# Urbanization and regional air pollution across South Asian developing countries - A nationwide land use regression for ambient $PM_{2.5}$ assessment in Pakistan

Yuan Shi[a], Muhammad Bilal[b]*, Hung Chak Ho[c], Abid Omar[d]

[a]Institute of Future Cities, The Chinese University of Hong Kong, Hong Kong SAR, China, shiyuan@cuhk.edu.hk

[b]School of Marine Sciences, Nanjing University of Information Science & Technology, Nanjing, 210044, China, muhammad.bilal@connect.polyu.hk

[c]Department of Urban Planning and Design, The University of Hong Kong, Hong Kong SAR, China, hcho21@hku.hk

[d]Pakistan Air Quality Initiative, Pakistan, abidomar@pakairquality.com

*Corresponding author:

Email: muhammad.bilal@connect.polyu.hk

Address: School of Marine Sciences, Nanjing University of Information Science and Technology, Nanjing 210044, China

**Abstract**

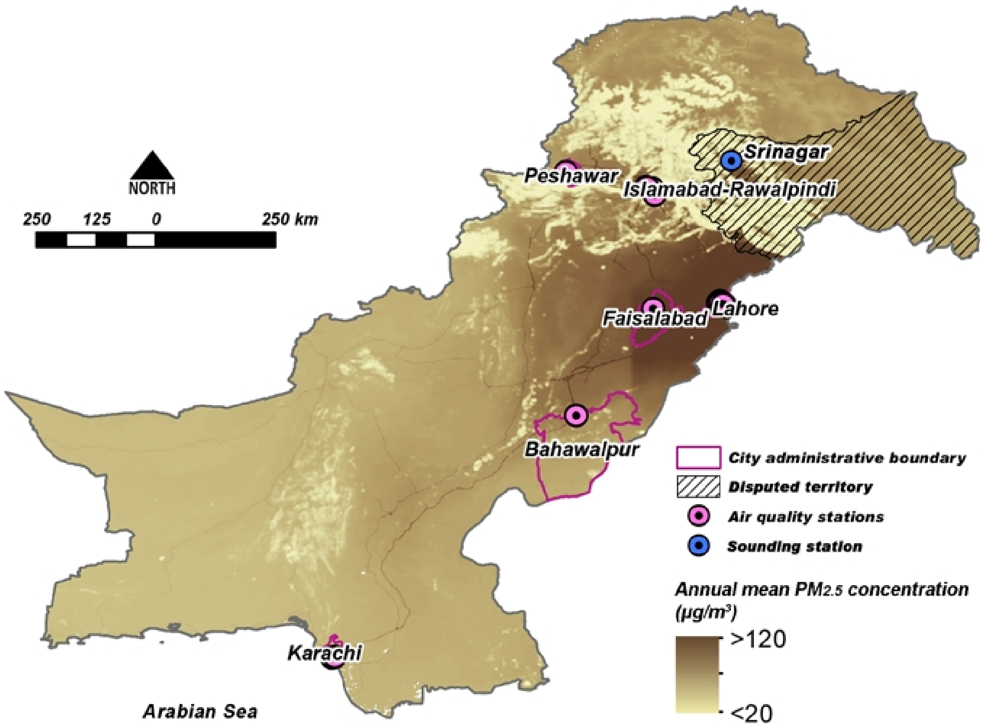
Rapid economic growth, urban sprawl, and unplanned industrialization has increased socioeconomic statuses but also decreased air quality in South Asian developing countries. Therefore, severe increase in air pollution has been a threat of local population, regarding health statuses, livability and quality of life. It is necessary to estimate fine-scale spatiotemporal distribution of ambient $PM_{2.5}$ in a national context so that the environmental planners and government officials can use it for environmental protocol development and policy-making. In this study, a spatiotemporal land use regression (LUR) model is developed to refine global air quality data to the national-scale ambient $PM_{2.5}$ exposure in a high-density country in South Asia - Pakistan. Combining with transport network, patterns of land use, local meteorological conditions, geographic characteristics, landscape characteristics, and satellite-derived data, our resultant model explains 54.5% of the variation in ambient $PM_{2.5}$ concentration level. Furthermore, tree coverage and road transport are identified to be two influential factors of the national-scale spatial variation of $PM_{2.5}$ in Pakistan, which implied that urbanization might be the major cause of air pollution across the country. In conclusion, our resultant LUR model as well as the spatial map of ambient $PM_{2.5}$ concentration level can be used as a supporting tool for national health risk management and environmental planning, and could also contribute to the air quality management and pollution reduction actions of Pakistan.

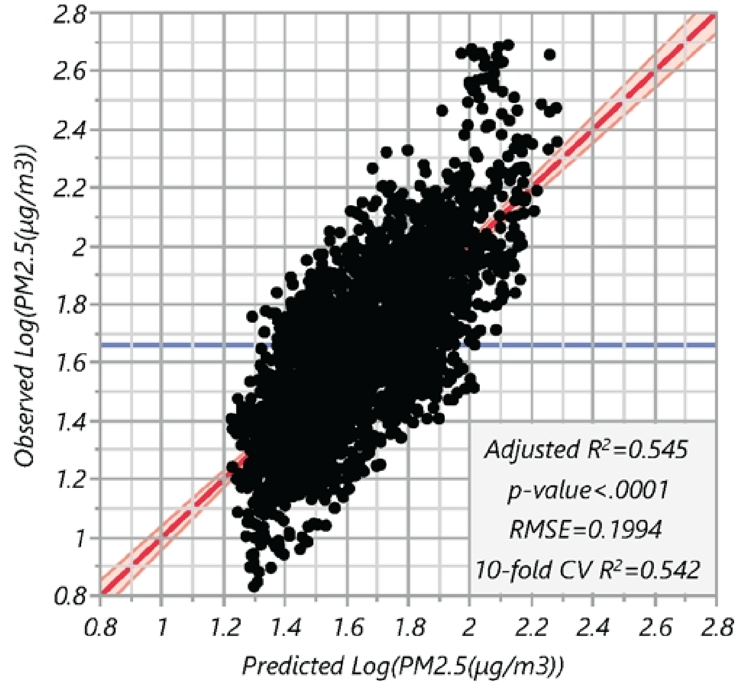
**Highlights**

- National-scale ambient $PM_{2.5}$ exposure assessment in Pakistan;
- A spatiotemporal land use regression $PM_{2.5}$ model was developed;
- Global air quality datasets were refined to local scenario;
- Tree coverage and road transport are critical factors of $PM_{2.5}$ spatial variation;
- Results can be used as reference for national health risk management of Pakistan.

**Keywords**

$PM_{2.5}$; Land use regression; Exposure; Urbanization; Pakistan


NORTH
250 125 0 250 km
Peshawar
Islamabad-Rawalpindi
Srinagar
Faisalabad
Lahore
Bahawalpur
Karachi
Arabian Sea
City administrative boundary
Disputed territory
Air quality stations
Sounding station
Annual mean PM2.5 concentration (μg/m³)
>120
<20


Observed Log(PM2.5(μg/m3))
Predicted Log(PM2.5(μg/m3))
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
Adjusted R²=0.545
p-value<.0001
RMSE=0.1994
10-fold CV R²=0.542

**Highlights**

- National-scale ambient $PM_{2.5}$ exposure assessment in Pakistan;
- A spatiotemporal land use regression $PM_{2.5}$ model was developed;
- Global air quality datasets were refined to local scenario;
- Tree coverage and road transport are critical factors of $PM_{2.5}$ spatial variation;
- Results can be used as reference for national health risk management of Pakistan.

# 1. Introduction

Air pollution has been regarded as one of the most serious environmental issues by the World Health Organization (WHO) (WHO, 2016). It is also a growing environmental problem in South Asian developing countries with rapid industrialization, such as Pakistan (UNEP, 2015). Among all major air pollutants, particulate matters (PM) with an aerodynamic diameter smaller than 2.5 μm ($PM_{2.5}$) have been identified as one of the biggest environmental risks due to its extreme harmfulness to human health (WHO, 2005). Exposure to high concentration levels of ambient $PM_{2.5}$ dramatically increase the health burdens of stroke mortality, cardiovascular disease, respiratory diseases, and lung cancer (Davidson et al., 2005; Russell and Brunekreef, 2009). Recent studies have also found that $PM_{2.5}$ can reduce the cognitive function of a person (Lavy et al., 2014; Power et al., 2011). According to WHO's report, health burdens due to human exposure to ambient $PM_{2.5}$ have been observed over an annual averaged ambient $PM_{2.5}$ concentration level higher than 10 μg/m$^3$. Therefore, this particular value has been set as a standard threshold by in WHO guideline.

## 1.1. Research Background and Relevant Studies

Pakistan is one of the biggest countries in South Asia. It is also a relatively high-density country with a population of more than 207.8 million. In addition, Pakistan is the 33$^{rd}$-largest country in the world (area: approximately 882,000 km$^2$). According to a previous study, fast economic growth, rapid urban sprawl, and unplanned industrialization are the major causes of severe air pollution issues across the country (Rasheed et al., 2015). Based on the aforementioned ambient $PM_{2.5}$ threshold, 100% of the total population of Pakistan are living in places with $PM_{2.5}$ concentrations that far beyond the level of resilience (Cohen et al., 2017), such that $PM_{2.5}$ has become one of the most serious air pollutants in the country (Colbeck et al., 2010). There are increasing debates and concerns about the negative environmental and health effects of poor air quality in Pakistan. Although Pakistan Environmental Protection Agency has started to take measures in monitoring air quality and reducing industrial pollution emissions (Saleem and Sughra, 2018), there are extraordinary increases in air pollution levels in Pakistan for the past several years due to the insufficient governmental intervention and ineffective control strategies and practical actions. Extraordinary rises in air pollution have appeared in the past few years, particularly during the wintertime (Javed et al., 2015; Mehmood et al., 2018). Poor air quality issues have been reported in many major cities of Pakistan, such as Karachi, Lahore, Quetta, Peshawar, Islamabad, and Rawalpindi (Rasheed et al., 2015). As a result, more than one-fifth of mortality in Pakistan each year are attributed to air pollution (Cohen et al., 2017). All the above facts degrade the living quality of citizens of Pakistan, especially those residents who are living in highly urbanized areas across the country.

As the increasing awareness of adverse impacts caused by air pollution, particularly the $PM_{2.5}$ environmental pollution, recent studies have been conducted to investigate the temporal change of $PM_{2.5}$ in Pakistan. For example, air quality monitoring and meteorological data in several major cities (Islamabad, Lahore, Peshawar, and Quetta) were collected and analyzed by a local study of Pakistan to quantify the current situations of air quality within the major urban environments in Pakistan (Rasheed et al., 2015). The results show that $PM_{2.5}$ mass concentration level is negatively correlated with meteorological parameters (air temperature and wind speed) and clearly associated with the traffic-related pollutants emissions. Specifically, the diurnal variation observed in all the cities suggests a strong association of $PM_{2.5}$ with vehicular traffic (Rasheed et al., 2015). As the road transport has been identified as a major emission source of PM and a major influential factor of the air quality in Pakistan, a literature-based desktop study was conducted to understand how the transport affects the urban air pollution in Pakistan (Ilyas, 2007). In this desktop study, the influential role of transport in urban air quality in Pakistan was discussed.

To further understand the main contributors of the PM in the urban environment of Pakistan, source apportionment was performed in Peshawar, a major city in Northern Pakistan (Alam et al., 2015). The results identified several major emission sources of PM: road and soil dust, emission from vehicular traffic, industrial activities, and household combustion. To evaluate the impact of biomass burning on ambient PM concentrations in an urban environment, the chemical characterization and mass closure of $PM_{2.5}$ was also investigated in urban sites of Karachi, a financial and industrial metropolis, the most populous city in Pakistan and also one of the largest megacities in the world (Shahid et al., 2016). Besides the local emission, regional influence is also a contributor to air pollution as a part of the prevailing monsoon circulation. The regional transportation of air pollutants makes the air quality condition even worse in Pakistan (Rasheed et al., 2015). All above studies show that $PM_{2.5}$ mass concentration in almost all major urban areas of Pakistan exceeding Pakistan's National Environmental Quality Standards both in long-term (annual average concentration of 25 $\mu g/m^3$) and in short-term (24 hourly average concentration of 40 $\mu g/m^3$) for ambient air quality (Ghauri et al., 2007).

As a result, the health burdens caused by environmental exposure to $PM_{2.5}$ in Pakistan have also drawn increasing attention. Health effect of $PM_{2.5}$ on daily morbidity in the previously mentioned megacity - Karachi was evaluated by collating on-site monitoring data with register-based data (i.e. daily hospital admissions and emergency room visit count) of cardiovascular disease data. The results indicate that morbidity of cardiovascular disease is strongly associated with the high concentrations > 150 $\mu g/m^3$, a level which is almost 4 times higher than the local 24 h air quality standard and 5 times higher than the WHO guideline. Specifically, this poor air quality commonly appears in many urban sites in Pakistan (Khwaja et al., 2012). Using a dataset of national mortality rate, health risk assessment has also been conducted in Islamabad, the capital city of Pakistan (Mehmood et al., 2018). Excessive mortality due to the environmental exposure of $PM_{2.5}$ was observed.

