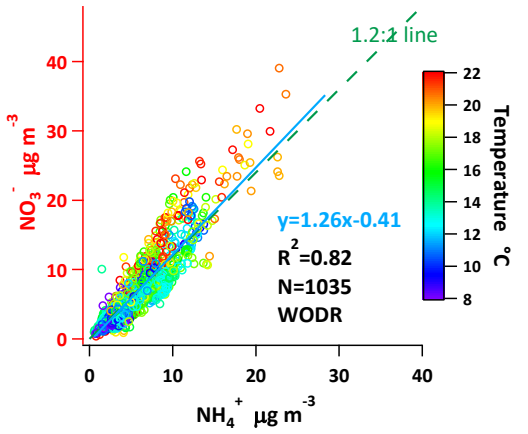
Scatter Plot



Scatter plot使用说明书

Manual for Scatter plot

吴晟

wucheng.vip@foxmail.com

2017-03-16

前言

散点图是一个方便的工具，可以最大限度地提高大气科学中数据可视化的效率。虽然有许多现有的通用数据可视化软件，但不能满足许多大气科学特定的研究目的，所以我开发自己的程序。 本程序包括WODR, Deming和York算法进行线性回归，这三种算法考虑了X和Y都包含不确定性（观测误差），对大气的应用而言更加客观地反映真实情况。它是基于Igor的，并且包含大量用于数据分析和图形绘图的有用功能，包括批量绘图，通过图形界面实现数据掩蔽，Z轴的颜色编码，根据数据或字符串进行过滤和分组。

关于程序的最新信息可以在我的网站上找到：

https://sites.google.com/site/wuchengust/

吴晟

2017-03-16

目录

1 关于数据结构的建议

如果数据的大小小于 100 万行，Excel 被建议用于存储数据。否则，建议使用.csv 文件。如果可能，将所有数据与同一时间线都放在一张表格上以实现最大化的效率，避免把它们都放在分离的表格，因为子集可以通过筛选取。建议的数据的结构如下图 1.1 所示。第一行是表头 （文本格式）。导入 Igor 后表头将成为 wave（Igor 中关于数列的概念，等同于 Excel 中的列） 的名称。表头中空格和其他非法字符 Igor 作为 wave 名称是不允许，将由"_"替换。数据分为三类：

1） 时间戳 （时间轴）

2） 数值数据 （如空气污染物的浓度）

3） 文本数据 （例如标签、 站点名称，后向轨迹聚类）

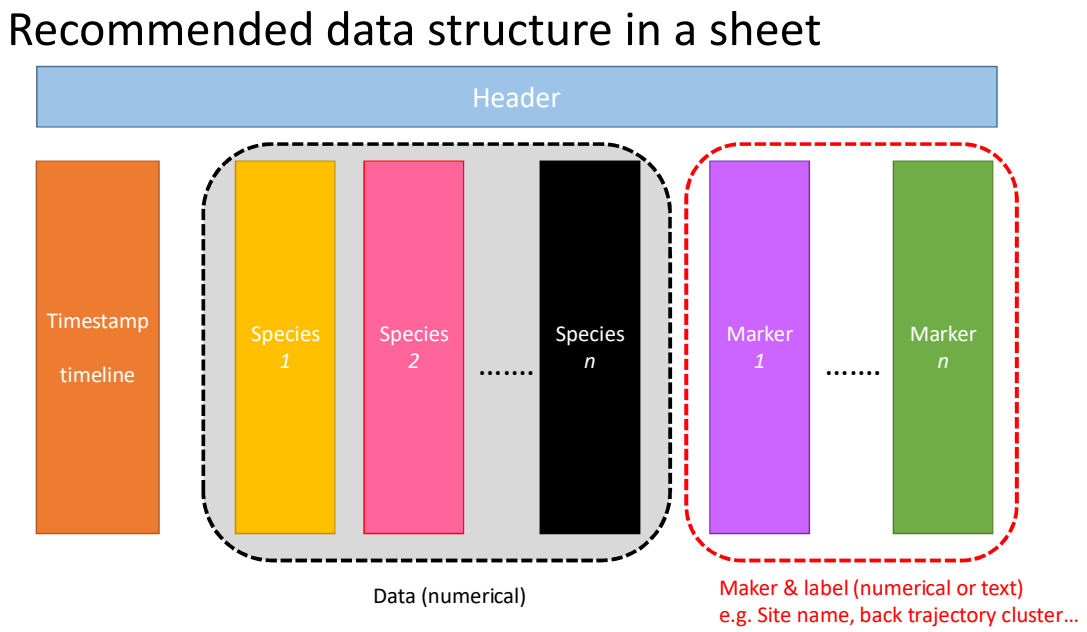


图 1.1 推荐在工作表中的数据结构

Excel 数据 （或.csv 文件）的一个实际例子如图 1.2 所示。应该指出的是，不同的数据列 （ wave ） 的顺序不一定是图 1.1 相同，可以混合使用三个类别，数据列的顺序没有限制。下例中，DateIndex属于时间轴，SampleID和Site属于Marker（文本数据），其余列属于数值列（污染物浓度值）

	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

图 1.2 Excel数据的一个实际例子（或.csv 文件）。

2 跟其他程序的总体比较

下表将本程序(散点图)与其他程序进行比较

软件	优势	不足之处
Excel	数据筛选	只能OLS线性回归, 不支持 Deming 回归 行数有限制(一百万行数据) 不能做数据遮掩 不支持Z轴颜色
SPSS	数据筛选	只能OLS线性回归, 不支持 Deming 回归 数据遮掩不能通过图形界面实现
Sigma Plot	支持Deming 回归	不支持数据筛选和数据遮掩

Origin	支持York回归 数据遮掩可以通过图形界面实现 支持Z轴颜色	不支持数据筛选
Scatter plot Igor program	支持OLS, Deming, Weighted orthogonal distance and York 回归。 支持数据筛选 数据遮掩可以通过图形界面实现 支持Z轴颜色 批量绘图	Igor普及率不及前四者

3 导入数据

3.1 在 MS excel 中的时间线示例

导入之前，数据在 excel 中，限制可以保存时间轴的格式如下所述。数据列中的时间轴必须遵循此格式"MM/DD/YY hh: mm"，位置必须是"英语 （美国）"，如图 3.1 所示。

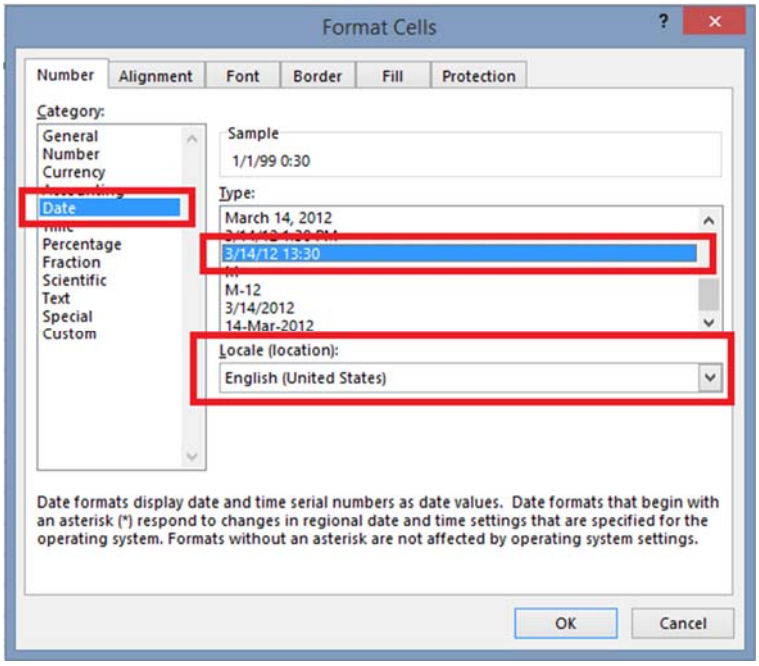
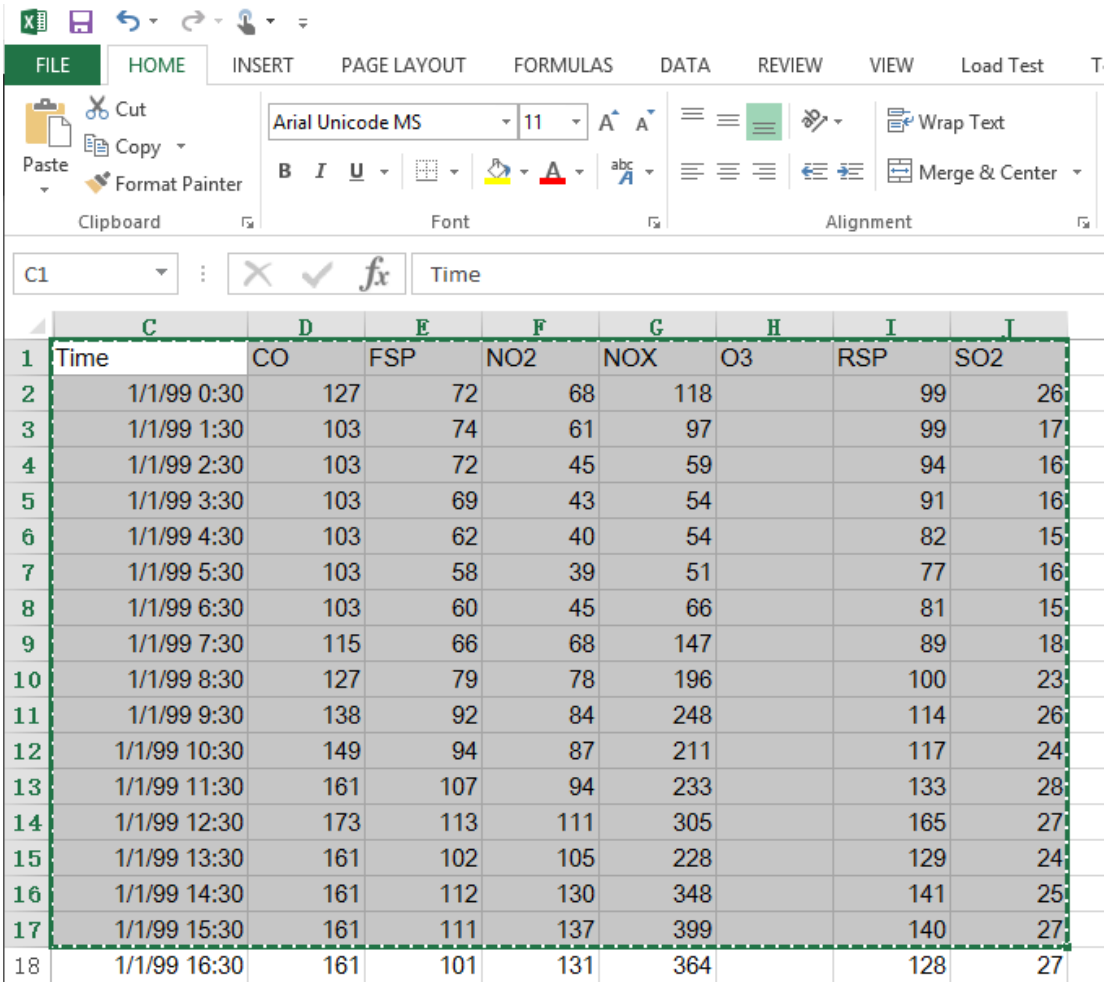


