Integrated Use of Machine Learning and Geostatistics for High Resolution Mapping of Ground-Level $PM_{2.5}$ Concentrations

Guofeng Cao
Department of Geosciences, Texas Tech University
Lubbock, TX USA
Guofeng.cao@ttu.edu

Ying Liu
Department of Geosciences, Texas Tech University
Lubbock, TX USA
Ying.liu@ttu.edu

Abstract

Fine particulate matter with aerodynamic diameters equal to or less than 2.5 micrometers ($PM_{2.5}$) is a major component of air pollutants. The adverse effects of $PM_{2.5}$ on public health have been well recognized. In this poster, we propose a statistical approach to integrate random forests, a commonly used machine learning algorithm, and regression kriging, a recognized geostatistical method, for high-resolution mapping of ground-level $PM_{2.5}$ concentrations. This approach jointly considers the heterogeneous geospatial variables that are closely related to the distribution of $PM_{2.5}$, including meteorological factors, socioeconomic development activities, and topographic information. The integration of the machine learning and geostatistical methods enables the effective modeling of the non-linear relationships between the $PM_{2.5}$ concentration and the predictor variables (via random forests) and the complex spatiotemporal effects (via kriging). Using this integrative approach, we produce a time-series (January 2014 to December 2014) monthly $PM_{2.5}$ concentration dataset at a spatial resolution of 500 m for the contiguous United States. The advantages of the proposed approach are discussed and highlighted with a performance comparison with a commonly used land use regression method.

Keywords: $PM_{2.5}$, air pollution, random forests, geostatistics, machine learning

1 Introduction

Fine particulate matter with aerosol dynamic diameters equal to or less than 2.5 micrometers ($PM_{2.5}$) has been identified as one of the three leading risk factors for human health. Exposure to $PM_{2.5}$ is estimated to cause 3.2 million premature deaths every year globally. To study and mitigate the adverse effects of $PM_{2.5}$ exposure on public health, accurately measuring ground-level PM2.5 concentrations is of essential importance.

The ground-based monitoring systems can provide accurate $PM_{2.5}$ measurements. However, no country in the world has yet established a monitoring network with a satisfying population coverage. Even in the United States (U.S.) the relatively developed $PM_{2.5}$ monitoring network with approximately 2,500 monitoring stations leaves many people living in suburban and rural areas unmonitored (Liu et al., 2018). To monitor the ground-level $PM_{2.5}$ concentrations at large geographical scales, remote sensing has proven to be a useful tool. A close relationship was found between the ground-level $PM_{2.5}$ concentration and satellite-observed variables, e.g., aerosol optical depth and land surface characteristics, which can be exploited to improve the $PM_{2.5}$ concentration mapping for large geographical areas. In this poster, we describe a statistical approach to integrate the advances in machine learning and geostatistics to exploit this close relationship for high resolution mapping of ground-level $PM_{2.5}$ concentrations.

Random forests (RF) is a commonly used machine learning method to capture the complex non-linear relationships between the response variable and related predictors (Breiman, 2001). In geostatistics literature, regression kriging (RK) represents a practical approach to integrate the linear regression and conventional kriging method for geostatistical mapping while accounting for spatial dependence (via kriging) of response variable and linear relationship with predict variables (via linear regression). Recently, a random forests-based regression kriging (RFRK) was proposed to replace the linear regression component in RK with random forests to consider the complex non-linear relationships between response and predictors (Hengel et al, 2015). The RFRK has shown performance advantages in mapping applications including mapping $PM_{2.5}$ concentrations (Liu et al., 2018). The previous study of Liu et al., (2018) adopted the RFRK method to refine a numerically derived dataset of $PM_{2.5}$ concentrations (with 10 km resolution) into a dataset with finer spatial resolution (1 km). Despite the improvements in spatial resolution and accuracy, the reliance on numerical models in Liu et al. (2018) limits the temporal resolution of results (at yearly). To mitigate this issue, this study extends the work of Liu et al. (2018) and applies the RFRK directly to in-situ $PM_{2.5}$ measurements and closely related meteorological variables and geographic variables (e.g., anthropogenic or socioeconomic development factors) to improve the ground-level $PM_{2.5}$ concentration mapping.

