sftrack, part II: Developer task force for a broad adoption

Mathieu Basille

2020-04-01

Warning We have renamed the package from `sftraj` to `sftrack` to properly reflect the purpose of the package, and the underlying semantical definitions.

Signatories

Project team

The **Project team** lists the core members of the work, who initiated the `sftrack` package and will be instrumental in progress and completion of the project (other R packages to which people contributed are indicated in parentheses):

- **Mathieu Basille**, Assistant Professor at the University of Florida, USA (`adehabitatHS`, `hab`, `rpostgis`, `rpostgisLT`)
- **Matt Boone**, Data Scientist at the University of Florida, USA (`refsplitr`)
- **Rocío Joo**, Postdoctoral Associate at the University of Florida, USA
- **Clément Calenge**, Statistical Analyst at the *Office français de la biodiversité*, France (`adehabitatMA`, `adehabitatHR`, `adehabitatHS`, `adehabitatLT`)
- **Emiel van Loon**, Assistant Professor at the University of Amsterdam, the Netherlands (`zoon`, `RNCEP`).

Consulted

The following R package developers have shown interest in working together to use `sftrack` for tracking data (relevant packages they have authored in parentheses):

- **Guillaume Bastille-Rousseau** (`lsmnsd`: Classify movement strategies using a latent-state model and NSD; `moveNT`: An R package for the analysis of movement data using network theory; `wildxing`: An R package for optimal positioning of widlife crossing structures)
- **David Cooley** (`gpx`: Converts GPX files to simple features)
- **Ross Dwyer** (`VTrack`: A collection of tools for the analysis of remote acoustic telemetry data)
- **Devin Johnson** and **Josh London** (`crawl`: Fit continuous-time correlated random walk models to animal movement data)
- **Ioannis Kosmidis** (`trackeR`: Infrastructure for running, cycling and swimming data from GPS-enabled tracking devices)
- **Jed Long** (`wildlifeDI`: Calculate indices of dynamic interaction for wildlife tracking data)
- **Alec Robitaille** (`spatsoc`: Group animal relocation data by spatial and temporal relationship)
- **Johannes Signer** (`amt`: Animal Movement Tools)

From the R Consortium:

- **Hadley Wickham**

The Problem

Movement defined broadly plays a central role in fields as diverse as transportation, sports, ecology, music, medicine and data science (Gudmundsson *et al.* 2012). Miniaturized tracking devices have become nearly ubiquitous, and resulted in an ever-increasing volume of *tracking data* (Joo *et al.* 2019). The Movement community in R has broadly embraced this new field, and created an entire ecosystem of tracking packages, with 58 packages that process, visualize and analyze tracking data (Joo *et al.* 2019, now available as a Tracking CRAN Task View (https://cran.r-project.org/web/views/Tracking.html) that will be updated twice a year).

However, there is a critical lack of standard infrastructure to deal with movement. As a matter of fact, half of the tracking packages work in isolation, not being linked to any other tracking package (Fig.1). After identifying this gap in dealing with tracking data in R, we started a long-term effort, to assess and structure the efforts of the community with respect to movement data. In particular, with support from the **R Consortium** (September 2019–March 2020), we developed the `sftrack` package (https://github.com/mablab/sftrack) to provide central classes and basic functions to build, handle, summarize and plot movement data.

The `sftrack` package, completely developed in the open, is now full-featured, and is ready for broad testing. This stepping stone in the development of `sftrack` already contributes solid foundations for the tracking ecosystem in R. There is a need now to deepen the connections with other packages, and work together with package developers for broad adoption of `sftrack`. We thus propose a two-step extension of the `sftrack` project, aiming at:

1. Working with developers of tracking packages to ensure linkage to `sftrack`, and enhance interoperability within the tracking ecosystem in R.
2. Establishing `sftrack` formally in the R ecosystem, with a submission to CRAN (https://cran.r-project.org/) and to rOpenSci (https://ropensci.org/).

