# Ostats package prospectus

*Quentin D. Read*

*April 08, 2020*

This document explains the current state of affairs with the `Ostats` package: why scientifically it is cool, what it currently does (with a quick example), what features it would be good to add, and some ideas for further reading.

## Scientific background (a.k.a. Why do we care?)

As part of an NSF EAGER grant to study within-species trait variation at different scales across the National Ecological Observatory Network (NEON), John Grady and I, assisted by co-authors, developed a way to look at trait overlap among co-occurring species, called the overlap statistic (O-stat). This is somewhat complementary to the T-statistics presented in a paper by Violle et al. (2012), which are implemented in R in the `cati` package. During our work, we also found that a basically identical statistic was proposed by Mouillot et al. (2005). Similar ideas have been developed recently by Blonder et al. (2018), who authored the R package `hypervolume`.

We originally used the statistic to look at body size distributions in rodents (Read et al. 2018). Recently, I was contacted by an Australian researcher who is using the code I posted as a supplement to our manuscript to look at body size distributions of reptiles in Australia. Also recently, my labmates from grad school asked me to help them analyze some data on the daily activity patterns of ants under different experimental warming treatments, potentially using this framework.

Because there is at least a little bit of interest in this approach, I thought it would be a good idea to write an R package so that people could calculate these statistics for themselves. This has the potential to really make our research have a "broader impact" at least in the research community, because it will lower the barrier for people to reproduce our research and apply the approach to their own systems.

Following is a description of what I have done so far, and ideas for what I think should be done next.

## Where are we now?

Right now, there is a very skeletal version of the package posted on GitHub at NEON-biodiversity/Ostats. It is basically just the code supplement from our 2018 manuscript in R package form. I made this for my own internal use for analyzing the ant data — once you have the package installed, you can just call `library(Ostats)` and all the functions are loaded for your use instead of having to source the individual script(s). R packages are actually not that hard to make, especially with the built-in package creation workflow in RStudio. It is really just a matter of pasting in the functions and running a few things to automatically generate some formatting around them.

There are only a few functions in the package right now, just enough to calculate one-dimensional trait overlap between pairs of species, get an average pairwise overlap for a community, and evaluate it against a null model. The main function is called `Ostats()` and the function that does the work of integrating the distributions and finding overlap is `pairwise_overlap()`. There are a few other helper functions.

You can install the development version of the package from GitHub by calling
`devtools::install_github('NEON-biodiversity/Ostats')`

## A quick example

Here I use the existing package to get the overlap statistic for a tiny subset of the NEON rodent data, taking two sites with very different overlaps. At one site (Harvard), most species have very similar body size distributions so we see high overlap, and at the other site (Jornada) the species have very different distributions so we see low overlap. The data are pulled directly from figshare where they are archived.

Load and clean the data, showing a sample of what it looks like:

```
library(tidyverse)
library(Ostats)

# Load data from web archive
dat <- read_csv('https://ndownloader.figshare.com/files/9167548')

# Keep only relevant part of data
dat <- dat %>%
  filter(siteID %in% c('HARV','JORN')) %>%
  select(siteID, taxonID, weight) %>%
  filter(!is.na(weight)) %>%
  mutate(log_weight = log10(weight))

# What does the data look like
dat %>%
  group_by(siteID, taxonID) %>%
  slice(1)
```

