同济大学
TONGJI UNIVERSITY

博士学位论文

建筑信息异构数据融合方法及混合能耗模型的建立

姓　　名：沙华晶

学　　号：1710230

所在院系：机械与能源工程学院

学科门类：工学

学科专业：供热、供燃气、通风及空调工程

指导教师：许鹏

二〇二一年五月

同济大学
TONGJI UNIVERSITY

A dissertation submitted to
Tongji University in conformity with the requirements for
the degree of Doctor of Philosophy

Building heterogeneous data integration method and the establishment of the hybrid energy prediction model

Candidate: Huajing Sha

Student Number: 1710230

School/Department: School of Mechanical Engineering

Discipline: Engineering

Major: HVAC & Gas Engineering

Supervisor: Peng Xu

学位论文版权使用授权书

本人完全了解同济大学关于收集、保存、使用学位论文的规定，同意如下各项内容：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以赢利为目的的前提下，学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：

年　　月　　日

同济大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年　　月　　日

摘要

建筑能耗预测是提高建筑能源利用效率、减缓全球变暖趋势的重要手段。在建筑的规划设计阶段，精确的能耗预测有助于实现能源设备的合理配置，在运行及改造阶段，能耗预测可以作为设计和选择合适的节能方法的工具，短期多步能耗预测可集成到基于模型预测的建筑系统运行控制中，预先优化系统运行方案，以实现调峰或降低运行能耗。目前主要有两类建筑能耗预测方法：（1）基于物理模型，（2）基于数据驱动模型。基于物理模型的建筑能耗预测借助能耗模拟软件，需经历信息收集、几何建模、系统建模、模型调试等几个过程，但由于参数及模型本身的不确定性，计算结果往往与实测值存在偏差。近几年，基于数据驱动模型的能耗预测方法得到了广泛关注，它从历史数据中挖掘信息，不需要繁复的建模过程，但能够取得令人满意的预测精度。目前绝大部分基于数据驱动模型的能耗预测模型训练因实际案例而异，以目标建筑历史能耗作为训练数据，模型只能反映该建筑的能耗使用特征，不能迁移到其他建筑。为了拓展数据驱动模型的应用场景，本课题以关键变量提取、数据融合方法为基础，创新性得提出了混合能耗预测模型的构建思路；并以酒店类建筑空调运行能耗为预测目标，通过已有数据资源训练得到了酒店类建筑的混合能耗预测模型。

首先，针对与建筑空调能耗相关的变量，分成负荷相关及系统相关两个层级，对各自变量分别进行敏感性分析和关键变量提取。在分析负荷相关变量时，除常规设计参数外，增加了表征施工质量的附加变量，引入楼板线性透过率、墙角线性透过率、玻璃线性透过率来描述冷桥效应；对于系统相关变量，除系统理想运行状态下的参数外，另外加入了换热盘管污垢系数、冷却塔填料阻塞率、风系统过滤器阻塞率、冷冻水供回水温差、设备效率等附加变量，用以描述系统处于低效运行状态下的特征。此外，为协助算例生成，本文开发了一个参数分析工具包，相比于现有工具功能更全面灵活。

其次，针对基本信息缺失，但有历史能耗数据记录的建筑，为确定其关键变量的取值，提出了数据融合方法，通过代理模型实现实测数据和模拟数据的融合。首先用模拟数据对实测数据进行修正，去除其中的异常值和噪声；而后利用经过修正的实测数据对建筑空调能耗关键变量进行贝叶斯推测，结合预设的关键变量先验分布和以实测能耗值为观测值的似然函数，得到关键变量的后验分布，选定分布的均值作为关键变量的推测值。于是，执行数据融合后的建筑拥有完整用能信息画像，以及能反应其用能概况的代理模型，基于代理模型对该建筑在时间和

空间两个维度进行用能数据填补，为后续混合能耗模型的构建提供数据基础。为验证数据融合算法的有效用，分别以经过校验的模拟建筑和经过详细调研的某五星级酒店为对象进行算法验证。结果表明，提出的数据融合算法能很好地修正存在异常值的实测数据，用经过修正的实测值推测出的关键变量值比较接近实际值。

再次，分析了建立混合能耗模型的三类特征参数，分别为室外气象相关特征、人员活动相关特征及建筑围护结构及系统相关特征。针对前两类特征，在初始特征的基础上构造了更高维的衍生特征；第三类特征即为前面分析得到的建筑空调能耗关键变量。据此，搭建了建筑信息及能耗数据库，该数据库的能耗数据标签同时包括了来自分项计量平台的细颗粒度能耗数据、能源审计报告的逐月能耗账单和基于模拟软件计算的能耗数据，这三类非结构化数据经过融合处理、颗粒度转换后形成了由一张建筑信息主表和对应的 6 张设备能耗数据副表组成的结构化数据，可以直接用来训练数据驱动能耗预测模型。另外，考虑进行能耗预测是时存在未知输入特征的情况，提出了能耗的非确定预测方法。

最后，基于实测和调研数据对冷机能耗混合预测模型和总制冷能耗混合预测模型分别进行模型训练和交叉测试。冷机能耗预测模型由分项计量平台逐时能耗数据和模拟数据融合后训练得到，其平均测试均方根误差变异系数（CV-RMSE）为 0.17，总制冷能耗预测模型由能源审计报告中的逐月账单数据和模拟数据融合后训练得到，由于推荐关键变量的观测数据点较少，增加了推测值的不确定性，其平均测试 CV-RMSE 为 0.28。另一方面，针对存在未知信息的能耗非确定预测方法进行了验证，结果表明，当缺失的特征越来越多时，模型的预测结果不确定性越来越大，给出的预测区间越来越宽泛；缺失的特征越重要将给模型带来越大的不确定性。

综上所述，本课题提出的数据融合方法，实现了模拟数据和实测数据的有机结合，取两者之所长降低了彼此的不确定性，从而得出完整的建筑用能画像；基于关键变量和数据融合方法建立的混合能耗模型实现了建筑基本信息与能耗值之间的非线性映射，可用于多场景能耗预测，为规划设计、系统优化运行提供有力支持。本课题建立的框架和方法同样适合于其他类型的建筑及冬季供热能耗。

关键词： 能耗预测，数据驱动模型，敏感性分析，数据融合，混合能耗模型

ABSTRACT

Building energy consumption prediction plays an import role for improving building energy efficiency and slowing down the global warming trend. Precise energy forecasting helps to achieve rational sizing of energy equipment during the architectural planning and design stage. While in the operation and reconstruction phase, energy consumption prediction can be used as a tool to achieve better energy conservation. Short-term multi-step based energy prediction model can be integrated into the building control system to optimize system operation strategies in advance, helping to realize the peak demand shaving and reduce the energy consumption. At present, there are mainly two kinds of building energy consumption prediction methods: (1) physical based model and (2) data-driven model. Physical based energy consumption prediction model needs to go through several processes such as information collection, geometric modeling, system modeling and model debugging with the help of energy simulation software. However, due to the uncertainty of parameters and the model itself, the simulated results often deviate from the measured values. In recent years, the data-driven energy prediction model has attracted widespread attention. It mines information from historical data and does not need complicated modeling process, but achieves satisfactory prediction accuracy none the less. However, most data-driven energy prediction models are developed case by case because they are trained using the historical energy consumption data of the target building. So the models only reflect the energy consumption characteristics of the target buildings and cannot be transferred to other buildings. In order to expand the applicability of data-driven models, this paper innovatively proposes the idea of hybrid energy prediction model which is able to predict the energy consumption of a building without historical energy record. This model takes key variables as input features and field-test energy data fused by simulated data as output.

First of all, sensitivity analysis and key variables extraction are carried out to select the variables that have major influence on building HVAC energy consumption. This process is conducted separately for load-level and system-level variables in order to reduce computation burden. During the stage of load-level key variable identification, variables representing the construction quality are added in addition to

the conventional design parameters. Linear transmittance of floor, linear transmittance of corner and linear transmittance of glass are introduced to describe the cold bridge effect. For system-level variables, in addition to the variables discussed when the system is operated on ideal condition, variables including fouling coefficient of heat exchange coil, blocking rate of cooling tower packing, blocking rate of air system filter, temperature difference between supply and return water, and low efficiency of equipment are added to describe the characteristics of the system under the low efficiency operation state. To assist the generation of computation samples for sensitivity analysis, a parameter analysis toolkit has been developed. It is more comprehensive and flexible than the existing tools.

Secondly, a data fusion method is proposed to deal with buildings with missing basic information. This method uses an agent model to fuse measured energy data and simulated data. Then the fused data are used to infer the values of missing key variables. To be specific, the simulated data are firstly used to correct the measured data to remove the outliers and noise. Then, the modified measured data were used to make Bayesian inference on the missing key variables. Combined with the pre-determined prior distribution of key variables and the measured energy consumption value as the likelihood, the posterior distribution of key variables is obtained. The mean value of the distribution was selected as the predicted value of the key variables. Therefore, the building gets a complete basic information and energy use portrain and an agent model that can reflect the energy use feature of this building after conducting data fusion. Based on the agent model, the energy use data of this building are extended in both time and space dimensions, which facilitates the development of the hybrid energy prediction model. In order to verify the effectiveness of the data fusion mechanism, a simulated hotel building model and a five-star hotel located in Shanghai are analyzed. The results show that the proposed data fusion algorithm can correct the measured data with outliers well, and the values of the missing key variables inferred from the corrected measured values are close to the true values.

Thirdly, three types of input features are selected for the establishment of the hybrid energy prediction model, including the outdoor meteorological parameters, the personnel activity features and the features representing building envelope and system characteristics. For the first two kinds of features, the more informative features are extended on the basis of the initial features. The third type of features are the key variables discussed in the previous sections. Accordingly, the database containing

structured data of aforementioned three types of features and corresponding daily energy consumption tag is established. The energy consumption data is mainly contributed from three sources including energy sub metering platform, energy audit report and simulated energy models. With the help of aforementioned data fusion algorithm, these three types of heterogeneous data are fused to construct the structure database. This database is composed of one primary table storing input features and six sub tables storing energy consumption data of six different HVAC plants. The primary table and sub tables are correlated using keys. In addition, a non-deterministic energy consumption prediction method is also proposed to cope with cases when the values of input features are hard to obtain.

Finally, the hybrid chiller energy prediction model and hybrid total cooling energy prediction model are developed and cross-tested. The hybrid chiller energy prediction model is trained using data from energy sub-metering platform and simulated energy models, the average CV-RMSE of this model is 0.17. While the hybrid total cooling energy prediction model is trained using data from energy audit report and simulated energy models, the average CV-RMSE of this model is 0.28. Besides, an experiment is conducted to validate the non-deterministic energy consumption prediction method. The result shows that the energy prediction uncertainty is higher when more input features are unknown and vice versa.

In conclusion, the data fusion method proposed in this paper successfully integrates simulated data and measured data to reduce the uncertainty of each other, and therefore to obtain a complete building basic information and energy use portrait. The hybrid energy models are then developed to map the nonlinear relationship from building and system characteristics to HVAC energy use. So the future energy use of a building can be predicted using this model as long as its key variables are known regardless of its historical energy use data is available or not. The method proposed in this paper is able to provide strong support for applications including building planning, design and optimized operation and control. The framework and method established in this paper are also suitable for other types of buildings and heating energy consumption prediction.

Key words: Energy consumption prediction, data-driven model, sensitivity analysis, data fusion, hybrid energy consumption model

目录

表格索引

图片索引

符号注释表

符号	含义	单位
NWWR	窗墙比（北）	
SWWR	窗墙比（南）	
EWWR	窗墙比（东）	
WWWR	窗墙比（西）	
AREA	建筑面积	m^2
NL	层数	
CR	体形系数	
WALLU	外墙传热系数	$\mathrm{W}/(m^2K)$
WSP	外墙热容	$\mathrm{J}/(kg\ K)$
RU	屋顶传热系数	$\mathrm{W}/(m^2K)$
WINU	窗玻璃传热系数	$\mathrm{W}/(m^2K)$
SHGC	窗玻璃太阳辐射得热系数	
WSA	外墙太阳辐射吸收系数	
RSA	屋顶太阳辐射吸收系数	
SPC	空调制冷设定温度	℃
SPH	空调供热设定温度	℃
LPD	照明功率密度	W/m^2
OPD	人员密度	P/m^2
INFIL	冷风渗透率	ACH
ST	内遮阳开启程度	
FLT	楼板线性透过率	$\mathrm{W}/(m\ K)$
GLT	玻璃线性透过率	$\mathrm{W}/(m\ K)$
CLT	墙角线性透过率	$\mathrm{W}/(m\ K)$
$sigma$	建筑平面形状系数	
A_{total}	建筑面积	m^2
A	建筑平面面积	m^2
NL	建筑层数	
CV-RMSE	均方根误差变异系数	
y_k	实测值	

$\widehat{y_k}$	预测值
η_k	模型的模拟值
Γ_k	输入参数带来的偏差
Δ_k	模型本身造成的偏差
ε	观测误差
x	可观测可控制的变量
θ	被推测的未知变量
t	模型中的未知参数
y	观测值
η	模拟值
δ	实际模型与真实过程之间的差异
π	分布函数
$k(x, x')$	径向基核函数
μ	高斯过程的均值函数
Cov	高斯过程的协方差函数
Σ_y	观测值协方差矩阵
I_n	n阶单位矩阵
λ	高斯过程模型的精度超参数
β	高斯过程模型超参数
α	高斯模型的平滑性超参数
r	周期因子
t	统计因子

第1章 绪论

1.1 研究背景

本课题来源于重点研发计划培育：基于大数据的绿色建筑管理技术。

1.1.1 建筑能耗预测

建筑行业在全球范围都是最大的一次能源消耗行业之一，占全球能源的 30%以上[1]。在中国所有的建筑服务系统中，超过 40%的能源用于支持暖通空调系统。相关研究表明，大多数暖通空调系统具有不同水平的节能潜力，从 15%到 30%不等[2]。从实用性和科学性的角度考虑，可采用包括更换高性能的围护结构材料、系统设备；采用节能新技术；优化建筑空调系统设计和运行以提高其运行效率以及使用能源控制和监控系统提高建筑能效等方法提高建筑能效。在上述提高建筑能效的方式中，都需要提前预知采用节能措施后的建筑能耗（即“能耗预测”）来判断所采用措施的有效性和经济性，因此进行准确的能耗预测是开展节能减排工作必不可少的环节。进行能耗预测主要有两种方法，一是基于物理模型；二是基于实测建筑能耗建立数据驱动模型进行预测[4]。物理模型依靠显式的热力学规律，其公式和计算机理易于理解。基于物理方法的模型也被称为白箱模型。广泛使用的仿真工具包括 EnergyPlus，DOE-2，TRNSYS，eQuest 等都是基于物理模型开发的。随着计算机技术的不断发展，上述建筑能耗模拟工具（BEM）已被广泛用于建筑优化设计，施工，运营管理等方面。这类 BEM 软件通常使用前向方法来进行建筑能耗模拟和性能分析，将建筑物几何形状，建筑材料，建筑物使用时间表，空调系统配置和控制策略作为输入参数。但是这些 BEM 软件普遍存在以下缺点：

（1） 建筑几何建模过程过于复杂；
（2） 输入参数太多，且不知道哪些参数对模拟结果的影响较大；
（3） 计算时间长，特别是对于几何外形或系统设置复杂的建筑[5]；
（4） 能耗模拟结果和实际能耗之间存在较大差异，如图 2 所示，即便是对于模型要求很高的 LEED 认证项目，其能耗模拟值与实测值之间也存在很大差异。

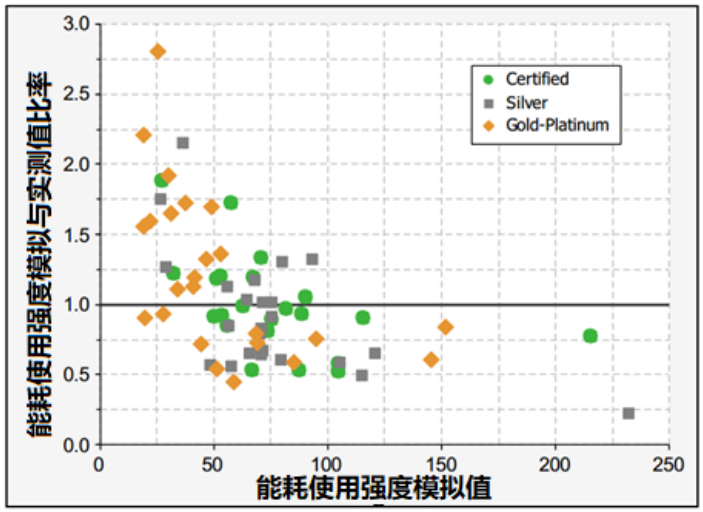


图 1.1 获 LEED 认证的建筑能耗模拟值与实测值对比

模拟结果与实测能耗之间的较大偏差是最常被抱怨的问题。这主要是由于计算机模型的不确定性，包括参数不确定性和模型本身不足[6]。模型不足是基于一个简单的事实，即没有物理模型是完美的，因为物理模型的建立通常需要将复杂问题进行近似和简化。另一方面，能耗模型需要大量参数作为模型输入，大多数参数都来自建筑设计文档，但是由于施工质量差、空调系统运维不当和设备老化等问题，某些参数常常偏离设计值。其他参数如人员在室率和新风渗透率本身具有很强的不确定性，并且难以直接测量。

数据驱动模型（也称为黑箱模型）的构建则不需要如此详细的数据。近年来，数据驱动方法因其高效、准确而受到越来越多的关注。数据驱动模型的最大优势之一是不需要构建复杂的物理模型。输入和输出变量之间的关系可以通过机器学习和人工智能等高级数据分析自动捕获。数据驱动建模使用统计方法来捕获输入、输出数据背后的关系，而不是用物理方程[7]。因此它特别适合解决复杂的非线性问题。数据驱动模型还具有内置误差项的优点，因为模型参数是使用现实世界中的大量观测数据来估计的，可以量化错误并估计置信水平。此外，数据驱动模型计算速度快，一旦构建好了数据驱动模型，就可以立即计算出输出，这比基于物理的模型更优越。但数据驱动模型的构建需要大量的训练数据，其性能也高度依赖于训练数据的质量，而且使用统计方法建立的模型也很难进行直观解释。

使用数据驱动模型进行能耗预测，根据训练数据的不同又可分为两种类型，一种是根据目标建筑自身的历史能耗进行未来能耗的预测，这种类型的预测问题建模相对简单，目前研究和应用比较成熟，建筑本身对能耗造成影响的特征（包括建筑围护结构特征、人员变化、运行策略等）可以不用反应的模型中，因为目标对象不变，一般来说上述特征参数在较短期的时间段内不会有太大变化。另一

种是利用其它建筑的能耗对目标建筑进行“迁移”预测[1]，这种类型的预测方法适用于没有历史能耗数据的建筑。这种预测模式需要考虑与建筑围护结构及内部系统相关的可能会对能耗造成影响的特征差异，因此要建立第二种类型的能耗模型，需要识别出对能耗有影响的特征，并确定特征对应的取值，作为数据驱动模型的输入，建立上述特征与能耗之间的映射关系。这是以往的研究中没有涉及到的。

基于数据驱动模型进行能耗预测的关键是大量且高质量的能耗数据。现场采集的建筑能耗，即实测数据，反映的是建筑用能的实际情况。但是由于计量装置和数据传输等方面的问题，采集到的数据质量不高，存在异常值、断点、噪声等多种问题，一直以来这是难以克服的问题。另外，要对一个楼宇进行全方位的监测，需要安装大量的测试装置，还需配备相应的采集系统，成本高昂，因此国家规定的能耗监测平台采集到的的数据种类很有限，仅包括照明、空调、设备、动力等几大类。相比之下，借助能耗模拟软件可以计算得到几乎任何想要的数据，完成各种各样的分析对比，并且能耗模拟软件的计算结果不存在缺失值、离群值等异常数据点。另一方面，虽然模拟软件由于输入参数的不确定性、以及模型本身的不确定性，其计算结果与真实值之间存在着偏差，但其计算结果的趋势是可靠的，比如能耗随温度的变化趋势、改变某些参数值后能耗将发生何种变化。

1.1.2 建筑运行智能监测系统的利用

2007 年，住建部、财政部发布了《关于加强国家机关办公建筑和大型公共建筑节能管理工作的实施意见》，此后我国各大城市陆续开始推动公共建筑节能监测及分项计量工作， 33 个省市陆续开展了公共建筑能耗监测平台建设工作。建筑能耗监测平台主要有三种：①应用于建筑单体的能源监测平台；②由政府、高校、园区等管理部门建立的区域型中心能源监测平台；③商业地产出于自身能耗管控需求建立的中心能源监测平台。这几类平台在数据的丰富度、数据存储标准的统一性、数据质量等方面都还存在不足。总体来说，目前国内外尚没有基于标准数据库、接入上百栋不同类型公建及数据类型、且嵌入完善的数据质量保障技术和机制的能耗监测平台的先例。经过多年运行，已积累了大量实时运行参数和能耗数据，随着纳入监测平台的建筑越来越多，数据量越来越大。目前，基于公共建筑能耗监测平台的应用实践主要包括以下几个方面[3]：

（1） 开发互联网用能管理产品，为用户提供建筑整体能耗分布情况及自身能耗水平；

（2） 推动节能改造工作和建筑用能标准的建立；

[1] 区别于深度学习算法中的“迁移学习”。

（3） 深化数据分析研究，包括建筑能耗预测、用能诊断、需求响应等。

就目前的数据使用情况而言，分项计量平台数据的使用还停留在比较初级的阶段，例如提供建筑整体能耗的分布情况、根据当前建筑水平从而建立用能标准，仅是对大量实测数据做简单的分析处理，得到较宏观的数据分布和建筑用能情况，并未深入到建筑内部，图 1.2 是根据上海市公共建筑分项计量平台数据汇总得到的各类建筑用能占比。另外，能耗分量计量平台采集的数据大多存在质量不佳的问题，导致很多数据无法得到有效利用，多地的分项计量能耗数据采集平台已经停止工作。

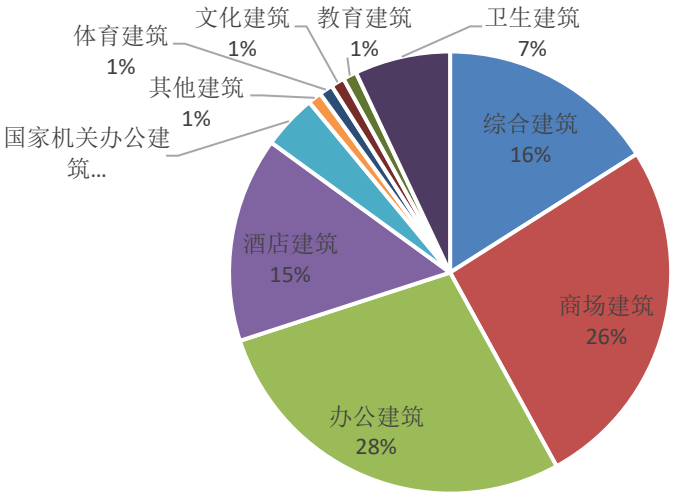


图 1.2 2017 年上海市能耗监测平台联网的建筑用能量占比情况

除了公共建筑能耗监测平台外，很多建筑还安装了建筑智能管理系统（BMS），监测的数据除了建筑能耗外还包括设备运行参数、环境参数等，相较于能耗监测平台的数据更加多为和复杂。与分项计量平台数据不同的是，每个建筑的 BMS 系统是独立的，不兼容，因此很难将其大规模合并，进行系统化应用。

来自建筑分项计量平台的能耗数据和建筑自身的运行数据构成了建筑大数据。很明显，建筑大数据呈现海量、复杂、多维、动态的特点。随着建筑信息技术飞速发展，建筑大数据密度、维度和复杂度不断提高，大量积累的建筑运行/能耗数据也带来了“数据灾难”和“维数灾难”。然而目前建筑领域与大数据挖掘领域仍存在知识鸿沟，现有的数据挖掘方法往往只基于历史数据对能耗进行静态分析，现有的预测模型和分析方法并不能对建筑大数据进行充分应用。如何科学合理地采集、存储、分析和应用这些数据成为亟待研究的课题。

综上所述，目前研究比较成熟的白箱模型和黑箱模型预测都存在不足，海量的实测数据未得到有效的利用，本课题提出的数据融合算法和混合能耗模型实现

了三者的有机结合，为建筑能耗预测提供了新的思路。

1.2 国内外研究现状

1.2.1 建筑大数据分析研究现状

与建筑机电系统相关的数据主要包括五个方面：

（1） 建筑本身物性参数，包括窗墙比、体形系数、墙体及窗户传热系数等；

（2） 系统设备参数，包括冷机水泵等设备的容量、效率等其他性能参数；

（3） 环境参数，包括室内外环境参数，例如温湿度、CO_2 含量、太阳辐射相关参数等；

（4） 系统运行参数，包括机组设备和供冷/热介质的状态参数，例如送风温度、冷机出水温度、流量等；

（5） 系统能耗参数，包括各个设备的耗能量，能源种类包含电、气等。

上述参数中，前两项为静态参数，建造完工后即确定了，后三项为动态参数，随时间不断改变，是时间序列参数。这些不断变化的数据中蕴含着很多信息，通过数据挖掘的方法可以发现建筑运行过程中隐藏的特征，进行故障诊断、负荷（能耗）预测等，相比于传统的基于经验的方法更精准高效，数据处理的一般过程如图 1.3 所示。

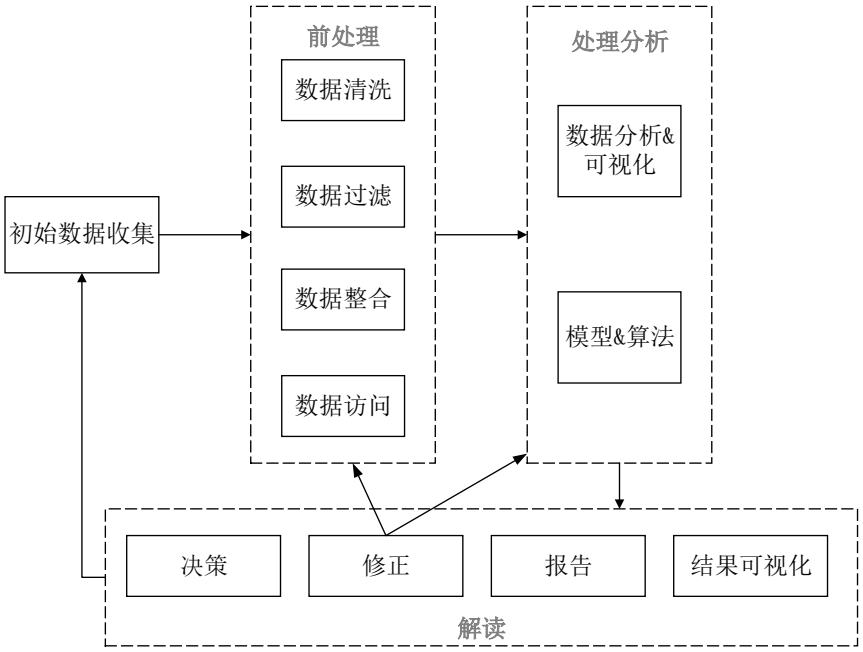


图 1.3 数据处理一般过程示意图

从技术角度讲，数据科学包括一系列的方法和工具，适用于不同的问题和情

况，其中比较广泛使用的包括分类、聚类、回归、关联规则挖掘、序列分析等。

分类是指基于每个对象集的特征对其赋予恰当的标签。支持向量机方法[8]是常用分类算法，建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折中，以求获得最好的泛化能力。决策树也是比较常用的数据分类方法，同时它还具有可视化的功能，可读性强[9]。随机森林、LightGBM、XGBoost 等是基于决策树模型的构建的集成算法（Ensemble Learning），通过组合多个弱分类器，最终结果通过投票或取均值，使得整体模型的结果具有较高的精确度和泛化性能，相比与单一模型，集成模型有更高的精度，因此在工程及算法比赛中被广泛使用。除此之外，贝叶斯分类、神经网络等方法也被用于分类[10]。

回归分析主要是用来建立输入和输出变量之间的映射关系，被广泛应用于预测，选择回归变量时一个很重要的条件是输入变量之间相互独立，否则会造成冗余[12]。回归算法可分为线性和非线性两类，线性算法较简单，只能用来表示输入输出之间的线性映射关系，常用算法有线性最小二乘法、贝叶斯线性回归等，但是由于现实情况中大部分情况输入输出的关系非常复杂，不是简单的线性关系能表达的，因此非线性回归更加应用广泛，上述提及的算法包括人工神经网络（ANN）、支持向量机回归(SVR)、以及基于树模型的集成模型都是常用的回归算法，其中集成模型在回归问题中相较于单一算法模型通常也有更好的表现。

聚类分析是寻找数据内部的分布结构，将其划分成若干簇[13]。常用的聚类算法可以分为：原型聚类、层次聚类和密度聚类。聚类算法通常是非监督学习，因此在没有额外信息的情况下，聚类通常是数据分析的第一步。 Han 和 Kamber 将聚类方法分为五类：分割法、层次法、密度法、网格法和模型法[14]。常用的聚类算法有 k 均值[15]、模糊 c 均值[16]、自组织网络（SOM）[17]、学习向量量化、高斯混合聚类、DBSCAN 等[18]。 对于时间序列数据的聚类处理方法，主要有三类，如下图所示[19]。聚类算法在各领域都有广泛应用，总结见表 1。

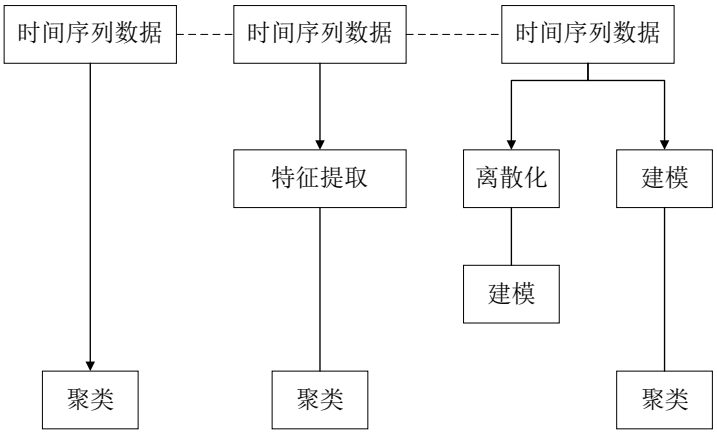


图 1.4 三种聚类方法（基于原始数据/特征提取/模型）

表 1 聚类算法在各领域应用案例

文献	聚类方法			算法	应用
	基于原始数据	基于特征	基于模型		
Golay 等[20]	√	/	/	模糊 C 均值	核磁共振图分析
Liao 等[21]	√	/	/	k-均值和模糊 C 均值	电池模拟
Liao 等[22]	√	/	/	k-中心遗传聚类	DNA 微阵列
Fu 等[23]	/	人为选取重要数据点	/	改进 SOM	香港股市分析
Owsley 等[24]	/	冬天区域的时频域表达	/	改进 k 均值	工具状态监测
Vlachos 等[25]	/	哈尔小波变换	/	改进 k 均值	/
Wilpon 等[26]	/	LPC 系数	/	改进 k 均值	孤立词识别
Biernacki 等[27]	/	/	高斯混合	组合学习	/
Kalpakis 等[28]	/	/	AR	分割算法	公共数据分析
Li 等[29]	/	/	连续马尔科夫模型	四层交织搜索	生物
Tran 和 Wagner 等[30]	/	/	高斯混合	模糊 c 均值	语者确认
Xiong 等[31]	/	/	ARMA 混	组合学习	公共数据分析

			合		

关联规则分析是从数据背后发现事物之间可能存在的关联或者联系，该方法通过分析两个变量同时出现的频次来来确定形如 A→B 的因果关系[32]。Apriori 算法及其改进形式是最常用的关联规则分析算法。支持度和置信度是判断关联规则是否显著的两个常用指标，关联规则挖掘就是从数据集合中挖掘出满足支持度和置信度最低阈值要求的所有关联规则。

时间序列（或称动态数列）是指将同一统计指标的数值按其发生的时间先后顺序排列而成的数列。时间序列分析的主要目的是根据已有的历史数据对未来进行预测[33]。时间序列分析是定量预测方法之一，包括一般统计分析(如自相关分析，谱分析等)，统计模型的建立与推断，以及关于时间序列的最优预测、控制与滤波等内容。

大数据分析技术已经被广泛应用于提升建筑机电系统能效和管理水平，其主要应用场景包括：负荷（能耗）预测、优化运行管理、系统故障诊断、能耗影响因素分析等。

建筑负荷（能耗）模式的发现非常复杂，因为相关因素很多，Kusiak 等人构建了人工神经网络来描述气象参数与建筑负荷之间的关系[34]。Yu 等人则提出用决策树的方法来建立建筑用能需求的预测模型，决策树相较于其他方法更加直观[35]。准确预测建筑峰值负荷对于合理配置机组和调整运行策略很重要。以往常用回归的方法预测负荷峰值，近几年有学者尝试用其他方法进行峰值预测。Li 等人建立了决策树来分析能耗峰值的外部影响因素，并通过建立不同气候条件下的回归模型来预测后一天的能耗峰值[36]。Yang 等人则用 C4.5 分类算法来分析室内外环境，并确定外部因素对室内热舒适的影响情况[37]。Fan 等人提出了预测能耗和峰值负荷的组合模型。该方法包括三个阶段，首先用特征选择和聚类的方法提取能耗异常部分，然后用递归特征剔除的方法选择每个子模型的优化输入参数，最后建立组合模型，其权值通过遗传算法确定[38]。

现在 BAS 系统已经广泛安装于公共建筑中，在建筑运行过程中，BAS 系统记录了大量数据，包括温度、湿度、流量、压力、设备运行状态等，这些数据可以被用来提取运行规则，以 IF-THEN 的形式展示[39]。 May-Ostendorp 等人用不同的分类算法从离线模型预测控制数据中提取了运行规则，在此基础上作者得出了建筑在制冷季的优化控制策略[40]。虽然有很多算法可以得到 IF-THEN 规则，但是大部分研究采用分类算法，特别是决策树。相较之下，另一种规则挖掘算法—关联规则的应用较少，这种算法可以发现数据中隐藏的关系，这些关系即便是有经验的运行人员也很难发现[41][42]。

建筑是一个惯性较大的系统，其运行数据前后联系很紧密，因此历史运行数据经常被用来进行故障诊断和预防。通过不断监测建筑运行数据，如果故障出现可以通过数据发现并且分析其如何影响其他设备，模式识别和回归算法在这方面比较常用。Capozzoli 等人基于照明能耗和总能耗历史数据构建了一个耗能设备自动故障诊断的简化方法，该方法结合了神经网络和其他离群值检测器[43]。Sedano 等人提出了一个类似的方法来发现保温材料的缺陷[44]。Daniel B. Araya 等人提出了一个基于能耗模式的异常监测方法 CCAD-SW，该方法采用重叠滑动窗口进行异常能耗监测；另外在 CCAD-SW 的基础上提出了综合异常监测方法（EAD），该方法集成了多种异常检测分类器，用实测数据验证表明就敏感度 EAD 比 CCAD-SW 提高了 3.6%，误报率降低了 2.7%[45]。Cheng Fan 等人提出了基于自编码的建筑能耗异常检测集成方法，并对不同的自编码类型和训练方法进行对比[46]。

供电公司经常借助于数据技术分析用户的用电习惯和特征，以此来提高自己的供配电效率。 Félix Biscarri 等人提出了一个对用户用电特征进行自动聚类的框架，在这个框架中聚类算法是自动选择的，新用户会根据其特征被赋予预先定义好的标签[47]。Enrico Carpaneto 等人总结了基于时域数据的负荷特征识别研究，但是该技术需要存储大量时域数据，于是作者提出了基于频域数据的特征识别方法，在保证识别精度的基础上可以大大降低数据存储压力[48]。

建筑是一个多维度的复杂系统，其能耗影响因素众多，从建筑本身形态、系统形式到运行方式，都会对能耗造成影响，如何从众多变量中找出主要影响因素也是目前研究的一个热点。Hao Zhou 等人通过数据挖掘的方法研究了天津住宅建筑供热能耗的影响因素。首先采用箱形图的方法进行异常值监测和能耗等级分类，然后采用信息增益率为供热能耗影响因素排序，最后用 4.5 决策树建立关系分类规则[49]。Hongting Ma 等人对中国北方的 119 懂公共建筑能耗进行了分析，得出办公、医院、学校这三类建筑的能耗分布特征，并基于能耗模拟软件 eQuest 对能耗影响因子进行了分析[50]。

1.2.2 基于数据驱动模型的建筑能耗预测研究现状

1.2.2.1 算法总结

如前所述模型的建立方法主要有两种，分别是前向模型和数据驱动模型。前向模型也称白箱模型，是基于对象系统的物理特征和物理规律（例如热质平衡、动量质量守恒等）建立的。建立前向模型需要对对象特征有深入的了解，从简单的关系开始构建描述整个系统的复杂模型。在暖通空调领域，白箱模型是非常常

用的建模手段，被用来建立冷机[80][81][82]、冷却塔、房间[83][84]、混合箱[85]、冷热盘管[86][87][88]、风机水泵[83]、传感器[82]等对象的模型。我们常用的能耗模拟软件本质上也是一种复杂白箱模型。基于数据驱动的模型也叫黑箱模型，其建立的基础是有对象系统运行过程中产生的大量试验数据，用统计的方法寻找变量和输出之间的映射关系，而不需要了解系统的工作原理，因此也被称为逆向模型。逆向建模方法特别适合于建立复杂系统的模型，因为这种系统往往很难用明确的物理过程描述。在建筑相关的各个研究领域逆向模型收到越来越广泛的关注，例如建筑设计[89]、能耗负荷预测（如表 2 所示）、建筑预测[90][91]、空调系统故障诊断[92][93]等。

建筑能耗预测与负荷预测非常相似，除了负荷预测用到的特征，还需考虑机电系统的特征。图 1.5 总结了进行能耗预测的常用方法，用于建筑能耗预测的数据驱动模型一般可分为单变量预测模型和多变量预测模型，如图 1.5 所示。单变量预测模型仅依靠时间序列的历史值来预测其未来值。它们不需要进行特性工程。代表性算法有自回归模型(AR)、移动平均模型(MA)和自回归积分移动平均算法(ARIMA)。一些研究表明，单变量预测模型不如其他数据驱动模型准确[125][126]，因为它们不能捕捉目标变量和外生变量之间的关系。多变量预测模型，顾名思义，是通过建立预测目标与多个变量之间的映射关系来预测未来的值。机器学习方法属于第二类。传统的机器学习模型(如人工神经网络、支持向量机和随机森林等)需要结构化数据来训练模型。合理的特征选择可以极大地帮助改善模型性能[127]。深度学习模型作为神经网络算法的一种，能够从原始数据中提取特征并自动降维生成特征[128]。

表 2 基于数据驱动模型的建筑负荷估算研究案例汇总

文献	特征变量	输出变量	模型	研究目的
[94]	体形系数、窗墙比、在室人员、单位照明/设备功率密度、室内空调设计温度	冷负荷指标 (W/m^2)	ANN	预计建筑最大负荷
[95]	窗户、墙体、地板面积，墙体及窗户构造、室内设计温度	热负荷指标 (W/m^2)	ANN	用最小的数据维度预测不同类型建筑的热负荷指标
[96]	平局室内空气温度、辐射温度、CO_2 含量、照度，照明设备能耗、室外干球	逐时制冷能耗 (kWh)	ANN	用少量输入参数预测逐时制冷能耗

	温度，相对湿度，降雨量，风速			
[97]	季节，保温，墙体厚度，传热系数	全年供能耗 (kWh)	ANN	针对被动太阳房建立墙体性能到制热能耗的映射关系
[98]	建筑长宽比，墙体总体传热系数，体形系数，总外墙面积，总窗户面积	供热负荷 (kW/m^2)	ANN	针对既有建筑，从几何形状、构造布局和天气情况等几方面进行能耗预测，并与能耗模拟软件（KEP – IYTE-ESS）进行对比
[99]	总建筑面积，层数，进深，长宽比，朝向，屋顶传热系数、颜色、反射率，窗户传热系数、遮阳系数，窗墙比	全年能耗(kWh)	MLR	建立快速能耗预测模型，供建筑师在设计前期进行分析
[100]	朝向，保温层厚度，窗墙比	供热能耗 (Wh)	ANN	在设计前期为围护结构设计参数的选取提供快速分析工具
[101]	朝向，屋顶保温系数，地板构造等	全年能耗 (kWh)	MLR	针对美国五个气候区的办公建筑进行快速能耗分析
[102]	相对紧凑度，墙体面积，屋顶面积，建筑总高度，朝向，窗户玻璃面积及分布	制冷/供热负荷 (kW)	ELM	分析不同机器学习模型在预测建筑能耗方面的差异
[103]	建筑形状，相对紧凑度，窗户玻璃面积，屋顶面积，墙体面积，朝向，高度，窗户分布	制冷/供热负荷 (kW)	ELM	分析各变量与建筑负荷的关系及其紧密程度

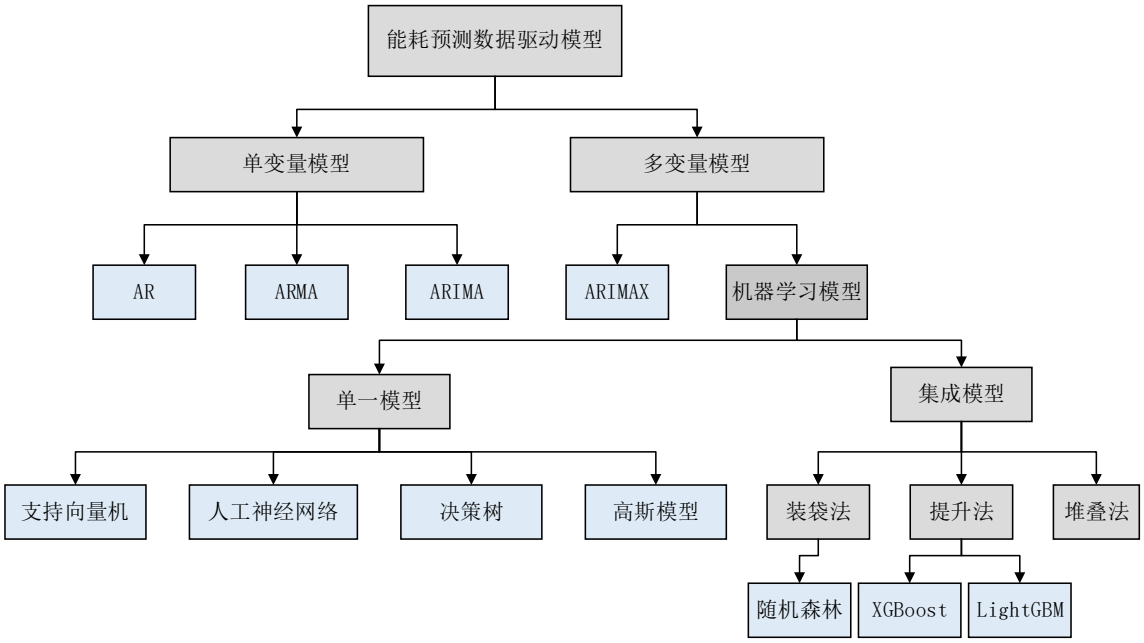


图 1.5 建筑能耗预测数据驱动模型算法总结

1.2.2.2 特征分析

对于机器学习模型的建立，最重要的一步是特征工程。特征是一个可能对预测目标产生影响的变量。对于建筑空调能耗的预测问题，特征参数可以是可观测的变量，如温度和太阳辐射或集合从原始数据，也可以是由若干特征经过组合得到的高维特征。经过精心选取和设计的特征往往含有更多的信息，因此可以帮助模型实现更高的预测精度，并且好的特征不需要借助复杂算法也能得到高性能，使用简单算法的效果是相近的。采用简单算法的好处是他们结构简单，消耗计算资源小，往往比复杂的算法有更高的稳定性、弹性和适应性。然而，据作者所知，现有的大部分研究论文更多地关注于复杂的模型，而不是特征工程。

总结来说，特征工程主要包括以下两个步骤:

（1） 寻找可能对预测目标有影响的初始特征。这一过程需要丰富的实践经验。除了使用可直接测量的特征，还使用从原始特征创建的新特征。以往的研究已经证实，使用集成特征可以获得比使用原始特征[129]更好的性能。

（2） 特征选择，也称为降维，一般有两种方法，一是从初始特征中选择对预测目标影响较大的若干特征，二是对初始特征进行转化去除特征中线性相关的信息。这一步也被称为特征提取。由于特征维数过大可能会导致维数诅咒[130]，冗余的特征可能会使模型性能下降[131]，因此降维是一个必要的步骤。

（1）初始特征选取

机器学习模型建立了输入参数与预测目标之间的映射关系。获取合适的输入参数是预测目标变化的驱动因素。建筑空调系统的能耗由多种因素决定，可分为以下四类:

- 室外气象参数
- 建筑及机电系统特点
- 室内环境
- 室内人数及活动情况

干球温度、相对湿度、露球温度、太阳辐射、风速等室外气象参数是影响建筑空调能耗的主要因素。几乎所有的研究都使用天气参数作为建筑空调能耗预测的输入特征。其中一些研究使用直接观测的天气参数，而另一些使用经过处理的天气参数，例如制冷度数日(CDD)和供热度数日(HDD)。度日数是一种衡量建筑冷热需求量的简化方法。它集成了平衡温度的信息，表明何时开启供能系统，从而在一定程度上提高了模型性能。

在室人员对建筑能耗的影响主要表现在人员数量和人员用能模式两个方面。人体是天然热源，人数越多散热越大。此外，办公设备能耗与在室人数高度相关。Wei 等[132]的研究表明，对于办公建筑而言，室内人员对能源使用的影响要比天气更重要。然而，在室人数和活动情况很难被统计。所以在大多数研究中，通常用时间指标(即一天的第几个小、一星期的第几天、工作日或周末、假日或非假日等)反映入住和活动情况。

除了上面描述的特征，历史能耗数据经常被用作输入特征。原因有两方面。第一个是建筑在相同的日类型下有相似的能源使用特性。办公建筑的能耗曲线一般以周为单位波动。Fan 等人[133]使用 7 天和 14 天前的能耗作为预测模型的输入。第二个原因是建筑的热惰性。前一时间步长天气情况和能耗可能会影响当前的能耗。它通常用于短期小粒度(例如小时)的能耗预测。Fan 等人[134]使用过去 24 小时的建筑冷负荷、室外温度和相对湿度作为输入特性。

应该注意的是，在选择输入特性时，数据泄漏是一个需要重点关注的问题。数据泄漏是指使用了在预测时实际无法获得的信息，从而导致预测分数高于实际情况[135]。严格地说，使用未来气象参数作为输入特征是一种数据泄漏。但由于准确的天气预报，它的影响可以被忽略。而 Ding 等[135]则使用无法提前获得的冷冻水次日流量作为预测冷负荷的输入特征，预测分数可能被高估了。

（2）特征选择

通常有两种类型的特征选择方法，第一种是从初始特征空间中选择一些重要的特征。下面列出的方法中过滤法、包裹法、嵌入法和敏感性分析属于这一类。

第二种是将最初的特征空间转换为新的特征空间，然后从新的特征空间中选择一个特征子集，该子集的每一维是相互独立的，但是用这种方法构造出的新特征的物理意义很难解释。

- 过滤法（Filter method）

过滤法利用信息理论或相关性分析对每个特征的重要性进行排序，选择其中得分高的特征[136]。Spearman 系数和 Pearson 系数是估计各输入特征与预测目标[138][148]之间相关性的两个常用准则，但这两个指标仅能反映参数之间的线性关系。

- 包裹法（Wrapper method）

包裹法以机器学习预测分数作为评价指标，度量所有可能的特征子集，并找到最优的特征子集[137]。包装法的重点是如何从初始特征空间中找到最优的特征子集。基于贪婪算法的 GreedyStepWise 包装法[140]、对特征空间进行全面搜索，直到当添加或删除任何剩余的特征对提高评估分数没有帮助时停止。这种方法在特征维数较大时，效率很低。有学者尝试了采用有效的搜索策略来降低计算复杂度，其中常用的是基于进化理念的进化算法，如遗传算法等。该算法被证明对求解复杂函数的最优或近优解是有效的[141]。Aurora 等人使用两种多目标进化搜索算法:ENORA 和 NSGA-II 来进行特征选择。Salcedo-Sanz 等人使用改进的 Harmony Search(HS)优化算法选择特征子集[142]。Boruta 算法是另一种有效的包裹[144]。它采用自上而下的搜索算法，通过比较原始特征集和随机打乱后特征集的重要性，将所有的特征标记为重要或是不重要。Huang 等人利用 Boruta 分析计算每个特征的重要性，并采用随机森林选择合适的子集[145]。Candanedo 等人使用 Boruta 包从初始的几十个变量[146]中查找所有相关特征。

- 嵌入法（Embedded method）

嵌入法不同于包裹法，它将特征选择过程与模型训练过程相结合。正则化和树模型属于该方法。正则化对每个模型参数增加惩罚，以减少模型的自由度，避免过拟合。Jain 等人对损失函数使用 L1 正则化(Lasso)，并优于不使用正则化[147]的 SVR 模型。Guo 等人也使用 Lasso[148]进行了特征选择。基于树的模型(如随机森林, LightGBM, XGBoost, CatBoost 等)不仅可以提供出色的预测性能，而且还可以得到特征重要性，作为特征选择的一种方式。Yuan 等[148]采用随机森林选取供暖能耗预测的前 10 个特征。

- 敏感性分析法（Sensitivty analysis, SA）

敏感性分析无论是对于建筑能耗模拟还是实测数据分析都有非常重要的作用。从多个不确定性因素中找出对目标结果有重要影响的敏感性因素，并分析、测算其对目标结果的影响程度和敏感性程度。图 1.6 展示了敏感性分析进行建筑

