FDA Final report

Chaeyun YEO, 248STG04

December 2024

1 Abstract

North Korea has fewer meteorological observation stations than South Korea, and in many cases, the data are not publicly available. Therefore, it is difficult to obtain meteorological data in regions without observation stations. In this study, meteorological data are analyzed using Spatial Functional Data Analysis. First, temperature is predicted using the Ordinary Kriging method, and the temperature of the Jangjin region in North Korea is estimated along with the mean squared error (MSE). Next, PM10 concentrations, which are not publicly available in North Korea, are predicted using a Cokriging method incorporating wind speed.

2 Introduction

2.1 Objectives of Analysis

1. There are a total of 27 meteorological observation stations in North Korea. As a result, residents in regions without nearby observation stations have relatively limited access to meteorological data. Therefore, the objective is to predict **temperature** in other regions of North Korea using Ordinary Kriging.

2. North Korea does not provide PM10 data. Therefore, the objective is to predict **PM10 concentrations**. Since wind is expected to influence PM10 concentrations, the analysis will be conducted using cokriging of wind speed and PM10.

2.2 Data Collection

1. **Temperature Data** : Temperature data observed at 27 locations in North Korea (Source: Korea Meteorological Administration Data Portal)
   - November 1, 2023 03:00 November 1, 2024 00:00 (3-hour intervals)

2. **Location data**：Latitude and longitude of locations in North Korea and South Korea (Source: KMA Weather Service)

3. **PM10 Data**：PM10 concentrations at 23 locations in South Korea (Source: Korea Meteorological Administration Data Portal)
   - December 10, 2023 (1-hour intervals)

4. **Wind Data**：Wind speed data at 23 locations in South Korea (Source: Korea Meteorological Administration Data Portal)
   - December 10, 2023 (1-hour intervals)

3 Prediction of Temperature

3.1 Data Preprocessing and EDA

- Temperature data and location data were preprocessed into the required formats. In addition, missing values were replaced with the average of the previous and subsequent time points.
  - Weather_array : Time × Location (2927 × 27)
  - coordinates : Location × (Latitude, Longitude) (27 × 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: Temperature data (weather_array) and location data (coordinates)

Figure 2 displays the locations of 26 stations in North Korea and the **Jangjin** region, which is used to evaluate model performance.

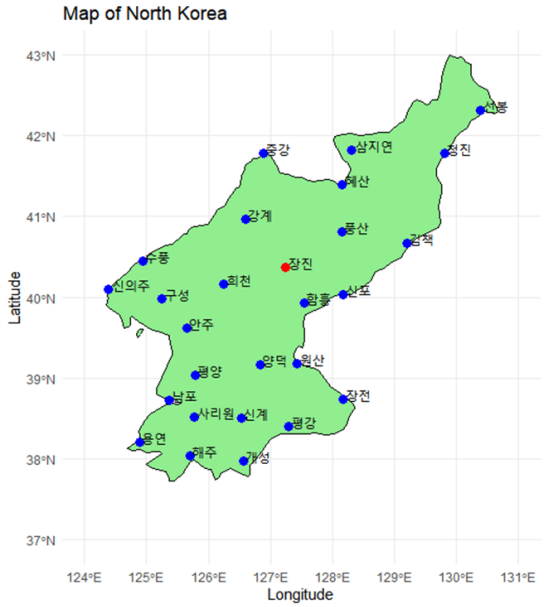

Figure 2: Locations of 26 stations in North Korea and the Jangjin site

- Relationship between altitude and temperature

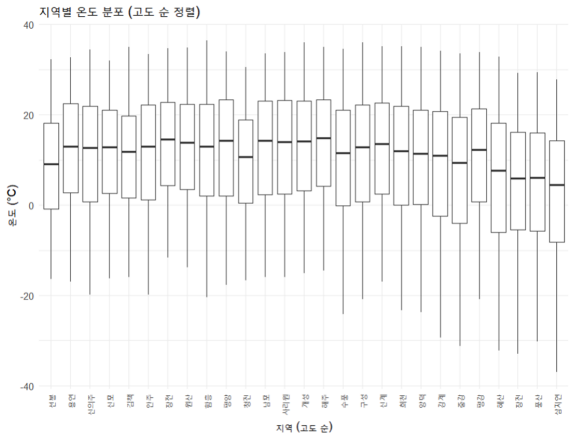

Figure 3: Relationship between altitude and temperature, showing that temperature decreases as altitude increases regardless of station location.

