同济大学
TONGJI UNIVERSITY

硕士学位论文

基于多源异构数据的办公建筑能耗预测方法

同济大学

姓　　名：郭明月

学　　号：1930255

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

二〇二二年三月

同济大学

同济大学
TONGJI UNIVERSITY

A dissertation submitted to

Tongji University in conformity with the requirements for

the degree of Master of Engineering

Energy consumption prediction method for office buildings based on multi-source heterogeneous data

同济大学

Candidate: Mingyue Guo

Student Number: 1930255

School/Department: School of Mechanical Engineering

Discipline: Civil Engineering

Major: Heat, Gas Supply, Ventilation and Air-conditioning Engineering

Supervisor: Prof. Peng Xu

March, 2022

同济大学

基于多源异构数据的办公建筑能耗预测方法

郭明月　同济大学

同济大学

同济大学

摘要

全球气候变暖严重威胁着人类的生存，全面降低能耗和碳排放、实现碳中和的目标已迫在眉睫。全球和我国的建筑业相关能耗大约占到了35%，降低建筑行业的能耗对全社会的节能事业至关重要，而建筑能耗预测是建筑节能和建筑与电网的供需匹配中不可或缺的部分。目前建筑能耗预测模型主要有白箱、黑箱、灰箱三种类型，这三种模型各有优缺点。白箱模型可解释性高但建模复杂，存在较多假设，输入参数过多，且某些输入参数数值无法测量或获取，往往与实际能耗相去甚远。黑箱模型学习了输入特征与历史能耗之间的关系，预测精度较高，但需要大量的历史数据进行训练，可解释性低，且模型的通用性不高，无法进行在完全无历史能耗时的跨建筑预测。现有的灰箱模型也无法很好的解决数据不足和模型通用性的问题。针对前述的输入变量多且难以获取以及无法进行跨建筑能耗预测的问题，本文提出了基于多源异构数据的办公建筑能耗预测方法，实现了多种场景下的能耗预测，尤其是能够进行无历史能耗场景下的跨建筑能耗预测。本文提出的预测方法包括以下步骤：关键变量的提取、关键变量缺失值的推测、多源数据的预处理和融合、多源异构数据库和能耗预测混合模型的建立。

首先，从负荷和机电系统两个方面选取了众多可能影响能耗的初始变量，从确定初始变量的取值范围并抽样，用python语言和其eppy库构建了快速建模工具批量生成算例，由此得到初始变量-能耗数据集，利用相关系数法（SRCC 和PRCC）、Morris法和XGBoost算法进行了对能耗有显著影响的关键变量的提取。

其次，针对实际情况中存在的关键变量值缺失的情况，进行了不同情景下的关键变量值推测。在存在历史能耗的情况下，采用遗传算法进行关键变量缺失值的推测，这种推测方法可用于数据库建立时的关键变量推测。在无历史能耗的情况下，采用 Apriori 关联规则挖掘算法进行关键变量缺失值的推测，这种推测方法适用于在进行无历史能耗的建筑能耗预测时的关键变量推测。对两种方法的推测准确度进行了校验和分析，关键变量缺失值推测的误差均在可接受范围内。

再次，搜集了来自能耗监测平台、节能审计报告、模拟三种来源的数据。对其缺失的关键变量信息进行了推测，对其能耗数据，根据不同的数据特点进行了预处理。对于来自能耗监测平台的能耗数据，对其进行了异常值和缺失值的处理。对来自节能审计报告的能耗数据，利用 K-Means 算法聚类得到的典型能耗曲线对其进行了颗粒度的细化。对于来自能耗模拟的数据，根据实测数据和模拟数据的偏差建立了两级模拟数据修正模型对模拟数据进行了修正，使其更接近于实测

数据。

最后，在上两步的基础上建立了多源异构数据库，并利用数据库进行了能耗预测混合模型的训练。多源异构数据库的输入为描述建筑物理信息的关键变量和气象参数、时序特征、能耗模拟修正值等其他变量，输出为能耗模型的预测目标，如空调制冷能耗。采用在能耗预测领域表现很好的组合树算法——lightGBM 进行混合模型的建模，并采用交叉验证的方式得到泛化误差较小的最佳超参数。利用实际建筑中的数据，对混合模型进行了确定性和不确定性的验证。

关键词：跨建筑能耗预测，数据驱动模型，机器学习，混合能耗预测模型

同济大学

ABSTRACT

Global warming becomes a serious threat to human existence, and the goal of reducing energy consumption and carbon emissions and achieving carbon neutrality is imminent. The energy consumption related to buildings in the world and in China accounts for about 35%. Reducing the energy consumption of buildings is crucial to the energy conservation of the whole society. The prediction of building energy consumption is the key to building energy conservation and the matching of supply and demand between buildings and the power grid. At present, there are three main types of mainstream methods for building energy consumption prediction: white box, black box, and gray box. Each of these three models has advantages and disadvantages. The white-box model has high interpretability, but the modeling is complex and there are many assumptions too many input parameters. Besides, the values of some input parameters cannot be measured or obtained. So, the simulation result is often far from the actual energy consumption. The black-box model learns the relationship between input features and historical energy consumption, and the prediction accuracy is high, but it requires a large amount of historical data for training, the interpretability is low, and the generality of the model is not high, and it cannot be carried out without historical energy consumption at all. The existing gray-box models also cannot solve the problem of data and model generality very well. In view of the aforementioned problems of too many input variables some of which may be difficult to obtain, and unable to used for cross-building energy consumption prediction, this paper proposes an office building energy consumption prediction method based on multi-source heterogeneous data. It can realize energy consumption prediction in various scenarios, especially for cross-building without historical energy consumption scenarios. It includes four steps: reprocessing, fusion of multi-source data, establishment of multi-source heterogeneous database and energy consumption prediction hybrid model

First, many initial variables that may affect energy consumption are selected from the two aspects of load and electromechanical system, and the value range of the initial variables is determined and sampled. A rapid modeling tool is built with python language and its eppy library to generate calculation examples in batches. A dataset of initial variables and their corresponding energy consumption was generated, and key

variables that had a significant impact on energy consumption were extracted using correlation coefficient method (SRCC and PRCC), Morris method and XGBoost algorithm.

Secondly, in view of the fact that the key variable value is missing in practice, the key variable value estimation under different scenarios is carried out. In the presence of historical energy consumption, the genetic algorithm is used to infer the missing values of key variables. This inference method is suitable for the inference of key variables when the database is being established. In the case of no historical energy consumption, the Apriori association rule mining algorithm is used to infer the missing values of key variables. This inference method is suitable for the prediction of key variables when predicting building energy consumption without historical energy consumption. The inference accuracy of the two methods was checked and analyzed, and the errors in the inference of missing values of key variables were all within the acceptable range.

Thirdly, data from three sources (energy consumption monitoring platform, energy saving audit report, and energy consumption simulation) are collected. The missing key variable information is speculated, and the energy consumption data is preprocessed according to different data characteristics. For the energy consumption data from the energy consumption monitoring platform, outliers and missing values are imputed. For the energy consumption data from the energy-saving audit report, the typical energy consumption curve obtained by clustering using K-Means algorithm is used to refine the granularity. For the data from the energy consumption simulation, a two-level simulation data correction model is established according to the deviation of the measured data and the simulated data, and the simulated data is corrected to make it closer to the measured data.

Finally, a multi-source heterogeneous database is established on the basis of the previous two steps, and the hybrid model of energy consumption prediction is trained using this database. The input of the multi-source heterogeneous database is the key variables describing the physical information of the building and other variables such as meteorological parameters, time series characteristics, energy consumption simulation correction value, etc., and the output is the prediction target of the energy consumption model, such as the energy consumption of air conditioning and refrigeration. The combination tree algorithm, lightGBM, which has a good performance in the field of energy consumption prediction, is used to develop the hybrid

model, and the optimal hyperparameters with small generalization error are obtained by cross-validation. The performance of the hybrid model is verified with the test data.

Key Words: Cross-building energy prediction, data-driven model, machine learning, hybrid energy prediction model

同济大学

目录

同济大学

同济大学

同济大学

第 1 章 引言

1.1 研究背景及意义

1.1.1 研究背景

根据世界气象组织发布的《2020 年全球气候状况》，近年来主要温室气体的浓度均在持续上升，二氧化碳的全球平均摩尔分数在 2021 年将达到或超过 414ppm，严重威胁着人类的生存[1]。要减小气候变化带来的影响，将全球温升控制在工业化前基线的 1.5℃以内，就迫切要求我们到 2050 年实现净零碳排放。建筑领域的能源消耗和碳排放在总体能源消耗和碳排放中占比较大，根据国际能源署的核算结果，2018 年全球建筑业相关的能耗占全球能耗的 35%；清华大学建筑节能中心的核算结果表示，2018 年我国建筑业相关能耗占中国全社会总能耗的 37%[2]。建筑绝大部分的能耗发生在建筑建造和建筑运行过程中，2018 年我国建筑建造能耗占中国全社会能耗的比例为 14%，建筑运行能耗占中国全社会能耗的比例为 23%。可以看出，降低建筑运行能耗可非常显著地降低全社会能耗。而在建筑能耗中，办公建筑面积及能耗占比均较大。从 2020 年上海市公共建筑能耗监测平台能耗数据分析来看，上海市国家机关办公建筑和办公建筑能耗在能耗监测平台联网建筑年用电量中占比最大，共计 32.1%，其占比如图 1 所示[3]。由此可见，实现办公建筑的节能是建筑节能事业中的重中之重。

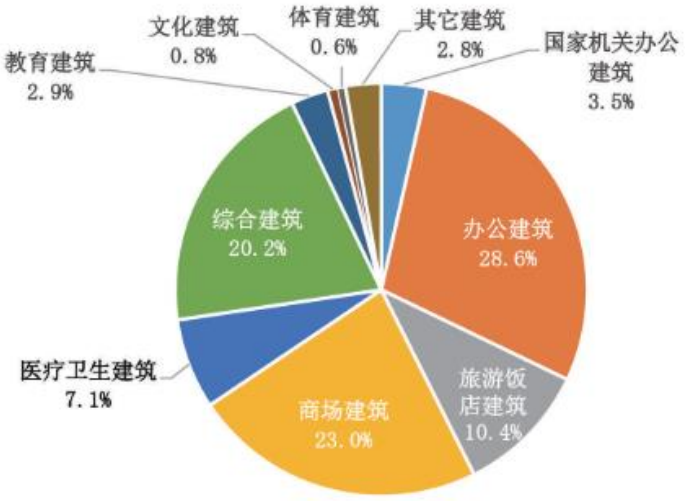


图 1.1 2020 年上海市各类建筑在能耗监测平台中年用电量占比情况[3]

为进行建筑用能监管和建筑节能，自 2007 年住建部、财政部发布《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》以来，各个省市陆续开展了公共建筑能耗监测平台建设工作[4][5]。截止到 2020 年 12 月 31 日，上海市共计 2017 栋公共建筑均安装了用能分项计量装置并实现了与能耗监测平台的联网，其中办公建筑占比最大，国家机关办公建筑和办公建筑的能耗共占所有建筑的 32.1%[3]。公共建筑能耗监测平台已积累了大量的建筑实时运行参数和分项能耗数据，为建筑能耗预测提供了数据基础。建筑能耗数据在绕日常管理、优化运行和节能改造等方面均有较大应用价值[6]，公共建筑能耗监测平台的主要应用包含以下几个方面：

1) 便于建筑节能的宣传，协助主管部门进行建筑能耗管理。

2) 推进建筑节能工作，有利于建筑节能改造和优化运行，为建筑用能标准和节能技术路线的制定提供依据，便于统计年度用能情况。

3) 利用建筑能耗数据进行深入分析，进行能耗预测，用能诊断，建筑需求响应等。

本课题研究内容属于上述的第三类应用，综合利用办公建筑数据进行多场景下的能耗预测。虽然建筑能耗平台数据量较为可观，但由于传感器故障、数据传输故障等原因，在数据质量上不够理想。办公建筑能耗数据难以利用的原因有二：第一，存在着较多缺失和异常数据，比如（1）长时间的死值；（2）单点数据异常；（3）分项数据之和与总用电量不相符；（4）某段时间数据缺失等情况。第二，能耗监测平台收集记录的主要是建筑运行的动态数据，而建造年份、窗墙比、围护结构热工参数、设备信息等基本的建筑静态物理信息无法获取，从而阻碍了建筑能耗的进一步分析。

除公共建筑能耗监测平台外，建筑能耗相关信息还可以从能源审计报告中获取。建筑能源审计报告数据的特点是：第一，数据颗粒度低，大多为逐月能耗账单数据，但数据更加真实可靠；第二，审计报告中有较为详细的建筑静态数据，如房间功能占比、窗墙比、围护结构热工参数、设备参数等物理信息。

目前虽然有大量的建筑用能数据，但进行建筑能耗预测往往依旧采用复杂且过于理想化的物理模型或只用该建筑的历史能耗建立数据驱动模型，海量数据并没有得到充分的利用。在当前的技术背景下，对于没有能耗监测系统的老旧建筑或没有历史数据的新建建筑无法建立数据驱动模型，如何将上述不同来源、不同特点的异构数据进行综合应用并进行跨建筑的能耗预测成为了亟待研究的课题。

1.1.2 研究意义

本文主要研究任务为办公建筑能耗预测，综合应用了多种来源的建筑信息和数据，采用静态物理数据和动态运行数据相结合，物理模型和数据驱动模型相结合的混合模型，可实现建筑不同数据条件下的能耗预测。

本节将从建筑能耗预测的意义，多源数据综合应用的意义和跨建筑预测的意义这三个方面来说明本课题的研究意义。

1) 建筑能耗预测的意义

建筑能耗绝大部分来自于建筑运行阶段的能耗，而建筑运行能耗大部分来自通风空调系统的运行能耗，大型公共建筑中的中央空调能耗占建筑能耗的40~60%[7]，建筑负荷和能耗的精确预测能够很大程度地降低建筑运行能耗，根据相关研究，不同水平的建筑约有 15~30%的节能潜力[8]。

在不同场景下，建筑能耗预测均可助力建筑节能。在设计阶段，建筑负荷预测有助于工程师在全负荷区间下进行设备选型，设备台数的确定等，使设计更加合理与节能。在运维阶段，大量研究表明合理的暖通空调系统的控制减少建筑13%~28%的用能量[9][10]，准确的建筑能耗预测可用于优化控制策略进行前馈式控制[11]，优化设备控制策略以提高运行能效，并且还可结合预测结果和实时运行数据进行设备维保时长的预判和故障诊断，保障系统节能可靠运行。在既有建筑节能改造时，能耗预测对于评估节能改造效果也发挥着重要的作用，由于处于改造前设备条件下的建筑能耗无法获取，要计算改造前后的节能量，只能通过能耗预测的方法获取改造前的能耗。随着可再生能源的发展，电网供应侧的电力来源变得复杂多样，除传统的煤电之外，光电、风电等可再生能源受时间、气候等影响较大，这部分电力稳定性较差，可再生能源的并网将对电网供给侧和需求侧的平衡带来新的挑战，这就要求需求侧能够实现更准确、更精细的逐时或逐日负荷预测，而不仅仅是逐月或逐年能耗总量的预测。可见，建筑能耗预测在建筑节能和建筑与电网的交互过程中意义重大。

2) 多源异构数据综合应用的意义

随着建筑智能化水平和大数据技术的提升，大量的建筑能耗数据可从多个途径获取。建筑数据可以分为两大类，一类是建筑建成后几乎不再变化的静态数据，如建筑面积、建筑穿墙比、围护结构热工参数等物理信息；另一类是建筑逐时能耗等运行动态数据。大部分的能耗监测平台或公开数据库中可获取的信息主要是动态数据，静态数据不够全面，通常只包含建筑类别、面积、层数等信息。

在国外，城市级别的建筑能耗数据库有：美国 BPD（Building performance database）建筑能耗数据库[12]、欧盟 BPIE 建筑信息数据库（Occupant IEQ Survey Reporting Tool）、英国 BUS-PROBE 数据库、爱尔兰 CER Smart Metering Project 数据库[13]等。除政府或科研机构公开的数据集外，也有学者对建筑能耗数据进行

同济大学

了搜集和整理，如 Clayton Miller 等人[14][15]整理的来自美国、英国等国家包含办公、学校、住宅、医疗等多种功能建筑的能耗数据。而在国内，办公建筑能耗相关的数据可获取途径有：由政府、高校等建立的较为大型的公共建筑能耗监测平台；园区或集团建立的能耗监测平台；建筑单体自身建立的能耗监测平台。但只有少量的能耗监测平台向外公布并持续更新数据。根据周浩[16]等人对各省市公共建筑能耗监测平台运行状况的调研结果，除上海市的国家机关办公建筑和大型公共建筑能耗监测中心的 web 平台外仍在更新数据和相关内容外，其余省市均搜索不到公开平台，并且这些平台往往主要是为政府或企业管理服务，并不为科研服务。

除可获取性差之外，数据集和能耗监测平台的数据存在这着数据质量差（普遍存在缺失值和异常值）、数据丰富程度低（建筑对应的静态数据较少）的问题，导致海量数据并没有得到充分利用。与能耗监测平台特点相反的是能源审计报告的数据，随着绿色建筑的提出和绿色建筑评价体系的建立，大部分公共建筑纷纷进行了能源审计和绿色建筑认证，能源审计报告中包含多且详细的能耗相关信息。能源审计报告中通常包含建筑总体的电耗账单数据和详细的建筑静态数据，电耗账单数据置信度高，但时间颗粒度较大，通常是逐月的电耗数据。可见，不同来源的数据丰富程度不同，数据质量不同，时间颗粒度也不同，目前我们虽然搜集了海量的数据，但并没有发挥其重要价值。将上述多源异构数据进行标准化处理，将静态数据和动态数据结合起来有助于进行建筑的用能特点，建筑能耗趋势及节能潜力的分析。

3) 混合模型的意义

能耗预测模型可分为：白箱模型（也称正演模型）、黑箱模型（也称数据驱动模型）和灰箱模型三类。白箱模型和黑箱模型各有优劣。白箱模型可解释性高，基于物理参数和物理公式，通过公式的求解得到温度、能耗等参数。由于白箱模型建模复杂，需输入非常详细的建筑物理参数，模型过于理想化无法涵盖施工、运行等带来的能耗偏差，在大数据背景下，黑箱模型逐渐代替传统的白箱模型受到越来越多学者和工程师的青睐。但常规的黑箱模型通常只能应用于有历史能耗的建筑，根据该建筑历史的用能特点来进行未来的能耗预测，并且对每栋建筑进行预测时需要重新建模，应用场景非常有限。本课题研究的混合模型类似灰箱模型，通过建立已知物理参数和未知物理参数的关联关系、物理参数和能耗数据的关系、融合模拟数据和实际数据来实现同类建筑跨建筑的能耗预测，扩展了能耗预测模型的应用范围。

1.2 国内外文献综述

本小节将从能耗模型、数据驱动模型和跨建筑能耗预测这三个方面介绍国内外的相关研究。

1.2.1 能耗模型综述

能耗预测模型按照模型建立方法可分为白箱模型（也称正演模型，物理模型）、黑箱模型（也称数据驱动模型）、灰箱模型（有的也称其为混合模型）三类[17]。

1) 白箱模型

白箱模型是通过分析负荷影响参数与负荷之间的物理关系，建立输入变量与输入变量之间的物理模型，通过求解物理模型的方法进行负荷预测。大多数能耗模拟软件均属于白箱模型，这种模拟方法需要对建筑部件和系统进行详细的描述，如建筑的几何尺寸、地理位置、围护结构传热特性、人员密度、人员和设备作息表、设备额定参数和运行曲线、空调系统类型和连接关系等。建筑能耗模拟开始于 20 世纪 60 时代中期，多年来发展出了大量的建筑能耗模拟软件[18]。常见的能耗模拟软件包括美国的 BLAST、DOE-2、EnergyPlus，欧洲的 ESP-r，日本的 HASP 和中国的 DeST，也包括 DesignBuilder、Energy-10、eQUEST 等具有成熟用户界面的逐时能耗模拟工具。此外 Dymola、TRNSYS 和 Matlab 也可用于建立建筑能耗模型[19]。

但是白箱模型也有其局限性，主要体现在以下几个方面：

第一，白箱模型往往需要建立建筑的几何模型、机电模型和物理关系，并且需要输入复杂且详细的围护结构、人员作息表、设备人员密度等信息。白箱模型建模需耗费大量的时间和精力，并且有的输入参数的具体数值难以获取，尤其是处于设计阶段的建筑，其输入参数只能依靠相似的建筑经验值进行设置。

第二，基于物理公式的白箱模型存在着大量的假设与简化，并且均是对理想状态进行模型，未考虑施工、运维过程带来的变化和误差，因此白箱模型的模拟结果往往和实际情况相去甚远。

第三，在拥有建筑运行数据后往往会对白箱模型进行校验，但白箱模型的校验往往难以操作，由于可用信息较少且输入参数过多且复杂，在校验过程中往往会出现多解的情况[20]。

2) 黑箱模型

在当今大数据背景下，黑箱模型受到越来越多学者和业界人士的青睐。黑箱模型是在输入变量和输出变量已知或存在较多观测值的情况下，运用统计分析或机器学习的方法，建立输入变量（天气参数、围护结构参数、空调系统参数等）

同济大学

和输出变量（能耗）的数学描述。

黑箱模型常见的工作流为：第一，搜集能耗预测相关数据。第二，数据清洗，包括异常数据处理，缺失值处理等。第三，特征工程，是将原始数据和各种信息转变为模型的输入变量的过程，简单来说就是确定模型的自变量。对于能耗预测这一任务来说，就是确定对建筑能耗有显著影响的变量。第四，算法的选择及模型训练，目前存在着多种算法均可实现能耗预测任务，但没有一种算法在任何场景下的性能都是最佳的，因此工程师因根据实际情况进行算法的选择。采用合适的算法应用前几个步骤建立的数据集进行黑箱模型的训练，训练过程中通常涉及到训练集和测试集的划分，模型超参数的调节等步骤。第五，进行能耗预测，输入相关输入变量值，应用步骤四建立的能够反映输入参数（与能耗相关的特征）和输出参数（能耗）关系的模型即可进行能耗预测。

在上述步骤中，数据清洗和特征工程是保障预测精度的基础，也是在实践过程中最耗费精力的部分。

在数据清洗方面，尽管我国多个省市均已建立了公共建筑能耗监测平台，但由于传感器长期未校准、传输信号故障等原因，能耗数据往往存在较多的缺失值和异常值。在进行数据分析之前，需首先对实际数据进行异常值的判断与缺失值的填补。异常值的判断方法有[21]：1）基于距离的方法，例如高英博等人在进行能耗预测前采用 K-Means 算法对上海某酒店建筑等能耗异常值进行了判断并采用 KNN（K-Nearest Neighbors）算法进行了能耗数据修复[22]。2）基于统计学的方法，如箱型图法，插值法，3δ准则法，标准偏差法等。3）基于决策树的方法，例如 Kim 等人[23]结合 K-Means 算法和孤立森林法对 9 栋建筑进行了能耗数据异常值的判断。缺失值的补全方法可分为：单变量（univariate）和多变量（multivariate）的方法。常用的单变量的方法有：平均值填补，向前填补，向后填补，线性插值，多项式插值，卡尔曼滤波器，滑动平均法等。多变量的方法有：KNN，随机森林（Random Forest），多重奇异谱分析（Multiple Singular Spectral Analysis），矩阵因子分解（Matrix Factorization）等方法。Brian 等人对比了不同缺失长度下 6 种缺失值填补方法的准确性和计算开销，结果表明，线性插值，KNN 和矩阵因子分解法是最有效的数据填补方法[24]。

在特征工程方面，特征工程对黑箱模型的性能起着决定性的作用，挖掘并构建与建筑能耗相关的变量作为黑箱模型的输入特征是建立黑箱模型的关键。特征工程包含：特征提取（feature extraction），特征创造（feature creation），特征选择（feature selection）三个步骤。特征提取指的是从文字、图像等非结构化数据中提取新信息作为特征。特征创造指的是把现有特征进行组合，或者互相计算得到新的特征，如把建筑外表面积和体积相除得到体形系数。特征选择指的是从所有

的特征中选择有重要的特征进行建模。常用的能耗预测黑箱模型的输入特征可分为：建筑物理相关参数（如建筑面积、层数、朝向、窗墙比、作息表等），气象参数（如室外干球温度、室外相对湿度、太阳辐射等），时间序列特征（如月份、是否节假日等）。但过多的特征往往会导致特征冗余并极大增加计算量。在某些算法下，特征的维度增多时所需要的计算量将会成指数增长，造成“维度灾难”，并且冗余的特征会导致模型精度降低。故在建模前需进行特征选择，提取出关键特征。特征选择的方法有：1）过滤法（filter），利用方差、卡方检验、皮尔森相关系数等方法去除掉包含信息较少或相关性较高的特征。2）嵌入法（embedded），是一种特征选择和模型训练同步进行的方法，利用随机森林，XGBoost 等基于决策树的算法进行特征重要程度的判断。3）包装法（wrapper），和嵌入法类似，但包装法通过使用一个目标函数而不是根据特征的重要性程度来进行特征选择。4）敏感性分析方法，包括 Morris 法，标准秩回归系数法（SRRC）和偏秩回归系数法（PRCC）等。5）降维算法，将初始变量映射到新的特征空间中，在新的空间中进行特征选择，即降维，常用的算法有 PCA、小波分解重构法等。上述特征选择方法往往结合起来使用，例如 Zhang 等人[25]首先利用专家知识进行特征的初步筛选，然后采用相关系数法去除掉与目标不相关的冗余特征，最后采用 MARS 作为包装法的算法进行了第三步特征选择；Yuan 等人[26]结合过滤法和嵌入法采用偏最小二乘回归(PLSR)、随机森林和支持向量机（SVM）进行了特征选择。不同的敏感性分析方法得到的特征重要性结果往往不同，例如，Li 等人[27]采用 SRCC、PRCC，Morris 和 Fast (Fourier Amplitude Sensitivity Test) 方法得到了不同的特征重要性排序结果。特征降维也可以有效地进行特征选择，例如 Ding 等人[28]结合过滤法中的相关系数法，采用小波分解和 PCA 进行了特征降维；朱明亚[29]在其博士论文中采用 PCA 对不同预测目标下的混合特征进行降维，提取出了不同目标、不同精度条件下的能耗预测最小变量集。此外，在实践过程中，一些数据竞赛平台（例如 Kaggle）上的参赛者采用 Leave one feature out（LOFO）的方法，将特征逐个添加到模型中，判断单个特对模型性能的影响，从而进行特征选择。

在数据驱动模型方法，常用的算法可分为：1）回归类的模型，例如岭回归，Lasso 回归，贝叶斯岭回归，支持向量回归等。2）树模型，通常采用组合树模型，例如随机森林，XGBoost，lightGBM，CatBoost 等强学习器。3）神经网络模型，例如多层感知机（MLP），ANN，RNN，LSTM 等。其中树模型因其准确性高，计算快的特点成为了数据比赛中较为主流的算法，在美国暖气和空调工程师学会（ASHRAE）举办的能耗预测比赛 Great Energy Predictor III competition 中，获奖的前五名选手均采用了树模型[30]。并且除采用单个模型外，他们均进行了模型的

同济大学

集成（ensemble），模型集成的方法有装袋法（bagging），提升法（boosting）和堆叠法（stacking）。

Chen 等人[31]总结了能耗预测模型中常用的特征和模型。从现有文献来看，数据驱动模型的特征仍不够丰富，鲜少涉及建筑物理相关参数。

3) 灰箱模型

白箱模型的建模和校准过程对建模者来说是一个巨大的挑战。白箱模型需要大量的基本参数作为输入，建模复杂且耗时久，需要较大的人力物力且模型预测精度并不高。黑箱模型采用统计学或机器学习的方式建立输入和输出变量之间的关系，但在不同的情景下，通常需要大量的历史数据来训练模型以实现准确预测负荷的目的，并且采用机器学习方法建立的模型并不具有很好的可解释性。为了解决上述困境，有学者提出了灰箱模型。灰箱模型既不像白箱模型那样需要详尽的建筑本身及内部设备的参数，也不像黑箱模型那样需要大量历史数据。能耗预测中的灰箱模型可分为两类[32]：采用黑箱算法来确定白箱模型的系数；结合白箱的模拟结果与统计学或机器学习算法来简化或替代白箱模型。

最为常用的混合模型为 RC 模型，由 Hassid S.在 1985 年初提出[33]。他提出由两个电阻和一个电容(即 2R1C 模型)来代表多层建筑围护结构的热性能。此后又发展出了 RC-S、3R2C 等模型用于不同的场景。一般有两种方法获得 R 和 C 的值[32]：1）正演法：由物理模型直接计算获得最佳取值。许多建筑模拟工具(如 EnergyPlus)均可以计算 R 和 C 值；2）逆向法：采用数据驱动算法进行曲线拟合得到 R 和 C 两个参数的值。参数估计采用的典型的算法有最小二乘回归[34][35][36]，系统辨识[37]等，有研究采用遗传算法对参数的取值进行了优化[38][39]。

Lam 等人[40][41]在 1997 年提出了应用白箱模拟数据训练黑箱模型的方法，即利用物理模拟软件得到大量模拟数据生成数据库，用前述数据库训练黑箱模型。黑箱算法可采用多元线性回归、人工神经网络和支持向量机等。其优点是可以快速得到模拟结果，而不需要建立物理模型进行模拟。

沙华晶[42]的博士论文中提出了将模拟数据、不同颗粒度的实测数据进行数据融合作为数据驱动模型的训练数据的方法，但该方法仅应用在了酒店建筑上，且在数据融合部分只是将模型数据和实测数据进行缩放和简单的修正，本研究对该混合模型进行了改进和扩展。

1.2.2 跨建筑能耗预测综述

大部分数据驱动模型的建立仍需大量的建筑历史能耗，但在实际中，需进行节能改造的旧楼宇或新建楼宇并没有详细的历史能耗数据。有学者研究采用迁移

学习的方法将一栋相似楼宇的能耗数据应用到另一栋楼宇的数据驱动模型中。数据驱动模型往往在一个数据集（源域，在跨建筑能耗预测中为相似建筑的能耗数据）上表现良好，到另一个数据集（目标域，在跨建筑能耗预测中为目标建筑的能耗数据）时需要进行重新训练[43]。但不管是在源域还是目标域上，要进行的预测任务是一致的，均是进行建筑的能耗预测，迁移学习可用来解决不同数据源下相似任务的问题[43]。迁移学习在能耗预测中的应用还比较少。Fan 等人[44]利用源域的数据进行神经网络模型的预训练，再利用目标域的数据进行特征提取或者隐藏层权重值的调节，Li 等人[45]采用相似的思路，进行 ANN 模型的预训练和权重调整。Tian 等人[46]根据欧式距离、Cosign 距离、曼哈顿距离、DTW 距离定义了数据的相似度，在相似建筑的初始建筑建立的 LSTM 模型基础上调整模型进行迁移学习。Ribeiro 等人[47]提取并去除能耗数据中的趋势项和季节项，使得一个标准的数据驱动模型可直接应用于多栋建筑。Qian 等人[48]利用迁移学习算法 TrAdaBoost 对源域中数据的权重进行调整。Fang 等人[49]提出了 LSTM-DANN 框架，利用对抗生成网络的思想提取出源域和目标域中共同的特征，再将其作为 LSTM 的输入特征进行迁移学习。Mocanu 等人[50]将强化学习与迁移学习结合以实现跨建筑的能耗预测。但上述文献中仍需用到少部分目标域建筑能耗数据，难以应用于完全没有历史数据的跨建筑能耗预测，并且上述研究并未将建筑本身的物理特性作为特征。

1.3 研究内容及技术路线

1.3.1 研究内容和术语解释

从文献综述中可以看出，因白箱模型复杂、过于理想化、未考虑到施工和运维阶段对能耗的影响的弊端，黑箱和灰箱模型受到了越来越多学者的关注。从现有研究来看，黑箱模型的输入变量挖掘仍不够，比较突出的是缺乏建筑物理相关特征。并且数据驱动模型和混合模型模型在跨建筑能耗预测上的表现并不佳，从课题组举办的跨建筑能耗预测比赛结果来看，所有队伍中最准确的空调分项预测误差 CV-RMSE 为 0.6337[51]，而现有的能耗预测迁移学习研究仍需少量的目标建筑历史能耗数据，限制了跨建筑能耗预测的应用范围。

故本论文要解决的问题有：

1) 哪些变量对能耗有较大的影响？
2) 在实际应用中，一些对能耗影响大的变量（如新风渗透率）的值难以获取，如何合理的表示这部分变量？

同济大学

3) 如何利用多源异构数据进行跨建筑的能耗预测？

为解决输入变量和跨建筑能耗预测的问题，在课题组的已有研究基础上，本课题聚焦于办公建筑，研究了基于多源异构数据的办公建筑能耗预测方法，综合应用批量建模生成的模拟数据，节能审计报告中建筑物理信息和颗粒度比较粗的逐月能耗数据，分项计量平台中颗粒度较细的数据，构建了不同场景下的能耗预测混合模型。

术语解释：

- **多源异构数据：**

本研究中的多源异构数据指的是来源不同，数据类型不同，数据颗粒度不同的不能直接合在一起使用的数据。本研究中使用到的数据主要来源于：模拟数据，建筑基本物理信息，节能审计报告的数据和能耗监测平台的数据。其中模拟数据、节能审计报告的能耗数据、能耗监测平台的数据时间分辨率（即颗粒度）不同，建筑基本物理信息需和时间序列数据进行匹配，均需要做不同程度的数据融合。

- **混合能耗模型：**

本研究中的混合能耗模型指的是用于建模的数据既包括模拟数据，又包括不同来源的实测数据的模型。

- **建筑负荷相关变量、机电系统相关变量：**

建筑负荷相关变量指的是在不考虑机电系统配置的情况下（例如在 EnergyPlus 中设置理想空调系统）的只与建筑物负荷相关的几何、热工等变量，例如建筑面积、层数、窗墙比等。机电系统相关变量指的是空调系统类型、水系统类型等与暖通空调系统相关的变量。

- **初始变量、关键变量：**

初始变量指的是所有与能耗可能相关的变量，是特征提取的基础变量。关键变量指的是从初始变量中提取的对建筑能耗影响较大的变量。

1.3.2 技术路线和文章安排

本课题研究的基于多源异构数据的办公建筑能耗预测方法包括三个部分：关键变量提取，数据融合，异构数据库和混合模型的建立。技术路线如图 1.2 所示。

第一部分是为了找出哪些变量对建筑能耗影响较大。将初始变量作为输入值，利用 python 和 eppy 库建立了能耗模型快速建模工具，批量生成初始变量对应的能耗模拟值，将初始变量中非关键变量的值按照公共建筑节能设计规范[52]和相关的规范及手册[53]进行设定，得到了关键变量作为输入参数的能耗模型快速建模工具，从而批量生成模拟数据。并利用不同的特征提取方式筛选出了对能耗影响较大的关键变量。

第二部分是为了得到缺失的关键变量的值和进行多源异构数据的数据融合。在数值缺失的关键变量推测方面，若存在逐月/逐日/逐时的能耗数据，将采用遗传算法对关键变量缺失值进行推测，若不存在能耗数据，则采用关联规则挖掘算法进行关键变量缺失值的推测。需要说明的是，在进行数据库建立的时候，由于所有的模型都存在历史能耗，故均采用遗传算法进行关键变量缺失值的推测，在混合模型训练好后进行预测时，可能存在无历史能耗的情况，此时可采用关联规则挖掘算法进行关键变量缺失值的推测。在多源异构数据融合方面，对于时间颗粒度的转换，首先将办公建筑的能耗进行聚类，得到不同类型的办公建筑的典型能耗曲线，再根据典型能耗曲线进行颗粒度较粗的能耗数据的填充，将能耗数据转换到目标颗粒度上；在模拟数据和实际数据的融合方面，由于模拟数据和实际数据的差异较大，采用了两级的数据驱动模型对模拟数据先进行修正，使其接近实际能耗数据，再将模拟数据作为混合模型的一个输入特征。

第三部分是为了进行不同场景下的能耗预测。用上述两部分得到的建筑关键变量和数据融合后的能耗数据构建多源异构数据库。将关键变量信息，时序特征（是否工作日等）和修正后的模拟数据作为输入，能耗数据作为预测值，进行不同场景下能耗预测模型的建立。

