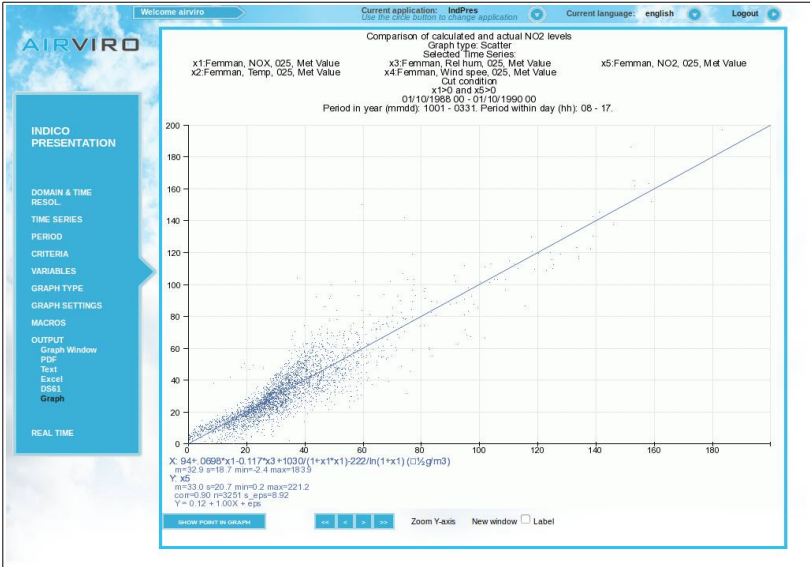




Airviro User's Reference



Working with Indico 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
4.01	May 2019	Upgrade	GS
5.01	Oct2019	Review	GS
5.01	Dec 2020	Review	DC

CONTENT

3.1 INTRODUCTION TO INDICO PRESENTATION.....	5
3.1.1 WHAT IS INDICO PRESENTATION?.....	5
3.1.2 HOW DOES INDICO PRESENTATION WORK?.....	6
3.2 GETTING STARTED.....	7
3.2.1 OVERVIEW OF THE INDICO PRESENTATION MAIN WINDOW.....	7
3.2.2 TIME SERIES DATABASE.....	7
3.2.3 SELECTING TIME SERIES.....	8
3.2.4 SELECTING TIME PERIOD.....	11
3.2.5 CONSTRAINING CASES (CRITERIA MENU).....	12
3.3 MATH EXPRESSION COMPILER (VARIABLES MENU).....	14
3.3.1 TRANSFORMING VARIABLES.....	17
3.3.1.1 <i>Handling missing values</i>	19
3.3.1.2 <i>Counting</i>	20
3.3.2 CREATING NEW VARIABLES.....	21
3.3.3 <i>Modifying a series by smoothing or differencing</i>	22
3.4 PRESENTING GRAPHS.....	24
3.4.1 CONTROLLING THE LAYOUT OF A GRAPH.....	26
3.4.2 DISPLAYING THE GRAPH.....	28
3.4.3 AVAILABLE GRAPH TYPES.....	30
3.5 REGRESSION MODELLING.....	42
3.5.1 LINEAR REGRESSION MODEL.....	43
3.5.1.1 <i>Fitting a curve</i>	51
3.5.2 BINARY LOGISTIC REGRESSION MODEL.....	52
3.6 FACTOR ANALYSIS.....	54
3.6.1 PRINCIPAL COMPONENT ANALYSIS.....	57
3.7 USING INDICO MACROS.....	59
3.8 INDICO REAL TIME.....	61
APPENDIX 3A EXPLOITING THE MATHEMATICAL FUNCTIONS FOR	

CALCULATION PARAMETERS.....	63
3A.1 LOGICAL FUNCTIONS.....	64
3A.2 TIME SHIFT FUNCTIONS.....	64
3A.2 SPECIAL VARIABLES.....	65
3.A.3 MATHEMATICAL FUNCTIONS.....	65
3.A.3.1 <i>Combining Formulae</i>	65
3.A.4 MISSING DATA VALUES.....	67
3.A.5. DEFINITION OF THE AIRVIRO AIR POLLUTION INDEX.....	68
APPENDIX 3B: THE STATIONS IN THE REFERENCE DATABASE.....	71
APPENDIX 3C: WAVED.....	73
3.C.1 INTRODUCTION.....	73
3.C.1.1 <i>What is Waved?</i>	73
3.C.1.2 <i>How does it work?</i>	73
3.C.2.OVERVIEW AND DEFINITIONS.....	73
3.C.3.GETTING STARTED.....	74
3.C.4. THE WAVED MENU IN EXCEL.....	74
3.C.5. DATABASE AND TIME RESOLUTION.....	75
3.C.6. IMPORT TO EXCEL FROM AIRVIRO.....	75
3.C.6.1. <i>An example of import to Excel</i>	77
3.C.6.2. <i>Limitations</i>	78
3.C.7. EXPORT FROM EXCEL TO AIRVIRO.....	78
3.C.7.1 <i>New station</i>	80
3.C.7.2. <i>New parameter</i>	81
3.C.7.3. <i>New instance</i>	82
3.C.7.4. <i>An example of export from Excel</i>	84
3.C.7.5. <i>Limitations</i>	85
3.C.7.6. <i>Setting up privileges for export from Excel</i>	85
3.C.7.7. <i>Pitfalls with export from Excel</i>	85
3.C.8. WAVED AS A DATABASE EDITOR.....	86
3.C.9. TECHNICAL SPECIFICATION.....	86

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. Together with Indico Real Time module, the most up to date data can be continuously displayed on the screen keeping you informed about the latest air quality situations.

In this manual 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 Airviro package.

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

3.1.1 What is Indico Presentation?

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 when 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 work?

This Airviro module runs on any PC or other devices running Internet Explorer 6 or later, Firefox or any other mozilla based browsers

After logging in to Airviro with your user-ID and password, a domain must first be selected, and then Indico Presentation module should be chosen from the available modules.. All data processing is made on the Airviro server and afterwards the results are sent to the web browser.

. Airviro version v5.00 or higher don't need Java Runtime Environment to run.

3.2 Getting started

Once Airviro has been properly installed on the server, you can start using it by typing the correct URL in your web browser over the Intranet/Internet.

After logging in to Airviro with your user-ID and password, a domain must first be selected, and then **Indico Presentation** should be chosen from the available modules.

At the top of the window, in Current application the name of the module actually selected is shown. Clicking on the [down arrow] button beside the name other modules can be selected.

Clicking on the Current language [down arrow] button it is possible to select the language to work with. This functionality is not yet implemented.

By clicking on the **Logout** [down arrow] button, the current module is closed and the Airviro login page is displayed instead.

3.2.1 Overview of the Indico Presentation main window

Once Indico Presentation has been loaded into the web browser, you will see the main menu options on the left side. To complete a setup you must go through all the submenus from Domain & Time resolution down to Output, except for Macros and Real Time that will be discussed later. It is preferable to work through the submenus sequentially, because some settings may depend on previous choices, e.g. settings in Criteria are based on the selections made under Time Series, entries in Graph Settings make reference to definitions in Variables etc.

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.

Incoming data may arrive every minute into a raw database to be filtered and post-processed to half-hour means, hourly means and daily means in a continuous process. Mean values and sums may be calculated and stored.

Incoming data may, on the other hand, arrive once a year from some other source to be imported into the time series database

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

All these data are gathered and organized into the time series database.

3.2.3 Selecting time series

In **DOMAIN & TIME RESOL**, you can select a domain and a time resolution to work with. The domain chosen after the login is selected by default. Different domains, will contain different time series and macros. See *Figure 3.2.3.1*.

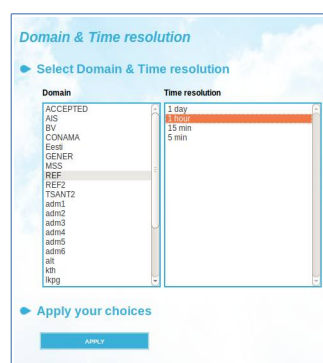


Figure 3.2.3.1 Domains and time resolutions.

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

In **TIME SERIES** a list of all stations -(both active or inactive)- and all observed parameters are displayed. This information should have been previously defined with Indico Administration. By clicking on a name in the station list box, a list of parameters measured by that station are displayed. The **[CLEAR]** button clears your selection. On the other hand, by clicking on a name in the parameter list box, all the stations measuring that parameter are displayed.

Stations and parameters can be sorted alphabetically or by key. Checking/unchecking **Reverse** refreshes the sort order accordingly. By ticking the check box **Active first**, active stations will be shown at the top of the station list.

Once you have selected both station and parameter, the other Instance, Attribute and Unit lists are automatically filled in with the corresponding data. 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. This letter is a code corresponding to a parameter type. Letter M or v indicates that you store a measured value and a status flag for the actual instance. M is the scale data and v is the raw value of the measurement as measured by the instrument. Letter O or P indicates that you store standard deviation and light intensity (for DOAS analysers) as well as the measured value.

