GIS 환경에서 통계적 공간분석 방안에 관한 연구

- 대기오염원과 호흡기 질환의 역학적 연구를 사례로 -

박기호, 유은혜
서울대학교 지리학과

1. 서론

1.1 연구 배경 및 목적

지금까지 GIS는 대부분 자료의 저장 · 검색과 간단한 조작과 같은 자료의 '운용(handling) 툴'로서 이용되어왔다. 그러나 점차 GIS가 가지고 있는 잠재적인 범용의 분석 도구로서의 가능성을 인정받으며, GIS 내에 지리적 자료의 분석적 이론과 모델링 기능과 같은 '통계적 공간분석' 기능에 대한 요구가 증가하고 있다[4]. 특히 위치 정보를 포함하고 있는 지리적 자료의 특성상 전통적인 통계 분석 기법의 적용은 통계적 모형의 유효성에대한 잘못된 판별력을 갖게함으로써 오인된 통계적 추론을 유도하기 쉽다[5] . 따라서 이와같은 지리적 자료의 분석에서는 반드시 '공간적 효과(spatial effect)'를 고려한 통계적 공간분석 기법이 요구된다[6].

최근 GIS 기술의 발달로 인해 이러한 문제가 점차 해결의 조짐을 보이고 있다. 즉, 기존의 통계 분석 모듈과 GIS와의 인터페이스를 마련함으로써, 현재 GIS에서 가장 취약한 기능인 통계적 분석 기능을 제공하고자 하는 노력이 다각적으로 이루어지고 있다.

이에 본 연구에서는 GIS와 통계 분석 모듈의 대표적인 연계 방안을 비교 · 분석하고, 통계분석 모듈인 S-plus[7]를 연계시킨 ArcView GIS[8]를 사용하여 지리적 자료의 '공간적 효과(spatial effect)'를 고려한 공간분석을 역학연구(Epidemiology)를 사례로 들어 수행하고자 한다.

1.2 자료와 분석 환경

사례 연구에 사용된 자료는 크게 두 가지로 분류할 수 있다. 1995년 당시 구별 행정 구역도와 당해 년도 센서스자료 및 천식 환자수와 같은 면 단위로 집계된 자료와 서울시의 20개소 대기오염 자동 측정망의 대기오염 자료이다. 센서스 자료와 상병 자료는 구별로 집계되었으며, 각 구별 총 인구수, 연령별 인구수(65세 이상, 14세 미만), 소득세 정보 등을 이용하였고, 연속 프로세스를 가정한 대기 오염 자료는 1995년 1년간 대기오염 자동 측정망에서 연속적으로 측정된 오존(O_3), 분진(TSP), 일산화탄소(CO), 이산화 황(SO_2), 이산화질소(NO_2) 등의 일별 평균치를 사용하였다.

한편, 본 연구에 사용된 통계적 분석 도구는 다양한 그래픽적 기능과 우수한 확장성을 가진 MathSoft' S-plus version 4.5와 이를 연계시킨 ESRI' ArcViewGIS version 3.0 이다.

2. GIS와 통계분석 모듈의 연계 방안

---

4) [Fisher, Henk, and Unwin, 1996]
5) [Fischer, Scholten, Unwin, 1996]; [Anselin, 1988]
6) [Cressie 1993; Anselin, 1994; Bailey and Gatrell 1995; Cook, 1996; Anselin and Bao 1997; Majure and Cressie, 1997]
7) MathSoft
8) ESRI

2.1 대표적인 GIS와 통계 모듈의 기술적 연계 방안

현재 대부분의 GIS 시스템에서 제공하고 있는 자료의 처리 · 집계 등과 같은 GIS의 분석 toolbox안에 통계적 분석 기능을 추가하기 위한 기술적인 연계 방안은 다음과 같이 약 세가지로 분류할 수 있다.