同济大学

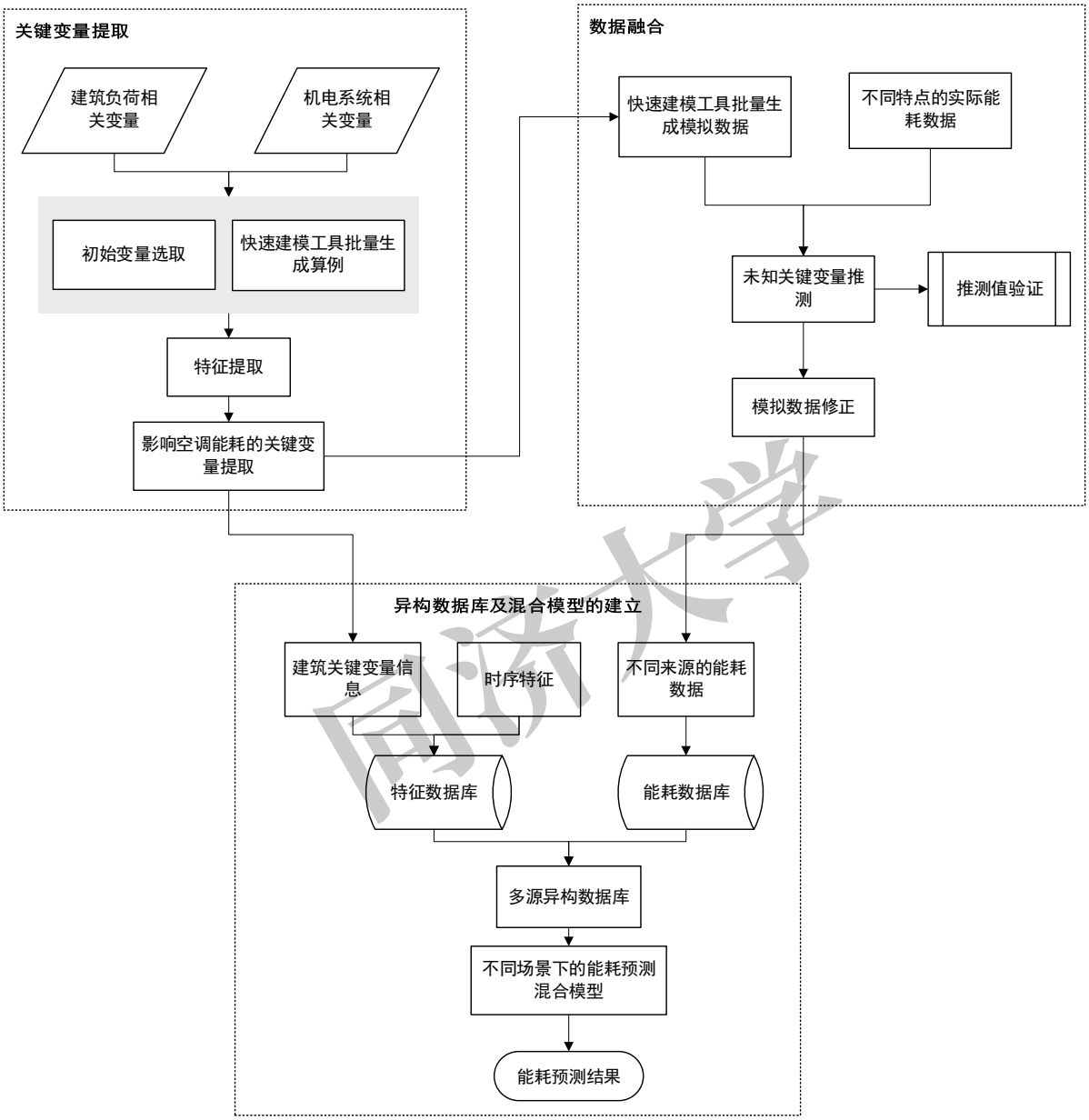


图 1.2 技术路线

本文第 2 章阐述了技术路线中第一部分关键变量的提取方法；第 3 章阐述了第二部分关键变量缺失值的推测过程和推测准确度分析；第 4 章介绍了第二部分多源异构数据的融合过程。第 5 章说明了第三部分多源异构数据库和混合模型的建立过程；第 6 章对混合模型进行了验证；第 7 章为本研究的结论与展望。

1.4 本章小结

本章节首先说明了本研究的研究背景和意义，建筑节能对我国的节能事业和实现双碳目标意义重大；然后对国内外能耗预测相关研究进行了综述，除传统的

白箱模型，黑箱模型和混合模型因其便利性和准确性受到了越来越多的关注，但也存在着输入特征挖掘不够充分，无法较好的实现跨建筑能耗预测的缺陷；最后提出了本文所研究的基于多源异构数据的办公建筑能耗模型的技术路线，通过综合考虑建筑负荷和机电系统相关变量并从中提取出关键变量的方法解决输入特征的问题，通过融合模拟数据、多种来源的实测能耗数据实现相似办公建筑的跨建筑能耗预测。

同济大学

同济大学

第 2 章 关键变量提取

2.1 概述

无论是白箱、黑箱还是灰箱能耗模型，从本质上来看均包括了三个部分：输入变量，输出变量以及输入变量与输出变量之间的关系。不管采用哪种软件，白箱模型的建立均需要大量且详细的输入变量，耗费大量时间和精力建立建筑几何和机电系统的物理描述，再经过复杂的迭代求解，最终才能得到输出变量值。而在实际应用中，一些输入变量值是难以获取的，例如新风渗透率，而白箱模型无法在变量缺失的情况下进行建模，这也是导致白箱模型计算结果置信度较低的原因之一。数据驱动模型的输入变量数量是可调整的，但较少的输入变量无法全面的反映对能耗的影响，过多的输入变量又会导致模型计算时间过长且将增加过拟合的风险。从理论上来讲，所有白箱模型需要输入的变量都与能耗有着或多或少的影响，但输入变量的增加往往会导致数据驱动模型所需的计算量急剧增长，将所有的相关变量均作为数据驱动模型的输入势必会造成“维度灾难”。各输入变量对能耗的影响并不是相同的，将对能耗有显著影响的变量作为输入可使模型在可接受的计算量内实现较为准确的能耗预测。本章节从可映射到白箱模型的初始模型中提取出了对办公建筑能耗有显著影响的关键变量。图 2.1 为关键变量的提取技术路线。首先，由于初始变量过多将会导致所需算例过多，且负荷相关变量和机电相关变量对能耗的影响不完全相同，故将初始变量分为建筑负荷相关变量和机电相关变量两类。其次，根据初始变量的抽样值，利用 python 语言和其 eppy 库[54]程序化地生成能耗模拟工具 EnergyPlus 所需的 IDF 文件，并调用 EnergyPlus 进行能耗模拟，得到制冷能耗和制热能耗。然后，分析初始变量对输出变量的影响。对于建筑负荷相关变量，采用敏感性分析方法从初始变量中提取出建筑负荷相关的关键变量，对于机电系统相关变量，由于这部分变量对能耗的影响并不仅限于峰值，对于这部分变量采用嵌入法进行关键变量的提取。

本章中的 2.2 和 2.3 节将会具体介绍两类初始变量的选取和关键变量提取过程。

同济大学

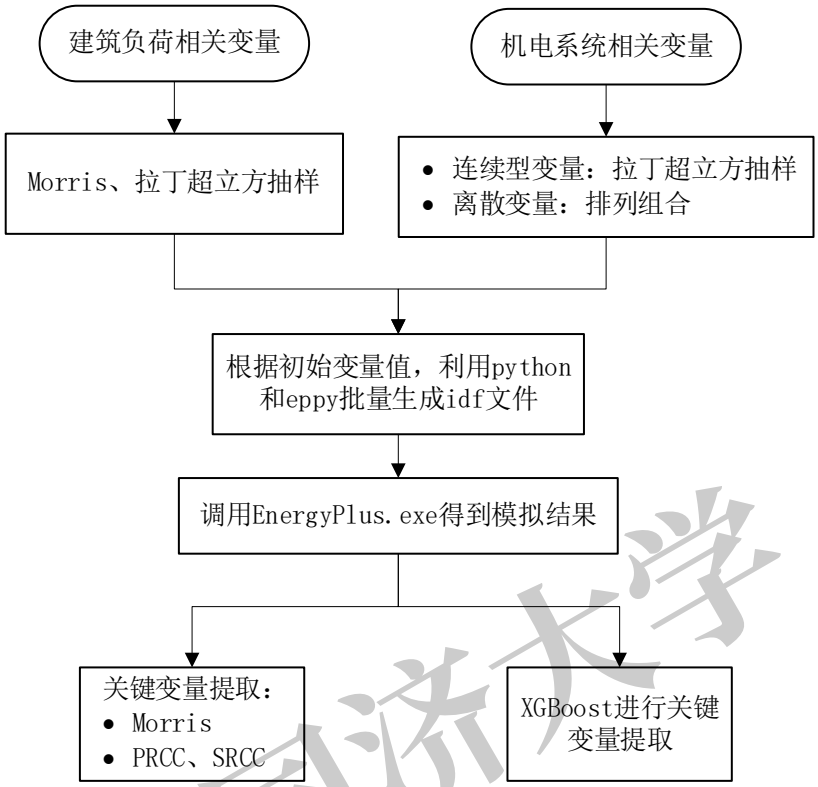


图 2.1 关键变量提取流程

同济大学

2.2 负荷部分关键变量

本节将介绍建筑负荷相关的关键变量的提取。负荷部分关键变量提取流程如图 2.1 左侧所示，大致可分为四个步骤：第一，进行确定初始变量及其取值范围；第二，在取值范围内，用 Morris 方法和拉丁超立方方法分别对初始变量进行抽样，生成一系列初始变量值的集合；第三，根据初始变量集合批量生成 EnergyPlus 的输入——IDF 文件，再调用 EnergyPlus 进行模拟计算，得到对应的“因变量”——能耗值；第四，将抽样得到的初始变量集合作为输入，能耗模拟值作为输出，分别用 Morris、标准秩回归系数法（SRRC）和偏秩回归系数法（PRCC）方法提取关键变量，其中用 Morris 方法提取关键变量时，需对应 Morris 方法抽样得到的初始变量集合作为输入，在使用其他方法提取关键变量时，对应拉丁超立方抽样得到的初始变量集合。下文介绍了负荷部分关键变量提取的各个步骤和关键变量提取结果。

2.2.1 负荷相关的初始变量选取及其取值范围

对于负荷相关的初始变量，本研究参考课题组沙华晶博士[42]在进行酒店建筑

关键变量提取时初始变量的选择，选取了几何参数、热工参数、运行参数、施工质量四个方面的 23 个初始变量，并根据工程经验、相关办公建筑节能规范和设计手册中对夏热冬冷地区的办公建筑的说明和规定进行初始变量取值范围的选取，初始变量名称及其取值范围等信息如表 2.1 所示，假设各变量均在其取值范围内为均匀分布。其中前三类的变量为能耗模拟中较为常见的变量，最后一类变量参考相关文献[55]，用楼板线性透过率、玻璃线性透过率和墙角线性透过率表征了建筑围护结构接缝存在的缝隙等施工质量不佳的情况。

表 2.1 建筑负荷相关初始变量及其取值

类别	初始变量名称	缩写	取值范围	单位
几何参数	北向窗墙比	NWWR	0.1~0.7	-
	南向窗墙比	SWWR	0.1~0.7	-
	东向窗墙比	EWWR	0.1~0.7	-
	西向窗墙比	WWWR	0.1~0.7	-
	建筑面积	AREA	20000~80000	m^2
	层数	NL	3~50	层
	体形系数	CR	0.1~0.5	-
热工参数	外墙传热系数	WALLU	0.09~0.5	$W/(m^2 \cdot K)$
	外墙热容	WSP	800~2000	$J/(kg \cdot K)$
	屋顶传热系数	RU	0.09~0.4	$W/(m^2 \cdot K)$
	外墙太阳辐射得热系数	WSA	0.1~0.9	-
	屋顶太阳辐射得热系数	RSA	0.1~0.9	-
	窗玻璃传热系数	WINU	1~2.7	$W/(m^2 \cdot K)$
	窗玻璃太阳辐射得热系数	SHGC	0.1~0.52	-
运行参数	制冷设定温度	SPC	21~29	℃
	供热设定温度	SPH	18~26	℃
	照明功率密度	LPD	3~20	W/m^2
	人员密度	OPD	0.05~1	P/m^2
	新风渗透率	INFIL	0.5~5	ACH
	内遮阳开启程度	ST	0.1~0.9	-

同济大学

类别	初始变量名称	缩写	取值范围	单位
施工质量	楼板线性透过率	FLT	0.007~1.842	-
	玻璃线性透过率	GLT	0.03~0.7	-
	墙角线性透过率	CLT	0.036~0.5	-

2.2.2 负荷相关初始变量的抽样

负荷相关的初始变量均为连续型变量，采用有效的抽样方式获取能够反映总体特性的样本点非常重要。由于后续采用的 Morris 法在进行敏感性分析时需采用与之对应的 Morris 抽样方法，本研究假设初始变量在其取值范围内均匀分布[27][56]，采用了 Morris 对应的抽样方法和拉丁超立方进行初始变量的抽样。

1) Morris 方法

Morris 抽样及敏感性分析方法是由 Max D. Morris 在 1991 年提出的适用于复杂模型的敏感性分析方法，因其计算成本小受到广泛的研究与应用[57][58][59]。Morris 方法基于元效应，即一次改变一个变量（One Factor At a Time, OAT），以此确定这一变量对输出值的影响。假设输出值 y 是由 k 个输入值 $x_1, x_2, x_3, \dots, x_k$(输入向量表示为向量 **X**)确定的，则第 i 个输入值的元效应（elementary effect， EE）为第 i 个输入值变化前后 y 值的相对变化量，其定义如式（2.1）所示：

$$EE_i(\boldsymbol{X}) = \frac{[y(x_1,x_2,\dots,x_{i-1},x_i+\Delta,x_{i+1},\dots,x_k)-y(\boldsymbol{X})]}{\Delta} \quad (2.1)$$

其中Δ可取 1/(p-1)[57]，p 为输入值的水平数。

由上可知，在用 Morris 进行敏感性分析时，样本中除第 i 项外，其他各项输入值均相同才能得到第 i 个输入值的元效应。故 Morris 法在实施时需用特定方法进行采样。**X** 中每一个变量均变化一次则完成一次轨迹，每次轨迹中包含 k+1 个样本数，每次轨迹的采样矩阵可表示为一个$(k+1)\times k$维的对角矩阵，如式 2.2 所示。

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \quad (2.2)$$

B 中每行表示 **X** 的一个样本值，1 表示改变的输入值，0 表示不变的输入值，

每行相较于上一行只改变 **X** 中的一个输入值，分析第 i 个元素所用的 B 中的两行如式（2.3）所示。

$$\mathrm{B(i)} = \begin{bmatrix} x_1\ x_2 \cdots x_{i-1}\ x_{i,1}\ x_{i+1} \cdots x_k \\ x_1\ x_2 \cdots x_{i-1}\ x_{i,2}\ x_{i+1} \cdots x_k \end{bmatrix} \tag{2.3}$$

在执行完 r 个轨迹的采样和分析后，y 对变量x_i的敏感度可表示为μ_i^*，如式（2.3）所示，元效应的波动大小可表示为σ_i，如式（2.4）。

$$\mu_i^* = \frac{\sum_{j=1}^{r} EE_i^j}{r} \tag{2.3}$$

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{r} \left| EE_i^j - \mu_i^* \right|^2}{r}} \tag{2.4}$$

若μ_i^*越大则输入变量x_i对 y 的影响越大，σ_i越大表示相对于x_i的元效应之间存在显著差异，变量间存在非线性效应和交互作用，σ_i越小，表示元效应不受其他变量影响。

2) 拉丁超立方抽样方法（Latin Hypercube Sampling LHS）

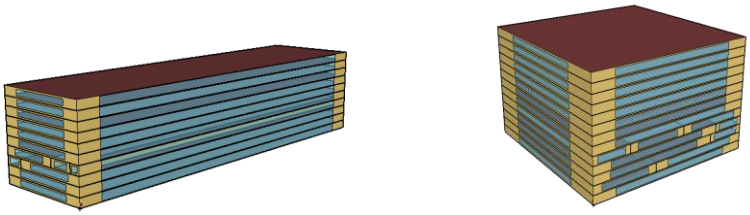
拉丁超立方是一种分层抽样方法，该方法可以高效的从样本空间中抽取相对较少的样本点，使其代表的整个样本空间信息最大化。拉丁超立方抽样是将样本空间划分成若干个区间，并从每个区间中抽取样本代表该区间。假设需在 n 维向量空间中抽取样本，拉丁超立方的抽样步骤为：第一，将每一维分成互不重叠且概率相等的 m 个区间（若概率分布函数为均匀分布，则每个区间长度相等）；第二，从每一维的每个区间中随机抽取样本，一旦样本从分层中抽取之后便不再从这一分层中抽样，确保每个分层中均有样本被抽取出，这些样本构成超立方体的总体；第三，将每一维中抽取出的样本组成 n 维向量。这种方法使得每个分层中均有样本代表，保证抽出的样本使得整个样本空间信息最大化。

2.2.3 负荷部分白箱模型的批量生成

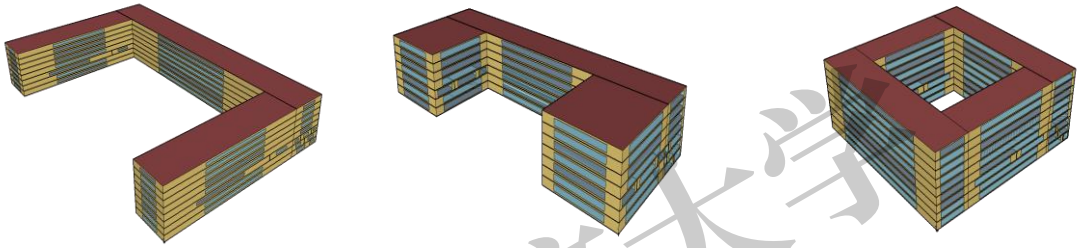
输入变量与输出变量之间的关系可通过多种方式建立，本研究选取 EnergyPlus 作为模拟引擎建立了输入变量与输出的对应关系。为减少建模成本、提高建模效率，本研究利用 python 语言和其 eppy 库进行批量建模生成 IDF 文件（IDF 是 EnergyPlus 采用的数据存储格式）。eppy 可实现 IDF 的信息进行批量读写、调用 EnergyPlus 进行模拟、读取模拟结果等。本研究将表 2.1 中的 23 个输入变量分别体现到 IDF 中的参数设置中。几何部分，参考沙华晶博士论文中[42]的方式，利用几何参数中的体形系数、层数等匹配常见的建筑外形，可生成的建筑外形有：长方形、正方形、U 形、回字形，如图 2.2 所示。由于有的综合类办公建筑中的功能有商业区和办公楼区域，根据办公建筑的特点，房间功能设置有：

同济大学

办公室、会议室、走廊、餐饮（食堂）和设备用房。



a) 长方形　　b) 正方形



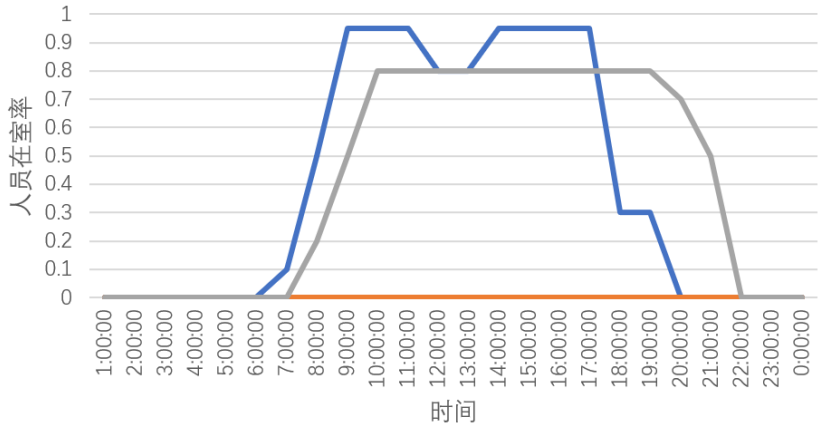
c) U 形（两侧为长方形）　　d) U 形（两侧为正方形）　　e) 回字形

图 2.2 可生成的建筑外形

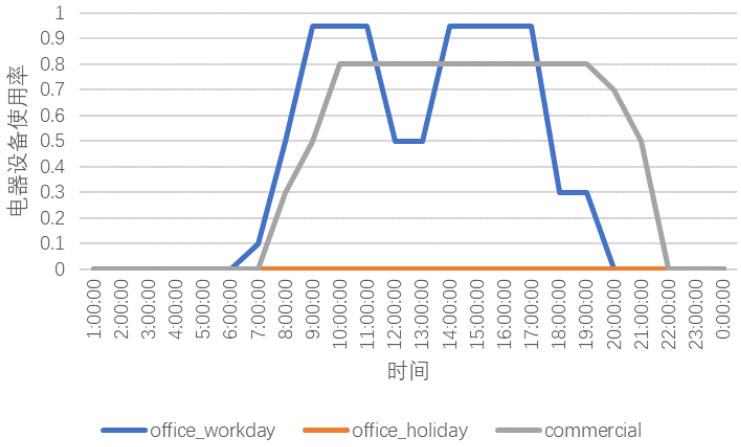
热工参数和运行参数部分，根据输入参数抽样值可较容易的在 IDF 中设置，施工质量相关的变量参考相关文献[55]，将楼板线性透过率、玻璃线性透过率和墙角线性透过率折算为墙体传热系数的增加。

其他 IDF 中需设置的输入值按照办公建筑的相关标准进行设置，商业和办公区域人员、电器设备、照明开关使用和新风的时间表根据《公共建筑节能设计标准》GB 50189-2015 进行设置，主要房间的作息表设置如图 2.3 所示。

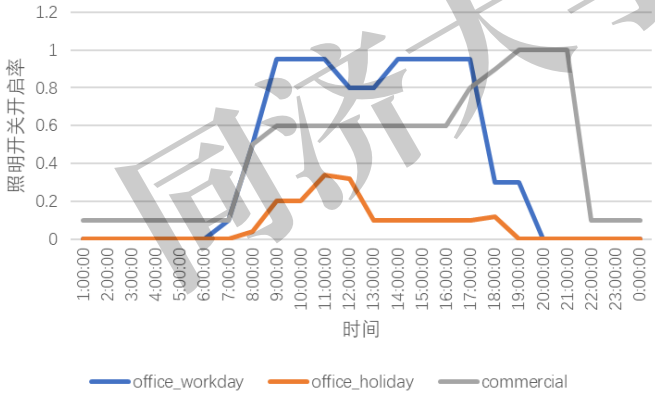
由于该部分只考虑负荷相关的变量，能耗模型的空调模型设置为 EnergyPlus 中的理想空调系统。



a) 人员在室率



b) 电器设备使用率



c) 照明开关开启率

d) 新风使用率

图 2.3 主要房间的时间表设置

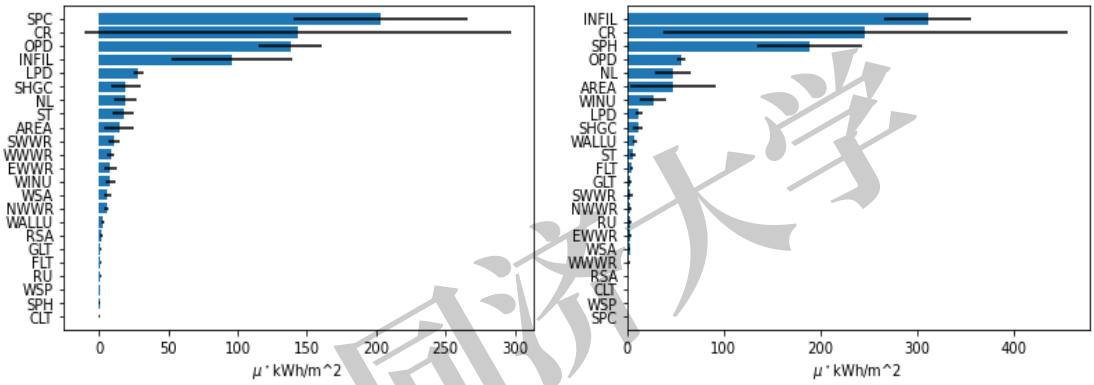
2.2.4 负荷部分关键变量的提取

对于负荷部分的变量，采用了 Morris 法、标准秩回归系数法（SRRC）、偏秩

回归系数法（PRCC）这三种敏感性分析方法进行关键变量的提取。

1) Morris 法

Morris 法的抽样与元效应分析过程已在 2.2.2 节进行了阐述，本课题进行了 8 个轨迹共 192 个样本的采样，使用 2.2.3 节所述的工具进行模型批量生成和模拟计算，从经验知识可知，能耗与面积成正相关关系，为弱化面积的影响，将单位面积的制冷能耗和供热能耗分别作为输出，得到的输入变量重要程度排序如图 2.4 所示。μ^*值越大则输入变量的重要程度越大，输入变量与缩写的对应关系见表 2.1。



a) 制冷能耗的输入变量重要性排序　　b) 供热能耗的输入变量重要性排序

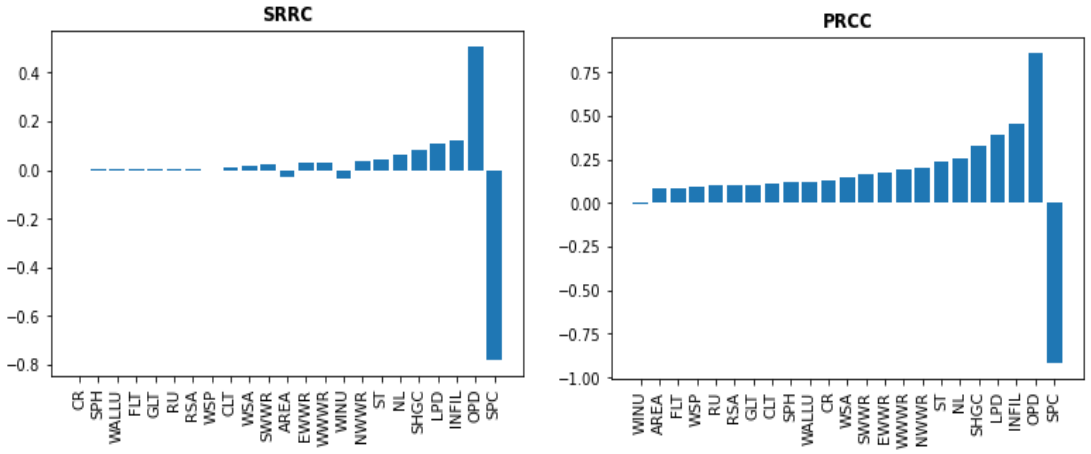
图 2.4 使用 Morris 法得到的输入变量重要性排序

2) 标准秩回归系数法（SRRC）和偏秩回归系数法（PRCC）

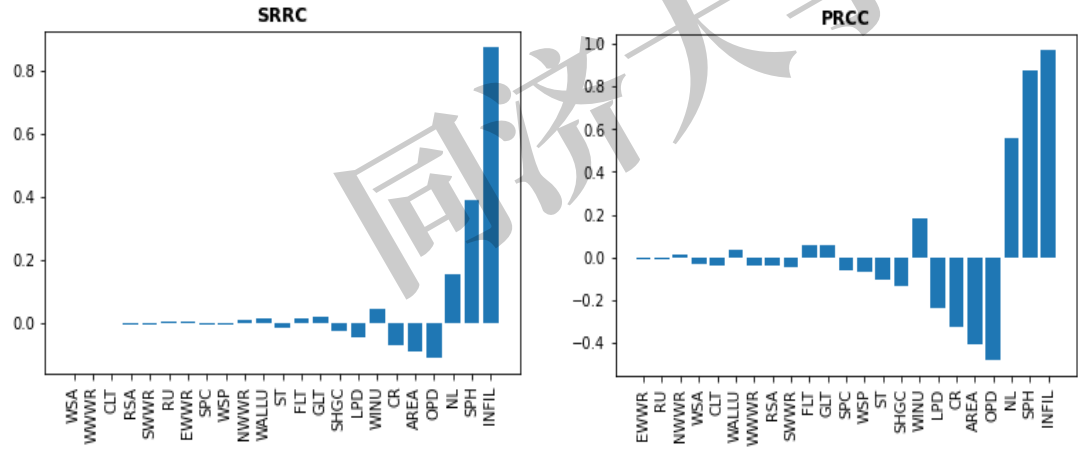
SRRC 和 PRCC 均属于敏感性分析方法中的回归分析法，也是能耗相关的敏感性分析研究中常用的方法[27][60][61]，是标准回归系数（SRC）和偏回归系数（PCC）的秩变换值。SRC 是线性回归模型中输入变量的系数，其绝对值越大，对模型的重要程度也越大，PCC 在相关性的基础上还考虑了偏相关性，但 SRC 和 PCC 通常适用于线性模型，对于能耗模型这样的非线性模型需要对其进行秩变换得到 SRRC 和 PRCC[42]。

SRRC 和 PRCC 的输入变量均采用 2.2.2 节所述的拉丁超立方法进行抽样，共抽取 3000 个样本，使用 2.2.3 节所述的工具进行模型批量生成和模拟计算，同样将单位面积制冷能耗和供热能耗分别作为输出，得到的输入变量重要程度排序如图 2.5 所示。其中 SRRC 或 PRCC 的绝对值越大则输入变量对输出变量而言更重要，输入变量与缩写的对应关系见表 2.1。

同济大学



a) SRRC 和 PRCC 指标下输入变量对制冷能耗的重要性排序



b) SRRC 和 PRCC 指标下输入变量对供热能耗的重要性排序

图 2.5 回归分析下（SRRC 和 PRCC 指标）得到的输入变量重要性排序

2.2.5 负荷部分关键变量结果分析

从图 2.4 和图 2.5 中可以看出 Morris 方法和回归分析法得出的重要性排序基本一致，无论是单位面积制冷能耗还是单位面积供热能耗，两种方法中重要性排名前五的输入变量中仅有一个变量有差异。表 2.2 归纳了两种方法中重要性排在前九的变量，从表中的结果可见，制冷能耗和供热能耗变量相似，前九的变量中只有一个变量不同：对于制冷能耗来说，窗墙比是较为重要的变量；而对于供热能耗来说，窗户传热系数是较为重要的变量。不过对于制冷和供热，关键变量重要性排序不同，对于制冷来说，最重要的变量为制冷设定温度，其次是体形系数、人员密度和新风渗透率；而对于供热来说，最重要的变量是新风渗透率，其次是体形系数、供热设定温度和人员密度。

同济大学

表 2.2 负荷相关关键变量汇总

输出变量	关键变量								
单位面积制冷能耗	制冷设定温度	体形系数	人员密度	新风渗透率	照明功率密度	太阳得热系数	层数	内遮阳开启程度	窗墙比
单位面积供热能耗	新风渗透率	体形系数	供热设定温度	人员密度	层数	面积	窗户传热系数	照明功率密度	太阳得热系数

2.3 机电系统部分关键变量

对于机电系统部分的关键变量提取，方法与负荷相关的关键变量提取类似。同样是四个步骤：第一，选取初始变量；第二，进行初始变量的抽样；第三，批量生成算例；第四，提取关键变量。不过由于机电系统的变量除了影响能耗的峰值或者均值外，对逐时值也有较大的影响，故采用 XGBoost 算法对于机电系统部分的关键变量进行提取。下文详细阐述了机电系统部分关键变量的提取。

2.3.1 机电系统相关初始变量选取及其取值范围

表 2.3 机电系统相关初始变量及其取值

类别	初始变量名称	缩写	取值范围	单位
离散变量	风系统类型	Terminal	定风量系统、变风量系统、风机盘管系统	-
	水系统类型	WS	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	-
连续变量	送风温度	SAT	8~18	℃
	冷冻水/热水供水温度*	CHWT	5~10/35~60	℃
	风机效率	FE	0.3~0.8	-
	水泵效率	PE	0.3~0.8	-
	冷机/热泵 COP	COP	3~7	-
	供回水温差*	TD	2~7(冷)/10~25(热)	℃
	冷却塔填料堵塞系数	TPR	0.5~1	-
	风管过滤器堵塞系数	FFR	1~2	-

同济大学

备注：上表中带*的部分为分析制冷能耗和供热能耗时有不同设置值的输入变量。

2.3.2 机电系统相关初始变量的抽样

机电系统部分初始变量既有离散变量又有连续变量，对离散变量进行排列组合，生成 9 种组合，对连续变量采用 2.2.2 节介绍的拉丁超立方方法在取值范围内进行抽样，抽取 800 个样本。连续变量的 800 个样本与离散变量的 9 种组合最终组合成 7200 个样本。

2.3.3 机电部分白箱模型的建立

机电系统部分同样利用 python 语言和其 eppy 库进行模型的批量建立，采用 EnergyPlus 作为模拟引擎建立输入变量与输出的对应关系。将负荷部分变量固定为规范或手册中的推荐值，将初始变量根据采样值设置到模型中。根据机电相关初始变量中不同的风系统、水系统类型建立不同的设备和环路，并对应的送风温度、供回水温度和设备效率的设置批量生成 IDF 文件。其中冷却塔填料堵塞系数由冷却塔换热系数的降低来表征，风管过滤器堵塞系数由风机曲线的变化来表征。

2.3.4 机电系统部分关键变量的提取

由于机电系统部分的变量对逐时能耗值和能耗曲线的形状影响较大，本部分采用逐时能耗值作为输出变量，初始变量采样值作为输入变量，采用 XGBoost 算法进行关键变量的提取，该方法属于 1.2.1 节中提到的嵌入法提取关键变量。

XGBoost 是陈天奇等人[62]开发的基于梯度提升树（Gradient Boosting Decision Tree, GBDT）模型。该算法采用 boosting 的思想，为防止过拟合，在建立决策树时往往会对数据集进行抽样建模，XGBoost 会在一个数据集上抽取样本作为训练数据建立多棵决策树，在新建决策树时加大之前模型中误差大的样本的权重，如此迭代多次，误差越大的样本权重越大，被抽中的概率也就越大，这样使得模型整体的训练往误差大的样本倾斜，更有可能降低这些样本的误差，从而使得新建的决策树预测效果逐渐提升。

机电系统相关的初始变量对逐时制冷能耗的重要性排序如图 2.6 所示（输入变量与缩写的对应关系见表 2.3）。其中对制冷能耗而言，最为重要的变量为供回水温差，其次是水系统类型。选取供回水温差、水系统类型、送风温度、风系统类型、冷冻水供水温度和 COP 作为后续研究的关键变量。

同济大学

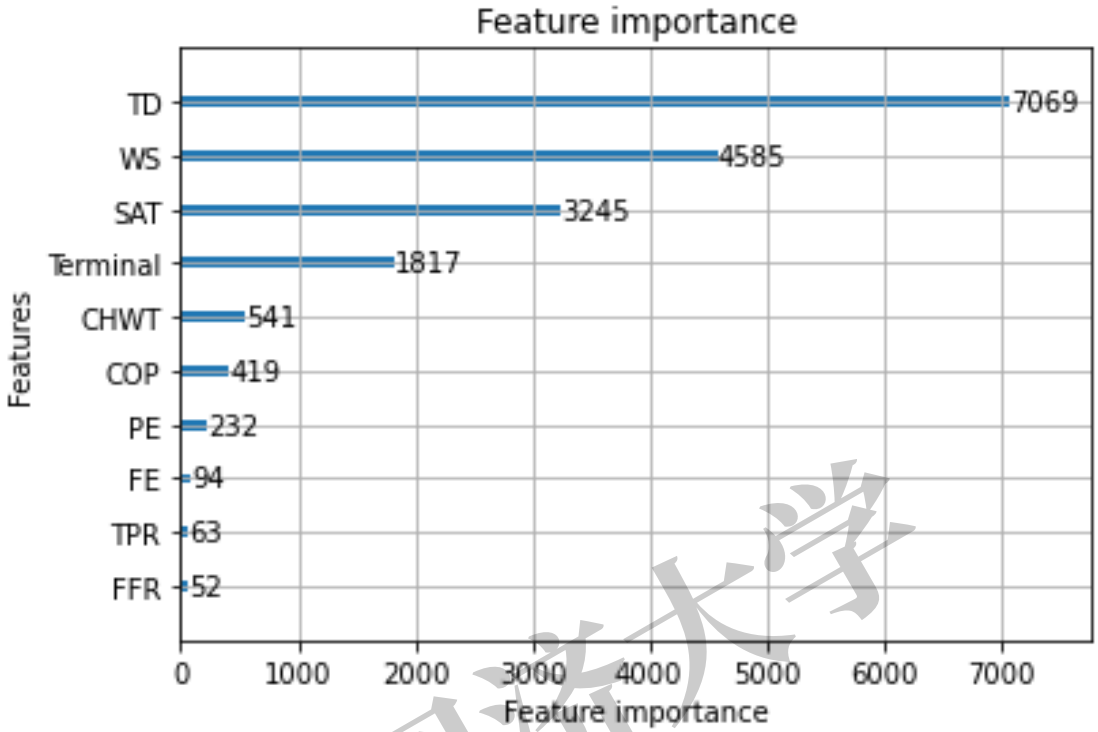


图 2.6 机电系统相关的初始变量的重要性排序

2.4. 能耗模拟值预测模型

为研究上述负荷相关变量、机电系统相关变量、天气参数和时序相关变量对能耗的影响，并得到后续研究中需用到的能耗模拟值预测模型，本节利用快速模拟工具（白箱模型）生成的模拟数据将能耗模型灰箱化得到用于关键变量缺失值推测的能耗模拟值预测模型。能耗模拟值预测模型使用 2.3 和 2.4 节中重要性排序靠前的变量、天气参数和时间序列相关特征作为输入变量，所有的输入变量如表 2.4 所示。将其中负荷相关和系统相关的连续变量采用拉丁超立方的方法抽取 300 个样本，再与离散变量进行排列组合，最终得到 2700 个样本。用 2.2.3 和 2.3.3 所述的快速建模工具进行建模并得到能耗模拟结果作为灰箱模型的输出。最后使用 2.3.4 节中提到的 XGBoost 算法进行建模，得到能耗模拟值预测模型。