The status flag is assigned during the quality control made when the incoming data is stored in the time series. The status condition options, can be seen in menu **CRITERIA >> STATUS CONDITIONS**.

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

Once you have chosen a station, parameter, instance and attribute from the lists, the time series is uniquely identified (for the current time resolution). Click on the **[ADD]** button to select the time series for further processing. You can select up to 64 time series .

[REMOVE SELECTED] button deletes a highlighted time series from the “Selected” list box . **[REPLACE]** button replaces a highlighted time serie in the “Selected” list box with another time series. . **[CLEAR ALL]** button deletes all the time series from the “Selected” list box. **[Show keys]** button displays the station keys for your “Selected” time series. Multiple key are backslash separated.

Up to 128 time series can be displayed in a graph Click on the **[APPLY]** button to your settings.

3.2.4 Selecting Time period

By default, Indico Presentation pre-selects the last week of data. The options within the **Period** menu, allows you to change the time frame. Date & Time formats available are: European, UK and US . Buttons [**<<**] and [**>>**] allow you to copy a datetime from one textbox into another. Buttons [...] display a calendar to select a date from, buttons [**-**] [**+**] allow you to change respectively the year, month, day, etc. Button [**PRESENT**] sets the current datetime from the server. (see *Figure 3.2.4.1*)

Click [**APPLY**] to save your settings. To check if there is any data for that period for the Time series selected , then, on the main menu, click on **GRAPH CONTENTS** .

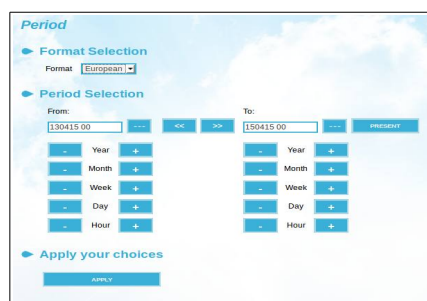


Figure 3.2.4.1 Period menu with start date and end date in European format

The hour starting at 00:00 and ending at 01:00 is written as 01. The hours are numbered from 01 to 24. This means that the hour starting 23:00 and ending 00:00 is written as 24. Make sure that all the stations are in the same time zone and preferably running without day light savings time (Winter time) . , otherwise you may get missing or duplicate data . , or 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 interface; they are always 00 here, which can be slightly confusing. If you specify a time period of one day for a time series with 15-minute resolution, the observations will be 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 for the From textbox and exclusive for the To textbox,

or in mathematical terms:

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

3.2.5 Constraining cases (CRITERIA MENU)

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

For instance, if you 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 identified by their name, all-inclusive. If you want observations from 22:00:00 until 22:59:59, you should then specify the **period within day** from 23 to 23. If you want all observations except that hour, specify the period from 24 to 22. See *Figure 3.2.5.1*.

Click on [**APPLY**] to save your settings. The option **Text** in the **OUTPUT** menu will display these settings as part of the header. It is instructive to look at 15minute data in this way. All criteria related to date and time or days is usually intended to exclude some data 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 using the option **Day type**. National holidays are specified in the resource file named `calendar.rf` in the Airviro server. .

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. These constraints or formulas should be written into the text box “**Condition Selection**”. The formula variables are named x1 to x64 according to the position of the time series in the “Selected” list box, (being x1 the one at the top of the list, x2, the following one, and so on) see *Figure 3.2.5.1*. You can compare your variables against constants or expressions in a Boolean formula. Statements can be combined with logical operators OR(), AND(&) or NOT(!). A quick help can be found in the **VARIABLES** menu, by clicking on the **Help** link. See further chapter 3.3 *Math expression compiler*.

Make sure that you have set your criteria with the correct units. If you have doubts about what unit is used, compare it against the parameter database, in 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 on how to write a condition formula. If the Boolean expression is false, all variables in that case will be left without a value.

Criteria

◆ Period Selection

Period within year

Active From 0101 To 1231

Period within day

Active From 01 To 24

Weekday Friday Day type Holiday (Sunday)

◆ Condition Selection

(x1>8 OR x1<2)&NOT EXIST(x2)

STATUS CONDITIONS CLEAR ALL

◆ Apply your choices

APPLY

Figure 3.2.5.1 Criteria menu with some criteria for constraining access to the selected cases.

Every data value has a status condition assigned to it, either by Indico Administration or by the external protocol. See Indico Administration manual for more details on how status conditions are assigned.

Time series can be constrained to data matching certain status conditions by clicking on the **[STATUS CONDITIONS]** button and then ticking on the required status codes. Data with status code "Checked – OK" and "Manually changed" are pre-selected by default, but can be unselected. Click on **[OK]** button to confirm 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 resets all criteria settings, including Status conditions.

3.3 Math expression compiler (VARIABLES MENU)

Time series can be displayed in a graph or used for statistical analysis. You can work separately with plot variables and statistical variables, using the same time series.

Your goal should be to extract as much information as possible from your data. This can be achieved by presenting descriptive statistics charts, exceedance statistics charts, distribution functions, correlation between stations; time series analysis for seasonal variation and trend using univariate or multivariate techniques in an attempt to improve your knowledge about a particular situation. In this way, you can get an understanding of the air quality situation, how it is related to different source areas and how the concentration varies according to the time of the day or season, or by meteorological factors. Then, you can use them to build a model that explains these variations.

In parallel, you can work with models in an attempt to describe known emissions of pollutants and their effect on the concentrations at a receptor point. 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 modelling.



Figure 3.3.1 Variables menu with three plot variables for two time series.

Synchronise 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. You can define the percentage of data that must at least be included for this calculation. This percentage is specified in the text box labelled **Required**

Ticking off **Show each scale separately** the scale labels for each parameter are drawn with the same color as the line colour of the corresponding time series. Leaving it unchecked, only one scale is drawn and will group the times series with the same scale together. Which scale goes with which time series will be indicated above each scale.

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** menu, all time series selected are listed.; See *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 x_1, x_2, \dots, x_{64} 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 and the formula if you want. This will have an effect on plotting of values. The **min-max** values refer to the scale in a graph; **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.

In the formula field you can enter arithmetic or conditional Boolean expressions. Examples of an arithmetic expressions are $x_1 * 1000$ or $\text{emax}(x_1 : x_5)$. An example of a conditional Boolean expression is $(x_1 > 0.65) ? x_1 : @$. The expression preceding the question mark is Boolean. If it is true, the plot variable gets the value x_1 , 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 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 as those 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 maximum 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 $\mu\text{g}/\text{m}^3$ into ppb(v), you need to know the 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

$$\text{var} = x_1 * 28.97 / (M * 1.2929) * \frac{x_2 + 13.273}{273.13} ,$$

if x1 is mass concentration in $\mu\text{g}/\text{m}^3$ and x2 is temperature in °C.

Another example is to use a regression model for some transformation process, e.g. from NO_x to NO₂, if it is validated.

2) If you want to 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.

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

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 $x < \{l_1, l_2, l_3, \dots, l_n\}$.

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

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 gather 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 x_1 , x_2 , x_3 are status codes for three observations. For each time step you get the number of undetectable concentrations.

If 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 server script 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:

$$\text{Trend} = (0.5 \cdot x1[-12] + x1[-11] + x1[-10] + \dots + x1[-1] + x1 + x1[1] + \dots + x1[11] + 0.5 \cdot x1[12]) / 24.$$

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

$$\nabla x_1 = \{x_1 - x_1[-1]\} = (1 - B)x_1,$$

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

$$\nabla_d x_1 = \{x_1 - x_1[-d]\} = (1 - B^d)x_1$$

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

3.4 Presenting graphs

In the **Graph Type** menu, you can select different presentation and 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 modelling* or chapter 3.6 *Factor analysis*.

In the Graph type menu, a list of the time series that you have already selected is

presented there for your convenience, because in some presentations or analysis types, you have to specify which variable is the dependent one. The dependent variable may also refer to a formula from the Variables menu.

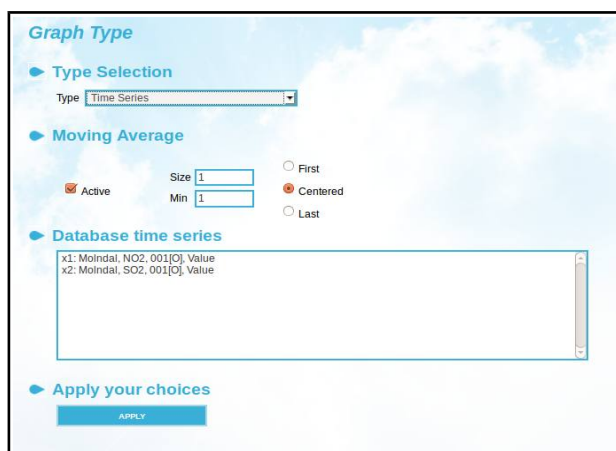


Figure 3.4.1 Graph type.

Different graph types will require different sets of data to be entered.

If you haven't previously defined a moving average with a formula, you can do it from the Graph type menu. Please be aware that if you do it at this stage, you won't be able to assess - by only looking at the text on the graph - if it has been applied or not.

