FDA Final report

Chaeyun YEO, 248STG04

December 2024

1 Abstract

북한은 기상 관측소가 남한에 비해 적은 편이고 공개하지 않는 경우도 많다. 따라서 관측소가 없는 지역에서는 기상 데이터를 얻기가 어렵다. 따라서 Spatial functional Data Anaylsis를 활용하여 기상 데이터를 분석해보고자 한다. 먼저, 온도를 예측해 보았고 이때 Ordinary kriging 기법을 사용하였다. 이를 통해 북한의 장진 지역의 온도를 예측해보고 mse값을 구하였다. 그 다음으로는 북한에서 공개하지 않는 PM10 농도를 예측해 보았는데, 이때 풍속을 활용해 Cokriging기법을 사용하였다.

2 Introduction

2.1 분석 목적

1. 북한의 기상 관측소는 총 27개가 존재한다. 따라서 다른 기상관측소가 없는 다른 지역의 주민들은 기상 데이터를 상대적으로 얻기 어렵다는 단점이 있다. 따라서 Ordinary kriging을 활용하여 북한의 다른 지역들의 **기온**을 예측하는 것이 목표이다.

2. 북한은 미세먼지 데이터를 제공하고 있지 않다. 따라서 북한의 **미세먼지 농도**를 예측하는 것이 목표이다. 미세먼지 농도를 예측하는데는 바람이 영향을 미칠 것이라 생각해, 풍속과 미세먼지 농도의 cokriging을 활용해 분석을 진행할 예정이다.

2.2 데이터 수집

1. **Temperature Data** : 북한 27개의 지점에서 관측된 기온 데이터 (출처 : 기상청 기상자료개방포털)
   - 2023년 11월 1일 03시 2024년 11월 1일 00시 (3시간 간격)

2. **Location data** : 북한 및 대한민국의 위도와 경도 (출처 : 기상청 날씨누리)

3. **PM10 Data** : 대한민국 23개 지점의 미세먼지 농도 (출처 : 기상청 기상자료 개방 포털)
   - 2023년 12월 10일 (1시간 간격)

4. **Wind Data** : 대한민국 23개 지점의 풍속 데이터 (출처 : 기상청 기상자료 개방 포털)
   - 2023년 12월 10일 (1시간 간격)

3 Prediction of Temperature

3.1 Data Preprocessing and EDA

- Temperature data와 location data를 형식에 맞게 전처리를 하였다. 또한 NA의 값은 이전 시간대와 이후 시간대의 평균으로 대체하여 입력하였다.
  - Weather_array : 시간 X 지점명 (2927 X 27)
  - coordinates : 지점명 X (위도, 경도) (27 X 2)

```
> head(weather_array)
                     강계 개성 구성 김책 남포 사리원 삼지연 선봉 수풍 신계 신의주 신포 안주
2023-11-01 3:00  10.6 17.0 14.9 13.4 19.2   18.0    4.4 13.0 13.6 17.0   16.7 14.6 17.5
2023-11-01 6:00  10.4 17.1 15.3 12.3 17.8   18.5    3.3 10.4 13.0 17.3   16.2 15.2 17.3
2023-11-01 9:00  10.8 17.4 15.1 14.2 17.7   18.6    5.1 11.3 13.5 18.0   16.4 14.1 18.0
2023-11-01 12:00 13.0 19.2 17.9 19.9 23.3   22.6    6.9 12.0 17.4 20.5   17.9 17.0 19.5
2023-11-01 15:00 16.2 19.7 20.2 19.2 22.5   23.6    5.6  8.3 20.3 21.4   20.6 18.8 19.9
2023-11-01 18:00 15.2 19.3 18.3 14.6 18.7   21.0    3.7  9.5 15.8 19.2   19.6 18.4 18.0
```

```
> head(coordinates)
       W.longitude N.latitude
선봉       130.4000    42.3167
삼지연      128.3167    41.8167
청진       129.8167    41.7833
중강       126.8833    41.7833
혜산       128.1667    41.4000
강계       126.6000    40.9667
```

Figure 1: 기온 데이터인 weather_array data와 위치 데이터인 coordinate data

Figure 2에는 북한의 26개 지점의 위치와 적합한 모형의 성능을 평가할 **장진** 지역을 지도에 표시해보았다.

