基于污染物类别、监测点位置和时间的北京空气质量分析

颜子杰 2014012217
张智博 2014012201
廖安迪 2013013006

一. 研究内容及意义

本文从污染物类别、监测点位置和时间三个维度对北京空气质量进行描述性研究，旨在从数据中发现一定的规律甚至一些违背常理的结果，由于水平有限恐不能分析的深入且全面，对某些现象也只是有相关的猜测不能说是完全的科学的解释。

本文的具体内容包括：数据可视化（某种污染物某段时间浓度曲线、两种污染物/监测点对比浓度曲线）、由已知数据对 AQI 的重新计算、各监测点空气质量情况、空气污染天数及分布情况等。

作为一个生活在北京的学生，对北京的空气状况自然十分重视，而在我们一再抱怨空气质量的同时如果能从较为客观的数据入手，是否会看到较平时不同的结果，这是我们出发的初衷。此外，北京每天每小时的污染物浓度的数据量很大、分类较多、数据为数值格式，因此是一个好的作为用 R 进行数据处理的样本。

尽管我们时间和水平有限，无法分析的面面俱到，但希望结果还是能让大家对北京的空气质量情况有一个大概、客观的认识，更希望能给大家一些启发。

二．数据来源及背景

我们采用的是 2014 年 1 月 1 日 0 时至 2016 年 11 月 26 日 23 时的北京空气质量历史数据，网址为 http://beijingair.sinaapp.com/。数据集中包含时间，污染物实时浓度，污染物 24 小时均值，空气质量指数（air quality index, AQI）及监测点名称。

在该数据集中，除 AQI 为计算所得外，其余数据均为测量所得。其中，时间包含年月日，精确到每小时；污染物实时浓度及污染物 24 小时均值所包含的指标如表 1 所示；监测点的详细信息如表 2 所示。

表 1 污染物指标的中文含义、单位及对应英文

污染物指标	污染物项目 P	单位	对应英文
污染物实时浓度	颗粒物（粒径<=2.5 um）1 小时平均	$\mu g/m^3$	PM2.5
	颗粒物（粒径<=10 um）1 小时平均	$\mu g/m^3$	PM10
	二氧化硫 1 小时平均	$\mu g/m^3$	SO2
	二氧化氮 1 小时平均	$\mu g/m^3$	NO2
	臭氧 1 小时平均	$\mu g/m^3$	O3
	一氧化碳 1 小时平均	mg/m^3	CO
污染物 24 小时均值	颗粒物（粒径<=2.5 um）24 小时平均	$\mu g/m^3$	PM2.5_24h
	颗粒物（粒径<=10 um）24 小时平均	$\mu g/m^3$	PM10_24h
	二氧化硫 24 小时平均	$\mu g/m^3$	SO2_24h
	二氧化氮 24 小时平均	$\mu g/m^3$	NO2_24h
	臭氧 24 小时平均	$\mu g/m^3$	O3_24h
	一氧化碳 24 小时平均	mg/m^3	CO_24h

表 2 监测点详细信息

城市环境评价点	郊区环境评价点	对照点及区域点	交通污染监测点
东四	房山	昌平定陵	前门东大街，前门交通点
天坛	大兴	京西北八达岭，京西北区域点	永定门内大街，永定门交通点
官园	亦庄	京东北密云水库，京东北区域点	西直门北大街，西直门交通点
万寿西宫	通州	京东东高村，京东区域点	南三环西路，南三环交通点
奥体中心	顺义	京东南永乐店，京东南区域点	东四环北路，东四环交通点
农展馆	昌平	京南榆垡，京南区域点	
万柳	门头沟	京西南琉璃河，京西南区域点	
北部新区	平谷		
植物园	怀柔		
丰台花园	密云		
云岗	延庆		
古城			

三． 数据预处理

我们获得的原始数据为一份份单独的 csv 文件，每份文件记录着当天的空气质量监测数据，对各份文件进行读取后所得的数据表如图 1 所示。

图 1 原始数据读取结果

	date	hour	type	东四	天坛	官园	万寿西宫	奥体中心	农展馆	万柳	北部新区	植物园	丰台花园	云岗	古城	房山	大兴
1	20140101	0	PM2.5	35	32	45	66	20	31	57	22	17	80	67	46	129	179
2	20140101	0	PM2.5_24h	53	50	53	59	51	51	65	39	32	63	47	42	88	100
3	20140101	0	PM10	114	110	151	175	90	117	152	70	62	182	167	126	NA	265
4	20140101	0	PM10_24h	131	124	147	151	221	153	174	115	109	171	145	125	NA	175
5	20140101	0	AQI	91	87	99	101	136	102	112	83	80	111	98	88	116	131
6	20140101	1	PM2.5	66	56	57	72	43	58	68	22	25	83	77	58	130	159
7	20140101	1	PM2.5_24h	51	48	50	56	49	50	62	36	31	60	47	39	87	100
8	20140101	1	PM10	154	126	150	174	348	154	179	62	52	185	183	156	NA	230
9	20140101	1	PM10_24h	133	124	148	150	229	154	173	112	108	171	147	126	NA	178
10	20140101	1	AQI	92	87	99	100	140	102	112	81	79	111	99	88	115	131
11	20140101	2	PM2.5	70	53	57	80	79	68	81	21	15	94	50	59	124	152
12	20140101	2	PM2.5_24h	49	45	47	54	47	48	59	33	29	57	45	38	86	97
13	20140101	2	PM10	144	119	151	192	423	162	240	63	46	207	159	145	NA	227
14	20140101	2	PM10_24h	133	124	148	149	235	155	174	110	107	171	147	127	NA	179
15	20140101	2	AQI	92	87	99	100	143	103	112	80	79	111	99	89	114	128

Showing 1 to 15 of 49,246 entries

之后为了处理方便，将各地点名改为了英文，并利用 tidyr 包中的 gather 函数将地点变量抽取出来单独作为一列，方便之后对数据进行筛选。初步整理之后的数据如图 2 所示。

图 2 数据初步整理结果

	date	hour	type	place	value
1	20140101	0	PM2.5	DongSi	35
2	20140101	0	PM2.5_24h	DongSi	53
3	20140101	0	PM10	DongSi	114
4	20140101	0	PM10_24h	DongSi	131
5	20140101	0	AQI	DongSi	91
6	20140101	1	PM2.5	DongSi	66
7	20140101	1	PM2.5_24h	DongSi	51
8	20140101	1	PM10	DongSi	154
9	20140101	1	PM10_24h	DongSi	133
10	20140101	1	AQI	DongSi	92
11	20140101	2	PM2.5	DongSi	70
12	20140101	2	PM2.5_24h	DongSi	49
13	20140101	2	PM10	DongSi	144
14	20140101	2	PM10_24h	DongSi	133
15	20140101	2	AQI	DongSi	92
16	20140101	3	PM2.5	DongSi	71
17	20140101	3	PM2.5_24h	DongSi	46

