

文章编号: 2095-2163(2020)11-0050-07

中图分类号: TU831.1

文献标志码: A

# 基于逐步回归的 PVT-M 建筑不舒适度小时数预测模型分析

钟胜利<sup>1</sup>, 林尧林<sup>2</sup>, 黄兴华<sup>1</sup>

(1 上海工程技术大学 机械与汽车工程学院, 上海 201620; 2 上海理工大学 环境与建筑学院, 上海 200093)

**摘要:** 本文以上海市一栋单层居住建筑为研究对象, 建立了集成 PVT-M 建筑不舒适度小时数的逐步回归预测模型。首先, 基于公开文献的调查, 总共选择了 34 个设计参数; 其次, 采用蒙特卡洛随机抽样方法并结合 DesignBuilder 计算软件, 得到了容量为 1000 的样本数据库; 最后, 利用数据挖掘软件 SPSS 建立了不舒适度小时数的逐步回归预测模型, 并对回归方程的有效性、适用性和准确性以及设计变量的重要性进行了分析。结果表明, 逐步线性回归方法是一种可靠的预测模型, 能对不舒适度小时数进行有效且准确的预测。就设计参数的影响程度而言, 夏季室内空调设定温度影响程度最大, 相变材料(Phase Change Material, PCM)类型和光伏倾角影响程度最小。

**关键词:** 不舒适度小时数; 逐步线性回归; 预测模型; 数据挖掘

## Analysis of PVT-M Building Discomfort Degree Hour Prediction Model Based on Stepwise Regression

ZHONG Shengli<sup>1</sup>, LIN Yaolin<sup>2</sup>, HUANG Xinghua<sup>1</sup>

(1 School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;

2 School of Environment and Architecture, University of Shanghai for Science and Technology, Shanghai 200093, China)

**[Abstract]** This paper takes a single-storey residential building in Shanghai as the research object, and establishes a stepwise regression prediction model that integrates PVT-M building discomfort hours. First, a total of 34 design parameters were selected based on the public literature survey; secondly, using the Monte Carlo random sampling method combined with the DesignBuilder calculation software, a sample database with a capacity of 1000 was obtained; finally, a data mining software SPSS was used to establish a sample database. A stepwise regression prediction model of comfort hours, and an analysis of the validity, applicability and accuracy of the regression equation and the importance of design variables. The results show that the stepwise linear regression method is a reliable predictive model, which can effectively and accurately predict the hours of discomfort. In terms of the degree of influence of the design parameters, the indoor air-conditioning setting temperature in summer has the greatest influence, and the phase change material (PCM) type and photovoltaic inclination have the least influence.

**[Key words]** discomfort degree hour; stepwise linear regression; prediction model; data mining

### 0 引言

随着经济的快速发展和人们生活水平的提高, 人们对改善居住环境条件的需求不断增加。然而, 在建筑能耗日益增长的趋势下, 大多数建筑设计者及研究人员往往关注的是如何降低建筑能耗, 却忽略了室内的环境质量, 这一现象容易产生病态建筑综合征(SBS)<sup>[1]</sup>。不舒适小时数作为衡量建筑性能的重要指标之一, 受建筑围护结构传热系数、窗墙比、建筑表面太阳热吸收率等诸多建筑热物理参数的影响<sup>[2]</sup>。

目前, 通过数据挖掘软件与模型预测技术, 可以获取影响不舒适小时数参数的基础数据, 并建立相关预测模型, 使设计者在建筑设计早期快速准确的获得室内热舒适情况, 从而为居住者提供一个健康

舒适的室内环境。

关于利用逐步线性回归方法建模和预测方面, 蒲清平等<sup>[3]</sup>通过 SPSS 软件建立了居住建筑能耗预测的逐步线性回归模型, 并对模型的拟合效果进行了检验。结果表明, 模型预测年能耗与实际统计年能耗符合度达 95% 左右, 说明模型具有较高的预测精度和较好的拟合效果; Amiri 等<sup>[4]</sup>采用逐步线性回归方法, 建立了建筑能耗预测模型, 并将预测与模拟结果进行对比分析。结果表明, 二者之间的误差是可以接受的, 同时指出该方法简单, 能够准确快速地对建筑能耗进行预测; Braun 等<sup>[5]</sup>利用逐步回归方法, 分别建立了燃气消耗和电力消耗预测模型, 并将预测值与实际值进行了比较。结果表明, 两个模型的预测值都是令人满意的; Hygh 等<sup>[6]</sup>采用逐步线

