Prediction of Air Pollution PM10 in the UK in 2021

Lenka Sefcakova

12 February 2022

Abstract

The air pollution levels in the UK in 2021 were estimated using three spatial interpolation techniques: K-nearest neighbors, inverse distance weighting, and Voronoi models.

The data was collected from 85 sites across the UK and the mean pollution level was obtained for each site. A grid of points was fitted over the map of the UK and the methods were employed to predict the pollution levels across the area.

The results were then compared and the best performing method was selected to create an ensemble model. In addition, the INLA model was used to obtain the statistical significance of the modeled predictions by providing upper and lower bounds.

This study highlights the importance of understanding air pollution and the methods that can be used to estimate the pollution levels in areas where continuous monitoring is not feasible.

1 Introduction

Continuous variables are hard to monitor in real world and we are forced to infer values for specific areas based on few observations. We will focus on techniques to interpolate and find air pollution levels across the UK in terms of one of the most used metrics PM10.

PM10 refers specifically to particles with a diameter of 10 micrometers or less. These particles can come from a variety of sources, including dust, soil, and emissions from industrial processes, transportation, and biomass burning.

We compare K-nearest neighbors, inverse distance weighting and Voronoi prediction performance over the UK area using only the pollution covariate, including possible improvements on the methods and reasoning why they are or are not suitable for given problem. To better understand the statistical significance we also implement the INLA model obtaining upper and lower bounds for modeled predictions.

2 Data set preparation and description

Data is collected from 85 sites across the UK for year 2021 and values are averaged to obtain the mean pollution level for a given site [3].

We use the openair package to collect data from UK Automatic Urban and Rural Network, Scottish, Welsh and Northern Ireland Air Quality Network, King's College London networks and locally managed sources in England.

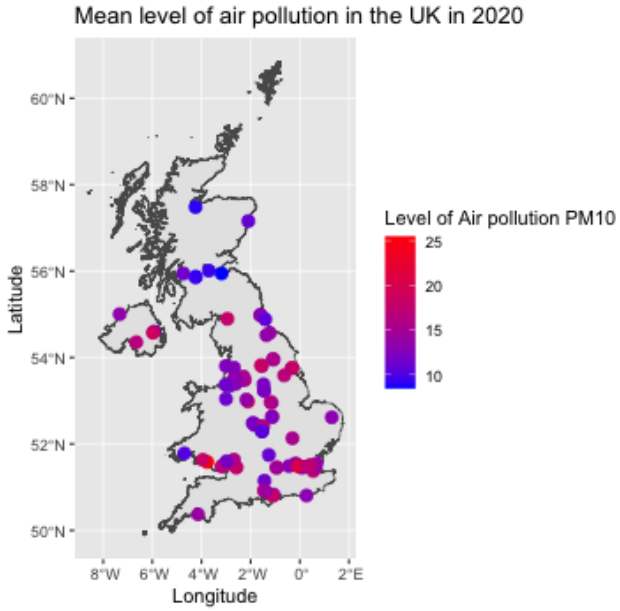


Figure 1: Pollution measurement at sites in 2021

Maximum pollution measured at 25.45 PM10 in Port Talbot Margam industrial zone, Wales and minimum value measured in Edinburgh St. Leonards, Scotland. The pollution distribution can be seen in Figure 2.

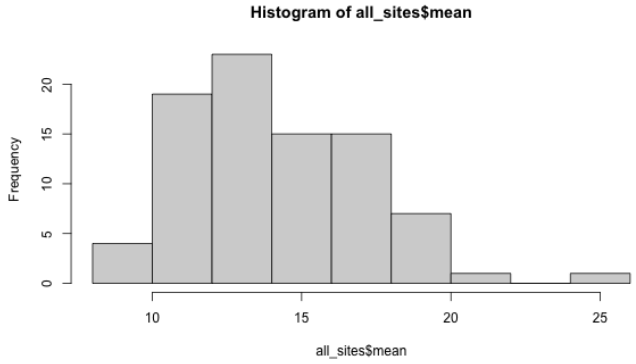


Figure 2: Pollution measurement distribution across sites in 2021

3 Methods

3.1 Spatial Interpolation

We will use spatial interpolation methods to predict the pollution levels across the UK. Namely K nearest neighbors, Inverse Distance Weighting and Voronoi models were used to create an ensemble model which has showed the best performance as indicated in Table 1. The methods are employed over a grid of points fitted over the map of the UK [2].

K-nearest neighbors method

Value for each grid point is calculated as the mean value of its K nearest neighbours with uniform weighting. Namely in our case K=25. The reason for this is that higher number of neighbors will result in smoother transitions between pollution levels across the map which is fitting for our problem. Note that this may introduce constant values over large areas when the area has only few observations.

