上海财经大学

期 末 论 文

题目：空气质量影响因素探究及预测分析

课程名称 数据分析与可视化

任科教师 刘鑫

姓　　名 马靖淳

院　　系 统计与管理学院

专　　业 数据科学与大数据技术

学　　号 2020111235

一、序言

(一) 研究背景

如今，气候变化已成为一个主要问题。空气正以前所未有的水平受到污染，从而增加了由于空气污染引起的疾病数量。根据 2017 年 10 月 27 日世界卫生组织国际癌症研究机构公布的致癌物清单，空气污染物属于致癌物的一类，因此预测空气质量的走势已经成为现今科学研究的热点。现在，是时候使用技术来评估这个问题并帮助人类。

(二) 空气质量数据集

数据来源是加州大学尔湾分校的机器学习社区，数据贡献者爬取了 2013 年 3 月 1 日到 2017 年 2 月 28 日的时间段北京地区 12 个国家控制的空气质量监测点的每小时空气污染物数据，其中包含 18 个字段如下表，除风向、监测点名称两个变量为定性变量外，其余变量均为定量变量。

表 1-1: 空气质量数据字段说明

No	数据序号	CO	CO 指数
year	年	O3	臭氧指数
month	月	TEMP	气温
day	日	PRES	气压
hour	时	DEWP	露点温度
PM2.5	PM2.5 指数	RAIN	降雨量
PM10	PM10 指数	wd	风向
SO2	SO2 指数	WSPM	风速
NO2	NO2 指数	station	监测点名称

二、空气质量影响因素分析

(一) 空气质量指数 AQI 的计算

AQI(Air Quality Index) 的中文名称为空气质量指数，具体是指根据细颗粒物 (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) 等六项参数综合得出的空气污染程度及空气质量状况的表述，AQI 是六项污染物空气质量分指数中的最大值。

1. 计算各污染物空气质量分指数

利用 mutate 函数计算各污染物空气质量分数，计算的数学公式如下：

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_p - BP_{Lo}) + IAQI_{Lo}$$

式中：

$IAQI_p$——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_{Hi}——附表一 (相应地区的空气质量分指数及对应的污染物项目浓度指数表) 中与 C_p 相近的污染物浓度限值的高位值

BP_{Lo}——附表一中与 C_p 相近的污染物浓度限值的低位值

$IAQI_{Hi}$——附表一中与 BP_{Hi} 对应的空气质量分指数

$IAQI_{Lo}$——附表一中与 BP_{Lo} 对应的空气质量分指数

2. AQI 计算及污染等级的划分

从各污染物的 IAQI 中选择最大值确定为 AQI，新增变量 AQI，当 AQI>50 时确定为首要污染物。再根据附表二 (空气质量指数 (AQI) 范围及相应类别) 判定污染等级，新增变量 classification。

