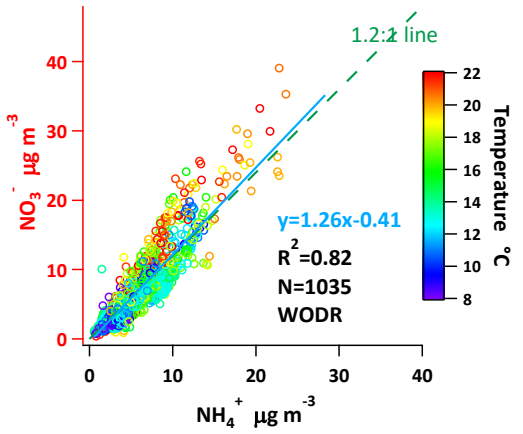
Scatter Plot



Scatter Plot使用说明书

Manual for Scatter Plot

吴晟

wucheng.vip@foxmail.com

2018-03-06

前言

Scatter Plot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。

有关Scatter Plot的评估和应用的更多细节，请参阅（**如果你在文章中用到了本软件，请引用以下文章**）

Wu, C. and Yu, J. Z.: Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting, **Atmos. Meas. Tech.**, 11, 1233-1250, doi:10.5194/amt-11-1233-2018, 2018.

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/

https://doi.org/10.5281/zenodo.832416

吴晟

2018-03-06

目录

0 Igor Pro 运行环境的安装 1
1 关于数据结构的建议 3
2 跟其他程序的总体比较 5
3 导入数据 6
3.1 在 MS excel 中的时间线示例 6
3.2 从Excel复制 7
3.3 将数据粘贴到Igor中 8
3.4 更新列表 9
3.5 指定时间轴 10
4 通用设置介绍 11
5 分页“Input”简介 13
6 分页“Linear regression ”简介: 15
6.1 数据按时间筛选 15
6.2 用数据进行筛选 16
6.3 使用图形界面进行数据遮掩 18
6.4 选择多个变量用作X&Y 22
6.5 时间变量作为Z轴 23
6.6 批量绘图 24
7 分页“Multiply Y time series”简介 26
8 分页“Percentile”简介 27

0 Igor Pro 运行环境的安装

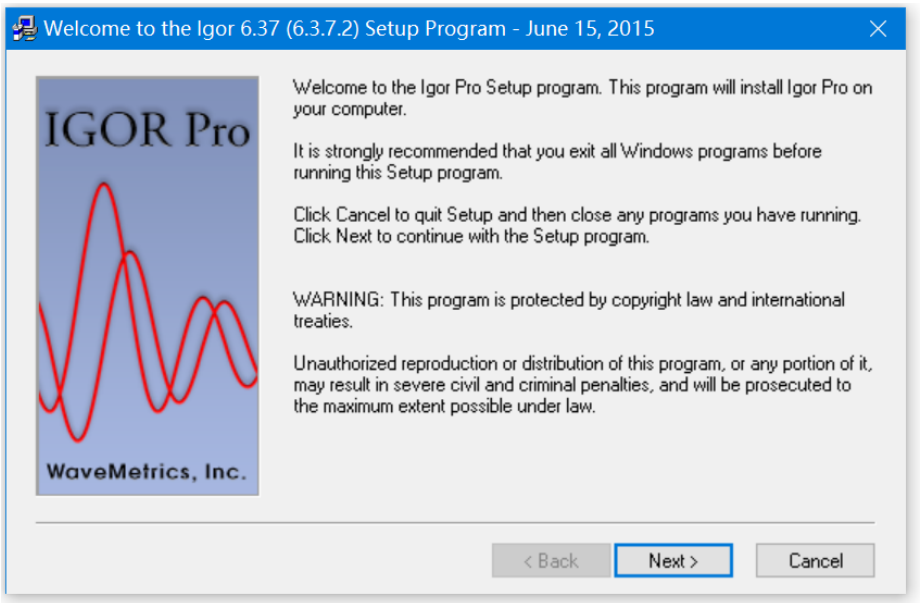
用户需要Igor Pro平台来运行Scatter Plot程序（pxp文件）。这类似于你需要安装微软 word来打开docx文件的场景。因此，您需要在计算机上安装Igor Pro。Igor Pro拥有Windows PC和Mac版本。

建议使用Igor Pro 6.x版。 Scatter Plot也可以在Igor Pro 7.x上运行，但在用户界面显示中存在一些问题，某些元素（按钮，下拉菜单）的比例失调，尤其是在计算机使用屏幕高DPI设置时。

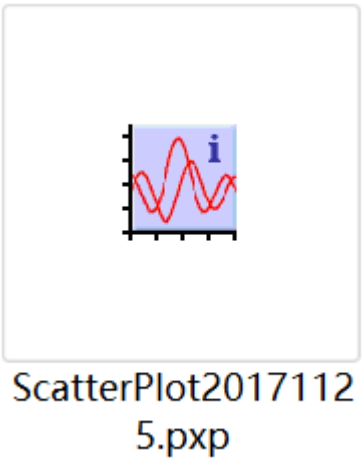
1) 从 https://www.wavemetrics.com/support/versions.htm 下载Igor Pro

2) 双击Igor Pro安装文件

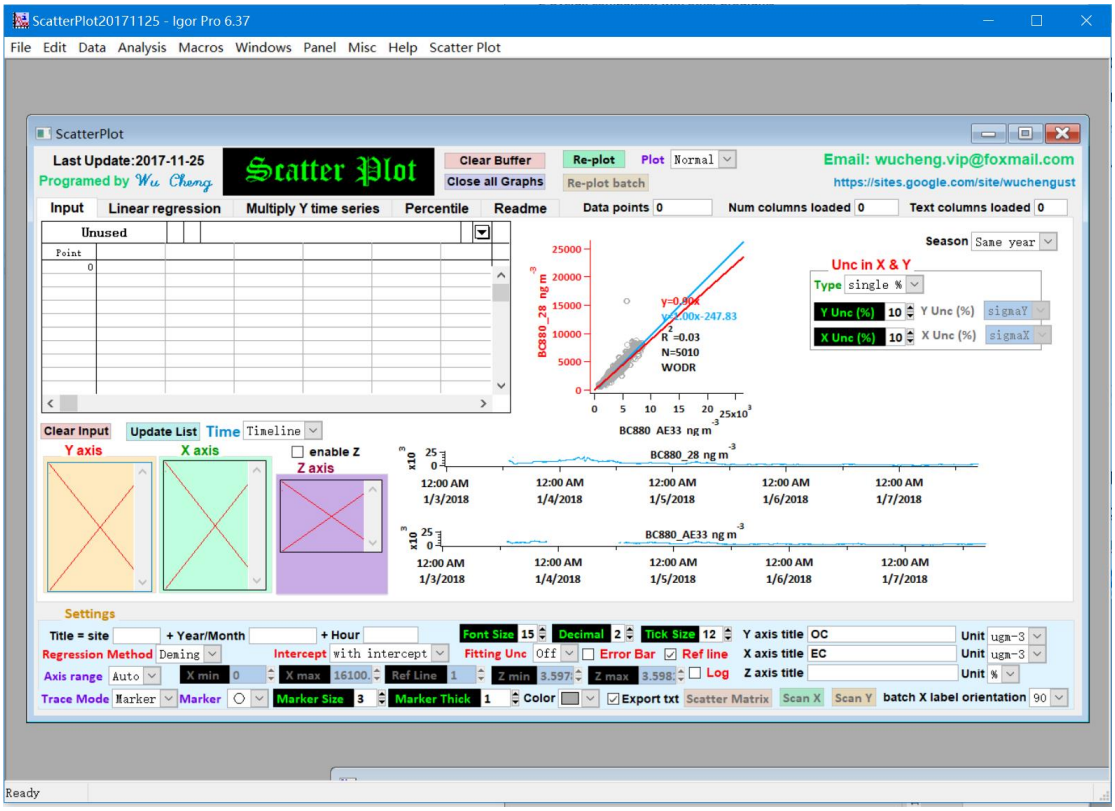
3) 按照步骤安装Igor Pro



4) 成功安装后，系统可以识别pxp文件(如下图图标那样)。如果您没有Igor Pro的激活码，演示版本Igor Pro会持续30天。 30天后，用户不能：a）导出图和文件; b）保存文件。



5) 要运行Scatter Plot, 只需双击pxp文件即可。如果要同时打开多个pxp文件，请使用光标选择pxp文件，然后按“Ctrl + Enter”。



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）

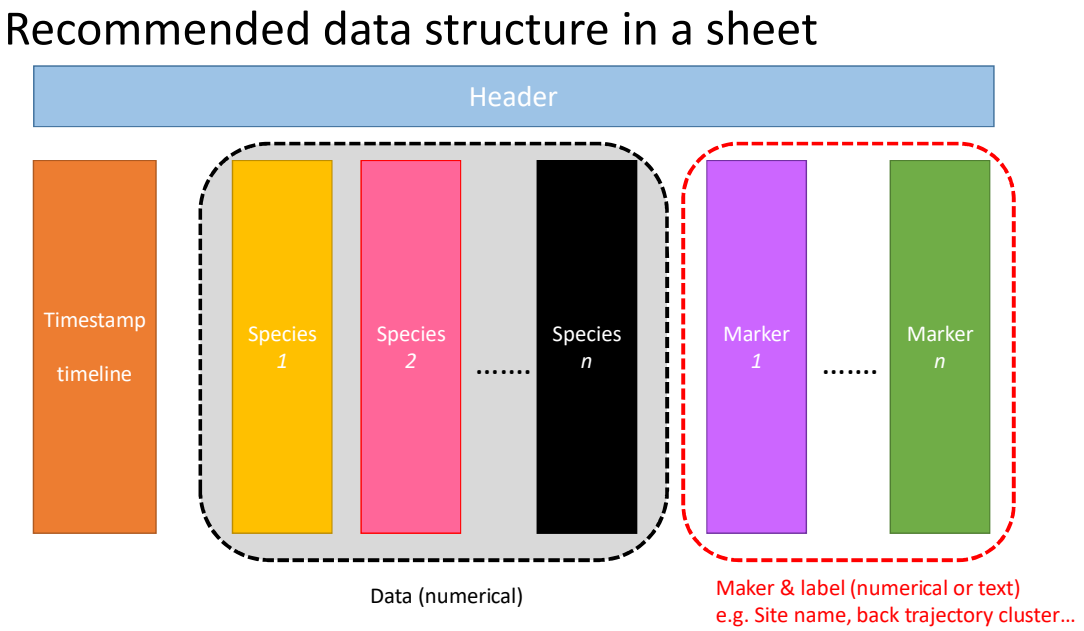


图 1.1 推荐在工作表中的数据结构

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(Scatter Plot)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。

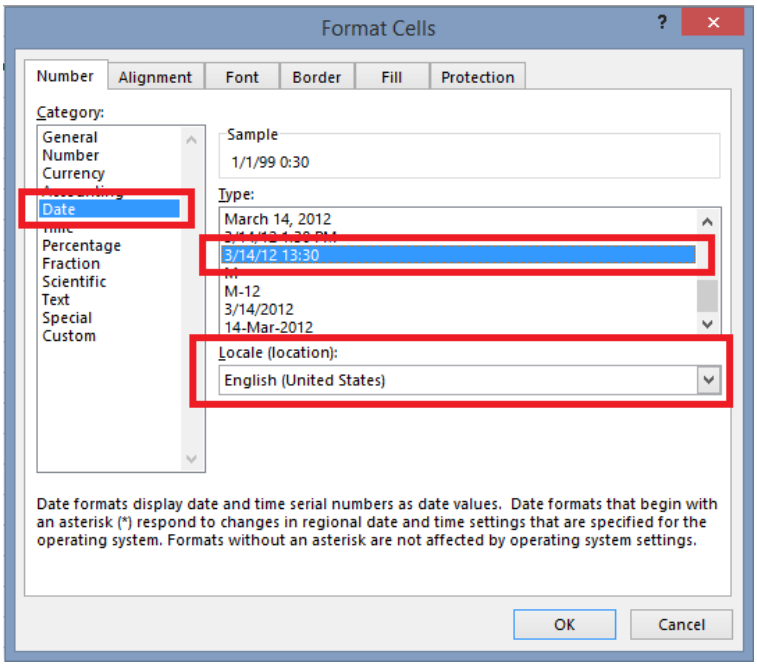
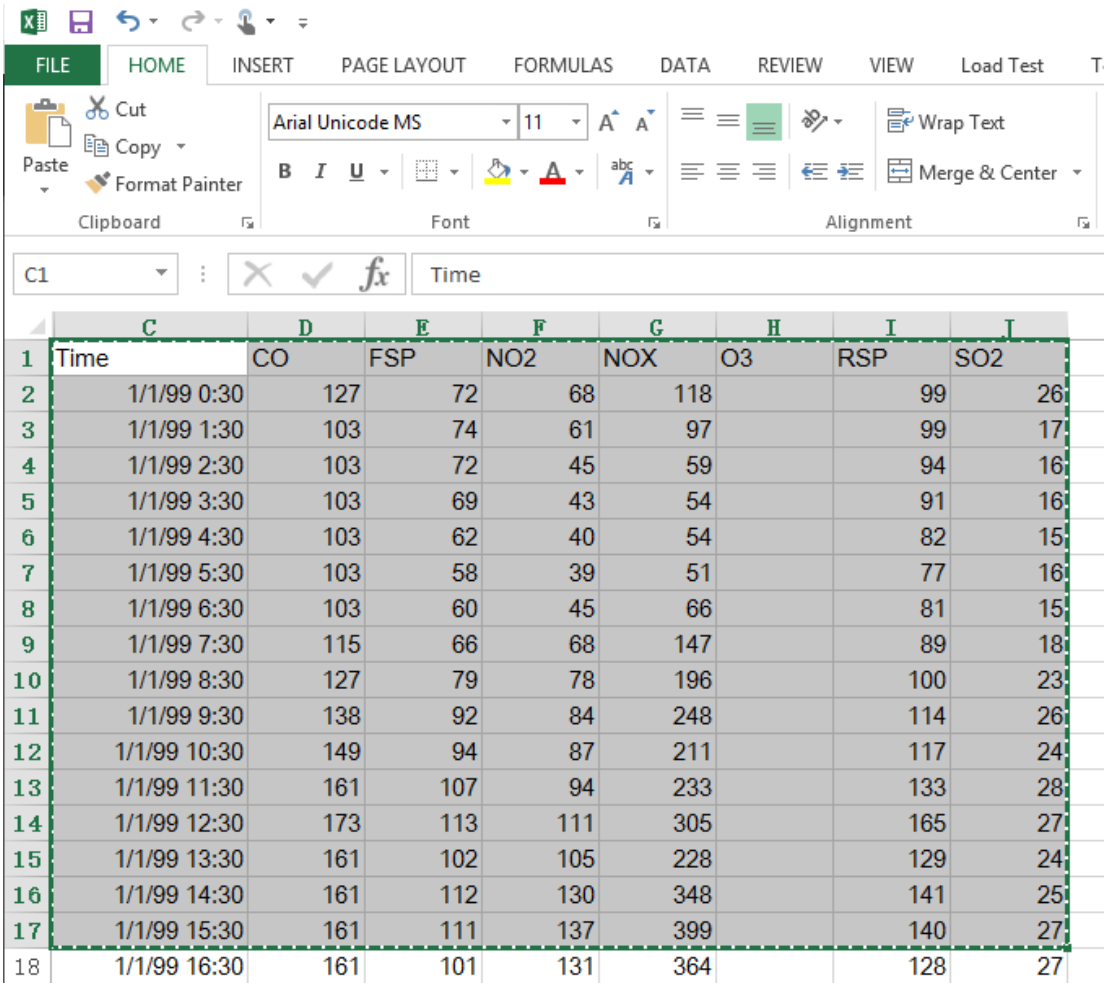


图3.1 时间轴列的MS Excel中的单元格格式配置

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

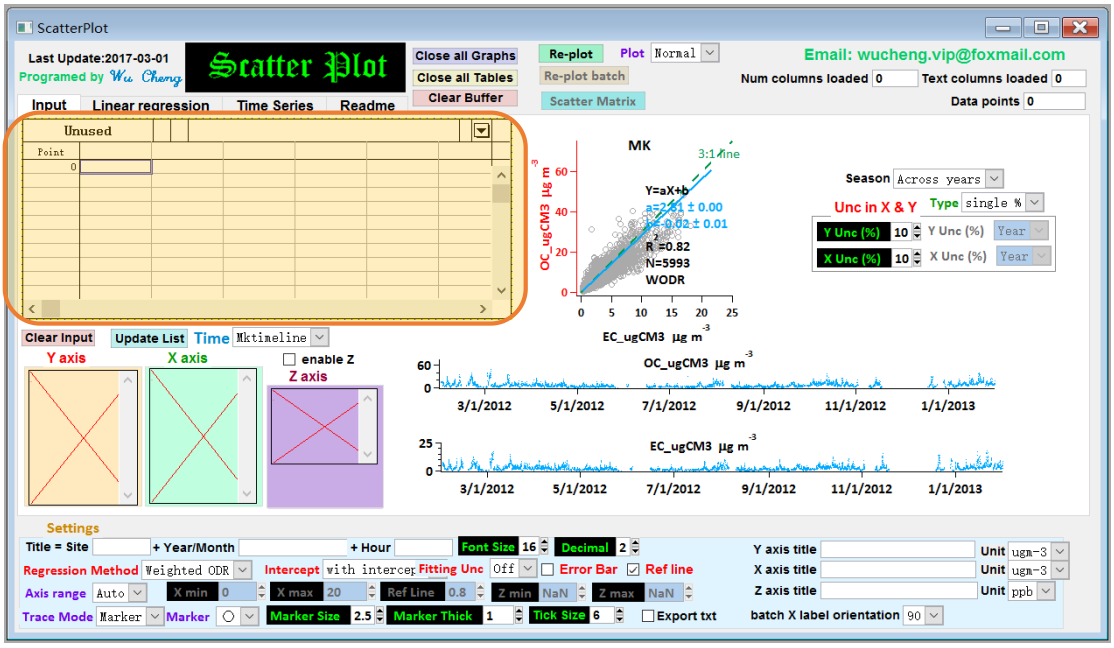


图3.3.1 粘贴数据之前Igor Pro中用户界面的示例.

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。

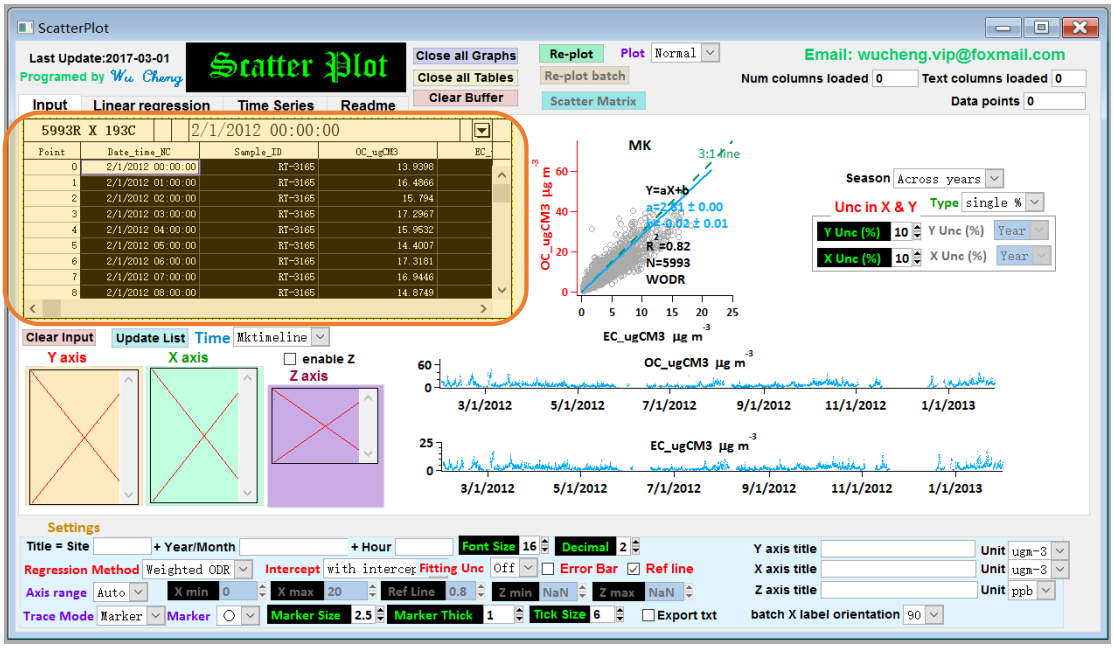


图3.3.2 粘贴数据后Igor Pro中的用户界面示例

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。

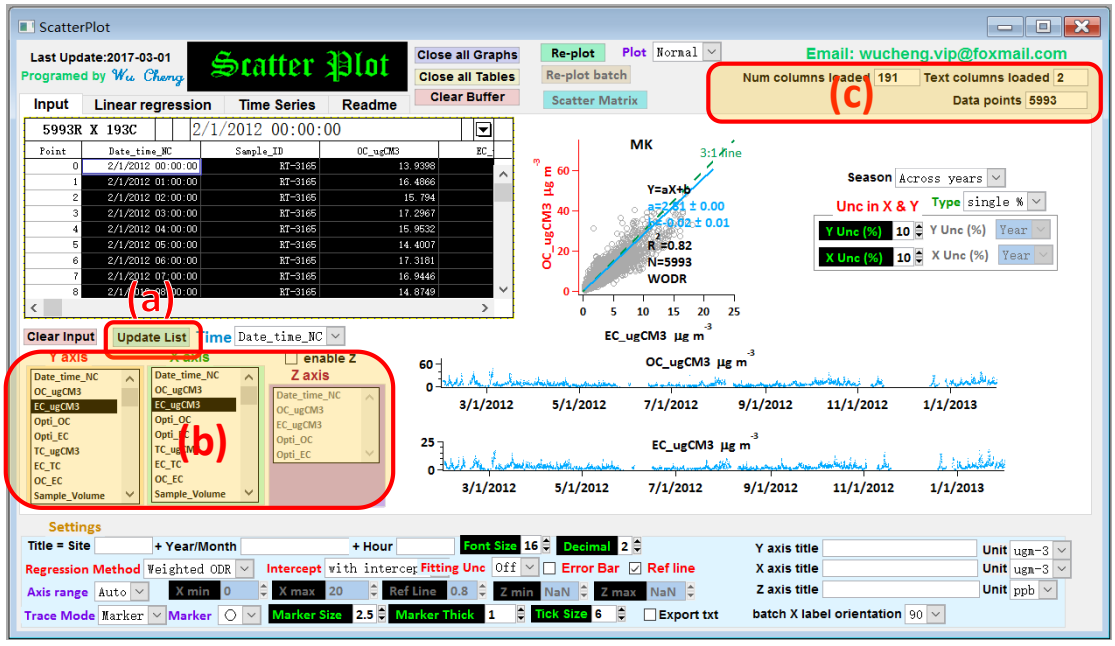


图3.4　Igor中的更新列表示例。

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。

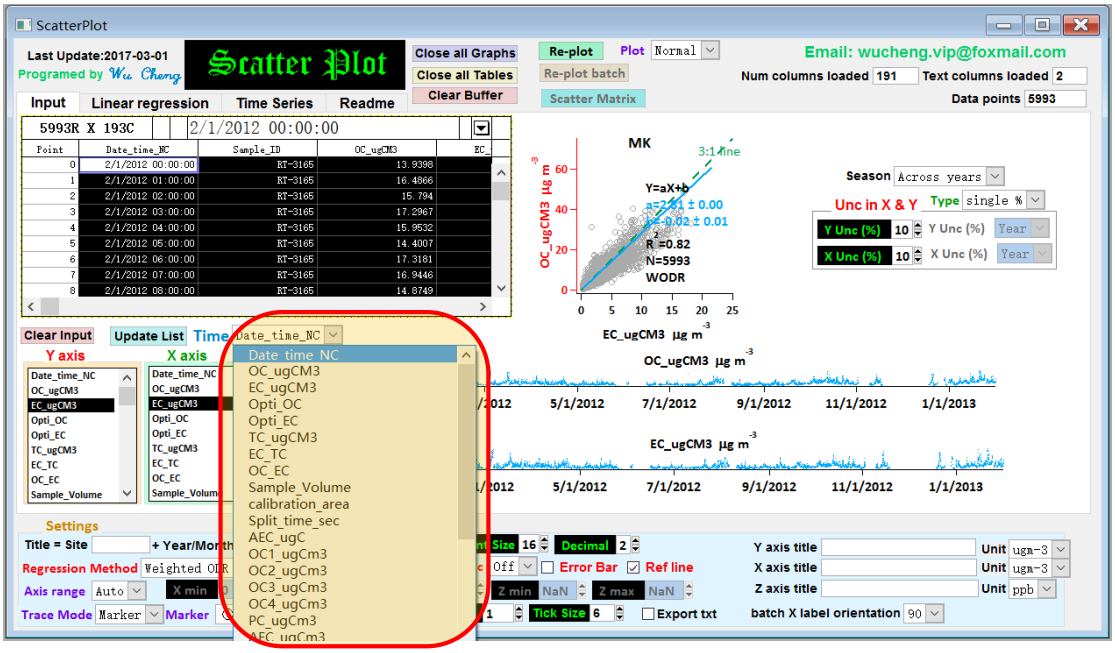


图3.5 在Scatter plot Igor程序中指定时间轴的示例

4 通用设置介绍

Scatter Plot 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

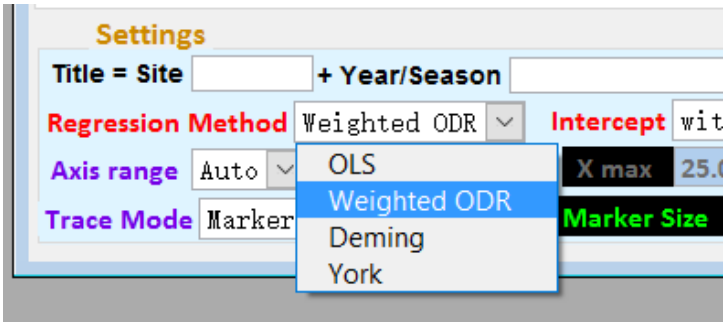
Clear Buffer：删除批处理图的缓存数据，避免程序文件过大

Replot：在当前选项卡中重绘图

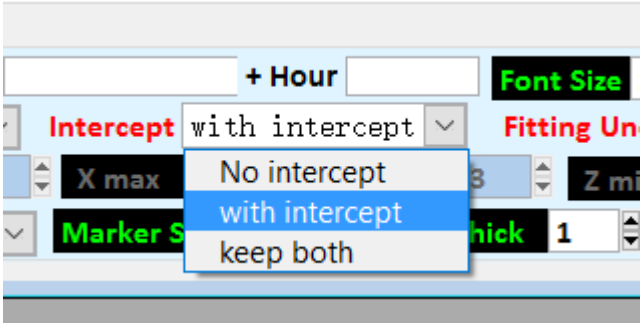
Plot option: Normal，只重绘当前选项卡; New window，也可以在一个新窗口中生成绘图，可以复制并粘贴到MS办公室; Export PNG，不仅在新窗口中生成绘图，还生成PNG文件; Export EMF，而不是PNG文件，EMF是一个矢量文件，可以无限放大。生成的PNG和EMF文件跟本Igor程序(.exp文件)存放在同一个文件夹。

