USER GUIDE:

EDDY

A PROGRAM FOR GROUNDWATER AND SURFACE WATER QUALITY TREND ANALYSES

Author
Margot Doucet

October 2016

USER GUIDE

EDDY: A PROGRAM FOR GROUNDWATER AND SURFACE WATER QUALITY TREND ANALYSES

Author
Margot Doucet
doucetmargot@hotmail.com

Contributors
B. Marc Adams (BGC Engineering Inc.)
Dr. Sharon Blackmore (BGC Engineering Inc.)
Dr. Gabriele Chiogna (Technische Universität München)

October 2016

Notice

This program has been developed in order to serve as a tool to assist scientists in decision-making and analysis. It is by no means intended as a substitute for human judgement, common sense, due diligence or as a check on the quality of data collected.

Executive Summary

The software EDDY has been developed with the aim to serve as a tool for scientists in assessing trends in groundwater and surface water quality parameters. Particular challenges face decision-makers when assessing environmental datasets for trends. While a large number of statistical trend analysis methods for environmental data are available, comprehensive guidance and software for best practices in computing trend analyses are not readily available. Further, data are rarely perfectly fitted to assumed distribution models, may be collected at sporadic or discontinuous intervals, often contain non-detects and can consist of underlying seasonal as well as autocorrelative patterns. This project was carried out with the goal of helping to address this documentation and software gap by assembling relevant documented literature and applying best practices in a prepared software package.

Documented methods for trend analyses on environmental data were first researched. The methods of maximum likelihood estimation (MLE) linear regression (using normal, lognormal and gamma distributions) as well as variations of the Mann-Kendall trend test (implemented for seasonal data as well as corrected for autocorrelated data) in combination with a determination of Theil-Sen and Akritas-Theil-Sen slopes were then chosen for implementation in a developed software program, as these are common methods which are capable of processing non-detect values common in environmental datasets.

The resulting software is able to read and process data sets containing non-detect values and "greater-than" values, as flagged in a dataset by the symbols "<" and ">". Linear regression, Mann-Kendall analyses as well as a determination of Akritas-Theil-Sen slope are then carried out, without substituting censored values and without resulting in the loss of data.

Table of Contents

Executive Summary III
Acknowledgements V
1 Introduction 1
2 Data Input Requirements 2
2.1 Minimum Data Requirements 2
3 Program Instructions 4
3.1 Installation 4
3.2 Initial Interface 4
3.3 Defining Fields 4
3.4 Analysis Setup 5
3.5 Output Interface: Linear Regression 6
3.6 Output Interface: Mann-Kendall Test 8
3.7 Output Interface: Seasonal Mann-Kendall Test 10
Bibliography 12
APPENDIX A - TECHNICAL BACKGROUND: 13
Nomenclature 14
Abbreviations 14
Mathematical Notations 14
1 Introduction 16
2 Literature Review 17
2.1 Censored Data: Handling Non-Detects 17
2.1.1 Substitution Methods 17
2.1.2 Maximum Likelihood Estimation (MLE) 18
2.2 Environmental Trend Analyses 18
2.2.1 Linear Regression 19
2.2.2 Mann-Kendall 21
2.2.3 Theil-Sen & Akritas-Theil-Sen Slope Estimation 24
3 Summary 26
Bibliography 27

Acknowledgements

This study project was carried out in partial fulfilment of the requirements of the Environmental Engineering MSc. Programme of the Technical University of Munich (TUM, Germany). Sincerest gratitude and thanks is expressed to B. Marc Adams and Dr. Sharon Blackmore of BGC Engineering as well as Dr. Gabriele Chiogna of the Technical University of Munich for their contributions, support, feedback and patience throughout the course of this project. Special thanks also to Paul Doucet, for his support as well as guidance in the software development and distribution phases.

Thanks.

1 Introduction

Increasingly, parties responsible for carrying out environmental monitoring as well as environmental regulators alike are becoming interested in changes in environmental quality parameters over time, rather than simply static parameter values in relation to a prescribed guideline. These trend analyses can be carried out in a number of ways (regression, Mann-Kendall, varying spatial and temporal averaging approaches, etc.). In addition, environmental datasets are particularly rarely well-behaved; with non-detects, outliers, skewness and/or seasonal patterns often present. Documentation and software for best practices for conducting environmental trend analyses are not, however, readily available. This project was thus carried out with the goal of helping to address this documentation and software gap by assembling relevant documented literature and applying current best practices in a prepared software package.

While there is a range of environmental data analysis software applications available, many tend to be designed for large-scale and long-term projects and involve high initial costs. Other free or low-cost programs are often inflexible. In addition, methods implemented in available software often do not reflect modern best trend analysis practices, as trend analysis capabilities are often included as an afterthought in software packages which focus rather on data management and/or on determining summary statistics. It was thus aimed that the software researched and developed in the scope of this project could provide a simple, rigorous and defensible tool for trend analyses on groundwater and surface water datasets.

2 Data Input Requirements

The software developed has been set up to read databases which are stored as Excel files (.xls or xlsx). Data entries should be stored as rows, where each column represents a data field. Though the user is free to choose the exact naming convention of the columns, the fields included should generally be, at a minimum:

- Sampling date
- Well or station ID
- Contaminant of concern concentration value (where non-detect values or greater than values are recognized by the symbols "<" and ">" next to a concentration value, where applicable)

Other considerations are:

- Group ID, when analyses are to be performed on groups of stations, rather than individual stations.

In parallel with the analysis data requirements, extreme importance should be stressed during data collection and database setup on consistency in nomenclature and formatting in order for analyses to be carried out smoothly. Stations ID's should be referred to exactly consistently (Station-1 or station 01 versus Station_01, for example) and **date formats must be consistent**.

2.1 Minimum Data Requirements

Linear Regression: USEPA (2013) has recommended that at least 10 observations should be available for hypotheses testing approaches. USEPA has further published in their *Unified Guidance for Statistical Analysis of Groundwater Monitoring at RCRA Facilities* (2009) that for linear regressions and for Mann-Kendall analyses, at least 8-10 observations should be available.

Mann-Kendall type trend tests can be carried out on data sets with as few as three observations. However, significant (at significance of 0.1) trends can only be deduced when at least four data points are available. When ties are present in a dataset however (equivalent concentration or sampling date values), n must be greater than 10.

For **seasonal Mann-Kendall** analyses, Helsel and Hirsch (2002) have stated that when the product of the number of seasons and the number of years of observations is at least 25, the

distribution of the S statistic can be well approximated by a normal distribution. Guidance regarding seasonal Mann-Kendall tests on data containing less than 25 sampling events and which may contain data ties is not readily available.

Mann-Kendall, Correction for autocorrelated data: since the correction is applied to the variance value of the S statistic (refer to Appendix A), the correction is only applicable when the S statistic approximately follows a normal distribution ($n>10$, or $n\geq25$ for a seasonal Mann-Kendall test)

3 Program Instructions

3.1 Installation

To run, run the file "Install_Eddy.exe" first (this only needs to be done once per pc) and then download and run the "Eddy.exe" file to launch the program. The program may be slow to start-up and to get from the first window to the second upon the first launch (less than ~20 seconds, but this can feel like a long time). Once past the first interface window, the user can then quickly switch between analyses. It is however advised to close one result window before opening a new one. If either of the programs fail to launch, try running as an administrator (right click -> "Run as administrator").

3.2 Initial Interface

In the initial interface (Figure 3.1), the user is prompted to select a file in which the data is contained. The software has been set up to read from .xls- or .xlsx-based databases (Microsoft Excel). Files containing multiple spreadsheets are supported, as the user is then prompted to specify the sheet which contains the data to be analysed.

Figure 3.1 Initial Interface.

