TTAinterfaceTrendAnalysis: A R GUI for Routine *T*emporal *T*rend *A*nalysis and Diagnostics

David Devreker & Alain Lefebvre

December 16, 2015

1 Introduction

Trend detection in hydrologic and water quality time series has received considerable attention in the recent past for pure research purposes on methodological development and/or assessment of ecological status [1,2]. Trend assessment is of particular interest because environmental changes have been sometimes unusual in the past few decades with consequences on the environment in general, including impacts on living resources and fishing management. Time-series are of importance to studies on the biological influence of anthropogenic effects and climatic changes, both in themselves and in providing a baseline and/or reference conditions for future investigations. These works could be particularly helpful to scientists and policymakers within the framework of European directives (Water Framework Directive (WFD 2000/60/EC), Marine Strategy Framework Directive (MSFD 2008/56/EC)) and regional sea convention (OSPAR convention [3]).

In this context and due to the heterogeneity, the variety of data sources and the variety of statistical analysis methodologies, there is a clear need for unifying methods of temporal trend assessment, for the purpose of a given directive or convention devoted to assess trends towards or from a good environment status. Consequently, the authors proposed the TTAinterfaceTrendAnalysis package with the aim to develop a standard procedure (routine) to assess temporal trend on database and which can be entirely done through a Graphical User Interface (GUI). Package containing temporal trend tests already exist in R language such as wq or pastecs, however they need command lines and are only ideal for advanced users who can deal with it. The TTAinterfaceTrend Analysis package uses the power of such tools but simplify their use (from database creation to results interpretation) through a GUI for standard, non-statistician user.

2 The GUI Organisation

The TTAinterface provides a GUI using Tcl/Tk interface. After loading the package the GUI start automatically. The GUI works through 4 successive panels:

- The panel '1-Data_management' focuses on the file and data management, it is the pre-processing part.
- The panel '2-Parameters_selection' focuses on the selection of the parameter and categories to analyse.
- The panel '3-TimeSeries_building' displays the option to build a regularised time series.
- The panel '4-Diagnostics/TrendAnalyses' focuses on diagnostic tools and statistics tests.

The right part of the GUI displays the results and different warning messages and advices. Help buttons are available on each panel to provide guidelines on how to use the options in their respective panel.

3 Files and data management panel (Panel 1)

3.1 Import data

The first panel of the GUI allows the importation of your database (1 in Figure 1) in the interface (a txt file). The program identifies each column as a function of its label and type (numerical, character, vector). In general, columns with numeric values are automatically identified as parameters. Other columns have to be manually labelled to facilitate the identification by the interface, such as sampling stations or depth. The User Guide, a pdf file, gives all information to build a compatible txt file. The package is also provides with a database (an Ifremer dataset of chlorophyll-a concentration monthly measured offshore Dunkerque in North of France) that can be useful as an example (2 in Figure 1).

Until the file is imported, panels 2 to 4 stay empty and panel 1 uncompleted. The other options will be available only when your data is imported (Figure 2).

3.2 Save directory

By default, all tables, plots and results of analysis display with the interface are saved in the same directory as your txt file. However if you want to save your results in a different folder, just click on the *Select your save directory* button (1 in Figure 2) and choose a new folder. The save path is display under the button.

