**Journal of Environment Protection and Sustainable Development**
Vol. 1, No. 5, 2015, pp. 245-254
http://www.aiscience.org/journal/jepsd
ISSN: 2381-7739 (Print); ISSN: 2381-7747 (Online)

# Exploring the Utility of the Random Forest Method for Forecasting Ozone Pollution in SYDNEY

**Ningbo Jiang*, Matthew L. Riley**

New South Wales Office of Environment and Heritage, Sydney, Australia

## Abstract

This paper explores the utility of an ensemble decision-tree method called *random forest*, in comparison with the classic classification and regression trees (CART) algorithm, for forecasting ground-level ozone pollution in the Sydney metropolitan region. Statistical forecasting models are developed to provide daily ozone forecasts in November-March for three subregions, i.e., Sydney east, Sydney south-west and Sydney north-west. The random forest models are evaluated in reference to the single decision-tree models developed from the classic CART algorithm. The results show that the random forest models outperform the CART models for forecasting high ozone pollution in Sydney south-west and Sydney north-west, the areas where the highest ozone pollution are observed. The random forest models also show a lift in forecasting skills in Sydney south-west if compared to the existing forecasting practice for the basin as a whole. These results suggest that random forest is a promising method for air quality forecasting in Sydney. This study promotes the application of a statistical ensemble approach to air quality forecasting.

## Keywords

Air Quality Forecast, Ozone Pollution, Decision Tree, Random Forest, Bagging, Boosting

Received: September 15, 2015 / Accepted: November 6, 2015 / Published online: December 20, 2015



## 1. Introduction

Air pollution increases risks of respiratory, cardiovascular and neurological diseases, particularly among the very young, the elderly and those with pre-existing health conditions (Barnett, 2012; WHO, 2006). Epidemiological studies suggest that health effects of air pollution are a problem even at low levels (Barnett et al., 2006; EPHC, 2010; Anderson et al., 2012; WHO, 2014). Sydney is the largest metropolitan region in New South Wales (NSW), accommodating over 60% of the State's population (ABS, 2011). Compared with other urban regions in the world, air quality in Sydney is considered to be relatively good (OEH, 2014). Concentrations of carbon monoxide (CO), nitrogen dioxide ($NO_2$), sulphur dioxide ($SO_2$) and lead (Pb) are stable and consistently meet the Australian national air quality standards (NEPC, 2003). However, (ground-level) ozone ($O_3$) and particle (PM10 and PM2.5, i.e., particulate matter less than 10 and 2.5 micrometers in diameter, respectively) levels can exceed the national standards from time to time, posing health risks to local communities and the environment (Jiang et al., 2015; EPA, 2012).

The NSW Office of Environment and Heritage (OEH) is responsible for providing information on air quality to local communities in the State (Jiang *et al*., 2015). In addition to operating an extensive ambient air quality monitoring network and providing communities with access to near real-time measurements (www.environment.nsw.gov.au/AQMS/hourly data.htm), OEH is also committed to providing air quality forecasts. Reliable air quality forecasts can help people take preventative actions to minimise personal health impacts (US EPA, 2003, 2015). Forecasts also support proactive emission reduction measures to avoid or mitigate air pollution on forecasted high pollution days. Forecasts can also help in

* Corresponding author
E-mail address: ningbo.jiang@environment.nsw.gov.au (Ningbo Jiang)

planning specialised campaign monitoring programs, optimising the use of expensive measurement resources (Zhang *et al*., 2012a).

OEH's existing air quality forecasting (AQF) has focused on ozone pollution in Sydney, and forecasts are issued in Air Quality Index (AQI) class at 4:00pm (local time) each day (www.environment.nsw.gov.au/AQMS/aqi.htm; Section 2.1). The operational process is largely manual and qualitative, with limited capability for forecasting particle pollution (Jiang *et al*., 2015). OEH plans to progressively advance its capability for forecasting air quality within the Greater Metropolitan Region and key regional areas in NSW, aiming to reduce the risks of harm to human health from air pollution episodes. In this endeavour, one of the initial tasks is to explore a statistical AQF tool that provides daily routine $O_3$ AQI forecast for three subregions in Sydney, i.e., Sydney east (E), Sydney south-west (SW), Sydney north-west (NW) (0). This development experience will prepare OEH with capability to apply more advanced forecasting systems and tools in the future.

