Confidence package: An introduction

Dennis J.J. Walvoort
Alterra – Wageningen University & Research Center
Wageningen, The Netherlands; e-mail: dennis.walvoort@wur.nl

Willem M.G.M van Loon
Rijkswaterstaat Water, Transport and Living Environment; Department of Information Management
Lelystad, The Netherlands; e-mail: willem.van.loon@rws.nl

2014-09-18

Contents

1 Introduction

This tutorial provides a brief introduction to the **confidence**-package. This package can be used to estimate the confidence of state classifications (*e.g.*, with the classification 'bad', 'moderate', 'good') produced using environmental indicators and associated targets. The implementation closely follows Baggelaar *et al.* (2010) where confidence intervals for the estimated multiyear averages are derived by assuming a Student's t distribution for the errors.

2 Input data

As input, the tool needs a table containing annual arithmetic average concentrations of chemical parameters or ecological quality ratio's (EQRs). An example of such a table is given below. In this tutorial, required column names are given in upper-case, and optional column names in lower-case. The tool itself treats column names case-insensitive to minimize the risk of errors by users unfamiliar with case-sensitive software. Users are free to add additional columns. These will be ignored by the tool.
The following columns need to be specified (the order of the columns is irrelevant)

OBJECTID
: unique identifier of a waterbody, *e.g.*, NL89I_os;

PAR
: parameter name, *e.g.*, Cadmium or BEQI2;

OBJECTID	PAR	YEAR	color	VALUE	TARGET	UNIT	transform
River X	Compound Y	2011	green/orange	0.63	0.60	ug/L	log
River X	Compound Y	2012	green/orange	0.87	0.60	ug/L	log
River X	Compound Y	2013	green/orange	0.55	0.60	ug/L	log
River Y	Compound Z	2011	green/orange	2.46	1.40	ug/L	log
River Y	Compound Z	2012	green/orange	1.50	1.40	ug/L	log
River Y	Compound Z	2013	green/orange	2.16	1.40	ug/L	log
River Y	Compound Z	2014	green/orange		1.40	ug/L	log
River Y	Compound Z	2015	green/orange	2.52	1.40	ug/L	log

YEAR
: year expressed as a four-digit integer (YYYY);

VALUE
: annual average value of `PAR`;

TARGET
: the target value for `PAR`, *e.g.*, the target value according to the European Water Framework Directive, or any other used environmental target;

UNIT
: the measurement unit of `PAR`. This unit should be the same for all records with the same `PAR` and pertains to both `VALUE` and `TARGET`;

transform
: data transformation applied to column `VALUE`. Allowed transformations are `log` and `logit`. In case no transformation is required, this field should either be omitted or contain one of the following synonyms: `NA`, `none`, or `identity`;

color
: The fill colors in the density plot to the left and right of `TARGET`. The following values are allowed: either `green/orange` or `orange/green`. The former is the default in case the color-column is missing.

3 Running the tool

The easiest way to execute the tool is by typing

```
> conf()
```

on the R-prompt. This will launch an interactive file selection dialogue that asks for the name of the comma separated values file (CSV) containing the input data. The format of this file is given in Section 2. After checking the contents of this file, the tool will estimate for each `OBJECTID` and `PAR`, the multiyear average of `VALUE`, the probability that this average exceeds `TARGET`, and the 90% confidence interval of this average. These results are reported as an HTML-document and a CSV-file. Both are stored in the same directory as the input file in a directory called 'output' with the current date-time stamp as postfix. This should prevent accidentally overwriting previous results. For example, if the name of the input file is

```
"my_directory/my_input_file.csv"
```

then the names of the output files will be:

```
"my_directory/outputYYYYmmddHHMMSS/output.csv"
"my_directory/outputYYYYmmddHHMMSS/output.html"
```

The first output file contains all results in CSV-format:

The second file is an HTML-report of the analysis, and will be automatically launched in a web browser. As an alternative, one may enter the filename directly

```
> conf("my_directory/my_input_file.csv")
```

This option may be convenient in case many input files have to be analysed. Suppose the names of these files have been stored in a character vector with the name `filenames`, then each file can be processed by running the following code:

```
> for (filename in filenames) {
+     conf(filename)
+ }
```

Apart from input files, the tool can also process `data.frames`. This option is very convenient when the input data are not available in an external CSV-file but are for instance stored in a database. In that case, the user has to run an SQL-query in R that results in a `data.frame` that complies with the format in Section 2. This `data.frame` can then be directly processed by the tool, circumventing the need to create external CSV-files.

```
> conf(my_data.frame)
```

The results will be stored in the current working directory.

