同济大学
TONGJI UNIVERSITY

硕士学位论文

（学术学位）

基于数据驱动的公共建筑综合体能耗预测方法

姓　　名：夏壮

学　　号：2130277

学　　院：机械与能源工程学院

学科门类：工学

一级学科：土木工程

二级学科：供热、供燃气、通风及空调工程

研究方向：建筑节能技术

指导教师：许鹏教授

联合培养单位：

二〇二四年三月

同济大学
TONGJI UNIVERSITY

A thesis submitted to

Tongji University in partial fulfillment of the requirements for

the degree of Master of Engineering

Data-Driven Method for Predicting Energy Consumption in Non-Residential Building Complexes

Candidate: Xia Zhuang

Student Number: 2130277

School/Department: School of Mechanical Engineering

Categories: Engineering

First-level Discipline/Degree: Civil Engineering

Second-level Discipline/Degree's Field: Heat, Gas Supply, Ventilation and Air-conditioning Engineering

Research Fields: Building Energy Saving Technology

Supervisor: Prof. Xu Peng

March 2024

摘要

目前正处在2030年“碳达峰”、2060年“碳中和”的重要阶段，高效、准确的能耗预测是能源管理和建筑节能研究的重要课题。目前建筑能耗预测主要有两种主流方法，分别是物理模型和数据驱动模型。数据驱动模型由于其高效灵活的建模过程，近年来越来越受研究者青睐。但是目前数据驱动的建筑能耗预测模型只针对某一特定单体建筑，所建立的预测模型难以迁移到其他建筑上，其方法对于新建建筑或没有历史能耗的建筑不可用。其次数据驱动模型往往需要大量的训练数据以保证较强的泛化能力，但在实际应用中往往存在数据不足的问题。并且目前能耗预测所研究的建筑领域以办公、商业和住宅等功能单一建筑为主，针对综合体这类多功能建筑的研究尚缺乏。为此，本研究提出了一种基于迁移学习的建筑综合体能耗预测方法。研究内容主要包括以下几个方面：

首先，本文通过敏感性分析提取了影响建筑能耗的关键变量。搜集了建筑负荷相关和系统相关的两类初始变量集，使用相关系数法（SRRC、PRCC）和随机森林两种方法进行敏感性分析，提取了对建筑能耗有显著影响的关键变量。

其次，本文提出了基于遗传算法和基于自编码器的关键变量推断算法来解决实际预测时部分建筑关键变量缺失的问题。对于有历史能耗的建筑，充分挖掘能耗和关键变量之间的关联关系，使用遗传算法推断最满足该关联关系的缺失变量值；对于没有历史能耗的建筑，通过降噪自编码器挖掘关键变量内部的关联关系，基于所学习到的关联关系推断缺失值。

再次，本文提出了基于条件生成对抗网络（CGAN）的建筑能耗数据增强方法，该方法结合了CGAN和XGBoost。一方面，为了保证生成器的生成效果，通过聚类归一化后的能耗数据，为每类数据训练专一的生成对抗网络模型，其对应生成器生成建筑全年能耗趋势曲线。另一方面，建立XGBoost全年总能耗预测模型，使用预测的全年总能耗对建筑全年能耗趋势曲线进行填充，获得最终的建筑全年能耗数据。本研究利用该方法生成了包含6000栋建筑的增强数据集。

最后，本文提出了基于迁移学习的建筑全年能耗预测方法。真实数据不足的情况下，传统模型往往会出现精度差、过拟合等问题。本研究通过迁移学习融合增强数据和真实数据，将增强数据集作为源域，真实数据集作为目标域，在增强数据集上建立LSTM预训练模型，将该预训练模型的结构迁移至真实数据集，用真实数据集微调模型参数。本文对比了迁移模型和仅在真实数据集直接训练的基础模型，结果显示，在仅拥有少量真实数据时，基于迁移学习所建立的预测模型

通过来自源域的知识迁移，预测精度显著优于未迁移模型，预测误差降低了 9.0%，并且还有效避免了过拟合。此外，逐月误差比对显示，迁移模型相比未迁移模型性能的提升主要来源于过渡季。

本课题提出的公共建筑综合体能耗预测方法，以高效准确的数据增强技术和基于迁移学习的数据融合方法缓解了跨建筑能耗预测真实数据不足的现状，弥补了当前综合体类跨建筑能耗预测领域研究的缺乏，其方法对于其他多功能建筑能耗预测的研究有一定的借鉴意义。

关键词：跨建筑能耗预测，数据驱动模型，机器学习，迁移学习，数据增强

Abstract

Currently, we are in a critical phase of achieving the goals of "peak carbon dioxide emissions by 2030" and "carbon neutrality by 2060." Efficient and accurate energy consumption prediction is a crucial issue in energy management and building energy conservation. At present, there are two mainstream methods for building energy consumption prediction: white-box models and data-driven models. The data-driven modeling process is more efficient and has become increasingly popular in recent years. However, current data-driven building energy consumption prediction models are only applicable to individual buildings, and the prediction models established are difficult to transfer to other buildings. These methods are not applicable to new buildings or buildings without historical energy consumption data. Additionally, data-driven models often require a large amount of training data to ensure strong generalization ability, but there is often a problem of insufficient data in practical applications. Furthermore, the current research on energy consumption prediction primarily focuses on office, commercial, and residential buildings, with a lack of studies on multifunctional buildings such as complexes. Therefore, this study proposes a building complex energy consumption prediction method based on transfer learning. The research mainly includes the following aspects:

Firstly, this paper extracted key variables affecting building energy consumption through sensitivity analysis. Two types of initial variable sets related to building load and system were collected. Sensitivity analysis was conducted using two methods: correlation coefficient-based methods (SRRC, PRCC) and random forest, to extract key variables significantly affecting building energy consumption.

Secondly, this paper proposes two algorithms, based on genetic algorithm and autoencoder, respectively, to infer key variables when some of them are missing during practical prediction. For buildings with historical energy consumption data, the genetic algorithm is employed to infer the missing variable values that best satisfy the correlation with energy consumption and key variables. For buildings without historical energy consumption data, a denoising autoencoder is utilized to uncover the internal correlations among key variables, and missing values are inferred based on the learned correlations.

Furthermore, an architecture energy consumption data augmentation method

based on Conditional Generative Adversarial Networks (CGAN) combined with XGBoost was proposed. On one hand, to ensure the effectiveness of the generator, energy consumption data is normalized through clustering, and a dedicated CGAN model is trained for each cluster to generate the annual energy consumption trend curve for each category. On the other hand, an XGBoost model is established for predicting the total annual energy consumption, and the predicted total annual energy consumption is used to fill in the annual energy consumption trend curve for buildings to obtain the final annual energy consumption data. Using this method, this study generated an augmented dataset containing 6000 buildings.

Finally, a method for predicting annual energy consumption in buildings based on transfer learning was proposed. Traditional models often suffer from issues like poor accuracy and overfitting when faced with insufficient real data. This study addresses this challenge by integrating augmented data with real data using transfer learning. The augmented dataset is utilized as the source domain, while the real dataset serves as the target domain. We establish an LSTM pre-trained model on the augmented dataset and transfer its structure to the real dataset, fine-tuning the model parameters using the real dataset. A comparison between the transfer model and the baseline model trained solely on real data is conducted. The results demonstrate that when only a limited amount of real data is available, the prediction model established based on transfer learning significantly outperforms the non-transfer model, with a 9.0% reduction in prediction error, while effectively avoiding overfitting. Furthermore, monthly error comparison indicates that the improvement in performance of the transfer model over the base model predominantly emanates from the transition season.

The proposed method for predicting energy consumption in public building complexes addresses the scarcity of real data in cross-building energy consumption prediction through efficient and accurate data augmentation techniques and a transfer learning-based data fusion approach. This method fills the gap in current research on cross-building energy consumption prediction in the field of complex buildings and holds certain reference value for the study of energy consumption prediction in other multifunctional buildings.

Key words: cross-building energy prediction, data-driven model, machine learning, transfer learning, data augmentation

目录

第一章 绪论

1.1 研究背景

1.1.1 研究背景

建筑业是现阶段我国最大的能源消耗行业之一。中国建筑节能协会能耗专委会发布的《中国建筑能耗研究报告 2022》指出，2020 年全国建筑全过程能耗总量占全国能源消费总量的 45.5%，其中建筑运行阶段能耗占比高达 46.8%[1]。在建筑能耗中，暖通空调系统占有很大的比重，具有很可观的节能潜力，在美国，这部分能耗约占建筑能耗的 50%[2]。距离 2030 年实现碳中和的目标还剩 6 年，建筑节能越来越成为一个关键的问题。自 2007 年住建部、财政部发布了《关于加强国家机关和大型公共建筑节能管理工作的实施意见》后，公共建筑能耗监测平台建设工作在全国迅速展开，截至目前已经有超过 33 个省市开展了公共建筑能耗监测平台建设工作，其中北京市、上海市、重庆市、天津市、深圳市、江苏省、山东省和安徽省等 8 个省市公共建筑能耗监测平台通过率验收，全国实施能耗监测建筑数量已经超过一万栋[3]。至 2021 年 12 月 31 日，上海市累计有 2143 栋公共建筑完成用能分项计量装置的安装并实现与能耗监测平台的数据联网，其中大型公共建筑 1933 栋[4]。关于建筑的能耗数据积累初具规模，如何基于大量能耗监测数据，挖掘出其中的隐藏信息，改善建筑用能情况成为目前需要解决的问题。

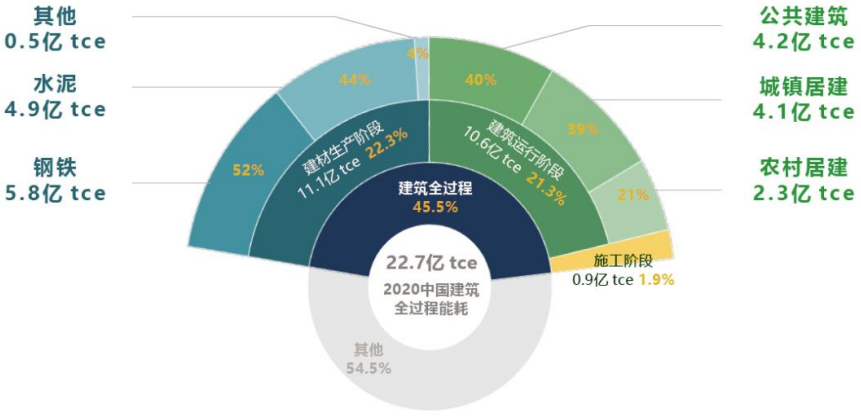


图 1.1 2020 年全国建筑全过程能耗占比情况

建筑能耗预测模型可以根据其数学原理主要分为两类：基于物理意义的白箱模型和数据驱动的黑箱模型 [5]。白箱模型根据热力学规律对目标进行详细的能量建模，可解释性较好，但建模过程复杂，需要专业领域知识，且建立白箱模型需

要大量建筑设计及运行参数的详细输入，不能提供准确的输入会导致模型预测性能较差。其次，白箱模型基于物理原理，模型本身存在着大量简化，且未考虑施工以及实际使用过程中带来的误差。而黑箱模型建模不需要非常专业的领域知识，建模过程更高效灵活[6]，且模型预测性能较好，近年来越来越受欢迎。

1.1.2 研究意义

建筑能耗在全国能源消费总量中占据相当大的比例，要实现“碳中和”、“碳达峰”的目标，建筑节能势在必行。建筑能耗预测是实现建筑节能的有效手段之一。在建筑设计阶段，精准的能耗预测对于合理配置能源设备至关重要。在建筑运行和改造阶段，能耗预测可作为设计和选择适当节能方法的工具。同时，短期多步能源预测可整合到基于模型预测的建筑系统运行控制中，通过预先优化系统运行方案，实现峰时调控或降低运行能耗。

数据驱动的建筑能耗预测模型由于灵活高效的建模过程，其优化问题在建筑能耗预测研究领域一直是热点。跨建筑的建筑全年能耗预测模型的建立，极大地提高了模型的适用性，使得预测模型不再仅适用于某一栋建筑而是适用于某一种建筑类型。基于迁移学习的跨建筑能耗预测模型则解决了模型训练真实数据不足的问题，大大提高了在少数据量下模型的预测精度。

1.2 研究现状

1.2.1 建筑能耗预测模型

数据驱动的能耗预测模型的目的是在一段时间内，学习历史数据与预测数据之间的映射关系，属于时间序列预测问题。时间序列预测主要有三种主流方法[7]，分别是传统时间序列预测模型、传统回归模型和深度学习模型。

1、时间序列模型

传统的时间序列模型用于识别能耗数据与其历史值之间的相关性，常见的时间序列模型有自回归模型(AR)、移动平均模型(MA)、自回归移动平均模型(ARMA)、自回归综合移动平均模型(ARIMA)和季节性自回归综合移动平均法（SARIMA）。Alduailij 等[8]评估了线性回归、动态回归、ARIMA 时间序列、指数平滑时间序列、人工神经网络和深度神经网络在建筑用电量的日前预测中的性能，结果显示 ARIMA 优于其他几种方法。Ediger 等[9]使用了 ARIMA 和 SARIMA 估算土耳其 2005 年至 2020 年的未来一次能源需求量。Abdelaal 等[10]基于 5 年的逐月用电数据建立了自回归综合移动平均模型，与之前根据相同数据开发的回归和归纳网络机器学习模型相比，ARIMA 模型所需的数据更少、系数更小、精度更高。Zhou 等[11]使用时间序列分析建立办公建筑能耗预测模型，该模型精度

较高。

尽管传统的时间序列模型在数据量较少的短期预测上表现良好，但由于其在处理非线性和高维问题上的缺点，这种模型很少能满足长期预测的需求[7]。当历史数据规模较大时，ARIMA 的计算成本非常高。近年来，研究者倾向于将传统时间序列模型与其他算法结合使用，以提高预测精度，而不再将其作为单独建立预测模型的首选方法。Liang 等[12]提出了一种将深度集成(DE)模型和自回归(AR)模型相结合的混合预测模型，将建筑能耗分解为全局部分和局部部分，使用 DE 模型预测全局部分，ARIMA 模型预测局部部分，该混合模型优于 ARIMA 模型、DE 模型和 LSTM 模型。Chou 等[13]提出了一种结合神经网络和 ARIMA 的混合模型用于电量预测。Guo 等[14]提出了 ARIMA-SVR 混合模型用于办公建筑的电量预测。

2、传统机器学习模型

如果我们认为数据是数据驱动模型的燃料，那么机器学习就是强大的引擎。机器学习位于计算机科学和统计学的交叉点，服务于人工智能(AI)和数据科学的核心，是发展最快的数据驱动技术领域之一[15]。传统回归模型、树模型和神经网络等模型近年来成为建筑负荷能耗预测的主流方法。其中回归模型包括岭回归，Lasso 回归，贝叶斯岭回归，支持向量回归等；树模型包括决策树、随机森林、XGBoost 等；神经网络模型包括多层感知机（MLP），ANN 等。Lim[16]等在多元回归分析的基础上，提出了考虑时滞现象的单位面积建筑冷负荷预测模型。Li[17]等建立了基于 SVM 的建筑小时冷负荷预测模型，其预测性能优于传统的反向传播神经网络。Ding 等[18]使用支持向量回归预测办公建筑短期能耗，并利用遗传算法对 SVR 参数进行优化，对于超短期能耗预测，建立了 GA-WD-SVR 预测模型。Ahmad 等[19]对比了随机森林和人工神经网络在预测酒店建筑逐时能耗上的预测性能，结果显示人工神经网络的表现略好于随机森林，均方根误差(RMSE)分别为 4.97 和 6.10。Sha 等[20]提出了一种简化的空调系统能耗预测模型，该模型仅需三个输入：度日、日类型和月份，对比了 SVR、ANN 和 MLR 的表现，其中 SVR 预测精度更高，该预测模型精度满足工程要求。

单一的机器学习算法面临着模型适应性的挑战[21]，随着建筑累积的历史数据的增多以及所考虑的预测目标影响因素的增多，预测变得更加复杂，相较于单一算法，将多种机器学习模型的预测结果相结合的集成学习的优势开始显露出来。Robinson 等[22]仅使用五个特征（主要建筑活动、楼层数、建筑面积、人员数量、制热天数、制冷天数）建立了梯度增强回归模型用于商业建筑能耗预测。Dong 等[23]利用决策树挖掘能源消费模式，采用集成学习方法建立各模式的能耗预测模型，集成学习方法显著优于支持向量回归和神经网络。Zhao 等[24]将时间特征和

气候特征分开分析，使用 LightGBM 建立建筑冷负荷预测模型，所提出的方法能有效提高建筑物冷负荷的预测精度。Wang 等[25]使用穷举法获得了回归树(RT)、支持向量回归(SVR)、多元线性回归(MLR)、集成 Bagging 树(EBT)、反向传播神经网络(BPNN)和通用回归神经网络(GRNN)的所有可能的模型组合，实验结果表明，包含所有备选基本模型的异构集成学习模型并不能保证最准确的预测性能。由任意基础模型组合而成的异构集成学习模型并不一定能提高预测精度。

3、深度学习模型

近年来随着深度学习的快速发展，深度神经网络模型开始用于建筑能耗预测领域。深度学习算法的结构比传统的预测算法更加复杂，因此，他们可以了解到更复杂的建筑能源负荷与相关变量之间的关系，这在某些情况下带来了更高的预测精度。Kim 和 Cho[26]提出 CNN-LSTM 网络模型，其中 CNN 层进行特征提取，LSTM 层用于建模时间序列分量中不规则趋势的时间信息，与传统的家庭用电量数据集预测方法相比，所提出的模型误差更小。Wang 等[27]分析了 7 个浅层学习、2 个深度学习和 3 个启发式算法在建筑热负荷预测上的性能，其中 LSTM 在短期负荷预测表现较好，XGBoost 在长期预测表现较好。Sala-Cardoso 等[28]建立了活动感知的 HVAC 电力需求预测，包括由 RNN 建立的占用预测模型和由自适应神经模糊推理系统建立的用电功率需求模型。Zhang 等[29]提出了一种基于长短期记忆网络和人工神经网络的混合预测方法，该混合模型在 1 小时前冷负荷预测中具有较高的预测精度。Lei 等[30]利用粗糙集理论提取影响建筑能耗的关键因素，并且对比了深度神经网络与反向传播神经网络、Elman 神经网络和模糊神经网络的预测结果，结果表明，粗糙集与深度神经网络相结合的方法准确率最高。

由于深层神经网络模型往往有大量超参数，超参数对模型性能至关重要，许多学者使用优化算法来进行超参优化，极大地提高了预测模型的性能。Qin 等[31]针对建筑能耗时间序列数据的非线性和非光滑特性，提出了一种结合变分模态分解(VMD)、模拟退火(SA)算法和深度信念网络(DBN)的短期混合建筑能耗预测模型。Yuan 等[32]采用布谷鸟搜索(cuckoo search, CS)算法对 WNN 的自由参数进行最优调整。Bui 等[33]将基于电磁学的萤火虫算法集成到人工神经网络中，通过优化初始参数集来提高其性能。

1.2.2 数据增强及数据融合

本质上，数据驱动模型主要依靠历史数据来训练参数，从而预测未来的数据。为了保证较强的泛化能力，通常要求训练数据驱动模型的样本数据具有全局学习空间的代表性。因此，实现数据驱动模型的二个基本假设是：一是训练数据应满足全局学习空间中独立且同分布的采样；二是必须有足够的训练样本来学习一个好的模型[34]。此外，随着预测模型的复杂度提高，训练一个深度学习模型所需要

的数据量也大大增加。然而在工程实践中，获得的样本数据往往不足以训练具有较强泛化能力的数据驱动模型。在这种情境下，采用数据增强策略成为一种可行的选择。数据增强是利用有限的训练数据实现数据驱动建模的有效方法之一[35]，主要用于解决机器学习和深度学习算法在有限数据集上的欠训练问题。这种方法已经成功应用于图像分类[36,37]、人脸识别[38]和自然语言处理[39]等领域。目前数据增强的手段主要有两种，第一种是通过对目标应用几何变换:平移、旋转、裁剪、翻转、缩放等，这种方法常用于图像分类和人脸识别。第二种方法是在现有的训练数据中加入噪声。

近年来，各种数据增强技术开始引入建筑领域[40][41]，部分学者尝试使用数据增强策略来缓解建筑能耗预测模型训练数据不足的困境。建筑能耗预测领域的数据增强主流方法主要有两种，一是融合模拟数据或者相似建筑真实数据进行增强。Lu 等[42]将基于 TRANSYS 建立目标建筑的基本仿真模型，并使用实测数据对该仿真模型进行标定。标定后的仿真模型作为数据增强源来对建筑真实能耗数据进行补充，所提出的数据增强策略显著提升了预测模型的精度。Amasyali 和 EIGohary[43]提出了一种混合机器学习方法，在模拟数据建立 GPR 模型学习天气因素，在真实数据集建立 SVR 模型学习人行为影响，再将前两个模型的输出作为集成模型的输入预测建筑制冷能耗，预测结果表明，所提出的混合预测方法优于传统的预测方法。Fang 等[44]通过建立 EnergyPlus 仿真模型通过改变人员占用、照明和设备强度来模拟边界情景对真实建筑能耗数据进行补充，显著提高了模型的鲁棒性和可靠性。

另一类是使用生成式模型学习真实数据分布来进行数据增强。Fan 等人[45]设计了两种条件变分自编码器对短期建筑能耗预测模型的建立进行数据增强，通过对 52 栋建筑物进行评估表明，条件变分自编码器能够生成高质量的合成数据样本，有助于提高短期建筑能耗预测的准确性。CV-RMSE 的平均性能提升率在 12%到 18%之间。Tian 等[46]利用生成对抗网络生成人工数据，使用原始数据和人工数据的混合数据集训练预测模型。基于两栋真实建筑对该数据增强方案进行验证，结果显示并行数据与原始数据具有相似的分布，混合数据训练的预测模型比仅使用原始数据训练的预测模型性能更好。

使用数据增强技术生成增强数据集后，高性能的数据融合方法也是数据增强策略成功应用的保证 [42]。数据融合是指将来自不同来源的数据集合、整合、分析，以提取更全面、准确的信息。数据融合可以提高模型的准确性和可靠性，减少模型的误差。常见的数据融合方法包括特征级融合、样本级融合和模型级融合。其中特征级融合是指将来自不同来源的特征进行整合；样本级融合是指将来自不同来源的样本进行整合；模型级融合是指将来自不同来源的模型进行整合。在获得

数据增强源后，高性能的数据融合方法也是数据增强策略成功应用的保证。在建筑能耗预测领域最常用的数据融合是样本级融合和模型级融合。

直接数据融合（属于样本级融合）是一种简单的方法，它是将增强后的数据直接添加到目标任务的不足训练集中[44,45,46]。然而，当增强数据与实际数据的分布差异较大时，直接数据融合往往性能不佳。基于迁移学习的数据融合比直接数据融合更有效[42]。近年来，迁移学习被用于实现增强数据与实际数据的融合[47]。迁移学习的主要思想不是直接使用增强数据，而是将从增强数据中学习到的规则或知识转移到目标建筑物中，以便在数据有限的情况下促进目标建筑物的学习过程。Ahn 和 Kim[48]提出了 TL-LSTM 的迁移学习框架，通过迁移学习融合模拟数据和实测数据，预测一栋办公建筑能耗，仅需要 24 h 的实测数据用于训练。Fang[49]等提出一个 Sim2Real 迁移学习框架，使用模拟数据集进行建筑能耗预测，实验案例表明，与基线模型相比，采用合适的迁移学习模型，所提出的 Sim2Real 迁移学习框架可以将 MAPE 1.06%到 29.05%。

迁移学习通常是指利用数据、任务、或模型之间的相似性，将在旧领域学习过的模型，应用于新领域的一种学习过程[50]。迁移学习放松了传统深度学习中的两个基本假设，即训练数据和测试数据可以有不同的分布，即使数据量很少也可以很好地完成任务[51]。迁移学习按照学习方法[52]可以分为基于样本的迁移学习、基于特征的迁移学习、基于模型的迁移学习和基于关系的迁移学习。基于样本的迁移学习就是指对不同的样本赋予不同的权重；基于特征的迁移学习是指寻找源域和目标域之间的共同特征空间；基于模型的迁移学习就是构建参数共享的模型；基于关系的迁移学习指在源域和目标域之间迁移关系知识，挖掘和利用关系进行类比迁移[82]。Ribeiro 等[53]提出了一种全新的迁移学习方法 Hephaestus，该方法考虑了域内的季节趋势。案例研究表明，所提出的方法可以将一所学校的能源预测性能提高 11.2%。Fang 等[54]提出了一种使用少量标记数据的建筑物预测模型，利用 LSTM 进行特征提取，DNN 通过域自适应来找到源建筑和目标建筑之间的域不变特征.。Fan 等[55]提出了 CNN-LSTM 的深度学习模型，结合迁移学习方法进行预训练和微调，将预训练模型中提取的知识分别用于特征提取和权值初始化。结果验证了迁移学习对短期负荷预测的重要性。Lu 等[56]提出了一种基于长短期记忆(LSTM)和多核最大平均差(MK-MMD)域自适应的多源迁移学习能量预测模型，采用动态时间规整(DTW)对源域进行选择。

1.3 研究内容

目前关于数据驱动的建筑能耗预测主要存在以下几个问题：

一是目前绝大部分数据驱动的能耗预测只针对某一单体建筑，依赖于来自同

一建筑的足够的历史数据来训练模型，所建立的预测模型难以迁移到其他建筑上。其方法对于新建建筑或没有历史能耗的建筑不可用。

二是目前建筑能耗预测研究的对象大多是功能单一的建筑（如办公建筑），很少有关于建筑综合体能耗预测的研究。建筑综合体一般包括办公区和商业区，根据功能特点分别采用不同的空调系统。多系统导致影响建筑能耗的参数成倍增加，如何优化特征工程，选择合适的关键变量成为提高模型精度、避免维度灾难的难点之一。

三是数据驱动模型往往需要大量的训练数据以保证较强的泛化能力，但在实际应用中往往存在数据不足的问题，这将严重影响数据驱动模型的预测性能。虽然目前响应国家号召，能耗监测平台建设工作在有序展开，但仍然有大量建筑未纳入能耗监测范畴，并且大部分建筑的能耗监测数据未得到很好的利用。理论上，模型训练数据通常要求样本数据能够代表全局学习空间，因此如果没有长期积累的涵盖建筑全部状态的能耗数据，建立的模型性能往往不佳。在建立针对某一类型建筑的跨建筑能耗预测模型时需要大量同类型建筑能耗数据集用于模型的训练，这对数据量的要求更大，这在现实中往往无法实现。

针对上述内容，本研究主要解决以下 3 个问题：

1) 综合体存在多套系统，多系统导致预测模型输入更加复杂，如何选择合适的关键变量提高模型精度、避免维度灾难？
2) 在部分关键建筑信息缺失的情况下如何进行能耗预测？
3) 跨建筑能耗预测模型的训练对数据量要求较高，如何解决真实数据不足的问题？

本课题在课题组已有研究的基础上，着眼于建筑综合体的全年能耗预测。本文技术路线如图 1.2 所示。

第二章首先搜集影响建筑能耗的初始变量集，主要包括建筑负荷相关变量和系统相关变量。接着通过敏感性分析获取影响建筑能耗的关键特征，作为后续预测模型的输入。

第三章为了解决实际建筑某些关键变量缺失导致预测模型“瘫痪”的情况，提出了基于遗传算法和基于自编码器的关键变量推断方法。其中，对于有历史能耗的建筑，基于能耗和关键变量间的关系进行缺失关键变量推断，对于无历史能耗的建筑，使用自编码器挖掘变量间的关系进行推断。

第四章提出了结合条件生成对抗网络和 XGBoost 的建筑全年能耗数据增强方法，并使用该方法生成了建筑能耗数据增强数据集。首先对模拟数据集聚类，然后对每一类数据建立一个生成对抗网络模型。为了保证数据增强的效果，生成对抗网络模型仅用于建筑能耗趋势曲线的生成，该曲线仅反映了建筑全年逐日能

耗数据的波动，并不代表具体的能源消耗值。XGBoost 模型预测建筑全年总能耗，用于填充生成器生成的能耗趋势曲线，最终获得建筑全年的用能数据。

第五章通过迁移学习融合增强数据集和真实数据，建立了预测精度较高的建筑能耗预测模型。基于增强数据集建立 LSTM 预训练模型，然后使用真实数据对该预训练模型的参数进行微调得到最终的综合体全年能耗预测模型。

第六章对所建立的能耗预测模型进行验证。选择十栋综合体的监测数据进行预测，并与仅使用真实数据训练的 LSTM 基础模型进行对比，结果显示本研究所建立的基于迁移学习的综合体全年能耗预测模型的预测性能更佳。

