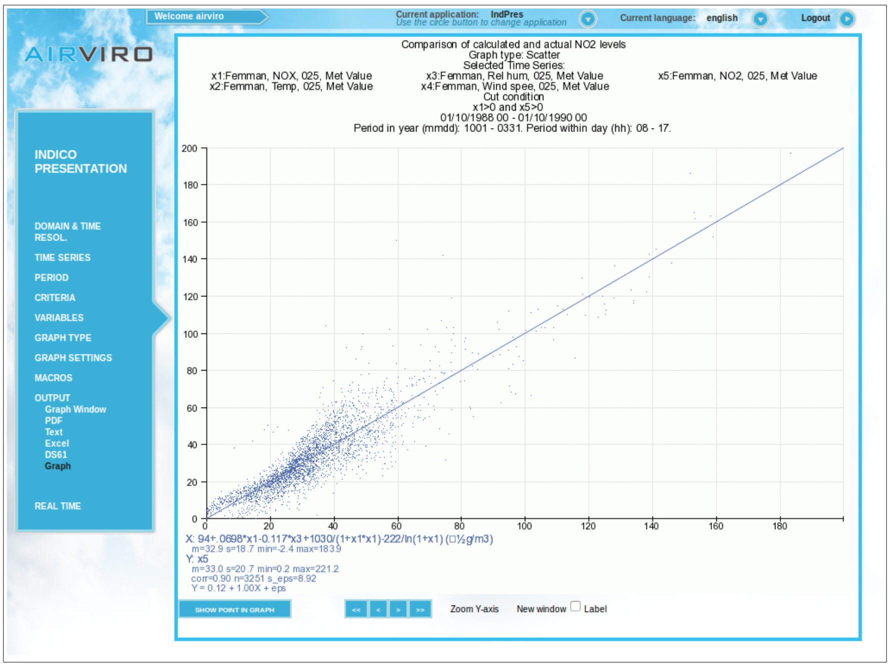
Apertum

Volume 3

Airviro User's Reference



Working with Indico Presentation

Time Series Analysis and Presentation

Time Series Analysis and Presentation

Amendments

Version	Date changed	Cause of change	Signature
3.11	September 2007	Upgrade	GS
3.12	January 2009	Upgrade	GS
3.13	January 2009	Upgrade	GS
3.20	April 2010	Upgrade	GS
3.21	December 2010	Upgrade	GS
3.21	July 2012	Review	GS
3.22	April 2013	Upgrade	GS
3.23	Nov 2014	Upgrade	GS
4.00	April 2015	New version	GS
4.00	Aug 2018	Review	GS

CONTENT

3.1 Introduction to Indico Presentation

Indico Presentation is a powerful tool for analysing data - either monitored data that have been collected automatically by the Indico Administration module, or other data imported using the Waved® or the ASCII interfaces in the system. There is also an optional Indico Real Time module, which shows a selection of the most up to date data and can run continuously of the screen keeping you informed about the latest air quality situations.

In this chapter you will find a fairly concise guide to using the various menus and subwindows followed by a number of examples and recommendations for using Indico Presentation. With some practice you will soon be a skilled user and find ways to work with your data that are more efficient than the recommendations made here. A more comprehensive guide to using the system is built into the on-line help that is provided as part of the package.

Some of the examples included here are given to show you how you can use the measured data to extend the interpretations from the simulation models of Airviro. All the examples included here are based on the Airviro (Göteborg) Reference Domain, included in all delivered Airviro systems.

3.1.1 What is Indico Presentation?

Indico Presentation is a powerful tool for presenting and analysing data in the time series database. With Indico Presentation, you can:

- Select one or more time series – measured, simulated or forecasted – for processing.
- Assess capture and status of the data.
- Constrain observations from further processing.
- Handle missing values by interpolation.

- Transform variables by computing, counting or recoding into categories.
- Find or eliminate trend and seasonal components by smoothing or differencing.
- Monitor diurnal, weekly and yearly variation.
- Plot time series data in a line chart, histogram or frequency distribution.
- Plot pairs of variables in scatter plots or polar diagrams.
- Fit a curve to pairs of variables.
- Set up a linear or binary logistic regression model to estimate concentrations or chemical reactions.
- Apply factor analysis or principal component analysis to structure data and avoid co-linearity.
- Automate the production by using macros.
- Automatically update diagrams as new data arrive.

When you become an experienced Indico user you will be able to use the Airviro system as an *integrated monitoring system*, i.e. extracting valuable information from the measured data and adding this information to the Airviro simulation models.

3.1.2 How does Indico Presentation client work?

Airviro has is a web based user interface. Airviro can be used from a PC or any other device running Internet Explorer 6 or later and Firefox.

After logging in on Airviro the Indico Presentation module can be selected. All data processing is made on the Airviro server and the results are transferred to the web browser.

Please note that JAVA JRE (run time plugin) must be installed and enabled in the web browser.

3.2 Getting started

Once Airviro has been properly installed on the *server*, you can begin using it by typing the correct URL in your web browser over Intranet/Internet. The web interface give you access to all Airviro modules, including **Indico Presentation**, Indico Administration, Indico Report, and Indico Validation among others.

After logging in with user-ID and password, the user can select a domain and choose a module.

In **Current application** you can see the actual Module that has been selected and you can use the circle button to choose another Module.

Click on the circle button besides **Current language** *to* select the working language. In version 4.0 this option is not available yet. With **Logout** you can close the airviro windows and return to the login page.

3.2.1 Overview of the Indico Presentation main window

When Indico Presentation has been selected, the user gets a list of available frames or submenus on the left-hand side. A complete setup includes working through all submenus from the Domain & Time resolution down to Output, excluding Macros and Real Time, which are previously defined automation objects. It is preferable to work through the submenus sequentially, because some settings may depend on earlier choices, e.g. settings in Criteria is done in terms of definitions made under Time Series, entries in Graph Settings refers to definitions in Variables etc. Submenus can be hidden or expanded by clicking on the associated frame.

3.2.2 Time series database

The time series database for a certain domain may contain a large number of measuring stations and parameters. The parameters can be related to mass concentrations of pollutants or meteorological data, traffic intensities, instrument readings of other kinds or quality control data from data loggers. For each parameter there is also a quality flag.

Data may arrive each second into a raw database to be filtered and condensed to half-hour means, hourly means or daily means in a continuous process. Mean values, peak values and standard deviations may be calculated and stored in the process or at the data logger.

Data may on the other hand arrive once per year from some other source to be imported with Indico Administration into the time series database.

Time series data can also be generated by the postprocessor menu option in Dispersion or by statistical forecasting in Aircast or from some meteorological agency.

All these data are gathered and organised into the time series database. In order to view or analyse the data, Indico Presentation has to select one or a few series.

3.2.3 Selecting time series

In the **DOMAIN & TIME RESOL.** frame, the user can select a domain and a time resolution to work with. It is not necessary to select the same domain as in the login process. If you change to another domain, you will get other time series and other macros related to the current domain. See *Figure 3.2.3.1.*

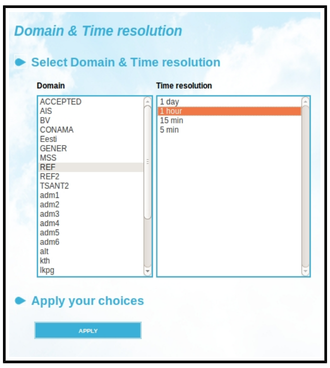


Figure 3.2.3.1 Indico Presentation window with available domains and time resolutions.

It is possible to work simultaneously with many instances of Indico Presentation. Many users can work on the same domain and one user can work on different domains or time resolutions without any substantial risk for interference.

In the **TIME SERIES** frame, you will see a list of all stations - active or inactive - in the station database and all observed parameters in the parameter database, regardless of which station that observes them. If you select one station by clicking on its name in the station list box, you will get a list of which parameters it observes. Clicking **CLEAR** releases the selection and creates a new list of all parameters in the parameter list box. If you, on the other hand, are interested in all stations that measure a certain parameter, start by clicking on the parameter in the parameter list box.

It is possible to sort the stations or the parameters in the list box alphabetically or by station key or parameter key by selecting sort key in the associated drop-down list box. Checking/unchecking **Reverse** rehashes the sort order accordingly. It is also possible to promote active stations by moving them to the top of the station list box by checking **Active first**. Sorting stations also by reverse death time creates a list of increasingly older stations. Click **CLEAR** to get a full list of stations. You may notice that some stations are preceded by an asterisk (*). This is to show that they are operational stations, i.e. they collect data automatically.

When you have selected both station and parameter, you will get a list of available instances of the actual parameter. The instance is used to differentiate between simultaneous measurements of the same parameter at the same site, e.g. if you measure at different levels above ground or if you are using more instruments or analytical functions to get an output.

A letter is shown in square brackets immediately following the instance. The letter is a code for parameter type. Letter M or Q indicates that you store a measured value and a status flag for the actual instance. M is used in a work database and Q is used in a validated database. Letter K or W indicates that you also store a peak value. Letter O or P indicates that you store standard deviation and light intensity (for DOAS analysers). Other letters may occur. See available variables in the attribute list box.

The status flag is assigned in the quality control in Indico Administration. For a table of status conditions, see under **CRITERIA >> STATUS CONDITIONS**.

Figure 3.2.3.2 Time Series frame with available stations and parameters in the REF database with 1 hour resolution. Up to 64 time series can be selected for further processing.

When you have clicked on station, parameter, instance and attribute, the time series is uniquely identified (for the current time resolution). Click **NEW** to select the time series for further processing. You can select up to 64 time series for further processing. If you click **CONTENTS**, a graph will appear to present data capture for selected variables during some period. If you let the haircross rest on the graph, you can read date and value in a label.

Please remember that the variables are numbered according to the order in which they are listed in the “Selected” list box. You can remove a highlighted selected variable from the list box with the **REMOVE SELECTED** button or replace it by identifying another time series and clicking **REPLACE**. If you click **CLEAR ALL**, all variables will be removed from the “Selected” list box.

Up to four variables at a time can be presented in a graph. When you are satisfied with your time series selections, click **APPLY** to save your settings.

3.2.4 Selecting cases

By default, Indico Presentation selects the last week of data. In the **Period** frame, you can change into any time period. Date and time can be presented in European, UK or US date format. If you want to set another start date, you can write the date in the **From** box. Alternatively, you can use the double arrows between the From and To boxes to transfer a date between the boxes. You can also use the double arrows adjacent to Year, Month, Week, Day or Hour to step forwards or backwards with one such time step. You can always reset the **To** box to present time by clicking **PRESENT**.

When you have selected a period, click **APPLY**. If you want to check data capture for the new period, go to Time series and click **CONTENTS** again. If you aren’t satisfied with the data capture, try setting another period with the arrows. It is possible that you have to change time resolution as well.

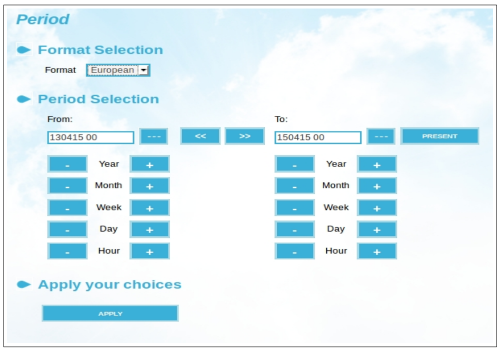


Figure 3.2.4.1 Indico Presentation Period frame with start date and end date in European format in the REF database with 1 hour resolution. There is no guarantee that data exist in the selected time period.

The hour starting at 00:00 and ending at 01:00 is named 01. The hours are numbered from 01 to 24. This means that the hour starting 23:00 and ending 00:00 is named 24. Please make sure that all time references in the system refers to the same time system, otherwise you may get in trouble with missing data or duplicate data each day. You can also get trouble with time series being out-of sync.

To get a time series, you would probably need at least two observations. A period starting at 01:00 and ending at 03:00 is two hours long, containing hourly observations named 02 and 03. You cannot specify minutes in the period frame; they are always 00 here, which can be slightly confusing. If you specify one day in a time series with 15-minute resolution, the observations are numbered from 0015 to 2360 in 15-minute steps. In principle, however, Indico Presentation is based on hourly elements.

The date and time given is inclusive in the From box and exclusive in the To box, or in mathematical terms:

$t \in [\ t_{From},.., t_{To}\ ]$

Also, you can input the date and time by pressing the button besides the date text box. Clicking it opens a calendar. *Figure 3.2.4.1*

3.2.5 Constraining cases

The selected variables form a set of observations that are simultaneous, but not necessarily from the same station. You can examine subsets of the cases by constraining access to data in various ways in the **CRITERIA** frame.

If you for instance want to study winter conditions first, you can limit the **period within the years** to only winter months (Dec, 1 to Feb, 29). Later on, you could make a new criterion for the opposite period (Mar, 1 to Nov, 31).

It is also possible to constrain by hours. The hours are accessed by their name, all-inclusive. If you want observations from 22:00:00 until 22:59:59, specify that you want the **period within day** from 23 to 23. If you want all observations except that hour, specify that you want the period from 24 to 22. See *Figure 3.2.5.1.*

You can take a look at the data you have selected by clicking **APPLY** and send the **OUTPUT** to a **Text** file, available from the menu on the left-hand side. It is instructive to look at 15minute data in this way. All criteria related to date and time or days tend to exclude cases entirely from further processing, i.e. cases that don't match are left out from the time series.

You may want to study **Weekdays** separately. You can choose to study Mondays – Thursdays as one class. If you are precise, you may want to exclude national holidays occurring on a weekday. This can be done in **Day type**. The national holidays are specified in the resource file calendar.rf during installation of Airviro. Use **man calendar.rf** at the server for a closer explanation of day types.

You can apply some constraint related to the observed data, e.g. to study only cases with low wind speed, low temperature or some other condition. The constraint should be written into the text box "**Condition Selection**". The variables are named x1 to x64 according to how the time series were entered in the "Selected" list box, see *Figure 3.2.5.1.* You can compare your variables with constants or expressions in a Boolean formula. Statements can be combined with logical operators OR(|), AND(&) or NOT(!). A quick reference of available functions can be found in the **VARIABLES** frame, following the **Help** link. See further chapter *3.3 Math expression compiler*.

Please make sure that you set criteria in the correct unit. If you have doubts about what unit is used, compare with the parameter database, which is available from Indico Administration. Alternatively, you can modify the expression by using arithmetic functions in your condition formula. See *Figure 3.2.5.1* below for an example of how to write a condition formula. If the Boolean expression is false, all variables in that case will be left without a value.

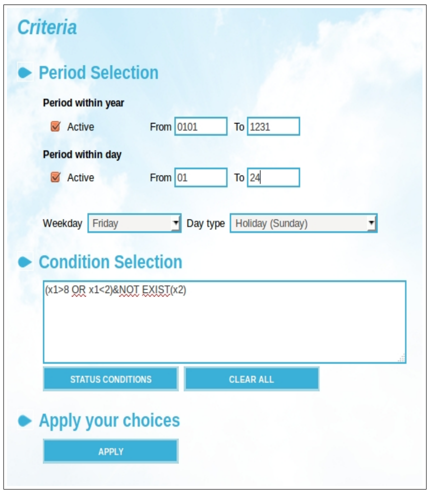


Figure 3.2.5.1 Indico Presentation Criteria frame with some criteria for constraining access to the selected cases.

Every data value is given a status condition, either by Indico Administration or by the external protocol. The status condition refers to absence of readings, readings below the detection limit or above the upper limit of measurement, similar values for a long time that cannot be distinguished within the measurement resolution, rapid changes or other error conditions.

You can constrain your time series to only use data associated with certain status conditions by selecting a set of status codes. This is useful in some analyses if data are bad. Click **STATUS CONDITIONS** to see or select among the status codes. Data that are "Checked – OK" or "Manually changed" are selected by default, but can be unselected. Click **APPLY** to commit your status conditions.

Variables that don't meet the specified status conditions will be left without a value, based on the individual reading.

The **CLEAR ALL** button will reset all criteria, including those already committed under Status conditions.

3.3 Math expression compiler

Once you have defined a number of time series and variables, you can work with the data to prepare it for presenting in a graph or for statistical analysis. You can work separately with plot variables and statistical variables, using the same time series.

The goal for working with data should be to extract as much information as possible from the measurements. This can be achieved by presenting descriptive statistics, exceedance statistics, distribution functions, correlation between stations; time series analysis for seasonal variation and trend using univariate or multivariate techniques in an attempt to improve the knowledge about the situation. With this knowledge, you can work systematically to get an understanding of the air quality situation, how it is related to different source areas, how the concentration varies by time of day or season or by meteorological factors to build a model that explains the variation.