性能分析的一般步骤：确定输入变量及其参数分布区间、建立能耗模型、运行模型、收集模拟数据、进行敏感性分析、结果展示。

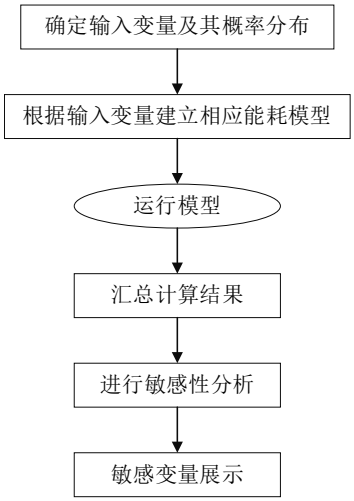


图 1.6 建筑性能敏感性分析典型流程

敏感性分析方法包括局部分析法和全局分析法，所谓局部分析法是仅改变目标变量的值，固定其他变量保持不变，来分析目标变量变化对于整体结果的影响。全局分析是所有变量一起变化，综合分析每个变量对于结果的影响程度。因此全局分析更可靠，但是其弊端是计算量庞大[51]。这两种方法在建筑性能分析中都被广发应用。常用的全局敏感性分析方法有回归法、Morris 法、FAST 法和 Sobol 法等。回归法由于其计算速度快，容易理解，在建筑能耗分析中是最被广泛使用的方法。诸多指标可以用来评估影响因子的重要性，包括标准回归系数（SRC）、偏相关系数（PCC）、标准秩回归系数（SRRC）、偏秩回归系数（PRCC）等。其中 SRC 和 PCC 只适用于线性模型，SRRC 和 PRCC 可以用于非线性模型。[52-57]。Morris 法又称为元效应方法，是一种简单而有效的方法，相对于其他算法，Morris 法的计算量相对较小，但是它不能定量给出每个输入参数对结果的影响大小。Morris 法可以从模型中包含的许多输入因素中筛选出一些重要的输入因素。μ^*和σ是其两个衡量指标，μ^*用于评估输入参数对结果的主要影响大小，σ用于评估输入参数之间的相互影响关系，具体的 Morris 方法的算法介绍见第二章。Sanchez 等人以能耗模拟软件 ESP-r 为计算工作，采用 Morris 方法对住宅建筑用能进行敏感性分析，分析了将一阶敏感型和二阶敏感性进行组合的作用。研究结果表明，一阶敏感性结果可以帮助分析变量之间的关联性，二阶敏感性结果除了可以对变量敏感型进行排序，还能明确变量对之间的相互影响。另外，该研究指出敏感性得到的结果不具备通用性，与分析目的、变量选取有很大关系[58]。Heo 等人在对既有建筑改造进行模型校验时，采用了 Morris 方法进行变量筛选[59]。

Heiselberg 等人在设计可持续建筑时采用了敏感性分析方法从众多变量中筛选出对建筑能耗影响最明显的几个变量，以便更具有针对性[60]。Hyun 等人研究了居住建筑自然通风量的不确定性，与常规的确定性设置进行对比，发现对于住宅建筑，其自然通风的不确定分布对建筑整体能耗影响很大，并且采用敏感性分析方法对不确定性因素的重要程度进行了排序[61]。FAST 法和 Sobol 法将输出结果的差异性拆分到每一个输入变量上，定量地给出输入变量敏感型高低排序，并考虑了变量之间的相互影响关系，但计算量较大。Mechri 等人用 FAST 方法来确定影响意大利典型办公建筑的关键设计变量。分析结果表明，围护结构透明面积比是最重要的因素[62]。Spitz 等人用 Sobol 方法进行了 6669 次模拟，确定了法国一栋住宅建筑影响能耗的 6 个主要因素[63]。

- 主成份分析（Principle component analysis, PCA）

与上述特征提取方法不同，PCA 通过将原始特征映射到变量线性不相关的低维空间来降低特征维数。Ding 等[64]采用 PCA 与小波分解重构、相关性分析相结合的方法，得到合理的模型输入。Li 等[65]分析了 PCA 对建筑负荷预测的效果。结果表明，核主成分分析与常规主成分分析相比，在无特征选择的情况下具有更好的性能。Yuldiz 等人[66]讨论了如何确定特征缩减空间的维数。Ruch 等人采用 PCA 来判定室外气象参数与建筑用电的关系[105]。Reddy 和 Claridge 比较了 PCA 和四参数线性模型的优劣[106]。但是应用 PCA 需要大量的观测数据，对于建筑实际运营数据来讲，由于项目运营周期和数据采集质量的问题，数据量常常不能满足条件。Lin 等人采用了最小绝对收缩和选择算子（LASSO）的方法来解决有限数据集下特征降维的问题[107]。

- 自编码器（Autoencoder）

Autoencoder 的作用与 PCA 类似，都是将高维数据经过处理映射到低维空间。Autoencoder 是一种监督算法，可以压缩输入数据的维数。它常用于深度学习网络中对大规模数据进行降维。然而，由于用于能耗预测的特征空间通常较小[137]，autoencoder 很少用于建筑能耗预测领域。Fan 等人[67]使用不同的方法(包括全链接自编码器、一维卷积自编码器和生成对抗网络)提取了相同数量的特征，并与传统方法比较了它们在建筑能耗预测方面的潜力。

1.2.3 异构数据融合研究现状

在大数据分析的研究范畴里，异构数据是指在类别和格式上差异比较大的数据组合，这类数据常常表现出语义模糊、质量低、冗余等特质，因此有很强的不确定性。现实场景中最常见的异构数据来自物联网（Internet of Things，IoT），物联网数据有三个特征，首先由于数据采集设备量大、空间位置分散，数据在类别

上表现出异构性。其次，物联网采集的数据体量大、持续时间长，需要存储大量的历史数据，因此数据的空间和时间可匹配性也是物联网数据的一个重要特征。最后，数据在采集、转换和传输过程中会引入大量噪声，降低数据质量。物联网数据的异构性通常表现在以下四个方面[68]：

1) 当数据描述的语言、标准不同时，表现出语法异构性；
2) 当对同一目标领域采用不同的建模或表述方法时，表现出语义异构性；
3) 当对同一事物采用不同的命名方式时，表现出术语异构性；
4) 当对同一事物采用不同的符号描述时，表现出符号异构性。

随着近几年建筑领域对能源管理意识的不断增强和相应的能源计量和管理法规的制定，有政府引导的城市级大型公建能耗管理平台陆续在城市分批次实施，一些大中型企业也开始搭建自己企业内部的能源管理平台，市场上如雨后春笋般出现了一批建筑能源管理系统，随着时间的推移积累了大量能耗数据和系统运营数据，但是每个建筑的能耗管理平台都有所差异，数据的种类、类型和存储方式、数据编码、标准和接口也不统一，无法保证数据的一致性和准确性，更无法进行综合、全面、深入的数据应用[69]。

为了解决数据异构的问题，自从 20 世纪中后期开始，人们就开始着手研究解决该问题的相关技术，根据集成研究的切入点和采取的信息模型来划分，异构数据融合方法可以分为结构化方法和语义方法两种。结构化方法主要用来解决数据结构的异构问题，目的在于统一信息结构，但不考虑信息的语义关联，该方法中的关键有效的方法是利用中间件—元数据（Metadata）对信息结构进行转化整合。所谓元数据，也被称为描述数据的数据，本质上是将异构数据进行连接的媒介。目前常用于存储和描述元数据的工具主要有关系型数据表、可标记扩展语言文本（Extensible makeup language，XML）、XML 结构定义（XML Schema Definition，XSD）等。语义方法针对的是信息的语义联系，从信息的局部语义模型描述上入手，通过对源数据的语义进行关联，从而完成对于全局信息的全面描述，进而完成以对于全局数据的查询和传输等功能，该方法中的关键有效的方法是使用本体和本体映射技术利用信息语义相似性进行计算，将相同语义的信息进行映射[70]。国外研究的结构化方法和语义方法的典型代表分别是：TSMMIS（斯坦福 IBM 多信息源管理）系统[71]和 MOMIS（多信息源中间件）系统[72]。

在工程应用领域，人们将不同信息源的数据进行整合，可以产生比单一信息源更加准确可靠的信息，降低不确定性。数据融合性及融合后数据不确定性评估的一般方案如下:

1) 对每个单独信息源相关的不确定性进行分类和估计；
2) 构建适当的融合方法，将不同信息源的输出组合起来；

3) 评估不同信息源的联合不确定性，并根据合并过程确定不确定性[73]。

目前常用的融合方法大多是基于统计理论，包括最小方差估计、极大似然估计[75]以及贝叶斯方法[76]等。其中，经典的数据融合算法—卡曼滤波器（Kalman Filter）属于贝叶斯方法，被广泛应用于多传感器数据的融合，因为传感器由于线路或网络等原因经常经常发生数据传输故障，单一传感器往往不可靠，因此需要多个传感器来降低不确定性。在暖通空调领域，数据融合的研究还比较稀少。Huang 等人把建筑冷负荷来自两个来源（直接测量：冷冻水流量和供回水温差，间接测量：冷机能耗和蒸发/冷凝温度）的测量值进行融合，得到了更为准确的冷负荷，并在此基础上对机组进行优化控制[77]。 Natasa Djuric 等人也用类似的方法对热泵的性能进行评估，其中直接测量为温度和压力测量值，间接测量为热泵的电力信号，结果表明通过数据融合方法得到的测量值比两个单一来源测量值更加可靠[78]。Huang 等人针对冷机负荷测量的两种数据融合方法进行了比较，第一种方法是基于模型的数据融合方法，另一种是基于多传感器的数据融合，结果表明当冷机负荷较温度且没有富余测量时，基于模型的融合方法效果更好，当存在富余测量时，基于多传感器的融合方法效果好[79]。

1.3 本课题的主要研究内容

1.3.1 课题的研究对象和概念说明

1.3.1.1 研究对象

本课题的题目为“建筑信息异构数据融合方法及混合能耗模型的建立”。主要解决新建建筑或是没有历史能耗数据记录的既有建筑进行能耗预测时存在的两个问题：

1) 利用能耗模拟软件进行能耗预测，准确度不高；

2) 不能基于自身的历史能耗数据建立黑箱模型进行预测。

本课题以公共建筑基本信息和空调系统实际运行能耗为研究对象，通过能耗模拟和数据分析相结合的方法从众多建筑基本信息中提取出对空调运行能耗造成重大影响的少数变量（本文称之为“关键变量”）；其次，关于如何得到某一建筑关键变量的取值，本课题提出了基于能耗模拟软件的模拟数据和实测能耗数据融合方法，实现两者的相互链接，采用模拟数据对实测实测数据进行修正，而后基于修正后的实测数据对未知关键变量的推测；接着，本课题提出了混合能耗模型的概念，并确定了模型结构（包括特征参数和算法），可以实现公共建筑实际空调运行能耗的快速预测。为了辅助混合能耗模型的开发，本课题建立了建筑

信息及空调能耗的异构数据库，该数据库同时包括了建筑信息关键变量与对应的空调系统实际运行能耗及模拟能耗作为补充，实现了不同来源数据、不同类型数据的结构化存储，可直接服务于混合能耗模型的训练和调试，数据库也可用于其他场景。另外，在实际工程实践中，很多建筑不能准确获得关键变量的值，为解决这个问题，本课题提出了空调能耗的非确定预测放法。最后本课题针对上述混合能耗模型进行了交叉测试，从准确度和稳定性两个方面验证模型的有效性。

本课题所述“建筑基本信息”包括建筑几何信息、围护结构热工信息、建筑及其机电系统使用信息、设备系统性能等所有可能对空调运行能耗造成影响的因素，具体参数详见第二章。本文所述“空调系统实际运行能耗”包括来自分项计量平台的逐时（或更细颗粒度）的能耗数据以及能源审计报告中的逐月用能统计值。

本课题以上海市星级酒店建筑为案例展示了数据融合方法及混合能耗模型的执行过程，其他类型公共建筑只需在酒店建筑的基础上对使用特征进行更改就可使用本课题提出的方法。

1.3.1.2 重要概念及说明

（1） 建筑信息异构数据

异构数据，简单来说是指数据集中包含不同的成分，可以是不同结构、不同格式、不同来源等。在本课题中，关于建筑信息的异构数据主要是指两个层面：

- 来源不同，本课题同时有能耗模拟数据和现场实测数据，实测数据又包含了来自分项计量平台和能源审计报告的数据，两者颗粒度不同；
- 结构不同，异构数据库中包含了建筑基本信息和能耗序列值，前者是静态数据，后者是时间序列值。

（2） 混合能耗模型

一般混合能耗模型是指采用不同建模方法（例如同时采用白箱和黑箱模型）或是黑箱模型中采用了多种算法结合（例如基于装袋法的随机森林、基于提升法的 GBDT，LightGBM 等）而成的模型，本课题所指混合能耗模型稍有不同，其采用的建模方法为数据驱动模型，但数据来源是上一条所述异构数据。

（3） 建筑信息关键变量

本课题所述“建筑信息”包括建筑围护结构本身、室内人员以及机电系统中，所有可能对空调运行能耗造成影响的因素，“关键变量”是指建筑信息中，对空调运行能耗造成显著影响的少数因素。

（4） 理想变量、附加变量

理想变量和附加变量是对建筑信息的分类，两者之间并没有特别严格的区分。

理想变量是指常规建模时需要考虑的参数，即其反应的是建筑及其设备系统均处于理想运行状态的情况，附加变量则是考虑了系统出现偏离理想状态的情况，例如盘管结垢、送回水温差过小等。

（5） 代理模型

本课题所述“代理模型”是针对实际建筑的简化模型，相对于实际建筑，代理模型的外形特征和分区设置大大简化，当代理模型的建筑信息关键变量与实际建筑一致时两者的空调能耗接近。代理模型产生的模拟数据被用于在数据融合中对实测建筑能耗进行修正。

1.3.2 课题的主要工作

根据国内外研究总结可以发现，通过能耗模拟工具计算建筑能耗常为人诟病其准确性不高；黑箱模型虽然具有较高的准确性，但只能用于有历史数据的建筑，对于新建建筑或是没有历史能耗记录的建筑不能使用。但另一方面，能耗模拟方法和实测能耗数据各自有其独特的优点。两者虽然是研究同一对象，但其研究路径基本没有交集。本课题希望打破这一局面，将能耗模拟方法和实测数据进行结合，提供建筑空调能耗预测的新思路。本课题最终采用数据驱动的方法来构建预测模型，因此需要明确构建数据驱动模型的训练数据，即其输入输出是什么。输入即影响建筑空调系统运行能耗的变量，输出即建筑空调运行能耗。为了实现这一目的，需要解决以下几个问题：

1) 问题一：对建筑空调系统运行能耗可能产生影响的因素很多，但其中最显著的影响因素（即关键变量）是哪些？
2) 问题二：目前来自分项计量平台的数据质量参差不齐，直接放入模型中进行训练会严重影响模型精度，那么该如何提高实测数据质量？建筑能耗分项计量平台仅对能耗数据（即数据驱动模型的输出变量）做采集，如何得到与之对应的建筑基本信息（即数据驱动模型的输入变量）？
3) 问题三：针对无历史能耗数据的建筑如何对其进行能耗预测？
4) 问题四：建立了数据驱动能耗预测模型后，在进行某一建筑的能耗预测时需要知道确切的输入变量取值，但某些情况下无法获取，如何解决存在未知特征时的能耗预测？

因此本文的研究工作将围绕以上问题展开，分为以下几个主要部分：

1) 提取建筑空调系统运行能耗关键变量。本文首先筛选了几十个影响空调能耗的初始变量，通过敏感性分析方法从初始变量集中提取若干关键变量；为了实现敏感性分析所需的数据量，本文开发了自动建模及参数分析工具，可以批量生成能耗模型，用于计算变量集在不同取值情况下的能耗，并根

据计算值进行敏感性分析。这里需要强调的是，由于气象参数是公认的影响空调能耗的关键因素，因此本文没有将气象参数纳入关键变量选取的工作范畴，而是直接将其作为关键变量用于混合能耗模型的建立。

2) 进行能耗模拟数据和实测数据的融合，可以同时实现实测数据的修正和建筑信息的填补。该数据融合方法首先利用模拟数据对实测数据进行异常值和噪声修正，然后基于贝叶斯理论，根据修正后的实测数据对确实的建筑信息进行推测，得到其期望值作为最终的推测值。本文通过虚拟建筑和实际建筑两个案例，从两个角度验证了数据融合算法的可行性和有效性。

3) 将建筑信息和空调运行能耗进行组合，搭建建筑信息异构数据库。该数据库中包含两部分：实际数据组和模拟数据组。实际数据组的每一组数据集对应一个实际建筑，数据包含通过上述数据融合方法得到的建筑信息关键变量和空调运行能耗值。模拟数据组包含了能耗模拟软件计算得到的大量模拟数据。基于上述建筑信息异构数据库建立混合能耗模型，实现空调系统运行能耗的预测。

4) 针对存在未知特征变量的场景，提出了基于关联规则算法的非确定能耗预测方法。

1.3.3 课题的技术路线和文章架构

本文共分为六章：

第一章介绍了本课题的研究背景，对目前国内外相关研究进行了总结分析，在此基础上提炼出目前该领域存在的问题和本文的研究内容。

第二章提出了影响建筑空调系统运行能耗关键变量的确定方法。首先选取所有可能对空调能耗产生影响的变量，对变量进行抽样并开发自动建模及参数分析工具批量生成算例，运用敏感性分析方法从中提取若干关键变量，最后通过案例验证了本章所提关键变量的有效性。

第三章提出了基于模拟能耗数据和实测能耗数据的融合方法，该方法通过模拟数据对实测数据进行修正，然后利用修正后的实测数据对关键未知变量进行推测。

第四章通过两个案例（分别是基于模拟和基于实测建筑能耗）展验证了数据融合方法的实践步骤及其有效性。

第五章提出了混合能耗模型的构建方法，通过机器学习方法描述建筑基本信息和空调运行能耗之间的映射关系，建立空调运行能耗预测模型，并搭建了建筑信息异构数据库。针对存在未知特征的情况，提出了非确定能耗预测方法。

第六章针对第五章提出的模型和算法进行了验证，采用的验证数据集包括了

上海地区的来自分项计量平台和审计报告的空调能耗数据，以及美国 CBECS 数据库的酒店空调能耗数据。

第七章总结了本研究的全部工作、创新点和局限性，并对后续工作进行了展望。

本课题的技术路线图见图 1.6。

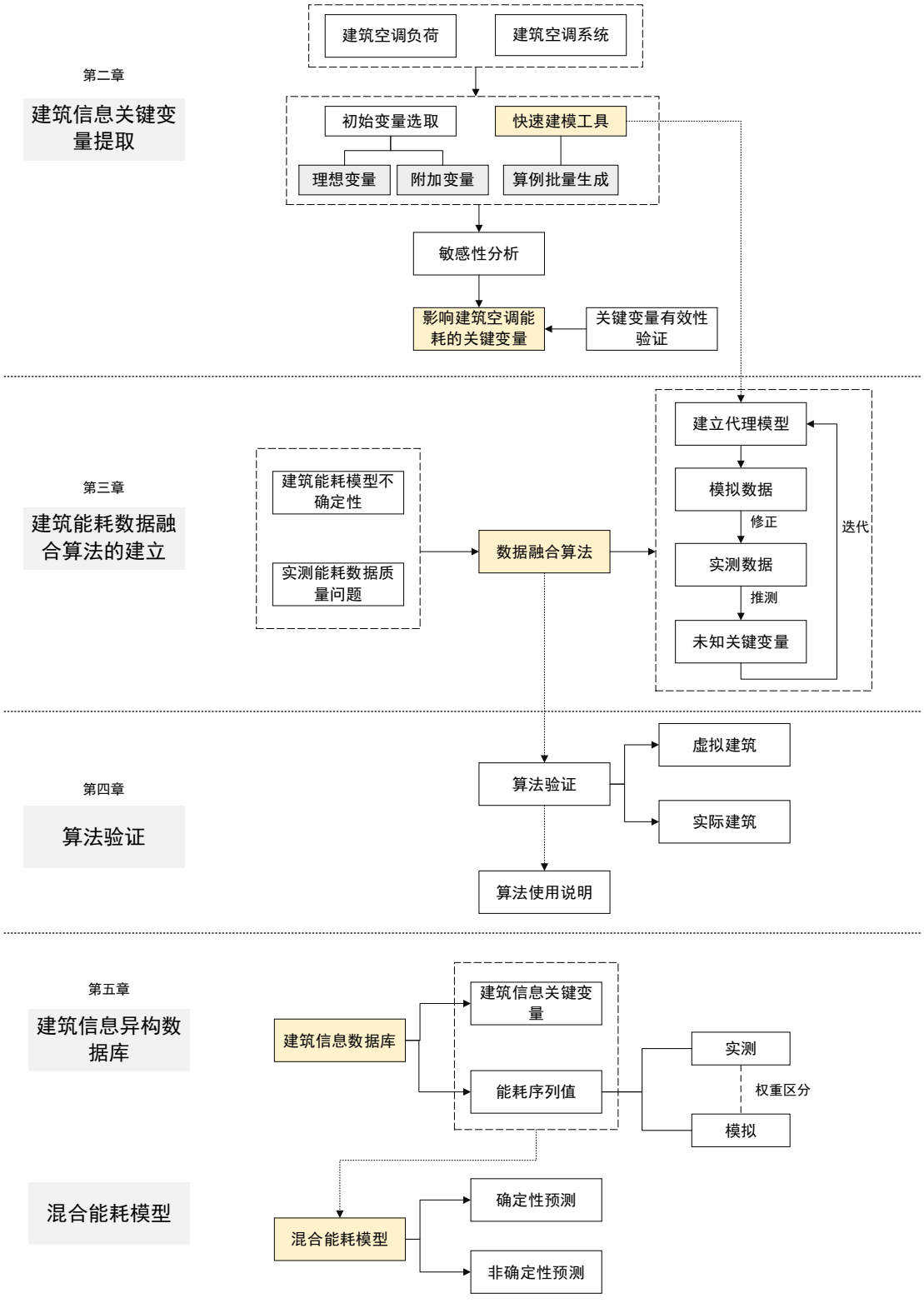

第二章
建筑信息关键变量提取
建筑空调负荷
建筑空调系统
初始变量选取
快速建模工具
理想变量
附加变量
算例批量生成
敏感性分析
影响建筑空调能耗的关键变量
关键变量有效性验证
第三章
建筑能耗数据融合算法的建立
建筑能耗模型不确定性
实测能耗数据质量问题
数据融合算法
建立代理模型
模拟数据
修正
实测数据
推测
未知关键变量
迭代
第四章
算法验证
算法验证
虚拟建筑
实际建筑
算法使用说明
第五章
建筑信息异构数据库
建筑信息数据库
建筑信息关键变量
能耗序列值
实测
权重区分
模拟
混合能耗模型
混合能耗模型
确定性预测
非确定性预测

图 1.7 课题技术路线图

1.4 本章小结

本章对目前建筑能耗模拟预测方法的应用情况及其缺点进行了介绍；另一方面，来自建筑分项计量平台的大数据并未得到有效利用。在此基础上，本课题提出了将模拟与实测数据相结合的方法构建建筑能耗混合模型，实现建筑空调系统运行能耗的确定性预测和非确定性预测方法。围绕上述要点，本课题指出了目前该领域存在的四个主要问题，在此基础上提炼出了本文的主要内容和技术路线。

第2章 建筑空调负荷及运行能耗关键变量的提取

2.1 概述

建筑空调运行能耗是由空调负荷和设备系统特征共同决定的。建筑空调负荷是进行系统设计、优化配置、高效运行等一系列后续工作和研究的前提。只有在准确计算空调负荷的前提下，才能保证其他工作的顺利开展。计算空调负荷有估算和详细计算两种，估算的依据是不同气候区、不同建筑类型的单位面积负荷指标，是长期的工程经验积累得来的。详细计算则根据建筑得热的机理，将负荷的形成用严谨的数学公式进行表达和计算。以空调冷负荷为例，包括了围护结构冷负荷、人体显热负荷和湿负荷、照明冷负荷、设备冷负荷等。其中围护结构冷负荷与建筑的外形结构、墙体窗户屋面等外围护结构的热工构造以及室外气象参数密切相关，因此详细计算建筑空调负荷时十分繁琐的过程。随着计算机科学的不断发展，利用计算机软件进行负荷计算越来越普及，其中比较常用的有美国能源部资助开发的 EnergyPlus，清华大学开发的 DeST 等。

空调系统是为了移除室内的冷热负荷而设计的，它向室内空间提供冷量或者热量以将室内维持在舒适的温湿度范围内。经过几十年的发展，空调系统的形式和种类越来越多，可以说没有任何两栋建筑的空调系统时一模一样的。但总体来说，公共建筑所采用的空调系统包括冷热源、输配系统、末端设备这三大部分。空调系统各设备的运行能耗的计算涉及到与负荷耦合迭代的过程，只能依靠计算机软件进行。

虽然计算机软件的使用降低了空调负荷和系统能耗计算的复杂程度，但还是存在以下几个问题：

- 输入参数太多，收集困难；
- 需要将建筑几何外形还原，耗时耗力；
- 没有区分各个参数的取值对计算精度的重要性。

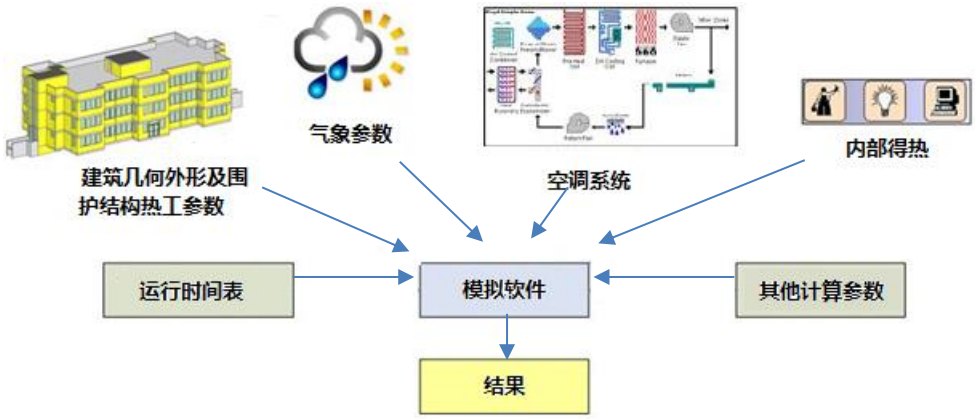


图 2.1 建筑空调运行能耗构成示意图

本章节的主要研究内容就是利用敏感性分析方法（图 2.2），从繁多的计算输入参数中选取少量的重要参数，也就是建筑空调负荷及运行能耗关键变量，这些变量保留了能解释空调能耗变化和差异的大部分信息，为后续的混合能耗模型建立做准备。本章节将分别针对建筑负荷和空调系统进行敏感性分析，这样操作的目的主要基于两点以下考虑：

1) 建筑负荷和系统运行能耗所涉及的变量太多，同时包括数值型和水平型，如果进行联合采样将造成样本量过大，难以实现。
2) 建筑负荷和系统运行能耗的计算可以单独进行，因此可以考虑分别分析空调负荷和运行能耗各自的关键变量。

2.2 敏感性分析

从数学角度讲，敏感性分析（Sensitivity analysis）是研究如何将模型输出变量的不确定度（可以是定量或其他形式）分配到各个输入变量上[108]。可以用来定量或定性分析每个输入变量的变化对于输出结果的影响大小，有助于了解并简化模型结构、提高模型精度并增加模型应用的可靠性。该方法在经济学、社会学、工程技术等领域都有十分广泛的应用。敏感性分析包括三大要素：模型、输入变量和输出变量。这里的“模型”是我们对真实世界中各种现象的假设和重构，在本课题研究的范畴内，“模型”即是人们对建筑冷热负荷的形成这一现象的重构。模型的建立可以是基于研究对象的内在规律，以数学和物理方程呈现，也可以是基于观察统计数据，用统计模型表达，本章节采用的模型属于第一种方法，借助 EnergyPlus 软件进行建筑负荷的敏感性分析。

敏感性分析可分为局部敏感性分析和全局敏感性分析。局部敏感性分析是固

定目标输入参数以外的其他输入参数，研究单个输入参数在局部范围内变化时对模型输出的影响。常见的局部敏感性分析方法是逐个变量法(One-variable-at-a-time approach, OAT)和微分分析法(Differential analysis, DA)。局部敏感性分析方法计算量小、操作简单，但是无法反应描述输入参数的联合空间分布形态，并且忽略了参数之间的相互作用。而全局敏感性分析研究的是所有模型输入变量同时变化时对模型输出的影响。常用的全局敏感性分析方法包括回归分析法、Morris 法、Sobol 法、FAST 法和 RSA 法等。本章节采用全局敏感性分析方法，下文全局敏感性分析方法的实现过程作简要阐述。

2.2.1 全局敏感性的一般执行过程

全局敏感性分析的操作方法有多种，最常见的方法是基于抽样的敏感性分析。基于抽样的敏感性分析是指对根据输入变量的分布(假设已知)进行抽样，然后根据抽样得到的输入变量组合值重复执行模型计算，得到对应的输出变量值。一般情况下，基于采样的敏感性分析执行步骤如下：

1) 确定输入变量和研究对象，建立可重复执行的模型。如前所述，该模型可以是基于物理机组的模型，也可以是基于观察数据的统计模型，但需要满足的条件是便于多次重复计算，因为后续过程需要基于抽样的输入变量组合进行大量的模型计算。
2) 为每个输入变量确定其取值范围和概率分布。这是比较关键的一步，如果变量的取值范围有误，分析结果也是不可靠的，输入变量的取值范围一般根据经验或文献资料。在缺少信息的情况下，概率分布一般可以用平均分布。
3) 在步骤 2 的基础上进行采样，得到一个输入变量矩阵。常用的抽样方法有随机采样、蒙特卡洛（Monto Carlo）采样、拉丁超立方（Latin Hypercube）抽样等，还有与后续采用的敏感性分析方法对应的特定抽样方法，如 Morris 法、FAST 法等。
4) 根据步骤 3 得到的输入变量样本矩阵代入模型，进行计算，得到对应的输出变量值。
5) 选择一种方法分析各个输入变量对输出变量的影响程度。这里可用的方法很多，具体介绍见下一小节。

2.2.2 全局敏感性分析方法

本文采用了 Morris 方法和回归分析法两种敏感性分析方法，因此本小节主要就这两种方法做介绍。

（1） Morris 分析法[51]

Morris 法又称为元效应方法，是一种简单而有效的方法，可以从模型中包含的许多输入因素中筛选出一些重要的输入因素。一个元效应的定义如下：

假设一个模型有k个独立的输入X_i，$i = 1, \dots, k$，这些独立变量在一个k维的单位体中进行变化，变化水平为p，也就是说输入变量空间被离散为p级的格子空间Ω。对于一个给定的输入变量X，其中第i维的元效应定义为：

$$EE_i = \frac{[Y(X_1,X_2,\dots,X_{i-1},X_i+\Delta,\dots,X_k)-Y(X_1,X_2,\dots,X_k)]}{\Delta} \quad (2\text{-}1)$$

其中，Δ的值取自于$\{\frac{1}{p-1}, \dots, 1-\frac{1}{p-1}\}$，$\mathrm{X} + e_i\Delta$仍处于$\Omega$中，$e_i$是在第$i$维上是1，其余维度为 0 的向量。

第i维输入变量的元效应分布是通过在Ω空间针对X随机抽样得到的，用F_i表示，即$EE_i \sim F_i$。又 Morris 提出的两个敏感性指标，μ和σ，就是F_i分布的均值和标准差。μ是该变量对结果整体影响的估计，而σ是对变量综合影响的估计，σ越大说明该变量对输出的影响受其他变量取值的影响大。由于在F_i包含正负值时，一些效应在计算μ值时会产生正负抵消，从而低估某一重要因素的敏感性，因此目前常用 Campolongo 提出的μ^*来代替μ，μ^*是元效应绝对值分布的均值，这个元效应分布定义为G_i，$|EE_i| \sim G_i$。

（2） 回归分析法

回归分析是另一种常用的敏感性分析方法，通过某种抽样策略生成输入变量X的多元样本，并利用模型计算相应的输出值。若模型是线性的，可用公式 2-2 表达[51]：

$$y_i = b_0 + \sum_j b_j x_{ij} + \varepsilon_i \quad (2\text{-}2)$$

其中，y_i表示模型输出，$i = 1, \dots, m$，b_j表示需要确定的模型参数，ε_i是误差。一个确定b_j的常用方法是最小二乘法，即通过使得$\sum_i \varepsilon_i^2$最小来求取b_j。于是，上述回归方程可以重写为式 2-3，式中的$b_j s_j/\hat{s}$被称为标准回归系数（SRC），可被用来衡量某一参数的重要性，用作敏感性指标。

$$(y - \bar{y})/\hat{s} = \sum_j (b_j s_j/\hat{s})(x_j - \bar{x_j})/\hat{s}_j \quad (2\text{-}3)$$

$$\bar{y} = \sum_i y_i/m \qquad \bar{x_j} = \sum_i x_{ij}/m \quad (2\text{-}4)$$

$$\hat{s} = [\sum_i (y_i - \bar{y})^2/(m-1)]^{1/2} \quad (2\text{-}5)$$

$$\hat{s}_j = [\sum_i \left(x_{ij} - \bar{x_j}\right)^2/(m-1)]^{1/2} \quad (2\text{-}6)$$

另一个敏感性指标是偏回归系数（PCC），是基于相关性和偏相关性的概念提出的。对于系列观测对象（x_{ij}, y_i），第i维输入变量X_i和输出变量Y之间的相关系数为：

$$r_{xiy}=\frac{\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)(y_i-\bar{y})}{[\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)^2]^{1/2}[\sum_{i=1}^{m}(y_i-\bar{y})^2]^{1/2}} \quad (2\text{-}7)$$

Y和X_i之间的偏相关系数通过定义两个回归模型来构造：

$$\hat{Y}=b_0+\sum_{h\neq j}b_h x_h \quad (2\text{-}8)$$

$$\hat{X}_j=c_0+\sum_{h\neq j}c_h x_h \quad (2\text{-}9)$$

进而再构造两个变量$Y-\hat{Y}$和$X_j-\hat{X}_j$，这两个变量之间的相关系数即为Y和X_j的偏相关系数。

需要注意的是，上述两个敏感性指标（SRC 和 PCC）只适用于分析线性模型，在分析非线性模型问题时需要进行秩变换，也就是将上述公式中的数据全部用相应的秩来替换。经过秩变换的敏感性指标被称为标准秩回归系数（SRRC）和偏秩回归系数（PRCC）[56]。

秩变换的操作过程是将N个观测值按升序排列，依次赋予秩序 1~N，对于相等的若干个观测值，需将其初始秩序相加求均值作为新的秩。例如，对于观测值序列（20.4，5.3，12.6，7，2，2），首先将其按升序排列（2，2，5.3，7，12.6，20.4），将其赋予秩序为（1，2，3，4，5，6），由于观测值中前两个值相同，因此其秩需求和取平均，即为（1.5，1.5，3，4，5，6），最后将其顺序调整至初始状态，所以最终的秩变换结果为（6，3，5，4，1.5，1.5）。

2.2.3 抽样方法

抽样是指根据变量分布特征从中选取可以反映数据集整体特征的有限个数据点，敏感性分析展现出来的信息非常依赖于计算模型所采用的样本点数量和位置，适合的抽样方法可以在保证精度的情况下减少数据点，从而减少计算时间，提高效率。常见的数据抽样方法包括以下几种：

（1） 正交试验（Orthogonal experimental design），正交试验根据正交性从全面试验中挑选出部分有代表性的点进行试验，这些有代表性的点具备了“均匀分散，齐整可比”的特点，是最为常用的采样方法。

（2） 蒙特卡洛抽样（Monte Carlo sampling，MC），蒙特卡洛抽样是一种随机模拟方法，其基本思想是用事件发生的“频率”来近似代替“概率”。首先用随机数生成器来构造事件的概率模型，然后从中进行抽样

（3） 拉丁超立方抽样（Latin hypercube sampling，LHS），拉丁超立方体抽样是一种分层抽样方法，可以从多维分布中生成近似随机样本点。相较于随机抽样方法更高效，它改进了采样策略能够做到以较小的采样规模获得较高的采样精度，在实际应用中很受欢迎。拉丁超立方抽样将k样本空间的每一维分割成n等分，样本空间即被分割成kn个小块，然后在每一个小块中

进行随机抽样，这种方法可以保证样本点分布相对均匀，防止出现伪随机的现象。

上述几种是比较通用的抽样方法，其抽样所得样本可用于计算回归系数（SRRC 和 PRCC）。但 Morris 分析法有其特有的抽样方法，其样本也不能用于其他敏感性分析方法。

在上一小节中提到，Morris 法的敏感性指标μ^*和σ是元效应G_i分布的均值和F_i分布的标准差。为了估计出F_i和G_i的分布，需要从每个F_i和G_i中抽取若干个样本点。假设需要抽取r个样本点，即需要计算r个元效应，由于每个元效应需要两个数据点，那么每一维输入则需要$2r$个数据点，k维输入空间需要$2rk$个数据点。Morris 提出了一种更有效的抽样方案，即在输入空间中建立r条由$k+1$个点构成的轨迹线，每一条轨迹线可以满足计算k个元效应的需求，每个维度一个，那么总的样本点可以减少为$r(k+1)$。

图 2.2 展示了三维空间中一条轨迹线的构造过程，首先在样本空间Ω中随机选取一点x^*，以此为基准点构造所有轨迹数据点，但x^*不属于轨迹线。第一个轨迹点$x^{(1)}$是在x^*的某一个或多个维度上增加Δ，$x^{(1)}$仍处于Ω中。第二个轨迹点$x^{(2)}$通过$x^{(1)}+e_i\Delta$或$x^{(1)}-e_i\Delta$得到，它与$x^{(1)}$尽在第i个维度上有差别。同样的，第三个轨迹点由$x^{(2)}+e_j\Delta$或$x^{(2)}-e_j\Delta$得到，其中$i\neq j$，以此类推直到完成构造一条$k+1$个点的轨迹线[51]。

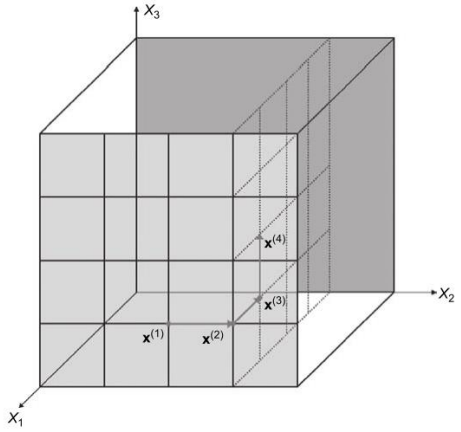


图 2.2 轨迹线构造示意图（$k=3$）

2.3 建筑空调负荷关键变量的提取

本小节以酒店建筑夏季冷负荷为研究对象，针对建筑空调负荷关键变量的选取做阐述，采用的方法是全局敏感性分析法，其流程图如图 2.3 所示，具体每一个步骤的实现过程和方法见后续小节。

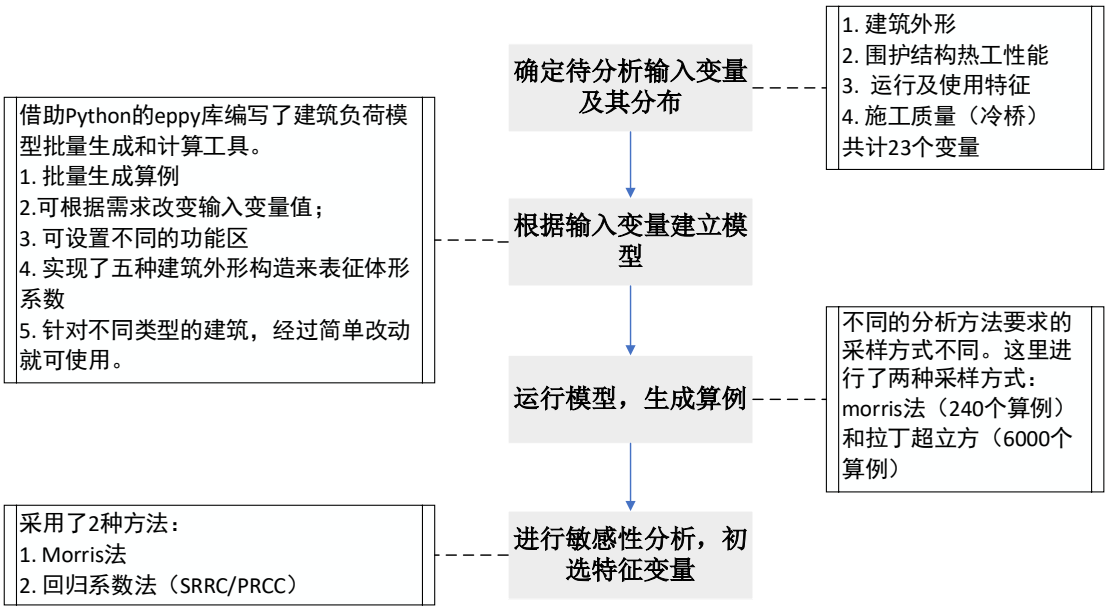


图 2.3 建筑空调负荷敏感性分析流程图

2.3.1 影响建筑空调负荷初始变量集的选取

对建筑空调负荷产生影响的因素众多，表 3 中罗列了 23 个变量，分为四类：建筑外形、围护结构热工性能、运行使用以及施工质量。每个变量的取值范围基于经验确定，在其取值范围内服从均匀分布。这四类变量中，前三类变量对于建筑负荷分析是比较常见的，这里不做展开分析，下面对第四类变量“施工质量”做一些补充说明。

本小节所述的施工质量主要指建筑围护结构保温不佳造成的“冷桥效应”，常见于建筑外墙转角、内外墙交角、楼屋面与外墙搭接角处、墙体与窗户接缝处，此处的传热系数相较于其他部位偏大，产生热量积聚。这里引入三个变量：楼板线性透过率、墙角线性透过率、玻璃线性透过率来描述上述容易产生冷桥效应处的热量传递，然后将这些部位额外产生的热量附加到整面墙体上，从总热量传递的角度考虑，相当于墙体的平均传热系数有所增加。因此在模拟计算时，采用公式 2-1 将上述线性透过率折算成墙体传热系数的增加[116]。

$$U_T = \frac{\sum(\psi \cdot L)}{A_{tot}} + U_0 \qquad (2\text{-}2)$$

其中，U_T表示考虑冷桥效应后的墙体综合传热系数，$\mathrm{W}/(m^2K)$；ψ表示线性透过率，$\mathrm{W}/(m\,K)$；L表示不同线性透过率所对应的长度，m，分别是楼板和墙面、玻璃与墙面、不同墙面之间的接触面长度；A_{tot}表示墙面面积，m^2；U_0表示墙面本身的传热系数，$\mathrm{W}/(m^2K)$。表 3 中初始变量的取值以夏热冬冷地区公共

建筑节能设计标准规定的值为中位数进行前后延伸得到。

表 3 建筑空调负荷敏感性分析初始变量集

种类	参数名称	缩写	范围取值	单位	说明
建筑外形	窗墙比（北）	NWWR	0.1~0.9		
	窗墙比（南）	SWWR	0.1~0.9		
	窗墙比（东）	EWWR	0.1~0.9		
	窗墙比（西）	WWWR	0.1~0.9		
	建筑面积	AREA	20000~80000	m^2	
	层数	NL	5~40		
	体形系数	CR	0.1~0.9		
围护结构热工性能	外墙传热系数	WALLU	0.09~11.1	$\mathrm{W}/(m^2K)$	
	外墙热容	WSP	800~2000	$\mathrm{J}/(kg\ K)$	
	屋顶传热系数	RU	0.09~4.8	$\mathrm{W}/(m^2K)$	
	窗玻璃传热系数	WINU	0.2~9	$\mathrm{W}/(m^2K)$	
	窗玻璃太阳辐射得热系数	SHGC	0.1~0.9		
	外墙太阳辐射吸收系数	WSA	0.1~0.9		
	屋顶太阳辐射吸收系数	RSA	0.1~0.9		
使用运行	空调制冷设定温度	SPC	22~28	℃	
	空调供热设定温度	SPH	18~24	℃	
	照明功率密度	LPD	3~15	W/m^2	
	人员密度	OPD	0.1~1	P/m^2	
	冷风渗透率	INFIL	0.5~5	ACH	
	内遮阳开启程度	ST	0.1~0.9		简化为内遮阳的透光率
施工质量	楼板线性透过率[116]	FLT	0.007~1.842	$\mathrm{W}/(m\ K)$	折算成墙体传热系数的增加
	玻璃线性透过率[116]	GLT	0.03~1.058	$\mathrm{W}/(m\ K)$	
	墙角线性透过率[116]	CLT	0.036~0.684	$\mathrm{W}/(m\ K)$	

2.3.2 建筑空调负荷计算模型算例生成

所谓算例生成，是构造样本点和输出结果的映射关系，通常用构建数学模型的方法完成，是敏感性分析的过程中的关键步骤。由于样本点数量可能成千上万，因此算例的构造和计算过程也是比较耗时的。本课题的研究对象是建筑空调负荷，

手动计算空调负荷繁琐且精确度不高，因此这里借助 EnergyPlus 来计算空调负荷，并基于 python 语言和 eppy 包开发了自动建模工具，方便批量建立和计算负荷模型。

eppy 是一种脚本语言，用于编辑 EnergyPlus idf 文件和 EnergyPlus 输出文件。eppy 是用编程语言 Python 编写的。因此，它充分利用了 python 中丰富的数据结构和习惯用法。用户可以使用 eppy 以编程方式导航、搜索和修改 EnergyPlus idf 文件[109]。它可以实现以下功能：

- 用几行 eppy 代码对 idf 文件进行大量更改；
- 对 idf 文件进行过滤式修改；
- 同时对多个 idf 文件进行修改；
- 从模拟结果中读取数据；
- 根据一个模拟结果生成另一个 idf 计算文件。

算例的批量生成过程如图 2.4 所示，首先通过抽样获取若干个计算样本点，根据样本点中描述几何外形的参数构建模型的几何部分，然后与 idf 基准文件相结合，根据样本点取值修改 idf 中相应的参数，生成 idf 文件，计算并保存结果，依次类推，直到完成所有样本点的计算。这里的基准 idf 文件保存着与样本取值无关的参数，包括模型计算的基本设置、不同功能区的使用时间表。

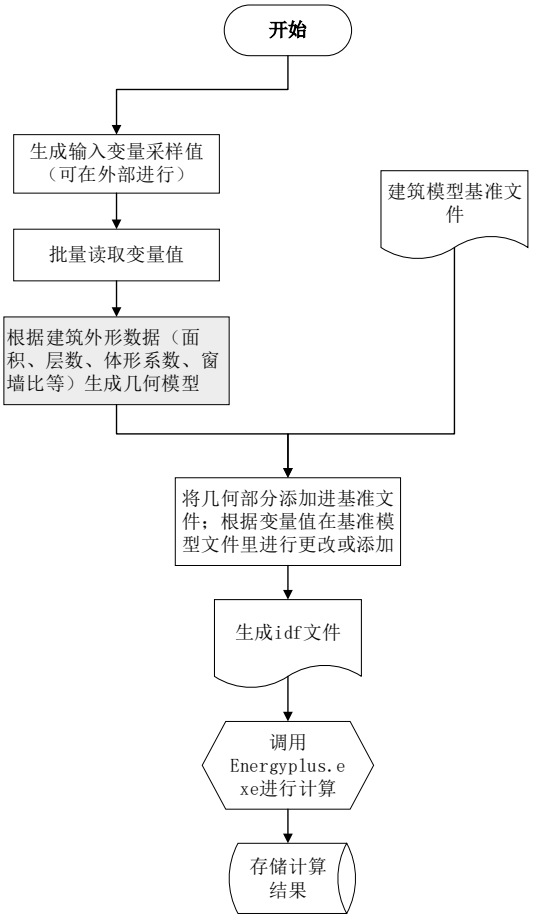


图 2.4 建筑空调负荷计算模型算例批量生成流程图

在以往的建筑负荷或能耗敏感型分析中，也需要进行算例的批量生成，但通常是针对同一个建筑对象，保持其模型外形和结构不变，只是改变部分参数值，因此不适用于涉及到针对建筑外形结构的分析，例如目前比较常用的参数分析软件 jEplus 软件。众所周知，体形系数一直被认为是影响建筑空调负荷的一个重要因素，要分析体形系数对于空调负荷的影响，需要比较不同体形系数的建筑能耗，但要根据需要的体形系数值去构建相应的模型外形是比较困难的，特别是在建筑面积、层数等其他参数确定的情况下。相较于以往的模型批量生成方法，本课题开发的自动建模工具更加灵活，可以根据既定的建筑面积、层数和体形系数取值来匹配合适的外形结构。目前该建模工具支持 5 种外形结构供选择，如图 2.5，从外形 a 到 e，建筑构造从紧凑型逐渐演变成松散型，工具能根据抽样得到的参数组自动匹配最合适的建筑外形，建立对应的几何模型，改变参数值，执行批量建模和计算。外形匹配算法设计思路为：

选取若干（本课题确定了 5 种方案）面积相同但周长不同的建筑平面形状，如图 2.6 所示，这 5 种外形的面积相同，周长按从小到大的顺序排列。在面积相同的情况下，周长越大表明与外界接触的面积越大，当建筑层数相同时，体型系

数越大。

构造系数 sigma 来表征平面形状特征，sigma 的计算公式为：

$$\text{sigma} = \text{C}/(\frac{\text{A}}{16}) \tag{2-3}$$

建立体型系数与 *sigma* 的函数关系：

$$\text{CR} = \text{f}(\text{sigma}, \text{A}_{\text{total}}, \text{NL}) \tag{2-4}$$

式中，C表示建筑平面周长（占地周长），A为建筑平面面积（占地面积），CR表示体型系数，A_{total}为总建筑面积，NL表示层数。

在进行计算是，模拟工具接收的参数是A_{total}，NL，CR，于是可以计算出对应的sigma。最后根据目标sigma寻找最匹配的建筑外形，实现几何模型的建立。

需要说明的是，课题的研究对象是酒店建筑，但工具是通用的，通过简单改动就能适合于不同的建筑类型，并可以通过设置不同功能分区的面积比例来实现功能区的划分。

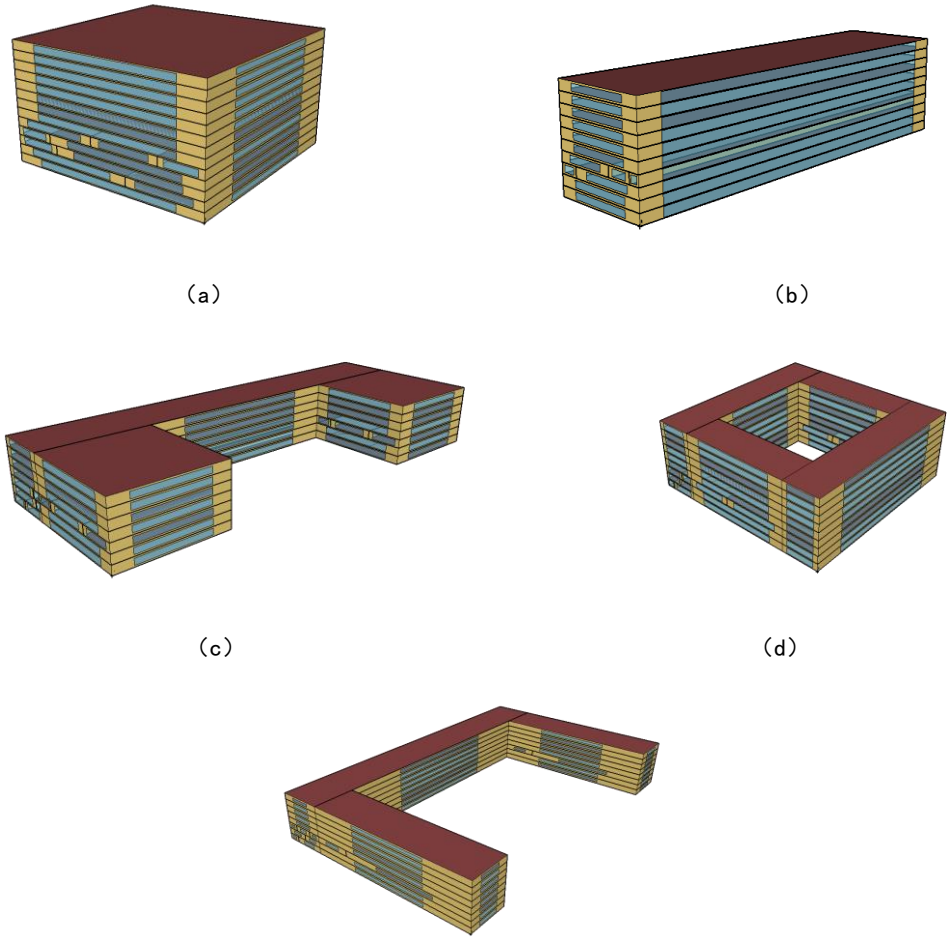


图 2.5 用于适配体形系数的 5 种外形结构

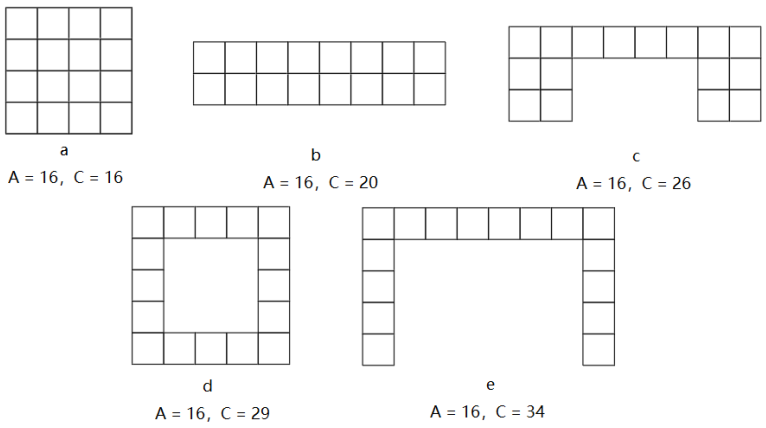


图 2.6 外形匹配算法平面示意图

2.3.3 建筑空调负荷敏感型分析及关键变量展示

敏感性分析的结果对初始条件很敏感。对于同样的输入参数，当目标输出参数发生变化时，敏感性分析结果可能会发生变化。因此，必须指定研究目标，在本节中，我们将冷负荷作为目标分析参数。分析模型对象的人员在室率、照明设备、空调启停等时间表见附录 A。

