Comparison of Spatial Interpolation Methods for Air Pollution Prediction: gstat and mgcv

by Hyesop Shin

Abstract Understanding the association between pollution exposure and the deleterious effect on population health is a vital precursor for modelling. However, there has been less comparisons of spatial interpolations for pollution predictions due to different assumptions and mathematical processes. In addition, SI has been provided in a coarse temporal scale which is difficult to understand the dynamics of exposure by mobility patterns and seasonal effects. This paper aims to compare spatial interpolation methods to predict air pollution at a finer temporal scale. 57 pollution stations around Seoul, S.Korea were collected for the comparison between Universal Kriging and Generalised Additive Model, with additional weights on road layers as an effect of roadside pollution fields. Neither of the interpolation methods was noticeably superior to the other, but the sparse station data meant that only very smoothed large-scale fields could be recovered, which did not accurately represent the extremes observed at individual stations..

Introduction

Population health has been seriously threatened by daily ambient air pollution in South Korea. Despite the efforts to legislate national pollution standards, daily peaks in recent winters and springs have already exceeded the standards countless times. During 5th-8th March 2019, the entire country experienced over 200μg/m^3 of PM_{10} and smog episodes; the national authorities warned everyone to reduce outdoor activities. As most of Seoul's areas tend to experience disastrous levels of pollution frequently, residents can be exposed to air pollution unconsciously, which in the long-term can lead to respiratory or cardiovascular ailments (Zhang and Batterman, 2013). Thus, understanding the spatial and temporal aspects of air pollution and its relationship with exposure is crucial.

Exposure research has exploited spatial interpolation (SI) to investigate the relationship between ambient air pollution and population exposure in a spatial context – "which places have high air pollution?" and "how many people can potentially become unwell from high episodes?". SI is a statistical method that can compute pollution fields over a wide area with a given set of point measurements. SI can mainly be split into a group that follows the assumption of spatial autocorrelation (e.g. Inverse Distance Weighted (IDW), Kriging), and a group on statistical inference e.g. generalised linear models and generalised additive models (Wood, 2019). Methods that take spatial autocorrelation into account assume that the values tend to be more similar when closer together, whereas methods that use (spatial) statistical inference delineate the inferential surfaces over a region by minimising the residuals of the model.

However, when air pollution monitoring stations such as those in Seoul are small in number compared to the size of the city, the estimation of the potential population at risk due to air pollution can be completely different depending on the measurement method used (Wu et al., 2019). Wong et al. (2004) used four spatial interpolation methods – spatial averaging, nearest neighbour, IDW, and Kriging - to estimate children's exposure to air pollution across the USA. The outcomes of the four methods only showed a small difference of PM_{10} and O_3 where the monitoring stations were denser, for example in the northeastern cities and urban California, but was difficult to predict in the mountainous regions and high-altitude zones. Aalto et al. (2013) compared a Generalised Additive Model (GAM) and an Ordinary Kriging (OK) to predict monthly mean temperature and precipitation and found that the GAM outperformed other methods by a small amount but the biased distribution of stations ("concentrated in the urban areas" of Finland) might have evened out the RMSE measures.

In addition, previous studies have provided tentative estimates of population risk based on annual or monthly statistics, but the aggregated figure lacks the potential for including acute injuries after an abrupt pollution rise, which may be more severe. However, there is likely to be a greater difference when the population at risk is measured at a finer temporal scale. This might support guidelines for surveillance of short-term exposure.

This chapter aims to compare spatial interpolation methods that can support estimating population exposure to daily air pollution in Seoul. This study compared Universal kriging (UK), Generalised additive model (GAM), UK with additional road effect, and GAM with additional road effect to model PM_{10}, and NO_2 in Seoul as an intermediate phase of pollution modelling. Compared to previous studies, this chapter generates outcomes on a 12-hour basis to understand the daily cycle of the

pollution over the city and superimposes road effects to take into account small scale variables that might be neglected in the typical spatial interpolation outcomes.