**作者简介:** 钟胜利(1995-), 男, 硕士研究生, 主要研究方向: 建筑节能; 林尧林(1976-), 男, 博士, 副教授, 主要研究方向: 可持续建筑与建筑节能。

**通讯作者:** 林尧林 Email: yaolinlin@gmail.com

**收稿日期:** 2020-08-13

性回归的方法,建立了 4 个城市不同气候区供热、制冷以及总能耗的预测模型,并与 EnergyPlus 模拟结果比较。结果显示,预测数据与模拟结果吻合较好,同时也表明在设计初期,线性回归可以作为一种有效的简化模型来代替能耗模拟模型。

在热舒适性建模与预测方面,孙斌等<sup>[7]</sup>利用 BP 网络、GA-BP 网络、RBF 网络及 Elman 网络,分别建立了热舒适性指标预测模型。结果表明,GA-BP 神经网络的预测性能最佳,并指出其对权值和阈值的优化是以训练时间为代价的;喻伟等<sup>[2]</sup>考虑到 14 个变量对建筑能耗和室内热舒适状况的影响,并建立了 GA-BP 网络模型。通过对样本数据进行训练和测试,验证了该模型具有较高的预测精度,同时表明人工神经网络预测精度受样本数据的影响;陆焯等<sup>[8]</sup>采用 PSO-RBF 的方法,建立了 PMV 指标预测模型,实现了对 PMV 指标的智能预测,并通过仿真计算表明,PSO-RBF 网络的预测误差精度提高了 79.5%,小于 RBF 网络;朱婵等<sup>[9]</sup>提出了一种基于改进的禁忌遗传算法神经网络的热舒适度预测模型(TGA-BPNN),通过仿真实验并与 BP 神经网络及遗传神经网络相比,TGA-BPNN 可以进一步提升模型预测的准确性,同时表明采用此方法存在算法运行时间长、空间复杂度大以及效率低等不足。

综上所述可以发现,利用逐步回归方法进行预测主要是针对建筑能耗,而对热舒适性等建筑环境领域的研究很少。回归模型不仅结构简单,而且可以达到准确可靠的预测效果。而对于热舒适性的预测普遍采用神经网络。然而,利用传统的神经网络进行预测时,其预测结果误差往往取决于样本数据。大部分文献都采用算法与神经网络结合的方式,来提高预测精度。但其结构复杂程度会随之增加,算法运行效率也会有所下降,即耗时又耗力。因此,本文采用逐步回归方法,建立了集成 PVT-M 建筑的不舒适度小时数预测模型,并对模型的准确度以及预测变量的重要性进行了分析。

## 1 建筑模型及参数变量

### 1.1 居住建筑模型

本文选取的居住建筑位于上海市,典型气候特征为夏季闷热,冬季湿冷。该气候区的建筑物必须满足夏季防热、通风降温要求,冬季应兼顾防寒取暖需求。

利用 DesignBuilder 建立了建筑模型,如图 1 所示。建筑面积 100 m<sup>2</sup>,高度 4m,为了建筑在冬季能获得更多的太阳辐射获得热量,建筑朝向采用了该

地区最佳的南偏东 15°方位。建筑采用光伏板、相变材料和特朗伯集热墙(Trombe wall)结构。光伏板布置在屋顶,主要提供室内用电设备的能源消耗;建筑南向为带有相变材料的特朗伯集热墙,分为内层、中间层和外层 3 层。内层墙体结构为面砖层、XPS 保温层、混凝土层、相变材料层、石膏抹灰层,墙体上开了两个通风孔,其主要作用是结合中间层的空气腔来实现建筑的自然通风,从而降低室内的冷热负荷,外层为玻璃幕墙。其它 3 面墙体都设置了外窗,且采取了遮阳措施。

