

基于因果分析算法的成都市AQI预测研究

杨 倩

成都理工大学数学科学学院, 四川 成都

收稿日期: 2024年12月23日; 录用日期: 2025年1月13日; 发布日期: 2025年1月26日

摘 要

本文旨在利用因果分析算法PCMCI对成都市空气质量指数(AQI)进行预测。随着城市化的加速, 空气污染问题日益严重, 因此准确预测AQI对于改善城市空气质量具有重要意义。本文基于成都市AQI预测的背景与现状, 介绍了PCMCI算法的基本原理及其在因果分析中的应用。随后, 对成都市2019至2023年的AQI数据进行因果分析, 识别出与AQI相关的因果变量, 并采用长短期记忆网络进行AQI预测。实验结果表明, 本文所提方法的预测结果的均方根误差RMSE相较于传统的ARIMA预测模型和单变量LSTM模型预测精度分别提升了22.14%和9.58%, 平均绝对百分比误差MAPE相较于传统的ARIMA预测模型和单变量LSTM模型预测精度分别提升了30.98%和11.59%。基于PCMCI提取的因果关系变量能显著提升AQI预测的准确性, 能够为成都市的空气质量提供有效的决策支持。

关键词

PCMCI, AQI, 长短期记忆网络, 因果分析

Study on AQI Forecasting in Chengdu Based on Causal Analysis Algorithm

Qian Yang

College of Mathematical Sciences, Chengdu University of Technology, Chengdu Sichuan

Received: Dec. 23rd, 2024; accepted: Jan. 13th, 2025; published: Jan. 26th, 2025

Abstract

The aim of this paper is to forecast the air quality index (AQI) in Chengdu by using the causal analysis algorithm PCMCI. With the acceleration of urbanization, the problem of air pollution is becoming more and more serious, so it is of great significance to accurately predict AQI for improving urban air quality. Based on the background and current situation of AQI prediction in Chengdu, this paper introduces the basic principles of PCMCI algorithm and its application in causal analysis.

Subsequently, the AQI data of Chengdu from 2019 to 2023 are causally analyzed to identify the causal variables related to AQI, and the long and short-term memory network is used for AQI prediction. The experimental results show that the root mean square error RMSE of the prediction results of the method proposed in this paper improves the prediction accuracy by 22.14% and 9.58% compared with the traditional ARIMA prediction model and univariate LSTM model, respectively, and the mean absolute percentage error MAPE improves the prediction accuracy by 30.98% compared with the traditional ARIMA prediction model and univariate LSTM model by 30.98% and 11.59%. The causal variables extracted based on PCMCI can significantly improve the accuracy of AQI prediction and provide effective decision support for air quality management in Chengdu.

Keywords

PCMCI, AQI, LSTM, Causal Analysis

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

1.1. 研究背景及意义

随着工业化和城市化的推进,空气污染已成为全球性问题,尤其在快速发展的城市。成都坐落于中国西南区域,位居四川盆地的中心地带,是西部地区重要的经济、文化和交通枢纽。近年来,由于城市工业化进程的推进、机动车尾气排放量的增加等多重因素,成都市的空气质量出现了较大的起伏,这一现象引起了社会各界的广泛关注[1]。AQI作为衡量空气质量的重要指标,直接影响居民的健康和生活质量。近年来,越来越多的研究开始关注AQI的预测,以期和政策制定和环境治理提供科学依据。传统的统计学和机器学习的预测方法往往基于变量的相关性分析,但相关关系不等于因果关系,忽视了变量之间内在的因果联系。本文旨在通过引入因果分析算法,结合长短期记忆网络,提升空气质量指数预测的精度。

1.2. 国内外研究现状

在大数据时代的背景下,针对时间序列维度不断增加、变量间关系愈发复杂这一问题,对复杂系统进行建模与分析并进行时间序列的趋势预测是重要的研究课题。由于传统的相关性分析无法处理复杂系统变量间的影响,在实际应用中难以满足分析要求,因此,因果分析的方法开始得到广泛关注和研究。作为人工智能的核心组成部分,机器学习凭借其强大的数据学习和潜在规律挖掘能力,在处理复杂、海量数据方面展现出独特的优势,涵盖了贝叶斯、决策树、随机森林、支持向量机、神经网络及集成学习等多种算法。当前,国内外众多学者正积极利用机器学习技术构建空气质量预测模型,以期提高预测的精准度。例如,李高等人[2]通过应用机器学习中的LSTM(长短期记忆网络)优化算法,成功构建了空气质量二次预测模型,相较于一次预报,该模型显著提升了预测数据与真实数据的匹配度;徐艳平等[3]则基于随机森林回归算法,结合气象参数,为重庆市构建了空气质量预测模型,利用三年的指标数据验证了模型的高预测精度,并深入分析了影响因子的相对重要性;郑红等[4]考虑气象因素、空间相关性和时间依赖性对AQI的影响,建立联合训练模型,以此提高预测的可靠性。朱菊香等[5]提出了一种基于智能算法和LSTM混合的模型,有效提高拟合度并降低预测的误差。

