A Satellite-based Model for Estimating $PM_{2.5}$ Concentration in a Sparsely Populated Environment Using Soft Computing Techniques

*Bijan Yeganeh[1, 2], Michael G. Hewson[3], Samuel Clifford[2, 4], Luke D. Knibbs[5], and Lidia Morawska[1, *]*

[1] International Laboratory for Air Quality and Health, Queensland University of Technology, Brisbane, Queensland 4001, Australia

[2] Centre for Air Quality and Health Research and Evaluation, Glebe, New South Wales 2037, Australia

[3] School of Education and the Arts, Central Queensland University, North Rockhampton, Queensland 4702, Australia

[4] Mathematical Sciences School, Queensland University of Technology, Brisbane, Queensland 4000, Australia

[5] School of Public Health, The University of Queensland, Herston, Queensland 4006, Australia

*** Corresponding author:** Professor Lidia Morawska, PhD

International Laboratory for Air Quality and Health

Queensland University of Technology, Brisbane, Queensland 4001, Australia

Phone: +61 7 3138 2616

Fax: +61 7 3138 9079

Email: l.morawska@qut.edu.au

Highlights

- We used comprehensive satellite-based, meteorological and geographical data to develop a satellite-based model for estimating the $PM_{2.5}$ concentration.
- Representative animations are created to visualize the spatiotemporal variation of the predictors.
- We applied the adaptive neuro-fuzzy inference system (ANFIS) for the first time as a core model to estimate the spatiotemporal variation of $PM_{2.5}$ concentration.
- We compared ANFIS with support vectors machine and back-propagation artificial neural network.
- Adaptive model identification technique has been used to identify the optimal predictive model.

Abstract

We applied three soft computing methods including adaptive neuro-fuzzy inference system (ANFIS), support vectors machine (SVM) and back-propagation artificial neural network (BPANN) algorithms for estimating the ground-level $PM_{2.5}$ concentration. These models were trained by comprehensive satellite-based, meteorological, and geographical data. A 10-fold cross-validation (CV) technique was used to identify the optimal predictive model. Results showed that ANFIS was the best-performing model for predicting the variations in $PM_{2.5}$ concentration. Our findings demonstrated that the CV-R^2 of the ANFIS (0.81) is greater than that of the SVM (0.67) and BPANN (0.54) model. The results suggested that soft computing methods like ANFIS, in combination with spatiotemporal data from satellites, meteorological data and geographical information improve the estimate of $PM_{2.5}$ concentration in sparsely populated areas.

Keywords: $PM_{2.5}$; Aerosol optical depth; ANFIS; SVM; BPANN; Australia.

Data availability

The type and source of the data set considered in this study.

Name of the data set	Data source (Developer) (All websites accessed on Jan 2016)	Data format	Software required	Data availability
OMI Near-UV AOD	Aura OMI AOD product via NASA Giovanni interface http://giovanni.sci.gsfc.nasa.gov/giovanni/?instance_id=omil2g	HDF / NetCDF files	ArcGIS	Freely available
Major road	PSMA Australia Transport and Topography product https://www.psma.com.au/products/transport-topography	ESRI shape files	" "	Price depends on the area of interest
Minor road	" "	" "	" "	" "
Industrial point source $PM_{2.5}$ emissions	Australia National Pollutant Inventory http://www.npi.gov.au/reporting/industry-reporting-materials	xml files	Microsoft Excel / R	Freely available
Australia population density	Australian Bureau of Statistics http://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.007	PNG ESRI Grid GeoTIFF	ArcGIS	" "
Australia land use classification	Australian Bureau of Statistics http://www.abs.gov.au/websitedbs/censushome.nsf/home/meshblockcounts	Excel spreadsheets / CSV files	Microsoft Excel / R / ArcGIS	" "
Elevation	U.S. Geological Survey https://www.usgs.gov/products/maps/topo-maps	PNG GeoTIFF	ArcGIS	" "
Normalized difference vegetation index	Terrestrial Ecosystem Research Network http://www.auscover.org.au/node/9	NetCDF files	" "	" "
Temperature **Rainfall** **Humidity** **Solar exposure**	Australian Bureau of Meteorology http://www.bom.gov.au/climate/maps/#tabs=Maps	ESRI Grid GIF	" "	" "

Software availability

The following software has been used in this study for statistical analysis, spatial data processing and map creation:

- R v.3.2.3 (R Foundation for Statistical Computing, Vienna, Austria)
- MATLAB R2014b (MathWorks Inc., Natick, USA)
- ArcGIS version 10.2 (ESRI Inc., Redlands, USA)

Note: No specific software component has been developed for this study.

1. Introduction

Exposure to fine particulate matter ($PM_{2.5}$, particles with aerodynamic diameter less than 2.5 μm) is a leading environmental risk factor associated with respiratory and cardiovascular morbidity and mortality (Franklin et al., 2007) and it is the twelfth-ranked contributor to the global burden of diseases (Forouzanfar et al., 2015).

Urbanisation increases the risk of being exposed to $PM_{2.5}$ (Han et al., 2015), and Australia, as one of the most urbanised countries in the world , is faced with adverse health effects of $PM_{2.5}$. To date, very little attention has been paid to the health effect of exposure to $PM_{2.5}$ in Australia. Some studies consistently suggest that $PM_{2.5}$ is associated with respiratory diseases and has significant effects on mortality (Barnett et al., 2005; Simpson et al., 2005), while conflicting results have been reported on cardiovascular health effects (Hinwood et al., 2006). These inconsistent results could be due to difficulties in assessing the Australian population exposure to $PM_{2.5}$.

Ground level aerosol measurement has been historically provided by ground monitoring networks, but there are high establishing and maintaining expenses associated with these measurements (Wu et al., 2012). The sparse ground $PM_{2.5}$ measurement network in Australia makes it difficult to evaluate the spatiotemporal variability of $PM_{2.5}$ and has significantly restrained the epidemiological studies on $PM_{2.5}$ health effects. Australia is the sixth largest country in the world by area while its population is quite small compared to the land size (Australian Government, 2015). Australia is one of the 10 least dense populated countries in the world (United Nations, 2015). The majority of the Australian population is living in the east and west coasts (Lunn et al., 2002). The population within these areas is concentrated in urban centres, particularly the capital cities (Australian Bureau of Statistics, 2012; Lunn et al., 2002).

Therefore, limited monitoring stations were established only in populated areas due to population distribution in Australia. Had such monitoring networks existed, there would have been no guarantee of an effective measurement of the spatiotemporal variation of $PM_{2.5}$, since it is changing on scales much smaller than monitoring networks density.

Estimates of air pollution exposure have been traditionally provided by assigning measurements derived from one (Chen et al., 2006) or several air pollution monitors (Barnett et al., 2005; Brook et al., 2010; Chan et al., 2006), allocating exposure using the nearest monitoring station (Lee et al., 2014) or using different proxies to estimate a local population's exposure (Hoffmann et al., 2007; Salam et al., 2008; Samet, 2007). There is potential for over-smoothing the exposure estimation and the results are likely to be biased with all these approaches (Jerrett et al., 2005a).

Satellite imagery is another important tool rapidly gaining interest in air pollution monitoring as it provides sequential observations over a broad area. Satellite sensors can be coupled with ground-based sensors at different spatiotemporal scales to reduce the limitations of surface monitoring station (Reis et al., 2015). Aerosol Optical Depth (AOD) is the most common parameter derived from satellite observations and applied to estimate $PM_{2.5}$. AOD describes the level of which aerosols attenuate the electromagnetic radiation at a given wavelength by absorption or scattering in an atmospheric column (Chudnovsky et al., 2012; Kaufman et al., 2002; NASA, 2013). The availability of satellite-derived AOD has helped to overcome the problems associated with sparse monitoring networks by providing observations where previously there were none (Hoff and Christopher, 2009; Reis et al., 2015).