Showing 1 to 9 of 10,822,245 entries

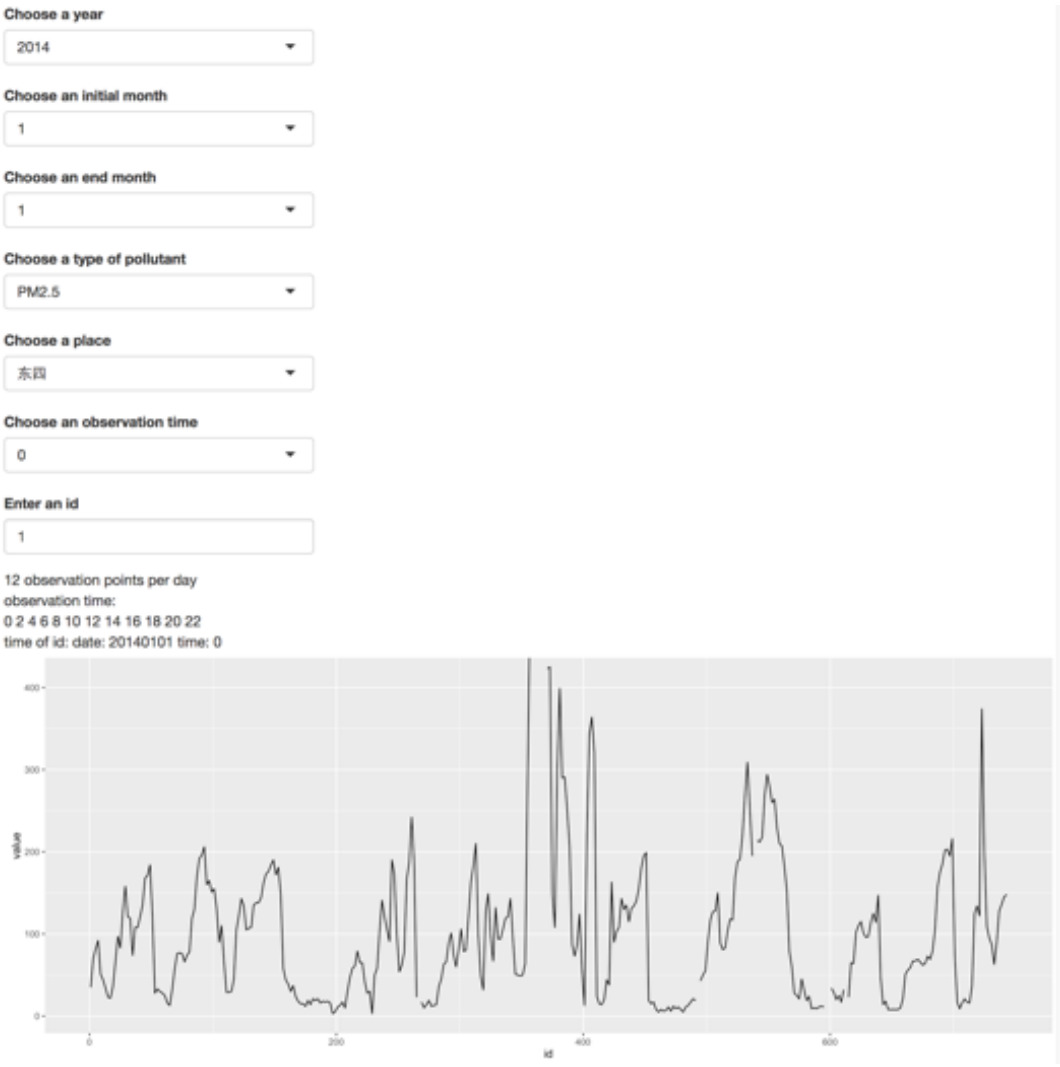
四. 数据可视化

在此次研究的问题中，我们主要目的之一为观察污染物随时间的变化关系，以及不同地点不同污染物的变化趋势，因而可视化部分主要针对这几点对数据进行处理，利用 ggplot2 进行绘图，再制作成为 shiny 页面，便于根据需要调整绘图变量，便于研究感兴趣的数据，为之后的工作提供基础。整个 shiny 页面分为四个部分，实现了不同的功能。

4.1 第一部分

第一张图的目的是对数据进行大体的预览，其界面如图 3 所示。通过选择观察的年月区间、污染物种类、测量地点，即可得到污染物随时间的变化曲线，利用此图可以大致反映污染物的变化趋势，从而找寻感兴趣的时间区间进行下一步研究。

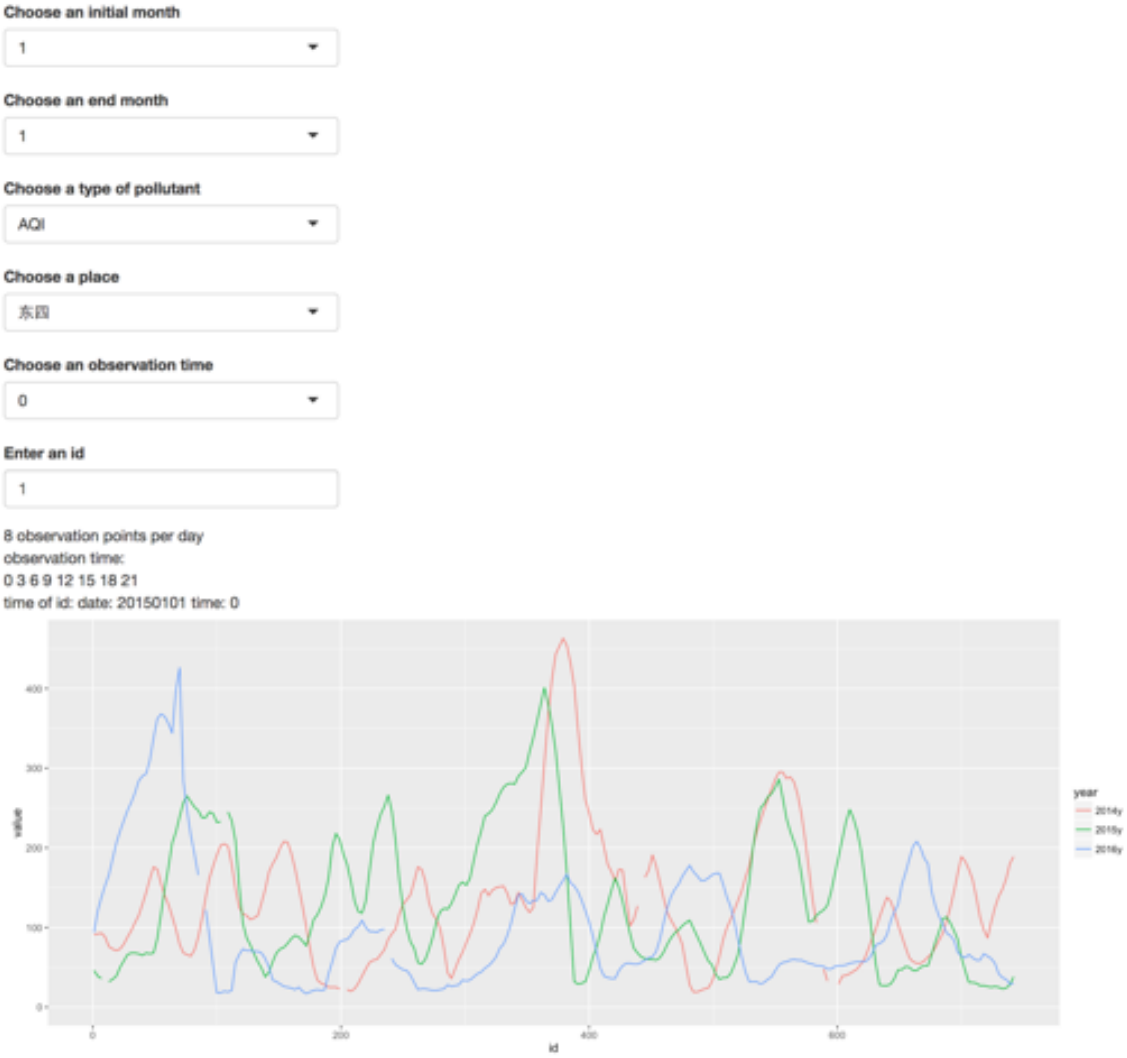
图 3 第一部分 shiny 界面



4.2 第二部分

通过第一张图我们对污染物的大致变化情况有了一定的了解，之后的第二张图实现的功能则是对不同年份的数据进行对比，找到共性的变化以及一些特殊的地方，其界面如图 4 所示。其中各窗格功能与第一部分类似，主要实现对三年数据的综合对比。

图 4 第二部分 shiny 界面



4.3 第三部分

为了对比不同观测点的空气质量数据，从而探究空气污染与地区间的关系，第三部分的思路是实现不同地区污染物随时变化的对比，其界面如图 5 所示。通过在界面中选择观测区间、观测地点和污染物种类，即可实现两地三年污染物随时变化情况的对比，以此发现两者间的联系与差别，从而进行进一步分析。

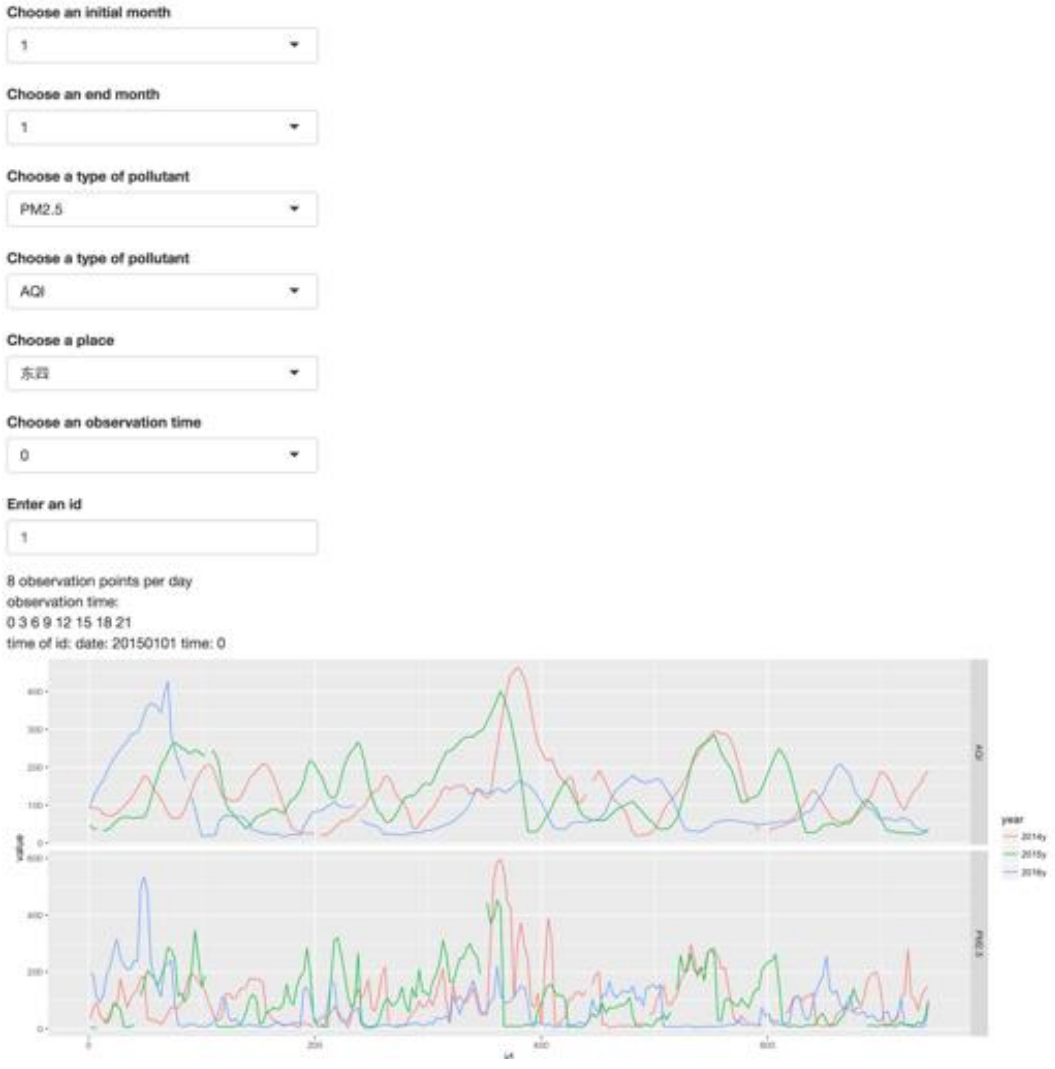
图 5 第三部分 shiny 界面

4.4 第四部分

除了对比地理位置因素，污染物种类也是一个很好的研究变量，此部分的目的是通过对比同一观测地点不同种类污染物的随时变化情况来比较各污染物的变化快慢及其峰值出现的次序等，以此推断各种污染物浓度间的联系。其界面如图 6 所示。

