R as a statistical engine for a water quality trend analysis web-service

P. Rustomji, B. Henderson, K. Mills, Q. Bai and P. Fitch

CSIRO Land and Water
CSIRO Mathematics, Informatics & Statistics

CSIRO

Motivation

Improve water quality condition and trend reporting in Australia by:

- harvesting existing statistical methods for water quality trend analysis
- assessing compliance or progress towards targets and guidelines and
- presenting these in a robust, scientifically supported and web-accessible tool.

CSIRO

Why is this important?

- Multiple trend analysis methods applied by States/Territories (or consultants) but they are not broadly available or presented in ways that makes adoption and regular use easy.
- A need to provide more robust and routinely available picture of water quality conditions.
- Assist in directing future investment in land and water management.
- Build awareness of the challenges and complexities in linking management actions with identifiable and desired environmental response.

CSIRO

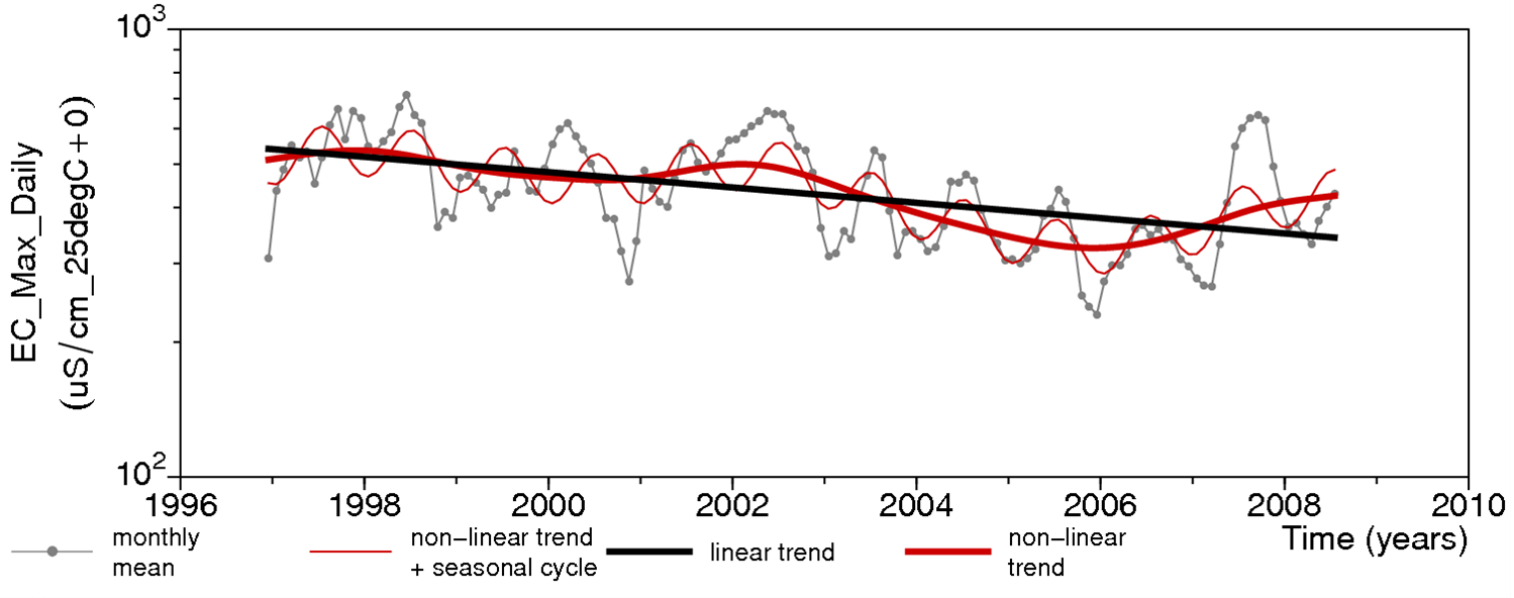
Trend Analysis Methods considered

- Seasonal Kendall's Tau slope estimate (Theil/Sen estimate)
  - non-parametric estimate of slope
  - related to Seasonal Kendall's Tau tests for monotonic change
  - flow adjustment possible but two step procedure
- Linear Regression & Generalised Additive Models
  - flexible framework for trend analysis that allows us to adjust for covariate effects
  - Linear time trend $\rightarrow$ linear regression
  - Nonlinear trend $\rightarrow$ GAMs (uses smoothing splines)

$$\underbrace{\log(EC_i)}_{\text{response}} = \beta_0 + \underbrace{\beta_1 \log(flow_i)}_{\text{flow effect}} + \underbrace{\beta_2 \sin(2\pi t_i) + \beta_3 \cos(2\pi t_i)}_{\text{seasonal cycle}} + \underbrace{\beta_4 t_i}_{\text{linear}} + \underbrace{s(t_{i,df})}_{\text{non linear}} + \epsilon_i$$

