Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level $PM_{2.5}$

Authors: Yongming Xu[1], Hung Chak Ho[2], Man Sing Wong[2,3], Chengbin Deng[4], Yuan Shi[5], Ta-Chien Chan[6], Anders Knudby[7]

1. School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science & Technology, Nanjing, China
2. Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong
3. Research Institute for Sustainable Urban Development, The Hong Kong Polytechnic University, Hong Kong
4. Department of Geography, State University of New York at Binghamton, Binghamton, New York, United States
5. School of Architecture, Chinese University of Hong Kong, New Territories, Hong Kong
6. Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan
7. Department of Geography, Environment and Geomatics, University of Ottawa, Ottawa, ON, Canada

Corresponding Author: Hung Chak Ho, Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong

Research Highlights

- Estimation of long-term spatially-continuous monthly $PM_{2.5}$ dataset
- Cubist outperforms other machine learning algorithms
- Several new predictors were employed to improve the estimation of $PM_{2.5}$
- $PM_{2.5}$ was estimated with a CV-RMSE of 2.64 $\mu g/m^3$

Abstract

Fine particulate matter ($PM_{2.5}$) has been recognized as a key air pollutant that can influence population health risk, especially during extreme cases such as wildfires. Previous studies have applied geospatial techniques such as land use regression to map the ground-level $PM_{2.5}$, while some recent studies have found that Aerosol Optical Depth (AOD) derived from satellite images and machine learning techniques may be two elements that can improve spatiotemporal prediction. However, there has been a lack of studies evaluating use of different machine learning techniques with AOD datasets for mapping $PM_{2.5}$, especially in areas with high spatiotemporal variability of $PM_{2.5}$.

In this study, we compared the performance of eight predictive algorithms with the use of multiple remote sensing datasets, including satellite-derived AOD data, for the prediction of ground-level $PM_{2.5}$ concentration. Based on the results, Cubist, random forest and eXtreme Gradient Boosting were the algorithms with better performance, while Cubist was the best (CV-RMSE=2.64 μg/m^3, CV-R^2=0.48). Variable importance analysis indicated that the predictors with the highest contributions in modelling were monthly AOD and elevation.

In conclusion, appropriate selection of machine learning algorithms can improve ground-level $PM_{2.5}$ estimation, especially for areas with nonlinear relationships between $PM_{2.5}$ and predictors caused by complex terrain. Satellite-derived data such as AOD and land surface temperature (LST) can also be substitutes for traditional datasets retrieved from weather stations, especially for areas with sparse and uneven distribution of stations.

1. Introduction

Fine particulate matter ($PM_{2.5}$) is one of the major dust-related air pollutants that can increase morbidity and mortality risks, especially for cardiovascular and respiratory issues (Atkinson et al., 2014). In order to reduce community health risks caused by environmental exposure, previous studies have commonly applied air quality data from single or a small number of monitoring stations to evaluate the temporal influences of $PM_{2.5}$ (Liu et al., 2018; Ostro et al., 2014; Wang et al., 2017), and have found positive association between $PM_{2.5}$ and chronic diseases. These results have helped pinpoint air pollution as a severe community health problem (Kan et al., 2012). However, sparse distribution of air quality monitoring stations across large areas reduces the ability to demonstrate the actual impact of $PM_{2.5}$ on all vulnerable populations.

Satellite remote sensing data can provide spatially continuous estimates of aerosol optical depth (AOD), providing an alternative method to map ground-level $PM_{2.5}$ across a large region. Since AOD from satellite images has complete spatial coverage and moderate spatial resolution, AOD measurement can fill in data for areas that lack monitoring stations. Multiple studies have been carried out to estimate $PM_{2.5}$ from satellite-derived AOD and other environmental variables (Lai et al., 2014; Saunders et al., 2014; Wu et al., 2015). Due to the spatio-temporal heterogeneity of AOD-$PM_{2.5}$ relationships, using AOD to directly represent ground-level $PM_{2.5}$ may be inappropriate, as has been reported by previous studies (Lee et al., 2011; Paciorek et al., 2008). Additional environmental predictors, such as geographical and meteorological variables, have also been incorporated in models to improve estimation performance (Hu et al., 2013; Kloog et al., 2011; Liu et al., 2009). To derive $PM_{2.5}$ from satellite-derived AOD and other predictors, various models have been developed. The most commonly used models include multiple linear regression (Lai et al., 2014; Liu et al., 2004; Saunders et al., 2014; Schaap et al., 2009; Yao et al., 2018a), mixed effect models (Just et al., 2015; Lee et al., 2011; Zheng et al. 2016; Xie et al., 2015), chemical transport models (Crouse et al., 2016; Wang & Chen, 2016; van Donkelaar et al., 2006) and geographically weighted regression (Chu et al., 2015; Chu et al., 2016; He and Huang, 2018; Jiang et al., 2017; Ma et al., 2014; Shi et al., 2018; Song et al., 2014; Wu et al., 2016; You et al., 2016). Recently, machine learning technology, which can fit complicated non-linear relationships in many dimensions, has also been employed to derive air-pollutant concentrations from remote sensing data (Chen et al., 2018; Deters et al., 2017, He & Huang, 2018, Yao et al., 2018b). Several machine learning methods, such as artificial neural networks, generalized boosting models, support vector machine and random forest, have also been used to generate models for estimating $PM_{2.5}$ (Di et al., 2016; Hu et al., 2017; Reid et al., 2015; Zhan et al., 2017). However, to date, studies with machine learning for estimating $PM_{2.5}$ are still rare in this field.

In order to better understand the potential of machine learning for $PM_{2.5}$ mapping, we developed an innovative approach to estimate spatial variability of $PM_{2.5}$ by using machine learning techniques with multiple predictors based on Moderate Resolution Imaging Spectroradiometer (MODIS) and re-analysis data. By using machine learning techniques, it can better characterize non-linear relationships for estimating air pollution based on all geophysical components. To enhance the ability to develop a spatiotemporal model for $PM_{2.5}$ prediction, the specific objectives of this study included 1) to develop a model for predicting $PM_{2.5}$ based on remote sensing data, re-analysis data and station observed air quality data; 2) to evaluate the prediction performance of different statistical methods, for determining the best model setting for estimating $PM_{2.5}$; and 3) to map the spatio-temporal distribution of $PM_{2.5}$ based on the best model. British Columbia of Canada was selected as the case of this study, because of its complex terrain and wildfire history that can significantly influence air quality across the province, including $PM_{2.5}$.

2. Study Area

British Columbia (BC) is the westernmost province of Canada (Fig. 1), and it is characterized by mountainous terrain and heavy forest cover. BC has traditionally been known for its clean environment. However, due to climate change, increasing frequency of wildfires has been observed in recent decades (Wildfire Management Branch, 2014; Wotton, 2010). Wildfires produce excessive smoke that can influence regional air quality and severely affect human health (Henderson et al., 2011; McLean et al., 2015; Krstic & Henderson et al., 2015). In order to minimize air pollution risk, a National Air Pollution Surveillance (NAPS) system with ground-based stations has been established across the province, monitoring temporal changes in air pollutants including the daily change in $PM_{2.5}$. However, due to the province's sprawling territory with complex terrain and a limited number of surveillance stations, station-based observation may not be able to adequately measure the $PM_{2.5}$ influencing all populated regions (McLean et al. 2015). The stations with data between 2001 and 2014 were sparsely distributed and clustered in the southern and central parts of BC. Therefore, combining satellite images to monitor the spatiotemporal changes in $PM_{2.5}$ across the province is essential.