The meteorological variables used in this study include total precipitation, mean temperature, average dew point temperature, and vapor pressure deficit. The geographic variables include remote sensing imagery of nighttime lights that has been shown as a reliable proxy of socioeconomic development, vegetation coverage (NDVI) and topography. The in-situ $PM_{2.5}$ measurements are collected from the air quality monitoring network of U.S. Environmental Protection Agency's (EPA) that are primarily located in populated urban areas and Interagency Monitoring of Protected Visual Environments (IMPROVE) that are mostly in remote areas and national parts.

2 Method

2.1. Random forests-based regression kriging

Given a location s_0, the RFRK estimation of the $PM_{2.5}$ concentration at s_0, $\hat{p}(s_0)$, can be written as:

$$\hat{p}(s_0) = f_{RF}(NTL_0, NDV_0, ELE_0, TEM_0, DEW_0, PRE_0, AIR_0;\ \hat{\beta}) + \hat{\varepsilon}(s_0)$$

where $f_{RF}(\cdot)$ represents the deterministic trend modeled by random forests with parameters $\hat{\beta}$, and *NTL*, *NDV*, *ELE*, *TEM*, *DEW*, *PRE*, *AIR* denote brightness of nighttime lights, NDVI, elevation, mean temperature, dew point temperature, precipitation, and air pressure respectively. $\hat{\varepsilon}(s_0)$ represents the spatially correlated residuals that can be modeled with kriging to account for complex spatial patterns. The *NTL*, *NDV*, *ELE*, *TEM*, *DEW*, *PRE*, *AIR* values of each monitoring station are extracted from the corresponding imagery datasets and then put into the RF model to establish the relationships between $PM_{2.5}$ concentration and covariates. The monthly average $PM_{2.5}$ concentration dataset for the contiguous U.S. during January 2014 to December 2014 are produced at a spatial resolution of 500 m × 500 m. The performance of the RFRK methods is compared with the commonly used Land Use Regression (LUR) model to show the advantages of the RFRK method.

2.2. Accuracy assessment

A 10-fold cross-validation method is adopted to quantitatively assess accuracy of the RFRK-$PM_{2.5}$ and the LUR-$PM_{2.5}$ concentration datasets. In each fold of the validation, 90% of the *in-situ* $PM_{2.5}$ measurements are selected to compose a training dataset for the RFRK and the LUR models and the remaining 10% of the $PM_{2.5}$ data are used for test.

3 Results

3.1. Model performance

Figure 1 illustrates the differences of the average RMSEs, MAEs, and R2s of the 10-fold cross-validations for each month between the RF and the LUR models. Except October, the RMSEs and MAEs of the RF model are apparently lower than those of the LUR model. For the October, the RMSE and MAE of the LUR model are slightly higher than those of the RF model. Compared with the LUR-$PM_{2.5}$ concentration dataset, accuracy of the RFRK- $PM_{2.5}$ concentration maps are more stable, demonstrated that the RMSEs and MAEs of the RF model maintain2.0μg/m3 (±0.05μg/m3) and 1.7μg/m3 (±0.1μg/m3) while those of the LUR model vary greatly across the 12 months (see Figure 1). Specifically, the LUR- $PM_{2.5}$ concentration dataset shows much larger errors in winter season. The better performance of the RF model can also be shown by the larger fitting R2 values for all of the 12 months. Except for April, the R2 values of the RF model are all higher than 0.6 while those of the LUR model are nearly all smaller than 0.4.

