北京空气质量分析报告

WISE 朱佳

空气质量与人体健康息息相关，近年来我国某些大城市空气污染问题严重，对人们的健康造成较大伤害。且雾霾在京津冀、长三角等地区频发，使空气质量问题受到公众的广泛关注。PM2.5、PM10 等空气污染物浓度指标，也成为社会重视的话题。

环境空气质量标准的建立，在环境质量管理、人体健康保护、生态环境安全维护方面发挥着积极作用。2012 年 2 月 29 日国家环保部发布了空气质量新标准——《环境空气质量标准》（GB3095－2012）、《环境空气质量指数(AQI)技术规定(试行)》（HJ 633-2012），空气质量指数（Air quality Index，AQI）替代原有的空气污染指数（Air Pollution Index，API），用以衡量环境空气质量。

目前全国各大城市的 AQI 监测体系已经基本建立，某些大城市各个观测站的数据也可每小时实时更新，大量关于 AQI 的历史数据积累下来。然而各个城市只是对实时 AQI 及日 AQI 进行发布，却没有对未来 AQI 的预报机制。如何利用 AQI 及 PM2.5、PM10 等各项污染物浓度数据对城市未来的 AQI 进行有效预测，对人们的生活、出行提供建议，是个很有意义的研究问题。

由于工业、地理和环境的多重影响，北京市的空气质量较差，而北京空气质量状况受到了广泛关注，尤其是居住、工作和学习在北京的人们。故本文将侧重点放在北京市的空气质量研究上，力图描述各项污染物浓度的数据特征，探索 AQI 的预测问题，为身处于北京的人们提供出行和户外运动的建议。

1. 数据说明

1.1 数据来源

由于中国环境监测总站只是实时地更新数据，并没有提供历史数据和下载历史数据的链接，故在PM2.5 监测网上使用 R 爬虫抓取数据。本文使用的数据为 2013 年 12 月 2 日至 2016 年 4 月 30 日，北京市每天 6 项污染物即细颗粒物（$PM_{2.5}$）、可吸入颗粒物（PM_{10}）、二氧化硫（SO_2）、二氧化氮（NO_2）、臭氧（O_3）、一氧化碳（CO）24 小时平均浓度值和当日 AQI，以及当日平均风级和主要风向数据，其中风级和风向数据存在缺失值。

表 1：数据类型、单位及说明

变量	单位	类型	说明
AQI	无	数值型	无量纲指数，用来衡量空气质量好坏，数值越大空气质量越差
PM2.5	μg/m³	数值型	
PM10	μg/m³	数值型	
SO2	μg/m³	数值型	
CO	mg/m³	数值型	
NO2	μg/m³	数值型	
O3	μg/m³	数值型	
风级	无	数值型	0，1，2，3
风向	无	字符型	东，南，西，北，东南，西南，东北，西北

1.2 相关定义

第一步：对照空气质量分指数（IAQI）及对应污染物浓度限值（表 2），以细颗粒物（PM2.5）、可吸入颗粒物（PM10）、二氧化硫（SO2）、二氧化氮（NO2）、臭氧（O3）、一氧化碳（CO）等各项污染物的实测浓度值（其中 PM2.5、PM10 为 24 小时平均浓度）分别计算得出空气质量分指数（Individual Air Quality Index，IAQI）。

$$IAQI_p = \frac{IAQI_H - IAQI_L}{BP_H - BP_L}\left(C_P - BP_L\right) + IAQI_L.$$

其中，

$IAQI_P$——污染物项目P的空气质量分指数；

C_P——污染物项目P的质量浓度值；

BP_H——表1中与C_P相近的污染物浓度限值的高位值；

BP_L——表1中与C_P相近的污染物浓度限值的低位值；

$IAQI_H$——表1中与BP_H对应的空气质量分指数；

$IAQI_L$——表1中与BP_L对应的空气质量分指数。

表2：空气质量分指数及对应污染物浓度限值

空气质量分指数	SO2-24h平均	SO2-1h平均	NO2-24h平均	NO2-1h平均	PM10-24h平均	CO-24h平均	CO-1h平均	O3-1h平均	O3-8h平均	PM2.5-24h平均
0	0	0	0	0	0	0	0	0	0	0
50	50	150	40	100	50	2000	5000	160	100	35
100	150	500	80	200	150	4000	10000	200	160	75
150	475	650	180	700	250	14000	35000	300	215	115
200	800	800	280	1200	350	24000	60000	400	265	150
300	1600		565	2340	420	36000	90000	800	800	250
400	2100		750	3090	500	48000	120000	1000		350
500	2620		940	3840	600	60000	150000	1200		500

第二步：从各项污染物的 IAQI 中选择最大值确定为 AQI，当 AQI 大于50 时，将IAQI 最大的污染物确定为首要污染物。

$$AQI = \max\{IAQI_1, IAQI_2, IAQI_3, \cdots, IAQI_n,\}.$$

其中，

IAQI ——空气质量分指数；

n ——污染物项目，此处 n=6。

1.3 AQI 分级

空气质量按照空气质量指数大小分为六级，相对应空气质量的六个类别，指数越大、级别越高说明污染的情况越严重，对人体的健康危害也就越大。AQI等级类别、对健康的影响及相应建议见表3。

表3 : AQI等级类别与相应建议

AQI	等级	类别	对健康的影响	措施建议
0~50	一级	优	空气质量令人满意，基本无空气污染	各类人群可正常活动
51~100	二级	良	空气质量可接受，但某些污染物可能对极少数异常敏感人群健康有较弱影响	极少数异常敏感人群应减少户外活动
101~150	三级	轻度污染	易感人群症状有轻度加剧，健康人群出现刺激症状	儿童、老年人及心脏病、呼吸系统疾病患者应减少长时间、高强度的户外锻炼
151~200	四级	中度污染	进一步加剧易感人群症状，可能对健康人群心脏、呼吸系统有影响	儿童、老年人及心脏病、呼吸系统疾病患者应减少长时间、高强度的户外锻炼，一般人群适量减少户外运动
201~300	五级	重度污染	心脏病和肺病患者症状显著加剧，运动耐受力降低，健康人群普遍出现症状	儿童、老年人及心脏病、肺病患者应停留在室内，一般人群减少户外运动
>300	六级	严重污染	健康人群运动耐受力降低，有明显强烈症状，提前出现某些疾病	儿童、老年人病人应停留在室内，避免体力消耗，一般人群避免户外运动

2. 描述性统计

图 1 为厦门、北京的 AQI 分布散点图，可以毫无悬念地看出从 2013 年 12 月始至 2016 年 4 月止，北京的 AQI 明显大于厦门 AQI 的值。而且，厦门的 AQI 几乎完全分布于 0 到 100 之间，可见厦门的空气质量几乎全为优或良，其 AQI 表现很均衡，AQI 没有较大的波动；而北京的 AQI 变化幅度很大，从 0 到 500 之间都有分布，其 AQI 大于 100 的天数不在少数，即北京的空气质量为轻度污染、中度污染、重度污染、严重污染甚至“爆表”的天数比例很高。而下文则聚焦在北京空气质量问题上。

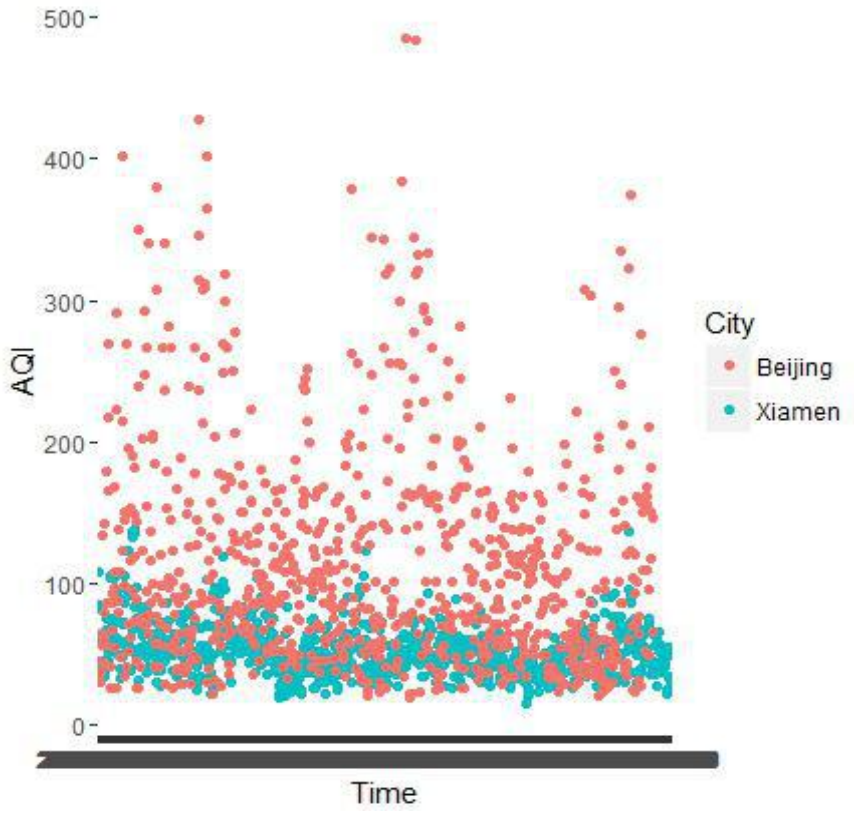


图 1：厦门市、北京市 AQI 散点图

2.1 北京 PM2.5、PM10 等 6 项污染物及 AQI 的数据特征

对于北京 PM2.5 总体分布情况，北京 PM2.5 浓度绝大部分分布于 0 至 200 μg/m³，小部分分布于 200 μg/m³至 400 μg/m³之间，在 2015 年 11 月至 12 月中，出现两个极大值，其值超过了 475 μg/m³，见图 2。