The altitude of observation stations affects temperature. Just as temperature decreases toward the polar regions, it is also evident that temperature decreases as altitude increases. Therefore, it is difficult to ignore the elevation of observation stations. In general, temperature decreases by 6.5 degrees per 1000 meters. Accordingly, temperatures were adjusted based on the relationship between altitude and temperature to create `adjusted_temperatures`. In other words, the temperatures at all stations were adjusted to a reference altitude of zero.

3.2 Preliminaries

In this analysis, a second-order isotropically stationary assumption is imposed, and a spatial functional random process is considered. The mean and variance are constant regardless of time or space, and the covariance between two locations depends only on the distance between them. In addition, the process exhibits identical statistical properties in all directions. That is, the following conditions are satisfied:

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

where $\|\cdot\|$ denotes the Euclidean distance.

3.3 Ordinary Kriging

The curve kriging predictor is defined as follows:

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

This approach is also referred to as Ordinary Kriging for Functional Data (OKFD).

The optimal kriging weights $\lambda_1, \ldots, \lambda_n \in R$ satisfy the following constraint:

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

Under this constraint, the optimal weights λ minimize the Mean Integrated Squared Error (MISE):

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

where

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

To estimate the trace-variogram, a moment-based estimator is introduced:

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

where $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$, and $|N(h)|$ denotes the number of elements in this set. When $N(h)$ is sparse, a tolerance range of $h \pm \epsilon$ is used to obtain a stable estimate. The estimated trace-variogram values are then used to compute the kriging weights λ_i, and $\chi_{s_i}(t)$ is replaced by $\tilde{\chi}_{s_i}(t)$ to construct the final predictor.

3.4 Application : Ordinary Kriging

3.4.1 Ordinary Kriging

First, temperature data were visualized to examine overall temperature trends.

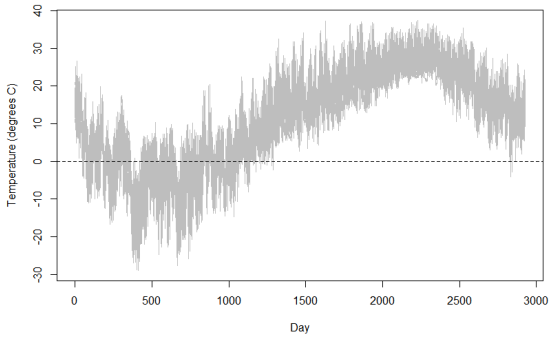

Figure 4: Temperature trends across 27 locations in North Korea

The figure shows temperature trends from November of last year to November of this year, where temperatures decrease during winter, increase during summer, and decrease again in recent months. Using this temperature data, predictions were performed via Ordinary Kriging. Since the objective is to predict temperature at the **Jangjin** region, the latitude and longitude of Jangjin were assigned to the `new.coord` variable. `Tempe.26` represents temperature data from 26 locations, and `coordinates` contains the corresponding latitude and longitude information.

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

The `okfd` function generates functional data based on spatially distributed observations and predicts values at new locations using the kriging method. To investigate the influence of the number of basis functions, the number of basis functions was set to 50, 100, 150, and 200.
In addition, two semivariogram models (exponential and spherical) were fitted to the empirical trace-semivariogram.

- Exponential:

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

- Spherical:

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

The exponential model exhibits spatial correlation that decreases gradually over relatively long distances, making it suitable when spatial dependence diminishes smoothly. In contrast, the spherical model shows strong correlation at short distances, followed by a rapid decrease beyond a certain range.
In total, $4 \times 2 = 8$ OKFD models were fitted, and model performance was compared using MSE values.

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: MSE according to the number of basis functions and covariance models

As the number of basis functions increases, computational cost also increases; however, predictive performance improves. Consequently, the model with 200 basis functions and a spherical trace-semivariogram demonstrated the best predictive performance.

