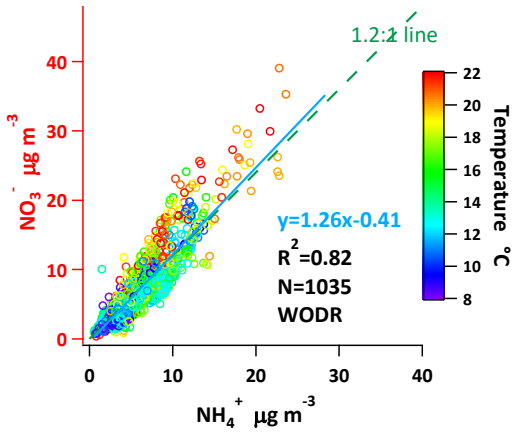
Scatter Plot



Manual for Scatter plot

Wu Cheng

wucheng.vip@foxmail.com

2017-03-16

Preface

Scatter plot is a handy tool to maximize the efficiency of data visualization in atmospheric science. Many existing generalized data visualization software are great, but could not satisfy many specify research purposes in atmospheric science so I develop my own program. The program includes Deming and York algorithm for linear regression, which consider uncertainties in both X and Y, that is more realistic for atmospheric applications. It is Igor based, and packed with lots of useful features for data analysis and graph plotting, including batch plotting, data masking via GUI, color coding in Z axis, data filtering and grouping.

The latest version of the program can be found on my website:
https://sites.google.com/site/wuchengust/

Wu Cheng

2017-03-16

Contents

1 Recommendation on data structure

Excel is recommended for storing data if the size of the data is less than 1 million rows. Otherwise, .csv file is recommended. If possible, put all data with the same timeline in a single sheet to maximize the efficiency, subset can be extracted by filtering rather than manually put them in separated sheets. The structure of the data are shown below in Figure 1.1. The first row is header (text). After import in Igor, header will become the name of waves (columns in Excel). Space and other illegal characters are not allowed in Igor as wave name and will be replaced by "_". Data falls into three categories:

1) Timestamp (timeline)

2) Numerical data (e.g. concentrations of air pollutants)

3) Text data (Makers, e.g. site name, back trajectories clusters)

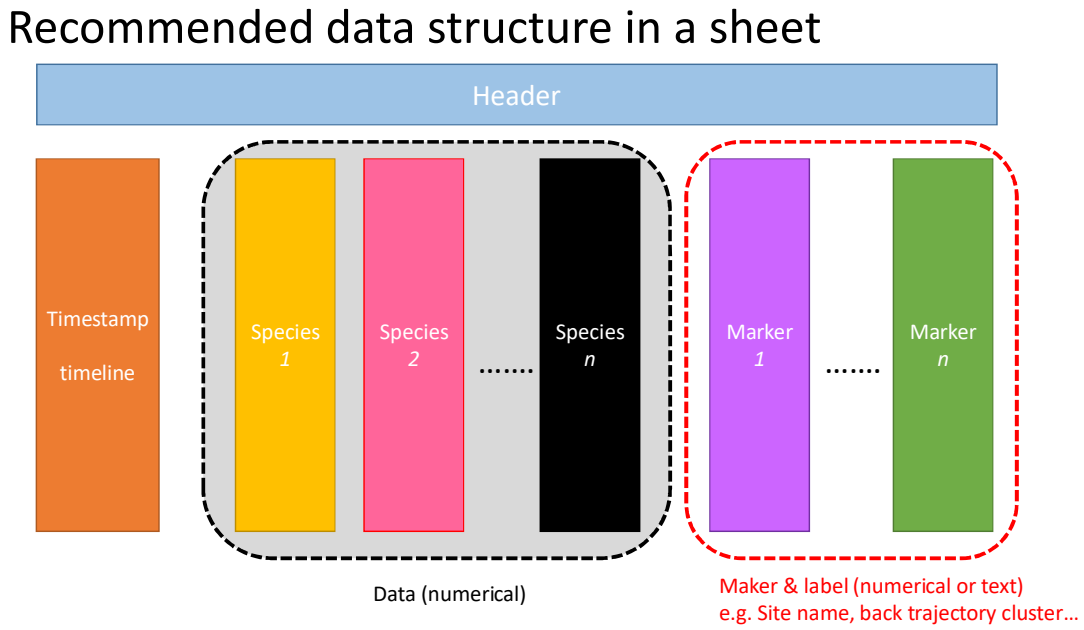


Figure 1.1 Recommended data structure in a sheet

An actual example of excel data (or .csv file) is shown in Figure 1.2. It should be noted that the order of different data columns (waves) is not necessarily the same as figure 1.1, the three categories can be mixed, there is no restrictions on data column order. As shown below, “DateIndex” is timeline, “Sample ID” and “Site” is text data (marker), rest columns are numerical data.

	A	B	E	F	G	H	I	J	K	L	CL
1	**DateIndex**	**Sample ID**	**TGC**	**QGC**	**NaIC_C**	**NH4_C**	**KIC_C**	**CLIC_C**	**NO3_C**	**SO4_C**	**Site**
2	1/13/11	MK110113	61.9167	69.2917	1.9802	6.4493	0.5765	0.8873	11.6865	10.2778	MK
3	1/25/11	MK110125	89.8333	101.3750	2.3110	10.9636	0.9455	0.9994	12.1388	22.2418	MK
4	1/27/11	MK110127	59.0417	66.6250	2.5072	5.8765	0.4537	0.8605	9.2997	10.7022	MK
5	1/31/11	MK110131	66.6667	73.9167	0.2254	7.7103	0.8675	0.4770	5.0206	16.8996	MK
6	2/5/11	MK110205	64.7500	73.0000	0.2102	8.3566	1.4050	0.1443	7.8915	17.7907	MK
7	2/9/11	MK110209	65.3333	72.7500	0.4780	9.0343	1.1588	0.4825	7.6093	19.9455	MK
8	2/11/11	MK110211	59.3750	65.8750	0.4283	6.6911	1.1546	0.1361	7.0313	13.4828	MK
9	2/15/11	MK110215	49.9583	52.3750	0.1801	6.4300	0.6436	0.6742	5.4804	13.6615	MK
10	2/23/11	MK110223	45.7083	48.7083	0.4691	4.6793	0.3861	0.2296	3.7037	10.7737	MK
11	2/25/11	MK110225	53.6667	63.6667	0.4837	6.7622	0.3832	0.3557	5.1990	15.1039	MK
12	3/1/11	MK110301	45.9167	53.5417	0.3452	5.0612	0.1890	0.2956	4.2997	10.6106	MK
13	3/10/11	MK110310	48.1667	53.3750	0.1575	3.2226	0.2605	0.2542	4.2285	11.9927	MK
14	3/13/11	MK110313	76.2500	79.7917	0.2476	10.6859	0.4086	0.1520	11.0401	20.7006	MK
15	3/25/11	MK110325	63.8750	70.8750	0.3131	7.1179	0.6857	0.2667	3.8859	17.7513	MK
16	3/29/11	MK110329	66.7500	74.0417	0.4628	6.7928	0.8272	0.2829	4.7515	15.8878	MK
17	3/31/11	MK110331	44.7917	50.7083	0.4477	4.2772	0.3880	0.2136	2.7727	10.1044	MK
18	4/9/11	MK110409	49.8750	56.9583	0.5887	6.7063	0.3916	0.3034	3.8906	14.8415	MK
19	4/12/11	MK110412	64.3333	74.2500	1.3417	7.3976	0.6255	0.1737	1.8103	21.5748	MK
20	4/18/11	MK110418	33.5417	44.2500	0.1214	3.0270	0.2772	0.0202	0.7076	8.1754	MK
21	4/24/11	MK110424	43.0417	54.2500	0.2250	4.8682	0.3361	0.0564	1.7277	13.0045	MK
22	4/30/11	MK110430	54.5833	61.6667	0.7725	6.1938	0.4845	0.0502	1.1593	20.6579	MK
23	5/6/11	MK110506	31.2917	41.9167	0.4799	3.4637	0.1624	0.0148	0.2584	10.8408	MK
24	5/18/11	MK110518	31.5000	38.9583	0.2331	3.0178	0.1924	0.0224	0.4353	8.6534	MK
25	5/20/11	MK110520	28.7083	34.8333	0.2930	2.7701	0.1236	0.0201	0.3417	8.0928	MK

Figure 1.2 An actual example of excel data (or .csv file).

2 Overall comparison with other programs

The following table compares scatter plot with other programs

Software	Advantage	Disadvantage
Excel	Powerful data filter	OLS only, no Deming regression row size limited to 1 million Can't mask data No color coding
SPSS	data filter	OLS only, no Deming regression Data masking can't by done by GUI
Sigma Plot	Deming regression	No data filter, can't mask data
Origin	York regression Data masking via GUI color coding	No data filter
Scatter plot Igor program	OLS, Deming, Weighted orthogonal distance and York Regression Data filter Data masking via GUI color coding Batch plotting	Require Igor to run the program

3 Import Data

3.1 Timeline example in MS Excel

Before import, data could be kept in Excel, restrictions on timeline format are described as follows. The timeline in data column **must** follow this format "MM/DD/YY hh:mm", Location **must** be “English (United States)”as shown in Figure 3.1.