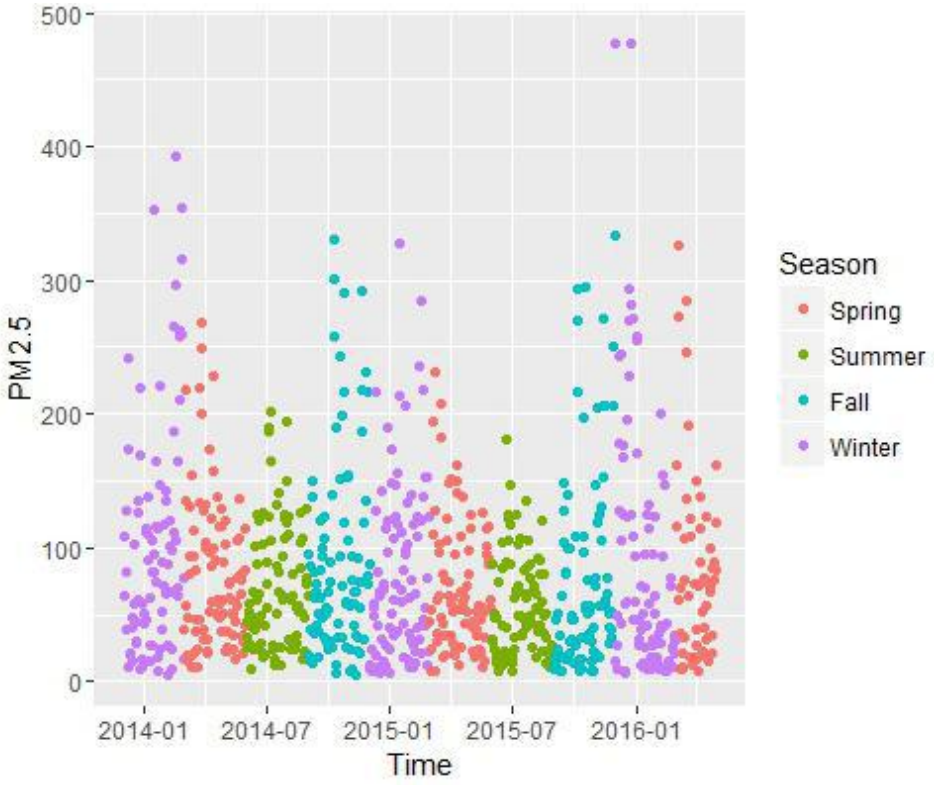


图 2：北京市 PM2.5 四季分布散点图

对于北京 PM2.5 的四季分布情况，由图 2 可看出夏季北京 PM2.5 浓度几乎全部低于 200 μg/m³，PM2.5 浓度高于 200 μg/m³ 的情况几乎全部分布在春季、夏季和冬季。夏季北京 PM2.5 浓度的平均值低于春季、秋季和冬季北京 PM2.5 浓度的平均值；其中，冬季 PM2.5 浓度平均值最高，春季次之，秋季紧随其后，见图 3。

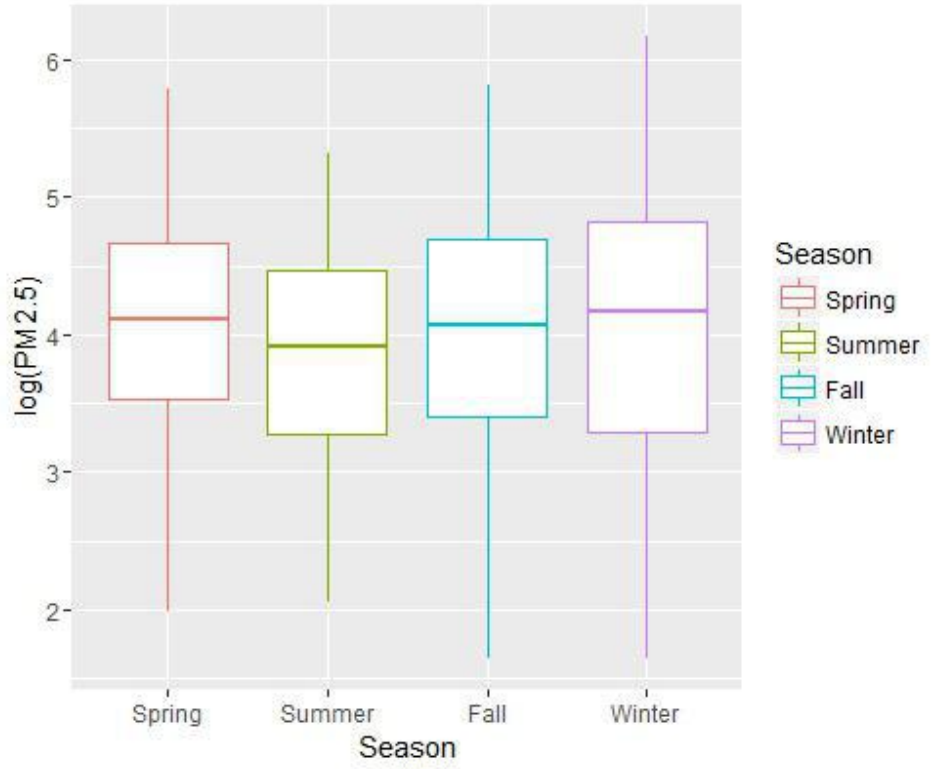


图 3：北京市 PM2.5 四季分布箱线图

对于北京 PM10 的总体分布情况，其分布与 PM2.5 的分布极为相似，绝大部分分布于 0 至 200 μg/m³，小部分分布于 200 μg/m³至 400 μg/m³之间，在 2014 年 2 月前后、2015 年 11 月和 2015 年 12 月前后，分别出现极大值，其值都等于甚至超过了 450 μg/m³，见图 4。对于北京 PM10 总体的分布情况。