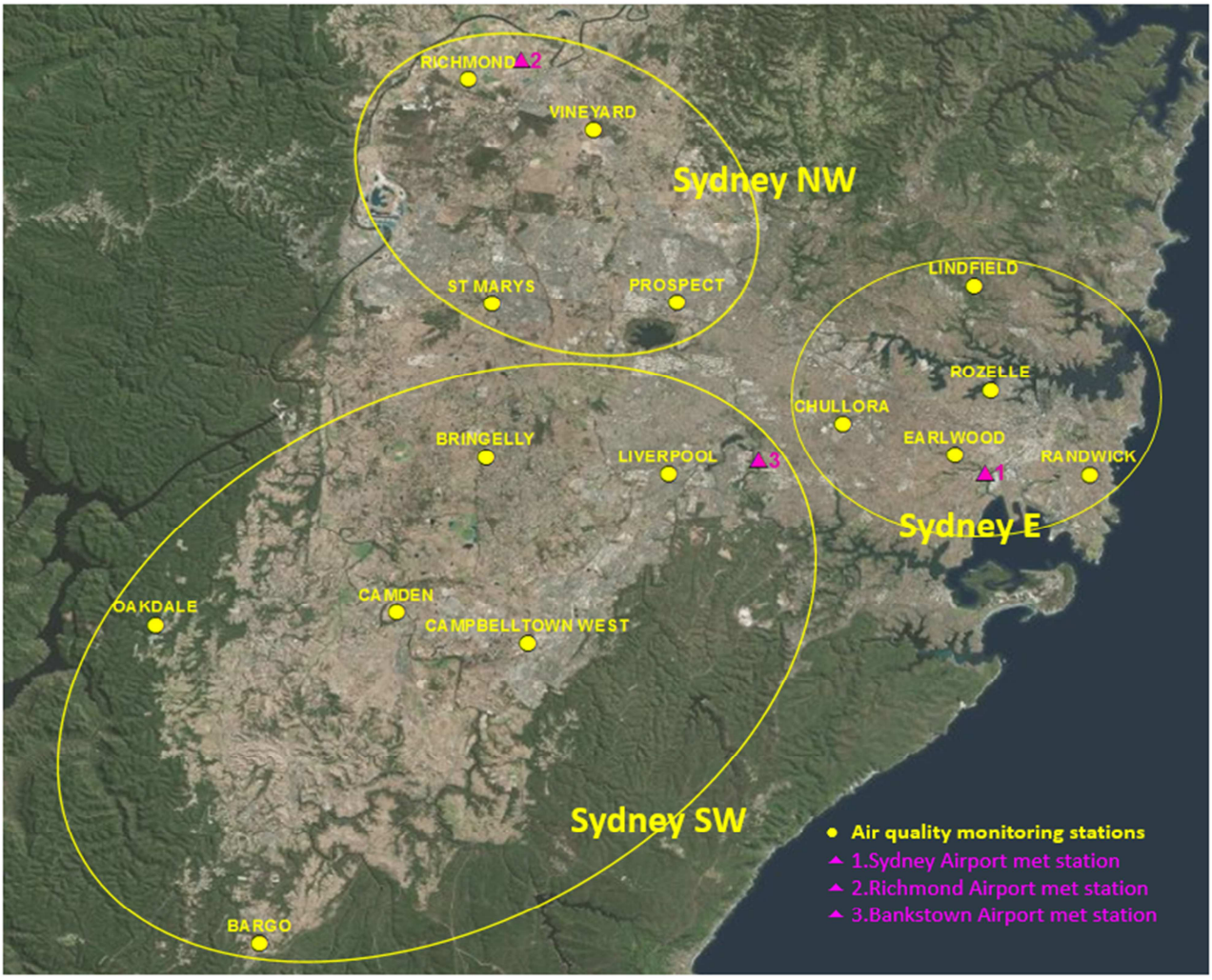


**Figure 1.** Schematic of three air quality subregions in the Sydney basin. Air quality monitoring stations are labelled in yellow, and three airport meteorological stations are labelled in pink.

Good reviews on the AQF systems and tools applied in different countries or regions can be found in Zhang *et al*. (2012a, 2012b), Honore *et al*. (2008), White (2011) and US EPA (2003). In summary, the existing AQF methods typically fall within four categories, i.e., empirical approach, statistical modelling approach, deterministic 3-D numerical modelling approach and hybrid approach, which feature with various levels of sophistication, forecasting skills and resource needs (Figure 2). For example, compared to more sophisticated numerical modelling, statistical models are computationally fast and suitable for describing the complex site-specific relations between concentrations of air pollutants and potential predictors. The commonly used statistical techniques addressing nonlinearities and interactive relationships include the decision tree method (e.g., classification and regression trees, CART), artificial neural networks (ANNs), nonlinear regression (NLR), fuzzy logic, and their variants (Zhang *et al*., 2012a).

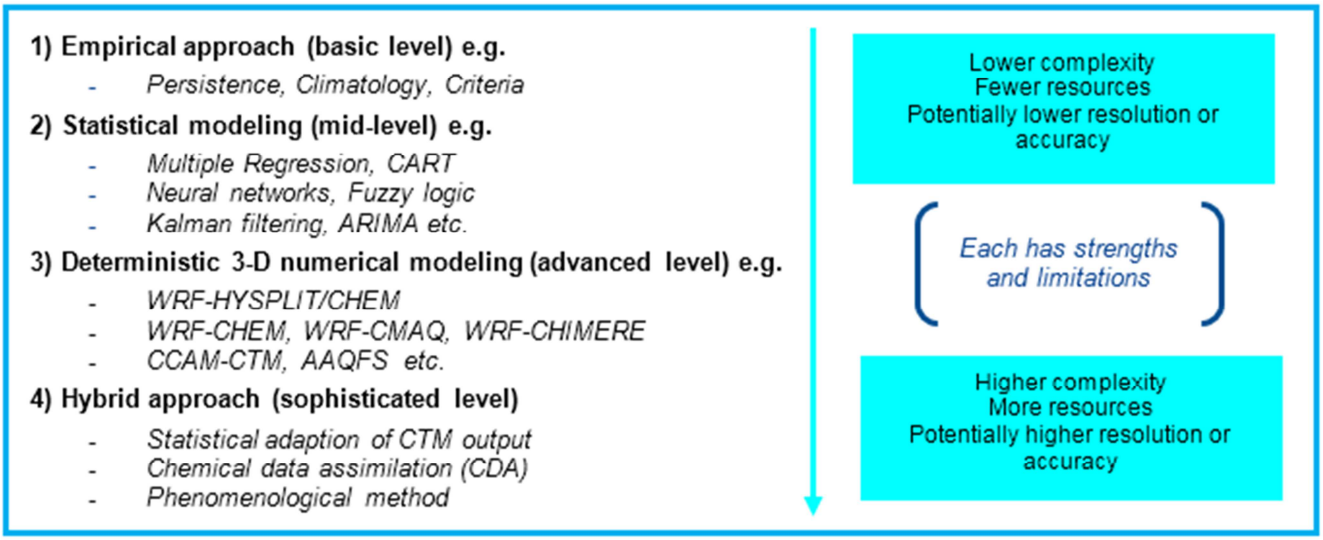


**Figure 2.** Examples of commonly-used forecasting methods.

Source: Jiang *et al*. (2015)