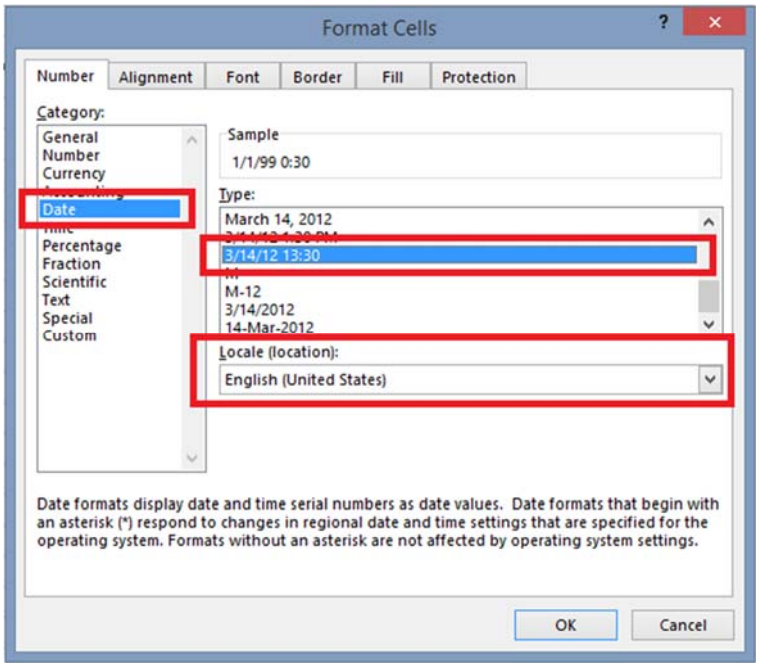
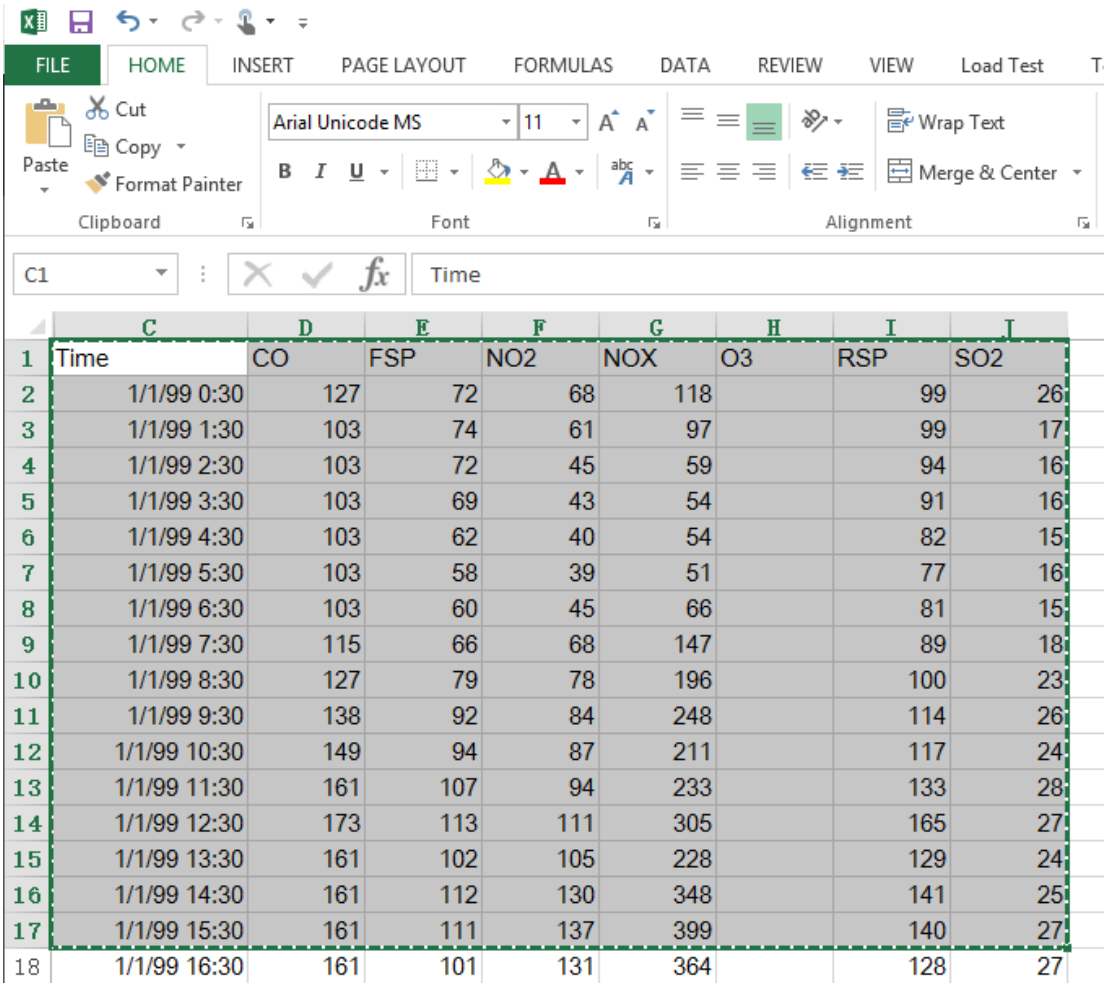


Figure 3.1 Cell format configuration in MS Excel for the timeline column

Make sure the cell format of the timeline column is exactly the same as shown in Figure 3.1, otherwise logr cannot recognize it.

3.2 Copy from Excel

Data can be imported via copy & paste from MS Excel, as shown in Figure 3.2. It is recommended to put timeline in the first column.



	C	D	E	F	G	H	I	J
1	Time	CO	FSP	NO2	NOX	O3	RSP	SO2
2	1/1/99 0:30	127	72	68	118		99	26
3	1/1/99 1:30	103	74	61	97		99	17
4	1/1/99 2:30	103	72	45	59		94	16
5	1/1/99 3:30	103	69	43	54		91	16
6	1/1/99 4:30	103	62	40	54		82	15
7	1/1/99 5:30	103	58	39	51		77	16
8	1/1/99 6:30	103	60	45	66		81	15
9	1/1/99 7:30	115	66	68	147		89	18
10	1/1/99 8:30	127	79	78	196		100	23
11	1/1/99 9:30	138	92	84	248		114	26
12	1/1/99 10:30	149	94	87	211		117	24
13	1/1/99 11:30	161	107	94	233		133	28
14	1/1/99 12:30	173	113	111	305		165	27
15	1/1/99 13:30	161	102	105	228		129	24
16	1/1/99 14:30	161	112	130	348		141	25
17	1/1/99 15:30	161	111	137	399		140	27
18	1/1/99 16:30	161	101	131	364		128	27

Figure 3.2 Example of data selection and copy (Ctrl + C) in MS Excel. The header of each column will be used as wave name in Igor.

3.3 Paste data into Igor

Put the cursor on up left corner and paste the data into the table in the Igor program interface (the highlighted area in orange) as shown in Figure 3.3.1

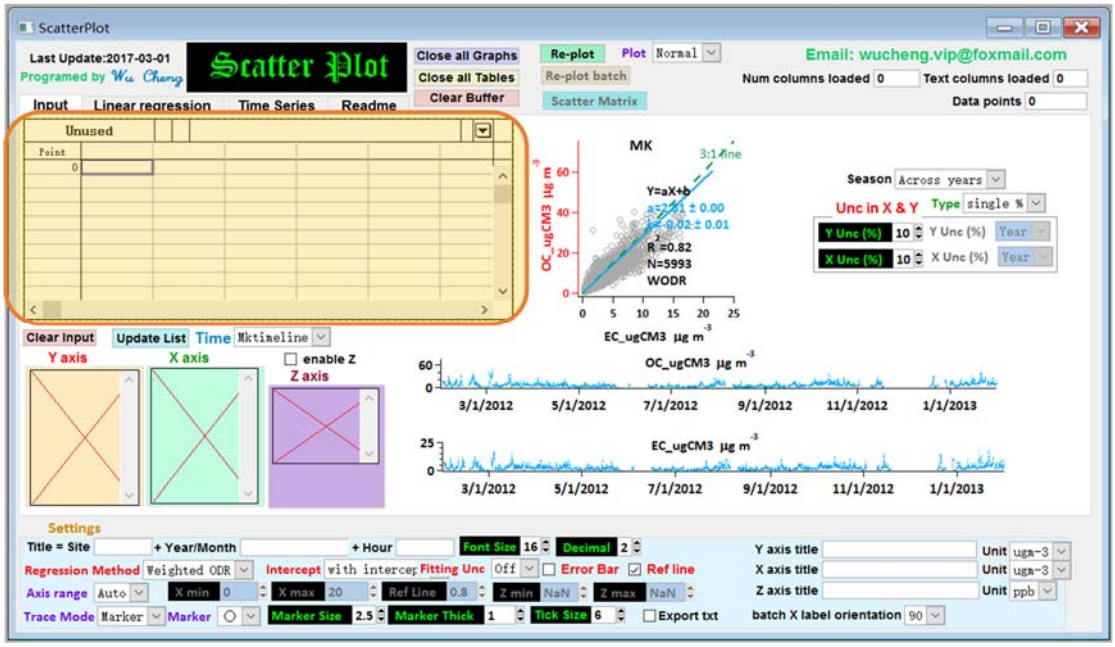


Figure 3.3.1 Example of user interface in scatter plot Igor Pro program before pasting data.

After applying paste (ctrl + V), the data will show up in the table area, make sure the timeline is recognized properly by Igor Pro. It should be noted that the index of data points starts from 0 in Igor.

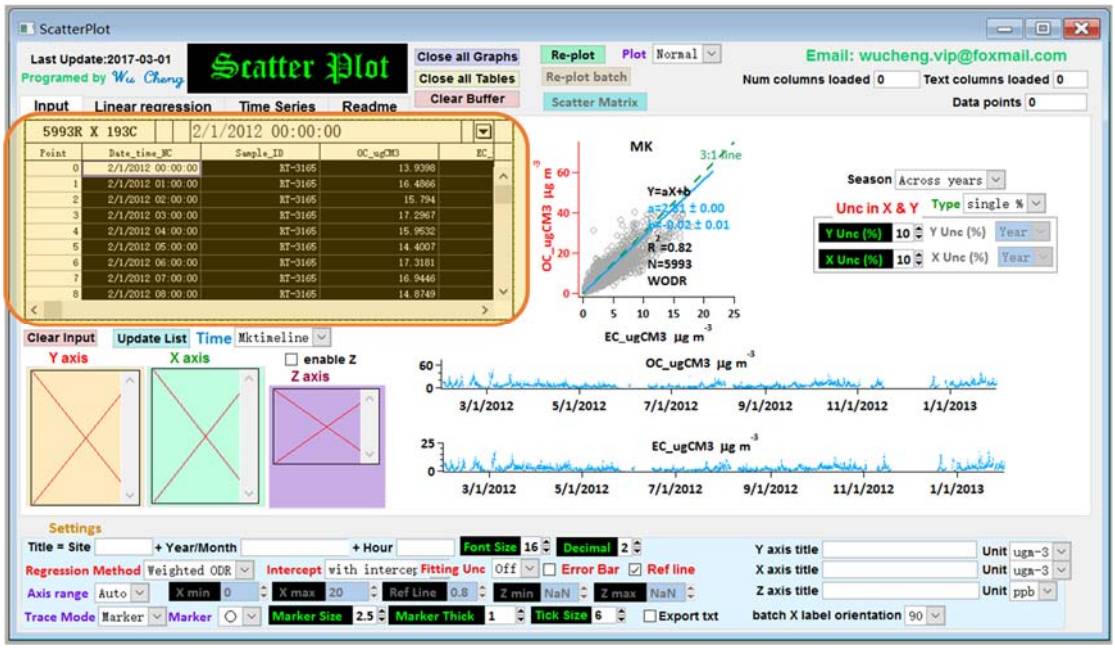


Figure 3.3.2 Example of user interface in scatter plot Igor Pro program after pasting data

3.4 Update list

Click “Update List” button (Figure 3.4, highlighted area a), then the list numerical data series (known as column in Excel and wave in Igor Pro) will be updated (Figure 3.4, highlighted area b). The statistics of loaded data are shown in Figure 3.4 highlighted area c, including number of numerical columns, text columns and data points (rows).

