이학석사학위 논문

비모수 베이지안을 이용한
한국 초미세먼지(PM2.5) 분석

Analysis of PM2.5 in Korea using
Non-parametric Bayesian statistical model

2024년 2월

인하대학교 대학원

통계학과(통계학전공)

정 태 민

이학석사학위 논문

비모수 베이지안을 이용한
한국 초미세먼지(PM2.5) 분석

Analysis of PM2.5 in Korea using
Non-parametric Bayesian statistical model

2024년 2월

지도교수 조 성 일

이 논문을 석사학위 논문으로 제출함

이 논문을 정태민의 석사학위논문으로 인정함.

2023년 12월

주심 유동현 (인)

부심 조성일 (인)

위원 온일상 (인)

요약

‘2019 세계 대기질 보고서’에 따르면 한국이 OECD 회원국 중 초미세먼지(PM2.5) 오염이 가장 심각하다. 또 초미세먼지가 건강에 치명적인 영향을 미치는 것은 많은 연구 통해 알려져 있다. 그래서 대한민국 정부는 초미세먼지를 줄이는 대책을 내놓았으나 초미세먼지 농도가 공간적으로 이질적인데도 불구하고 일괄적인 정책을 펴고 있다. 본 논문에서는 초미세먼지 농도가 공간적으로 이질적임을 군집화를 통해 반영하고 초미세 먼지 농도를 추론하여 초미세먼지 대응에 도움을 주고자 한다. 초미세먼지 자료로는 월별 초미세먼지 농도 평균인데, 시군구에 따른 162개의 지역과 2019년 5월부터 2022년 11월까지 총 43개월의 자료가 있다. 지역별 초미세먼지 농도에 대해 디리슈레 과정 가정하였고, 공간적 특성을 반영하기 위해 로지스틱 막대 분할 과정을 사용하였다. 시간에 따른 추론을 하고자 각 모수들에 가우스 과정을 가정했다. 100번의 시뮬레이션을 통해 검증하였고, 한국 초미세먼지 자료에 대해 공간적으로 군집화하였고, 각 군집에 따른 모수들을 추정하였다.

주요용어: 비모수 베이지안, 디리슈레 과정, 가우스 과정, 시공간 자료 분석, 로지스틱 막대분할 과정, 초미세먼지

abstract

According to the '2019 World Air Quality Report', Korea has the most serious pollution of PM2.5 (particles less than 2.5 micrometers in diameter) among OECD. In addition, it is known through many studies that PM2.5 has a fatal effect on health. So, the government of the Republic of Korea has come up with measures to reduce PM2.5, but despite the spatially heterogeneous concentration of PM2.5, a comprehensive policy is being implemented. In this paper, we reflect the spatial heterogeneity of PM2.5 concentration through clustering and infer the PM2.5 concentration to help respond to PM2.5. The data is the average concentration of PM2.5 per month, and there are 162 regions according to cities, counties, and districts, and data for a total of 43 months from May 2019 to November 2022. The Dirichlet process was assumed for the concentration of PM2.5 by region, and the logistic stick-breaking process was used to reflect spatial characteristics. Gaussian process was assumed for each parameter to make inference over time. It was verified through 100 simulations, spatially clustered for PM2.5 data in Korea, and estimated parameters for each cluster.

Keywords: Nonparametric Bayes, Dirichlet process, Gaussian process, spatio-temporal data analysis, logistic stick-breaking process, PM2.5

Contents

1 서론

대기오염물질 중 미립자 물질(Particulate Matter; PM)은 그 크기에 따라 미세먼지(Particulate Matter less than $10\mu m$ in diameter; PM10)와 초미세먼지(Particulate Matter less than $2.5\mu m$ in diameter; PM2.5)로 구분할 수 있다. 그중 초미세먼지는 호흡기 질환(Qiu 등, 2012) 심혈관 질환 발생위험을 높이는 것으로 보고되었다 (Ito 등, 2011). 2006년부터 2010년까지 서울시민을 대상으로 했을 때 65세 이상 연령집단에서 미세먼지(PM10) 농도와 초미세먼지(PM2.5)의 농도가 $10\mu g/m^3$ 증가 시 심혈관계 초과 사망 발생위험을 각각 0.80% (95% CI : 0.39∼1.21%), 1.75% (95% CI : 0.91∼2.59%) 높이는 결과를 보이는 등 (Bae, 2014) 초미세먼지는 건강에 심각한 영향을 끼친다. 에어비주얼이 출간한 '2019 세계 대기질 보고서'에 따르면 한국이 OECD 회원국중 초미세먼지 오염이 가장 심각하다. 이와 같이 대한민국의 초미세먼지 심각성과 건강에 치명적인 영향을 미치는 것으로 인해 대한민국 정부는 2027년까지 초미세먼지를 30% 감축하여 전국 연평균 농도를 2021년 $18\mu g/m^3$에서 $13\mu g/m^3$까지 낮추는 것을 국정과제로 삼아 고농도 미세먼지 대응을 강화하기로 한 바 있다. 하지만 미세먼지의 분포는 공간적으로 상이하며 그 발생도 지역별로 다르게 기인하는데도 불구하고 미세먼지 저감을

위한 정책은 차별성 없이 이루어지고 있기 때문에 미세먼지의 공간적 이질성을 반영한 연구가 필요하다 (Jeon 등, 2018). 따라서 본 논문에서는 이러한 이유 즉, 대한민국의 초미세먼지 심각성과 초미세먼지가 건강에 치명적이라는 것으로 인해 한국 초미세먼지에 대해 통계적 분석을 하려고 한다. 특히 공간적 이질성을 반영하고 초미세먼지 농도에 대해 추론하기 위한 통계분석을 하려고 한다.