2. 理论基础

2.1. PCMCI 算法原理

因果网络学习方法是一种将变量作为节点，然后通过统计指标计算和推断，将有因果关系的节点通过有向线段连接起来的因果分析方法。常用的因果网络学习方法有 PC、FCI 等。针对现在系统中变量个数较多，且含有无关变量的问题，采用两阶段策略的因果网络学习方法被广泛使用。通常，这类方法首先利用特征选择方法对原始数据进行预处理，达到降低计算复杂度和减弱干扰信息影响的目的。然后，基于上一阶段的结果对变量间的因果关系进行探究。

其中，PCMCI 是一种具有代表性的两阶段因果网络学习方法，其在计算复杂度、一致性等方面都有着很好的效果。接下来对 PCMCI 的基本原理进行详细介绍。

PCMCI 是一种基于约束的因果推断算法，由 Runge 在 2019 年提出[6]。PCMCI 方法基于潜在结果框架，通过比较实际观察到的结果与潜在结果之间的差异，来推断因果效应。其基本核心是对因果关系中的潜在因果历程进行建模和分析。

PCMCI 包括 PC 和 MCI 两个阶段：第一步执行 PC 算法对所有的变量 $X_t^j \in \{X_t^1, \dots, X_t^N\}$ 中找到 X_t^j 的因果父节点 $P(X_t^j)$ 的估计值 $\hat{P}(X_t^j)$ ，即估计变量间的瞬时影响，得到因果骨架图。第二步，执行瞬时条件独立性检验(MCI 算法)，检验 $X_{t-\tau}^i$ 和 X_t^j 之间是否存在着瞬时条件独立性。由此就可以获知父节点 X_t^j 与时序上滞后的子节点 $X_{t-\tau}^i$ 之间是否存在着联系[7]。

$$\text{MCI}:(X_{t-\tau}^i \perp X_t^j | P(X_t^j)\{X_{t-\tau}^i\}, P(X_{t-\tau}^i)) \quad (1)$$

2.2. LSTM 模型

LSTM (长短期记忆网络)是一种常用于处理序列数据的深度学习模型(图 1)，与传统的 RNN (循环神经网络)相比，LSTM 引入了三个门(输入门、遗忘门、输出门)和一个记忆单元(cell state) [8]，这些机制使得 LSTM 能够更好地处理序列中的长期依赖关系，并可以有效解决传统 RNN 在处理长时间序列数据时的梯度消失和梯度爆炸问题[9]。

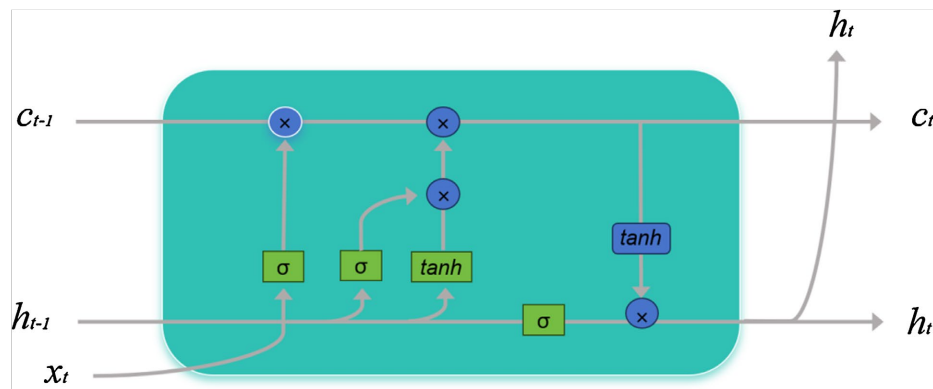


Figure 1. LSTM network structure
图 1. LSTM 网络结构

LSTM 神经网络大体结构与 RNN 神经网络结构相同，其计算过程如式(2)至(8)所示，其中 f 、 i 、 o 分别为遗忘门、输入门、输出门； C 代表短期记忆， h 代表长期记忆。 σ 为激活函数； W 是由单位向量转换到门向量的变换权矩阵(W_f 、 W_i 、 W_o 、 W_c)； x 作为当前输入； b 是输入层各门获得的向量特征(b_f 、 b_i 、 b_o 、 b_c)。