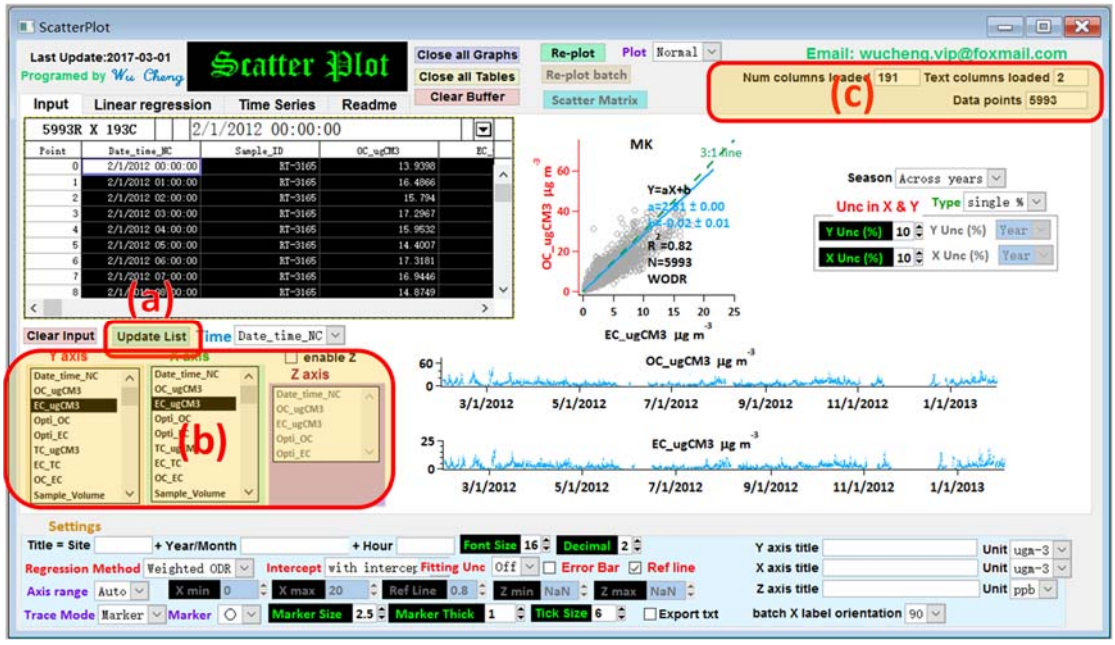


Figure 3.4 Example of Update list in Igor.

3.5 Specify timeline

The next step is to tell the program which column is the timestamp. It can be done by using the pop-up menu as shown in Figure 3.5.

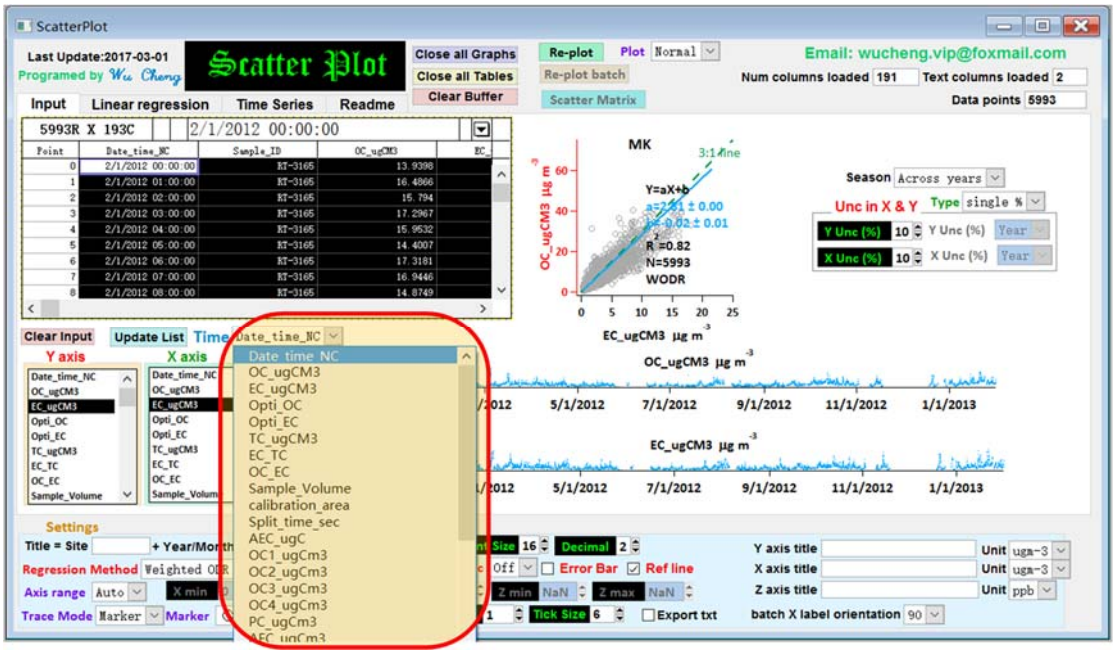


Figure 3.5 Example of specifying timeline in Scatter plot Igor program.

4 Introduction to the general settings

General settings of scatter plot Igor program are shown in Figure 4.1, which include,

Close all Graphs: Close all Graphs in the new windows

Close all Tables: Close all Tables in the new windows

Clear Buffer: erase data for batch plots and keep file size of the program

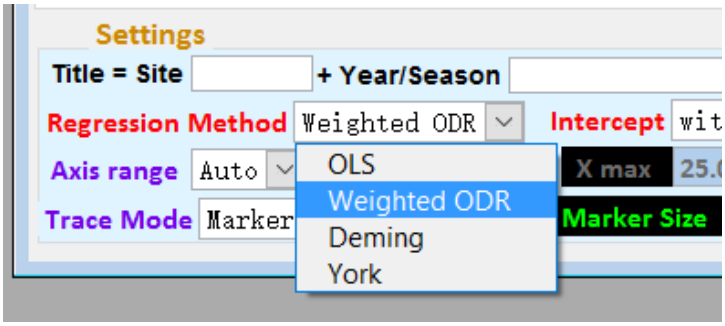
Replot: replot the figure in the current tab

Plot option: Normal, only replot the current tab; New window, also generate the plot in a new window that can be copy&paste to MS office; Export PNG, not only generate plots in new windows, but also a PNG file; Export EMF, instead of PNG file, EMF is a vector file that can zoom in infinitely. The exported PNG and EMF files are placed in the same folder that igor .exp file (this Histbox program) is placed.

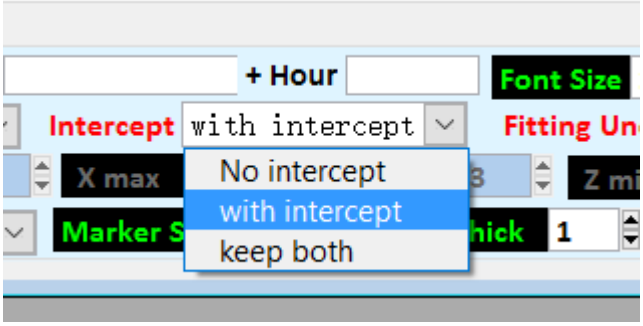
Title: Title of plot, contain three fields (Site, Year/Month, Hour and species), these fields will use auto naming when doing a scan if they left as blanks

X/Y/Z title: If left as blank, wave name will be adopted, otherwise will be overridden by user input.

Regression method: Ordinary Least Squared (OLS), Weighted orthogonal distance regression (WODR), Deming regression, York regression. The OLS only consider errors in Y, while the later three consider uncertainties in both Y and X.

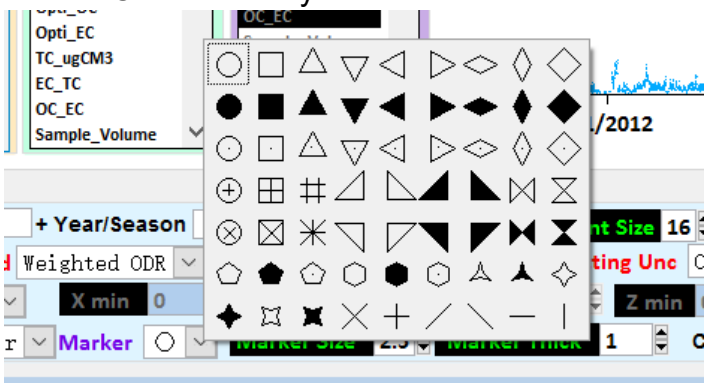


Intercept: If "No intercept" is selected, regression will be performed through origin (Not available for Deming and York Regression). If "with intercept" is selected, all regression methods are available. If "keep both" is selected, both with and without intercept regression will be performed (Not available for Deming and York Regression).

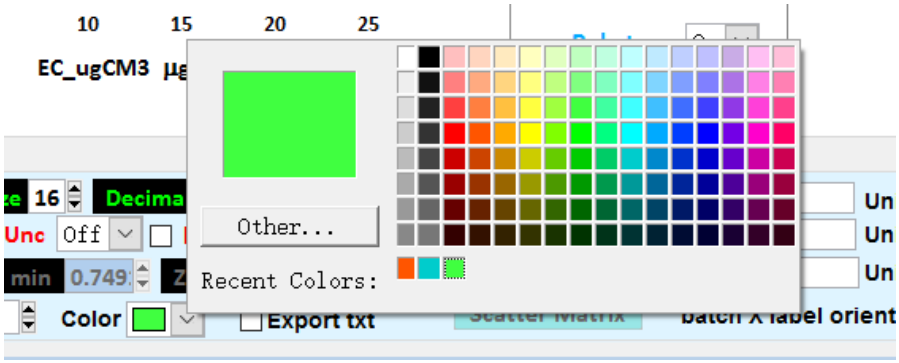


Axis range: turn on or off auto scale

Unit: use pre-set unit on conc. axis
Decimal: decimal point
Font Size: control font size in figure
Trace Mode: Choose between dot and Marker for display the data points on scatter plot.
Maker: Choose the symbol for the marker



Marker Size: Control the size of the marker.
Color: Set marker color (only works when Z axis if **off**)



Ref Line: set the ratio of Y:X for the reference dash line

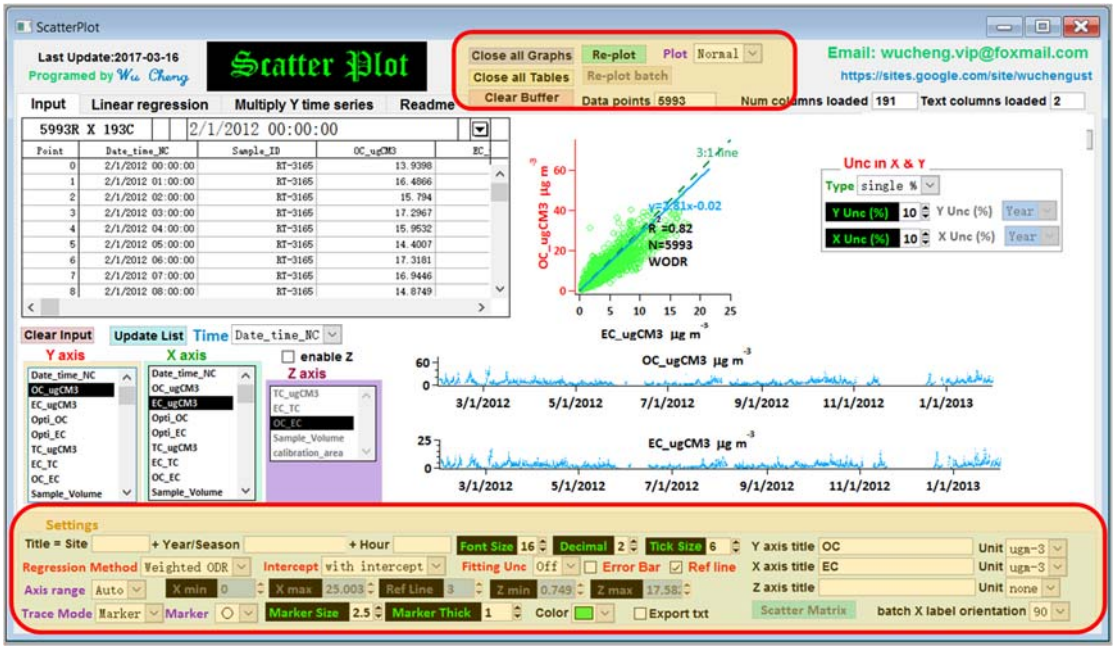


Figure 4.1 Example of general settings in scatter plot Igor program.