3.3 Defining Fields

Once data have been loaded from the specified spreadsheet (may take a few seconds for large databases), the user is then taken to the next interface (Figure 3.2) in order to define the fields of the database which are required for the trend analyses. A sample of the data is displayed in the interface in order to assist with defining the fields. Here, the user specifies the number of header rows to be read and defines the required fields "Station ID's", "Sampling Date", "Date Format" as well as the analysis parameter. Gaps or blank entries in a data set are supported. Blanks or text characters which are stored in the analysis parameter field (with, of course, the exception of the symbols '<' or '>') will be ignored. Further options (not required) include:

- Maximum length of sampling event: In the case that individual sampling events span more than one day (for example, a bi-annual sampling program which runs over the course of a week, twice per year), the user has the option here to enter a threshold for

which sampling events should be grouped together. This is only important in the case where groups of wells are to be analysed together, but some sampling events may have occurred a day or two apart. For example, for a biannual sampling program where samples from a site could be collected over a period of seven days, twice a year, the value here would be seven in order to ensure that samples collected only seven days apart are considered as ties in the time domain.

- Station Grouping: Here, the user can set groups of stations, for which analyses will also be carried out. These can be defined in a separate field within the database, or a maximum of four groups can be defined manually within the interface. The user can also take advantage of this feature to group together the same station which has been inconsistently labelled, for example "P01" and "P-01".

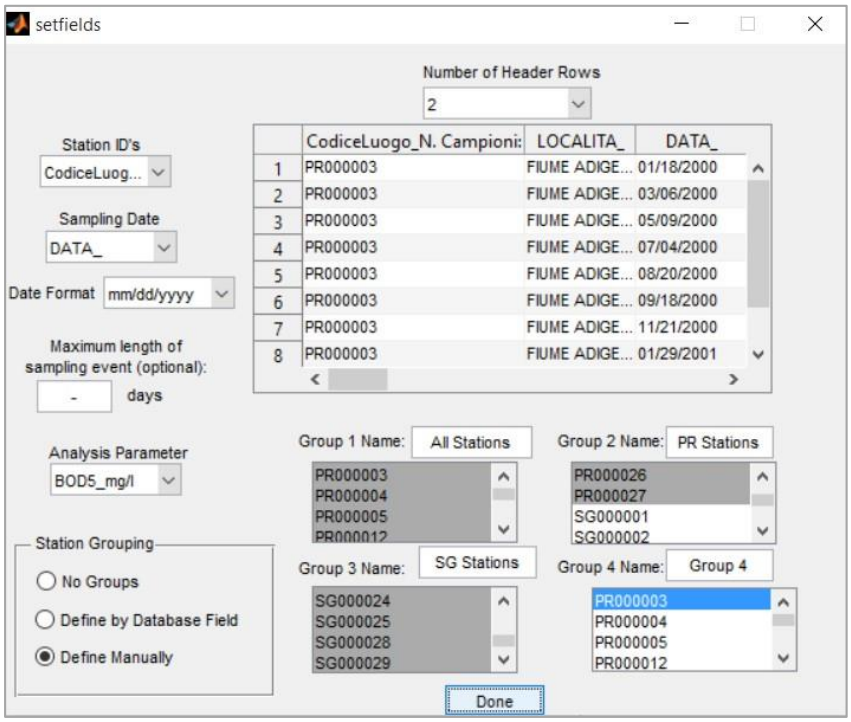


Figure 3.2 Interface for defining database fields.

3.4 Analysis Setup

In the next interface (Figure 3.3), the user is prompted to select whether an MLE linear regression or a Mann-Kendall test is to be carried out. In the case of a Mann-Kendall test, the user can then define whether the test should be carried out as a normal test or as a seasonal Mann-Kendall test. When a seasonal Mann-Kendall test is selected, the seasons must also be defined. Season options include: monthly, per season (Winter, Spring, Summer and Fall), two seasons per year (Winter/Spring and Summer/Fall) or up to twelve user-defined seasons. In the analysis setup

interface, the user must also specify on which stations, or station groups, the analyses are to be carried out and can set the desired significance level for the trend tests (default is 0.05).

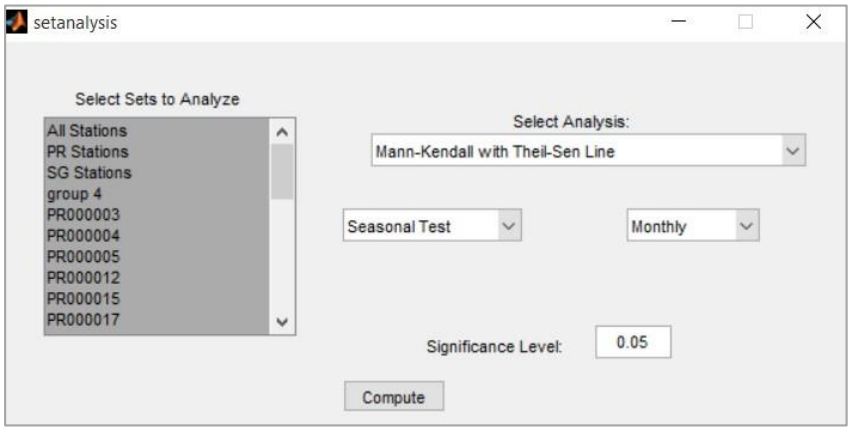


Figure 3.3 Analysis setup interface.

3.5 Output Interface: Linear Regression

When the user has selected a linear regression trend analysis and proceeded with the analysis, the linear regression output will display a graph with three lines, as in Figure 3.4. The three lines correspond to linear regressions on the dataset assuming a (1) normal, (2) lognormal and (3) gamma distribution of the data residuals. In the plot, detected data values, non-detect values (plotted at their detection limit) as well as the fitted MLE linear model are displayed.

The interface further displays the p-value of the corresponding test and, if the p-value is lower than the user-specified significance level, the slope of the trend as estimated by the model. Note that for the normal and gamma distributions, the units of the regression slope value are in (input units)/year, while the units for the lognormal distribution are in % change/year.

In order to view how well the residuals fit each of the specified distributions, the user can click on the "View Residual Probability Plot" control button and a new window will appear with a probability plot of the residuals (Figure 3.5). With the probability plots, the user can assess which of the distributions represents a better fit for the data model.

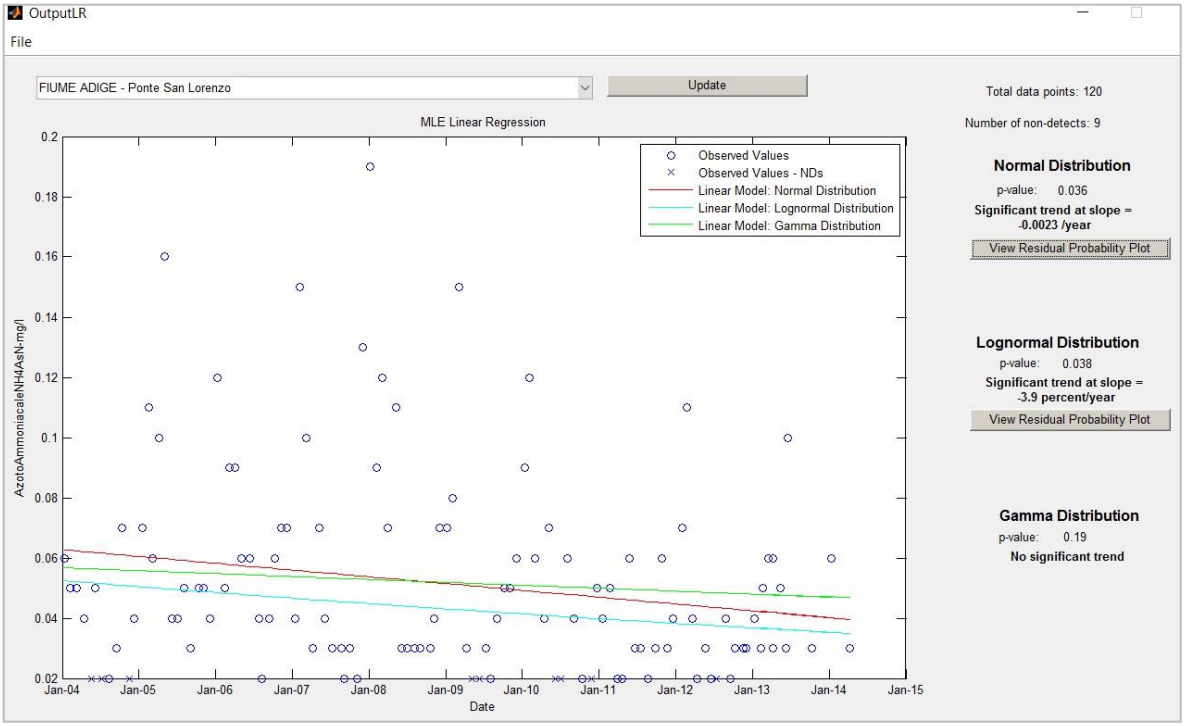


Figure 3.4 Linear regression output interface.

