성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)

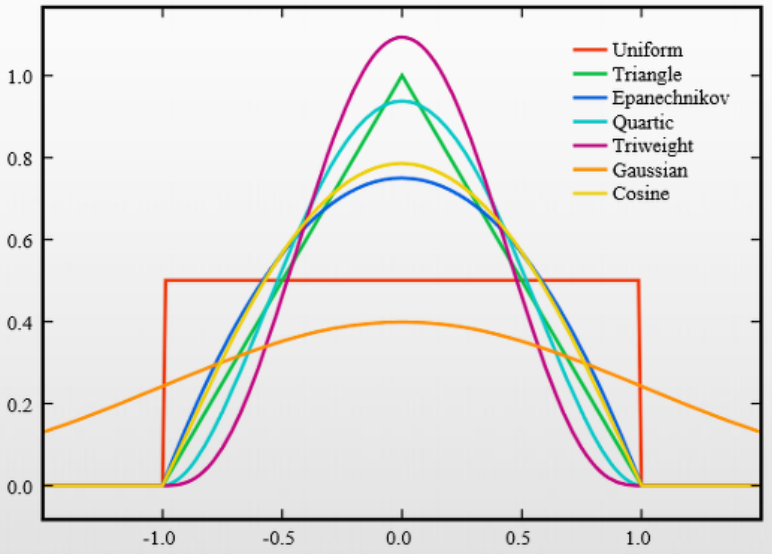
위치기반데이터 분석(Location based Data Analytics)

공간통계 분석(2)

Kernel Density Estimation

커널함수(Kernel function)를 이용한 밀도추정 방법

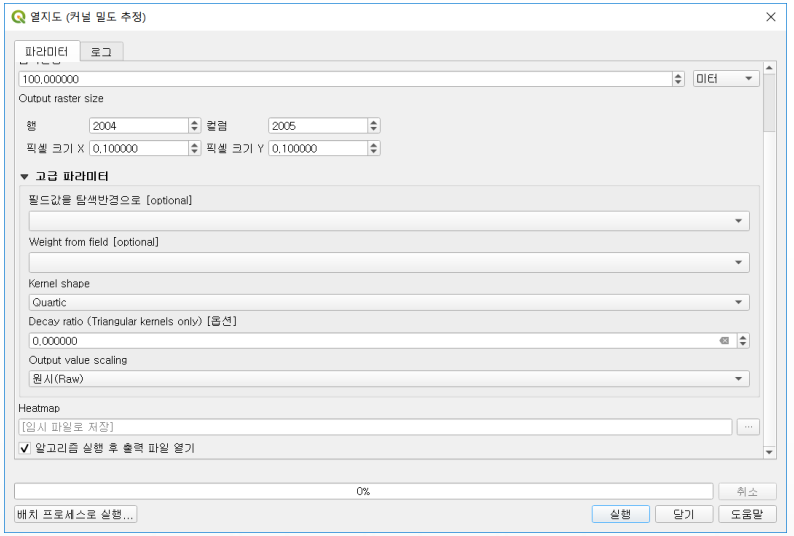
- 데이터는 어떠한 현상을 관측한 값이다 . 통계학에서는 관측치를 바탕으로 모집단의 특성을 유추하려고 하며 따라서 데이터는 모집단의 샘플이 된다.
- 불행히도 이러한 샘플은 현실에 대한 값만을 나타낼 뿐 모집단의 성질을 설명해 주지 않는다
- 샘플을 통한 변수의 성질을 알기 위해 다양한 통계적 기법이 동원되어 현상에 대한 특성을 유추
- 밀도추정 : 데이터와 변수의 관계를 파악하는 방법, 데이터로 부터
- 변수가 가질 수 있는 모든 값의 밀도(확률)을 추정
- 커널함수 : 원점을 중심으로 대칭이면서 적분값이 1인 non_negative 함수로 Gaussian, Quantic, uniform 함수 등이 대표적인 커널 함수들이다.
- GIS에서 커널 밀도추정 방법을 활용하여 밀도지도를 생성하여 활용



