Statistics and Application 统计学与应用, 2026, 15(5), 88-97
Published Online May 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.155109

Hans汉斯

基于Energy Efficiency数据集的能耗建模研究

李　洋

南开大学统计与数据科学学院，天津

收稿日期：2026年4月12日；录用日期：2026年5月4日；发布日期：2026年5月14日

摘　要

随着建筑行业向绿色、低碳化转型，精准预测建筑能耗成为提升能源利用效率的关键。本文以UCI机器学习库中的建筑能效(Energy Efficiency)数据集为研究对象，深入探讨了不同抽样策略对统计推断精度及机器学习预测模型性能的影响。研究首先系统分析了数据集的8个特征变量与2个能耗目标变量(采暖负荷Y1与制冷负荷Y2)之间的关联性。随后，设计并实施了简单随机抽样、基于K-Means聚类的分层抽样以及整群抽样三种方案，并从均值相对误差、统计稳定性和设计效应等维度对抽样质量进行了量化评估；在此基础上，本文利用不同抽样背景下的样本构建了多输出回归模型。实验结果表明：分层抽样通过捕捉建筑高度和玻璃窗面积等核心变量的结构化分布，其样本代表性显著优于其他方法，能够有效降低抽样方差。在模型性能方面，基于分层抽样训练的预测模型在R^2和RMSE指标上均表现最优。研究结论证实，科学的抽样设计不仅能提高有限样本下的数据质量，更能显著增强后续回归分析的泛化能力与可靠性，为建筑能效领域的统计分析提供了方法论支持。

关键词

建筑能效，抽样策略，分层抽样，回归建模，统计评估

A Study on Energy Consumption Modeling Based on the Energy Efficiency Dataset

Yang Li

School of Statistics and Data Science, Nankai University, Tianjin

Received: April 12, 2026; accepted: May 4, 2026; published: May 14, 2026

Abstract

As the construction industry transitions toward green and low-carbon development, the accurate prediction of building energy consumption has become a key factor in improving energy utilization efficiency. This paper takes the Energy Efficiency dataset from the UCI Machine Learning Repository as

文章引用: 李洋. 基于 Energy Efficiency 数据集的能耗建模研究[J]. 统计学与应用, 2026, 15(5): 88-97.
DOI: 10.12677/sa.2026.155109

its research object, deeply exploring the impact of different sampling strategies on the accuracy of statistical inference and the performance of machine learning regression modeling. The dataset covers eight input variables reflecting building geometric characteristics and two energy consumption target variables: Heating Load (Y1) and Cooling Load (Y2). The study first identified the correlations between variables through exploratory data analysis (EDA), and subsequently designed and implemented three sampling schemes: simple random sampling, stratified sampling based on K-Means clustering, and cluster sampling. Through metrics such as relative mean error, statistical stability, and the design effect (Deff), the paper systematically and quantitatively evaluated the representativeness of samples from each scheme. On this basis, multi-output regression prediction models were constructed using samples obtained from the different sampling backgrounds. The experimental results indicate that stratified sampling, by capturing the structural distribution of core variables such as building height and glazing area, demonstrates significantly better sample representativeness than other methods. It effectively reduces sampling variance and enhances the information efficiency of the data. In terms of model performance, the model trained on stratified samples achieved the best results across both R^2 and RMSE indicators. The findings confirm that scientific sampling design not only improves data quality under limited sample conditions but also significantly strengthens the generalization capability and reliability of regression analysis, providing methodological support for statistical research in the field of building energy efficiency.

Keywords

Building Energy Efficiency, Sampling Strategy, Stratified Sampling, Regression Modeling, Statistical Evaluation

1. 引言

在全球气候变化与资源日益匮乏的背景下，提升建筑能效已成为各国实现可持续发展的核心战略[1]。据统计，建筑运行能耗在某些国家已占据全社会总能耗的 30%以上，因此，精准预测建筑的采暖负荷(Y1)与制冷负荷(Y2)对于优化暖通空调(HVAC)系统设计及降低建筑碳排放具有至关重要的意义[2]。建筑能效的表现受到多种几何设计变量的复杂驱动。研究表明，建筑的体型特征直接决定了其热损失效率，例如相对紧凑度(X1)作为衡量热交换效率的关键指标，其数值越高通常意味着更小的热量散失潜力[3]。表面积(X2)与总高度(X5)则从宏观上界定了建筑围护结构与外部大气环境接触的广度与风压暴露强度[4]。