초미세먼지 분석에 관한 선행연구로 베이지안 방법으로는 Beloconi 등 (2018)은 가우스 과정의 공분산 행렬에 공간 상관관계를 넣어서 모델링하는 지리 통계모형을 적용했다. 또 Saez와 Barcelo (2021)은 통합 중첩 라플라스 근사(integrated nested laplace approximation; INLA)을 이용한 가우스 마르코프 랜덤 장(gaussian markov random field; GMRF) 모델을 적용했다. Sahu와 Mardia (2005)은 보간된 칼만 필터 (kriged Kalman filter) 모델을 이용한 초미세먼지 예측 모형을 제안했다. 빈도론 방법으로는 Doreswamy 등 (2020)은 랜덤포레스트(Random forest), 경사도 부스팅 회귀(Gradient boosting regression; GBR)와 의사결정나무(decision tree)를 이용한 초미세먼지 예측 모델들을 비교하였고, Zhang와 Yang (2022)은 분할 정복 패러다임(divide-and-conquer paradigm)을 이용해 군집분석을 하였고, Rendana 등 (2022)은 계층적 군집화 분석(Hierarchical clustering analysis; HCA)을 이용해 군집분석을 하였다.

본 논문의 분석 목적은 한국 초미세먼지 자료에 통계모형을 적합시키고 군집분석을 하여 초미세먼지 농도에 따라 우리나라 지역을 분할하고 초미세먼지 농도를 추론하는 것이다. 기존의 연구는 초미세먼지에 대해 통계모형을 적합하거나 군집화 둘 중 하나만 이용해 분석했다면 본 논문의 연구는 모형 적합을 통한 추론과 군집화를 동시에 할 수 있다. 또 이에 따 RMSE(root mean squared error), MAE(mean absolute error) 등 모형 성능 지표를 계산할 수 있어 다른 모형과의 수치적 비교가 가능하는 장점도 있다.

초미세먼지 자료 분석에 사용한 통계모형은 비모수 베이지안 방법인 디리슈레 과정(Dirichlet process; DP)이다. 디리슈레 과정은 분포가정에 대한 의존도를 줄이고 다양한 통계 모형에 대해 로버스트한 결과를 제공한다. (Ferguson, 1973). 또 공간적으로 종속된 자료를 군집화하기 위해 근접한 자료가 함께 군집이 될 가능성이 높은 로지스틱 막대 분할 과정(Logistic stick-breaking process; LSBP)을 사용하였다 (Len 등, 2011). 시간에 따른 추론을 하고자 각 모수에 가우스 과정(Gaussian process; GP)을 가정하였다. 시공간 자료 분석에서 모수들에 가우스 과정을 가정하는 아이디어는 요인분석의 맥락에서 제시되었다 (Luttinen와 Ilin, 2009). 다음과 같은 모형은 Ding에 의해 시공간 자료 분석에 효과적임이 입증되었다 (Ding 등, 2012).

2장은 초미세먼지 자료에 시공간적 특성을 보기 위한 자료 분석을 한 결과이다. 3장에서는 모형에 사용된 이론을, 4장에서는 자료에 적용한 비모수 베이지안 통계 모형에 관해 설명했다. 5장은 시뮬레이션 설계와 그에 따른 분석 결과이고, 6장은 초미세먼지 자료에 통계 모형을 적합한 결과이다. 7장은 본 연구의 결론이다.

2 자료 설명

한국 초미세먼지 자료는 KOSIS(https://kosis.kr)에 있는 2019년 5월부터 2022년 11월까지 총 43개월간 162개 지역에 따른 월평균 초미세먼지 농도를 사용하였다. 이는 전국에 있는 510여 개의 대기오염 측정망을 통해 관측한 것으로 162개 지역의 월별 초미세먼지 평균을 나타낸다. 결측치들은 도평균으로 대체하였다. y_{it}를 초미세먼지 농도라고 했을 때, i는 지역 번호로 $i = 1, 2, \ldots, 162$이고, t는 시간 번호로 $t = 1, 2, \ldots, 43$이다.

2.1 전체적 공간 자기상관 분석

공간 상관관계란 공간 정보가 관측치에 영향을 미치는지 여부로 공간적으로 인접한 관측치들이 유사한 값을 갖거나 상반되는 값을 갖는 것을 말한다. 공간 상관관계가 있는지 파악하기 위한 방법으로 가장 많이 쓰이는 것은 전체적 모란의 I (Global Moran's I) 통계량이다. 이는 '공간적으로 무작위 패턴을 보인다(공간적 상관관계가 없다)'라는 귀무가설

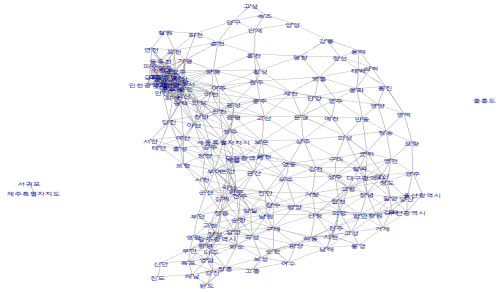
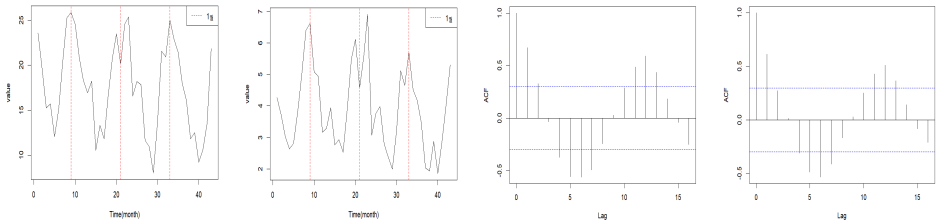


Figure 1: 공간가중치

하에서 z-검정을 통해서 다음과 같이 계산된다.

$$I_t = \frac{N}{W}\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_{it}-y_{\cdot t})(y_{jt}-y_{\cdot t})}{\sum_{i=1}^{N}(y_{it}-y_{\cdot t})^2},\ W = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \qquad (1)$$

$y_{\cdot t}$는 t시점에서 모든 지역의 초미세먼지 평균이고, $N = 162$이다. w_{ij}는 공간가중치로 근접한 지역에 대한 네트워크 형태다. 초미세먼지 자료 분석에서는 50km 이하의 지역에 대해 공간가중치를 1, 아니면 0이라고 설정했다. 이렇게 설정한 공간가중치를 그림으로 표현하면 Figure 1과 같다. z-검정을 통해 테스트했을 때 모든 시간에 대해 전체적 모란의 I의 p-value는 모두 1e-6 미만이다. 따라서 모든 시간에 대해 공간적 상관관계가 있다고 할 수 있다.