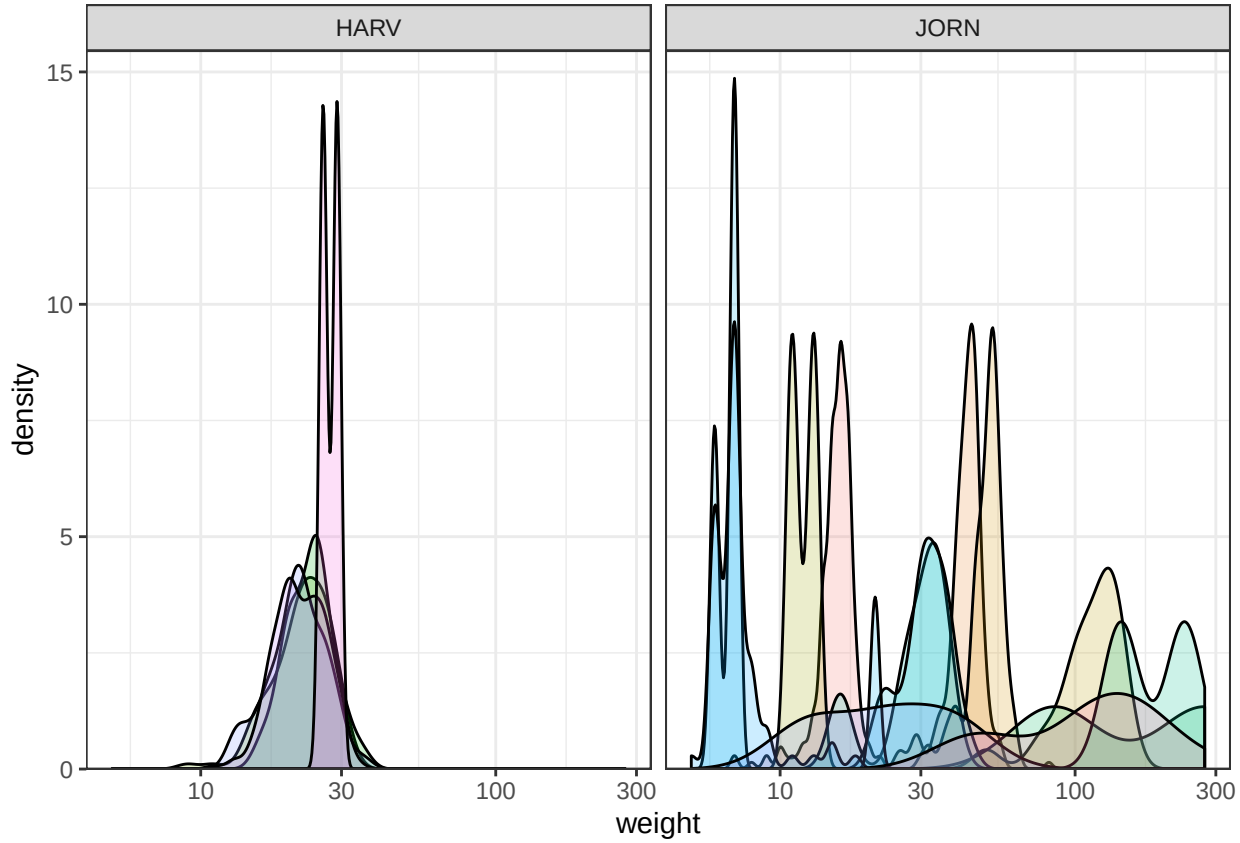
```
## # A tibble: 20 x 4
## # Groups:   siteID, taxonID [20]
##    siteID taxonID weight log_weight
##    <chr>  <chr>    <dbl>      <dbl>
##  1 HARV   MYGA      19        1.28
##  2 HARV   NAIN      17        1.23
##  3 HARV   PELE      36        1.56
##  4 HARV   PEMA      23.5      1.37
##  5 HARV   PEME      23        1.36
##  6 HARV   PESP      29        1.46
##  7 JORN   CHPE      16        1.20
##  8 JORN   DIME      47        1.67
##  9 JORN   DIOR      43        1.63
## 10 JORN   DISP     122.       2.09
## 11 JORN   MUMU      13        1.11
## 12 JORN   NEAL      85        1.93
## 13 JORN   NEMI     143        2.16
## 14 JORN   ONAR      33        1.52
## 15 JORN   ONLE      27.5      1.44
## 16 JORN   PEFA      21        1.32
## 17 JORN   PEFL       6        0.778
## 18 JORN   PELE      39        1.59
## 19 JORN   PGSP      14        1.15
## 20 JORN   SIHI     170        2.23
```

Run O-stats on the data.

```
# nperm = 2 means do not bother with null models
Ostats_example <- Ostats(traits = as.matrix(dat$log_weight),
```

```
                    sp = factor(dat$taxonID),
                    plots = factor(dat$siteID),
                    nperm = 2)
```

This results in O-stat of 0.8946416 for Harvard (high overlap), and 0.018304 for Jornada (low overlap). Plotting the distributions shows that this makes sense.