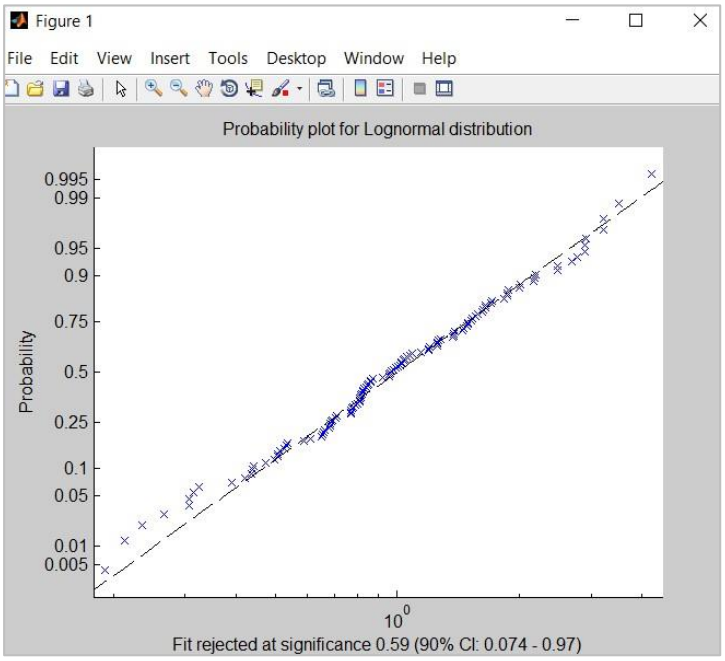


Figure 3.5 Probability plot for specified distribution (in this case: lognormal).

In the probability plot output, the user is provided, at the bottom of the graph, with a range of p-values resulting from a Chi-squared test which has tested the goodness of fit of the residuals to the respective model. The Chi-squared test has been carried out using the following number equal-probability bins:

- 3, for very small data sets ($n \leq 15$)
- round(n/5), for data sets where $15 < n < 50$
- 10, for $n \geq 50$.

The Chi-square test evaluates the null-hypothesis that the residuals follow the specified trend and rejects this hypothesis only at the specified significance. (i.e. for a typical desired significance of 0.05, it would be concluded that the given distribution does **not** fit the residuals when the outputted range is below 0.05. If the range is above 0.05, such as in Figure 3.5, it is not rejected that the given distribution fits the data residuals.)

A range is outputted in this interface as a result of the uncertainty associated with the non-detect or "greater than" observations. The Chi-square test is in fact carried out 1000 times, each time with random values between 0 and the detection limit (distributed according to the assumed distribution) replacing each of the non-detect values and values above the maximum detection limit replacing "greater than" values, where applicable. The outputted value represents the median significance of the goodness of fit test from all 1000 trials. A range representing the mid-90% interval of these 1000 p-values is also outputted.

Due to limitations of the Chi-square goodness-of-fit test for small sample sizes, it is recommended to evaluate the distribution fit using **both** the Chi-square test results and probability plots, especially when n<50.

3.6 Output Interface: Mann-Kendall Test

In the output interface for the Mann-Kendall trend test, a single graph is displayed which contains the observed values as well as the computed Akritas-Theil-Sen slope for the dataset. To the right of the graph, both the computed Akritas-Theil-Sen slope and a range of Theil-Sen slopes are displayed.

In computing the Theil-Sen slope, non-detects have been replaced by random values (uniform distribution) between 0 and the respective detection limits and "greater than" values have been replaced by random values between the max. detection limit and 2 x the max. detection limit. The resulting Theil-Sen line has then been computed 100 times. When censored values are present, a range is then output representing the mid-90% interval of these 100 trials.

It is noted that the use of the uniform distribution and 2 x the max. detection limit is an arbitrary approach. It has been included as a sort of sensitivity analysis and for comparison/research purposes. ***When censored values are present in a dataset, the Akritas-Theil-Sen slope should be reported, not the Theil-Sen slope.***

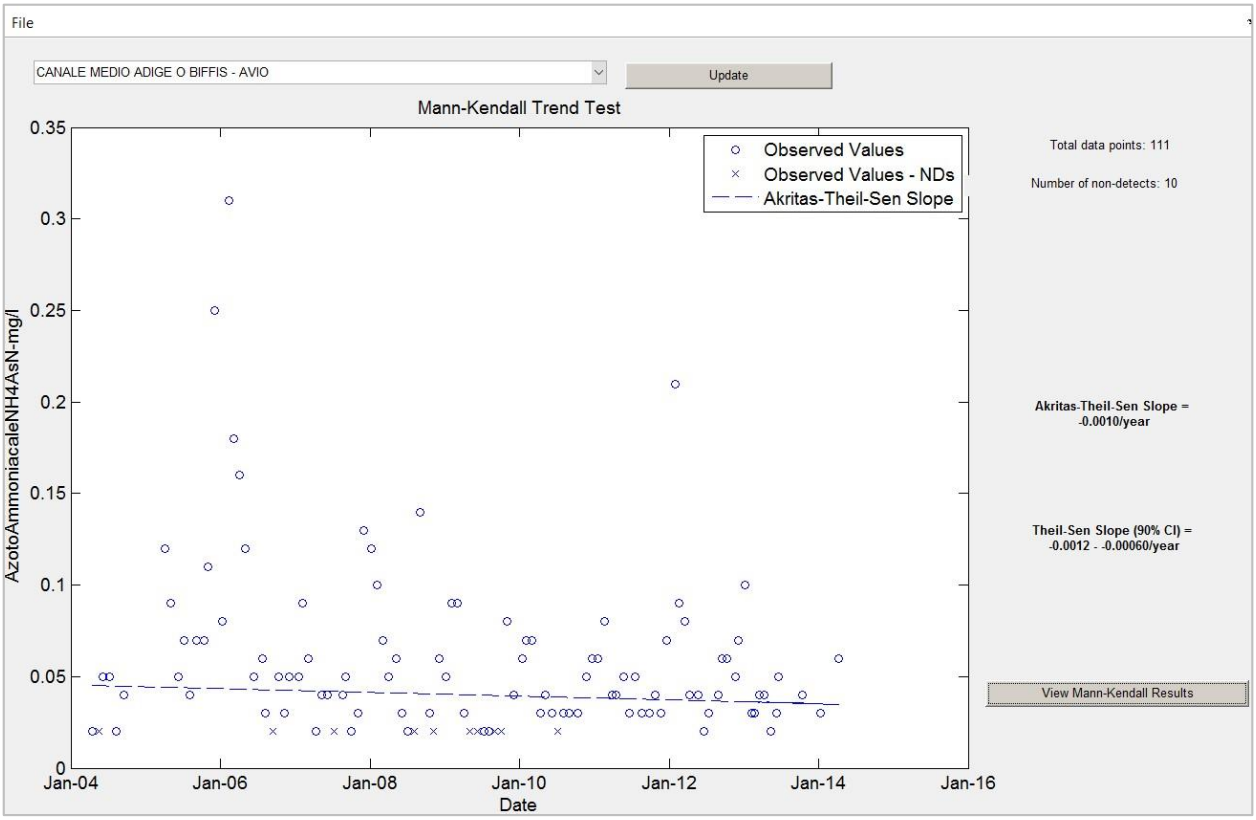


Figure 3.6 Mann-Kendall trend test output interface.