3. Data and Methods

3.1 Selection of predictors for $PM_{2.5}$ mapping

According to previous studies, AOD has strong positive relationships with ground-level $PM_{2.5}$ concentrations (Engel-Cox et al., 2004; Mukai et al., 2006; Wang & Christopher, 2003; Xin et al., 2014), and some studies have applied satellite-derived AOD to map $PM_{2.5}$ (Chu et al., 2016). Therefore, AOD was the first predictor for $PM_{2.5}$ mapping. In this study, AOD data were retrieved from MOD04_3K, a 3-km near-real-time aerosol dataset derived from TEAAR/MODIS.

The $PM_{2.5}$-AOD relationship can be a multivariate function of a wide range of influencing factors (Lary et al., 2015; Natunen et al., 2010; Song et al., 2014; van Donkelaar et al., 2006). For example, meteorological and geographical predictors can be the parameters of co-predicting $PM_{2.5}$ concentrations (Jiang et al., 2017; Liu et al., 2009; Ma et al., 2014; Reid et al., 2015; You et al., 2016). Built on the literature, the following parameters may contribute to $PM_{2.5}$ prediction: humidity, temperature, albedo, normalized difference vegetation index (NDVI), height of the planetary boundary layer (HPBL), wind speed, distance to the ocean, elevation, and calendar month. Therefore, we constructed the input datasets for modelling as follows.

Considering the bias which sparse distribution of weather stations may produce in data representing spatial variations in temperature and humidity, 26855 images of MODIS land surface temperature product (MOD11A1) and 44336 images of MODIS water vapor product (MOD05_L2) were used as alternatives to air temperature and relative humidity for better spatial representativeness. In brief, MOD11A1 is a 1-km daily land surface temperature (LST) product derived from TERRA/MODIS, and MOD05_L2 is a 1-km near-real-time water vapor product derived from TERRA/MODIS.

In addition, NDVI and albedo were derived based on MODIS products: the MODIS vegetation index product (MOD13A3), a 1-km monthly vegetation index product derived from TERRA/MODIS; and the MODIS albedo product (MCD43B3), a 1-km 8-day albedo product derived from TERRA/MODIS and AQUA/MODIS. For the mapping purpose, all MODIS datasets were re-projected to the Albers projection, resampled to 1-km spatial resolution, and averaged for each month.

Finally, HPBL and wind speed were derived from NCAR/NCEP re-analysis data, which provides the corresponding data on a monthly basis. Elevation was derived from a digital elevation model (DEM) dataset of the Shuttle Radar Topography Mission (SRTM). Distance to the ocean was calculated by buffer analysis based on the coastal boundary of BC.

Based on the satellite-derived products and re-analysis data, a total of 10 predictors were employed to estimate ground-level $PM_{2.5}$ concentration across BC: monthly AOD, monthly vapor, monthly LST, monthly NDVI, monthly albedo, monthly HPBL, monthly wind speed, elevation, distance to ocean and calendar month (Table 1).

It is known that the relationship between environmental predictors and $PM_{2.5}$ may vary across space (Hu et al., 2013; Song et al., 2014), as well as time. We did not include spatial predictors (e.g. latitude, longitude) other than “distance to ocean”, and we did not use spatially weighted models such as geographically weighted regression, because of the limited insight that can be gained from using such predictors/models, and the limited transferability such models will have to other geographical regions.

3.2 Model development with machine learning algorithms

Association between $PM_{2.5}$ concentration of air quality monitoring stations and the values of predictors retrieved by the locations of stations were first established for each machine learning model in order to estimate the spatial distributions of ground-level $PM_{2.5}$ concentrations. In this study, ground-level $PM_{2.5}$ concentrations for modelling were retrieved from 63 stations of the NAPS network operated by Environment Canada, with hourly $PM_{2.5}$ data between 2001 and 2014 across BC. Since several stations within this study period did not provide temporal-continuous observations, or even had significant data gaps in temporal observation, we averaged hourly $PM_{2.5}$ data on a daily basis, then converted the daily information to the monthly average $PM_{2.5}$ concentrations based on all valid daily values.

These monthly average $PM_{2.5}$ values across BC province were then applied to the following statistic algorithms to construct the regression models: 1) multiple linear regression (MLR), 2) Bayesian Regularized Neural Networks (BRNN), 3) Support Vector Machines with Radial Basis Function Kernel (SVM), 4) Least Absolute Shrinkage and Selection Operator (LASSO), 5) Multivariate Adaptive Regression Splines (MARS), 6) Random forest (RF), 7) eXtreme Gradient Boosting (XGBoost), and 8) Cubist.

MLR is a widely used algorithm in remote sensing applications because of its simplicity, but it relies on several assumptions concerning data distributions, and its performance depends on meting these assumptions as well as the linearity of the modeled relationship (Helsel and Hirsch 1992). BRNN is a back-propagation network that based on a mathematical technique named Bayesian regularization to convert nonlinear regression into "well-posed" problems (Burden and Winkler, 2008). It is more robust than standard back-propagation neural networks. SVM was originally developed for classification by constructing separating hyperplanes to define decision boundaries, and later expanded for regression. To map samples to high dimension space, kernel functions were introduced. The radial basis function showed its advances of handling nonlinear problems and fewer tunable parameters (Hsu, 2003; Bennett and Campbell, 2000). LASSO is a regularization and variable selection method which shrinks coefficients by forcing some less important coefficients to zero (Tibshirani, 1996). It can improve the model interpretability and reduce overfitting. MARS is a fully automated method based on the divide-and-conquer strategy, in which the training dataset is split into piecewise linear segments (splines) (Friedman, 1991). RF is an ensemble-based decision tree approach, which consists of a combination of decision trees fitted by randomly selected subsets of training samples. Final predictions produced by RF model are determined by the average of the results of all the trees (Breiman, 2001). XGBoost is an ensemble tree method which follows the principle of Gradient boosting framework (Friedman, 2001), and uses regularization techniques to control overfitting and model complexity (Chen and Guestrin, 2016). Cubist is a rule-based tree model, which produces multiple linear regression models in the terminal nodes of trees based on the M5 theory (Quinlan, 1992; RuleQuest, 2018). A prediction at the terminal node is made by the corresponding linear regression model and is smoothed by combining with predictions from nearest-neighbor nodes within the tree to improve prediction accuracy (Houborg & McCabe, 2018). In addition, Cubist also constructs multiple tree models (called committees), each of which consists of a set of rule-based models (John et al., 2018). Predictions from all the committees are averaged to produce the final prediction.

Except for the widely-used traditional MLR algorithm, others were machine learning algorithms, which can effectively fit nonlinear and complex relationships between outcomes and predictors (Ngufor et al. 2015). In this study, the complex terrain of the study area can form a nonlinear relationship between ground-level $PM_{2.5}$ concentrations and all predictors, for which machine learning models may provide better results.

In order to optimize the $PM_{2.5}$ estimation, parameter values were adjusted in each machine learning model with a fitting process, based on the determination of the best parameters by cyclic testing with committees of 1, 5, 10, 20, 50, and neighbors of 0, 1, 5, 9. In addition, predictions of $PM_{2.5}$ concentrations with all machine learning models were conducted with the R (R Development Core Team).