Although the air quality in Pakistan is severe and has caused considerable negative health effects, the majority of the citizens in Pakistan still do not have enough consciousness of self-protection from the toxic smog due to the lack of social awareness. More seriously, despite that there are already many efforts have been made to investigate the poor air quality issue, a usable and reliable reference for environmental exposure to air pollution is currently still not available for either the public health professions or the general public in Pakistan for further actions in the reduction of health risks caused by air pollution. In addition, the influential factor of regional air pollution across South Asian developing countries such as Pakistan is still underestimated.

### 1.2. Research Objectives

Under the above background, the main objective of the present study is to provide a national spatial estimation of the ambient $PM_{2.5}$ exposure in Pakistan in a relatively fine-scale, based on the re-adjustment and refinement of existing datasets (e.g. global satellite-derived air quality datasets, local monitoring data). Based on the resultant model, influential factors of the national-scale spatial variation of $PM_{2.5}$ will also be identified and quantified. Additionally, a series of spatial maps of ambient $PM_{2.5}$ concentration level matching local scenario will be generated based on the resultant LUR model, which is useful to the national health risk management and environmental planning. This procedure of refining global air quality data from LUR could also contribute to the air quality management and pollution reduction actions of Pakistan.

## 2. Material and methods

### 2.1. Analytical method - Land Use Regression (LUR)

Land Use Regression (LUR) was applied to estimate the spatiotemporal variations of ambient $PM_{2.5}$ exposure from the refinement of the global dataset in this study. LUR is a promising geospatial data integration technique that has been widely applied to estimate pollution surfaces for environmental exposure assessment usually at the urban scale but increasingly also at larger scales (Eeftens et al., 2012; Ross et al., 2006; van Nunen et al., 2017). This data-driven analytic method can estimate spatial distributions of air pollution concentration with high-resolution and high accuracy (Knibbs et al., 2014) based on an aggregation of multiscale datasets generated by spatial buffering. Currently, most of the previous LUR models with fine-scale outputs are still developed at the city-scale (Larson et al., 2009; Rivera et al., 2012). Recent studies have applied LUR to predict global $PM_{2.5}$ exposure (Donkelaar et al., 2010; van Donkelaar et al., 2016). However, the results are either in a coarse-scale, or either may not be representative of a specific country because of the selection of global monitoring stations for air quality mapping.

As national environmental assessment and public health management have been given a higher priority, spatial prediction of regional-scale air pollution becomes critically important, especially for those South Asian developing countries that are experiencing severe air pollution issues. In Pakistan, despite the efforts being made by different sectors of the society, there are still both a lack of well-developed and coordinated network of surface measurements of $PM_{2.5}$ and lack of the precise spatially-resolved emission inventory and other geospatial datasets that can be directly used for the estimation of ground-level $PM_{2.5}$ exposure. In that case, by persevering with the limited data and make the best use of open-source datasets that are currently available, this study, for the first time, applied the LUR modelling technique to provide a national-scale spatial estimation of the ambient $PM_{2.5}$ exposure for Pakistan. Specifically, using the ground-level $PM_{2.5}$ monitoring data from the Pakistan Air Quality Initiative (PAQI پاکی) as the dependent variable, and combing with a comprehensive set of predictor variables: meteorological and sounding data in regional basis, land use/land cover products derived from high-resolution satellite datasets, and satellite-derived Aerosol Optical Depth (AOD) based-dataset, this LUR study aims to provide a spatiotemporal $PM_{2.5}$ exposure estimation result with a fit-for-purpose accuracy.

## 2.2. Summary of datasets

Ground-level daily average $PM_{2.5}$ monitoring data at 15 air quality monitoring stations were used as the dependent variable of LUR modelling. Considering that the dependent variable dataset contains both spatial and temporal variability, the following multiple sets of different data sets are used as independent variables: spatial datasets including the land use/land cover, geographical features, and AOD data derived from satellite observation; temporal-resolved datasets including meteorological data and sounding data.

### 2.2.1. Ground-level $PM_{2.5}$ observation – the dependent variable

Currently, there is a lack of well-developed and coordinated network of surface measurements of $PM_{2.5}$ to act as the dependent variable in the statistical analysis. Therefore, in this study, we acquired $PM_{2.5}$ monitoring data from PAQI. PAQI provides community-driven air quality data to increase social awareness. As reference-standard air quality data is not available for Pakistan, PAQI has crowd-sourced air quality data using a nationwide network of low-cost air quality monitors. These monitors are the proven IQAir AirVisual Pro air quality monitors, and have been functional across Pakistan since 2016. The IQAir Air Visual Pro is a standalone device measuring fine particles $PM_{2.5}$ and $PM_{10}$, $CO_2$, temperature, and relative humidity. It uses a propriety $PM_{2.5}$ sensor (AVPM25b) based on the nephelometer light-scattering principle to measure particulate matter, and are calibrated from 0.3 to 2.5 µm. It also uses the A33 quad-core Cortex microprocessor, has its own internal data logging function (4 GB flash storage), and communicates through Wifi. The data can be downloaded as a .csv file using the SMB protocol or published live in an Airvisual cloud. The sampling interval is 10 seconds. A laboratory evaluation by the AQ-SPEC (Air Quality Performance Evaluation Center) can be found at http://www.aqmd.gov/aq-spec/sensordetail/iqair---airvisual-pro. The measurements are done in continuous real-time, though for analytical purposes hourly-average data is utilized.

For the present study, hourly observations of ground-level $PM_{2.5}$ concentration monitored between October 2016 and September 2018 at a total of 15 air quality monitoring stations in Pakistan (shown in Figure 1), distributed in seven major cities, which are Bahawalpur, Faisalabad, Islamabad, Karachi, Lahore, Peshawar, and Rawalpindi. Noted that the time periods of available observations are slightly varying between the 15 stations) is acquired. The daily average of $PM_{2.5}$ concentration was calculated and used as the dependent variable of LUR modelling in order to be collated with the temporal resolution of sounding data. The log-transformation was performed for the $PM_{2.5}$ concentration observation data, as the daily averaged $PM_{2.5}$ data does not have a normal distribution, which is similar to some representative previous LUR studies (Eeftens et al., 2012).

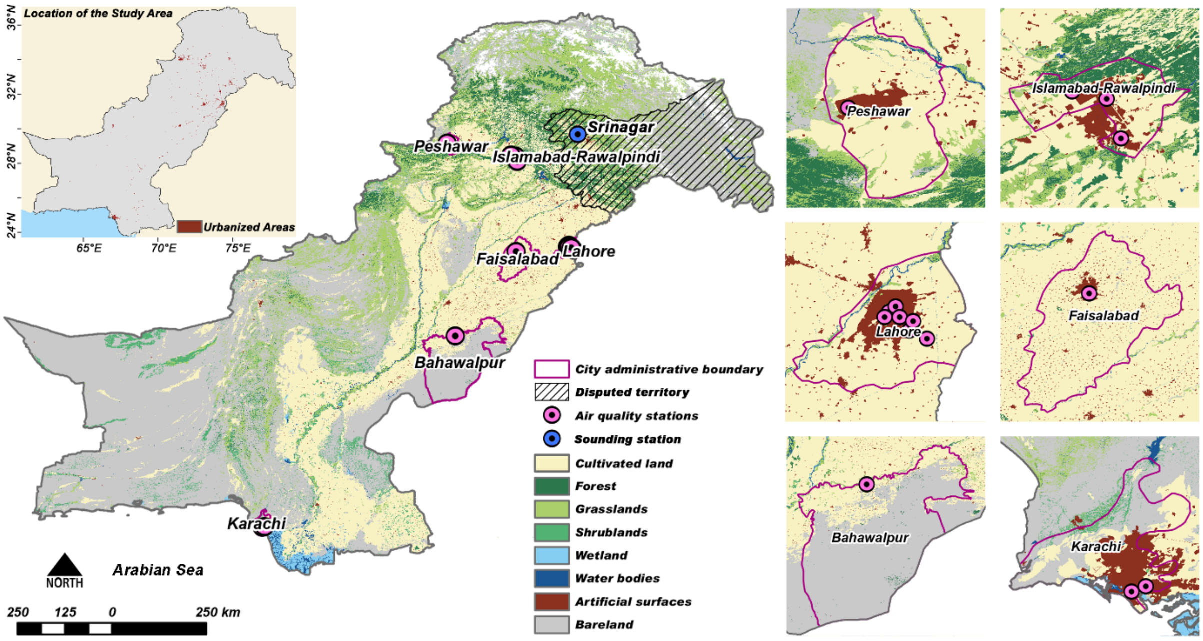



***Figure 1.*** *The locations of the 15 air quality monitoring stations in Pakistan available for this study.*

### 2.2.2. Meteorological data and sounding data

Regional climatic condition is an influencing factor of air quality (Crumeyrolle et al., 2014). It alters the atmospheric condition as such affects the movement and spatial distribution of air pollutants. For example, in Pakistan, a significant relationship has been found between $PM_{2.5}$ and temperature by a previous investigation of urban air pollutants emission patterns in the city Lahore (Haider et al., 2017). It has been investigated that the variation in atmospheric stability also strongly affects the vertical distribution of aerosol (Lee et al., 2011). Therefore, in this study, both commonly-used meteorological data (air temperature and relative humidity monitored at the same locations of the air quality monitoring stations, daily averages were calculated), and a group of atmospheric sounding indexes were used as candidate predictor variables (Table S-1, supplementary material**Error! Reference source not found.**). The sounding data used in this study is provided by the Department of Atmospheric Science, the University of Wyoming at their website: http://weather.uwyo.edu/upperair/sounding.html (which provides all relevant information about the sounding data). The sounding station is located in Srinagar, a capital city of Jammu and Kashmir (Region: Southeast Asia, Station Number: 42027). The station provides sounding data for every 24 hours at the time of hour 0000 (00Z) Greenwich Mean Time (GMT/ UTC) which is the local time at 5:30 am. This station is selected to represent an overall atmospheric condition for each day as it is relatively close to five cities with available air quality monitoring stations (Faisalabad, Islamabad, Lahore, Peshawar, and Rawalpindi). The location of the sounding station is labeled in Figure 1.

#### 2.2.3. Land use/land cover and geographical features

Currently, there is also a lack of precise spatially-resolved emission inventory and other geospatial datasets that can be tested for significance as predictor independent variables. It has been observed that the changes in spatial land use and land cover (LU/LC) also strongly affects the regional climatic condition of the urbanized area in Pakistan (Arshad et al., 2019). Therefore, LU/LC was also quantitatively measured by combining various of remote sensing data sources and open map service, which are High-resolution Multi-temporal Mapping of Global Urban Land 2015 (Liu et al., 2018) and GlobeLand30 (GLC30) (Jun et al., 2014) and OpenStreetMap (OSM). Specifically, the fraction of impervious surface was calculated based on Global Urban Land 2015; tree coverage ratio and water coverage ratio were calculated based on GlobeLand30. A buffering method that has been commonly-adopted by LUR studies (Hoek et al., 2008; Ryan and LeMasters, 2007) was used to analyze the land use of the study area. The land use area of commercial, industrial, residential, and retail land use within a series of round buffers (see section 2.3 and Table S-1, supplementary material) was calculated based on OSM. Same with all previous LUR studies, population density is also used, as human activities are the most direct source of air pollution. The geographical location (longitude, latitude, and elevation) of air quality monitoring stations are also adopted as candidate predictors, as the regional transportation affects the air quality in Pakistan

(Rasheed et al., 2015). Transport is one of the major emission sources of PM. Traffic-related PM emission is often estimated based on national emission inventories. However, a previous study (Ilyas, 2007) has indicated that the emission inventory of Pakistan is not a reliable source because it only takes vehicular exhausts into account, and it cannot represent either the actual vehicle populations or the in-use conditions, due to the inadequate emission factors. Besides, it has been found that there are discrepancies between emission inventories provided by different organizations (Ilyas, 2007). In that case, a commonly-used alternative in many LUR studies – the road network was adopted to represent the spatial emission of transport. Road length within the buffer area of air quality stations was used as the measure.

#### 2.2.4. Satellite-observation based dataset

The air quality monitoring stations available for this study are relatively limited in terms of the amount and sparsely distributed in the spatial context. This fact possibly introduces large uncertainties in the $PM_{2.5}$ estimation for unmonitored regions. The satellite-derived Aerosol Optical Depth (AOD) has been a popularly-used input to overcome this issue and provide robust spatial estimation due to its advantage of spatial coverage (Chu and Bilal, 2019). Incorporating satellite data also enables the consideration the regional impacts of $PM_{2.5}$ emitted and transported from the outside of Pakistan (i.e. the impact of long-range transport from the emission sources in neighboring countries usually does not fully reflect in national geographical dataset, which can still be captured by satellite in AOD dataset). Therefore, the Global Annual $PM_{2.5}$ Grids from MODIS, MISR, and SeaWiFS AOD with GWR, v1 (van Donkelaar et al., 2018) were also used as candidate predictor variables.