5 Tab “Input” Introduction

(a) User can choose which species to be X, Y and Z (Z can be switch on or off). By choosing different X Y Z combinations, the scatter plot and the two time series plots will be updated instantly so user can take a quick look at the dataset

(b) Unc in X&Y

Uncertainty setting for WODR, Deming and York regression. Two types of input is available, "Single %" means user just need to provide a single number to indicate the relative uncertainty in X and Y. "Input Data" means user need to provide weighting for individual data points (a separate wave in the form of standard deviation). User need to use the pop-up menu to indicate the corresponding weighting wave for Y and X.

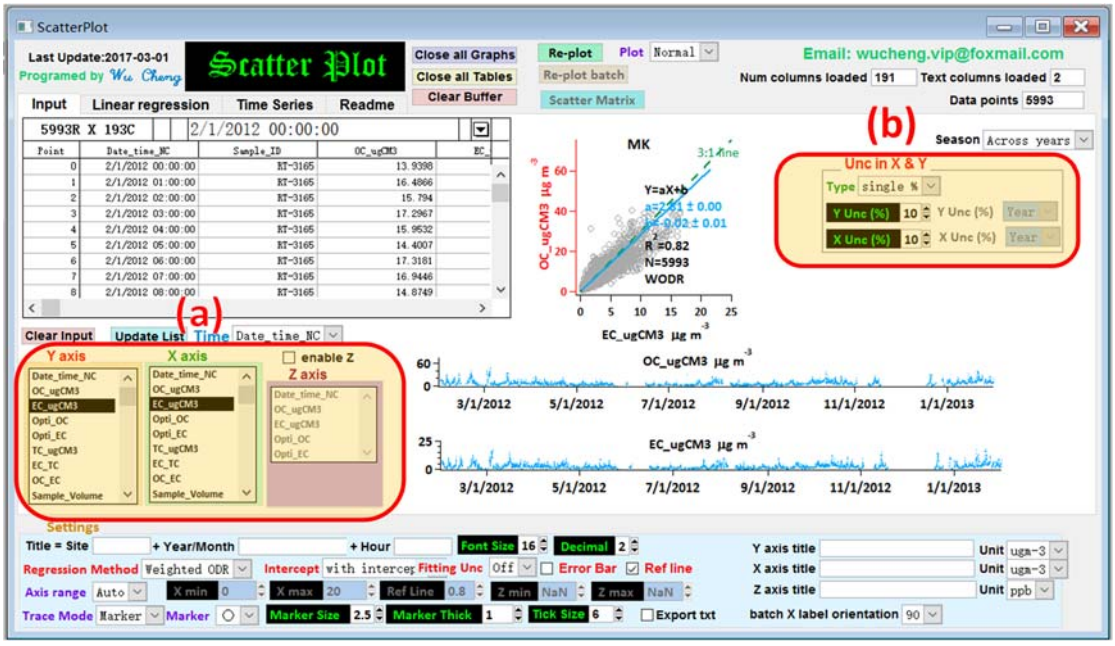
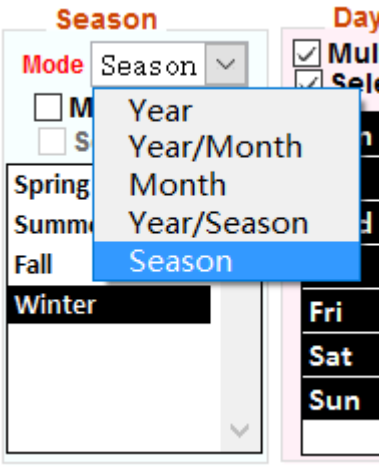


Figure 5.1 Scatter plot settings

6 Tab “Linear regression ” Introduction:

6.1 data filter by time

Three type of time scales are used for data filtering: YSM (year season month), Dow (day of week) and Hour (0:00~23:00)



(a) YSM can be further divided into five scenarios: Year; Year/Month; Month; Year/Season; Season. Season is defined by the highlighted area (d) using the first day of each season as cut-off. Two options are available, across year and same year. For example, if select across year, 1999 Dec and 2000 Jan are grouped together as 1999 winter. Multiply selection is possible using the shift key during selection.

(b) Dow (day of week). Multiply selection is possible using the shift key during selection.
(c) Hour (0:00~23:00). Multiply selection is possible using the shift key during selection.

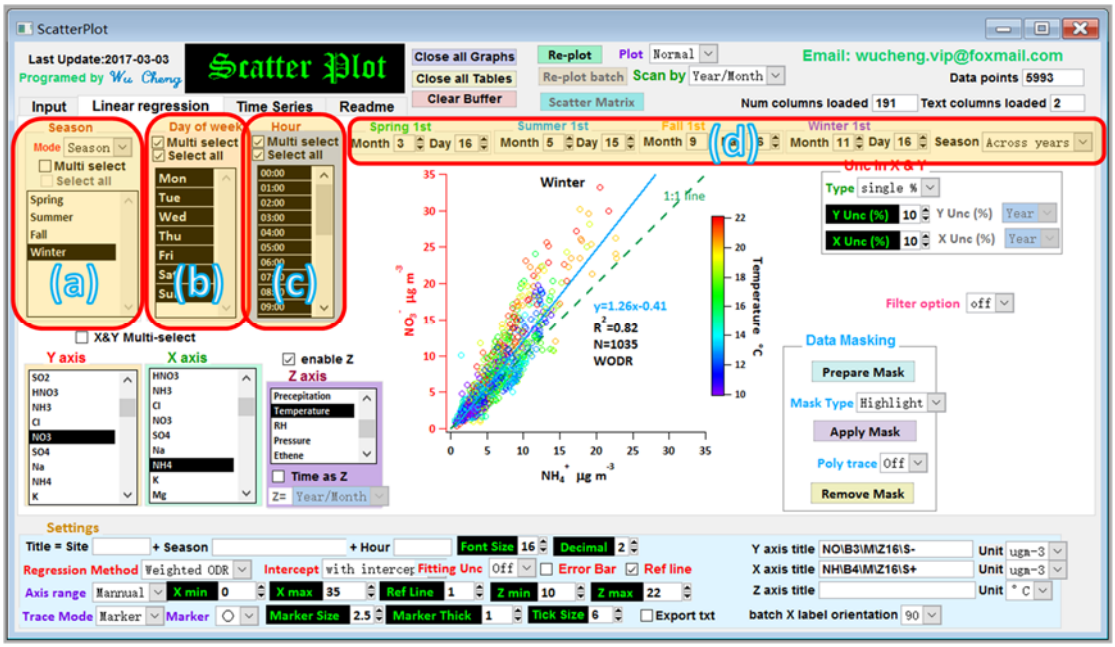


Figure 6.1.1 List boxes for time filtering

6.2 Data filter by data

Three types of data filtering are possible: Text data by list, Numerical data by List and Numerical data by range.

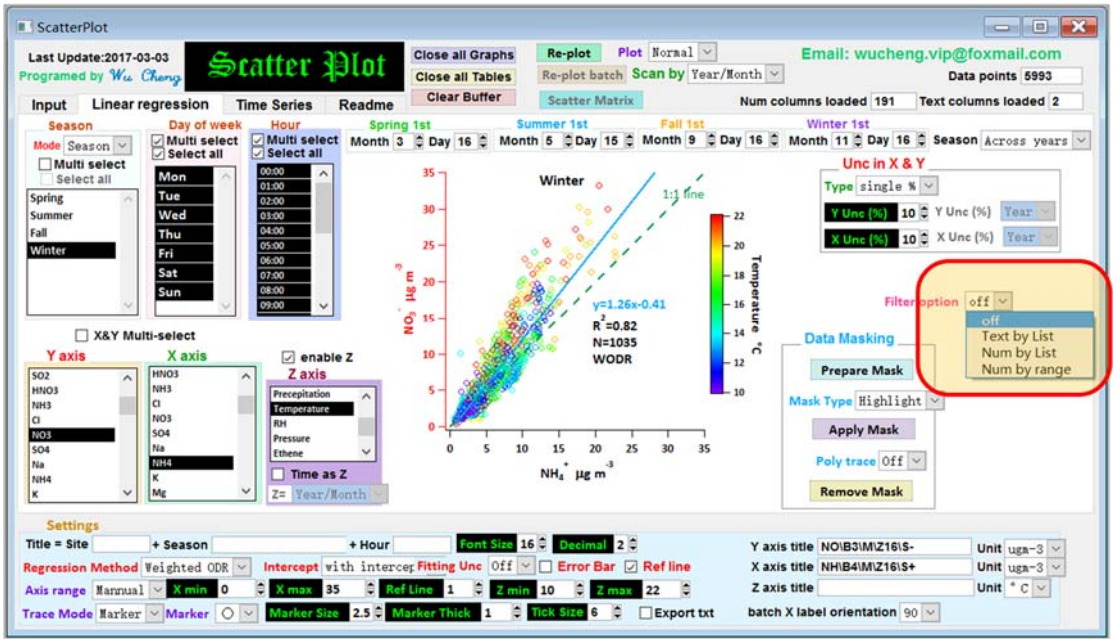


Figure 6.2.1 Data filter by data

(a) **Text by list**: say a column provide the back trajectories grouping info that includes C1~C4, using this function, a subset (e.g. as shown below, only C4 in winter) is plotted. Multiply selection is possible.

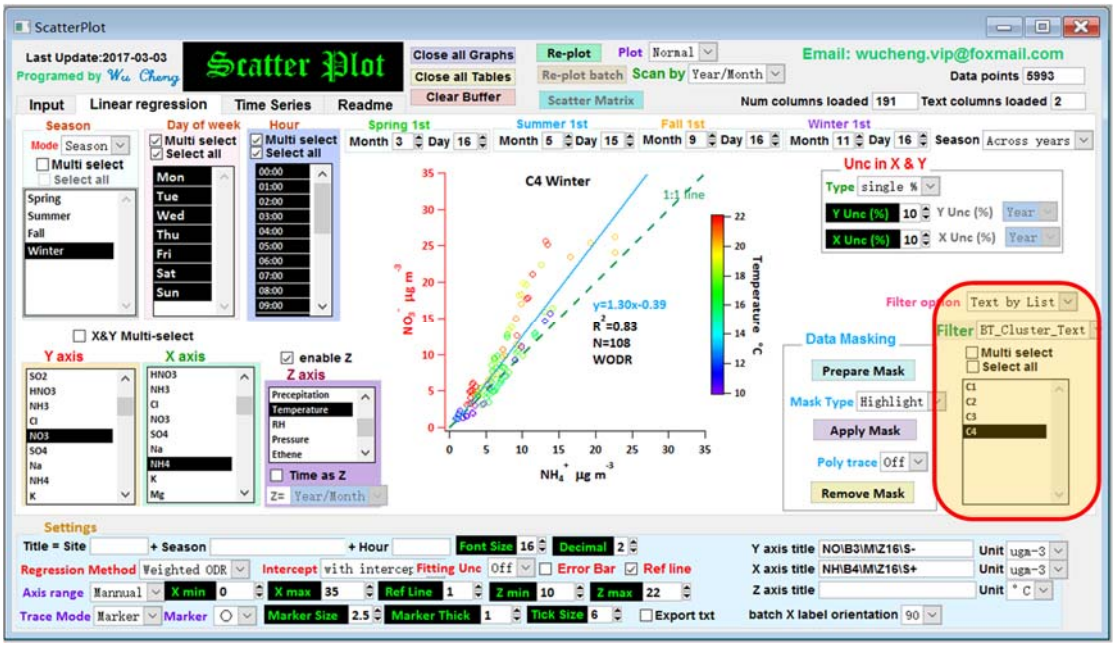


Figure 6.2.2 Data filter by data – Text by List

(b) **Num by List**: a column with numerical values can be used a filter for data grouping. It's useful when the number of unique values are much smaller than the total counts. For example, data is grouped by RH as shown below.

