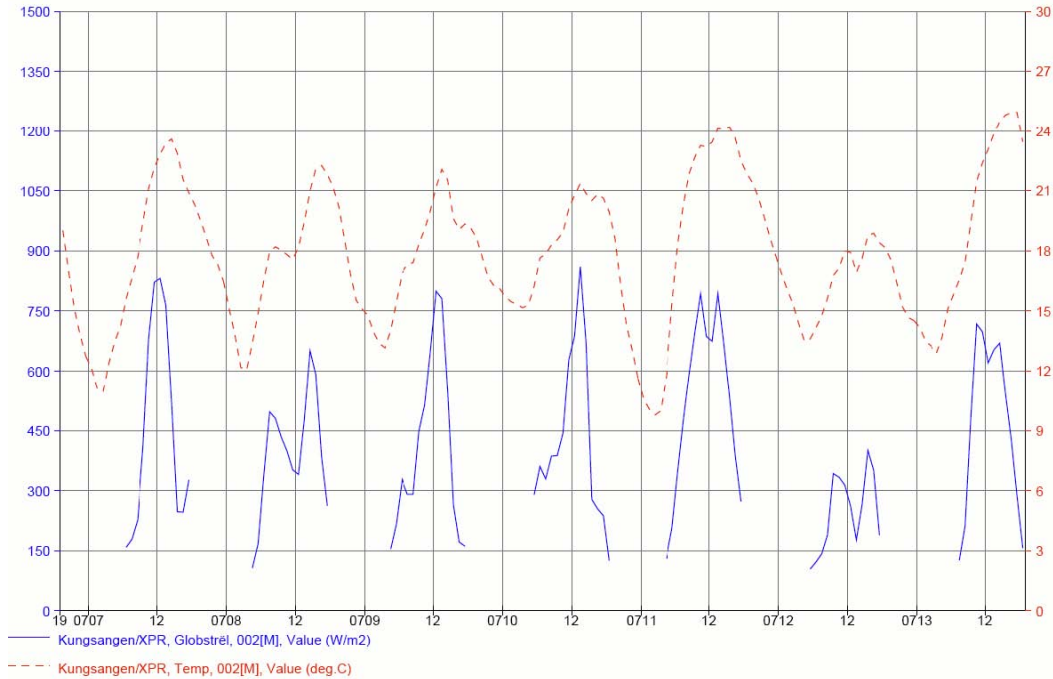


Airviro User's Reference



**Working with Indico Presentation Client
Time Series Analysis and Presentation**

Working with Indico Presentation Client

Time Series Analysis and Presentation

Amendments

Version	Date changed	Cause of change	Signature

Contents

12.1 Introduction to Indico Presentation	7
12.1.1 What is Indico Presentation?	7
12.1.2 How does Indico Presentation client work?	8
12.2 Getting started	9
12.2.1 Overview of the Indico Presentation main window	9
12.2.2 Time series database	10
12.2.3 Selecting time series	10
12.2.4 Selecting cases	12
12.2.5 Constraining cases	13
12.3 Math expression compiler	15
12.3.1 Transforming variables	16
12.3.1.1 Handling missing values.....	17
12.3.1.2 Counting	18
12.3.2 Creating new variables	19
12.3.3 Modifying a series by smoothing or differencing	19
12.4 Presenting graphs	21
12.4.1 Controlling the layout of a graph	22
12.4.2 Displaying the graph	23
12.4.3 Available graph types	25
1 The time series graph	25
2 The frequency distribution graph	25
3 The scatter graph	27
4 The Breuer diagram	28

5 The mean/sector diagram	29
6 The frequency/sector diagram.....	30
7 The diurnal variation diagram.....	30
8 The weekly variation diagram.....	31
9 The annual variation diagram.....	32
12.5 Regression modelling	33
12.5.1 Linear regression model	34
12.5.1.1 Fitting a curve.....	40
12.5.2 Binary logistic regression model	41
12.6 Factor analysis	43
12.6.1 Principal component analysis	45
12.7 Using Indico macros	47
12.8 Real Time Graph	49

12.1 Introduction to Indico Presentation

Indico Presentation is - together with Indico Admin, Indico Editor, Indico Report and Indico Validation - part of the integrated IIndico package. The extra “I” in IIndico stands for Internet, which indicates that IIndico is a web-distributed tool in the IAirviro system.

The IIndico package is one of the three main tools in IAirviro Air Quality Management, intended for acquiring, storing, editing, presenting, analysing, reporting and exporting time series data.

12.1.1 What is Indico Presentation?

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

- · Select one or more time series – measured, simulated or forecasted – for processing.
- Assess capture and status of the data.
- Constrain observations from further processing.
- Handle missing values by interpolation.
- Transform variables by computing, counting or recoding into categories.
- Find or eliminate trend and seasonal components by smoothing or differencing.
- Monitor diurnal, weekly and yearly variation.
- Plot time series data in a line chart, histogram or frequency distribution.
- Plot pairs of variables in scatter plots or polar diagrams.
- Fit a curve to pairs of variables.
- Set up a linear or binary logistic regression model to estimate concentrations or chemical reactions.

- · Apply factor analysis or principal component analysis to structure data and avoid collinearity. · Automate the production by using macros.
- · Automatically update diagrams as new data arrive.

An experienced Indico user will be able to use the IAirviro system as an *integrated monitoring system*, i.e. extract valuable information from the measured data and add this information to the emission database or simulation models of IAirviro.

12.1.2 How does Indico Presentation client work?

Indico Presentation is a JAVA-based user interface to the time series database and shell programs at the Airviro server. Indico Presentation client can be executed from a PC or other device running Internet Explorer 6 or Netscape Navigator 7 (or higher).

After logging in, an IAirviro user can process data from a time series database at the server, according to settings in Indico Presentation. All processing of data is done at the Airviro server. Results are transferred over a TCP/IP connection for presentation in Indico.

Output is generated to an SVGA-capable display as an HTML-picture - including some interactive elements, as a pdf-file or as plain ASCII text. Printing is generated by the web browser or with third-party software, e.g. Adobe® Reader 6.

A searchable help system is provided in the form of a pdf-file (this document).

Please be aware that the Indico Presentation client requires no installation at the PC side. Anybody who knows the IAirviro web URL may reach the database. It is therefore important to change passwords regularly and to set privilege levels for the users with great concern to limit unauthorised access. Using an Intranet solution behind a firewall instead of Internet prevents access for the world-wide web, if properly set up.

12.2 Getting started

When IAirviro has been properly installed at the Airviro server, it is possible to jump to your **IAirviro** URL with a web browser over Intranet/Internet. The web interface controls all web modules, including **Indico Presentation**, Indico Admin, Indico Report, Indico Editor and Indico Validation among others.

After logging in with user-ID and password, the user is presented with a list of available domains and web modules. When they have been selected, the title bar changes its

caption to reflect your choices. See figure 12.2.1.1.

12.2.1 Overview of the Indico Presentation main window

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

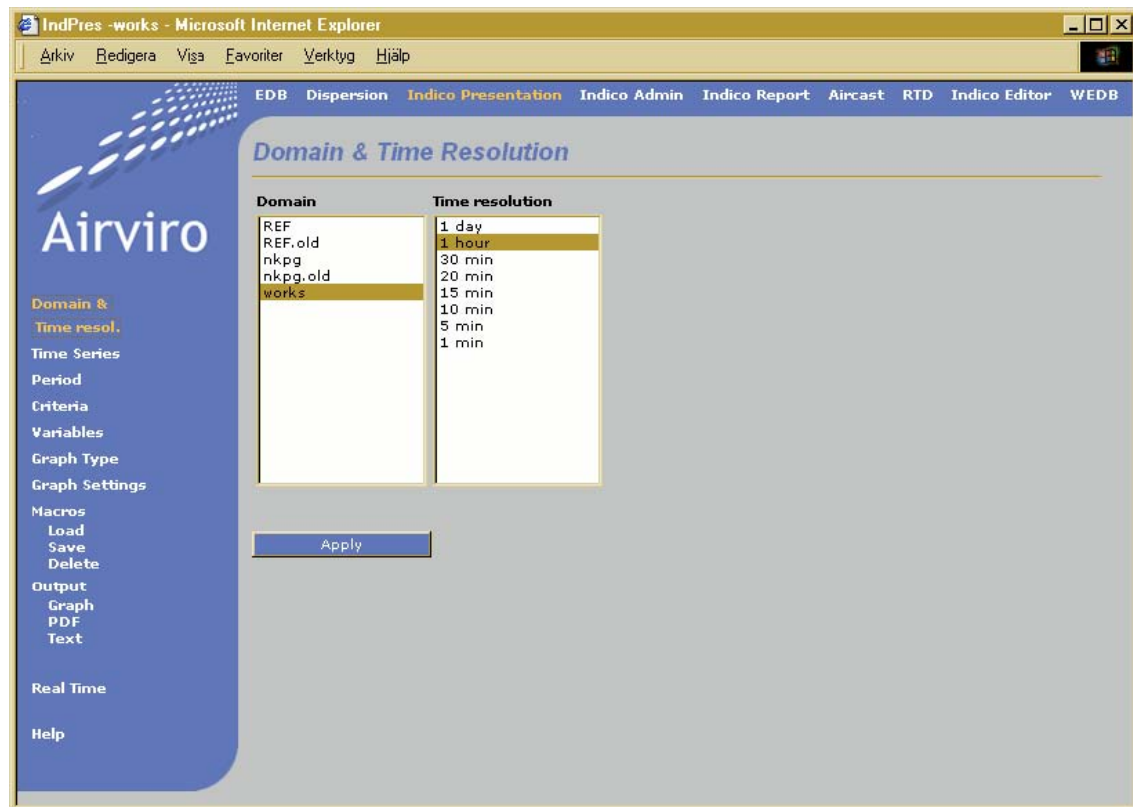


Figure 12.2.1.1 Indico Presentation window with available domains and time resolutions.

12.2.2 Time series database

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

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

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

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

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

12.2.3 Selecting time series

In the **Domain & Time resolution** frame, the user selects database project and one available resolution. It is not necessary to select the same domain as in the login process. If you change to another domain, you will get other time series and other macros related to the current domain. See figure 12.2.1.1 above.

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

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

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

When you have selected both station and parameter, you will get a list of available instances of the actual parameter. The instance is used to differentiate between simultaneous measurements of the same parameter at the same site, e.g. if you measure

at different levels above ground or if you are using more instruments or analytical functions to get an output.

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

The status flag is assigned in the quality control in Indico Admin. The status can be changed if you work with Indico Editor. Please be aware that manual changes in the time series database can be rolled back, if necessary. For a table of status conditions, see under **Criteria >> Status conditions**.