Data

This article explores the spatial and temporal patterns of pollution in Seoul as well as the extent to which background and roadside stations are similar or different. The NIER (National Institute of Environmental Research) provided 6 pollutants, PM_{10}, $PM_{2.5}$, NO_2, CO, SO_2, O_3, which were released as an hourly aggregation. This study selected PM_{10} as the main source that can causally result in human cancer and NO_2 as the main source that harms respiratory symptoms, whether constantly or instantaneously (Europe, 2013).

```
options(scipen = 100)
library(tidyverse)
library(sf)
library(mgcv)
library(raster)

load("../Data/no2.RData")
stations <- read_sf("../Data/stations_10km.shp")
stations_df <- stations %>% filter (F.R == "Fixed")%>% st_set_geometry(NULL)
seoul <- read_sf("../Data/Seoul_City.shp") %>% as_Spatial() %>% fortify()
no2.winter <- merge(no2.win.bk, stations_df, by.x = c("Station.ID", "X", "Y"), by.y = c("Station", "X", "Y"))
```

Pollution data were collected from two different types of stations: background stations (installed on the rooftops of district offices), and roadside stations (close to the major roads which are strongly affected by nearby traffic). This study took into account 57 urban background stations and 19 roadside stations that were within 10km from the city boundary (see Figure 1). Amongst these 76 stations, Seoul itself had 25 background stations and 15 roadside stations. Roadside data were retrieved from the 19 stations installed on the roads near the city centre, 8-lane junctions, and a highway entrance. The possible download period was between 01/01/2010 01:00 and 01/01/2018 00:00. Units are measured in ppb (parts per billion) for NO_2, and $\mu g/m^3$ for PM_{10}.

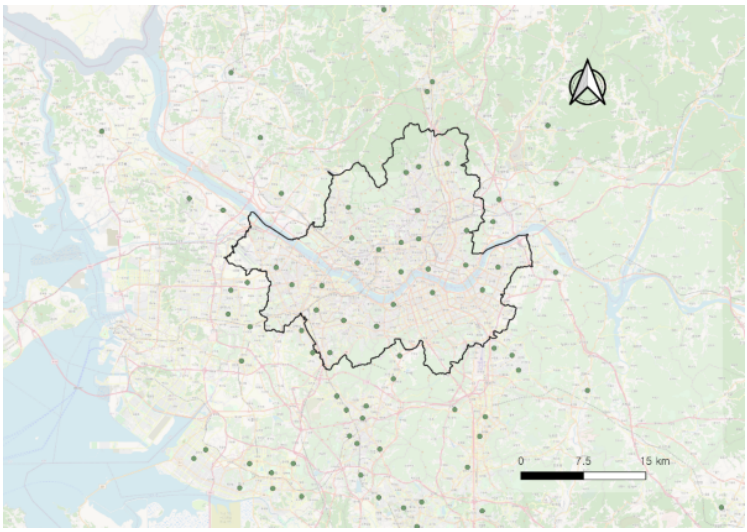

Figure 1: 57 Background pollution stations that are considered for spatial interpolation (boundary area: Seoul)

Kriging: Conceptual Framework

Kriging is a geostatistical method that interpolates an unknown location using the characterized mean and variance structures (**??**). It assumes that the overall mean, distance, and variance of all observations are spatially autocorrelated from Tobler's First Law of Geography: "Everything is related to everything else, but nearer things are more related than distant things". For example, pollution

concentrations less than 1 metre apart are likely to be similar, but less likely when the distance becomes larger. This context of the variability between the two values and their grouped distances (e.g. 0-100m, 101-200m) can be structurally applied over the region with a conceptual semivariogram. Once the semivariogram is produced, then the kriged map is produced together with an error map.

To create a kriged map, there are a few steps to follow: 1) modelling an empirical semivariogram, 2) fitting the empirical model with a mathematical model, and 3) choosing the type of kriging according to the fit of the assumptions (i.e. data normality, stationarity, and whether the data has trends).

Understanding Semivariograms Spatial autocorrelation (or spatial dependency) is assessed by a semivariogram. Semivariograms apply the squared differences of measurements against distances between pairs of data points (see Equation~1). This means that all the pollution values at the 57 stations are compared between one another, and once calculated, the semivariogram is drawn. The conceptual semivariance can be estimated as below.

$$\gamma(h) = \frac{1}{2N}(h) \sum_{i=1}^{N(h)} Z(s_i) - Z(s_i + h)^2 \tag{1}$$

- N(h)= number of pairs of observation points with distance h
- Z = field that holds a height or magnitude value for each point (z-values)
- $Z(s_i)$and Z(Si+h) = sample data pairs at a distance h (**?**)

Directions are not considered in the current formula; however, this can be taken into account depending on the aspect of the study, for example, if the wind direction is dominantly affecting the pollution concentrations.