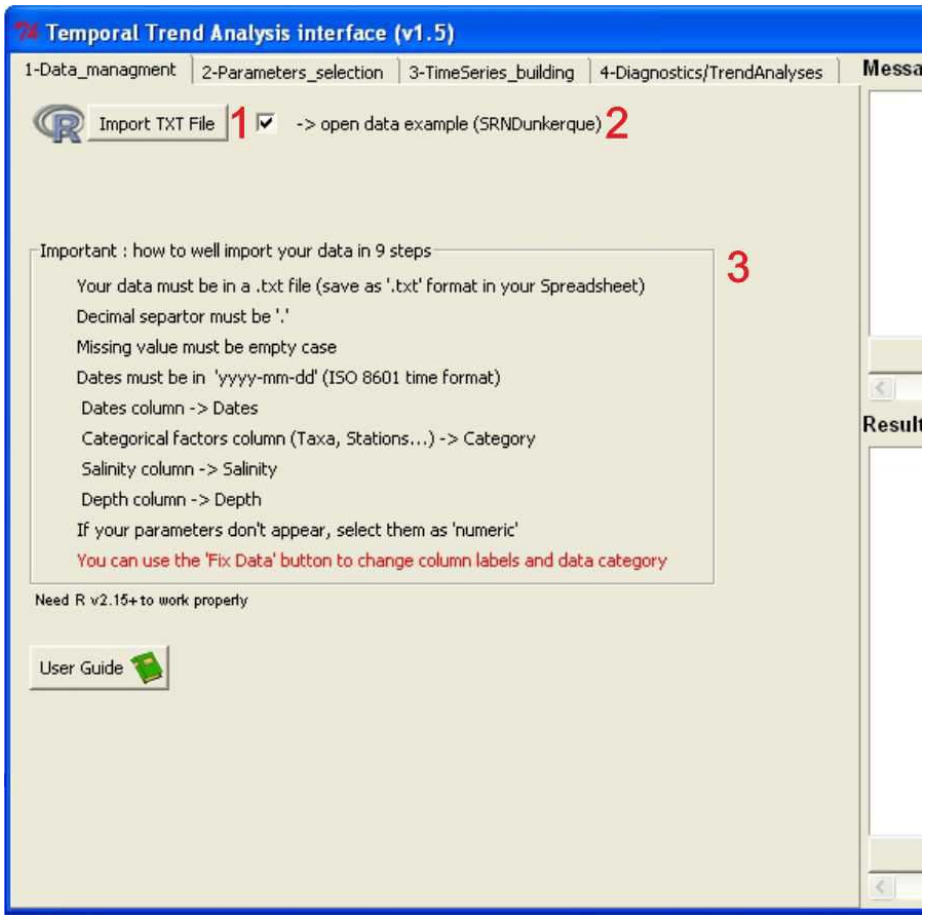

Figure 1: Panel 1 (Data_management) with the import options (1 and 2) and importation advices (3).

From this path, the programme automatically creates an arborescence to save your files based on the options you choose to perform analyses (Figure 3). Other options, such as months, salinity and depths, are not added in order to limit the arborescence declination and keeping a clear save path, therefore the user has to be careful to not overwrite its files (by changing the save directory destination) if these options are changed between two analyses.

3.3 Editing the data

In case there are importations issues, the data can be edited by using the *Fix Data* button (2 in Figure 2). This button called the fix() function who is similar to the one present in R commander (*Edit* command) [4]. Once edited, the new dataset is automatically saved (FileName_fixed.txt) and read by the interface.