In parallel, you can work with models in an attempt to describe the known emissions and their effect on the concentrations at a receptor. It is quite possible to use the measurements for assessing the quality of the emission database and to find emission areas where the quality of data has to be improved. This is done with inverse air pollution modeling.

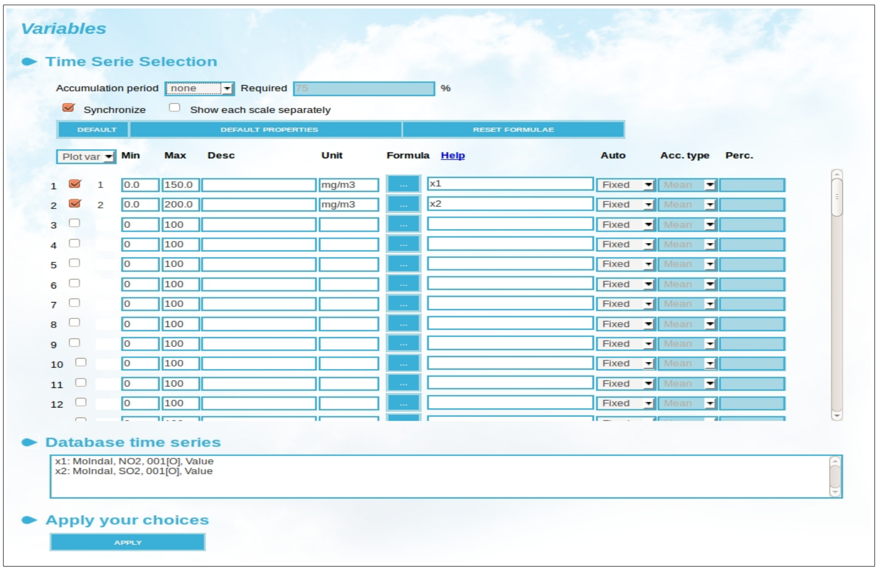


Figure 3.3.1 Indico Presentation Variables frame with three plot variables for two series.

It is a cumbersome work to reach this level, and it includes many steps along the way to build up a system of this kind.

Synchronize has to do with accumulation. Accumulation is like creating a new time resolution. The Accumulation Period setting affects the data that will be accumulated from the database.

For instance, if you select accumulation for daily values and synchronized is checked, each value will be integrated over one full day, from midnight to midnight. If it is not checked, it could be any 24-hour period, depending on the start time. The user can define the percentage of data that must at least be included for this calculate. The text box **Required** is in **%**.

Checking **Show each scale separately** marks each parameter’s scale with the same color as the line colour of the corresponding time series. Leaving it unchecked will group times series with the same scale together. Which scale that belongs to which time series will be indicated above each scale.

For this purpose, Indico Presentation has many tools to help along the way. Take a look at the mathematical functions; fill functions; arithmetic functions, relational, logical and other functions available for working with time series data.

In the **Variables** frame, you see a list of all time series that were selected; compare with the bottom text box in *Figure 3.3.1*. If you didn’t check the box **Keep settings for variables** under **Time series**, you will get a simple list of variables x1, x2,…, x64 with their associated units of measurement and default min-max values from the parameter database. You can change the min-max values, the associated unit of measurement or the formula as you wish. This will have an effect on plotting of values. The **min-max** values refer to the scale in a plot; **unit** refers to the text on the ordinate axis. You can change the order in which variables are plotted, by unchecking them all and then check a new plot order with up to four variables in one graph. A new variable can be defined (variable 3, plot variable 1 in the Figure) by entering a function in the formula field next to the variable.

You can enter arithmetic or conditional Boolean expressions in the formula field. Examples of an arithmetic expressions are x1*1000 or emax(x1:x5). An example of a conditional Boolean expression is (x1>0.65)?x1:@. The expression preceding the question mark is Boolean. If it is true, the plot variable gets the value x1, otherwise it gets the value @, a special sign for not-a-number. It is possible to copy expressions with Ctrl-C (copy) and Ctrl-V (paste) between formula fields or other electronic documents. Long expressions will scroll horizontally.

For a full list of functions and operators, see *Appendix 3A* and *Appendix E4 Calculation Formulae* in *Airviro Specification, part II.* You can also look at the quick reference help following the **Help** link.

There are three buttons that are used to reset your settings to their default values. The **DEFAULT** button resets the number of variables and the plot order to what was selected under Time series. The **DEFAULT PROPERTIES** button resets min-max values and unit to the values defined in the parameter database. The **RESET FORMULAE** button resets the formula to the variable itself.

In the **Auto max/ min** box you can change the scale in a plot: fixed, max auto or auto. These options have the following meanings. **Fixed**: the maximum and minimum values are the same of the box max/min. **MaxAuto**: the system automatically adjusts the maximum variable. **Auto** the system automatically adjust the minimum and maximum variable.

In the **Acc.type** box you can select options to calculate accumulation. These are mean (sum of all the observation values ÷ number of observations) , min (the minimum value between the hours n), max (the minimum value between the hours n), sum (sum accumulated from the n hours backward), n°values (number considered for each value) and perc n.(percentile).

Order statistics provide a way of estimating proportions of the data that should fall above and below a given value, called a percentile. The pth percentile is a value, Y(p), such that at most (100p)% of the measurements are less than this value and at most 100(1- p)% are greater. The 50th percentile is called the median. (median)

Percentiles split a set of ordered data into hundredths. For example, 70% of the data should fall below the 70th percentile

3.3.1 Transforming variables

You can transform a variable by computing or recoding into categories. If you want to change unit from µg/m^3 into ppb(v), you need to know air density as a function of temperature (+pressure and moisture), the molecular mass of air (28.97u) and the molecular mass M of the substance.

1) By computing, the volume ratio in ppb(v) is then

$$\text{var} = x1 * 28.97 / (M * 1.2929) * \frac{x2 + 13.273}{273.13} ,$$

if x1 is mass concentration in µg/m^3 and x2 is temperature in °C.

Another example is to use a regression model for some transformation process, e.g. from NO_X to NO_2, if it is validated.

2) If you want do divide the material into groups, you have number the groups by some instructive value - the mean value or some code, e.g. Beaufort scale. You can use the conditional operator **?:** to recode the variable.

```
var = (x1>0&x1<0.25)?0:(x1<1.55?1:(x1<3.35?2:(x1<5.45?3:(x1<7.95?4:(x1<10.75?5:@))))),
```

if x1 is wind speed in m/s. The scale continues up to 32.7 m/s, which is 12 Beaufort. Missing values can be coded with not-a-number.

If you want to divide source areas into sectors for different stations, you can define the upper and lower wind direction limits for a source sector - for each station - with this recode function.

Other examples can be to divide the atmospheric stability into stability classes or construct a ventilation index for wind speed and boundary layer height.

In a simplified way, the above division into groups can be accomplished with the descriptor function. The **desc** function returns values {1,2,3,…,n} if $x < \{l_1,l_2,l_3,\dots,l_n\}$.

var = desc(x1,0.25,1.55,3.35,5.45,7.95,10.75,13.85,17.15,20.75,24.45,28.45,32.65,50)

3) Another way to recode values is to interpret concentrations as an index, using piece-wise linear functions. The US EPA has defined a Pollutant Standards Index, ranging from 0 to 500 for five different substances.

The index is Good (<50), Moderate (<100), Unhealthy for sensitive groups (<150), Unhealthy (<200), Very unhealthy (<300) or Hazardous (>300).

For sulfur dioxide, the breakpoints are 0.035 ppm, 0.145 ppm, 0.225 ppm, 0.305 ppm, 0.605 ppm and 1.005 respectively for hourly values.

This is recoded as:

var = API(x1,0.035,50, 0.145,100, 0.225,150, 0.305,200, 0.605,300,1.005,500),

if x1 is the hourly concentration in ppm, otherwise the concentration has to be computed first. The formula for concentration in ppm can be written in the first position of API.

For more information about the **API** function, see *Appendix 3A* and *Appendix E4.6 The Air Pollution Index in Airviro Specification, part II*.

3.3.1.1 Handling missing values

Some statistical analyses require that all values in a time series are present. If this is not the case, you can use the math expression compiler to estimate missing values, if they are not too many.

There are three built-in functions that can be applied to the time series to fill missing values with a guessed value. The fill functions are quite simple; you define the variable and the size of the filter.

Sustain(x,n) fills in missing values by copying the nearest previous value. The function requires that at least one value before the current time is within n time steps.

Interpol(x,n) fills in missing values by linear interpolation of the nearest valid surrounding values. The function requires that at least one value before and one value after the current time is within n time steps.

Interps(x,n) fills in missing values by linear interpolation of the nearest valid surrounding values. If there is only one value before or one value after the current time within n time steps, the function copies that value. The function requires that at least one value before or one value after the current time is within n time steps.

Apart from these functions, it is possible to use the centered moving average function **eaver**(x1[-1], x1,x1[1]) to get a continuous series. Other moving averages can be defined with different size and lag.

It is also possible, under stationary conditions, to define an autoregressive univariate model, which can be fitted with stepwise regression. These functions can be invoked for missing values using the **exist**(x) function and a conditional statement.

It is in principle possible - but complicated - to use differencing methods and mathematical functions to define a seasonal Box-Jenkins model, which can be invoked for missing values. This will give the best estimate for missing values, including the stochastic error of the time series.

3.3.1.2 Counting

If you want to count exceedances for an observation, you can use the **reep** function. Simply fetch all your monitored channels into the variables x1..x64 and compare the observations with some threshold value.

var = reep(x1,110)+reep(x2,40)+reep(x3,0.5)…,

where x1, x2, x3 are three different substances that are compared with a guideline value. For each time step, you will get the number of exceedances as a value.

It is more complicated if you want to count status conditions. If you want to check how many observations that fall below the detection limit, you could check status flag 4.

var = (x1==4?1:0)+(x2==4?1:0)+(x3==4?1:0)…,

where x1, x2, x3 are status codes for three observations. For each time step you get the number of undetectable concentrations.

f you want to calculate the total sum of some variable during a period, you can set the environment variable **INDICO_SUM** at the server. When **INDICO_SUM** is set, the total sum of a series will be plotted together with other descriptive statistics below the graph.

If you, on the other hand, measure some flow - traffic or emission rate - in vehicles/h or kg/h, you can specify an integration unit in the environment variable **INDICO_INT**. The integration unit should be expressed in seconds, followed by a blank and the unit, e.g. “3600 h” or “86400 d” for one hour or one day respectively (corresponding to the rate unit).

This will print the total number of vehicles or the total emission during the examined period, together with other descriptive statistics below the graph.

The mentioned environment variables can be set by the Airviro system administrator during the AIRVIRO installation or after a user request.

3.3.2 Creating new variables

Plot variables and statistical variables that are defined with a formula don’t have a name; they are only referred to by their formula or by plot order in a graph.

If you want to use a complex variable in another formula, you have to include the whole expression in the new formula. There are occasions when you would prefer to use a short name for the complex variable, e.g. if it is part of a polynomial, where the complex variable is used repeatedly.

You should avoid as long as possible to store new variables for purpose of analysing, but if it is absolutely necessary, you can export the variable and import it to the time series database as a new parameter or instance.

If you decide to do this, please make sure that you save the formula in a macro for future reference, see section 3.7 for information about using macros. You have to define the new time series in Indico Administration to allow for import, if it isn’t already defined. Use a prefix like ‘mod’ to indicate that the parameter isn’t directly measured, i.e. modNO2. Alternatively, you can use a new instance or a dummy variable.

Exporting can be done in ASCII format by sending the output to a text file. Don’t forget to set or export any available status codes or other additional attributes.

You can use Waved® (optional excel interface to Indico Module) in your PC or a script at the server to import the new time series into the time series database. Ask your Airviro system administrator for help.

Later on, you can apply the macro to another period, if the conditions still are valid.

3.3.3 Modifying a series by smoothing or differencing

When you analyse a time series, you should always plot the data first. If there are discontinuities in the series, it should be broken into homogeneous sequences.

You may find that the data can be decomposed into a trend component, a seasonal component and a stationary random noise component. If that is the case, you may want to estimate the trend and the seasonal variation. The trend doesn’t have to be linear.

The trend can be estimated by applying a moving average filter chosen to eliminate the seasonal component and to dampen noise. If the period is even, say 24, you can use a centered moving average like:

Trend= (0.5*x1[-12]+x1[-11]+x1[-10]+…+x1[-1]+x1+x1[1]+…+x1[11]+0.5*x1[12]) / 24.

If the period is odd, you can use a simple centered moving average for smoothing.

This is a low-pass filter that attenuates noise but allows linear trend functions to pass without distortion.

By clever choice of weights, you can design a filter which is effective in attenuating noise and also allows a larger class of trend functions to pass undistorted through the filter. See further in Kendall and Stuart, The advanced theory of statistics, Volume 3, chapter 46: Trend and seasonality. One example of a filter that allows polynomials up to fourth order to pass without distortion is the Spencer 21-point formula:

$$\text{Trend} = \sum_{i=-10}^{10} a_i X_{t+i} \text{ , where}$$

$$[a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}] = \frac{1}{350} * [,60,57,47,33,19,6, -,2 -,5 -,5 -,3 -1]$$

The second step is to estimate the seasonal component. To do this, you have to identify which phase you are observing. One way to do this is to number the observations in an absolute series, related to date and time. When you have done that, you can select one phase at a time and compute the average deviations from the trend. If the sum of average deviations for all phases differ from zero, the seasonal component should be corrected by subtracting the normalized deviation. Finally, the trend is re-estimated by subtracting the seasonal component from the series and by applying a moving average as above.

Another method, which doesn't require an absolute date and time, is to apply a difference operator

$$\nabla x1 = \{x1 - x1[-1]\} = (1 - B)x1,$$

where B is a backward shift operator. The difference operator and backward shift operator can be applied repeatedly as a polynomial to eliminate the trend term by differencing. If you have seasonal data, you can introduce a lag-difference operator

$$\nabla_d\, x1 = \{x1 - x1[-d]\} = (1 - B^d)x1$$

to eliminate the seasonal and the trend term by repeated differencing.

3.4 Presenting graphs

In the **Graph Type** frame, you can select between nine different presentation types and five different analysis types.

The presentation types are:

- Time series graph
- Filled time series:
- Bar chart
- Frequency distribution graph
- Scatter plot
- Polar plots: Breuer, Mean/sector and Freq/sector

Seasonal variation charts: Diurnal, Weekly and Annual The different charts will be explained in section 3.4.3 according the list above.

For the statistical analysis types, see chapter *3.5 Regression modeling* or chapter *3.6 Factor analysis*.

In the Graph type frame, you can also see a list of the time series that you have already selected. It is presented there for your convenience, because in some presentation or analysis types, you have to specify which variable that is dependent. The dependent variable may also refer to a formula from the Variables frame.

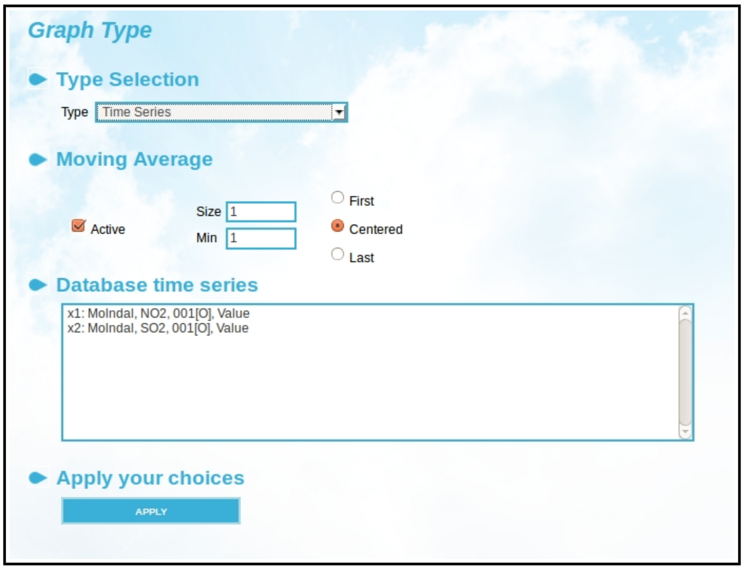


Figure 3.4.1 Graph type.