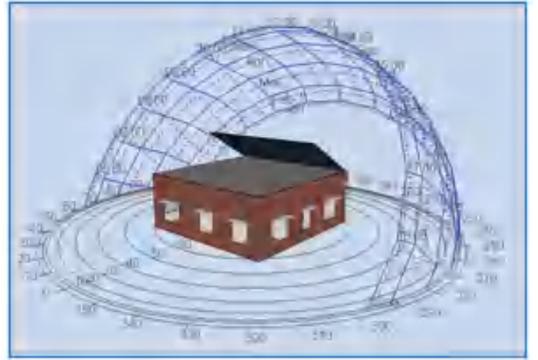


图 1 建筑模型图

Fig. 1 Building model diagram

### 1.2 参数变量

#### 1.2.1 自变量

本文选取的 34 个参数变量都是查阅相关文献以及规范标准获取的,参数涉及窗墙比、保温层厚度、混凝土厚度、太阳热吸收率、外窗类型、遮阳类型、夏季室内空调和冬季室内采暖温度设定值、相变材料的类型、厚度以及相变温度、光伏板的倾角和面积、Trombe 墙的空腔厚度、幕墙厚度和通风口面积等。具体变量类型及取值范围见表 1。

表 1 中:G1-G16 表示 16 种不同的外窗类型;L1-L9 表示 9 种不同长度厚 10 cm 的悬挑混凝土板;P1-P5 表示 5 种不同的相变材料;W6-W10 表示 5 种不同厚度的玻璃幕墙,这些变量均属于离散型变量。

#### 1.2.2 因变量

本文选取的目标函数为不舒适度小时数,可分为夏季不舒适度小时数和冬季不舒适度小时数,其数学表达式为<sup>[10]</sup>:

$$DH = \sum_{i=1}^{8760} (T_1 - 26) + \sum_{i=1}^{8760} (18 - T_2) \quad (1)$$

式中,  $T_1$  为全年高于 26 °C 的室内空气温度,  $T_2$  为全

年低于 18℃ 的室内空气温度。

表 1 变量类型及数值范围

Tab. 1 Variable type and numeric range

变量类型		数值范围
窗墙比/%	东 ( $X_1$ )	[10,60]
	西 ( $X_2$ )	[10,60]
	北 ( $X_3$ )	[10,60]
保温层厚度/mm	东 ( $X_4$ )	[10,300]
	南 ( $X_5$ )	[10,300]
	西 ( $x_6$ )	[10,300]
	北 ( $X_7$ )	[10,300]
	屋顶 ( $X_8$ )	[10,300]
混凝土厚度/mm	东 ( $X_9$ )	[50,300]
	南 ( $X_{10}$ )	[50,300]
	西 ( $X_{11}$ )	[50,300]
	北 ( $X_{12}$ )	[50,300]
	屋顶 ( $X_{13}$ )	[50,300]
太阳热吸收率(-)	东 ( $X_{14}$ )	[0.1,0.9]
	西 ( $X_{15}$ )	[0.1,0.9]
	北 ( $X_{16}$ )	[0.1,0.9]
	屋顶 ( $X_{17}$ )	[0.1,0.9]
外窗类型	东 ( $X_{18}$ )	G1-G16
	西 ( $X_{19}$ )	G1-G16
	北 ( $X_{20}$ )	G1-G16
遮阳类型	东 ( $X_{21}$ )	L1-L9
	西 ( $X_{22}$ )	L1-L9
	北 ( $X_{23}$ )	L1-L9
相变温度/℃	熔化温度 ( $X_{24}$ )	[18,28]
	凝固温度 ( $X_{25}$ )	[15,25]
PCM 类型(-)	$X_{26}$	P1-P5
PCM 厚度/mm	$X_{27}$	[5,100]
玻璃幕墙厚度/mm	$X_{28}$	W6-W10
冬季室内采暖温度/℃	$X_{29}$	[18,22]
夏季室内空调温度/℃	$X_{30}$	[24,26]
光伏倾角/(°)	$X_{31}$	[0,90]
光伏面积/m <sup>2</sup>	$X_{32}$	[10,100]
空腔厚度/mm	$X_{33}$	[100,500]
通风口面积/m <sup>2</sup>	$X_{34}$	[0.01,0.25]

