基于TSOLR和ISOLR模型的分类时间序列中复杂多重季节性的建模与预测

Anirban Ghosh和Raju Maiti*

印度统计研究所，203 B.T. Road，加尔各答700108，印度。

arXiv:2603.03954v1 [stat.ME] 4 Mar 2026

摘要

多重季节性已在连续时间序列中得到广泛研究，例如使用TBATS等模型进行电力需求预测。然而，其在分类时间序列（如空气质量指数（AQI）数据）中的处理仍然有限。分类AQI往往在多个频率上呈现明显的季节模式，而标准模型无法捕捉这些模式。本文提出一个框架，受TBATS方法启发，利用傅里叶级数和示性函数对多重季节性进行建模。该方法兼顾空气质量指数类别的有序性，同时显式捕捉周、月和年季节周期。模拟研究表明，在所提模型下参数估计具有经验一致性。我们进一步利用来自加尔各答的真实分类AQI数据说明其适用性，并将预测性能与马尔可夫模型和机器学习方法进行比较。结果表明，我们的方法能够有效捕捉复杂的季节动态，并提高预测准确度。所提方法为分析呈现多重季节模式的分类时间序列提供了一个灵活且可解释的框架，在空气质量监测、能源消耗及其他环境领域具有潜在应用。

关键词：空气质量指数；预测；分类时间序列；多重季节性；傅里叶级数；*TBATS*模型。

1 引言

季节性是指时间序列中以固定间隔呈现的系统性重复模式。对于连续值序列，如果存在一个占主导地位的重复周期，例如小时电力需求中的日模式或温度数据中的年度周期，就会出现单一季节性。然而，更复杂的时间序列通常具有多重季节性，即不同的季节周期以不同频率共存。电力消费就是一个典型例子，它可能同时呈现日、周和年度季节模式。TBATS（Livera 等，2011）及其变体（Makatjane 和 Mmelesi，2024；Thayyib 等，2023；Wang 等，2025）等模型，专门用于捕捉连续时间序列中的此类多重季节性结构。

当时间序列是分类数据时，例如将每日空气质量指数（AQI）数据分为优、满意、中等、差、很差和严重，季节性的定义和处理就不那么直接明了。在这种情况下，季节性可理解为类别概率分布在不同时间周期中的系统性变化。例如，在许多印度城市，

*通讯作者：Raju Maiti，印度统计研究所，印度加尔各答 B.T. Road 203号，邮编700108。电子邮箱：raju.isical@gmail.com

与季风季节相比，冬季出现不健康空气质量指数等级的概率更高。同样，在更精细的时间分辨率下，也可能观察到类别分布的日度或周度模式。由此进一步而言，分类时间序列中的多重季节性是指这类重复模式在多个时间尺度上并存，例如空气质量指数类别中日度、周度和年度周期的并存。

对分类时间序列中的季节性进行建模面临独特挑战，因为为连续型数据设计的标准方法无法直接应用。现有方法包括用于捕捉类别间转移概率的马尔可夫链模型、用于分类响应的广义线性模型（例如，加入季节性协变量的逻辑回归），以及较新的机器学习方法，如随机森林或循环神经网络。然而，这些方法往往无法明确考虑多个季节周期。将傅里叶级数展开或示性函数等灵活表示方式纳入分类模型，为捕捉多重季节性结构提供了一个有前景的方向，这一思路借鉴了 TBATS 等模型在连续数据领域取得的成功。

为应对这些挑战，我们提出了针对分类时间序列的 TBATS 框架扩展。具体而言，我们采用比例优势序数逻辑回归模型（Agresti，2002）来处理空气质量指数类别的有序特性，同时将傅里叶级数项作为协变量，以灵活表示日、周和年频率下的多种季节模式。这种设计保留了季节效应的可解释性，因为傅里叶系数直接反映了各个周期的贡献，同时确保与序数数据结构保持一致。

作为基于傅里叶的表示的替代方法，也可以使用示性函数对多重季节性进行建模，直接编码重复出现的时间段。例如，可以为一天中的每个小时、一周中的每一天或一年中的每个月分别引入虚拟变量，从而以完全非参数的方式捕捉季节效应。与利用正弦函数平滑近似季节周期的傅里叶级数不同，示性函数能够表示突变和不规则的季节形态，而这些现象在实践中十分常见，例如特定节庆日或冬季期间空气质量指数等级突然飙升。这种方法能更灵活地表示不同类别间的异质季节效应，但通常需要更多参数，若不加以谨慎的正则化，还可能导致过拟合。尽管如此，基于指示变量的建模为傅里叶展开提供了互补视角，尤其适用于季节性结构高度不规则或难以用平滑谐波项近似的情形。

本文其余部分安排如下。第2节全面回顾空气质量指数建模与预测技术方面的现有研究。第3节介绍数据集，并进行探索性分析以阐明建模方法的动机。第4节介绍并详细说明所提出的有序逻辑回归模型。第5节详细阐述所提模型的估计过程。第6节总结若干用于分类时间序列数据建模与预测的现有方法。第7节讨论如何根据各种自关联度量选择模型阶数。第8节报告评估所提模型估计精度与预测能力的模拟实验结果。第9节展示空气质量指数数据分析结果。最后，第10节总结全文，并提出未来研究方向。

2 文献综述

许多研究采用时间序列技术对空气质量进行建模，其中大多数将空气质量指数视为连续变量，并应用自回归积分滑动平均模型、向量自回归模型和状态空间方法等模型（V. Gupta et al., 2023; H. Liu et al., 2021; Y. Liu et al., 2024; Sethi and Mittal, 2020）。尽管这些方法适用于连续型数据，但出于公共卫生传播的需要，空气质量指数通常以有序类别的形式报告，因此连续值模型的结果不易直接解读。关于分类时间序列模型的研究仍相对有限（M. Liu et al., 2022; Weiß and Swidan, 2025）。经典

马尔可夫链模型已被广泛用于捕捉空气质量指数类别之间的转移，但随着状态空间增大，其参数数量会呈指数级增长。为缓解这一问题，研究者提出了多种扩展模型，例如离散ARMA（DARMA）过程（Jacobs and Lewis, 1978a; Jacobs and Lewis, 1978b, 1978c）、混合转移分布（MTD）模型（Berchtold and Raftery, 2002; Raftery, 1985）、Pegram算子型ARMA过程（Biswas and Song, 2009; Pegram, 1980）以及状态转换DARMA模型（Weiß, 2020）。然而，这些方法大多未能处理城市AQI序列的显著特征——多重季节性，也无法灵活纳入气象变量等外生驱动因素。

与此同时，随机森林、支持向量机和深度学习等机器学习方法在空气质量预测方面展现出强大的预测性能（例如，Desai等，2025；N. S. Gupta等，2023；Mishra and Gupta，2024）。然而，这些模型常因可解释性有限且缺乏对季节性结构的明确表示而受到批评。相比之下，有序逻辑回归（Agresti，2002；Fokianos and Kedem，2003）为有序分类结果建模提供了具有统计学依据且可解释的框架，同时能够自然地纳入温度、湿度、交通强度和工业活动等外部协变量。该方法的灵活性也允许纳入季节性成分，例如通过傅里叶展开或示性函数实现。尽管潜力可观，但使用有序回归模型对分类空气质量指数序列中的多重季节性进行显式建模，仍缺乏充分研究；在加尔各答这类人为因素与气候周期均对空气质量动态产生显著影响的城市环境中，尤其如此。

3 加尔各答每日空气质量指数数据

空气质量指数（AQI）是用于每日报告空气质量的指标。在印度，空气质量指数根据中央污染控制委员会（CPCB）确定的12种空气污染物浓度计算。对于每种污染物，CPCB都会根据其测量浓度计算一个子指数。总体空气质量指数取所有可用污染物各自子指数中的最大值。关于印度空气质量指数计算方法的详细说明，见《了解空气污染：指南》（2022年）。

3.1 研究区域

本研究聚焦于印度西孟加拉邦首府加尔各答的空气质量。加尔各答是印度人口第三多的大都市区，仅次于Delhi和Mumbai，总面积为206.08 km^2，人口约450万，人口密度约为每平方公里30,000人^2（https://en.wikipedia.org/wiki/Kolkata）。从地理位置上看，该市位于 $22°34′3″$N和 $88°22′12″$E。图1展示了加尔各答在印度境内的地理位置。作为该国主要的工业和经济中心之一，它为区域发展作出了重要贡献。在全国监测的12种污染物中，只有颗粒物 PM_{10}和 $PM_{2.5}$在加尔各答冬季的浓度超过了许可限值（例如，参见《了解空气污染：指南》（2022年））。

3.2 AQI数据集

本研究采用加尔各答多年期的每日分类AQI数据，数据自2019年起可用。2019—2024年的数据收集自中央污染控制委员会（CPCB）和西孟加拉邦污染控制委员会（WBPCB）运营的官方空气质量监测站。这些监测站每日记录污染物浓度。数据可通过国家空气质量指数（NAQI）门户网站的AQI数据资料库公开获取（https://airquality.cpcb.gov.