CSIRO

Example


EC_Max_Daily
(uS/cm_25degC+0)
10^3
10^2
1996
1998
2000
2002
2004
2006
2008
2010
Time (years)
monthly mean
non–linear trend + seasonal cycle
linear trend
non–linear trend

What We Did

Provide a web service that performs trend analyses of water quality data, using R as the statistical engine for analysis and visualisation

- Microsoft .NET is used to construct the web service
- Text files for data and parameter input
- Server calls R scripts using Rscript.exe
- Analysis is contained within Sweave files
  - Report template in LaTeX interspersed with R code
- Sweave'd files (*.tex) are compiled using pdflatex.exe to produce a pretty PDF report
- PDF graphics files converted to PNG format using Ghostscript

- User can download data and graphs.

CSIRO

Advantages:

1. Makes R available to a larger audience (no direct R programming experience required).
2. Reference R objects in report using Sweave **`Sexpr{}`**.
3. Include interpretative statements tailored to the statistical results (using the LaTeX **`ifthenelse`** package in conjunction with **`Sexpr{}`** statements) e.g.
   "The flow adjusted linear trend is -14.45 units change per unit time. The significance level (p-value) for this trend is <0.001 which means the likelihood of such a trend occurring by chance is less than 1 in 1000."
4. Harness typesetting capabilities of TeX to produce a high quality PDF report.
5. Access mapping capabilities of GoogleMaps.
6. Internet-wide accessibility.
7. Can be called by other applications (e.g. from Microsoft Excel).

CSIRO

```
C:/>rscript.exe %WQSAR_SCRIPT_PATH%\\mastertrend.r \path\to\output_dir\1234 --slave

---- contents of mastertrend.r ----
outpath <- commandArgs(TRUE)            #first and only argument is the path to the output directo
uniquenum <- basename(foo[1])           #get last part of directory name
par.file<- paste(uniquenum,"parameter_file.txt",sep="-") #parameter_file name
inputfile <- paste(uniquenum,"single_file.csv",sep="-")  #data_file name
sp <- Sys.getenv("WQSAR_SCRIPT_PATH")   #path to code

#read in input parameter file
f <- function(.file){source(.file,local=TRUE);as.list(environment())}
ipf <- f(par.file)

#now call the Sweave files that actually do stuff....
try(Sweave(paste(sp,"\\routines\\BEGIN_ROUTINE.Rnw",sep=""),
output=paste(uniquenum,"-BEGIN_ROUTINE.tex",sep=""),debug=FALSE,quiet=FALSE))

if(ipf$gam.method == TRUE){  #if GAM analysis was chosen...
 try(Sweave(paste(sp,"\\routines\\GAM_method.Rnw",sep=""),
 output=paste(uniquenum,"-GAM_method.tex",sep=""),debug=FALSE,quiet=FALSE)) }

if(ipf$lin.method == TRUE){  #if linear regression was chosen...etc
 try(Sweave(paste(sp,"\\routines\\LINEAR_method.Rnw",sep=""),
 output=paste(uniquenum,"-LINEAR_method.tex",sep=""),debug=FALSE,quiet=FALSE)) }
---- end mastertrend.r ----

::NOW MERGE OUTPUT FILES READY FOR PDFLATEX COMPILATION
C:/>copy /Y latex-preamble1.tex /A + 1234-BEGIN_ROUTINE.tex /A + 1234-data_summary.tex /A + 1234-1
1234-GAM_method.tex /A + 1234-NONPAR_method.tex /A + 1234-END_ROUTINE.tex /A %uniquenum%-%fileend%

:: then compile...
C:/>pdflatex.exe --quiet --job-name=%uniquenum%-%fileend% "%uniquenum%-%fileend%.tex"
::Voila!
```