对污染程度 classification 进行统计获得结果如图 2-1，污染等级为 1、2 占总体中的绝大多数，污染程度最严重的类别总数相对较少。

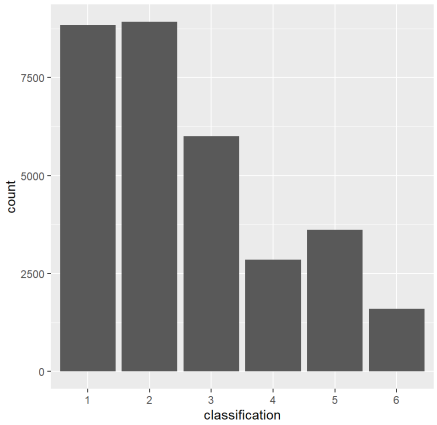


图 2-1: 污染程度统计图

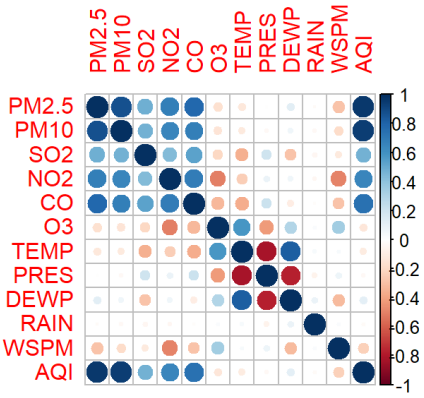


图 2-2: 字段相关性展示图

3. 字段相关性分析

计算并展示数据集中所有定量变量的两两相关性结果如图 2-2 所示，由图片可以看出污染物 PM2.5、PM10、SO2、NO2、CO 之间存在着较强的正相关性。另外，气温及露点温度与气压之间存在较强的负相关性，也比较符合物理常识。再关注最后一行，分析 AQI 与其他变量的相关性，可

以看出 AQI 与 PM2.5 和 PM10 的相关性最强，认为 PM2.5 和 PM10 主要影响空气污染程度，另外 AQI 与 CO、NO2、SO2 的相关性也较强，但与 O3 相关性很弱，说明臭氧不是主要污染物。

(二) 影响空气质量因素分析

1. 风的影响

如图 2-3 所示，该可视化图选择风速 WSPM 作为 x 轴，AQI 作为 y 轴，分析 16 幅不同风向 (wd) 图形，可以发现样本的分布情况较为相似，说明风向几乎对污染物含量不产生什么影响，这是因为风向只影响着污染物的扩散方向，不影响污染物总量。

而分析每幅图中不同颜色 (classification) 点的分布情况可以看出，风速越小时（越靠近左侧）污染程度越高的点分布越集中，说明随风速增大，污染程度逐渐降低，这是因为风速的大小决定着污染物的扩散和稀释状况，通常情况下，污染物在大气中的浓度与平均风速成反比。

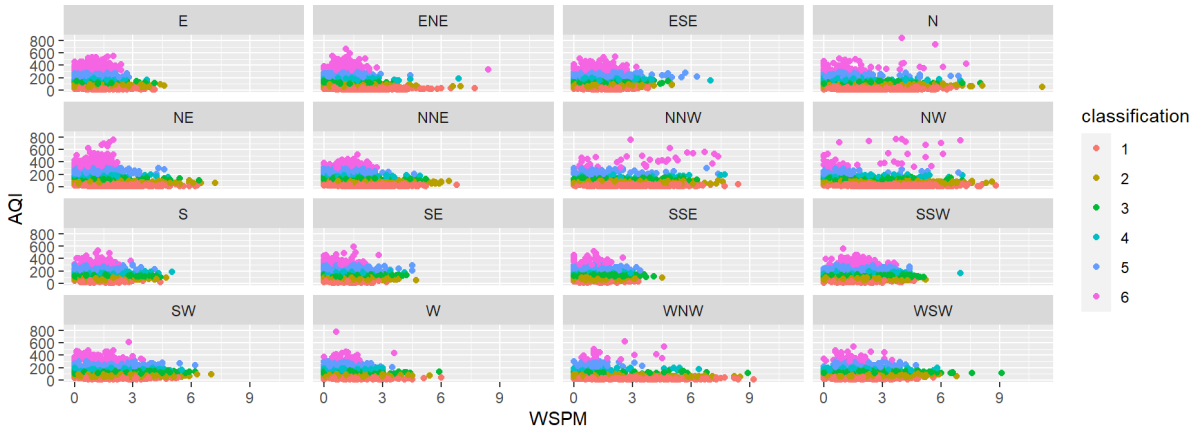


图 2-3: 风速、风向与 AQI 关系图

2. 降水的影响

如图 2-4 左图所示，横轴代表污染程度 (classification)，条形图颜色代表降水的多少 (RAIN)，随污染程度上升，红色部分（降雨量为 0）占比逐渐增大，说明降水可以缓解污染程度，这是因为各种形式的降水，特别是降雨，能有效地吸收、淋洗空气中各种污染物。