## 2 建立预测模型

### 2.1 数据采集

充足的样本量是保证预测模型稳定性和准确性的关键。为了建立不舒适度小时数预测模型,需要

建立一个以建筑设计参数为输入,以不舒适度小时数为输出的数据库。本文采用了 20 世纪 40 年代由 S.Ulam 提出的蒙特卡洛抽样方法(MCM)<sup>[11]</sup>,MCM 是一种随机模拟抽样方法,其工作原理如下:

- (1) 构造或描述概率过程;
- (2) 实现从已知概率分布抽样;
- (3) 建立各种估测量。

利用该方法对选取的 34 个变量进行抽样,最终确定了 1 000 个样本。通过仿真软件 DesignBuilder 对样本进行模拟,来获取不舒适度小时数。

### 2.2 逐步线性回归模型(SLR)

线性回归分析方法已被普遍应用于不同建筑的性能预测。S.Asadi 等<sup>[12]</sup>发现多元线性回归模型在建筑设计阶段的早期应用,可以提高能源效率和减少排放。逐步线性回归模型(SLR)属于线性回归的一种,由于变量个数和回归模型的复杂性会对模型拟合优度产生显著影响,逐步线性回归可以采用正向选择和逆向淘汰相结合的方法实现自动选择自变量,从而确定自变量对因变量的影响程度大小。其模型描述如下:

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_p x_{ip} + \varepsilon. \quad (2)$$

式中, $\beta_0$  为回归常数, $\beta_1, \beta_2, \beta_3, \dots, \beta_p$  为回归系数,通过最小二乘法确定回归系数,使平方和误差最小。

## 3 逐步线性回归结果分析

### 3.1 逐步线性回归方程

利用 IBM SPSS Modeler 数据挖掘软件建立了不舒适度小时数逐步回归模型,模型结构如图 2 所示。采用步进(条件:当候选变量中最大  $F$  值的概率  $\leq 0.05$  时,引入相应变量;在引入方程的变量中,最小  $F$  值的概率  $\geq 0.1$  时,则剔除该变量)的方法,选择进入或除去的自变量。在 34 个建筑设计参数中,逐步回归方法建立的不舒适度小时数回归模型保留了 22 个参数。

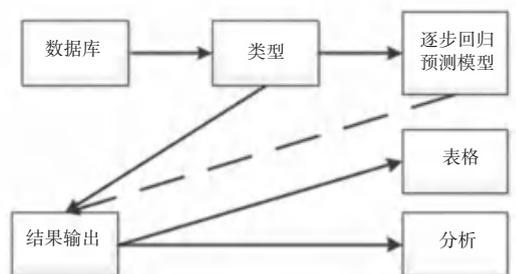


图 2 逐步回归模型结构图

Fig. 2 Stepwise regression model structure diagram  
在回归模型中,一般  $P \leq 0.05$  则认为具有显著

性, 根据未标准化系数  $B$  值可以得到式 (3) 所示的不舒适度小时数回归方程。不舒适度小时数回归模型变量的回归系数以及显著性  $P$  值见表 2。从表中可以看出, 不舒适度小时数回归模型的变量回归系数所对应的  $P < 0.05$ , 说明模型的自变量和因变量之间有明显的线性关系, 建立的回归方程是有效的。

$$y = -3267.725 - 784.885 X_1 - 254.772 X_2 - 305.261 X_3 - 0.203 X_4 + 0.251 X_5 - 0.174 X_6 - 0.368 X_7 - 0.694 X_8 - 0.203 X_{12} - 60.210 X_{14} - 142.576 X_{17} - 3.000 X_{18} - 2.498 X_{19} - 7.507 X_{20} + 216.340 X_{21} +$$

$$7.369 X_{22} + 5.729 X_{23} + 3.854 X_{24} + 6.667 X_{26} - 123.494 X_{29} + 446.592 X_{30} - 0.368 X_{31}. \quad (3)$$

在进行线性回归分析时, 共线性会使参数估计不稳定。方差膨胀因子 (VIF) 可以检测多重共线性, 它和容差 (Tolerance) 互为倒数关系, 当  $VIF \geq 10$  时, 说明变量之间有严重的多重共线性, 其值越接近 1, 变量之间多重共线性越弱。从表 2 统计的数值来看, 不舒适度小时数回归模型相关变量的 VIF 均在 1 附近, 说明这些变量之间共线性较弱。