This paper explores the potential utility of a statistical ensemble method called *random forest* in comparison with the classic CART algorithm, for forecasting ozone pollution in three Sydney subregions (Figure 1). The random forest method adopts a statistical ensemble principle, and is essentially a generalisation of the CART algorithm through application of the bagging (bootstrap aggregation) technique (Han and Kamber, 2008; Hastie *et al*., 2009; Section 2.6). As an initial attempt, the present study was focused on warm months (November-March), when $O_3$ pollution is more significant compared to other months of the year (Jiang *et al*., 2013). Section 2 describes the analytic framework for model development, while Section 3 gives a brief summary on the model performance evaluation results. At last, the paper is concluded in Section 4.

## 2. Analytic Framework

This section describes the analytic framework for constructing three statistical models that provide short-term (24 hours) forecast of $O_3$ pollution for three subregions in Sydney. The description is centred on the model building process, response variables, predictive variables, input data, definition of model verification statistics, and the *rpart* (version 4.1-8) and *randomForest* (version 4.6-7) packages in R (Therneau *et al.* 2014; Liaw and Wiener, 2013).

### 2.1. Model Building Process Overview

The schematic of the model building process is shown in Figure 3. The input data for model building consisted of observational records obtained from the OEH (in-house) Air Quality Database or the Australian Bureau of Meteorology (BOM) (details in Section 2.4). The data were quality assured (QA) following the standard OEH QA procedure, with invalid or suspected values excluded from the present study. The full input data set was split into two subsets, i.e., the training and testing sets. The training set was used for model construction, while the testing set was used for assessing the forecasting skills of the models.

The model development was undertaken on the R for Windows version 3.1.0 platform, in particular through the use of two statistical packages, *rpart* and *randomForest*. The rpart version 4.1-8 package, an implementation of the classic CART algorithm, was applied to derive a *single-tree* reference (benchmark) model for each subregion. In parallel, the randomForest version 4.6-7 package, an implementation of the random forest algorithm, was used to build a two-phase decision-tree ensemble model for the same subregion. A simple majority vote is taken as model prediction. The random forest algorithm was applied twice following the boosting principle (Shapire et al., 1998), with Phase 2 focusing on data points incorrectly predicted in Phase 1. A brief summary of the CART and random forest techniques is given in Sections 2.5 and 2.6, and further discussions on these topics can be found in Han and Kamber (2008), Therneau *et al*. (2014) or Breiman *et al*. (2013). Information on the *rpart* and *randomForest* packages in R is available on The Comprehensive R Archive Network website http://cran.r-project.org/, in particular in the user manuals by Therneau *et al*. (2014) and Liaw and Wiener (2013).

The performance of the *decision-tree ensemble* models was compared with the single-tree (rpart) benchmark models for

each subregion. Six performance statistics were computed on both the training and testing data sets, i.e., probability of detection (POD), false alarm rate (FAR), critical success index (CSI), accuracy (A), bias (B) and model skill (details in Section 2.7). Based on the performance evaluation, the final AQF models were determined.

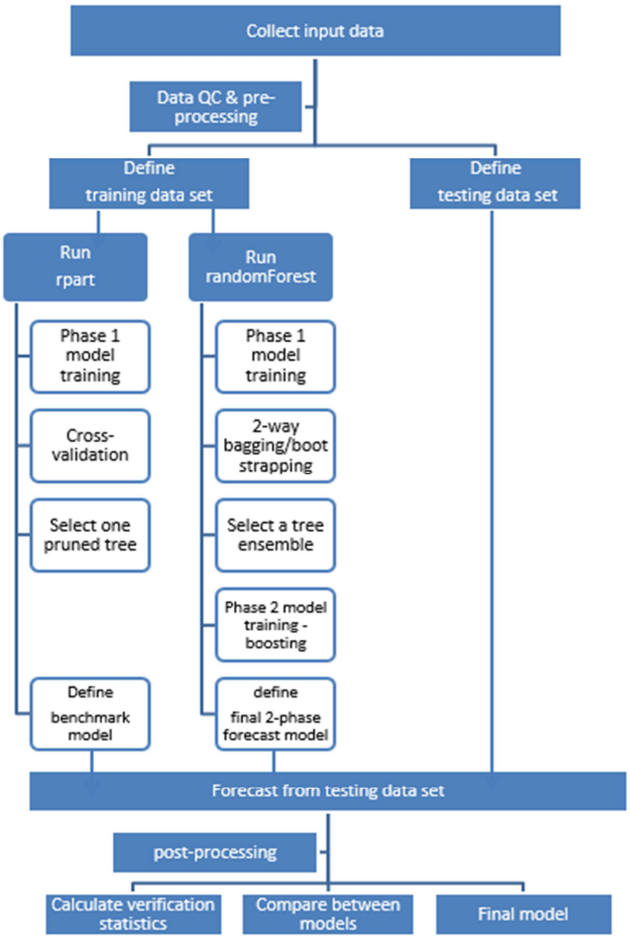


**Figure 3.** Schematic of the model building process.

## 2.2. Response Variables

In NSW, air quality is reported as AQI in six classes, very good (AQI ϵ [0, 33]), good (AQI ϵ [34-66]), fair (AQI ϵ [67-99]), poor (AQI ϵ [100-149]), very poor (AQI ϵ [150-199]) or hazardous (AQI >= 200) (Jiang *et al*., 2014a). In particular, when ozone pollution exceeds the relevant national air quality standards (i.e., 100 ppb for 1-hour average $O_3$ concentration and 80 ppb for 4-hour rolling average $O_3$ concentration), the AQI is classified as poor, very poor or hazardous, depending on the magnitude of the exceedance. In reality, when compared to those in the very good or good class, the number of days with $O_3$ pollution falling in poor or higher class is generally small in Sydney. In the present study, for statistical validity (the need for sufficiently large sample sizes) and also practical relevance (the poor or higher class pollution poses increased risks to human health), the six AQI classes were collapsed into the following three AQI categories:

Category 1 (good): AQI ϵ [0, 66], i.e., *very good* or *good* AQI class is forecasted to be good;

Category 2 (fair): AQI ϵ [67, 99], i.e., *fair AQI* class is still forecasted to be fair;

Category 3 (poor): AQI >= 100, i.e., *poor*, *very poor* or *hazardous* class is forecasted to be poor.

A site AQI category for a day was derived from the 1-hour average or 4-hour rolling average $O_3$ AQI values/classes at that site on that day. A subregional maximum 1-hour AQI category for a day was then defined as the highest AQI category across all sites in that subregion on that day. Hence, three categorical response variables were defined for the forecasting models, one for each subregion (Table 1).

## 2.3. Predictive Variables

Previous studies have shown that both synoptic and local-scale meteorological conditions have significant influence on air quality conditions in Sydney (e.g., Leighton and Spark, 1997; Katestone Scientific, 1997; Hart *et al*., 2006; Jiang *et al*., 2013). Based on Hart *et al*. (2006) recommendations and also taking into account of the availability of meteorological forecast data, meteorological variables for three airport sites, i.e., Sydney Airport, Bankstown Airport and Richmond Airport (Figure 1), were included in the pool of predictive variables for building the single-tree or random-forest models (Table 1). A few considerations are worth noting:

1) Present day and previous day maximum 1-hour $O_3$ and $NO_2$ concentrations for three subregions were included in the model building process to account for possible pollution accumulation and spatial transport effects.

2) The index for day of the week was included to (at least partially) take into account the weekly cycle in pollutant emissions (e.g., NOx emissions from the urban traffic corridors).

3) Meteorological data for two time points of the day, 6:00 and 15:00 AEST, were included in order to take into account the effect of diurnal variability in local circulation features (e.g., sea breezes and drainage flows; Jiang *et al*., 2013).

4) Mixing height and rainfall were not included in the predictive variable pool. These data were not readily available at the time of this study.

5) Solar radiation was not included during model development, with cloud cover being used as proxy. The BOM does not undertake direct solar radiation measurement.

**Table 1.** Predictive and response variables for building the $O_3$ AQI category forecasting models.

| Response variables | Predictive variables | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Type** | **Site** | **Level** | **Hour (AEST)** | **Variables** | **Day*** | **Data availability in forecast mode** |
| subregional day +1 max 1-hour AQI category: 1) Sydney SW; 2) Sydney NW; 3) Sydney E | Local meteorology | Sydney Airport | Surface | 6:00, 15:00 | wind speed, u, v, T, Td, MSLP | +1 | BOM or in-house |
| | | | 850hPa | 6:00, 15:00 | wind speed, u, v, T, Td | +1 | BOM or in-house |
| | | | Upper | 6:00, 15:00 | cloud cover (average) | +1 | BOM or in-house |
| | | Bankstown Airport | Surface | 6:00, 15:00 | wind speed, u, v, T, Td | +1 | BOM or in-house |
| | | Richmond Airport | Surface | 6:00, 15:00 | wind speed, u, v, T, Td | +1 | BOM or in-house |
| | | Between-site | Surface | 15:00 | delta T: difference in T between Sydney Airport and Bankstown Airport | +1 | Model build-in |
| | | Between-site | Surface | 15:00 | delta T: difference in T between Sydney Airport and Richmond Airport | +1 | Model build-in |
| | | Sydney Airport | Surface | 15:00 | derived variable: v*T | +1 | Model build-in |
| | | Bankstown Airport | Surface | 15:00 | derived variable: v*T | +1 | Model build-in |
| | | Richmond Airport | Surface | 15:00 | derived variable: v*T | +1 | Model build-in |
| | Synoptic catalogue | NSW | Surface | 10:00 | index of synoptic type defined with the Jiang et al. (2014b) method | +1 | BOM and in-house |
| | Air quality | Sydney SW | Surface | daily | subregional maximum 1-hour $O_3$ | 0, -1 | In-house |
| | | | Surface | daily | subregional maximum 1- hour $NO_2$ | 0 | In-house |
| | | Sydney NW | Surface | daily | subregional maximum 1- hour $O_3$ | 0, -1 | In-house |
| | | | Surface | daily | subregional maximum 1- hour $NO_2$ | 0 | In-house |
| | | Sydney E | Surface | daily | subregional maximum 1- hour $O_3$ | 0, -1 | In-house |
| | | | Surface | daily | subregional maximum 1- hour $NO_2$ | 0 | In-house |
| | Other | NSW | Surface | daily | index for day of the week | +1 | Calendar |

* +1: next day; 0: current day; -1: previous day; T: air temperature; Td: dew point; u: west-east wind component; v: south-north wind component; MSLP: mean sea level pressure.