### 2.3. Spatial buffer scheme

To be consistent with other existing LUR models, based on literature (Knibbs et al., 2014; Knibbs et al., 2018; Novotny et al., 2011), a total of 22-circular buffers was generated with radii of 100 m, 200 m, 300 m, 400 m, 500 m, 600 m, 700 m, 800 m, 1000 m, 1200 m, 1500 m, 1800 m, 2000 m, 2500 m, 3000 m, 3500 m, 4000 m, 5000 m, 6000 m, 7000 m, 8000 m, and 10000 m. All predictor variables are summarized in Table S-1 (in the supplementary material).

### 2.4. Model development and validation

In this study, the LUR model was developed for the log-transformed $PM_{2.5}$ concentration using linear regression approaches which is similar to many previous LUR studies. There is a total of 324 potential predictor variables in this study, which is a relatively large number of potential predictor variables that need to be examined. "A Distance Decay REgression Selection Strategy (ADDRESS)" - a systematic method has been developed for optimizing the variable selection process for LUR modeling (Su et al., 2009). The study introduces the process of constructing distance-decay curves and the criteria for identifying optimized buffer distance for spatial covariates for LUR modelling. By referencing this method of constructing distance-decay curves, in our study, we performed a pre-selection process. Briefly speaking, the correlation between log-transformed $PM_{2.5}$ concentrations against each of the buffer-based spatial covariate was calculated and used to construct distance-decay curves. Typically, it is expected that the variable has the highest correlation to $PM_{2.5}$ concentrations at its optimized buffer radii. To avoid the correlation overestimation caused by the possible over-aggregation of the spatial data, the distance at the largest slope change should be chosen as the optimized buffer radii on curves that continue to rise with buffer distance. Multiple optimized buffer distances were chosen if multiple peaks appear in the curve. This condition indicates that the spatial covariates possibly influences the pollution level at different spatial scales. Performing this step allows us to find the optimized buffer distance of a spatial variable. The process was repeatedly performed for all the distance decay curves such that a sub-group of variables was chosen.

To build up the LUR model, the chosen sub-group of variables resulted from the previous variable selection process, together with other point-based spatial variables and temporal variables (meteorological data and sounding data) were used as the input of a forward-direction stepwise regression using the minimum Bayesian information criterion (BIC) criteria. Collinearity between variables was checked with the variance inflation factor (VIF). Variables were excluded if it significantly collinear with existing predictors (evaluated as VIF $\geq 2$) or the p-values exceeded $\alpha = 0.05$. The adjusted $R^2$ was used to represent the explained variance by the LUR model.

For the validation of the resultant model, we conducted the 10-fold cross-validation (10-fold CV) to compare the difference between the monitored and the estimated concentration. The root-mean-square error ($RMSE$) and the $R^2$ of 10-fold cross validation (10-fold CV $R^2$) (Burman, 1989) were used to validate the resultant LUR models. All response data were randomly divided into ten subsets, with nine subsets used as the training dataset and the other one subset used as validation datasets. This process was repeated ten times until all data have been used as validation data once.

# 3. Results

## 3.1. Variable pre-selection

By constructing distance-decay curves, all spatial covariates were examined to explore their relationship with pollution levels. It was found that there are four spatial covariates have clear relationships with the $PM_{2.5}$ concentration level: tree coverage ratio (TREE), water coverage ratio (WATER), the length of motorway and trunk roads (RDTRU), and the length of tertiary roads (RDTER). TREE has negative correlations with $PM_{2.5}$ level at buffer distance ranges from 400 m to 10000 m. Tree coverage ratio within the buffer distance of 100 m, 200 m, 300 m are found to be zero for all monitoring locations. Similarly, WATER is negatively correlated $PM_{2.5}$ level at all buffers between the distance from 1000 m to 10000 m. There is no water coverage was found within the buffer distance < 800 m of all monitoring locations. The rest of spatial covariates were not chosen as input predictor variables for stepwise regression because there are no clear patterns found in their correlations with pollution level (the correlation changed between positive and negative with the increase of buffer distance). It is noticed that most of the spatial covariates that do not have clear correlations with the pollution level are OSM polygon layer-based covariates. For example, the industrial land use area extracted from OSM does not have a significant positive correlation with the pollution level, which is unexpected. A possible explanation of the abnormal findings is the incompleteness of the OSM data (Haklay, 2010) which introduces errors in the spatial predictor variable data. Figure 2 illustrates the distance-decay curves of the four chosen spatial covariates.

Table ***1*** summarizes all chosen predictor variables used as the input of stepwise modeling (spatial covariates and optimized buffer distance).

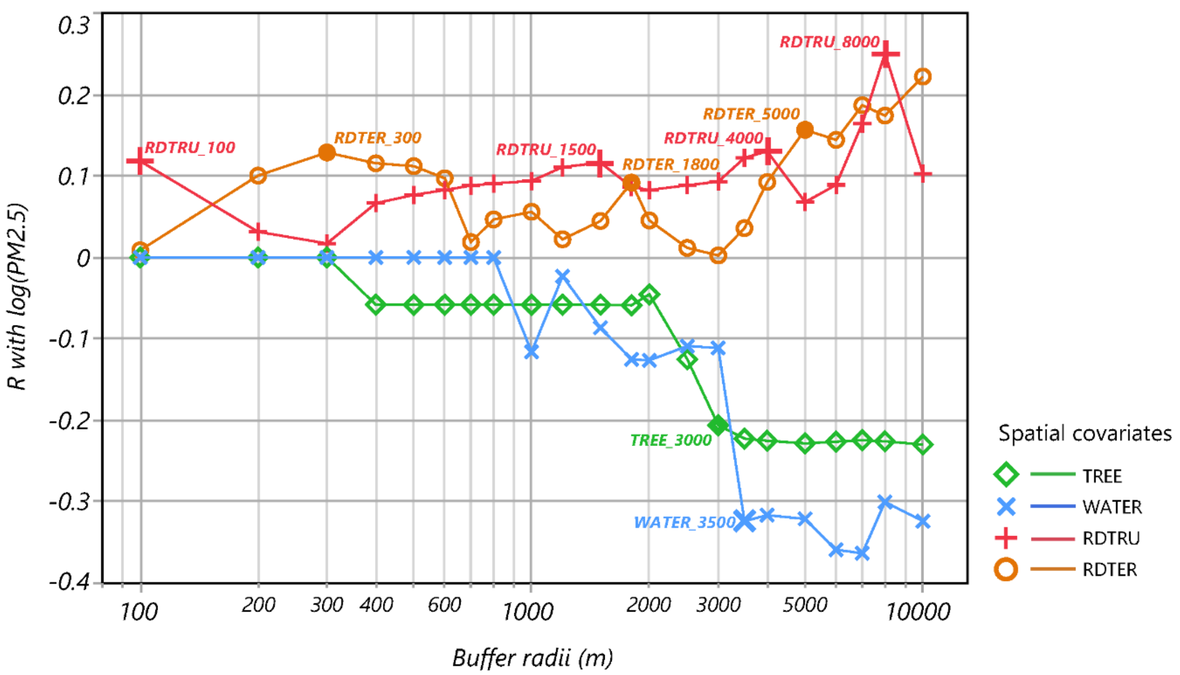



***Figure 2.*** *The distance-decay curves of the four chosen spatial covariates.*

***Table 1.*** *Summary of all chosen predictor variables (spatial covariates and their corresponding optimized buffer distance) used as the input of stepwise modelling.*

| **Spatial covariates** | **Chosen buffer radii (m)** | **R with log($PM_{2.5}$)** | **$R^2$ with log($PM_{2.5}$)** |
|---|---:|---:|---:|
| TREE | 3000 | -0.206 | 0.043 |
| WATER | 3500 | -0.324 | 0.105 |
| RDTRU | 100 | 0.118 | 0.014 |
| RDTRU | 1500 | 0.116 | 0.014 |
| RDTRU | 4000 | 0.130 | 0.017 |
| RDTRU | 8000 | 0.251 | 0.063 |
| RDTER | 300 | 0.129 | 0.017 |
| RDTER | 1800 | 0.092 | 0.008 |
| RDTER | 5000 | 0.157 | 0.025 |

### 3.2. Resultant LUR model

The resultant LUR model includes five predictor variables: annual mean daily air temperature (TEMP), lifted index (LIFT), annual mean $PM_{2.5}$ estimation with GWR gridded at 0.01° (PMGWR), tree coverage ratio within the buffer radii of 3000 m (TREE_3000), and the total length of motorway and trunk roads within the buffer radii of 1500 m (RDTRU_1500). The resultant model explains 54.5% of the variance in the log-transformed $PM_{2.5}$ concentration (adjusted $R^2 = 0.545$) and also has a reasonable 10-fold CV $R^2$ of 0.542. The two values are quite close, which indicates that the resultant model is not a statistical coincidence and provides an unbiased estimation. Figure 3 and Table 2 show the model performance and summaries all predictor variables that included by the resultant LUR model.

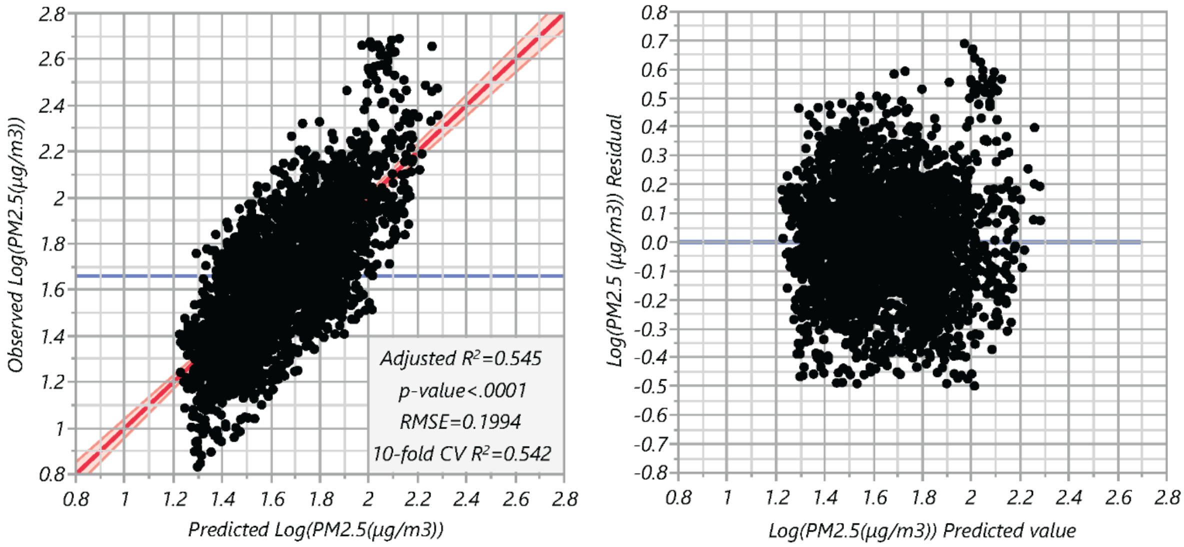



***Figure 3.*** *Observed log-transformed $PM_{2.5}$ concentration on predicted value and residual (left: actual by predicted plot; right: residual by predicted plot).*

***Table 2.*** *Summary of all included predictor variables of the resultant LUR model of log-transformed $PM_{2.5}$ concentration.*

| Predictor Variables | Unstandardized coefficients | Std Error | t Ratio | Significance level | Lower 95% | Upper 95% | VIF |
|---|---|---|---|---|---|---|---|
| TEMP | -5.42E-03 | 1.02E-03 | -5.32E+00 | <.0001 | -7.42E-03 | -3.42E-03 | 1.52E+00 |
| LIFT | 2.97E-02 | 1.30E-03 | 2.29E+01 | <.0001 | 2.72E-02 | 3.22E-02 | 1.51E+00 |
| PMGWR | 9.29E-03 | 2.17E-04 | 4.29E+01 | <.0001 | 8.86E-03 | 9.71E-03 | 1.06E+00 |
| TREE_3000 | -2.95E+00 | 1.94E-01 | -1.53E+01 | <.0001 | -3.33E+00 | -2.57E+00 | 1.02E+00 |
| RDTRU_1500 | 5.25E-03 | 1.02E-03 | 5.14E+00 | <.0001 | 3.25E-03 | 7.25E-03 | 1.06E+00 |
| Model intercept | 1.40E+00 | 3.11E-02 | 4.51E+01 | <.0001 | 1.34E+00 | 1.46E+00 | n/a |

### 3.3. Interpretation of resultant model

LIFT was included in the resultant LUR model and positively correlated with the log-transformed $PM_{2.5}$ concentration level. The lifted index (LIFT) is a simple and straightforward sounding index which has been commonly used for estimating the atmospheric stability (Galway, 1956). Stable atmospheric conditions lead to the accumulation of air pollutants and build up poor air quality scenarios. Therefore, this inclusion is reasonable because a higher value of LIFT indicates higher atmospheric stability as such indicates a more severe air pollution scenario. Air temperature which represents seasonal changes is also included by the model and negatively correlated with $PM_{2.5}$ level. This is consistent with the previous findings that $PM_{2.5}$ mass concentration level is negatively correlated with air temperature (Rasheed et al., 2015). This also reflects the fact that local emission sources become dominant to the air quality in Islamabad, Rawalpindi, and Lahore (the three cities locate at the eastern side of Pakistan territory and contain the majority of the monitoring stations) during winter time due to the reversion of the monsoon flow (Rasheed et al., 2014).

