APPLYING HYBRID FEATURE SELECTION METHODS FOR STATISTICAL MODELLING OF ROADSIDE PARTICLE CONCENTRATIONS ($PM_{2.5}$ AND PNC)

A. SULEIMAN[1], M.R. TIGHT[2], & A.D. QUINN[2]
[1] Department of Civil Engineering, Faculty of Engineering, Bayero University Kano, Nigeria
[2] Department of Civil Engineering, School of Engineering, University of Birmingham, UK

ABSTRACT

The task of selecting a predictor variable to include in statistical models is enormous. A model built with fewer predictor variables can be more interpretable and less expensive than the one built with many input variables. In this study, the effects of hybrid feature selection methods (genetic algorithms [GA] and simulated annealing (SA) each combined with random forests [RF]) in improving the efficiency of five variants of multiple linear regression models in the prediction of roadside $PM_{2.5}$ and particle number count (PNC) concentrations are investigated. The GA-RF and SA-RF selected 9 and 16 variables, respectively, of the 27 predictor variables in the $PM_{2.5}$ training data. Thirteen variables were selected by the GA-RF of the 25 possible variables in the PNC training data, while the SA-RF selected 13 variables. The methods selected variables that are nearly the same especially for predicting PNC, while for the $PM_{2.5}$ models the SA-RF selected 16 variables and the GA-RF selected only 10 variables. The hybrid feature selection methods eliminated most of the correlated variables, especially the background pollutants and the traffic variables. Whereas the temporal variables and the meteorological variable have been selected in all the cases considered. The statistical performance of the linear models with the selected variables is similar to those developed using the entire predictor variables. The actual benefit derived from this study is the successful reduction in the number of predictor variables by more than half in most of the cases considered. The reduction in the number of variables will eventually result in the reduction of the operational and computational cost of the models without possibly compromising the predictive performance of the models. Also, the reduction in the number of variables will enhance interpretability.

Keywords: air quality, genetic algorithms (GA), particulate matter, random forests (RF), simulated annealing (SA), statistical modelling.

1 INTRODUCTION

The choice of which predictor variable to include in a model is one of the difficult task air quality modellers often encountered when dealing with statistical and machine learning methods. This issue becomes apparent as the data are increasingly becoming large and multidimensional, and also the computational efficiencies of computing machines are being multiplied. A model that is built with fewer predictor variables can be more interpretable and less expensive compared with the one built with many input variables. The air quality predictor variables are often costly to measure and maintained for a very long time. Moreover, the models tend to be less efficient when built with so many correlated predictor variables. Therefore, the air quality models developed with less predictor variables are likely to be less expensive than those with the higher number of predictor variables. The immediate solution to this problem is to optimise the use of the predictor variables so that fewer variables are used without compromising the efficiency of the intended model. Feature selection techniques are invoked for this purpose such that more interpretable and relatively cheaper models are obtained. Some modelling methods like ensemble regression trees have built-in mechanisms for feature selection. However, simpler methods, such as multiple linear regression (MLR)

ISSN: 2398-2640 (paper format), ISSN: 2398-2659 (online), http://www.witpress.com/journals
DOI: 10.2495/EI-V3-N2-101-111

and its variants and more sophisticated methods such as artificial neural networks, support vector machines and lots more, require feature selection as part of their modelling process.

The feature selection methods can be broadly categorised into filter and wrapper methods. The wrapper methods consider the relationships between the predictor variables and response variables during their selection process, while the filter methods select their variables without regard for the response variables. The advantage of the wrapper methods over filter method is that it reduces the number of predictor variables such that more efficient and interpretable models can be obtained. They used subsets of predictor variables as inputs while considering the performance of the models as the output to be optimised [1]. The advantage of filter methods is that they are faster. However, they do not consider the efficiency of the models during the selection process. The wrapper methods are slow, thereby requiring more computational effort than the filter methods. Moreover, there is also a risk of overfitting when using wrapper methods as they aggressively search the dataset. In this work, the two wrapper methods namely genetic algorithms (GAs) and simulated annealing (SA) in combination with a random forests (RF) algorithm are considered. The effects of the hybrid feature selection methods in improving the efficiency of the five popular linear regression methods are investigated. The methods include MLR, partial least square regression (PLSR), principal component regression (PCR), stepwise regression and lasso/elastic-net regressions.