本章节采用 Morris 方法和回归分析法进行敏感性分析，Morris 方法对 23 个输入参数在其范围内采样，得到 240 个采样点，快速模拟工具执行了 240 次模拟，得到每个采样点对应的模拟冷负荷，Morris 方法的敏感性分析结果如图 2.7 所示。图中某一变量对应的数值的绝对值越大（即图中柱越长），说明该变量对空调负荷变化越敏感。

回归方法采用拉丁超立方体采样方法对 23 个输入参数进行采样，生成 6000 个样本。由于建筑热系统是高度非线性的，因此计算 SRRC 和 PRCC 作为敏感性指标，结果如图 2.8 所示，各参数指标的绝对值表示其重要性。两个回归 SRRC 和 PRCC 提供了相同的结果。

不难发现，两种方法的分析结果呈现出高度的一致性，对夏热冬冷地区酒店类建筑夏季冷负荷产生重要影响的关键变量为夏季空调设定温度（SPC）、人员密度（OPD）、新风渗透率（INFIL）、照明功率密度（LPD）。但是对于体形系数（CR）这个指标，两种方法给出了不同结果，这里对两种方法的结果取并集，取夏季空调设定温度（SPC）、人员密度（OPD）、新风渗透率（INFIL）、照明功率密度（LPD）、体形系数（CR）作为影响空调负荷的关键变量。需要说明的是，关键变量的数量是可以根据实际情况调整的，选取的量越多越能精确表征负荷的变化，但是关键维度太高也会造成后续关键变量推测的不确定性和计算量大大增

加，因此需要根据实际情况进行权衡判断。

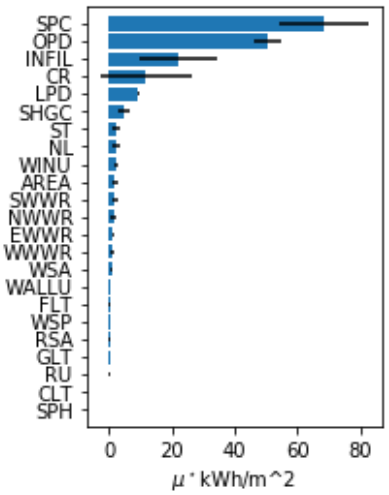


图 2.7 各变量的μ^*指标量化排列

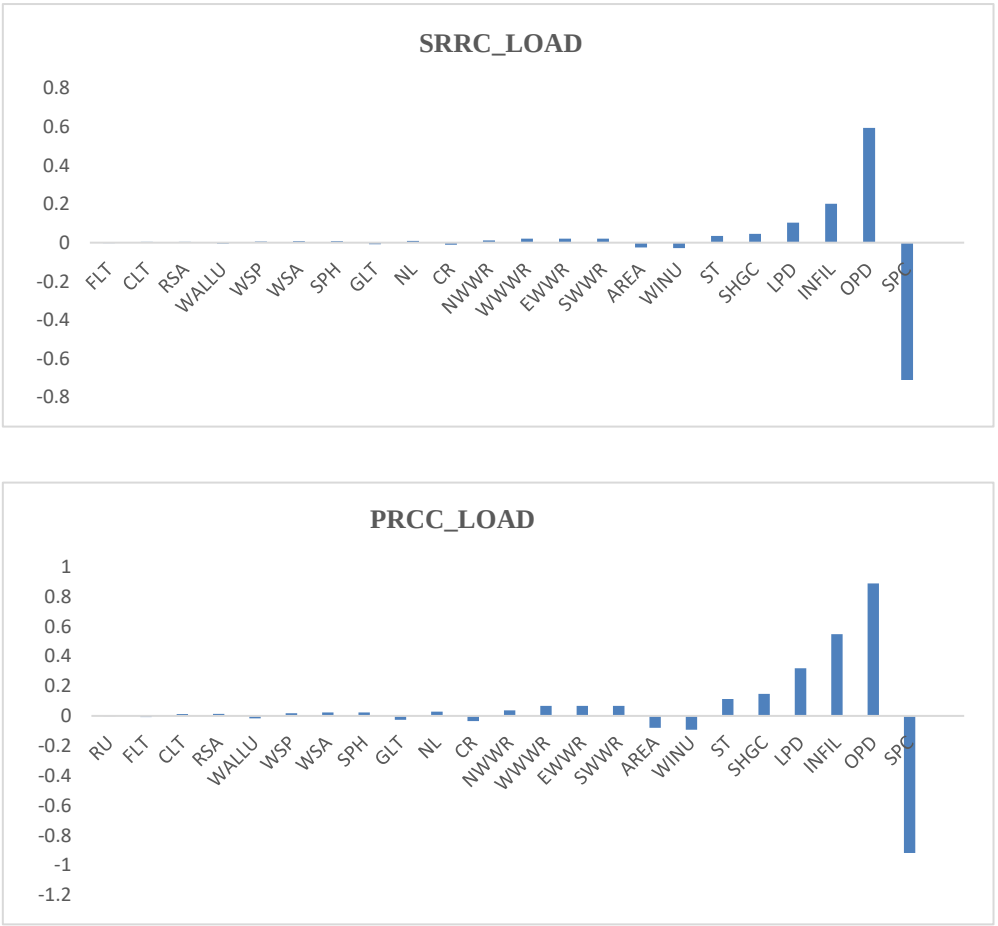


图 2.8 秩回归指标 SRRC 和 PRCC 量化排列

2.3.4 建筑空调负荷关键变量有效性验证

为了更进一步探讨各个变量对于冷负荷的影响程度，下面从预测拟合的角度给出更加直观的展示。具体来讲，是用机器学习的方法建立输入变量与输出（空调负荷）之间的映射关系，并对冷负荷进行预测。由于各个变量对冷负荷的影响程度不同，因此采用不同的变量组合作为输入参数将会得到不同的预测精度。若采用关键变量作为模型的输入特征，模型应该取得较高的预测精度。

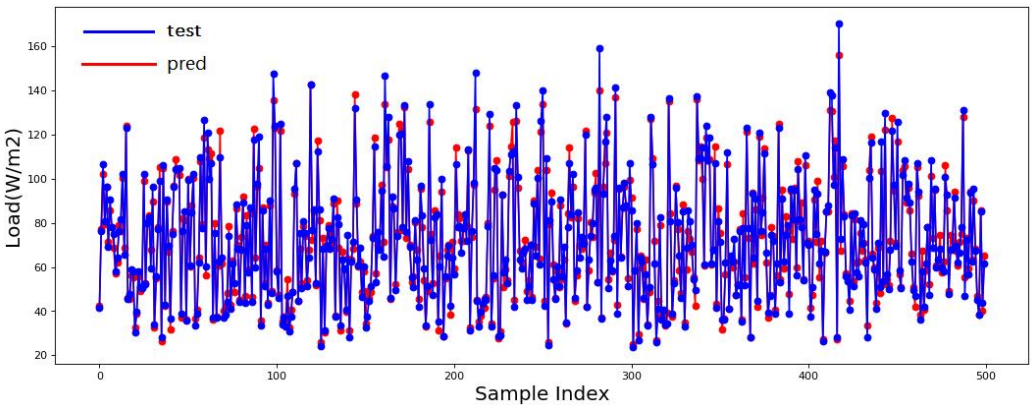
建立预测模型的机器学习算法有多种，例如人工神经网络、支持向量机、随机森林等。本课题采用随机森林算法建立预测模型，算法介绍见 5.3.2 小节。

本小节建立了两个对比模型，一个是将所有初始变量作为模型输入特征，二是仅以关键变量作为模型输入特征。训练和测试数据集来自采用回归系数法进行敏感性分析时得到的 6000 个样本点及对应计算负荷组合的数据集。两个模型的测试结果如图 2.9-2.10 所示。预测精度衡量指标为 CV-RMSE，计算公式为：

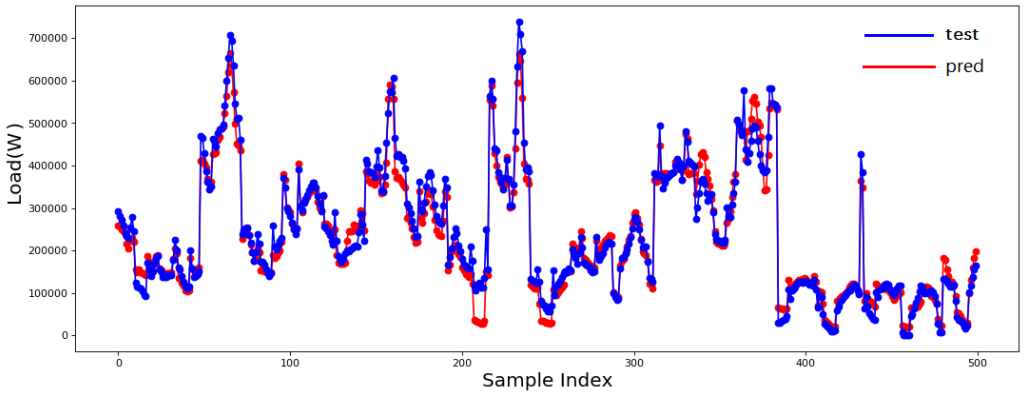
$$\mathrm{CV-RMSE}=\sqrt{\frac{\sum_{k=1}^{n}(y_k-\widehat{y_k})^2}{n}}\Big/\frac{\sum_{k=1}^{n}y_k}{n} \tag{2-1}$$

其中，y_k为实测值，$\widehat{y_k}$为预测值，n为算例个数。

预测结果表明，采用全部初始变量作为模型输入特征进行预测时，无论对设计负荷预测结果还是对小时负荷预测结果，预测误差都非常低。这是合理的，因为预测目标的变化完全可以用初始变量来解释。第二个模型只使用 5 个关键变量作为输入特征。如图 2.10 所示，模型的预测误差与第一个模型相比也很低。因此，本组对比试验表明，关键变量保留了描述建筑空调冷负荷变化的大部分信息。

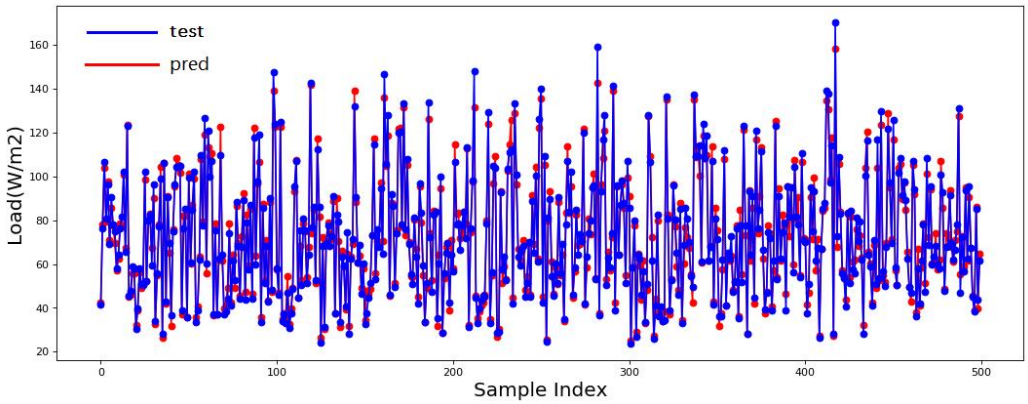


（a）设计冷负荷预测（CV-RMSE = 0.004%）

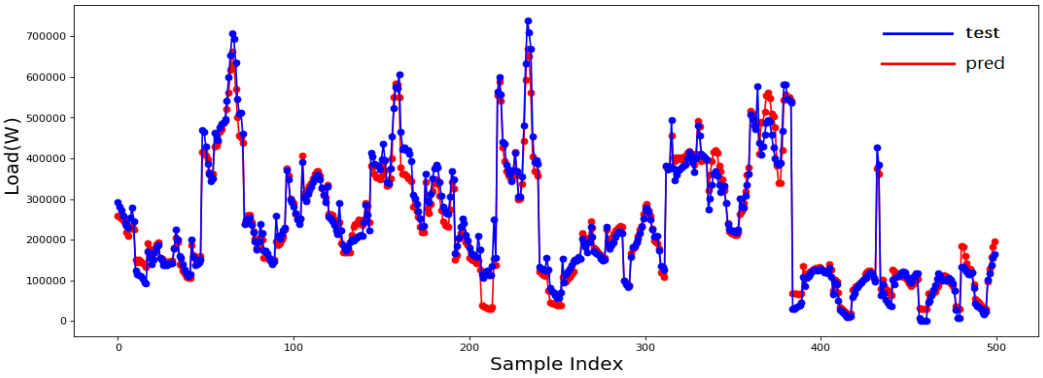


（b）逐时冷负荷预测（CV-RMSE = 0.0003%）

图 2.9 基于初始变量的负荷预测结果



（a）设计冷负荷预测（CV-RMSE = 0.08%）



（b）逐时冷负荷预测（CV-RMSE = 0.0074%）

图 2.10 基于关键变量的负荷预测结果

2.4 建筑空调系统运行能耗关键变量的提取

与上一小节进行负荷关键变量的选取类似，本小节针对空调系统的运行能耗选择关键变量。空调系统运行能耗主要取决于冷负荷需求和冷量供给系统，影响

冷负荷需求的关键变量已经在上述小节中进行，因此本小节着重针对冷量供给系统，即空调系统，进行关键变量的选取，采用全局敏感性和局部敏感性相结合的方法。分析的对象包括冷机能耗、水泵能耗、风机能耗、冷却塔能耗，由 EnergyPlus 计算得到。

2.4.1 影响建筑空调系统运行能耗初始变量集

在理想情况（即系统设计合理、运维得当、系统没有故障）下，空调系统的运行能耗主要取决于系统本身的特性，比如系统类型、设备能效等，但是在实际工程中，大部分建筑空调系统处于不理想的运行状态，例如大流量小温差、设备能效偏低、风管漏风等，导致系统实际运行能耗偏大。但是在能耗模拟时，一般默认系统运行正常，参数按照设计文件设置，不会考虑上述种种导致能耗偏离的“非正常”因素，所以从这个角度出发可以理解为什么能耗模型计算结果与实际运行能耗存在偏差。因此本课题在分析空调系统运行能耗的影响因素时，除了考虑空调系统本身特征外，还加入了描述系统运行状况不佳的变量，分别称之为“一般变量”和“附加变量”，汇总见表 4。其中，设备能效参数（风机效率、水泵效率、冷机 COP）即是系统的基本特征参数，也是可以反映系统运行效率低下的“附加变量”。

表 4 空调运行能耗敏感性分析初始变量集

变量名称		取值范围	单位
一般变量	风系统类型	定风量系统、变风量系统、风机盘管系统	
	水系统类型	一次泵定流量系统、一次泵变流量系统、一次泵定流量二次泵变流量系统	
	送风温度	8~18	^oC
	冷冻水供水温度	5~10	^oC
	风机效率	0.3~0.8	
	水泵效率	0.3~0.8	
	冷机 COP	3~7	
附加变量	冷冻水供回水温差	1~6	^oC
	换热盘管污垢系数	0~200	m^2K/W
	冷却塔填料堵塞率	0.5~1	
	风系统过滤器堵塞率	1~2	

2.4.2 建筑空调运行能耗算例生成

与建筑空调负荷计算模型的算例生成方法类似，空调系统算例的生成也是借助于 python 和 eppy 包批量生成 idf 文件，并进行计算，分别得到冷机能耗、水泵能耗、风机能耗和冷却塔能耗。需要说明的是，由于此处不涉及建筑负荷，因此空调系统的变化是基于同一建筑进行的，建筑特征和人员、照明等设置均保持不变。

由于描述空调系统基本特征的“一般变量”比较常见，此处不作说明。下面针对“附加变量”的设置和计算进行一些补充解释。

（1） 冷冻水供回水温度

“大流量小温差”是中央空调系统运行过程中的常见问题，一般冷冻水供回水温度为 7/12^oC，供回水温差 5^oC，但相关调研表明很多建筑的冷冻水供回水温差只有 2~3^oC，有时甚至只要 1^oC，造成实际水流量偏大，水泵的能耗大大增加[117]。因此本课题将“冷冻水供回水温度”作为“附加变量”之一来描述空调系统的非正常运行工况。

（2） 换热盘管污垢系数

空调箱表冷器中的冷热盘管会由于表面结垢等原因而造成传热系数（UA）的衰减，进而导致整体换热性能恶化，供回水温差减小，冷冻水流量增加，这里用换热盘管污垢系数来描述，换热盘管总的污垢系数是水侧和空气侧污垢系数的总和，计算公式见式 2-3：

$$R_{foul} = r_{air}/A_{air} + r_{water}/A_{water} \tag{2-3}$$

其中，R_{foul}是换热盘管总污垢系数，K/W；

r_{air}是空气侧污垢系数，m^2K/W；

r_{water}是水侧污垢系数，m^2K/W；

A_{air}是盘管空气侧面积，m^2；

A_{water}是盘管水侧面积，m^2；

于是，带有污垢的盘管传热系数计算公式为：

$$UA_{fouled} = 1/\left[1/UA_{air} + R_{foul} + 1/UA_{water}\right] \tag{2-4}$$

其中，UA_{fouled}是带有污垢的盘管传热系数，W/K；

UA_{air}是盘管空气侧的传热系数，W/K；

UA_{water}是盘管空气侧的传热系数，W/K；

R_{foul}是换热盘管总污垢系数，K/W；

于是可以通过分别定义盘管空气侧和水侧的污垢系数来得到盘管结垢对整体换热性能的影响，由于考虑到空气接触盘管时已经经过过滤器过滤，因此这里

不考虑空气侧的污垢系数，进考虑水侧污垢系数，取值范围见表 4。

（3） 冷却塔填料堵塞率

冷却塔填料堵塞是冷却塔运行过程中的常见故障，主要是由于循环水池中的水一般采用自来水或者地下水，水的硬度都比较大，矿物质成分比较高，随着水分慢慢蒸发，水中的矿物质浓度越来越高，发生钙化反应，或是由于杂物掉落进冷却塔中。被杂质堵塞的冷却塔的换热性能会严重下降，本课题通过定义冷却塔换热系数的削减程度来反应阻塞的严重情况。

$$UA_{tower,f} = UA_{tower,ff} \times F_{UA} \tag{2-5}$$

其中，$UA_{tower,f}$是被阻塞的冷却塔换热系数，W/K；

$UA_{tower,ff}$是无杂质的冷却塔换热系数，W/K；

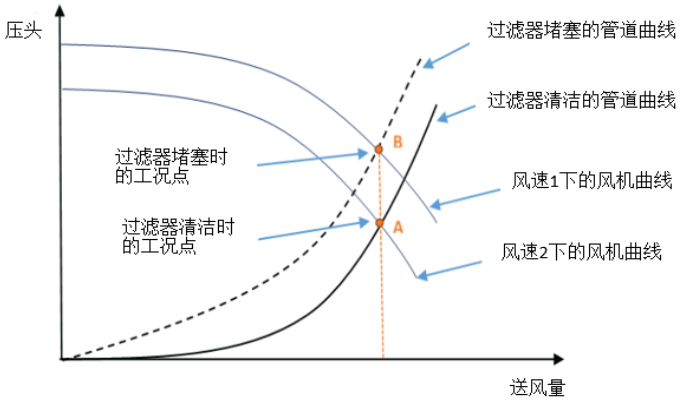
F_{UA}是冷却塔填料堵塞率。

（4） 风系统过滤器堵塞率

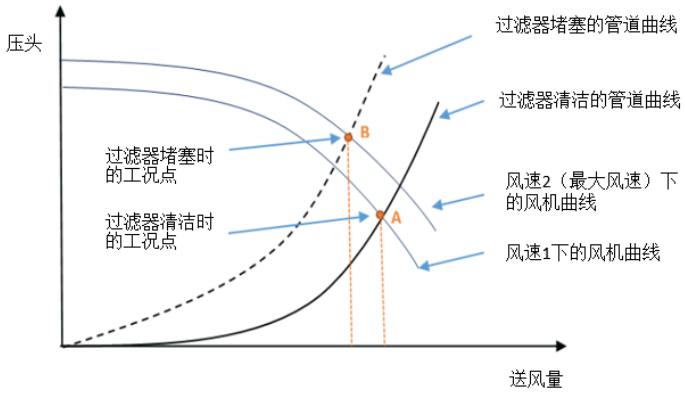
空调箱过滤器长时间不清洗或其他异物进入空调箱会使过滤器堵塞，造成整个系统阻力的增加，改变原有管道阻力曲线，进而改变风机的运行状态点。具体来讲，风系统过滤器堵塞会造成风机压头增大，能耗增加，出口空气温度上升，还可能造成系统风量减小，影响其他部件（例如冷热盘管）的工作状态。

在模拟计算时，通过定义风机压头的增加率和风机的风量-压头曲线来计算在过滤器被阻塞时风机的运行性能。在过滤器被堵塞时，风机的运行工况可能出现以下三种变化情况：

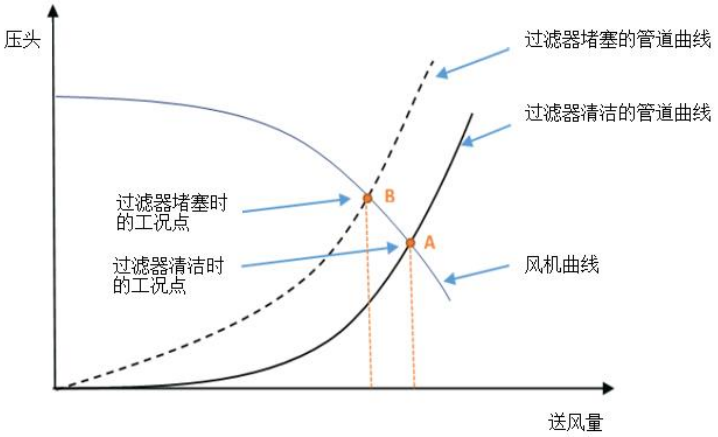
a) 风机变频，风量保持不变，风机转速增加（图 2.11-a）。风机从工况点 A 移动到工况点 B，压头增加，功率增加，出口空气温度增加。
b) 风机变频，风机转速增加到最大值也不足以维持原风量，风量减小（图 2.11-b）。风机从工况点 A 移动到工况点 B，压头增加，功率根据风量变化情况可能增加也可能减小。
c) 风机定频，风量减小（图 2.11-c）。与情况（b）类似，风机从工况点 A 移动到工况点 B，压头增加，功率根据风量变化情况可能增加也可能减小。



（a）



（b）



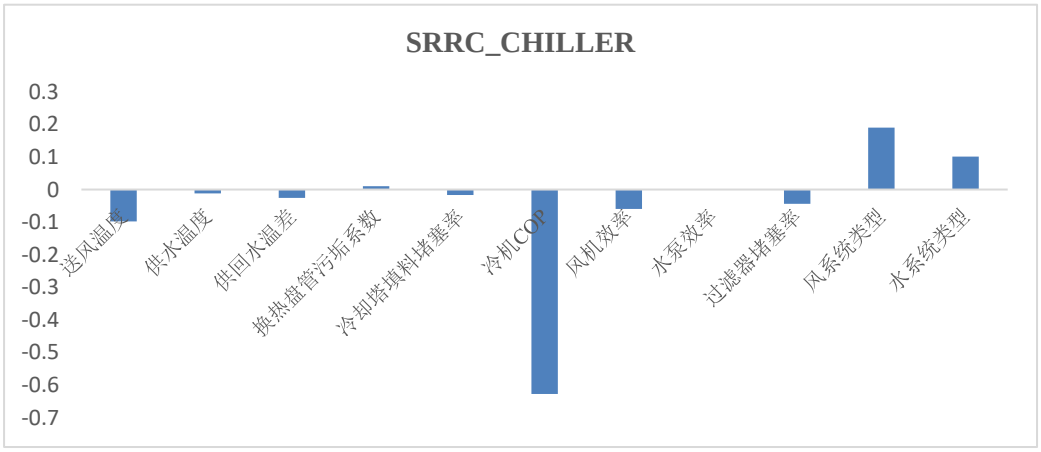
（c）

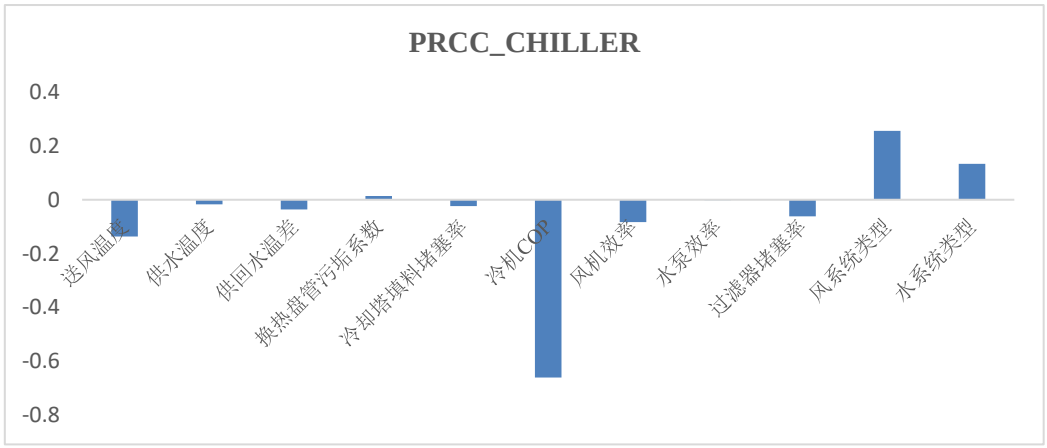
图 2.11 风系统过滤器堵塞时风机运行工况的变化

2.4.3 建筑空调运行能耗敏感性分析及关键变量展示

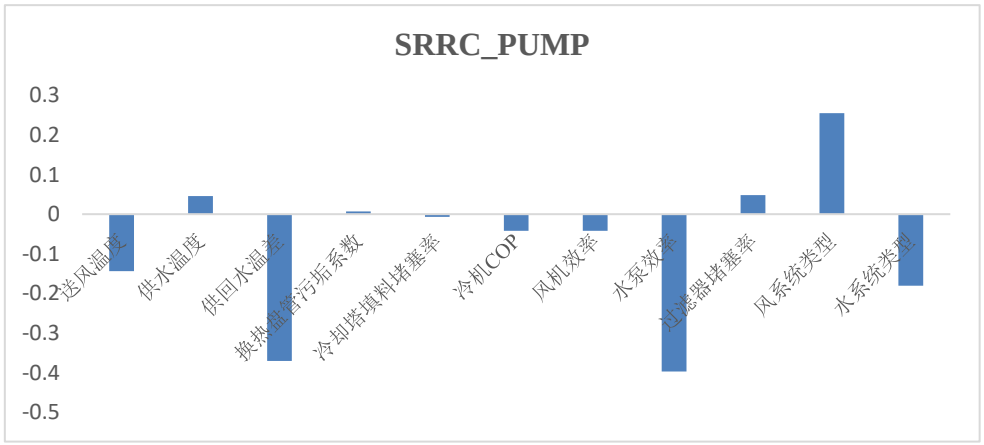
由表 4 可以看出，空调运行能耗初始变量中同时包含了数值型变量和非数值

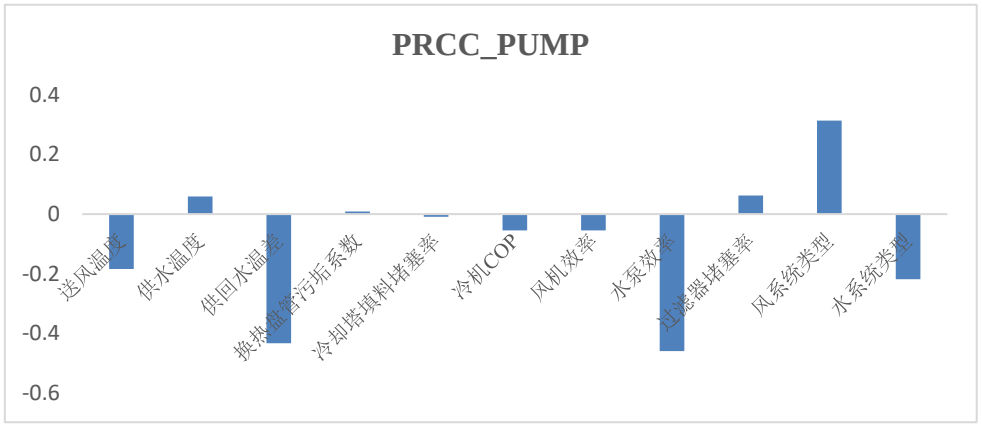
型变量，数值型变量采用拉丁超立方方法进行采样 600 个，非数值型变量（风系统类型、水系统类型）采用排列组合可得到 9 种组合形式。将数值型采样点和非数值型采样点直接排列组合成 5400 个样本点，在此基础上分别用 SRRC 和 PRCC 两个敏感性量化指标分别对冷机能耗、风机能耗、水泵能耗、冷却塔能耗的影响变量进行敏感性分析和量化，分析结果见图 2.12，其中某一变量对应的柱形越长，代表该变量越敏感，与其数值的正负无关。不难发现 SRRC 和 PRCC 两种方法得到的分析结果高度一致。根据各变量对不同分项能耗敏感性的高低确定各分项能耗的关键变量，总结见表 5。从表中可以看出，附加变量（供回水温差、过滤器堵塞率、冷却塔填料堵塞率）均对设备能耗具有高度敏感性，因此在分析实际设备运行能耗时必须将上述附加变量考虑进去，但这是过往相关研究所欠缺的。



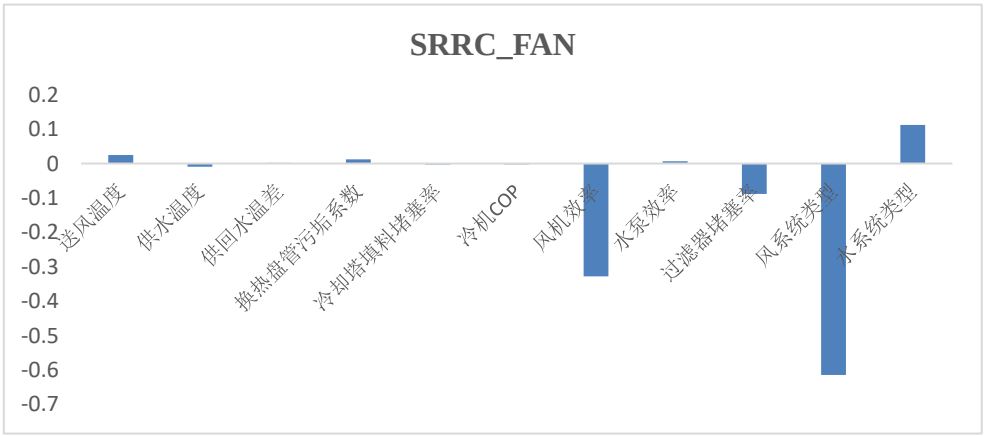


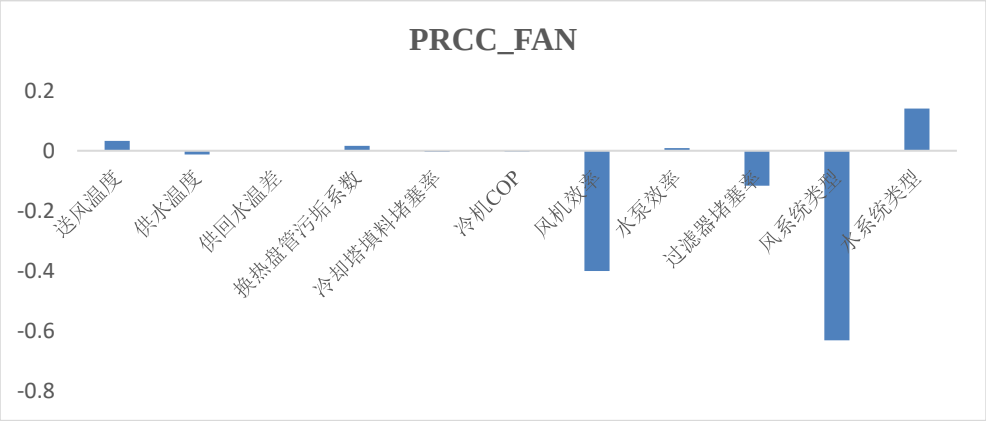
（a）影响冷机能耗的变量敏感性量化



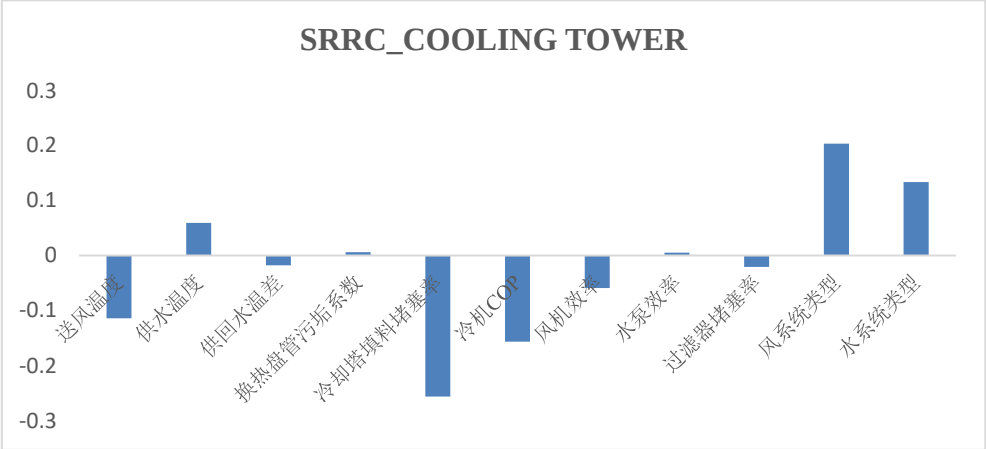


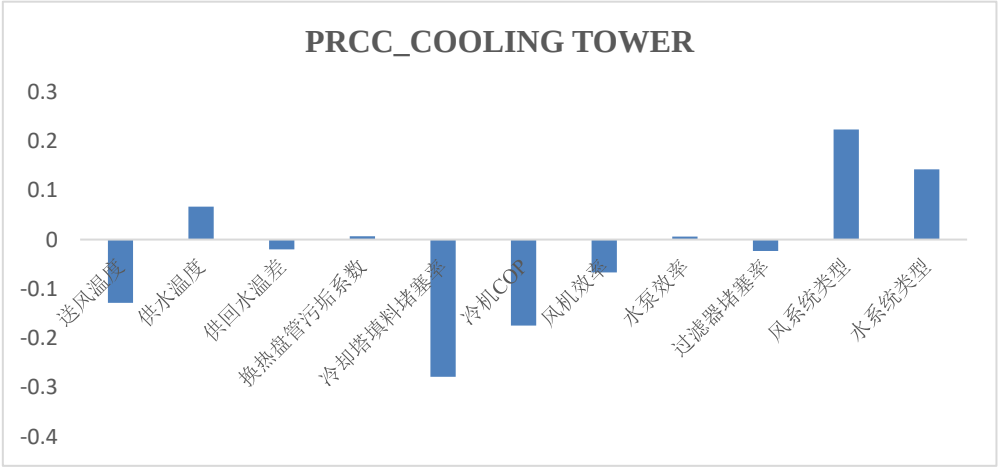
（b）影响水泵能耗的变量敏感性量化





（c）影响风机能耗的变量敏感性量化





（d）影响冷却塔能耗的变量敏感性量化

图 2.12 影响各制冷设备能耗的变量敏感性量化

表 5 各分项能耗的关键变量汇总

	冷机能耗	水泵能耗	风机能耗	冷却塔能耗
送风温度	△	△	△	△
冷冻水供水温度				△
冷冻水供回水温差		▲		
换热盘管污垢系数				
冷却塔填料堵塞率				▲
冷机 COP	▲			▲
风机效率			▲	△
水泵效率		▲		
风系统过滤器堵塞率			▲	
风系统类型	▲	▲	▲	▲
水系统类型	▲	▲	△	▲

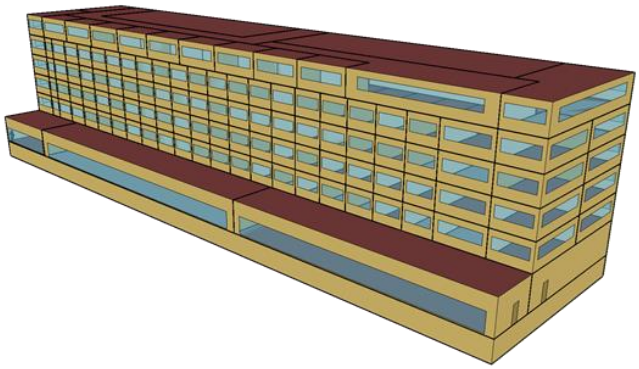
注：黑色三角形表示该变量有强敏感性，灰色表示敏感性较弱，无三角标识的变量认为其对该分项能耗不敏感。

2.4.4 建筑空调运行能耗关键变量有效性验证

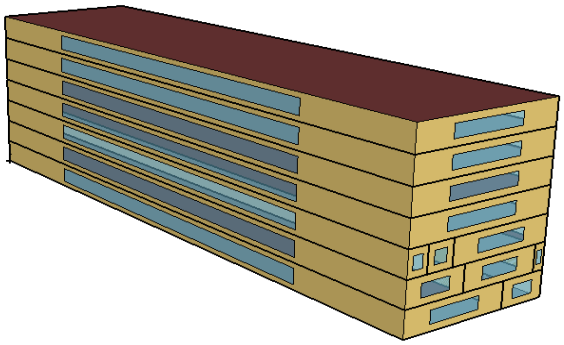
本小节用模拟的方法验证了利用本文所述方法提取建筑空调运行能耗关键变量的有效性。这里从美国 DOE 商业典型建筑模型库中选取酒店建筑模型为基准模型，因为这个模型库是根据大范围调研建立而成的，能够体现各类建筑的基本特征。在基准模型的基础上，本文对某些参数进行了改动，得到一个基准模型和四个对比模型，基准模型即为原典型酒店建筑模型，对比模型 I-IV 对模型参数进行了不同程度的改变。其中，对比模型 I 改变非关键变量的值，但保持关键变量（表中加粗部分视为关键变量）的取值不变；对比模型 II 在模型 I 的基础上改变了部分关键变量取值；而对比模型 III 在模型 II 的基础上进一步改变部分关键变量取值；直至在对比模型 IV 中，关键变量和非关键变量的取值都改变了。从模型 I 至模型 IV，所改变关键变量的敏感度逐渐增强。基准模型和对比模型的变量取值如表 6 所示，与基准模型相比，各参数值的偏差约基准值的 10%，并且所有变量的改变对输出目标（这里考虑冷机能耗）的变化方向是一致的。基准模型和四个对比模型的冷机电量均由 EnergyPlus 计算。

总冷机能耗偏差对比如图 2.14 所示，逐日冷机能耗偏差对比如图 2.15 所示。不难发现，从对比模型 I 到对比模型 IV，随着更多关键变量取值的偏移，其冷机能耗与基准模型冷机能耗偏差越来越大，并且所偏移的关键变量敏感度越高，冷

机能耗的偏差越大。换一个角度分析，从对比模型 IV 到对比模型 I，可以看出，随着越来越多关键变量取值的纠正，对比模型的偏差越来越小，即模型的精度越来越高，但模型 III 相较于模型 IV 的精度提升大于模型 I 相较于模型 II 的精度提升。也就是说，起决定性作用的是敏感度高的少数头部关键变量，弱敏感变量对提升模型精度的贡献是有限的。关键变量的选取没有严格的规则，关键变量越多，模型精度自然越高，但其边际效应呈递减趋势，因此关键变量数目的确定需要根据实际情况做权衡判断。



（a）基准模型



（b）对比模型

图 2.13 基准模型与对比模型外形对比

表 6 基准模型和对比模型参数设置表[2]

类型	参数名	变量值					单位
		基准模型	对比模型 I	对比模型 II	对比模型 III	对比模型 IV	
负荷	NWWR	0.26	0.26	0.33	0.33	0.33	

[2] 表中阴影部分表示该关键变量的取值与上一个对比模型保持一致。在这个对比实验中，系统类型作为关键变量之一，因其在实际工程中这是很容易得到的信息，因而在各对比模型中是保持一致的。

相关变量	**SWWR**	0.37	0.37	0.44	0.44	0.44	
	EWWR	0.24	0.24	0.31	0.31	0.31	
	WWWR	0.24	0.24	0.31	0.31	0.31	
	AREA	11345	11345	11345	11345	11345	m^2
	NL	6	6	6	6	6	
	CR	0.56	0.56	0.56	0.5	0.45	
	WALLU	0.698	1.2	1.2	1.2	1.2	$W/(m^2K)$
	WSP	2000	1640	1640	1640	1640	$J/(kg\ K)$
	RU	0.228	0.56	0.56	0.56	0.56	$W/(m^2K)$
	WINU	2.37	2.37	2.37	3	3	$W/(m^2K)$
	SHGC	0.39	0.39	0.39	0.48	0.48	
	WSA	0.7	0.9	0.9	0.9	0.9	
	RSA	0.7	0.9	0.9	0.9	0.9	
	SPC	24	24	24	24	23	℃
	SPH	21	22	22	22	22	℃
	LPD	12	12	12	13	13	W/m^2
	OPD	20	20	20	20	15	m^2/P
	INFIL	0.2	0.2	0.2	0.2	0.25	ACH
	ST	0.9	0.9	0.7	0.7	0.7	
	FLT	0.007	0.1905	0.1905	0.1905	0.1905	$W/(m\ K)$
	GLT	0.03	0.1328	0.1328	0.1328	0.1328	$W/(m\ K)$
	CLT	0.036	0.1008	0.1008	0.1008	0.1008	$W/(m\ K)$
系统相关变量	**AST**	风机盘管系统	风机盘管系统	风机盘管系统	风机盘管系统	风机盘管系统	
	WST	一次泵定流量二次泵变流量	一次泵定流量二次泵变流量	一次泵定流量二次泵变流量	一次泵定流量二次泵变流量	一次泵定流量二次泵变流量	
	SAT	12.8	12.8	13.8	13.8	13.8	^oC
	CWST	6.7	6.2	6.2	6.2	6.2	^oC
	FEffi	0.6	0.55	0.55	0.55	0.55	
	PEffi	0.9	0.85	0.85	0.85	0.85	
	COP	5.7	5.7	5.7	5.7	4.7	
	TDW	8.3	7.8	7.8	7.8	7.8	^oC

	CFF	100	120	120	120	120	m^2K/W
	CTFR	0.5	0.65	0.65	0.65	0.65	
	AFFR	1	1.3	1.3	1.3	1.3	

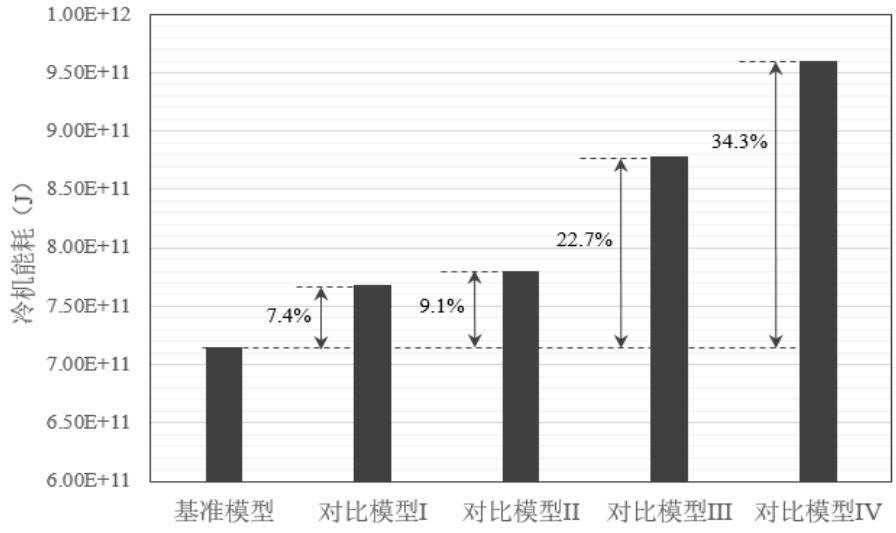


图 2.14 基准模型与对比模型冷机总能耗对比

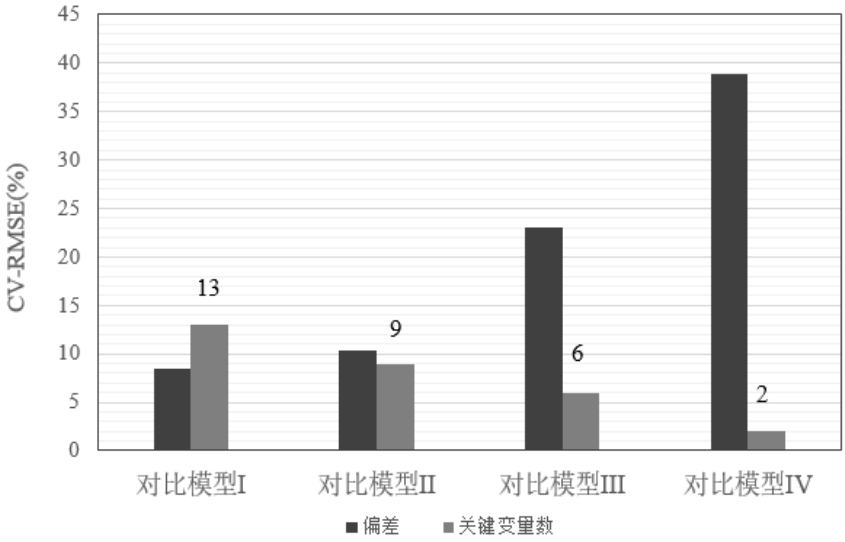


图 2.15 基准模型与对比模型冷机逐日能耗偏差对比

2.5 关键变量自动提取工具

如上所述，关键变量的提取是比较复杂但执行程序相对固定的过程，因此本课题考虑将其打包成一个软件工具包。本软件提供用户三个输入窗口，分别为建筑类型、气候区和敏感目标。建筑类型的可选项包括办公、酒店、商场、住宅类

型（目前仅支持酒店建筑）；气候区可选项包括夏热供冷地区、夏热冬暖地区、寒冷地区、严寒地区；敏感目标可选项包括冷负荷、热负荷、冷机能耗、锅炉能耗、风机能耗、水泵能耗、冷却塔能耗。本程序直接接收来自用户界面的输入参数，启动后根据建筑类型执行基准模型匹配搜索和初始变量集生成算法，然后执行其他模块调用算法，形成批量计算模型文件并执行计算，将计算结果进行存储，最后调用关键分析模块的结果进行变量的敏感度展示。模型生成模块包括三个子模块，分别为数据读取模块、建筑几何模型生成模块、系统模型生成模块。数据读取模块读取来自主控程序传出的变量集参数表，根据参数值生成对应的的几何模型和系统模型，形成计算文件，传给主控程序。本模块采用敏感性算法进行关键变量的分析和提取。敏感性分析的结果对初始条件很敏感。对于同样的输入参数，当目标输出参数发生变化时，敏感性分析结果可能会发生变化。因此，必须指定敏感分析目标，这里读取用户输入的“敏感目标”进行分析。软件采用 Morris 方法和回归分析法两种同时进行敏感性分析，对两种算法的结果进行综合，最后将结果回传到主控程序进行展示。软件架构及数据流图如图 2.16 所示，部分代码件附录 B。

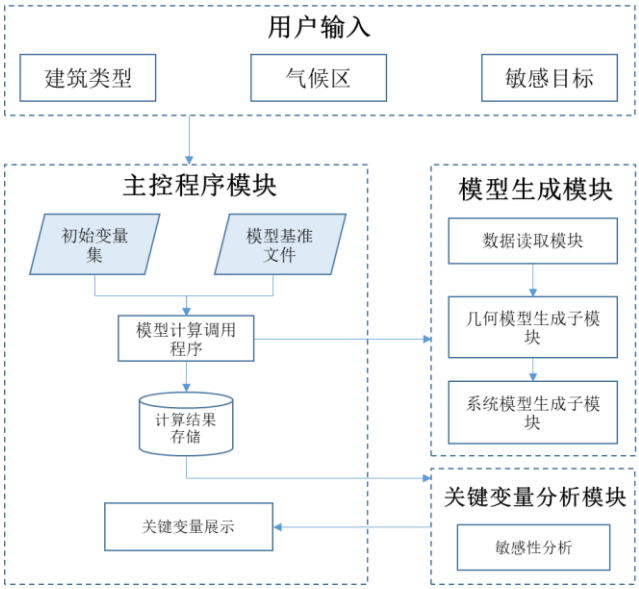


图 2.16 关键变量自动提取工具架构及数据流动图

2.6 本章小结

对建筑空调运行能耗可能存在的影响因素很多，在利用传统模拟工具计算建筑能耗时需要用户输入大量的参数，但是每个参数对于能耗的影响程度是不同的，在时间精力有限的情况下，应该重点考虑对能耗影响最重要的若干参数。另一方面，在利用数据驱动模型建立能耗预测模型时，模型输入特征应当包括能够解释