3.3 Model evaluation

10-fold cross-validation was performed to evaluate the accuracy of all machine learning models. Data were first randomly divided into 10 subsets, with one of the subsets used as the validation dataset and the remaining used as training datasets; then repeating 10 times until all subsets have been used as validation datasets once. Root-mean-square error (CV-RMSE) and coefficient of determination (CV-R^2) based on the comparison of validation and training data were used to evaluate the accuracy of each machine learning model. While the best model for $PM_{2.5}$ estimation was determined based on the accuracies, variable importance analysis was also conducted to evaluate the contributions of each predictor in $PM_{2.5}$ estimation, based on the determination of percentage increase in mean square error (%IncMSE) of each model relative to the original error, after a predictor was randomly permuted. A higher value of %IncMSE indicated higher importance of this corresponding predictor to the estimation.

4. Results

4.1 Empirical relationship between $PM_{2.5}$ and AOD

A total of 1242 records of observed data of ground-level $PM_{2.5}$ concentrations were retrieved from stations with effective monthly AOD values based on location. In brief, $PM_{2.5}$ concentrations of this subset ranged from 1.26μg/m^3 to 51.14μg/m^3, with an average of 5.26μg/m^3 and a median of 4.58μg/m^3. This indicated a clean environment with low air pollution during the study period across BC, except in a few extreme cases. Based on the observed data, the extremes in $PM_{2.5}$ concentration samples were observed in August 2003 and August 2010, when there were wildfire events (e.g. 2003 Okanagan Mountain Park Fire) across BC.

A positive but poor correlation was observed based on evaluation of an empirical relationship between observed $PM_{2.5}$ and satellite-derived AOD (Fig. 2), with a correlation coefficient (R) of 0.34 (P-value < 0.01), a clustering of data was found with AOD value less than 0.8 and $PM_{2.5}$ value less than 15μg/m^3. Observed data with moderate or high values were scattered, possibly due to the complexity of the atmospheric conditions and landscapes across BC. Similar evidence has also been found in a previous study, which demonstrated a non-linear relationship between geophysical environment and air temperature across BC (Xu et al., 2014). Therefore, the use of simple linear regression for ground-level $PM_{2.5}$ estimation is insufficient and inaccurate, and nonlinear multivariate models should be adopted to predict $PM_{2.5}$ under consideration of relevant atmosphere-surface interactions.

4.2 Model performance

Parameters of machine learning models were optimized with the fitting process, by cyclic testing with a given parameter range and step size. Based on the results of optimized models, CV-RMSE ranged from 2.64 μg/m^3 to 3.24 μg/m^3 and CV-R^2 ranged from 0.22-0.49 (Table 2). Among all, RF, XGBoost and Cubist were the models with better performance, while Cubist had the best performance determined by CV-RMSE. With 20 committees and 5 neighbors as optimal parameters, CV-RMSE and CV-R^2 of Cubist were 2.64 μg/m^3 and 0.48. In contrast, MLR method had the lowest performance (CV-RMSE=3.24 μg/m^3 and CV-R^2=0.22), indicating its poor capability of capturing complex relationships for the study area.

For the best model, the predicted and observed values were well aligned with the line of best fit (Fig. 3), indicating the high accuracy of $PM_{2.5}$ estimation with Cubist. However, underestimation was also found for observed data with high $PM_{2.5}$ values (> 20μg/m^3), possibly due to the small sample size, resulting in inability to robustly predict these high-value data with a decision-based machine learning algorithm. Moreover, average deviation of $PM_{2.5}$ estimation was 0.07μg/m^3, slightly higher than the deviation of observed values. These results show that lower $PM_{2.5}$ concentration in observed data may result in overestimation, while higher values in observed data might result in underestimation during prediction.

4.3 Variable importance analysis

Based on the variable importance analysis, the predictors with highest contributions to the Cubist model were monthly AOD and elevation. %IncRMSE without monthly AOD as predictor was 12.14%, possibly due to its strong association between AOD and ground-level air quality. %IncRMSE without elevation as a predictor was 9.26%, also suggesting a high importance in $PM_{2.5}$ estimation because of the influences of complex terrain in BC, with great variations in altitude between the coast and interior. However, there shall be several factors which contributed to the importance of elevation for predictions of $PM_{2.5}$: areas with high elevation are inclined to suffer from wildfires; areas with low elevation tend to be influenced by human activities. As AOD is an important predictor in the models, elevation may be used to correct for model predictions. In addition, %IncRMSE of monthly albedo, monthly LST and calendar month ranged from 4% to 6%. Predictors with the least importance were monthly wind speed, monthly HPBL, monthly vapor and monthly NDVI, with a range of %IncRMSE between 2% and 4%.

4.4 Determination of location-based error

To further determine the spatial variability of error, RMSEs were extracted by the location of each station (Fig. 4). Most stations had RMSEs lower than 2.0μg/m^3, while the stations with the lowest RMSEs were in southeastern, western and southwestern BC. In contrast, high errors were found at stations located in central and central-southern parts of BC, with RMSEs ranging from 3.0 - 4.0μg/m^3 or even higher. Compared with the DEM, these stations with higher RMSEs were in mountainous valleys with high $PM_{2.5}$ concentrations. Estimation errors of these stations were mostly negative, indicating an underestimation of ground-level $PM_{2.5}$ across these valleys. These were also aligned with previous findings (Fig. 3) that observed data with higher $PM_{2.5}$ may introduce a higher chance of underestimation based on the Cubist model in this study.

5 Discussion

5.1 Spatiotemporal variability of ground-level $PM_{2.5}$ concentration

Based on the average concentrations of ground-level $PM_{2.5}$ between 2001 and 2014 (Fig. 6), considerable spatial heterogeneity was found across BC. Generally, northern and northeastern BC were areas with lower $PM_{2.5}$ concentrations (< 4 μg/m^3), while mountainous regions across western BC were areas with higher concentrations of $PM_{2.5}$ (5-6 μg/m^3). We also observed several extreme cases in mountainous valleys of BC (>7 μg/m^3). One reason for this spatiotemporal variability might be associated with wildfires, as this was a major source of ambient $PM_{2.5}$ across mountainous BC. Previous studies have found a particular deposition process of $PM_{2.5,}$ emitted from biomass burning, with long-distance transport (Ward et al., 1991; Sapkota et al., 2005). We should emphasize that terrain can play an influential role in the deposition, due to the aerodynamic characteristics of $PM_{2.5}$ and the topographical effect on wind flow. For example, the mountainous topography of BC, with its irregular terrain, can result in uneven distribution of air pressure that further influences near-surface wind. The effect of local terrain on $PM_{2.5}$ dispersion due to its impact on wind dynamics has also been found in another study in mountainous areas (Shi et al., 2017). A considerable fraction of $PM_{2.5}$ is therefore expected to be trapped by the leeward side of mountains, valleys, canyons and basins (Steyn et al., 2013) under the typical transport process of air pollutants. Urban areas with high aerodynamic surface roughness may also have influence similar to this topographical effect on the deposition of $PM_{2.5}$ from wildfires (Landsberg, 1981). These findings indicate that regions across BC with lower altitude and with poorer air dispersion due to topographical effects may be areas with higher $PM_{2.5}$

concentration. In addition, these facts may also partly explain the lower contribution of monthly coarse spatial resolution (2.5 degree latitude x 2.5 degree longitude) and monthly wind speed in modelling based on variable importance analysis, while another reason may be the coarse spatial resolution (2.5 degree) of predictors derived from NCEP/NCAR re-analysis data. Due to this resolution, it cannot represent micro-scale topographical effects on air pollution transport and deposition. Some mountain valleys in BC have high temperatures and little rainfall during the summer, and become dry enough to have near-desert conditions with substantial amounts of dust suspended in the atmosphere, which is also contributed to the high PM2.5 concentrations of valleys. An isolated cluster of high $PM_{2.5}$ in the Greater Vancouver Area and its surrounding regions was also observed, which has not been shown in other BC cities. This can be attributed to the large population and corresponding industrial, traffic and domestic emissions over this region.