2 METHODS

2.1 Study Area and Data

This study used historical data on traffic, pollutants and the meteorological variables collected between 2007 and 2012 in the case of $PM_{2.5}$ and 1-year data in the case of particle number count (PNC). The breakdown of data is shown in Fig. 1. The meteorological data were collected from Heathrow airport weather station which is believed to be representative of the meteorological condition of London. While the traffic and the pollutant data were collected from the Marylebone road air quality monitoring station through London air quality network [2].

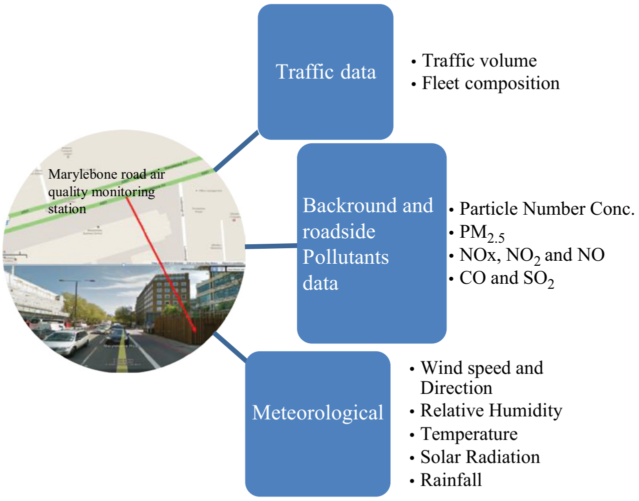


Figure 1: Study area and data.

2.2 Statistical modelling methods

The statistical methods used in this study include MLR, stepwise regression, Lasso regression, elastic-net regression, PCR and PLSR. These methods were selected because of their popularity in many air quality studies and other environmental studies [3–8]. Besides their popularity, the other four methods were considered based on their individual improvement over ordinary MLR. The objective of linear regression is to find a plane that minimises the sum-of-squared errors (SSE) between the observed and the predicted response [see eqn (1)]:

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{1}$$

where y_i is the outcome and $\hat{y}_i$ is the model prediction of that sample's outcome

Mathematically, the optimal plane can be shown to be:

$$\beta = \left(X^T X \right)^{-1} X^T y \tag{2}$$

where X is the matrix of predictor variables and y is the response vector. Equation (2) is a vector that contains the coefficients for each predictor. The MLR is easy to model and interpret the relationships between the predictor and the response variables. However, despite its simplicity, MLR is not robust in handling the trade-off between bias and the variance in the least square estimates, and it minimises only the bias component. Also, it describes only linear relationships and it cannot handle a case when a number of samples are greater than the number of predictors resulting in overfitting and consequently poor predictions on future observations not used in model training [9]. Moreover, it has difficulty in dealing with highly correlated variables [1]. Although the remaining methods are also linear in nature, they were developed with various improvements over ordinary MLR, the basis which formed their inclusion in this study.

PLSR and PCR methods use principal component analysis to transform the feature space into new sets of uncorrelated variables. The principal component analysis reduces the dimensionality of the input space which decreases the requirements for capacity and memory and increases efficiency given the processes taking place in a smaller dimension. The newly created variables are expected to have low noise sensitivity and are introduced to cater to the problem of handling highly correlated variables in MLR [10]. The main disadvantages of PCA are that the covariance matrix is difficult to be accurately evaluated and even the modest invariance could not be captured by the PCA unless the information is explicitly provided in the training data [10]. The difference between the two methods is that PLSR estimates its latent variables in consideration of their effect on the response variable, while the PCR estimates its components without consideration of the response variables.