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。

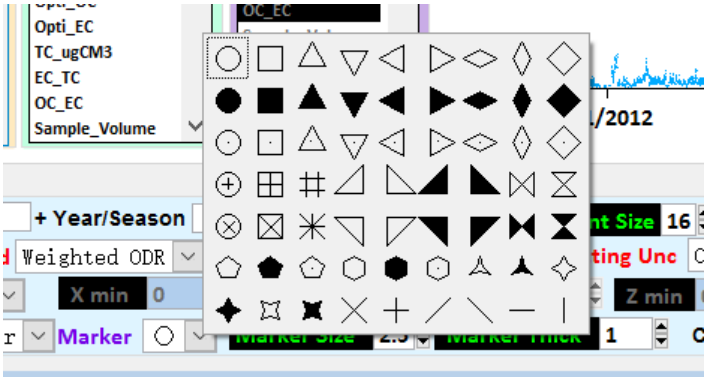


Intercept：如果选择No intercept，则将通过原点进行回归（不适用于Deming和York回归）。如果选择“with intercept”，则所有回归方法都可用。如果选择“keep both”，则将执行有和无截距回归（不适用于Deming和York回归）。



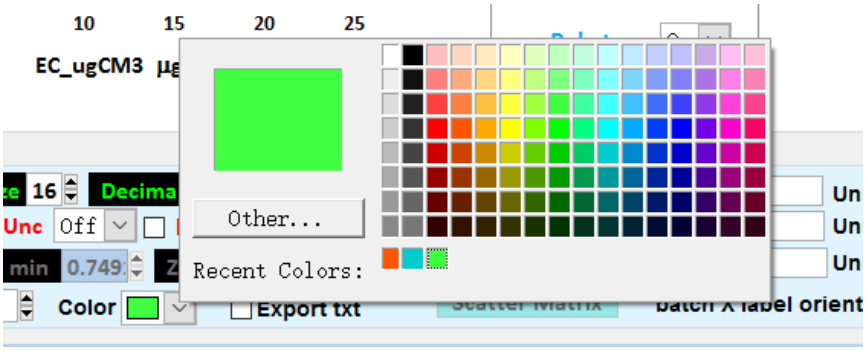
Axis range：打开或关闭XY轴的自动标度

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

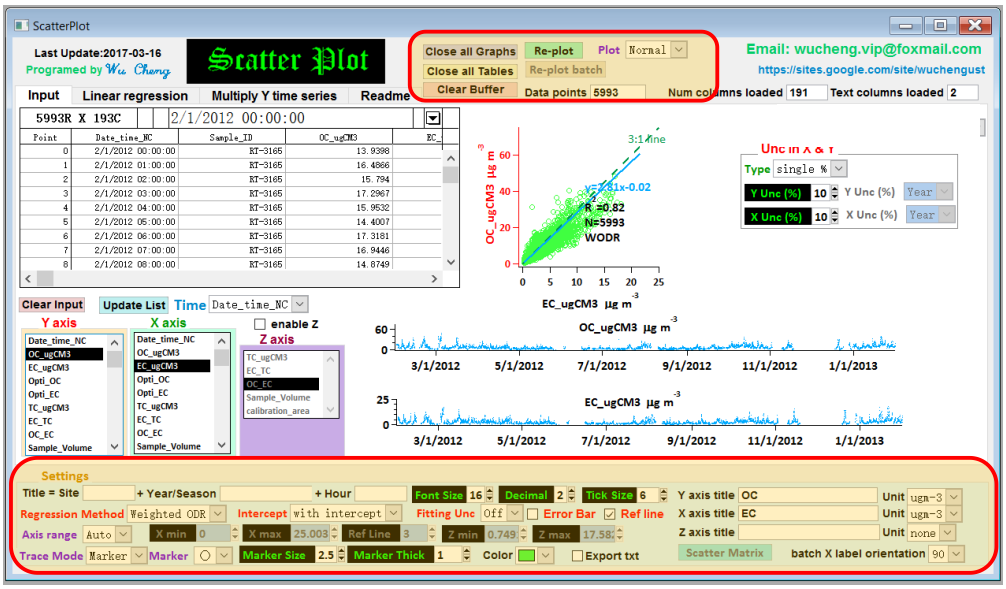


图4.1 Scatter plot Igor 程序中的常规设置示例。

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。

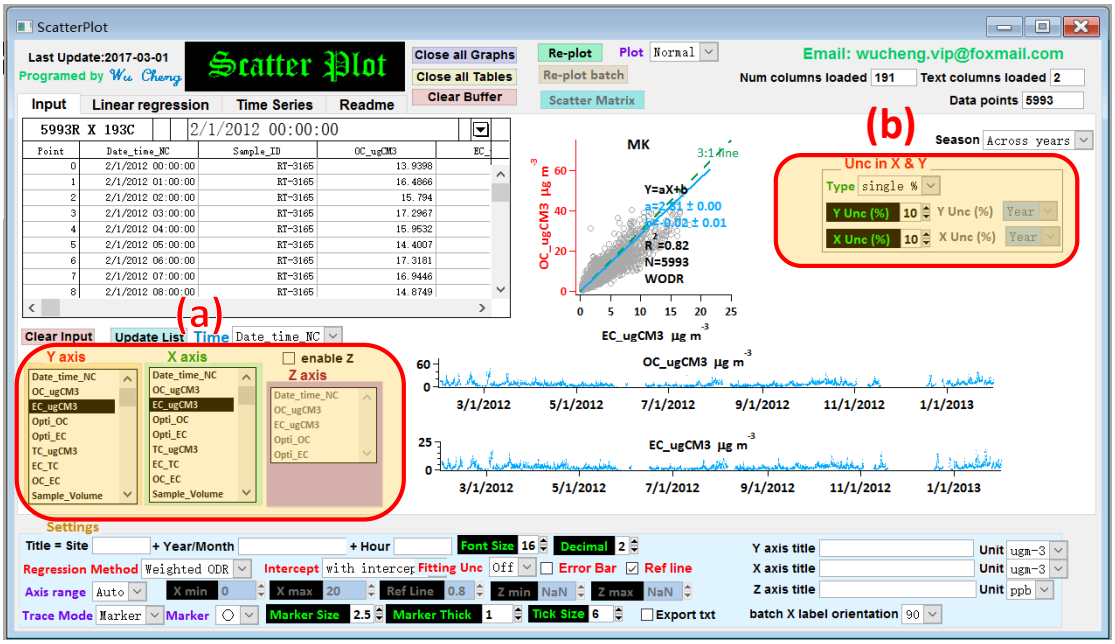


图5.1 Scatter plot 设置

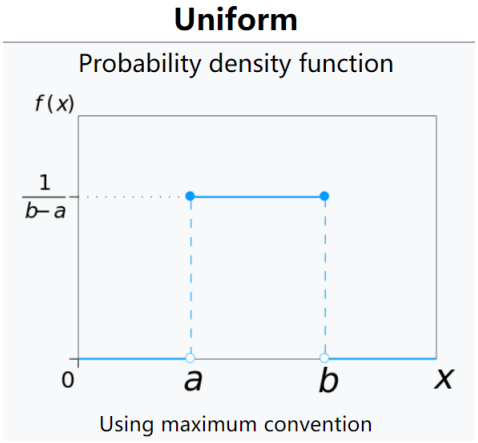
In Deming regression, the key parameter λ is the ratio of the weights:

$$\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$$

and the weights are:

$$\omega(X_i) = \frac{1}{{\sigma_{X_i}}^2}\ , \omega(Y_i) = \frac{1}{{\sigma_{Y_i}}^2}$$

σ_{X_i} and σ_{Y_i} are the standard deviations of the error in measurement of X_i and Y_i respectively. For example, say data point X_i has a ±m% uncertainty, which follow a uniform distribution (in the range of [a,b]). the variance of the uniform distribution becomes



$$\begin{aligned}{\sigma_{X_i}}^2 &= \frac{1}{12}(b-a)^2 \\ &= \frac{1}{12}\left(X_i + m \times X_i - (XX_i - m \times X_i)\right)^2 \\ &= \frac{1}{12}(2 \times m \times X_i)^2 = \frac{(mX_i)^2}{3}\end{aligned}$$

As a result, the standard deviation of the error can be written as

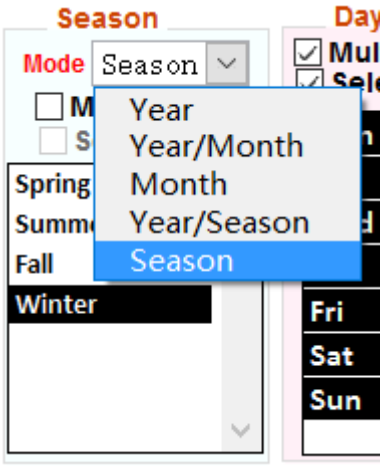
$$\sigma_{X_i} = \frac{mX_i}{\sqrt{3}}$$

In "Input Data" mode, σ_{X_i} and σ_{Y_i} are required as measurement error input for WODR and YR. σ_{X_i} and σ_{Y_i} are also used to calculate λ for Deming regression.