A variety of methods have been used to investigate the quantitative relationship between satellite-derived AOD and ground-level $PM_{2.5}$ measurements. These studies mainly fall into two major classes: numerical-based methods and empirical observation-based methods (Lin et al., 2014).

Numerical-based models, including dispersion and chemical transport models, are still under development due to the uncertainties regarding the definition of source inventories, and chemical and dynamical processes of aerosols in atmosphere (Gupta and Christopher, 2009b; Kondragunta et al., 2008). Empirical observation-based methods rely on the relationship between air quality measurements and different observations (Maciejewska et al., 2015). Several techniques have been used to describe this relationship including simple regression (Chu et al., 2003), multiple regression (Dirgawati et al., 2015; Gupta and Christopher, 2009b; Li et al., 2011), geostatistical methods (Jerrett et al., 2005b; Kunzli et al., 2005), generalized additive models (GAM) (Strawa et al., 2013), land use regression (Henderson et al., 2007; Kloog et al., 2011; Knibbs et al., 2014; Liu et al., 2009), and hybrid approaches (Beckerman et al., 2013b; Lindstrom et al., 2011). Soft computing refers to computational techniques which are able to achieve optimal solutions for analysing complicated phenomena at reasonable costs (Carnevale et al., 2016; Kruse et al., 2013; Ovaska, 2004). In recent years, soft computing techniques such as support vector machine (SVM) (Moazami et al., 2016; Reid et al., 2015; Yeganeh et al., 2012), Bayesian models (Corani and Scanagatta, 2016; McBride et al., 2007), *k*-nearest neighbours (kNN) (Reid et al., 2015), and artificial neural network (ANN) (Al-Alawi et al., 2008; Gupta and Christopher, 2009a; Ordieres et al., 2005; Wu et al., 2012) have been gaining popularity in air quality modelling because of their high flexibility and well documented prediction abilities. However, other soft comuting methodes like adaptive neuro-fuzzy inference system (ANFIS), which is accepted as an efficient and robust method for multivariate analysis, have not been used for modelling the spatiotemporal variations of $PM_{2.5}$ concentrations.

Although most of the aforementioned methods can be applied to determine the relationship among AOD and $PM_{2.5}$, imposing a specific method could make it difficult to select the best predictive model. Hence, adaptive model identification approach is used to choose the most efficient model by using cross-validation technique rather than fitting a specific model to the dataset (Reid et al., 2015; Syed, 2011).

Few studies have investigated the relationship between $PM_{2.5}$ and satellite-based AOD in Australia (Gupta et al., 2007; Gupta et al., 2006; Meyer et al., 2008). While other studies have recommended the meteorological and geographical factors incorporation to the AOD–$PM_{2.5}$ relationship to improve models' performance (Chudnovsky et al., 2014; Liu et al., 2009), there is a clear need to conduct an Australian study to develop a satellite-based model investigating significant geographical and meteorological factors including humidity, planetary boundary layer, and wind speed and direction.

In this study, we aimed to improve the estimate of $PM_{2.5}$ concentration by using remotely-sensed AOD in conjunction with comprehensive meteorological and geographical data. Three different soft computing algorithms were applied to estimate the monthly average exposure to $PM_{2.5}$ in the South-east Queensland (SEQ) region of Australia, from 2006 to 2011. In turn, an adaptive model identification approach was used to choose the optimal model from ANFIS and other soft computing methods: SVM and BPANN, by using 10-fold cross-validation (Pandey et al., 2013; Syed, 2011). We ultimately used the model with the best predictive ability to estimate spatiotemporal variation of $PM_{2.5}$ in this sparsely populated area with dense vegetation cover.

2. Materials and Methods

2.1. Study location and ground-level $PM_{2.5}$ measurements

SEQ is a region in the state of Queensland, Australia, which covers 22,420 km^2 and is home to 3.05 million people out of the state's population of 4.58 million based on the 2011 Australian census (Australian Bureau of Statistics., 2012). The study area consists of Brisbane, the state's capital city, as well as other urban and rural centres including Ipswich, Logan City, Gold Coast, Sunshine Coast, and the Lockyer Valley. Motor vehicle emissions and industrial boilers are identified as major sources of $PM_{2.5}$ in SEQ (Queensland Government., 2014). The Queensland government and other agencies are responsible for regulatory aerosol monitoring in SEQ. We obtained quality-assured 24 h ground-level $PM_{2.5}$ measurements from January 2006 to December 2011. During the study period, $PM_{2.5}$ was measured at 8 monitoring sites across SEQ (supplement, page S3). We used monthly averages of the daily measured $PM_{2.5}$, and the inclusion criteria for a given month was that less than 5% of the daily measurements were missing.

2.2. Land use data

We obtained data on anthropogenic and natural land use variables as spatial predictors that were possible predictors of measured $PM_{2.5}$ concentrations. The selected land use variables, summarised in Table 1, were examined to discover which ones improved the prediction of $PM_{2.5}$ (Knibbs et al., 2014). They included proxies for emissions from traffic, point sources and changing land cover conditions.

The impacts of vegetation cover and its phenological state on the relationship between the $PM_{2.5}$ and satellite AOD were also examined in the present study. Normalized difference vegetation index (NDVI) is used to provide a measure of greenness and vegetation cover. NDVI was found to be an effective predictor for pollutant concentrations in previous studies (Chudnovsky et al., 2014; Dirgawati et al., 2015; Su et al., 2009). The monthly mean NDVI data were derived from an Advanced Very High Resolution Radiometer (AVHRR) sensor carried on the National Oceanic and Atmospheric Administration (NOAA) satellite and processed by the Australian Bureau of Meteorology (BoM) at a spatial resolution of 1 km (Bureau of Meteorology, 2015).

2.3. Satellite data

Daily global gridded observations of AOD at a resolution of 0.25 degrees latitude and longitude are derived from the Ozone Monitoring Instrument (OMI) aboard the Aura satellite (Levelt et al., 2006). Aura crosses the equator in a sun-synchronous polar orbit for the daylight ascending orbit (Torres et al., 2007), and it passes over SEQ at approximately 14:00 local time. We download the monthly average OMI AOD level 2 Near-UV AOD and single Scattering Albedo product (OMAERUVG.003 at 342.5 nm) from NASA Giovanni interface for each month from 2006–2011.

Table 1. Independent variables included as potential predictors of $PM_{2.5}$

Variables (units)	Spatial resolution	Point or buffer	Data source
OMI Near-UV AOD	0.25 degrees Lat/Lon	Point	Aura OMI AOD product via NASA Giovanni interface
Distance to coast (m)	-	Point	ArcGIS geoprocessing tools
Distance to port (m)	-	Point	" "
Distance to airport (m)	-	Point	" "
Distance to nearest major road	-	Point	" "
Distance to nearest minor road	-	Point	" "
Airport (present/not present)	-	Buffer	" "
Major road (m)	-	Buffer	PSMA Australia Transport and Topography product
Minor road (m)	-	Buffer	" "
Industrial point source $PM_{2.5}$ emissions (kg/yr)	-	Buffer	Australia National Pollutant Inventory
Time (Julian month)	-	Point	ArcGIS geoprocessing tools
Population density (person/km^2)	1 × 1 km^2	Point	Australian Bureau of Statistics
Land use by type (% area)[b]	Mesh block[c]	Buffer	Australian Bureau of Statistics
Elevation (m)	30 m	Point	U.S. Geological Survey
Normalized difference vegetation index	1 × 1 km^2	Point	Terrestrial Ecosystem Research Network, Australian Bureau of Meteorology, AusCover project and NASA NOAA satellite
Mean daily maximum temperature (°C)	5 × 5 km^2	Point	Australian Bureau of Meteorology
Mean daily minimum temperature (°C)	5 × 5 km^2	Point	" "
Rainfall (mm)	5 × 5 km^2	Point	" "
Humidity (hPa)	5 × 5 km^2	Point	" "
Solar exposure (MJ/m^2)	6 × 6 km^2	Point	" "
Planetary boundary layer height (m)	3 × 3 km^2	Point	Derived from Weather Research and Forecasting model
U-component of wind speed (m/s)	3 × 3 km^2	Point	" "
V-component of wind speed (m/s)	3 × 3 km^2	Point	" "
Wind speed (m/s)	3 × 3 km^2	Point	" "
Wind direction (Degrees)	3 × 3 km^2	Point	" "

[a] 22 Circular buffers were generated with radii of 50 m, 100 m, 200 m, 300 m, 400 m, 500 m, 600 m, 700 m, 800 m, 900 m, 1000 m, 1200 m, 1500 m, 1800 m, 2000 m, 2500 m, 3000 m, 3500 m, 4000 m, 5000 m, 7500 m, and 10,000 m (Novotny et al., 2011).