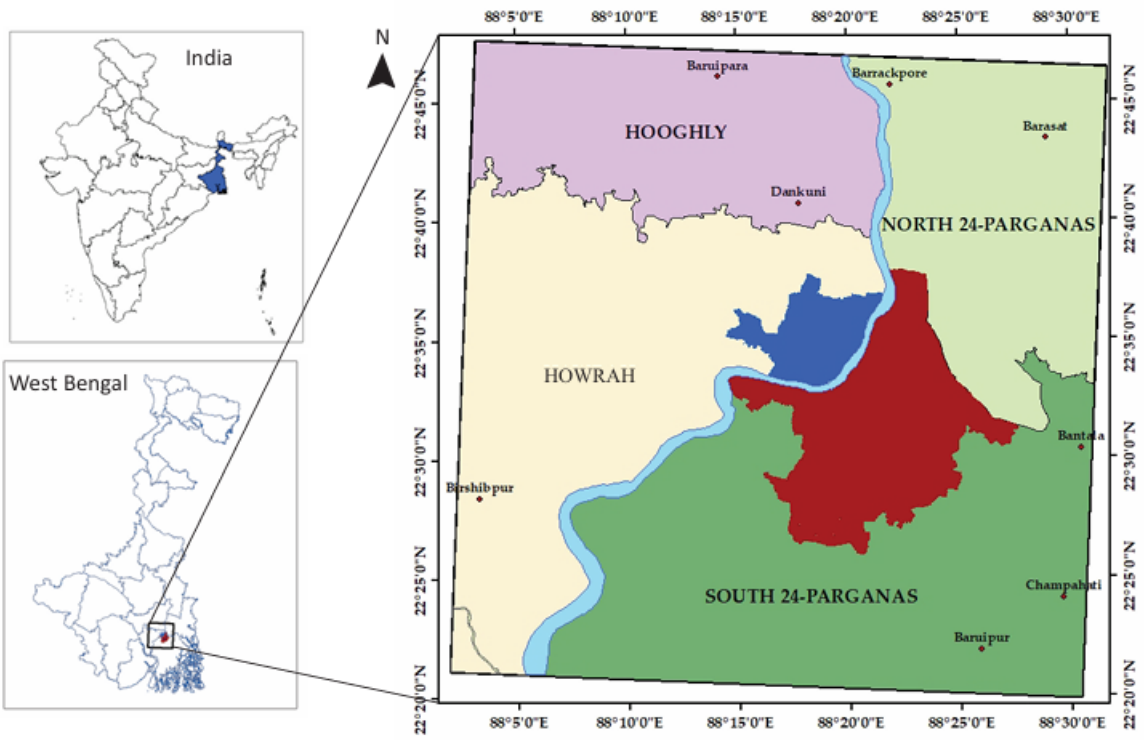


图1：印度加尔各答市（红色阴影区域）位置图（Dutta 等，2025）

in/ccr/#/caaqm-dashboard-all/caaqm-landing/aqi-repository），为时间序列分析和预测提供了可靠依据。

3.3 数据预处理与准备

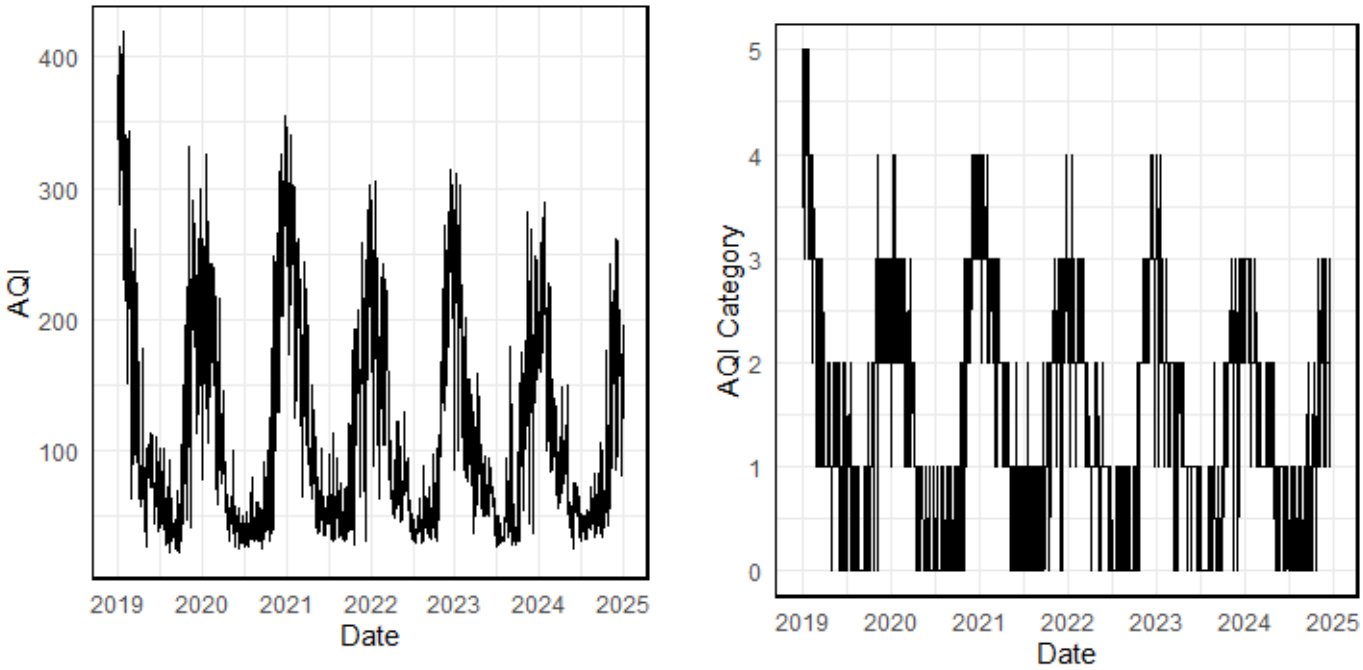
在拟合模型之前，首先对原始空气质量指数数据进行预处理，以处理缺失或不一致的条目。随后，按照国家空气质量指数分类方案（https://www.iitk.ac.in/new/national-air-quality-index），将最终的每日空气质量指数值映射为标准类别标签，如表1所示。接着，将这些类别标签编码为数值型有序形式，以便进行统计建模。2019年至2024年加尔各答的清理后可直接使用的每日空气质量指数数据，以及用于数据分析的相关 R 和 Python 代码，均可在作者的 GitHub 存储库中获取：https://github.com/ganirban004/aqi-logistic-paper.git。

表1：国家空气质量指数分类方案

24小时平均空气质量指数值	空气质量类别	Code
≤ 50	Good	0
51-100	满意	1
101-200	中等	2
201-300	Poor	3
301-400	很差	4
≥ 401	严重	5

3.4 探索性分析

通过探索性数据分析，了解加尔各答空气质量指数类别的分布和时间模式。图2a和图2b分别展示了原始数值型空气质量指数时间序列及其转换后的类别型时间序列。两种表示形式都呈现出全年明显的空气质量季节性模式。数值型空气质量指数呈现出明显的下降趋势，表明空气质量逐步改善。在类别型序列中（图2b），未出现“严重”类别；值得注意的是，最近两个冬季也没有“很差”等级的日子。这些观察结果共同表明，近年来加尔各答的空气质量显著改善。

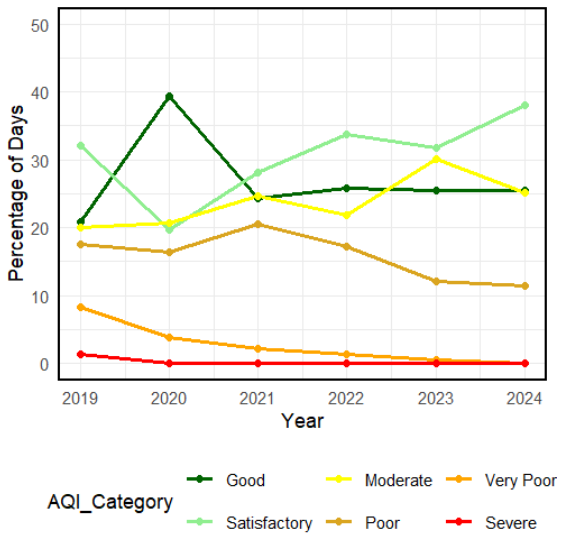


(a) 加尔各答2019—2024年的每日空气质量指数（空气质量指数）时间序列（数值越高表示空气质量越差）
(b) 加尔各答2019年至2024年的每日分类空气质量指数时间序列

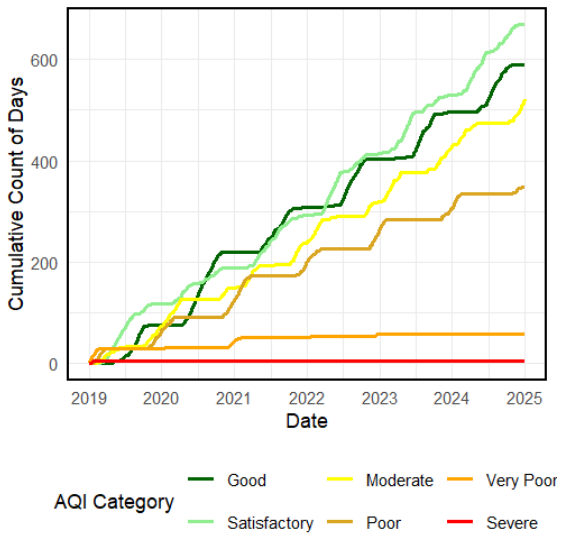
图2：加尔各答2019—2024年的每日空气质量指数时间序列（原始值和分类值），展示空气污染水平随时间的变化，并呈现明显的季节性和偶发性波动

图3a描绘了加尔各答空气质量随时间的变化，展示了2019年至2024年不同空气质量指数类别中每日占比的年度分布。研究时期内，“满意”天数的比例呈现明显上升趋势，表明空气质量逐步改善，但2020年除外。2020年新冠肺炎疫情封锁期间，“优”空气质量指数天数的比例大幅上升，这归因于工业和车辆活动减少。相比之下，“差”和“很差”类别的占比稳步下降；值得注意的是，如图2b所示，2019年后未再记录到“严重”空气质量指数天数。图3b展示了不同空气质量指数类别变化率随时间的演变（有关变化率演化图的更多信息，请参见 Weiß，2020）。该变化率演化图的斜率给出了相应类别的估计边际概率。近年来，“很差”和“严重”类别的曲线几乎趋于平坦。总体而言，这些发现也反映出近年来加尔各答空气质量显著且持续改善，与图2所示结果一致。

频数图显示　　图4a展示了整个期间各空气质量等级的占比　　t



（a）2019年至2024年加尔各答分类空气质量指数等级的年度趋势



(b) 2019—2024年加尔各答分类空气质量指数等级每日变化率演变图

图3：2019—2024年加尔各答分类空气质量指数等级的年度趋势及变化率演变图

研究时期（2019—2024年）。表4b所示的转移概率矩阵展示了研究时期内连续两天之间从一种空气质量类别转变为另一种类别的概率。对角线上的较高数值表明同一类别具有较强的时间持续性，而非零的非对角线元素则反映了空气质量状况发生短期变化的可能性。

3.5 使用强度度量的月份-类别热力图

为可视化分类AQI数据的季节模式，我们使用强度度量构建月份-类别热力图。设 $Y_t \in \{0,1,2,3,4,5\}$ 表示第 t天的AQI类别，编码为优$= 0$、满意 $= 1$、中等 $= 2$、差 $= 3$、很差 $= 4$和严重 $= 5$。设 $m \in \{1,\cdots,12\}$表示观测月份。对于每个月份 m 和AQI类别 j，我们将出现频次定义为

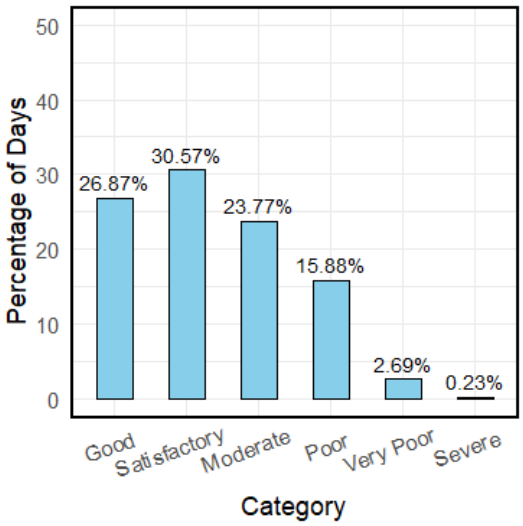
$$f_{m,c} = \sum_{t \in \text{Month } m} \mathbb{I}(Y_t = j),$$

其中， $\mathbb{I}(\cdot)$ 为指示函数。或者，也可以将 $f_{m,\ c}$ 除以月份 m 的总天数，再乘以 100，以此计算每个类别所占天数的百分比。

热力图以月份为横轴、空气质量指数类别为纵轴，每个单元格的颜色深浅表示相应的天数百分比。图5展示了这种可视化效果。颜色越深，表示百分比越高，能够直观呈现不同月份空气质量等级的分布情况。这种图形表示有效地突显了季节模式，例如冬季差和很差的空气质量日更为频繁，而整个季风季节则以中等或优的空气质量日为主。

