Authors Response to Review

Our responses to the review are written in **blue**.

Response to Editor

We have received the reports from our advisors on your manuscript, "Interactive Graphics for Visually Diagnosing Forest Classifiers in R", Based on the advice received, the Editor feels that your manuscript could be accepted for publication should you be prepared to incorporate minor revisions. When preparing your revised manuscript, you are asked to carefully consider the reviewer comments which are attached, and submit a list of responses to the comments. Your list of responses should be uploaded as a file in addition to your revised manuscript.

Thank you. We appreciate the chance to do a further revision, and hope that these changes are satisfactory.

Response to Reviewer 2

We appreciate all of your feedback which help to improve our paper.

Summary of the paper

The paper introduces an app for visual exploration of tree-based ensemble models in three steps. First, the app enables the user to investigate how well the model performs when it comes to predicting each class in its observation-level functionality. Second, the tree-level section allows the user to select a tree based on a variable importance measure in one of the first three nodes and visualize the way the first three splits are made in that tree. Finally, the app visually compares the predictive performance of the model to a benchmark one, which is illustrated by comparing a projection-pursuit-based random forest to a standard one.

The paper and its accompanying app seem to be a good fit for Computational Statistics, as they apply interesting visualization techniques to a tree-based ensemble model that will surely prove useful for practitioners.

However, for this potential to materialize the authors should put more emphasis on the limitations of their approach. In particular, the app focuses on the first three nodes of the trees in question, which would be much less informative in settings requiring deep trees. **The app was modified to adress this issue. In the new application the user can select specific nodes to see (from one to all the nodes). If there are a lot of nodes will be better to select 4 or less each time to see better the complete display but the application is not limited to see just the first three nodes.**

Furthermore, plots such as the parallel coordinate plot would lose their appeal in settings with thousands of variables. A brief discussion of circumstances limiting the usability of the app would address these concerns and allow users to easily determine whether the app is suitable for their application.

The discussion has been extended to include limitations.

Substantive comments

1. p. 8, line 42: The issue of visualizing the tetrahedron is arguably not of first-order importance as it is not as helpful for classification problems with more than four classes. It might be more useful to focus

exclusively on the side-by-side jittered dotplot as this approach generalizes in a more straightforward way to problems with more classes.

As we mentioned in page 7 the ternary plot can be generalized to more than three dimensional space, with G classes, the ternary plot idea is generalized to a G-1 dimensional simplex. Then if we have 4 classes the generalized ternary plot is a tetrahedron as shown in Figure 5 but if we have more than four classes we can also generalize the ternary plot idea. In the fishcatch data used in this paper to illustrate the shiny app there are 7 classes and we generalized the ternary plot to a 6-dimensional simplex as is shown in the last panel of Figure 8 (6-didmensional simplex shown pairwise), using tourr package you can see better the 6-dimensional simplex in different rotations https://vimeo.com/707047875. The structure is complex because a 6-dimensional simplex has 7 vertices, 21 edges and 35 faces.

2. p. 11, line 16: As section 3 discusses all other plots produced by the app it would be useful if it also included an overview of the parallel coordinate plot that only appears in section 4, so the meaning of that plot is not immediately clear (e.g. what is displayed on the vertical axis).

   We have included more explanaition about the parallel coordinate plot in Section 4 where this figure appears and added a corresponding reference for further details. In the vertical axis the values for each variable are standardized to easily compare them.

3. p. 13, line 38: I find the choice of the interaction driver for the model tab puzzling. I would argue that the key statistic distinguishing the trees in the ensemble is OOB error so a more natural design choice would be to include jittered points in the OOB error boxplot so that the user can explore the well or poorly performing trees in more detail. I do not have a good intuition for why the user would be interested in exploring trees based on variable importance.

   The second tab was modified based on your helpful comments. We have included three main changes in the second tabset, first we have included jittered points in the OOB error boxplot where the user can select trees based in their performance. Second, we have included a panel to select the nodes to see in the importance variable plot, the mosaic plot and the density plot. The panel allows you to select more or less than three nodes and the mentioned plots are updated acordingly. Finally, we rearranged some of the plots to make the selection and interaction more intuitively presented in Figure 10.

4. p. 15, line 35: The OOB error plot is somewhat hard to read due to some classes being harder to classify than others, and a potential solution to that would be using a different scale on the vertical axis (e.g. percent of highest OOB error for a class).

   We explore different alternatives as you suggested but the result was not clearly better. We have removed the points and just keep thiner lines which improve the visualization. Also in the app, because the interactivity, it is easy to see the oob error values mousing over the plot which in the static version it is not possible.