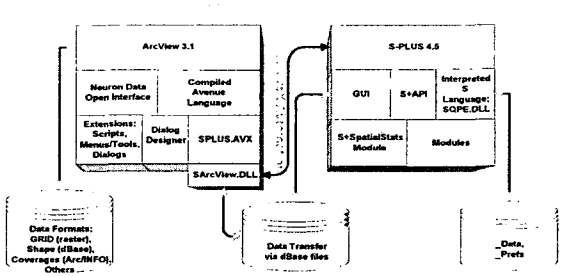
(1) loose-coupling : 가장 일반적으로 이루어지는 연계 방식으로서 동일한 파일 포맷을 따르는 자료나 명령어의 상호 교환 메커니즘에 의해 두 시스템의 연계가 이루어진다.

(2) close-coupling : 보다 진보된 연계 방식으로서 소스코드나 스크립팅 언어를 이용하여 두 시스템간의 인터페이스를 마련한다. 그러나 현재까지는 기술적 통계 분석 수준에서 머물고 있는 수준이다.

(3) fully-integrated : 가장 이상적인 연계 방안이나 현 시스템의 전반적인 구조 재편성이 필요하기 때문에 현실적 구현 가능성이 희박하다.

2.2. ArcView GIS와 S-plus의 연계

S-plus와 ArcView는 일종의 'loose coupling'의 연계방식 즉, [그림 2-1]에서 확인할 수 있는것과 같이 ActiveX Automation과 dBase 또는 이미지 파일 형태의 자료 교환 메커니즘에 의해 구동된다. 이 결과 사용자들은 S-plus의 다양한 일반 통계 기능과 S+spatialStat 모듈의 공간 통계 기능을 ArcView의 인터페이스를 통해 접근 · 이용할 수 있다.



[그림 2-11] 연계 메카니즘

그러나 서로 독립적인 개발환경과 목표하에 개발된 두 시스템의 근본적인 차이로 인해 진정한 동적 상호작용을 원활히 제공할 수 없다. 즉, S-plus의 'S'나 ArcView의 'Avenue'와 같은 자체적인 스크립트 언어에 대한 외부의 제어가 불가능하고, S-plus의 풍부한 그래픽과 통계 분석처리 기능을 제한적으로만 사용할 수 있을 뿐 아니라 S-plus에서 GRID, Coverage, Polygon과 같은 공간 객체에 직접 접근할 수 없기 때문에 S-plus에서 처리한 다양한 공간 통계 결과물을 ArcView에서 활용할 수 없는 한계가 있다.

2.3 공간 통계 기법에 관한 연구

지리적 자료는 공간 객체의 유형에 따라 점 · 선 · 면 등으로 구별되며, 속성 자료의 특성에 따라서는 이산적(discrete) 또는 연속적(continuous) 자료로 구분된다. 사실상 이와같은 자료의 특성에 따른 공간 통계의 이론과 기법들은 통계학외의 지질학이나 지리학, 경제학 등과 같은 분야에서 일찍부터 연구되어 왔다.