For some presentation types, a sub-frame will appear to prompt for scatter or polar settings.

If you haven't already defined a moving average with a formula, you can apply it from the Graph type frame. Please be aware that if you apply it from here, you cannot decide - by looking at the text on the graph - if it has been applied or not.

The **Moving Average** defined in the Graph type frame is slightly different, since you can specify the number of required time steps in the **Min** field. It is up to your personal preferences if you want to apply moving averages and from where it should be done. Some smoothing methods imply the use of repeated moving averages or median values. If you don't want to apply a moving average, make sure that the **Size** of the filter is 1 time step, otherwise it might be applied by mistake. The function is activated by checking the **Active** check box. If the function is activated, it will be applied on all plot variables.

Click **APPLY** to commit your settings in Graph Type.

3.4.1 Controlling the layout of a graph

In the **Graph Settings** frame, you can specify a heading in the **Graph Title** field. You can change the line style or marker style for each plot variable in a chart. It is allowed to present up to twelve plot variables in a chart. You can set line type, line width and line colors.

First, select the plot variable to under **Plots**. The topmost option button refers to plot variable 1, the next one refers to variable 2 etc. When you set line or marker type, width and color, the image next to the selected option button will change appearance according to your settings. When you are satisfied with your settings for the first variable, continue with the next plot variable by selecting another option button.

If you want to plot individual observations, you should select a marker. Single observations or observations surrounded by missing values are otherwise invisible in the plot, since a line requires at least two observations to be drawn. If you select the large marker, you can use line width to change its size. The small dot is not scalable.

As a general rule, you can decide if missing values should be left blank in a graph, or if the line should interpolate between existing values. This is selected with the **Missing values** option button.

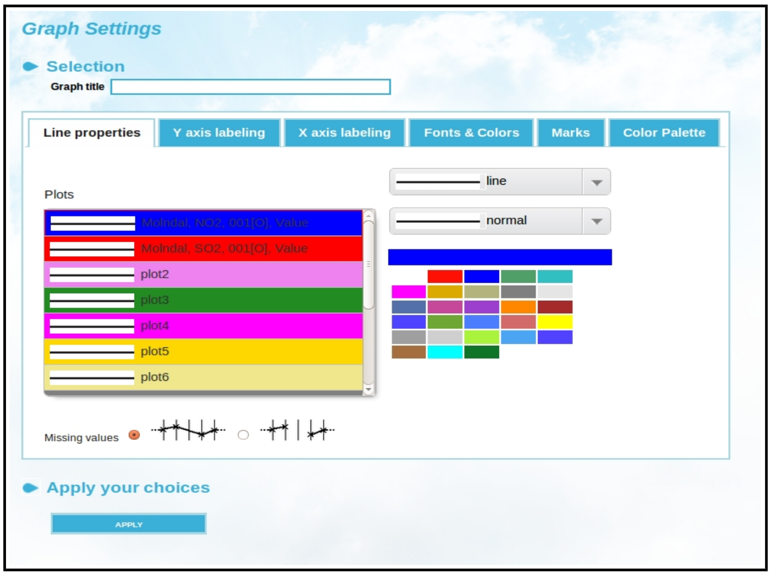


Figure 3.4.1.1 Graph Settings frame with title, layout and line properties.

The **Y-axis labelling** window let's you choose between automatic or manual scale. If manual is selected it is possible to define the levels and their labels for the y- axis. Labels can be specified for the levels themselves (with or without a number) and for the intervals between the levels. You can enter up to 31 levels for the y-axis.

The **X-axis labelling** window let's you divide the x-axis in a number of intervals. A label can be specified for each interval. You can enter up to 32 levels for the x-axis.

In **Fonts & Colors** you set these features for different parts of the graph. **Background** will display the background with the color chosen in the color palette. **Background graph** will display the background graph with the color chosen in the color palette, and so on.

Horizontal/vertical stripes are drawn instead of straight horizontal/vertical lines if **Horizontal/Vertical stripes** are selected. The **Horizontal/vertical lines** are lines divisions on each axis.

If you want information in a subtitle about selected time series, criteria and graph type, you have to check the **Header** box. If you want a description that explains the plot variables, check the **Footers** box. For descriptive statistics in the footnote, check the **Statistics** box.

Frame is the perimeter of the graph area. You can select the color. If you want an ordinate axis and abscissa, check the X and Y **Axis** boxes.

Further to the layout, you can add reference levels as horizontal lines for the time series and seasonal variation charts. First you have to choose which plot variable you will associate your reference levels with. This is done in the **Reference** drop-down list. Next, you can enter up to four reference levels – **Marks**. You have to select a mark in the adjacent check-box to activate the reference line in your chart.

It is possible to change the **colour palette** for each project

3.4.2 Displaying the graph

The graph can be displayed in a new window by selecting **Output – Graph Window** in the left-hand menu or be shown in the working area using **Outout – Graph**. See *Figure 3.4.2.1* below.

The graph is interactive so that you can **zoom** the ordinate axis or the abscissa to change the content of the graph. The zoomed graph can be opened in a new window, if you want to keep the original min-max and from-to settings.

If you let the hair cross rest on the graph, you can read date and value on a label. Alternatively, you can read date and value in a text area that appears when you click **SHOW POINT IN GRAPH**.

With **Zoom Out** you can zoom out to a time period that is the double of the one shown. To zoom in itn time, place the hair cross in the graph to the start of the period that you want to zoom in to, left click and drag the mouse to stretch a rectangle. Release the button and a new graph with the selcted time period is displayed.

Also, you can use the **arrows** to move the graph forward or backward in time. The simple arrow is used to move the shown time period half a period forward or backward. The double arrow moves the time period a whole time period.

If you want to create a high-quality graph suitable for printing , you can write it in Adobe® PDF-format by selecting **Output – PDF** in the left-hand menu.

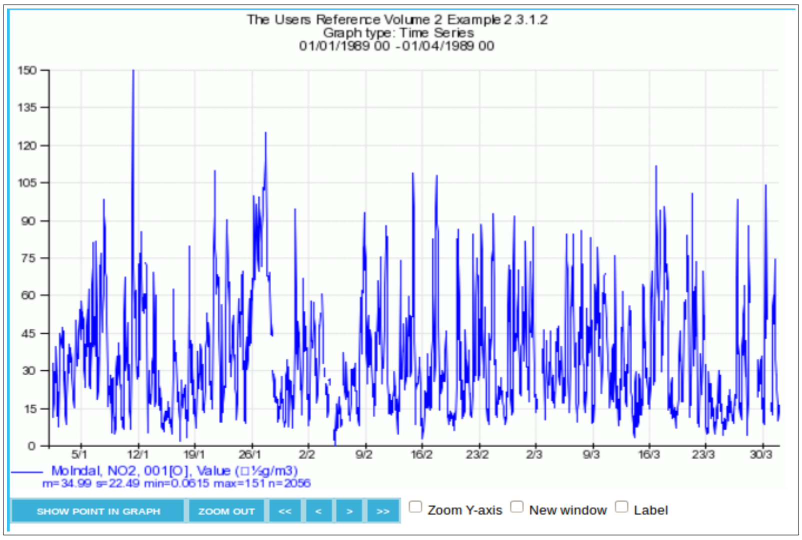


Figure 3.4.2.1 Interactive Output graph in GIF format.

When you write the graph in PDF-format, you need Adobe® Reader® or some other Adobe program to look at the graph. If you use Adobe Reader 6, you can take a snapshot of the graph and paste it into a Word document. You can zoom in the graph and print to any printer, using Adobe functions.

Text font and special characters can be changed if you use Adobe® Acrobat®. With Acrobat, you can also export the graph in gif format.

You cannot save or use the PDF-file without an Adobe program, except if you can locate the temporary PDF-file, which is saved under Temporary Internet Files in your profile directory.

The time series can be exported to other programs by sending the output to an ASCII file. Select **Output – Text** in the left-hand menu.

Also, the time series can be exported to excel format by selecting **Output** – **Excel** in the left-hand menu (only version 3.13 or higher).

The DS61 output can be used if the time series database has been configured to allow data to be deleted with a reason. It must exist a configuration file on the server that lists the reasons for data to be deleted. If configured, the DS61 output opens two windows: The first one containing normal data and the second the data that was deleted and the reason why it was deleted.

3.4.3 Available graph types

1. The time series graph

This graph type is a multiple line chart with a date variable on the abscissa and up to twelve variables on the ordinate axis. Both axes are linear and continuous. An example of a time series graph is seen in *Figure 3.4.3.1.* A legend may appear below the graph with descriptive statistics about mean value, standard deviation, span and number of valid cases.

It is important that you use an appropriate time period and time resolution to avoid cramming. If more than one variable is presented, you can try a different offset and ordinate scale to get a readable graph.

Applying a moving average to one or more variables will filter away high-frequency components and leave a smoothed line showing short-time trend. If you have much data in your view, as in *Figure 3.4.2.1*, your understanding of the variation will benefit from applying a smoothing filter.

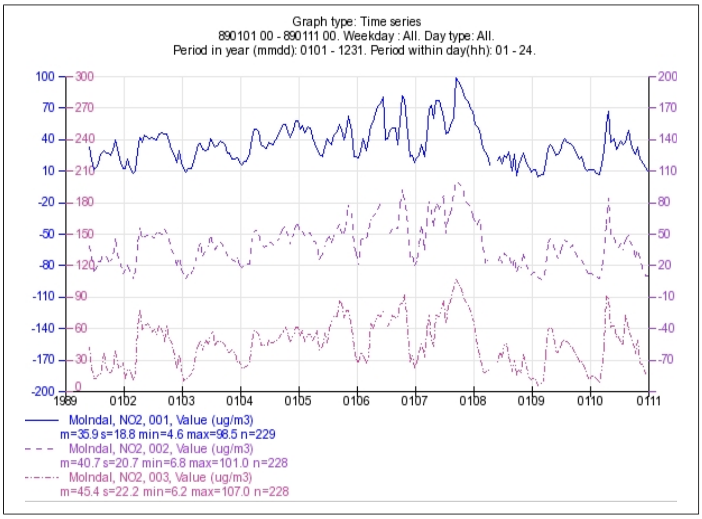


Figure 3.4.3.1 Time series graph with three channels at different offset.

2. Bar Chart

This graph type consists of vertical bars (rectangles) for each value.

You can show up to 4 plot variables in the Bar chart.

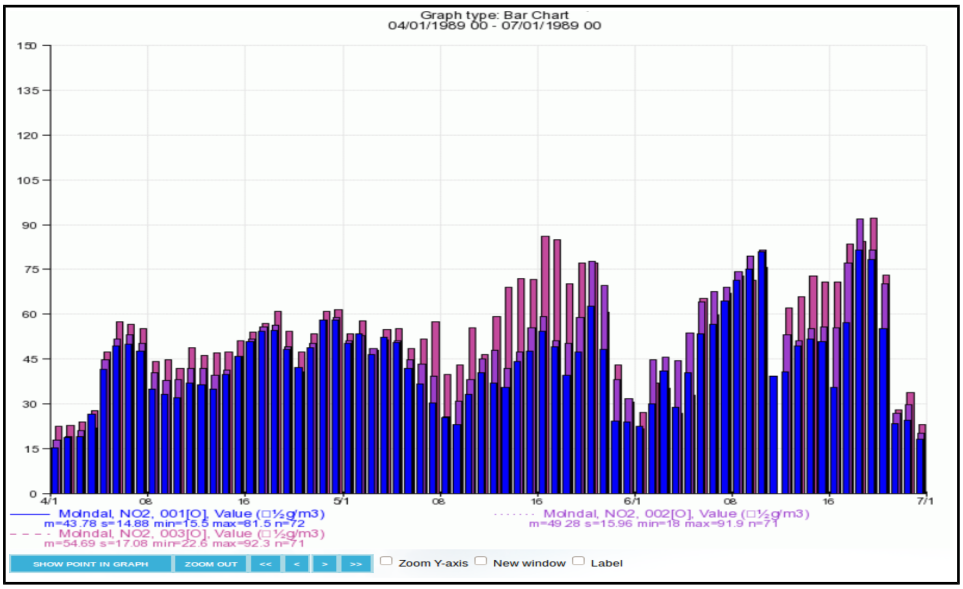


Figure 3.4.3.2 Bar Chart showing values for NO2 at different instances.

3. Filled time series

This graph type is also a time series graph with a time scale on the abscissa and one variable on the ordinate axis. Both axes are linear and continuous. An example of a time series graph is seen in *Figure 3.4.3.1.* A legend may appear below the graph with descriptive statistics such as mean value, standard deviation, span and number of valid cases.

The main difference with Time Series graph is that Filled Time Series displays Time Series values colored according their scale and user settings.

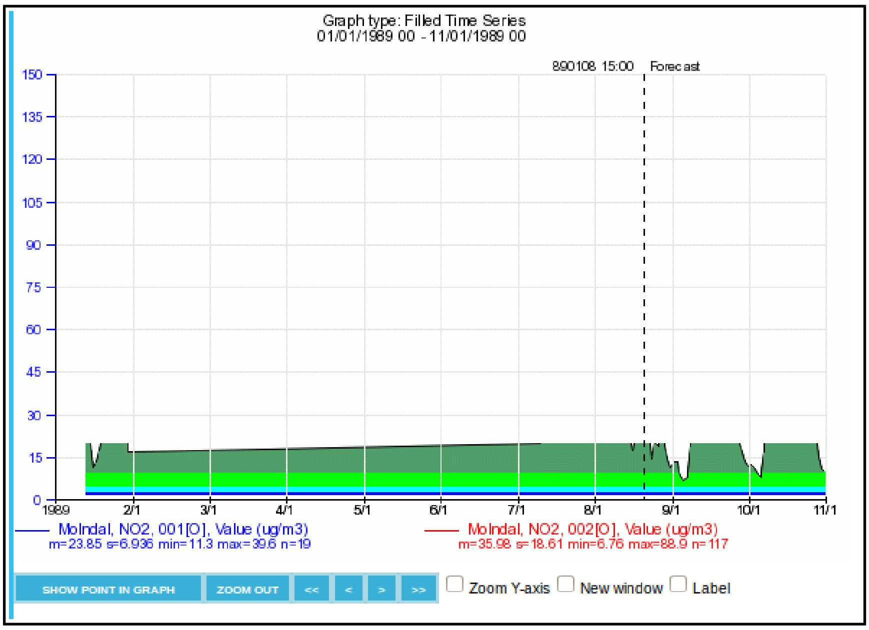


Figure 3.4.3.3 Filled Time Series showing colored values for NO2 at Molndal station.

4. The frequency distribution graph

This graph type is divided into two parts - a percentage histogram and a cumulative distribution chart. In the histogram, the ordinate is always linear from 0 to 100%. The examined variable is divided into 10 discrete classes according to the min-max settings in the Variables frame.

If you want to group your data into other classes, it is always possible to recode your data as in section 3. 3.1.

You can have up to four plot variables in the frequency distribution chart. The color of the histogram bar is the same as for the associated line. Multiple bars are clustered, but are careful - they can have different scales.

In the cumulative distribution chart, the ordinate is always square root-distributed for percentiles from 100 to 0%. The abscissa is the same as in the histogram – linear and continuous for the examined variable. For mass concentrations, it is sometimes interesting to transform into a logarithmic scale, since some substances are log normally distributed.

If you know the cumulative distribution functions, it is possible to calculate extreme values and exceedance statistics with recurrence times and more. However, be careful with the effects of sampling time, which tend to filter away peak values.

In the cumulative distribution chart, you can read the observed median, 90%-ile, 95%-ile, 98%-ile, 99 and 99.9%-ile etc. during some period. Many national standards have limit values related to these percentiles.

An example of a frequency distribution graph is seen in *Figure 3.4.3.2*.

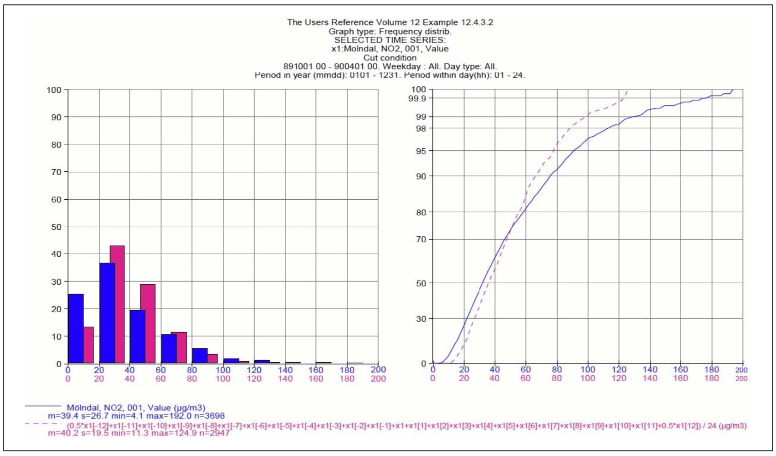


Figure 3.4.3.4 Frequency distribution graph with percentage histogram and cumulative distribution chart. The effect of applying a 24-h moving average can be seen by comparing plot variable 1 and 2.