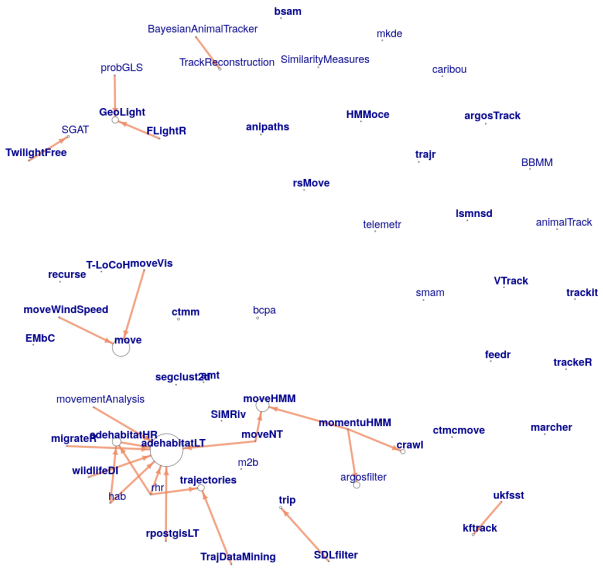

Figure 1: Network representation of the dependency between 58 tracking packages in R. Solid arrows go towards the package the others depend on. From Joo et al. 2019.

The proposal

Overview

During the first stage of the work (September 2019 until March 2020), we developed all features of the `sftrack`, which now provides foundations to build, handle, summarize, and plot tracking data. As the package is now ready for use, we will work with R package developers to help them integrate and adjust `sftrack` solutions into their data flow (June–August). Secondly, we will use this collaborative development to establish the version 1.0.0, which defines the public API. This major stepping stone will be submitted to to CRAN and to rOpenSci.

Detail

In the first stage of the work, we defined a precise data model for tracking data, which recognizes the duality of locations, in the form of geographic (x, y, z) and temporal (t) coordinates, and steps, i.e. the straight-line segment connecting two successive locations. Throughout the process, we have focused on providing solid documentation of the package and its specifications, targeting both users and package developers, which is essential for a broad adoption of the `sftrack` package. The development of the package happened in the open on GitHub (https://github.com/mablab/sftrack), where we have received feedback from the community in the form of use cases (https://github.com/mablab/sftrack/issues?q=label%3Ause-case).

In the next stage of this project, we will work with developers to finalize the implementation, to ensure that `sftrack` meets all needs of the movement community. We have ongoing discussions with 9 developers of 10 packages, ranging from import of GPX files to animal movement ecology, through human tracking devices for running, cycling and swimming. Specific elements that will require further development or fine-tuning include the grouping structure (coined `burst`, where the individual identification of the moving object is required, and can be completed by any number of arbitrary factors) and the error measure (typically the error in GPS positions), both of which have no established standard.

We will work with developers of R tracking package (including, but not limited to the one mentioned above) to integrate `sftrack` in their data flow. How we will conduct the work will depend on the role and complexity of tracking data in each package:

- Input data only (e.g. `crawl` or `spatsoc`): work with the developers on their package initial function(s) to make them `sftrack`-ready (preferably in addition to the original input format).
- Simple data classes (e.g. `amt::track_xy/amt::track_xyt`): work with the developers to directly replace their classes with the one offered by `sftrack`.
- Complex data classes (e.g. `adehabitatLT::ltraj` and `trackeR::trackeRdata`): For these packages, the work will be conducted in two phases: 1) develop bi-directional converters from and to `sftrack` classes (which should be lossless both ways); 2) ultimately replace these data classes with the new ones offered by `sftrack`.

The second stage of work will entirely be focused on establishing `sftrack` in the R tracking ecosystem. We will first prepare the package for submission to CRAN, then focus on the preparation and submission of `sftrack` to rOpenSci. The first step will ensure that the package stays up-to-date with R and other package dependencies, while the second step will guarantee that `sftrack` adheres to a best practice standards for R and reproducible science. Packages submitted to rOpenSci go through detailed peer-review by other developers—the rOpenSci certification (through their badge) will assure `sftrack` is easy to use, well documented, and well coded.