Through the button labelled "View Mann-Kendall Results", the user can view the results of the Mann-Kendall test as well as a plot of the lag-values versus the respective autocorrelations on the data ranks. The autocorrelation plot should help the user to assess whether the corrected or the uncorrected test results are most appropriate for the dataset. This displayed autocorrelation function is the function which has been used to compute the corrected Mann-Kendall trend test. When autocorrelation in the dataset is plausible and evidenced by the autocorrelation plot, the corrected Mann-Kendall result is more appropriate for the dataset.

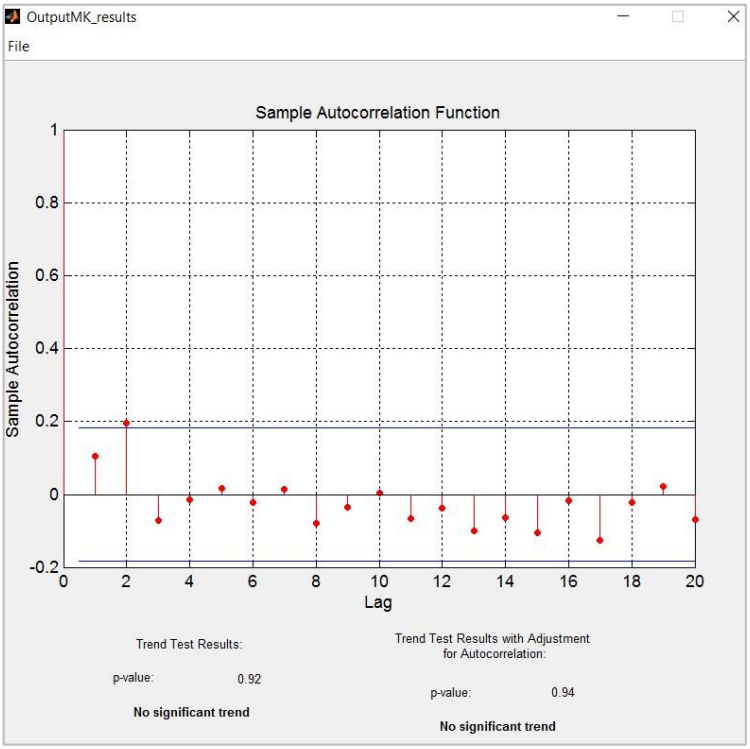


Figure 3.7 Mann-Kendall results interface as well as autocorrelation function of data ranks used to compute Mann-Kendall test corrected for autocorrelation.

3.7 Output Interface: Seasonal Mann-Kendall Test

The interface for the output of the seasonal Mann-Kendall test (Figure 3.8) is very similar to the interface for the Mann-Kendall test, except the results refer to a seasonal trend analysis and Theil-Sen/Akritas-Theil-Sen slope estimates have been computed for each of the respective seasons. In the autocorrelation function display (Figure 3.9), a separate plot is shown for each season, since the corrected variance values in the seasonal test are summed for each of the seasons. The autocorrelation functions displayed thus represent the functions which have been used to compute the corrected variance for each of the seasons.

In computing the Theil-Sen slope for each season, non-detects have been replaced by random values (uniform distribution) between 0 and the respective detection limits and “greater than” values have been replaced by random values between the max. detection limit and 2 x the max. detection limit. The resulting Theil-Sen line has been computed 100 times. When censored values are present, a range is then output representing the mid-90% interval of these 100 trials. Data pairs in a set which have been collected in the same season and year are excluded in computing the Theil-Sen and Akritas-Theil-Sen slopes and are considered ties in computing the Mann-Kendall test.

It is noted that the use of the uniform distribution and 2 x the max. detection limit is an arbitrary approach. It has been included as a sort of sensitivity analysis and for comparison/research purposes. ***When censored values are present in a dataset, the Akritas-Theil-Sen slope should be reported, not the Theil-Sen slope.***

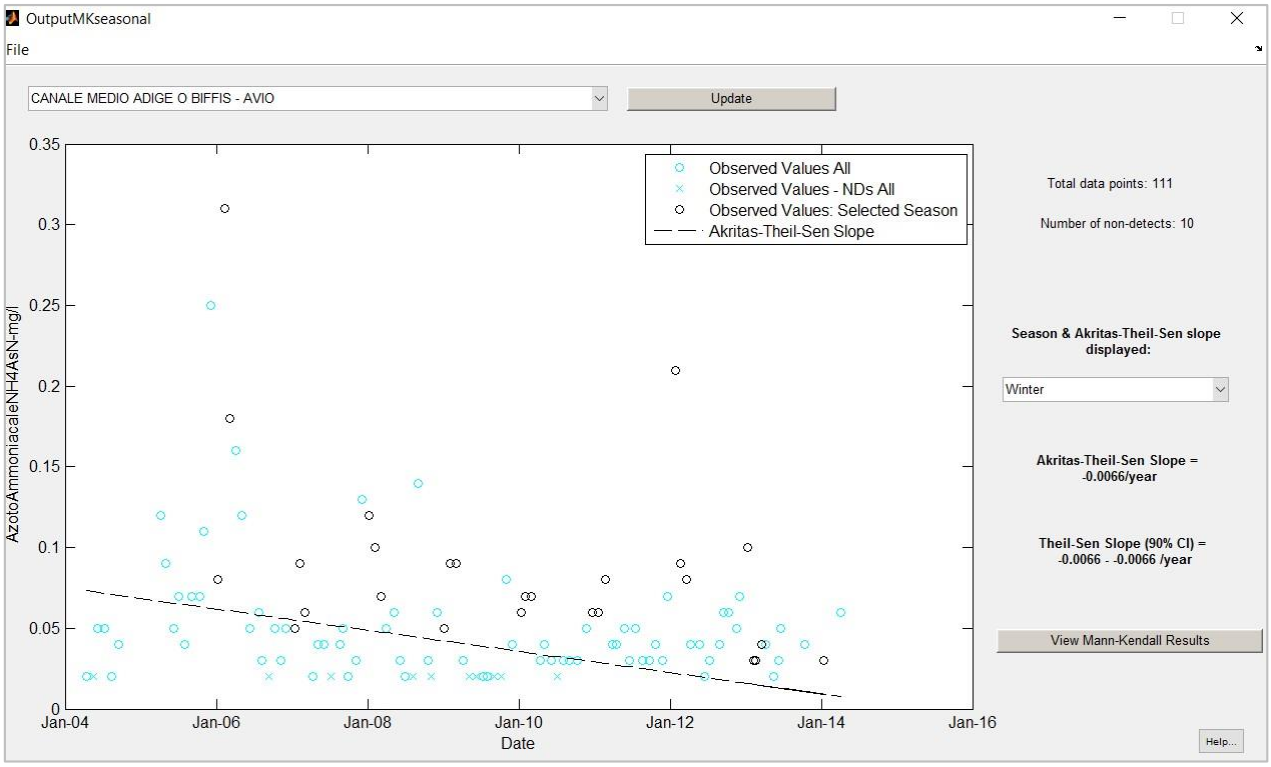


Figure 3.8 Seasonal Mann-Kendall trend test output interface.