3. 气温气压的影响

如图 2-4 右图所示，横轴代表气温变化 (TEMP)，各条形图颜色代表污染程度 (classification)，由于气温在 0 度以下及 40 度以上天气相对较少，重点关注 0 40 度这一区间的污染程度变化。关注污染程度最严重的粉色列，发现其随气温升高，分布逐渐减少，与其对比的是土黄色列，代表污染程度为 2 时的总数，其随温度升高逐渐增多。这是因为气温低时大气扩散度和稀释度降低，导致污染物浓度较高；而当气温较高时，相对湿度大，因而污染物浓度较低。

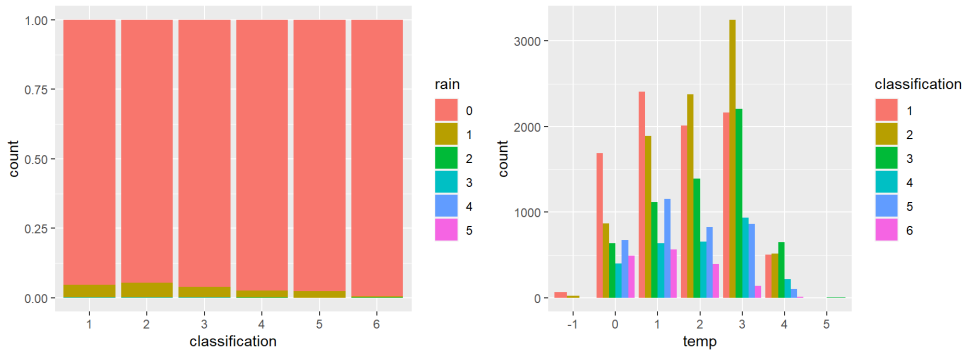


图 2-4: 左图：各污染程度降水比例图，右图：污染程度随气温变化图

(三) 污染物的时间变化

根据图 2-5 左上角图所示，可以看出各污染物在 2017 年显著下降，根据背景调查发现，污染物下降得益于控制燃煤锅炉、提供更清洁的家用燃料以及产业结构调整等措施。分析右上角图，在冬季 11-1 月份，污染物含量显著上升，这是因为北京在冬季会大量燃烧煤进行供暖。

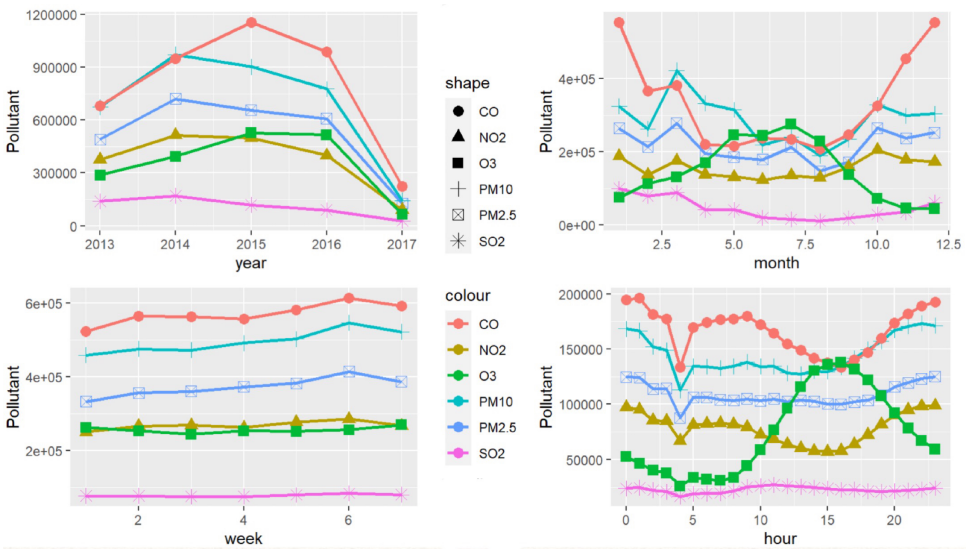


图 2-5: 污染物的时间变化图

分析图 2-5 左下角随星期变化图，可以发现各污染物变化比较平稳，在工作日略有上升，周末略有下降。再分析右下角随小时变化图，在营业时间（上午 9:00 到晚上 7:00）这一区间，其他气体除臭氧外与 PM2.5 的变化比较一致，含量比较稳定，从 20 点以后各气体污染比重加大，据研究是由于一些工厂会在半夜排放污染气体。