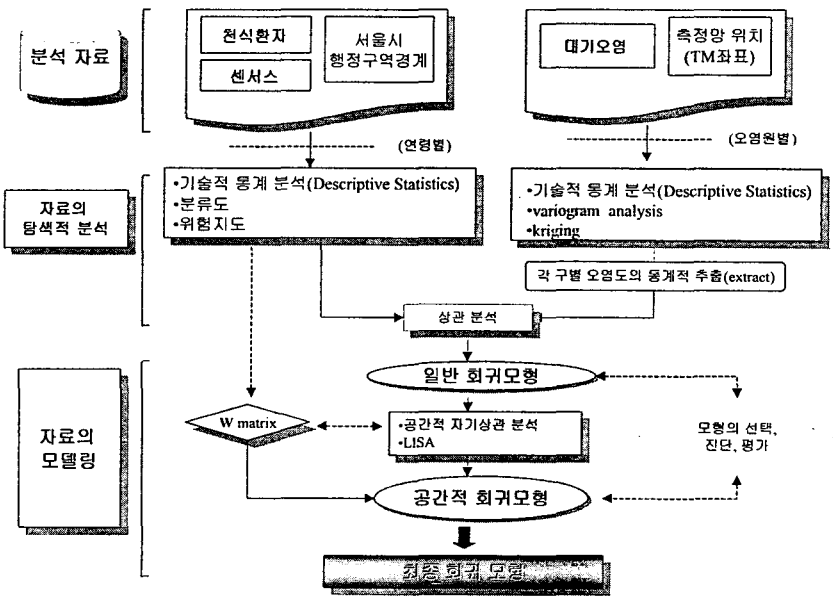
GIS의 '공간분석 컴포넌트'로서 통계적 공간분석 기능을 추가하기위해서는 GIS와의 연계를 통해 지리적 자료의 '공간성'을 보다 명확히 분석할 수 있는 정교한 분석 기법에 관한 연구가 필요하다. 최근 면 단위(areal unit) 자료의 국지적 공간 패턴에 관한 통계적 분석 기법인 LISA(Local Indicator of Spatial Association)가 바로 그 대표적인 예라고 할 수 있다. 이외에도 Nearest Neighbour Methods와 K-Functions과 같은 탐색적 자료분석(Exploratory Spatial Data Analysis)이나 Geostatistical and spatial econometric modelling에 이르는 자료의 모형 생성을 통한 공간분석 기법들을 들 수 있다.

3. 사례 연구

3.1 분석의 틀

본 사례연구는 대기 오염이 호흡기 질환에 미치는 영향을 평가하기 위한 기초 연구로서, 서울시의 25개 행정구 단위로 집계된 천식 환자와 대기오염의 분포 패턴을 탐색하고, 그 상관성을 추적하는데 목적이 있다.

분석 대상이 되는 자료와 그 처리과정은 그림 [3-1]과 같이 도식적으로 요약될 수 있다.



[그림 3-1] 분석 흐름도

3.2 분석 과정 및 결과

3.2.1 대기 오염원의 분포 패턴

1995년 1년 동안 집계된 서울시의 20개소 측정소 별 대기오염 농도로부터 각 구별 오염도를 얻기위해서는 우선 관측 지점을 기반으로 서울시 전역에 걸친 오염도 표면을 추정해야 한다. 추정 방법으로는 공간 통계학[9]이론에 기반을 둔 보다 체계적이며 객관적인 보간 방법인 크리깅(kriging)[10]을 사용하였다.

일단 기술적 통계 분석과 semivariogram analysis를 통해 자료의 분포적 특성을 파악한 후, 이를 기반으로 각각의 오염원별 특성에 따라 최상의 이론적 모형을 적합시켜 Kriging을 통해 오염 표면을 산출하였다. 이렇게 산출된 오염 표면과 오염 contour, 그리고 행정 구역도와의 중첩을 통해 각 구별 오염원의 통계치를 산출하였다. 한편, 오염 물질간의 상관 분석결과 No_2와 O_3, CO와 TSP 등이 강한 상관성을 나타낸다.

3.2.2 천식 환자의 분포 패턴

연령별 천식 환자의 지역적 분포 패턴을 파악하기위해 ArcView의 분류도와 S-plus의 EDA(Exploratory Data Analysis)툴을 이용하였다. 한편, 천식 환자는 전체 인구에 비해 상대적으로 드물게 발생하는 사건이기 때문에 상대적 위험도, 포아송 확률지도, 베이시언 추정도와 같은 모집단을 고려한 탐색적 분석 기법을 적용하였다.

[표 3-1] 천식 환자의 탐색적 자료 분석

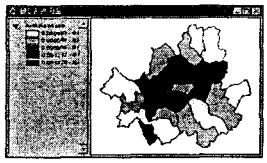
	14세 이하 인구	65세 이상 인구	전 연령 인구
원 자료(분류도)	동대문 > 중랑구	송파구 > 도봉구	동대문 > 중랑구
상대적 위험도 (Relative Risk map)	동대문 > 중랑구	도봉구 > 중구	동대문 > 중랑구
포아송 확률지도 (poisson probability map)	동대문 > 중랑구	도봉구 > 중구	동대문 > 중랑구
베이지언추정도 (Bayes estimation map)	동대문 > 중랑구	도봉구 > 중구	동대문 > 중랑구
인구 수	송파구> 노원구	성북구 > 송파구	송파구> 노원구

