

灰色预测与SVR组合模型实现空气质量预测

刘世婕, 金红伟, 刘艳艳, 张子含

北京邮电大学世纪学院, 北京

收稿日期: 2026年2月1日; 录用日期: 2026年3月2日; 发布日期: 2026年3月16日

摘要

气象因素不仅直接影响空气质量, 还增加了空气质量预测(AQP)的难度。然而, 现有方案在学习同时涉及时间序列空气污染物和气象因素的模式方面效果不佳。本研究首先探讨了灰色预测与SVR组合模型在AQP中的优越性。为了分析气象因素在AQP中的作用, 我们选取PM_{2.5}、NO₂、PM₁₀等7项核心指标作为特征变量, 通过Pearson相关性分析与Lasso回归完成特征筛选, 保留所有有效特征以保障预测全面性。模型构建中, 利用灰色预测GM(1, 1)模型处理数据不确定性与小样本特性, 挖掘整体变化趋势; 结合SVR算法精准捕捉数据中复杂非线性关联, 实现优势互补。同时采用训练集R²得分、历史数据R²得分、平均绝对误差、中值绝对误差、可解释方差五项评价指标对模型进行全面性评估。实验结果显示, 模型训练集与历史数据R²得分均达0.9745, 平均绝对误差为3.1439, 中值绝对误差为2.0044, 可解释方差值达0.9747, 且预测值与真实值对比曲线高度吻合。结果表明, 该组合模型拟合精度高、抗干扰能力强、泛化性能稳定, 能有效捕捉空气质量波动规律, 该模型在预测空气质量方向起着重要作用。

关键词

灰色预测, SVR, GM(1, 1), 空气质量, 评价指标

Realizing Air Quality Prediction with the Combined Model of Grey Prediction and SVR

Shijie Liu, Hongwei Jin, Yanyan Liu, Zihan Zhang

Century College, Beijing University of Posts and Telecommunications, Beijing

Received: February 1, 2026; accepted: March 2, 2026; published: March 16, 2026

Abstract

Meteorological factors not only directly affect air quality but also increase the difficulty of air quality prediction (AQP). However, existing schemes perform poorly in learning patterns that simultaneously involve time-series air pollutants and meteorological factors. This study first explores the superiority

of the combined model of grey prediction and SVR in AQP. To analyze the role of meteorological factors in AQP, we select seven core indicators such as PM_{2.5}, NO₂, and PM₁₀ as feature variables. Feature selection is completed through Pearson correlation analysis and Lasso regression, retaining all effective features to ensure comprehensive prediction. In model construction, the grey prediction GM (1, 1) model is utilized to handle data uncertainty and small sample characteristics, mining overall trends; combined with the SVR algorithm, it accurately captures complex nonlinear associations in the data, achieving complementary advantages. Meanwhile, five evaluation metrics including R² score of the training set, R² score of historical data, mean absolute error, median absolute error, and explained variance are used to comprehensively evaluate the model. The experimental results show that both the R² score of the model training set and the historical data reach 0.9745, with an average absolute error of 3.1439, a median absolute error of 2.0044, and an explained variance of 0.9747. Furthermore, the curve comparing the predicted values and the actual values is highly consistent. The results indicate that this combined model has high fitting accuracy, strong anti-interference capability, and stable generalization performance, effectively capturing the fluctuation patterns of air quality. This model plays a significant role in predicting air quality.

Keywords

Grey Prediction, SVR, GM (1, 1), Air Quality, Evaluation Index

Copyright © 2026 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

空气质量对个体健康与日常活动的深远影响, 已成为社会各界的普遍共识。随着社会科学技术的持续革新与商业化进程的不断推进, 空气质量不可避免地出现波动, 这一问题也因此引发公众与相关政府部门的广泛关切。为积极应对这一挑战, 多国近年来陆续搭建起覆盖广泛的空气质量监测网络。2013年10月17日, 世界卫生组织(WHO)下属的国际癌症研究机构(IARC)正式将户外空气污染列为人类致癌物, 这一认定进一步推动了环境科学、统计学及计算机科学领域对空气质量预测研究的关注与投入, 相关研究热度持续攀升[1]。

当前, 减少碳排放、改善空气质量、应对气候变化已成为制约经济社会可持续发展的核心挑战。准确的空气质量预测是气象服务体系的重要构成, 而对空气质量的动态管控, 在追踪污染物扩散轨迹、研判空气质量变化趋势等方面发挥着关键作用。提升空气质量预测的精准度, 将为实现清洁空气与碳中和目标提供有力支撑, 因此, 充分挖掘算法模型在空气质量预测中的应用价值, 已成为当下亟待推进的重要课题[2][3]。