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。

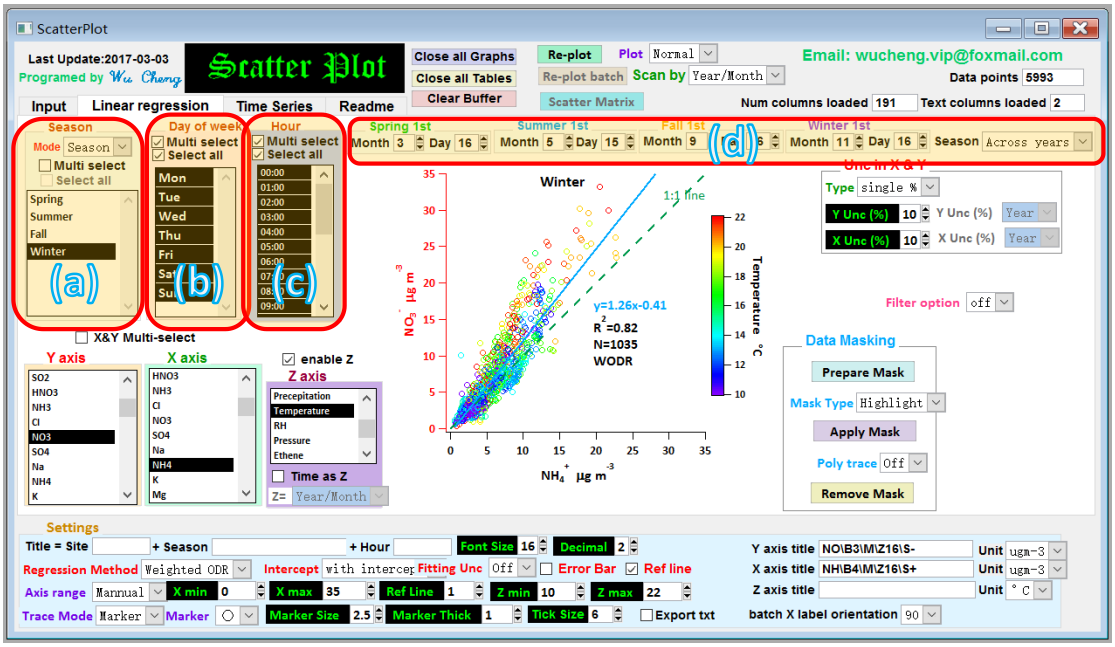


图 6.1.1 时间筛选的列表框

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

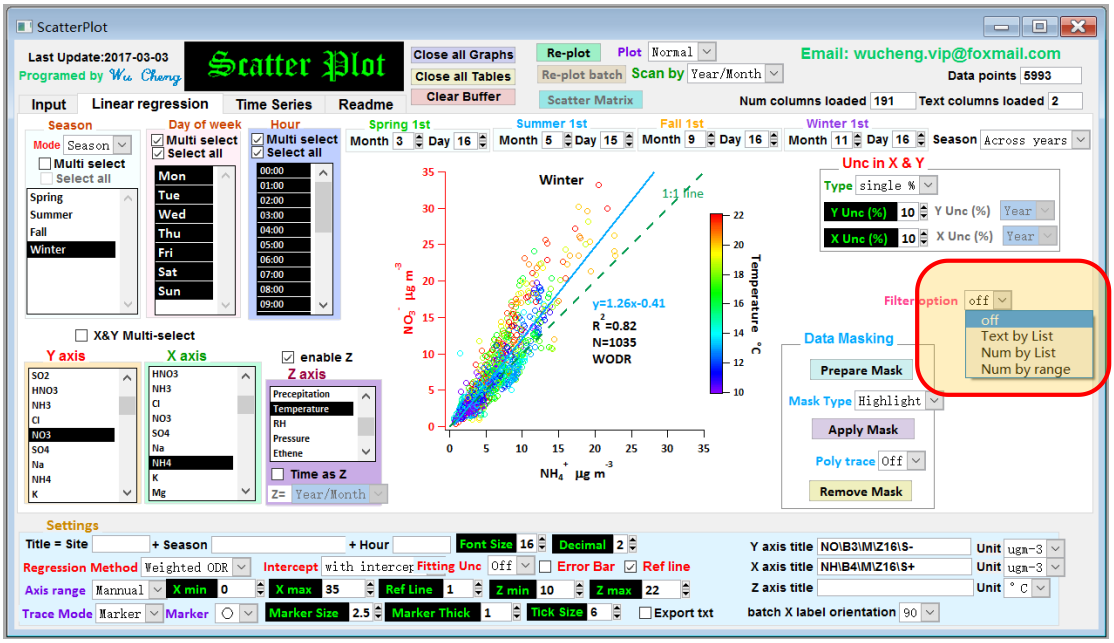


图 6.2.1 按数据过滤数据

(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。

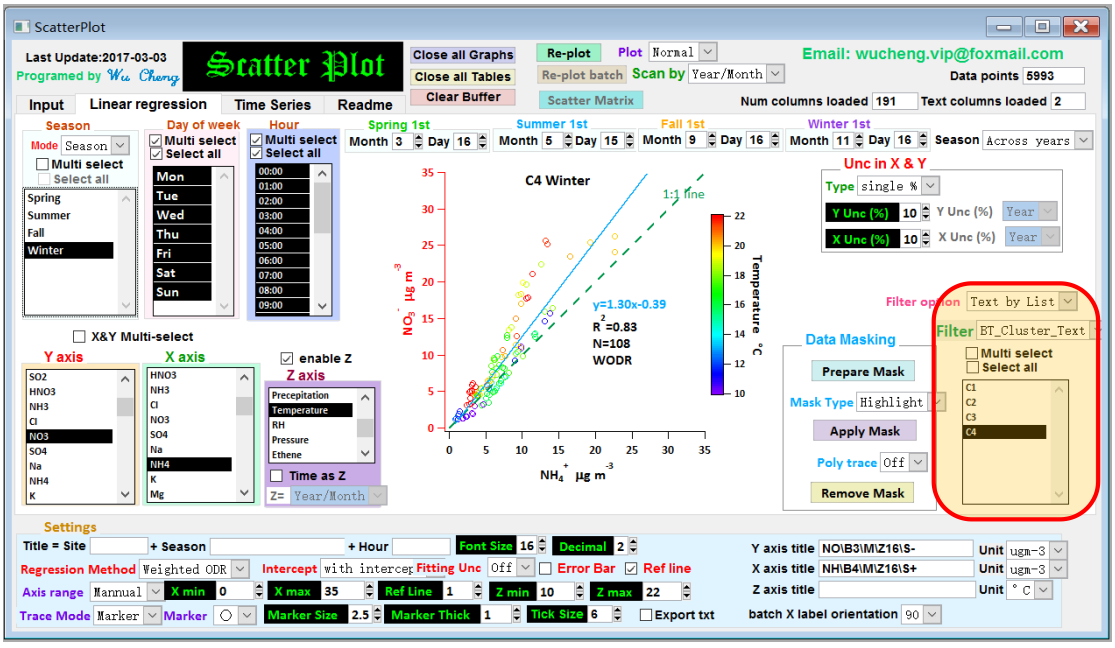


图6.2.2 按文本数据列表过滤数据 - Text by List

(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。

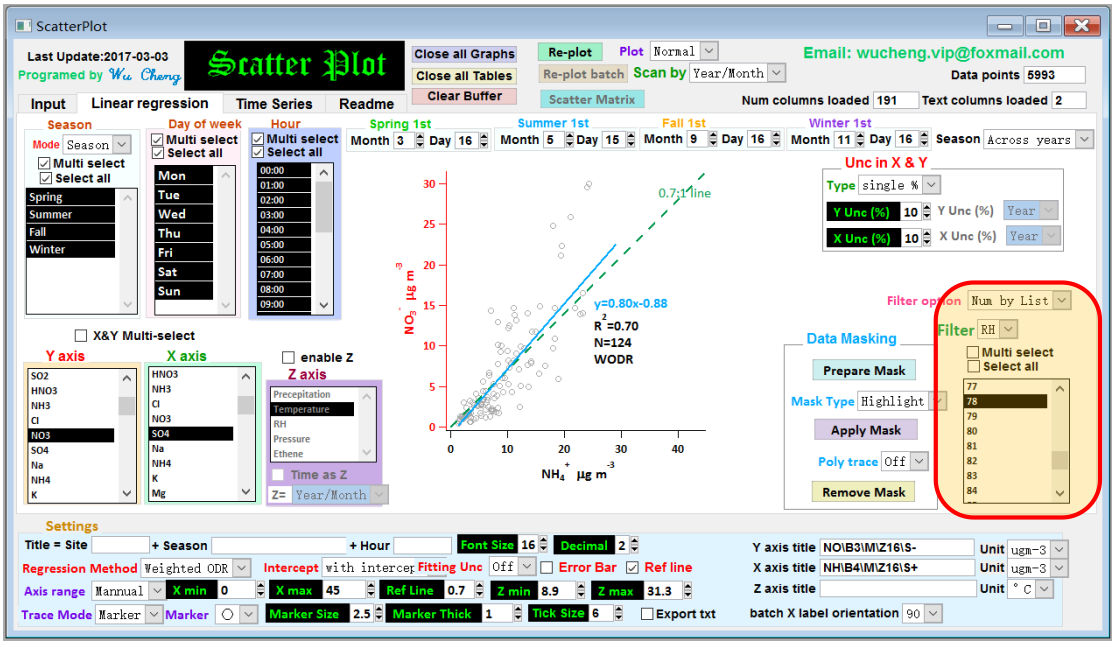


图 6.2.3 按数据过滤数据– Num by List

(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。

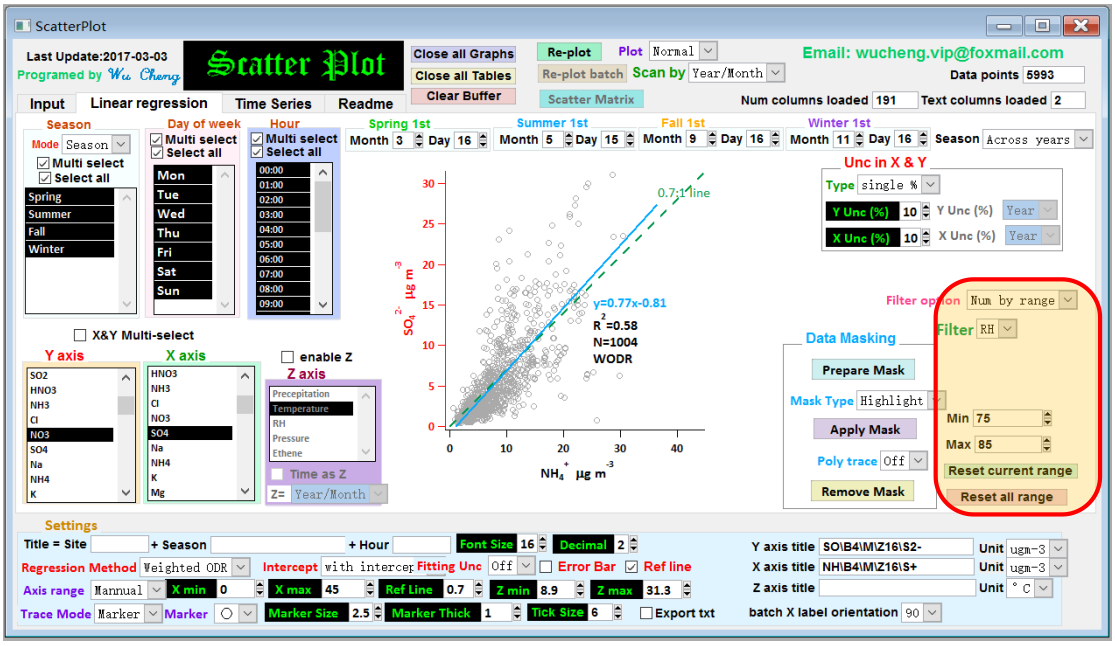


图 6.2.4 按数据范围过滤– Num by range

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。

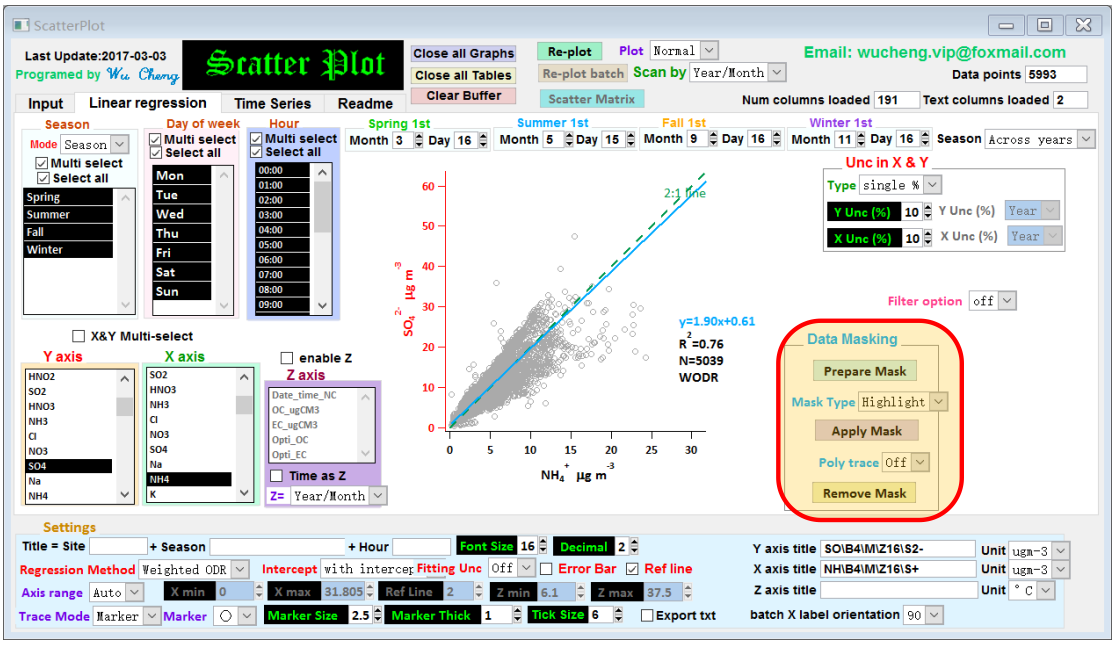


图 6.3.1 数据遮掩功能区一览

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。

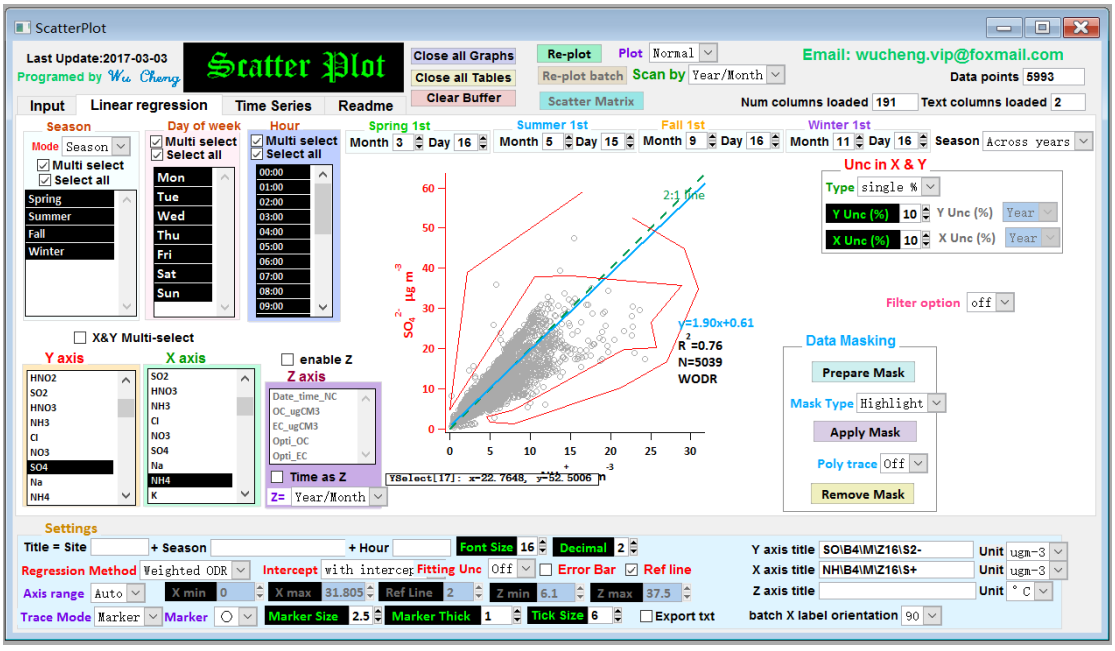


图 6.3.2 用光标开始进行多边形的绘制。

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。

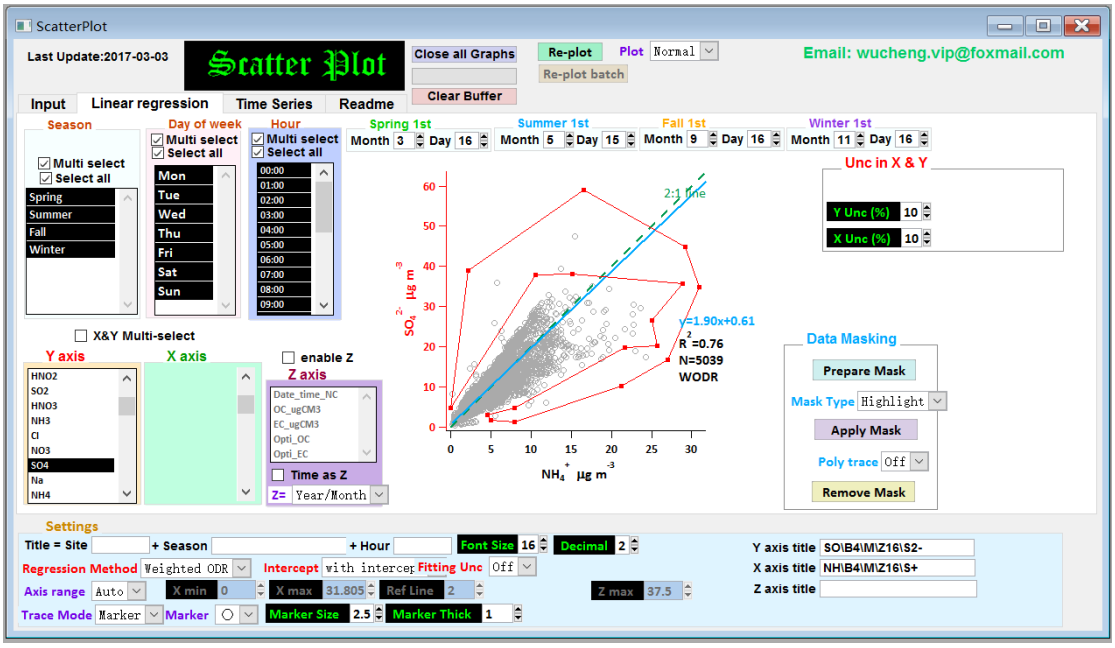


图6.3.3 闭合的多边形的示例

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。

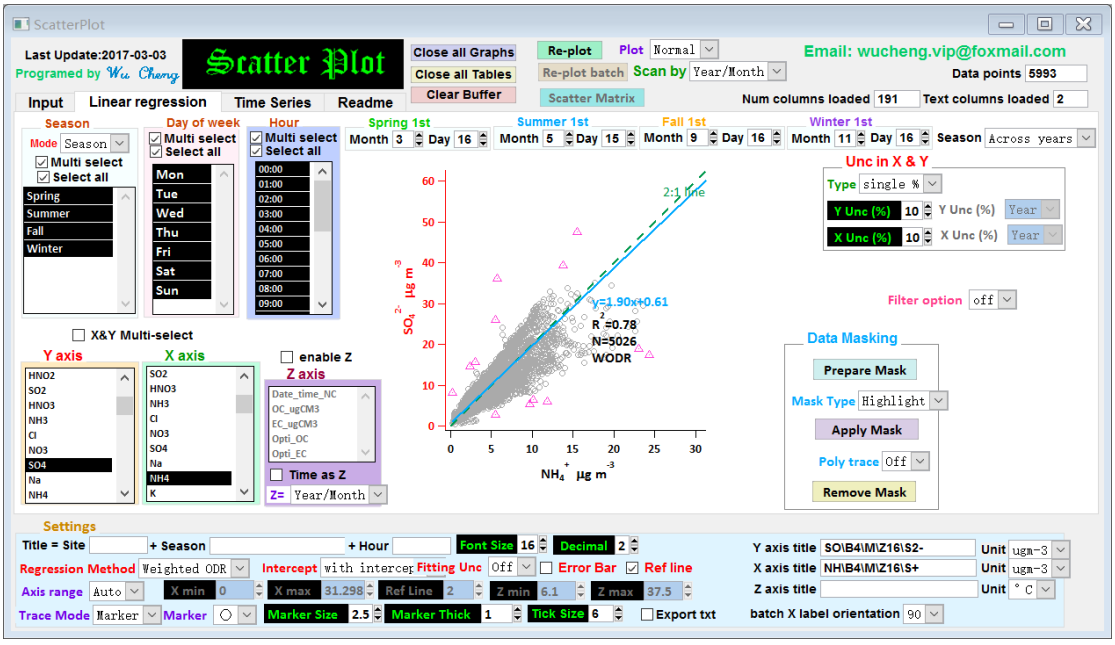


图6.3.4 在"Mask type"中选择“Highlighted”的示例。

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。

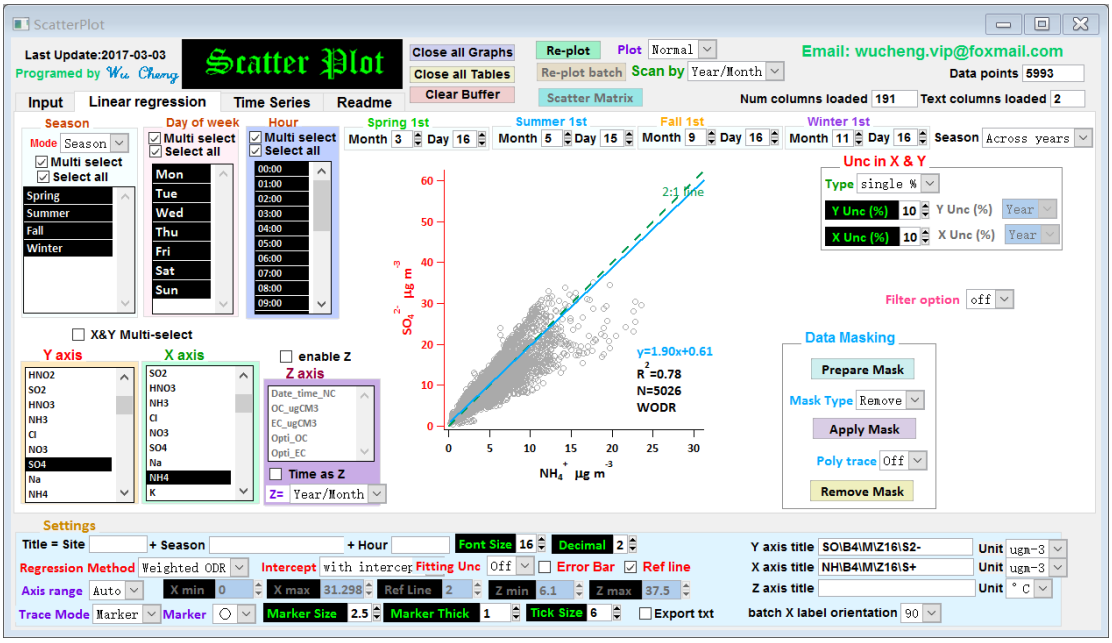


图 6.3.5 在"Mask type"中选择“Remove”的示例。

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）。

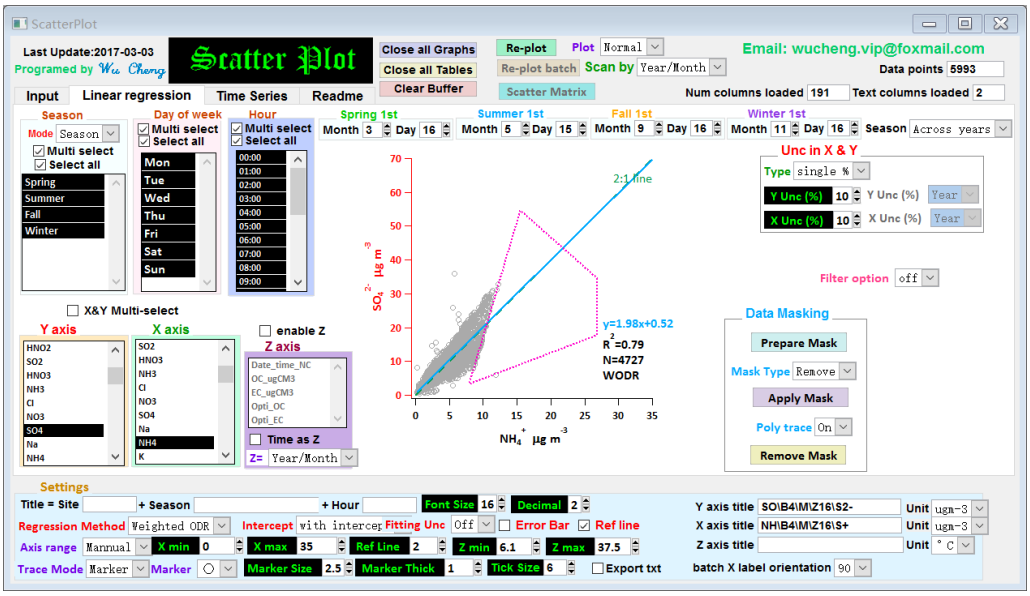


图 6.3.6 在"Trace type"中选择“On”的示例.

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。

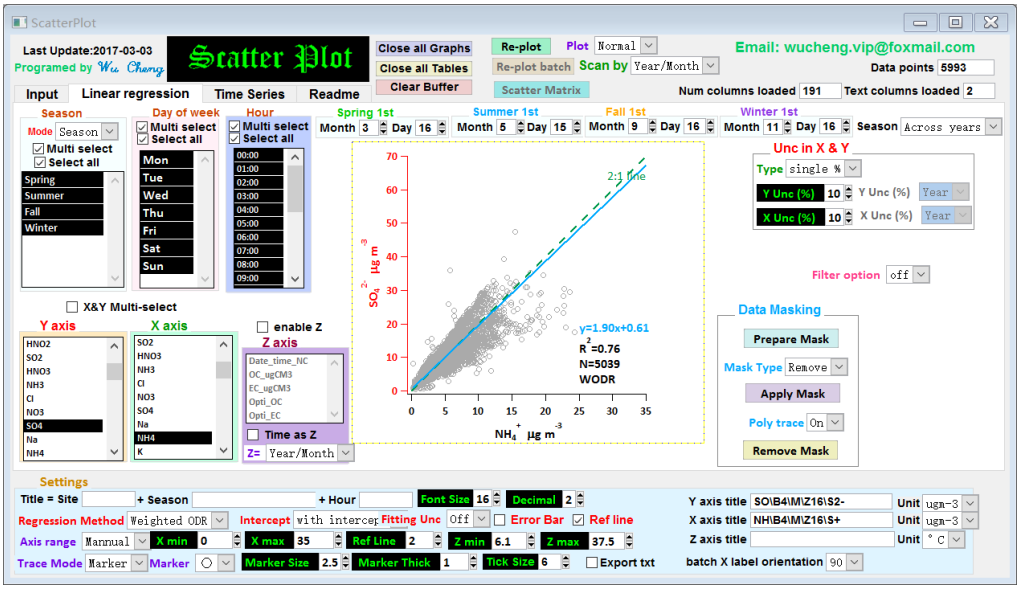


图 6.3.7 应用“Remove Mask”后的示例。

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。

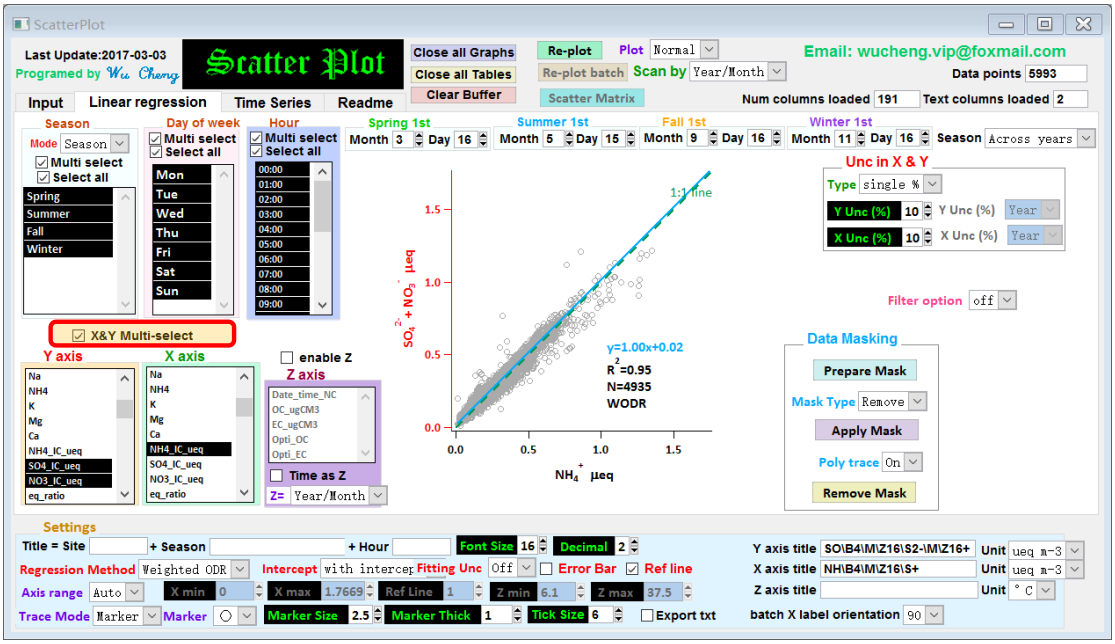


图 6.4.1 选择多个变量用作X&Y的示例.

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。

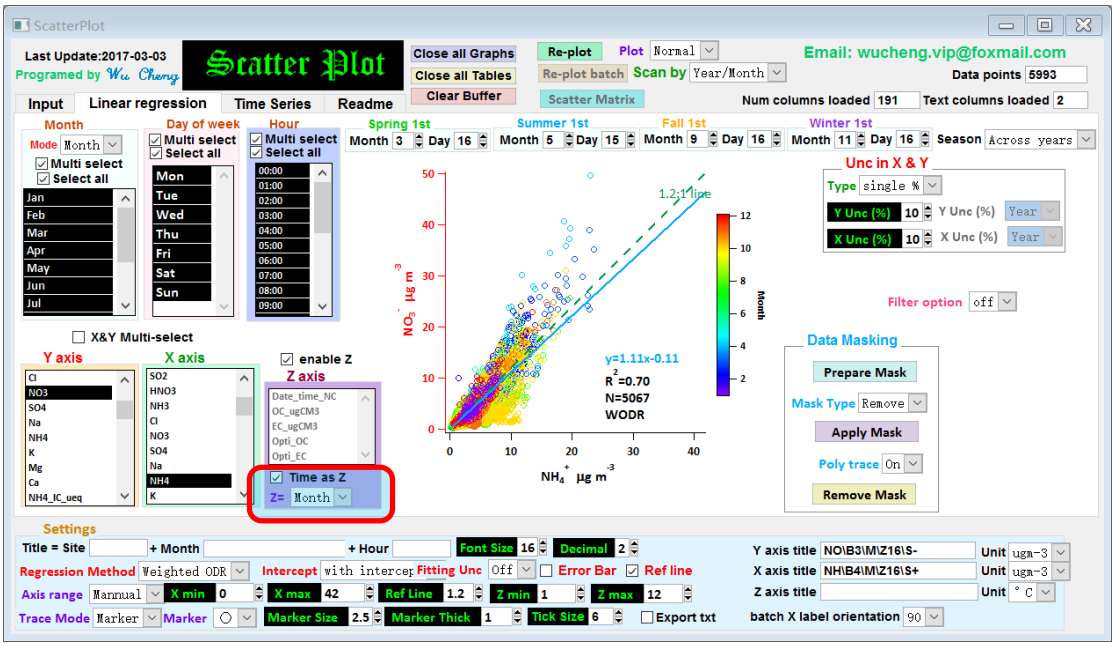


图 6.5.1 使用月作为Z轴颜色编码的示例。

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。

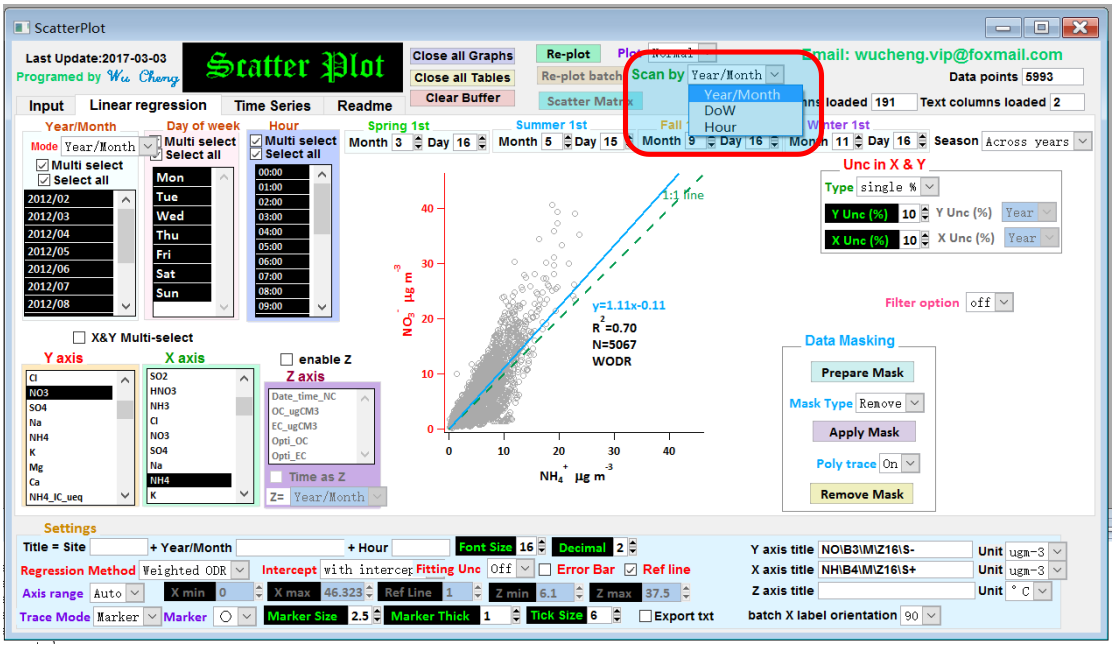


图 6.6.1 按时间维度进行批量绘图的设置

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。

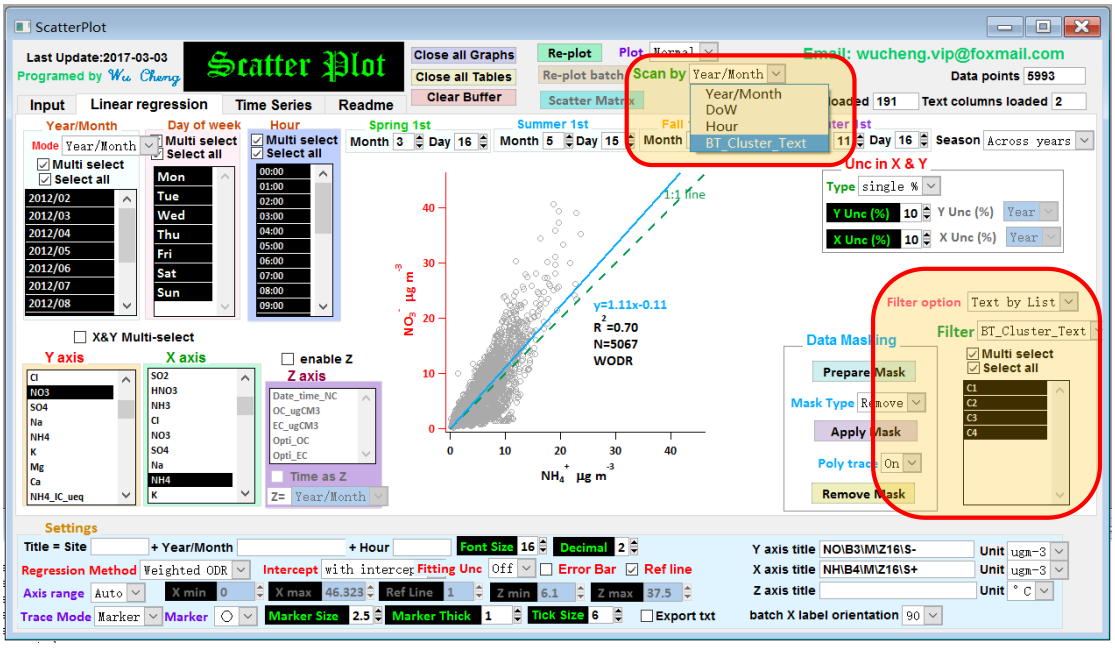
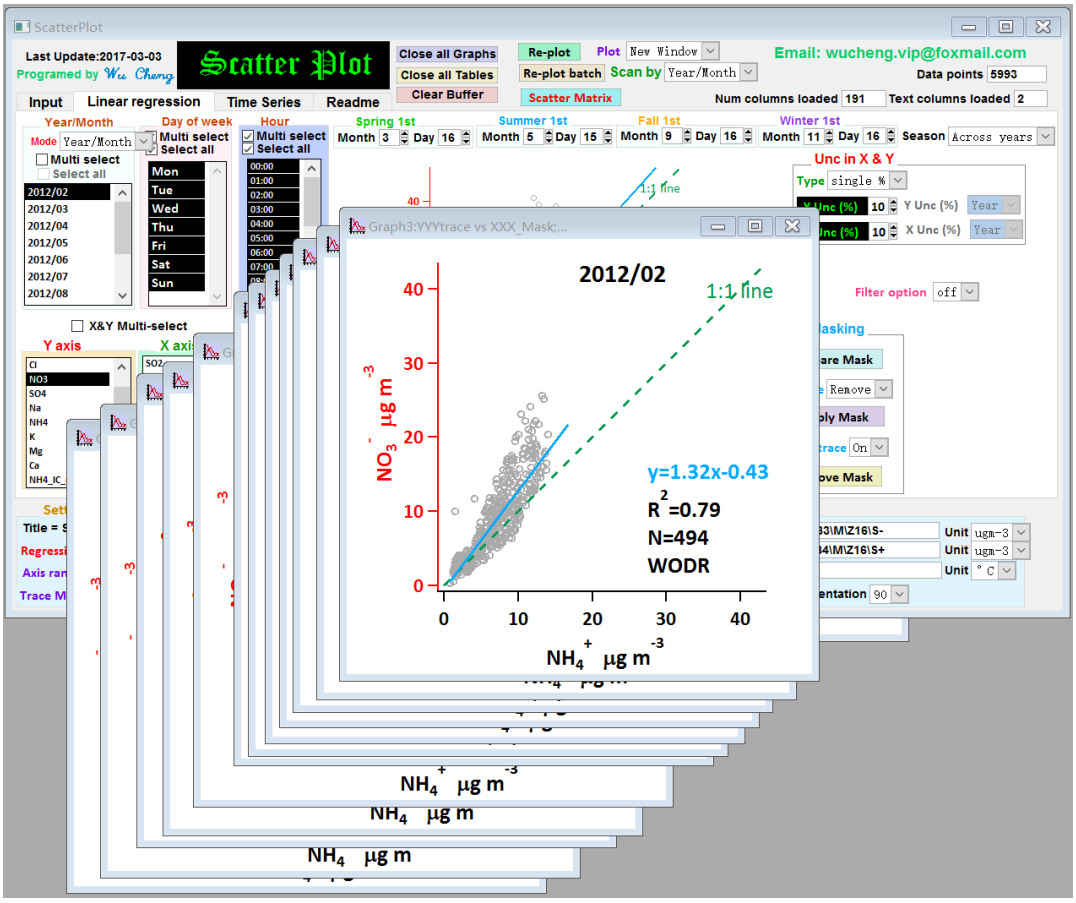


图6.6.2 通过文本标记进行批量绘图

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）。



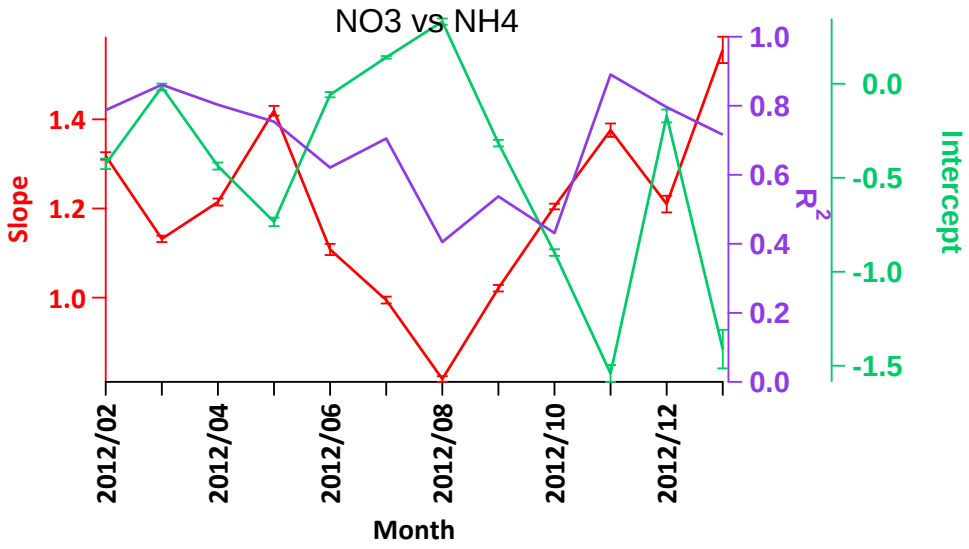


图 6.6.3 根据年/月执行批量绘图的范例。

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。

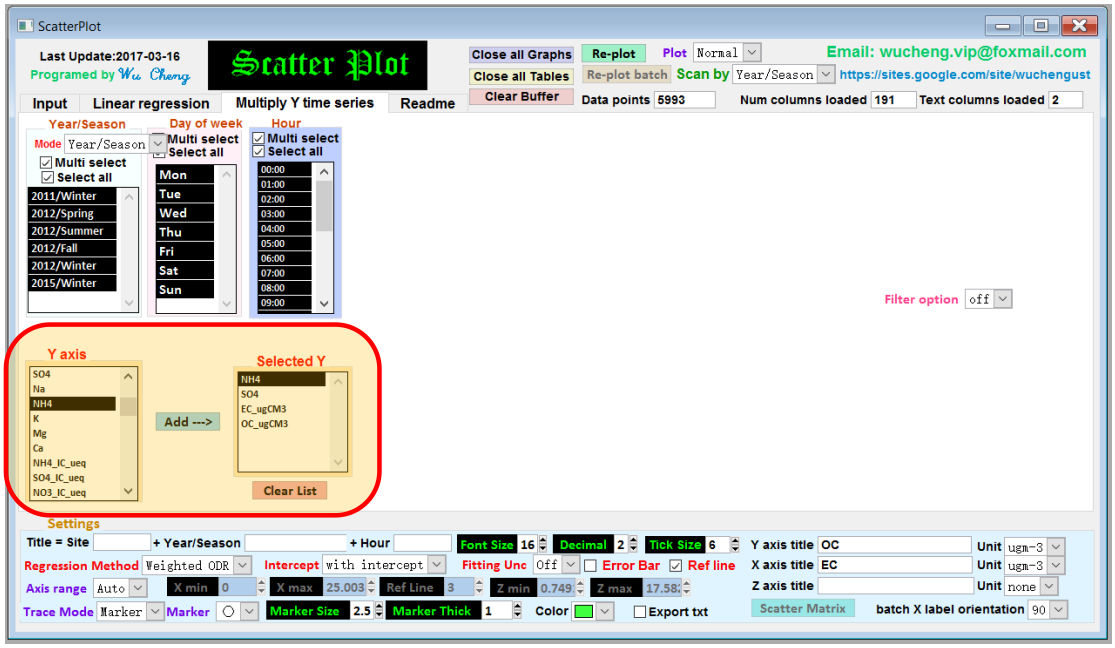


图7.1通过“Add”按钮选择所需Y的示例

“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。

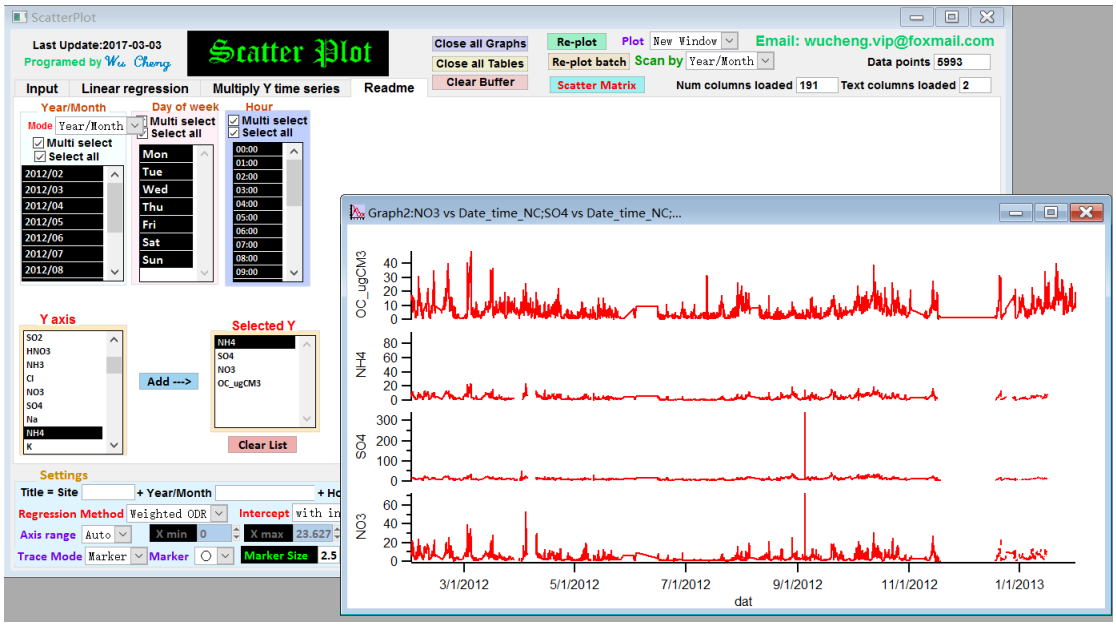


图7.2 点击”Re-plot”后在新窗口中生成多变量Y时间序列图的示例。

8 分页“Percentile”简介

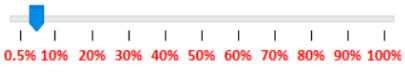
在EC示踪法中，具有特定OC / EC百分位数的子集通常用于回归来确定(OC/EC)$_{pri}$。

Step: 百分位数子集的步长. 例如，0.005代表0.5%的间隔。

Slope Min & Max: 设置左图的斜率范围（y轴）

Intercept Min & Max: 设置左图的截距范围（y轴）

Replot: 计算子集上的逐步回归。例如，以0.5%的步长，从OC / EC 0.5%～100%的所有子集上计算线性回归。



：这个滑动条是为了选择百分位数。该选择仅在应用“Replot”后才可用。

Redraw: 在新窗口中创建绘图(plot option: new window).