将本研究与朱明亚博士论文中提取的办公建筑能耗模型的最小变量集相比较，结果是较为一致的。

同济大学

表 2.4 能耗模拟值预测模型的输入变量

类别	参数	缩写	取值范围	单位
负荷相关	制冷设定温度	SPC	21~29	℃
	供热设定温度	SPH	18~26	℃
	体形系数	CR	0.1~0.5	-
	人员密度	OPD	0.05~1	P/m^2
	新风渗透率	INFIL	0.5~5	-
	照明功率密度	LPD	3~20	W/m^2
	太阳得热系数	SHGC	0.1~0.52	-
	层数	NL	3~50	层
	内遮阳开启程度	ST	0.1~0.9	-
	窗墙比	WWR	0.1~0.7	-
	面积	AREA	20000~80000	m^2
系统相关	供回水温差	TD	2~7(冷)/10~25(热)	℃
	送风温差	SATD	4~10	℃
	冷冻水供水温度/热水供水温度	CHWT	5~10/35~60	℃
	冷机/热泵 COP	COP	3~7	-
	风系统类型	Terminal	定风量系统、变风量系统、风机盘管系统	-
	水系统类型	WS	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	-
天气参数	干球温度	DryT	-	℃
时间序列特征	每年的月	Month	1~12	-
	每月的日	Day	-	-
	每日的小时	Hour	1~24	-
其他	建筑编号	ID	-	-
	商业部分面积占比	COR	0.1~0.5	-

同济大学

2.5 本章小结

许多变量均会对能耗产生影响，对于白箱模型，需要耗费大量时间精力输入详细的参数，对于数据驱动模型，维度太高的输入变量会导致计算量呈指数增加。但各个变量对模型的重要性程度是不同的，本研究将影响能耗的变量分成了负荷相关和机电系统相关的变量两类，2.2 节采用 Morris 和相关系数法（SRCC 和

PRCC）提取出负荷相关的关键变量，2.3 节采用 XGBoost 算法提取了机电系统相关的关键变量。2.4 节还将 EnergyPlus 的能耗模拟模型灰箱化，得到了能耗模拟值预测模型，该模型可用于第 3 章中有历史能耗的关键变量缺失值推测。

同济大学

第 3 章 关键变量缺失值推测

3.1 概述

尽管第 2 章中进行了影响空调能耗的关键变量提取，大大减少了能耗预测的输入变量。但在实践中，某些关键变量值是难以测量和获取的，例如新风渗透率几乎无法测量和获取，绝大部分既有建筑的现存资料中也没有体形系数这一参数。本章节要解决的问题便是此类关键变量缺失值的推测，以补全输入变量。关键变量与能耗值是存在联系的，并且某些关键参数之间存在着一定的关联关系，本章节通过挖掘关键变量与关键变量之间，以及关键变量与能耗之间的关系，找到最佳的关键变量值作为该变量的推测值。本章涉及到的关键变量缺失值推测有两种场景：

第一种场景为存在历史能耗数据时的关键变量缺失值推测，已知的历史数据可能是逐时、逐日或者逐月值。在这种场景下将采用遗传算法进行缺失值的推测，将误差最小的模拟数据对应的关键变量值作为最佳关键变量推测值。

第二种场景为无历史能耗数据的关键变量缺失值推测，在这种场景下无法构建输入与输出之间的关系，将利用关联规则挖掘关键变量之间的联系，从而进行关键变量缺失值的推测。

3.2 有历史能耗的关键变量推测-基于遗传算法

在有一定历史能耗数据的场景下，采用遗传算法进行关键变量值缺失值的寻优，将模拟能耗与实际能耗间的误差最小时对应的关键变量值作为最优的关键变量推测值（即最终的关键变量推测结果），这种推测场景可应用于混合模型的数据库建立过程和有历史能耗时的能耗预测。

遗传算法（Genetic Algorithm， GA）是受到生物进化论中遗传和变异的启发，通过模拟自然进化过程进行高效最优解搜寻的方法，该方法的特点是能够在优化空间中自适应调整搜索方向，不需要输入确定的规则，其核心规则是选择、交叉和变异。其执行步骤为：首先随机选取由多个个体组成的初始种群，然后通过随机选择、交叉和变异步骤产生一系列更优的个体以更新种群，使新的种群在搜索空间中到达更优的区域，再通过一代又一代的进化（迭代），最终收敛到最适合环境的个体，从而得到最优解。遗传算法与本研究结合时，首要任务是确定

以下几个要素：决策变量、目标函数和约束条件，遗传算法寻优的决策变量为有缺失值的关键变量，遗传算法优化的目标是最小化关键变量对应的能耗模拟值与实际历史能耗值之间的误差—CV_RMSE，即目标函数为模拟能耗和实际能耗的CV_RMSE。CV_RMSE 是常用的误差衡量指标，其计算公式如式（3.1）所示。约束条件决定了遗传算法的搜索空间，各个变量的取值范围由表 2.1 和表 2.3 确定。

$$\mathrm{CV_{RMSE}} = \sqrt{\frac{\sum_{k=1}^{n}(y_k-\widehat{y_k})^2}{n}} \Big/ \frac{\sum_{k=1}^{n} y_k}{n} \tag{3.1}$$

其中，y_k为实际值，$\widehat{y_k}$为实际值的观测数据，n为观测点个数。

利用遗传算法进行关键变量缺失值推测的技术路线如图 3.1 所示。灰色部分为 2.4 节建立的将白箱模型灰箱化的能耗模拟值预测模型，该模型建立了关键变量和能耗之间的关系，其作用是在遗传算法寻优过程中，输入迭代的关键变量缺失值时该模型能够快速得到能耗预测值。在执行关键变量推测前，需确定待推测的关键变量及其取值范围，待推测的关键变量根据实际情况确定，包含离散型和数值型两类变量，其取值范围根据表 2.1、表 2.3 和表 2.4 确定。在进行遗传算法寻优时，首先，初始化待推测关键变量以补全输入参数，并利用能耗模拟值预测模型快速得到逐时的能耗模拟值。再根据历史能耗的颗粒度进行模拟数据颗粒度的转换，由于能耗模拟值预测模型的输出颗粒度为最细的逐时值，在进行颗粒度转换（颗粒度由细变粗）时仅需在指定时间内进行数据的加和。然后，计算实际数据和模拟数据之间的误差 CV_RMSE，将其作为遗传算法的目标函数。经过多次迭代，目标函数 CV_RMSE 会逐渐减小直至收敛，将收敛时的关键变量推测值视为最优的关键变量值，并将其作为最终的关键变量推测值。

同济大学

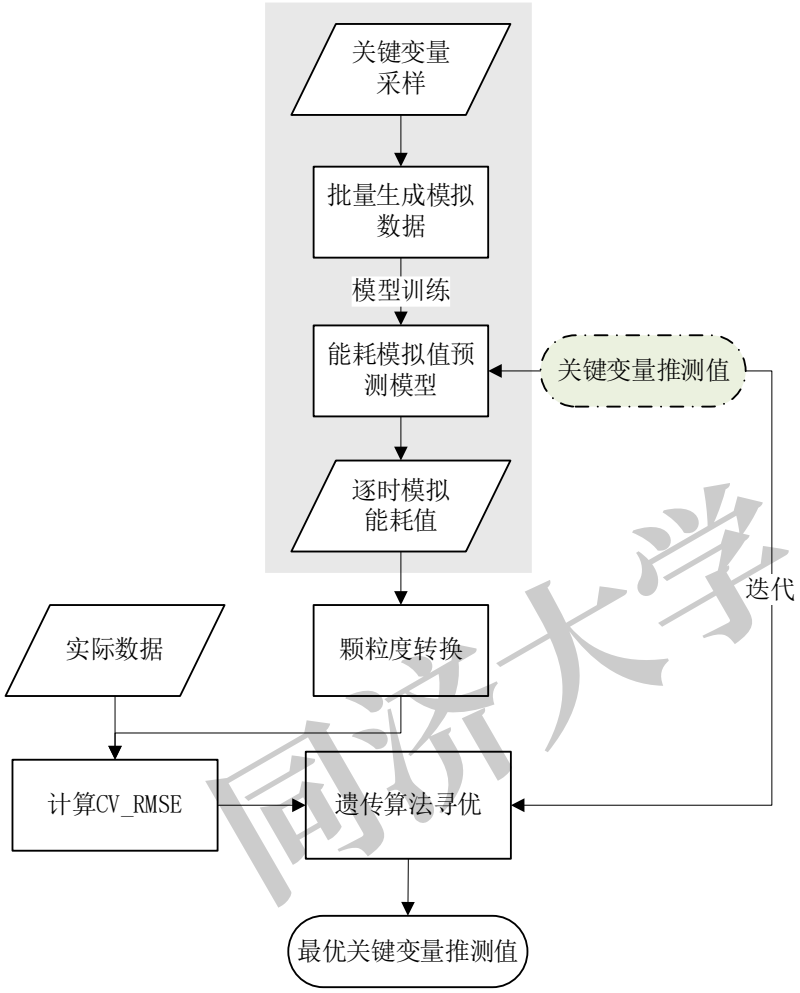


图 3.1 有历史能耗时的关键变量缺失值推测技术路线

下文对遗传算法推测制冷能耗的关键变量缺失值进行了准确性验证。从 2.4 节所述的样本中采用留出法（hold-out）划分了训练集和测试集。留出法是将整个数据集划分为两个互斥的子集，将其中一个作为训练集，另外一个作为测试集的数据划分方法。训练集指的是用于训练能耗模拟值预测模型的数据集，测试集指的是假设某些关键变量值缺失以验证关键变量值推测结果的数据集。为保证验证效果，训练集和测试集的划分要尽可能地保持其数据分布的一致，故本研究的测试集是从全样本中分层随机抽取的 15 个样本。

3.2.1 和 3.2.2 节将介绍验证集中单个变量缺失和多个变量缺失时的推测结果，由于类别型变量（如风系统类型）的值通常是已知的，本章主要分析了数值型变量的预测误差和误差的分布，数值型变量的误差采用如式（3.2）所示的相对误差来表示。

$$\delta = \frac{|y_k - \widehat{y_k}|}{y_k} \tag{3.2}$$

其中，y_k是实际值，$\widehat{y_k}$是实际值的观测数据。

同济大学

3.2.1 单个变量缺失时的推测结果验证

由于本研究获取的能耗数据对应的关键变量均有或多或少的缺失值，本节采用 3.1 节所述的测试集进行关键变量推测算法的验证和关键变量推测值准确性的评价。在实际情况下，存在的历史能耗数据可能是不同颗粒度的，本节在已知逐时或逐日历史能耗两种情景下对单个关键变量缺失值推测的准确度进行了探究。

在已知逐时能耗的情景下，假设测试集中某一个关键变量数值缺失，其他变量均已知，此时采用遗传算法对测试集中单个关键变量缺失值推测的相对误差分布如图 3.2 所示。图中横坐标为变量名称的缩写（缩写与变量名称的对应见表 2.1、表 2.3 和表 2.4），纵坐标为相对误差，由于此种情景下相对误差普遍偏小，在绘图时将纵坐标的刻度进行了对数处理。误差分布由图 3.2 所示的箱线图表示，箱线图中的每一个箱子代表在其横坐标对应的变量缺失但其他变量均已知的情况下，测试集的多个样本中该变量推测的相对误差分布，箱体表示第一个四分位数（Q1）和第三个四分位数（Q3）间的距离（即四分位距 IQR），箱体中的横线表示中位数，箱体越短表示推测误差分布越集中；箱体外上下的两条横线表示上下限值，上限值为 Q3+IQR，下限值为 Q1-IQR，在上下限值外的点可视为离群值，在图 3.2 中用菱形的点表示。从图 3.2 中可以看出，在逐时能耗数据存在的情况下，对单个变量的缺失值推测精度整体较高，平均误差都在 1%以下。对制冷设定温度（SPC）的缺失值推测准确度最高，其平均相对误差最小且误差分布也较为集中；而对层数（NL）的缺失值推测的平均相对误差最大，其误差分布也相对较为分散，且有较多的离群点；对体形系数（CR）缺失值推测的最大相对误差最大，且其相对误差的离群点最多也最远，说明对这一参数的缺失值推测效果不稳定。

在已知逐日能耗的情景下，同样假设测试集中某一个关键变量数值缺失，其他变量均已知，此时采用遗传算法对测试集中单个关键变量缺失值推测的相对误差分布如图 3.3 所示。从图 3.3 中可以看出，在已知逐日能耗时，各个变量的预测准确度规律变化不大，同样是制冷设定温度（SPC）为预测最准确的变量，层数（NL）为平均误差最大的变量，单个变量缺失值的推测误差整体上大于已知逐时能耗的推测误差，这是由于更细颗粒度的能耗数据中包含了更多的信息。平均相对误差的差异并不是非常明显，但是已知逐日能耗情景下的推测误差分布更分散，离群值也更多，这表明推测结果的不确定性更大，推测效果更不稳定。

两种情景下单个变量的推测误差见表 3.1。

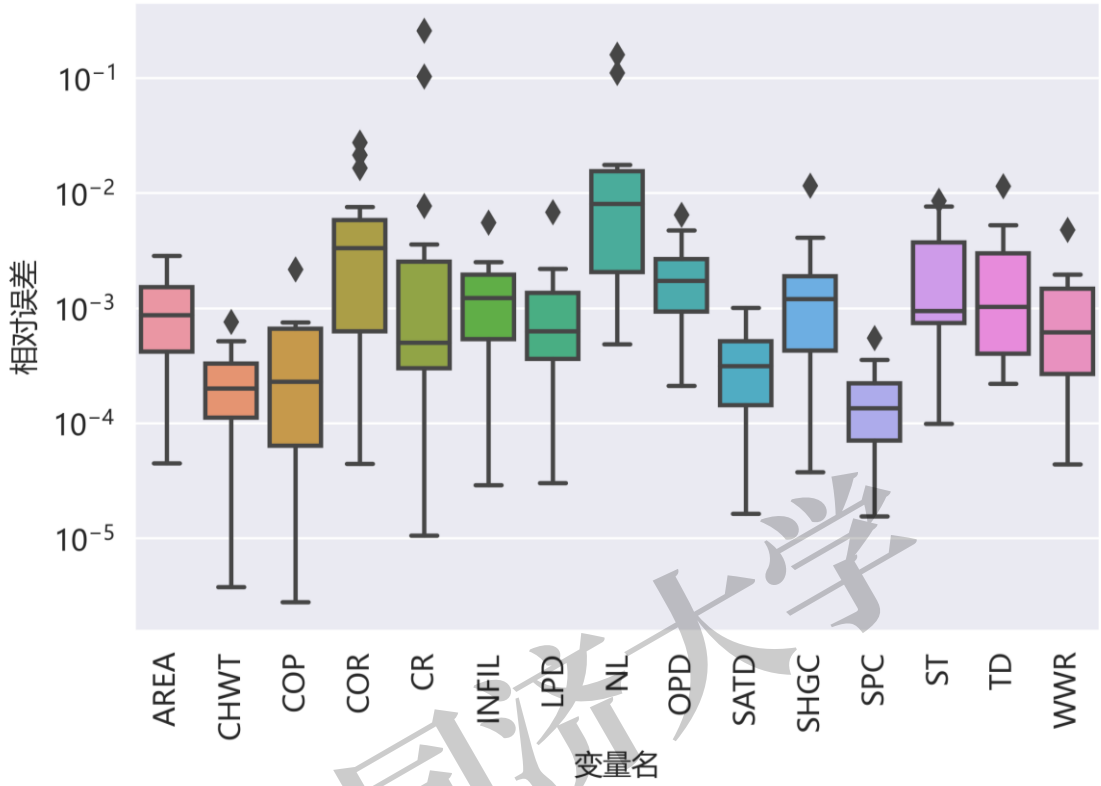


图 3.2 已知逐时能耗情景下，对单个变量缺失值的推测误差分布

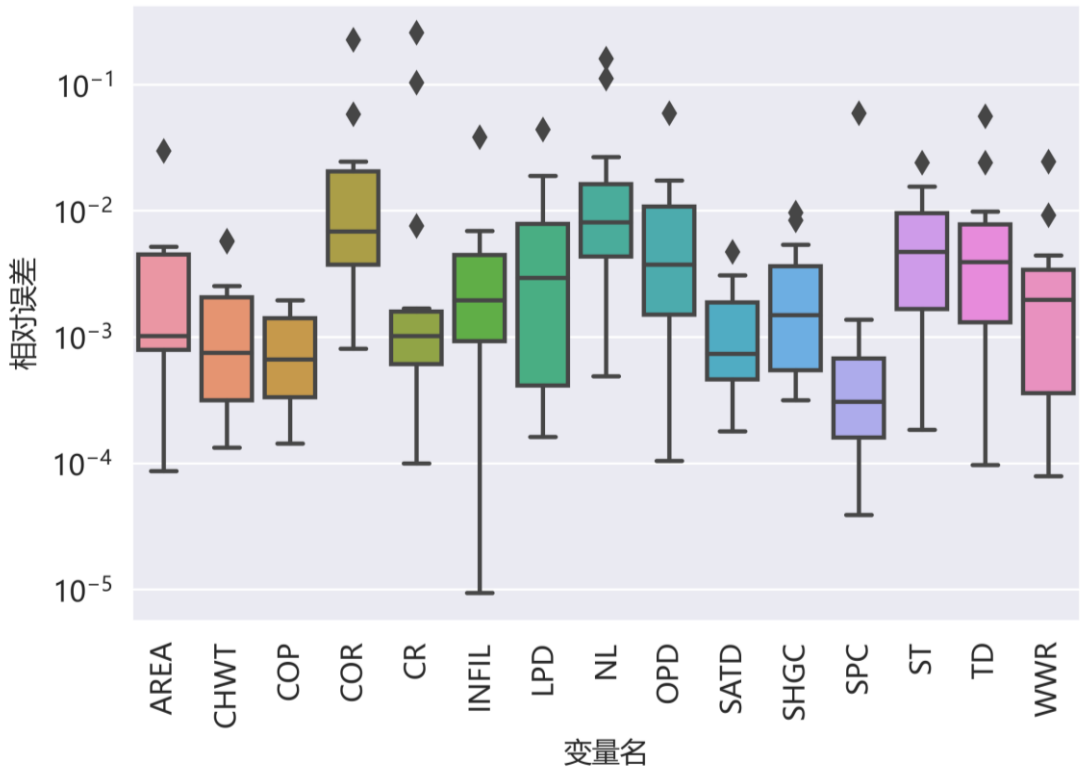


图 3.3 已知逐日能耗情景下，对单个变量缺失值的推测误差分布

同济大学

表 3.1 单个变量缺失值的推测相对误差

变量名	缩写	平均误差-逐时	最大误差-逐时	平均误差-逐日	最大误差-逐日
面积	AREA	0.001063	0.002858	0.003924	0.02988
冷冻水供水温度	CHWT	0.000251	0.000759	0.001596	0.005787
冷机/热泵 COP	COP	0.000423	0.00217	0.000832	0.001953
商业与办公面积比	COR	0.006145	0.027648	0.026507	0.227161
体形系数	CR	0.025143	0.257099	0.025197	0.257027
新风渗透率	INFIL	0.001532	0.005579	0.004964	0.038328
照明功率密度	LPD	0.001273	0.006851	0.007184	0.043941
层数	NL	0.024886	0.160121	0.025817	0.160121
人员密度	OPD	0.002067	0.006474	0.009105	0.059157
送风温差	SATD	0.000354	0.001011	0.001357	0.00473
太阳得热系数	SHGC	0.001876	0.01162	0.002711	0.00967
制冷设定温度	SPC	0.000178	0.000551	0.004368	0.059541
内遮阳开启程度	ST	0.00239	0.008609	0.00664	0.024045
冷冻水供回水温差	TD	0.002274	0.011465	0.008447	0.05589
窗墙比	WWR	0.001081	0.004808	0.003578	0.024458

3.2.2 多个变量缺失值时的推测结果验证

在实际应用中通常会存在多个变量缺失的情况，本节采用 3.1 节所述的测试集，对比了已知逐时或逐日历史能耗两种情景下不同数量的关键变量缺失值推测误差。

以缺失制冷设定温度（SPC）和 COP 为例，介绍已知逐时或逐日历史能耗两种情景下两个变量缺失时的推测精度。假设测试集输入变量中的 SPC 和 COP 数值缺失，其他变量已知，采用遗传算法对这两个变量的推测相对误差如图 3.4 所

示。图 3.4 左边为已知逐时能耗情景下的推测误差分布，右边为已知逐日能耗情景下的推测相对误差分布，可见在两种情景下进行两个变量缺失值推测的平均相对误差均小于 0.001，逐日场景下的推测误差与逐时场景下的推测误差相比稍微高一些，且不确定性更大，但仍在可接受的范围内。

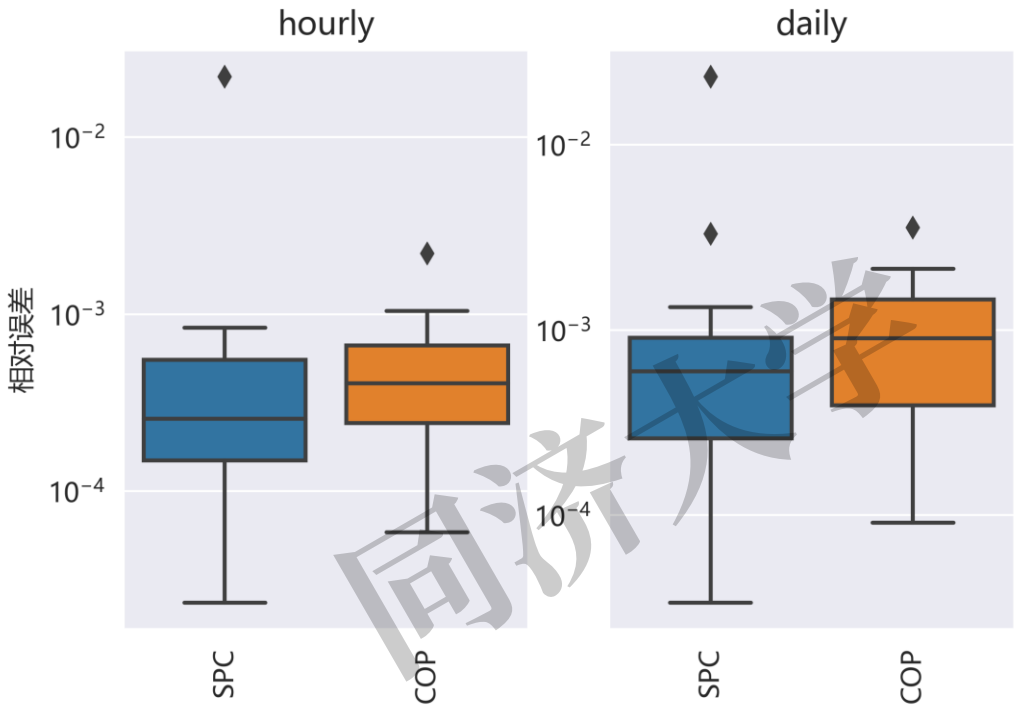


图 3.4 两种场景下，两个变量缺失时的推测误差

在上述示例基础上，以缺失制冷设定温度（SPC）、COP 和人员密度（OPD）为例，介绍已知逐时或逐日历史能耗两种情景下三个变量缺失时的推测精度。假设测试集输入变量中的 SPC、COP 和 OPD 数值缺失，其他变量已知，采用遗传算法对这三个变量的推测相对误差如图 3.5 所示。可见在两种情景下进行三个变量缺失值推测的平均相对误差均小于 0.01，误差均在可接受的范围内，其中 SPC 的推测误差受能耗数据时间颗粒度的影响最大。

同济大学

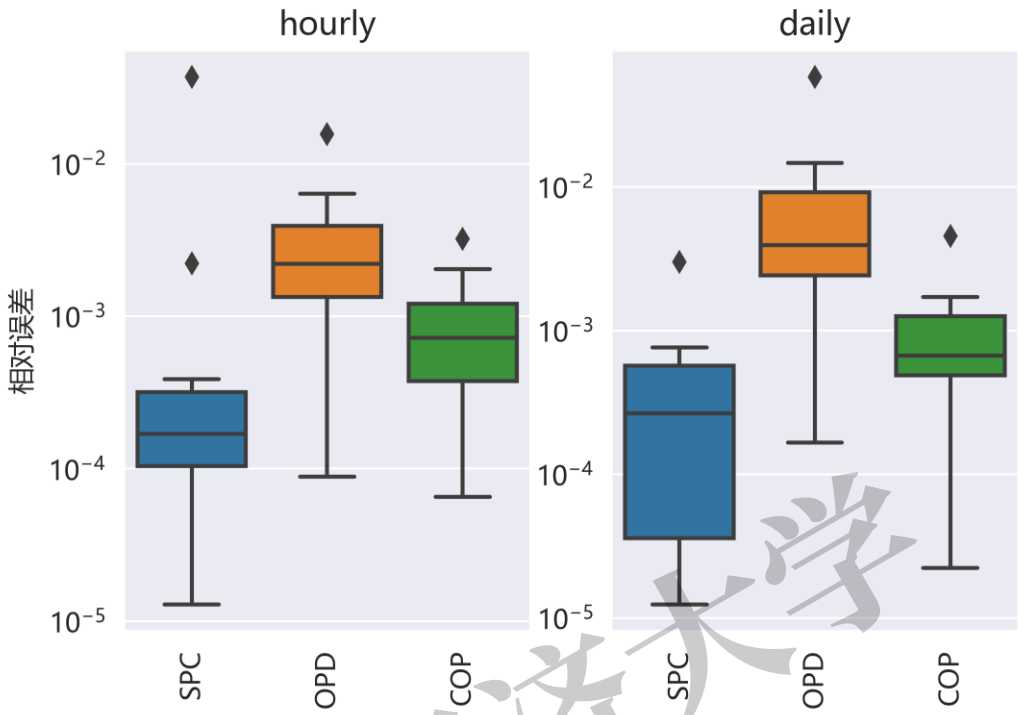


图 3.5 两种场景下，三个变量缺失时的推测误差

在上述示例基础上，再增加一个缺失变量，以缺失制冷设定温度（SPC）、人员密度（OPD）、送风温度（SATD）和 COP 为例，介绍已知逐时或逐日历史能耗两种情景下四个变量缺失时的推测精度。假设测试集输入变量中的 SPC、OPD、SATD 和 COP 数值缺失，其他变量已知，采用遗传算法对这四个变量的推测相对误差如图 3.6 所示。由图可见，在两种情景下进行四个变量缺失值推测的平均相对误差仍小于 0.01，误差仍在可接受范围内，与缺失三个变量相比，四个变量缺失时推测的相对误差分布明显更加分散（箱线图中的箱体变长，且存在更多离群值），这表明信息量减少时，推测的不确定性会增大。与逐时场景相比，图中逐日场景下的箱体整体向上偏移，推测的误差整体变大，离群值也增多，受能耗数据颗粒度影响最小的是 SATD，虽然逐日场景下，SATD 推测的离群值增多，但平均相对误差反而有些许降低，误差分布也更为集中。

同济大学

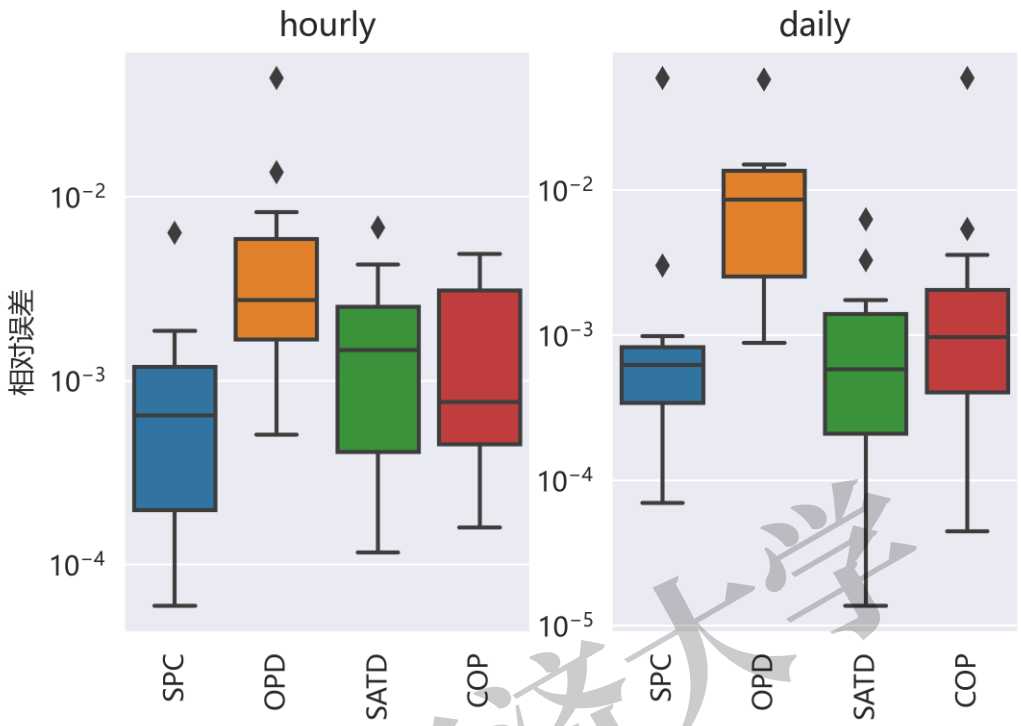


图 3.6 两种场景下，四个变量缺失时的推测误差

从上述结果看，采用遗传算法进行缺失变量的推测准确性是较高的，最后本研究旨在探究已知变量极少的极端情况下，假设在所有的数值型变量中只已知五个变量，分析同时推测其他所有数值型变量时的推测误差。同样采用 3.1 节所述的测试集，假设仅已知制冷设定温度（SPC）、层数（NL）、面积（AREA）、商业部分面积占比（COR）、冷冻水供水温度（CHWT）这五个数值较为容易获取的变量，采用遗传算法对同时其他变量进行推测的相对误差见图 3.7，平均相对误差如表 3.2 所示。可以看出，在此极端情况下，各推测变量平均相对误差的最大值在 0.5 左右，各推测变量相对误差的中位数的最大值在 0.2 左右，推测误差最大的变量为 OPD 和 SHGC，在这种情况下推测结果最好的是 CR、SATD 和 COP。从图 3.7 可见，在推测 OPD、LPD、SHGC 和 TD 时，少数几个样本的相对误差存在大于 1 的情况，但这几个变量的推测误差平均值和中位数均不大于 0.5，在实际应用时应降低对这几个参数推测值的置信度。

同济大学

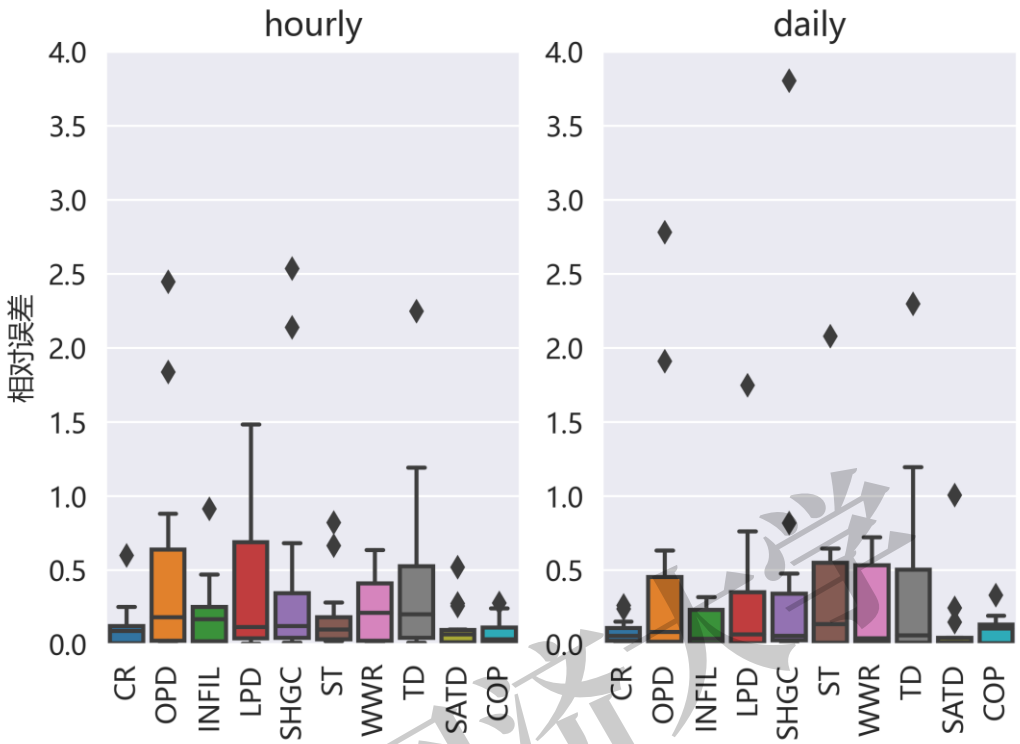


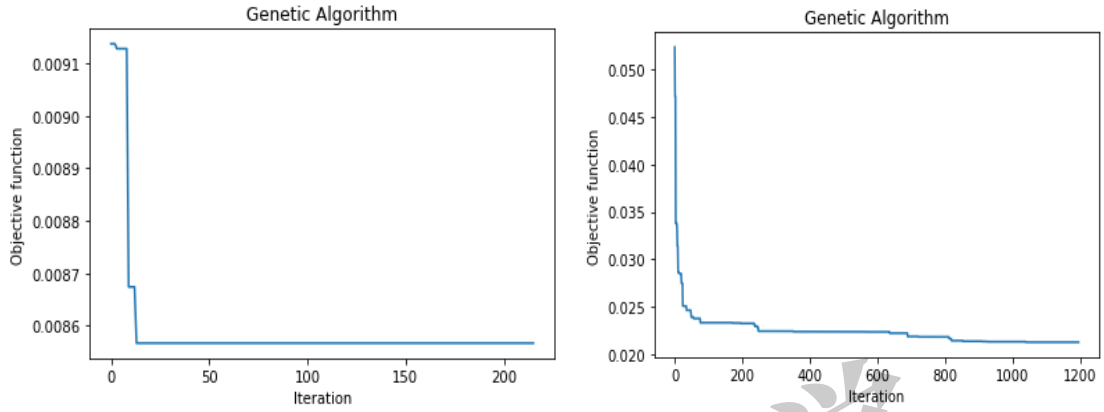
图 3.7 两种场景下，已知五个变量，推测其他所有变量时的推测误差

表 3.2 极端情况下同时推测多个变量的相对误差

变量名	缩写	相对误差平均值-逐时	相对误差平均值-逐日	相对误差中位数-逐时	相对误差中位数-逐日
体形系数	CR	0.11	0.079	0.08	0.05
人员密度	OPD	0.50	0.48	0.18	0.08
新风渗透率	INFIL	0.21	0.10	0.16	0.03
照明功率密度	LPD	0.37	0.27	0.11	0.06
太阳得热系数	SHGC	0.46	0.42	0.12	0.05
内遮阳开启程度	ST	0.19	0.35	0.09	0.13
窗墙比	WWR	0.23	0.50	0.21	0.03
冷冻水供回水温差	TD	0.44	0.37	0.20	0.05
送风温差	SATD	0.10	0.11	0.06	0.01
冷机/热泵 COP	COP	0.07	0.09	0.03	0.10