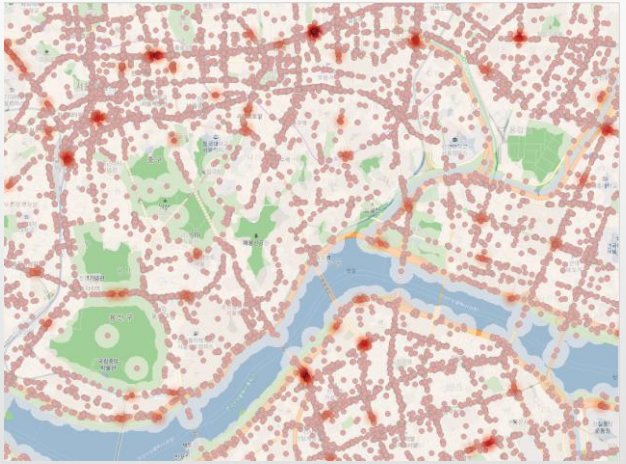
Kernel Density Estimation

열지도(heatmap) 실습

- 서울시 교통사고지점 데이터를 통해 교통사고가 많은 지점에
- 대한 HeatMap을 생성
- sago_2013.shp 열기
- 공간처리 툴박스에서 보간법 -> 열지도(커널 밀도 추정)
- Weight from field[optional]에 사망자 수로 가중치를 부여
- 필셀크기는 20m 단위



탐색반경 500m



탐색반경 200m

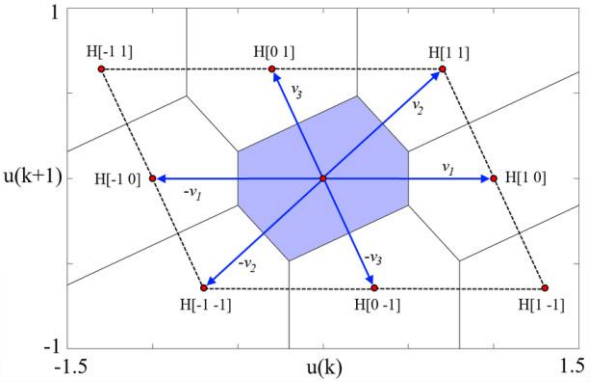
패턴분석

1. 티센폴리곤(Thiessen Polygon)

- 미국의 기상학자 Alfred H. Thiessen에 의해 명명됨
- Voronoi 다각형을 통해 알려진 값의 집합을 주어진 영역으로 나누는 기하학적 방법

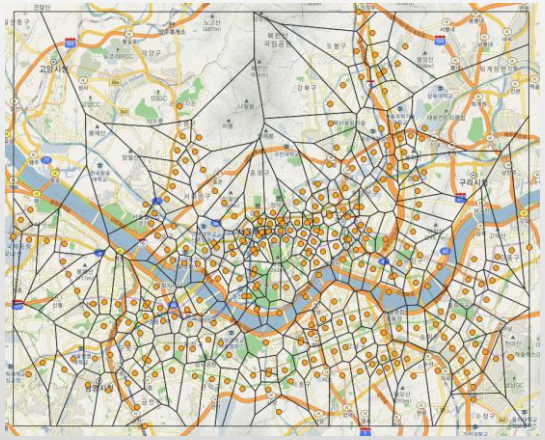
*보로노이 다각형 : 점이 공간에 배치되어 있을때 공간을 각각의 점에 가장 가까운 영역에서 분할했을 때 얻어지는 다각형

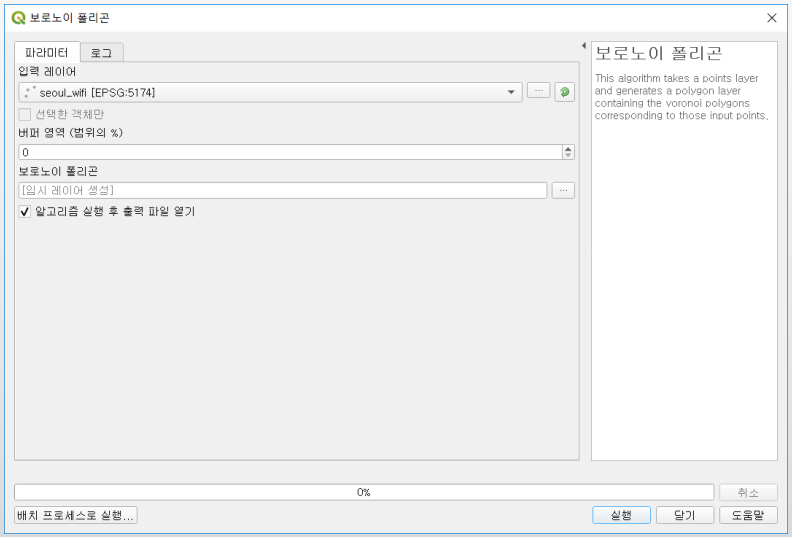
*Delaunay triangles : 삼각형을 정의하는 3개의 점을 제외한 어떤 점도 삼각형을 감싸고 있는 원에 포함되지 않게 삼각형