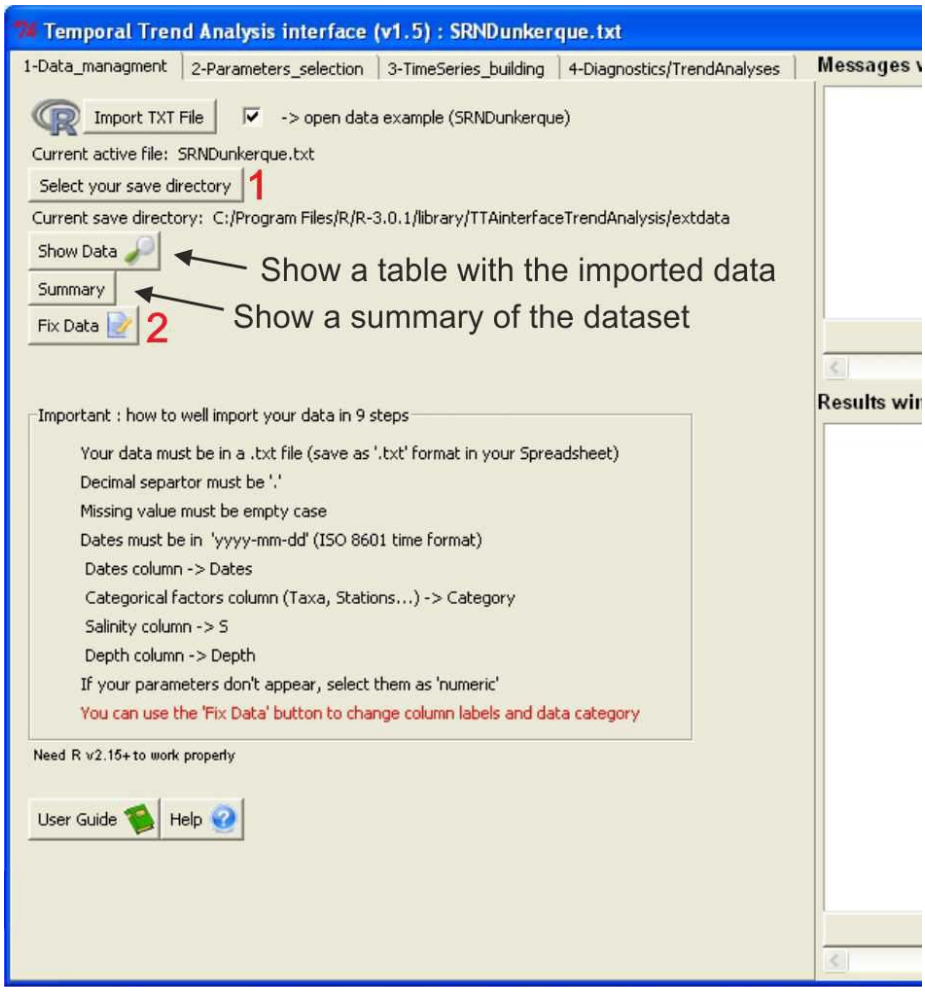

Figure 2: Panel 1 (Data_management) with all options available. The top panel displays the name of the imported data file.

Unfortunately data type cannot be saved in a txt file. Editing the data will not change your save directory.

4 Parameters selection panel (Panel 2)

If columns are correctly labelled, lists, sliders and frames should be automatically filled with appropriate values (Figure 4).

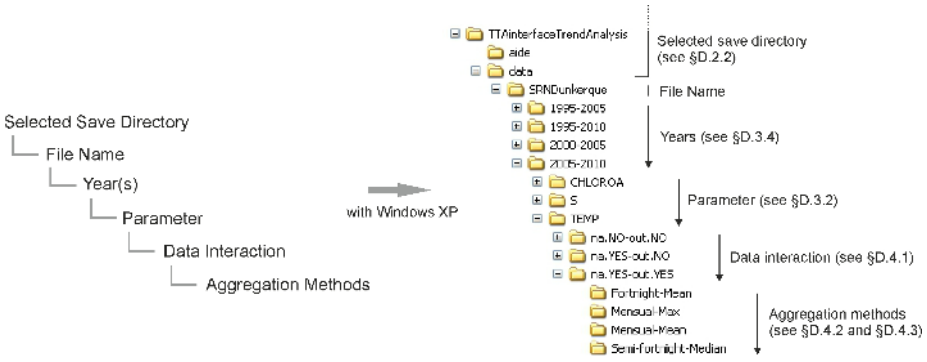

Figure 3: Example of a created save path arborescence based on options selected.

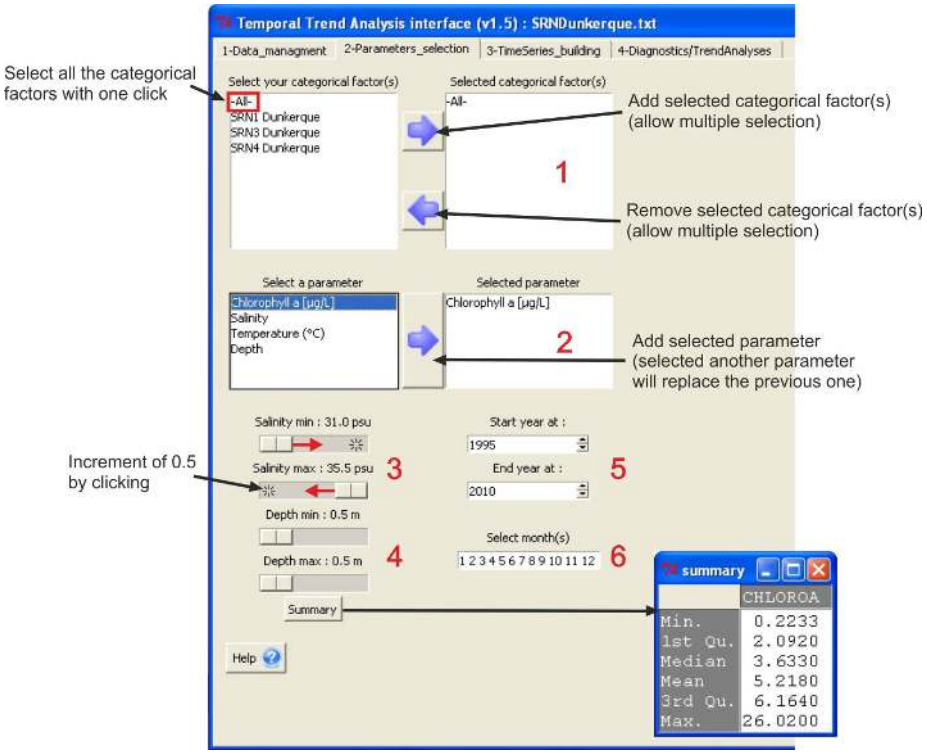

Figure 4: Panel 2 (Parameters_selection) of the TTAinterfaceTrendAnalysis package.