9) 'Spatial statistics' 또는 'Geostatistics'

10) 크리깅(kriging)은 확률변수의 공간적 변화성을 설명하는 변동도(variogram)에 따라 선택된 가중치와 자료간의 선형결합형태로 임의의 공간상 위치에서의 값을 추정하는 선형추정방법중의 하나로서 IDW나 Spline method 등과 같은 여타의 추정방법에 비해 훨씬 유연하고 정교한 공간 추정방법 중의 하나이다.

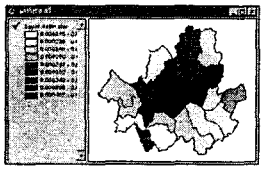
[표 3-1]에서도 알 수 있듯이 14세 이하 인구에서는 동대문구와 중랑구에서 많은 발생률(clustering)을 보였으며, 65세 이상에서는 도봉구가 많은 발병률을 보이고 있다.

[그림 3-2] 베이지언 추정도에 의한 위험 지역 추출



[그림 13-2-a] 14세 이하

[그림 3-2-b] 65세 이상



[그림 3-2-c] 전 연령

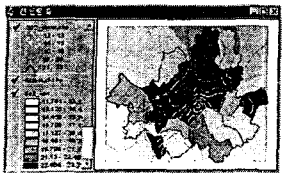
3.2.3 오염원과 천식 환자 수의 상관 분석

구별 오염도의 통계치와 연령별 상병 환자수간의 상관 분석(Pearson's correlation)은 [표 3-2]와 같이 전체적으로 그다지 유의한 결과를 나타내지는 않는다. 그러나 일부 오염원의 오염표면과 천식 환자의 위험도를 중첩 시킨 지도의 시각적인 비교를 통해 알 수 있듯이, 동대문을 중심으로 TSP나 SO_2 오염 표면간의 상관성을 확인할 수 있다.

[표 3-2] 오염원과 천식 환자수의 Pearson' Correlation Test

	O3	NO2	CO	SO2	TSP
O3	1.000000	0.752951	0.388288	0.273914	0.092761
NO2	0.752951	1.000000	0.420827	0.082122	0.248606
CO	0.388288	0.420827	1.000000	0.346200	0.413335
SO2	0.273914	0.082122	0.346200	1.000000	0.187781
TSP	0.092761	0.248606	0.413335	0.187781	1.000000
14 세 이하	0.015297	0.179262	**0.300722**	0.213237	0.253938
65 세 이상	0.122441	0.094675	0.065067	**0.212502**	**0.250310**
전 연령	0.088590	0.184253	0.222905	**0.325709**	0.159866

[그림 3-3] 중첩 지도



[그림 3-3-a] SO_2와 전 연령층

[그림 3-3-b] CO와 14세 이하

[그림 3-3-c] TSP와 65세 이상

3.2.4. 오염 회귀 모형(Ⅰ) (Non-spatial Regression Model)

천식 발병에는 직업이나 소득 수준과 같은 사회·경제적 지표, 유전적 요인, 대기 오염 이외의 많은 교란 변인등이 작용을 하기 때문에 3.2.3의 상관 분석결과에서 확인 할 수 있듯이 오염원과 천식 환자 수사이의 직접적인 상관성을 기대하기는 힘들다. 이에 본 연구에서는 먼저 각 구별 소득세를 설명 변수로 설정하여 기본적인 경제적 변인을 제어하였다.

한편, [그림 3-2]의 위험도와 S-plus의 다양한 EDA 결과에 의하면, 14세 이하와 전 연령층의 천식 환자가 지역적으로 상당히 왜곡 분포를 이루고 있음을 확인할 수 있다. 따라서 이러한 자료에 대해 회귀 모형 산출을 위해서는 "Freeman-Turkey Square-Root[11]"와 같은 자료의 변형이 필요하다. 또한 14세 이하와 전 연령층에 대해서는 동대문구를 이례적인 지역으로 처리하여 회귀모형에서 제외하였다.