5. The scatter graph

This graph type can plot a dependent variable against up to four other plot variables, one at a time in an XY scatter plot. The variables are simultaneous pairs.

The correlation coefficient R between the two series is calculated and a regression line is fitted in the scatter plot. The y-intercept and the slope of the regression line are presented if you have included Statistics in the layout. The square of the correlation coefficient is a measure of how much of the variation in Y that is explained by the plot variable. By comparing the standard error s_eps in the regression line with the standard deviation s for the dependent variable Y, you will get an opinion of how much that remains to be explained by other variables.

If you check the **Regression line** box in the **Scatter** sub-frame, the regression line will be included in the scatter plot.

You can transform the plot variables if you want to improve the correlation between the series. Make sure that the selected sample is homogeneous and control outliers to get a more representative correlation. You should probably not apply a moving average in the scatter plot, because it would blur the correlation between the variable pairs.

See *Figure 3.4.3.3* for an example of a scatter plot. The color and size of the markers can be defined in Graph settings.

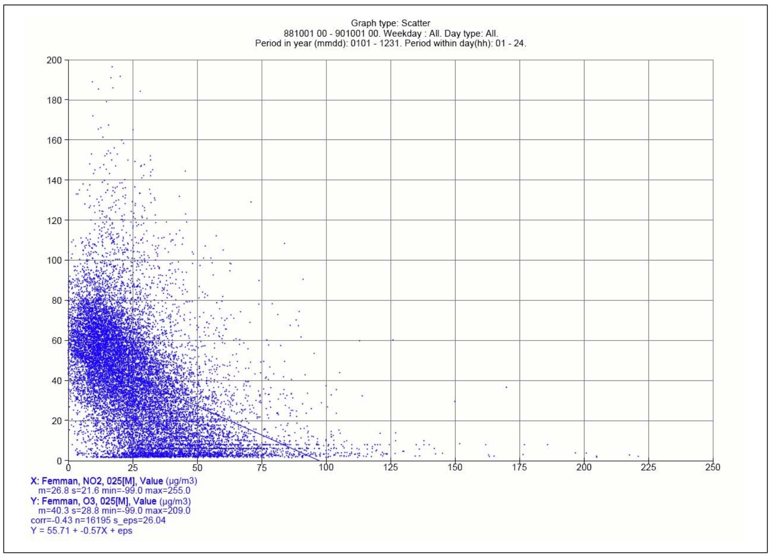


Figure 3.4.3.5 Scatter plot with ozone vs NO_2 concentrations. A regression line is included in the plot.

6. The Breuer diagram

This graph type is a polar diagram where up to four variables can be plotted against simultaneous wind direction. Each observation is plotted with clockwise angle in degrees from north and distance from the center in a polar coordinate system according to the scale in min-max in the Variables frame. Negative values in some wind direction are allowed, e.g. temperature.

The circle is split up into sectors of arbitrary size. A user-selected percentile (quantile) is displayed with an arc in each sector. If the sector size is indivisible with the full 360-degree circle, the size of the last sector will increase.

The Breuer diagram can be used as a pollution rose, which points out the direction to major sources. By combining measuring stations, you can get more bearings to the sources, making it possible to localize source areas. When interpreting a pollution rose, it is important to remember that a small source located near the measuring site can give high concentrations, which are not representative for a larger area.

If you multiply the concentration by wind speed in a formula to present the pollutant flux, you may get a clearer opinion of the direction to various sources.

It is possible to use regression techniques to construct pollution roses by combining 24-hourly samples with hourly wind measurements with good results, if the sources are continuous. See *Cosemans, G. and Kretschmar, J, 2003: Pollution roses for 24h averaged pollutant concentrations by regression. Proc. 8th Int. Conf. On Harmonisation within Atmospheric Dispersion Modeling for Regulatory Purposes*, which also hints on methods to select optimum sector size.

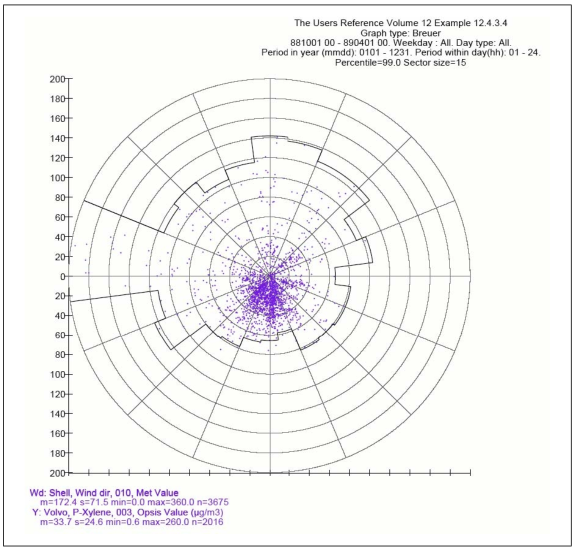


Figure 3.4.3.6 Breuer diagram with P-Xylene concentrations contra wind direction in a polar coordinate system. The 99%-ile concentration in each sector is indicated by an arc.

7. The mean/sector diagram

This graph type is another polar diagram, which presents the arithmetic mean value of the plot variable as a radius vector.

You can get an even sharper direction to the source by using this diagram, particularly if you combine it with a non parametric regression estimator. In principle, you have to apply a sliding window with a known shape over contiguous wind directions to form an average concentration. One suggestion is to calculate the mean as:

$$\bar{C}(\theta, \Delta\theta) = \frac{\sum_{i=1}^{n} C_i \, K((\theta - W_i)/\Delta\theta)}{\sum_{i=1}^{n} K((\theta - W_i)/\Delta\theta)}$$

where θ is the examined wind direction, W_i is the actual wind direction, $\Delta\theta$ is the width and K is the shape of the sliding window, which could be a Gaussian kernel like:

$$K(x) = (2¶)^{-1/2} \exp(-0.5 x^2)$$

Other shapes could be the Epanechnikov kernel $\{K(x) = 0.75(1-x^2), -1<x<1\}$, or a simple function returning the value 1 inside the window and 0 outside.

The technique is known as a Nadaraya-Watson estimator. See *R.C.Henry et al.* In *Atmospheric Environment 36 (2002) 2237-2244.* Optimal window width can be calculated by cross validation regression.

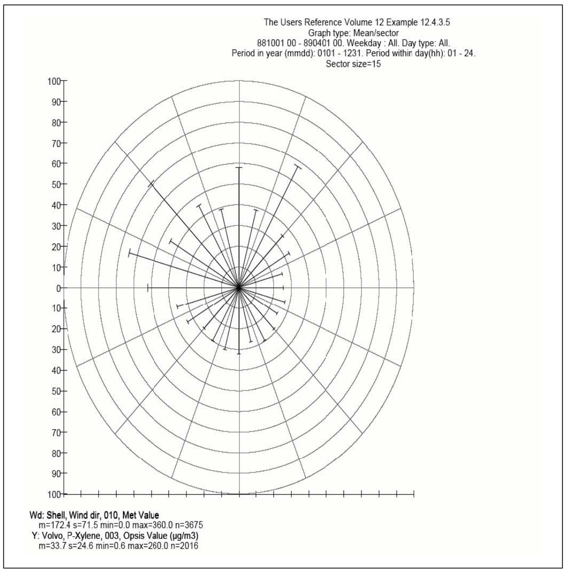


Figure 3.4.3.7 Mean/sector diagram with P-Xylene concentration vs. wind direction.

In the same way as in the Breuer diagram, you can calculate the pollution flux instead of the concentration to get a more distinct presentation.

8. The frequency/sector diagram

This type graph is simply a wind rose showing the relative frequency of winds in the different sectors. You can select sector width and scale on the radial axis and change colour of the radius vector.

You need at least one plot variable, but the frequency/sector diagram always uses the specified wind direction variable (in degrees from north) to calculate and present a wind rose. The wind direction is the compass direction that the wind is coming from.

If the sector width is indivisible with the full 360-degree circle, the radius vector will be plotted in the direction of the midpoint of the integer divided sector. Winds in the exceeding fraction will be omitted, so the total sum can be lower than 100%.

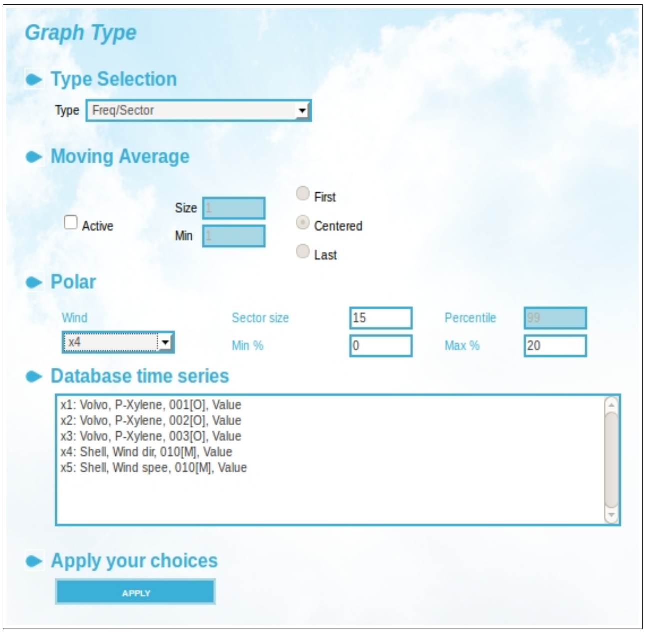


Figure 3.4.3.8 Graph Type frame with settings for a wind rose in the Polar sub frame.

9. The diurnal variation diagram

This graph type is a simple line chart showing the arithmetic mean for the examined variable(s) grouped by hour.

Some sources have a typical diurnal pattern. If you divide the observations into different sectors, you may be able to recognise the diurnal pattern from different source types in the observations. Please remember that the wind often has a diurnal pattern, so it can be a good idea to put the concentrations on flux form by multiplying with wind speed.

If the observed trend is small compared with the seasonal variation, the diurnal variation diagram is the easiest way to present this component.

Standard deviations, number of cases in each group or other summary statistics are not easily calculated, but for a sample of known size, it is possible to write formulas.

The moving average should of course not be used in the diurnal variation diagram, since it would destroy the purpose of the graph.

The diagram is slightly difficult to read, since the hours are numbered from 1 to 24, while the abscissa is scaled from 0 to 24. This is because hour 01 represents 0:00 to 1:00.

If you write the output to Text, you will find more concise information, also including standard deviation, min/max values and number of time steps.

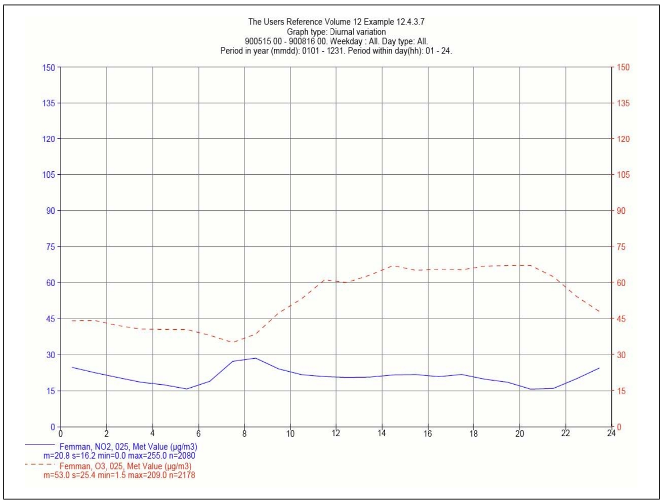


Figure 3.4.3.9 Diurnal variation diagram with anticorrelated NO_2 and ozone concentrations.

10. The weekly variation diagram

This graph type is a simple line chart showing the arithmetic mean for the examined variable(s) grouped by weekday.

Days are numbered from 1 to 7, representing Monday through Sunday. In the graph, you should probably use markers instead of lines to present the mean value. Each marker should be shifted one half time step to the right.

If you write the output to Text, you will again find information about standard deviation, min/max values and number of observations in each group.

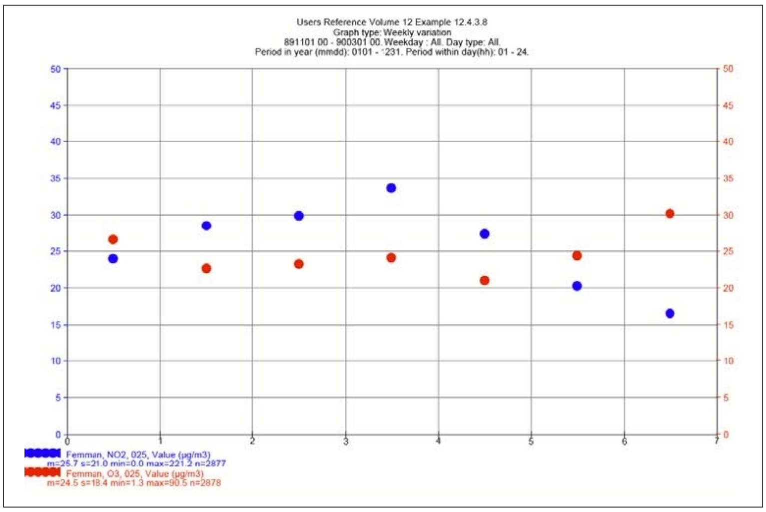


Figure 3.4.3.10 Weekly variation diagram with anticorrelated NO_2 and ozone concentrations. During weekends, the NO_2 concentration is lower than during weekdays.

11. The annual variation diagram

This graph type is very much like the diurnal and weekly variation diagrams, except that data are grouped per calendar month, showing seasonal variations.

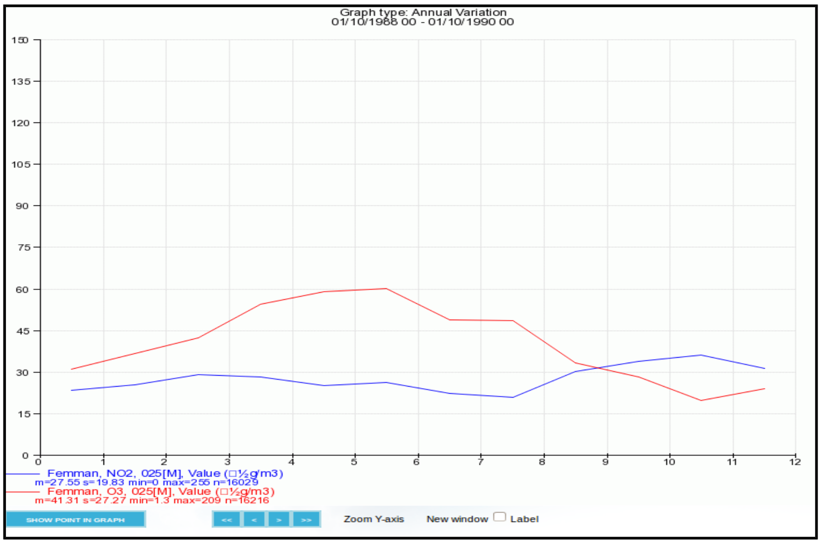


Figure 3.4.3.11 Annual variation diagram with NO_2 and ozone concentrations. Ozone has a very strong seasonal variation, since more ozone is formed in strong sunshine.

3.5 Regression modeling

In the **GRAPH TYPE** frame, there are five different analysis types. All systems don't have access to the statistical analysis types. The statistical analysis types are:

- Multiple Linear stepwise Regression based on Forward selection (MLRF F-criteria)
- Multiple linear stepwise regression with cross validation (MLRF crossvalid)
- Regression Estimation of Event Probability (REEP)
- Factor Analysis
- Principal Component Analysis

The first three methods will be described in this chapter, while the others are described in chapter *3.6 Factor analysis*.

All statistical analyses use statistical variables, as defined in the **VARIABLES** frame. Up to 64 statistical variables can be defined, regardless of the number of time series that have been selected.

