Available online at www.elixirpublishers.com (Elixir International Journal)

Statistics

Elixir Statistics 89 (2015) 37217-37220

ISSN: 2229-712X

Modelling the stochastic boosted regression trees (brt)'salgorithm on ground level ozone (o3) concentration

Asri M.A.M[1], Yahaya, N.Z[2] and Sabri Ahmad[1]

[1]Department of Mathematics, School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu.

[2]Department of School of Fundamental and Liberal Education, Universiti Malaysia Terengganu.

ARTICLE INFO

Article history:

Received: 22 April 2015;

Received in revised form:

23 December 2015;

Accepted: 28 December 2015;

Keywords

Stochastic,

Boosted Regression Trees,

Ozone,

Modelling Algorithm,

R Statistics.

ABSTRACT

This paper aim to discuss the preliminary study of modelling algorithm setting of Stochastic Boosted Regression Trees (BRT) using R statistical software. The ground level ozone (O3) concentrations from a Malaysia's air quality monitoring station was used as a case study. The new approach; BRT developed by Friedman (2001), Friedman (2002) and adopted by Yahaya (2013) in air pollution studies. BRT is differ substantially from traditional regression-based approaches. BRT are constructed of multiple regressions models which finally produce a single 'best iteration' model to optimize prediction performance. Sensitivity testing of model been carried out to determine the best parameters' setting which are number of trees (nt) (1000-10000), learning rates (lr) (0.005-0.5), and interaction depth (tc) (1-10) to suits the ozone data. Results indicated that the BRT analysis algorithm best modeled with the best combination of parameters lr=0.001, tc=5 and nt=5301 that achieves minimum predictive error (minimum error for predictions). The algorithm model would crucially best provided clear benefit for air pollution study and their ability to model complex variable interactions and non-linear effects.

Introduction

Background

Previous research demonstrated that a Boosted Regression Trees (BRT) model that has stochasticity (SGM) can be used to produce the best model that can deal with a high level of complexity and large datasets to produce a substantial outcomes (Yahaya et al., 2011a; Yahaya et al., 2011b) and Yahaya (2013). Boosting is a general method that attempts as the name suggests to 'boost', i.e., improve the model accuracy of any given learning algorithm. Boosting was originally developed by Schapire, who first proved polynomial-time boosting algorithms in 1989, as reported in Schapire (1998). The Ada Boost algorithm was examined from a statistical perspective by Friedman et al. (2000) and Hastie et al. (2001). Friedman (2002) made a minor modification to the gradient boosting algorithm to include randomness as an essential part of the boosting procedure and then introduced Stochastic Gradient Boosted Regression Trees (SBRT) algorithm. The difference in SBRT algorithm is that at each iteration a subsample of the training data is selected at random and without replacement from the full training data set. This led to a series of theoretical and practical advances in our understanding of boosting, which realized the power and potential of boosting as a general method for approximation based on additive models. Regression trees models have been used in a few fields such as in ecological studies (De'ath et al., 2007; Leathwick et al., 2006), in the social sciences (Kriegler, 2007) and also in remote sensing (Lawrence et al., 2004).

A BRT approach uses two algorithms: regression trees from the classification and the regression tree (decision tree) group of models, and boosting improves these models by combining them to create a best model. This approach laid the foundations for a new generation of boosting algorithms, which are called Stochastic Boosted Regression Trees (SBRT). Stochasticity is controlled through a 'bag fraction' that specifies the proportion of data to be selected at each step. The default bag fraction is 0.5, meaning that in each iteration 50% of the data are drawn at random, without replacement, from the full training set. Optimal bag fractions can be established by comparing predictive performance and model-to-model variability under different bag fractions. There are five tuning parameters that need to be controlled (in addition to the distribution): the training sample size relative to the training population (bag.fraction), the number of iterations (nr), the learning rate (lr), the maximum tree depth (interaction depth), and the number of observations in each terminal node in developing the optimum number of iterations that need to be tested for different predictors. It has been proven that the BRT algorithm is a suitable method that can deal with different types of variables and able in handling big dataset.