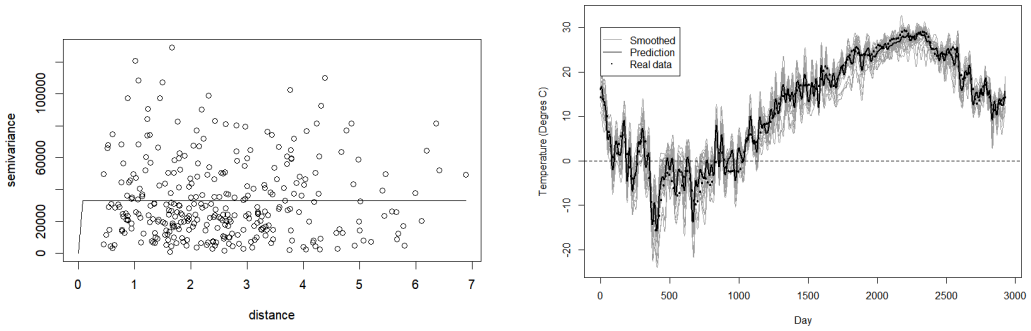

Figure 5: Semivariogram of the spherical covariance model (Left) and comparison between predicted and observed temperatures (Right)

From the semivariogram in Figure 5, it can be observed that independence between locations emerges rapidly as distance increases. This is likely due to the relatively small geographic size of North Korea. The comparison between predicted and observed temperature curves on the right indicates that the predictions closely follow the observed values.

3.4.2 Cross-validation analysis

The `okfd.cv` function performs cross-validation for Ordinary Kriging for Functional Data (OKFD) and is used to select optimal parameters. Accordingly, cross-validation was conducted by varying the number of basis functions. The values of `nbasis` were set to 5, 105, 205, 305, and 405.

```
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)
```

The results are summarized as follows:

- k.opt: 405, the optimal number of basis functions
- krig.CV: An array of kriging predictions computed across all cross-validation iterations
- MSE.CV: MSE values computed for each number of basis functions
- MSE.CV.opt: 21.24, the optimal MSE value

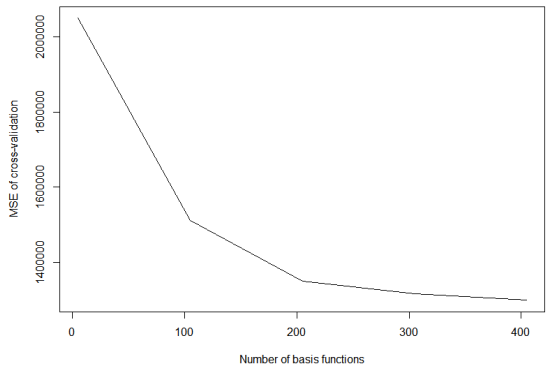

Figure 6: Decrease in MSE as the number of basis functions increases

Figure 6 shows that MSE decreases as the number of basis functions increases, indicating improved predictive accuracy. However, larger numbers of basis functions result in substantially higher computational cost.

Using OKFD, temperature prediction achieved strong performance, with an RMSE value close to 4.5. Although increasing the number of basis functions reduces MSE, it also increases computational burden, highlighting the importance of selecting an appropriate number of basis functions. The source code for the `okfd` and `okfd.cv` functions is provided separately.

4 Prediction of PM10

PM10 concentrations are strongly influenced by wind speed. When wind speed is high, PM10 concentrations tend to decrease, whereas low wind speed leads to higher concentrations. Therefore, PM10 concentrations in North Korea are predicted by incorporating wind speed. Meteorological data from South Korea (PM10 and wind speed) are used to

fit the model, which is then applied to North Korea. Ultimately, the objective is to predict PM10 concentrations using wind speed data collected at different time points over a single day.

4.1 Data Description

- **PM10 (scalar data)**: Represents PM10 concentrations measured at each location in South Korea.
- **Wind (functional data)**: Wind speed data measured at each location in South Korea.
- **Coord**: Spatial location information for each site (South Korea and North Korea).

4.2 Cokriging

Cokriging is a generalization of kriging developed to handle multivariate spatial prediction. Cokriging exploits the covariance between two or more correlated random fields to provide more accurate predictions of the primary variable. This method has been shown to outperform kriging in terms of both cost efficiency and predictive accuracy. In addition, cokriging can handle not only multivariate random field data but also functional random field data, which are considered in this analysis.
In this study, a cokriging approach is employed to predict the scalar primary variable PM10 by using wind speed, a functional random field, as a secondary variable. Based on this framework, the relationship between PM10 and wind curves is analyzed using the CLM. The proposed cokriging predictor is an extension of the classical cokriging predictor, where functional data are represented using basis functions and incorporated into the prediction framework.