3.6 空气质量指数数据的多重季节性

在此，我们通过类别比例、月份-类别热力图和恢复时间分析，探讨了数据在不同频率下的季节性。



(a) 整个研究时期（2019–2024）加尔各答空气质量指数类别的频率分布

(b) 整个研究时期（2019—2024年）空气质量类别的转移概率矩阵

Cat.	优	满意	Mod.	Poor	V.P.	Sev.
Good	0.80	0.19	0.00	0.00	0.00	0.00
Sat.	0.17	0.70	0.13	0.00	0.00	0.00
Mod.	0.01	0.16	0.68	0.15	0.00	0.00
Poor	0.00	0.01	0.22	0.70	0.07	0.00
V.P.	0.00	0.00	0.00	0.41	0.53	0.07
Sev.	0.00	0.00	0.00	0.00	0.80	0.20

图4：空气质量指数类别的频率分布，以及分类AQI时间序列的转移概率矩阵。

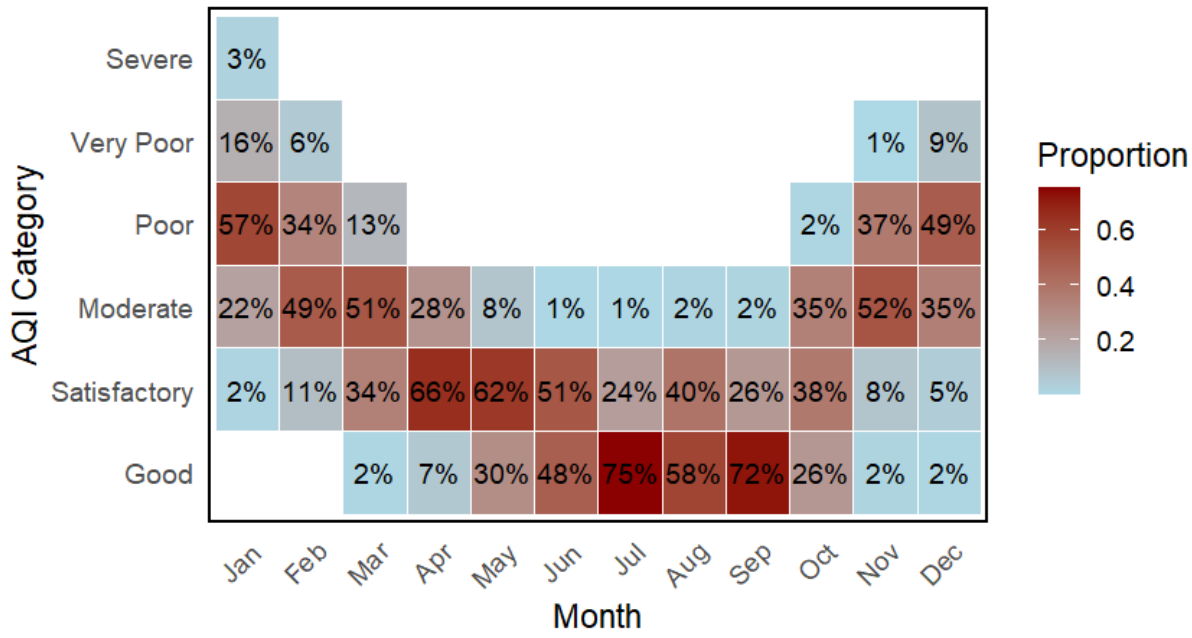


图5：使用强度度量的空气质量指数类别月份-类别热力图

3.6.1 季节模式：夏季 vs 季风季 vs 秋季 vs 冬季 r

图6所示的图表展示了夏季、季风季、秋季和冬季这四个季节的空气质量变化。较差的空气质量等级（例如“中等”、“差”和“很差”）在冬季和季风后季节更为常见，而较清洁的空气状况（例如“优”和“满意”）在季风季更为常见，凸显了季节对污染动态的影响。季风季节的空气质量通常较好，这可能是由于降雨的净化作用。这些反复出现的季节性波动表明，在对时间序列进行建模时，需要纳入季节性成分或虚拟变量。

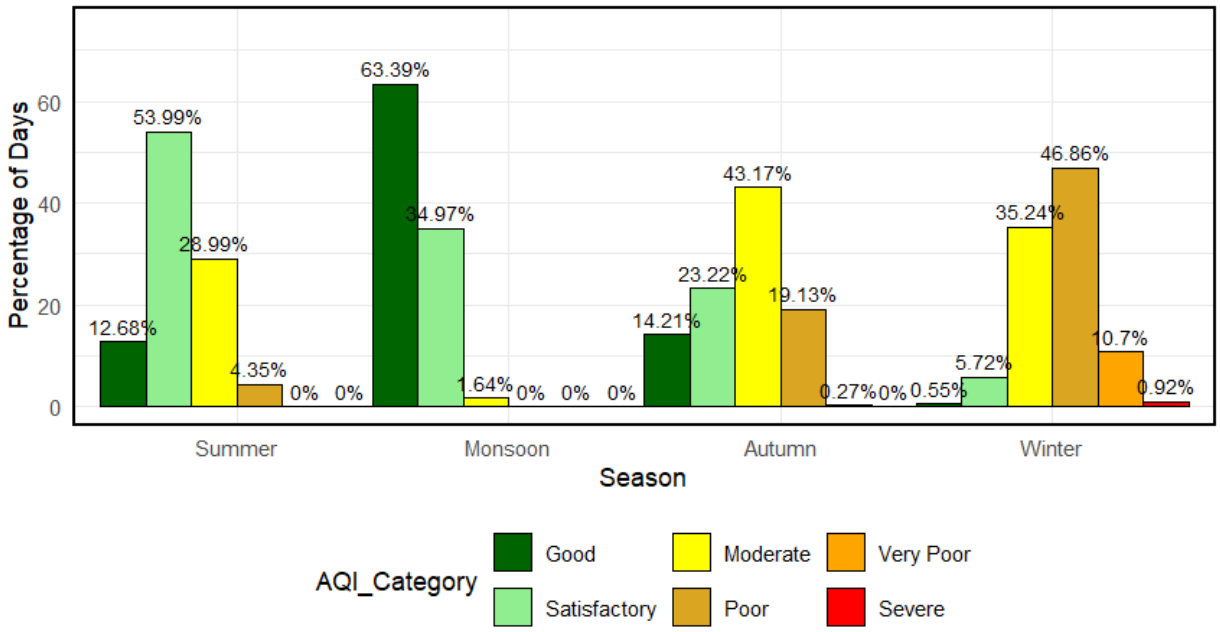


图6：季　　节分布：加尔各答2019年至2024年分类空气质量指数等级

3.6.2 季节模式：工作日 vs 周末

图7展示了工作日和周末的空气质量指数类别分布，其中存在细微但有意义的差异。周末属于“满意”类别的天数比例较高（32.75%），高于工作日（29.69%），这表明周末的空气状况略好。与此同时，工作日“优”类别的天数比例（27.59%）略高于周末（25.08%）。“中等”和“差”类别在两类日期中的出现频率相对一致，而“很差”和“严重”等危害最大的类别出现频率较低，且工作日的出现次数略多。这些发现表明，尽管一周内的空气质量总体相近，但周末的空气状况略有改善，这可能是由于交通和工业活动减少所致。

3.6.3 季节性模式：节庆与非节庆

加尔各答的空气污染呈现显著的季节性变化，在排灯节期间明显加剧。本研究考察排灯节对该市空气质量的具体影响。为进行分析，我们将研究时期划分为以下三个连续阶段：

- **排灯节前时期（第 −22 天至第 −8天）：** 此期间指排灯节前的15天，

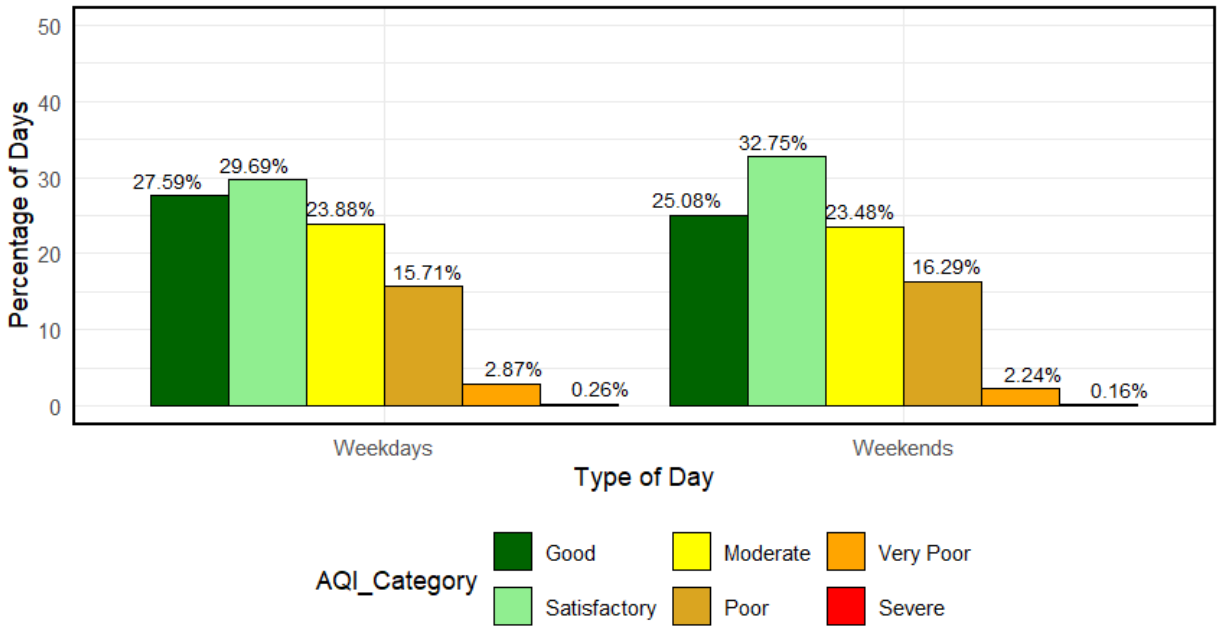


图7：加尔各答工作日和周末空气质量指数类别的频率分布（2019年—2024年）

也就是排灯节前22天至排灯节前8天。

$$\text{Pre-Diwali Period} = [\text{Diwali} - 22 \text{ days},\ \text{Diwali} - 8 \text{ days}]$$

- **排灯节期间（第 −7 天至第 +7天）：** 这是一个以排灯节为中心的15天对称窗口，包括排灯节前7天和排灯节后7天。

$$\text{排灯节期间} = [\text{排灯节} - 7\text{天},\ \text{排灯节} + 7\text{天}]$$

- **排灯节后期间（第 +8 天至第 +22天）：** 此期间包括排灯节窗口之后的15天，即排灯节后第8天至第22天。

$$\text{Post-Diwali Period} = [\text{Diwali} + 8 \text{ days},\ \text{Diwali} + 22 \text{ days}]$$

这种连续的结构有助于系统分析排灯节前、期间及之后的空气质量趋势，并区分烟花爆竹排放和季节变化的影响。图8所示的图表突出显示，在排灯节期间和排灯节后期间，“中等”“差”和“很差”空气质量日的频率明显增加，表明节庆相关排放导致空气质量在短期内显著恶化。