Minor comments

1. p. 6, line 17: introducing the G_s notation is unnecessary as it is not used in the remainder of the paper (also, the meaning of the subscript s is not explained).

   I have removed G_s, as you notices it is unnecessary here

2. p.6, line 21: it should be "Alternatively, it can be computed" instead of "Alternatively, can be computed".

   Done

3. p. 6, line 33: it should be "The plot in the right" instead of "The plot at right".

   Done

4. p. 8, line 20: it should be "into a three-dimensional space" instead of "into three dimensional".

   Done

5. p.8, line 49: crabs are "classified", not "predicted".

   Done

6. p. 10, line 40: the phrase "one of the biggest difficulties for the app in order manage linking" should be rewritten to, for example, "one of the biggest difficulties when it comes to the app managing linking".

   Done

7. p. 12, line 46: using the phrase "some trees could be removed from the forest" instead of "some trees could be pruned from the forest" would make this sentence clearer.

   Done

8. p. 14, line 49: "a receiver" instead of "an receiver".

   Done

9. p. 16, line 32: it would be useful to include a legend of the class colors.

   Class color legends were not included in the interactive web app because the interactivity allows you to get the class name when you mouse over each plot.

10. p. 17, line 6: " will perform" instead of "will performs".

    Done

Response to Reviewer 4

I think the paper overall is good. The comments below are a combination of comments on the paper itself and some suggestions for the plotting in the software (which don't have to be changed for the paper). I think also that some of the details of the plots need to be explained better.

MAIN COMMENTS:

Section 4 in general: since the big point here is that the software is interactive, I think the discussion in this section would be a lot more useful if you illustrated step by step with screenshots an example where your interactive drill-down allows you to quickly discover things. For instance, say there are some outliers in plot A. Then you click on them and see some further details in plot B that help you understand what is going on. In this section you show a bunch of screenshots and label their capabilities, but you don't actually demonstrate an end-to-end example of the interactivity. For instance, in Figure 9 you give an example of a set of actions, why don't you actually show us the results and the changes that occur in a sequence of plots? Furthermore, I would show some actual examples of screenshots of what shows when you hover over observations.

In Section 4 we have added more details and examples of use and interpreting results to address your comment.

CONTENT COMMENTS:

1. Page 2: In describing Random Forests, in addition to randomly sampling variables to try to split on in each node, don't they also initially sample a subset of variables for each tree? If so, you should note what settings you use in this paper (how many variables are selected for each tree). Also, if the variables are randomly sampled at each node, rather than simply choosing the best projection possible for, say, any pair of variables, what effect does this have on the importance measures? What effect is there to determining the number of variables to use in each projection (e.g. 3 instead of 2)? Is this something you have experimented with? I think the paper could also benefit from a simple preliminary

example of, say, two features and two classes and showing the plot of the linear combination resulting from projection pursuit. You could even include an image from the original paper.

A random forest has two sources of randomness: variable sampling and bootstrap samples. The variable sampling is used for node partition, where each split is produced from a random subset of variables. There is not an initial sample of variables for each tree.

We did not formally explore the sensitivity of the random variable sample size for the measure of variable importance. In the algorithm you can select the proportion of variables to be sampled at each node. In the crab and fishcatch examples presented in the paper two variables were sampled at each node.

We have included an image from the original paper with the boundary that would result from a classification tree fitted using the PPtree algorithm to take into account your last request. The plot shows a simple example from simulated data with two features and three classes. The partitions generated by the PPtree algorithm are oblique to the axis, incorporating the association between the two variables.

2. Page 4 line 48: should O_k also have a subscript for the variable j? Also $\hat{y}_{ikj}$

   Subscript j for O_k is not needed because for every variable j O_k will be the same and not change depending on j, the same for $\hat{y}_{ik}$. In the only case subscript j can be included is in $\tilde{y}_{ik}$ to indicate the permuted variable but for simplicity we just keep the formula as it is in the original PPforest paper (https://www.tandfonline.com/doi/abs/10.1080/10618600.2020.1870480)

3. Page 5 Figure 1: You mention that the variables should be standardized for the coefficients to be interpretable as importances and be comparable. But in the upper right corner, the plots of the variables distributions, for instance for node 2, are not in the standardized range (node 2 has an x-axis that goes up to 5 and 6). **As you mentioned the data should be standardized for the coefficients to be interpretable as importances. We have fixed the density plot and the x axis accordingly.**

   Is the distribution being plotted here the value of the linear combination for each class, rather than a single variable? **Yes, the density plots shows the data projection at each node. This is now made clear in the caption.**

   Is there something going on here with the default x limits for doing KDE? **The x limits were fixed.**

   As a side note, you show that the first cut perfectly separates the blue and orange species. You may want to consider having the bounds for the KDE be the observed range of the linear combination for each class so that the density plots won't overlap in the plot for node 1, which they seem to do more than they should if the separation is perfect.