[b] Four different land use classes were investigated including industrial, commercial, residential, and open space (which contains the agricultural land, parks, and water bodies (Rose et al., 2010)).

[c] Mesh Block is the smallest geographic unit defined by the Australian Statistical Geography Standard for which the Census data is available (Australian Bureau of Statistics, 2011), and can be variable in size.

2.4. Meteorological Data

We obtained surface meteorological parameters including mean maximum and minimum temperature, rainfall, and humidity from high-quality spatial climate data-sets developed by BoM which provides gridded climatological maps for each month of the year (Jones et al., 2009). In addition, monthly solar exposure maps are also obtained from BoM during the study period.

Planetary boundary layer height (PBLH), wind direction (WD) and wind speed (WS) can play a critical role in the transport and dilution of $PM_{2.5}$ (Harrison et al., 1997); hence, special attention was paid to these parameters in this study. The Weather Research and Forecasting model (WRF) was used to calculate these parameters as at 2:00 pm local time at a spatial resolution of 3 km to match the over-pass time of the Aura satellite. Details on the WRF configuration are provided in the supplement (page S3-S7).

2.5. Modelling approach

Following similar studies (Knibbs et al., 2014; Novotny et al., 2011), 22 circular buffers were created around each monitoring site to obtain local and more remote sources of $PM_{2.5}$ (Table 1). Certain variables were calculated within a buffer (e.g., land use type, road length) while others were extracted at each monitoring site (e.g., wind speed, humidity, temperature). In total, 194 independent variables were obtained including 18 point variables and 176 buffer variables (8 variables calculated at 22 buffers each).

2.6. Statistical analysis

In this study, there were 194 predictor variables to choose from, hence choosing the optimum subset was a complicated process and needed to be carefully conducted. In many soft computing and data mining tasks, there can be some irrelevant variables which may affect the derived statistical relationship between the dependent variable and the other relevant predictors. A common solution to overcome this problem is to use a variable selection process which can help to select a subset of the most relevant and representative predictors from input predictors. The Least Absolute Shrinkage and Selection Operator (Lasso) is a well-known method which is widely used to suppress or shrink variables to select the most relevant predictor variable set. Lasso-type variable selection method was used in this study since it was successfully adopted in many applications (Hu et al., 2015; Li and Shao, 2015; Tibshirani, 2011).

Following Beelen et al. (2013), we only included potential predictor variables with less than 10% null values and centred and standardised some independent variables to improve model convergence and make the parameter estimates more interpretable. Subsequently, all remaining predictor variables were evaluated, and variables with p-value greater than 0.10 or variance inflation factor (VIF) greater than 6 were removed in order to avoid multicollinearity (see Table S2). As suggested by other studies (Beelen et al., 2013; Henderson et al., 2007; Novotny et al., 2011; Vienneau et al., 2013), if two buffer sizes of a particular variable were found to be collinear, donut ring buffers (so called concentric adjacent rings) were replaced with original circular buffers and the analysis was redone. Ring buffers (i.e., annulus) were calculated by differencing the circular buffers.

In this study, we employed the soft computing techniques ANFIS, BPANN and SVM. The soft computing techniques employed here are explained in the supplement, page S8-S17. The input variables were composed of different types of data, including land use, meteorological, and satellite data. We matched the selected variables with the $PM_{2.5}$ measurements at 8 monitoring sites during the study period (72 months) which resulted in more than 550 observation sets in total. These observations were divided into training, validation, and test subsets. The majority of the observations (70%) were used for training the models. In order to avoid over-training, 15% of the observations were used for validation and checking the model's generalisation (Wu et al., 2012). Finally, the remaining 15% of the observations were employed as the test subset to estimate the $PM_{2.5}$ concentration by the models.

In this study, 10-fold cross validation (CV) method was applied to evaluate the performance of the BPANN, ANFIS, and SVM models and identify the optimal model for estimating the $PM_{2.5}$ concentration. This method has the ability to examine the model's predictive ability (Beckerman et al., 2013a). This examination was accomplished by randomly splitting the data into 10 equal-sized folds. Subsequently, one of the folds was used to test the model and the remaining 9 folds were used to train the model (Kim, 2009; Refaeilzadeh et al., 2009). This process was repeated 10 times for each candidate model while all folds were used as the test subset and the 10 results were averaged to obtain the overall CV-R^2 and CV-RMSE. The best predictive model was selected from those with the smallest CV-RMSE and highest CV-R^2 (Dirgawati et al., 2015).

Bland-Altman plot was also used to examine the agreement between the observations and predictions. In this plot, X axis shows the average of the model predictions and observations, and Y axis represents the difference between these values. Bland-Altman plot also provides statistical limits by calculating the average and mean and the standard deviation (*sd*) of the differences between observations and predictions (Giavarina, 2015). These limits were used to evaluate the agreement between observations and model predictions. For more explicit information on Bland-

Altman plot see Giavarina (2015). Figure 1 illustrates the overall research process used in this study.

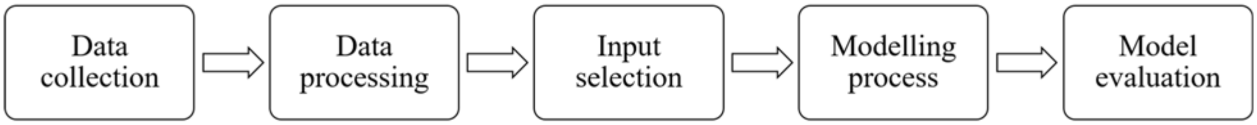


Figure 1. General research process for estimating $PM_{2.5}$ concentration.

We used R v.3.2.3 (R Foundation for Statistical Computing, Vienna, Austria) and MATLAB R2014b (MathWorks Inc., Natick, USA) for all statistical and soft computing analyses and ArcGIS version 10.2 (ESRI Inc., Redlands, USA) for spatial data processing and map creation.

3. Results

3.1. Modeling results and evaluation

In this study, a wide range of ground-based $PM_{2.5}$ measurements, land use, meteorological, and remotely-sensed AOD data were employed to estimate the $PM_{2.5}$ concentration using soft computing techniques. In following section, the agreement between predicted and observed $PM_{2.5}$ concentration is evaluated. 10-fold cross validation is also used to compare the potential of different algorithms for estimating $PM_{2.5}$ concentration.

The variable selection results showed that 16 variables were the most effective predictors of $PM_{2.5}$ concentration. The variables most correlated with the outcome were firstly humidity, followed by maximum temperature, AOD and then the length of the major roads. The results of the variable selection process are provided in the supplement (Table S2).

Using the testing dataset for each of the developed models, $PM_{2.5}$ concentrations were then predicted. A summary of the observed and predicted $PM_{2.5}$ concentrations is presented in Fig. 2.

