Response to reviewer comments "SWMPr: An R package for retrieving, organizing, and analyzing environmental data for estuaries", by M. W. Beck.

I thank the reviewers for providing thoughtful comments on my manuscript and the SWMPr package. My response to these comments are shown in italics. Page and paragraph numbers refer to the original manuscript.

Reviewer 1:

Main text

I didn't find it obvious from the text whether the setstep and comb functions would handle missing data with interpolation or NA. (It's NA, as it turns out.) Perhaps it is because I don't do a lot of work in this domain, though.

This was clarified in the revision, page 6 paragraph one:

'Both functions treat missing data as NA values, either for observations that exceed the allowable tolerance for the differ argument of setstep or for time series that do not overlap given the method argument passed to comb.'

Package performance

Some of the functions can spend a surprising amount of time parsing. For example, the call: all_params_dtrng('hudscwq', dtrng = c('09/01/2013', '10/01/2013')) took 19 seconds on my machine, and running under a profiler showed that about 12-3 of those seconds were spent parsing the data, not downloading it. The SWMPr::parser code calls htmlTreeParse, which is where all that time is being spent; I found that changing it to xmlTreeParse made almost all of that time go away. The only difference I noticed was that htmlTreeParse converts all of the tag names to lowercase, while xmlTreeParse does not. This can be easily fixed by having the parser() function change the 'out' data frame's names to lowercase before returning. With these two modifications, the elapsed time drops to 6-7 seconds–the bulk of the time now being spent downloading the data (as far as I can tell).

That is a fantastic suggestion and I've made these changes in the development repository of the package. I noticed an immediate increase in performance with the import functions. I will push the new version to CRAN if or when the paper is accepted for publication.

shiny_comp app

The Shiny app that uses Leaflet is using a deprecated API. The new API is documented at https://rstudio.github.io/leaflet/ and I think it would significantly simplify the code. I can lend a hand if the author wants.

I would be interested in working together to improve the code. I made the app by borrowing

an example from the rstudio GitHub repo (https://github.com/rstudio/leaflet/tree/master/inst/legacy/examples/population) and tweaking the code for my application. I admit that I have a fuzzy understanding of using leaflets maps reactively and I have always had trouble editing the code as a result. Simplification with the new API would be great not only for performance, but also to help me better understand reactive programming.

The Leaflet map could use a color legend, which is possible with the new Leaflet API. (It could use a radius legend too, but that is not yet a feature of the Leaflet package.)

There is a feature in the new Leaflet API that lets you have a label http://leaflet.github.io/Leaflet.label/ for each circle that is either always shown, or shown on mouseover. This is a nicer experience for the user than having to click. This feature has not yet made it to CRAN, but it is available in the GitHub master branch of Leaflet.

It would be interesting if the Leaflet app could let you select multiple stations and see the data plotted together.

Again, these are all great suggestions but I would probably need assistance to implement these changes with the new API.

shiny_summary app

In the file https://github.com/fawda123/swmp_summary/blob/master/server.R, the plotInput should not be a reactive(), but a regular function that takes no arguments. This is because reactives cache their values, and should not be executed for their side effects (in this case, plotting). As the app is written now, there doesn't appear to be an actual user-visible bug, but a change as simple as setting the plot width to "100%" would introduce a bug (I'd expect it to fail to redraw the plot when the window was resized).

Yes, the plot fails to render on shinyapps when the window is resized. I had not noticed the bug so I appreciate the comment. I've changed the code in server.R to make plotInput a regular function. I've redeployed the app and the bug has been fixed.

Reviewer 2:

First, I did not see any information about how SWMPR treats censored data, i.e., measurements that can be reported only as less than or greater than an analysis or reporting limit. For example, measurements of nutrient parameters could include values that are below analytical detection limits. A considerable amount of research has shown that the assumptions and methods used to analyze data that include censored values can influence the conclusions of the analyses (for example, see Helsel 2012). Considerable attention is given in the manuscript, appropriately, to various quality assurance/quality control functions and flags, and I realize not every aspect of QA/QC needs to be covered explicitly in the manuscript. However, I believe that the topic of censored values is

important enough that the paper would be improved by providing explicit information about how censored values may be coded, retrieved, and ultimately used in various data analyses carried out by SWMPr.

I fully agree that considering censored data is a necessary aspect of processing and analyzing water quality data. Censored data, either above or below the detection limit, are given unique QAQC flags during initial data processing through the Centralized Data Management Office (CDMO). Censored values in the imported data can be identified using functions in the SWMPr package. The default behavior of the qaqc function is to remove censored values, but the user can also retain these values by changing the function arguments. A data frame showing the number of observations for a given QAQC flag can also be retrieved using the qaqcchk function. I have revised the text to more clearly describe the presence of censored values in SWMP data and how SWMPr can be used to assess their distribution (page 4, paragraph 3):

"SWMP data often contain observations above or below the detection limit for the sensor or laboratory method used to quantify the parameters. For example, nutrient data exceeding the high sensor range are assigned a QAQC flag of -5, whereas data below the low sensor range are assigned a QAQC flag of -4. The presence of censored data is non-trivial and can influence the types of analyses that are appropriate for the time series (Helsel, 2012). A detailed discussion of methods for evaluating censored data is beyond the scope of the manuscript and existing methods for R are provided by other packages (e.g., cents, McLeod et al., 2014). However, the functions in SWMPr can be used to identify censored data based on the appropriate QAQC flag for a given parameter. Viewing this information can be helpful for determining how to further process the data with the qaqc function or alternative methods outside of SWMPr. The qaqcchk function returns a data.frame of the number of observations for a parameter that are assigned to all QAQC flags, including those for censored data. SWMP data should not be analyzed without viewing this information to determine an appropriate method to address data with questionable QAQC flags."

Second, in the "Applications using the SWMPr package" section near the end of the manuscript, simple linear regression (SLR) is identified as a technique used to summarize trends in a parameter over time. This particular statistical method may or may not be appropriate for a given data set. Alternative trend analysis methods exist, including several that treat censored values appropriately (again, Helsel 2012 is a useful reference). It is unclear from the manuscript what alternatives to SLR may be available in SWMPr (e.g., maximum likelihood estimation, Akritas-Theil-Sen nonparametric regression, others). Therefore, clarifying text that either identifies the availability of other trend analysis procedures in SWMPr or provides caveats to the use of SLR (e.g., used only for initial exploratory analysis) should be added to this section.

I agree that the use of SLR is a simple method that may not be appropriate for all datasets provided by SWMP. The application is meant to be exploratory and I have revised the text to clarify that SLR is a simple and potentially inadequate method for trend analysis, particularly with censored data. The text description on the application was also revised.

Page 12, paragraph one: 'More robust methods for evaluating trends are currently not provided by the application and the use of simple linear regression is meant for exploratory purposes only.'

Added to app: 'Please note that the use of simple regression to identify trends is for exploratory purposes only and may not be appropriate for all datasets.'

Reference: Helsel, D.R. 2012. Statistics for Censored Environmental Data Using Minitab and R. 2nd edition, John Wiley & Sons, Inc. Hoboken, New Jersey. 324pp.

This reference was added for the text additions above.