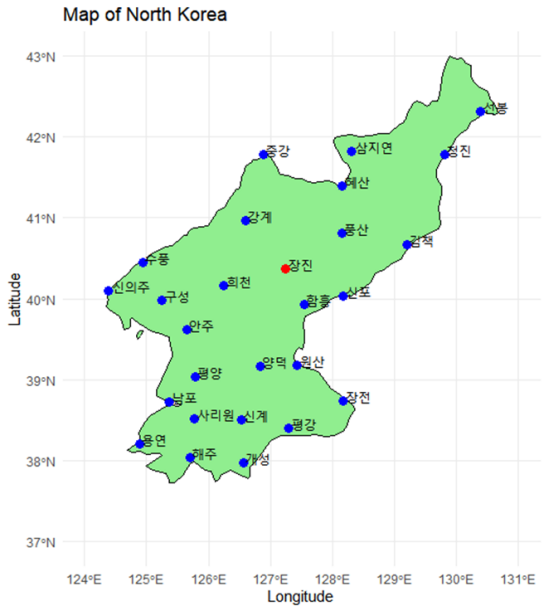

Figure 2: 북한의 26개 지점의 위치 및 장진 위치

- 고도와 기온의 관계

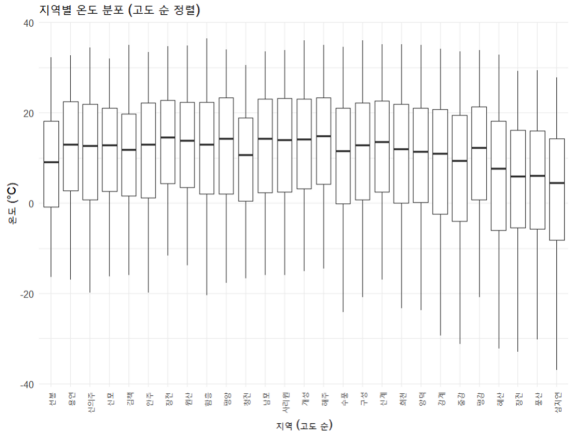

Figure 3: 고도와 기온의 관계, 관측 지점 위치와 관계 없이 고도가 높아질 수록 기온이 떨어지는 것을 확인할 수 있다.

관측 지점의 고도는 기온에 영향을 미친다. 북극으로 갈 수록 온도가 낮아지는 것이 자명하듯이, 고도가 높아질 수록 온도가 떨어진다는 사실도 명백하다. 따라서 관측 지점의 해발고도를 무시하는 것은 어려워 보인다. 보통 1000m당 기온은 6.5도 감소한다. 따라서 해발고도와 온도의 관계를 계산해 온도를 보정한 adjusted_temperatures를 생성하였다. 즉, 모든 지점의 관측소가 고도가 0인 기준으로 온도를 보정하였다.

3.2 Preliminaries

이번 분석에서는 second-order isotropically stationary를 가정한 후, spatial functional random process를 다룬다. 시간이나 공간에 관계없이 평균과 분산이 일정하고, 두 지점 간의 공분산이 그들 간의 거리에만 의존한다. 또한 모든 방향에 대해 동일한 통계적 특성을 가진다. 즉, 다음 조건을 만족하는 경우이다:

- (i) $E[\chi_s(t)] = m(t)$ 및 $\mathrm{Var}[\chi_s(t)] = \sigma^2(t)$ $(\forall s \in D, \forall t \in T)$,
- (ii) $\mathrm{Cov}[\chi_s(r), \chi_v(t)] = C(\|s - v\|, r, t)$ $(\forall s, v \in D, \forall r, t \in T)$,

여기서 $\|\cdot\|$는 유클리드 거리이다.

3.3 Ordinary Kriging

curve kriging predictor 모형은 다음과 같다:

$$\hat{\chi}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i \chi_{s_i}(t), \quad t \in T.$$

이는 함수형 데이터의 일반 크리깅 (OKFD, ordinary kriging for functional data) 이라고도 불린다.