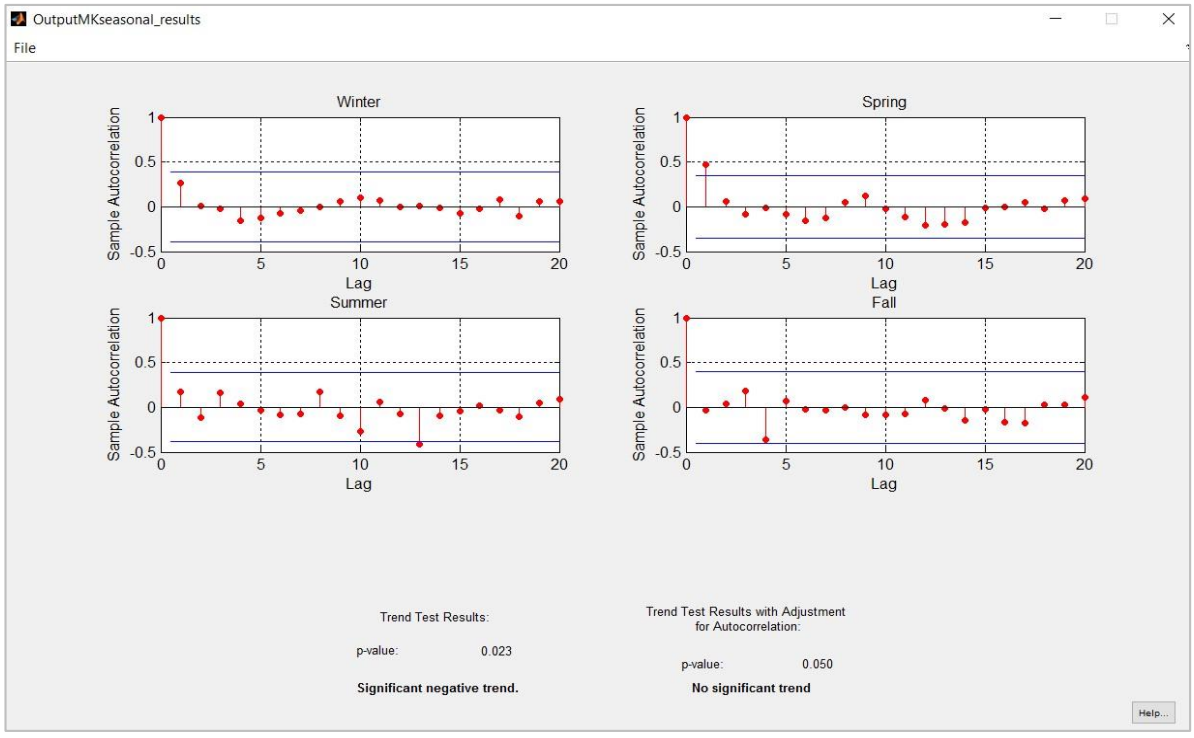


Figure 3.9 Autocorrelation function of data ranks used to compute corrected seasonal Mann-Kendall test.

Bibliography

Helsel, D. R., & Hirsch, R. M. (2002). *Statistical Methods in Water Resources.* Retrieved from Techniques of Water-Resources Investigations of the United States Geological Survey: http://water.usgs.gov/pubs/twri/twri4a3/

Jones, W. R., Spence, M. J., Bowman, A. W., Evers, L., & Molinari, D. A. (2014). A software tool for the spatiotemporal analysis and reporting of groundwater monitoring data. *Environmental Modelling & Software*, 242-249.

Ofungwu, J. (2014). *Statistical Applications for Environmental Analysis and Risk Assessment.* Hackettstown, NJ: Wiley.

USEPA. (2009). *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities - Unified Guidance.* United States Environmental Protection Agency. Retrieved 12 20, 2015, from http://www3.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/sitechar/gwstats/unified-guid.pdf

USEPA. (2013). *ProUCL Version 5.0.00 Technical Guide.* United States Environmental Protection Agency.

APPENDIX A - TECHNICAL BACKGROUND:

EDDY

A PROGRAM FOR GROUNDWATER AND SURFACE WATER QUALITY TREND ANALYSES

Author
Margot Doucet

Contributors
B. Marc Adams (BGC Engineering Inc.)
Dr. Sharon Blackmore (BGC Engineering Inc.)
Dr. Gabriele Chiogna (Technische Universität München)

October 2016

Nomenclature

Abbreviations

KM	Kaplan-Meier
MLE	Maximum Likelihood Estimation
ROS	Regression on Order Statistics
TUM	Technical University of Munich
USEPA	United States Environmental Protection Agency

Mathematical Notations

b_0, b_1	Parameters (intercept, slope) of linear regression
cov	Coefficient of variation
f, F	Normal probability density function and cumulative distribution function
G^2	Log-likelihood test statistic
k	Shape parameter of gamma distribution
L	Log-likelihood
m	Number of seasons in seasonal Mann-Kendall test
n,n_d,n_{nd},n_{md}	Number of observations (general, uncensored, left-censored and right-censored)
N_c, N_D	Number of concordant and discordant pairs in a dataset, respectively
$\frac{n}{n_S^*}$	Correction factor due to autocorrelation
p,P	Gamma probability density function and cumulative distribution function
Q	Theil-Sen slope

$R_{t(X)}$, $R_{t(Y)}$	Ranks of variables *x* and *y* at time *t*, respectively
S	Mann-Kendall test statistic
$S(d^2)$	Sum of squared differences
t	Time
t_i, u_i	Extent of data ties in the time domain and concentration domain, respectively
T_i	Pair-wise slope *i*
$Var^*[S]$	Variance of S corrected for autocorrelation
y	Analysis parameter, typically concentration
Z	Normalized Z statistic
$\rho_s(i)$	Autocorrelation function of observation ranks
θ	Scale parameter of gamma distribution
σ	Standard deviation
π	Probability of concentration being above greater than or equal to specified value

1 Introduction

The focus of this Project was to develop a rigorous and defensible tool to assist in conducting trend analyses on environmental datasets. In order to accomplish this, a literature review was first carried out in order to identify best practices in trend analysis. From this, methods were selected which were to be implemented in the developed software application. The software with the selected trend analysis capabilities has then been developed in Matlab R2013b and compiled for distribution as a .exe file.

2 Literature Review

2.1 Censored Data: Handling Non-Detects

Particular complexity in carrying out trend analyses on water quality datasets is attributed to the frequent presence of non-detect values in datasets. When the exact value of a data point is not known, but rather, information is known regarding some range within which the data point falls (i.e. less than the detection limit or, in some cases, greater than the reporting limit), the data is referred to as censored. More specifically, data points which are defined as being less than a given detection limit are referred to as left-censored. Right-censored values, or values with are only known to be above a certain reporting limit (> laboratory method limit), are less common in environmental datasets but also sometimes present. Additional complexity is added when multiple differing detection limits are present within a single dataset, which is often the case in large environmental monitoring programs which span several decades.

The approach with which these data points are dealt with should not be overlooked. Helsel (2012), for example, has written "the worst practice when dealing with censored observations is to exclude or delete them". Helsel (2012) further strongly discourages the use of simple substitution methods (0, detection limit, or half of the detection limit, for example) in trend analysis. Simple substitution methods can add false signals to the data that were not previously there and can obscure information that is present. This can result in inaccurate and unreproducible results when arbitrary substitution methods are employed. Furthermore, statistical trend detection methods which can be carried out without resorting to the substitution of censored values are well documented. These are further detailed in the following sections.

2.1.1 Substitution Methods

Previous United States Environmental Protection Agency (USEPA) statistical guidance (USEPA, 2004) had endorsed the use of substituting non-detects with half of the reporting limit when less than 15 % of the samples in a set consisted of non-detects. However, more recent USEPA publications (USEPA, 2013a) do not recommend the use of substitution methods, regardless of how many non-detects are in a sample set. In this USEPA (2013a) document, no particular method for dealing with non-detects is instead provided as a recommendation for trend analyses. For statistical analyses involving the determination of distribution parameters, USEPA (2013a) instead proposes the Kaplan–Meier estimation (KM) or regression on order statistics (ROS) in order to determine distribution parameters of a dataset, both methods of which cannot meaningfully be applied in trend analysis.

Other common substitution methods include replacing all non-detect values by the single highest, lowest or median detection limit or by the value of 0 (Helsel, 2012). The choice of which of these substitution methods to choose has remained subjective. Because of this, modern practices are tending toward avoiding trend analysis methods which require substituted values and instead employing more robust methods, which are capable of preserving censored values.

