# Spatio Temporal Data Analysis

# Week 1: Overview and Introduction

## What is Spatial Data?

통계학이 다루는 자료들의 종류는 정말로 다양하다. 그 중, 가장 기본적이면서 현실에서 잘 일어나지 않는 데이터 구조는, $i.i.d$ 데이터이다.

*i.i.d: identically, independently distributed data*

$n$개의 데이터에 대해서, 이들이 모두 동일한 분포에서, 그리고 독립적으로 생성되었다는 가정이다. 예를 들어, $n$명의 졸업생의 GPA 데이터가 있을 때, 이 학생들의 GPA가 서로 독립적으로 생성되었다고 본다. 근데 과연 이러한 과정이 올바를까? 서로 친한 학생들끼리 많이 놀거나 같이 공부를 열심히 한다면 GPA에 영향이 있을 것이다. 즉, $n$개의 데이터가 독립이라고 볼 수 없다.
그렇다면 현실과 조금 동떨어진 가정을 기반으로 통계학에서는 많은 논의를 시작하는 것일까? 이는 아마도 $i.i.d$ 가정을 한다면, 논의가 쉬워지기 때문이다. 즉, 분포를 유도해서 추론을 하는데 용이하기 때문이다. 따라서 $i.i.d$ 가정을 하고, 추후에 이 가정을 하는 것이 정말로 올바른 것인지 사후적으로 살펴보는 것이다.

사후적으로 살펴보았을 때, $i.i.d$ 가정을 하는 것이 어느 정도 맞았다면, 그 가정을 유지한채 논의를 확장하면 되고, 그것이 아니면 데이터 구조에 따라서 다른 방법을 생각해야 한다. 예를 들어, longitudinal data를 생각해보자. subject 별로 반복 측정 자료에 대해서 $i.i.d$ 가정을 하는 것은 비현실적이다. 한 subject 내에서 관측된 데이터들은 분명 correlate 되어 있을 것이기 때문이다. Spatial Data도 데이터 간의 독립을 가정하는 것은 무리가 있고 spatially dependency structure가 있다고 가정한다.
여기서 spatial dependecy는, 데이터가 geographically reference 되었다는 것을 의미한다. 예를 들어, 중국의 지역별 우한 바이러스의 발생 빈도 데이터를 생각해보자. 우한 바이러스는 중국의 후베이성, 우한 도시에서 처음 발병했는데, 초기에 후베이성을 중심으로 확진자들이 급속도로 증가했다. 동시에, 아프리카에서 우한 바이러스가 발생한 빈도를 보자. 거의 0에 가까울 것이다. 즉 데이터가 후베이성을 중심으로 급격히 증가하는 형태를 갖는데, 데이터가 서로 독립이라고 보기는 사실상 불가능하고 공간에 따라서 서로 correlate 되어 있다고 보는 것이 합리적이다. 바로 이러한 dependency structure을 갖는 데이터가 Spatial data이다.

## Importance of Dependence

데이터 간에 존재하는 dependence을 무시하고 독립 가정을 한다면 어떤 문제가 발생할까? 직관적으로는, 실제 존재하는 variability을 무시하는 것이므로, variability가 작다고 간주하여 underestimate 할 것이라고 예상할 수 있다. 실제로 이런 직관이 맞다. $n$개의 데이터가 $i.i.d$인 경우를 생각해보자;

$Z(1), \cdots, Z(n) \overset{iid}{\sim} N(\mu, \sigma^2)$. $\mu$에 대한 estimator인 $\bar{Z}$의 CI는 $\left(\bar{Z} - 1.96\frac{\sigma}{n}, \bar{Z} + 1.96\frac{\sigma}{n}\right)$이다.
반면에, $Z(i)$ 간의 covariance strucure을 $Cov(Z(i), Z(j)) = \sigma^2\rho^{|i-j|},\ \rho \in (0,1)$로 설정한 경우를 생각해보자. 이때 $Var(\bar{Z})$는 아래와 같다.

$$Var(\bar{Z}) = \sum_{i,j} Cov(Z(i),\ Z(j))/n^2 = \frac{\sigma^2}{n}\left(1 + 2\left(\frac{\rho}{1-\rho}\right)\left(1 - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2 \frac{(1-\rho^{n-1})}{n}\right)$$

예를 들어 $n = 10, \rho = 0.26$이면 $Var(\bar{Z}) = \frac{\sigma^2}{n}1.608$이고 $\mu$에 대한 CI는 $\left(\bar{Z} - 2.485\frac{\sigma}{n}, \bar{Z} + 2.485\frac{\sigma}{n}\right)$이므로 독립을 가정했을 때보다 더 넓다. 즉, dependence를 무시한다면 estimator가 underestimate될 우려가 있다는 것이다.