최적의 크리깅 가중치 $\lambda_1, \ldots, \lambda_n \in R$는 다음 조건을 만족한다:

$$\sum_{i=1}^{n} \lambda_i = 1.$$

이때 최적의 가중치 λ는 MISE를 최소화 한다.

$$\text{MISE}(s_0) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\|s_i - s_j\|) + 2\sum_{i=1}^{n} \lambda_i \gamma(\|s_i - s_0\|)$$

where

$$\gamma(h) = \frac{1}{2} E\left[\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 \, dt\right], \quad h = \|s_i - s_j\|.$$

trace-variogram을 추정하기 위해, 모멘트 방법을 사용하여 다음의 추정량을 도입한다:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 \, dt,$$

여기서 $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$이며, $|N(h)|$는 이 집합의 원소 개수를 나타낸다. $N(h)$가 드문 경우, $h \pm \epsilon$ 범위를 이용해 안정적인 추정치를 얻는다. 추정된 trace-variogram 값을 크리깅 가중치 λ_i 계산에 적용하고, $\tilde{\chi}_{s_i}(t)$로 $\chi_{s_i}(t)$를 대체하여 최종적인 predictor를 구성한다.

3.4 Application : Ordinary Kriging

3.4.1 Ordinary Kriging

먼저, 기온을 시각화하여 기온의 추세를 살펴보았다.

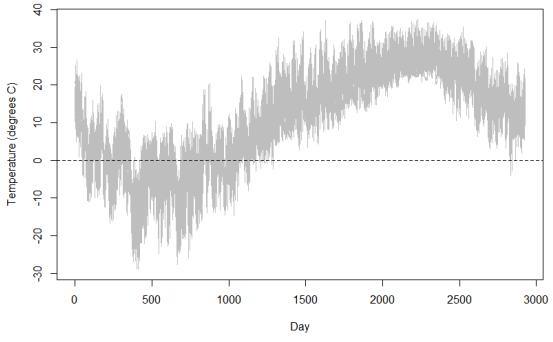

Figure 4: 북한 27개 지점의 기온 추세 그래프

이는 작년 11월 부터 올해 11월 까지의 기온 그래프인데, 겨울에 기온이 떨어졌다가 여름에 오르고, 최근 다시 기온이 떨어지는 것을 확인할 수 있다. 이 기온 데이터를 Ordinary Kriging을 활용해 예측을 해 보았다. 북한의 **장진** 지역을 예측하는 것을 목표로 하기 때문에,

new.coord 변수에는 장진의 위도, 경도 값을 넣어주었다. Tempe.26은 26개 지점의 기온 데이터이고 coordinates는 26개 지점의 위도 경도 데이터이다.

```
okfd.res<-okfd(coord=coordinates,data=Tempe.26,argvals=argvals,
    nbasis=nbasis,new.coord=coord.0)
```

okfd함수는 공간적으로 분포된 데이터를 바탕으로 함수 데이터를 생성하고 kriging 기법을 통해 새로운 위치에서의 값을 예측한다. 얼마나 많은 수의 basis function이 영향을 주는지 알아보기 위해 basis의 수는 50, 100, 150, 200로 지정하였다.
또한 2개의 semivariogram models (exponential,spherical) 을 empirical trace-semivariogram에 적용시켰다.

- Exponential:

$$\phi(h) = \exp\left(-\frac{h}{\rho}\right)$$

- Spherical:

$$\phi(h) = \begin{cases} 1 - \frac{3h}{2\rho} + \frac{h^3}{2\rho^3}, & \text{if } h \leq \rho \\ 0, & \text{if } h > \rho \end{cases}$$

Exponential 모델은 공간적 상관성이 비교적 긴 거리까지 서서히 감소하는 특성을 가진다. 이 모델은 상관성이 급격하게 감소하지 않고 점진적으로 감소하는 경우에 적합하다. Spherical 모델은 가까운 거리에서 높은 상관성을 보이다가, 일정 거리 이상에서는 상관성이 급격히 줄어드는 특성을 가진다.
이렇게 총 4 x 2 = 8 OKFD모델을 적용시켰다. 각 모델의 비교는 MSE값을 사용해 비교하였다.

