-169.

Anselin, L., 1995, Local indicators of spatial association—LISA, *Geographical analysis,* 27(2), 93-115.

Anselin, L., 2003, *GeoDa 0.9 User's Guide,* Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics, University of Illinois.

Anselin, L., Lozano, N., and Koschinsky, J., 2006, Rate transformations and smoothing, *Urbana,* 51, 61801.

Assunção, R. M. and Reis, E. A., 1999, A new proposal to adjust Moran's *I* for population density, *Statistics in Medicine,* 18(16), 2147-2162.

Atchley, W. R., Gaskins, C. T., and Anderson, D., 1976, Statistical properties of ratios. I. Empirical results, *Systematic Biology,* 25(2), 137-148.

Bailey, T. C. and Gatrell., A. C., 1995, *Interactive Spatial Data Analysis*, UK: Longman Scientific & Technical.

Barnes, P., 1982, Methodological implications of non-normally distributed financial ratios, *Journal of Business Finance* & *Accounting*, 9(1), 51-62.

Berglund, S. and Karlström, A., 1999, Identifying local spatial association in flow data, *Journal of Geographical Systems,* 1(3), 219-236.

Besag, J. and Newell, J., 1991, The detection of clusters in rare diseases, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 143-155.

Bivand, R., Müller, W. G., and Reder, M., 2009, Power calculations for global and local Moran's *I*, *Computational Statistics* & *Data Analysis,* 53(8), 2859-2872.

Boscoe, F. P., McLaughlin, C., Schymura, M. J., and Kielb, C. L., 2003, Visualization of the spatial scan statistic using nested circles, *Health* & *Place,* 9(3), 273-277.

Burra, T. J., M., Burnett, R. T., Anderson, M, 2002, Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario, *The Canadian Geographer,* 46(2), 160-171.

Carroll, M. C., Reid, N., and Smith, B. W., 2008, Location quotients versus spatial autocorrelation in identifying potential cluster regions, *The Annals of Regional Science,* 42(2), 449-463.

Ceapraz, I. L., 2008, The concepts of specialisation and spatial concentration and the process of economic integration: theoretical relevance and statistical measures: the case of Romania's regions, *Romanian Journal of Regional Science,* 2(1), 68-93.

Coffey, W. and Shearmur, R., 2001, The identification of employment centres in Canadian metropolitan areas: The example of Montreal, 1996, *The Canadian Geographer,* 45(3), 371-386.

Craglia, M., Haining, R., and Wiles, P., 2000, A comparative evaluation of approaches to urban crime pattern analysis, *Urban Studies,* 37(4), 711-729.

Cromley, R. G. and Hanink, D. M., 2012, Focal location quotients: specification and applications, *Geographical Analysis,* 44(4), 398-410.

De Smith, M. J., Goodchild, M. F., and Longley, P., 2013, *Geospatial Analysis: a Comprehensive Guide to Principles,Techniques and Software Tools, 4th Edition,* The Winchelsea Press.

Duque, J. C., Aldstadt, J., Velasquez, E., Franco, J. L., and Betancourt, A., 2011, A computationally efficient method for delineating irregularly shaped spatial clusters, *Journal of Geographical Systems,* 13(4), 355-372.

Fernhaber, S. A., Gilbert, B. A., and McDougall, P. P., 2008, International entrepreneurship and geographic location: an empirical examination of new venture internationalization, *Journal of International Business Studies,* 39(2), 267-290.

Feser, E., Sweeney, S., and Renski, H., 2005, A descriptive analysis of discrete US industrial complexes, *Journal of Regional Science,* 45(2), 395-419.

Fischer, M. M. and Getis, A. eds., 2010, *Handbook of Applied Spatial Analysis: Software Tools, Methods and Application*, Springer.

Forstall, R. L. and Greenre, P., 1997, Defining job concentrations: the Los Angeles case, *Urban Geography,* 18, 705-739.