The purpose of applying a statistical analysis can be to design a statistical model. The model can be used to find the input function, the transfer function or the output function. Often, we are interested in describing a dependent variable as a function of other variables.

The regression model can in general be described as:

$$Y = b_0 + b_1 X + b_2 X + \dots + b_n X_n + \varepsilon,$$

Where Y is called the predictand, representing NO_2 or any variable of interest. The X's, i.e. X_1, X_2, …, X_n, are called predictors, representing any variable like NO_x, ozone, wind speed or some transformation of primary variables like ventilation index, pollution index, stability index, zonality index or some mathematical transformation like ln(x).

The coefficients b_0, b_1,…,b_n are called regression coefficients. Linear regression estimates the coefficients of the linear equation to best predict the value of the predictand during the observed period. If the observed period is longer than one time step, it is seldom possible to find coefficients that perfectly describe the predictand. The model is usually approximate, leaving a residual error, which is denoted as e in the equation above. When the coefficients have been correctly estimated, the residual error is minimized.

For each model we can define a number of statistics to evaluate the model performance:

- Correlation between Y and its model estimate (perfect fit = 1, useless = 0)
- Standard error of the estimate (standard deviation of e)
- Explained variation of Y (square of the correlation, 0-100%)

In order to estimate the values of b_0 – b_n you need a dataset including not less than 10 times the number of predictors (preferably 100-1000 times).

3.5.1 Linear regression model

The multiple linear stepwise regression based on forward selection is explained in some detail in *E1.1 The Stepwise Regression Scheme* in *Airviro Specification, part II*. Its use is best explained by a thorough example of how to build a statistical model.

When simulating dispersion, the transformations of substances due to chemical reactions are often difficult to compute. First of all because a proper mathematical scheme describing the coupled system of chemical non-linear equations are extremely compute-intensive, implying the need of a supercomputer. Secondly, it is often doubtful if initial conditions can be correctly described, e.g. the initial distribution of chemical substances needed for the calculations.

For climatological simulations, i.e. with all types of weather and emission scenarios, it is not necessary to include a non-stationary chemical model if we only want to identify mean ambient air concentrations or perhaps extreme cases. We only need a statistical model able to properly describe the mean and the distribution function in the chemical transformation process.

In the example, we shall demonstrate the principles of how to set up a statistic model describing the relation of NO_2 to NO_x, using stepwise regression.

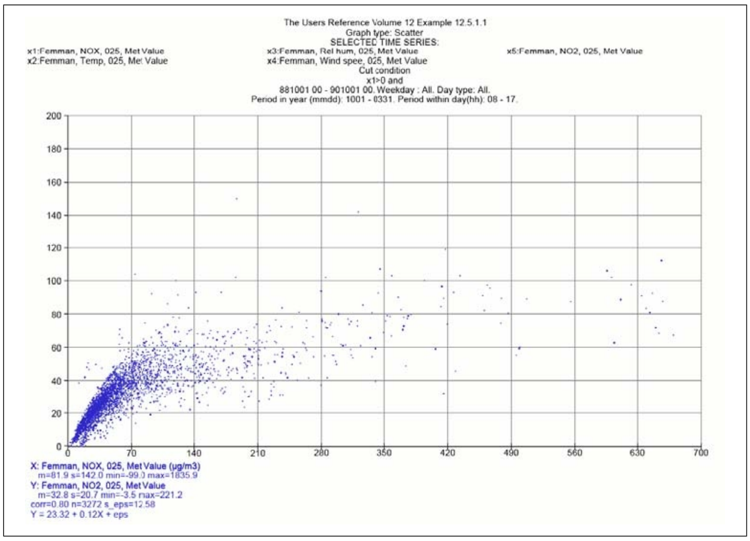


Figure 3.5.1.1 Scatter plot of NO_2 vs. NO_x concentrations.

In *Figure 3.5.1.1* you can see a scatter plot of NO_2 as a function of NO_x. For small values of NOx (<100 µg/m^3) the ratio NO_2/NO_x is 50%-70%, but for large NO_x concentrations (>500 µg/m^3) the ratio is approximately 15%.

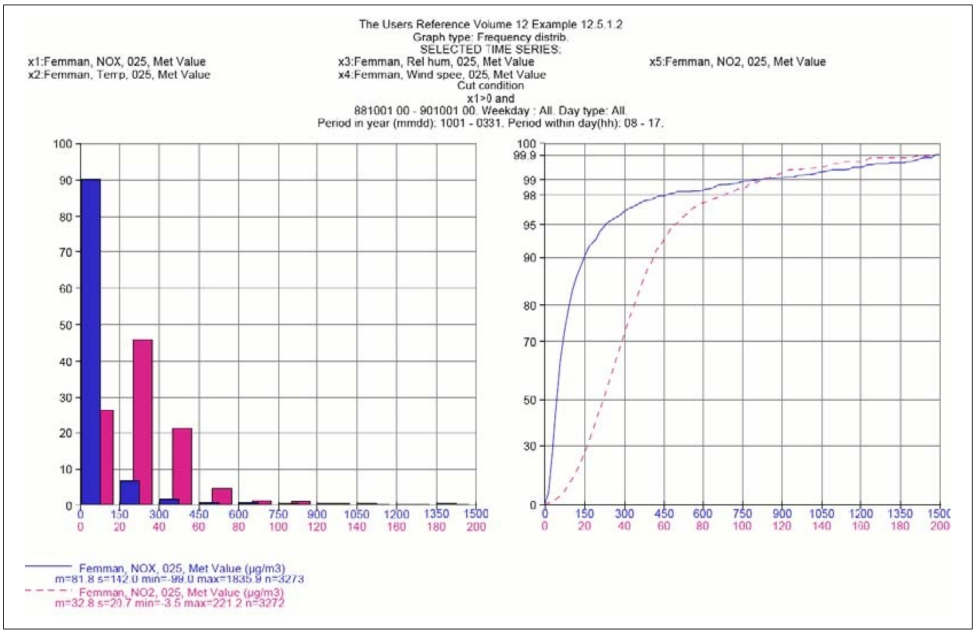


Figure 3.5.1.2 Frequency distribution graph with NO2 and NOx concentrations.

From the frequency distribution graph in *Figure 3.5.1.2* you can see that the cumulative distribution of NO_x differs from the distribution of NO_2. Consequently, any linear model relating NO_2 to NO_x would not be able to describe the basic features of the NO_2 variations.

We can see in the scatter plot that the relation seems to be logarithmic or inversely proportional so we arrange the predictors as:

$$\ln(1+NO_x),\ 1/(1+NO_x),\ NO_x^{0.8},\ NO_x$$

and add three additional linear predictors from temperature, relative humidity and wind speed.

Settings for statistical variables are done in the **VARIABLES** frame, see *Figure 3.5.1.3* below.

The MLRF F-criterion scheme is selected in the **GRAPH TYPE** frame. You can set criteria for including and excluding predictors in the fields F-in and F-out. **F-in** is the probability that you include a variable that is not correlated with the predictand. If you have no prior knowledge of the selected predictors, you can use a low probability like 0.01 (1%). If you on the other hand believe that the predictors should be included, you can use a higher value like 0.05, meaning that the predictors will be included at each step with rejection at the 5% probability level in the one-sided Fisher distribution.

F-out is the criterion that a variable already included in an earlier step should be rejected in a later step. You should require a high probability that the variable is insignificant before rejecting it, since it has already been selected in the scheme as having better fit than subsequent predictors. A recommended value for F-out is 0.10 but even higher values like 0.25 can be used, which means that you can retain highly insignificant variables. It is possible to enter predictors into the scheme without stepwise selection.

Figure 3.5.1.3 Defining statistical variables in the Variables frame.

With the settings in *Figure 3.5.1.3*, there are seven predictors. In the Graph type frame, you also define which variable that is dependent. The **dependent variable** should not be included among the predictors, but it should be written as a formula among the variables. See *Figure 3.5.1.4.*

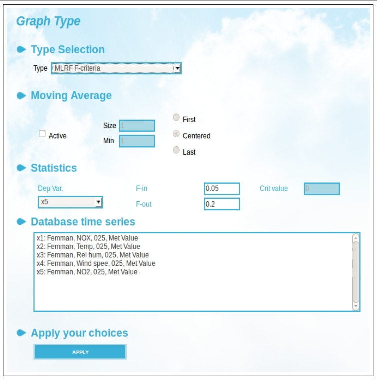
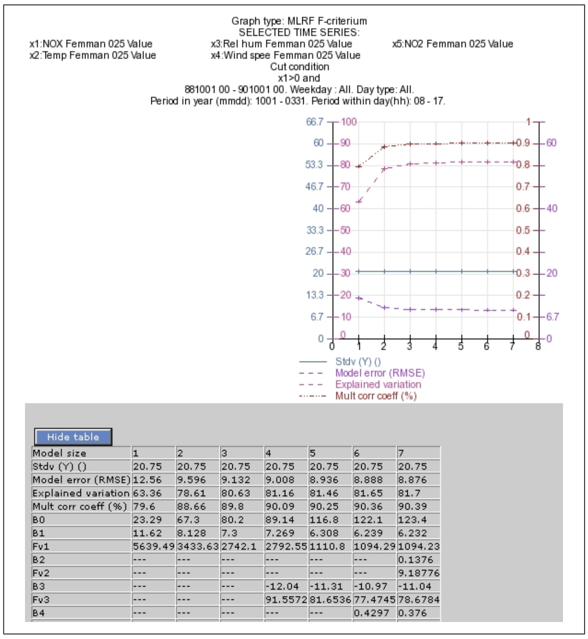


Figure 3.5.1.4 Defining dependent variable and F-criteria in the Graph type frame.

With these settings you can run the stepwise regression model **MLRF F-criteria**.

The result is presented as in *Figure 3.5.1.5* below.



Model size	1	2	3	4	5	6	7
Stdv (Y) ()	20.75	20.75	20.75	20.75	20.75	20.75	20.75
Model error (RMSE)	12.56	9.596	9.132	9.008	8.936	8.888	8.876
Explained variation	63.36	78.61	80.63	81.16	81.46	81.65	81.7
Mult corr coeff (%)	79.6	88.66	89.8	90.09	90.25	90.36	90.39
B0	23.29	67.3	80.2	89.14	116.8	122.1	123.4
B1	11.62	8.128	7.3	7.269	6.308	6.239	6.232
Fv1	5639.49	3433.63	2742.1	2792.55	1110.8	1094.29	1094.23
B2	---	---	---	---	---	---	0.1376
Fv2	---	---	---	---	---	---	9.18776
B3	---	---	---	-12.04	-11.31	-10.97	-11.04
Fv3	---	---	---	91.5572	81.6536	77.4745	78.6784
B4	---	---	---	---	---	0.4297	0.376

Figure 3.5.1.5 Regression model performance for stepwise increasing model size.

From the result you can see that predictor 2 is included in the first step, predictor 1 in the second step, predictor 4 and 3 in the next steps etc. The regression coefficients are presented if you output the result to a graph, together with critical F-values for significance. You can also see the multiple correlation coefficient R, total explained variation R^2 and standard error of the estimate, which can be compared with the standard deviation of the predictand Y.

It is possible to decide model size based on this information, but it is advisable to carry out a cross validation with **MLRF crossvalid** before the decision.

The multiple linear stepwise regression with cross validation is explained in some detail in *E1.2 Validation of the Regression Model* in *Airviro Specification, part II*.

Simply explained, the original data will be divided into a number of subsets to be used systematically both as basic data and as test data. The purpose of this procedure is to warn against problems like over fitting of data.

Choose **MLRF Crossvalid** in the Graph type frame and **APPLY**. The result, as a graph will be similar to *Figure 3.5.1.6.* It shows the standard error and the explained variation R^2 for each model size. Of course, in a regression model we would like to have the model error as low as possible and the explained variation as high as possible. In this case you will notice that the standard error drastically increases for model size 7, and the explained variation decreases. This is cau

sed by over fitting of the data and it would be unwise to use model size 7 in this case.

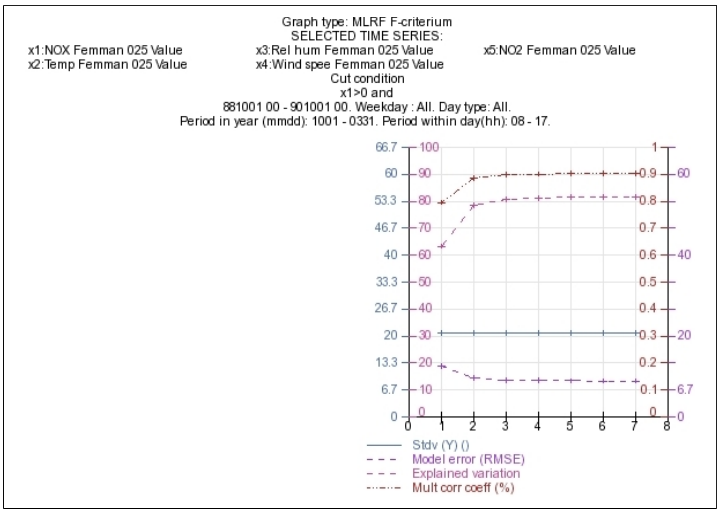


Figure 3.5.1.6 Regression model performance for cross validated models.

Looking at *Figure 3.5.1.5* and *3.5.1.6* we decide to use the parameters based on model size 4 (the improvement in performance from model size 4 to 5 is not pronounced). The coefficients for the predictors can be obtained from the table in the Output graph. Click on **SHOW TABLE** and read the coefficients and correct order of the predictors, comparing with the order of statistical variables in the Variables frame (listed to the right of active check boxes).

If you want to examine the correlation between the dependent variable and the model estimate, you can enter the linear regression model as a formula into a plot variable and produce a scatter plot with a regression line. See an example in *Figure 3.5.1.7*. The Figure shows that the observed and predicted NO_2 concentrations seem to be non-biased and distributed along a straight line with a standard error of 8.7 µg/m^3.

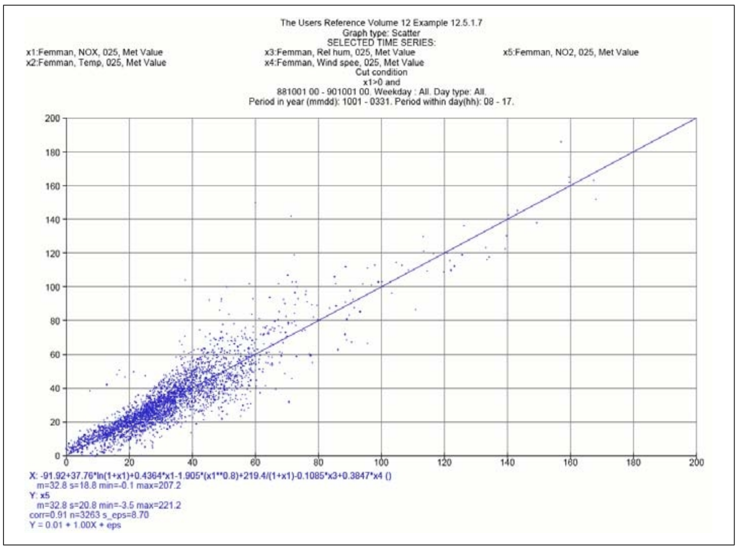


Figure 3.5.1.7 Scatter plot of measured contra estimated NO_2 concentrations.

The next step is to check if the model is capable of describing the cumulative distribution of NO_2 in a proper way. Present the two variables in a frequency distribution graph. The result in *Figure 3.5.1.8* shows that the NO_2 distribution from the model is very close to the curve from observed NO_2.

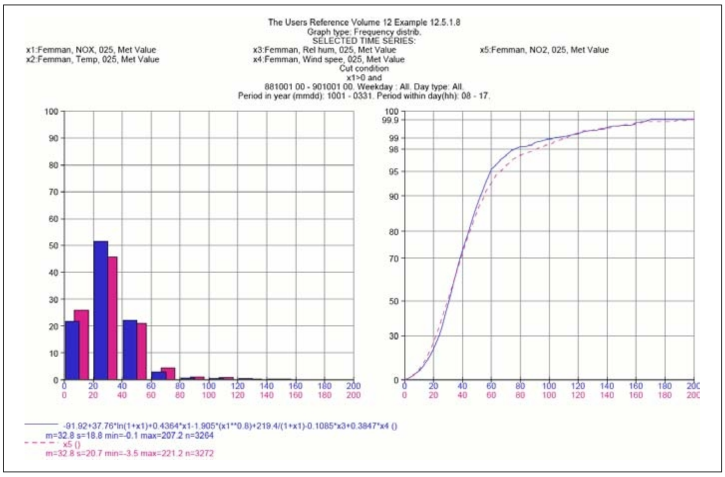


Figure 3.5.1.8 Frequency distribution graph with measured and estimated NO_2 concentrations.

