Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-173-RC2, 2019
© Author(s) 2019. This work is distributed under
the Creative Commons Attribution 4.0 License.

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Assessing the impact of Clean Air Action Plan on Air Quality Trends in Beijing Megacity using a machine learning technique" *by* Tuan V. Vu et al.

Anonymous Referee #2

Received and published: 11 April 2019

In this paper, the authors attempt to unravel the impacts of the Clean Air Action Plan on Air Quality trends in Beijing. Clearly this is an important area of study and the methods used do show some potential. However, before publication in any journal there are a number of issues that need addressing and, perhaps, re-thinking. It is currently not clear if these issues might impact on derived conclusions and enable reproducibility of the presented work.

There are a number of grammatical issues throughout the paper and I have only listed a few here. These should be looked at by the authors whilst re-framing their work.

Section 2: 2.1 Data Sources The authors note the use of met data from Beijing Airport.

How representative is this data of all sites studied? Im a little concerned this forms an important factor in determining the general applicability of the model. As the paper by Grange and Carslaw 2019 shows, the selection of wind directions, for example, can have significant impact on model fidelity if a site is affected by specific geography.

2.2 Modelling

Rather than referring to variables 'such as', please be specific in all cases.

You state that the 'regression model is an ensemble-model which consists of hundreds of individual decision tree models'. Please clearly state the number and how hyper-parameters were derived.

You state you used 'e.g. 70% of the all data [correct - of all the data]'. Is this an example or is this the actual training portion you used? I think this is clarified later on but please refrain from vague statements in describing any model development workflow.

It is customary to combine a single random sampling strategy with K-folds [e.g. 5] validation. Has this been used? If not, why?

If random sampling, how do you know if using different initial seeds in any random number generator leads to better or worse results? I cant see any code sharing so cant check this - please see a further comment on this.

The authors talk about an 'enhanced' normalisation procedure. Please explain more clearly how this is different from the original paper by Grange et al 2018. I will admit, that paper isnt as clear as it could be, but they do provide the model base. As far as I can tell, both studies only re-sample weather data. Also there is no discussion of classification into back trajectories, for example, or estimated boundary layer heights etc. If these products are not used, how is this study an enhancement? In some ways I struggle to see how section 2 'weather normalisation' is significantly different from the Grange et al approach. If they are different, they need clearly stating why - perhaps even with a visual workflow/table for each - and a comparison on findal data products

from both. The title of the paper leads me to believe this is a new technique.

line 104 - concentrations of an air pollutant and it[s] predictor variables - please correct

line 116: 'These time variables ' - do you mean parameters that vary with time or the time variable?

line 119 [equation with no label] - what is the significance of year 'i'? Is this defined on, say, the Unix epoch?

line 134: 'To validate the model for unseen data sets, a test data set which represents 30% of entire data sets[set] is input into the random forest model which has been constructed from training data sets.' This is a confusing statement. The test and training sets refer to both features and predicted variable. Thus, only features are 'input into the model'? Please re-phrase this. In fact, I would suggest you consider using the term 'features' when referring to variables to which you are fitting the model.

line 140: 'A weather normalization technique predicts the concentration of an air pollutant at a specific measured time point but with various meteorological conditions (termed as "weather normalised concentration").' Do you mean to state that this technique predicts the concentrations of an air pollutant as a function of meteorological factors alone?

line 142: 'Both time variable (month, hour) and meteorological parameters, except the trend variable were re-sampled randomly and was added into the random forest model as input variables to predict the concentration of a pollutant'. This is a confusing statement when referred to 'adding'. What do you mean by adding? On top of pre-existing variables?

Section 3.4 Please explain why, in a few cases, normalised values are higher than original.

Section 3.5 'Our results confirmed that the "Action Plan" has been highly effective'. Please define 'highly effective'.

Code/data availability: The current paper has no statement on this. The authors need to meet the current data and code sharing standards provided by Copernicus: https://www.atmospheric-chemistry-and-physics.net/about/data_policy.html https://peerj.com/articles/cs-86/ Indeed, there are currently many uncertain aspects of this study which could be resolved by clear code sharing and documentation.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-173, 2019.