The **Moving Average** defined in the Graph type menu 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 text box **Size** 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 on [**APPLY**] to commit your settings in Graph Type.

3.4.1 Controlling the layout of a graph

In the **Graph Settings** menu, you can specify a graph header 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 graph. You can set line type, line width and line colours.

Firstly, select the variable to graph 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 its appearance according to your settings. Once 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 graph, 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.

The **Missing values** option allows you to leave blanks in the graph where missing values are found, or to interpolate values to create a continuous line. *Figure 3.4.1.1 Graph Settings* .

The **Y-axis labelling** window allows you to choose between automatic and 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 allows you to 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 colour chosen in the colour palette. **Background graph**

will display the background graph with the colour chosen in the colour 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 colour. If you want an ordinate axis and abscissa, check the **X-Axis** and **Y-Axis** boxes.

Also, 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 .

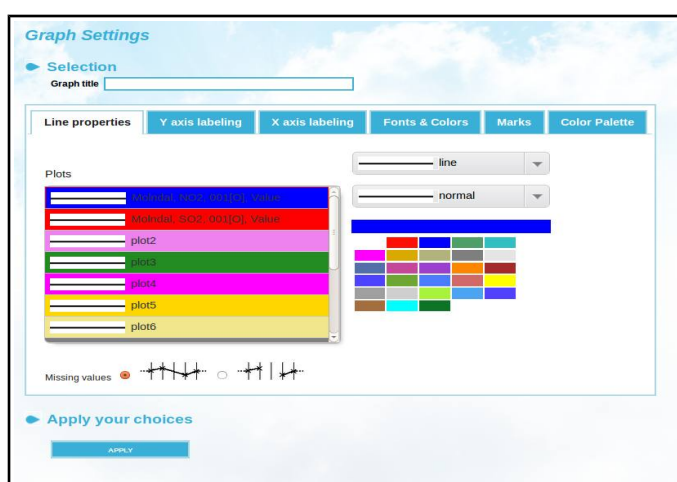


Figure 3.4.1.1 Graph Settings.

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 **Output – Graph**. See *Figure 3.4.2.1* below.

The graph is interactive so that you can **zoom** in/out 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 leave the hair cross mouse pointer over the graph, a label showing the date and value will appear. 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 place the hair cross mouse pointer on the graph at the beginning of the period that you want to zoom in , left click and drag the mouse to stretch a rectangle. Release the button and a new graph with the selected time period is displayed.

Also, you can use the **arrows** to move the graph variables 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 do it in Adobe® PDF-format by selecting **Output – PDF**

To display a graph in PDF-format, you need Adobe® Reader® or other similar Adobe program installed. 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 it , using Adobe functions.

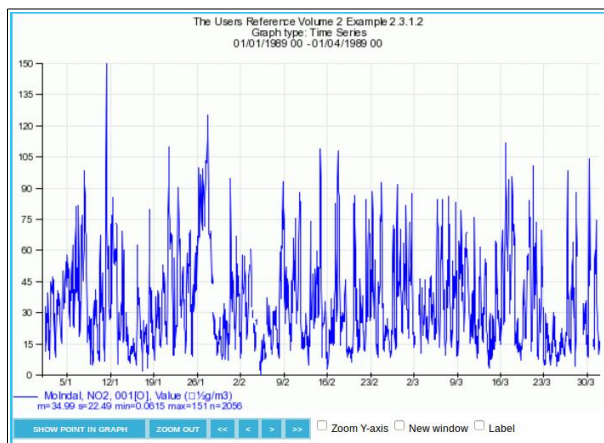


Figure 3.4.2.1 Interactive Output graph in GIF format.

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, unless you can find 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 menu.

Also, the time series can be exported to excel format by selecting **Output – Excel**

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. Time series graph

This is a multiple line graph where the date is represented on the X-axis and up to twelve variables on the Y-axis. Both axes are linear and continuous (see *Figure 3.4.3.1.*) A legend below the graph shows 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 change the offset and Y-axis scale in order 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 display too much data in your graph, as in *Figure 3.4.2.1*, your understanding of the variation will benefit from applying a smoothing filter.

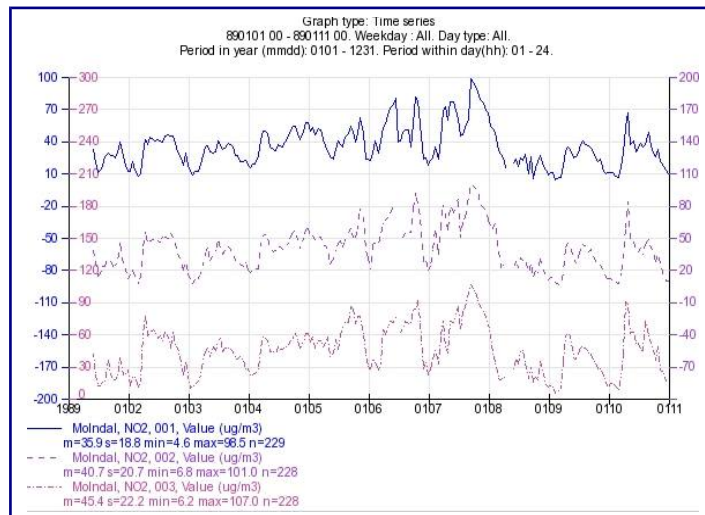


Figure 3.4.3.1 Time series graph with three channels at different offset.

2. Bar Chart

It consists of vertical bars (rectangles) for each value.

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

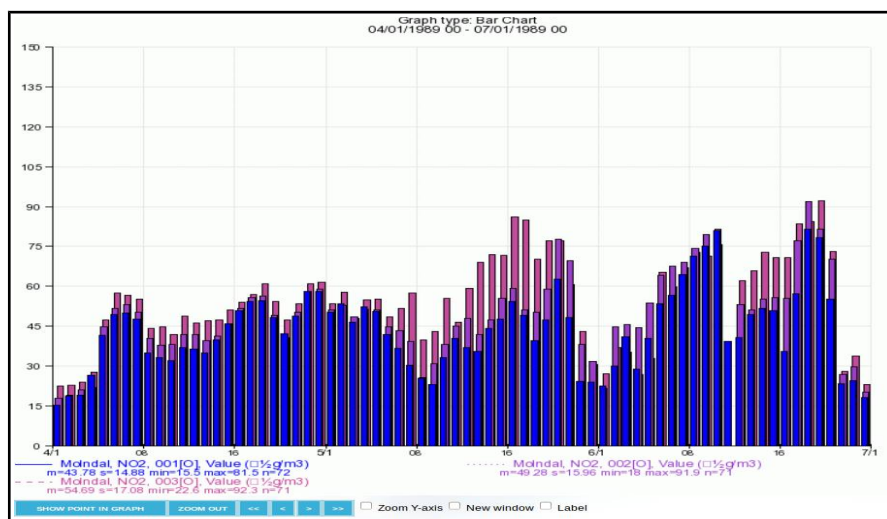


Figure 3.4.3.2 Bar Chart showing values for NO2 at different instances.

3. Filled time series

This is a time series graph with a time scale on the X-axis and one variable on the Y-axis. Both axes are linear and continuous. (see *Figure 3.4.3.1*). A legend below the graph shows 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 coloured according their scale and user settings.

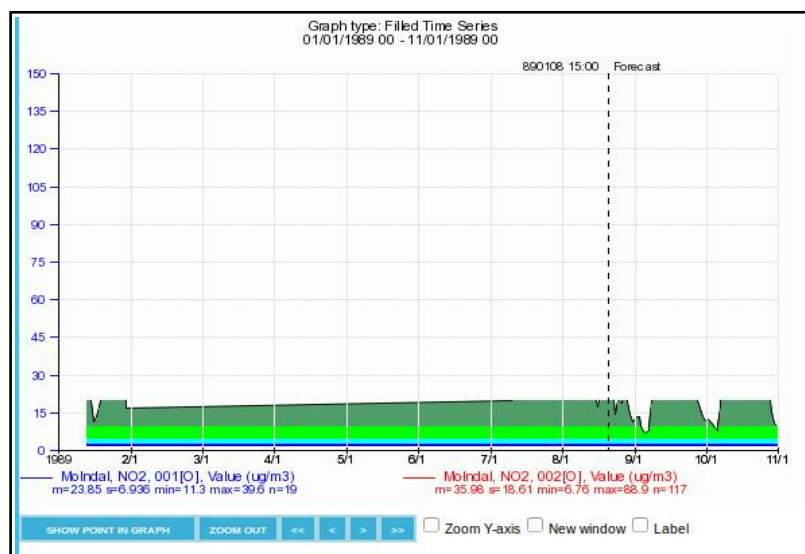


Figure 3.4.3.3 Filled Time Series showing coloured values for NO2 at Molndal station.

4. 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 menu.

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 display up to four plot variables in the frequency distribution graph. The colour of the histogram bar is the same as for the associated line. Multiple bars are clustered together, but they can have different scales.