b_o 、 b_c)。 c_t 是细胞状态。

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (3)$$

$$\text{Sigmoid} = \frac{1}{1 + e^{-1}} \quad (4)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (5)$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (6)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (7)$$

$$h_t = o_t * \tanh(c_t) \quad (8)$$

3. 数据介绍

3.1. 数据来源

本文采用 2019 年 1 月 1 日到 2023 年 12 月 31 日成都市的 6 种污染气体的日平均质量浓度数据和每日空气质量指数进行实例分析, 包含 $\text{PM}_{2.5}$ ($\mu\text{g}/\text{m}^3$)、 PM_{10} ($\mu\text{g}/\text{m}^3$)、 SO_2 ($\mu\text{g}/\text{m}^3$)、 NO_2 ($\mu\text{g}/\text{m}^3$)、 CO ($\mu\text{g}/\text{m}^3$)、 O_3 ($\mu\text{g}/\text{m}^3$) 和 AQI 值, 数据来源于空气质量在线监测分析平台。除去缺失值和极端值, 2023 年成都市的 AQI 数据序列情况如图 2 所示。

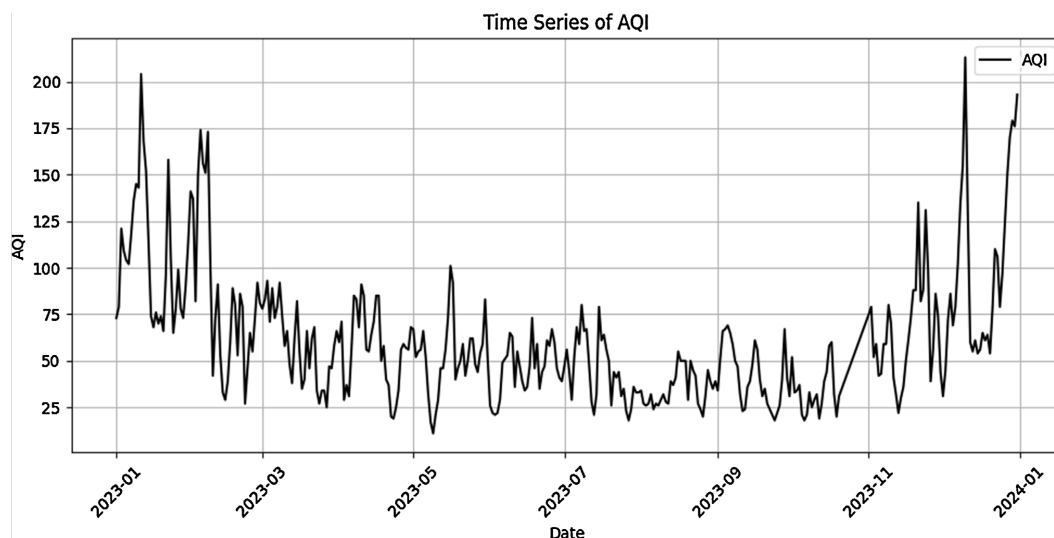


Figure 2. AQI data series for Chengdu in 2023

图 2. 2023 年成都市的 AQI 数据序列

3.2. 数据预处理

2019~2023 年成都 $\text{PM}_{2.5}$ 、 PM_{10} 、 SO_2 、 NO_2 、 CO 、 O_3 和 AQI 数据中存在的缺失值, 本文采用插值法 [10] 来填补数据集中的缺失值, 取数据框中缺失值的上一条数值和下一条数值的平均值代替原缺失值。

同时为降低由于不同特征之间的量纲差异而导致的预测误差, 对整个数据进行归一化处理, 公式为:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (9)$$

