文章编号:1671-7872(2025)06-0669-09

基于支持向量机的夏热冬冷地区农村住宅能耗混合预测模型

刘峻江,孙亚东,黄志甲
(安徽工业大学 建筑工程学院,安徽 马鞍山 243032)

摘要: 针对夏热冬冷地区农村住宅建筑能耗预测困难的问题,提出一种基于支持向量机(support vector machine, SVM)的混合预测模型。通过采集典型农村住宅的建筑参数、气象参数、行为参数、设备参数及年能耗数据构建初始数据集,采用包含显著性分析、共线性分析、随机森林敏感性分析和后向逐步回归方法的递进筛选框架,从29个候选变量中筛选出10个关键变量,显著降低模型复杂度。通过融合白箱模型理论计算数据与黑箱模型实测数据构建SVM的预测混合模型,并采用基于网格搜索与交叉验证的联合策略优化模型关键参数以提高模型性能。验证结果表明:本文模型决定系数(R^2)为0.914,均方根误差变异系数(CVRMSE)为0.163,在保证预测精度的同时实现了模型复杂度的最优平衡。本研究提出的变量筛选与数据融合策略,有效解决了该地区农村住宅因设计参数缺失和能耗数据不足导致的预测难题。
关键词: 农村住宅;夏热冬冷地区;变量筛选;支持向量机;能耗预测;混合模型;机器学习;建筑能效
中图分类号: TU 111.195 **文献标志码:** A **doi:** 10.12415/j.issn.1671-7872.25024

Hybrid Prediction Model for Rural Residential Energy Consumption in Hot-summer and Cold-winter Regions Based on Support Vector Machine

LIU Junjiang, SUN Yadong, HUANG Zhijia
(School of Architectural Engineering, Anhui University of Technology, Maanshan 243032, China)

Abstract: To address the challenge of energy consumption forecasting for rural residential buildings in hot-summer and cold-winter areas, a hybrid model based on support vector machine (SVM) was proposed. An initial dataset was constructed by collecting building parameters, meteorological parameters, behavioral parameters, equipment parameters, and annual energy consumption data from typical rural residences. A progressive screening framework incorporating significance analysis, collinearity analysis, random forest sensitivity analysis, and backward stepwise regression method was employed to select 10 key variables from 29 candidate variables, significantly reducing model complexity. A hybrid SVM prediction model was established by integrating theoretical calculation data from white-box models with measured data from black-box models, and a joint strategy combining grid search and cross-validation was adopted to optimize key model parameters for performance enhancement. The validation results demonstrate that the proposed model achieves a coefficient of determination (R^2) of 0.914 and a coefficient of

收稿日期: 2025-03-13
基金项目: 国家自然科学基金项目(51608001)
作者简介: 刘峻江(1999—),男,江苏连云港人,硕士生,主要研究方向为建筑节能与绿色建筑技术。
通信作者: 黄志甲(1963—),男,安徽安庆人,博士,教授,博士生导师,主要研究方向为建筑节能与绿色建筑技术。
引文格式: 刘峻江,孙亚东,黄志甲. 基于支持向量机的夏热冬冷地区农村住宅能耗混合预测模型[J]. 安徽工业大学学报(自然科学版),2025,42(6):669-677.

variation of root mean square error (CVRMSE) of 0.163, maintaining prediction accuracy while realizing optimal balance in model complexity. The variable screening and data fusion strategies developed in this study are proved to effectively address the prediction challenges caused by missing design parameters and insufficient energy consumption data in rural residences of this region.

Keywords: rural buildings; hot-summer and cold-winter regions; variable selection; support vector machine(SVM); energy consumption prediction; hybrid model; machine learning; building energy efficiency

在乡村振兴战略持续推进和农村居民生活水平不断提升的双重背景下，农村住宅能耗呈现出显著增长态势，然而其能耗预测却面临基础数据双重缺失的现实挑战：一方面，建筑围护结构参数等物理信息记录不完整；另一方面，用能行为数据的系统采集也较为欠缺。在传统预测方法中，基于物理方程的白箱模型因高度依赖精确的建筑参数而难以适用于农村住宅的实际情况；而纯粹数据驱动的黑箱模型则因可用数据稀疏而易出现过拟合问题，泛化能力明显不足。这一技术瓶颈不仅制约了建筑节能领域差异化政策工具箱的构建，也影响了光伏屋顶、被动式改造等低碳技术在农村地区的有效推广。因此，建立适用于低数据条件的新型能耗预测范式，已成为协调农村民生改善与碳减排目标之间矛盾的关键切入点，也是推动“双碳”目标向乡村层面深化实施的重要基础性工作。

当前研究表明白箱与黑箱模型通过整合物理原理与数据驱动方法的优势，在参数缺失和数据稀疏条件下展现出显著潜力[1]，但现有方法仍存在明显局限。例如：Su 等[2] 提出的贝叶斯回归–热惯性校正模型通过等效温度变量表征建筑热惯性，弥补了物理参数缺失对预测精度的影响，但需 18 个输入变量 (参数缺失率达 38%)，且预测精度仍有限，其模型决定系数 R^2=0.85，均方根误差变异系数 CVRMSE=0.22；Qiao 等[3] 结合 SARIMA 季节特征与支持向量机 (support vector machine, SVM) 的非线性拟合能力，将输入变量精简至 15 个 (R^2=0.88, CVRMSE=0.19)，在历史数据有限的场景下提升了建筑能耗预测的鲁棒性，但对参数缺失高达 42% 的适应能力仍显不足；Liang 等[4] 基于领域知识分解的混合模型虽提升至 R^2=0.90，但需 22 个高复杂度输入变量 (缺失率高达 55%)。这些模型普遍面临维度输入高与缺失容忍低的困境。值得注意的是，袁鹏丽等[5] 针对农村住宅用电行为的研究表明，通过分解供暖与非供暖季负荷特征，混合模型可有效捕捉居民用能行为的时空差异性，为农村场景的模型构建提供了重要参考。此外，动态与静态数据融合、物理方程与机器学习耦合[6] 等方法均通过整合多维度信息，在降低对完整参数依赖的同时，也提高了模型对农村住宅的能耗预测性能。针对以上问题，构建一种基于 SVM 的白箱–黑箱混合模型，对夏热冬冷地区农村住宅建筑能耗进行预测，以期为推动农村建筑节能降碳实践提供理论基础和技术支撑。