In the cumulative distribution graph, the Y-axis is always square root-distributed for percentiles from 100 to 0%. The X-axis 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 values, 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*.

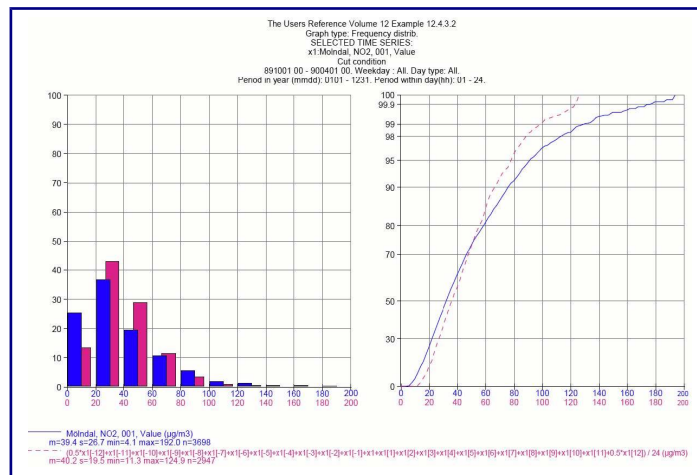


Figure 3.4.3.4 Frequency distribution graph with percentage histogram and cumulative distribution chart. The effect of applying a 24-h moving average can be seen by comparing plot variable 1 and 2.

5. Scatter graph

This type allows you to graph a dependent variable against up to four other variables, one at a time in an XY scatter graph. 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 idea of how much that remains to be explained by other variables.

If you check the **Regression line** box in the **Scatter** section, the regression line will be included in the scatter graph.

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 3.4.3.5* for an example of a scatter plot. The colour and size of the markers can be defined in Graph settings.

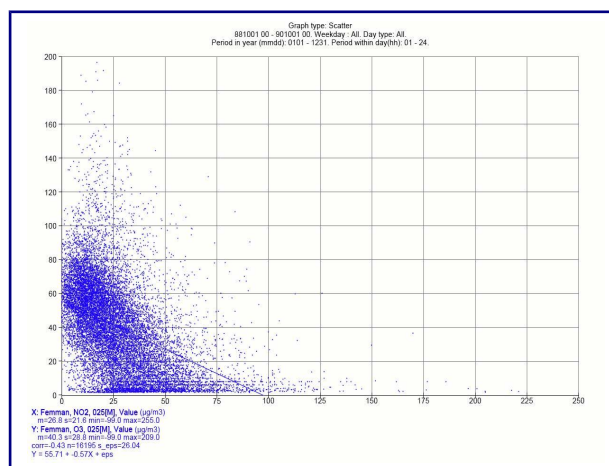


Figure 3.4.3.5 Scatter plot with ozone vs NO₂ concentrations. A regression line is included in the graph.

6. Breuer diagram

This is a polar diagram where up to four variables can be plotted against 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 menu. Negative values for some wind directions 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 be increased.

The Breuer diagram can be used as a pollution rose, which points out to the direction of the 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 idea 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.

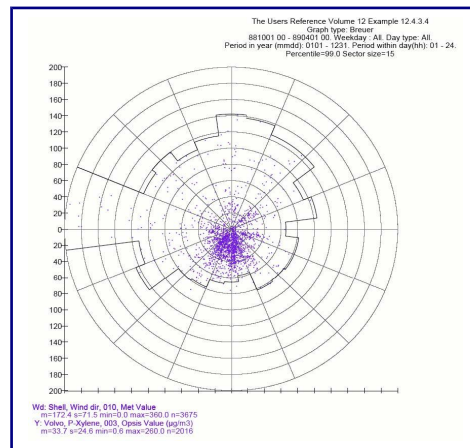


Figure 3.4.3.6 Breuer diagram with P-Xylene concentrations vs wind direction in a polar coordinate system. The 99%-ile concentration in each sector is indicated by an arc.

7. Mean/sector diagram

This 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:

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

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:

$$K(x) = (2\pi)^{-1/2} \exp(-0.5x^2)$$

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.

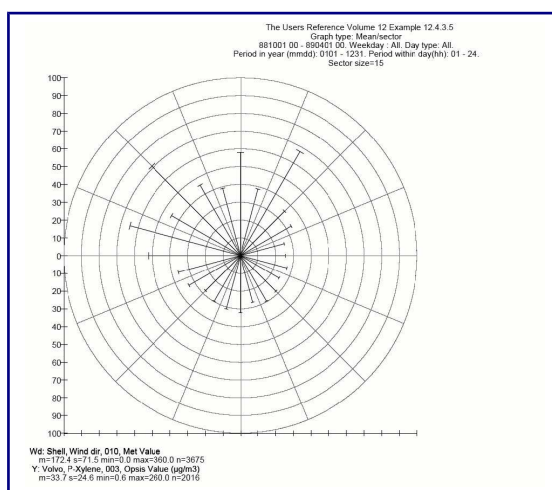


Figure 3.4.3.7 Mean/sector diagram with P-Xylene concentration vs. wind direction.

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. Frequency/sector diagram

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

At least one plot variable is needed, but the frequency/sector diagram always uses the specified wind direction variable (in degrees from north) to calculate and display a wind rose. The wind direction is the compass direction where 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 is lower than 100%.

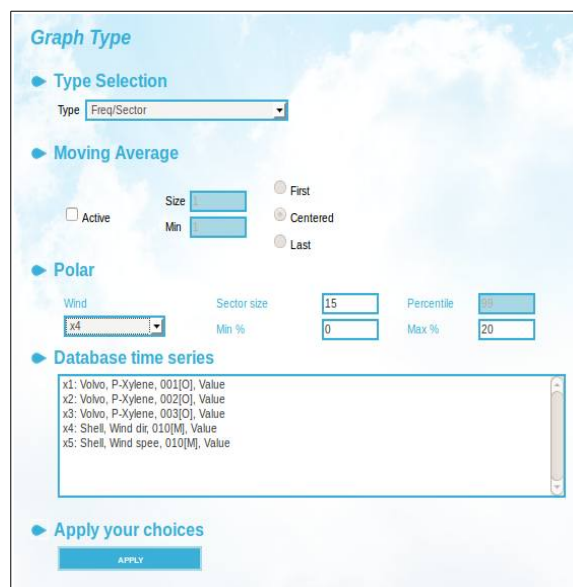


Figure 3.4.3.8 Graph Type menu with settings for a wind rose in the Polar section

9. Diurnal variation diagram

This 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 daytime pattern from different source types. Please remember that the wind often has a diurnal pattern, so it can be a good idea to put the concentrations on flux form multiplying by the wind speed.

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

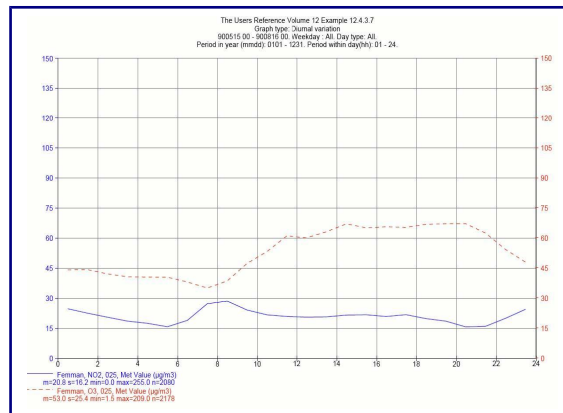


Figure 3.4.3.9 Diurnal variation diagram with anticorrelated NO₂ and ozone concentrations.

10. 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.

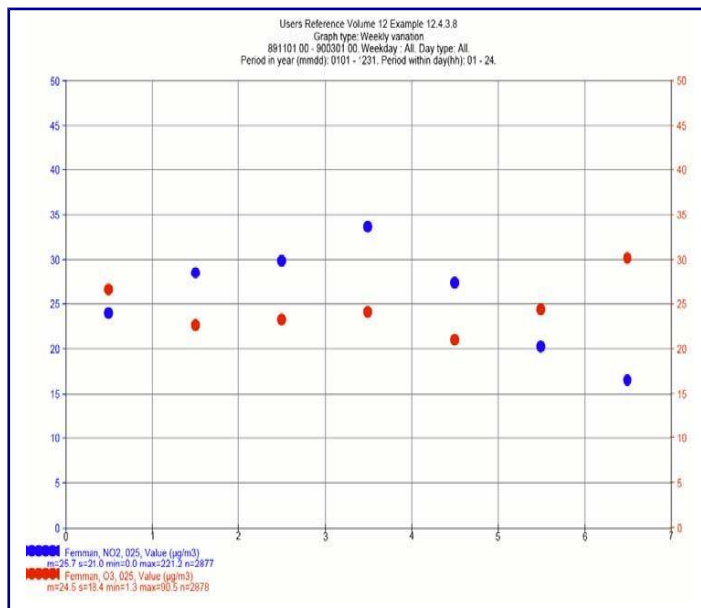


Figure 3.4.3.10 Weekly variation diagram with anticorrelated NO₂ and ozone concentrations. During weekends, the NO₂ concentration is lower than during weekdays.