通过纵向对比两种污染物随时变化的曲线，比较其峰值位置、变化快慢等可定性得出两者间的一些关系，如果将其中一个选为 AQI，还可探究何种污染物与 AQI 变化趋势最吻合，从而得到出 AQI 影响较大的污染物种类。

图 6 第四部分 shiny 界面



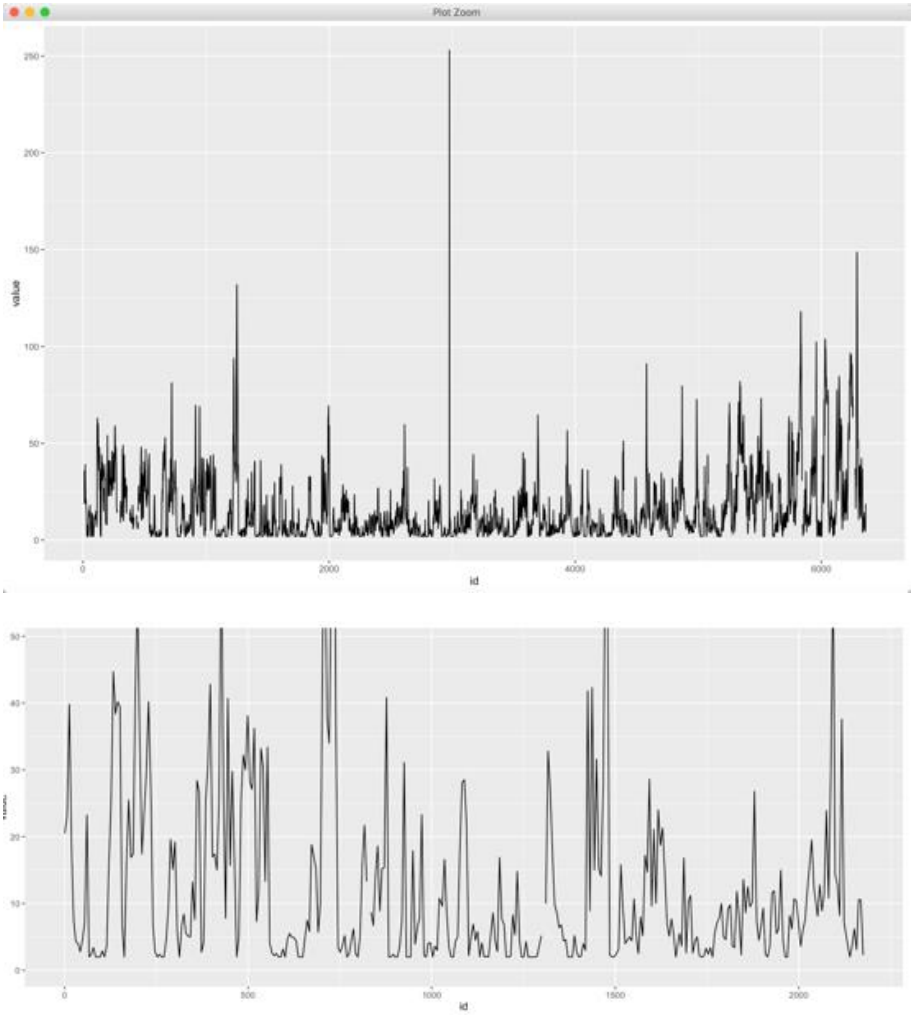
4.5 实现方法及细节

在各部分的绘图工作中，基本方法为用输入的条件利用 subset 函数对总体数据进行筛选，得到符合要求的数据集，再用 geom_path 函数绘制变化曲线，用 aes(color)或 facet_grid 实现曲线对比，最后通过 observe 函数实现动态变化。

在此过程中，有以下几点进行了改进：

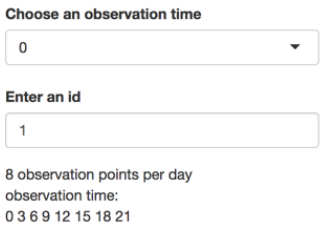
1) 为了避免数据中部分异常值影响绘图坐标轴范围的选取（部分异常值远大于其他数据，使得坐标轴过长，导致大部分数据的变化难以观察），因而在使用 ggplot2 绘图时添加了 coord_cartesian 函数，将坐标轴上限设置为四倍数据平均值与最大值取小，以此减小异常值带来的误差。（大于四倍均值的数据均属于空气污染较为严重的情况，此时比较具体峰值的大小已无太大意义，因此可对坐标轴上限进行限制）。更改前后对比如图 7 所示。

图 7 更改前后对比图



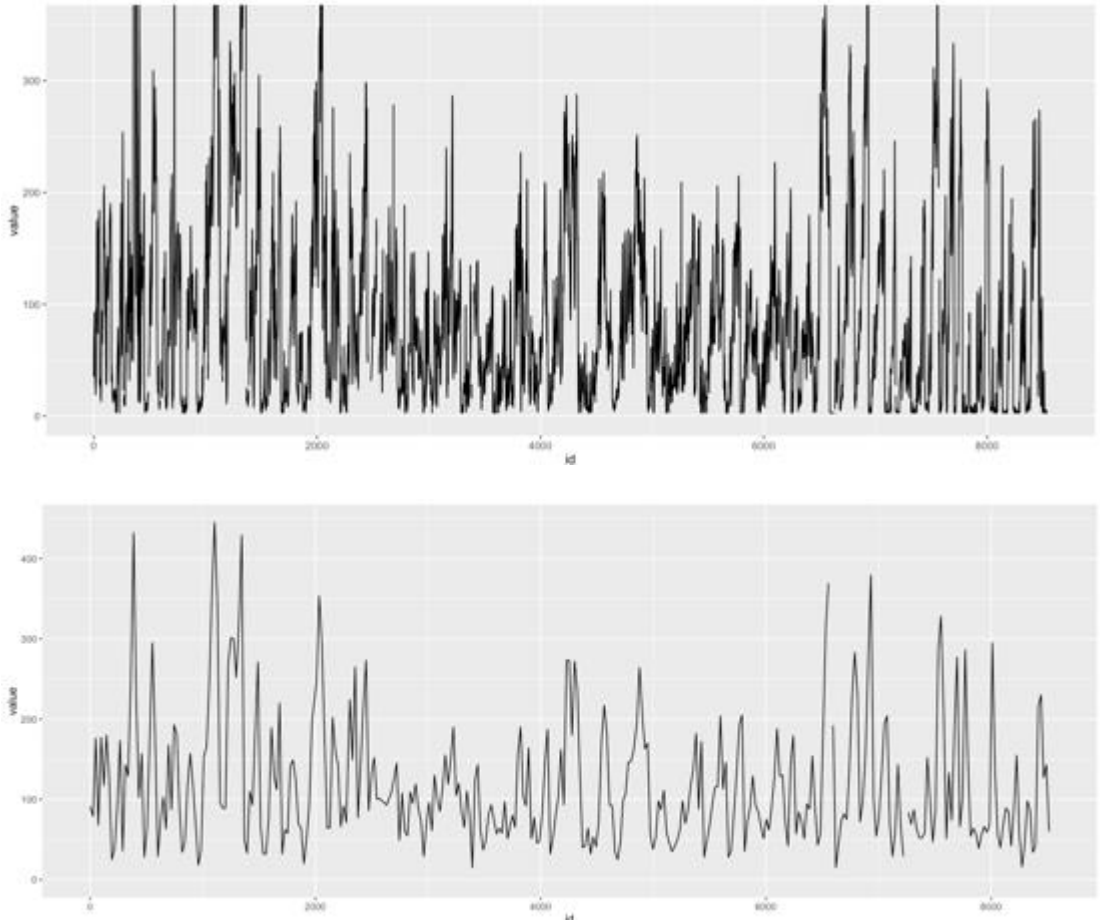
2) 由于所使用数据采样较密（一天 24 个样本），所以当选取的绘图时间段较长时，图像上的数据点过多，图像过于密集，无法看出变化趋势。针对这一情况，我们采取的措施是根据选择的时间长短不同，调整每天的取样个数，只取部分数据点作为代表（例如当绘制一个月的变化曲线时，一天取 12 个代表点，绘制一年变化曲线时，一天只取一个代表点）。与此同时，为了保留数据的全部信息，避免只能对比一天的某个时间点，因而加入了采样点选择功能，用户可选择每天的观测时间，从而得到不同时间的观测曲线。窗口如图 8 所示。