图3.1 时间轴列的MS Excel中的单元格格式配置

确保时间轴列的单元格格式与图2.1所示的完全相同，否则Iogr无法识别它

3.2 从Excel复制

数据可以通过从Excel Excel复制和粘贴导入，如图3.2所示。建议将时间轴放在第一列。



	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

图3.2　MS Excel中数据选择和复制（Ctrl + C）的示例。每列的表头将用作Igor中的wave名称。

3.3 将数据粘贴到Igor中

将光标放在左上角，将数据粘贴到Igor程序界面中的表格（高亮橙色区域），如图3.3.1所示

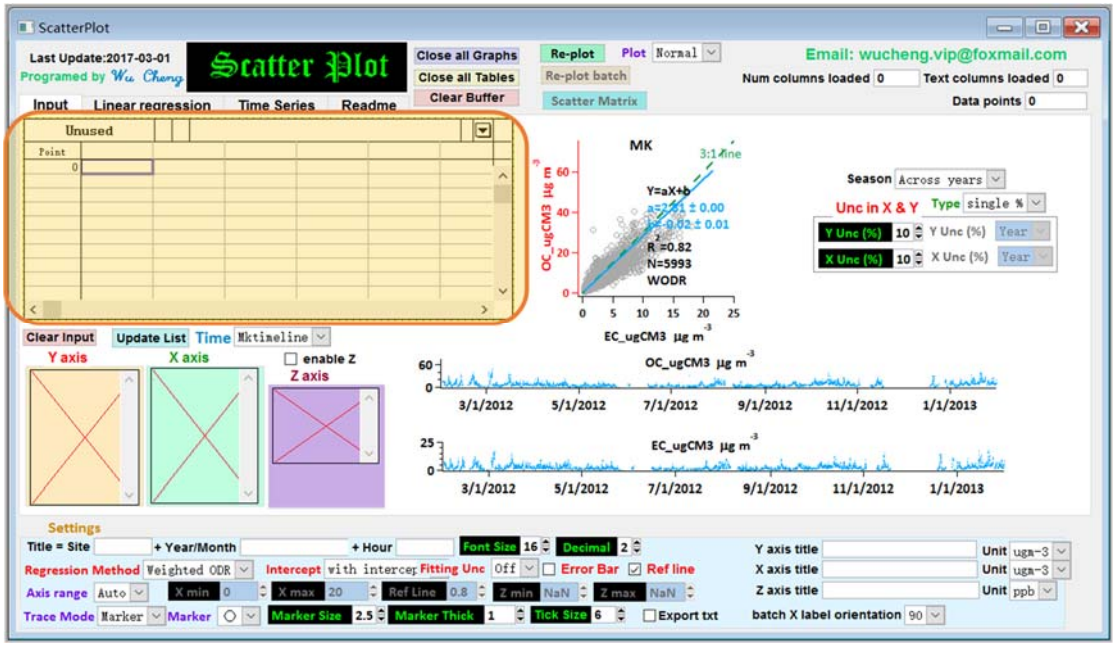


图3.3.1 粘贴数据之前Igor Pro中用户界面的示例.

应用粘贴（ctrl + V）后，数据将显示在表格区域中，确保时间线被Igor Pro正确识别。应当注意，数据点的索引从Igor中的0开始。

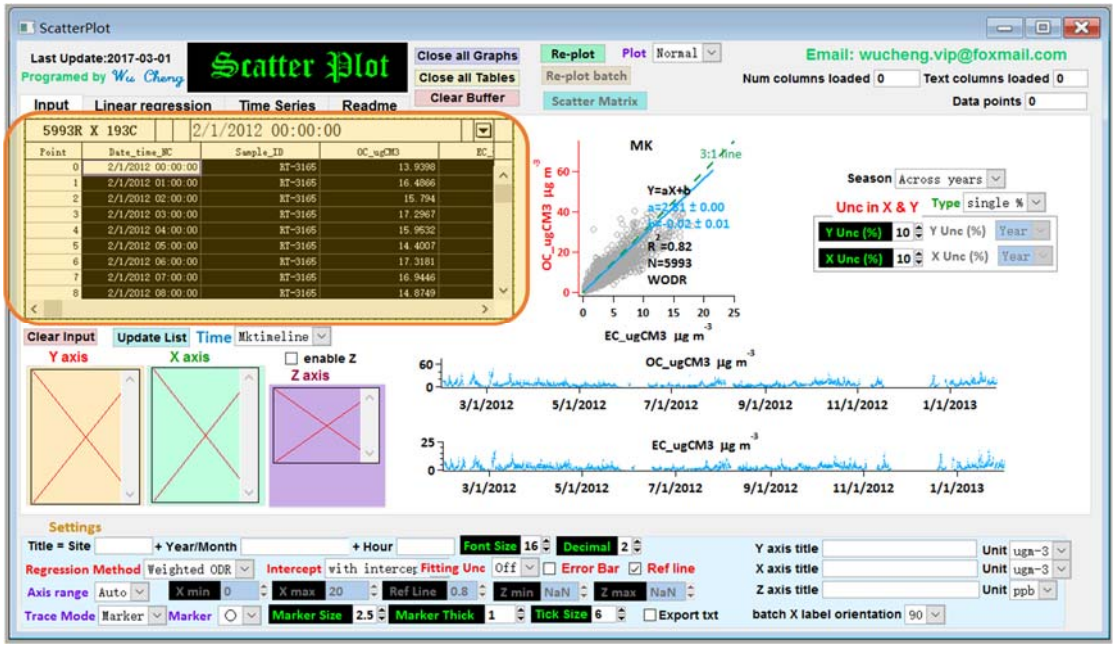


图3.3.2 粘贴数据后Igor Pro中的用户界面示例

3.4 更新列表

(a) 点击“Update List” 按钮(图3.4，高亮显示的区域a)