CSIRO

WQSAR

Water Quality Statistical Analysis and Reporting Tool

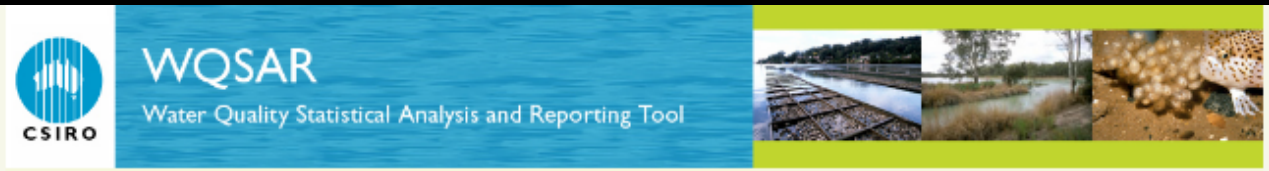
Home | Trend analysis | Multi-site analysis

Welcome to the Water Quality Statistical Analysis and Reporting (WQSAR) suite of tools.

You can use these tools to analyse water quality trends and generate reports on these for freshwater and estaurine systems all over Australia. Trends may also be compared across different monitoring stations in a multi-site analysis.

There are three different tools available. Two of these perform single site water quality trend analyses but use different client application types: a Web application and an add-in for Excel 2007. They retrieve, inspect and analyse the data and produce a statistical analysis report.

The multi-site analysis Web application takes completed analyses of station data and compares them, also producing a report.



Welcome to the Water Quality Statistical Analysis and Reporting Tool (WQSAR).
You can use this tool to track water quality trends and generate reports on these for freshwater and estuarine systems all over Australia; as well as comparing results from water quality multiple monitoring stations (multi-site analysis).

Please read how to use this tool before you start.

Please select a service (source of water quality data), either:

- from the drop-down list below
- or by entering your own service URL.

(Learn more about selecting or setting up a service).

Choose a service from the drop-down list:

OR ☐ **Enter own service**

NEXT

WQSAR

Water Quality Statistical Analysis and Reporting Tool

Select station and operating mode

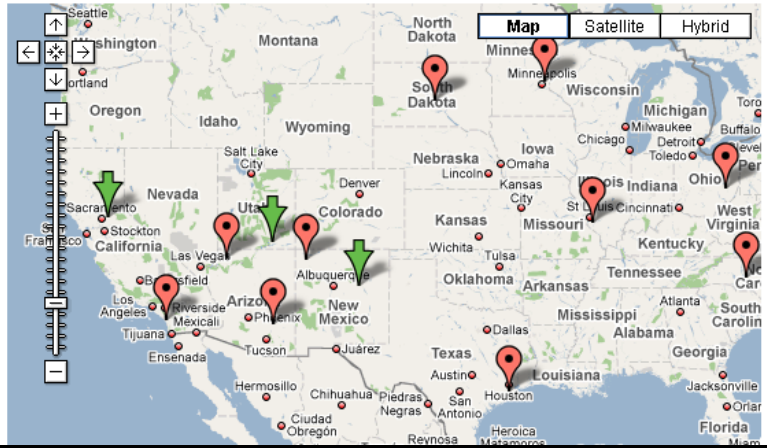
BACK NEXT

Select station from known identifying information

(Learn more about selecting a station).

Select station ID: ODM:9315000 [GREEN RIVER AT GREEN RIVER]

OR zoom in on the map below to select station.



WQSAR

Water Quality Statistical Analysis and Reporting Tool

Start over (Home) | HELP | Glossary

Home | Choose a Station | Choose Data | Inspect Data | Final Data Analysis | Get Results

Set up Water Quality Data

(Learn more about choosing data).

Choose a flow (if available) and water quality variable.

Flow variables (choose one): Discharge

Water quality variables (choose one): Solids, total

Date Measured From: 1/01/1941 12:00:00 AM

Date should be in Australian format, with optional time (dd/mm/yyyy hh:mm:ss AM/PM with this spacing). For example: 22/07/2008 9:11:00 AM. Or use the Date/Time chooser."

1/01/1941 12:00 AM

Date Measured To: 31/12/1984 12:00:00 AM

Date should be in Australian format, with optional time (dd/mm/yyyy hh:mm:ss AM/PM with this spacing). For example: 22/07/2008 9:11:00 AM. Or use the Date/Time chooser."

31/12/1984 12:00 AM

Description: Solids, total was measured in milligrams per liter. There were 16071 observations.