对于北京 PM10 的四季分布情况，春季 PM2.5 浓度平均值最高，冬季次之，夏季的 PM2.5 浓度平均值最低，见图 5。

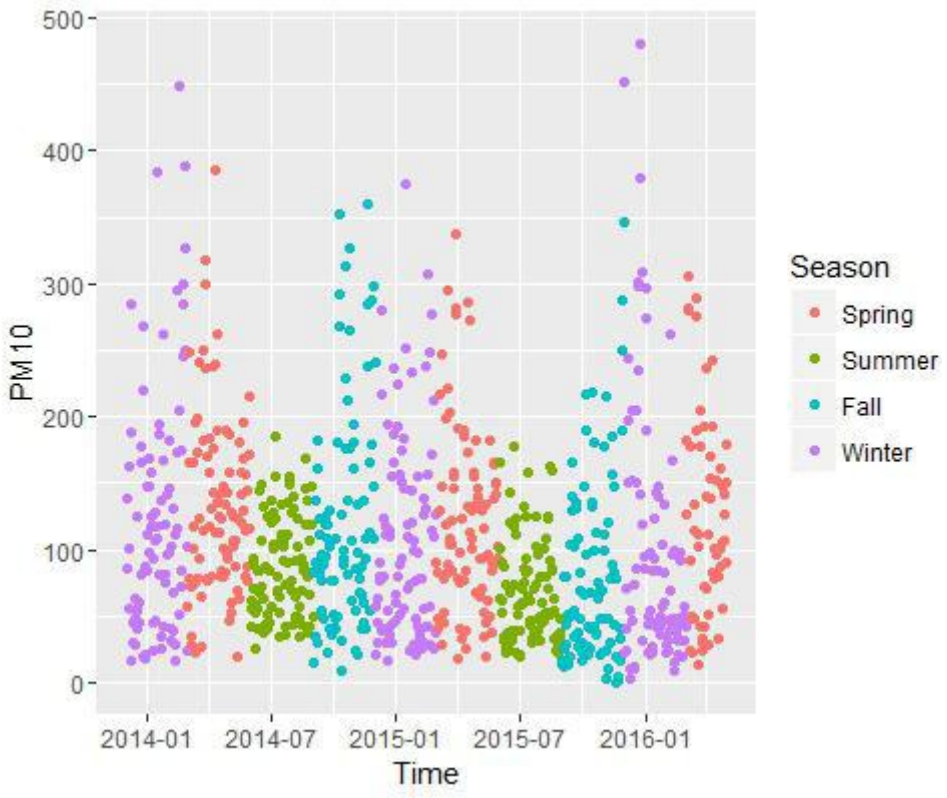


图 4：北京市 PM10 四季分布散点图

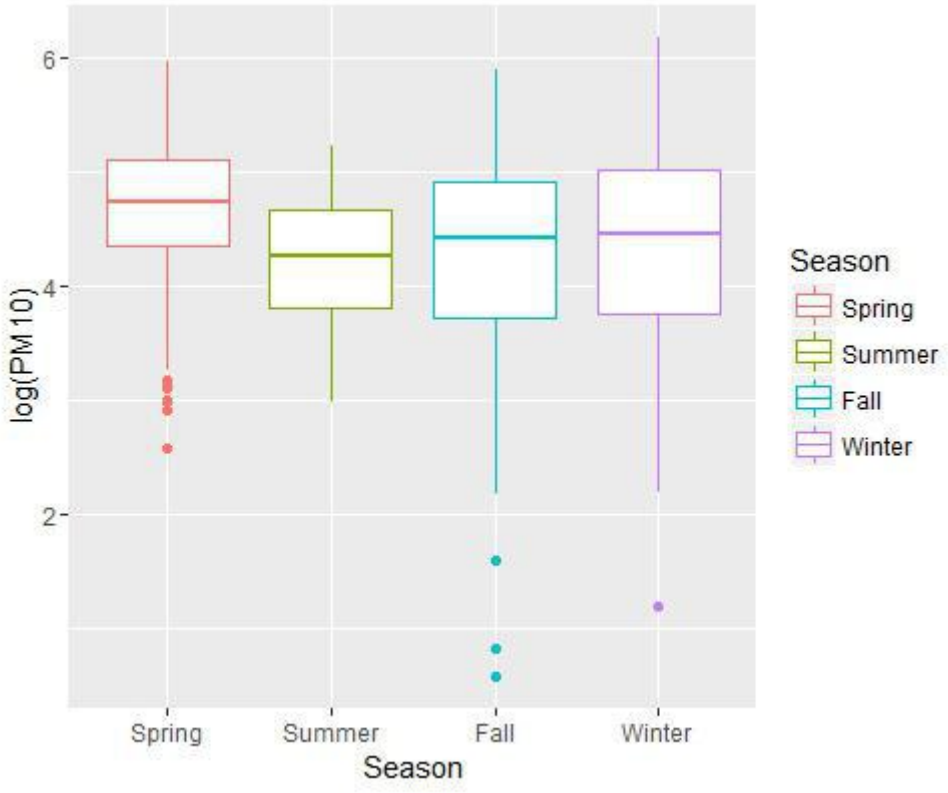


图 5：北京市 PM10 四季分布箱线图

同样的，北京 AQI 整体分布和四季分布，与 PM2.5 和 PM10 的分布十分相似，此处不再赘述。见图 6，图 7。

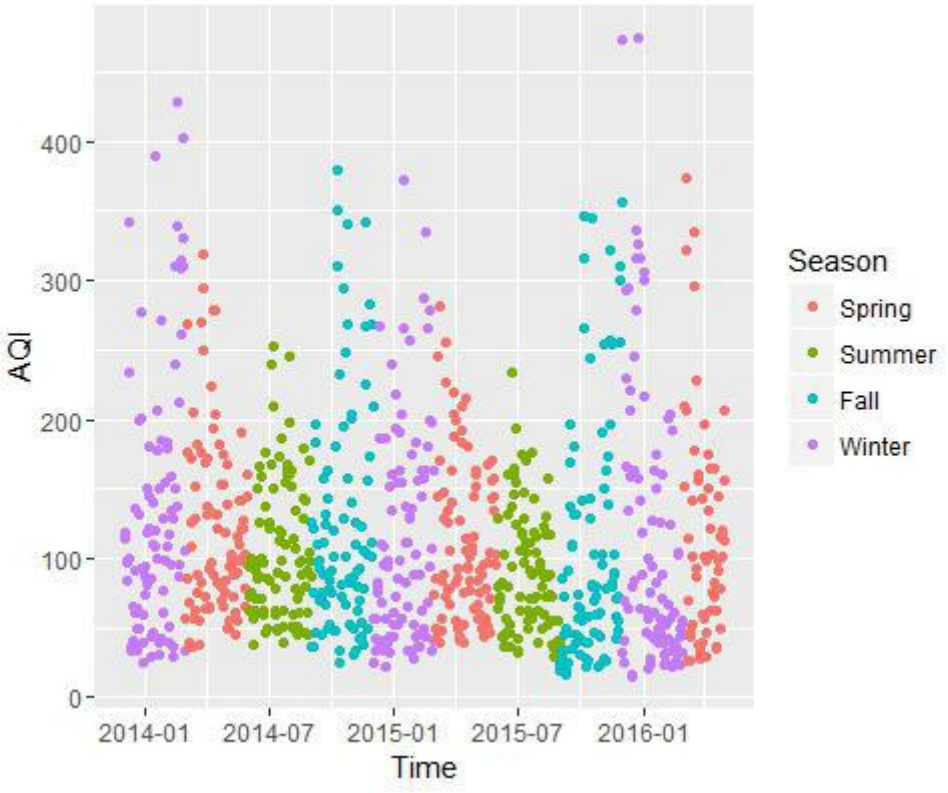


图 6：北京市 AQI 四季分布散点图

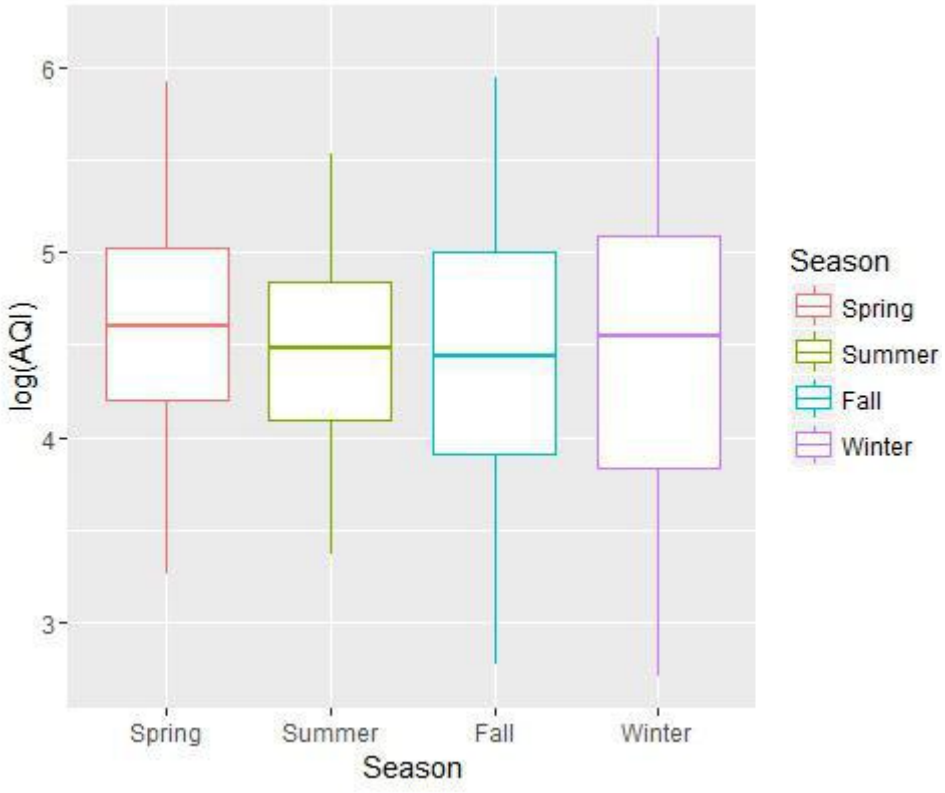


图 7：北京市 AQI 四季分布箱线图

以下分别为 SO2、CO、NO2 和 O3 四季分布箱线图，见图 8，图 9，图 10 和图 11。值得一提的是 O3 的季节分布，夏季 O3 浓度平均水平最高，而臭氧的浓度升高是光化学烟雾污染的标志。所谓光化学污染，即在高温、低湿、低风速气象条件下，大气中的挥发性有机物和氮氧化物等一次污染物在阳光（紫外光）的作用下发生光化学反应，生成高浓度臭氧及过氧乙酰硝酸酯、醛、酮、酸、细粒子气溶胶等二次污染物，形成一次污染物和二次污染物共存的污染现象。臭氧则是光化学污染的一种重要的污染物。光化学污染一般在夏季午后高发，因此，应减少在该时段外出。