图 3.8 展示了单个变量缺失和多个变量缺失时利用遗传算法进行变量值推测的收敛曲线对比，图中横坐标为迭代步数，纵坐标为目标函数值，即实际值和模

拟值的 CV_RMSE。可以看出推测多个变量所需的迭代步数远远大于单个变量推测时所需的迭代步数，且目标函数的绝对值也远高于单个变量推测时的目标函数。



a) 单个变量推测　　b) 多个变量推测

图 3.8 单个变量和多个变量推测收敛情况对比

3.3 无历史能耗的关键变量推测-基于关联规则

在不存在历史能耗数据的场景时，拥有的信息非常有限，无法根据关键变量与能耗之间的关系进行关键变量的推测。在这种情况下，将挖掘关键变量之间的关联规则，利用已知关键变量进行缺失关键变量值的推测。这种推测方法可在通常没有历史能耗数据的跨建筑能耗预测时应用。

在实际中，同一栋建筑的某些特征变量之间存在一定的联系，比如建筑楼层越大往往建筑面积也越大，建筑建造年份越早通常建筑内设备的效率越低且设备老旧程度越大，冷机的 COP 越低可能建筑未进行节能改造，建筑内其他设备的效率较低的概率也较大。在无法获取变量与能耗之间的联系时，本节通过关联规则挖掘算法挖掘已有数据中变量与变量之间关联规则，根据该关联规则进行关键变量缺失值的推测。关联规则挖掘基于频繁项集、事件 A、B 同时发生的概率和事件 A、B 发生的条件概率等，假设 A 和 B 是 2 不相交的项集，即$A \cap B = \emptyset$，其关联规则的强度可以用二者的支持度（support）和置信度（confidence）来衡量。支持度的公式如式（3.3）所示，表示了事件发生的概率，置信度的公式如式（3.4）所示，表示了事件 B 发生时，事件 A 发生的条件概率。

$$\text{Support}(A, B) = P(A, B) = \frac{num(AB)}{num(allsamples)} \tag{3.3}$$

$$\text{Confidence}(AB) = P(A|B) = \frac{P(AB)}{P(B)} \tag{3.4}$$

满足最小支持度的项集即为频繁项集，最小支持度是人为规定的阈值，它代表了该项集在统计学上的最低重要性，支持度大于最小支持度的项集才能成为频繁项集。最小置信度也是人为规定的阈值，它代表了关联规则的最低可靠性，只有支持度和置信度均大于最小值（即规定的阈值）时，该关联规则才会被视为强规则。本节采用 Apriori 算法来进行频繁项集和关联规则的挖掘。该算法基于频繁项集的两条性质：

a) 频繁项集的所有非空子集也为频繁项集；

b) 若 A 项集不是频繁项集，则其他事件或项集与 A 的并集也不是频繁项集。

若数据集过大时，通过遍历进行频繁项集的搜索会导致计算量过大，Apriori 算法利用上两条性质，根据支持度进行扫描和剪枝（即去掉支持度小于最小支持度的项），从而得到频繁项集

频繁项集中可能有多个元素，包含 k 个元素的项集称为 k-项集。在进行 k 值不同的频繁项集的搜索时，Apriori 算法进行元素个数为 k 的频繁项集搜索后，令 k = k+1，生成候选频繁 k+1 项集。

由于本研究中第 2 章得到的模拟能耗数据库中的关联变量是拉丁超立方抽样后组合得到的，不具有实际意义，而真实数据中所有关键变量值均存在的情况极少，本节采用的数据集为 3.2 节中根据历史能耗的数据补全关键变量后的数据集（实际建筑数据集），基于该数据集进行实际建筑中关键变量的关联规则挖掘，在应用时，无需历史能耗数据，直接利用该关联规则，通过已知的关键变量，进行缺失的关键变量值的推测。

本研究中 Apriori 算法挖掘关键变量关联规则从而进行缺失变量值推测的步骤为：

1) 连续变量离散化，由于基于 Apriori 的关联规则挖掘算法仅适用于离散型变量，故首先进行连续变量的离散化，根据连续型关键变量的取值范围进行等距离的离散化。

2) 基于上述的实际建筑数据集，进行各变量间关联规则的挖掘，设置最小支持度为 0.4，最大的 k 值为 4，进行 2 项、3 项和 4 项频繁项集的挖掘。

3) 若缺失关键变量和已知关键变量之前存在关联关系，则根据关联规则表进行缺失值的推测，若已知关键变量大于 3 项，则根据将已知变量进行 2 项及 3 项的排列组合，进行多次推测，然后取多次推测的最大值和最小值作为关键变量推测值的上下限。

4) 进行推测的离散关键变量值的还原，取与该值最近的两个离散值作为该被推测值的上下限，并在该上下限区间内采用拉丁超立方抽样，得到被推测值的

分布。若不存在关联关系则假设其在取值范围内均匀分布，并在该区间内进行拉丁超立方抽样，得到被推测值的分布。

本节介绍的利用 Apriori 关联规则挖掘算法，基于已知关键变量进行缺失关键变量值的推测为不确定推测，得到的结果为在推测变量值所在的区间内的抽样结果。

本研究对上述无历史能耗场景下，采用关联规则挖掘算法进行缺失值推测的结果进行了验证。假设缺失可能性较高的几个参数：体形系数（CR）、人员密度（OPD）、新风渗透率（INFIL）、送风温差（SATD）同时缺失，其他变量已知，采用上述 Apriori 关联规则挖掘算法对缺失值进行推测。表 3.3 展示了该四个参数的缺失值推测结果，其结果表示为在推测缺失值可能存在的区间内的抽样结果。从表 3.3 中可见，缺失变量的实际值均在推测区间内，推测结果较为可信。

表 3.3 采用关联规则进行缺失值推测的结果

变量缩写	**CR**	**OPD**	**INFIL**	**SATD**
实际值	0.282	0.0732	0.68	8.1
推测区间内的抽样结果	0.255	0.0625	1.15	8.9
	0.275	0.2125	0.75	8.5
	0.285	0.1125	0.55	8.3
	0.335	0.2875	1.45	9.9
	0.345	0.2625	1.05	8.1
	0.315	0.1875	0.95	9.5
	0.305	0.0875	1.35	9.3
	0.265	0.1375	0.65	9.1
	0.325	0.2375	0.85	8.7
	0.295	0.1625	1.25	9.7

3.4 本章小结

目前的能耗监测平台往往只储存了建筑能耗数据，对于能耗预测非常重要的建筑本体的体形系数和窗墙比等关键信息往往是缺失的，并且新风渗透率等信息往往是不可得也不可测的。在进行多源异构数据库的建立和建筑能耗预测时，我们通常面临着关键变量数值缺失的情况。本节介绍了在有历史能耗数据和无历史能耗数据两种场景下的关键变量推测方法，第一种方法常用于建筑能耗数据充足的数据库建立过程，第二种方法常用于在进行完全没有历史能耗数据的跨建筑或新建建筑能耗预测时。3.2 和 3.3 节分别从算法、数据集、推测过程等方面对关键变量缺失值推测进行了阐述。3.2 节介绍了在不同颗粒度的能耗数据下，通过遗传算法进行关键变量缺失值的推测过程及推测结果。通过遗传算法有效地寻找

同济大学

到模拟数据和实测数据 CV_RMSE 最小时的关键变量组合。关键变量的相对推测误差随着能耗数据的颗粒度的增大和缺失变量的增大逐渐增大，在缺失少数几个变量时（4 个及以下），在各种颗粒度下的关键变量推测误差均较小，在仅已知少数关键变量（已知 5 个易得的关键变量）的极端情况下，同时推测其他所有关键变量的平均相对误差也是可接受的，但存在着些许离群值，此时建议降低推测变量的置信度。3.3 节介绍了在无历史能耗情况下，通过 Apriori 关联规则挖掘算法挖掘各个关键变量之间的关联规则，根据各个关键变量之间的关联规则进行缺失关键变量的推测的过程，从验证结果来看，缺失变量的实际值均在推测得到的区间内，结果较为可信。

同济大学

第 4 章 多源异构数据的能耗数据融合

4.1 概述

随着建筑智能化水平的提高和建筑运行节能概念的普及，越来越多的公共办公建筑进行了节能审计和节能改造，大型公共建筑均与能耗监测平台进行了联网。本研究从多个途径搜集了办公建筑的能耗数据，这些数据包括：

1) 能耗监测平台的分项能耗数据，能耗监测平台上可获取的信息如图 4.1 所示；
2) 节能审计报告中的建筑物理信息与能耗月账单数据；
3) 第 2 章快速模拟工具生成的能耗模拟数据。

图 4.1 能耗审计平台上可获取的信息

上述三类数据各自有着不同的特性，在进行综合利用之前需要进行不同的处理：

1) 能耗监测平台的数据：能耗监测平台的数据较能耗审计数据而言时间颗粒度较细，分项计量数据为逐日数据。但由于无法避免传感器异常或传输信号异常的情况，此类数据的异常值和缺失值较多，并且对于建筑的描述较少，通常只有建筑面积、层数等信息。对于这部分数据需进行异常值的识别和修复、缺失值的填补以及缺失的关键变量的推测。

2) 节能审计报告数据：此类数据中建筑相关的信息较为丰富，质量较高的节能审计报告中有窗墙比、建筑外围护结构传热系数等参数，审计报告中的能耗数据来自于月账单，可信度较高。但此类能耗数据往往是逐月数据，时间颗粒度较大。对于这部分数据需进行数据颗粒度的转换以及少部分关键变量的推测。

3) 模拟数据：这部分模拟数据来自将关键变量作为输入，利用第 2 章中的介绍的快速建模工具模拟计算生成的数据。模拟数据的特点是：建筑能耗数据详细且颗粒度灵活可调，并且有与之对应的建筑信息相关的变量。但由于能耗模型存在很多简化，建模也没有考虑到运行施工带来的不确定性，故模拟数据的可信度较低。本研究建立模型对实测数据和模拟的偏差进行刻画，利用该偏差模型进

同济大学

行模拟数据的修正。

下文将详细阐述对上述三类能耗数据的处理。

4.2 能耗监测平台的分项计量数据处理

对能耗监测平台的分项能耗数据需进行异常值的识别和修复以及缺失值的填补。对分项能耗数据的处理流程如图 4.2 所示.

第一步，进行异常值的判断。异常值包括离群值和异常波动值，例如长时间的死值。离群值的判断采用箱线图的判断方法进行异常值的识别。箱形图可用于反映原始数据分布的特征及进行多组数据分布特征的比较，此外，还可以有效地判断数据中的异常值，是实际使用过程中较为常见的方法。在箱形图中，异常值被定义为小于$U-1.5IQR$或大于$L+1.5IQR$的值 (其中 U、L分别为上下四分位数，IQR 为四分位距，是上四分位数与下四分位数之差，其间包含了全部观察值的一半)。异常波动值的判断利用采用 python 开发的 tsod 模块[63]进行突变值和死值的判断。当一栋建筑的异常数据量占比超过 15%时，则认为该建筑异常数据过多，则将该建筑从数据中剔除。

第二步，进行异常楼宇的剔除。有的建筑即使异常值较小，但单位面积能耗值和各分项能耗占比与经验值差异太大也应被视为数据异常的楼宇。具体来说，以下三类建筑将会被视为数据异常的建筑：1）在能耗时间序列的 1/3 以上的时间中，能耗监测平台上得到的建筑总能耗数据与分项能耗数据之和的误差大于 1 的建筑；2）空调分项的最大值和总能耗的最大值的比值小于 25%的建筑；3）单位面积全年能耗强度不在 50~150 kWh/m^2·a 的建筑。

第三步，进行异常值的修复和缺失值的补全。这里将识别出来的异常数据去除，并与缺失值一起填补。填补缺失值的方法是：首先寻找与待填补建筑能耗密度最相似并且缺失值处有正常数据的相似建筑，然后用待填补建筑能耗密度与相似建筑能耗密度的比值乘以相似建筑同一时间的能耗值作为填补值。缺失值的填补效果如图 4.3 所示。

本研究共搜集了 105 栋办公建筑的分项计量数据，经过异常值的判断和异常楼宇的剔除步骤后，仅剩下 30 栋数据较为正常的建筑可供后续研究使用。

同济大学

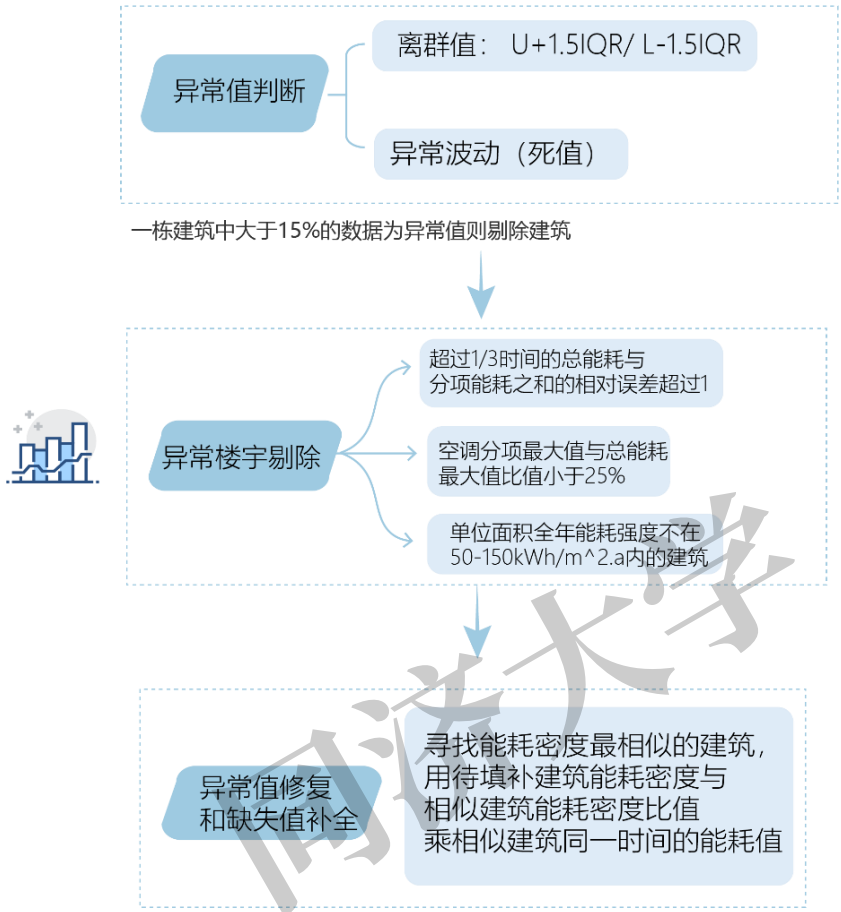
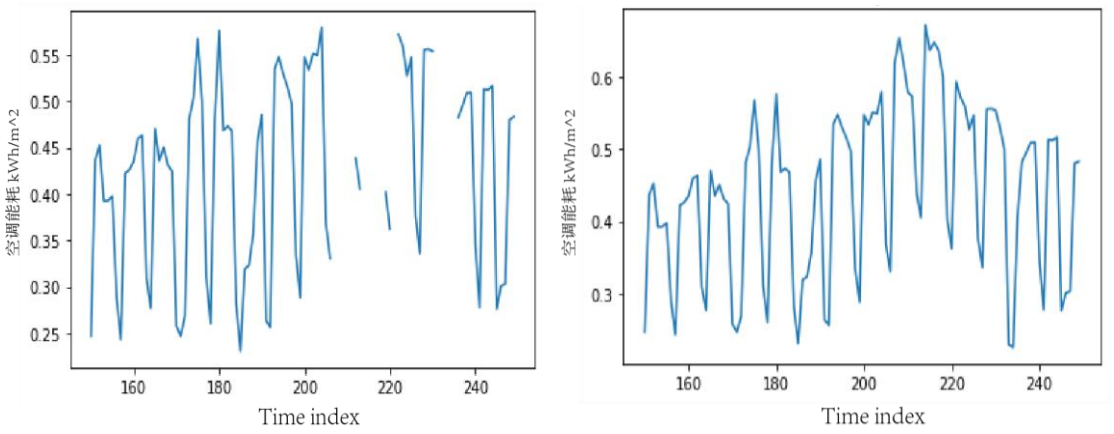


图 4.2 分项能耗数据异常值和缺失值处理流程



a) 填补前　　b) 填补后

图 4.3 缺失值填补效果

4.3 节能审计报告的能耗数据处理

来自节能审计报告的能耗数据置信度较高，几乎不存在缺失值和异常值，但

这部分数据时间颗粒度较粗，通常是逐月数据。但要进行数据融合需首先保证数据的颗粒度是一致的，而对这部分数据进行的处理主要是时间颗粒度的细化。本研究从逐日数据中提取出典型逐日能耗曲线，再用典型逐日能耗曲线进行逐月数据的填充。由于数据量的限制，有节能审计报告的楼宇均不存在逐日分项计量数据，故本部分采用的数据均来自与能耗计量平台，在进行颗粒度转换结果的验证时，将逐日的分项计量平台的数据折算为逐月数据，作为时间颗粒度转换的输入数据，最终将时间颗粒度细化后的逐日数据与原始的逐日数据进行对比。

本研究采用 K-Means 聚类的方法进行典型逐日能耗曲线的提取。K-Means 算法是无监督的基于距离的分类算法，它将把包含 n 个样本的数据集分为 k 个无交集的簇，簇中所有数据的均值为该簇的质心。K-Means 算法的执行步骤是：

第一步，初始化质心，将随机抽取 k 个样本作为初始的质心。

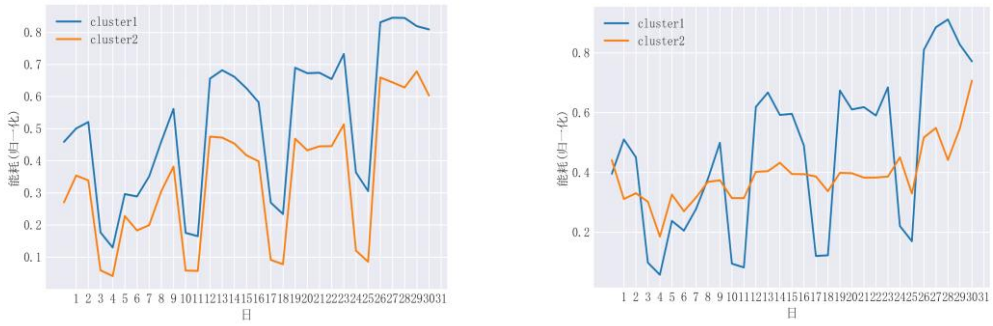
第二步，进行质心迭代求解。首先，将每个样本归到距离他们最近的质心，生成 k 个簇；然后在每个簇内，重新计算该簇的质心，再重新进行样本的归类。

最后，多次迭代后，每个簇的质心不再发生改变时则停止迭代，完成聚类。

对 K-Means 算法的聚类结果影响较大的是距离的度量方法和 k 值的选择。

- 距离度量方法

常用的距离度量方法有欧氏距离、马式距离、曼哈顿距离和余弦距离等。由于时序数据的特殊性，本研究采用的是 DTW（Time series Warping）来度量两个时间序列的相似性（即距离），该方法在声音处理等时序数据中已有较多的应用，在能耗时序数据中也有较多应用[64]。DTW 尤其适用于不同长度、不同节奏的时间序列的相似度度量，有的时间序列整体的波形形状相似，但在存在着相位的差异，该方法将时间序列进行 warping 扭曲，使得两个时间序列相位对齐，从而判断两个时间序列的波形是否相似。图 4.4 表示了在同一数据集下（办公建筑 7 月的能耗数据），采用欧式距离和 DTW 作为相似度度量值的聚类结果，可以看出，采用欧式距离只能将能耗绝对值相近的建筑聚为一簇，而采用 DTW 作为距离度量指标可将曲线形状相似的建筑聚为一簇。本研究的目的是聚类出不同波形的典型逐日能耗曲线，故采用 DTW 作为相似度度量值是合理的。



a) 欧氏距离聚类结果　　b) DTW 聚类结果

图 4.4 采用不同距离度量方式的聚类结果对比

- K 值的选择：基于轮廓系数

K 值的最佳取值采用轮廓系数来确定，轮廓系数可以同时衡量簇内稠密程度（簇内相似度）和簇间离散程度（簇间差异度），是常见的评估聚类效果的指标，单个样本的轮廓系数定义式如式（4.1）所示。轮廓系数范围是(-1,1)，其中轮廓系数越接近 1 则表示样本与其所在的簇中的样本约相似，与其他簇中的样本越不相似。轮廓系数越接近-1，则表示当前样本与簇外的样本更相似。若一个簇中的大部分样本具有较高的轮廓系数，则该簇整体的轮廓系数会较高，整个数据集的平均轮廓系数也较高，可认为此时聚类的效果较为理想，故将平均轮廓系数最高的 k 值作为最佳的 k 值。

$$s = \frac{b-a}{\max(a,b)} \tag{4.1}$$

其中，a 为样本与其所在簇中的其他样本的相似度，等于该样本与簇中其他样本的平均距离，b 为样本与其他簇中样本的相似度，等于该样本与其他簇中样本的平均距离。

结合本课题的数据集，采用上述聚类方法进行时间颗粒度细化的流程是：

第一步，将 4.2 节中数据清洗后的能耗值按照式(4.2)进行最大最小归一化。

$$x' = \frac{x-min}{max-min} \tag{4.2}$$

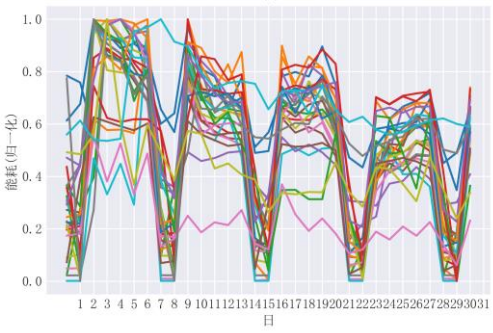
式中，x'表示的是归一化之后的值，x表示归一化前的数值，min表示该数据的最小值，max表示该数据的最大值。

第二步，采用上述的 K-Means 算法对逐日能耗数据时间序列按照月份进行聚类，将 DTW 作为时间序列的距离（相似度）度量指标，对于 k 值的确定，采用轮廓系数最接近 1 下的 k 值。这一步得到了每个月的典型逐日能耗曲线。由于本课题获取的部分办公建筑冬季采用锅炉供暖，而夏季几乎均采用空调制冷，这导致制冷季和供热季聚类结果差异较大，下文将在冬夏季各选取一个月进行聚类

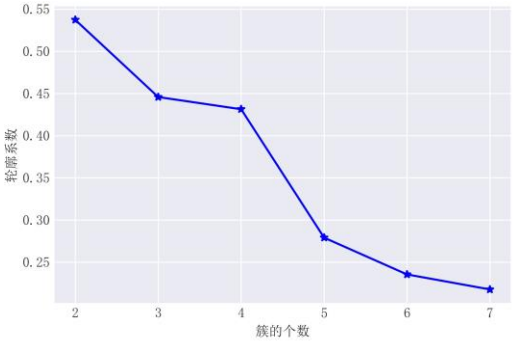
过程的说明，其他月份的聚类结果见附录 A，其中除 5、10、11 月聚为三类外，其他月份均聚为两类。

对于夏季的能耗时序数据聚类，选取 8 月的数据为例阐述聚类过程与结果，如图 4.5 所示，a) 为所有办公建筑 8 月份归一化之后的逐日空调能耗；b) 展示了簇的个数（即 k 值）对平均轮廓系数的影响，可见当簇的个数为 2 时，平均轮廓系数最大，故选取 2 作为簇的个数，将制冷能耗聚为两类；c) 与 d) 分别为归类为第一簇和第二簇的建筑的空调能耗；e) 为两簇的聚类中心，即两类建筑的典型逐日能耗曲线。分析同一簇中的建筑特性可知，大部分第一簇中的建筑房间功能以办公为主，商铺等功能的占比较小，第二簇中的建筑有较大的商业面积，大部分为综合建筑。

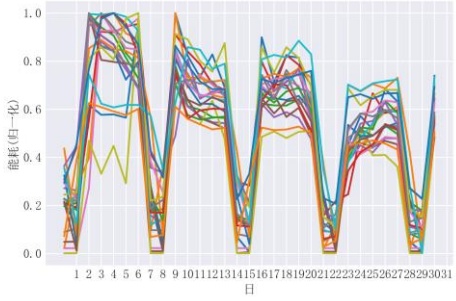
对于冬季的能耗时序数据聚类，选取 1 月的数据为例阐述聚类过程与结果，如图 4.6 所示，a) 为所有办公建筑 1 月份归一化之后的逐日空调能耗；b) 展示了簇的个数（即 k 值）对平均轮廓系数的影响，同样选取 2 作为簇的个数，将空调能耗聚为两类；c) 与 d) 分别为归类为第一簇和第二簇的建筑的空调能耗；e) 为两簇的聚类中心，即两类建筑的典型逐日能耗曲线。分析冬季同一簇中的建筑特性可知，第二簇中的建筑电耗极低，分析存在不消耗电能的热源对建筑进行供热或者是热负荷非常小的商业建筑，而第一簇中的建筑电耗较高，为热负荷较大的建筑。



a）所有办公建筑 8 月空调能耗

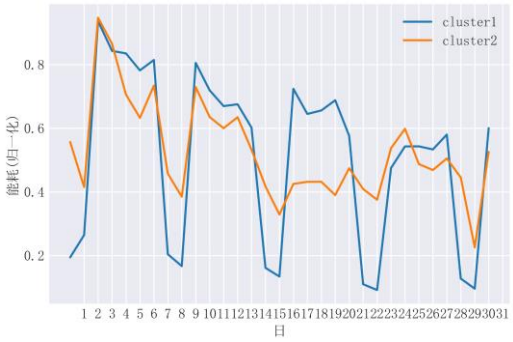


b) 簇的个数对平均轮廓系数的影响



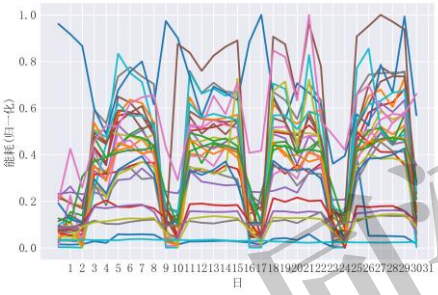
c) 第一簇中建筑的空调能耗

d) 第二簇中建筑的空调能耗

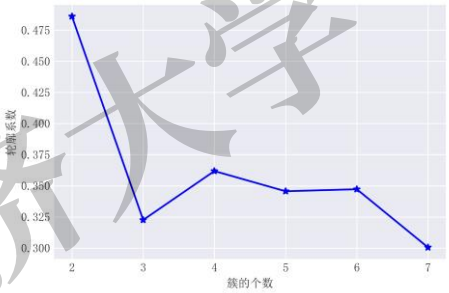


e) 两簇的聚类中心

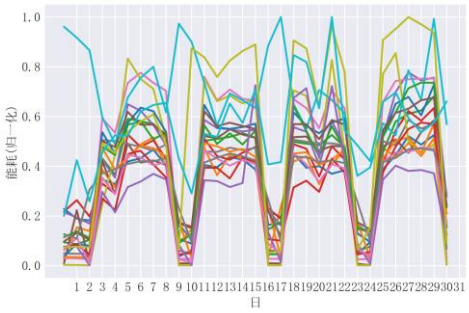
图 4.5 8 月空调能耗典型逐日能耗聚类



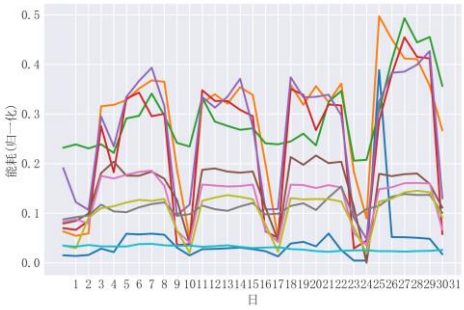
a) 所有办公建筑 1 月空调能耗



b) 簇的个数对平均轮廓系数的影响



c) 第一簇中建筑的空调能耗



d) 第二簇中建筑的空调能耗

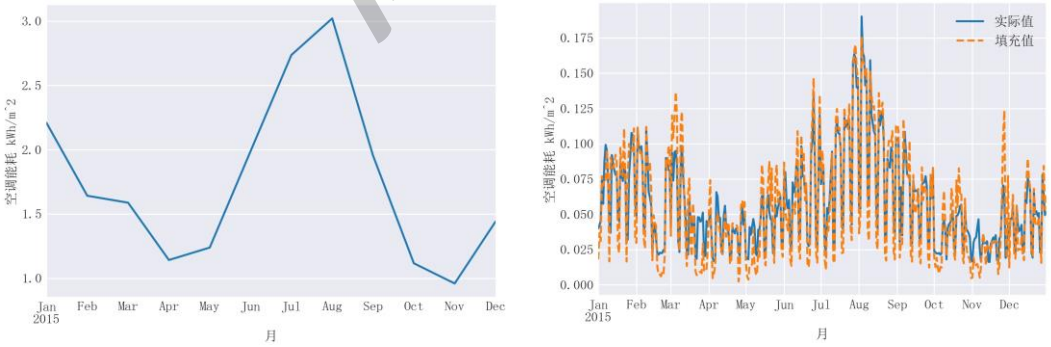
e) 两簇的聚类中心

图 4.6 1 月空调能耗典型逐日能耗聚类

同济大学

第三步，分析聚类结果，得到每簇的用能特性，每月每簇的用能特性描述见附录 A 中的表 A-1。将符合该簇用能特性的颗粒度较大的能耗数据用该簇的聚类中心曲线（即典型能耗曲线）进行填充。分析同一簇中的建筑用能特性可知，在夏季聚为两簇的月份，其中一簇中大部分的建筑房间功能以办公为主，商铺等功能的占比较小，另一簇中的建筑有较大的商业面积，为综合建筑；在冬季和过渡季聚为两簇的月份，其中一簇中的电耗非常低，这类建筑的热源为锅炉等不消耗电能供热的热源或者为热负荷很低的综合建筑，另一簇建筑为采用空调等消耗电能进行供热且有较大热负荷的办公或综合建筑。其余聚为三簇的月份则根据冷热源是否消耗电能，负荷大小进行分类。

图 4.7 展示了位于上海的某栋实际建筑的颗粒度细化结果，该建筑存在逐日的能耗数据，将逐日数据转换为逐月数据（每个月的逐日能耗之和）后的能耗曲线如图 4.7 a) 所示，用聚类得到的典型逐日能耗曲线对逐月数据进行填充后的结果与实际数据的对比如图 4.7 b) 所示。本研究中的填充精度采用 3.2.1 节中提到的 CV_RMSE 来量化，该示例建筑的填充精度 CV_RMSE 为 0.0549，具有较高的精度，可满足本研究的需求。



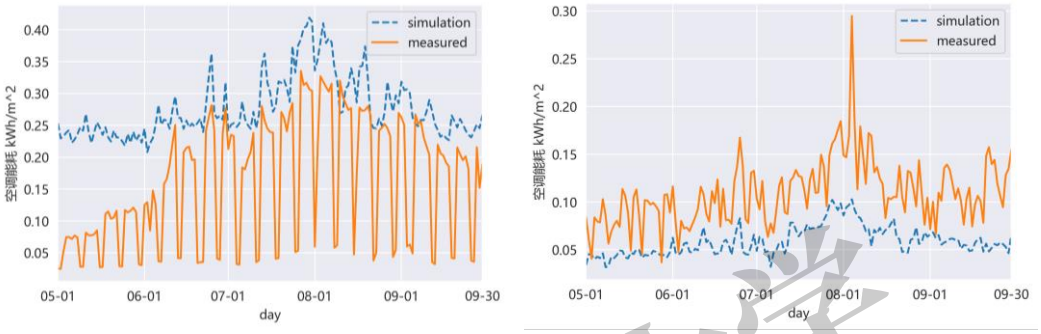
a) 实际逐月能耗曲线　　b) 填充前后的逐日能耗曲线

图 4.7 颗粒度转换结果：逐月数据填充为逐日数据

4.4 模拟数据处理

这部分模拟数据来自于 2.4 节中将关键变量作为输入，由快速模拟工具生成的模拟数据。模拟数据的灵活性较高，可生成不同颗粒度的分项及总能耗的数据，但由于白箱模型的参数大多为设计参数，且模型也是在理想情况下做了适当简化，故模拟数据的置信度不高，与实际能耗相比的差异较大。图 4.8 展示了两栋实际建筑制冷季逐日的模拟空调能耗和实测空调能耗的对比，由图可见，二者存在较

大的差异，且不同建筑的误差模式是不同的，模拟数据可能偏高也可能偏低。在混合模型中加入的模拟数据需在数据融合阶段进行修正，使其尽量接近实测数据以提高模型准确度。本节将介绍模拟数据的修正过程和结果，修正后的模拟数据将作为混合模型中该建筑对应的一个特征，借助修正后的模拟数据可大幅提高跨建筑能耗预测的精度。



a) 模拟数据大于实测数据　　b) 模拟数据小于实测数据

图 4.8 两栋实际建筑模拟数据和实测数据的差异

模拟数据修正的技术路线如图 4.9 所示，第一步，利用有实测数据的建筑建立输入（模拟数据+关键变量+天气及时序参数）-输出（实测数据）数据库；第二步，训练数据数据驱动模型来拟合实测数据和模拟数据存在的偏差；第三步，输入待修正建筑的模拟数据、关键变量和天气及时序参数，利用第二步中训练好的数据驱动模型输出修正后的模拟数据。

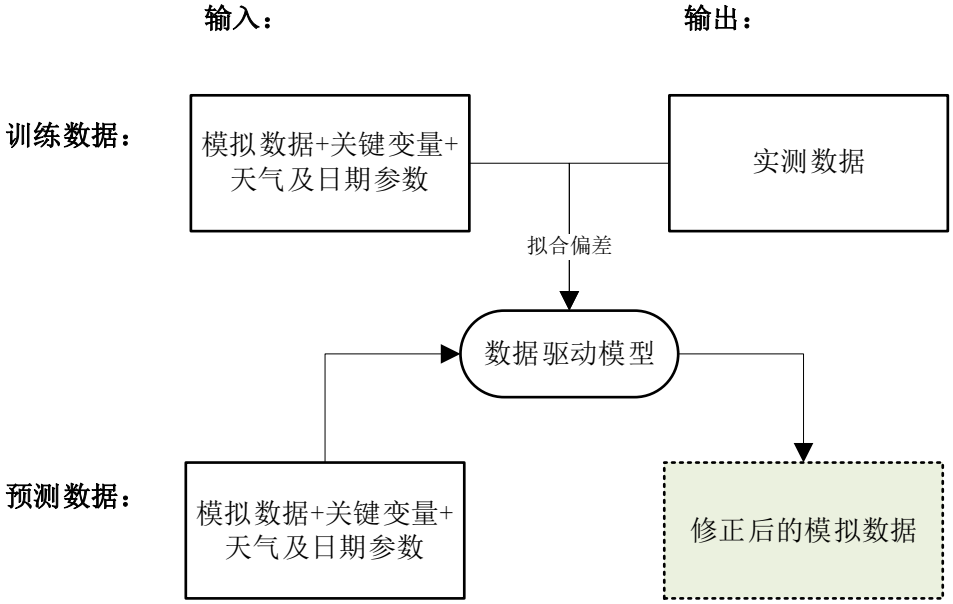
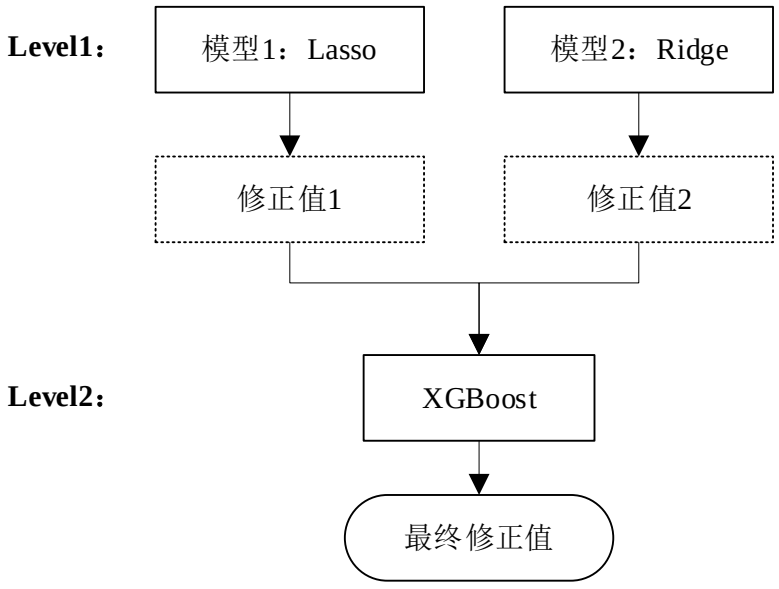