Practically, only few studies use BRT in the environmental sciences and we could find no intensive study related to air pollution, particularly in which ozone is reported, except those by Carslaw and Taylor (2009), Yahaya et al. (2011a) and Yahaya et al. (2011b) and Yahaya (2013). The previous study conducted by Carslaw and Taylor (2009). Their study explored the use of BRT to understand the source characteristics of NOx at a location of high source complexity at London Heathrow airport, UK. The most recently published work that has applied BRT in the context of air pollution is by Yahaya (2013) that explore the used of particle number count concentration applied to the Stochastic Boosted RegressionTrees using R Software in the analysis. The extensive work was done in Leeds, United Kingdom and appears to be the first to explore and use BRT to analyse pollution data which is particle number count concentrations ([PNC]), traffic data and meteorological conditions intensively.

A wide range of methods have been used to study ozone in the atmospheric environment such as Classification and

Tele:

E-mail addresses: imanasri90@gmail.com

Regression Trees or CART (Breidman, 1984), which allows the estimation of different trends in different clusters. The latest ozone study that uses CART as a suitable technique for forecasting the daily exceeds of ozone standards was established by Italian law (Bruno et al., 2004). However, the CART approach has a limitation in that the analysis is more focused on using the episode selection method, which was used in Bruno et al. (2004) for air quality modelling of ozone and also for performing integrated control strategies analysis. The stochastic element in the BRT algorithm and the tuning parameters that control the iterations in developing the BRT model (from single models) gives this model an advantage over CART and has the potential to offer a robust approach for the analysis of O3 formation process and for its prediction. It has also been stated that although some ozone prediction models using BRT have been developed there is still a significant need for more accurate models (Ghazali et al., 2010; Ghazali et al., 2009).

Compared to other techniques, the BRT method has the ability to deal with complex data and has the following advantages/benefits:

i. A BRT model can provide a much smoother gradient, analogous to the fit achieved when using the gradient boosted machine (gbm) in the R software package. It can also handle sharp discontinuities, which is an important advantage when modelling the distributions of ozone that only occupy a small proportion of the sample environmental space;

ii. A BRT model can select the relevant variables, fit accurate functions, and it can automatically identify and identify the influence of various variables as well as interactions between them;

iii. A BRT model can predict the predictors from any excluded subsets of data or from any other datasets.

Methodology and Data Analysis

Model Development Processes

The use of the Boosted Regression Trees technique involved in this work comprises the meteorological parameters (prevailing wind variables) data collected from an air quality station. The 1-hour forty-eight month data (four calendar year) from one station and meteorological variables, data were recorded from one air quality monitoring stations.

Table 2.1. Variables names and descriptions used to model ozone boosting algorithm setting.

Respond Name	Variable names	Description and units
1	O3	Ozone (ppm)
		Predictor
2	nox	Nox (ppm)
3	no	
4	No2	
5	Ambient	Ambient Temperature (^{o}C)
6	Humidity	Humidity Percentage (%)
7	Ws	Wind Direction and Wind Speed 10m (km/hr) Average
8	Wd	Wind Direction (^{o}) and Wind Speed (m/s)

As tabulated in table 2.1, it is shown the list of variables names and corresponding descriptions for the overall data set used in the BRT analyses.

The model was fitted in R 3.0.2 software(R Development Core Team, 2008) using the gbm package version 1.6-3.1(Ridgeway, 2010). For all settings other than those mentioned, author used the defaults in gbm. The three model fitting typically needs the specification of three main parameters which are the number of trees (nt), learning rate(lr) and tree complexity(tc). The used of CV for selecting optimal settings is becoming increasingly common, (Hastie, 2001), by led by the mechine learning(ML) which focus on the predictive success. The process of obtaining the BRT algorithm and the analyses are concluded into three stages. But for this paper considered only first stage by author.

Ozone Boosting Algorithm Setting

For the Ozone boosting algorithm, importance features of BRT as applied in this study are as follows. First stage, the process is stochastic where it's includes a random or probabilistic component.

The aim here is to find the best combination of parameters (lr, tc and nt) that achieves minimum predictive error (minimum error for predictions to independent samples). The stochastic gradient machine technique, which introduces some randomness into the boosted model and usually improves the accuracy and speed, and reduces the over fitting (Friedman, 2002), was also applied in this model. At this stage the sequential model-fitting process builds on trees fitted where a series of BRT analyses using gradient boosted model (gbm) in different settings were performed to explore gbm before determining the tree complexity(tc), number of trees (nt) and learning rates (lr), also known as the shrinkage parameter that control the setting of the model.