使用Tab“Percentile”的步骤

1）设置步长，斜率和截距范围

2）点击“Replot”

3）使用滑杆检查结果。例如，要显示10%子集的回归结果，请将滑动条拖动到10%，右侧的散点图将相应更新。紫色的数据点代表回归的选定子集，灰色的数据点代表未使用的数据。

4）导出绘图（在新窗口/导出到文件），点击“Redraw”。

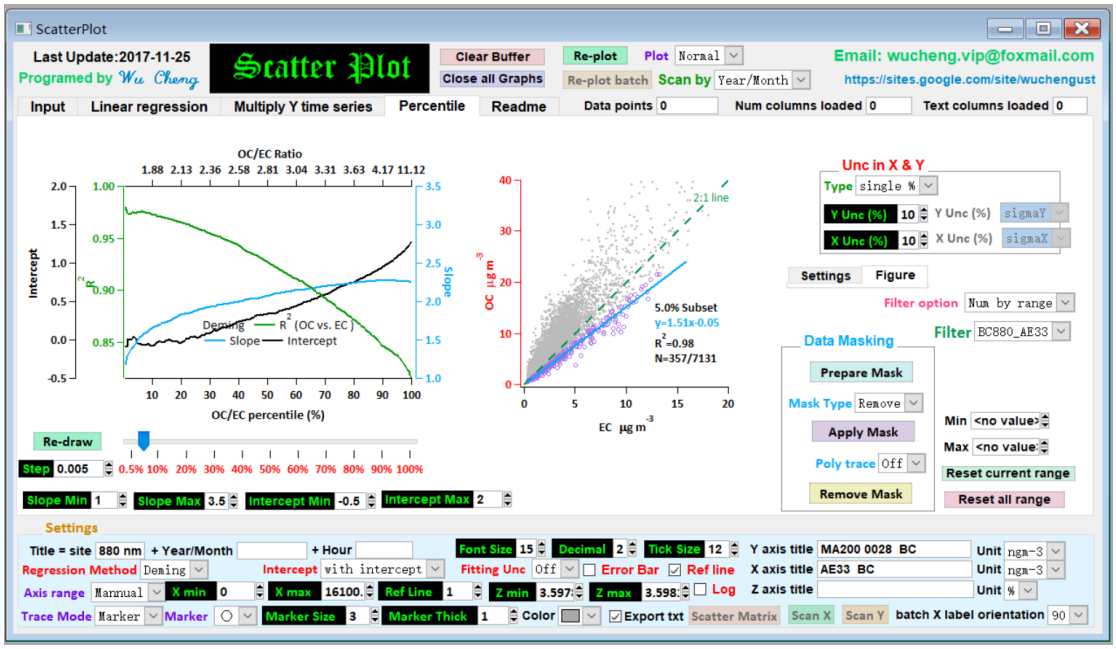


图 8.1 分页 “Percentile”.

Atmos. Meas. Tech., 11, 1233–1250, 2018
https://doi.org/10.5194/amt-11-1233-2018

Atmospheric Measurement Techniques
Open Access
EGU

Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting

Cheng Wu[1,2] **and Jian Zhen Yu**[3,4,5]

[1]Institute of Mass Spectrometer and Atmospheric Environment, Jinan University, Guangzhou 510632, China
[2]Guangdong Provincial Engineering Research Center for On-Line Source Apportionment System of Air Pollution, Guangzhou 510632, China
[3]Division of Environment, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China
[4]Atmospheric Research Centre, Fok Ying Tung Graduate School, Hong Kong University of Science and Technology, Nansha, China
[5]Department of Chemistry, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Correspondence: Cheng Wu (wucheng.vip@foxmail.com) and Jian Zhen Yu (jian.yu@ust.hk)

Received: 15 August 2017 – Discussion started: 27 September 2017
Revised: 6 February 2018 – Accepted: 6 February 2018 – Published: 2 March 2018

Abstract. Linear regression techniques are widely used in atmospheric science, but they are often improperly applied due to lack of consideration or inappropriate handling of measurement uncertainty. In this work, numerical experiments are performed to evaluate the performance of five linear regression techniques, significantly extending previous works by Chu and Saylor. The five techniques are ordinary least squares (OLS), Deming regression (DR), orthogonal distance regression (ODR), weighted ODR (WODR), and York regression (YR). We first introduce a new data generation scheme that employs the Mersenne twister (MT) pseudorandom number generator. The numerical simulations are also improved by (a) refining the parameterization of nonlinear measurement uncertainties, (b) inclusion of a linear measurement uncertainty, and (c) inclusion of WODR for comparison. Results show that DR, WODR and YR produce an accurate slope, but the intercept by WODR and YR is overestimated and the degree of bias is more pronounced with a low R^2 XY dataset. The importance of a properly weighting parameter λ in DR is investigated by sensitivity tests, and it is found that an improper λ in DR can lead to a bias in both the slope and intercept estimation. Because the λ calculation depends on the actual form of the measurement error, it is essential to determine the exact form of measurement error in the XY data during the measurement stage. If a priori error in one of the variables is unknown, or the measurement error described cannot be trusted, DR, WODR and YR can provide the least biases in slope and intercept among all tested regression techniques. For these reasons, DR, WODR and YR are recommended for atmospheric studies when both X and Y data have measurement errors. An Igor Pro-based program (Scatter Plot) was developed to facilitate the implementation of error-in-variables regressions.

1 Introduction

Linear regression is heavily used in atmospheric science to derive the slope and intercept of XY datasets. Examples of linear regression applications include primary OC (organic carbon) and EC (elemental carbon) ratio estimation (Turpin and Huntzicker, 1995; Lin et al., 2009), MAE (mass absorption efficiency) estimation from light absorption and EC mass (Moosmüller et al., 1998), source apportionment of polycyclic aromatic hydrocarbons using CO and NO_x as combustion tracers (Lim et al., 1999), gas-phase reaction rate determination (Brauers and Finlayson-Pitts, 1997), inter-instrument comparison (Bauer et al., 2009; Cross et al., 2010; von Bobrutzki et al., 2010; Zieger et al., 2011; Wu et al., 2012; Huang et al., 2014; Zhou et al., 2016), inter-species analysis (Yu et al., 2005; Kuang et al., 2015), analytical protocol comparison (Chow et al., 2001, 2004; Cheng et al., 2011; Wu et al., 2016), light extinction budget reconstruction (Malm et al., 1994; Watson, 2002; Li et al., 2017), com-

parison between modeling and measurement (Petäjä et al., 2009), emission factor study (Janhäll et al., 2010), retrieval of shortwave cloud forcing (Cess et al., 1995), calculation of pollutant growth rate (Richter et al., 2005), estimation of ground-level $PM_{2.5}$ from MODIS data (Wang and Christopher, 2003), distinguishing OC origin from biomass burning using K^+ as a tracer (Duan et al., 2004) and emission type identification by the EC / CO ratio (Chen et al., 2001).

Ordinary least squares (OLS) regression is the most widely used method due to its simplicity. In OLS, it is assumed that independent variables are error-free. This is the case for certain applications, such as determining a calibration curve of an instrument in analytical chemistry. For example, a known amount of analyte (e.g., through weighing) can be used to calibrate the instrument output response (e.g., voltage). However, in many other applications, such as inter-instrument comparison, X and Y (from two instruments) may have comparable degrees of uncertainty. This deviation from the underlying assumption in OLS would produce biased slope and intercept when OLS is applied to the dataset.

To overcome the drawback of OLS, a number of error-in-variables regression models (also known as bivariate fittings; Cantrell, 2008) or total least-squares methods (Markovsky and Van Huffel, 2007) arise. Deming (1943) proposed an approach by minimizing sum of squares of X and Y residuals. A closed-form solution of Deming regression (DR) was provided by York (1966). Method comparison work of various regression techniques by Cornbleet and Gochman (1979) found significant error in OLS slope estimation when the relative standard deviation (RSD) of measurement error in "X" exceeded 20 %, while DR was found to reach a more accurate slope estimation. In an early application of the EC tracer method, Turpin and Huntzicker (1995) realized the limitation of OLS since OC and EC have comparable measurement uncertainty and thus recommended the use of DR for $(OC / EC)_{pri}$ (primary OC to EC ratio) estimation. Ayers (2001) conducted a simple numerical experiment and concluded that reduced major axis regression (RMA) is more suitable for air quality data regression analysis. Linnet (1999) pointed out that when applying DR for inter-method (or inter-instrument) comparison, special attention should be paid to the sample size. If the range ratio (max / min) is relatively small (e.g., less than 2), more samples are needed to obtain statistically significant results.

In principle, a best-fit regression line should have greater dependence on the more precise data points rather than the less reliable ones. Chu (2005) performed a comparison study of OLS and DR specifically focusing on the EC tracer method application and found that the slope estimated by DR is closer to the correct value than OLS but may still overestimate the ideal value. Saylor et al. (2006) extended the comparison work of Chu (2005) by including a regression technique developed by York et al. (2004). They found that the slope overestimation by DR in the study of Chu (2005) was due to improper configuration of the weighting parameter, λ. This λ value is the key to handling the uneven errors between data points for the best-fit line calculation. This example demonstrates the importance of appropriate weighting in the calculation of best-fit line for error-in-variables regression model, which is overlooked in many studies.

In this study, we extend the work by Saylor et al. (2006) to achieve four objectives. The first is to propose a new data generation scheme by applying the Mersenne twister (MT) pseudorandom number generator for evaluation of linear regression techniques. In the study of Chu (2005), data generation is achieved by a varietal sine function, which has limitations in sample size, sample distribution, and nonadjustable correlation (R^2) between X and Y. In comparison, the MT data generation provides more flexibility, permitting adjustable sample size, XY correlation and distribution. The second is to develop a nonlinear measurement error parameterization scheme for use in the regression method. The third is to incorporate linear measurement errors in the regression methods. In the work by Chu (2005) and Saylor et al. (2006), the relative measurement uncertainty (γ_{Unc}) is nonlinear with concentration, but a constant γ_{Unc} is often applied on atmospheric instruments due to its simplicity. The fourth is to include weighted orthogonal distance regression (WODR) for comparison. Abbreviations and symbols used in this study are summarized in Table B1 for quick reference.

2 Description of regression techniques compared in this study

Ordinary least squares (OLS) method

OLS only considers the errors in dependent variables (Y). OLS regression is achieved by minimizing the sum of squares (S) in the Y residuals (i.e., distance of AB in Fig. S1 in the Supplement):

$$S = \sum_{i=1}^{N} (y_i - Y_i)^2, \tag{1}$$

where Y_i are observed Y data points, while y_i are regressed Y data points of the regression line. N represents the number of data points that is used for regression.

Orthogonal distance regression (ODR)

ODR minimizes the sum of the squared orthogonal distances from all data points to the regressed line and considers equal error variances (i.e., distance of AC in Fig. S1):

$$S = \sum_{i=1}^{N} \left[(x_i - X_i)^2 + (y_i - Y_i)^2 \right]. \tag{2}$$

Weighted orthogonal distance regression (WODR)

Unlike ODR, which considers even error in X and Y, weightings based on measurement errors in both X and Y are considered in WODR when minimizing the sum of squared orthogonal distance from the data points to the regression line (Carroll and Ruppert, 1996) as shown by AD in Fig. S1:

$$S = \sum_{i=1}^{N} \left[(x_i - X_i)^2 + (y_i - Y_i)^2 / \eta \right], \quad (3)$$

where η is the error variance ratio, which determines the angle θ shown in Fig. S1. Implementation of ODR and WODR in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) was done by the computer routine ODRPACK95 (Boggs et al., 1989; Zwolak et al., 2007).

Deming regression (DR)

Deming (1943) proposed the following function to minimize both the X and Y residuals as shown by AD in Fig. S1,

$$S = \sum_{i=1}^{N} \left[\omega(X_i)(x_i - X_i)^2 + \omega(Y_i)(y_i - Y_i)^2 \right], \quad (4)$$

where X_i and Y_i are observed data points and x_i and y_i are regressed data points. Individual data points are weighted based on errors in X_i and Y_i,

$$\omega(X_i) = \frac{1}{\sigma_{X_i}^2}, \ \omega(Y_i) = \frac{1}{\sigma_{Y_i}^2}, \quad (5)$$

where σ_{X_i} and σ_{Y_i} are the standard deviation of the error in measurement of X_i and Y_i, respectively. The closed-form solutions for slope and intercept of DR are shown in Appendix A.

York regression (YR)

The York method (York et al., 2004) introduces the correlation coefficient of errors in X and Y into the minimization function.

$$S = \sum_{i=1}^{N} \left[\omega(X_i)(x_i - X_i)^2 - 2 r_i \sqrt{\omega(X_i)\omega(Y_i)} \right.$$
$$\left. (x_i - X_i)(y_i - Y_i) + \omega(Y_i)(y_i - Y_i)^2 \right] \frac{1}{1 - r_i^2}, \quad (6)$$

where r_i is the correlation coefficient between measurement errors in X_i and Y_i. The slope and intercept of YR are calculated iteratively through the formulas in Appendix A.

Summary of the five regression techniques is given in Table S1 in the Supplement. It is worth noting that OLS and DR have closed-form expressions for calculating slope and intercept. In contrast, ODR, WODR and YR need to be solved iteratively. This need to be taken into consideration when choosing regression algorithm for handling huge numbers of data.

A computer program (Scatter Plot; Wu, 2017a) with a graphical user interface (GUI) in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) was developed to facilitate the implementation of error-in-variables regression (including DR, WODR and YR). Two other Igor Pro-based computer programs, Histbox (Wu, 2017b) and Aethalometer data processor (Wu, 2017c), are used for data analysis and visualization in this study.

3 Data description

Two types of data are used for regression comparison. The first type is synthetic data generated by computer programs, which can be used in the EC tracer method (Turpin and Huntzicker, 1995) to demonstrate the regression application. The true "slope" and "intercept" are assigned during data generation, allowing quantitative comparison of the bias of each regression scheme. The second type of data comes from ambient measurement of light absorption, OC and EC in Guangzhou for demonstration in a real-world application.

3.1 Synthetic XY data generation

In this study, numerical simulations are conducted in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) through custom codes. Two types of generation schemes are employed: one is based on the MT pseudorandom number generator (Matsumoto and Nishimura, 1998) and the other is based on the sine function described by Chu (2005).

The general form of linear regression on XY data can be written as

$$Y = kX + b, \quad (7)$$

where k is the regressed slope and b is the intercept. The underlying meaning is that, Y can be decomposed into two parts. One part is correlated with X, and the ratio is defined by k. The other part of Y is constant and independent of X and regarded as b.

To make the discussion easier to follow, we intentionally avoid discussion using the abstract general form and instead opt to use a real-world application case in atmospheric science. Linear regression had been heavily applied on OC and EC data, here we use OC and EC data as an example to demonstrate the regression application in atmospheric science. In the EC tracer method, OC (mixture) is Y and EC (tracer) is X. OC can be decomposed into three components based on their formation pathway:

$$\mathrm{OC} = \mathrm{POC_{comb}} + \mathrm{POC_{non\text{-}comb}} + \mathrm{SOC}, \quad (8)$$

where $\mathrm{POC_{comb}}$ is primary OC from combustion. $\mathrm{POC_{non\text{-}comb}}$ is primary OC emitted from non-combustion activities. SOC is secondary OC formed during atmospheric aging. Since $\mathrm{POC_{comb}}$ is co-emitted with EC and well correlated with each other, their relationship can be parameterized

as

$$\mathrm{POC_{comb}} = (\mathrm{OC}/\mathrm{EC})_{\mathrm{pri}} \times \mathrm{EC}. \quad (9)$$

By carefully selecting an OC and EC subset when SOC is very low (considered as approximately zero), the combination of Eqs. (8) and (9) becomes

$$\mathrm{POC} = (\mathrm{OC}/\mathrm{EC})_{\mathrm{pri}} \times \mathrm{EC} + \mathrm{POC_{non\text{-}comb}}. \quad (10)$$

The regressed slope of POC (Y) against EC (X) represents $(\mathrm{OC}/\mathrm{EC})_{\mathrm{pri}}$ (k in Eq. 7). The regressed intercept become $\mathrm{POC_{non\text{-}comb}}$ (b in Eq. 7). With known $(\mathrm{OC}/\mathrm{EC})_{\mathrm{pri}}$ and $\mathrm{POC_{non\text{-}comb}}$, SOC can be estimated by

$$\mathrm{SOC} = \mathrm{OC} - ((\mathrm{OC}/\mathrm{EC})_{\mathrm{pri}} \times \mathrm{EC} + \mathrm{POC_{non\text{-}comb}}). \quad (11)$$

The data generation starts from EC (X values). Once EC is generated, $\mathrm{POC_{comb}}$ (the part of Y that is correlated with X) can be obtained by multiplying EC by a preset constant, $(\mathrm{OC}/\mathrm{EC})_{\mathrm{pri}}$ (slope k). Then the other preset constant $\mathrm{POC_{non\text{-}comb}}$ is added to $\mathrm{POC_{comb}}$ and the sum becomes POC (Y values). To simulate the real-world situation, measurement errors are added on X and Y values. Details of synthesized measurement error are discussed in the next section. Implementation of data generation by two types of mathematical schemes is explained in Sect. 3.1.2 and 3.1.3, respectively.

3.1.1 Parameterization of synthesized measurement uncertainty

Weighting of variables is a crucial input for errors-in-variables linear regression methods such as DR, YR and WODR. In practice, the weights are usually defined as the inverse of the measurement error variance (Eq. 5). When measurement errors are considered, measured concentrations ($\mathrm{Conc._{measured}}$) are simulated by adding measurement uncertainties ($\varepsilon_{\mathrm{Conc.}}$) to the true concentrations ($\mathrm{Conc._{true}}$):

$$\mathrm{Conc._{measured}} = \mathrm{Conc._{true}} + \varepsilon_{\mathrm{Conc.}}, \quad (12)$$

where $\varepsilon_{\mathrm{Conc.}}$ is the random error following an even distribution with an average of 0, the range of which is constrained by

$$-\gamma_{\mathrm{Unc}} \times \mathrm{Conc._{true}} \le \varepsilon_{\mathrm{Conc.}} \le +\gamma_{\mathrm{Unc}} \times \mathrm{Conc._{true}}. \quad (13)$$

The γ_{Unc} is a dimensionless factor that describes the fractional measurement uncertainty relative to the true concentration ($\mathrm{Conc._{true}}$). γ_{Unc} could be a function of $\mathrm{Conc._{true}}$ (Thompson, 1988) or a constant. The term $\gamma_{\mathrm{Unc}} \times \mathrm{Conc._{true}}$ defines the boundary of random measurement errors.

Two types of measurement error are considered in this study. The first type is $\gamma_{\mathrm{Unc-nonlinear}}$. In the data generation scheme of Chu (2005) for the measurement uncertainties ($\varepsilon_{\mathrm{POC}}$ and $\varepsilon_{\mathrm{EC}}$), $\gamma_{\mathrm{Unc-nonlinear}}$ is nonlinearly related to $\mathrm{Conc._{true}}$:

$$\gamma_{\mathrm{Unc-nonlinear}} = \frac{1}{\sqrt{\mathrm{Conc._{true}}}}, \quad (14)$$

and thus Eq. (13) for POC and EC becomes

$$-\frac{1}{\sqrt{\mathrm{POC_{true}}}} \times \mathrm{POC_{true}} \le \varepsilon_{\mathrm{POC}} \le +\frac{1}{\sqrt{\mathrm{POC_{true}}}} \times \mathrm{POC_{true}}, \quad (15)$$

$$-\frac{1}{\sqrt{\mathrm{EC_{true}}}} \times \mathrm{EC_{true}} \le \varepsilon_{\mathrm{EC}} \le +\frac{1}{\sqrt{\mathrm{EC_{true}}}} \times \mathrm{EC_{true}}. \quad (16)$$

In Eq. (14), the γ_{Unc} decreases as concentration increases, since low concentrations are usually more challenging to measure. As a result, the $\gamma_{\mathrm{Unc-nonlinear}}$ defined in Eq. (14) is more realistic than the constant approach, but there are two limitations. First, the physical meaning of the uncertainty unit is lost. If the unit of OC is $\mu g\,m^{-3}$, then the unit of $\varepsilon_{\mathrm{OC}}$ becomes $\sqrt{\mu g\,m^{-3}}$. Second, the concentration is not normalized by a consistent relative value, making it sensitive to the X and Y units used. For example, if $\mathrm{POC_{true}} = 0.9\,\mu g\,m^{-3}$, then $\varepsilon_{\mathrm{POC}} = \pm 0.95\,\mu g\,m^{-3}$ and $\gamma_{\mathrm{Unc}} = 105\,\%$, but by changing the concentration unit to $\mathrm{POC_{true}} = 900\,ng\,m^{-3}$, $\varepsilon_{\mathrm{OC}} = \pm 30\,ng\,m^{-3}$ and $\gamma_{\mathrm{Unc}} = 3\,\%$. To overcome these deficiencies, we propose to modify Eq. (14) to

$$\gamma_{\mathrm{Unc}} = \sqrt{\frac{\mathrm{LOD}}{\mathrm{Conc._{true}}}} \times \alpha, \quad (17)$$

where LOD (limit of detection) is introduced to generate a dimensionless γ_{Unc}. α is a dimensionless adjustable factor to control the position of γ_{Unc} curve on the concentration axis, which is indicated by the value of γ_{Unc} at LOD level. As shown in Fig. 1a, at different values of α ($\alpha = 1$, 0.5 and 0.3), the corresponding γ_{Unc} at the same LOD level would be 100, 50 and 30 %, respectively. By changing α, the location of the γ_{Unc} curve on x axis direction can be set, using the γ_{Unc} at LOD as the reference point. Then Eq. (7) for POC and EC becomes

$$-\sqrt{\frac{\mathrm{LOD_{POC}}}{\mathrm{POC_{true}}}} \times \alpha_{\mathrm{POC}} \times \mathrm{POC_{true}} \le \varepsilon_{\mathrm{POC}} \le +\sqrt{\frac{\mathrm{LOD_{POC}}}{\mathrm{POC_{true}}}} \times \alpha_{\mathrm{POC}} \times \mathrm{POC_{true}}, \quad (18)$$

$$-\sqrt{\frac{\mathrm{LOD_{EC}}}{\mathrm{EC_{true}}}} \times \alpha_{\mathrm{EC}} \times \mathrm{EC_{true}} \le \varepsilon_{\mathrm{EC}} \le +\sqrt{\frac{\mathrm{LOD_{EC}}}{\mathrm{EC_{true}}}} \times \alpha_{\mathrm{EC}} \times \mathrm{EC_{true}}. \quad (19)$$

With the modified $\gamma_{\mathrm{Unc-nonlinear}}$ parameterization, concentrations of POC and EC are normalized by a corresponding LOD, which maintains unit consistency between $\mathrm{POC_{true}}$ and $\varepsilon_{\mathrm{POC}}$ and $\mathrm{EC_{true}}$ and $\varepsilon_{\mathrm{EC}}$ and eliminates dependency on the concentration unit.

