**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，即空⽓的⼲净或污染程度，它专注于⼈们在呼吸受污染空⽓⼏个⼩时或⼏天内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，空⽓污染⽔平越⼤，健康问题也越⼤。例如，50及以下的AQI值代表良好的空⽓质量，⽽超过300的AQI值代表危险的空⽓质量。对于污染物，100 的AQI值通常对应于环境空⽓浓度，该浓度相当于保护公共卫⽣的短期国家环境空⽓质量标准⽔平。AQI 值在 100 或以下通常被认为是令⼈满意的。对于每个⼈来说，特别是某些敏感⼈群，当 AQI 值超过 100 时，空⽓质量是不健康甚至是有害的。AQI 分为六类，每个类别都对应于不同级别的健康问题，这些级别让⼈们可以轻松快速确定社区空⽓质量是否达到不健康的⽔平。AQI 的计算需要从空⽓监测器或模型中获得的特定平均期内的空⽓污染物浓度。总⽽⾔之，浓度和时间代表了空⽓污染物的剂量，再由流⾏病学研究确定了与给定剂量对应的健康影响。空⽓污染物的效⼒各不相同，⽽⽤于从空⽓污染物浓度转换为 AQI 的功能也因污染物⽽异。其空⽓质量指数值的每个范围都分配了一个空⽓质量级别和一个标准化的公共卫⽣建议。AQI 可能会因空⽓排放增加，例如在交通⾼峰时段、空⽓污染物缺乏稀释的情况下AQI会增加，而停滞的空⽓通常也会导致空⽓污染留在当地，导致污染物浓度⾼。

     就像天⽓预报让你知道当日应该穿什么样的衣服，是否需要带伞出门等等，使⽤ AQI 预测可以帮助你规划当日的户外活动，让你知道你什么时候应该减少户外活动或者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，⽐如先前患有肺部或⼼脏病的⼈、参与需要⼤量或长期户外活动或者⼯作的⼈、⽼年⼈和⼉童等可以提前预警。许多国家的机构或公共卫⽣组织也都会对 AQI 进⾏监控和预测。在空⽓质量⾮常差的时期，如空⽓污染事件，当 AQI 表明急性暴露可能会对公共卫⽣造成重⼤损害时，各机构可能会援引应急计划，允许他们命令主要排放来源（如燃煤⾏业）减少排放，直到危险条件缓解。

    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、最低问题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、数据来源与指标设计

本报告数据来源于中国天气网的可信数据来源抓取获得，希望通过某地区各类天⽓相关数据，分析影响空⽓质量指数的因素。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

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。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

表2：数据概要描述表

（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。

图表, 直方图

描述已自动生成

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。

（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。

图表, 直方图

描述已自动生成图2：自变量（最高气温、最低气温、风力）直方图

图表, 箱线图

描述已自动生成

图3：最高温度分布对应的AQI

图表, 箱线图

描述已自动生成

图4：最低温度分布对应的AQI

风⼒对 AQI 有明显的线性影响，如图 5，风⼒越⼤，AQI 越低。

图表, 箱线图

描述已自动生成

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。

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 | | |

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下表 4。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。

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 | | |

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。

五、模型应⽤

图形用户界面

描述已自动生成

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水平，帮助人们更好地理解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，具体实现是将可以将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。

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。