The mean observed $PM_{2.5}$ concentration for the testing dataset is 6.77 μg/m^3. All three models approached this value within a numerical range of -0.02 to +0.38 μg/m^3. The non-parametric Wilcoxon test was performed to check if there was any significant difference between the predicted and observed mean $PM_{2.5}$ concentrations of each model. The test on all models yielded p-values greater than 0.01, showing an insignificant difference between the predicted and observed $PM_{2.5}$ concentration at 1% significant level. A comparison of the predicted values demonstrated that the ANFIS model predicted values were slightly closer to the full range of the observed monitoring data than the SVM and BPANN models. In general, Fig. 2 shows that all models reliably calculated the average and range of $PM_{2.5}$ concentration; therefore, predicted-observed plots are used to evaluate the predictive abilities of the models (Fig. 3).

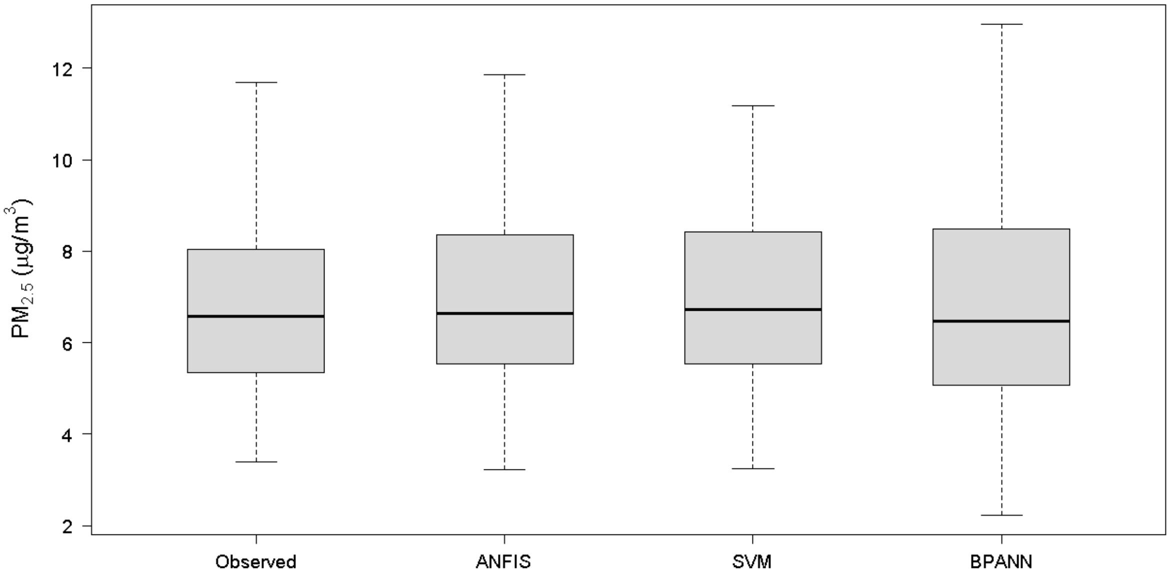


Figure 2. Summary of $PM_{2.5}$ model predictions on the testing dataset.

We compared the observed $PM_{2.5}$ concentrations to the predicted values of the ANFIS, SVM, and BPANN models. The ANFIS's predicted-observed plot indicates that the values are more equally scattered across the line of agreement at the low and high $PM_{2.5}$ concentrations whereas the SVM model under-predicts and over-predicts these values, respectively. In addition, the predicted-observed plot shows relatively weak correlation between the BPANN's predictions and actual observations.

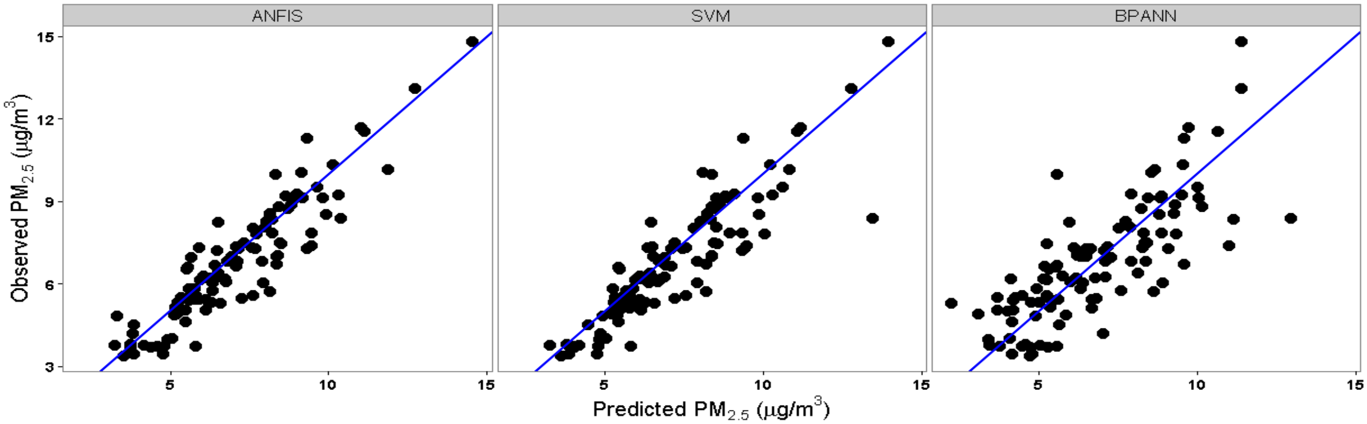


Figure 3. Scatter plots of observed vs. predicted $PM_{2.5}$ for the optimal model fitting on the testing dataset using ANFIS, SVM, and BPANN, respectively. Blue line indicates the line of agreement (y = x).

Table 2 compares the R^2 and RMSE for model fitting and cross validation. For the model fit the R^2 values are 0.61, 0.73, and 0.84 for the BPANN, SVM and ANFIS models, respectively. The RMSE values are 1.57 μg/m^3, 1.36 μg/m^3, and 0.94 μg/m^3 for the BPANN, SVM, and ANFIS models, respectively. Comparing the model fittings and cross validation, CV-R^2 decreases by just 0.03 and CV-RMSE increases by 0.85 μg/m^3 for the ANFIS model indicating negligible model overfit. The CV-R^2 of the SVM and BPANN decreased by 0.06 and 0.07, respectively indicating both models overfit more than ANFIS.

Table 2. R-squared and RMSE for model fittings vs. cross validation

	R^2	**RMSE ($\mu g/m^3$)**	**CV-R^2**	**CV-RMSE ($\mu g/m^3$)**
ANFIS	0.84	0.94	0.81	1.79
SVM	0.73	1.36	0.67	2.02
BPANN	0.61	1.57	0.54	2.11

Our findings demonstrated that the CV-R^2 of the ANFIS (0.81) was higher than that of the SVM (0.67) and BPANN (0.54) model. Also, the CV-RMSE of the ANFIS model (1.79 $\mu g/m^3$) was lower than that of the SVM (2.02 $\mu g/m^3$) and BPANN (2.11 $\mu g/m^3$) model. Compared to SVM and BPANN models, the ANFIS model had higher accuracy without causing more overfit.

Bland-Altman analysis was used to evaluate the agreement between the observation and predictions of ANFIS and SVM since both models showed promising performance in the testing stage (Figure 4). The Bland-Altman plots demonstrated low bias in both models; however, the ANFIS model had slightly tighter agreement than the SVM with fewer large residuals.

