Bias-Correcting Daily Satellite-Retrieved AOD For Air Quality Research

Wyatt G. Madden†, Yang Liu‡, Howard H. Chang†

†Department of Biostatistics & Bioinformatics, Emory University,
‡Department of Environmental Health, Emory University

EMORY

Introduction

Fine particulate matter ($PM_{2.5}$) is a major air pollutant that is associated with adverse health outcomes. Typically $PM_{2.5}$ is measured at ground level monitoring stations, however these stations are globally sparse and provide poor coverage for many areas that experience high air pollution levels. The National Aeronautics and Space Administration (NASA) selected the Multi Angle Imager for Aerosols (MAIA) proposal in 2016 to improve this coverage and better understand how respirable particulate matter ($PM_{2.5}$) affects human health.[1] The MAIA project will deploy satellites to collect global 1-km spatial resolution Aerosol Optical Depth (AOD) data. AOD strongly correlates with $PM_{2.5}$, and thus can be used to better understand adverse health outcomes by employing bias-correcting methods.

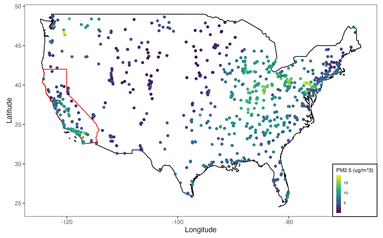

Fig. 1: PM2.5 observations on 2018-10-08 at all AQI monitors in the contiguous United States.

While bias-correcting satellite or numerical model simulations with ground-level monitoring data is a common task in air quality modeling, Bayesian uncertainty quantification is not widely adopted in part due to lack of user-friendly implementations. Here we present a Bayesian geostatistical regression model framework and corresponding R package (`grmbayes`) to efficiently estimate the relationship between satellite/simulation air quality data and ground monitor collected $PM_{2.5}$, and predict $PM_{2.5}$ at locations for which only satellite/simulation data is available. This software will provide scientists and practitioners with a user-friendly tool to employ Bayesian spatio-temporal methods when incorporating MAIA data in air quality research.

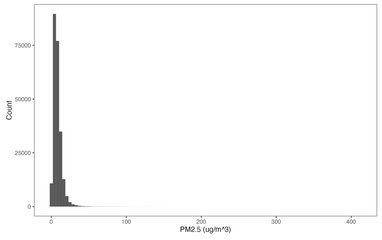

Fig. 2: Histogram of PM2.5 observations across all AQI monitors and days in 2018.

Data

Our data consists of daily $PM_{2.5}$ observations from 2018 at 973 ground-level air quality monitors in the contiguous United States.(Figure 1) These data were chosen due to the large number of major wildfire events, most notably the Camp Fire in California, which corresponded with high $PM_{2.5}$ variability.(Table 1, Figure 2) While our motivation and exposition assumes regressing on MAIA AOD data, we utilize 12km gridded Community Multiscale Air Quality (CMAQ) Chemical Transport Model (CTM) simulation data for our examples throughout this work without loss of generality.[2] MAIA AOD data are matched to each monitor by date and location (nearest neighbor) for model fitting while full data sets can be used for $PM_{2.5}$ prediction.

Table 1: Monthly summary statistics of $PM_{2.5}$ observations across all monitors and days.

Month	Mean	Median	Min	Max	Sd
Jan	9.3	7.8	0.1	318.8	7.9
Feb	8.2	6.9	0.1	71.5	5.6
Mar	6.6	5.8	0.1	74.0	4.2
Apr	6.9	6.3	0.1	55.1	3.9
May	7.6	7.1	0.1	156.0	3.9
Jun	7.9	7.3	0.1	167.8	4.5
Jul	9.4	8.2	0.1	146.0	6.3
Aug	12.4	10.1	0.1	261.0	11.0
Sep	6.9	6.1	0.1	76.2	4.1
Oct	6.3	5.6	0.1	43.0	3.8
Nov	9.7	6.7	0.1	411.7	14.2
Dec	8.8	7.3	0.1	70.6	6.3

Model

The geostatistical regression model is given by

$$PM_{2.5}(s,t) = \alpha_0(s,t) + \alpha_1(s,t)X(s,t) + \epsilon(s,t)$$

where $PM_{2.5}(s,t)$ and $X(s,t)$ are the fine particulate matter concentration and the aerosol optical depth respectively, at location s and time t.[3] The $\alpha_0(s,t)$ and $\alpha_1(s,t)$ parameters are the intercept and slope of the regression model composed of the following spatial and temporal effects:

$$\alpha_0(s,t) = \beta_0(s) + \beta_0(t) + \gamma_0 Z_0$$
$$\alpha_1(s,t) = \beta_1(s) + \beta_1(t) + \gamma_1 Z_1$$

where $\beta_i(s) \sim NNGP(0, \tau_i^2 K_i)$, $\beta_i(t)$ are modeled as first-order random walks, and γ_i are fixed effects for spatial or spatio-temporal varying covariates Z_i. The K_i kernal of the Nearest Neighbor Gaussian Process (NNGP)[4] is assumed Matérn(ν_i, θ_i), with $\nu_i \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$, with a pre-specified number of neighbors. Weakly-informative priors are placed on all parameters and MCMC is employed to sample from the posterior distribution of the model parameters. For computational efficiency, both NNGP and regular GP are assessed in addition to discretization of the spatial process range parameter θ_i.