Furthermore, CV-RMSE of this study was lower than previous research in other areas (Liu et al., 2009; Song et al., 2014; Kloog et al., 2014; You et al., 2015; Reid et al., 2015; Liu et al., 2005), partially indicating better air quality of BC compared to other regions. In contrast, a lower CV-R^2 was found, which may be the result of extreme wildfire events in BC leading to data with high $PM_{2.5}$ concentration values as outliers in modelling.

5.2 Advantages and Limitations

In this study, optimization of machine learning models can effectively reduce the sensitivity of the model tree to data noise with uncertainty; while the evaluation of eight machine learning algorithms for modelling indicated that ensemble machine learning can improve the accuracy of ground-level $PM_{2.5}$ prediction. In addition, weather stations were generally designed under government protocols, resulting in a sparse and uneven distribution. This, as well as the strong variation in topography across the study area, makes it unsuitable to apply conventional geostatistical methods such as spatial interpolation for mapping the spatial variability of environmental variables (e.g. temperature and humidity), while these maps should be the input layers for air quality prediction. In this study, we provided an alternative, in which the use of LST and atmospheric water vapor derived from satellite images can be substitutes for temperature and humidity maps.

There were areas with data missing from the prediction (Fig. 6). These were mainly the high-altitude areas covered with perennial snow, because the Dark Target algorithm for AOD retrieval was designed for areas with lower surface reflectance under a clear sky. For areas with high surface reflectance values (e.g. snow coverage and desert), null values of AOD data would be found. In addition, AOD values surrounding the missing data were generally high, because AOD in such areas could be easily overestimated by the Dark Target algorithm, especially in areas with high surface brightness and low vegetation coverage (Levy et al., 2010). These became the areas with missing values of $PM_{2.5}$ concentration across snow coverage in this study, and there were extremely high values of $PM_{2.5}$ concentration surrounding these areas with missing data, especially those areas just below the snowline with lower vegetation coverage. The issue of missing data is especially noticeable in winter, as mountainous BC was covered by snow, resulting in high surface reflectance, and this area was also constantly covered by clouds due to the relatively humid weather in wintertime, resulting in spatiotemporal incompleteness of $PM_{2.5}$ estimation.

In addition, the $PM_{2.5}$ concentration over BC showed high values both in western high mountains and the Fraser River Delta. The principal sources of $PM_{2.5}$ is likely different between these areas. In mountain areas high $PM_{2.5}$ concentration is mostly caused by wildfires, while in the Fraser River Delta high $PM_{2.5}$ concentration is caused by human activity. Due to the lack of the chemical characteristics of particulate matter, we cannot perform a chemical analysis of fine particulate matter over these regions. Further study with field measurement should be applied to observe personal and ambient exposure of $PM_{2.5}$ from multiple sources. However, this future study will be limited by the accessibility of field measurement and the potential bias from indoor-outdoor exchange of air pollution.

6 Conclusions

In this study, we evaluated the abilities of machine learning techniques to estimate the monthly concentrations of ground-level $PM_{2.5}$ between 2001 and 2014, based on eight algorithms with predictors derived from remote sensing and meteorological re-analysis data. Predictions from these algorithms were evaluated by a 10-fold cross-validation, with CV-RMSE ranging from 2.64μg/m^3 to 3.25μg/m^3 and CV-R^2 ranging from 0.23-0.49. Among all, Cubist had the best performance (CV-RMSE=2.64μg/m^3, CV-R^2=0.48). A series of maps were produced for representing the monthly $PM_{2.5}$ concentrations across BC, which can be reference information on intra-province air pollution over 14 years for further air quality monitoring and public health surveillance. In conclusion, selection of appropriate machine learning algorithms for modelling can improve the accuracy in $PM_{2.5}$ estimation, while using satellite-derived data as predictors can minimize the spatial bias compared with use of traditional datasets retrieved from weather stations.

Recently, deep learning technology has attracted much attention in various fields.

Compared with conventional machine learning technology, deep learning can provide better accuracy but requires a large amount of training data (Camilleri and Prescott, 2017; Ravì et al., 2017). Due to the limited number of air quality stations, there are not enough samples to sufficiently train deep learning models. Therefore it is a big challenge to adopt deep learning technology to map $PM_{2.5}$ at the present stage. In the future, if the big training data requirement of deep learning can be resolved, it is expected to achieve improved estimation of $PM_{2.5}$ concentration from remote sensing data. The method used in this study with the combination of machine learning and multi-source variables was a preliminary attempt to map $PM_{2.5}$ concentration with the currently available data and suitable machine learning methods. The method proposed in this paper could also be applied to other complex terrain regions with sparse distributed air quality stations. Due to the limitation of AOD retrieval algorithms, the remotely sensed AOD data have coarse spatial resolutions. Re-analysis data have even coarser resolutions. The low spatial resolution of datasets restricts the application of this method on a small scale (e.g. city scale).

Acknowledgments

This work was supported by the Social Sciences Foundation of the Ministry of Education of China (Grant No. 17YJCZH205) and the National Key Research and Development Program of China (2017YFB0503903-4). We would like to thank the Land Processes Distributed Active Archive Center (LPDAAC) and Level-1 and Atmosphere Archive & Distribution System (LAADS) for providing MODIS data, US Geological Survey (USGS) for providing SRTM/DEM data, and National Oceanic and Atmospheric Administration (NOAA)/ Earth System Research

Laboratory (ESRL) for providing NCEP Reanalysis data. Man Sing Wong thanks the support in part by a grant from the General Research Fund (project ID: 15205515); and a grant of PolyU 1-ZVFD from the Research Institute for Sustainable Urban Development, the Hong Kong Polytechnic University. We also thank the two reviewers for their valuable comments and suggestions.

References

Atkinson, R. W., Kang, S., Anderson, H.R., Mills, I.C., Walton, H.A., 2014. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. Thorax 69, 660–665

B.C. Wildfire Management Branch. 2014. Proactive Wildfire Threat Reduction. Accessed June 15, 2017. http://docs.openinfo.gov.bc.ca/d63519414a_response_package_fnr-2014-00274.pdf

Bennett, K.P., Campbell, C., 2000. Support vector machines: hype or hallelujah? SIGKDD Explor. 2, 1–13.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Burden, F., Winkler, D., 2008. Bayesian regularization of neural networks. Methods Mol. Biol. 458, 25–44.

Camilleri, D. Prescott, T., 2017. Analysing the limitations of deep learning for developmental robotics. In: Biomimetic and Biohybrid Systems. 6th International Conference, Living Machines 2017, Stanford, CA, USA.

Chen, B., Song, Y., Jiang, T., Chen, Z., Huang, B., Xu, B., 2018. Real-time estimation of population exposure to PM2.5 using mobile-and station-based big data. Int. J. Environ. Res. Public Health 15, 573.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-789.

Chu, H.J., Huang, B., Lin, C.Y., 2015. Modeling the spatio-temporal heterogeneity in the PM10-PM2.5 relationship. Atmos. Environ. 102, 176–182.

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., Liu, F., Tian, L., Zhu, Z., Xiang, H., 2016. A review on predicting ground PM2.5 concentration using satellite aerosol optical depth. Atmosphere 7, 129.

