mStatGraph Matlab Package

Plotting and Statistical Analysis for Oceanographers, Meteorologists and Ecologists

USER MANUAL

By

Humberto L. Varona

Version 1.7

May, 2023

Software to analyze, compare and validate analysis and reanalysis datasets with an observed dataset (DSCompare).

Overview

Software to carry out statistical analysis, plot figures and maps of spatial distribution of physical and biogeochemical parameters, calculation of the parameters of the carbonate system of rivers and of MLD, ILD and BLT.

Version

1.7

Release date

April, 15th 2023

License

MIT

DOI

10.5281/zenodo.8152683

Download URL

https://zenodo.org/deposit/ 8152683

Cite as

Varona, Humberto L., Noriega, C., Araujo, J., Lira, S. M. A., Araujo, M., & Hernandez F. (2023). Plotting and Statistical Analysis for Oceanographers, Meteorologists and Ecologists (mStatGraph). Version 1.7. Zenodo. https://doi.org/10.5281/zenodo.8152683

How to install

Matlab 2021b compatible software

- Open Matlab.
- Go to APP tab.

- Click on the "Install App" button.
- Select the mStatGraph.mlappinstall file.
- In the Install dialog click on the "Install" button.

Create a directory and copy into it the following databases: etopo_2.mat, gshhs_f_coast.mat, gshhs_f_rivers.mat, and wdb_f_borders.mat.

How to run

Type in the Matlab command window:

>> mStatGraph<Enter>

or find mStatGraph in the APP tab of Matlab.

Statistical tests

- *The minimum is the smallest value observed in a dataset. It represents the lowest point or the lowest measurement recorded.*

- *The maximum is the largest value observed in a dataset. It represents the highest point or the highest measurement recorded.*

- *Amplitude refers to the range between the minimum and maximum values in a dataset. It provides a measure of the spread or variability of the data.*

- *The mean, also known as the average, is the sum of all values in a dataset divided by the total number of observations. It is a measure of central tendency that represents the typical or average value of the data.*

- *The trimedian is a robust measure of central tendency that provides a compromise between the median and mean. It is calculated by taking the average of the median and the midrange (the average of the minimum and maximum values).*

- *The standard deviation is a measure of the dispersion or variability of the data points around the mean. It quantifies how much the individual data points deviate from the mean. A higher standard deviation indicates greater variability, while a lower standard deviation indicates less variability.*

- *The standard error measures the variability of the sample mean. It represents the standard deviation of the sample means that would be obtained if multiple samples were taken from the same population. The standard error provides an estimate of the precision or reliability of the sample mean as an estimate of the population mean.*

- *Variance is a measure of the dispersion of data points around the mean. It is calculated as the average of the squared differences between each data point and the mean. A*

higher variance indicates greater variability in the data, while a lower variance indicates less variability.

- *The coefficient of variation (CV) is a relative measure of variability that expresses the standard deviation as a percentage of the mean. It is calculated by dividing the standard deviation by the mean and multiplying by 100. The CV allows for the comparison of the variability of different datasets with varying means.*

- *The median is the middle value in a dataset when the data is arranged in ascending or descending order. It represents the value that separates the higher half from the lower half of the data. The median is a measure of central tendency that is not affected by extreme values.*

- *The mode is the value or values that occur most frequently in a dataset. It represents the peak or the most common value(s) in the distribution.*

- *The mode frequency is the number of times the mode value(s) appear in a dataset. It indicates the count or frequency of the most frequently occurring value(s).*

- *Skewness measures the asymmetry of a distribution. It quantifies the extent to which the data is skewed to the left or right. Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side.*

- *Kurtosis measures the degree of peakedness or flatness of a distribution. It quantifies the tails and outliers in the data. Positive kurtosis indicates a relatively peaked distribution with heavier tails, while negative kurtosis indicates a flatter distribution with lighter tails.*

- *Quartiles divide a dataset into four equal parts, each containing 25% of the data. The first quartile (Q1) represents the 25th percentile, the second quartile (Q2) represents the median, and the third quartile (Q3) represents the 75th percentile.*

- *The interquartile range (IQR) is a measure of statistical dispersion. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). The IQR represents the range of the middle 50% of the data and provides a measure of the spread that is less influenced by extreme values.*

