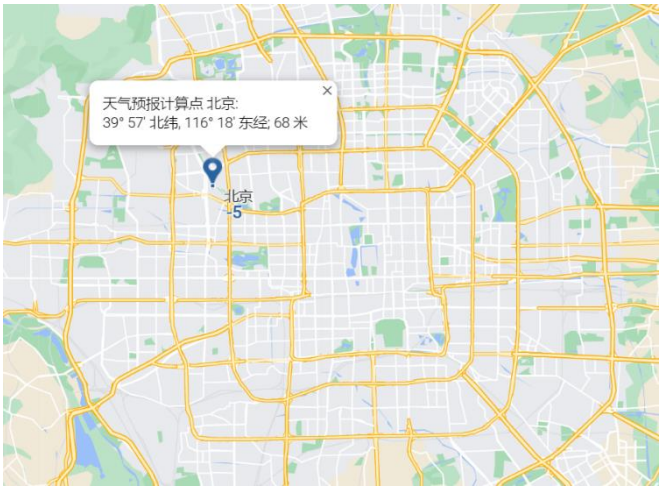
基于北京地区历年气候数据的统计规律

1　引言

气候统计是气象学中一个重要而复杂的研究领域，旨在深入了解和解释气象要素的时空分布规律。作为中国的首都，北京地区的气象状况对社会、经济和人类活动具有深远的影响。在这个独特而复杂的城市气象环境中，气温作为关键的气象要素之一，直接关系到居民的生活、农业、工业和城市规划。

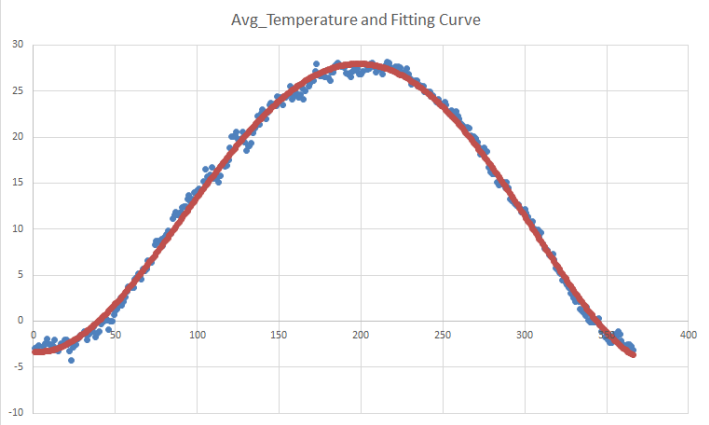
本研究聚焦于北京地区的气象分布特征，主要依据北京气象局（其地理位置如下图所示）对 2005 年到 2023 年的气象观测数据进行统计分析[1]，旨在揭示气温在时间和空间上的变化规律。在气象局给出的历年气象数据表中，包括有气温、空气质量指数（AQI）、气压风向风速等数据。北京地区的气象系统受到山脉、城市化和季风等多重因素的影响，使得该地区的气温表现出独特的动态特征。通过深入研究这些特征，我们可以更好地理解北京地区的气象规律，为应对气候变化、城市规划和生态环境保护提供科学依据。



2　气温的统计与预测

2.1　气温统计量的建立

根据我们对气温的深入分析，气温可被视为一个随机变量，其与年份、季节和昼夜紧密相联，呈现为季节性变化的表现，同时气温在一天内亦表现出显著的波动，成为后续内容的关键统计指标。设变量y表示年份，d表示一年中的日期，t表示一天中的时刻。因此，我们将温度变量定义为$T(y,d,t)$。鉴于温度的随机性，为了研究温度随日期的整体变化趋势，有必要进行平均处理。一个关键的统计量为$Ey[Et[T(y,d,t)]]$，其反映了温度整体变化的趋势。其中下标为t表示对t求期望，为y表示对y求期望。我们对表中数据进行统计分析，得到以下的分布图。



2.2　对气温的预测

基于前述对温度T的讨论，我们可将温度T重新表达为$T = (T - T_t) + (T_t - T_{t,y}) + T_{t,y}$。在此，我们使用简记符号，其中$T_{t,y} = \mathrm{E}y[\mathrm{E}t[T(y,d,t)]]$，$T_t = \mathrm{E}t[T(y,d,t)]$。符号$T_{t,y}$表示温度在一年内整体变化趋势的中心，$T_t - T_{t,y}$ 表示日内平均值相对于中心值的偏差，而$T - T_t$ 表示每一时刻相对于日内平均值的偏差。