1 数据采集与模型构建

针对夏热冬冷地区农村住宅能耗预测，本文采用数据驱动与理论计算的混合建模方法，通过多源数据采集、融合与特征筛选构建输入变量集，最终建立优化的 SVM 能耗预测模型。

1.1 数据采集

为构建适用于夏热冬冷地区农村住宅的能耗混合预测模型，需首先识别关键能耗影响因素并解决数据缺失条件下的输入变量优化问题。本研究通过整合实地调查与理论计算两类数据，建立包含建筑参数、气象参数、设备参数和行为参数[7]4 个维度的综合数据集。其中：实地调查数据主要包括建筑参数 (如建筑年代、建筑面积等)、行为参数 (如人口数、家庭收入等) 以及设备参数 (如家电数、采暖空调数等)；理论计算数据涵盖建筑参数 (如建筑面积、窗地比等)、气象参数 (如温度、太阳辐射等) 及设备参数 (设备效率)。通过对上述两类数据的整合，形成如图 1 所示的数据集，从而为后续模型构建提供全面可靠的数据基础。

1.1.1 实地调查数据

马鞍山市地处安徽省东部，属于典型的夏热冬冷地区 IIA 子区，其气候表现为夏季高温高湿而冬季寒冷干燥的显著特征，具体表现为夏季月均气温不低于 28 ℃、冬季月均气温不高于 5 ℃，且年温差变化较为显著，完全符合气候分区的标准定义，因此本研究选取该市作为代表性样本区域；在此基础上，本文通过实地调查采集了当地农村住宅的四类基础数据，包括建筑参数、气象参数、行为参数、设备参数，并同步收集了相应的住宅年度能耗数据，用于后续模型构建与分析。

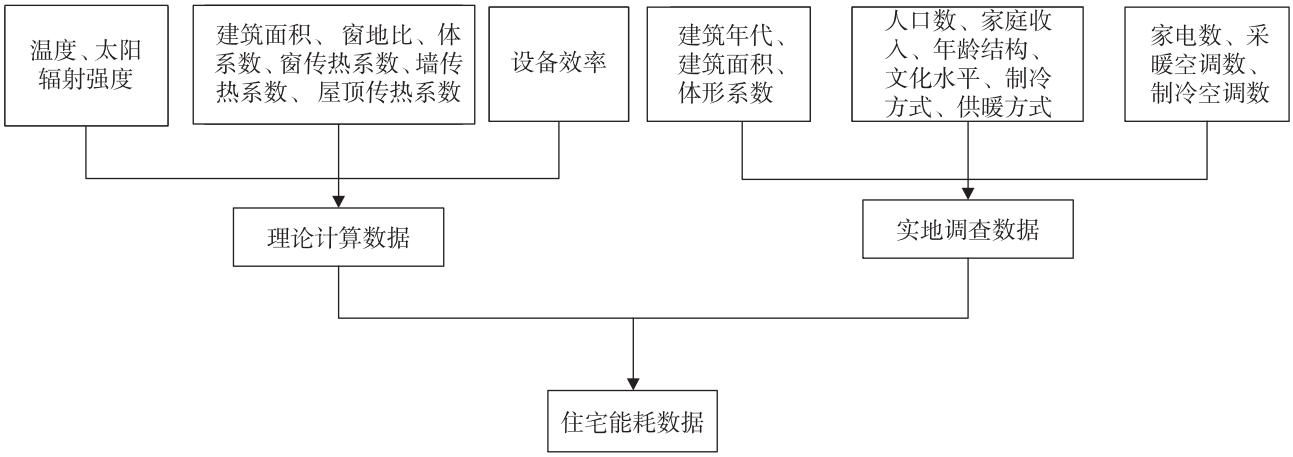

图1　农村住宅能耗数据集构建框架

Fig. 1　Construction framework for rural residential energy consumption dataset

为确保研究数据的可靠性, 本文采用多维度验证策略对采集的数据进行检验。首先基于 G*Power 3.1 软件进行样本效能分析, 设定效应量 f^2=0.15、显著性水平 α=0.05、预测变量数 k=10 个, 经计算得出最小样本需求量为 105 户, 实际获取有效样本量为 205 户, 经异常值筛除后保留样本量 200 户, 最终统计效能达 0.99, 完全满足研究要求。其次开展样本代表性验证, 将采集数据与马鞍山市 2020 农村普查数据进行比对 (关键指标如表 1), 数据显示: 老年人口比例 (22.4% vs. 25.7%)、低收入家庭占比 (66.5% vs. 61.2%) 及 2000 年前建造住宅比例 (92.0% vs. 88.0%) 等核心指标均无统计学显著差异 (卡方检验 χ^2=2.34, p=0.12), 证实样本具有区域代表性。最后进行系统的信效度检验, 信度方面采用 Cronbach's α 系数 (信度系数 α=0.79) 和重测信度 (intraclass correlation coefficient, ICC=0.72) 评估内部一致性, 效度方面通过探索性因子分析 (exploratory factor analysis, EFA) 验证结构效度 (累积方差解释率达 68%), 采用效标关联法分析显示住户自报电耗与电力公司记录的 Pearson 相关系数 r=0.65(p<0.01), 证实数据质量达到建模标准。

