This is an interesting article for the readers of the BSSA journal. The authors propose a new approach based on machine learning techniques to determine the importance of validation metrics for ground motion simulation. The field of ground motion simulation validation (GMSV) is relatively new and of interest to both seismologists who develop simulation models and engineers who would like to use these simulations in engineering applications. One of the most challenging questions to address is “which validation metric(s) is best to use?” as there are so many different metrics in this field and so many different opinions. It has been difficult to address this question because the answer varies depending on the application of the simulations (e.g., GMPE development, PSHA, structural analysis response of interest, etc). This paper is interesting because it provides a statistical approach to reduce the number of validation metrics and select a few representatives from a bigger set without compromising the final validation results. The authors have done a great job providing background on the topic. However, their application is very narrow, they only consider the 11 Anderson2004+ parameters (as they’ve noted there are other parameters) for one simulation method for one earthquake, and as a result their conclusions are only valid for this particular case. They have mentioned this limitation in the conclusions, but the abstract and introduction do not explain the limitations and the application-dependency of the proposed approach. Both Abstract and Introduction need modification to specify the application for which the results and conclusions are true (i.e., the application here is to narrow down the 11 Anderson2004+ parameters to select fewer parameters that give similar results as the original 11 for the simulations they chose in this paper). Currently, the Abstract and Introduction give the impression that the conclusions are true for any engineering application or simulation method. Furthermore, there should be a short discussion in the paper on how the users can apply the same method for different applications, for example for structural responses such as drift ratio or responses that are sensitive to high-frequency content of simulations (theoretically, this method should be applicable to any response of interest if they replace the GOF scores used in this study? If true this should be mentioned as it will be of interest to others).

More specific comments are provided below:

1. The terminologies used in the paper are not clear in many cases. For example, what are “features” “attributes” “dimensions” “trees, nodes and leafs” “class”? It was very difficult to read the section on Data Analysis Method and Introduction since these terminologies were not clearly defined or described in terms of the variables used in this paper. I was confused what “data” and “results” were referring to: GOF score data? GOF metric data? GOF category? Something other than GOF? I suggest that the authors streamline the paper and the terminologies to be consistent and clear in the final version.
2. The terms “common practice” or “common standards” have been mentioned several times. Because GMSV is such a new field (and I don’t think there is consensus, I for one don’t agree with some statements but I agree that others may), I think these terms should be removed. Just say this is your opinion, or it is based on the references that you have listed.
3. In abstract, “semi-supervised” and “supervised” learning techniques are mentioned, but in the paper these terms are not explained, what part of the learning techniques is semi-supervised and what does that mean?
4. In introduction, the difference between “verification” and “validation” should be discussed.
5. In abstract and introduction, what does a “direct application” mean? it’s not clear. Similarly, in Validation Metrics section: “direct quantitative comparisons” and “indirect statistical analysis”
6. Page 4, line 53: add “relatively” before “well understood”
7. The two questions posed in the introduction lines 59-61: I don’t think the first one is really the question, all parameters should be considered if anyone has thought of them as important (i.e., has proposed them). The second question is the main question and the one being addressed in this paper.
8. You are treating these metrics as “independent variables”, but as you have also noted many of them are not. For example, Arias intensity and energy are correlate. Does this compromise your conclusions? If not explain why in the Results section.
9. Page 7, line 133: delete the last sentence
10. Page 14, line 310: you say “… the latter can essentially be discarded…” I don’t understand this statement, wouldn’t this mean that both C1 and C7 are important to consider since they are independent of each other (one gives information that the other one doesn’t)?
11. Page 14 line 323: It’s not clear how you “replicate data”, provide more detail.
12. I don’t think “confusion matrix” given in tables is explained in the paper!
13. Page 17, lines 391-394: this sentence should be re-worded, what is an “accurate conclusion”? You should be very specific that your conclusion is that instead of using all 11 validation metrics you can select a few in order to get the same categorization as all 11 metrics would have given you.
14. Page 18, line 405: this is implying that fig 10 is similar to fig2! I personally cannot see this in the figures! They look very different. It might be better to regenerate fig 11a but with different color scales that only vary by four choices, then the comparison would be much easier.
15. Figure 8: I don’t know what you mean by y-label “decision attributes”. Do you mean “the number of GOF metrics”? why there is no data point for 1 or 11?
16. The conclusion basically says that response spectrum is the most important validation metric. This is not surprising to any engineer or modeler. The big question is what other parameters describe the differences between a simulation and a recorded motion if their response spectra matched? Or in other words, what are the other important validation parameters aside from response spectrum (many have argued that’s duration, but no statistical proof). Can you use your method to answer this question? Can you condition your machine learning process on Sa being the same (for example use a subset of data with similar Sa or artificially make them equal), then decide what’s the next important parameter? Would the result be the same as in your conclusions? Or would it be influenced by removal of Sa as a parameter (this is possible since there are correlations between Sa and other metrics).

Typos:

1. Page 2, line 8: change focused to focus
2. Page 2, line 20: change classifiers to classifier
3. Page 13, line 185: change measure to measured
4. Page 16, line 354: the last C5 should be C6?