4. 实验分析

4.1. 因果分析

首先，选择 AQI 作为目标变量，通过 PCMCI 算法的因果推断分析得到与目标变量最相关的特征集合，因果分析结果如图 3 所示。

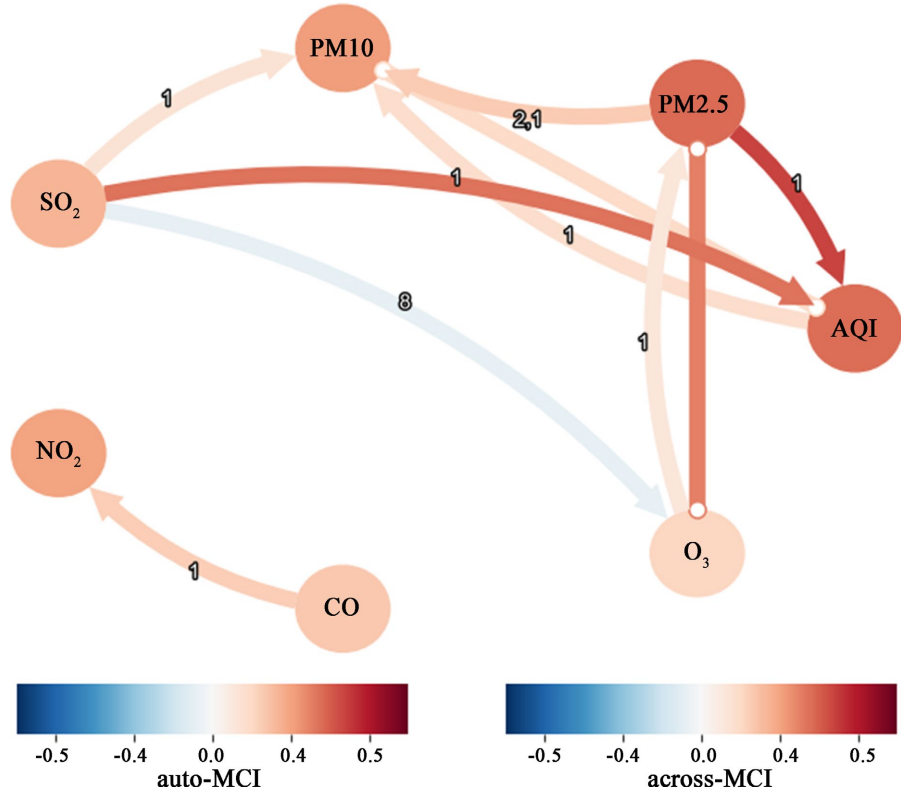


Figure 3. Causal analysis of AQI dataset
图 3. AQI 数据集因果关系分析

根据上图所示的因果关系分析结果，选择 PM_{2.5} 和 SO₂ 作为 AQI 的驱动变量。另外可以看到因果关系图中有一些不定向边，这是由于 PCMCI 算法自身的局限性导致的，本文在此不讨论相关算法的优化改进。

4.2. LSTM 模型的 AQI 预测

将模型数据分为 2 部分，总数据随机抽取 80% 的数据作为训练集，总数据的 20% 作为测试集，以最终预测的 AQI 值与实际真值进行对比。(图 4)

为了验证模型预测的准确性，分别采用只考虑 AQI 历史数据单个变量的 LSTM 模型、ARIMA 模型和包含因果分析的 PCMCI-LSTM 组合模型，对 AQI 值预测进行仿真分析。为了更好地对模型的预测精度进行对比评估，选择均方根误差(RMSE)和平均绝对百分比误差(MAPE)作为评估指标，计算公式为：