第七章总结了本研究的结论、创新点和局限性，并对后续工作进行了展望 。

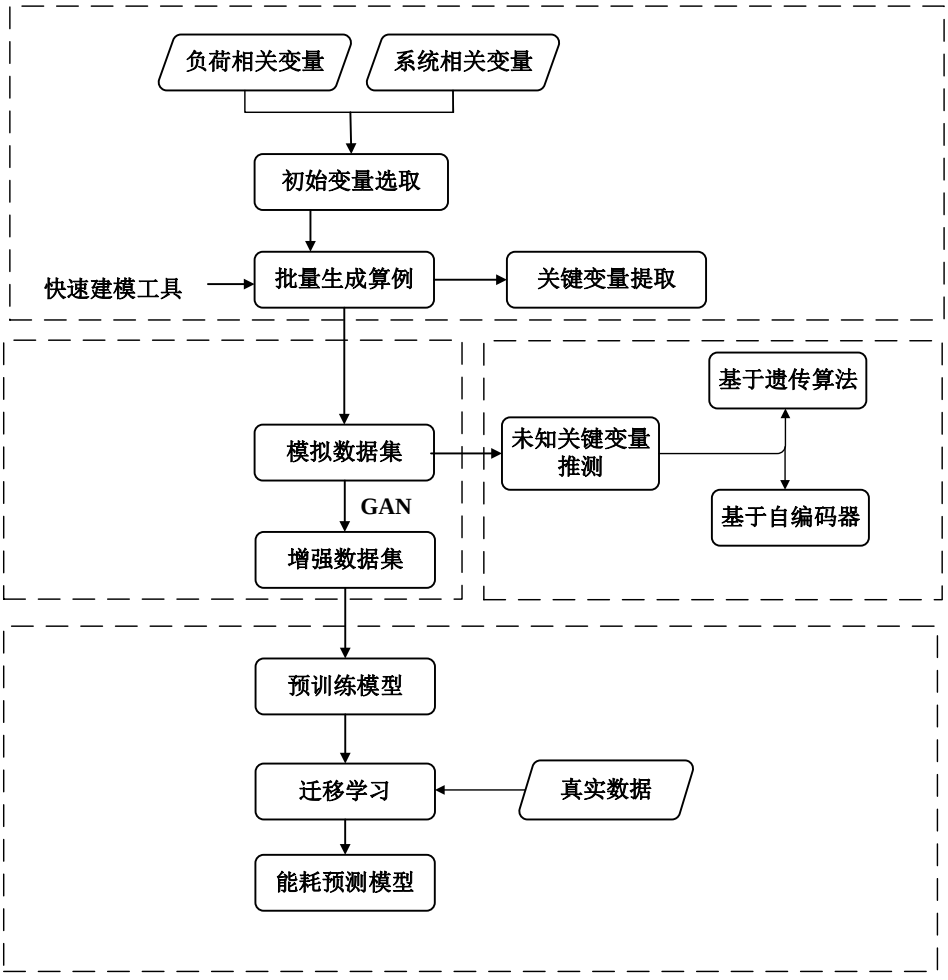


图 1.2 本文技术路线

术语解释：

- 初始变量、关键变量

  初始变量是指本文搜集的与建筑能耗相关的所有变量；关键变量是指初始变量中对能耗数据影响较大的变量。

- 能耗趋势曲线
  能耗趋势曲线是指仅反映建筑全年用能趋势，不反映建筑具体用能大小的曲线。
- 基础模型、预训练模型、迁移模型
  基础模型是指在真实数据集上训练得到的模型；预训练模型是指在增强数据集上训练得到的模型；迁移模型是指在真实数据集上微调后的预训练模型。

1.4 本章小结

本章对目前数据驱动的建筑能耗预测模型和数据增强及融合的研究现状进行了综述，并总结了目前建筑能耗预测存在的三个主要问题，即缺乏面向综合体的能耗预测模型、输入变量缺失问题和数据不足问题。在此基础上确立了将数据增强方法用于补充真实建筑能耗数据的不足，并且使用迁移学习融合增强数据和真实数据的技术路线。

第二章 建筑能耗模型关键变量提取

2.1 概述

影响建筑能耗的因素有许多种，但是实际在做预测模型时，如果将所有相关的因素均作为输入，反而会带来各种问题：一是使用过多特征会导致模型非常复杂，学习了过多样本的噪声进而导致过拟合，尤其是当样本数量较少时；二是输入特征过多，需要耗费大量计算资源和时间；三是繁杂的输入特征，特征之间可能具有高度相关性使模型变得不稳定且难以解释。因此在建立能耗预测模型的第一步往往是通过敏感性分析等方法进行特征选择，舍弃对预测变量影响较小的特征，在减小数据获取难度的同时降低过拟合风险和节约计算成本，提高模型的可解释性。本章节介绍了本研究建筑综合体全年能耗预测模型的关键变量提取的过程，具体流程见图 2.1。

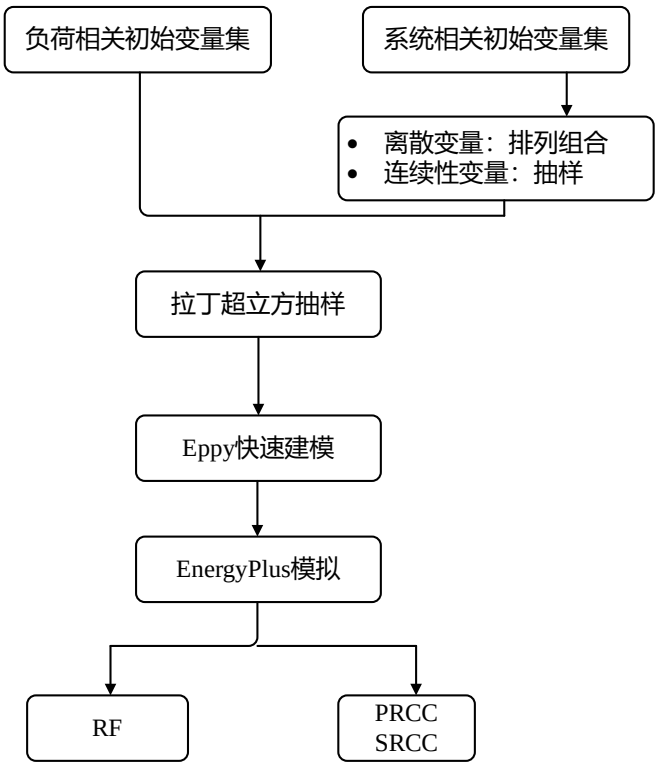


图 2.1 关键变量提取技术路线

2.2 建筑负荷相关关键变量提取

2.2.1 初始变量集选取

本文在进行初始变量集的选取时，参考课题组沙华晶博士论文[57]中酒店建筑能耗预测模型的初始变量集，选择建筑几何外形参数、建筑围护结构热工性能、

建筑日常使用情况以及施工质量四类参数。本文建筑负荷相关初始变量集选取如表 2.1 所示。

表 2.1 建筑负荷初始变量集

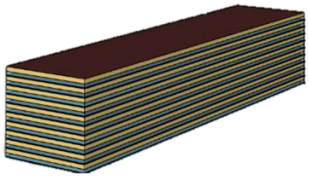
类别	变量名称	缩写	取值范围	单位
建筑几何外形	东向窗墙比	EWWR	0.1-0.9	-
	南向窗墙比	SWWR	0.1-0.9	-
	西向窗墙比	WWWR	0.1-0.9	-
	北向窗墙比	NWWR	0.1-0.9	-
	层数	NL	4-60	层
	建筑面积	AREA	20000-200000	m^2
	体形系数	CR	0.1-0.5	-
建筑围护结构热工性能	外墙传热系数	WALLU	0.09-0.5	$W/(m^2 \cdot K)$
	外墙热容	WSP	800-2000	$J/(kg \cdot K)$
	屋顶传热系数	RU	0.09-0.4	$W/(m^2 \cdot K)$
	窗玻璃传热系数	WINU	0.2-0.9	$W/(m^2 \cdot K)$
	窗玻璃太阳辐射得热系数	SHGC	0.1-0.9	-
	外墙太阳辐射吸收系数	WSA	0.1-0.9	-
	屋顶太阳辐射吸收系数	RSA	0.1-0.9	-
建筑使用情况	空调制冷设定温度	SPC	21-28	℃
	空调制热设定温度	SPH	18-26	℃
	设备照明功率密度	LPD	3-15	W/m^2
	人员密度	OPD	0.1-1	人/m^2
	冷风渗透率	INFIL	0.5-5	ACH
	内遮阳开启程度	ST	0.1-0.9	-
施工质量	楼板线性透过率	FLT	0.007-1.842	W/(m K)
	玻璃线性透过率	GLT	0.03-1.058	W/(m K)
	墙角线性透过率	CLT	0.036-0.684	W/(m K)

2.2.2 算例批量生成及模拟

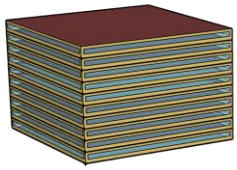
本文为了构建敏感性分析数据集，通过拉丁超立方采样得到 3000 组初始变量组合，再通过白箱的方法构建负荷部分初始变量与建筑能耗数据之间的对应关

系。拉丁超立方采样(Latin hypercube sampling, LHS)最早由 McKay 等[57]提出，是一种从多元参数分布中近似随机采样的方法，属于分层采样技术。拉丁超立方抽样确保了样品的结构与整体结构相对相似，并且样品是均匀的[58]。与蒙特卡罗采样相比，它可以大大减少模拟样本的数量，提高计算效率[59]。

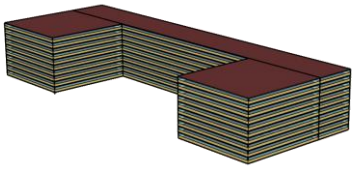
为了实现白箱模型的快速生成，本文基于 python 语言和 eppy 库，读取采样得到的参数集批量构建用于建筑全年能耗数据模拟的 IDF 文件，调取 Energyplus 内核执行建筑能耗批量模拟。白箱模型中建筑外形参考沙华晶博士毕业论文，根据参数中的面积、层数、体形系数匹配五种常见的建筑几何（长方形、正方形、U 形、回字形），如图所示。

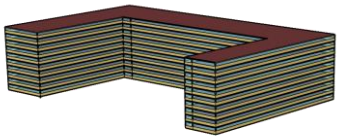


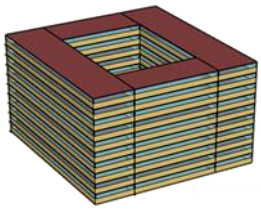
a) 长方形



b) 正方形







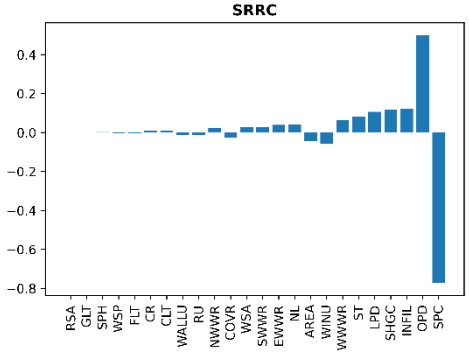
c) U 形（两侧为长方形）d) U 形（两侧为正方形）e) 回字行

图 2.2 建筑外形

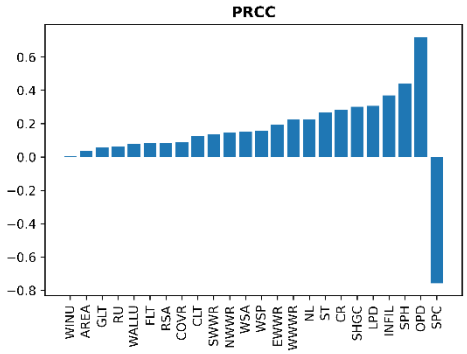
2.2.3 建筑负荷关键变量提取

本文负荷部分关键变量提取分别以制冷能耗和制热能耗为目标，采用标准秩回归系数（SRRC）、偏秩回归系数（PRCC）以及随机森林三种方法。

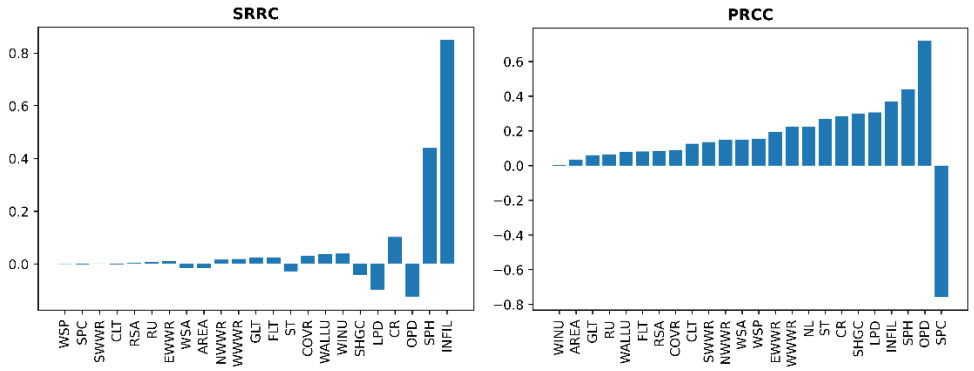
1）标准秩回归系数和偏秩回归系数



（a）



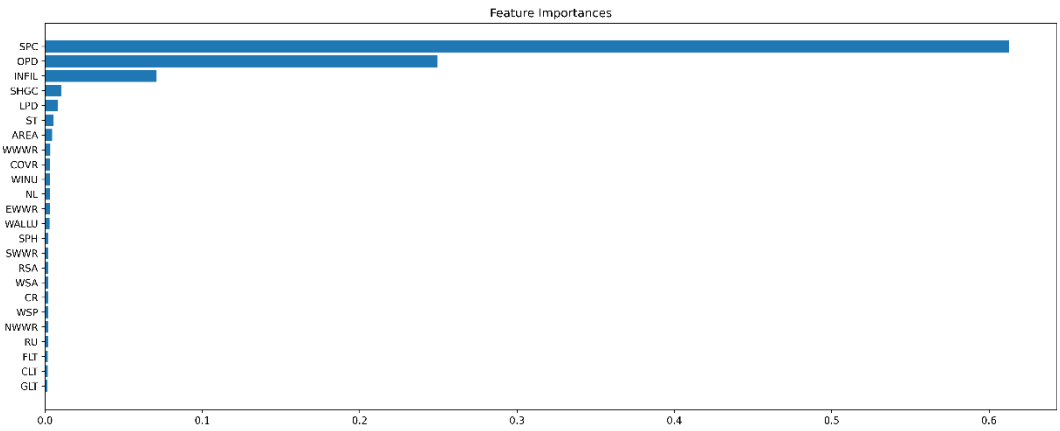
（b）



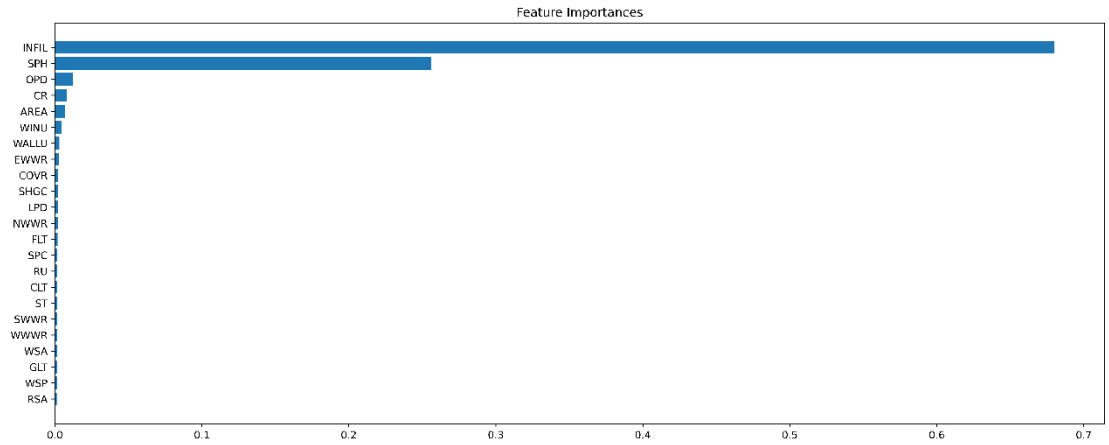
（c） （d）

图 2.3 SRRC、PRCC 关键变量分析（其中（a）（b）以制冷季能耗为目标，（c）（d）以制热季能耗为目标）

2）随机森林



（a）制冷季关键变量提取



（b）制热季关键变量提取

图 2.4 Random Forest 关键变量提取

3）建筑负荷部分关键变量提取结果

根据上述敏感性分析结果，选择标准秩回归系数、偏秩回归系数和随机森林方法中重要性排名前 5 的变量的并集，包括制冷设定温度、制热设定温度、人员密度、照明设备密度、冷风渗透率、体形系数、太阳得热系数、窗墙比和内遮阳开启程度。

2.3 空调系统相关关键变量提取

2.3.1 初始变量集选取

参考沙华晶博士[60]和郭明月硕士[61]的毕业论文，空调系统初始变量集选择如表 2.2。

表 2.2 空调系统相关初始变量集

类别	变量名称	缩写	取值范围	单位
系统类型参数	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、二次泵变流量系统	-
系统运行参数	送风温差	SATD	4-10	℃
	冷冻水供水温度	CHWT	5-10	℃
	热水供水温度	HWT	50-65	℃
	风机效率	FE	0.3-0.8	
	水泵效率	PF	0.3-0.8	
	主机 COP	COP	3-7	
	冷冻水供回水温差	CTD	2-7	℃
	热水供回水温差	HTD	8-15	℃
	冷却塔填料堵塞率	CFBR	0.5-1	-
	风系统过滤器堵塞率	FFBR	1-2	-

算例批量生成及模拟过程同 2.2.2 节，此处不再赘述。

2.3.2 系统相关变量敏感性分析结果

1）标准秩回归系数和偏秩回归系数

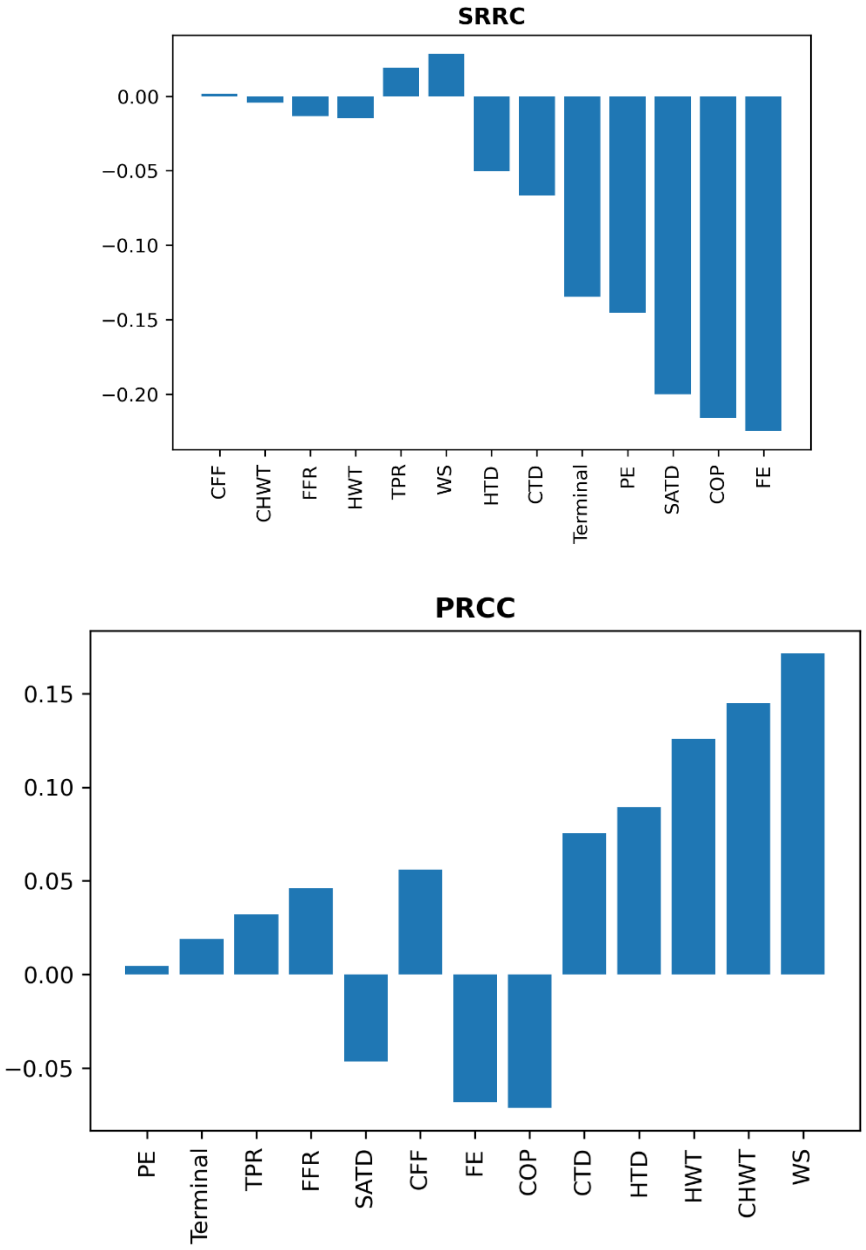


图 2.5 SRRC、PRCC 关键变量分析

2）随机森林

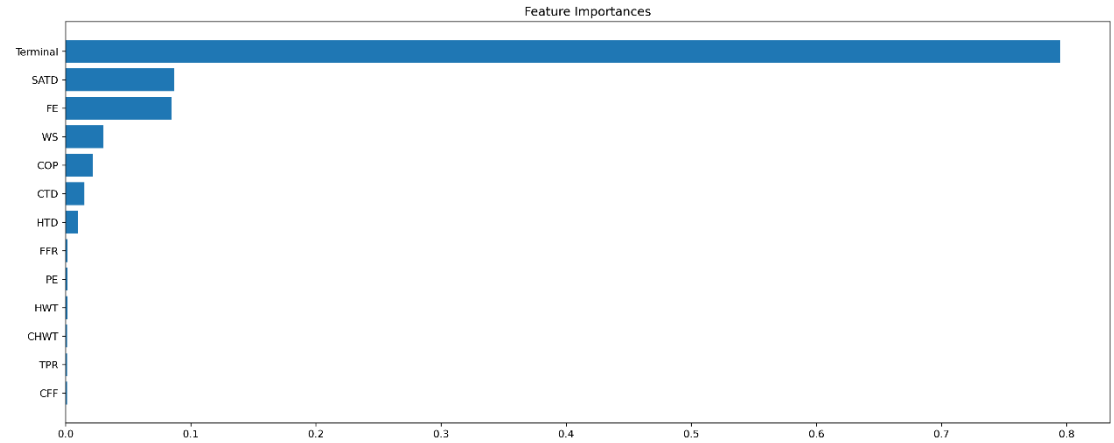


图 2.6 随机森林关键变量分析

根据上述敏感性分析结果，选择标准秩回归系数、偏秩回归系数和随机森林方法中重要性排名前 5 的变量的并集，包括风系统类型、水系统类型、主机 COP、冷冻水供回水温差、热水供回水温差、冷冻水温度、热水温度、风机效率和水泵效率。

2.4 本章小结

影响建筑能耗的因素有许多，将所有相关变量均用于模型预测是不现实且不可行的。为此，本章首先根据专家知识搜集了影响建筑能耗的初始变量集，并将其分为建筑负荷相关变量和空调系统相关变量两大类。建筑负荷相关变量主要包括建筑围护结构参数和内热参数；空调系统相关变量主要指系统特性参数。随后采用相关系数法（SRCC、PRCC）和随机森林两种方法，对两大类变量进行敏感性分析。根据两种方法的敏感性分析结果，通过取并集的方式提取了对建筑能耗影响较大的特征，用于后续模型预测。

第三章 模型关键变量缺失值推断

3.1 概述

虽然近年来公共建筑能耗监测平台建设工作迅速展开，但是这类数据监测平台的主要监测目标是建筑能耗，对建筑围护结构以及系统运行相关信息的监测较少，这就导致在进行建筑能耗预测时，部分关键变量无法获取。在关键变量缺失的情况下无法进行准确的能耗预测，因此对预测建筑的缺失关键变量进行推断是十分必要的。本文针对建筑历史能耗是否可以获取提出了两套关键变量推断方案。

针对有历史能耗的建筑，基于历史能耗和关键变量之间的关系，应用遗传算法进行缺失关键变量推断。对于没有历史能耗的建筑，利用降噪自编码器挖掘关键变量内部的相关关系，对缺失值进行推测。

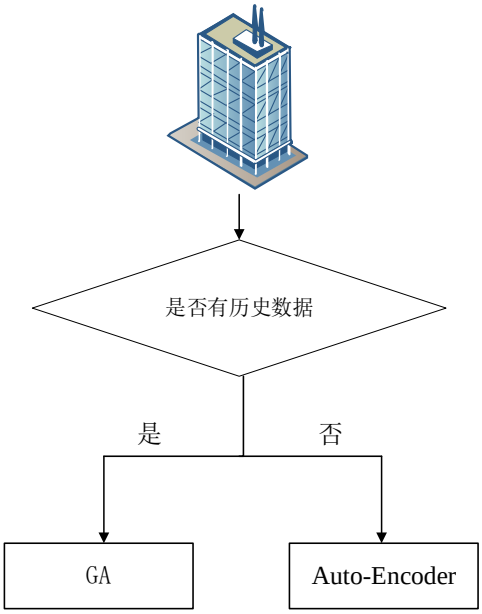


图 3.1 缺失关键变量推断方法

3.2 有历史能耗的建筑关键变量缺失值推断

3.2.1 方法

1、能耗预测模型的建立

利用第二章所述的快速模拟方法，批量模拟 3000 栋建筑综合体模型，得到模拟数据集。通过留出法，选出 80 栋建筑用于后续方法验证部分，余下 2920 栋模型建筑数据建立 XGBoost 能耗预测模型。该模型的输入为第二章所提取的关

键变量加上天气特征和时间特征，如表 3.1 所示，输出为建筑的单位面积逐日能耗。训练集测试集比例为 4：1。在训练集上使用贝叶斯搜索对 XGBoost 模型的学习率、最大深度、最小子节点权重、gmma 和列采样率进行参数寻优。模型性能评估参数为 R^2 和均方根误差（Mean Squared Error）。

表 3.1 能耗预测模型输入

参数类别	参数名称	缩写	单位
建筑负荷相关变量	制冷设定温度	SPC	℃
	制热设定温度	SPH	℃
	人员密度	OPD	人/m^2
	设备照明密度	LPD	W/ m^2
	冷风渗透率	INFIL	ACH
	体形系数	CR	/
	太阳得热系数	SHGC	/
	窗墙比	WWR	/
	内遮阳开启程度	ST	/
系统相关变量	风系统类型	Terminal	/
	水系统类型	WS	/
	主机 COP	COP	/
	冷冻水供回水温差	CTD	℃
	热水供回水温差	HTD	℃
	冷冻水温度	CHWT	℃
	热水温度	HWT	℃
	风机效率	FE	/
	水泵效率	PE	/
天气参数	干球温度	DryT	℃
	相对湿度	RH	/

	风速	Wind	m/s
时间标签	每年的月	month	/
	每月的日	day	/
	星期几	week	/
	是否是工作日	workday	/

值得一提的是，由于综合体一般有多个空调系统（对应不同的功能分区），为了降低特征维度，对系统运行参数按照不同功能分区的面积比例进行加权获得最终建筑维度的参数值。该方法被应用于本文综合体能耗预测模型的建立。以典型功能分区为商业区和办公区为例，x_1、x_2分别为办公区和商业区面积占比，$value_1$、$value_2$分别为办公区和商业区系统对应参数值。$value$最终作为模型输入。

$$value = x_1 \times value_1 + x_2 \times value_2 \quad （1\text{-}1）$$

最终得到的 XGBoost 预测模型 R2 为 0.997，均方根误差为 0.00492，在新建筑上的预测效果如图所示。所建立的预测模型用于基于遗传算法的关键变量推断中。