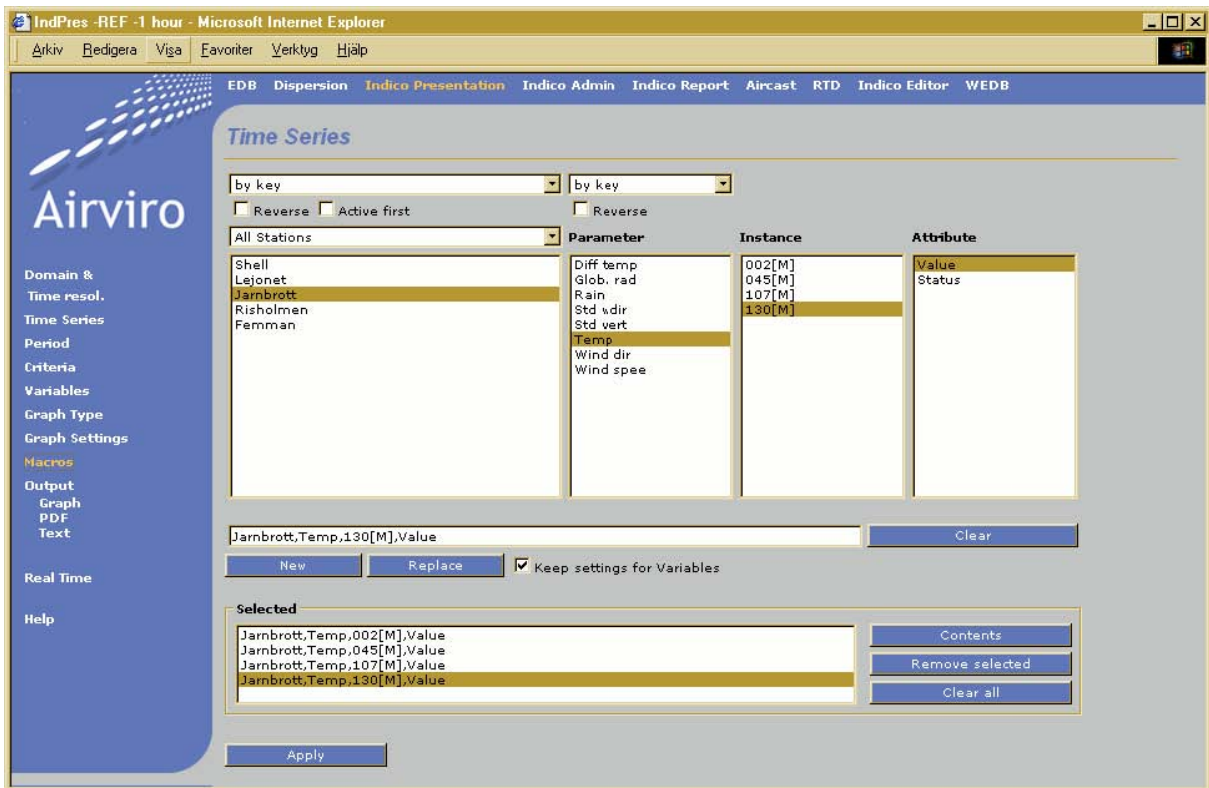


Figure 12.2.3.1 Time Series frame with available stations and parameters in the REF database with 1 hour resolution. Up to 15 time series can be selected for further processing.

When you have clicked on station, parameter, instance and attribute, the time series is uniquely identified (for the current time resolution). Click **New** to select the time series for further processing. You can select up to 15 time series for further processing. If you click **Contents**, a graph will appear to present data capture for selected variables during some period. If you let the haircross rest on the graph, you can read date and value in a label.

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

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

12.2.4 Selecting cases

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

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



Figure 12.2.4.1 Indico Presentation Period frame with start date and end date in European format in the REF database with 1 hour resolution. There is no guarantee that data exist in the selected time period.

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

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

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

$$t \hat{I} [t, \dots, t$$

From To

12.2.5 Constraining cases

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

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

It is also possible to constrain by hours. The hours are accessed by their name, all-inclusive. If you want observations from 22:00:00 until 22:59:59, specify that you want the **period within day** from 23 to 23. If you want all observations except that hour, specify that you want the period from 24 to 22. See figure 12.2.5.1 below.

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

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

You can apply some constraint related to the observed data, e.g to study only cases with low wind speed, low temperature or some other condition. The constraint should be written into the text box "**Condition formula**". The variables are named x1 to x15 according to how the time series were entered in the "Selected" list box, see figure 12.2.3.1. You can compare your variables with constants or expressions in a Boolean formula. Statements can be combined with logical operators OR(), AND(&) or NOT(!). A quick reference of available functions can be found in the **Variables** frame, following the **Help** link. See further chapter *12.3 Math expression compiler*.

Please make sure that you set criteria in the correct unit. If you are unsure of which unit that is used, compare with the parameter database, which is available from Indico Admin. Alternatively, you can modify the expression by using arithmetic functions in your condition formula. See figure 12.2.5.1 below for an example of how to write a condition formula. If the Boolean expression is false, all variables in that case will be left without a value.

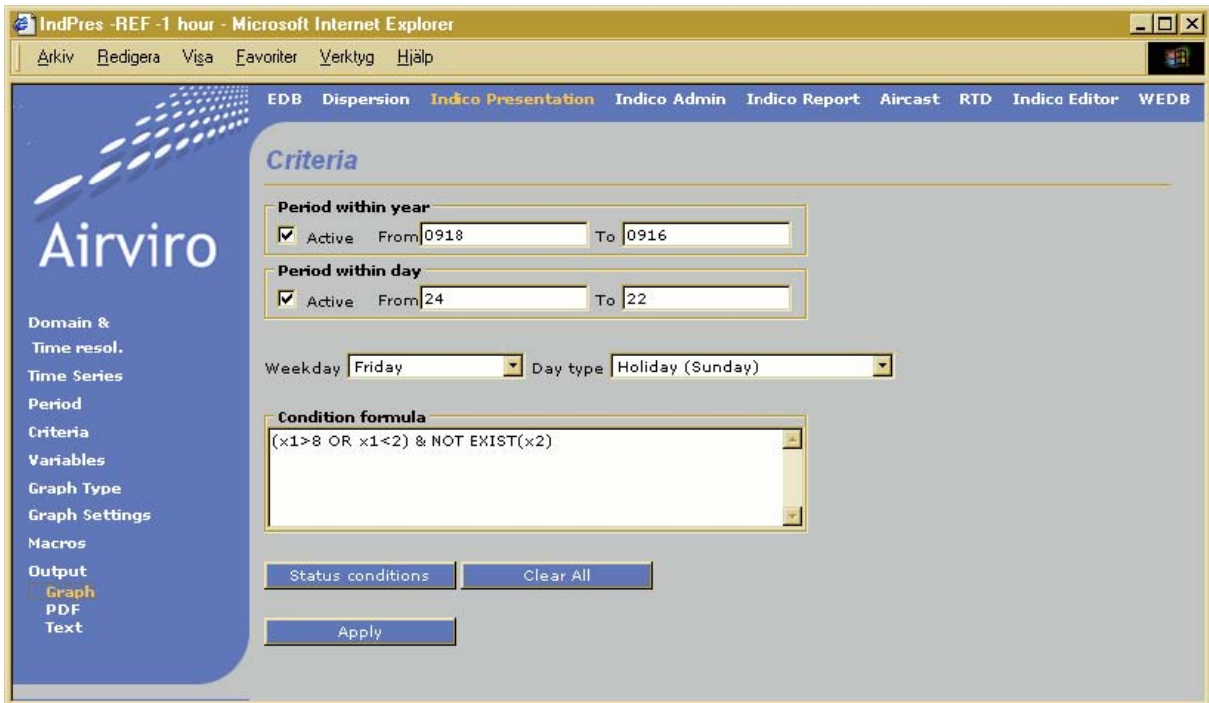


Figure 12.2.5.1 Indico Presentation Criteria frame with some criteria for constraining access to the selected cases.

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

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

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

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

12.3 Math expression compiler

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

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

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

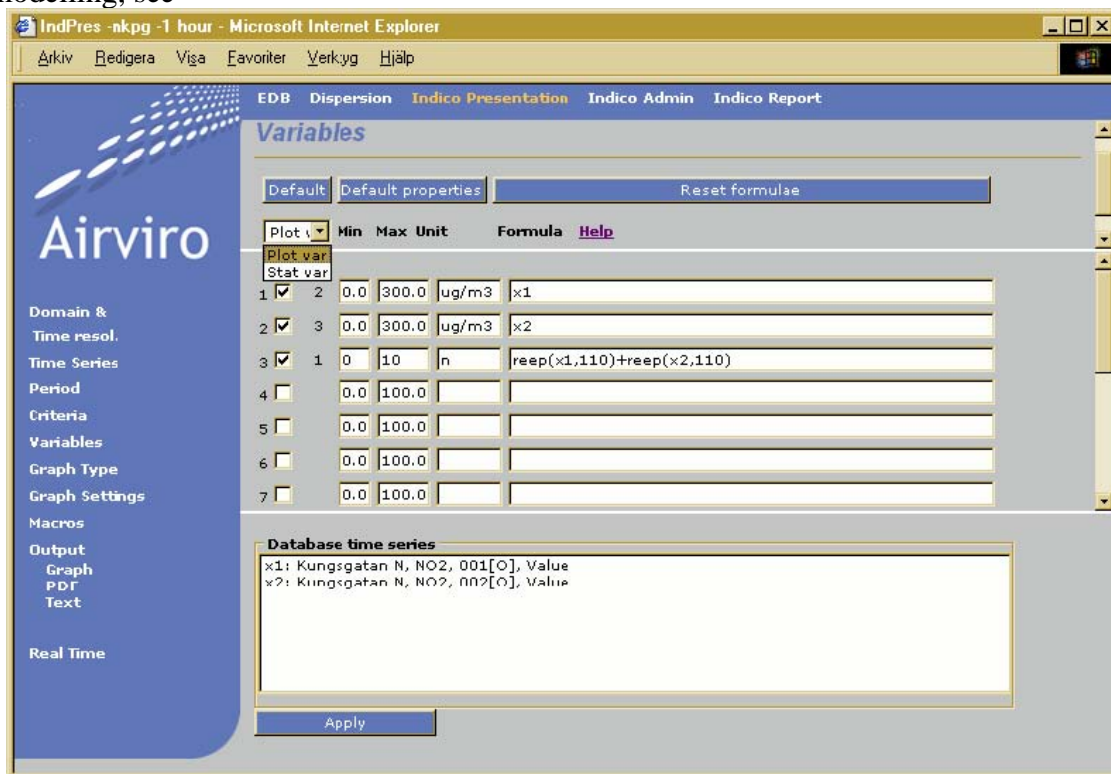


Figure 12.3.1 Indico Presentation Variables frame with three plot variables for two series.

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

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

In the **Variables** frame, you see a list of all time series that were selected; compare with the bottom text box in figure 12.3.1. 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_{15} with their associated units of measurement and default min-max values from the parameter database. You can change the min-max values, the associated unit of measurement or the formula as you wish. This will have an effect on plotting of values. The **min-max** values refer to the scale in a plot; **unit** refers to the text on the ordinate axis. You can change the order in which variables are plotted, by unchecking them all and then check a new plot order with up to four variables in one graph. A new variable can be defined (variable 3, plot variable 1 in the figure) by entering a function in the formula field next to the variable.

