Spreadsheet for automatic processing of water quality data: 2010 update - Calculation of percentiles and tests for seasonality

C. Daughney

GNS Science Report 2010/42
August 2010

GNS SCIENCE
TE PŪ AO

BIBLIOGRAPHIC REFERENCE

Daughney, C. 2010. Spreadsheet for automatic processing of water quality data: 2010 update – Calculation of percentiles and tests for seasonality, *GNS Science Report* 2010/42 19 p.

C. Daughney, GNS Science, PO Box 30368, Lower Hutt

© Institute of Geological and Nuclear Sciences Limited, 2010
ISSN 1177-2425
ISBN 978-0-478-19789-1

CONTENTS

FIGURES

TABLES

1.0 INTRODUCTION

1.1 Previous versions of the spreadsheet

In 2005, a spreadsheet was developed for automatic processing of water quality data obtained through the National Groundwater Monitoring Programme (Daughney, 2005). This spreadsheet was designed to automatically process water quality results for many different analytes in each of many different samples collected from many different sites. The spreadsheet was developed because software with comparable functionality was not commercially available. Excel was selected to ensure portability. There has been some criticism of Excel's built in higher-level statistical functions (McCullogh and Wilson, 2002), and so these were avoided. Specific functions and capabilities of the spreadsheet are as follows:

- Default settings permit processing of data from up to 199 different analytes, up to 399 different sites, up to 199 samples from each site, and up to a total of 4999 different samples. The spreadsheet can also process data with any number of censored values (i.e. results reported as being less than some detection limit), and there can be several different detection limits for each analyte, even for samples from a single site.
- A subset of the data can be selected for analysis by exclusion of samples collected before and/or after user-specified dates.
- Charge balance error (CBE) is calculated separately for each sample. The user can define acceptable limits for CBE and control whether or not samples with CBE outside acceptable limits are excluded from subsequent data analysis.
- Potential outliers (extreme values) are identified using a non-parametric method. The user can control the sensitivity of outlier detection, and the user can control whether or not outliers are excluded from subsequent data analysis.
- Distributional parameters (e.g. median, average, median absolute deviation, standard deviation, etc.) are calculated on a per-site and per-analyte basis, using a log-probability regression method valid for datasets in which up to 80% of results are below the detection limit.
- Temporal trends in the dataset are identified on a per-site and per-analyte basis, using the Mann-Kendall test. The user can perform either a non-seasonal or a seasonal trend test. For seasonal trend tests, the year can be divided into up to 12 different seasons, and the user can control the Julian day on which the first season starts. Trend magnitudes are quantified using Sen's Slope Estimator and a linear regression method.
- A macro can be used to automatically step through the sites and analytes listed in the input data.

Two years later, the spreadsheet was updated with additional functionality and renamed "2007 NGMP Calculator". The most significant changes to the updated spreadsheet, identified as Version NS-3, are as follows (Daughney, 2007):

- Input data are now provided by the user as an external Excel workbook. There is no longer any need to copy and paste the input data into the spreadsheet.
- The macro automatically copies results of the calculations performed for each site and

analyte into a separate worksheet, permitting convenient import into a report or into another software package for graphing, mapping, etc. The output data are no longer included anywhere in the spreadsheet itself.

- Pick lists are now used to select the current site and analyte. The user no longer has any requirement to refer to site or analyte index numbers.
- Settings for the calculations of CBE are now accessed directly from the Control sheet. Specifically, the user must use a pick list to select the analyte name that is to be used to correspond to each major ion. The user also has the ability to change default settings for the ion's gram formula weight and charge. For example, this allows the user to account for cases where nitrate might be reported in the input data as nitrate, with a gram formula weight of 62 g/mol, or as nitrate-nitrogen, with a gram formula weight of 14 g/mol.
- The plotted date range in the chart on the Control sheet is now automatically adjusted whenever a new site is selected via the pick list. This ensures that the chart displays the data correctly in time series, rather than in any other Excel format.

1.2 Newly implemented features

This report provides a description the most recent version of the spreadsheet, which is named "2007 NGMP Calculator Vers NS-4". The latest version of the spreadsheet is almost identical to Version NS-3, hence the decision to retain "2007" in the file name. However, there are two new capabilities in Version NS-4:

- The minimum, maximum and the 5^{th}, 25^{th}, 75^{th} and 95^{th} percentiles are now included in the reported distributional parameters, along with the median, average, median absolute deviation and standard deviation. As in the previous version of the spreadsheet, these distributional parameters are calculated on a per-site and per-analyte basis using the log-probability regression method.
- Seasonality is assessed with the Kruskal-Wallis test. As in the previous version of the spreadsheet, the user can opt to divide the year into up to 12 different seasons, and the user can control the Julian day on which the first season starts. The same season definitions are used for the Mann-Kendal trend test and for the Kruskal-Wallis test for seasonality. Both tests are also performed with using the same user-specified confidence level.

1.3 Organisation, terminology and syntax of this report

This report consists of three main sections. This first section provides background information pertaining to previous and current versions of the 2007 NGMP Calculator spreadsheet. The second section provides an overview of the theory upon which the newly implemented capabilities of spreadsheet are based. Example calculations are also included in the second section of the report. The third section of the report, which is reproduced from Daughney (2007), gives instructions for use of the spreadsheet.

An effort has been made to use consistent terminology throughout this report. Much of the terminology follows standard conventions in hydrogeology and geochemistry. For example:

Site	A location at which groundwater has been collected for analysis.
Sample	The actual groundwater that has been collected.
Analyte	A parameter that has been measured in a sample, such as electrical conductivity, sodium concentration, etc.
Result	The actual value of a given analyte in a given sample, e.g. 5 g m^{-3}, <0.01 g m^{-3}, 300 μS cm^{-1}, etc.