近年来, 空气质量预测(AQP)领域涌现出数值预测法、统计法、机器学习法及深度学习法等多种技术路径。数值预测法依赖多监测站动态数据, 通过气象与化学模型模拟污染物排放、迁移及扩散过程, 但存在耗时长、对领域知识要求高的局限; 统计法(如 ARIMA、MLR、VAR)基于数据驱动预测, 却难以捕捉复杂底层模式; 传统机器学习法(如 ANN、SVM)无需复杂数据分析, 能学习非线性关系且性能优于统计法, 但其处理长时序数据的能力不足[4]-[6]。

本文创新灰色预测联合 SVR 的 AQP 预测模型, 适用于小样本数据集, 能挖掘整体趋势并弱化噪声, 实现短时间内的空气质量预测, 且效果较好。我们的模型面对大气环境多因素耦合情况具有较强的抗干扰能力与稳定性。

2. 数据处理以及特征提取

本文主要进行空气质量预测，数据采用 2025 年 1 月~6 月沈阳市的各项空气指标，空气质量受多种因素影响，分析数据包含以下关键指标：

空气质量指数(AQI)：综合反映空气质量状况的无量纲指数；

质量等级：基于 AQI 划分的空气质量类别(如优、良、轻度污染等)；

细颗粒物(PM2.5)：空气动力学直径 $\leq 2.5 \mu\text{m}$ 的颗粒物；

可吸入颗粒物(PM10)：空气动力学直径 $\leq 10 \mu\text{m}$ 的颗粒物；

一氧化碳(CO)：有毒气体污染物；

二氧化氮(NO₂)：主要气态污染物之一；

二氧化硫(SO₂)：典型大气污染物；

8 小时平均臭氧(O₃_8 h)：光化学污染的标志性污染物。

本文采用相关性分析以及 Lasso 回归方法进行特征提取，相关性分析可快速排除与目标变量弱相关的噪声特征，减少后续模型计算量；Lasso 回归则进一步从剩余特征中筛选出对预测最关键的变量，在保证预测精度的同时简化模型结构。对于空气质量预测场景，这种阶梯式筛选既能保留核心影响因素，又能避免过拟合，提升模型的泛化能力与计算效率。

2.1. 相关性分析

相关性分析是对两个或多个具备相关性的特征元素开展分析，以此衡量特征因素间相关密切程度的方法。在统计学领域，常运用 Pearson 相关系数来进行相关性分析。Pearson 相关系数用于度量两个特征 X 和 Y 之间线性相关的强弱，是较为简单的一种相关系数，通常用 r 或 ρ 表示，取值范围在 $[-1,1]$ 之间。并且，它具有一个关键特性，即不会因特征的位置或大小变化而改变，比如将 X 变为 $a+bX$ ， Y 变为 $c+dY$ (其中 a 、 b 、 c 、 d 为常数)，二者的相关系数依然保持不变。

若两个向量 $X = [x_1, x_2, \dots, x_n]$ ， $Y = [y_1, y_2, \dots, y_n]$ ，它们之间的 Pearson 相关系数为

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2}} \quad (1)$$

当 $0 < r < 1$ 时， X 和 Y 呈现正相关关系；当 $-1 < r < 0$ 时， X 和 Y 呈现负相关关系；若 $r = 0$ ，则 X 和 Y 不相关；若 $r = 1$ ， X 和 Y 完全正相关；若 $r = -1$ ， X 和 Y 完全负相关。且 $|r|$ 越接近 1， X 和 Y 的相关性越大。

如表 1，该相关系数矩阵是通过训练 1 月份空气质量数据生成，其中行与列均对应 7 个变量分别为(PM2.5、NO₂、PM10、SO₂、CO、空气质量、O₃_8 h)，表格数据直观呈现变量间相关性强度与方向，对角线值均为 1(变量自身完全相关)，非对角线值为两两变量相关系数。根据表格分析：7 项指标呈现鲜明的协同关联特征：多数污染物间呈正相关，但是 AQI 与所有污染物呈负相关，这就会导致我们认为出现异常值，根据研究，我们提供的数据 AQI 的数值越小代表空气质量越好，故该负相关完全合理，并非数据异常。

污染物协同性突出，例如从表格中易发现 PM2.5、NO₂、O₃_8 h 构成核心正相关组，三者相关系数均在 0.896 以上，其中 PM2.5 与 O₃_8 h 达 0.976，说明三者污染来源高度同源，受共同污染源或气象条件影响，浓度变化高度同步。PM10 与 SO₂ (0.853)呈强正相关，体现颗粒物与气态污染物的密切关联。