Note that 20 Fold CV is used to evaluate model performance.

Voronoi method

We create a Voronoi diagram over the set of our observations by dividing the space into polygons such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than to any other, additionally the polygons are convex. The whole polygon then assumes the value of the interior point.

Inverse Distance Weighting method

This method is similar to KNN however all points are considered to evaluate each grid point pollution level. The main feature of this method is that when computing the grid point value the weighted average of all other points is considered. The weights are inversely proportional to the distance of each of the points to the point that is being estimated.

$$w_i = \frac{1/d_i^\beta}{\sum_i^n (1/d_i^\beta)}$$

where our choice of beta is $\beta = 1$. Multiple values were assessed in terms of RMSE.

Ensemble method

In order to combine all models mentioned above we calculate each grid value as the weighted average of the respective model predictions, weighted by their inverse RMSE test score ratios.

3.2 INLA

We assume that the pollution levels $\mathbf{Y}_i$ at location $\mathbf{s}_i$ follow a Gaussian distribution with mean μ_i and variance σ, where μ_i is a sum of the intercept and a spatial random effect, with no other covariates included [1].

$$Y_i \sim N(\mu_i, \sigma^2),\ i = 1, 2, \ldots, n$$

$$\mu_i = \beta_0 + Z(\boldsymbol{s_i}).$$

To be able to predict a continuous variable over area we first create a triangulated mesh of the region and a projection matrix A. The prediction is then computed over a grid of values over the UK.

Since we are computing the posteriors we will be able to obtain an actual 95% confidence intervals for predictions.

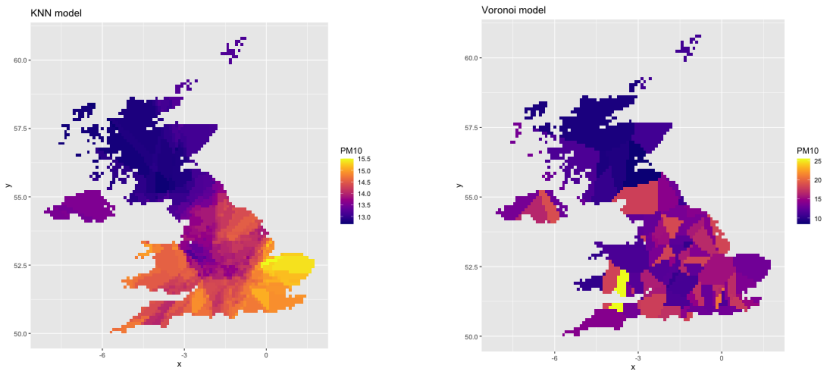
4 Results and interpretation

4.1 Spatial Interpolation

The results show that models conclude substantially different pollution levels across areas.

In Figure 3 we see London to be predicted with low air quality where as the Voronoi model predicts this area within the mean values of the overall spectrum. We can also see that Scotland has good air quality according to both models.

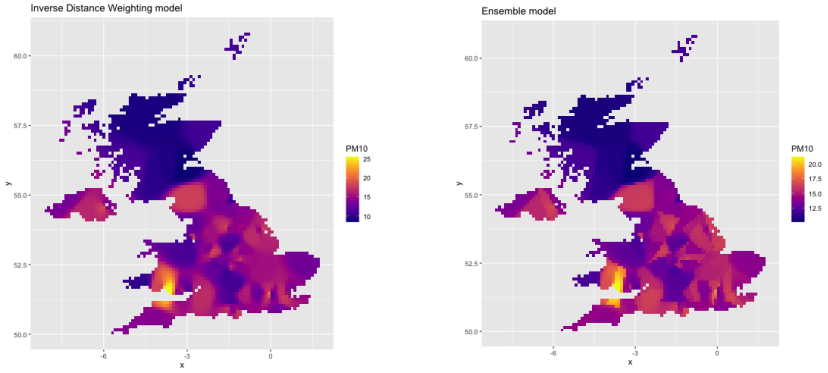
In both cases the predicted values for ares form visible tiles i.e. the predictions are not viewed as continuous, in Voronoi model predictions this being a more prominent feature as expected from the model architecture mentioned in previous section.



(a) KNN (b) Voronoi

Figure 3: KNN (left) and Voronoi (right) predicted values for UK region

The inverse distance weighting model predictions resemble the Voronoi model with a more continuous nature. Since the IDW model was the best performing out of the three models used in the ensemble we can see that the Ensemble model predictions carry a lot of its structure as seen in Figure 4.



