Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2020-473-RC1, 2020

Atmospheric Measurement Techniques Discussions

Interactive comment on "Towards low-cost and high-performance air pollution measurements using machine learning calibration techniques" *by* Peer Nowack et al.

Carl Malings (Referee)

cmalings@alumni.cmu.edu

Received and published: 28 December 2020

General Comments

This paper presents an application of machine learning techniques to the calibration of data from low-cost sensors. It particularly focuses on (1) the effects of different subsets and combinations of inputs, including transformed inputs, on the resulting calibration performance, (2) the comparison of regularized Ridge regression to un-regularized regression, and of Gaussian Process regression to the more common Random Forest approach, and (3) the transferability of performance between locations, especially in cases where there is a greater range of concentrations at one location as compared

to another. Overall, the paper presents and explains the issues well, will a good description of the motivation and methods. The discussion of the machine learning approaches in particular shows a good understanding of these techniques. The results presented are mainly in line with previous work in this area and help to highlight and provide greater context for some of these issues, in particular the question of transferability of calibrations between locations. The paper is fairly well written, and I believe it is suitable for publication, provided some steps are taken to clarify certain aspects and statements (as outlined below).

Specific Comments

Lines 1-8: This background information can probably be condensed to 1 or 2 sentences within the abstract.

Lines 18-19: For the sentence "In particular, none of the methods is able to extrapolate to pollution levels well outside those encountered at training stage.", I believe it should say "...none of the non-parametric methods..." or "...none of the non-linear methods...", since you later state that the linear Ridge regression is able to extrapolate. Alternatively, if you mean that the methods are able to extrapolate but may not do so well, I suggest phrasing that as "...none of the methods is able to extrapolate well to pollution levels...".

Line 76: I would recommend removing the "1 – residual sum of squares/total sum of squares" part of this sentence, as this is more of a calculation formula than a definition of the term. Instead, I would suggest including this as a numbered equation in your paper, e.g., in the results section.

Figure 2: I believe this is the first time "AirPublic" is mentioned in the context of the sensor nodes. I suggest that this be explicitly stated as the maker of the sensor nodes in the body of the paper where the sensor nodes are described.

Line 301: Same comment as for line 76.

Line 316: “logarithmic plus exponential” can be ambiguous, i.e., did you use both as separate inputs, or add them together? I would instead phrase this as “both logarithmic and exponential”.

Lines 386-389: I would also suggest mentioning the importance of measuring potential interferents, like ozone and NO, since this seems to be indicated by your results as well and is a separate issue from the temperature and humidity effects.

Lines 449-451: It is not clear to me why dividing the data based on time in this way would guarantee the largest variability in pollutant concentrations.

Lines 510-512: You should specify whether this statement (in particulate the concentration range given) refers to NO2, PM10, or both.

Technical Corrections

Lines 18-19: I believe that “. . .none of the methods is able. . .” should be “. . .none of the methods are able. . .”.

Lines 56-57: This sentence is rather grammatically complicated; I would suggest revising it and/or splitting it up into several sentences.

Lines 79-80: You refer to the “r2” (lower-case R, non-superscript 2) metric here, is this the same as the coefficient of determination?

Line 102: “plantowers” should be capitalized. You may also want to indicate that this is the manufacturer.

Line 112: Missing period.

Line 123: Extra space before period.

Table 1: “varies” should be “vary”.

Lines 262-264: Data are plural.

Figure 4: Again, “r2” is used here.

Line 403: “r2” is being used again.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2020-473, 2020.