   We explore your suggestion (changing the bound for the KDE to observe range) but the result was not better so we have kept the original.

   Furthermore, if you have any control over the color scheme (if you don't I would recommend adding this as a variable), it might be a good idea in this paper to have the colors for the blue and orange species be blue and orange, and have differing shading patterns to distinguish male from female, or something similar; perhaps you can choose more similar colors within both species. Or, better yet, can you use different plotting symbols, say triangles vs circles to distinguish the species, and shaded vs hollow to distinguish the sex?

   The selected palette (Dark2) is a qualitative palette suited to mapping a categorical data because it is colorblind friendly. It is a little unfortunate that the categories in the crabs data are colors, which makes the explanation confusing. We have changed the category names to be simply O and B so that there is less confusion.

   Also, can you include labeling in your plot that the sub-plots are node ID (e.g., put the label "node" above the 3 density plots). **We have included in each sub-plots "Node" before the number and also update this in the app.**

What do the numbers in the grey boxes at the terminal nodes represent? **The numbers in the gray boxes are the predicted class, we fixed the plot to show the class names**

4. Figures 1 and 2: I see that the potential range for coefficient values for the projection pursuit seem to be between -1 and 1. I imagine this is a feature of the algorithm. Perhaps include in your description of projection pursuit a note about this fact, for those unfamiliar, and mention what optimization criteria is used in it.

   Yes, the coefficient values are between -1 and 1 because the length of the projection vector is 1. We find the linear combinations, a which maximize $\frac{a^T B a}{a^T W a}$, where B and W are the between and within group sums of squares. We get $\hat{a}_1 = e_1$ where $e_1 ... e_s$ are eigenvectors of $W^{-1}B$. In this case we find the 1-D projection optimizing the LDA index. We have included in Section 3 LDA and PDA indexes which are optimized to get the linear projections to make it more clear how the coefficients are computed.

5. Page 6: If you use a notation for VI for global variable importance (you define it as the average across k trees), I'd just define it here as an equation and use it in the text (e.g in line 19).

   Formulas for all VI were defined in equations and mentioned in the text. It is clearer now.

6. Page 6 line 48: does MDS need a citation? Also, can you connect the results in this discussion to the RF graphics, e.g. in terms of same variables with high RF importance being important in MDS? Otherwise, this whole discussion seems unnecessary. If you are doing it simply to understand the data, I would move this section to right after you introduce the data, before you show the other graphics. Here, it seems like a tangent. Also, are these MDS and tetrahedral plots part of your software or do you perform them separately? I would say so explicitly in either case.

   A reference for MDS has been included. In Section 3 we describe the mapping of diagnostics to visualizations. MDS is presented in Subsection 3.3 (Similarity of cases) the key connection with the forest is that we are using the proximity matrix information which is one of the diagnostics described for PPforest or RF and we use this information to compute the MDS. This visualization can be useful to identify in which cases the model is performing well and which it is having problems.

   MDS and generalized ternary plot (tetrahedral plot in case there are four clases) are both implemented in the app and we have added a comment in the paper to clarify this.

7. Page 7 Figure 2: why does the left plot need to be a dot plot rather than, say, a single box or violin plot for each variable? I don't understand what role the x axis for each variable plays here. If it doesn't play any role (e.g. is just for the jittering), then don't use it.

   We selected a jittered side-by-side dot plot in this case instead of a box plot because we want to see the values for each tree, and it makes it more sensible when linking observations between multiple plots. Jittering spreads the data on the categorical variable axis to better read the distribution.

8. Page 9, Figure 4: is the dimension of the simplex always at most 3 or G-1? What would it look like with more classes? I'd clarify whether this only works for your example or whether it can always be drawn.

   he dimension of the simplex is $G - 1 = 3$ (Figure 5 revised version), since in crab data there are 4 classes. Ternary plot can be generalized to more than three dimensional – with $G > 4$ classes, the ternary plot idea is generalized to a $G - 1$ dimensional simplex. In the fishcatch data there are 7 classes and the ternary plot to a 6-dimensional simplex as it is shown in pairwise plots in Figure 6. Using the tourr package you can see the 6-dimensional simplex in a rotation through 6D, see https://vimeo.com/707047875. The structure is complex because a 6-dimensional simplex has 7 vertices, 21 eddges and 35 faces. This has been clarified in the paper.