各指标关联存在差异：比如 PM2.5 与 PM10、SO₂ 与 CO 等呈中等正相关，CO 与其他污染物多为弱相关，可能其排放源更分散，对整体污染协同影响较小。AQI 与 SO₂、PM10(相关系数-0.639、-0.544)呈

中等负相关，是影响大气质量的核心污染物；与其余污染物呈弱负相关，影响相对温和。

Table 1. Correlation coefficient matrix

表 1. 相关性系数矩阵

相关系数矩阵	PM2.5	NO ₂	PM10	SO ₂	CO	AQI	O ₃ _8 h
PM2.5	1.000	0.896	0.780	0.607	0.429	-0.326	0.976
NO ₂	0.896	1.000	0.695	0.409	0.299	-0.272	0.939
PM10	0.780	0.695	1.000	0.853	0.665	-0.544	0.748
SO ₂	0.607	0.409	0.853	1.000	0.653	-0.639	0.527
CO	0.429	0.299	0.665	0.653	1.000	-0.398	0.388
AQI	-0.326	-0.272	-0.544	-0.639	-0.398	1.000	-0.326
O ₃ _8 h	0.976	0.939	0.748	0.527	0.388	-0.326	1.000

2.2. 热力图分析

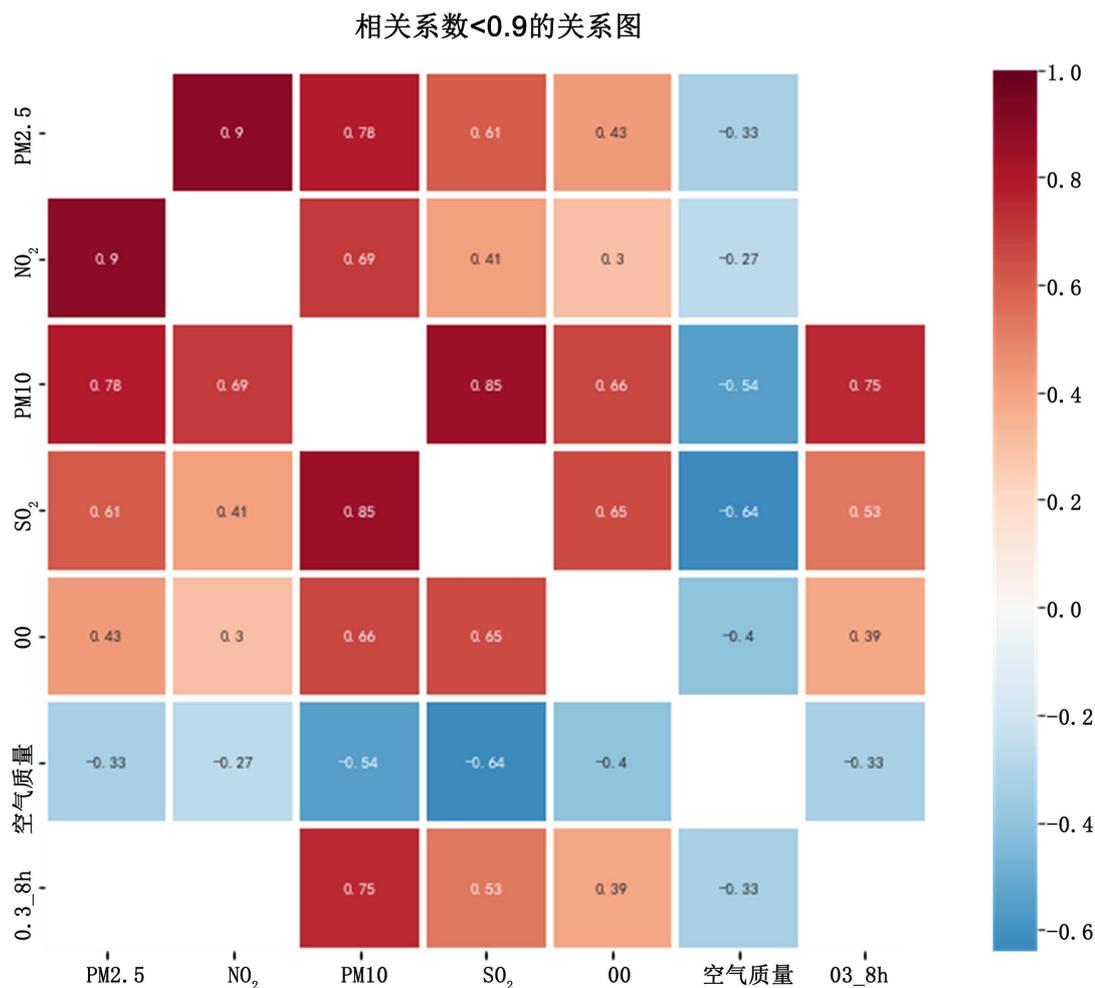


Figure 1. Feature heatmap

图 1. 特征热力图

同时我们根据表格对应生成了热力图也可以直观地说明上述相关性，如图 1：

深红色区域：代表相关系数接近 1.0，说明对应变变量间呈强正相关，即一个变量增大时，另一个变量也极有可能随之增大。

深蓝色区域：代表相关系数接近-0.6，说明对应变变量间呈较强负相关，即一个变量增大时，另一个变量很可能随之减小。

浅色系(浅红、浅蓝等)区域：代表相关系数接近 0，说明对应变变量间线性相关性很弱，变量间通过线性关系相互影响的趋势不明显。

从图中色块分布来看：

左上角等区域存在较多深红色、红色色块，表明这些区域对应的变量间正相关性较强。

右下方等区域存在蓝色色块，表明这些区域对应的变量间负相关性较强。

中间偏下等区域存在较多浅色系色块，表明这些区域对应的变量间线性相关性较弱。

我们生成热力图的目的是把抽象的数值关系转化为直观的视觉信息，让复杂的关联特征一目了然，不需要逐行的核对数字，直接通过颜色深浅快速分析出相关性的强弱问题，提高分析效率，同时热力图进一步支撑了我们的核心结论。

2.3. Lasso 回归

本文的模型中采用了 Lasso 回归，该方法属于正则化方法的一种，是压缩估计。它通过构造一个惩罚函数得到一个较为精炼的模型，使得它压缩一些系数，同时设定一些系数为零，保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。

