sftrack v1.0: Stable API for a broad adoption

Mathieu Basille

2020-04-02

Note: 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 the progress and completion of the project (other R packages to which people contributed are indicated in parentheses):

- **Matthew E. 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`)
- **Mathieu Basille**, Assistant Professor at the University of Florida, USA (`adehabitatHS`, `hab`, `rpostgis`, `rpostgisLT`).

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)
- **Théo Michelot** (`moveHMM`: Animal Movement Modelling using Hidden Markov Models)
- **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

Miniaturized tracking devices have become nearly ubiquitous, and resulted in an ever-increasing volume of *tracking data* in fields as diverse as transportation, sports, ecology, music, medicine and data science (Gudmundsson *et al.* 2012). The movement community in R has broadly embraced this new field, and created an entire ecosystem of 58 tracking packages that process, visualize and analyze tracking data (Joo *et al.* 2020). This intense activity has led to the need of 1) a modular solution to handle tracking data in R; 2) links with other packages to streamline the use of tracking data in R. We developed the `sftrack` package (https://github.com/mablab/sftrack), with support from the **R Consortium** (September 2019–March 2020), to answer the first point. We now request additional funding to move a step further and create a more integrated tracking ecosystem across fields in R.

There is a critical lack of standard infrastructure to deal with movement in R. As a matter of fact, half of the tracking packages work in isolation, not being linked to any other tracking package (Fig.1). We developed the package `sftrack` to fill up this gap: `sftrack` provides central classes and basic functions to build, handle, summarize and plot movement data. `sftrack` is part of a global effort, to assess and structure the efforts of the tracking community in R. For instance, we created a Tracking CRAN Task View describing all tracking packages (updated twice a year, available at: https://cran.r-project.org/view=Tracking).

After the first stage of development, completely in the open, the `sftrack` package is now full-featured, and is ready for broad testing. This milestone 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 an extension of the `sftrack` project aiming at reaching the stable API of the package (i.e. version 1.0) by:

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, through submissions to CRAN (https://cran.r-project.org/) and to rOpenSci (https://ropensci.org/).

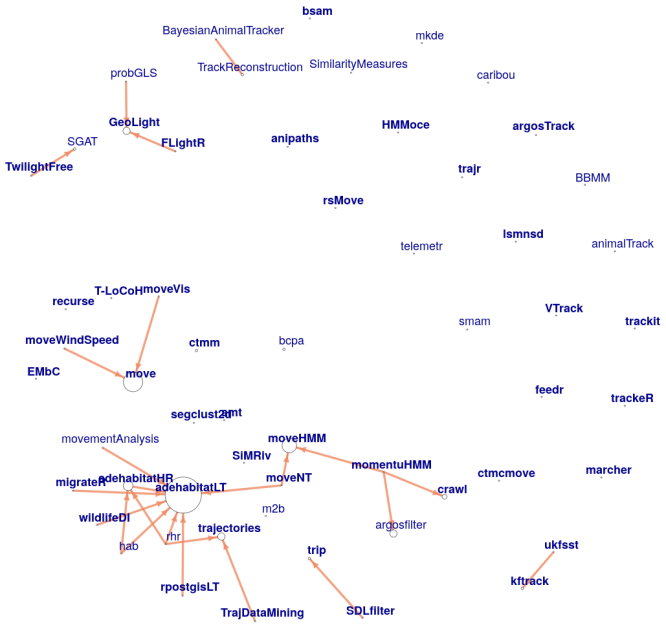

Figure 1: Network representation of the dependency between 58 tracking packages in R. Solid arrows go towards the package the others depend on. From Joo et al. 2020.

The proposal

Overview

During the first stage of the work (September 2019 until March 2020), we developed all features of the `sftrack` package, which now provides foundations to build, handle, summarize, and plot tracking data. As the package is now ready for use, we will submit it to CRAN, and we will work with other tracking package developers for the integration of `sftrack` into their data flow. Secondly, we will use this collaborative development to establish the version 1.0, which defines the stable public API. This major milestone will then be submitted to rOpenSci.

Detail

In the first stage of the work, we defined a precise data model that recognizes the duality of tracking data, in terms 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 thorough 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 10 developers of 11 packages, in addition to 3 developers of 4 packages from the core team. Together, these 15 tracking packages (`adehabitatLT`, `adehabitatHR`, `amt`, `crawl`, `gpx`, `hab`, `lsmnsd`, `moveHMM`, `moveNT`, `rpostgisLT`, `spatsoc`, `trackeR`, `VTrack`, `wildlifeDI`, `wildxing`) represent more than a quarter of the tracking ecosystem in R, ranging from import of GPX files to animal movement ecology, through human tracking devices for running, cycling and swimming, and connection to a spatially-enabled PostgreSQL database. 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 measurement (typically the error in GPS positions), both of which have no established standard.

We will work with developers of R tracking packages (including, but not limited to the ones 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:

- Ad-hoc input data not defined in a class (e.g. `crawl` or `spatsoc`): We will work with the developers on their package initial function(s) to make them `sftrack`-ready (possibly in addition to the original input format);
- Simple data classes (e.g. `amt::track_xy`/`amt::track_xyt`): We will 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) we will develop bi-directional converters from and to `sftrack` classes (which should be lossless both ways); 2) we will work with the developers to ultimately replace the original data classes with the new ones offered by `sftrack`.