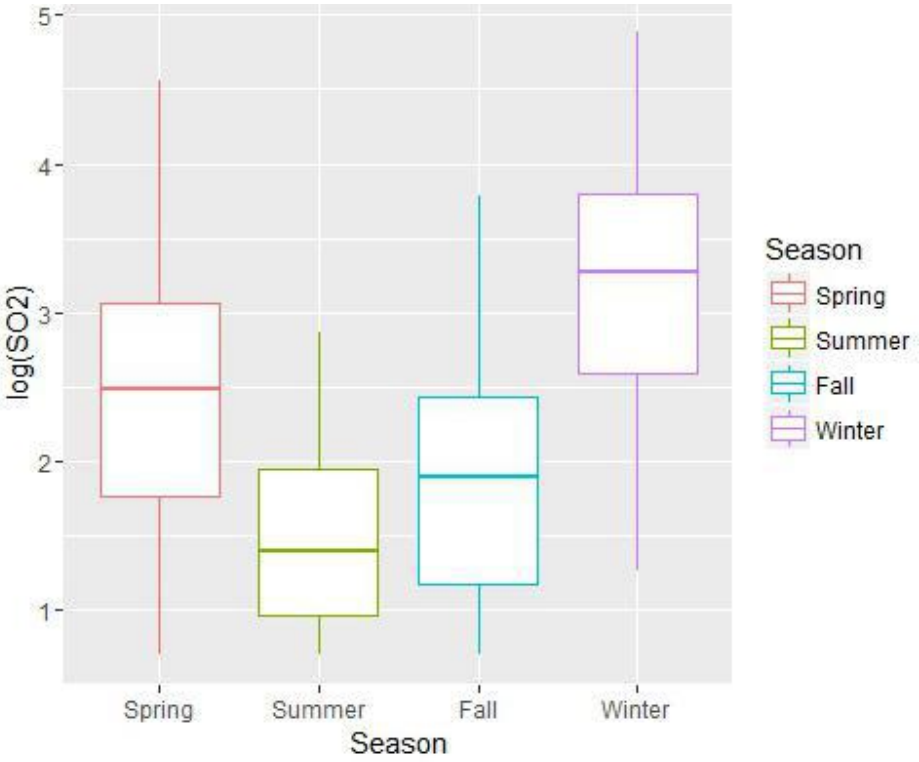


图 8：北京市 SO2 四季分布箱线图

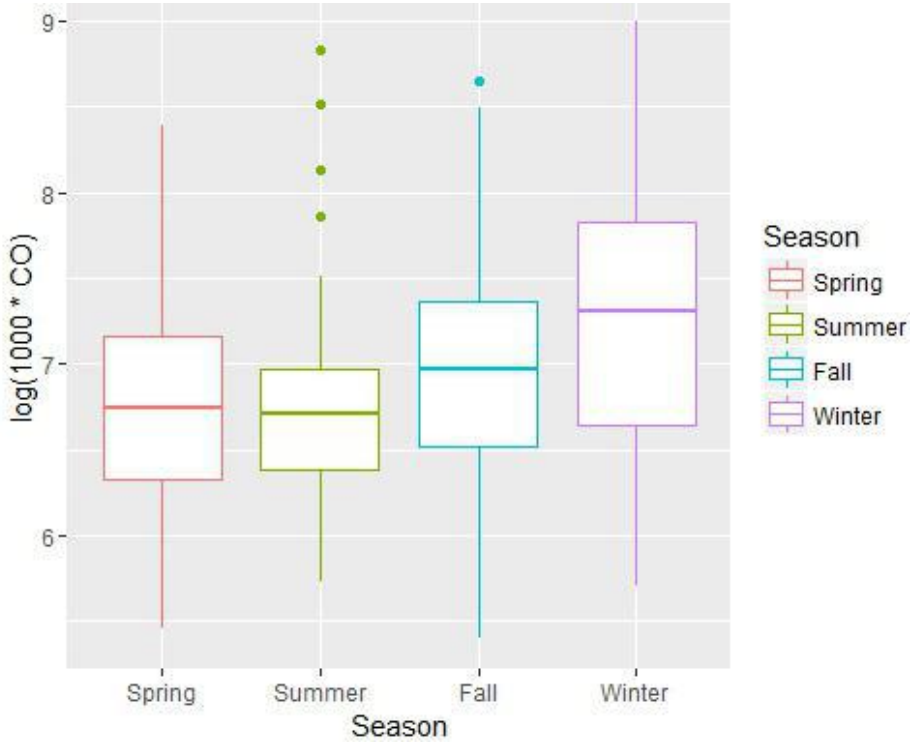


图 9：北京市 CO 四季分布箱线图

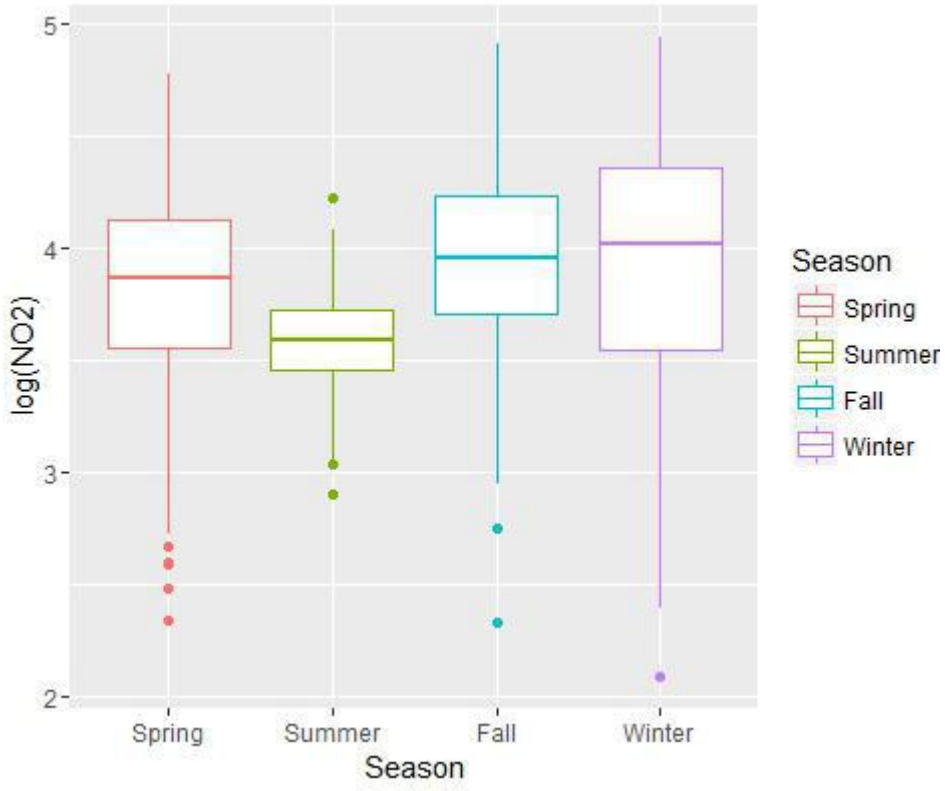


图 10：北京市 NO2 四季分布箱线图

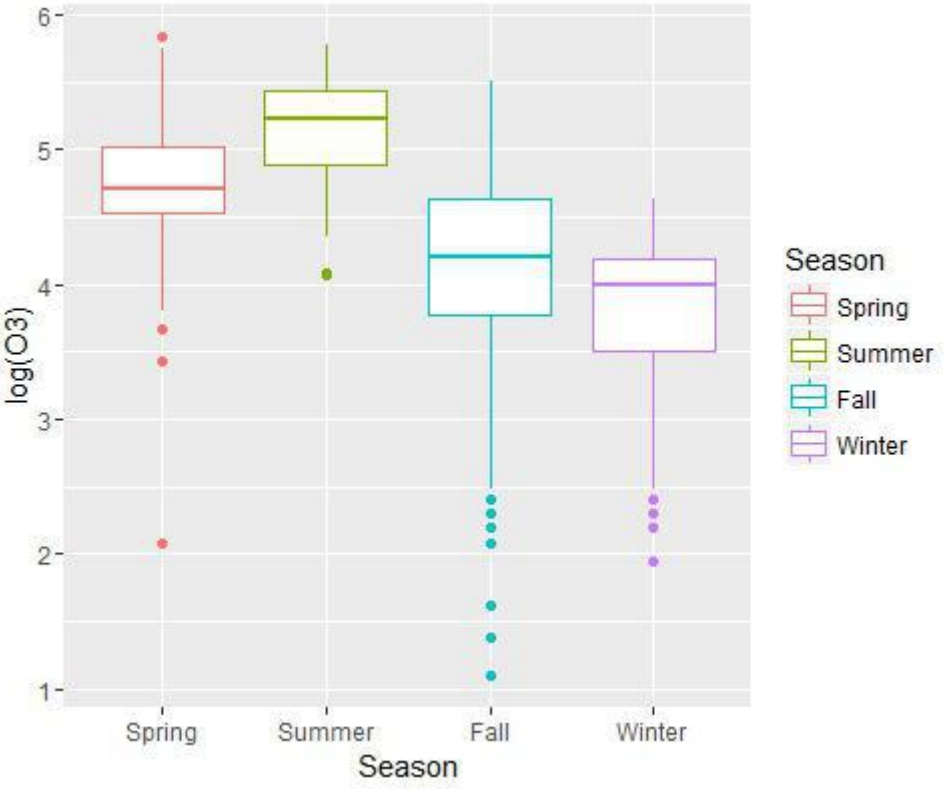


图 11：北京市 O3 四季分布箱线图

2.2 北京空气质量等级逐年及季节分布

对于北京空气质量等级的总体分布情况，等级为优良的天数占比分别为 19.2% 和 34.3% ，占据样本数据总天数的 53.5% 。而空气质量等级为轻度、中度、重度甚至严重污染的天数占比总和为 46.5% ，其中轻度污染和中度污染天数占比之和为 34.3% ，与等级为良的占比相同，重度污染和严重污染天数占比总和则为 12.2% 。可见，在 2013 年 12 月 2 日至 2016 年 4 月 30 之间，生活在“帝都”人们有大约一半时间在被污染的空气中度过，这些空气被污染的天数中，空气质量很差即为严重污染和重度污染的天数占比高达 35.8% ！见表 4，图 12。

表 4：北京空气质量等级频数及频率分布表