4.1 The categories

The categories to be analysed can be selected and removed using the arrows between the two selection boxes (1 in Figure 4). This supports multiple selection

using the *Ctrl* key or by dragging the cursor. All categories can be analysed at once by selecting -All- in the left box. The data of all selected categories are melt for analysis.

4.2 The parameters

The process of selecting the parameter to be analysed is the same as for the categories, except that only one parameter can be selected (it does not support multiple selection) (2 in Figure 4). To replace a parameter already selected by another one, just select the new parameter in the left box and click on the arrow, it will automatically replace the previous one in the right box.

4.3 Depth and salinity

In some cases analyses have to be performed at specific depth or salinity (which characterise the studied water masses). There are 4 sliders to select these salinities and depths (3 and 4 in Figure 4). By default these sliders display the maximum and minimum values of salinity and depth in your dataset (if they exist). By keeping these values unchanged all data are taken into account for analysis, including data at missing salinity and depth. If modified, data at missing salinity or depth are excluded from the analysis. Analysis can be performed at a unique depth or salinity by giving the same min and max values.

4.4 Years and months

As for salinity and depth, years and months to be analysed can be modified. By default the two lists in panel 2 display first and last years and the months present in your dataset (5 and 6 in Figure 4). Years can be modified just by clicking on the arrows or by typing it. Months can be deleted or added (the order is irrelevant) and there must be a space between the months. This can be useful to process data for a given period of a year, for example, to compare the productive period (in terms of phytoplankton development) versus the non-productive period of an area such as within the WFD.

5 Time series building panel (Panel 3)

Temporal trend analyses generally need regularised time series to be performed. The third panel focused on time series regularisation (Figure 5).

5.1 Missing values and outliers

The option *Replace missing values* (1 in Figure 5) replace the missing values from the time series (and not from the raw data) by values calculated from the aggregated data in two successive steps.