We will use this process to finalize the public API and establish `sftrack` in the R tracking ecosystem. We will first prepare the package for early submission to CRAN, then later focus on the preparation and submission of `sftrack` to rOpenSci ((version 1.0)). 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 (DONE)

The `sftrack` 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 is just finished, has provided solid technical foundations for tracking data in R. Our project has already been received very positively on Twitter, which we use to channel our communication. We explained our approach and decisions in a series of 4 blog posts (https://mablab.org/tags/sftraj/). We used them to call for feedback from the community, and received the perspective on `sftrack` from 10 developers of 11 packages (outside of the core team), who expressed interest in working together. Altogether, the feedback we received only confirmed the need for a central tracking package in R. We will keep working in the open, and collect feedback from the whole movement community.

Technical delivery

We will deliver the following products during the 6-month period of the grant:

- **Submission to CRAN** [month 1]: `sftrack` is currently installable from the GitHub repository, and passes all CRAN checks. However, 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. In this first stage, we will thus submit the package to CRAN.
- **Data class converters** [month 1–2]: We will provide converters to classes from the main tracking packages, such as `adehabitatLT::ltraj` and `trackeR::trackeRdata`, integrated into `as_sftrack` through method dispatch. We will develop bidirectional conversion, and work towards lossless conversion both ways. Depending on the specific requirements of other classes, this will require adjusting the classes already developed in `sftrack`.
- **Integration of `sftrack`** [month 2–4]: In parallel, we will work with developers willing to fully integrate the solution offered by `sftrack` into their package data flow. This step may require substantial rewriting of other packages — while we will not take the lead on this task, we will provide specific guidance and feedback, and potentially adjust `sftrack` to match the needs of each package involved at this stage.
- **Submission to rOpenSci** [month 5–6]: In the last two months, we will go one step further and actually prepare `sftrack` for rOpenSci, which will coincide with the stable public API of version 1.0. rOpenSci emphasizes (and basically enforces) a package's quality, fit, documentation, and clarity. `sftrack` 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 extra visibility, and guarantee its overall usefulness and usability.

Other aspects

The work on `sftrack` will happen completely in the open. We will keep communicating 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 collect use cases, bug issues and suggestions for enhancements. We have already successfully presented our work on `sftrack` at the 2020 Moving2Gather conference (held in Rennes, France, in March 2020), and will consider other venues once the situation with COVID-19 allows us to travel normally. For instance, we have submitted an abstract for the useR! 2020 conference in July 2020 (St. Louis, MO, USA).

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 be highly collaborative, as we will work with other tracking package developers. We have already established working connections with several of them, notably through our efforts in the movement community over the last two years.

In particular, we request $10,000 in salary for Matt Boone (1.74 months over the course of the project), Data Scientist at the University of Florida. Matt has a solid experience in R which complements that of other team members, and is the lead coder of `sftrack`. 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 1.74 months of Matt Boone, Data Scientist (Biological Scientist II) in the MabLab at the University of Florida.

- Salary: $7,189
- Fringe rate (39.10 %): $2,811
- **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 `sftrack` becomes the adopted standard for tracking data in R, both from a user and a developer perspective. Successful work with other tracking package developers to integrate `sftrack` into their data flow will be the most important outcome. If our efforts are successful, `sftrack` will become the central package in the ecosystem of R tracking packages (i.e. provide a central node in Fig.1).

Measuring success

Success will be measured essentially from the adoption of the `sftrack` package:

- From a developer perspective, the number of packages that will progressively rely on `sftrack` will be a direct measure of success: We will measure the number of packages linked to `sftrack` (`Depends` and `Imports`);
- From a user perspective, the number of downloads will tell us how popular the package becomes: We will track CRAN package download statistics, from the RStudio CRAN mirror.

Future work

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

- Continuing support to developers of R tracking packages to help them integrate `sftrack` classes or develop conversion tools from their own custom classes to `sftrack` classes;
- Dynamic visualization of trajectories, allowing step by step exploration of trajectories with keyboard and mouse control (based on the solution provided in `rpostgisLT`, see https://cran.r-project.org/web/packages/rpostgisLT/vignettes/explorePgtraj.html).

Key risks

Now that the `sftrack` package is feature complete, there is no foreseen technical difficulty. The main risk is actually linked to successfully reaching out to other package developers. The R movement has so far responded very positively to our invitations, and several tracking package developers from a range of applications have already expressed their interest in working together on the integration of `sftrack`.

In addition, core project team members have all the expertise required for the deliverables:

- **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 part of rOpenSci (https://docs.ropensci.org/refsplitr/). As a matter of fact, Matt even wrote a detailed blog post about the whole review process at rOpenSci (https://mablab.org/post/ropensci/).

References

- 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. (2020). Navigating through the R packages for movement. Journal of Animal Ecology, 89:248–267. http://dx.doi.org/dcnf Pre-print available at: https://arxiv.org/abs/1901.05935