图 8 采样点选择功能



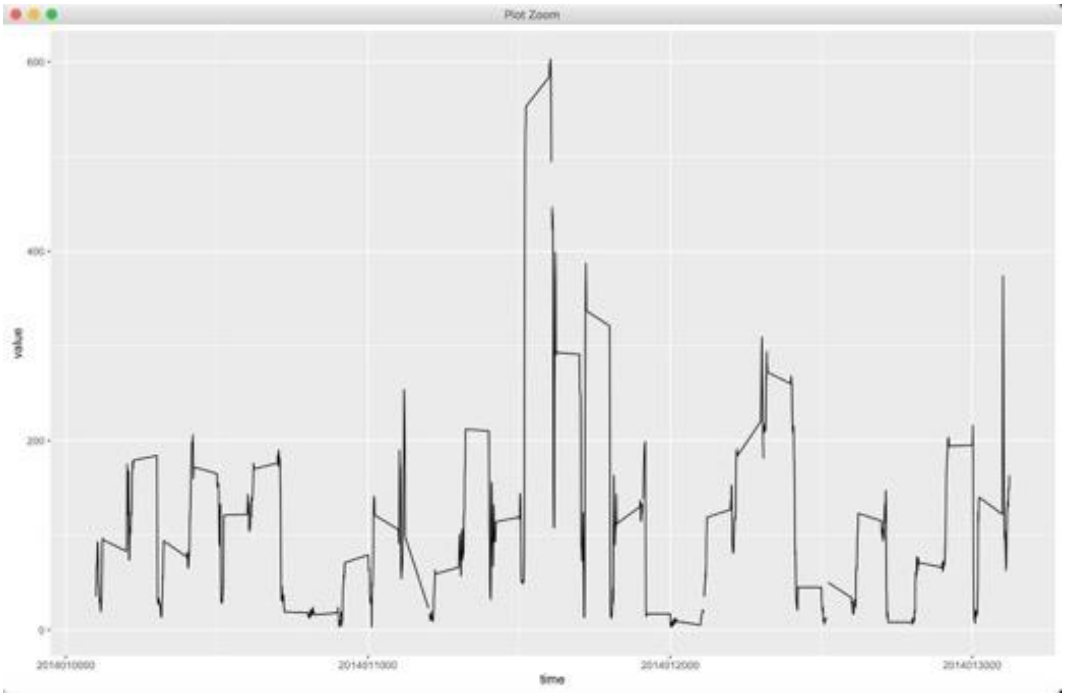
下方输出的文本信息可告知用户每天的观测点个数以及观测时间。使用此方法前后的效果对比如图 9 所示。

图 9 采样点选取方法改进前后效果对比图



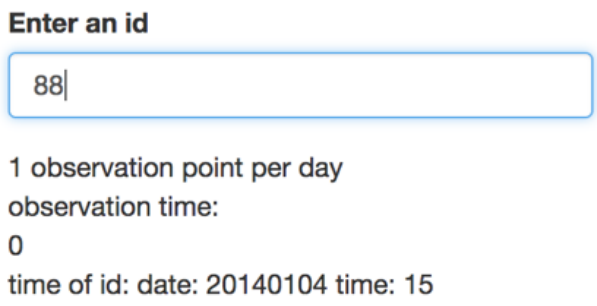
3) 在绘图过程中，由于数据的时间变量为日期，如果直接以时间作为横轴进行绘图会导致系统将时间变量当做连续变量（即实际上 1 月 31 日（0131）后为 2 月 1 日（0201），而系统则会认为 0132-0200 均为缺失值），由此得到的曲线如图 10 所示。

图 10 时间作为连续变量的绘图效果



因而在绘图前，我们首先将筛选出的数据按照时间顺序依次加上”id”,以 id 为横坐标进行绘制，同时，为了获得图中 id 与具体时间的对应关系，在 shiny 中加入了一个窗口，可以通过输入 id 得到该 id 对应的时间，从而实现了横坐标与时间的一一对应关系，最终结果如图 11 所示。

图 11 id 与日期的对应



五. AQI 重新计算

该部分主要通过 AQI 的重新计算出发，与原始数据集中的 AQI 进行比较，从而尝试判断北京空气污染的主要污染物。

根据环境保护部发布的环境空气质量指数（AQI）技术规定（试行），污染物项目 P 的空气质量分指数IAQI_p的计算公式为：

$$\mathrm{IAQI}_p = \frac{\mathrm{IAQI}_{Hi} - \mathrm{IAQI}_{Lo}}{\mathrm{BP}_{Hi} - \mathrm{BP}_{Lo}}(C_P - \mathrm{BP}_{Lo}) + \mathrm{IAQI}_{Lo}$$

其中，

IAQI_p代表污染物项目 P 的空气质量分指数；

C_P代表污染物项目 P 的质量浓度值；

BP_{Hi}代表图 12 中与C_P相近的污染物浓度限值的高位值；

BP_{Lo}代表图 12 中与C_P相近的污染物浓度限值的低位值；

IAQI_{Hi}代表图 12 中与BP_{Hi}对应的空气质量分指数；

IAQI_{Lo}代表图 12 中与BP_{Lo}对应的空气质量分指数。

空气质量指数AQI的计算公式为：

$$\mathrm{AQI} = \max\{\mathrm{IAQI}_p\}, \mathrm{p} = 1,2,\ldots \mathrm{n}。$$

因此，通过已有的污染物实时浓度及平均浓度按照以上两个公式重新计算，理论上应该得到与数据集中的 AQI 相近的结果。我们猜测，不同污染物对空气质量分数 AQI 的贡献可能不同，其中 PM2.5 及 PM10 可能是突出最为贡献的污染物。因此，我们采用 PM2.5、PM2.5_24h、PM10 及 PM10_24h 进行空气质量指数 AQI_NEW 的计算，与数据集中的 AQI 进行比较。

图 12 污染物项目浓度限值

空气质量分指数（IAQI）	污染物项目浓度限值									
	二氧化硫（SO_2）24 小时平均/（μg/m³）	二氧化硫（SO_2）1 小时平均/（μg/m³）(1)	二氧化氮（NO_2）24 小时平均/（μg/m³）	二氧化氮（NO_2）1 小时平均/（μg/m³）(1)	颗粒物（粒径小于等于10μm）24 小时平均/（μg/m³）	一氧化碳（CO）24 小时平均/（mg/m³）	一氧化碳（CO）1 小时平均/（mg/m³）(1)	臭氧（O_3）1 小时平均/（μg/m³）	臭氧（O_3）8 小时滑动平均/（μg/m³）	颗粒物（粒径小于等于2.5μm）24 小时平均/（μg/m³）
0	0	0	0	0	0	0	0	0	0	0
50	50	150	40	100	50	2	5	160	100	35
100	150	500	80	200	150	4	10	200	160	75
150	475	650	180	700	250	14	35	300	215	115
200	800	800	280	1 200	350	24	60	400	265	150
300	1 600	(2)	565	2 340	420	36	90	800	800	250
400	2 100	(2)	750	3 090	500	48	120	1 000	(3)	350
500	2 620	(2)	940	3 840	600	60	150	1 200	(3)	500

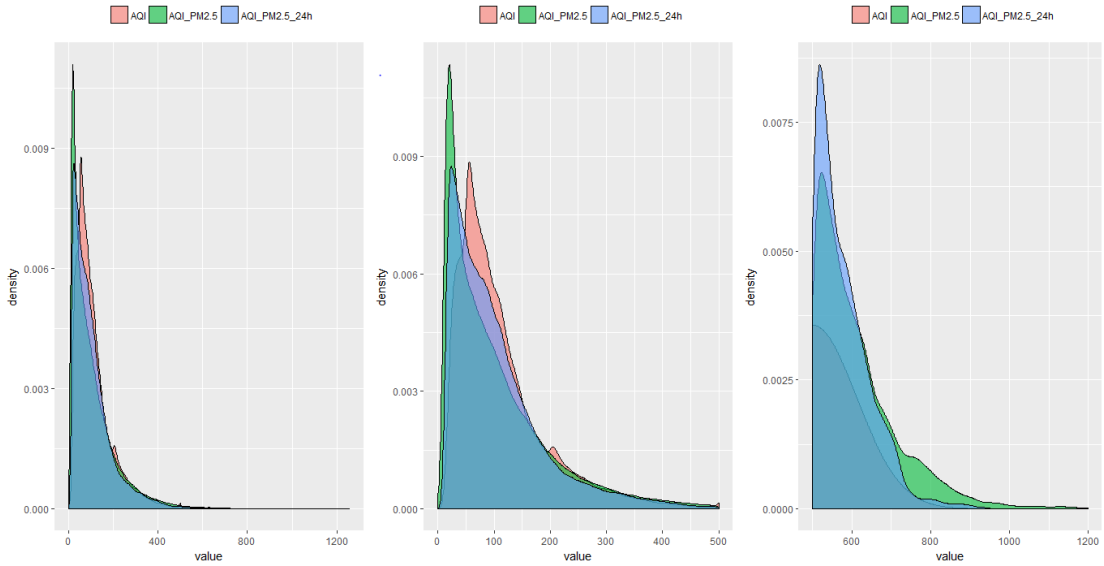
5.1 PM2.5 及 PM2.5_24h

1) 总体分布

在剔除异常值之后，将由 PM2.5 及 PM2.5_24h 所计算出的 AQI_NEW 和原始数据集中的 AQI 的概率密度曲线进行比较，如图 13(左)所示，PM2.5_AQI_NEW 或 PM2.5_24h_AQI_NEW 大体趋势与原始 AQI 一致，但存在一定差异。由于平日里 AQI 的最大值被人为设置为 500，因此我们将 AQI=500 作为分界线，分别观察当 AQI<500 及 AQI>500 时，PM2.5 及 PM2.5_24h 所计算出的 AQI_NEW 与原始 AQI 的差异情况。