Stepwise regression is a popular modification of MLR with variable selection property which combines the backward and forward procedure. The predictor variables are tested for addition or removal from multivariate regression models using forward and backward stages, respectively. The variables are retained or dropped based on their statistical significance. Lower and upper boundaries of p-values of F-statistics are set such that for a variable to be kept in the model or removed must satisfy those boundaries [11]. The stepwise regression has an advantage in avoiding the collinearity issues of the MLR [12]. The major limitations of stepwise regression consist of bias in parameter estimation, inconsistencies among model selection algorithms and dependence on a single best model [13]. The stepwise regression method has been applied in many studies involving air quality [4, 14–17].

The Lasso/elastic-net [18], although not so popular in the air quality studies [19, 20], are forms of penalised regressions aimed at reducing the variances in the least square estimates by using bias-variance trade-off. The penalty is added to the sum of the squared errors as the estimates become large. This trade-off between variance and the bias ensures a modest reduction in the mean squared errors (MSEs) which translates to a better estimate. The estimates shrink to zero when the penalty becomes large. Therefore, feature selection becomes possible as the predictors with zero coefficients are discarded. The ability of the Lasso/elastic–net methods to carry out feature selection could help improve their predictive ability.

2.3 Hybrid feature selection

The hybrid feature selection methods referred here combine the powers of search algorithms (GA and SA) each and that of the RF to give GA-RF and SA-RF. This combination is aimed at using the capabilities of the search algorithms in finding the possible subsets of the predictor variables that will optimise the out-of-bag errors estimated by the RFs.

A GA is one of the methods that mimics the biological evolutionary processes [21]. The algorithms are based on the biological reproduction principles where the training datasets are considered to represent the population, and the data subsets are considered as individual candidates (chromosomes) that undergo reproduction process to produce offspring. The implementation of feature selection using GA by [22] in R software [23] is adopted in this work. The algorithm carries out repeated search in the feature space within the resampling iterations. Initially, the resampling method is specified in the control function, and then the entire GA process is implemented separately on each sample. Here the 10-fold cross-validation repeated five times was adopted as the resampling method for the external performance. Therefore, for the first fold, the search is conducted on the nine–tenth of the training data while estimating the external performance with the remaining tenth. The optimal number of generations is determined using the external performance since it does not take part in the search process. However, during the search, there is a need for the internal performance to guide the search, and this is determined using another resampling within the selected data. This procedure has the potential of overfitting the estimates; that is why the external performance is used for the selection of the final predictors.

The SA method is a global search technique that mimics the metal cooling process [24]. The algorithm randomly makes small changes to the initially selected subset of predictor variables. The perturbed subset is then used to create a model, and the initial error is estimated. The same procedure is repeated, and the error for the new model is compared with the previous error. If the performance of the new model is better than the previous model, then the current set of predictors is accepted. Otherwise, a probability of acceptance is determined based on the difference between the performance of the two models and the current iteration of the search. The probability is estimated such that it decreases as the number of iterations becomes large, making it difficult for a suboptimal model to be accepted. The process is repeated until the specified number of iterations is reached, and the optimal combination of predictors is determined. The estimates of the internal and external performance follow the same procedure as that of the GA method.

RF method is one of the variants of ensemble learning techniques designed to improve the prediction accuracy of regression trees [25]. In this method, bagged regression trees are built using bootstrapped subsets of the training data so that the final model is the average of all the individual trees. The out-of-bag errors of the individual trees are estimated using the remaining samples of the training data left during the resampling process. The averaging of the trees