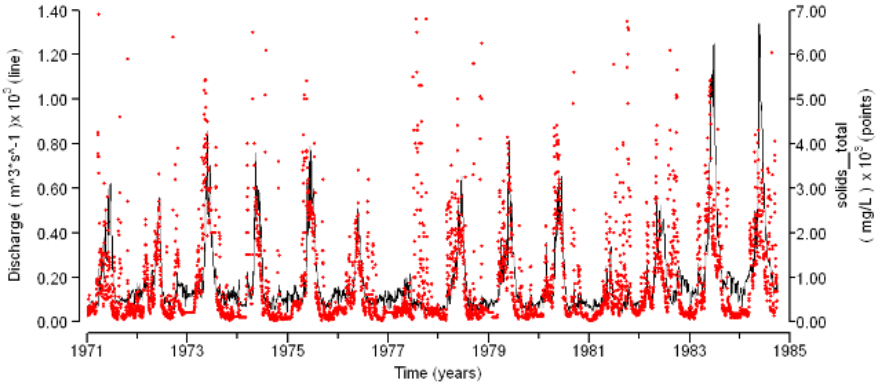
BACK NEXT

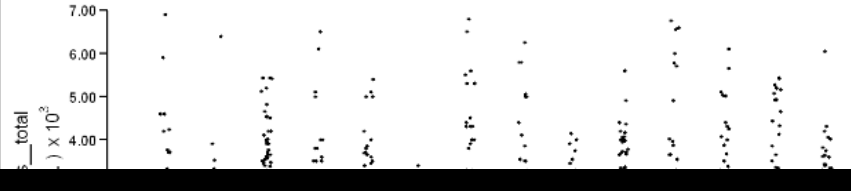
View and Save Graphs and their Associated Data

BACK NEXT

The data exploration graphs below will help you to assess the suitability of raw data for analysis. Links to download the graphs and data/parameter files are available beneath the graphs.

(Learn more about inspecting the data).





WQSAR
Water Quality Statistical Analysis and Reporting Tool

CSIRO

Choose methods and transformations

(Learn more about setting up the analysis parameters or read a report discussing the uses of the trend methods below).
NOTE that data run through analysis may be visible to other users.

Trend methods - choose one or more:

- ☑ Linear regression
- ☑ Seasonal Kendall's Tau / Mann-Kendall Test
- ☑ Generalised Additive Model

Description of the methods:

A parametric linear regression model is used to estimate the linear trend, after allowing for seasonal effects and possibly the effect of flow where available.

Seasonal Kendall's Tau is a nonparametric method for testing for monotonic trend based on the Kendall rank correlation. Seasonal Kendall's Tau is an extension that accounts for seasonal effects.

GAM is a semi-parametric regression model used to estimate a flexible non-linear trend, after allowing for seasonal effects and possibly the effect of other factors.

Select Data Transformation:

Water quality variable transform: None

☑ Replace zeros in water quality attribute data column with no-data values.

(Generalised Additive Model Only) Select Flexibility in Non-Linear Trend:

Flexibility for non-linear time trend: Medium

WQSAR
Water Quality Statistical Analysis and Reporting Tool

Start over (Home) | HELP | Glossary

Home | Choose a Station | Choose Data | Inspect Data | Final Data Analysis | Get Results

Get Report and Associated Data

You have reached the end of the data analysis and can download a report (and/or zipped contents) of the analysis.

(Learn more about understanding the report).

Download report: 1693-wqsar_report.pdf

Download one page summary: 1693-wqsarreport-onepage.pdf

Download everything (zip file of report, graphs and data): 1693.zip

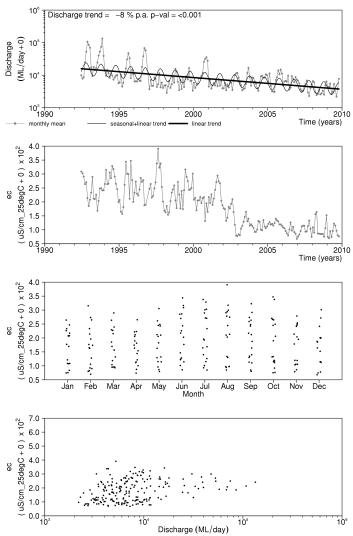
BACK

Water Quality Trend Analysis for Station 123456 Euston at Euston

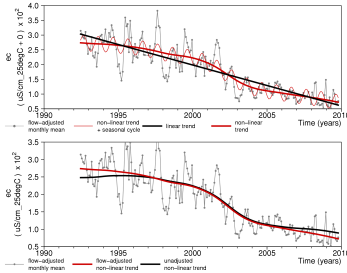
start date	1992-06-06
end date	2009-10-21
duration (days)	6012
catchment area (km^2)	
longitude (degrees East)	
latitude (degrees South)	
operating agency	
water quality variable	electrical conductivity
units of electrical conductivity	uS.cm.25degC
% days with water quality data	100
units of flow	MEGALITRES.PER.DAY
% days with flow data	100

Table 1: Summary of station details.