在微观构造方面，墙体面积(X3)与屋顶面积(X4)的热工性能及热阻特性是维持室内热平衡的物质基础[5]。与此同时，玻璃窗面积(X7)及其在不同立面上的空间分布(X8)则是影响室内太阳辐射得热与光热平衡的最活跃因素[6]。此外，建筑的朝向(X6)决定了光热获取的周期性波动，其与几何变量之间的非线性耦合关系增加了能耗预测的难度[7]。虽然回归分析与机器学习模型已广泛应用于该领域，但高精度数据的采集往往面临实测成本高、周期长等实际障碍。

因此，如何在有限的样本容量下，通过科学的抽样策略(Sampling Strategy)提取最具代表性的数据点，成为统计学与数据科学交叉领域的研究重点[8]。传统的简单随机抽样(SRS)在处理具有显著结构化特征(如不同高度等级或玻璃比例)的建筑数据集时，往往容易产生样本偏移[9]。为了提升估计精度，学者们开始尝试引入 K-Means 聚类技术辅助的分层抽样(Stratified Sampling)，利用建筑几何特征的相似性进行层级划分，从而有效降低统计方差并提升数据的信息密度[10]。通过引入设计效应(Deff)指标，研究者可以

从量化角度评估特定抽样方案相对于随机抽样的增益效率[11]。

目前，关于抽样设计如何具体影响多输出回归模型预测稳健性的研究仍需进一步深化[12]。本文立足于 UCL 机器学习库的建筑能效数据集，系统探讨不同抽样方案对模型性能的底层驱动作用[13]。通过对比各方案在 R^2 与 RMSE 指标上的表现，本文旨在证明合理的抽样设计在增强回归分析泛化能力中的核心价值，从而为绿色建筑能效评估提供一套稳健的统计方法论支持[14]。

2. 数据来源

Table 1. Dataset metrics introduction
表 1. 数据集指标介绍

特征代码	变量名称	英文描述
X1	相对紧凑度	Relative Compactness
X2	表面积	Surface Area
X3	墙体面积	Wall Area
X4	屋顶面积	Roof Area
X5	总高度	Overall Height
X6	朝向	Orientation
X7	玻璃窗面积	Glazing Area
X8	玻璃窗分布	Glazing Area Distribution
Y1	采暖负荷	Heating Load
Y2	空调负荷	Cooling Load

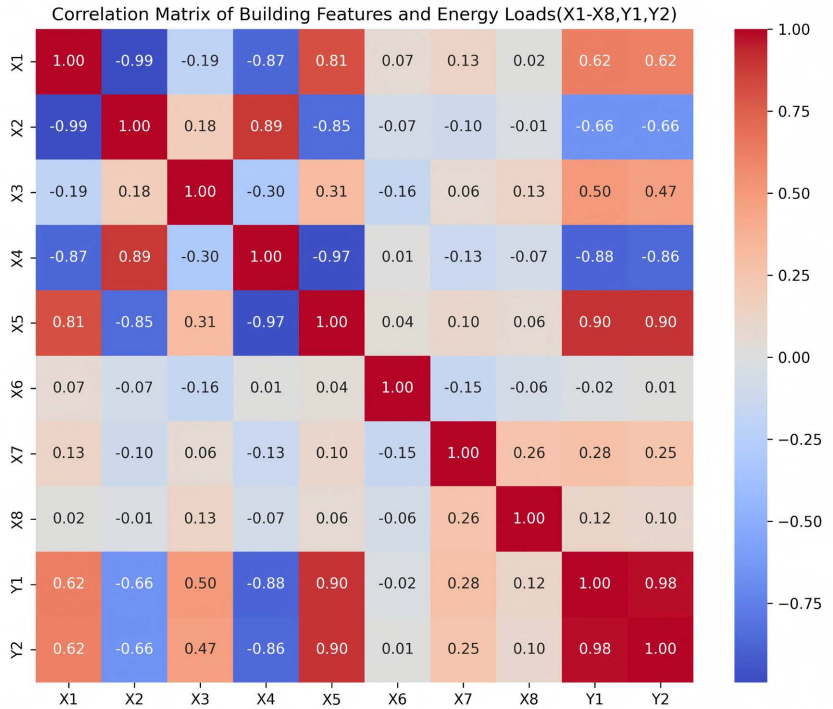

Figure 1. Correlation analysis
图 1. 相关性分析

本论文采用的是 Energy Efficiency 数据集，这份数据集源自 UCL 机器学习库，由 Athanasios Tsanas 和 Angeliki Xifara 共同开发，专门用于研究建筑几何参数与能源消耗之间的复杂关系。该数据集并非通过实地测量获得，而是利用专业的建筑仿真软件 Ecotect 对 12 种不同建筑形状的住宅建筑进行模拟生成的。模拟过程中，研究者通过系统地改变建筑的各种物理参数(如玻璃窗比例、分布、朝向等)，生成了 768 组具有代表性的实验样本。这种受控的生成方式使得数据非常干净，非常适合用于回归分析和算法基准测试。