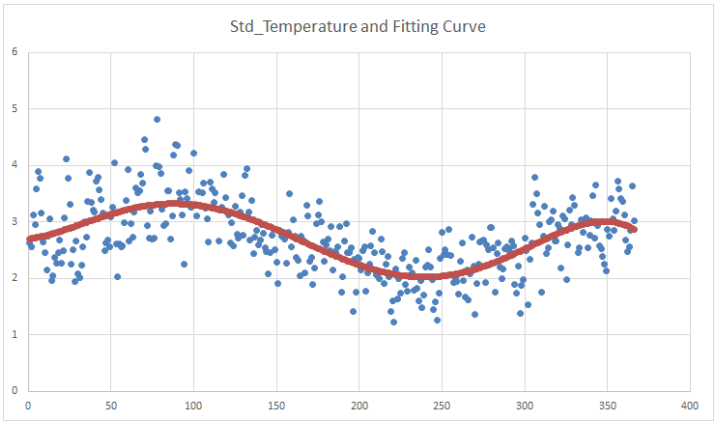
假设不同年份同一天的日平均温度服从正态分布，即$T_t \sim N(\mu,\sigma)$。我们的数据集跨度为 2006 年至 2023 年，涵盖了 17 个同一天的气象数据。以此数据为基础，我们能够近似估计T_t的分布。这也使我们有望推测未来某一天的温度，即提供温度的 95%置信区间。

在实际实施阶段，我们从数据集中抽取前 16 个数据进行分布估计。随后，利用数据集中的第 17 个数据进行验证。仅当正确预测的天数在 95%*365 的范围

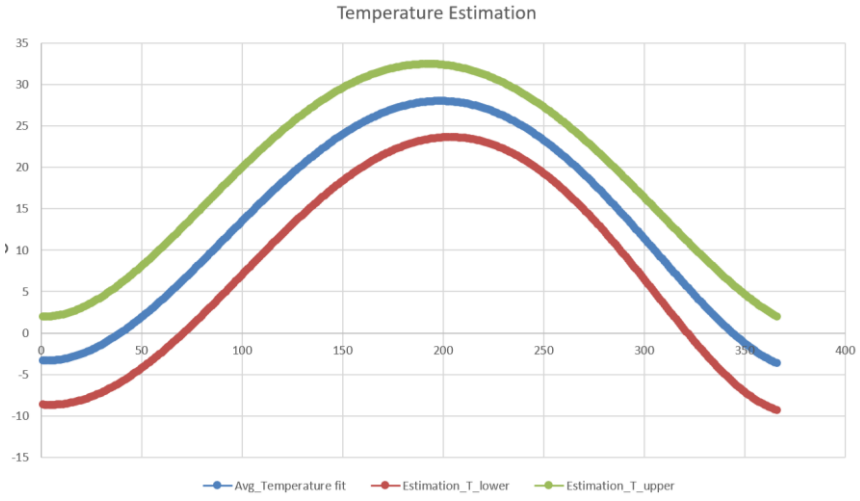
内时，方能证实该模型的有效性。这一验证过程旨在评估模型对未来数据的泛化能力，并通过对更广泛范围的观测结果的准确性验证，确保其在复杂现实环境中的可行性。

$$\mu=\sum_{i=1}^{i=16}\frac{x_i}{16}\ ,\quad {\sigma_{16}}^2=\frac{1}{15}\sum_{i=1}^{16}(x_i-\bar{x})^2$$

事实上${\sigma_{16}}^2$是基于 16 个数据点给出的方差的无偏估计。但是在实际操作过程中发现，${\sigma_{16}}^2$本身的分布的方差较大，因此需要对多日的${\sigma_{16}}^2$求平均以估计其期望，以此作为正态分布方差σ^2的最终估计，即$\sigma^2=\mathrm{E}({\sigma_{16}}^2)$。从而 95%置信区间可以表示为$[\mu-u_{0.025}\sigma,\mu+u_{0.025}\sigma]$。在下图中，蓝色的散点表示${\sigma_{16}}^2$的分布，红线表示平均值。