The optimal iteration or learning rate, controls the rate at which model complexity is increased. The number of splits, termed the tree complexity tc), in gbm in each tree and the number of trees must also be identified at this stage. These two parameters (lr and tc) then determine the number of trees (nt) required for optimal prediction. The algorithm for simulating the Ozone dataset needs to be addressed for initial model development. In this work model statistics were derived using an independent dataset called database model data from four-year datasets. The dataset consisting of the Ozone independent database and 5 variables, in referring to Table 2.1, were fitted using BRT models with varying values of nt (100 – 10,000) and lr (0.1 – 0.0001) and interaction depth (5,10).

```
1
2  #recall the dataset
3  brto3<-read.csv("kemaman_brt.csv",header=T,na.strings="NA")
4  names(brto3)
5  summary(brto3)
6  # Load the  packages
7  library(sp)
8  library(rJava)
9  library(raster)
10 library(dismo)
11 library(lattice)
12 library(survival)
13 library(gbm)
14 library(reshape)
15 library(caret)
16 #working with First Test gbm1
17 #gbm1 test on the PNC and all met(roof top ws/wd)variables plus the traffic
18 # and excluded the time, julian day and the day
19
20 #nomber of tree = 8065, lr =0.001 and interaction depth 5
21 gbm1 <- gbm(o3 ~ co + so2 + nox + no + no2 + pm10 +
22 wd +ws + humidity + ambient,
23             data = brto3,
24             distribution="gaussian",
25             n.trees= 3580,              # number of trees
26             shrinkage=0.001,           # shrinkage or learning rate, 0.001
27             interaction.depth=5,       # 1: additive model, 2: two-way inte
28             bag.fraction = 0.5,        # subsampling fraction, 0.5 is proba
29             train.fraction = 0.5,      # fraction of data for training, fir
30             cv.folds = 10,               # do 5-fold cross-validation
31             keep.data = TRUE,
32             verbose = TRUE,
33             n.minobsinnode = 10)
34 summary(gbm1)
35 summary(gbm1,n.trees=1) # based on the first tree
36 best.iter <- gbm.perf(gbm1,method="cv")
37 summary(gbm1,n.trees=best.iter) # based on the estimated best number of tre
```

Figure 1. The Best Modelled BRT Algorithm Setting using the Ozone dataset

The *gbm* offers three methods for estimating the optimal number of iterations after the gbm model has been fitted, an independent test set (*test*), out-of-bag estimation (*OOB*), and v-fold cross validation (*cv*) to identify the optimal number of trees.

The optimal number of iterations based on the independent test set method uses a single holdout base dataset, which is similar to Friedman's MART software (Ridgeway, 2010). The BRT model with a different of *number of trees (nt)*, from 1000 to 10000 of *nt*, were simulated for datasets with the *nt* = 10000, *lr* = 0.001, the independent depth = 5, and CV fold = 10. The best number of iterations for cross-validation with a minimum square error from CV simulations was indicated and the number of trees of 10000 simulations is also performed to make it clear. The performance prediction of the model was done by estimating the optimal number of boosting iterations for a *gbm* object by obtaining the boosting model with minimum predictor error using the cross-validation method as discussed by De'ath (2007) graphically and from the best iteration output. A sample of the cross-validation performance plot to identify the best iteration or number of trees is shown in Figure 2 for example, in this work what is the best fit for the dataset. As illustrated in Figure 3.1, it is shown that the best methods for estimating the optimal number of iterations after the gbm model has been fitted is independent test set (test) with the least number of trees of 5301 by the maximum number of trees simulated in the algorithm which is 10000.

The different learning rates for values of 0.1, 0.01, 0.05 and 0.001 were examined to determine the best *learning rate* to fit in the dataset. The value of *lr* = 0.01 was found too fast for both *tc* values, and at each addition of trees above the minimum 100 trees, the predictive deviance increased, indicating that over fitting occurred almost immediately. Finally, author found that the best predictive performance of lr is 0.001 fitted, which suits thousands of datasets. A similar shrinkage value (*lr*) of 0.001 was also suggested by (Ridgeway, 2010) and applied by Carslaw and Taylor(2009) in tackling with air pollution dataset. The data are randomly partitioned into 10 subsets. Nine of the subsets are used to build the models and model performance statistics are calculated from the remaining data sets using the least squares error method. In the stochastic gradient boosting machine, the stochastic technique is controlled through the bag fraction that specifies the proportion of data to be selected at each step. The default bag fraction of 0.5, means that 50% of the data are drawn randomly at each iteration without replacement from the full training set data. According to Elith *et al.* (2008), the stochasticity improves predictive performance, reducing the variance of the final model, by using only a random subset of data to fit each new tree(Friedman, 2002). It was found from the iterations, that the best parameter setting to be simulated in the algorithm are number of trees of 3580, learning rate of 0.001, and tree complexity of 5 with the lowest RMSE of 10.

