Environmental Pollution 235 (2018) 272–282

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

# Improve ground-level $PM_{2.5}$ concentration mapping using a random forests-based geostatistical approach[☆]

Ying Liu [a, b], Guofeng Cao [a, b, *], Naizhuo Zhao [b], Kevin Mulligan [a, b], Xinyue Ye [c]

[a] Department of Geosciences, Texas Tech University, Lubbock, TX 79409, USA
[b] Center for Geospatial Technology, Texas Tech University, Lubbock, TX 79409, USA
[c] Department of Geography, Kent State University, Kent, OH 44240, USA

ARTICLE INFO

*Article history:*
Received 18 April 2017
Received in revised form
18 December 2017
Accepted 19 December 2017

*Keywords:*
$PM_{2.5}$
Remote sensing
Random forests
Geostatistics
Air pollution

ABSTRACT

Accurate measurements of ground-level $PM_{2.5}$ (particulate matter with aerodynamic diameters equal to or less than 2.5 μm) concentrations are critically important to human and environmental health studies. In this regard, satellite-derived gridded $PM_{2.5}$ datasets, particularly those datasets derived from chemical transport models (CTM), have demonstrated unique attractiveness in terms of their geographic and temporal coverage. The CTM-based approaches, however, often yield results with a coarse spatial resolution (typically at 0.1° of spatial resolution) and tend to ignore or simplify the impact of geographic and socioeconomic factors on $PM_{2.5}$ concentrations. In this study, with a focus on the *long-term* $PM_{2.5}$ distribution in the contiguous United States, we adopt a random forests-based geostatistical (regression kriging) approach to improve one of the most commonly used satellite-derived, gridded $PM_{2.5}$ datasets with a refined spatial resolution (0.01°) and enhanced accuracy. By combining the random forests machine learning method and the kriging family of methods, the geostatistical approach effectively integrates ground-based $PM_{2.5}$ measurements and related geographic variables while accounting for the non-linear interactions and the complex spatial dependence. The accuracy and advantages of the proposed approach are demonstrated by comparing the results with existing $PM_{2.5}$ datasets. This manuscript also highlights the effectiveness of the geographical variables in long-term $PM_{2.5}$ mapping, including brightness of nighttime lights, normalized difference vegetation index and elevation, and discusses the contribution of each of these variables to the spatial distribution of $PM_{2.5}$ concentrations.



## 1. Introduction

Environmental and disease experts widely recognize that ambient particulate matter with aerodynamic diameters $\leq 2.5$ micrometers ($PM_{2.5}$) adversely affects public health (Brunekreef and Forsberg, 2005; Pope et al., 2002). Lim et al. (2013) estimated that 3.2 million premature deaths occur due to high $PM_{2.5}$ exposure every year. For the year of 2013, ambient particulate matter was identified as a leading risk for disease burden globally (Brauer et al., 2012). Health and environmental organizations in many countries (e.g., United States Environmental Protection Agency, Centers for Disease Control and Prevention, Canada Regional Air Quality Deterministic Prediction System) have paid particular attention to monitoring changes in ground-level $PM_{2.5}$ concentrations. However, no country has yet established a $PM_{2.5}$-monitoring system that covers its entire population with a sufficiently high density of monitoring stations. Even in the United States (U.S.) and other developed countries, the relatively dense $PM_{2.5}$ monitoring networks tend to cluster in urban areas leaving many people in suburban and rural areas unmonitored. This thus makes it difficult to study the effects of $PM_{2.5}$ exposure on human health and the impacts of anthropogenic activities on $PM_{2.5}$ concentrations over large areas at relatively fine spatial and temporal scales (Zou et al., 2016a).

To address this issue, remote sensing has been successfully used to estimate ground-level $PM_{2.5}$ concentrations. Wang and Christopher (2003) and Engel-Cox et al. (2004) explored the close relationship between ground-level $PM_{2.5}$ concentrations and aerosol optical depth (AOD) derived from Moderate Resolution Imaging Spectroradiometer (MODIS). This close relationship has

☆ This paper has been recommended for acceptance by Eddy Y. Zeng.

