Atmospheric Chemistry and Physics Discussions

Atmos. Chem. Phys. Discuss., referee comment RC2
https://doi.org/10.5194/acp-2021-751-RC2, 2022

Comment on acp-2021-751

Anonymous Referee #2

Referee comment on "Technical note: Investigating sub-city gradients of air quality: lessons learned with low-cost PM2.5 and AOD monitors and machine learning" by Michael Cheeseman et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2021-751-RC2, 2022

The authors propose a Random Forest model to predict sub-city-scale $PM_{2.5}$ concentrations. The studied case is wintertime in Denver, captured by CEAMS' low-cost sensor network on the one hand, and EPA's reference monitors on the other. A permutation metric is applied to conclude predictor importance, with a special interest in AOD.

While this is an interesting approach to quantify the influence of various drivers, I would like to point out some insufficiently discussed choices in applying the methods that might compromise the results.

Main concerns:

- From line 283 I conclude that the model was trained and tested on the same dataset that was used to tune the hyperparameters beforehand. Therefore, the test data can't strictly be considered unseen. The extent to which this limits the detection of overfitting and therefore validity of the results should at least be discussed. Potential overfitting is also implied by the authors' lack of confidence in the predictive skills of their model for new data (lines 464-466).
- Caveats in the analysis of predictor importance. A citation introducing and discussing the permutation metric seems to be missing. To my knowledge, the current gold standard to deduce predictor importance are Shapley-value based methods, due to their favorable theoretical properties. Therefore, it would be nice to justify the choice (presumably computational cost?). Especially the presence of a competitor like RH, that apparently got an unfair advantage by the correction factor (lines 428-430), seems to call for a metric where subsets of predictors are left out in the training. It is also questionable how well models trained on highly autocorrelated data are suited for the importance analysis, as stated in lines 314-316. Further justification is needed.

Minor concerns:

- Further investigation of the impact of interpolating the data could be insightful.
- To me, the main purpose of the paper is partly unclear. While transparency about the training and tuning process is important, the extensive explanation of Random Forests, cross validation and parameter tuning seems a bit convoluted for a paper whose foremost goal is to investigate the impact of different factors on the spatiotemporal variability of PM_5, and not necessarily to serve as a guide on applying RF models.

Technical notes:

- Line 160: consistency in use of special characters in "Angstrom."
- Line 262: missing hyphen in "over- or underfitting"
- Line 278: "depth of 15, 2 samples needed" – as far as I know, starting a clause with a symbol is considered bad style and also interrupts the reading flow here
- Table 2: The explanation for min_samples_leaf seems misleading, since leaf nodes aren't split. Do you mean the minimum samples stored in a leaf?
- Line 289: "This process was repeated until a distribution of each error statistic was created" makes it sounds as if there was an absolute threshold on how often to repeat a process before you can apply statistics. Maybe rather something like: "…repeated to create a distribution…"?
- It seems counterintuitive that the shuffled folds entail more autocorrelation than the consecutive ones. A very brief explanation or some numbers in the supplementary material could be helpful. On a positive note, I appreciate the topic is addressed at all.