Spatio Temporal Data Analysis

Week 7: Introduction to Areal/Lattice Data

1. Introduction

여태까지의 포스팅은 spatial data 중 point referenced data (geostatistical data)를 다루었다. 즉, 공간 의존성을 가지되, 공간 정보가 점으로 주어지는 데이터에 대해서 살펴보았다. 이러한 데이터는 공간에 대한 정보가 (위도,경도)로 주어진다. 예를 들어서, 위도와 경도가 (50,50)인 지역과 (52,53)인 지역의 온도, 바람의 세기 등이 주어진다. point referenced data에서는 데이터 간의 공간 의존성을 정의할 때, distance을 어떻게 정의하는지가 중요했다. 여태까지 살펴본 바로는, 단순히 euclidean distance을 구하기도 했고, 또는 지구의 curvature을 반영한 두 지점의 거리를 계산하기도 했다. 이번 포스팅에서 살펴볼 데이터의 형태는 Areal/Lattice 데이터이다. 사실 공간 의존성이 있는 데이터에 대해서는 첫 주차에서 살펴보았는데, 다시 한번 Areal 데이터에 대해서 간략하게 살펴보자.

Areal 데이터는 공간에 대한 정보가 점, 즉 (위도, 경도)로 주어지는 것이 아니라 지역으로 주어진다. 즉, 서대문구의 미세먼지 지수, 영등포구의 미세먼지 지수, 또는 더 세부적으로 연희동, 연남동, 문래동의 미세먼지 지수 등의 형태로 주어지는 것이다. 구체적으로 문래동의 어느 위도, 경도의 미세먼지가 주어지는 것이 아니다. 이러한 데이터의 특성 때문에 데이터들 간의 공간 의존성을 정의하는 것도 point referenced 데이터와는 조금 차이를 보인다. point referenced 데이터에서는, 데이터들 간의 거리를 정의하는 것이 key라고 했는데, areal 데이터에서는 데이터들의 이웃을 어떻게 정의하느냐가 중요하다. 이는 조금만 생각해보면 자연스러운 결과인데, areal 데이터에서는 구체적인 위도, 경도가 주어지는 것이 아니라 넓은 지역에 대한 데이터가 주어지기 때문에 이러한 지역들 사이의 거리를 정의하기가 애매하다. 지역의 중점간의 거리로 정의할 수 있겠지만, 이는 많은 bias를 발생시킬 것이다. 따라서 areal 데이터에서는 지역의 이웃을 정의하는 것이다.

그러면 이웃을 어떤식으로 정의할까? 아래 예시를 살펴보자.

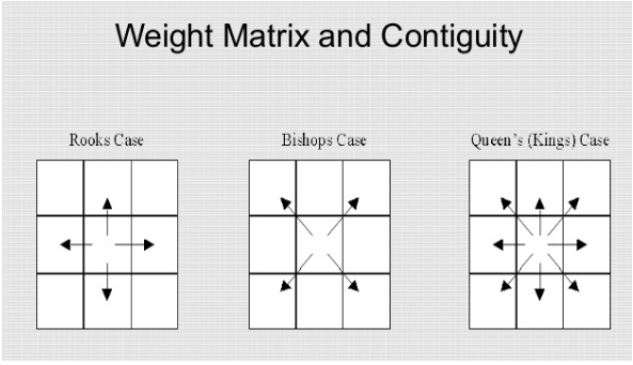

그림 1

만약 지역이 그림 1과 같이 regular하게 정의되어 있다면, 동서남북, 대각선 등으로 이웃을 정의할 수 있다. 하지만 지역이 이런 식으로 이쁘게 정의되는 데이터는 많지 않을 것이다. 아래의 그림을 보자.

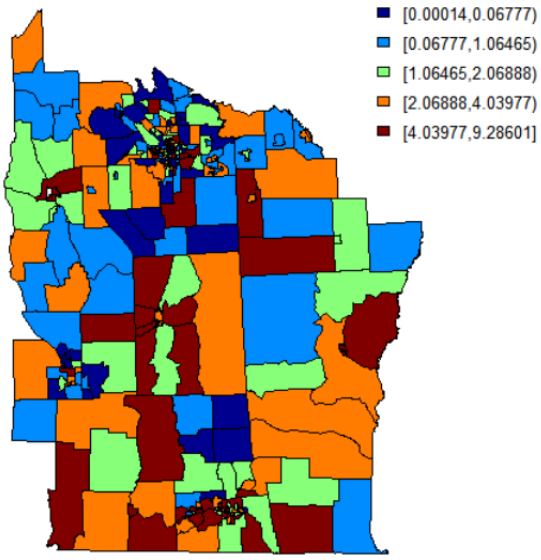

그림 2

그림 2는 뉴욕 주의 지역별 백혈병 발생 빈도를 색으로 표현한 것이다. 이와 같이 지역의 모양이 제각각일 때는 이웃을 제대로 정의해줘야 할 것이다.