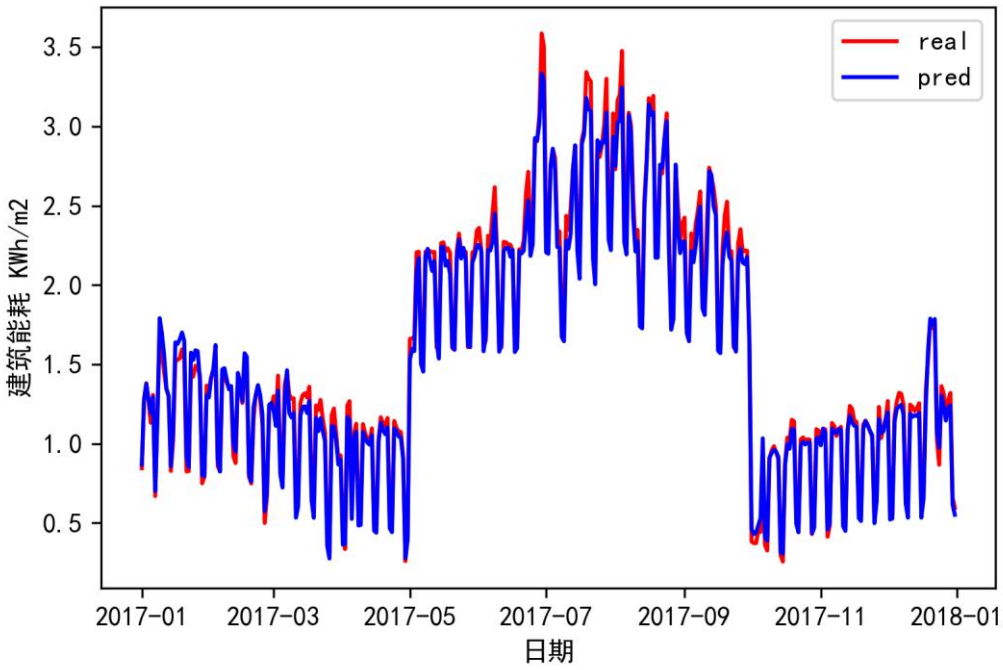


图 3.2 XGBoost 预测效果展示

2、基于遗传算法的关键变量推断

对于有历史能耗的建筑，进行关键变量推断时应尽可能利用已知信息。本节所提出的方法基于关键变量同历史能耗之间的强相关性，通过遗传算法寻找最符合历史能耗趋势的缺失关键变量。遗传算法由 Holland 于 20 世纪 70 年代开发，是一种基于“自然选择、适者生存”的高度并行、随机、自适应的优化算法[62]。本文所提出的方法中，遗传算法的决策变量是目标建筑缺失的关键变量，约束是根据先验知识给出的缺失变量的可行域以及预测模型中所学到的关键变量与能耗之间的关系，优化目标是最小化建筑能耗预测值和真实历史能耗之间的绝对值误差。图 3.2 为基于遗传算法的有历史能耗建筑的缺失关键变量推断的技术路线。具体步骤如下：

- 给定缺失参数的可行域，在可行域内生成初始种群即缺失变量推断值集合；
- 每一个个体连同其他已知的关键变量输入能耗预测模型，得到能耗预测值，计算预测值与建筑历史能耗之间的绝对值误差（即适应度）。
- 改变缺失变量推断值（交叉、变异），重复上一步。当绝对值误差收敛或者到达最大迭代次数，跳出寻优输出最终关键变量推断值。

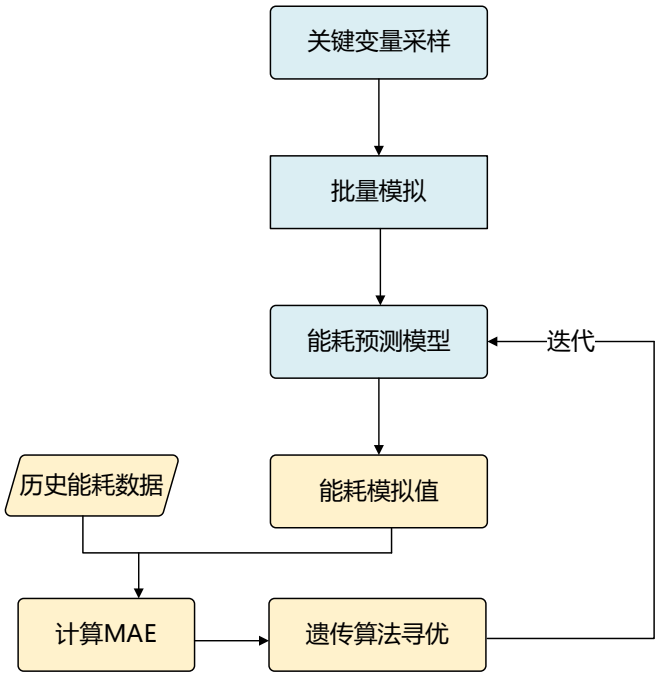


图 3.3 基于遗传算法的缺失关键变量推断技术路线

3.2.2 单变量缺失推断结果

为了验证本节所提出的缺失变量推断方法，使用该方法对 80 栋模拟建筑的缺失变量进行推断，计算各个变量推断结果的误差百分比。选择 11 个典型关键

变量进行推断验证，包括制冷设定温度（SPC）、体形系数（CR）、人员密度（OPD）、冷风渗透率（INFIL）、照明密度（LPD）、窗户太阳得热系数（SHGC）、窗墙比（WWR）、建筑面积（AREA）、冷冻水供回水温差（CTD）、送风温差（SATD）和主机性能系数（COP）。

推断过程的收敛曲线如下图 3.4 所示。

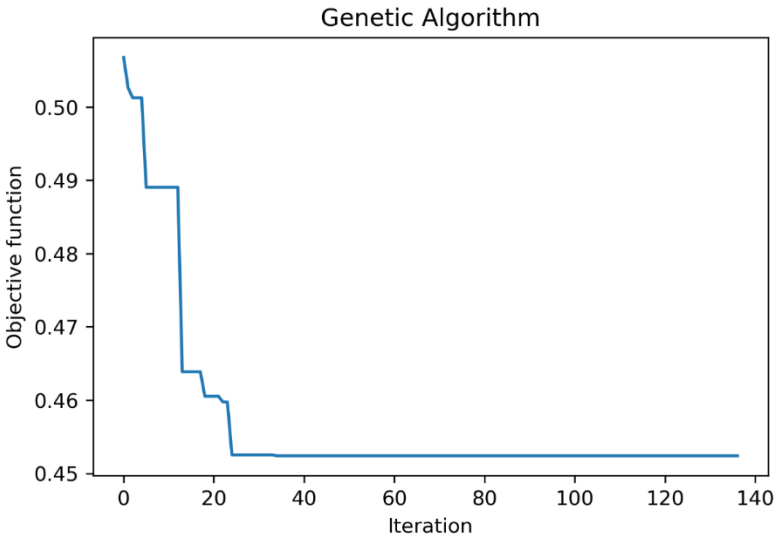


图 3.4 单变量推断误差收敛曲线示例

误差百分比的计算公式如下：

$$error = \frac{|x-\tilde{x}|}{x} \tag{3.1}$$

其中x为缺失变量真值，$\tilde{x}$为缺失变量推断值， $error$ 为误差百分比。

表 3.2 单变量推断误差百分比

缺失变量	符号	平均误差	最小误差	最大误差
制冷设定温度	SPC	0.0038	5.54E-6	0.044
体形系数	CR	0.0219	7.91E-6	0.135
人员密度	OPD	0.017	3.65E-6	0.162
冷风渗透率	INFIL	0.010	6.35E-7	0.128
照明密度	LPD	0.031	4.12E-5	0.264
窗户太阳得热系数	SHGC	0.052	2.38E-4	0.466
建筑面积	AREA	0.121	9.67E-5	0.357
冷冻水供回水温差	CTD	0.045	2.02E-4	0.201
送风温差	SATD	0.035	1.52E-6	0.398
冷机/热泵 COP	COP	0.014	7.63E-5	0.156

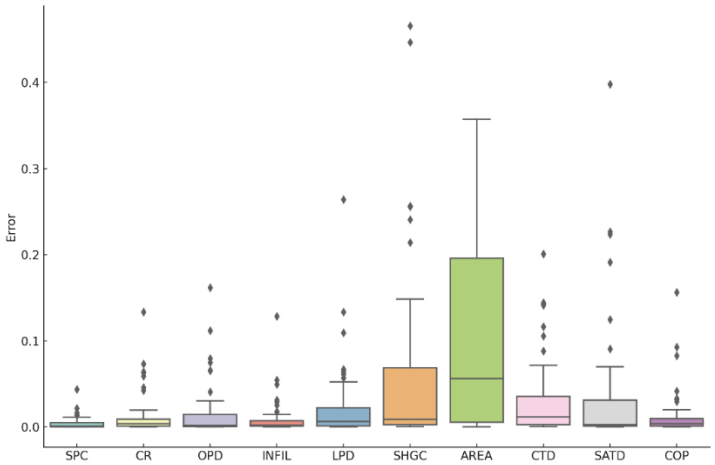


图 3.5 基于遗传算法的单变量推断误差

可以看出，基于遗传算法的关键变量推断的效果良好，除建筑面积以外所有变量的误差均值在 10%以内。推断误差最小的是制冷设定温度，误差均值仅为 0.38%；推断误差最大的是建筑面积，均值误差为 12.1%。由于建筑综合体的面积大小不一，上至几十万平米，下至几万平米，因此建筑面积参数的可行域较大，这给变量推断带来了很大困难，导致建筑面积的推断误差偏大。另一方面，相较于其他关键变量，建筑面积同建筑能耗的相关性稍低。

此外通过图 3.5 也可以发现，与建筑能耗的相关性越强的参数，推断的准确性也越高，如制冷设定温度、人员密度、体形系数、主机性能系数等，而建筑面积、窗墙比、太阳得热系数这些在第二章变量提取中重要性排序稍靠后的参数，推断的准确性也低一些。这是因为同建筑能耗相关性越低的参数，它对建筑能耗的影响也越小，那么当这类参数在一定范围内变动时对建筑能耗的变动也较小。

3.2.3 多变量缺失推断结果

对于现实建筑，获取关键变量时存在多量缺失十分常见，因此本节对建筑多个关键变量缺失的推断进行验证。

假设建筑的制冷设定温度和窗墙比同时缺失，利用基于遗传算法的关键变量推断方法进行推断，由图 3.6 可以看出来，SPC 的推断误差较小，均低于 10%，窗墙比的推断误差稍大，均值误差在 10%以内，但是存在离群值。

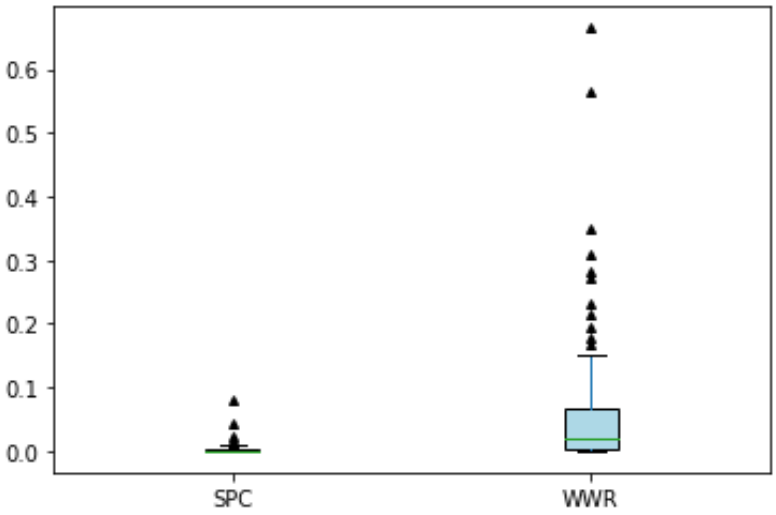


图 3.6 SPC 和 WWR 同时缺失的推断结果

当缺失变量增加到 3 个，即 SPC、WWR 和 COP 同时缺失，推断结果如图 3.7，可以看出，SPC 和 WWR 相比双变量缺失时，推断误差略有增大。图 3.8 和 3.9 显示，当缺失变量的个数继续增加，变量推断误差也在不断增大。这是因为当多个变量缺失时，每个缺失的变量对建筑能耗有着不同的影响（有些正相关有些负相关），那么就会出现不同变量值的组合使建筑呈现相似的能耗表现，推断结果也就很有可能偏离真实值。但是当推断出的缺失变量组合满足历史能耗规律，我们便可认为这些变量组合也会满足未来该建筑的用能规律。也就是说，尽管所推断出的这些缺失变量的值与建筑真值不同，但它们共同作用的效果同建筑真实用能情况相近。

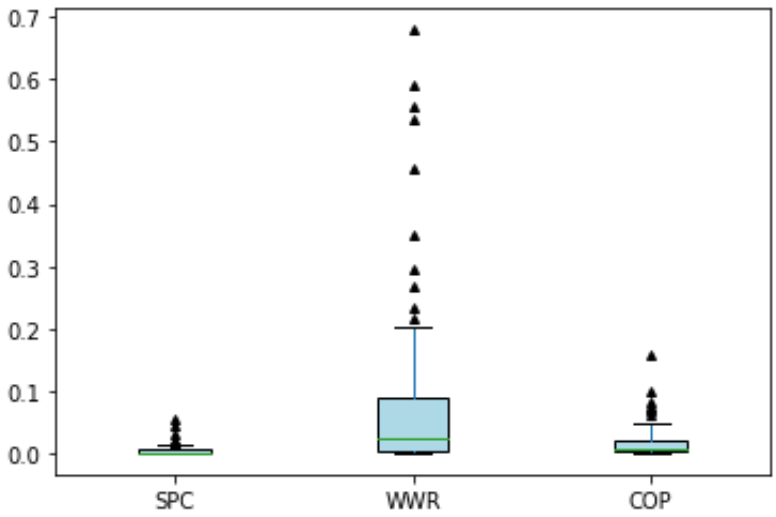


图 3.7 SPC、WWR 和 COP 同时缺失的推断结果

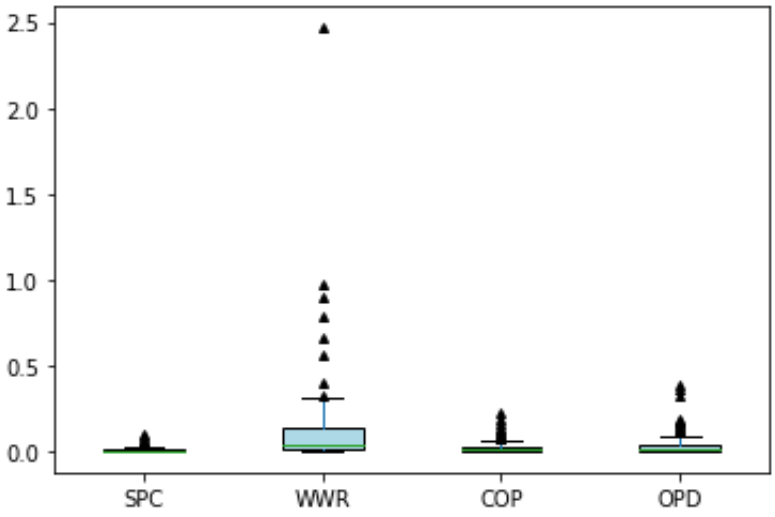


图 3.8 SPC、WWR、COP 和 OPD 同时缺失的推断结果

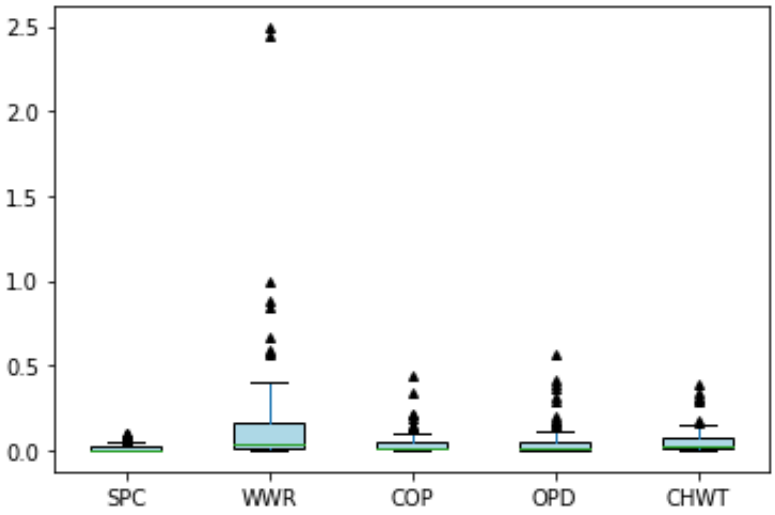


图 3.9 SPC、WWR、COP、OPD 和 CHWT 同时缺失的推断结果

3.3 无历史能耗建筑的关键变量缺失值推断

3.3.1 方法

当建筑历史能耗缺失时，无法根据历史能耗推断出符合历史能耗趋势的关键变量组合，只有从关键变量组合本身出发进行推断。

自编码器（Auto-Encoder）是一种无监督学习模型，最早由 Rumelhart 等人在《Nature》上提出[63]。它利用输入数据本身作为监督来指定到神经网络学习一个映射关系来重构输出[64]。自编码器包括两部分，分别是编码器和解码器。编码器的作用是把输入编码为隐层变量，解码器的作用就是将隐层变量还原到初始维度。编码器的输入维度和解码器的输出维度相同以保证输入重构。自编码器的原

理如下。

$$h_1 = \sigma_e(W_1 x + b_1)$$

$$y = \sigma_d(W_2 h_1 + b_2)$$

其中h_1为隐层向量，y 为输出向量，W 为权重系数，b 为偏置系数，σ是激活函数。

降噪自编码器（Denoising Auto-Encoder）由 Vincent 等人[65]于 2008 年提出，就是一种通过引入噪声来增加编码鲁棒性的自编码器。降噪自编码器的主要研究目标是隐层表达对被局部损坏的输入信号的鲁棒性[66]。对于一个向量，我们对这个向量加入一定比例的噪声得到一个被损坏的向量，将被损坏后的变量送入自编码器的输入段并要求它通过编码加解码两个步骤重构出无损的原始输入。降噪自编码器在异常数据诊断、图像去噪[67][68]等得到广泛应用。Shao 等[69]将降噪自编码器用于旋转机械的故障诊断，使用自编码器来降低背景噪声的影响，学习真实震动数据的鲁棒重建。

本节利用降噪自编码器挖掘关键变量组合内部的关系，重构缺失的关键变量。使用的降噪自编码器网络结构如图 3.10 所示，输入维度和输出维度为关键变量总数。编码器通过逐渐增加层级，学习输入数据的抽象和高级特征，较深的层可以捕获更抽象和高级的特征。解码器逐渐减少层级逐渐减少的解码器层可以有助于去除输入中的噪音。

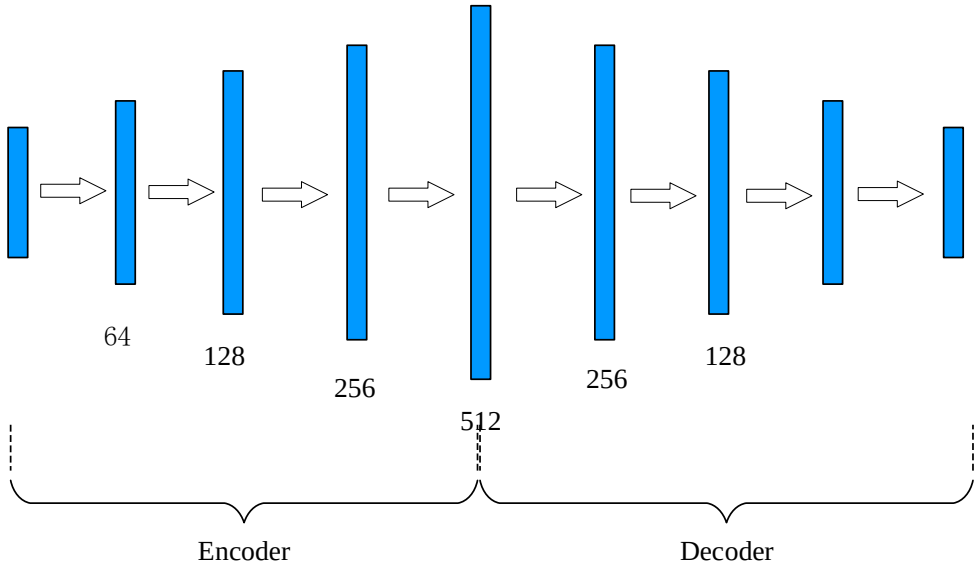


图 3.10 降噪自编码器的结构

3.3.2 单变量缺失推断结果

为了验证本节所提出的基于自编码器的缺失关键变量推断方法，使用该方法对 10 栋实测建筑的缺失变量进行推断，并计算误差百分比，计算公式如 3.2.2 小

节所述。选择典型缺失变量制冷设定温度、体形系数、照明密度、冷冻水供水温度和主机性能系数，具体推断结果如图 3.11 所示。从图中可以看出，单变量缺失时，推断误差在 30%以下，冷冻水供水温度的推断平均误差最大，制冷设定温度和主机性能系数的平均推断误差最小。

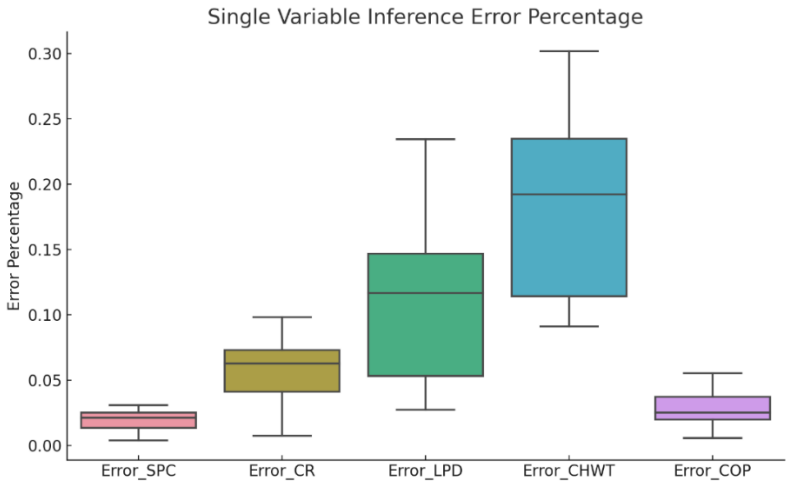


图 3.11 单变量缺失推断误差百分比

3.3.3 多变量缺失推断结果

使用本节所提出的方法，对 10 栋建筑多变量缺失的情况进行推断，结果如图 3.12-3.15 所示。从图中可以看出，多变量缺失时推断误差均在 30%以下，随着缺失变量的增加，推断误差增加，且箱线图箱子高度增加，这表明推断结果的不确定性有所增加。

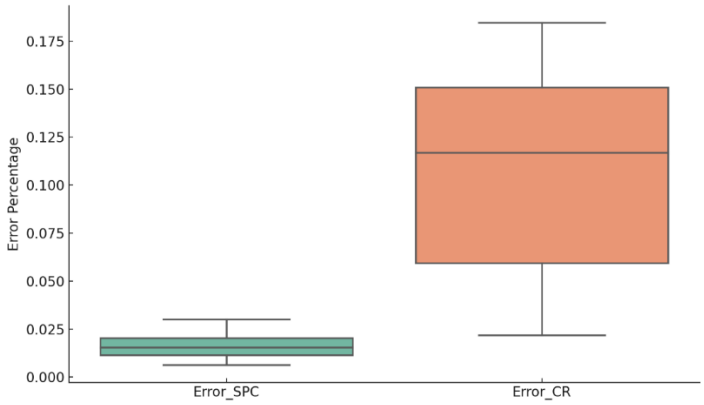


图 3.12 SPC 和 CR 缺失推断结果

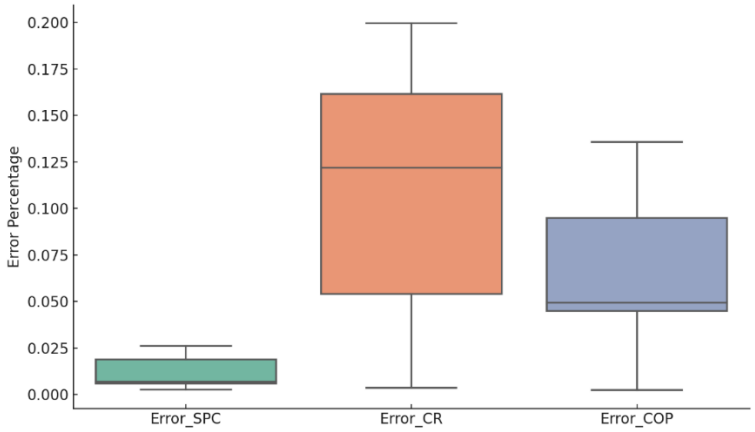


图 3.13 SPC、CR 和 COP 缺失推断结果

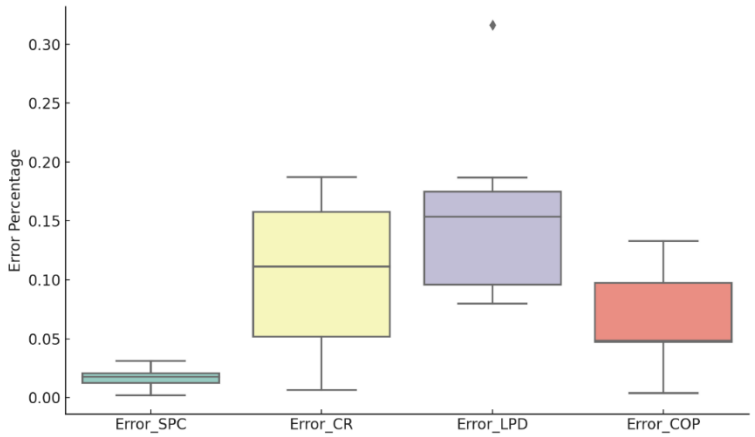


图 3.14 SPC、CR、LPD 和 COP 缺失推断结果

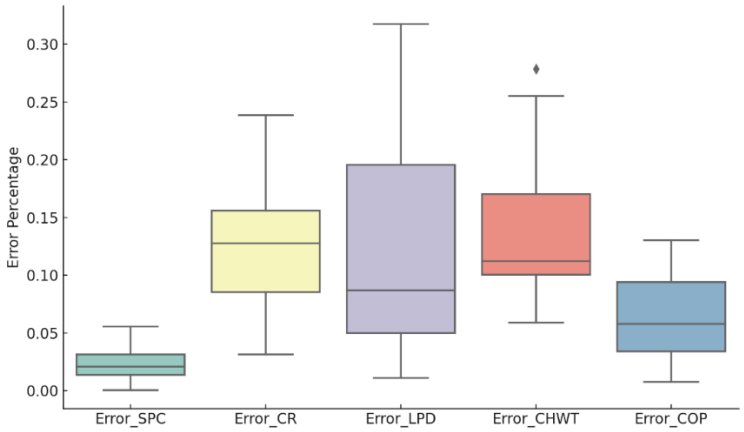


图 3.15 SPC、 CR、LPD、CHWT 和 COP 缺失推断结果

3.4 本章小结

本文在前一章提取了对建筑能耗影响较大的关键变量，但这其中的一些关键变量，现实中由于数据丢失或者未监测等原因难以获取，这给模型预测工作带来极大难度。为了解决上述困境，本章节提出了两种不同情境下的建筑缺失关键变量推断方法。针对有历史能耗的建筑，提出了基于遗传算法的关键变量推断方法。以缺失关键变量作为决策变量，通过遗传算法寻找对应模拟能耗与实测能耗偏差最小的缺失关键变量的取值。针对没有历史能耗的建筑，提出了基于自编码器的关键变量推断方法。通过自编码器学习关键变量之间的关联关系，根据其关联关系进行缺失关键变量的推断。前者经验证平均推断误差低于 20%，后者经验证平均推断误差低于 30%，证明本文所提出的缺失关键变量推断算法的可靠性。

第四章 基于生成对抗网络的数据增强方法

4.1 概述

数据驱动模型在挖掘实际建筑能耗特征和提高预测精度方面具有优势，但需要大量的训练数据来保证较强的泛化能力。数据是数据驱动模型的燃料，然而在实际应用中往往存在数据不足的问题，这将严重影响数据驱动模型的预测性能。从机器学习到深度学习，随着算法模型复杂度变高，训练好一个模型所需要的数据量也越来越大。但是实际中，数据获取有以下几个困境。首先，所能够获得的数据并不能覆盖整个样本空间，或是存在着严重的数据不平衡问题；其次，由于传输、存储或数据质量问题导致的数据短缺也会降低训练样本的多样性，这将训练样本限制在局部学习空间，导致数据驱动模型泛化能力弱[42]；最后对于建筑能耗数据来说，能耗模拟常常是数据的一大来源，但能耗模拟耗时耗力且对专业领域知识要求高。数据增强是一种通用的独立于模型的数据端解决方案[70]。数据增强的目的是通过从相关任务中借用实例来增加样本种类和样本量，以便于目标预测任务的建模。

本章节提出了基于条件生成对抗网络（CGAN）的建筑能耗数据增强方法。技术路线如图 4.1 所示。本章节技术路线主要包括两大部分，分别是生成对抗网络模型和全年总能耗预测模型的建立，其中前者用于建筑用能趋势曲线的生成，后者用于预测建筑全年总能耗，用于填充趋势曲线。CGAN 模型基于模拟数据集训练得到，具体步骤如下：为了得到性能更佳的生成对抗网络模型同时解决数据不平衡问题，本研究首先对模拟数据集中的建筑全年的逐日能耗趋势曲线进行聚类，接着为每一类数据训练对应的 CGAN 模型，CGAN 模型的条件为建筑的关键变量。