Monthly Mean Data Plots

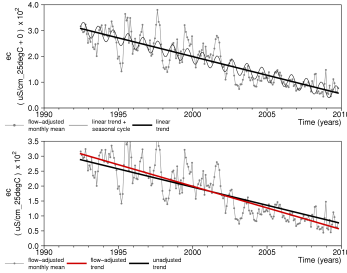


GAM Trend Analysis



The flow adjusted linear trend is -14.35 units change per unit time. The significance level (p-value) for this trend is <0.001 which means the likelihood of such a trend occurring by chance is approximately less than 1 in 1000. The significance level (p-value) of the non-linear trend is 0.03996. This value is less than 0.05 and indicates the non-linear trend is likely to be adding additional important information beyond that captured by the linear trend alone.

Linear Trend Analysis



The flow adjusted linear trend is -14.45 units change per unit time. The significance level (p-value) for this trend is <0.001 which means the likelihood of such a trend occurring by chance is less than 1 in 1000.

Non-Parametric Trend Analysis

The flow adjusted linear trend from the seasonal Kendall slope analysis is -14.48 units change per unit time. A 95% confidence interval for this trend is [-15.18,-13.84]. As this interval does not intersect 0 the trend is statistically significant.

	GAM	GAM_noflow	LINEAR	LINEAR_noflow	NONPAR	NONPAR_noflow
units change per unit time	-14.35	-12.26	-14.45	-12.16	-14.48	-12.43
S.E. (change per unit time)	1.006	0.9328	1.2	1.237		
Lower 95% C.I. (trend)	-16.36	-14.13	-16.85	-14.63	-15.18	-12.94
Upper 95% C.I. (trend)	-12.34	-10.39	-12.05	-9.688	-13.84	-11.84
Significance of linear trend (p-val)	<0.001	<0.001	<0.001	<0.001		
Significance of non-linear trend (p-val)	0.03996	0.003127				
Residual standard error (transformed scale)	35.37	37.07	38.68	42.04		
Autocorrelation	0.5693	0.5383	0.6472	0.6486		
R-squared (%)	79.31	77.16	74.63	69.88		
Shapiro-Wilk test of non-normality (p-val)	0.2171	0.03645	0.1947	0.008045		
Seasonal peak (months from 01-JAN)	8.329	7.575	8.39	7.541		
Number of outliers deleted (months)	0	0	0	0		
Notes	Model with auto-correlated errors fitted	Model with auto-correlated errors fitted				

Consistency between trend methods increases support for the trend identified. Where there are differences this indicates that other features may be important. In particular check whether the non-linearity in the trend appears to be important. Differences between the trends with and without flow indicate that changes in flow account for some of the change in the water quality parameter of interest.

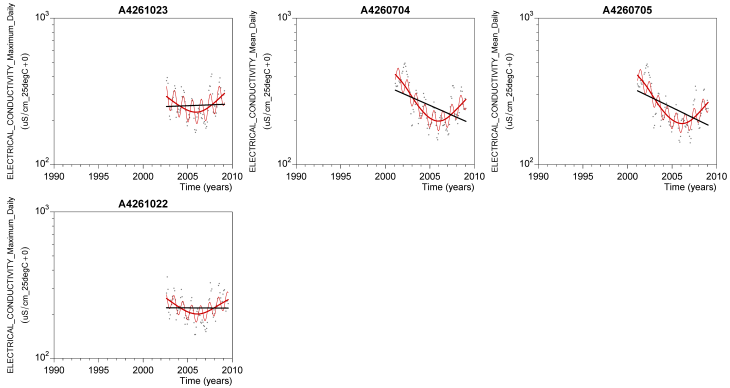


Figure 3: Time series plots of water quality variable and fitted trends. Grey dots show monthly mean observations. Thick black lines indicated fitted linear trends. Thin red line indicates fitted non-linear trend + seasonal cycle; thick red line shows non-linear trend.

