# 

NRV R Package Guidance Document

DRAFT

Prepared for:

Lorraine Brekke

Northwest Territories Cumulative Impact Monitoring Program

Environment and Natural Resources Conservation, Assessment and Monitoring

PO Box 1320

7th Floor, Scotia Building, 5012 50th Avenue

Yellowknife, Northwest Territories X1A 2L9

Prepared in March 2024 by:

Jess Kidd

Reviewed by: Heidi Swanson

University of Waterloo

200 University Avenue

Waterloo, Ontario N2L 3G1

**Table of Contents**

[1.0 Introduction 1](#_Toc162626265)

[2.0 Data Preparation 2](#_Toc162626266)

[2.1 Data 2](#_Toc162626267)

[2.2 Data Preparation 2](#_Toc162626268)

[2.3 Data Frame Formatting 2](#_Toc162626269)

[3.0 Using the NRV Package in R Studio 4](#_Toc162626270)

[3.1 Install NRV R Package in R Studio 4](#_Toc162626271)

[3.2 Create NRV Table 7](#_Toc162626272)

[3.3 Create NRV Time Series Plots 11](#_Toc162626273)

# 1.0 Introduction

This document includes all the information people need to use the NRV R Package. The NRV R Package includes the core NRV function that calculate the Natural Range of Variation (NRV) lower and upper thresholds for surface water quality data.

The goal of the NRV package is to help government personnel and consultants calculate baseline values for water quality parameters using the Natural Range of Variation (NRV) method developed by the Northwest Territories Cumulative Impact Monitoring Program (NWT CIMP). This NRV method calculates the lower and upper thresholds for individual water quality parameters using either the Tukey Inner Fence (TIF) method for parameters with <25% results below the detection limit or the Median±2MAD (M2MAD) method for parameters with ≤50% results below the detection limit. If datasets contain samples from potentially impacted sites, then the M2MAD method—which is more conservative than the TIF method—is applied to all parameters, including those with <25% results below the detection limit.

The NRV package provides the NRV function that calculates NRV thresholds for data frames that contain surface water quality results for one or more sites. It is important to note that the analysis will run for each unique parameter label (i.e., all sites within a dataset are pooled for NRV analysis). Users must remove any site from the dataset that is not meant to be included in the NRV calculation.

The NRV function produces a table of the NRV thresholds and associated summary statistics (Section 3.2). The NRV function requires you to format your data frame with specific column names. Please refer to Section 2.3 Data Frame Formatting below and the example data frames in the R Package to see how to format your data frame. The NRV Package also includes a NRVTimeSeries function that produces time series plots that display the NRV thresholds calculated by the NRV function (Section 3.3). The NRV function must be run before the time series function so that the required data tables with the NRV thresholds are available.

If you just want to test the NRV package function, then please skip section 2.0. The R package contains an example data frame you can use in Section 3.0 to test the package function.

# 2.0 Data Preparation

## 2.1 Data

The dataset you prepare for the NRV calculation should ONLY include the data you plan to calculate the NRV thresholds with. It is important to identify if all data are from reference sites (i.e., unimpacted) or if some are from potentially impacted sites. The NRV function default is set to “reference,” but the user can change it to “impacted” if any sites are potentially impacted.

## 2.2 Data Preparation

The following are important steps to perform to prepare your data for the NRV threshold calculation:

1. Average all duplicate samples.
2. Remove any censored value (i.e., value below the laboratory minimum detection limit) with a detection limit greater than any measured value (i.e., value above the laboratory minimum detection limit).
3. All detection limits that were less than all measured values should be standardized to the same detection limit. We recommend using the lowest detection limit because this technique is least likely to introduce errors into your dataset.

For example, if you have two censored values with detection limits of 0.1 and 0.2, then the detection limit would be standardized to 0.1. The substitution method of one-half the detection would yield values of 0.05 for both samples. It is possible that both samples were 0.05. However, if we instead standardize the detection limits to the highest detection limit, then the substitution method would yield values of 0.1 for both samples. We know at least one of these results is an error because we know the one sample was <0.1.