	优	良	轻度污染	中度污染	重度污染	严重污染
频数	169	302	193	109	75	33
频率	0.192	0.343	0.219	0.124	0.085	0.037

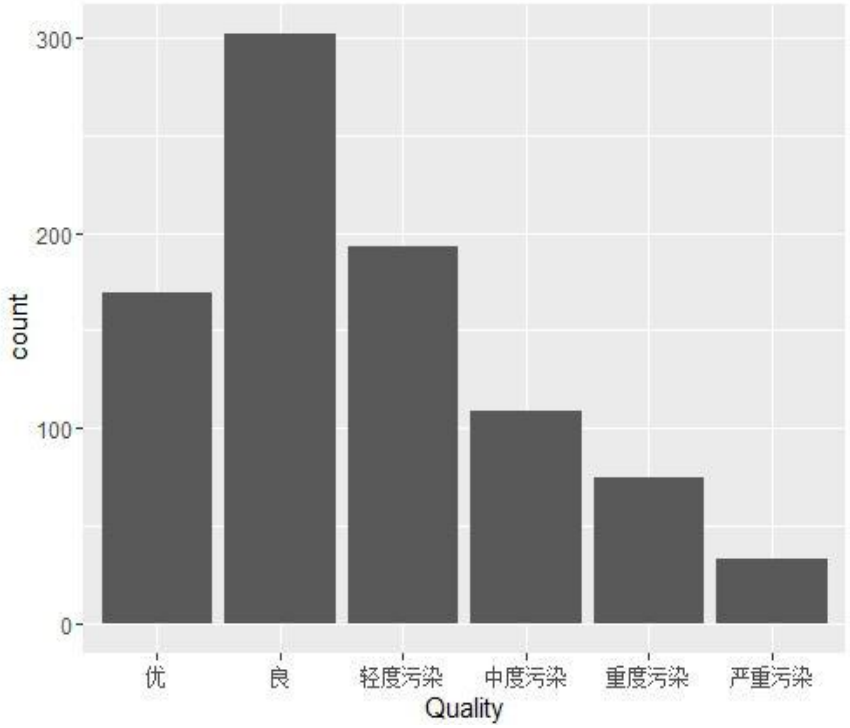


图 12：北京空气质量等级频数分布直方图

对于北京空气质量等级的季节分布情况，北京冬季中度污染、重度污染和严重污染天数占比 29.9%，在四个季节中占比最高，春季次之，占比为 25.3% ，夏季最低，占比 15.8%。究其原因在于，每年 11 月份至次年 3 月份，北京城市开始进入供暖期，不少研究表明煤的燃烧是 PM2.5 的一项重要来源，而目前整个北方地区的供暖主要依靠燃煤，而且不少地方使用劣质煤。见图 13。

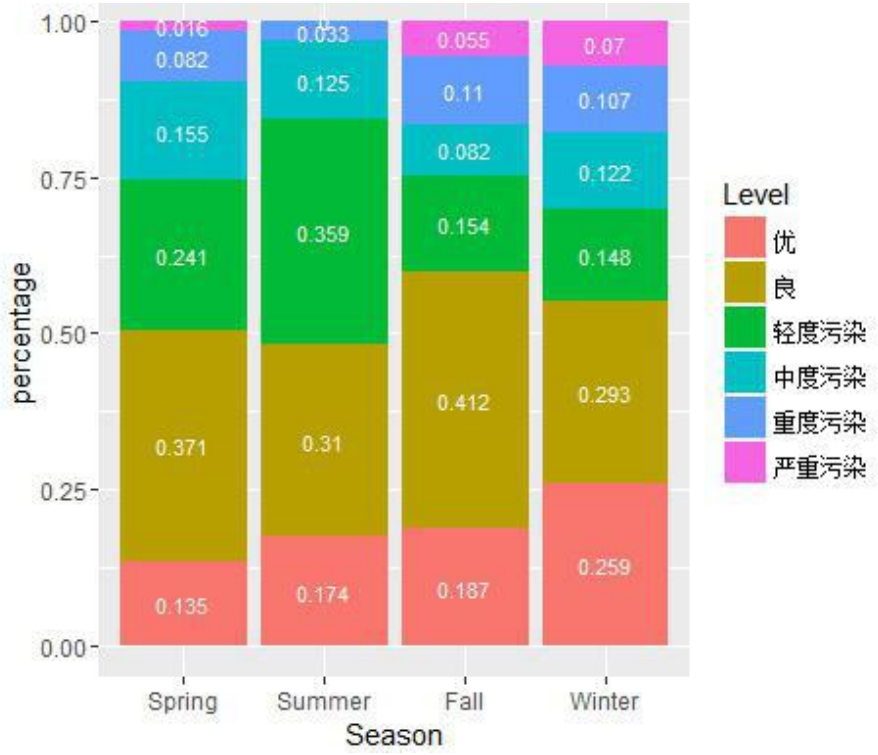


图 13：北京四季空气质量等级频率分布直方图

对于北京 2014 年至 2016 年空气质量等级分布情况，北京的优良空气质量等级占比有逐年提升的趋势，从 2014 年的 50.4%，上升至 2015 年的 53.3%，到 2016 年的 59.5%，分别提高 2.9%，9.4%。而三年间中度、重度和严重污染占比持平，为 24.7%，污染较为严重的天数没有降低，污染物排放并没有得到有效的减少和治理。见图 14。