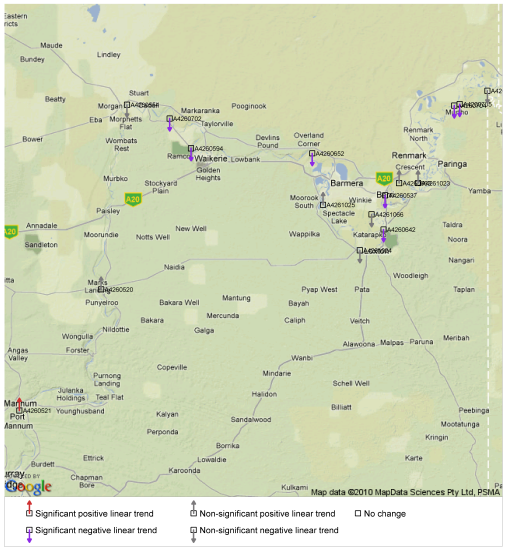
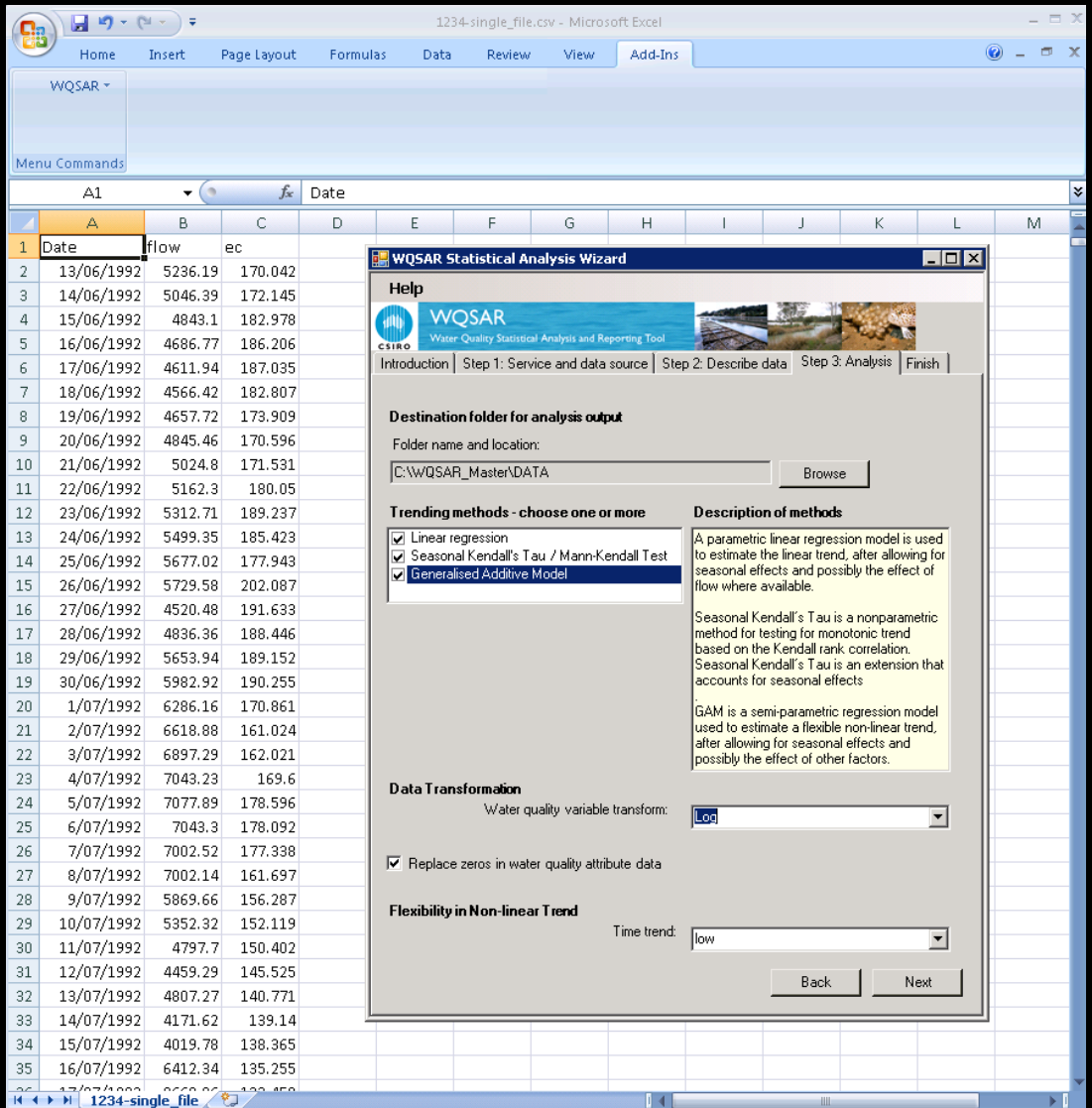