数据集包含了 8 个用于描述建筑特征的输入变量(特征)以及 2 个反映能效表现的目标变量(标签) (见表 1)，相关性分析结果如图 1。

3. 研究方法

3.1. 简单随机抽样

利用 Python 对 Energy Efficiency 原始数据集(总体容量 N = 768)进行了无放回随机抽样，根据实验设计，设定抽样比例为 12.5%，最终抽取的样本量为 n = 96。为了确保实验结果的可重复性，在代码实现中通过设定了固定的随机种子，这种方式能够消除随机波动对不同模型性能比较带来的干扰。

为了验证所选样本是否能有效代表总体特征，本研究设计了基于均值对比的量化评估流程，对数据集中的 8 个自变量(X1~X8)及 2 个能耗目标变量(Y1 采暖负荷、Y2 制冷负荷)分别计算总体均值(μ)与样本均值($\bar{x}$)。通过计算相对误差(Relative Error)来衡量抽样偏差，计算公式如下：

$$\text{Relative Error} = \frac{\bar{x} - \mu}{\mu}$$

3.2. 分层抽样

在分层前，首先对数据集的 8 个几何特征变量(X1~X8)进行标准化处理。针对连续变量采用标准分数(Z-Score)转换，以消除不同物理量纲对后续聚类计算的干扰，确保各特征在构建分层依据时具有平等的权重。同时本方案打破了传统单一变量分层的局限，采用 K-Means 聚类算法自动识别总体的内在结构。通过迭代计算不同聚类数下的惯性值(Inertia)与轮廓系数(Silhouette Score)，确定最优的层数划分，从而将总体(N = 768)的建筑模型划分为若干个高内聚、低耦合的子群(即抽样层)。

在完成层划分后，本方案遵循按比例分配(Proportional Allocation)原则进行抽样。根据每个聚类簇在总体中所占的权重，设定 12.5%的统一抽样比例，从各层中独立进行无放回随机抽样，最终形成总规模为 n = 96 的分层样本。这一过程通过固定随机种子保证了实验的可重复性，且同样通过计算相对误差(Relative Error)来衡量抽样偏差。

3.3. 整群抽样

整群抽样方法将具有特定物理属性组合的建筑群体视为一个整体单元，模拟了在有限条件下对特定类别建筑进行全量调查的场景。本文以建筑能耗影响最为显著的两个离散变量——窗墙比(X7)与朝向(X6)作为划分群的基准。通过对这两个变量进行笛卡尔积组合，将总体的观测记录划分为若干个互斥的物理特征群。这种划分方式旨在通过特征耦合形成自然的“群”单元，以测试当样本集中于某些特定建筑物理形态时对全局推断的影响。在实施过程中，研究并非直接抽取个体记录，而是以“特征群”为抽样单元。利用随机数发生器从所有生成的特征群中随机抽取若干个完整的群，通过设定随机种子确保了筛选过程的随机性与实验的可重复性，并根据抽样比例要求，使选定群内的个体总数尽可能接近预设的样本容量。一旦某个特征群被选中，该群内的所有建筑模型观测记录(包含 X1~X8 及 Y1~Y2)均被全量纳入样本集。这种“群内全查、群间抽样”的模式，使得样本在群内保持了高度的局部相关性，从而能够有效评估在建筑特征分布不均或存在局部聚集现象时，抽样推断的偏差情况。

3.4. 回归建模

在完成样本抽取后，本研究构建了回归模型以量化预测建筑特征对采暖负荷(Y1)与制冷负荷(Y2)的影响，并以此评估不同抽样策略对模型泛化能力的作用。本文采用线性回归(Linear Regression)和随机森林(Random Forest)作为核心建模算法，以 X1 至 X8 作为输入特征，同时为了严谨评估建模效果，采用了双重指标评价体系。首先使用决定系数(R^2)来衡量模型对数据变异性的解释程度；其次引入均方根误差(RMSE)来直观反映预测值偏离实际能耗的具体量级。

针对两个目标变量，分别计算其在抽样样本上的表现，从而捕捉模型在不同预测维度上的精度差异；本文的一项关键流程是“全量验证”对比，除了评估模型在抽样测试集上的表现外，代码还实现了将基于少量样本训练的模型直接应用于全量数据集进行推理。通过对比“抽样 R^2 和 RMSE”与“全局 R^2 和 RMSE”的差异，量化分析不同抽样策略下模型的外推能力与统计稳健性，进而验证科学抽样对提升能效预测可靠性的价值。