进一步地，最终的数据处理结果以图像呈现，其中红色曲线和绿色曲线分别表示置信下界和置信上界，而蓝色曲线则代表置信区间的中心。通过使用 2023 年的数据进行验证，我们观察到有 336 天落在该置信区间内，占比达到 92%。这一验证结果与 95%相近，因此我们可以近似认为该模型是准确的。这样的结果表明了模型对于新数据的预测具有相对高的可靠性和泛化能力。



2.3 正态分布的合理性

前文的论述中有一个重要的假设，即认为不同年份同一天的日平均温度服从正态分。下面采用假设检验的方法证明这一假设的合理性。

采用 Kolmogorov-Smirnov 假设检验方法（K-S）[2]，K-S 检验是一种非参数检验方法，用于确定两个样本是否来源于同一个分布，或一个样本是否来源于某个已知分布。这种检验特别适用于样本大小较小的情况。K-S 检验的基本原理是比较累积分布函数（CDF）。对于两个样本的情况（双样本 K-S 检验），它比较两个经验分布函数（EDF）的最大绝对差异。对于单样本 K-S 检验，它比较样本的 EDF 与理论 CDF 的最大差异。

本问题中该检验的原假设 H_0 是假定样本数据服从正态分布，而备择假设 H_1 假设样本数据不服从正态分布。

K-S 检验的具体做法如下：

1，获取样本集$\{X_1, X_2, \dots, X_n\}$和待检验的理论分布函数$F(x)$。

2，计算经验分布函数（EDF）$F_n(x)$ ，它是一个阶跃函数，对于给定的 x ，$F_n(x)$是样本中小于或等于 x的比例。即

$$F_n(x_i) = \frac{\text{样本中 } x_i \text{以下的数据点数}}{\text{总样本数}}$$

3，计算 K-S 统计量：$D_n = sup_x|F_n(x) - F(x)|$，即 EDF 和 CDF 之间的最大绝对差值，用于衡量两个分布之间的差异。

4，统计显著性通过计算D_n值的概率来确定。在大样本极限下，这个概率可以通过 Kolmogorov 分布近似。近似计算 P 值：

$$P \approx 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2D_n^2}$$

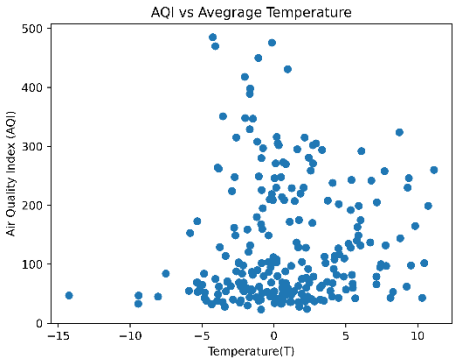
根据 K-S 价额检验的方法，使用 2023 年的数据可以计算出$Dn = 0.0636$，$P = 0.0995$。这里计算出来了的 P 值大于 0.05，因此没有理由拒绝温度服从正态分布的原假设。