- *Percentiles are values that divide a dataset into hundred equal parts. The Pth percentile represents the value below which a given percentage of the data falls. For example, the 25th percentile represents the value below which 25% of the data falls.*

- *Mann-Kendall test is a non-parametric test used to assess the presence of trends in a time series data. This test is widely employed in various fields, including oceanography, to analyze changes or variations in variables such as water temperature, salinity, or sea level. The test compares the rank differences between successive data pairs and calculates the test statistic known as Kendall's Tau. A positive value of Tau indicates an increasing trend, while a negative value indicates a decreasing trend.*

- *Mann-Whitney test, also known as the Mann-Whitney U-test, is a non-parametric test used to compare two independent samples and determine if there are significant*

differences between them. In oceanography, this test could be used to compare two groups of data, such as the concentrations of a specific chemical compound in two different areas or time periods. The test is based on comparing the rankings of the data in both groups and calculates a test statistic called U. If the U value is sufficiently small or large, it can be concluded that there are significant differences between the samples.

- *Two-sample t-test is a parametric test used to compare the means of two independent groups. In oceanography, this test could be applied to compare the mean of a physical or chemical variable at two different locations or time points. The t-test compares the differences between the means of the groups, taking into account the variability within the groups. Upon conducting the test, a t-test statistic and a p-value are obtained. If the p-value is less than a predefined threshold (usually 0.05), it is considered that there are significant differences between the means of the groups.*

- *The Kolmogorov-Smirnov one-sample test is a non-parametric test used to evaluate if a sample of data follows a specific probability distribution. In oceanography, this test could be used to assess whether data from a variable, such as water temperature, fits a normal distribution or another theoretical distribution. The test compares the empirical cumulative distribution function of the data with the theoretical cumulative distribution function and calculates a test statistic based on the maximum absolute difference between the two functions.*

- *The Kolmogorov-Smirnov two-sample test is a non-parametric test used to compare the distributions of two independent samples. In oceanography, this test could be employed to determine if the distributions of two physical or chemical variables are different, for example, comparing the distributions of oxygen concentration in two different zones or time periods. The test compares the cumulative distribution functions of the two samples and calculates a test statistic based on the maximum absolute difference.*

- *The Kruskal-Wallis test is a non-parametric test used to compare the medians of three or more independent groups. In oceanography, this test could be utilized to analyze if there are significant differences in a biological or chemical variable among different locations or time periods. The test is based on the ranks of the data and calculates a test statistic known as H. If the H value is sufficiently large and the associated p-value is less than a predetermined threshold, it is concluded that there are significant differences between the medians of the groups.*

- *The Dunn test is a multiple comparisons procedure used as a post-hoc test after finding significant differences in the Kruskal-Wallis test. It allows identifying which specific groups have significant differences from each other. It provides a correction for multiple tests and calculates adjusted test statistics.*

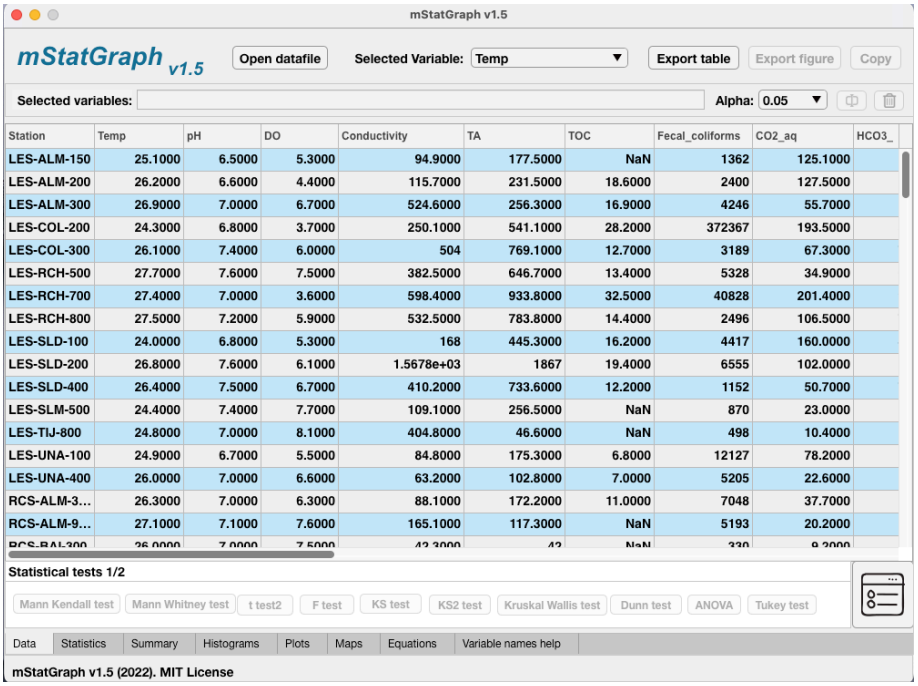
- *ANOVA is a parametric test used to compare the means of three or more independent groups. In oceanography, it can be applied to analyze if there are significant differences in a physical or chemical variable among different locations or time periods. The test decomposes the total variability in the data into components due to differences between groups and within groups. It calculates an F-test statistic and an associated p-value. If the p-value is less than a predetermined threshold, it is concluded that there are significant differences between the means of the groups.*