Crouse, D.L., Philip, S., van Donkelaar, A., Martin, R.V., Jessiman, B., Peters, P.A., Weichenthal, S., Brook, J.R., Hubbell, B., Burnett, R.T., 2016. A new method to jointly estimate the mortality risk of long-term exposure to fine particulate matter and its components. Sci. Rep. 6, 18916.

Deters, J.K., Zalakeviciute, R., Gonzalez, M., Rybarczyk, Y., 2017. Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters. J. Elect. Comput. Eng. 2017, 1-14

Di, Q., Koutrakis, P., Schwartz, J., 2016. A hybrid prediction model for PM2.5 mass and components using a chemical transport model and land use regression. Atmos. Environ. 131, 390–399.

Engel-Cox, J.A., Holloman, C.H., Coutant, B.W., Hoff, R.M., 2004. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality.

Atmos. Environ. 38, 2495–2509.

Friedman, J.H., 1991. Multivariate Adaptive Regression Splines. Ann. Stat. 19, 1–67.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.

He, Q., Huang, B., 2018. Satellite-based high-resolution PM2.5 estimation over the Beijing-Tianjin-Hebei region of China using an improved geographically and temporally weighted regression model. Environ. Pollut. 236, 1027–1037.

Helsel, D. R., Hirsch, R. M., 1992. Statistical Methods in Water Resources, 296–299. Amsterdam: Elsevier.

Henderson, S.B., Brauer, M., MacNab, Y.C., Kennedy, S.M., 2011. Three measures of forest fire smoke exposure and their associations with respiratory and cardiovascular health outcomes in a population-based cohort. Environ. Health Perspect. 119, 1266–1271.

Houborg, R., McCabe, M.F., 2018. A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning. ISPRS J. Photogramm. Remote Sens. 135, 173–188.

Hsu, C.W., Chang, C.C., Lin, C.J., 2003. A practical guide to support vector classification.

Hu, X., Waller, L.A., Al-Hamdan, M.Z., Crosson, W.L., Estes, M.G., Jr, Estes, S.M., Quattrochi, D.A., Sarnat, J.A., Liu, Y., 2013. Estimating ground-level PM2.5 concentrations in the southeastern U.S. using geographically weighted regression. Environ. Res. 121, 1−10.

Hu, X., Belle, J.H., Xia, M., Wildani, A., Waller, L., Strickland, M., Liu Y., 2017. Estimating pm2.5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944.

Jiang M., Sun W., Yang G., Zhang D., 2017. Modelling seasonal GWR of daily PM2.5 with proper auxiliary variables for the Yangtze River Delta. Remote Sens. 9, 346.

John, R., Chen, J., Giannico, V., Park, H., Xiao, J., Shirkey, G., Ouyang, Z., Shao G., Laforteza, R., Qi, J., 2018. Grassland canopy cover and aboveground biomass in Mongolia and Inner Mongolia: Spatiotemporal estimates and controlling factors. Remote Sens. Environ. 213, 34–48.

Kan, H., Chen, R., Tong, S., 2012. Ambient air pollution, climate change, and population health in China. Environ. Int. 42, 10–19.

Kloog, I., Sorek-Hamer, M., Lyapustin, A., Coull, B., Wang, Y., Just, A. C., Schwartz, J., Broday, D. M., 2015. Estimating daily pm 2.5, and pm 10, across the complex geo-climate region of Israel using MAIAC satellite-based AOD data. Atmos. Environ. 122, 409–416.

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. Atmos. Environ. 45, 6267–6275.

Krstic, N., Henderson, S.B., 2015. Use of MODIS data to assess atmospheric aerosol before, during, and after community evacuations related to wildfire smoke. Remote Sens. Environ. 166, 1–7.

Lai, H.K., Tsang, H., Thach, T.Q., Wong, C.M., 2014. Health impact assessment of exposure to fine particulate matter based on satellite and meteorological information. Environ. Sci. Process. Impact 2014, 16, 239–246.

Landsberg, H.E., 1981. The urban climate (Vol. 28). Academic Press.

Lary, D.J., Lay, T., Sattler, B., 2015. Using machine learning to estimate global PM2.5 for environmental health studies, Environ. Health Insights 9, 41–52.

Lee, H.J., Chatfield, R.B., Strawa, A.W., 2016. Enhancing the applicability of satellite remote sensing for PM2.5 estimation using MODIS deep blue AOD and land use regression in California, United States. Environ. Sci. Technol. 50, 6546–6555.

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

Levy, R.C., Remer, L.A., Kleidman, R.G., Mattoo, S., 2010. Global evaluation of the collection 5 modis dark-target aerosol products over land. Atmospheric Chem. Phys., 10, 10399–10420.

Liu, J., Li, W., Wu, J., Liu, Y. 2018. Visualizing the intercity correlation of PM2.5 time series in the Beijing-Tianjin-Hebei region using ground-based air quality monitoring data. PloS one, 13, e0192614.

Liu, Y., 2014. Mapping annual mean ground-level PM2.5 concentrations using multiangle imaging spectroradiometer aerosol optical thickness over the contiguous United States. J. Geophys. Res. 109, D22.

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. Environ. Health Perspect. 117, 886–892.

Liu, Y., Franklin, M., Kahn, R., Koutrakis, P., 2007. Using aerosol optical thickness to predict ground-level PM 2.5 concentrations in the St. Louis area: a comparison between MISR and MODIS. Remote Sens. Environ. 107, 33–44.

Ma, Z., Hu, X., Huang, L. Bi, J., Liu, Y., 2014. Estimating ground-level PM2.5 in China using satellite remote sensing. Environ. Sci. Technol. 48, 7436–7444.

McLean, K.E., Yao, J., Henderson, S.B., 2015. An evaluation of the British Columbia Asthma Monitoring System (BCAMS) and PM2.5 exposure metrics during the 2014 forest fire season. Int. J. Environ. Res. Public Health 12, 6710–6724.

Mukai, S., Sano, I., Satoh, M., Holben, B.N., 2006. Aerosol properties and air pollutants over an urban area. Atmos. Res. 82, 643–651.

Natunen, A., Arola, A., Mielonen, T., Huttunen, J., Komppula, M., Lehtinen, K.E.J., 2010. A multi-year comparison of PM2.5 and AOD for the Helsinki region. Boreal Environ. Res. 15, 544–552

Ngufor, C., Murphree, D., Upadhyaya, S., Madde, N., Kor, D., Pathak, J., 2015. Effects of plasma transfusion on perioperative bleeding complications: a machine learning approach. Stud. Health Technol. Inform. 216, 721–725.

Ostro, B., Malig, B., Broadwin, R., Basu, R., Gold, E.B., Bromberger, J.T., Derby, C., Feinstein, S., Greendale, G. Jackson, E., Kravitz, H.M., Matthews, K.A., Sternfeld, B., Tomey, K., Green, R.R., Green. R., 2014. Chronic PM2.5 exposure and inflammation: determining sensitive subgroups in mid-life women. Environ. Res. 132, 168–175.

Paciorek, C.J., Liu, Y., Moreno-Macias, H., Kondragunta, S., 2008. Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level PM2.5. Environ. Sci. Technol. 42, 5800–5806.

R Core Development Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.Z. 2017. Deep learning for health informatics. IEEE J. Biomed. Health Inform. 21, 4–21.

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. Environ. Sci. Technol. 49, 3887–3896

RuleQuest., 2018. Data mining with Cubist, https://www.rulequest.com/cubist-info.html

Sapkota, A., Symons, J.M., Kleissl, J., Wang, L., Parlange, M.B., Ondov, J., Breysse, P.N., Buckley, T.J., 2005. Impact of the 2002 Canadian forest fires on particulate matter air quality in Baltimore City. Environ. Sci. Technol. 39, 24–32.