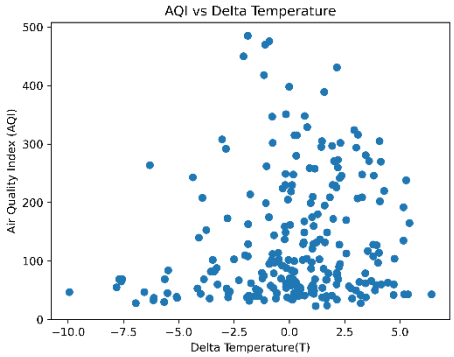
3　空气质量的相关性研究

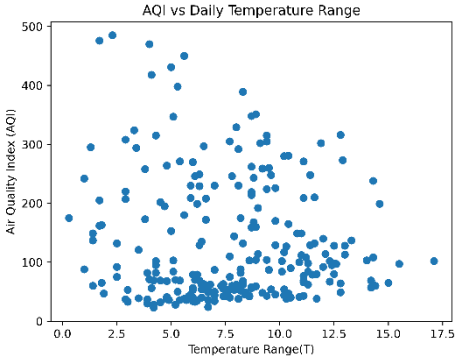
3.1　温度与空气质量的关系

最初的考虑是日均温度与空气质量指数（AQI）可能存在关联性。因此，我们绘制了以下左侧所示的散点图以探究二者之间的相关性。然而，观察发现并未呈现出明显的相关性。在进一步思考中，我们认识到日均温度本身受到总体温度分布中心的影响。因此，我们考虑消除这部分影响，即通过研究$T_t - T_{t,y}$和 AQI 之间的相关性。我们重新制作了中间所示的散点图，然而结果依旧未呈现出明显的相关性。最后尝试考虑一日之内的最大温差$\Delta = Tmin - Tmax$和 AQI 进行相关性分析，得到右侧的散点图，最终结果仍未呈现出明显的相关性

这一分析结果表明，日均温度与空气质量指数之间的关系可能受到更为复杂和多元的因素影响，超出了单一线性相关性的范畴。在后续的研究中，我们将进一步深入探讨可能影响这一关系的其他因素，以更全面地理解气象与空气质量之间的潜在关联。





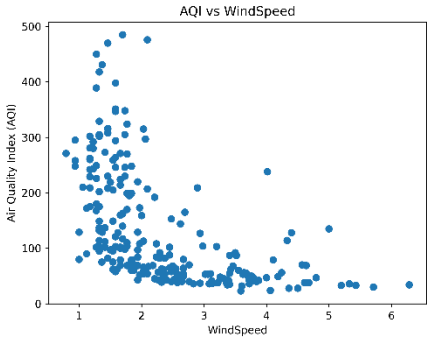


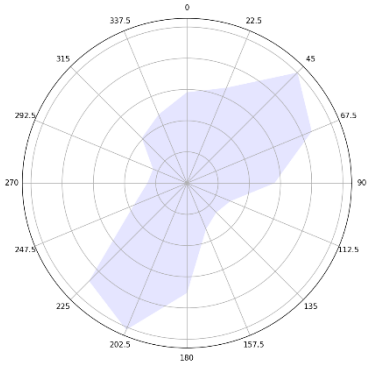
3.2　风速对空气质量的影响

风速的刻画包括风向和风速两个方面。若仅考虑风速大小对空气质量指数（AQI）的影响，我们通过对风速进行平方平均得到统计量$\overline{v^2}$。

$$\overline{v^2}_{(y,d)} = \sum_t v_{(y,d,t)}{}^2$$

基于该统计量，我们进一步考察其与 AQI 之间的相关性，并绘制了如下的散点图。图示结果表明二者之间存在一定的负相关关系，特别是在大风天气下，空气质量往往维持较低水平。



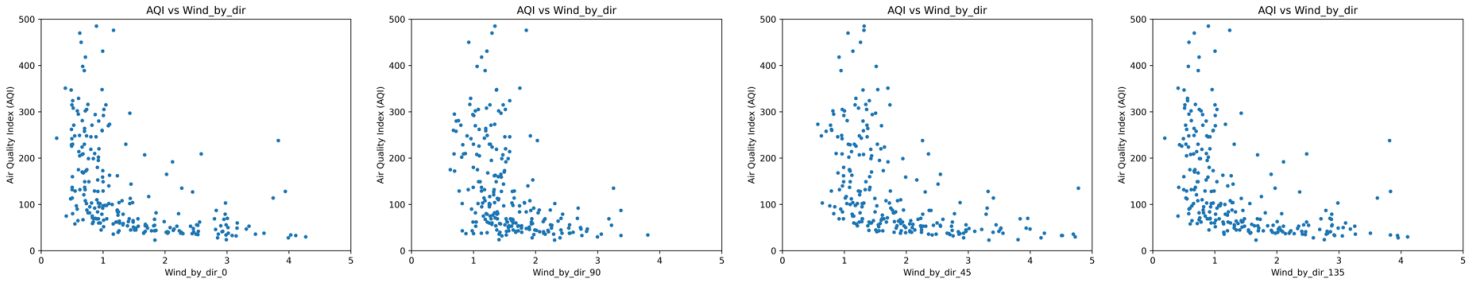


作为验证，在研究过程中，我们还绘制了风向的频数分布图。观察发现，风向频数在东北方向和西南方向相对较高。结合北京地区的地形特征，值得注意的是北京西北面有山脉，这一地理条件在该方向上抑制了气流的生成。因此，大部分风向主要集中在东北风和西南风。研究结果与实际情况相符，进一步印证了数据的合理性。