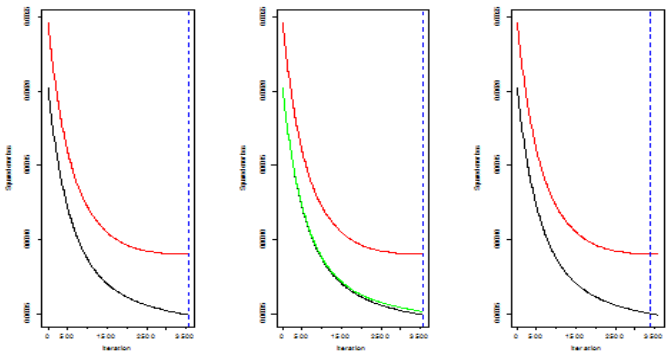


Figure 3.1 Out-of-bag estimation (*OOB*), v-fold cross validation (*cv*), and independent test set (*test*) respectively.

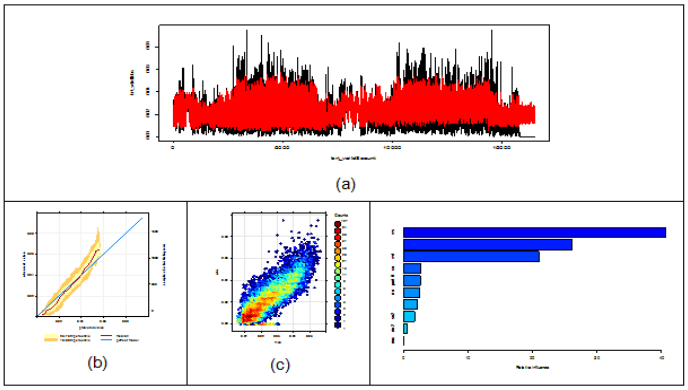
Model Performance and results

Carslaw and Taylor (2009) and Yahaya (2013) evaluate the predictive performance of candidate models through 10-fold CV. This can be done by model fit statistics, such as the R^2 and the residual sum of squares. In this study, we adopt the latest method available for model evaluation and comparison of models, as proposed in Willmott et al. (1985) who suggest that comparative analysis begins with an evaluation by means of the statistical analysis and graphic representation of the relationships between the modelled and observed variables. This method is also proposed by Derwent et al. (2010). The correlation of determination is R^2 = 0.722 between the observations and the fitted model obtained from this analysis show how well the BRT model fits.

The Index of Agreement (IOA) which based spans between -1 and +1between the modelled and observed data was found to above 0.6 for all this station, with 0.797. These scores are very good and therefore the developed model hasan acceptable IOA value.The correlation of coefficient (R) and R^2 between the observations and the fitted model obtained from this analysis show how well the BRT model fits as sown in Table….

Table Variables influence of the formation of Ozone

default	n	FAC2	MB	MGE	NMB	NMGE	RMSE	r	COE
all data	1653 9	0.7970	0.001894	0.00651	0.091307	0.31378	0.00835	0.8499	0.4664



Conclusion

This work is among the first to explore ozone data using the stochastics method in running the Boosted Regression Trees. Compared to other techniques, BRT method could explains the variability/ability to deal with complex data and the advantages/benefits; BRT trees provide a much smoother gradient, analogous to the fit compared to traditional method such as the least squares technique or multiple regressions. BRT can also handle sharp discontinuities, BRT has the ability to select relevant variables, fit accurate functions and automatically identifies and selects the model interactions and, BRTs can be examined to show how dependent variables respond to individual model variables.

Reference

ABDUL-WAHAB, S.A., BAKHEIT, C.S., AL-ALAWI, S.M. 2005. Principle Component and Multiple Regression Analysis in Modelling of Ground-level Ozone and Factors Affecting Its Concentrations. *Environmental Modelling & Softwares20*, p. 1263-1271.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A, & STONE. 1984. Classification and Regression Trees. *Wadsworth International Group.* Belmore, California, USA.

BRUNO, F., COCCHI, D. & TRIVISANO, C. 2004. Forecasting daily high ozone concentrations by classification trees. *Environmetrics*, 15, 141-153.