Project plan

Start-up phase

The `sftraj` package is meant to be a cornerstone for the development of a more cohesive Movement community in R. The first stage of the work, which just started, will deliver solid technical foundations for trajectories in R. Our project has already been received very positively on Twitter, which we use to channel our communication. Two blog posts (https://mablab.org/post/sftraj-intro/ and https://mablab.org/post/sftraj-plan/) explained what we are doing, and why we need feedback from the community. Interestingly, in a very short period of time, we already received the perspective on `sftraj` from developers of 4 tracking packages (`amt`, `crawl`, `spatsoc`, and `wildlifeDI`), which comes in addition to our early discussions with 5 developers from three other tracking packages (`VTrack`, `moveHMM` and `momentuHMM`). Altogether, feedback received only confirms the need for a central trajectory package in R.

Technical delivery

Progress will focus on the delivery of four products of increasing complexity. They will be addressed one by one, as they all build on top of each other:

- **Vignette of the package** [month 1]: Writing a vignette will provide a reproducible tutorial which will guide through the use of `sftraj`. A well-written vignette is now essential for users to get started and see the benefits of using a package.
- **Submission to CRAN** [month 1]: During the first stage of the work, `sftraj` will become full-featured and installable from GitHub. In this second stage, we will ensure that the package is ready for CRAN and finally submit it there. Publication on CRAN will allow for easy installation on any computer, as well as enabling the dependency system of R packages, which are both mandatory for broad adoption.
- **Submission to rOpenSci** [month 2]: After the package has been submitted to CRAN, we will go one step further and actually prepare it for rOpenSci. rOpenSci emphasizes (and basically enforces) a package's quality, fit, documentation, and clarity. `sftraj` would fit in the special category "geospatial data". Becoming the first official rOpenSci tracking package (none of the 58 existing tracking packages are currently on rOpenSci) would give the package an extra exposure, and guarantee its overall usefulness and usability.
- **Submission of a companion manuscript to the R Journal** [month 3–4]: Finally, as the quality of `sftraj` increases with the previous steps, we will work on a manuscript aimed at developers to provide the rationale behind development decisions for the package, and detail its implementation. In practice, this manuscript will provide package developers the necessary information for other packages to be able to rely on classes from `sftraj`.

Other aspects

While the work on `sftraj` will happen completely in the open, we will keep communicating on our progress, directly on Twitter as well as with more detailed blog posts on https://mablab.org/. We will watch for feedback, notably through the GitHub repository, which will remain constantly open.

In parallel, we will use scientific conferences (which we already planned to attend, using other funding) to introduce `sftraj` and present our work, such as useR! 2020 (which will be held in St. Louis, Missouri in July 2020), the International Statistical Ecology Conference (ISEC, which will be held in Sydney Australia, in June 2020), or Moving2Gather (which will be held in Rennes, France in March 2020).

Requirements

People

This is the same core project team than for the first stage of the work, which will lead the work for all deliverables. This stage will likely be less collaborative than the first stage, as it focuses on writing documentation and packaging the code, although all contributors will be welcome.

In particular, we request $10,000 in salary for 1.79 months (over the course of the project) of Matt Boone, Data Scientist at the University of Florida. Matt has a solid experience in R which complements that of other team members, and will be the main developer and contributor to the codebase. Having Matt fully dedicated guarantees successful completion of the project.

Processes

We will follow the same principles of openness than before, relying on a community-based code of conduct that aims to be inclusive, and a work that will happen entirely publicly, using the GitHub repository.

Tools & Tech

No technical constraint is foreseen. All team members are already equipped with enough computer power to work on the project. The development platform (GitHub) is already set up and public, and will remain open to the entire Movement community.

Funding

We request a total of $10,000 to support 1.79 months of Matt Boone, Data Scientist (Biological Scientist II) in the MabLab at the University of Florida.