总之，分类时间序列分析清楚地表明，加尔各答的空气质量指数类别存在多种季节性。因此，恰当地建模这些季节性成分，对于提高污染评估的准确率以及支持有效的公共卫生和缓解策略至关重要。

4 所提模型

针对澳大利亚的电力消费数据，Livera 等，2011提出了 TBATS 模型，用以捕捉多重季节性和趋势。TBATS 模型旨在处理复杂的

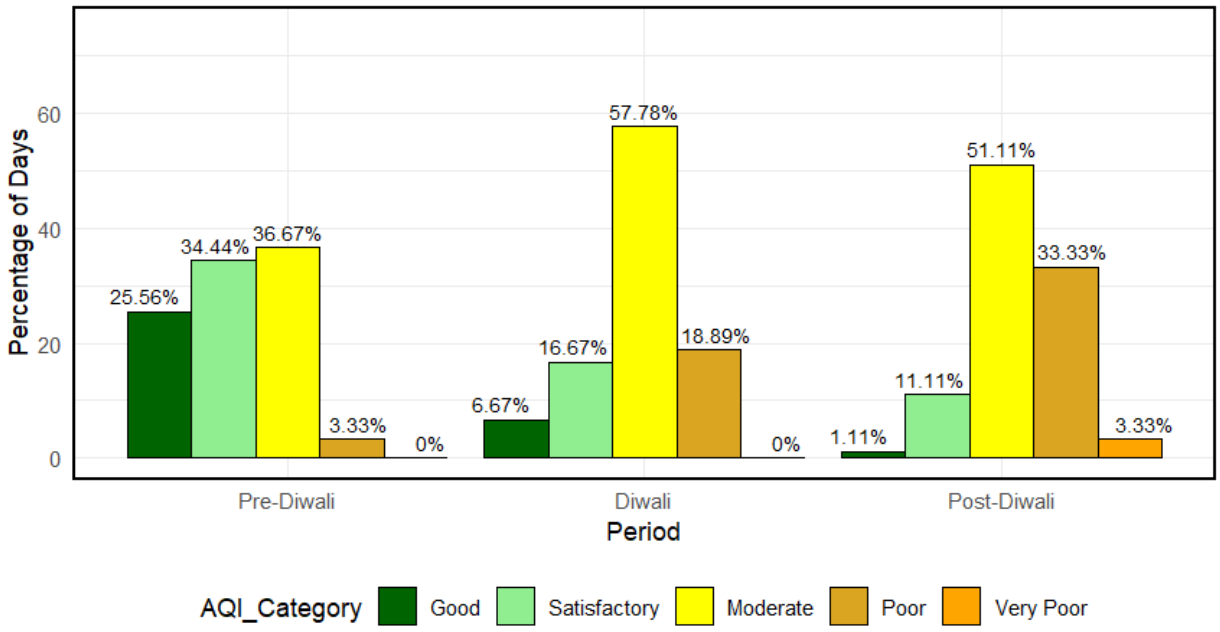


图8：加尔各答排灯节前、排灯节期间和排灯节后期间空气质量指数类别的频率分布（2019—2024年）

具有多重季节性、非线性变换和短期依赖的时间序列。TBATS 是三角函数季节性、Box–Cox 变换、ARMA 误差、趋势和季节性成分的缩写。设 y_t 为时间序列在时间点 t的观测值。首先对原始时间序列 y_t 应用 Box-Cox 变换，得到

$$y_t^{(\omega)} = \begin{cases} \dfrac{y_t^{\omega} - 1}{\omega} & \text{for } \omega \neq 0 \\ \log y_t & \text{for } \omega = 0 \end{cases}$$

其中 ω 是 变换参数。随后，将变换后的序列建模为

$$y_t^{(\omega)} = l_t + \phi b_t + \sum_{i=1}^{T} s_{i,t-P_i} + \varepsilon_t,$$

其中， l_t 是局部水平， b_t 是趋势分量，阻尼参数为 $\phi \in (0, 1)$， $s_{i,\ t}$ 是季节性成分， T 是季节性成分的数量， ε_t 服从 ARMA(p, q) 过程。每个季节性成分均使用基于傅里叶展开的三角函数项表示。对于季节周期 P_i，其表示形式为

$$s_{i,t} = \sum_{k=1}^{K_i} \left[\beta_{1i}^{(k)} \cos\left(\frac{2\pi k t}{P_i}\right) + \beta_{2i}^{(k)} \sin\left(\frac{2\pi k t}{P_i}\right) \right]$$

其中， K_i 是为季节周期 i选定的谐波数量。状态向量 $\beta_{1i}^{(k)}$ 和 $\beta_{2i}^{(k)}$ 根据递归三角函数更新动态演变。TBATS模型的原始形式能够有效捕捉连续值时间序列中的多重季节性、非线性趋势和短期依赖，例如澳大利亚的电力需求（Livera等，2011）。通过使用用于三角函数季节性的傅里叶项、Box–Cox变换和ARMA误差结构，TBATS可以同时对多个时间尺度上的复杂周期模式进行建模。然而，该模型无法直接应用于分类时间序列，因为TBATS假定观测值为连续值且误差为加性，而分类空气质量指数数据则是离散且有序的。

为弥合这一差距，我们提出将TBATS理念扩展至有序时间序列。具体而言，我们采用比例优势模型（Agresti，2002）处理有序响应，并将傅里叶级数项作为协变量，以捕捉空气质量指数数据中隐含的多重季节模式。因此，我们将这一框架称为有序逻辑回归中的三角季节性（TSOLR）模型。该方法保留了TBATS的核心优势，即显式建模多个季节周期，同时兼顾空气质量指数类别离散且有序的特性。

令 $Y_t \in \{0, 1, \cdots, m\}$ 表示第 t 天的空气质量指数类别。我们假设底层潜在过程受到多重季节性的影响，这与 TBATS 框架类似，但适用于有序回归情形。我们采用比例优势模型（Agresti, 2002），其中在时间 t 处属于类别 j 或以下类别的累积概率为

$$\operatorname{logit}\Big(P(Y_t \le j \mid \mathcal{F}_{t-1}, \boldsymbol{X_t})\Big) = \theta_j - \eta_t, \quad j = 0, 1, \cdots, m-1 \tag{1}$$

其中，$\mathcal{F}_t$包含截至时间点 t 的所有过去信息，即 $\mathcal{F}_t = \{(y_{t'}, \boldsymbol{X_{t'}})\ t' = 1, 2, \cdots, t\}$；$\theta_j$ 为类别特定阈值参数，η_t 为线性预测器。具体而言，为了将多重季节性作为协变量通过傅里叶项来捕捉，η_t 的形式为

$$\eta_t = \sum_{i=1}^{T} \sum_{k=1}^{K_i} \left[\beta_{1i}^{(k)} \cos\left(\frac{2\pi k t}{P_i}\right) + \beta_{2i}^{(k)} \sin\left(\frac{2\pi k t}{P_i}\right)\right] + \boldsymbol{X_t}^{\top}\boldsymbol{\gamma},$$

其中，P_i 是第 i^{th} 个季节性成分的季节周期（例如，在日数据中，$P_i = 7$ 表示周季节性，$P_i = 365$表示年度季节性等），K_i 是针对特定季节周期 i 选取的傅里叶谐波数。为避免过拟合并保持模型简单且易于解释，我们假设 $K_i = 1$，这样通过单个正弦—余弦曲线即可很好地捕捉季节性。β_{1i} 和 β_{2i}（由于 $K_i = 1$，省略上标（k））是对应于 i^{th} 季节性成分的傅里叶系数。项 $\boldsymbol{X_t}$ 包含额外的外生协变量，$\boldsymbol{\gamma}$ 为相应的参数向量。于是，TSOLR模型的形式为

$$\operatorname{logit}\Big(P(Y_t \le j \mid \mathcal{F}_{t-1}, \boldsymbol{X_t})\Big) = \theta_j - \sum_{i=1}^{T} \left[\beta_{1i} \cos\left(\frac{2\pi t}{P_i}\right) + \beta_{2i} \sin\left(\frac{2\pi t}{P_i}\right)\right] + \boldsymbol{X_t}^{\top}\boldsymbol{\gamma} \tag{2}$$

这种表述使有序空气质量指数模型能够沿用 TBATS 明确建模多个季节性成分的理念，同时仍适用于离散、有序的结果。还可以在 η_t 中加入滞后类别指标或自回归项，以纳入短期依赖，这与TBATS模型中的 ARMA误差类似。此外，也可以在模型中加入诸如 $|\sin\left(\frac{2\pi t}{P_i}\right)|$ and $|\cos\left(\frac{2\pi t}{P_i}\right)|$ 这样的项，以获得非负季节效应。

另一种经典方法是通过指示变量来捕捉多重季节性；与傅里叶级数相比，指示变量是离散函数。我们将这一框架称为有序逻辑回归中的指示变量季节性（ISOLR）模型。对于每日空气质量指数数据，存在三种季节性：每周季节性、季节性以及节日季节性（排灯节与非排灯节）。因此，模型（1）中的线性预测器η_t 具有以下形式

$$\eta_t = \beta_1 \cdot Summer_t + \beta_2 \cdot Monsoon_t + \beta_3 \cdot Winter_t + \beta_4 \cdot Diwali_t + \beta_5 \cdot Weekend_t + \beta_6 \cdot Y_{t-1}$$

其中，若第 t 天分别处于夏季、季风季和冬季，则 $Summer_t$、$Monsoon_t$ 和 $Winter_t$ 取值为1，否则为0。秋季被视为基准类别。同样，若第 t 天处于15天排灯节期间，则 $Diwali_t$ 取值为1；若

第 t 天为星期六或星期日，否则为0。 Y_{t-1} 是第 t天的滞后1期值。于是ISOLR模型变为

$$\log\left(\frac{P(Y_t \le j)}{P(Y_t > j)}\right) = \theta_j - \Big(\beta_1 \cdot Summer_t + \beta_2 \cdot Monsoon_t + \beta_3 \cdot Winter_t + \beta_4 \cdot Diwali_t + \beta_5 \cdot Weekend_t + \beta_6 \cdot Y_{t-1}\Big) \tag{3}$$

5 参数估计

为了估计所提模型的参数，通常的做法是最大化对数似然函数

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}) = \log \prod_{t=1}^{N} \prod_{j=0}^{m} P(Y_t = j)^{\mathbb{I}(Y_t = j)} = \sum_{t=1}^{N} \sum_{j=0}^{m} \mathbb{I}(Y_t = j) \log P(Y_t = j)$$

其中

$$P(Y_t = j) = \begin{cases} P(Y_t \le j) & \text{for } j = 0 \\ P(Y_t \le j) - P(Y_t \le j-1) & \text{for } j = 1, 2, \cdots, m-1 \\ 1 - P(Y_t \le m-1) & \text{for } j = m \end{cases}$$

and

$$P(Y_t \le j) = \begin{cases} \dfrac{\exp(\theta_j - \eta_t)}{1 + \sum_{i=0}^{m-1} \exp(\theta_i - \eta_t)} & \text{if } j = 0, 1, \cdots, m-1 \\ \dfrac{1}{1 + \sum_{i=0}^{m-1} \exp(\theta_i - \eta_t)} & \text{if } j = m \end{cases}$$