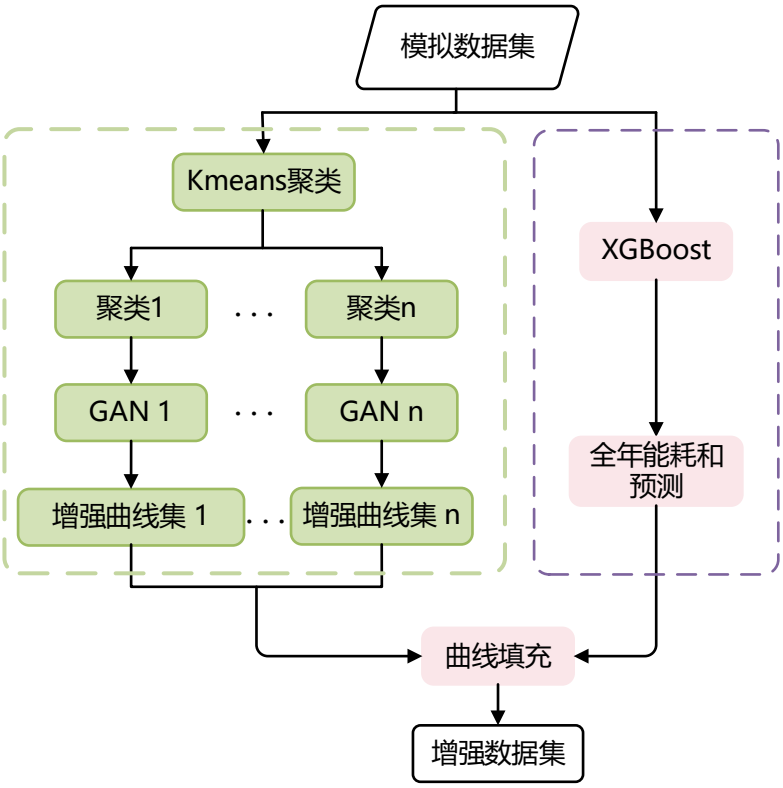


图 4.1 基于 CGAN 的数据增强技术路线

4.2 基于条件生成对抗网络的年能耗曲线生成

4.2.1 数据预处理

生成对抗网络模型的训练数据为通过 EnergyPlus 获取的模拟数据集。由于神经网络模型不能处理非数值型变量，因此将风系统类型、水系统类型这种类别型数据通过独热编码转换为数值型数据。为了加快收敛并且改善训练稳定性，把建筑特征及能耗数据进行-1 到 1 归一化。

4.2.2 K 均值聚类

由于建筑体量及围护结构空调系统等参数的不同，建筑全年能耗差别较大，但是仅考虑能耗曲线趋势，那么同一气候区建筑之间的全年用能曲线有很大的相似性。为了保证生成对抗网络模型的性能以及最终得到的增强数据集的平衡性，我们对归一化后的建筑全年能耗曲线使用 K-means 算法进行聚类。KMeans 本质上是一种基于欧式距离度量的数据划分方法，它的基本思想是通过迭代寻找 K 个簇（Cluster）的划分方案，使得聚类结果对应的损失函数最小。公式 4-1 为 K-means 的损失函数，其中$X = \{X_1, X_2, \dots, X_n\}$为数据集，$C = \{C_1, C_2, \dots, C_k\}$为聚类中心。

$$J(X,C) = \sum_{i=1}^{k}\sum_{j=1}^{n} d(C_i, X_j) \tag{4-1}$$

聚类数 K 值的选择一般基于实验和多次实验结果。本研究使用肘部法则确定最佳聚类数，尝试了聚类数从 1 到 10，绘制出的肘部图如 4.2 所示。根据肘部图结果，选定最佳聚类数为 3。聚类出的典型能耗曲线如图 4.3 所示。第一类曲线是最常见的夏热冬冷地区的典型曲线，呈现出明显的季节性能耗变化，符合上海地区夏季高温高湿和冬季寒冷干燥的气候特征。第二类曲线夏季能耗显著高于冬季，可能表明建筑内热较大或者夏季制冷设定温度很低，冬季制热设定温度较高。第三类曲线冬季制热能耗高于夏季制冷能耗，这可能是因为建筑外围护结构保温性能差，窗墙比高，建筑内热小或者冬季制热设定温度低，夏季制冷温度较高等原因。

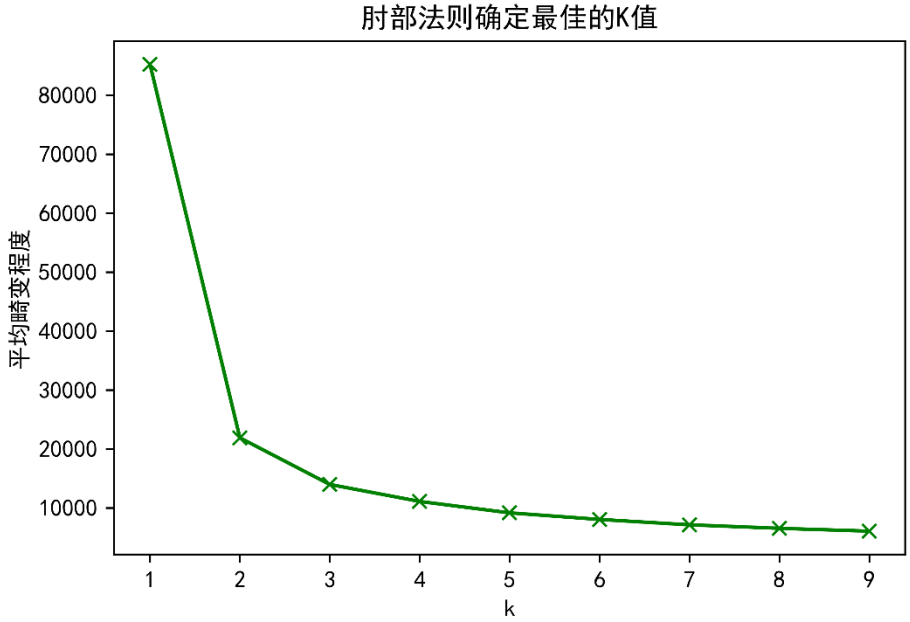
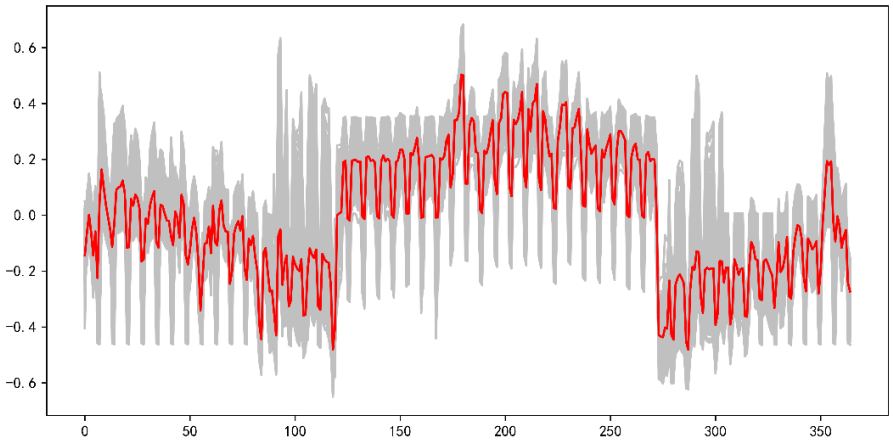
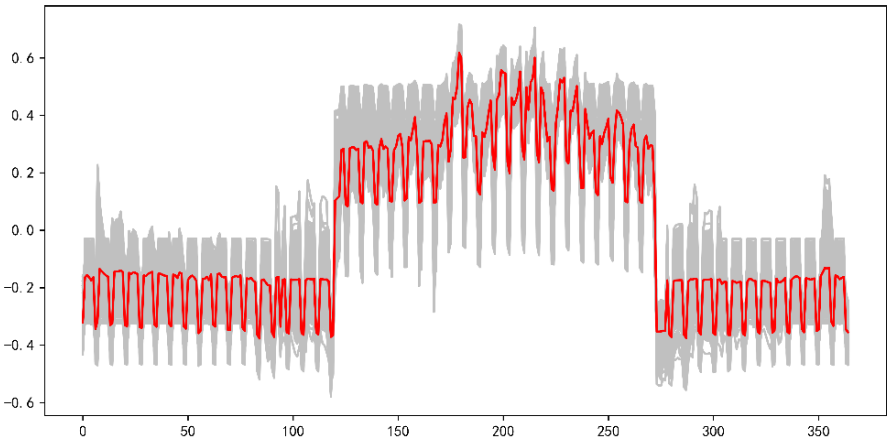


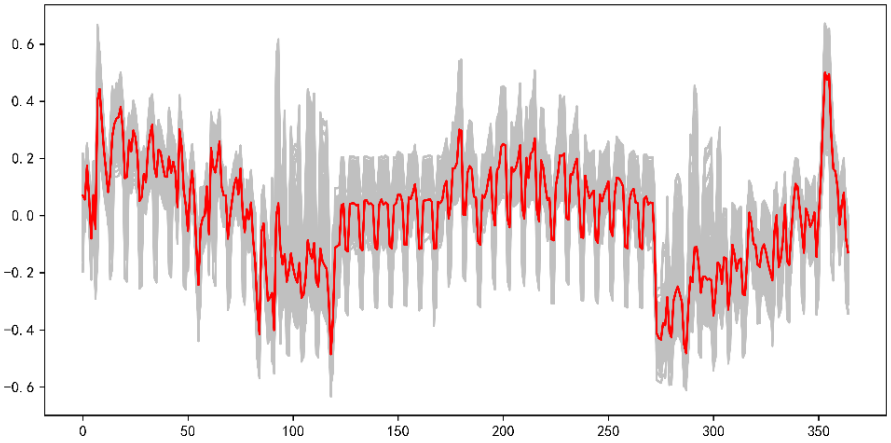
图 4.2 肘部法则



（a）第一类



（b）第二类



（c）第三类

图 4.3 K-means 聚类结果

4.2.3 条件生成对抗网络模型

传统的生成模型由于泛化能力较弱，通常需要计算马尔科夫链，这严重影响了网络性能和生成结果。2014 年，Goodfellow 等人[71]提出生成对抗网络（GAN）。该网络由判别器和生成器组成。生成器将随机噪声转换为预期输出，而判别器则用于鉴别生成器提供的输出是真的还是假的。生成器和判别器在训练过程中相互对抗，最终达到纳什平衡。GAN 的目标函数如公式 4-2 所示。

$$min_G max_D L_{GAN}(D,G) = E_{x\sim p_{data}(x)}[logD(x)] + E_{z\sim p_z(z)}[log(1-D(G(z)))] \quad (4\text{-}2)$$

式中 G 是生成器，D 是判别器，$p_{data}(x)$表示真实数据分布，$p_z(z)$表示生成器生成数据分布。当生成数据分布和真实数据分布达到一致时，即$p_{data}(x) =$

$p_z(z)$，达到纳什平衡

条件生成对抗网络是原始生成对抗网络模型的一种变体，由 Mirza 等人[72]在 2014 年提出。条件生成对抗网络通过在生成器和鉴别器中引入额外的标签进行训练，这就使得生成的数据不仅仅是从数据分布中随机抽取的，而是在额外条件约束下生成的。CGAN 的训练目标和传统的 GAN 相似，都是让生成器生成逼近真实数据分布的数据，同时让鉴别器更加准确识别数据的真伪。不同之处在于 CGAN 要考虑条件的匹配度，生成的数据不仅要符合真实数据分布，还要满足给定的额外条件。条件生成对抗网络的目标函数如式 4-3 所示。其中 y 为额外增加的标签。

$$min_G max_D L_{CGAN}(D,G) = E_{x\sim p_{data}(x)}[logD(x|y)] + E_{z\sim p_z(z)}[log(1 - D(G(z|y)))] \quad (4\text{-}3)$$

图 4.4 为本文所用条件生成对抗网络的原理图。从图中可以看出，建筑静态参数、天气参数和时间标签被当作额外的条件输入到生成器中，指导生成器的产生同这些参数相符合的能耗趋势曲线。也就是说，生成器的输入包括随机噪声，建筑静态参数、天气参数、时间标签，输出为建筑全年能耗数据；鉴别器的输入为建筑全年能耗数据，输出为对该数据是否真实的判断，1 为真，0 为假。注意，这里提到的建筑全年能耗数据均为进行-1 到 1 归一化后的数据，即建筑全年能耗趋势数据。

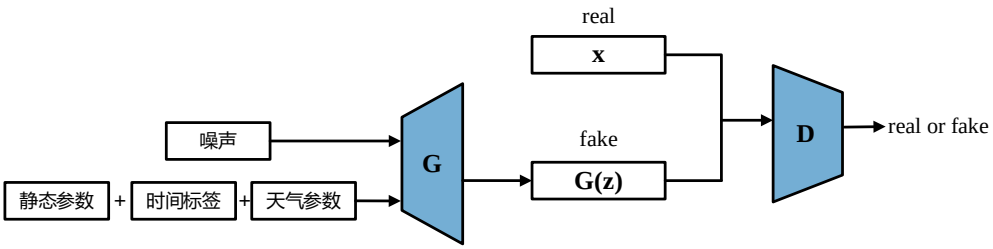


图 4.4 条件生成对抗网络原理图

由于 3.2.1 聚类得到的三类能耗曲线的差别较大，如果仅训练一个生成对抗网络模型，会导致该模型所获取的真实数据分布范围较混乱，导致数据生成效果差。为了尽可能保证生成器的性能，为每一类别能耗曲线建立了一个生成对抗网络模型。下面以第一类能耗曲线的生成对抗网络模型为例，详细介绍模型的建立和训练过程。

模型的生成器由 7 个全连接层和 1 个池化层组成，为了加快收敛速度，最后一层的激活函数为 Tanh，其余层激活函数为 ReLU 函数。鉴别器由全连接层构成，最后一层的激活函数为 Sigmoid 函数，其余层激活函数为 LeakyReLU 函数。

模型优化器使用 Adam，Adam 在避免梯度消失以及自适应学习率上有显著优势。生成器学习率为 0.0001，鉴别器学习率为 0.00005，使用学习率衰减技术，每 50 步学习率衰减为原来的十分之一。模型损失函数选择二元交叉熵损失。模型训练迭代 200 步基本达到收敛，模型训练过程的收敛曲线如图 4.5 所示，在 200 步内，生成器和鉴别器的误差基本稳定，从收敛曲线中也可以看出训练过程也是生成器和鉴别器博弈学习的过程。鉴别器的稳定误差接近 0.5，这意味着生成器生成的人工数据能够欺骗鉴别器，生成的数据与原始数据的分布几乎相同。

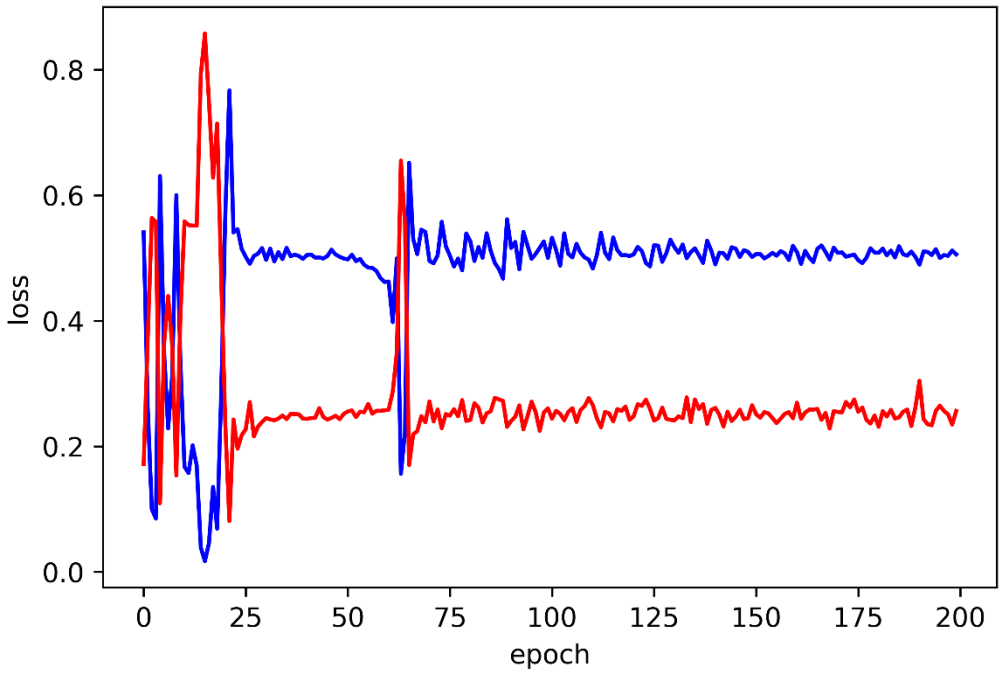


图 4.5 训练过程收敛曲线（红色为生成器损失，蓝色为判别器损失）

模型训练的最终目标是希望生成器生成的数据同原始数据尽可能相似，因此使用均方根误差评估模型生成器的生成效果。式 4-4 中，y_i代表增强前能耗数据，$\hat{y}_i$代表增强后能耗数据。经计算，训练集的均方根误差为 0.0133，测试集的均方根误差为 0.0162，这反映出生成器的数据生成效果良好，可以用于数据增强。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4-4}$$

第一类能耗数据生成效果如图 4.6 所示。从图中可以看出，数据增强在应对模拟数据中连续极度相似趋势时，可以一定程度上通过深度学习挖掘能耗曲线的内部规律来对能耗数据重新生成。

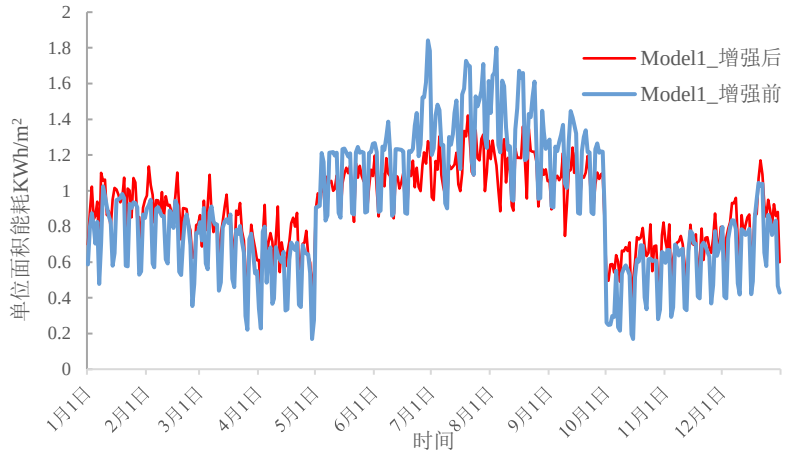


图 4.6 第一类能耗曲线数据增强效果（蓝色为模拟数据，红色为增强数据）

另外两类数据的生成对抗网络模型的建立过程同上述相似，不再赘述，仅展示数据生成结果，如图 4.7 和图 4.8 所示

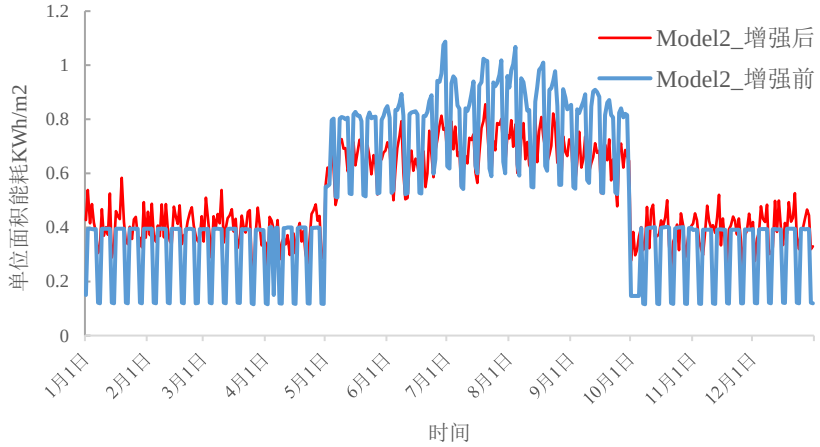


图 4.7 第二类能耗曲线数据增强效果

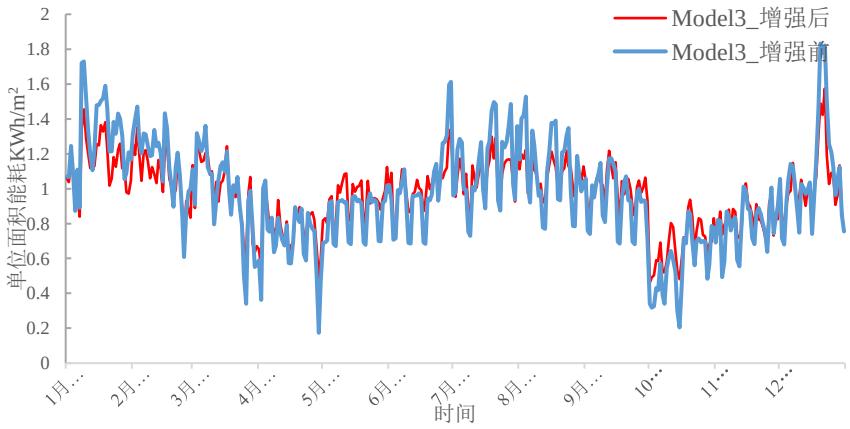


图 4.8 第三类能耗曲线数据增强效果

4.3 全年总能耗预测模型

使用条件生成对抗网络进行建筑全年能耗趋势曲线的生成之后，我们可以得到建筑全年的用能趋势，用能趋势仅仅反映用能规律，并不能反映建筑的用能大小。本文提出根据建筑参数预测建筑全年总能耗，然后对能耗趋势曲线进行填充得到最终建筑全年的能耗曲线。

选择一个有效的算法对预测模型的性能至关重要。为了选择合适的算法模型，本章节进行了算法审查工作。算法审查是进行模型选择的最有效的方法之一[73]。审查算法前没有办法判断哪个算法对数据集最有效、能够生成最优模型。

全年总能耗预测是一个典型的回归问题，模型的输入包括第二章所提取的所有关键变量，输出为建筑的全年总能耗。数据总量为 3000 组，全部为模拟数据，训练集测试集按照 8：2 的比例划分。

1. 算法审查

本章节选择线性回归（LR）、支撑向量机（SVR）、K 近邻（KNN）、分类与回归树（Cart）、随机森林（Random Forest）和极致梯度提升（XGBoost）这几个典型回归算法进行审查。模型审查在训练集上使用 5 折交叉验证，审查指标为净均方根误差。所有数据进行 0-1 归一化。审查结果如图 4.9 所示。从图中可以看出集成模型（Ramdom Forets 和 XGBoost 性能远优于普通回归模型），最终选择 XGBoost 算法来建立全年总能耗预测模型。

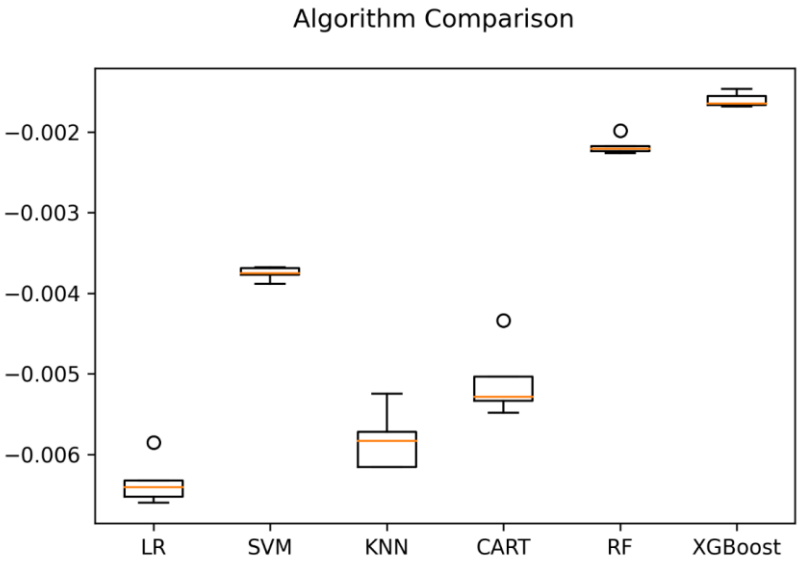


图 4.9 算法审查结果

XGBoost 算法由陈天奇（Tianqi Chen）和卡洛斯·格斯特林（Carlos Guestrin）在 2016 年发表的论文《XGBoost: A Scalable Tree Boosting System》中正式提出[74]。XGBoost 是一种集成学习算法，在建筑能耗预测[75]及故障诊断[76]等多个领

域均表现出优异的性能，在各种机器学习竞赛中占据主导地位。

2. 超参优化

本文使用贝叶斯搜索算法对 XGBoost 模型的超参数进行优化，测试集上采用五折交叉验证策略。贝叶斯搜索算法是一种基于贝叶斯统计原理的优化算法。它主要用于在不确定的搜索空间中找到最优解，尤其适用于那些评估成本较高的函数[77]。优化的超参数包括学习率 learning_rate、树的最大深度 max_depth、最小叶子节点样本权重和 min_child_weight、gamma 和特征采样比例 colsample_bytree。由于决策树模型在进行节点分割时基于特征阈值，与特征具体大小无关，因此不需要对数据进行归一化。参数寻优结果如表 4.1 所示。

表 4.1 xgboost 模型贝叶斯超参寻优结果

参数	learning_rate	max_depth	min_child_weight	gamma	colsamle_bytree
范围	[0.05,0.10,0.15, 0.20,0.25, 0.30]	[3,4,5,6,8,10,12,15]	[1,3,5,7]	[0.1,0.1,0.2,0.3,0.4]	[0.3,0.4,0.5,0.7,0.8,0.9]
结果	0.15	6	3	0.1	0.8

3. 结果

模型评价指标选择均方根误差和 R^2。MSE 的计算公式同 4-4，R^2 的计算公式如下，其中SSE_{res}是残差平方和，即实际观测值与模型预测值之差的平方和，SS_{tot}是总平方和，即实际观测值与其平均值之差的平方和。

$$R^2 = 1 - \frac{SSE_{res}}{SS_{tot}}$$

最终得到全年总能耗的预测均方根误差为 2164.06，R^2为 0.9308。

4.4 数据增强

生成对抗网络模型和全年总能耗预测模型的训练完成后便可进行数据增强，具体增强步骤如图 4.10 所示。图中模型 A 为 3.2.1 节基于模拟数据集所训练到全年能耗预测模型，模型 B 为 4.3 节建立的全年总能耗预测模型。具体步骤如下首先，根据事先定义好的变量区间，对建筑的关键变量进行拉丁超立方采样；接着，将关键变量输入 3.2.1 所建立的 XGBoost 模型预测该建筑的全年能耗曲线。计算该全年能耗数据同各聚类中心的距离，选择最近类别数据训练得到的生成器用于生成能耗趋势曲线；然后，将关键变量作为 CGAN 模型和全年总能耗预测模型的输入，得到该建筑的全年能耗趋势曲线和全年总能耗；最后，使用全年总能耗对该建筑的全年趋势曲线进行填充得到该建筑全年逐日能耗数据。

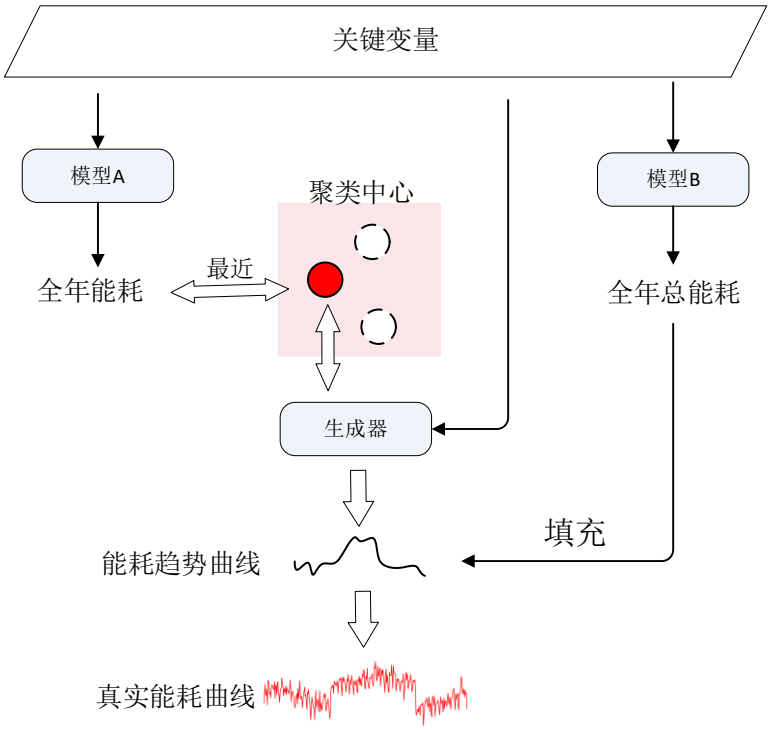


图 4.10 数据增强流程

4.5 本章小结

数据是数据驱动模型的燃料。建筑能耗预测模型发展至今，最终限制其预测性能的往往是数据问题。这其中最常见也是最难解决的数据问题之一便是数据不足。跨建筑能耗预测对数据量的要求更高，它不仅仅要求某栋建筑的历史能耗数据充足，而是要求有足够多栋建筑的历史能耗数据。为此，本章提出了基于条件生成对抗网络的数据增强方法。该方法分成两部分。第一部分是通过条件生成对抗网络进行能耗趋势曲线生成。首先为了保证数据生成效果，通过能耗数据聚类，为三类能耗趋势曲线分别训练生成对抗网络模型。该模型可根据建筑静态参数、时间标签以及天气特征生成对应的能耗趋势曲线。趋势曲线为归一化后的真是能耗曲线。第二部分是建立 XGBoost 全年总能耗预测模型，利用该模型预测的全年总能耗对趋势曲线进行填充，得到建筑的全年逐日能耗曲线。本章节所提出的方法相比白箱模拟，可以更加快速高效生成大量增强数据。基于本章节所提出的方法，为每类能耗曲线生成 2000 栋建筑的增强数据，共 6000 栋用于后续预测模型的开发。