You can enter arithmetic or conditional Boolean expressions in the formula field. Examples of an arithmetic expressions are $x_1 * 1000$ 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 E4 Calculation Formulae in Airviro Specification, part II*. You can also look at the quick reference help following the **Help** link.

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

12.3.1 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 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

$$\text{var} = \frac{x_2 + 13.273}{13.273 \cdot (97.28 M 2929.1 *)}$$

if x_1 is mass concentration in $\mu\text{g}/\text{m}^3$ and x_2 is temperature in $^{\circ}\text{C}$.

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 to number the groups by some instructive value - the mean value or some code, e.g. Beaufort scale. You can use the conditional operator **?:** to recode the variable.

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

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

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

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

In a simplified way, the above division into groups can be accomplished with the descriptor function. The **desc** function returns values {1,2,3,...,n} if $x < \{I_1, I_2, I_3, \dots, I_n\}$.

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

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

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

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

This is recoded as:

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

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

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

12.3.1.1 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.

12.3.1.2 Counting

If you want to count exceedances for an observation, you can use the **reep** function. Simply fetch all your monitored channels into the variables x1..x15 and compare the observations with some threshold value.

$$\text{var} = \text{reep}(x1,110)+\text{reep}(x2,40)+\text{reep}(x3,0.5)\dots,$$

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.

$$\text{var} = (x1==4?1:0)+(x2==4?1:0)+(x3==4?1:0)\dots,$$

where x1, x2, x3 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.

Environment variables can be set by the Airviro system administrator.

12.3.2 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 12.7 for information about using macros. You have to define the new time series in Indico Admin 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) at your PC or a script at the server to import the new time series into the time series database. Ask your Airviro system administrator for help.

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

12.3.3 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:

$$\text{Trend} = \frac{(0.5 \cdot x_1[-12] + x_1[-11] + x_1[-10] + \dots + x_1[-1] + x_1[1] + \dots + x_1[11] + 0.5 \cdot x_1[12])}{24}$$

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:

$$\text{Trend} = \sum_{i=-10}^{10} X a_i, \text{ where}$$

$$[a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}] = \frac{1}{350} [6, 18, 33, 47, 57, 60, -2, -5, -5, -3, -1]$$

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 normalised deviation. Finally, the trend is reestimated 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 $\tilde{N}x_1 = \{x_1 - x_1[-1]\} = (1 - B)x_1$, 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 $\tilde{N}_d x_1 = \{x_1 - x_1[-d]\} = (1 - B^d)x_1$ to eliminate the seasonal and the trend term by repeated differencing.

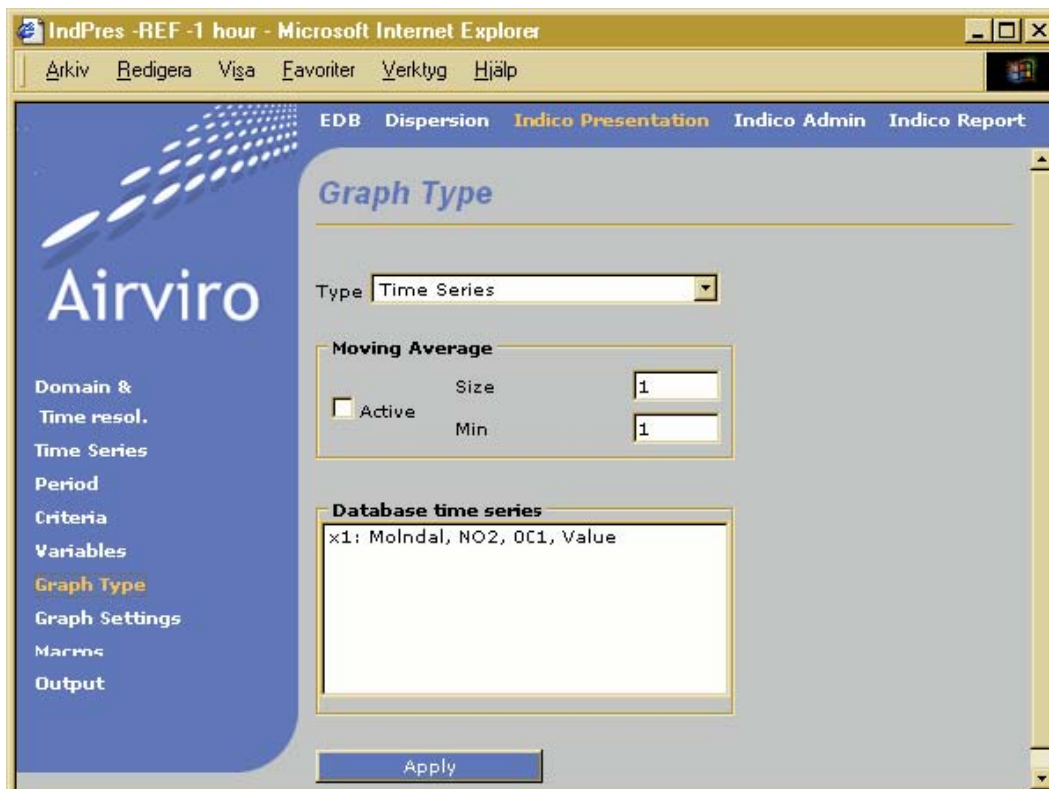
12.4 Presenting graphs

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

The presentation types are:

1. 1. Time series graph
2. 2. Frequency distribution graph
3. 3. Scatterplot
4. 4. Polar plots: Breuer, Mean/sector and Freq/sector
5. 5. Seasonal variation charts: Diurnal, Weekly and Annual The different charts will be explained in section 12.4.3 according the list above.

For the statistical analysis types, see chapter 12.5 *Regression modelling* or chapter 12.6 *Factor analysis*.



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

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

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

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

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

12.4.1 Controlling the layout of a graph

In the **Graph Settings** frame, you can specify a heading in the **Title** field.

You can control the layout of the presentation by checking chart elements from the **View** subframe. 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 legend that explains the plot variables, check the **Footers** box. For descriptive statistics in the footnote, check the **Statistics** box.

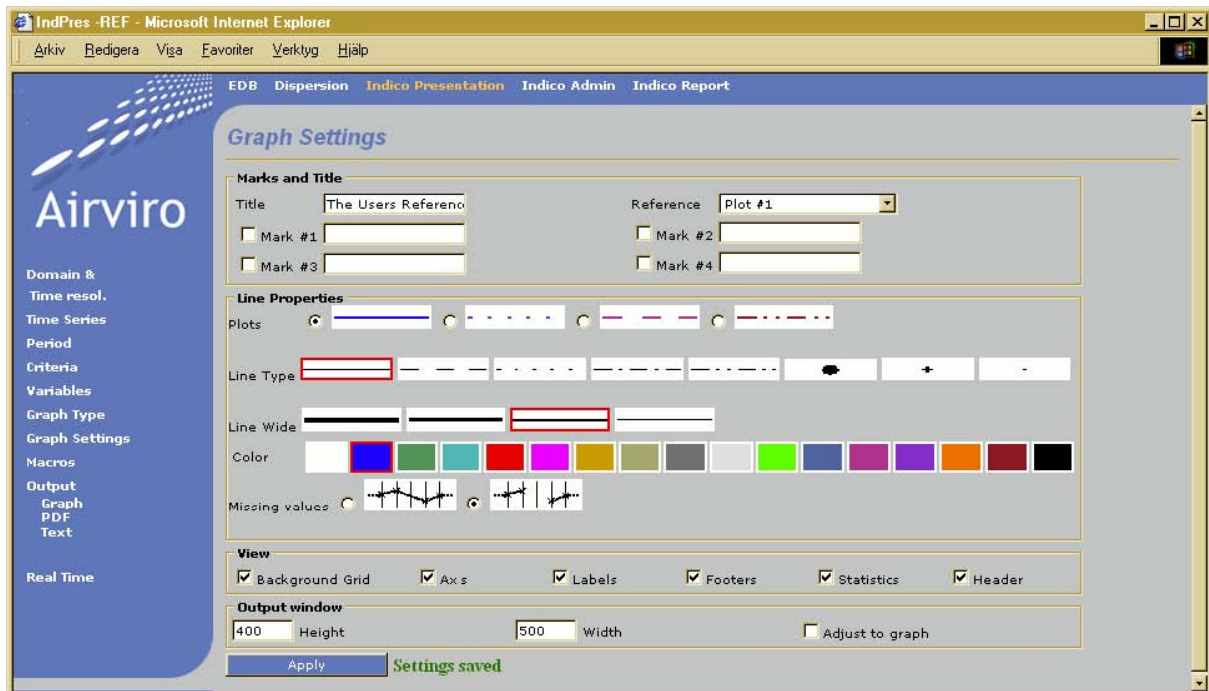


Figure 12.4.1.1 Graph Settings frame with title, layout and line properties.

If you want an ordinate axis and abscissa, check the **Axis** box. If you check the **Label** box, you will get a fair number of tick marks on the axes. The spacing is calculated according to your settings for min and max in the Variables frame. If you check **Background grid**, you will get grid lines in your chart.

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

You can change the line style or marker style for each plot variable in a chart. It is allowed to present up to four plot variables in a chart. In the **Line properties** subframe, you can set line type, line width and line color.

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

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

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

12.4.2 Displaying the graph

The graph can be written to an interactive HTML-window by selecting **Output – Graph** in the left-hand menu. The size of the window – in pixels - can be set in the **Output window** subframe. There you can also adjust the size of the output window to the graph. The HTML-graph is opened in a new window. See figure 12.4.2.1 below.

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

If you let the haircross rest on the graph, you can read date and value on a label. Alternatively, you can read date and value in a text area that appears when you click **Show point in graph**.

In the HTML-window, you can save the graph as an HTML-document. You can also print the graph to any connected printer, using ordinary web browser functions.