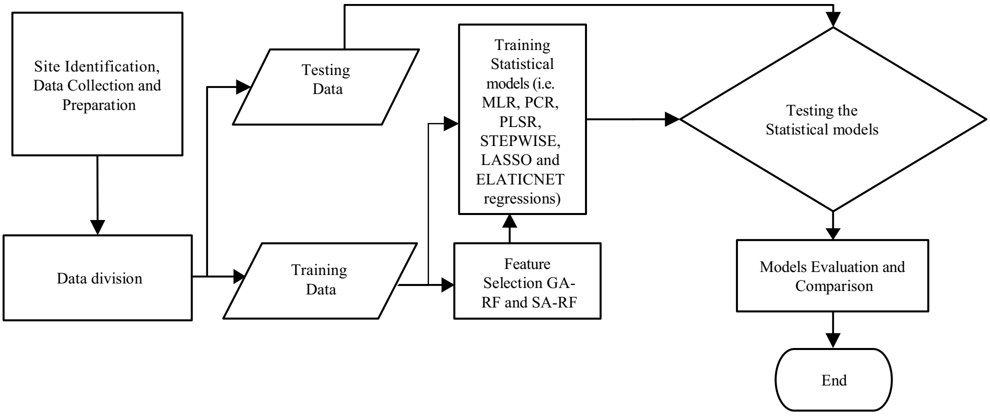

Site Identification, Data Collection and Preparation
Data division
Testing Data
Training Data
Training Statistical models (i.e. MLR, PCR, PLSR, STEPWISE, LASSO and ELATICNET regressions)
Feature Selection GA-RF and SA-RF
Testing the Statistical models
Models Evaluation and Comparison
End

Figure 2: Flow chart for the statistical modelling process.

reduces the overall variance in their estimates. However, the trees are correlated in one way or the other, which limits the reduction of the overall variance in the estimates. RF method seeks to de-correlate the trees by introducing randomness in the tree-building process. The algorithm first selects the predictor variables at random and then selects the best predictors out of the random samples to partition the data. This process reduces the variance in the estimates of the individual trees and in turn reduces the overall variance in the final estimate.

2.4 Modelling

The process began with the selection of the monitoring site with the available data. The data collected were then processed to identify the nature and completeness of the data (see Fig. 2).

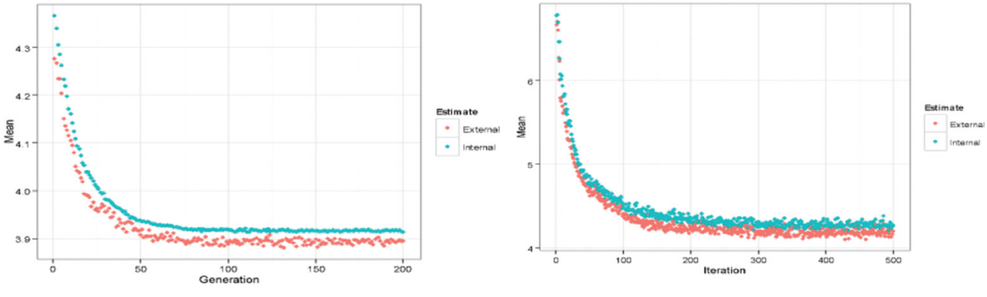
The next step was the determination of the most relevant predictor variables among the various predictor variables collected for the modelling using the hybrid feature selection methods. The data with the appropriate number of inputs were then divided into the training and testing dataset through the use of K-fold cross-validation. The next step was the training and testing of the statistical models. The models were then evaluated using the model performance metrics provided in *openair* by Carslaw and Ropkins [26] in addition to the scatter plots and conditional quantile plots. The final task is the comparison of the performances of the models for each target pollutant.

3 RESULTS

3.1 Application of hybrid feature selection to the statistical models.

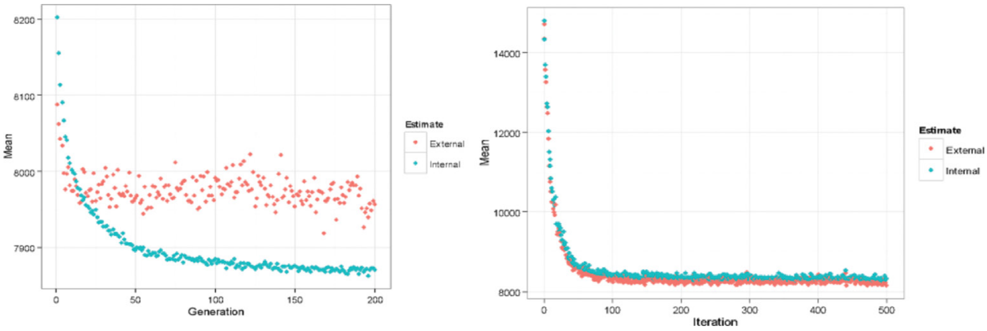
The genetic algorithms combined with random forests (GA-RF) and simulated annealing combined with random forests (SA-RF) were applied to the samples drawn from the training data for predicting $PM_{2.5}$ and PNC. Out-of-bag RMSE and R^2 were used as measures of the internal performance, while 10-fold cross-validation repeated five times was the resampling methods used to estimate the RMSE and R^2 for the external performances. The external and internal performance of the feature selection for $PM_{2.5}$ and PNC models are shown in Figs. 3 and 4.