第五章 基于迁移学习的能耗预测模型的建立

5.1 概述

迁移学习是进行数据融合的有效手段之一。迁移学习的核心是，找到源领域和目标领域之间的相似性，并加以合理利用[78]。迁移学习很好地缓解了大数据与少标注和大数据与弱计算之间的矛盾。目前迁移学习已经广泛应用于计算机视觉[79],[80]、文本分类[81]和医疗健康等领域。一般将具有足够数据的域定义为源域D_S，将在源域上执行的任务定义为源域任务D_T。我们想要处理但数据不足的域和任务将定义为目标域T_S和目标任务T_T。对于本研究来说，目的是通过迁移学习融合增强数据集和真是数据集，那么源域D_S就是第四章生成的增强数据集，目标域T_S是真实建筑能耗数据，来自上海某能耗监测平台，源域任务和目标域任务相同，均为建立建筑全年能耗预测模型。如图 5.1 为本章的研究路线，首先使用增强数据集训练了一个长短时记忆网络模型，接着将该预训练模型迁移到目标域，使用真实数据集对该模型的参数进行微调。

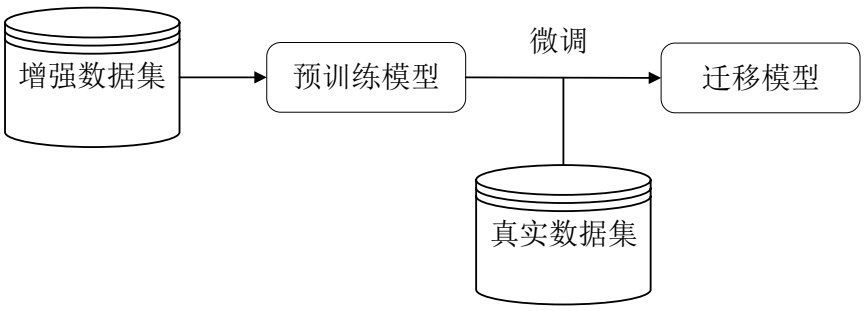


图 5.1 本章研究路线

5.2 基于增强数据集的预训练模型建立

5.2.1 算法介绍

一般来说，神经网络模型分为前馈神经网络（FNN）和递归神经网络（RNN）。前馈网络从输入层向输出层单向传播，无记忆性。而循环神经网络使用自带反馈的神经元，具有内部记忆和循环，可以将前一个时间步长的信息循环回网络，即时间步之间的信息共享[82]。然而 RNN 在处理长间隔序列时由于局部误差的传播会带来“梯度消失”或“梯度爆炸”问题。为了克服 RNN 的上述缺点，Hochreiter 和 Schmidhuber 基于“记忆块(门单元)”的概念为 RNN 设计了一种新的架构，称为长短期记忆（LSTM）[83]。LSTM 在 RNN 的基础上，引入隐藏层自连接的记忆

单元和门控机制来解决“梯度消失”或“梯度爆炸”问题[84]。自连接的记忆单元使模型可以捕获时间序列中的长期依赖关系。三个门控单元，包括输入门、遗忘门和输出门，使模型能够分别从存储单元中写入、更新、忘记和读取信息。

LSTM 的内部详细的结构图如图 5.3 所示。在时间步长 t，遗忘门计算前一时间步长同一存储单元传入信息H_{t-1} 需要保存的比例，如式 5-1；输入门计算当前新信息X_t中需要与前一个时间步长输出H_{t-1}合并的比例，如式 5-2，合并后的信息的定义为时间步长 t 的记忆单元状态，如式 5-3；输出门根据激活函数来控制存储单元的输出，如式 5-4 所示。最终在时间步长 t，当前 LSTM 单元的隐藏状态和最终输出如式 5-5 和 5-6 所示。

$$f_t = \sigma(W_{fx} \cdot X_t + W_{fh} \cdot H_{t-1} + B_f) \tag{5-1}$$

$$i_t = \sigma(W_{ix} \cdot X_t + W_{ih} \cdot H_{t-1} + B_i) + \tanh(W_{cx} \cdot X_t + W_{ch} \cdot H_{t-1} + B_c) \tag{5-2}$$

$$C_t = f_t * C_{t-1} + i_t * g_t \tag{5-3}$$

$$o_t = \sigma(W_{ox} \cdot X_t + W_{oh} \cdot H_{t-1} + B_o) \tag{5-4}$$

$$H_t = o_t * \tanh(C_t) \tag{5-5}$$

$$y_t = (W_{hy} * H_t + B_y) \tag{5-6}$$

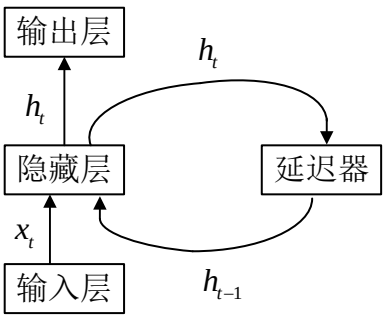
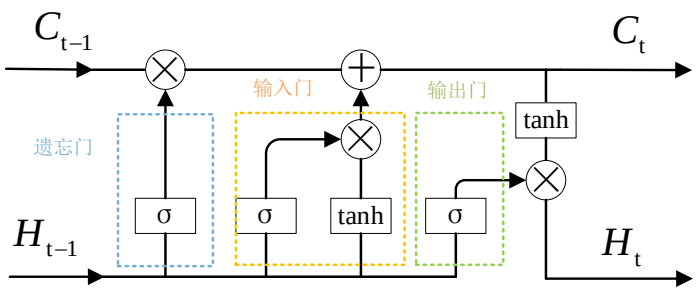


图 5.2 RNN 结构图



tanh tanh函数　　σ sigmoid函数

图 5.3 LSTM 结构图

目前，LSTM 模型非线性时间序列中得到广泛应用，在建筑能耗预测中表现出优异的性能。Zhou 等[85]对比了 SVR、BPNN 和 LSTM 在居民用户和商业用户建筑负荷预测方面的效果发现，LSTM 的预测效果更佳。Somu 等[82]提出了 ISCOA-LSTM 模型用于建筑能耗预测，采用改进的正弦余弦优化算法对 LSTM 的超参数进行优化，经验证该预测结果稳定、准确。Karijadi 和 Chou[86]提出了一种基于经验模态分解的混合 RF-LSTM 的建筑能耗预测方法，该方法使用 LSTM 预测分解后的分量，该模型的表现远优于传统机器学习模型。

本文的任务是建筑全年逐日能耗预测，属于非线性时间序列预测问题。参考 LSTM 在以往相似任务中的优异性能[87]，本文决定选用 LSTM 模型。

5.2.2 数据来源

预训练模型的数据为第四章生成的 6000 栋建筑的增强数据集。数据集包含建筑特征（静态参数、天气参数、对应的时间标签）和全年单位面积能耗，数据颗粒度为逐日。数据集具体内容如表 5.1 所示。

表 5.1 增强数据集详细介绍

数据数量	6000 栋建筑全年逐日数据	
特征	建筑负荷相关变量	制冷设定温度
		制热设定温度
		人员密度
		设备照明密度
		冷风渗透率
		体形系数
		太阳得热系数
		窗墙比
		内遮阳开启程度
	系统相关变量	风系统类型
		水系统类型

		主机 COP
		冷冻水供回水温差
		热水供回水温差
		冷冻水温度
		热水温度
		风机效率
		水泵效率
	天气参数	干球温度
		相对湿度
		风速
	时间标签	每年的月
		每月的日
		星期几
		是否是工作日
能耗	全年逐日单位面积能耗	

5.2.3 模型设置

模型的输入为建筑静态参数、天气参数以及时间标签。为了利用 LSTM 模型捕获时间序列中的长期依赖关系的优势，单栋建筑的训练数据输入处理为 n*365 维，其中 n 为特征数，输出为全年逐日单位面积能耗。训练集、验证集、测试集的划分比例为 6：2：2，其中训练集和验证集用于模型训练以及超参优化，测试集用于评估模型预测性能。

模型训练中，优化器选择 Adam。Adam 优化器结合了动量法和自适应学习率技术来进行梯度估计修正和梯度方向优化，在深度学习模型中一直有较好的表现。训练损失函数为 MSELoss。为了加快收敛减少震荡，训练中使用学习率衰减技术，训练进行到 50%和 75%，学习率执行一次衰减，衰减率为 0.9（学习率衰减为原来的十分之一）。

5.2.4 超参优化

本文基于 Ray.tune 对 LSTM 的超参数进行寻优。Ray.tune 是一个标准的超参优化工具，它基于 Ray 分布式计算框架，支持 Pytorch、Tensorflow、Keras 等多个训练框架。Ray.tune 进行超参优化的过程即是在一定的搜索空间内寻找最优的一组超参数，使得该组参数对应的训练任务有最大的准确率。LSTM 模型的隐层数量、层数、drop out 大小影响模型的结构设置；学习率影响模型收敛速度，过大则容易产生震荡，过小则收敛过慢；批量大小不影响随机梯度的期望，但影响随机梯度的方差，批量越大，随机梯度的方差越小，引入噪声也越小，训练越稳定，计算效率就高，但批量小，模型泛化就会更好；优化器的权重衰减参数主要作用是在损失函数中加入 L2 正则化项，来避免过拟合。因此，本文使用 Ray.tune 对 LSTM 模型的隐层数量、层数、drop out 大小、学习率、批次大小以及优化器的权重衰减进行组合寻优。优化空间采样次数为 200，单次训练的最大学习步长为 120 次，参数寻优的结果如表 5.2 所示。

表 5.2 LSTM 模型进行优化的超参数

参数	隐层大小	堆叠 lstm 层数	drop out	学习率	批次大小	权重衰减
搜索空间	[26,24,20,16,12,10,8,6,5,4]	[1,2,3]	[0.1,0.2,0.3,0.4,0.5]	Loguniform (1e-5,1e-2)	[16,32,64,128]	[0.0002,0.0004,0.0006]
寻优结果	20	2	0.2	0.00944	16	0.0002

5.2.5 模型预测性能

图 5.4 展示了模型在经过超参数优化后的训练收敛曲线。观察结果表明，模型在 120 次迭代步骤后达到基本收敛。具体来说，模型在训练集上的均方误差（MSE）为 0.01500，而在验证集和测试集上的均方误差分别为 0.01945 和 0.01813。图 5.5 展示了模型在三类增强数据上预测效果，可以看出三类数据预测精度均较高，模型在整个增强数据集上具有稳健的泛化能力。

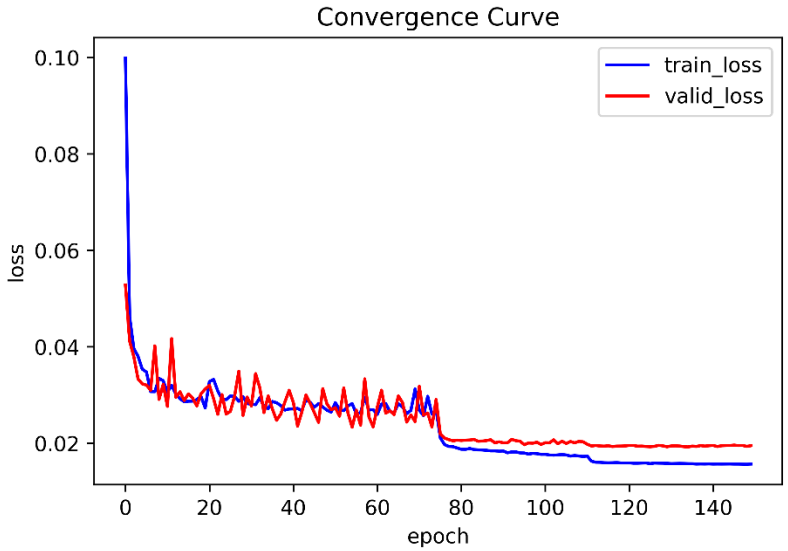
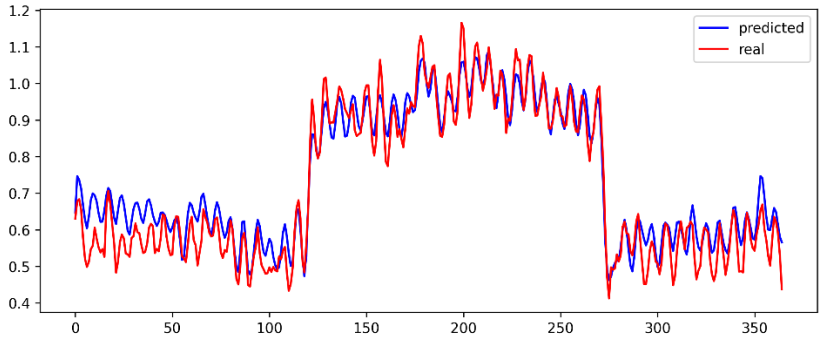
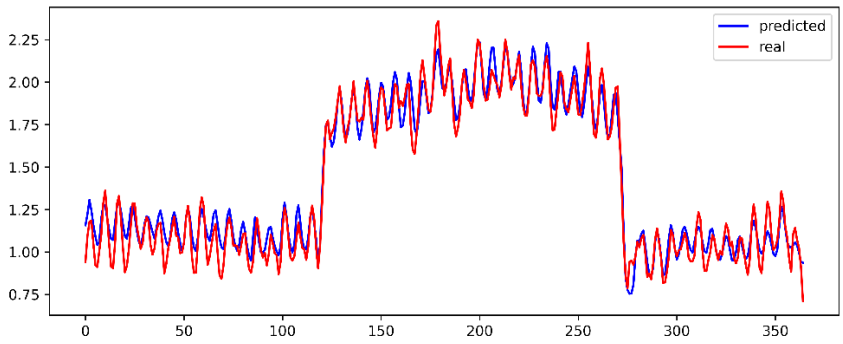


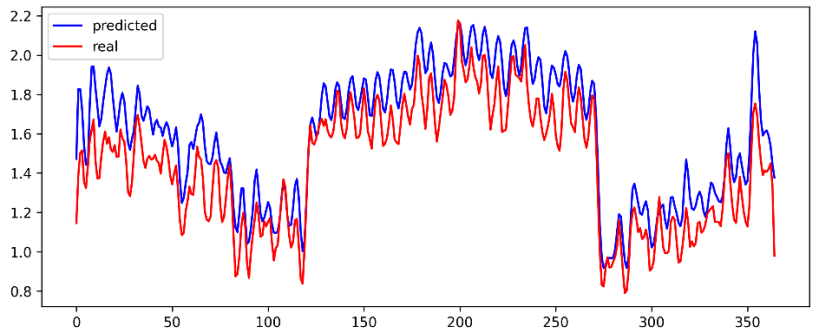
图 5.4 训练过程收敛曲线



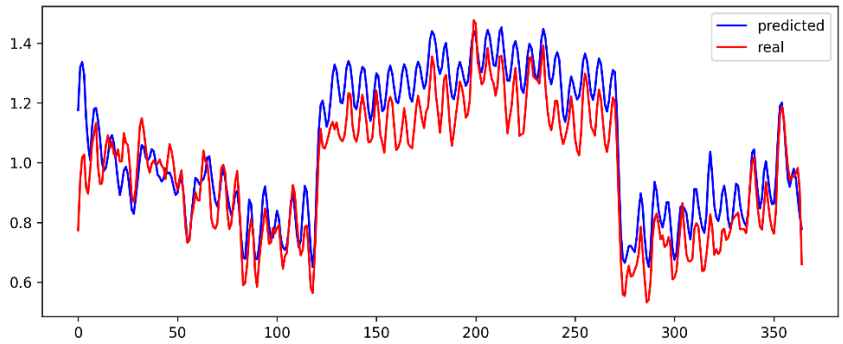
（a）



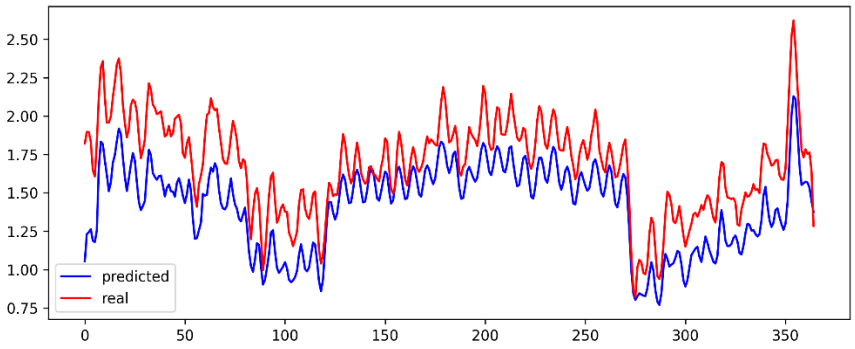
（b）


predicted
real
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0
50
100
150
200
250
300
350

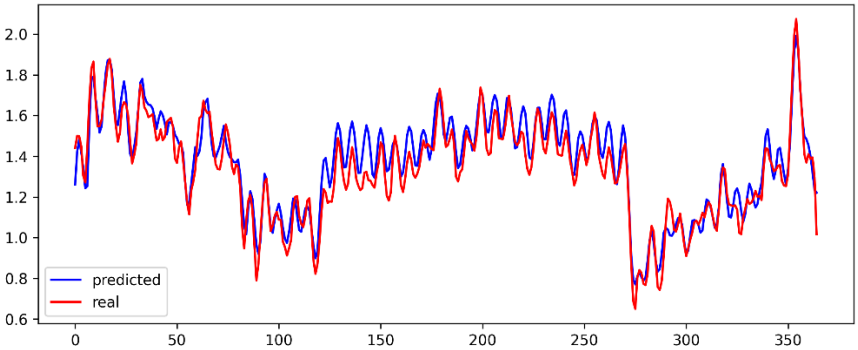
（c）


predicted
real
1.4
1.2
1.0
0.8
0.6
0
50
100
150
200
250
300
350

（d）


predicted
real
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0
50
100
150
200
250
300
350

（e）


predicted
real
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0
50
100
150
200
250
300
350

（f）

图 5.5 模型预测结果示例

5.3 基于真实数据集的预训练模型调优

5.3.1 数据来源

本研究所使用的建筑能耗数据源自上海市某能耗监测平台，其页面如图 5.6 所示。能耗检测平台所提供的数据主要包括建筑基本参数（包括建筑编号、地上层数、地下层数、建筑面积、建筑类型、空调系统形式），设备清单列表，气象参数，度日数，分项计量能耗（动力、照明、空调和其他能耗）等。其中可提供的能耗数据的颗粒度有每 15min、逐时、逐日和逐月。该能耗监测平台提供了 2017 年至 2023 年的能耗监测数据。本文所使用的数据集涵盖 112 栋建筑综合体，主要囊括了 2017 年、2018 年和 2019 年的能耗数据，其他年份数据考虑到疫情以及时间久远等因素不予采用。由于监测平台所提供的来自各设备的逐时分项能耗数据差，存在大量异常，本文使用逐日能耗数据。

图 5.6 能耗监测平台页面

针对存在特征缺失的建筑数据，本研究采用了第二章所述的基于遗传算法的关键变量推断方法以推断缺失值。对真实数据集做了同增强数据集相同的处理：对这部分建筑的特征数据执行了-1 至 1 的归一化处理（参见公式 5-1），并将能耗数据转换为单位面积能耗。数据集按照 7：3 的比例被划分为训练集和测试集。其中，训练集用于模型参数的训练，而测试集则用于评估模型的性能。

$$x_i = \frac{x_i - \frac{(x_i)max - (x_i)min}{2}}{(x_i)max - (x_i)min} \tag{5-1}$$

5.3.2 模型迁移

本研究采用了迁移预训练模型结构的迁移策略。预训练模型的参数被用作迁移模型的初始化参数，随后使用实测建筑数据集对模型的参数进行微调。

1. 模型设置

模型选择了 Adam 优化器，并采用绝对误差损失函数（L1Loss）。相较于均方误差损失，绝对误差损失在预测性能上展现出更佳的效果。考虑到数据集的规模以及模型迁移所涉及的有限超参数，对迁移模型的超参数进行了手动调整。经过优化，模型的批次大小定为 16，学习率设置为 0.0005，权重衰减系数为 0.0003。此外，为了加快收敛减少震荡，本研究采用了学习率衰减策略，即每 50 个迭代步骤后将学习率降低至原来的十分之一。

2. 训练结果

如图 5.7 所示，迁移模型在训练过程中的收敛曲线表明，训练至 75 步后，训练误差和测试误差均趋于稳定，且二者之间的差距较小，这表明模型没有出现过拟合现象。最终，该模型在训练集上的平均绝对误差（MAE）为 0.04836，在测试集上的 MAE 为 0.05077，表明模型在预测建筑能耗方面具有良好的准确性。

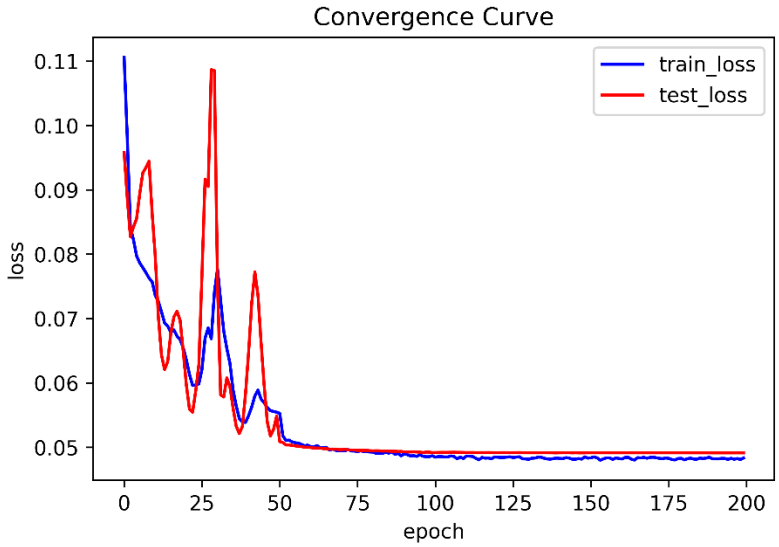


图 5.7 迁移模型收敛曲线

5.4 本章小结

在真实数据不足的情况下，仅使用少量数据训练得到的模型精度差、泛化能力差、容易过拟合。本文通过第四章生成了 6000 栋建筑的增强数据集，但是第

四章模型的训练基于模拟数据集，所获取的增强数据无法避免地同真实数据存在差别，那么下面需要解决的问题便是如何融合增强数据集和数据量较少的真实数据集。本章提出了基于迁移学习的能耗预测模型的建立过程，通过迁移学习融合增强数据和真实数据。首先基于增强数据集建立预训练模型，由于增强数据集数据充足，预训练模型的结构及超参经过优化，可以充分学习关键变量与能耗之间的非线性映射关系。随后将该模型的结构迁移到真实数据集，使用建筑实际能耗数据集对模型参数进行调整。最终建立的迁移模型在测试集上预测性能良好，MAE 为 0.05077。

第六章 模型验证

6.1 概述

第五章详细介绍了基于迁移学习的能耗预测模型的建立过程。本章节首先对本文所提出的基于迁移学习的能耗预测方法进行了验证，展示了模型在真实数据集上的预测表现以及在 5 栋真实建筑上的具体预测结果。其次通过对比迁移模型与基础 LSTM 模型在另外 5 栋真实建筑上的能耗预测结果，验证了本文所提出迁移策略的有效性。

本文使用平均绝对误差（MAE）和均方根误差（RMSE）两个指标来评价基础模型和迁移模型的性能。

平均绝对误差是预测值与实际值之间绝对偏差的平均值，计算公式如式 6-1 所示。平均绝对误差越小，说明预测值同真值之间的偏差越小。

$$MAE = \frac{1}{n}|y_i - f(x_i)| \tag{6-1}$$

式中，y_i为实际值，$f(x_i)$为对应预测值，n 为数据集中样本数。

均方根误差是预测值与实际值偏差平方的均值的平方根，计算公式如式 6-2 所示。均方根误差反映了样本的离散程度。均方根误差越小，说明预测值同真值之间的偏差越小。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2} \tag{6-2}$$

6.2 迁移模型预测结果

本文所建立的跨建筑能耗预测模型在对一栋新建筑进行预测时，无需重新训练模型。迁移模型在训练集上的平均绝对误差（MAE）为 0.04836，在测试集上的 MAE 为 0.05077，表明模型在预测建筑能耗方面具有良好的准确性。图 6.1 为迁移模型在测试集上的绝对误差分布图，图中显示迁移模型在测试集上的 MAE 均在 0.1 以内，集中分布在 0.04 左右。

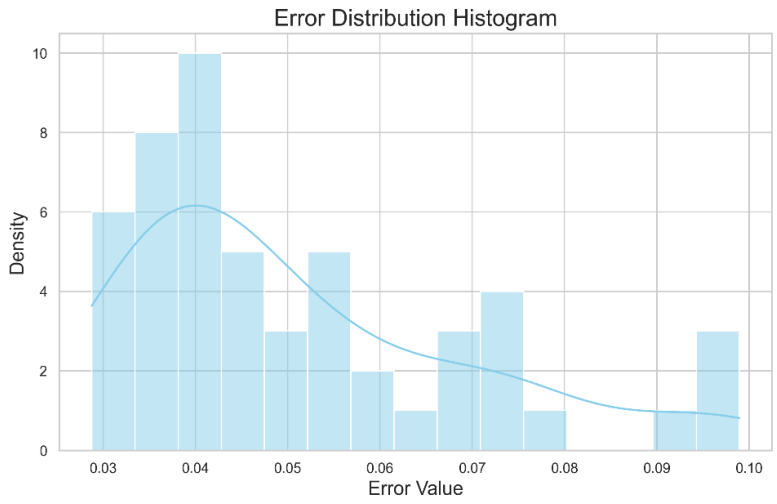


图 6.1 迁移模型测试集误差分布图

表 6.1 为验证迁移模型预测性能的 5 栋真实建筑的基本概况。表 6.2 和图 6.2 展示了迁移模型在五栋建筑上的全年能耗预测结果，可以看出，迁移模型在这五栋建筑上的预测效果较好，在建筑 1 上的预测效果最差，但是其对应 MAE 也仅为 0.03597。此外，通过图 6.2 还可以发现，以建筑 4 为例，在 12 月和 1 月之间存在 3-5 天能耗低谷，本研究所构建的迁移模型未能有效捕捉这一异常用能情况。这一局限性源于模型输入特征中未包含能够反映异常用能的相关信息。本研究所建模型主要通过时间标签间接表征建筑内的人员占用情况，但这些时间标签无法捕捉到异常的占用模式。例如，如果某栋建筑内的公司安排了集体出游活动，在出游期间建筑的能耗可能会显著下降，而本研究构建的预测模型无法准确预测此类情况下的能耗变化。