表 1　样本与普查数据的卡方检验匹配性分析

Tab. 1　Matching analysis using the Chi-square test for sample and census data

特征	样本分布/%	普查分布/%	χ^2检验p值	一致性结论
年龄：老年(>60岁)	22.4	25.7	0.12	无显著差异
家庭年收入：<5万元	66.5	61.2	0.08	无显著差异
建筑年代：2000年前	92.0	88.0	0.21	无显著差异

1.1.2　理论计算数据

在建筑能耗预测领域, 传统的环境控制能耗算法主要包括当量满负荷小时数法、BIN 法、度日数法等方法。其中, 度日数法通过建立室内外温差与能耗之间的线性回归模型进行估算, 在数据可获得性有限且对预测精度要求中等的应用场景中具有独特优势, 因而被广泛采用。基于此, 本研究采用度日数法计算理论制冷和供暖能耗。根据文献 [8–9] 的规范定义,将采暖度日数 (heating degree days, HDD) 与制冷度日数 (cooling degree days, CDD) 分别界定为冬季供暖与夏季制冷需求的量化指标; 结合本课题组[10] 关于皖南农村住宅适应性热舒适的研究成果, 并参考李俊鸽等[11] 建立的夏热冬冷地区热舒适气候适应性模型, 最终确定 HDD 的室内基准温度为 14 ℃, CDD 为 30 ℃, 该参数设置考虑了当地居民动态热舒适特征。具体的能耗计算模型表述如下:

$$Q_h = \frac{D_h KSAH}{\eta_h} \tag{1}$$

$$Q_c = \frac{D_c KSAH}{\eta_c} \tag{2}$$

式中: D_h, D_c 分别为采暖与制冷度日数; K 为建筑综合传热系数; η_h, η_c 分别为供暖与制冷设备效率, 取值依据设备类型及使用占比的调研结果进行设定; S 为建筑体形系数; A 为建筑面积; H 为建筑层高。

1.2　实测与理论数据融合

基于信息互补性原理[12], 本研究融合理论计算数据 (建筑面积、气温和建筑围护结构热工性) 与实地调查数据构建模型输入数据集, 其中理论制冷与

理论供暖能耗作为混合模型的关键输入变量。针对数据缺失问题，采用差异化插补策略：对连续变量采用中位数或均值填补，对分类变量采用多重插补法处理；在数据预处理阶段[13]，对正态分布变量实施 *Z*-score 标准化，对异常值数据采用 Robust 标准化，并通过 Kolmogorov–Smirnov 检验确保插补后数据分布的一致性，将变量间最大相关性差异控制在 5% 以内，最终形成时空对齐的融合数据集，有效克服了单一数据源的局限性。

1.3 变量筛选四步法

传统变量筛选方法 (如逐步回归法、单指标筛选法) 存在模型稳定性差、无法处理复杂关系等不足[14]。为此，本文提出四步递进式变量筛选法，通过显著性、共线性、敏感性及后向逐步回归分析递进筛选，有效提升建筑能耗预测中变量筛选的可靠性。

1.3.1 显著性分析

基于 Simca 软件中的 PLS 模型[15] 对 14 项能耗影响因素进行评估，这些因素包括建筑面积、体形系数、建筑年代、人口数、家庭收入、年龄结构、文化水平、制冷方式、供暖方式、家电数、供暖空调数、制冷空调数、理论制冷能耗、理论供暖能耗；通过变量投影重要性指标[16](variable importance in projection, VIP) 量化各自变量的贡献度，针对农村住宅能耗数据特有的行为异质性和测量噪声特性，参考 Chong 等[17] 的探索性分析策略，将 VIP 阈值优化设置为 0.8，该阈值较常规标准 (VIP>1.0) 显著提高了潜在影响因素的保留率；对于 VIP>0.8 的变量，表示其对因变量的影响显著，反之影响不显著。

1.3.2 共线性分析

为进一步提升模型稳定性，需对显著变量进行共线性诊断。采用 SPSS26 分析软件对显著性筛选后的变量进行多重共线性分析，通过计算方差膨胀因子 (variance inflation factor, VIF) 评估变量间的共线性程度。其计算公式如下：

$$\mathrm{VIF}=\frac{1}{1-C^2} \tag{3}$$

式中 C^2 为自变量与自变量之间的相关系数。针对农村住宅能耗数据的特点，参考 O'brien[18] 的行为科学研究准则，本文采用 VIF<10 的阈值，该阈值设置既有效控制了共线性干扰，又保留了具有重要理论价值的变量。当检测到 VIF>10 的变量时，表示该变量与其他自变量相关，直接删除冗余变量或采用其他指标替代原有共线性的自变量，最后获得无严重共线性的自变量。

1.3.3 敏感性分析

在完成共线性诊断的基础上，采用随机森林法[19] 进行敏感性分析，通过计算各变量的均方误差减少量 (increase in mean squared error, %InMSE) 来量化其对能耗预测模型的贡献度。该方法通过随机置换特征值后观测模型均方误差 (mean squared error, MSE) 的变化程度来评估变量重要性，%InMSE 值越低表明该变量对模型输出的影响越显著[20]。相较于传统参数敏感性分析方法，随机森林法能够有效捕捉变量间的非线性交互作用，尤其适合于处理农村住宅能耗数据中存在的复杂特征关系[19]，从而为后续变量筛选与模型精简提供科学依据。

1.3.4 后向逐步回归分析

基于敏感性所获变量排序结果，针对低影响力变量存在的模型精度贡献有限但计算成本较高的问题，采用逐步回归法进行变量优化选择：以包含全部显著变量的全模型为起点，通过迭代过程逐步删除贡献最弱的变量，直至满足预设的停止准则[21]。这种自上而下的筛选策略在充分保留高敏感性变量信息贡献的基础上，有效识别并移除了冗余变量，最终在维持预测精度基本不变的前提下，将模型输入维度压缩至关键变量集合，从而通过降低特征空间复杂度与共线性干扰，显著增强模型结构稳定性与运算效率。