Saunders, R.O., Kahl, J.D.W., Ghorai, J.K., 2014. Improved estimation of PM2.5 using Lagrangian satellite-measured aerosol optical depth. Atmos. Environ. 91, 146–153.

Schaap, M., Apitley, A., Timmermans, R.M.A., Koelemeijer, R.B.A., de Leeuw G., 2009. Exploring the relation between aerosol optical depth and PM2.5 at Cabauw, the Netherlands. Atmos. Chem. Phys. 9, 909–925.

Shi, Y., Ho, H.C., Xu, Y., Ng, E., 2018. Improving satellite aerosol optical Depth-PM2.5 correlations using land use regression with microscale geographic predictors in a high-density urban context. Atmos. Environ. Doi: 10.1016/j.atmosenv.2018.07.021.

Shi, Y., Lau, K.K.L., Ng, E., 2017. Incorporating wind availability into land use regression modelling of air quality in mountainous high-density urban environment. Environ. Res. 157, 17–29.

Song, W., Jia, H., Huang, J., Zhang, Y., 2014, A satellite-based geographically weighted regression model for regional PM2.5 estimation over the Pearl River Delta region in China. Remote Sens. Environ. 154, 1–7.

Steyn, D.G., De Wekker, S.F., Kossmann, M., Martilli, A., 2013. Boundary layers and air quality in mountainous terrain. In Mountain Weather Research and Forecasting. Springer Netherlands, pp. 261–289.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B Stat. Methodol. 58, 267–288.

van Donkelaar A., Martin R.V., Park R. J., 2006. Estimating ground-level pm2.5 using aerosol optical depth determined from satellite remote sensing. J. Geophys. Res. 111, D21.

Wang, B., Chen, Z., 2016. High-resolution satellite-based analysis of ground-level PM2.5 for the city of Montreal. Sci. Total Environ. 541, 1059–1069.

Wang, J., Christopher, S.A., 2003. Intercomparison between satellite derived aerosol optical thickness and PM2.5 mass: implications for air quality studies, Geophys. Res. Lett. 30, 2095.

Wang, Y., Shi, L., Lee, M., Liu, P., Di, Q., Zanobetti, A., Schwartz, J.D., 2017. Long-term exposure to PM2.5 and mortality among older adults in the Southeastern US. Epidemiology 28, 207–214.

Ward, D.E., Hardy, C.C., 1991. Smoke emissions from wildland fires. Environ. Int. 17, 117-134.

Xie, Y., Wang, Y., Zhang, K., Dong, W., Lv, B., Bai, Y., 2015. Daily estimation of ground-level PM2.5 concentrations over Beijing using 3km resolution MODIS AOD. Environ. Sci. Technol, 19, 12280–12288.

Wotton, B.M., Nock, C.A., Flannigan, M.D., 2010. Forest fire occurrence and climate change in Canada. Int. J. Wildland Fire 19, 253–271.

Wu, J., Li, J., Peng, J., Li, W., Xu, G., Dong, C., 2015. Applying land use regression model to estimate spatial variation of PM2.5 in Beijing, China. Environ. Sci. Pollut. Res. Int. 22, 7045-7061.

Wu, J., Yao, F., Li, W., Si, M., 2016. VIIRS-based remote sensing estimation of ground-level PM2.5 concentrations in Beijing-Tianjin-Hebei: A spatiotemporal statistical model. Remote Sens. Environ., 184, 316–328.

Xin, J., Zhang, Q., Wang, L., Gong, C., Wang, Y., Liu, Z., Gao, W., 2014. The empirical relationship between the PM2.5 concentration and aerosol optical depth over the background of North China from 2009 to 2011. Atmos. Res. 128, 179–188.

Yao, F., Si, M., Li, W., Wu, J., 2018a. A multidimensional comparison between MODIS and VIIRS AOD in estimating ground-level PM2.5 concentrations over a heavily polluted region in China. Sci. Total Environ 618, 819-828.

Yao, J., Raffuse, S.M., Brauer, M., Williamson, G.J., Bowman, D.M., Johnston, F.H., Henderson, S.B., 2018b. Predicting the minimum height of forest fire smoke within the atmosphere using machine learning and data from the CALIPSO satellite. Remote Sens. Environ. 206, 98–106.

You, W., Zang, Z., Pan, X., Zhang, L., Chen, D., 2015. Estimating pm2.5 in Xi'an, China using aerosol optical depth: a comparison between the MODIS and MISR retrieval models. Sci. Total Environ. 505, 1156–1165.

You, W., Zang, Z., Zhang, L., Li, Y., Pan, X., Wang, W., 2016. National-scale estimates of ground-level PM2.5 concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD. Remote Sens. 8, 184.

Zheng, Y., Zhang, Q., Liu, Y., Geng, G., He, K., 2016. Estimating ground-level PM2.5 concentrations over three megalopolises in China using satellite-derived aerosol optical depth measurements. Atmos. Environ. 124, 232–242.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M.L., Shen, X., Zhu, L., Zhang, M., 2017. Spatiotemporal prediction of continuous daily PM2.5, concentrations across China using a spatially explicit machine learning algorithm. Atmos. Environ. 155, 129–139.