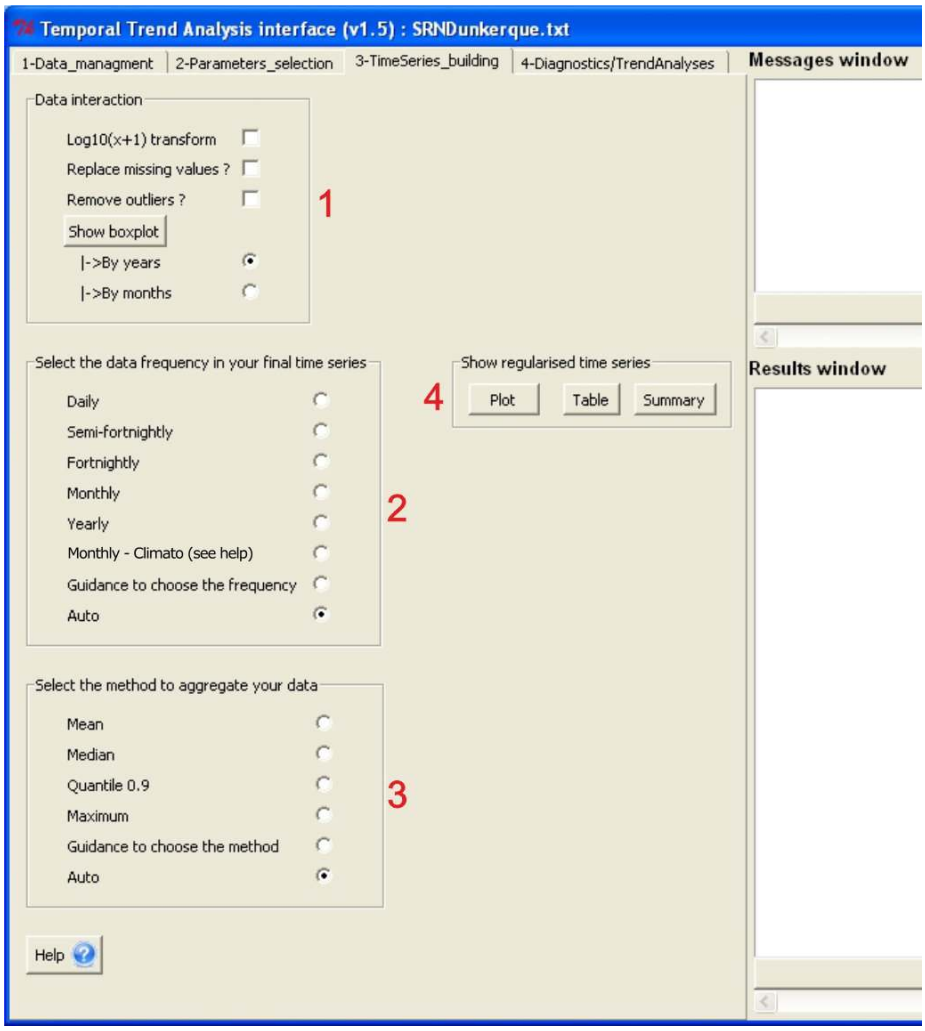

Figure 5: Panel 3 (TimeSeries_building) with default options allowing the temporal trend analysis to be performed.

- As time series generally present strong autocorrelation, missing values can be estimated (predicated) from linear regression of values around the missing value, however this is relevant only when few data are missed, long period of missing values cannot be replaced using this method (like seasonal fluctuations)
- When missing values are present over a long period, they can be replaced by the median of data from the same cycle (e.g. month, week, year, depending on the time step chosen), inversely this method is less relevant

than regression for shorter period of missing values (it loses the dependency due to autocorrelation).

The present interface uses a combination of both methods; the linear regression method to replace missing values in 3 successive units of time and the median method for longer period of missing values. The median method acts first to reduce the lag between missing values and to allow the regression method more frequently. Missing values at the beginning and at the end of the series are replaced using the median method if possible or are ignored.

Data distribution frequently contains outliers; these outliers are due, for example, to error of measurements or extreme natural event. In some cases these outliers can greatly influence statistical analysis comparatively to the rest of values and it should be interesting to remove them. The second option present in the frame (1 in Figure 5) allows you to remove these outliers and to save them in a separate txt file (in case you need to identify them). The method used to identify outliers is the boxplot method by years [5]. The *Show boxplot* button displays the box and whiskers plot with outliers by months or by years, this is the boxplot() function of the graphics package.

Both options, missing values and outliers, can be checked together or independently; outliers will be always removed first and missing values in second places.

5.2 Time step selection

To build a regularised time series, the interface will aggregate raw data from the same period (day, week, month, year) using an appropriate method (mean, median, max value). These different options are available in frame 2 and 3 of the third panel (2 and 3 in Figure 5).

The second frame in panel 3 *Select the data frequency in your final time series* allows the selection of the time step, using radio button, at which the programme aggregates the data (2 in Figure 5). Eight options are available; the 5 first options are classic frequencies: daily, semi-fortnightly, fortnightly, monthly and yearly. Monthly - Climato aggregate all data by month, all years including.

It is better to choose a time step in relation with the theoretical sampling frequency of your data in order to keep the maximum of information without creating too many missing values. The option *Guidance to choose the time step* suggests a balanced choice by computing the mean time and the minimum and maximum period that separates two successive measurements in your database. This method is inspired from the pastecs package [6]. Arbitrary, if mean time between two measurements is under 5 days the interface advice the daily time step; semi-fortnight time step for [5-10[days ; fortnight time step for [10-23[days; monthly time step for [23-60[days and over 60 days annual time step is advice. Monthly - Climato time step is only available in manual choice. You are free to follow these suggestions or to select another time step. The auto

option (default option) will automatically apply the advice without displaying the suggestion.

5.3 Method of aggregation

The third frame *Select the method to aggregate your data* (3 in Figure 5) allow the method with which data will be aggregated at the time step previously set to be chosen. Four methods are available: by averaging the data (Mean), by selecting the median of the data, by selecting the quantile 90% of the data or the maximum of the data of the same time step. The guidance option will also suggest the method that best fits the original data distribution. The interface compares data distribution obtained with each method (at the selected time step) with the raw data distribution using an ANOVA with Dunnett's post-hoc test. The comparison with the highest p-value (less significant difference) determines the best method. The auto option (default option) automatically applies the advice without displaying the suggestion.