(a) 시간에 따른 평균 변화량 (b) 시간에 따른 표준편차 변화량 (c) 시간의 따른 평균의 자기상관함수 (d) 시간의 따른 표준편차의 자기상관함수

Figure 2: 시간에 따른 공간의 초미세먼지 평균과 표준편차와 자기상관계수

2.2 자기상관함수

시계열 특성을 파악하기 위해 가장 많이 사용되는 지표로 자기상관함수가 있다. 이는 시간 차이(time lag)에 따라 값들이 어떤 상관이 있는지 보기 위한 함수로 다음과 같이 정의된다.

$$r_k = \frac{\sum_{t=k+1}^{T}(y_{\cdot t} - \bar{y})(y_{\cdot(t-k)} - \bar{y})}{\sum_{t=1}^{T}(y_{\cdot t} - \bar{y})^2} \tag{2}$$

$y_{\cdot t}$는 t시점에서 모든 지역의 초미세먼지 평균이고, $T = 43$이다. r_k는 자기상관계수이고 k는 시간 차이를 뜻한다. Figure 2은 시간에 따른 평균과 표준편차의 변화량과 자기상관함수를 그린 그림이다. Figure 2의 (c)와 (d)에서 수평선은 95% 신뢰구간이다. $k = 6$과 $k = 12$부근에서 95% 신뢰구간을 초과하는 것을 볼 수 있다. 이는 시간에 따른 주기성이

있다고 할 수 있고, 12개월의 주기를 갖는다.

3 비모수 베이지안

비모수 베이지안 방법은 모수적 방법보다 유동적인 모형을 만드는 것이 가능한데 그 이유는 모수적 방법은 변수를 하나의 분포만으로 가정하기 때문에 자료의 특성을 지나치게 단순화시킬 수 있지만 비모수 베이지안 방법은 임의의 유동적 분포를 설정하기 때문이다 (Noh 등, 2014). 이 장에서는 본 논문에서 사용된 통계모형인 디리슈레 과정(Dirichlet process; DP), 로지스틱 막대분할 과정(logistic stick-breaking process; LSBP), 가우스 과정(Gaussian process; GP)에 대해 설명하고자 한다.

3.1 디리슈레 과정

디리슈레 과정은 비모수 베이지안 방법 중 하나로 임의분포에 대한 사전분포를 가정한다. 임의분포가 추정되는 과정에서 디리슈레 분포가 사용되고, 무한대까지 확장할 수 있는 모형이기 때문에 디리슈레 과정이라고 한다. 디리슈레 분포의 특성을 이용해 군집화가 가능하고 비모수 방법이기 때문에 유연한 모델링이 가능하다. 디리슈레 과정을 많이 사용하는 이유로는 켤레 사전분포로 계산하기 용이하다는 것과 일반적인 모형으로 확장이 쉽기 때문이다 (Dunson, 2010). 디리슈레 과정을 이용해서 1부터 N까지 θ_n에 대해 임의분포를 가정할 때 다음과 같이 표현할 수

있다.

$$\theta_n \mid G \overset{iid}{\sim} G, n = 1, \ldots, N,$$
$$G \sim DP(\alpha, G_0) \tag{3}$$

α와 G_0는 초모수(hyperparameters)이다. α는 질량모수(mass parameter)로 군집의 개수에 영향을 미친다. α가 커질수록 군집의 개수가 많아지는 경향이 있고 표본 분포가 기반분포 G_0에 집중되는 성격이 약해지므로 표본 분포의 다양성이 증가한다. G_0는 기저분포(base measure)로 G의 기대분포이면서, θ에 가정되는 분포 집합이다. G_0의 영역을 유한개의 집합 $(T_1, \ldots, T_k)$으로 나누면 식 (3)을 식 (4)처럼 표현할 수 있는데 여기서 $K \to \infty$ 이면 G가 디리슈레 과정을 따른다고 할 수 있다.

$$(G(T_1), \ldots, G(T_K)) \sim Dir(\alpha G_0(T_1), \ldots, \alpha G_0(T_K)). \tag{4}$$

$(T_1, \ldots, T_k)$은 G_0를 유한개의 영역으로 나눴을 때의 집합이다. 디리슈레 과정의 장점으로 계산이 단순하다고 했다. 다음과 같이 $(\theta_1, \ldots, \theta_N)$가 주어졌을 때 켤레 사전분포의 특성을 이용하여 G의 사후분포 $G \mid$

$(\theta_1, \ldots, \theta_N)$를 다음과 같이 구할 수 있다.

$$(G(T_1), \ldots, G(T_K)) \mid (\theta_1, \ldots, \theta_N) \sim Dir(G(T_1) + \sum_{i=1}^{N} I(\theta_i \in T_1), \ldots, G(T_K) + \sum_{i=1}^{N} I(\theta_i \in T_K)). \tag{5}$$

I는 지시함수로 괄호 안의 조건을 만족하면 1, 아니면 0을 반환한다.

3.2 디리슈레 혼합모형

디리슈레 혼합모형(Dirichlet process mixture model; DPM)이란 디리슈레 과정을 이용하여 변수를 군집화하고 추정하기 위한 모형이다. 무한한 수의 군집을 허용하여 사전에 군집의 개수를 알 수 없는 경우에 특히 유용하다. y가 반응변수로 주어지고 θ를 모수로 갖는 임의의 분포 F를 따른다고 할 때, 디리슈레 과정을 이용한 혼합모형은 다음과 같이 표현할 수 있다.

$$p(y) = \int f(y|\theta) dG(\theta), \tag{6}$$

$$G \sim DP(\alpha, G_0) \tag{7}$$

f는 분포 F에 대한 확률질량함수이다. 디리슈레 과정은 임의분포에 대한 사전분포이기 때문에 자료의 확률밀도함수 추정이 가능하다. 이를 풀어서 쓰면 다음과 같이 표현할 수 있다.

$$y_n \sim F(\theta_n), n = 1, \ldots, N \tag{8}$$

$$\theta_n \sim G \tag{9}$$

$$G \sim DP(\alpha, G_0) \tag{10}$$

디리슈레 과정을 따르는 G는 이산형이라서 연속형 자료의 분포를 모형화하는데 적합하지 않지만 위의 식과 같이 디리슈레 혼합모형을 통해 해결이 가능하다. 오히려 디리슈레 과정의 이산성을 활용해 자료의 분할이 가능하고 이를 통해 군집분석을 할 수 있다.