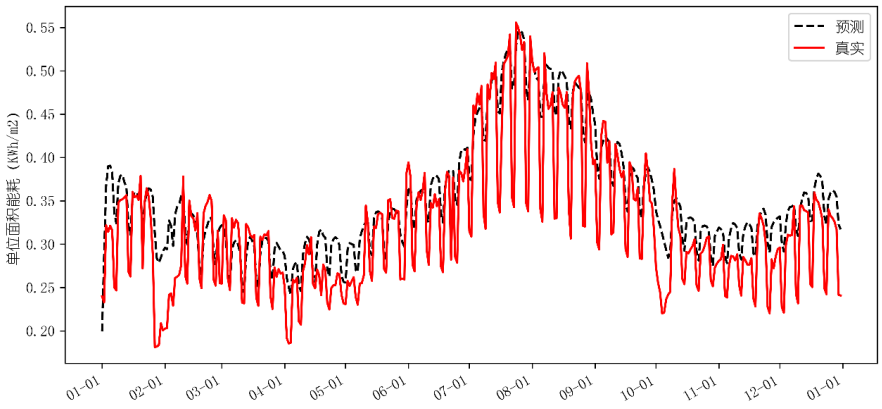
表 6.1 建筑基本概况

建筑编号	建筑面积（m^2）	建筑类型	建筑业态
建筑 1	106759	综合体	1 至 7 层为商场，集中式全空气系统；8 至 31 层为办公区，风机盘管系统。
建筑 2	79822	综合体	1 至 7 层为商场，集中式全空气系统；8 至 31 层为办公区，风机盘管系统。
建筑 3	96485	综合体	总共 20 层，风机盘管系统。
建筑 4	134332	综合体	1 至 7 层为商场，集中式全空气系统；8 至 16 层为办公

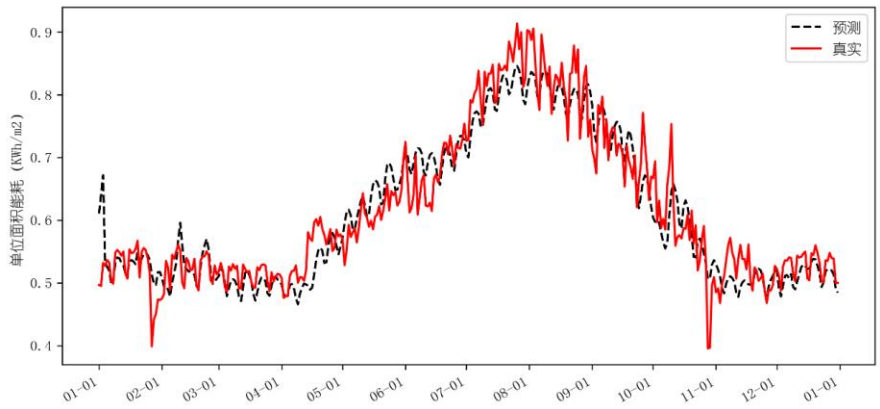
建筑 5	254000	综合体	区，分体式空调或 VRV 的局部式机组系统。1 至 9 层为商场，集中式全空气系统；10 至 35 层为办公区，风机盘管系统。

表 6.2 迁移模型在五栋建筑上的预测结果

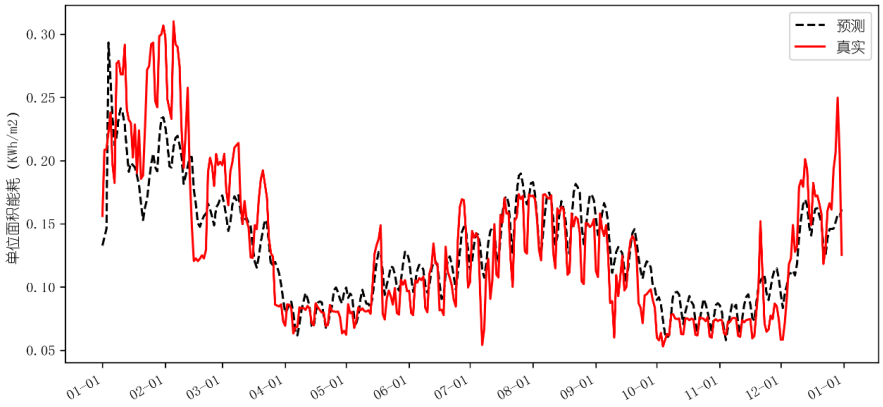
建筑编号	MAE	RMSE
建筑 1	0.03597	0.04661
建筑 2	0.03005	0.03932
建筑 3	0.02176	0.02877
建筑 4	0.02314	0.03580
建筑 5	0.02704	0.03482



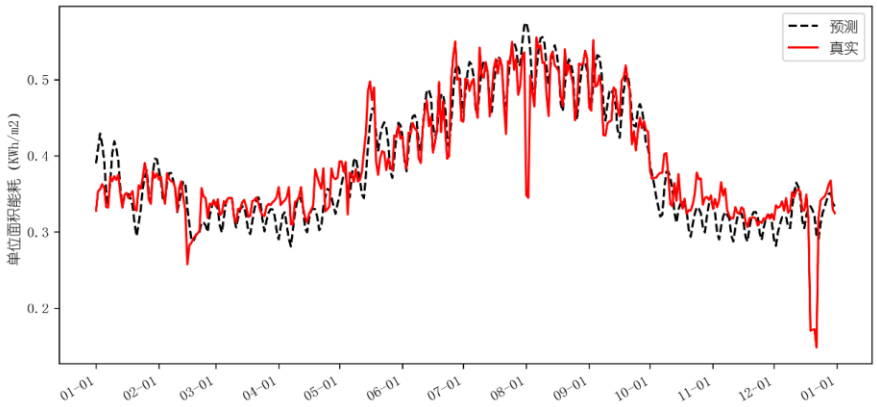
(a)建筑 1



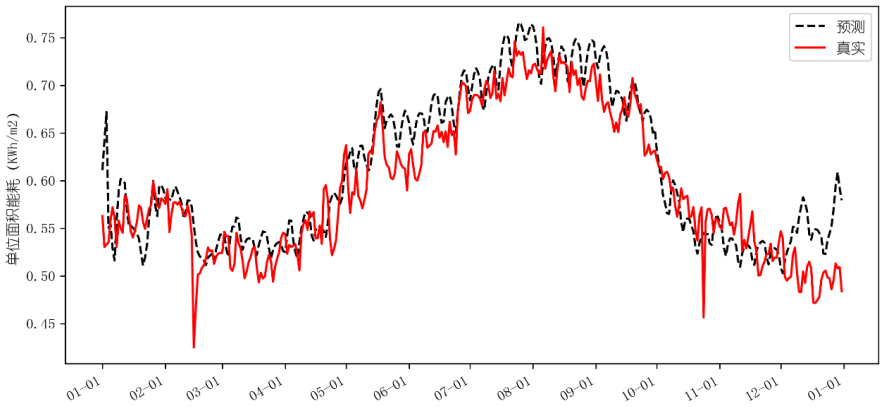
(b)建筑 2



(c)建筑 3



(d)建筑 4



(e)建筑 5

图 6.2 迁移模型在五栋建筑上的能耗预测结果

6.3 迁移模型预测不确定性

不确定性是指知识有限，无法准确描述状态的一种状态，广泛存在于工程过程中，如预测过程、建模、控制和管理[88]。Wang[89]将建筑能耗模拟领域的不确定性分为参数不确定性和模型不确定性。Yu 等[90]引入了三种能源预测模型的不

确定性来源：人为、建筑和天气因素。本文所建立的能耗预测模型的不确定性主要包括输入不确定性和模型不确定性。其中输入不确定性是指预测模型的输入并不一定同建筑真实情况完全相符。模型不确定性又称为认知不确定性，是由于模型的参数不确定。目前建筑能耗领域不确定分析中对输入参数不确定量化的研究较多，但绝大部分研究基于蒙特卡洛方法，即事先为每个输入变量定义一个分布，通过抽样技术运行大量模拟来获得输出分布。

本研究采用如下方法衡量建筑能耗预测的不确定性：首先计算模型测试集样本误差的标准差，标准差代表了模型预测的不确定性水平，即模型的预测偏离真实值的平均程度；其次确定置信水平，本文选择置信水平为 95%；将标准差乘以所选置信水平对应的临界值，以计算置信区间的宽度；最后，将计算得到的置信区间宽度添加到模型的预测结果中。图 6.3 为模型预测结果的不确定性分析，其中橙色曲线为建筑真实能耗，蓝色曲线为模型预测值灰色区域为模型预测结果的 95%置信区间。

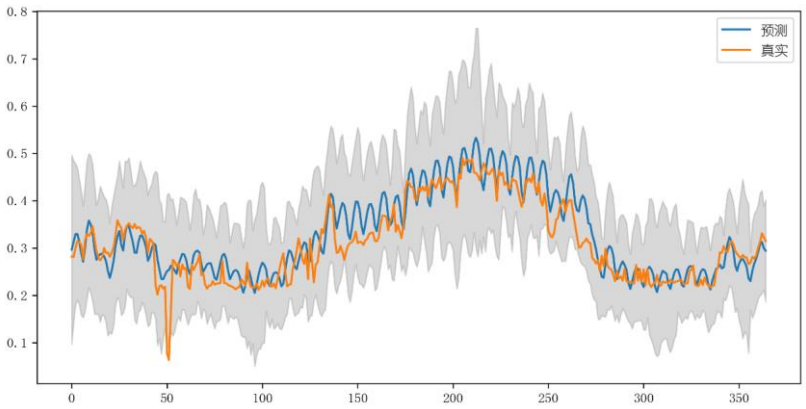


图 6.3 模型预测结果的 95%置信区间

6.4 迁移模型和基础模型的对比验证

为了验证本文所提出的迁移策略的有效性，本文基于实测数据集建立了基础预测模型，并对比迁移模型和基础模型的预测效果。基础模型所用算法与迁移模型相同，均为长短时记忆网络；模型输入与迁移模型相同，包括建筑静态参数、天气参数以及时间标签，输出为全年逐日单位面积能耗。与迁移模型不同的是，基础模型的全程训练仅使用真实数据集。

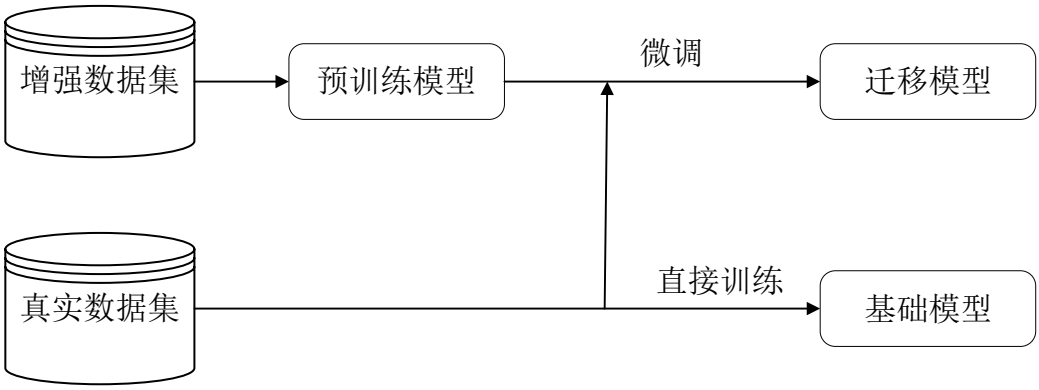


图 6.4 迁移模型和基础模型的对比

6.4.1 基础模型的建立

1. 数据集划分

真实数据集在 5.3.1 已详细介绍过，这里不再赘述。建立基础模型时，真实数据集按照 6：2：2 的比例划分为训练集、验证集、测试集，其中训练集和验证集用于模型训练以及超参优化，测试集用于评估模型预测性能。

2. 模型设置

模型优化器选择 Adam。训练损失函数为 L1Loss，L1Loss 是 pytorch 中的绝对误差损失，它计算的是预测值和真实值之间的绝对差的平均值。为了加快收敛减少震荡，训练中使用学习率衰减技术，训练进行到 50%和 75%，学习率执行一次衰减，衰减率为 0.9（学习率衰减为原来的十分之一）。

3. 超参优化

本文基于 Ray.tune 对基础预测模型的超参数进行寻优。优化空间采样次数为 200，单次训练的最大学习步长为 120 次，参数寻优的结果如表 6.3 所示。

表 6.3 基础模型超参寻优结果

参数	隐层大小	堆叠 lstm 层数	drop out	学习率	批次大小	权重衰减
搜索空间	[26,24,20,16,12,10,8,6,5,4]	[1,2,3]	[0.1,0.2,0.3,0.4,0.5]	Loguniform (1e-5,1e-2)	[16,32,64,128]	[0.0002,0.0004,0.0006]
寻优结果	24	3	0.1	0.00270	16	0.0002

6.4.2 基础模型和迁移模型预测结果对比

1. 整体预测性能比较

表 6.4 展示了基础模型和迁移模型在训练集及测试集上的预测性能。从表中

可以看出迁移模型在测试集上的平均 MAE 比基础模型低了 9.0%，迁移模型的预测准确性更高。此外，基础模型训练误差和验证误差差别较大，训练集平均 MAE 比测试集低了 30.5%，模型有过拟合趋势。这可能是由于数据集额外划分了验证集用于模型调参，导致训练数据量减少。图 6.5 为基础模型在测试集上的绝对误差分布图，图中显示基础模型在测试集上的 MAE 在 0.15 以内，而迁移模型在测试集上的 MAE 均在 0.1 以内，这表明基础模型的预测鲁棒性更差。基础模型在测试集上的 MAE 集中分布在 0.05 左右，迁移模型误差集中分布在 0.04 左右，迁移模型的预测精度更高。

表 6.4 基础模型和迁移模型预测 MAE 对比

模型	训练集	验证集	测试集
基础模型的平均 MAE	0.03877	0.05917	0.05579
迁移模型的平均 MAE	0.04836	/	0.05077

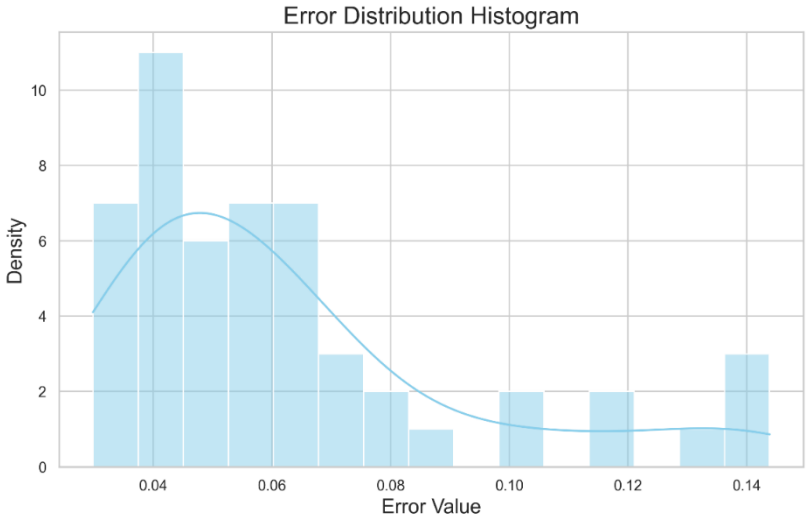


图 6.5 基础模型测试集误差分布

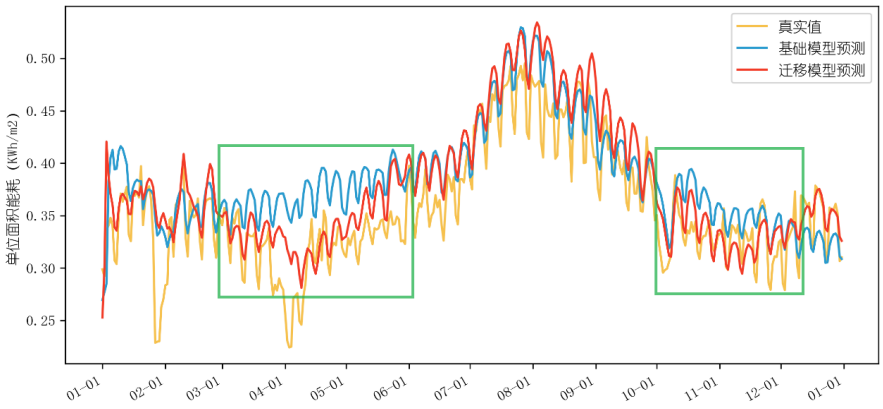
2. 典型建筑预测性能比较

为了直观展示基础模型和迁移模型在相同建筑上的预测性能差异，表 6.6 和图 6.6 对比了迁移模型和基础模型在 5 栋真实建筑上的全年能耗预测表现。为了区别于 6.2 节中的 5 栋建筑，这里把建筑编号为 a 到 e。表 6.5 为 5 栋建筑的基本信息概况。从图 6.6 和表 6.6 可以看出，迁移模型在 5 栋真实建筑上的 MAE 和 RMSE 均比基础模型低，其预测性能优于基础模型。迁移模型在 5 栋建筑上的 MAE 和 RMSE 均低于 0.1。此外，不难看出相比于其他四栋建筑，迁移模型和基础模型在建筑 b 上的预测效果较差，这可能是由于建筑 b 办公区对外租赁，具体租赁情况不太稳定导致用能波动较大，进而给预测带来困难。即使在这种情况下，

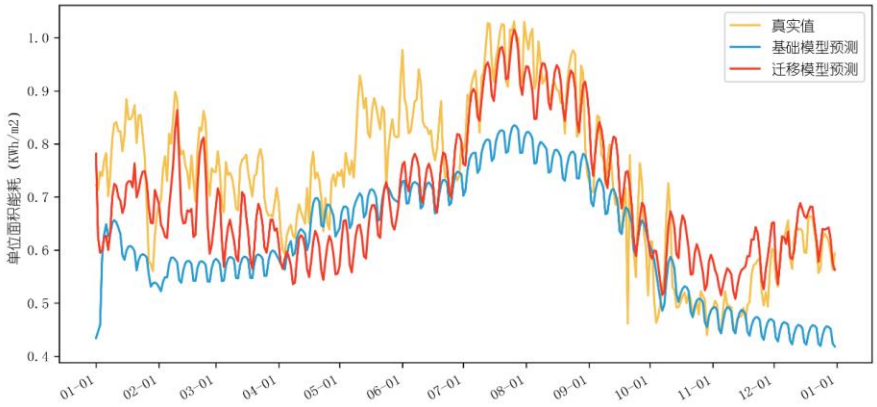
迁移模型在建筑 b 上的 MAE 和 RMSE 均在 0.1 以内，而基础模型的 MAE 和 RMSE 均超过 0.1。

表 6.5 建筑基本概况

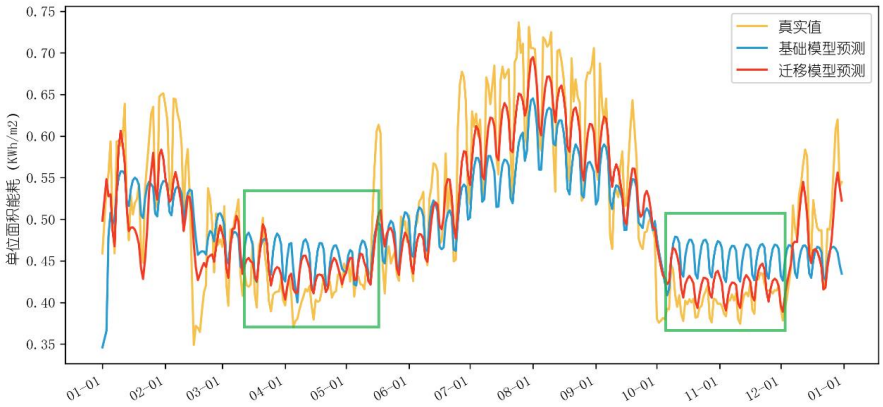
建筑编号	建筑面积（m2）	建筑类型	建筑业态
建筑 a	79245	综合体	1 至 7 层商场，集中式全空气系统；8 至 24 层为办公楼，风机盘管系统。
建筑 b	33986	综合体	1 至 6 层为商场，集中式全空气系统；7 至 30 层为办公区，风机盘管系统。
建筑 c	56475	综合体	1 至 6 层为商场，集中式全空气系统；7 至 25 层为办公区，风机盘管系统。
建筑 d	90000	综合体	1 至 7 层商场，集中式全空气系统；8 至 16 为办公区，多联机系统。
建筑 e	43680	综合体	1 至 6 层为商场，集中式全空气系统；7 至 24 层为办公区，风机盘管系统。



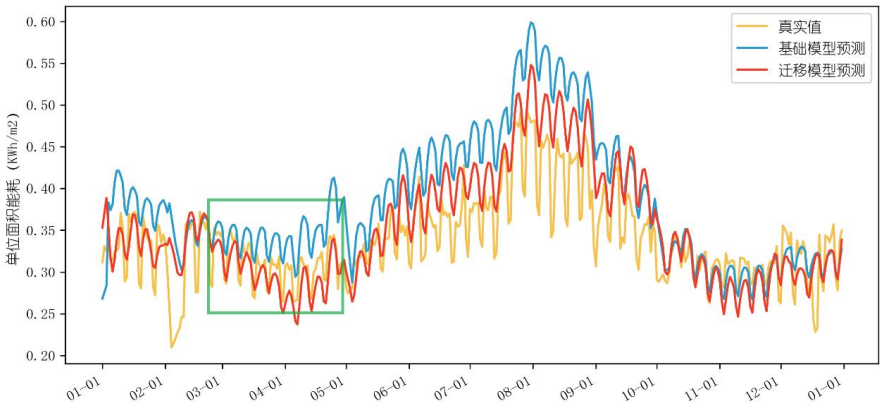
(a)建筑 a



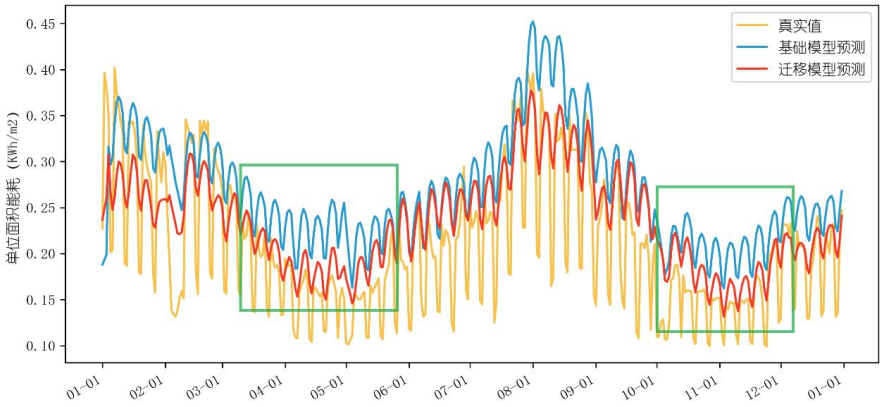
(b)建筑 b



(c)建筑 c



(d)建筑 d



(e)建筑 e

图 6.7 迁移模型和基础模型在 5 栋建筑上的预测结果比对

表 6.6 迁移模型和基础模型在 5 栋建筑上的预测误差对比

建筑编号	迁移模型		基础模型	
	MAE	RMSE	MAE	RMSE
建筑 a	0.02687	0.03416	0.03400	0.04172
建筑 b	0.07812	0.09638	0.1190	0.1408
建筑 c	0.03376	0.04247	0.05107	0.06213
建筑 d	0.03041	0.03933	0.05078	0.06326
建筑 e	0.04277	0.05305	0.06526	0.07512

此外，观察图 6.7 中的绿色矩形框选部分发现，相比基础模型，迁移模型在过渡季的预测曲线同建筑的真实曲线更加贴合，这表明迁移模型相比基础模型在过渡季表现更优。为了定量证明这一结论，表 6.7~表 6.11 统计了建筑 a~建筑 e 逐月的 MAE 和 RMSE。从表中可以看出，迁移模型相比基础模型性能的提升，主要来源于过渡季。具体来说，对于建筑 a，迁移模型在 3 至 5 月和 10 月、11 月的预测误差低于基础模型，4 月份迁移模型的 MAE 甚至比基础模型低了 60.7%；对于建筑 c，迁移模型在 3 至 5 月和 10 月、11 月的 MAE 和 RMSE 比基础模型低了 30%左右；对于建筑 d，迁移模型在 3 至 5 月预测误差小于基础模型，虽前者在 10 月、11 月预测误差高于后者，但 MAE 和 RMSE 差距均在 0.001 左右；对于建筑 e，迁移模型在 3 至 5 月和 10 月、11 月的 MAE 和 RMSE 均比基础模型低了 30%以上。对于建筑 b，迁移模型在 4、5 月份和 10、11 月份的预测误差反而比基础模型稍大，但在其他月份的预测误差小于基础模型，这很可能也是由于建筑 b 用户复杂，内部用能波动复杂导致的。

表 6.7 建筑 a 逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1 月	0.03320	0.04916	0.04784	0.06365	0.01464	0.01449
2 月	0.02121	0.02869	0.01513	0.01833	-0.00608	-0.01036
3 月	0.02244	0.02764	0.04152	0.04824	0.01908	0.0206
4 月	0.02794	0.03689	0.07108	0.07636	0.04314	0.03947
5 月	0.03226	0.03542	0.05038	0.05167	0.01812	0.01625
6 月	0.03274	0.03469	0.03230	0.03519	-0.00044	0.0005
7 月	0.02600	0.03011	0.02133	0.02616	-0.00467	-0.00395
8 月	0.04377	0.05100	0.03056	0.03586	-0.01321	-0.01514
9 月	0.03198	0.03679	0.01825	0.02137	-0.01373	-0.01542
10 月	0.01916	0.02259	0.03722	0.04016	0.01806	0.01757
11 月	0.01572	0.01887	0.02236	0.02560	0.00664	0.00673
12 月	0.01545	0.02006	0.02304	0.02731	0.00759	0.00725

表 6.8 建筑 b 逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1 月	0.10750	0.11270	0.19471	0.22243	0.08721	0.10973
2 月	0.08265	0.09147	0.21566	0.22078	0.13301	0.12931
3 月	0.08807	0.09448	0.15196	0.15637	0.06389	0.06189
4 月	0.09192	0.10475	0.05145	0.05938	-0.04047	-0.04537
5 月	0.1756	0.18634	0.13271	0.14326	-0.04289	-0.04308
6 月	0.08799	0.10516	0.10319	0.12456	0.0152	0.0194
7 月	0.03583	0.04321	0.15625	0.16096	0.12042	0.11775
8 月	0.04514	0.05768	0.13790	0.14517	0.09276	0.08749
9 月	0.06045	0.07996	0.05647	0.07411	-0.00398	-0.00585
10 月	0.08409	0.09367	0.03429	0.04994	-0.0498	-0.04373
11 月	0.06232	0.06587	0.04212	0.05980	-0.0202	-0.00607
12 月	0.02371	0.02947	0.16547	0.16758	0.14176	0.13811

表 6.9 建筑 c 逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1 月	0.04780	0.06757	0.06395	0.09118	0.01615	0.02361

2 月	0.05441	0.06296	0.06125	0.07132	0.00684	0.00836
3 月	0.02272	0.02491	0.03421	0.03911	0.01149	0.0142
4 月	0.02418	0.02881	0.04011	0.04470	0.01593	0.01589
5 月	0.03211	0.04461	0.03465	0.04702	0.00254	0.00241
6 月	0.03980	0.05031	0.03918	0.05746	-0.00062	0.00715
7 月	0.04009	0.04737	0.07519	0.08611	0.0351	0.03874
8 月	0.04326	0.05052	0.06980	0.07941	0.02654	0.02889
9 月	0.03581	0.04311	0.03564	0.04492	-0.00017	0.00181
10 月	0.03345	0.03865	0.05774	0.06091	0.02429	0.02226
11 月	0.01355	0.01554	0.04691	0.04887	0.03336	0.03333
12 月	0.02823	0.03509	0.06062	0.07169	0.03239	0.0366

表 6.10 建筑 d 逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1 月	0.03427	0.07138	0.05654	0.06767	0.02227	-0.00371
2 月	0.03820	0.05136	0.04618	0.06384	0.00798	0.01248
3 月	0.01497	0.01840	0.02726	0.03118	0.01229	0.01278
4 月	0.02144	0.02572	0.04796	0.05086	0.02652	0.02514
5 月	0.01947	0.02571	0.04764	0.05372	0.02817	0.02801
6 月	0.03961	0.04485	0.07764	0.08073	0.03803	0.03588
7 月	0.03628	0.04227	0.08873	0.09163	0.05245	0.04936
8 月	0.06363	0.06937	0.1145	0.1178	0.05087	0.04843
9 月	0.04330	0.05041	0.05400	0.05904	0.0107	0.00863
10 月	0.01649	0.02452	0.01614	0.02370	-0.00035	-0.00082
11 月	0.02364	0.02661	0.01429	0.01573	-0.00935	-0.01088
12 月	0.02789	0.03503	0.02578	0.03437	-0.00211	-0.00066

表 6.11 建筑 e 逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1 月	0.06449	0.09378	0.06327	0.07932	-0.00122	-0.01446
2 月	0.05927	0.06619	0.06476	0.08374	0.00549	0.01755
3 月	0.03176	0.03921	0.05608	0.06173	0.02432	0.02252
4 月	0.02873	0.03357	0.07475	0.07620	0.04602	0.04263