Gangnon, R. E. and Clayton, M. K., 2001, A weighted average likelihood ratio test for spatial clustering of disease, *Statistics in Medicine,* 20(19), 2977-2987.

Giuliano, G. and Small, K. A., 1991, Subcenters in the Los Angeles region, *Regional Science and Urban Economics,* 21(2), 163-182.

Goovaerts, P. and Jacquez, G., 2004, Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York, *International Journal of Health Geographics,* 3(1), 14.

Goovaerts, P. and Jacquez, G. M., 2005, Detection of temporal changes in the spatial distribution of cancer rates using local Moran's I and geostatistically simulated spatial neutral models, *Journal of Geographical Systems,* 7(1), 137-159.

Han, S. S. and Qin, B., 2009, The spatial distribution of producer services in Shanghai, *Urban Studies,* 46(4), 877-896.

Hanson, C. E. and Wieczorek, W. F., 2002, Alcohol mortality: a comparison of spatial clustering methods, *Social Science* & *Medicine,* 55(5), 791-802.

Hinman, S. E., Blackburn, J. K., and Curtis, A., 2006, Spatial and temporal structure of typhoid outbreaks in Washington, DC, 1906-1909: evaluating local clustering with the G_i^* statistic, *International Journal of Health Geographics,* 5(1), 13.

Huang, L., Stinchcomb, D. G., Pickle, L. W., Dill, J., and Berrigan, D., 2009, Identifying clusters of active transportation using spatial scan statistics, *American Journal of Preventive Medicine,* 37(2), 157-166.

Jackson, M. C., Huang, L., Luo, J., Hachey, M., and Feuer, E., 2009, Comparison of tests for spatial heterogeneity on data with global clustering patterns and outliers, *International Journal of Health Geographics,* 8(1), 55.

Jacquez, G. M. and Greiling, D. A., 2003, Local clustering in breast, lung and colorectal cancer in Long Island,New York, *International Journal of Health Geographics,* 2(1), 3.

Johnson, G. D., 2004, Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling, *International Journal of Health Geographics,* 3(1), 29.

Kane, G. and Meade, N., 1998, Ratio analysis using rank transformation, *Review of Quantitative Finance and Accounting*, 10(1), 59-74.

Knox, E. G. 1989, Detection of clusters, in Elliott, P. ed., *Methodology of Enquiries into Disease Clustering,* London: Small Area Health Statistics Unit, 17-20.

Kulldorff, M., 1997, A spatial scan statistic, *Communications in Statistics-Theory and Methods,* 26(6), 1481-1496.

Kulldorff, M., Tango, T., and Park, P. J., 2003, Power comparisons for disease clustering tests, *Computational Statistics* & *Data Analysis,* 42(4), 665-684.

Liu, Z., 2013, Global and local: measuring geographical concentration of China's manufacturing industries, *The Professional Geographer*, (in press).

Lloyd, C. D., 2010, Exploring population spatial concentrations in Northern Ireland by community background and other characteristics: an application of geographically weighted spatial statistics, *International Journal of Geographical Information Science,* 24(8), 1193-1221.

McLaughlin, C. C. and Boscoe, F. P., 2007, Effects of randomization methods on statistical inference in disease cluster detection, *Health* & *Place,* 13(1), 152-163.

Messner, S. F., Anselin, L., Baller, R. D., Hawkins, D. F., Deane, G., and Tolnay, S. E., 1999, The spatial patterning of county homicide rates: An application of exploratory spatial data analysis, *Journal of Quantitative Criminology,* 15(4), 423-450.

Nødtvedt, A., Guitian, J., Egenvall, A., Emanuelson, U., and Pfeiffer, D. U., 2007, The spatial distribution of atopic dermatitis cases in a population of insured Swedish dogs, *Preventive Veterinary Medicine,* 78(3), 210-222.

Openshaw, S., Charlton, M., Craft, A. W., and Birch, J. M., 1988, Investigation of leukaemia clusters by use of a geographical analysis machine, *The Lancet,* 331(8580), 272-273.