$$P(c_n = c \mid \{c_j\}_{j \neq n}, y_n, \theta) \propto \begin{cases} \frac{\#(c)}{N-1+\alpha} F(y_n \mid \theta_n), & existing\, cluster \\ \frac{\alpha}{N-1+\alpha} \int F(y_n \mid \theta) dG_0(\theta), & new\, cluster \end{cases} \tag{11}$$

식 (8∼10) 통해 식 (11)과 같이 유도가 가능하다. c는 디리슈레 과정으로 인해 할당되는 군집번호다.

$$P(\theta_c \mid y_c) \propto \prod_{i:c_i=c} F(y_i, \theta_c) G_0(\theta_c) \tag{12}$$

y_c는 c번 군집에 해당하는 관측치이고, θ_c는 c번 군집에 해당하는 모수이다. 식 (11)와 식 (12)에 의해 디리슈레 혼합모형이 업데이트 된다.

3.3 막대분할 과정

디리슈레 과정을 표현하는 방법 중 막대분할 표현방식(stick-breaking representation)이 있다. 막대분할 표현방식에 의하면 디리슈레 과정에서 추출된 임의의 확률분포 G는 다음과 같이 표현할 수 있다.

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\tilde{\theta}_k}(\theta) \tag{13}$$

$\delta_{\theta_k}(\theta)$는 퇴화분포(point mass)로 $\theta = \tilde{\theta}_k$에서의 질량이다. k는 군집번호고 π_k는 $\sum_k \pi_k = 1$을 만족하는 확률벡터로 k번째 군집에 들어갈 확률이다. 여기서 디리슈레 과정을 따르는 G로 부터 추출된 θ_i는 동일한 값을 가지는 θ_i끼리 k를 통해 묶이게 된다. 이때 그 묶음을 군집이라고

정의할 수 있다.

3.4 로지스틱 막대분할 과정

로지스틱 막대분할 과정이란 막대분할 과정의 변형으로 막대분할에 사용되는 비율을 로지스틱 함수를 사용하여 정의하는 디리슈레 과정의 표본추출을 위한 스킴(scheme)이다. 로지스틱 막대분할 과정은 공간적인 특성이 있는 자료에 대해 근접할수록 같은 군집에 들어갈 가능성 높게 군집화한다.

$$G(\theta) = \sum_{k=1}^{\infty} w_k(s_i)\delta_{\tilde{\theta}_k}(\theta) \tag{14}$$

$$w_k(s_i) = p_k(s_i)\prod_{h=1}^{k-1}[1 - p_h(s_i)] \tag{15}$$

$$p_k(s_i) = \sigma(g_k(s_i)),\ for\, k = 1, \ldots, K-1,\ p_K(s_i) = 1 \tag{16}$$

$$g_k(s_i) = \sum_{j=1}^{J} \beta_{kj}\mathcal{K}(s_i, \tilde{s}_j;\ \psi_k) + \beta_{k0} \tag{17}$$

$$\mathcal{K}(s_i, \tilde{s}_j;\ \psi_k) = \exp\left(-\|s_i - \tilde{s}_j\|^2/\psi_k\right) \tag{18}$$

로지스틱 막대분할 과정을 위의 식과 같이 표현할 수 있다. s_i는 i번째 관측치의 공간 좌표이고, $w_k(s_i)$는 i번째 관측치가 k번 군집에 들어갈

확률로 $\sum_{k=1}^{\infty} w_k(s_i) = 1$이다. $p_k(s_i)$는 막대분할에서의 막대를 자르는 비율로 0에서 1 사이의 값을 갖는다. σ는 시그모이드 함수로 $\sigma(x) = \exp(x)/(1+\exp(x))$이다. $\mathcal{K}$는 핵함수로 2차원인 공간 정보를 1차원으로 줄여 계산이 가능하게 한다. $\tilde{s}_j$는 핵중심점(kernel center)으로 초모수(hyperparameter)이며 사전에 정의된다. J는 핵중심점의 개수이다. ψ_k는 척도 모수(scale parameter)로 공간 정보에 대한 척도를 조절해 준다.

3.5 가우스 과정

가우스 과정은 확률 과정(stochastic process)으로 함수에 대한 모델링 및 예측에 사용되는 확률 모델이다. 따라서 함수에 대해 분포를 가정을 하는데 이것은 다변량 정규분포의 무한차원으로의 확장으로도 볼 수 있다. 관측된 변수가 각각 가우스 분포를 따른다고 하면 이를 하나의 다변량 가우스 분포로 볼 수 있다. 그러면 관측되지 않은 새로운 포인트에서 확률적 방식으로 함수 값에 대한 예측을 할 수 있다. 두 점 (x, x')에 대하여 정의되는 $f(\cdot)$의 공분산 구조를 핵함수라고 한다. 이를 $k(x, x')$

라고 하면 다음과 같이 표현이 가능하다 (Lim, 2022).

$$f(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot)) \tag{19}$$

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_m) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_m) \\ \vdots & \ddots & \vdots \\ k(x_m,x_1) & \cdots & k(x_m,x_m) \end{bmatrix} \right) \tag{20}$$

m은 f함수의 평균 함수이다. 가우스 과정은 핵함수를 통해 변수 간의 관계를 설명할 수 있고 조건부 가우스 분포가 가우스 분포를 따르는 특성을 이용하여 변수의 사후분포를 추정할 수 있다.

$$\begin{bmatrix} f \\ f' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m \\ m' \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma' \\ \Sigma'^T & \Sigma'' \end{bmatrix} \right), \tag{21}$$

$$f'|f \sim \mathcal{N}(m' + \Sigma'^T \Sigma^{-1}(f-m), \Sigma'' - \Sigma'^T \Sigma^{-1} \Sigma') \tag{22}$$

f는 관측된 값이고 f'는 예측하는 값이다. m는 f의 평균이고 m'은 f'의 평균이다. Σ는 핵함수로 만들어진 f에 대한 공분산이고, 이때 핵함수의 파라미터들이 학습된다. Σ'와 Σ''는 이 학습된 파라미터들로 만들어진 공분산이다.