The total length of motorway and trunk roads within the buffer radii of 1500 m (RDTRU_1500) is included by the resultant model, but other traffic-related variables with a much larger buffer (i.e. RDTRU_4000, RDTRU_8000, RDTER_5000) are excluded by the stepwise regression in spite of their higher correlation coefficient with $PM_{2.5}$ level in the distance decay curves. This is consistent with the finding in previous LUR studies that the maximum influenced buffer distance for traffic-related covariates is 1500 m (Henderson et al., 2007; Jerrett et al., 2004). The inclusion of this traffic-related predictor variable also verifies the previous statement that $PM_{2.5}$ is clearly associated with the traffic-related pollutants emissions, and road transport is a major emission source of PM and a major influential factor of the air quality in Pakistan (Ilyas, 2007; Rasheed et al., 2015). It has been reported that the lack of vegetation has been held responsible for the poor urban air quality in Pakistan (Colbeck et al., 2010), which supports the inclusion of the tree coverage ratio within the buffer radii of 3000 m (TREE_3000) as a predictor variable.

Figure 4 shows the spatial estimation of the annual mean $PM_{2.5}$ concentration level in 2016 based on the resultant LUR model. The spatial data of annual mean air temperature of 2016 used for the mapping is extracted from a recently released global dataset of air temperature derived from satellite remote sensing and weather stations (Hooker et al., 2018). By overlapping the $PM_{2.5}$ spatial estimation with population distribution data of Pakistan, the previous finding that all population of Pakistan is living in the condition with $PM_{2.5}$ concentration levels higher than the standard value of WHO guideline (Cohen et al., 2017) was verified. The pollution level in city Lahore and Faisalabad were found to be more severe which possibly because of the combination of the lack of large-scale vegetation coverage and the regional influence of long-range transport of $PM_{2.5}$ from emission sources outside the country. These results are supported by previous studies which showed that biomass burning in the neighboring region significantly affected the air quality conditions over Lahore and Faisalabad and increased the $PM_{2.5}$ level (Khokhar et al., 2016; Tariq et al., 2015). High $PM_{2.5}$ concentration levels > 150 μg/m$^3$ which has been correlated with many serious health burdens were also observed.

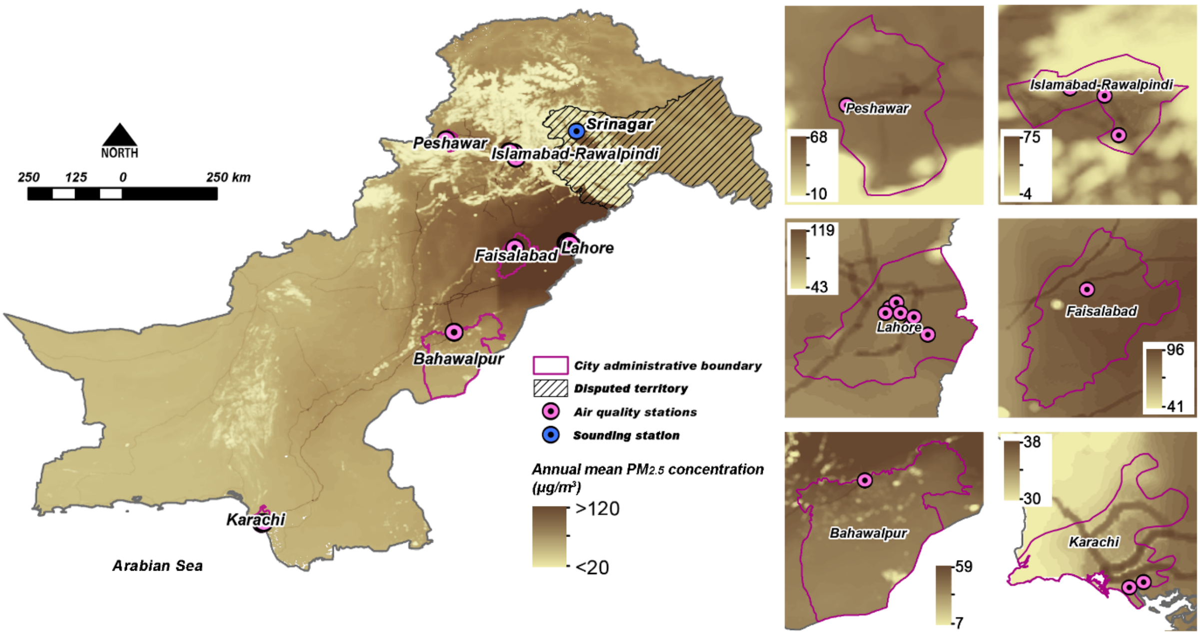



***Figure 4.*** *Spatial estimation of the annual mean $PM_{2.5}$ concentration level in 2016 based on the resultant LUR model. The color bar in both the national map and the maps of each city are optimized separately for better visualization of the spatial variability.*

### 3.4. Observing seasonal variation from the resultant model

Noticeable seasonal changes in $PM_{2.5}$ concentration levels have been clearly observed in Pakistan (Javed et al., 2015; Mehmood et al., 2018), which is mainly caused by seasonal changes in meteorological conditions. Pakistan, particularly the northeastern inland part of the nation's territory, has four seasons, which are the warm and rainy summer (June to August), dry autumn (September to November), cold and dry winter (December to February) and spring (March to May). The seasonal variation in the coastal area (the southwestern side of the country) is slightly different: winter (January to March), pre-monsoon (April to June), monsoon (July to September), and post-monsoon (October to December) (Khan, 1991). As a fundamental part of the seasonal changes in meteorological conditions, the monsoon reversion is clearly reflected in seasonal alternation. As predictor variables in the resultant model, both the air temperature (TEMP) and the sounding index lifted index (LIFT) directly reflect the seasonal alternations of meteorological conditions. Therefore, the seasonal alternation and changes in meteorological conditions are already included by the resultant model. Spatial maps of the seasonal average of $PM_{2.5}$ concentration in Pakistan have also been produced to reflect the seasonal variation (Figure 5). The $PM_{2.5}$ concentration level seasonal difference in the seasonal maps is consistent with the observation in previous studies (Javed et al., 2015; Mehmood et al., 2018).

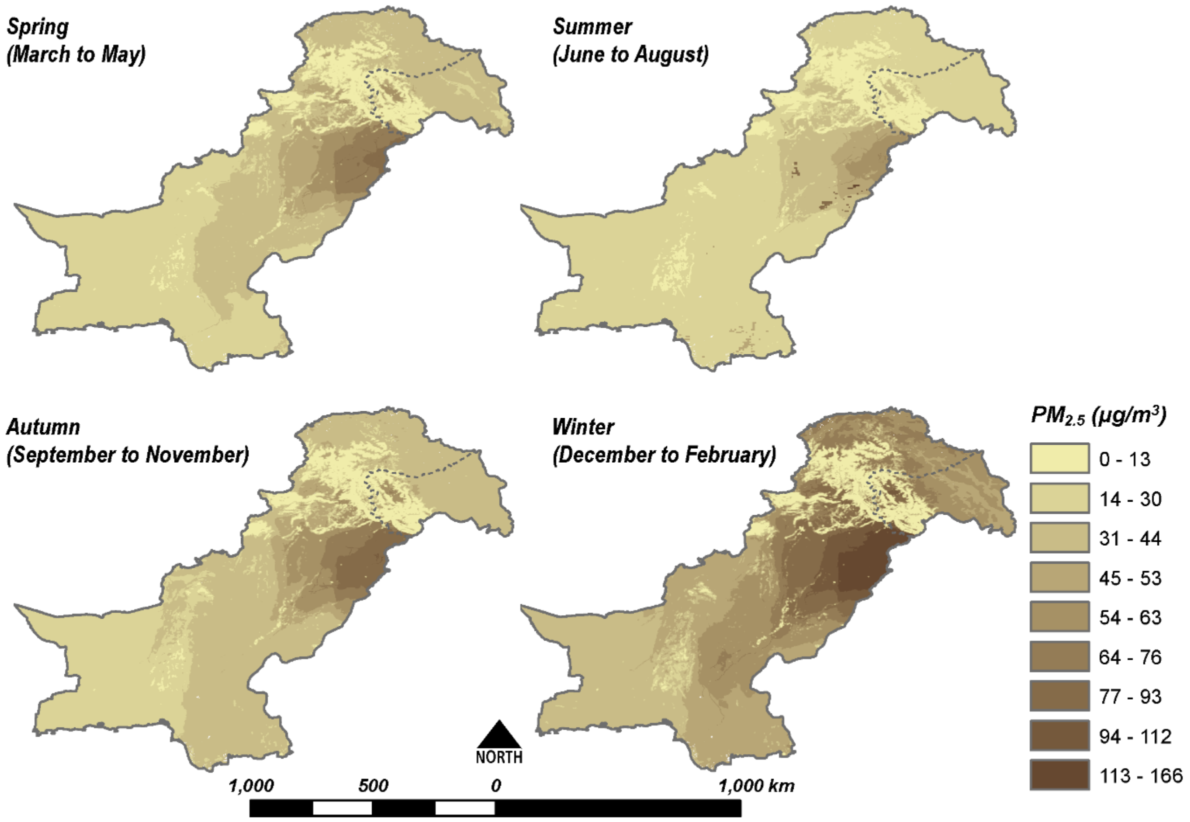



***Figure 5.*** *Spatial estimation of the seasonal mean $PM_{2.5}$ concentration level in 2016 based on the resultant LUR model. A unified color bar is used for all seasonal maps.*

## 4. Discussion

### 4.1. Refinement of the existing $PM_{2.5}$ dataset

There are two major refinements. First, the resultant land use regression model could be used for spatiotemporal estimation of $PM_{2.5}$ for a period-of-interest as long as the input data are available. The existing $PM_{2.5}$ spatial data is an annual average map. After the refinement, the spatial maps of $PM_{2.5}$ produced by this study can be temporal-resolved (Figure 5). Second, through the use of land use regression modelling and integration of local data (vegetation distribution, road network), the existing global annual average $PM_{2.5}$ dataset (which is based on geographical and climate space weighted regressions at a coarse spatial resolution) is refined to a scenario that more representative to the local condition of Pakistan.

From the viewpoint of the application in public health risk assessment, a comparison is made between the existing dataset and the refinement of $PM_{2.5}$ spatial estimation by using a 7-point bipolar scale (from high exposure to low exposure) and natural breaks classification. The $PM_{2.5}$ data is spatially aggregated based on administrative boundaries. As a result, each administrative zone has a score of the level of exposure (Figure 6). There are noticeable differences between the two datasets. Compared with the refined dataset, the existing dataset might overestimate the exposure level in the northern mountainous area and underestimates the exposure level in the southern coastal area of the country.

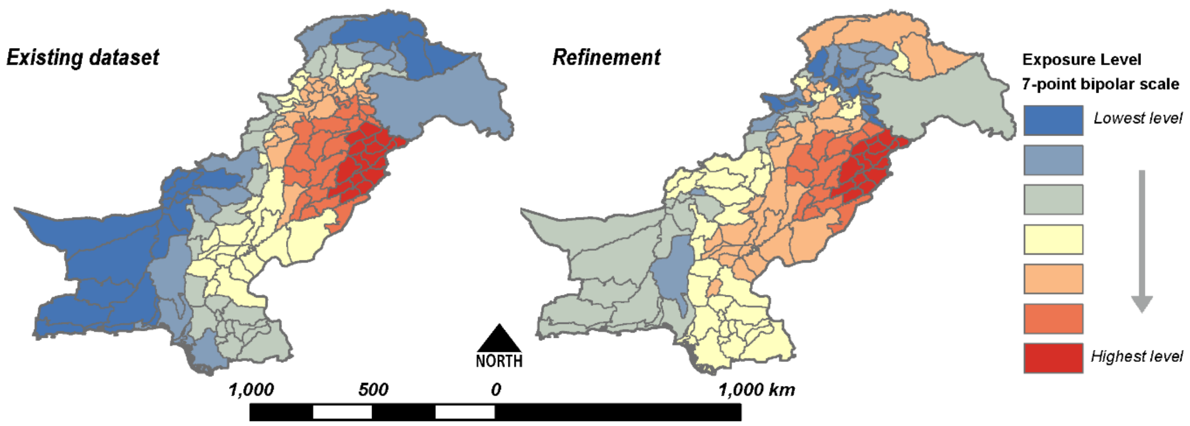



***Figure 6.*** *The classification map of the existing dataset and the refinement of $PM_{2.5}$ spatial estimation based on a 7-point bipolar scale using natural breaks classification.*

### 4.2. Implications of results on preferred socioeconomic development trajectories for air quality management of cities in Pakistan

Based on the results, this study implied that urbanization might be the key influential of air pollution across a South Asian developing country. Specifically, a mega network of motorway and trunk roads not only increased the number of vehicles but also can increase the frequency and duration of vehicles on the road network, because motorway and trunk roads are usually designed for long-distance traffic. More importantly, trunk road is a specific type of road for freight traffic, in which the major vehicle on the roads can be a heavy goods vehicle (HGV). It is known that HGV operated by heavy-duty diesel can release a large amount of $PM_{2.5}$, in which the expansion of trunk roads across the country has no doubt to be continuing to negatively influence the air quality in Pakistan. In details, as a developing country, the logistic network for freight traffic in Pakistan should be car-dependent but not flight-dependent, since road network can provide the most sustainable strategy to deliver goods from one city to another city. This transportation network is essential because there is a significant urban/rural difference in Pakistan. In order to reach both mega cities and small towns within the country, the road network has become more important than the past. Based on this, the use of long-distance vehicles especially HGV is expected to be increased in the future across this country, while this can further worsen the air quality in the urbanized area. In contrast, although the industrial sector has accounted for approximately 24% of GDP in Pakistan, it is dominated by the light industry (e.g. Cotton textile production). Therefore, the emission of air pollutants from factories in Pakistan may be relatively low compared to those factories for heavy industries such as chemical products and heavy metals. This may also somewhat explain why industrial lands may not be the contributors of $PM_{2.5}$ in this country since the negative effects of urbanization in Pakistan might be driven by the necessity of transportation within the country.