Uniform distribution has been used in previous studies (Cox et al., 2003; Chu, 2005; Saylor et al., 2006) and is adopted in this study to parameterize measurement error. For

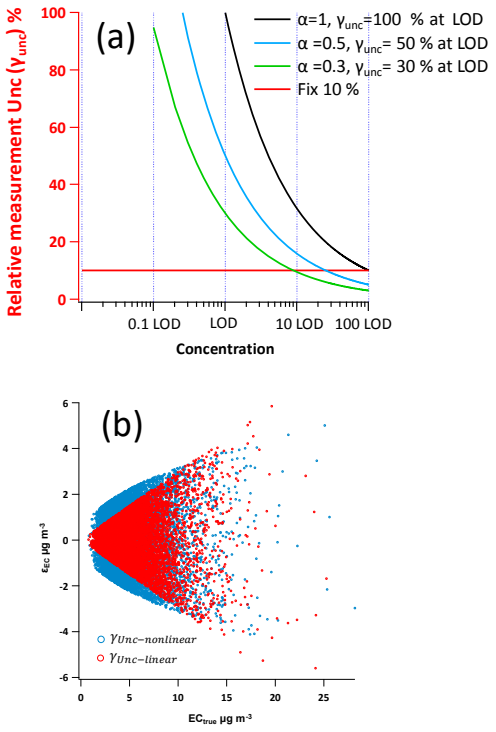


Figure 1. (a) Example $\gamma_{\mathrm{Unc-nonlinear}}$ curves by different α values (Eq. 17). The x axis is concentration (normalized by LOD) in log scale and the y axis is γ_{Unc}. Black, blue and green line represents α equal to 1, 0.5 and 0.3, respectively, corresponding to the $\gamma_{\mathrm{Unc-nonlinear}}$ at LOD level equals to 100, 50 and 30 %, respectively. The red line represents $\gamma_{\mathrm{Unc-linear}}$ of 10 %. **(b)** Example of measurement uncertainty generation of $\gamma_{\mathrm{Unc-nonlinear}}$ and $\gamma_{\mathrm{Unc-linear}}$. The blue circles represent $\gamma_{\mathrm{Unc-nonlinear}}$ following Eq. (17) ($\mathrm{LOD_{EC}} = 1$, $a_{\mathrm{EC}} = 1$). The red circles represent $\gamma_{\mathrm{Unc-linear}}$ (30 %).

a uniform distribution in the interval $[a, b]$, the variance is $\frac{1}{12}(a-b)^2$. Since $\varepsilon_{\mathrm{POC}}$ and $\varepsilon_{\mathrm{EC}}$ follow a uniform distribution in the interval as given by Eqs. (18) and (19), the weights in DR and YR (inverse of variance) become

$$\omega(X_i) = \frac{1}{\sigma_{X_i}^2} = \frac{3}{\mathrm{EC_{true}} \times \mathrm{LOD_{EC}} \times \alpha_{\mathrm{EC}}^2}, \quad (20)$$

$$\omega(Y_i) = \frac{1}{\sigma_{Y_i}^2} = \frac{3}{\mathrm{POC_{true}} \times \mathrm{LOD_{POC}} \times \alpha_{\mathrm{POC}}^2}. \quad (21)$$

The parameter λ in Deming regression is then determined:

$$\lambda = \frac{\omega(X_i)}{\omega(Y_i)} = \frac{\mathrm{POC_{true}} \times \mathrm{LOD_{POC}} \times \alpha_{\mathrm{POC}}^2}{\mathrm{EC_{true}} \times \mathrm{LOD_{EC}} \times \alpha_{\mathrm{EC}}^2}. \quad (22)$$

Besides the $\gamma_{\mathrm{Unc-nonlinear}}$ discussed above, a second type measurement uncertainty parameterized by a constant proportional factor, $\gamma_{\mathrm{Unc-linear}}$, is very common in atmospheric applications:

$$-\gamma_{\mathrm{POCunc}} \times \mathrm{POC_{true}} \le \varepsilon_{\mathrm{POC}} \le +\gamma_{\mathrm{POCunc}} \times \mathrm{POC_{true}}, \quad (23)$$

$$-\gamma_{\mathrm{ECunc}} \times \mathrm{EC_{true}} \le \varepsilon_{\mathrm{EC}} \le +\gamma_{\mathrm{ECunc}} \times \mathrm{EC_{true}}. \quad (24)$$

where γ_{POCunc} and γ_{ECunc} are the relative measurement uncertainties, e.g., for relative measurement uncertainty of 10 %, $\gamma_{\mathrm{Unc}} = 0.1$. As a result, the measurement error is linearly proportional to the concentration. An example comparison of $\gamma_{\mathrm{Unc-nonlinear}}$ and $\gamma_{\mathrm{Unc-linear}}$ is shown in Fig. 1b. For $\gamma_{\mathrm{Unc-linear}}$, the weights become

$$\omega(X_i) = \frac{1}{\sigma_{X_i}^2} = \frac{3}{(\gamma_{\mathrm{ECunc}} \times \mathrm{EC_{true}})^2}, \quad (25)$$

$$\omega(Y_i) = \frac{1}{\sigma_{Y_i}^2} = \frac{3}{(\gamma_{\mathrm{POCunc}} \times \mathrm{POC_{true}})^2}, \quad (26)$$

and λ for Deming regression can be determined:

$$\lambda = \frac{\omega(X_i)}{\omega(Y_i)} = \frac{(\gamma_{\mathrm{POCunc}} \times \mathrm{POC_{true}})^2}{(\gamma_{\mathrm{ECunc}} \times \mathrm{EC_{true}})^2}. \quad (27)$$

3.1.2 *XY* data generation by the Mersenne twister generator following a specific distribution

The Mersenne twister (MT) is a pseudorandom number generator (PRNG) developed by Matsumoto and Nishimura (1998). MT has been widely adopted by mainstream numerical analysis software (e.g., MATLAB, SPSS, SAS and Igor Pro) as well as popular programing languages (e.g., R, Python, IDL, C++ and PHP). Data generation using MT provides a few advantages: (1) frequency distribution can be easily assigned during the data generation process, allowing straightforward simulation of the frequency distribution characteristics (e.g., Gaussian or lognormal) observed in ambient measurements; (2) the inputs for data generation are simply the mean and standard deviation of the data series and can be changed easily by the user; (3) the correlation (R^2) between X and Y can be manipulated easily during the data generation to satisfy various purposes; and (4) unlike the sine function described by Chu (2005), which has a sample size limitation of 120, the sample size in MT data generation is highly flexible.

In this section, we will use POC as Y and EC as X as an example to explain the data generation. Procedure of applying MT to simulate ambient POC and EC data can be found in our previous study (Wu and Yu, 2016). Details of the data generation steps are shown in Fig. 2 and described below. The first step is generation of $\mathrm{EC_{true}}$ by MT. In our previous study, it was found that ambient POC and EC data follow a lognormal distribution in various locations of the Pearl River Delta (PRD) region. Therefore, lognormal distributions are adopted during $\mathrm{EC_{true}}$ generation. A range of average concentration and relative standard deviation (RSD) from ambient samples is considered in formulating the lognormal

distribution. The second step is to generate POC_{comb}. As shown in Fig. 2, POC_{comb} is generated by multiplying EC_{true} with $(OC / EC)_{pri}$. Instead of having a Gaussian distribution, $(OC / EC)_{pri}$ in this study is a single value, which favors direct comparison between the true value of $(OC / EC)_{pri}$ and $(OC / EC)_{pri}$ estimated from the regression slope. The third step is generation of POC_{true} by adding $POC_{non\text{-}comb}$ onto POC_{comb}. Instead of having a distribution, $POC_{non\text{-}comb}$ in this study is a single value, which favors direct comparison between the true value of $POC_{non\text{-}comb}$ and $POC_{non\text{-}comb}$ estimated from the regression intercept. The fourth step is to compute ε_{POC} and ε_{EC}. As discussed in Sect. 3.1.1, two types of measurement errors are considered for ε_{POC} and ε_{EC} calculation: $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$. In the last step, $POC_{measured}$ and $EC_{measured}$ are calculated following Eq. (12), i.e., applying measurement errors on POC_{true} and EC_{true}. Then $POC_{measured}$ and $EC_{measured}$ can be used as Y and X, respectively, to test the performance of various regression techniques. An Igor Pro-based program with a GUI was developed to facilitate the MT data generation for OC and EC. A brief introduction is given in the Supplement.

3.1.3 *XY* data generation by the sine function of Chu (2005)

Besides MT, inclusion of the sine function data generation scheme in this study mainly serves two purposes. First, the sine function scheme was adopted in two previous studies (Chu, 2005; Saylor et al., 2006), the inclusion of this scheme can help to verify whether the codes in Igor for various regression approaches yield the same results from the two previous studies. Second, the crosscheck between results from sine function and MT provides circumstantial evidence that the MT scheme works as expected.

In this section, XY data generation by sine functions is demonstrated using POC as Y and EC as X. There are four steps in POC and EC data generation as shown by the flowchart in Fig. S2. Details are explained as follows. (1) The first step is to generate POC and EC (Chu, 2005):

$$POC_{comb} = 14 + 12\left(\sin\left(\frac{x}{\tau}\right) + \sin(x - \phi)\right), \quad (28)$$

$$EC_{true} = 3.5 + 3\left(\sin\left(\frac{x}{\tau}\right) + \sin(x - \phi)\right), \quad (29)$$

where x is the elapsed hour ($x = 1, 2, 3 \ldots n$; $n \leq 120$), τ is used to adjust the width of each peak, and ϕ is used to adjust the phase of the sine wave. The constants 14 and 3.5 are used to lift the sine wave to the positive range of the y axis. An example of data generation by the sine functions of Chu (2005) is shown in Fig. 3. Dividing Eq. (28) by Eq. (29) yields a value of 4. In this way the exact relation between POC and EC is defined clearly as $(OC / EC)_{pri} = 4$. (2) With POC_{comb} and EC_{true} generated, the second step is to add $POC_{non\text{-}comb}$ to POC_{comb} to compute POC_{true}. As for $POC_{non\text{-}comb}$, a single value is assigned and added to all

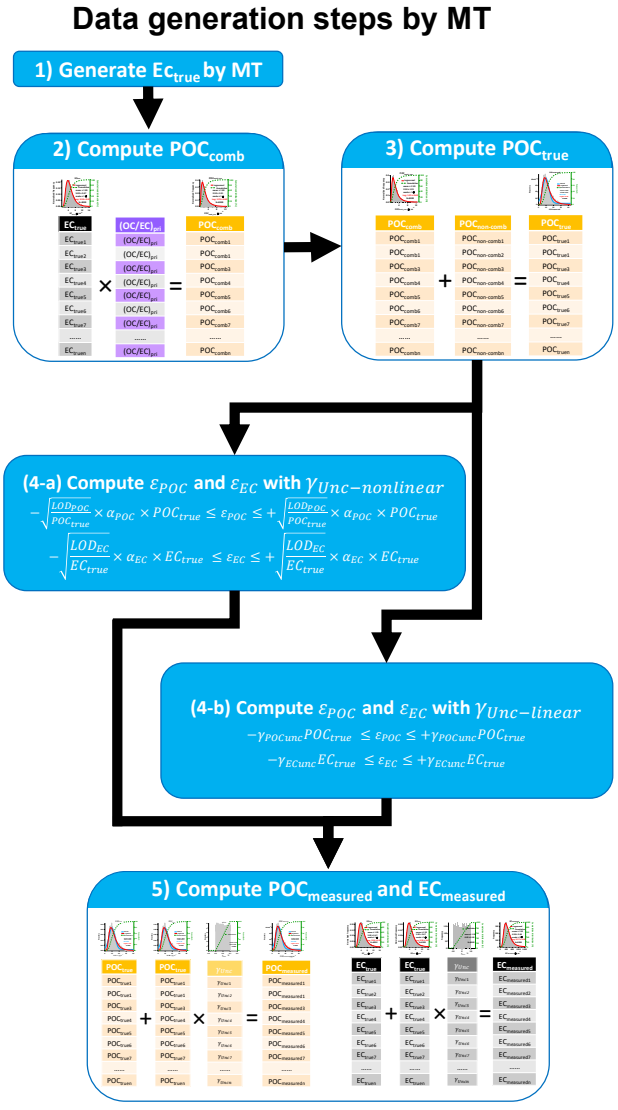


Figure 2. Flowchart of data generation steps using MT.

POC following Eq. (10). Then the goodness of the regression intercept can be evaluated by comparing the regressed intercept with preset $POC_{non\text{-}comb}$. (3) The third step is to compute ε_{POC} and ε_{EC}, considering both $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$. (4) The last step is to apply measurement errors on POC_{true} and EC_{true} following Eq. (12). Then $POC_{measured}$ and $EC_{measured}$ can be used as Y and X, respectively, to evaluate the performance of various regression techniques.

3.2 Ambient measurement of σ_{abs} and EC

Sampling was conducted from Feb 2012 to Jan 2013 at the suburban Nancun (NC) site (23°0′11.82″ N, 113°21′18.04″ E), which is situated on top of the highest peak (141 m a.s.l.) in the Panyu district of Guangzhou. This site is located at the geographic center of Pearl River

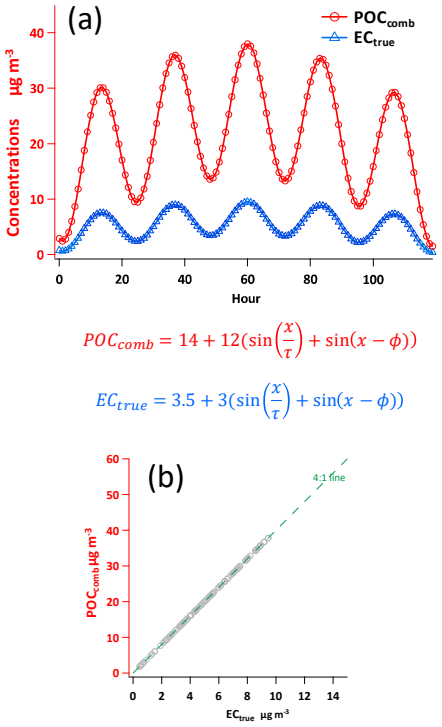


Figure 3. POC_{comb} and EC_{true} data generated by the sine functions of Chu (2005). **(a)** Time series of the 120 data points for POC_{comb} and EC_{true}. **(b)** Scatter plot of POC_{comb} vs. EC_{true}.

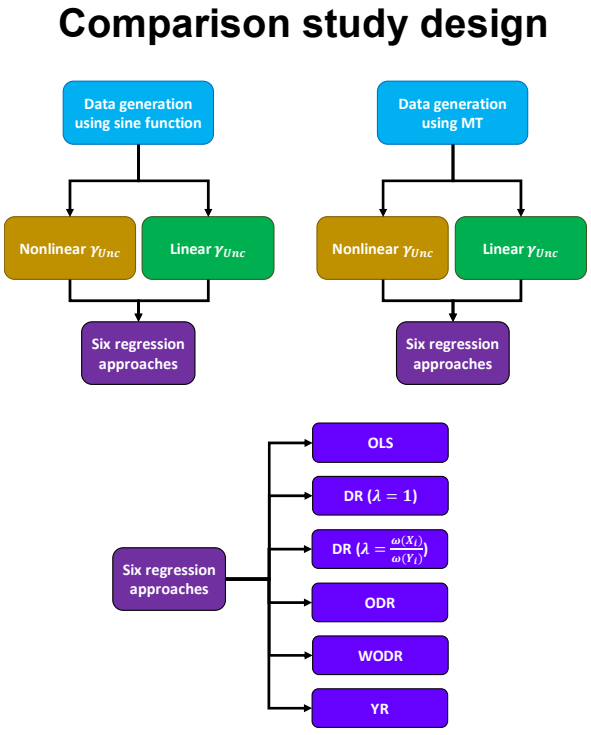


Figure 4. Overview of the comparison study design.

Delta region (PRD), making it a good location for representing the average atmospheric mixing characteristics of city clusters in the PRD region. Light absorption measurements were performed by a 7λ Aethalometer (AE-31, Magee Scientific Company, Berkeley, CA, USA). EC mass concentrations were measured by a real time ECOC analyzer (model RT-4, Sunset Laboratory Inc., Tigard, Oregon, USA). Both instruments utilized inlets with a 2.5 µm particle diameter cutoff. The algorithm of Weingartner et al. (2003) was adopted to correct the sampling artifacts (aerosol loading, filter matrix and scattering effect; Collaud Coen et al., 2010) in Aethalometer measurement. A customized computer program with GUI, Aethalometer data processor (Wu et al., 2018), was developed to perform the data correction and detailed descriptions can be found in https://sites.google.com/site/wuchengust. More details of the measurements can be found in Wu et al. (2018).

4 Comparison study using synthetic data

In the following comparisons, six regression approaches are compared using two data generation schemes (Chu sine function and MT) separately, as illustrated in Fig. 4. Each data generation scheme considers both $\gamma_{\mathrm{Unc-nonlinear}}$ and $\gamma_{\mathrm{Unc-linear}}$ in measurement error parameterization. In total, 18 cases are tested with different combination of data generation schemes, measurement error parameterization schemes, true slope and intercept settings. In each case, six regression approaches are tested, i.e., OLS, DR ($\lambda = 1$), DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$, ODR, WODR and YR. In commercial software (e.g., OriginPro®, SigmaPlot®, GraphPad Prism®), λ in DR is set to 1 by default if not specified. As indicated by Saylor et al. (2006), the bias observed in the study of Chu (2005) is likely due to $\lambda = 1$ in DR. The purpose of including DR ($\lambda = 1$) in this study is to examine the potential bias using the default input in many software products. The six regression approaches are considered to examine the sensitivity of regression results to various parameters used in data generation. For each case, 5000 runs are performed to obtain statistically significant results, as recommended by Saylor et al. (2006). The mean slope and intercept from 5000 runs is compared with the true value assigned during data generation. If the difference is < 5 %, the result is considered unbiased.

4.1 Comparison results using the dataset of Chu (2005)

In this section, the scheme of Chu (2005) is adopted for data generation to obtain a benchmark of six regression approaches. With different setup of slope, intercept and γ_{Unc}, six cases (Cases 1–6) are studied and the results are discussed below.

Table 1. Summary of six regression approaches comparison with 5000 runs for 18 cases.

Data generation						Results by different regression approaches											
						OLS		DR $\lambda = 1$		DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$		ODR		WODR		YR	
Case	Data scheme	True Slope	True Intercept	R^2 (X, Y)	Measurement error	slope	intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept
1		4	0	0.67 ± 0.03	$LOD_{POC} = 1$, $LOD_{EC} = 1$,	2.94 ± 0.14	5.84 ± 0.78	4.27 ± 0.27	-1.45 ± 1.36	4.01 ± 0.25	-0.04 ± 1.28	4.27 ± 0.27	-1.45 ± 1.36	3.98 ± 0.22	1.12 ± 1.02	3.98 ± 0.22	1.12 ± 1.02
2		4	3	0.67 ± 0.04	$a_{POC} = 1$, $a_{EC} = 1$.	2.95 ± 0.15	8.83 ± 0.80	4.32 ± 0.28	1.28 ± 1.43	4.01 ± 0.26	2.94 ± 1.34	4.32 ± 0.28	1.28 ± 1.43	3.99 ± 0.23	3.98 ± 1.05	3.99 ± 0.23	3.98 ± 1.05
3	Chu	4	0	0.95 ± 0.01	$LOD_{POC} = 0.5$, $LOD_{EC} = 0.5$, $a_{POC} = 0.5$, $a_{EC} = 0.5$.	3.83 ± 0.08	0.95 ± 0.40	4.03 ± 0.09	-0.18 ± 0.44	4 ± 0.09	0 ± 0.44	4.03 ± 0.09	-0.18 ± 0.44	4 ± 0.08	0.12 ± 0.37	4 ± 0.08	0.12 ± 0.37
4		4	0	0.78 ± 0.02	$LOD_{POC} = 1$, $LOD_{EC} = 0.5$, $a_{POC} = 1$, $a_{EC} = 1$.	3.39 ± 0.15	3.34 ± 0.75	4.3 ± 0.21	-1.66 ± 1.06	4 ± 0.19	-0.03 ± 0.99	4.3 ± 0.21	-1.66 ± 1.06	4 ± 0.17	0.33 ± 0.81	4 ± 0.17	0.33 ± 0.81
5		4	0	0.69 ± 0.04	$\gamma_{Unc} = 30\%$	3.32 ± 0.20	3.77 ± 0.90	4.75 ± 0.30	-4.14 ± 1.36	4.01 ± 0.25	-0.04 ± 1.13	4.75 ± 0.30	-4.14 ± 1.36	4 ± 0.18	-0.01 ± 0.59	4 ± 0.18	-0.01 ± 0.59
6		4	3	0.66 ± 0.04		3.31 ± 0.22	6.79 ± 1.02	4.95 ± 0.31	-2.26 ± 1.48	3.99 ± 0.26	3.05 ± 1.22	4.95 ± 0.31	-2.26 ± 1.48	4.01 ± 0.20	2.72 ± 0.74	4.01 ± 0.20	2.72 ± 0.74
7		4	0	0.76 ± 0.01		3.22 ± 0.03	4.3 ± 0.14	4.17 ± 0.04	-0.94 ± 0.18	4 ± 0.03	0 ± 0.17	4.17 ± 0.04	-0.94 ± 0.18	3.96 ± 0.03	1.21 ± 0.13	3.96 ± 0.03	1.21 ± 0.13
8		4	3	0.75 ± 0.01		3.22 ± 0.03	7.29 ± 0.14	4.2 ± 0.04	1.88 ± 0.18	4 ± 0.03	3 ± 0.18	4.2 ± 0.04	1.88 ± 0.18	3.97 ± 0.03	4.11 ± 0.13	3.97 ± 0.03	4.11 ± 0.13
9		0.5	0	0.76 ± 0.01	$LOD_{POC} = 1$, $LOD_{EC} = 1$,	0.43 ± 0.00	0.36 ± 0.02	0.46 ± 0.01	0.23 ± 0.03	0.5 ± 0.01	0 ± 0.03	0.46 ± 0.01	0.23 ± 0.03	0.5 ± 0.00	0 ± 0.01	0.5 ± 0.00	0 ± 0.01
10		0.5	3	0.56 ± 0.01	$a_{POC} = 1$, $a_{EC} = 1$.	0.43 ± 0.01	3.36 ± 0.03	0.5 ± 0.01	3.02 ± 0.04	0.49 ± 0.01	3.05 ± 0.04	0.5 ± 0.01	3.02 ± 0.04	0.51 ± 0.01	2.73 ± 0.03	0.51 ± 0.01	2.73 ± 0.03
11		1	0	0.76 ± 0.01		0.87 ± 0.01	0.72 ± 0.05	1 ± 0.01	0 ± 0.06	1 ± 0.01	0 ± 0.06	1 ± 0.01	0 ± 0.06	1 ± 0.01	0 ± 0.02	1 ± 0.01	0 ± 0.02
12	MT	1	3	0.66 ± 0.01		0.87 ± 0.01	3.72 ± 0.05	1.09 ± 0.01	2.52 ± 0.07	0.99 ± 0.01	3.07 ± 0.06	1.09 ± 0.01	2.52 ± 0.07	1.01 ± 0.01	2.71 ± 0.04	1.01 ± 0.01	2.7 ± 0.04
13		4	0	0.76 ± 0.01	$\gamma_{Unc} = 30\%$	3.48 ± 0.04	2.87 ± 0.18	4.53 ± 0.05	-2.94 ± 0.24	4 ± 0.05	0 ± 0.22	4.53 ± 0.05	-2.94 ± 0.24	4 ± 0.03	0 ± 0.09	4 ± 0.03	0 ± 0.09
14		4	3	0.73 ± 0.01		3.48 ± 0.04	5.87 ± 0.19	4.67 ± 0.05	-0.67 ± 0.26	3.98 ± 0.05	3.08 ± 0.23	4.67 ± 0.05	-0.67 ± 0.26	4.02 ± 0.03	2.68 ± 0.11	4.02 ± 0.03	2.68 ± 0.11
15		0.5	0	0.54 ± 0.01		0.4 ± 0.01	0.55 ± 0.03	0.45 ± 0.01	0.26 ± 0.03	0.5 ± 0.01	0.01 ± 0.03	0.45 ± 0.01	0.26 ± 0.03	0.52 ± 0.01	-0.23 ± 0.02	0.52 ± 0.01	-0.23 ± 0.02
16		0.5	3	0.40 ± 0.01		0.4 ± 0.01	3.54 ± 0.04	0.5 ± 0.01	2.98 ± 0.04	0.5 ± 0.01	3 ± 0.04	0.5 ± 0.01	2.98 ± 0.04	0.52 ± 0.01	2.65 ± 0.04	0.52 ± 0.01	2.65 ± 0.04
17		1	0	0.65 ± 0.01		0.8 ± 0.01	1.07 ± 0.04	1 ± 0.01	0 ± 0.05	1 ± 0.01	0 ± 0.05	1 ± 0.01	0 ± 0.05	1 ± 0.01	0 ± 0.04	1 ± 0.01	0 ± 0.04
18		1	3	0.59 ± 0.01		0.8 ± 0.01	4.07 ± 0.05	1.07 ± 0.01	2.62 ± 0.07	1 ± 0.01	3 ± 0.06	1.07 ± 0.01	2.62 ± 0.07	1.02 ± 0.01	2.84 ± 0.05	1.02 ± 0.01	2.84 ± 0.05

4.1.1 Results with $\gamma_{\text{Unc-nonlinear}}$

A comparison of the regression techniques results with $\gamma_{\text{Unc-nonlinear}}$ (following Eqs. 18 and 19) is summarized in Table 1. LOD_{POC} , LOD_{EC}, α_{POC} and α_{EC} are all set to 1 to reproduce the data studied by Chu (2005) and Saylor et al. (2006). Two sets of true slope and intercept are considered (Case 1: slope = 4, intercept = 0; Case 2: slope = 4, intercept = 3) to examine if any results are sensitive to the nonzero intercept. The R^2 (POC, EC) from 5000 runs for both Case 1 and 2 are 0.67 ± 0.03.

As shown in Fig. 5, for the zero-intercept case (Case 1), OLS significantly underestimates the slope (2.95 ± 0.14), while it overestimates the intercept (5.84 ± 0.78). This result indicates that OLS is not suitable for errors-in-variables linear regression, consistent with similar analysis results from Chu (2005) and Saylor et al. (2006). With DR, if the λ is properly calculated by weights $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$, unbiased slope ($4.01 \pm 0.25$) and intercept ($-0.04 \pm 1.28$) are obtained; however, results from DR with $\lambda = 1$ show obvious bias in the slope (4.27 ± 0.27) and intercept (-1.45 ± 1.36). ODR also produces biased slope (4.27 ± 0.27) and intercept (-1.45 ± 1.36), which are identical to results of DR when $\lambda = 1$. With WODR, unbiased slope (3.98 ± 0.22) is observed, but the intercept is overestimated (1.12 ± 1.02). Results of YR are identical to WODR. For Case 2 (slope = 4, intercept = 3), slopes from all six regression approaches are consistent with Case 1 (Table 1). The Case 2 intercepts are equal to the Case 1 intercepts plus 3, implying that all the regression methods are not sensitive to a nonzero intercept.

For Case 3, $\text{LOD}_{\text{POC}} = 0.5$, $\text{LOD}_{\text{EC}} = 0.5$, $\alpha_{\text{POC}} = 0.5$, $\alpha_{\text{EC}} = 0.5$ are adopted (Table 1), leading to an offset to the left of $\gamma_{\text{Unc-nonlinear}}$ (blue curve) compared to Case 1 and 2 (black curve) in Fig. 1. As a result, for the same concentration of EC and OC in Case 3, the $\gamma_{\text{Unc-nonlinear}}$ is smaller than in Cases 1 and 2 as indicated by a higher R^2 (0.95 ± 0.01 for Case 3, Table 1). With a smaller measurement uncertainty, the degree of bias in Case 3 is smaller than in Case 1. For example, OLS slope is less biased in Case 3 (3.83 ± 0.08) compared to Case 1 (2.94 ± 0.14). Similarly, the slope (4.03 ± 0.09) and intercept (-0.18 ± 0.44) of DR ($\lambda = 1$) exhibit a much smaller bias with a smaller measurement uncertainty, implying that the degree of bias by improperly weighting in DR, WODR and YR is associated with the degree of measurement uncertainty. A higher measurement uncertainty results in larger bias in slope and intercept.

An uneven LOD_{POC} and LOD_{EC} is tested in Case 4 with $\text{LOD}_{\text{POC}} = 1$, $\text{LOD}_{\text{EC}} = 0.5$, $\alpha_{\text{POC}} = 0.5$, $\alpha_{\text{EC}} = 0.5$, which yield an R^2(POC, EC) of 0.78 ± 0.02. The results are similar to Case 1. For DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$ unbiased slope and intercept are obtained. For WODR and YR, unbiased slopes are reported with a small bias in the intercepts. Large bias values are observed in both the slopes and intercepts in Case 4 using OLS, DR ($\lambda = 1$) and ODR.

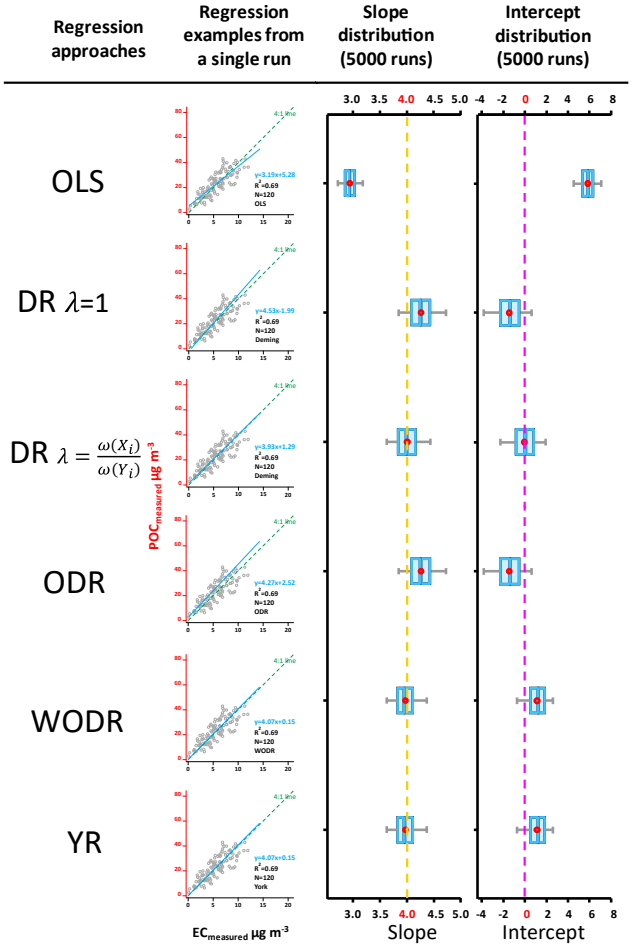


Figure 5. Regression results on synthetic data, Case 1 (Slope = 4, Intercept = 0, $\text{LOD}_{\text{POC}} = 1$, $\text{LOD}_{\text{EC}} = 1$, $a_{\text{POC}} = 1$, $a_{\text{EC}} = 1$, R^2 (POC, EC) = 0.67 ± 0.03). The scatter plots demonstrate regression examples from a single run. The box plots show the distribution of regressed slopes and intercepts from 5000 runs of six regression approaches. The dashed lines in orange and pink represent true slope and intercept, respectively.