* Corresponding author. Department of Geosciences, Texas Tech University, Lubbock, TX 79409, USA
*E-mail address:* guofeng.cao@ttu.edu (G. Cao).

https://doi.org/10.1016/j.envpol.2017.12.070
0269-7491/

been exploited for $PM_{2.5}$ concentration mapping from both statistical and geophysical perspectives. Most statistical methods are derived from a regression setting (e.g., land use regression, geographically weighted regression) where remotely sensed AOD and other related covariates (e.g., meteorological variables and land use information) are used as predictors to fit the $PM_{2.5}$ measurements from monitoring stations (Hoek et al., 2008; Ma et al., 2014). These methods are flexible in incorporating measurements of diverse source of variables, but the availability of the *in situ* $PM_{2.5}$ measurements limits the application only to areas with a high density of monitoring stations.

The geophysical models, on the other hand, quantify the relationship between the total column of AOD and ground-level $PM_{2.5}$ concentrations based on atmospheric chemical transport models (CTM). These CTMs describe the diffusion of chemical compositions in atmospheric dynamics by combining assimilated meteorology and emission inventory. Given a location and a time, a CTM can generate vertical profiles of aerosol distribution, from which the relationship between ground-level $PM_{2.5}$ concentrations and AOD, or the $PM_{2.5}$/AOD ratio, can be approximately derived. This relationship was first applied to the MISR (Multi-angle Imaging SpectroRadiometer) AOD retrievals with the GEOS (Goddard Earth Observing System)–Chem CTM (http://www.geos-chem.org) over the U.S. (Liu et al., 2004), and then extended globally using MODIS and MISR AOD retrievals (van Donkelaar et al., 2006). This CTM-based approach has proven to be effective in practice and has been exploited extensively. Most notably, van Donkelaar et al (van Donkelaar et al., 2010) combined AOD retrievals of MODIS and MISR, and used this relationship to develop a set of long-term ground-level $PM_{2.5}$ concentration maps with global coverage at 0.1° spatial resolution. Using AOD retrievals from MISR and SeaWiFS (Sea-Viewing Wide Field-of-View Sensor), Boys et al. (2014) applied the $PM_{2.5}$/AOD ratio to generating a unified fifteen-year global time series of gridded $PM_{2.5}$ concentrations. In a similar fashion, van Donkelaar et al (van Donkelaar et al., 2015a) improved the results of van Donkelaar et al (van Donkelaar et al., 2010) by combining three different satellite AOD sources, including an optimal estimation (OE) of AOD derived from MODIS (van Donkelaar et al., 2013), and AOD retrievals of MISR (Diner et al., 2005) and SeaWiFS (Hsu et al., 2013). The CTM output can also be jointly used with other related variables as predictors in statistical methods to improve the $PM_{2.5}$ estimation e.g., (Datta et al., 2016; Denby et al., 2008; Hamm et al., 2015; Shaddick et al., 2018).