실습

- Seoul_subwaystation.shp 파일을 통해 지하철의 최소단위 영역을 파악해 보기
- 공간처리 툴박스 -> 벡터 도형 -> 보로노이 폴리곤



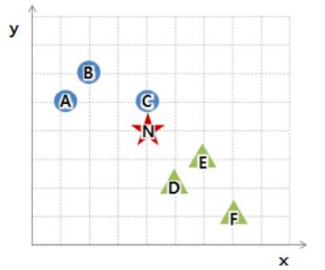


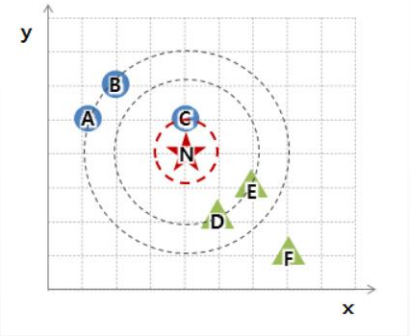
패턴분석

2. K-Nearest neighbor

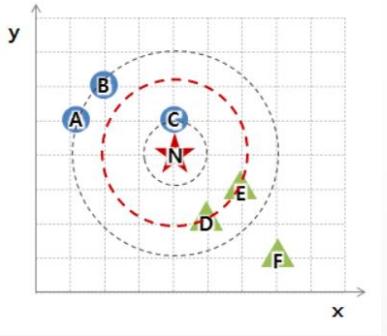
- 패턴인식에서 분류나 회귀에 사용되는 비모수 방식으로 가장 간단한 기계 학습 알고리즘에 속함
- 거리함수에 의해 가장 가까운 이웃 K개를 찾는다

데이터	x좌표	y좌표	그룹
A	1	5	●
B	2	6	●
C	4	5	●
D	5	2	▲
E	6	3	▲
F	7	1	▲
N	4	4	?

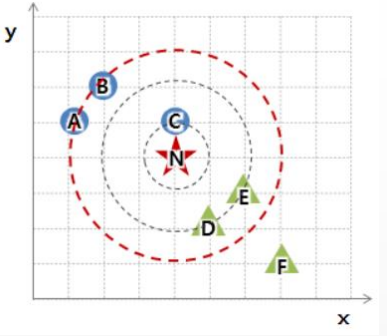




K=1인경우



K=3인경우

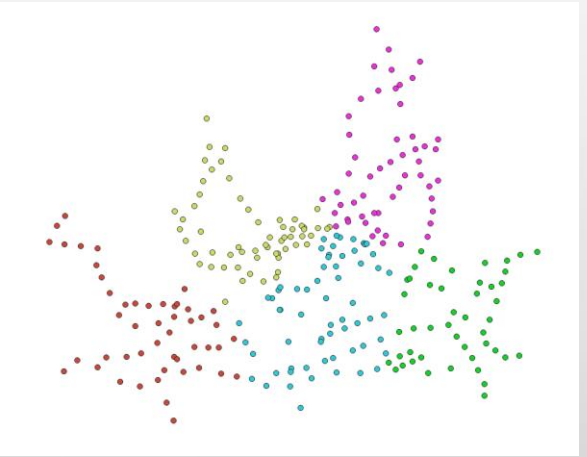


K=5인경우

실습

- Seoul_subwaystion_TM.shp을 이용하여 서울시 지하철역사를 5개의 지역적 그룹으로 패턴분석

공간처리 툴박스 -> 벡터분석-> K-means 클러스터링



공간적 자기 상관관계(Spatial autocorrelation)