Lasso 参数定义如下：

$$\hat{\beta}(\text{lasso}) = \left\| y - \sum_{j=1}^p x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (2)$$

其中， λ 控制模型复杂程度， λ 越大，对包含较多特征的线性模型惩罚越重，能得到特征更少的模型；确定 λ 可采用交叉验证法，选取交叉验证误差最小的 λ 值，再用全部数据拟合模型。Lasso 回归常用于特征选择，可将部分不重要的特征参数压缩至 0，实现稀疏化建模。

Table 2. Correlation coefficients of Lasso regression

表 2. Lasso 回归相关系数

对应大气指标	Lasso 相关系数	系数正负
PM2.5	0.13013	正
NO ₂	0.05838	正
PM10	0.00436	正
SO ₂	-0.01676	负
CO	0.00301	正
AQI	-0.00944	负
O ₃ _8 h	27.88165	正

我们选取 Lasso 回归(L1 正则化回归)的目的是特征筛选与模型简化，首先通过 Lasso 回归筛选对目标值有显著解释能力的特征，剔除冗余、无关指标，避免模型过拟合；其次通过正则化参数 α 控制特征惩罚强度，保留关键特征的同时，降低模型复杂度，提升后续建模(GM11 + SVR)的效率与稳定性；最后量化各大气指标对目标值的影响权重(即表格中 Lasso 相关系数)，明确核心影响指标，为预测建模提供针

对性依据。

训练本文模型选取最优正则化参数 $\alpha = 0.0001$ ，根据表 2 生成的 Lasso 回归相关系数，7 项指标相关系数均为非零，全量保留无剔除，表明了这 7 个指标均对目标值存在线性解释能力，无冗余特征。整体结论与上述相关性以及热力图表示具有一致性。

3. 模型

本文采用灰色预测和 SVR (支持向量回归) 构建空气质量预测模型，主要是通过两者的优势实现互补。灰色预测模型常应用于处理数据的不确定性与小样本特性，而 SVR 模型擅长于捕捉数据中的非线性规律，最终提升预测精度。

我们将灰色预测和 SVR 这两个模型组合，主要是因为其适用于我们所处理的空气质量数据集的特性，即样本量有限、影响因素复杂且存在非线性关联的特点。又因为空气质量数据受工业排放、气象条件、交通尾气等多种因素影响，往往存在数据缺失、波动剧烈的不确定性，且部分区域监测站点数量有限，难以获取大规模连续样本，应用于单一预测模型中。

3.1. 灰色预测算法

灰色预测法[7]用于对含有不确定因素的系统进行预测，在建立灰色预测模型前，我们要先对原始时间序列做数据处理，处理后所得的时间序列叫生成列。其以灰色模型为基础，其中 GM (1, 1) 模型最为常用，下面我们给与 GM (1, 1) 模型建立步骤[8]：