Substitution methods are, however, still common practise since not all analysis methods are suited for censored data (such as least-squares linear regression or Theil-Sen estimation) or, when analysis methods are suited for censored data, standard commercial software may not be coded to process them. In this case, Helsel (2012) recommends at least first re-censoring the dataset at the highest reporting limit – an important step when multiple censoring levels are present in a single set.

2.1.2 Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is a parametric approach to handling non-detects (Helsel, 2012). As a parametric approach, using MLE to handle non-detects requires that the data be fitted to a particular distribution. The data thus needs to be checked against a given distribution (normal, lognormal or gamma distribution, for example) before conclusions can be drawn from MLE results.

In MLE, information required consists of: (1) known data points above reporting limits, (2) the proportion of data below each reporting limit and (3) the mathematical formula for an assumed distribution. MLE then computes distribution parameters of (3) to best fit (1) and (2). This method is commonly employed in statistical tests when values of interest are the defining distribution and the distribution parameters of the data, rather than the data points themselves.

2.2 Environmental Trend Analyses

Trend analysis methods can be classified as either parametric or nonparametric. A method is described as parametric if it assumes that the data belong to a specified probability distribution (often normal or lognormal), which is characterized by defining parameters (e.g. mean and standard deviation) (Ofungwu, 2014). A nonparametric method is an approach which does not require data to belong to any particular distribution, typically relying on relative ranking of data points. Parametric methods in trend analysis include linear regression, while nonparametric methods include Mann-Kendall, Theil-Sen, Spearman's rho and binary logistic regression.

2.2.1 Linear Regression

One of the most widely used techniques in trend detection is linear regression. It is used to relate a response variable (e.g. concentration in mg/l) to one or several explanatory variables (such as time) through a linear model. In linear regression, the traditional least squares method (with substituted censored values) or the MLE method for censored data can be used to estimate the model parameters of slope and intercept. Chung (1990), as well as Thomson and Nelson (2003), have both compared substitution methods (various multiplications of the detection limit and half of the detection limit, respectively) with MLE for regression and found that MLE outperformed substitution methods in estimating model slopes when both were performed on sets containing censored data.

Ofungwu (2014) as well as the USEPA (2009) recommend that linear regression be carried out when at least 8 to 10 data values are present. In MLE, the procedure requires (1) a defining objective function which describes the agreement between the data and the model and (2) a model defined by parameters to be optimized by (1), which attempts to relate y (concentration) to t (time). In a linear model, (2) thus takes the form $y = b_o + b_1 t$, where b_0 and b_1 are the parameters to be optimized by the objective function (1) and y actually refers to the mean (or expected) value of y, which is conditional on t. The objective function is known as the log-likelihood function (L) and is defined by the assumed distribution of the dataset. In MLE, the parameters b_0 and b_1 are optimized such that the likelihood of making the observations, which make up the dataset, is maximized. This method is inherently suited for censored values, as likelihoods can also be determined for censored values with a given distribution.

Using the relationship $y_{expected} = b_o + b_1 t$ for the expected value of the distribution, the log-likelihood function to be maximized can be defined, for normally-distributed residuals containing censored data as:

$$L = \sum_{i=1}^{n_d} \log(f_{(y_i|b_o + b_1 t_i, \sigma)}) + \sum_{i=1}^{n_{nd}} \log(F_{(y_i|b_o + b_1 t_i, \sigma)}) + \sum_{i=1}^{n_{md}} \log(1 - F_{(y_i|b_o + b_1 t_i, \sigma)}) \quad (2.1)$$

Where the first term represents the log of the probability given a normal probability density function (f) of making n_d uncensored observations given a mean of $b_o + b_1 t_i$ and a standard deviation σ, while the second and third terms are the sum of the log of the probabilities based on the cumulative distribution function (F) of making n_{nd} left-censored observations, and n_{md} right-censored observations where the y_i's in this case represent the detection limits. The MLE linear

regression method finds the values of b_0, b_1 and σ which maximize the log-likelihood (L). For a lognormal distribution assumption, y values are first log-transformed.

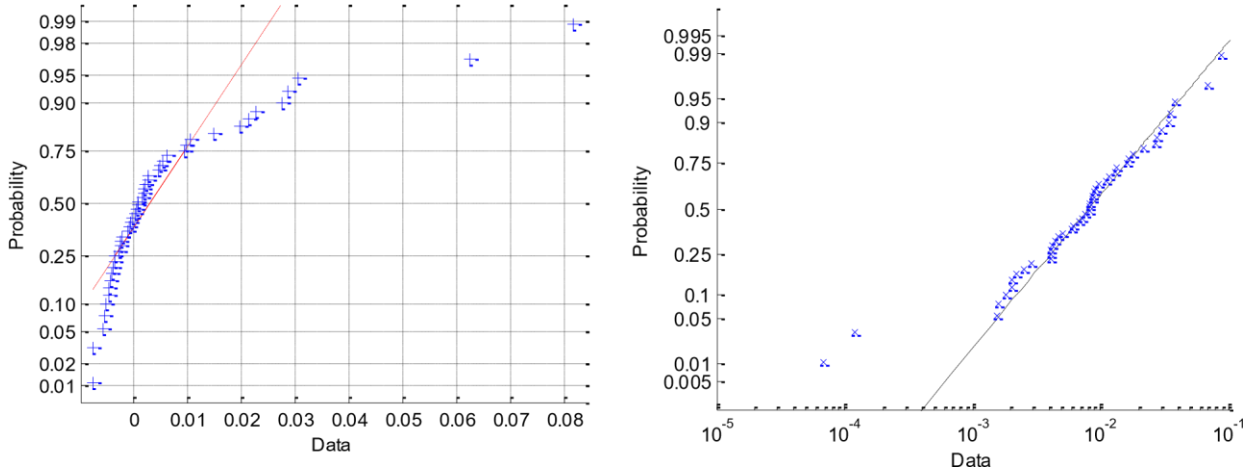
For regression using a gamma distribution assumption on the residuals, the relationship from the gamma distribution of mean = $b_o + b_1 t_i = k\theta$, where k and θ are the defining shape and scale parameters of the gamma distribution, can be used. Assuming a constant coefficient of variation (cov), the log-likelihood for a gamma distribution on residuals can be derived as:

$$L = \sum_{i=1}^{n_d} log\left(p\left(y_i \middle| \frac{1}{cov^2}, cov^2(bo + b_1 t_i)\right)\right) + \sum_{i=1}^{n_{nd}} log\left(P\left(y_i \middle| \frac{1}{cov^2}, cov^2(bo + b_1 t_i)\right)\right) + \sum_{i=1}^{n_{md}} log\left(1 - P\left(y_i \middle| \frac{1}{cov^2}, cov^2(bo + b_1 t_i)\right)\right) \quad (2.2)$$

Where $p(y|k,\theta)$ is the probability distribution function of the gamma distribution and $P(y|k,\theta)$ is the cumulative distribution function of the gamma distribution.

Linear regression will provide a representative estimate of model slope and intercept if the following conditions are met: the data are linear, the residuals (distances between observations and fitted model) follow the given distribution and the residuals have a constant standard deviation (or in the case of gamma, a constant coefficient of variation) along the range of the independent variable (Helsel, 2012).

The assumption of the distribution of residuals can be checked once model parameter estimates have been made. Analysis on residuals can be carried out by subtracting the expected value from the observed, for a normal distribution assumption, and by dividing the observed value by the expected, in the case of the lognormal or gamma distribution assumptions. Distribution assumptions can be checked through a probability plot of residuals, such as depicted in Figure A-2.1. When the plot of residuals fails to adequately fit the normal distribution assumption, the analysis should be repeated with another distribution assumption (eg. lognormal or gamma). A Chi-square test can also be performed on residuals in order to evaluate the distribution assumption fit.



(a) Normal distribution (b) Lognormal distribution

Figure A-2.1 Probability plot (percent of data versus standardized residuals) example from MLE regression estimates with varying distribution assumptions. In this example, the lognormal distribution is better suited for MLE regression of the dataset, though two residual outliers are noted. Crosses represent cumulative probabilities based on observed residuals, lines represent theoretical cumulative probabilities based on the assumed distribution.