4. 结果与分析

4.1. 简单随机抽样

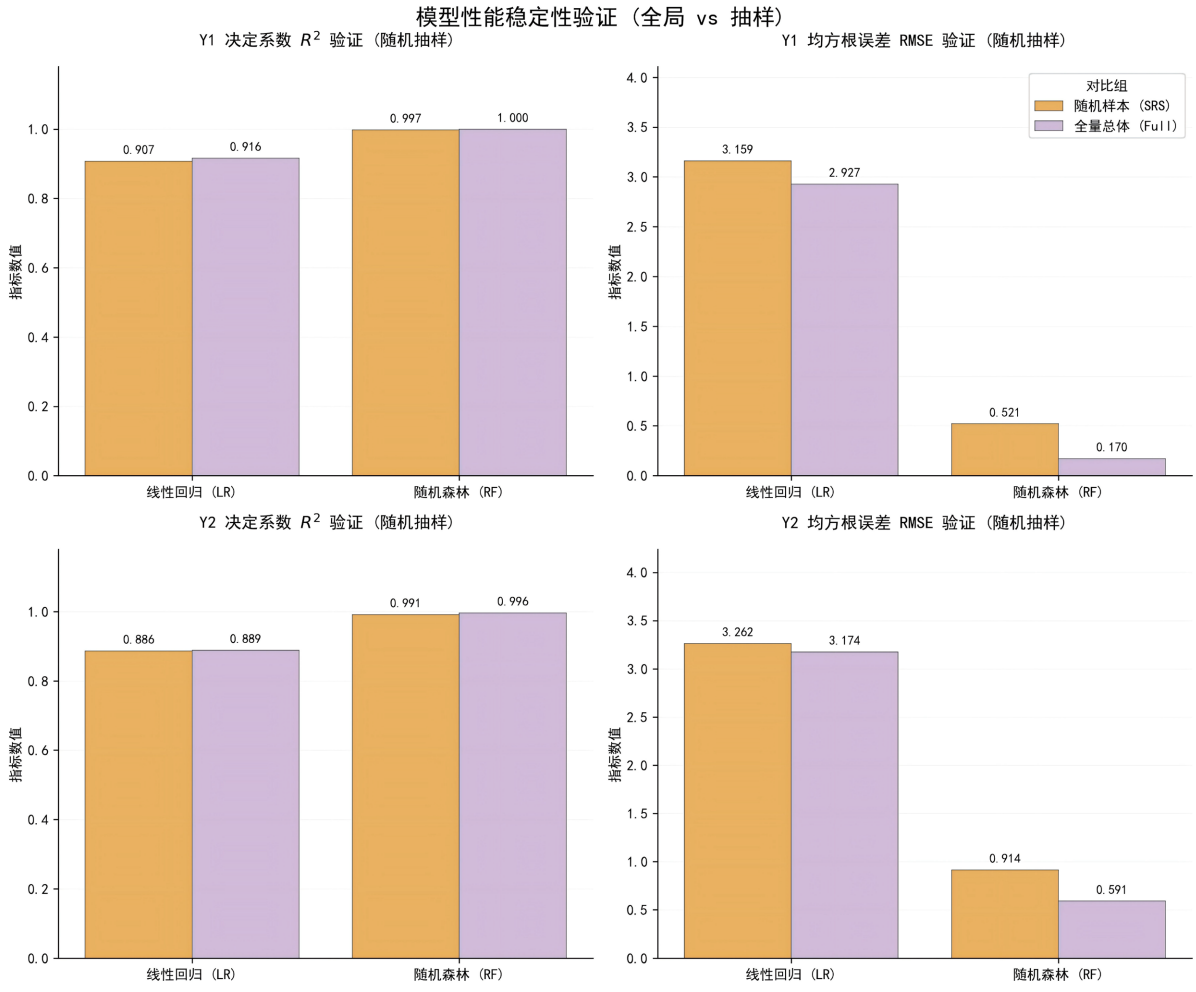

Figure 2. Model performance of simple random sampling
图 2. 简单随机抽样模型性能

Table 2. Relative error of simple random sampling
表 2. 简单随机抽样的相对误差

指标	Total Mean	Sample Mean	Relative Error
X1	0.764	0.769	0.006
X2	671.708	667.114	−0.007
X3	318.500	319.521	0.003
X4	176.604	173.797	−0.016
X5	5.250	5.359	0.021
X6	3.500	3.333	−0.048
X7	0.234	0.224	−0.042
X8	2.813	2.875	0.022
Y1	22.307	23.031	0.032
Y2	24.588	25.284	0.028