5 月	0.04548	0.05021	0.06730	0.07167	0.02182	0.02146
6 月	0.05230	0.06204	0.05777	0.06487	0.00547	0.00283
7 月	0.04386	0.05423	0.06100	0.07215	0.01714	0.01792
8 月	0.04997	0.06563	0.09513	0.10976	0.04516	0.04413
9 月	0.05117	0.06397	0.06754	0.07630	0.01637	0.01233
10 月	0.04514	0.05107	0.06784	0.07091	0.0227	0.01984
11 月	0.02421	0.02939	0.05741	0.06001	0.0332	0.03062
12 月	0.02451	0.03393	0.04959	0.05694	0.02508	0.02301

6.4 本章小结

为了验证本文所建立的基于迁移学习的预测模型的优越性，本章详细介绍了迁移模型在真实数据集上的预测表现，并与直接在真实数据集上训练的基础 LSTM 模型进行对比。验证结果显示，本文所提出的基于迁移学习的能耗预测模型在测试集上的平均 MAE 仅为 0.05077，比基础模型低了 9.0%。并且迁移模型在测试集上的误差分布显示，MAE 均在 0.1 以内，模型预测鲁棒性优于基础模型。此外，通过逐月误差比对，迁移模型相比基础模型，性能提升主要来源于过渡季。综上，本文所提出的基于迁移学习的公共建筑综合体能耗预测方法有效性得到验证。

第七章 结论与展望

7.1 主要结论

本文提出了基于迁移学习的建筑综合体能耗预测方法，从众多影响建筑能耗的因素中提取了关键因素，并提出了基于遗传算法和自编码器的缺失关键变量推断算法。利用数据增强技术获取增强数据集，通过迁移学习融合增强数据集和真实数据集，建立了预测精度较高的综合体全年能耗预测模型。下文将逐一介绍本研究的结论与成果：

1）本文第二章提取了影响综合体能耗的关键变量。建筑能耗受诸多因素影响，本文将其划分为建筑负荷相关和空调系统相关两大类，通过 SRRC、PRCC 以及随机森林的方法进行敏感性分析，提取影响综合体能耗的关键变量作为所建立预测模型的输入。经过敏感性分析，最终选定的负荷相关关键变量包括制冷设定温度、制热设定温度、人员密度、照明设备密度、冷风渗透率、体形系数、太阳得热系数、窗墙比和内遮阳开启程度；最终选定的系统相关关键变量包括风系统类型、水系统类型、主机 COP、冷冻水供回水温差、热水供回水温差、冷冻水温度、热水温度、风机效率和水泵效率。

2）本文第三章提出了基于遗传算法和基于降噪自编码器的关键变量缺失推断算法。对于有历史能耗的建筑，基于历史能耗与关键变量之间的关联关系，使用遗传算法推断最符合历史能耗趋势的缺失关键变量值。对于没有历史能耗的建筑，使用降噪自编码器挖掘各关键变量之间的关系进行缺失值推断。验证结果显示，基于遗传算法的关键变量推断算法在单变量缺失推断中平均误差均在 15%以内，多变量推断误差均值在 20%以内，但存在离群值；基于自编码器的关键变量推断算法在单变量缺失推断和多变量缺失推断中均值误差均在 20%以内。

3）本文第四章提出了基于条件生成对抗网络的数据增强方法。把归一化后的模拟数据聚成 3 类，基于每一类数据建立不同的生成对抗网络模型，建筑关键变量作为生成器的条件。使用训练完成的生成器进行建筑能耗趋势曲线生成，并利用 XGBoost 模型预测的建筑全年总能耗填充能耗趋势曲线，完成数据增强过程。可视化生成器生成效果并与原模拟能耗对比显示，生成器生成效果优异。

4）本文第五章建立了基于迁移学习的建筑能耗预测模型，融合了增强数据集和真实数据集。首先基于增强数据集训练 LSTM 预训练模型，接着把

该预训练模型迁移到真实数据集，使用真实数据对该预训练模型的参数进行微调。该模型在训练集上的平均绝对误差（MAE）为 0.04836，在测试集上的 MAE 在 0.1 以内，集中分布在 0.04 左右，均值为 0.05077。经过对比，基于迁移学习所建立的模型优于仅使用真实数据训练的基础模型，前者在测试集上的预测误差比后者低了 9.0%。此外逐月预测误差比较显示，迁移模型在过渡季预测表现优于基础模型。

7.2 主要贡献

1) 本文提出了基于条件生成对抗网络的数据增强方法，利用聚类来保证了生成对抗网络模型的性能。本文提出的数据增强方法，可以快速生成大量建筑能耗数据，相比白箱模拟生成数据更加高效，解决了数据驱动预测模型数据不足的问题。
2) 本文建立了基于迁移学习的建筑能耗预测模型。利用迁移学习融合了增强数据和真实数据，缓解真实数据缺失导致深度学习预测模型效果差的现象。此外通过迁移学习微调的方式，有效避免了模拟数据存在偏差，直接融合影响预测模型效果的问题。

7.3 研究的局限性与展望

由于数据量和研究时间的限制，本研究仍有一些值得改进的地方：

1) 建筑空调系统的运行是一个非常复杂的问题，系统的连接方式及控制逻辑难以参数化表示，故本文的初始变量集仅根据专家知识选择，无法囊括所有建筑能耗影响因素。
2) 本文的预测模型无法准确预测突发情况下的建筑能耗。本文在进行初始变量集的选取时，主要包括建筑负荷相关变量和空调系统相关变量两大类，并未包括一些非常规因素，比如建筑由于一些不可控因素临时停业、疫情关闭等，因此该模型无法准确预测这种突发情况下的能耗。
3) 由于真实数据缺失，本研究在第三章基于遗传算法的关键变量推断中所使用的能耗预测模型是利用模拟数据集训练得到的，这给对推断结果产生了负面影响。因为模拟数据同真实数据有偏差，那么基于模拟数据所训练得到的预测模型学习到的是模拟数据中关键变量和能耗之间的非线性关系，这跟实际情况可能会存在偏差。在进行关键变量推断时，依据的正是模型中所学到的关键变量和能耗之间的关系，这会对推断结果的准确性带来负面影响。

参考文献

[1] 中国建筑节能协会能耗专委会,中国建筑能耗研究报告 2022, 2022.
[2] PéREZ-LOMBARD L, ORTIZ J, POUT C. A review on buildings energy consumption information [J]. Energy and Buildings, 2008, 40(3): 394-8.
[3] 支建杰,吴蔚沁. 公共建筑能耗监测平台数据应用的探讨[C]//.2018, 城市发展与规划论文集.[出版者不详], 2018:428-434.
[4] 吴蔚沁,王何斌,徐强等.2021 年上海市公共建筑能耗监测平台数据分析[J].上海节能,2022(09):1096-1104.
[5] ZHANG L, WEN J, LI Y, et al. A review of machine learning in building load prediction [J]. Applied Energy, 2021, 285.
[6]VIROTE J, NEVES-SILVA R. Stochastic models for building energy prediction based on occupant behavior assessment [J]. Energy and Buildings, 2012, 53: 183-93.
[7] ZHOU D, MA S, HAO J, et al. An electricity load forecasting model for Integrated Energy System based on BiGAN and transfer learning [J]. Energy Reports, 2020, 6: 3446-61.
[8] ALDUAILIJ M A, PETRI I, RANA O, et al. Forecasting peak energy demand for smart buildings [J]. The Journal of Supercomputing, 2020, 77(6): 6356-80.
[9] EDIGER V Ş, AKAR S. ARIMA forecasting of primary energy demand by fuel in Turkey [J]. Energy Policy, 2007, 35(3): 1701-8.
[10] ABDEL-AAL R E, AL-GARNI A Z. Forecasting monthly electric energy consumption in eastern Saudi Arabia using univariate time-series analysis [J]. Energy, 1997, 22(11): 1059-69.
[11] ZHOU R, PAN Y, HUANG Z, et al. Building Energy Use Prediction Using Time Series Analysis [Z]. 2013 IEEE 6th International Conference on Service-Oriented Computing and Applications. 2013: 309-13.10.1109/soca.2013.14
[12] LIANG, X., et al., Domain knowledge decomposition of building energy consumption and a hybrid data-driven model for 24-h ahead predictions[J]. Applied Energy, 2023. 344.
[13] CHOU, J.-S. and A.S. TELAGA. Real-time detection of anomalous power consumption [J]. Renewable and Sustainable Energy Reviews, 2014, 33: 400-11.
[14] GUO N, CHEN W, WANG M, et al. Appling an Improved Method Based on ARIMA Model to Predict the Short-Term Electricity Consumption Transmitted by the Internet of Things (IoT) [J]. Wireless Communications and Mobile Computing, 2021, 2021: 1-11.
[15] JORDAN M I, MITCHELL T M. Machine learning: Trends, perspectives, and prospects [J]. Science, 2015, 349(6245): 255-60.
[16] LIM H S, KIM G. Prediction model of Cooling Load considering time-lag for preemptive action in buildings [J]. Energy and Buildings, 2017, 151: 53-65.
[17] LI Q, MENG Q, CAI J, et al. Applying support vector machine to predict hourly cooling load in the building [J]. Applied Energy, 2009, 86(10): 2249-56.
[18] DING Y, ZHANG Q, YUAN T. Research on short-term and ultra-short-term cooling load prediction models for office buildings [J]. Energy and Buildings, 2017, 154: 254-67.
[19] AHMAD M W, MOURSHED M, REZGUI Y. Trees vs Neurons: Comparison between

random forest and ANN for high-resolution prediction of building energy consumption [J]. Energy and Buildings, 2017, 147: 77-89.
[20] SHA H, XU P, HU C, et al. A simplified HVAC energy prediction method based on degree-day [J]. Sustainable Cities and Society, 2019, 51.
[21]WANG Z, SRINIVASAN R S. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models [J], Renew. Sustain. Energy Rev. 75 (2017) 796–808.
[22] ROBINSON C, DILKINA B, HUBBS J, et al. Machine learning approaches for estimating commercial building energy consumption [J]. Applied Energy, 2017, 208: 889-904.
[23] DONG Z, LIU J, LIU B, et al. Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification [J]. Energy and Buildings, 2021, 241.
[24] ZHAO R, WEI D, RAN Y, et al. Building Cooling load prediction based on LightGBM [J]. IFAC-PapersOnLine, 2022, 55(11): 114-9.
[25] WANG Z, LIANG Z, ZENG R, et al. Identifying the optimal heterogeneous ensemble learning model for building energy prediction using the exhaustive search method [J]. Energy and Buildings, 2023, 281: 112763.
[26] KIM T-Y, CHO S-B. Predicting residential energy consumption using CNN-LSTM neural networks [J]. Energy, 2019, 182: 72-81.
[27] WANG Z, HONG T, PIETTE M A. Building thermal load prediction through shallow machine learning and deep learning [J]. Applied Energy, 2020, 263.
[28] SALA-CARDOSO E, DELGADO-PRIETO M, KAMPOUROPOULOS K, et al. Activity-aware HVAC power demand forecasting [J]. Energy and Buildings, 2018, 170: 15-24.
[29] ZHANG C, LI J, ZHAO Y, et al. A hybrid deep learning-based method for short-term building energy load prediction combined with an interpretation process [J]. Energy and Buildings, 2020, 225.
[30] LEI L, CHEN W, WU B, et al. A building energy consumption prediction model based on rough set theory and deep learning algorithms [J]. Energy and Buildings, 2021, 240.
[31] QIN Y, ZHAO M, LIN Q, et al. Data-Driven Building Energy Consumption Prediction Model Based on VMD-SA-DBN [J]. Mathematics, 2022, 10(17).
[32] YUAN Z, WANG W, WANG H, et al. Combination of cuckoo search and wavelet neural network for midterm building energy forecast [J]. Energy, 2020, 202.
[33] BUI D-K, NGUYEN T N, NGO T D, et al. An artificial neural network (ANN) expert system enhanced with the electromagnetism-based firefly algorithm (EFA) for predicting the energy consumption in buildings [J]. Energy, 2020, 190.
[34] QIAN F, GAO W, YANG Y, YU D. Potential analysis of the transfer learning model in short and medium-term forecasting of building HVAC energy consumption [J]. Energy. 2020;193.
[35] CHEPURKO N, MARCUS R, ZGRAGGEN E, et al. ARDA: automatic relational data augmentation for machine learning[J]. Proc VLDB Endowment 2020, 13(9): 1373-87.
[36] PEREZ L, WANG J. The effectiveness of data augmentation in image classification using deep learning[J]. 2017.
[37] NANDHINI ABIRAMI R, DURAI RAJ VINCENT P M, SRINIVASAN K, et al. Deep CNN and Deep GAN in Computational Visual Perception-Driven Image Analysis [J]. Complexity, 2021,

2021: 1-30.
[38] CAO X, WIPF D, WEN F, et al. A Practical Transfer Learning Algorithm for Face Verification [C]. proceedings of the 2013 IEEE International Conference on Computer Vision, F 1-8 Dec. 2013, 2013.
[39] SHORTEN C, KHOSHGOFTAAR T M, FURHT B. Text Data Augmentation for Deep Learning [J]. J Big Data, 2021, 8(1): 101.
[40] OTTONI A L C, DE AMORIM R M, NOVO M S, et al. Tuning of data augmentation hyperparameters in deep learning to building construction image classification with small datasets [J]. International Journal of Machine Learning and Cybernetics, 2023, 14(1): 171-86.
[41] TANG Z, LI S, KIM K S, et al. Multi-Output Gaussian Process-Based Data Augmentation for Multi-Building and Multi-Floor Indoor Localization; proceedings of the 2022 IEEE International Conference on Communications Workshops (ICC Workshops), F 16-20 May 2022, 2022 [C].
[42] LU Y, TIAN Z, ZHANG Q, et al. Data augmentation strategy for short-term heating load prediction model of residential building [J]. Energy, 2021, 235.
[43] AMASYALI K, EL-GOHARY N. Hybrid approach for energy consumption prediction: Coupling data-driven and physical approaches [J]. Energy and Buildings, 2022, 259.
[44] FANG H, TAN H, KOSONEN R, et al. Study of the Data Augmentation Approach for Building Energy Prediction beyond Historical Scenarios [J]. Buildings, 2023, 13(2).
[45] FAN C, CHEN M, TANG R, et al. A novel deep generative modeling-based data augmentation strategy for improving short-term building energy predictions [J]. Building Simulation, 2021, 15(2): 197-211.
[46] TIAN C, LI C, ZHANG G, et al. Data driven parallel prediction of building energy consumption using generative adversarial nets [J]. Energy and Buildings, 2019, 186: 230-43.
[47] BELLAGARDA A, CESARI S, ALIBERTI A, et al. Effectiveness of neural networks and transfer learning for indoor air-temperature forecasting [J]. Automation in Construction, 2022, 140.
[48] AHN Y, KIM B S. Prediction of building power consumption using transfer learning-based reference building and simulation dataset [J]. Energy and Buildings, 2022, 258.
[49] FANG X, GONG G, LI G, et al. Transferability investigation of a Sim2Real deep transfer learning framework for cross-building energy prediction [J]. Energy and Buildings, 2023, 287.
[50] 王晋东, 迁移学习简明手册[M], 北京, 中国科学院计算技术研究所, 2019.
[51] ZHUANG F, QI Z, DUAN K, et al. A comprehensive survey on transfer learning [J]. Proc IEEE 2021, 109(1): 43–76.
[52] PAN S J, YANG Q. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-59.
[53] RIBEIRO M, GROLINGER K, ELYAMANY H F, et al. Transfer learning with seasonal and trend adjustment for cross-building energy forecasting [J]. Energy and Buildings, 2018, 165: 352-63.
[54] FANG X, GONG G, LI G, et al. A hybrid deep transfer learning strategy for short term cross-building energy prediction [J]. Energy, 2021, 215.
[55] FAN C, SUN Y, XIAO F, et al. Statistical investigations of transfer learning-based methodology for short-term building energy predictions [J]. Applied Energy, 2020, 262.
[56] LU H, WU J, RUAN Y, et al. A multi-source transfer learning model based on LSTM and

domain adaptation for building energy prediction [J]. International Journal of Electrical Power & Energy Systems, 2023, 149.
[57] MCKAY M D, BECKMAN R J, CONOVER W J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code [J]. Technometrics, 2000, 42(1): 55-61.
[58] DING T-T, LIU S-S, WANG Z-J, et al. A novel mixture sampling strategy combining latin hypercube sampling with optimized one factor at a time method: A case study on mixtures of antibiotics and pesticides [J]. Journal of Hazardous Materials, 2024, 461.
[59] VAN GRIENSVEN A, MEIXNER T, GRUNWALD S, et al. A global sensitivity analysis tool for the parameters of multi-variable catchment models [J]. Journal of Hydrology, 2006, 324(1-4): 10-23.
[60] 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.
[61] 郭明月. 基于多源异构数据的办公建筑能耗预测方法[D]. 同济大学机械与能源工程学院, 2022.
[62] HOLLAND J H. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence [M]. 1992.
[63] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323(6088): 533-6.
[64] HINTON G E. Connectionist learning procedures [J]. Artificial Intelligence, 1989, 40(1): 185-234.
[65] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders [Z]. Proceedings of the 25th international conference on Machine learning. Helsinki, Finland; Association for Computing Machinery. 2008: 1096–103.10.1145/1390156.1390294
[66] 袁非牛,章琳,史劲亭等.自编码神经网络理论及应用综述[J].计算机学报,2019,42(01):203-230.
[67] XIANG Q, PANG X L. Improved Denoising Auto-encoders for Image Denoising [Z]. 2018 11TH INTERNATIONAL CONGRESS ON IMAGE AND SIGNAL PROCESSING, BIOMEDICAL ENGINEERING AND INFORMATICS (CISP-BMEI 2018). 2018
[68] MA H Q, MA S P, XU Y L, et al. Deep Marginalized Sparse Denoising Auto-Encoder for Image Denoising [Z]. 2017 2ND INTERNATIONAL CONFERENCE ON COMMUNICATION, IMAGE AND SIGNAL PROCESSING (CCISP 2017). 2018.10.1088/1742-6596/960/1/012033
[69] SHAO H, JIANG H, WANG F, et al. An enhancement deep feature fusion method for rotating machinery fault diagnosis [J]. Knowledge-Based Systems, 2017, 119: 200-20.
[70] NANDHINI ABIRAMI R, DURAI RAJ VINCENT P M, SRINIVASAN K, et al. Deep CNN and Deep GAN in Computational Visual Perception-Driven Image Analysis [J]. Complexity, 2021, 2021: 1-30.
[71] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Commun ACM, 2020, 63(11): 139–44.
[72] MIRZA M, OSINDERO S. Conditional Generative Adversarial Nets [J]. ArXiv, 2014, abs/1411.1784.
[73] 魏贞原, 机器学习 Python 实践[M], 北京，电子工业出版社, 2018.

[74] CHEN T, GUESTRIN C. XGBoost [Z]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785-94.10.1145/2939672.2939785
[75] WANG Z, HONG T, PIETTE M A. Building thermal load prediction through shallow machine learning and deep learning [J]. Applied Energy, 2020, 263.
[76] CHAKRABORTY D, ELZARKA H. Advanced machine learning techniques for building performance simulation: a comparative analysis [J]. Journal of Building Performance Simulation, 2019, 12(2): 193-207.
[77] SNOEK J., LAROCHELLE H., ADAMS R. P. Practical Bayesian Optimization of Machine Learning Algorithms [J]. Advances in Neural Information Processing Systems, 2012, 25, 2951–2959.
[78] YAN R, ZHAO T, REZGUI Y, et al. Transferability and robustness of a data-driven model built on a large number of buildings [J]. Journal of Building Engineering, 2023, 80.
[79] MENSINK T, UIJLINGS J, KUZNETSOVA A, et al. Factors of Influence for Transfer Learning Across Diverse Appearance Domains and Task Types [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 9298-314.
[80] CODY T, ADAMS S, BELING P A. Empirically Measuring Transfer Distance for System Design and Operation [J]. IEEE Systems Journal, 2022, 16(3): 4962-73.
[81] ŞAHIN G, DIRI B. The Effect of Transfer Learning on Turkish Text Classification; proceedings of the 2021 29th Signal Processing and Communications Applications Conference (SIU), F 9-11 June 2021, 2021 [C].
[82] SOMU N, M R G R, RAMAMRITHAM K. A hybrid model for building energy consumption forecasting using long short term memory networks [J]. Applied Energy, 2020, 261.
[83] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8): 1735-80.
[84] SRIVASTAVA S, LESSMANN S. A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data [J]. Sol Energy, 2018, 162: 232–47.
[85] ZHOU D, MA S, HAO J, et al. An electricity load forecasting model for Integrated Energy System based on BiGAN and transfer learning [J]. Energy Reports, 2020, 6: 3446-61.
[86] KARIJADI I, CHOU S-Y. A hybrid RF-LSTM based on CEEMDAN for improving the accuracy of building energy consumption prediction [J]. Energy and Buildings, 2022, 259.
[87] YAN H, QIN Y, XIANG S, et al. Long-term gear life prediction based on ordered neurons LSTM neural networks [J]. Measurement, 2020, 165: 108205.
[88] HUANG G, CHOW T-T. Uncertainty shift in robust predictive control design for application in CAV air-conditioning systems [J]. Building Services Engineering Research and Technology. 2011, 32(4): 329-343.
[89] WANG Q. Accuracy, validity and relevance of probabilistic building energy models [D], 2016.
[90] YU J, CHANG W-S, DONG Y. Building Energy Prediction Models and Related Uncertainties: A Review [J]. Buildings, 2022, 12(8).

致谢

行文至此，三年读研生涯至尾声，一切尽在不言中。感谢三年来许鹏老师的教导，对我影响颇深，感谢课题组师兄师姐师弟师妹的帮助，感谢家人永远支持我的一切决定。最后，感谢黄乾晋同学的陪伴与支撑，让枯燥的学术生活变得充满活力与光彩。感恩相遇。

毕业快乐！我们都有光明的未来！

个人简历、在读期间发表的学术成果

个人简介：

夏壮，女，1998 年 8 月生。

2021 年 7 月毕业于同济大学 建筑环境与能源应用工程专业 获学士学位。

2021 年 9 月入同济大学攻读硕士研究生

已发表论文：

[1] XIA Z, GUAN H, QI Z, et al. Multi-Zone Infection Risk Assessment Model of Airborne Virus Transmission on a Cruise Ship Using CONTAM [J/OL] 2023, 13(9):10.3390/buildings13092350.

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**《基于数据驱动的公共建筑综合体能耗预测方法》**，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： 夏北

日期： 2024年 3 月 20 日

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： 夏北　　**指导教师签名：** 许[illegible]

日期： 2024年 3 月 20 日　　**日期：** 2024年 3 月 20 日

七、学位论文答辩委员会决议

姓名	夏壮	学号	2130277	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	2024.3.19	答辩地点	开物馆A413.
论文题目	基于数据驱动的公共建筑综合体能耗预测方法				

答辩委员会共 4 人，经表决，4 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。

☐ 申请人可在一年内修改论文，申请重新答辩一次。

☐ 建议不授予申请人硕士学位。

是否推荐为同济大学优秀硕士学位论文：☐ 是 ☑ 否

夏壮同学的硕士学位论文研究了建筑综合体的能耗预测方法，研究成果对综合体建筑的节能设计与运行具有实际意义。

论文采用敏感性分析方法筛选了影响综合体能耗的关键变量；研究了基于遗传算法和降噪自编码器的缺失关键变量的推断算法；基于条件生成对抗网络对建筑能耗数据进行数据增强；最终建立了融合了增强数据集和真实数据集的迁移学习建筑能耗预测模型。

论文研究目标明确，结构完整，条理清晰，工作量饱满，模型及论证合理，体现了作者具有扎实的基础理论和专业知识，具备了从事科学研究工作和解决实际工程问题的能力。

夏壮同学在答辩过程中，表述完整，思路清晰，能正确回答答辩委员提出的问题。论文达到了工学硕士学位论文要求。经无记名投票表决，四位答辩委员中，四位同意并建议授予夏壮同学工学硕士学位。

答辩委员会主席签名：苏醒　　2024 年 3 月 19 日

	职务	姓名	职称	单位	签名
答辩委员会成员签名	主席	苏醒	副教授	同济大学	苏醒
	委员	许鹏	教授	同济大学	许鹏
	委员	李沛伟	副教授	同济大学	李沛伟
	委员	叶霖	副教授	同济大学	叶霖
	委员				
秘书		曾令杰	助理教授	同济大学	曾令杰

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
硕士学位论文答辩

题　目：基于数据驱动的公共建筑综合体能耗预测方法

硕士生：夏壮

导　师：许鹏　教 授

2022年03月16日

同济大学硕士学位论文答辩

主　席：	苏　醒	教　　授	同济大学
委　员：	李铮伟	副 教 授	同济大学
	叶　蔚	副 教 授	同济大学
	许　鹏	教　　授	同济大学
秘　书：	曾令杰	助理教授	同济大学

答辩人：　夏壮

2022年03月16日

基于数据驱动的公共建筑综合体能耗预测方法

答辩人：夏壮　　指导老师：许鹏教授

二〇二四年三月

目录

CONTENTS

1. 引言

研究背景

建筑运行能耗占比居高不下

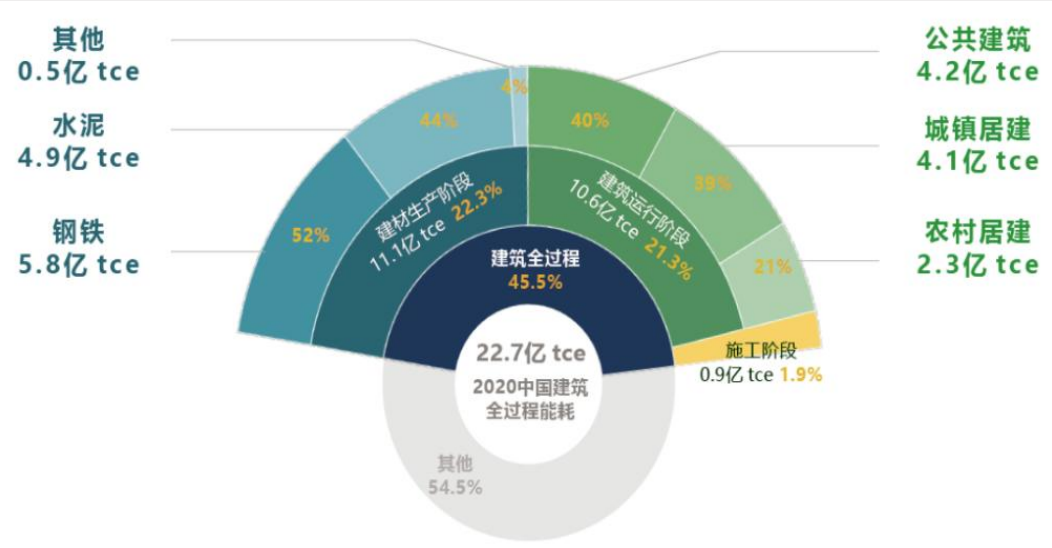


图1.1 2020 年全国建筑全过程能耗占比情况

综合建筑年用电量占比较大

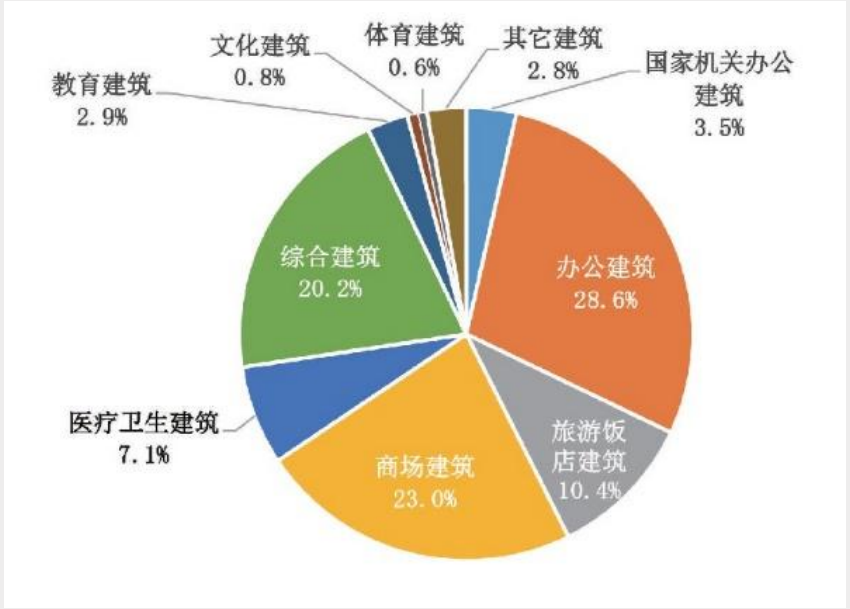


图1.2 2020年上海市公共建筑能耗监测平台能耗数据分析

>> 建筑综合体节能意义重大！

研究背景

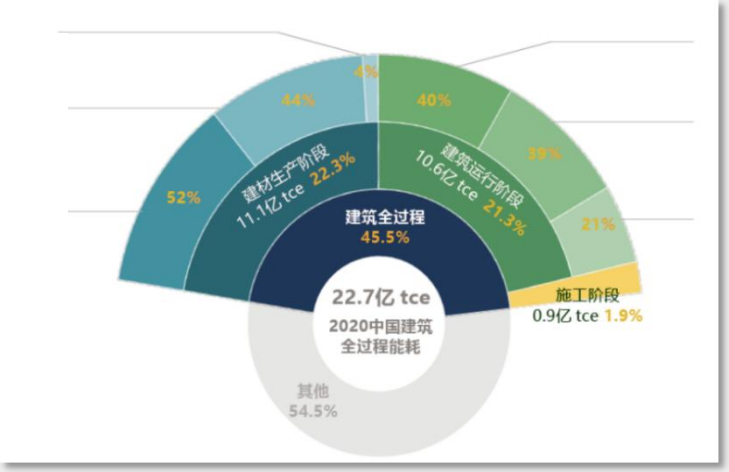


图1.1 2020 年全国建筑全过程能耗占比情况

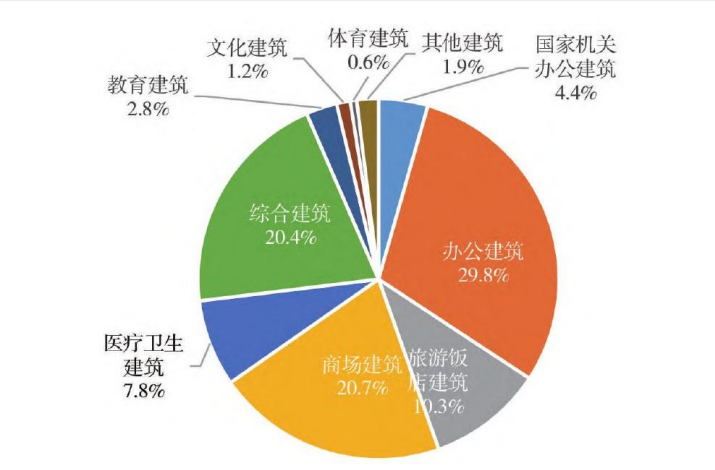
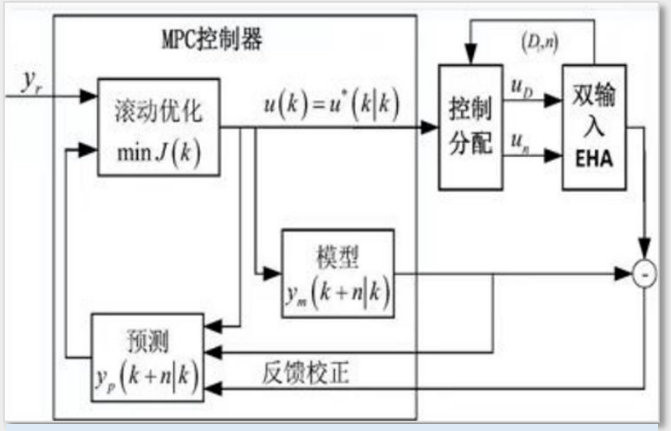