## 2.3 Data Frame Formatting

A critical first step before using the NRV R Package is to ensure your dataset includes all required columns. The required columns must be labelled as they are here for the function to identify each column required for the NRV threshold calculations. There are **eight** required columns, which are discussed in detail below.

1. **Site**: The “Site” column contains the site ID for where the sample was collected from.
2. **Date**: The “Date” column contains the data when the sample was collected. The dates will be presented in the YYYY-MM-DD format (e.g., 2024-03-25).
3. **Parameter**: The “Parameter” column will contain the name of the parameter that was analyzed. Ensure all parameter names are spelled exactly the same throughout your dataset. The NRV thresholds will be calculated for each unique parameter name. If you are calculating the NRV thresholds for both total and dissolved fractions, we recommend distinguishing them by adding the fraction to the parameter name (e.g., arsenic\_total, phosphorus\_dissolved). It may also be useful to label the parameters with their units of measurement (e.g., mercury\_total\_ug\_L and mercury\_total\_ng\_L). Remember, the names should NOT contain spaces, only underscores “\_”.
4. **DL**: The “DL” column contains the laboratory minimum detection limit value. If no detection limit value is available, then this column will be NA. This column should only contain numerical values and not any units of measurement.
5. **RDC**: The “RDC” column (Result Detection Condition) identifies if the result is below the detection limit. In this column, each row should be either “DET” (i.e., detected) for all results that were above the laboratory minimum detection limit OR “BDL” (below detection limit) for all samples that were below the laboratory minimum detection limit.
6. **ResultRaw**: The “ResultRaw” column contains either the raw value for the laboratory results that were above the detection limit AND the value of the detection limit for all results below the detection limit. This column should only contain numerical values and not any units of measurement.
7. **ResultCalc:** The “ResultCalc” column contains the data on which the summary statistics and NRV calculations will be performed. It must contain the raw values for all results above the detection limit AND the value you choose to represent the results below the detection limit (e.g., half the detection limit).This column should only contain numerical values and not any units of measurement.
8. **ResultCalcLog:** The “ResultCalcLog” column contains the data on which the summary statistics and NRV calculations will be performed if the Shapiro-Wilk test results indicate log-transformation is necessary. This column must contain the log-transformed values of the ResultCalc column. The transformation MUST be completed using the natural logarithm. To achieve this in R, use the function log(). This column should only contain numerical values and not any units of measurement.

# 3.0 Using the NRV Package in R Studio

## 3.1 Install NRV R Package in R Studio

1. Open your R Studio and click File – New Project – Existing Directory

Navigate to the folder that contains the R package files (it should be labelled natural-range-of-variation-source-master). If you have received this as a zip.folder, make sure you have saved it as an uncompressed folder on your computer.

1. Your R Workspace should look like the image below with all of the R package files listed in the bottom right **Files** window.

A screenshot of a computer

Description automatically generated

The figure below highlights 3 key areas of your R studio workspace, the **Console**, **Environment**, and **Files.**

A screenshot of a computer

Description automatically generated

1. You need to install and load the development tool kit. You can skip the installation step if you’ve already installed devtools. ***Note***: If this is your first time downloading the devtools package then it may take several minutes and expect to see a lot of text appearing in your **Console**. You will know the download is complete when the blue arrow appears at the bottom of your **Console**. In the **Console** (where the “>” symbol is), type the following (or copy/paste) and then hit the enter key:

install.packages("devtools")

library("devtools")

A screenshot of a computer

Description automatically generated

1. Now you need to install and load the NRV package. ***Note*:** When you install the R package, you may receive a message that lists all of the recent updates to packages used within the NRV package. Simply enter the number of updates it lists to continue. In the **Console**, type the following:

install()

Now, load the NRV Package by typing the following in the **Console**:

library(NRV)

## 3.2 Create NRV Table

The NRV function generates an NRV Table that includes the following for each unique parameter in your data frame: summary statistics, the percent of samples with results below the detection limit, the p-values of the Shapiro-Wilk test performed on untransformed and log-transformed data, the NRV method selected, and the upper and lower NRV thresholds.