在简单随机抽样(SRS)的结果中，根据表 2 显示，大部分特征变量的相对误差控制在较小范围，如相对紧凑度(X1)为 0.006，表面积(X2)为−0.007，墙体面积(X3)为 0.003，而屋顶面积(X4)为−0.016；然而，受随机性影响，部分关键变量出现了较明显的偏移，如总高度(X5)的误差为 0.021，朝向(X6)与玻璃窗面积(X7)的误差分别扩大至−0.048 和−0.042，这直接导致目标变量采暖负荷(Y1)与制冷负荷(Y2)的相对误差分别达到 0.032 和 0.028。在模型性能方面(见图 2)，针对 Y1，线性回归(LR)在样本上的 R^2 为 0.907 (全局 0.916)，但 RMSE 从全局的 2.927 上升至 3.159；随机森林(RF)在样本上的 R^2 为 0.997，其 RMSE (0.521) 显著高于全局水平(0.170)。针对 Y2，LR 与 RF 的样本 R^2 分别为 0.886 和 0.991，其 RMSE 分别为 3.262 (全局 3.174)和 0.914 (全局 0.591)。虽然 SRS 在决定系数 R^2 上展现了较强的统计稳健性，能较好地还原变量间的总体关联，但在处理如建筑高度、朝向和玻璃窗面积等具有显著结构化特征的变量时，由于随机抽样缺乏对核心特征分布的针对性捕捉，容易产生样本偏移。这种偏差在 RMSE 指标上被放大，反映出模型在预测精度上的损失，证实了基础随机抽样在提取高效率信息及维持小样本预测稳健性方面存在局限。

4.2. 分层抽样

Table 3. Relative error of stratified sampling
表 3. 分层抽样的相对误差

指标	Total Mean	Sample Mean	Relative Error
X1	0.764	0.753	−0.014
X2	671.708	679.875	0.012
X3	318.500	323.604	0.016
X4	176.604	178.135	0.009
X5	5.250	5.250	0
X6	3.500	3.521	0.006
X7	0.234	0.255	0.088
X8	2.813	2.833	0.007
Y1	22.307	22.908	0.027
Y2	24.588	24.991	0.016