如图 13(中)所示，当 0<AQI<150 时，PM2.5_AQI_NEW 及 PM2.5_24h_AQI_NEW 所对应的峰值小于原始 AQI 所对应的峰值，说明当空气质量良好时，PM2.5 或 PM2.5_24h 均不是影响 AQI 的主要因素；当 150<AQI<500 时，PM2.5_AQI_NEW 及 PM2.5_24h_AQI_NEW 与原始 AQI 差异较小，概率分布曲线较为接近，说明在空气质量一般时，PM2.5 或 PM2.5_24h 均是影响 AQI 的主要因素。如图 13(右)所示，当 500<AQI<1200 时，PM2.5_AQI_NEW 及 PM2.5_24h_AQI_NEW 所对应的概率大于原始 AQI 所对应的概率，由 PM2.5 及 PM2.5_24h 所引起的空气严重污染的概率大大高于官方确认的“爆表”概率。

图 13 PM2.5/PM2.5_24h 所得 AQI_NEW 及原始 AQI 的概率密度图

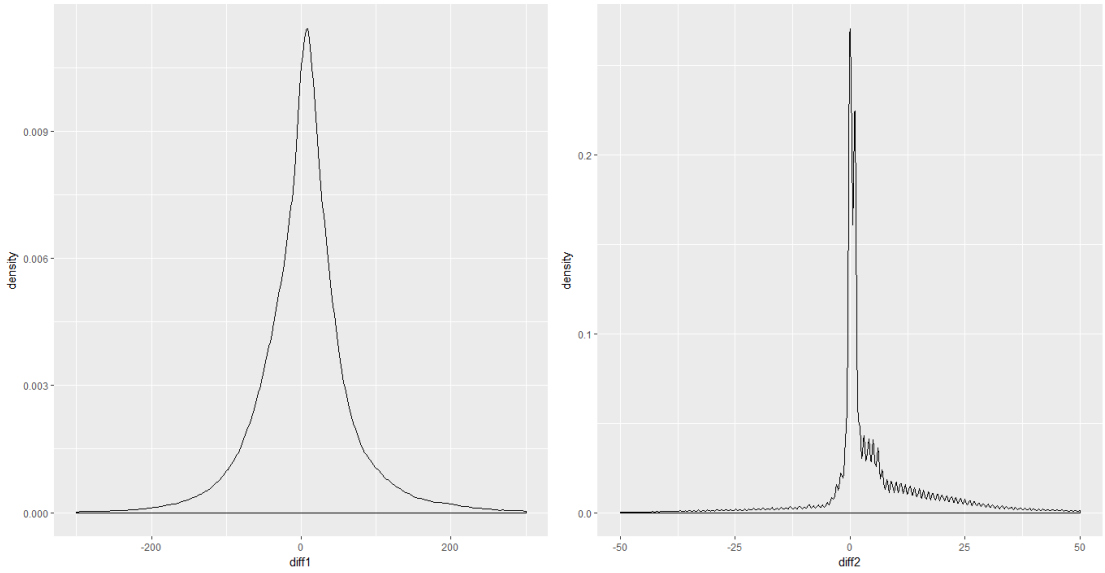


2) 差值比较

将 PM2.5_AQI_NEW 及 PM2.5_24h_AQI_NEW 与原始 AQI 分别作差，得出 diff1 及 diff2，因为样本量足够大，因此可以将 diff1 及 diff2 视为正态分布进行考察，如图 14 所示。对于 PM2.5_AQI_NEW 与原始 AQI 所得的 diff1 而言，平均值为 4.64，方差为 66.83；而对于 PM2.5_24h_AQI_NEW 与原始 AQI 所得的 diff2 而言，平均值为 5.27，方差为 19.46。

可以发现，diff1 及 diff2 的平均值均为正数，说明使用 PM2.5 或 PM2.5_24h 所计算的 AQI_NEW 总体高于原始 AQI，这一结果能够从图 2 中得到解释，即 PM2.5 或 PM2.5_24h 所计算的 AQI_NEW 在空气污染严重时概率较高，因此拉高了总体平均值。此外，diff2 的方差远小于 diff1 的方差，说明使用 PM2.5_24h 所计算的 AQI_NEW 较为稳定，因而与原始 AQI 更为接近。

图 14 diff1 及 diff2 的概率密度图



3) 正确匹配

当 PM2.5_AQI_NEW 及 PM2.5_24h_AQI_NEW 与原始 AQI 的差值 diff 足够小时，可以视为一次正确匹配。为寻找正确匹配所适合的误差阈限，我们将 diff 1 及 diff2 在 0，±1，±2，±3，±4，±5 时的频数进行统计，如表 3 所示。当误差阈限严格为 0 时，PM2.5 所计算的 AQI_NEW 只有 1.02%能够与原始 AQI 正确匹配，而 PM2.5_24h 所计算的 AQI_NEW 有 21.05%能够与原始 AQI 正确匹配；随着误差阈限的放松，diff1 及 diff2 的累积正确匹配率均在上升，当误差阈限放宽至±5 时，PM2.5 所计算的 AQI_NEW 有 10.78%能够与原始 AQI 正确匹配，而 PM2.5_24h 所计算的 AQI_NEW 有 59.10%能够与原始 AQI 正确匹配。这一结果再次证明使用 PM2.5_24h 所得出的 AQI_NEW 更为接近原始 AQI。

表 3 diff1 及 diff2 的累积正确频次及匹配率

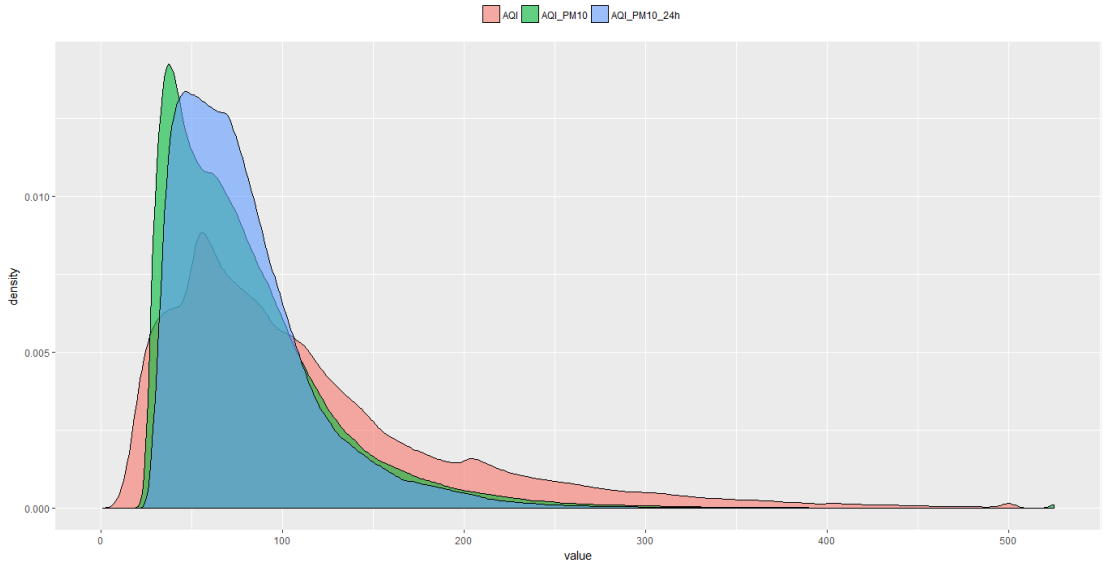
Diff 数值	Diff1 累积正确频次	Diff2 累积正确频次	Diff1 正确匹配率	Diff2 正确匹配率
0	8955	184807	0.01019961	0.2104923
±1	26197	367230	0.02983798	0.4182693
±2	43470	413862	0.04951166	0.4713824
±3	60704	453806	0.06914092	0.5168780
±4	77764	487533	0.08857200	0.5552926
±5	94650	518889	0.10780490	0.5910066

5.2 PM10 及 PM10_24h

1) 总体分布

在剔除异常值之后，将由 PM10 及 PM10_24h 所计算出的 AQI_NEW 和原始数据集中的 AQI 的概率密度曲线进行比较，如图 15 所示，PM10_AQI_NEW 或 PM10_24h_AQI_NEW 与原始 AQI 差异较大。当 0<AQI<100 时，PM10_AQI_NEW 及 PM10_24h_AQI_NEW 所对应的概率大于原始 AQI 所对应的概率；而当 100<AQI<500 时，PM10_AQI_NEW 及 PM10_24h_AQI_NEW 所对应的概率小于原始 AQI 所对应的概率。这一结果表明，PM10 及 PM10_24h 所引起的空气质量污染普遍较轻，并且 PM10 及 PM10_24h 对于严重空气污染时的贡献不大。