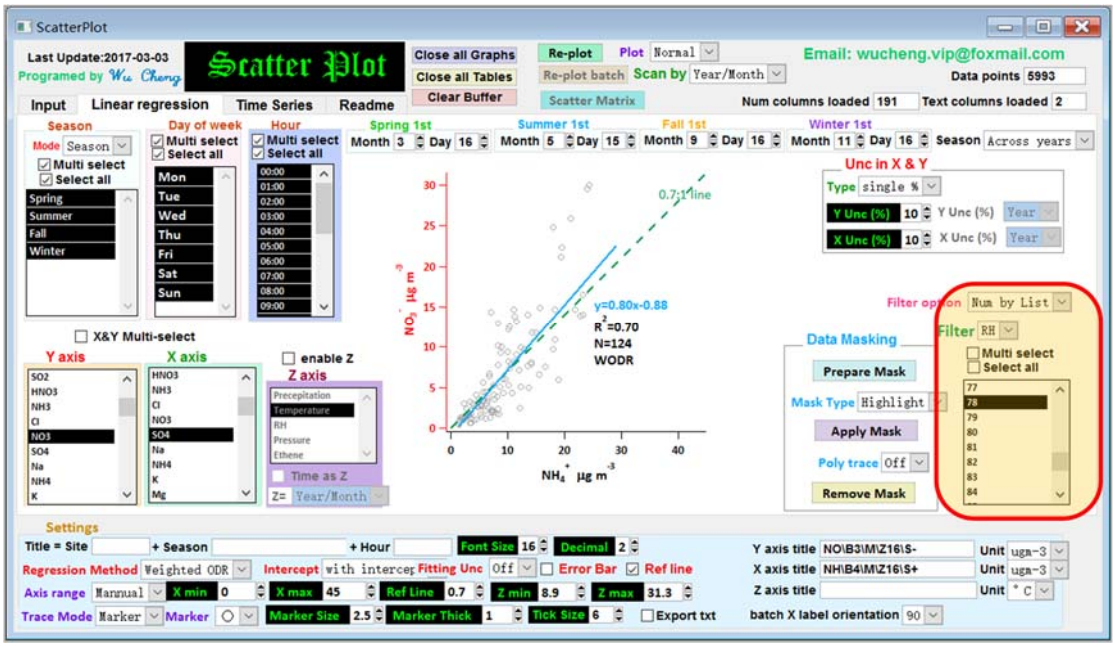


Figure 6.2.3 Data filter by data – Num by List

(c) **Num by range**: A range (defined by min and max) can be used for individual columns to extract a subset. When multiply columns are used, intersection of these conditions are applied. Following is an example of using 75<RH<85.

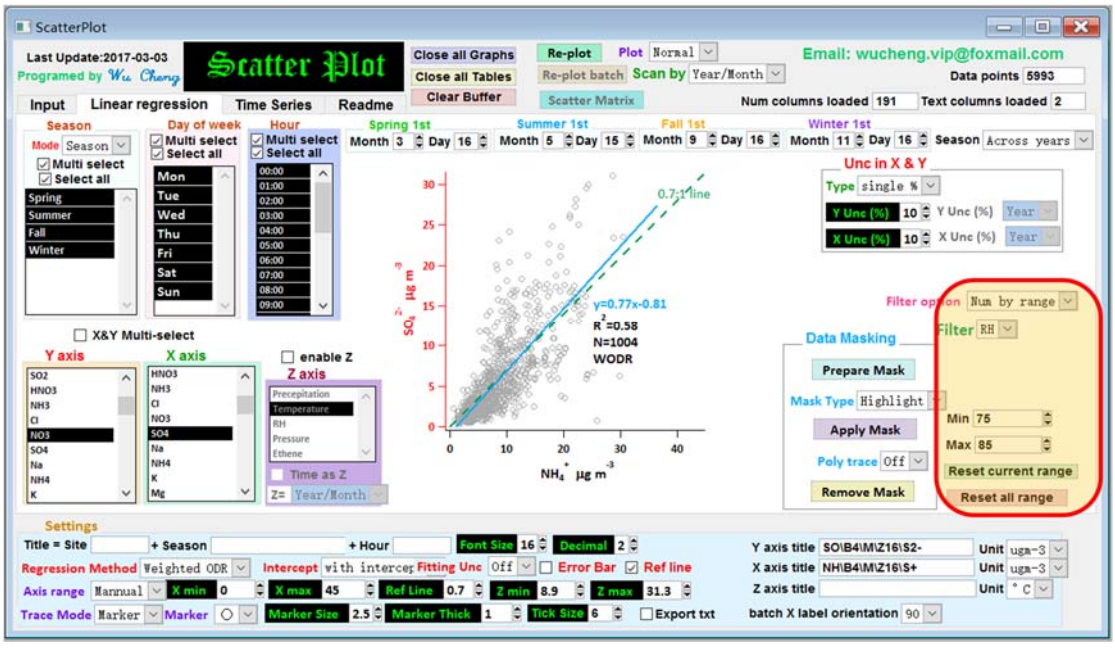


Figure 6.2.4 Data filter by data – Num by range

6.3 Data masking via GUI

Data Masking feature to exclude unwanted data points for regression. Data Masking can be applied using graphic user interface.

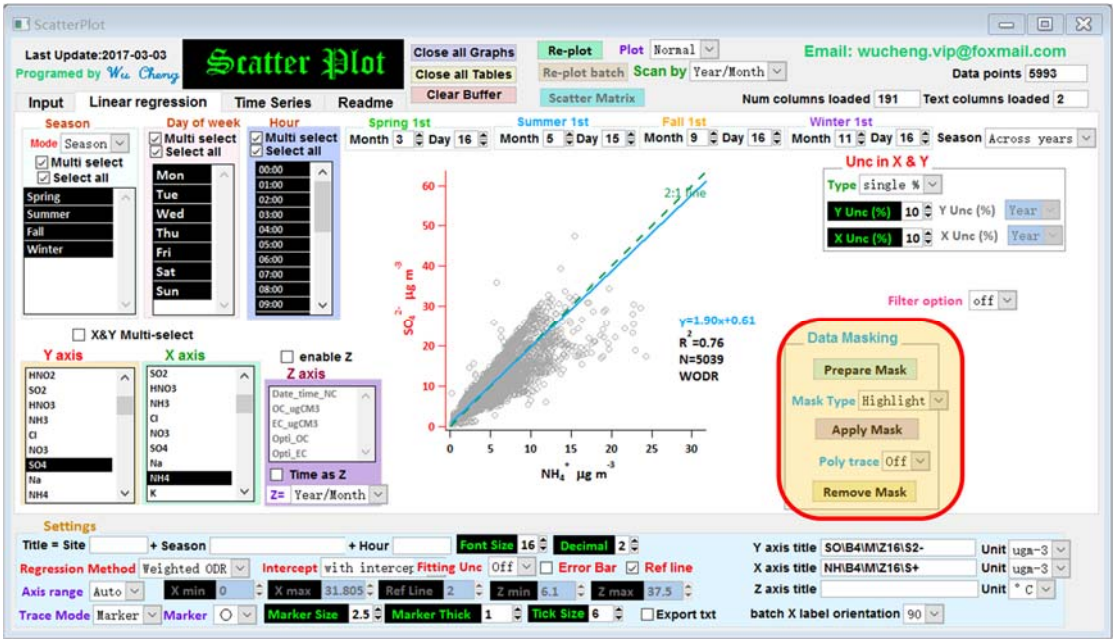


Figure 6.3.1 Data Masking overview

Firstly, click the "Prepare Mask" button. Then a polygon can be drawn using the cursor. The polygon is defined by way points.

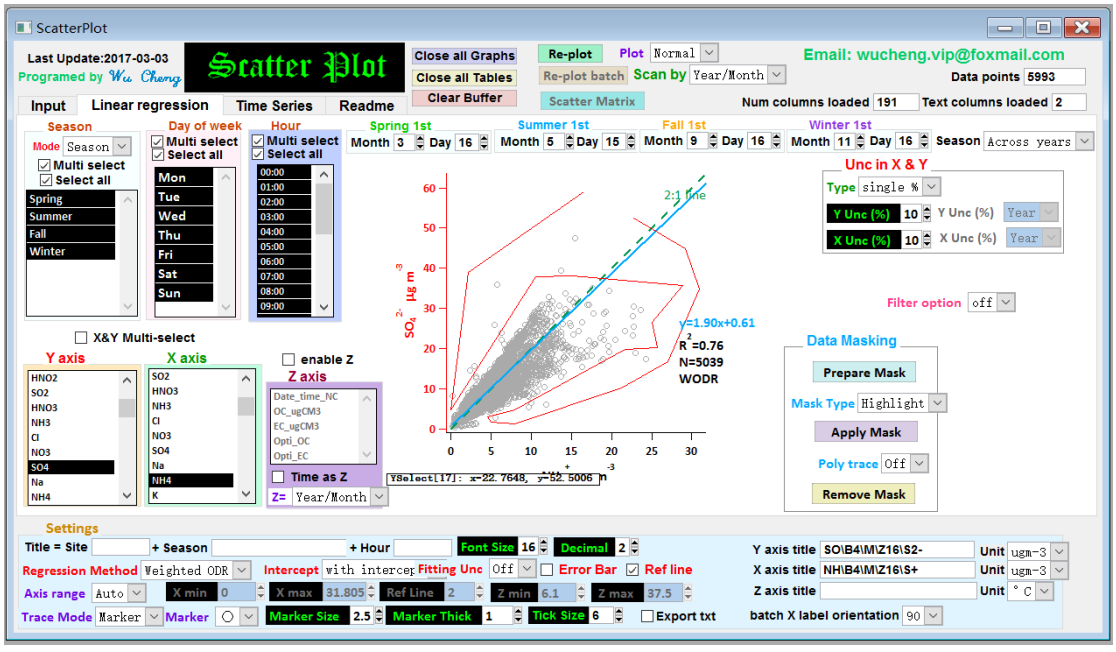


Figure 6.3.2 Start a polygon drawing

Make sure the polygon is closed as shown in Figure 6.3.3, each way points will be shown in the form of square markers.