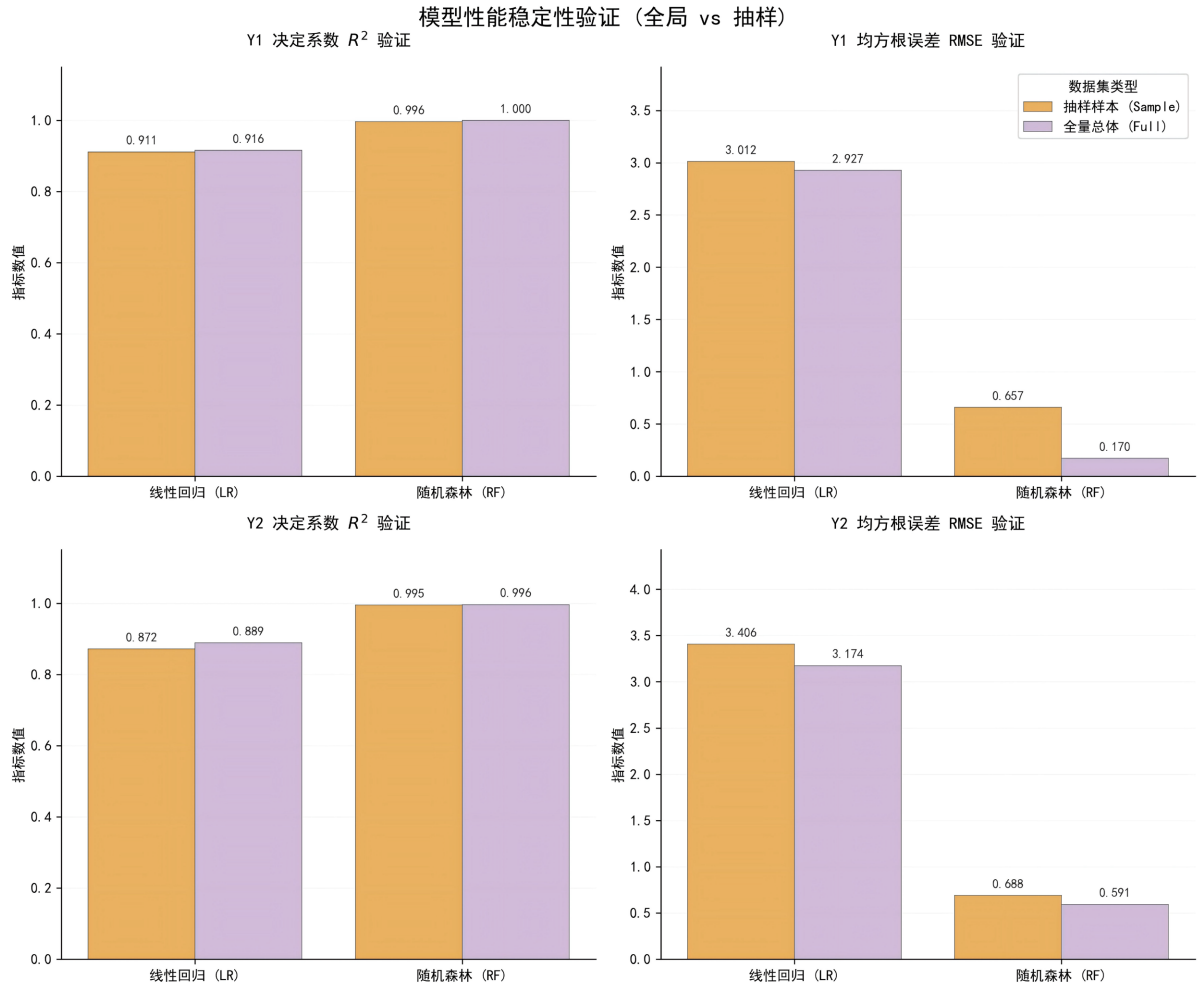

Figure 3. Model performance of stratified sampling
图 3. 分层抽样的模型性能

在分层抽样的量化结果中，根据表 3 显示，通过 K-Means 聚类对建筑几何特征进行层级划分，样本展现了极高的代表性。各特征变量相对误差极低，其中总高度(X5)的相对误差为 0，精准还原了总体的结构化分布；相对紧凑度(X1)为−0.014，表面积(X2)为 0.012，屋顶面积(X4)仅为 0.009，玻璃窗面积(X7)虽为 0.088，但目标变量采暖负荷(Y1)与制冷负荷(Y2)的相对误差分别优化至 0.027 和 0.016。在模型性能验证上(见图 3)，针对 Y1，线性回归(LR)在样本上的 R^2 达到 0.911，极度接近全局的 0.916，其 RMSE (3.012)较全局(2.927)仅有微小波动；随机森林(RF)的样本 R^2 为 0.996，RMSE 为 0.657。针对 Y2，LR 与 RF 的样本 R^2 分别为 0.872 和 0.995，RMSE 分别为 3.406 (全局 3.174)和 0.688 (全局 0.591)。从结果可以看出，分层抽样通过捕捉建筑高度和窗墙比等核心变量的结构化分布，有效降低了抽样方差并提升了信息效率。相比简单随机抽样，该方案在 R^2 的稳定性及 RMSE 的控制上表现更优，证明了科学的分层设计能保障样本在关键物理属性上的一致性，显著增强模型在受限样本下的泛化能力与稳健性。

4.3. 整群抽样

在整群抽样的量化分析中，根据表 4 显示，由于以玻璃窗面积(X7)与朝向(X6)的组合作为抽样单元，样本在关键变量上表现出显著偏移，其中 X7 的相对误差高达 0.387，导致目标变量采暖负荷(Y1)与制冷

负荷(Y2)的相对误差分别扩大至 0.100 和 0.066。实验观察到针对 Y1 的线性回归模型在样本上的 R^2 达到 0.925，甚至略高于全局的 0.916，这一异常升高现象被诊断为“同质化引起的伪高拟合风险”。这种现象源于整群抽样“群内全查”的特性，使得样本在局部具有高度相关性，虽然能较好地拟合特定特征组合下的能耗规律，但这种局部过度代表与全局信息缺失的并存，导致模型在处理非选定特征群的数据时稳健性较差。由图 4 可知，该风险在 RMSE 指标上得到了进一步验证，随机森林模型在 Y1 上的样本内误差(0.250)明显高于全局水平(0.170)，反映了模型在预测精度上的实际损失。因此，整群抽样中 R^2 的异常表现成为展示该方法统计局限性的典型案例，警示了在处理建筑能效这种强特征依赖数据时，因抽样偏差导致的“性能虚标”与泛化能力匮乏的风险。