如模型（1）所述，关于 $\boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_{m-1})$ 和 $\boldsymbol{\beta}$。在满足某些正则条件的情况下，最大似然估计量渐近正态这一事实已得到充分确立，其形式为

$$\begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} \overset{a}{\sim} \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{\beta} \end{bmatrix}, \mathcal{I}^{-1}(\boldsymbol{\theta}, \boldsymbol{\beta})\right).$$

观测费希尔信息矩阵 $\mathcal{I}(\theta)$ 通过对数似然函数在最大似然估计处的海森矩阵取负得到：

$$\hat{\mathcal{I}}(\boldsymbol{\theta}, \boldsymbol{\beta}) = -\nabla^2 \ell(\boldsymbol{\theta}, \boldsymbol{\beta})\big|_{(\boldsymbol{\theta}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})}$$

该海森矩阵通过数值方法计算，并作为估计参数的方差-协方差矩阵，随后可用于构造估计参数的置信区间。对于参数 $(\boldsymbol{\theta}, \boldsymbol{\beta})$，使用对数似然函数 $\ell(\boldsymbol{\theta}, \boldsymbol{\beta})$ 的有限差分来近似偏导数。计算有限差分时的步长经过选择，以平衡数值精度和稳定性。R 中 `MASS` 软件包的 `polr()` 函数实现了比例优势累积逻辑回归模型。

6 现有模型

为了比较 TSOLR 和 ISOLR 模型的预测性能，本节列举了一些用于建模分类时间序列的现有方法，包括一些基于马尔可夫链的模型，以及一些机器学习和深度学习模型。

6.1 马尔可夫链模型

建模分类时间序列的一种传统方法是使用 p阶马尔可夫链，其中转移概率定义为：

$$P(Y_t = j \mid Y_{t-1}, \cdots, Y_{t-p}), \quad j = 0, 1, \cdots, m.$$

然而，这种方法会导致参数数量呈指数级增长。对于一个阶数为 p、包含 $(m+1)$ 个类别的马尔可夫链，参数数量为 $m(m+1)^p$。例如，当 $m = 4$ 且 $p = 2$时，需要 100 个参数，计算量很大。此外，该模型无法纳入外部协变量。

6.2 混合转移分布（MTD）模型

为降低复杂度，Raftery 于 1985 年提出了 MTD 模型，随后 Berchtold and Raftery 于 2002 年对其进行了改进。该模型将转移概率表示为一步转移的加权混合：

$$P(Y_t \mid Y_{t-1}, \cdots, Y_{t-p}) = \sum_{k=1}^{p} \lambda_k P(Y_t \mid Y_{t-k}),$$

其中 $\lambda_k \geq 0$ 且 $\sum \lambda_k = 1$。MTD 模型将参数数量减少至 $m(m+1) + p - 1$。例如，当 $m = 4$ 且 $p = 2$时，参数数量降至 21。与马尔可夫链模型一样，这种方法也不考虑任何外部协变量的信息。

6.3 Pegram 自回归（PAR）过程

Pegram 于 1980 年提出了一种称为 PAR(p) 过程的简化方法，它是离散ARMA模型的一个特例（Jacobs and Lewis, 1978a；Jacobs and Lewis, 1978b, 1978c）。PAR过程定义如下：

$$P(Y_t = j \mid Y_{t-1}, \cdots, Y_{t-p}) = \sum_{k=1}^{p} \phi_k \mathbb{I}(Y_{t-k} = j) + \left(1 - \sum_{k=1}^{p} \phi_k\right) \pi_j,$$

其中，$\phi_k \in (0, 1)$，且 $\sum_{k=1}^{p} \phi_k < 1$, π_j 是类别 j 的边际概率，$\mathbb{I}(\cdot)$是指示函数。然而，该模型未纳入外部协变量。

6.4 用于有序数据的决策树和随机森林 a

标准的基于树的方法可以通过采用二次熵指数等有序分裂准则，或施加单调约束，轻松加以调整（Janitza et al., 2016）。随机森林（RF）是一种基于集成的学习方法，通过组合在数据和特征的随机子集上训练的多棵树，扩展了传统的决策树模型。令 $\mathcal{D} = \{(\boldsymbol{X_t}, Y_t)\}_{t=1}^{N}$ 表示训练数据集，其中 $\boldsymbol{X_t} = (Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p}, Z_{1t}, Z_{2t}, \cdots, Z_{qt})^\top$ 表示预测变量向量，其中包含 p 个滞后值和 q 个外部变量。为与 ISOLR 和 TSOLR 模型进行公平比较，采用两种备选的季节性表示（如指示变量和三角函数项）纳入外部协变量，并分别进行评估。对于每个 K 自助法样本 $\mathcal{D}_b^{(k)}$，均拟合一棵分类树 $h_k(X)$，在每个分裂节点使用随机选取的预测变量子集，以降低树与树之间的相关性。最终的随机森林预测通过多数投票得出：

$$\hat{y} = \arg\max_{j \in \mathcal{C}} \sum_{k=1}^{K} \mathbb{I}\big(h_k(X) = j\big),$$

其中 $\mathcal{C} = \{0, 1, 2, \cdots, m\}$ 是空气质量指数类别的集合，$\mathbb{I}(\cdot)$ 是指示函数 $\mathrm{n}.$

随机森林以能够抵御过拟合、建模非线性和复杂关系的能力，以及即使面对高维分类特征也能保持出色的预测性能而著称。然而，其集成特性降低了可解释性并增加了计算成本，给时变类别数据的理论推断和因果分析带来了挑战。

6.5 支持向量机（SVM）

支持向量机（SVM）是广泛应用于分类问题的强大监督学习算法（Awad和Khanna，2015）。对于分类时间序列 $\{Y_t\}$，时刻 t 的预测变量向量可能同时包含以整数表示的滞后响应和外生协变量，即 $\boldsymbol{X_t} = (Y_{t-1}, \cdots, Y_{t-p}, Z_{1t}, \cdots, Z_{qt})^\top$。与决策树框架一样，我们考虑协变量的两种备选形式：一种使用季节指示变量，另一种使用三角函数项，并在每种设定下分别拟合SVM模型。给定训练数据 $\mathcal{D} = \{(\boldsymbol{X_t}, Y_t)\}_{t=1}^N$，其中 $Y_t \in \{-1, +1\}$，标准SVM通过求解以下问题来寻找分离超平面 $\boldsymbol{w}^\top \boldsymbol{X} + b = 0$

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{t=1}^{N} \xi_t, \quad \text{subject to} \quad Y_t(\boldsymbol{w}^\top \boldsymbol{X_t} + b) \geq 1 - \xi_t,\ \xi_t \geq 0,$$

其中，$C > 0$ 用于调节间隔最大化与误分类之间的平衡。对于有序取值的 Y_t （例如空气质量指数类别），有序支持向量机通过引入一系列有序的平行超平面来推广这一框架，

$$\boldsymbol{w}^\top \boldsymbol{X_t} - b_1 = 0,\ \boldsymbol{w}^\top \boldsymbol{X_t} - b_2 = 0,\ \ldots,\ \boldsymbol{w}^\top \boldsymbol{X_t} - b_{m-1} = 0,$$

确保预测类别遵循其自然顺序。通过将滞后值和外部变量都纳入 $\boldsymbol{X_t}$，有序支持向量机能够捕捉短期时间动态，以及协变量驱动的类别时间序列预测效应。

SVM在高维特征空间中表现出很强的泛化能力，并且能够有效避免过拟合。不过，其性能很大程度上取决于核函数和正则化参数的选择。此外，计算复杂度会随样本量增加；而且，SVM无法自然地捕捉固定滞后结构以外的序列动态，因此对长程时间依赖的建模能力有限。

6.6 循环神经网络（RNN）

循环神经网络（RNN）及其先进变体，例如长短期记忆网络（LSTM）和门控循环单元（GRU）架构，广泛用于对序列数据中的时间依赖性进行建模（Lindemann等人，2021）。与传统的前馈网络不同，RNN包含反馈连接，使信息能够在多个时间步之间持续传递。在这种设定下，时刻 t 的预测变量序列可能同时包含以整数表示的滞后类别指标和外生变量，即 $\boldsymbol{X_t} = (Y_{t-1}, \cdots, Y_{t-p}, Z_{1t}, \cdots, Z_{qt})^\top$。给定输入序列$\{\boldsymbol{X_t}\}_{t=1}^N$，标准RNN按照以下方式更新其隐藏状态:

$$\boldsymbol{h}_t = \sigma\left(W_h \boldsymbol{h}_{t-1} + W_x \boldsymbol{X_t} + \boldsymbol{b_h}\right),$$

其中，$\boldsymbol{h_t}$ 表示时间 t的隐藏表示，$\sigma(\cdot)$是非线性激活函数，而 W_h、W_x、$\boldsymbol{b_h}$ 是模型参数。长短期记忆网络通过引入门控机制改进了这一框架

这些门控机制用于调节信息流并缓解梯度消失问题。具体而言，对于每个时间步 t，长短期记忆网络会计算

$$
\begin{aligned}
\boldsymbol{f_t} &= \sigma(W_f \boldsymbol{X_t} + U_f \boldsymbol{h_{t-1}} + \boldsymbol{b_f}),\\
\boldsymbol{i_t} &= \sigma(W_i \boldsymbol{X_t} + U_i \boldsymbol{h_{t-1}} + \boldsymbol{b_i}),\\
\tilde{\boldsymbol{c_t}} &= \tanh(W_c \boldsymbol{X_t} + U_c \boldsymbol{h_{t-1}} + \boldsymbol{b_c}),\\
\boldsymbol{c_t} &= \boldsymbol{f_t} \odot \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \odot \tilde{\boldsymbol{c_t}},\\
\boldsymbol{o_t} &= \sigma(W_o \boldsymbol{X_t} + U_o \boldsymbol{h_{t-1}} + \boldsymbol{b_o}),\\
\boldsymbol{h_t} &= \boldsymbol{o_t} \odot \tanh(\boldsymbol{c_t}),
\end{aligned}
$$

其中，$\boldsymbol{f_t}$、$\boldsymbol{i_t}$ 和 $\boldsymbol{o_t}$ 分别表示遗忘门、输入门和输出门，$\odot$ 表示逐元素乘法，$\boldsymbol{c_t}$ 表示记忆单元状态。与决策树和支持向量机一样，长短期记忆网络也使用两类协变量进行拟合，即指示变量和作为季节性成分的三角函数项。

对于有序分类输出，可以用有序感知损失函数（如有序交叉熵或基于累积链接的目标函数）替换或修改 Softmax 输出层，以保留类别之间固有的顺序。这类架构能够有效捕捉非线性动态和长期时间依赖，因此非常适合预测空气质量指数类别随时间的变化。然而，这些模型需要进行大量超参数调优，才能兼顾模型复杂度与泛化能力。由于存在序列依赖，训练过程可能需要大量计算；与其他深度学习方法一样，尽管这类模型具有很强的预测能力，其可解释性仍然是一项挑战。

这些机器学习方法为有序分类时间序列建模提供了灵活的框架，可根据数据特征、可解释性要求和计算方面的考量进行选择。

7 自关联 度量分类时间序列的序列依赖

自相关函数（ACF）是衡量数值时间序列中序列依赖的标准工具。ACF图还被严格用于确定模型的适当阶数。然而，由于缺乏自然的算术结构，它不适用于分类数据。为了研究分类时间序列中的依赖关系，尤其是在数据为名义型或有序时，人们提出了几种类似ACF的度量方法。