cov.model	**# of Basis = 50**	**# of Basis = 100**	**# of Basis = 150**	**# of Basis = 200**
spherical	26.8	24.2	22.4	21.2
gaussian	26.7	24.8	23.2	22.1

Table 1: number of basis와 cov.model에 따른 MSE

basis가 늘어나면 계산 속도는 증가하지만 예측력이 좋아지는 것으로 파악할 수 있었다. 결과적으로, basis가 200개 일때, 그리고 spherical trace-semivariogram일때 가장 예측을 잘 하는 것으로 파악할 수 있었다.

Figure 5의 semivariogram을 확인해보면 거리가 조금 멀어졌을 때 지점 간의 독립이 빠르게 보이는 것을 확인할 수 있다. 이는 북한의 크기가 크지 않기 때문이라고 판단이 된다. 오른쪽의 예측과 실제 그래프를 확인해보면, 실제와 예측이 크게 차이가 나지 않는 것으로 확인할 수 있다.

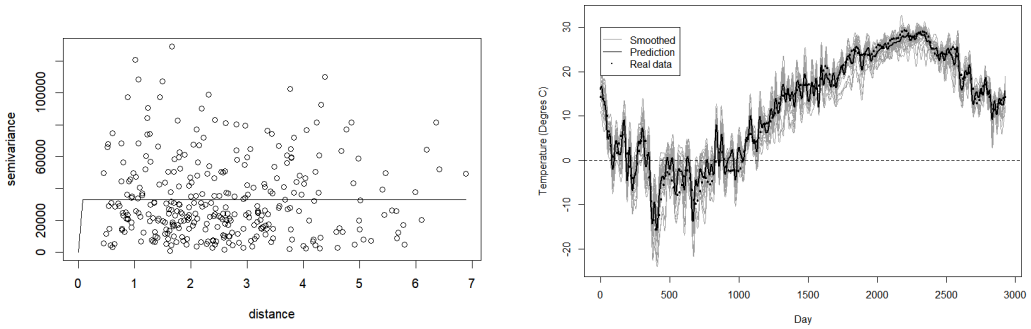

Figure 5: spherical cov model의 semivariogram(Left), 예측과 실제 기온 그래프(Rright)

3.4.2 Cross-validation analysis

okfd.cv 함수는 Ordinary Kriging for Functional Data (OKFD)의 cross-validation을 수행하여 최적의 매개변수를 선택하는 데 사용된다. 따라서 nbasis의 수를 바꿔가면서 cross-validation을 해보았다. nbasis의 수는 5, 105, 205, 305, 405 로 지정하였다,

```
n<-dim(Tempe.26)[1]
argvals<-seq(1,n, by=1)
array.nbasis <- seq(5,by=100, length = 5)
okfd.cv.res <- okfd.cv(coord=coordinates, data=Tempe.26, argvals
   =argvals, array.nbasis=array.nbasis, max.dist.variogram=NULL,
   nugget.fix=NULL)
```

결과는 다음과 같다.

- k.opt: 405, 최적의 기저 함수 개수
- krig.CV: 모든 교차 검증 반복에서 계산된 크리깅 예측값 배열.
- MSE.CV: 각 기저 함수 개수에 대해 계산된 MSE 값.
- MSE.CV.opt: 21.24, 최적의 MSE 값.

Figure 6을 보면, nbasis가 증가할 수록 MSE값이 줄어들어 nbasis가 높은 것이 좋지만, nbasis가 높아지면 계산량이 너무 많아진다는 단점이 있다.

이렇게 OKFD를 활용하여 기온을 예측한 결과, RMSE의 값은 거의 4.5 근처로 예측력이 좋은 것을 확인할 수 있다. basis function의 갯수가 증가할 수록 MSE의 값은 작아지지만, 계산량이 증가하므로, 적정한 basis function의 갯수를 정하는 것이 중요하다. okfd와 okfd.cv function의 코드는 따로 첨부하였다.