设原始非负单调时间序列为 $X^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n))$ ，其中 n 为样本数量， $x^{(0)}(k) > 0$ ($k = 1, 2, \dots, n$)，GM (1, 1) 模型的具体建立步骤如下：

1) 一次累加生成：对原始序列 $X^{(0)}$ 进行一次累加操作，生成新的时间序列 $X^{(1)}$ ，弱化原始数据的噪声与波动，挖掘整体变化趋势， $X^{(1)} = (x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n))$ ，其中：

$$x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i) \quad (k = 1, 2, \dots, n)$$

2) 构建紧邻均值生成序列：对一次累加序列 $X^{(1)}$ 构建紧邻均值生成序列 $Z^{(1)}$ ， $Z^{(1)} = (z^{(1)}(2), z^{(1)}(3), \dots, z^{(1)}(n))$ ，其中：

$$z^{(1)}(k) = \frac{1}{2}(x^{(1)}(k) + x^{(1)}(k-1)) \quad (k = 2, 3, \dots, n)$$

3) 建立 GM (1, 1) 灰色微分方程：以 $X^{(0)}$ 为因变量， $Z^{(1)}$ 为自变量，建立一阶线性灰色微分方程： $x^{(0)}(k) + ax^{(1)}(k) = b$ 其中， a 为发展系数，反映序列的变化趋势； b 为灰作用量，反映系统的输入输出关系。

4) 求解模型参数：采用最小二乘法求解参数向量 $\hat{\theta} = (a, b)^T$ ，计算公式为：

$$\hat{\theta} = (B^T B)^{-1} B^T Y$$

其中， B 为数据矩阵， Y 为常数向量：

$$B = \begin{pmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{pmatrix}, Y = \begin{pmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{pmatrix}$$

5) 建立预测模型: 对灰色微分方程求导, 得到白化微分方程: $\frac{dx^{(1)}(t)}{dt} + ax^{(1)}(t) = b$

求解白化微分方程, 得到一次累加序列 $X^{(1)}$ 的预测模型:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a} \right) e^{-ak} + \frac{b}{a} (k=1, 2, \dots, n)$$

6) 逆累加还原: 对一次累加预测序列 $\hat{x}^{(1)}$ 进行一次累减还原, 得到原始序列 $X^{(0)}$ 的预测值 $\hat{x}^{(0)}$, 即为 GM (1,1) 模型的最终预测结果:

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) (k=1, 2, \dots, n)$$

灰色预测算法中 GM (1, 1) 模型是其最常见的模型, 适配性较强, 在小样本数据集以及不确定性的数据集中预测效果具有一定的优势, 同时与传统预测模型相比, GM (1, 1) 有效的避免了小样本数据集下偏差大的问题。此外, 该模型结构简洁、计算难度较低、可操作性较强, 只包含了单变量的一阶线性微分方程, 降低了时间成本。但是, 其预测精确度并没有因为其结构简单有所降低, 反而能快速的捕捉到空气质量短期的变化规律, 误差小, 响应快实现精准输出。所以在本文中空气质量预测具有一定的适配性。

3.2. SVR 算法

SVR (Support Vector Regression, 支持向量回归) 是一种融合支持向量思想的回归分析方法, 核心在于通过特定的损失规则和支持向量选择, 实现对数据的拟合与回归预测[9]。

SVR 在进行数据拟合时, 并非追求模型输出与真实值完全一致(即传统回归中“无偏差”的目标), 而是引入“ ε -不敏感损失”的核心思路: 允许模型预测值与样本真实值之间存在一定范围(最多为 ε) 的偏差。在该偏差范围内, 认为预测结果准确且不计算损失, 仅当偏差超出 ε 时, 才对超出部分计算损失, 以此平衡模型的拟合精度与泛化能力。

SVR 的分析基于给定的训练数据集, 其标准形式为[10]:

$D = (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 其中 $x_i \in R^m$ 为第 i 个样本的 m 维特征向量, $y_i \in R$ 为第 i 个样本的真实输出值, n 为样本数量, SVR 的标准优化问题为:

$$\begin{cases} \min_{\omega, b, \xi_i, \hat{\xi}_i} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\xi_i + \hat{\xi}_i) \\ y_i - (\omega^T x_i + b) \leq \varepsilon + \xi_i \\ (\omega^T x_i + b) - y_i \leq \varepsilon + \hat{\xi}_i \\ \xi_i \geq 0, \hat{\xi}_i \geq 0, i=1, 2, \dots, n \end{cases}$$

