proceedings

MDPI

Proceedings

Daily Concentrations of $PM_{2.5}$ in the Valencian Community Using Random Forest for the Period 2008–2018 †

Soledad Natacha Represa [1,2,*], Jesús Palomar-Vázquez [2], Andrés Porta [1] and Alfonso Fernández-Sarría [2]

[1] Centro de Investigaciones del Medioambiente, Universidad Nacional de La Plata, Consejo Nacional de Investigaciones Científicas y Técnicas, La Plata 1900, Argentina

[2] Geo-Environmental Cartography and Remote Sensing Group, Department of Cartographic Engineering, Geodesy and Photogrammetry, Universitat Politècnica de València, València 46022, Spain

* Correspondence: solrepresa@quimica.unlp.edu.ar

† Presented at the II Congress in Geomatics Engineering, Madrid, Spain, 26–27 June 2019.

Published: 15 July 2019

Abstract: Fine particulate matter ($PM_{2.5}$) is a global problem that affects the population health and contributes to climate change. Remote sensing provides useful information for the development of air quality models. This work aims to obtain a daily model of $PM_{2.5}$ levels in the Valencian Community with a resolution of 1 km for the period 2008–2018. MODIS-MAIAC images, meteorological parameters of the MERRA-2 project, land cover information and ground level measurements of $PM_{2.5}$ levels were analysed with Random Forest. The verification of the model was carried out using cross-validation repeated ten times, and an evaluation of a test set with 20% of the collected information. The final model was used to generate maps of the daily concentrations of $PM_{2.5}$ for the area of the Valencian Community throughout the study period.

Keywords: $PM_{2.5}$; LUR; Random Forest; MODIS; MERRA-2

1. Introduction

Atmospheric aerosols are a critical compound in the atmosphere due to their influence on climate change and population health [1]. Remote sensing has contributed significantly to the air quality study in order to capture the spatial-temporal variation of pollutants. Previous papers have shown that the amount of light absorbed or scattered by suspended particles, aerosol optical depth (AOD), is a relevant parameter for estimating $PM_{2.5}$ at ground level [2,3].

Moderate Resolution Imaging Spectroradiometer (MODIS) products are widely used in atmospheric models, as they have a daily review and a convenient spatial resolution for regional and local studies [2]. The recent Multi-angle Implementation of Atmospheric Correction (MAIAC) algorithm presents new opportunities for the development of atmospheric aerosol models [3]. These images employ the MODIS Aqua and Terra data and improve spatial resolution from 25 to 1 km [3].

Land use regression (LUR) models have had a broad application in different parts of the world. The LURs incorporate satellite images and meteorological and land use information as predictors to model PM_{10} (particulate matter with diameter < 10 μm) and $PM_{2.5}$ (particulate matter with diameter < 2.5 μm). Recent work shows positive results in the use of Random Forest (RF) for LUR models [2].

The air of the Valencian Community is affected by the growth of the vehicle fleet, industrial production, Sahara dust events and biomass burning smoke. An air quality model would allow the risk of exposure to be estimated. Previous works presented the correlation between daily $PM_{2.5}$ ground-measures and AOD MODIS for this region [4,5]. The aim of this work is to apply an RF model, using 15 atmospheric variables and characteristic for land use to estimate the daily $PM_{2.5}$ ground-concentration at the 1 km grid for 2008–2018 in Valencia Community.

2. Materials and Methods

2.1. Study Area

The Valencian Community is to the east of the Iberian Peninsula, on the Mediterranean coast (Figure 1). The population is mainly concentrated in urban centres, in particular, in the metropolitan areas of Valencia and Alicante. Segura et al. [6] found a significant relationship between black smoke and the number of emergency admissions for heart disease in Valencia city.

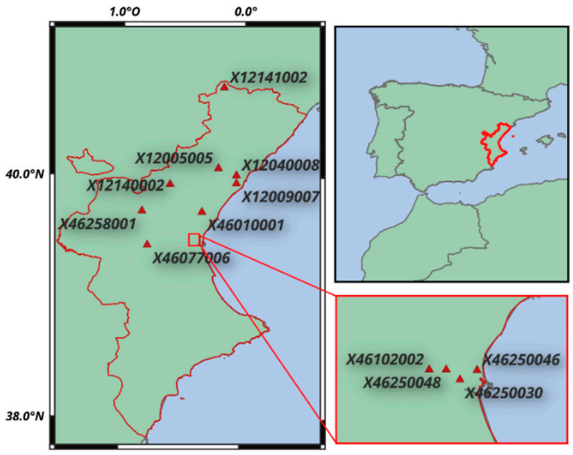

Figure 1. Valencian Community and the $PM_{2.5}$ monitoring stations.