11. Annual variation diagram

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

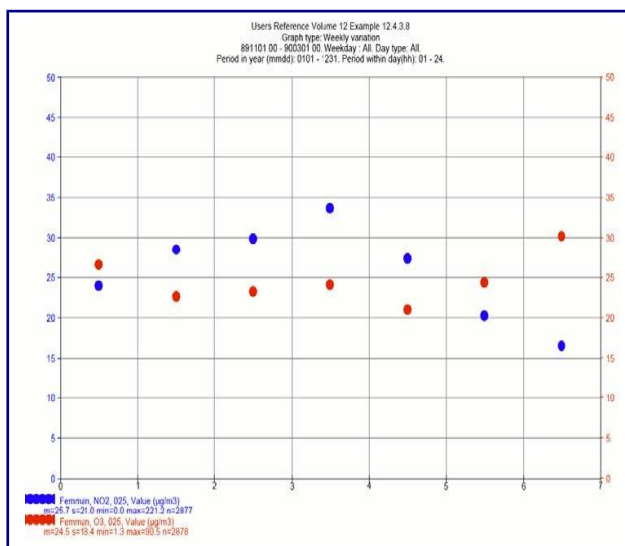


Figure 3.4.3.11 Annual variation diagram with NO₂ and ozone concentrations. Ozone has a very strong seasonal variation, since more ozone is formed in strong sunshine.

3.5 Regression modelling

In the **GRAPH TYPE** menu, there are five different analysis types. Not all airviro installations 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** menu. 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 is 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 be described as:

$$Y = b_0 + b_1X + b_2X + \dots + b_n X_n + \epsilon,$$

Where Y is called the predictand, representing NO_2 or any variable of interest. The X 's, i.e. X_1, X_2, \dots, 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, \dots, 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.

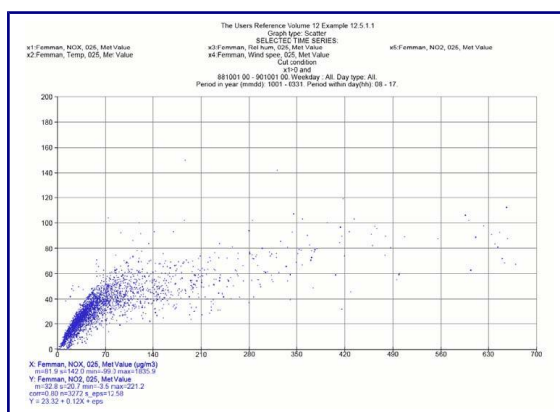


Figure 3.5.1.1 Scatter plot of NO_2 vs. NO_x concentrations.

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 NO_x ($<100 \mu\text{g}/\text{m}^3$) the ratio NO_2/NO_x is 50%-70%, but for large NO_x concentrations ($>500 \mu\text{g}/\text{m}^3$) the ratio is approximately 15%.

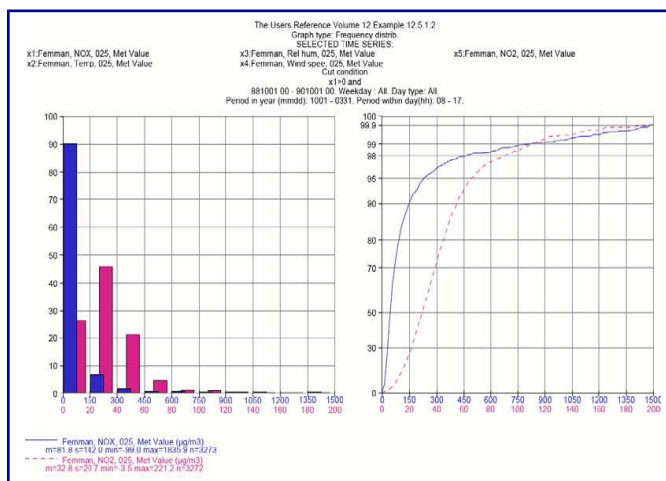


Figure 3.5.1.2 Frequency distribution graph with NO₂ and NO_x concentrations.

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 NO₂. Consequently, any linear model relating NO₂ to NO_x would not be able to describe the basic features of the NO₂ 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** menu, see Figure 3.5.1.3 below.

The MLRF F-criterion scheme is selected in the **GRAPH TYPE** menu. 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.

The screenshot shows the 'Variables' configuration window. Under 'Time Serie Selection', there are checkboxes for 'Synchronize' (checked) and 'Show each scale separately'. Below this is a table with columns: Plot var, Min, Max, Desc, Unit, Formula, Help, Auto, Acc. type, and Perc. The table contains 12 rows. The first two rows are selected (checked):

Plot var	Min	Max	Desc	Unit	Formula	Help	Auto	Acc. type	Perc.
1 x1	0.0	150.0		mg/m3	x1		Fixed	Fixed	
2 x2	0.0	200.0		mg/m3	x2		Fixed	Fixed	
3	0	100					Fixed	Fixed	
4	0	100					Fixed	Fixed	
5	0	100					Fixed	Fixed	
6	0	100					Fixed	Fixed	
7	0	100					Fixed	Fixed	
8	0	100					Fixed	Fixed	
9	0	100					Fixed	Fixed	
10	0	100					Fixed	Fixed	
11	0	100					Fixed	Fixed	
12	0	100					Fixed	Fixed	

Below the table, the 'Database time series' section shows:

```
x1: Molindal_NO2_001[0]. Value
x2: Molindal_SO2_001[0]. Value
```

An 'APPLY' button is located at the bottom of the window.

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*.

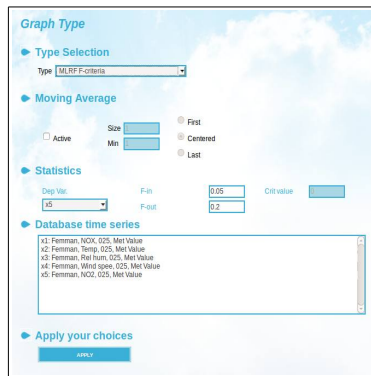


Figure 3.5.1.4 Defining dependent variable and F-criteria in the Graph type frame.

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

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

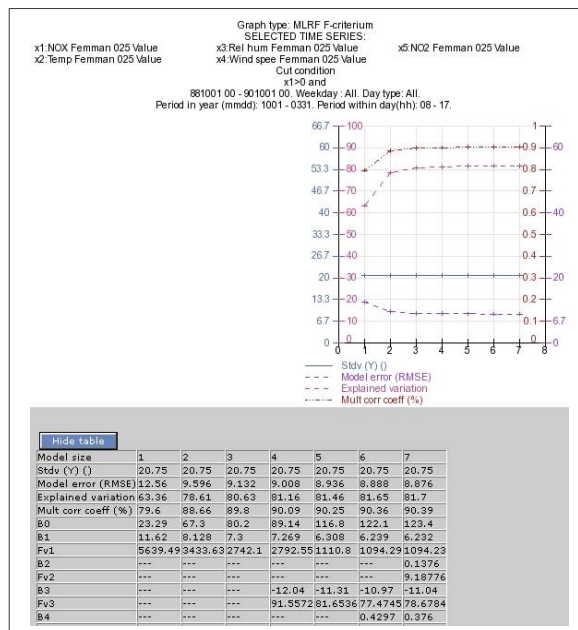


Figure 3.5.1.5 Regression model performance for stepwise increasing model size.

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 menu and press [**APPLY**]. The graph displayed 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 caused by over fitting of the data and it would be unwise to use model size 7 in this case.

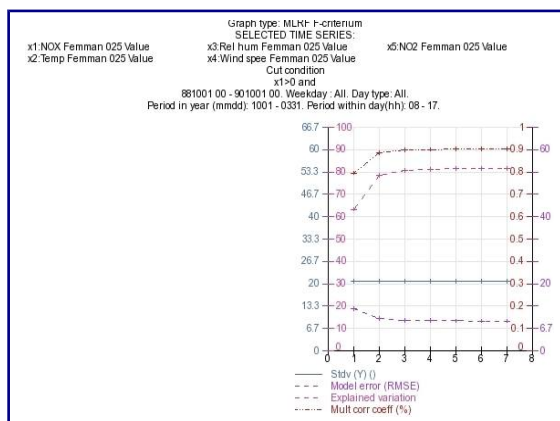


Figure 3.5.1.6 Regression model performance for cross validated models.

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 menu (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 NO₂ concentrations seem to be non-biased and distributed along a straight line with a standard error of 8.7 µg/m³.

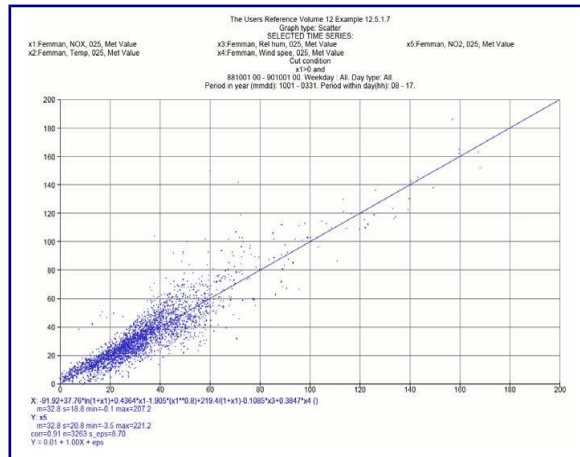


Figure 3.5.1.7 Scatter plot of measured contra estimated NO₂ concentrations.