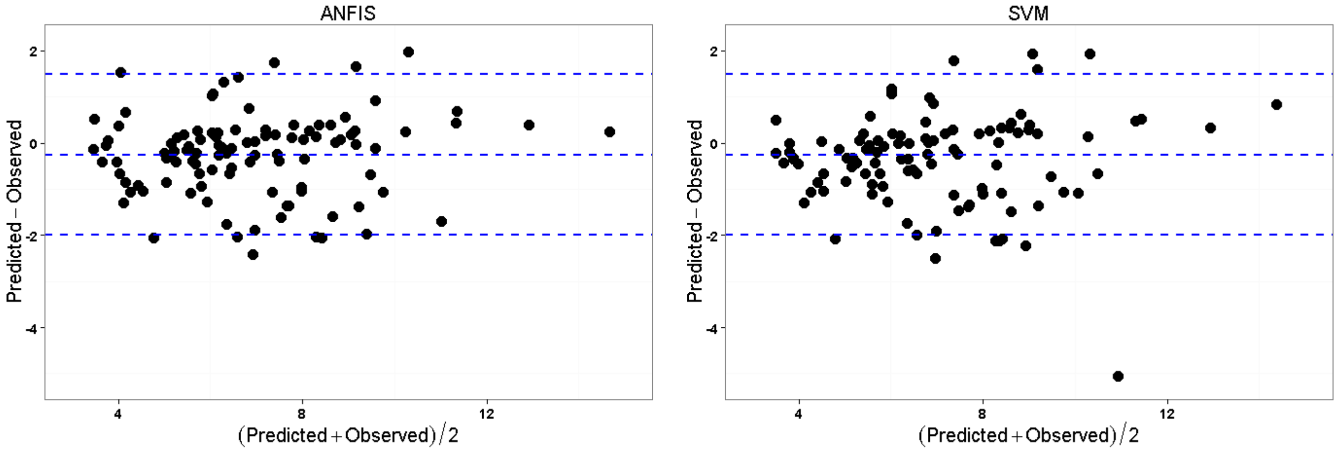
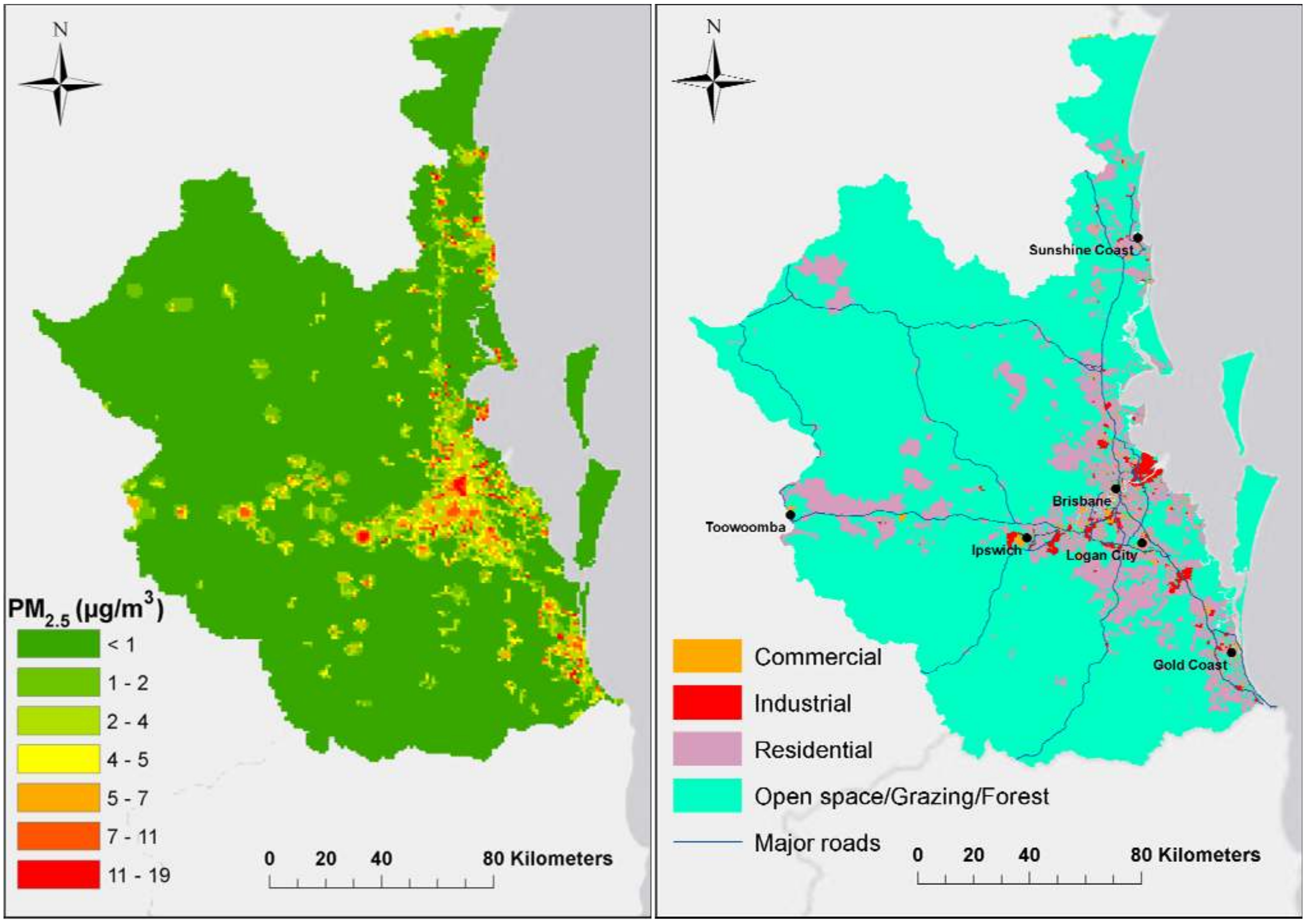


Figure 4. Bland-Altman plots of predicted and observed $PM_{2.5}$ concentrations ($\mu g/m^3$)

3.2. Application of the model

The predicted values of PM2.5 concentration in September 2011, selected from the 6-year study period, using ANFIS model for the study area 1 km grid is presented in Figure 5a.



(a) (b)

Figure 5. a) Monthly average PM2.5 concentration in September 2011 predicted by ANFIS model b) land use map of SEQ

Figure 5b illustrates land use map of SEQ. Concentrations ranged from less than 2 to 19 µg/m^3. Areas with higher concentrations (7 to 19 µg/m^3) corresponded to cities and major towns. Higher concentrations were predicted in locations with extensive adjacent industrial areas and major roads. This pattern was observed in all 6 cities of the study area. The highest levels were predicted for the three largest cities: Brisbane (with 1.977 million people), Gold Coast (494,500) and Logan (287,474).

4. Discussion

We employed soft computing techniques to improve concentration estimates for $PM_{2.5}$ using satellite, meteorological and land use predictor variables in South-east Queensland, Australia. The ANFIS model utilized in this work was the first attempt to apply it for spatiotemporal modelling of $PM_{2.5}$. Using cross validation technique, the ANFIS model was found to have the best performance compared to SVM and BPANN models, and better agreement with the observed data. The results provide estimates of monthly $PM_{2.5}$ concentrations for SEQ from 2006 to 2011.

ANFIS is a hybrid system that combines the strengths of fuzzy logic and artificial neural network (Jang, 1993; Lin and Lee, 1991), which provides a robust and accurate method for predicting $PM_{2.5}$ concentration over the range of observations used in this study.

In this research, WRF used to calculate PBLH, and WS. Both parameters were highly associated with $PM_{2.5}$ concentration across ANFIS runs. In addition, daily maximum temperature had higher importance compared to the daily minimum temperature. This might be because daily maximum temperature is temporally coincident with the Aura satellite overpass time for the study area.