4 모형

자료 분석에 사용한 모형은 크게 공간 효과 모형과 시간 효과 모형으로 나눌 수 있다. 공간 효과 모형은 공간에 따라 군집한 디리슈레 과정 부분이고 시간 효과 모형은 시간에 따라 모수들이 어떻게 변하는지에 관한 가우스 과정 부분이다.

4.1 공간 효과 모형

공간모형으로는 로지스틱 막대분할 과정을 사용하였다. K는 최대 군집의 개수로 $K = 20$으로 설정했다.

$$y_{it} \sim N(\mu_{it}, \sigma_t^2),\ \mu_{it} \sim \sum_{k=1}^{K} w_{kt}(s_i)\delta_{\mu_{kt}^*},\ i = 1, \ldots, N,\ t = 1, \ldots, T \quad (23)$$

$$w_{kt}(s_i) = p_{kt}(s_i) \prod_{h=1}^{k-1} [1 - p_{ht}(s_i)] \quad (24)$$

$$p_{kt}(s_i) = \sigma(g_{kt}(s_i)),\ for\, k = 1, \ldots, K-1,\ p_{Kt}(s_i) = 1 \quad (25)$$

$$g_{kt}(s_i) = \sum_{j=1}^{J} \beta_{kjt} \mathcal{K}(s_i, \tilde{s}_j;\ \psi_k) + \beta_{k0t} \quad (26)$$

다음과 같은 모형을 사용한 이유는 공간적으로 군집화를 하기 위함이다. 식 (23)를 통해 μ_{it}가 군집으로 묶이게 된다. $g_{kt}(s_i)$를 핵함수를 선형모형으로 만든 값을 넣었고, 이는 근접한 공간에 대해 비슷한 $w_{kt}(s_i)$를 갖게하고, 같은 군집에 들어갈 확률을 높인다. $w_{kt}(s_i)$는 k군집에 들어갈 확률로 $\sum_{k=1}^{K} w_{kt}(s_i) = 1$이다.

4.2 시간 효과 모형

시간 효과 모형으로는 가우스 과정을 적용하였다.

$$\beta_{kj:} \sim \mathcal{N}(\mu_{\beta_{kj}}, \Sigma_{\beta_{kj}}),\ \beta_{k0:} \sim \mathcal{N}(\mu_{\beta_{k0}}, \Sigma_{\beta_{k0}}), \tag{27}$$

$$\mu^*_{k:} \sim \mathcal{N}(\mu_{\mu_k}, \Sigma_{\mu_k}),\ \log(\sigma^2{}_{:}) \sim \mathcal{N}(\mu_{\sigma^2}, \Sigma_{\sigma^2}), \tag{28}$$

$$\mu_{\beta_{kj}}, \mu_{\beta_{k0}}, \mu_{\mu_k}, \mu_{\sigma^2} \sim N(0, 1000) \tag{29}$$

모든 가우스 과정의 핵함수는 $k(x, x') = c_0 {c_1}^{(x-x')^2} + c_2 {c_3}^{\sin^2(\pi|x-x'|/12)}$ 를 사용하였다. c_0와 c_2에는 $IG(10^{-3}, 10^{-3})$과 같이 역감마분포를 사전분포로 c_1,c_3에는 $N_{(0,1)}(0.5, 1)$과 같이 잘린 정규분포를 사전분포로 하였다. 이러한 핵함수를 사용한 이유는 12개월을 주기로 주기성을 가지고 있기 때문이다. 시간적 연관성을 반영하기 위해 모수들에 시간 효과를 주었는데 이때 가우스 과정을 사용한 이유는 시간적으로 복잡한 연관성을

가지고 있어도 핵함수를 조정함으로 그 연관성을 반영할 수 있기 때문이다. 이로 인해 시간적인 추세와 주기성 모두 모형에 반영할 수 있다.

5 시뮬레이션

이 시뮬레이션은 공간의 인접성에 따라 군집을 잘 형성하는지와 시간에 따른 추세를 잘 반영하는지, 각 군집 별로 적절한 모수를 추정하는지 보기 위한 시뮬레이션이다.

5.1 시뮬레이션 설계

시뮬레이션에서 생성한 자료는 0부터 20 사이의 50개의 1차원 공간 정보와 9개의 시간 정보를 가지고 있다. s_i는 0부터 20을 50개로 균등하게 쪼갠 벡터 (0.4, 0.8,...,19.6, 20)이고, 시간은 1에서 9까지의 정수이다.($t = 1, \ldots, 9$) 그리고 μ_{it}와 σ_t^2는 다음 식을 따르도록 하였다.

$$\mu_{it} = \begin{cases} 20, & \text{if } 5 + \frac{5}{8}(t-1) < s_i < 10 + \frac{5}{8}(t-1) \\ 1, & \text{o.w} \end{cases} \tag{30}$$

$$\sigma_t^2 = 1 \tag{31}$$

시뮬레이션에 사용한 모형은 앞에서 설명한 공간 효과 모형과 시간 효과 모형을 합친 것으로 다음과 같다.

$$y_{it} \sim N(\mu_{it}, \sigma_t^2), \tag{32}$$

$$\mu_{it} \sim \sum_{k=1}^{K} w_{kt}(s_i)\delta_{\mu_{kt}^*},\ i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{33}$$

$$w_{kt}(s_i) = p_{kt}(s_i) \prod_{h=1}^{k-1} [1 - p_{ht}(s_i)], \tag{34}$$

$$p_{kt}(s_i) = \sigma(g_{kt}(s_i)),\ for\, k = 1, ..., K-1,\ p_K(s_{it}) = 1, \tag{35}$$

$$g_{kt}(s_i) = \sum_{j=1}^{J} \beta_{kjt} \mathcal{K}(s_i, \tilde{s}_j;\ \psi_k) + \beta_{k0t}, \tag{36}$$

$$\beta_{kj:} \sim \mathcal{N}(\mu_{\beta_{kj}}, \Sigma_{\beta_{kj}}),\ \beta_{k0:} \sim \mathcal{N}(\mu_{\beta_{k0}}, \Sigma_{\beta_{k0}}), \tag{37}$$

$$\mu_{k:}^* \sim \mathcal{N}(\mu_{\mu_k}, \Sigma_{\mu_k}),\ \log(\sigma^2_{:}) \sim \mathcal{N}(\mu_{\sigma^2}, \Sigma_{\sigma^2}), \tag{38}$$

$$\mu_{\beta_{kj}}, \mu_{\beta_{k0}}, \mu_{\mu_k}, \mu_{\sigma^2} \sim N(0, 1000), \tag{39}$$

$$\psi_k \sim Unif(0.05, 5) \tag{40}$$

이때 사용한 핵중심점인 $\tilde{s}_j$는 0부터 20을 5개로 균등하게 쪼갠 벡터(4, 8, 12, 16, 20)로 설정하였고 공간 좌표에 대한 척도 모수인 ψ_k는 0.05에서 5 사이의 균일분포를 가정하였다.