The next step is to check if the model is capable of describing the cumulative distribution of NO₂ in a proper way. Present the two variables in a frequency distribution graph. The result in *Figure 3.5.1.8* shows that the NO₂ distribution from the model is very close to the curve from observed NO₂.

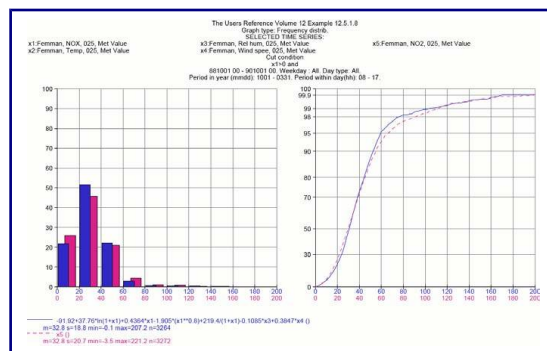


Figure 3.5.1.8 Frequency distribution graph with measured and estimated NO₂ concentrations.

We have thereby shown that a statistical model can be designed to describe the relation of NO₂ to NO_x as a non-biased pure or causal model, having a distribution function similar to the distribution of NO₂. 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 NO₂. 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	$Y = b_0 + b_1 x^3$	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 $\exp(b_0 + b_1 x)$; with the power transform, you can find predictors like $b_0 x^{b_1}$. 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 menu 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:

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 is excluded.

For categorical predictors, which can be ordinal or nominal, you may have to normalize 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:

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

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 on **[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**.

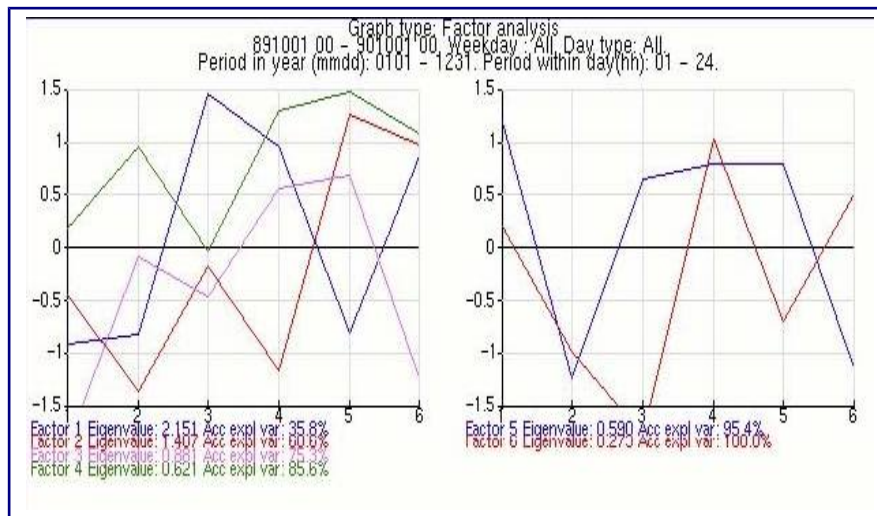


Figure 3.6.1. Factor analysis .

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:

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

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

All the options selected in the different Indico presentation configuration menus can be saved as “macros” to be re-used any time.

Macros are saved in folders,. Each user has his own folder, a common folder and some other folders may also be created. The system administrator decides who is allowed to save macros in the common folder (setting up the corresponding Indico.WriteGroup.user privilege in priv.rf). You can always save macros in your own folders, but usually not in other users' folders, although it is possible to load macros from any 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. You can add or 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 a folder
- Select a macro from the list.
- Select a time period:
 - **Time from macro:** It uses the same period that was set when the macro was saved
 - **Keep current period:** It keeps the period currently set in the menu **Period.**

- **Latest 24 hrs:** Period is set to the last 24hrs from the present period.
 - **Today:** It is the period between 00 and 23hrs (00:00 ≤ x < 00:00).
 - **Yesterday:** Period is set to the 24 hrs of the yesterday day.
 - **Latest 7 days:** Period is set to the last 7days from the present period.
 - **This month:** Period is set to the present month.
 - **Previous month:** Period is set to the previous month to present month.
 - **This year:** Period is set to the present year.
- Press the **[LOAD]** button.

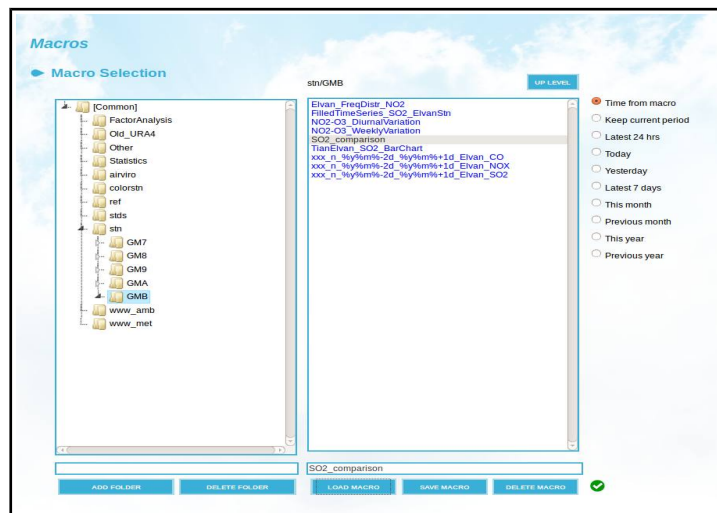


Figure 3.7.1. Macros.

An important macro is the one named **default**. This macro is loaded when Indico Presentation is started. Here you can define line styles and colours, status conditions you want to use, etc. The macro default in the user's own folder will be loaded if it does exist, otherwise the default macro from the folder **[Common]** will be used instead.

When a macro is loaded, the current settings in the different menus are overwritten with those taken from the macro.

Macro names that begin with the string “Auto” are used by Indico Real Time, included in the Indico Presentation module.

3.8 Indico Real Time

The Indico Real Time Graph automatically displays a selection of predefined macros sequentially, one after the other for a number of seconds before the next graph is displayed. Graphs are updated as new data arrive.

Once you are satisfied with the configuration of your graph, save it as an Indico macro. The macro should begin with the string “Auto” to be recognized by Indico Real Time. All macros beginning with the string “Auto” are displayed in alphabetical order. You can also control the order in which the macros are displayed by including a number in the macro name, immediately following the string “Auto”.

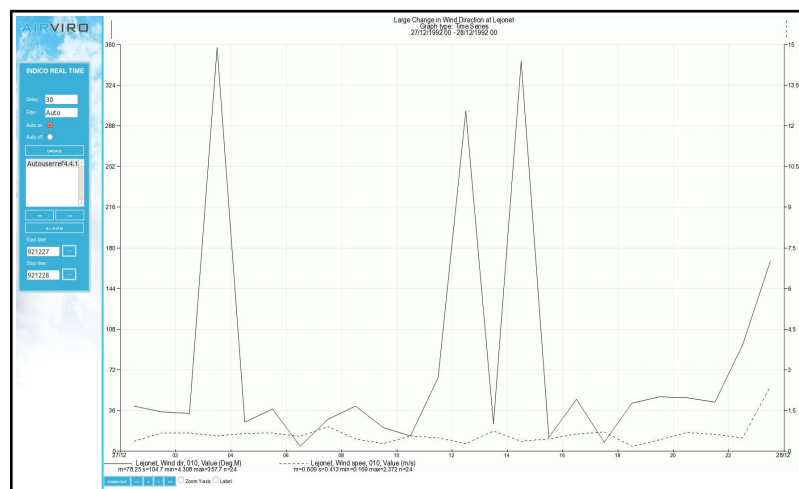


Figure 3.8.1. Real Time.

Different users can create different Auto macros. However, all Auto macros that have been saved 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 is logged in to Airviro

In the Indico Real Time menu, enter the number of seconds in the input box **Delay** for each macro graph to be displayed. Default is 30 seconds. See *Figure 3.8.1*. Use the input box **Filter** to specify the beginning string of the macro names that you want to appear in the list

Once 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 graph will automatically resize each time the browser window is expanded or shrunk. The size of the window can also be specified in the **GRAPH SETTINGS** menu.

Use the forward [**>>**] and backward [**<<**] buttons to manually force the next or previous graph to be displayed. The option buttons **Auto off** and **Auto on** will let you stop/start displaying the next graph automatically.

The macro name of the macro currently displayed is highlighted in the list. Clicking on any name in the macro list will display the associated graph. Press the [**UPDATE**] button to update your settings and force the macro to refresh the graph

The time period displayed can be modified entering a different date in the **Start time** and **Stop time** text boxes.