3.2. Spatiotemporal variations

Figure 2 displays the RFRK-derived monthly average $PM_{2.5}$ concentration images for 12 months of 2014 at the 500 m × 500 m spatial resolution. For each of the given months, the $PM_{2.5}$ concentrations are apparently higher in the Eastern part than in the Western part. The large population densities in the Eastern region of the contiguous U.S. lead to more emissions $PM_{2.5}$ from commercial and industrial activities and therefore associate with high $PM_{2.5}$ levels. California Valley is an exception in the West region. The California Valley, especially in the winter season (i.e., November, December, and January), has apparently higher $PM_{2.5}$ concentrations than its surrounding regions. The closed terrain of the California Valley greatly constrains diffusion and dispersion of air pollutants and so leads to high $PM_{2.5}$ concentrations. Furthermore, the dominant meteorological conditions of California valley in winter are cool and moist with low wind speed, which promotes the formations of the secondary $PM_{2.5}$ components (Liu et al, 2017).

4 Conclusion

This study highlights the potential of combining machine learning and geostatistical methods on mapping concentrations of air pollutants. Compared with the commonly used LUR model, the RFRK method can simultaneously consider the non-linear relationship with predictor variables (with RF) and the complex spatial effects (with kriging) in a practical and effective manner. Additionally, we also show the effectiveness of the selected geographical variables in ground-level $PM_{2.5}$ concentration mapping, particularly the brightness of NTL extracted from the VIIRS-DNB monthly image composites as a comprehensive indicator of human activities. In the future, we plan to exploit the capabilities of deep learning methods (e.g. recurrent neural network) to replace the present random forests in the hybrid model to produce more accurate $PM_{2.5}$ concentration maps at finer spatiotemporal resolutions.

Reference

- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., Sila, A., et al. (2015). Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PloSOne*, 10(6), e0125814.
- Liu, Y., Zhao, N., Vanos, J. K., & Cao, G. (2017). Effects of synoptic weather on ground-level PM2. 5 concentrations in the United States. *Atmospheric Environment*, 148, 297-305.
- Liu, Y., Cao, G., Zhao, N., Mulligan, K., & Ye, X. (2018). Improve ground-level $PM_{2.5}$ concentration mapping using a random forests-based geostatistical approach. *Environmental Pollution*, 235, 272-282.

Figure 1: Accuracy comparisons between RFRK and LUR: RMSEs, MAEs, and R^2s of the RFRK model and the LUR model in the 100-time cross-validation during each month in the year 2014.

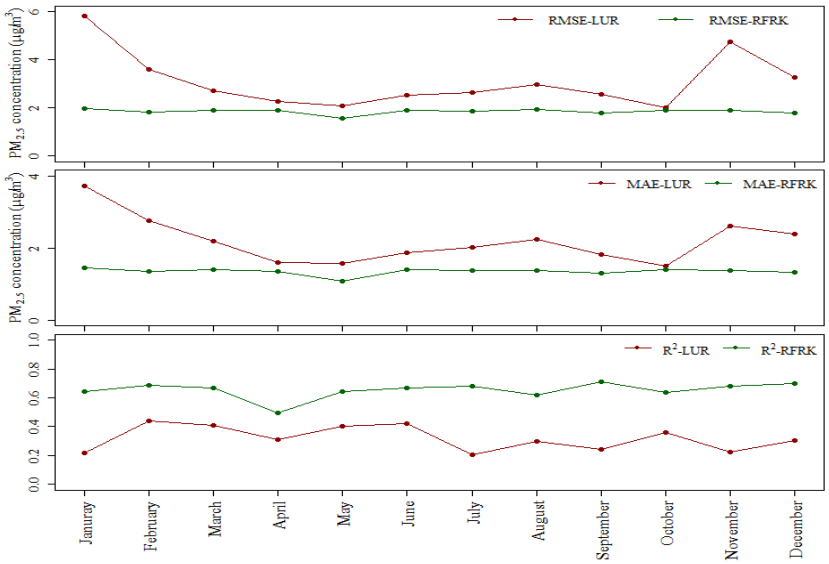

Figure 2: Monthly RFRK-$PM_{2.5}$concentration dataset for the contiguous United States during the year of 2014