## 2.4. Input Data

The following subsections describe the type, sources and preparation of the input data (forecast data are not discussed in this paper) used for training and testing statistical forecasting models.

### 2.4.1. Data Description

*Air quality data (source: OEH)*

Air quality data were extracted from the OEH Air Quality Database. These include 1-hour average and 4-hour rolling average $O_3$ AQI values, 1-hour average $NO_2$ and $O_3$ concentrations, and 1-hour average nephelometer readings (NEPH hereafter) from 14 air quality monitoring stations in Sydney for November-March in 2007-2014 (Figure 1). Two new sites, Campbelltown West and Camden, are not included in the study due to short records. The historical records from the Macarthur site were used as proxy data for the Campbelltown West site, since these two sites are separated by less than 1 km and are within the same local air shed.

*Local meteorological data (source: BOM)*

Local meteorological data were obtained from BOM, including hourly records of wind speed, wind direction, air temperature (T) and dew point (Td) at the surface level for the Bankstown Airport and Richmond Airport meteorological stations and at both the surface and 850 hPa levels for the Sydney Airport meteorological station, for November-March (warm months) in 2007-2014. Also included were the 6:00 and 15:00 AEST cloud cover and mean sea level pressure (MSLP) records from the Sydney Airport station for the same period.

*Synoptic catalogue (source: in-house)*

A synoptic catalogue was derived in house by applying a classification method called self-organising map (SOM) to the twice-daily (10:00 and 22:00 AEST) NCEP/NCAR 1000hPa geopotential height reanalysis for November to March in 1958-2014 (Kalnay et al., 1996). The classification procedure was detailed in Jiang *et al*. (2012; 2014b). A set of 12 typical synoptic patterns was identified for east Australia, and each of the twice daily geopotential height maps was categorised to one of the 12 types using the Euclidean distance metric. In the

present study, only the 10:00 AEST categorisation for the period 2007-2014 was used to build the single-tree or random forest models.

### 2.4.2. Data Preparation

Wind speed and wind direction data were used to derive the west-east (u) and south-north (v) wind components (Table 1). Hourly $O_3$ AQI data were used to derive the subregional daily maximum $O_3$ AQI categories as described in Section 2.2; in a similar fashion, the hourly $NO_2$ and $O_3$ concentration data were used to derive daily subregional maximum 1-hour average $NO_2$ and $O_3$ concentrations. The nephelometer readings were used to derive daily maximum 1-hour average NEPH values for each subregion. For each subregion, data for days with NEPH>= $2.1 \times 10^{-4}$/m (i.e., *low visibility* or *smoky* days; the cut-off value is equivalent to a distance of around 9 km in visibility) were excluded from the model building process, so as to suppress the potential effects from wildfires or planned hazard reduction burns on $O_3$ pollution. This treatment led to less than 1% data loss.

*Training and testing data sets*

In model building, one may choose two-way or three-way data split in order to undertake model training, evaluation and testing. A three-way split involves splitting the input data into three subsets, often termed as *training set*, *evaluation set*, and *testing set*. A two-way split leads to only training and testing tests. Given that the number of records (in days) here was not very large, we chose a two-way split through random sampling (Table 2), i.e., forming a training set (2/3 of the full set) and a testing set (1/3 of the full set). The training data set was used for model construction, while the testing data set was applied to determine the model performance (note: the model *evaluation* was undertaken using the cross-validation in rpart but bootstrapping in randomForest; Sections 2.5 and 2.6).

**Table 2.** Training and testing data sets: sample sizes (days) by AQI category for each subregion.

| Subregion | Data set | Category 1 | Category 2 | Category 3 | Subtotal | Grand total |
|---|---|---|---|---|---|---|
| Sydney SW | Training set | 363 | 103 | 26 | 492 | 719 |
| | Testing set | 185 | 36 | 6 | 227 | |
| Sydney NW | Training set | 397 | 84 | 12 | 493 | 721 |
| | Testing set | 188 | 34 | 6 | 228 | |
| Sydney E | Training set | 449 | 38 | 2 | 489 | 713 |
| | Testing set | 204 | 19 | 1 | 224 | |

## 2.5. Single Decision Tree and the *rpart* Package in R

A decision tree is a flowchart-like tree structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of decisions rules (Han and Kamber, 2008). The CART algorithm (Breiman *et al*., 1984) adopts a non-backtracking approach, in which a decision tree is developed by continuously splitting the response variable data (e.g., AQI categories) into two (most dissimilar) groups based on a single value of a selected predictive variable. The selected predictive variable is identified as the variable having the highest correlation with the response variable. The splitting of the data set and tree development continues until the data in each group are sufficiently uniform.

The *rpart* (recursive partitioning) package is an implementation of the CART algorithm in R (Therneau *et al*., 2014). It was used to derive a benchmark (reference) *single-tree* model for each subregion, which was then used to assess the performance of a tree ensemble model developed from the randomForest package (Section 2.6) for the same subregion. When applying rpart to this study, the *gini* index was used as the splitting criterion function. If a data set T contains examples from n classes, the gini index, denoted as *gini (T)*, is defined as

$$gini(T) = 1 - \sum_{k=j}^{n} p^2{}_{\mathrm{j}}$$

where $p_j$ is the proportion of records in class $j$ in T. A 15-fold cross-validation (CV) was implemented in the model training in order to estimate prediction errors and undertake tree pruning.