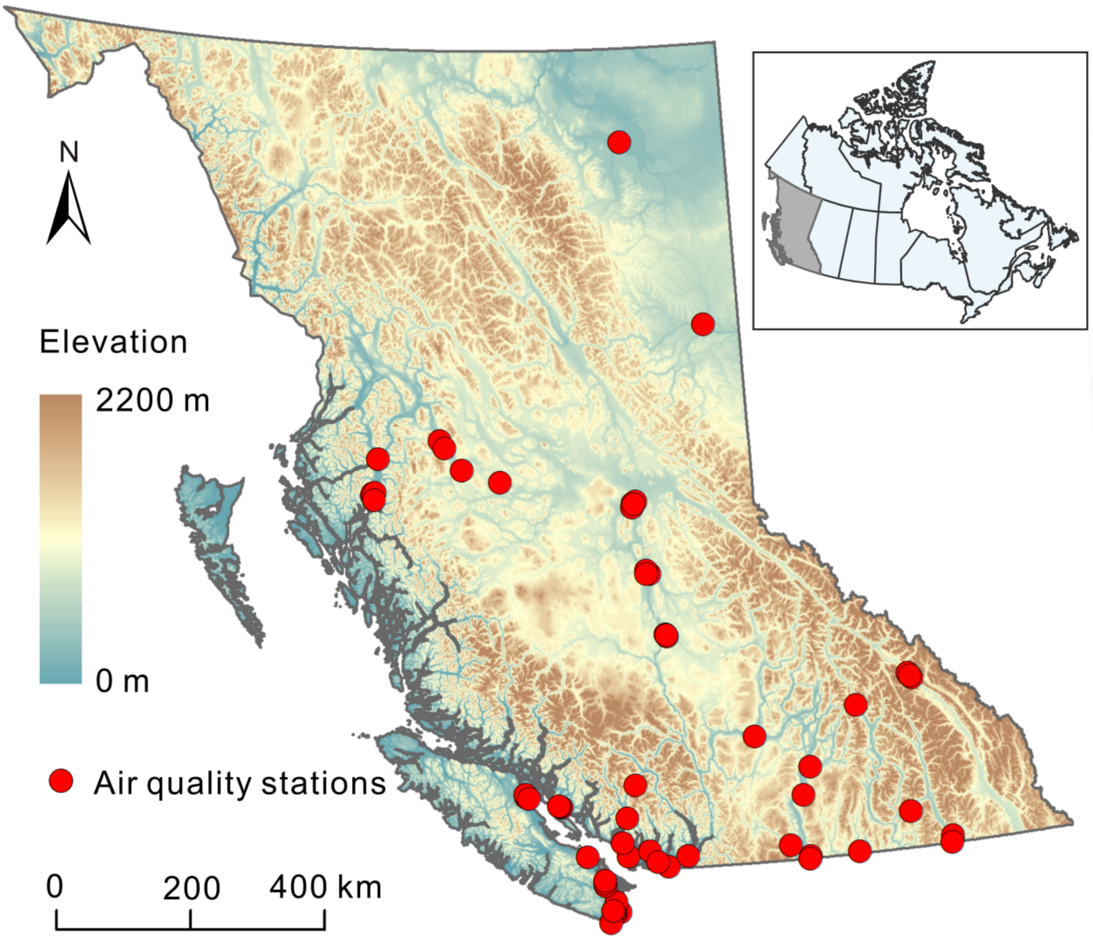

Fig. 1. Study Site. Red dots represent the location of air quality stations across BC.

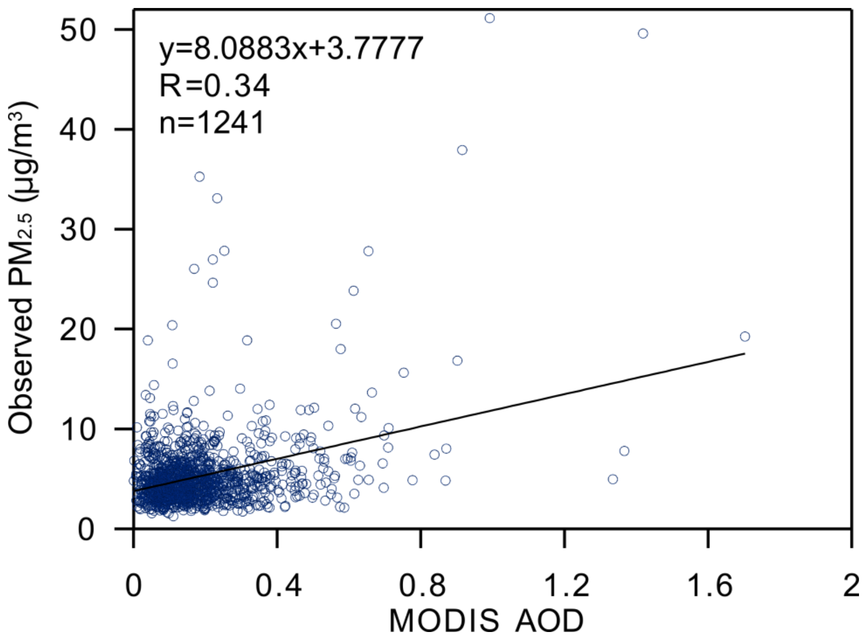

Fig.2 Empirical relationship between $PM_{2.5}$ and AOD. X-axis indicated the AOD values derived from MODIS dataset. Y-axis indicated the $PM_{2.5}$ retrieved from the air quality stations.

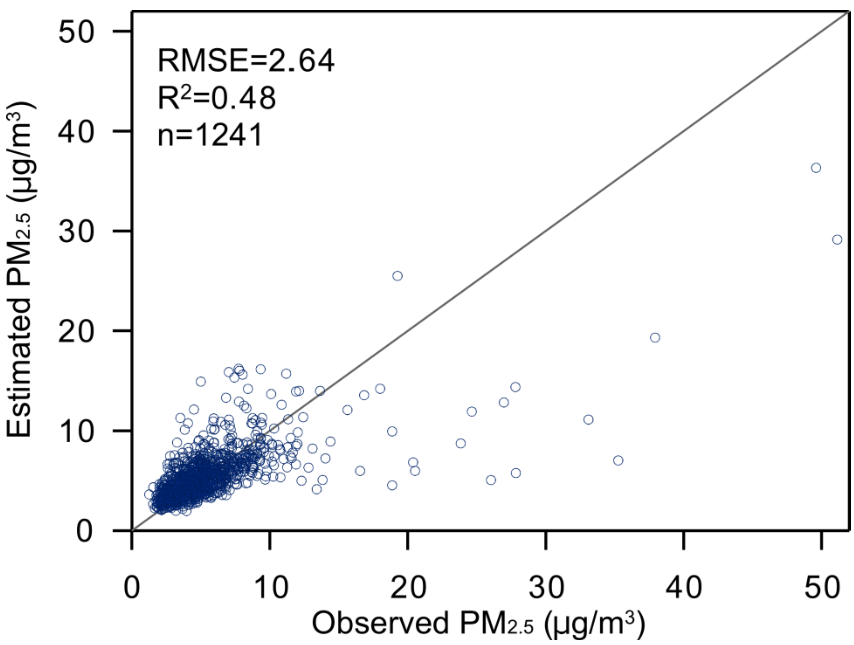

Fig. 3 Comparison between observed and estimated $PM_{2.5}$ using Cubist.

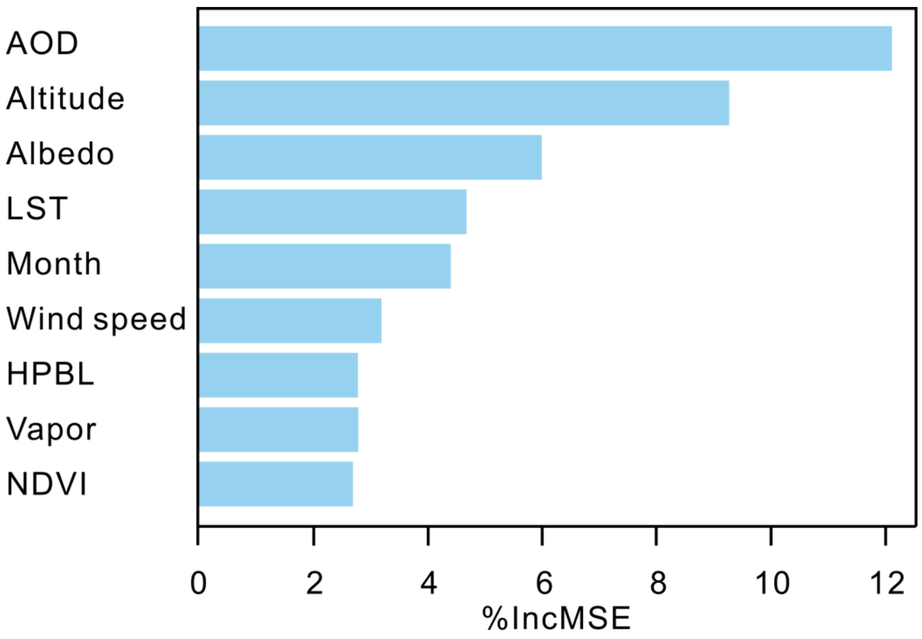

Fig. 4 Variable importance analysis (Cubist Model). Y-axis indicated the predictors for predicting $PM_{2.5}$. X-axis indicated the percentage increase in mean square error (%IncMSE) without using the corresponding predictor.