图 4.9 模拟数据修正原理

在上述第二步中，由于不同建筑模拟数据和实测数据的误差模式不同，单个

模型很难实现精确度较高的模拟数据修正。为加强用于修正模拟数据的数据驱动模型的泛化能力，提升模拟数据修正精度，本研究基于同一个数据集建立不同类型的模型进行模拟值的修正后，再利用集成学习（ensemble learning）的思想，集成不同模型的建模结果。集成算法可分为：装袋法（bagging）、提升法（boosting）、和堆叠法（stacking）三类，本研究结合了装袋法和堆叠法，将同一类算法建立的模型用装袋法集成，不同类型的算法建立的模型用堆叠法集成。

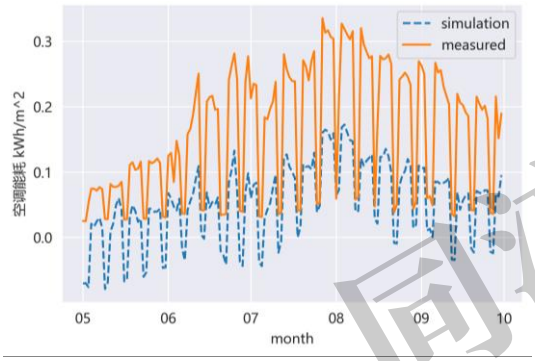
集成算法集成的各个模型被称为基模型。装袋法指的是利用各个算法独立地建立基模型，各个基模型是互不相关的，然后对每个模型的输出结果取（加权）算术或几何平均。堆叠法具有多层结构，每一层中有一个或多个模型，通常第一层的模型为采用不同算法或在不同数据集上独立建立的基模型，后一层将前一层模型的输出结果作为输入，用训练数据进行后一层模型的训练，从而减少预测误差。图 4.10 展示了本研究中模型集成的结构，本研究中采用的堆叠法具有两层结构，并在堆叠法中嵌套了装袋法，第一层采用 Lasso 回归和 Ridge 回归（岭回归）两种模型。Lasso 回归和 Ridge 回归均是在多元线性回归基础上衍生出来的回归模型。为解决多元线性回归的多重共线性问题，Lasso 和 Ridge 回归分别在多元线性回归的损失函数基础上加上了正则项——L1 和 L2 范式。由于回归模型对输入变量的绝对值大小非常敏感，故在第一层的回归模型训练之前需进行输入参数的处理，对于数值型变量，将其进行归一化处理，对于类别型变量将进行独热编码等编码处理。将这两个模型的输出结果（即基模型的模拟数据修正结果）进行加权算术平均，再将加权平均值作为第二层模型的输入，采用 2.3 节所述的 XGBoost 算法进行第二层模型的训练，第二层模型的输出即为最终的模拟数据修正值。



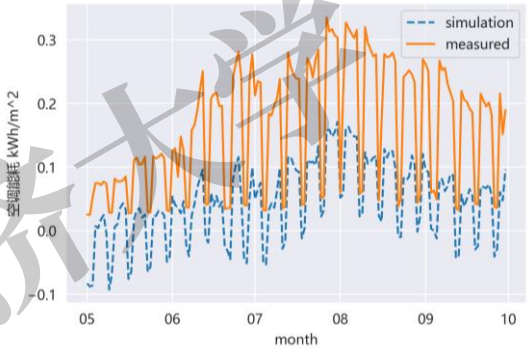
同济大学

图 4.10 模型集成结构

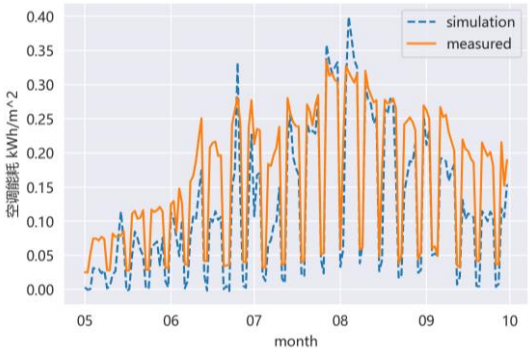
本文以图 4.8 a) 所示的实际建筑能耗数据来说明模拟数据修正的过程和修正结果。第一层模型中 Lasso 和 Ridge 回归的修正结果如图 4.11 a) 和 b) 所示，这两种模型使得模拟数据的曲线形状于真实数据的曲线更加接近，但模拟数据和实际数据的绝对值仍有差异。将第一层模型的修正结果取算术平均，得到第一层模型用装袋法集成后的模拟数据修正结果。将该结果作为第二层模型的输入，得到的模拟数据修正结果如图 4.11 c) 所示。修正前模拟数据和实际数据的 CV_RMSE 为 0.89，经过修正后模拟数据与实际数据的 CV_RMSE 为 0.364，可见，本节的模型大幅提高了模拟数据的准确度，使其更接近于实测数据。



a) 第一层模型 Lasso 回归的修正结果



b) 第一层模型 Ridge 回归修正结果



c) 第二层模型 XGBoost 的修正结果

图 4.11 模拟数据修正结果展示

4.5 本章小结

在大数据时代，建筑能耗数据可通过多个途径获取，但多个途径获取的能耗数据往往时间颗粒度和置信度存在差异，要将多源数据进行综合利用需首先进行数据融合。本章阐述了对三种不同来源的能耗数据的处理：

对于来自能耗监测平台的分项能耗数据，其特点是时间颗粒度较细但异常值

缺失值较多。本研究利用箱线图和相似建筑的能耗值对其进行了异常值的识别和修复以及缺失值的填补。

来自节能审计报告的能耗数据，其特点是置信度较高，几乎不存在缺失值和异常值，但这部分数据时间颗粒度较粗，通常是逐月数据。本研究利用 K-Means 算法提取了典型能耗曲线，对其进行了时间颗粒度的细化。

对于来自白箱模型的模拟数据，其特点是灵活性较高，可生成不同颗粒度的分项及总能耗的数据，但数据的置信度不高，与实际能耗相比的差异较大。本研究利用既有模拟数据又有实际数据的建筑建立了数据集，集成多种类型的数据驱动模型对模拟数据进行了修正，大幅减小了模拟数据与实测数据的差异。

同济大学

第 5 章 多源异构数据库和混合模型建立

5.1 概述

前文提到，白箱模型和黑箱模型各有优劣。白箱模型基于物理公式，可解释性高且可根据需要输出多种模拟结果；但第一，建模复杂，部分输入参数不可获取，第二，模型简化较多，也未考虑实际施工和运维过程中带来的模型偏差，预测准确度较低。黑箱模型预测准确度较高，但输入特征过多将导致“维度灾难”，且需要大量的历史数据进行训练，对于无历史数据的建筑难以建立数据驱动模型。并且，无论是黑箱模型还是白箱模型，在进行某栋楼或某个建筑群的能耗预测时都需要进行重新建模或训练，无法进行跨建筑的能耗预测，不具备通用性。

本研究针对上述两种模型的缺点，提出了混合模型，旨在解决输入变量多且部分不可获取和无历史数据的跨建筑能耗预测这两个问题。本章在第 2 章到第 4 章的前期工作基础上，构建了用于训练混合模型的多源异构数据库，再通过机器学习算法建立了可用于完全无历史能耗的跨建筑能耗预测的混合模型。

具体来说，为解决输入变量多且部分不可获取的问题，第 2 章从众多可能影响能耗的变量中提取了对能耗有显著影响的关键变量。为解决关键变量存在数值缺失或不可获取的问题，第 3 章采用了遗传算法和 Apriori 关联规则挖掘算法分别进行了有历史能耗和无历史能耗场景下，关键变量缺失值的推测。通过第 2 章和第 3 章的工作，本研究构建了完整的与建筑能耗强相关的建筑物理信息，该信息将作为混合能耗模型的特征变量。为解决多源异构数据因数据结构、时间颗粒度等不同而无法进行综合利用的问题，第 4 章针对来自能耗检测平台、节能审计报告的实测数据和快速模拟工具生成的模拟数据分别进行了相应的预处理，进行了数据融合，使得各种数据有相同的结构和时间颗粒度。通过第 4 章的工作，本研究构建了建筑能耗数据信息，再结合第 2、3 章构建的建筑物理信息、根据专家知识判断的可能影响能耗的外部信息（建筑建造年份等）、建筑所在地的天气信息（干球温度等）、与建筑功能和人行为相关的时间序列信息（是否工作日等），构建了多源异构数据库。值得一提的是，在第 2 章和第 3 章得到了关键变量信息后，每一栋楼均可利用快速能耗模拟工具得到模拟数据，该模拟数据可能与实测数据相去甚远，但经过 4.4 节的修正后，模拟数据与实测数据相似度将大大提高。由于该模拟数据是根据建筑物理相关的关键变量生成的，并利用相似建筑的模拟数据与实测数据建立的数据驱动模型进行了修正，该模拟数据结合了建筑物理信

同济大学

息和相似建筑的历史能耗数据，包含了丰富的信息，该修正后的模拟数据可作为混合模型的输入特征。

最后将上述多源异构数据库作为训练数据，采用组合树算法建立了能耗预测混合模型。

需明确的是，本文中的混合模型是采用机器学习算法进行训练的模型，但模型训练所用到的数据为数据融合后的多源异构数据，即既有来自白箱模型的模拟数据，也有从不同途径获取的实测数据。混合模型建立过程中最重要的是数据集的建立（即模型的输入输出值，建立过程包括了数据预处理、特征工程等步骤）和模型的训练，故下文将对这两个方面进行详细阐述，介绍多源异构数据库的建立和混合模型的训练。

5.2 多源异构数据库的建立

多源异构数据库是混合模型训练的基础，数据驱动模型本质上是利用各种机器学习算法建立数据库中输入和输出之间的关系，数据的质量和特征工程对模型的准确性有着至关重要的作用。下文将从混合模型的输入变量（即特征）和输出变量（即预测目标：如空调能耗）两个方面展开。

5.2.1 输入特征

特征对黑箱模型的性能起着决定性的作用，黑箱模型就是建立特征与预测目标之间的关系的过程，理想的特征应体现对能耗的所有影响，特征工程往往是黑箱模型建立过程中较为关键的一步。如 1.2.1 节所述，特征工程包括：特征提取、特征创造和特征选择。本研究的输入特征确定思路为：首先进行特征提取和特征创造，搜集全面的可能与预测目标有关的特征；然后进行特征选择，过多的相关性低的特征将导致模型过拟合，故需对特征进行选择，保留下有用的特征。由于嵌入法中采用不同的算法选择的特征差异较大，本研究采用 Kaggle 数据建模比赛中较为常用且非常有效的特征选择方法——LOFO（Leave One Feature Out）进行特征选择，其执行步骤为：首先用所有的变量进行建模，然后每次去除掉一个变量，若去除变量后交叉验证的准确度提升则确定去掉该变量，反之保留该变量。图 5.1 汇总了本研究中提取和创造的特征，包括建筑特征、气象特征、时序特征和模拟特征这四类，表中最后一列“是否采用”表示了在特征选择步骤中是否保留了该变量，“是”表示在最终建立混合模型时采用了该变量。

下文将对建筑特征、气象特征、时序特征和模拟特征分别进行说明。

建筑特征类的变量包括第 2 章提取的关键变量、建筑编号、建造年份和商业部分面积占比。关键变量代表了对能耗有显著影响的建筑物理信息，包括建筑负荷相关变量和建筑系统相关变量，具体见表 2.4 中负荷相关和机电系统相关部分的变量。此外还包括表 2.4 中的建筑编号和商业面积占比。建筑编号代表了对应的能耗数据属于哪栋建筑，为表示类别的变量，其绝对值没有实际意义。商业部分面积占比是商业部分面积和建筑总面积的比值，由于商业部分和办公部分的人员作息、照明功率密度、设备功率密度等存在显著差异，故能耗水平也存在较大差异，本研究用商业部分面积占比作为输入特征以衡量建筑功能分区的差异。关键变量缺失值根据是否存在历史能耗选择第 3 章中对应的算法进行推测并填补。除关键变量外，建筑特征还包括建造年份，建造年份接近的建筑通常有相似的墙体传热系数、设备老化程度等，从而有相似的能耗水平，故将该变量纳入输入特征中。

气象特征包括了直接从气象站获取的变量（干球温度、相对湿度、风速和降雨量）和计算得到的变量。值得说明的是计算得到的变量。空调能耗，尤其是新风能耗与焓值直接相关，故本研究通过干球温度和相对湿度计算得到焓值，将其作为输入特征。

时序特征主要是从时间序列中提取的特征。建筑能耗受人员行为的影响很大，办公建筑的人员作息和人员行为会影响空调是否开启、灯光和设备密度等，而办公建筑的人员作息和人员行为又很大程度上受时间序列的影响，如工作日人员密度普遍大于周末和节假日，故本研究用从时间序列中提取的时序特征，如是否节假日等，作为输入特征。

模拟特征为本文研究的混合模型中较为特殊的特征，它是由建筑的关键变量作为输入，通过第 2 章中的快速模拟工具生成的，再由 4.4 节中模拟数据修正模型修正后的模拟数据。由于该模拟数据是根据建筑物理相关的关键变量生成的，并利用相似建筑的模拟数据与实测数据建立的数据驱动模型进行了修正，故其既包含了建筑本身的物理信息，还包含了相似建筑的能耗信息，将其作为特征可很好地提升跨建筑预测时的预测精度。

同济大学

表 5.1 特征汇总

类别	特征名称	特征说明	是否采用
建筑特征	第2章中提取的关键变量，见表 2.4 中负荷相关、机电系统相关和其他特征中的变量	其中制冷设定温度/制热设定温度和冷冻水/热水供水温度这两类变量根据预测目标选取相应的变量（如预测制冷季的能耗时选用制冷设备温度和冷冻水供水温度），缺失值利用第3章的方法进行填补。	是
	建造年份	可能可以大致反映建筑的能耗水平	是
气象特征	干球温度	从气象站获取	是
	相对湿度	从气象站获取	是
	风速	从气象站获取	否
	降雨量	从气象站获取	否
	焓值	根据干球温度和相对湿度创造的特征	是
时序特征	年份	该数据点的时间所在的年份	否
	每年的第几月	该数据点的时间位于每年的第几月	是
	每月的第几日	该数据点的时间位于每月的第几日	逐日及更细的颗粒度时预测时采用
	每年的第几周	该数据点的时间位于每年的第几周	否
	每日的第几小时	该数据点的时间位于每日的第几小时	逐时预测时采用
	是否节假日	会影响人行为，进而影响能耗	是
	该小时是否为工作时间	会影响人行为和空调是否正常开启，进而影响能耗	逐时预测时采用
模拟特征	模拟数据修正值	第 2 章中的快速模拟工具生成的，再由 4.4 节中模拟数据修正模型修正后的模拟数据	是

5.2.2 输出：能耗数据

能耗为本研究的预测目标，也是混合模型的输出变量。除输入特征外，能耗数据本身的数量和质量也对预测模型的表现有很大的影响。首先需要有足够多数量的数据，机器学习算法才能学习到输入与输出之间的关系。其次，作为机器学习模型的学习目标，能耗数据需要足够准确（即质量足够高）才能保证模型学到正确的输入与输出之间的关系。由于模拟数据和实测数据始终存在一定的偏差，

若将模拟数据作为学习目标可能影响混合模型的准确度，而若是将其作为输入特征，模型在学习时将会去判定该特征的重要程度，故本研究仅将模拟数据用作混合模型的一个输入特征。多源异构数据库的能耗数据来自于能耗检测平台和节能审计报告，并根据第 4 章所述的方法分别对这两类数据进行前处理，对于来自能耗监测平台的能耗数据进行异常值处理和缺失值填补，对于来自节能审计报告的能耗数据进行时间颗粒度的转换。

5.3 混合模型的建立

在建立了输入-输出数据（即多源异构数据库）后即可利用机器学习算法进行模型的建立。有多种算法可以进行模型的建立，1.2.1 节中对常用的算法进行了综述，各种算法均有其优缺点，并不一定所有算法都适用于此种能耗预测场景。在算法选择时，本研究参考了 ASHRAE 举办的 Great Energy Predictor III competition 能耗预测比赛获奖者所用的算法，其中前五名均采用了组合树模型[30]，在课题组举办的“能耗侦探”比赛中，准确度最高的队伍也采用了组合树模型中的 XGBoost 算法[51]，故本节所述的混合模型选用组合树模型进行建模。组合数模型将多个决策树模型进行组合，在用此种算法进行建模时，算法本身会进行参数的优化，但建模者需要进行超参数的设置，超参数决定了模型的复杂程度等。组合树模型涉及到的超参数较多，在模型训练过程中需要花费大量时间进行超参数的调节，为防止模型过拟合，往往采用交叉验证的方法挑选出精度高、泛化能力也较强（偏差和方差均较小）的一组超参数。故本节的混合模型建立的流程为：首先采用交叉验证方法将训练数据划分为多组训练集和验证集；再在同一组超参数下用每组训练集的数据进行建模，并预测测试集的能耗值，进行误差的计算，混合模型中采用的误差衡量指标为式（3.1）所表示的 CV_RMSE；然后进行超参数的调节，选取交叉验证中的多组验证集的平均 CV_RMSE 最小的一组超参数作为模型的最佳超参数；最后在最佳超参数下用所有的训练数据进行混合模型的训练。下文将详细介绍选取的组合树模型以及数据集的划分和模型调参。

5.3.1 算法的介绍

本研究选用的算法属于组合树模型，为基于决策树的集成算法。决策树算法根据特征进行分支，以减少叶子节点上的信息熵，组合树模型采用装袋法或提升法将多个决策树模型进行集成。除了 2.3.4 中介绍的 XGBoost 之外，本研究采用了一种速度更快，占用资源更少的 lightGBM 算法。LightGBM 是微软提出的一

同济大学

种快速的、分布式的梯度提升树算法，其在 XGBoost 的基础上进行了优化，在不降低其准确率的前提下，显著提升了算法的速度和训练需占用的内存。LightGBM 在传统的梯度提升树（Gradient Boosting Decision tree, GBDT）算法的基础上进行了优化，通过改进采样方法、进行互斥特征的捆绑、对叶子生长的深度进行限制等方法提升了训练效率，并且 lightGBM 支持直接使用类别型变量，并可以高效并行的计算。由于以上的优点，lightGBM 受到了广泛的应用，在能耗预测比赛中也有大量的队伍采用了 lightGBM 算法。

5.3.2 数据的划分和模型调参

数据驱动模型的基石是数据，数据除了用于模型的训练外，还需用于模型的评估，为测试模型在完全未知的数据集上的表现，训练数据和测试数据是不能重合的，故本研究将搜集到的所有数据划分为训练数据和测试数据。由于模型训练过程中还需进行复杂的超参数调节，故还需将训练数据划分为训练集和验证集。多源异构数据库中所有的有效数据划分如图 5.1 所示。下文将详细解释训练数据划分的原因和方式。

同济大学

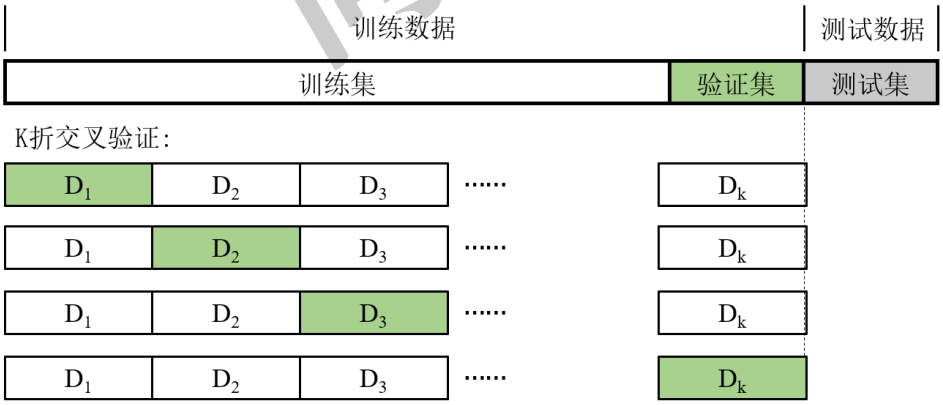


图 5.1 本研究中有效数据的划分方式示意图

在机器学习领域，常常采用泛化误差来衡量模型在未知数据集上的准确率。一个理想的机器学习模型需有更可能小的泛化误差，在训练数据和算法确定的情况下，可采用调节超参数的方式降低误差，使得模型在未知数据集上有很高的准确性，这也是本节对混合模型进行调参的目的。泛化误差由方差、偏差和噪声组成，方差可理解为在模型在不同数据集上表现出来的稳定性，偏差可理解为模型在单个数据集上预测的准确度，方差和偏差往往是此消彼长的，在机器学习领域被称为“方差-偏差困境”。过拟合指的是模拟对训练数据进行了过度学习，并没有描述数据的整体特征，导致在训练数据上表现很好而在新的测试数据上表现很

差。树模型是一种非常容易过拟合的模型，一般来说，模型越复杂，过拟合可能性会越高，而模拟越简单，则可能导致欠拟合（模型对数据之间的关系学习能力较弱），调参的目标便是在过拟合和欠拟合之间寻找一个合适的模型复杂度。

通过以上背景知识我们可以知道，只在单个训练集上训练模型是没有办法得到泛化误差小的模型的，故本研究采用 k 折交叉验证的方法进行训练集和验证集的划分，在同一组超参数下进行多个训练集上的模型训练，选取交叉验证中的多组验证集的平均 CV_RMSE 最小的一组超参数作为模型的最佳超参数，最后在最佳超参数下用所有的训练数据进行混合模型的训练。

K 折交叉验证的示意图如图 5.2 所示，首先将训练数据 D 分成 k 份，然后不重复的取其中一份作为测试集，其他的数据作为训练集用于模型的训练，再计算每个模型在其测试集上的 CV_RMSE，最后将 k 个 CV_RMSE 取平均作为最终作用判断模型效果的 CV_RMSE。本研究中取的 k 为数据库中建筑的个数。

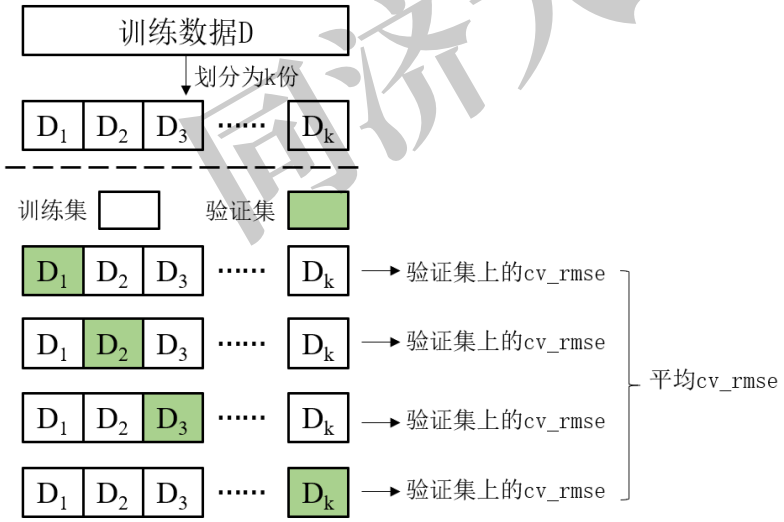


图 5.2 K 折交叉验证示意图

组合树模型需要进行调节的参数主要有以下几类：1）控制树的生长的参数（如叶子节点个数，树的深度等）；2）防止过拟合的参数（如正则项等）；3）集成算法涉及到的参数（如树的棵树等）；4）其他相关参数（如随机树的种子等）。模型的表现与参数的选择密切相关，调参方法有多种，本研究选取了最为常用的是网格搜索法进行调参，这种方法是在超参数的取值空间中进行超参数的组合，建立超参数网格，再根据每一个组合进行模型的训练，选取模型表现最好的超参数组合作为最终的超参数。最后在最佳的超参数下，用所有的训练数据进行模拟训练得到最终的混合模型。

同济大学

5.4 本章小结

基于多源异构数据的混合模型结合了白箱模型和黑箱模型的特点，在输入特征中既考虑了建筑物理相关的参数，又考虑了建筑运行使用的参数，并且模型建立过程中利用的数据既包含白箱的模拟数据，又包含了实测数据。本文研究的混合模型主要解决了两个问题：1）白箱模型输入变量过多且部分变量缺失的问题，2）目前预测模型不够通用，无法进行跨建筑预测的问题。本章是一个“集大成”的部分，综合应用了第 2 章、第 3 章和第 4 章的研究成果建立了多源异构数据库，然后进行了混合模型的建立。

对于多源异构数据库的建立，多源异构数据库包括输入特征和输出的预测目标两个部分。对于输入特征部分，通过第 2 章提取的关键变量和第 3 章对关键变量缺失值的推测得到了部分建筑物理相关的混合模型输入特征，结合其余可从气象站或时间序列中提取的特征和第 4.4 节中修正的模拟数据，得到了完整的多源异构数据中的输入特征部分，再通过特征选择确定了混合模型最终的输入特征。对于输出的预测目标部分，能耗数据来自于能耗检测平台和节能审计报告，通过 4.2 和 4.3 节中介绍的分别对这两个数据进行预处理的方法可得到高质量、相同时间颗粒度的能耗数据。

对于混合模型的建立，本研究采用在能耗预测竞赛中预测准确度较高的队伍所用到的组合树模型进行建模，采用 5 折交叉验证的方式进行超参数的优化，以得到泛化误差较小的混合模型。

同济大学

第 6 章 混合模型有效性验证

6.1 概述

本章对本文研究的基于多源异构数据的办公建筑能耗预测方法进行了验证。为防止信息泄露并测试混合模型在未知数据集上的泛化能力，本研究将搜集到的所有数据划分为训练数据和测试数据，训练数据是混合模型建立过程中将用到的数据，训练数据包括训练集和验证集，测试数据为混合模型训练好之后用于评估模型性能的数据，这部分数据不能用于模型训练，否则会导致评估结果偏高。本研究所有搜集到的有效数据的划分方式如图 5.1 所示，先将所有有效数据划分为训练数据和测试数据，再将训练数据按照图 5.2 所述的交叉验证方法划分为训练集和验证集。

本研究提出的办公建筑能耗预测方法可应用于多种场景，可根据是否有历史能耗数据，关键变量提取时的研究对象（如空调制冷能耗，空调供热能耗，建筑全楼宇能耗等）建立对应场景的混合模型。

由于数据的限制，本研究对办公建筑制冷能耗预测的效果进行了验证和展示，包括关键变量为确定性输入的确定性预测和关键变量为一个区间的不确定性预测。但值得说明的是，本研究所提出的方法可用于楼宇总能耗、制热能耗等作为预测目标的预测任务

6.2 数据说明

本研究共搜集了 105 栋来自能耗监测平台的办公建筑数据，经过异常楼宇的剔除后保留了 30 栋建筑的数据。从能耗监测平台的办公建筑信息示例如图 6.1 和图 6.2 所示，其中图 6.2 b) 中的白色部分为缺失的能耗数据。从能耗监测平台获取的可用数据包括建筑基本参数（建筑类型、地上层数、地下层数、总面积、建筑类型、空调类型等）和分项能耗数据（提供逐日、逐时、逐月颗粒度下的分项计量数据），由于逐时的各设备的能耗数据较差，存在大量异常值和死值，本文中验证了可获取的最细的时间颗粒度（逐日能耗）下的预测结果。

除能耗监测平台的数据外，本研究还搜集了 18 栋来自节能审计报告的办公建筑能耗数据，部分建筑的节能审计报告如图 6.3 所示。

同济大学

节能审计报告的建筑信息较为丰富，为说明可获得的建筑信息，本文选取了一栋建筑进行详细的信息说明。图 6.4 展示了建筑的外观图，建筑位置于上海市黄浦区，总建筑面积为 29531.93 平方米，建筑高度为 93.7 米。建筑地上层数为 22 层，标准层面积为 1333 平方米，裙楼部分有商业用房，裙楼部分为办公写字楼。地下层数为 2 层，设置有地下车库和设备机房。建筑于 2007 年竣工并投入使用。建筑围护结构部分，外墙墙体材料为粘土块，局部为玻璃幕墙（双层中空镀膜 low-e 玻璃），外墙与屋面均无保温措施，建筑未采用外遮阳，有窗帘、玻璃贴膜等内遮阳措施。建筑机电系统部分，该建筑空调系统为中央空调风机盘管+新风系统，夏季冷源和冬季热源均采用 7 台风冷螺杆热泵机组，3 台负责低区，4 台负责高区，其中 2 台型的额定制冷量为 700kW，额定功率为 200kW，3 台的额定制冷量为 568.6kW，额定功率为 160kW，剩余 2 台的额定制冷量为 445.7kW，额定功率为 140kW。无特殊情况下采用 4 用 3 备，此外还有一台水源热泵机组。在屋顶设有 1 台冷却塔，冷却塔的额定功率为 1.5kW。地下车库、监控室、物业办公室等区域则采用分体空调及 VRV 机组。空调输送系统配有 11 台冷冻水泵，8 台的额定功率为 15kW，3 台的额定功率为 18.5KW；冷却水泵共 3 台，其额定功率为 7.5kW。该建筑空调系统的控制方式为人工控制，开启时间为 9：00~18：00（周一至周五）9：00~13：00（周六）（商铺一般全年供应冷热源）。

对于每组数据中的能耗数据，根据其特点分别用第 4 章所述的方法进行数据处理。对于每组数据中的建筑特征数据，本研究所需的特征为表 5.1 所罗列的用于混合模型的特征，若存在特征缺失，则根据第 3 章所述的方法进行变量缺失值推测。经过上述步骤后，本研究完成了多源异构数据库的构建。

上述办公建筑可分为三种类型：裙楼存在商业用途的综合类办公建筑、租赁型办公建筑和政府办公建筑，由于综合类办公建筑的用能特点与其余两类存在较大差异，需要进行分别建模。本研究的主要研究对象为办公建筑，故主要研究了以办公为主的租赁型办公建筑和政府办公建筑的能耗预测，上述搜集的数据中有 19 栋来自于租赁型办公建筑和政府办公建筑。

同济大学

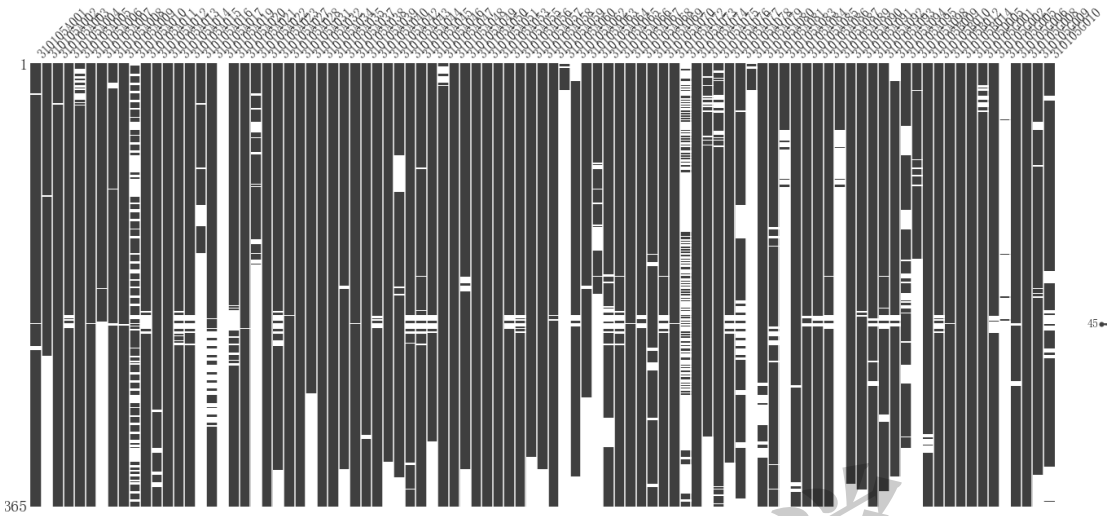
建筑编号	建筑名称	建筑地址	地上层数	地下层数	总面积	建筑类型	空调类型
310105A008	6号地块		6	1	77563.5400	办公建筑	集中式全空气系统
310105A009	DFSJDS		24	1	26883.8100	办公建筑	风机盘管+新风系统
310105A022	JHJMDS		20	2	37632.0000	办公建筑	风机盘管+新风系统
310105A010	GBCF-I		15	3	38881.3200	办公建筑	风机盘管+新风系统
310105A004	JQDS		19	2	24911.0000	办公建筑	风机盘管+新风系统
310105A034	KLDS		18	1	16111.0000	办公建筑	风机盘管+新风系统
310105A035	CNQZF		20	2	49118.3000	其它建筑	风机盘管+新风系统
310105A038	DZJR		35	2	49933.0000	其它建筑	风机盘管+新风系统
310105A042	ZYKJY		20	1	21908.0000	办公建筑	分体式空调或VRV的
310105A028	DFWJDS		25	2	47630.0000	办公建筑	分体式空调或VRV的

图 6.1 能耗监测平台中获取的办公建筑基本信息（图中模糊了建筑地址等敏感信息）

分项名称	时间	能耗值(kWh)
空调用能	2021-07-01	5868.85
空调用能	2021-07-02	5011.48
空调用能	2021-07-03	550.17
空调用能	2021-07-04	131.60
空调用能	2021-07-05	8482.19
空调用能	2021-07-06	8197.97
空调用能	2021-07-07	9869.94
空调用能	2021-07-08	8086.24

a) 能耗监测平台中获取的办公建筑能耗信息

同济大学



b) 能耗数据缺失情况（白色为缺失值）

图 6.2 办公建筑能耗信息

同济大学

图 6.3 节能审计报告节选（模糊敏感信息）

图 6.4 测试建筑外观

同济大学

图 6.5 测试建筑设备实际图

6.3 测试建筑的能耗预测结果

在存在大量历史数据下，可用机器学习算法建模，但此种场景下建立的模型往往是不具备通用性，在预测另一栋建筑的能耗时需要重新建模，且不适用于无历史能耗的场景。本文研究的混合模型与其他的模型相比最大的特点在于：可实现无历史能耗场景下的跨建筑能耗预测。

在无预测要求颗粒度下的历史数据时，其余已知信息可能存在三种情况：

1) 关键变量值已知，无任何历史能耗；
2) 关键变量存在缺失值，有与预测颗粒度一致但数据点极少的历史数据或比预测颗粒度粗（例如来自节能审计报告的逐月能耗数据）的历史数据；
3) 关键变量存在缺失值，无任何历史能耗。

在第 2）种情况下，可采用 3.2 节所述的方法补全关键变量，得到确定的关键变量值。在第 3）种情况下，可采用 3.3 节所述的方法补全关键变量，得到的为一个关键变量值的推测区间（或在区间内的采样点）。在第 1）种或第 2）种情况下，因混合模型输入的关键变量为一个确定的数值，本文称这种情景下的能耗

预测为确定性预测。在第 3）种情况下，因关键变量推测值为一个区间，本文称此时的能耗预测为不确定性预测。

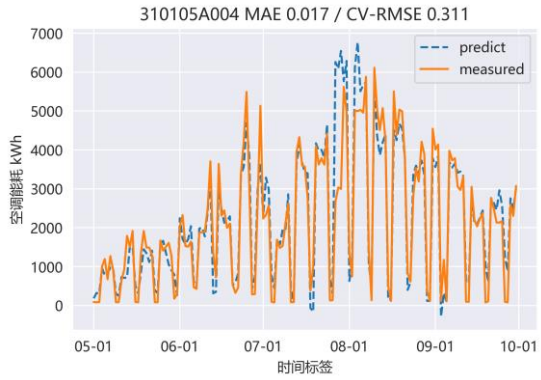
要利用混合模型进行能耗预测，首先要进行输入的确定，其中可能存在缺失值的为关键变量值。在第 1）种场景下，所有关键变量均已知，无需进行推测。在第 2）种场景下，存在关键变量缺失值的建筑能耗预测步骤为：第一，首先采用 3.2 节所示的遗传算法补全关键变量。第二，以关键变量作为输入，利用第 2 章介绍的快速建模工具得到对应的模拟值，并用 4.4 节建立的模拟数据修正模型对该模拟数据进行修正。第三，根据气象参数和时间序列信息得到混合模型所需的其余特征。最后进行能耗预测。在第 3）种场景下，与第 2）种情况的类似，主要区别在第一步，此时采用 3.3 节所示的基于 Apriori 关联规则挖掘算法的关键变量缺失值推测算法补全关键变量。由于基于关联规则得到的关键变量推测值为一个区间，并在区间中进行抽样作，进行不确定性的能耗预测。