表 2 逐步回归系数  
Tab. 2 Stepwise regression coefficient

模型	变量	未标准化系数		标准化系数 Beta	显著性	共线性统计	
		B	标准误差			容差	VIF
DH	(常量)	-3267.725	226.002		0.000		
	夏季室内空调温度	446.592	8.189	0.686	0.000	0.983	1.017
	冬季室内采暖温度	-123.494	4.045	-0.384	0.000	0.980	1.021
	东窗墙比	-784.885	31.913	-0.309	0.000	0.982	1.019
	屋顶保温层厚度	-0.694	0.057	-0.153	0.000	0.977	1.023
	北窗墙比	-305.261	31.990	-0.120	0.000	0.980	1.021
	东外遮阳	16.340	1.830	0.113	0.000	0.971	1.030
	西窗墙比	-254.772	31.393	-0.102	0.000	0.980	1.020
	屋顶吸收率	-142.576	20.106	-0.090	0.000	0.975	1.026
	北外窗类型	-7.507	1.028	-0.092	0.000	0.977	1.023
	北外墙保温层厚度	-0.368	0.056	-0.083	0.000	0.970	1.031
	西外遮阳	7.369	1.766	0.053	0.000	0.976	1.024
	南外墙保温层厚度	0.251	0.056	0.056	0.000	0.981	1.020
	东外墙保温层厚度	-0.203	0.056	-0.046	0.000	0.987	1.013
	北外遮阳	5.729	1.803	0.040	0.002	0.978	1.023
	西外墙保温层厚度	-0.174	0.056	-0.039	0.002	0.982	1.018
	东外窗类型	-3.000	1.009	-0.037	0.003	0.983	1.017
	东外墙吸收率	-60.210	20.118	-0.038	0.003	0.984	1.017
	北外墙混凝土厚度	-0.203	0.065	-0.040	0.002	0.968	1.033
	西外窗类型	-2.498	0.996	-0.031	0.012	0.989	1.011
	熔化温度	3.854	1.598	0.030	0.016	0.983	1.017
	光伏倾角	-0.368	0.179	-0.026	0.040	0.964	1.037
	PCM 类型	6.667	3.297	0.026	0.043	0.971	1.029

### 3.2 回归模型拟合优度检验

回归方程的拟合优度检验, 是检验样本数据聚集在样本回归直线周围的密集程度, 从而判断回归方程对样本数据的代表程度。拟合优度检验一般采用调整决定系数  $R^2$  实现, 该统计量的值越接近于 1,

拟合优度越好,  $R^2$  可由式 (4) - 式 (8) 计算得到。

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}, \quad (4)$$

$$y = \frac{1}{n} \sum_{i=1}^n y_i, \quad (5)$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2, \quad (6)$$

$$SS_{reg} = \sum_i (f_i - \bar{y})^2, \quad (7)$$

$$SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2. \quad (8)$$

式中,  $SS_{reg}$  为回归平方和;  $SS_{res}$  为残差平方和;  $SS_{tot}$  为总平方和;  $y_i$  为真实值;  $f_i$  为预测值;  $\bar{y}$  为平均值。

不舒适度小时数模拟值与 SLR 预测值的回归如图 3 所示。可以看出, 模拟和预测的数据结果有很好的一致性, 不舒适度小时数回归模型的  $R^2$  为 0.845, 显示出较好的拟合效果。

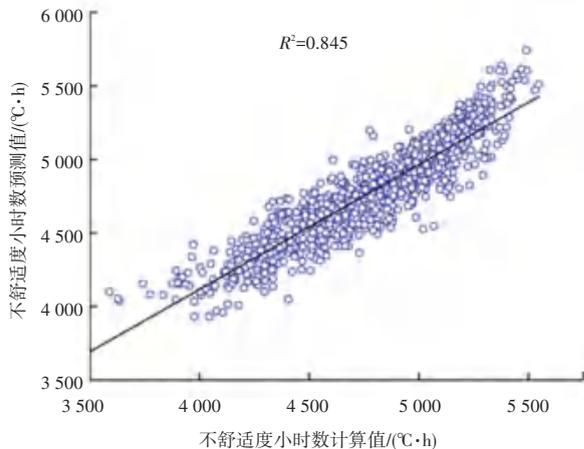


图 3 不舒适度小时数模拟值与 SLR 预测值回归图

Fig. 3 Regression diagram of simulated value of discomfort degree hour and SLR predicted value

