`rshift` STARS manual - regime shift analysis for paleoecological data
v2.2.0

A. H. Room, F. Franco-Gaviria, D. H. Urrego

January 5, 2023

1 Introduction

STARS (Sequential T-test Analysis of Regime Shifts) is an algorithm which detects regime shifts in an ecosystem [Rodionov, 2004] - i.e., where the ecosystem undergoes a persistent change in some measure. Regime shifts can happen as a direct result of extrinsic factors like fire or tidal disturbance, or by the breaking of a 'threshold factor' intrinsic to the ecosystem itself.
This document will illustrate the theory behind STARS, as well as a guide on how to easily deploy it in R using the `rshift`[1] package.

2 Theory

STARS is based on the Student's t-test[2]; it goes along the data set, comparing each new observation against every entry in the current 'regime'. If an entry deviates significantly from the regime average, it becomes a hypothesised 'shift point', and the algorithm then tests if this shift is persistent. If it is, then an RSI (regime shift index) is created to quantify the size of the shift, and the algorithm starts over using the shift point as the start of a new regime. Otherwise, it rejects the shift point, adds it to the current regime, and continues the search.
You will need a table with two columns; an ordinated proxy or variable (such as pollen, temperature, or charcoal), and some form of time series, like a list of depths for samples or a chronology. From there, it goes like this:

1. Firstly, we choose an 'l' value. This is essentially your predicted length for each regime, and has knock-on effects for the sensitivity of the algorithm. If you choose a bad value, don't worry - you can always change it and start over. 10-15 is usually a good starting point. We also need a probability for statistical significance p - usually 0.05 (a 5% chance that a positive result is a fluke).

[1] found at https://github.com/alexhroom/rshift
[2] which can be problematic; see section 4

2. We then calculate the difference between mean values that would be significantly different - i.e., how much your value has to shift between two adjacent observations for it to be considered a new regime. The formula for this is

$$\text{diff} = t_p(2l-2) \cdot \sqrt{2\sigma_l^2/l}$$

There are a couple new things here: $t_p(2l-2)$ is the critical value for a two-tailed t-test, at probability level p, with $2l-2$ degrees of freedom. We don't need to know what all these things mean just to use STARS - just find $2l-2$, google a t-table, and grab the value from it.
σ_l^2 is an average 'rolling variance' - the variance of each adjacent group of 10 values. To find this, we calculate the variance for values 1 to 10, 2 to 11, 3 to 12, and so on until we reach the last ten, then find the mean.

3. We calculate the initial mean of your proxy for regime 1 $\overline{x_{R1}}$, and find the range of acceptable values in this regime, which is the range from $\overline{x_{R1}} - \text{diff}$ to $\overline{x_{R1}} + \text{diff}$. Then, for each value starting with the $l+1$'th value, check if it is outside the acceptable range. If it is, we consider this as our candidate shift point. Hereon, this candidate sample will be called j.

4. We then calculate the Regime Shift Index (RSI) to test if it is a real shift point. The equation for this is:

$$RSI_{i,j} = \sum_{i=j}^{j+l-1} \frac{x_i^*}{l \cdot \sigma_l}.$$

I will explain the process of how to use this, but first it is important to note that here, the direction of the shift becomes important. x_i^* is equal to $x_i - (\overline{x_{R1}} + \text{diff})$ if the shift is up (i.e. $x_j > \overline{x_{R1}} + \text{diff}$), or $\overline{x_{R1}} - \text{diff} - x_i$ if the shift is down. (so $x_j < \overline{x_{R1}} - \text{diff}$) σ_l is trivially the square root of σ_l^2, which you calculated in step 2. Furthermore, if the RSI value turns negative at any point during that sum, stop calculating and go onto step 5.
Of course, looking at that equation one might be confused on what to do with it, so I will explain what it means:
i is an index number; it stands for every value in the data between j and $j+l-1$. This means, starting from the j'th value of the data:
find x_i, the i'th value of your proxy/variable. Then, use it to calculate x_i^*. Divide this by $l \cdot \sigma_l$, and add it to a running total. Repeat with $i+1$, $i+2$ and so on until you reach $j+l-1$. Each time you add a new value to your total, check if it is positive or negative; if negative, stop calculating and go onto step 5.

