数据科学导论期末大作业

一．数据来源及数据背景

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

	监测点	监测点全称
城市环境评价点	东四	东城东四
	天坛	东城天坛
	官园	西城官园
	万寿西宫	西城万寿西宫
	奥体中心	朝阳奥体中心
	农展馆	朝阳农展馆
	万柳	海淀万柳
	北部新区	海淀北部新区
	植物园	海淀北京植物园
	丰台花园	丰台花园
	云岗	丰台云岗
	古城	石景山古城
郊区环境评价点	房山	房山良乡
	大兴	大兴黄村镇

	亦庄	亦庄开发区
	通州	通州新城
	顺义	顺义新城
	昌平	昌平镇
	门头沟	门头沟龙泉镇
	平谷	平谷镇
	怀柔	怀柔镇
	密云	密云镇
	延庆	延庆镇
对照点及区域点	定陵	昌平定陵
	八达岭	京西北八达岭，京西北区域点
	密云水库	京东北密云水库，京东北区域点
	东高村	京东东高村，京东区域点
	永乐店	京东南永乐店，京东南区域点
	榆垡	京南榆垡，京南区域点
	琉璃河	京西南琉璃河，京西南区域点
	定陵	昌平定陵
交通污染监测点	前门	前门东大街，前门交通点
	永定门内	永定门内大街，永定门交通点
	西直门北	西直门北大街，西直门交通点
	南三环	南三环西路，南三环交通点
	东四环	东四环北路，东四环交通点

二．AQI 重新计算

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 1 中与C_P相近的污染物浓度限值的高位值；

BP_{Lo}代表图 1 中与C_P相近的污染物浓度限值的低位值；

IAQI_{Hi}代表图 1 中与BP_{Hi}对应的空气质量分指数；

IAQI_{Lo}代表图 1 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 1 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

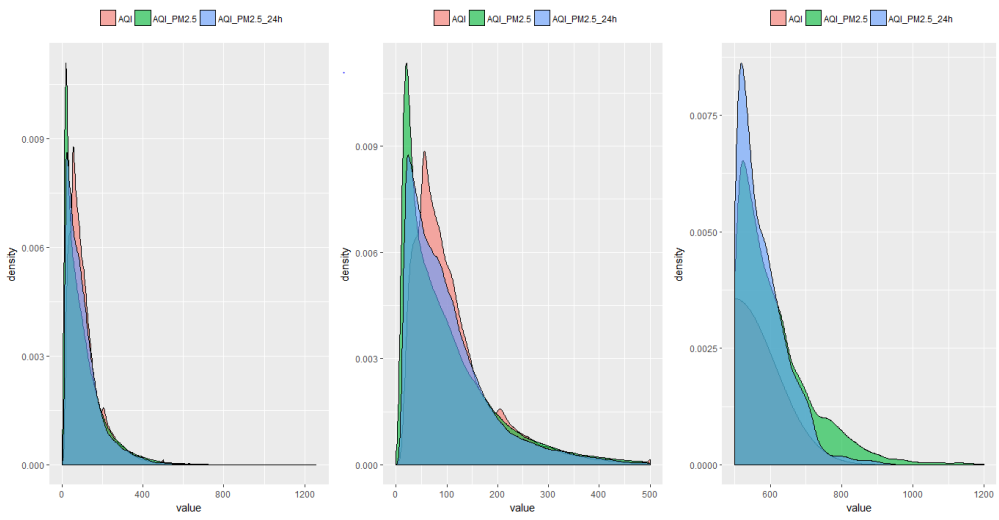
1. PM2.5 及 PM2.5_24h

1.1 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 2(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 2(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 2(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 2 PM2.5/PM2.5_24h 所得 AQI_NEW 及原始 AQI 的概率密度图