(b) 然后列表数值数据系列（在Excel中称为列和Igor Pro中的波）将被更新（图2.4.1，高亮显示区域b）。

(c) 加载数据的统计信息如图2.4.1高亮显示的区域c所示，包括数字列，文本列和数据点（行）的数量。

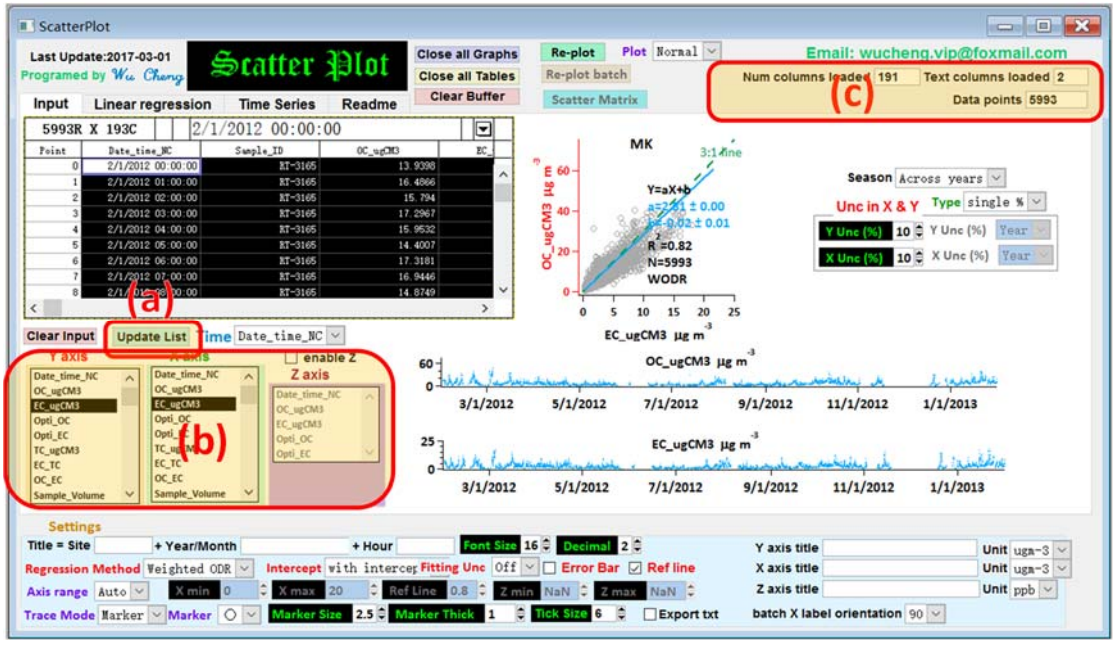


图3.4　Igor中的更新列表示例。

3.5 指定时间轴

下一步是告诉程序哪个列是时间戳。它可以通过使用弹出菜单来完成，如图3.5所示。

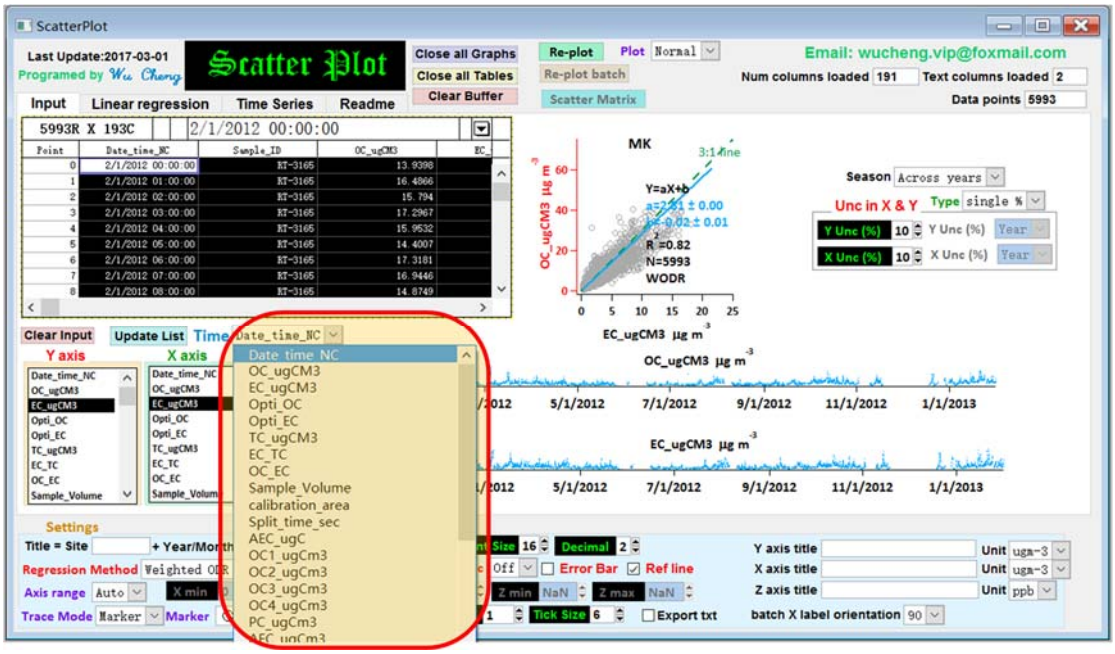


图3.5 在Scatter plot Igor程序中指定时间轴的示例

4 通用设置介绍

散点图Igor程序的一般设置如图4.1所示，其中包括，

Close all Graphs: 关闭新窗口中的所有图形

Close all Tables: 关闭新窗口中的所有表

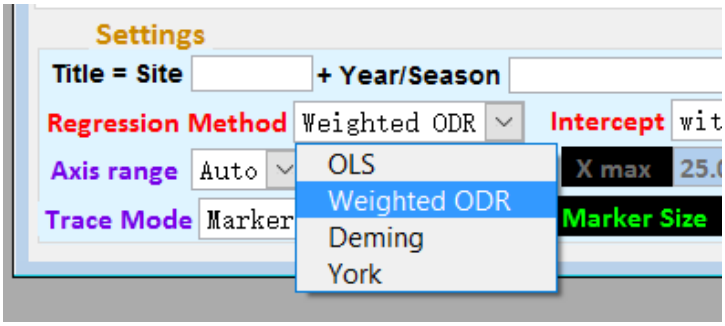
Clear Buffer: 删除批处理图的缓存数据，避免程序文件过大

Replot: 在当前选项卡中重绘图

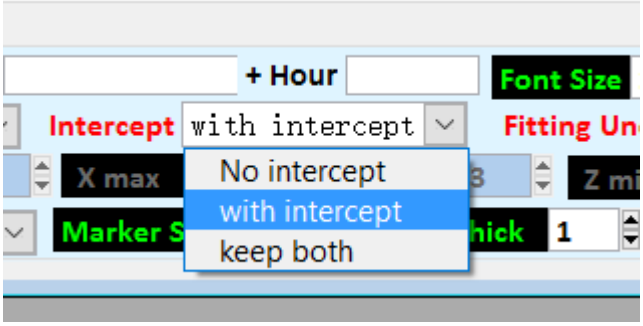
Plot option: Normal，只重绘当前选项卡; New window，也可以在一个新窗口中生成绘图，可以复制并粘贴到MS办公室; Export PNG，不仅在新窗口中生成绘图，还生成PNG文件; Export EMF，而不是PNG文件，EMF是一个矢量文件，可以无限放大。生成的PNG和EMF文件跟本Igor程序(.exp文件)存放在同一个文件夹。

Title: 绘图的标题，包含三个字段（采样站，年季月，小时和物种），如果它们留空,这些字段将在扫描时使用自动命名，否则会使用用户输入的字符。

Regression method: 最小二乘法 (OLS), 带权重的正交距离回归 (WODR), Deming 回归, York 回归. OLS仅考虑Y中的错误，而后三者考虑Y和X中的不确定性。



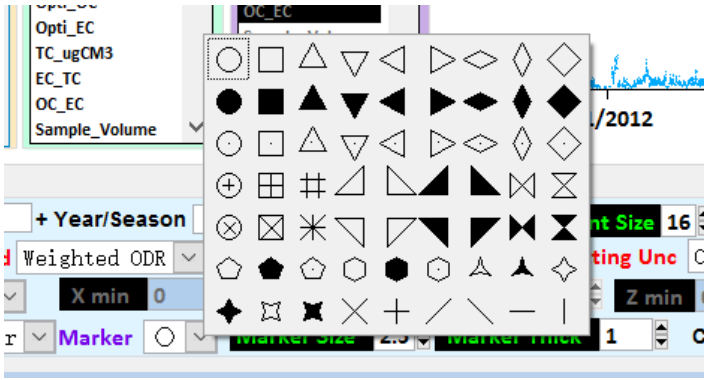
Intercept: 如果选择No intercept，则将通过原点进行回归（不适用于Deming和York回归）。如果选择“with intercept”，则所有回归方法都可用。如果选择“keep both”，则将执行有和无截距回归（不适用于Deming和York回归）。



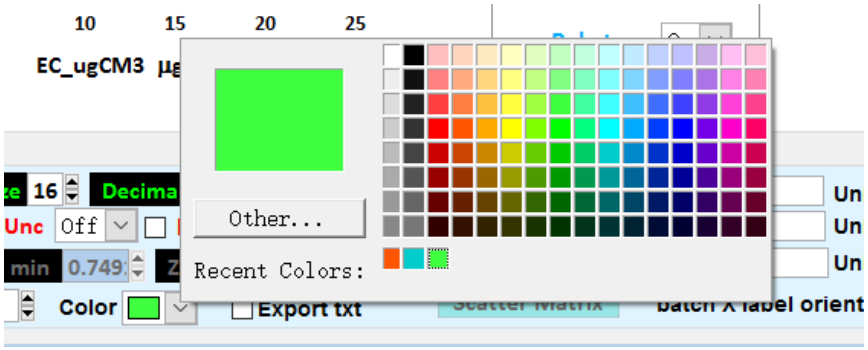
Axis range: 打开或关闭XY轴的自动标度

Unit: 为浓度轴使用预设单位

Decimal：显示几位小数点
Font Size：控制字体大小
Trace Mode:散点图上显示数据点的形式可以在点和标记之间进行选择。
Maker：选择数据点标记的符号



Marker Size：选择数据点标记的符号尺寸大小
Color：设置marker颜色（仅在Z 轴**关闭时有效**）



Ref Line：置参考虚线的Y：X的比值

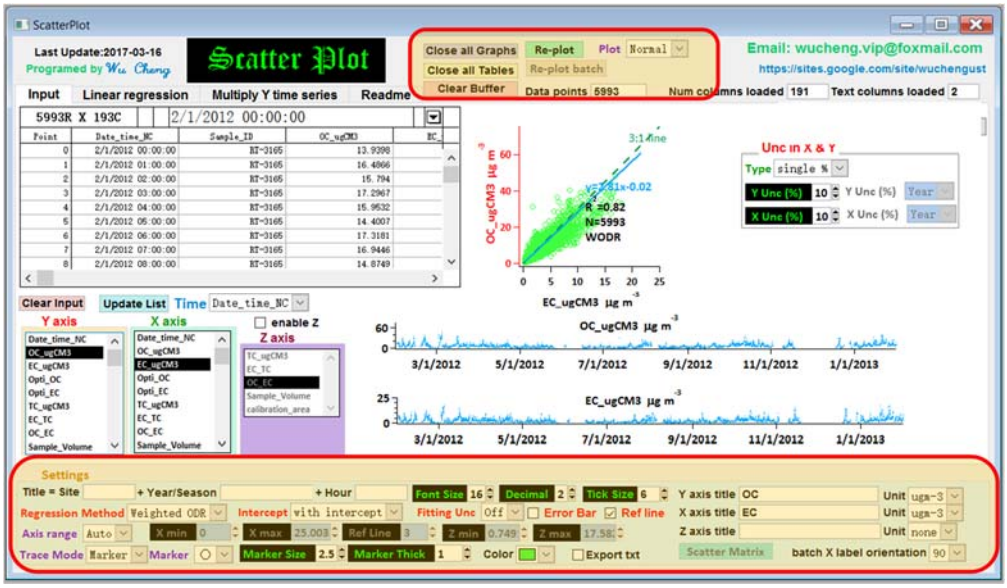


图4.1 Scatter plot Igor 程序中的常规设置示例。

5 分页“Input”简介

（a）用户可以选择哪些变量为X，Y和Z（Z可以打开或关闭）。 通过选择不同的X Y Z组合，散点图和两个时间序列图将立即更新，故而用户可以快速查看数据

（b）X&Y中的不确定性(误差)

WODR，Deming和York回归的不确定性设置。 有两种类型的输入可用，“Single%”意味着用户只需要提供一个数字来指示X和Y中的相对不确定性。Input Data”意味着用户需要为单个数据点提供误差权重(单独一列数据的形式，内容为标准差）。 用户需要使用弹出菜单来指定Y和X的相应的权重wave。

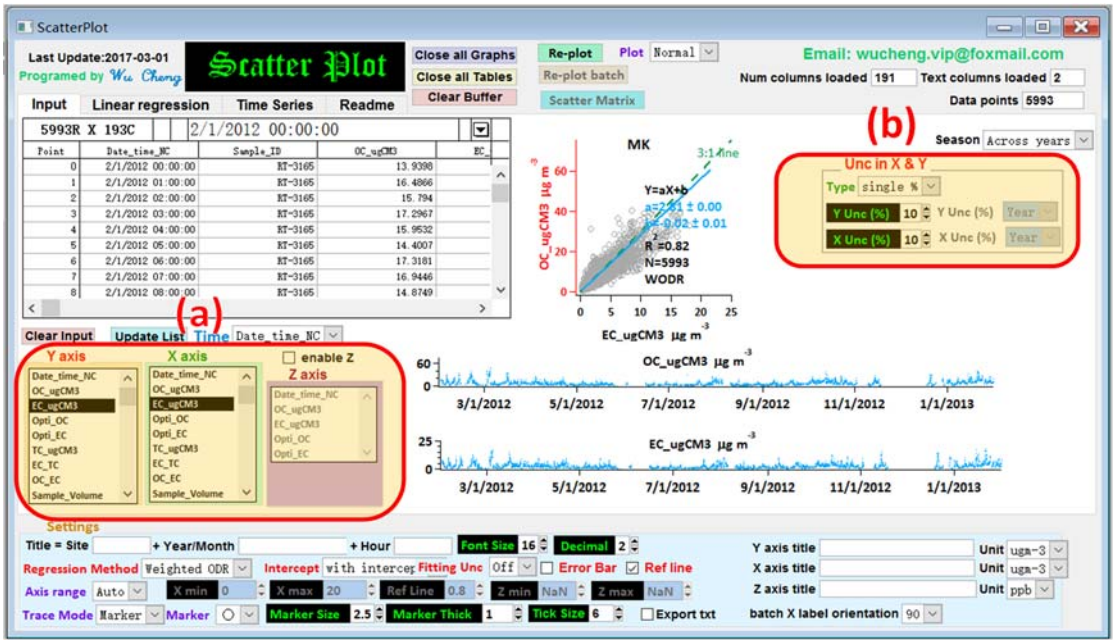
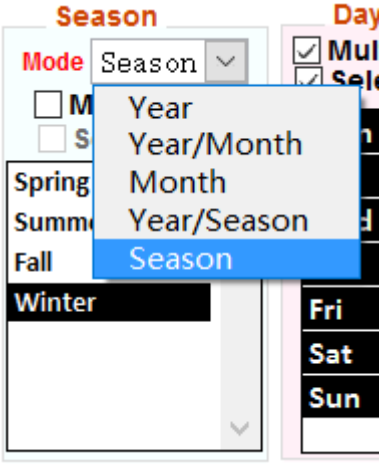