Excel syntax is employed throughout this report. All references to worksheets and cell addresses follow Excel syntax and are highlighted using bold text. For example, cell addresses are cited using the worksheet name followed by an exclamation point, the column letter and the row number (e.g. **Control!L11, Control!A3**, etc.). Cell arrays are described using these conventions and a colon (e.g. **Control!L11:L15**).

2.0 THEORY FOR NEWLY IMPLEMENTED FEATURES

Readers are referred to Daughney (2005) and Daughney (2007) for theory pertinent to functions implemented in earlier versions of the NGMP Calculator spreadsheet. For example, Daughney (2007) provides an overview of the theory relevant to calculation and assessment of the Charge Balance Error (CBE), calculation of distributional parameters such as the median, average, etc., identification of outliers, and trend testing.

2.1 Calculation of percentiles

Several methods have been applied to estimate percentiles and other distributional parameters (e.g. median, average, etc.) for datasets containing censored results (i.e. results reported as being below some detection limit). Simple substitution methods are widely used. For these substitution methods, censored values are replaced with either zeros or with some fraction of the detection limit (e.g. 0.75 times the detection limit, van Trump and Miesch, 1977; 0.55 times the detection limit, Sanford et al., 1993). However, such methods perform poorly when the number of censored values exceeds about 10% of the dataset (Gliet, 1985; Gilliom and Helsel, 1986; Helsel and Cohn, 1988). Gilliom and Helsel (1986) and Sanford et al. (1993) have shown that a lognormal probability method allows for robust estimation of distributional parameters for datasets with a single censoring threshold. For hydrochemical data, several different censoring levels are often reported for a single element from a single site. Such different censoring levels reflect a change in the analytical detection limit, due to a change in the analytical method, or because analyses were conducted by different laboratories.

The Versions NS-3 and NS-4 of the 2007 NGMP Calculator spreadsheet calculate distributional parameters using the method of Helsel and Cohn (1988), which is essentially an extension of the lognormal probability method of Gilliom and Helsel (1986), adapted for datasets with multiple censoring levels. The method of Helsel and Cohn (1988) involves seven calculation steps, which are described in detail by Daughney (2007). In simple terms, the method uses information about the range of results that are above the detection limit to estimate the most likely values for results that are below the detection limit (Figure 1). It is important to note that the log probability regression method is relatively ineffective for calculation of distributional parameters for heavily censored (> 80%) datasets (Gilliom and Helsel, 1986; Helsel and Cohn, 1988).

Distributional parameters can be calculated once the results that are below the detection limit have been replaced with numerical estimates (Figure 1). Versions NS-3 and NS-4 of the 2007 NGMP Calculator spreadsheet use standard Excel functions to calculate the median, average and standard deviation. Both versions of the spreadsheet also calculate the median absolute deviation (MAD) following the method of Helsel and Hirsch (1992):

$$MAD = median|x_i - \bar{x}|$$

where x_i is the *i*th result and $\bar{x}$ is the median of all *i* results.

Version NS-4 also uses standard Excel functions to calculate the 5th, 25th, 75th and 95th percentiles (Figure 1). The first step in calculating percentiles, by any method, is to sort all *n* results for the analyte from the site in question into order from smallest to largest, and then compute the rank (*r*) associated with the percentile of interest (e.g. *p* = 95 for the 95th percentile). Excel uses the following formula:

$$r = 1 + \frac{p(n-1)}{100}$$

For example, for the results in the rightmost column of Figure 1 (*n* = 26), if we want to calculate the 25th percentile (*p* = 25), the corresponding rank is *r* = 7.25. Accordingly, the value of the 25th percentile is somewhere between the 7th and 8th largest results (1.338E-4 and 1.923E-4, respectively). Since *r* = 7.25, we interpolate ¼ of the way between the 7th and 8th largest results to determine that the 25th percentile is 1.48E-4.

It is important to note that different software packages use different equations to calculate *r*, and no one equation is universally accepted as correct (Scarsbrook and McBride, 2007). The equation used by Excel always produces lower values of r compared to the equations used by other software packages, and so the value of any percentile calculated by Excel will be less than the value of the same percentile calculated by other commonly used programs.

Version NS-4 also reports the minimum and maximum value for the currently selected analyte and site. The maximum is determined using the standard Excel function. If there are no censored results for the analyte/site of interest, the minimum is also determined using the standard Excel function. If there are any censored results, the minimum is reported as being less than the lowest detection limit. For the example shown in Figure 1, even though there is an uncensored result of 0.005, the minimum is reported as <0.01, because the latter may be the smaller of the two. This method of reporting the minimum also ensures consistency with the estimated percentile values. For the example shown in Figure 1, it is consistent for the reported minimum to be <0.01, because the estimated 5th percentile is 2.42E-5 (assuming the estimation of the 5th percentile is robust, the minimum could be taken as not just less than 0.01, but also less than 2.42E-5).

For some hydrochemical datasets, *all* samples may be below the method detection limit for particular analytes. Clearly, in these cases, it is not reliable or even possible to apply the log probability regression method using only samples from the site in question. When all results are censored, the 2007 NGMP Calculator Version NS-4 will report the minimum as being less than the lowest detection limit, and the maximum and percentiles will be reported as being less than the highest detection limit. If there are no results available for the analyte/site

in question, the minimum, maximum and percentiles will be reported as “ND” (not determined).