- *The Tukey test, also known as the Tukey's Honestly Significant Difference (HSD) test, is used as a post-hoc test after finding significant differences in ANOVA. It allows for the comparison of all possible pairs of means. This test helps identify which specific groups have significantly different means. It provides a correction for multiple comparisons and calculates adjusted test statistics.*

The aforementioned statistical tests play crucial roles in different aspects of data analysis and interpretation within the field of oceanography. They provide researchers with powerful tools to investigate trends, compare groups or distributions, examine relationships, and reduce complex datasets, ultimately contributing to a deeper understanding of the intricate dynamics of the ocean environment.

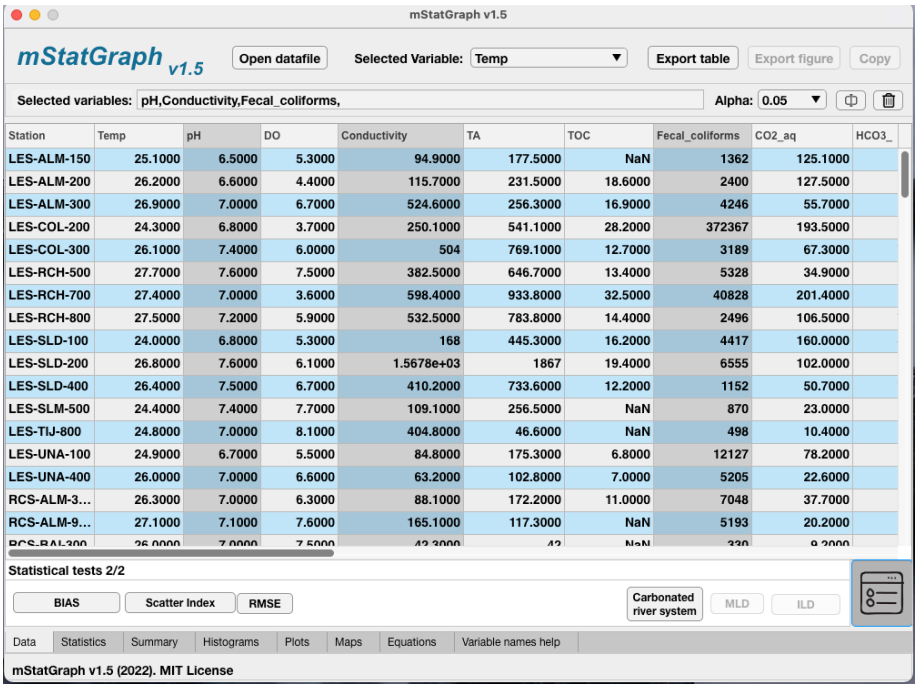
Operation mode

mStatGraph allows you to load files with ".dat" extension with comma-separated columns and the first row containing the name of the variables. The variables will be displayed in the table located in the "Data" tab. In this tab there are buttons for all the statistical tests that are implemented, as well as options for the computation of parameters of physical oceanography and the carbonate system in rivers. For each statistical test the significance level can be set (0.01, 0.05, and 0.10). The variables (columns) can be deleted with button  and renamed with button  .



The variables can be selected by clicking on each column, for more than one column it will be necessary to combine the click with the Win (command) key for Linux and Windows (MacOS) operating systems. In the editbox "Selected variables" all selected variables will appear separated by comma. The variables Longitude, Latitude, Depth, Time, and Station are not selectable.