Using the $PM_{2.5}$ training data, the training performance of the GA-RF measured by RMSE and R^2 were estimated to be 3.88 μg/m^3 and 0.87, respectively. Nine variables were selected



(a) (b)

Figure 3: External and internal performances of (a) GA-RF and (b) SA-RF feature selection for $PM_{2.5}$.



(a) (b)

Figure 4: External and internal performances of (a) GA-RF and (b) SA-RF feature selection for PNC.

out of the 27 possible variables in the training data. For the SA-RF, which selected 16 variables of the 27, the training RMSE and R^2 were 4.10μg/m^3 and 0.87, respectively. For the PNC training data, the training RMSE and R^2 for the GA-RF were estimated to be and 7918.43 number/cm^3 and 0.94, respectively. Thirteen variables were selected of the 25 possible variables in the training data. SA-RF also selected 13 variables of the 25 variables for training PNC models and the training RMSE and R^2 were 8152.82 number/cm^3 and 0.93, respectively. The internal and external performances follow the same pattern as shown in Figs. 3 and 4. However, the GA-RF shows that the external performance is slightly higher than the internal performance for the $PM_{2.5}$ case while it is much higher for the PNC data. This trend is expected since the internal performance procedure has some chance of overfitting the data. For the SA-RF method, the external performance is slightly better than the internal in the both cases. Table 1 shows the variables selected by the hybrid feature selection methods for the $PM_{2.5}$ and PNC models.

The methods selected the variables that are nearly the same especially for predicting PNC, while for the $PM_{2.5}$ models, the SA-RF selected 16 variables and the GA-RF selected only 9 variables (see Table 1). The general pattern in their selection is that they have eliminated most of the correlated variables, especially the background pollutants and the traffic variables.

Table 1: Variables selected by hybrid feature selection methods.

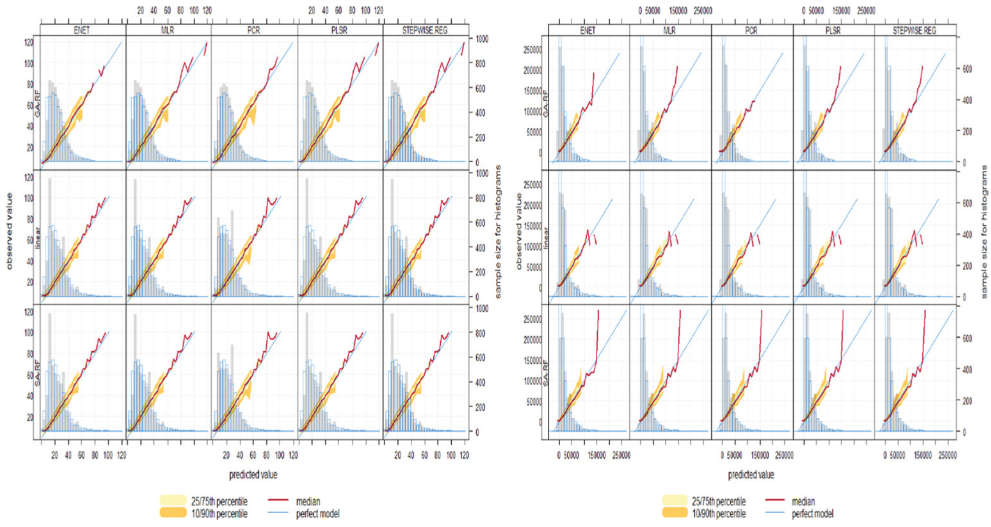
Predictor variables	GA-RF selected variables for $PM_{2.5}$ models	SA-RF selected variables for $PM_{2.5}$ models	GA-RF selected variables for PNC models	SA-RF selected variables for PNC models
Articulated HGV[8]				
B. pressure[1]	✓	✓		
Bus and coach				
CO				
CO.bg[2]		✓		
Day of the week	✓	✓	✓	✓
Diesel car		✓		
Hour of the day		✓	✓	✓
Julian day	-	-	✓	✓
LGV[11]				
Month of the year	✓	✓		
Motorcycle		✓		✓
NO			✓	✓
NO.bg[3]			✓	
NO_2				
NO_2bg[4]	✓		✓	✓
NO_x	✓	✓	✓	✓
NO_x.bg[5]				
Petrol car				✓
PM.bg[6]	✓	✓	✓	✓
R. humidity[7]	✓	✓	✓	✓
Rainfall		✓		
Rigid HGV[8]		✓		
SO				
SO_2.bg[9]		✓		
Solar Rad[10]			✓	
Taxi				
Temperature	✓	✓	✓	✓
Wind direction			✓	✓
Wind speed	✓		✓	✓
Year		✓		✓