## Types of Spatial Data

일반적으로 세 종류의 spatial data가 있다.

1. Geostatistical data

   points에서 '연속적'으로 변하는 spatial preocess이다. 여기서 points는 말 그대로 특정 지점을 의미한다. 예를 들어서, 위도, 경도가 (50,50)인 지점의 관측 첫날 온도가 25도이고 10일동안 관측한다고 하자. (위도, 경도, 온도는 임의로 정했다) 또, 위도, 경도가 (50,60)인 지점의 관측 첫날 온도가 28도이고 마찬가지로 10일동안 관측한다고 하자. 이렇게 어느 '지점'에서 연속적으로 관찰된 데이터를 Geostatistical data 또는 Point Reference data라고 한다.

2. Lattice (areal) data

   Geostatistical data와 유사한데, 관측 단위가 특정 지점이 아니라 '지역'을 의미한다. 예를 들어, 서울의 기온과 부산의 기온을 10일 동안 관측하며 생기는 데이터는 Lattice data이다. 예를 들어 서대문구 신촌동 xx로 xxxx과 강남구 서초동 xx로 xxxx 등, 서울의 특정 지점 10곳의 데이터를 10일동안 관측했다면 이는 Geostatistical data이고, 이 10곳의 데이터를 평균내서 하나의 데이터로 만들고 동일한 형태의 데이터를 전국 도에서 관찰한다면, 이는 Lattice data인 것이다.

3. Spatial point process

   Geostatistical data와 Lattice data가 어떤 위치에서의 value에 관심이 있었다면, Spatial point process는 위치 자체에 관심을 가진다. 예를 들어, 어떤 점이 모여 있는 것을 보고 이 현상이 정말 어떤 pattern을 가지고 cluster을 형성한 것인지, 또는 random noise인지가 주요 연구 분야 중 하나이다.

## Basic Notations for Spatial Models

앞으로 Spatial models에 많이 쓰일 notation을 미리 정의하겠다.

- $d$차원에 있는 spatial data가 따르는 spatial stochastic process를 $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset R^d\}$라고 한다. 여기서 차원은 spatial data가 존재하는 공간이다. 예를 들어 위도, 경도인 spatial data는 2차원이다. 보통, $d = 2, 3$이다. spatial stochastic process에서 관측되는 realization을 sample path라고 한다. 보통, $Z = (Z(s_1), \cdots, Z(s_n))^T$를 관측한다.
- $\mathbf{s}$: process $Z(\mathbf{s})$가 관측되는 location이다.
- $Z(\mathbf{s})$는 $\mathbf{s}$에서의 spatial process value이다.
- $\mathbf{s}$는 idex set $D$에 따라서 변화한다.
- 데이터는 random spatial field에서의 realization sample이라고 간주된다.

## Spatial Distance

위에서 우한 바이러스 예시를 살펴보았는데, 후베이성 인근의 확진자 수는 아프리카 국가의 확진자 수보다 훨씬 많다. 이는 후베이성과 아프리카 국가의 거리가 멀기 때문인데, 이 거리라는 개념을 spatial data에서 어떻게 정의하는지 알아보자.

- Geodesic Distance
  두 (위도, 경도) point가 주어졌을 때, 이 두 지점간의 거리를 단순 euclidean distance로 구하면 지구의 curvature를 고려하지 않은, 부정확한 측정을 하게 된다. 따라서 이를 반영한 거리가 Geodesic Distance이다. $R$은 지구의 평균 반지름이다.

$$D = R\, arccos[sin(lat1)sin(lat2) + cos(lat1)cos(lat2)(lon1 - lon2)],\ R = 6371$$

- Chordal Distance
  $\mathbf{u}_1 = (x_1, y_1, z_1),\ \mathbf{u}_2 = (x_2, y_2, z_2)$, 두 벡터의 euclidean distance을 구하는데, $x_i, y_i, z_i$는 아래와 같이 둔다.

$$x_i = Rcos(lat_i)cos(lon_i),\ y_i = Rcos(lat_i)sin(lon_i),\ z_i = Rsin(lat_i)$$