关于影响空气质量指数因素的分析

刘宁哲

摘要：本报告主要研究天气相关因素对空气质量指数（AQI）的影响，为预测 AQI 提供理论依据。报告首先构建若干天气相关因素的指标，再建立线性回归模型探索各指标对 AQI 的影响。模型结果表示，风力与 AQI 显著相关。在此基础上，进一步对 AQI 预测的应用场景进行探讨。

一、背景介绍

空气质量指数（AQI）的用途是报告每日空气质量，即空气的干净或者污染程度，它专注于人们在呼吸受污染空气一段时间后可能遇到的健康影响，可以让人们了解每日呼吸的空气是否安全。AQI 被视为一个衡量标准，AQI 值越高，空气污染情况越严重，对人们的健康影响程度也就越大。例如，50 及以下的 AQI 值代表良好的空气质量，超过 300 的 AQI 值代表危险的空气质量。在中国，通常来讲，AQI 值在 100 及以下被认为是令人满意的，相应的，对于某些敏感人群，AQI 值超过 100 时的空气质量就是不健康甚至是有害的。AQI 按照不同的程度可分为六个级别，这些级别让人们可以轻松快速的确定空气质量是否达到不健康的水平。

就像天气预报让你知道当日应该穿何种衣服，或者是否需要带伞出门那样，使用 AQI 预测可以帮助你规划当日的户外活动，让你知道什么时候应该减少外出，或者至少佩戴口罩进行防护以降低你吸入的空气污染量。对于敏感人群，例如患有肺病或者心脏病的人、老年人和儿童等，也能够起到关键的提前预警作用。许多国家的卫生机构或者组织也会对 AQI 进行监控和预测，在空气质量非常差的时候，各组织甚至可能会动用应急计划，例如命令主要的排放来源如燃煤工业、汽车出行等进行停工或者停驶，直到危险情况解除。所以，AQI 是一个非常重要的指数。

基于此，本报告以从中国天气网抓取的 500 条 AQI 数据为因变量，以最高温度、最低温度以及风力作为自变量，探究哪些变量是影响天气质量指数的重要因素，为用户预判空气质量情况提供帮助。

二、数据来源和指标设计

本报告数据来源于中国天气网的可信数据来源，通过抓取技术手段获得。希望通过各地区的天气相关数据，分析影响 AQI 的因素。此次研究的因变量为空气质量指数 AQI，同时收集了最高温度、最低问题以及风力作为自变量，具体变量说明见表 1。

（1）AQI：空气质量指数。该指标能够对空气质量进行定量描述。根据中国对空气质量级别的划分，AQI 超过 300 属于中度污染，超过 400 即属于重度污染情况，都能说明所研究城市的空气质量一般。

（2）最高温度：当日最高的温度。该指标描述了当条数据对应日期的最高气温。高温天气会产生下沉气流，会使空气从上往下不断循环流动，污染物有可能会被带到高空，对空气质量造成一定的影响。

（3）最低温度：当日最低的温度。该指标描述了当条数据对应日期的最低气温。低温会使得气压下降，使得空气中的元素凝结或者下沉，所以气温的降低有可能对空气质量有着一定的影响。

（4）风力：当日的风力情况。通常情况下，风速越大，越有利于空气污染物的稀释扩散。但在北方的冬春干燥季节，地面沙尘较多，如果风速过大，反而会卷起地面的尘粒，形成大风扬沙，严重破坏空气质量。所以风力和空气情况密不可分，可作为一个重要因素进行研究。

表1：变量说明表

变量类型	变量名	变量说明	取值范围
因变量	AQI	数值型变量	15~429
自变量	最高温度	数值型变量，单位：摄氏度	-2~38
	最低温度	数值型变量，单位：摄氏度	-10~27
	风力	数值型变量，单位：级	1~5