1.4 基于 SVM 的能耗预测混合模型的构建

针对农村住宅能耗数据有限的特点，本研究通过核函数改进和参数优化策略提升 SVM 在数据有限条件下的预测性能。传统 SVM 虽能通过核函数实现非线性映射，但在处理多源异构的建筑能耗数据时对理论计算数据蕴含的先验知识利用不足，为此构建融合数据驱动与理论模型的混合预测框架。

1.4.1 核函数的改进设计

选用径向基核函数 (radial basis function, RBF) 作为基础核函数，其表达式如下：

$$K_{\mathrm{RBF}}(x_i,\ x_j)=\exp(-\gamma\|x_i-x_j\|^2) \tag{4}$$

式中：$K_{\mathrm{RBF}}(x_i,x_j)$ 为径向基核函数；x_i,x_j 为输入特征，表示样本 i 和样本 j 的特征值；γ 为核函数的带宽参数。

RBF 核的优势在于能够灵活拟合复杂非线性关系，尤其适用于捕捉农村住宅能耗数据中建筑参数与行为参数的交互效应。而理论能耗作为基于物理模型计算得出的关键影响因素，为有效融合此类先验知识以强化模型可解释性和对物理规律的遵循性，文中引入物理–数据混合核函数，将理论计算能耗作为先验信息嵌入核空间，其具体表达式如下：

$$K_{\mathrm{hybrid}}=\lambda K_{\mathrm{RBF}}(x_i,x_j)+(1-\lambda)K_{\mathrm{phys}}(Q_{\mathrm{h}},Q_{\mathrm{c}}) \tag{5}$$

式中：λ 为混合权重，文中将其视为待优化的超参数；K_{RBF} 为径向基核函数；K_{phys} 为基于理论制冷与供暖

能耗的线性核函数; Q_h, Q_c 分别为理论供暖与理论制冷能耗。

1.4.2　参数优化策略

为提高模型性能并确保泛化能力, 采用基于网格搜索与交叉验证的联合策略对模型关键参数进行优化, 主要步骤如下:

1) 数据集划分。采用分层随机抽样将融合数据集按 8∶2 划分为训练集和独立测试集, 这一方法确保了关键特征在两组间的分布一致性, 从而避免了数据划分可能带来的评估偏差[22]。

2) 参数空间探索。采用网格搜索算法在由惩罚系数 C、核函数参数 σ^2、不敏感损失参数 ε 所构成的三维参数空间中进行穷举寻优。

3) 性能评估与模型数选择。于训练集上采用三折交叉验证来评估不同参数组合的性能, 以综合性能最优为目标的函数作为评估指标, 并记录每一折验证结果的平均值作为最终性能。

4) 早停机制。为提升优化效率, 引入早停机制以监控验证集上的损失变化, 当连续五轮网格点搜索后验证损失均未出现显著下降时, 则提前终止当前参数路径的深度搜索, 并将计算资源转向其他潜在更优的参数区域。

5) 最优参数确定。当网格搜索全部完成或提前终止后, 最终选择在交叉验证中能同时实现最低平均 CVRMSE 与最高平均 R^2的参数组合, 作为最优模型配置。

综上, 本文构建模型的预测流程如图 2。

1.5　模型评价指标

决定系数 (R^2) 用于量化模型对能耗变化的解释能力 (0~1 区间, 越接近 1 表明拟合优度越高), 均方根误差变异系数 (coefficient of variation of root mean square error, CVRMSE) 反映预测值与真实值的标准化离散程度, 平均相对误差 (mean relative error, MRE) 直接表征预测偏差的百分比水平 (工程应用要求<18%[23])。这 3 个指标分别从方差解释度、绝对误差规模和相对误差幅度角度评估预测结果与建筑实际运行能耗的吻合程度, 可有效衡量模型的准确性和有效性。文中选取以上 3 个评价指标来评估预测模型的性能。其计算公式分别如下:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2} \tag{6}$$

$$\mathrm{CVRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}}{y_i} \tag{7}$$

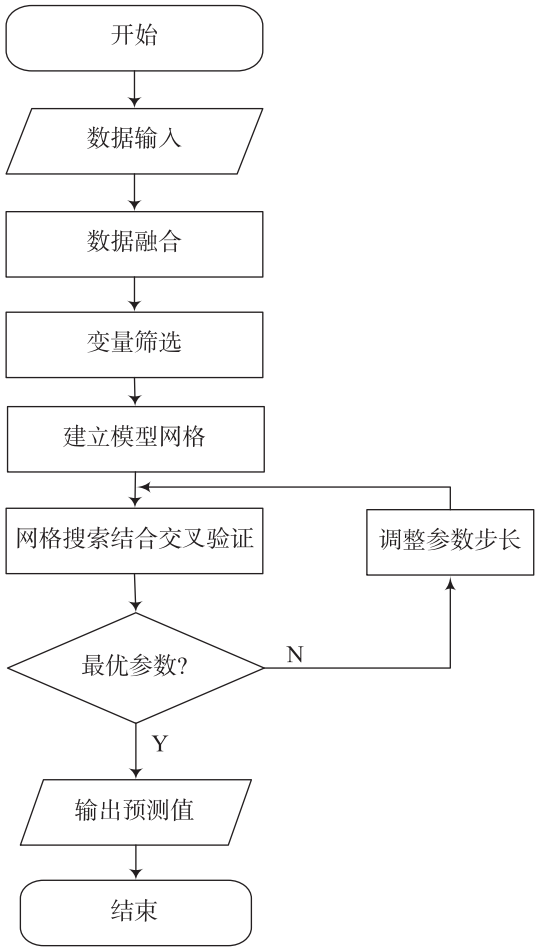

图2　本文构建模型的流程图

Fig. 2　Flowchart of the model construction process in this study

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{8}$$

式中: i 代表样本序号, 取值范围为从 1~n; y_i 为真实的建筑能耗; $\hat{y}_i$为预测的建筑能耗; $\bar{y}_i$为平均建筑能耗; n 为样本个数。