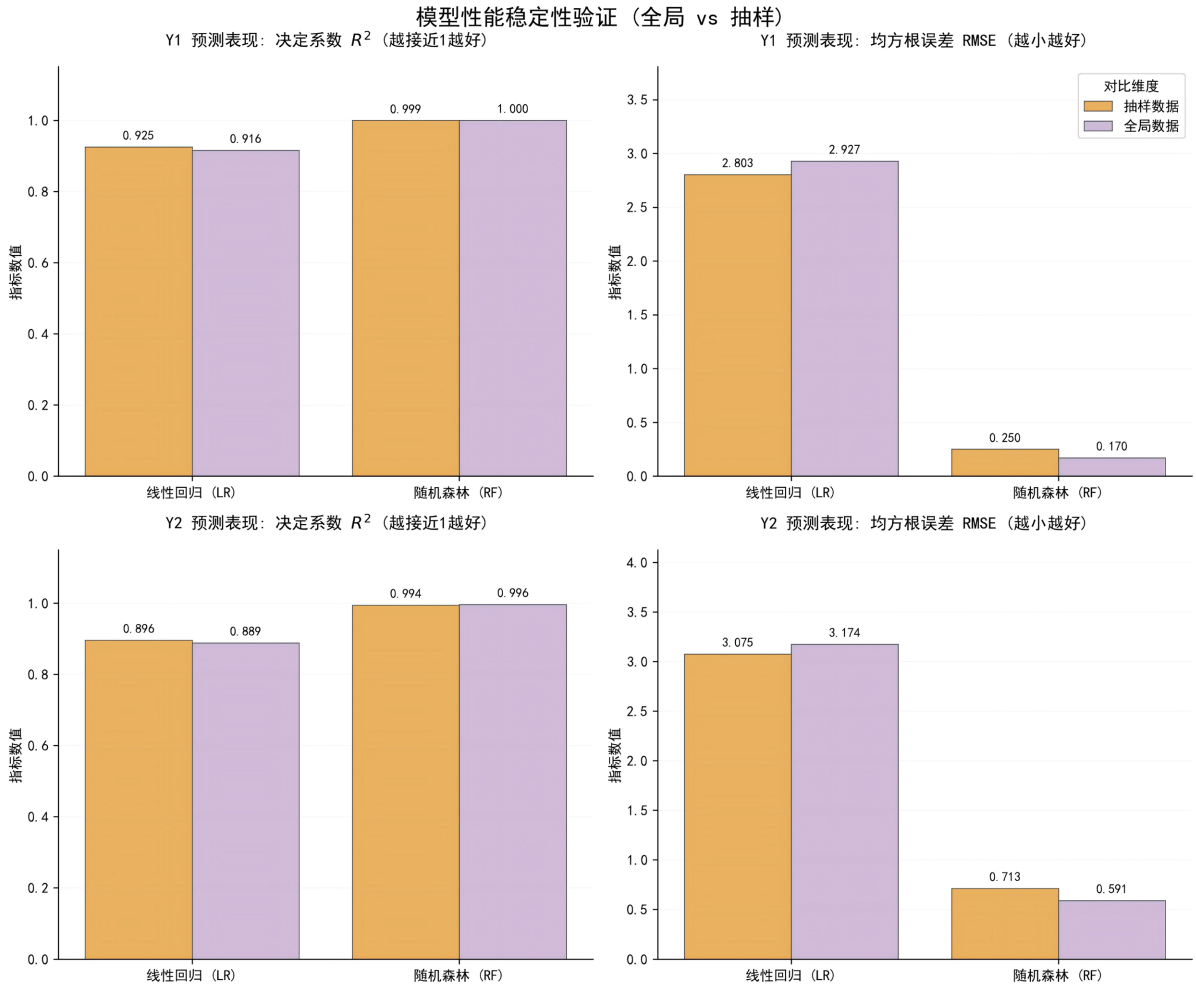

Figure 4. Model performance of cluster sampling
图 4. 整群抽样模型性能

Table 4. Relative error of cluster sampling
表 4. 整群抽样的相对误差

指标	Total Mean	Sample Mean	Relative Error
X1	0.764	0.769	0.007
X2	671.708	666.859	−0.007

续表

X3	318.500	317.990	−0.002
X4	176.604	174.435	−0.012
X5	5.250	5.356	0.021
X6	3.500	3.500	0
X7	0.234	0.325	0.387
X8	2.813	3.094	0.100
Y1	22.307	24.545	0.100
Y2	24.588	26.219	0.066