4.1.2 Results with $\gamma_{\text{Unc-linear}}$

Cases 5 and 6 represent the results from using $\gamma_{\text{Unc-linear}}$ and are shown in Table 1. γ_{Unc} is set to 30 % to achieve an R^2 (POC, EC) of 0.7, a value close to the R^2 in studies of Chu (2005) and Saylor et al. (2006). In Case 5 (slope = 4, intercept = 0), unbiased slopes and intercepts are determined by DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$, WODR and YR. OLS underestimates the slope (3.32 ± 0.20) and overestimates intercept (3.77 ± 0.90), while DR ($\lambda = 1$) and ODR overestimate the slopes (4.75 ± 0.30) and underestimate the intercepts (-4.14 ± 1.36). In Case 6 (slope = 4, intercept = 3), results similar to Case 5 are obtained. It is worth noting that although the mean intercept (3.05 ± 1.22) of DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$

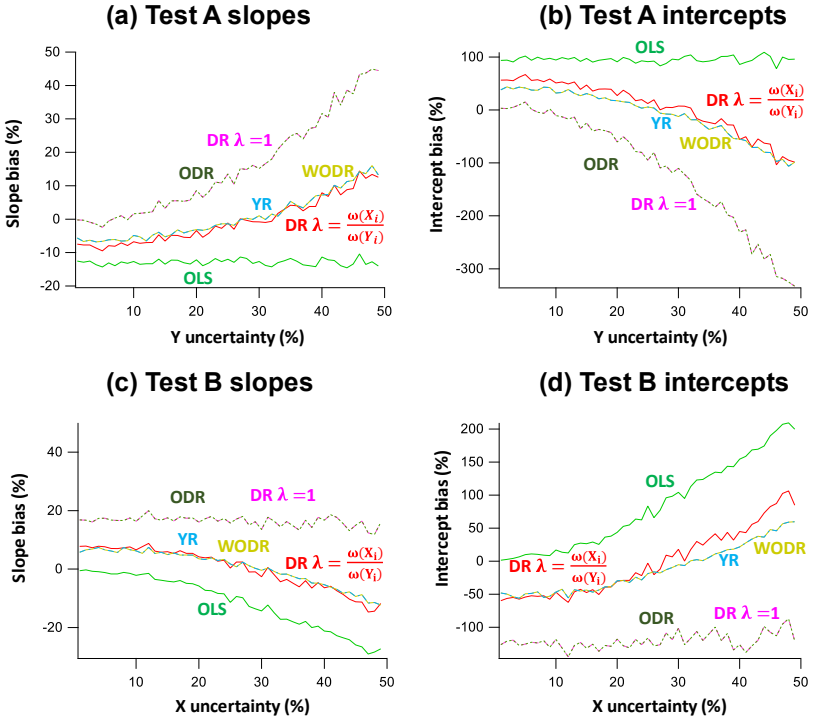


Figure 6. Slope and intercept biases by different regression schemes in two test scenarios (A and B) in which the assumed error for one of the regression variables deviates from the actual measurement error. In Test A data generation, $\gamma_{\mathrm{Unc\_X}}$ is fixed at 30 % and $\gamma_{\mathrm{Unc\_Y}}$ is varied between 1 and 50 %. In Test B, $\gamma_{\mathrm{Unc\_X}}$ varies between 1 and 50 % and $\gamma_{\mathrm{Unc\_Y}}$ is fixed at 30 %. The assumed measurement error for regression is 10 % for both X and Y. **(a)** Slope biases as a function of $\gamma_{\mathrm{Unc\_Y}}$ in Test A. **(b)** Intercept biases as a function of $\gamma_{\mathrm{Unc\_Y}}$ in Test A. **(c)** Slope biases as a function of $\gamma_{\mathrm{Unc\_X}}$ in Test B. **(d)** Intercept biases as a function of $\gamma_{\mathrm{Unc\_X}}$ in Test B.

is closest to the true value (intercept = 3), the deviations are much larger than for WODR (2.72 ± 0.74).

4.2 Comparison results using data generated by MT

In this section, MT is adopted for data generation to obtain a benchmark of six regression approaches. Both $\gamma_{\mathrm{Unc-nonlinear}}$ and $\gamma_{\mathrm{Unc-linear}}$ are considered. With different configuration of slope, intercept and γ_{Unc}, 12 cases (Cases 7–18) are studied and the results are discussed below.

4.2.1 $\gamma_{\mathrm{Unc-nonlinear}}$ results

Cases 7 and 8 use data generated by MT and $\gamma_{\mathrm{Unc-nonlinear}}$ with results shown in Table 1. In Case 7 (slope = 4, intercept = 0, $\mathrm{LOD_{POC}} = 1$, $\mathrm{LOD_{EC}} = 1$, $\alpha_{\mathrm{POC}} = 1$, $\alpha_{\mathrm{EC}} = 1$), unbiased slope (4.00 ± 0.03) and intercept (0.00 ± 0.17) is estimated by DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$. WODR and YR yield unbiased slopes (3.96 ± 0.03) but overestimate the intercepts (1.21 ± 0.13). DR ($\lambda = 1$) and ODR report slightly biased slopes (4.17 ± 0.04) with biased intercepts (-0.94 ± 0.18). OLS underestimates the slope (3.22 ± 0.03) and overestimates the intercept (4.30 ± 0.14). In Case 8 (slope = 4, intercept = 3, $\mathrm{LOD_{POC}} = 1$, $\mathrm{LOD_{EC}} = 1$, $\alpha_{\mathrm{POC}} = 1$, $\alpha_{\mathrm{EC}} = 1$), DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$ provides unbiased slope (4.00 ± 0.03) and intercept (3.00 ± 0.18) estimations. WODR and YR report unbiased slopes (3.97 ± 0.03) and overestimate intercepts (4.11 ± 0.13). OLS, DR ($\lambda = 1$) and ODR report biased slopes and intercepts.

To test the overestimation/underestimation dependency on the true slope, Case 9 (slope = 0.5, intercept = 0, $\mathrm{LOD_{POC}} = 1$, $\mathrm{LOD_{EC}} = 1$, $\alpha_{\mathrm{POC}} = 1$, $\alpha_{\mathrm{EC}} = 1$) and Case 10 (slope = 0.5, intercept = 3, $\mathrm{LOD_{POC}} = 1$, $\mathrm{LOD_{EC}} = 1$, $\alpha_{\mathrm{POC}} = 1$, $\alpha_{\mathrm{EC}} = 1$) are conducted and the results are shown in Table 1. Unlike the overestimation observed in Cases 1–8, DR ($\lambda = 1$) and ODR underestimate the slopes (0.46 ± 0.01) in Case 9. In Case 10, DR ($\lambda = 1$), DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$ and ODR report unbiased slopes and intercepts. Cases 11 and 12 test the bias when the true slope is 1, as shown in Table 1. In Case 11 (intercept = 0), all regression approaches except OLS can provide unbiased results. In Case 12, all regression approaches report unbiased slopes except OLS, but DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$ is the only regression approach that reports unbiased intercept.

These results imply that if the true slope is less than 1, the improper weighting ($\lambda = 1$) in Deming regression and ODR without weighting tends to underestimate slope. If the true

slope is 1, these two estimators can provide unbiased results. If the true slope is larger than 1, the improper weighting ($\lambda = 1$) in Deming regression and ODR without weighting tends to overestimate slope.

4.2.2 $\gamma_{\text{Unc-linear}}$ results

Cases 13 and 14 (Table 1) represent the results from using $\gamma_{\text{Unc-linear}}$ (30 %) and data generated from MT. For Case 13 (slope = 4, intercept = 0), DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$, WODR and YR provide the best estimation of slopes and intercepts. DR ($\lambda = 1$) and ODR overestimate slopes (4.53 ± 0.05) and underestimate intercepts (-2.94 ± 0.24). For Case 14 (slope = 4, intercept = 3), DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$, WODR and YR provide an unbiased estimation of slopes. But DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$ is the only regression approach reporting unbiased intercept (3.08 ± 0.23). Cases 15 and 16 are tested to investigate whether the results are different if the true slope is smaller than 1. As shown in Table 1, the results are similar to Cases 13 and 14, i.e., that DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$ can provide unbiased slope and intercept while WODR and YR can provide unbiased slopes but biased intercepts. Cases 17 and 18 are tested to see if the results are the same for a special case when the true slope is 1. As shown in Table 1, the results are similar to Cases 13 and 14, implying that these results are not sensitive to the special case when the true slope is 1.

4.3 The importance of appropriate λ input for Deming regression

As discussed above, inappropriate λ assignment in the Deming regression (e.g., $\lambda = 1$ by default for much commercial software) leads to biased slope and intercept. Besides $\lambda = 1$, inappropriate λ input due to improper handling of measurement uncertainty can also result in bias for Deming regression. An example is shown in Fig. S3. Data are generated by MT with following parameters: slope = 4, intercept = 0, and $\gamma_{\text{Unc-linear}}$ (30 %). Figure S2a and b demonstrate that when an appropriate λ is provided (following $\gamma_{\text{Unc-linear}}$, $\lambda = \frac{\text{POC}^2}{\text{EC}^2}$), unbiased slopes and intercepts are obtained. If an improper λ is used due to a mismatched measurement uncertainty assumption $\left(\gamma_{\text{Unc-nonlinear}}, \lambda = \frac{\text{POC}}{\text{EC}}\right)$, the slopes are overestimated (Fig. S3c, 4.37 ± 0.05) and intercepts are underestimated (Fig. S3, -2.01 ± 0.24). This result emphasizes the importance of determining the correct form of measurement uncertainty in ambient samples, since λ is a crucial parameter in Deming regression.

In the λ calculation, different representations for POC and EC, including mean, median and mode, are tested as shown in Fig. S4. The results show that when X and Y have a similar distribution (e.g., both are lognormal), any of mean, median or mode can be used for the λ calculation.

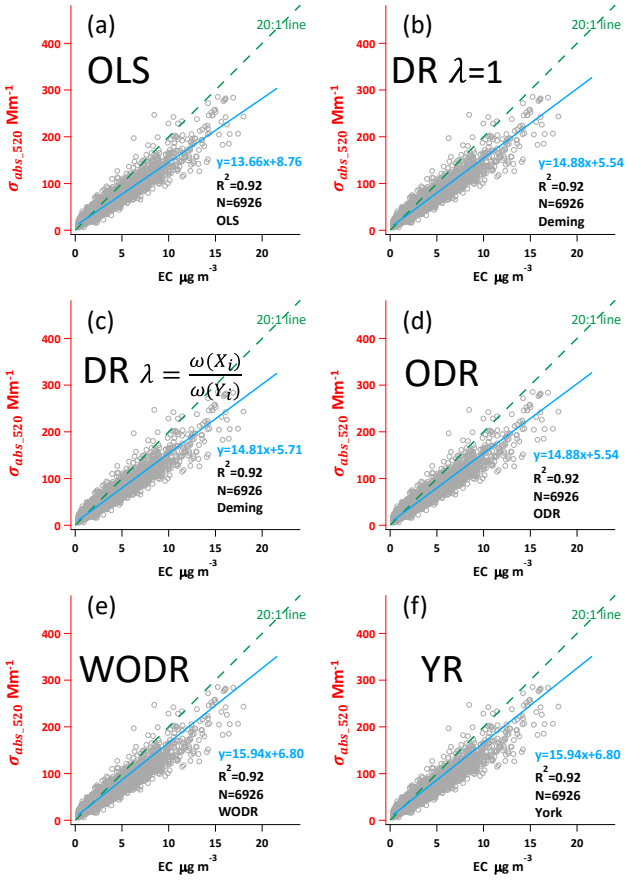


Figure 7. Regression results using ambient σ_{abs520} and EC data from a suburban site in Guangzhou, China.

4.4 Caveats of regressions with unknown X and Y uncertainties

In atmospheric applications, there are scenarios in which a priori error in one of the variables is unknown, or the measurement error described cannot be trusted. For example, in the case of comparing model prediction and measurement data, the uncertainty of model prediction data is unknown. A second example is the case in which measurement uncertainty cannot be determined due to the lack of duplicated or collocated measurements and as a result, an arbitrarily assumed uncertainty is used. Such a case was illustrated in the study by Flanagan et al. (2006). They found that in the Speciation Trends Network (STN), the whole-system uncertainty retrieved by data from collocated samplers was different from the arbitrarily assumed 5 % uncertainty. Additionally, the discrepancy between the actual uncertainty obtained through collocated samplers and the arbitrarily assumed uncertainty varied by chemical species. To investigate the performance of different regression approaches in these cases, two tests (A and B) are conducted.

In Test A, the actual measurement error for X is fixed at 30 %, while γ_{Unc} for Y varies from 1 to 50 %. The assumed measurement error for regression is 10 % for both X and Y. Results of Test A are shown in Fig. 6a and b. For OLS, the slopes are underestimated (−14 to −12 %) and intercepts are overestimated (90–103 %) and the biases are independent of variations in $\gamma_{\mathrm{Unc\_Y}}$. ODR and DR ($\lambda = 1$) yield similar results with overestimated slopes (0–44 %) and underestimated intercepts (−330–0 %). The degree of bias in slopes and intercepts depends on the $\gamma_{\mathrm{Unc\_Y}}$. WODR, DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$ and YR perform much better than other regression approaches in Test A, with a smaller bias in both slopes (−8–12 %) and intercepts −98–55 %).

In Test B, $\gamma_{\mathrm{Unc\_Y}}$ is fixed at 30 % and $\gamma_{\mathrm{Unc\_X}}$ varies between 1 and 50 %. The results of Test B are shown in Fig. 6c and d. The assumed measurement error for regression is 10 % for both X and Y. OLS underestimates the slopes (−29 to −0.2 %) and overestimates the intercepts (2–209 %). In contrast to Test A in which slope and intercept biases are independent of variations in $\gamma_{\mathrm{Unc\_Y}}$, the slope and intercept biases in Test B exhibit dependency on $\gamma_{\mathrm{Unc\_X}}$. The reason for this is that OLS only considers errors in Y and X is assumed to be error-free. ODR and DR ($\lambda = 1$) yield similar results with overestimated slopes (11–18 %) and underestimated intercepts (−144 to −87 %). The degree of bias in slopes and intercepts is relatively independent on the $\gamma_{\mathrm{Unc\_X}}$. WODR, DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$ and YR performed much better than the other regression approaches in Test B, with a smaller bias in both slopes (−14–8 %) and intercepts (−59–106 %).

The results from these two tests suggest that, if one of the measurement errors described cannot be trusted or a priori error in one of the variables is unknown, WODR, DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$ and YR should be used instead of ODR and DR ($\lambda = 1$) and OLS. This conclusion is consistent with results presented in Sect. 4.1 and 4.2. This analysis, albeit crude, also suggests that, in general, the magnitude of bias in slope estimation by these regression approaches is smaller than those for intercept. In other words, slope is a more reliable quantity compared to intercept when extracting quantitative information from linear regressions.

5 Regression applications to ambient data

This section demonstrates the application of the six regression approaches on a light absorption coefficient and EC dataset collected in a suburban site in Guangzhou. As mentioned in Sect. 4.4, measurement uncertainties are crucial inputs for DR, YR and WODR. The measurement precision of Aethalometer is 5 % (Hansen, 2005), while EC by the RT-ECOC analyzer is 24 % (Bauer et al., 2009). These measurement uncertainties are used in DR, YR and WODR calculation. The dataset contains 6926 data points with an R^2 of 0.92.

As shown in Fig. 7, the y axis is light absorption at 520 nm (σ_{abs520}) and the x axis is EC mass concentration. The regressed slopes represent the mass absorption efficiency (MAE) of EC at 520 nm, ranging from 13.66 to 15.94 $\mathrm{m^2\ g^{-1}}$ by the six regression approaches. OLS yields the lowest slope (13.66 as shown in Fig. 7a) among all six regression approaches, consistent with the results using synthetic data. This implies that OLS tends to underestimate regression slope when mean Y to X ratio is larger than 1. DR ($\lambda = 1$) and ODR report the same slope (14.88) and intercept (5.54); this equivalency is also observed for the synthetic data. Similarly, WODR and YR yield identical slope (14.88) and intercept (5.54), in line with the synthetic data results. The regressed slope by DR ($\lambda = 1$) is higher than DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$, and this relationship agrees well with the synthetic data results.

Regression comparison is also performed on hourly OC and EC data. Regression on OC / EC percentile subset is a widely used empirical approach for primary OC / EC ratio determination. Figure S5 shows the regression slopes as a function of OC / EC percentile. OC / EC percentile ranges from 0.5 to 100 %, with an interval of 0.5 %. As the percentile increases, SOC contribution in OC increases as well, resulting in decreased R^2 between OC and EC. The deviations between six regression approaches exhibit a dependency on R^2. When percentile is relatively small (e.g., < 10 %), the differences between the six regression approaches are also small due to the high R^2 (0.98). The deviations between the six regression approaches become more pronounced as R^2 decreases (e.g., < 0.9). The deviations are expected to be even larger when R^2 is less than 0.8. These results emphasize the importance of applying error-in-variables regression, since ambient XY data more likely has an R^2 less than 0.9 in most cases.

As discussed in this section, the ambient data confirm the results obtained in comparing methods with the synthetic data. The advantage of using the synthetic data for regression approaches evaluation is that the ideal slope and intercept are known values during the data generation, so the bias of each regression approach can be quantified.

6 Recommendations and conclusions

This study aims to provide a benchmark of commonly used linear regression algorithms using a new data generation scheme (MT). Six regression approaches are tested, i.e., OLS, DR ($\lambda = 1$), DR $\left(\lambda = \frac{\omega(X_i)}{\omega(Y_i)}\right)$, ODR, WODR and YR. The results show that OLS fails to estimate the correct slope and intercept when both X and Y have measurement errors. This result is consistent with previous studies. For ambient data with R^2 less than 0.9, error-in-variables regression is needed to minimize the biases in slope and intercept. If measurement uncertainties in X and Y are determined during

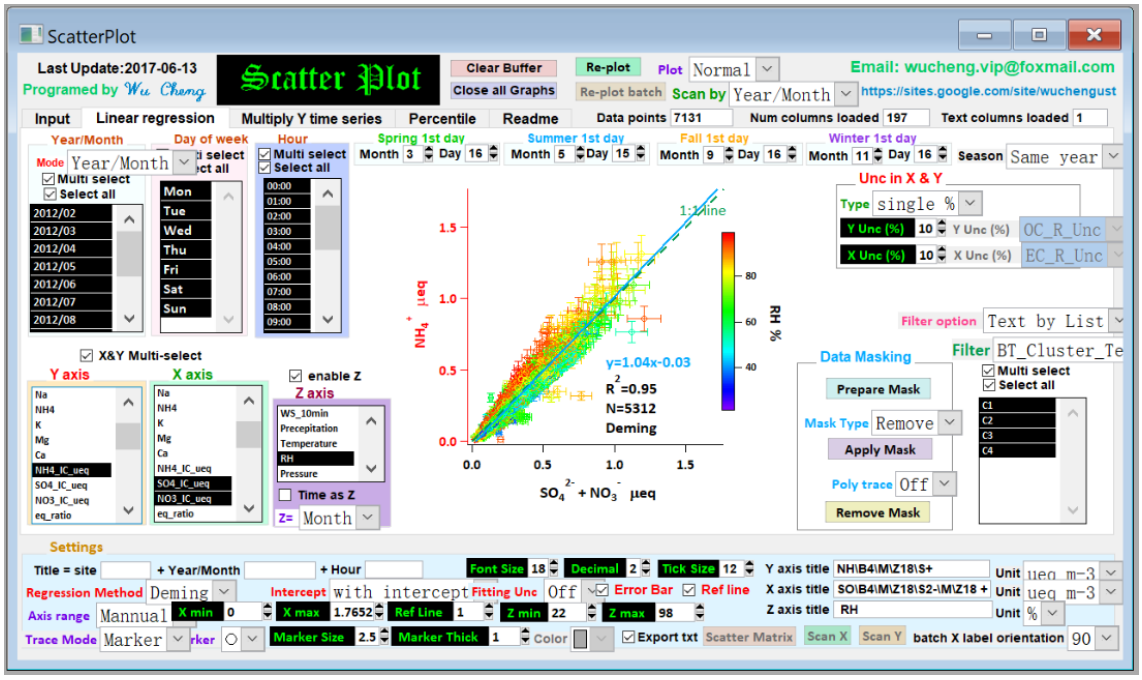


Figure 8. The user interface of the Scatter Plot Igor program. The program and its operation manual are available from https://doi.org/10.5281/zenodo.832417.

the measurement, measurement uncertainties should be used for regression. With appropriate weighting, DR, WODR and YR can provide the best results among all tested regression techniques. Sensitivity tests also reveal the importance of the weighting parameter λ in DR. An improper λ could lead to biased slope and intercept. Since the λ estimation depends on the form of the measurement errors, it is important to determine the measurement errors during the experimentation stage rather than making assumptions. If measurement errors are not available from the measurement and assumptions are made on measurement errors, DR, WODR and YR are still the best option that can provide the least bias in slope and intercept among all tested regression techniques. For these reasons, DR, WODR and YR are recommended for atmospheric studies when both X and Y data have measurement errors.

Application of error-in-variables regression is often overlooked in atmospheric studies, partly due to the lack of a specified tool for the regression implementation. To facilitate the implementation of error-in-variables regression (including DR, WODR and YR), a computer program (Scatter Plot) with a GUI in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) was developed (Fig. 8). It is packed with many useful features for data analysis and plotting, including batch plotting, data masking via GUI, color coding in the z axis, data filtering and grouping by numerical values and strings. The Scatter Plot program and user manual are available from https://sites.google.com/site/wuchengust and https://doi.org/10.5281/zenodo.832417.

Data availability. OC, EC and σ_{abs} data used in this study are available from the corresponding authors upon request. The computer programs used for data analysis and visualization in this study are available in Wu (2017a–c).

Appendix A: Equations of regression techniques

Ordinary least squares (OLS) calculation steps.

First calculate average of observed X_i and Y_i.

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{A1}$$

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} \tag{A2}$$

Then calculate S_{xx} and S_{yy}.

$$S_{xx} = \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2 \tag{A3}$$

$$S_{yy} = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 \tag{A4}$$

OLS slope and intercept can be obtained from

$$k = \frac{S_{yy}}{S_{xx}}, \tag{A5}$$

$$b = \overline{Y} - k\overline{X}. \tag{A6}$$

Deming regression (DR) calculation steps (York, 1966).

Besides S_{xx} and S_{yy} as shown above, S_{xy} can be calculated from

$$S_{xy} = \sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right), \tag{A7}$$

DR slope and intercept can be obtained from

$$k = \frac{S_{yy} - \lambda S_{xx} + \sqrt{\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S_{xy}^2}}{2S_{xy}}, \tag{A8}$$

$$b = \overline{Y} - k\overline{X}. \tag{A9}$$

York regression (YR) iteration steps (York et al., 2004).

Slope by OLS can be used as the initial k in W_i calculation.

$$W_i = \frac{\omega(X_i)\,\omega(Y_i)}{\omega(X_i) + k^2\omega(Y_i) - 2kr_i\sqrt{\omega(X_i)\,\omega(Y_i)}} \tag{A10}$$

$$U_i = X_i - \overline{X} = X_i - \frac{\sum_{i=1}^{N} W_i X_i}{\sum_{i=1}^{N} W_i} \tag{A11}$$

$$V_i = Y_i - \overline{Y} = Y_i - \frac{\sum_{i=1}^{N} W_i Y_i}{\sum_{i=1}^{N} W_i} \tag{A12}$$

Then calculate β_i.

$$\beta_i = W_i\left[\frac{U_i}{\omega(Y_i)} + \frac{kV_i}{\omega(X_i)} - \left[kU_i + V_i\right]\frac{r_i}{\sqrt{\omega(X_i)\,\omega(Y_i)}}\right] \tag{A13}$$

Slope and intercept can be obtained from

$$k = \frac{\sum_{i=1}^{N} W_i\beta_i V_i}{\sum_{i=1}^{N} W_i\beta_i U_i}, \tag{A14}$$

$$b = \overline{Y} - k\overline{X}. \tag{A15}$$

Since W_i and β_i are functions of k, k must be solved iteratively by repeating Eqs. (A11) to (A15). If the difference between the k obtained from Eq. (A15) and the k used in Eq. (A11) satisfies the predefined tolerance ($\frac{k_{i+1}-k_i}{k_i} < e^{-15}$), the calculation is considered as converged. The calculation is straightforward and usually converged in 10 iterations. For example, the iteration count on the dataset of Chu (2005) is around 6.

Appendix B: Summary of abbreviations and symbols

Abbreviation/symbol	Definition
α	a dimensionless adjustable factor to control the position of γ_{Unc} curve on the concentration axis
b	intercept in linear regression
β_i, U_i, V_i, W_i	intermediates in York regression calculations
γ_{Unc}	fractional measurement uncertainties relative to the true concentration (%)
DR	Deming regression
$\varepsilon_{\mathrm{EC}}, \varepsilon_{\mathrm{POC}}$	absolute measurement uncertainties of EC and POC
EC	elemental carbon
$\mathrm{EC}_{\mathrm{true}}$	numerically synthesized true EC concentration without measurement uncertainty
$\mathrm{EC}_{\mathrm{measured}}$	EC with measurement error ($\mathrm{EC}_{\mathrm{true}} + \varepsilon_{\mathrm{EC}}$)
λ	$\omega(X_i)$ to $\omega(Y_i)$ ratio in Deming regression
k	slope in linear regression
LOD	limit of detection
MT	Mersenne twister pseudorandom number generator
OC	organic carbon
OC / EC	OC to EC ratio
$(\mathrm{OC}/\mathrm{EC})_{\mathrm{pri}}$	primary OC / EC ratio
$\mathrm{OC}_{\mathrm{non\text{-}comb}}$	OC from non-combustion sources
ODR	orthogonal distance regression
OLS	ordinary least squares regression
POC	primary organic carbon
$\mathrm{POC}_{\mathrm{comb}}$	numerically synthesized true POC from combustion sources (well correlated with $\mathrm{EC}_{\mathrm{true}}$), measurement uncertainty not considered
$\mathrm{POC}_{\mathrm{non\text{-}comb}}$	numerically synthesized true POC from non-combustion sources (independent of $\mathrm{EC}_{\mathrm{true}}$) without considering measurement uncertainty
$\mathrm{POC}_{\mathrm{true}}$	sum of $\mathrm{POC}_{\mathrm{comb}}$ and $\mathrm{POC}_{\mathrm{non\text{-}comb}}$ without considering measurement uncertainty
$\mathrm{POC}_{\mathrm{measured}}$	POC with measurement error ($\mathrm{POC}_{\mathrm{true}} + \varepsilon_{\mathrm{POC}}$)
$\sigma_{X_i}, \sigma_{Y_i}$	the standard deviation of the error in measurement of X_i and Y_i
r_i	correlation coefficient between errors in X_i and Y_i in YR
S	sum of squared residuals
SOC	secondary organic carbon
τ	parameter in the sine function of Chu (2005) that adjusts the width of each peak
ϕ	parameter in the sine function of Chu (2005) that adjusts the phase of the curve
WODR	weighted orthogonal distance regression
$\overline{X}, \overline{Y}$	average of X_i and Y_i
YR	York regression
$\omega(X_i), \omega(Y_i)$	inverse of σ_{X_i} and σ_{Y_i}, used as weights in DR calculation.