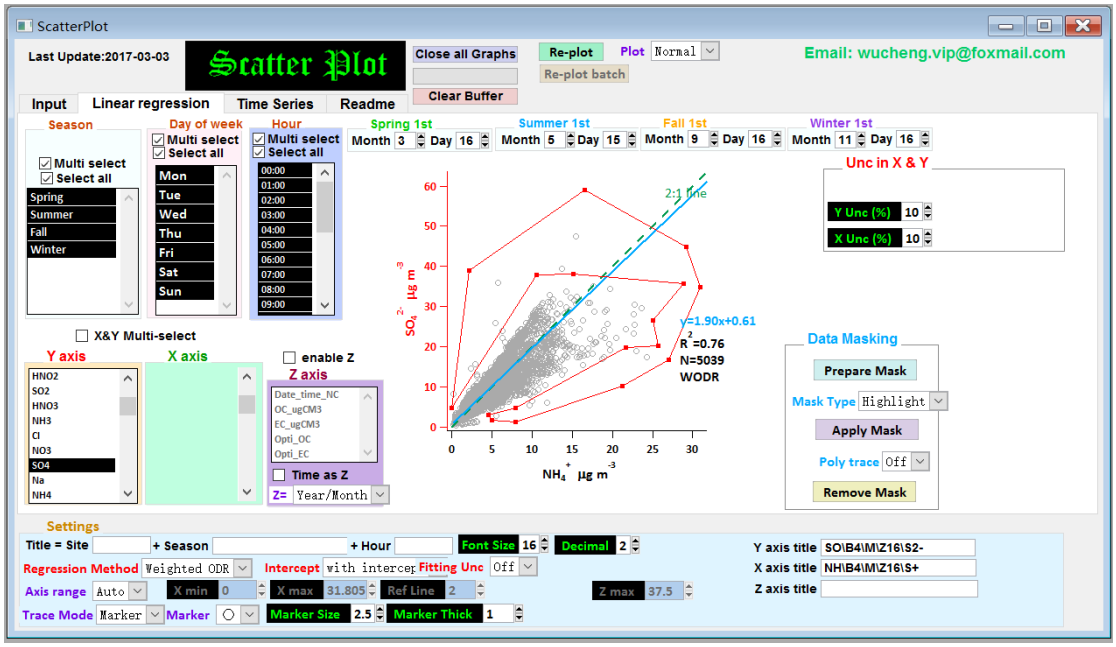


Figure 6.3.3 An example of closed polygon.

Once polygon is finished, choose "Mask type". Following is an example of choosing “Highlighted”, then click "Apply Mask" button. Unwanted data points are marked as pink triangles. Data points inside the polygon are excluded for regression (N changed from 5039 in Figure 6.3.3 to 5026 in Figure 6.3.4). For this particular example, the removal of unwanted data points didn’t affect the slope and intercept, but R^2 do improve from 0.76 to 0.78.

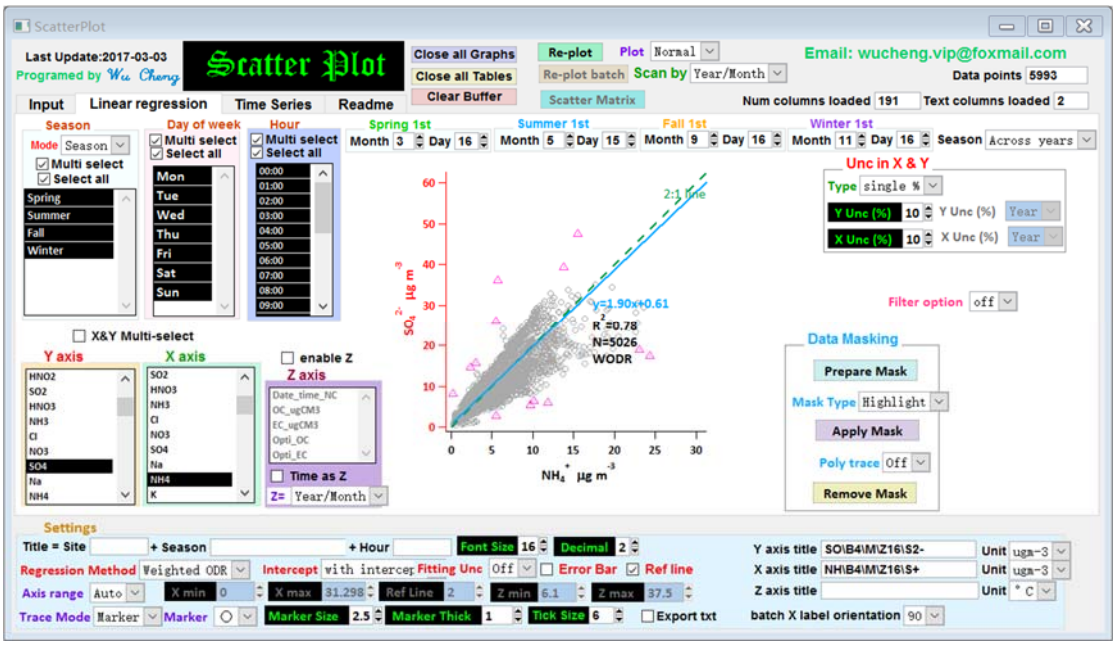


Figure 6.3.4 An example of choosing “Highlighted” in "Mask type".

Following is an example of choosing “Remove” in "Mask type", then click "Apply Mask" button. Unwanted data points are removed. Data points inside the polygon are excluded for regression (N changed from 5039 in Figure 6.3.3 to 5026 in Figure 6.3.5). For this particular example, the removal of unwanted data points didn’t affect the slope and intercept, but R^2 do improve from 0.76 to 0.78.

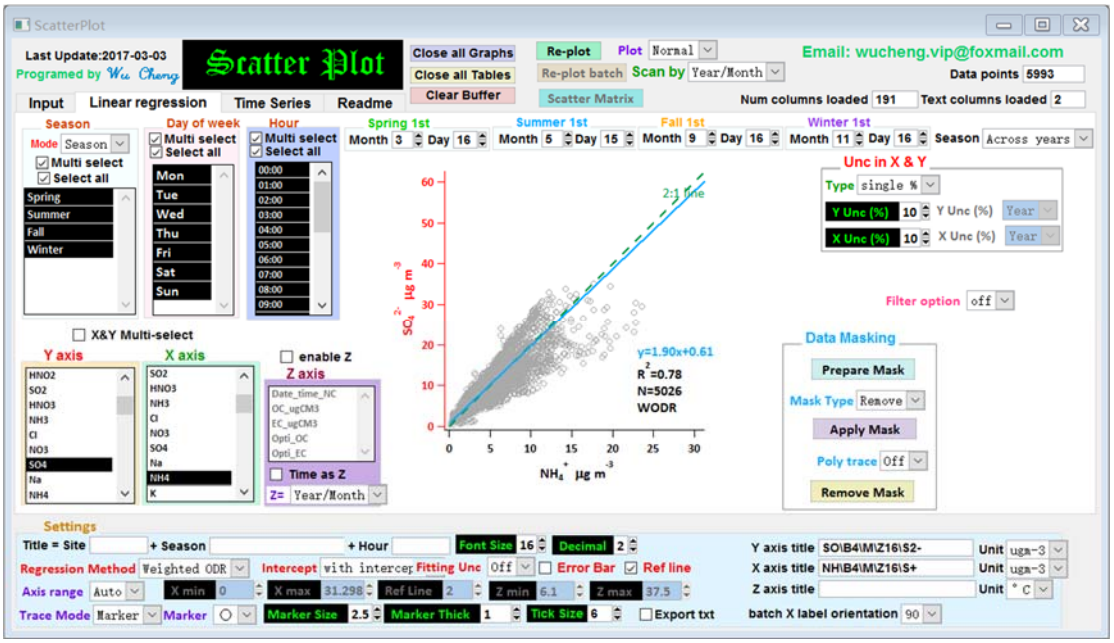


Figure 6.3.5 An example of choosing “Highlighted” in "Mask type".

Following is an example of choosing “On” in "Trace type”, then click "Apply Mask". Unwanted data points are removed and polygon is shown in dash line. Data points inside the polygon are excluded for regression (N changed from 5039 in Figure 6.3.3 to 4727 in Figure 6.3.6). For this particular example, the removal of unwanted data points result in changed slope (1.90->1.98) and intercept (0.61->0.52).

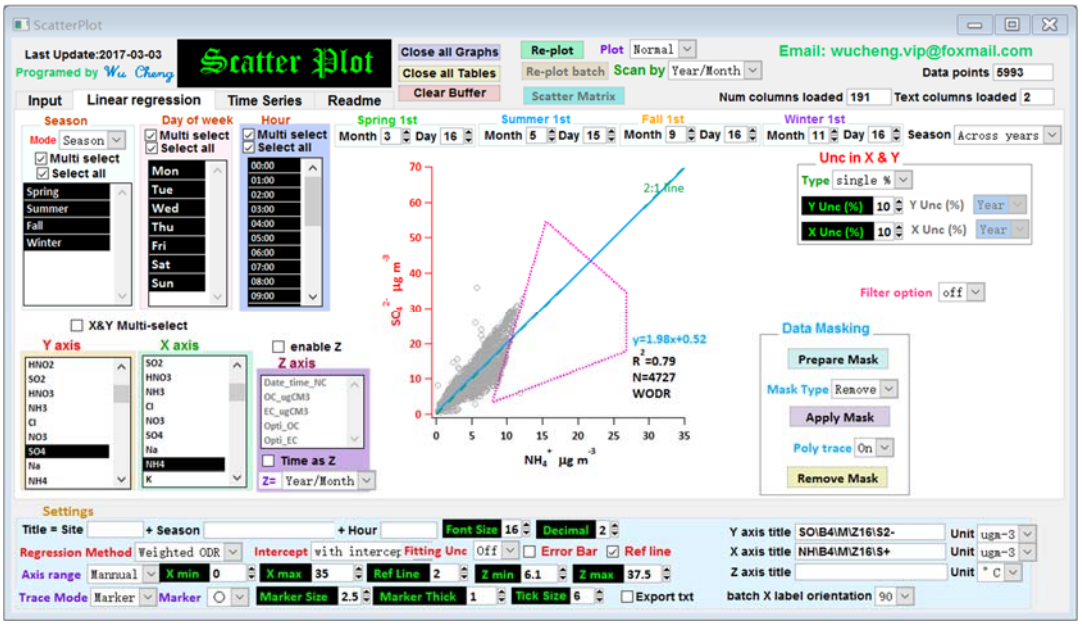


Figure 6.3.6 An example of choosing “On” in "Trace type".

To reset data masking, click “Remove Mask”, then all data masking will be wiped as shown below.