2　模型预测与结果分析

基于所构建的混合模型框架, 首先明确模型的初始输入变量; 随后通过显著性分析、共线性分析与敏感性分析, 识别关键能耗驱动因素并确定其影响权重, 同时结合后向逐步回归法对输入维度进行精简; 最后评估不同变量组合下 SVM 模型的预测性能, 并通过超参数优化确定最终的最优模型配置。

2.1　模型输入变量

基于图 1 所示的农村住宅能耗数据集构建框架, 选取 29 个关键变量作为预测混合模型的初始输入变量。其中调查数据为黑箱模型的输入, 计算数据为白箱模型的输入。经数据标准化和特征对齐的融合预处理, 最终形成的变量体系如表 2。

表 2 能耗预测混合模型输入变量选取

Tab. 2 Selection of input variables for energy consumption prediction hybrid model

数据来源	变量名称	输入变量
实地调查	建筑面积	X_1
	体形系数	X_2
	建筑年代	X_3(1950—1970年)，X_4(1971—1980)，X_5(1981—1990年)，X_6(1991—2000年)，X_7(2001—2010年)，X_8(2010年以后)
	人口数	X_9
	家庭收入	X_{10}
	年龄结构	X_{11}(少年，<18岁)，X_{12}(青年，18~60岁)，X_{13}(老年，>60岁)
	文化水平	X_{14}(小学及以下)，X_{15}(初中)，X_{16}(高中)，X_{17}(大学)，X_{18}(其他)
	制冷方式	X_{19}(风扇)，X_{20}(空调)，X_{21}(风扇+空调)
	供暖方式	X_{22}(空调)，X_{23}(电暖器)，X_{24}(空调+电暖器)
	家电数	X_{25}
	采暖空调数	X_{26}
	制冷空调数	X_{27}
理论计算	制冷能耗	X_{28}
	供暖能耗	X_{29}

注：X_{18} 包括未接受过全日制学历教育，但通过自学、家庭教育等途径获得知识的群体；接受过非全日制学历教育的职业技能培训 (如短期技工培训、家政服务培训等)，但未取得国家认可学历证书的群体。

2.2 模型变量筛选结果

采用 1.4.1 所述方法对表 2 所示的变量进行显著性分析，其 VIP 分析结果见图 3。

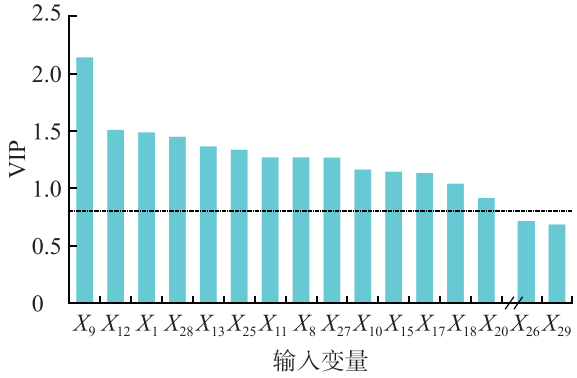

图3 变量的显著性筛选结果

Fig. 3 Outcome of significant variable screening

在图 3 中，断点之前变量的 VIP 值均大于 0.8，断点之后变量的 VIP 值均小于 0.8；所有变量中，X_2，X_3~X_8，X_{22}~X_{24}，X_{26}，X_{29} 均未对农村住宅能耗产生显著影响；X_9(人口数)、X_{12}(青年)、X_1(建筑面积)、X_{28}(理论制冷能耗) 显著影响农村住宅能耗，说明农村住宅建筑能耗与人口数、青年人数、建筑面积具有很强的相关性；X_{28} 显著影响能耗，说明农村地区夏季使用空调的频率一定程度上是符合季节变化规律的。

进一步地，对初步筛选出的 13 个变量 (X_9, X_{12}, X_1, X_{28}, X_{13}, X_{25}, X_{11}, X_{27}, X_{10}, X_{15}, X_{17}, X_{18}, X_{20}) 进行多重共线性分析，结果见表 3。表 3 表明：X_1(建筑面积) 和 X_{28}(理念制冷能耗) 的 VIF 值显著大于 10，说明两者存在严重共线性。经分析，X_{28} 同时包含气象和建筑参数，对总建筑能耗的影响更大且已涵盖 X_1 特征，故保留 X_{28} 而剔除冗余变量 X_1，最终确保所有变量的 VIF 值均满足小于 10 的共线性控制标准。

表 3 显著变量的共线性分析结果

Tab. 3 Collinearity analysis result of significant variables

显著影响变量	VIF
X_{17}	1.430
X_{18}	1.729
X_{20}	1.746
X_{11}	1.846
X_{10}	2.199
X_{27}	2.675
X_{25}	2.802
X_{15}	2.870
X_9	2.967
X_{13}	3.778
X_{12}	5.063
X_{28}	45.270
X_1	45.607

剔除 X_1 后对剩余变量进行共线性分析，结果见表 4。由表 4 可知：剔除 X_1 后剩余变量的 VIF 均小于 10，表明筛选后的显著影响变量间已不存在严重的多重共线性问题。

表 4 剔除 X_1 后剩余变量的共线性分析结果

Tab. 4 Collinearity analysis result of remaining variables after excluding X_1

显著影响变量	VIF
X_{17}	1.428
X_{18}	1.680
X_{28}	1.694
X_{20}	1.744
X_{11}	1.824
X_{10}	2.152
X_{27}	2.674
X_{15}	2.783
X_{25}	2.798
X_9	2.965
X_{13}	3.677
X_{12}	5.031

