Referee contribution to public discussion - Geochronology

invitation received: 2022-11-18 | today: 2022-11-22

1 Contribution summary

In a nutshell, the manuscript presents a numerical feature implementation in R and a code example in Python to map similarities/dissimilarities between distributional age data. The 'new' metric is the Wasserstein-2 distance, which is somewhat tested against the Kolmogorov-Smirnov distance metric.

2 Major comments

2.1 Presentation quality

Sincerely, I enjoyed reading the manuscript. It is a concise and neatly written technical paper that presents reasoning and implementation of a numerical metric to map age/grains-size distributions in R and Python. The graphical quality of the figures and tables is good. Still, the axes labelling could be improved for readers unfamiliar with multidimensional scaling.

Minor point: The title seems to promise more than the manuscript delivers. "Comparing . . . ". The manuscript does not compare something. The presented "comparison" is a performance test of the Wasserstein-2 distance and the Kolmogorov-Smirnov distance. I think the title should reflect better what the manuscript tries to achieve: a presentation of an alternative metric in the realm of multidimensional scaling.

2.2 Scientific significance

The general idea of the manuscript fits within the scope of GChron. However, I am in a little bit of doubt about whether it justifies requesting a peer-review procedure and a peer-review publication. The numerical metric presented here is not new, and the manuscript does not (yet?) show significant scientific progress. The implementation in R appears limited to a few code lines. Perhaps under different circumstances, the implementation in R would have remained a single line in a news files, along with a few lines in the package manual or an entry in a science blog.

Having said that, on the other hand, I understand that the authors of such software solutions usually spend countless hours on coding and maintaining valuable scientific code used over and over for free by others. At some point, they must cash in their effort to get the credit they deserve. Nonetheless, I believe the manuscript content is not significant enough, and the authors should put a little more effort into it. For instance:

- As a non-expert in multidimensional scaling, I feel the manuscript would benefit from more context. The formal description is sufficient and easy to follow, but the likely impact of this manuscript seems low except for having announced a 'new' feature. In other words: How does this new measure perform for real samples and their (new) interpretation? Section 4 reads interesting, but was a new conclusion reached? Did it lead to better (e.g., more accurate, more precise) results, or did the geoscientific interpretation essentially remains the same? If the latter is the case, perhaps you can present a real case underlining the point you want to make better.
- The manuscript comes without a proper discussion. Section 4 is an application example that includes elements of a discussion. However, for a scientific manuscript, I would expect to see more. In particular I would like to see a discussion about the question: Does it likely change the outcome of studies working with this 'new' metric.

- The synthetic data outlines the general problem you want to address. I suggest leading with an example based on a case study where the Kolmogorov-Smirnov distance did not perform as expected for the reasons you have mentioned.

2.3 Scientific quality

The scientific quality of the manuscript is good and valid. I found a few minor inconsistencies though, but nothing out of the ordinary (see detailed comments below).

3 Detailed comments

- L111: I've played a bit with the proposed synthetic data and found that it depends to some extent on the standard deviation. A more narrow standard deviation for the same fixed mean values leads to more complex KS-distance patterns. The higher the degree of overlap (higher standard deviation), the more conclusive the KS distance becomes. Perhaps you can add a few lines about it in the text.
- L150-L175: I think this paragraph can be improved in order to provide a better experience to readers.
  - I had to download the example files from an external repository, but I was expecting to find everything up and running as a supplement to the manuscript; a minor issue, though.
  - I was expecting the R and the Python code snippets to do somewhat the same; just for the two different languages. Instead, the R code loads a CSV file with eight datasets, and the Python code imports only two datasets. The R code does considerably more. It will be easier to understand if the example code lines lead to the same output (they do if I limit the R code example to the datasets of the Python code).
  - The R code snippet produces a plot output. However, if I reduce the dataset, it fails. This appears to be a bug in the package 'IsoplotR' because it returns an uncontrolled error:

```
DZ <- IsoplotR::read.data("scandinavia_short.csv", method = "detritals")
DZ
```

```
## $Byskealven
##  [1] 1507 1769 1762 1077 1246  943 1776 1453 1129 1875 1847 1792 1286 1870 1798
## [16] 1811 1806 1016 1590 1798 1834 1794  989 1457 1832 1856 1794 1787 1809 1875
## [31] 1790 1816 1740 1878 1698 1739 1593 1811 1803 1868 1795 1710 1805 1419 1635
## [46] 1800 1606 1622 1865 1813  968 1627 1497 1812 1782 1823 1813 1387 1623
##
## $Vefsna
##  [1] 1677 1113  987 2501  500 1655  977 1748 1526 1401 1882 1025 1192 1129 1164
## [16] 1522 1652 1734 1157  499 1795  486 1475 1791 1331  580 1043  455 1590 1102
## [31] 1038  561
##
## attr(,"class")
## [1] "detritals"
```

```
# example 1. calculate the W2 distance matrix for the Scandinavian dataset:
d <- IsoplotR::diss(DZ, method = "W2")
```

```
## [1] "W2"
```

```
d
```

```
##        Byskealven
## Vefsna   490.0072
```

```
# example 2. apply MDS to the Scandinavian data set:
try(IsoplotR::mds(DZ, method = "W2"))
```

```
## [1] "W2"
## Error in cmdscale(d, k) : 'k' must be in {1, 2, ..  n - 1}
```