This button appears on many of the tabs and allows you to activate more options for statistical tests, computation of parameters or to change the labels on the figures.

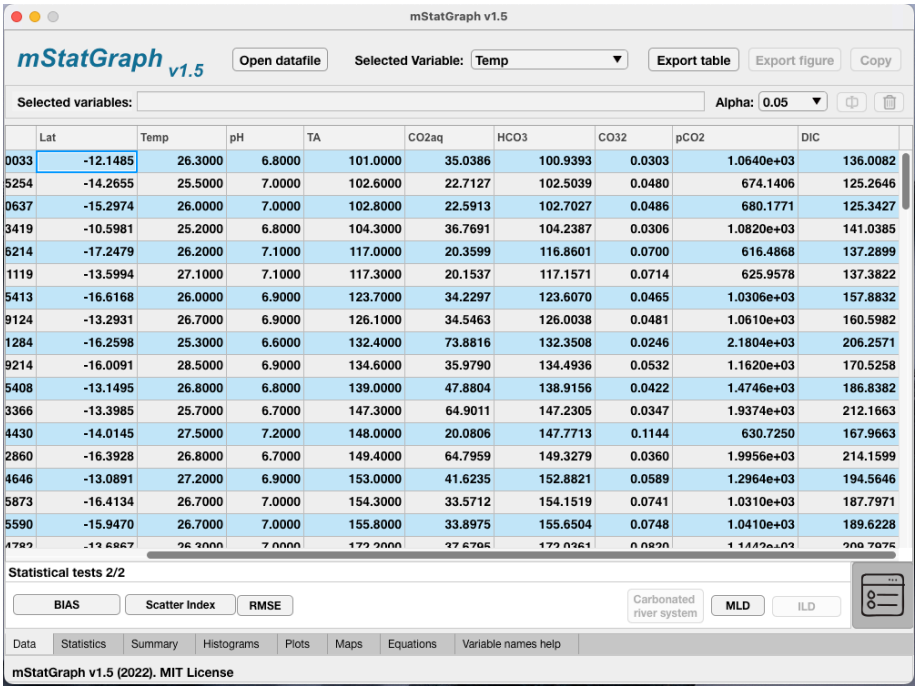


When the data is loaded, mStatGraph sorts the variables by name. In the following table we show some of the basic variables.

Variables	Description
Station	Collection station name
Lon Long Longitude	Geographic longitude
Lat Latitude	Geographic latitude
Depth Level	Depth or atmospheric level
Time Date Datetime	Time/Date of sampling or analysis or reanalysis data
Temp Tmp Pottemp Temperaure SSS	Ocean or air temperature
Salt Salinity SSS	Ocean salinity
uCurr vCurr	Velocity components of the current

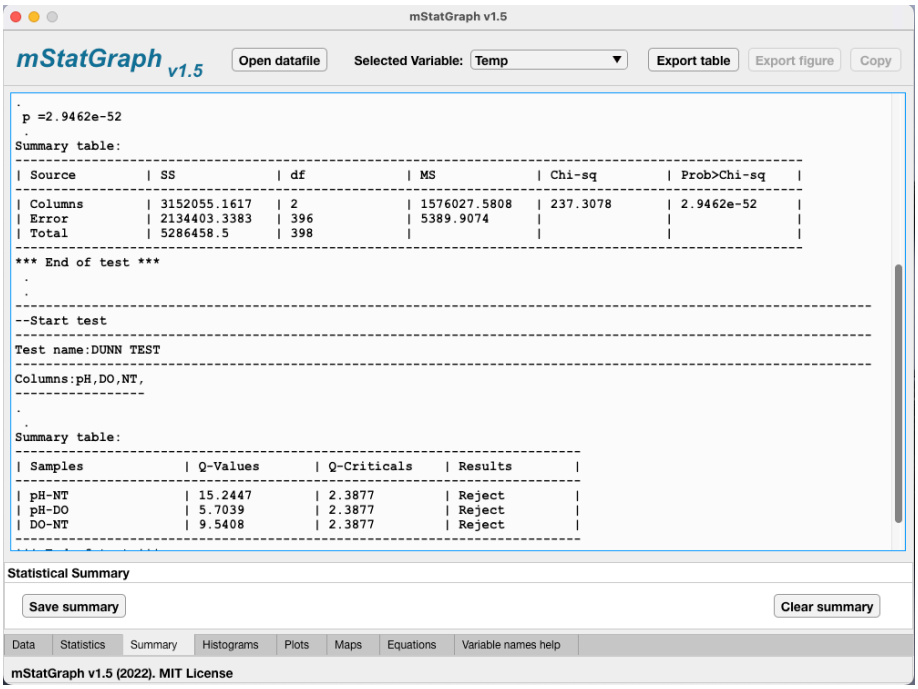
uWind vWind	Velocity components of the wind
TA	Total alkalinity of ocean or river water
pH PH	pH of ocean or river water