[1]B.Pressure = Barometric Pressure; [2]CO.bg = Background CO; [3]NO.bg = Background NO; [4]NO_2 = Background NO_2; [5]NOx.bg = Background NO_x; [6]PM.bg = Background Particulate Matter; [7]R. Humidity = Relative Humidity; [8]HGV = Heavy Goods Vehicles; [9]SO_2.bg = Background SO_2; [10]Solar Rad = Solar Radiation; [11]LGV = Light Goods Vehicles

Whereas the temporal variables and the meteorological variables are selected for the both $PM_{2.5}$ and PNC. These variables were less significant for the linear models when run without feature selection. Their inclusion here suggests that they might have a non-linear relationship between the predictor variables or their correlation with other predictors makes it impossible for the linear models to discover their true relationships with the response variables.

The test performance of the models was measured using fraction of predictions within a factor or two (FAC2), mean bias (MB), normalised mean gross error (NMGE), root mean squared error (RMSE) and coefficient of efficiency (COE). For the $PM_{2.5}$, they were found to be 0.97, –0.04, 0.15, 5.25 and 0.67, respectively. Also, for the PNC model, the FAC2, MB, NMGE, RMSE and COE values were 0.96, 129.87, 0.17, 9943.52 and 0.71, respectively. The performance was largely similar between the models developed using the complete set of variables and those selected by the GA-RF and SA-RF. However, in some instances, the models with the selected variables tend to have lower RMSE and MB values, but these differences are too little to consider it an advantage over the normal linear models. The actual benefit derived from this exercise is the successful reduction in the number of predictor variables by more than half in most of the cases considered. The reduction in the number of variables will eventually result in the reduction of the operational and computational cost of the models without possibly compromising the predictive performance of the models. Since the performance measures used did not show much disparity in models, it is necessary to use graphical methods to further evaluate the performance of the models as shown in Figs. 5 and 6.

The conditional quantile plots in Fig. 5 show that the models developed with GA-RF captured the higher values slightly better than the other two models. The same feature is also reflected in Fig. 6, which further revealed that GA-RF has fewer predictions outside the FAC2 boundaries.



(a) (b)

Figure 5: Conditional quantile plots comparing the performance of (a) $PM_{2.5}$ and (b) PNC models.