图5.1 Scatter plot 设置

6 分页“Linear regression ”简介:

6.1 数据按时间筛选

三种类型的时间标度用于数据过滤：YSM（年季月），Dow（星期几）和小时（0：00∽23：00）



(a) YSM可以进一步分为五种情况：年;年/月;月;年/季;季节。季节由突出显示的区域（d）定义，使用每个季节的第一天作为分野。有两个选择是可用的，跨年和同年。例如，如果选择跨年，1999年12月和2000年1月分组在一起作为1999年冬季。在选择期间可以使用shift键进行多项选择。

(b) Dow （星期几）。在选择期间可以使用shift键进行多项选择。

(c) Hour （0：00∽23：00）。在选择期间可以使用shift键进行乘法选择。

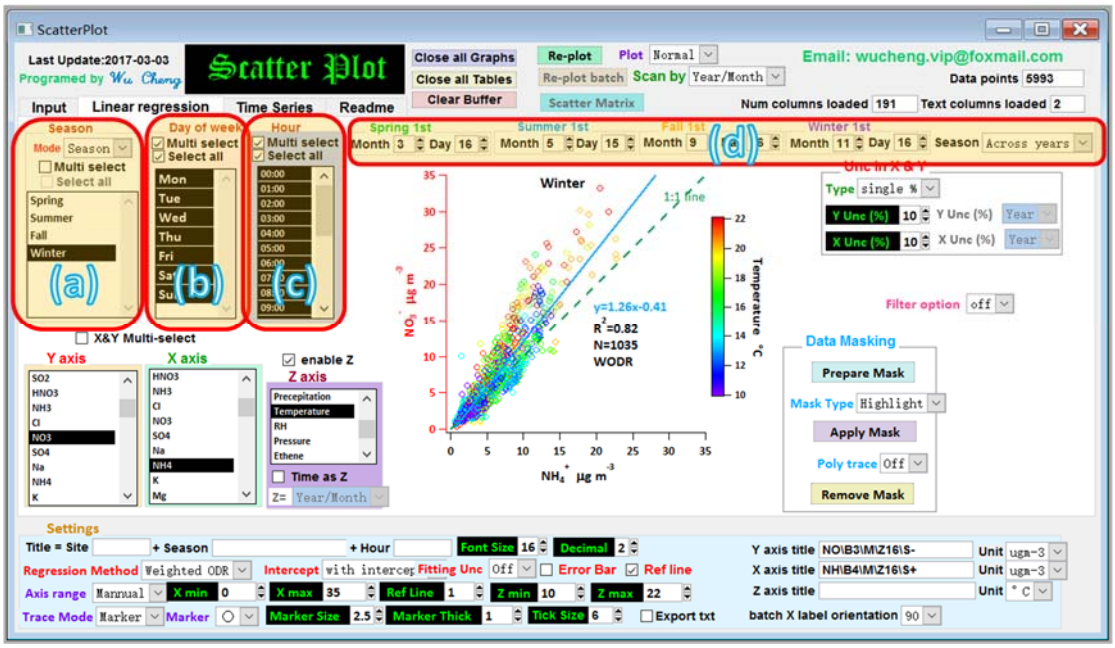


图 6.1.1 时间筛选的列表框

6.2 用数据进行筛选

可以有三种类型的数据过滤：按列表的文本数据，按列表的数字数据，按范围的数字数据

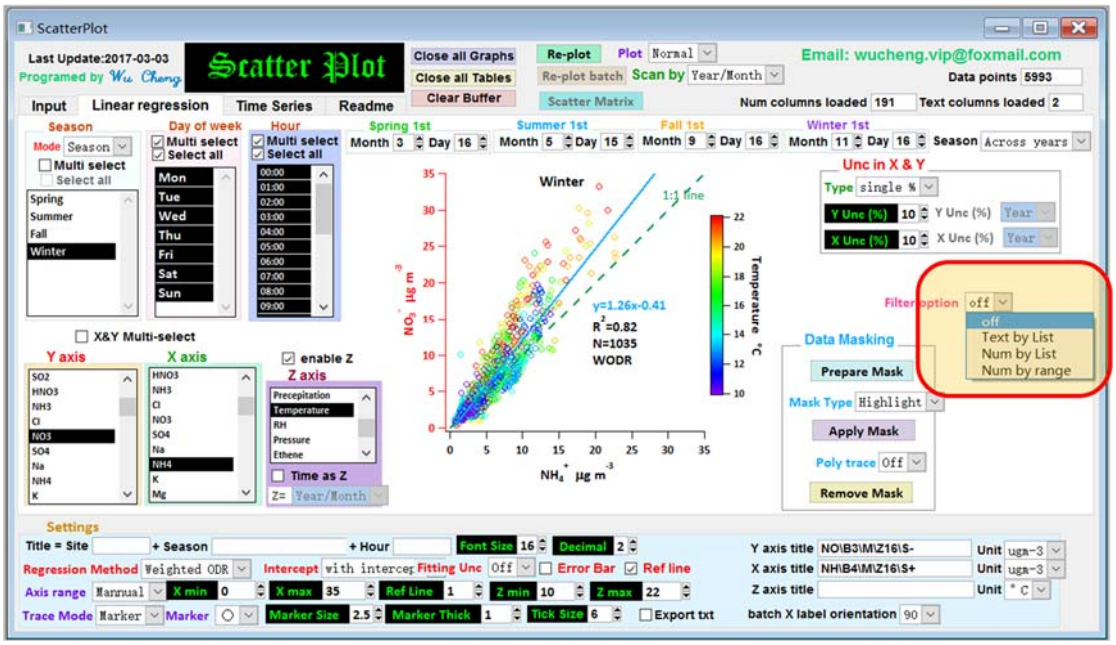


图 6.2.1 按数据过滤数据

(a) **Text by list**: 例如，提供包括C1∽C4的后面轨迹分组信息的列，使用该函数，绘制子集（例如，如下所示，仅C4）。可以进行多项选择。

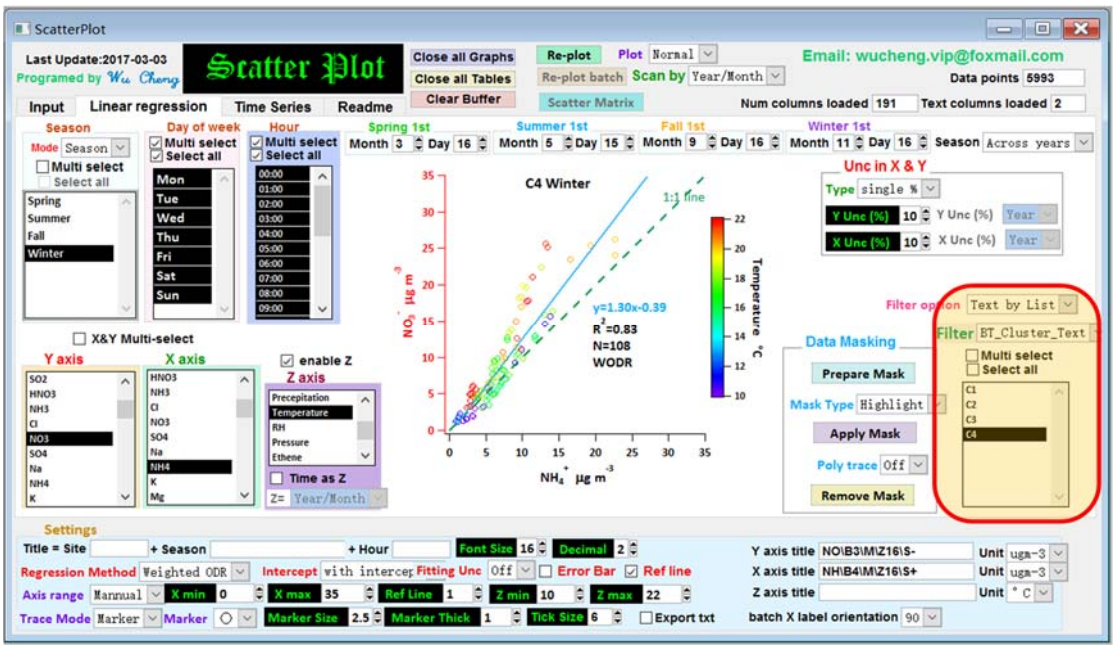


图6.2.2 按文本数据列表过滤数据 - Text by List

(b) **Num by List**: 具有数值的列可以用作数据分组的过滤器。当唯一值的数量远小于总计数时，它很有用。例如，数据按照如下所示的RH分组。

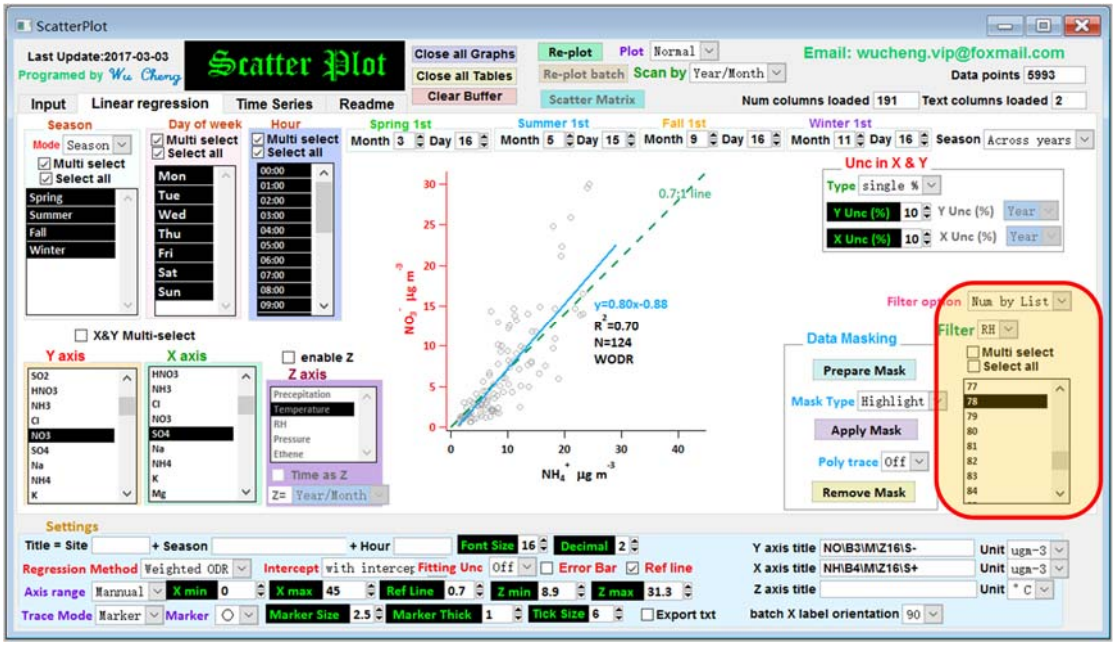


图 6.2.3 按数据过滤数据– Num by List

(c) **Num by range**: 范围（由min和max定义）可用于单个列以筛选子集。当使用多个列时，会取这些条件的交集。 以下是使用75 <RH <85的实例。