表 3 不舒适度小时数残差统计表

Tab. 3 Residual statistical table of discomfort degree hour

模型	回归统计值	最小值	最大值	平均值	标准偏差
DH	预测值	3791.5938	5741.6099	4764.2056	342.14055
	残差	-523.69379	522.80048	.00000	144.74873
	标准预测值	-2.843	2.857	.000	1.000
	标准残差	-3.578	3.572	.000	.989

### 3.3 回归模型误差分析

为了评价不舒适度小时数逐步回归模型的准确度, 采用相对误差 (RE) 这一指标来进行衡量, 其数学表达式如下:

$$RE = \frac{R_p - R_s}{R_s}. \quad (9)$$

式中,  $R_p$  为利用 SPSS 软件线性回归的预测值,  $R_s$  是利用 DesignBuilder 仿真软件的计算值。

表 4 给出了 1 000 组预测样本数据的相对误差范围。由此可见, 利用 SLR 方法预测的不舒适度小时数最大值和最小值分别为 5741.61 和 3791.59, 相对误差的最大值和最小值分别为 16.03 和 -10.32,

一个好的线性回归模型必须满足相关的所有假设, 其中包括线性、独立性、正态性、方差齐性等。图 4 给出了不舒适度小时数回归模型的标准化残差正态概率 P-P 图。由图可见, 所有的点都比较靠近对角线, 且结合残差统计表 3 得到的不舒适度小时数回归的标准偏差为 0.989 ( $< 2$ ), 说明回归模型的残差是呈正态分布的。

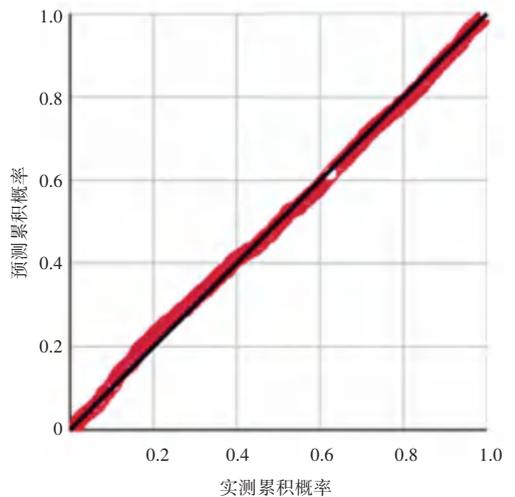


图 4 不舒适度小时数标准化残差正态概率 P-P 图

Fig. 4 P-P graph of standardized residual normal probability of discomfort degree hour

再结合图 5 统计的不舒适度小时数相对误差分层梯度范围可以得出: 相对误差范围在 10% - 20% 的样本只有 8 组, 占样本总数的 0.8%, 绝大多数样本相对误差范围小于 10%, 其中相对误差小于 2.5% 更是达到了一半以上, 说明利用 SLR 能达到对不舒适度小时数较好的预测效果。