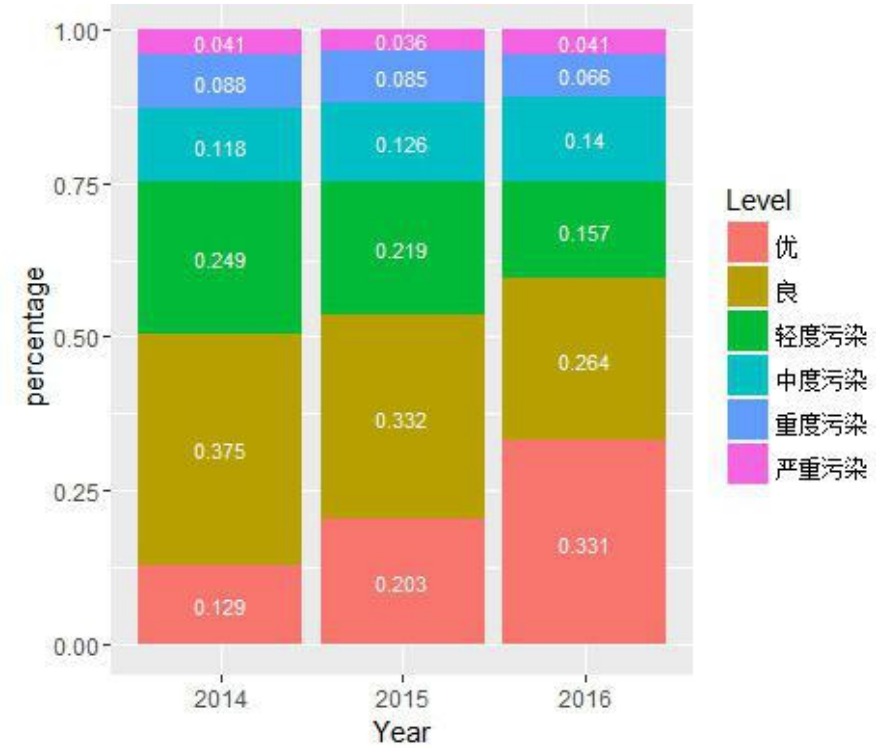
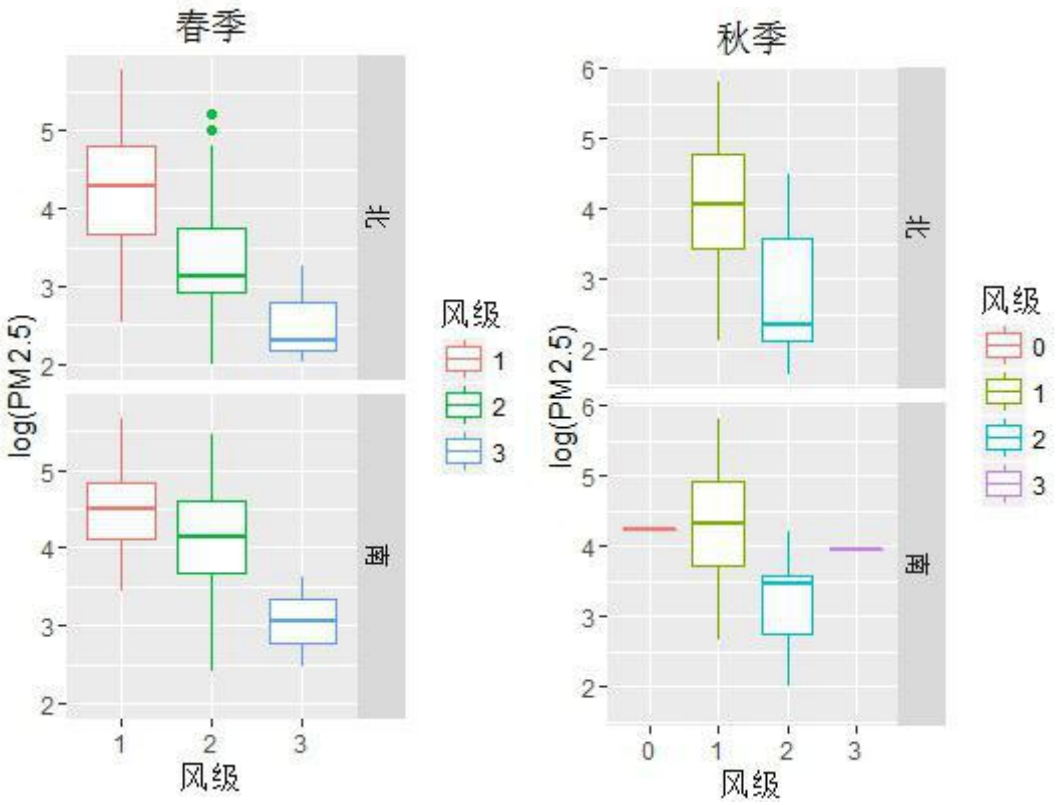


图 14：北京 2014 年至 2016 年空气质量等级频率分布直方图

2.3 风力、风级对北京空气污染质量等级的影响

北京地势西北高东南低，西部、北部、东北部三面环山，东南部是一片缓缓向渤海倾斜的平原。因此，风向、风级的不同，会对不同季节的 PM2.5、PM10 浓度以及 AQI 造成影响。考虑到北京地形因素和风向指标，故将西风、西北风、东北风和北风归为“北风”类，将东风、东南风、西南风和南风归为“南风”类。PM2.5 浓度受风向和风级影响，在相同的季节中，风级越大，PM2.5 浓度相对越低；当季节和风级相同时，“北风”即西风、西北风、东北风和北风，相比于“南风”即东风、东南风、西南风和南风，更有利于北京 PM2.5 的扩散，见图 15。而风向和风级对 AQI 的影响与 PM2.5 的结果类似，见图 16。



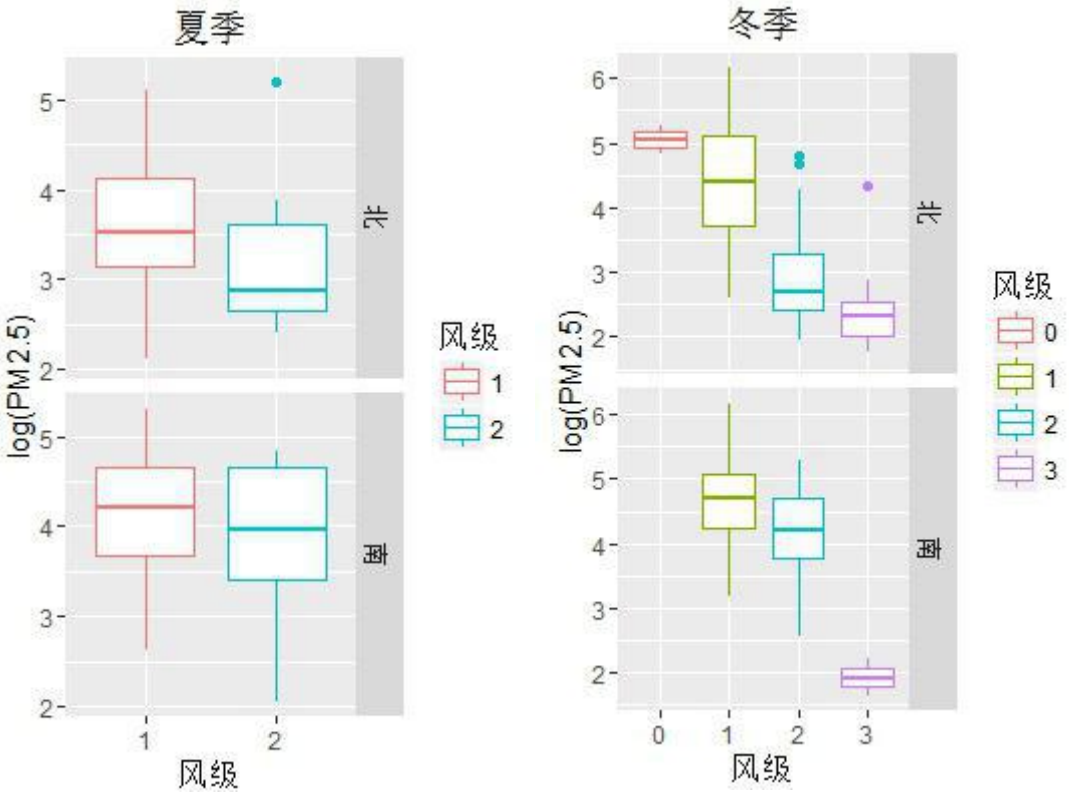
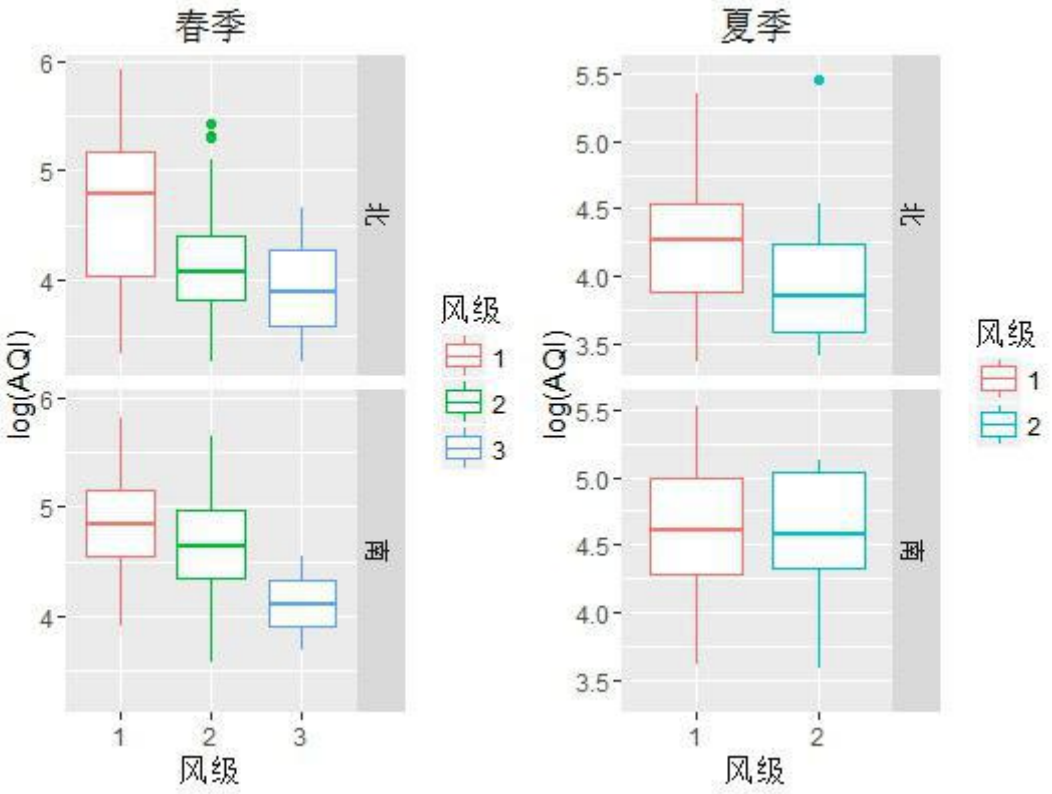