Therefore, the association between tree coverage and $PM_{2.5}$ in Pakistan can further be expressed as a consequence of increased impervious surfaces due to urbanization. Specifically, the development of motorway and trunk roads must have to interrupt the natural environment. This can induce deforestation across the country. This deforestation can be further enhanced because of the economic growth and energy consumption of the country (Ahmed et al., 2015). For example, biofuel burning is a known problem in Pakistan (Tahir et al., 2010), and informal mining has been threatening the natural environment in the country for decades (Lahiri-Dutt and Brown, 2017), not to mention that government-controlled mining sites can be found nationwide. Combining all these factors above, deforestation has become a great threat to air quality control due to a lack of natural greenery to reduce $PM_{2.5}$.

Based on these facts, policies for sustainable development should be established in this country for air quality improvement. Specifically, these policies should at least include the following three factors: 1) integration of urban design with greenery along with the road network, 2) afforestation of abandoned mining sites and 3) afforestation of sites after biofuel burning. These policies should be delivered in both top-down and bottom-up basis. Specifically, the bottom-up strategies can be a community engagement among local residents, non-governmental organization, and government sectors to establish afforestation programs together. Such approaches have somewhat taken been places in this country. For examples, the 2010 agreement of International Union for Conservation of Nature to raise a mangroves plantation of over 25 hectares along the Karachi coast, Pakistan; and the "Billion Tree Tsunami" project launched in 2014 to restore 350,000 hectares of forests and degraded land to surpass its Bonn Challenge commitment. Although action plans for afforestation across this country have been established, the magnitude for such actions still needs to be further increased.

Moreover, the study outputs could also contribute to national health risk management. The study outputs can be directly used in environment-related risk assessment in GIS. Taking the Crichton's conceptual definition of risk triangle as an example (Crichton, 1999), the Crichton's risk triangle transfers a risk into three dimensions, which corresponds to three data layers in GIS: the hazard layer, the exposure layer, and the vulnerability layer. This study directly outputs the hazard layer.

### 4.3. Limitations and future works

In this study, there were only a few 15 ground-level air quality monitoring locations across the country that can be used to refine the global dataset to the local scenario. As these air quality monitoring locations are sparsely and unevenly distributed in the study area, this may still introduce bias and uncertainty in the spatial estimation. However, since this study is a refinement of the existing global $PM_{2.5}$ dataset by integrating local data, the bias should be lower than either directly using global datasets or predicting ambient $PM_{2.5}$ exposure by solely using air quality data from sparsely distributed monitoring locations. For future studies, besides the long-term monitoring data from fixed air quality monitoring network, data collected in short-term air quality sampling or mobile monitoring campaigns can also be used to substantially enrich the spatial coverage of ambient $PM_{2.5}$ observations (Brantley et al., 2014; van Nunen et al., 2017). However, since Pakistan is a developing country with a great urban/rural difference, nationwide mobile monitoring campaigns to measure air quality across the whole country would still be extremely difficult. In that case, our approach in this current study is appropriate to provide a cost-effective solution to deliver national-scale air quality outputs and exposure assessment by refining the global datasets.

Another limitation of the study is that only open map service was used for the extraction of detailed intraurban land utilization and road networks. It has been widely noticed that OSM has limitations in its spatial data completeness and positional accuracy (Haklay, 2010), particularly in polygon feature layers. Therefore, in our future works, more accurate geoinformation of the study area (for example, a spatial-resolved industrial emission inventory from local authorities) should be acquired and used for improving the estimation of spatial $PM_{2.5}$ distribution. It should also be noted that the resultant model only has a limited capacity to estimate the regional influence caused by long-range pollution transport. In the next step of the study, incorporating geoinformation of neighboring regions and mesoscale meteorological modelling results into LUR modelling would be helpful to enhance the prediction capacity of the resultant model. Last but not least, it should be noticed that, to help with the local government, there are much more coordination and communication need to be done despite that the study outputs have the potential of helping with the local government.

## 5. Conclusions

In this study, a spatiotemporal LUR model is developed to refine global air quality datasets to the local scenario of ambient $PM_{2.5}$ exposure in Pakistan. Multiple open-source and publicly available datasets were used for model development, which means that the model development process of the present LUR study could be transferred and adopted by other regions for the development of cross-comparable LUR models. The research findings also show that tree coverage and road transport are two influential factors of the national-scale spatial variation of $PM_{2.5}$. This finding implies that Pakistan's current efforts in environmental protection (i.e., the effort in reforestation and transport pollution emission reduction) are in a good direction and need to be continued. Based on the resultant LUR model, a spatial map of ambient $PM_{2.5}$ concentration level matching local scenario was generated, which could provide useful spatial information to the national health risk management and also has a great potential of helping local authorities on the air quality management, and contributing to the pollution reduction actions of Pakistan (Matthew, 2001; Qadir, 2002; Shaikh et al., 2016).

## Acknowledgements

This research is supported by the General Research Fund (GRF No.14610717) from the Research Grants Council (RGC) of Hong Kong and the Special Project of Jiangsu Distinguished Professor (1421061801003, 1421061901001) from the Jiangsu Provincial Department of Education, China. The authors wish to thanks the Department of Atmospheric

Science, the University of Wyoming, especially Dr. Larry Oolman, for providing the atmospheric sounding indices data. The authors appreciate reviewers for their insightful comments and constructive suggestions on our research work. The authors also want to thank editors for their patient and meticulous work for our manuscript.

## Reference


Ahmed, K., Shahbaz, M., Qasim, A., Long, W., 2015. The linkages between deforestation, energy and growth for environmental degradation in Pakistan. Ecological Indicators 49, 95-103.

Alam, K., Rahman, N., Khan, H.U., Haq, B.S., Rahman, S., 2015. Particulate matter and its source apportionment in Peshawar, Northern Pakistan. Aerosol Air Qual. Res 15, 634-647.

Arshad, A., Zhang, W., Zaman, M.A., Dilawar, A., Sajid, Z., 2019. Monitoring the impacts of spatio-temporal land-use changes on the regional climate of city Faisalabad, Pakistan. Annals of GIS 25, 57-70.

Brantley, H.L., Hagler, G.S.W., Kimbrough, E.S., Williams, R.W., Mukerjee, S., Neas, L.M., 2014. Mobile air monitoring data-processing strategies and effects on spatial air pollution trends. Atmos. Meas. Tech. 7, 2169-2183.

Burman, P., 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 76, 503-514.

Chu, H.-J., Bilal, M., 2019. PM2.5 mapping using integrated geographically temporally weighted regression (GTWR) and random sample consensus (RANSAC) models. Environ Sci Pollut Res 26, 1902-1910.

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., Pope, C.A., Shin, H., Straif, K., Shaddick, G., Thomas, M., van Dingenen, R., van Donkelaar, A., Vos, T., Murray, C.J.L., Forouzanfar, M.H., 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet 389, 1907-1918.

Colbeck, I., Nasir, Z.A., Ali, Z., 2010. The state of ambient air quality in Pakistan—a review. Environ Sci Pollut Res 17, 49-63.

Crichton, D., 1999. The risk triangle, in: Ingleton, J. (Ed.), Natural Disaster Management. Tudor Rose, London.

Crumeyrolle, S., Chen, G., Ziemba, L., Beyersdorf, A., Thornhill, L., Winstead, E., Moore, R., Shook, M., Hudgins, C., Anderson, B., 2014. Factors that influence surface PM 2.5 values inferred from satellite observations: perspective gained for the US Baltimore–Washington metropolitan area during DISCOVER-AQ. Atmospheric Chemistry and Physics 14, 2139-2153.

Davidson, C.I., Phalen, R.F., Solomon, P.A., 2005. Airborne Particulate Matter and Human Health: A Review. Aerosol Sci. Technol. 39, 737-749.

Donkelaar, A.v., Martin, R.V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., Villeneuve, P.J., 2010. Global Estimates of Ambient Fine Particulate Matter Concentrations from Satellite-Based Aerosol Optical Depth: Development and Application. Environ. Health Perspect. 118, 847-855.

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dedele, A., Dons, E., de Nazelle, A., 2012. Development of land use regression models for PM2. 5, PM2. 5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46, 11195-11205.

Galway, J.G., 1956. The Lifted Index as a Predictor of Latent Instability. Bulletin of the American Meteorological Society 37, 528-529.

Ghauri, B., Lodhi, A., Mansha, M., 2007. Development of baseline (air quality) data in Pakistan. Environ. Monit. Assess. 127, 237-252.

Haider, R., Yasar, A., Tabinda, A.B., 2017. Urban Emission Patterns at a Semi-Arid Site in Lahore, Pakistan. Polish Journal of Environmental Studies 26.

Haklay, M., 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. Environment and Planning B: Planning and Design 37, 682-703.

Henderson, S.B., Beckerman, B., Jerrett, M., Brauer, M., 2007. Application of Land Use Regression to Estimate Long-Term Concentrations of Traffic-Related Nitrogen Oxides and Fine Particulate Matter. Environ. Sci. Technol. 41, 2422-2428.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42, 7561-7578.

Hooker, J., Duveiller, G., Cescatti, A., 2018. A global dataset of air temperature derived from satellite remote sensing and weather stations. Scientific Data 5, 180246.

Ilyas, S.Z., 2007. A review of transport and urban air pollution in Pakistan. Journal of Applied Sciences and Environmental Management 11.

Javed, W., Wexler, A.S., Murtaza, G., Ahmad, H.R., Basra, S.M., 2015. Spatial, temporal and size distribution of particulate matter and its chemical constituents in Faisalabad, Pakistan. Atmósfera 28, 99-116.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., Giovis, C., 2004. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol 15, 185-204.

Jun, C., Ban, Y., Li, S., 2014. China: Open access to Earth land-cover map. Nature 514, 434.

Khan, F.K., 1991. A geography of Pakistan: environment, people and economy. Oxford University Press.

Khokhar, M.F., Yasmin, N., Chishtie, F., Shahid, I., 2016. Temporal Variability and Characterization of Aerosols across the Pakistan Region during the Winter Fog Periods. Atmosphere 7, 67.

Khwaja, H.A., Fatmi, Z., Malashock, D., Aminov, Z., Kazi, A., Siddique, A., Qureshi, J., Carpenter, D.O., 2012. Effect of air pollution on daily morbidity in Karachi, Pakistan. Journal of Local and Global Health Science 2012.

Knibbs, L.D., Hewson, M.G., Bechle, M.J., Marshall, J.D., Barnett, A.G., 2014. A national satellite-based land-use regression model for air pollution exposure assessment in Australia. Environ. Res. 135, 204-211.

Knibbs, L.D., van Donkelaar, A., Martin, R.V., Bechle, M.J., Brauer, M., Cohen, D.D., Cowie, C.T., Dirgawati, M., Guo, Y., Hanigan, I.C., Johnston, F.H., Marks, G.B., Marshall, J.D., Pereira, G., Jalaludin, B., Heyworth, J.S., Morgan, G.G., Barnett, A.G., 2018. Satellite-Based Land-Use Regression for Continental-Scale Long-Term Ambient PM2.5 Exposure Assessment in Australia. Environ. Sci. Technol. 52, 12445-12455.

Lahiri-Dutt, K., Brown, H., 2017. Governing the ungovernable? Reflections on informal gemstone mining in high-altitude borderlands of Gilgit-Baltistan, Pakistan. Local Environment 22, 1428-1443.

Larson, T., Henderson, S.B., Brauer, M., 2009. Mobile monitoring of particle light absorption coefficient in an urban area as a basis for land use regression. Environ. Sci. Technol. 43, 4672-4678.