预测目标变化的变量，理想情况是将所有对预测目标有影响的参数都作为模型特征，但这是不现实的，原因有两方面，一是参数的搜集非常费时费力，二是模型特征维数过大易造成“维度灾难”，导致模型性能下降。综合两方面考虑，对影响空调能耗关键变量的进行提取是十分有必要的。

本章节采用了基于模拟的敏感性分析方法对空调能耗关键变量进行提取。进行敏感性分析首先需要确定初始变量，空调能耗同时由负荷和系统运行两方面决定，如果将两个维度的初始变量综合考虑，会使得样本点数量剧增，计算成本大大增加。因此，本文分别进行了负荷相关和系统相关的关键变量提取。在分析系统相关的变量时，本文不仅考虑了系统理想运行状态下的参数，即“理想变量”，还包括了系统偏离理想运行状态时的情况，加入了“附加变量”用以描述系统处于低效运行状态下的特征。

另外，为了协助完成敏感分析过程中的算例生成步骤，本章节开发了快速建模及参数分析工具，并打包成关键变量自动提取工具，该工具的开发语言为 python，基于 EnergyPlus 计算内核和 eppy 包。相比于现有的参数分析工具，本课题开发的工具更加灵活、功能更加全面。

最后，本章节从两个角度分别验证了本章节所提取的负荷相关的关键变量和系统相关的关键变量的合理性。结果表明，关键变量保留了大部分信息，用少数关键变量可以描述空调能耗的变化情况。

本章节以上海地区酒店建筑为例，提取了 5 个空调负荷相关关键变量，并针对系统中不同设备分别给出了相应的关键变量。需要说明的是，由于敏感性分析结果很容易受到边界条件选择的影响，因此本章节的结论不能直接移植到其他情况（例如气候区改变，或是其他类型的建筑），但方法论是通用的。另外需要注意的是，具体关键变量数量的确定需要根据实际情况综合权衡精度和计算量的冲突。

第3章 建筑空调能耗数据融合算法的建立

3.1 概述

通过第二章节的分析，我们可以通过少数几个关键变量描述建筑空调能耗变化和差异，大大简化了参数搜集工作。但是，对于大部分既有建筑而言，由于图纸和设计信息的缺失，确定这些关键变量的值是很困难的，需要通过仪器仪表对能源使用情况进行实测，并对建筑及其设备系统进行现场审计，这种方法能提供建筑的实际运行数据，但费事费力，且测量过程中难免会有误差。另一方面，我们还可以通过建立计算机能耗仿真模型来计算建筑能耗，这种方法成本低，仿真模型的建立同样面临输入参数难以获取的问题，这是影响能耗模拟结果的一个重要因素。

上述两种能耗获取和分析方法各具有缺点，具体来说，建筑能耗实测数据反映的是建筑用能的实际情况，可靠度较高，但由于传感器和传输系统故障的原因，采集的数据质量不高，经常出现离群值、缺失数据点。相比之下，模拟软件由于输入参数的不确定性、以及模型本身的不完备性，其计算结果与真实值之间存在着偏差，即存在“系统误差”，但是能耗仿真模型提供的计算结果较为稳定，不存在异常。也就是说，仿真模型计算得到的能耗随气象条件和使用特征变化的趋势是可靠的。

基于上述两类数据的基本特征，本章节提出一种将建筑空调能耗模拟数据和实测数据进行融合的通用方法，该方法用经过修正的能耗数据推测出未知关键变量的值。如此，一个复杂的建筑可以简化为若干关键变量和能耗序列值，从而构建复杂建筑的简单描述方法（即建筑信息及用能画像），便于格式化存储和处理，为后面的混合能耗模型建立奠定基础。该算法将建筑能耗实测数据和模拟数据进行融合，将两类数据互相补充修正，用模拟数据修正实测数据的异常数据，再用实测数据纠正模拟数据的“系统误差”，推断出建筑及空调系统未知关键变量的分布情况，并引入循环迭代融合机制以使能耗数据更加准确可靠。

本章节将从计算机能耗模型的不确定性、实测数据的误差来源以及数据融合算法原理三方面展开进行论述，具体的案例分析在第四章节进行展示。

3.2 建筑空调能耗数据融合算法

3.2.1 建筑能耗模型的不确定性

计算机模型的应用在各行各业都非常普遍，但是没有模型是可以完全复刻真实的物理过程的，因为模型本身包含着多种不确定性，归纳来讲，模型的不确定性主要包括以下几点：

- 输入参数的不确定。这是导致模型结果结果产生偏差的主要原因，一个模型可能需要非常多的输入参数，要确定每个参数的准确值是十分困难的事情。以能耗模型为例，有些参数本身是很难准确测量的，例如新风渗透率、人员在室率；还有些参数可以测量，但测量本身存在偏差，例如温度、流量等；另外，在建立能耗模型时，参数取值一般会参考设计文件，但在实际建造施工时可能会由于各种各样原因使实际值偏离设计值，例如围护结构由于施工质量不佳产生冷桥效应，使围护结构的传热系数偏大等。
- 模型本身的不完备。世界上没有一个模型是完美的，即使我们能得到所有参数的准确值，模型的计算结果与真实值依然有偏差，这就是由于模型的不完备造成的。因为所谓模型就是对真实物理现象的简化和抽象，在模型的建立过程中已经引入了许多“近似”，所以模型与实际情况天然存在偏差。
- 模型代码的不确定。由于模型大多是非线性的，很难直接求出解析解，因此复杂模型的求解一般用数值分析的方法求出近似解。

3.2.2 建筑能耗实测数据质量问题

公共建筑安装能耗计量监测系统是提高节能意识、排查能耗异常、提高设备运行能效的重要手段。中国自开展公共建筑能耗计量工作以来，已经积累了大量的能耗数据，但使用情况不容乐观，其中一个重要原因是数据质量问题。总的来说，电耗数据异常只要存在于四个方面：数值缺失、出现负值、等比例偏离、环境扰动异常，其成因及表现特征见表 7。第一类错误是没有采集到数据或数据为零，这是非常常见的错误，主要是由于采集系统的硬件故障引起的，包括传感器的故障或是传输线路的故障。第二类故障是数据出现负值，可能是由于系统接线失误（线接反了）造成的。第三类问题是数据序列连续地等比例偏低，主要是由于三相电流/电压中缺少了某些相位值。第四类错误主要是由于计量设备周围的环境引起的，使得测量值大幅偏离正常区间，与计量设备本身无关。除此之外，

实测电耗数据还存在噪声、调教偏离等情况[114]。

表 7 公共建筑能耗计量数据常见问题

错误种类	原因	表现
数据缺失或值为零	数据传输系统失效	无数据传输至平台
	测量仪表失效	没有数据记录或值为零
	数据采集系统失效	没有数据记录或值为零
数据值为负数	电流反向流动	实测耗电量为负数
比例偏差	电流或电压缺相	实测耗电量趋势合理但绝对值偏低
数据扰动	环境干扰	短时间内大幅偏离实际值

3.2.3 建筑能耗实测数据与模拟数据融合算法

为了得到一个建筑完整的基本信息及用能画像描述，需要对缺失的关键变量进行填补和推断，因此本小节提出了数据融合算法。该算法以贝叶斯推断（Bayesian inference）为框架，用实测能耗数据对未知关键变量进行估计，考虑到实测数据的不确定性，引入模拟数据对其进行修正，算法流程如图 3.1 所示。

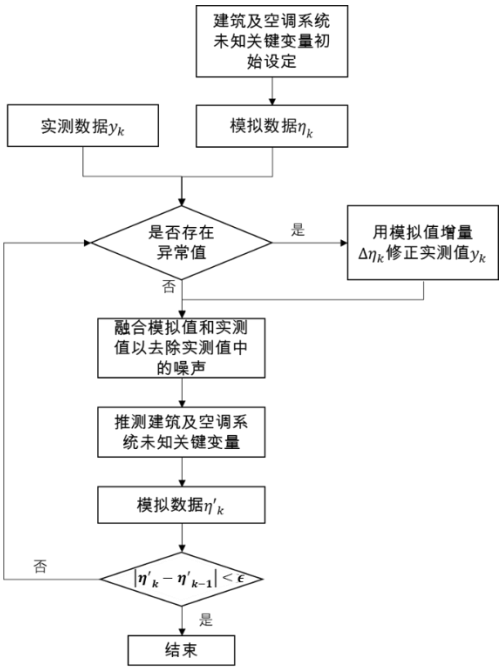


图 3.1 建筑能耗实测数据与模拟数据融合算法流程图

模拟数据来自于“代理能耗模型”（以下简称“代理模型”）。根据第二章的分析结果可知，建筑的能耗是由部分关键变量决定的，当非关键变量偏离实际值时不会对能耗造成很大影响，因此可以将非关键变量进行弱化，用一个简化的仿真模型去描述实际建筑，即所谓的“代理模型”，只要当代理模型和实际建筑的关键变量一致时，其能耗接近。由于代理模型易于操作，可以进行多次参数分析和数值计算，因此可以通过分析代理模型得到关键变量的值。代理模型的建立应满足以下三个条件：

1） 代理模型的建筑面积、层数应与目标建筑一致；
2） 代理模型的供能空间面积配比以及人员在室率、照明设备、空调启停等时间表设置能反映目标类型建筑及其机电系统的使用运行特征，本文研究的星级酒店建筑模型时间表设置如附录 A；
3） 气象参数使用实测数据而非典型气象年数据。

在本文研究范畴内，在数据融合过程进行之前，已知的数据是建筑能耗及部分关键变量，这一初始条件的设置也符合实际情况（已知某建筑的逐时能耗或电费账单，以及部分建筑信息）。由于部分关键变量值未知，因此可根据经验值或统计数据对未知的关键变量进行初始值设定，建立代理模型，得到模拟数据。此处能耗模型的搭建可以采用第二章节设计的模型快速生成工具。接着针对实测数据的异常值和噪声进行处理，详细描述见 3.4 小节。然后根据未知关键变量的先验值和修正后的实测能耗观测值，采用贝叶斯定理对未知变量进行推测，得到未知关键变量的后验分布。这里的未知变量推测方法借鉴了 Kennedy 和 O’Hagan 提出的方法[110]（简称 KOH 法），详见 3.5 小节。由于该方法一开始设定的未知关键变量值是根据经验确定，可能偏离真实值较大，因此模型的模拟能耗值也与实际值相差较大，据此进行实测数据修正的结果不理想，因此推测得到的关键变量可能也与实际值偏差较大，因此引入了循环迭代的机制，把推测得到的关键变量期望值代入模型再次进行计算、修正等步骤，直到前后两次的模拟计算值偏差小于阈值或是后次迭代后的结果相较于前次没有提升，即认为关键变量的推测收敛。

关于衡量误差的指标有多种，数据处理中常用的指标包括平均绝对误差（MAE），均方根误差（RMSE），均方根误差变异系数（CV-RMSE）,其计算公式分别为：

$$\mathrm{MAE} = \frac{\sum_{k=1}^{n}|y_k - \widehat{y_k}|}{n} \tag{3-1}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{k=1}^{n}(y_k - \widehat{y_k})^2}{n}} \tag{3-2}$$

$$\mathrm{CV-RMSE}=\sqrt{\frac{\sum_{k=1}^{n}(y_k-\widehat{y_k})^2}{n}}\Big/\frac{\sum_{k=1}^{n}y_k}{n} \tag{3-3}$$

其中，y_k是真实数据，$\widehat{y_k}$是预测数据。不难看出 MAE 和 RMSE 的值均与对比数据的绝对值有关，即有量纲的，不适用于不同数据集之间的横向对比，而 CV-RMSE 是无量纲的，对数据集的大小范围没有限制，因此本课题将以 CV-RMSE 作为对比指标。

3.3 基于模拟数据的实测能耗数据修正

由于传感器和传输系统的问题，实测数据普遍存在质量差的问题，目前在工程及学术界一般采用统计方法[150]或机器学习模型[151]对数据进行修正。常用的基于统计的数据修正或填补方法包括（1）根据数据集的均值或众数对异常值进行修正，（2）3σ法则。这种方法仅用简单的统计量捕捉数据集的分布特征，因而精度不高。基于机器学习模型的修正方法相对精度较高，该方法首先用数据集中的正常值建立分类或回归模型，再用模型进行预测，用于填补缺失值或修正异常值，常用的模型包括 k 近邻（KNN）、自组织网络（SOM）、多层感知机（MLP）、支持向量机（SVM）等，但这种方法的缺点同样也很明显，当数据集不够大或是异常值数量较多时，建立的模型本身精度较差，因此据此进行数据修正的效果较差。

根据 3.2 小节的讨论，模拟数据和实测数据各具有缺点，但两者可以相互弥补，因此本小节提出了基于模拟数据的实测能耗数据修正方法，将修正后的数据用于未知关键变量的推测。这里做两点假设：

1) 模型本身造成的偏差是稳定的，因此模型可以反映建筑能耗随外界条件的变化情况；
2) 测量值仅包含部分异常值、离群值和缺失值，不存在系统误差。

3.3.1 异常值修正

异常值在实测数据中是非常常见的，主要指大幅偏离正常数据分布的数据点，偏离值可为正或负，因此缺失值也可被当成异常值处理，本小节提出的的异常值修正方法适用于通常意义上的异常数据点以及缺失点，即 3.2.2 小节论述的四类异常数据。由于能耗模拟数据能较为准确的反映出建筑能耗随外界环境的变化趋势，并且不存在异常值，因此这里采用能耗模拟数据序列的变化来监测和修正实测数据中的异常值。

首先，根据根据经验或已知信息确定第二章节论述的关键变量取值，建立目标建筑的代理能耗模型，进行计算，得到与实测数据颗粒度一致的模拟数据序列。

然后，计算各时刻实测数据和模拟数据分别对应于上一时刻的增长值Δy_k，$\Delta \eta_k$，d_k，以及两者的差值（式 3-4~3-6）：

$$\Delta y_k = y_k - y_{k-1} \tag{3-4}$$

$$\Delta \eta_k = \eta_k - \eta_{k-1} \tag{3-5}$$

$$d_k = |\Delta \eta_k - \Delta y_k| \tag{3-6}$$

其中，k表示时间序列，y_k表示观测值，η_k表示对应模型的模拟值。如果d_k大于事先设定的阈值ε，那么就认为该时刻的实测值是异常的，并通过式 3-7~8 进行修正：

$$y'_k = y_{k-1} + \tau \Delta \eta_k \tag{3-7}$$

$$\tau = \frac{1}{N} \sum_1^N y_i / \eta_i \tag{3-8}$$

其中判断是否出现异常的阈值ε需要根据测量仪表的误差范围确定，本课题确定ε的取值为 $2\Delta \eta_k$。

3.3.2 噪声值去除

测量噪声也称为随机误差，或统计不确定性。它与系统误差有本质区别。系统误差是指由于仪器的错误校准或测量中没有考虑到的一些物理影响等缺陷而引起的测量误差，原则上可以检查和纠正。而测量噪声是不可避免的，降低测量噪声最常用的方法是多次测量取均值。但是对于本课题的研究对象—建筑能耗，不具备可重复测量性，安装多个测量设备成本太高。针对这个问题，本小节用模拟数据来降低实测能耗数据的随机误差。

由于随机误差成正态分布，因此对于静态测量可以将多次测量值求和，然后取平均值来消除随机误差，但在本课题的研究对象—能耗是动态变化的，多次测量取均值后虽然消除了随机误差，但也抹去了动态特征，因此需要借助没有随机误差的能耗模拟值来重构数据序列的动态特性。

本方法首先定义一个移动窗口，包含实测数据和模拟数据对应时刻的各N个序列值，分别用y_i^m和η_i^m表示，m表示当前移动窗口，i表示当前移动窗口中第i个值，$1 \le i \le N$。目标值y_k，η_k位于窗口的中间位置（对于前$2/N$和后$2/N$个值做相应的位置前移或后移，以保证窗口中数据个数为N）。对该窗口中的实测数据和模拟数据分别进行求和平均，消除随机误差（式 3-8、3-9）：

$$\overline{y_k} = (\sum_{i=1}^N y_i^m)/N \tag{3-8}$$

$$\overline{\eta_k} = (\sum_{i=1}^N \eta_i^m)/N \tag{3-9}$$

其次，由于模拟数据的序列增量比实测数据更加可靠，因此在$\overline{y_k}$的基础上增加由模拟数据重构的动态特性，得到修正后的数据$y_{f,k}$（式 3-11）：

$$y_{f,k} = \overline{y_k} + \eta_k - \overline{\eta_k} \quad (3\text{-}10)$$

3.4 基于贝叶斯的建筑及空调系统未知关键变量推断

在前面两小节中，本课题利用模拟数据对实测数据进行了去异常、填补缺失值、降噪等修正工作，完成了数据融合的前半部分内容。在第二章本课题分析得到了建筑及空调系统的关键变量，关键变量结合建筑用能数据就可以完整描述一栋建筑的用能特征，即建筑信息及用能画像。但在实际工程上，大部分建筑仅采集了能耗数据，建筑信息关键变量是缺失的，也很难获取，本小节利用经过修正后的能耗实测数据对建筑及空调系统关键未知变量进行推断，采用的方法是KOH 法，该方法基于贝叶斯理论，能同时降低模型和输入参数的不确定性。KOH 法的统计学框架如式 3-11 所示，包含了三部分不确定因素：（1）参数不确定；（2）模型不完备；（3）观测误差。

$$y(x) = \eta(x,\theta) + \delta(x) + \epsilon \quad (3\text{-}11)$$

其中，x表示可观测可控制的变量，例如天气参数等，θ表示需要被推测的未知变量，例如新风渗透率、人员密度等，$y(x)$表示观测值，$\eta(x,\theta)$表示模拟值，$\delta(x)$表示实际模型与真实过程之间的差异，即模型不完备性，ϵ表示观测误差。在上述 KOH 框架中，$\eta(x,\theta)$和$\delta(x)$分别用高斯过程回归模型来描述。

3.4.1 贝叶斯推断

3.4.1.1 贝叶斯推断

贝叶斯推断是推论统计的一种方法，该方法使用贝叶斯定理，将后验概率（考虑相关证据或数据后，某事件或参数的概率）推导为先验概率（考虑相关证据或数据前，某事件或参数的概率）和似然函数（由观测数据获得）的结果。贝叶斯推断可以用来做参数估计。

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \propto P(y|\theta)P(\theta) \quad (3\text{-}12)$$

其中，$P(\theta|y)$是参数θ考虑证据或数据后的后验概率；$P(\theta)$是未考虑证据或数据时的先验概率；$P(y|\theta)$是似然函数，似然函数描述了模型参数值的在不同取值组合下，得到特定观测值的概率，一般可以通过最大化似然函数，即极大似然估计来对未知参数进行估计。 $P(y)$是观测值的概率，由于与待推测变量θ无关，一般可以忽略。在本课题的研究范畴中，θ即为待推测的变量，y是已知的观测数据，即能耗数据。

在贝叶斯推断中，参数的先验概率分布扮演很重要的角色，它表达了我们对于参数的初步判断，知道其取值范围和大概的分布，通常根据经验或过往研究获取，但也有很多情况我们对待估计参数知之甚少。一般来说，先验概率分布可以根据先验知识的多寡分为三种：强信息先验（Informative Priors）和弱信息先验（Weakly Informative Priors）。

（1） 强信息先验

待推测变量θ的先验知识一般来自调研实测或经验积累。对于本课题研究内容，可以对目标建筑进行现场调研来确定待推测变量θ的先验分布。当我们对待推测变量θ有一定的了解时，需要将所了解信息包含进θ的先验分布，并且随着验证数据的不断增加，可以将前一次的后验分布作为当前推测的先验分布，实现分层递进式的贝叶斯推断。例如，如果当前模型是在获取了更多数据的情况下对前期版本改进，那么前期模型的后验分布可以作为当前模型的先验。按照这样的思路，模型的每一个版本更新都不是从无到有的建立过程，而是对前期工作的改进。如果前后的验证数据是相近的，那么θ的后验分布精度将被逐渐提高；反之，如果前期验证数据有误，那么通过后期验证数据的不断更新和增加，θ的后验分布会偏离先验分布。因此当有多组数据时，采用分层式的贝叶斯推断可以取得不错的效果。

（2） 弱信息先验

弱信息先验是指对参数进行必要的规定，以防止结果出现算法错误或违背常识。相比于强信息先验，采用弱信息先验得到的结果更多的从验证数据中获取信息，这样可以防止先验知识由于掺杂了过多人为主观判断而对后验分布造成偏差。但在这种情况下，弱信息先验需要更多的验证数据支持，为了获取更精确的后验分布，先验分布需要根据样本容量进行调整，这一操作有较高的技巧性。

综上所述，贝叶斯推断是结合了先验知识和验证数据得到的综合效果，当我们对先验知识比较确信时，可以将参数的先验分布设置成强信息先验分布，例如方差较小正态分布；反之，则需要更多的依赖验证数据，这时仅需要对参数的先验做一个较弱的规定，例如将先验分布设置成某一区间内的均匀分布。

3.4.1.2 贝叶斯推断后验分布的数值近似

对于多参数的推测，贝叶斯后验分布往往是高维非线性的，很难求出解析解，因此常用的方法是通过数值方法对后验分布进行采用，进而得到参数的近似后验概率密度分布。其中最常用的方法是马尔科夫链-蒙特卡罗采样法（Markov chain Monte Carlo，MCMC）。MCMC 方法其实是采用某一类方法的统称，这类方法的特点是同时采用了马尔科夫链和蒙特卡洛采样。马尔可夫链是随时间以“无记忆”

方式演化的随机过程，称为马尔可夫特性。马尔可夫特性意味着一个马尔可夫链的状态转移到另一个状态的概率只取决于系统的前一个状态，而不是它的整个历史过程。另一方面，蒙特卡罗抽样通过重复随机抽样来解决确定性问题常用的 MCMC 方法是 Metropolis-Hastings 算法，Gibbs 算法和 Hamiltonian 蒙特卡洛算法。MCMC 算法简单来讲，就是随机产生样本，然后根据某一规则选择留下或舍弃该样本。但并不是所有的 MCMC 算法都是一样的。对于参数较多的复杂模型，简单的方法，如 Metropolis-Hastings 算法和 Gibbs 算法可能需要很长时间才能收敛到目标分布，主要是这些方法采用低效随机游动探索参数空间。当模型参数是连续的而不是离散的，Hamiltonian 蒙特卡洛算法采用一阶梯度信息来加速收敛过程，因此更高效，适用于多维参数情况。但用户使用 Hamiltonian 蒙特卡洛算法时需要指定两个参数，迭代步长和步数，选择不恰当会导致算法的效率急剧下降。本文采用的是 Hamiltonian 蒙特卡洛算法的一种变形，称为 No-U-Turn Sampler（NUTS）。NUTS 的一个特点是不需要指定迭代步长和步数，算法可以实现自动调整[122]。判断采样序列是否收敛是采用 MCMC 的重要一环，目前一般采用 Gelman[123]提出定量收敛指标$\sqrt{R}$（比例缩小得分）来判断 MCMC 计算是否是收敛，其计算公式为[124]：

$$\sqrt{R}=\sqrt{\frac{g-1}{g}+\frac{q+1}{q\times g}\frac{B}{W}} \tag{3-13}$$

其中，g每个参数的采样迭代次数，q为用于评价的序列数，$\frac{B}{g}$为q个序列的平均值的方差，W为q个序列的方差的平均值。需要计算每个待估计参数的$\sqrt{R}$值，若接近于 1 则表示该参数的后验分布趋于收敛[124]。

在大多数情况下，由 MCMC 计算得到参数后验分布呈偏正态分布，取后验分布的均值$\hat{\theta}=E(\theta|y)=\int_{\theta}\theta f(\theta|y)d\theta$为最终的参数估计值[122]。

3.4.2 高斯过程回归

在本课题中，输入变量与模拟值的映射关系$\eta(x,t)$和输入变量与误差的映射关系$\delta(x)$都用高斯过程回归模型来描述。高斯过程回归模型是机器学习算法的一种[112]。对于一般的回归统计模型，我们通常会假设回归函数$f(x)$的形式，例如与自变量呈线性关系或二次、三次函数关系，然后采用最小二乘法来得到函数的参数值。高斯过程回归则不同，它并没有对函数形式做任何具体的规定，而是把它当成一个随机变量。要了解高斯过程回归模型首先需要知道高斯过程。高斯过程是定义在连续域（时间或空间）上的无限多个随机变量所组成的随机过程，高斯过程的任意随机变量的线性组合的联合分布都服从多维正态分布，高斯过程有

其均值和协方差函数完全确定，协方差函数也被称为核函数$k(x,x')$。我们一般假设高斯过程的均值为0，因此观测值之间的联系仅由其核函数确定。核函数的形式有多种，其中比较受常用的是径向基核函数（RBF Kernel）：

$$k(x,x')=\sigma^2\exp[\frac{-(x-x')^2}{2l^2}] \tag{3-13}$$

对于径向基核函数，最大允许的方差由σ^2确定。如果x和x'相接近，那么$k(x,x')$取最大值，即$f(x)$与$f(x')$相关性高；反之如果x和x'相差很远，$k(x,x')\approx 0$，即$f(x)$与$f(x')$不相关。

对于高斯过程回归问题，已知训练集（X,Y），X,Y分别为 n 维的多维变量，需要预测在新的自变量x_*所对应的值y_*。假设Y服从联合正态分布，那么有：

$$Y\sim N(0,K) \tag{3-14}$$

$$K=\begin{bmatrix}k(x_1,x_1)k(x_1,x_2) & \dots & k(x_1,x_n)\\ k(x_2,x_1)k(x_2,x_2) & \dots & k(x_2,x_n)\\ \dots & \dots & \dots\\ k(x_n,x_1)k(x_n,x_2) & \dots & k(x_n,x_n)\end{bmatrix} \tag{3-15}$$

将训练集观测数据Y和预测值y_*合并，它们同样服从联合正态分布，于是有：

$$\begin{bmatrix}Y\\ y_*\end{bmatrix}\sim N\left(0,\begin{bmatrix}K & K_*^T\\ K_* & K_{**}\end{bmatrix}\right) \tag{3-16}$$

$$K_*=[k(x_*,x_1)\ \ k(x_*,x_2)\dots k(x_*,x_n)] \tag{3-17}$$

$$K_{**}=k(x_*,x_*) \tag{3-18}$$

可以求出：

$$y_*|Y\sim N(K_*K^{-1}Y,K_{**}-K_*K^{-1}K_*^T) \tag{3-19}$$

从式（3-19）可以看出采用高斯过程回归得到的预测值不是一个确定的数值，而是概率分布，由此产生的一个突出优势就是高斯过程回归不仅能模拟任何黑盒函数，还能模拟不确定性，这是其他机器学习方法所无法实现的。对于y^*的最优估计是其均值$K_*K^{-1}Y$，其不确定性是方差$K_{**}-K_*K^{-1}K_*^T$。

3.4.3 KOH 法

对于一个真实的物理过程，以用计算机模型$\eta(x,t)$去模拟，其中t是模型中的未知参数，t的真值为θ，那么θ的取值可以由 KOH 法求得。KOH 法最早由 Marc C.Kennedy 和 Anthony O'Hagan 提出[110]，该方法以贝叶斯定理为基础，用高斯过程回归去拟合计算机模型$\eta(x,\theta)$和模型与真实过程之间的误差$\delta(x)$（式 3-12），被用于计算模型的参数校准和预测。其中

对于不存在系统误差，与真实过程之间仅有观测误差ϵ而言，有：

$$y(x)=\eta(x,\theta)+\epsilon \tag{3-20}$$

假设ϵ服从正态分布，$\epsilon\sim N(0,\sigma^2)$，那么$y(x)-\eta(x,\theta)\sim N(0,\sigma^2)$，于是有似

然函数：

$$L(y|\eta(\theta)) \propto \exp\{-\frac{1}{2}(y-\eta(\theta))^T\Sigma_y^{-1}(y-\eta(\theta))\} \quad (3\text{-}21)$$

其中，$y=(y(x_1),y(x_2),\dots,y(x_n))^T$，$\eta(\theta)=(\eta(x_1,\theta),\eta(x_2,\theta),\dots,\eta(x_m,\theta))^T$，$n$和$m$分别是观测值和模拟值的个数。$\Sigma_y$是观测值的协方差矩阵，$\Sigma_y=\sigma^2\cdot I_n$。

根据贝叶斯定理，$\pi(\theta|y)\propto L(y|\eta(\theta))\times\pi(\theta)$。这里$\pi(\theta)$是$\theta$的先验分布，假设$\pi(\theta)\sim N(0.5,0.25^2)$，那么$\theta$的后验分布为：

$$\begin{aligned}\pi(\theta|y) &\propto L(y|\eta(\theta))\times\pi(\theta)\\ &\propto \exp\{-\frac{1}{2}\big(y-\eta(\theta)\big)^T\Sigma_y^{-1}\big(y-\eta(\theta)\big)-\frac{1}{2(0.25^2)}(\theta-0.5)^2\}\end{aligned} \quad (3\text{-}22)$$

如果$\eta(\theta)$是线性方程，那么$\pi(\theta|y)$可以较容易的算出解析解，但对于本课题研究内容，建筑的负荷或能耗与诸多因素呈复杂非线性关系，因此这里采用 3.4.1.2 节中介绍的 MCMC 方法近似出θ的后验分布。

在上述分析过程中，计算机模型$\eta(x,t)$是一个未知函数，KOH 法用高斯过程回归模型来对其进行建模。一个典型的高斯过程由其均值方程$\mu(x,t)$和协方差方程$Cov((x,t),(x',t'))$唯一确定。根据 KOH 法，均值方程$\mu(x,t)$被认为是一个常数，协方差方程采用如下形式（RBF 协方差方程）：

$$Cov\big((x,t),(x',t')\big)=$$

$$\frac{1}{\lambda_\eta}\exp\left\{-\sum_{k=1}^{p}\beta_k^\eta|x_{ik}-{x'}_{ik}|^\alpha-\sum_{k'=1}^{l}\beta_{p+k'}^\eta|t_{ik}-{t'}_{ik'}|^\alpha\right\} \quad (3\text{-}23)$$

其中，x是p维向量，t是l维向量，λ_η是高斯过程模型的精度超参数，$\beta_1^\eta,\dots,\beta_{p+l}^\eta$是高斯过程模型的相关性超参数，$\alpha$是高斯模型的平滑性超参数。这些超参数都需要先指定先验分布，后根据贝叶斯定理求出其后验分布。

现在将观测值$y=(y(x_1),y(x_2),\dots,y(x_n))^T$和计算机模型的模拟值$\eta(t)=(\eta(x_1^*,t_1^*),\eta(x_2^*,t_2^*),\dots,\eta(x_m^*,t_m^*))^T$进行联合，得到$z=(y^T,\eta^T)^T$，相应的输入参数为$(x_1,\theta),(x_2,\theta),\dots,(x_n,\theta)$和$(x_1^*,t_1^*),(x_2^*,t_2^*),\dots,(x_m^*,t_m^*)$。于是，对于$z$的似然函数为：

$$L(z|\theta,\mu,\lambda_\eta,\beta^\eta,\Sigma_y)\propto|\Sigma_z|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(z-\mu I_{n+m})^T\Sigma_z^{-1}(-\mu I_{n+m})\} \quad (3\text{-}24)$$

$$\Sigma_z=\Sigma_\eta+\begin{pmatrix}\Sigma_y & 0\\ 0 & 0\end{pmatrix} \quad (3\text{-}25)$$

其中，I_{n+m}是$n+m$维的单位向量。Σ_y是$n\times n$维的协方差矩阵，$\Sigma_y=I_n/\lambda_\epsilon$。$\Sigma_\eta$是根据式（3-23）求得的$(n+m)\times(n+m)$维协方差矩阵。

在上述分析过程之上，KOH 法又进一步考虑计算机模型与真实过程之间的偏差$\delta(x)$，并且也用高斯过程回归模型进行建模，其均值函数为 0，协方差函数

为：

$$Cov(x,x') = \frac{1}{\lambda_\delta}\exp\left\{-\sum_{k=1}^{p}\beta_k^\delta |x_{ik} - x'_{ik}|^{\alpha_\delta}\right\} \quad (3\text{-}26)$$

考虑偏差$\delta(x)$的似然函数与式（3-24）形式一致，区别仅在于Σ_z：

$$\Sigma_z = \Sigma_\eta + \begin{pmatrix} \Sigma_y + \Sigma_\delta & 0 \\ 0 & 0 \end{pmatrix} \quad (3\text{-}27)$$

Σ_δ是根据式（3-26）求得的$n \times n$维协方差矩阵。

最后，根据贝叶斯定理求出待推测参数θ和其他超参数的后验分布：

$$\pi\left(\theta,\mu,\lambda_\eta,\lambda_\delta,\beta^\eta,\beta^\delta \middle| z\right) \propto$$

$$L(z|\theta,\mu,\lambda_\eta,\lambda_\delta,\beta^\eta,\beta^\delta,\Sigma_y)\pi(\theta)\pi(\mu)\pi(\lambda_\eta)\pi(\lambda_\delta)\pi(\beta^\eta)\pi(\beta^\delta) \quad (3\text{-}28)$$

关于超参数的先验分布$\pi(\mu),\pi\left(\lambda_\eta\right),\pi(\lambda_\delta),\pi(\beta^\eta),\pi(\beta^\delta)$，均值函数$\mu = 0$，另外文献[113]中给出了其他超参数的参考先验分布：

$$\lambda_\eta \sim Gamma(a = 10, b = 10) \quad (3\text{-}29)$$

$$\lambda_\delta \sim Gamma(a = 10, b = 0.3) \quad (3\text{-}30)$$

$$\lambda_\epsilon \sim Gamma(a = 10, b = 0.03) \quad (3\text{-}31)$$

$$\left(-\frac{\beta^\eta}{4}\right) \sim Beat(a = 1, b = 0.3) \quad (3\text{-}32)$$

$$\left(-\frac{\beta^\delta}{4}\right) \sim Beat(a = 1, b = 0.3) \quad (3\text{-}33)$$

关于带推测变量θ的先验分布，考虑采用弱信息先验，根据经验或相关规范取定范围，其分布则设置为均匀分布。

3.5 本章小结

建筑空调能耗的数据有两个渠道，一是通过现场调研审计或者数据实时采集系统获取的实测数据，二是通过能耗模拟软件进行计算得到的模拟数据，两者各有优缺点，实测数据能反映建筑的实际用能情况，但往往存在误差、异常值等，并且现场采集数据需要大量的人力物力，采集的数据量有限，并不能反映所有工况下建筑的用能情况；而利用能耗模拟软件则能通过计算大量案例快速得到任何建筑类型任何工况下的用能特点，但是由于输入参数的不确定以及模型本身的不完备性，计算结果虽然能较准确地反映能耗变化趋势，但往往与实际值存在系统偏差。

基于上述情况，本章节提出了一种将实测数据和模拟数据进行融合的方法，该方法通过“代理模型”将两类数据进行链接。根据第二章的研究结论，建筑空调能耗主要是由若干关键变量决定的，因此一个完整的建筑用能信息画像由两部

分组成：（1）影响建筑用能的建筑基本信息，（2）建筑能耗时间序列值，可以为逐时、逐日或是逐月等多种颗粒度。我们认为，在关键变量取值一致的情况下，“代理模型”的计算结果与实际建筑的用能情况是接近的（但由于不确定性的存在，两者依然存在偏差）。但目前绝大部分既有建筑，其关键变量是缺失的，部分关键变量甚至很难现场测量，部分关键变量由于系统运行维护不当，已偏离设计值。在本章节的数据融合算法框架中，研究对象是这类建筑基本信息缺失，但有历史能耗数据记录的建筑，首先建立“代理模型”，用其模拟数据对实测数据进行修正，去除其中的异常值和噪声；其次，利用经过修正的实测数据对建筑空调能耗关键变量进行推测，该推测算法基于贝叶斯理论，结合预设的关键变量先验分布和由实测数据得到的似然函数，根据 MCMC 方法抽样得到关键变量的后验分布，选定分布的期望作为关键变量的推测值。在完成数据融合后，我们得到了该建筑的完整用能信息画像和能反应其用能情况的“代理模型”，于是可以基于“代理模型”对该建筑在时间和空间两个维度进行数据填补，得到丰富的数据资源，用于后续混合能耗模型的构建。

第4章 建筑空调能耗数据融合算法的验证

4.1 概述

上一章节针对建筑能耗数据融合算法的原理进行了详细阐述，这是一个实测数据和模拟数据通过迭代的方式相互修正以提高数据准确性，降低关键参数不确定性的过程。在本章节中，我们将用两个案例、从两个角度对上述数据融合方法的执行过程进行演示，并验证算法的可行性。第一个案例中的基准建筑来自于美国能源部（DOE）发布的商业建筑典型模型库[115]。这个模型库的开发是为了支持美国建筑相关标准的制定，模型库中包含 16 个典型商业建筑类型（包括：大、中、小型办公楼，独立零售，沿街商铺，大、小型酒店，饭店等），模型参数是基于广泛的调研得到的，因此用该模型库中的模型来代表实际建筑具有较高的可行性。第一个案例将该典型酒店建筑模型在上海气象条件下的能耗作为研究对象，在模拟得到的能耗数据上人为添加异常值和噪声来代替实测数据，然后基于这一人为处理得到的实测数据开展后续的融合过程，下面将人为“加噪”处理的能耗模拟数据称之为“虚拟实测数据”，采用“虚拟实测数据”的原因有几下两点：

（1） 未知关键变量和真实的能耗值是已知的，可用来评估算法的可行性；

（2） 相比于真实的实测数据，“虚拟实测数据”灵活性更强，数据的多少、种类和颗粒度可以自由调节。

第二个案例选取了上海市一栋真实存在的五星级酒店建筑作为研究对象，来探讨本文所提算法在实际工程背景下的应用效果和价值。

4.2 基于“模拟实测数据”的融合算法验证

4.2.1 模型及数据描述

该酒店建筑模型如图 4.1 所示，建筑面积为 $13345m^2$，地下一层，地上六层，分区包括地下室、商业区、大堂、咖啡厅、储藏室、设备间、客房、餐厅、宴会厅等，其模型参数设置如表 8 所示。其 4-11 月份制冷能耗模拟数据和加入异常值、噪声之后的“虚拟实测数据”见图 4.2。加入异常值和噪声的目的是为了使本案例更接近实际情况，验证融合算法的有效性。

在本案例中，已知的参数包括建筑外形参数、空调系统类型、冷热源配置、人均新风量、人员密度、照明功率密度，其他参数未知，其中需要推测的未知关键变量包括：新风渗透率（INFIL）、冷机性能系数（COP）和制冷设定温度（SPC）。

用于推测关键变量的数据为冷机能耗数据。

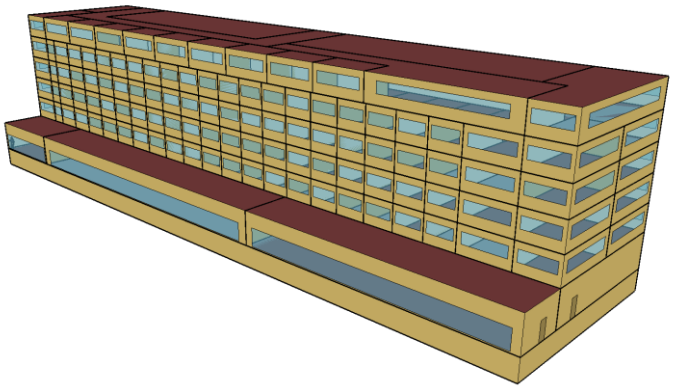


图 4.1 酒店建筑模型几何外形

表 8 酒店建筑模型参数设置

种类	参数	实际模型取值	单位
建筑外形	窗墙比（北）	26	%
	窗墙比（南）	36.7	%
	窗墙比（东）	24.5	%
	窗墙比（西）	24.5	%
	建筑面积	13345	m^2
	层数	7	
	紧凑系数	0.56	
围护结构热工性能	外墙传热系数	0.698	$\mathrm{W}/(m^2K)$
	外墙热容	2000	$\mathrm{J}/(kg\ K)$
	屋顶传热系数	0.228	$\mathrm{W}/(m^2K)$
	窗玻璃传热系数	3.064	$\mathrm{W}/(m^2K)$
	窗玻璃太阳辐射得热系数	0.244	
	外墙太阳辐射吸收系数	0.7	
	屋顶太阳辐射吸收系数	0.7	
空调系统	风系统类型	客房采用风机盘管加新风系统，其他房间采用变风量系统	
	送风温度	12.8	℃
	水系统类型	一级泵定流量二级泵变流量	
	空调冷冻水供水温度	7.22	℃

	冷机 COP	5.7	
	水泵效率	0.9	
	风机	0.6	
使用运行	空调制冷设定温度	24	℃
	空调供热设定温度	21	℃
	照明功率密度	12	W/m^2
	人员密度	0.125	P/m^2
	新风量	0.0094	$m^3/s/P$
	冷风渗透率	0.3	ACH

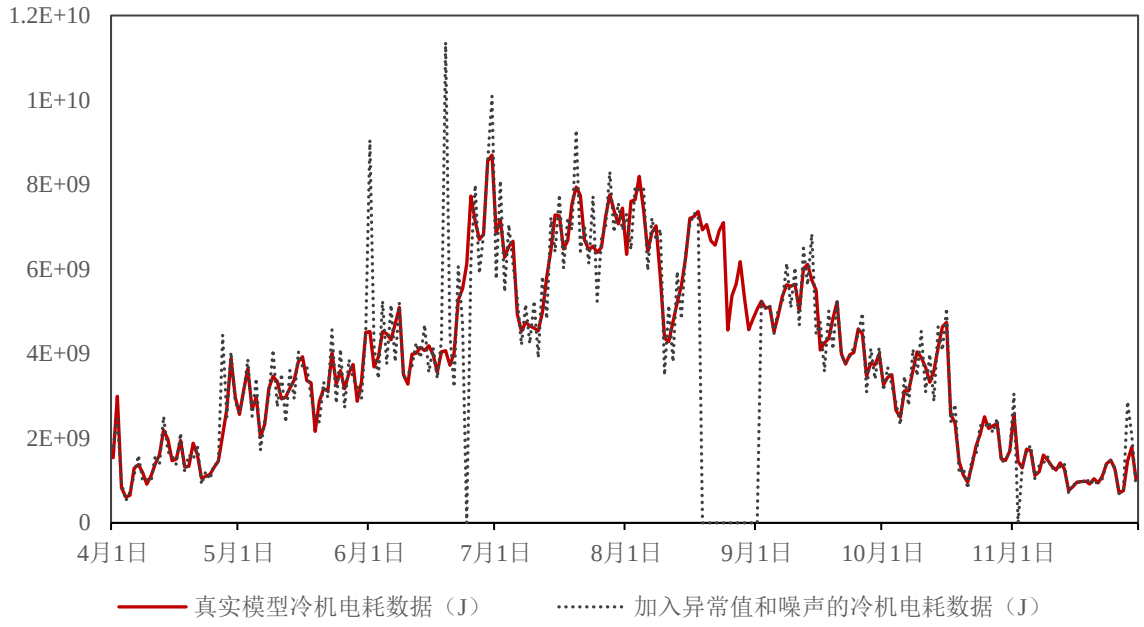


图 4.2 酒店建筑模型冷机模拟数据及“虚拟实测能耗数据”

4.2.2 模拟建筑关键变量推测

根据图 3.1 给出的数据融合算法流程，首先需要对目标建筑建立简化代理模型，然后基于简化代理模型的模拟数据和“虚拟实测数据”进行多次迭代和融合修正。

4.2.2.1 代理模型的建立

这里的代理模型的建立借助第一章中设计的自动建模工具，模型外形和初始参数设置见图 4.3 和表 9 所示，模型的外形结构是自动建模工具根据面积、层数、

体形系数自动匹配生成的。根据 4.2.1 小节中的假设，模型外形参数和空调系统类型、冷热源配置、人均新风量是已知的，因此这些参数与真实模型一致，其余参数则是根据经验选取的，在此基础上计算得到的代理模型的冷机电耗见图 4.4。可以看出，由于部分参数是任意选取的，所以代理模型能耗与带异常值和噪声的“虚拟实测数据”相差较大，“虚拟实测数据”整体较代理模型计算值偏小，并且很多数据突变点。将代理模型能耗与真实模型能耗进行比较，可以发现两组数据虽然整体差异较大，但其变化趋势比较接近，这一现象也验证了上一章节中论述的用“模拟数据”修正“实测数据”的假设。

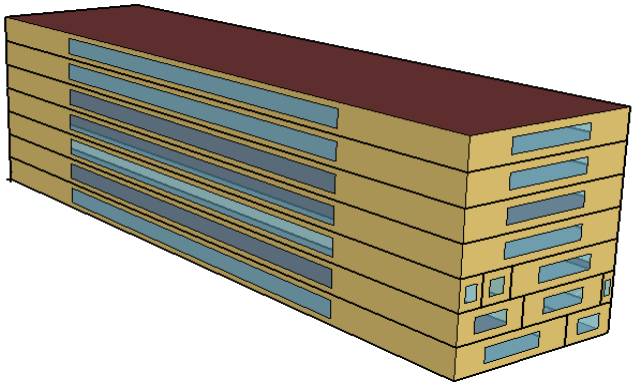


图 4.3 代理模型建筑外形

表 9 真实模型参数设置和代理模型参数设置对比

种类	参数	实际模型取值	代理模型初始取值	单位
建筑外形	窗墙比（北）	26	25	%
	窗墙比（南）	36.7	30	%
	窗墙比（东）	24.5	25	%
	窗墙比（西）	24.5	25	%
	建筑面积	13345	13345	m^2
	层数	地下一层，地上六层	地下一层，地上六层	
	紧凑系数	0.56	0.5	
围护结构热工性能	外墙传热系数	0.698	0.75	$W/(m^2K)$
	外墙热容	2000	1500	$J/(kg\ K)$
	屋顶传热系数	0.228	0.2	$W/(m^2K)$
	窗玻璃传热系数	3.064	3	$W/(m^2K)$
	窗玻璃太阳辐射得热系数	0.244	0.538	
	外墙太阳辐射吸收系数	0.7	0.7	

	屋顶太阳辐射吸收系数	0.7	0.7	
空调系统	空调风系统	客房采用风机盘管加新风系统，其他房间采用变风量系统	客房采用风机盘管加新风系统，其他房间采用变风量系统	
	送风温度	12.8	14	℃
	空调水系统	一级定流量二级泵变流量	一级定流量二级泵变流量	
	冷冻水供水温度	7.22	7	℃
	冷机 COP	5.7	*4.3*	
	水泵效率	0.9	*0.8*	
	风机效率	0.6	*0.8*	
使用运行	空调制冷设定温度	24	*26*	℃
	照明功率密度	12	12	W/m^2
	人员密度	0.125	0.125	P/m^2
	新风量	0.0094	0.0094	$m^3/s/P$
	冷风渗透率	1.2	0.8	ACH

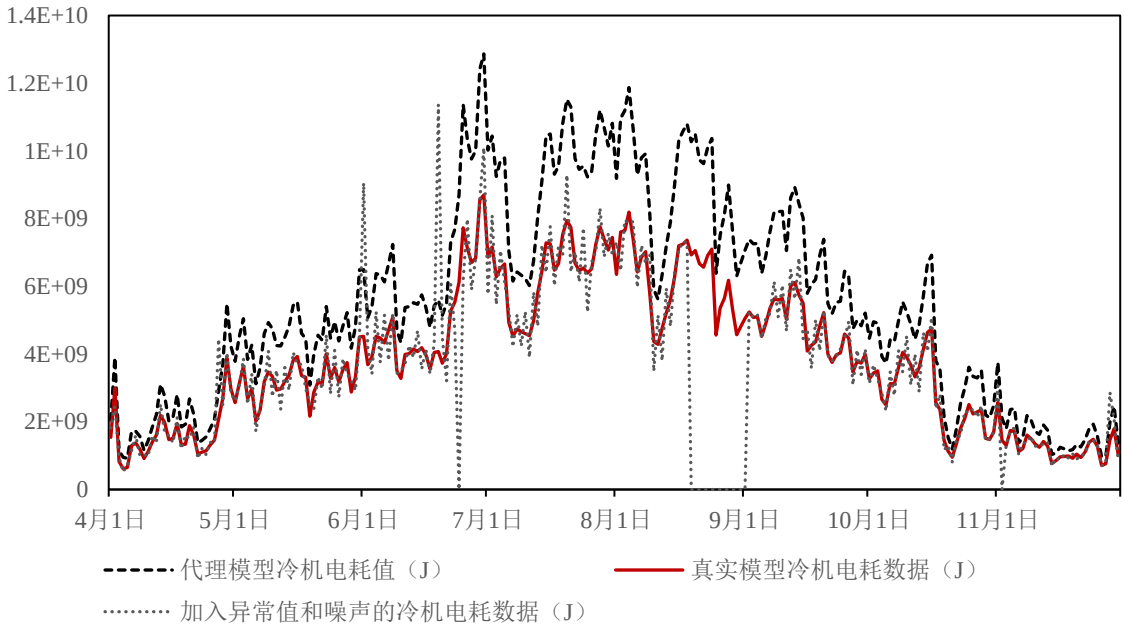
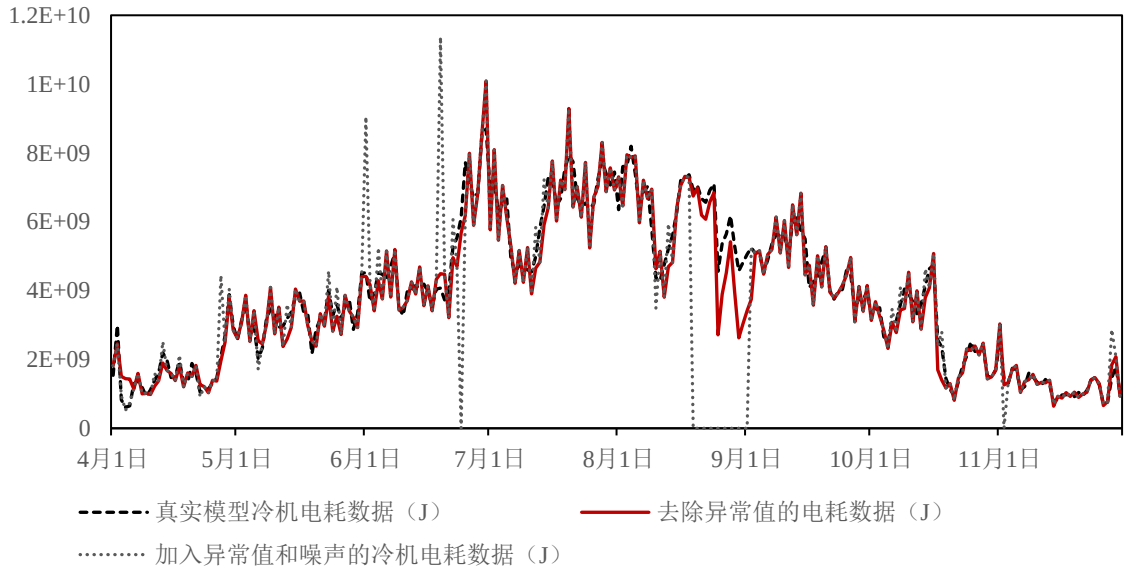