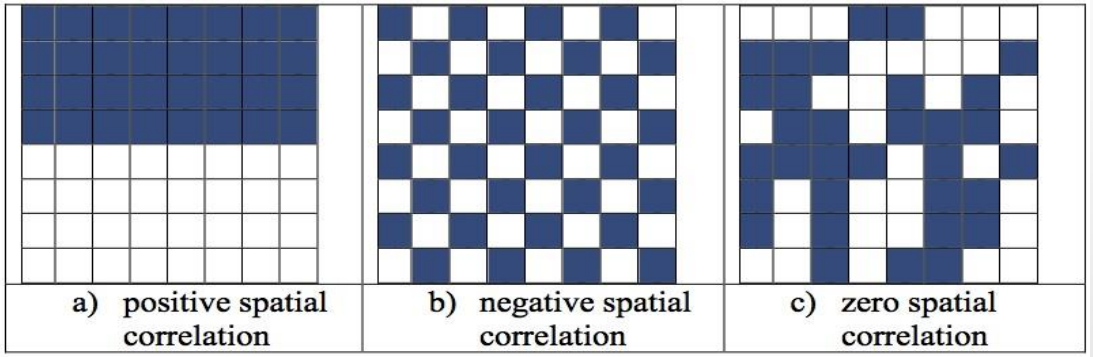
공간적으로 인접할수록 유사한 특성을 지니게 되는 상관성을 분석

1. 공간적 자기상관

- 특정 현상이 그 지역의 다른 현상과 연관하여 나타나기 보다는 해당지역에서 나타나고 있는 현상자체가 다른 지역과 연관되어 있다는 추론
- Moran's I, Geary C 지수, Getis Ord Gi, Local Moran's I 분석 등이 사용

2. Global Moran's I

- 일련의 피쳐와 관련 속성이 주어지면 표현된 패턴이 클러스터화, 분산, 무작위인지의 여부를 평가



a) I = 1

b) I = −1

c) I = 0

Moran's I	Patten
I > 0	공간객체는 공간상에서 군집되어 있으며 비슷한 속성을 가짐
I = 0	공간객체는 공간상에서 랜덤하게 존재하며 비슷한 속성을 가지지 않음
I < 0	공간객체는 공간상에서 분산되어 존재하며 높은 값과 낮은 값들이 점철되어 있음

공간적 자기 상관관계(Spatial autocorrelation)

3. Local Moran's I

- Local Morans's I 통계를 사용하여 통계적으로 중요한 핫스팟, 콜드스팟, 및 공간 아웃라이어를 식별
- 입력피쳐에 대해 로컬 모란 I 지수, z-score, p-value, COType 과 같은 특성을 가진 새로운 출력 클래스를 생성
- z-score 양의 값이 클수록 Hot spot이 생성되며 음의 값이 작을수록 Cold spot이 생성
- P-value는 관찰된 데이터의 검정통계량이 귀무가설을 지지하는 정도의 확률 표현 p-value 가 낮을수록 관찰된 데이터의 신뢰성이 높다고 할 수 있다

* 귀무가설 : 설정한 가설이 진실일 확률이 극히 적어 처음부터 버릴 것이 예상되는 가설

- COType

HH (cluster of high values) -> 핫스팟

LL(cluster of low values) -> 콜드스팟

HL -> 나는 높지만 내 주변이 낮은 경우

LH -> 나는 낮지만 내 주변이 높은 경우

Spatial Outlier (공간적 이상치)

공간적 자기 상관관계(Spatial autocorrelation)

4. Geary's C

- 같은 형상의 인접 관측이 상관관계가 있는 경우에 공간적 자기 상관 측정하는 방법
- Moran's I와 역으로 관계되어 있으며 Moran's I는 공분산의 개념에 기초한 지수인 반면, Geary's C는 분산의 개념에 기초
- 공간가중치에 의해 이웃으로 정의된 분석단위들이 얼마나 유사한 속성 값을 갖는지를 직접적으로 비교하는 척도

Geary's C	Patten
0 < C < 1	공간객체는 공간상에서 군집되어 있으며 비슷한 속성을 가짐
C = 1	공간객체는 공간상에서 랜덤하게 존재하며 비슷한 속성을 가지지 않음
C > 1	공간객체는 공간상에서 분산되어 존재하며 높은 값 낮은 값 점철되어 있음

5. Getis-Ord Gi*

- 연구지역내의 공간객체의 z-score(통계학적으로 의미 있는 정도)를 계산하여 높은 값과 낮은 값들의 집중도를 보여줌
- 핫스팟 분석도구로 사용
- z-score의 양의 값이 클수록 높은 값의 더 강렬한 클러스터링(hot spot)
- z-score의 음의 값이 작아질 수록 낮은 값의 더 강렬한 클러스터링(cold spot)