

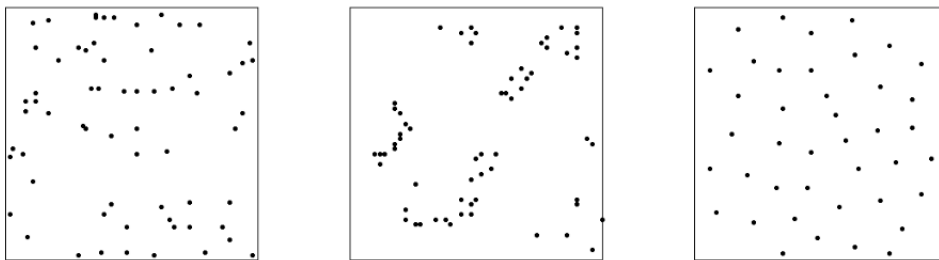
SPATIO TEMPORAL DATA ANALYSIS

WEEK 9: MODELING SPATIAL PATTERNS

공간 데이터는 세 종류로 나눌 수 있는데, point referenced, areal, spatial point process가 그것들이다. point referenced 데이터는 데이터의 정확한 위치 정보를 가지고 있다. 즉, 데이터가 발생한 위도, 경도 위치 정보를 담고 있다. areal 데이터는 위도, 경도 위치를 담고 있지 않고 '지역' 정보를 담고 있다. 예를 들어, 서울의 각 구별 미세먼지 데이터는 정확한 위도, 경도가 아닌, 서대문구, 영등포구 등 지역별 미세먼지 정보를 담고 있기 때문에 areal 데이터라고 볼 수 있다. 이번에 살펴볼 데이터는 마지막으로 남은 spatial point process이다. 이는 데이터들의 분포가 spatial dependence을 가지고 있는지에 관심이 있다. 즉, 데이터별로 response variable이 있었던 point referenced, areal 데이터와는 달리, 데이터들이 어떻게 분포되어있는지 자체에 관심이 있다. 물론 이 데이터들이 추가적인 정보를 가지고 있을 수 있다. 예를 들어, 서울 서대문구의 상점들의 위치에 대한 공간 모델링을 하고 싶을 때, 상점들의 위치 뿐만 아니라 매출액 등의 정보도 사용할 수 있다. 이제 spatial point process에 대해 자세히 알아보자.

먼저 spatial point process와 spatial point patterns의 차이를 짚고 넘어가자. spatial point process는 point patterns을 발생시키는, 내재하는 매커니즘이다. 따라서 관측된 것은 point patterns이고 이를 설명하기 위해 spatial point process 모델을 사용한다.

point process에 패턴이 없다면, complete spatial randomness (CSR)이라고 부른다. CSR은 spatial point process의 null model이라고 생각하면 된다. CSR이 아니라면, clustering, regularity와 같은 특징을 가진다.



위 그림은 왼쪽부터 차례대로 CSR, clustering, regularity의 특징을 보이는 데이터이다. CSR은 아무런 패턴이 없고 clustering은 점들이 모여있으며 regularity는 점들이 일정한 간격을 띄고 분포되어 있다.

point process model에 대한 추론은 앞서 살펴본 세 종류의 공간 데이터 중, 가장 어려운 문제이다. 전과 마찬가지로 기술적인 측면과 복잡한 모델링, 두 가지가 있는데 우선 기술적인 측면부터 살펴보자.

Homogeneous Poisson Process

- CSR을 만족한다. 따라서 null model에 쓰인다.
- $N(A)$ 를 지역 A 에서 일어나는 사건의 수, $|A|$ 을 그 지역의 크기라고 하자.
- Homogeneous Poisson Process는 아래와 같이 정의된다.

$$N(A) \sim \text{Poisson}(\lambda|A|)$$

여기서 λ 는 intensity parameter로 이 값이 클수록 $N(A)$ 의 값이 커지니까 더 많은 데이터가 지역 A 에서 생성된다.

- 또는 HPP가 CSR을 만족하므로, 아무 패턴이 없는 데이터라고 볼 수 있다. 즉, uniform 분포를 따르는 n 개의 iid 샘플을 지역 A 에서 뽑는 것과 동일하다.