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

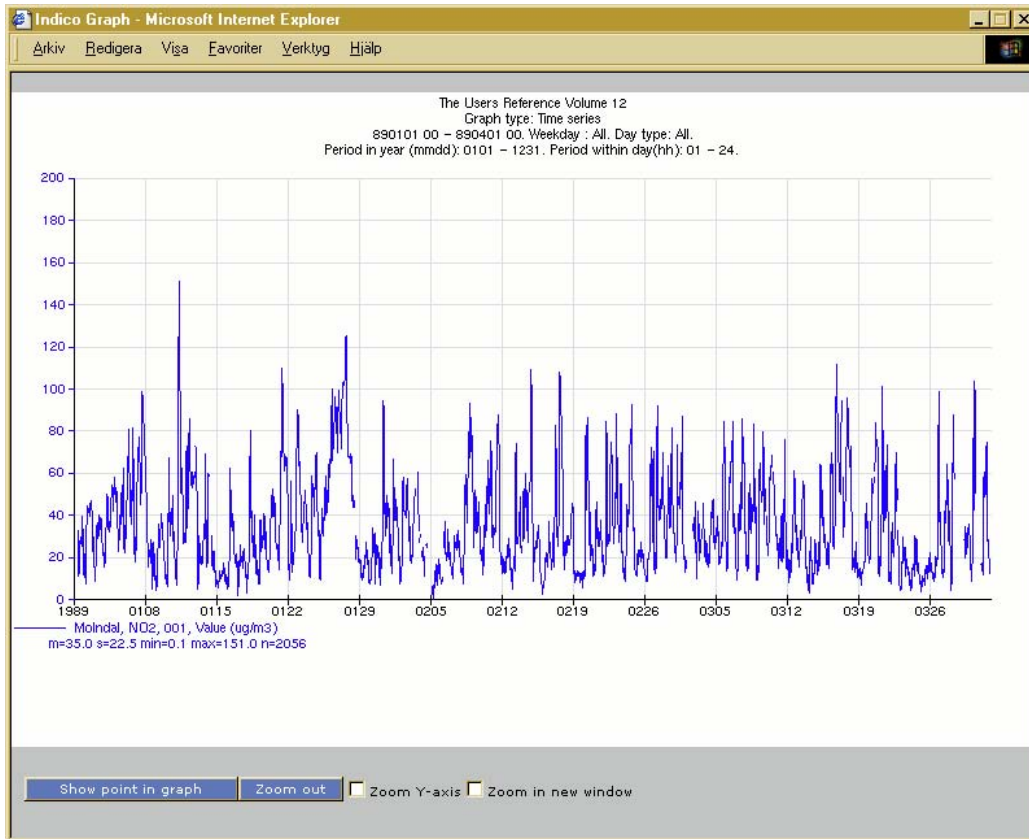


Figure 12.4.2.1 Interactive Output graph in HTML-format.

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

Text font and special characters can be changed if you use Adobe Acrobat. With Acrobat, you can also export the graph in jpeg, tiff or png format. See figure 12.4.3.1 for an example of output in PDF-format.

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

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

12.4.3 Available graph types

1 The time series graph

is a multiple line chart with a date variable on the abscissa and up to four variables on

the ordinate axis. Both axes are linear and continuous. An example of a time series graph is seen in figure 12.4.3.1. A legend may appear below the graph with descriptive statistics about mean value, standard deviation, span and number of valid cases.

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

Applying a moving average to one or more variables will filter away high-frequency components and leave a smoothed line showing short-time trend. If you have much data in your view, as in figure 12.4.2.1, your understanding of the variation will benefit from applying a smoothing filter.

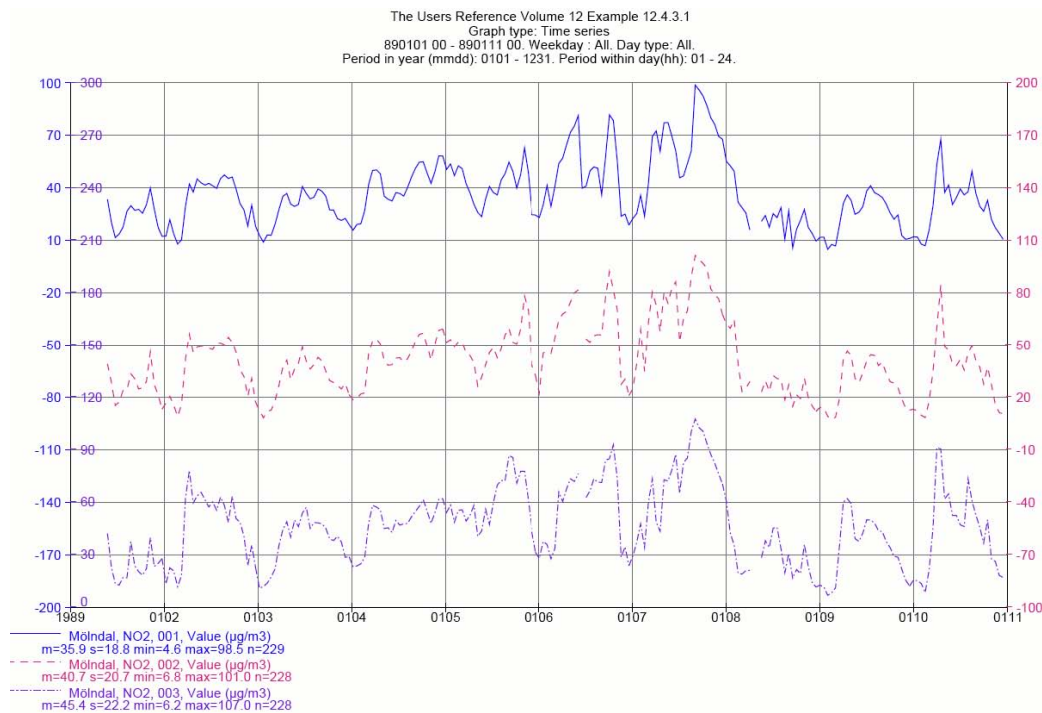


Figure 12.4.3.1 Time series graph in PDF-format with three channels at different offset.

2 The frequency distribution graph

is divided into two parts - a percentage histogram and a cumulative distribution chart. In the histogram, the ordinate is always linear from 0 to 100%. The examined variable is divided into 10 discrete classes according to the min-max settings in the Variables frame. If you want to group your data into other classes, it is always possible to recode your data as in section 12 3.1.

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

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

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

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

An example of a frequency distribution graph is seen in figure 12.4.3.2.

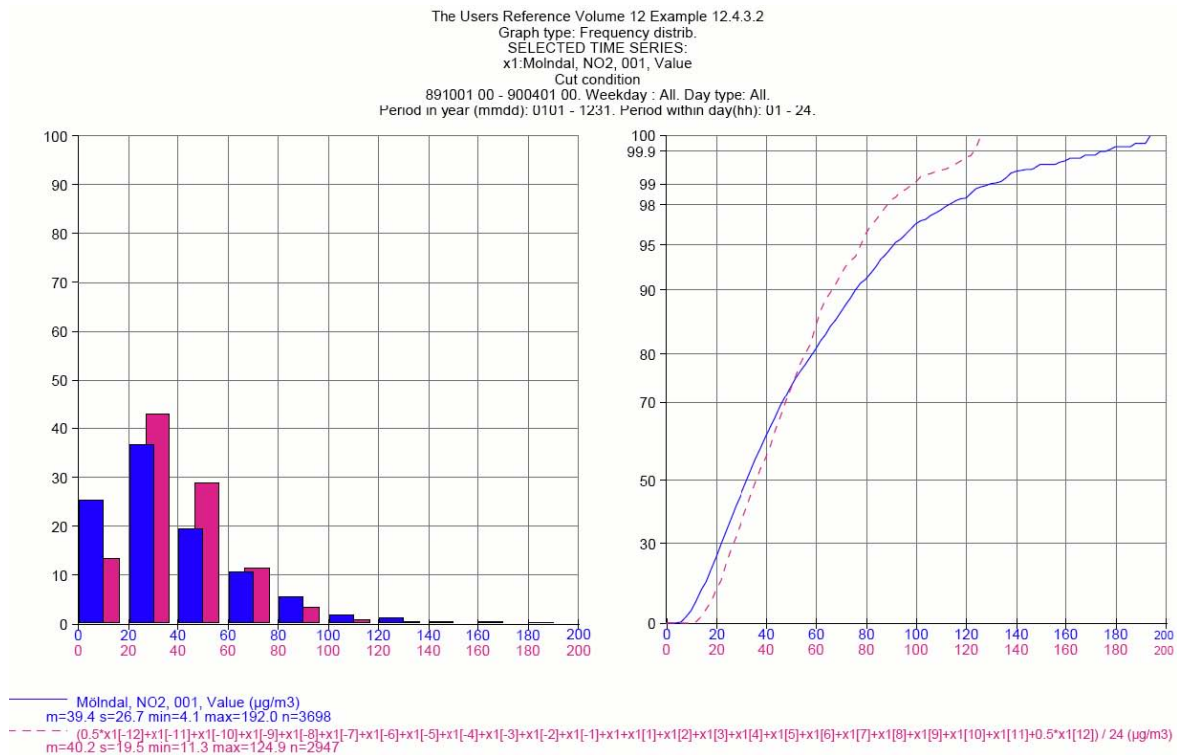


Figure 12.4.3.2 Frequency distribution graph with percentage histogram and cumulative distribution chart. The effect of applying a 24-h moving average can be seen by comparing plot variable 1 and 2.

3 The scatter graph

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

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

If you check the **Regression** box in the **Scatter** subframe, the regression line will be included in the scatterplot.

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 12.4.3.3 for an example of a scatterplot. The color and size of the markers can be defined in Graph settings.

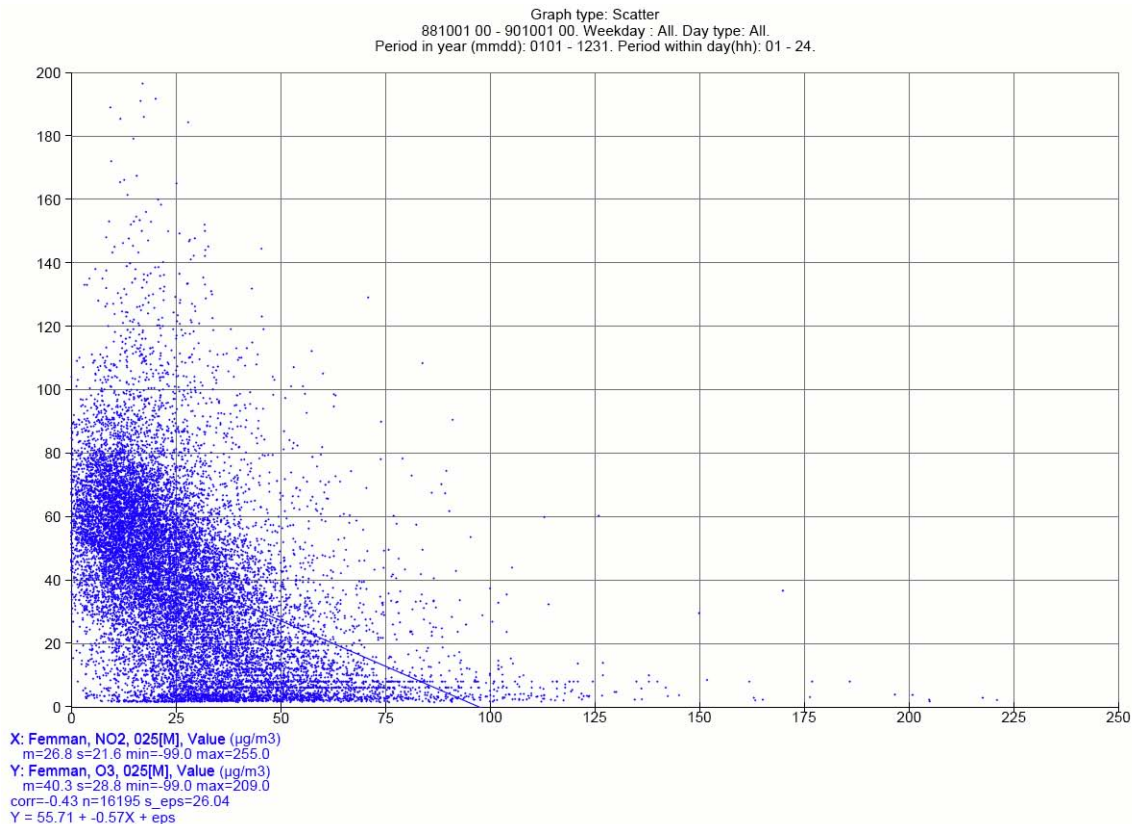


Figure 12.4.3.3 Scatterplot with ozone contra NO_2 concentrations. A regression line is included in the plot.