2.2 Test for seasonality

Version NS-4 of the 2007 NGMP Calculator uses the non-parametric Kruskal-Wallis test to assess seasonality (Helsel and Hirsch, 1992). The term ‘seasonality’ refers to a difference in results for samples collected in at least one of up to 12 user-defined seasons, compared to results for samples collected in other seasons. For example, this test could be used to determine whether chloride concentrations at a particular monitoring are systematically higher in spring compared to summer, autumn and/or winter. Note that the Kruskal-Wallis test indicates whether *at least one* of the seasons differs from the others in terms of reported results for the parameter of interest; a positive test does not suggest that all seasons are different from each other. Note that the the Mann-Kenkall trend test, which is also performed by the Calculator, can account for seasonality. However, there may be a seasonal pattern in the data without a consistent trend over time. The Kruskal-Wallis test was implemented to test the statistical difference between results between seasons, and is independent of the Mann-Kendall trend test.

The first step of the test is to sort all *n* results for the analyte/site of interest into order from smallest to largest. A rank *r* is assigned for each observation *i*. The average rank $\overline{r}$ for all *n* results is (*n* + 1)/2. Each result is then partitioned into a separate season, based on the date of sample collection, and the number of seasons and their date boundaries as defined by the user (see Daughney (2007) for procedures used to define seasons). The ranks for all individual observations *i* falling within a particular season *j* are used to calculate the average rank for that season $\overline{r_j}$:

$$\overline{r_j} = \frac{\sum_{i=1}^{n_j} r_{ij}}{n_j}$$

The test statistic K_j is then calculated from the group rank $\overline{r_j}$ for each season:

$$K_j = n_j \left[\overline{r_j} - \frac{n+1}{2} \right]^2$$

Finally, an overall *K* statistic is determined by summation of K_j across all seasons and weighting for sample size:

$$K = \frac{12}{n(n+1)} \sum_{j=1}^{k} K_j$$

The null hypothesis is that all seasons have the same distribution of results, whereas the alternate hypothesis is that the distribution of at least one season differs. The null hypothesis is rejected if:

$$K \geq \chi^2_{(1-\alpha, k-1)}$$

i.e. if K is greater than or equal to the 1-α quantile of the chi-square distribution having k-1 degrees of freedom, where α is the confidence level to be used in the test (e.g. α = 0.05 for the 95% confidence level) and k is the total number of seasons defined by the user.

Example assessments of seasonality for two analytes are presented in Figure 2 and Table 1. Here, each result is assigned to one of four seasons, based on the date of sample collection. Season 1 is assumed to start on January 1, and all four seasons are assumed to be of equal length. The average rank for each analyte in each season is shown in Table 1. For example, the average ranks for Result A in Seasons 1, 2, 3 and 4 are 14.77, 23.10, 31.80 and 22.64, respectively. The overall value for the Kruskal-Wallis test statistic is 9.974, and the corresponding p value for the chi-square distribution is 0.019. Thus, we would conclude that values of Result A are systematically different for at least one season, compared to all the others, at a confidence level of 95% (i.e. α = 0.05 and $p < \alpha$). The box-whisker plot in Figure 2 shows that values of Result A are higher for Season 3 compared to the other seasons. By contrast, at the 95% confidence level, the Kruskal-Wallis test does not reveal any systematic differences for Result B between the seasons as defined (p value for chi-square distribution is 0.143, and $p > \alpha$).

3.0 USE

Use of 2007 NGMP Calculator Vers NS-4 is essentially exactly the same as the use of version NS-3, which is described by Daughney (2007). In summary:

1. The user must supply the input data in a separate file, formatted according to certain guidelines (see Section 3.1).
2. The user must enter certain parameters on the **Control** sheet, to govern the way the calculations are to be performed, and to select the site and analyte considered (Section 3.2).
3. Calculations will be performed automatically, and the results of the calculations will be displayed on the **Control** worksheet (Section 3.3).
4. If desired, a macro can be used to automatically step through the analytes and sites, and to copy the results of the calculations into a new spreadsheet file (Section 3.4).

3.1 Data input

The user must supply the input data in a separate input file. The user is prompted to browse for the file whenever the 2007 NGMP Calculator is opened, or whenever the user clicks the button on the **Control** worksheet ("Click to open new data file"). In its default form, the 2007 NGMP Calculator spreadsheet can process data from up to 399 different sites, up to 199 different analytes, and up to 199 samples per site and up to 4999 different samples overall. A version capable of processing a larger amount of input data is available from the author.

The 2007 NGMP Calculator spreadsheet can also process data with any number of 'less thans' for each analyte, and there can be several different detection limits for each analyte, even for samples from a single site. The 2007 NGMP Calculator will not operate properly if the input data are not correctly formatted (see Table 2). The following formatting rules must be observed:

1. Row 1 is a header row. Row 2 and higher must contain analytical results. Each row contains results from a single sample; blank rows are not permitted.
2. Column A must be a site identification number. Row 1 is a header row, so the entry in cell **A1** must be text, although any characters can be used (e.g. Feature_ID, ID, #, Site Number, etc.). Rows 2 and higher must contain integer entries in Excel's number format. A unique site identification number is required for each site, and all samples from the same site must have the same identification number. Any numeric identifiers can be used, but it generally appropriate to choose sequential numbers, e.g. 1, 101, 5001, etc. for the first site, 2, 102, 5002, etc. for the second site, 3, 103, 5003, etc. for the third site, and so on. Blank rows and empty cells are not permitted.
3. Column B must be a site name or alias. Row 1 is a header row, so the entry in cell **B1** must be text, although any characters can be used (e.g. Name, Site_Name, Alias, etc.). Rows 2 and higher must contain the names of individual sites. Any numbers or text characters can be used in the site name, but all samples from the same site must use the same site name. Blank rows and empty cells are not permitted.
4. Column C must be the date the sample was collected. Row 1 is a header row, so the entry in cell **C1** must be text, although any characters can be used (e.g. Date, Sample_Start_Date, Date-Time, etc.). Rows 2 and higher must contain date entries; any date, time, numeric or comparable custom format can be used, but Excel's text format is not appropriate. Blank rows and empty cells are not permitted.
5. The remaining columns are for results, one analyte to each column. Row 1 is a header row, so the entries in Row 1 from Column D, E, F, and so on must be text, although any characters can be used (e.g. Nitrate, NO3, NO3-N, Dissolved Nitrate are all acceptable). The columns can contain the analytes in any order. Blank cells are not permitted for analyte names. Duplicate analyte names are not permitted (i.e. no two analytes can have the same name). Rows 2 and higher must contain entries for the analytical results. Analytical results should be listed in g m^{-3} in order for the CBE calculations to work correctly. (In fact any units can be used for any analyte, and where necessary for CBE calculations the user can enter a formula weight that encompasses a unit conversion factor. However, while this is mathematically possible, it is somewhat more cumbersome for the user to keep track of, and hence consistent use of g m^{-3} for all concentrations is recommended). Entries of zero are permitted but should not be used to represent results reported as being below some detection limit ('less thans') or cases where a sample wasn't analysed for a particular parameter. 'Less thans' must be in text format as <0.001, <0.05, etc; a space between the < symbol and the first number is optional. If a sample wasn't analysed for a particular analyte, the cell should be left blank. Any other text entries are not permitted. Note that the 2007 NGMP Calculator does not accept 'greater thans', and so entries such as >500 are not permitted. For the spreadsheet to function, 'greater thans' must be replaced with numbers, e.g. >500 can be replaced with 500. This is not strictly statistically proper, so the user must carefully inspect the results of the calculations for any biases caused by replacements of this nature.
6. The input data worksheet does not need to be sorted, but it is recommended for simplicity to sort the input data in ascending order by site identification number and then by sample collection date.

Whenever a new input file is opened, its data formatting will be automatically checked. A macro is used to ensure that, for example, all entries in Column A are numeric, all entries in

Column C are dates, and there are no blank or duplicate analyte names in Row 1. If the file format is appropriate, then the user is given a message to this effect. If the file format is inappropriate, the user is made aware of this and details regarding the formatting errors are pasted into a new worksheet in the input data file.

3.2 Entry of control parameters

Once the input data have been loaded, the user must enter values on the **Control** sheet to govern the way the calculations are to be performed, and to select the site and analyte to be considered.

1. The pick lists on rows 4 and 5 of columns A to B must be used to select the site and analyte for which calculations are to be performed.
2. Dates must be entered into Cells **D6** and **D7**. These define a 'date window', and samples collected before the date specified in Cell **D6** or after the date specified in Cell **D7** will be excluded during all calculations.
3. A text entry of either Y or N must be entered into Cell **D8**. The entry in this cell determines whether or not results identified as outliers will be excluded from the calculations. If outlying results are excluded, they will not be considered for any calculations, including assessment of distributional parameters, identification of trends, assessment of seasonality, and so on.
4. A positive number must be entered into Cell **D9**. Outliers are defined as being more than *x* times the MAD away from the median, where *x* is the number that the user has entered in Cell **D9**. Experience shows that a value of 3 is usually appropriate, although values from 2 to 4 can be used to identify a larger or smaller proportion of sites as outliers, respectively.
5. A text entry of either Y or N must be entered into Cell **D10**. The entry in this cell determines whether or not samples with CBE outside acceptable limits will be considered during the calculations. Any samples that are excluded on the basis of CBE will be excluded for all calculations, including calculation of distributional parameters, identification of trends, and assessment of seasonality.
6. A numeric entry must be supplied in Cell **D11** to define the threshold for acceptable CBE. For example, a value of 5 would indicate that any sample with CBE below -5% or above 5% might be excluded from subsequent calculations, depending on the entry (Y or N) in Cell **D11**. For reference, a value of 5 is recommended by Freeze and Cherry (1979).
7. The entry in Cell **D12** defines the Julian day on which the first season is assumed to start, for the purpose of the seasonal tests. The entry in this cell must be a positive number between 1 and 365. For example, values of 1, 15, 48 or 200 would indicate that the first season starts on January 1, January 15, February 18 or September 8, respectively. This option is included in the 2007 NGMP Calculator to allow the user to define seasons that do not necessarily correspond to particular weeks or months of the Julian calendar, but may instead correspond to cycles of rainfall intensity or temperature.
8. The entry in Cell **D13** determines the number of seasons considered for the Mann-Kendall trend test and the Kruskal-Wallis test for seasonality. This entry must be a positive integer greater than or equal to 1. A value of 1 results in non-seasonal tests (i.e. all samples are assumed to have been collected in the same season). A value greater

than 1 results in a seasonal testing, where all samples are assigned to categories based on the sample collection date, and the trend and seasonality tests are performed by comparing analytical results for samples collected in each season only to results for other samples collected in the same season. For example, a value of 2 would create two seasons for the trend and seasonality tests, a value of 4 would create four seasons, and so on.

9. The entry in Cell **D14** defines the confidence interval used to identify trends and to assess seasonality. The entry in this cell must be a positive number between 0 and 1. It is usually appropriate to select a value such as 0.01, 0.05 or 0.1, to allow for identification of trends and assessment of seasonality at confidence levels of 99%, 95% and 90%, respectively.
10. The user must provide information to be used for the calculation of CBE. The pick lists in Cells **B17:B30** are used to select the analyte name in the input file that corresponds to each ion in the CBE calculation (Br, Ca, Cl, etc, as listed in Cells **A17:A30**). If the ion is to be excluded from CBE calculations, or if it is not included in the input file, select "N/A" (i.e. "not applicable") from the pick list. Next, enter the values to be used for the gram formula weight and charge for each ion. For example, for Br, the typical gram formula weight would be 79.9 g/mol and the charge would be -1 (negative numbers must be entered for anion charge and positive numbers for cation charge). If desired, at any stage the default settings for CBE can be restored by clicking the box in Row 13.