其中:

ω 为特征系数向量, b 为偏置项;

$C > 0$ 为惩罚系数, C 越大, 对预测偏差超出 ε 的样本惩罚越重;

$\varepsilon \geq 0$ 为不敏感系数;

ξ_i 、 $\hat{\xi}_i$ 为松弛变量, 分别表示样本真实值超出上边界、下边界的偏差程度。

通过拉格朗日对偶变换, 可求解得到 SVR 的回归预测模型为:

$$f(x) = \sum_{i=1}^n (\alpha_i - \hat{\alpha}_i) K(x_i, x) + b$$

其中, α_i 、 $\hat{\alpha}_i$ 为拉格朗日乘子, 且 $\alpha_i \geq 0$ 、 $\hat{\alpha}_i \geq 0$ 。

SVR 是一种高效的机器学习方法, 在非线性数据拟合中体现出显著的优势。此外, 其泛化能力较强, 抗过拟合效果好。SVR 基于结构风险最小化原则, 通过引入惩罚系数控制模型复杂度, 无需依赖大量样本训练, 即便在样本有限或存在少量异常值的场景下, 也能保持稳定的预测效果, 与我们试验中采用小样本数据同样适配。

同时, SVR 鲁棒性优良, 对噪声数据敏感度低, 能够有效规避空气质量数据中异常监测值的干扰, 保证预测精度; 同时还与 GM (1, 1) 等灰色预测模型结合互补, 弥补单一模型短板, 进一步提升整体预测性能, 为我们整体研究提供核心价值。

3.3. GM (1, 1)-SVR 组合模型的运行流程

本研究新增组合模型的详细运行流程, 并明确核心数学表达式, 构建的 GM (1, 1)-SVR 组合预测模型, 以 GM (1, 1) 模型为基础, 挖掘空气质量时间序列的整体变化趋势, 再以 GM (1, 1) 模型的预测残差为研究对象, 利用 SVR 算法拟合残差的非线性规律, 最终将 GM (1, 1) 模型的预测值与 SVR 模型的残差修正值相加, 得到组合模型的最终预测结果, 核心数学表达式:

$$\hat{x}^{(0)}(k) = \hat{x}_{GM}^{(0)}(k) + \hat{e}(k)$$

其中:

$\hat{x}^{(0)}(k)$: 组合模型对原始序列的最终预测值;

$\hat{x}_{GM}^{(0)}(k)$: GM (1, 1) 模型对原始序列的初步预测值;

$\hat{e}(k)$: SVR 模型对 GM (1, 1) 预测残差的拟合值(残差 $e(k) = x^{(0)}(k) - \hat{x}_{GM}^{(0)}(k)$)。

具体运行流程如下:

数据预处理: 对原始空气质量监测数据进行归一化处理, 消除量纲差异, 提升模型训练效率;

GM (1, 1) 模型预测: 将预处理后的时间序列数据输入 GM (1, 1) 模型, 得到空气质量指标的初步预测值 $\hat{x}_{GM}^{(0)}$, 并计算预测残差 $e(k) = x^{(0)}(k) - \hat{x}_{GM}^{(0)}(k)$;

SVR 模型残差拟合: 以空气质量指标的特征向量为输入, 以 GM (1, 1) 模型的预测残差 $e(k)$ 为输出, 训练 SVR 残差修正模型, 得到残差的预测值 $\hat{e}(k)$;

组合模型预测结果: 将 GM (1, 1) 模型的初步预测值与 SVR 模型的残差预测值相加, 得到组合模型的最终预测值;

结果反归一化: 对组合模型的预测值进行反归一化处理, 还原为空气质量指标的实际物理值, 得到最终的空气质量预测结果。

4. 评价指标与实验设置

4.1. 评价指标

为全面、客观地评估 GM (1, 1)-SVR 组合模型的预测性能, 本研究选取五项核心评价指标, 从拟合精度、误差水平、方差解释能力三个维度开展模型评估, 各指标的定义与计算公式如下, 所有指标均基于 2025 年 1 月沈阳市空气质量监测数据训练生成:

平均绝对误差: 是衡量预测值与实际值偏差程度的直观指标, 计算所有样本预测值与实际值绝对偏差的算术平均值, 能够反映模型预测的平均误差水平, 计算公式为:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

中值绝对误差：通过先计算所有样本预测值与实际值的绝对偏差，再取这些偏差值的中位数，用于衡量模型预测偏差的稳健性水平，计算公式为：

$$MedAE = median\left(|y_1 - \hat{y}_1|, |y_2 - \hat{y}_2|, \dots, |y_n - \hat{y}_n|\right)$$