4 The Breuer diagram

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

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

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

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

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

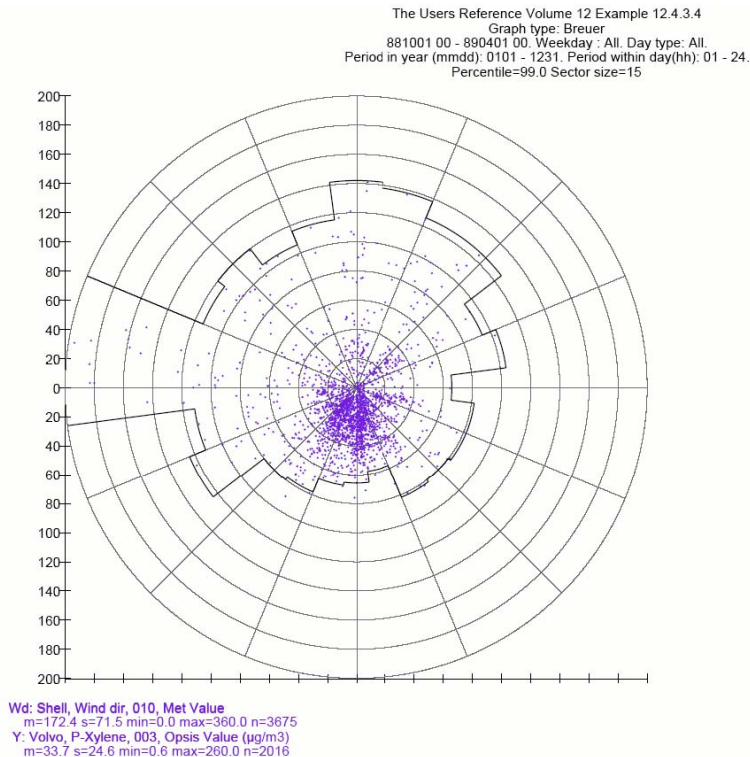


Figure 12.4.3.4 Breuer diagram with P-Xylene concentrations contra wind direction in a polar coordinate system. The 99%-ile concentration in each sector is indicated by an arc.

5 The mean/sector diagram

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 nonparametric 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:

$$C(q, Dq) = \frac{\sum_{i=1}^n K((q - W_i) / Dq)}{\sum_{i=1}^n 1}$$

where q is the examined wind direction, W_i is the actual wind direction, Dq is the width and K is the shape of the sliding window, which could be a Gaussian kernel like:

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

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.

The Users Reference Volume 12 Example 12.4.3.5
 Graph type: Mean/sector
 881001 00 - 890401 00, Weekday : All, Day type: All.
 Period in year (mmd): 0101 - 1231, Period within day(hh): 01 - 24.
 Sector size=15

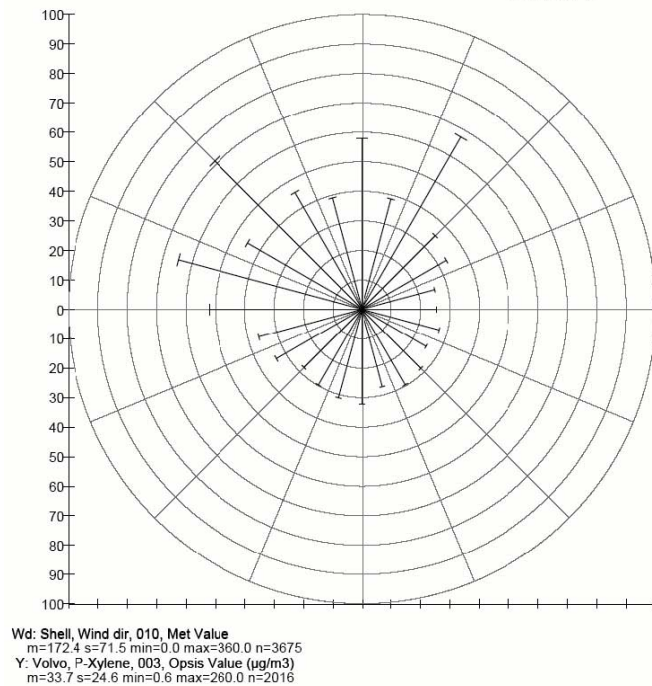


Figure 12.4.3.5 Mean/sector diagram with P-Xylene concentration contra wind direction. In the same way as in the Breuer diagram, you can calculate the pollution flux instead of the concentration to get a more distinct presentation.

6 The frequency/sector diagram

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

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

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

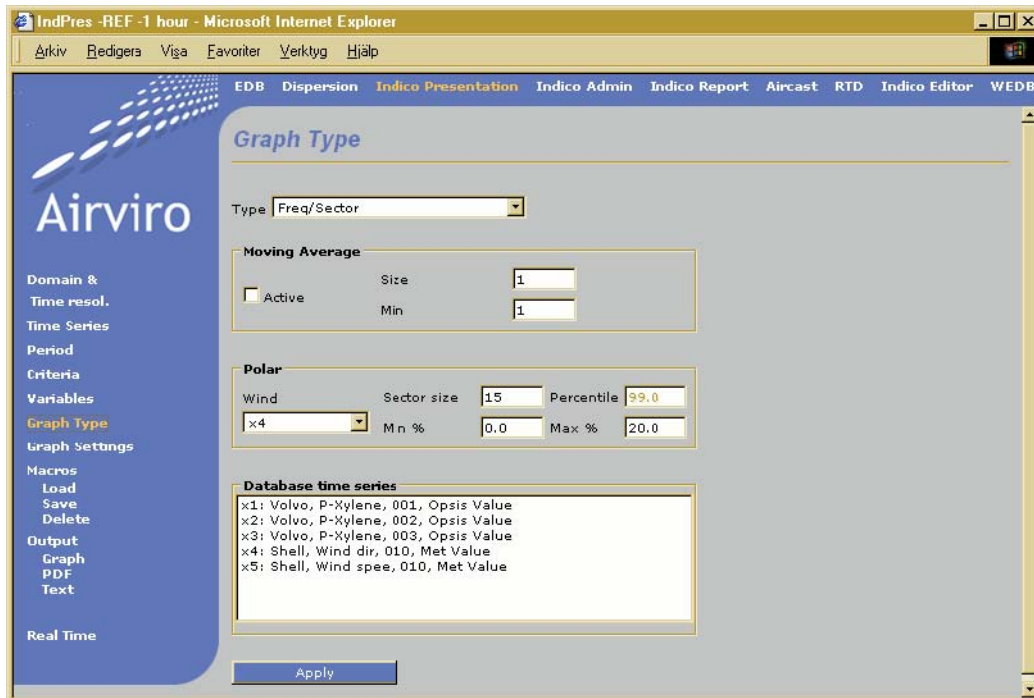


Figure 12.4.3.6 Graph Type frame with settings for a wind rose in the Polar subframe.

7 The diurnal variation diagram

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

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

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

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

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

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

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

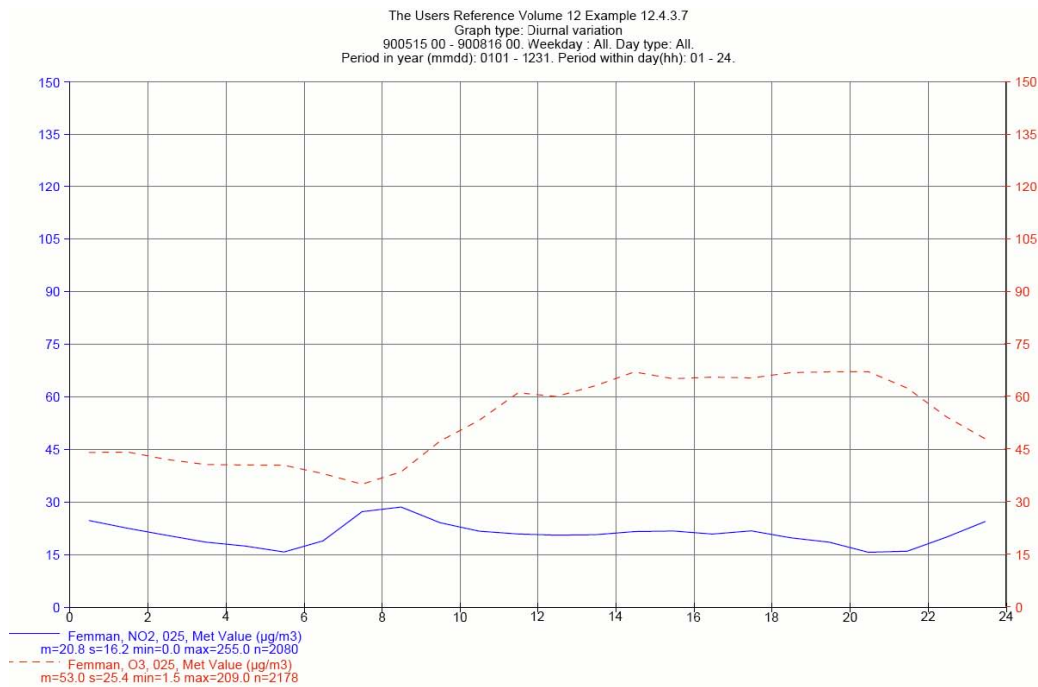


Figure 12.4.3.7 Diurnal variation diagram with anticorrelated NO₂ and ozone concentrations.

8 The weekly variation diagram

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.

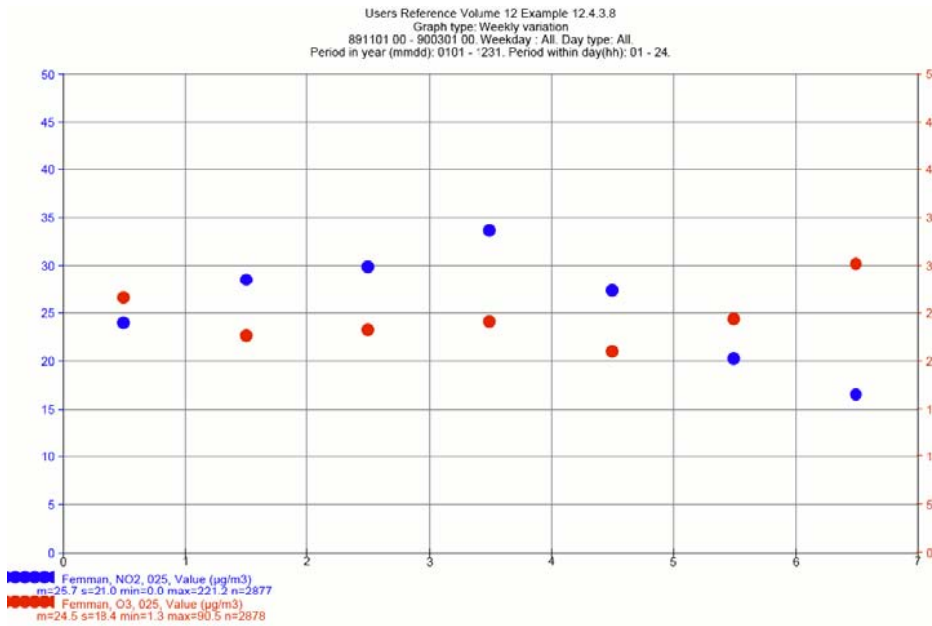


Figure 12.4.3.8 Weekly variation diagram with anticorrelated NO₂ and ozone concentrations. During weekends, the NO₂ concentration is lower than during weekdays.

