Review of Stamatis et al.: “Development and Application of a Supervised Pattern Recognition Algorithm for Identification of Fuel-Specific Emissions Profiles”

General Comment:

The researchers trained an unsupervised machine learning model utilizing PCA, k-means clustering, and LDA to identify fuel types by their monoterpenoid combustion emission profiles. They then apply this classification model to synthetic mixtures and to real field data. They find performance of the PR scheme declines as synthetic mixtures come closer in composition, but that in the case of field data, performance was good despite the presence of unknown fuel types.

Basically, there just needs to be a bit more separation and depth into the methodology given the complexity of it. All methodology should be explained up front and then results considered. More depth on methods would mean things like why PCA was the decided dimension reduction technique, and why use only 2 PCs as input to your cluster analysis (would this have led to overfitting of the data?).

Nonetheless, using machine learning to build intricate fuel type fingerprints is very interesting and I can recommend this work for publication following these minor revisions.

Specific Comments:

Line 151: Is the LOD mentioned here the same as the “detection limit” mentioned in line 115? What is this detection limit specifically? Is this in reference to the instrumentation LOD used across these different campaigns?

Line 161: What’s the rationale behind performing the manual selection at all? Is the idea that this demonstrates that your feature selection is more sophisticated than effectively just guessing?

Section 3.2.1, Line 177: Fig. 2 seems unclear to me in what it’s describing. My understanding is having fewer PCs describe more variance indicates better separation, but I think that could be made clearer (or perhaps I am altogether wrong). For instance, why mention only PCs 1-4 in line 178, why is PC5 not considered? Why are PCs 3-4 mentioned when only PCs 1-2 are used in the PR model?

Line 181: The automated criteria result in “more distinct and more consistent fuel profiles for each family,” but I don’t quite see how that’s exemplified in Figures 3 and 4.

Line 250: What is C_{st} in eqn. 4?

Technical Corrections:

Line 56: “…including to identify” should be “…including the ability to identify”

Line 92-93: Add comma, “…FIREX FL16, a broad…”

Line 119: Add comma, “In this work, an analysis of variance…”

Line 155: Add comma, “During the FIREX-16 FL16 study, …”

Line 170: Add comma, “In this application, …”

Line 192: Change “from the retained PCs (PC1 and PC2)” to “from these retained PCs”

Line 195: Add comma, “In this study, …”

Line 317: Remove comma after “algorithm”

Line 336: Add “of” to “27 out 28”

Figures 14-16 would be better conveyed were they vertically stacked.