设 $Y_t \in \{0, 1, \cdots, m\}$ 为分类时间序列。定义滞后 h 下的联合概率为 $p_{ij}^{(h)}$ =P(Y$_t = i$， $Y_{t-h} = j$)，边际概率为 $p_{i\cdot}^{(h)}$ = P(Y$_{t-h} = i$) = $\sum_{j=0}^{m} p_{ij}^{(h)}$ 和 $p_{\cdot j}^{(h)}$ = P(Y$_{t-h}$ =j) = $\sum_{i=0}^{m} p_{ij}^{(h)}$ 。在此设定下，Weiß和G¨ob（2008）总结了各种自关联度量，这些度量列于表2中。在本文分析中，计算这些自关联度量不仅是为了评估分类时间序列中序列依赖的强度，也是为了指导后续时间序列建模中适当模型阶数的选择。

8 模拟研究

在本节中，我们开展大量模拟研究，以检验TSOLR模型的一致性与预测性能。数据分别由ISOLR模型和TSOLR模型在多种参数配置下生成，具体见后续各节。估计值通过第5节所述的最大似然估计（MLE）获得。

我们考虑三种不同的样本量（n）：500、1000 和 10000。每种模拟设置均重复 1000 次，以确保估计结果稳定可靠。

表2：分类时间序列的各种自关联度量

度量	定义	范围
Cohen's κ	$\kappa(h) = \frac{\sum_{i=0}^{m} p_{ii}^{(h)} - \sum_{i=0}^{m} p_{i\cdot}^{(h)} p_{i\cdot}^{(h)}}{1 - \sum_{i=0}^{m} p_{i\cdot}^{(h)} p_{i\cdot}^{(h)}}$	$[-\frac{\sum_{i=0}^{m} p_{i\cdot}^{(h)} p_{i\cdot}^{(h)}}{1 - \sum_{i=0}^{m} p_{i\cdot}^{(h)} p_{i\cdot}^{(h)}}, 1]$
克雷默优 ν	$\nu(h) = \sqrt{\frac{1}{nm} \sum_{i=0}^{m} \sum_{j=0}^{m} \frac{(N_{ij}^{(h)} - E_{ij})^2}{E_{ij}}}$	$[0, 1]$
古德曼-克鲁斯卡尔优 τ	$A_{\nu}^{(\tau)}(h) = \frac{\sum_{i,j=0}^{m} \frac{p_{ij}^{(h)} p_{ij}^{(h)}}{p_{\cdot j}^{(h)}} - \sum_{i=0}^{m} p_{i\cdot}^{(h)} p_{i\cdot}^{(h)}}{1 - \sum_{i=0}^{m} p_{i\cdot}^{(h)} p_{i\cdot}^{(h)}}$	$[0, 1]$
Goodman-Krushkal 的 γ	$\gamma = \frac{\text{一致对数量} - \text{不一致对数量}}{\text{一致对数量} + \text{不一致对数量}}$	$[-1, 1]$
皮尔逊度量 X^2	$X_n^2 = n \sum_{i=0}^{m} \sum_{j=0}^{m} \frac{(p_{ij}^{(h)} - \hat{p}_{i\cdot}^{(h)} \hat{p}_{\cdot j}^{(h)})^2}{\hat{p}_{i\cdot}^{(h)} \hat{p}_{\cdot j}^{(h)}}$	$[0, n(m-1)]$
互信息	$A_{\nu}^{(u)}(h) = -\frac{\sum_{i,j=0}^{m} p_{ij}^{(h)} \ln\left(\frac{p_{ij}^{(h)}}{p_{i\cdot}^{(h)} p_{\cdot j}^{(h)}}\right)}{\sum_{i=0}^{m} p_{i\cdot}^{(h)} \ln p_{i\cdot}^{(h)}}$	$[0, 1]$

8.1 经验一致性

为检验 TSOLR模型（2）的经验一致性，我们根据模型（4）生成不同样本量的样本，其中 $P_1 = 100$、$P_2 = 5$ 和 $P_3 = 100$；模型（4）是模型（2）的一个特例。

$$\log\left(\frac{P(Y_t \le j)}{P(Y_t > j)}\right) = \theta_j - \Big[\sum_{i=1}^{2}\Big\{\beta_{1i} \cdot \sin\left(\frac{2\pi t}{P_i}\right) + \beta_{2i} \cdot \cos\left(\frac{2\pi t}{P_i}\right)\Big\} + \beta_{13} \cdot |\sin\left(\frac{2\pi t}{P_3}\right)| + \beta_{23} \cdot |\cos\left(\frac{2\pi t}{P_3}\right)| + \gamma Y_{t-1}\Big]. \tag{4}$$

在数据生成过程中，考虑了两种不同的参数设定：参数为（θ_0、θ_1、β_{11}、β_{21}、β_{12}、β_{22}、β_{13}、β_{23}、γ）=（-0.84、0.84、0、2、0、0.7、0、0.2、1.9）的三类别情形，以及参数为（θ_0、θ_1、θ_2、β_{11}、β_{21}、β_{12}、β_{22}、β_{13}、β_{23}、γ）=（-1.1、0、1.1、0.6、-1.1、1.4、-0.8、-0.1、-0.4、-0.3）的四类别情形。表3和表4汇总的模拟结果显示，随着样本量增加，估计参数的标准误会降低。这一趋势为最大似然估计量的经验一致性提供了有力证据，从而证实 TSOLR模型在各种数据规模和类别配置下均具有可靠性。

8.2 预测性能

为评估 TSOLR模型的预测性能，我们将 TSOLR模型与 ISOLR模型的预测性能进行比较。

我们从模型 (3) 中生成不同规模的样本，其中包含三种季节性指标和一个滞后响应。具体而言，季节性结构包括一个三水平分类变量 S_{1t}，用于表示季节（例如，季节 1、季节 2 和季节 3），以及两个二元指标 S_{2t} 和 S_{3t}，分别对应节庆期间和日期类型。由于季节变量有三个水平，我们创建两个虚拟变量 $S_{1t}^{(1)}$ 和 $S_{2t}^{(2)}$，分别对应季节 1 和季节 2，如下所示。

$$S_{1t}^{(1)} = \begin{cases} 1 & \text{if } t \in [100k+1, 100k+30] \\ 0 & otherwise \end{cases}$$

表 3：三类别情形下参数的平均估计值及括号中的标准误；真实参数值为 (θ_0, θ_1, β_{11}, β_{21}, β_{12}, β_{22}, β_{13}, β_{23}, γ) =(−0.84, 0.84, 0, 2, 0, 0.7, 0, 0.2, 1.9)

n	500	1000	10000
θ_0	−1.12（0.56）	−0.97（0.33）	−0.85（0.10）
θ_1	0.60（0.58）	0.74（0.34）	0.83（0.10）
β_{11}	0.00 (0.29)	0.00 (0.18)	0.00 (0.06)
β_{21}	2.27 (0.69)	2.12 (0.40)	2.01 (0.11)
β_{12}	0.00 (0.29)	−0.01 (0.19)	0.00 (0.06)
β_{22}	0.81 (0.38)	0.75 (0.23)	0.70 (0.07)
β_{13}	−0.01 (0.21)	0.00 (0.14)	0.00 (0.04)
β_{23}	0.21 (0.20)	0.21 (0.14)	0.20 (0.05)
γ	1.88 (0.21)	1.89 (0.15)	1.90 (0.04)

表4：真实参数值（θ_0, θ_1, θ_2, β_{11}, β_{21}, β_{12}, β_{22}, β_{13}, β_{23}, γ） = （−1.1, 0, 1.1, 0.6, −1.1, 1.4, −0.8, −0.1,−0.4, −0.3）在四类别情形下的参数平均估计值，括号内为标准误

n	500	1000	10000
θ_0	−1.14 (0.16)	−1.12 (0.12)	−1.10 (0.04)
θ_1	−0.02 (0.15)	−0.01 (0.11)	0.00 (0.03)
θ_2	−0.02 (0.16)	−0.01 (0.11)	0.00 (0.05)
β_{11}	0.61 (0.13)	0.61 (0.09)	0.60 (0.03)
β_{21}	−1.13 (0.14)	−1.11 (0.10)	−1.10 (0.03)
β_{12}	1.43 (0.15)	1.42 (0.10)	1.40 (0.03)
β_{22}	−0.82 (0.14)	−0.81 (0.09)	−0.80 (0.03)
β_{13}	−0.10 (0.13)	−0.10 (0.09)	−0.10 (0.03)
β_{23}	−0.41 (0.12)	−0.40 (0.09)	−0.40 (0.03)
γ	−0.33 (0.08)	−0.31 (0.06)	−0.30 (0.02)

and

$$S_{1t}^{(2)} = \begin{cases} 1 & \text{if } t \in [100k+31, 100k+70] \\ 0 & otherwise \end{cases}$$

对于 $k = 0$、1、2、$\cdots$、$\frac{n}{100} - 1$。类似地，S_{2t}和 S_{3t}定义如下。

$$S_{2t} = \begin{cases} 1 & \text{if } t \in [5k+4, 5k+5] \\ 0 & otherwise \end{cases}$$

and

$$S_{3t} = \begin{cases} 1 & \text{if } t \in [100k+81, 100k+85] \\ 0 & otherwise \end{cases}$$

对于 $k = 0$，1，2，$\cdots$，$\frac{n}{100} - 1$。也就是说，每个100天周期（在本模拟中相当于1年周期）的前30天被指定为季节 1，接下来的40天（即第 31^{st} 天至第 70^{th} 天）被指定为季节 2，其余30天被指定为季节 3。同样，S_{2t} 在每个5天周期（在本模拟中相当于1周周期）的第4天和第5天取值为1，而 S_{3t} 在每个100天周期的第81^{st} 天至第 85^{th} 天取值为1。因此，数据生成模型设定为

$$\log\left(\frac{P(Y_t \le j)}{P(Y_t > j)}\right) = \theta_j - \left(\beta_1 \cdot S_{1t}^{(1)} + \beta_2 \cdot S_{1t}^{(2)} + \beta_3 \cdot S_{2t} + \beta_4 \cdot S_{3t} + \beta_5 \cdot Y_{t-1}\right). \tag{5}$$

在数据生成过程中，考虑了两种不同的参数设定：三类别情形的参数为（θ_0，θ_1，β_1，β_2，β_3，β_4，β_5）=（-0.8，0.78，0.2，-1.4，0.3，0.9，1.1），四类别情形的参数为（θ_0，θ_1，θ_2，β_1，β_2，β_3，β_4，β_5）=（-0.9，0.02，1.2，-0.4，-1.3，0.5，0.7，1.2）。使用占数据80%的训练集来拟合模型（4）和（5），并使用占数据20%的测试集进行一步超前预测，计算准确率和加权 F1 分数，权重为测试集中不同类别所占的比例。结果见表5和表6。这些结果表明，尽管ISOLR模型优于TSOLR模型，但差异非常小。此外，随着样本量的增加，差异也在减小。这证实了TSOLR模型在预测具有季节效应的有序分类数据方面表现足够良好。