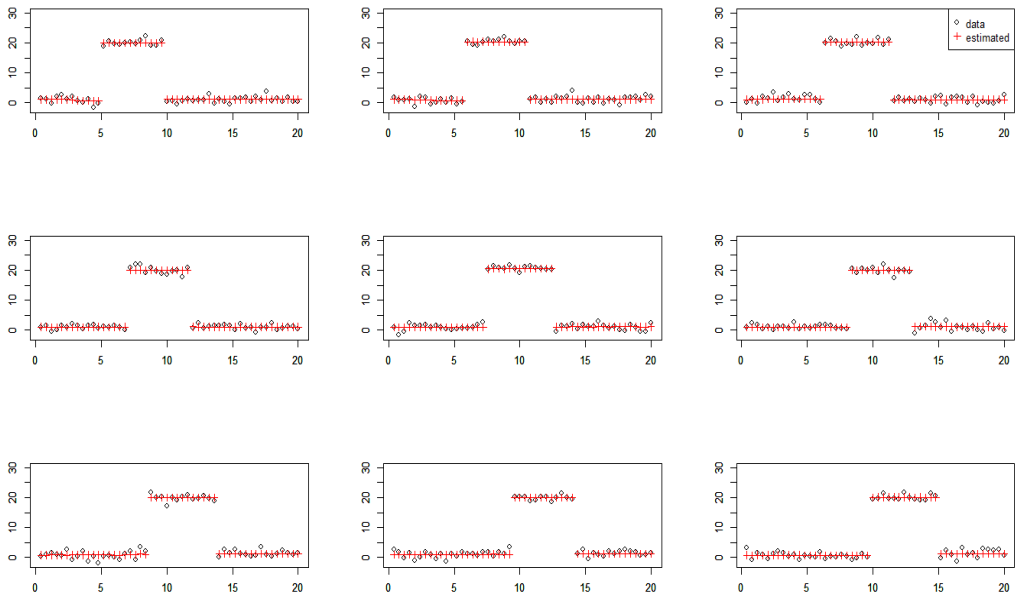


Figure 3: 시뮬레이션

Table 1: 시뮬레이션 결과

	$train$	$N_{miss}=1$	$N_{miss}=2$	$N_{miss}=3$	$N_{miss}=4$	$N_{miss}=5$
log_likelihood	-1.3773	-11.7386	-10.1091	-12.4665	-13.0377	-14.2535
정분류율	1	0.937	0.9426	0.9252	0.9265	0.9167

5.2 시뮬레이션 결과

시뮬레이션에 베이지안 시공간 모형을 적용했을 때의 적합 결과는 Figure 3과 같다. 이를 100번 반복하여 시뮬레이션하였고, 시간에 대해 보간한 결과도 담았다. 시간에 대한 보간은 각 모수들을 양옆 시간에 대해 평균 내어 계산하였다. 결과는 Table 1과 같다. N_{miss}는 보간한 시간의 개

수이다. 2에서 8 사이의 숫자 중에서 무작위로 N_{miss}개 만큼 시험자료를 뽑은 뒤에 그에 해당하지 않는 훈련자료로 모형을 학습시켰다. 이렇게 적합된 모수들로 시험자료를 보간하였다. log_likelihood는 자료와 적합된 값을 비교하여 평균내어 계산하였고 정분류율은 데이터가 맞게 분류된 비율로 정의하였다. $N_{miss} = 1, \ldots, 5$에서는 시험자료에 대해서만 결과를 산출했다. 마르코프 연쇄 몬테칼로(Markov Chain Monte Carlo; MCMC)를 이용한 샘플링을 하였고 10000개의 표본을 뽑아 5000개를 버리고 사용하였다. Table 1의 결과를 얻기 위해 사용한 대표값으로는 MCMC 표본의 중앙값으로 하였다. 시뮬레이션은 R 패키지 'nimble'을 통해 실행했고, 이는 collapsed gibbs sampler를 이용한 결과이다.

6 초미세먼지 자료에 모형 적용 결과

초미세먼지 자료에 대해 시뮬레이션과 같은 모형을 적용하였다. 핵중심점인 $\tilde{s}_j$는 다음의 20개 지역의 중심점으로 하였다.(서울시, 경기도, 강원도, 충청북도, 충청남도, 전라북도, 전라남도 ,경상북도, 경상남도, 제주도, 삼척, 포천, 부산시, 대구시, 포항시, 상주시, 대전시, 광주시, 여수시, 속초시) 핵중심점을 선정한 기준은 대도시이거나 지역적 특성이 있거나 우리나라에 골고루 배치되도록 하였다. Figure 4는 핵중심점을 지도에 표시한 그림이다. 속초와 삼척의 경우 영동지방이라는 지역적 특성이 있어 다른 지역보다 미세먼지가 적은 경향이 있어서 핵중심점에 추가하였고, 나머지 지역은 대도시이거나 핵중심점을 골고루 배치하기 위해서 추가하였다. MCMC 표본은 30000개를 뽑았고 처음 20000개의 표본은 버려서 추론하였다. Figure 5는 초미세먼지 자료를 적합한 결과를 그림으로 나타낸 것이다. 왼쪽의 그림은 초미세먼지 농도에 따라 자료를 지도에 표시한 그림이고 가운데 그림은 모형에 적합된 값을 지도에 나타낸 그림이다. 오른쪽 그림은 디리슈레 과정에 의해 군집분석된 결과이다. Table 2는 R 패키지에 있는 베이지안 시공간 자료 분석 모형들과 본 논문에서 소개한 모형을 성능 지표를 통해 비교한 결과이다. 여기서 MSE는 Mean Squared Error, RMSE는Root Mean Squared Er-