图 4.4 融合处理前三类能耗数据对比

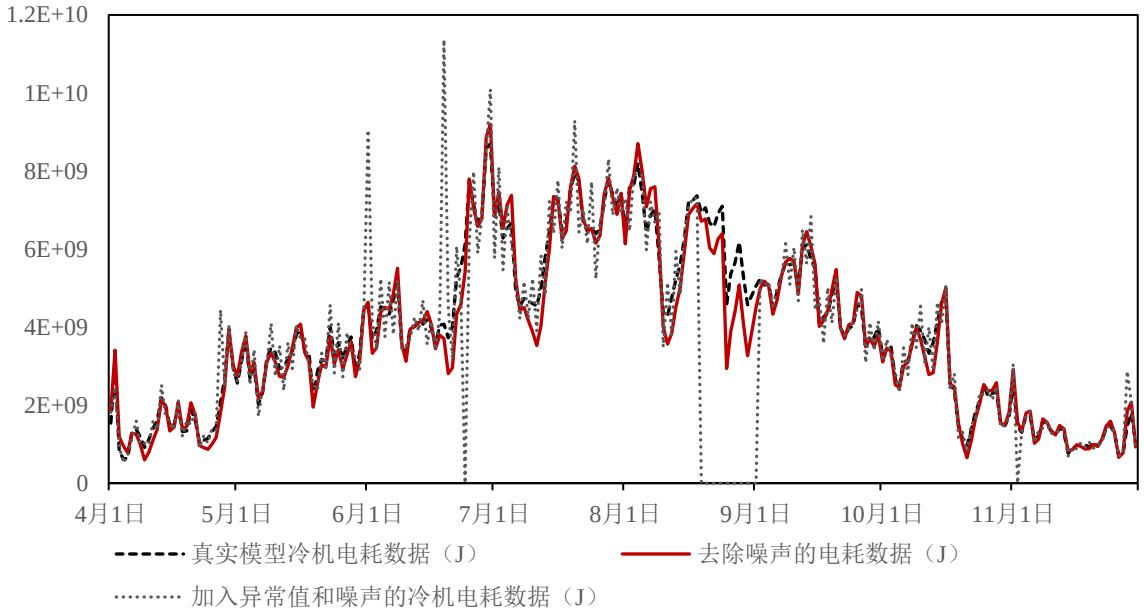
4.2.2.2 I次迭代融合

在第一次数据融合迭代过程中，代理模型的参数是根据经验设置的，其冷机

电耗数据如图 4.4 中的“代理模型冷机电耗数据”所示，可以看出与“真实模型冷机电耗数据”有较大的偏差，但是两者的变化趋势是相同的。下面将基于此数据，根据 3.4、3.5 小节中论述的方法对“模拟实测数据”进行异常值、噪声检测和修正，修正结果如图 4.5 所示。经过“去异常值”修正后，“实测数据”中的过大、过小值均得到了修正，缺失值也得到了填补，经过修正后的数据整体趋势与“真实模型冷机电耗数据”吻合度很高，但是仔细观察可以发现数据不够平滑，也就是存在“噪声”。于是我们进行了第二步的“去噪声”修正，如图 4.5（b）所示，与仅经过“去异常值”的数据进行对比发现，经过“去噪声”修正后，数据整体更加平滑，与“真实模型冷机电耗数据”有了进一步的提高。除个别点外，大部分数据点与真实模型数据相吻合。由此可以证明本课题提出的基于模拟数据对实测数据进行修正的方法具有较好的可行性和实用性。即使在模拟数据与真实数据存在较大差异的情况下，该方法依然可以取得较好的修正效果；当然，若模拟数据与真实数据的差异较小，其修正效果更好，因此我们进行了 II 次迭代融合。在 II 次迭代融合过程中，采用基于 I 次迭代融合得到的关键变量带入代理模型进行计算得到模拟数据，相比于 I 次迭代融合前，此时得到的模拟数据与“真实模型”数据之间的差异较小，对“模拟实测数据”的修正效果更好。



（a）去除异常值



（b）去除噪声

图 4.5 I次迭代“模拟实测数据”修正结果

在完成“虚拟实测数据”的修正工作后，将基于修正后的“实测数据”进行建筑及空调系统关键未知变量的推测。此过程需要进行较为复杂的数据处理和准备工作，处理流程如图 4.6 所示。

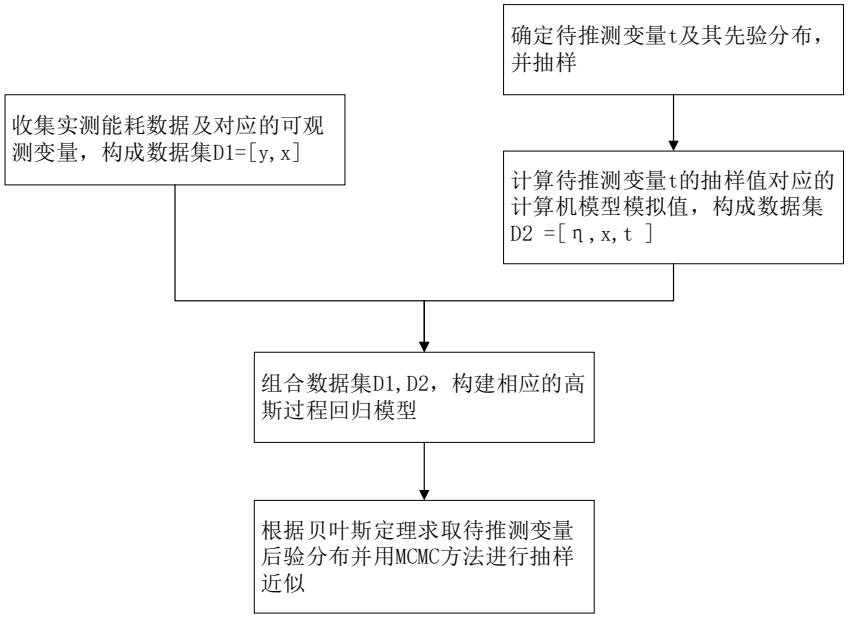
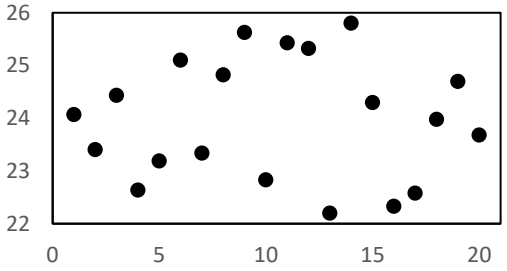


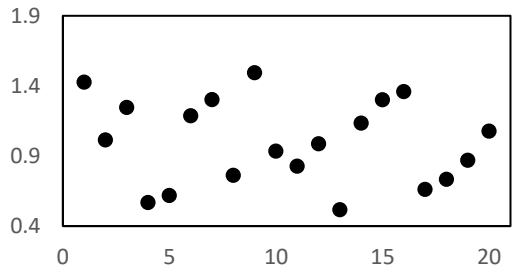
图 4.6 关键变量推测算法执行流程图

首先需要确定待推测关键未知变量的先验分布，并抽样。由于在本案例中我们并未进行调研等活动，对未知变量知之甚少，为了避免过多的主观判断对推测

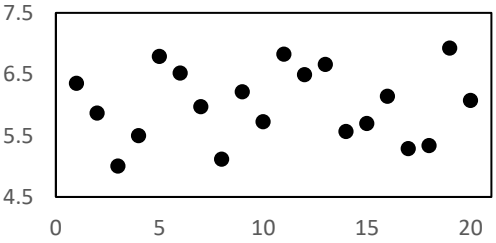
造成误差，我们这里采用弱信息先验，即仅根据工程经验和相关节能设计标准确定每个变量的取值范围，其概率分布均设置成均为分布。然后采用拉丁超立方采样方法对变量进行联合采样，采样数为 20，每个变量的采样值分布如图 4.7 所示。



（a）制冷设定温度（SPC）先验分布采样



（b）冷风渗透率（INFIL）先验分布采样

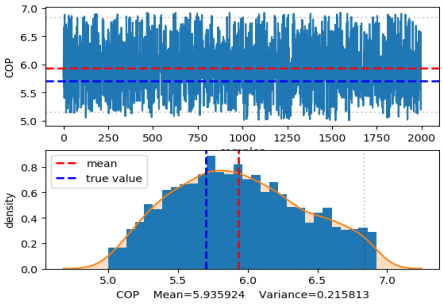


（c）冷机效率（COP）先验分布采样

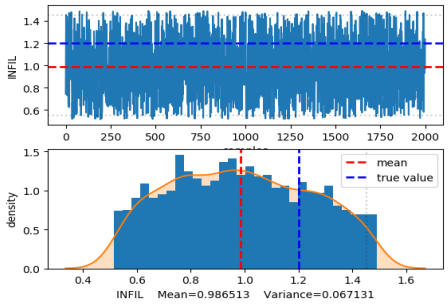
图 4.7 待推测参数先验分布采样

接下来需要针对采样后的参数计算对应的能耗值，进行批量模拟计算，即把上一步采样得到的变量值带入代理模型中进行计算，得到每一组采样数据对应的冷机电耗数据序列，这里的数据序列时间跨度需要与实测数据一致，即从 4 月 1 日到 11 月 30 日。将模拟得到的冷机电耗数据、气象参数、未知变量采样值一一对应组成矩阵结构，这部分数据就是高斯过程回归模型的训练数据。需要说明的是，由于 KOH 法的计算量很大，当训练数据增加时，计算时间将以指数倍增长，因此出于时间因素考虑，我们仅从 4 月 1 日~11 月 30 日这 244 个数据点中均匀抽取 25 个点进行组成训练数据进行后面的计算。

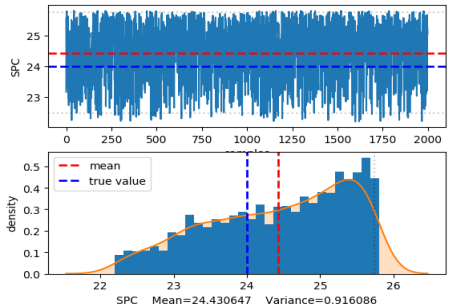
最后我们将由计算机代理模型计算得到的训练数据和经过修正后的实测数据进行组合，根据 KOH 方法对未知变量进行推测。如前所述，由于关键未知变量的后验分布无法用解析的方法直接求出，因而通过 MCMC 采样以分布的形式呈现，3 个关键变量的后验分布如图 4.8 所示，分布期望值即认为是关键变量推测值。可以发现，I次迭代后的关键变量与真实值已较为接近，为了进一步逼近真实值，再进行II次迭代。



（a）冷机 COP 推测值后验分布



（b）冷风渗透率推测值后验分布



（c）制冷设定温度推测值后验分布

图 4.8 I次迭代关键变量推测值后验分布

4.2.2.3 II次迭代融合

将 I 次迭代后的关键关键变量推测值代入代理模型再次进行计算，可以发现代理模型的冷机电耗数据与“真实模型”的数据已十分接近。利用代理模型模拟数据再次对“真实模型”实测数据进行修正，得到修正结果如图 4.9，与上一小节的修正结果相比，修正后数据更进一步接近“真实模型”数据，但改善的幅度并不大。同样地，重复图 4.6 所述执行步骤，基于II次迭代的修正数据进行关键变量推测。与I次迭代的区别是，II次迭代推测过程采用的待推测变量先验分布的取值区间进一步缩窄（表 10）。另外，其选取的实测数据是从整体实测能耗序列值抽取了与I次迭代所采用数据不重合的 25 个数据点，这种做法可以更全面得涵盖整体实测数据的特征。

表 10 未知关键变量先验分布取值范围对比

未知变量名称	符号	I次迭代先验分布	II次迭代先验分布
冷机效率	COP	[5, 7]	[5.5, 6.5]
冷风渗透率	INFIL	[0.5, 1.5]	[0.8, 1.2]

空调制冷设定温度	SPC	[22, 26]	[23, 25]

根据表 11 可以发现，II次迭代后，代理模型的冷机能耗计算值与实测值的偏差相较于I次迭代后结果有轻微上升，图 4.11 的未知变量推测分布图也表明了 II 次跌倒并没有使推测精度上升，因此可以认为此时关键变量值已经收敛了，仅进行 I 次推测即可。收敛结果与真实值之间仍存在差异主要是由于代理模型与真实模型之间并不是完全吻合所导致的。

另外，通过实验可以发现，第一次迭代后得到的数据（包括修正后的实测数据和关键变量的推测值）都已经比较接近实际值，第二次迭代的改进效果不明显甚至有可能使推测结果劣化，但计算量和花费的时间较多，因此在应用时也需要权衡精度与计算量的关系，根据实际情况进行确定时候需要进行二次迭代甚至更多次的迭代计算。

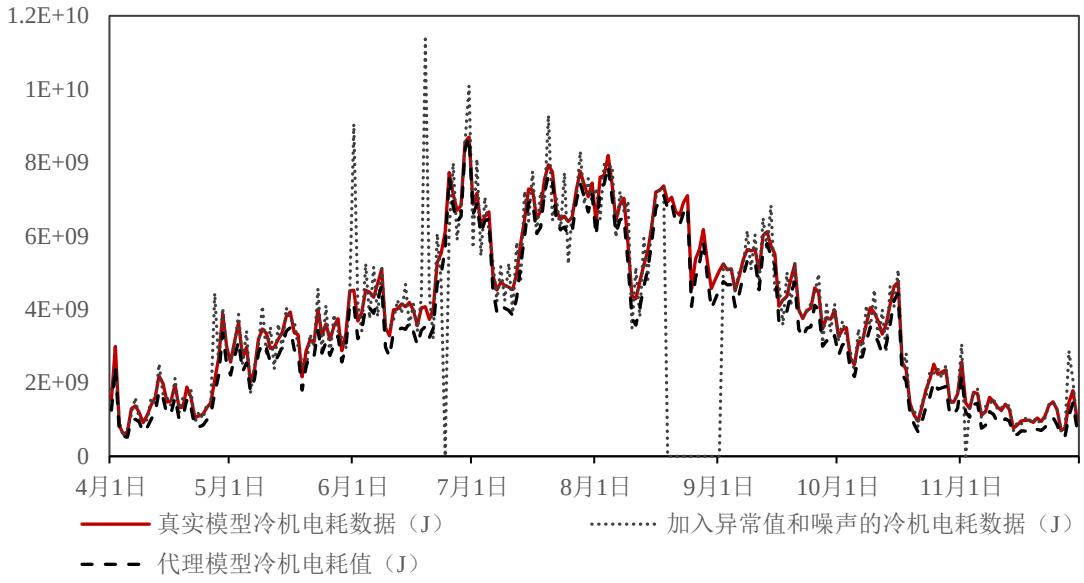
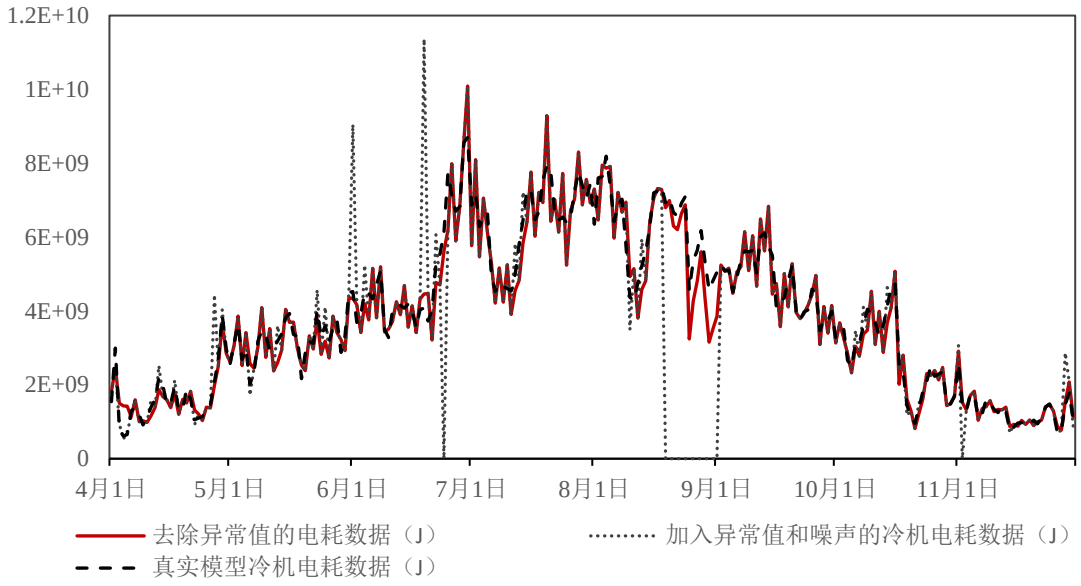
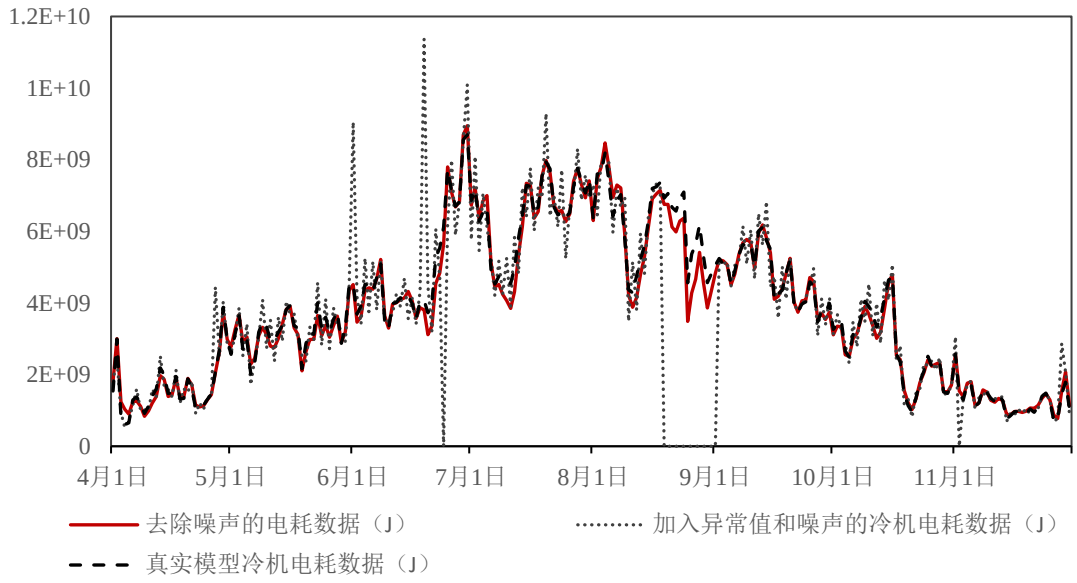


图 4.9 I次迭代后三类能耗数据对比

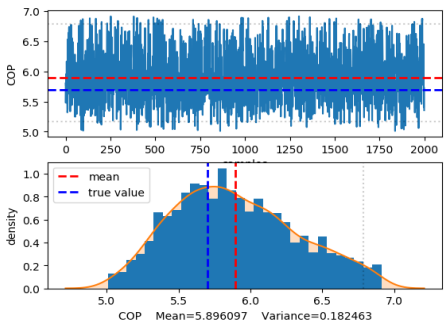


（a）去除异常值

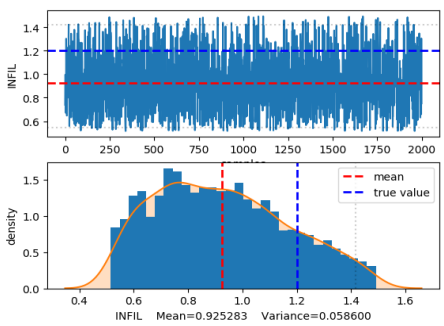


（b）去除噪声

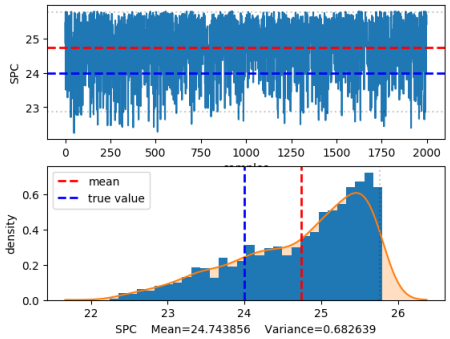
图 4.10 II次迭代“模拟实测数据”修正结果



（a）冷机 COP 推测值后验分布



（b）冷风渗透率推测值后验分布



（c）制冷设定温度推测值后验分布

图 4.11 II次迭代关键变量推测值后验分布

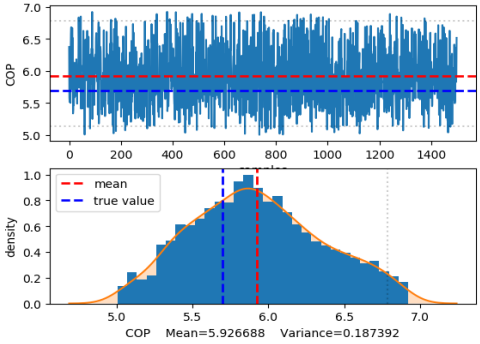
表 11 基于“模拟实测数据”的关键变量推测结果

	I次迭代	II次迭代	真实值
冷机 COP	5.9	5.9	5.7
冷风渗透率（次/小时）	0.99	0.95	1.2
制冷设定温度（℃）	24.4	24.7	24
能耗数据偏差（CV-RMSE）	13.1%	13.5%	—

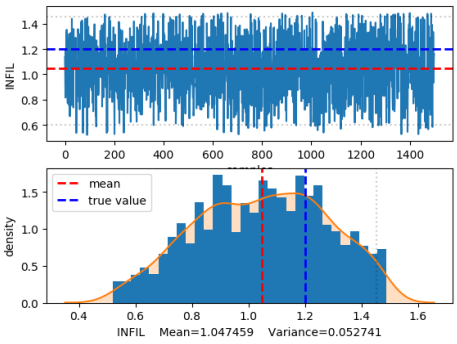
4.2.3 关于算法应用的几点探讨

4.2.3.1 观测数据量对推测结果的影响

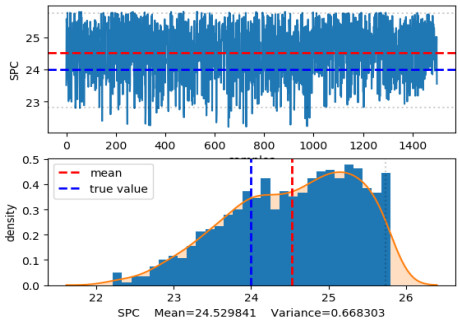
在 KOH 法中观测数据的数量在一定程度上影响了推测结果的准确度，在上一小节中，我们从 244 个观测数据点中均匀抽取了 25 个数据点进行未知变量值推测，为了分析训练数据量对推测结果的影响，这里我们将I次迭代过程的数据点增加至 60 个数据点，推测结果见图 4.12。可以发现，当数据量增加时，未知变量的推测值与真实值之间的偏差变小，越敏感的变量其偏差程度越小，相较于其他两个敏感变量而言，在训练数据量增大时冷风渗透率的准确度增加更大。并且，增加数据量后，后验分布的方差有所减小，说明推测值的不确定性（用后验分布的方差衡量）降低了。但 MCMC 方法对后验分布进行采样非常消耗计算资源，增加训练数据会大幅增加计算时间，在只采用 25 个数据点的情况下，计算时间为 3 小时，增加至 60 个数据点时，计算时间增加至将近 12 个小时，因此在确定训练数据量时需要权衡计算精度与计算时间，一般推荐 20~30 个观测点用于推测。



（a）冷机 COP 推测值后验分布



（b）冷风渗透率推测值后验分布

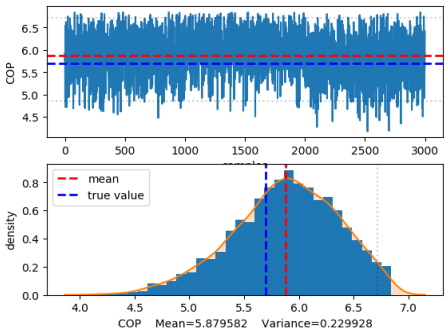


（c）制冷设定温度推测值后验分布

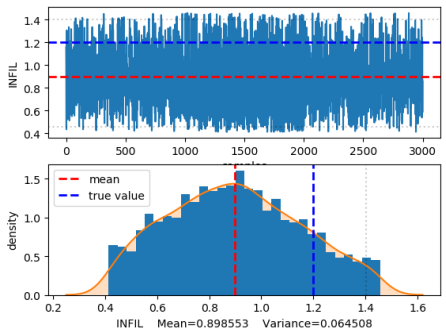
图 4.12 增加观测数据量时未知变量推测结果

4.2.3.2 待推测未知变量数量对推测效果的影响

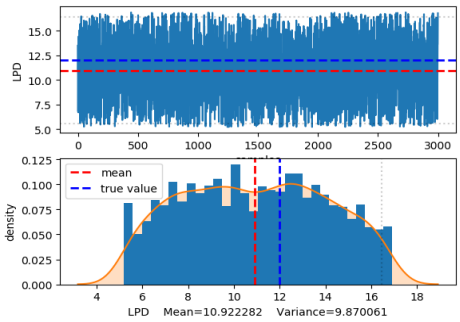
在实际应用过程中常会遇到待推测变量较多，但观测数据量较少的情况。本小节进行了对比实验，将待推测未知变量的数目增加至 5 个，分别为冷机 COP、冷风渗透率、照明功率密度、人员密度、制冷设定温度，观测数据仍然是 25 个数据点，推测结果如图 4.13 所示。相较于仅有 3 个未知变量的情况（图 4.10），未知变量的推测结果与实际值偏差变大了，并且不确定性增加了。因此在观测数据量充足的情况下，应采用与未知关键变量对应的分项设备能耗作为进行推测的观测值。



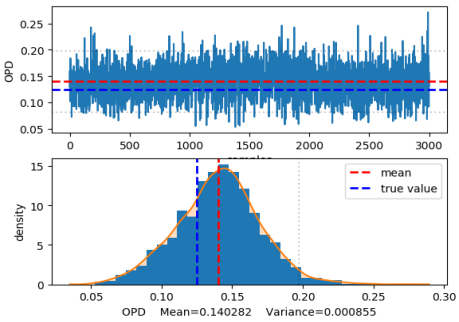
（a）冷机 COP 推测值后验分布



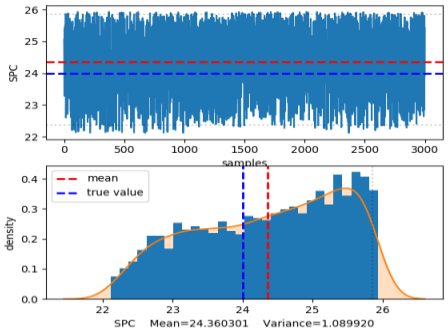
（b）冷风渗漏率推测值后验分布



(c) 照明功率密度推测值后验分布



(d) 人员密度推测值后验分布



(e) 制冷设定温度推测值后验分布

图 4.13 未知变量数目增加时推测结果

4.3 基于实际建筑能耗数据的融合算法验证

为了进一步验证本文所提数据融合及未知变量推测方法在实际工程中的可行性，本小节将基于上海市一五星级酒店实测能耗数据和能源审计报告所提供的建筑基本信息对本文所提算法进行验证。本小节使用的能耗数据包括冷机能耗和水泵能耗数据，据此两组数据进行未知变量的推测。

4.3.1 建筑基本信息描述

该五星级酒店位于上海市徐汇区，总建筑面积为 27500m^2，地下三层、地上 16 层，总客房数 190 间，年平均入住率为 59.8%（图 4.14）。2000 年竣工，2001 年投入使用。该酒店于 2015 年安装分项计量表具，已实现对照明、空调、动力、特殊用电的分项计量。酒店功能分区包括大堂、餐厅、多功能厅、酒吧、咖啡厅、办公及客房等。

外墙采用 240 厚 MU10 多孔砖，M5 混合砂浆砌筑，无保温层；主客房与隔墙采用 120 厚、陶粒砼空心小砌块，以 M5 混合砂浆砌筑，卫生间、电梯前室、

楼梯间及前室采用 120 厚 MU7.5 多孔承重砖，M5 混合砂浆砌筑，其他内墙采用 200 厚陶粒砼空心小砌块，M5 混合砂浆砌筑。外窗、幕墙均为 Low-e 双层中空玻璃，规格为 5mm+12mm+5mm，窗框材料为铝合金。客房、餐厅及多功能厅均有窗帘内遮阳。

图 4.14 酒店建筑外形及外窗

该酒店采用中央空调系统，冷源包括冷水机组包括 3 台特灵冷水机组，单台制冷量为 1050kW，夏季全天 24 小时运行。空调水系统为四管制，二次泵系统，4 台定流量一次泵（三用一备），3 台变流量二次泵（两用一备），4 台冷却水泵（三用一备），3 台冷却塔，设备清单见附录 D。空调风系统根据场所特点及需求的不同进行设置，客房、办公、部分餐厅采用风机盘管加新风系统，其余采用全空气系统。由于本文仅针对夏季工况进行研究，因此不对酒店的制热系统进行详细阐述。

以上是对酒店基本信息的描述，但上述部分信息是较难获取的，比如围护结构构造、设备清单及参数，因此本小节只从中选取了部分容易获取的参数用于建立代理模型，包括：建筑面积、层数及包含的功能区；空调系统拓扑结构（如 4.15 所示）；围护结构参数假设未知，按照上海市公共建筑节能设计规范设置，代理

模型中围护结构的性能参数选取参考相关节能标准。

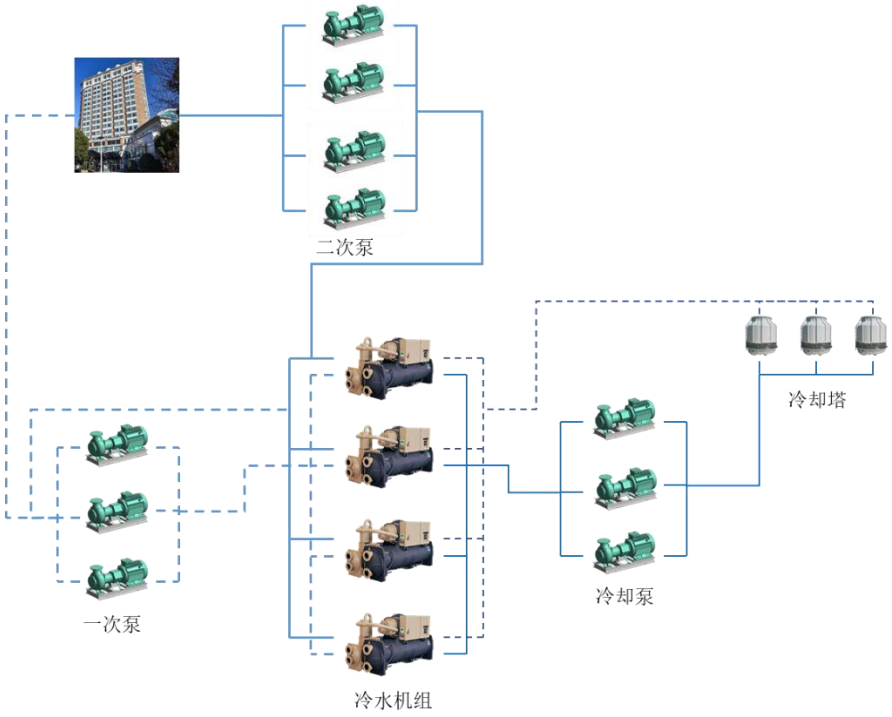


图 4.15 供冷系统图

另外根据该酒店的能源审计报告，我们同时也得以知道另外一些影响空调能耗的参数，包括：

（1） 平均照明功率密度为 6.5W/m^2；

（2） 冷机实测性能系数（COP）为 4.1；

（3） 水泵平均效率为 64.3%；

（4） 冷冻水供回水温差曲线如图 4.16，取其平均为 2.2^oC，存在大流量小温差现象；

（5） 夏季客房平均温度为 23~24^oC，餐厅、厨房平均温度为 23~25^oC。

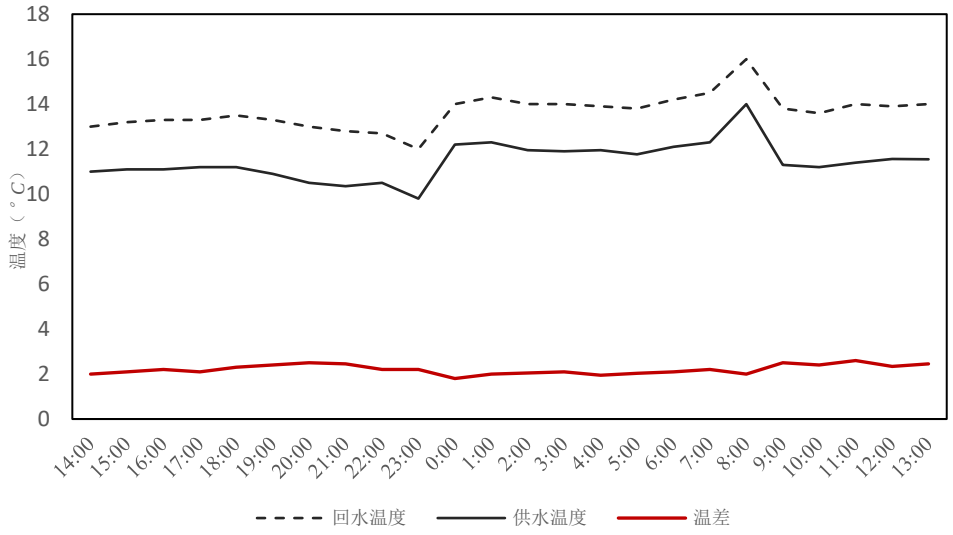


图 4.16 冷冻水供回水温度及温差

上述参数需要通过特定专业测试才能确定，费时费力，在进行大数据分析时，不可能对每一栋建筑都进行类似的专业测评，因此在本案例中将其作为待推测的未知关键变量，利用本文提出的方法，结合分项计量平台获取的能耗数据对其进行推测。若推测结果与实测值相吻合，则说明了本文所提方法的合理性和可行性。

本案例所用到的能耗数据包括冷机电耗数据和水泵电耗数据，时间跨度为2019 年 5 月 1 日~9 月 24 日，颗粒度为逐日，来自酒店安装的分项计量平台，能耗数据见图 4.17。

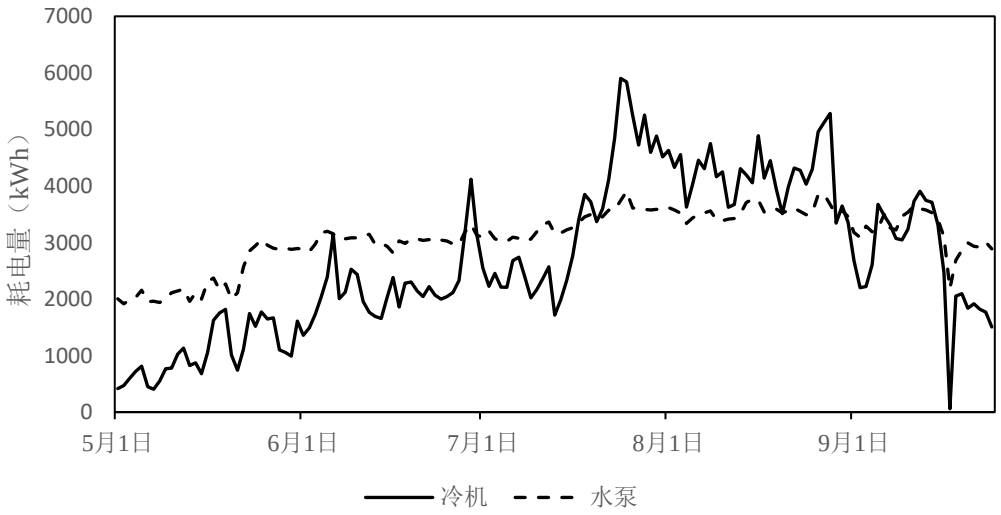


图 4.17 酒店冷机及水泵实测逐日耗电量

4.3.2 实际建筑关键变量推测

与 4.2 小节中阐述的推测过程类似，这里采用同样的步骤对实际建筑的关键变量进行推测，最大的区别在于 4.2 小节中采用的是模拟建筑，因此所有的参数是已知的，可以为建立代理模型提供更多的参考，但对于已经投入运营的实际建筑中由于设计文件的缺失，大部分数据是未知的，只能根据目测（如窗墙比）或相关规范（如围护结构热工参数）进行近似定量取值，但根据第二章的分析，影响建筑空调能耗的关键变量并不多，大多数变量属于非关键变量，这些变量偏离实际值并不会对计算空调能耗产生很大影响，这也正是本文分析的立足点。

首先需要根据已知的建筑基本信息构建代理模型，根据 4.3.1 小节的描述，表 12 罗列总结了实际建筑基本参数和代理模型参数初始值，代理模型的外形结构如图 4.12 所示。

其中需要推测的未知关键变量包括（A）照明功率密度、冷风渗透率冷、冷机 COP、空调制冷设定温度；（B）水泵效率、冷冻水供回水温差。A 类变量利用冷机能耗数据进行推测，B 类数据采用水泵能耗数据进行推测。

表 12 实际建筑基本信息和代理模型参数设置对比

种类	参数	实际建筑取值	代理模型初始取值	单位
建筑外形	窗墙比（北）	未知	40	%
	窗墙比（南）	未知	40	%
	窗墙比（东）	未知	40	%
	窗墙比（西）	未知	40	%
	建筑面积	27500	27500	m^2
	层数	地下 3 层，地上 16 层	地下 3 层，地上 16 层	
	紧凑系数	0.48	0.5	
围护结构热工性能	外墙传热系数	0.67[3]	0.75	$\mathrm{W}/(m^2K)$
	外墙热容	1350[3]	1500	$\mathrm{J}/(kg\ K)$
	屋顶传热系数	0.65[3]	0.7	$\mathrm{W}/(m^2K)$
	窗玻璃传热系数	1.77[3]	2.2	$\mathrm{W}/(m^2K)$
	窗玻璃太阳辐射得热系数	0.6[3]	0.5	
	外墙太阳辐射吸收	未知	0.7	

[3] 由设计文件计算得到，假设未知

	系数			
	屋顶太阳辐射吸收系数	未知	0.7	
空调系统	风系统类型	办公、客房、部分餐厅采用风机盘管加新风系统，其他房间采用全空气定风量系统	办公、客房、部分餐厅采用风机盘管加新风系统，其他房间采用全空气定风量系统	
	送风温度	未知	12.8	℃
	水系统类型	一级定流量二级变流量	一级定流量二级变流量	
	冷冻水供水温度	7	7	℃
	冷机 COP	4.1	5	
	水泵效率	0.643	0.8	
	冷冻水供回水温差	2.2	5	℃
	风机	0.6	0.8	
使用运行	空调制冷设定温度	24	26	℃
	照明功率密度	6.5	10	W/m^2
	人员密度	未知	0.025	P/m^2
	新风量	0.0094	0.0094	$m^3/s/P$
	冷风渗透率	未知	0.5	ACH

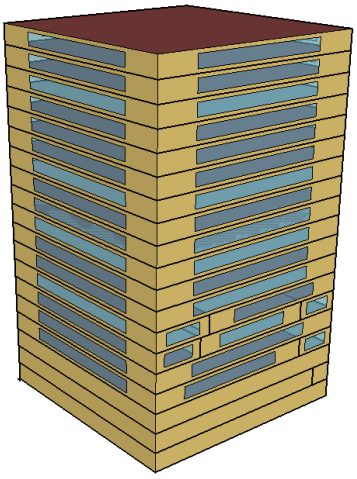


图 4.18 代理模型建筑外形

4.3.2.1 基于冷机电耗数据的 A 类关键变量推测

根据表 12 的参数设置，我们建立了代理能耗模型，其冷机模拟能耗如图 4.19 中的黑色虚线所示，黑色实线为分项计量平台采集的实际冷机电耗数据，两者对比可以发现，其整体变化趋势相似，但幅度有所偏差，实测数据中存在少数异常点。首先我们根据 3.4 小节所述方法对实测数据进行了修正，修正结果为图 4.19 中的红色虚线。为了减少计算量，对修正后的冷机电耗数据均匀采样 25 个数据点来作为推测关键变量的后验数据，两次推测结果如图 4.20-4.21 及表 13 所示，除了冷风渗透率的实测值未知，其余三个变量的推测值与实测值比较接近，验证了本课题所提数据融合算法的有效性。

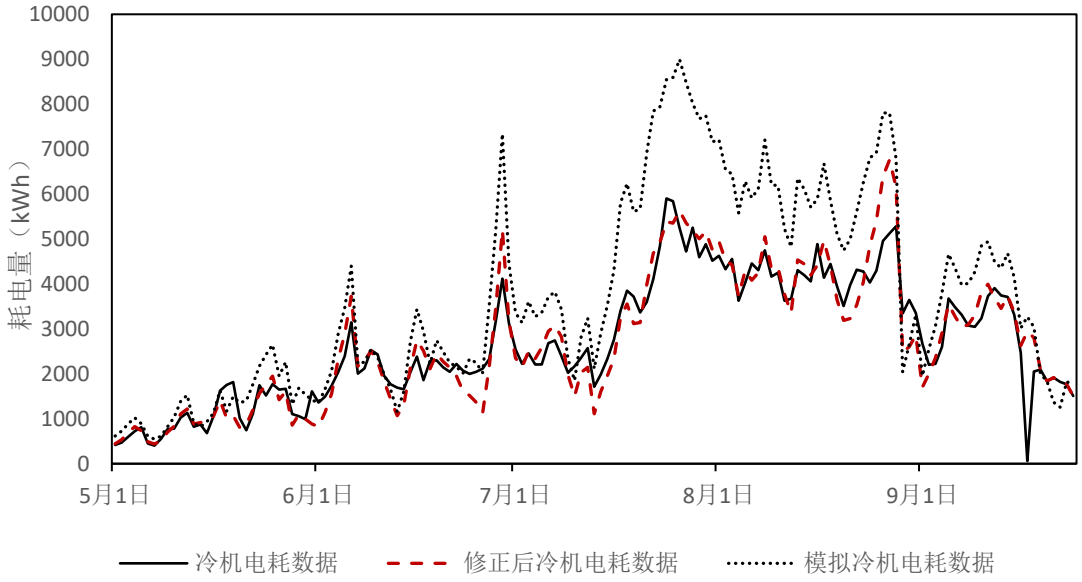
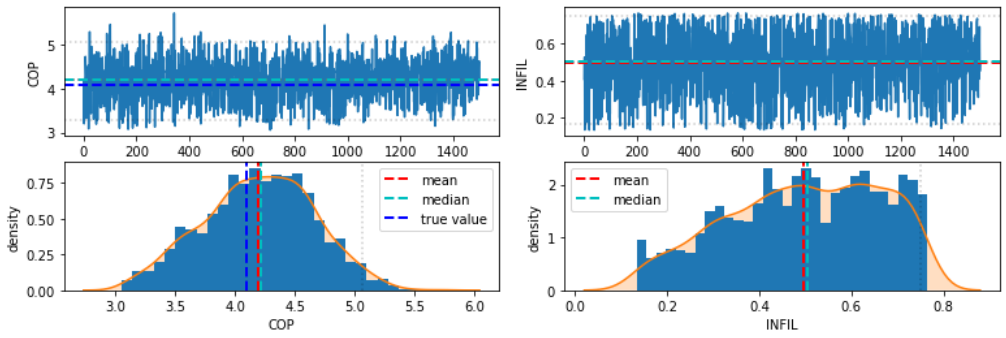
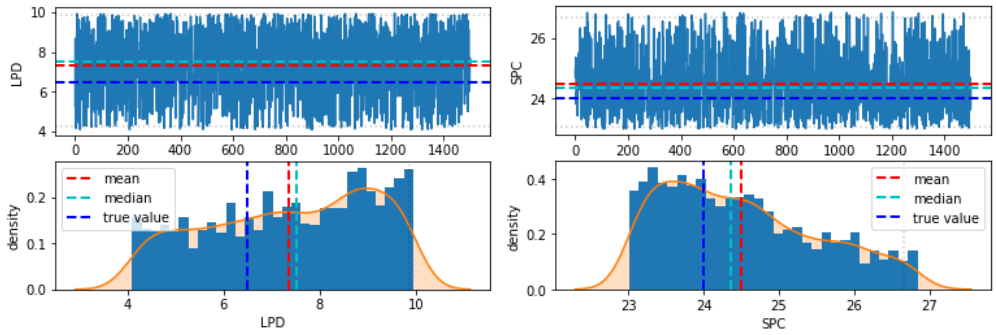


图 4.19 冷机实测电耗初始值及处理后数据

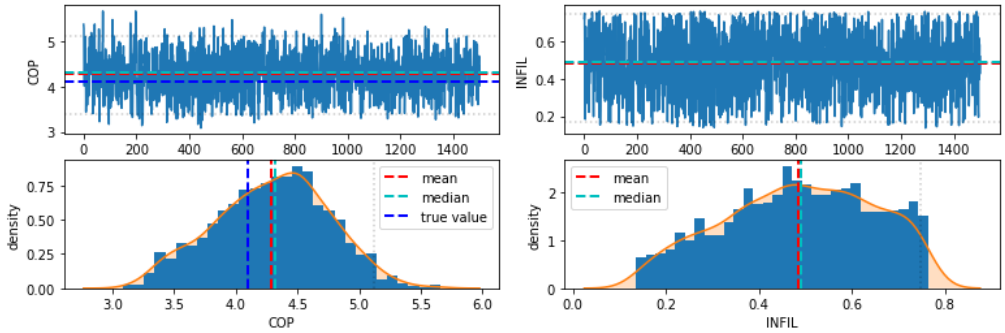


（a）冷机 COP　　（b）冷风渗透率

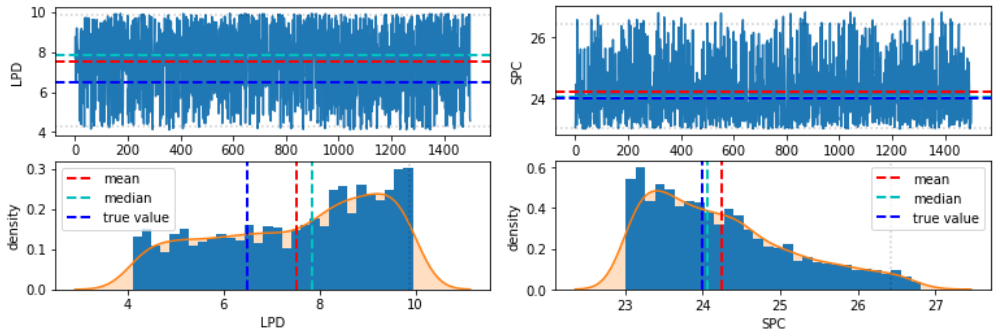


（c）照明功率密度　　（d）制冷设定温度

图 4.20 I次迭代关键变量推测值后验分布



（a）冷机 COP　　（b）冷风渗透率



（c）照明功率密度　　（d）制冷设定温度

图 4.21 II次迭代关键变量推测值后验分布

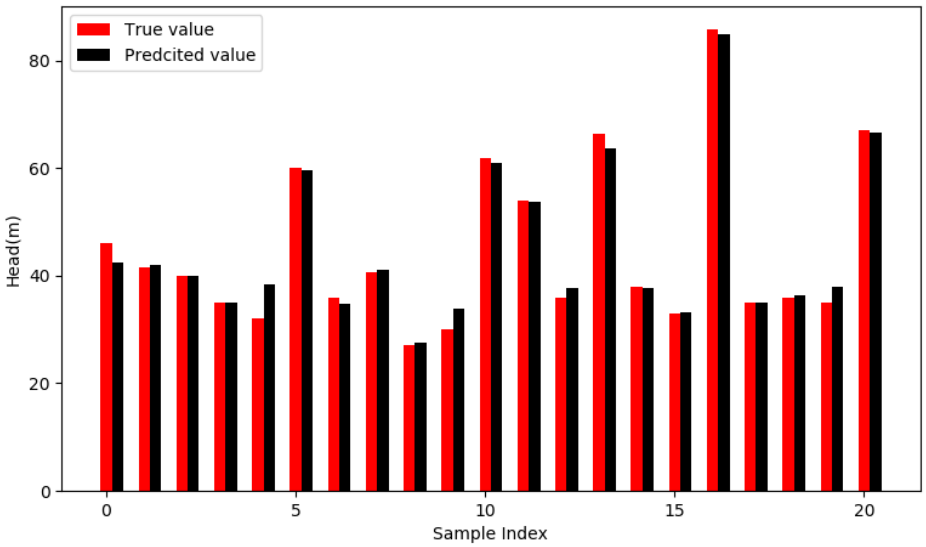
表 13 实际建筑数据融合结果

	I次迭代	II次迭代	真实值
冷机 COP	4.2	4.3	4.1
冷风渗透率（次/小时）	0.49	0.48	—
照明功率密度（W/m^2）	7.4	7.3	6.5

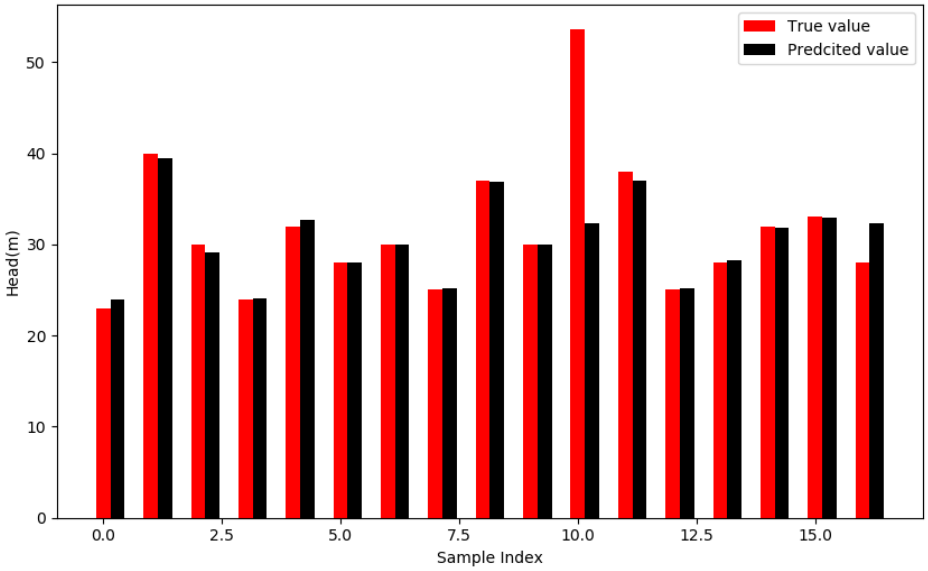
制冷设定温度（℃）	24.5	24.2	24
能耗数据偏差（CV-RMSE）	11.5%	10.9%	—

4.3.2.2 基于输配系统及冷却塔电耗数据的 B 类关键变量推测

在建筑空调系统中，风机/水泵的运行能耗主要有以下几个参数决定：额定风量/水流量、压头、效率。对于既有建筑，在缺乏设备样本数据的情况下，风机/水泵流量可根据设计工况的空调负荷进行估算。并且在建筑外形布局确定的情况下，风量/水量及压头是基本确定的。这里本课题在建立代理模型时，水泵扬程可根据建筑形态和高度、系统类型进行估算，本文利用建筑面积、楼宇地上地下层数及空调水系统类型（一次泵/二次泵）作为数据特征，水泵扬程作为数据标签进行训练得到了用于预测建筑水泵扬程的黑箱模型，当数据缺失时可用此黑箱模型进行水泵扬程的估算。算法可以分别估算冷冻水泵和冷却水泵的扬程。对于二次泵系统，算法计算得到的是总扬程，通常一级泵用于克服机房内管道阻力，这部分阻相对稳定，本文一般取 15mH_2O，剩下的即为二级泵扬程。图 4.22 为估算结果与实际数据的对比图。图 4.23 为 B 类关键变量的推测结果，其中水泵效率的推测值为 0.7，实测值 0.64，供回水温差推测值为 2.1℃，实测值为 2.2℃，较为接近。