In order to test the significance of the computed slope (b_1), the maximum log-likelihood obtained from the given model can be compared to the maximum log-likelihood which is obtained without the linear model, in other words, when a slope of 0 is applied (Helsel, 2012). Once log-likelihood values are determined for both the "null model" and the linear model ($L(0)$ and $L(b)$, respectively), an overall test statistic (G^2) can be computed by (Helsel, 2012):

$$G^2 = 2((b) - L(0)) \tag{2.3}$$

A significance level is then determined by comparing G^2 to a Chi-square distribution with one degree of freedom (Helsel, 2012).

2.2.2 Mann-Kendall

The Mann-Kendall trend test is a nonparametric test which is commonly applied in environmental sciences. The method is also inherently applicable and commonly employed for censored data with a single detection limit and is suited for data which do not meet requirements for a linear regression test. Helsel (2012) also demonstrates how the Mann-Kendall test can be adapted for data with multiple detection limits. However, this capability is seldom coded in most analysis software packages. In this case when multiple detection limit capabilities are not available as part of the analysis in a given software, Helsel (2012) has recommended to first re-censor the data at the highest detection limit.

The Mann-Kendall test can be generally considered a test as for whether observed data tend to increase or decrease over time. In addition to being able to handle non-detects and not rely on linearity or normality assumptions, the Mann-Kendall trend test produces reproducible results regardless of whether the original data is transformed. The test determines whether, within a specified significance level, the dataset is predominantly increasing, decreasing, or remaining steady over time. It does not, however, produce an estimate of the magnitude of such changes over time. The Mann-Kendall test is computed by first determining the *S* test statistic as (Helsel and Hirsch, 2002):

$$S = N_c - N_d \tag{2.4}$$

where N_c refers to the number of concordant pairs in a dataset (where the concentration increases as time increases) and N_D refers to the number of discordant pairs in a dataset (where the concentration decreases as time increases) and *S* has properties (Helsel and Hirsch, 2002):

$$E[S] = 0 \tag{2.5}$$

$$Var[S] = \frac{n(n-1)(2n+5)}{18} \tag{2.6}$$

when no data ties are present in a dataset, or (Gilbert, 1987):

$$\begin{aligned} Var[S] = \frac{1}{18}\{n(n-1)(2n+5) - \sum t_i(t_i - 1)(2t_i + 5) - \sum u_i(u_i - 1)(2u_i + 5)\} \\ + \frac{1}{9n(n-1)(n-2)}\{\sum t_i(t_i - 1)(t_i - 2)\}\{\sum u_i(u_i - 1)(u_i - 2)\} \\ + \frac{1}{2n(n-1)}\{\sum t_i(t_i - 1)\}\{\sum u_i(u_i - 1)\} \end{aligned}$$

for *n* observations, in datasets containing ties of extent t_i in the time domain and u_i in the concentration domain. Ties refer here to data pairs (or triplets, etc.) which cannot be considered concordant nor discordant. This refers to, for example, three samples which where taken on the same day (where *t*=3) or for the tie between the two concentration values <10 and 7, for example (where *u*=2). When *n*>10, a normalized *Z* statistic is then computer as (Helsel and Hirsch, 2002):

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{Var[S]}} & if\ S > 0 \\ 0 & if\ S = 0 \\ \dfrac{S+1}{\sqrt{Var[S]}} & if\ S < 0 \end{cases} \tag{2.7}$$

The test for trend then compares this *Z* value to the critical, two-tailed test Z value for a given significance level ($Z_{critical}$ = 1.96, for example, for a significance value of 0.05) (Ahmad et al., 2015). A two-tailed test is typically used, as the alternative hypothesis being tested is whether there is either an upward or a downward trend present in the data. If the absolute value of the computed *Z* value is greater than the $Z_{critical}$ value, then the hypothesis that there is no trend is rejected and a trend is said to be identified (increasing for positive *S* values and decreasing for negative *S* values).

When $n \leq 10$, the test is performed by finding the exact p value, given *S* and *n*, as published by Helsel & Hirsh (2002, Table B8 or Gilbert, 1987 Table A18). It is noted, however, that published p-values for data sets involving ties where $n \leq 10$ are not readily available. In these cases, the use of published tables such as Helsel & Hirsh (2002, Table B8 or Gilbert, 1987 Table A18) would result in inaccuracies.

Mann-Kendall Correction for Autocorrelated Data (Hamed & Rao, 1998)

Hamed & Rao (1998) proposed a correction to the original Mann-Kendall test in order to account for autocorrelation, when it is present in a dataset. The correction involves a modification of the variance to *Var*[S]*, such that (Hamed & Rao, 1998):

$$Var^*[S] = Var[S]\frac{n}{n_S^*} \tag{2.8}$$

$$\frac{n}{n_S^*} = 1 + \frac{2}{n(n-1)(n-2)}\ x\ \sum_{i=1}^{n-1}(n-i)(n-i-1)(n-i-2)\rho_s(i) \tag{2.9}$$

where $\rho_s(i)$ is the autocorrelation function of the ranks of the observations and $\frac{n}{n_S^*}$ is a correction factor due to the autocorrelation in the data (Hamed & Rao, 1998). The normalized Z-statistic is then calculated from *Var*[S]* as in the normal Mann-Kendall test. Since data rankings are required, the determination of the autocorrelation function is not inherently suited for datasets which include censoring at multiple levels. Though it results in some loss of data, this can be overcome by re-censoring the data at the highest detection limit. Since the correction is carried out on the variance, the correction is only applicable when the normal distribution assumption for the S statistic can be used ($n > 10$, or $n \geq 25$ for seasonal Mann-Kendall tests).

Seasonal Mann-Kendall

Environmental data often exhibit seasonal cycles or fluctuations which have been taken into account in the Seasonal Mann-Kendall test, as presented by Hirsch et al. (1982). The Seasonal

Mann-Kendall test involves breaking the full dataset into subsets, as categorized by "seasons" (typically months, seasons, or as defined by a sampling interval). The test is recommended when evidence of seasonality exists or is suspected and it is meaningful to test seasonal trends separately. Otherwise, a normal Mann-Kendall test is recommended (Hipel & McLeod, 1994). It is computed similarly to the Mann-Kendall test, where (Helsel & Hirsch, 2002):

$$S_k = \sum_{i=1}^{m} S_i \tag{2.10}$$

$$Var[S_k] = \sum_{i=1}^{m} Var[S_k] \tag{2.11}$$

for m seasons. The selection of a length of a season should be such that there is data available for most of the seasons in record (Helsel & Hirsh, 2002). If samples are collected on a monthly basis, for example, there should be 12 seasons per year and if they are collected quarterly, four (Helsel & Hirsh, 2002). As in the Mann-Kendall test, the Z statistic is then computed and compared to a $Z_{critical}$ value at a specified significance level.

For datasets with an inconsistent sampling interval, Helsel & Hirsch (2002) have recommended that in the case where there are a few instances where no value exists for some season of the year and several samples are available for another season, data can be collapsed into one value per season by taking the median, where applicable. Helsel & Hirsch (2002) then further recommended that, when a systematic change in sampling frequency exists in a dataset, the seasons considered should correspond to the lowest sampling frequency, and the single representative season value should be taken as the value which occurred closest to the mid-point of the season, in order to avoid introducing a trend in the computed variance. Another option in this case, as detailed for the Mann-Kendall test, is to use all values within the same season, but adjust the variance for the ties in the temporal domain (Hipel & McLeod, 1994).

2.2.3 Theil-Sen & Akritas-Theil-Sen Slope Estimation

Theil-Sen type slope estimations are nonparametric methods which, unlike Mann-Kendall tests, produce estimates of the magnitude of slope coefficients. As this test produces an estimate of a single slope coefficient, it is best suited to trends which are approximately linear. When this is not the case, Theil-Sen tests can also be performed on sets of transformed data, such as by using the logarithms of data values (Helsel & Hirsch, 2002). The use of the test also assumes that the trend residuals are independent (Helsel & Hirsch, 2002). USEPA (2009) recommends that this test be carried out when there are at least 10 data values.