5. if RSI has gone negative, the test has failed. We consider RSI for x_j to be 0. Add x_j to the values for regime 1 (i.e. use it when calculating $\overline{x_{R1}}$, and go back to step 3. If it remained positive, then your final value is the RSI value for x_j, and j is the start point for a new regime.

6. Calculate the value of $\overline{x_{R2}}$, i.e. the mean of all the values you used for the RSI calculation. We then start the search for regime 3 from $j+1$, rather than from l values along like before; this is so we still detect shifts even if regime 2 is a different length to l. Repeat steps 3 through 6 until you have processed all the data.

If you would like an example of this in action, see Rodionov's original paper [Rodionov, 2004], which shows how it works with annual PDO data for temperature in January.

3 Using STARS with `rshift`

Of course, all this grunt work is very boring. With modern technology, we can make computers do it for us instead! There is a function for this built into the `rshift` R package.

3.1 Using `Rodionov()`

The `rshift` command for STARS is `Rodionov()`. It takes 7 variables, so in RStudio will look more like `Rodionov(data, col, time, l, prob = 0.95, startrow = 1, merge = FALSE)`. These are the 7 inputs you can put into the function, but only the first 4 are mandatory (the rest default to what's after the = sign).

- `data` - the dataset in R that you're using. This should NOT be in quote marks.
- `col` - the column that your ordinated proxy/variable data is in. The name of it MUST be in quote marks.
- `time` - the column containing your chronology/depth for each entry. This MUST also be in quote marks.
- `l` - the cutoff length, as explained in step 1 of the theory.

The other 3 variables are optional:

- `prob` - the significance probability, in the form $1-p$. Defaults to $p = 0.05$.
- `startrow` - where the algorithm should start from, if you want to skip the first few rows of the dataset. Defaults to 1.
- `merge` - changes the result to be either a regime-shift only table (if FALSE), or an addition to the original table (if TRUE). If merge = FALSE (default), produces a 2-column table of 'time' (the time value for each regime shift) and 'RSI' (the RSI for each regime shift). If merge = TRUE, returns the original dataset with an extra RSI column, giving the RSI for each time unit - 0 for non-shift years.

So, as an example: say my dataset is called 'lake_data'. I've taken pollen data, and organised it using a DCA, so my DCA data is in a column of the dataset called 'DCA1'. The chronology of the sample is in a column called 'Age', and I choose l to equal 10. I would then type the following:

```
Rodionov(lake_data, "DCA1", "Age", 10)
```

and it would return a list of points in the chronology where it has detected a regime shift.

3.2 Visualising your data

`rshift` also contains `RSI_graph`, a way of visualising the results of a STARS test. Using our previous lake_data, we run `Rodionov()` with `merge` set to TRUE, save it as RSI_lake_data with the `<-` operator, then use `RSI_graph()` to create a graph.

```
RSI_lake_data <- Rodionov(lake_data, "DCA1", "Age", 10, merge = TRUE)
RSI_graph(RSI_lake_data, "DCA1", "Age", "RSI"))
```

It takes the data, variable column and chronology like before, but you must also specify the name of the column with RSI values in (which `Rodionov()` automatically creates as 'RSI').