기본적인 교란 요인을 제거한 모형에 3.2.3의 상관 분석에서 어느정도 유의성을 드러낸 오염원을 설명 인자로 추가하여 회귀 분석을 실시하였다. 분석 결과는 [표 3-3]에 정리되어 있다. [표 3-3]에서 알 수 있듯이 전 연령 층에 대해 CO는 약 10%의 설명력을, 14세 이하의 연령 층에 대해 TSP는 약 16%의 설명력을 보이며, 65세 이상의 연령에 대해 SO_2는 약 11%를 설명하고 있다.

---

11) count 자료가 근본적으로 내포하고 있는 "mean-dependence와 skewedness"를 최소화할 수 있는 자료의 변형 방식으로써 그 식은 다음과 같다.

$Y_i = \sqrt{1000}(\sqrt{S_i/n_i} + \sqrt{(S_i+1)/n_i})$

[표 3-3] 오염 회귀 모형(I)

전 연령	y = 711.9665 -1.2901X R^2 : 0.311, RMSE : 76.31, df : 23 **y = 443.20 -1.3X + 15.2X1** (X1 : SO2) R^2: 0.41, RMSE : 72.05 df : 22
14 세 이하	y = 469.72 - 1.17X R^2 : 0.267 , RMSE : 76.91 df : 23 **y = 292.42 - 1.50X + 15.2X1** (X1 : CO) R^2 : 0.42, RMSE : 70.21 df : 22
65 세 이상	y = 357.5 - 0.21X R^2 : 0.03, RMSE : 44.71 df : 23 **y = 781.60 - 0.24X - 8.12X1** (X1 : TSP) R^2 : 0.10, RMSE : 44.94 df : 22

(X : 소득세)

3.2.5 공간적 자기상관[12] // 공간적 상관[13] 분석

공간적 자기상관은 지도에 나타난 공간적 패턴을 일반화하고 정량화하기위한 척도로서 뿐만 아니라 공간적 프로세스를 통계적으로 모형화하기위한 일종의 탐색적 분석과정으로 이용될 수 있다. 특히, 일반적인 다중 회귀모형[14]의 잔차에 대한 공간적 자기상관 분석은 연구 대상 지역의 프로세스를 선형 공간 회귀 모형(spatial autoregressive process)에 적합시키기위한 준거(Criteria)설정 과정으로서 최적의 통계적 예측 모형[15]을 산출하기위해서는 반드시 필요한 분석 과정이다.

한편, 연구 대상지역의 공간적 자기상관은 공간 구조의 정의 유형[16](Spatial Neighbor)과 가중치[17](Weight)에의해 많은 영향을 받는다. 따라서 본 연구에서도 이러한 점을 고려하여 차수[18]와 거리[19]를 차별적으로 정의한 공간 이웃 구조에 근거해 3.2.4의 회귀 모형에서 추출된 잔차와 Freeman-Turkey 변형을 거친 천식환자 수에대해 공간적 자기상관 분석을 실시하였다. 지도를 통한 분포 패턴에서 예측할 수 있듯이 천식환자 수의 경우 14세 이하와 전 연령층에대해서는 유의한 수준의 공간적 자기상관을 확인할 수 있었으며, 일반 다중회귀 모형의 잔차에 대한 분석결과는 [표 3-4]와 같다.