3.3 Interpretation of results

Once the user has made the desired entries on the **Control** worksheet, calculations will be performed automatically. Results are displayed in two forms. First, the actual input data for the analyte and site in question are displayed in Columns E to I:

1. A graph displays the data for the selected analyte at the selected site as a function of sample collection date. The graph will only show data points for samples collected within the user-specified date window. Samples that were not analysed for the analyte in question are always plotted as zeros. Censored results are plotted at ½ the detection limit. In any case where more than one result is available for a single date, the graph displays the average. If the user has opted to exclude outliers or samples with CBE outside acceptable limits, results for these samples will be displayed as zero on the graph, but these results are not used for subsequent calculations.
2. Rows 16 and upwards display the actual raw data from the input file. Column E indicates the row in the input file corresponding to the sample collected on the date displayed in Column F. Column G displays the calculated CBE (%); note that if the sample in question was not analysed for any of the major ions (Na, K, Ca, Mg, HCO_3, Cl, SO_4), then the result "ND" (i.e. "not determined") will be displayed ("ND" will also be displayed if the pick list in Column B for any of the aforementioned major ions is set to N/A). Column H shows the results of the outlier assessments, and Column I displays the actual reported result from the input file.

The results of the calculations are also summarised in Cells **J1:J30**:

1. The light blue cells (Rows 1 to 3) present the site identification number, site name and analyte name.

2. The light yellow cells (Rows 4 to 6) list the total number of samples collected from the currently selected site, the number of samples collected within the user-specified date window (based on the dates in Cells **D6** and **D7**), and the number of results actually used in the calculation of distributional parameters, trends, etc. The number of results used for the calculations may be less than the total number of samples within the date window if 1) some samples were not analysed for the currently selected analyte, and/or 2) the user has opted to exclude some samples as outliers or based on CBE.

3. The green cells (Rows 7 to 17) summarise the distributional parameters, including the minimum, maximum, median, average, MAD, standard deviation (SD) and selected percentiles, for the currently selected analyte and site. If fewer than two results are available for the currently selected analyte and site, the MAD, SD and percentiles will be reported as ND (not determined). If more than two results are available but if all of these results are censored, the distributional parameters will be reported as being less than the highest detection limit (e.g. <0.01), and the MAD and SD will be reported as ND. The Pearson correlation coefficient (*r*) is also presented to describe the strength of the regression used to estimate distributional parameters for censored data. If the dataset includes any censored values, the *r* value should be greater than 0.8 for the calculated values of the distributional parameters to be meaningful. If fewer than three results are available for the currently selected analyte, the *r* value will be listed as ND (not determined), and if all of the results for the currently selected analyte are above the detection limit, the *r* value will be listed as zero.

4. The orange-brown cells (Rows 18 to 26) summarise the results of the Mann-Kendall trend test and the seasonality assessment. In Cell **K18**, a result of N, DECR, INCR, or ND will be displayed to denote the absence of a statistically significant trend, a significant decreasing trend, a significant increasing trend, or a case where the trend test is 'not determined' due to lack of data, respectively. Cells **K19** and **K20** display the magnitude of the trend (units per year) for the currently selected analyte, based on Sen's slope estimator and linear regression, respectively. If fewer than 2 results are available or if all of the available results are censored, trend magnitudes will be listed as ND (not determined). This array of cells presents values for the variables *p*, *n*, *S* and *Z* defined in Daughney (2007), corresponding to the p value at which the trend is significant, total number of results upon which the trend test is based, Kendall's S statistic, and its Z-scored value, respectively. For robust identification of trends, *n* should be at least 10. Cell **J25** displays the result of the Kruskal-Wallis test for seasonality, and Cell **J26** shows its corresponding p value. In the event that the user has defined only one season (i.e. if Cell **D13** = 1), both of these cells will be reported as "N/A".

5. The pink cells (Rows 27 to 36) provide more detailed information about the input data, such as the number of censored results and the number of zeros. Cell **J27** reports the number of samples within the date window for which CBE cannot be calculated because one or more of the major ions were not analysed (Na, K, Ca, Mg, HCO_3, Cl, SO_4). Cells **J28** to **J30** report the number of samples collected within the date window for which the CBE result is OK (within acceptable limits), Low (anion excess) or High (cation excess), respectively. Cell **J31** reports the number of potential outliers within the date window, based on the user-entered value in Cell **D9**. Cell **J32** reports the number of zeros identified in the input data, within the user-specified date window; blank cells and censored results are not considered in this tally. Cell **J33** reports the number of samples collected within the date window that were not analysed for the currently selected analyte. Cells **J34** to **J36** display the total number of uncensored results, censored

results, and the percentage of censored results used to calculate distributional parameters, trends, etc. If the user has opted to exclude outliers, samples with CBE outside acceptable limits, or samples based on collection date, these choices will be reflected in the values displayed in Cells **J34** to **J36**.

3.4 Operation of the spreadsheet by macro

If desired, a button-triggered macro ("Click to process all sites and analytes") can be used to automatically step through the analytes and sites, and to copy the results of the calculations to a new spreadsheet. This macro can process about 23 site/analyte combinations per minute. For example, an input file containing 100 sites and 30 analytes would require about $100 \times 30 / 23 = 130$ minutes to process completely. The actual processing speed will depend on the specifications of the computer being used. While the macro is running, the screen will be frozen. The macro can be interrupted at any time by hitting the Esc key twice.