9 The annual variation diagram

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



Figure 12.4.3.9 Annual variation diagram with NO₂ and ozone concentrations. Ozone has a very strong seasonal variation, since more ozone is formed in strong sunshine.

12.5 Regression modelling

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

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

The first three methods will be described in this chapter, while the others are described in chapter 12.6 *Factor analysis*.

All statistical analyses use statistical variables, as defined in the **Variables** frame. Up to 15 statistical variables can be defined, regardless of the number of time series that have been selected.

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

The regression model can in general be described as:

$$Y = b_0 + X_1 b_1 + X_2 b_2 + \dots + X_n b_n + e$$

12 nn

where Y is called the predictand, representing NO₂ or any variable of interest. The X's, i.e. X₁, X₂, ..., 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 b₀, b₁, ..., 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 minimised.

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 ϵ)
- 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).

12.5.1 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 transformation 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.

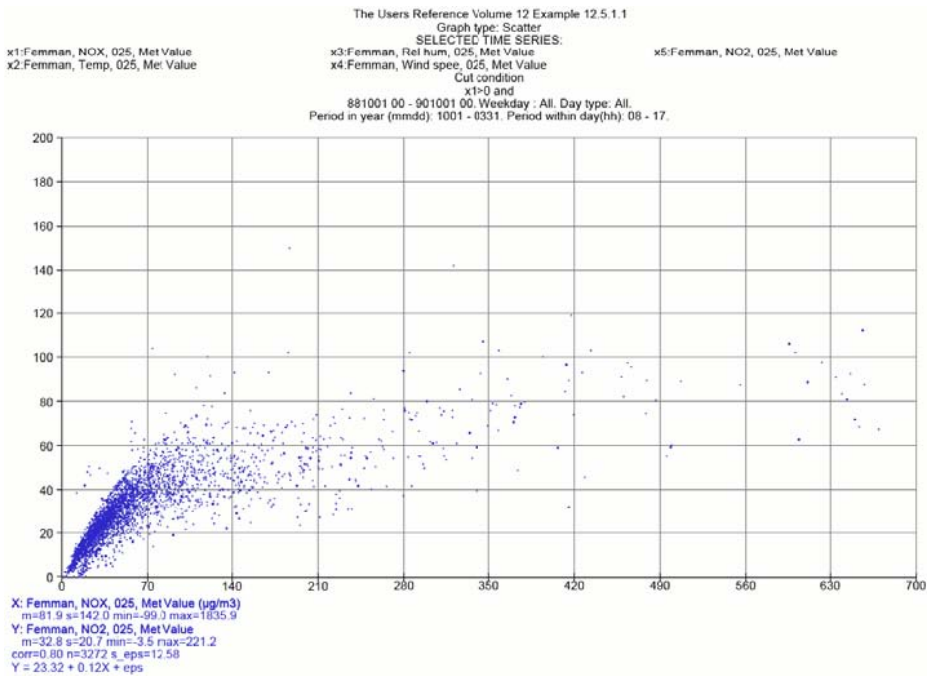


Figure 12.5.1.1 Scatterplot of NO₂ contra NO_x concentrations.

In figure 12.5.1.1 you can see a scatterplot of NO₂ as a function of NO_x. For small values of NO_x (<100 µg/m³) the ratio NO₂/NO_x is 50%-70%, but for large NO_x concentrations (>500 µg/m³) the ratio is approximately 15%.

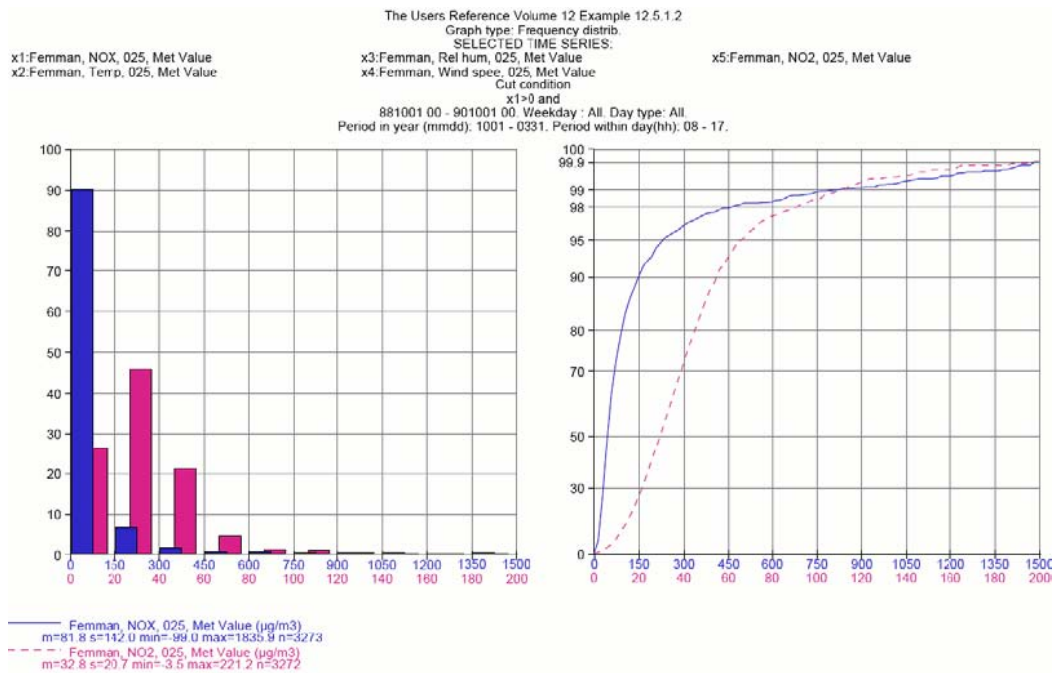


Figure 12.5.1.2 Frequency distribution graph with NO₂ and NO_x concentrations.

From the frequency distribution graph in figure 12.5.1.2 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 scatterplot that the relation seems to be logarithmic or inversely proportional so we arrange the predictors as:

$$\ln(1+\text{NO}_x), 1/(1+\text{NO}_x), \text{NO}_x^{0.8}, \text{NO}_x$$

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

Settings for statistical variables are done in the **Variables** frame, see figure 12.5.1.3 below.

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

F-out is the criterion that a variable already included in an earlier step should be rejected in a later step. You should require a high probability that the variable is insignificant before rejecting it, since it has already been selected in the scheme as having better fit than subsequent predictors. A recommended value for F-out is 0.10 but even higher values like

0.25 can be used, which means that you can retain highly insignificant variables. It is possible to enter predictors into the scheme without stepwise selection.

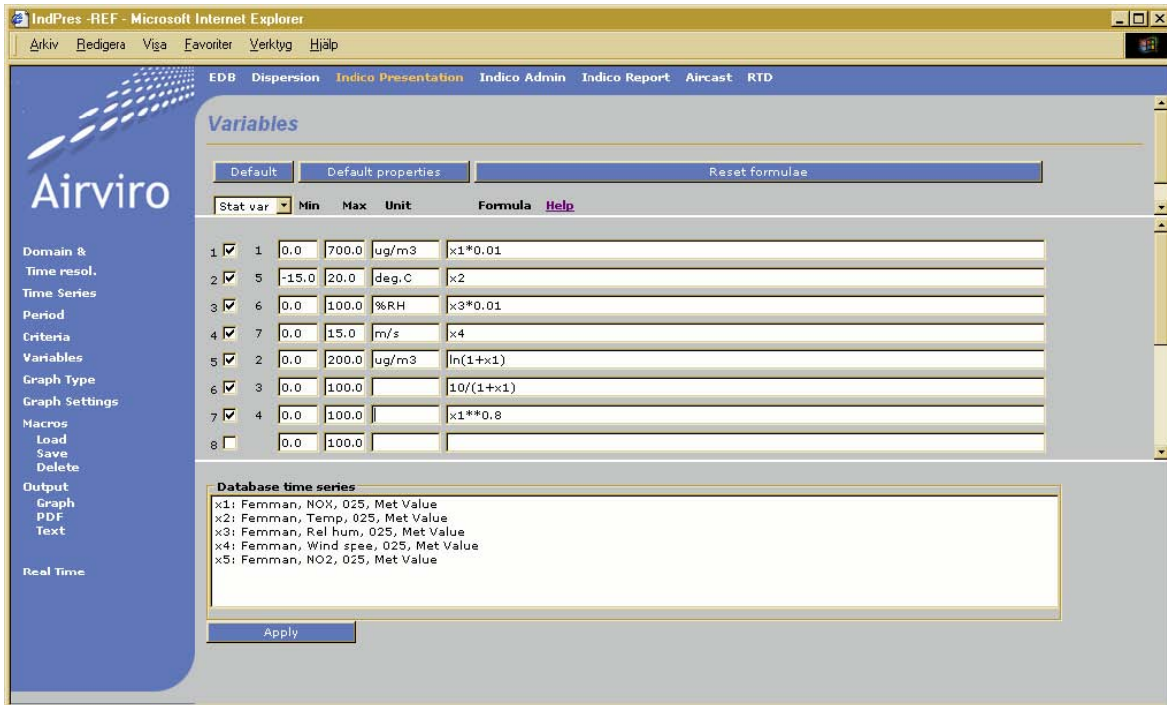


Figure 12.5.1.3 Defining statistical variables in the Variables frame.

With the settings in figure 12.5.1.3, 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 12.5.1.4.