Zooming in can be made using the mouse to select a rectangle over the graph. To zoom out press the button [**ZOOM OUT**]. Use the buttons [**<<**] [**<**] and [**>**] [**>>**] to shift the period displayed by a whole period or half a period backwards and forwards.

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 $x_3[-1]$ refers to the time series selected as x_3 , shifted by -1 time unit. As an example, if x_3 is plotted together with $x_3[-12]$ using the hourly database, then the values for $x_3[-12]$ will be the same as those for x_3 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: $\ln(x)$, $\log(x)$, $\exp(x)$, $\text{int}(x)$, $\text{abs}(x)$, $\text{sqrt}(x)$

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

Inverse trigonometric functions:

$\arcsin(x)$, $\arccos(x)$, $\arctan(x)$, $\text{arccot}(x)$

Hyperbolic functions:

$\sinh(x)$, $\cosh(x)$, $\tanh(x)$, $\coth(x)$

Fill functions:

$\text{interpol}(x,n)$ 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.

$\text{sustain}(x,n)$ 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.

$\text{interps}(x,n)$ is the same as $\text{interpol}()$, but also makes a

constant extrapolation if only one of the surrounding

values around the current time is within n time steps.

Miscellaneous functions: $\text{reep}(x,a)$ (=0 if $x < a$ else =1)

$\text{aver}(x_1, x_2, \dots)$ Mean value of x_1, x_2, \dots

$\text{aver}(x_1:x_5)$ Mean value of $x_1 - x_5$)

$\text{min}(x_1, x_2, \dots)$ Minimum value of x_1, x_2, \dots

$\text{max}(x_1, x_2, \dots)$ Maximum value of x_1, x_2, \dots

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

$\text{min}(x_1+4, 0, \ln(x_2 - x_1))$ the minimum value of several functions

$\text{max}(x_1:x_3[1], x_1:x_3, x_1:x_3[-1])$ the maximum value of x_1, x_2, x_3 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:

$$API(x) = I_j + \frac{I_{j+1} - I_j}{C_{j+1} - C_j} \times (x - C_j) \text{ for } C_j \leq x \leq C_{j+1} - C_j$$

$$j + 1$$

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:

$$\text{api}(x, c_1, i_1, c_2, i_2, \dots, c_n, i_n)$$

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:

$$\text{desc}(Y, l_1, l_2, \dots, l_n)$$

gives one of the values $1, 2, 3, \dots, n$ if $Y > l_1, l_2, \dots, \text{ or } l_n$. 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

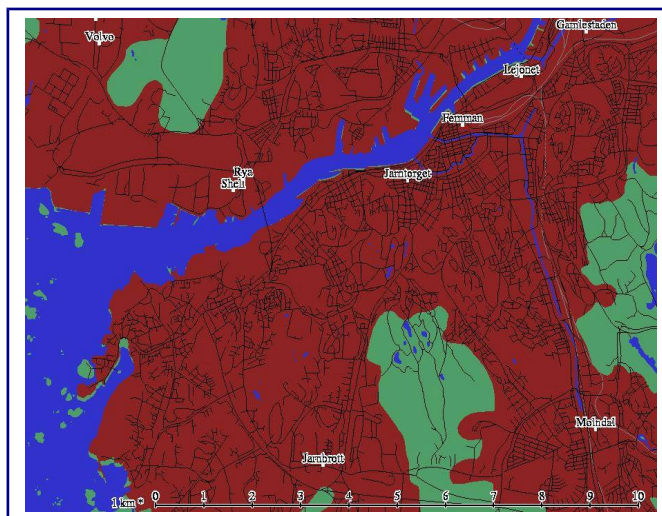
With the above definition of API a so called subindex can be created for each substance, where the break points can be defined in different ways depending on the chosen substance.

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

$$\text{Total API} = \max(\text{API1}, \text{API2}, \dots)$$

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

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®, is a registered trademark by Microsoft Corporation,

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 box will appear the first time Waved is used in a working session in order to allow you 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 options are available:

- Import to Excel from Airviro: Displays the import dialog box.
- Export from Excel to Airviro: Shows the export dialog box.
- 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 use. The existing time resolutions for the selected database are shown in the time resolution list. Select a time resolution and press [OK]. See *Figure 3C.1. Database and Time Resolution*.

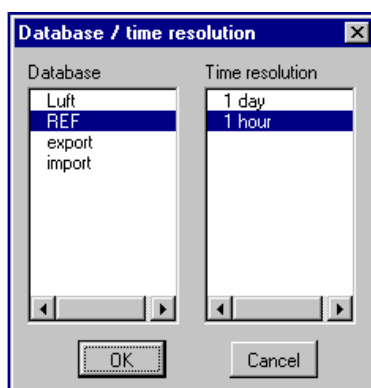


Figure 3C.1. Database and Time Resolution.

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

3.C.6. Import to Excel from Airviro

In **Import to Excel**, the Airviro dialog box allows you to select the time series to be

imported.

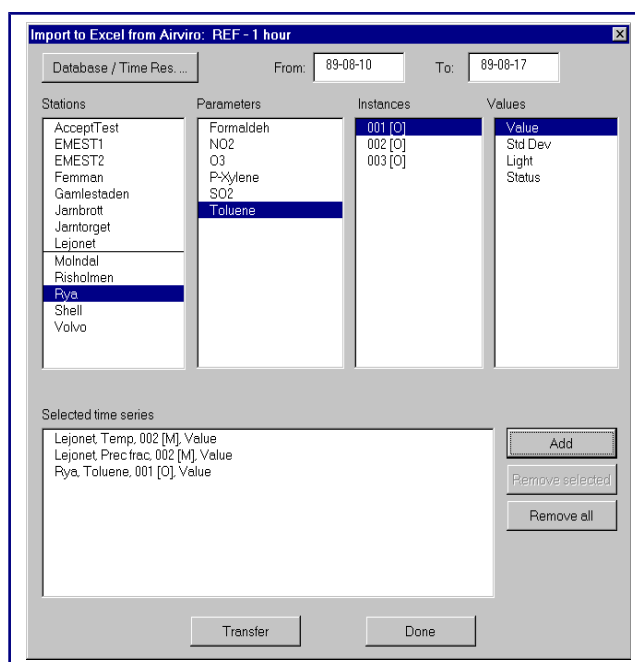


Figure 3C.2. Import from TSDB.

Import to Excel in the Airviro dialog box of Waved is organised in three sections: The top, the middle and the bottom. (Figure 3C.2. Import from TSDB.)

In the top section you select the database, the time resolution and the time period of data that will be transferred from Airviro to Excel. The from/to dates format is the local Windows format. It can be changed from the control panel.

Time series are specified in the middle section. A time series is determined by a station, a parameter, an instance and a value. All of them must be selected in order to specify a time series. Only stations that have data in the time series database are listed. When a station is selected the parameters measured for that station are listed. When a parameter is selected the instances available for that parameter and station are shown. Finally when an instance is selected the values in the time series database for the selected instance, parameter and station are displayed.

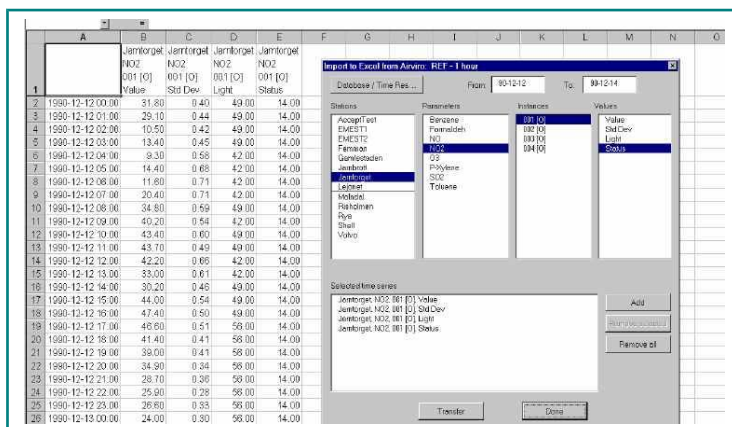
The bottom section contains the selected time series. To add a specified time series to the selected ones press **[Add]**. To remove it from the selection press **[Remove selected]**, to remove all the selected time series, press **[Remove all]**.

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

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

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

The following example is taken from Gothenburg domain. Time series data from the DOAS station Järntorget, which is located in the middle of Gothenburg, will be imported to Excel from Airviro. The data imported will be Value, Standard deviation, Light and Status for NO2.



1. Determine the active cell in Excel from where the data will 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 box will appear 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 the date 90-12-12 in the **From** textbox and 90-12-14 in the **To** textbox . NOTE: The date format will depend 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 for Std Dev, Light and Status.
9. Press [**Transfer**] to import the data to Excel .

3.C.6.2.Limitations

The maximum number of rows in an Excel spreadsheet is limited to 65535. This number corresponds to approximately 7 years of hourly data.