Ord, J. K. and Getis, A., 1995, Local spatial autocorrelation statistics: distributional issues and an application, *Geographical Analysis,* 27(4), 286-306.

Pfeiffer, D. U., Robinson, T., Stevenson, M., Stevens, K. B., Rogers, D. J., and Clements, A. C., 2008, *Spatial Aalysis in Epidemiology* New York: Oxford University Press.

Ratcliffe, J. H. and McCullagh, M. J., 1999, Hotbeds of crime and the search for spatial accuracy, *Journal of Geographical Systems,* 1(4), 385-398.

Riguelle, F., Thomas, I., and Verhetsel, A., 2007, Measuring urban polycentrism: a European case study and its implications, *Journal of Economic Geography,* 7, 193-215.

Rogerson, P. A., 1999, The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic, *Geographical Analysis,* 31(1), 130-147.

Rogerson, P. A., 2005, A set of associated statistical tests for spatial clustering, *Environmental and Ecological Statistics,* 12(3), 275-288.

Rogerson, P. and Yamada, I., 2009, *Statistical Detection and Surveillance of Geographic Clusters*, CRC Press.

Scardaccione, G., Scorza, F., Las Casas, G., and Murgante, B., 2010, Spatial autocorrelation analysis for the evaluation of migration flows: the Italian case, *Computational Science and Its Applications-ICCSA 2010*, 62-76.

Tango, T., 1995, A class of tests for detecting 'general' and 'focused' clustering of rare diseases, *Statistics in Medicine,* 14(21-22), 2323-2334.

Tango, T. and Takahashi, K., 2005, A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics,* 4(1), 11.

Thomas, A. J. and Carlin, B. P., 2003, Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering, *Statistics in Medicine,* 22(1), 113-127.

Torabi, M. and Rosychuk, R. J., 2011, An examination of five spatial disease clustering methodologies for the identification of childhood cancer clusters in Alberta, Canada, *Spatial and Spatio-temporal Epidemiology,* 2(4), 321-330.

Trevelyan, B., Smallman-Raynor, M., and Cliff, A. D., 2005, The spatial structure of epidemic emergence: geographical aspects of poliomyelitis in north-eastern USA, July-October 1916, *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168(4), 701-722.

Waller, L. A. and Gotway, C. A., 2004, *Applied Spatial Statistics for Public Health Data*, John Wiley & Sons.

Walter, S., 1992, The analysis of regional patterns in health data I. Distributional Considerations, *American Journal of Epidemiology,* 136(6), 730-741.

Wang, F., 2004, Spatial clusters of cancers in Illinois 1986-2000, *Journal of Medical Systems,* 28(3), 237-256.

Yao, Z., Tang, J., and Zhan, F. B., 2011, Detection of arbitrarily-shaped clusters using a neighbor-expanding approach: A case study on murine typhus in South Texas, *International Journal of Health Geographics,* 10(1), 23.

Yeshiwondim, A. K., Gopal, S., Hailemariam, A. T., Dengela, D. O., and Patel, H. P., 2009, Spatial analysis of malaria incidence at the village level in areas with unstable transmission in Ethiopia, *International Journal of Health Geographics,* 8(1), 5.

Zhang, T. and Lin, G., 2008, Identification of local clusters for count data: a model-based Moran's I test, *Journal of Applied Statistics,* 35(3), 293-306.

Zhang, T. and Lin, G., 2009, Spatial scan statistics in loglinear models, *Computational Statistics* & *Data Analysis,* 53(8), 2851-2858.

교신: 조대헌, 151-742, 서울특별시 관악구 관악로1, 서울대학교 사범대학 지리교육과(이메일: dhncho@gmail.com)

Correspondence: Daeheon Cho, Department of Geography Education, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea (e-mail: dhncho@gmail.com)

최초투고일 2013. 10. 9
수정일 2013. 10. 18
최종접수일 2013. 10. 24