2. Basic Setup

Define Proximity (or Weight) Matrix

이제 $\mathbf{W}$ $(n \times n)$을 정의해보자. 이 행렬은 i, j가 얼마나 가까운지, 그리고 그 관계가 얼마나 강한지를 나타내야 한다. 예를 들어서,

- w_{ij}: 지역 i, j간의 거리의 역수
- w_{ij}: 지역 i, j가 인접하면 1, 그렇지 않으면 0
- w_{ij}: 지역 i, j의 거리가 미리 정해진 δ보다 크면 1, 그렇지 않으면 0
- w_{ij}: 지역 j가 i의 가장 가까운 m개의 이웃 중 하나일 때 1, 그렇지 않으면 0

어떤 기준으로 갈 것인지는 AIC, BIC등을 보고 판단한다.

To see Whether there is Spatial Dependence

point referenced 데이터에서 EDA 느낌으로 variogram을 살펴보았다. 이는 point referenced 데이터에 공간 의존성이 있는지 알아보기 위해 수행하였다. areal 데이터에서도 variogram과 비슷한 역할을 하는 것이 있다. 바로 Geary's C와 Moran's I가 바로 그것이다.

- Geary's C

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{2\left(\sum_{i \neq j} w_{ij}\right)\sum_i (Y_i - \bar{Y})^2}$$

Geary's C는 무조건 양수이고 만약 areal units들이 독립이라면 1이다. 즉 C 값이 작을수록 areal units간에 공간 의존성이 크다. 이를 이용해서 검정을 하는데 귀무가설을 공간 의존성이 없다고 두고, 귀무가설 하에 근사적으로 정규분포를 따르므로 검정을 할 수 있다.

- Moran's I

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (Y_i - \bar{Y})^2}$$

Moran's I는 양수일 필요가 없고, 음의 공간 의존성도 나타낸다. 즉, 절댓값이 클 수록 공간 의존성이 강하다. 예를 들어 I가 음의 방향으로 작다면, 거리가 멀어질수록 두 지역간의 공간 의존성이 작아지는 것을 의미한다. C는 양수만 가지므로 상관관계로 해석하기가 애매하지만 I는 양수, 음수 모두 가질 수 있으므로 더 상관관계처럼 해석할 수 있다. I도 근사적으로 귀무가설 하에 독립을 따른다.

C, I는 $\mathbf{W}$를 어떻게 정의하느냐에 매우 의존한다. 그리고 관측치가 non-iid인 이유는 공간 의존성 때문이 아니라 이분산 등의 이유일 수도 있으니 주의하자.

Markov Random Fields

Areal 데이터를 본격적으로 살펴보기 이전에 markov random fields을 살펴보자. 만약 random vector $\mathbf{Y}$가 가우시안 분포를 따르고 아래의 조건부 독립을 만족한다면 Gaussian Markov random field라고 부른다.

$$Y_i \perp Y_j \mid \mathbf{Y}_{-\{i,j\}}$$

즉, Y_i, Y_j는 이 둘을 제외한 $\mathbf{Y}_{-\{i,j\}}$가 주어졌을 때 독립이다.
GMRF와 관련된 개념으로 precision matrix가 있다. 알고 있겠지만, precision matrix는 공분산 행렬 $\mathbf{\Sigma}$의 역행렬이다; $Q = \mathbf{\Sigma}^{-1}$. precision matrix을 먼저 정의하면 역행렬을 계산할 필요가 없다. GMRF를 이용하면, 아래의 fact가 성립한다고 한다.

$$Y_i \perp Y_j \mid \mathbf{Y}_{-\{i,j\}} \iff Q_{ij}$$

따라서 GMRF을 따르는 random vector $\mathbf{Y}$는 아래의 다변량 정규분포를 따른다.

$$(2\pi)^{-n/2} |\mathbf{Q}|^{1/2} exp\left\{-\frac{1}{2}(\mathbf{Y} - \mu)' \mathbf{Q} (\mathbf{Y} - \mu)\right\}$$

만약 $\mathbf{Q}$가 singular 하다면, 즉 그 역행렬이 존재하지 않는다면 우리는 여전히 GMRF을 얻지만, 결합분포가 improper하다. 따라서 데이터를 모델링하는데에는 부적절하지만 베이지안 계층 모형에서 사후 분포가 propoer하다면 prior로 사용해도 괜찮다.
GMRF는 세 개의 조건부 독립 성질을 가지고 있는데, 이는 모두 동치이다.