12) Spatial Autocorrelation 또는 Global Spatial Association
13) Spatial Association
14) Non-Spatial Regression
15) "reculsive and BLUE(Best Linear Unbiased Estimation) procedure" (Ripley, 1981, p. 100)
16) order(first order, second order ...), distance, direction ...
17) row-standardization, length of common boundary, distance function ...
18) first neighbor, second neighbor
19) 5500 // 6000 // 6500 // 7000 // 7500 // 8000 m (일반적으로 인접한 이웃 중심까지의 거리는 5500 이상)

[표 3-4] Spatial Autocorrelation (Moran'I)

전 연령 (5500 m)	Moran'I = 2.05 , p-value = 0.39 Moran'I = 0.45 , p-value = 0.652
14 세 이하 (5500 m)	Moran'I = 2.77, p-value = 0.005 Moran'I = 1.5 , Normal p-value = 0.133
65세 이상(5500 m)	Moran'I = 0.9 , p-value = 0.366 Moran'I = 0.23 , p-value = 0.821

(천식환자수; 회귀 모형의 잔차)

한편, 이상의 global한 스케일의 공간적 자기상관 분석과 달리 local한 스케일의 공간적 상관분석 (LISA[20])을 통해 살펴본 각 연령별 지역적 상관 구조와 위험 지역은 다음과 같이 요약될 수 있다. LISA에서 유의한 결과를 보이는 지역들은 [표 3-2]와 같다.

[표 3-5] LISA를 통한 위험 지역 추출

		Local Moran	Local Geary
전 연령	동대문구	0.85(1.97)	14.8(3.35)
	중랑구	0.12(2.07)	2.84(1.68)
	중구	-0.57(-2.21)	7.94(-2.21)
65세 이상	송파구	0.90(1.78)	2.27(1.78)
	도봉구	0.57(1.68)	2.84(1.67)
	종로	-0.93(-2.78)	8.06(-2.78)
14세 이하	동대문구	0.5414(3.469)	2.789(1.91)
	중랑구	2.026(1.909)	13.56(3.49)
	중구	-0.666(-1.864)	7.08(-1.86)

전 연령과 14세 이하에 대해 동대문구와 중랑구와 같이 평균 발병률보다 높은 지역은 Local Moran이나 Local Geary 모두 (++)로서 정적인 공간적 상관성을 보여주고 있다. 반면, 중구의 경우 각각의 지수가 (-,+)로 부적인 공간적 자기상관을 보임을 알 수 있다. 이러한 지수의 변이를 통해 중구를 중심으로한 주변 지역의 천식환자 발병률의 국지적인 변이 패턴을 찾아볼 수 있다.

3.2.6 오염 회귀 모형(II) (Spatial Regression Model)

공간적 자기상관 분석을 통해 밝혀진 바와같이 일정한 경향성을 보이는 천식환자 자료의 공간 프로세스를 모형화하기위해서는 Autoregressive Covariance 모형에 기반을 둔 공간 회귀 모형이 보다 적합할 것으로 판단된다.

한편, 공간 회귀 모형을 산출하기위해서는 모형의 설명 변수, 공간 구조(Spatial Neighbor), 가중치(Weight), Covariance Family[21] 등을 고려해야한다.

20) Local Indicator of Spatial Association

본 연구에서는 각 연령층에 대해 단일 오염원을 설명변수로하고, 거리와 차수를 공간 구조로하며, CAR 모형을 기본 모형으로하는 각각의 공간 오염 회귀 모형을 산출하였다. [표 3-4]의 공간적 자기상관 분석 결과를 통해 확인할 수 있듯이 공간적 자기상관 구조를 보이는 전 연령과 14세 이하의 회귀모형은 일반 회귀모형의 계수와 상당한 차이를 보이고 있다.

[표 3-6] 연령별 공간 회귀 모형

전 연령	y = 440.62 -1.30X + 15.39X1 (X1 : SO2) rho = 0.01 , Log-likelihood = -139.2
14 세 이하	y = 113.62 - 0.0001X + 31.18 (X1 : CO) rho = 0.37 , Log-likelihood = -140.2
65 세 이상	y = 782.18 - 0.24X - 8.12X1 (X1 : TSP) rho = -0.01 , Log-likelihood = -133.8

한편, 다중 오염원을 설명 변수로 추가할 경우 최적의 다중 공간 회귀 최적의 모형을 선택하기위해서는 'Likelihood Ratio Test[22)]'를 이용할 수 있다.