Prior studies mostly considered land use parameters mainly focused on roadways and population-related data. We also evaluated industrial point source emissions, port and airport locations as potential land use predictors. Land use parameters used in our study, may not be important predictors for short term events (e.g. bush fire episode) but our findings revealed that they are of the strong predictors for $PM_{2.5}$ estimation in a long term period.

Data sets with different spatial resolutions have been used in our study. The resolution of NDVI and WRF outputs for example are finer than the OMI sensor data. Individually, OMI AOD data are not spatially fine enough for estimating the $PM_{2.5}$ exposure in epidemiological studies. However, method of combining different buffer sizes of land use parameters with meteorological and satellite-based data enabled the model to integrate fine and more spatially coarse data sets to estimate $PM_{2.5}$ concentration and provide more informative results for epidemiological studies. Our results also corroborated with Reis et al. (2015) who suggested that the incorporation of 'big data' derived from different sources provides new opportunities for data-intensive models to improve the estimates of population exposures to air pollution. Another important finding was that the highest concentrations (10 to 20 $\mu g/m^3$) were estimated in the Brisbane, Gold Coast and Logan City which have the highest population density in SEQ (Fig. 5); hence, the increased risk of population expose to higher concentration of $PM_{2.5}$. Although the average $PM_{2.5}$ concentration was below the WHO guideline (25 $\mu g/m^3$), epidemiological studies have demonstrated that $PM_{2.5}$ exposure to even lower concentrations is associated with high health risks (Burnett et al., 2014). Based on the shape of the $PM_{2.5}$ exposure-response curves derived by Burnett et al. and used in the Global Burden of Disease studies (Burnett et al., 2014), characterizing with a steep slope for concentrations ranging from 10 to 20 $\mu g/m^3$, exposure to $PM_{2.5}$ concentration between 10 to 20 $\mu g/m^3$ highly increases the relative risk of stroke and chronic obstructive pulmonary disease; therefore, even $PM_{2.5}$ concentration below the WHO guideline could not be considered safe.

Different methodologies make it difficult to compare our results to other studies, however we have attempted to compare our results with two studies which have demonstrated the ability of remotely sensed AOD and meteorological data to predict $PM_{2.5}$ concentration (Gupta and Christopher, 2009a; Wu et al., 2012). Both studies used BPANN method to estimate the spatiotemporal variation of $PM_{2.5}$, and reported R^2 lower than 0.61. Our model exhibited better correlations than these models, which could be due to either: (1) the comprehensive input variables used or (2) the more robust modeling algorithms used. We also compared our methodology with a study of the global burden of disease 2013 conducted by Brauer et. al. (2016) which combined ground measurements, chemical transport model outputs, and satellite-based data to provide global estimates of $PM_{2.5}$ concentration. Although, chemical transport model simulations were unavailable for our study area, our model was still able to capture 81% of $PM_{2.5}$ variations.

Previous research demonstrated that PBLH and WS significantly affect the $PM_{2.5}$-AOD relationship. Our results support these findings, but also demonstrate that incorporating other spatial and spatiotemporal data as well as road density, land use types, NDVI, and industrial point sources improves the model's performance.

5. Conclusions

Three different soft computing methods were applied to develop a satellite-based model for estimating the spatiotemporal variation of $PM_{2.5}$. ANFIS performed very well compared to SVM and BPANN. It exhibited satisfactory performance with CV-R^2, and CV-RMSE equals to 0.81, and 1.79 μg/m^3, respectively. It provides estimates of monthly $PM_{2.5}$ concentrations during 2006-2011. The modelling approach used in this study is highly applicable to similar settings anywhere in the world assuming that researchers have access to data sets equivalent to these used in our study. WRF, and its underlying boundary condition data, is a community model available to the world research fraternity. The NASA Giovanni data is available to anyone in the research community who can be registered with NASA as a data user. It is expected that researchers will have access to all other similar data sets or proxies in their own national research and information collection institutions, hence this method could be applied in other regions that experience $PM_{2.5}$ exposure. We hope that our approach will be beneficial for epidemiological studies and other researches seeking spatially accurate estimates of $PM_{2.5}$ with few monitoring stations. It is certainly feasible to develop a model with higher spatial resolution which is a direction of our future research. Further analysis such as global sensitivity and uncertainty analyses (GSUA) can also be done to assess input factor importance and interaction (Lüdtke et al., 2008; Saltelli et al., 2008). Recently developed Unified-Weather Research and Forecasting model (NU-WRF) can be employed to obtain more accurate estimates of meteorological parameters compared to WRF model (Peters-Lidard et al., 2015). Data management remains a major challenge for empirical modelling as it requires to store, process and analyse large data sets containing different types of data from multiple sources. Recently, new information models have been developed to facilitate the data management and validation in observation-based studies (Horsburgh et al., 2016; Shu et al., 2016).

Acknowledgment

A PhD scholarship to Bijan Yeganeh has been provided by the Centre for Air Quality & Health Research and Evaluation (National Health and Medical Research Council Centre of Research Excellence). We thank the scientists and staff of NASA for the Aura mission as well as the Netherlands Agency for Aerospace Programs and the Finnish Meteorological Institute for the OMI sensor. We also thank the Australian Government, Bureau of Meteorology, NASA/NOAA, Australian National Pollutant Inventory, Australian Bureau of Statistics, CSIRO and the AusCover project for land use, AOD, NDVI and related data sets. We thank the Queensland Government for $PM_{2.5}$ data. WRF was processed on the National Computational Infrastructure (NCI) facility in Canberra, Australia, which is supported by the Australian Government.

References

Al-Alawi, S.M., Abdul-Wahab, S.A., Bakheit, C.S., 2008. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. Environmental Modelling & Software 23(4) 396-403.

Australian Bureau of Statistics, 2011. Australian Statistical Geography Standard, Volume 1. , Main Structure and Greater Capital City Statistical Areas: Canbrra, Australia, p. 15.

Australian Bureau of Statistics, 2012. Year Book Australia: Canberra , Australia, pp. 237-251.

Australian Bureau of Statistics., 2012. Population Change in South-East Queensland: Australia.

Australian Government, 2015. Our Natural Environment: Australia.

Barnett, A.G., Williams, G.M., Schwartz, J., Neller, A.H., Best, T.L., Petroeschevsky, A.L., Simpson, R.W., 2005. Air pollution and child respiratory health: a case-crossover study in Australia and New Zealand. American Journal of Respiratory and Critical Care Medicine 171(11) 1272-1278.

Beckerman, B.S., Jerrett, M., Martin, R.V., van Donkelaar, A., Ross, Z., Burnett, R.T., 2013a. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. Atmospheric Environment 77 172-177.

Beckerman, B.S., Jerrett, M., Serre, M., Martin, R.V., Lee, S.-J., van Donkelaar, A., Ross, Z., Su, J., Burnett, R.T., 2013b. A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of PM2.5 in the Contiguous United States. Environmental Science & Technology 47(13) 7233-7241.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.-Y., Künzli, N., Schikowski, T., Marcon, A., 2013. Development of NO_2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe–the ESCAPE project. Atmospheric Environment 72 10-23.

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R.V., Dentener, F., Dingenen, R.v., Estep, K., Amini, H., Apte, J.S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P.K., Knibbs, L.D., Kokubo, Y., Liu, Y., Ma, S., Morawska, L., Sangrador, J.L.T., Shaddick, G., Anderson, H.R., Vos, T., Forouzanfar, M.H., Burnett, R.T., Cohen, A., 2016. Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013. Environmental Science & Technology 50(1) 79-88.

Brook, R.D., Rajagopalan, S., Pope, C.A., Brook, J.R., Bhatnagar, A., Diez-Roux, A.V., Holguin, F., Hong, Y., Luepker, R.V., Mittleman, M.A., 2010. Particulate matter air pollution and cardiovascular disease an update to the scientific statement from the American Heart Association. Circulation 121(21) 2331-2378.