4 Results

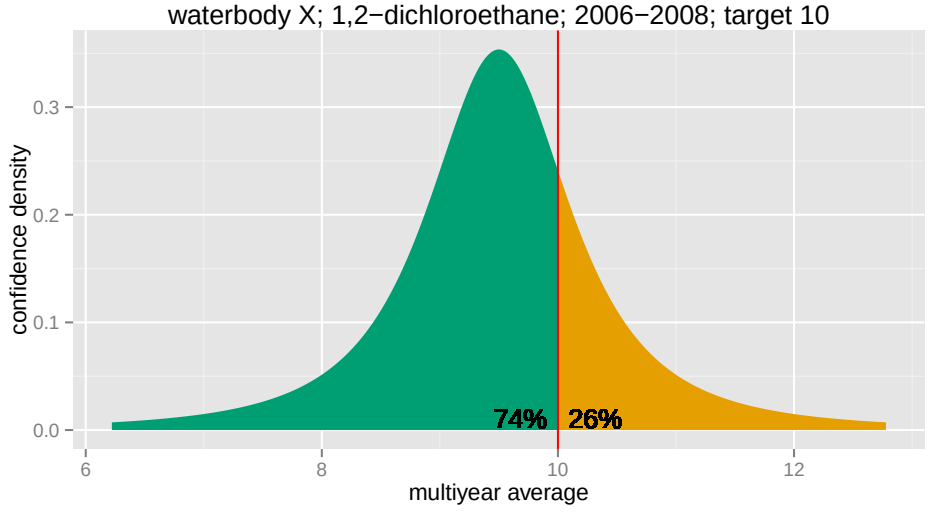
In this section, the results will be presented that have been produced by running the tool on the DCA data set. This data set has been shipped with the **confidence**-package and is given in the table below:

OBJECTID	PAR	color	char	comp	sampdev	YEAR	VALUE	TARGET	UNIT
waterbody X	1,2-dichloroethane	green/orange		water		2006	9.00	10	ug/l
waterbody X	1,2-dichloroethane	green/orange		water		2007	10.80	10	ug/l
waterbody X	1,2-dichloroethane	green/orange		water		2008	8.70	10	ug/l

Applying the `conf`-function to these data results in an HTML-document with the following table:

OBJECTID	PAR	PERIOD	MYA	TARGET	PROB_lt	PROB_gt	q05	q95
waterbody X	1,2-dichloroethane	2006-2008	9.50	10	0.74	0.26	7.59	11.41

This table gives for each `OBJECTID`, `PAR`, and `PERIOD`, the multiyear average (`MYA`), the lower bound (`q05`) and upper bound (`q95`) of the confidence interval of `MYA`, and the probability that `MYA` is greater (`PROB_gt`) or less (`PROB_lt`) than `TARGET`. These data are also given in the figure below.



5 Sample data

The package ships with three sample datasets. The user may wish to analyse these data in order to get familiar with the tool. These datasets are:

metal : simulated (*in silico*) metal contents in two arbitrary rivers for the period 2011-2015;

DCA : data presented by Baggelaar *et al.* (2010) representing annual average 1,2-dichloroethane concentrations in a specific waterbody for the years 2006, 2007, and 2008;

EQR : data presented by Baggelaar *et al.* (2010) representing ecological quality ratio's for macrofauna in a specific waterbody for the years 2003, 2006, and 2009.

The following annotated code block illustrates how to load the `EQR` dataset:

```
> # load confidence package
> # Note: this has to be done only once, at the start of an R session
> library(confidence)
>
> # load ecological quality ratio's
> data(EQR)
>
> # print these data to the screen
> EQR

     OBJECTID PAR        color char  comp sampdev YEAR VALUE TARGET  UNIT
1 waterbody x EQR orange/green   NA water      NA 2003  0.10    0.6 NODIM
2 waterbody x EQR orange/green   NA water      NA 2006  0.24    0.6 NODIM
3 waterbody x EQR orange/green   NA water      NA 2009  0.19    0.6 NODIM
  transform
1     logit
2     logit
3     logit
```

The other sample data sets can be loaded in a similar way.

6 References

Baggelaar, P., O. van Tongeren, R. Knoben, and W. van Loon, 2010. Rapporteren van de betrouwbaarheid van KRW-beoordelingen (in Dutch, English translation: Reporting the accuracy of WFD-assessments). H_2O 16: 21–25