1. Import your data frame into R Studio. ***Note***: your data frame MUST be formatted exactly as outlined in **Section 2.3** or the NRV functions will not work.

Here, we will guide you on using one of the example data frames that comes with the NRV Package. To use your own data frame, simple replace the name of the example data frame with the name of yours as it appears in your **Environment**.

To access the example data, click the “ExampleData” file in **Files**. Then click “NRV\_example” and click “ImportDatset…”. When the Import Text Data window appears just click “Import” in the bottom right corner. See sample images below.

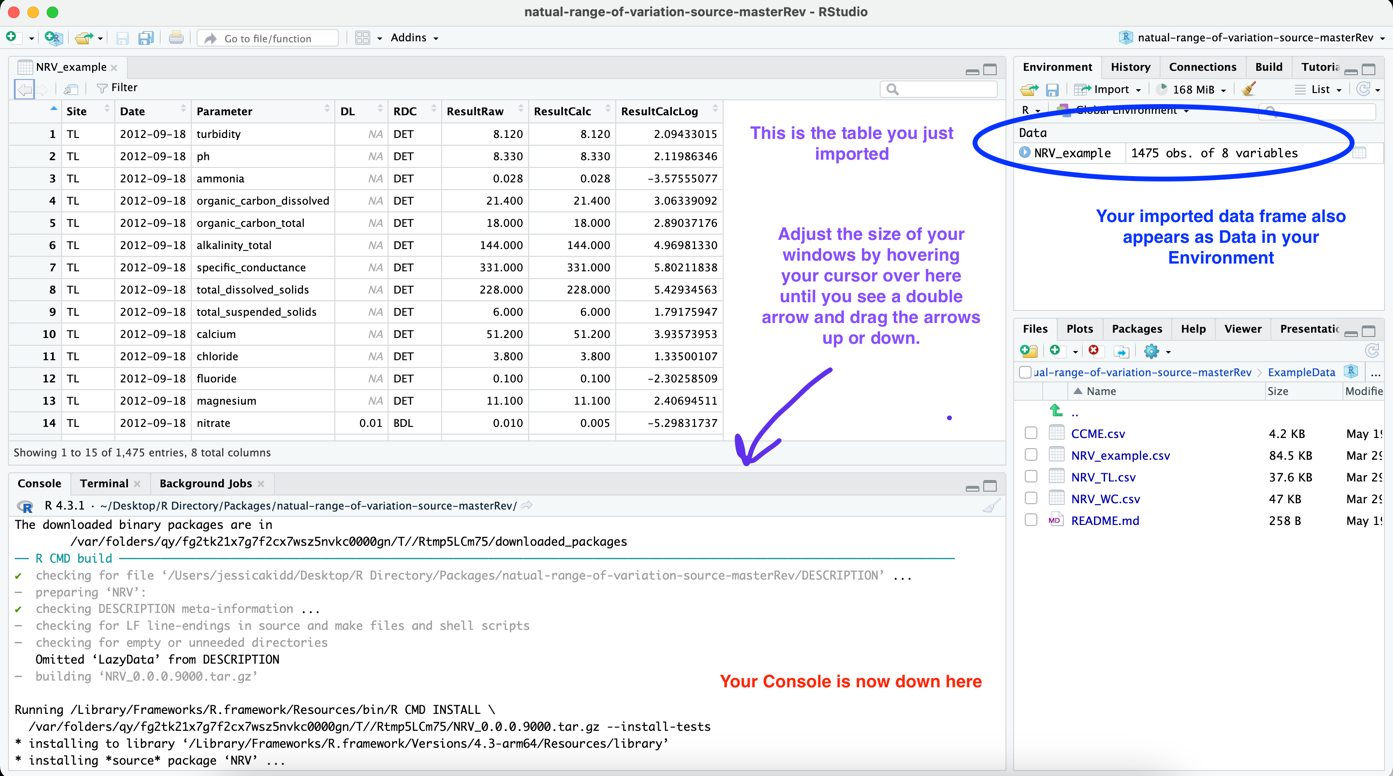
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The example data frame NRV\_example should now appear in your **Environment** and a new window will appear in the top left corner of your R Workspace displaying the contents of your data frame. You can adjust the size of your Console that is now in the bottom left corner of your R workspace where you will continue to type your commands. See example image below.