Diagnosing CSR

통계량 U 의 관측된 값을 u_1 이라고하자. 마치 t test에서 하나의 test statistics을 얻은 것과 같은 느낌이다. 그런데 t test에서는 이의 표본 분포를 구하기가 매우 쉬웠지만 여기서는 그러기 어렵다. 따라서 s 개의 test statistics의 표본을 귀무가설 하에서 독립적으로 뽑는다. 즉, CSR가정 하에서 s 개의 test statistics, u_1, \dots, u_s 을 뽑으면,

$$P(u_1 \text{ has rank } j) = 1/s$$

실제 관측된 것은 u_1 이고 나머지는 모두 같은 분포에서 뽑았으므로 u_1 이 어떠한 rank를 가지든, 그 확률은 모두 동일하다. 이러한 아이디어로 CSR을 테스트하는 것이다. 그렇다면, 검정을 하는데 쓰이는 test statistics의 구체적인 형태를 살펴보자.

test statistics을 도출하기 위해, 사용하는 것 중 하나가 empirical cdf이다. empirical cdf는

$$\hat{F}(x) = \frac{\sum 1\{X_i \leq x\}}{n} = \frac{\#\{X_i \leq x\}}{n}$$

이는 population cdf의 불편 추정량이다.

$$F(x) = P(X_i \leq x) = \int_{-\infty}^x f(y)dy = \int_{-\infty}^{\infty} 1\{y < x\} f(y)dy$$

또 데이터들 간의 정의된 거리를 이용한다.

- Pairwise distances: $s_{ij} = \|x_i - x_j\|, i \neq j$
- Nearest neighbor distances: $t_i = \min_{j \neq i} s_{ij}$

- Empty space distances: $d(u) = \min_i \|u - x_i\|$, $i \neq j$. 여기서 u 는 관측되지 않은 값, x_i 는 관측된 값이다.

먼저 empty space distances와 ECDF을 이용하여 CSR을 확인하는 방법을 알아보자.

Empty Space Distances

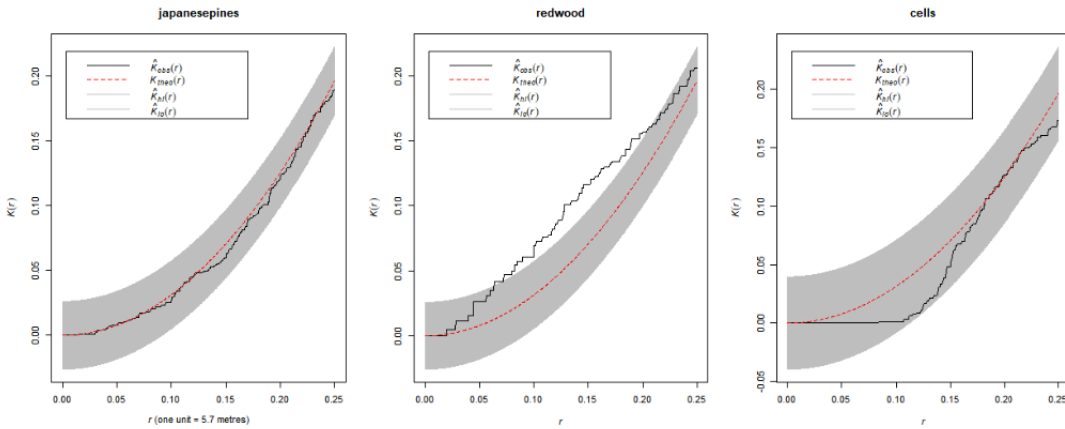
관심있는 공간 전체에서 $d(u)$ 을 살펴보기 힘들니까, 특정 지역 A 에 대해서만 $d(u)$ 을 살펴보고 $d(u)$ 와 관련된 process 있다고 생각한다. 아래 cdf을 생각해보자.

$$\begin{aligned} F_u(r) &= P(d(u) \leq r) \\ &= P(\text{at least one point within radius } r \text{ of } u) \\ &= 1 - P(\text{no points within radius } r \text{ of } u) \end{aligned}$$