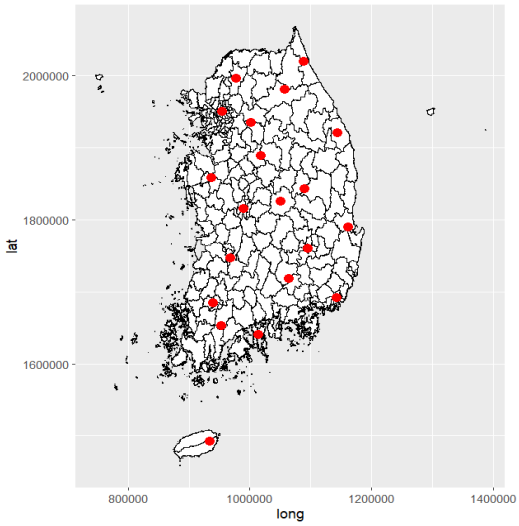


Figure 4: 핵중심점 지도

Table 2: 모형 비교

	spTimer	spBayes	INLA	spTDyn	my
MSE	12.4820	80.1505	32.3299	12.6411	**1.9852**
RMSE	3.5330	8.9527	5.6859	3.5554	**1.4090**
MAE	2.7145	7.1462	4.6099	2.7310	**1.0900**
MAPE	17.2037	47.0777	31.2526	17.2878	**7.1993**
BIAS	0.2241	**0.0006**	0.0072	0.2054	-0.0089
rBIAS	0.0127	**0.0000**	0.0004	0.0116	-0.0005
CRPS	2.1603	5.1543	4.2317	2.1702	**1.0723**

ror, MAE는 Mean Absolute Error, MAPE는 Mean Absolute Percentage Error, BIAS는 Bias, rBIAS는 Relative Bias, CRPS는 Continuous Ranked Probability Score를 뜻한다. 다음 지표들은 모두 낮을수록 좋은

모형이다.

$$MSE = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \hat{y}_{it})^2 \tag{41}$$

$$RMSE = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \hat{y}_{it})^2} \tag{42}$$

$$MAE = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}|y_{it} - \hat{y}_{it}| \tag{43}$$

$$MAPE = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{|y_{it} - \hat{y}_{it}|}{y_{it}} \tag{44}$$

$$BIAS = \frac{1}{NT}(\sum_{i=1}^{N}(\sum_{t=1}^{T}\hat{y}_{it} - y_{it}) \tag{45}$$

$$rBIAS = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{y}_{it} - y_{it})}{\sum_{i=1}^{N}\sum_{t=1}^{T}y_{it}} \tag{46}$$

$$CRPS = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\int_{-\infty}^{\infty}(F_{it}(x) - \mathbf{1}_{\{x>y_{it}\}})^2 dx,\ F_{it}(x) : CDF\ of\ N(\mu_{it}, \sigma_t^2). \tag{47}$$

여기서 $\hat{y}_{it} = \mu_{it}$이다.

BIAS와 rBIAS를 제외한 지표에서 본 논문에서 제시한 모형이 가장 좋은 성능을 보였다. BIAS와 rBIAS가 낮다는 것은 적합 값의 평균과 실제값의 평균이 비슷한 것을 의미할 뿐 좋은 모형임을 의미하는 것은

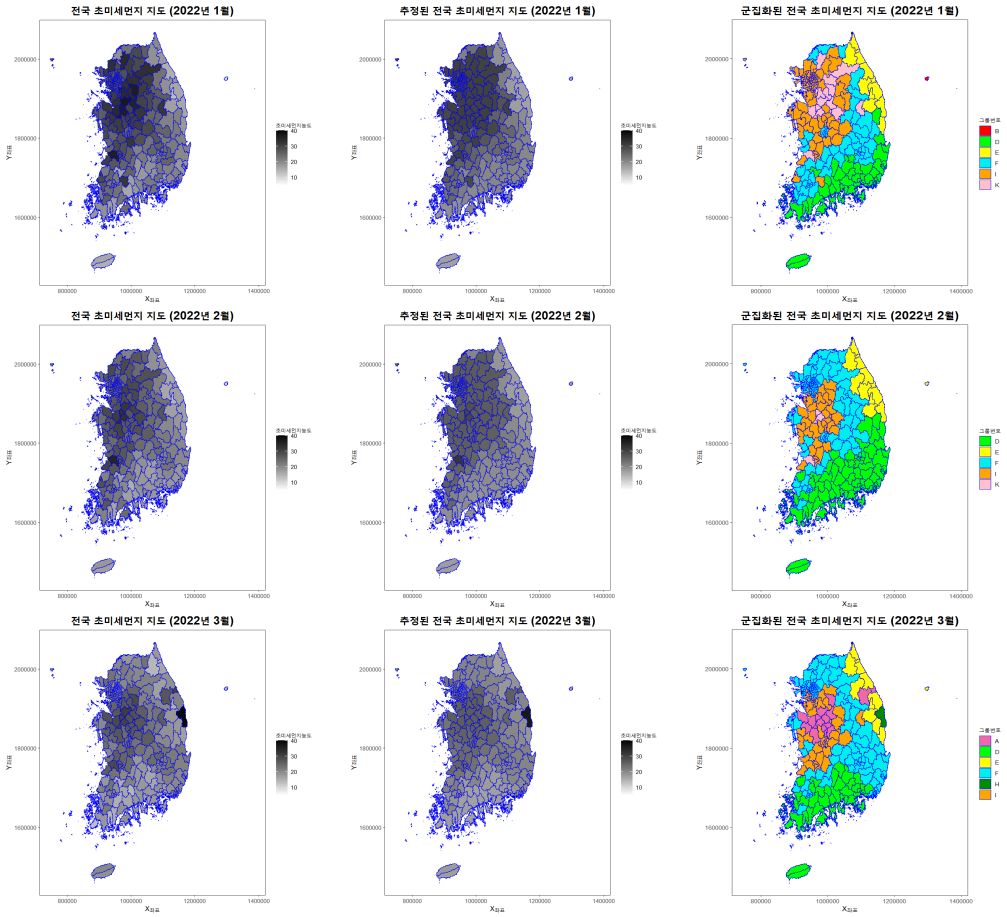


Figure 5: 초미세먼지 자료 모형 적용 결과

아니다.