We have thereby shown that a statistical model can be designed to describe the relation of NO_2 to NO_x as a non-biased pure or causal model, having a distribution function similar to the distribution of NO_2. This model can be used as a complement to the dispersion model of the REF application for estimating seasonal mean as well as extreme values of NO_2. The statistical model should not be applied on other domains unless validation studies have been made.

3.5.1.1 Fitting a curve

It is not always easy to find good predictors, but the available time series can be examined systematically in a search for them. If you want to fit a dependent variable to only one independent variable, you can explore some terms of a polynomial. In a scatter plot, you can plot the dependent variable against some transform of the independent variable. It is also possible to transform the dependent variable to find power law functions etc.

The following transforms are suggested:

Linear	$Y = b_0 + b_1 x$	no transform of x
Logarithmic	$Y = b_0 + b_1 \ln(x)$	logarithmic transform of x

Inverse $Y = b_0 + b_1 \ 1/x$ inverse transform of x

Quadratic $Y = b_0 + b_1 \ x^2$ quadratic transform of x

Cubic $Y = b_0 + x \ b_1 \ x^3$ cubic transform of x.

Growth $\ln(Y) = b_0 + b_1 x$ logarithmic transform of Y, no transform of x

Power $\ln(Y) = \ln(b_0) + b_1 \ln(x)$ logarithmic transform of Y and x

S $\ln(Y) = b_0 + b_1 \ 1/x$ logarithmic transform of Y, inverse transform of x

Exponential $\ln(Y) = \ln(b_0) + b_1 x$ logarithmic transform of Y, no transform of x.

With the growth transform, you can find predictors like $\exp(b_0+b_1x)$; with the power transform, you can find predictors like $b_0 * x^{b1}$. With the S transform, you can find predictors like $\exp(b_0+b_1/x)$ and with the exponential transform you can find predictors like $b_0 * \exp(b_1x)$.

In the scatter plot, the intercept, b_0 or $\ln(b_0)$, and the slope, b_1, are expressed in the statistical information below the regression line. You can test different transforms in a scatter plot before including them in a regression analysis.

If you further want to describe a process that is time-dependent, you may use the lag function to form autoregressive predictors. Together with moving averages and differencing, you can form a pure stochastic model with good performance.

If you have good physical reason for including other predictors from your measurements, they can be combined in various ways with each other to form indexes or transforms.

3.5.2 Binary logistic regression model

The REEP model (Regression Estimation of Event Probability) is based on the stepwise regression method with forward selection, but the predictand is transformed to a binary variable, i.e. a variable that has the value 0 or 1 (False or True). The transformation is based on the **Crit value** in the Graph type frame. If the value of the predictand is less than the criterion, it will be transformed to a False value, otherwise it is True.

The REEP procedure can in principle be applied on categorical or binary predictors. A categorical predictor is some variable divided into categories, e.g. Beaufort wind speed classes, wind sector or some other category. A binary predictor could be the presence of snow, daylight, temperature inversion, decoupling, thunderstorms or some other phenomenon like sports events that attract much traffic etc. The binary predictor has the value 0 or 1 (False or True). Categorical and binary predictors can be created by recoding some available time series with the conditional operator or some other function.

An important case is if you want to verify a statistical model against measured data for some threshold value, e.g. the National Standard. To do this, you should select your dependent variable in the Graph type frame and write the National Standard in the **Crit value** field. Next you have to transform your statistical model, which you have probably written as a plot variable. The formula is similar to **reep** ($b_0+b_1x_1+b_2x_2$, *crit value*).

If you run the **REEP** model with the recoded model, you will get a contingency table that shows you how well the estimate agrees with the reference about exceeding the national standard:

MS: 1	PROG=0	PROG=1
OBS=0	61	6
OBS=1	7	26

In the above case, the model agrees with predictand at 61+26 cases of totally 100, which means that chances are 87% that the model will give a correct answer if the national standard is exceeded or not.

If you have many candidates for the statistical model, you can test them one at a time with the REEP model to get best agreement in a particular concentration interval.

You can add more binary predictors to see if the performance could be improved. In that case you will get a contingency table for each model size, showing how the odds change. You can find the regression coefficients in the **Output graph** by clicking **Show table**.

Categorical predictors can be used in the REEP analysis, but each category can also be transformed into a binary predictor. If you decide to do this, bear in mind that no binary predictor can be an exact linear combination of other predictors; one category must be left out, for mathematical reasons. It doesn’t matter which category that is excluded.

For categorical predictors, which can be ordinal or nominal, you may have to normalise the categories into the interval]0,1[.

For more information, see “Miller, R.G. (1964): Regression estimation of event probabilities. Technical Report No 1. The Travellers Weather Research Center, Inc., Hartford, Conn.” or “Glahn, H.R., Murphy, A.H., Wilson, L.J, and Jensenius, J.S. (1991): Lectures presented at the WMO training workshop in the interpretation of NWP products in terms of local weather phenomena and their verification. PSMP Report Series No 34, WMO.”

3.6 Factor analysis

The investigation of basic relationships between air quality and other aerometric variables by statistical means is complicated by the highly intercorrelated nature of variations in the data. The fact that many variables tend to rise and fall more or less in tandem presents problems for statistical analysis and interpretation. Factor analysis and the associated principal component analysis can overcome the technical difficulties and at the same time provide valuable insight into the underlying chemical and physical properties of the atmosphere. Principal component analysis is a special case of factor analysis, but both refer to a method of multivariate linear statistical analysis.

It is potentially dangerous to run a multiple regression analysis on intercorrelated variables. Meteorological and air quality data are often highly intercorrelated. Ordinary multiple regression has been shown to significantly overestimate the importance of two pollution related variables.

The basic idea of factor analysis is to transform a set of intercorrelated variables into a set of independent, uncorrelated variables, by means of orthogonal transformations (rotations).

The first step is to standardize the original time series $x_1(t)..x_K(t)$. The standardization means that for each series we determine the ensemble mean value and the standard deviation:

$$m_i = 1 / M \sum_{t=1}^{M} x_i \quad \text{and} \quad \sigma_i = 1 / (M-1) \sum_{t=1}^{M} (x_i - m_i)^2$$

where M is the number of time steps in the series. The standardized variables $z_i(t)$ are given by:

$$z_i(t) = \frac{x_i - m_i}{\sigma_i}$$

Whereby all standardized predictors have the same mean value (0) and the same standard deviation (1). The values used in the factor analysis are also made dimensionless by this transformation. The factor analysis model is:

$$z_i(t) = \sum_{n=1}^{N} f_n(t) \cdot h(i) \quad \text{for } i=1,2,\ldots,K$$

Let $f_n(t)$ denote an orthonormal factor and $h_n(i)$ the eigenvector corresponding to the factor f_n. The original standardized variables have now been transformed to a number of new variables, f_n, and the contribution of each factor to the original series is described by the eigenvector h_n.

In the decomposition of original data into factors, a constraint of fastest possible convergence is applied. This implies that the first factor chosen is the one that alone explains as much variation in the original variables as possible. The number of factors can be as many as the number of original variables (N = K). If the original variables are highly intercorrelated, we will probably end up with a number of factors that are less than the original number (N < K), which is what we want to achieve.

In order to run the **Factor analysis**, select your independent statistical variables in the variables frame, go to the Graph type frame and select **Factor analysis**. Click **APPLY** and send the result to **Output Graph**. See *Figure 3.6.1.* You will find a component matrix plot with eigenvectors for each variable. You can also see a component matrix with similar information in tabular form if you click **Show table**.

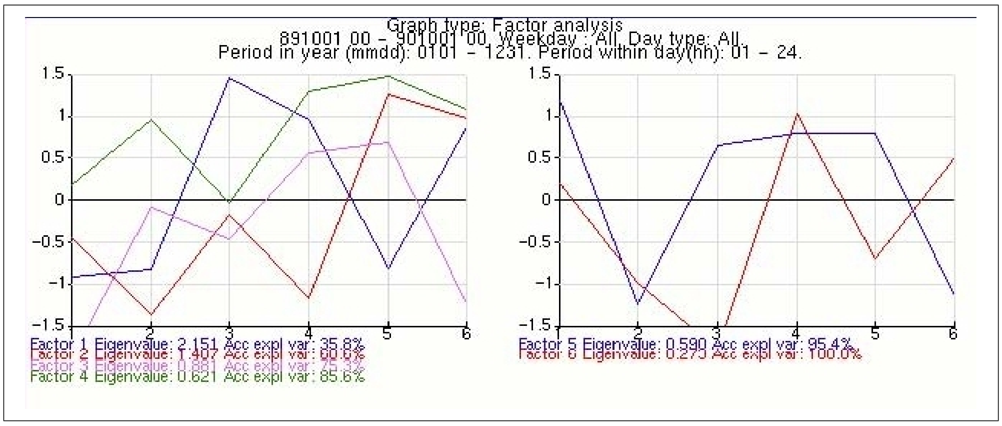


Figure 3.6.1. Factor analysis .

The first factor is a linear combination of your standardized statistical variables with weights according to the plotted (or tabulated) eigenvectors. In the table you can read the explained variance for each statistical variable.

The second factor is another linear combination of your standardized statistical variables. You can read the accumulated explained variance both in the table and in the statistical information below the graph.

The eigen value is a diagnostic measure of the co linearity of the factor with your original data. The eigen values are decreasing in size for each factor. The factor analysis will structure your data in this way, finding the typical variation among your time series and attempt to compress the common variation into a number of factors and describe to what extent the factors can explain the variation in each original time series.

You must decide how many factors you should retain, e.g. factors with an eigen value above some threshold; maybe 1. All factors that you retain can be introduced into a multiple linear regression analysis to get a good model fit for a dependent variable with as few independent factors as possible, but remember that the eigenvectors are based on standardized variables. Mean value and standard deviation can be found in the descriptive statistics below a time series graph.

3.6.1 Principal component analysis

The principal component analysis is similar to the implementation of factor analysis above, but the variables are not standardized, only adjusted to give each predictor the mean value 0:

$$z_i(t) = x_i - m_i$$

The principal component analysis model is:

$$z_i(t) = \sum_{n=1}^{N} a_n(t) \cdot g_n(i) \quad \text{for } i=1,2,\ldots,K$$

Let $a_n(t)$ denote an orthonormal amplitude function and $g_n(i)$ the eigenvector corresponding to the amplitude function a_n. The original adjusted variables have now been transformed to a number of new variables, a_n, and the contribution of each amplitude function to the original series is described by the eigenvector g_n.

In order to run the Principal component analysis, select your statistical variables in the Variables frame, go to the **Graph type** frame and select **Principal component analysis**. Click **APPLY** and send the result to **Output Graph**. You will find a component matrix plot with eigenvectors for each variable. You can also see a component matrix with similar information in tabular form if you click **Show table**. See *Figure 3.6.2.*

Hide table

	Factor	1	2	3	4
	Eigenvalue	2.612	0.150	0.019	-0.000
Var	Acc.expl.var	93.9%	99.3%	100.0%	100.0%
1	Eigen.vec	1.40	1.02	-0.06	-1.00
	Expl.var	97%	100%	100%	100%
2	Eigen.vec	0.49	-1.66	-0.06	-1.00
	Expl.var	60%	100%	100%	100%
3	Eigen.vec	-0.87	0.29	1.47	-1.00
	Expl.var	97%	98%	100%	100%
4	Eigen.vec	-1.02	0.35	-1.35	-1.00
	Expl.var	98%	99%	100%	100%

Figure 3.6.2. Principle component analysis table with eigenvectors for each variable in the amplitude function (factor).

The first amplitude function is a linear combination of your adjusted statistical variables with weights according to the plotted (or tabulated) eigenvectors. In the table you can read the explained variance for each statistical variable.

The second amplitude function is another linear combination of your adjusted statistical variables. You can read the accumulated explained variance both in the table and in the statistical information below the graph.

The eigen value is a diagnostic measure of the co linearity of the amplitude functions with your original data. The eigen values are decreasing in size for each function. The principal component analysis will structure your data in this way, finding the typical variation among your time series and attempt to compress the common variation into a number of amplitude functions and describe to what extent they can explain the variation in each original time series.

You must decide how many amplitude functions you should retain, e.g. functions with an eigen value above some threshold; maybe 1. All functions that you retain can be introduced into a multiple linear regression analysis to get a good model fit for a dependent variable with as few independent amplitude functions as possible, but remember that the eigenvectors are based on adjusted variables. The mean value for each series can be found in the descriptive statistics below a time series graph.

3.7 Using Indico macros

In the previous chapters we have given a number of recommendations for working with time series data presentation and statistical analysis. There are many ways to set selection criteria, make transformations, create statistical models or even self-adapting filters. Perhaps you have also made a layout of the graph. After all this work you realize that it would be convenient to save all these settings, to be able to use them again in the future.

In Indico the settings for a graph can be saved in a macro. Macros are stored in folders, each user have their own folder, a common folder and some other folders may also be created. The system administrator decides who is allowed to store macros in the common folder (setting the corresponding Indico.WriteGroup.user privilege in priv.rf). Users can always save macros in their own folders, but usually not in other folders, although it is possible to load macros from other folder.

It is very simple to save your settings:

- · On the menu, choose **MACROS**
- · Select a folder. The **Common folder** is always the root folder and can not be deleted. The user can add o delete new folders
- Specify a name for the macro in the text box under the **Macro** list to the right.
- Press the save macro button..

To load a macro:

- Select the folder where the macro is saved from the left list.
- Select the macro from the right list.
- Decide if you want to change period:
  - — **Time from macro:** the same period that is used during the saved in this macro.
  - — **Keep current period:** is the present period.
  - — **Latest 24 hrs:** are the last 24hrs from the present period.
  - — **Today:** is the period between 00 and 23hrs (00:00 ≤ x < 00:00).
  - — **Yesterday:** are the 24 hrs of the yesterday day.

    Latest 7 days: are the last 7days from the present period.
  - — **This month:** is the present month.
  - — **Previous month:** is a previous month to present month.
  - — **This year:** is a present year.
- Press the load macro button.

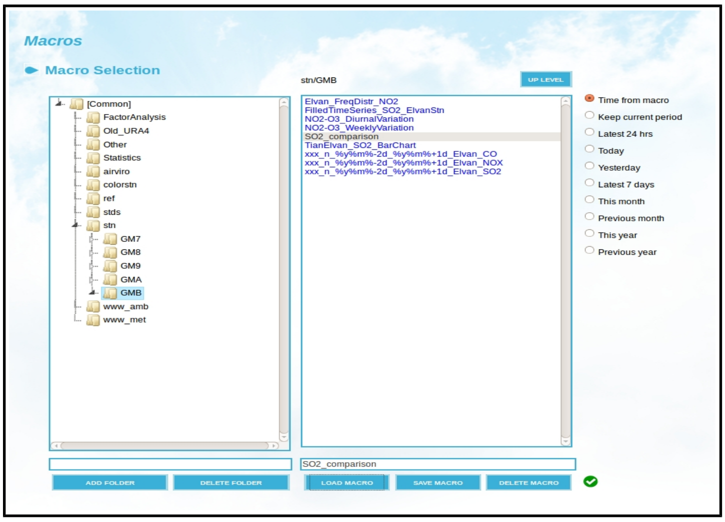


Figure 3.7.1. Macros.

An important macro is the one called **default**. This is loaded when Indico Presentation is started. Here you can define line styles and colors that you like best, status conditions you want to use, etc. If the default macro exists under your user name folder, it will be used, otherwise the default macro stored under **[Common]** will be used.

When a macro has been loaded your macro, the whole application will be in the same state as when it was saved.

Macros that start with the string “Auto” are used by the Real Time Graph in Indico Presentation module.

3.8 Indico Real Time

The Indico Time Graph displays a selection of predefined macros, one after the other, and is updated as new data arrive. Each macro is shown for a number of seconds before the next graph is drawn.

Once you are satisfied with your graph, save it as an Indico macro. The macro must start with the string “Auto” to be recognized by Indico Real Time. All macros starting with the string “Auto” are displayed in alphabetical order. You can also control the order in which the macros are displayed by using a number in the macro name, immediately following the "Auto" string.