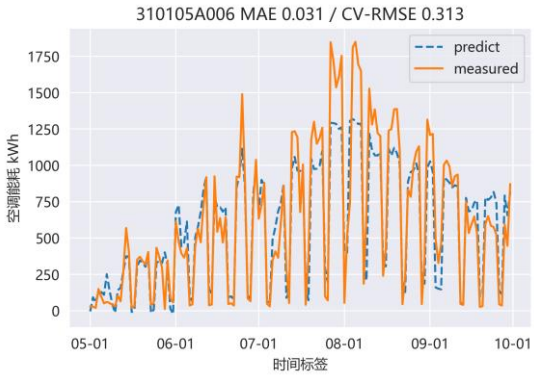
下文将对确定性和不确定性两种预测方式进行结果验证。

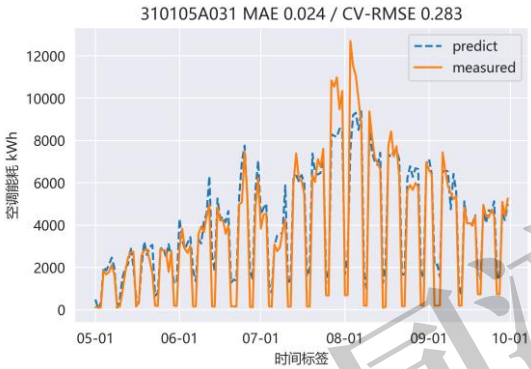
6.3.1 确定性预测

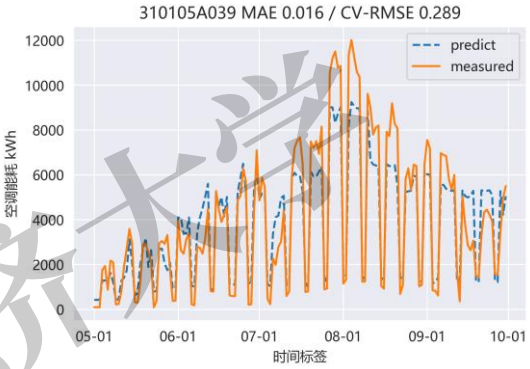
为保证模型和超参数的普适性，应在固定模型超参数的情况下进行验证。本节的验证方法为：首先，采用交叉验证的方式确定超参数，选择平均 CV_RMSE 最小的超参数组合；然后，固定这一组超参数，将多组建筑的能耗数据作为测试数据，其余建筑的能耗数据作为训练数据，建立模型，并用该模型对测试建筑进行能耗预测。图 6.6 展示了其中 10 栋建筑的测试结果，每一子图的顶部展示了预测的相对误差（MAE）和 CV_RMSE。可见，总体的预测结果较好，但有的建筑预测误差较大，本研究对误差较大的建筑信息和其他建筑信息进行了比对，预测误差较大的建筑存在着与大部分训练建筑不同的特点。如编号为 310105A086 的预测结果，由于其 7 月-9 月节假日仍存在较高的能耗，而其他办公建筑节假日能耗较低，故导致该楼 7 月-9 月节假日的预测效果不佳。再比如编号为 310105A053 的预测结果，由于该楼整体的制冷能耗峰值较小，能耗随天气变化也较小，而其他办公建筑能耗随天气变化比较明显，故该楼气温变化大的日期预测效果不佳。这表明，在进行跨建筑能耗预测时，还需更加详细的对办公建筑进行分类，采用与目标建筑相似性高的建筑数据作为训练建筑才能更准确地进行跨建筑的能耗预测。本研究再将用能模式不太一致的建筑去除，重新进行超参数的调节和模型训练，图 6.7 展示了测试建筑的误差分布，得到的预测误差 CV_RMSE 大部分在 0.3 左右，最大为 0.5128。

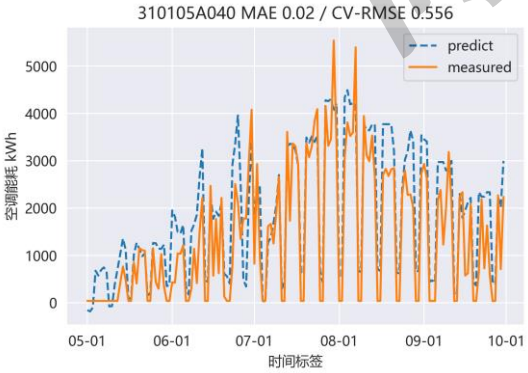
同济大学

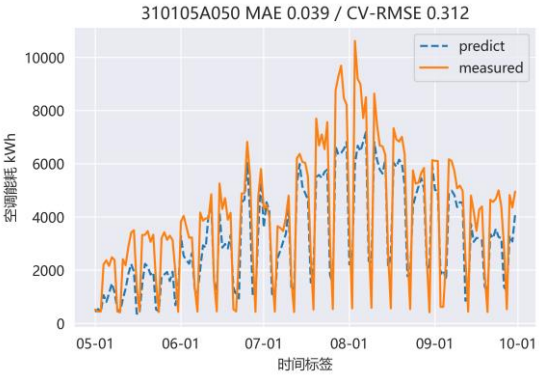

310105A004 MAE 0.017 / CV-RMSE 0.311
predict
measured
空调能耗 kWh
时间标签

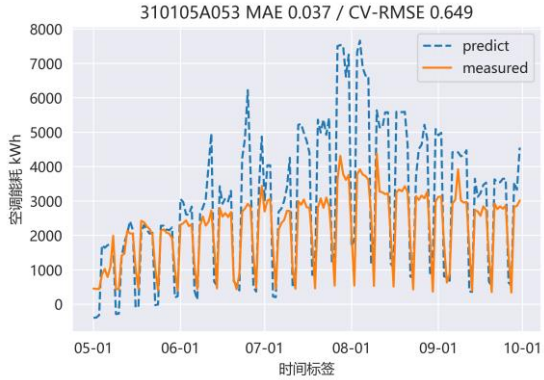

310105A006 MAE 0.031 / CV-RMSE 0.313
predict
measured
空调能耗 kWh
时间标签

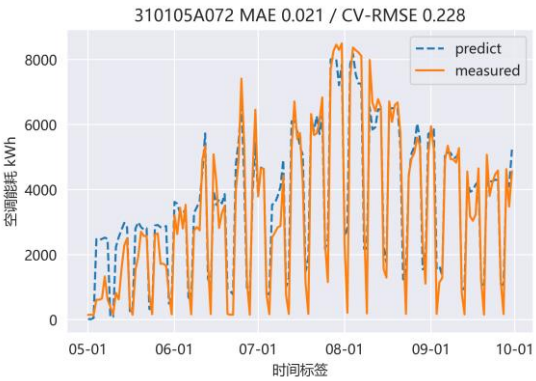

310105A031 MAE 0.024 / CV-RMSE 0.283
predict
measured
空调能耗 kWh
时间标签


310105A039 MAE 0.016 / CV-RMSE 0.289
predict
measured
空调能耗 kWh
时间标签


310105A040 MAE 0.02 / CV-RMSE 0.556
predict
measured
空调能耗 kWh
时间标签


310105A050 MAE 0.039 / CV-RMSE 0.312
predict
measured
空调能耗 kWh
时间标签


310105A053 MAE 0.037 / CV-RMSE 0.649
predict
measured
空调能耗 kWh
时间标签


310105A072 MAE 0.021 / CV-RMSE 0.228
predict
measured
空调能耗 kWh
时间标签

同济大学

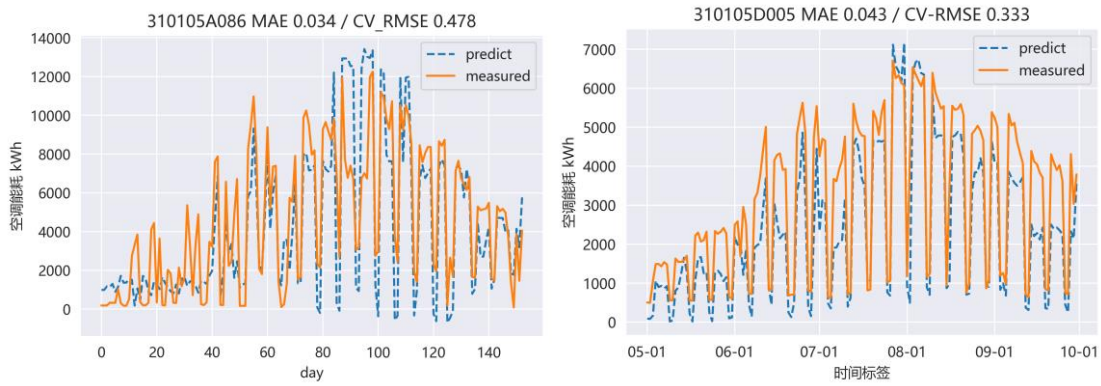


图 6.6 确定性预测部分结果展示

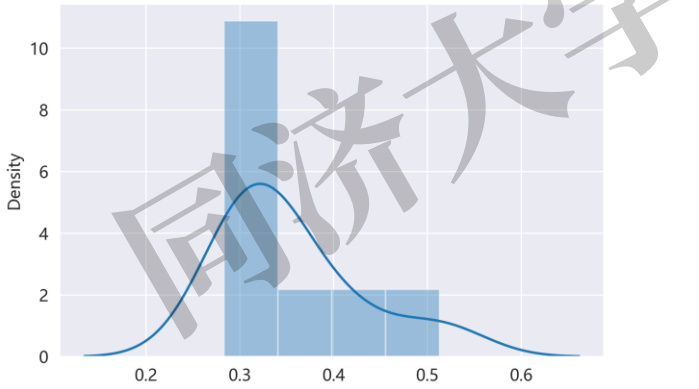


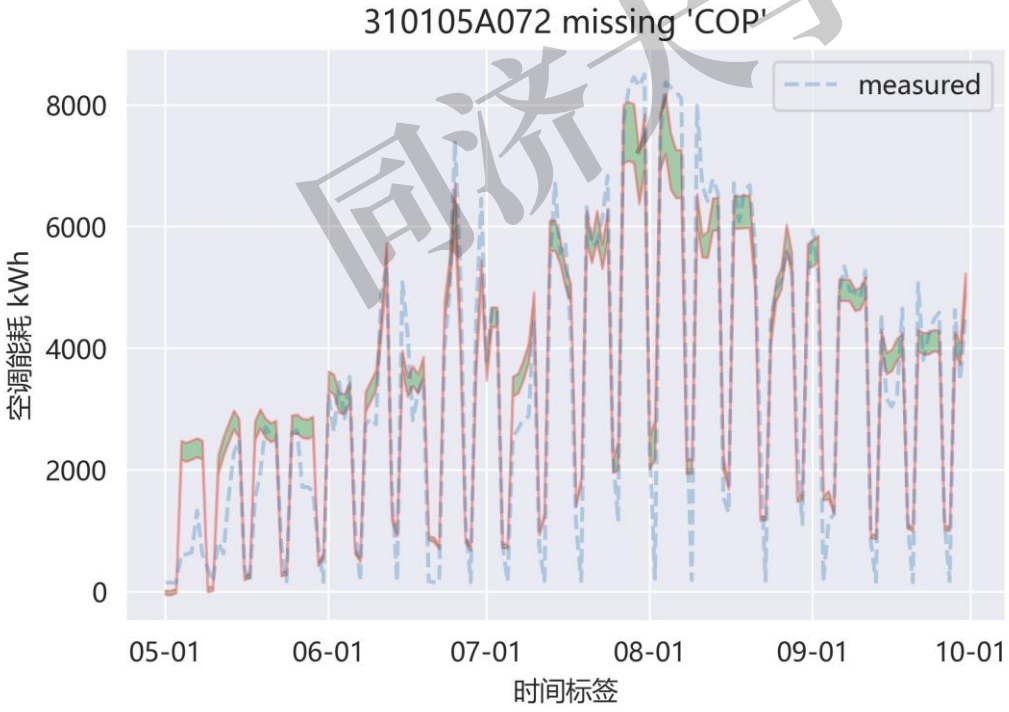
图 6.7 预测误差分布

6.3.2 不确定性预测

本节对上文所述的第 3）种情况下，推测的关键变量为一个区间时的能耗预测结果进行了验证。为分析关键变量的不同缺失情况下的制冷能耗预测的误差和不确定性，本节选取图 6.3 中编号为 310105A072 的建筑作为研究对象，对单个和多个不同变量缺失时的预测结果进行了分析。对于单个变量缺失，图 6.8 a)和 b) 分别展示了 COP 或新风渗透率（INFIL）缺失时的预测结果，其中阴影部分表示模型给出的能耗预测区间，虚线表示实测结果。可见在仅 COP 缺失时，存在少量的不确定性，而在 INFIL 缺失时，不确定性非常小，其可能原因为 INFIL 的推测误差很小，且给出的推测范围也较小，故对能耗预测结果的不确定性影响也较小。图 6.9 展示了 COP 和 INFIL 同时缺失时的不确定预测结果，其不确定性与只有 COP 缺失时类似。在此基础上，本节接着探索了 COP、INFIL 和太阳得热系数（SHGC）同时缺失时的预测结果，如图 6.10 所示，预测的不确定性与 COP 和 INFIL 同时缺失的不确定性相比略有增大，但增大幅度较小。于是再此

同济大学

基础上再加大了缺失变量的数量，图 6.11 显示了在 INFIL、照明功率密度（LPD）、SHGC、内遮阳开启程度（ST）、窗墙比（WWR）、送风温差（SATD）和 COP 同时缺失的情况下的预测结果，结果表明，即使缺失变量数量缺失增多，预测的不确定性也突增，但预测误差整体增大（如图 6.11 中 5 月第一周的预测能耗较 6.9 和 6.10 中误差增大很多），这表明随着缺失变量的增多，虽然预测的整体准确度明显下降。对于不确定性并未明显增大的原因可能是：某些缺失变量的推测误差较小或某些相较于时序特征和其他关键变量，某些变量的重要性程度较低。为验证上述猜想，在 COP 和 INFIL 缺失的基础上，增加了对制冷能耗影响较大且推测误差较大的楼层（NL）作为缺失变量，其预测结果如图 6.12 所示，可以看出，此时推测的误差和不确定性显著增大。结果表明，除缺失变量的数量外，缺失变量的推测准确度和对预测目标的重要程度会显著地影响预测结果和不确定性。



a) COP 缺失时预测结果

同济大学

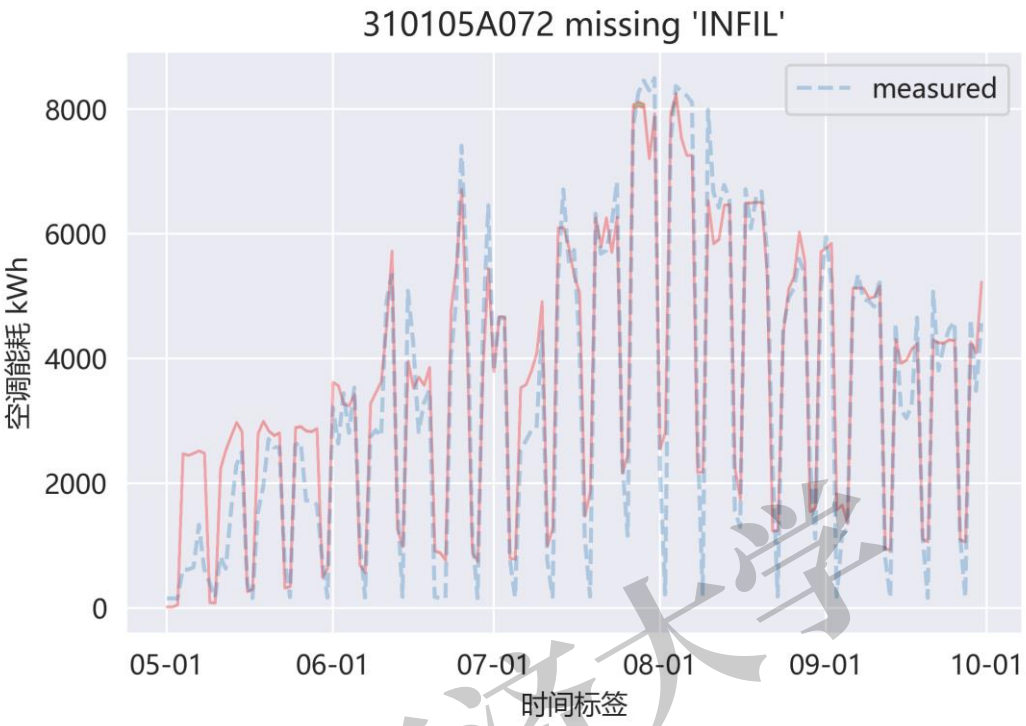


图 6.8 缺失单个变量时的不确定预测

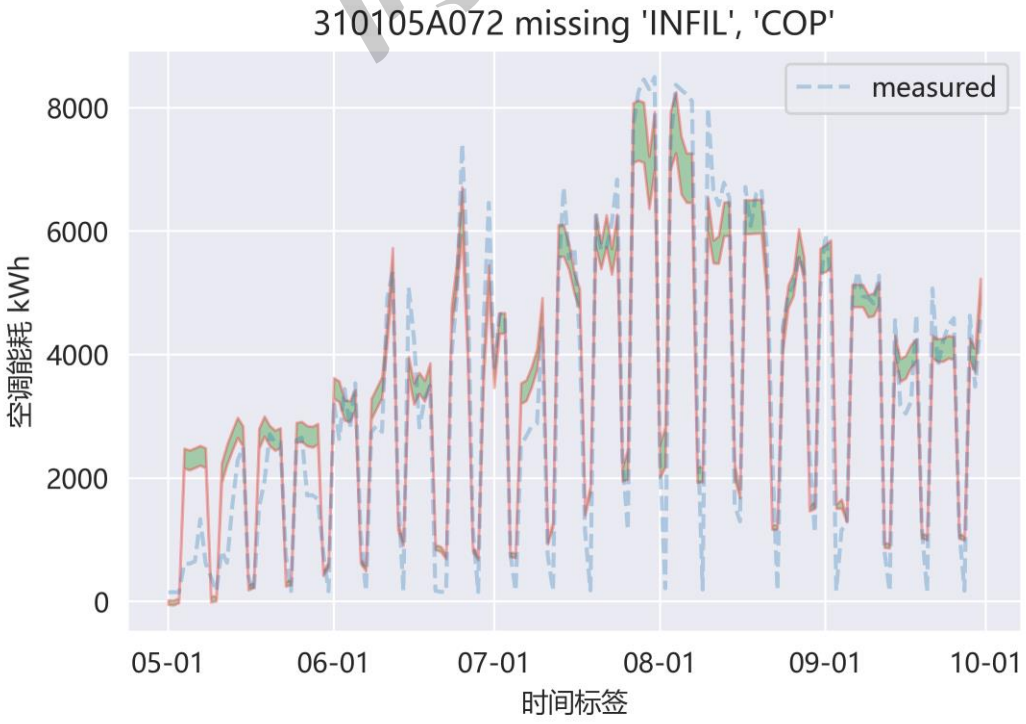


图 6.9 COP 和 INFIL 同时缺失时的预测结果

同济大学

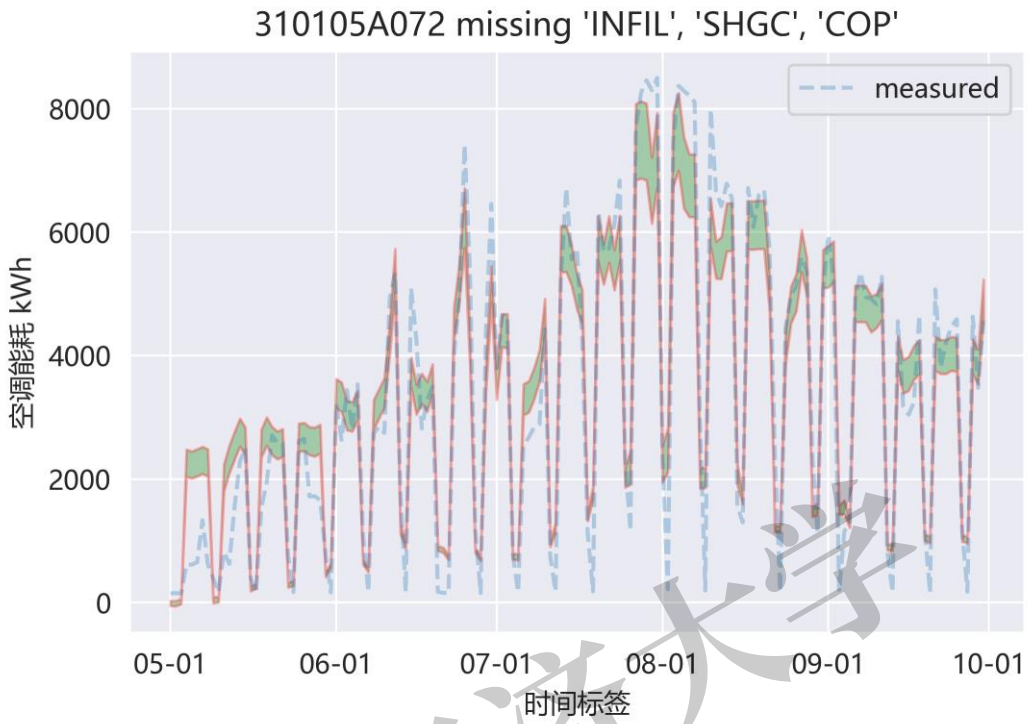


图 6.10 COP、INFIL 和 SHGC 同时缺失时的预测结果

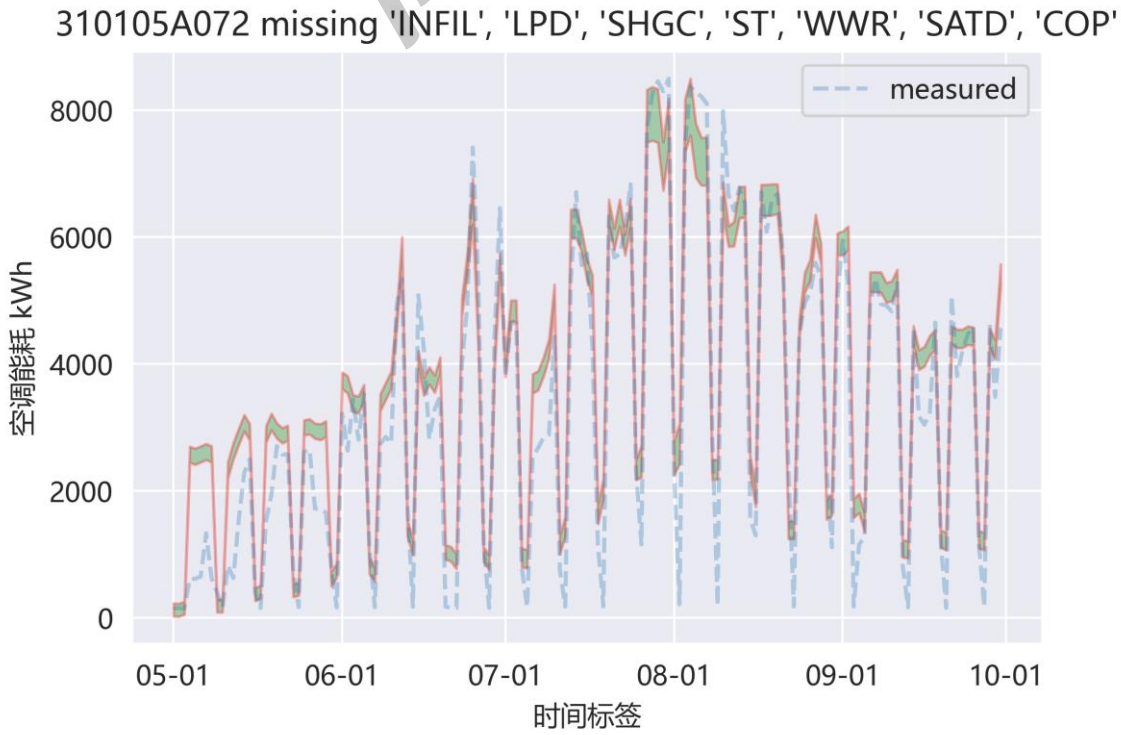


图 6.11 INFIL、LPD、SHGC、ST、WWR、SATD 和 COP 同时缺失时的预测结果

同济大学

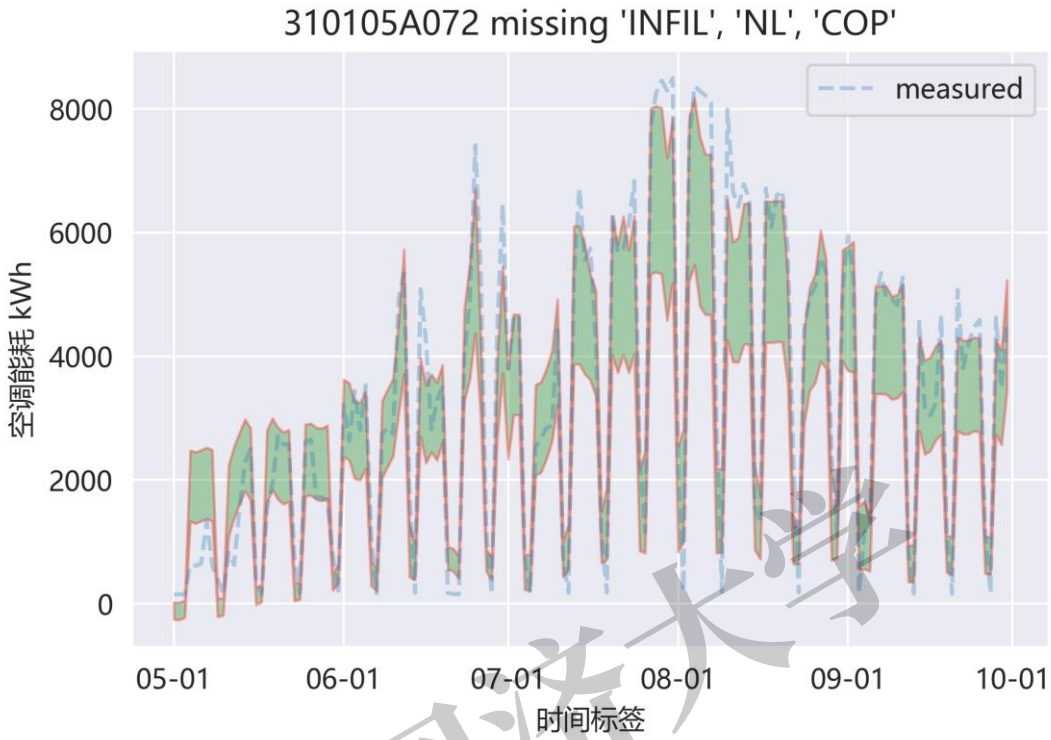


图 6.12 INFIL、COP 和 NL 同时缺失时的预测结果

同济大学

6.4 本章小结

本章对本文研究了基于多源异构数据的办公建筑能耗预测方法进行了验证，包括确定性和不确定性预测的验证。首先介绍了各种预测方法的使用情景和验证方法，主要是用于训练和测试的数据划分。其次交代了多源异构数据库中涉及到的数据来源、数据构成和对应的处理方法。最后进行了不同情况下的混合模型预测结果验证。在关键变量为确定性输入的确定性预测下，展示了多栋建筑的预测结果，除与训练建筑存在明显差异的建筑外，大部分建筑的预测误差 CV_RMSE 大部分在 0.3 左右，与训练建筑存在明显差异的建筑预测结果存在较大偏差，这表明应对办公建筑进行进一步的细分。在关键变量为一个区间的不确定性预测下，探究了缺失变量的数量、缺失变量的推测误差及缺失变量对预测目标的重要程度对预测误差和不确定性的影响，结果表明除缺失变量的数量外，缺失变量的推测准确度和对预测目标的重要程度也会显著地影响预测结果和不确定性。

同济大学

第 7 章 结论与展望

7.1 主要结论与成果

本文的研究内容为基于多源异构数据的办公建筑能耗预测方法，综合应用了多种来源的建筑信息和数据，结合静态物理数据和动态运行数据，从众多可能影响能耗的变量中提取了关键变量并进行了关键变量缺失值的推测，建立了物理模型和数据驱动模型相结合的混合模型，与现有能耗模型相比，本文研究的能耗预测混合模型提高了模型的通用性，可实现很好的实现跨建筑的能耗预测，大大扩展了能耗模型的适用情景。

能耗预测混合模型的建立包含：关键变量提取，关键变量缺失值推测，多源能耗数据融合，多源异构数据库建立和办公建筑能耗预测混合模型建立这几个步骤。下文将逐一介绍本研究的结论与成果：

1) 本文第 2 章提取了对办公建筑空调能耗有显著影响的关键变量。无论是白箱模型还是黑箱模型，都需要较多的输入变量。白箱模型的输入变量是不可或缺的，而黑箱模型的输入变量直接影响了预测准确度，较少的输入变量不能全面的反映能耗的影响因素，较多的输入变量将会导致模型计算量过大并且与能耗关联度不大的变量将会增大模型过拟合的可能。在第 2 章中使用相关系数法（SRCC 和 PRCC）、Morris 法和 XGBoost 算法分别提取出负荷相关和机电系统相关对空调能耗影响较大的关键变量，制冷和供热能耗提取到的关键变量比较类似，负荷相关的关键变量包括：制冷设定温度、体形系数、新风渗透率等，机电系统相关的关键变量包括：供回水温差、水系统类型、送风温度等。表 2.2 和图 2.6 分别展示了负荷和系统相关的关键变量。此外，在进行关键变量提取过程中，本研究利用 python 语言和其 eppy 库[54]建立了办公建筑的能耗模拟工具，能够灵活快速地批量化生成建筑能耗模拟所需的 EnergyPlus 模型。

2) 本文第 3 章对不同情景下的关键变量缺失值进行了推测。在实践中，某些关键变量值是难以测量和获取的（如新风渗透率），本文第 3 章通过能耗与关键变量的关系及已知变量与未知关键变量的关系提出了关键变量缺失值的推测方法。在存在历史能耗的情况下通过遗传算法进行关键变量缺失值的推测，这通常用在数据库的建立过程；在不存在历史能耗的情况下采用 Apriori 关联规则挖掘算法进行关键变量缺失值的推测，这通常用在对无历史能耗的建筑进行能耗预测的情景下。第 3 章还分析了已知逐日或逐月能耗的情景下，缺失一个或多个变

量时的推测准确性。大体来说，已知逐日能耗时的推测误差和误差分布均小于已知逐月能耗时的误差和误差分布，缺失的变量越多，推测误差越高，在仅已知 5 个关键变量的情况下，多栋测试建筑的关键变量推测的平均误差在可接受范围内，但存在些许离群值。

3) 本文第 4 章对不同来源的数据进行了不同的预处理，进行了数据融合。由于不同来源的数据具有不同的特点，各种数据的差异也较大，无法进行直接综合利用，第 4 章根据三类不同数据的不同特点进行了数据预处理。来自能耗审计平台的数据，其颗粒度较小但数据质量不高，对其进行了异常值和缺失值的处理。来自节能审计报告的数据，其可靠度高但颗粒度较粗，通过聚类提取典型的能耗曲线，对其进行了时间颗粒度的细化或转换，经验证该方法对能耗数据时间颗粒度的细化效果较好。对于快速建模工具生成的模拟数据，其数据较为丰富且细致，但与实测数据差异较大，通过两级的数据驱动修正模型对进行了修正，使其更加贴近实际数据。

4) 本文第 5 章在前文的工作基础上建立了多源异构数据库和办公建筑能耗预测混合模型。本文提出的办公建筑能耗预测方法可实现多种场景下的能耗预测，尤其是能够在没有任何历史能耗数据的场景下使用。本文第 6 章进行了不同情况下的混合模型预测结果验证。在关键变量为确定性输入的确定性预测下，展示了多栋建筑的预测结果，除与训练建筑存在明显差异的建筑外，大部分建筑的预测误差 CV_RMSE 在 0.3~0.4 之间，与训练建筑存在明显差异的建筑预测结果存在较大偏差，这表明应对办公建筑进行进一步的细分。在关键变量为一个区间的不确定性预测下，探究了缺失变量的数量、缺失变量的推测误差及缺失变量对预测目标的重要程度对预测误差和不确定性的影响，结果表明除缺失变量的数量外，缺失变量的推测准确度和对预测目标的重要程度也会显著地影响预测结果和不确定性。

7.2 主要贡献

本研究在课题组已有研究[42]基础上探索了办公建筑基于多源数据的能耗预测方法。与已有研究相比，本研究的提升点如下：

1) 在关键变量提取部分，对快速建模工具进行了优化，利用 python 语言和其 eppy 库对机电模型部分进行了更详细的模型描述，并对影响逐时能耗（而非峰值能耗）的变量进行了关键变量提取。

2) 采用遗传算法显著降低了关键变量缺失值推测部分的运行时间。

3) 采用更严谨与准确的方法进行数据融合。采用两级数据驱动修正模型对

模拟数据进行修正，使用 K-Means 聚类的方法提取典型能耗曲线以细化节能审计报告的能耗数据，使这两种数据质量更高，更接近实测数据。

4) 验证了除酒店建筑外，能耗预测混合模型在办公建筑上的有效性。

与现有研究相比，本文的贡献有：

1) 从众多可能对能耗产生影响的变量中提取了关键变量，并通过第 2 章所述的不同场景下的关键变量推测方法解决了实际中经常会出现的关键变量缺失的问题。

2) 对于不同特点的能耗数据提出了对应的预处理方式，综合应用了多种来源的数据成为了可能。使用 K-Means 聚类的方法提取典型能耗曲线以细化节能审计报告的能耗数据，并验证了颗粒度细化的准确性。采用两级数据驱动修正模型对模拟数据进行修正，使其更切近实测数据，并在混合模型中将既包含物理信息又包含相似建筑能耗信息的修正后的模拟数据作为模型的输入特征之一。

3) 实现了办公建筑在完全无历史能耗情景下的能耗预测。

7.3 局限性与展望

由于数据量和研究时间的限制，本研究仍有一些值得进一步研究的内容：

1) 关于关键变量的提取，由于机电系统连接方式和控制逻辑难以参数化表示，机电系统相关的变量通常是类别型变量（如冷机启停控制方式），而离散的类别型变量的排列组合将会大大增加样本数据，故本文中机电系统相关的初始变量仅选取了从专家知识上判断对能耗有显著影响且容易量化的变量。机电系统相关初始变量的选择还需更深入的探索。

2) 由于数据量的限制，本文研究的基于多于异构数据的办公建筑能耗预测方法仅在办公建筑的逐日颗粒度上进行了验证，并且对于审计报告和能耗监测平台中可得到的对能耗预测也许有很大帮助的机电设备信息（如设备的额定功率等）没有加以利用。在后续研究中还需考虑更多的特征变量，将全楼宇建筑能耗、设备能耗等作为研究目标进行更细致和更广泛的研究与验证。

3) 对于关键变量缺失值推测带来的不确定性需进一步进行研究。

建筑能耗预测对节能事业至关重要，除有很强的科研价值外，还广泛应用在建筑节能量评估，模型预测控制等实践中。随着可再生能耗的兴起，电网供应侧的不确定性大大增加，电网供应侧和需求侧需进行供需匹配，这对需求侧的用电负荷预测提出了新的要求，而建筑作为主要的用电终端，建筑的能耗预测显得更为重要。而建筑能耗预测影响因素及不确定性较大，尤其受到人行为的影响较大，白箱模型和黑箱模型在能耗预测方面各有千秋，均无法完全满足所有场景下的预

同济大学

测要求，故混合模型是一种非常有潜力的能耗预测模型。笔者坚持工学专业的科学研究的终点应当是能为实践服务的工具，而不只是论文，故在接下来的工作中，笔者将致力于将本文的研究内容具象为可供用户使用的工具，一方面用户丰富了可供选择的工具，一方面也可检验和优化混合模型。笔者希望深入该领域的研究，能够为下一代的能耗模拟引擎略尽绵薄之力。

同济大学

参考文献

[1] WMO. State of the Global Climate 2020[EB/OL]//WMO. (2021). https://library.wmo.int/doc_num.php?explnum_id=10786.

[2] 清华大学建筑节能研究中心. 中国建筑节能年度发展研究报告[M]. 中国建筑工业出版社, 2020.

[3] 吴蔚沁, 徐强, 冯君等. 2020年上海市公共建筑能耗监测平台能耗数据分析[J/OL]. 上海节能, 2021. DOI:10.13770/j.cnki.issn2095-705x.2021.07.003.

[4] 支建杰, 吴蔚沁. 公共建筑能耗监测平台数据应用的探讨[J]. 2018城市发展与规划论文集, 2018: 428-434.

[5] 周新军. 我国能耗监测管理现状及未来发展趋势[J/OL]. 当代经济管理, 2014, 36: 46-50. DOI:10.13253/j.cnki.ddjjgl.2014.02.009.html.