3.5 Common problems

The Calculator spreadsheet will not function properly unless the input data worksheet is formatted according to the rules listed in Section 3.1. Ensure that the numeric data are actually formatted using Excel's number format, rather than the text format. Likewise, dates must be expressed in Excel's number, date or comparable custom format rather than the text format. Remember that the 2007 NGMP Calculator spreadsheet can only process up to 199 different analytes or up to 399 different sites. The input data must not contain more than 199 samples per site, and not more than 4999 different samples (rows) overall. Note also that the input data must not contain any 'greater thans'; only numeric entries, blank cells, or 'less thans' are permitted.

4.0 ACKNOWLEDGEMENTS

Robert Reeves and Magali Moreau-Fournier (GNS Science, Wairakei) are thanked for helpful discussions, and assistance with the day-to-day operation of the National Groundwater Monitoring Programme and its database, without which development of this spreadsheet would not have been possible. Ed Mroczek (GNS Science, Wairakei) is also thanked for helpful discussions and support during the development of an early prototype of this spreadsheet. Warwick Smith (GNS Science, Avalon) and Jasim Adam (ACE Training, Wellington) are thanked for assistance with programming to create Version NS-3 of this. Rob van der Raaij and Magali Moreau-Fournier (GNS Science, Avalon and Wairakei) are thanked for reviewing this report.

5.0 REFERENCES

Daughney, C. J. 2005. Spreadsheet for Automatic Processing of Water Quality Data: Theory, Use and Implementation in Excel. GNS Science Report 2005/35. 84 p.

Daughney, C. J. 2007. Spreadsheet for Automatic Processing of Water Quality Data: 2007 Update. GNS Science Report 2007/17. 18 p.

Freeze, R. A., Cherry, J. A. 1979. Groundwater. Prentice Hall, New Jersey. 604 p.

Gilliom, R. J., Helsel, D. R. 1986. Estimation of distributional parameters for censored trace level water quality data: 1. Estimation techniques. Wat. Resources Res. 22: 135-146.

Gliet, A. 1985. Estimation for small normal data sets with detection limits. Env. Sci. Tech. 19: 1201-1206.

Helsel, D. R., Cohn, T. A. 1988. Estimation of descriptive statistics for multiply censored water quality data. Wat. Resources Res. 24: 1997-2004.

Helsel, D. R., Hirsch, R. M. 1992. Statistical Methods in Water Resources. Studies in Environmental Science v. 49, Elsevier, Amsterdam. 529 p.

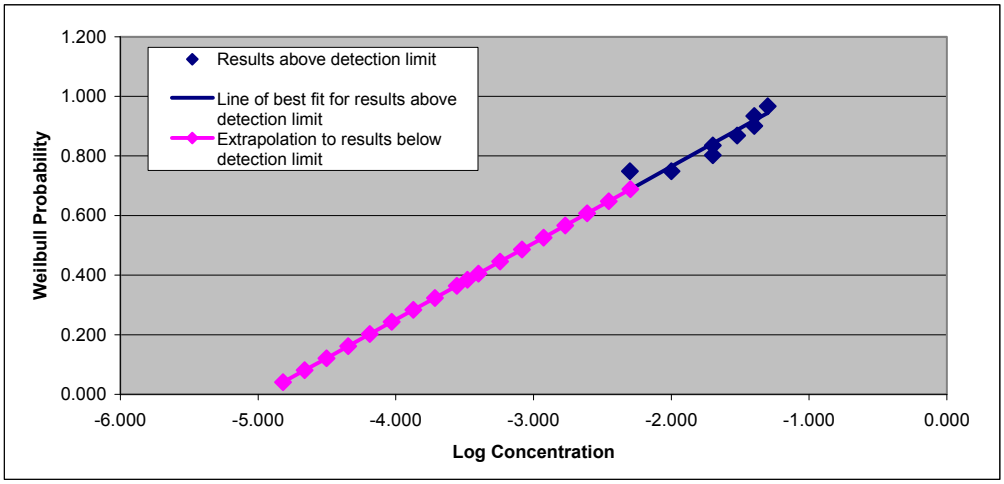
Langmuir, D. 1997. Aqueous Environmental Geochemistry. Prentice Hall, New Jersey. 600 p.

McCullough, B. D., Wilson, B. 2002. On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. Comp. Stat. Data Anal. 40: 713-721.

Sanford, R. F., Pierson, C. T., Crovelli, R. A. 1993. An objective replacement method for censored geochemical data. Math. Geol. 25: 59-90.

Scarsbrook, M. R., McBride, G. B. 2007. Best practice guidelines for the statistical analysis of freshwater quality data. NIWA Client Report HAM2007-088.

van Trump, G. Jr., Miesch, A. T. 1977. The US Geological Survey RASS-STATPAC system for management and statistical reduction of geochemical data. Comput. Geosci. 3: 475-488.