通过显著性检验和共线性分析的双重筛选, 最终确定 12 个显著影响且独立的能耗影响变量。基于随机森林法的敏感性分析 (训练集∶测试集=7∶3) 显示 (如图 4), 上述 12 个变量的特征重要性排序为 X_9>X_{12}>X_{28}>X_{25}>X_{10}>X_{11}>X_{18}>X_{15}>X_{13}>X_{17}>X_{27}>X_{20}。X_{25}(家电数) 对住宅能耗影响较大, 在夏热冬冷地区复合气候参数 (X_{28}) 和热工性能参数 (X_{25}) 对农村住宅能耗具有显著影响。该地区虽存在冬季采暖需求, 但相较于北方严寒地区, 其采暖强度较低且具有明显的间歇性特征。实地调研显示, 当地农户多采用空调、电暖器等灵活采暖方式, 且仅在极端低温时段使用, 这种用能模式与参数敏感性分析结果相互印证, 共同解释了 X_{28}(理论制冷能耗) 和 X_{25}(家电数) 在能耗预测模型中的重要性。

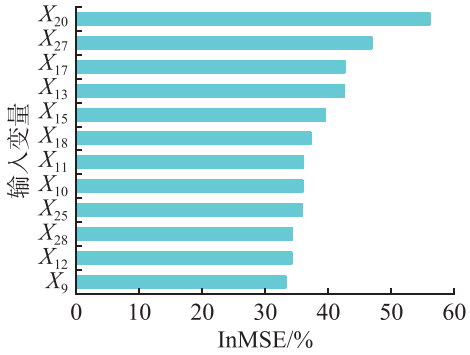

图4　基于 %InMSE 的变量敏感性排序结果

Fig. 4　Variable sensitivity ranking based on %InMSE

进一步采用回归法剔除次要变量, 最终生成 12 组不同维度的输入组合, 结果见表 5。

表 5　模型输入变量组合

Tab. 5　Combination of model input variables

模型编号	变量组合
12	X_{20}, X_{27}, X_{17}, X_{13}, X_{15}, X_{18}, X_{11}, X_{10}, X_{25}, X_{28}, X_{12}, X_9
11	X_{27}, X_{17}, X_{13}, X_{15}, X_{18}, X_{11}, X_{10}, X_{25}, X_{28}, X_{12}, X_9
10	X_{17}, X_{13}, X_{15}, X_{18}, X_{11}, X_{10}, X_{25}, X_{28}, X_{12}, X_9
9	X_{13}, X_{15}, X_{18}, X_{11}, X_{10}, X_{25}, X_{28}, X_{12}, X_9
8	X_{15}, X_{18}, X_{11}, X_{10}, X_{25}, X_{28}, X_{12}, X_9
7	X_{18}, X_{11}, X_{10}, X_{25}, X_{28}, X_{12}, X_9
6	X_{11}, X_{10}, X_{25}, X_{28}, X_{12}, X_9
5	X_{10}, X_{25}, X_{28}, X_{12}, X_9
4	X_{25}, X_{28}, X_{12}, X_9
3	X_{28}, X_{12}, X_9
2	X_{12}, X_9
1	X_9

2.3　SVM 超参数优化结果

在模型优化过程中, 采用后向逐步回归法确定变量组合后需通过系统化的超参数寻优策略, 综合评估不同输入变量组合对模型预测性能的影响, 最终在保证预测精度的同时实现模型简化。基于 SVM 模型的优化标准设定为: 在决定系数 (R^2)>0.8 且均方根误差变异系数 (CVRMSE)<0.3 的条件下, 追求 R^2 最大化和 CVRMSE 最小化的参数组合, 该标准可作为模型可靠性的判定依据。图 5 所示为不同变量组合对模型能耗预测性能的影响。

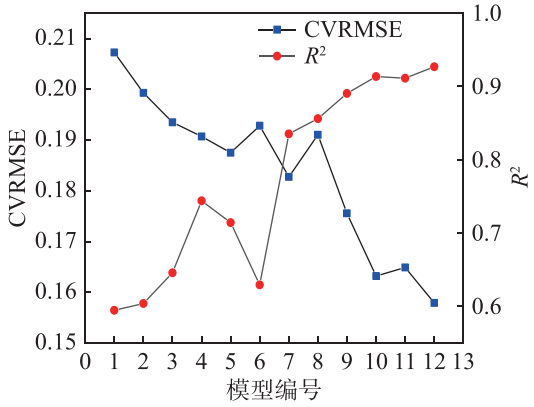

图5　不同变量组合对模型能耗预测性能的影响

Fig. 5　Impact of variable combinations on building energy prediction performance

由图 5 可见: 随着输入变量增加, 模型决定系数 (R^2) 呈上升趋势, 而 CVRMSE 呈先降后升再降的趋势; 其中模型编号 10 仅采用 10 个关键变量, 同时实现了 CVRMSE 最小值 (0.163) 和较高 R^2(0.914) 的优化目标。综合考虑预测精度与模型复杂度, 该模型达成了二者的最优权衡。因此, 最终从初始 29 个变量中筛选确定 X_{17}, X_{13}, X_{15}, X_{18}, X_{11}, X_{10}, X_{25}, X_{28}, X_{12}, X_9 作为混合模型的输入变量集。

3　模型验证

为评估本文混合模型在实际应用中的预测性能, 选用最优模型 (模型编号 10) 开展验证: 通过对比模型预测值与实际能耗数据的吻合程度验证模型预测精度; 基于预测误差的分布分析模型系统性偏差; 借助外推测试评估其在数据分布外的泛化能力。