表 4 不舒适度小时数相对误差范围

Tab. 4 Relative error range of discomfort degree hour

模型	计算值	预测值	相对误差/%	
DH	最大值	5547.30	5741.61	16.03
	最小值	3267.90	3791.59	-10.32

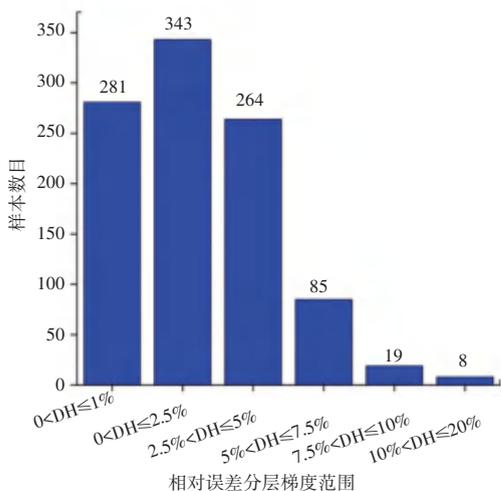


图 5 不舒适度小时数相对误差分层梯度范围

Fig. 5 Relative error stratified gradient range of discomfort degree hour

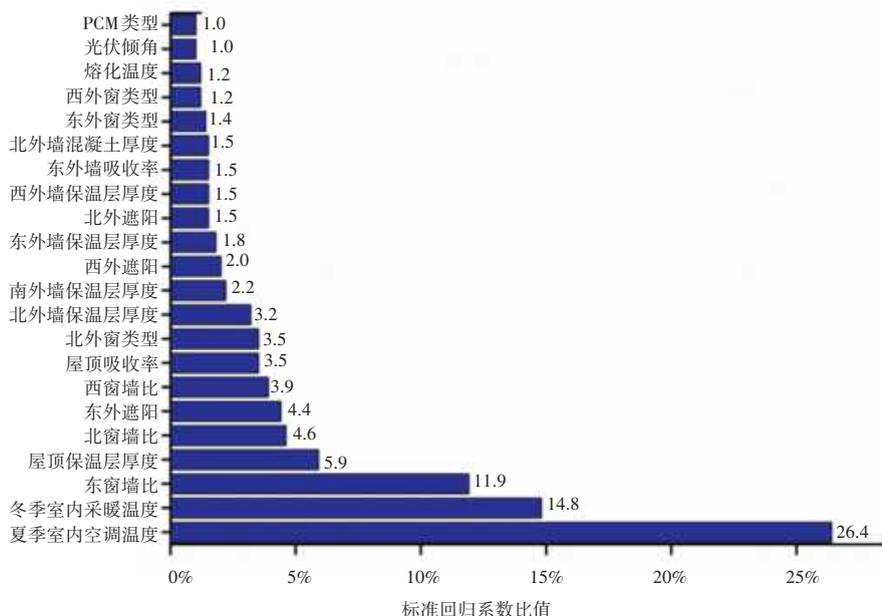


图 6 预测变量影响程度大小

Fig. 6 The degree of influence of predictive variables

## 4 结束语

本文利用逐步回归方法,建立了集成 PVT-M 建筑不舒适度小时数模型。通过对模型分析,可以得到以下结论:

(1)在选取的 34 个设计变量中,利用逐步回归方法建立的不舒适度小时数回归方程中保留了 22 个参数,同时通过显著性 P 值验证了方程的有效性。

(2)在模型拟合优度方面,不舒适度小时数 SLR 模型的回归系数为 0.845,说明计算数据与预测数据之间具有较高的线性拟合度。

(3)在回归模型误差方面,相对误差范围小于 2.5%的样本数占半数以上,只有极少一部分样本相对误差范围较大,说明 SLR 是一种可行的模型预测

## 3.4 预测变量重要性分析

依据表 2 中列出的 22 个参数变量,为了分析每个自变量对不舒适度小时数的影响程度大小,采用单个自变量标准化系数值的绝对值与方程相关的所有自变量的绝对值之和的比值作为评价标准。

通过表 2 的标准化系数,可以计算得到每个自变量所占比例大小,其统计结果如图 6 所示。从图中可以看出,夏季室内空调温度对不舒适度小时数的影响程度最大,其次为冬季室内采暖温度和东向窗墙比。前三者标准回归系数所占比值分别为 26.4%、14.8% 和 11.9%,PCM 类型和光伏倾角对不舒适度小时数的影响程度最小,所占比例只有 1%。

方法,能实现对不舒适度小时数的准确预测。

(4)在预测变量重要性方面,对不舒适度小时数影响程度最大的为夏季室内空调温度,影响程度最小的为 PCM 类型和光伏倾角。

## 参考文献

- [1] 叶剑军,李景广,葛曹燕,等. 室内环境预测不满意度评价方法研究[J]. 暖通空调,2010,40(10):57-61.
- [2] 喻伟,李百战,杨明宇,等. 基于人工神经网络的建筑多目标预测模型[J]. 中南大学学报(自然科学版),2012,43(12):4949-4955.
- [3] 蒲清平,李百战,喻伟. 重庆城市居住建筑能耗预测模型[J]. 中南大学学报(自然科学版),2012,43(4):1551-1556.
- [4] AMIRI S S, MOTTAHEDI M, ASADI S. Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the U. S. [J]. Energy and Buildings, 2015, 109: 209-216.