## 2.6. Decision-Tree Ensemble and the *random Forest* Package in R

In standard trees (from CART), each node is split using the best split option among all variables. The random forest algorithm adopts an ensemble modelling approach by applying the bagging technique (Han and Kamber, 2008), where an ensemble of classification and/or regression trees are derived from random subsets of the data, using a subset of randomly chosen predictors for each split in each classification tree (Liaw and Wiener, 2002). The algorithm is able to better examine the contribution and behaviour that each predictor has, even when one predictor's effect would be overshadowed by more significant competitors in simpler models. Studies showed that random forest performs very well compared with other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against over-fitting (Breiman, 2001).

The *random Forest* package in R provides an interface to the

Fortran implementation of the random forest algorithm by Breiman and Cutler (Liaw and Wiener, 2013). This package was used to develop ensemble tree models for the three subregions in Sydney. The gini index was used as the classification function (as in the rpart package), and the boostrap aggregation (bagging) technique was applied to estimate the classification errors in order to evaluate the forest models. A simple majority vote from the multiple trees is taken as model prediction.

The random forest algorithm was applied twice to formulate a dual-forest forecasting model for each subregion. This was motivated by the idea of boosting (Han and Kamber, 2008; Shapire et al., 1998). In the first phase, the algorithm was applied to the full training data set (for the subregion) to generate an initial forest model, which consists an ensemble of 300 decision trees. This model tends to provide reliable forecasts for days falling into the good AQI category if compared with observations. In the second (boosting) phase, based on the output from the initial forecasting model, records in the training data set that were not correctly predicted were isolated to form a new set of training data. The random forest algorithm was then applied to this new data subset alone to obtain an ensemble of 500 trees for the subregion. This model appears to provide better (boosted) forecasts for days that fall into the fair or poor AQI category if compared with observations. If a day was forecasted to be good in the second phase or in both phases, that day was assigned to the good category. However, if a day forecasted to be fair or poor in the second phase or in both phases, that day was assigned to the fair or poor category. Hence, two forest models applied in tandem to generate an AQI category forecast for a specific day.

### 2.7. Model Performance Statistics

***Table 3.*** Interpretation of six performance statistics. Source: adapted from US EPA (2003).

| Statistics | Description |
|---|---|
| Accuracy (A) | Percent of forecasts that correctly predicted the event or non-event. |
| Bias (B) | Indicates, on average, if the forecasts are under-predicted (false negatives) or over-predicted (false positives). |
| False alarm rate (FAR) | Percent of times a forecast of high pollution did not actually occur. |
| Critical success index (CSI) | How well the high-pollution events were predicted; it is unaffected by a large number of correctly forecasted, low-pollution events. |
| Probability of detection (POD) | Ability to predict high-pollution events. |
| Skill | Percentage of improvement in the accuracy of a random forest model with respect to the rpart reference/benchmark model |

As mentioned earlier, the model performance was assessed using six statistical indicators defined for categorical forecasts. These are accuracy (A), bias (B), false alarm rate (FAR), critical success index (CSI), probability of detection (POD) and forecast skill (Table 3). For simplicity, the performance statistics were evaluated for the three AQI category forecasts at the cut-off between fair and poor (i.e., exceeding or not exceeding the national standards for ozone) following the method described in US EPA (2003). In other words, the evaluation was focused on assessing the model performance for forecasting ozone exceedance days in Sydney subregions.

## 3. Results and Discussion

From the analytic framework, two models were constructed for each subregion, one consisting of a single decision tree developed with the rpart package, the other an ensemble of decision trees developed with the randomForest package. Tables 4 and 5 show the performance statistics (Section 2.7) for each subregion and for the rpart (benchmark) models and the randomForest models, calculated on the training and testing data sets, respectively. The calculation was undertaken at the cut-off between fair and poor, focusing on assessing the model performance for forecasting ozone exceedance days. The forecasting skills of the randomForest models were derived in reference to the rpart benchmark models for different subregions.

The forecasting accuracies (As) of these statistical models vary across the three subregions. The models demonstrate considerably higher forecasting accuracies for Sydney E (92-99%), yet similar (if not higher) accuracies for both Sydney SW (85-94%) and Sydney NW (8-96%), if compared to the existing OEH forecasting practice (87%; where the daily AQI category was forecasted for the Sydney basin as a whole). Of note is the generally similar accuracies between the randomForest models and the rpart benchmark models on the testing data set, although the randomForest models tend to perform better than the rpart models on the training data set. This lack of difference (for the testing data) between two types of models in Sydney E is largely determined by the fact that ozone levels in this subregion tend to fall (almost all the time) under the good or (to a lesser degree) fair AQI category, with poor air quality (exceedance) days mostly observed within Sydney SW and/or Sydney NW (Table 2).

The rpart benchmark models have slightly higher POD (probability of detection) and CSI (critical success index) levels for Sydney SW, if compared to the existing forecasting practice; however, they are unable to predict high pollution events in Sydney NW. This is true for both the training and testing data sets. In contrast, the randomForest models provide significantly improved POD (>66%) and CSI (>44%) levels for both Sydney SW and Sydney NW, compared to the rpart models (0-47% and 0-38% for POD and CSI, respectively)

and also the existing OEH practice (32% and 26% for POD and CSI, respectively). While the false alarm rates (FARs) are similar (43-45%), the forecasting bias (B) tend to be higher for the randomForest models (up to 1.5, i.e., over-predicting) than the rpart models (up to 0.8, i.e., under-predicting). The statistical models tend to outperform the existing OEH forecasting practice. Overall, the randomForest models appear to outperform the rpart benchmark models for Sydney SW and Sydney NW - this is considered to be important, as ozone exceedance occurs more often in the south-west and north-west of Sydney (EPA, 2012).