The Supplement related to this article is available online at https://doi.org/10.5194/amt-11-1233-2018-supplement.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (grant no. 41605002, 41475004 and 21607056), NSFC of Guangdong Province (grant no. 2015A030313339), Guangdong Province Public Interest Research and Capacity Building Special Fund (grant no. 2014B020216005). The author would like to thank Bin Yu Kuang at HKUST for the discussions on mathematics and Stephen M. Griffith at HKUST for the valuable comments.

Edited by: Willy Maenhaut
Reviewed by: two anonymous referees

References

Ayers, G. P.: Comment on regression analysis of air quality data, Atmos. Environ., 35, 2423–2425, https://doi.org/10.1016/S1352-2310(00)00527-6, 2001.

Bauer, J. J., Yu, X.-Y., Cary, R., Laulainen, N., and Berkowitz, C.: Characterization of the sunset semi-continuous carbon aerosol analyzer, J. Air Waste Manage., 59, 826–833, https://doi.org/10.3155/1047-3289.59.7.826, 2009.

Boggs, P. T., Donaldson, J. R., and Schnabel, R. B.: Algorithm 676: ODRPACK: software for weighted orthogonal distance regression, ACM T. Math. Software, 15, 348–364, https://doi.org/10.1145/76909.76913, 1989.

Brauers, T. and Finlayson-Pitts, B. J.: Analysis of relative rate measurements, Int. J. Chem. Kinet., 29, 665–672, https://doi.org/10.1002/(SICI)1097-4601(1997)29:9<665::AID-KIN3>3.0.CO;2-S, 1997.

Cantrell, C. A.: Technical Note: Review of methods for linear least-squares fitting of data and application to atmospheric chemistry problems, Atmos. Chem. Phys., 8, 5477–5487, https://doi.org/10.5194/acp-8-5477-2008, 2008.

Carroll, R. J. and Ruppert, D.: The use and misuse of orthogonal regression in linear errors-in-variables models, Am. Stat., 50, 1–6, https://doi.org/10.1080/00031305.1996.10473533, 1996.

Cess, R. D., Zhang, M. H., Minnis, P., Corsetti, L., Dutton, E. G., Forgan, B. W., Garber, D. P., Gates, W. L., Hack, J. J., Harrison, E. F., Jing, X., Kiehi, J. T., Long, C. N., Morcrette, J.-J., Potter, G. L., Ramanathan, V., Subasilar, B., Whitlock, C. H., Young, D. F., and Zhou, Y.: Absorption of solar radiation by clouds: Observations versus models, Science, 267, 496–499, https://doi.org/10.1126/science.267.5197.496, 1995.

Chen, L. W. A., Doddridge, B. G., Dickerson, R. R., Chow, J. C., Mueller, P. K., Quinn, J., and Butler, W. A.: Seasonal variations in elemental carbon aerosol, carbon monoxide and sulfur dioxide: Implications for sources, Geophys. Res. Lett., 28, 1711–1714, https://doi.org/10.1029/2000GL012354, 2001.

Cheng, Y., Duan, F.-K., He, K.-B., Zheng, M., Du, Z.-Y., Ma, Y.-L., and Tan, J.-H.: Intercomparison of thermal–optical methods for the determination of organic and elemental carbon: influences of aerosol composition and implications, Environ. Sci. Technol., 45, 10117–10123, https://doi.org/10.1021/es202649g, 2011.

Chow, J. C., Watson, J. G., Crow, D., Lowenthal, D. H., and Merrifield, T.: Comparison of IMPROVE and NIOSH carbon measurements, Aerosol. Sci. Technol., 34, 23–34, https://doi.org/10.1080/027868201300081923, 2001.

Chow, J. C., Watson, J. G., Chen, L. W. A., Arnott, W. P., and Moosmuller, H.: Equivalence of elemental carbon by thermal/optical reflectance and transmittance with different temperature protocols, Environ. Sci. Technol., 38, 4414–4422, https://doi.org/10.1021/Es034936u, 2004.

Chu, S. H.: Stable estimate of primary OC / EC ratios in the EC tracer method, Atmos. Environ., 39, 1383–1392, https://doi.org/10.1016/j.atmosenv.2004.11.038, 2005.

Collaud Coen, M., Weingartner, E., Apituley, A., Ceburnis, D., Fierz-Schmidhauser, R., Flentje, H., Henzing, J. S., Jennings, S. G., Moerman, M., Petzold, A., Schmid, O., and Baltensperger, U.: Minimizing light absorption measurement artifacts of the Aethalometer: evaluation of five correction algorithms, Atmos. Meas. Tech., 3, 457–474, https://doi.org/10.5194/amt-3-457-2010, 2010.

Cornbleet, P. J. and Gochman, N.: Incorrect least-squares regression coefficients in method-comparison analysis, Clin. Chem., 25, 432–438, 1979.

Cox, M., Harris, P., and Siebert, B. R.-L.: Evaluation of measurement uncertainty based on the propagation of distributions using Monte Carlo simulation, Meas. Tech., 46, 824–833, https://doi.org/10.1023/B:METE.0000008439.82231.ad, 2003.

Cross, E. S., Onasch, T. B., Ahern, A., Wrobel, W., Slowik, J. G., Olfert, J., Lack, D. A., Massoli, P., Cappa, C. D., Schwarz, J. P., Spackman, J. R., Fahey, D. W., Sedlacek, A., Trimborn, A., Jayne, J. T., Freedman, A., Williams, L. R., Ng, N. L., Mazzoleni, C., Dubey, M., Brem, B., Kok, G., Subramanian, R., Freitag, S., Clarke, A., Thornhill, D., Marr, L. C., Kolb, C. E., Worsnop, D. R., and Davidovits, P.: Soot particle studies – instrument intercomparison – project overview, Aerosol. Sci. Technol., 44, 592–611, https://doi.org/10.1080/02786826.2010.482113, 2010.

Deming, W. E.: Statistical Adjustment of Data, Wiley, New York, 1943.

Duan, F., Liu, X., Yu, T., and Cachier, H.: Identification and estimate of biomass burning contribution to the urban aerosol organic carbon concentrations in Beijing, Atmos. Environ., 38, 1275–1282, https://doi.org/10.1016/j.atmosenv.2003.11.037, 2004.

Flanagan, J. B., Jayanty, R. K. M., Rickman, J. E. E., and Peterson, M. R.: PM2.5 speciation trends network: evaluation of whole-system uncertainties using data from sites with collocated samplers, J. Air Waste Manage., 56, 492–499, https://doi.org/10.1080/10473289.2006.10464516, 2006.

Hansen, A. D. A.: The Aethalometer manual, Berkeley, California, USA, Magee Scientific, 2005.

Huang, X. H., Bian, Q., Ng, W. M., Louie, P. K., and Yu, J. Z.: Characterization of PM2. 5 major components and source investigation in suburban Hong Kong: A one year monitoring study, Aerosol. Air. Qual. Res., 14, 237–250, https://doi.org/10.4209/aaqr.2013.01.0020, 2014.

Janhäll, S., Andreae, M. O., and Pöschl, U.: Biomass burning aerosol emissions from vegetation fires: particle number and mass emission factors and size distributions, Atmos. Chem. Phys., 10, 1427–1439, https://doi.org/10.5194/acp-10-1427-2010, 2010.

Kuang, B. Y., Lin, P., Huang, X. H. H., and Yu, J. Z.: Sources of humic-like substances in the Pearl River Delta, China: positive matrix factorization analysis of $PM^{2.5}$ major components and source markers, Atmos. Chem. Phys., 15, 1995–2008, https://doi.org/10.5194/acp-15-1995-2015, 2015.

Li, Y., Huang, H. X. H., Griffith, S. M., Wu, C., Lau, A. K. H., and Yu, J. Z.: Quantifying the relationship between visibility degradation and PM2.5 constituents at a suburban site in Hong Kong: Differentiating contributions from hydrophilic and hydrophobic organic compounds, Sci. Total Environ., 575, 1571–1581, https://doi.org/10.1016/j.scitotenv.2016.10.082, 2017.

Lim, L. H., Harrison, R. M., and Harrad, S.: The contribution of traffic to atmospheric concentrations of polycyclic aromatic hydrocarbons, Environ. Sci. Technol., 33, 3538–3542, https://doi.org/10.1021/es990392d, 1999.

Lin, P., Hu, M., Deng, Z., Slanina, J., Han, S., Kondo, Y., Takegawa, N., Miyazaki, Y., Zhao, Y., and Sugimoto, N.: Seasonal and diurnal variations of organic carbon in PM2.5 in Beijing and the estimation of secondary organic carbon, J. Geophys. Res., 114, D00G11, https://doi.org/10.1029/2008JD010902, 2009.

Linnet, K.: Necessary sample size for method comparison studies based on regression analysis, Clin. Chem., 45, 882–894, 1999.

Malm, W. C., Sisler, J. F., Huffman, D., Eldred, R. A., and Cahill, T. A.: Spatial and seasonal trends in particle concentration and optical extinction in the United-States, J. Geophys. Res., 99, 1347–1370, https://doi.org/10.1029/93JD02916, 1994.

Markovsky, I. and Van Huffel, S.: Overview of total least-squares methods, Signal Process., 87, 2283–2302, https://doi.org/10.1016/j.sigpro.2007.04.004, 2007.

Matsumoto, M. and Nishimura, T.: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, ACM T. Model. Comput. S., 8, 3–30, https://doi.org/10.1145/272991.272995, 1998.

Moosmüller, H., Arnott, W. P., Rogers, C. F., Chow, J. C., Frazier, C. A., Sherman, L. E., and Dietrich, D. L.: Photoacoustic and filter measurements related to aerosol light absorption during the Northern Front Range Air Quality Study (Colorado 1996/1997), J. Geophys. Res., 103, 28149–28157, https://doi.org/10.1029/98jd02618, 1998.

Petäjä, T., Mauldin III, R. L., Kosciuch, E., McGrath, J., Nieminen, T., Paasonen, P., Boy, M., Adamov, A., Kotiaho, T., and Kulmala, M.: Sulfuric acid and OH concentrations in a boreal forest site, Atmos. Chem. Phys., 9, 7435–7448, https://doi.org/10.5194/acp-9-7435-2009, 2009.

Richter, A., Burrows, J. P., Nusz, H., Granier, C., and Niemeier, U.: Increase in tropospheric nitrogen dioxide over China observed from space, Nature, 437, 129–132, https://doi.org/10.1038/nature04092, 2005.

Saylor, R. D., Edgerton, E. S., and Hartsell, B. E.: Linear regression techniques for use in the EC tracer method of secondary organic aerosol estimation, Atmos. Environ., 40, 7546–7556, https://doi.org/10.1016/j.atmosenv.2006.07.018, 2006.

Thompson, M.: Variation of precision with concentration in an analytical system, Analyst, 113, 1579–1587, https://doi.org/10.1039/AN9881301579, 1988.

Turpin, B. J. and Huntzicker, J. J.: Identification of secondary organic aerosol episodes and quantitation of primary and secondary organic aerosol concentrations during SCAQS, Atmos. Environ., 29, 3527–3544, https://doi.org/10.1016/1352-2310(94)00276-Q, 1995.

von Bobrutzki, K., Braban, C. F., Famulari, D., Jones, S. K., Blackall, T., Smith, T. E. L., Blom, M., Coe, H., Gallagher, M., Ghalaieny, M., McGillen, M. R., Percival, C. J., Whitehead, J. D., Ellis, R., Murphy, J., Mohacsi, A., Pogany, A., Junninen, H., Rantanen, S., Sutton, M. A., and Nemitz, E.: Field inter-comparison of eleven atmospheric ammonia measurement techniques, Atmos. Meas. Tech., 3, 91–112, https://doi.org/10.5194/amt-3-91-2010, 2010.

Wang, J. and Christopher, S. A.: Intercomparison between satellite-derived aerosol optical thickness and PM2.5 mass: Implications for air quality studies, Geophys. Res. Lett., 30, 2095, https://doi.org/10.1029/2003gl018174, 2003.

Watson, J. G.: Visibility: Science and regulation, J. Air Waste Manage., 52, 628–713, https://doi.org/10.1080/10473289.2002.10470813, 2002.

Weingartner, E., Saathoff, H., Schnaiter, M., Streit, N., Bitnar, B., and Baltensperger, U.: Absorption of light by soot particles: determination of the absorption coefficient by means of aethalometers, J. Aerosol. Sci., 34, 1445–1463, https://doi.org/10.1016/S0021-8502(03)00359-8, 2003.

Wu, C.: Scatter Plot, https://doi.org/10.5281/zenodo.832417, 2017a.

Wu, C.: Aethalometer data processor, https://doi.org/10.5281/zenodo.832404, 2017b.

Wu, C.: Histbox, https://doi.org/10.5281/zenodo.832411, 2017c.

Wu, C. and Yu, J. Z.: Determination of primary combustion source organic carbon-to-elemental carbon (OC / EC) ratio using ambient OC and EC measurements: secondary OC-EC correlation minimization method, Atmos. Chem. Phys., 16, 5453–5465, https://doi.org/10.5194/acp-16-5453-2016, 2016.

Wu, C., Ng, W. M., Huang, J., Wu, D., and Yu, J. Z.: Determination of Elemental and Organic Carbon in PM2.5 in the Pearl River Delta Region: Inter-Instrument (Sunset vs. DRI Model 2001 Thermal/Optical Carbon Analyzer) and Inter-Protocol Comparisons (IMPROVE vs. ACE-Asia Protocol), Aerosol. Sci. Technol., 46, 610–621, https://doi.org/10.1080/02786826.2011.649313, 2012.

Wu, C., Huang, X. H. H., Ng, W. M., Griffith, S. M., and Yu, J. Z.: Inter-comparison of NIOSH and IMPROVE protocols for OC and EC determination: implications for inter-protocol data conversion, Atmos. Meas. Tech., 9, 4547–4560, https://doi.org/10.5194/amt-9-4547-2016, 2016.

Wu, C., Wu, D., and Yu, J. Z.: Quantifying black carbon light absorption enhancement with a novel statistical approach, Atmos. Chem. Phys., 18, 289–309, https://doi.org/10.5194/acp-18-289-2018, 2018.

York, D.: Least-squares fitting of a straight line, Can. J. Phys., 44, 1079–1086, https://doi.org/10.1139/p66-090, 1966.

York, D., Evensen, N. M., Martınez, M. L., and Delgado, J. D. B.: Unified equations for the slope, intercept, and standard

errors of the best straight line, Am. J. Phys., 72, 367–375, https://doi.org/10.1119/1.1632486, 2004.

Yu, J. Z., Huang, X.-F., Xu, J., and Hu, M.: When Aerosol Sulfate Goes Up, So Does Oxalate: Implication for the Formation Mechanisms of Oxalate, Environ. Sci. Technol., 39, 128–133, https://doi.org/10.1021/es049559f, 2005.

Zhou, Y., Huang, X. H. H., Griffith, S. M., Li, M., Li, L., Zhou, Z., Wu, C., Meng, J., Chan, C. K., Louie, P. K. K., and Yu, J. Z.: A field measurement based scaling approach for quantification of major ions, organic carbon, and elemental carbon using a single particle aerosol mass spectrometer, Atmos. Environ., 143, 300–312, https://doi.org/10.1016/j.atmosenv.2016.08.054, 2016.

Zieger, P., Weingartner, E., Henzing, J., Moerman, M., de Leeuw, G., Mikkilä, J., Ehn, M., Petäjä, T., Clémer, K., van Roozendael, M., Yilmaz, S., Frieß, U., Irie, H., Wagner, T., Shaiganfar, R., Beirle, S., Apituley, A., Wilson, K., and Baltensperger, U.: Comparison of ambient aerosol extinction coefficients obtained from in-situ, MAX-DOAS and LIDAR measurements at Cabauw, Atmos. Chem. Phys., 11, 2603–2624, https://doi.org/10.5194/acp-11-2603-2011, 2011.

Zwolak, J. W., Boggs, P. T., and Watson, L. T.: Algorithm 869: ODRPACK95: A weighted orthogonal distance regression code with bound constraints, ACM T. Math. Software, 33, 27, https://doi.org/10.1145/1268776.1268782, 2007.

Supplement of Atmos. Meas. Tech., 11, 1233–1250, 2018
https://doi.org/10.5194/amt-11-1233-2018-supplement

Atmospheric Measurement Techniques
Open Access
EGU

Supplement of

Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting

Cheng Wu and Jian Zhen Yu

Correspondence to: Cheng Wu (wucheng.vip@foxmail.com) and Jian Zhen Yu (jian.yu@ust.hk)

This document contains three supporting tables, nine supporting figures.

1 Comparison of three York regression implementations

A variety of York regression implementations are compared using the Pearson's data with York's weights according to York (1966) (abbreviated as "PY data" hereafter). The dataset is given in Table S2.Three York regression implementations are compared using the PY data, including spreadsheet by Cantrell (2008), Igor program by this study and a commercial software (OriginPro™ 2017). The three York regression implementations yield identical slope and intercept as shown in the highlighted areas (in red) in Figure S6. These crosscheck results suggest that the codes in our Igor program can retrieve consistent slopes and intercepts as other proven programs did.

2 Impact of two primary sources in OC/EC regression

A sampling site is often influenced by multiple combustion sources in the real atmosphere. In section 1 and 2 of the main text we evaluate the performance of OLS, DR, WODR and YR in scenarios of two primary sources and arbitrarily dictate that the $(OC/EC)_{pri}$ of source 1 is lower than that of source 2. By varying f_{EC1} (proportion of source 1 EC to total EC) from test to test, the effect of different mixing ratios of the two sources can be examined. Two scenarios are considered (Wu and Yu, 2016): two correlated primary sources and two independent primary sources. Common configurations include: EC_{total}=2 μgC m^{-3}; f_{EC1} varies from 0 to 100%; ratio of the two OC/EC_{pri} values (γ_{pri}) vary in the range of 2~8. Studies by Chu (2005) and Saylor et al. (2006) both suggest ratio of averages (ROA) being the best estimator of the expected primary OC/EC ratio when SOC is zeroed. Since the overall OC/EC_{pri} from the two sources varies by γ_{pri}, ROA is considered as the reference OC/EC_{pri} to be compared with slope regressed by of OLS, DR, WODR and YR. The abbreviations used for the two primary sources study are listed in Table S3.

2.1 Impact of two correlated primary sources

Simulations considering two correlated primary sources are performed, to examine the effect on bias in the regression methods. The basic configuration is: $(OC/EC)_{pri1}$=0.5, $(OC/EC)_{pri2}$=5, γ_{Unc}=30%, N=8000, intercept=0, and the following terms are compared:

ratio of averages (ROA here refers to the ratio of averaged OC to averaged EC, which is considered as the true value of slope when intercept=0), DR, WODR, WODR' (through origin) and OLS. As shown in Figure S7, when R^2 (EC1 vs. EC2) is very high, DR, WODR and WODR' can provide a result consistent with ROA. If the R^2 decreases, the bias of the slope and intercept in DR and WODR is larger. OLS constantly underestimates the slope.

2.2 Impact of two independent primary sources

Simulations of two independent primary sources are also conducted. If RSD_{EC1}=RSD_{EC2}, slopes and intercepts may be either overestimated or underestimated (Figure S8), and the degree of bias depends on the magnitude of RSD_{EC1} and RSD_{EC2}. Larger RSD results in larger bias. Uneven RSD between two sources leads to even more bias (Figure S8 a and b). The degree of bias also shows dependence on γ_pri. If γ_pri decreases, the bias becomes smaller (FigureS8 c~f). These results indicate that the scenario with two independent primary sources poses a challenge to $(OC/EC)_{pri}$ estimation by linear regression.

For the EC tracer method, if EC comes from two primary sources and contribution of the two sources is comparable, the regression slope is no longer suitable for $(OC/EC)_{pri}$ estimation and the subsequent SOC calculation, and making EC a mixture that violates the property of a tracer. For such a situation, pre-separation of EC into individual sources by other tracers (if available) by the Minimum R Squared (MRS) method can provide unbiased SOC estimation results (Wu and Yu, 2016).

3 Igor programs for error in variables linear regression and simulated OC EC data generation using MT

An Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) based program (Scatter plot) with graphical user interface (GUI) is developed to make the linear regression feasible and user friendly (Figure 8). The program includes Deming and York algorithm for linear regression, which considers uncertainties in both X and Y, that is more realistic for atmospheric applications. It is packed with many useful features for data analysis and plotting, including batch plotting, data masking via GUI, color coding in Z axis, data filtering and grouping by numerical values and strings.

Another program using MT can generate simulated OC and EC concentration through user defined parameters via GUI as shown in Figure S9.

Both Igor programs and their operation manuals can be downloaded from the following links:

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

https://doi.org/10.5281/zenodo.832417

References

Cantrell, C. A.: Technical Note: Review of methods for linear least-squares fitting of data and application to atmospheric chemistry problems, Atmos. Chem. Phys., 8, 5477-5487, 10.5194/acp-8-5477-2008, 2008.

Chu, S. H.: Stable estimate of primary OC/EC ratios in the EC tracer method, Atmos. Environ., 39, 1383-1392, 10.1016/j.atmosenv.2004.11.038, 2005.

Saylor, R. D., Edgerton, E. S., and Hartsell, B. E.: Linear regression techniques for use in the EC tracer method of secondary organic aerosol estimation, Atmos. Environ., 40, 7546-7556, 10.1016/j.atmosenv.2006.07.018, 2006.

Wu, C. and Yu, J. Z.: Determination of primary combustion source organic carbon-to-elemental carbon (OC/EC) ratio using ambient OC and EC measurements: secondary OC-EC correlation minimization method, Atmos. Chem. Phys., 16, 5453-5465, 10.5194/acp-16-5453-2016, 2016.

York, D.: Least-squares fitting of a straight line, Can. J. Phys., 44, 1079-1086, 10.1139/p66-090, 1966.

Table S1. Summary of the five linear regression techniques.

Approach	Sum of squared residuals (SSR)	Calculation
Ordinary least squares (OLS)	$\boldsymbol{S = \sum_{i=1}^{N} (y_i - Y_i)^2}$	closed form
Orthogonal distance regression (ODR)	$\boldsymbol{S = \sum_{i=1}^{N} [(x_i - X_i)^2 + (y_i - Y_i)^2]}$	iteration
Weighted orthogonal distance regression (WODR)	$\boldsymbol{S = \sum_{i=1}^{N} [(x_i - X_i)^2 + (y_i - Y_i)^2 / \eta]}$	iteration
Deming regression (DR)	$\boldsymbol{S = \sum_{i=1}^{N} [\omega(X_i)(x_i - X_i)^2 + \omega(Y_i)(y_i - Y_i)^2]}$	closed form
York regression (YR)	$\boldsymbol{S = \sum_{i=1}^{N} \left[\omega(X_i)(x_i - X_i)^2 - 2r_i\sqrt{\omega(X_i)\omega(Y_i)}(x_i - X_i)(y_i - Y_i) + \omega(Y_i)(y_i - Y_i)^2\right] \frac{1}{1 - r_i^2}}$	iteration

Table S2. Pearson's data with York's weights according to York (1966).

X_i	$\omega(X_i)$	Y_i	$\omega(Y_i)$
0	1000	5.9	1
0.9	1000	5.4	1.8
1.8	500	4.4	4
2.6	800	4.6	8
3.3	200	3.5	20
4.4	80	3.7	20
5.2	60	2.8	70
6.1	20	2.8	70
6.5	1.8	2.4	100
7.4	1	1.5	500

Table S3. Abbreviations used in two primary sources study.

Abbreviation	Definition
EC_1,EC_2	EC from source 1 and source 2 in the two sources scenario
f_{EC1}	fraction of EC from source 1 to the total EC
ROA	ratio of averages (Y to X, e.g., averaged OC to averaged EC)
γ_pri	ratio of the $(OC/EC)_{pri}$ of source 2 to source 1
RSD	relative standard deviation
RSD_{EC}	RSD of EC
ε_{EC} , ε_{OC}	measurement uncertainty of EC and OC
γ_{unc}	relative measurement uncertainty
$\gamma_{\_RSD}$	the ratio between the RSD values of $(OC/EC)_{pri}$ and EC

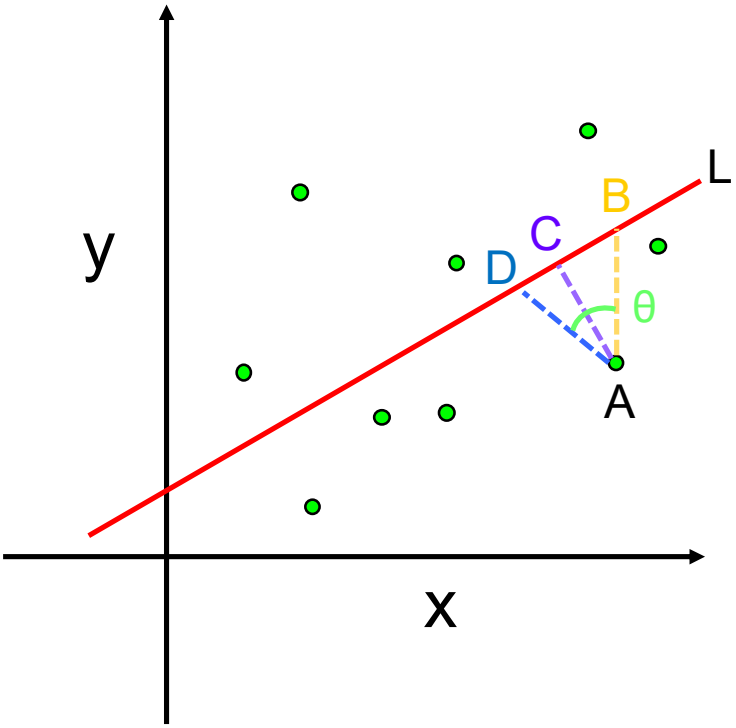


Figure S1. Relationships between data point A and fitting line L. Fitting line by OLS minimizes the distance of AB (AB is perpendicular to the x axis). Fitting line by ODR and DR ($\lambda = 1$) minimizes the distance of AC (AC is perpendicular to L). Fitting line by WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR minimizes the distance of AD. AD has a θ degree angle relative to AB and the θ depends on the weights of measurement errors in Y and X.