三、基于 PM2.5 的回归分析

据前文分析，可以看出 AQI 与 PM2.5 的相关性最强，并且根据科学研究其是对人类影响最大的污染物，故现对 PM2.5 进行回归分析，衡量空气的污染程度。

(一) 基于 PM2.5 的线性回归

根据前文污染物随时间变化图，可以发现 PM10 与 PM2.5 变化轨迹几乎一致，并且两者相关性极强，所以去除 PM10 对 PM2.5 进行回归分析。对数据进行标准化后获得线性回归结果如图 3-1：

```
Call:
lm(formula = as.matrix(y_train) ~ ., data = X_Train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9500 -0.2889 -0.0538  0.1907  7.0966

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.002660   0.003261  -0.816  0.41457
month       -0.056285   0.003848 -14.627  < 2e-16 ***
day          0.010538   0.003279   3.214  0.00131 **
hour        -0.011449   0.003544  -3.231  0.00124 **
SO2          0.096203   0.004339  22.170  < 2e-16 ***
NO2          0.338707   0.005667  59.772  < 2e-16 ***
CO           0.556977   0.005234 106.417  < 2e-16 ***
O3           0.173880   0.005140  33.828  < 2e-16 ***
TEMP        -0.140471   0.009521 -14.753  < 2e-16 ***
PRES        -0.005381   0.006538  -0.823  0.41044
DEWP         0.290130   0.008365  34.685  < 2e-16 ***
RAIN        -0.014284   0.003236  -4.414 1.02e-05 ***
WSPM         0.079376   0.004302  18.449  < 2e-16 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.5202 on 25439 degrees of freedom
Multiple R-squared:  0.7228,  Adjusted R-squared:  0.7227
F-statistic:  5528 on 12 and 25439 DF,  p-value: < 2.2e-16
```

图 3-1: 线性回归结果图

```
Step:  AIC=-33256.85
as.matrix(y_train) ~ month + day + hour + SO2 + NO2 + CO + O3 +
    TEMP + DEWP + RAIN + WSPM

       Df Sum of Sq    RSS    AIC
<none>              6884.0 -33257
- day    1      2.70 6886.7 -33249
- hour   1      2.94 6887.0 -33248
- RAIN   1      5.20 6889.2 -33240
- month  1     65.26 6949.3 -33019
- TEMP   1     68.87 6952.9 -33005
- WSPM   1     92.48 6976.5 -32919
- SO2    1    134.45 7018.5 -32767
- O3     1    309.50 7193.5 -32140
- DEWP   1    354.30 7238.3 -31982
- NO2    1    971.01 7855.0 -29900
- CO     1   3064.30 9948.3 -23887
```

图 3-2: 变量选择结果图

回归结果表明，模型整体通过 F 检验具有显著性。另外，除 PRES 外，其余变量均通过 α=0.05 参数检验，具有显著性。而由变量相关图分析得，一些变量之间相关性显著，故认为变量之间可能存在多重共线性问题，现进行对多重共线性的检验。

采用特征根判定法进行判断，计算获得条件数 k 值为 152.94231>100，根据经验判断当 100 ≤ k ≤ 1000 时认为设计矩阵 X 存在较强的多重共线性。

采用 stepwise 方法重新进行变量选择，获得结果如图 3-2，剔除了变量 PRES 后，模型整体显著，重新计算条件数 k 值为 46.06486<100。最终计算 test-MSE 结果为 0.3087812。