图 15 PM10/PM10_24h 所得 AQI_NEW 及原始 AQI 的概率密度图

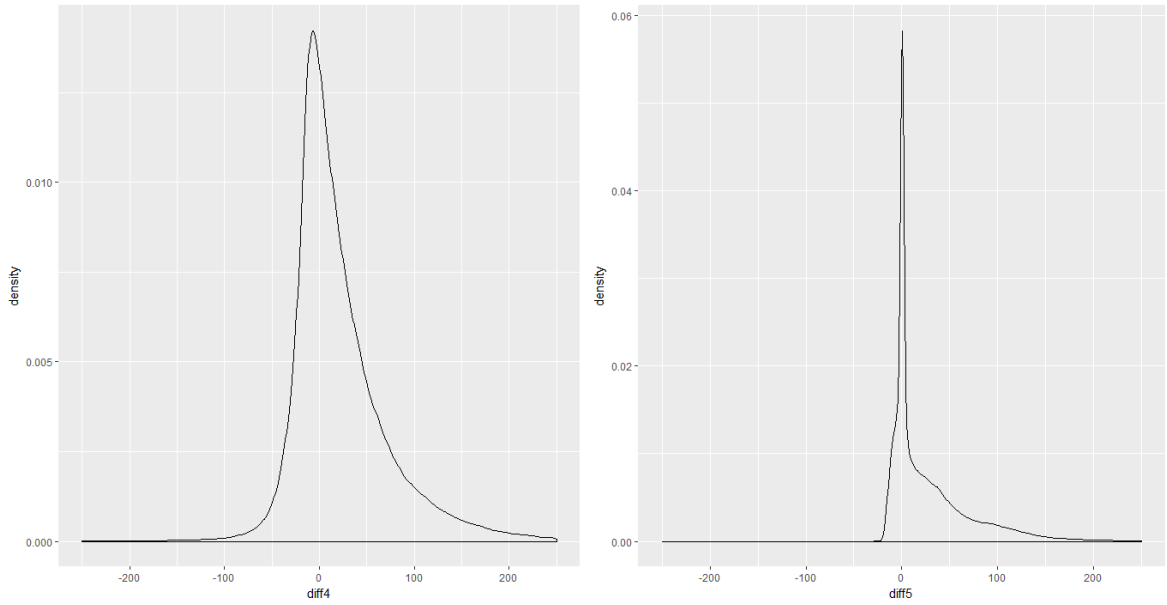


2) 差值比较

将 PM10_AQI_NEW 及 PM10_24h_AQI_NEW 与原始 AQI 分别作差，得出 diff4 及 diff5，因为样本量足够大，因此可以将 diff4 及 diff5 视为正态分布进行考察，如图 16 所示。对于 PM10_AQI_NEW 与原始 AQI 所得的 diff4 而言，平均值为 22.43，方差为 54.12；而对于 PM10_24h_AQI_NEW 与原始 AQI 所得的 diff5 而言，平均值为 26.28，方差为 41.63。

可以发现，diff4 及 diff5 的平均值均为正数，说明使用 PM10 或 PM10_24h 所计算的 AQI_NEW 总体高于原始 AQI，这一结果并不能很好地被图 15 解释。此外，diff3 与 diff4 的均值与方差均与标准正态分布相差较大，说明无论是 PM10 还是 PM10_24h 所计算的 AQI_NEW 均与原始 AQI 相去甚远。

图 16 diff4 及 diff5 的概率密度图



3) 正确匹配

当 PM10_AQI_NEW 及 PM10_24h_AQI_NEW 与原始 AQI 的差值 diff 足够小时，可以视为一次正确匹配。为寻找正确匹配所适合的误差阈限，我们将 diff 4 及 diff5 在 0，±1，±2，±3，±4，±5 时的频数进行统计，如表 4 所示。当误差阈限严格为 0 时，PM10 所计算的 AQI_NEW 只有 0.86%能够与原始 AQI 正确匹配，而 PM10_24h 所计算的 AQI_NEW 有 9.88%能够与原始 AQI 正确匹配；随着误差阈限的放松，diff4 及 diff5 的累积正确匹配率均在上升，当误差阈限放宽至±5 时，PM10 所计算的 AQI_NEW 有 9.32%能够与原始 AQI 正确匹配，而 PM10_24h 所计算的 AQI_NEW 有 26.56%能够与原始 AQI 正确匹配。这一结果证明，使用 PM10_24h 所得出的 AQI_NEW 与使用 PM10 所得出的 AQI_NEW 相比，稍微接近原始 AQI 一些。

表 4 diff4 及 diff5 的累积正确频次及匹配率

Diff 数值	Diff4 累积正确频次	Diff5 累积正确频次	Diff4 正确匹配率	Diff5 正确匹配率
0	7513	86795	0.008549012	0.09876368
±1	22561	171380	0.02567207	0.1950126
±2	37656	186999	0.04284861	0.2127854
±3	52350	202960	0.05956885	0.2309474
±4	67291	218177	0.07657015	0.2482627
±5	81933	233414	0.09323123	0.2656008

5.3 小结

通过对于 AQI 的重新计算，我们能够得出以下结论。

1) PM2.5 及 PM2.5_24h 是主要空气污染物，而 PM10 及 PM10_24h 是次要空气污染物，且它们的正确匹配率 PM2.5_24h>PM10_24h> PM2.5> PM10，其中 PM2.5_24h 正确匹配率为 59.10%，PM10_24h 正确匹配率为 26.56%；

2) 使用 PM2.5 及 PM2.5_24h 进行空气质量指数 AQI_NEW 计算所得的结果分布在两端，在空气质量良好时，PM2.5 或 PM2.5_24h 均不是影响 AQI 的主要因素，在空气质量一般时，PM2.5 或 PM2.5_24h 均是影响 AQI 的主要因素，而在空气质量恶劣时，由 PM2.5 及 PM2.5_24h 是罪魁祸首的概率大大高于官方“爆表”概率；

3) 使用 PM10 及 PM10_24h 进行空气质量指数 AQI_NEW 计算所得的结果分布更为集中，在空气质量良好时，PM10 或 PM10_24h 很可能是影响 AQI 的主要因素，而且其他情况时，PM10 或 PM10_24h 均不是影响 AQI 的主要因素。

六．AQI 的时空分布分析

该部分主要从空间和时间两个维度出发，考察 AQI 在北京地区的变化情况。

6.1 监测点间相关度

除了各种污染物，我们还收集了北京各空气质量监测点的数据，由于局限在北京的行政区划这个尺度很小，从观测的数据上来看似乎没有必要设置如此多的监测站（尤其是市区），我们先按照气象局的四种分类，看各个监测点从 2014 年的数据的相关度。