For example, if the file contains geographical longitude and latitude, the "Maps" tab will be activated, allowing the plotting of maps with the sampling stations or the spatial distribution of each parameter; if it contains temperature, pH, TA, the option to compute $CO_2(aq)$, HCO_3^-, CO_3^{2-}, and DIC will be activated. If the file contains data of current velocity or wind velocity components, current and wind rose figures can be plotted.

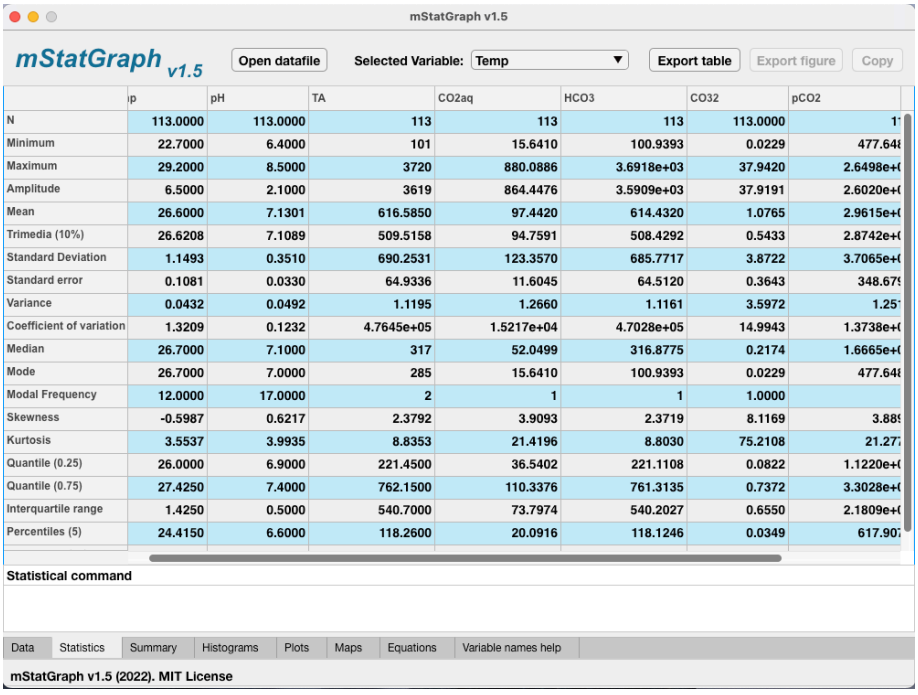


To compare two variables, at least two variables (columns) must be selected, thus automatically activating: Mann Whitney test, two-samples t-test, two-samples F-test, RMSE, Scatter index, and bias. If more than two are selected in addition to the above mentioned tests, the Kruskal Wallis test (the Dunn test will only be activated after the Kruskal-Wallis test) and ANOVA (the Tukey test will only be activated after the ANOVA test) will be activated.