3.2.7 회귀 모형의 진단

추정된 회귀모형의 신뢰도를 위해서는 회귀 모형에 대한 잔차분석을 실시해야 한다. 한편, 공간 회귀 모형의 잔차는 CAR, SAR, MA와 같은 covariance model에 따라 상이하게 나타난다. 그러나 각각의 모형에 대한 잔차는 정규성, 등질성(homogeneity), 이례지점 등과 같은 일반적인 잔차분석을 거친다. 각 연령별 공간 회귀 모형에 대한 잔차분석 결과 몇몇 이례적인 지점을 제외하고 가정에 크게 위배되지 않았다.

4. 결론

본 연구는 크게 두가지의 의의를 가지고 있다.

첫째, '공간성'을 고려한 일련의 공간 통계 분석기법을 적용한 결과 기존의 시계열적 분석과 달리 천식환자와 대기오염원의 지역적 분포패턴을 파악할 수 있었으며, 일반적인 회귀모형에서 간과하고 있는 '공간적 변이'를 고려한 새로운 통계적 예측 모형을 산출할 수 있었다. 특히, 14세 이하의 천식환자의 경우 지역적 발병률이 상당한 공간적 자기상관을 보이고 있었으며, 이러한 요인을 고려한 통계모형을 적합시킨 공간회귀모형과 일반회귀모형에 의한 예측치를 비교해 볼 때 공간 회귀모형에의해 보다 설명력 있는 모형을 산출할 수 있었다.

둘째, S-plus와 ArcView GIS의 연계 방식을 사례로 GIS 환경에서 통계적 공간분석 가능성을 확인할 수 있었다. 그러나 이와같은 'loose coupling' 방식의 연계는 두 시스템간의 동적인 상호작용을 지원하지 못하기 때문에 진정한 통계적 분석 기능을 완벽히 구현하기가 어렵다. 따라서 지속적인 연계 방안에 관한 후속 연구가 필요하며, 동시에 지리적 자료에 적합한 '통계적 공간 자료 분석 기법'에 관한 끊임없는 관심과 연구가 이루어져야 할 것이다. 특히, 방대한 자료를 다루게 되는 GIS의 본질적인

21) CAR, SAR, MA

22) $U^2 = 2\times\frac{(n-p-r)}{n}\times(L_p - Lp + r)$ (Cressie, 1993)

특성을 고려해 볼 때 ESDA와 같은 자료 분석과정은 사용자의 수준에 관계 없이 사용될 수 있을 뿐 아니라 컴퓨팅 자원을 합리적으로 이용할 수 있는 방안으로서 GIS 의 통계적 분석 기능 중에서도 반드시 필요한 기능이라고 사려된다.

ㅁ 참고 문헌

Anselin, L. (1998) "Exploratory spatial data analysis in a geocomputational environment"

Bailey, T.C. and Gartell (1995) "Interactive spatial data analysis" Harlow, Longman Scientific and Technical.

Getis and Ord (1996) "Local Spatial Statistics : an overview", Spatial Analysis : Modelling in a GIS Environment (Eds) Paul Longley and Michael Batty.

Griffith, D.A (1993) "Spatial Regression Analysis on the PC : Spatial Statistics Using SAS"

Issaks, E.H. and Srivastava, R.M. 1989, "Applied Geostatistics", oxford University Press, New York.

Trevor, C. Bailey, (1994) "A review of statistical spatial analysis in geographical information systems", Spatial Analysis and GIS. (Eds) Stewart Fotheringham and Peter Rogerson, Taylor & Francis.

Upton and Fingleton(1985) "Spatial Data Analysis by example", Point pattern and Quantitative data Vol(1), Wiley, Toronto Singapore, Brisbane, New York, Chichester.

Ripley, B.D. (1981) "Spatial Statistics", John Wiley, Chichester, Chapters 6-8.

Stephen, P. Kauzny, Silvia, C. Vega, Tamre,P. Cardoso, Asice, A. Shelly, (1996) S+SpatialStats User's Manual Version 1.0 MathSoft, Inc.