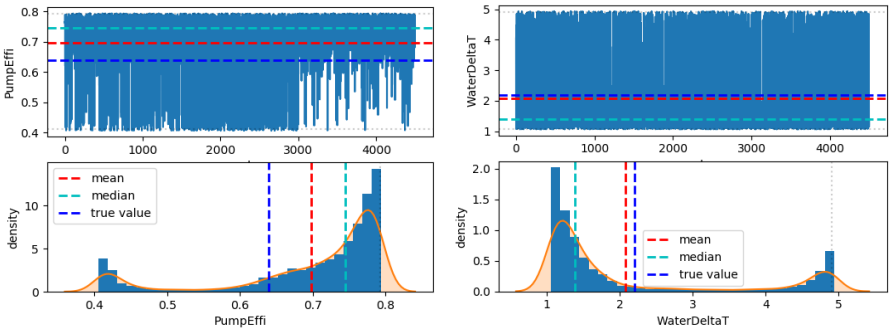


（a）冷冻水泵



（b）冷却水泵

图 4.22 水泵杨程估算结果对比



（a）水泵效率　　（b）供回水温差

图 4.23 B 类关键变量推测值后验分布

4.4 本章小结

本章节是针对第三章数据融合算法的执行操作过程进行了详细阐述，并通过两个案例，从两个角度对其进行了可行性验证，考虑到部分关键变量在实际建筑中很难获取，本章节分别以模拟建筑和实测建筑为研究对象进行案例分析。第一个案例的研究对象是经过校验的模型，因此其可以反应建筑的实际情况。基于模型进行案例分析的原因是，模型十分完备的建筑用能信息画像，每一个关键变量的真实值都可以获取得到，因此可以用来验证关键变量推测的准确性。另外，本文在模型的模拟数据中人为增加了异常值和数据噪声得到“虚拟实测数据”，用

来验证数据修正算法的可行性。第二个案例的研究对象是进行了深度能源审计的上海一五星级酒店，其用能信息画像较为完备，用来验证数据融合算法在真实场景中的融合效果。两个案例的验证结果表明，本课题提出的数据融合算法能很好地修正实测数据，并推测出关键变量的值。需要说明的是，在关键变量确实较为严重的情况下，需要更多实测数据作为观测值进行推测，并且推测的不确定性会增加。因此推荐优先采用分项计量数据对各个分项或设备对应的关键变量分别进行推测。

第5章 建筑混合能耗预测模型的建立

5.1 概述

如前所述，预测建筑空调能耗主要有两种方法：基于白箱模型的能耗模拟方法和基于黑箱模型的机器学习方法。白箱模型基于物理规则建立输入输出变量之间的关联公式，黑箱模型则是根据输入输出数据构建它们之间的统计分布规律。能耗白箱模型的校验，和黑箱模型的训练都需要数据来支撑。对于既有建筑，可以利用其历史运行数据预测未来能耗，有研究表明在历史数据质量较好的前提下，预测既有建筑的未来能耗可以有较高的精度。这是目前大多数能耗预测数据驱动模型的研究场景。但是对于新建建筑或是没有记录历史运行数据的建筑，对其进行能耗预测是比较棘手的问题。即便是对于设计参数很完备的建筑，采用白箱模型法建立能耗模型进行预测也不能得到令人满意的结果，因为设计参数往往与运行参数偏差较大。这里引入另一种研究较少的能耗预测驱动模型，对象是没有历史运行或能耗数据的建筑，采用同类其他建筑的能耗进行“迁移”预测。众所周知，数据驱动模型最重要的是特征工程。所谓特征是指对被观察对象有重大影响的可观测变量或特性，它可以解释观测变量的变化。对于迁移预测模型，训练数据集包括不同建筑的能耗数据，因此模型特征需要包括能够表征与建筑及其机电系统相关的，对空调能耗产生重大影响的特性，也就是本文前面几个章节得到的建筑信息关键变量。通过建立建筑信息关键变量与能耗值之间的映射关系，就可以对新建筑进行预测。

为了解决这个问题，本章节建立了建筑信息及能耗数据库，该数据库包含了影响空调能耗的关键变量及对应的能耗值，以关键变量作为输入特征、能耗作为输出预测值建立黑箱模型，该模型训练完成后可在没有历史数据的情况下进行建筑能耗的“迁移”预测，解决新建建筑的能耗预测问题。在前面三个章节中，详细介绍了建筑空调能耗关键变量的选择方法以及关键变量的推测填补方法，经过这两个步骤，我们可以将一栋复杂的建筑去除冗余信息，仅用少数几个特征变量去描述其能耗特性，大大简化了建筑能耗分析的复杂程度，为本章节建立能耗预测模型进行了铺垫。

另一方面，很多公共建筑开展了深度能源审计来建立建筑画像，但是对于大数据分析而言，单栋建筑的价值很有限，我们不可能根据某一栋建筑的特征推测出其他建筑的特征，但是当数据规模足够大之后，就可以通过数据分析从中发现某些共有特征，这也是建立建筑信息及能耗数据库的意义。

5.2 混合能耗模型特征及算法的确定

目前对基于数据驱动的能耗预测模型研究绝大部分是针对单个建筑开展的，并且要求有充足的历史能耗数据用以训练机器学习模型，这种方法建立的模型不具备通用性。本课题提出的混合能耗模型解决了这一问题，通过融合不同来源的数据，挖掘影响建筑空调能耗的关键性因素，搭建通用性的能耗预测模型。需要说明的是，与机器学习领域通常意义上的混合模型不同，本文中混合能耗模型的“混合”是指用以训练机器学习模型的数据来源不同，经过融合以后形成了建筑信息及能耗数据库，在此基础上训练得到的模型为混合能耗模型。训练机器学习模型两个主要内容是特征工程和模型训练（包括算法选择和模型调参），本小节将分别就这两部分进行阐述。

5.2.1 输入特征的确定

机器学习通过数据建立了输入输出变量之间的映射关系。特征工程是建立机器学习模型最重要的步骤，构建简洁但信息含量丰富的特征组能大大提高模型的性能。本课题的研究对象是建筑空调能耗，其影响因素主要有三大类：

1) 室外气象参数
2) 室内人员活动特征
3) 建筑围护结构特性及机电系统特性和运行情况

因此混合模型的输入特征包括上述三类参数。下面分别对以下三类参数所构成基础特征及衍生特征进行详细阐述。

5.2.1.1 室外气象参数相关特征

干球温度、相对湿度、露球温度、太阳辐射、风速等室外气象参数是影响建筑暖通空调能耗的主要因素。几乎所有的能耗预测研究都使用天气参数作为建筑物暖通空调能耗预测的输入特征。部分学者使用直接观测的天气参数，而另一些则使用经过处理的天气参数，例如冷却度数日(CDD)和加热度数日(HDD)。度日数是衡量冷热需求的一个简单方法，它结合了平衡温度的信息，在一定程度上提高了模型性能[119]。文献[119]表明干球温度和相对湿度是所有气象参数中对能耗变化影响最显著的变量，因此本课题采用这两个变量作为输入特征。

另一方面，由于建筑热惯性的影响，室内环境的变化往往滞后于室外环境的变化。因此，一些研究采用预测目标时间之前的气象参数作为输入特征。确定滞后时间步长需要计算偏自相关系数（PACF）。但对于时间滞后较大的情况(如 6 小时)，模型中会包含太多的前若干时刻步长的气象参数(4*6=24，如果使用 4 个气象参数作为输入特征)，可能会造成维数过大。本课题引入经过平滑处理的气象

参数作为模型输入特征，以平衡滞后和特征维数的冲突。光滑方法是 Savitzky-Golay 平滑滤波器[120]。Savitzky-Golay 滤波器是用于平滑数字信号的常用手段，可以消除高频波动，保持数据趋势和变化。平滑过程称为卷积，通过线性最小二乘法[121]将相邻数据点的连续子集用多项式拟合。此外，本课题还计算了平滑后的温度数据的第一次和第二次差分作为输入特征来表征温度变化。

5.2.1.2 室内人员活动相关特征

室内人员的数量也是影响建筑空调能耗的主要因素之一。但是，但在大多数情况下，精确的人流数据难以获取。因此，代表人员在室率的特征通常用分类特征表示，本课题称之为“时间标签”，如一天的第几小时(表示 1-24)、一星期的第几天(表示 1-7)、一个月份的第几天(表示 1-31)。此外，不同日类型(工作日、周末、假日等)的人员活动特征是不同的，对于大多数建筑类型，不同日类型的能耗变化信息也可以用 0-1 的分类特征来表示。

另外，从历史能耗本身的使用规律也可以提取出一些能表征人员活动特点的特征，本课题提出了周期因子和统计量因子的这两个特征。

（1） 周期因子

人类活动趋向于每周重复，使建筑能耗遵循相似的周期。此外，同一日类型的建筑能耗特性相似。例如，前一个星期一的能量消耗可以用来预测下一个星期一的能量消耗。在本文中，我们利用同类建筑的历史能耗数据构建了周期因子来表征每个工作日的能源消耗。，计算公式如下：

$$r_{i,j} = \frac{\bar{e}_{i,j,}}{\bar{e}} ,(i = 1,\dots,7,\ j = 1,\dots,12) \tag{5-1}$$

式中，$r_{i,j}$表示第 j个月第i个工作日对应的周期因子，$\bar{e}_{i,j}$表示在预测日期之前第 j个月第i个工作日的平均能耗，$\bar{e}$为日均能耗。

（2） 统计因子

统计因子从同类建筑历史能耗数据集中提取了均值、中位数、最大值、最小值、偏度和标准差等统计信息。计算公式为：

$$t_{i,j} = T(Y_{i,j}) \tag{5-2}$$

式中，$t_{i,j}$表示第 j个月第i个工作日对应的统计因子，T代表了均值、中位数、最大值、最小值、偏度和标准差等统计量的计算公式，$Y_{i,j}$表示在预测日期之前对应月份对应工作日的历史数据集。

5.2.1.3 建筑围护结构及设备系统相关特征

这类特征变量即为本课题第二~四章节所确定的关键变量。本课题研究的输出对象为制冷能耗，包括了冷机、水泵、空调箱、冷却塔四类设备，根据第二章的研究结论，不同的设备能耗对应的关键变量是不同的，针对不同的设备能耗需要分别建立能耗模型，每个能耗模型对应的输入特征也不尽相同，结合前面章节的研究结果，不同设备的能耗模型对应的输入参数总结如表 14。

表 14 混合能耗模型特征表

关键变量	预测目标				
	总能耗	冷机能耗	水泵能耗	空调箱能耗	冷却塔能耗
建筑面积	✓	✓	✓	✓	✓
入住率	✓	✓	✓	✓	✓
地上层数	✓	✓	✓	✓	✓
地下层数	✓	✓	✓	✓	✓
冷风渗透率	✓	✓	✓	✓	✓
照明功率密度	✓	✓	✓	✓	✓
人员密度	✓	✓	✓	✓	✓
空调设定温度	✓	✓	✓	✓	✓
体型系数	✓	✓	✓	✓	✓
冷冻水供回水温差	✓		✓		
冷却塔填料堵塞率	✓				✓
冷机 COP	✓	✓			
风机效率	✓			✓	
水泵效率	✓		✓		
风系统过滤器堵塞率	✓			✓	
风系统类型	✓	✓	✓	✓	✓
水系统类型	✓	✓	✓		✓

另外，考虑到模拟数据存在一定偏差，我们希望模型更倚重实测数据，即能反应出建筑实际运行能耗的特征，因此设置了数据训练权重，实测数据的权重大，模拟数据的权重小，并且随着实测数据的不断充实，模拟数据的权重将被不断降低直至不再被采用。

5.2.2 算法选择

目前常用的机器学习算法有很多种，但并没有某一个算法在任何场景下都是性能最好的，即不同的算法有各自的适用场景。在工程实践及机器学习比赛中，组合树模型（例如随机森林、LightGBM 等）表现很好，但在训练数据集较小的情况下，简单算法（例如线性回归）的性能反而但更好，复杂算法容易过拟合。因此本文选取了 8 种常用的机器学习算法，包含了结构简单和复杂的常用算法，同时进行测试，选取最适用于本文所述场景的算法。

（1） 多元线性回归

多元线性回归试图通过对观测数据拟合一个线性方程来建立若干变量之间的线性关系。拟合回归线最常用的方法是最小二乘法。这种方法通过最小化每个观测数据点到线垂直偏差的平方和来计算拟合线性方程。线性回归是应用最广泛的回归分析算法。这是因为线性依赖于其未知参数的模型比非线性依赖于其参数的模型更容易拟合，并且因为所得到的估计量的统计特性更容易确定。但是线性回归无法拟合复杂的非线性物理过程。

（2） Lasso 回归

Lasso 回归是一种使用压缩（shrinkage）估计的线性回归，在线性回归的损失函数基础上加上惩罚项（L1 正则项）。压缩是指数据向某一中心点(如平均值)收缩。Lasso 回归倾向得到更简单、稀疏的模型(即参数较少的模型)。相比于普通线性回归，Lasso 不易过拟合，非常适合拟合特征中存在多重共线性的模型，也经常用于自动化变量选择或参数消除。类似的还有 Ridge 回归和 ElasticNet，都是普通线性回归的的扩展，附加了一个惩罚参数，旨在最小化模型复杂性或减少最终模型中使用的特征数量。

（3） K 近邻回归

K 近邻算法是机器学习算法集中比较简单易懂的，但在某些问题上被证明十分有效。K 近邻算法既可以用于分类问题，也可以用于回归问题。该算法使用“特征相似度”来预测未知数据点的值，即根据该点与训练集中的点的距离来判断相似程度。计算距离有很多种方法，其中最常见的有欧几里德距离、曼哈顿距离和汉明距离（用于计算分类特征）。在完成测量训练集中各点与新观测点之间的距离后，下一步需要选择最近的 K 个点，过大或是过小的 K 值都会造成模型性能劣化，K 的最优值一般根据交叉验证法得到。

（4） 支持向量机回归（SVR）

与支持向量机分类算法类似，SVR 同样由两条间隔线，如图 5.1 所示，红色直线为拟合线，两旁的黑色直线为间隔线，SVR 要求尽可能多的预测点落在两条

间隔线之间，同时允许一部分点在区间之外，其偏差用ξ表示，但这部分偏差要求尽可能小。与普通线性回归不同的是，SVR 的目标函数是使系数最小化，更具体地说，是系数向量的 L2 范数最小化，而不是平方误差。因此 SVR 的目标函数为:

$$min(||w_i||^2 + C\sum_{i=1}^{n}\xi_i) \quad (5\text{-}1)$$

约束条件为：

$$|y_i - w_i x_i| \le \varepsilon + |\xi_i| \quad (5\text{-}2)$$

其中 C 为弹性因子，是一个可以调整的超参数。随着 C 增加，算法容忍度增加，接受更多的点落在间隔线之外。当 C 接近 0 时，容忍度也接近 0，此时模型由于弹性不够易过拟合。

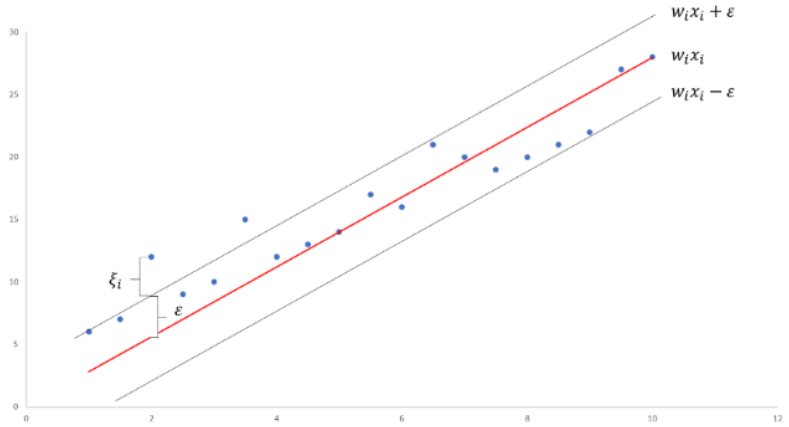


图 5.1 支持向量机回归算法原理示意图

（5） 回归决策树

决策树算法是一种用于分类和回归的非参数监督学习方法。其目标是通过学习从数据特征推导出的简单决策规则，创建一个预测目标变量值的模型。决策树算法以树结构的形式建立回归或分类模型。该算法以从上到下的形式将数据集分解成越来越小的子集，不断地生成新的子树结构，最终形成一个具有决策节点和叶节点的树。一个决策节点有两个或多个分支，每个分支代表被测试属性的值。叶节点表示对数值目标的决定。相比于其他机器学习算法，可以同时处理分类型数据和数值型数据；决策树有较强的可解释性，并且不需要进行数据预处理（标准化、归一化等）。但是决策树算法容易过拟合，而且不稳定，当数据集中存在噪声时会学习到完全不一样的树结构，这个问题可以通过组合树模型解决。

（6） 随机森林

在机器学习算法集中，存在一类集成学习算法，这种算法根据集成学习理论由若干弱学习器（也称为基学习器）组合而成，具有比弱学习器更低的偏差（bias）

和方差（variance）。目前有三种常用的集成方法：

- 装袋法（Bagging）：该方法常考虑采用同质弱学习器，相互独立并行学习，最后并根据某种确定性的平均过程将弱学习器的进行组合（例如取均值）。
- 提升法（Boosting）：该方法同样采用同质的若学习器，与装袋法不同的是，提升法采用串联迭代的方法组合多个弱学习器。每个弱学习器着重训练在上一个弱学习器中表现不好的样本点。
- 堆叠法（Stacking）：堆叠法与装袋法、提升法主要有两点不同。首先堆叠法通常考虑异质的弱学习器(不同的学习算法)，其次，堆叠法的组合方式不同，它采用一个元模型将弱学习器的结果进行组合。例如，我们可以采用线性回归、支持向量回归和决策树回归作为弱学习器，然后将它们的预测结果作为另一个元模型（例如神经网络算法）输入，该元模型的输出即为最终的预测预测结果。

随机森林回归是一种使用集成学习方法进行回归的监督学习算法。集成学习方法是一种将多个机器学习算法的预测结合起来，从而做出比单一模型更准确预测的技术。随机森林是基于决策树算法建立的组合树模型。其预测结果是基于装袋法综合了所有决策树的结果得到的，构建过程如下：首先利用有放回方法从原始训练集中随机抽取 n 次样本，构建 n 个决策树；对于单个决策树模型，每次分裂时根据信息增益、信息增益比或是基尼指数选择最好的特征进行分裂，直到该节点的所有训练样本都属于同一类；将生成的多颗决策树组成随机森林。对于分类问题，按照多棵树分类器投票决定最终分类结果；对于回归问题，由多颗树预测值的均值决定最终预测结果。随机森林回归模型功能强大且准确，在许多问题上表现出色，包括具有非线性关系的特征。缺点容易发生过拟合。

（7） LightGBM

LightGBM 的全称是 Light Gradient Boosted Machine，是最初由微软开发的分布式梯度提升框架，是基于梯度提升数算法(Gradient Boosting Decision Tree，GBDT) 改进，在工程项目和专业比赛上都取得了很不错的成绩。GBDT 是机器学习中一个十分有效的模型，基于提升法设计的集成模型，该模型具有训练效果好、不易过拟合等优点。但 GBDT 在特征维数高、数据量大的情况下，效率和可伸缩性仍然不能令人满意。一个主要原因是对于每个特征，该算法需要扫描所有的数据实例来估计所有可能的分割点的信息增益，这是非常耗时的。LightGBM 在不损害 GBDT 精度的基础上增加了单边梯度采样（Gradient-based One-Side Sampling, GOSS）和互斥稀疏特征绑定两项技术（Exclusive Feature Bundling, EFB）。使用单边梯度采样可以减少大量只具有小梯度的数据实例，使用互斥稀疏特征绑定可以将许多互斥的特征绑定为一个特征，达到降维的目的。因此

LightGBM 具有精度高，计算速度快的特点。

（8） CatBoost

CatBoost 也是基于梯度提升的集成算法，性能优于许多现有的梯度提升算法（例如 XGBoost，LightGBM 等）。CatBoost 与其他梯度提升算法的一个主要区别是，CatBoost 采用了对称树结构，有助于减少计算时间。CatBoost 嵌入了自动将类别型特征处理为数值型特征的创新算法，能够高效合理地处理分类型特征。CatBoost 还利用了基于特征之间联系而形成的组合分类特征，丰富了特征维度。另外，CatBoost 采用的 ordered boost 方法避免了梯度估计的偏差，进而解决了预测偏移的问题。

本章节采用的回归性能衡量指标包括 RMSE，CV-RMSE，计算公式为：

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{k=1}^{n}(y_k-\widehat{y_k})^2}{n}} \tag{5-1}$$

$$\mathrm{CV-RMSE}=\sqrt{\frac{\sum_{k=1}^{n}(y_k-\widehat{y_k})^2}{n}}\Big/\frac{\sum_{k=1}^{n}y_k}{n} \tag{5-2}$$

其中，y_k为实测值，$\widehat{y_k}$为预测值，n为算例个数。

5.2.3 模型调参

在初选确定适合本应用场景的机器学习算法后还需要对该算法的超参数进行调试取优。与在训练中学习的模型参数不同，模型超参数由建模者在训练前设置，并控制模型的实现。在线性回归模型的训练中学习的权值是参数，而随机森林中的树的数量是模型的超参数，因为这是由建模者设置的。超参数可以看作是模型设置。这些设置需要针对每个问题进行调整，因为针对一个特定数据集的最佳模型超参数并不会对所有数据集都是最佳的。超参数调优(也称为超参数优化)的过程意味着为机器学习模型找到最佳的超参数值组合。超参数调优属于优化问题的一种，因此一般需要考虑以下几个方面：

- 目标函数：接受超参数并返回建模者试图最小化或最大化的分数的函数。
- 域：要开展搜索的超参数值的集合，一般需要建模者根据经验手动确定。
- 算法：在目标函数中选择下一组要评估的超参数的方法。
- 历史结果：包含超参数组合和目标函数的结果分数，算法需要使用过去的结果来决定下一个进行评估计算的超算数组合。

目前常用的调参方法主要包括以下几种：

- 手动调参：基于直觉、经验、猜测选择超参数，用超参数训练模型，并对验证数据进行评分。重复这个过程，直到对测试结果感到满意。
- 网格搜索：建立超参数值的网格，针对每个组合，训练一个模型并对验证数

据进行评分。在这种方法中，尝试每一个超参数值的组合，这种方法耗时长，比较低效。

- 随机搜索：建立超参数值网格，随机选择组合对模型进行训练和评分。搜索迭代的次数是根据时间或资源设置的。
- 自动超参数调优：使用梯度下降、贝叶斯优化或进化算法等方法来引导搜索最佳超参数。

本文采用随机搜索的方法进行超参数调优。与网格搜索相比，随机搜索的效率惊人。随机搜索通常会在更少的迭代中找到一个“足够接近”的值。网格搜索花费了太多的时间来评估超参数搜索空间中没有希望的区域，因为它必须评估网格中的每一个单独的组合。相比之下，随机搜索可以更好地探索搜索空间，因此通常可以在更少的迭代中找到超参数的良好组合。

为了评估每个超参数值组合，我们需要在一个验证集上对它们进行评分。超参数不能在测试数据上进行调优。在评估最终模型时，我们只能使用测试数据一次。当在真实数据上部署时，测试数据用于作为模型性能的评价，因此我们不希望用测试数据优化模型。因此，正确的方法是使用验证集。然而，与其将有价值的训练数据分割成单独的训练和验证集，一般使用 k 折交叉验证，k 由建模者自行确定，在本文中使用 5 折交叉验证。除了保留训练数据之外，这比使用单个验证集能耗更好地估计模型测试集上的泛化性能。

5.3 建筑信息及能耗数据库的建立

5.3.1 数据库的选择

为了存放用于建立混合能耗模型的训练数据，本课题搭建了建筑信息即能耗数据库。数据库又被称为数据管理系统，是按照数据结构来组织、存储和管理数据的仓库。数据库类型有很多种，按照存储模型划分，主要可分为网状数据库、关系型数据库、树状数据库和面向对象数据库等，其中关系型数据库(Relational Database Management System)比较常用，常见的商业数据库，例如 Oracle、MySQL 等都属于关系型数据库。所谓关系型数据库，是建立在关系模型基础上的数据库，借助于集合等数学概念和方法来处理数据库中的数据，所谓的“关系型”可以理解为“表格”的概念, 一个关系型数据库由一个或多个表格组成。关系型数据库的特点是：

- 数据以表格的形式存储和展现；
- 每一行为数据名称，每一列为数据名称对应的数据域；

- 若干行列组成一张表单，若干表单组成数据库。

本文的数据库工具是当下最流行的开放源码 SQL 数据库管理系统—MySQL，该数据库具有体积小、速度快、开放源代码、界面友好等特点，同时能为多种编程语言（包括 C、C#、C++、Python 等）提供接口 API 进行创建、访问、管理等操作。本课题采用 Python 语言对数据库进行各种操作。Python 的标准数据库接口为 Python DB-API，支持 Python 与多种数据库连接。DB-API 是一个规范，它定义了一系列必须的对象和数据库存取方式。Python 与 MySQL 数据库的接口为 MySQLdb。

5.3.2 数据库结构设计

数据库建立的目的是为了建立能耗预测模型，为了能够实现新建建筑的能耗预测，需要以影响建筑空调能耗的因素作为输入特征变量，能耗值作为输出预测值（在本课题中，统一将时间颗粒度设置为逐日），建立两者之间的映射关系，即回归模型。根据第二章的敏感性分析，建筑空调能耗的变化仅取决于少数几个关键变量，因此为了避免黑箱模型的特征过多而造成维度灾难，我们仅把关键变量以及与天气、人员活动相关的参数作为预测模型的输入特征变量，存储到数据库中。在数据库中，共有 7 张表，分别用于存储建筑关键变量值，对应城市或气候区的气象参数时间序列值以及空调分项能耗的时间序列值，表与表之间用建筑编号和所在地点作为“键”进行关联，如图 5.3 所示。建筑信息表是主表，每一行数据代表一栋建筑，建筑用编号唯一标识，建筑的分项（冷机、风机、水泵、冷却塔）及总空调能耗分别存储于对应的表中，以“建筑编号”为主键与建筑信息表链接。建筑信息表和天气日表用“所在城市”进行链接。能耗日表和天气日表用“时间”进行链接。

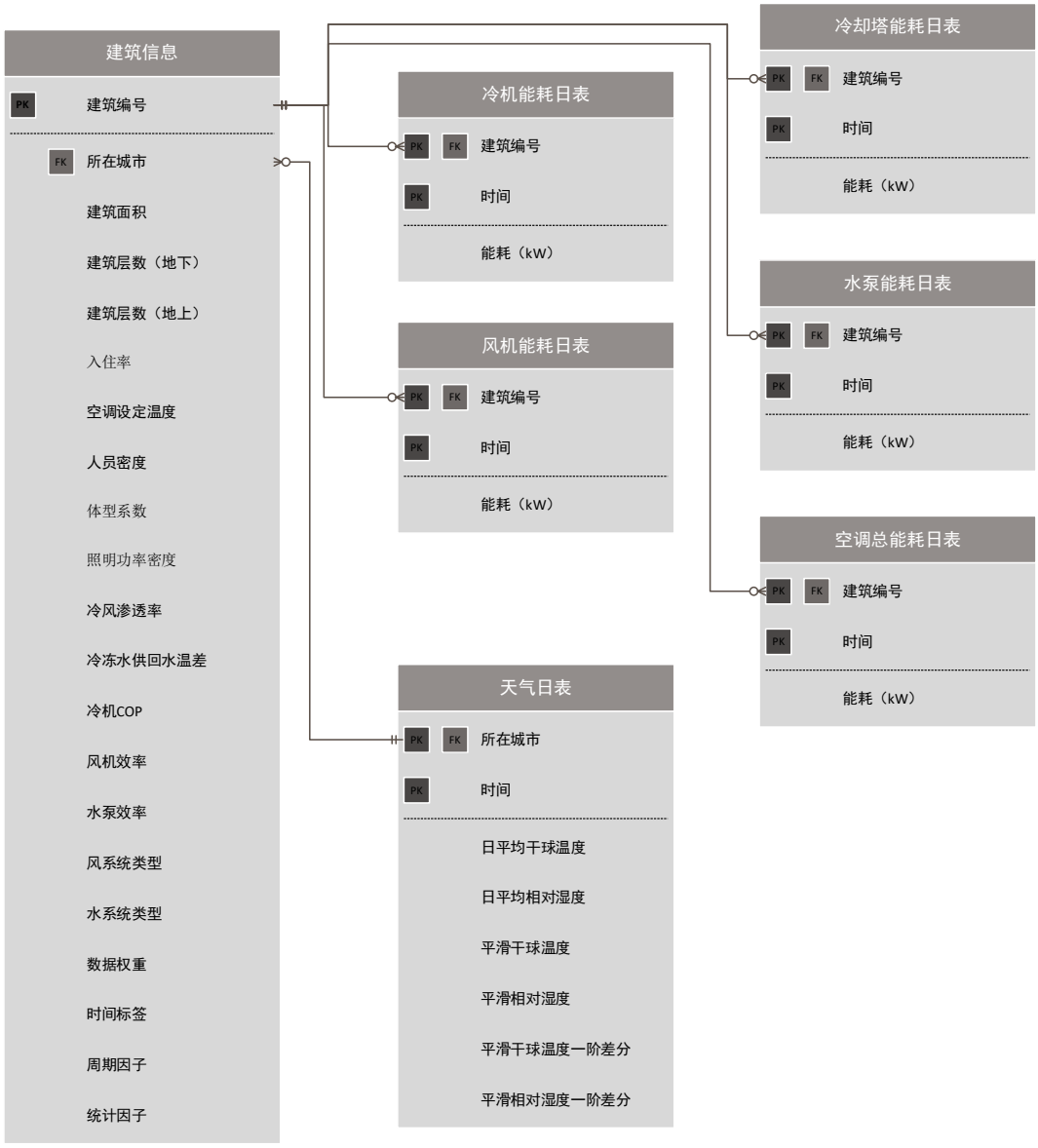


图 5.2 数据库表格关系

5.3.3 数据库数据源预处理

对于黑箱模型而言，训练数据的质量和数量直接决定了模型的精度。训练数据质量差会造成模型误差大，训练数据数量太少会造成模型欠拟合，即在训练集和测试集上模型预测结果都比较差。为了解决上述问题，数据库包含了三类数据：（I）建筑能耗分项计量数据、（II）能源审计账单、以及（III）能耗模型模拟数据。前两类数据的加入是为了解决准确度问题，第三类数据是为了解决数据量问题。本课题共累计收集了 26 组酒店建筑的相关数据，其中 6 组来自分项计量平台，20 组来自能源审计报告，以及 100 个模拟建筑样本计算得到的数据。这三

类数据有以下特点：

（1） I类数据有细颗粒度的能耗时间序列值（15min），但由于传感器故障或采集系统的问题，往往数据质量不高，存在很多异常值，需要预处理。另外，该类数据缺乏对应的建筑基本信息。

（2） II类数据来自能耗审计报告，能源审计报告则会对建筑及其设备系统进行详细勘察，包括围护结构、照明电器设备的使用、空调系统的运行状况及设备效率等，因此可以提供较为详细的建筑基本信息，但审计报告中通常只包含建筑逐月总用电量，数据颗粒度很粗。

（3） III类数据有完整的建筑信息和能耗时间序列值，并且数据量很大，但其置信度不高。

为了解决上述问题从而建立完备的模型训练数据库，本课题提出了表 15 所示解决方案。需要说明的是，本文中 I、II 类数据描述的是不同建筑、不同分项的数据，因此两者的建模过程是分开的，没有交集。但若是两者数据描述的是同一个建筑，两者数据可结合，提升数据融合及关键变量推测的准确度。

表 15 模型数据库数据预处理方案

问题	数据类型	解决方案
数据质量差	I	基于模型的数据融合方法（详见第三、四章）
关键变量值缺失	I、II	贝叶斯关键变量推测（详见第三、四章）
能耗数据颗粒度大	II	基于代理模型的数据颗粒度转换（见图 5.2）
数据置信度不高	III	设置数据训练权重

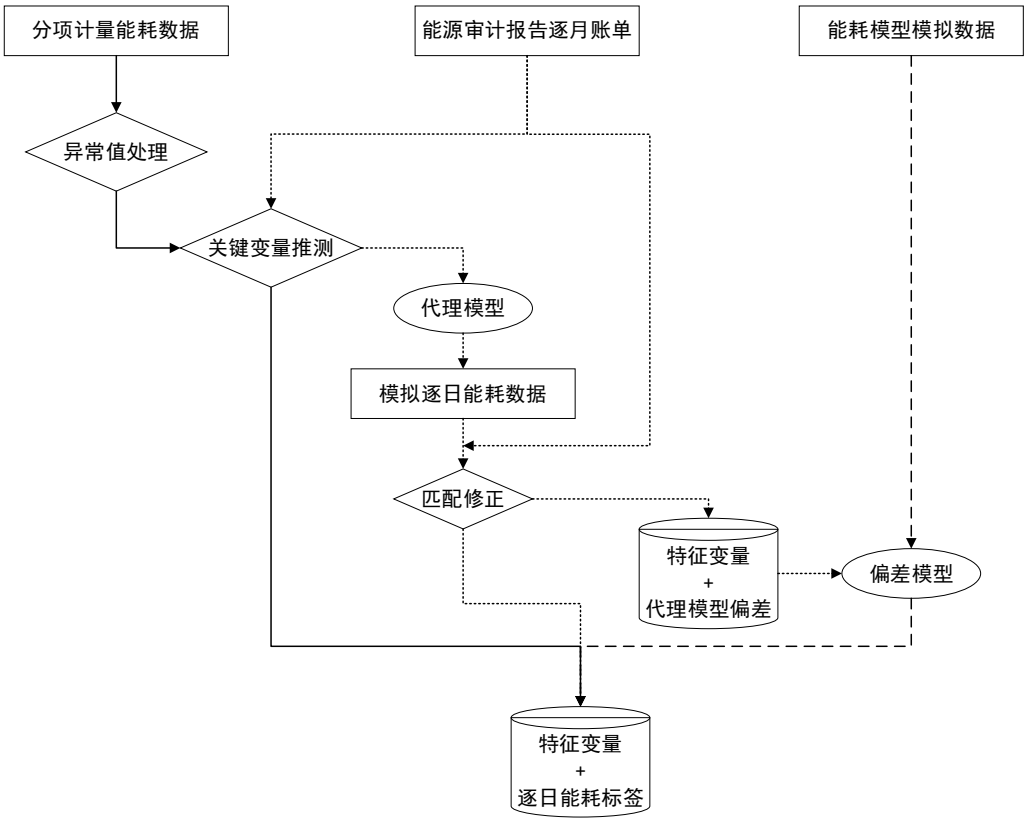


图 5.3 建筑信息数据库数据预处理流程

本课题建立的建筑能耗数据库要求不同来源的能耗数据均以同一颗粒度进行存储，其构建流程如图 5.3 所示。来自分项计量平台的数据最小颗粒度为 15min，进行累加即可得到逐日数据。能源审计报告的数据颗粒度为逐月，本课题提出的数据颗粒度转换方法为，首先基于逐月数据推测出完整的关键变量值，再基于代理模型生成逐日能耗数据并进行修正，用这种方法可生成多种颗粒度的数据格式，本课题将数据颗粒度统一为逐日的格式进行存储和建模。这里需要进行修正的原因是模型计算值与真实过程之间存在偏差（即 3.6.3 节中的$\delta(x)$），这里我们进一步建立了特征变量和模型偏差之间的映射关系，从而可以对模拟数据进行一定程度的纠偏。

由于目前阶段我们收集到的实测数据量较少，容易使模型过拟合，因此加入了一定量的模拟数据。这部分模拟数据集的输入特征取值是有拉丁超立方抽样得到的，抽样范围参考第二章初始变量集的取值范围，对应的能耗取值由能耗模拟软件 EnergyPlus 计算得到。

5.4 特征缺失情况下的能耗非确定预测

在实际预测场景中，很多情况下不能准确获得所有关键变量的值，即模型的输入变量是缺失的。为了扩展混合能耗模型的适用场景，提升算法鲁棒性，本小节提出了在特征缺失情况下进行非确定能耗预测的方法。该方法给出的预测结果不是确定的一个数值，而是一个区间，更有利于帮助用户制定决策。该方法基于关联规则挖掘算法（Apriori），算法设计基础是考虑建筑特征之间存在关联规则，比如早期建造的建筑照明功率密度比较高，设备老化效率低；经过节能改造的建筑各方面性能提升；冷机水泵等设备效率低预示着运维管理水平低，因此可以推断出其他设备效率也比较低。

该方法原理如图 5.4 所示，首先根据数据库中已有的关键变量数据筛选强有力的关联规则关系，在进行新建筑能耗预测时，若关联规则库中未知的特征与已知特征存在某种关联，则根据两者之间的规则进行未知特征的分布预测，否则根据数据库中该特征的均值建立均匀分布，然后根据得到的未知变量分布进行抽样，将所有抽样结果代入混合能耗模型中进行批量计算得到能耗预测值的分布范围。

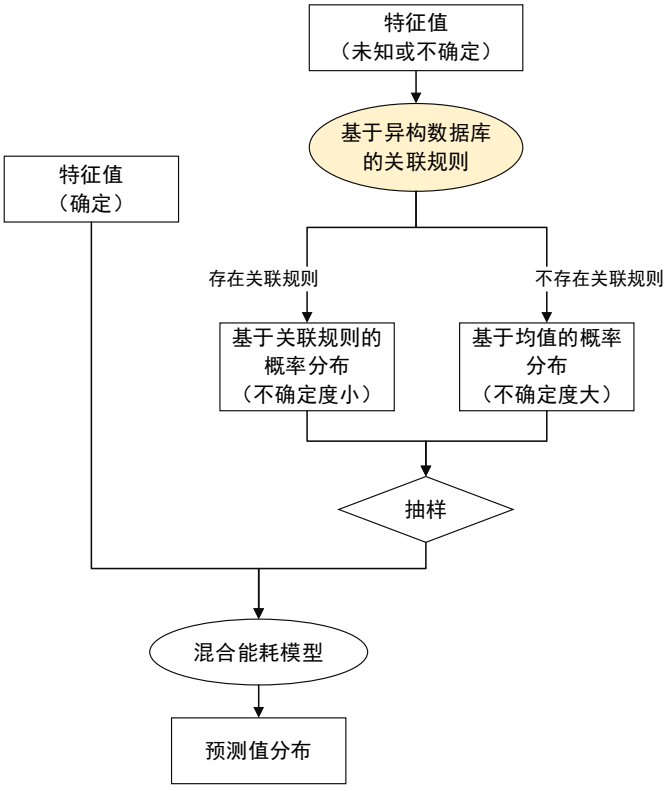


图 5.4 特征缺失情况下能耗不确定预测流程图

Apriori 算法是 R. Agrawal 和 R. Srikant 提出的一种在数据集中寻找布尔关联规则频繁项集的方法。由于该算法使用了频繁项集属性的先验知识，因此其名称

为 Apriori。首先需要定义关联分析的几个基本概念：

（1） 支持度：关联规则 A→B 支持度是指事件 A 和 B 同时发生的概率，$support = P(AB)$。

（2） 置信度：指发生事件 A 的基础上发生 B 的概率，$confidence = P(B|A) = P(AB)/P(A)$

（3） k-项集：包含k个元素的事件，满足某一最小支持度阈值的事件称为频繁k-项集。

简单来说，Apriori 算法分为两个步骤：

$Step\ 1$ 通过迭代，找出数据库中所有的频繁项集，确定频繁项集的阈值由用户自行设定；

$Step\ 2$ 利用频繁项集构造出满足用户最小信任度的规则。

具体步骤如图 5.5 所示：

首先扫描一遍数据集，从中生成 1-项集C1。接着调用 Scan 函数扫描C1，过滤不满足最小支持度的项集，最后留下的项集就是频繁项集L1。第二轮迭代中，只需要对上一轮迭代产生的频繁项集进行新的组合即可，然后接着调用Scan函数检查新组合的支持度是否满足最小支持度要求，将不满足的新组合给过滤。如此循环，直到没有新组合可生成为止。其中，C1，C2，…，Ck分别表示 1-项集，2-项集，…，k-项集；L1，L2，…，Lk 分别表示对应项集经过“过滤”后的频繁项集；Scan：表示数据项扫描函数，该函数过滤不满足最小支持度的项集。连接步：分为两种情况，第一是从数据集中生成C1，第二是根据Lk − 1 生成Ck。简单地说，连接步就是产生项集的过程。剪枝步：剔除不满足最小支持度的项集。

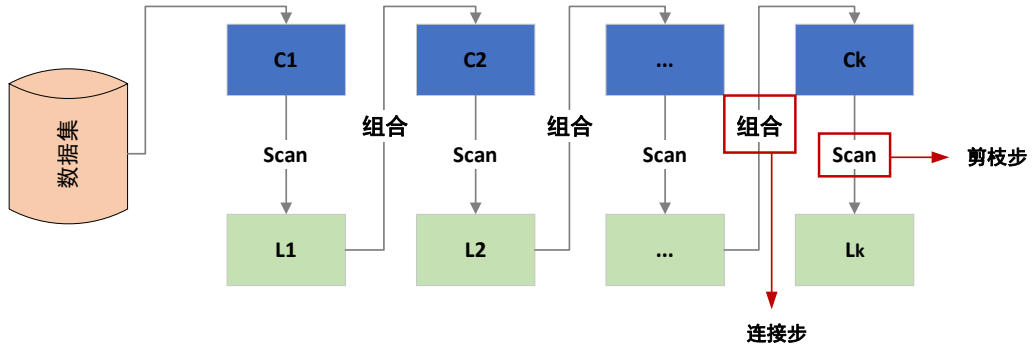


图 5.5 Apriori 算法示意图

上述 Apriori 算法只适用于离散变量，为了适用于连续型变量，本课题首先将连续变量进行离散化，然后根据上述步骤进行关联规则提取，推测得到未知特征的项集后还原变量连续取值范围。由于缺乏未知特征的分布信息，本课题假设其为均匀分布，利用拉丁超级方进行抽样得到若干未知特征的取样点，代入混合

能耗模型进行计算预测。

5.5 本章小结

本章节首先分析了两类采用数据驱动模型进行能耗预测的应用场景，第一种是基于已知的目标建筑历史能耗数据进行未来能耗预测，对于这一类能耗预测问题的研究比较成熟。另一种是目标建筑没有任何历史数据可以参考，这一预测问题，本章节提出了利用同类型其他建筑的能耗数据进行“迁移”预测的方法，并且构建了混合能耗模型。混合能耗模型是由不同来源、不同性质的数据集训练而成的能够反映建筑空调系统实际运行能耗的数据驱动模型。

首先，本章节基于前三章节研究得到的建筑信息关键构建了建筑信息及能耗数据库，该数据库同时包括了来自分项计量平台的细颗粒度能耗数据、能源审计报告的逐月能耗账单和基于模拟软件计算的能耗数据，这三类非结构化数据经过融合处理、颗粒度转换后形成了由一张建筑信息主表和对应的 6 张设备能耗记录副表组成的结构化数据，可以直接用来训练机器学习模型。针对不同的设备能耗预测模型，本章节给出了相应模型输入特征，并对比分析了 8 个常用的机器学习算法以及模型调参方法。最后，考虑进行能耗预测时存在未知输入特征的情况，本章节给出了能耗的非确定预测方法，该方法首先根据连续变量的关联规则算法进行未知特征的范围推测并抽样出若干样本点，然后根据混合能耗模型预测出目标建筑的能耗区间。

第6章 建筑混合能耗预测模型有效性验证

6.1 概述

上一章节详细阐述了混合能耗模型的建立过程，本章节将基于第五章所述方法，根据实测数据建立总制冷能耗预测模型和冷机能耗预测模型，采用“交叉测试”的方法验证模型的预测精度。“交叉测试”方法能够同时验证模型的精度和稳定性（即在不同数据集上的表现），避免由于数据选择的偶然性对模型评价造成偏差。另外，针对第五章提出的存在未知特征非确定能耗预测方法，本章节基于两个不同的数据集进行了结果展示和验证。

6.2 混合能耗模型建立及交叉测试

借鉴 k 折交叉验证的思路，为了进行更全面的验证模型有效性，本文提出了“交叉测试”的方法，如图 6-1 所示。由于本文的建筑信息及能耗数据库是有若干建筑的能耗数据组成的，为了验证本文所提出的关键变量提取方法及混合能耗模型的有效性，我们进行了n次测试，n为数据库中所包含的建筑数量。每一次以其中一栋建筑的数据作为测试集，其余建筑能耗数据作为训练集建立模型。与传统的仅随机取一组数据作为测试集不同，“交叉测试”法能够依次评价模型在不同数据集上的预测效果，采用这种方法可以更真实地展示模型性能，避免由于随机选取测试数据集带来模型评价的偏差。

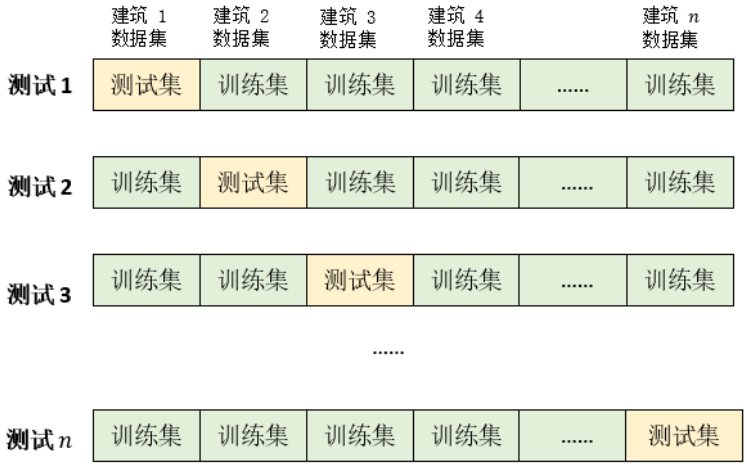


图 6.1 混合能耗模型交叉测试示意图

6.2.1 数据集描述

如 5.3.3 节所述，用于进行训练集测试的数据集包含三部分数据：

1) 6 栋上海市星级酒店建筑的冷机逐日能耗数据；
2) 20 栋上海市星级酒店建筑的总制冷能耗逐日数据（包括冷机、冷却塔、水泵、空调箱的能耗总和）；
3) 100 个模拟建筑样本，其中每个样本贡献 184 个数据点。

其中，冷机的能耗数据直接来自腾天分项能耗计量平台，如图 6.2 所示。该平台提供的数据包括建筑基本信息、能源设备清单、室外气象参数、度日数、分项能耗等。建筑基本信息数据提供了建筑类型、暖通空调系统类型、建筑面积、楼层数等建筑基本信息。在 6.2.2 小节中将基于该分项计量平台的冷机逐日能耗数据和模拟能耗数据建立冷机的混合能耗模型并进行交叉测试。

总制冷能耗数据是根据能源审计报告的逐月账单推测的逐日数据，如图 6.3 所示。在 6.2.3 小节中将基于来自能源审计报告的总制冷能耗逐日数据和模拟能耗数据建立总制冷能耗混合能耗模型并进行交叉测试。

本文训练混合能耗模型是，设置实测数据权重为 0.7，模拟数据权重为 0.3。

图 6.2 腾天分项能耗计量平台界面

图 6.3 能源审计报告（部分）

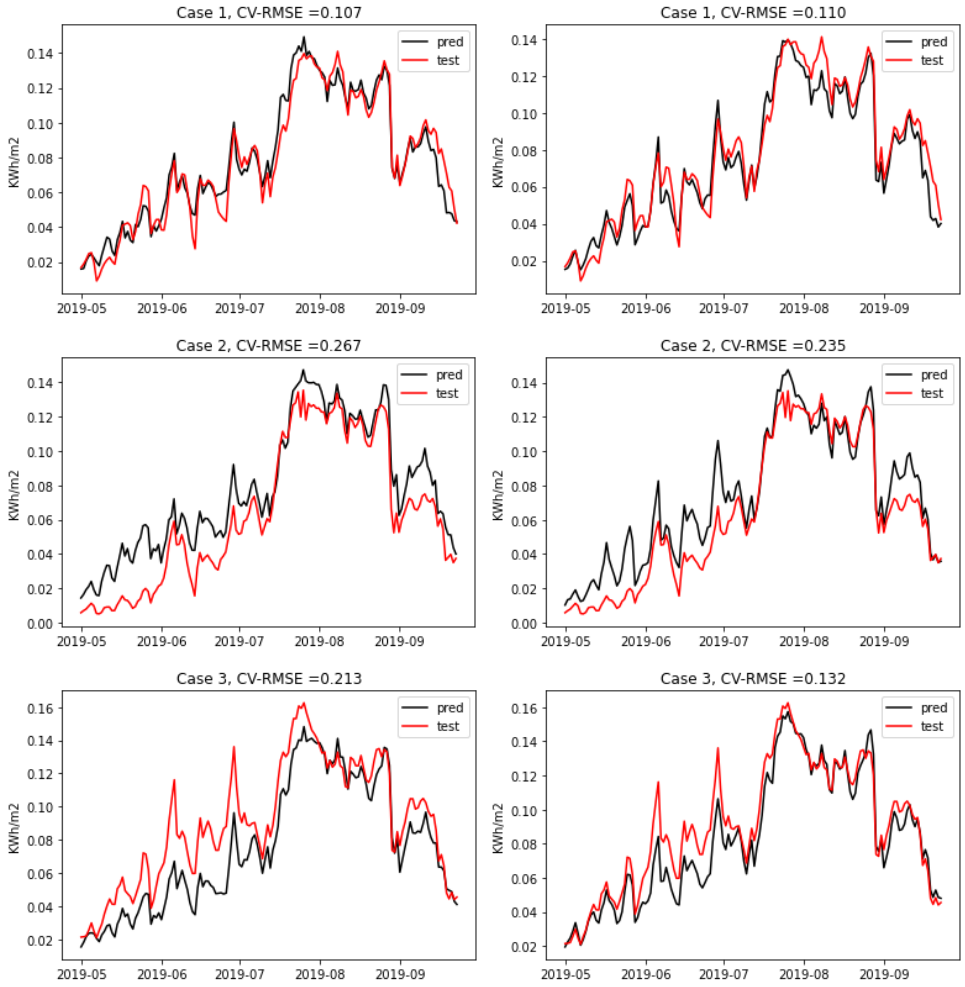
6.2.2 基于冷机能耗的模型测试

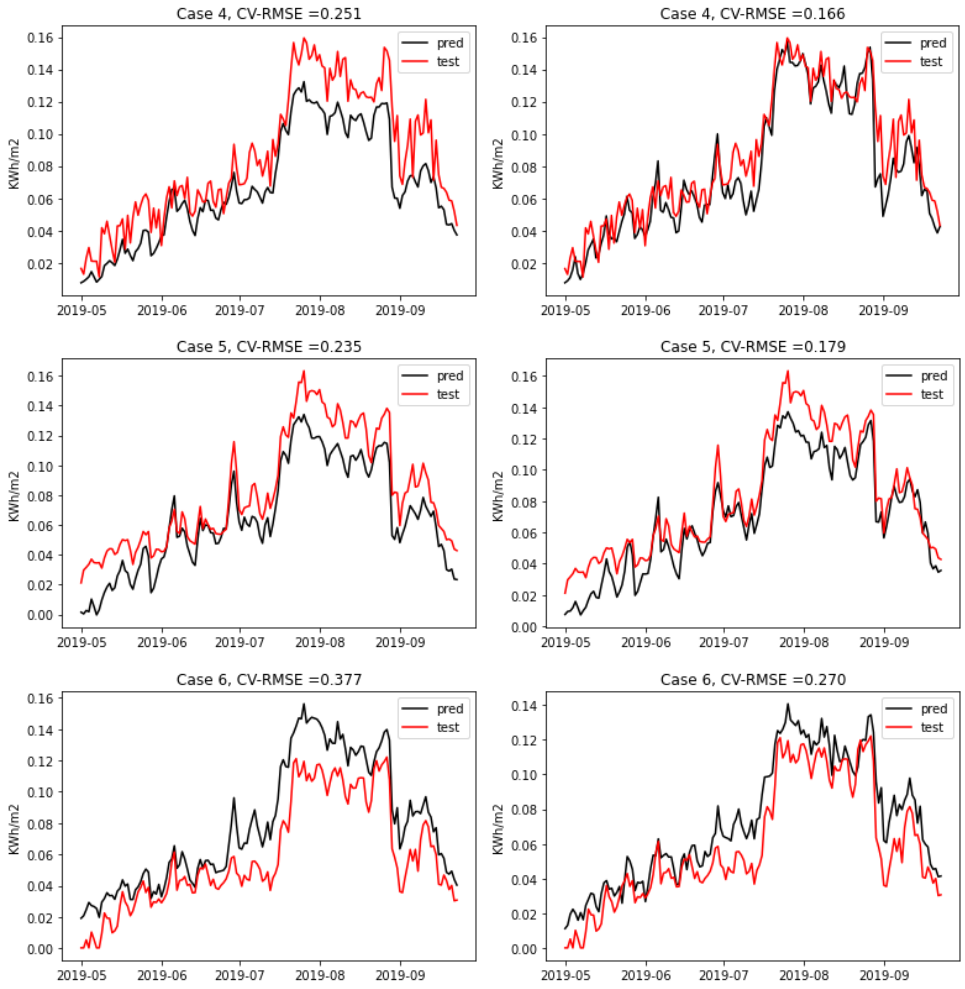
根据第 5.3 节所述方法，本小节首先根据已有冷机能耗数据集进行模型预选，备选模型和预测性能如表 16 所示，可以发现组合树模型的表现优于其他算法，选用其中表现最好的 CatBoost 算法进行建模和超算数调优。针对数据库中的 6 栋建筑冷机数据进行交叉测试，测试的结果如图 6-2 所示。另外，本文针对是否采用模拟数据进行了对比测试，图 6-2 中（a）组数据不包含模拟数据，（b）组数据包含了模拟数据，可以发现，（b）组模型的预测效果大概率优于（a）组，根据图 6-3 绘制的预测指标分布图，（b）组模型的预测指标明显低于