- pairwise markov property: $Y_i \perp Y_j \mid \mathbf{Y}_{-\{i,j\}}$
- local markov property: $Y_i \perp \mathbf{Y}_{-\mathcal{N}(i)} \mid \mathbf{Y}_{\mathcal{N}(i)}$, where $\mathcal{N}(i)$ is the set of neighbors of i

- global markov property: $\mathbf{Y}_A \mid \mathbf{Y}_B \mid \mathbf{Y}_C$ for all nonempty disjoint sets A, B, C where C separates A, B.

Brook's Lemma

이 Lemma는 areal 데이터의 모델에 쓰이는 정리이므로 짚고 넘어가자.

- Let $p(\mathbf{y})$be the joint density of $\mathbf{Y} \in \mathbb{R}^n$. Let $(y_1^*, \cdots, y_n^*)$ be an any fied point in the support of $p(\cdot)$. Thenm

$$p(\mathbf{y}) \propto \frac{p(\mathbf{y})}{p(\mathbf{y}^*)} = \frac{p(y_1 \mid y_2, \cdots, y_n)p(y_2 \mid y_1^*, y_3, \cdots, y_n)}{p(y_1^* \mid y_2, \cdots, y_n)(y_2^* \mid y_1^*, y_3, \cdots, y_n)} \times \frac{p(y_n \mid y_1^*, \cdots, y_{n-1}^*)}{p(y_n^* \mid y_1^*, \cdots, y_{n-1}^*)}$$

  joint density가 조건부 분포들의 값과 상수 $p(\boldsymbol{y}^*)$로 결정된다.

- Brook's Lemma는 조건부 분포를 알고 있다면 결합 분포를 유도할 수 있음을 보여준다.

Brooks' Lemma는 결합 분포를 알고 싶은데, 조건부 분포에만 접근할 수 있는 상황에 사용된다. 예를 들어서,

$$X \mid Y \sim N(\rho Y, 1-\rho^2),\ Y \mid X \sim N(\rho X, 1-\rho^2)$$

와 같이 조건부 분포만 주어졌을 때, 어느 값의 x_0, y_0에 대해서

$$\frac{p(x,y)}{p(x_0,y_0)} = \frac{p(x \mid y)p(y \mid x_0)}{p(x_0 \mid y)p(y_0 \mid x_0)}$$

가 성립하니까 $x_0 = y_0 = 0$으로 두고 결합 분포를 구할 수 있다.

3. Models for Areal Data

Intrinsic Autorevressive Models (IAR)

areal 데이터에 대한 접근 중 하나로 IAR이 있다. 결합 분포가 아니라 조건부 분포를 알고 있다고 하자.

$$Y_i \mid Y_j, j \neq i \sim N\left(\sum_j b_i y_j, \tau_i^2\right),\ i = 1, \cdots, n$$

Brooks' Lemma를 이용해서 결합 분포를 얻기 전에, 아래의 notation을 정의하자.

- $w_{i+} = \sum_j w_{ij}$, the row sums of $\mathbf{W}$
- $b_{ij} = w_{ij}/w_{i+}$
- $\tau_i^2 = \tau^2/w_{i+}$

이를 이용하면

$$Y_i \mid Y_j, j \neq i \sim N\left(\sum_j w_{ij} y_j / w_{i+}, \tau^2 / w_{i+}\right),\ i = 1, \cdots, n$$

여기에 Brook's Lemma를 이용하면

$$f(\mathbf{y}) \propto exp\left\{-\frac{1}{2\tau^2}\mathbf{y}^{'}\left(\mathbf{D}_w - \mathbf{W}\right)\mathbf{y}\right\}, where\ \left[\mathbf{D}_w\right]_{ii} = w_{i+}$$

여기서 $\left(\mathbf{D}_w - \mathbf{W}\right)\mathbf{1} = 0$이므로 $\mathbf{Q} = \mathbf{D}_w - \mathbf{W}$은 singular하다. 따라서 IAR은 데이터를 모델링하기에는 적절하지 않은데, 어떻게 $\mathbf{Q}$를 바꿔야 proper 분포가 될까?