- Salary: $7,369
- Fringe rate (35.70 %): $2,631
- **Total award = $10,000**

Summary

Salary to support a Data Scientist is requested to have one person committed to the project, who will dedicate set chunks of time to the work. This seems required to ensure project completion in the proposed timeline. Almost 8 weeks of funded work over the course of the project is a reasonable amount of time, which matches proposed deliverables.

Success

Definition of done

This project will be successful if `sftraj` becomes the standard for tracking data in R, both from a user and a developer perspective. Outsourcing code for low level trajectory classes from other packages to `sftraj` will be the most important outcome. In this stage of the work, success is directly related to the deliverables, namely the vignette, submission of the package to CRAN, submission of the package to rOpenSci, and submission of a companion manuscript to the R Journal by the end of the 4-month period.

Measuring success

Beyond delivering the four products in the planned time frame, success will be measured essentially from the adoption of the package:

- From a user perspective: number of downloads (e.g. using RStudio download statistics);
- From a developer perspective: adoption in tracking packages that will depend on `sftraj`.

Future work

Three axes for further development will be targeted after this stage of work:

- Provide support to developers of R tracking packages to help them develop conversion tools from their own custom classes to classes from `sftraj`.
- Dynamic visualization of trajectories, allowing keyboard- and mouse-controlled exploration of trajectories, step by step (based on the solution provided in `rpostgisLT`).
- Developing tools to clean and interpolate trajectories, based on specific filters and assumptions (e.g. maximum speed allowed, or adding missing locations by interpolation, etc.).

Key risks

The main risk is actually linked to the completion of a functional package at the end of the first stage of the work. Despite long delay to actually start the work due to administrative difficulties, we do not foresee further delay in delivering the foundations of the `sftraj` package: our team is experienced both in developing R packages and working on tracking data, and the R Movement community is responding very positively to our invitations.

Specifically for this stage of the work, core project team members have all the expertise required for the deliverables:

- **Vignette:** Clément Calenge has produced very elaborate vignettes for each package of the `adehabitat` series. The vignette for `adehabitatLT` has notably been recognized very positively in our recent survey (Joo *et al.*, 2019), with 88.6% of respondents expressing that the documentation was either good (allowing the user to do everything they wanted and needed to do with the package) or excellent (allowing users to do even more than what they initially planned because of the excellent quality of the information).
- **Submission to CRAN:** Mathieu Basille (`rpostgis`, `rpostgisLT`), Clément Calenge (`adehabitatMA`, `adehabitatHR`, `adehabitatHS`, `adehabitatLT`) and Emiel van Loon (`zoon`, `RNCEP`) all have experience preparing and submitting R packages to CRAN.
- **Submission to rOpenSci:** Matt Boone has recently been through the process for the package `refsplitr` (https://github.com/embruna/refsplitr), for which he was the lead coder. `refsplitr` is now in the latest stages of review with rOpenSci, and will be added to their list of packages very soon. As a matter of fact, Matt even wrote a detailed blog post about the whole review process at rOpenSci (https://mablab.org/post/ropensci/).
- **Companion manuscript for the R Journal:** Mathieu Basille has worked on a similar approach for the package `rpostgis`, which is detailed in Bucklin & Basille (2018), a manuscript that was prepared for the R Journal right after the package reached a stable state and was published on CRAN.

References

- Bucklin, D., & Basille, M. (2018). rpostgis: linking R with a PostGIS spatial database. The R Journal, 10(1), 251–268. https://doi.org/c7fc

- Gudmundsson, J., Laube, P., & Wolle, T. (2011). Computational Movement Analysis. *In* Kresse W, & Danko D. M. (Eds.), Springer Handbook of Geographic Information (pp. 423–438), Springer-Verlag Berlin Heidelberg. https://dx.doi.org/10.1007/978-3-540-72680-7_22
- Joo, R., Boone, M. E., Clay, T. A., Patrick, S. C., Clusella-Trullas, S., & Basille, M. (2019). Navigating through the R packages for movement. Journal of Animal Ecology (early view). https://doi.org/10.1111/1365-2656.13116
  Pre-print available at: https://arxiv.org/abs/1901.05935