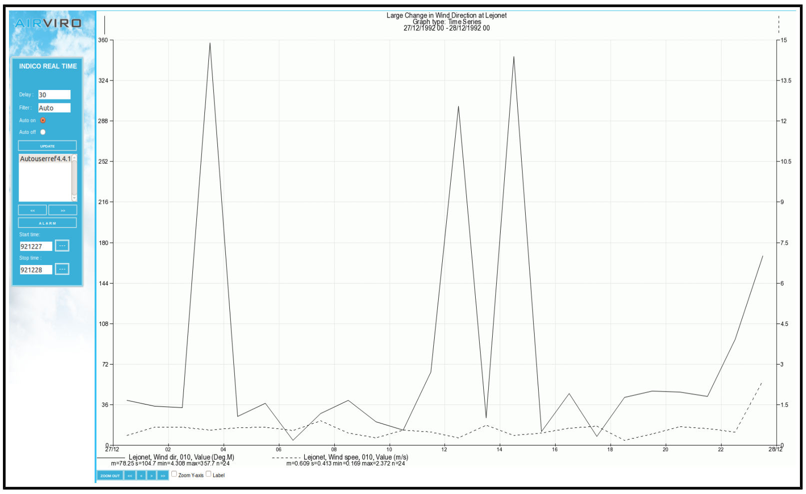


Figure 3.8.1. Real Time.

Different users can create different Auto macros. However, all Auto macros that have been created in the [Common] folder will be treated as belonging to all users. Indico Real Time will therefore display all Auto macros from the [Common] folder first, followed by all auto macros found in the folder of the user that started Indico Real Time.

In the Real Time graph, you can set the number of seconds that each macro will be displayed for in the **Delay** field. See *Figure 3.8.1.* If you want to select another set of macros then the macros starting with Auto, you can use **Filter** to specify what the macro names that will be used should start with.

When you have selected your macros and set the delay time, you can just leave the Real Time graph running on the screen to keep an eye on the measurements. In this way, you will be the first to notice if pollution levels start rising or if data collection has stopped working. If you are not pleased with the result, you can just alter the settings in the Auto macro in Indico Presentation and store it again. The size of the graph is changed by clicking on the border, holding it down and dragging the frame to desired size. You can also define the size of the window in the **GRAPH SETTINGS** frame.

You can use the forward “**>>**” and backward “**<<**” buttons to manually force the next or previous graph to be drawn. The option buttons **Auto off** and **Auto on** will let you stop Real Time from drawing the next graph, if you want to examine some detail closer.

The currently displayed macro is highlighted in the macro list. You can, at any time, click in the macro list to display the associated graph. You can also click the **UPDATE** button to force the macro to search the time series database again.

If you change time resolution in the domain, your macros will still be available, but remember that the settings may be dependent on time resolution.

The time period used can e changed in the Start time and Stop time text boxes.

Zooming can be made using the mouse to select a rectangle over the graph.

Appendix 3A Exploiting the Mathematical Functions for Calculation Parameters

Just looking at the measured data is not enough for you to draw the conclusions you would like to. In Indico Presentation you can process your data in a variety of ways, to build explanatory models and test your hypotheses.

The data accessed from the time series database can be transformed using mathematical analytical functions, which can be combined using algebraic as well as logical expressions, to end up with mathematical models.

The operators available are listed in the following sections: arithmetic and relational functions

-	Negation
+, -, *, /,	Standard operators
^, **	Power
? :	Conditional,e.g.(x1>0?x1:0)If x1>0 then use the value x1 else use the value 0

EQ (==) Equal to

NE (!=) Not equal to

GT (>) Greater than

GE (>=) Greater than or equal to

LT (<) Less than

LE (<=) Less than or equal to

3A.1 Logical functions

AND (&) And OR (|) Or NOT (!) Not

3A.2 Time Shift Functions

The expression x3[-1] refers to the time series selected as x3, shifted by -1 time unit. As an example, if x3 is plotted together with x3[-12] using the hourly database, then the values for x3[-12] will be the same as those for x3 but will be displayed with a time shift of 12 hours.

The general syntax is:

$$x_n[d]$$

where n is the number of the time series and d is the time shift required (d can be positive or negative).

3A.2 Special Variables

Th	Contains the number of the hour [1..24] for the current value
Td	Contains the number of the day [1..31] for the current value.
Tm	Contains the number of the month [1..12] for the current value.
Ty	Contains the year [YYYY] for the current value.
NTh	Contains the number of the current hour [1..24].
NTd	Contains the number of the current day [1..31].
NTm	Contains the number of the current month [1..12].
NTy	Contains the current year [YYYY].

3.A.3 Mathematical Functions

3.A.3.1 Combining Formulae

Algebraic functions: ln(x), log(x), exp(x), int(x), abs(x), sqrt(x)

Trigonometric functions: sin(x), cos(x), tan(x), cot(x)

Inverse trigonometric functions:	arcsin(x), arccos(x), arctan(x), arccot(x)
Hyperbolic functions:	sinh(x), cosh(x), tanh(x), coth(x)
Fill functions:	interpol(x,n) fills in missing values for x by
	interpolation of the nearest surrounding values.
	Requires that at least one value before and at least one
	value after the current time is within n time steps.
	sustain(x,n) fills in missing values for x by copying the
	nearest previous value. Requires that at least one value
	before the current time is within n time steps.
	interps(x,n) is the same as interpol(), but also makes a
	constant extrapolation if only one of the surrounding

values around the current time is within n time steps.

Miscellaneous functions: reep(x,a) (=0 if x<a else =1)

aver(x1,x2,...) Mean value of x1,x2...

aver(x1:x5) Mean value of x1 - x5)

min(x1,x2,...) Minimum value of x1, x2, ...

max(x1,x2,...) Maximum value of x1, x2, ...

It is of course possible to combine all of these functions to produce very complex functions such as:

min(x1+4, 0, ln(x2 - x1)) the minimum value of several functions

max(x1:x3[1], x1:x3, x1:x3[-1]) the maximum value of x1, x2, x3 looking at values for the current hour, the last hour and the next hour.

3.A.4 Missing Data Values

What happens when data is missing? Usually, if a variable is undefined for a particular point in time, then any function of that variable will also be undefined at that particular point in time.

However, this is not always the best solution. Consider the function **min(x1,x2,x3)**. If x1 is missing but x2 and x3 are not, **min** still returns the value undefined, whereas it would be preferable in some cases if it returned the value **min(x2,x3)** instead.

To get around this problem, three new functions have been created called **eaver**, **emin** and **emax** which work in exactly the same way as aver, min and max, except that these functions are only undefined if **all** of their parameters are undefined. So, if a function **emin(x1,x2,x3)** has been defined, and x1 and x2 are missing, **emin** just returns the value of x3.

Along with these a Boolean function has been created called exist, where for a time series x, exist(x) takes the value 1 if x exists and 0 otherwise.

3.A.5. Definition of the Airviro Air Pollution Index

The API (Air Pollution Index) is a mathematical function which transforms a level of a particular substance to an index value using the following function:

$API(x) = I_j + \frac{I_{j+1} - I_j}{} \times (x - c_j)$ for $c_j \leq x \leq c_{j+1} - c_j$

$c_j + 1$

where x is the measured concentration of a substance (rounded to an integer), and the c_j and I_j are the break points on the stepwise linear function which defines the relation between the concentration and index values.

In the following table the linear relation between the concentration values and index values is shown for five different substances, which has been prescribed by the United States EPA in the index known as PSI (Pollutant Standard Index).

Substance	PM	SO2	CO	O3	NO2	PSI value
Unit	µg/m3	µg/m3	µg/m3	µg/m3	µg/m3	%
Sampling Period (hours)	24	24	8	1	1	
	50	80	5	120	-	50
	150	365	10	235	-	100
	350	800	17	400	1130	200
	420	1600	34	800	2260	300
	500	2100	46	1000	3000	400
	600	2620	57.5	1200	3750	500

In Airviro the following mathematical function has been defined:

$api(x,c_1,i_1,c_2,i_2,...,c_n,i_n)$

where x is the database parameter, and the pairs c_j, i_j are the break points which specify the stepwise linear function which defines the API-function. An arbitrary number of break points can be defined, but the origin is not specified ((0,0) is by default used for the first break point). A minimum of one break point (c,i) must be defined.

The mathematical function:

$desc(Y,I_1,I_2,...,I_n)$

gives one of the values 1,2,3,...,n if $Y > I_1,I_2,...,$ or I_n. An arbitrary number of intervals can be defined but there must be at least one.

The USEPA uses the following descriptive words:

Lower value in PSI	Upper value in PSI	Descriptor category
0	50	Good
51	100	Moderate
101	199	Unhealthful
200	299	Very Unhealthful
300		Hazardous

With the above definition of API a so called subindex can be created for each substance, where the break points can be defined in different ways depending on the chosen substance.

In order to create a composite index according to the USEPA the highest subindex is chosen, i.e.

Total API = max(API1,API2,...)

By using the function **emax** the composite index can be decided.

Appendix 3B: The Stations in the Reference Database

The following table shows a summary taken from the station database for the reference database. Example, see *Figure 3B.1*.The first column shows the station key, the internal name for the station. The second column shows the station name, which is the name that is used in Indico Presentation and also the Indico data collection module. These names are shown on the map beside their locations. Example, see *Figure 3B.2*.The final column shows the type of measuring equipment that is used at the station.

See *Users Reference Volume 6: Using the Indico Administration Module* for more information about the station database.

Station Key	Station Name	Type of station
GO1	Gamlestaden	DOAS
GO2	Molndal	DOAS
GO3	Rya	DOAS
GO4	Volvo	DOAS
GO5	Jarntorget	DOAS

Station Key	Station Name	Type of station
GM1	Shell	Meteorological mast
GM2	Lejonet	Meteorological mast
GM3	Jarnbrott	Meteorological mast
GM5	Femman	Conventional point monitoring

Figure 3B.1. Stations : key, name and type .

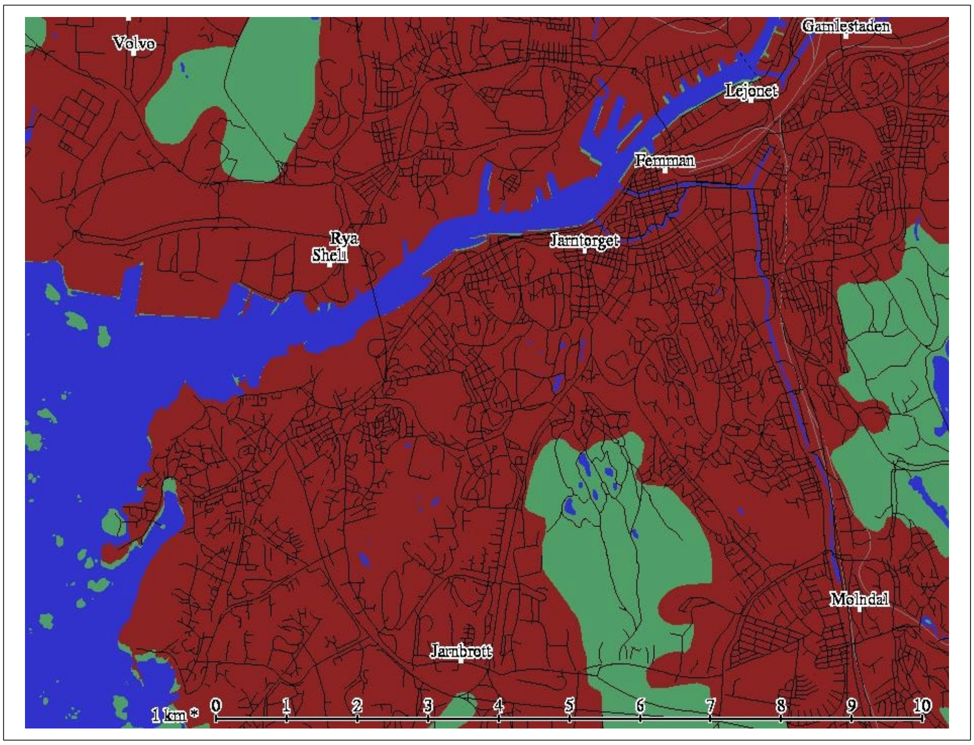


Figure 3B.2. Stations Map.

Appendix 3C: Waved

3.C.1 Introduction

3.C.1.1 What is Waved?

Waved is a tool that integrates the Airviro time series database into MS Excel®. The following tasks can be easily performed:

- With Waved you can use the whole power of MS Excel® with data from the fast and compact time series database of Airviro. Once you have the data in your MS Excel® workbook, you can either use the excellent reporting features of MS Excel® or easily cut and paste the data to other reporting tools.

- With Waved you can store any time series data in the Airviro database. You are not limited to the storage of data collected with Airviro.

- With Waved it is easier to edit data. You just export the data to MS Excel®, make the changes there and import the data back to Airviro.

3.C.1.2 How does it work?

Dialogs for Airviro time series database access are added to the MS Excel® interface. Just choose import or export, select a number of time series from Airviro and the transfer will take place instantly. No storage device, no difficult commands to get data to and from Airviro, just a few clicks. The transfer of data between MS Excel® and Airviro is done either directly through the local area network or by a dial-up network modem connection.

3.C.2.Overview and definitions

Waved uses the same structure as the time series database of Airviro. The Airviro time series database consists of four sub tables: Station, parameter, instance and value type as well as the time series database itself. Each value in the Airviro time series database references these tables. A set of measurement values in the time series database that references the same station, parameter, instance and value type is called an existing time series.

Example: All the measurement values for the station Femman, parameter NOX, instance 010 [M] and value type Status is a time series. The values referencing the station Femman, parameter NOX instance 010 [M] and the value type Value is another time series.

When Waved is started, all the existing times series are loaded into Waved. All the stations, parameters, instances and value types are loaded as well. MS Excel®, a registered trademark by Microsoft Corporation, will be referred to as Excel in the rest of the document.

3.C.3.Getting Started

The following steps are needed in order to use Waved:

- Install Waved on your computer.
- Start Excel. Click on **Complements**.
- To transfer data from Airviro to Excel click on the **Import to Excel from Airviro** in the **Waved** drop down menu. The **Login** dialog will appear the first time Waved is used in a working session in order to allow the user to enter the Airviro user and password.

The procedure to transfer data from Excel to Airviro is very similar.

3.C.4. The Waved menu in Excel

When Waved is installed the Waved menu is added in the Excel menu bar.

The following menu items are available:

- Import to Excel from Airviro: Displays the import dialog.
- Export from Excel to Airviro: Shows the export dialog.
- Host: shows the IP adress from server to connect to.
- About: Displays information about Waved. (Current Version, Developer, etc.)

3.C.5. Database and time resolution

Click on the database you want to select. The existing time resolutions for that database are shown in the time resolution list. Select one of the time resolutions. See *Figure 3C.1. Database and Time Resolution.*

Figure 3C.1. Database and Time Resolution.

The database and time resolution can be set either from the import or the export dialog. If the database and/or time resolution is changed in the export dialog, it will change in the import dialog as well and vice versa.

3.C.6. Import to Excel from Airviro

In the **Import to Excel** from Airviro dialog the time series that are selectable are the existing time series. By selecting a station, all the parameters for existing time series for that station are shown. Then, by selecting a parameter, all the instances for existing time series for that station and parameter are shown. Finally, by selecting an instance, all the value types for existing time series for that station, parameter and instance are shown. See *Figure 3C.2. Import from TSDB.*

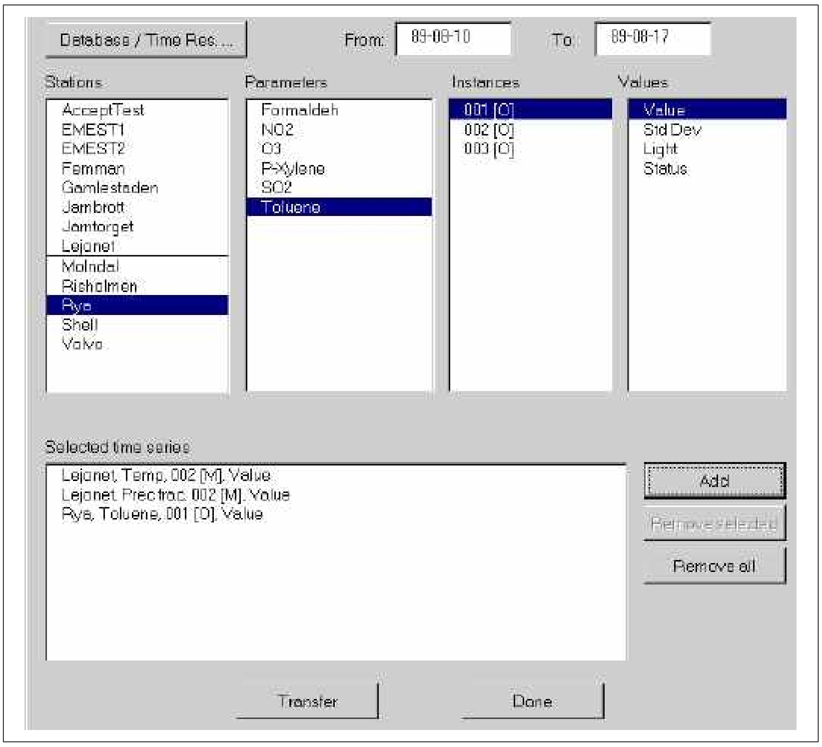


Figure 3C.2. Import from TSDB.