4.2.1 CLM (Co-Regionalized Linear Model)

In cokriging, the Co-Regionalized Linear Model (CLM) is used to estimate auto-correlation and cross-correlation functions. CLM models spatial dependence between variables and improves prediction accuracy by capturing interactions among different variables. It is particularly useful for multivariate and functional data, and in this analysis, it is used to model the relationship between PM10 and wind speed.

4.2.2 Cokriging Predictor Using Functional Secondary Variables

When the number of secondary variables is assumed to be very large, these variables can be replaced by functional variables. This allows the definition of a cokriging predictor that uses functional secondary variables within a random field framework.
For a functional linear model with a scalar response, it is assumed that $m \to \infty$ in Equation

(1). In this case, the parameters β_{ij} defined in Equation (1) can be replaced by functions $\beta_i(t)$, and the secondary variables $X_2(s_i), \ldots, X_m(s_i)$ can be replaced by a functional variable $X_{s_i}(t)$.
Therefore, the cokriging predictor for $X_1(s)$ at an unsampled location s_0 can be expressed as follows:

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

where $X_1(s_0)$ denotes the predicted value of the primary variable at the unsampled site s_0, λ_i represents the contribution of the scalar variable $X_1(s_i)$ to the prediction, $X_{s_i}(t)$ is the functional variable observed at site s_i, and $\beta_i(t)$ is the functional parameter that represents the weight of $X_{s_i}(t)$ in the prediction $(i = 1, \ldots, n)$.

To estimate the parameters, an approach based on B-spline and Fourier-type basis functions is adopted. The functional variables and parameters are expanded as follows:

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

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

where $\boldsymbol{\phi}(t) = \begin{bmatrix} \phi_1(t) & \ldots & \phi_k(t) \end{bmatrix}^T$ denotes the vector of basis functions, and $\mathbf{a}_i$ and $\boldsymbol{\beta}_i$ are vectors of coefficients estimated via least squares.

By substituting the basis function expansions into Equation (2), the predictor can be rewritten as:

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

where β_{ij} is defined in Equation (3), and $\mathbf{a}_i$ is given by:

$$\mathbf{a}_i = \begin{bmatrix} a_{i1}, \ldots, a_{ij}, \ldots, a_{ik} \end{bmatrix}^T. \tag{5}$$

4.2.3 Cokriging Prediction of PM10 Using wind Curves

Various basis functions can be used to smooth time-series data. In this study, Fourier basis functions were employed because they effectively capture periodic patterns. To smooth the data, a Fourier basis with $k = 7$ was used, as shown in Figure 7. The figure indicates that certain regions exhibit relatively high wind speeds.
As a result, each observation site has a PM10 value and a corresponding set of seven coefficients obtained by fitting the wind data using the Fourier basis.
Based on these coefficients, a CLM (Cross-Linear Model) was fitted. In this process, a Gaussian model was used to describe all simple variograms and cross-variograms.

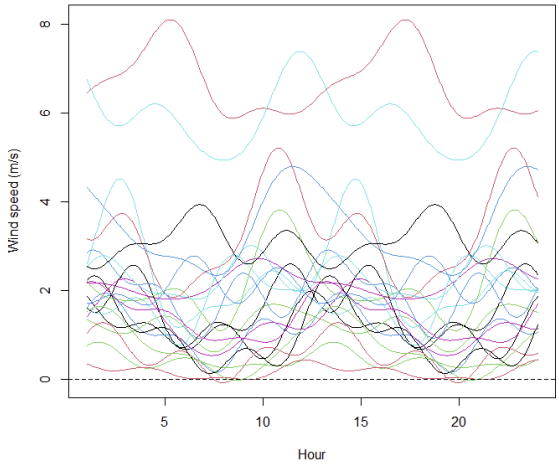

Figure 7: Wind data transformed into functional data

```
# Linear model of coregionalization (spatial correlation
   structure modeling)
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) # calculate variogram
plot(v, xlab="Distance") # visualize variogram
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.lmc fits the linear model of coregionalization (LMC)
plot(v, g.fit, xlab="Distance", ylab="Simple and cross
   semivariance", main ="")
names(g.fit)
```

```
g.fit$model
```