5.4 Visualised your regularised time series

The fourth frame *Show regularised time series* (4 in Figure 5) display the newly build regularised time series through a plot or a table that will be saved. The table display column labels which vary as a function of the time step you selected.

6 Diagnostics, statistics and results (Panel 4)

6.1 Diagnostic tools

The options present in the first frame of the forth panel *Diagnostics (optional)* (Figure 6) are not required to perform temporal trend analysis but give additional information that can be useful to explain some patterns in the time series.

6.1.1 Spectrum analysis

The spectrum analysis option displays a periodogram of the regularised time series (z). Use the spectrum() function of the stats package.

6.1.2 Autocorrelation

The autocorrelation option computes and plots the autocorrelation function on the regularised time series (z) with confidence interval at 0.95. This is the acf() function of the stats package.

```
> acf(z, lag.max = ((nrow(TimeSerie))/2), na.action = na.pass)
```

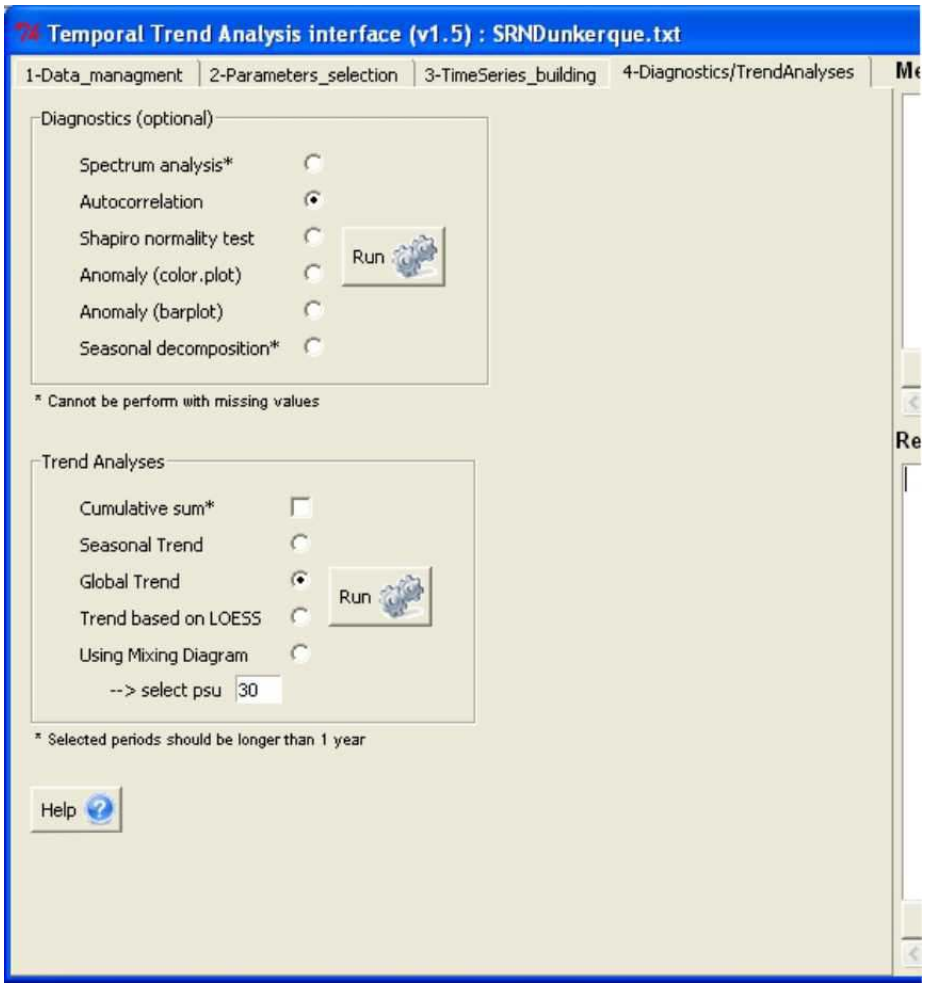

Figure 6: Panel 4 (Diagnostics/TrendAnalyses) with default options allowing diagnostics or temporal trend analysis to be performed.