（a）组。这验证了在小数据集下，融合模拟数据能够提高模型的性能。加入模拟数据的混合能耗模型，其平均交叉测试 CV-RMSE 为 0.17，R2 为 0.87，这个结果对于逐日实测能耗预测来说是比较令人满意的。

表 16 基于冷机能耗数据模型预选对比

序号	算法名称	RMSE
1	CatBoost	0.0111
2	LightGBM	0.012
3	随机森林	0.0126
4	决策树	0.0167
5	线性回归	0.0229
6	岭回归	0.0551
7	支持向量机	0.0479
8	K 近邻回归	0.0397





（a）训练集不含模拟数据　　（b）训练集包含模拟数据

图 6.4 冷机能耗预测值与测试值对比

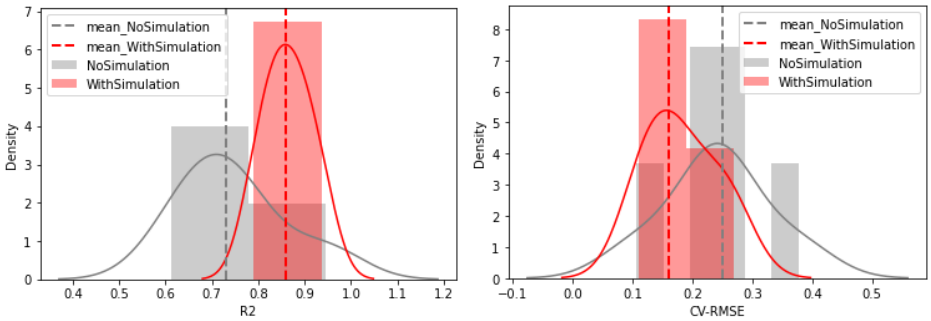


图 6.5 基于不同数据集的冷机能耗预测指标（R2 和 CV-RMSE）分布对比

本课题加入模拟建筑的目的是为了充实数据量，因为数据驱动模型的预测效果与训练数据量密切相关。但另一方面，并不是加入的模拟数据量越多越好，因为模拟数据的分布与实测数据有差异，过多模拟数据的加入会使得数据整体分布过多向其倾斜，从而降低对实测数据的预测精度。图 6.6 展示了加入不同数量模

拟建筑样本后的模型预测精度，可以发现，当模拟建筑样本数量为 100 时（此时模拟数据点的数量为实测数据点数量的 21 倍），预测精度最高，随着模拟数据的进一步扩充，模型精度反而有下降趋势，因此在实际应用过程中需要注意这一点。

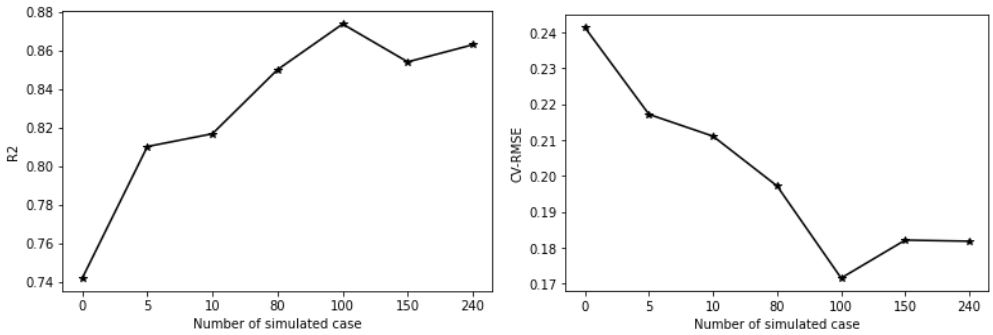


图 6.6 加入不同数量模拟数据后模型平均预测精度对比

6.2.3 基于总制冷能耗的模型测试

同样地，本小节首先根据已有制冷能耗数据集进行模型预选，备选模型和预测性能如表 17 所示，虽然各算法优劣对比与上一个案例有所不同，但组合树模型的表现依然优于其他算法，选用其中表现最好的 CatBoost 算法进行建模和超算数调优。针对数据库中的 20 栋建筑冷机数据进行交叉测试，测试的结果如图 6.7 所示，其中（a）组数据不包含模拟数据，（b）组数据包含了模拟数据。可以发现，本案例的模型预测效果明显逊于基于冷机能耗数据的模型，原因主要有两个方面，

（1） 本案例是根据审计报告的逐月能耗数据进行关键变量推测，数据量少，推测结果不如数据量大的情况；

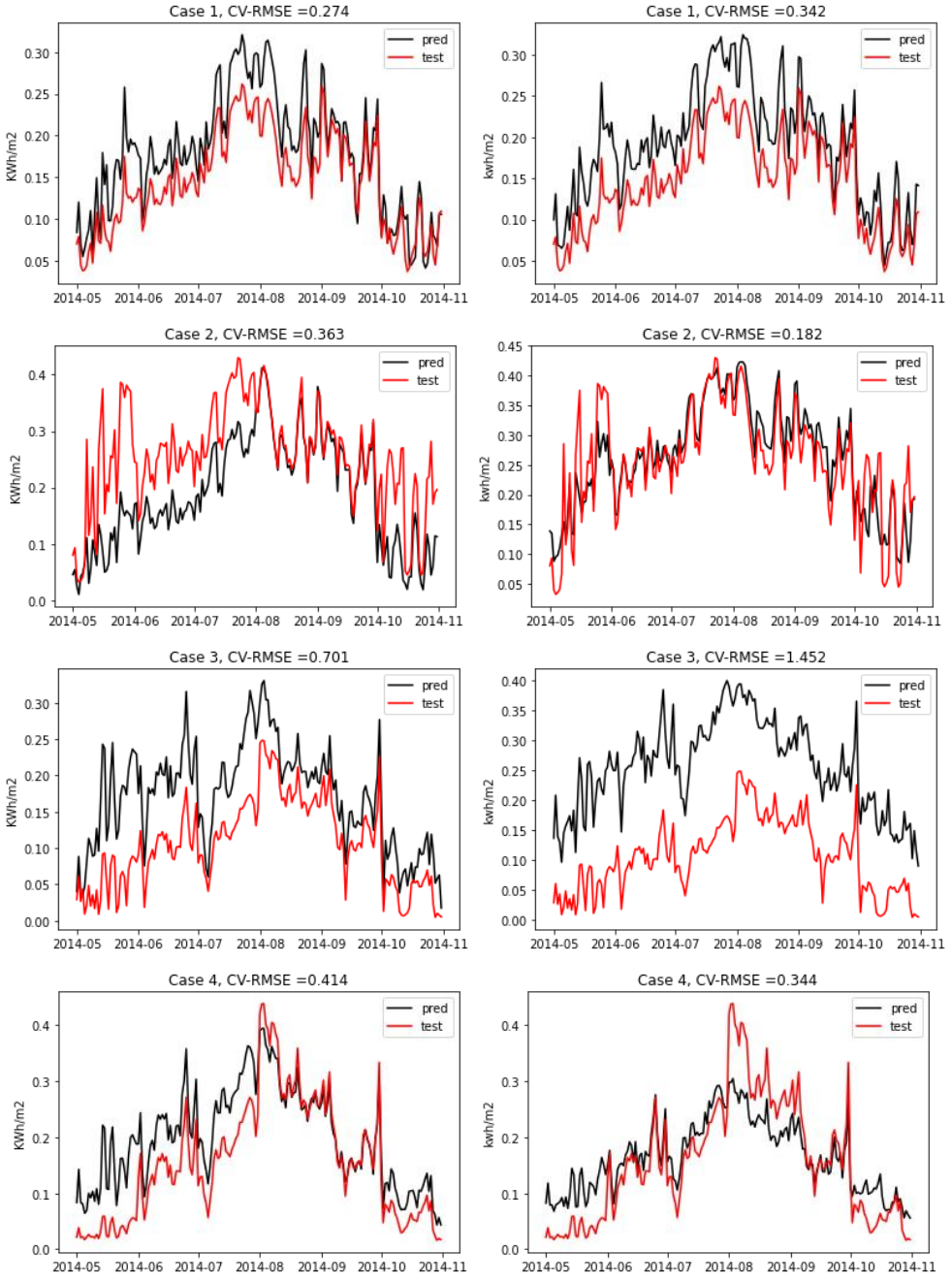
（2） 本案例需要推测影响总制冷能耗的变量，变量数相对来说比较多。

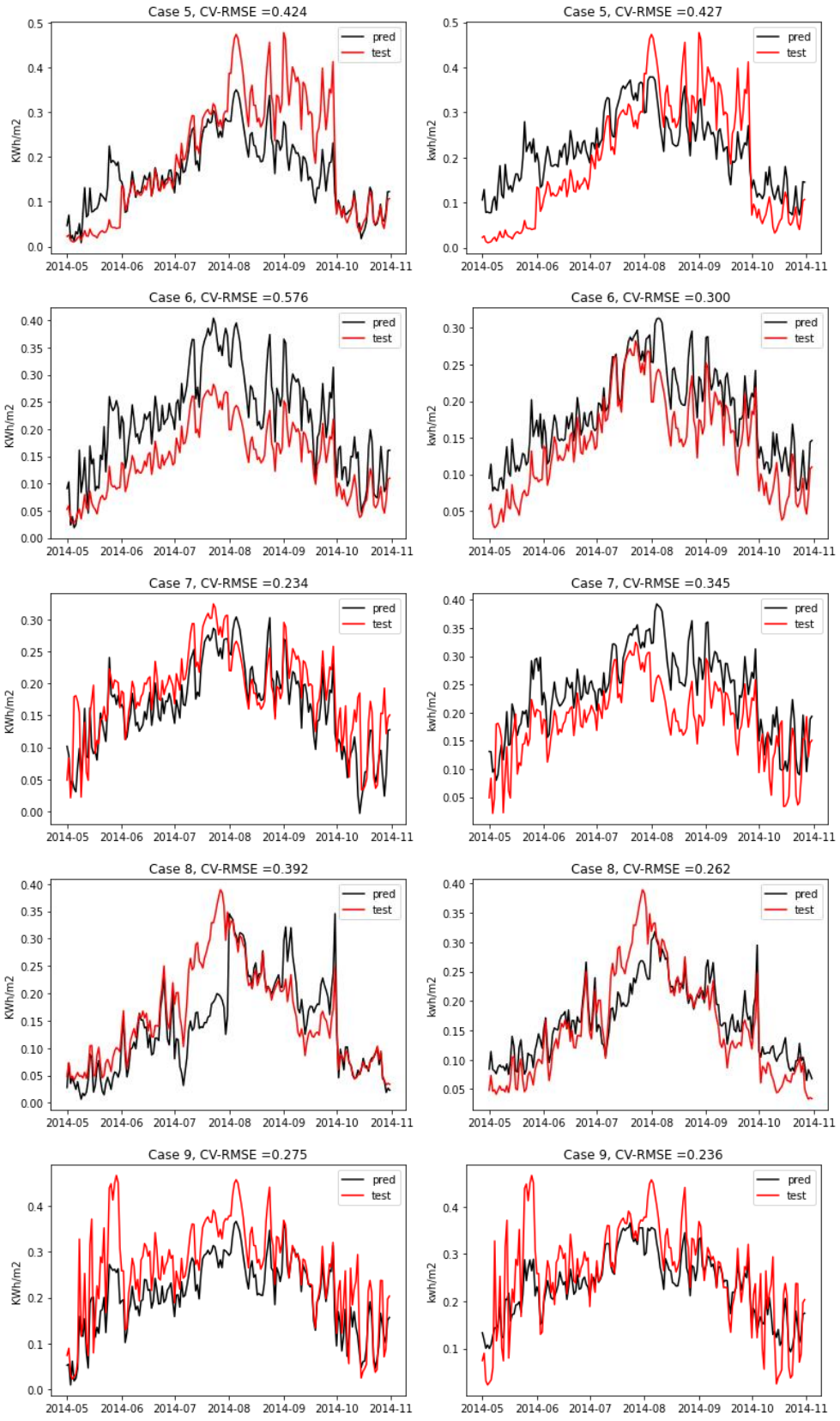
在上述两个因素的影响下，导致模型的性能不如冷机能耗预测模型。另外，根据图 6.9 绘制的预测指标分布图，（b）组模型中除了一个离群点，其余样本的测试效果比较理想，整体预测效果优于（a）组样本，但（b）组模型的预测指标均值与（a）组相差不大，主要是因为受离群值的影响，这个现象说明在实际数据足够的情况下，模拟数据起到的模型性能作用比较小，而且可能会使模型变得不稳定。

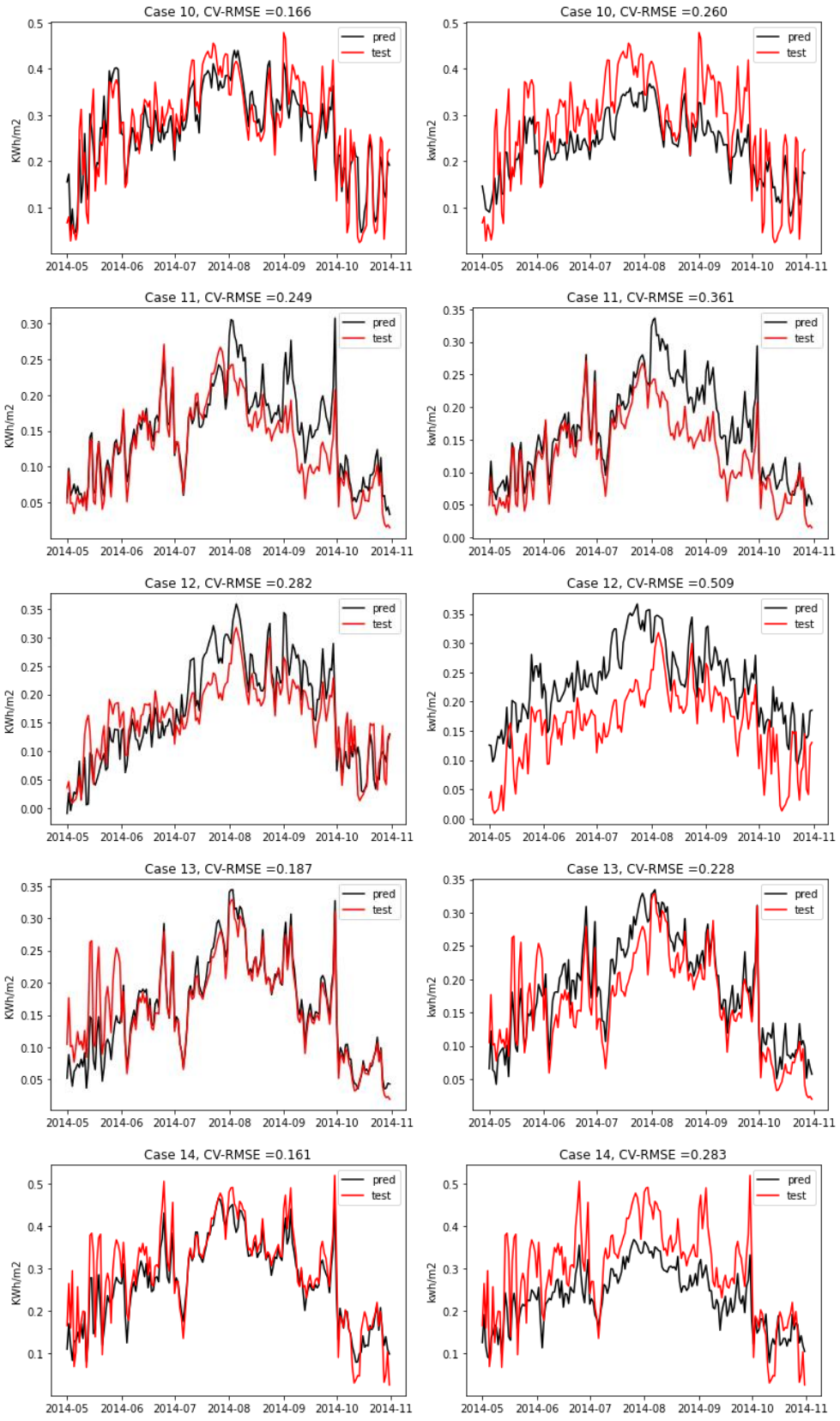
表 17 基于总制冷能耗数据模型预选对比

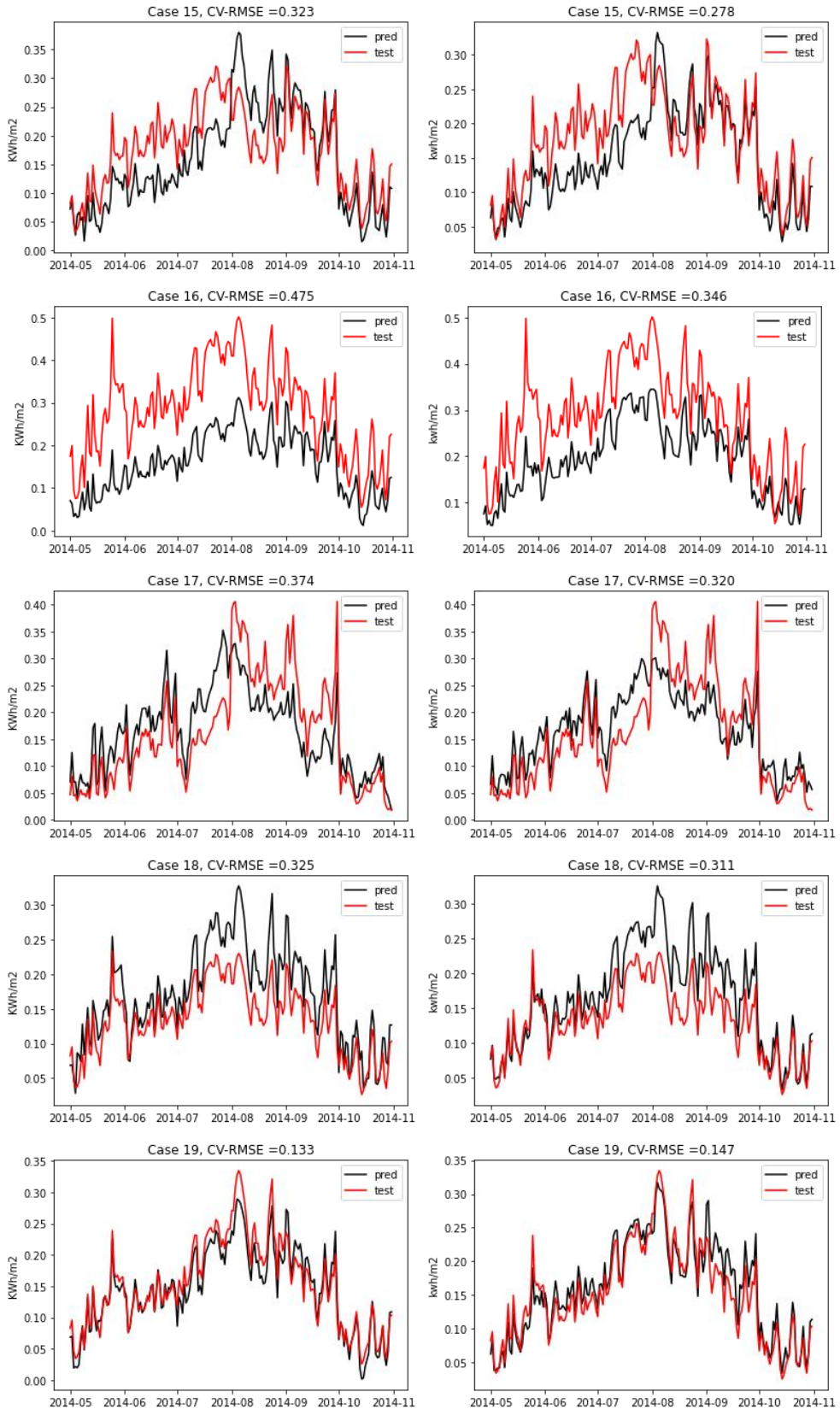
序号	算法名称	RMSE
1	CatBoost	0.0321

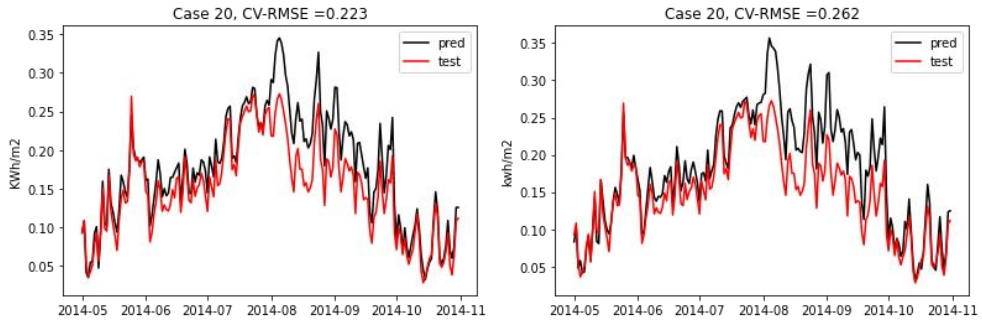
2	LightGBM	0.0334
3	随机森林	0.0352
4	决策树	0.0456
5	线性回归	0.0448
6	岭回归	0.1039
7	支持向量机	0.1018
8	K 近邻回归	0.0428




Case 5, CV-RMSE =0.424
Case 5, CV-RMSE =0.427
Case 6, CV-RMSE =0.576
Case 6, CV-RMSE =0.300
Case 7, CV-RMSE =0.234
Case 7, CV-RMSE =0.345
Case 8, CV-RMSE =0.392
Case 8, CV-RMSE =0.262
Case 9, CV-RMSE =0.275
Case 9, CV-RMSE =0.236
pred
test
KWh/m2
kwh/m2
2014-05
2014-06
2014-07
2014-08
2014-09
2014-10
2014-11


Case 10, CV-RMSE =0.166
Case 10, CV-RMSE =0.260
Case 11, CV-RMSE =0.249
Case 11, CV-RMSE =0.361
Case 12, CV-RMSE =0.282
Case 12, CV-RMSE =0.509
Case 13, CV-RMSE =0.187
Case 13, CV-RMSE =0.228
Case 14, CV-RMSE =0.161
Case 14, CV-RMSE =0.283
pred
test
KWh/m2
kwh/m2
2014-05
2014-06
2014-07
2014-08
2014-09
2014-10
2014-11


Case 15, CV-RMSE =0.323
Case 15, CV-RMSE =0.278
Case 16, CV-RMSE =0.475
Case 16, CV-RMSE =0.346
Case 17, CV-RMSE =0.374
Case 17, CV-RMSE =0.320
Case 18, CV-RMSE =0.325
Case 18, CV-RMSE =0.311
Case 19, CV-RMSE =0.133
Case 19, CV-RMSE =0.147
pred
test
KWh/m2
kwh/m2
2014-05
2014-06
2014-07
2014-08
2014-09
2014-10
2014-11



（a）训练集不含模拟数据　　（b）训练集包含模拟数据

图 6.7 冷机能耗预测值与测试值对比

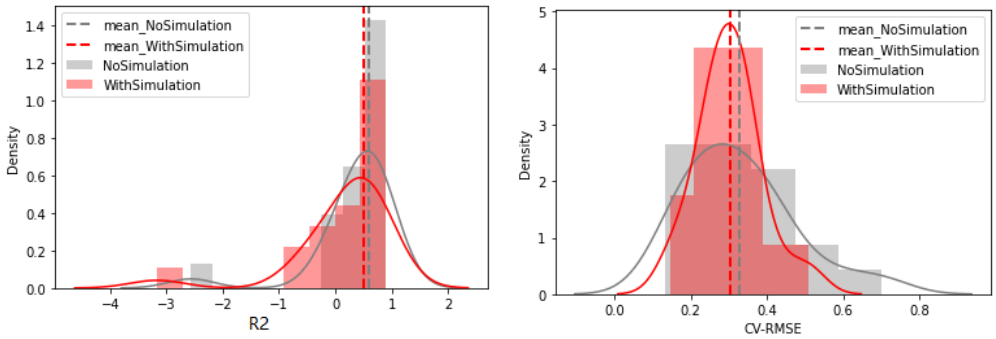


图 6.8 基于不同数据集总制冷能耗预测指标（R2 和 CV-RMSE）分布对比

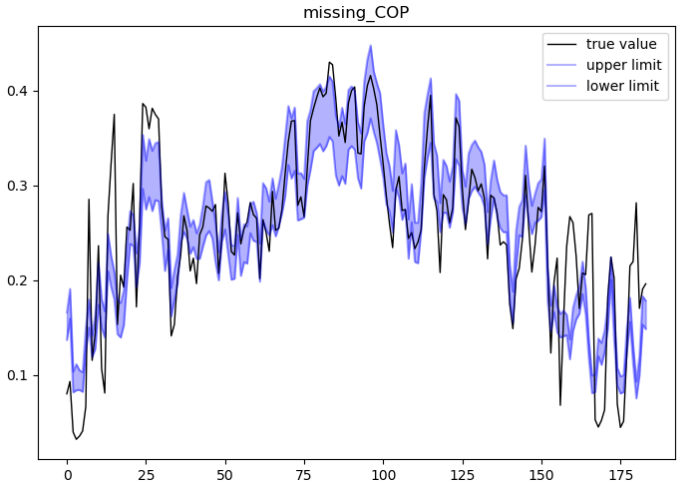
6.3 存在未知特征的能耗分布预测验证

本小节将针对 5.4 小节提出的特征缺失情况下的能耗非确定预测算法进行验证，数据集即为 6.2.2 小节所使用的数据集，由于文章篇幅所限，本小节从原始数据集中取一个建筑样本的数据为测试集（包括 184 的数据点），其余作为训练集。

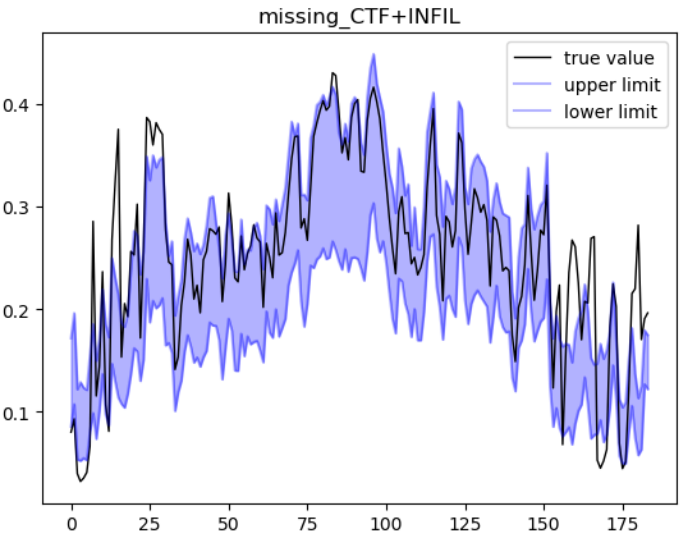
在这个测试案例中，本小节从原数据集中随机抽取一个建筑样本作为测试集。为了展示在有特征缺失的情况下能耗模型可以给出的预测结果，作者将分三次从原特征集中隐去部分特征，然后利用非确定能耗预测算法首先得到特征变量的关联规则库，然后根据规则库推测未知特征取值区间，最后带入混合能耗模型得到能耗的预测分布。

提取出的关联规则库见附录 E，模型预测结果如图 6.9 所示，图中黑色曲线表示的是实际能耗值，蓝色阴影区表示模型给出的能耗预测区间。在三次测试中，特征的信息缺失程度从低到高排列，（a）组仅缺失 COP 信息，（b）组缺失冷却塔堵塞率和渗透系数，（c）COP、冷却塔堵塞率、风机效率和渗透系数。从结果

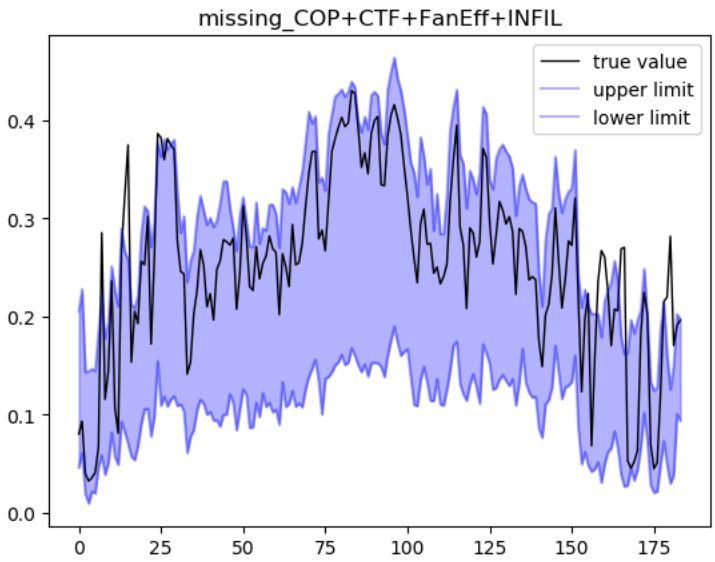
图中可以看出，随着信息缺失程度的越来越严重，模型给出的预测范围越来越宽泛，说明不确定性在变强。但实际能耗曲线基本被包含在预测区间内，说明根据关联规则算法推测出的未知特征分布区间是可信的。因为只有在特征变量取值准确的情况下，模型计算得到的能耗预测区间才不会与实际值出现很大偏差。



（a）COP 缺失时能耗预测结果



（b）冷却塔堵塞率和渗透系数缺失时能耗预测结果



（c）冷却塔堵塞率、风机效率和渗透系数缺失时能耗预测结果

图 6.9 同特征缺失时的能耗非确定预测

6.4 模型应用场景及方法说明

至此，本文涉及的酒店建筑混合能耗模型的构建及验证已全部完成，由于可用数据的限制，目前仅包括以下四个模型：

1) 冷机能耗预测模型
2) 总制冷能耗预测模型
3) 冷机能耗非确定预测模型
4) 总制冷能耗非确定预测模型

上述模型可直接用于夏热冬冷地区预测新建建筑或既有建筑的冷机或总制冷能耗值。其中，前两个模型用于当所有输入关键变量已知的场景，预测结果为确定的能耗序列值；后两个模型用于存在部分输入关键变量未知的场景，预测结果为能耗序列值区间（即每个时间点的预测结果包括最大值和最小值）。预测各制冷设备能耗所需输入变量总结如表 18，由于缺乏水泵、空调箱、冷却塔实测能耗数据，本文未建立相应模型。

需要说明的是，对于其他气候区或其他类型建筑的能耗预测，不能简单套用本文所建模型，因为关键变量提取方法对边界条件敏感，其他建筑类型或气候区建筑对应的关键变量存在区别。需根据本文所述关键变量提取方法、数据融合方法和混合能耗模型建立方法重新建立。

表 18 预测不同制冷设备能耗所需输入关键变量

关键变量	预测目标

	总能耗	冷机能耗	水泵能耗	空调箱能耗	冷却塔能耗
冷风渗透率	✓	✓	✓	✓	✓
照明功率密度	✓	✓	✓	✓	✓
人员密度	✓	✓	✓	✓	✓
空调设定温度	✓	✓	✓	✓	✓
体型系数	✓	✓	✓	✓	✓
冷冻水供回水温差	✓		✓		
冷却塔填料堵塞率	✓				✓
冷机 COP	✓	✓			
风机效率	✓			✓	
水泵效率	✓		✓		
风系统过滤器堵塞率	✓			✓	
风系统类型	✓	✓	✓	✓	✓
水系统类型	✓	✓	✓		✓
干球温度	✓	✓	✓	✓	✓
湿球温度	✓	✓	✓	✓	✓
时间标签	✓	✓	✓	✓	✓
周期因子	✓	✓	✓	✓	✓
统计因子	✓	✓	✓	✓	✓
数据权重	✓	✓	✓	✓	✓

6.5 本章小结

本章节基于实际数据和模拟数据建立建筑信息及能耗数据库开展了混合能耗模型的建立及验证工作，并验证的非确定能耗预测方法的合理性。由于实测数据的局限性，本章节仅建立了冷机能耗预测模型和总制冷能耗预测模型。其中冷机能耗数据样本较少，当数据集缺少模拟数据补充时，仅依靠实测冷机能耗数据建立的能耗预测模型平均 CV-RMSE 约为 0.25，R2 为 0.73，融合模拟数据训练的模型平均 CV-RMSE 下降至 0.17，R2 提高至 0.86，模型性能有了明显的提升。但模拟数据的数量并不是越多越好，本案例中模拟数据量为实测数据量的 21 倍时，模型性能最优。对于总制冷能耗数据，其实测数据比较充实，未结合模拟数据训练得到的模型预测精度为 0.3 左右，加入模拟数据后模型平均 CV-RMSE 降

低至 0.28，仅有微弱的提升，并且出现了预测不稳定的情况。因此，模拟数据在实测数据缺乏的情况下有助于提升模型的整体性能，但当随着实测数据的不断充实，应该减少模拟数据的使用，避免由于模拟数据本身的偏差对模型性能造成不好的影响，这也是本文提出训练权重系数的目的。

另一方面，不难发现总制冷能耗预测模型的性能劣于冷机能耗预测模型，主要原因是由于用于推测的观测值较少，但需要推测的变量个数较多，造成了推测结果的不确定性增加，从而影响了最终模型的预测性能。基于上述实验结果，应优先采用分项计量数据分别对各分项设备能耗对应的关键变量进行推测，降低关键变量推测结果的不确定度。

针对存在未知特征的能耗非确定预测，本章节基于之前建立的混合能耗模型进行了验证，人为从中删除若干特征，可以发现当缺失的特征越来越多时，模型的预测结果不确定性越来越大，给出的预测区间越来越宽泛。并且缺失的特征越重要将给模型带来越大的不确定性。

第7章 结论与展望

能耗预测是实现区域能源高效规划、建筑能耗管理、优化建筑系统控制的重要手段。相对于传统的采用能耗模拟软件、基于物理模型的能耗预测，越来越多的能耗预测研究转而采用数据驱动模型。一方面，因为采用了真实的数据进行模型训练，数据驱动模型更加精准；另一方面，数据驱动模型的预测速度更快。但是，目前绝大部分基于数据驱动模型的能耗预测研究仅仅是针对单一建筑展开的，而且要求其必须存在历史能耗数据。如此，基于某一建筑建立的模型不能迁移至其他场景，没有历史能耗记录的建筑无法进行预测。更重要的是，随着能耗计量平台的越来越普及，大量的能耗数据被采集存储，虽然这些数据被集中存放，但却没有相互链接，无法发挥大数据应有的价值。因此，本课题对建筑数据的融合进行了研究，使建筑能耗大数据在时间和空间两个维度得到关联耦合。

7.1 主要结论

本课题开展了数据融合算法及混合能耗模型的理论研究工作，并以酒店建筑为具体对象进行研究结论的展示和验证。针对 1.3.2 小节提出的 4 个问题，现得出主要结论及成果如下：

（1）可能影响建筑空调系统的因素有很多，本课题在选取初始变量时，本文不仅考虑了建筑及系统理想工况及运行状态下的参数，即“理想变量”，还包括了系统偏离理想运行状态时的情况，加入了“附加变量”用以描述施工质量不到位及系统处于低效运行状态下的特征。其次本文基于敏感性分析方法分别进行了空调负荷相关和系统相关的关键变量提取，从 34 个初始变量中分别针对空调负荷及空调设备的能耗提取了若干关键变量。上海地区酒店建筑空调能耗关键变量总结如表 19。对比分析结果表明，仅采用负荷关键变量拟合逐时负荷的 CV-RMSE 为 0.0074%，拟合设计负荷的 CV-RMSE 为 0.08%；采用本文所选负荷及系统关键变量计算冷机总能耗的偏差为 9.1%，逐日能耗偏差为 10.3%，从而验证了关键变量能表征建筑负荷及系统能耗的大部分信息。

另外，本文以 Python 语言及 eppy 包为工具开发了建筑能耗参数分析工具，该工具集成了模型生成、批量计算、结果存储、敏感性分析等模块，能够针对不同类型的建筑分析不同参数（包括形体类、性能类、系统设计类及运行类等）对空调系统能耗的影响，相对于现有的参数分析工具功能更加全面。本课题着重分析上海地区酒店建筑，其他地区、其他类型建筑的关键变量可用此工具包计算得

到。

表 19 影响建筑空调能耗关键变量汇总表

关键变量	预测目标				
	总能耗	冷机能耗	水泵能耗	空调箱能耗	冷却塔能耗
冷风渗透率	✓	✓	✓	✓	✓
照明功率密度	✓	✓	✓	✓	✓
人员密度	✓	✓	✓	✓	✓
空调设定温度	✓	✓	✓	✓	✓
体型系数	✓	✓	✓	✓	✓
冷冻水供回水温差	✓		✓		
冷却塔填料堵塞率	✓				✓
冷机 COP	✓	✓			
风机效率	✓			✓	
水泵效率	✓		✓		
风系统过滤器堵塞率	✓			✓	
风系统类型	✓	✓	✓	✓	✓
水系统类型	✓	✓	✓		✓

（2）本课题提出了数据融合方法，分为两步，首先根据建筑能耗模拟数据修正实测数据，解决了实测数据质量不佳的问题；其次利用修正后的实测数据推测建筑信息关键变量的取值，从而得到与能耗数据相对应的建筑基本信息，构建完整的建筑信息及用能画像，用于建立混合能耗预测模型。其次，本课题分别以经过校验的模拟建筑和经过现场调研的实际酒店建筑为对象对数据融合方法进行验证，结果表明本课题所提数据修正算法能消除异常值、缺失值和噪声值，关键变量推测算法能够较准确地推测出关键变量的取值。

（3）为了将捕捉建筑基本信息与空调设备运行能耗之间的关系，建立无历史能耗数据场景下的能耗预测模型，本课题基于数据融合方法得到的建筑信息及用能画像搭建了建筑信息及能耗数据库，以结构化的的形式存储建筑信息及用能数据，在此基础上讨论并最终确定了混合能耗模型的模型结构、特征构成、训练数据获取方法及训练算法。分别根据 6 栋建筑的冷机逐时能耗数据和 20 栋建筑的能源审计报告的逐月电耗账单分别建立了冷机能耗预测模型和总制冷能耗模型预测模型。冷机能耗数据样本较少，结果表明当数据集缺少模拟数据补充时，仅依靠

实测冷机能耗数据建立的能耗预测模型平均 CV-RMSE 约为 0.25，融合模拟数据训练的模型平均 CV-RMSE 下降至 0.17，模型性能有了明显的提升。但模拟数据的数量并不是越多越好，本案例中模拟数据的数据为实测数据的 21 倍时，模型测试性能最优。对于总制冷能耗数据，其实测数据比较充实，未结合模拟数据训练得到的模型预测 CV-RMSE 为 0.3 左右，加入模拟数据后模型平均 CV-RMSE 为 0.28，仅有微弱的提升，并且出现了预测不稳定的情况。说明模拟数据在实测数据缺乏的情况下有助于提升模型的整体性能，但随着实测数据的不断充实，应该减少模拟数据的使用，避免由于模拟数据本身的偏差使模型性能降低。

（4）针对存在未知特征的应用场景，本课题建立了能耗非确定预测方法，验证结果表明当缺失的特征越来越多时，模型的预测结果不确定性越来越大，给出的预测区间越来越宽泛。并且缺失的特征越重要将给模型带来越大的不确定性。

7.2主要创新点

创新点 1：建筑能耗模拟数据与实测数据数据融合算法

本课题通过寻找建筑空调能耗关键变量，以代理模型为依托，实现了建筑模拟数据与实测数据的融合，修正了实测能耗数据，并使其时间维度得到拓展，同时推测出关键变量的值，完善了建筑用能画像。

创新点 2：无历史能耗数据场景下的混合能耗预测模型

基于数据融合算法，本课题实现了不同类型、不同来源的能耗数据的结构化，并建立了包含建筑信息关键变量及能耗数据的异构数据库。在此基础上建立混合能耗模型，该模型创新性地融入了模拟数据以解决实测数据量不足的问题，实验表明模拟数据的填充可显著提高模型精度和稳定性。

创新点 3：输入特征缺失场景下能耗非确定预测方法

本课题基于关联规则算法，建立未知特征与已知特征之间的关联关系，推测未知特征的分布区间，从而实现非确定能耗预测，拓展了混合能耗模型的适用场景。

7.3研究的局限性与展望

7.3.1 局限性

（1）建筑空调系统的运行是多设备联动的复杂过程，本课题选取的描述建筑空

调系统运行能耗的参数较粗略，例如仅用 COP 描述冷机效率，没有深入细分到冷机的配置、控制策略等。在后续的研究中将结合工程实际对系统重大用能设备进行更细致的参数分析。

（2）本课题采用的关键变量提取方法计算量较大，对计算设备性能要求较高，并且在观测数据缺失的情况下推测值的不确定性较大，这也是造成总制冷能耗预测模型性能不佳的原因，在后续的研究中将寻找更加高效的方法。

（3）针对关键变量推测值的不确定性，以及基于此建立的混合模型的不确定性，本课题未进行对其进行量化研究。

（4）由于目前已有数据的量不够多，本课题仅针对夏热冬冷地区酒店建筑冷机和总制冷能耗建立了预测模型，在后续的研究中将对其他类型的建筑、其他设备分别建立预测模型。

7.3.2 展望

建筑能耗预测由于其在诸多领域扮演着重要角色一直是研究热点，人工智能及大数据分析方法的快速发展也使能耗预测取得了长足的进步，数据驱动模型是近年来能耗预测的主要方法，但绝大部分是针对有历史能耗数据的场景。一些研究者意识到这个问题，建立了深度学习模型从大量不同建筑的能耗数据中挖掘有用信息，尝试用迁移学习的方法将从一部分建筑数据中学习得到的信息应用于预测其他建筑的能耗，但归根结底上述方法都是纯黑箱模型，黑箱模型有一个重要弊端是其可解释性差，相较于白箱模型，黑箱模型训练得到的参数很难与物理参数进行关联。因此本课题尝试将白箱与黑箱模型结合起来，利用白箱模型的可解释性强和黑箱模型精度高的特点，建立了混合模型，尝试从新的维度对能耗预测展开研究。

建筑空调系统的能耗是受多重因素综合影响的，室外气象状况及建筑本身及机电系统特征对其的影响已较明确，但室内人员对能耗的影响非常复杂，人员活动特征有较强的随机性，并且会对包括照明、设备、空调等分项都造成影响，大多数能耗预测相关研究仅用水平型变量对其简化处理，本课题亦是如此，对于粗颗粒度预测的场景（如逐日、逐月能耗预测）影响不大，但若要细化到逐时甚至更细颗粒度的高精度预测，则需要对人员特征展开深入的研究。

另外，本课题及大部分数据驱动模型能够预测的能耗仅仅是建筑整体的能耗，相较于白箱模型能够输出各分区能耗，适用的场景有限，但这一方向鲜有人涉足。作者将在以后的科研工作中开展持续相关研究工作。

致谢

时光荏苒，白驹过隙，不知不觉已是我再入同济的第四个冬天，此时的我双手托着下巴，脑子里回忆着过往的点点滴滴。2017 年 29 岁“高龄”的我辞去了老家的清闲工作，带着满满的期待和雄心壮志回到同济开始攻读博士学位。这一路有欢笑有泪水，结识了很多优秀的人，经历了挫折，更收获了成长。在此，我要由衷地感谢曾经支持和帮助过我的人。

首先，非常感谢我的导师许鹏教授，非常庆幸能够遇到这样一位老师。许老师是我博士阶段的导师，也是硕士阶段导师。2011 年我报考同济大学暖通专业研究生，以笔试第二的成绩入围复试。有位师兄建议我联系当时刚回国任教的许老师，因为知道的人不多，录取的希望比较大。在复试的前夕，四平路校区综合楼咖啡厅，我见到了许老师，“和善儒雅”是他给我的第一印象。在后来的学习科研过程中，许老师的教育培养方式让我的能力和自信心得到了很大的提升。在把握课题项目大方向的前提下，许老师充分信任我们，也尊重我们的意见和建议，给以我们足够的空间和自由度，抓大放小，适当鞭策，并且对失误充分包容，极大地发挥了我们的个人能动性。而且，许老师心态开放，热爱新鲜事物，经常会分享给我们他了解到的科研前沿动态，带领我们探索更新更有价值的研究领域。另外，因为我在攻读博士期间怀孕生子，耽误了很多工作，但许老师并没有因此而斥责或是催促，更多的是理解和关怀，再次献上诚挚的感谢。

其次，感谢课题组所有的小伙伴，这是一群可爱、善良又非常优秀的人，我们性格迥异但相处融洽。与他们在一起吃饭、游玩、谈天说地是最开心放松的时刻，让我忘却课题进展不顺带来的压力和苦恼，与他们讨论问题时常带给我新的灵感和思路。

最后，感谢我的家人，你们的每一句鼓励、每一次肯定和支持都给了我前行的勇气。

深深地感谢评阅我论文的老师们，感谢你们在百忙之中审阅我的论文，感谢你们的意见和建议。

最后，再次感谢所有给过我帮助的老师、同学、朋友和家人们。

2021 年 5 月 12 日

参考文献

[1] Ürge-Vorsatz D, Cabeza LF, Serrano S, Barreneche C, Petrichenko K. Heating and cooling energy trends and drivers in buildings [J]. Renew Sustain Energy Rev 2015, 15:85-98.

[2] Pérez-Lombard L, Ortiz J, Pout C. A review on buildings energy consumption information [J]. Energy Build 2008,40:394-398.

[3] 支建杰，吴蔚沁. 公共建筑能耗监测平台数据应用的探讨[C]. 2018 城市发展与规划论文集.

[4] ASHRAE. ASHRAE Handbook—Fundamentals [M]. Refrig Am Soc OfHeating Air-Conditioning Eng Atlanta, GA, USA 2009.

[5] Hong T, Buhl F, Haves P, Selkowitz S, Wetter M. Comparing computer run time of building simulation programs [J]. Build Simul 2008,128: 210-213.

[6] Kennedy MC, O'Hagan A. Bayesian calibration of computer models [J]. J R Stat Soc Ser B Stat Methodol 2001.

[7] Sha H, Xu P, Yang Z, Chen Y, Tang J. Overview of computational intelligence for building energy system design [J]. Renew Sustain Energy Rev 2019, 108: 76-90.

[8] Wu X, Kumar V, Ross QJ, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining [J]. Knowl Inf Syst 2008,14: 1-37.

[9] Bennett KP, Campbell C. Support vector machines: hype or hallelujah? [J] SIGKDD Explor Newsl 2000,2:1-13.

[10] Ng AY, Jordan MI. On discriminative vs. Generative classifiers: A comparison of logistic regression and naive bayes [J]. Adv Neural Inf Process Syst 2002:841-848.

[11] Cook TR. Macroeconomic Forecasting in the Era of Big Data [M], 2020.

[12] Goldstein M, Chatterjee S, Price B. Regression Analysis by Example [M]. J R Stat Soc Ser A 1979.

[13] Jain AK, Murty MN, Flynn PJ. Data clustering: A review [J]. ACM Comput. Surv., 1999, 31:5-16.

[14] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques [M]. 2012.

[15] MacQueen J. Some methods for classification and analysis of multivariate observations [J]. Proc. fifth Berkeley Symp. Math. Stat. Probab., 1967,3:281-291.

[16] Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters [J]. J Cybern 1973, 3:32-57.

[17] Kohonen T. The Self-Organizing Map [J]. Proc IEEE 1990,78: 1464 -1480.

[18] 周志华.机器学习[M].清华大学出版社，2016.

[19] Warren Liao T. Clustering of time series data - A survey [J]. Pattern Recognit 2005,38: 2:15.

[20] Golay X, Kollias S, Stoll G, Meier D, Valavanis A, Boesiger P. A new correlation-based fuzzy logic clustering algorithm for fMRI [J]. Magn Reson Med 1998,12: 249-260.

