Review of assessing the reproducibility of discriminant function analyses

Paper description. In this paper, Andrew and coauthors attempt to assess how variation in data curation practices affect our ability to reproduce reanalysis of published data. I think this paper (and related efforts along these lines) is extremely important to show the value proposition for researchers for why they should to take a few extra steps to document and share their data at the time of publication after which time it becomes increasingly harder if not impossible to do so (which as Michener points out is when the authors have the most knowledge about their datasets).

Major points

On line 47, I would suggest using 'underlying data' rather than accompanying data. The former sounds more substantial, as in the data on which the conclusions of the study are based, rather than data that support or accompany the paper. I'll leave it up to the authors to decide how to take this.

For the protocol described in lines 60-67, it would be worth mentioning a few more reasons why data might be hard to associate with an analysis or actually be usable. Data curation is a catch all that can mean many things along a spectrum of usability. This would be a great opportunity to clarify what you mean here. Descriptive columns names, a separate metadata file, even better if validated against some schema, stored in a persistent repository with identifiers etc.

Although the paper is really not about suggesting solutions, but more of a case study, it would greatly improve the paper if you did point people to the right resources. A couple of years ago Ethan White et al wrote a paper *Nine simple ways to make it easier to (re)use your data* (http://library.queensu.ca/ojs/index.php/IEE/article/view/4608). It's a fairly short paper but quite effective as an introduction to someone looking to immediately improve their data curation practices.

On 83, the authors mention another purpose of the paper as investigating *the role of changing curation standards affecting reproducibility.* In curious if this is really true. Have such practices themselves changed significant over this period? From the perspective of trivial computational reproducibility, minimal metadata (not even data validated to a specific standard) are not technology agnostic to some sense.

I also found it rather encouraging to see no relationship between publication data and data issues, however this is likely just an artifact of the criteria used to select the papers. The criteria (an analysis which has remained relatively unchanged over the past few decades, as with data collection methodology, and data size) do not really allow for exploring reproducibility challenges beyond the more trivial problems (data discrepancy, insufficient metadata, incorrect data). I would suggest highlighting two specific issues in the discussion: a) If reproducibility is so hard with such a limited scope, complex analyses with more moving parts, larger datasets, and complex software dependencies would be hopeless. Briefly mentioning such additional challenges in the discussion would be helpful. b) The discussion does not get into much detail about why metadata matters beyond column labels. Mentioning something about standards (e.g. EML if speaking to this audience) that allow for structured metadata would be really helpful in this context. There are many reproducibility challenges that will simply go beyond issues with column labels.

The methodology and criteria used to assess the accuracy of reproducibility at various levels, and the use of DFA as a measure all seemed quite appropriate to me.

Minor points

This is a bit meta, but I would love for the datasets you used to be shared somewhere as a repository. Perhaps on one of the persistent ones, or at the least a GitHub repo with a Zenodo DOI? Or something similar to Vines 2014 (*The analysis code and data are available on Dryad under DOI number 10.5061/dryad.q3g37.*)

Again in the interest of reproducibility, it would be helpful to include the output of either `sessionInfo()` or from the `devtools` package `session_info()` at the end of your R script.

I find the code a bit hard to read, especially with spacing. It would be worth putting it through `tidy_source` from the `formatR` package. I couldn't run the code myself without any input data (I don't have access to the raw data in the Google Drive. Perhaps the data are available via), but there are obvious syntax errors:

e.g.

```
newdat0=read.csv(file.path(year,datadir,data"infile"),na.strings=c("NA","na","",".","#N/A","#DIV/0!","#
```

Super minor points

In the section spanning lines 104-133 I would suggest bolding the various categories you assign the papers to ('Incorrect data file', 'Data discrepancy', 'Insufficient metadata').

Line 234: There needs to be a space after estimation.

I enjoyed the contribution.

Sincerely, Karthik (further discussion is welcome over email, karthik.ram@gmail.com)