(a) IDW (b) Ensemble

Figure 4: Inverse Distance Weighting (left) and Ensemble (right) predicted values for UK region

Finally the results of cross validation can be seen in Table 1. As mentioned before the ensemble method has showed to be the most accurate of the four options with the lowest RMSE value 2.59. The Voronoi method has the highest value RMSE = 3.52, presumably caused by the lack of flexibility of the model to reflect the true continuous value of our target variable.

	voronoi	near.neigh	IDW	ensemble
1	3.52	2.73	2.71	2.59

Table 1: RMSE values obtained by 20 fold CV for respective methods used

With RMSE 2.59 being relatively high no concrete conclusions should be drawn from the data, however all models agree on Scottish air quality to be the best from the overall areas. Areas around Bristol Channel (Cardiff, Bath etc.) with industrial zones have poor air quality with high number of particles in the air.

4.2 INLA

The Fixed Effects estimated by the model can be found in Table in 2, with standard deviation 2.32 and mean 13.73. The optimal model has Marginal log-Likelihood of -230.19.

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
b0	13.731	2.328	7.942	13.82	19.14	13.898	0.002

Table 2: INLA Prediction summary Fixed Effects

In Figure 5 we see upper and lower bound as well as the mean (expected) value of the model within 95% confidence interval. Since the model is more conservative than the spatial interpolation methods the results are less refined (the values in the plots have small variance within one plot). The plots still suggest aforementioned polluted areas as well as higher air quality in Scotland, with the exception of the upper bound which provides a rather pesimistic prediction.

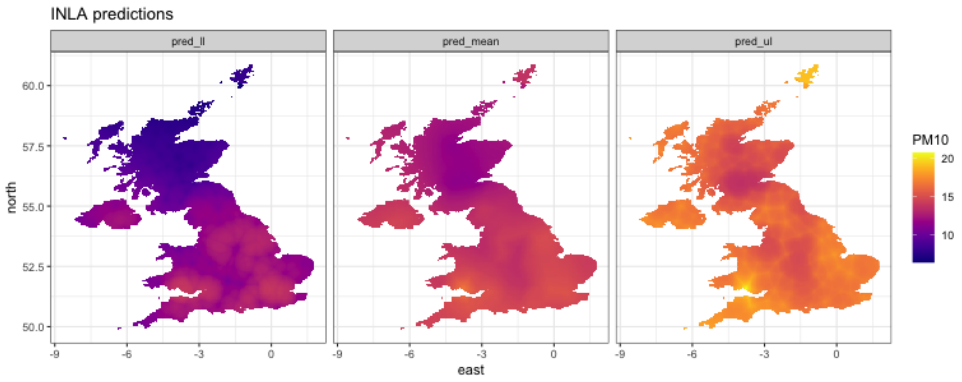


Figure 5: INLA SPDE model 95% confidence interval prediction, lower bound (left), upper bound (right) and mean in the middle.

5 Discussion

The results obtained in this study show the potential of using spatial interpolation methods for predicting air pollution levels in terms of PM10 in the UK. The models used, K-nearest neighbors, inverse distance weighting and Voronoi, were able to provide estimates for the mean PM10 levels across the entire UK. The ensemble model that was created by combining the best models from each method performed even better, providing a more accurate estimate of the air pollution levels.

One of the strengths of this study is the use of the INLA model to obtain the statistical significance of the results. The upper and lower bounds that were obtained from the model can be used to determine the level of uncertainty associated with the predictions. This information can be useful for policymakers and decision-makers when it comes to addressing air pollution and mitigating its impact.

It is important to note that there were limitations to this study. The data used was collected from 85 sites across the UK, so it is possible that the results may not be generalisable to other regions or countries. Additionally, no additional covariates were added to the model other than the observed pollution levels. Adding covariates such as area type (urban/industrial/rural), temperature, wind etc. might improve the predictions significantly.

In conclusion, this study provides a useful starting point for exploring the potential of spatial interpolation methods for predicting air pollution levels. The results show that the methods used in this study are capable of providing accurate estimates for PM10 levels across the UK, but more research is needed to validate the results and explore alternative methods. The information provided by this study can be used to support decision-making related to air pollution and to help mitigate its impact on human health and the environment.

References

[1] Paula Moraga. *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Chapman & Hall/CRC Biostatistics Series, 2019.

[2] Paula Moraga. *GeoStatistical Data Interpolation*. URL: `https://www.paulamoraga.com/book-gds/41-geostatisticaldata-interpolation.html`.

[3] *UK Air: Air Quality in the UK*. 2023. URL: `https://uk-air.defra.gov.uk/`.