Both the rpart and randomForest models fail to identify the (small number of) poor air quality days for Sydney E (indicated by zero POD and CSI values). However, this is expected and is not considered to be significant model under-performing, since poor air quality is rarely observed in this subregion due mainly to its coastal location. Overall, compared to the rpart benchmark models, the randomForest models provide increased forecasting skills on the testing data set for Sydney SW (4.0), but similar or slightly decreased skills for Sydney NW (-0.3) and Sydney E (-2.7).

**Table 4.** Evaluation on the training data set: performance statistics for the randomForest and rpart based models, in comparison with those for the existing OEH forecasting practice. A 15-fold cross-validation (CV) is applied in rpart, while two-way bagging/bootstrapping is applied in randomForest. Evaluation cut-off: between fair and poor.

| Region/subregion | Model | A | POD | FAR | CSI | B | Skill** |
|---|---|---|---|---|---|---|---|
| Sydney | CAS* | 86.7 | 32.4 | 45 | 25.6 | 0.6 | |
| Sydney SW | rpart | 89.9 | 46.9 | 31.8 | 38.5 | 0.7 | |
| | randomForest | 94.3 | 84.6 | 12.0 | 75.9 | 1.0 | 4.9 |
| Sydney NW | rpart | 88.2 | 0.0 | NA | 0.0 | 0.0 | |
| | randomForest | 96.3 | 66.7 | 33.3 | 50.0 | 1.0 | 9.2 |
| Sydney E | rpart | 95.1 | 0.0 | NA | 0.0 | 0.0 | |
| | randomForest | 98.6 | 0.0 | NA | 0.0 | 0.0 | 3.7 |

*Evaluated using OEH's forecast data for the Sydney basin as a whole over 2008-2013 (Jiang *et al*., 2014a).
**Skill of the randomForest model is calculated in reference to the rpart benchmark model

**Table 5.** Evaluation on the testing data set: performance statistics for the randomForest and rpart based models, in comparison with those from the existing OEH forecasting practice. Evaluation cut-off: between fair and poor.

| Region/subregion | Model | A | POD | FAR | CSI | B | Skill** |
|---|---|---|---|---|---|---|---|
| Sydney | CAS* | 86.7 | 32.4 | 45 | 25.6 | 0.6 | |
| Sydney SW | rpart | 84.7 | 41.7 | 44.4 | 31.3 | 0.8 | |
| | randomForest | 88.1 | 83.3 | 44.4 | 50.0 | 1.5 | 4.0 |
| Sydney NW | rpart | 87.6 | 0.0 | NA | 0.0 | 0.0 | |
| | randomForest | 87.3 | 66.7 | 42.9 | 44.4 | 1.2 | -0.3 |
| Sydney E | rpart | 95.0 | 0.0 | NA | 0.0 | 0.0 | |
| | randomForest | 92.4 | 0.0 | NA | 0.0 | 0.0 | -2.7 |

* Evaluated using OEH's forecast data for the Sydney basin as a whole over 2008-2013 (Jiang *et al*., 2014a).
** Skill of the randomForest model is calculated in reference to the rpart benchmark model

## 4. Summary and Conclusion

This paper has explored the utility of the random forecast algorithm in comparison with the classic CART algorithm for forecasting ozone pollution in Sydney. The results have shown that both methods have a potential for forecasting ozone pollution in Sydney with an accuracy at least similar to (if not better than) the existing OEH forecasting practice (where ozone pollution is forecasted for the Sydney region as a whole). In general, however, the random forest models appear to outperform the CART models for forecasting ozone exceedance days, as indicated by high probability of detection and critical success index, particularly in the southwest and northwest subregions where highest pollution are observed. Moreover, the random forest models also show a lift in forecasting skills for the southwest subregion if compared to the existing practice.

In conclusion, the present study demonstrates that random forest is a promising method for air quality forecasting in Sydney. In a wider sense, most existing statistical AQF practice adopts *single-model development approach*. This paper promotes the application of a *statistical ensemble approach* to air quality forecasting. It is acknowledged that the statistical models obtained from this study can be improved in many different ways, for example, by including mixing height and rainfall data in the model building process or removing those less critical predictive variables from the models. Moreover, a reduction in false alarm rates is another area of future research. One aim of air quality forecasts is to improve health outcomes by advising people when to take preventative actions to minimise their exposure to high pollution - to ensure that such outcomes are realised, it is important to minimise false alarms, thereby ensuring that advice to take preventative action is considered seriously by the population. Also, how the random forest method compares to other techniques, such as

neural networks, non-linear multiple regression and Bayesian updating, for developing statistical AQF models can be examined in our future research for the same study region or a wider context.

## Acknowledgements

Thanks to colleagues in the Climate and Atmospheric Science Branch, New South Wales Office of Environment and Heritage (NSW OEH) for supporting the conduction of this project, in particular to Dr. Yvonne Scorgie for facilitating this study, Mr. David Salter for providing air quality data, and Dr. Lisa Chang for providing Figure 1. The Australian Bureau of Meteorology (BOM) provided the meteorological data for three airport meteorological stations in Sydney.

## References

[1] ABS. 2011. 2011 Census QuickStats. Australian Bureau of Statistics. Access to www.censusdata.abs.gov.au/census_services/getproduct/census/2011/quickstat/1GSYD?opendocument&navpos=220 on 1st May, 2014.

[2] Anderson JO, Thundiyil JG, Stolbach A. 2012. Clearing the Air: a Review of the Effects of Particulate Matter Air Pollution on Human Health. Journal of Medical Toxicology 8: 166-175.

[3] Barnett AG. 2012. Air pollution trends in four Australian cities 1996-2011. Air Quality and Climate Change 46(4): 28-33.