즉, 관측되지 않은 어떤 한 점 u 을 기준으로 반지름이 r 인 원을 만들었을 때, 그 원의 영역에 점이 하나도 떨어지지 않을 확률은 intensity parameter λ 와 그 원의 넓이 $|A| = \pi r^2$ 을 이용하여, $\frac{(\lambda \pi r^2)^0 \exp\{-\lambda \pi r^2\}}{0!}$ 로 주어지고, $F(r) = 1 - \exp\{-\lambda \pi r^2\}$ 이다. 이 확률은 u 에 의존하지 않는다. 또한 n 을 조건으로 하여, λ 을 $\hat{\lambda} = n/|A|$ 로 추정한다면 비교를 위해 유용한 베이스라인을 얻을 수 있다. 이렇게 true ECDF에 대한 추정값을 얻어서, empty space distances로 구한 ECDF와 비교를 한다. 즉, u_1, \dots, u_m 에 대해 $d(u)$ 에 기반하여

$$\hat{F}(r) = \frac{1}{m} \sum_{j=1}^m 1\{d(u_j) \leq r\}$$

을 구하고 $F(r) = 1 - \exp\{-\lambda \pi r^2\}$ 와 비교한다. 이때 simulation envelopes을 만들기 위해서 Monte Carlo Sampling을 이용한다. 지역 $|A|$ 에서 n 개의 위치 표본을 uniformly 뽑는다. 이 과정을 s 번 반복하여 rank 기반 confidence band을 생성한다.



위 그림은 왼쪽부터 차례대로 CSR, clustering, regular 가정을 만족하는 데이터에 ECDF인 $\hat{F}(r)$ 과 CSR 가정 하에서의 이론적 cdf인 $F(r)$ 을 비교한 그림이다. 좌측 데이터가 CSR을 만족하는데,

이론적 cdf와 가장 유사한 형태를 보였다. 가운데 데이터는 clustering을 만족하는데, ECDF가 이론적 cdf보다 큰 값을 가짐을 알 수 있다.

Nearest-neighbor distances

empty space distances와 유사하게 진행하면 된다. $t_i = \min_{j \neq i} \|X_i - X_j\|$ 을 이용하여 이론적 cdf인 $G(r) = P(t_i \leq r)$ 와 ECDF인 $\hat{G}(r) = \frac{1}{n} \sum_{i=1}^n 1\{t_i \leq r\}$ 을 비교한다.

여태까지 살펴본 Homogeneous Poisson Process은 CSR과 동일한 의미로, intensity parameter가 지역에 의존하지 않았다. 즉 지역 A 의 poisson 분포 $N(A)$ 는 평균이 $\mu(A) = \int_A \lambda dx = \lambda|A|$ 이다. 또한 $N(A) = n$ 으로 주어질 때, 이 n 개의 데이터는 지역 A 에서 n 개의 표본을 균등하게 생성하는 것과 동일함을 살펴보았다. 이제는 지역에 따라서 intensity parameter가 다른 process을 보자.

Inhomogeneous Poisson Process

$$N(A) \sim \text{Poisson}(\mu(A))$$

$$\mu(A) = \int_A \lambda(x) dx$$

즉, 지역에 따라서 intensity parameter가 다르기 때문에 한 지역에서 균등하게 데이터를 생성하는 것과는 다른 의미를 가진다. 보통 균등하게 데이터를 생성하고, 확률에 따라서 데이터를 지우거나 유지하여 Inhomogeneous Poisson Process의 표본을 만든다.

- $\lambda_{max} = \max_{x \in A} \lambda(x)$
- λ_{max} 을 intensity parameter로 하여 homogeneous process 표본을 얻는다.
 - $n \sim \text{Poisson}(\lambda_{max}|A|)$
 - Draw $X_1, \dots, X_n \sim \text{Unif}(A)$ independently
- $i = 1, \dots, n$ 에 대해서 X_i 을 확률에 따라서 유지한다.
 - $p(x_i) = \lambda(x_i)/\mu(A)$, $\mu(A) = \int_A \lambda(x) dx$

근데 현실에서는 $\lambda(x)$ 을 모르므로 추정해야한다. 하지만 관측된 데이터가 한 개 뿐이라서 Quadrat counting, kernel density 방법을 사용한다.

다음에는 Inhomogeneous Poisson Process을 모델링하는 방법에 대해서 살펴본다. likelihood에 기반한 방법과, 베이지 방법에 대해서 자세하게 살펴본다.