Lavy, V., Ebenstein, A., Roth, S., 2014. The impact of short term exposure to ambient air pollution on cognitive performance and human capital formation. National Bureau of Economic Research.

Lee, H., Liu, Y., Coull, B., Schwartz, J., Koutrakis, P., 2011. A novel calibration approach of MODIS AOD data to predict PM2. 5 concentrations. Atmos. Chem. Phys 11, 7991-8002.

Liu, X., Hu, G., Chen, Y., Li, X., Xu, X., Li, S., Pei, F., Wang, S., 2018. High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google Earth Engine Platform. Remote Sensing of Environment 209, 227-239.

Matthew, R.A., 2001. Environmental stress and human security in Northern Pakistan. Environmental Change and Security Project Report 7, 17-31.

Mehmood, T., Tianle, Z., Ahmad, I., Li, X., Shen, F., Akram, W., Dong, L., 2018. Variations of PM2. 5, PM10 mass concentration and health assessment in Islamabad, Pakistan, IOP Conference Series: Earth and Environmental Science. IOP Publishing, p. 012031.

Novotny, E.V., Bechle, M.J., Millet, D.B., Marshall, J.D., 2011. National satellite-based land-use regression: NO2 in the United States. Environ. Sci. Technol. 45, 4407-4414.

Power, M.C., Weisskopf, M.G., Alexeeff, S.E., Coull, B.A., Spiro, A., Schwartz, J., 2011. Traffic-Related Air Pollution and Cognitive Function in a Cohort of Older Men. Environ. Health Perspect. 119, 682-687.

Qadir, N.F., 2002. Air quality management in Pakistani cities: Trends and challenges. Better Air Quality in Asian and Pacific Rim Cities, 16-18.

Rasheed, A., Aneja, V.P., Aiyyer, A., Rafique, U., 2014. Measurements and analysis of air quality in Islamabad, Pakistan. Earth's Future 2, 303-314.

Rasheed, A., Aneja, V.P., Aiyyer, A., Rafique, U., 2015. Measurement and analysis of fine particulate matter (PM2. 5) in urban areas of Pakistan. Aerosol Air Qual. Res 15, 426-439.

Rivera, M., Basagaña, X., Aguilera, I., Agis, D., Bouso, L., Foraster, M., Medina-Ramón, M., Pey, J., Künzli, N., Hoek, G., 2012. Spatial distribution of ultrafine particles in urban settings: A land use regression model. Atmos. Environ. 54, 657-666.

Ross, Z., English, P.B., Scalf, R., Gunier, R., Smorodinsky, S., Wall, S., Jerrett, M., 2006. Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. Journal of Exposure Science and Environmental Epidemiology 16, 106-114.

Russell, A.G., Brunekreef, B., 2009. A Focus on Particulate Matter and Health. Environ. Sci. Technol. 43, 4620-4625.

Ryan, P.H., LeMasters, G.K., 2007. A review of land-use regression models for characterizing intraurban air pollution exposure. Inhalation Toxicol. 19, 127-133.

Saleem, S., Sughra, T., 2018. Pakistan moves to curb urban air pollution after high court ruling. Thomson Reuters Foundation.

Shahid, I., Kistler, M., Mukhtar, A., Ghauri, B.M., Ramirez-Santa Cruz, C., Bauer, H., Puxbaum, H., 2016. Chemical characterization and mass closure of PM10 and PM2.5 at an urban site in Karachi – Pakistan. Atmos. Environ. 128, 114-123.

Shaikh, S.A., Hongbing, O., Nisar, M.S., Khan, K., 2016. Environmental Justice in Pakistan: Issues, Policies, and Solutions. New Horizons 10, 41.

Su, J.G., Jerrett, M., Beckerman, B., 2009. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. Sci. Total Environ. 407, 3890-3898.

Tahir, S.N.A., Rafique, M., Alaamer, A.S., 2010. Biomass fuel burning and its implications: Deforestation and greenhouse gases emissions in Pakistan. Environ. Pollut. 158, 2490-2495.

Tariq, S., ul-Haq, Z., Ali, M., 2015. Analysis of optical and physical properties of aerosols during crop residue burning event of October 2010 over Lahore, Pakistan. Atmospheric Pollution Research 6, 969-978.

UNEP, 2015. The Environment and Climate Change Outlook of Pakistan. United Nations Environment Programme.

van Donkelaar, A., Martin, R., Brauer, M., Hsu, N., Kahn, R., Levy, R., Lyapustin, A., Sayer, A., 2018. Global Annual PM2. 5 Grids from MODIS MISR and SeaWiFS Aerosol Optical Depth (AOD) with GWR 1998–2016. NASA Socioeconomic Data and Applications Center (SEDAC).

van Donkelaar, A., Martin, R.V., Brauer, M., Hsu, N.C., Kahn, R.A., Levy, R.C., Lyapustin, A., Sayer, A.M., winker, D.M., 2016. Global Estimates of Fine Particulate Matter Using a Combined Geophysical-Statistical Method with Information from Satellites. Environ. Sci. Technol. 50, 3762.

van Nunen, E., Vermeulen, R., Tsai, M.-Y., Probst-Hensch, N., Ineichen, A., Davey, M., Imboden, M., Ducret-Stich, R., Naccarati, A., Raffaele, D., Ranzi, A., Ivaldi, C., Galassi, C., Nieuwenhuijsen, M., Curto, A., Donaire-Gonzalez, D., Cirach, M., Chatzi, L., Kampouri, M., Vlaanderen, J., Meliefste, K., Buijtenhuijs, D., Brunekreef, B., Morley, D., Vineis, P., Gulliver, J., Hoek, G., 2017. Land Use Regression Models for Ultrafine Particles in Six European Areas. Environ. Sci. Technol. 51, 3336-3345.

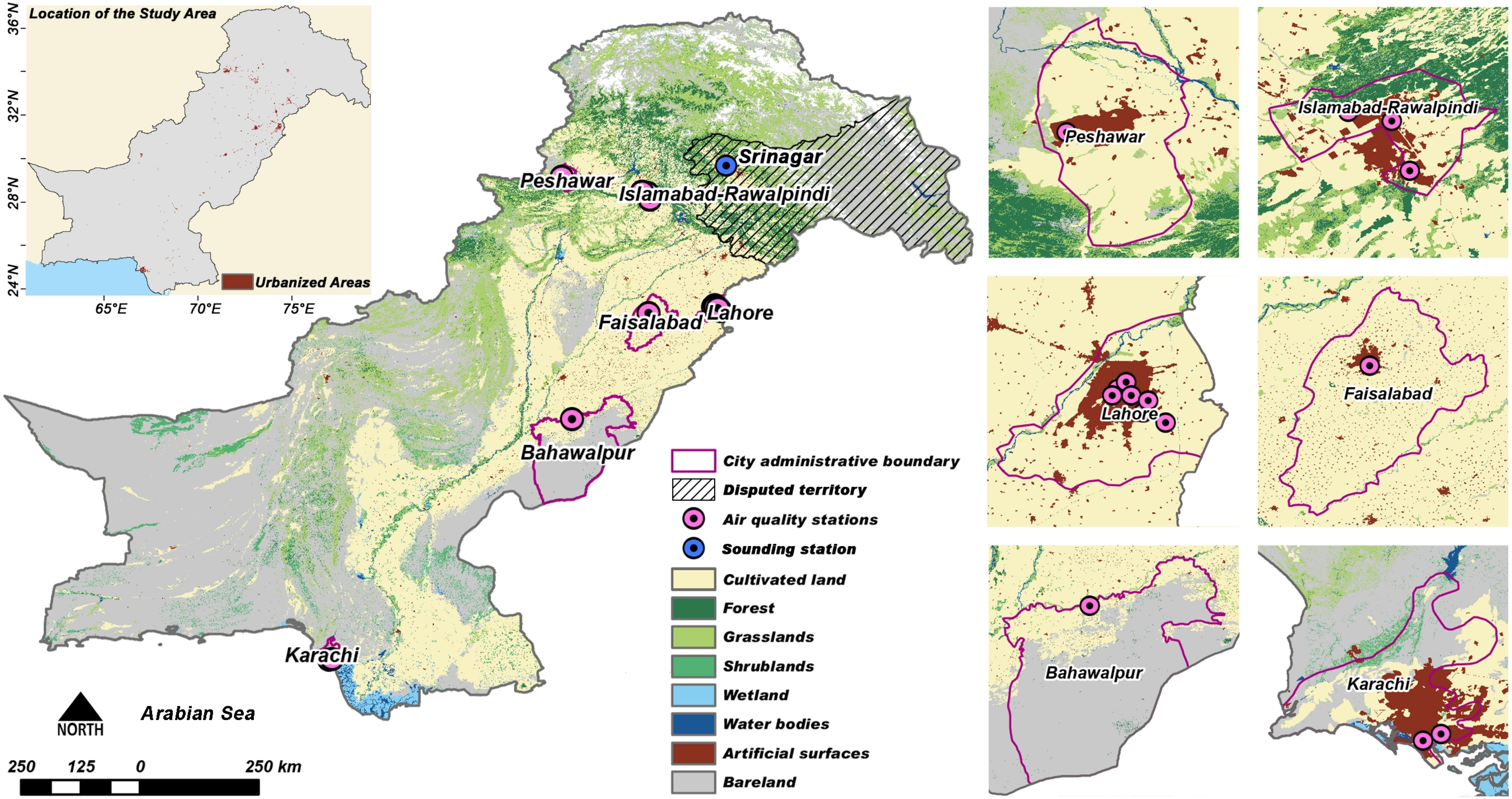
WHO, 2005. “WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen, Dioxide and Sulfur Dioxide. Global update 2005, Summary of Risk Assessment,”WHO/SDE/PHE/OEH/06.02. World Health Organization.

WHO, 2016. World health statistics 2016: monitoring health for the SDGs sustainable development goals. World Health Organization, Geneva, Switzerland.