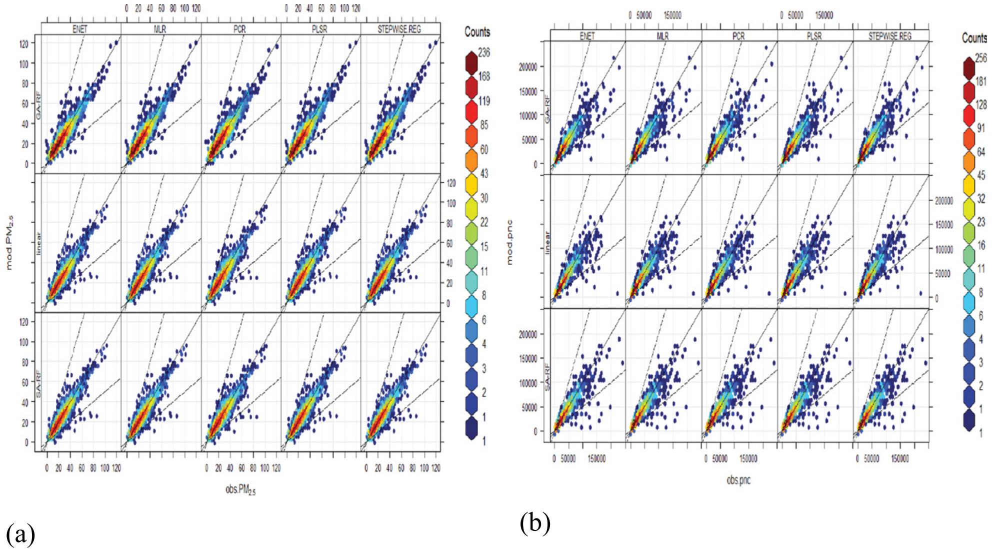


Figure 6: Scatter plots comparing the performance of (a) PM$_{2.5}$ and (b) PNC models.

The PNC models developed with the features selected by the GA-RF and SA-RF have shown more data coverage than the linear model (see Fig. 5). However, the higher concentrations were poorly predicted by the models. The SA-RF linear models performed poorer than the GA-RF linear models in that respect (see Fig. 6). The scatter plots show that the SA-RF linear models have more of its prediction outside the FAC2 boundaries than the remaining two model types.

4 CONCLUSIONS

This study investigates the effect of using two hybrid feature selection methods (GA-RF and SA-RF) on the prediction performance of five statistical models. The results obtained show that there was a slight increase in performance when using feature selection before statistical modelling. Moreover, using the hybrid feature selection resulted in a remarkable reduction in the number of predictor variables which consequently reduces the operational and computational costs of the models. The main advantage of using the linear models lies in their ability to produce models that can be interpreted, but this quality is often diminished when the predictor variables are many and are correlated. The feature selection methods successfully selected variables that are less correlated and are quite small in number, and that will enhance interpretability. However, where the relationship between the predictor variables and the response variables are nonlinear as is the case in air quality modelling, the models might not capture the underlying relationships. These shortcomings limit the use of the linear models to the only prediction rather than to be used for analysing air quality problems based on the relationships of the variables expressed by the models. Therefore, invoking methods that are more sophisticated in handling nonlinear relationships will offer more benefit than using the linear methods if the prediction performance is the primary goal.

ACKNOWLEDGEMENTS

The authors would like to thank Petroleum Technology Development Fund (PTDF) for funding this research and also the London Air Quality network and Environmental research group at Kings College London for their help in providing data and Openair package for the data analysis.

REFERENCES

[1] Kuhn, M. & Johnson, K., *Applied Predictive Modeling*, Springer, 2013.

[2] LondonAir. (2013, 03/04/2013). *London Air quality Network*. Available: http://www.londonair.org.uk/london/asp/datadownload.asp

[3] Benas, N., Beloconi, A. & Chrysoulakis, N., Estimation of urban PM10 concentration, based on MODIS and MERIS/AATSR synergistic observations. *Atmospheric Environment*, **79**, pp. 448–454, Nov 2013.

[4] Chen, Y. Y., Shi, R. H., Shu, S. J. & Gao, W., Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis. *Atmospheric Environment*, **74**, pp. 346–359, Aug 2013.

[5] de Paula, P. H. M., Mateus, V. L., Araripe, D. R., Duyck, C. B., Saint'Pierre, T. D. & Gioda, A., Biomonitoring of metals for air pollution assessment using a hemiepiphyte herb (Struthanthus flexicaulis). *Chemosphere*, **138**, pp. 429–437, Nov 2015.

[6] Deka, P., Bhuyan, P., Daimari, R., Sarma, K. P. & Hoque, R. R., Metallic species in PM10 and source apportionment using PCA-MLR modeling over mid-Brahmaputra Valley. *Arabian Journal of Geosciences*, **9**, May 2016.