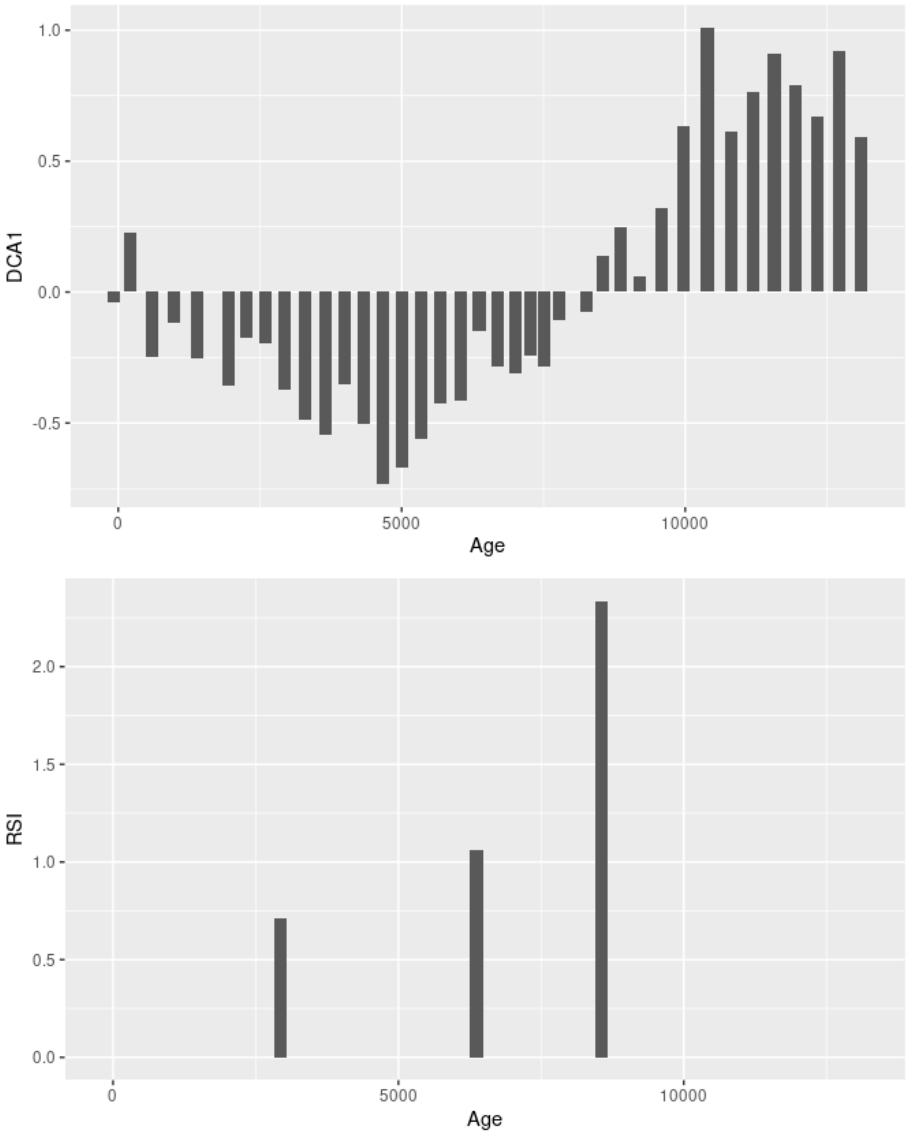

Figure 1: The visualisation produced by `RSI_graph()`

4 Evaluating use of STARS

There are some considerations one should make when using STARS to measure regime shifts. The main one is that the t-test works on the assumption that your observations are independent and normally distributed - i.e. they follow a bell curve where the mean equals the median. For variables like temperature, this is a fair enough assumption, but for other proxies like pollen it may be less reasonable to assume this. In those cases, one may want to consider using a different method such as Lanzante [1996]'s L-Method, based on the Mann-Whitney U test. This is also accessible in `rshift` using the `Lanzante()` function.

Another concern might be with the choice of l - it is rather arbitrary (although for a lot of particularly 'bad' values it simply will not work at all), and one should keep an eye on how it may affect the validity and objectivity of their findings.

References

John R Lanzante. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 16(11):1197–1226, 1996.

Sergei N Rodionov. A sequential algorithm for testing climate regime shifts. *Geophysical Research Letters*, 31(9), 2004.