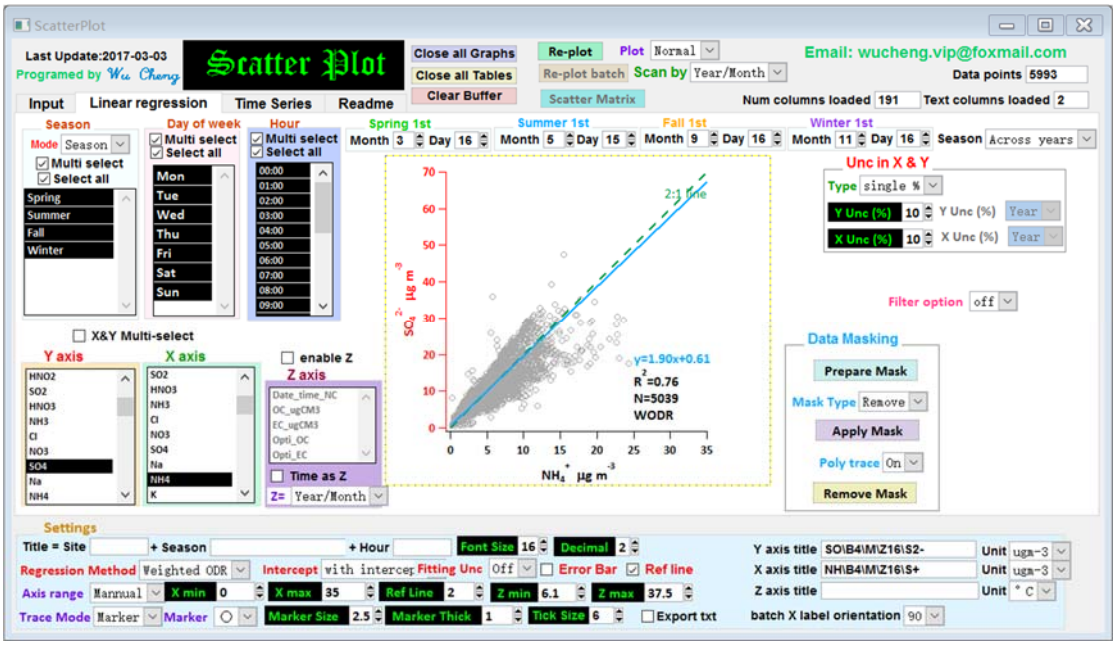


Figure 6.3.7 An example after “Remove Mask” applied.

6.4 Multiply selection in X&Y

Sometimes X and Y are not one to one variables, but multiply to one or multiply to multiply. Multiply selection in X&Y function allows user to use multiply variables in both X and Y via their sum. Multiply waves in X and Y can be selected using "shift" key with cursor. The sum of selected waves in X and Y are used for regression. Following is an example of QA/QC for ion chromatograph data of aerosols. Sum of sulfate and nitrate is used as Y against ammonium as X to check the charge balance of ions.

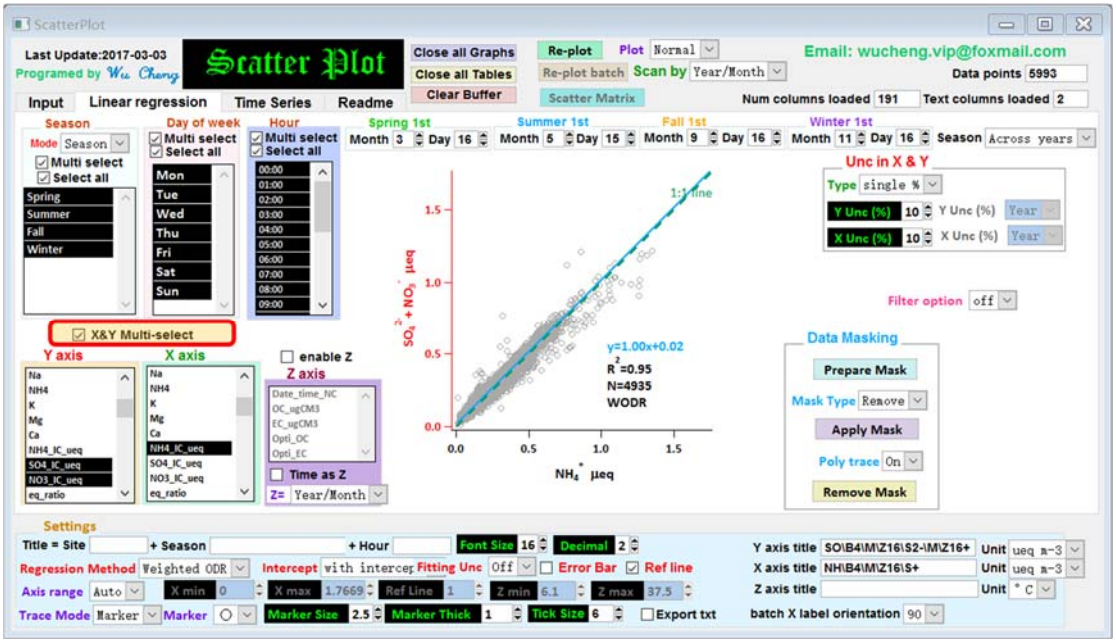


Figure 6.4.1 An example of multiply selection in X&Y.

6.5 Time as Z

Besides using user direct input variables as Z axis, derived variables including YSM (year season month), Dow (day of week) and Hour (0:00~23:00) can be used as Z.

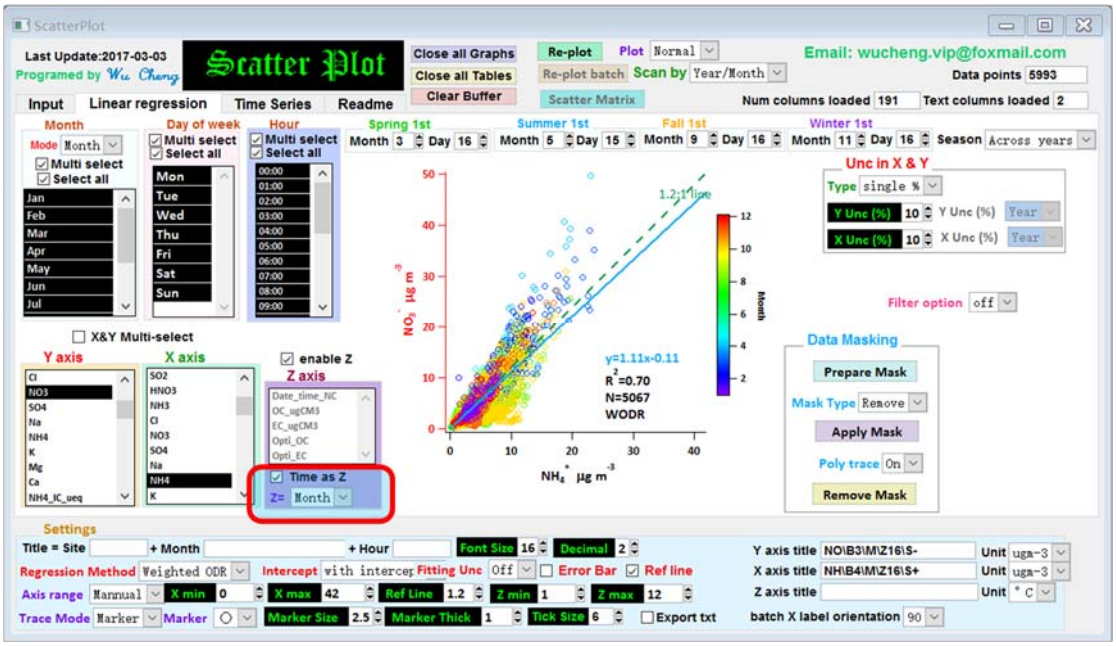


Figure 6.5.1 An example of using month as Z axis color coding.

6.6 Batch plotting

When plot option is other than “normal”, then batch plotting is activated. Batch plotting can be done on three time aspects (Scan by): Year/season/month, day of week, hour, which is corresponding to the three list boxes for data grouping by time.

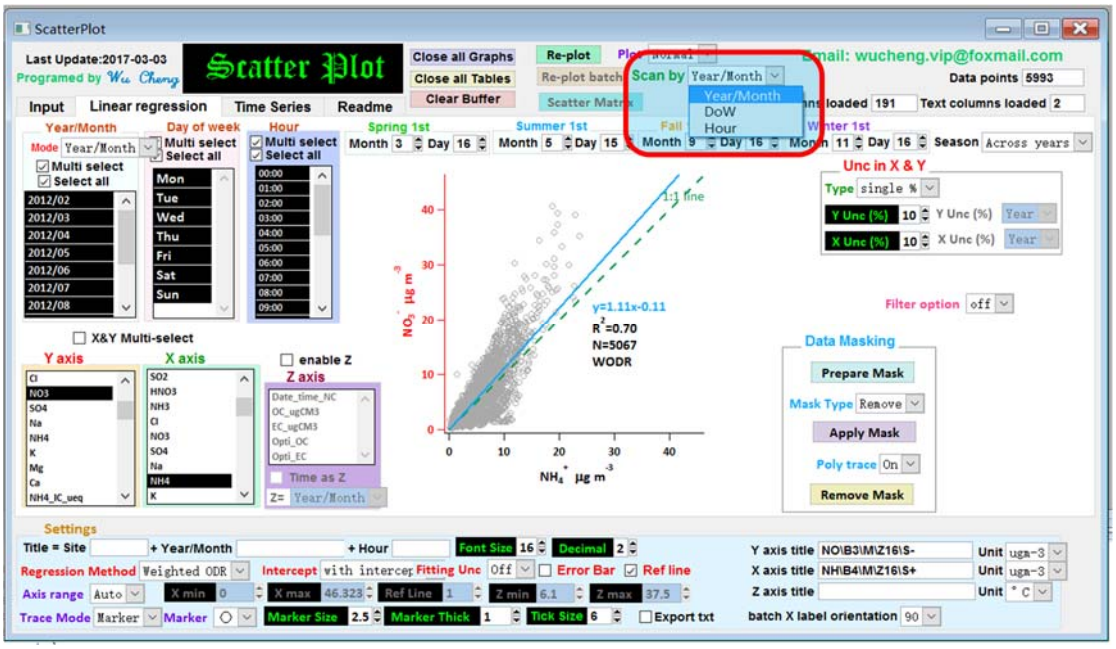


Figure 6.6.1 Settings of batch plotting by time

The fourth way is scan by text markers. When data filter is activated using “Text by list”, a 4th option will show up in the “Scan by” pop-up menu.