2.2. Data Sets

The hourly concentrations at the ground level of $PM_{2.5}$ were download from the *Valencian Network for Monitoring and Control of Atmospheric Pollution of the Generalitat Valenciana* for the period between 1 January 2008 to 30 September 2018 (http://www.agroambient.gva.es/va/web/calidad-ambiental). During this time, 24 stations measured $PM_{2.5}$ continuously, of which only 12 stations had a percentage of missing values less than 30% (Figure 1). For this work, we calculate the average $PM_{2.5}$ concentrations between the hours of the Aqua and Terra satellite overpass.

MODIS-MAIAC products were downloaded from the *Level-1 and Atmosphere Archive and Distribution System* website (https://ladsweb.modaps.eosdis.nasa.gov/) [2]. AOD measurements were calibrated with Aerosol Robotic Network (AERONET) data Level 2.0 (http://aeronet.gsfc.nasa.gov/) [4]. The fraction of the artificial surface was estimated for each pixel using the information provided by the Corine Land Cover project for the year 2012 [7]. The terrain elevation was obtained from the *Consultative Group for International Agricultural Research* Consortium for Spatial Information GEOPortal (http://srtm.csi.cgiar.org) with a resolution of 90 m at the equator [8].

Finally, atmospheric conditions data was download from the *NASA's Goddard Earth Sciences Data and Information Services Center* website (https://disc.gsfc.nasa.gov/). The Modern Era-Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) is a global atmospheric reanalysis using the Goddard Earth Observing System Model, Version 5 with its Atmospheric Data Assimilation System, at a spatial resolution of 0.5° × 0.625° [9].

2.3. Statistical Analysis

The RF model was trained with 80% of the collected information and evaluated with the remaining 20%. The model was built using $PM_{2.5}$ observations as dependent variables. The predictor variables of the model were: (1) atmospheric variables: aerosol optical depth (AOD), surface pressure (PS), relative humidity (RH), surface temperature (T), surface wind component u (U), surface wind component v (V), black carbon surface mass concentration (BCSMASS), dimethyl-sulfide surface mass concentration

(DMSSMASS), dust surface mass concentration (DUSMASS), SO_4 surface mass concentration (SO4SMASS), sea salt surface mass concentration (SSSMASS25), total precipitation (PRECTOT), high cloud cover (CLDHGH), low cloud cover (CLDLOW); (2) land use: fraction of artificial surface (CLC_1); (3) terrain elevation (DEM). The data were centered and scaled prior to being incorporated into the model.

The model verification contains a 10-fold cross validation (cv). The feature importance of each variable is then calculated after the model fitting process. Analyses were performed using the R language [10] and the "caret" package for RF model [11]. The maps were made with the software QGIS [12].

3. Results and Discussion

The ground $PM_{2.5}$ average for the entire period was 8.3 μg.m^{-3}. The months with the highest concentration were February (10.9 μg.m^{-3}) and March (10.4 μg.m^{-3}) and the lowest May (6.13 μg.m^{-3}) and June (7.22 μg.m^{-3}). Station X46250048 was the site with the most elevated average $PM_{2.5}$ levels (11.8 μg.m^{-3}). This station is in a busy urban area of the city of Valencia. The lowest station (X12141002, 6.13 μg.m^{-3}) is situated in Viver, a small village of fewer than 2000 inhabitants (Figure 1).

The top RF accuracy with a *ntree* = 10 was with *mtry* = 7. The variables PRECTOT, CLDHGH and AOD were the most significant predictors that contribute to the model construction process. PRECTOT and CLDHGH have a pronounced influence on the deposition and dispersion of pollutants. Based on these results, we ran the daily model for the rest of the Community of Valencia. Figure 2a presents a map example of average $PM_{2.5}$ concentrations for the years 2008.

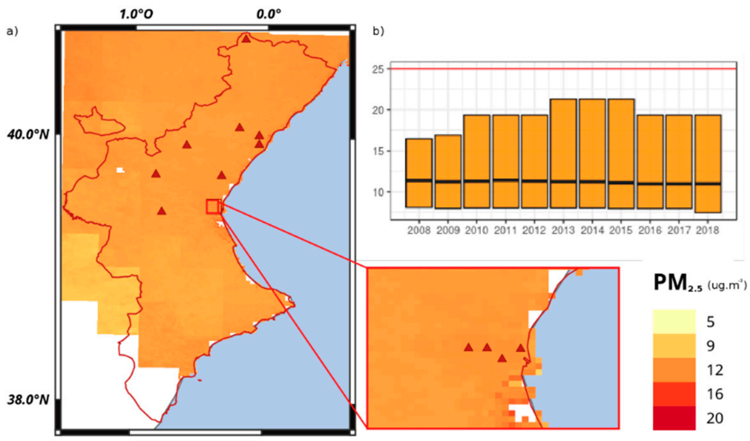

Figure 2. (**a**) Example of average $PM_{2.5}$ concentrations map in 2008 for the Valencian Community and the city of Valencia. (**b**) Range of annual mean concentrations modelled for the Valencia Community (2008–2018).