Reported concentration	Results above detection limit		Extrapolated for results below detection limit		Reported and extrapolated concentration
	Log concentration	Weibull probability	Weibull probability	Log concentration	
0.005	-2.301	0.749			0.005
<0.01			0.040	-4.819	1.515E-05
<0.01			0.081	-4.662	2.179E-05
<0.01			0.121	-4.504	3.132E-05
<0.01			0.162	-4.347	4.502E-05
<0.01			0.202	-4.189	6.472E-05
<0.01			0.243	-4.031	9.305E-05
<0.01			0.283	-3.874	1.338E-04
<0.01			0.324	-3.716	1.923E-04
<0.01			0.364	-3.558	2.764E-04
<0.01			0.405	-3.401	3.974E-04
<0.01			0.445	-3.243	5.713E-04
<0.01			0.486	-3.086	8.213E-04
<0.01			0.526	-2.928	1.181E-03
<0.01			0.567	-2.770	1.697E-03
<0.01			0.607	-2.613	2.440E-03
<0.01			0.648	-2.455	3.508E-03
<0.01			0.688	-2.297	5.043E-03
0.01	-2.000	0.749			0.01
<0.02			0.385	-3.480	3.314E-04
0.02	-1.699	0.802			0.02
0.02	-1.699	0.835			0.02
0.03	-1.523	0.868			0.03
0.04	-1.398	0.901			0.04
0.04	-1.398	0.934			0.04
0.05	-1.301	0.967			0.05

Minimum	<0.01
5th Percentile	2.42E-05
25th Percentile	1.48E-04
Median	1.00E-03
75th Percentile	8.76E-03
95th Percentile	0.04
Maximum	0.05
Average	8.92E-03
MAD	9.74E-04
Standard Deviation	1.48E-02

Figure 1. Example estimation of concentrations for censored results based on the method of Helsel and Cohn (1988). Reported concentrations, both above and below the detection limit, are shown in the left-most column. Weibull probabilities are derived for the results above the detection limit, and then a regression line is fit to derive the relationship to log concentration. This regression line is then extrapolated to estimate the highest probability values of log concentration for all results reported as below the detection limit (see Daughney (2007) for explanation of calculations). Finally, distributional parameters, including percentiles, are calculated from the combination of the reported and extrapolated concentrations (for the uncensored results and replacement estimates for the censored results, respectively, as shown in the right-most column).

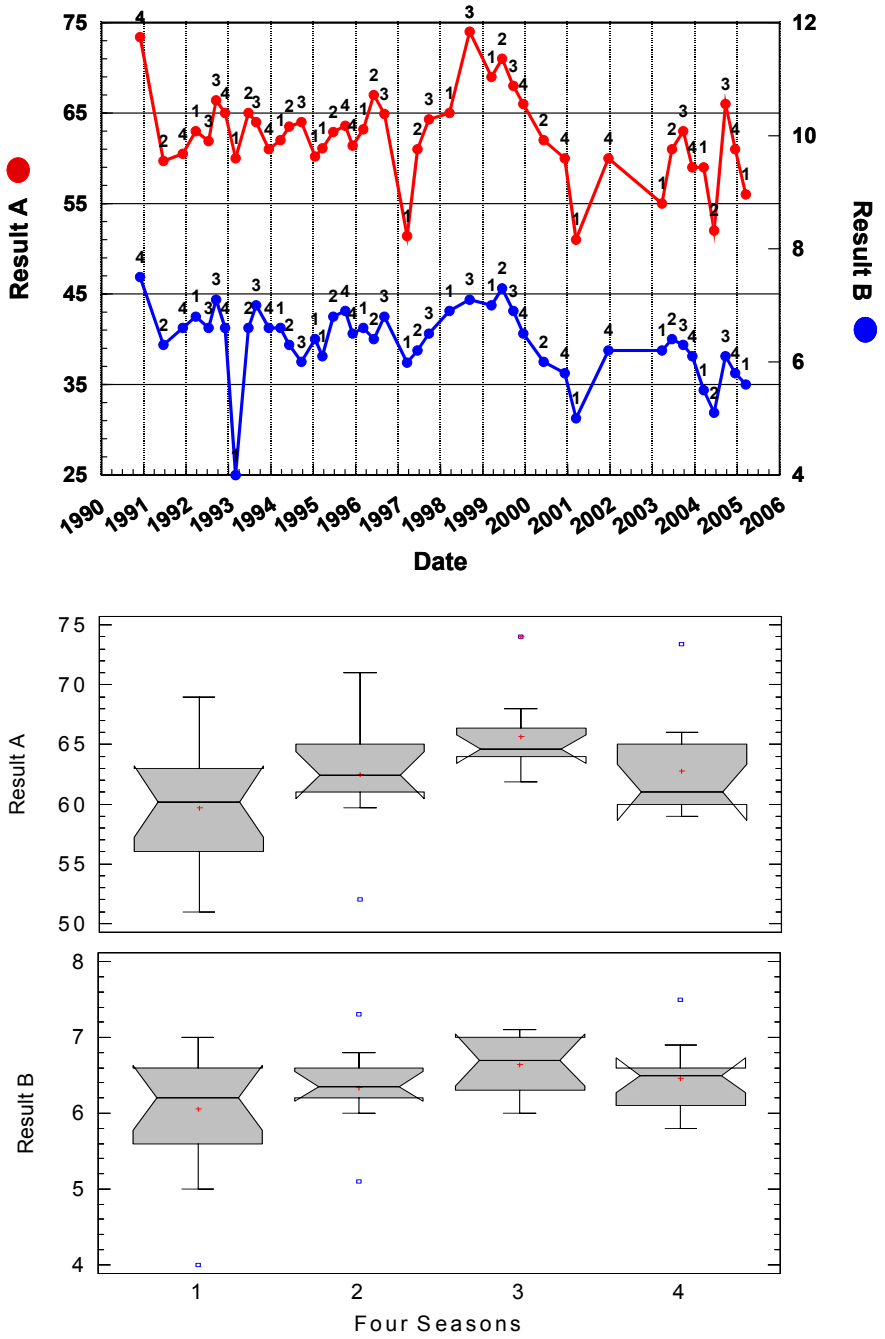


Figure 2. Example assessment of seasonality for two parameters. Top plot shows time series with each result assigned to one of four seasons as indicated by data point labels (season 1 is assumed to start on Julian day 1, i.e. January 1). Lower box-whisker plots compare the range of reported results within each season. The rectangular part of each "box" extends from the 25th to the 75th percentile, covering the centre half of each season's distribution of results. The centre lines and plus signs within each box indicate the medians and the means, respectively. The "whiskers" extend from the box to the minimum and maximum values in each sample, except for any outside points that lie more than 1.5 times the interquartile range above or below the box, and are plotted separately. Also included on the plots are notches covering a distance above and below each median. If the two notches for any pair of medians overlap, there is not a statistically significant difference between the medians at the 95% confidence level. For the examples shown here, there is a significant seasonal difference for Result A (higher in Season 3 than in other season), whereas there are no significant seasonal differences for Result B. Different results for the assessment of seasonality could be obtained with more/fewer seasons defined, or if the boundary dates between the seasons were changed.