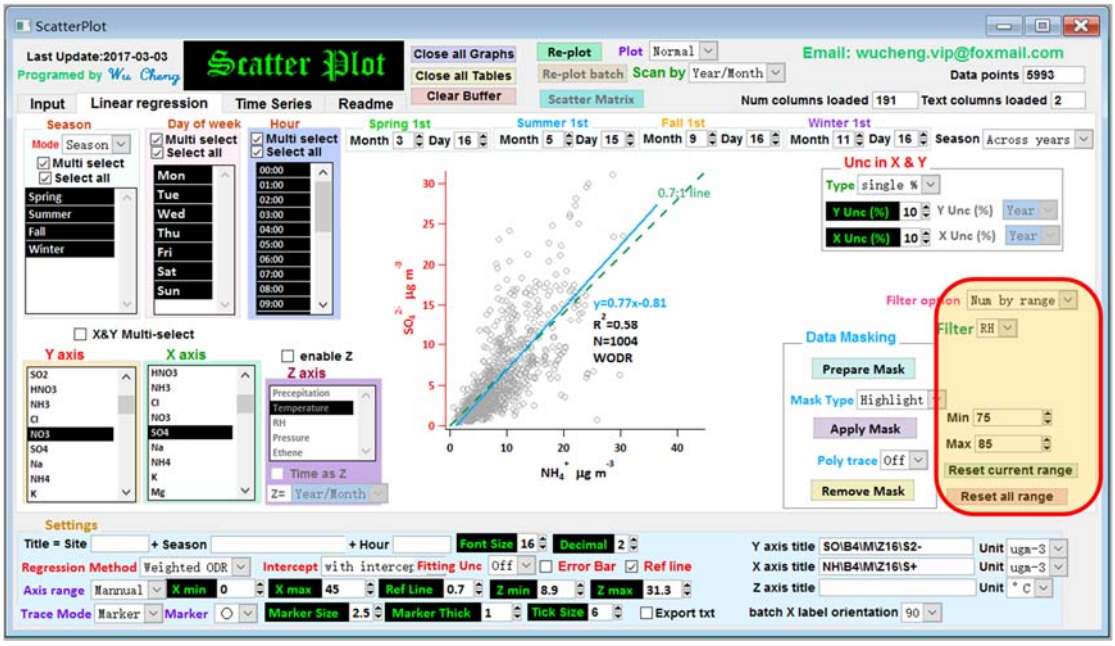


图 6.2.4 按数据范围过滤– Num by range

6.3 使用图形界面进行数据遮掩

数据遮掩功能可排除不需要的数据点再进行线性回归。 本程序可以使用图形用户界面直接实现，大大提高了易用性。

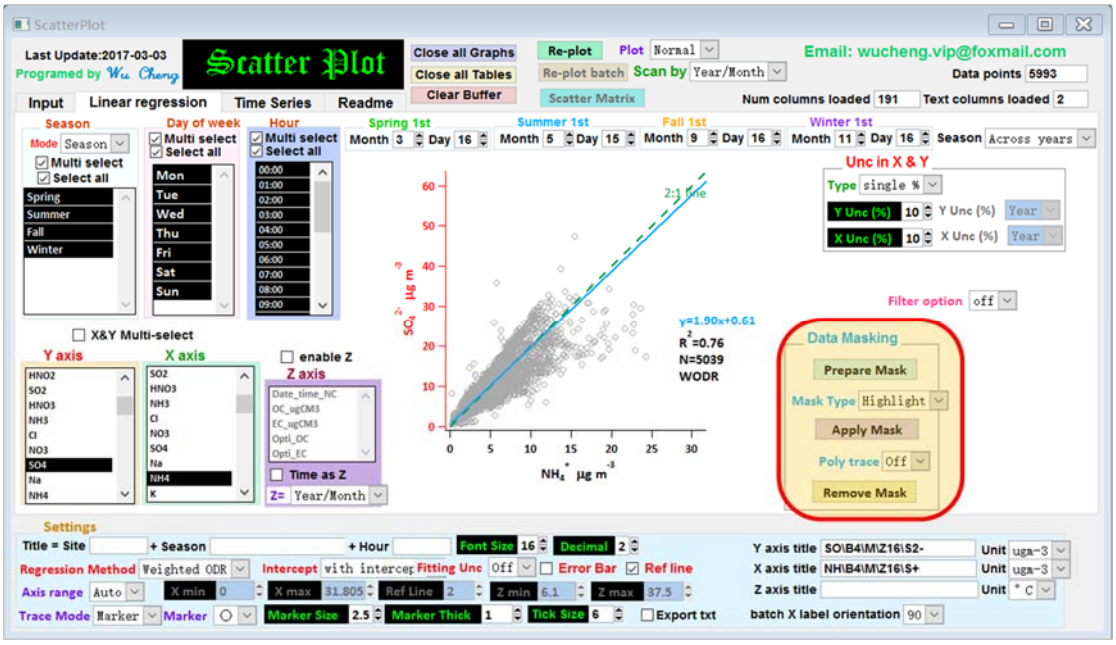


图 6.3.1 数据遮掩功能区一览

首先，点击"Prepare Mask" 按钮。 然后可以使用光标绘制多边形。 多边形由路径点定义。

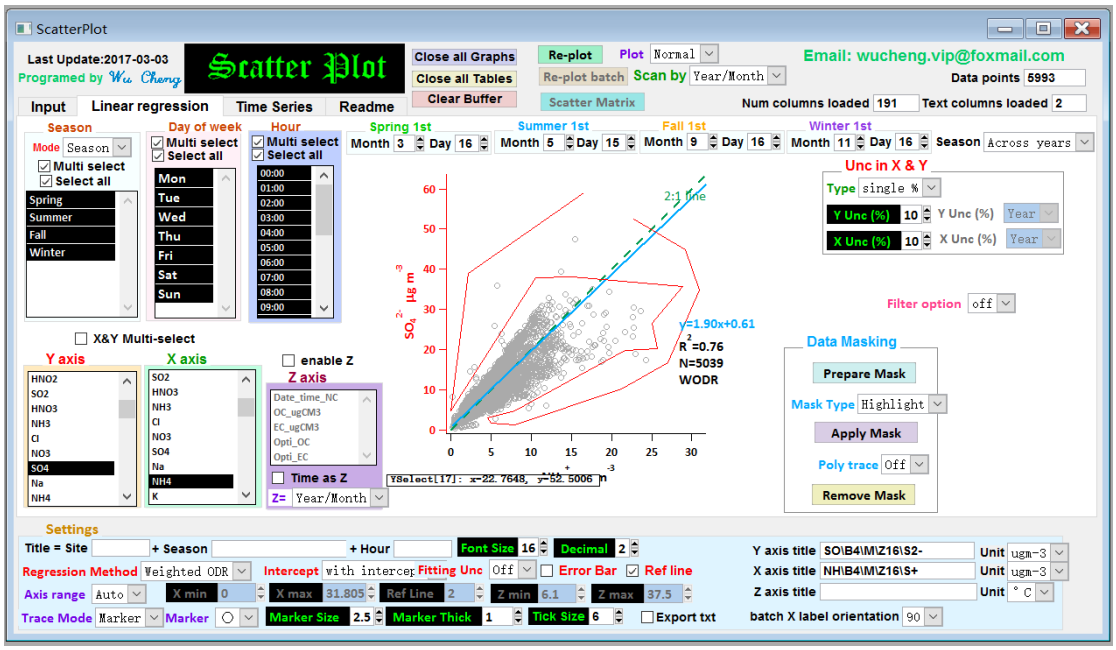


图 6.3.2 用光标开始进行多边形的绘制。

确保多边形闭合，如图6.3.3所示，每个点将以方形标记的形式显示。

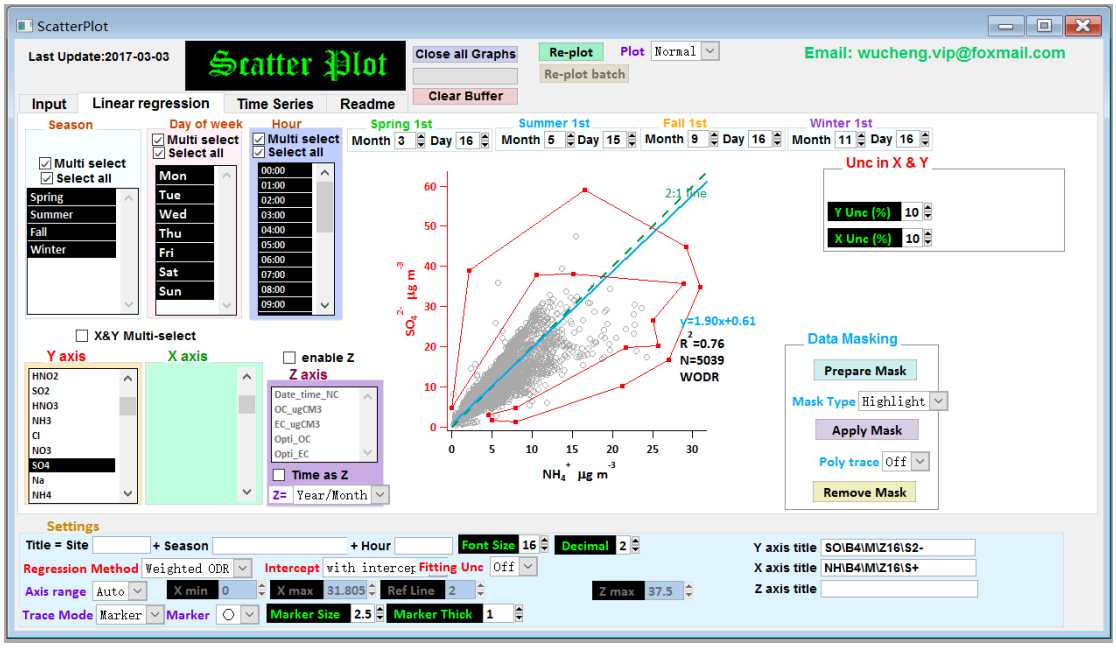


图6.3.3 闭合的多边形的示例

一旦多边形完成，选择“Mask type”。 以下是选择“Highlighted”的示例，然后单击“Apply Mask”按钮。 不需要的数据点会被标记为粉红色三角形。 这样就实现了排除多边形内的数据点进行回归（注意，数据点个数N从图6.3.3中的5039变为图6.3.4中的5026）。 对于这个特定的例子，去除不需要的数据点没有改变斜率和截距，但R^2确实从0.76提高到0.78。