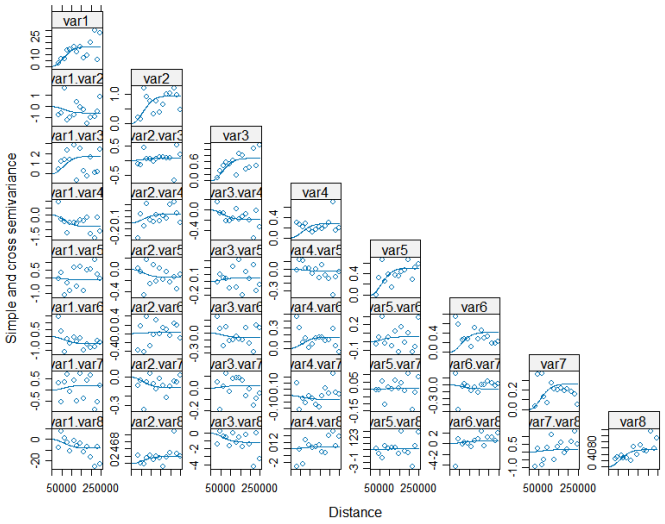

Figure 8: Visualization of the semivariogram representing spatial correlation

In Figure 8, the x-axis represents the distance between data points, and the y-axis represents the semivariance for each variable or combination of variables. In general, semivariance tends to increase as distance increases. Cross semivariance measures the correlation between two variables. This visualization allows assessment of how semivariance changes with distance and how well the fitted model represents the observed data.

The following code uses the fitted `g.fit` model to predict PM10 concentrations over 10,000 locations covering 27 regions in North Korea. Cokriging prediction and prediction variance were computed at each grid point.

```
# Analyze North Korean location data (coord)
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"]
```

The summary of predicted PM10 concentrations and prediction variances across 10,000 locations in North Korea is as follows:

```
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 9 visualizes the predicted PM10 concentrations and prediction variances over North Korea. Since predictions were performed on a grid of 10,000 points covering the entire latitude and longitude range of North Korea, it is difficult to identify specific locations; however, the overall spatial distribution of PM10 concentrations can be observed.

Although some regions show relatively low PM10 concentrations, overall PM10 levels in North Korea appear to be relatively high.

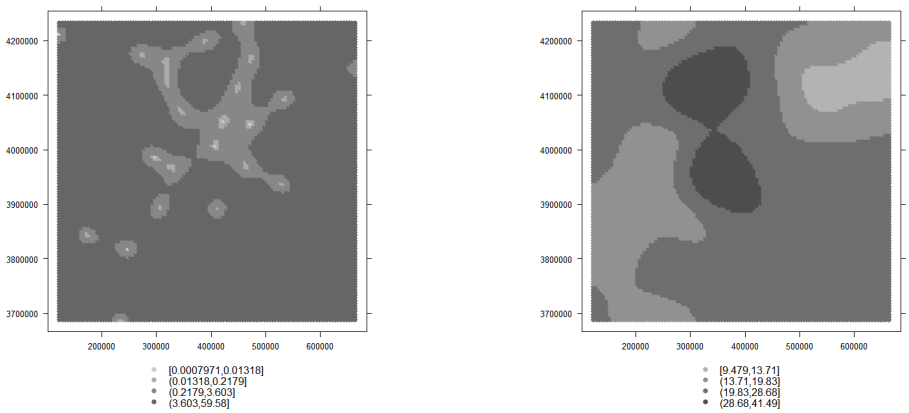

Figure 9: Predicted PM10 concentrations (Left) and prediction variance (Right)

5 Conclusion

In this study, temperature and PM10 concentrations in North Korea were predicted using OKFD and Cokriging. Due to the relatively small geographic size of North Korea, spatial data analysis did not yield substantially stronger results compared to other analyses; however, this study demonstrates that meteorological data can be predicted using relatively accessible geographic information. In the case of OKFD, the low MSE values indicate that the analysis is practically meaningful. For PM10 prediction, although RMSE could not be computed due to the lack of observed PM10 data in North Korea, the analysis is meaningful in that PM10 concentrations were predicted using location and wind speed data.

In addition to OKFD, various spatial functional data analysis methods such as PWFK (Pointwise Functional Kriging) and FKTM (Functional Kriging Total Model) exist. In future work, these models could be fitted and compared to evaluate which method yields the best predictive performance.

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.