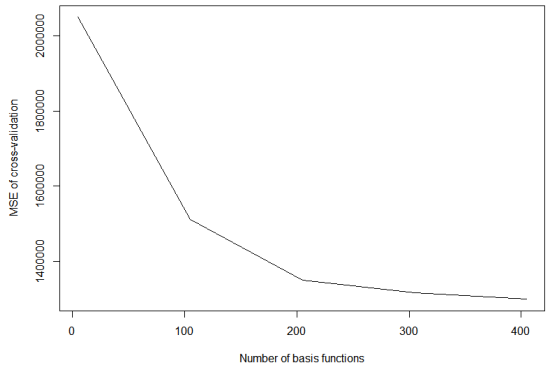

Figure 6: number of basis function이 증가할수록 MSE값이 감소하고 있다.

4 Prediction of PM10

미세먼지는 풍속의 영향을 많이 받는데. 만약에 풍속이 강하다면 미세먼지 농도는 줄어들고, 풍속이 약하다면 농도는 심해진다. 따라서 풍속을 고려해, 북한의 미세먼지 농도를 예측해보려고 한다. 대한민국의 기상 데이터 (미세먼지, 풍속) 를 활용하여 모델을 적합시킨 후, 이를 북한에 적용시키려 한다. 결론적으로는, 하루동안 시간대별로 수집한 풍속 데이터를 활용해 미세먼지 농도를 예측하는 것이 목표이다.

4.1 Data Description

- **PM10 (scalar data)**: 대한민국의 각 지점에서 측정된 미세먼지 농도를 나타냄.
- **Wind (functional data)**: 대한민국의 각 지점에서 측정된 풍속 데이터.
- **Coord**: 각 지점의 공간적 위치 정보. (대한민국, 북한)

4.2 Cokriging

Cokriging은 kriging의 일반화로, 다변량 공간 예측을 처리하기 위해 개발되었다. Cokriging은 두 개 이상의 상호 상관된 random field 간의 공분산을 활용하여 주 변수를 더 정밀하게 예측할 수 있다. 이 방법은 비용 및 예측 정확성 측면에서 kriging보다 우수한 것으로 나타났다. 또한 Cokriging은 단순히 다변량 random field data뿐 아니라 functional random field data를 다룰 수 있으며 이번 데이터 분석에서도 functional random field data를 다룰 예정이다. 이번 분석에서는 풍속인 functional random field data를 보조 변수 (secondary variables)로 활용하여 스칼라 주 변수 PM10를 예측하는 Cokriging 기법을 사용한다. 이를 기반으로 PM10과 Wind 곡선을 CLM을 사용하여 연관성을 분석하는 방법론을 사용한다. 해당 Cokriging predictor는 기존 Cokriging predictor의 확장 형태로, functional data를 basis function로 표현한 후, 이를 통해 예측한다.

4.2.1 CLM (Co-Regionalized Linear Model)

Cokriging에서 Co-Regionalized Linear Model (CLM)은 auto-correlation function 및 cross-correlation function를 추정하는 데 사용된다. CLM은 변수 간의 공간 상관 관계를 모델링하며, 서로 다른 변수들 간의 상호작용을 통해 예측의 정확도를 높이는 역할을 한다. 특히 다변량 데이터나 functional 데이터를 처리할 때 유용하게 활용되며, 이번 분석에서는 PM10과 Wind 간의 관계를 모델링하는 데 사용된다.