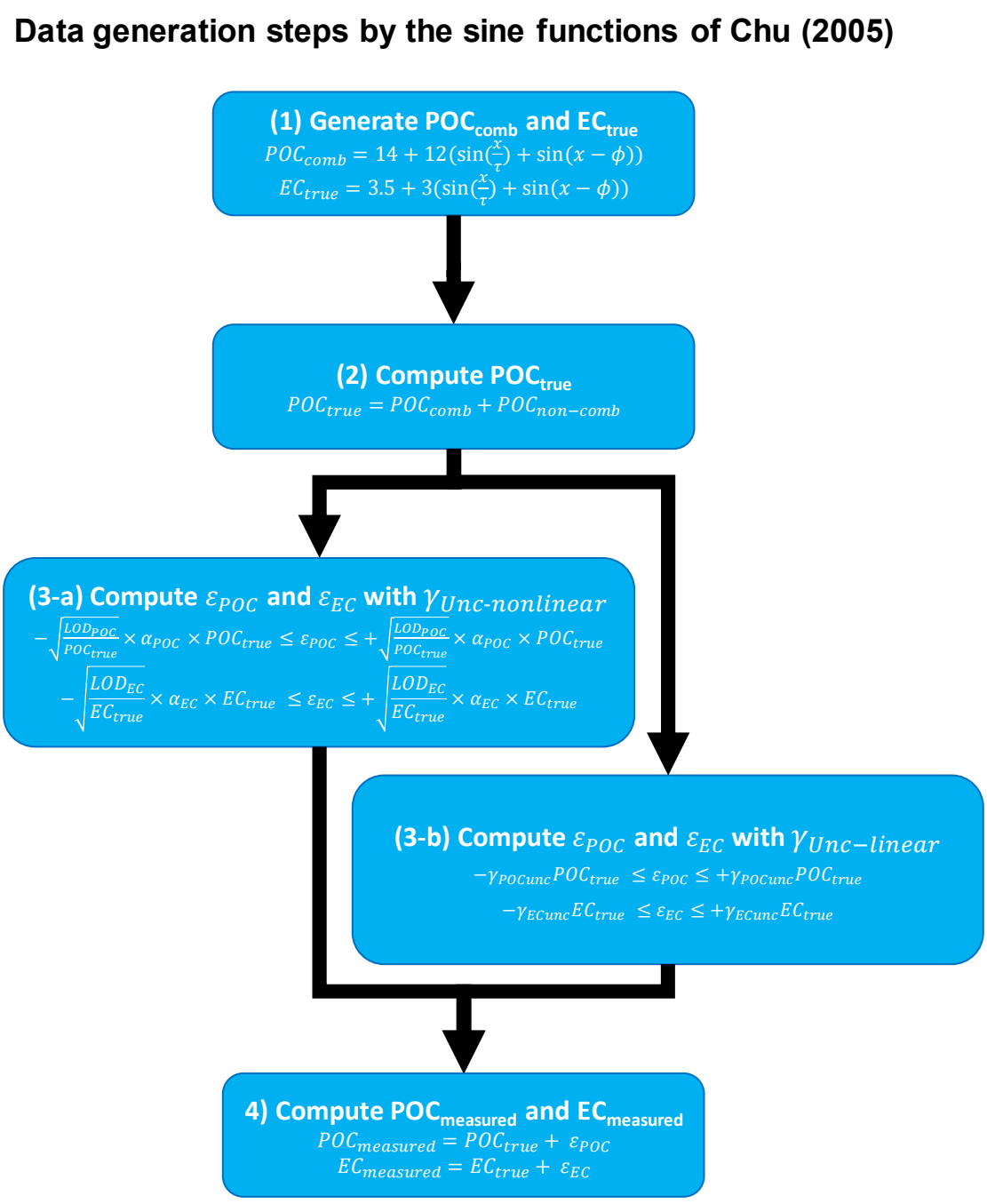


Figure S2. Flowchart of data generation steps using the sine functions of Chu (2005).

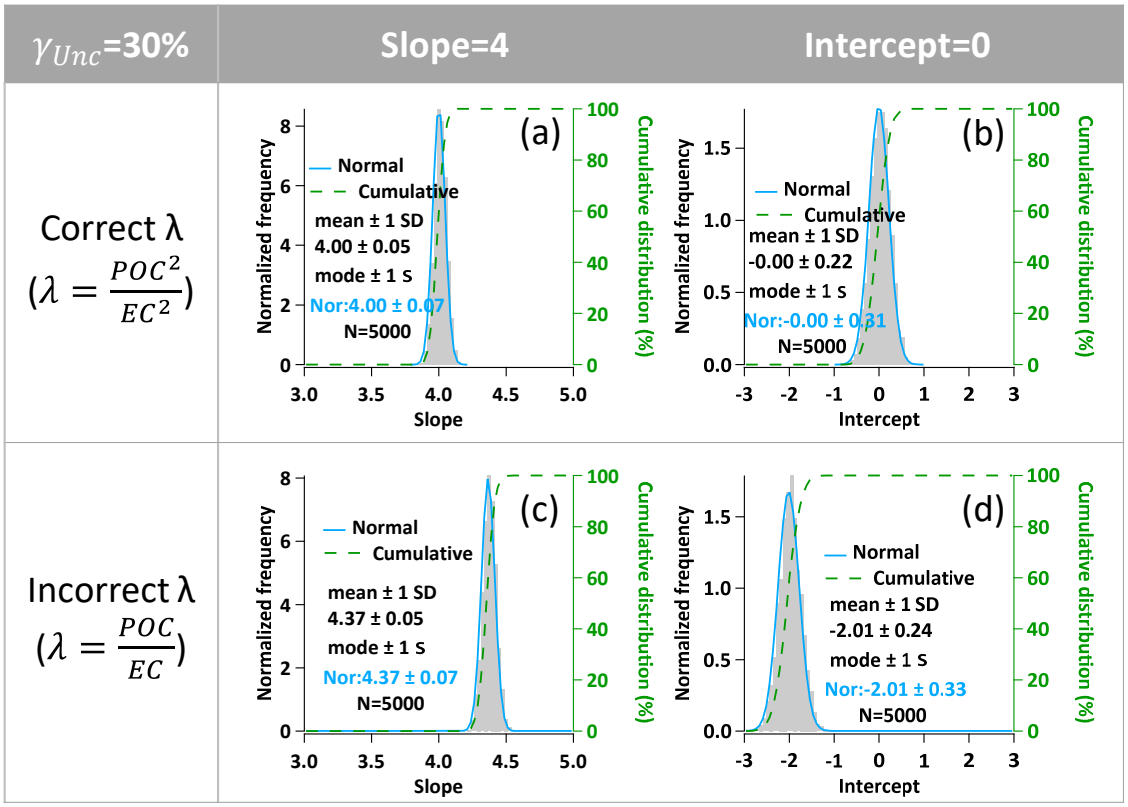


Figure S3. Example of bias in slope and intercept due to improper λ assignment. Data generation: Slope=4, Intercept=0; linear γ_{Unc} (30%). (a)&(b) Slopes and intercepts when proper λ is input following linear γ_{Unc} $(\lambda = \frac{POC^2}{EC^2})$; (c)&(d) Slopes and intercepts when improper λ is input following non-linear γ_{Unc} $(\lambda = \frac{POC}{EC})$.

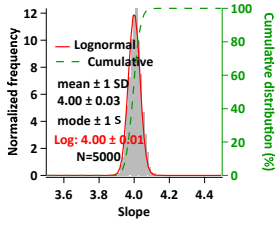
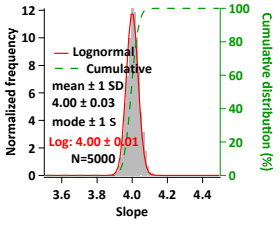
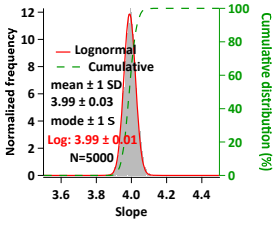
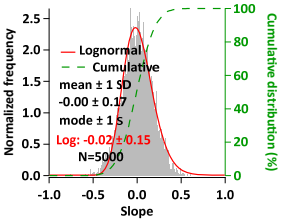
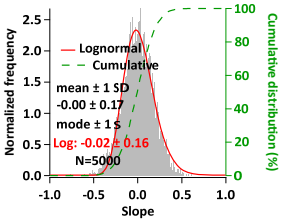
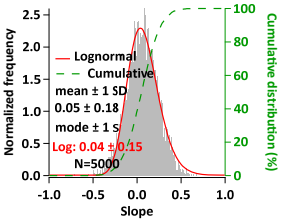
	λ by mean	λ by median	λ by mode
Slope=4			
Intercept=0			

Figure S4. Sensitivity tests of λ calculated by mean, median and mode.

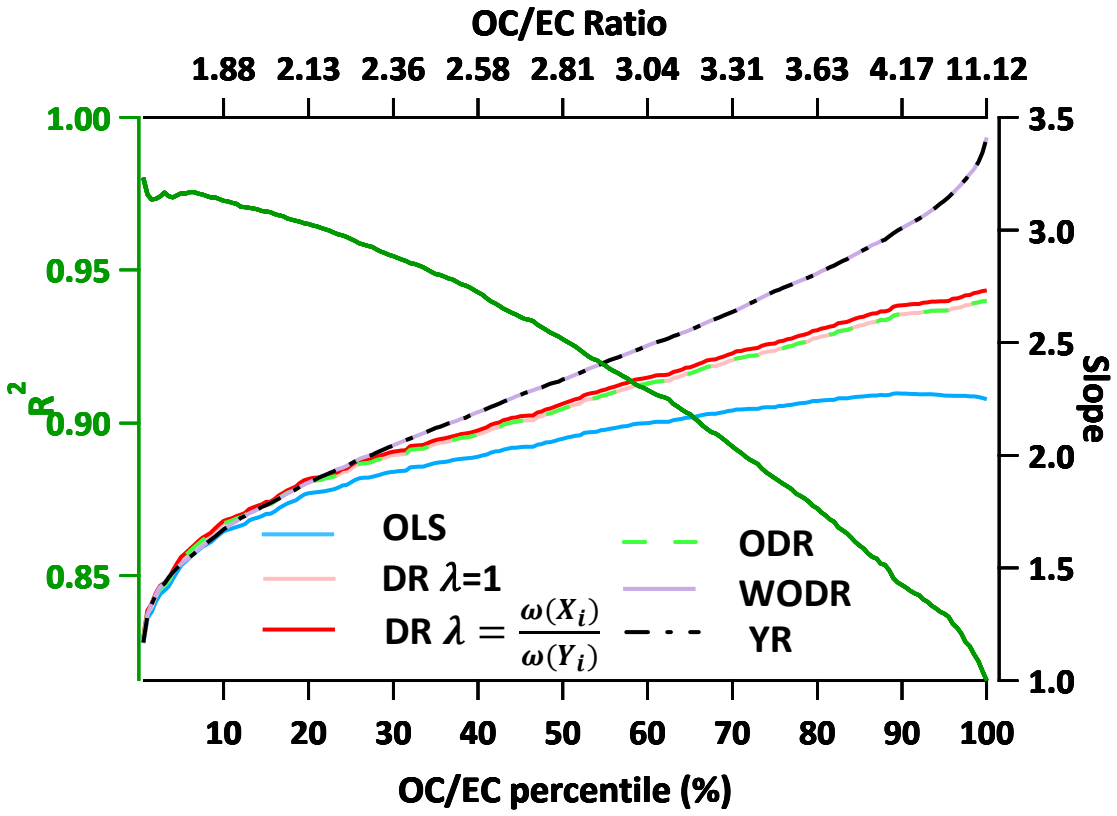


Figure S5. Regression slopes as a function of OC/EC percentile. OC/EC percentile range from 0.5% to 100%, with an interval of 0.5%.

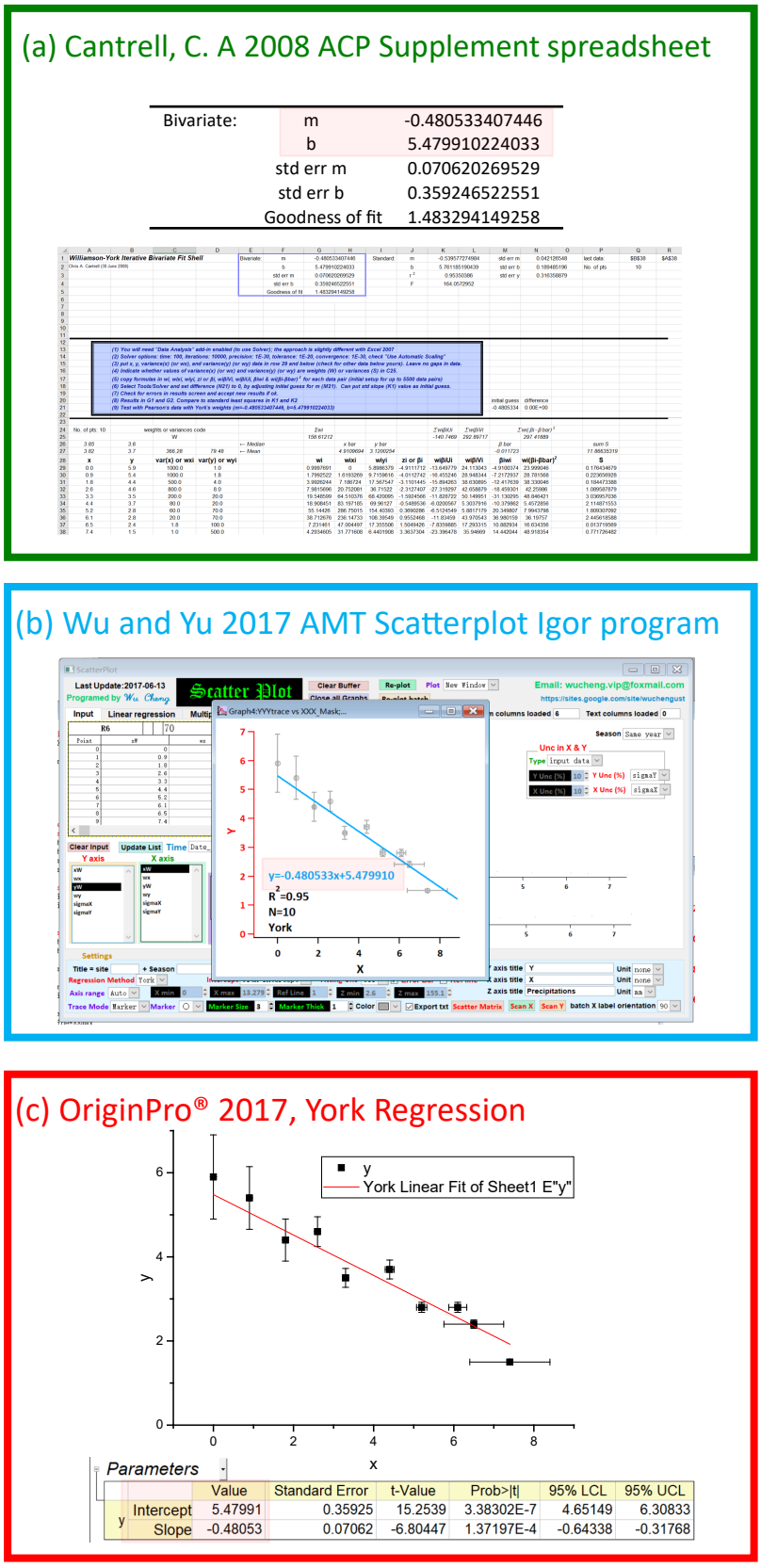


Figure S6. York regression implementations comparison using data shown in Table S2, including (a) spreadsheet by Cantrell (2008), (b) Igor program by this study and (c) a commercial software (OriginPro® 2017).

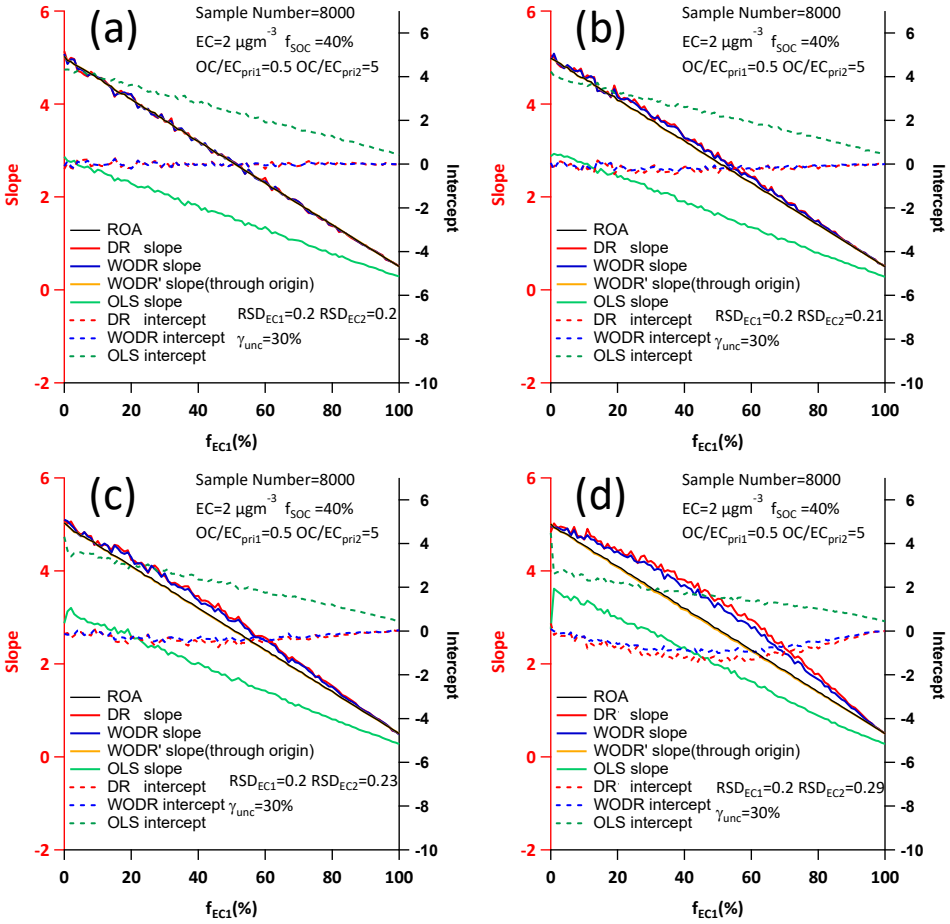


Figure S7. Study of two correlated sources scenario by different R^2 between the two sources. (a) $R^2 = 1$ (b) $R^2 = 0.86$ (c) $R^2 = 0.75$ (d) $R^2 = 0.49$.

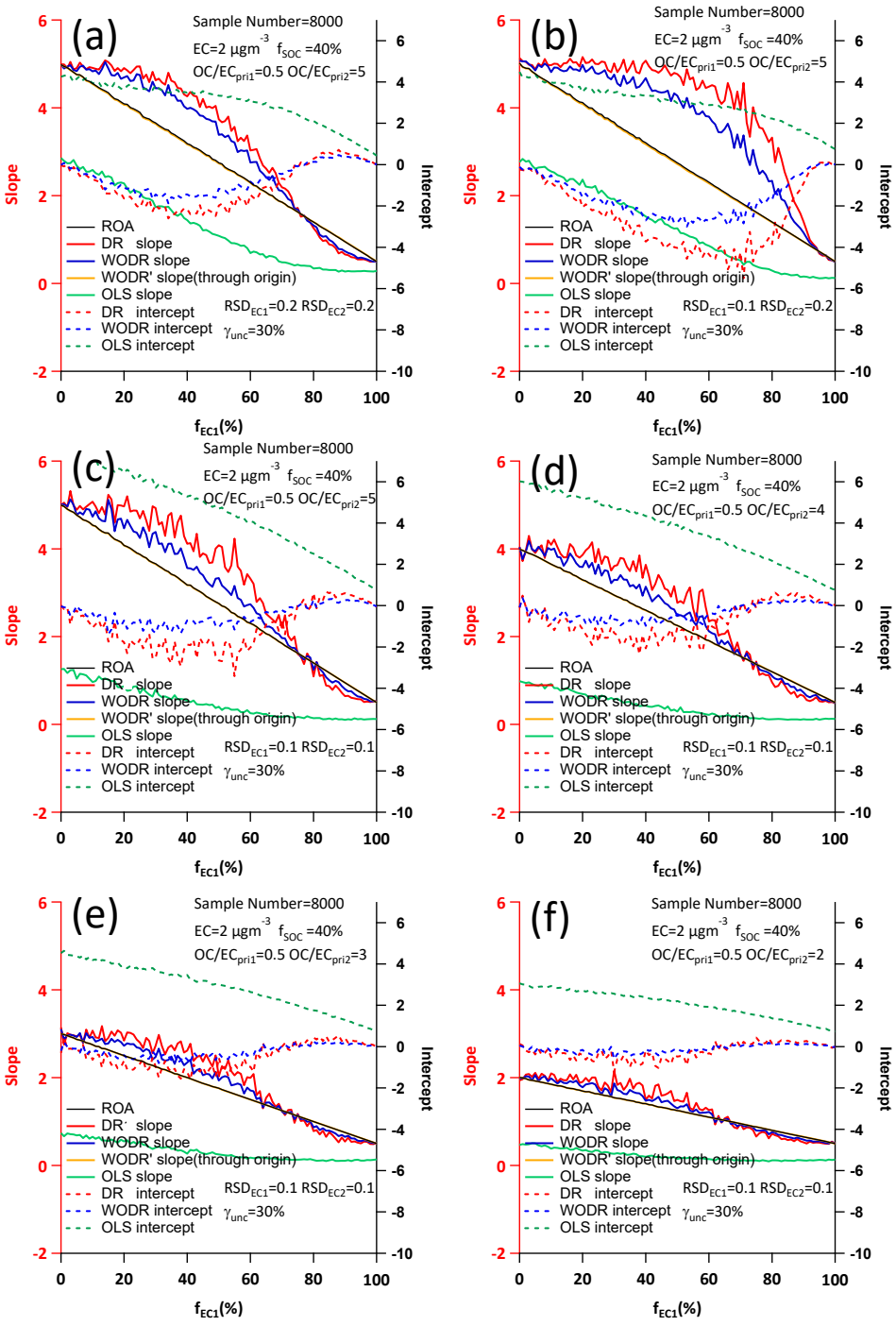
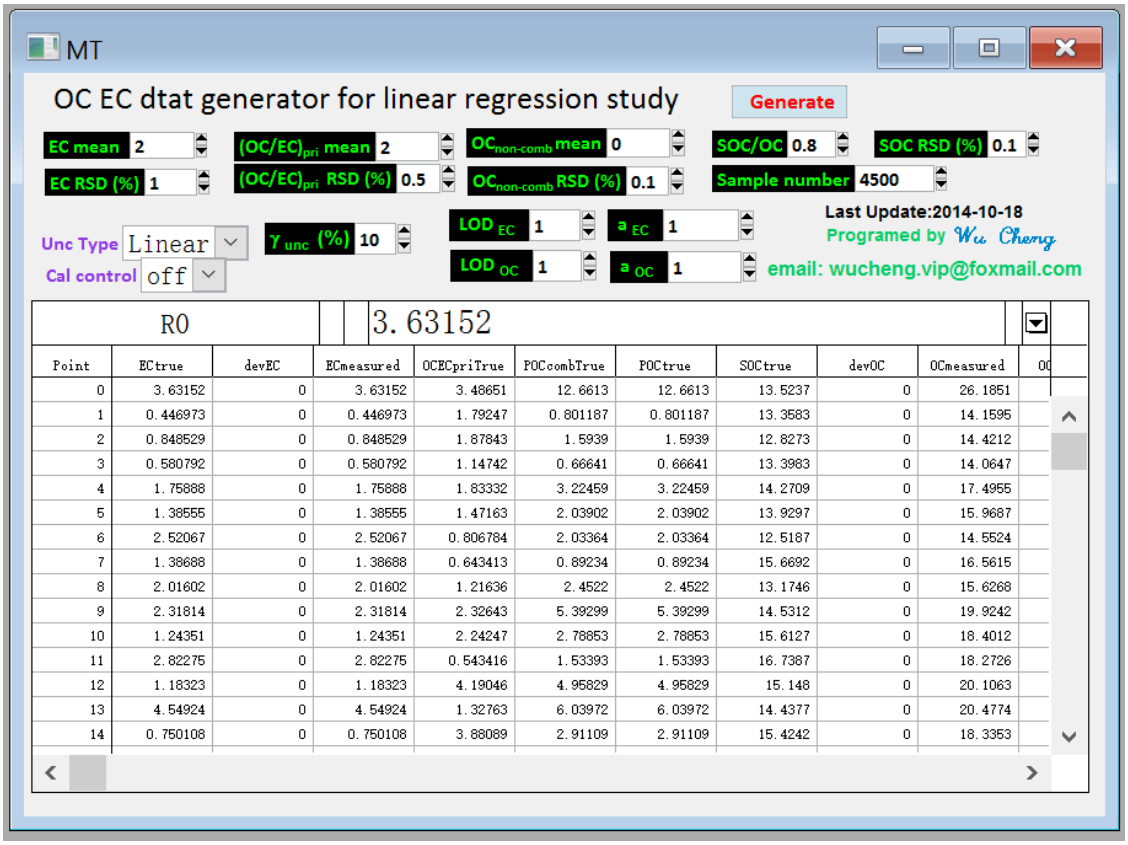


Figure S8. Study of two independent sources scenario by different parameters. (a) γ_pri=10, RSD_{EC1}=0.2, RSD_{EC2}=0.2 (b) γ_pri=10, RSD_{EC1}=0.1, RSD_{EC2}=0.2 (c) γ_pri=10, RSD_{EC1}=0.1, RSD_{EC2}=0.1 (d) γ_pri=8, RSD_{EC1}=0.1, RSD_{EC2}=0.1(e) γ_pri=6, RSD_{EC1}=0.1, RSD_{EC2}=0.1 (f) γ_pri=4, RSD_{EC1}=0.1, RSD_{EC2}=0.1.



Point	ECtrue	devEC	ECmeasured	OCECpriTrue	POCcombTrue	POCtrue	SOCtrue	devOC	OCmeasured	OC
0	3.63152	0	3.63152	3.48651	12.6613	12.6613	13.5237	0	26.1851	
1	0.446973	0	0.446973	1.79247	0.801187	0.801187	13.3583	0	14.1595	
2	0.848529	0	0.848529	1.87843	1.5939	1.5939	12.8273	0	14.4212	
3	0.580792	0	0.580792	1.14742	0.66641	0.66641	13.3983	0	14.0647	
4	1.75888	0	1.75888	1.83332	3.22459	3.22459	14.2709	0	17.4955	
5	1.38555	0	1.38555	1.47163	2.03902	2.03902	13.9297	0	15.9687	
6	2.52067	0	2.52067	0.806784	2.03364	2.03364	12.5187	0	14.5524	
7	1.38688	0	1.38688	0.643413	0.89234	0.89234	15.6692	0	16.5615	
8	2.01602	0	2.01602	1.21636	2.4522	2.4522	13.1746	0	15.6268	
9	2.31814	0	2.31814	2.32643	5.39299	5.39299	14.5312	0	19.9242	
10	1.24351	0	1.24351	2.24247	2.78853	2.78853	15.6127	0	18.4012	
11	2.82275	0	2.82275	0.543416	1.53393	1.53393	16.7387	0	18.2726	
12	1.18323	0	1.18323	4.19046	4.95829	4.95829	15.148	0	20.1063	
13	4.54924	0	4.54924	1.32763	6.03972	6.03972	14.4377	0	20.4774	
14	0.750108	0	0.750108	3.88089	2.91109	2.91109	15.4242	0	18.3353	

Figure S9. MT Igor program. OC and EC data following log-normal distribution can be generated for statistical study purpose (no time series information). User can define mean and RSD of EC, $(OC/EC)_{pri}$, SOC/OC ratio, measurement uncertainty, sample size, etc. MT Igor program can be downloaded from the following link: https://sites.google.com/site/wuchengust.