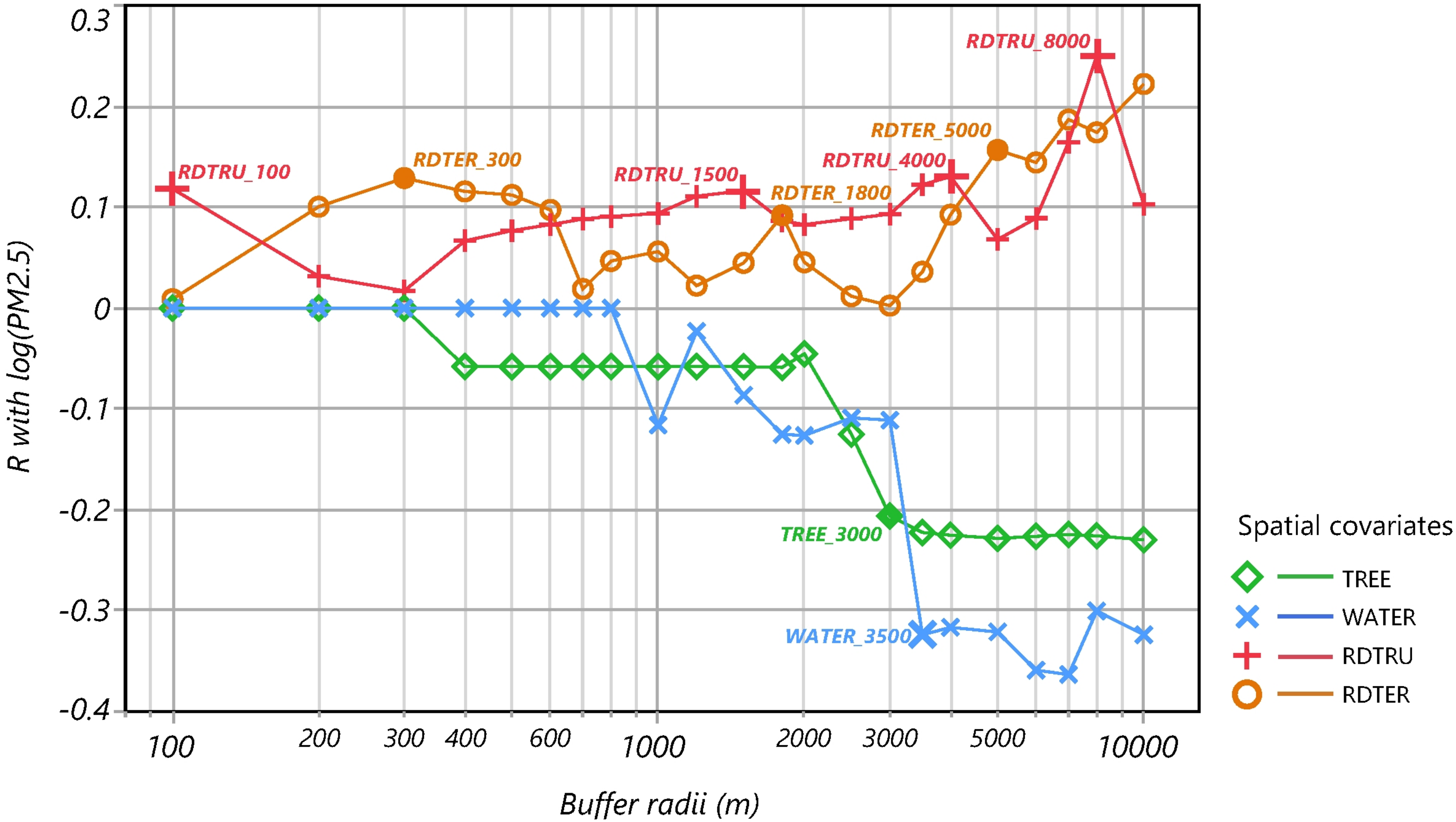
## List of Figures



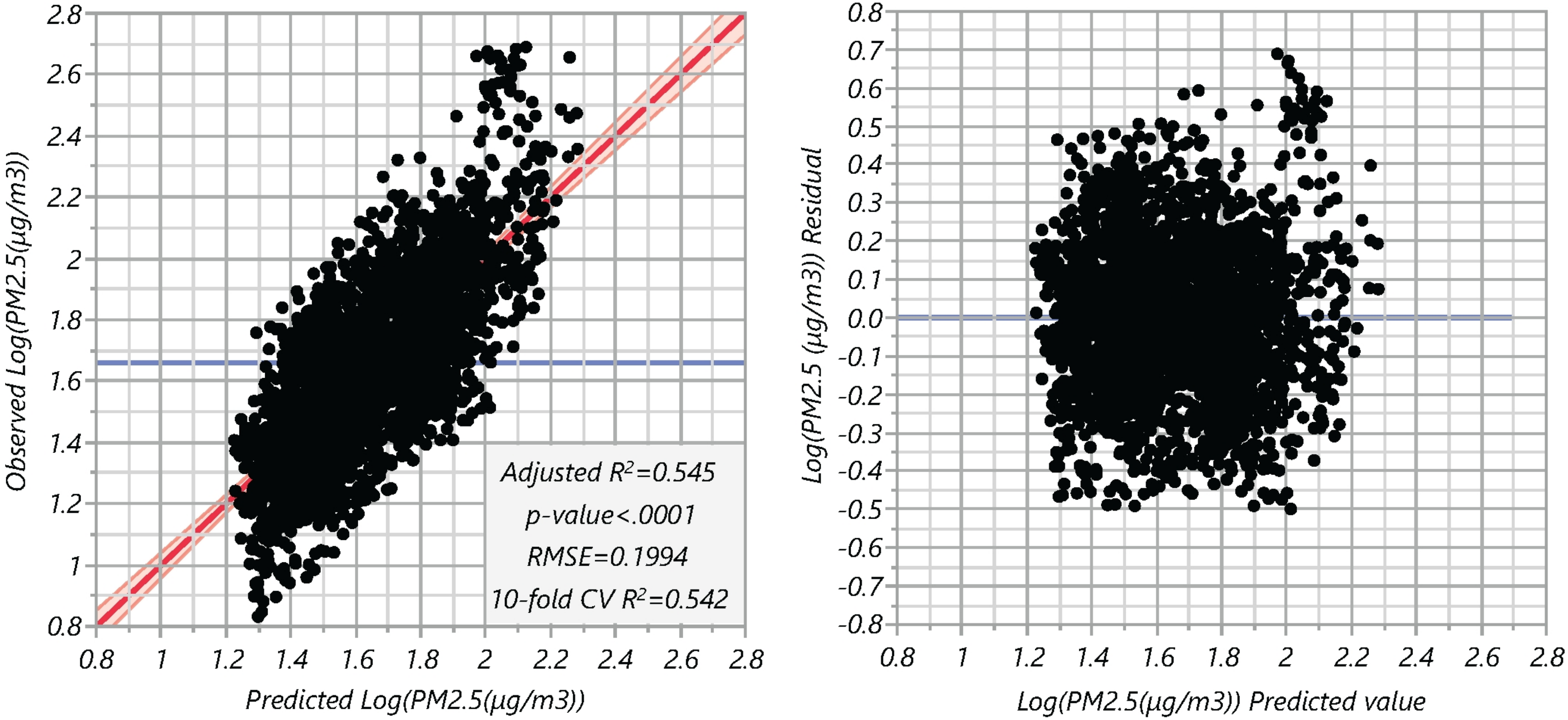
## List of Tables



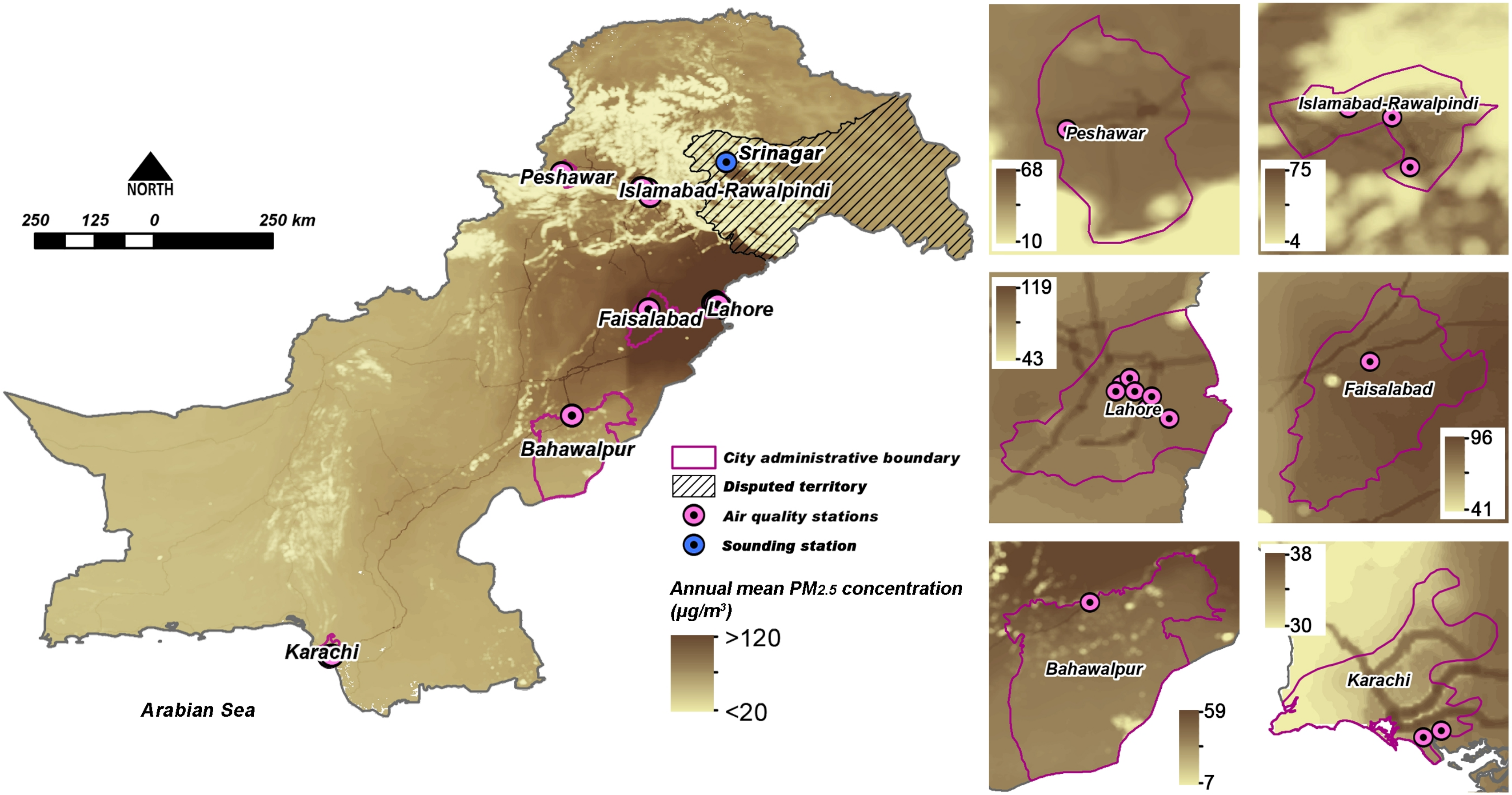
Location of the Study Area
36°N
32°N
28°N
24°N
65°E
70°E
75°E
Urbanized Areas
Peshawar
Islamabad-Rawalpindi
Srinagar
Faisalabad
Lahore
Bahawalpur
Karachi
Arabian Sea
NORTH
250
125
0
250 km
City administrative boundary
Disputed territory
Air quality stations
Sounding station
Cultivated land
Forest
Grasslands
Shrublands
Wetland
Water bodies
Artificial surfaces
Bareland



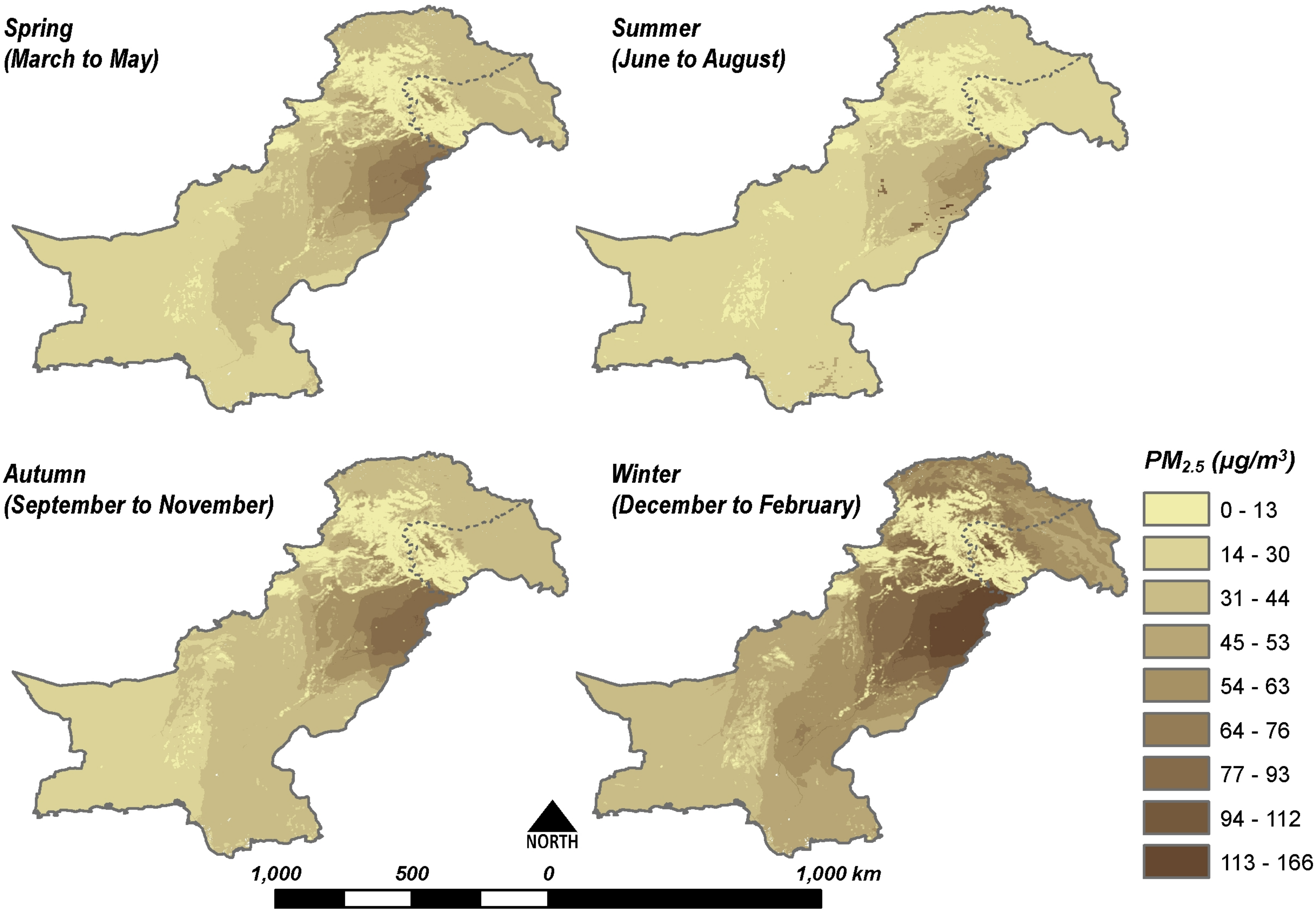
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
R with log(PM2.5)
100
200
300
400
600
1000
2000
3000
5000
10000
Buffer radii (m)
RDTRU_100
RDTER_300
RDTRU_1500
RDTER_1800
RDTRU_4000
RDTER_5000
RDTRU_8000
TREE_3000
WATER_3500
Spatial covariates
TREE
WATER
RDTRU
RDTER