此外，通过对文献的详细查阅，我们了解到在某些情况下，风速和空气质量的关系体现为风速在特定方向投影的方均速率与空气质量的负相关。因此，我们考虑引入统计量 $\overline{v_{(\theta)}{}^2}$。

$$\overline{v_{(\theta,y,d)}{}^2} = \sum_t v_{(y,d,t)}{}^2 \cos(\theta_t - \theta)^2$$

在不同的 θ 值下，我们绘制了 AQI 与 $\sqrt{\overline{v_{(\theta)}{}^2}}$ 的关系图，然而，观察结果显示在不同角度下两者的分布均未呈现出明显的相关性。因此，我们得出结论，在北京地区，无法通过该理论进行风速和空气质量的有效分析。



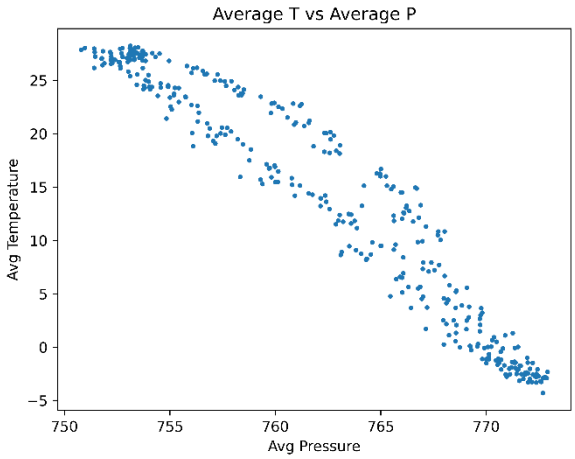
4 压强相关研究

4.1 温度和压强的关系

在相同地点，大气压强与温度之间应呈现显著的负相关关系。我们可以运用理想气体状态方程进行定性分析。地表气压实质上是地表垂直上方气体总质量所产生的效果。从全球范围来看，我们可以合理地假设不同区域的气压处于近似平衡状态。依据理想气体状态方程，温度升高导致气体密度降低，由此使得地表以上的气体总质量减小，进而引起压强的减小。

4.2 数据验证

与前文分析相似，我们分别提取同一天的温度和压强的平均值，分别记为$T_{y,t}$和$P_{y,t}$，并以它们为横纵坐标绘制图表。从下图可以明显看出，温度和压强之间存在显著的正相关性。实际结果与上述理论分析相符，进一步印证了其合理性。



5 总结

本论文基于概率论与数理统计的方法探究气象学领域的关键问题。首先，通过对气温的分布进行深入分析，我们揭示了其时间变化规律，并成功实现了对未来气温的预测。其次，对空气质量的研究考察了多个影响因素，重点关注了气温、风速和风向的作用。通过概率论与数理统计的手段，我们深入了解了这些要素对空气质量的综合影响，为空气质量预测提供了科学依据。最后，我们分析了温度和压强之间的关系，通过理论分析和实证研究，揭示了它们之间的密切联系。

未来的研究方向包括但不限于以下几个方面。首先，可以进一步挖掘气象学中其他重要气象要素的概率分布特征，扩展研究范围，提高模型的综合性能。其次，可以考虑引入更多影响空气质量的因素，如颗粒物浓度和湿度等，以建立更为全面的预测模型。此外，可以采用更加先进的统计模型，比如使用人工神经网络对更多的参数进行回归。综合而言，概率论与数理统计在气象学研究中的应用仍有广阔的发展空间，为更好地理解和应对气象变化提供更为精准的工具和方法。

参考文献

[1] Reliable Prognosis. 北京历史天气[EB/OL].(2024-1-13)[2024-1-17]. https://rp5.ru/%E5%8C%97%E4%BA%AC%E5%8E%86%E5%8F%B2%E5%A4%A9%E6%B0%94_

[2] Fasano, Giovanni, and Alberto Franceschini. "A multidimensional version of the Kolmogorov–Smirnov test." Monthly Notices of the Royal Astronomical Society 225.1 (1987): 155-170.