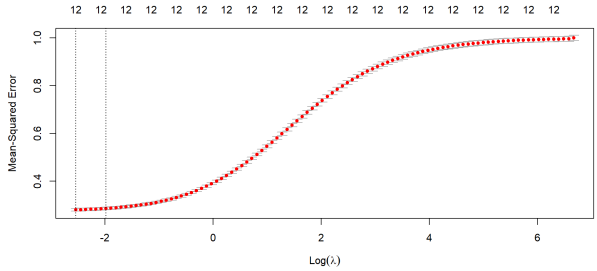
(二) 基于 PM2.5 的岭回归

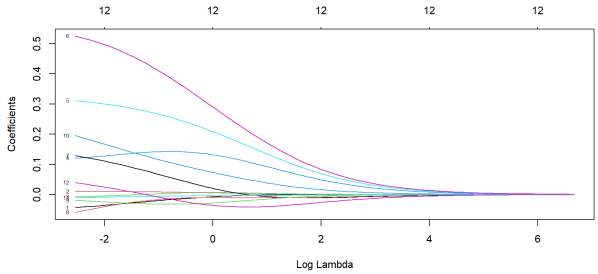
因变量之间存在多重共线性问题，现进行岭回归估计，获得结果如图 3-3，根据上图可以看出最优 λ 值大致位置，当取最优 λ 值为 0.07883541 时，获得各变量参数如图 3-3 下方，岭回归不具有变量选择功能，但对参数值进行了调节，最终 test-MSE 结果为 0.3135175。

(三) 基于 PM2.5 的 lasso 回归

现进行 lasso 回归估计，获得结果如图 3-4，根据上图可以看出最优 λ 值大致位置，当取最优 λ 值为 0.000793614 时，获得各变量参数如图 3-4 下方，lasso 可以进行变量选择，根据图 3-4 中间图

可看出最终选择变量数量为 7 个，有 5 个变量的值几乎接近于 0，最终 test-MSE 结果为 0.3088395。

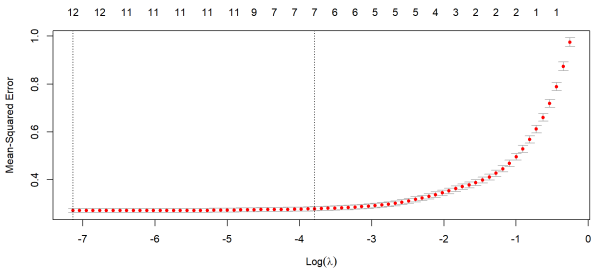


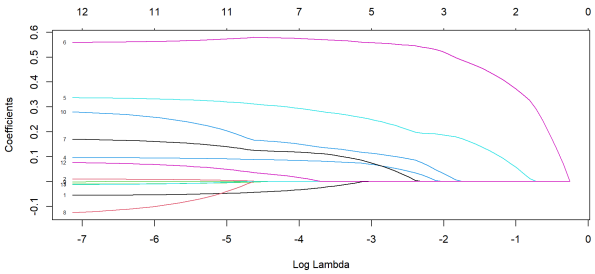


13 x 1 sparse Matrix of class "dgCMatrix"

	s0
(Intercept)	-0.002766401
month	-0.039262358
day	0.010718445
hour	-0.008428215
SO2	0.113249935
NO2	0.316958916
CO	0.519796737
O3	0.131598259
TEMP	-0.057799528
PRES	-0.016265471
DEWP	0.195927440
RAIN	-0.008776871
WSPM	0.042603563

图 3-3: 岭回归结果图





13 x 1 sparse Matrix of class "dgCMatrix"

	s0
(Intercept)	-0.002646740
month	-0.055660640
day	0.009739074
hour	-0.010756517
SO2	0.095695021
NO2	0.336234638
CO	0.559171371
O3	0.169356521
TEMP	-0.126522911
PRES	-0.002524300
DEWP	0.280212998
RAIN	-0.012926118
WSPM	0.075734829