Results indicate a high predictive capability of the RF model, with an extremely high R^2 (0.89). The predictions for the test set have a good fit, with a root-mean-square error (RMSE) equal to 2.29 and a mean absolute error equal (MAE) to 0.67. In turn, the errors in estimations occur when modelling the highest concentrations. These may be a consequence of two factors. On the one hand, high values are under-represented in the data set because high concentrations are particular events in the time series. On the other hand, the elevated measurements could be due to situations strongly influenced by events of micro-scale or of short temporal duration. Problems of immeasurability appear when modelling a point data from variables registered in portions of area (pixel).

The European Directive 2008/50/EC fixes the annual concentration for $PM_{2.5}$ in 25 μg.m^{-3}. The model does not show areas that exceeded this annual level (Figure 2b). The year with the smallest spatial variation was 2008, with a minimum annual modelled concentration of 8 μg.m^{-3} and a maximum of 16 μg.m^{-3}, while the year with the most considerable variation was 2015 (range: 7–21 μg.m^{-3}).

4. Conclusions

This study proposed a daily concentration model of $PM_{2.5}$ based on the RF for the Valencia Community (Spain). The method used AOD MAIAC measures, MERRA-2 products and land cover information to simulate ground $PM_{2.5}$ values. Based on the evaluation of the 10-fold cross-validation method and the test set verification method, the model performs very well. With this data, we were able to predict ~90% of the temporal and spatial variability of $PM_{2.5}$. More RF trees and an exhaustive analysis of predictor variables will bring benefits to $PM_{2.5}$ simulations in the future. This work may provide support for air quality management and may also give evidence for epidemiological studies.

Author Contributions: Conceptualization, A.P.; Methodology, formal analysis, and writing—original draft preparation, S.N.R.; writing—review, editing and supervision, J.M.P.V. and A.F.-S.

Funding: National Agency for Scientific and Technological Promotion (Argentina) supported this research (PICT-2017-1670).

Conflicts of Interest: Authors declare no conflict of interest.

References

1. WHO (World Health Organization). Available online: https://www.who.int/en/newsroom/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed on 5 May 2019).
2. Stafoggia, M.; Bellander, T.; Bucci, S.; Davoli, M.; de Hoogh, K.; De'Donato, F.; Gariazzo, C.; Lyapustin, A.; Michelozzi, P.; Renzi, M.; et al. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environ. Int.* **2019**, *124*, 170–179.
3. Lyapustin, A.; Wang, Y.; Korkin, S.; Huang, D. MODIS Collection 6 MAIAC algorithm. *Atmos. Meas. Tech.* **2018**, *11*, 5741–5765.
4. Represa, N.S.; Fernández-Sarría, A.; Porta, A.; Vázquez, J.P. Assessment of satellite aerosol optical depth to estimate particulate matter distribution in Valencia city. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 9141–9144.
5. Segura, S.; Estellés, V.; Utrillas, M.P.; Martínez-Lozano, J.A. Long term analysis of the columnar and surface aerosol relationship at an urban European coastal site. *Atmos. Environ.* **2017**, *167*, 309–322.
6. Ballester, F.; Tenias, J.M.; Perez-Hoyos, S. Air pollution and emergency hospital admissions for cardiovascular diseases in Valencia, Spain. *J. Epidemiol. Community Health* **2001**, *55*, 57–65.
7. EEA (European Environmental Agency). *Corine Land Cover Technical Guide–Addendum 2000*; Technical Report No. 40; EEA: Copenhagen, Denmark, 2013.
8. Jarvis, A.; Reuter, H.I.; Nelson, A.; Guevara, E. Hole-Filled Seamless SRTM Data V4, International Centre for Tropical Agriculture (CIAT). 2008. Available online: http://srtm.csi.cgiar.org (accessed on 1 April 2019).
9. Gelaro, R.; McCarty, W.; Suárez, M.J.; Todling, R.; Molod, A.; Takacs, L.; Randles, C.A.; Darmenov, A.; Bosilovich, M.G.; Reichle, R.; et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *J. Clim.* **2017**, *30*, 5419–5454.
10. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019. Available online: https://www.R-project.org/ (accessed on 1 May 2019).
11. Kuhn, M. 2019. caret: Classification and Regression Training. R Package Version 6.0-82. Available online: https://rdrr.io/cran/caret/ (accessed on 1 June 2019).
12. QGIS Development Team. QGIS Geographic Information System. Version 2.18.15. Open Source Geospatial Foundation Project. Available online: http://qgis.osgeo.org (accessed on 1 November 2018).