7 결론

초미세먼지가 건강에 치명적인 영향을 미침에 따라 초미세먼지에 대한 관심은 계속 커지고 있고 정부도 그에 따른 대책을 계속 강구하고 있다. 본 논문에서는 초미세먼지 농도 자료에 대해 공간 분할과 초미세먼지 농도 추론을 하기 위한 통계모형을 소개하였다. 또 통계모형을 적용하고자 하는 초미세먼지 자료를 분석해 본 결과 공간적으로 인접한 지역에 대해서 초미세먼지 발생이 비슷한 경향이 있다는 것을 모란의 I 통계량을 통해 확인하였다. 그리고 자기상관계수를 시간 지연(time lag) 별로 확인한 결과 초미세먼지 발생량이 주기적으로 높고 낮아짐을 볼 수 있었고 그 주기가 12개월임을 확인하였다. 그래서 공간 인접성과 시간 주기성을 고려한 통계모형이 적합하다. 공간 인접성을 위해 로지스틱 막대분할 과정을 이용한 디리슈레 과정을 공간 효과 모형으로 사용하였다. 또한 시간 효과를 고려하기 위해 디리슈레 과정에 사용한 모수에 가우스 과정을 가정했다. 시간 주기성과 시간적인 추세를 고려하기 위해 가우스 과정의 핵함수로 방사형 기저함수와 주기함수를 합하여 사용하였다. 시뮬레이션을 통해 논문에서 제시한 모형이 자료를 잘 분할함을 확인하였다. 사후분포 추론으로는 MCMC 표본추출을 사용하였고 여러 척도를 통해 추정치와 실제값을 비교하였다. 그리고 같은 자료에 다른

모형을 적용함으로 논문에서 제시한 모형과 다른 모형의 성능을 비교하였다. 그 결과 본 논문의 통계모형이 다른 모형에 비해 좋은 성능을 보였고, 공간에 대해 군집한다는 점에서도 유익하다. 공간에 따른 이질성을 보이는 한국 초미세먼지에 대해 초미세먼지 농도에 따른 공간 군집화와 초미세먼지 농도 추론이 공간적으로 이질적인 초미세먼지에 대응하는 데 사용되기를 희망한다.

8 참고문헌

Qiu H, Yu IT, Tian L, Wang X, Tse LA, Tam W, and Wong TW (2012). Effects of Coarse Particulate Matter on Emergency Hospital Admissions for Respiratory Diseases: A Time-Series Analysis in Hong Kong, *Environmental Health Perspectives*, **120**.

Ito K, Mathes R, Ross Z, Nádas A, Thurston G, and Matte T (2011). Fine Particulate Matter Constituents Associated with Cardiovascular Hospitalizations and Mortality in New York City, *Environmental Health Perspectives*, **119**, 813–840.

Bae HJ (2014). Effects of Short-term Exposure to PM10 and PM2.5 on Mortality in Seoul, *Journal of Environmental Health Sciences*, **40**, 346–354.

Jeon CH, Cho DH, and Zhu L (2018). Exploring the Spatial Heterogeneity of Particulate Matter (PM10) using Geographically Weighted Ridge Regression (GWRR), *Journal of the Korean Cartographic Association*, **18**, 91–104.

Beloconi A, Chrysoulakis N, Lyapustin A, Utzinger J, and Vounatsou P

(2018). Bayesian geostatistical modelling of PM10 and PM2.5 surface level concentrations in Europe using high-resolution satellite-derived products, *Environment International*, **121**, 57–70.

Saez M and Barceló MA (2021). Spatial prediction of air pollution levels using a hierarchical Bayesian spatiotemporal model in Catalonia, Spain, *Environmental Modelling & Software*, **151**.

Sujit KS and Kanti VM (2005). A Bayesian kriged Kalman model for short-term forecasting of air pollution levels, *Journal of Applied Statistics*, **54**, 223–244.

Doreswamy, Harishkumar KS, Yogesh KM, and Gad I. (2020). Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models, *Procedia Computer Science*, **171**, 2057–2066.

Zhang L and Yang G (2022). Cluster analysis of PM2.5 pollution in China using the frequent itemset clustering approach, *Environmental Research*, **204**.

Rendana M, Idris WMR, and Rahim SA (2022). Clustering analysis of PM2.5 concentrations in the South Sumatra Province, Indonesia, using the Merra-2 Satellite Application and Hierarchical Cluster Method,

AIMS Environmental Science, **9**, 754–770.

Ferguson TS (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, **1**, 209–230.

Ren L, Du L, Carin L, and Dunson DB (2011). Logistic Stick-Breaking Process, *Journal of Machine Learning Research*, **12**, 203–239.

Luttinen J and Ilin A (2009). Variational Gaussian-process factor analysis for modeling spatio-temporal data, *Advances in Neural Information Processing Systems*, 1177–1185.

Ding M, He L, Dunson D, and Carin L (2012). Nonparametric Bayesian Segmentation of a Multivariate Inhomogeneous Space-Time Poisson Process, *Bayesian Analysis*, **7**, 813–840.

Noh HS, Park JS, Sim GS, Yu JE, and Chung YS (2014). Nonparametric Bayesian Statistical Models in Biomedical Research, *The Korean Journal of Applied Statistics*, **27**, 867—889.

Woo HY and Kim YH (2018). Noise reduction algorithm for an image using nonparametric Bayesian method, *The Korean Journal of Applied Statistics*, **31**, 555—572.

Meguelati K, Fontez B, Hilgert N, and Masseglia F (2019). Dirichlet Process Mixture Models made Scalable and Effective by means of Massive Distribution, *SAC '19: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 502–509.

Lim HC (2022). Gaussian Process Regression and Its Application to Mathematical Finance, *Journal for History of Mathematics*, **35**, 1–18.

Valpine P, DanielTurek, Paciorek CJ, Anderson-Bergman C, Lang DT, and Bodik R (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE, *Journal of Computational and Graphical Statistics*, **26**, 403–413.