[7] Guo, X. Y., Li, C., Gao, Y., Tang, L., Briki, M., Ding, H. J., *et al.*, Sources of organic matter (PAHs and n-alkanes) in $PM_{2.5}$ of Beijing in haze weather analyzed by combining the C-N isotopic and PCA-MLR analyses. *Environmental Science-Processes & Impacts*, **18**, pp. 314–322, 2016.

[8] He, H. D., Lu, W. Z., & Xue, Y. Prediction of particulate matters at urban intersection by using multilayer perceptron model based on principal components. *Stochastic Environmental Research and Risk Assessment*, **29**, pp. 2107–2114, Dec 2015.

[9] James, G., Witten, D. & Hastie, T., An Introduction to Statistical Learning: With Applications in R. ed, 2014.

[10] Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M. & Hooman, A., An Overview of Principal Component Analysis. *Journal of Signal and Information Processing*, **4**, p. 173, 2013.

[11] Singh, K. P., Gupta, S., Kumar, A. & Shukla, S. P. Linear and nonlinear modeling approaches for urban air quality prediction. *Science of the Total Environment*, **426**, pp. 244–255, Jun 2012.

[12] Chen, Y., Shi, R., Shu, S. & Gao, W., Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis. *Atmospheric Environment*, **74**, pp. 346–359, 8// 2013.

[13] Whittingham, M. J., Stephens, P. A., Bradbury, R. B. & Freckleton, R. P., Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, **75**, pp. 1182–1189, 2006.

[14] Banerjee, T., Singh, S. B. & Srivastava, R. K., Development and performance evaluation of statistical models correlating air pollutants and meteorological variables at Pantnagar, India. *Atmospheric Research*, **99**, pp. 505–517, Mar 2011.

[15] Brown, T., Dassonville, C., Derbez, M., Ramalho, O., Kirchner, S., Crump, D., *et al.*, Relationships between socioeconomic and lifestyle factors and indoor air quality in French dwellings. *Environmental Research*, **140**, pp. 385–396, 7// 2015.

[16] Diaz-de-Quijano, M., Joly, D., Gilbert, D. & Bernard, N., A more cost-effective geomatic approach to modelling PM10 dispersion across Europe. *Applied Geography*, **55**, pp. 108–116, 12// 2014.

[17] Krivtsov, V., Howarth, M. J. & Jones, S. E., Characterising observed patterns of suspended particulate matter and relationships with oceanographic and meteorological variables: Studies in Liverpool Bay. *Environmental Modelling & Software*, **24**, pp. 677–685, Jun 2009.

[18] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, pp. 301–320, 2005.

[19] Simons, K., De Smedt, T., Van Nieuwenhuyse, A., Buyl, R. & Coomans, D., Ensemble post-processing is a promising method to obtain flexible distributed lag models. *Air Quality, Atmosphere & Health*, pp. 1–12, 2016.

[20] Suleiman, A., Tight, M. R. & Quinn, A. D. Hybrid Neural Networks and Boosted Regression Tree Models for Predicting Roadside Particulate Matter. *Environmental Modeling & Assessment*, pp. 1–20, 2016.

[21] Fouskakis, D. & Draper, D., Stochastic optimization: a review. *International Statistical Review*, **70**, pp. 315–349, 2002.

[22] Kuhn, M., The caret Package. 2012.

[23] R Development Core Team, "R 3.2. 1," ed: R Project for Statistical Computing Vienna, Austria, 2015.

[24] Lin, S.-W., Tseng, T.-Y., Chou, S.-Y., & Chen, S.-C., A simulated-annealing-based approach for simultaneous parameter optimization and feature selection of back-propagation networks. *Expert Systems with Applications*, **34**, pp. 1491–1499, 2// 2008.

[25] Breiman, L., Random forests. *Machine learning*, **45**, pp. 5–32, 2001.

[26] Carslaw, D. C. & Ropkins, K., openair — An R package for air quality data analysis. *Environmental Modelling & Software*, **27–28**, pp. 52–61, 2012.