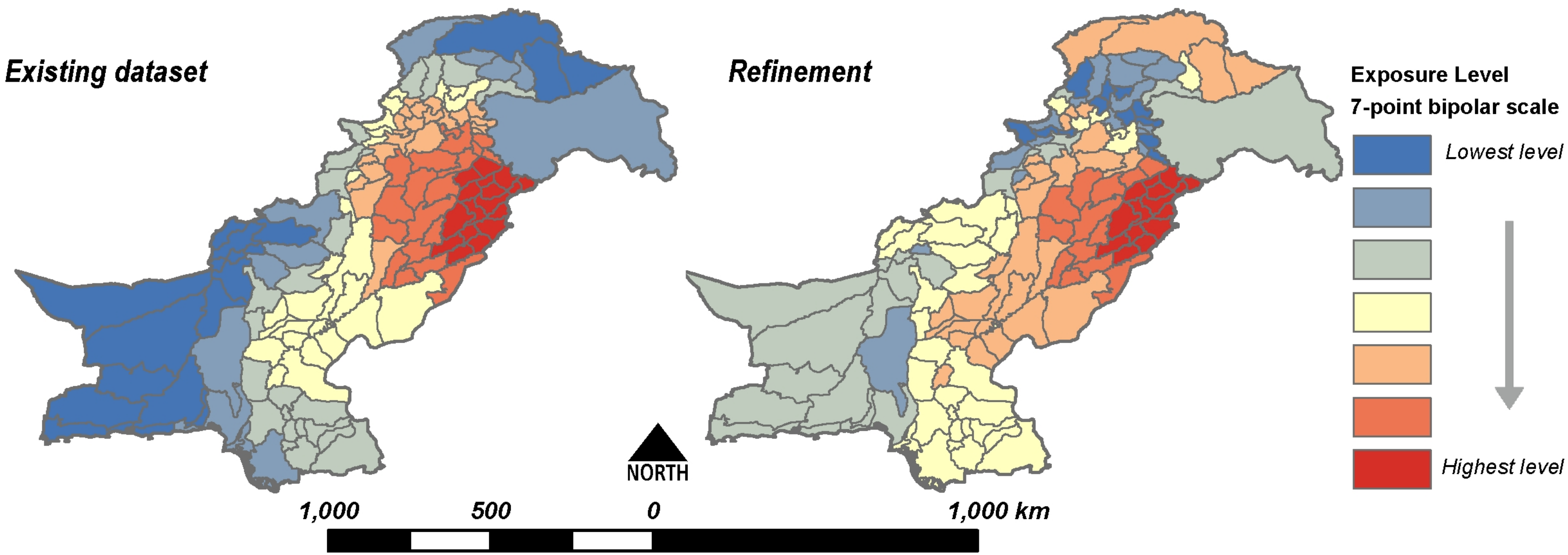


Adjusted R²=0.545
p-value<.0001
RMSE=0.1994
10-fold CV R²=0.542
Observed Log(PM2.5(μg/m3))
Predicted Log(PM2.5(μg/m3))
Log(PM2.5 (μg/m3)) Residual
Log(PM2.5(μg/m3)) Predicted value



NORTH
250 125 0 250 km
Peshawar
Islamabad-Rawalpindi
Srinagar
Faisalabad
Lahore
Bahawalpur
Karachi
Arabian Sea
City administrative boundary
Disputed territory
Air quality stations
Sounding station
Annual mean PM2.5 concentration (µg/m³)
>120
<20
68
10
75
4
119
43
96
41
59
7
38
30


Spring
(March to May)
Summer
(June to August)
Autumn
(September to November)
Winter
(December to February)
PM2.5 (µg/m3)
0 - 13
14 - 30
31 - 44
45 - 53
54 - 63
64 - 76
77 - 93
94 - 112
113 - 166
NORTH
1,000
500
0
1,000 km


Existing dataset
Refinement
Exposure Level
7-point bipolar scale
Lowest level
Highest level
NORTH
1,000
500
0
1,000 km

## ENVPOL_2020_1991R1 CRediT Author Statement

**Yuan Shi:** Methodology, Investigation, Formal analysis, Writing - Original Draft Preparation, Funding acquisition

**Muhammad Bilal:** Conceptualization, Validation, Writing - Review & Editing, Funding acquisition

**Hung Chak Ho:** Writing - Review & Editing

**Abid Omar:** Data Curation, Resources

# Urbanization and regional air pollution across South Asian developing countries - A nationwide land use regression for ambient $PM_{2.5}$ assessment in Pakistan

## - Supplementary material –

Yuan Shi[a], Muhammad Bilal[b]*, Hung Chak Ho[c], Abid Omar[d]

[a]Institute of Future Cities, The Chinese University of Hong Kong, Hong Kong SAR, China, shiyuan@cuhk.edu.hk

[b]School of Marine Sciences, Nanjing University of Information Science & Technology, Nanjing, 210044, China, muhammad.bilal@connect.polyu.hk

[c]Department of Urban Planning and Design, The University of Hong Kong, Hong Kong SAR, China, hcho21@hku.hk

[d]Pakistan Air Quality Initiative, Pakistan, abidomar@pakairquality.com

*Corresponding author:

Email: muhammad.bilal@connect.polyu.hk

Address: School of Marine Sciences, Nanjing University of Information Science and Technology, Nanjing 210044, China

***Table S-1.*** *Summary of meteorological, geographic predictor variables and satellite-based observations. Notes: [a] The 22-circular buffer scheme was created with radii of 100 m, 200 m, 300 m, 400 m, 500 m, 600 m, 700 m, 800 m, 1000 m, 1200 m, 1500 m, 1800 m, 2000 m, 2500 m, 3000 m, 3500 m, 4000 m, 5000 m, 6000 m, 7000 m, 8000 m, and 10000 m (Knibbs et al., 2014; Knibbs et al., 2018; Novotny et al., 2011); [b] Average or sum of the independent variable within buffer.*

| **Variable (units)** | **Abbreviation code** | **Spatial /temporal Resolution** | **Point or buffer [a] (average or sum [b])** | **Data source** |
|---|---|---|---|---|
| *Spatial datasets - Land use/land cover and geographical features* | | | | |
| Elevation (m) | ELEV | 30 m | Point | SRTM 1-ArcSecond Global Digital Elevation Model: https://earthexplorer.usgs.gov/ |
| Longitude | LONG | Vector | Point | Location of air quality monitoring stations |
| Latitude | LAT | Vector | Point | Location of air quality monitoring stations |
| Built-up areas/Impervious surfaces (%) | BUILT | 30 m | Buffer | High-resolution Multi-temporal Mapping of Global Urban Land 2015 (Liu et al., 2018): http://www.geosimulation.cn/GlobalUrbanLand.html |
| Tree coverage ratio (%) | TREE | 30 m | Buffer | GlobeLand30 (GLC30) (Jun et al., 2014) 2010 Data: http://www.globallandcover.com/ Lands covered with trees, with vegetation cover over 30%, including deciduous and coniferous forests, and sparse woodland with cover 10-30%. |
| Water coverage ratio (%) | WATER | 30 m | Buffer | GlobeLand30 (GLC30) (Jun et al., 2014) 2010 Data: http://www.globallandcover.com/ Water bodies in land area, including river, lake, reservoir, fish pond. etc. |
| Commercial land use area ($m^2$) | LUCOM | Vector | Buffer (total area) | OpenStreetMap (OSM): openstreetmap.org |
| Industrial land use area ($m^2$) | LUIND | Vector | Buffer (total area) | As above. |
| Residential land use area ($m^2$) | LURES | Vector | Buffer (total area) | As above. |
| Retail land use area ($m^2$) | LURET | Vector | Buffer (total area) | As above. |
| Distance to coast (km) | COAST | Vector | Point | Measured using 'Near' tool in ArcGIS (excludes inland waterbodies) |
| Population density (persons/$km^2$) | POPD | 250 m | Buffer | GHS_POP_GPW4_GLOBE_R2015A European Commission, Joint Research Centre (JRC); Columbia University, Center for International Earth Science Information Network - CIESIN (2015): GHS population grid, derived from GPW4, multitemporal (1975, 1990, 2000, 2015). European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/jrc-ghsl-ghs_pop_gpw4_globe_r2015a |
| Motorway and trunk roads (km) | RDTRU | Vector | Buffer (total length) | OpenStreetMap (OSM): openstreetmap.org |
| Primary roads (km) | RDPRI | Vector | Buffer (total length) | As above. |
| Secondary roads (km) | RDSEC | Vector | Buffer (total length) | As above. |

| | | | | |
|---|---|---|---|---|
| Tertiary roads (km) | RDTER | Vector | Buffer (total length) | As above. |
| Ordinary roads (km) | RDORD | Vector | Buffer (total length) | As above. (Includes: living street, residential, and service roads) |
| Count of bust stations and stops | BUSST | Vector | Buffer (count) | As above. |
| *Temporal-resolved datasets - Meteorological data and sounding data* | | | | |
| Annual mean daily average Air Temperature (°C) | TEMP | Daily | Point (at the same location of air quality monitoring stations) | Provided by Pakistan Air Quality Initiative (PAQI پاکی). |
| Annual mean daily average Relative Humidity (%) | RH | Daily | Point (at the same location of air quality monitoring stations) | As above. |
| Bulk Richardson Number | BRCH | Daily | Temporal variable (daily averaged value for all sounding data) | Department of Atmospheric Science, University of Wyoming at their website: http://weather.uwyo.edu/upperair/sounding.html (Station Number: 42027) |
| Bulk Richardson Number using CAPV | BRCV | Daily | Temporal variable | As above. |
| Convective Available Potential Energy (J/kg) | CAPE | Daily | Temporal variable | As above. |
| CAPE using virtual temperature (J/kg) | CAPV | Daily | Temporal variable | As above. |
| Convective Inhibition (J/kg) | CINS | Daily | Temporal variable | As above. |
| CINS using virtual temperature (J/kg) | CINV | Daily | Temporal variable | As above. |
| Cross totals index | CTOT | Daily | Temporal variable | As above. |
| K index | KINX | Daily | Temporal variable | As above. |
| Pressure of the Lifted Condensation Level (hPa) | LCLP | Daily | Temporal variable | As above. |
| Temperature of the Lifted Condensation Level (K) | LCLT | Daily | Temporal variable | As above. |
| Lifted index | LIFT | Daily | Temporal variable | As above. |
| LIFT computed using virtual temperature | LIFV | Daily | Temporal variable | As above. |
| Mean mixed layer mixing ratio (g/kg) | MLMR | Daily | Temporal variable | As above. |
| Mean mixed layer potential temperature (K) | MLPT | Daily | Temporal variable | As above. |

| | | | | |
|---|---|---|---|---|
| Total precipitable water (mm) | PWAT | Daily | Temporal variable | As above. |
| Showalter index | SHOW | Daily | Temporal variable | As above. |
| SWEAT index | SWET | Daily | Temporal variable | As above. |
| Total totals index | TTOT | Daily | Temporal variable | As above. |
| Vertical totals index | VTOT | Daily | Temporal variable | As above. |
| *Satellite-observation based spatial dataset* | | | | |
| Annual mean $PM_{2.5}$ estimation gridded at 0.01° ($\mu g/m^3$)* | PMGWR | 1 km | Point (the value in the corresponding cell of the satellite image) | Global Annual $PM_{2.5}$ Grids from MODIS, MISR and SeaWiFS Aerosol Optical Depth (AOD) with GWR, v1 (van Donkelaar et al., 2018): http://sedac.ciesin.columbia.edu/data/set/sdei-global-annual-gwr-pm2-5-modis-misr-seawifs-aod/data-download |

## Reference


Jun, C., Ban, Y., Li, S., 2014. China: Open access to Earth land-cover map. Nature 514, 434.

Knibbs, L.D., Hewson, M.G., Bechle, M.J., Marshall, J.D., Barnett, A.G., 2014. A national satellite-based land-use regression model for air pollution exposure assessment in Australia. Environ. Res. 135, 204-211.

Knibbs, L.D., van Donkelaar, A., Martin, R.V., Bechle, M.J., Brauer, M., Cohen, D.D., Cowie, C.T., Dirgawati, M., Guo, Y., Hanigan, I.C., Johnston, F.H., Marks, G.B., Marshall, J.D., Pereira, G., Jalaludin, B., Heyworth, J.S., Morgan, G.G., Barnett, A.G., 2018. Satellite-Based Land-Use Regression for Continental-Scale Long-Term Ambient PM2.5 Exposure Assessment in Australia. Environ. Sci. Technol. 52, 12445-12455.

Liu, X., Hu, G., Chen, Y., Li, X., Xu, X., Li, S., Pei, F., Wang, S., 2018. High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google Earth Engine Platform. Remote Sensing of Environment 209, 227-239.

Novotny, E.V., Bechle, M.J., Millet, D.B., Marshall, J.D., 2011. National satellite-based land-use regression: NO2 in the United States. Environ. Sci. Technol. 45, 4407-4414.

van Donkelaar, A., Martin, R., Brauer, M., Hsu, N., Kahn, R., Levy, R., Lyapustin, A., Sayer, A., 2018. Global Annual PM2. 5 Grids from MODIS MISR and SeaWiFS Aerosol Optical Depth (AOD) with GWR 1998–2016. NASA Socioeconomic Data and Applications Center (SEDAC).