CARSLAW, D. C. & TAYLOR, P. J. 2009. Analysis of air pollution data at a mixed source location using boosted regression trees. *Atmospheric Environment*, 43, 3563-3570.

DE'ATH, G. 2007. Boosted Trees for Ecological Modeling and Prediction. *Ecology*, 88, 243-251.

DEPARTMENT OF ENVIRONMENT ANNUAL REPORT. 2009. Department of Environment Malaysia

DEPARTMENT OF ENVIRONMENT ANNUAL REPORT. 2011. Department of Environment Malaysia

FRIEDMAN, J. H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29, 1189-1232.

FRIEDMAN, J. H. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38, 367-378.

GARDNER, M. W. & DORLING, S. R. 2000. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment*, 34, 21

GHAZALI, N.A., RAMLI, N.A., YAHAYA, A.S. 2009. A Study to investigate and Model the Transformation of Nitrogen Dioxide into Ozone Using Time Series Plot. *European Journal of scientific Research* 37(2), p. 192-205.

GHAZALI N. A., RAMLI N. A., YAHAYA A. S., YUSOF N. F., SANSUDDIN N., MADHOUN W. A. 2010. Transformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regression techniques. *Environmental Monitoring & Assessment 165(1)* p. 475-489. *Doi: 10.1007/s10661-009-0960-3*.

HASSANZADEH, S., HOSSEINIBALAM, F. & OMIDVARI, M., (2008). Statistical methods and regression analysis of stratospheric ozone and meteorological variables in Isfahan. *Physica A: Statistical Mechanics and its Applications*, 387(10), 2317-2327.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. 2001. *The Elements of Statistical Learning: Data mining, Inference, and Prediction.* , Springer-Verlag, New York.

HECK, W. W., Taylor, O. C. and Tingey, D. T. 1988. Assessment of Crop Loss from Air Pollutants. Elsevier Applied Science, London. 552 pp.

Hubbart, M.C., Cobourn, W.G. 1998. Development of a Regression Model to Forecast Ground-Level Ozone Concentration in Louisville, Ky. Atmospheric Environment. 32(14/15), P. 2637-2647

HUANG, L.S., & SMITH, R.L. 1999. Meteorologically-Dependent trends in urban ozone. Florida State University (FSU) Technical Report M-916, National Institute of Statistic Science, Research Triangle Park, NC 27709.

KRIEGLER, B. 2007 Cost-sensitivive Stochastic Gradient Boosting within a Quantitative Regression Framework. PhD Thesis. University of California, LA

LEATHWICK, J. R., ELITH, J., FRANCIS, M. P., HASTIE, T. & TAYLOR, P. 2006. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, 321, 267-281.

LAWRENCE, R., BUNN, A., POWELL, S. & ZAMBON, M. 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90, 331-336

MONKS, P. S., CARPENTER, L. J., PENKETT, S. A., AYERS, G. P., GILLETT, R. W., GALBALLY, I. E., AND MEYER, C. P. 1998. Fundamental ozone photochemistry in the remote marine boundary layer: The SOAPEX experiment, measurement and theory, Atmospheric Environment, 32, 3647–3664.

R DEVELOPMENT GROUP. 2008. R: A Language and environment for Statistical Computing. *In:* COMPUTING, R. F. F. S. (ed.). Vienna, Austria.

ROBESON AND STEYN. 1990. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations Atmospheric Environment. Part B. Urban Atmosphere

Volume 24, Issue 2, 1990, Pages 303–312

WANG, x., MANNING, W., FENG, Z. & ZHU, Y. 2007. Ground-level Ozone in China: Distribution and Effects on Crops Yields. Environmental Pollution 147, p. 394-400.

YAHAYA, N.Z., TATE, J.E., TIGHT, M.R. 2011a. Studying Particle Number Concentrations (PNC) in an Urban Street Canyon: Using Boosted Regression Trees BRT. Presented and in a proceeding book of the International Conference on Humanities, Social Sciences and Science Technology 2011 held on 27-28 June 2011 in Manchester University UK

YAHAYA, N.Z. TATE, J.E., TIGHT, M.R. 2011b. Analyzing Roadside Particle Number Concentrations using Boosted Regression Trees (BRT). Presented and in a abstract book of the European Aerosol Conference 2011 held on 4-9th September 2011 in Manchester University.

YAHAYA. 2013. Study of the Temporal and Spatial Variations of Ultra-Fine Particles in the Urban Environment. PhD Thesis. Institute for Transport Studies, University of Leeds, United Kingdom.