6.1.3 Shapiro normality test

The *Shapiro normality test* tests the null hypothesis that a sample came from a normally distributed population (Null hypothesis: follow a normal distribution, thus if the p-value is lower than the chosen alpha level (0.05 in our program), the sample does not follow a normal distribution). This is the shapiro.test() function of the stats package.

6.1.4 Anomaly (color.plot)

The anomaly option computes time series anomalies as $X_{ij} - X_i$, with X_{ij} value of the parameter X at the period i of the year j and X_i the median of the parameter X for the period i (all year mixed). The option produce a contour plot with the areas between the contours filled in solid colour. Red colours show positive anomaly and blue colours negative anomalies. White areas occur when there are missing values. This option works only with time series build at monthly, semi-fortnight and fortnight time step. Use the filled.contour() function of the graphics package. For more informations about color plots see also [7].

6.1.5 Anomaly (bar.plot)

Displays a bar plot that show the anomalies of the time series calculated for each time step. Each anomaly is the difference between the value at the time step and the median of the entire regularized time series. Values that are under this median are negative anomalies (blue bars in the figure) and values over this median are positive anomalies (red bars in the figure). Use the barplot() function of the graphics package.

6.1.6 Seasonal decomposition

This option decomposes and plots the regularised time series into seasonal, trend and irregular components using loess. This is the function stl() of the stats package.

```
> stl(z2, s.window="periodic", t.window=(F*10), na.action=na.fail)
```

6.2 Temporal trend tests

The second frame of the fourth panel displays the available tests to perform the temporal trend analysis (Figure 6). Significant results are display in **bold** (p.value <0.05).

6.2.1 Seasonal Trend

The Seasonal Trend option allows performing a Seasonal Kendall test on the time series with details of trend as a function of the time step selected (Figure 7). This is the seasonTrend() function of the wq package. For more information about Kendall test see [8,9].

6.2.2 Global Trend

Same as above but gives the general trend without detail. This is the seaKen() function of the wq package.

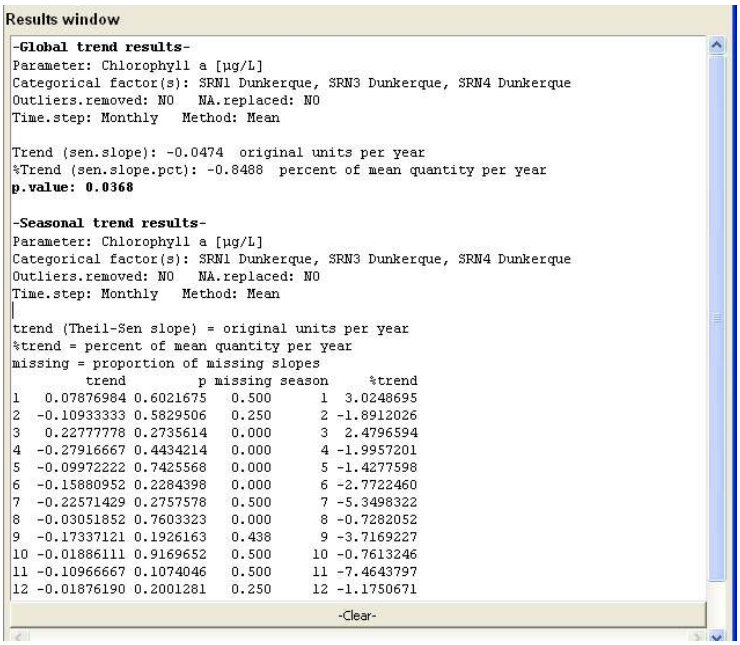

Figure 7: Display of the results of the Seasonal Trend analysis (monthly scale) saved in a txt file named 'OriginalName_Seasonal Trend_Parameter.txt'.

6.2.3 Cumulative sum

This option plots a cumulative sum curve of the time series and allows to manually identify changes in the pattern (shift, trend) in the time series. This is the local.trend() function of the pastecs package. For more information about this function see also [10]. Once the periods are identified, the programme performs the Global or Seasonal Trend test (as selected by the user) on each period. The cumulative sum curve is automatically calculated from the time series with missing values removed (cannot work with missing values), however the trend calculations are performed on the time series build with your own options (so even with no replacement of missing values).