图 3-4: lasso 回归结果图

三个模型之中经变量选择后的 linear model 具有最小的 test-MSE，其次是 lasso 模型，表现最差的是岭回归模型。

四、基于 LSTM 模型预测 PM2.5 浓度及 AQI

(一) 模型简介

LSTM(Long Short-Term Memory) 是长短期记忆网络，是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。LSTM 是一种特殊的 RNN（循环神经网络），主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。简单来说，就是相比普通的 RNN，LSTM 能够在更长的序列中有更好的表现，且能解决 RNN 梯度消失或者梯度爆炸的问题。

由图 4-1 可以看出，RNN 只有一个传递状态 h^t，LSTM 有两个传输状态，一个 c^t(cell state) 和一个 h^t(hidden state)。另外，LSTM 模型中存在 sigmoid 和 tanh 两种激活函数，而不是选择统一一种激活函数，这是因为 sigmoid 用在了各种 gate 上，产生 0 1 之间的值，一般只有 sigmoid 最直接

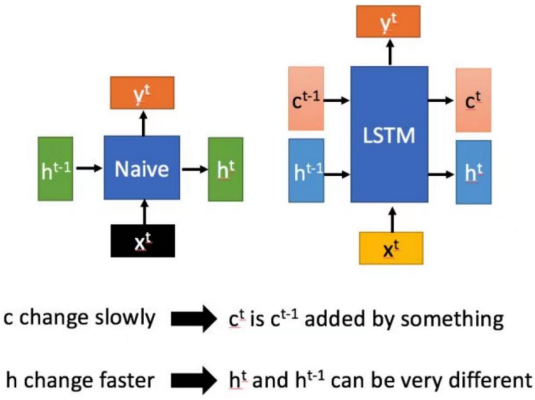


图 4-1: RNN 与 LSTM 输入输出对比图 (图源网络)

了；tanh 用在了状态和输出上，是对数据的处理。

(二) 模型应用与分析

1. 数据预处理

利用 python 中 datetime 包，将年、月、日、小时四项进行合并，变成单独的一个时间字符串，比如转换后结果为 “2013-03-01 00:00:00”。与上一节回归中一样，因 PM10 与 PM2.5 变化轨迹几乎一致，并且两者相关性极强，所以去除 PM10 保留其余定量变量代入模型中进行计算。

2. RNN 方法预测

采用 RNN 方法进行预测，50epoch 后获得结果如图 4-2，最终获得损失函数值为 5043.9761，均方误差为 52.5923。

```
Epoch 42/50
273/273 [==============================] - 1s 3ms/step - loss: 4076.4517 - mae: 46.9997 - val_loss: 5207.8350 - val_mae: 53.0668
Epoch 43/50
273/273 [==============================] - 1s 4ms/step - loss: 4034.1675 - mae: 46.7748 - val_loss: 5146.3511 - val_mae: 53.0692
Epoch 44/50
273/273 [==============================] - 1s 4ms/step - loss: 3988.1782 - mae: 46.4873 - val_loss: 5133.3799 - val_mae: 52.9704
Epoch 45/50
273/273 [==============================] - 1s 4ms/step - loss: 3960.6548 - mae: 46.3188 - val_loss: 5111.4072 - val_mae: 52.9195
Epoch 46/50
273/273 [==============================] - 1s 3ms/step - loss: 3923.3315 - mae: 46.1288 - val_loss: 5132.8486 - val_mae: 53.0027
Epoch 47/50
273/273 [==============================] - 1s 3ms/step - loss: 3878.2271 - mae: 45.8934 - val_loss: 5104.8232 - val_mae: 52.8555
Epoch 48/50
273/273 [==============================] - 1s 3ms/step - loss: 3845.8645 - mae: 45.6801 - val_loss: 5136.4546 - val_mae: 52.7081
Epoch 49/50
273/273 [==============================] - 1s 3ms/step - loss: 3820.3840 - mae: 45.5468 - val_loss: 5134.8521 - val_mae: 52.7614
Epoch 50/50
273/273 [==============================] - 1s 3ms/step - loss: 3771.7722 - mae: 45.2687 - val_loss: 5043.9761 - val_mae: 52.5923
```