三、描述分析

接下来我们进行描述分析，检查数据质量，并初步判断各解释性自变量与 AQI 之间的关联，为后续的建模研究进行铺垫。

（一）数据概要

表 2：数据概要描述表

指标/参数	数量	平均数	标准差	最小值	中位数	最大值
最高温度	500	19.484	11.0885371	-11	22	38
最低温度	500	9.262	10.8123619	-16	11	27
风力	500	1.946	0.9449178	1	2	5
AQI	500	96.69	63.7431115	15	82	429

观察各数据的数据分布，通过计算得到的结果如表2 所示。从中可以发现以下情况，（1）该城市 AQI 中位数为82，均值为 97，空气质量等级处于“良”的水平，空气污染治理效果较好；（2）该城市温度差异较大，最高温度最大值达38℃，最小值仅-11℃，最低温度最大值达27℃，最小值仅-16℃，极端情况差异十分

明显，最大温差可达 18℃，最小温差仅 2℃。（3）该城市风力处于比较和缓的状态，平均风力为 1.95 级，风力中位数为2级。

（二）因变量

AQI 数据其直方图分布如图 1 左图所示，呈现明显的右偏分布，样本数据长尾效应比较明显，说明有少量样本在这个变量上取值非常大，远远大于其他样本，而且存在部分超过400的数据。所以考虑对该变量做对数变换，变换后的直方图如图 1 右图所示，分布情况有了明显改善，更接近正态分布，利于模型估计。