图 17 北京气象监测点地图

（其中蓝色为城区监测点、红色为交通污染监测点、绿色为郊区监测点、黄色为对照点）

（https://drive.google.com/open?id=1l6Th-r01AeeaASOr2goC6YAHP-0&usp=sharing）

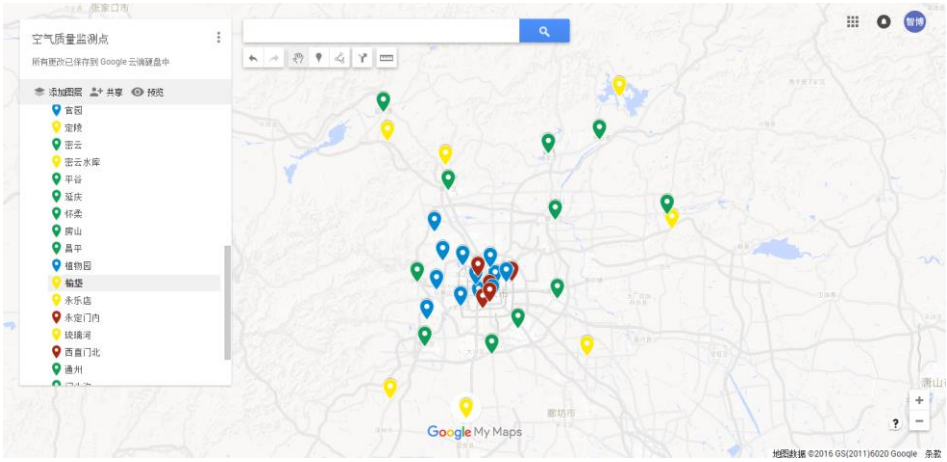


图 18 北京城区监测点 AQI 相关度

	DongSi	TianTan	GuanYuan	WanShouXiGong	AoTiZhongXin	NongZhanGuan	WanLiu	BeiBuXinQu	ZhiWuYuan	FengTaiHuaYuan	YunGang	GuCheng
DongSi	1.0000000	0.9786893	0.9781402	0.9683042	0.9683466	0.9744123	0.9677351	0.9003987	0.9312432	0.9552235	0.9458401	0.9533809
TianTan	0.9786893	1.0000000	0.9794511	0.9803542	0.9658685	0.9796642	0.9684587	0.9009202	0.9307146	0.9660325	0.9518741	0.9562909
GuanYuan	0.9781402	0.9794511	1.0000000	0.9725636	0.9711917	0.9769723	0.9792273	0.9070988	0.9475465	0.9647328	0.9580148	0.9675580
WanShouXiGong	0.9683042	0.9803542	0.9725636	1.0000000	0.9561676	0.9741412	0.9622477	0.8960183	0.9154092	0.9712836	0.9523334	0.9515307
AoTiZhongXin	0.9683466	0.9658685	0.9711917	0.9561676	1.0000000	0.9694992	0.9695244	0.9008952	0.9394268	0.9477154	0.9404357	0.9530542
NongZhanGuan	0.9744123	0.9796642	0.9769723	0.9741412	0.9694992	1.0000000	0.9690029	0.9081222	0.9337033	0.9656833	0.9526215	0.9513224
WanLiu	0.9677351	0.9684587	0.9792273	0.9622477	0.9695244	0.9690029	1.0000000	0.9247230	0.9573965	0.9654224	0.9644225	0.9723363
BeiBuXinQu	0.9003987	0.9009202	0.9070988	0.8960183	0.9008952	0.9081222	0.9247230	1.0000000	0.9107928	0.9060957	0.9188679	0.9151934
ZhiWuYuan	0.9312432	0.9307146	0.9475465	0.9154092	0.9394268	0.9337033	0.9573965	0.9107928	1.0000000	0.9160230	0.9413239	0.9476209
FengTaiHuaYuan	0.9552235	0.9660325	0.9647328	0.9712836	0.9477154	0.9656833	0.9654224	0.9060957	0.9160230	1.0000000	0.9611494	0.9544093
YunGang	0.9458401	0.9518741	0.9580148	0.9523334	0.9404357	0.9526215	0.9644225	0.9188679	0.9413239	0.9611494	1.0000000	0.9650426
GuCheng	0.9533809	0.9562909	0.9675580	0.9515307	0.9530542	0.9513224	0.9723363	0.9151934	0.9476209	0.9544093	0.9650426	1.0000000

图 19 北京交通污染监测点 AQI 相关度

	QianMen	YongDingMenNei	XiZhiMenBei	NanSanHuan	DongSiHuan
QianMen	1.0000000	0.9792821	0.9597133	0.9733418	0.9584166
YongDingMenNei	0.9792821	1.0000000	0.9706436	0.9867575	0.9705349
XiZhiMenBei	0.9597133	0.9706436	1.0000000	0.9607579	0.9584366
NanSanHuan	0.9733418	0.9867575	0.9607579	1.0000000	0.9635648
DongSiHuan	0.9584166	0.9705349	0.9584366	0.9635648	1.0000000

从整体上来看，北京城区及北京交通污染各监测点 AQI 数据的相关度非常高，因此在一定的误差范围中我们可以在很大程度上将其归为一类进行更高层级的比较。

图 20 北京郊区监测点 AQI 相关度

	FangShan	DaXing	YiZhuang	TongZhou	ShunYi	ChangPing	MenTouGou	PingGu	HuaiRou	MiYun	YanQing
FangShan	1.0000000	0.9503246	0.9379396	0.9088426	0.8736571	0.8396068	0.8764011	0.8515381	0.8418462	0.8456380	0.7989592
DaXing	0.9503246	1.0000000	0.9745094	0.9304698	0.8535201	0.8132674	0.8353160	0.8483030	0.8086379	0.8172880	0.7818342
YiZhuang	0.9379396	0.9745094	1.0000000	0.9496681	0.8904150	0.8362007	0.8529025	0.8738141	0.8448614	0.8513731	0.7986541
TongZhou	0.9088426	0.9304698	0.9496681	1.0000000	0.8874443	0.8288951	0.8385863	0.8651553	0.8387747	0.8460803	0.7906477
ShunYi	0.8736571	0.8535201	0.8904150	0.8874443	1.0000000	0.9096329	0.9067525	0.9263940	0.9602343	0.9524188	0.8649466
ChangPing	0.8396068	0.8132674	0.8362007	0.8288951	0.9096329	1.0000000	0.9303840	0.8711529	0.9237013	0.9096052	0.9017017
MenTouGou	0.8764011	0.8353160	0.8529025	0.8385863	0.9067525	0.9303840	1.0000000	0.8692884	0.9209058	0.9116114	0.8899286
PingGu	0.8515381	0.8483030	0.8738141	0.8651553	0.9263940	0.8711529	0.8692884	1.0000000	0.9131157	0.9077171	0.8299477
HuaiRou	0.8418462	0.8086379	0.8448614	0.8387747	0.9602343	0.9237013	0.9209058	0.9131157	1.0000000	0.9642224	0.8872547
MiYun	0.8456380	0.8172880	0.8513731	0.8460803	0.9524188	0.9096052	0.9116114	0.9077171	0.9642224	1.0000000	0.8823452
YanQing	0.7989592	0.7818342	0.7986541	0.7906477	0.8649466	0.9017017	0.8899286	0.8299477	0.8872547	0.8823452	1.0000000

图 21 北京对照监测点 AQI 相关度

	DingLing	BaDaLing	MiYunShuiKu	DongGaoCun	YongLeDian	YuDai	LiuLiHe
DingLing	1.0000000	0.8591953	0.9022999	0.8578584	0.7508679	0.7128201	0.6913444
BaDaLing	0.8591953	1.0000000	0.8205655	0.7160218	0.5886479	0.5753281	0.5349108
MiYunShuiKu	0.9022999	0.8205655	1.0000000	0.8411144	0.6744505	0.6409942	0.6063910
DongGaoCun	0.8578584	0.7160218	0.8411144	1.0000000	0.8461463	0.8062576	0.7367157
YongLeDian	0.7508679	0.5886479	0.6744505	0.8461463	1.0000000	0.9277473	0.8690661
YuDai	0.7128201	0.5753281	0.6409942	0.8062576	0.9277473	1.0000000	0.8902167
LiuLiHe	0.6913444	0.5349108	0.6063910	0.7367157	0.8690661	0.8902167	1.0000000

相比来说，郊区及对照监测点 AQI 相关度就不那么好，而郊区各点相关度又整体高于对照监测点的相关度，从数据及地理位置我们可以初步判断这与各监测点的相对位置有关，如果我们计算郊区监测点与其对应的对照点的相关系数可以发现相关度很高。

因此在分析前先按照地理位置分组、筛选了五组进行对比：市区监测点（13 个的平均值）；西北部监测点（延庆+八达岭）；东北部监测点（密云+密云水库）；东部监测点（东高村+平谷）；西南部监测点（琉璃河+榆垡）。其中每组内的相关系数都在 0.9 以上。

6.2 各地区分年度 AQI 比较

由于要分析各地的污染情况，由上节我们可以看到 AQI 基本可以代表空气的污染状况了，因此在数据处理时，我们取每天 AQI 的平均值代表整天的 AQI 值，取整个组内各点数据的平均值代表这一组的值，下图为整理后的数据。

图 22 整理后数据

	date	市区	西北	东北	东部	西南
1	20140101	96.36538	94.45833	72.79545	86.20833	161.70833
2	20140102	124.43910	79.72917	76.81818	87.43478	194.70833
3	20140103	135.15385	86.72917	93.73913	106.70833	176.45833
4	20140104	106.91346	92.54167	70.81250	61.84783	132.41667
5	20140105	173.20192	135.29167	144.66667	109.95833	243.10417
6	20140106	131.02885	117.37500	131.25000	135.60417	175.91667
7	20140107	186.99666	182.97917	178.72727	180.97826	215.06250
8	20140108	51.55128	41.93750	34.35417	49.41667	83.33333

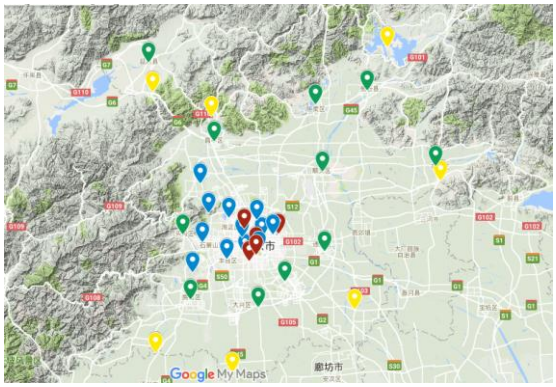
	date	type	AQI
1	20140101	市区	96.36538
2	20140102	市区	124.43910
3	20140103	市区	135.15385
4	20140104	市区	106.91346
5	20140105	市区	173.20192
6	20140106	市区	131.02885
7	20140107	市区	186.99666
8	20140108	市区	51.55128

首先粗略的分析，对各组计算这三年的平均 AQI 大概有一个整体的判断，发现西南部的平均 AQI 最高，而西北和东北部的平均 AQI 最低，市区和东部较高。从地理条件来看，北京西部、北部地形以山地为主，西北、东北及东部地区（尤其是监测点）多面环山，位置较为封闭，再加上身处远郊空气质量自然较为良好，相比而言西南监测点位置较为开阔，容易受到外部（尤其是河北地区）影响，因此即使在远郊空气质量都不如城区。为了验证此猜想，将位置空旷、同样在远郊的永乐店（位于京东南，之前不在任何一组中）单独拿出，计算其平均 AQI=130，因此验证了我们的猜想。

表 5 三年平均 AQI

type	平均 AQI
西南	133.814
市区	115.2848
东部	103.091
西北	91.11063
东北	89.39118

图 23 各监测点地形图



直观的来看，北京各组的年平均 AQI 都在逐年下降，所以从官方的平均数据来看北京的空气质量是逐年提高的，其中尤其是西南部的空气质量提升效果较为显著。此外由于缺失了 2016 年 12 月的数据，故 2016 年算得的平均值应比实际值低，但不会有较大的影响。