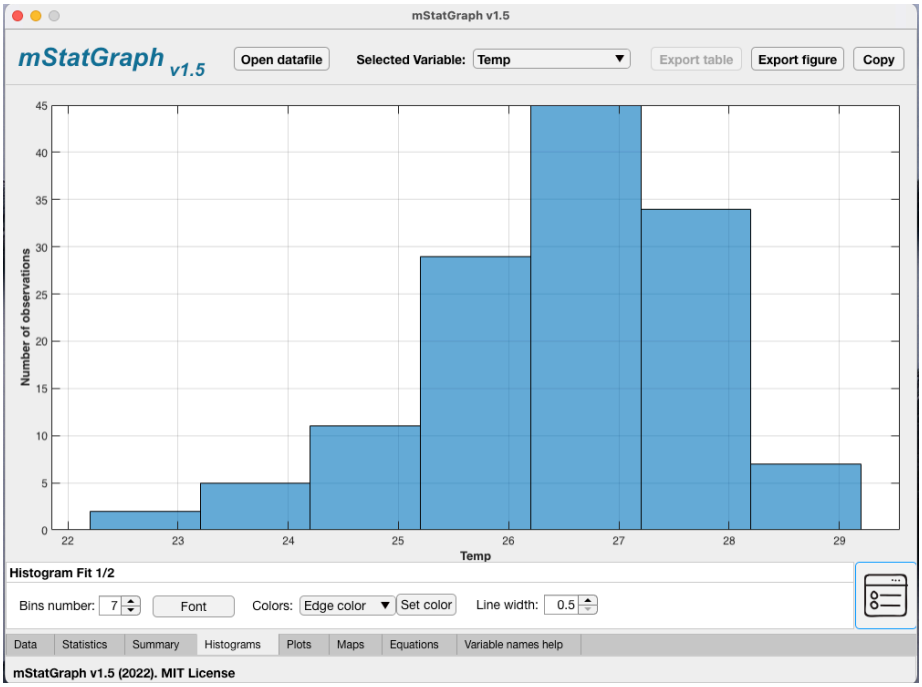
All statistical tests and computation of ocean parameters will be included in a summary found in the "Summary" tab. Summaries can be saved in a ".txt" file. The summary can be cleared using the "Clear summary" button.



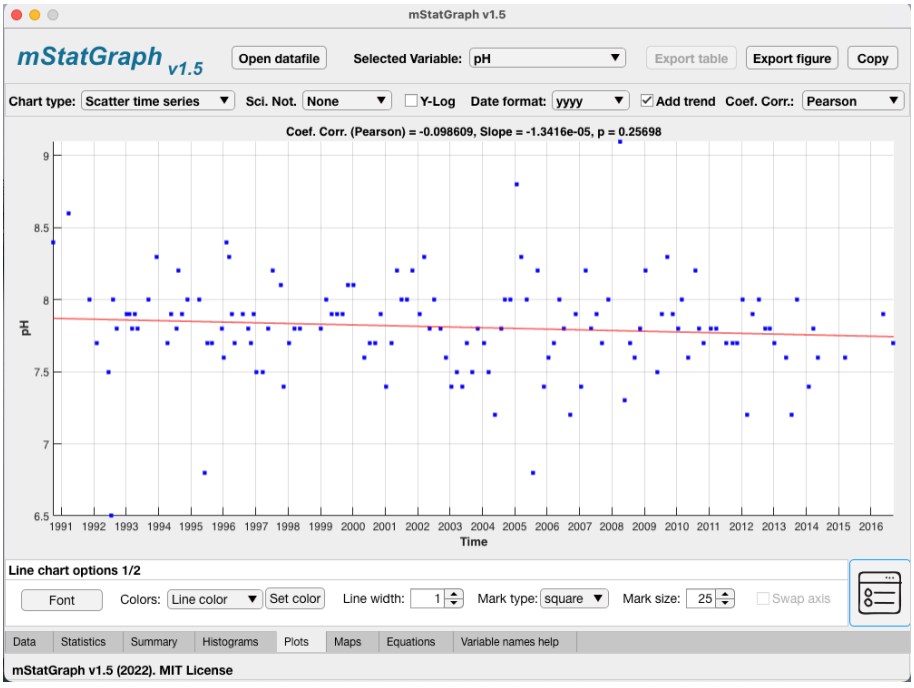
The basic statistics of all variables are computed automatically after the data file is loaded and displayed in the table under the "Statistics" tab. These statistics can be saved in a ".txt" file.