[6] 刘芳, 马晓雯. 基于公共建筑能耗监测系统的节能管理应用研究[J/OL]. 建筑节能, 2015, 6: 0-4. DOI:10.3969/j.issn.1673-7237.2015.06.023.

[7] 本刊编辑部. 解密高效机房系统[J]. 机电信息, 2020, 28: 18-28.

[8] Pérez-Lombard L, Ortiz J, Pout C. A review on buildings energy consumption information[J/OL]. Energy and Buildings, 2008, 40(3): 394-398. DOI:10.1016/j.enbuild.2007.03.007.

[9] Westphalen D, Koszalinski S, Little A D. Energy Consumption Characteristics of Commercial Building HVAC Systems Volume I : Chillers , Refrigerant Compressors , and Heating Systems And[M]//Building: I. 2001.

[10] Gyalistras D, Gwerder M, Oldewurtel F, et al. Analysis of Energy Savings Potentials for Integrated Room Automation[J/OL]. 10th REHVA World Congr. Clima, 2010: 8. http://www.sysecol2.ethz.ch/OptiControl/Lit/Gyal_10_Proc-Clima2010.pdf.

[11] Drgoňa J, Arroyo J, Cupeiro Figueroa I, et al. All you need to know about model predictive control for buildings[J/OL]. Annual Reviews in Control, 2020, 50(September): 190-232. DOI:10.1016/j.arcontrol.2020.09.001.

[12] Department of Energy, LBNL. Building Performance Database[EB/OL]. (2021)[2022-01-06]. https://bpd.lbl.gov/.

[13] CER. Data from the Commission for Energy Regulation[EB/OL]. (2012)[2022-01-06]. https://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

[14] Miller C, Meggers F. The Building Data Genome Project: An open, public data set from non-residential building electrical meters[J/OL]. Energy Procedia, 2017, 122: 439-444. https://doi.org/10.1016/j.egypro.2017.07.400. DOI:10.1016/j.egypro.2017.07.400.

[15] Miller C, Kathirgamanathan A, Picchetti B, et al. The Building Data Genome Project 2 , energy meter data from the ASHRAE Great Energy Predictor III competition[J/OL]. 2020: 1-13. DOI:10.1038/s41597-020-00712-x.

[16] 周浩, 田昕, 林波荣等. 北京市公共建筑能耗数据应用服务方案探讨[J/OL]. 2020: 22-31. DOI:10.16116/j.cnki.jskj.2020.16.004.

[17] Foucquier A, Robert S, Suard F, et al. State of the art in building modelling and energy performances prediction: A review[J/OL]. Renewable and Sustainable Energy Reviews, 2013, 23: 272-288. DOI:10.1016/j.rser.2013.03.004.

[18] 潘毅群等. 实用建筑能耗模拟手册[M]. 北京: 中国建筑工业出版社, 2013.

[19] Harish V S K V, Kumar A. A review on modeling and simulation of building energy systems[J/OL]. Renewable and Sustainable Energy Reviews, 2016, 56: 1272-1292. http://dx.doi.org/10.1016/j.rser.2015.12.040. DOI:10.1016/j.rser.2015.12.040.

[20] Raftery P, Keane M, O'Donnell J. Calibrating whole building energy models: An evidence-based methodology[J/OL]. Energy and Buildings, 2011, 43(9): 2356-2364. DOI:10.1016/J.ENBUILD.2011.05.020.

[21] 李梅香, 彭惠旺, 陈毅兴等. 商场建筑运行能耗实测数据修复方法研究[J]. 建筑节能(中英文), 2021, 49(5): 37-45.

[22] 高英博, 李德英, 顾中煊等. 能耗预测导向的建筑能耗异常数据识别与修复[J]. 科学技术与工程, 2019, 19(35): 1671-1815.

同济大学

[23] Kim J, Naganathan H, Moon S Y, et al. Applications of Clustering and Isolation Forest Techniques in Real-Time Building Energy-Consumption Data: Application to LEED Certified Buildings[J/OL]. Journal of Energy Engineering, 2017, 143(5): 04017052. DOI:10.1061/(asce)ey.1943-7897.0000479.

[24] Cho B, Dayrit T, Gao Y, et al. Effective Missing Value Imputation Methods for Building Monitoring Data[J/OL]. Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020, 2020: 2866-2875. DOI:10.1109/BigData50022.2020.9378230.

[25] Zhang L, Wen J. A systematic feature selection procedure for short-term data-driven building energy forecasting model development[J/OL]. Energy and Buildings, 2019, 183: 428-442. https://doi.org/10.1016/j.enbuild.2018.11.010. DOI:10.1016/j.enbuild.2018.11.010.

[26] 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/OL]. Sustainable Cities and Society, 2019, 50(March): 101643. https://doi.org/10.1016/j.scs.2019.101643. DOI:10.1016/j.scs.2019.101643.

[27] Li H, Wang S, Cheung H. Sensitivity analysis of design parameters and optimal design for zero/low energy buildings in subtropical regions[J/OL]. Applied Energy, 2018, 228(June): 1280-1291. https://doi.org/10.1016/j.apenergy.2018.07.023. DOI:10.1080/00401706.1991.10484804.

[28] Ding Y, Zhang Q, Yuan T, et al. Model input selection for building heating load prediction: A case study for an office building in Tianjin[J/OL]. Energy and Buildings, 2018, 159: 254-270. http://dx.doi.org/10.1016/j.enbuild.2017.11.002. DOI:10.1016/j.enbuild.2017.11.002.

[29] 朱明亚. 办公建筑能耗预测模型最小变量集构成方法[D]. 同济大学机械与能源工程学院, 2019.

[30] Miller C, Arjunan P, Kathirgamanathan A, et al. The ASHRAE Great Energy Predictor III competition: Overview and results[J/OL]. Science and Technology for the Built Environment, 2020, 26(10): 1427-1447. DOI:10.1080/23744731.2020.1795514.

[31] Chen Y, Guo M, Chen Z, et al. Physical energy and data-driven models in building energy prediction: A review[J/OL]. Energy Reports, 2022, 8: 2656-2671. https://doi.org/10.1016/j.egyr.2022.01.162. DOI:10.1016/j.egyr.2022.01.162.

[32] Li Y, O'Neill Z, Zhang L, et al. Grey-box modeling and application for building energy simulations - A critical review[J/OL]. Renewable and Sustainable Energy Reviews, 2021, 146(May): 111174. https://doi.org/10.1016/j.rser.2021.111174. DOI:10.1016/j.rser.2021.111174.

[33] Hassid S. A linear model for passive solar calculations: Evaluation of performance[J/OL]. Building and Environment, 1985, 20(1): 53-59. DOI:10.1016/0360-1323(85)90032-0.

[34] Hazyuk I, Ghiaus C, Penhouet D. Model Predictive Control of thermal comfort as a benchmark for controller performance[J/OL]. Automation in Construction, 2014, 43: 98-109. DOI:10.1016/J.AUTCON.2014.03.016.

[35] Dewson T, Day B, Irving A D. Least squares parameter estimation of a reduced order thermal model of an experimental building[J/OL]. Building and Environment, 1993, 28(2): 127-137. DOI:10.1016/0360-1323(93)90046-6.

[36] Zhang D, Xia X, Cai N. A dynamic simplified model of radiant ceiling cooling integrated with underfloor ventilation system[J/OL]. Applied Thermal Engineering, 2016, 106: 415-422. DOI:10.1016/J.APPLTHERMALENG.2016.06.017.

[37] Ogunsola O, Song L. Review and evaluation of using R-C thermal modeling of cooling load prediction for HVAC system control purpose[C]//PROCEEDINGS OF THE ASME INTERNATIONAL MECHANICAL ENGINEERING CONGRESS AND EXPOSITION. 2012: 735-743.

[38] Wang S, Xu X. Simplified building model for transient thermal performance estimation using GA-based parameter identification[J/OL]. International Journal of Thermal Sciences, 2006, 45(4): 419-432. DOI:10.1016/J.IJTHERMALSCI.2005.06.009.

[39] Wang S, Xu X. Parameter estimation of internal thermal mass of building dynamic models using genetic algorithm[J/OL]. Energy Conversion and Management, 2006, 47(13-14): 1927-1941. DOI:10.1016/J.ENCONMAN.2005.09.011.

[40] Lam J C, Hui S C M, Chan A L S. Regression analysis of high-rise fully air-conditioned office buildings[J/OL]. Energy and Buildings, 1997, 26(2): 189-197. DOI:10.1016/S0378-7788(96)01034-1.

[41] Lam J C, Hui S C M. Sensitivity analysis of energy performance of office buildings[J/OL]. Building and Environment, 1996, 31(1): 27-39. DOI:10.1016/0360-1323(95)00031-3.
[42] 沙华晶. 建筑信息异构数据融合方法及混合能耗模型的建立[D]. 同济大学机械与能源工程学院, 2021.
[43] Panigrahi S, Nanda A, Swarnkar T. A Survey on Transfer Learning[J/OL]. Smart Innovation, Systems and Technologies, 2021, 194: 781-789. DOI:10.1007/978-981-15-5971-6_83.
[44] Fan C, Sun Y, Xiao F, et al. Statistical investigations of transfer learning-based methodology for short-term building energy predictions[J/OL]. Applied Energy, 2020, 262(November 2019): 114499. https://doi.org/10.1016/j.apenergy.2020.114499. DOI:10.1016/j.apenergy.2020.114499.
[45] Li A, Xiao F, Fan C, et al. Development of an ANN-based building energy model for information-poor buildings using transfer learning[J/OL]. Building Simulation, 2021, 14(1): 89-101. DOI:10.1007/s12273-020-0711-5.
[46] Tian Y, Sehovac L, Grolinger K. Similarity-Based Chained Transfer Learning for Energy Forecasting with Big Data[J/OL]. IEEE Access, 2019, 7: 139895-139908. DOI:10.1109/ACCESS.2019.2943752.
[47] Ribeiro M, Grolinger K, Elyamany H F, et al. Energy & Buildings Transfer learning with seasonal and trend adjustment for cross-building energy forecasting[J/OL]. Energy & Buildings, 2018, 165: 352-363. https://doi.org/10.1016/j.enbuild.2018.01.034. DOI:10.1016/j.enbuild.2018.01.034.
[48] Qian F, Gao W, Yang Y, et al. Potential analysis of the transfer learning model in short and medium-term forecasting of building HVAC energy consumption[J/OL]. Energy, 2020, 193: 116724. https://doi.org/10.1016/j.energy.2019.116724. DOI:10.1016/j.energy.2019.116724.
[49] Fang X, Gong G, Li G, et al. A hybrid deep transfer learning strategy for short term cross-building energy prediction[J/OL]. Energy, 2021, 215: 119208. https://doi.org/10.1016/j.energy.2020.119208. DOI:10.1016/j.energy.2020.119208.
[50] Mocanu E, Nguyen P H, Kling W L, et al. Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building transfer learning[J/OL]. Energy & Buildings, 2016, 116: 646-655. http://dx.doi.org/10.1016/j.enbuild.2016.01.030. DOI:10.1016/j.enbuild.2016.01.030.
[51] Xiao T, Xu P, He R, et al. Status quo and opportunities for building energy prediction in limited data Context—Overview from a competition[J/OL]. Applied Energy, 2022, 305(September 2021): 117829. https://doi.org/10.1016/j.apenergy.2021.117829. DOI:10.1016/j.apenergy.2021.117829.
[52] 中华人民共和国住房与城乡建设部. GB 50189-2015 公共建筑节能设计标准[A]. 中华人民共和国, 2015.
[53] 陆耀庆. 实用供热空调设计手册[M]. 第二版. 北京: 中国建筑工业出版社, 2008.
[54] Philip S, Tanjuatco L. Eppy Tutorial[EB/OL]. (2021). https://eppy.readthedocs.io/en/latest/Main_Tutorial.html.
[55] Morrison Hershfield Limited. Building envelope thermal bridging guide[M]. VERSION1.6. Vancouver, BC: BC Hydro Power Smart, 2021.
[56] Heiselberg P, Brohus H, Hesselholt A, et al. Application of sensitivity analysis in design of sustainable buildings[J/OL]. Renewable Energy, 2009, 34(9): 2030-2036. http://dx.doi.org/10.1016/j.renene.2009.02.016. DOI:10.1016/j.renene.2009.02.016.
[57] Morris M D. Factorial sampling plans for preliminary computational experiments[J/OL]. Technometrics, 1991, 33(May, 1991): 161-174. DOI:10.1177/001872086700900503.
[58] Campolongo F, Cariboni J, Saltelli A. An effective screening design for sensitivity analysis of large models[J/OL]. Environmental Modelling and Software, 2007, 22(10): 1509-1518. DOI:10.1016/j.envsoft.2006.10.004.
[59] 沙华晶, 许鹏, 钟文智等. 建筑空调能耗关键变量通用提取方法及工具的开发[J/OL]. 土木与环境工程学报(中英文), 2021. DOI:10.11835/j.issn.2096-6717.2021.044.
[60] Yildiz Y, Arsan Z D. Identification of the building parameters that influence heating and cooling energy loads for apartment buildings in hot-humid climates[J/OL]. Energy, 2011, 36(7): 4287-4296. DOI:10.1016/j.energy.2011.04.013.
[61] Tian W, De Wilde P. Uncertainty and sensitivity analysis of building performance using probabilistic climate projections: A UK case study[J/OL]. Automation in Construction, 2011, 20(8): 1096-1109. http://dx.doi.org/10.1016/j.autcon.2011.04.011.

同济大学

DOI:10.1016/j.autcon.2011.04.011.
[62] Chen T, Guestrin C. XGBoost : A Scalable Tree Boosting System[C/OL]//the 22nd ACM SIGKDD International Conference. 2016: 785-794. DOI:https://doi.org/10.1145/2939672.2939785.
[63] Andersson H. tsod 0.1.4[EB/OL]. (2022). https://pypi.org/project/tsod/.
[64] Chen Z, Chen Y, Xiao T, et al. A novel short-term load forecasting framework based on time-series clustering and early classification algorithm[J/OL]. Energy and Buildings, 2021, 251(September): 111375. https://doi.org/10.1016/j.enbuild.2021.111375. DOI:10.1016/j.enbuild.2021.111375.

同济大学

附录 A 各个月份聚类结果

表 A-1 每月每簇的用能特性

月份	簇的个数	每簇用能特性
1	2	Cluster1：热源不消耗电能的建筑或热负荷极低的综合建筑 Cluster2：建筑热源消耗电能且有一定热负荷的办公建筑
2	2	Cluster1：热源不消耗电能的建筑或热负荷极低的综合建筑 Cluster2：建筑热源消耗电能且有一定热负荷的办公建筑
3	2	Cluster1：热源不消耗电能的建筑或热负荷极低的综合建筑 Cluster2：建筑热源消耗电能且有一定热负荷的办公建筑
4	2	Cluster1：过渡季不开空调的办公建筑 Cluster2：过渡季开空调的办公建筑
5	3	Cluster1：过渡季开空调的办公建筑 Cluster2：过渡季开空调但能耗密度较低的办公建筑 Cluster3：过渡季不开空调的办公建筑
6	2	Cluster1：办公为主的建筑 Cluster2：综合建筑
7	2	Cluster1：办公为主的建筑 Cluster2：综合建筑
8	2	Cluster1：办公为主的建筑 Cluster2：综合建筑
9	2	Cluster1：办公为主的建筑 Cluster2：综合建筑
10	3	Cluster1：综合建筑 Cluster2：过渡季能耗密度低的普通办公建筑 Cluster3：过渡季能耗密度高的高级办公建筑
11	3	C Cluster1：过渡季开空调但能耗密度较低的建筑 Cluster2：过渡季不开空调的建筑或热源不消耗电能的建筑 Cluster3：过渡季开空调，建筑热源消耗电能且有一定热负荷的办公建筑
12	2	Cluster1：热源不消耗电能的建筑或热负荷极低的综合建筑 Cluster2：建筑热源消耗电能且有一定热负荷的办公建筑

每月的逐日制冷能耗聚类曲线下图所示：

同济大学

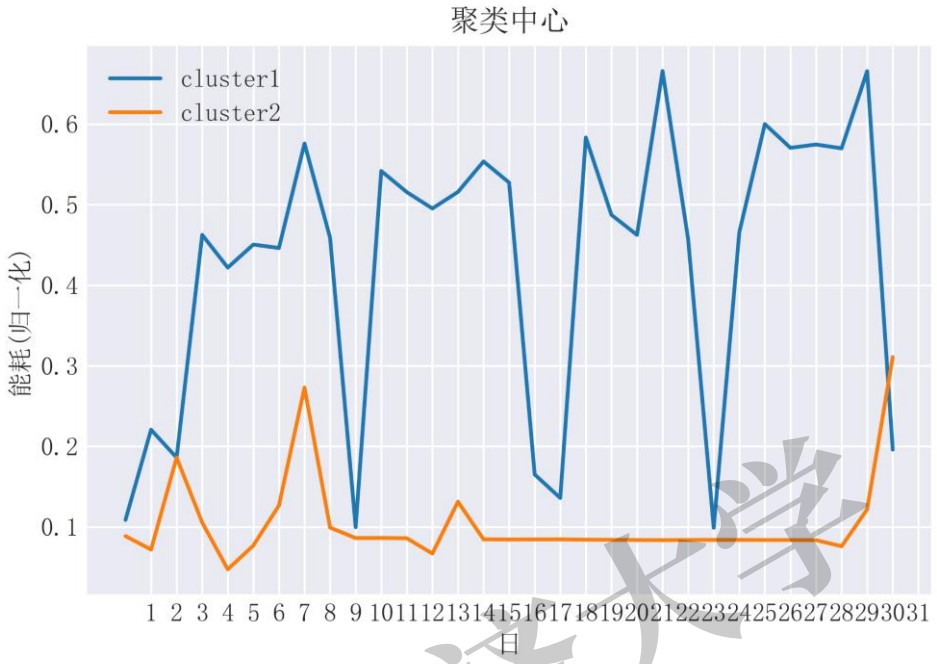


图 A-1 1 月份空调能耗聚类结果

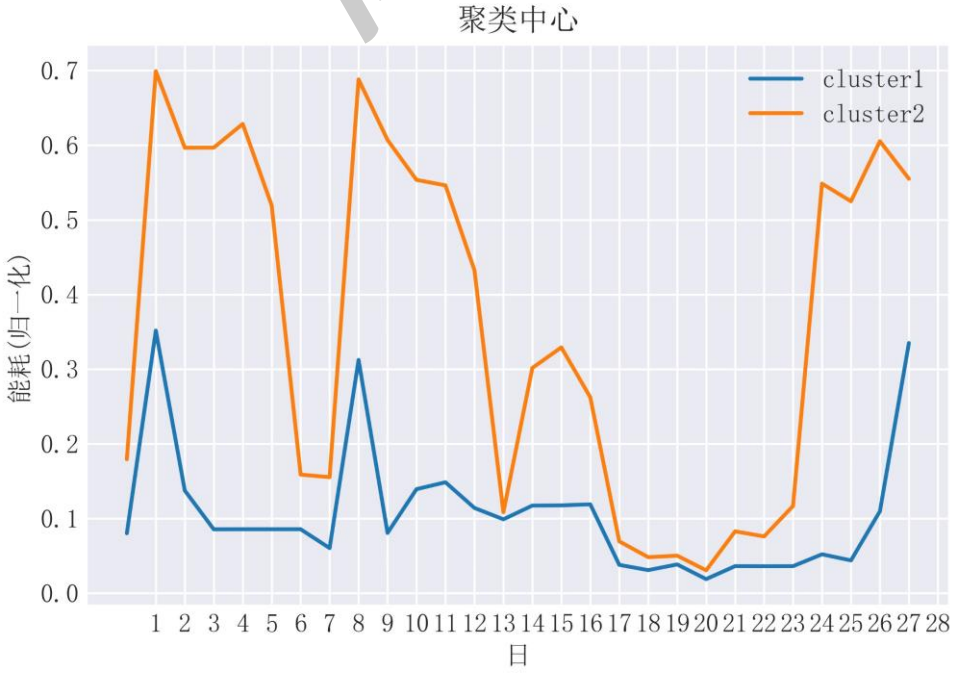


图 A-2 2 月份空调能耗聚类结果

同济大学

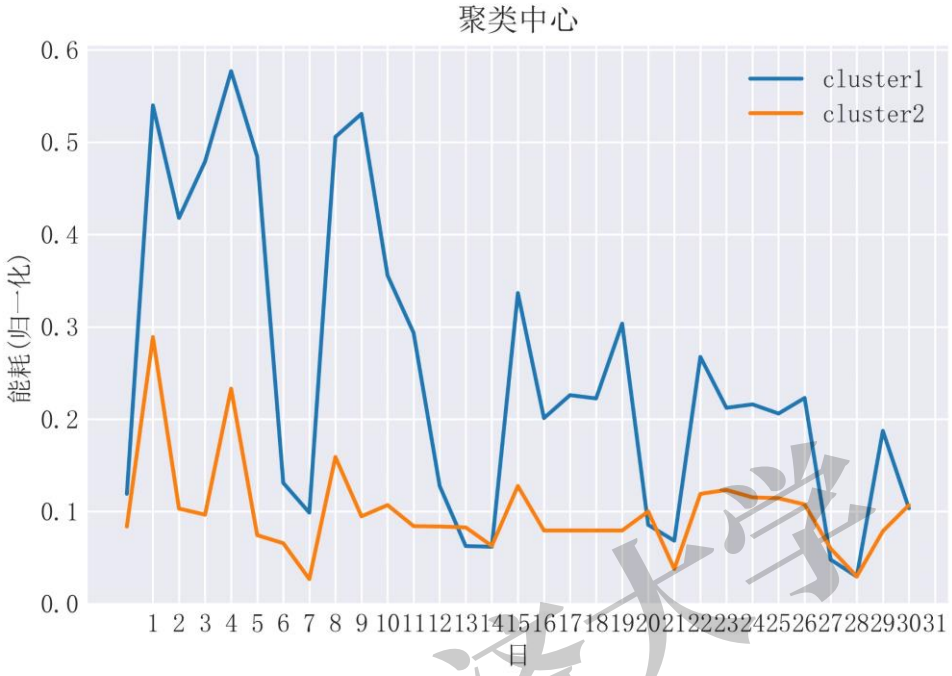


图 A-3 3 月份空调能耗聚类结果

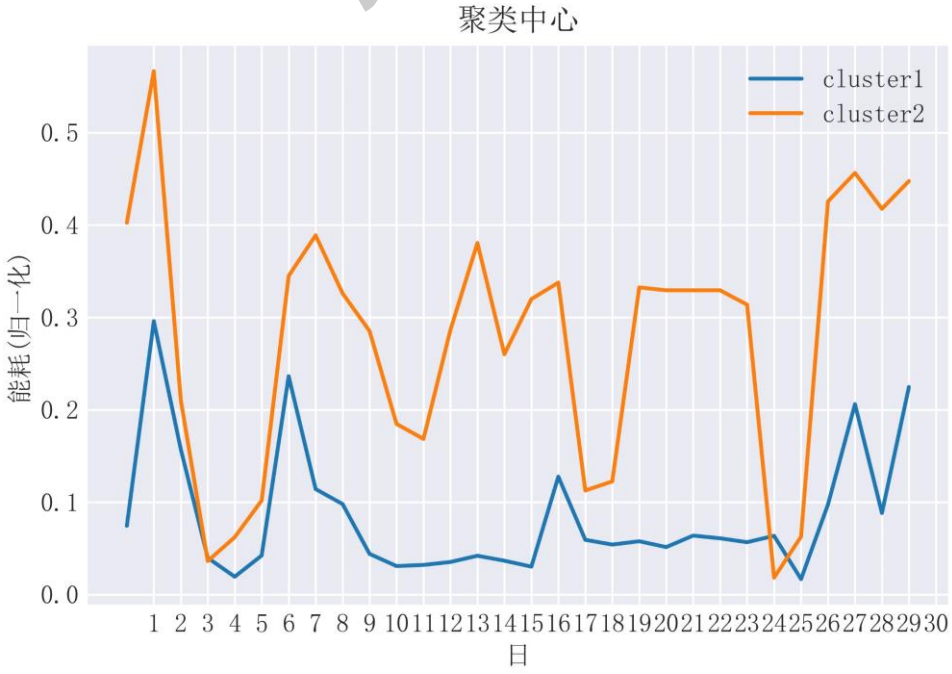


图 A-4 4 月份空调能耗聚类结果

同济大学

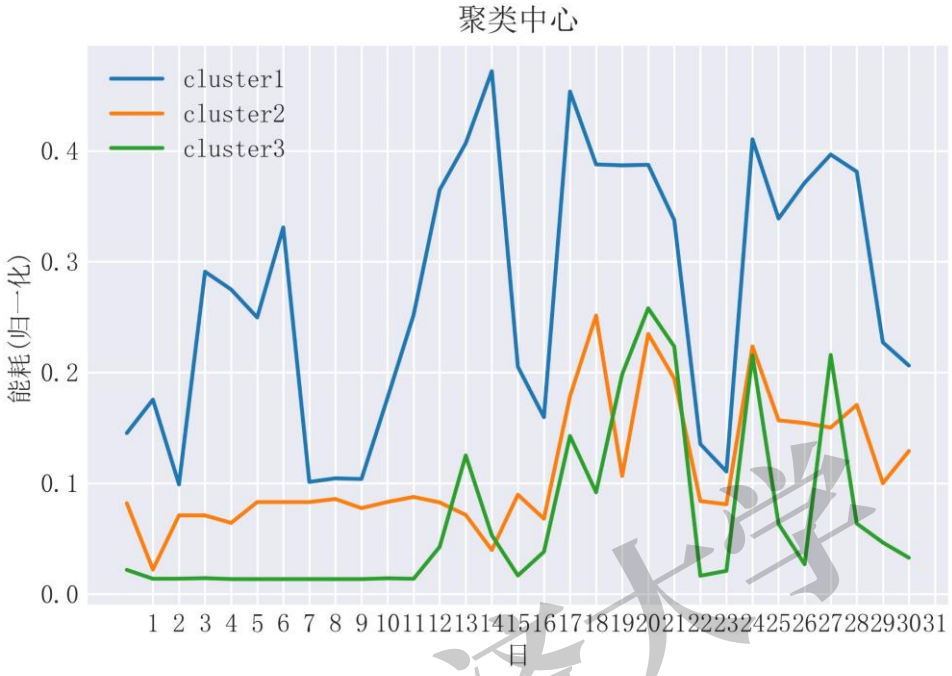


图 A-5 5 月份空调能耗聚类结果

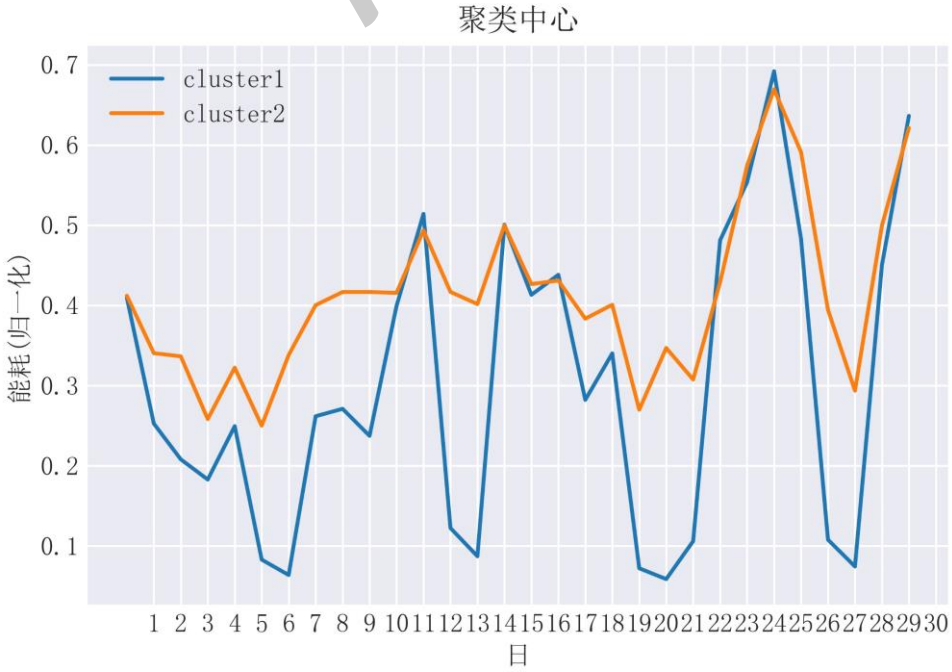


图 A-6 6 月份空调能耗聚类结果

同济大学

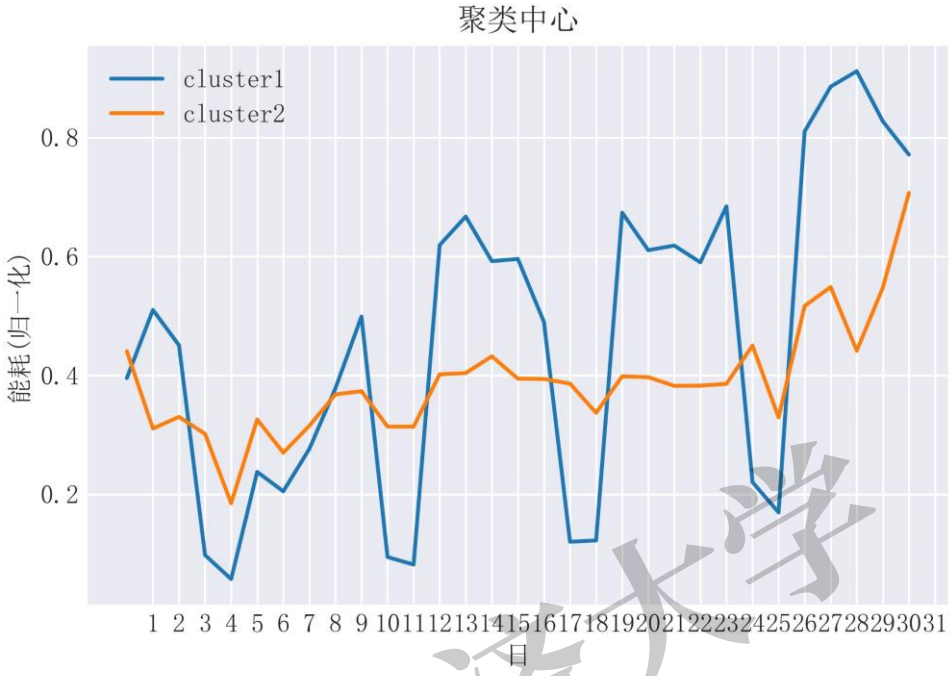


图 A-7 7 月份空调能耗聚类结果

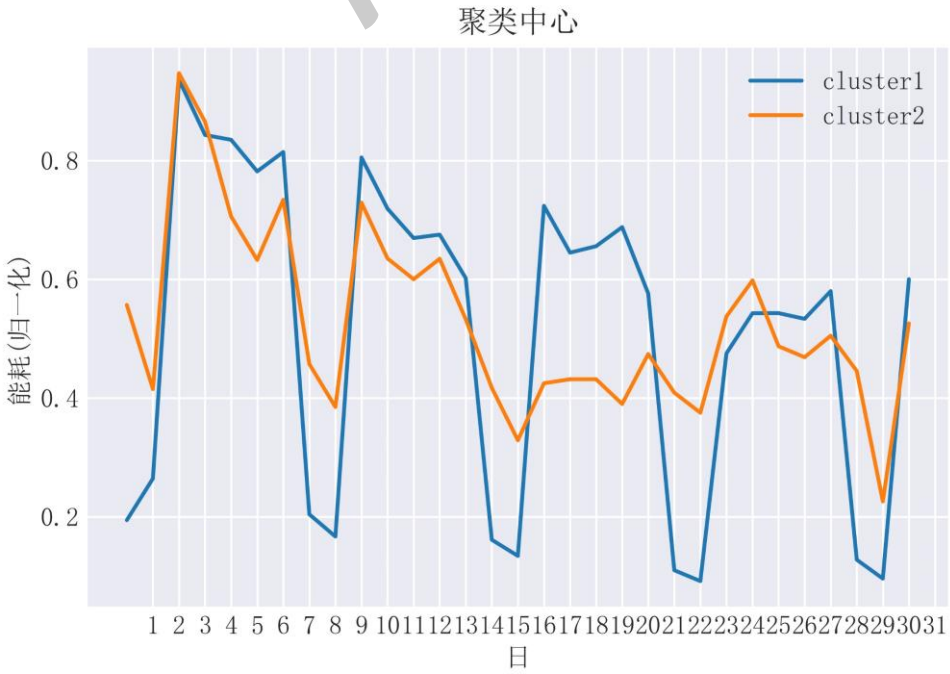


图 A-8 8 月份空调能耗聚类结果

同济大学

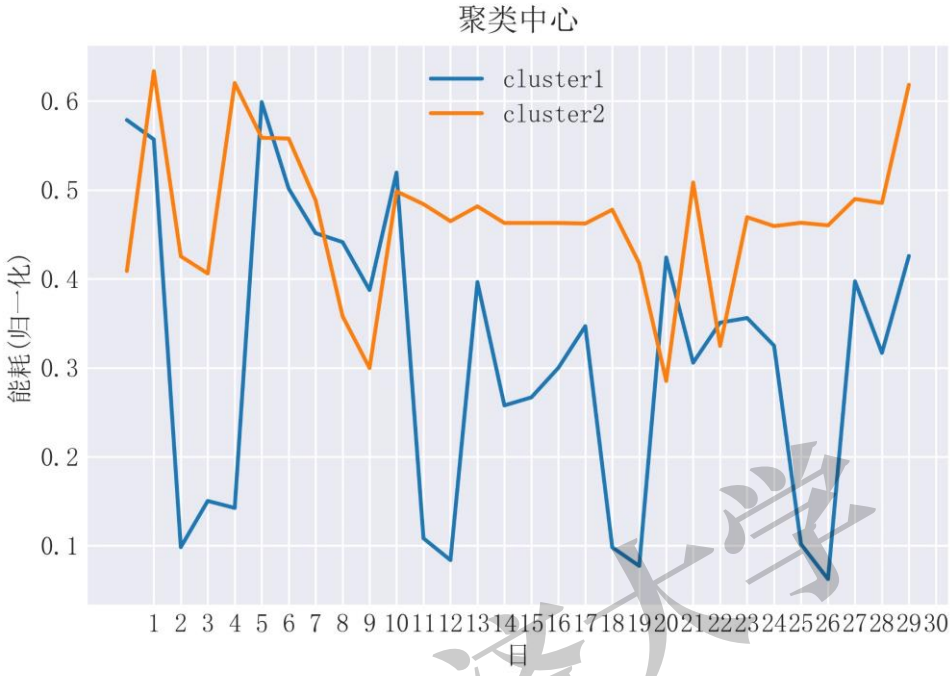


图 A-9 9 月份空调能耗聚类结果

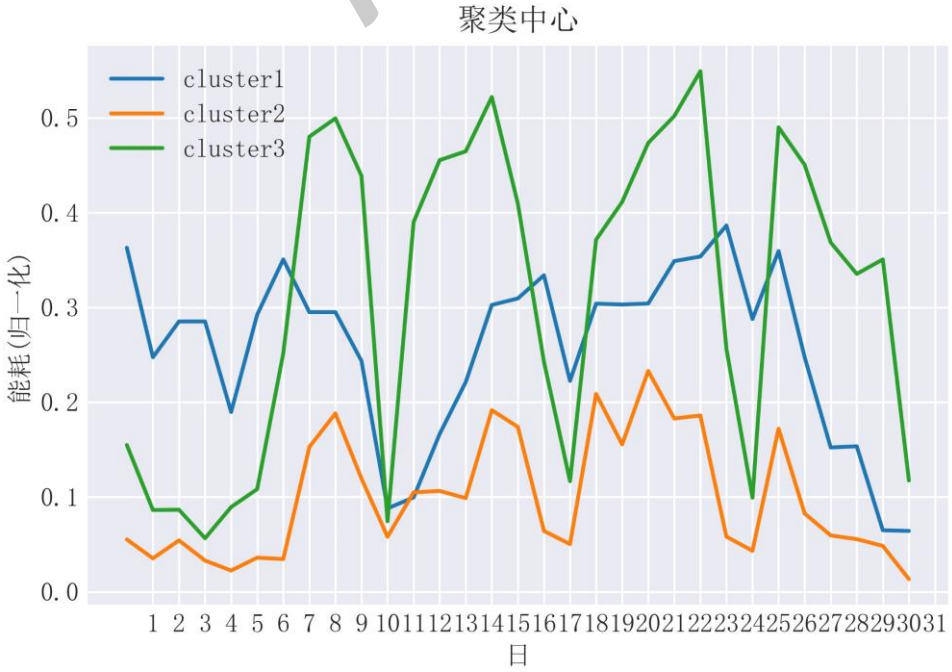


图 A-10 10 月份空调能耗聚类结果

同济大学

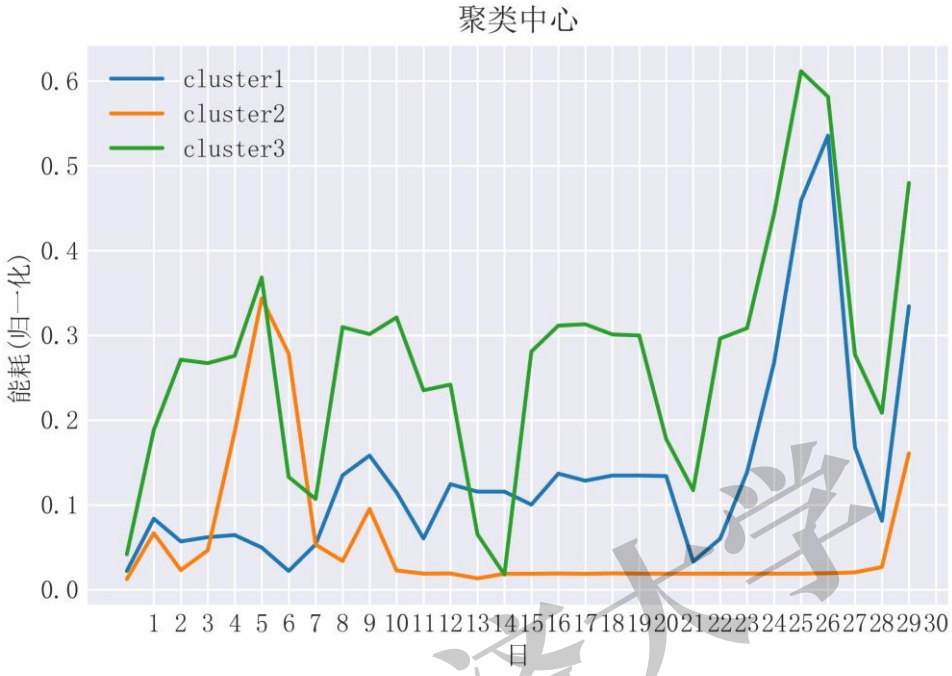


图 A-11 11 月份空调能耗聚类结果

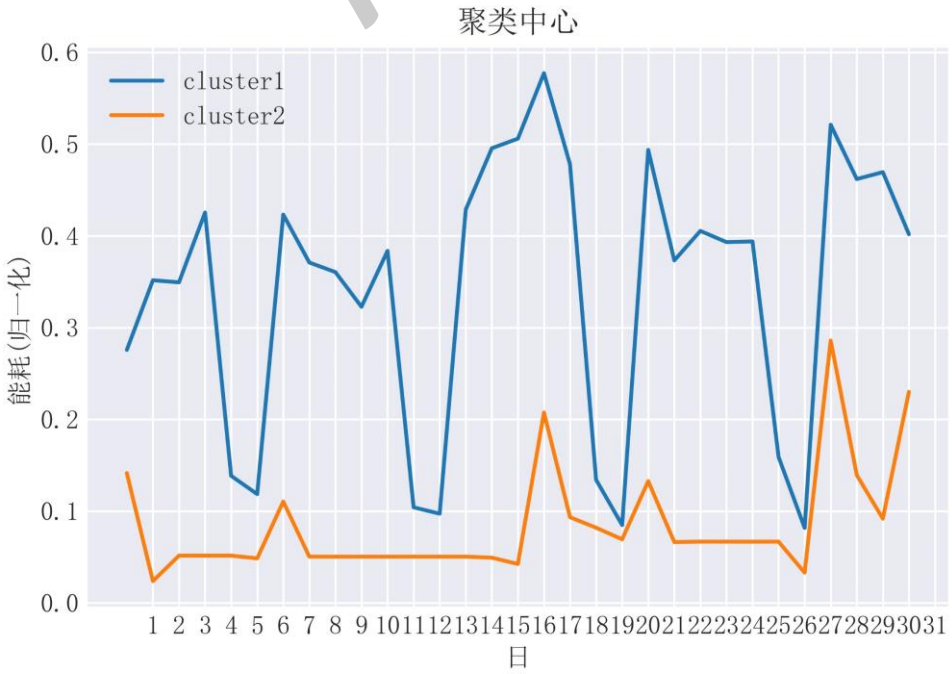


图 A-12 12 月份空调能耗聚类结果

同济大学

附录 B 关键变量提取部分代码节选

```
# 主程序
import os
dirname, filename = os.path.split(os.path.abspath(__file__))
os.chdir(dirname)
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 as prcc
def generate_idf_main(ds,refidfname,refname, iddfile):
    if bld_func == 'Hotel':
        room_function = ['lobby', 'service', 'dinning', 'kitchen', 'meeting', 'room']
        room_ratio = [0.1, 0.1, 0.075, 0.03, 0.025, 0.67]
        func =pd.DataFrame({'function':room_function, 'ratio':room_ratio})
    elif bld_func == 'OfficeBuilding':
        officebld_type = "Alloffice"
        if officebld_type == "Alloffice" :
            room_function = ['lobby', 'service', 'dinning', 'kitchen', 'meeting',
'office']
            room_ratio = [0.1, 0.1, 0.075, 0.025, 0.13, 0.57]
            func =pd.DataFrame({'function':room_function,
'ratio':room_ratio})
    #generate idf file of each sample, run simulation and store results
```