Figure 4: Map of linear water quality trends from GAM analysis (without flow variable)


1234-single_file.csv - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-Ins
WQSAR
Menu Commands
A1
Date
Date
flow
ec
13/06/1992 5236.19 170.042
14/06/1992 5046.39 172.145
15/06/1992 4843.1 182.978
16/06/1992 4686.77 186.206
17/06/1992 4611.94 187.035
18/06/1992 4566.42 182.807
19/06/1992 4657.72 173.909
20/06/1992 4845.46 170.596
21/06/1992 5024.8 171.531
22/06/1992 5162.3 180.05
23/06/1992 5312.71 189.237
24/06/1992 5499.35 185.423
25/06/1992 5677.02 177.943
26/06/1992 5729.58 202.087
27/06/1992 4520.48 191.633
28/06/1992 4836.36 188.446
29/06/1992 5653.94 189.152
30/06/1992 5982.92 190.255
1/07/1992 6286.16 170.861
2/07/1992 6618.88 161.024
3/07/1992 6897.29 162.021
4/07/1992 7043.23 169.6
5/07/1992 7077.89 178.596
6/07/1992 7043.3 178.092
7/07/1992 7002.52 177.338
8/07/1992 7002.14 161.697
9/07/1992 5869.66 156.287
10/07/1992 5352.32 152.119
11/07/1992 4797.7 150.402
12/07/1992 4459.29 145.525
13/07/1992 4807.27 140.771
14/07/1992 4171.62 139.14
15/07/1992 4019.78 138.365
16/07/1992 6412.34 135.255
1234-single_file
WQSAR Statistical Analysis Wizard
Help
CSIRO
WQSAR
Water Quality Statistical Analysis and Reporting Tool
Introduction
Step 1: Service and data source
Step 2: Describe data
Step 3: Analysis
Finish
Destination folder for analysis output
Folder name and location:
C:\WQSAR_Master\DATA
Browse
Trending methods - choose one or more
Linear regression
Seasonal Kendall's Tau / Mann-Kendall Test
Generalised Additive Model
Description of methods
A parametric linear regression model is used to estimate the linear trend, after allowing for seasonal effects and possibly the effect of flow where available.
Seasonal Kendall's Tau is a nonparametric method for testing for monotonic trend based on the Kendall rank correlation. Seasonal Kendall's Tau is an extension that accounts for seasonal effects
.
GAM is a semi-parametric regression model used to estimate a flexible non-linear trend, after allowing for seasonal effects and possibly the effect of other factors.
Data Transformation
Water quality variable transform:
Log
Replace zeros in water quality attribute data
Flexibility in Non-linear Trend
Time trend:
low
Back
Next

1234-single_file.csv - Microsoft Excel

Picture Tools

Home | Insert | Page Layout | Formulas | Data | Review | View | Add-Ins | Format

WQSAR

Menu Commands

Picture 1

	A	B	C
1	Date	flow	ec
2	13/06/1992	5236.19	170.042
3	14/06/1992	5046.39	172.145
4	15/06/1992	4843.1	182.978
5	16/06/1992	4686.77	186.206
6	17/06/1992	4611.94	187.035
7	18/06/1992	4566.42	182.807
8	19/06/1992	4657.72	173.909
9	20/06/1992	4845.46	170.596
10	21/06/1992	5024.8	171.531
11	22/06/1992	5162.3	180.05
12	23/06/1992	5312.71	189.237
13	24/06/1992	5499.35	185.423
14	25/06/1992	5677.02	177.943
15	26/06/1992	5729.58	202.087
16	27/06/1992	4520.48	191.633
17	28/06/1992	4836.36	188.446
18	29/06/1992	5653.94	189.152
19	30/06/1992	5982.92	190.255
20	1/07/1992	6286.16	170.861
21	2/07/1992	6618.88	161.024
22	3/07/1992	6897.29	162.021
23	4/07/1992	7043.23	169.6
24	5/07/1992	7077.89	178.596
25	6/07/1992	7043.3	178.092
26	7/07/1992	7002.52	177.338
27	8/07/1992	7002.14	161.697
28	9/07/1992	5869.66	156.287
29	10/07/1992	5352.32	152.119
30	11/07/1992	4797.7	150.402
31	12/07/1992	4459.29	145.525
32	13/07/1992	4807.27	140.771
33	14/07/1992	4171.62	139.14
34	15/07/1992	4019.78	138.365
35	16/07/1992	6412.34	135.255
36	17/07/1992	8669.06	132.458