The **Import to Excel** from Airviro dialog of Waved is organised in three parts: The top, the middle and the bottom.

The top allows the user to select database and time resolution and to enter the time period that will be transferred from Airviro to Excel. The format of the from/to dates is the local Windows format that can be changed in the control panel.

An existing time series is specified in the middle part. A time series specification consists of a station, a parameter, an instance and a value. All of them must be selected in order to add the specification to the collection of selected time series. The specification is made from left to right. The stations shown are the ones that have data in the time series database. When a station is selected the parameters in the time series database for that station are shown. When a parameter is selected the instances in the time series database for that parameter and station is shown. Finally when an instance is selected the values in the time series database for that instance, parameter and station are shown.

The bottom part contains the selected time series. To add a specified time series to the selected ones press **Add.** To remove a selected time series, select it and press **Remove selected**, to remove all selected time series, press **Remove all**.

When **Transfer** is pressed the selected time series will be transferred from Airviro to Excel. The data is inserted into the active work sheet in Excel starting with the active cell.

The first row of the inserted data contains column heading for each imported time series. The following rows contain a date/time stamp and the data for each of the time series.

3.C.6.1. An example of import to Excel

The following example is from Gothenburg. Time series data from the DOAS station Järntorget, that is located in the middle of Gothenburg, is imported to Excel from Airviro. The imported time series are Value, Standard deviation, Light and Status for NO2.

1. Position the active cell in Excel where the data should be inserted.

2. Select **Waved** and **Import to Excel** from Airviro in the Excel menu bar.

3. If this is the first time Waved is used in the Excel session the **Login** dialog will be shown and the user must enter a password and optionally a user name.

4. The **Import to Excel from Airviro** dialog is shown.

5. Press **Database / Time Res** and select the database Luft and the time resolution 1 hour.

6. Enter 90-12-12 in the **From** text field and 90-12-14 in the **To** text field. NOTE: The format of the date is depending on your regional settings in the **Control panel**.

7. Select Järntorget in the station list, NO2 in the parameter list and 001 [O] in the instance list.

8. Select Value in the value list and click Add. Repeat the manoeuvre for Std Dev, Light and Status.

9. Press **Transfer** to import the data to Excel from Airviro.

3.C.6.2.Limitations

The number of rows in an Excel spreadsheet is limited to 65535. This corresponds to a little bit more than 7 years of hourly data.

3.C.7. Export from Excel to Airviro

In the **Export from EXCEL** to Airviro dialog the stations, parameters, instances and values that are shown in the lists are the ones from existing time series. It works just like the **Import to Excel** from Airviro with the following exception: By clicking **New station**, **New Parameter** and **New Instance** it is possible to select stations, parameters and instances that are not a member of any existing time series. NOTE: It is not possible to create new stations, parameters or instances, only to select stations, parameters and instances that are not part of an existing time series.

Example: The parameters NOX and NO2 exist in the parameter database. The stations

Femman and Järntorget exists in the station database. Two time series exists: Femman, NOX,010 [M], Value and Järntorget, NO2, 010 [M], Value. The stations shown in the station list are Femman and Järntorget. By clicking Femman in the station list, NOX is shown in the parameter list. NO2 does not appear in the parameter list because it does not exist a time series for the station Femman that contains NO2. However, by clicking on **New station**, a dialog is shown that contains all the parameters in the parameter database, i e NOX and NO2. By selecting NO2 and click **OK** the parameter NO2 is added to the parameter list in the **Export from EXCEL** to Airviro dialog. No instances are shown because no time series exist with the station Femman and the parameter NO2. It is necessary to choose New Instance in order to add one.

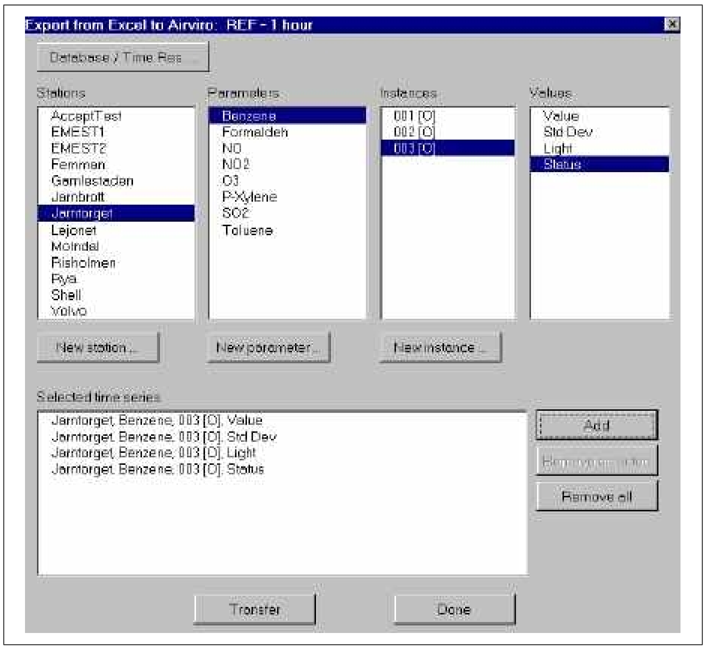


Figure 3C.3. Export from TSDB.

The **Export from Exce**l to Airviro dialog of Waved is organized in three parts: The top, the middle and the bottom.

The top allows the user to select the Airviro database to export to and the time resolution.

A time series is specified in the middle part. A time series specification consists of a station, a parameter, an instance and a value. All of them must be selected in order to add the specification to the collection of selected time series. The specification is made from left to right. The stations shown are the ones that have data in the time series database. If **New station** is clicked a dialog is shown where the user can select any existing station. When a station is selected the parameters in the time series database for that station are shown. A parameter in an existing time series for the selected station can be chosen in the parameter list or the user can click on New Parameter in order to choose any parameter in the parameter database. When a parameter is selected the instances in the time series database for that parameter and station are shown. The instance can be selected from instance list of from the **New Instance** where any instance can be chosen. Finally, when an instance is selected the values in the time series database for that instance, parameter and station are shown.

The **New station**, **New parameter** and **New Instance** allow the creation of time series that do not exist in the time series database. See *Figure 3C.3. Export from TSDB.*

The bottom part contains the selected time series. To add a specified time series to the selected ones press **Add**. To remove a selected time series, select it and press **Remove selected**, to remove all selected time series, press **Remove all**.

When **Transfer** is pressed, the selected time series will be transferred from Excel to Airviro. The *active cell* in the Excel document must be the *first* date / time cell. The second column should contain the data of the first selected time series and so on.

NOTE: The date / time column must be in the Excel date format.

3.C.7.1 New station

Only the stations with existing time series are shown in the station list of the Export from **Excel to Airviro** dialog. To export data to a station that does not have any data in the time series database click on **New station**. A dialog containing all the stations in the station database of Airviro is shown. Select the wanted station and click on **OK**. The selected station immediately occurs at the bottom of the station list in the **Export from Excel to**

Airviro dialog. See *Figure 3C.4.New station.*

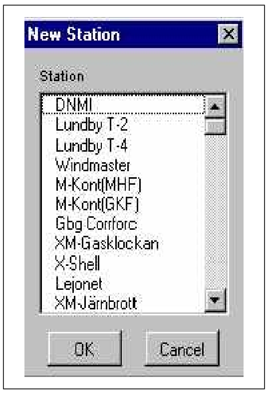


Figure 3C.4.New station.

3.C.7.2. New parameter

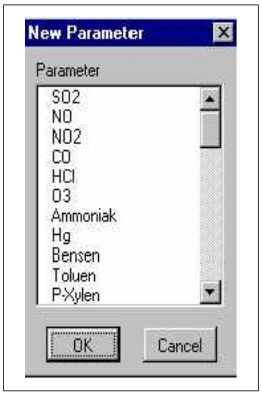


Figure 3C.5. New parameter.

For the selected station only the parameters with existing time series are shown in the station list of the **Export from Excel to Airviro** dialog. To export data to a station and parameter combination that does not have any data in the time series database click on **New paramete**r. See *Figure 3C.5. New parameter.*

A dialog containing all the parameters in the parameter database of Airviro is shown. Select the wanted parameter and click on OK. The selected parameter immediately occurs at the bottom of the parameter list in the **Export from Excel to Airviro** dialog.

3.C.7.3. New instance

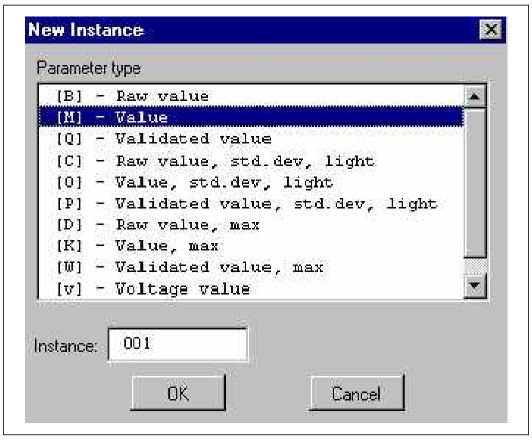


Figure 3C.6.New instance.

For the selected station and parameter only the instances of existing time series are shown in the station list of the **Export from Excel to Airviro** dialog. To export data to a station and parameter combination that does not have any data in the time series database click on **New instance**. A dialog containing all the possible instances in Airviro is shown. Select the wanted instance and click on OK. The selected instance immediately occurs at the bottom of the instance list in the **Export from Excel to Airviro** dialog. See *Figure 3C.6.New instance.*

Normally data should be imported to the [M] - Value parameter type. The table below shows the use of the other parameter types:

Parameter type	Three step database only	OPSIS only	NILU only	Calibration database only	Number of parameters (excluding status)
[M] - Value					1
[B] - Raw value	X				1
[Q] - Validated value	X				1
[C] - Raw Value, std. dev, light	X	X			3
[O] - Value, std. dev, light		X			3
[P] - Validated value, std. dev, light	X	X			3
[D] - Raw value, max	X		X		2
[K] - Value max			X		2
[W] - Validated value, max	X		X		2
[v] - Voltage value				X	1
[V] - Voltage value, max			X	X	2

Figure 3C.7. Parameter types.

The table (*Figure 3C.7. Parameter types)* shows which parameter types that should be used only when the three step database is used, when OPSIS values are exported from Excel, when NILU values are exported from Excel and when the calibration database is used. The last column states the number of parameters that must be exported.

Example: For the [O] parameter type: value, standard deviation and light must be exported at the same time.

3.C.7.4. An example of export from Excel

1. Position the active cell in Excel at the first row of the date / time column,

2. Select Waved and **Export from Excel to Airviro** in the Excel menu bar.

3. If this is the first time Waved is used in the Excel session the user must enter password and optionally a user name.

4. The **Export from Excel to Airviro** dialog is shown.

5. Press **Database / Time Res** and select the database Luft and the time resolution 1 hour.

6. Select Järntorget in the station list, NO2 in the parameter list and 001 [O] in the instance list.

7. Select Value in the value list and click **Add**. Repeat this step or Std Dev, Light and Status.

8. Press **Transfer** to export the data from Excel to Airviro. See *Figure 3C.8.Example export.*

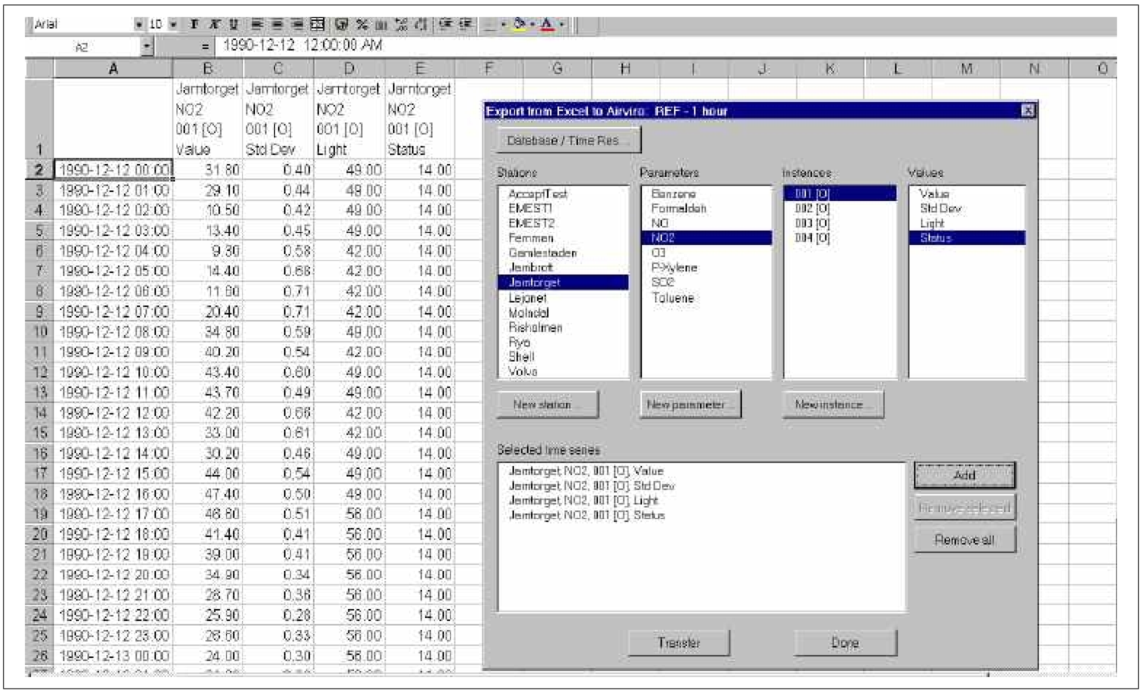


Figure 3C.8.Example export.

3.C.7.5. Limitations

All the values of a parameter type must be exported at the same time.

3.C.7.6. Setting up privileges for export from Excel

The same privileges are valid for Waved as for the Airviro time series editor. The privileges for the Airviro time series editor are set up in the file *priv.rf*, normally located in*/usr/airviro/rsrc*.

3.C.7.7. Pitfalls with export from Excel

These are the most common errors that are made when exporting data from Excel to Airviro:

1. The active cell in Excel must be the first row of the date / time column.

2. The date / time column must be in Excel date format.

3. The database that data was imported to, was not in the scan list of avdbm. Check it with *avstat.*

CAUTION: It is very easy to export data from Excel to the Airviro Time Series database using **Waved**. Care must be taken so that data is not corrupted. The best way to avoid corrupted databases is probably to set up a parallel import database.

3.C.8. Waved as a database editor

The design of Waved enables it to be used as a database editor. The selection of time series made in one of the dialogs (export, import) is transferred to the other dialog when it is opened.

The steps to perform in order to use Waved as a database editor are:

1. Open the **Import to Excel** from Airviro dialog.

2. Select the time series that you want to edit. Remember to export the status if you want to set it to status 15, manually changed.

3. Transfer the data to Excel by pressing **Transfer**.

4. Close the **Import to Excel from Airviro** dialog and make your changes to the data.

5. Set the first date /time row as the active cell in Excel.

6. Open the **Export from EXCEL** to Airviro dialog.

7. Press **Transfer.** The data is exported to Airviro.

3.C.9. Technical specification

Waved includes the following features:

- No installation of any programs on the Airviro workstation. The only installation of software is done in the client PC as a plug-in module to Excel.
- Extremely fast transfer of data due to the fact that the low level C interface of Excel is used. One year of hourly data for four stations is exported to Excel in a few seconds.
- Export of 32 simultaneous time series.
- Import of 32 simultaneous time series.
- Allows creation of new time series. A time series can be created for any existing station and parameter. New instances can be created.