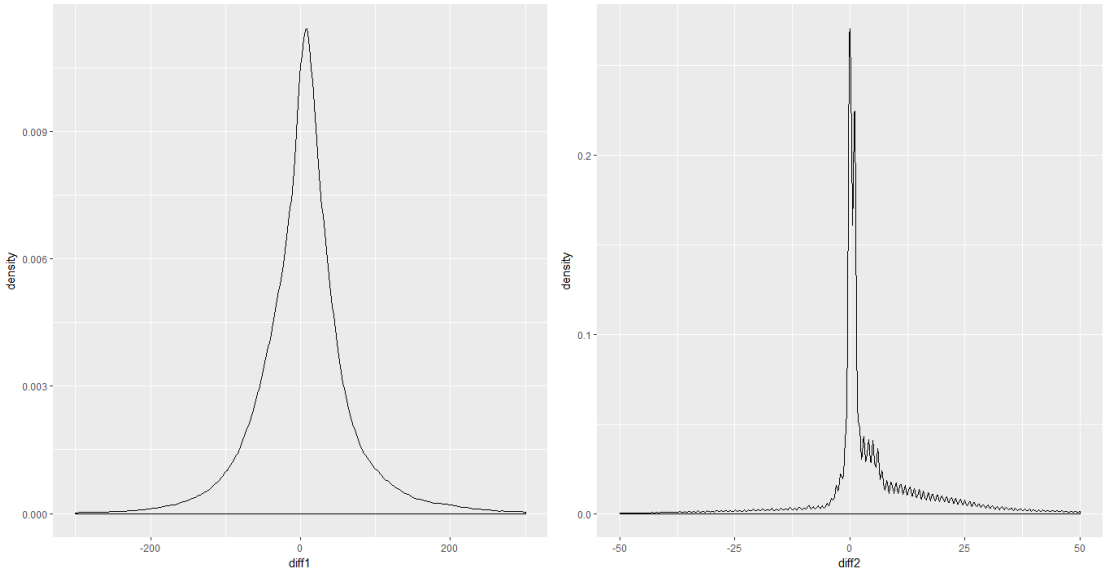


1.2 差值比较

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 3 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 3 diff1 及 diff2 的概率密度图



1.3 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 累积正确匹配率	Diff2 累积正确匹配率
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

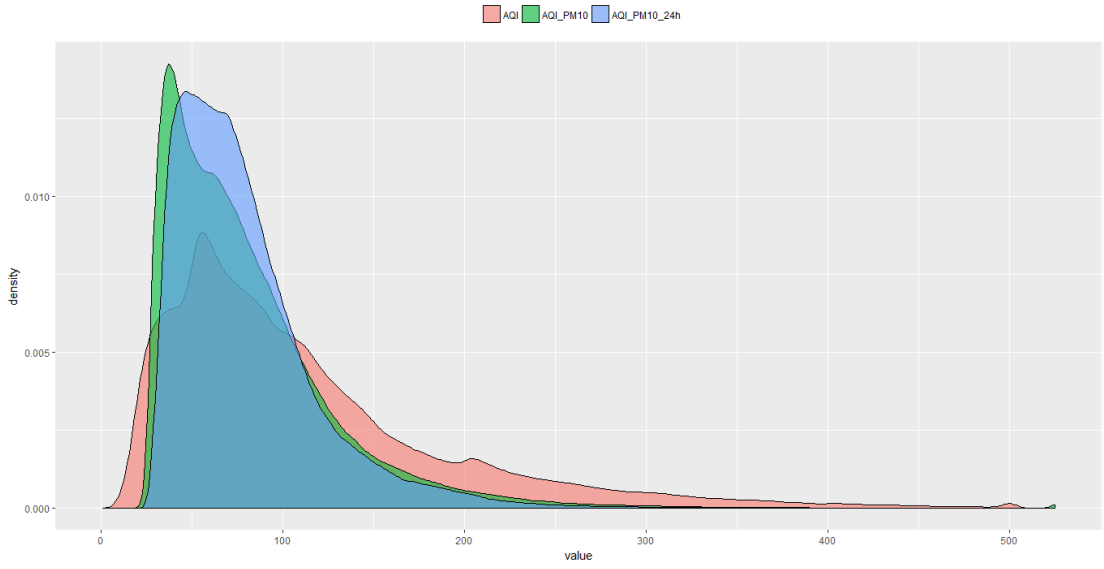
±4	77764	487533	0.08857200	0.5552926
±5	94650	518889	0.10780490	0.5910066

2. PM10 及 PM10_24h

2.1 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 4 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 4 PM10/PM10_24h 所得 AQI_NEW 及原始 AQI 的概率密度图

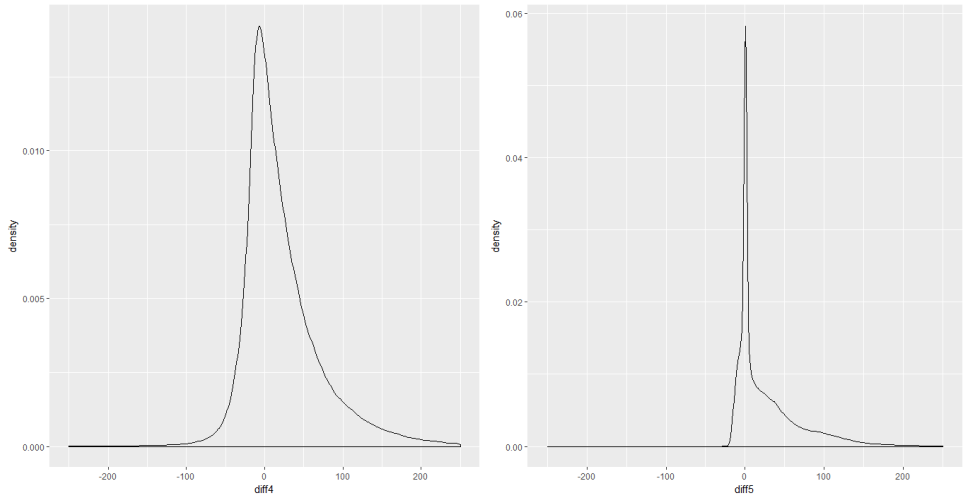


2.2 差值比较

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 5 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 4 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 5 diff4 及 diff5 的概率密度图



2.3 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 累积正确匹配率	Diff5 累积正确匹配率
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

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。