图 15：风向、风级对北京四季 PM2.5 的影响



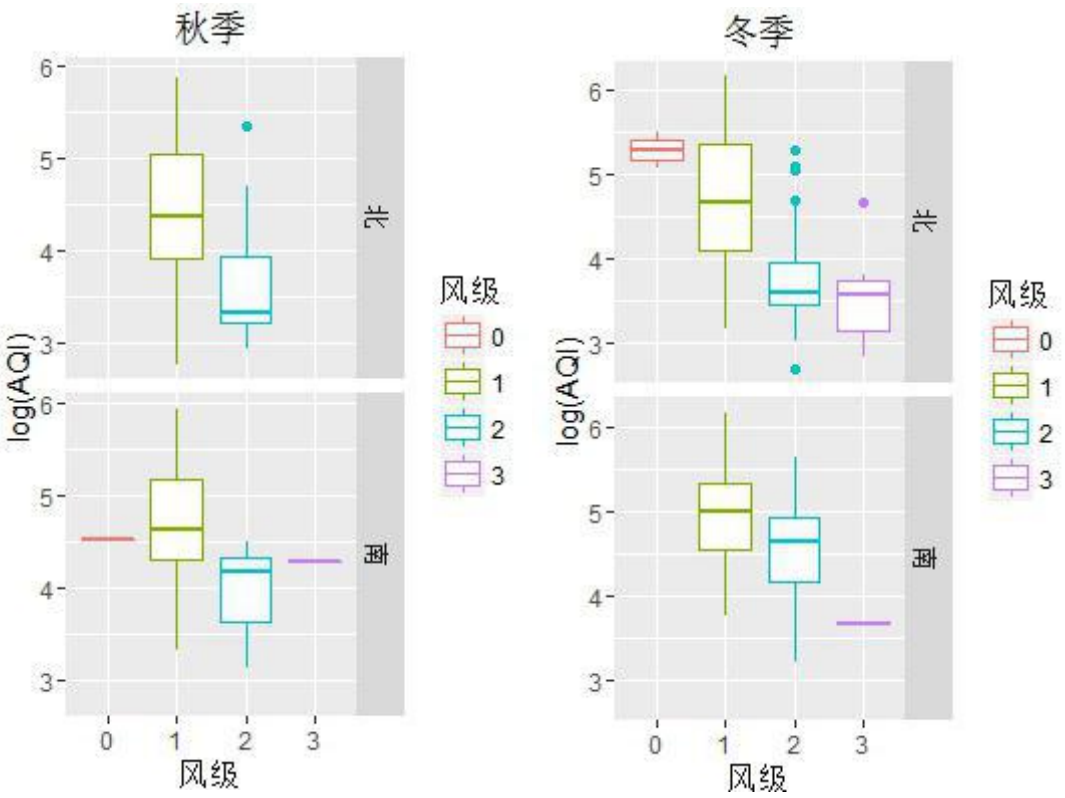


图 16：风向、风级对北京四季 AQI 的影响

3. 时间序列分析

若使用经典线性回归模型来处理 AQI 的预测问题，比如预测明天的 AQI 值，则需要已知明天的 PM2.5、PM10、SO2、NO2、CO 和 O3 浓度等数据。因此线性回归模型在预测 AQI 时是不适用的。而将北京每日 AQI 值看作时间序列数据是合理的，故使用时间序列模型来处理。从图 17 看出 AQI 具有一定的周期性，这与前文的分析是吻合的。

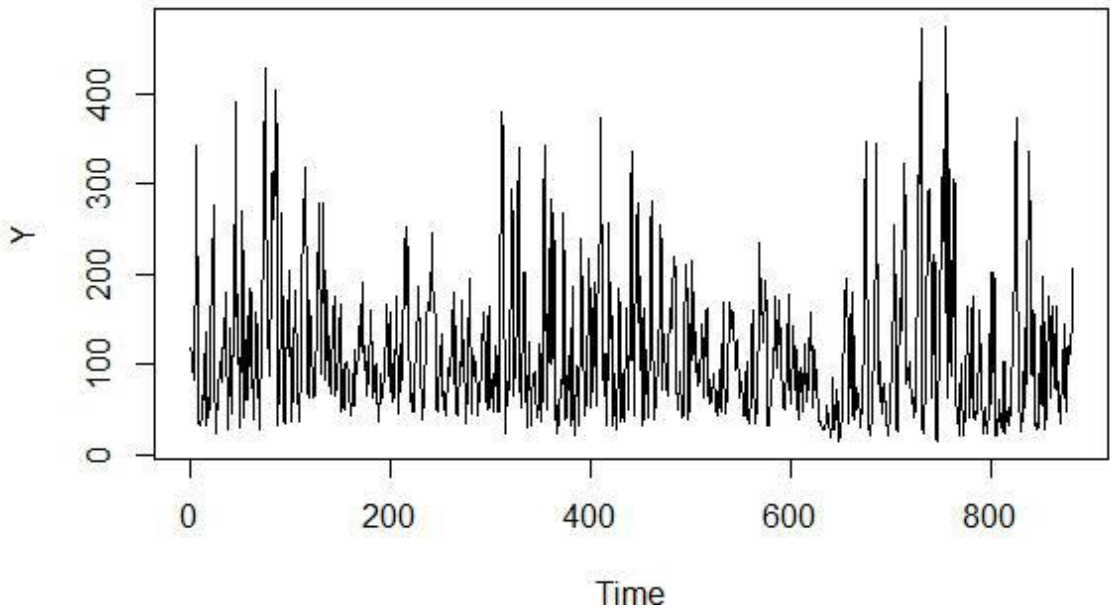


图 17：AQI 时序图

3.1 平稳性检验

在处理时间序列数据之前首先要进行平稳性检验，此处使用 ADF 检验(Augmented Dickey-Fuller Test)。ADF 检验的原假设为该时间序列不平稳。R 运行结果显示 p 值为 0.01，拒绝原假设，故 AQI 为平稳时间序列。

3.2 模型参数选择

由前文分析可知，AQI 数据具有季节效应，故将季节变量转化为哑变量，并与 AQI 进行线性回归，将 AQI 与线性回归拟合值的差进行时间序列分析。这里我们选择 ARMA(p,q)模型来处理。