图 4-2: RNN 结果

3. LSTM 方法预测

在采用 LSTM 模型预测 PM2.5 浓度的过程中，需要对选择多少历史数据进行预测进行衡量。选择之前过多的数据比如说前一个月的数据进行预测，可能出现较早的信息与新的值预测无关，从而

产生信息冗余；而选择过少的数据，比如仅仅采用前一天或前几个小时的数据进行预测，预测结果也不是很好。最终选择采用前五天的数据对第六天的数据进行预测，50epoch 后结果如图 4-3，最终损失函数值为 2966.3672，均方误差为 34.5268，结果相较于 RNN 方法有显著提升。

```
Epoch 42/50
273/273 [==============================] - 78s 286ms/step - loss: 3830.6843 - mae: 39.3504 - val_loss: 3691.6575 - val_mae: 40.3123 - lr: 0.0010
Epoch 43/50
273/273 [==============================] - 77s 282ms/step - loss: 3861.9287 - mae: 40.0499 - val_loss: 3608.7966 - val_mae: 40.2206 - lr: 0.0010
Epoch 44/50
273/273 [==============================] - 78s 284ms/step - loss: 3456.3359 - mae: 36.8220 - val_loss: 3389.7278 - val_mae: 37.5160 - lr: 0.0010
Epoch 45/50
273/273 [==============================] - 78s 284ms/step - loss: 3425.8489 - mae: 36.5117 - val_loss: 3884.2761 - val_mae: 41.0319 - lr: 0.0010
Epoch 46/50
273/273 [==============================] - 78s 284ms/step - loss: 3328.1504 - mae: 35.9220 - val_loss: 3219.0481 - val_mae: 36.4861 - lr: 0.0010
Epoch 47/50
273/273 [==============================] - 78s 286ms/step - loss: 3352.6699 - mae: 36.2944 - val_loss: 3294.2966 - val_mae: 37.4533 - lr: 0.0010
Epoch 48/50
273/273 [==============================] - 78s 285ms/step - loss: 3144.3279 - mae: 34.7289 - val_loss: 3071.8643 - val_mae: 35.4447 - lr: 0.0010
Epoch 49/50
273/273 [==============================] - 79s 288ms/step - loss: 2986.4802 - mae: 33.5247 - val_loss: 2940.8044 - val_mae: 34.4791 - lr: 0.0010
Epoch 50/50
273/273 [==============================] - 78s 285ms/step - loss: 3079.2251 - mae: 34.3783 - val_loss: 2966.3672 - val_mae: 34.5268 - lr: 0.0010
```

图 4-3: LSTM 结果

对 LSTM 模型与 RNN 模型训练过程进行可视化，图 4-4 展示的是各 epoch 的 Loss 对比图，图 4-5 展示的是各 epoch 的 MSE 变化情况，可以看出验证集在训练过程中都有一定波动变化，但 LSTM 训练效果明显好于 RNN 模型。图 4-6 的是用 LSTM 模型新预测的 PM2.5 代入公式重新计算获得的 AQI 与真实数据计算 AQI 的对比折线图，可以看出预测结果与实际结果变化趋势一致且很接近。而图 4-7 展示的是 RNN 方法预测结果与实际结果对比，可以看出 RNN 方法预测结果都相对偏大，可能是导致 MSE 偏高的原因。