Bureau of Meteorology, 2015. Normalised Difference Vegetation Index: Australia.

Burnett, R.T., Pope, C.A., Ezzati, M., Olives, C., Lim, S.S., Mehta, S., Shin, H.H., Singh, G., Hubbell, B., Brauer, M., 2014. An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. Environ Health Perspect. 122(4) 397-403.

Carnevale, C., Finzi, G., Pederzoli, A., Turrini, E., Volta, M., 2016. Lazy Learning based surrogate models for air quality planning. Environmental Modelling & Software 83 47-57.

Chan, C.-C., Chuang, K.-J., Chien, L.-C., Chen, W.-J., Chang, W.-T., 2006. Urban air pollution and emergency admissions for cerebrovascular diseases in Taipei, Taiwan. European heart journal 27(10) 1238-1244.

Chen, L., Verrall, K., Tong, S., 2006. Air particulate pollution due to bushfires and respiratory hospital admissions in Brisbane, Australia. International journal of environmental health research 16(03) 181-191.

Chu, D.A., Kaufman, Y., Zibordi, G., Chern, J., Mao, J., Li, C., Holben, B., 2003. Global monitoring of air pollution over land from the Earth Observing System□Terra Moderate Resolution Imaging Spectroradiometer (MODIS). Journal of Geophysical Research: Atmospheres (1984–2012) 108(D21).

Chudnovsky, A.A., Koutrakis, P., Kloog, I., Melly, S., Nordio, F., Lyapustin, A., Wang, Y., Schwartz, J., 2014. Fine particulate matter predictions using high resolution Aerosol Optical Depth (AOD) retrievals. Atmospheric Environment 89 189-198.

Chudnovsky, A.A., Lee, H.J., Kostinski, A., Kotlov, T., Koutrakis, P., 2012. Prediction of daily fine particulate matter concentrations using aerosol optical depth retrievals from the Geostationary Operational Environmental Satellite (GOES). Journal of the Air & Waste Management Association 62(9) 1022-1031.

Corani, G., Scanagatta, M., 2016. Air pollution prediction via multi-label classification. Environmental Modelling & Software 80 259-264.

Dirgawati, M., Barnes, R., Wheeler, A.J., Arnold, A.-L., McCaul, K.A., Stuart, A.L., Blake, D., Hinwood, A., Yeap, B.B., Heyworth, J.S., 2015. Development of Land Use Regression models for predicting exposure to NO2 and NOx in Metropolitan Perth, Western Australia. Environmental Modelling & Software 74 258-267.

Forouzanfar, M.H., Alexander, L., Anderson, H.R., Bachman, V.F., Biryukov, S., Brauer, M., Burnett, R., Casey, D., Coates, M.M., Cohen, A., 2015. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet 386(10010) 2287-2323.

Franklin, M., Zeka, A., Schwartz, J., 2007. Association between $PM_{2.5}$ and all-cause and specific-cause mortality in 27 US communities. Journal of Exposure Science and Environmental Epidemiology 17(3) 279-287.

Giavarina, D., 2015. Understanding Bland Altman analysis. Biochemia medica 25(2) 141-151.

Gupta, P., Christopher, S.A., 2009a. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. Journal of Geophysical Research: Atmospheres (1984–2012) 114(D20).

Gupta, P., Christopher, S.A., 2009b. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. Journal of Geophysical Research: Atmospheres (1984–2012) 114(D14).

Gupta, P., Christopher, S.A., Box, M.A., Box, G.P., 2007. Multi year satellite remote sensing of particulate matter air quality over Sydney, Australia. International Journal of Remote Sensing 28(20) 4483-4498.

Gupta, P., Christopher, S.A., Wang, J., Gehrig, R., Lee, Y., Kumar, N., 2006. Satellite remote sensing of particulate matter and air quality assessment over global cities. Atmospheric Environment 40(30) 5880-5892.

Han, L., Zhou, W., Li, W., 2015. Increasing impact of urban fine particles (PM2.5) on areas surrounding Chinese cities. Scientific reports 5 12467.

Harrison, R.M., Deacon, A.R., Jones, M.R., Appleby, R.S., 1997. Sources and processes affecting concentrations of PM10 and PM2.5 particulate matter in Birmingham (UK). Atmospheric Environment 31(24) 4103-4117.

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. Environmental Science & Technology 41(7) 2422-2428.

Hinwood, A., De Klerk, N., Rodriguez, C., Jacoby, P., Runnion, T., Rye, P., Landau, L., Murray, F., Feldwick, M., Spickett, J., 2006. The relationship between changes in daily air pollution and hospitalizations in Perth, Australia 1992–1998: a case-crossover study. International journal of environmental health research 16(1) 27-46.

Hoff, R.M., Christopher, S.A., 2009. Remote sensing of particulate pollution from space: have we reached the promised land? Journal of the Air & Waste Management Association 59(6) 645-675.

Hoffmann, B., Moebus, S., Möhlenkamp, S., Stang, A., Lehmann, N., Dragano, N., Schmermund, A., Memmesheimer, M., Mann, K., Erbel, R., 2007. Residential exposure to traffic is associated with coronary atherosclerosis. Circulation 116(5) 489-496.

Horsburgh, J.S., Aufdenkampe, A.K., Mayorga, E., Lehnert, K.A., Hsu, L., Song, L., Jones, A.S., Damiano, S.G., Tarboton, D.G., Valentine, D., 2016. Observations Data Model 2: A community information model for spatially discrete Earth observations. Environmental Modelling & Software 79 55-74.

Hu, Z., Follmann, D.A., Miura, K., 2015. Vaccine design via nonnegative lasso-based variable selection. Statistics in medicine 34(10) 1791-1798.

Jang, J.-S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. Systems, Man and Cybernetics, IEEE Transactions on 23(3) 665-685.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., Giovis, C., 2005a. A review and evaluation of intraurban air pollution exposure models. Journal of Exposure Science and Environmental Epidemiology 15(2) 185-204.

Jerrett, M., Burnett, R.T., Ma, R., Pope, C.A., Krewski, D., Newbold, K.B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E.E., Thun, M.J., 2005b. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology 16(6) 727-736.

Jones, D.A., Wang, W., Fawcett, R., 2009. High-quality spatial climate data-sets for Australia. Australian Meteorological and Oceanographic Journal 58(4) 233.

Kaufman, Y.J., Tanré, D., Boucher, O., 2002. A satellite view of aerosols in the climate system. Nature 419(6903) 215-223.

Kim, J.-H., 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics & Data Analysis 53(11) 3735-3745.

Kloog, I., Koutrakis, P., Coull, B.A., Lee, H.J., Schwartz, J., 2011. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. Atmospheric Environment 45(35) 6267-6275.

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. Environmental research 135 204-211.

Kondragunta, S., Lee, P., McQueen, J., Kittaka, C., Prados, A.I., Ciren, P., Laszlo, I., Pierce, R.B., Hoff, R., Szykman, J.J., 2008. Air quality forecast verification using satellite data. Journal of Applied Meteorology and Climatology 47(2) 425-442.

Kruse, R., Pasi, G., Alonso, J.M., 2013. Introduction to the Soft Computing and Intelligent Data Analysis Minitrack, 46th Hawaii International Conference on System Sciences, p. 1384.

Kunzli, N., Jerrett, M., Mack, W.J., Beckerman, B., LaBree, L., Gilliland, F., Thomas, D., Peters, J., Hodis, H.N., 2005. Ambient air pollution and atherosclerosis in Los Angeles. Environ Health Perspect. 113(2) 201-206.

Lee, J.-H., Wu, C.-F., Hoek, G., de Hoogh, K., Beelen, R., Brunekreef, B., Chan, C.-C., 2014. Land use regression models for estimating individual NOx and NO2 exposures in a metropolis with a high density of traffic roads and population. Science of The Total Environment 472 1163-1171.