4.2.2 Cokriging Predictor Using Functional Secondary Variables

secondary variables의 개수가 매우 많다고 가정하면, 이러한 변수들은 functional variable로 대체될 수 있으며, 이를 통해 random field의 secondary variables을 사용하는 cokriging predictor를 정의할 수 있다.
scalar response를 가지는 functional linear model에 대해서, 식 (1)에서 $m \to \infty$로 가정한다.
이때, 식 (1)에 정의된 매개변수 β_{ij}는 함수 $\beta_i(t)$로 대체될 수 있으며, secondary variables $X_2(s_i), \ldots, X_m(s_i)$는 functional variable $X_{s_i}(t)$로 대체될 수 있다.
따라서, unsampled location s_0에서의 $X_1(s)$에 대한 cokriging predictor는 다음과 같은 형태를 가진다:

$$X_1(s_0) = \sum_{i=1}^{n} \lambda_i X_1(s_i) + \sum_{i=1}^{n} \int \beta_i(t) X_{s_i}(t)\, dt, \tag{1}$$

여기서 $X_1(s_0)$는 unsampled site s_0에서의 primary variable의 예측값을 나타내며, λ_i는 예측에 대해 scalar 변수 $X_1(s_i)$가 미치는 영향을 제공하고, $X_{s_i}(t)$는 site s_i에서의 functional variable이며, $\beta_i(t)$는 $X_{s_i}(t)$가 예측에 미치는 가중치를 나타내는 functional parameter이다. ($i = 1, \ldots, n$).

매개변수 추정을 위해 B-splines 및 Fourier 유형의 basis functions를 기반으로 하는 접근법을 사용한다. Functional variables와 parameters는 다음과 같이 확장된다:

$$X_{s_i}(t) = \sum_{j=1}^{k} a_{ij}\phi_j(t) = \mathbf{a}_i^T \boldsymbol{\phi}(t), \tag{2}$$

$$\beta_i(t) = \sum_{j=1}^{k} \beta_{ij}\phi_j(t) = \boldsymbol{\beta}_i^T \boldsymbol{\phi}(t), \tag{3}$$

여기서 $\boldsymbol{\phi}(t) = \begin{bmatrix} \phi_1(t) & \ldots & \phi_k(t) \end{bmatrix}^T$는 basis functions의 벡터이며, $\mathbf{a}_i$, $\boldsymbol{\beta}_i$는 least squares를 통해 추정된 coefficients의 벡터이다.

Basis function expansions을 식 (2)에 대입함으로써, predictor는 다음과 같이 변환된다:

$$X_1(s_0) = \sum_{i=1}^{n} \lambda_i X_1(s_i) + \sum_{i=1}^{n} \lambda_i \mathbf{a}_i^T \mathbf{W} \boldsymbol{\beta}_i = \sum_{i=1}^{n} \lambda_i X_1(s_i) + \sum_{i=1}^{n} \sum_{j=1}^{k} \beta_{ij} a_{ij}, \tag{4}$$

여기서 β_{ij} 는 식 (3) 에서 정의된 값이며, $\mathbf{a}_i$ 는 다음과 같이 주어진다:

$$\mathbf{a}_i = \Big[a_{i1}, \ldots, a_{ij}, \ldots, a_{ik}\Big]^T. \tag{5}$$

4.2.3 Cokriging Prediction of PM10 Using wind Curves

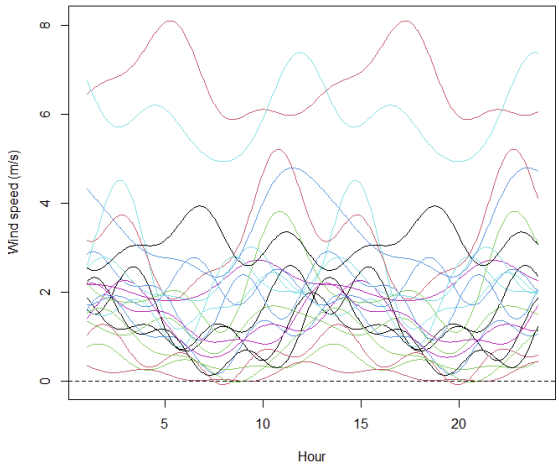

Figure 7: functional data로 변환한 wind data

시계열 데이터를 smoothing하기 위해 다양한 basis function이 사용될 수 있는데, 주기적 변동을 잘 설명할 수 있는 Fourier basis를 사용하였다. 데이터를 smoothing하기 위해 k = 7 인 Fourier basis를 사용하였고, 이는 Figure 7에서 확인할 수 있다. 해당 그래프를 보면 특정 지역은 높은 풍속을 보이는 것을 확인할 수 있다.
결과적으로, 관측소 마다 PM10 값을 가지며, Fourier basis로 wind data 를 fitting한 결과 얻어진 7개의 coefficients로 구성된 데이터를 가지게 된다.
이를 기반으로 CLM(Cross-Linear Model)을 피팅하였다. 이때 모든 simple variogram 및 cross-variogram을 설명하기 위해 Gaussian 모델을 사용하였다.

```
# Linear model of coregionalization 공간적( 상관구조모델링  )
k=nbasis
coeficients<-matrix(datafd$coefs,nrow=k)
coeficients
coef<-t(coeficients)
coef<-cbind(coord,coef)
coef
datcok<-cbind(coef, pm10)
datcok<-as.data.frame(datcok)
datcok
n2<-paste(expression(phi),1:(k+1), sep="")
names(datcok)<-c("x","y",n2)
coordinates(datcok)= ~x+y
```

```
g<-NULL
for (i in 1:(k+1))
{
  g <- gstat(g,formula= as.formula(paste(n2[i],"~1")), data=
     datcok)
}
v <- variogram(g) # 을variogram 계산
plot(v, xlab="Distance") # 을variogram 시각화
g <- gstat(g, model=vgm(4000, "Gau", 80000), fill.all=TRUE)
g.fit <- fit.lmc(v,g,   fit.lmc=TRUE, correct.diagonal = 1.01)
# fit.함수는lmc 선형모델의를  LMC 적합시키는함수
plot(v, g.fit, xlab="Distance", ylab="Simple and cross
   semivariance", main ="")
names(g.fit)
g.fit$model
```

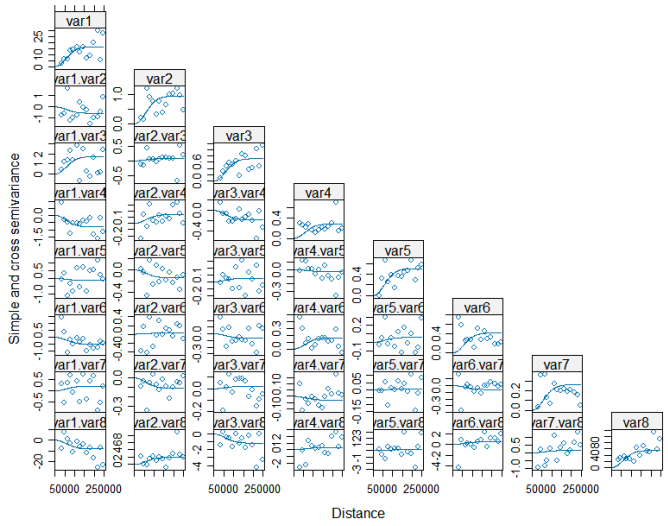

Figure 8: spatial correlation을 나타내는 semivariogram 시각화

Figure 8의 X축은 데이터 포인트 간의 거리이고 Y축은 각 변수 또는 변수 조합에 대한 세미분산 값이다. 일반적으로 거리 증가에 따라 세미분산이 증가하는 경향이 있다. Cross semivariance은 두 변수 간의 상관 관계를 측정한다. 거리 증가에 따라 세미분산이 어떻게 변화하는지, 그리고 모델이 실제 데이터와 얼마나 잘 맞는지 평가할 수 있다.

아래의 코드는 위에서 적합시킨 g.fit 모델을 활용하여 북한의 27개 지점을 커버하는 10,000개의 지역의 미세먼지 (PM10) 농도를 예측하는 코드이다. 각 점에서 cokriging 예측을 수행하였고 예측 분산을 추정하였다.

```
# 북한의지점  location 데이터(coord)를 가져와분석
grid = expand.grid(Longitude=seq(min(coord[,1]),max(coord[,1])
    ,5000), Latitude=seq(min(coord[,2]),max(coord[,2]),5000))
grid = as.data.frame(grid)
summary(grid)
names(grid) = c("x","y")
coordinates(grid)=~x+y
prediction= predict(g.fit, newdata=grid)
summary(prediction)
names(prediction)

# PM10 prediction

pm10_prediction=prediction["var8.pred"]
```

