Annales Geophysicae Discussions

Ann. Geophys. Discuss., referee comment RC1
https://doi.org/10.5194/angeo-2021-33-RC1, 2021


# Comment on angeo-2021-33

Anonymous Referee #1

Referee comment on "Unsupervised classification of simulated magnetospheric regions" by Maria Elena Innocenti et al., Ann. Geophys. Discuss., https://doi.org/10.5194/angeo-2021-33-RC1, 2021

Anonymous review of Unsupervised classification of simulated magnetospheric regions by Innocenti et al for Annales Geophysicae

In the manuscript under review, the authors have applied machine learning, more specifically self-organizing maps (SOMs), to the question of identifying the magnetospheric regions in which space plasma measurements were made. They propose that their method could be applied to deciding which parts of large measurement databases be downlinked at high resolution, a very real issue for space missions. In this study, they apply the SOM method and K-means clustering to simulation data generated by the OpenGGCM-CTIM-RCM code, a global MHD simulation of the Earth's magnetosphere. They show the method is capable of successfully classifying several magnetospheric plasma regions, and perform comparisons of input data preparation on classification results.

The manuscript is well written, describes new science results, and is of high interest to the community. The application does not contain brilliant breakthroughs as such, but is a valuable addition and a useful parameter study. There are only a few clarifications and improvements to the discussion to suggest before recommending publication, which I have divided below into major and minor suggestions.

Major points:

L94-95: Please explicitly state that the boundary conditions vary with time. Are all boundary cells surrounding the simulation domain always at the same value, varying identically with time?

L157: Please clarify the data point selection. Is it randomized? Is there any selection based on Y or Z?

Figures and analysis: Although the polar plane is very interesting in many ways and should indeed play a role in the study, the equatorial plane could be considered even more relevant, particularly considering the orbits of recent significant space missions. I strongly urge that you present also equatorial plane plots early on to show the strength of the method.

Table 2: Instead of multiplying values in the latter two feature sets, could all sets perhaps be re-normalized to a norm of 1? This would make comparison easier.

Kneedle determination of optimal cluster count: did you attempt changing the number of clusters to see how robust the method is? Would setting k=6 merge clusters 1 and 4? How would this change the misclassified cluster 5 points at the bow shock for F1?
What about setting k=5, would clusters 1,4 and 5 merge? I believe a written description of such a test without figures would suffice.

Could you please add some quantitative estimate of the agreement between different feature sets? This would be particularly useful if you also included comparison against the pure K-clustering approach, to show how much improvement SOMs bring to the table. For this purpose, I think manually merging some clusters (e.g. 1,4,5) would be acceptable.

L372: Since you talk of lessons learned, I was surprised that you did not describe attempts using the logarithm of magnitude of B, instead going with the quasi-arbitrary clipping procedure. Please try out log(B) if not already attempted, and at least briefly report the results.

Conclusions: I would like to see some more discussion about relevance and limitations. An important point to discuss, albeit outside the scope of the current paper, is the identification of small structures inside larger domains (e.g. the tail current sheet within the boundary layer) and the points of transition between domains. The latter might arise naturally from this method, the former not so much. Some mention of this should be included. Also, there has been no discussion of the drastic difference between MHD simulation descriptions of plasma parameters and the values measured by spacecraft or hybrid simulation methods, namely kinetic effects, noise, and/or instrumentation limitations. Some of these are touched upon in Amaya(2020), but they are quite relevant to the discussion here as well, if this method is indeed to become a first step in any actual classification approach.

Minor points:

Abstract: I would recommend you mention comparisons against the K-clustering event here, as well as the use of PCA to reduce input data.

L14: Introduction to machine learning: Perhaps some more general, canonical paper could be cited?

L30: how has the data magnitude changed when compared with Cluster, the Van Allen probes, ISEE-1, etc?

L90: Please clarify in better detail the simulation set-up. Is this the domain size of the whole simulation, or the portion of it shown in figure 1? In section 4 you state you use only the relevant portion of the simulation for post-processing SOM analysis. What is the spatial resolution of the simulation?

L119: The indicated magnetic field clipping value does not make much sense. Is the intention to keep the same ratio between the three components and the original signs, but re-scale the magnetic field vector to a magnitude of 100 nT? Please rephrase.

L129: is the lattice really of type R^2? The visualizations show a hexagonal grid.

L131: I recommend you briefly mention that from available plasma variable you select n features for the SOM method to use, so that the R^n notation is meaningful.

L135: Could this be better described (to the layman) as altering the feature values of the code word so that the distance ws for the winning element becomes smaller? Similarly, consolidation of terms might make it easier to read, e.g. data entry vs input data point vs input point - these are probably all the same thing, i.e. a list of features associated with a point of measurement.

L143: is the numerator of the exponent in formula (3) the integer lattice neighbor distance, up to a value of sigma(tau)?

L201: Is the K-means clustering performed based on the final code words of each node?

L282: I would suggest briefly mentioning the sunward inner magnetosphere misclassification already here.

L286: This should probably be Bx, not Bz?

L299: Since the cluster numbering is arbitrary, you could perhaps re-order the colors to match the earlier ordering to assist the reader.

Figure 7: What time value is this? (for the caption, not just the text)

L337: Please clarify if these are trained with t0+210 or with mixed time data?

Figures 8 and 9, main text: The ordering, going from the top row of Fig8 to two rows of Fig9, back to the bottom row of Fig8, and then to the last row of Fig9 is counter-intuitive. This could surely be improved.

L367-368: Please briefly mention what F7 and F8 added

L416: More comparison with Amaya(2020) of the results and potential future avenues would be good - it was only very briefly cited before.

L447: Do you have any references to indicate what direction non-linear feature correlation analysis in deciding a dimensionality reduction could take? What about the addition of non-local features, such as spatial and temporal derivatives, curls etc?

L448: Are dynamic SOMs a reasonable approach to automatic classification, or do they require user validation after every re-learning?