[21] Kuo HC, Lee TL, Huang JP. Cluster analysis on time series gene expression data [J]. Int J Bus Intell Data Min 2010, 5:56-76
[22] Liao, Bolt, Forester, Hailman, Hansen, Kaste, et al. Understanding and projecting the battle state [J]. 23rd Army Sci Conf Orlando, FL 2002:2-5.
[23] [1] Fu TC, Chung FL, Ng V, Luk R. Pattern discovery from stock time series using self-organizing maps [J]. Work Notes KDD2001 Work Temporal Data Min 2001:26-29.
[24] Owsley LMD, Atlas LE, Bernard GD. Self-organizing feature maps and hidden Markov models for machine-tool monitoring [J]. IEEE Trans Signal Process 1997, 45:2-18.
[25] Vlachos M, Lin J, Keogh E. A wavelet-based anytime algorithm for k-means clustering of time series [J]. Proc Work Clust 2003, 3: 32-57.
[26] Wilpon JG, Rabiner LR. A modified k-means clustering algorithm for use in isolated word recognition [J]. IEEE Trans Acoust 1985, 13: 587-594.
[27] Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood [J]. IEEE Trans Pattern Anal Mach Intell 2000, 22: 5-21.
[28] Kalpakis K, Gada D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series[J]. Proc. - IEEE Int. Conf. Data Mining, ICDM, 2001, 273-280.
[29] Li C, Biswas G, Dale M, Dale P. Building models of ecological dynamics using hmm based temporal data clustering – a preliminary study[J]. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2001: 53-62.
[30] Tran D, Wagner M. Fuzzy c-means clustering-based speaker verification [J]. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2002: 318-324
[31] Xiong Y, Yeung DY. Mixtures of ARMA models for model-based time series clustering [J]. Proc. - IEEE Int. Conf. Data Mining, ICDM, 2002: 717-720.
[32] Agrawal R, Imieliński T, Swami A. Mining Association Rules Between Sets of Items in Large Databases [J]. ACM SIGMOD Rec 1993, 2: 207-216.
[33] Hamilton JD. Time series analysis [M]. Princeton University Press, 1994.
[34] Kusiak A, Li M, Zhang Z. A data-driven approach for steam load prediction in buildings [J]. Appl Energy 2010, 87: 925-933.
[35] Yu Z, Haghighat F, Fung BCM, Yoshino H. A decision tree method for building energy demand modeling [J]. Energy Build 2010, 42: 1637-1646.
[36] Li Z. An empirical study of knowledge discovery on daily electrical peak load using decision tree [J]. Adv. Mater. Res., 2012, 3: 4898-4902.
[37] Gao Y, Tumwesigye E, Cahill B, Menzel K. Using data mining in optimisation of building energy consumption and thermal comfort management [J]. 2nd Int. Conf. Softw. Eng. Data Mining, SEDM 2010: 434-439.
[38] Fan C, Xiao F, Wang S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques [J].

Appl Energy 2014, 127: 1-10.
[39] Gao Y, Tumwesigye E, Cahill B, Menzel K. Using data mining in optimisation of building energy consumption and thermal comfort management [J]. 2nd Int. Conf. Softw. Eng. Data Mining, SEDM 2010: 434-439.
[40] May-Ostendorp PT, Henze GP, Rajagopalan B, Corbin CD. Extraction of supervisory building control rules from model predictive control of windows in a mixed mode building [J]. J Build Perform Simul 2013, 6: 199–219.
[41] Ahmed A, Ploennigs J, Gao Y, Menzel K. Analyse building performance data for energy-efficient building operation [J]. Proceedings of the 26th international conference on managing IT in construction, Istanbul, Turkey, 2009.
[42] Kusiak A, Li M, Tang F. Modeling and optimization of HVAC energy consumption [J]. Appl Energy 2010, 87: 3092-3102.
[43] Capozzoli A, Lauro F, Khan I. Fault detection analysis using data mining techniques for a cluster of smart office buildings[J]. Expert Syst Appl 2015, 42: 4324-4338.
[44] Sedano J, Curiel L, Corchado E, De La Cal E, Villar JR. A soft computing method for detecting lifetime building thermal insulation failures [J]. Integr. Comput. Aided. Eng., 2010, 17: 103-115.
[45] Araya DB, Grolinger K, ElYamany HF, Capretz MAM, Bitsuamlak G. An ensemble learning framework for anomaly detection in building energy consumption [J]. Energy Build 2017, 144: 191-206.
[46] Fan C, Xiao F, Zhao Y, Wang J. Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data [J]. Appl Energy 2018, 211: 1123-1135.
[47] Biscarri F, Monedero I, García A, Guerrero JI, León C. Electricity clustering framework for automatic classification of customer loads [J]. Expert Syst Appl 2017, 86: 54-63.
[48] Carpaneto E, Chicco G, Napoli R, Scutariu M. Electricity customer classification using frequency-domain load pattern data [J]. Int J Electr Power Energy Syst 2006, 28: 13-20.
[49] Zhou H, Lin B, Qi J, Zheng L, Zhang Z. Analysis of correlation between actual heating energy consumption and building physics, heating system, and room position using data mining approach [J]. Energy Build 2018, 166: 73-82.
[50] Ma H, Du N, Yu S, Lu W, Zhang Z, Deng N, et al. Analysis of typical public building energy consumption in northern China [J]. Energy Build 2017, 136: 139-150.
[51] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global Sensitivity Analysis [M]. The Primer. 2008.
[52] Hygh JS, DeCarolis JF, Hill DB, Ranji Ranjithan S. Multivariate regression as an energy assessment tool in early building design [J]. Build Environ 2012, 57: 165-175.
[53] Hopfe CJ, Hensen JLM. Uncertainty analysis in building performance simulation for design support [J]. Energy Build 2011, 43: 2798-2805.

[54] Tian W, Choudhary R. A probabilistic energy model for non-domestic building sectors applied to analysis of school buildings in greater London [J]. Energy Build 2012, 54: 1-11.
[55] Moran F, Natarajan S, Nikolopoulou M. Developing a database of energy use for historic dwellings in Bath, UK [J]. Energy Build 2012, 55: 218-226.
[56] Tian W, De Wilde P. Uncertainty and sensitivity analysis of building performance using probabilistic climate projections: A UK case study [J]. Autom Constr 2011, 20: 1096-1109.
[57] de Wilde P, Tian W. Identification of key factors for uncertainty in the prediction of the thermal performance of an office building under climate change [J]. Build Simul 2009, 2: 157–174.
[58] Garcia Sanchez D, Lacarrière B, Musy M, Bourges B. Application of sensitivity analysis in building energy simulations: Combining first- and second-order elementary effects methods [J]. Energy Build 2014, 68: 741-750.
[59] Heo Y, Choudhary R, Augenbroe GA. Calibration of building energy models for retrofit analysis under uncertainty [J]. Energy Build 2012, 47: 550-560.
[60] Heiselberg P, Brohus H, Hesselholt A, Rasmussen H, Seinre E, Thomas S. Application of sensitivity analysis in design of sustainable buildings [J]. Renew Energy 2009., 34: 2030-2036.
[61] Hyun SH, Park CS, Augenbroe GLM. Analysis of uncertainty in natural ventilation predictions of high-rise apartment buildings [J]. Build Serv Eng Res Technol 2008, 29: 311-326.
[62] Mechri HE, Capozzoli A, Corrado V. USE of the ANOVA approach for sensitive building energy design [J]. Applied Energy 2010, 87:3073–83.
[63] Spitz C, Mora L, Wurtz E, Jay A. Practical application of uncertainty analysis and sensitivity analysis on an experimental house [J]. Energy Build 2012, 55: 459-470.
[64] Ding Y, Zhang Q, Yuan T, Yang K. Model input selection for building heating load prediction: A case study for an office building in Tianjin [J]. Energy Build 2018, 159: 254-270.
[65] Li X, Ding. L, Lv J, Xu G, Li J, A novel hybrid approach of KPCA and SVM for building cooling load prediction [J]. 2010 Third International Conference On Knowledge Discovery and Data Mining, 2010: pp. 522-526.
[66] Yildiz B, Bilbao JI, Sproul AB. A review and analysis of regression and machine learning models on commercial building electricity load forecasting [J]. Renew Sustain Energy Rev 2017, 73: 1104-1122.
[67] Fan C, Sun Y, Zhao Y, Song M, Wang J. Deep learning-based feature engineering methods for improved building energy prediction [J]. Appl Energy 2019, 240: 35-45.
[68] Wang L. Heterogeneous Data and Big Data Analytics [J]. Autom Control Inf Sci 2017, 3: 8-15.
[69] 范珑, 杨成德. 基于分项计量的建筑能源管理系统的发展和应用[C]. 第八届国际绿色建筑与建筑节能大会论文集.

[70]姚香菊. 基于本体的异构数据集成技术的研究[M]. 东华大学, 2015.
[71]Chawathe S, Garcia-Molina H, Hammer J. The TSLMMIS Project: Integration of Heterogeneous Information Sources [J]. Journal of Intelligent Information System, 1994: 117-132.
[72]Sollazzo T, Handschuh S, Staab S, Frank M, Stojanovic N. Semantic Web Service Architecture — Evolving Web Service Standards toward the Semantic Web [J]. Inf Sci (Ny) 2001: 425-429.
[73]Huang G, Wang S, Xiao F, Sun Y. A data fusion scheme for building automation systems of building central chilling plants [J]. Autom Constr 2009, 18: 302-309.
[74]D.B. Ozyurt, R.W. Pike, Theory and practice of simultaneous data reconciliation and gross error detection for chemical processed [J]. Computers and Chemical Engineering, 2004, 28: 381–402.
[75]Grewal MS, Andrews AP. Kalman Filtering: Theory and Practice Using MATLAB [M]. California State University at Fullerton, 2008.
[76]Gan Q, Harris CJ. Comparison of two measurement fusion methods for Kalman-filter-based multisensor data fusion [J]. IEEE Trans Aerosp Electron Syst 2001, 37: 273-279.
[77]Huang G, Sun Y, Li P. Fusion of redundant measurements for enhancing the reliability of total cooling load based chiller sequencing control [J]. Autom Constr 2011, 20: 789-798.
[78]Djuric N, Huang G, Novakovic V. Data fusion heat pump performance estimation [J]. Energy Build 2011, 43: 621-630.
[79]Huang G, Sun Y, Wang S. Building instantaneous cooling load fused measurement: Multiple-sensor- based fusion versus chiller-model-based fusion [J]. Build Serv Eng Res Technol 2013, 34:177-194.
[80]Wang S, Wang J, Burnett J. Mechanistic model of centrifugal chillers for HVAC system dynamics simulation [J]. Build Serv Eng Res Technol 2000, 21: 73-83.
[81]Kohlenbach P, Ziegler F. A dynamic simulation model for transient absorption chiller performance. Part I: The model [J]. Int J Refrig 2008, 31: 217-225.
[82]Bendapudi S, Braun JE, Groll EA. Dynamic model of a centrifugal chiller system - Model development, numerical study, and validation [J]. ASHRAE Trans., 2005.B.
[83]Tashtoush B, Molhim M, Al-Rousan M. Dynamic model of an HVAC system for control analysis [J]. Energy 2005, 30: 1729-1745.
[84]Chen Y, Treado S. Development of a simulation platform based on dynamic models for HVAC control analysis [J]. Energy Build 2014, 68: 376-386.
[85]Mossolly M, Ghali K, Ghaddar N. Optimal control strategy for a multi-zone air conditioning system using a genetic algorithm [J]. Energy 2009, 34: 58-66.
[86]Jin GY, Cai WJ, Wang YW, Yao Y. A simple dynamic model of cooling coil unit [J]. Energy Convers Manag 2006, 47: 15-16.
[87]Wang YW, Cai WJ, Soh YC, Li SJ, Lu L, Xie L. A simplified modeling of cooling coils for control and optimization of HVAC systems [J]. Energy Convers Manag 2004, 45: 2915-2930.

[88] Nassif N, Moujaes S, Zaheeruddin M. Self-tuning dynamic models of HVAC system components [J]. Energy Build 2008, 40: 1709-1720.
[89] Sönmez NO. A review of the use of examples for automating architectural design tasks [J]. CAD Comput Aided Des 2018, 96: 13-30.
[90] Yildiz B, Bilbao JI, Sproul AB. A review and analysis of regression and machine learning models on commercial building electricity load forecasting [J]. Renew Sustain Energy Rev 2017. doi:10.1016/j.rser.2017.02.023.
[91] Raza MQ, Khosravi A. A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings [J]. Renew Sustain Energy Rev 2015, 73: 1104-1122.
[92] Widodo A, Yang BS. Support vector machine in machine condition monitoring and fault diagnosis [J]. Mech Syst Signal Process 2007, 21: 2560-2574.
[93] Van Every PM, Rodriguez M, Jones CB, Mammoli AA, Martínez-Ramón M. Advanced detection of HVAC faults using unsupervised SVM novelty detection and Gaussian process models [J]. Energy Build 2017, 149: 216-224.
[94] Gao L, Bai H, Lu Y. Development of an expert system for the calculation of building's design cooling load [J]. Proc. World Congr. Intell. Control Autom., 2003.
[95] Kalogirou S, Neocleous C, Schizas C. Building Heating Load Estimation Using Artificial Neural Networks [J]. Proc 17th Int Conf Parallel Archit Compil Tech 1997.
[96] Martellotta F, Ayr U, Stefanizzi P, Sacchetti A, Riganti G. On the use of artificial neural networks to model household energy consumptions [J]. Energy Procedia, 2017, 126: 250-257
[97] Kalogirou SA, Bojic M. Artificial neural networks for the prediction of the energy consumption of a passive solar building [J]. Energy 2000, 25: 479-491.
[98] Turhan C, Kazanasmaz T, Uygun IE, Ekmen KE, Akkurt GG. Comparative study of a building energy performance software (KEP-IYTE-ESS) and ANN-based building heat load estimation [J]. Energy Build 2014, 85: 115-125.
[99] Hygh JS, DeCarolis JF, Hill DB, Ranji Ranjithan S. Multivariate regression as an energy assessment tool in early building design [J]. Build Environ 2012, 57: 165-175.
[100] Ekici BB, Aksoy UT. Prediction of building energy consumption by using artificial neural networks [J]. Adv Eng Softw 2009, 40: 356-362.
[101] Catalina T, Virgone J, Blanco E. Development and validation of regression models to predict monthly heating demand for residential buildings [J]. Energy Build 2008, 40: 1825-1832.
[102] Sekhar Roy S, Roy R, Balas VE. Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of MARS and ELM [J]. Renew Sustain Energy Rev 2018, 82: 4256-4268.
[103] Kumar S, Pal SK, Singh RP. A novel method based on extreme learning machine to predict heating and cooling load through design and structural attributes [J]. Energy Build 2018, 176: 275-286.

[104] Vidal R, Ma Y, Sastry SS. Generalized principal component analysis [J]. Interdiscip. Appl. Math., 2016, 27:1945-1959.
[105] D. Ruch , L. Chen , J.S. Haberl , D.E. Claridge. A change-point principal component analysis (CP/PCA) method for predicting energy usage in commercial buildings: the PCA model [J]. J. Solar Energy Eng. 1993, 115: 77-84 .
[106] Kissock JK, Reddy TA, Claridge DE. Ambient-temperature regression analysis for estimating retrofit savings in commercial buildings [J]. J Sol Energy Eng Trans ASME 1998, 120: 168-176.
[107] Lin B, Wang D, Chen Z, Freihaut J. Inverse energy model development via high-dimensional data analysis and sub-metering priority in building data monitoring [J]. Energy Build 2018, 172: 116-124.
[108] Saltelli A, Tarantola S, Campolongo F, Ratto M. Sensitivity Analysis in Practice [M]. 2002.
[109] https://pypi.org/project/eppy/
[110] Higdon D, Kennedy M, Cavendish JC, Cafeo JA, Ryne RD. Combining field data and computer simulations for calibration and prediction [J]. SIAM J Sci Comput 2005, 26: 448-466.
[111] Hoffman MD, Gelman A. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo [J]. J Mach Learn Res 2014.
[112] Seeger M. Gaussian processes for machine learning [J]. Int J Neural Syst 2004, 15: 1593-1623.
[113] Chong A, Menberg K. Guidelines for the Bayesian calibration of building energy models [J]. Energy Build 2018, 174: 527-547.
[114] Fu Y, Li Z, Feng F, Xu P. Data-quality detection and recovery for building energy management and control systems: Case study on submetering [J]. Sci Technol Built Environ 2016, 22: 798-809.
[115] https://www.energycodes.gov/development/commercial/prototype_models
[116] Morrison Hershfield Limited. Building envelope thermal bridging guide [R]. BC Hydro Power Smart 2016.
[117] 胡晓俊. 中央空调系统“大流量小温差”现象的原因分析及改进措施[J]. 上海节能，2013,(02): 44-48
[118] https://www.eia.gov/consumption/commercial/about.php
[119] Sha H, Xu P, Hu C, Li Z, Chen Y, Chen Z. A simplified HVAC energy prediction method based on degree-day [J]. Sustain Cities Soc 2019, 51: 1-10.
[120] Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures [J]. Anal Chem 1964.
[121] https://en.wikipedia.org/wiki/Savitzky%E2%80%93Golay_filter
[122] 吴云标，迟艺侠. 基于贝叶斯 MCMC 方法的洪水频率分析及不确定性评估[J]. 安徽工业大学学报(自然科学版), 2018, 35(01): 66-72.
[123] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences [R].
[124] 王建平，程声通，贾海峰. 基于 MCMC 法的水质模型参数不确定性研究[J]. 环境科学, 2006,(01): 24-30.

[125] Zhou C, Fang Z, Xu X, Zhang X, Ding Y, Jiang X, et al. Using long short-term memory networks to predict energy consumption of air-conditioning systems [J]. Sustain Cities Soc 2020, 55: 2-15.
[126] Zhao L, Liu Z, Mbachu J. Energy management through cost forecasting for residential buildings in New Zealand [J]. Energies 2019, 12: 10-19.
[127] Brownlee J. Discover Feature Engineering, How to Engineer Features and How to Get Good at It. Sept 26 2014.
[128] Schmidhuber J. Deep Learning in neural networks: An overview [J]. Neural Networks 2015, 61: 85-117.
[129] Chandrashekar G, Sahin F. A survey on feature selection methods [J]. Comput Electr Eng 2014, 40: 16-28.
[130] Pardalos PM. Approximate dynamic programming: solving the curses of dimensionality [M]. 2007.
[131] Trunk G V. A Problem of Dimensionality: A Simple Example [J]. IEEE Trans Pattern Anal Mach Intell 1979, 1: 306 - 307.
[132] Wei Y, Xia L, Pan S, Wu J, Zhang X, Han M, et al. Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks [J]. Appl Energy 2019, 240: 276-294.
[133] Fan C, Xiao F, Wang S, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques [J]. Appl Energy 2014, 127: 1-10.
[134] Fan C, Xiao F, Zhao Y. A short-term building cooling load prediction method using deep learning algorithms [J]. Appl Energy 2017, 195: 222-233.
[135] Ding Y, Zhang Q, Yuan T, Yang F. Effect of input variables on cooling load prediction accuracy of an office building [J]. Appl Therm Eng 2018.
[136] Sondhi P. Feature construction methods: a survey [J]. Sifaka Cs Uiuc Edu 2010.
[137] Sun Y, Haghighat F, Fung BCM. A review of the-state-of-the-art in data-driven approaches for building energy prediction [J]. Energy Build 2020, 221: 2-16.
[138] Cai M, Pipattanasomporn M, Rahman S. Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques [J]. Appl Energy 2019, 236: 1078-1088.
[139] Ding Y, Zhang Q, Yuan T, Yang K. Model input selection for building heating load prediction: A case study for an office building in Tianjin [J]. Energy Build 2018, 159: 254-270.
[140] Russell S , Norvig P , Artificial Intelligence: A Modern Approach, 2nd ed. [M]. Prentice-Hall, 2003 .
[141] Daniel A. Evolutionary Computation for Modeling and Optimization [M]. 2006.
[142] Salcedo-Sanz S, Muñoz-Bulnes J, Portilla-Figueras JA, Del Ser J. One-year-ahead energy demand estimation from macroeconomic variables using computational intelligence algorithms [J]. Energy Convers Manag 2015, 99: 62-

71
[143] González-Vidal A, Jiménez F, Gómez-Skarmeta AF. A methodology for energy multivariate time series forecasting in smart buildings based on feature selection [J]. Energy Build 2019, 196: 71-82.
[144] https://cran.r-project.org/web/packages/Boruta/index.html
[145] Wang Z, Wang Y, Srinivasan RS. A novel ensemble learning approach to support building energy use prediction [J]. Energy Build 2018, 159: 109-122.
[146] Candanedo LM, Feldheim V, Deramaix D. Data driven prediction models of energy use of appliances in a low-energy house [J]. Energy Build 2017, 140: 81-97
[147] Jai n R., Damoulas T, Kontokosta C.E. Towards data-driven energy consumption forecasting of multi-family residential buildings: feature selection via the lasso [J]. Comput Civ Build Eng 2014, 12: 3-17.
[148] Guo Y, Wang J, Chen H, Li G, Liu J, Xu C, et al. Machine learning-based thermal response time ahead energy demand prediction for building heating systems [J]. Appl Energy 2018, 221: 16-27.
[149] Yuan P, Duanmu L, Wang Z. Coal consumption prediction model of space heating with feature selection for rural residences in severe cold area in China [J]. Sustain Cities Soc 2019, 50: 34-47.
[150] Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning [J]. Appl Artif Intell 2003, 17: 519-533.
[151] García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR. Pattern classification with missing data: A review [J]. Neural Comput Appl 2010, 19: 263-282.

附录 A 星级酒店建筑代理模型时间表设置

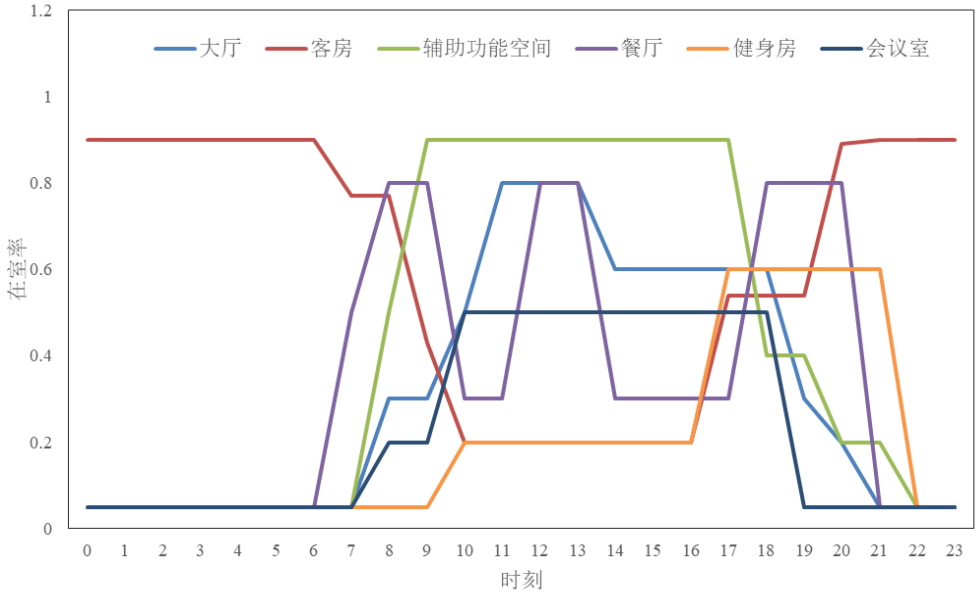


图 A.1 人员在室率时间表

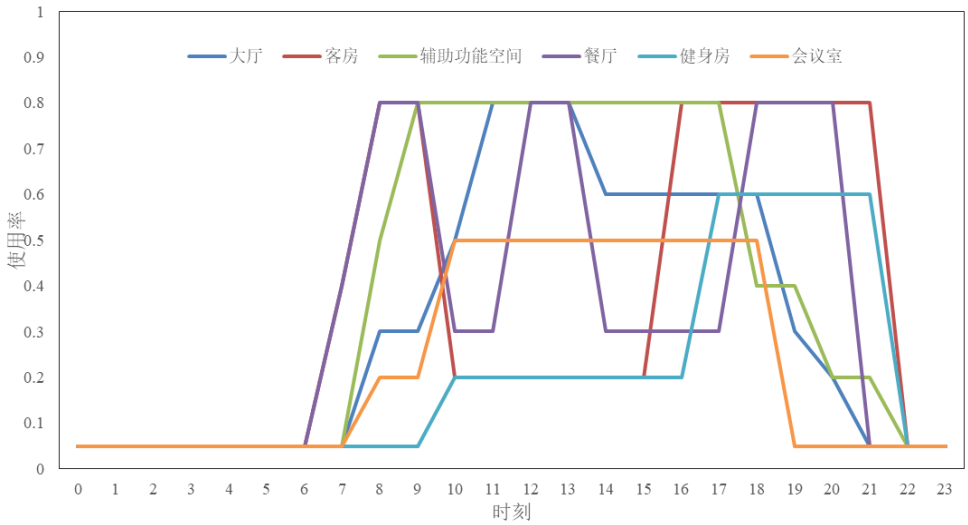


图 A.2 照明设备使用率时间表

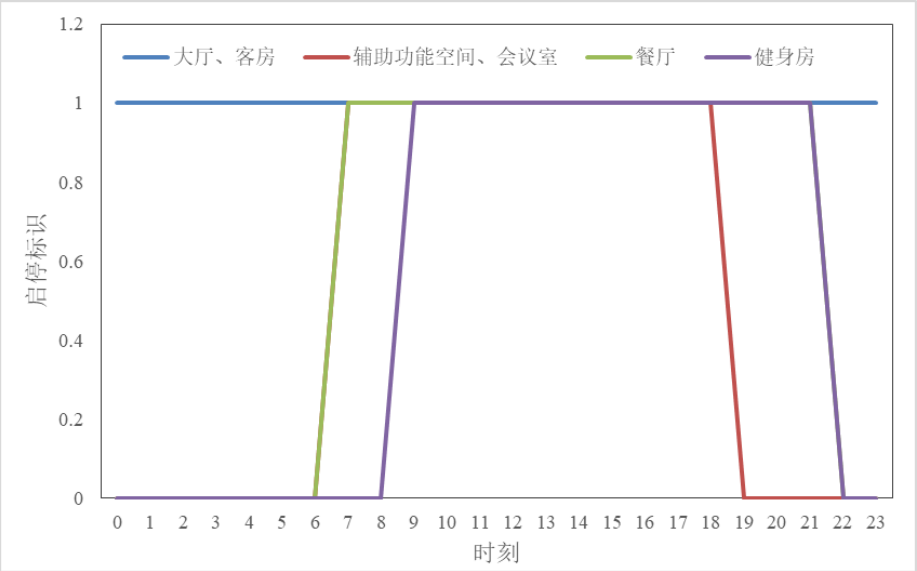

1.2
1
0.8
0.6
0.4
0.2
0
启停标识
大厅、客房
辅助功能空间、会议室
餐厅
健身房
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
时刻

图 A.3 空调启停时间表

附录 B 关键变量提取工具代码节选

```
#主程序
import os
import pandas as pd
from eppy.modeleditor import IDF
import matplotlib.pyplot as plt
from SALib.analyze import morris
from SALib.plotting.morris import horizontal_bar_plot
import SampleGenerate
import IDFGenerate
import OutputCollect
import RankTransform
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression as LR
import PRCC
dirname, filename = os.path.split(os.path.abspath(__file__))
os.chdir(dirname)
bld_func = 'Hotel'
weather_file = 'CHN_Shanghai.Shanghai.583620_CSWD.epw'
sampling_method    = 'LHS'
LHS_sample_num = 3000
Morris_sample_num = 300
#generate samples
if sampling_method == 'LHS':
    ds = SampleGenerate.sampling_LHS(num = LHS_sample_num)
elif sampling_method == 'Morris':
    ds = SampleGenerate.sampling_Morris(num = Morris_sample_num)
……..
## Sensitity analysis
## Morris method
if sampling_method == 'Morris':
```

```
        ds = pd.read_excel('param_values_morris_real_CR.xlsx')
        ds1 = (ds - ds.mean()) / ds.std()
        param_values = ds1.values
        problem = SampleGenerate.Morris_problem
        Si = morris.analyze(problem, param_values, output, conf_level=0.95,
                                    print_to_console=True,
                                    num_levels=8)
        # Returns a dictionary with keys 'mu', 'mu_star', 'sigma', and 'mu_star_conf'
        # e.g. Si['mu_star'] contains the mu* value for each parameter, in the
        # same order as the parameter file
        fig, (ax1, ax2) = plt.subplots(1, 2)
        horizontal_bar_plot(ax1, Si, {}, sortby='mu_star', unit=r"kWh/m^2")
## regression method
elif sampling_method == 'LHS':
        ds = pd.read_excel('param_values_LHS_real_CR.xlsx')
        ds['output'] = output
        # rank transformtion
        for i in ds.columns:
            x = ds[i].copy()
            ds[i] = RankTransform.transform(x)
        # standardize
        scaler = StandardScaler()
        ds_scaler = scaler.fit_transform(ds)
        ## calculate SRRC
        X=ds_scaler.loc[:,:-1]
        y=ds_scaler.loc[:,-1]
        linreg = LR()
        model=linreg.fit(X, y)
        SRRC = linreg.coef_
        plt.figure()
        plt.bar(x = ds.columns, height = SRRC)
        ## calculate PRCC
        PRCC_matric = PRCC.partial_corr(ds.values)
        PRCC = list(PRCC_matric[:-1,-1])
```

```
    plt.figure()
    plt.bar(x = ds.columns, height = PRCC)
```

#采样程序

```
###generate sample for sensitivity analysis using both HL and morris method
import sys
import os
import pandas as pd
import numpy as np
from SALib.analyze import morris
from SALib.sample.morris import sample
from smt.sampling_methods import LHS
dirname, filename = os.path.split(os.path.abspath(__file__))
os.chdir(dirname)
# name of initial variables
names = ['NWWR', 'SWWR', 'EWWR', 'WWWR', 'AREA', 'NL', 'CR', 'WALLU',
                    'WSP', 'WSA', 'RU', 'RSA', 'WINU', 'SHGC', 'SPC', 'SPH', 'LPD','OPD', 'INFIL',
'FLT', 'GLT', 'CLT', 'ST']
……
# Morris sampling
def Morris_problem():
    problem = {
      'num_vars': len(names),
      'names': names,
      'groups': None,
      'bounds': bounds
     }
    return problem
def Morris_sampling(num_levels = 8, N_morris=300):
    problem = Morris_problem()
    param_morris = sample(problem, N=N_morris, num_levels=num_levels,
                                optimal_trajectories=8)
    df_param_morris = pd.DataFrame(param_morris, columns = problem['names'])
```

```
    df_param_morris.to_excel("param_values_morris.xlsx" )
    return df_param_morris
# Latin Hypercube sampling
def sampling_LHS(num = 3000):
    xlimits = np.array(bounds)
    sampling = LHS(xlimits=xlimits)
    N_LHS=num
    param_LHS = sampling(N_LHS)
    df_param_LHS = pd.DataFrame(param_LHS, columns = names)
    df_param_LHS.to_excel("param_values_LHS.xlsx" )
     return df_param_LHS
```

#模型生成程序

```
import sys
import os
import pandas as pd
import numpy as np
from eppy import modeleditor
from eppy.modeleditor import IDF
import math
def Layer_Allocation(area,nl,func,A0):
    zone = pd.DataFrame(columns = ['layer','function','area', 'ratio'])
    Ar = A0       #本次划分时该楼层的剩余面积
    j = 1; n = 0
    r = 0
    for i in range(len(func.index)):
        sub_area = area * func.iat[i,1]
        while j<=nl:
            if sub_area == 0:
                break
            if Ar <= sub_area:
                r = r + round(Ar/A0,4)
                zone.loc[n] = [j, func.iat[i,0], Ar, r]
                n = n + 1; j = j + 1; r = 0
```

```
                    sub_area = sub_area - Ar
                    Ar = A0
                    continue
                else:
                    r = r + round(sub_area/A0,4)
                    zone.loc[n] = [j, func.iat[i,0], sub_area, r]
                    n = n + 1
                    Ar = Ar - sub_area
                    break
    for i in range(len(zone.index)):
        if zone.iat[i,3] > 0.95 : zone.iat[i,3] = 1
    return zone
……..
def generate_idf(NWWR, SWWR, EWWR, WWWR, AREA, NL, CR, WALLU,
                 WSP, WSA, RU, RSA, WINU, SHGC, SPC, SPH, LPD,OPD, INFIL, FLT, GLT,
CLT, ST, func):
    iddfile = "Energy+.idd"
    IDF.setiddname(iddfile)
    fname1 = "E+RefModel/Ref_Model.idf"
    IDF.setiddname(iddfile)
    idf_T1 = IDF(fname1)
     #add zones
    A0 = round(AREA / NL, 4)
    zone0 = Layer_Allocation(AREA,NL,func,A0)
    zone = pd.DataFrame(columns = ['layer','function','area','ratio'])
    for j in range(len(zone0.index)):
        if zone0.iat[j,2]/A0 > 0.05:
            zone = zone.append(zone0.loc[j,:])     #去掉面积太小的区域
    zone_idf = idf_T1.idfobjects['ZONE']
    for j in range(len(zone.index)):
        idf_T1.newidfobject('ZONE')
        zone_idf[-1].Name = zone.iat[j,1] + '_' + str(zone.iat[j,0])
    #relate building compactness to geometry type 1-5
```

```
#把每层等分成 16 小块，排列组合得到不同的形状，每个形状的特征参数 sigma 是：一层周长/sqrt（一层面积/16）
Geo_type_p = [16,20,26,29,34]
sigma = (CR*A0*NL - A0) / ((3.3*NL) * (A0/16)**0.5)
delta = [abs(c-sigma) for c in Geo_type_p]
Geo_type = delta.index(min(delta)) + 1
    CR_real.append(((A0/16)**0.5 * Geo_type_p[Geo_type-1] * 3.3 * NL + A0) / AREA)
……
return(idf_T1,CR_real)
```

附录 C 模拟与实测数据融合算法代码节选

```
#模拟数据修正实测数据
import pandas as pd
import os
import numpy as np
import matplotlib as plt
dirname = r'E:\01-Phd\01-SimulationEngine\02-Process\03-Integration\TestwithSimulation'
os.chdir(dirname)
data_field = pd.read_csv('HotelChillerElectricitywithNoise.csv')
data_simu = pd.read_csv('HotelChillerElectricityNoNoise_simulated.csv')
data_obs_para = pd.read_excel('ObservationParameter.xlsx')
def Outlier_Remove(data_outlier,data_smooth):
    k = 1
    for i in range(len(data_outlier)-1):
        delta = data_outlier[i+1] - data_outlier[i]
        delta_base = data_smooth[i+1] - data_smooth[i]
        if abs(delta) > data_outlier[i] * 0.4: #data_outlier[i] * 0.7:
            k = data_outlier[i]/data_smooth[i]
            data_outlier[i+1] = data_outlier[i] + delta_base * k
    return data_outlier
def Noise_Remove(data_noise, data_smooth, N):
    k = 1
    A = []
    for i in range(1,N):
        A.append(i)
    for i in range(len(data_noise)-N):
        k = data_noise[i]/data_smooth[i]
        mean_noise = sum(data_noise[i:i+N]) / N
        mean_smooth = sum(data_smooth[i:i+N]) / N
        data_noise[i] = mean_noise + (data_smooth[i] - mean_smooth) * k
    return data_noise
```

```
def get_sample(df):
    frames = []
    for i in range(24):
        frames.append(df[i*10+2 : i*10+3])
    df_sampled = pd.concat(frames)
    return df_sampled
data_processed = pd.DataFrame(columns = ['Outlier_Removed','Noise_Removed'])
data_outlier_removed = Outlier_Remove(data_field.iloc[:,0],data_simu.iloc[:,0])
data_processed['Outlier_Removed'] = data_outlier_removed
data_noise_removed = Noise_Remove(data_outlier_removed,data_simu.iloc[:,0],6)
data_processed['Noise_Removed'] = data_noise_removed
data_field.iloc[:,0] = data_outlier_removed #data_noise_removed
data_field = get_sample(data_field.join(data_obs_para))

#关键变量推测
import numpy as np
import pystan as ps
import pickle as pk
from matplotlib import pyplot as plt
import seaborn as sns
import os
import pandas as pd
# read in field and computer simulation data
os.chdir(r'E:\01-Phd\01-SimulationEngine 20200312\02-Process\04-
Integration\TestwithSimulation\daily\5_para_to_be_cal')
DATACOMP = np.genfromtxt('DATACOMP_5p_20.csv', delimiter = ',', skip_header=1)
DATAFIELD = np.genfromtxt('DATAFIELD_5p_20.csv', delimiter = ',', skip_header=1)
y = DATAFIELD[:, 0] # observed output
xf = DATAFIELD[:, 1:] # observed input
(n, p) = xf.shape
eta = DATACOMP[:, 0] # simulation output
xc = DATACOMP[:, 1:(p+1)] # simulation input
tc = DATACOMP[:, (p+1):] # calibration parameters
(m, q) = tc.shape
```

```
x_pred = xf # design points for predictions
n_pred = x_pred.shape[0] # number of predictions
# standardization of output y and eta
eta_mu = np.nanmean(eta) # mean value
eta_sd = np.nanstd(eta) # standard deviation
y = (y - eta_mu) / eta_sd
eta = (eta - eta_mu) /eta_sd
# Put design points xf and xc on [0,1]
x = np.concatenate((xf,xc), axis=0)
x_min = np.nanmin(x, axis=0)
x_max = np.nanmax(x, axis=0)
xf = (xf - x_min) / (x_max - x_min)
xc = (xc - x_min) / (x_max - x_min)
x_pred = (x_pred - x_min) / (x_max - x_min)
# Put calibration parameters t on domain [0,1]
tc_min = np.nanmin(tc, axis=0)
tc_max = np.nanmax(tc, axis=0)
tc = (tc - tc_min) / (tc_max - tc_min)
# create data as dict for input to Stan
stan_data = dict(n=n, m=m, n_pred=n_pred, p=p, y=y, q=q,
 eta=eta, xf=xf, xc=xc, x_pred=x_pred, tc=tc)
# run model in stan
model = ps.StanModel(file=r"E:\01-Phd\01-SimulationEngine\02-Process\04-
Integration\TestwithSimulation\daily\bcWithoutPred.stan")
fit = model.sampling(data=stan_data, iter=2000, chains=3, n_jobs=1)
# plot traceplots and posterior probability histograms
fit.plot()
summary_dict = fit.summary()
df = pd.DataFrame(summary_dict['summary'],
                                columns=summary_dict['summary_colnames'],
                                index=summary_dict['summary_rownames'])
# extract predictions, excluding warm-up and
# extract posterior distibution of calibrating parameters
```

```
tf = fit.extract()["tf"]
tf = tf * (tc_max - tc_min) + tc_min
sns.set_style('darkgrid')
sns.distplot(tf[:,2])
```

附录 D 数据融合算法验证案例设备清单

表 D-1 冷机设备清单

所述系统	设备名称	品牌	规格型号	数量	制冷量	电机功率
				台	（*KW*）	（*KW*）
制冷机房	螺杆式冷水机组	特灵	RTHC1D1R0E0D1L3E1LFR0000	3	1050	196
制冷机房	螺杆式冷水机组	特灵	RTWD120C3A02C3B2BA2AN1A0N	1	439.2	93

表 D-2 水泵设备清单

所述系统	设备名称	品牌	规格型号	数量	额定参数	电机功率
				台		（*KW*）
制冷机房	冷冻水泵	BELL&GOSSETT	5BC-8.875-BF	4	杨程 12*m*，流量 180m^3/h	7.5
制冷机房	冷冻水泵	BELL&GOSSETT	6E-10.750-BF	3	杨程 20*m*，流量 270m^3/h	22
制冷机房	冷却水泵	BELL&GOSSETT	VSC-11.500-BFRHR	3	杨程 25*m*，流量 330m^3/h	29.4
制冷机房	冷却塔	-	-	3	流量 195m^3/h	5.5

表 D-3 空调末端设备清单

所述系统	设备名称	品牌	数量	额定风量	电机功率
			台	（m^3/h）	（*KW*）
主楼	空调箱	新晃	2	3600	2
主楼	空调箱	新晃	1	4000	2
主楼	空调箱	新晃	1	4500	2.5
主楼	空调箱	新晃	3	5000	2.5
主楼	空调箱	新晃	1	6000	3

主楼	空调箱	新晃	3	7000	3
主楼	空调箱	新晃	5	8000	3.5
主楼	空调箱	新晃	1	9000	4
主楼	空调箱	新晃	1	15000	4.5
主楼	空调箱	新晃	1	20000	5
主楼	空调箱	新晃	1	21500	5.5
主楼	空调箱	新晃	1	25000	6
主楼	空调箱	风机盘管	821	600	0.062

附录 E 关联规则库（频繁项数=2）

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	length
frozenset({'Area_2'})	frozenset({'CTF_2'})	0.43	0.74	0.35	0.80	1.08	0.03	1.30	2
frozenset({'Area_2'})	frozenset({'ChillerType_0'})	0.43	0.91	0.43	1.00	1.10	0.04	inf	2
frozenset({'Area_2'})	frozenset({'FanEff_2'})	0.43	0.78	0.35	0.80	1.02	0.01	1.09	2
frozenset({'COP_1'})	frozenset({'ChillerType_0'})	0.30	0.91	0.30	1.00	1.10	0.03	inf	2
frozenset({'COP_2'})	frozenset({'ChillerType_0'})	0.48	0.91	0.43	0.91	1.00	0.00	0.96	2
frozenset({'CTF_2'})	frozenset({'ChillerType_0'})	0.74	0.91	0.74	1.00	1.10	0.06	inf	2
frozenset({'ChillerType_0'})	frozenset({'CTF_2'})	0.91	0.74	0.74	0.81	1.10	0.06	1.37	2
frozenset({'CTF_2'})	frozenset({'FanEff_2'})	0.74	0.78	0.61	0.82	1.05	0.03	1.23	2
frozenset({'CTF_2'})	frozenset({'INFIL_1'})	0.74	0.74	0.65	0.88	1.19	0.11	2.22	2
frozenset({'INFIL_1'})	frozenset({'CTF_2'})	0.74	0.74	0.65	0.88	1.19	0.11	2.22	2
frozenset({'LPD_2'})	frozenset({'CTF_2'})	0.61	0.74	0.57	0.93	1.26	0.12	3.65	2
frozenset({'NL_B_1'})	frozenset({'CTF_2'})	0.39	0.74	0.35	0.89	1.20	0.06	2.35	2
frozenset({'CTF_2'})	frozenset({'SPC_1'})	0.74	0.70	0.61	0.82	1.18	0.09	1.72	2
frozenset({'SPC_1'})	frozenset({'CTF_2'})	0.70	0.74	0.61	0.88	1.18	0.09	2.09	2
frozenset({'WSysType_1'})	frozenset({'CTF_2'})	0.43	0.74	0.39	0.90	1.22	0.07	2.61	2
frozenset({'FanEff_2'})	frozenset({'ChillerType_0'})	0.78	0.91	0.70	0.89	0.97	-0.02	0.78	2
frozenset({'INFIL_1'})	frozenset({'ChillerType_0'})	0.74	0.91	0.70	0.94	1.03	0.02	1.48	2
frozenset({'LPD_1'})	frozenset({'ChillerType_0'})	0.30	0.91	0.30	1.00	1.10	0.03	inf	2
frozenset({'LPD_2'})	frozenset({'ChillerType_0'})	0.61	0.91	0.61	1.00	1.10	0.05	inf	2

frozenset({'LR_0'})	frozenset({'ChillerType_0'})	0.39	0.91	0.39	1.00	1.10	0.03	inf	2
frozenset({'NL_A_3'})	frozenset({'ChillerType_0'})	0.35	0.91	0.35	1.00	1.10	0.03	inf	2
frozenset({'NL_B_1'})	frozenset({'ChillerType_0'})	0.39	0.91	0.35	0.89	0.97	-0.01	0.78	2
frozenset({'NL_B_2'})	frozenset({'ChillerType_0'})	0.43	0.91	0.43	1.00	1.10	0.04	inf	2
frozenset({'OD_2'})	frozenset({'ChillerType_0'})	0.65	0.91	0.57	0.87	0.95	-0.03	0.65	2
frozenset({'PumpEff_1'})	frozenset({'ChillerType_0'})	0.30	0.91	0.30	1.00	1.10	0.03	inf	2
frozenset({'PumpEff_2'})	frozenset({'ChillerType_0'})	0.61	0.91	0.57	0.93	1.02	0.01	1.22	2
frozenset({'SPC_1'})	frozenset({'ChillerType_0'})	0.70	0.91	0.70	1.00	1.10	0.06	inf	2
frozenset({'WSysType_1'})	frozenset({'ChillerType_0'})	0.43	0.91	0.39	0.90	0.99	-0.01	0.87	2
frozenset({'WSysType_2'})	frozenset({'ChillerType_0'})	0.39	0.91	0.39	1.00	1.10	0.03	inf	2
frozenset({'INFIL_1'})	frozenset({'FanEff_2'})	0.74	0.78	0.61	0.82	1.05	0.03	1.23	2
frozenset({'LPD_2'})	frozenset({'FanEff_2'})	0.61	0.78	0.57	0.93	1.19	0.09	3.04	2
frozenset({'LR_1'})	frozenset({'FanEff_2'})	0.35	0.78	0.35	1.00	1.28	0.08	inf	2
frozenset({'NL_A_2'})	frozenset({'FanEff_2'})	0.39	0.78	0.35	0.89	1.14	0.04	1.96	2
frozenset({'NL_B_1'})	frozenset({'FanEff_2'})	0.39	0.78	0.39	1.00	1.28	0.09	inf	2
frozenset({'PumpEff_2'})	frozenset({'FanEff_2'})	0.61	0.78	0.57	0.93	1.19	0.09	3.04	2
frozenset({'WSysType_1'})	frozenset({'FanEff_2'})	0.43	0.78	0.39	0.90	1.15	0.05	2.17	2
frozenset({'LPD_2'})	frozenset({'INFIL_1'})	0.61	0.74	0.57	0.93	1.26	0.12	3.65	2
frozenset({'NL_B_1'})	frozenset({'INFIL_1'})	0.39	0.74	0.35	0.89	1.20	0.06	2.35	2
frozenset({'SPC_1'})	frozenset({'INFIL_1'})	0.70	0.74	0.57	0.81	1.10	0.05	1.39	2
frozenset({'WSysType_1'})	frozenset({'INFIL_1'})	0.43	0.74	0.39	0.90	1.22	0.07	2.61	2
frozenset({'LPD_2'})	frozenset({'SPC_1'})	0.61	0.70	0.52	0.86	1.23	0.10	2.13	2
frozenset({'WSysType_1'})	frozenset({'LPD_2'})	0.43	0.61	0.39	0.90	1.48	0.13	3.91	2
frozenset({'LR_0'})	frozenset({'SPC_1'})	0.39	0.70	0.39	1.00	1.44	0.12	inf	2

frozenset({'NL_A_2'})	frozenset({'OD_2'})	0.39	0.65	0.35	0.89	1.36	0.09	3.13	2
frozenset({'NL_B_2'})	frozenset({'OD_2'})	0.43	0.65	0.35	0.80	1.23	0.06	1.74	2
frozenset({'WSysType_1'})	frozenset({'PumpEff_2'})	0.43	0.61	0.35	0.80	1.31	0.08	1.96	2
frozenset({'WSysType_1'})	frozenset({'SPC_1'})	0.43	0.70	0.39	0.90	1.29	0.09	3.04	2

个人简历、在读期间发表的学术论文与研究成果

个人简历：

沙华晶，女，1989 年 8 月生

2007.9~2011.6，南京工业大学，建筑环境与设备工程专业，学士

2011.9~2014.3，同济大学，供热、供燃气、通风及空调工程专业，硕士

2014.4~2015.6，奥雅纳工程咨询（上海）有限公司，绿色建筑工程师

2015.7~2017.8，双良节能股份有限公司，研发工程师

2017.9~至今，同济大学，博士研究生在读

已发表论文：

SCI:

[1] **Huajing Sha**, Peng Xu, Meishun Lin, Cheng Peng, Qiang Dou. Development of a multi-granularity energy forecasting toolkit for demand response baseline calculation. Applied Energy (Accepted). (JCR Q1/IF: 9.1)

[2] **Huajing Sha**, Peng Xu, Zhiwei Yang, Yongbao Chen, Jixu Tang. Overview of computational intelligence for building energy system design. Renewable & Sustainable Energy Reviews 2019. (JCR Q1/IF: 12.3)

[3] **Huajing Sha**, Peng Xu, Chonghe Hu, Zhiling Li, Zhe Chen, Yongbao Chen. A simplified HVAC energy prediction method based on degree-day. Sustainable Cities and Society 2019. (JCR Q2/IF: 5.268)

[4] Yongbao Chen,Zhe Chen,Peng Xu, Weilin Li, **Huajing Sha**, Zhiwei Yang, Guowen Li, Chonghe Hu. Quantification of electricity flexibility in demand response: Office building case study. Energy 2019. (JCR Q1/IF: 6.082)

会议:

[1] **Huajing Sha**, Peng Xu, Zhiwei Yang. IFC based semi-automated design tool for HVAC central system: A general framework. IOP Conf. Ser. Earth Environ. Sci., 2019. （012074）

[2] **Huajing Sha**, Peng Xu, Meishun Lin, Cheng Peng, Qiang Dou. A Framework of hybrid building energy forecasting model. Engineering Procedia（Accepted）.

中文CSCD：

[1] **沙华晶**，许鹏，钟志文，李云飞. 建筑空调能耗关键变量通用提取方法及工具的开发. 土木与环境工程学报(中英文)（录用）

其他：

[2] 陈智博，**沙华晶**，许鹏，奚培峰. 中国公共建筑的建筑典型模型建立. 节能改造与技术，2020年第2期.

发明专利：

用于集中式空调系统自动设计的信息处理方法与流程（实审中），公开号：109408884A

参与学术会议：

[1] 2018 年 12 月, 4th Asia Conference of International-Building-Performance-Simulation-Association (ASIM) 香港，中国

[2] 2019 年 4 月，暖通空调模拟学术年会，成都，中国

[3] 2019 年 7 月，第一届华人能源与人工环境国际学术会议，成都，中国

[4] 2020 年 12 月，12 th International Conference on Applied Energy, Bangkok/Virtual

研究及项目经历：

[1] 建筑大数据分析方法（横向），主要参与人

[2] 数字化建筑负荷模型研究（横向），主要参与人

[3] 基于大数据挖掘的建筑能效与故障诊断诊断平台研究（横向），主要参与人

[4] 基于动态特性的建筑遮阳节能效果评价标准算法的开发（横向），主要参与人

[5] 应对气候改变的低碳技术的适应性分析（纵向），主要参与人

[6] 基于 BIM 的绿色建筑运营优化关键技术研发（国家重点研发项目），参与人

[7] 空调产品基准模型分析（横向），参与人

[8] 空调等制冷产品节能技术及潜力分析（横向），参与人

获奖经历：

[1] 2017，上海市标准化优秀学术成果二等奖《基于动态特性的建筑遮阳节能效果评价标准算法的开发》

[2] 2019，博士生国家奖学金