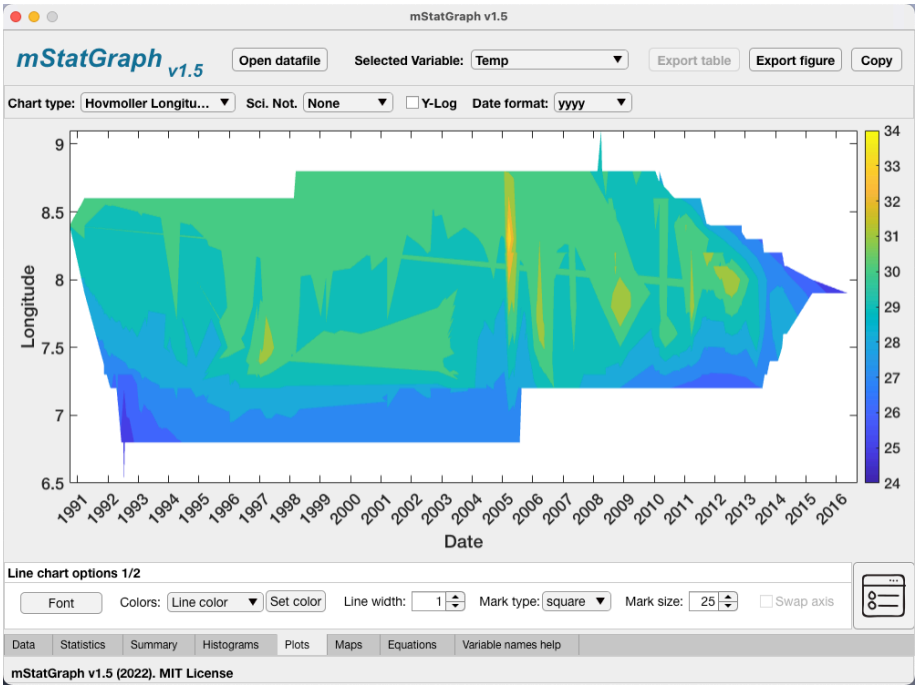
Example of a histogram (Variable selected: Temperature):



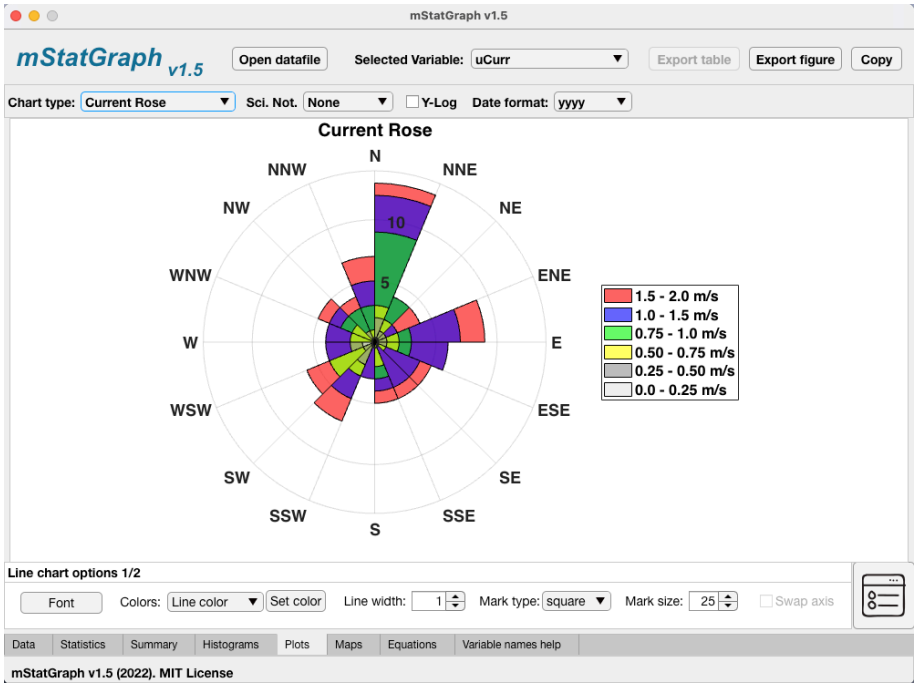
Example of a trend analysis (Variable selected: pH):