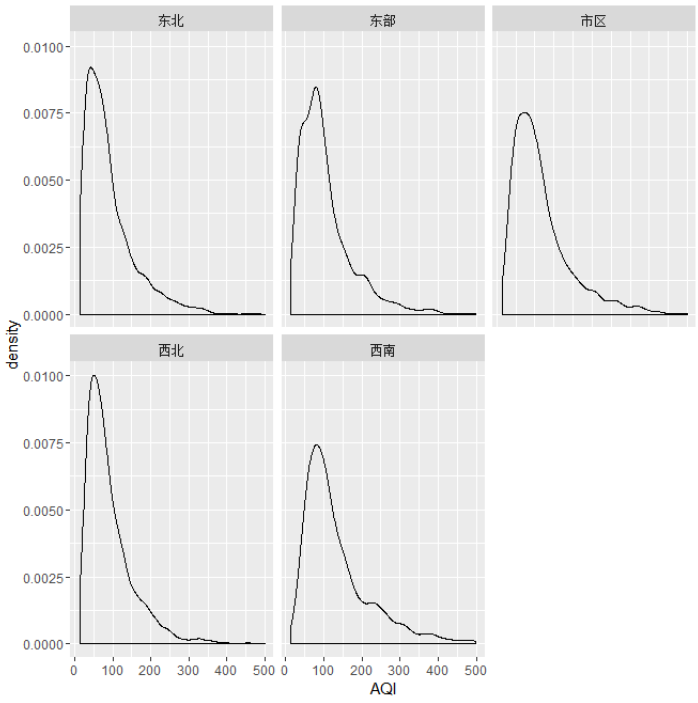
表 6（从左至右）2014、2015、2016 北京各部平均 AQI

type	平均AQI
西南	146.4219
市区	122.4254
东部	110.4177
西北	101.0625
东北	99.37767

type	平均AQI
西南	136.093
市区	120.4684
东部	107.5557
东北	89.09378
西北	87.77721

type	平均AQI
西南	117.706
市区	102.4915
东部	90.06508
西北	81.65763
东北	79.12743

图 24 北京各部 AQI 分布



6.3 严重污染的天数分析

此外，我们筛选出平均 AQI>300 的数据，各地区三年 AQI>300 的天数见下表，此外通过计算，2014、2015、2016 年 AQI>300 的天数分别为 31、36、11 天，由于 2016 年缺失了 12 月的数据，而从往年的数据中可以发现 12 月 AQI>300 的天数往往最多，因此由此不能推断出 2016 年 AQI>300 的天数比往年少。此外，AQI>200 的天数为 222 天，AQI>100 的天数为 634 天，因此尽管严重污染天数不是很多，但每年近乎有 2/3 的天数北京处在污染的空气环境中。

表 7 北京各部 AQI>300 的天数

东北	东部	市区	西北	西南
16	22	37	13	68

北京严重污染主要分布在每年的 1、2、10、11、12 几个月份，从晚秋到初春，其中又以 11、12 月为主，初步猜测与温度及供暖相关，尤其是从气象学分析，冬天内陆城市出现逆温现象较为严重，致使空气难以流通导致污染物堆积，再加上冬季机动车燃料不易完全燃烧，使得排放增加，冬季降水少，对空气中污染物的冲刷效果不明显。而且我们发现，严重污染天气往往是连续出现的，在冬季尤为严重，这也符合我们之前的分析，如果联系上天气状况其实可以发现基本上都是晴天。相比来说，烟花的影响看起来就小得多，根据 《北京市烟花爆竹安全管理规定》，在农历正月初一至十五可以燃放烟花爆竹，而三年内只有 5 天在春节期间 AQI>300，尽管从理论上来说燃放烟花爆竹对空气质量的影响不会很小，但是考虑到总量不够大以及政策管控的影响，在现实中燃放烟花爆竹不能算为主要污染来源。

而 2014 年 11 月的 APEC 蓝并没有对当年整个 11 月的空气质量造成很大贡献，反而在 APEC 结束后迎来了一波污染高潮。

此外尽管我们发现 2015 年整体空气质量较 2014 年有提升，但 2015 年 12 月的空气质量让人堪忧，超过半个月的时间空气质量达到严重污染，尤其在月末几乎没有好天气。相比而言 2016 年 2 月空气质量较前两年有所提升，而 3 月污染较为严重，初步猜测是由于 2016 年春节较前两年略晚，从季节情况和人类的作息两个角度考虑，应该只是峰值延迟了。

表 8 北京 AQI>300 的具体日期

20140116	20150104	20160102
20140117	20150105	20160103
20140124	20150109	20160110
	20150110	20160128
20140215	20150115	
20140216	20150116	20160208
20140221	20150124	
20140222		20160303
20140223	20150214	20160304
20140224	20150215	20160305
20140225	20150222	20160317
20140226		20160318
	20150307	
20140327	20150317	20161126
20141009	20151006	
20141010	20151007	
20141011	20151017	
20141019		
20141020	20151114	
20141025	20151115	
	20151128	
20141119	20151129	
20141120	20151130	
20141121		
20141122	20151201	
20141126	20151202	
20141127	20151207	
	20151208	
20141209	20151209	
20141210	20151210	
20141223	20151213	
20141227	20151214	
20141228	20151221	
20141229	20151222	
20141230	20151223	
	20151224	
	20151225	
	20151226	
	20151229	
	20151230	

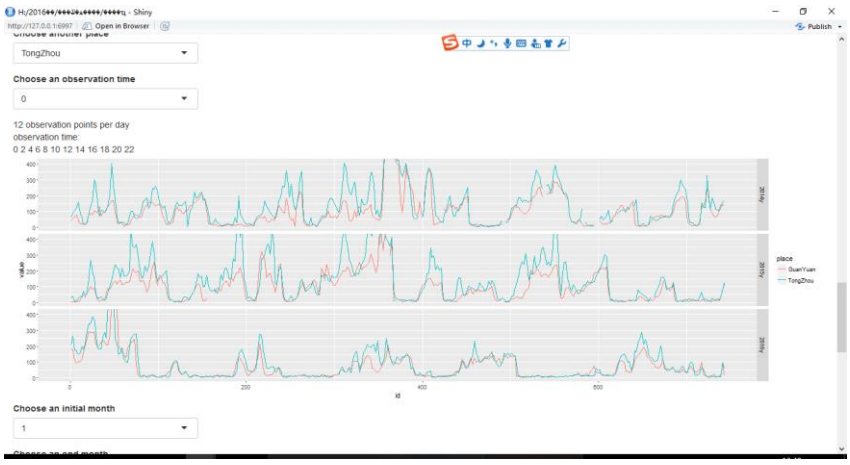
6.4 公园与交通污染点对比

我们取官园和西直门北两个监测点进行对比，其中官园是公园而西直门是交通污染点，从直观上来看，西直门北的污染应该比官园严重，但是从我们有的数据来看，二者的空气质量相差无几，尤其是在图中红点位置（2014 年七月），官园的空气质量甚至不如西直门北。我们排除计量等误差，那么为什么在有机动车行驶的要道和没有机动车行驶的地方空气质量几乎相同呢，难道说尾气排放对空气质量没有影响。实际上是有的，但是机动车直接的 PM2.5 排放量实际上是很少的，尤其是像近期较为火热的直接测量尾气的 PM2.5 值有的竟然低于空气中的 PM2.5 含量，而机动车对 PM2.5 的贡献主要来自于氮氧化物和碳氢化合物在大气中的二次反应，而这种反应的尺度很大，因此不会只局限在路边，所以看起来公园的空气质量并没有多好。

图 25 官园 vs 西直门北

此外我们还将官园和通州两地的监测点进行对比，分别代表着城区和近郊。在之前的分析中发现，地势较为平坦开阔的远郊的空气质量差，通州地势平坦开阔、距城区不远不近，因此拿来对比。从图中可以很明显的看到，通州的 AQI 普遍高于官园，因此以我们的标准来看，通州作为近郊空气质量不如城区的官园。

图 26 官园 vs 通州



6.5 小结

1) 北京的空气质量在逐年（缓慢）变好，但空气质量仍不乐观。

2) 远郊、近郊的空气质量不一定比城区好，还要考虑地形、位置等因素（实际还应考虑各种类型的车的排放及车辆分布，比如远郊会有重型卡车其排放量远远高于各类私家车，此文没有提及）。

3) 城区内公园的空气质量不一定比交通污染点的好，甚至不如。

4) 目前来看，北京的空气污染集中在冬季且极为严重，尤其是 11、12 月份，此外按照规律来看，2017 年从 3 月开始空气质量将有所好转，严重污染主要分布在 2 月，全年的 2/3 会有至少轻度污染。

5) 用 AQI 的弊端：如前所述，AQI 的最终值只反映了主要污染物的情况，但是无法体现所有污染物的情况，一方面不同污染物对人体的危害情况和预防方式都不同，会使得空气质量指数变得模糊，另一方面由于只体现了首要污染物的 IAQI，因此在 AQI 相同的情况，也会有不同程度的污染状况，比如在 AQI 相同的情况下，PM2.5 和 CO 同时为首要污染物的污染程度肯定高于只有 PM2.5 为首要污染物的情况。但是由于各种观测的污染物之间有一定的相关关系，所以用 AQI 还是可以粗略的代表空气质量情况的。

七. 结语

以上为我们的报告，若其中存在不科学、不准确的地方恳请指正，最后感谢俞声老师一学期的指导。