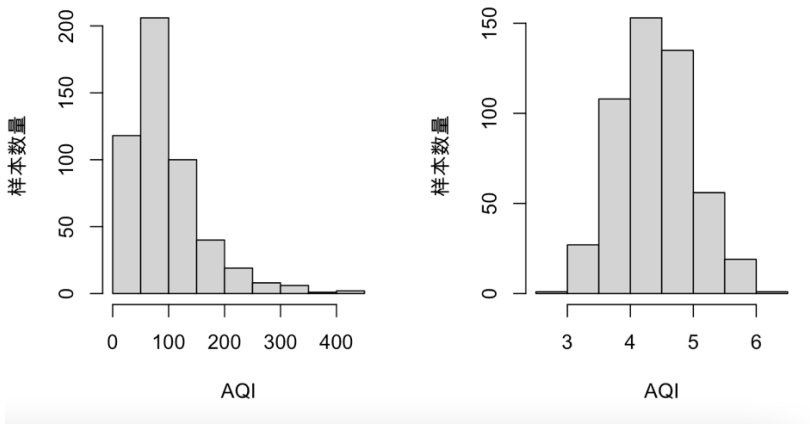

图 1：AQI 直方图

（三）自变量

自变量数据直方图分布如图 2 所示，观察最高气温、最低气温、风力数据分布的情况，未发现明显的数据异常。

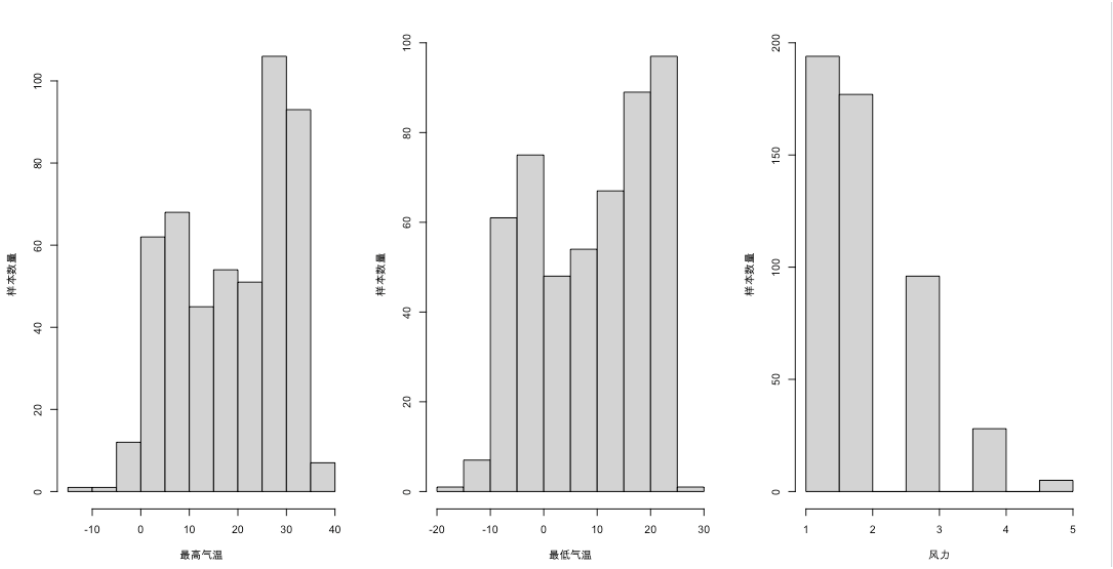

图 2：自变量（最高气温、最低气温、风力）直方图

（四）相关分析

进一步，针对自变量与因变量的相关关系进行描述分析。通过图 3 及图 4 可见，不同空气质量对应的最低温度和最高温度分布均呈现出明显不同。总的来说，伴随着空气质量的下降，最低温度和最高温度均呈现先升后降的趋势。空气质量在:“良”和“轻度污染”的情况下，无论是最低温度还是最高温度均较高；当空气质量在“重度污染”和“严重污染”的情况下，其温度分布下降明显，相较其余情况均偏低。