In the Theil-Sen test, a slope is computed for every possible data pair, and infinite slopes (for samples collected at the same time) are excluded (USEPA, 2009). The median slope of each pair is then taken as the Theil-Sen slope coefficient. In other words, (Ahmad et al., 2015):

$$T_i = \frac{y_j - y_k}{j - k} \tag{2.12}$$

$$Q = Media[T] \tag{2.13}$$

Where y_j and y_k represent data values at times j and k, respectively. Since this computation requires the determination of slopes between all data pairs, this measure is not inherently suited to handle non-detect values. In order to apply this method to a dataset containing non-detects, the use of some substitution method is required.

The Akritas-Theil-Sen is a nonparametric regression method, published in Akritas et al. (1995) and is capable of handling censored values without requiring substitution. It has also been shown to outperform Buckley-James regression, commonly used for nonparametric regression of censored data (Helsel, 2012).

In order to compute the Akritas-Theil-Sen slope, first an initial estimate of the slope is set and this is subtracted from the *Y* (observation) values to create *Y residuals* (where $residual_i = Y_i - slope*X_i$). Kendall's *S* statistic is then computed between the *Y residuals* and the *X* variable (time). An iterative search is then conducted to find the slope that will produce an S of zero (Helsel, 2012). Akritas et al. (1995) specifically describe the slope value as the midpoint between the highest known slope which produces a positive S value and the lowest known slope which produces a negative S value. As with the Theil-Sen slope estimator, the Akritas-Theil-Sen method assumes that the data best fit a linear slope model. If this is not the case, the dataset can be transformed prior to analysis.

3 Summary

Due to their common use in environmental analyses and their ability to be adapted to handle non-detect values (as well as non-detects with varied censoring levels), linear regression using MLE, Mann-Kendall, seasonal Mann-Kendall and Theil-Sen/Akritas-Theil-Sen were chosen to be included in the developed software package. The software program was then implemented in Matlab 2013b, prepared for user-suitability and flexibility and then packaged for distribution. The resulting software is able to read and process data sets containing non-detect values and "greater-than" values, as flagged in a dataset by the symbols "<" and ">".

Further development in the direction of this project could be aimed at adding additional features to the developed software. Since it is desirable for users to be able to perform a variety of trend analysis approach methods, future developments could examine also implementing binary logistic regression as well as Spearman's Rho test.

By implementing a variety of well-documented trend analysis approaches, confidence in analysis results can be increased, and the developed software tool EDDY presents a tool for implementing varied trend analysis approaches.

Bibliography

Ahmad, I., Tang, D., Wang, T., Wang, M., & Wagan, B. (2015). Precipitation Trends over Time Using Mann-Kendall and Spearman's rho Tests in Swat River Basin, Pakistan. *Advances in Meteorology*. Retrieved 01 10, 2016, from http://www.hindawi.com/journals/amete/2015/431860/

Akritas, M. G., Murphy, S. A., & LaValley, M. P. (1995). The Theil-Sen Estimator With Doubly Censored Data and Applications to Astronomy . *Journal of the American Statistical Association*, 170-177.

Barnett, V. (2004). *Environmental Statistics: Methods and Applications.* Chichester: Wiley.

Bergenroth, B., Rineer, J., Munoz, B., Cooter, W., Young, D., & Atkinson, D. (2007). *Google Earth and Statistical Trends Analysis Tools.* Retrieved from RTI International: :www.epa.gov/storet/archive/conf/2007_Bergenroth_StatisticalTrendAnalysisSTORET.p p t

Bolks, A., DeWire, A., & Harcum, B. J. (2014). Baseline Assessment of Left-Censored Environmental Data Using R. *Tech Notes 10*.

Chandler, R., & Scott, M. (2011). *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences.* Wiley.

Chung, C. F. (1990). Regression Analysis of Geochemical Data with Observation Below Detection Limits. *Computer Applications in Resource Estimation*, 421-433.

ESdat. (2015, 11 02). *ESDat Environmental Data Management Software*. Retrieved from http://www.esdat.net/

Hamed, K. H., & Ramachandra Rao, A. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 182-196.

Hamed, K. H., & Rao, A. R. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 182-196.

Helsel, D. R. (2012). *Statistics for Censored Environmental Data Using Minitab and R.* Denver: Wiley.

Helsel, D. R., & Hirsch, R. M. (2002). *Statistical Methods in Water Resources.* Retrieved from Techniques of Water-Resources Investigations of the United States Geological Survey: http://water.usgs.gov/pubs/twri/twri4a3/

Hipel, K., & McLeod, A. (1994). *Time Series Modelling of Water Resources and Environmental Systems (out-of-print).* Amsterdam: Elsevier. Retrieved 01 13, 2016, from http://www.stats.uwo.ca/faculty/mcleod/1994Book/default.htm

Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, 107-121.

Jones, W. R., Spence, M. J., Bowman, A. W., Evers, L., & Molinari, D. A. (2014). A software tool for the spatiotemporal analysis and reporting of groundwater monitoring data. *Environmental Modelling & Software*, 242-249.

Millard, S. P. (2013). *EnvStats: An R Package for Environmental Statistics.* New York: Springer.

Mozejko, J. (2012). Detecting and Estimating Trends of Water Quality Parameters. In *Water Quality Monitoring and Assessment.* Retrieved 11 15, 2015, from http://cdn.intechopen.com/pdfs-wm/35048.pdf

Nova Metrix LLC. (2015, 11 02). *Hydro GeoAnalyst*. Retrieved from Features: http://www.novametrixgm.com/environmental-data-management-software/hydrogeoanalyst#features

Nychka, D., Piegorsch, W. W., & Cox, L. H. (1998). Case Studies in Environmental Statistics. In *Lecture Notes in Statistics.* New York: Springer.

Ofungwu, J. (2014). *Statistical Applications for Environmental Analysis and Risk Assessment.* Hackettstown, NJ: Wiley.

Panneton, M., & Robillard, P. (1972). The exact disribution of Kendall's S with ties in one ranking. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 321-323.

QGC Pty Limited. (2013). *Methodology for Groundwater Level and Quality Trend Analysis, Rev 1.*

Scientific Software Group. (2015, 11 02). *Chempoint/Chemstat*. Retrieved from http://www.scientificsoftwaregroup.com/pages/product_info.php?products_id=70

State of Oregon Department of Environmental Quality. (2015, 10 29). *Trend Analysis and Presentation*. Retrieved from http://www.deq.state.or.us/lab/wqm/docs/TrendAnalysisCD.pdf

Taylor, J. M. (1987). Kendall's and Spearman's correlation coefficients in the presence of a blocking variable. *Biometrics*, 409-419. Retrieved from http://www.uvm.edu/~pdodds/files/papers/others/1987/taylor1987a.pdf

Thomson, M. L., & Nelson, K. P. (2003). Linear regression with Type I interval- and left-censored response data. . *Environmental and Ecological Statistics*, 221-230.

United States Environmental Protection Agency. (2013). *ProUCL Version 5.0.00 - User Guide.* Washington.

USEPA. (2004). *ProUCL Version 3.0 User Guide.* Las Vegas: United States Environmental Protection Agency. Retrieved 01 18, 2016, from http://www.epa.gov/sites/production/files/201503/documents/proucl_v3.0_user.pdf

USEPA. (2009). *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities - Unified Guidance.* United States Environmental Protection Agency. Retrieved 12 20, 2015, from http://www3.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/sitechar/gwstats/unified-guid.pdf

USEPA. (2013). *ProUCL Version 5.0.00 Technical Guide.* United States Environmental Protection Agency.

USEPA. (2013). *ProUCL Version 5.0.00 User Guide: Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations.* Edison, NJ: US EPA.