6.2.4 Trend based on LOESS

In this method, a loess smoothed curve of the regularised time series is considered in order to perform a Global Trend test. This is the loess() function of the stats package.

```
>  Loess <- loess(param ~ time, Regularised.data, family="gaussian",
+                 span=0.25, control = loess.control(surface = "direct"),
+                 na.action=na.exclude)
>  tsLoess <- ts(predict(Loess),
```

```
+                start=(min(Regularised.data$YEARS)), deltat=freq)
```

6.2.5 Mixing Diagram

To consider the temporal trend of nutrient concentration in a salinity gradient, a widely-used method consists of using monthly normalised concentration of nutrient at fixed salinity (generally 30) instead of raw data to perform temporal analyses [11]. To normalise, a monthly linear regression is done between raw salinity and nutrient concentration (one regression per month). From these linear regression equations, normalised concentrations of nutrient are estimated at the salinity you enter in the text box 'select psu'. Thus, a monthly time series is built using the new normalised concentrations instead of the aggregated raw data (this test is independent from the time step and aggregation method selected on panel 3). A Global Trend analysis is performed on this time series.

All mixing diagrams are saved for each month and year but only the final result is displayed by the interface. A txt table containing all normalised concentration of nutrient per month/year is also generated and saved.

7 Summary

The TTAinterfaceTrendAnalysis gives the possibility to perform temporal trend analysis through powerful tools developed in R with the ease of a GUI for standard user. Of course the advantage of GUI is limited for the advanced user, some argument of complex function are settled and some advanced options that R packages offer are not available through this GUI. However this is not what the TTAinterfaceTrendAnalysis is developed for. The importance of GUI creation to open R tools to standard user was already discussed by [12] and [13].

This tool was recently proposed to the Intersessional Correspondence Group on Eutrophication (ICG EUT) within the regional sea convention OSPAR for application by Contracting Parties, on a voluntary basis, in the Common Procedure to help to assess the eutrophication status of the OSPAR maritime area. This tool is also well adapted for analysis of ecological and environmental data at large conditionally to have the minimum prerequisite in the data frame (columns with category names, dates and parameter). The GUI can also help with training sessions on R or for educational statistical purposes.

References

[1] Smetacek V. and Cloern J.E., *On Phytoplankton Trends*, Science(319),1346–1348, 2008

[2] Goberville E., Beaugrand G., Sautour B. and Treguer P., *Early Evaluation of Coastal Nutrient over-Enrichment: New Procedures and Indicators*, Marine Pollution Bulletin(62), 1751–1761, 2011.

[3] OSPAR Commission, *Integrated Report 2003 on the Eutrophication Status of the OSPAR Maritime area based upon the First Application of the Comprehensive Procedure. OSPAR Eutrophication Series, publication 189/2003. OSPAR Commission, London.*, 2003

[4] Fox J., *The R Commander: A Basic-Statistics Graphical User Interface to R*, Journal of Statistical Software(14), 2005.

[5] WGSAEM, *Report of the Working Group on Statistical Aspects of the Environmental Monitoring (WGSAEM)*, ICES CM 2000/D:1, Agenda item 4, 2000.

[6] Grosjean P. and Ibanez F., *Package for Analysis of Space-Time Ecological Series. PASTECS version 1.2-0 for R v.2.0.0 & version 1.0-1 for S+2000 rel 3*, 2004.

[7] Hirsch R.M., Slack J.R. and Smith R.A., *Techniques of Trend Analysis for Monthly Water Quality Data*, Water Resources Research(18), 107–121, 1982.

[8] Hirsch R.M. and Slackh J.R., *A Non Parametric Trend Test for Seasonal Data with Serial Dependence*, Water Resources Research(20), 727–732, 1984

[9] Ibanez F., Fromentin J.M. and Castel J., *Application of the Cumulated Function to the Processing of Chronological Data in Oceanography*, Comptes Rendus-Academie des Sciences Paris Série 3 (316), 745–745, 1993.

[10] OSPAR Commission, *Common Assessment criteria, their (region specific) Assessment Levels and Guidance on their Use in the Area Classification within the Comprehensive Procedure of the Common Procedure. OSPAR 02/8/2-E.*, 2002.

[11] Cleveland W.S., *Visualizing Data*, Summit, New Jersey, U.S.A., 1993.

[12] Unwin A., *Oscars and Interfaces*, Journal of Statistical Software(49), 1–18, 2012.

[13] Valero-Mora P.M. and Ledesma R.D., *Graphical User Interface for R*, Journal of Statistical Software(49), 1–8, 2012.