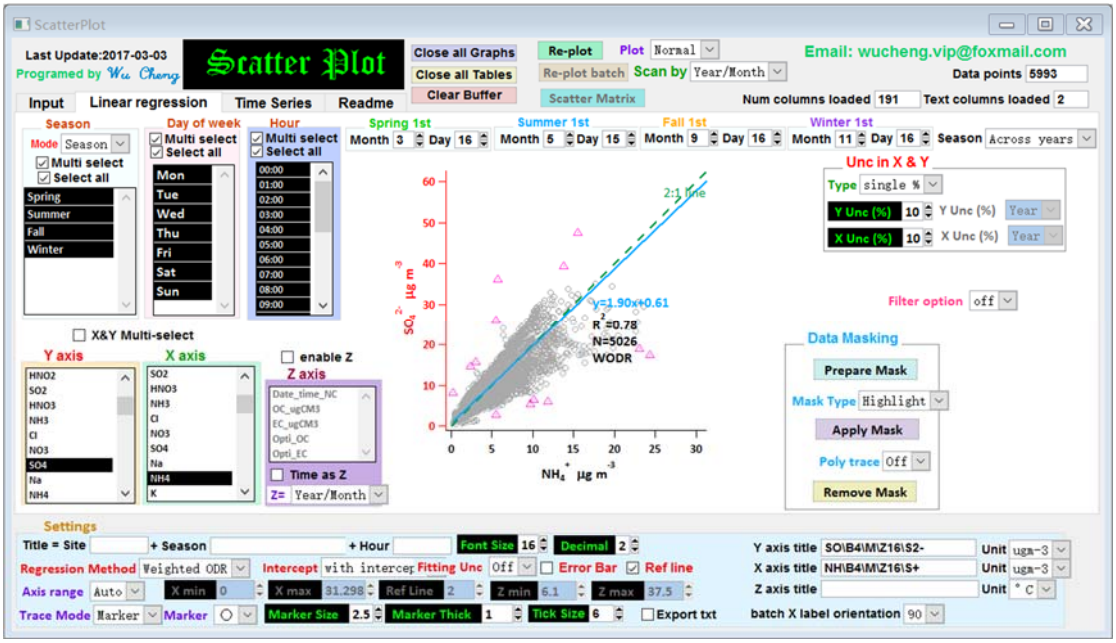


图6.3.4 在"Mask type"中选择“Highlighted”的示例。

以下是在“Mask type”中选择“Remove”的示例，然后单击“Apply Mask”按钮。 删除不需要的数据点。 这样就实现了排除多边形内的数据点进行回归（注意，数据点个数N从图6.3.3中的5039变为图6.3.5中的5026）。 对于这个特定的例子，去除不需要的数据点不影响斜率和截距，但R^2确实从0.76提高到0.78。

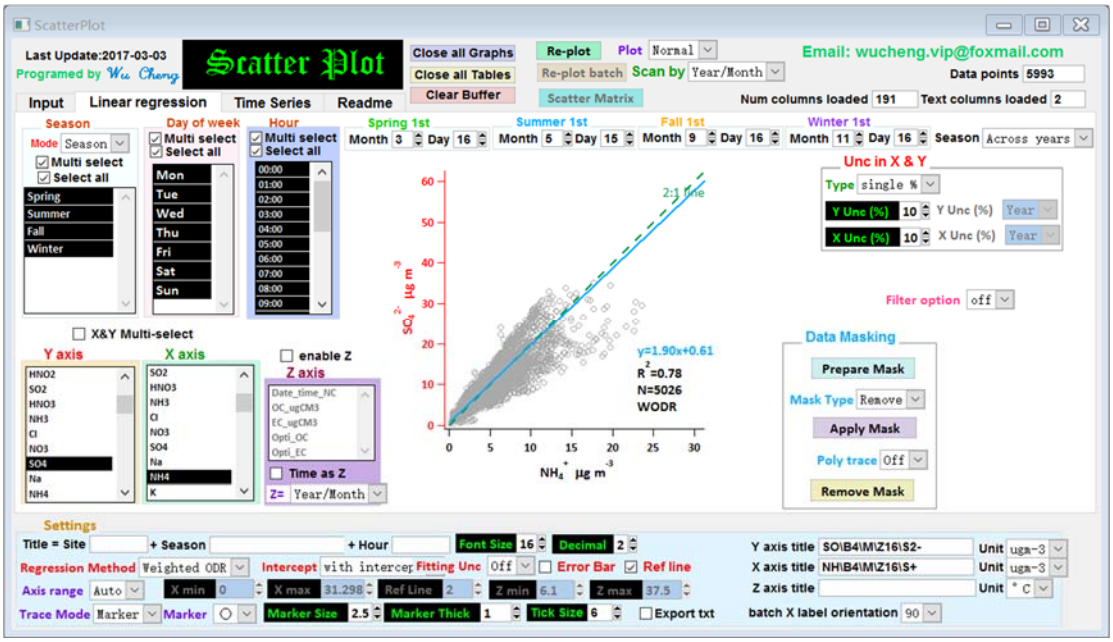


图 6.3.5 在"Mask type"中选择“Remove”的示例。

以下是在“Trace type”中选择“On”的示例，然后单击“Apply Mask”。除去不需要的数据点，并以虚线显示多边形。多边形内的数据点被排除回归（注意，数据点个数N从图6.3.3中的5039变为图6.3.6中的4727。 对于这个特定的例子，去除不需要的数据点改变了斜率（1.90-> 1.98）和截距（0.61-> 0.52）。

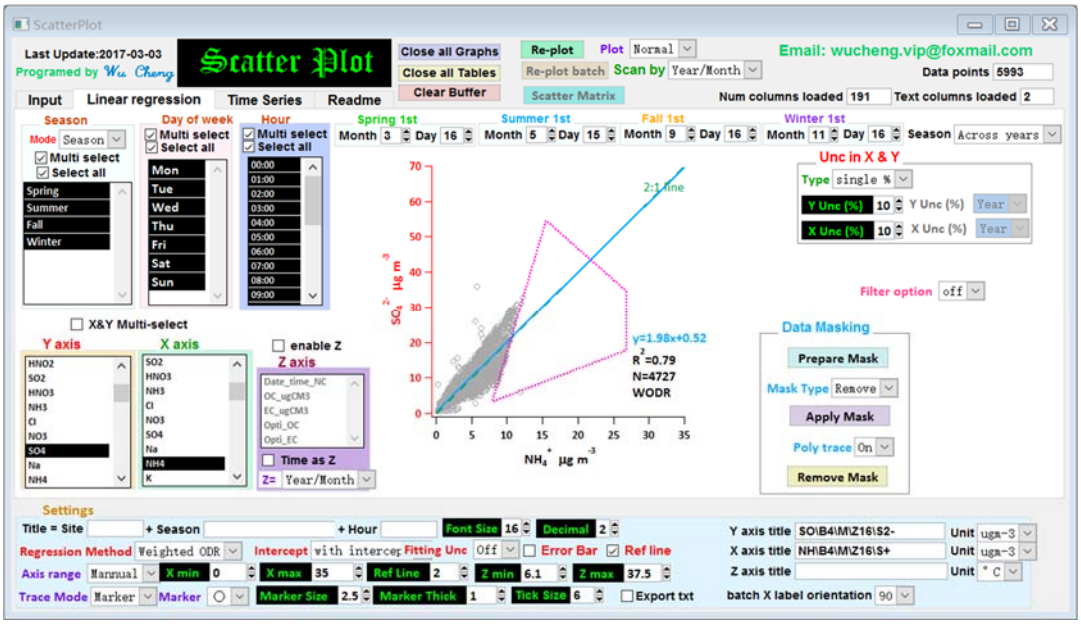


图 6.3.6 在"Trace type"中选择“On”的示例.

要重置数据遮掩，请单击“Remove Mask”，然后将擦除所有数据屏蔽，如下所示。

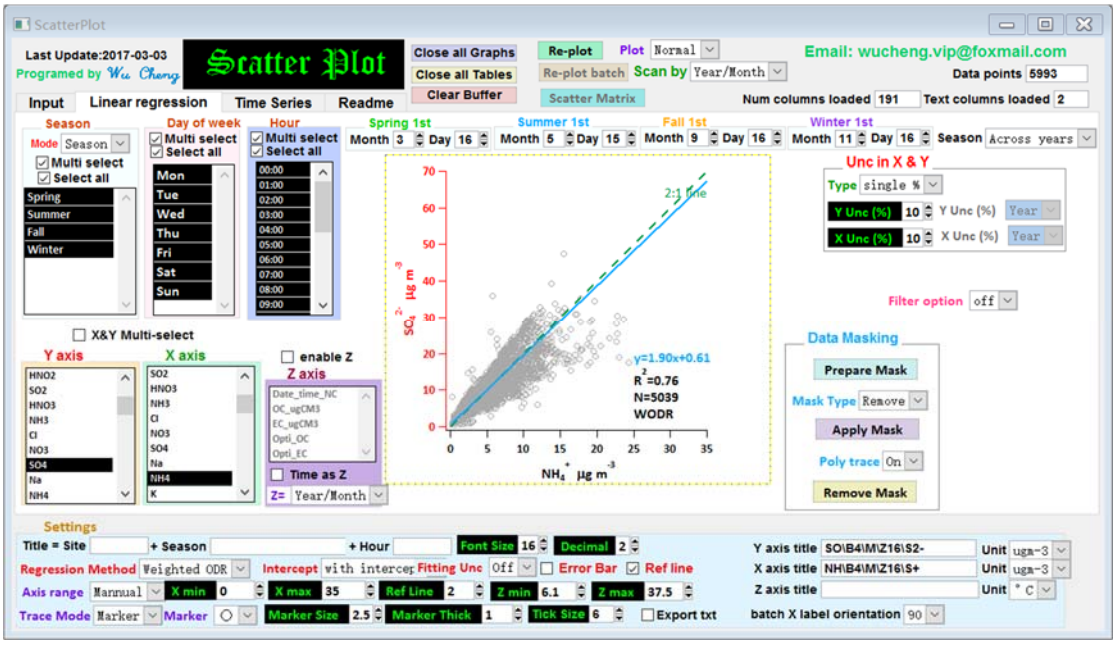


图 6.3.7 应用“Remove Mask”后的示例。

6.4 选择多个变量用作X&Y

有时X和Y不是一对一的变量，有可能是多对一或者多对多。 在X和Y变量中多项选择允许用户通过使用它们的加和用作X和Y。 可以使用“shift”键和光标选择X和Y中的多项变量（wave）。 X和Y中各自所选的总和将被用于线性回归。以下是气溶胶的离子色谱数据的QA / QC示例。 使用硫酸盐和硝酸盐的总和作为Y，将铵根离子作为X，以检查离子的电荷平衡。