북한의 만개의 지점에서 미세먼지 농도를 예측한 값과 예측 분산을 요약하면 다음과 같다.

```
summary(pm10_prediction_var) #
Object of class SpatialPointsDataFrame
Coordinates:
        min       max
x  123341.3  663341.3
y 3687451.0 4232451.0
Is projected: NA
proj4string : [NA]
Number of points: 11990
Data attributes:
    var8.var
 Min.   : 0.0008
 1st Qu.: 8.5624
 Median :30.6784
 Mean   :30.7585
 3rd Qu.:53.8249
 Max.   :59.5803
```

Figure 9는 북한 지역의 미세먼지 예측값과 예측 분산을 지도로 나타낸 것이다. 북한 위도 경도를 모두 커버하는 만개의 그리드를 생성해 예측을 진행했기 때문에, 구체적인 지역 위치는 찾기 어렵지만 대략적인 미세먼지 농도 분포를 알 수 있다.

중간 중간 미세먼지 농도가 낮은 곳도 보이지만, 전반적으로 북한의 미세먼지 농도는 꽤 높은 것을 확인할 수 있다.

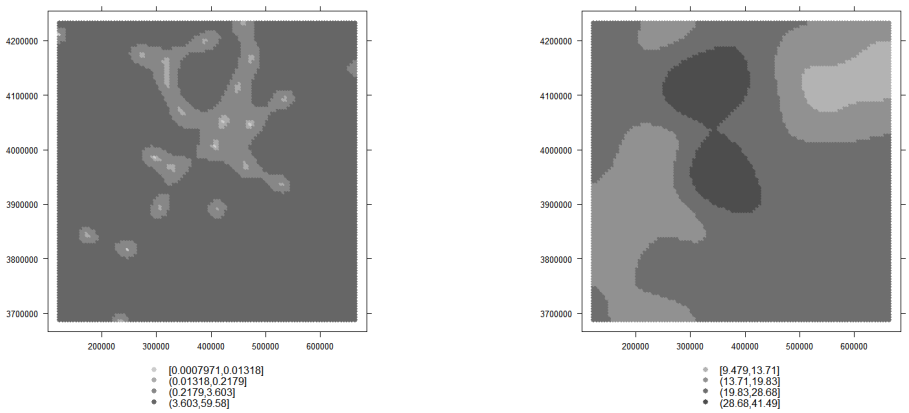

Figure 9: 미세먼지 예측 값(Left), 예측 분산 (Rright)

5 Conclusion

이렇게 북한의 기온과 미세먼지 농도를 OKFD와 Cokriging을 활용하여 예측해 보았다. 북한이 크지 않아서, 공간적 데이터 분석이 다른 분석에 비해 매우 유의미한 결과를 도출하진 못했지만, 비교적 얻기 쉬운 지리적 정보로 기상 데이터를 예측할 수 있다는 점에서 의의가 있다. 또한 OKFD의 경우 MSE값이 낮게 나와, 실질적으로 의미있는 분석이라고 할 수 있다. 미세먼지 농도 예측에서는, 비록 북한의 미세먼지 농도가 없어 RMSE값을 비교해보진 못했지만, 위치 데이터와 풍속 데이터로 미세먼지를 예측했다는 점에서 의의가 있다.
공간 데이터 분석에는 OKFD 뿐만 아니라 PWFK (Pointwise Functional Kriging), FKTM (Functional kriging total model) 등 여러 방법이 있는데 다음 기회에 이러한 모델들도 적합해 보고 어느 모형이 가장 예측력이 좋은지 검증해보고 싶다.

References

[1] Jorge Mateu., Ramón Giraldo. (2022). *Geostatistical Functional Data Analysis*. John Wiley Sons Ltd.

[2] Ramón Giraldo.,Luis Herrera., Víctor Leiva (2020). *Cokriging Prediction Using as Secondary Variable a Functional Random Field with Application in Environmental Pollution*. Mathematics.