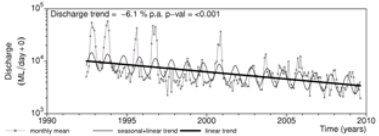
CSIRO Water Quality Statistical Analysis and Reporting Tool

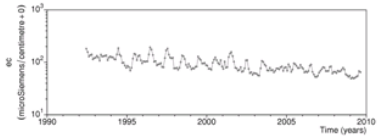
Water Quality Trend Analysis for Station Torrumbarry
NULL

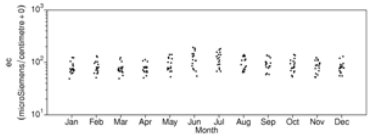
start date	1992-06-13
end date	2009-08-14
duration (days)	6176
catchment area (km^2)	
longitude (degrees East)	not specified
latitude (degrees South)	not specified
operating agency	not specified
water quality variable	electrical conductivity
units of electrical conductivity	microSiemens.centimetre
% days with water quality data	100
units of flow	ML.day
% days with flow data	100

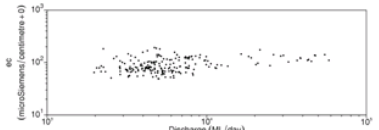
Table 1: Summary of station details.

Monthly Mean Data Plots

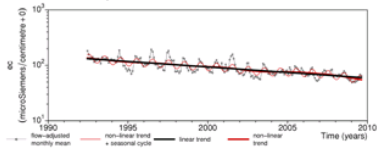


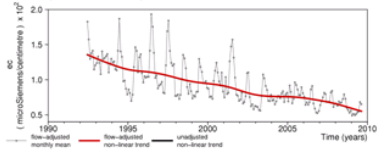






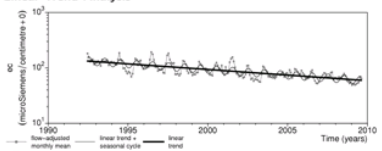
GAM Trend Analysis

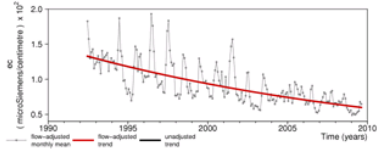




The flow adjusted linear trend is -4.517 percent change per annum. The significance level (p-value) for this trend is <0.001 which means the likelihood of such a trend occurring by chance is approximately less than 1 in 1000. The significance level (p-value) of the non-linear trend is 0.1352. This value is greater than 0.05 and indicates the non-linear trend is not likely to be adding additional important information beyond that captured by the linear trend alone.

Linear Trend Analysis





The flow adjusted linear trend is -4.548 percent change per annum. The significance level (p-value) for this trend is <0.001 which means the likelihood of such a trend occurring by chance is less than 1 in 1000.

Non-Parametric Trend Analysis

The flow adjusted linear trend from the seasonal Kendall slope analysis is -4.528

Acknowledgements

CSIRO's Water for a Healthy Country Flagship, Australian Government's Caring for our Country program, the Bureau of Meteorology and the Northern Australian Sustainable Yields project.

Plus lots of R and LaTeX packages . . .

CSIRO Land and Water
CSIRO Mathematics, Informatics & Statistics

Paul Rustomji

Phone: +61 2 9710 6915
Email: paul.rustomji@csiro.au
Web: wron.net.au/WebApps/WQSARPortal/Home.aspx

www.csiro.au

Thank you

Contact Us
Phone: 1300 363 400 or +61 3 9545 2176
Email: enquiries@csiro.au Web: www.csiro.au

CSIRO