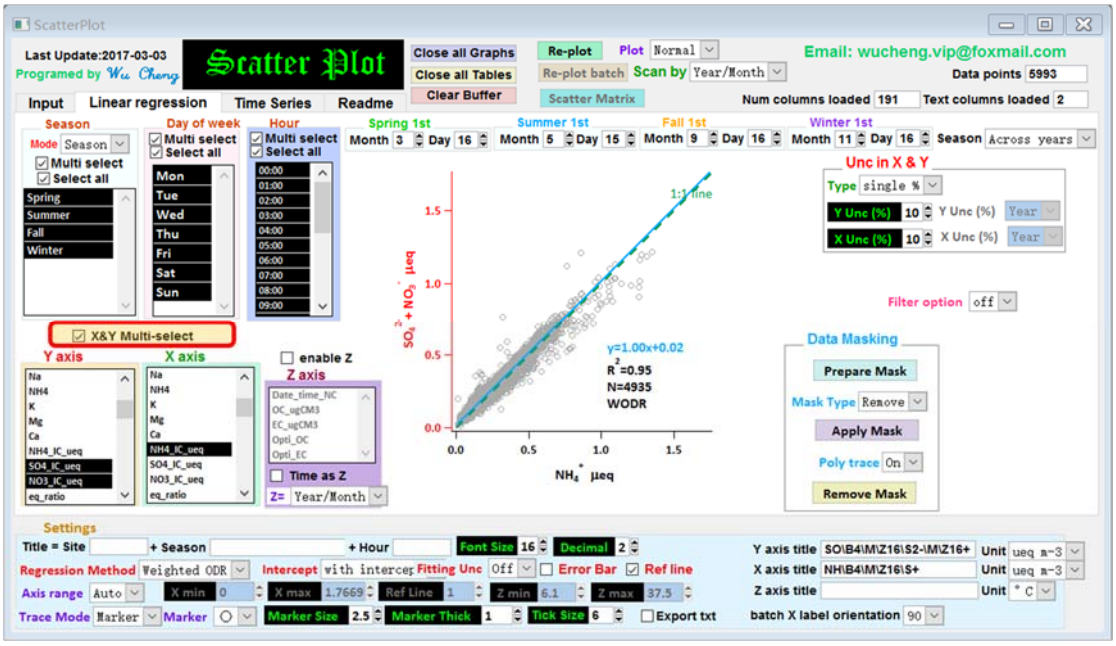


图 6.4.1 选择多个变量用作X&Y的示例.

6.5 时间变量作为Z轴

除了使用用户直接输入变量作为Z轴，包括YSM（年季节月），Dow（星期几）和小时（0：00～23：00）的派生变量可以用作Z。

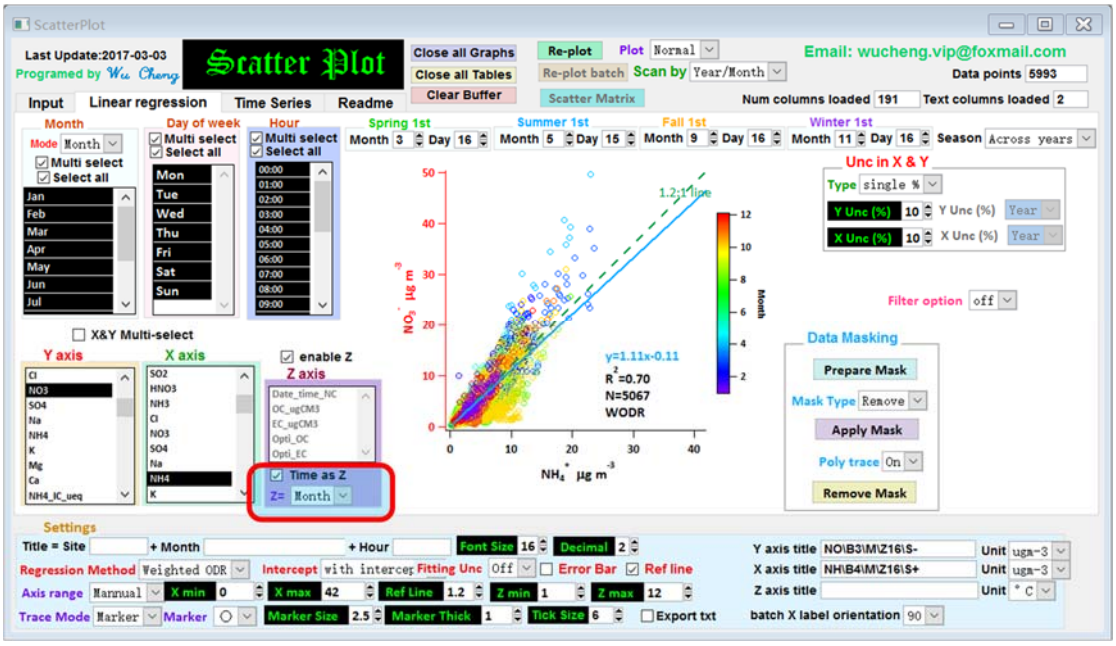


图 6.5.1 使用月作为Z轴颜色编码的示例。

6.6 批量绘图

当绘图选项不是“normal”时，则批处理绘图被激活。 批量绘图可以在三个时间维度（Scan by）进行：年/季/月，星期几，小时，这对应于按时间进行数据分组的三个维度的列表框。