3.1　模型适应性和鲁棒性验证

模型预测与 200 户实测能耗数据对比如图 6, 误差分析结果如图 7。图 6 显示: 实测数据 (E_a) 与预测值 (E_p) 趋势高度一致, 数据点紧密分布于拟合曲线周围 (相关系数 r^2=0.85), 验证了模型预测性能良好。图 7 表明: 绝多数样本的能耗预测相对误差在 30% 以内, 经计算其平均 MRE 为 14.57%, 满足农村节能改造的 (MRE≤18%[23]) 标准要求; 仅 17 个样本相对误差超过 30%, 异常样本量占比 (8.5%) 较低, 验证了模型优异的适应性和工程适用性。

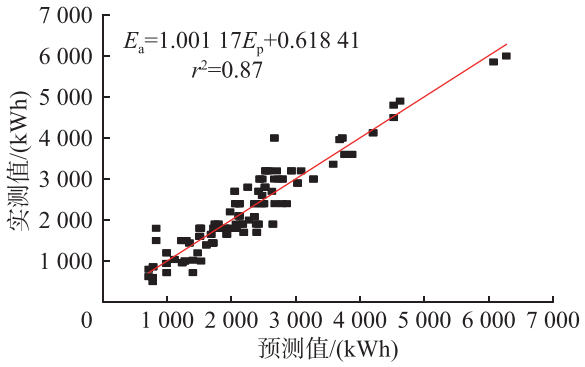

图6 模型预测能耗与实测数据的相关性验证结果

Fig. 6 Validation of correlation between model-predicted and measured energy consumption data

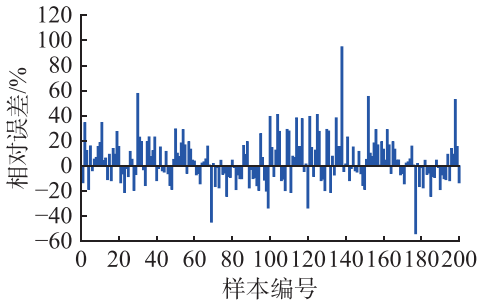

图7 模型预测能耗和与实测数据的相对误差分布

Fig. 7 Distribution of relative errors between model-predicted and measured energy consumption

3.2 模型可靠性验证

为评估模型预测结果的可靠性，采用 Bland–Altman 分析法对预测数据与实测值间的差异进行检验，结果如图 8。计算结果显示：平均差 (0.41 kWh) 接近 0，证实模型整体无显著系统偏差；一致性界限±1.96 标准差[24] 为−9.73~10.55 kWh，约 95% 的样本差值落在此区间，表明模型预测误差在可接受范围，验证了模型在常规能耗场景下的可靠性。

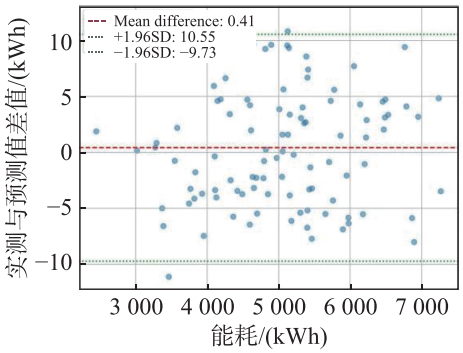

图8 模型预测能耗与实测数据的 Bland–Altman 一致性分析

Fig. 8 Bland–Altman agreement analysis between between model-predicted and measured energy consumption

3.3 模型泛化性验证

为验证模型的分布内泛化性，通过分层抽样确保训练集 (n=160) 与测试集 (n=40) 的变量的分布一致性，基于训练集构建的 SVM 混合模型在测试集上进行预测，结果如表 6。

表 6 分布内泛化验证指标数据对比表

Tab. 6 Comparison table of generalization validation indicator data within distribution

样本集	R^2	CVRMSE	MRE/%
训练集	0.921	0.158	13.5
测试集	0.914	0.163	14.6
差值/%	−0.7	+3.2	+1.1

由表 6 可看出：模型在测试集上表现优异，R^2 差异仅为 0.7%(测试集 0.914，训练集 0.921)，CVRMSE (测试集 0.163，训练集 0.158) 相对偏差为 3.2%；变量筛选 (输入参数从 29 个精简至 10 个) 后模型仍保持高精度 (R^2=0.914，CVRMSE=0.163)，充分验证模型对农村能耗数据稀疏性与变量缺失的强适应能力。

4 结论

通过采集夏热冬冷典型地区农村住宅的实测与理论数据，构建融合白箱与黑箱模型的多源数据集；采用四步变量筛选法优化输入参数，并基于 SVM 算法建立能耗预测混合模型，得到如下主要结论：

1) 在建筑参数信息缺失和实际能耗数据稀缺条件下，通过融合理论与易获实测数据构建混合预测框架，有效解决了农村地区数据稀缺问题。

2) 四步变量筛选法 (PLS−VIP 显著性、VIF 共线性、随机森林敏感性和后向回归分析) 从 29 个初始变量中遴选出 10 个关键变量，显著降低了模型复杂度，且预测性能优异 (R^2=0.914，CVRMSE=0.163)，实现了精度与效率的最佳平衡。

3) 模型验证结果表明预测平均相对误差 (14.57%) 满足农村节能改造要求 (≤18%)，Bland−Altman 分析显示 95% 样本预测误差分布在−9.73~10.55 kWh 一致性界限内且平均偏差仅 0.41 kWh，测试集与训练集的性能差异微小 (决定系数差异仅 0.7%，均方根误差变异系数差值 3.2%)，证实模型能有效适应农村住宅数据稀疏性和行为异质性特征，具备较强的泛化能力。

本文建立的模型在夏热冬冷 IIA 子区展现出优异的工程适应性、鲁棒性、可靠性与泛化能力，可为农村住宅节能提供可靠的技术支撑。但模型在实时预测和多地域适用性方面仍存在局限，未来将通过物联网动态监测和多气候区验证，进一步提升模型的泛化性与实时预测性能。

参考文献:

[1] DONG B, LI Z X, MAHBOBUR RAHMAN S M, et al. A hybrid model approach for forecasting future residential electricity consumption[J]. Energy and Buildings, 2016, 117:341−351.