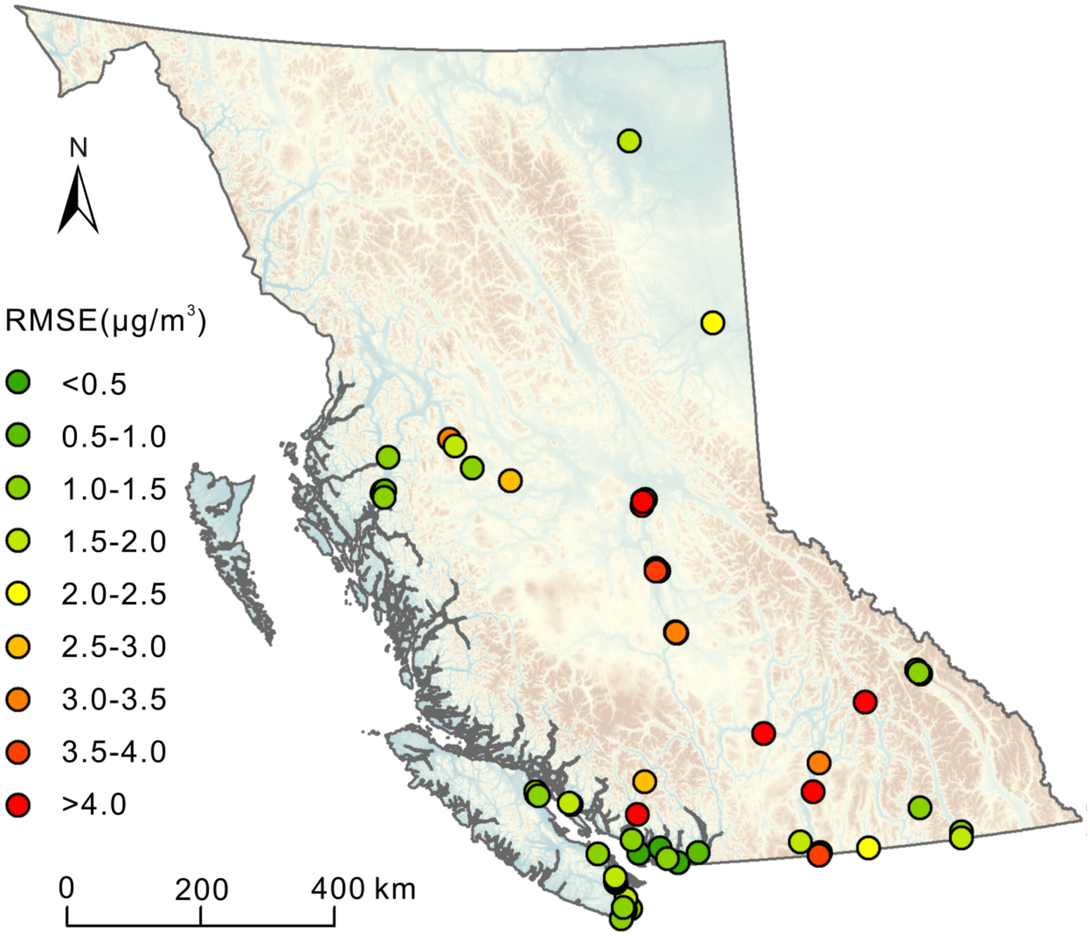

Fig. 5 Location-based root mean square error (RMSE) of estimated $PM_{2.5}$. Red indicated an air quality station with higher RMSE, and green indicated a station with lower RMSE after a comparison with observed data.

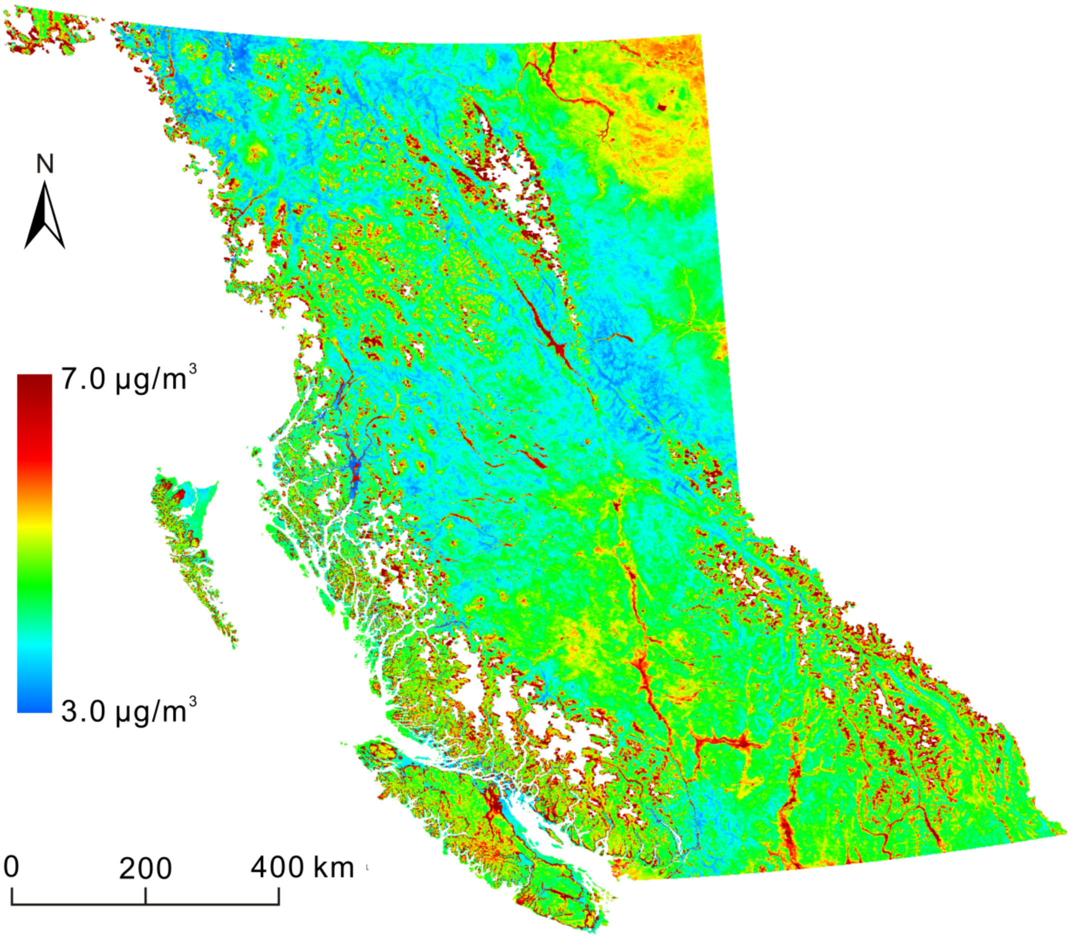

Fig. 6 Average of ground-level $PM_{2.5}$ concentration across BC (2001-2014)

Table 1 Information on datasets used for PM2.5 estimation

Dataset	Spatial resolution	Temporal resolution	Scenes	Derived predictors
MOD04_3k	3km	Daily	25350	AOD
MOD05_L2	1km	Daily	22198	Vapor
MOD11A1	1km	Daily	25369	LST
MOD13A3	1km	Monthly	1677	NDVI
MCD43B3	1km	16 days	6394	albdo
NCAR/NCEP re-analysis	2.5º	Monthly	/	HPBL, wind speed
SRTM DEM	90m	/	/	elevation

Table 2 Accuracy of $PM_{2.5}$ prediction of each machine learning model.

Model	CV-RMSE (μg/m^3)	CV-R^2
MLR	3.24	0.22
BRNN	3.04	0.31
SVM	3.13	0.30
LASSO	3.20	0.24
MARS	3.05	0.31
RF	2.67	0.49
XGBoost	2.71	0.46
Cubist	2.64	0.48