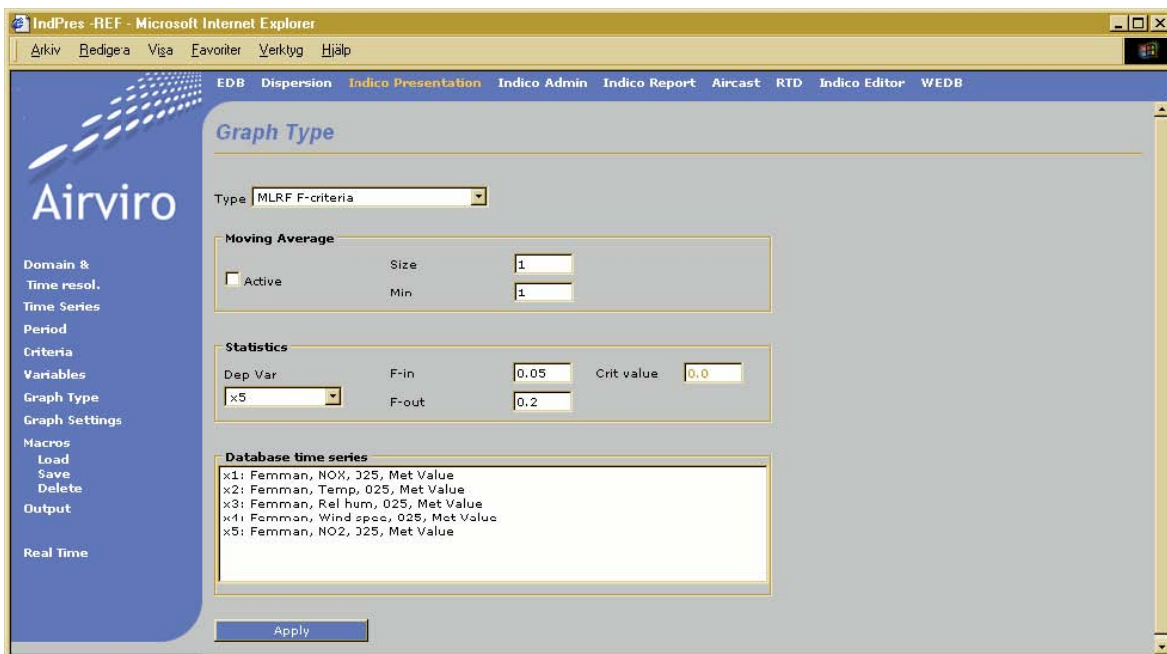


Figure 12.5.1.4 Defining dependent variable and F-criteria in the Graph type frame.

With these settings you can run the stepwise regression model **MLRF F-criterion**. The result is presented as in figure 12.5.1.5 below.

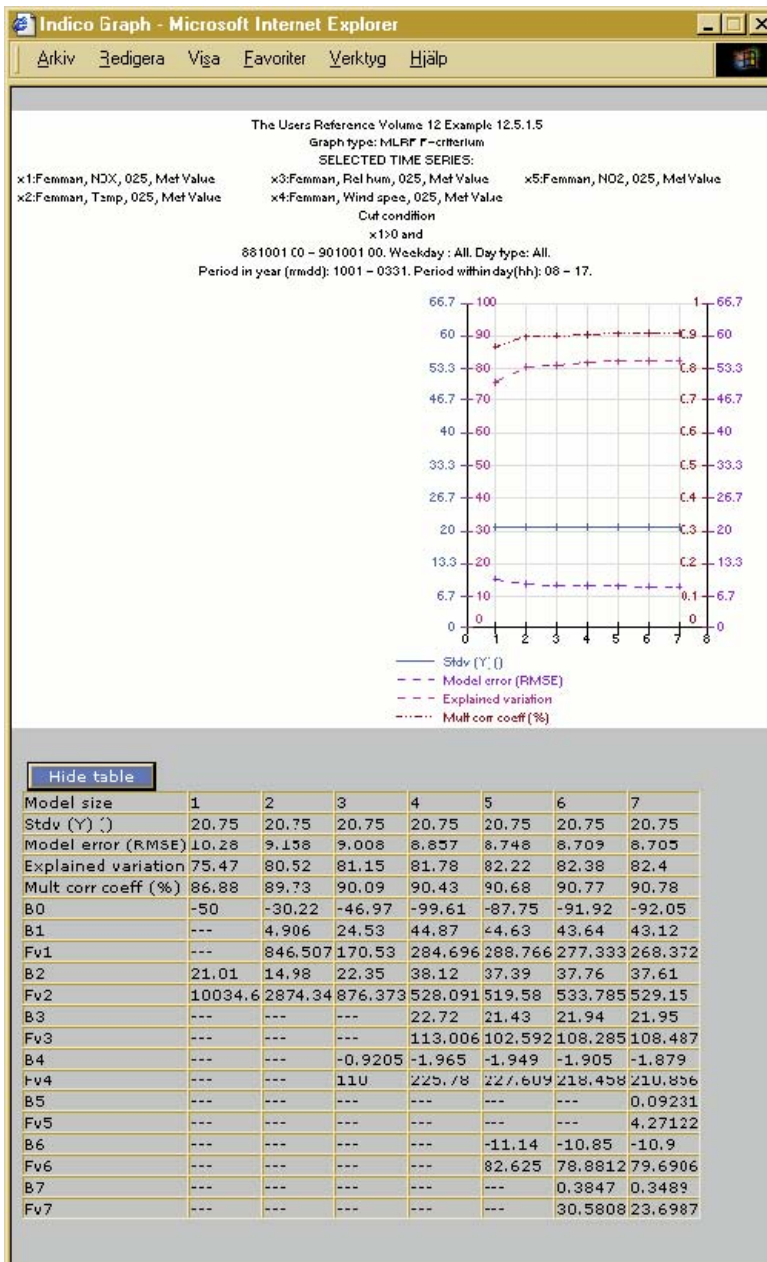


Figure 12.5.1.5 Regression model performance for stepwise increasing model size.

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^2 , 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 cross validation** 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 overfitting of data.

Choose **MLRF Crossvalid** in the Graph type frame and **Apply**. The result, as a graph will be similar to figure 12.5.1.6. 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 overfitting of the data and it would be unwise to use model size 7 in this case.

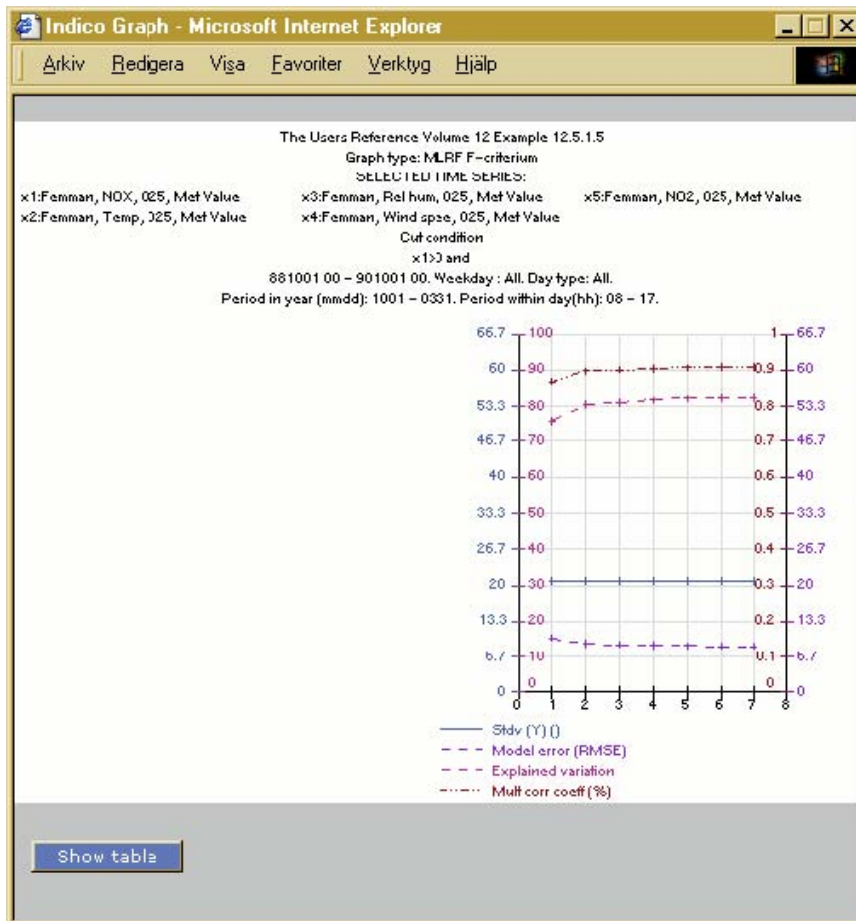


Figure 12.5.1.6 Regression model performance for crossvalidated models.

Looking at figure 12.5.1.5 and 12.5.1.6 we decide to use the parameters based on model

size 4 (the improvement in performance from model size 4 to 5 is not pronounced). The coefficients for the predictors can be obtained from the table in the Output graph. Click on **Show table** and read the coefficients and correct order of the predictors, comparing with the order of statistical variables in the Variables frame (listed to the right of active check boxes).

If you want to examine the correlation between the dependent variable and the model estimate, you can enter the linear regression model as a formula into a plot variable and produce a scatterplot with a regression line. See an example in figure 12.5.1.7. 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.

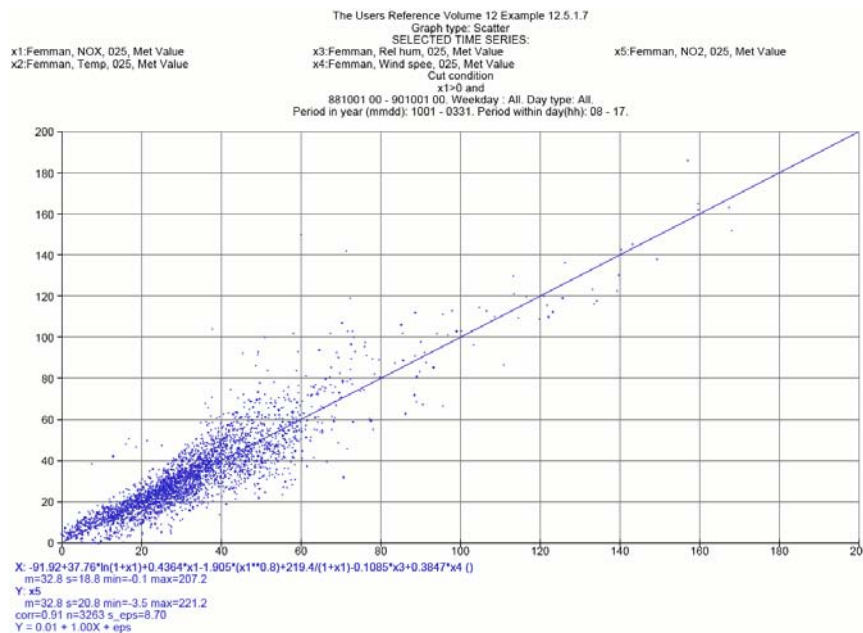


Figure 12.5.1.7 Scatterplot of measured contra estimated NO₂ concentrations.

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 12.5.1.8 shows that the NO₂ distribution from the model is very close to the curve from observed NO₂.