Conditional Autoregressive Models (CAR)

$\mathbf{Q} = \mathbf{D}_w - \rho\mathbf{W}$로 둔다면, positive definite가 되고, proper 분포를 만들 수 있다. 여기서 ρ는 추정해야하는 모수이다. 만약에 ρ가 0이라면 독립이고, 1이라면 IAR에서의 improper 분포가 된다.
따라서 $\rho \in [0, 1]$이 공간 의존성을 결정하는, CAR 모델을 정의할 수 있다.

$$Y \sim N\left(0, \left(\mathbf{D}_w - \rho\mathbf{W}\right)^{-1}\right)$$

이러한 CAR 모델은 아래의 두 문제점이 있다. CAR 모델의 조건부 분포는

$$Y_i \mid Y_j, j \neq i \sim N\left(\rho\sum_j w_{ij} y_j / w_{i+}, \tau^2 / w_{i+}\right),\ i = 1, \cdots, n$$

인데, Y_i의 조건부 평균을 해석하는데 어려움이 있다.
더 심각한 문제는 ρ의 추정치이다. ρ는 공간 의존성을 결정하는 모수라고 했는데, ρ가 거의 1에 가까워야지 Moran's I의 값이 1에 가까워지기 때문이다. 그런데, ρ가 1에 가까워질수록 IAR의 특성을 닮아가고, $\mathbf{Q}$가 singular에 가까워져서 불안정해질 것이다. 공간 의존성을 나타내기 위해 Moran's I의 값이 어느 정도 1에 근접해야하는데, 그럴수록 ρ는 1에 가까워지는 것이다. 대안으로 SAR 모델이 제안되었다.

Simultaneous Autoregressive Model (SAR)

마치 시계열의 autoregressive model과 유사한 형태를 지닌다. 각 Y_i의 평균이 어떤 고정된 $\mathbf{B}$에 대해서, Y의 선형 결합이라고 가정한다.

$$\mathbf{Y} = \mathbf{B}\mathbf{Y} + \epsilon,\ \therefore \mathbf{Y} = (\mathbf{I} - \mathbf{B})^{-1}\epsilon$$

$$\epsilon \sim N(0, \tilde{\mathbf{D}}),\ where\ \left(\tilde{\mathbf{D}}\right)_{ii} = \sigma_i^2$$

$$\therefore \mathbf{Y} \sim N(0, (\mathbf{I} - \mathbf{B})^{-1}\tilde{\mathbf{D}}\left((\mathbf{I} - \mathbf{B})^{'}\right)^{-1}$$

대부분의 사람들은 $\mathbf{B} = \rho\mathbf{W}$와 $\tilde{\mathbf{D}} = \sigma^2\mathbf{I}$ (등분산)을 사용한다. 이러한 SAR 모델은 full condtions에 대한 분포가 없어서, 깁스 샘플링을 사용하기에 무리가 있다. SAR의 단점을 극복하기 위해서,

Matern GP를 근사하는 방법이 제안되었다.

GMRF via SPDE

Lindbergh[1]는 Stochastic Partial Differential Equation의 해가 Gasussian Process with a Matern Covariance임을 밝혔다. SPDE의 형태는 아래와 같다.

$$(\kappa^2 - \Delta)^{\alpha/2} Y(s) = \sigma W(s),\ s \in \mathbb{R}^2$$

이러한 SPDE의 stationary solution이 바로 GP인데, solution을 finite differences로, 여기서 differences는 이웃 구조와 일치하게 조정하여 살짝 다르게 만들어보자. $\alpha = 2$로 두면,

$$(\kappa^2 \tilde{\mathbf{C}} + \mathbf{G})\mathbf{Y} = N(0, \tilde{\mathbf{C}})$$

를 얻고, $\tilde{\mathbf{C}}, \mathbf{G}$는 neighborhood structure에 의존하고 sparse한 행렬이다. 이를 적용하여 $\mathbf{Y}$의 분포를 유도하면,

$$\mathbf{Y} \sim N\left(0, \tilde{\mathbf{Q}}^{-1}\right),\ where\ \tilde{\mathbf{Q}} = (\kappa^2 \tilde{\mathbf{C}} + \mathbf{G})' \tilde{\mathbf{C}}^{-1} (\kappa^2 \tilde{\mathbf{C}} + \mathbf{G})$$

그런데 $\tilde{\mathbf{Q}}$가 sparse하지 않으므로, $\tilde{\mathbf{C}}$의 row sums으로 구성된 diagonal matrix $\mathbf{C}$로 바꾸어 $\mathbf{Q} = (\kappa^2 \tilde{\mathbf{C}} + \mathbf{G})' \mathbf{C}^{-1} (\kappa^2 \tilde{\mathbf{C}} + \mathbf{G})$을 유도한다. 이제 $\mathbf{Q}$는 sparse하고 계산하기 쉽다.
결론은 Matern GP를 근사하는 모델이 GMRF via SPDE이고, 이는 공분산의 모수를 CAR, SAR 보다 더 잘 조절할 수 있다. 또한 CAR 모델과 마찬가지로 precision matrix을 직접적으로 모델링하기 때문에, 역행렬을 취할 필요도 없다.

[1] An explicit link between Gaussian fields andGaussian Markov random fields: the stochasticpartial differential equation approach, Lindgren et al.