9. Page 10, Figure 5: similar remark as before. I think a violin plot or something similar for each class, rather than a jittered dot plot, is more correct. The x-axis for each class is a "fake" axis. I think jittered dot plots should only be used with scatterplots to make sure individual dots can be distinguished.

   We are using points because it is important to see the individual cases and the jitter is to overcome overplotting on the categorical variable. It is not recommended using jittered scatterplots because it would change the real values of your data.

10. Page 10, line 40: This sentence is muddled "As Sievert (2020) describes, one of the biggest difficulties for the app in order manage linking between plots is the data structure management for each widget". Furthermore, what does "widget" refer to here? Use a more descriptive term.

    At their core Shiny widgets are mini-applications created using the shinyApp function. We have changed the language of the paragraph to help with this terminology.

11. Page 10, line 45: If you are just going to demonstrate the interactivity, wouldn't it be better to just continue with the original crab dataset, or have one dataset throughout? Unless there is something particularly interesting about this dataset.

    Yes, fishcatch data are more difficult to classify and then more interesting to show some of the results. For simplicity to describe the main plot of the app we use crab data but to describe the app we prefer to present a more complex problem, more classes, and umbalanced data.

12. Page 12 Figure 6: in the enhanced PCP, you say all variable pairs are shown at least once as neighbors. Is there a way, either through ordering or adding dividing lines, etc. to help a reader find the pair they are interested in? Or is there some sorting done so that, the pairs are ordered from left to right in terms of some metric of "interesting-ness"? There seems to be no order in the given plot. Do you automatically pick the number of dimensions to show in MDS? Should the tetrahedral plots here have some kind of axis labels? See Figure 7 where they do.

    A PCP can be ordered in different ways, and we have modified the PCP to include variable order based on the mean value. As you suggested, we have also change the grid lines to keep only the vertical grid in gray. In the MDS we only used two dimensions but can be more if deemed more appropriate for a particular problem. We have included labels in the generalized ternary plot.

13. Page 13 line 46: I wasn't aware that there was an option in RF to prune under-performing trees (as opposed to nodes within a tree) in an ensemble. Please elaborate or include a reference. Also, in this case, maybe a different word than "prune", such as "discard".

    This comment is in page 12, In the original RF there is not an option to prune under-performing trees. The comment is just to mention a possible use for the information we get from the app, also we are presenting the app using PPforest and this can be a possible extension for our algorithm.

14. Page 16 Figure 10: in the upper LH corner, where you select the classes, do these appear as text that you type or delete, or as buttons you can toggle on and off? Seems like the latter would be better. Also, the bottom two plots have different x axes, so it is hard to compare them (if they should be compared, the axes should be the same; perhaps you could even plot them on the same plot?). Furthermore, there are no titles for either, and they have the same axis labels so I have no idea what the difference is between them.

    In the upper LH corner we are using the function 'selectizeInput()' from shiny R package which creates a select list that can be used to choose a single or multiple items from a list of values, you do not have to type the class names you just select the names from a list. For the bottom two plots (variable importance plots) we have fixed the axis and include titles in each plot, so it is easier to compare them.

STYLISTIC COMMENTS

1. Should out of bag be written OOB rather than oob?

   We have changed to using OOB now.

2. Page 3 line 29: when you say "proportionately predicted", do you mean something specific, like that the class probabilities are proportionate to their frequency in the data?

   Cases that are predicted to be multiple classes, without a clear majority in any of them, indicate that those cases are difficult to classify. The paragraph was not clear and we modify it to:" Cases that are predicted to be multiple classes, with similar class probabilities, indicate difficult to classify observations".

3. Page 3 line 38: These two sentences don't seem to work together: "Each model uses samples of variables. So that, the accuracy of the models can be compared when the variable is included or omitted." Do you mean "Thus" rather than "So that"?

   Yes, should be Thus. It is fixed.

4. Page 4: "Specie" in Figure 1 caption is misspelled.

   Done

5. Page 11 line 12: For the sentence beginning "The data feeding. . . ", separate each clause with semicolons rather than commas.

   Done

6. Page 11 line 39: say "ID" or "index" rather than "id" ; same for, say, page 13 line 44

   Done

7. Page 11 line 49: I would modify the final sentence to say that "all variable pairs are shown next to each other once" rather than "all variables are neighbors".

   Done