R Package: grmbayes

We provide an intuitive interface for fitting the geostatistical regression model with the following features:

- **Spatial Process**: Select either GP or NNGP (with m number of neighbors)
- **Random Effects**: Select either additive or multiplicative random effects for spatial and/or temporal components.
- ***θ* Discretization**: Discretize the spatial process range parameter θ_i for spatial intercept and/or spatial slope. Choose levels, and either Gibbs or Metropolis-Hastings updating schemes.
- **Cross Validation**: Choose number of folds, and cross validation type (out of 'ordinary', 'spatial', 'spatial clustered' or 'spatial buffered' with a corresponding buffer size).
- **Covariance Kernal**: Select Matérn(θ, ν) covariance function with $\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$, or input user-defined covariance kernal.
- **Covariates**: Include additional regression covariates.

Performance

First we assess the appropriateness of the approximations designed to improve computational efficiency. Setting the number of neighbors to 10 and discretizing the θ parameters into 20 equally spaced levels across the range of feasible values determined from a non-discretized test run, we fit the models for all combinations of NNGP/GP and discretization schemes for 119 monitors within California. All models perform nearly identically (Table 2), suggesting that the approximations are appropriate.

Table 2: California in-sample RMSE for all combinations of spatial process and θ discretization schemes.

	θ Discretization		
Spatial Process	Gibbs	MH	None
GP	8.80	8.8	8.8
NNGP	8.81	8.8	8.8

Next we assess the performance of the model on the full contiguous United States data set for the three settings of Matérn covariance $\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$, utilizing five 10-fold cross-validation formulations: ordinary, spatial, spatial clustered, and spatial buffered (with buffer size of 35km and 100km).(Table 3) These results suggest that $\nu = \frac{1}{2}$ (exponential covariance kernal) is most appropriate for these data, and that there is considerable information gained from the spatial components of the model.

Table 3: Full U.S. RMSE (95% Prediction Interval Coverage Probability) for all combinations of spatial process and θ discretization schemes.

	Matérn ν Parameter		
Cross Validation Type	0.5	1.5	2.5
Ordinary	5.18 (0.97)	5.18 (0.97)	5.18 (0.97)
Spatial	5.61 (0.98)	5.88 (0.98)	5.9 (0.98)
Spatial Buffered (35km)	5.79 (0.98)	6.03 (0.98)	6.02 (0.98)
Spatial Buffered (100km)	6.12 (0.98)	6.16 (0.98)	6.16 (0.98)
Spatial Clustered	6.35 (0.98)	6.43 (0.98)	6.43 (0.98)

Prediction

Predictions are made by kriging the nearest neighbor spatial effects from the $PM_{2.5}$ monitor data set for each AOD location, and combining this with temporal effects. (Figures 4, 3)

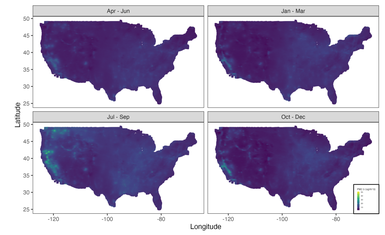

Fig. 3: PM2.5 mean posterior predictions by season at all prediction locations in the contiguous United States.

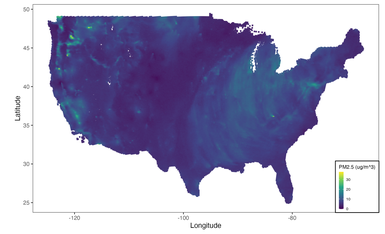

Fig. 4: PM2.5 predictions on 2018-10-08 at all prediction locations in the contiguous United States.

Acknowledgements

This project is partially supported by the NASA/JPL EVI-3 Multi-Angle Imager for Aerosols (MAIA) Proposal, and the motivating dataset is provided by Dr. Vaidyanathan at the US CDC Climate and Health Program.

References

[1] Yang Liu and David J. Diner. "Multi-Angle Imager for Aerosols: A Satellite Investigation to Benefit Public Health". In: *Public Health Reports* 132.1 (2017), pp. 14–17.

[2] Daewon Byun and Kenneth L. Schere. "Review of the Governing Equations, Computational Algorithms, and Other Components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System". In: *Applied Mechanics Reviews* 59.2 (Mar. 2006), pp. 51–77.

[3] Howard Chang, Xuefei Hu, and Yang Liu. "Calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentrations via statistical downscaling". In: *Journal of exposure science & environmental epidemiology* 24 (Dec. 2013).

[4] Abhirup Datta et al. "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets". In: *Journal of the American Statistical Association* 111 (June 2014).