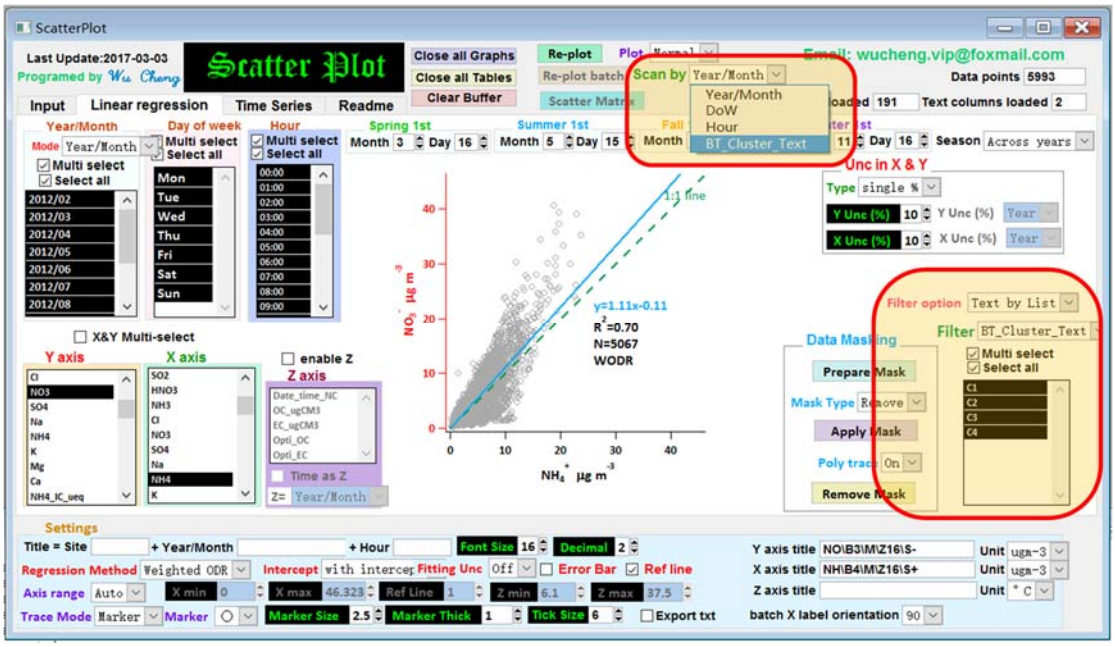
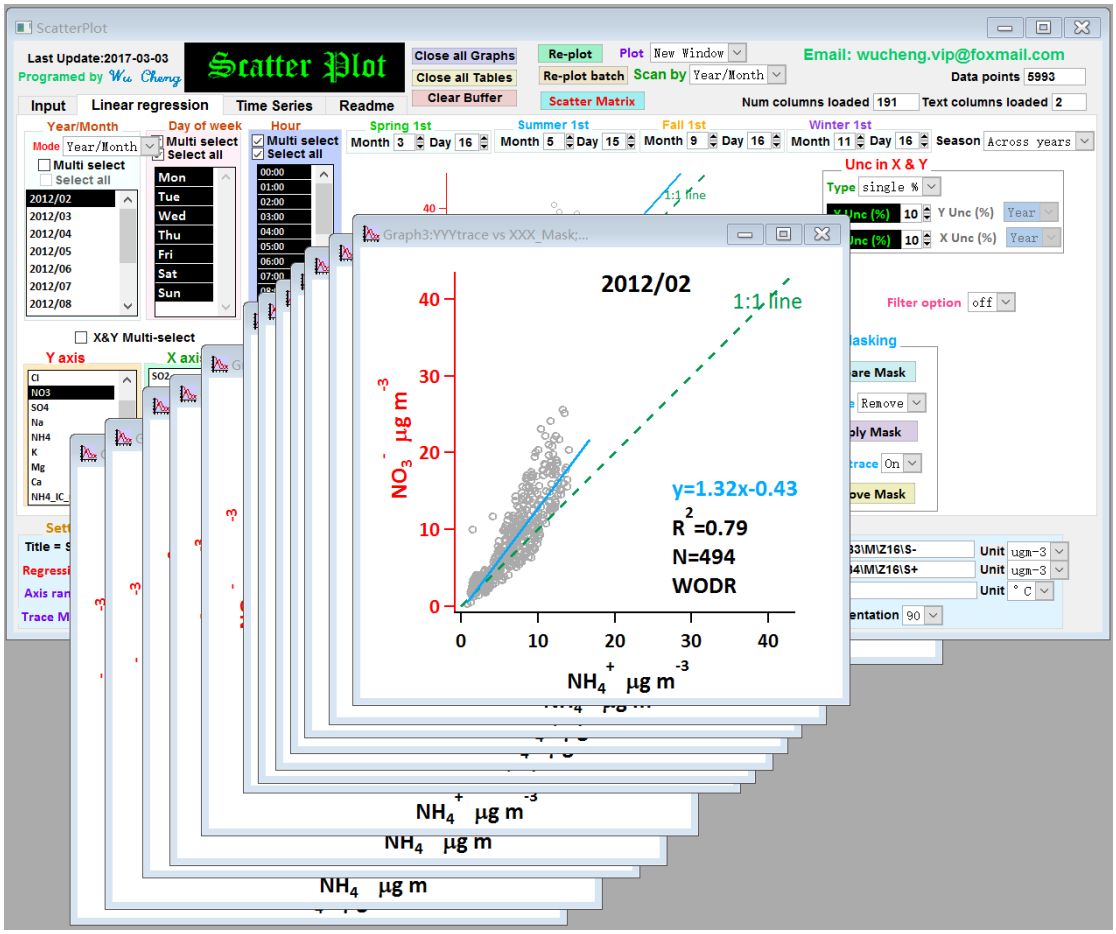


Figure 6.6.2 Batch plotting by text filtering

Following is an example of implementation of batch plotting, scan by year/month (12 months). Besides individual scatter plots, a plot summarizing the variations of slope, intercept and R^2 vs. year/month will also be given. As shown below, nitrate is sensitive to temperature (vaporization), as a result the slope during summer time (Jun-Sep) is much lower than the winter time (Dec-Feb).



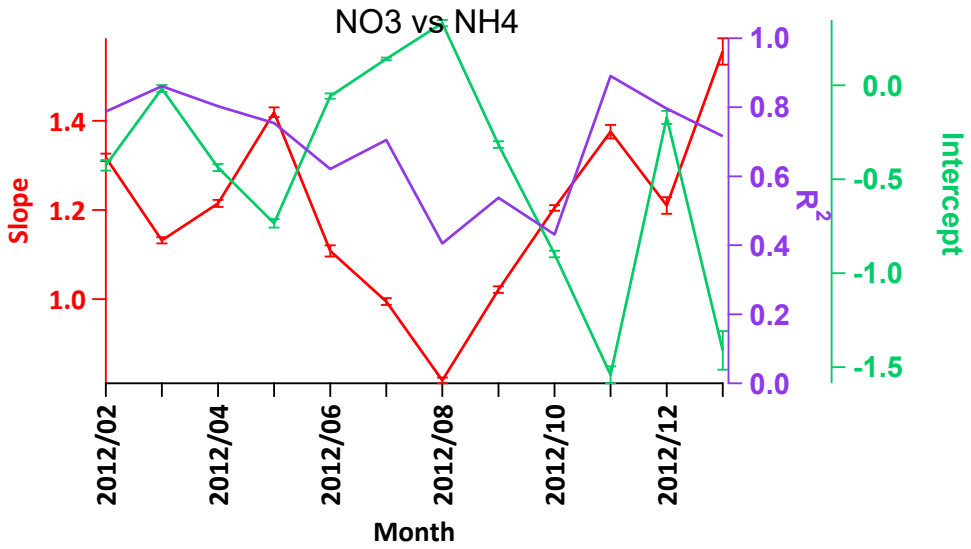


Figure 6.6.3 An example of batch plotting by year/month.

7 Tab “Multiply Y time series” Introduction

Multiply Y time series plot is commonly used for presenting temporal variations of various pollutants. As shown below, desired Y can be selected using the “Add” button.

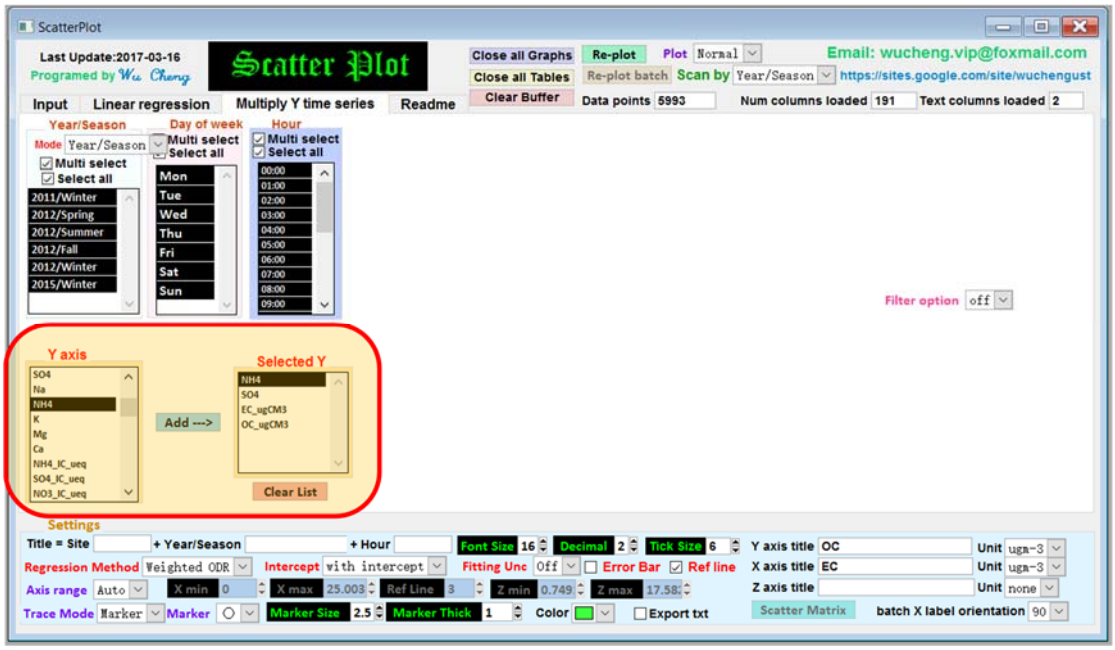


Figure 7.1 Example of selecting desired Y by the “Add” button.

“Plot option” should be set to new. Then click “Re-plot”, the graph will be generated in a new window. User can set the color and line shape, axis title in the new window. The main purpose is to save the time in setting the % portion in Y direction for individual axis.

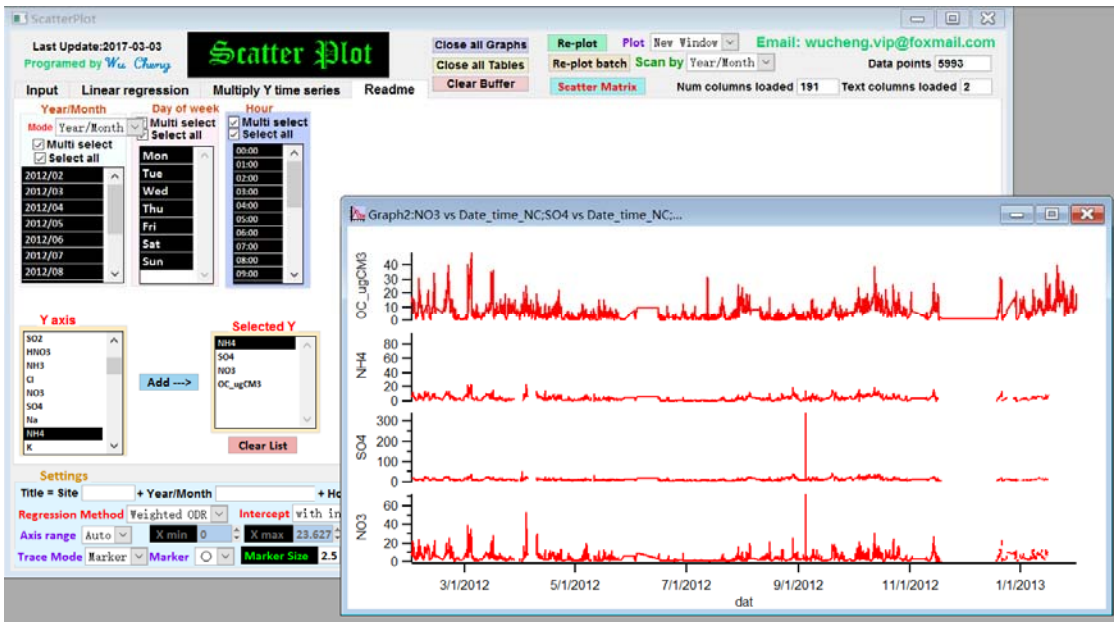


Figure 7.2 Example of multiply Y time series plot in a new window.