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.

Geostatistics and Statistical Inference Models

Recent literature of spatial statistics has shown advanced models such as a bayesian fused spatio-temporal kriging or mathematically sound equations to fit the domain contents. However, this article compares the typical Kriging and GAM that appears on a two dimensional space.

Similarities

The distribution of the data should follow the bell curve The robustness of data is only meaningful when the basic assumption that the data is to follow the normal distribution is met. Hence, it is always essential to explore the data.

Mean and the variation are important Both Universal Kriging and GAM work with the mean value. In universal kriging, it is important to find the average across the space

Differences

- Distance vs shape of the Spline
- Fitting the Semivariogram versus Adjusting the Knots and Wiggliness
- Computing Time

Empirical Experiment using `gstat` and `mgcv`

- Using 57 stations surrounding Seoul, this article runs two SI models
- We use NO_2 which chemically reacts well with O_3 thus is sensitive day and night, and PM_{10} which is less sensitive to hourly differences but more sensitive to seasonal patterns
-

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]

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