可解释方差值：衡量预测模型对实际数据波动规律的解释能力，反映预测值方差与实际值方差的契合程度，计算公式为：

$$EVS = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

R 方值：又称拟合优度，是衡量回归或预测模型对数据拟合程度的核心指标，反映预测值能够解释实际值变化的比例。计算公式为：

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

4.2. 实验设置

实验数据：2025 年 1 月~6 月沈阳市空气质量监测的 7 项核心指标数据，其中以 2025 年 1 月数据为训练集，用于模型的训练与参数调优；以 2025 年 2 月~6 月数据为历史测试集，用于验证模型的泛化能力。

实验环境：Python、机器学习库 Scikit-learn、数值计算库 Numpy、Pandas，可视化库 Matplotlib。

5. 结论

Table 3. Evaluation indicator values

表 3. 评价指标数值

评价指标	指标数值
训练集 R ² 得分	0.9745
历史数据 R ² 得分	0.9745
平均绝对误差	3.1439
中值绝对误差	2.0044
可解释方差	0.9747

表 3 评价指标数值是基于沈阳 2025 年 1 月份空气质量数据训练模型生成，通过五项核心评价指标进行全面评估模型性能，结果表明模型预测效果优异、稳定性强。

分析以上数据，训练集与历史数据 R² 得分均达 0.9745，说明模型对该时段数据拟合精度极高；可解释 97.45% 的空气质量指标变化，说明模型的泛化能力稳定；平均绝对误差为 3.1439，中值绝对误差为 2.0044，两项误差指标数值较小，是在可接受的范围之内，并且中值绝对误差比平均绝对误差较低些，体现出模型的抗干扰能力；可解释方差值达 0.9747，表明模型可以较精准的捕捉该阶段空气质量的波动规律。综上可得，该模型适配沈阳 2025 年 1 月空气质量预测场景，性能可靠。

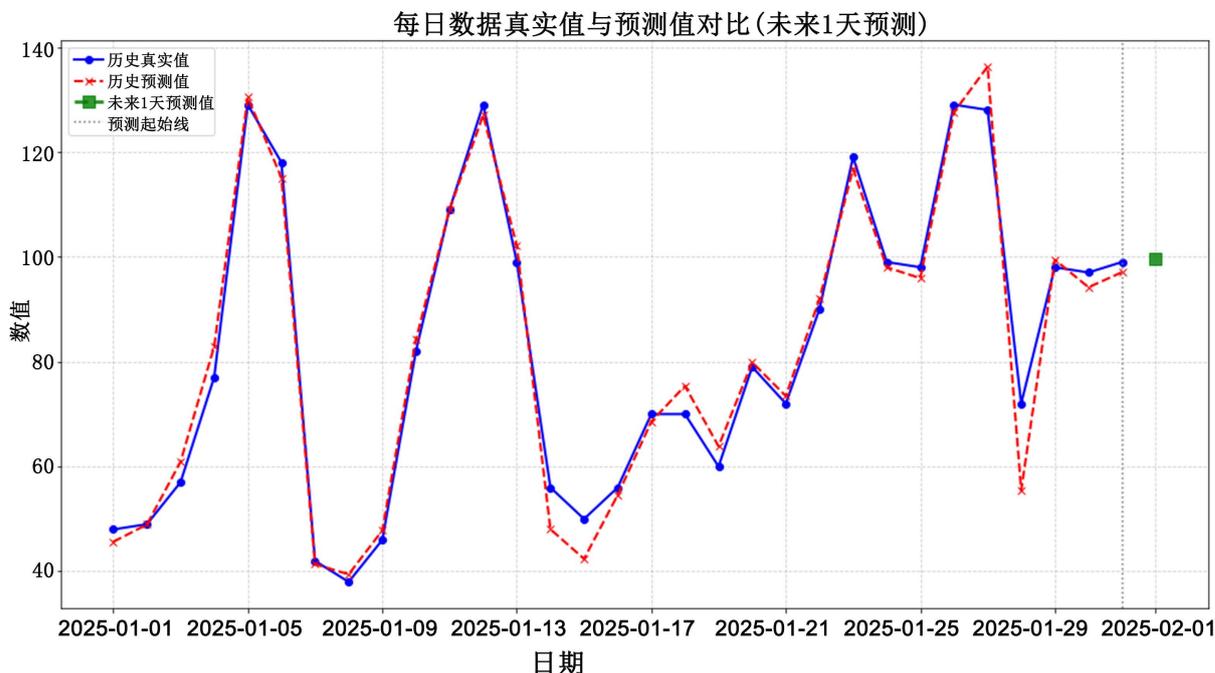


Figure 2. Comparison chart of actual values and predicted values

图 2. 真实值与预测值对比图

以上“每日数据真实值与预测值对比(未来 1 天预测)”图表(如图 2), 是通过沈阳 2025 年 1 月份训练灰色预测和 SVR 模型生成, 可以直观的展示拟合结果, 进一步进行分析。蓝色实线为历史真实值, 红色虚线为历史预测值, 绿色方块为预测后 1 天(2 月 1 日)预测值。