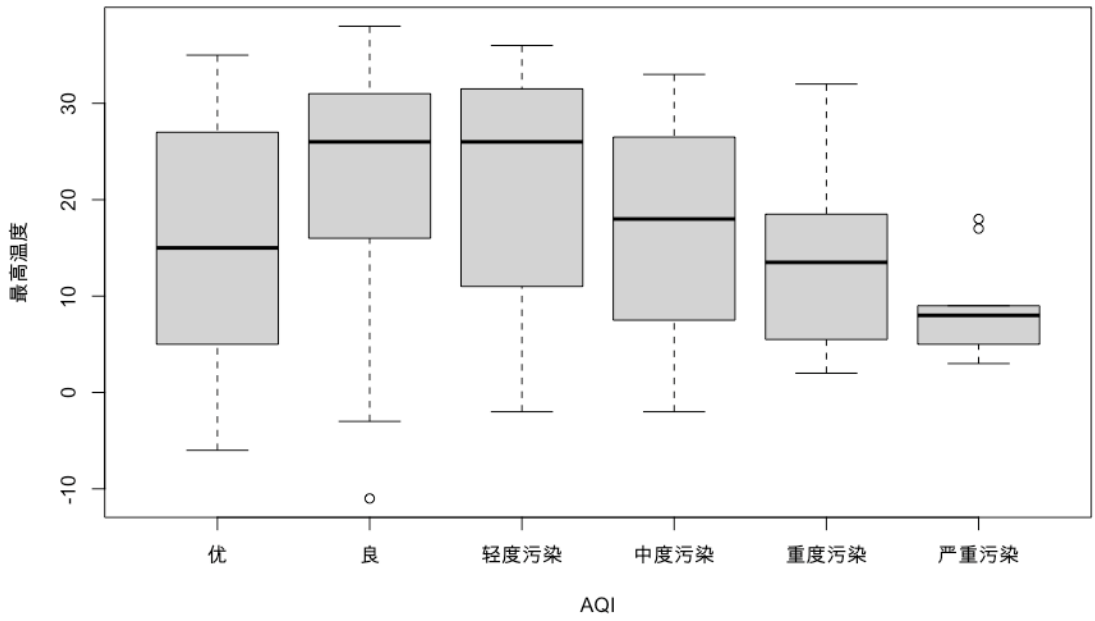

图 3：最高温度分布对应的 AQI

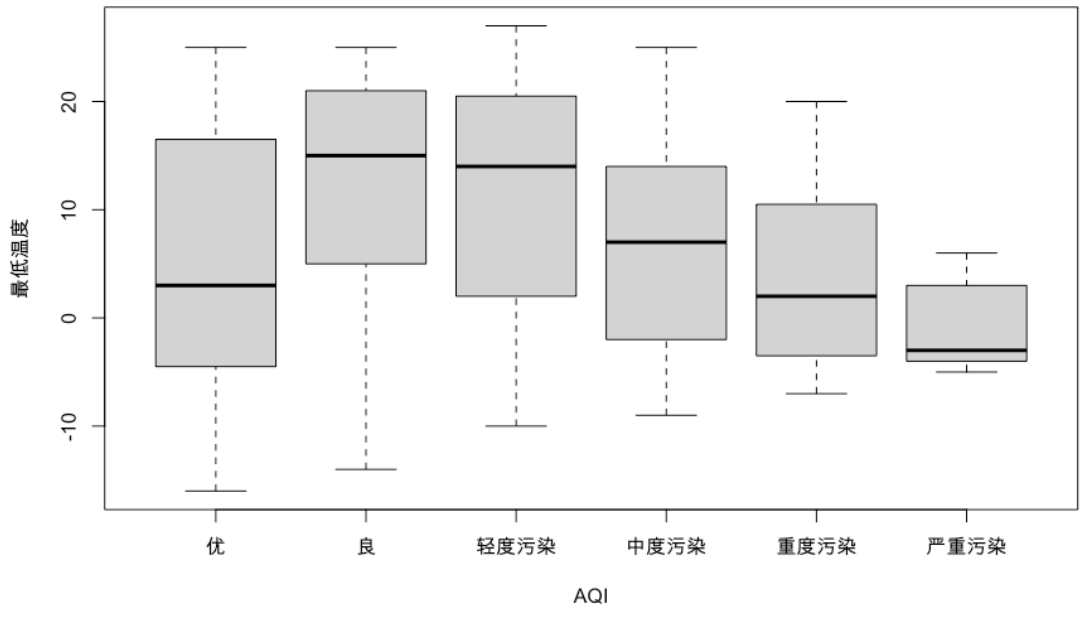

图 4：最低温度分布对应的 AQI

风力对 AQI 有明显的线性影响，风力越大，AQI 越低。如图 5 所示。

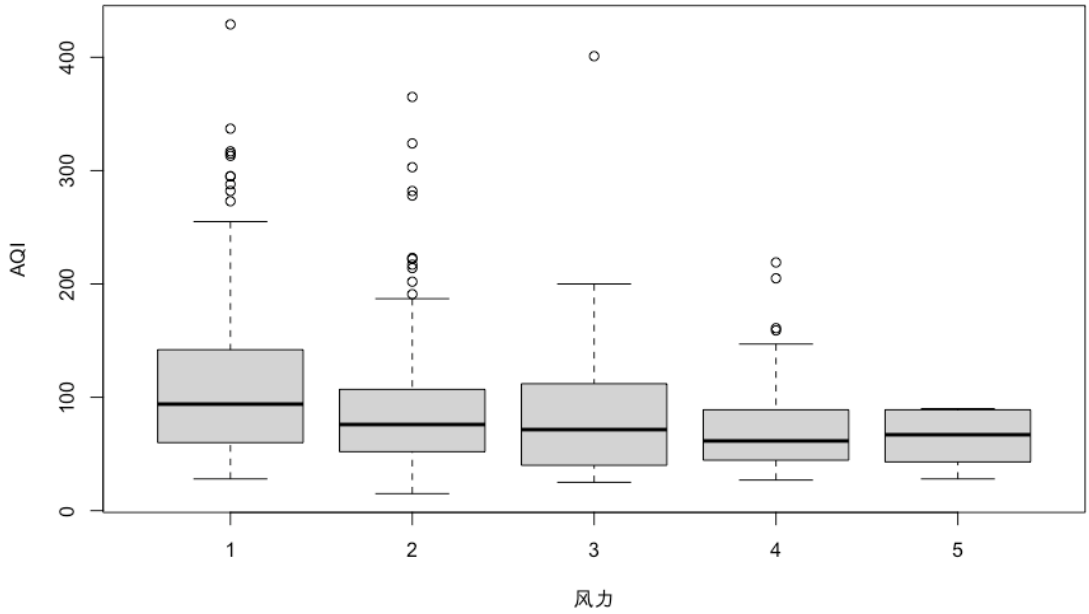

图 5：风力对应的 AQI

四、建模分析

在描述分析的基础上，通过建立线性回归模型进一步分析因变量空气质量指数 AQI 和各个自变量（最高气温、最低气温、风力）之间的关系。

首先，对数据建立回归全模型，相关参数估计及检验结果如表 3 所示。模型整体的 F-检验高度显著（P 值<0.001），这说明至少有一个解释变量跟对数 AQI 显著相关。模型调整后的判决系数为 0.059。对表 3 中的每一个自变量的 t-检验结果进行分析，发现风力因素较为显著，而最高气温、最低气温都不显著。这与描述分析中的情况比较契合，最高气温和最低气温呈现先升后降的波动趋势，而风力越大则 AQI 越低。这也与直观感受相符合，风力越大越可能吹散空气中的污染物使得 AQI 提升，同时，在多重共线性检验中，最高气温、最低气温的膨胀因子都较大，说明自变量之间可能存在多重共线性，在描述分析的部分，我们也发现最高气温和最低气温相似的变化趋势，t-检验结果的不显著可能与多重共线性有关。

然后，我们根据 AIC 和 BIC 准则进行模型选择，得到精简的回归模型结果如表4 所示。本报告发现，AIC 准则和 BIC 准则下模型选择的结果一致，都只保留了风力这个自变量。从显著性上来看，结论与全回归模型一致，即风力对于 AQI 有着显著的影响。

综合上述模型构建与选择，在控制其他变量不变的情况下，我们可以通过本次回归可以得出以下结论：

1）首先，在控制其他变量不变的情况下，风力是影响 AQI 的最重要因素，这可能由于风力越大，空气中的杂质被吹走或者稀释的可能性越大，AQI 降低，意味着风力一定程度决定了空气质量。