  - I would not call the Python code "implementation", which, to me, implies new, unique software code. Instead, the Python code is a minimum running example showing how W_2 values can be calculated in Python using existing libraries.
  - It is probably evident to the authors that both attempts, R and Python, lead to similar results, but this should be made clear to the readers by quoting the output for each code snippet. Or, if this is too obvious, remove the output after the Python code.
  - L152: If I look into the R code (file `mds.R`), I read in line 199 of the code: `#modified after the wasserstein1d function of the transport package`. It is normal to look up open-source code of others, however, if it helped for the own implementation and since the code line in question seems identical, credit should be given in the manuscript to authors of the package `'transport'` (Schuhmacher et al. 2022)

```
## transport file transport1d.R, line 6
return(mean(abs(sort(b)-sort(a))^p)^(1/p))

## Isoplot, file mds.R, line 199
out <- mean(abs(sort(y)-sort(x))^p)^(1/p)
```

- L164: Please consider adding the example data to the manuscript or the R package
- L167+ (footnote): The repository `pvermees/IsoplotRbeta` does not exist, but I guess the branch `beta` was meant and it should read: `remotes::install_github("pvermees/IsoplotR@beta")`

3.1 Comments on figures and tables

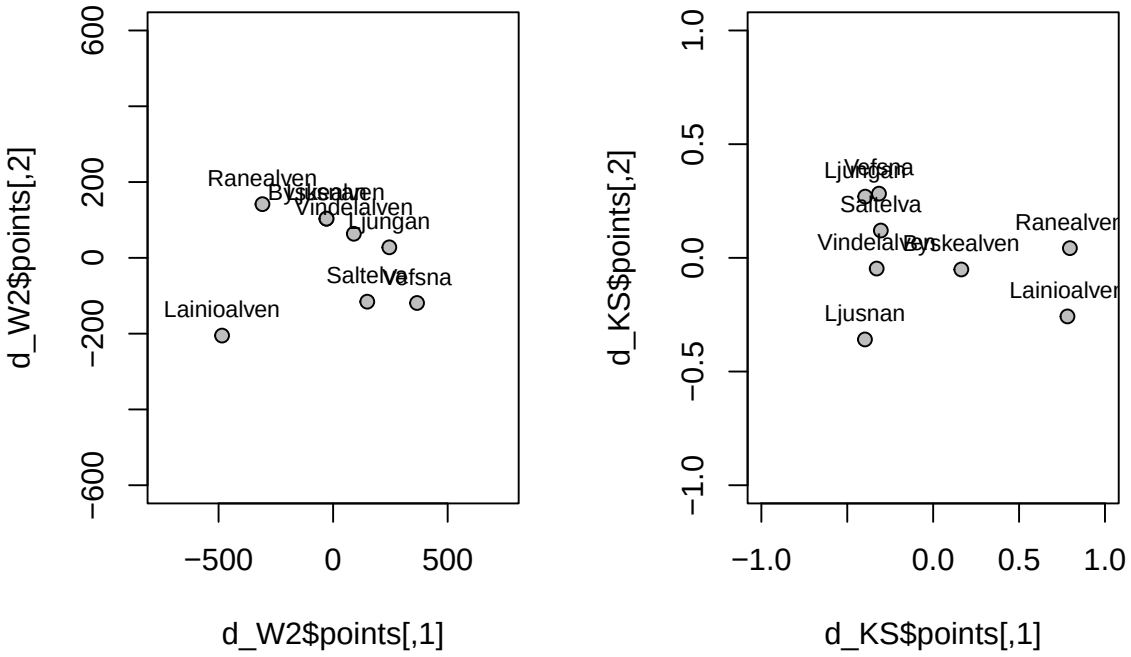
- Figure 1:
  - I suggest using dashed lines for the "semi-transparent colour lines" for better readability.
  - How did you modify the data to "aid illustration"? It appears that you have shifted the 'Byskealven' dataset by 1 Ma. This should be mentioned. If so, how does it affect the W_1 distance? Your pink area becomes considerably smaller if the non-shifted dataset is used. Perhaps you have a better dataset at hand; one that does not need such manipulation.
- Figure 2:
  - The upper plot would benefit from y-axis labelling
- Figure 3:
  - Similar to my comments above, please label the plot axes. I do not doubt that readers familiar with this kind of analysis will have no problem understanding what is shown, but please consider the GChron's broader audience.
  - Something is at odds with the lower figures (j and k), if I try to reproduce them with `IsoplotR::mds()` and the example data from GitHub. Figure 3j does not look as in the manuscript, because 'Ljusnan' and ' Byskealven' seem to be no different. Figure 3k is somewhat mirrored. Below my code, I first used `IsoplotR():mds()`, here more manually to show the calculation steps. The mirrored figure is not a big deal because the interpretation should not change, but it should be presented as the users would see it running the code.