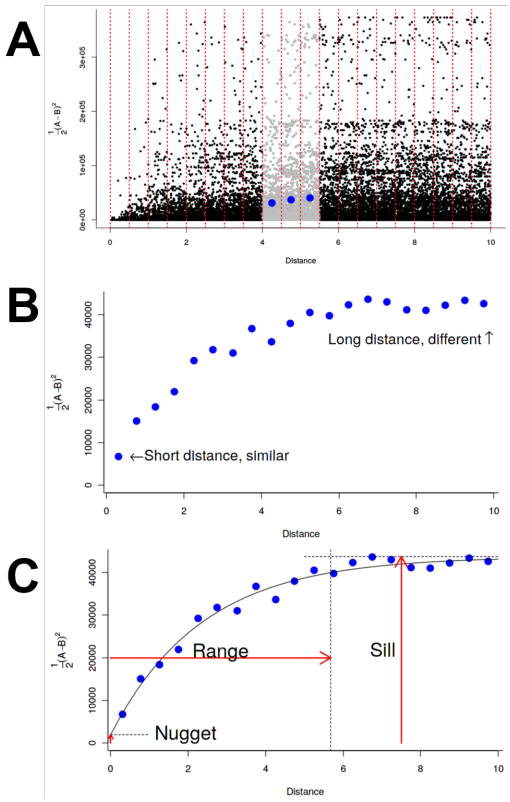

Figure 2: Conceptual process of kriging measurements (captured from Datacamp.com lectures)

Figure 2 explains the process to generate a semivariogram. First, the distance against half-squared distances between all pair points is plotted in the variogram cloud (see Figure 2a). The dots in the variogram cloud are allocated to an arbitrary number of imaginary bins (10 bins are used for this study). The mean of each bin becomes a representative point which normally has an upward trend as the distance increases to some extent and then levels off (see Figure 2b). This means that the closer the distance, the more similar the values are, and vice versa. Finally, an asymptotic curve, termed

semivariogram, is drawn through the points (see Figure~2c). Range[1], partial sill[2], and nugget[3] are the key parameters inside the model, and the kriged map is drawn based on the semivariogram (Law and Collins, 2019).

```
myVario <- list()
myList <- list()

for(i in 1:20){
 myVario[[length(myVario)+1]] <- variogram(top.jan[[i]] ~ 1, top.jan, cutoff = 30000, width = 3000)
  myList[[i]]  <- fit.variogram(myVario[[i]],
                                vgm(psill = 100,
                                    nugget= 15,
                                    model = "Ste"),
                                fit.kappa = TRUE, fit.method = 6)
}
```

- Here we set the width and the cutoff points

Bibliography

J. Aalto, P. Pirinen, J. Heikkinen, and A. Venäläinen. Spatial interpolation of monthly climate data for Finland: comparing the performance of kriging and generalized additive models. *Theoretical and Applied Climatology*, 112(1-2):99–111, 2013. ISSN 0177-798X. [p1]

W. Europe. Proximity to roads, NO2, other air pollutants and their mixtures. In *Review of evidence on health aspects of air pollution–REVIHAAP Project: Technical Report [Internet]*. WHO Regional Office for Europe, 2013. [p2]

M. Law and A. Collins. *Getting to know ArcGIS PRO*. Esri press, 2019. ISBN 1589485378. [p4]

D. W. Wong, L. Yuan, and S. A. Perlin. Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Science & Environmental Epidemiology*, 14(5):404–415, 2004. ISSN 1559-064X. doi: 10.1038/sj.jea.7500338. [p1]

S. Wood. Simplified integrated nested Laplace approximation. *Biometrika*, 2019. [p1]

C.-Y. Wu, J. Mossa, L. Mao, and M. Almulla. Comparison of different spatial interpolation methods for historical hydrographic data of the lowermost Mississippi River. *Annals of GIS*, 25(2):133–151, apr 2019. ISSN 1947-5683. doi: 10.1080/19475683.2019.1588781. URL https://doi.org/10.1080/19475683.2019.1588781. [p1]

K. Zhang and S. Batterman. Air pollution and health risks due to vehicle traffic. *Science of the Total Environment*, 450-451:307–316, 2013. ISSN 00489697. doi: 10.1016/j.scitotenv.2013.01.074. [p1]

Hyesop Shin
MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Berkeley Square, 99 Berkeley Street, Glasgow, G3 7HR
https://www.gla.ac.uk/researchinstitutes/healthwellbeing/staff/hyesopshin/
hyesop.shin@glasgow.ac.uk

[1]The distance at which a spatial correlation exists.
[2]The upper limit of the semivariogram is the sill. A Sill minus a nugget is termed a partial-sill
[3]Represents short scale randomness or noise in the regionalised variable [@Camana2020]