表 5：三类别情形下，使用真实参数值（θ_0, θ_1, β_{11},β_{21}, β_{12}, β_{22}, β_{13}, β_{23}, γ）=（-0.8, 0.78, 0.2, -1.4, 0.3, 0.9, 1.1）进行预测的平均准确率与加权 F1 分数

n	ISOLR模型		TSOLR模型	
	准确率	加权F1	准确率	加权F1
500	67.53%	63.16%	66.89%	62.09%
1000	67.29%	63.18%	66.92%	62.26%
10000	67.50%	63.83%	67.20%	63.00%

表 6：四类别情形下使用真实参数值（θ_0, θ_1, θ_2,β_{11}, β_{21}, β_{12}, β_{22}, β_{13}, β_{23}, γ）=（-0.9, 0.02, 1.2, -0.4, -1.3, 0.5, 0.7, 1.2）进行预测的平均准确率与加权 F1 分数

n	ISOLR模型		TSOLR模型	
	准确率	加权F1	准确率	加权F1
500	79.07%	72.61%	78.89%	72.45%
1000	79.05%	72.46%	78.96%	72.35%
10000	79.28%	72.34%	79.21%	72.44%

9 数据分析结果

9.1 通过自关联度量进行阶数选择

如表2所示，计算了多种逐滞后依赖度量，以评估自相关及潜在的模型阶数。图9中的曲线清晰地展示了自关联呈下降趋势，表明随着滞后增加，时间依赖性逐渐减弱。这些发现指导了后续分析中对一阶模型的选择。

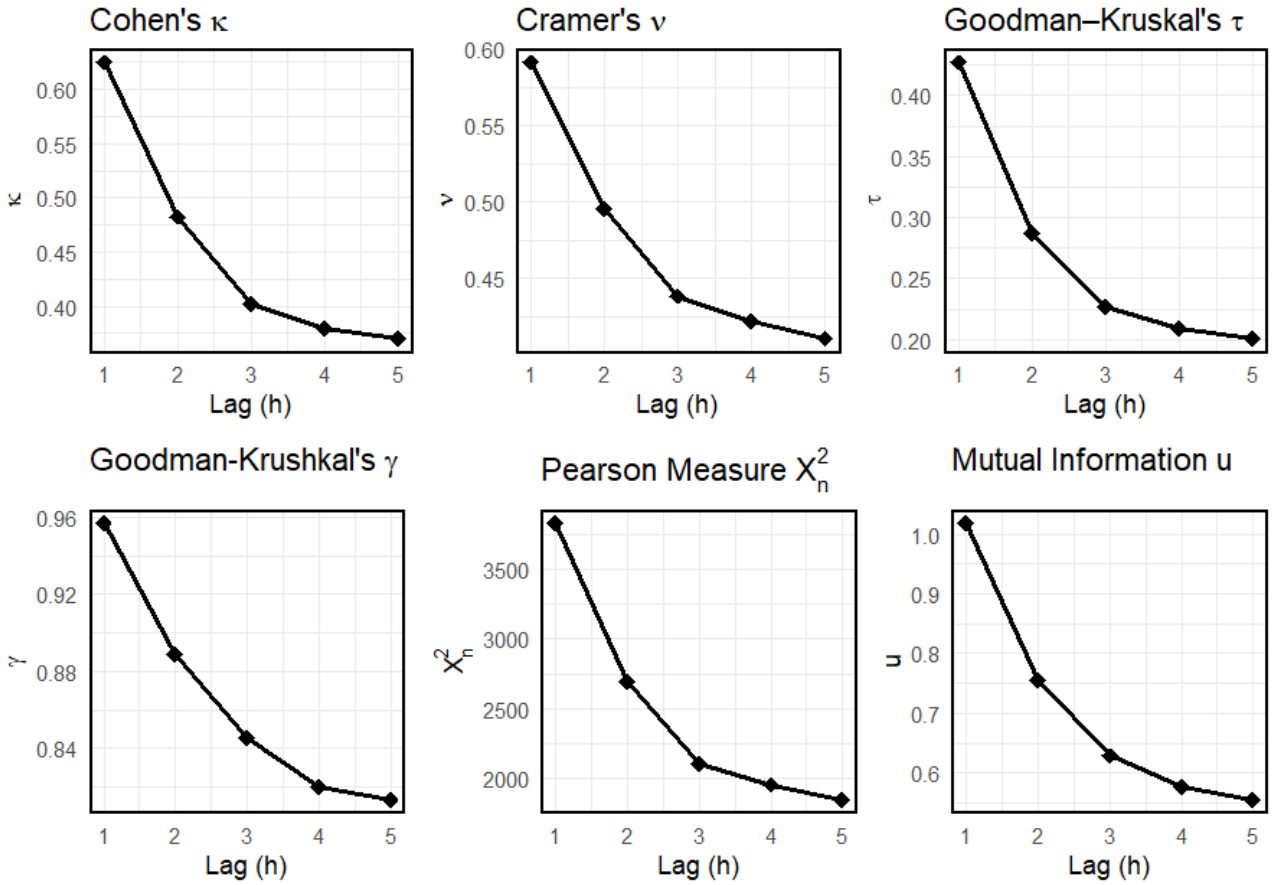


图9：分类AQI时间序列在不同滞后值下计算得到的自关联度量

9.2 预测比较

为评估各种分类AQI预测方法的有效性，我们使用2019年至2023年的数据作为训练集，并将2024年的数据作为测试集。我们评估了第4节和第6节讨论的模型。每个模型都基于季节性信息和当前AQI状态，预测次日的AQI类别。ISOLR模型按照模型（3）的设定，在其线性预测器中纳入离散季节性指标。该模型具有很强的可解释性，并能捕捉主要季节模式所产生的分类效应。TSOLR模型则按照模型（2）的设定，使用正弦项以更平滑的方式建模季节性。该模型既能捕捉每周和年度周期成分，也能捕捉排灯节等较高频率的事件。

为了补充这些模型，我们还实现了 MTD 和 PAR 模型等经典时间序列模型。这些模型不考虑季节项等外部协变量。使用这些模型进行预测时，仅依据一阶滞后值。

除了这些模型，我们还实现了随机森林、支持向量机和长短期记忆网络等机器学习和深度学习模型。这些模型分别使用指示变量和三角函数变量，并结合滞后一期 AQI 类别进行训练。尽管这些模型并非专为有序结果设计，但它们具备灵活的非线性分类能力，可作为有用的基准方法。

表 7 所示的加尔各答预测结果表明，TSOLR模型的整体表现最佳，准确率为 74.04%，加权 F1 分数为 74.07%。在所有比较模型中，包括传统时间序列模型（MTD(1)、PAR(1)）、机器学习方法（随机森林、支持向量机）和深度学习模型（长短期记忆网络），基于傅里叶级数的逻辑回归和支持向量机模型始终优于基于指示变量的对应模型。值得注意的是，随机森林使用指示特征时表现良好，但改用傅里叶特征后，其表现显著下降。图 10 展示了测试集的实际空气质量指数类别以及不同模型预测的类别。图 10 的最后三个面板中，1 和 2 分别代表基于指标的模型和基于三角级数的模型。这些结果凸显了基于傅里叶级数的季节性编码在捕捉潜在时间模式方面的有效性；尤其是与逻辑回归或支持向量机等模型结合时，可成为此情境下分类时间序列预测的合适选择。

表 7：加尔各答数据分析的预测结果

模型	准确率	加权F1
ISOLR	72.95%	72.96%
TSOLR	**74.04%**	**74.07%**
MTD(1)	72.68%	72.68%
PAR(1)	72.68%	72.68%
随机森林（基于指标）	71.31%	71.38%
随机森林（基于傅里叶级数）	65.85%	66.02%
支持向量机（基于指标）	72.68%	72.68%
支持向量机（基于傅里叶级数）	73.50%	73.51%
长短期记忆网络（基于指标）	72.88%	72.89%
长短期记忆网络（基于傅里叶级数）	73.70%	73.71%

10 讨论

本研究使用日数据，探讨有序逻辑回归模型在空气质量指数（AQI）等级短期预测中的有效性。TBATS模型（Livera 等，2011）采用三角傅里叶级数展开，以捕捉连续时间序列数据中多重且可能复杂的季节模式。受这一框架启发，我们将其基本思想拓展至序数情境，在有序逻辑回归模型（TSOLR）中纳入三角季节性成分，从而在分类时间序列中实现平滑变化的季节性结构。与此同时，我们还开发了另一种模型设定，以季节指示变量（ISOLR）取代傅里叶项，使模型能够捕捉突发或离散的季节变化。这两种互补的模型形式为表征有序空气质量指数数据中渐进和突发的季节变化提供了灵活的方法。通过同时采用指示变量和傅里叶级数项纳入季节性成分，所提模型为捕捉空气质量指数数据中固有的时间动态和周期性提供了灵活的框架。