5. 结论与展望

本文围绕建筑能效数据集，系统对比了简单随机抽样、基于 K-Means 聚类的分层抽样及整群抽样对能耗预测模型的影响。研究发现，分层抽样在提升模型性能与统计代表性方面具有显著优势。通过对建筑高度、玻璃窗面积等关键物理特征进行分层设计，该方案有效捕获了数据集的结构化分布，使得基于其训练的线性回归与随机森林模型在 R^2 与 RMSE 指标上均优于其他抽样策略，验证了科学抽样在增强模型稳健性中的核心作用。简单随机抽样虽具备良好的统计中性，但在小样本量下容易因随机波动导致核心变量偏移，造成预测精度受损。而整群抽样由于群内个体的高度同质化，在面对特征依赖性强的建筑能效数据时，极易产生代表性偏差，限制了模型的泛化能力。

未来的研究可从以下三个维度进一步深化：首先，在抽样准则上，可探索结合深度学习的特征选择算法，通过自动识别高维特征间的非线性交互效应，构建更为动态和精准的分层框架；其次，在算法验证方面，建议引入支持向量机(SVM)、梯度提升树(XGBoost)等更多类型的回归模型，以全面评估抽样方案在复杂模型下的适用性；最后，应尝试将研究结论推广至不同气候区、不同建筑类型的多源能耗数据集中，以验证该抽样评价体系的普适性，为实际工程中的高效数据采集提供理论支撑。

综上所述，本研究通过对三种抽样策略的深入对比验证，系统地揭示了数据抽取方式对建筑能效预测模型可靠性的内在影响，这一研究成果不仅为有限资源下的高效数据采集提供了实证依据，也为未来构建更具适应性与精准度的绿色建筑智慧能源管理系统奠定了方法论基础。

参考文献

[1] Tsanas, A. and Xifara, A. (2012) Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools. *Energy and Buildings*, **49**, 560-567. https://doi.org/10.1016/j.enbuild.2012.03.003

[2] Catalina, T., Virgone, J. and Blanco, E. (2008) Development and Validation of Regression Models to Predict Monthly Heating Demand for Residential Buildings. *Energy and Buildings*, **40**, 1825-1832. https://doi.org/10.1016/j.enbuild.2008.04.001

[3] Ourghi, R., Al-Anzi, A. and Krarti, M. (2007) A Simplified Analysis Method to Predict the Impact of Shape on Annual Energy Use for Office Buildings. *Energy Conversion and Management*, **48**, 300-305. https://doi.org/10.1016/j.enconman.2006.04.011

[4] Pacheco, R., Ordóñez, J. and Martínez, G. (2012) Energy Efficient Design of Building: A Review. *Renewable and Sustainable Energy Reviews*, **16**, 3559-3573. https://doi.org/10.1016/j.rser.2012.03.045

[5] Al-Sanea, S.A. and Zedan, M.F. (2011) Improving Thermal Performance of Building Walls by Optimizing Insulation Layer Distribution and Thickness for Same Thermal Mass. *Applied Energy*, **88**, 3113-3124. https://doi.org/10.1016/j.apenergy.2011.02.036

[6] Susorova, I., Tabibzadeh, M., Rahman, A., Clack, H.L. and Elnimeiri, M. (2013) The Effect of Geometry Factors on

Fenestration Energy Performance and Energy Savings in Office Buildings. *Energy and Buildings*, **57**, 6-13. https://doi.org/10.1016/j.enbuild.2012.10.035

[7] Lindén, A., Carlsson-Kanyama, A. and Eriksson, B. (2006) Efficient and Inefficient Aspects of Residential Energy Behaviour: What Are the Policy Instruments for Change? *Energy Policy*, **34**, 1918-1927. https://doi.org/10.1016/j.enpol.2005.01.015

[8] Cochran, W.G. (1977) Sampling Techniques. 3rd Edition, John Wiley & Sons.

[9] Lohr, S.L. (2021) Sampling: Design and Analysis. CRC Press. https://doi.org/10.1201/9780429298899

[10] Magoulès, F. and Zhao, H.X. (2016) Data Mining and Machine Learning in Building Energy Analysis. Wiley. https://doi.org/10.1002/9781118577691

[11] Kish, L. (1965) Survey Sampling. John Wiley & Sons.

[12] Bourdeau, M., Zhai, X.Q., Nefzaoui, E., Guo, X. and Chatellier, P. (2019) Modeling and Forecasting Building Energy Consumption: A Review of Data-Driven Techniques. *Sustainable Cities and Society*, **48**, Article ID: 101533. https://doi.org/10.1016/j.scs.2019.101533

[13] Chou, J.S. and Bui, D.K. (2014) Modeling Heating and Cooling Loads by Artificial Intelligence for Energy-Efficient Building Design. *Energy and Buildings*, **82**, 437-446. https://doi.org/10.1016/j.enbuild.2014.07.036

[14] 金勇进. 抽样技术[M]. 北京: 中国人民大学出版社, 2015.