3.C.7. Export from Excel to Airviro

In the **Export from EXCEL** to Airviro dialog box the stations, parameters, instances and values from existing time series are displayed. It works just like **Import to Excel** from Airviro except that by clicking on **New station**, **New Parameter** and **New Instance** it is possible to select stations, parameters and instances for which no time series exist in the database. NOTE: It is not possible to create new stations, parameters or instances.

Example: The parameters NOX and NO2 are defined in the parameter database and stations Femman and Järntorget in the station database. Two time series are stored: 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 any time series for station Femman that contains NO2. However, by clicking on **New station**, a dialog box containing all the parameters in the parameter database will appear, ie NOX and NO2. By selecting NO2 and clicking on **[OK]** the parameter NO2 is added to the parameter list in the **Export from EXCEL to Airviro** dialog box. As no time series exist for station Femman and parameter NO2, therefore no instances will be listed. It is necessary to press **[New Instance]** in order to add one.

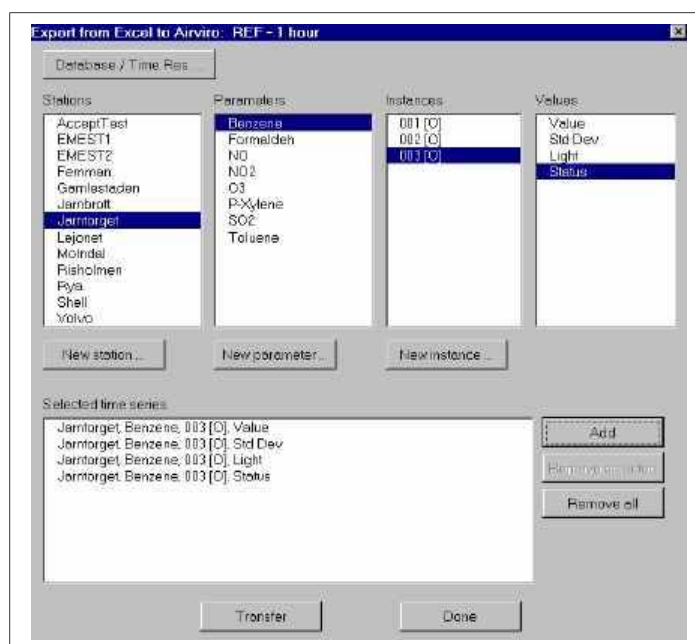


Figure 3C.3. Export from TSDB.

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

In the top section you select the Airviro database to export to and the time resolution.

Time series are specified in the middle section. A time series is determined by a station, a parameter, an instance and a value. All of them must be selected in order to specify a time series. Only stations that have data in the time series database are listed. Clicking on **[New station] button** a new dialog box will pop up where you can select any existing station.

When a station is selected the parameters measured for that station are listed. In the parameters list either an existing parameter can be selected or otherwise you can click on **[New Parameter]** in order to select any parameter in the parameter database. When a parameter is selected, the instances in the time series database for that parameter and station are listed. An instance can be selected either from the instance list or from the **New Instance list** 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 listed.

[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 section contains the selected time series. To add a specified time series to the selected ones press **[Add]**. To remove it from the selection press **[Remove selected]**, to remove all the selected time series, press **[Remove all]**.

When **Transfer** is pressed, the selected time series will be copied 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 stations with existing time series are shown in the station list of the **Export from Excel to Airviro** dialog box. To export data to a station that does not have any data in the time series database click on **[New station]**. A dialog box listing all the stations in the station database of Airviro will be displayed. Select a station and click on **[OK]**. The selected station is immediately added at the bottom of the station list in the **Export from Excel to Airviro** dialog box. See *Figure 3C.4.New station*.

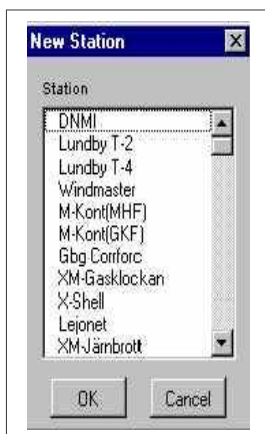


Figure 3C.4. New station.

3.C.7.2. New parameter

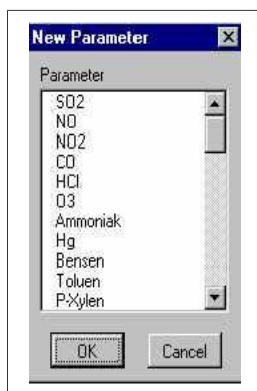


Figure 3C.5. New parameter.

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

A dialog box listing all the parameters in the parameter database of Airviro is displayed. Select a parameter and click on [**OK**]. The selected parameter is immediately added at the bottom of the parameter list in the **Export from Excel to Airviro** dialog box.

3.C.7.3. New instance

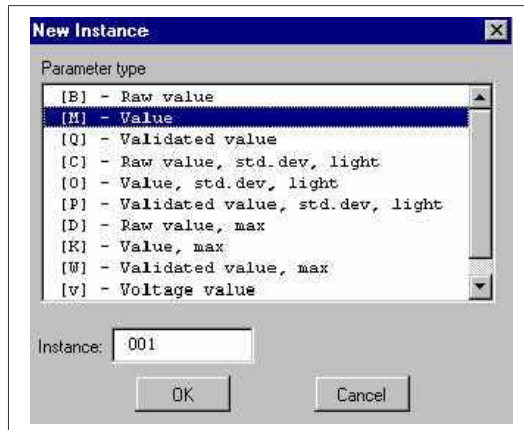


Figure 3C.6.New instance.

For the selected station and parameter only instances of existing time series are listed in the **Export from Excel to Airviro** dialog box. 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 listing all the possible instances in Airviro is displayed. Select an instance and click on **[OK]**. The selected instance is immediately added at the bottom of the instance list in the **Export from Excel to Airviro** dialog box. See *Figure 3C.6.New instance*.

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

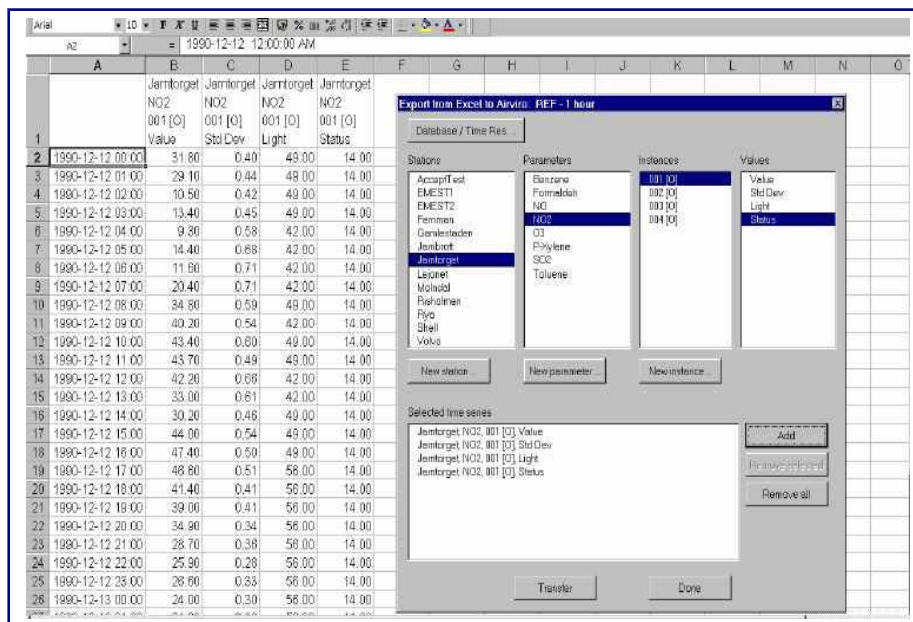
Parameter type	Three step database only	OPSIS only	NILU only	Calibration database only	Number of parameters (excluding status)
[M] - Value					1
[B] - Raw value	X				1
[Q] - Validated value	X				1
[C] - Raw Value, std. dev, light	X	X			3
[O] - Value, std. dev, light		X			3
[P] - Validated value, std. 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. In Excel ,set the active cell for 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, you must first enter your password and optionally a user name.
4. The **Export from Excel to Airviro** dialog box is displayed.
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*.

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*, usually located in the folder */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 into which data was imported, was not in the scan list of avdbm. Check it out 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 time series selected in one of the dialog boxes (export, import) is mirrored in the other dialog box when it opens.

To use Waves as a database editor these steps must be followed:

1. Open the **Import to Excel** from Airviro dialog box.
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 box 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** Airviro dialog box.
7. Press [**Transfer**]. The data is exported to Airviro.

3.C.9. Technical specification

Waved includes the following features:

- . Waved only require to be installed in the client PC as a plug-in module to Excel. It is not required to install any other software in the Airviro server
- The Excel low level C interface used allows an extremely fast transfer of data . A whole year of hourly data for four stations can be exported to Excel in just a few

seconds.

- Allows to export up to 32 simultaneous time series.
- Allows to import up to 32 simultaneous time series.
- Allows the creation of new time series. A time series can be created for any existing station and parameter. New instances can be created as well.