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

$$\text{MAPE} = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

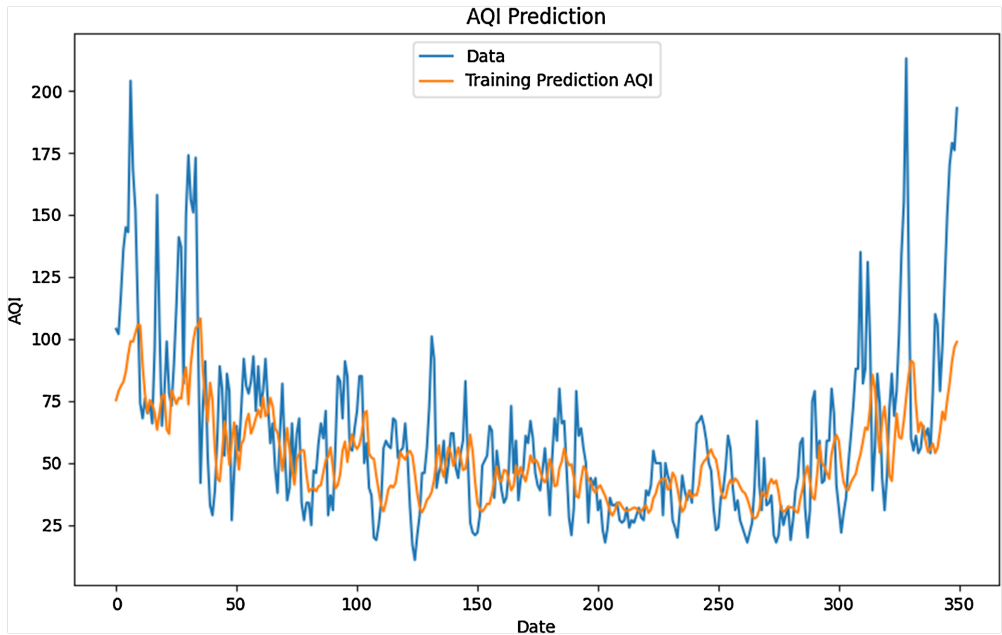


Figure 4. AQI prediction results
图 4. AQI 预测结果

各个模型的预测误差结果如表 1 所示。

由表 1 可知，包含因果分析的 PCMCI-LSTM 组合模型的预测精度显著高于对比模型。该模型预测 AQI 值的 MAPE 和 RMSE 均低于其他方法，总体来说 AQI 总体预测有较大提升。

Table 1. Prediction error comparison results
表 1. 预测误差对比结果

模型	RMSE	MAPE
ARIMA	36.7132	0.7651
单变量 LSTM	31.6128	0.5973
PCMCI-LSTM	28.5832	0.5281

5. 结论

本研究成功地将因果分析算法应用于成都市空气质量指数的预测中，并通过与深度学习模型 LSTM 的结合，实现了预测精度的显著提升。利用 PCMCI 算法，我们有效地识别了与 AQI 存在因果关系的变量，克服了传统预测方法中仅依赖相关性分析的局限性。实验结果表明，基于因果关系的变量选择能够更准确地反映影响 AQI 的关键因素，从而提高了预测的准确性。这一研究不仅为成都市的空气质量提供了有力的决策支持，也为其他城市的空气质量预测研究提供了新的思路和方法。未来，我们将继续探索和优化因果分析算法在空气质量预测中的应用，以期为实现更加精准和有效的环境管理贡献力量。

参考文献

[1] 朱艺轩. 基于机器学习的成都市空气质量预测[J]. 信息记录材料, 2024, 25(7): 160-162.

[2] 李高平, 邱治邦, 苗加庆, 等. 基于 LSTM 的空气质量预测模型[J]. 西南民族大学学报(自然科学版), 2023, 49(1): 67-73.

-
- [3] 徐艳平, 陈义安. 基于随机森林回归和气象参数的城市空气质量预测模型——以重庆市为例[J]. 重庆工商大学学报(自然科学版), 2021, 38(6): 118-124.
- [4] 郑红, 程云辉, 胡阳生, 等. 基于 MLP&ST 模型的空气质量预测[J]. 应用科学学报, 2022, 40(2): 302-315.
- [5] 朱菊香, 谷卫, 任明煜, 等. 基于 SWT-ISSA-LSTM 的地铁空气质量预测建模[J]. 国外电子测量技术, 2023, 42(7): 164-174.
- [6] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. and Sejdinovic, D. (2019) Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets. *Science Advances*, **5**, eaau4996.
<https://doi.org/10.1126/sciadv.aau4996>
- [7] 迟赫. 因果推断视角下碳排放权价格影响因素分析[D]: [硕士学位论文]. 北京: 中央财经大学, 2023.
- [8] 杨丽, 吴雨茜, 王俊丽, 等. 循环神经网络研究综述[J]. 计算机应用, 2018, 38(S2): 1-6+26.
- [9] 栗然, 马涛, 张潇, 等. 基于卷积长短期记忆神经网络的短期风功率预测[J]. 太阳能学报, 2021, 42(6): 304-311.
- [10] 路凯丽, 杨露, 李涛. 基于集成深度学习模型的空气质量指数预测[J]. 南京信息工程大学学报, 2024, 16(1): 56-65.