同济大学

```
for i in range(len(ds.index)):      #len(ds.index)
    NWWR = ds.loc[i]['NWWR']
    SWWR = ds.loc[i]['SWWR']
    EWWR = ds.loc[i]['EWWR']
    WWWR = ds.loc[i]['WWWR']
    AREA = ds.loc[i]['AREA']
    NL = ds.loc[i]['NL']
    CR = ds.loc[i]['CR']
    WALLU = ds.loc[i]['WALLU']
    WSP = ds.loc[i]['WSP']
    RU = ds.loc[i]['RU']
    WINU = ds.loc[i]['WINU']
    SHGC = ds.loc[i]['SHGC']
    SPC = ds.loc[i]['SPC']
    SPH = ds.loc[i]['SPH']
    LPD = ds.loc[i]['LPD']
    OPD = ds.loc[i]['OPD']
    INFIL = ds.loc[i]['INFIL']
    FLT = ds.loc[i]['FLT']
    GLT = ds.loc[i]['GLT']
    CLT = ds.loc[i]['CLT']
    ST = ds.loc[i]['ST']
    WSA = ds.loc[i]['WSA']
    RSA = ds.loc[i]['RSA']

    idf_T1,CR_real = IDFGenerate.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, refidfname, refname, iddfile)

    os.makedirs(sampling_method + 'Samples\Model' + str(i))
    idf_T1.saveas(sampling_method + 'Samples\Model' + str(i) + '\Model' +
str(i) + '.idf')
    idf = IDF(sampling_method + 'Samples\Model' + str(i) + '\Model' + str(i)
+ '.idf', weather_file)
```

同济大学

```
            idf.run(output_directory = sampling_method + 'Samples\Model' + str(i),
readvars = True, output_prefix = 'Model' + str(i), output_suffix = 'C')
            print(str(i+1) + 'Samples are finished')

        ds['CR'] = CR_real
        ds.to_excel('param_values_'+sampling_method+'_real_CR.xlsx')


    def sensitity_analysis(location,sampling_method):
        ## Morris method
        if sampling_method == 'Morris':
            ds = pd.read_excel('param_values_morris_real_CR.xlsx',index_col =
'Unnamed: 0')
            ds1 = (ds - ds.mean()) / ds.std()
            param_values = ds1.values
            bounds = SampleGenerate.bound(location)
            problem = SampleGenerate.Morris_problem(bounds)
            Si = morris.analyze(problem, param_values, output.values,
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', index_col =
"Unnamed: 0").dropna(how = "all")
            ds['output'] = output
            ds.to_excel('param_values_LHS_real_CR_output.xlsx')
```

```
# 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[:,:-1]
y=ds_scaler[:,-1]
linreg = LR()
model=linreg.fit(X, y)
SRRC = linreg.coef_
srrc_result = pd.DataFrame(data = SRRC, index= ds.columns.drop("output"), columns = ["SRRC"])
srrc_result['temp_sort'] = abs(srrc_result['SRRC'])
srrc_result = srrc_result.sort_values(by=['temp_sort']).drop(columns=['temp_sort'])
plt.figure()
plt.title('SRRC',fontsize='large', fontweight='bold')
plt.xticks(rotation=90, fontsize=10)
plt.bar(list(srrc_result.index), srrc_result["SRRC"])

## calculate PRCC
PRCC_matric = prcc.partial_corr(ds.values)
PRCC = list(PRCC_matric[:-1,-1])
prcc_result = pd.DataFrame(data = PRCC, index= ds.columns.drop("output"), columns = ["PRCC"])
prcc_result['temp_sort'] = abs(prcc_result['PRCC'])
prcc_result = prcc_result.sort_values(by=['temp_sort']).drop(columns=['temp_sort'])
plt.figure()
plt.title('PRCC',fontsize='large', fontweight='bold')
```

同济大学

```
            plt.xticks(rotation=90, fontsize=10)
            plt.bar(list(prcc_result.index), prcc_result["PRCC"])


    def path_defination():
        iddfile = r"S:\install\EnergyPlusV8-9-0\Energy+.idd"
        refname = "E+RefModel/"
        refidfname = "Ref_Model_Summer_HSCW"
        return iddfile, refname, refidfname


    if __name__ == "__main__":
        iddfile, refname, refidfname = path_defination()
        BldId = int(input("请输入建筑编号：\n0：Officebuilding；\n1: Hotel\n"))
        if(BldId == 0):
            bld_func = 'OfficeBuilding'
        elif(BldId == 1):
            bld_func = 'Hotel'
        LocationId = int(input("请输入建筑所处热工分区编号：\n0：HSCW(Shanghai)；\n1：Cold(Beijing)\n"))
        if LocationId == 0 :
            weather_file = r'weatherdata\CHN_Shanghai.Shanghai.583620_CSWD.epw'
            location = "HSCW"
        elif LocationId == 1:
            weather_file = r'weatherdata\CHN_Beijing.Beijing.545110_CSWD.epw'
            location = "Cold"
        SeasonId = int(input("请输入负荷类型编号：\n0: cooling load;\n1: heating load\n"))
        if SeasonId == 0:
            if BldId == 1:
                refmodel = "Ref_Model_Hotel.idf"
            else: #office
                if LocationId == 0: #Shsnghai
                    refmodel = "Ref_Model_Summer_HSCW.idf"
```

同济大学

```
                elif LocationId == 1: #Beijing:
                    refmodel = "Ref_Model_Summer_Cold.idf"

        elif SeasonId == 1:
            if BldId == 1:
                refmodel = "Ref_Model_Hotel.idf"
            else: #office
                if LocationId == 0: #Shanghai
                    refmodel = "Ref_Model_Year_HSCW.idf"
                elif LocationId == 1: #Beijing:
                    refmodel = "Ref_Model_Year_Cold.idf"
        sampling_method = int(input("分析方法编号\n0：Morris；\n1：LHS\n"))
        if sampling_method == 0 :
            sampling_method = 'Morris'
        elif sampling_method == 1:
            sampling_method = 'LHS'
        print("正在进行采样")
        global CR_real
        if(sampling_method == "LHS"):
            LHS_sample_num = 3000
            ds = SampleGenerate.sampling_LHS(location, num =
LHS_sample_num)
        elif (sampling_method == "Morris"):
            Morris_sample_num = 300
            num_levels = 10
            ds = SampleGenerate.sampling_Morris(location, num_levels,
Morris_sample_num)
        ##generate idf file of each sample, run simulation and store results
        ##generate samples
        print("正在进行模型生成")
        generate_idf_main(ds,refmodel,refname, iddfile)
        ## get simualtion results
        if SeasonId == 0:
            output = OutputCollect.output(sampling_method, SA_index =
```

同济大学

```
'Cool_sum')
        elif SeasonId == 1:
            output = OutputCollect.output(sampling_method, SA_index =
'Heat_sum')
        ## Sensitity analysis
        print("正在进行敏感性分析")
        sensitity_analysis(location,sampling_method)

    #批量生成 IDF 代码节选
    import time
    from multiprocessing import process
    from multiprocessing import Pool

    import pandas as pd
    import numpy as np
    from smt.sampling_methods import LHS
    import itertools
    from eppy.modeleditor import IDF
    import sys
    import os
    import math
    dirname, filename = os.path.split(os.path.abspath(__file__))
    os.chdir(dirname)

    #%% HVAC part IDF generate
    def layer_recognize(idf_T0):
        zone_idf = idf_T0.idfobjects["ZONE"]
        surfaces = idf_T0.idfobjects['BuildingSurface:Detailed']
        storey = []
        for zone in zone_idf:
            for surface in surfaces:
                if zone.Name == surface.Zone_Name and "floor" in surface.Name:
                    zone.Z_Origin = surface.Vertex_1_Zcoordinate
                    if (surface.Vertex_1_Zcoordinate not in storey):
```

同济大学

```
                storey.append(surface.Vertex_1_Zcoordinate)
    return storey, idf_T0


def euipment_schedule(idf_T1):
    schedule_idf = idf_T1.idfobjects['SCHEDULE:COMPACT']
    idf_T1.newidfobject('SCHEDULE:COMPACT')
    schedule_idf[-1].Name = "system_avail"
    schedule_idf[-1].Schedule_Type_Limits_Name = "On/Off"
    schedule_idf[-1].Field_1 = "Through: 12/31"
    schedule_idf[-1].Field_2 = "For: AllDays"
    schedule_idf[-1].Field_3 = "Until: 8:00"
    schedule_idf[-1].Field_4 = "0"
    schedule_idf[-1].Field_5 = "Until: 22:00"
    schedule_idf[-1].Field_6 = "1"
    schedule_idf[-1].Field_7 = "Until: 24:00"
    schedule_idf[-1].Field_8 = "0"


    schedule_idf = idf_T1.idfobjects['SCHEDULE:COMPACT']
    idf_T1.newidfobject('SCHEDULE:COMPACT')
    schedule_idf[-1].Name = "HTG_AVAIL_SCHE_equip"
    schedule_idf[-1].Schedule_Type_Limits_Name = "On/Off"
    schedule_idf[-1].Field_1 = "Through: 3/31"
    schedule_idf[-1].Field_2 = "For: AllDays"
    schedule_idf[-1].Field_3 = "Until: 24:00"
    schedule_idf[-1].Field_4 = "1"
    schedule_idf[-1].Field_5 = "Through: 4/30"
    schedule_idf[-1].Field_6 = "For: AllDays"
    schedule_idf[-1].Field_7 = "Until: 24:00"
    schedule_idf[-1].Field_8 = "0"
    schedule_idf[-1].Field_9 = "Through: 9/30"
    schedule_idf[-1].Field_10 = "For: AllDays"
    schedule_idf[-1].Field_11 = "Until: 24:00"
```

同济大学

```
        schedule_idf[-1].Field_12 = "0"
        schedule_idf[-1].Field_13 = "Through: 10/31"
        schedule_idf[-1].Field_14 = "For: AllDays"
        schedule_idf[-1].Field_15 = "Until: 24:00"
        schedule_idf[-1].Field_16 = "0"
        schedule_idf[-1].Field_13 = "Through: 12/31"
        schedule_idf[-1].Field_14 = "For: AllDays"
        schedule_idf[-1].Field_15 = "Until: 24:00"
        schedule_idf[-1].Field_16 = "1"
……
def chiller(idf_T1, COP, Source):
        Chiller = idf_T1.idfobjects['HVACTemplate:Plant:Chiller']
        #Chiller Template
        idf_T1.newidfobject('HVACTemplate:Plant:Chiller')
        Chiller[-1].Name =    'chiller1'
        Chiller[-1].Chiller_Type =    'ElectricCentrifugalChiller'
        Chiller[-1].Nominal_COP = COP
        if Source == 'heat pump':
                Chiller[-1].Condenser_Type = 'AirCooled'
        else:
                Chiller[-1].Condenser_Type = 'WaterCooled'
        Chiller[-1].Sizing_Factor = '1.1'
        return idf_T1


def tower(idf_T1):
        Tower = idf_T1.idfobjects["HVACTemplate:Plant:Tower"]
        #Tower Template
        idf_T1.newidfobject('HVACTemplate:Plant:Tower')
        Tower[-1].Name =    'Tower1'
        Tower[-1].Tower_Type =    'SingleSpeed'
        Tower[-1].Sizing_Factor = 1.1
        return idf_T1
```

同济大学

```
def fault(idf_T1, TPR):
    fault_tower = idf_T1.idfobjects['FaultModel:Fouling:CoolingTower']
    idf_T1.newidfobject('FaultModel:Fouling:CoolingTower')
    fault_tower[-1].Name = 'CTFouling'
    fault_tower[-1].Cooling_Tower_Object_Type = 'CoolingTower:SingleSpeed'
    fault_tower[-1].Cooling_Tower_Object_Name = 'Tower1'
    fault_tower[-1].Reference_UA_Reduction_Factor = TPR

    fault_air = idf_T1.idfobjects['FaultModel:Fouling:AirFilter']
    fault_coil = idf_T1.idfobjects['FaultModel:Fouling:Coil']
……
```

同济大学

附录 C 关键变量推测部分代码节选

```
#遗传算法部分
from dateutil.parser import parse
import pandas as pd
import numpy as np
import xgboost as xgb
from geneticalgorithm import geneticalgorithm as ga
from sklearn.preprocessing import MinMaxScaler,StandardScaler
import joblib
from sklearn.metrics import explained_variance_score, mean_absolute_error as MAE, mean_squared_error as MSE, r2_score
import copy, time, datetime
import matplotlib.pyplot as plt
import os
import math
def not_onecode(dataset):
    for i in list(dataset.index):
        if dataset["WS"][i] == "primary_constant":
            dataset.iloc[i,0] = 0
        elif dataset["WS"][i] == "primary_varying":
            dataset.iloc[i,0] = 1
        elif dataset["WS"][i] == "secondary_varying":
            dataset.iloc[i,0] = 2
        if dataset["Terminal"][i] == "FCU":
            dataset.iloc[i,1]    = 0
        elif dataset["Terminal"][i] == "CAV":
            dataset.iloc[i,1] = 1
        elif dataset["Terminal"][i] == "VAV":
            dataset.iloc[i,1] = 2
        if dataset["Source"][i] == "heat pump":
            dataset.iloc[i,2] = 0
```

同济大学

```
                elif dataset["Source"][i] == "boiler_chiller":
                    dataset.iloc[i,2] = 1
        dataset = dataset.apply(pd.to_numeric,errors = 'ignore')
        return dataset
    #把待推测的那一行变成逐时值
    def concat_feature(building, weather):
        feature_repeat = pd.DataFrame()
        for i in range(weather.shape[0]):
            feature_repeat = pd.concat([feature_repeat,building]).reset_index(drop = True)
        feature_repeat = pd.concat([feature_repeat,weather], axis = 1).reset_index(drop = True)
        return feature_repeat

    #归一化所有变量
    def para_scaler(samples_and_inference_df):
        samples_and_inference_df_scaler = copy.deepcopy(samples_and_inference_df)
        scaler_np = MinMaxScaler().fit_transform(samples_and_inference_df.iloc[:,3:-1])
        samples_and_inference_df_scaler.iloc[:,3:-1] = scaler_np[:,:]
        return samples_and_inference_df_scaler
    def get_predict_weather_fearture(predict_year, predict_weather):
        weather_feature = pd.DataFrame(columns = ['DryT','Hour','Month','Day'])
        weather_feature['DryT'] = predict_weather['温度（℃）']
        weather_feature['Hour'] = [predict_weather['time'][i].hour for i in predict_weather.index]
        weather_feature['Month'] = [predict_weather['time'][i].month for i in predict_weather.index]
        weather_feature['Day'] = [predict_weather['time'][i].day for i in predict_weather.index]
        return weather_feature
    #找到需要推断的变量
    def find_inference(param_df):
```

同济大学

```
        #param_df = pd.DataFrame([params])
        to_inference = param_df.columns[param_df.isna().any()].tolist()
        real_val = ['SPC', 'CR', 'OPD', 'INFIL', 'LPD', 'SHGC', 'ST', 'WWR', 'AREA',
'COR', 'TD', 'SATD', 'CHWT', 'COP']
        int_val = ['WS', 'Terminal', 'Source', 'NL']
        to_inference_type = []
        for i in to_inference:
            if i in real_val:
                to_inference_type.append('real')
            elif i in int_val:
                to_inference_type.append('int')
        return to_inference, to_inference_type
    def fake_data_generate(day_num):
        fake_date = []
        for i in range(day_num):
            for j in range(14):
                fake_date.append(i)
        return fake_date
    def fc(param_inference_iter):
        #将 param_inference 进行归一化, 并赋值
        for p in inference_param:
            if ('NL' == p):
                houly_inference_feature['NL'] = np.log1p
(param_inference_iter[inference_param.index('NL')])
            elif ( 'AREA' == p):
                houly_inference_feature['AREA'] = np.log1p
(param_inference_iter[inference_param.index('AREA')])
            else:
                houly_inference_feature[p] =
param_inference_iter[inference_param.index(p)]
        dtest = xgb.DMatrix(houly_inference_feature)
        predict_value = bst.predict(dtest)
        if granularity == 'daily':
            predict_value = pd.DataFrame(predict_value)
```

同济大学

```
            predict_value = daily(predict_value, -1, 153)
        obj = MAE(ground_truth_y, predict_value)
        return obj
    #%% main
    if __name__ == "__main__":
        #params:dic
        path = 'D:\graduate\dissertation\code\DataFusion\GA\hold_out'
        file_list = os.listdir(path)
        for file_name in file_list:
            #需要设置的参数
            season = 'cooling' #cooling, heating, transition
            granularity = 'daily'    #hourly; daily; monthly
            meter_data = False
            if meter_data == True:
                predict_year = 2015
                predict_weather = pd.read_excel(shanghai weather.xlsx')
                params_df = pd.read_excel(r'to_inference_field.xlsx')
            else:
                params_df = pd.read_excel(path + "\\" + file_name)
            #需要设置的参数
            IDs = params_df['ID']
            for i in params_df.columns:
                if params_df[i].dtype == 'O':
                    params_df[i] = pd.to_numeric(params_df[i],errors='coerce')
            bst = joblib.load(r" xgboost.dat") = joblib.load(r"
modify_simulation.dat")
            result_df = copy.deepcopy(params_df)
            obj = []
            #推测需要推测的模型中的参数
            for ID in IDs:
                print(ID)
                if meter_data == False:
                    ground_truth_y = pd.read_excel(r'GA\hold_out.xlsx',
index_col = 0)
```

同济大学

```
                ground_truth_y = ground_truth_y[ID].reset_index(drop =
True)
            else:
                ground_truth_y = pd.read_excel(r'\data_HVAC.xlsx',
index_col = 0)
                ground_truth_y = np.log1p(ground_truth_y.iloc[:, ID-
1801]*3600000)
            if season == 'cooling':
    = MinMaxScaler().fit_transform(ground_truth_y.values.reshape(-1,1))
                if granularity == 'daily':
                    if meter_data == True:
                        ground_truth_y =
ground_truth_y[120:273].reset_index(drop = True)
                    else:
                        ground_truth_y = daily(ground_truth_y,0,365)
            #把需要推测的 sample 的参数(已归一化)变成逐时值
            param_df = params_df[params_df['ID'] == ID]
            param_df.drop('ID', axis = 1, inplace = True)
            inference_param, inference_param_type =
find_inference(param_df)
            param_df['NL'] = param_df['NL'].apply(np.log1p)
            param_df['AREA'] = param_df['AREA'].apply(np.log1p)
            houly_inference_feature = get_hourly_inference_para(param_df,
ID)
            #遗传算法模型
            varbound= get_bounds(inference_param).values
            vartype= np.array([inference_param_type]).T
            algorithm_param = {'max_num_iteration': 800,\
                               'population_size':100,\
                               'mutation_probability':0.1,\
                               'elit_ratio': 0.01,\
                               'crossover_probability': 0.5,\
                               'parents_portion': 0.3,\
                               'crossover_type':'uniform',\
```

```
                    'max_iteration_without_improv':100}
            model=ga(function=fc,\
                dimension=len(inference_param),\
                variable_type_mixed=vartype,\
                variable_boundaries=varbound,\
                algorithm_parameters=algorithm_param)
            model.run()
            inference_result = model.output_dict['variable']
            obj.append(model.output_dict['function'])
            result_df.loc[result_df[result_df['ID']  ==  ID].index[0],
inference_param] = inference_result
        dif(result_df, file_name)
```

同济大学

附录 D 数据融合部分代码节选

```
import datetime
import joblib
import copy
import matplotlib.dates as mdate
from sklearn.linear_model import BayesianRidge, LinearRegression, ElasticNet, Ridge, Lasso, LassoCV
from sklearn import svm
from sklearn.model_selection import KFold
import lightgbm as lgb
import matplotlib.dates as mdates
from chinese_calendar import is_workday, is_holiday
import catboost
from catboost import CatBoostRegressor
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import ConstantKernel, RBF
sns.set_style('darkgrid', {'font.sans-serif':['SimHei', 'Arial']})#设置图表背景颜色字体
plt.rcParams['font.sans-serif']=['Microsoft YaHei']


def corr():
    features_with_max = pd.read_excel(r'D:\graduate\dissertation\code\DataFusion\ModifySimulationData\field_result_field_temp.xlsx', index_col = 0)
    features_with_max.drop('ID', axis = 1, inplace = True)
    features_with_max.corr()
    plt.matshow(features_with_max.corr(method = 'pearson'))
    plt.xlabel('features')
    plt.title('pearson')
    scale_ls = range(features_with_max.shape[1])
    plt.xticks(scale_ls,features_with_max.columns, rotation = 90)
```

同济大学

```
        plt.yticks(scale_ls,features_with_max.columns)
        plt.colorbar()

    def daytype(date):
        daytype = []
        for d in date:
            if is_workday(d):
                daytype.append(1)
            else:
                daytype.append(0)
        daytype_df = pd.DataFrame(daytype, index = date,columns = ['day_type'])
        return daytype_df

    def split_train_test(X_all, y_all, scaled = False, exID = '310105A050'):
        test_ID = exID
        test_order = ID_order_dic[test_ID]
        order_index = orders.index(test_order)
        if scaled == True:
            test_order =
MinMaxScaler().fit_transform(np.array(list(ID_order_dic.values()))).reshape(-
1,1))[order_index][0]
        X_train = X_all[round(X_all.ID,4) != round(test_order,4)]
        X_test = X_all[round(X_all.ID,4) == round(test_order,4)]
        y_train = y_all.drop(test_ID,axis = 1).values.reshape(-1,1)
        y_test = y_all[test_ID].values.reshape(-1,1)
        return X_train, X_test, y_train, y_test

    def feature_scaling(features):
        features_scaling = pd.DataFrame(MinMaxScaler().fit_transform(features),
columns = features.columns)
        features_scaling['simulation_value'] = features['simulation_value'].values *
1000
        features_scaling['AREA'] = features['AREA'].values
        features_scaling['NL'] = features['NL'].values
```

同济大学

```
        return features_scaling


    #linear models
    def linear_model(X_train, y_train, X_test, y_test, method):
        Id = X_test['ID'].values[0]
        X_train = X_train.drop(['ID','Month','Day'],axis = 1)
        X_test = X_test.drop(['ID','Month','Day'],axis = 1)
        if method == 'Lasso':
            clf = LassoCV(eps=0.001
                          ,n_alphas=2000
                          ,cv=5 #交叉验证的折数
                          ).fit(X_train, y_train)
            predict = clf.predict(X_test)
            modify_result_visualize(target_test,predict, method = "Lasso" + str(Id),
is_save = False)
            mae = MAE(target_test,predict)
        if method == 'Ridge':
            clf = BayesianRidge().fit(X_train, y_train)
            predict = clf.predict(X_test)
            modify_result_visualize(target_test,predict, method = "BaysianRidge"
+ str(Id), is_save = False)
            mae = MAE(target_test,predict)
        return clf, predict, mae


    def cv_rmse(y,yhat):
        return np.sqrt(MSE(y,yhat))/np.mean(y)


    def XGB(features_train,features_test, target_train, target_test):
        dtrain = xgb.DMatrix(features_train,target_train)
        dtest = xgb.DMatrix(features_test)
        #dfull = xgb.DMatrix(features,target)
        #tune parameters
        axisx =    np.arange(0.1, 1, 0.1)
        rs = pd.DataFrame()
```

```
var = pd.DataFrame()
ge = pd.DataFrame()
for i in axisx:
    num_round = 500
    t1 = datetime.datetime.now()
    print("*******number : "+str(i) +"********")
    params = {'verbosity':0 #并非默认
              ,'objective':'reg:squarederror' #并非默认
              ,"subsample":1
              #,"max_depth":6
              ,"max_depth":21
              ,"eta":0.6
              #,"gamma":0
               ,"gamma":0
              ,"lambda":1.2
              ,"alpha":0
              ,"colsample_bytree":0.4
              ,"colsample_bylevel":0.4
              ,"colsample_bynode":0.7
              #,"nfold":5
              }
    #bst = xgb.train(params, dtrain, i)
    cvresult = xgb.cv(params, dtrain, num_round,5, metrics = "mae")
    print(cvresult.mean())
    print(datetime.datetime.now() - t1)
```

同济大学

致谢

硕士三年的时光，想起来很长，过起来很短。

大四毕设刚来到同济的时候是一个寒冷的冬天，对嘉定的第一印象是阴沉的天，枯槁的树，刺骨的风。以为会很难适应研究生的生活，但好在来到了一个温馨可爱又富有活力的课题组，如今写下这段文字的时候已是阳春三月，温暖和煦的吹风，含苞待放的花朵，还有充满感激和期冀的我。

最要感谢的是导师——许鹏教授，故事的开始是一次电话面试，当时就被许老师贴近生活又充满想象力的面试题目圈粉，后来非常有幸进入了 A434 课题组，跟着许老师和优秀的同门师兄弟姐妹一起学习。一个一个的科研任务、一次一次不同主题的头脑风暴和组会讨论，让我不断地成长和学习，拓宽了视野并加深了对世界与知识的好奇心。如何对复杂任务进行拆解以及要保持对事物的好奇心是我将终生学习和受益的两点。非常感谢许老师三年来的教导与鼓励，除了在学术之外，许老师还注重培养思维和表达方面的能力，并且独创的“沙戴李”理论和早上九点的打卡真的让我养成了早睡早起的好习惯！

其次，要感谢的是课题组的各位同学，谢谢非常照顾我的师兄师姐师弟师妹和同门。课题组的每个同学也都有其闪光之处，每次组会上大家都有新奇的观点和想法，三人行必有我师焉。科研之余的桌游局、电影局、K 歌局以及紧张刺激的吃饭局（抢菜局）带来了很多的欢声笑语，将会是我一生的记忆。

再次，要感谢远程陪伴我的朋友们，能够做我的树洞，分享喜悦和悲伤，并时常在低谷时鼓励我。

还要感谢的是我亲爱的家人，感谢我的父母一直以来对我的爱和支持。

感谢提供数据的腾天节能技术有限公司和其他单位，感谢帮助过我的朋友和同学们，感谢遇到的每一个老师和他们辛苦的教学与指导。

人生就像一段长长的火车旅程，一路上总有人要上车下车，不管遇到什么风景我们总是要奔赴下一程，谢谢在同济的点点滴滴，好的坏的都将是美好的回忆。

2022 年 3 月

同济大学

个人简历、在读期间发表的学术论文与研究成果

个人简历：

郭明月，女，1997 年 11 月生。

2019 年 6 月毕业于重庆大学 建筑环境与能源应用工程专业 获学士学位。

2019 年 9 月入同济大学读硕士研究生。

已发表论文：

[1] **Mingyue Guo,** Peng Xu, Tong Xiao, Ruikai He, Mingkun Dai, Shelly L. Miller. Review and comparison of HVAC operation guidelines in different countries during the COVID-19 pandemic. Building and Environment 2021;187:107368. **(SCI, ESI 高被引论文)**

[2] Yongbao Chen, **Mingyue Guo,** Zhisen Chen, Zhe Chen, Ying Ji. Physical energy and data-driven models in building energy prediction: A review. Energy Reports, 2022, 8: 2656–2671. **(SCI)**

[3]**郭明月**, 许鹏, 肖桐, 何睿凯, 戴明坤. 应对新型冠状病毒国内外暖通相关指南对比.暖通空调,2020,50(11):13-20.

[4]王鸿鑫, 许鹏, **郭明月**, 顾洁帆, 肖桐, 陈喆, 杨一昆. 基于 BIM 的能耗模拟软件功能的测评分析.建设科技,2019(16):15-19.

已公开发明专利：

[1] 许鹏,**郭明月**,何睿凯,陈志森,陈喆,陈永保.一种管道井寻优算法[P]. 上海市：CN112241563A,2021-01-19.

软件著作权：

[1] **郭明月**,许鹏,王鸿鑫,顾洁帆.建筑信息模型向建筑能耗模型转换软件

研究及项目经历：

[1] 2019 年 9 月~2021 年 6 月：基于 BIM 的绿色建筑运营优化关键技术研发 (项目编号：2018YFC0705900)，国家重点研发项目(十三五)，主要参与人

[2] 2020 年 6 月~2021 年 6 月：暖通空调设计自动化研究（一期），横向课题，参与人

在校期间获得奖励：

2020 年 11 月　　亚大奖学金三等奖

2021 年 11 月　　硕士研究生国家奖学金

2021 年 11 月　　同济大学优秀学生

同济大学

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《基于多源异构数据的办公建筑能耗预测方法》，是本人在导师指导下，独立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：郭明月

日期：2022 年 3 月 21 日

同济大学

同济大学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；允许论文被查阅和借阅。学校有权将本学位论文的全部或部分内容授权编入有关数据库出版传播，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于（在以下方框内打“√”）：

☐ 保密，在_____年解密后适用本授权书。

☑ 不保密。

学位论文作者签名：郭明月　　指导教师签名：许鹏

日期：2022 年 3 月 21 日　　日期：2022 年 3 月 22 日

七、学位论文答辩委员会决议

姓　　名	郭明月	学　　号	1930255	所在学科/专业	供热供燃气通风及空调工程
指导教师	许鹏	答辩日期	2022年3月16日	答辩地点	腾讯会议
论文题目	基于多源异构数据的办公建筑能耗预测方法				

郭明月同学的硕士学位论文《基于多源异构数据的办公建筑能耗预测方法》研究了一种基于多种来源的数据预测办公室建筑能耗的方法。

该方法提取了影响办公建筑能耗的关键变量和缺失变量；分析了来源于能耗监测平台、节能审计报告、快速模拟工具等三种途径的数据的特性，构建了多源异构数据库；基于办公建筑能耗预测混合模型，建立了一种不依赖历史数据的能耗预测方法。该方法在实际建筑中进行验证，效果较好。此项研究成果对工程实践具有一定的参考意义。

论文目标明确，框架完整，思路清晰，模型及论证合理，结论可信，体现了作者具有较扎实的理论基础和专业知识，具备独立从事科学研究工作和解决实际工程问题的能力。

郭明月同学答辩过程阐述清晰，思路明确，能正确回答答辩委员提出的问题，论文达到了硕士学位论文水平。经无记名投票表决，五位答辩委员中，五位同意建议授予其工学硕士学位，并推荐其申请同济大学优秀硕士学位论文。

答辩委员会主席签名：于航　　2022 年 3 月 16 日

答辩委员会共 5 人，经表决，5 人建议授予申请人硕士学位。根据《同济大学学位授予工作细则》[注]（在□内划“√”）：

☑建议授予申请人硕士学位。
□建议不授予申请人硕士学位。
□申请人可在一年内修改论文，申请重新答辩一次。

☑推荐为同济大学优秀硕士学位论文。

答辩委员会成员签名	职　务	姓　名	职　称	单　位	签　名
	主　席	于航	教授	同济大学	于航
	委　员	许鹏	教授	同济大学	许鹏
		潘家英群	教授	同济大学	潘家英群
		苏醒	副教授	同济大学	苏醒
		李铮伟	副教授	同济大学	李铮伟
	秘　书	叶蔚	副教授	同济大学	叶蔚

注：根据《同济大学学位授予工作细则》第十一条规定：

1. 申请人获得全体答辩委员会成员三分之二以上（含）同意票，为建议授予申请人硕士学位；

2. 申请人获得全体答辩委员会成员二分之一以上（含）、三分之二以下（不含）同意票，申请人可在一年内修改论文，申请重新答辩一次；

3. 申请人获得全体答辩委员会成员二分之一以下（不含）同意票，为建议不授予申请人硕士学位。

同济大学