在 R 中使用 auto.arima 函数进行模型参数 p，q 的选择，运行结果显示，最终选择的模型为 ARMA(1,1)。最终的模型为：$\tilde{Y}_t = 0.3503\tilde{Y}_{t-1} + \varepsilon_t + 0.3614\varepsilon_{t-1}$。此处，$\tilde{Y}_t$ 表示去除季节效应后的 AQI 。

表 5：时间序列分析结果

	AR1	MA1
	0.3503	0.3614
SE	0.0497	0.0494

3.3 AQI 的预测

我们用 3.2 训练出的模型进行 AQI 的预测，这里我们预测 2016 年 5 月 1 日至 5 月 5 日这 5 天的 AQI 值，结果见表 6，图 18。结果显示，预测效果并不理想。关键原因在于，此处只使用了历史的 AQI 的数据，并没有使用其他解释变量，如 PM2.5、PM10 等污染物的浓度。

表 6：AQI 预测结果与真实结果比较

	2016.5.1	2016.5.2	2016.5.3	2016.5.4	2016.5.5
预测值	170	135	123	119	117
真实值	240	80	56	53	92

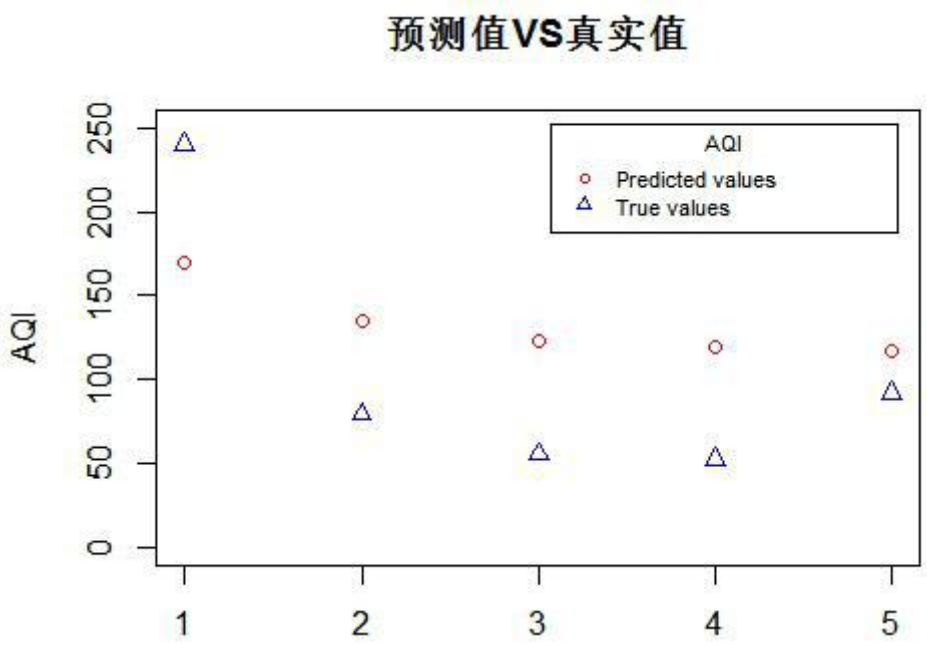


图 18：AQI 五天预测值与真实值比较图

3.4 模型改进

由于只使用了历史的 AQI 的数据，并没有使用其他解释变量，导致预测结果不佳。因此，模型的改进方法是先对 PM2.5、PM10 等 6 项污染物进行时间序列分析，对 6 项污染物进行预测后，再由公式计算出 AQI 的的值。

对 PM2.5、PM10、SO2、CO、NO2 和 O3，这 6 项污染物浓度分别进行平稳性检验，结果显示 p 值分别为 0.01、0.01、0.02197、0.01、0.01 和 0.1582。所以，只有 O3 是非平稳的时间序列，对 O3 取差分后再验证平稳性。差分 O3 序列 p 值为 0.01，是平稳的。

然后，对 PM2.5、PM10、SO2、CO、NO2 和差分 O3 时间序列数据进行建模，与前文 3.2 相似。而此处，使用 auto.arima 函数可以直接选出 ARIMA(p,d,q) 中的最优模型参数 p、d、q，故直接使用 O3 浓度数据，而不是差分 O3 数据。运行出的结果见表 7。

表 7：6 项污染物浓度时间序列分析结果

	模型	AR1	AR2	AR3	AR4	MA1	MA2	MA3	MA4	MA5
PM2.5	ARMA(1,1)	0.3458				0.3568				
PM10	ARMA(1,1)	0.2805				0.3424				
SO2	ARMA(2,5)	0.99	-0.0096			-0.3916	-0.3363	-0.1973	-0.111	0.1735
CO	ARMA(1,1)	0.3206				0.3609				
NO2	ARMA(1,1)	0.336				0.3073				
O3	ARIMA(4,1,2)	-0.4803	0.4042	-0.0711	-0.1245	0.0295	-0.8104			

接下来就要对这 6 项污染物浓度进行预测。同样，预测 2016 年 5 月 1 日至 5 月 5 日这 5 天的 6 项污染物浓度。结果见表 8。

表 8-1：PM2.5 预测结果与真实结果比较

预测值	125.99	93.97	82.89	79.06	77.74
真实值	191.6	50.5	23.8	28.5	53.2

表 8-2：PM10 预测结果与真实结果比较

预测值	151.39	132.60	127.33	125.86	125.44
真实值	176.9	100.5	65.2	57.8	112.1

表 8-3：SO2 预测结果与真实结果比较

预测值	12.42	11.31	11.43	13.18	12.49
真实值	24.3	2.5	2.2	3.8	14.4

表 8-4：CO 预测结果与真实结果比较

预测值	1.23	1.06	1.01	0.99	0.99
真实值	2.26	0.525	0.425	0.541	0.942

表 8-5：NO2 预测结果与真实结果比较

预测值	55.49	51.93	50.73	50.33	50.20
真实值	42.5	20.1	27.8	49.5	45.7

表 8-6：O3 预测结果与真实结果比较

预测值	210.77	173.31	162.89	151.72	162.66
真实值	283	103	107	141	141

由表 8 各项污染物预测值，再由 AQI 计算公式，计算出 AQI 的值，见表 9、图 19。

表 9：模型改进后 AQI 预测结果与真实结果比较

	2016.5.1	2016.5.2	2016.5.3	2016.5.4	2016.5.5
预测值	165	123	109	105	103
真实值	240	80	56	53	92

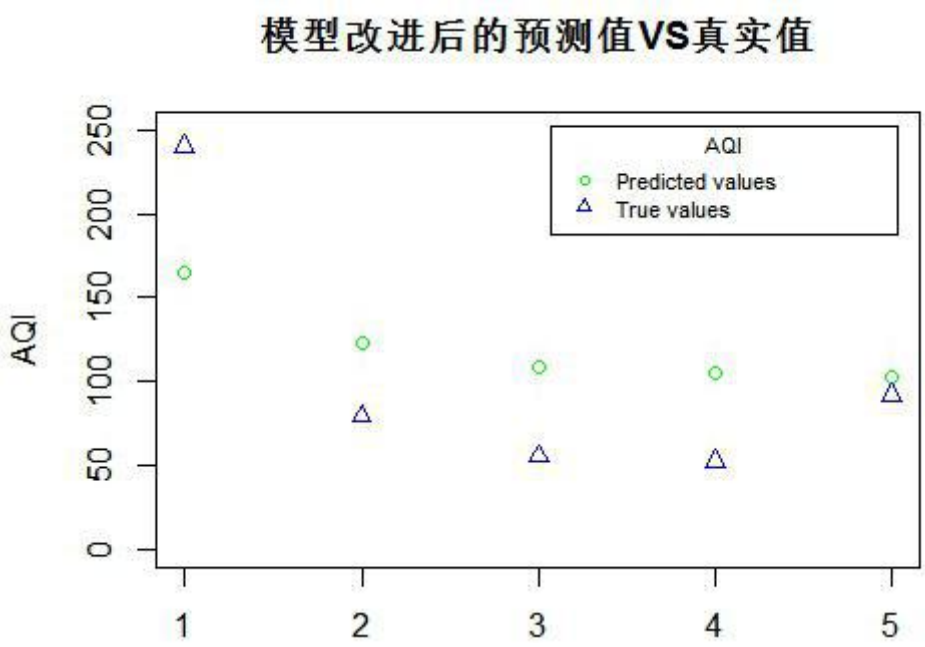


图 19：模型改进后 AQI 五天预测值与真实值比较图

比较图 18 和图 19 可知，改进后的模型确实增加了预测的精度。由图 19 可知，预测效果虽然有所改善，但是结果依然不甚理想。时间有限，便止步于此。