Response to Reviewers - Article 4780

cubble: An R Package for Organizing and Wrangling Multivariate Spatio-temporal Data

H. Sherry Zhang, Dianne Cook, Ursula Laa, Nicolas Langrené, Patricia Menéndez

2024-01-02

We provide an overall summary of the changes made to the paper and the package and then we make a point-by-point explanation of how we addressed the reviewers' comments.

Response to B-review_2_2

Sec. 1

- Typo: "Section 5 discuss**es** the paper" **Fixed.**

Sec. 2.2

- Consider to also show the print or str of the input data sets station and meteo to make clear how the connection is established between both tables. **The input data are now printed.**
- Typo on page 5: "there are **a** number of methods" **Fixed.**

Sec. 2.3

- Fig. 1: The time dimension runs backwards in the data cubes, that might be irritating. **Fixed.**
- Typo in caption: "To focus on **t**he temporal" **Fixed.**

Sec. 2.5

- Typo: "it is a matter of choice ~~on~~ which structure to use given the application" **Fixed.**

Sec. 3.1:

- Typo: "that can be used to match~~ed~~ the selected time series across locations. **Fixed.**
- Typo: "The function ~~on~~ calculates" **Fixed.**
- Typo: "it is possible to ~~can~~ include" **Fixed.**
- Is a matching based on index (date and time) only also possible in cubble?

Sec. 4:

- The first paragraph is hard to read. It could possibly be restructure as a numbered list, or description listing, or simply adding roman numbers in the text, e.g.: "Five examples are chosen to illustrate different aspects of the cubble package: (i) creating a cubble object from two Coronavirus (COVID) data tables with the challenge of having different location names, (ii) using spatial transformations . . . " **Done.**
- It is nice to mention the challenges of all examples, as interested users of cubble could directly jump to the relevant example for their own use case.

Sec. 4.1

- I suggest to be even more explicit about warnings to make clear, that they are expected and not due to malfunctioning code: "Discrepancies are flagged (see warnings in the R output below) when creating the cubble object." **Done.**

Sec. 4.3

At the very beginning of the example, please clearly motivate the two-step approach. **Added. The two-step approach is motivated because ideal matches require both spatial proximity and temporal similarity in trend.** I thought I had understood it, but am again puzzled. What I understand is that you pick (i) the 10 closest pairs of met stations and river stations and then (ii) additionally require a temporal similarity (in terms of peaks). But how is the temporal offset treated and controlled in these cases? **see dot point 3 below** I also see some logic in an approach where you (i) select 10 spatially closest met stations per river station (i.e. setting `spatial_n_each` = 10) and (ii) pick the highest data similarity between met and river stations within each group. Here, it might be possible, that close-by met and river stations are less related than apparently further apart stations due to local terrain and catchment boundaries. The motivation of the example is crucial and might raise awareness for unintended matches in individual user data. Please clearly describe what this example is meant to illustrate to the reader and potential user of cubble. **The proposed strategy to pick pairs has been considered, which motivates the arguemnt spaital_n_each.**

- Note that R> print(`res_sp`, n = 20) only shows ten rows as `res_sp` only has ten at that point in the script. I'd suggest to drop that argument n = 20 here. **Done.**
- After the use of `bind_rows()` the matched stations are stored in rows 2i-1 and 2i, i = 1, . . . , 8. I believe that this notion of how to write pairs in the cubble should be mentioned and explained.
- How is temporal offset handled, due to delay in the river network from upstream precipitation. Can that be controlled during the temporal matching? Can that be illustrated in the figure? **This can be controlled by the arguemnt temporal_window, default to 5. The figure shows 1 year worth of daily data and an offset of 5 days would be indifferentiable in the plot.**
- Fig. 4 caption:
  - Typo: "These four station **pairs** show**n** on the map (a) and **as** time **series plots** (b)" **Fixed.**
  - Typo: "reflect ~~percipitation~~ precipitation" **Fixed.**

Sec. 4.5:

- Fig. 6: "The bottom row first selects the lowest temperature in August in the seasonal display **from the first row**" **Fixed.**

Sec. 5:

- "It provides a much better interface to the vast array of statistical and machine learning model architecture." Much better than ..? **Changed to: It provides the infrastructural interface to build ...**