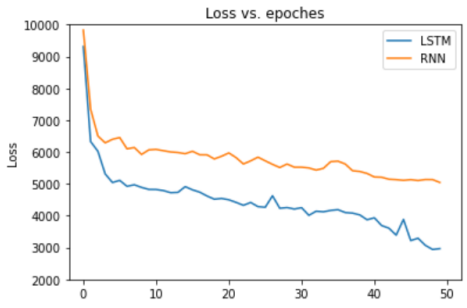


图 4-4: 两个模型损失函数值变化图

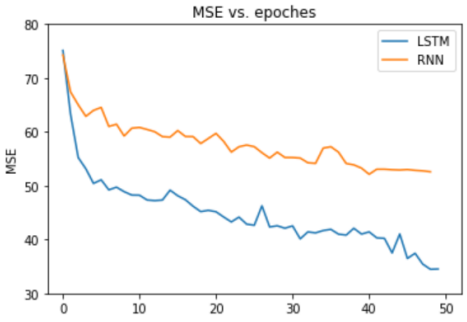


图 4-5: 两个模型均方差值变化图

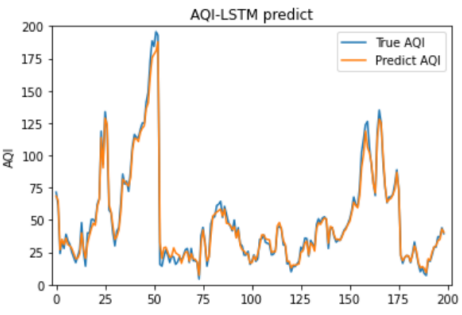


图 4-6: LSTM 预测 AQI 对比折线图

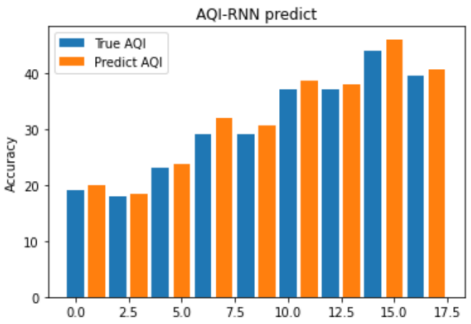


图 4-7: RNN 模型预测 AQI 对比条形图

参考文献

[1] 杨柳. 基于深度学习的空气质量数据智能质控的研究与应用 [D]. 中国科学院大学 (中国科学院沈阳计算技术研究所),2022.DOI:10.27587/d.cnki.gksjs.2022.000011.

[2] 赵小明, 顾珂铭, 张石清. 面向深度学习的空气质量预测研究进展 [J]. 计算机系统应用,2022,31(11):49-59.DOI:10.15888/j.cnki.csa.008847.

[3] 徐洪珍, 宋文琳, 韦诗㊣, 王强. 一种基于混合深度学习模型的空气质量预测方法 [P]. 江西省: CN115293269A,2022-11-04.

五、附录

(一) 附表一　相应地区的空气质量分指数及对应的污染物项目浓度指数表

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
说明：	(1) 二氧化硫（SO_2）、二氧化氮（NO_2）和一氧化碳（CO）的 1 小时平均浓度限值仅用于实时报，在日报中需使用相应污染物的 24 小时平均浓度限值。 (2) 二氧化硫（SO_2）1 小时平均浓度值高于 800 μg/m³ 的，不再进行其空气质量分指数计算，二氧化硫（SO_2）空气质量分指数按 24 小时平均浓度计算的分指数报告。 (3) 臭氧（O_3）8 小时平均浓度值高于 800 μg/m³ 的，不再进行其空气质量分指数计算，臭氧（O_3）空气质量分指数按 1 小时平均浓度计算的分指数报告。									

(二) 附表二　空气质量指数 (AQI) 范围及相应类别

空气质量指数	空气质量指数级别	空气质量指数类别及表示颜色		对健康影响情况	建议采取的措施
0～50	一级	优	绿色	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	黄色	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	橙色	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	红色	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、呼吸系统疾病患者避免长时间、高强度的户外锻练，一般人群适量减少户外运动
201～300	五级	重度污染	紫色	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、老年人和心脏病、肺病患者应停留在室内，停止户外运动，一般人群减少户外运动
＞300	六级	严重污染	褐红色	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、老年人和病人应当留在室内，避免体力消耗，一般人群应避免户外活动