1. Before using the NRV function, you must first decide if your data frame contains only data from reference sites or contains data from sites that are potentially impacted. The NRV function by default assumes you are using only data from reference sites.

If you are working with only data collected from reference sites, then type the following in the Console to create the NRV Table:

NRV(NRV\_example)

If you are working with only data collected from potentially impacted sites, then type the following in the Console to create the NRV Table:

NRV(NRV\_example, data\_type="impacted")

The results\_table (i.e., NRV Table) will appear in your R Workspace. Type the following function to also save it as a data frame in your **Environment**.

NRV\_Table<- NRV(NRV\_example)

The NRV Table is now saved in your Environment as NRV\_Table. See example image below.

A screenshot of a computer

Description automatically generated

You can now export this table to your computer as a .csv file.

write.csv(NRV\_Table, "NRV\_Table.csv", row.names = TRUE)

This code will export the .csv file to your R Working Directory. To see where your R Working Directory is on your computer, type the following:

getwd()

Below are the details of the NRV Summary Table output:

**Parameter**: name of parameter the calculations were performed on.

**n**: the total number of samples used to perform the calculations.

**percentNonDetect**: the percentage of samples with results below the laboratory minimum detection limit (i.e., “BDL” in the RDC column of your data frame).

**MEAN**: The mean value of the parameter’s CalcResult values.

**MIN**: The minimum value of the parameter’s CalcResult values.

**MAX**: The maximum value of the parameter’s CalcResult values.

**SD**: The standard deviation of the parameter’s CalcResult values.

**P25**: 25th percentile (i.e., 1st Quartile) of the parameter’s CalcResult values.

**P50**: 50th percentile of the parameter’s CalcResult values.

**P75**: 75th percentile (i.e., 3rd Quartile) of the parameter’s CalcResult values.

**P90**: 90th percentile of the parameter’s CalcResult values.

**S-W**: The p-value of the Shapiro-Wilk test performed on the parameter’s CalcResult values.

**S-WLog**: The p-value of the Shapiro-Wilk test performed on the parameter’s CalcResultLog values.

**NRVMethod**: The NRV Method that was performed on the parameter values. If the data contained <25 BDL data, then it will be TIF or TIFLog. If the data contained 25-49% BDL data, then it will be M2M or M2MLog. The Log method is used only if the SW value is <0.05 AND the S-WLog value is ≥0.05. If both are <0.05, then the untransformed method is used. ***Note***: The M2MAD methods are used for all parameters if the data\_type is set to “impacted”.

**lowerThreshold**: The calculated lower NRV threshold.

**upperThreshold**: The calculated upper NRV threshold.

## 3.3 Create NRV Time Series Plots

The NRV Package contains the function NRVTimeSeries that will automatically create time series plots for all parameters that display the lower and upper NRV thresholds.

To create time series plots of all the parameters that have calculated NRV thresholds, you must enter the following information into the function:

* 1. The name of your data frame that contains the data you want displayed on the time series plot. In this example, this data frame is NRV\_example.
  2. The name you have assigned the results\_table from the NRV function. In this example, this table is NRV\_table.
  3. The season the data were collected in. In this example, the season is Open-Water

To run the function, type the following:

NRVTimeSeries(NRV\_example, NRV\_Table, season = "Open-Water")

Below is an example of what the time series plots will look like. TL and WC are the site names from the data frame.

A graph with red and blue dots

Description automatically generated

The plots will all appear automatically in the bottom right-hand corner of you R Workspace under the **Plots** tab.

A screenshot of a computer

Description automatically generated

Once you select Save as Image a new window “Save Plot as Image” will appear with options to select where the image is saved on your computer and to adjust the size of the image. Adjusting the image size is important to ensure the resolution is suitable for a report. Typically, a width of 1900 and heigh of 800 is sufficient.

A screenshot of a computer screen

Description automatically generated