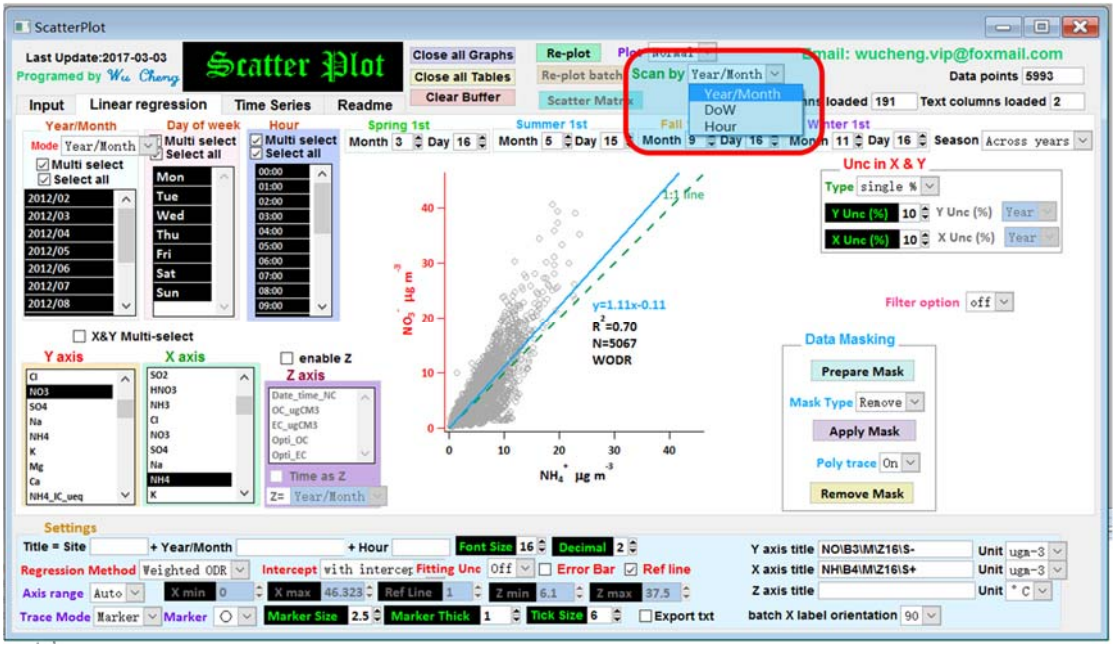


图 6.6.1 按时间维度进行批量绘图的设置

第四种方式是通过文本标记进行扫描。 当使用“Text by list”激活数据筛选器时，第4个选项将显示在“Scan by”弹出菜单中。

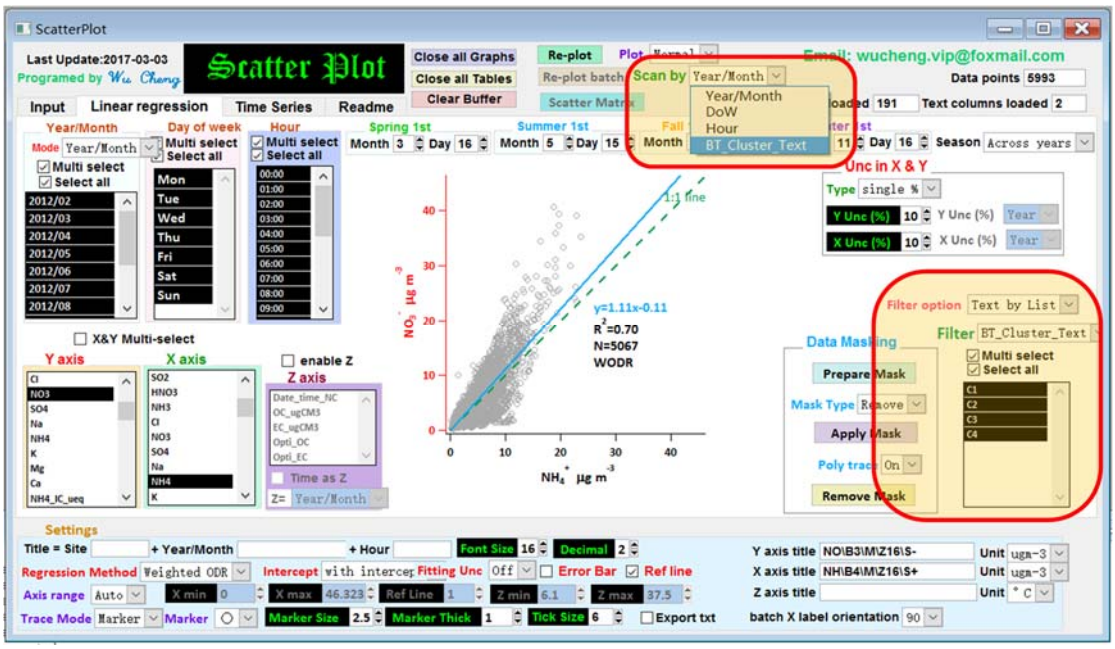
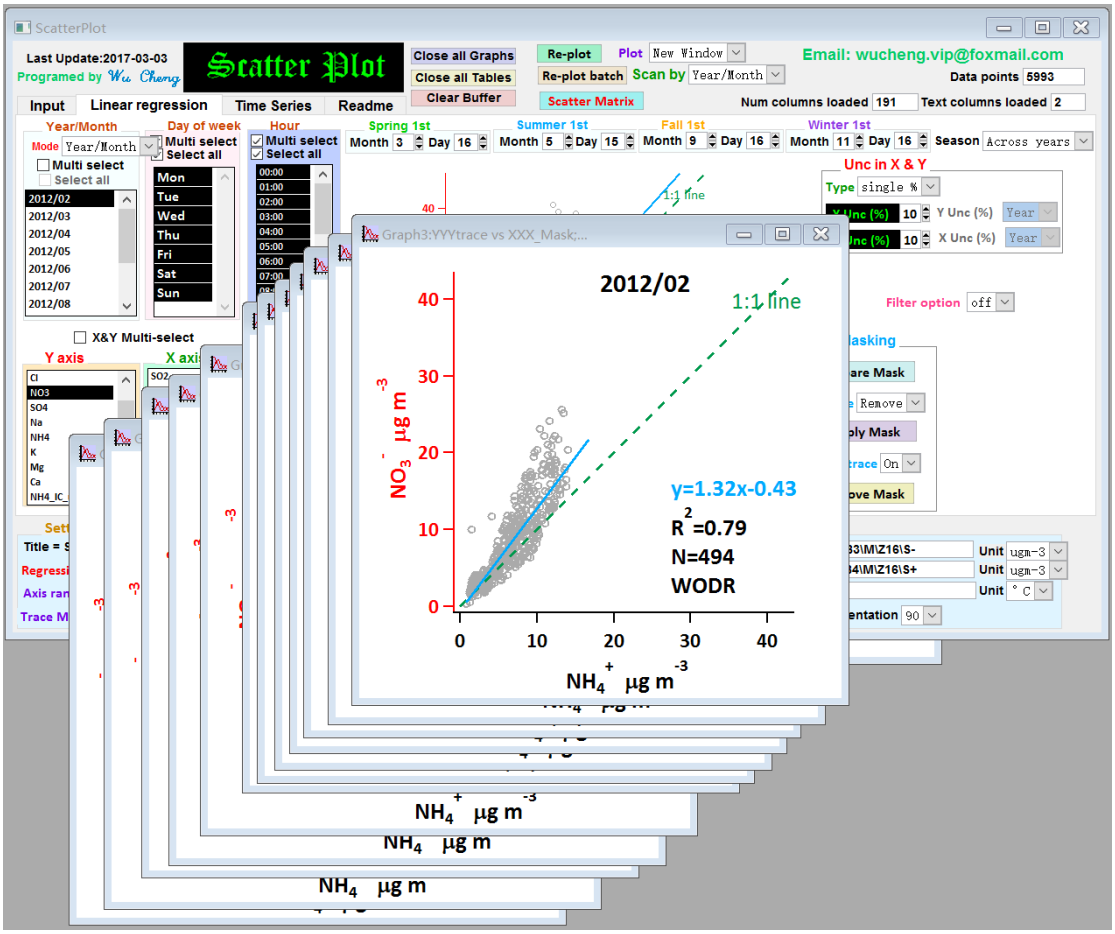


图6.6.2 通过文本标记进行批量绘图

以下是实施批量绘图的示例，按年/月（12个月）扫描。 除了单个散点图，还将给出总结斜率，截距和R^2随年/月的变化的图。 如下图所示，硝酸盐对温度（挥发）敏感，因此夏季（6月 - 9月）的斜率远低于冬季（12月 - 2月）。



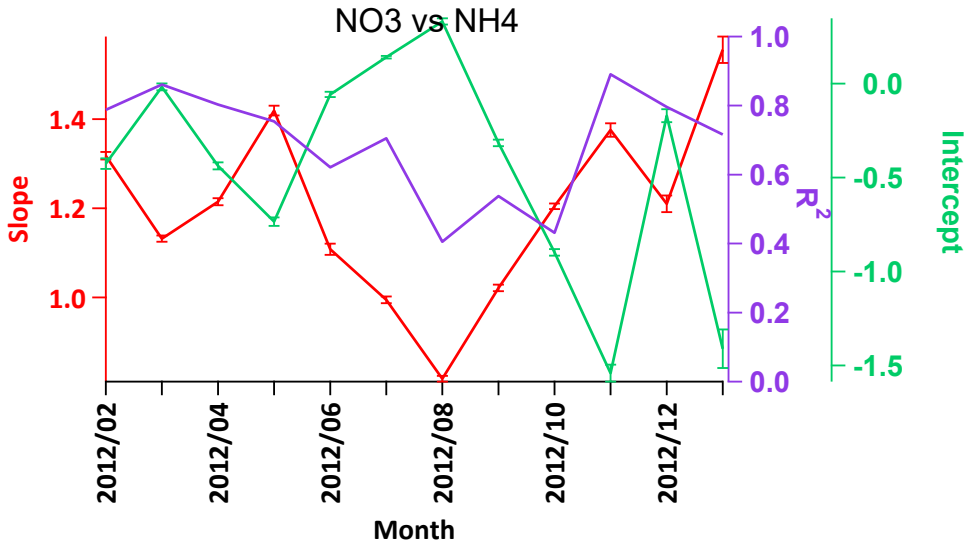


图 6.6.3 根据年/月执行批量绘图的范例。

7 分页“Multiply Y time series”简介

多变量Y时间序列图通常用于呈现各种污染物的时间变化。 如下所示，可以使用“添加”按钮选择所需的Y。

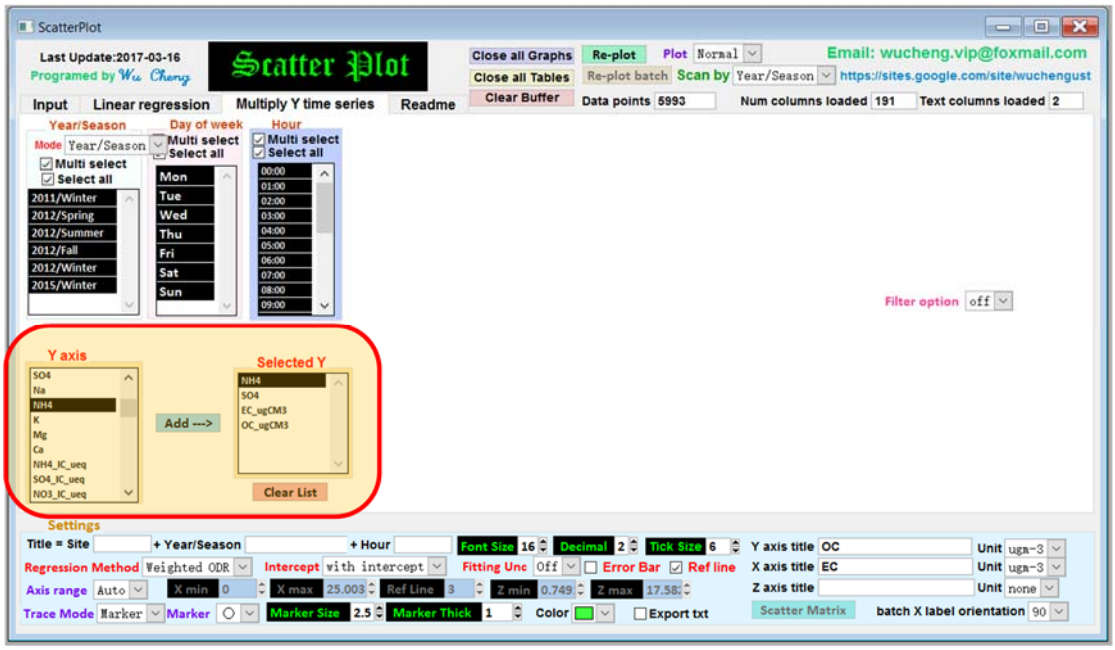


图7.1通过“Add”按钮选择所需Y的示例

“Plot option”应设置为”new”。 然后，点击”Re-plot”将在新窗口中生成图形。而后用户可以在新窗口中设置颜色和线形，轴标题等外观。 本功能主要目的是节省为单个轴设置Y方向的份额需要耗费的时间。

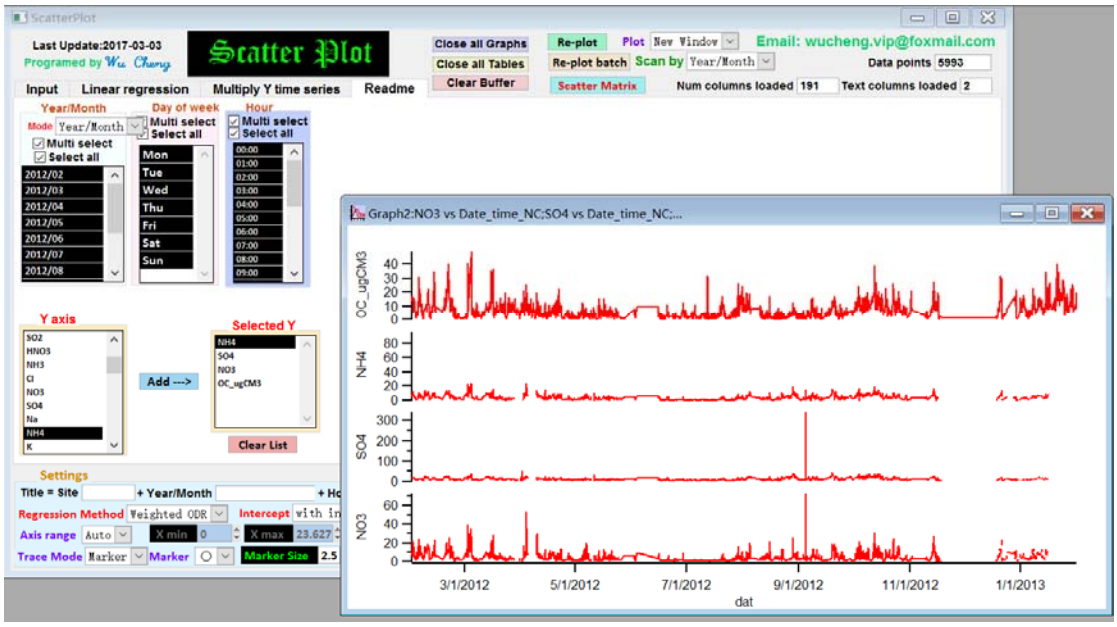


图7.2 点击”Re-plot”后在新窗口中生成多变量Y时间序列图的示例。