```
DZ <- IsoplotR::read.data("scandinavia.csv", method = "detritals")
## calculate xy values for plots
d_W2 <- IsoplotR::diss(
  x = DZ,
  method = "W2") |> MASS::isoMDS(trace = FALSE)
```

```
## [1] "W2"
```

```
d_KS <- IsoplotR::diss(
  x = DZ,
  method = "KS") |> MASS::isoMDS(trace = FALSE)
```

```
## [1] "KS"
```



- I am not sure how this table helps the readers. The information given is of no particular relevance to the text. Either extend the table and add additional information or remove the table. Later I saw in the text why the 2nd column was in the table. I still think you could add more details. Otherwise, it appears relatively trivial.

4 Additional comments

The following comments refer to something I discovered along the way. It was not considered for my recommendation to the editor.

- I have not used `'IsoplotR'` before, but it appears to be an interesting and mature package. However, something I noticed missing while searching for the new feature implementation was a NEWS file, as it is custom for R packages. The authors may want to consider adding such a file because it helps users understand package changes without inspecting code changes. This is something I find very useful for scientific software packages.
- The R code example returns a default plot (`IsoplotR::mds()`). Perhaps it is obvious to users familiar with the package, but I found it confusing having numbers on the x and y axis but no axes labels. I guess what is shown is some scaled distance. In Section 4, you have explained it better, but personally, I prefer to understand figures without reading the text.
- The default plot `IsoplotR::mds()`returns only the names, but the location in the plot remains a little bit obscure because the default plot setting is `pch = NA`. I find it awkward because it reduces the value of the plot when what is shown remains too vague to be positioned correctly. In particular, because the `text()` position in R can be very different from what you would expect depending on the setting in `pos`.
- In the function `IsoplotR::plot.MDS(),` the standard plot output has the argument `asp = 1` hard coded. This is at least unexpected and leads to inconclusive plot behaviour‘ when `ylim` and `xlim` are modified.

References

Schuhmacher, Dominic, Björn Bähre, Carsten Gottschlich, Valentin Hartmann, Florian Heinemann, and Bernhard Schmitzer. 2022. *transport: Computation of Optimal Transport Plans and Wasserstein Distances*. https://cran.r-project.org/package=transport.