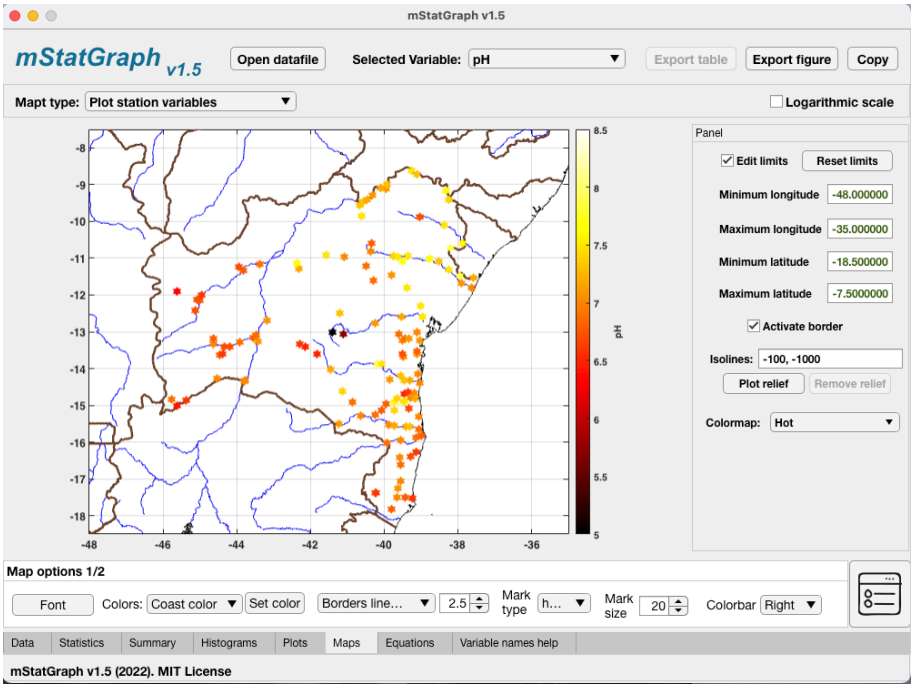
Hovmöller example (Variable selected: Temperature):



Example of marine currents rose:

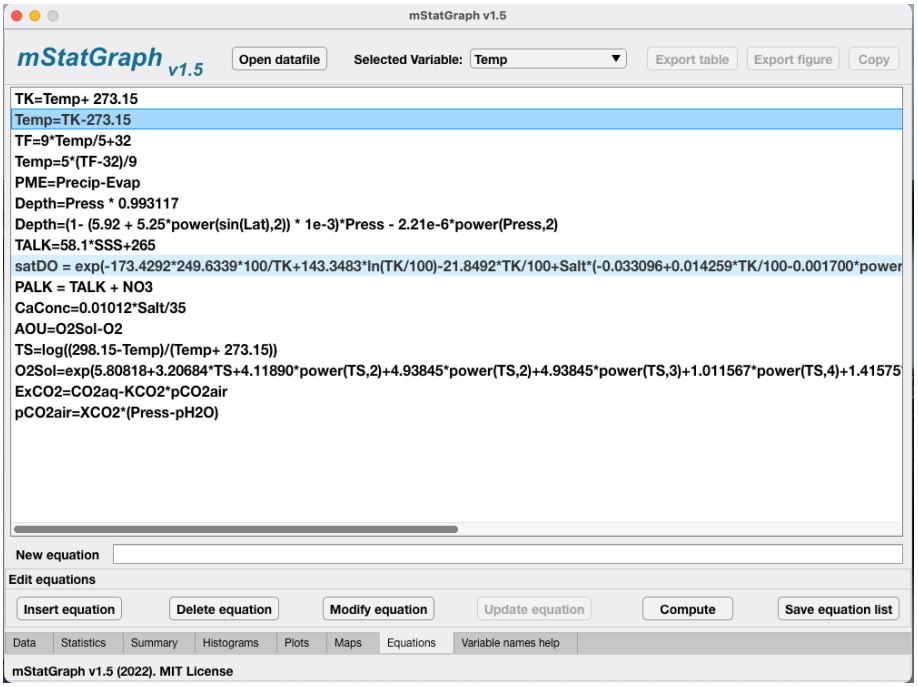


Example of a map with the spatial distribution of pH sampling in rivers:



In the "Maps" tab you can customize the maps before export by changing the geographic coordinates of the boundaries, adding state or country borders, changing the shape of the sampling point, you can add bathymetry in the case of ocean data, and you can change the colormap.

mStatGraph allows you to enter equations to calculate new parameters or make unit conversions, these equations only allow as independent variables those found in the table of the "Data" tab and you can use all the predefined functions in MATLAB 2021b. The process is simple, write the equation in the "New equation" editbox and click on the "Insert equation" button, this way a new equation will be added to the list, this list can be saved through the "Save equation list" button. To calculate any equation just select it and click on the "Compute" button, this will add a new column to the table in the "Data" tab with the variable name to the left of the "=" sign.



The last tab ("Variable name help") lists the allowed variable names so that mStatGraph can classify them.