## What could we do moving forward?

There are more features we could add to the package, and also some semi-tedious but necessary things that need to be done to bring the package up to a good quality standard. This is a non-exhaustive list.

### Cool new features to add

- I am currently working on analyzing time-of-day data with this framework (the times of day that ants foraged under different temperature conditions). As it turns out, the raw overlap statistic does not really work here because times are basically angles on a circle, not points on a number line that go off in both directions. So it would be cool to add support to the package for doing the overlap statistic on circular data, and possibly plots as well.
- I think it would really improve the package to include some built-in plotting functions (probably done in `ggplot2`). I am envisioning having some diagnostic plots that you could call up by running `plot(Ostats_object)` where `Ostats_object` is the output of the function `Ostats()`. Then there would also be different built-in plotting functions to visualize the distributions.

- Right now, the only null model included in the package is a very "naive" null model that just jumbles species and traits up completely at random. A lot of ecologists think this is not that great of a null model, and that there should be some kind of stratified or constrained sampling. It would be nice to include options for different null models.
- Another major shortcoming of our current statistic is that it only works in one dimension (one trait at a time). It would be nice, but maybe too much work, to extend it so that it works in multiple dimensions. There is already a lot of work out there on trait hypervolume overlap, so this might be redoing some work that's already been done.

### Necessary chores

- Improve the code to be in line with best programming practices:
  - The functions are currently very restrictive and opaque in terms of what format of input they accept. Some of this is a relic of my original idea of trying to make the `Ostats` function have the same structure as the `Tstats` function in the `cati` package. We should write code that will convert different possible types of input in a standardized format
  - There are a lot of almost redundant functions that could be condensed into one function with options.
  - The functions should include error-catching code that will return informative error messages and/or warnings.
- Write documentation for the functions and for the package as a whole:
  - Help documentation for all the functions
  - Package description
  - Vignette(s) or demo(s) that would include example data to demonstrate how the package works.

### End goal

The R package is a great product in its own right. But if we get it to a good enough place, to get even more "credit" for putting in the work, we can write it up into a paper and submit it to a journal. The easiest journal to target would be the Journal of Open Source Software, which is pretty easy to get R packages published in, and is both open access and free to publish in. That is more of a way to document the software, and give people something to cite and not really read by ecologists. If we end up doing a lot more "actual ecology" in this process, it *might* be worth it to write up a full manuscript for an ecological methods journal such as Methods in Ecology and Evolution. That would be a lot of additional work so I am a little wary of that, since I doubt I would have a ton of time to commit to that.

## References

### Works cited

Blonder, B. (2018). Hypervolume concepts in niche- and trait-based ecology. Ecography, 41, 1441–1455.

Mouillot, D., Stubbs, W., Faure, M., Dumay, O., Tomasini, J.A., Wilson, J.B., et al. (2005). Niche overlap estimates based on quantitative functional traits: a new family of non-parametric indices. Oecologia, 145, 345–353.

Read, Q.D., Grady, J.M., Zarnetske, P.L., Record, S., Baiser, B., Belmaker, J., et al. (2018). Among-species overlap in rodent body size distributions predicts species richness along a temperature gradient. Ecography, 41, 1718-1727.

Violle, C., Enquist, B.J., McGill, B.J., Jiang, L., Albert, C.H., Hulshof, C., et al. (2012). The return of the variance: intraspecific variability in community ecology. Trends in Ecology & Evolution, 27, 244–252.

## Further reading

On the science, the works cited are just a start. It's definitely worth looking into other papers that they cite. These ideas go back to MacArthur, Hutchinson, and other old-school ecologists.

On R package development, I should say that I am not very experienced in R package development myself. So far, I have worked on maintaining and developing new features for an R package that is on CRAN called rslurm. I relied on Hadley Wickham's (of ggplot, tidyverse, and RStudio fame) book on writing packages in R to get up to speed with things. The book is available online free here.