观察可发现代表“历史预测值”的红色虚线与代表“历史真实值”的蓝色实线高度重合, 精准跟踪了真实值的所有峰值、谷值及波动趋势, 如 1 月 5 日、13 日、25 日的峰值处拟合度几乎完美, 印证了灰色预测与 SVR 模型对历史数据的拟合精度极高, 抗干扰能力强, 与此前 R^2 得分 0.9745 的量化结论完全吻合。对于未来 1 天的预测(2 月 1 日), 绿色方块标记的预测值与当前真实值的变化趋势保持一致, 进一步体现了灰色预测与 SVR 模型的稳定泛化能力。

整体来看, 灰色预测与 SVR 模型在历史拟合与未来预测中均表现出色, 能够精准捕捉沈阳 1 月空气质量的复杂波动规律, 虽然整体拟合较好, 但是也会发现个别数值拟合效果不好, 分析因空气质量易受偶发事件影响, 比如周末或者节假日可能会导致空气质量的数据异常, 拟合效果不好。分析以上结果, 未来, 我们将在该灰色预测与 SVR 模型中加入参数, 控制节假日预测空气质量拟合的波动性。

参考文献

- [1] Sushmitha and Vijayalaxmi (2026) Air Quality Prediction Using Computational Intelligence. In: Iglesias, A., *et al.*, Eds., *Information Systems for Intelligent Systems*, Springer, 206-215. https://doi.org/10.1007/978-3-032-12987-1_20
- [2] Joshi, S., Parikh, A., Gandhi, R. and Bhowmick, K. (2025) Meteorology Driven Air Quality Prediction. In: Gunjan, V.K. and Zurada, J.M., Eds., *Proceedings of 5th International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, Springer, 15-26. https://doi.org/10.1007/978-981-97-8861-3_2
- [3] Kumar, S., Mishra, A., Pandey, H.K., Tiwari, U.K. and Harnisha, M. (2025) Air Quality Prediction Using Ensemble Learning. In: Panda, S.K., *et al.*, Eds., *Computing, Communication and Learning*, Springer, 60-72. https://doi.org/10.1007/978-3-031-79041-6_5
- [4] Iqbal, A. and Mukherjee, N. (2025) A Systematic Review and Comparative Study of Machine Learning Techniques for

- Air Quality Prediction. *Water, Air, & Soil Pollution*, **236**, Article No. 789. <https://doi.org/10.1007/s11270-025-08410-3>
- [5] Abbasi, M.T., Alesheikh, A.A., Lotfata, A. and Azizi, Z. (2025) Hybrid Graph Convolutional Networks for Air Quality Prediction: A Systematic Review of Foundations, Challenges, and Opportunities. *International Journal of Environmental Science and Technology*, **22**, 13327-13362. <https://doi.org/10.1007/s13762-025-06599-8>
- [6] Desai, T., Kapadia, S., Halani, M., Zinzuwadia, P., Shah, K., Shah, M., *et al.* (2025) Comparative Analysis of Machine Learning Algorithms for Air Quality Index Prediction. *Machine Learning for Computational Science and Engineering*, **1**, Article No. 14. <https://doi.org/10.1007/s44379-025-00014-2>
- [7] Zeng, B. and Shi, Z. (2024) Grey Prediction Methods and Their Applications. Springer.
- [8] 李燕斌, 张久菊, 肖俊明. 基于指数平滑法的灰色预测模型[J]. 中原工学院学报, 2015(4): 1-4.
- [9] 郑洋洋, 白艳萍, 续婷. 基于 SARIMA-SVR 组合模型的质量指数预测[J]. 河北工业科技, 2019(6): 436-441.
- [10] S, A., Debnath, M.K. and R, K. (2024) Statistical and Machine Learning Models for Location-Specific Crop Yield Prediction Using Weather Indices. *International Journal of Biometeorology*, **68**, 2453-2475. <https://doi.org/10.1007/s00484-024-02763-w>