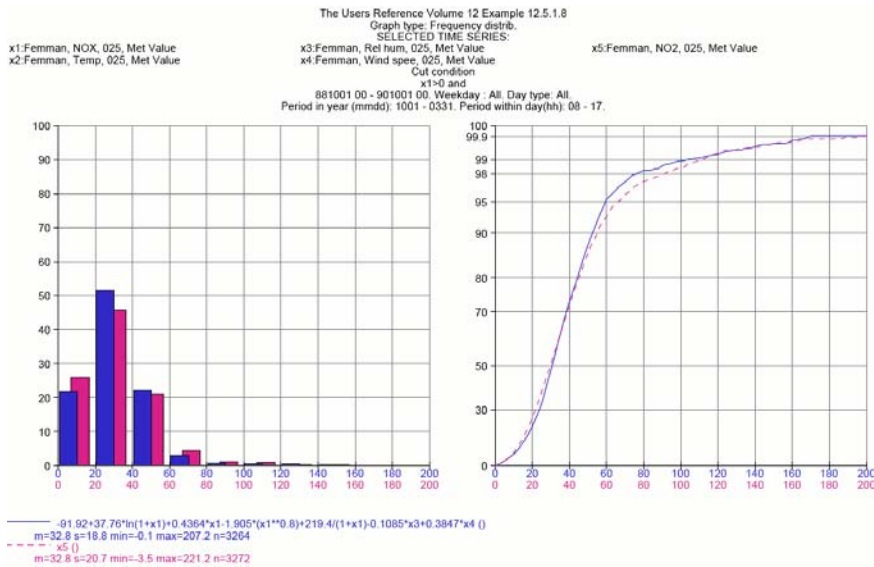


Figure 12.5.1.8 Frequency distribution graph with measured and estimated NO₂ concentrations.

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.

12.5.1.1 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 scatterplot, 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 $Y = b_0 + x b_1$ no transform of x

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

Inverse $Y = b_0 + b_1$ inverse transform
of x

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

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

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

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

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

x Exponential $\ln(Y) = \ln(b) + b x$ 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 $b_0 \exp(b_1 x)$.

In the scatterplot, 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 scatterplot 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.

12.5.2 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; see section 12.3.1.

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

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

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

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

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

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

Categorical predictors can be used in the REEP analysis, but each category can also be transformed into a binary predictor. If you decide to do this, bear in mind that no binary predictor can be an exact linear combination of other predictors; one category must be left out, for mathematical reasons. It doesn't matter which category that is excluded. For categorical predictors, which can be ordinal or nominal, you may have to normalise the categories into the interval]0,1[.

For more information, see "Miller, R.G. (1964): Regression estimation of event probabilities. Technical Report No 1. The Travelers 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."

12.6 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:

$$m_i = \frac{1}{M} \sum_{t=1}^M x_i \text{ and } s_i = \sqrt{\frac{1}{M-1} \sum_{t=1}^M (x_i - m_i)^2}$$

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

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

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:

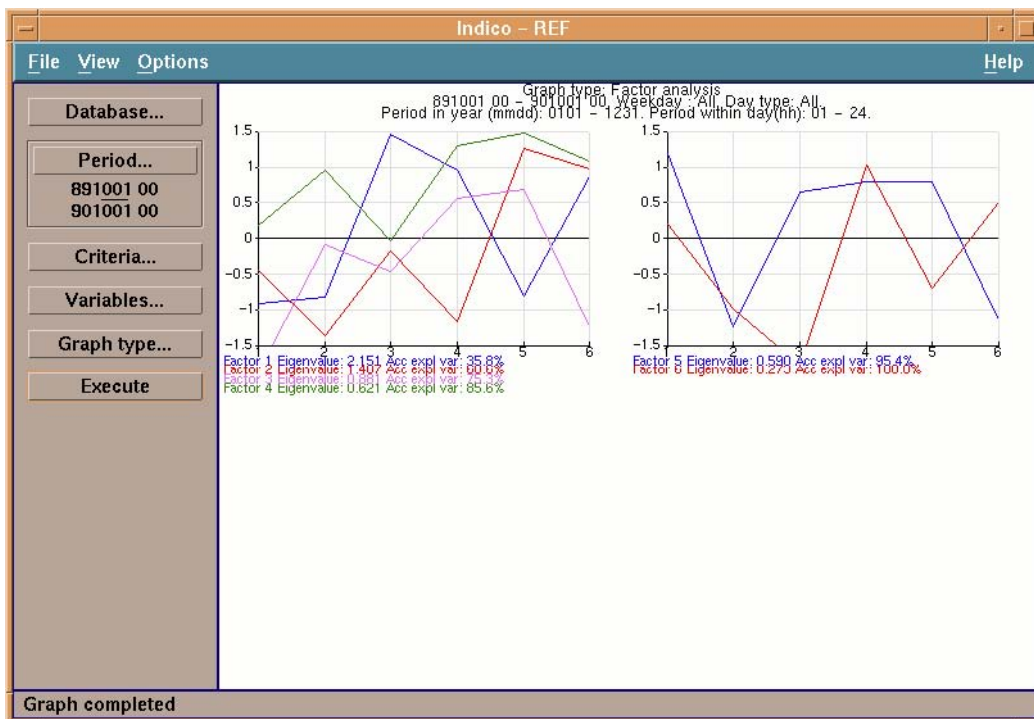
$$z_i(t) = \sum_{m=1}^N f_m(t) \times h_m(i) \text{ for } i=1,2,\dots,K$$

$n=1$

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

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

In order to run the **Factor analysis**, select your independent statistical variables in the variables frame, go to the Graph type frame and select **Factor analysis**. Click **Apply** and send the result to **Output Graph**. See figure 12.6.1. 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**.



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 eigenvalue is a diagnostic measure of the collinearity of the factor with your original data. The eigenvalues 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 eigenvalue 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.

12.6.1 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:

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

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 12.6.2.

Hide table						
Factor	1	2	3	4	5	
Eigenvalue	0.000	0.000	0.000	0.000	0.000	
Var	Acc. expl. var	0.0%	0.0%	0.0%	0.0%	0.0%
1	Eigen. vec	0.00	0.00	0.00	0.00	0.00
	Expl. var	0%	0%	0%	0%	0%
2	Eigen. vec	0.00	0.00	0.00	0.00	0.00
	Expl. var	0%	0%	0%	0%	0%
3	Eigen. vec	0.00	0.00	0.00	0.00	0.00
	Expl. var	0%	0%	0%	0%	0%
4	Eigen. vec	0.00	0.00	0.00	0.00	0.00
	Expl. var	0%	0%	0%	0%	0%
5	Eigen. vec	0.00	0.00	0.00	0.00	0.00
	Expl. var	0%	0%	0%	0%	0%

Figure 12.6.2 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 eigenvalue is a diagnostic measure of the collinearity of the amplitude functions with your original data. The eigenvalues 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 eigenvalue 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.

12.7 Using Indico macros

In the previous chapters we have given a number of recommendations for working with presentation and statistical analysis of time series. There are many possibilities to set selection criteria, make transformations, create statistical models or even self-adapting filters. Perhaps you have also made a layout of the graph. After all this work you realise

that it would be convenient to save all these settings, to be able to use them another time.

In principle, a setup of Indico is equivalent with defining a macro, an automation object or a settings file. It is very simple to save your settings:

- On the left-hand menu, choose **Macro**
- Choose **Save**
- Select your username from the list
- Enter a unique and meaningful name for your settings
- Click **Apply**



If your username doesn't exist in the macro group list, you can create it, or some other subgroup by clicking **Create new subgroup**.

Macros are stored in groups, where there is a group for each user in a domain, a common group and some other groups. The system administrator decides who is allowed to store macros in the common group (using `Indico.WriteGroup.user` in `priv.rf`). Users can always save macros in their own group, but usually not in other groups, although it is possible to load macros from other groups.

You can have your settings back at any later time:

- On the left-hand menu, choose **Macro**
- Choose **Load**
- Select a group or select a blue macro name from the list
- Decide if you want to change

period · Click **Apply**



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

When you have loaded your macro, you can output the result to a graph or change your settings.

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

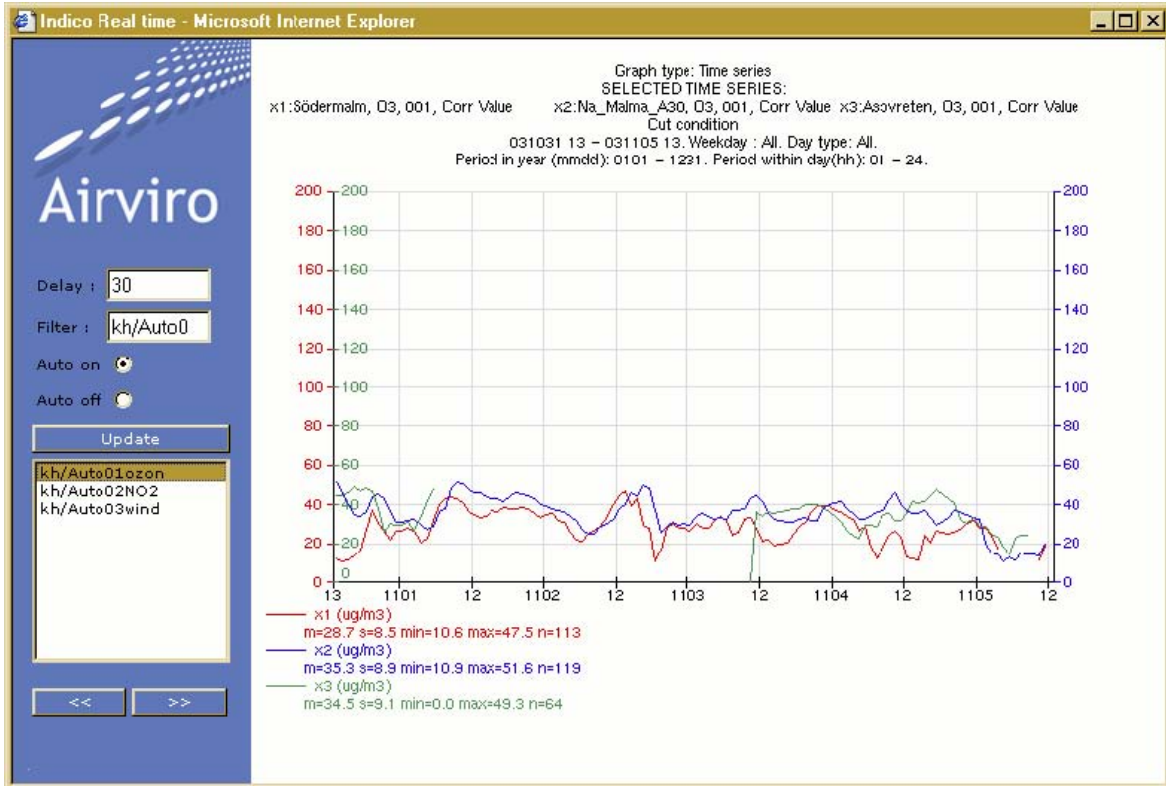
12.8 Real Time Graph

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

You first need to choose **Time Series** data to Real Time Graph that are collected frequently. In **Period**, it is not important which time you choose in the **To** field, it is the duration of the period that is taken into account. This is because Real Time Graph always uses the current time as stop time.

Once you are satisfied with your graph, save it as an Indico macro. The macro must start

with the string “Auto” to be recognised by Real Time Graph. All macros starting with the string “Auto” are displayed in alphabetical order. You can also control the order in which the macros are displayed by using a number in the macro name, immediately following the “Auto” string.



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

In the Real Time graph, you can set the number of seconds that each macro will be displayed for in the **Delay** field. See figure 12.8.1. If you want to select only a subset of all Auto macros, you can use a **Filter** to select your group or some macro number.

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

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

closer.

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

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