2）其次，最低气温、最高气温与 AQI 没有呈现明显的相关，且不显著。可说明这两个自变量在本次模型中对 AQI 没有直接的解释能力。

3）目前数据量过少，部分变量分组样本不均衡，导致模型拟合程度和解释性一般。后续可以通过增加数据量以发现更多具有普适性的观点。

表 3：回归全模型表

变量	回归系数	标准误差	Pavalue
最高温度	0.004	0.004	0.261
最低温度	-0.002	0.004	0.571
风力	-0.064	0.013	<0.001
F 校验	<0.001		
Rsquare	0.064		
adj-Rsquare	0.059		

表 4：线性回归 AIC 模型选择结果

变量	回归系数	标准误差	Pavalue
风力	-0.141	0.027	<0.001
F 校验	<0.001		
Rsquare	0.054		
adj-Rsquare	0.048		

五、模型应用

利用本次分析结果，我们可以通过风力水平来预测空气质量等级，帮助人们更好地理解那些影响空气质量的关键因素。例如，尝试在天气类产品中加入空气质量预测功能，具体实现可以将 AQI 预测值加入使用界面，通过用户所选地区及所选日期的风力数据，展现当时当地的次日 AQI 预测数据，使用者可以提前一天知道次日的空气质量情况，进而做出对应的动作来保护自己的健康。

图 6:AQI 预测产品原型图

六、结论与展望

本报告以空气质量指数 AQI 为因变量，以构建描述天气特征的各项指标作为自变量，包括最低气温、最高气温、风力等 3 个变量。通过线性回归模型研究了 AQI 与各自变量间的相关关系。本报告主要结论为：风力与 AQI 显著相关，风力越大，AQI 越低。在此基础上，我们对不同地区不同时间的 AQI 进行评估预测，希望带来实际的产品化方案。本报告目前模型的拟合程度一般，预测能力还有待提升，需要挖掘更多解释因素和更多、更均衡的样本来解释和预测 AQI 空气质量水平。