ISOLR模型可以直观地解释分类季节变量的影响，例如不同的季节、周末以及排灯节等节日期间。相比之下，TSOLR模型能够捕捉

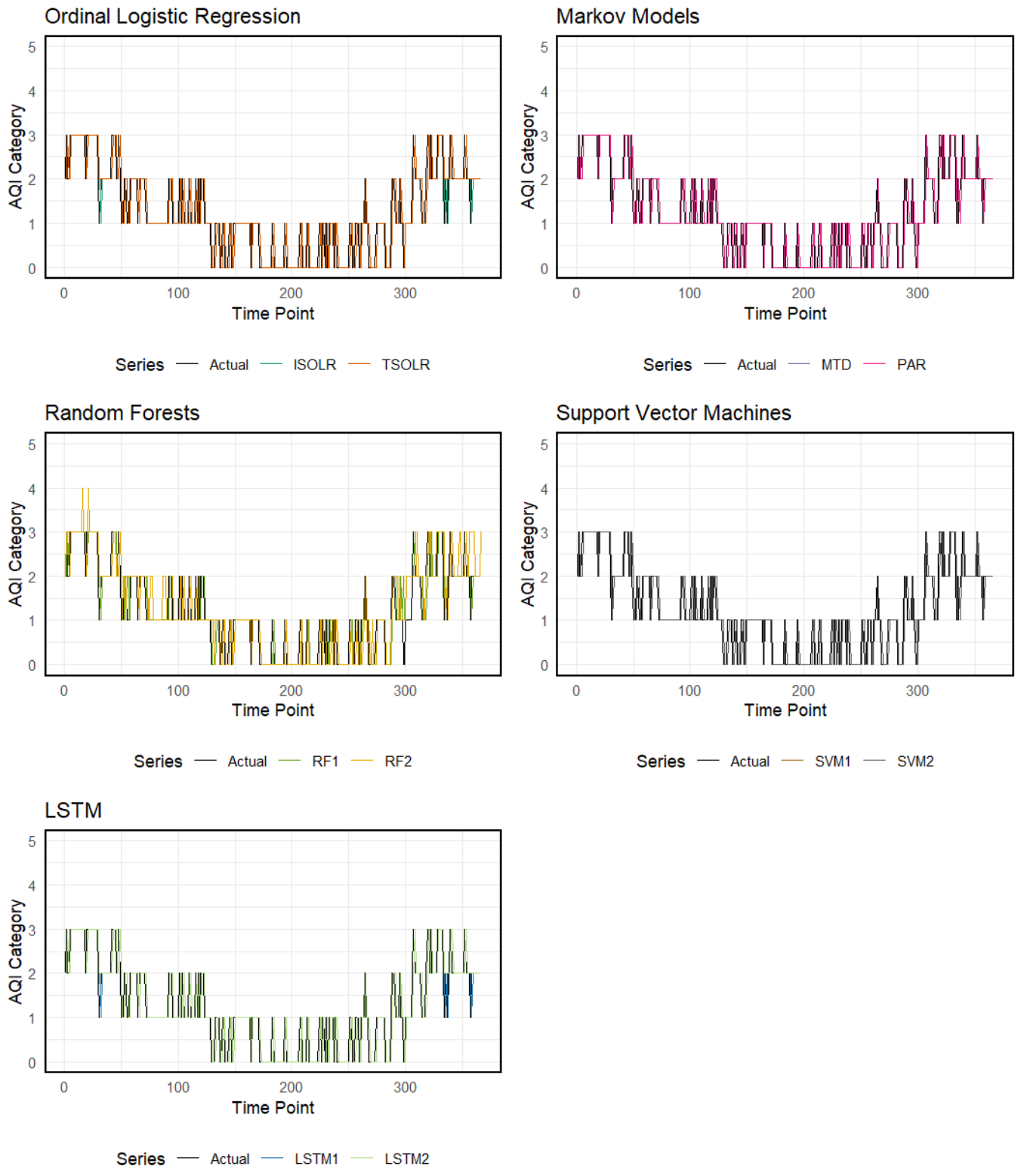


图 10：各种模型预测的空气质量指数类别及真实类别；1 和 2 分别表示基于指标的模型和基于傅里叶级数的模型

连续的季节周期和更高频率的波动，使年度和每周变化的估计更加平滑，其中也包括节日期间的剧烈峰值。

我们进行了全面的模拟实验，结果表明，采用最大似然估计法得到的模型参数具有经验一致性。此外，尽管 ISOLR模型优于 TSOLR模型，但两者差异很小，且随着样本量增加，差异逐渐缩小。这表明，TSOLR模型在预测具有多重季节性的有序分类时间序列方面表现合理。

对加尔各答空气质量指数数据进行分类时间序列分析表明，两种方法在提前1天预测方面均有效；但在多个预测时域上的整体预测准确率方面，基于傅里叶的模型略优于基于指示变量的模型。这一优势可能源于其捕捉细微且复杂季节性结构的能力，尤其是对高频事件的捕捉。与随机森林和支持向量机（SVM）等机器学习模型的比较显示，尽管这些方法表现不俗，所提出的序数逻辑回归模型还具有可解释性强、概率结构规范以及易于通过结构化季节项融入先验知识等优点。使用自关联度量进行滞后阶数选择，进一步提升了建模过程的效果，因为它能够识别时间上的有意义依赖关系。

本分析为后续研究开辟了若干方向。首先，可以扩展所提模型以纳入空间依赖，例如，利用空间相关随机效应对多个监测站的空气质量指数数据进行联合建模。其次，加入更多时变外生变量，如风速、湿度和颗粒物（$PM_{2.5}$和/或 PM_{10}）浓度，可能增强预测能力。第三，采用部分优势模型或非比例优势模型放宽比例优势假设，当预测变量效应因空气质量指数阈值而异时，可能会提高模型的灵活性。

此外，对所提模型采用贝叶斯估计，并为季节效应或傅里叶系数设置层次先验，可在数据含噪或稀疏时提供稳健的推断。最后，将所提框架整合到面向城市环境政策的实时决策支持系统中，将有助于污染控制和公共卫生规划，发挥实际效用。

致谢

作者谨向 OpenAI 致谢。本稿使用了由 OpenAI 提供的 ChatGPT GPT-4（https://chatgpt.com），并输入了“请润色该段落，使其表达清晰并纠正语法错误”等提示，以协助进行语言编辑、语法纠正并提升稿件表达的清晰度。本文未使用任何生成式人工智能或大语言模型生成原创观点、分析数据或得出结论等任何智识或分析内容。作者还谨向匿名审稿人致谢，感谢他们提出宝贵意见和建设性建议，这些意见和建议极大地提升了本文的质量和清晰度。

经费声明

本研究未获得外部经费资助。

利益冲突声明

作者声明，与本研究相关的利益冲突不存在。

数据可用性声明

支持本研究结论的数据可在 GitHub 上获取，网址为 https://github.com/ ganirban004/aqi-logistic-paper.git。这些数据来源于印度中央污染控制委员会的网站（https://airquality.cpcb.gov.in/ccr/#/ caaqm-dashboard-all/caaqm-landing/aqi-repository），该网站面向公众开放。

参考文献

Agresti, A. (2002).分类数据分析. John Wiley & Sons. Awad, M., & Khanna, R. (2015). 用于分类的支持向量机。载于高效学习机器：面向工程师和系统设计人员的理论、概念与应用 （第 39–66 页）。Apress. Berchtold, A., & Raftery, A. (2002). 用于高阶马尔可夫链和非高斯时间序列的混合转移分布模型。*Statistical Science*，*17*(3)，328–356。Biswas, A., & Song, P. X.-K. (2009). 离散值 ARMA 过程。*Statistics & probability letters*，*79*(17)，1884–1889。Desai, T., Kapadia, S., Halani, M., Zinzuwadia, P., Shah, K., Shah, M., & Prajapati, M. (2025). 用于空气质量指数预测的机器学习算法比较分析。*Machine Learningfor ComputationalScience andEngineering*，*1*(1)，14。https : / / doi. org / 10 . 1007 / s44379 - 025 - 00014-2 Dutta, A., Rahman, M. H., Suresh, R., Sharma, V. P., Sharma, S., Sharma, P., & Jacob, J. (2025).加尔各答和豪拉市政公司辖区大气黑碳浓度的时空变化 （技术报告）。能源与资源研究所（TERI），新德里。https : / / www . teriin . org / policy - brief / spatiotemporal - variations - atmospheric - black - carbon - concentrations-kolkata-and-howrah Fokianos, K., & Kedem, B. (2003). 分类时间序列的回归理论。*Statistical Science*，*18*。https://doi.org/10.1214/ss/1076102425 Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). 使用机器学习技术预测空气质量指数：比较分析。环境与公共卫生杂志，*2023*年(1)，4916267。https://doi.org/https://doi.org/10.1155/2023/ 4916267 Gupta, V., Kapadia, S., & Bhadane, C. (2023). 印度空气质量的时间序列分析与预测。*2023*年第五届电气、计算机与通信技术国际会议（*ICECCT*），1–5。https://doi.org/10.1109/ICECCT56650.2023.10179673 Jacobs, P. A., & Lewis, P. A. W. (1978a). 混合分布生成的离散时间序列。I：相关性与游程性质。英国皇家统计学会会刊。*B*辑（方法论），*40*(1)，94–105。Jacobs, P. A., & Lewis, P. A. W. (1978b). 混合分布生成的离散时间序列。II：渐近性质。英国皇家统计学会会刊。*B*辑（方法论），*40*(2)，222–228。Jacobs, P. A., & Lewis, P. A. W. (1978c).混合分布生成的离散时间序列。*III*：自回归过程（*DAR(p)*）（技术报告）。海军研究生院，加利福尼亚州蒙特雷。加利福尼亚州蒙特雷。海军研究生院。Janitza, S., Tutz, G., & Boulesteix, A.-L. (2016). 用于有序响应的随机森林：预测与变量选择。*Computational Statistics & Data Analysis*，*96*，57–73。https://doi.org/https: // doi.org/10.1016/j.csda.2015.10.005 Lindemann, B., M¨uller, T., Vietz, H., Jazdi, N., & Weyrich, M. (2021). 用于时间序列预测的长短期记忆网络综述 [14届CIRP智能计算会议

制造工程，7月15日至17日 2020]。*Procedia CIRP*，*99*，650–655。https://doi.org/https: //doi.org/10.1016/j.procir.2021.03.088 Liu, H., Yan, G., Duan, Z., & Chen, C. (2021). 空气质量时间序列预测的智能建模策略：综述。应用软计算，*102*，106957。https://doi.org/https://doi.org/10.1016/j.asoc.2020.106957 Liu, M., Li, Q., & Zhu, F. (2022). 使用灵活的分类自回归模型对空气质量等级建模。随机环境研究与风险评估，*36*(9)，2835–2845。https://doi.org/10.1007/s00477-021-02164-0 Liu, Y., Wen, L., Lin, Z., Xu, C., Chen, Y., & Li, Y. (2024). 基于时间序列的空气质量历史相关模型。《科学报告》。https://doi.org/https://doi.org/10.1038/s41598-024-74246-2 Livera, A. M. D., Hyndman, R. J., & Snyder, R. D. (2011). 使用指数平滑法预测具有复杂季节性模式的时间序列。美国统计学会会刊，*106*(496)，1513–1527。https://doi.org/10.1198/jasa.2011.tm09771 Makatjane, K., & Mmelesi, K. (2024). 提高风险度量预测的模型准确度：集成方法的应用。《应用经济学杂志》，*27*(1)，2395775。https://doi.org/ 10.1080/15140326.2024.2395775 Mishra, A., & Gupta, Y. (2024). 使用深度学习算法预测空气质量指数的比较分析。《空间信息研究》，*32*(1)，63–72。https://doi.org/10.1007/s41324-023- 00541-1 Pegram, G. G. S. (1980). 多滞后马尔可夫链自回归模型。《应用概率杂志》，*17*(2)，350–362。Raftery, A. E. (1985). 高阶马尔可夫链模型。英国皇家统计学会会刊。*B*辑（方法论），*47*(3)，528–539。Sethi, J. K., & Mittal, M. (2020). 使用单变量和多变量时间序列模型分析空气质量。*2020*年第十届云计算、数据科学与工程国际会议（*Confluence*），823–827。https://doi.org/10.1109/Confluence47617.2020.9058303 Thayyib, P., Thorakkattle, M. N., Usmani, F., Yahya, A. T., & Farhan, N. H. (2023). 使用 TBATS、ETS、神经网络和混合时间序列模型预测印度商品及服务税收入。《*Cogent*经济与金融》，*11*(2)，2285649。https://doi.org/10.1080/23322039.2023.2285649了解空气污染：指南。2022年。https://www.wbpcb.gov.in/files/Th-12-2022-12-55- 16Understanding%20Air%20Pollution.pdf Wang, L., Guo, Y., & Liang, C. (2025). 探究基于聚类装袋的天气注意力在 GARCH-MIDAS 框架下对石油市场波动率的影响。《应用经济学》，*57*(38)，5839–5858。https://doi.org/10.1080/00036846.2024.2370494 Weiß, C. H. (2020). 分类时间序列的状态转换离散 ARMA 模型。《熵》，*22*(4)，458。Weiß, C. H., & G¨ob, R. (2008). 分类时间序列中序列依赖的测量。《*AStA*统计分析进展》，*92*(1)，71–89。Weiß, C. H., & Swidan, O. (2025). 分类时间序列的加权离散 ARMA 模型。《时间序列分析杂志》，*46*(3)，505–529。