[4] Barnett AG, Williams GM, Schwartz J, Best TL, Neller AH, Petroeschevski AL, Simpson RW. 2006. The effects of air pollution on hospitalizations for cardiovascular disease in elderly people in Australia and New Zealand cities. Environmental Health Perspectives 114(7): 1018-1023.

[5] Breiman L. 1996. Bagging predictors. Machine Learning 24 (2): 123–140.

[6] Breiman L. 2001. Random forests. Machine Learning 45(1): 5–32.

[7] Breiman L, Cutler A, Liaw A, Wiener M. 2013. Package "randomForest" version 4.6-7. Access to The Comprehensive R Archive Network http://cran.r-project.org/ on 23rd June, 2014.

[8] Breiman L, Friedman J, Olshen RA, Stone CJ. 1984. Classification and Regression Trees. 1st Edition. Chapman & Hall: USA.

[9] EPA. 2012. New South Wales State of Environment 2012. NSW Environment Protection Authority: Sydney, Australia.

[10] EPHC. 2010. Expansion of the Multi-city Mortality and Morbidity Study, final report by University of the Sunshine Coast, University of Queensland, Department of Environmental Protection Western Australia, Environment ACT, Environment Protection Authority Victoria, NSW Health, New Zealand Ministry for the Environment, Queensland Health to the Environment Protection and Heritage Council, Canberra. Environment Protection and Heritage Council (EPHC) [www.scew.gov.au/archive/air/#multi].

[11] Han J, Kamber M. 2006. Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufmann Publishers: San Francisco, CA.

[12] Hart M, De Dear R, Hyde R. 2006. A synoptic climatology of tropospheric ozone episodes in Sydney, Australia. International Journal of Climatology 26: 1635-1649.

[13] Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning: data mining, inference, and prediction. Springer: USA.

[14] Honore C, Ung A, Corbet L, Malherbe L. 2008. CITEAIR II Common Information to European Air: Good Practice Guide to Urban Air Quality Forecast. European Union (EU) European Regional Development Fund Regional Initiative Project.

[15] Jiang N, Betts A, Quigley S, 2013. Visualisation of climate and air quality data using self-organising maps. In proceedings of 21st International Clean Air and Environment Conference (CASANZ), 7-11 September 2013, Sydney.

[16] Jiang N, Luo K, Beggs PJ, Cheung K, and Scorgie Y. 2014b. Insights into the implementation of synoptic weather-type classification using self-organizing maps: an Australian case study. International Journal of Climatolology doi: 10.1002/joc.4221

[17] Jiang N, Riley M, Scorgie Y, Betts B, Kirkwood J, Duc H, Trieu T, Salter D, Ji F, Chang L. 2015. Enhancing air quality forecast in New South Wales. In proceedings of 22nd International Clean Air and Environment Conference (CASANZ), 20-23 September 2015: Melbourne.

[18] Jiang N, Salter D, Dutt U, Alan B. 2014a. 2008-2013 air quality forecast performance evaluation. CAS Technical paper CAR-Tech-001. New South Wales Office of Environment and Heritage: Australia.

[19] Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society 77: 437–471. DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

[20] Katestone Scientific. 1997. Anthropogenic Influences on Australian Urban Airsheds. A report to the Australian Academy if Technological Science and Engineering: Brisbane, Australia.

[21] Leighton RM, Spark E. 1997. Relationship between synoptic climatology and pollution events in Sydney. International Journal of Biometeorology 41: 76–89.

[22] Liaw A, Wiener M, 2002. Classification and Regression by randomForest. R News 2/3: 18-20.

[23] Liaw A, Wiener M, 2013. Package 'randomForest' Version 4.6-7. Access to The Comprehensive R Archive Network http://cran.r-project.org/ on 1st July, 2014.

[24] NEPC. 2003. National Environment Protection (Ambient Air Quality) Measure – As varied 7 July 2003. Environment Protection & Heritage Council, Level 5, 81 Flinders Street, Adelaide, SA 5000, Australia.

[25] OEH. 2014. Air Quality Trends in Sydney. New South Wales Office of Environment and Heritage: Sydney, Australia.

[26] Shapire R, Freund Y, Bartlett P, Lee W. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics 26:1651–1686.

[27] Therneau T, Atkinson B, Ripley B. 2014. Rackage "rpart" Version 4.1-8. Access to The Comprehensive R Archive Network http://cran.r-project.org/ on 23rd June, 2014.

[28] US EPA. 2003. Guidelines for developing an air quality (ozone and PM2.5) forecasting system. EPA-456/R-03-002. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Part, North Carolina.

[29] US EPA. 2015. How effective are air quality alerts in reducing adverse effects in the real world. U.S. Environmental Protection Agency website: http://www.epa.gov/o3healthtraining/aqi.html#alerts, accessed on 8 May 2015.

[30] White J. 2011. EPA AIRNow Program. Air Quality Index Reporting and Forecasts. Presentation at the 1st Workshop on Satellite Observations for Air Quality Management. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Part, North Carolina.

[31] WHO. 2006. WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide. Global Update 2005. Summary of Risk Assessment. World Health Organization, Geneva.

[32] WHO. 2014. Ambient (outdoor) air quality and health. World Health Organisation http://www.who.int/mediacentre/factsheets/fs313/en/ accessed on 17 April, 2015.

[33] Zhang Y, Bocquet M, Mallet V, Seigneur C, Baklanov A. 2012a. Real-time air quality forecasting, part I: History, techniques, and current status. Atmospheric Environment 60: 632-655.

[34] Zhang Y, Bocquet M, Mallet V, Seigneur C, Baklanov A. 2012b. Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. Atmospheric Environment 60: 656-676.