Levelt, P.F., Van den Oord, G.H., Dobber, M.R., Mälkki, A., Visser, H., De Vries, J., Stammes, P., Lundell, J.O., Saari, H., 2006. The ozone monitoring instrument. Geoscience and Remote Sensing, IEEE Transactions on 44(5) 1093-1101.

Li, C., Hsu, N.C., Tsay, S.-C., 2011. A study on the potential applications of satellite data in air quality monitoring and forecasting. Atmospheric Environment 45(22) 3663-3675.

Li, Q., Shao, J., 2015. Regularizing lasso: a consistent variable selection method. Statistica Sinica 25 975-992.

Lin, C.-T., Lee, C.S.G., 1991. Neural-network-based fuzzy logic control and decision system. IEEE Transactions on computers 40(12) 1320-1336.

Lin, C., Li, Y., Yuan, Z., Lau, A.K.H., Li, C., Fung, J.C.H., 2014. Using satellite remote sensing data to estimate the high-resolution distribution of ground-level PM2.5. Remote Sensing of Environment 156 117-128.

Lindstrom, J., Szpiro, A.A., Sampson, P.D., Sheppard, L., Oron, A.P., Richards, M., Larson, T., 2011. A Flexible Spatio-Temporal Model for Air Pollution: Allowing for Spatio-Temporal Covariates. Berkeley Electronics Press.

Liu, Y., Paciorek, C.J., Koutrakis, P., 2009. Estimating regional spatial and temporal variability of PM2.5 concentrations using satellite data, meteorology, and land use information.

Lüdtke, N., Panzeri, S., Brown, M., Broomhead, D.S., Knowles, J., Montemurro, M.A., Kell, D.B., 2008. Information-theoretic sensitivity analysis: a general method for credit assignment in complex networks. Journal of The Royal Society Interface 5(19) 223-235.

Lunn, H., Johnston, C., Flavel, R., 2002. The Vision and Living Skills Research Project: Meeting the challenge of intervention in urban and rural communities, 11th ICEVI World Conference: Noordwijkerhout, the Netherlands.

Maciejewska, K., Juda-Rezler, K., Reizer, M., Klejnowski, K., 2015. Modelling of black carbon statistical distribution and return periods of extreme concentrations. Environmental Modelling & Software 74 212-226.

McBride, S.J., Williams, R.W., Creason, J., 2007. Bayesian hierarchical modeling of personal exposure to particulate matter. Atmospheric Environment 41(29) 6143-6155.

Meyer, C.M., Luhar, A.K., Mitchell, R.M., 2008. Biomass burning emissions over northern Australia constrained by aerosol measurements: I—Modelling the distribution of hourly emissions. Atmospheric Environment 42(7) 1629-1646.

Moazami, S., Noori, R., Amiri, B.J., Yeganeh, B., Partani, S., Safavi, S., 2016. Reliable prediction of carbon monoxide using developed support vector machine. Atmospheric Pollution Research 7(3) 412-418.

NASA, 2013. Global Change Master Directory, Earth Science Keywords, Aerosol and Warming, 8th ed: USA.

Novotny, E.V., Bechle, M.J., Millet, D.B., Marshall, J.D., 2011. National satellite-based land-use regression: NO2 in the United States. Environmental Science & Technology 45(10) 4407-4414.

Ordieres, J., Vergara, E., Capuz, R., Salazar, R., 2005. Neural network prediction model for fine particulate matter (PM2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environmental Modelling & Software 20(5) 547-559.

Ovaska, S.J., 2004. Computationally Intelligent Hybrid Systems: The Fusion of Soft Computing and Hard Computing (IEEE Press Series on Computational Intelligence). Wiley-IEEE Press.

Pandey, G., Zhang, B., Jian, L., 2013. Predicting submicron air pollution indicators: a machine learning approach. Environmental Science: Processes & Impacts 15(5) 996-1005.

Peters-Lidard, C.D., Kemp, E.M., Matsui, T., Santanello, J.A., Kumar, S.V., Jacob, J.P., Clune, T., Tao, W.-K., Chin, M., Hou, A., 2015. Integrated modeling of aerosol, cloud, precipitation and land processes at satellite-resolved scales. Environmental Modelling & Software 67 149-159.

Queensland Government., 2014. Air Quality Monitoring. Queensland Government: Australia.

Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validation, Encyclopedia of database systems. Springer, pp. 532-538.

Reid, C.E., Jerrett, M., Petersen, M.L., Pfister, G.G., Morefield, P.E., Tager, I.B., Raffuse, S.M., Balmes, J.R., 2015. Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning. Environmental Science & Technology 49(6) 3887-3896.

Reis, S., Seto, E., Northcross, A., Quinn, N.W., Convertino, M., Jones, R.L., Maier, H.R., Schlink, U., Steinle, S., Vieno, M., 2015. Integrating modelling and smart sensors for environmental and human health. Environmental Modelling & Software 74 238-246.

Rose, N., Cowie, C., Gillett, R., Marks, G.B., 2010. Validation of a spatiotemporal land use regression model incorporating fixed site monitors. Environmental Science & Technology 45(1) 294-299.

Salam, M.T., Islam, T., Gilliland, F.D., 2008. Recent evidence for adverse effects of residential proximity to traffic sources on asthma. Current opinion in pulmonary medicine 14(1) 3-8.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global sensitivity analysis: the primer. John Wiley & Sons.

Samet, J.M., 2007. Traffic, air pollution, and health. Inhalation toxicology 19(12) 1021-1027.

Shu, Y., Liu, Q., Taylor, K., 2016. Semantic validation of environmental observations data. Environmental Modelling & Software 79 10-21.

Simpson, R., Williams, G., Petroeschevsky, A., Best, T., Morgan, G., Denison, L., Hinwood, A., Neville, G., Neller, A., 2005. The short□term effects of air pollution on daily mortality in four Australian cities. Australian and New Zealand Journal of Public Health 29(3) 205-212.

Strawa, A., Chatfield, R., Legg, M., Scarnato, B., Esswein, R., 2013. Improving Retrievals of Regional PM2.5 Concentrations From MODIS and OMI Multi-Satellite Observations, AGU Fall Meeting Abstracts, p. 0300.

Su, J., Jerrett, M., Beckerman, B., 2009. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. Science of The Total Environment 407(12) 3890-3898.

Syed, A.R., 2011. A review of cross validation and adaptive model selection, Department of Mathematics and Statistics. Georgia State University.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(3) 273-282.

Torres, O., Tanskanen, A., Veihelmann, B., Ahn, C., Braak, R., Bhartia, P.K., Veefkind, P., Levelt, P., 2007. Aerosols and surface UV products from Ozone Monitoring Instrument observations: An overview. Journal of Geophysical Research: Atmospheres (1984–2012) 112(D24).

United Nations, 2015. World Population Prospects: The 2015 Revision. Department of Economic and Social Affairs, Population Division.

Vienneau, D., de Hoogh, K., Bechle, M.J., Beelen, R., van Donkelaar, A., Martin, R.V., Millet, D.B., Hoek, G., Marshall, J.D., 2013. Western European land use regression incorporating satellite-and ground-based measurements of NO2 and PM10. Environmental Science & Technology 47(23) 13555-13564.

Wu, Y., Guo, J., Zhang, X., Tian, X., Zhang, J., Wang, Y., Duan, J., Li, X., 2012. Synergy of satellite and ground based observations in estimation of particulate matter in eastern China. Science of The Total Environment 433 20-30.

Yeganeh, B., Motlagh, M.S.P., Rashidi, Y., Kamalan, H., 2012. Prediction of CO concentrations based on a hybrid partial least square and support vector machine model. Atmospheric Environment 55 357-365.