While the geophysical approach has proved to be effective, it is subject to considerable limitations. First, due to the high computational complexity, the CTM is only computationally feasible at a relatively coarse spatial resolution. The resolution of $PM_{2.5}$/AOD ratio generated by the GEOS-Chem CTM usually ranges from $4° \times 5°$–$0.25° \times 0.3125°$ (or roughly 25 km–28 km) and the LOTOS-EUROS can possibly generate results at $1/8° \times 1/16°$ (Manders-Groot et al., 2016). Most of the currently available satellite AOD retrievals are restricted to 0.1°–0.18° of spatial resolution (MODIS at 0.1°, MISR at 0.18° and SeaWiFS at 0.14°), although higher resolution of AOD (e.g., Multi-Angle Implementation of Atmospheric Correction at 0.01°) starts becoming available. Van Donkelaar et al (van Donkelaar et al., 2010; van Donkelaar et al., 2015a) interpolated the $PM_{2.5}$/AOD ratio at $2° \times 2.5°$ to roughly match the resolution of AOD retrievals and generated $PM_{2.5}$ concentration datasets at spatial resolution of 0.1°. However, this spatial resolution (i.e., 0.1°) remains insufficient for applications at the city or finer geographic scales. Fig. 1 shows the annual average $PM_{2.5}$ concentrations (van Donkelaar et al., 2015a) of the contiguous U.S. for the year 2005, where Dallas-Fort Worth, TX and Philadelphia, PA, the 9th and the 15th largest metropolis in the U.S., only occupy 53 and 47 pixels respectively. Secondly, the CTM tends to under-utilize the valuable ground-based $PM_{2.5}$ measurements and can lead to large discrepancies from the *in situ* measurements (van Donkelaar et al., 2015b). Thirdly, the input data sources used in the CTMs are incomplete and imperfect. The GEOS-Chem CTM model typically has two types of inputs: the observations of meteorological variables at different heights and the pollutant estimations from lists of major emission sources. The meteorological variables are often observed from satellites and the measurements can be contaminated by noises and sensor calibrations. More importantly, the GEOS-Chem model relies on the U.S. National Emission Inventory (NEI) for major $PM_{2.5}$ emission sources. The chemical emissions in the NEI cannot capture the whole picture of human activities that affect the $PM_{2.5}$ concentrations, and despite the regular updates of the NEI, changes of chemical emission sources cannot be incorporated in a timely manner. The quality issues associated with the emission data sources and the meteorological measurements can inevitably propagate into the GEOS-Chem results.

The primary goal of this study is to mitigate the above limitations in the long-term, satellite-derived estimation of $PM_{2.5}$ concentrations for the contiguous U.S. Besides the meteorological variables and chemical emission inventories that the CTM utilizes, geographic variables (e.g., socioeconomic development) have been shown to closely relate to the long term ground-level $PM_{2.5}$ concentrations and can therefore be used to improve long-term $PM_{2.5}$ concentration mapping (van Donkelaar et al., 2015b; van Donkelaar et al., 2016; Di et al., 2016; Liu et al., 2009; Ross et al., 2007). Recent studies have shown that the quantitative relationships between the $PM_{2.5}$ concentrations and the geographic variables are nonlinear, and likely to vary across different regions (Fang et al., 2016; Zou et al., 2016b, 2016c) and different geographic scales (Zhai et al., 2016). Accordingly, in this study, we adopt a geostatistical approach, namely random forests-based regression kriging (RFRK), to refine the accuracy and spatial granularity of the CTM and satellite-derived datasets by integrating *in situ* $PM_{2.5}$ measurements with closely related geographic variables. In the RFRK, random forests (RF) regression (Breiman, 2001), a robust nonlinear and nonparametric estimator, is used to model the relationship between $PM_{2.5}$ concentrations and the related geographic variables, and kriging, a commonly used geostatistical method for spatial interpolation (Chiles and Delfiner, 2009), is applied for residuals of RF regression to take account of the spatial dependence. The integration of the random forests machine learning and kriging family of methods gives the RFRK framework many appealing advantages, which will be discussed in the following sections.

In the remainder of the manuscript, we first introduce the data sources and verify the correlations between ground-measured $PM_{2.5}$ concentrations and the selected geographic variables including the brightness of nighttime lights (NTL), Normalized Difference Vegetation Index (NDVI), and the elevation of the contiguous U.S. We then discuss the methodology including the RFRK method and how the RFRK was used to refine one of the most accurate and commonly used long-term gridded annual $PM_{2.5}$ concentration datasets ($0.1° \times 0.1°$) for the contiguous U.S (van Donkelaar et al., 2015a) to a finer resolution ($0.01° \times 0.01°$) for the years 2000–2013. To highlight the advantages of our approach, we compare our results with another recent $0.01° \times 0.01°$ $PM_{2.5}$ concentration dataset based on a geographically weighted regression (GWR) approach (van Donkelaar et al., 2015b). In addition, to demonstrate the effectiveness of using the brightness of NTL, NDVI, and elevation to improve long-term $PM_{2.5}$ concentrations, we repeat the process using only a subset of the geographic variables. Lastly, the contributions of the adopted geographic variables in the refining process are discussed by investigating the importance and their contributions to the long-term $PM_{2.5}$ estimations.