[2] SU H L, DUAN M M, ZHUANG Z, et al. Building energy consumption prediction method based on Bayesian regression and thermal inertia correction[J]. International Journal of Renewable Energy Development, 2024, 13(1):71−79.

[3] QIAO Q Y, YUNUSA−KALTUNGO A, EDWARDS R. Hybrid method for building energy consumption prediction based on limited data[C]//2020 IEEE PES/IAS Power Africa. Nairobi, Kenya. IEEE, 2020:1−5.

[4] LIANG X B, CHEN S L, ZHU X, et al. Domain knowledge decomposition of building energy consumption and a hybrid data-driven model for 24 h ahead predictions[J]. Applied Energy, 2023, 344:121244.

[5] 袁鹏丽, 端木琳, 王宗山. 赤峰市某镇农村人员生活用电设备使用行为研究 [J]. 建筑科学, 2022, 38(10): 108−115, 28.
YUAN P L, DUANMU L, WANG Z S. Study on the usage behavior for domestic electrical appliances of rural households in Chifeng[J]. Building Science, 2022, 8(10):108−115, 228.

[6] ZOU R , YANG Q L, XING J C, et al. Research on public building energy consumption prediction method based on hybrid analysis of dynamic and static data[C]//2021 China Automation Congress. Beijing, China: IEEE, 2021:7961−7966.

[7] 李明财, 田喆, 曹经福, 等. 气候变化与建筑节能 [M]. 北京: 气象出版社, 2019.
LI M C, TIAN Z, CAO J F, et al. Climate Change and Building Energy Efficiency[M]. Beijing: China Meteorological Press, 2019.

[8] 龙惟定, 潘毅群, 王暫. 碳中和城市建筑能源系统 (3): 负荷篇 [J]. 暖通空调, 2022, 52(9):1−14.
LONG W D, PAN Y Q, WANG X. Building energy system of carbon neutrality cities(3): load[J]. Heating Ventilating & Air Conditioning, 2022, 52(9):1−14.

[9] 潘毅群, 李玉明, 张洁, 等. 实用建筑能耗模拟手册 [M]. 北京: 中国建筑工业出版社, 2013.
PAN Y Q, LI Y M, ZHANG J, et al. Practical Handbook of Building Energy Consumption Simulation [M]. Beijing: China Architecture & Building Press, 2013.

[10] 黄志甲, 徐萌, 吴州琴, 等. 徽州民居自然通风潜力评估 [J]. 建筑科学, 2023, 9(2):130−134,201.
HUANG Z J, XU M, WU Z Q, et al. Natural ventilation potential evaluation of Huizhou traditional dwellings[J]. Building Science, 2023, 9(2):130−134,201.

[11] 李俊鸽, 杨柳, 刘加平. 夏热冬冷地区人体热舒适气候适应模型研究 [J]. 暖通空调, 2008, 8(7):20−24.
LI J G, YANG L, LIU J P. Adaptive thermal comfort model for hot summer and cold winter zone[J]. Heating Ventilating & Air Conditioning, 2008, 8(7):20−24.

[12] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 3:1157−1182.

[13] THARWAT A, SCHENCK W. Active learning for handling missing data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(2):3273−3287.

[14] GORZAŁCZANY M B, RUDZIŃSKI F. Energy consumption prediction in residential buildings: an accurate and interpretable machine learning approach combining fuzzy systems with evolutionary optimization[J]. Energies, 2024, 17(13):3242.

[15] OLU−AJAYI R, ALAKA H, OWOLABI H, et al. Data-driven tools for building energy consumption prediction: a review[J]. Energies, 2023, 16(6):2574.

[16] MAHIEU B, QANNARI E M, JAILLAIS B. Extension and significance testing of variable importance in projection (VIP) indices in partial least squares regression and principal components analysis[J]. Chemometrics and Intelligent Laboratory Systems, 2023, 242:104986.

[17] CHONG I G, JUN C H. Performance of some variable selection methods when multicollinearity is present[J]. Chemometrics and Intelligent Laboratory Systems, 2005, 78(1/2):103−112.

[18] O'BRIEN R M. A caution regarding rules of thumb for variance inflation factors[J]. Quality & Quantity, 2007, 41(5):673−690.

[19] WANG Z Y, ZHANG W H, XU W J, et al. The prediction algorithm of energy consumption for cattle growth based on random forest[C]//2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence. Chongqing, China: IEEE, 2023:510−513.

[20] YUAN P, DUAN M, WANG Z. Analysis of factors influencing heating energy consumption in rural houses based on actual energy consumption data[J]. Building Science, 2020, 36:28−37.

[21] MAULIDINA F, RUSTAM Z, HARTINI S, et al. Feature optimization using backward elimination and support vector machines (SVM) algorithm for diabetes classification[J]. Journal of Physics: Conference Series, 2021, 1821(1): 012006.

[22] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1):267−288.

[23] 袁鹏丽. 北方农村区域住宅终端能耗预测方法及模型 [D]. 大连: 大连理工大学, 2020.
YUAN P L. Prediction Methods and Models for Residential End-use Energy Consumption in Rural Northern China[D]. Dalian: Dalian University of Technology, 2020.

[24] BLAND J M, ALTMAN D G. Statistical methods for assessing agreement between two methods of clinical measurement[J]. Lancet, 1986(8476):307−310.

责任编辑: 何莉