Table 1. Example assessment of seasonality for two parameters. Corresponding time series plots and box-whisker plots are shown in Figure 1. The left table shows sample collection date, results for two parameters, and assigned season. For the latter, it is assumed that the year is divided into four seasons of equal length and the first season starts on January 1. The two tables on the right show for each parameter the number of results per season, the average rank per season, computed K value, p value, and the outcome of the seasonality test.

Date	Result A	Result B	Season
30/11/1990	73.4	7.5	4
19/06/1991	59.7	6.3	2
3/12/1991	60.5	6.6	4
24/03/1992	63	6.8	1
13/07/1992	61.9	6.6	3
15/09/1992	66.4	7.1	3
2/12/1992	65	6.6	4
2/03/1993	60	4	1
21/06/1993	65	6.6	2
26/08/1993	64	7	3
13/12/1993	61	6.6	4
23/03/1994	62	6.6	1
7/06/1994	63.5	6.3	2
20/09/1994	64	6	3
16/01/1995	60.2	6.4	1
20/03/1995	61.1	6.1	1
21/06/1995	62.9	6.8	2
3/10/1995	63.6	6.9	4
5/12/1995	61.4	6.5	4
6/03/1996	63.2	6.6	1
5/06/1996	67	6.4	2
4/09/1996	64.9	6.8	3
19/03/1997	51.4	5.985	1
17/06/1997	61	6.2	2
23/09/1997	64.3	6.5	3
19/03/1998	65	6.9	1
9/09/1998	74	7.1	3
16/03/1999	69	7	1
15/06/1999	71	7.3	2
20/09/1999	68	6.9	3
15/12/1999	66	6.5	4
7/06/2000	62	6	2
6/12/2000	60	5.8	4
13/03/2001	51	5	1
18/12/2001	60	6.2	4
24/03/2003	55	6.2	1
18/06/2003	61	6.4	2
23/09/2003	63	6.3	3
10/12/2003	59	6.1	4
17/03/2004	59	5.5	1
15/06/2004	52	5.1	2
21/09/2004	66	6.1	3
13/12/2004	61	5.8	4
15/03/2005	56	5.6	1

Result A				
Season	1	2	3	4
n	13	10	10	11
Av Rank	14.77	23.10	31.80	22.64
K_j	776.94	3.60	864.90	0.20
Seasonality Test Results				
K (overall)				9.974
p value, chi-squared distribution				0.019
Is significant seasonality detected?				YES

Result B				
Season	1	2	3	4
n	13	10	10	11
Av Rank	18.23	19.40	29.90	23.64
K_j	236.94	96.10	547.60	14.20
Seasonality Test Results				
K (overall)				5.423
p value, chi-squared distribution				0.143
Is significant seasonality detected?				NO

Table 2. Example of properly formatted input data. Each site has a unique identification number, and all samples from the same site have the same identification number. The worksheet is sorted in order by ascending identification number (Column A) and then by ascending sample collection date (Column C). All identification numbers are in number format, and all dates are in date format. Each analyte has a name listed in Row 1; there are no blanks and no two analytes have the same name. All of the cells that represent analytical results contain either numbers, blanks, or "less thans". There are no "greater thans" or any other type of text entry. The entire input sheet contains less than 199 analytes, less than 399 sites, less than 199 samples for each site, and less than 4999 rows of data overall.

	A	B	C	D	E	F	G
1	**ID**	**Name**	**Date**	**Calcium**	**Cl, Total**	**NO3-N**	
2	2003	Site 1	09/25/90	13.6	8.8	3.3	
3	2003	Site 1	11/30/90	15.1		2.3	
4	2003	Site 1	03/26/91	15.5	12.4	3.03	
5	2004	Breyers	09/25/90		12.1	<0.05	
6	2004	Breyers	11/30/90	21	12.4	<0.01	
7	2005	Johnson's Spring	11/30/90	37.4	4.8		
8	2005	Johnson's Spring	03/26/91	42.3	5.2	1.2	
9	2005	Johnson's Spring	06/19/91		5.6	<0.04	
10	2005	Johnson's Spring	09/23/91		5.9	2.4	
11	2005	Johnson's Spring	12/03/91	44	5.3	2.3	
12							

www.gns.cri.nz

Principal Location

1 Fairway Drive
Avalon
PO Box 30368
Lower Hutt
New Zealand
T +64-4-570 1444
F +64-4-570 4600

Other Locations

Dunedin Research Centre
764 Cumberland Street
Private Bag 1930
Dunedin
New Zealand
T +64-3-477 4050
F +64-3-477 5232

Wairakei Research Centre
114 Karetoto Road
Wairakei
Private Bag 2000, Taupo
New Zealand
T +64-7-374 8211
F +64-7-374 8199

National Isotope Centre
30 Gracefield Road
PO Box 31312
Lower Hutt
New Zealand
T +64-4-570 1444
F +64-4-570 4657