图1.2 2022年与能耗监测平台联网的建筑年用电量占比情况

系统设备选型



系统控制优化

建筑节能改造

需求响应

>> 建筑综合体能耗预测意义重大

文献综述

数据驱动的建筑能耗主流预测模型主要有三大类：

◘ 传统时间序列模型：

优点：解释性强

考虑季节性和周期性变化

缺点：对数据假设严格

处理非线性和高维问题表现差

◘ 传统机器学习模型

优点：适用性广泛

灵活性高

缺点：手动特征工程

难以捕捉长期依赖关系

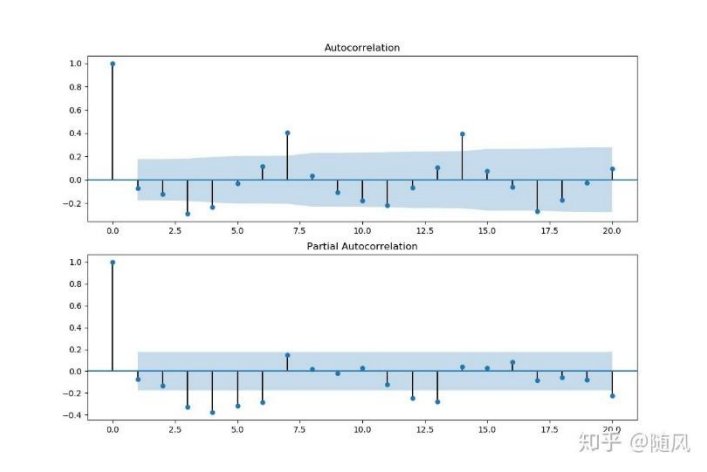
◘ 深度学习模型

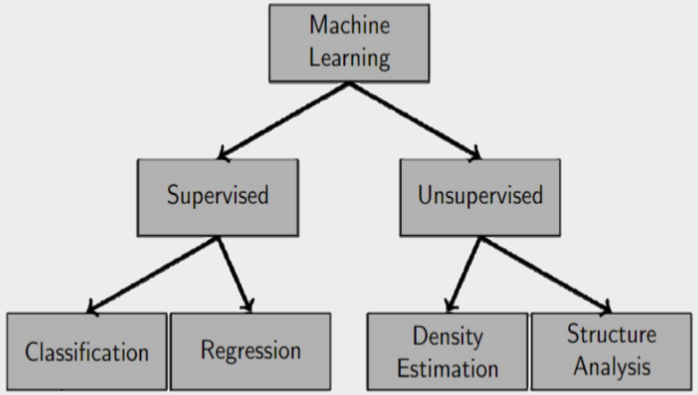
优点：能够处理大规模数据

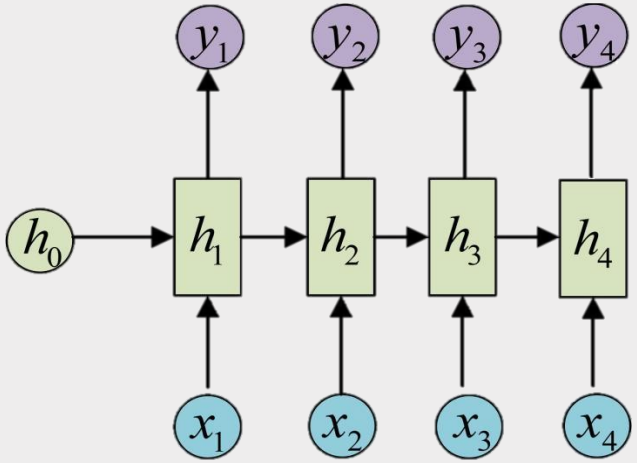
自动特征学习

缺点：数据需求量大

调参复杂





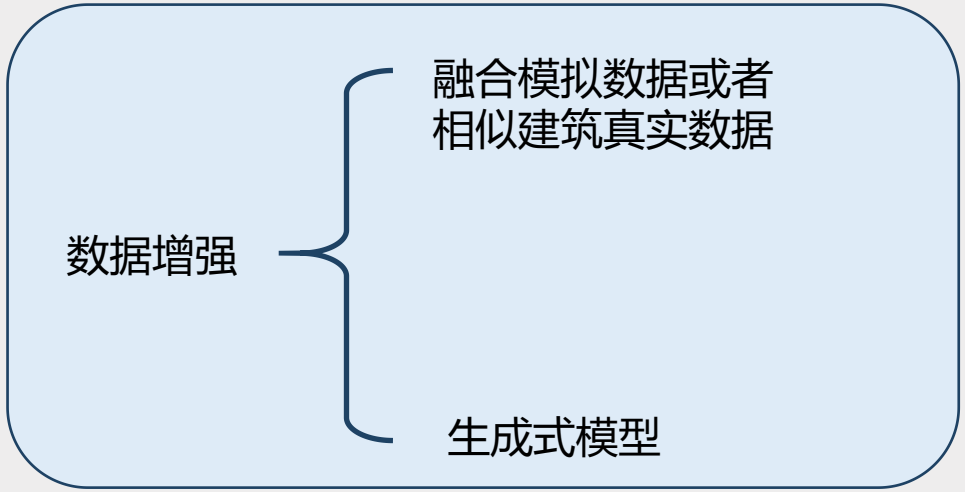


文献综述

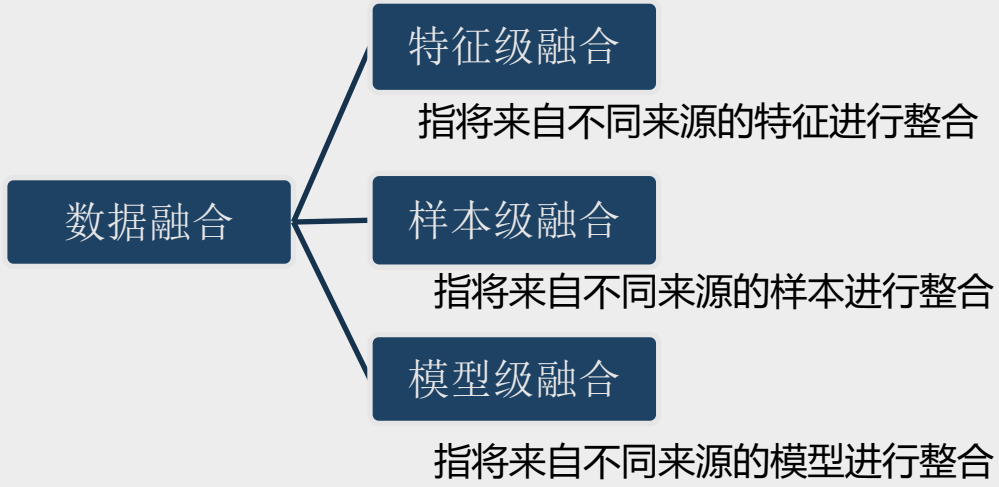
实现数据驱动模型的2个基本假设是：

- 训练数据应满足全局学习空间中独立且同分布的采样；
- 必须有足够的训练样本来学习一个好的模型

数据增强是利用有限的训练数据实现数据驱动建模的有效方法之一，主要用于解决机器学习和深度学习算法在有限数据集上的欠训练问题。



建筑能耗预测领域数据融合主要有两种方式，一是直接数据融合，而是通过迁移学习融合。直接数据融合是一种简单的方法，它是将增强后的数据直接添加到目标任务的不足训练集中。然而，当增强数据与实际数据的分布差异较大时，直接数据融合往往性能不佳。基于迁移学习的数据融合比直接数据融合更有效。



研究意义

现有研究的不足

01 目前建筑能耗预测研究的对象大多是功能单一的建筑，很少有关于建筑综合体能耗预测的研究。

02 目前绝大部分数据驱动的能耗预测只针对某一单体建筑，依赖于来自同一建筑的足够的历史数据来训练模型，所建立的预测模型难以迁移到其他建筑上。

03 真实数据不足，难以满足深度学习所需要的数据量导致所训练出的模型预测效果差

本研究的意义

综合体能耗预测

弥补目前市面上关于多功能建筑能耗预测研究的空缺

跨建筑能耗预测

跨建筑的建筑全年能耗预测模型的建立，极大地提高了模型的适用性，使得预测模型不再仅适用于某一栋建筑而是适用于某一种建筑类型。

数据增强+迁移学习

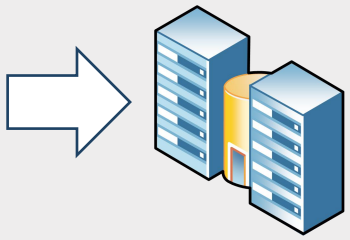
数据增强解决了模型训练真实数据不足的问题
迁移学习解决了真实数据和增强数据融合的问题

技术路线

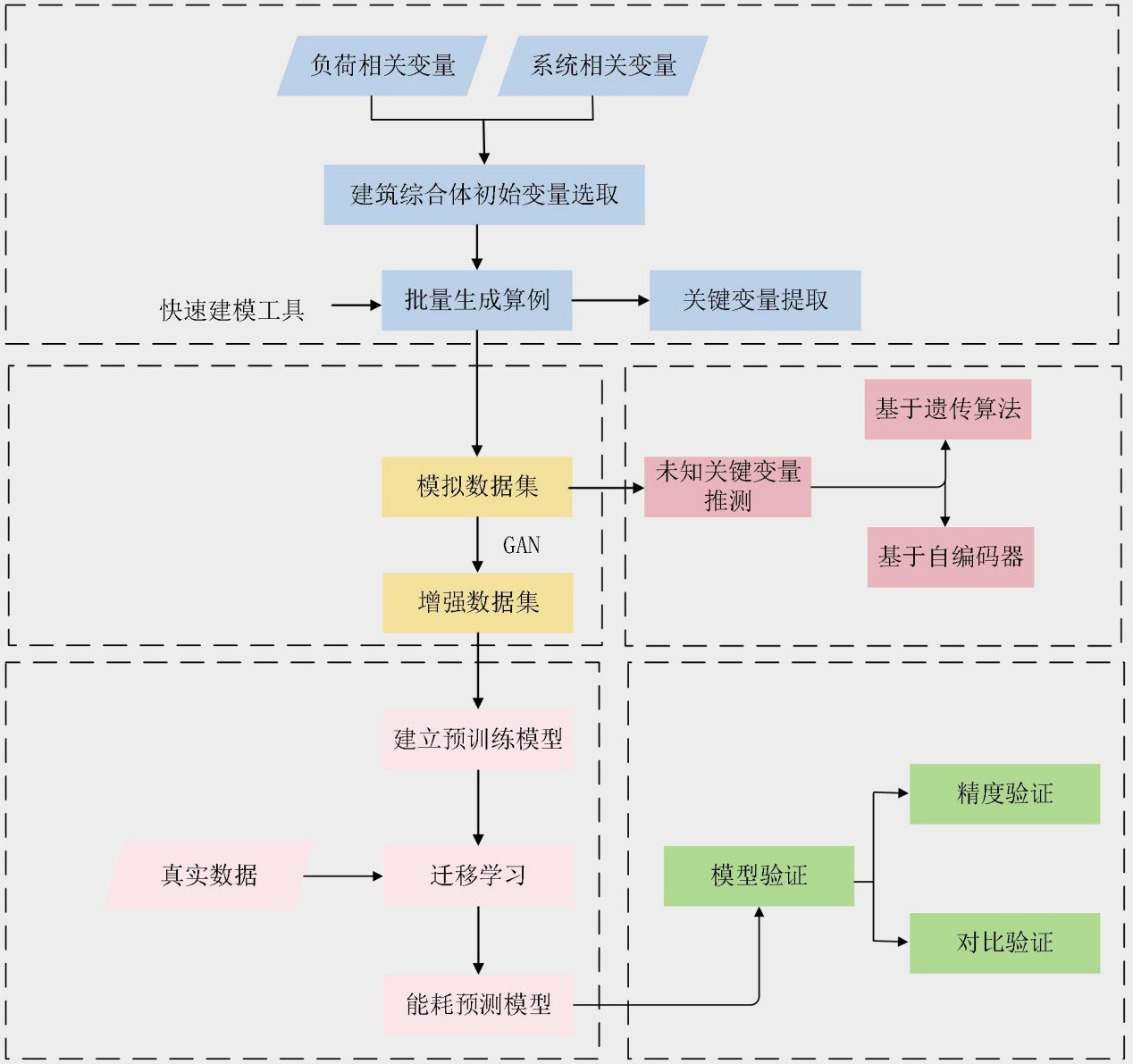
本研究需要解决的问题

1) 综合体存在多套系统，多系统导致预测模型输入更加复杂，如何选择合适的关键变量提高模型精度、避免维度灾难？
2) 在部分关键建筑信息缺失的情况下如何进行能耗预测？
3) 跨建筑能耗预测模型的训练对数据量要求较高，如何解决真实数据不足的问题？

① 有/无历史能耗
② 建筑静态参数缺失



建筑能耗分布



2. 关键变量提取

2. 关键变量提取

- **负荷相关变量：**

  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变量

- **系统相关变量：**

  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：
所有与能耗可能相关的变量，是特征提取的基础变量

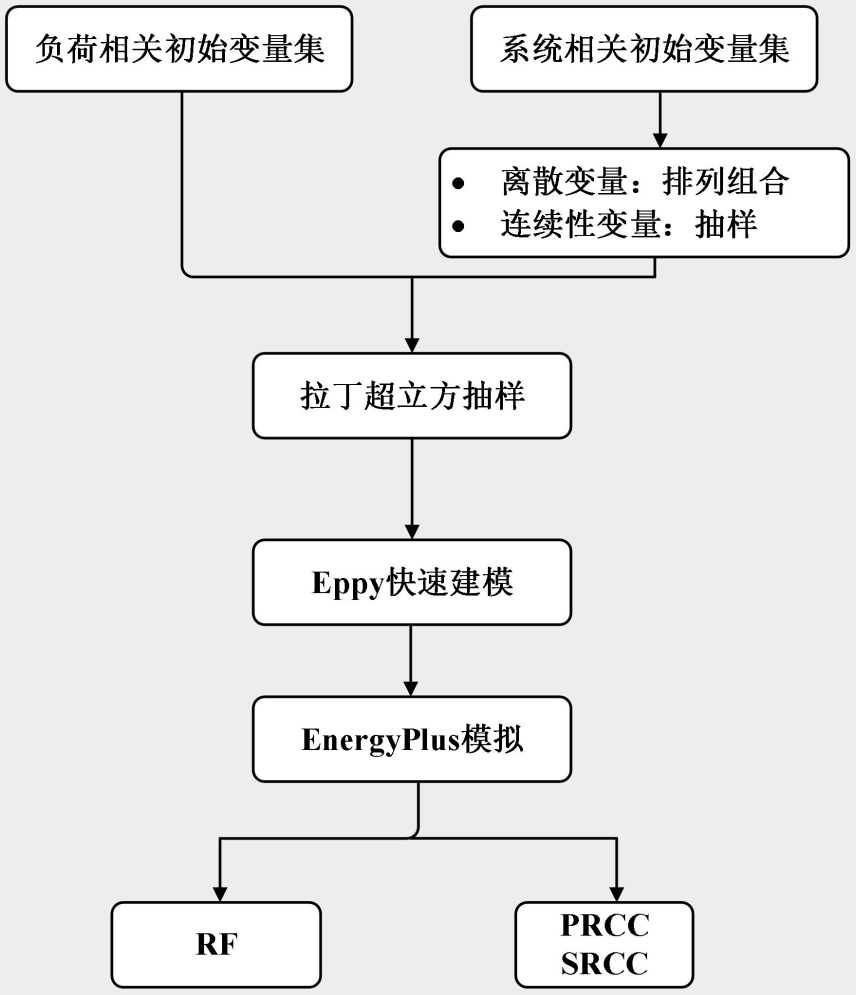


图2.1 关键变量提取技术路线

2. 关键变量提取

负荷相关关键变量

- 确定初始变量集及其范围
- 拉丁超立方抽样得到输入变量
- 批量生成IDF文件并模拟
- 关键变量提取

类别	变量名称	缩写	取值范围	单位
建筑几何外形	东向窗墙比	EWWR	0.1-0.9	-
	南向窗墙比	SWWR	0.1-0.9	-
	西向窗墙比	WWWR	0.1-0.9	-
	北向窗墙比	NWWR	0.1-0.9	-
	层数	NL	4-60	层
	建筑面积	AREA	20000-200000	m^2
	体形系数	CR	0.1-0.5	-
建筑围护结构热工性能	外墙传热系数	WALLU	0.09-0.5	$W/(m^2·K)$
	外墙热容	WSP	800-2000	J/(kg·K)
	屋顶传热系数	RU	0.09-0.4	$W/(m^2·K)$
	窗玻璃传热系数	WINU	0.2-0.9	$W/(m^2·K)$
	窗玻璃太阳辐射得热系数	SHGC	0.1-0.9	-
	外墙太阳辐射吸收系数	WSA	0.1-0.9	-
	屋顶太阳辐射吸收系数	RSA	0.1-0.9	-
建筑使用情况	空调制冷设定温度	SPC	21-28	℃
	空调制热设定温度	SPH	18-26	℃
	设备照明功率密度	LPD	3-15	W/m^2
	人员密度	OPD	0.1-1	人/m^2
	冷风渗透率	INFIL	0.5-5	ACH
	内遮阳开启程度	ST	0.1-0.9	-
施工质量	楼板线性透过率	FLT	0.007-1.842	W/(m K)
	玻璃线性透过率	GLT	0.03-1.058	W/(m K)
	墙角线性透过率	CLT	0.036-0.684	W/(m K)

表2.1 负荷相关初始变量集

2. 关键变量提取

负荷相关关键变量

- ✓ 确定初始变量集及其范围
- 拉丁超立方抽样得到输入变量
- 批量生成IDF文件并模拟
- 关键变量提取

拉丁超立方采样(Latin hypercube sampling, LHS)最早由McKay等提出，是一种从多元参数分布中近似随机采样的方法，属于分层采样技术。拉丁超立方抽样确保了样品的结构与整体结构相对相似，并且样品是均匀的。

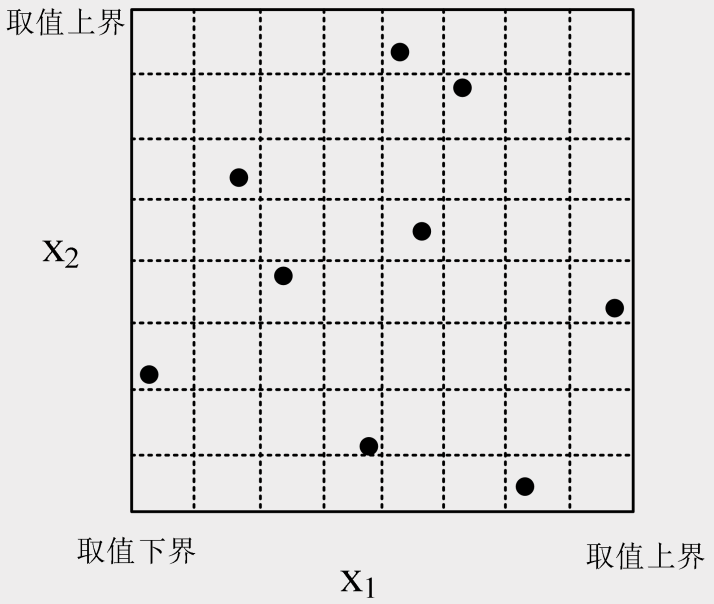


图2.2 LHS原理示意图

2. 关键变量提取

负荷相关关键变量

- ✓ 确定初始变量集及其范围
- ✓ 拉丁超立方抽样得到输入变量
- 批量生成IDF文件并模拟
- 关键变量提取

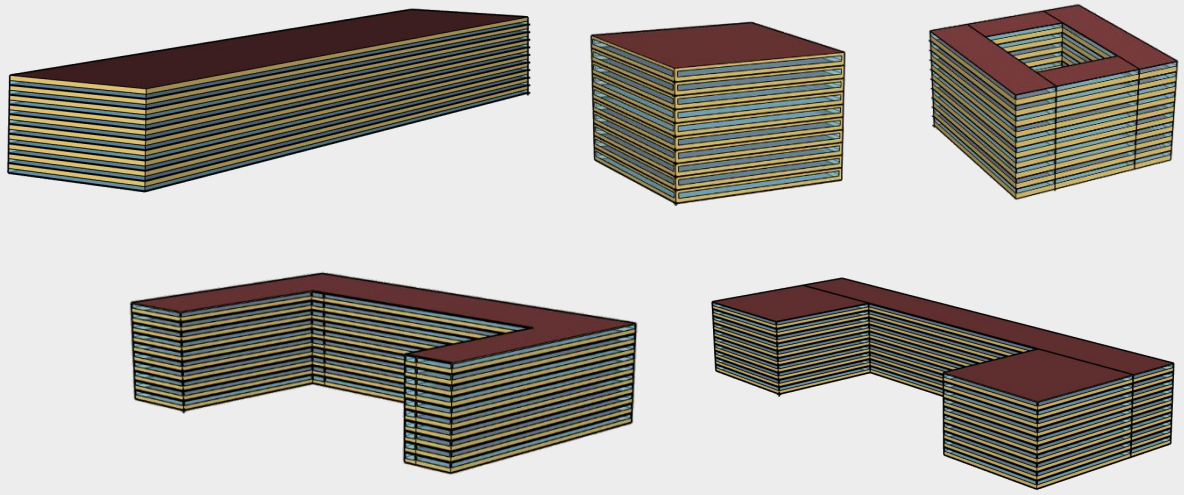
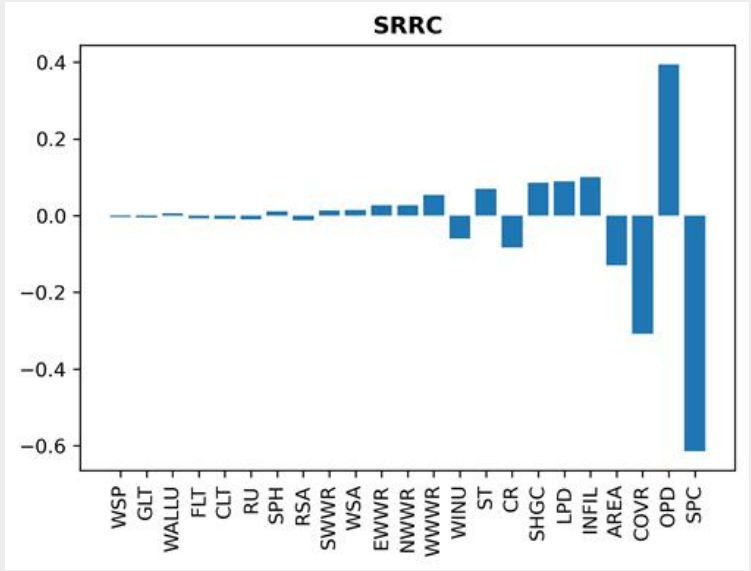


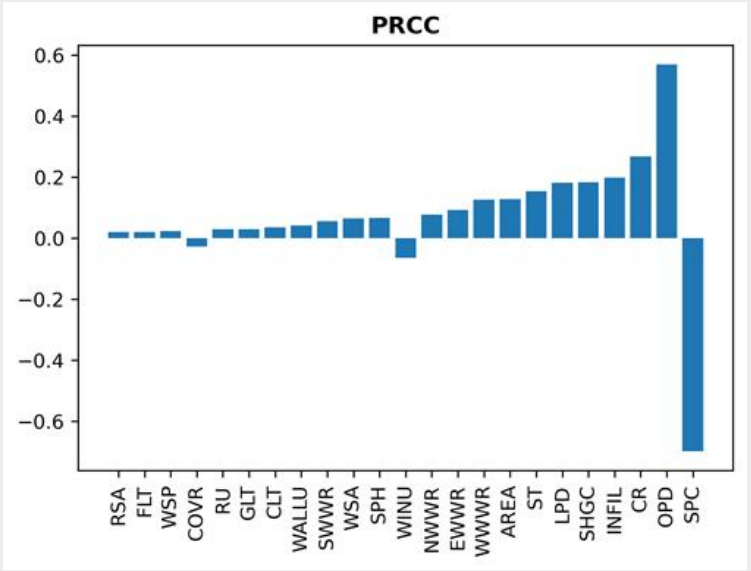
图2.3 可批量生成的建筑外形（匹配建筑体形系数）

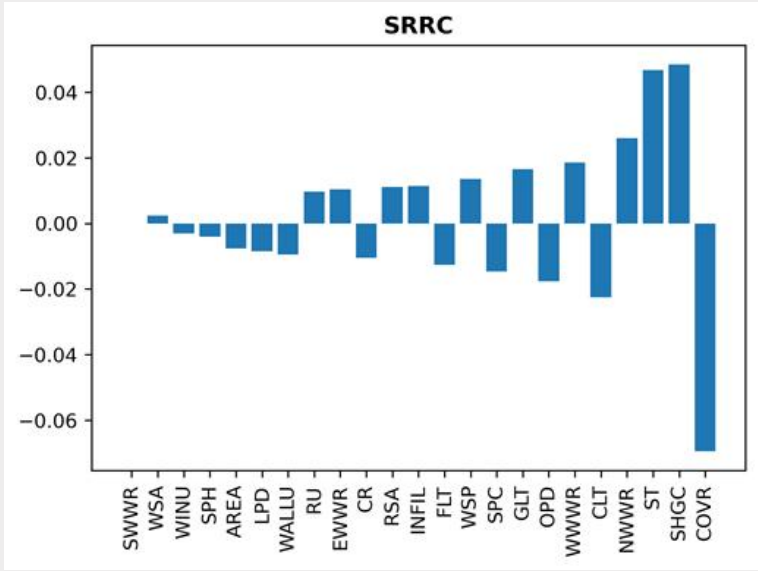
2. 关键变量提取

负荷相关关键变量

- ✓ 确定初始变量集及其范围
- ✓ 拉丁超立方抽样得到输入变量
- ✓ 批量生成IDF文件并模拟
- 关键变量提取：相关系数法







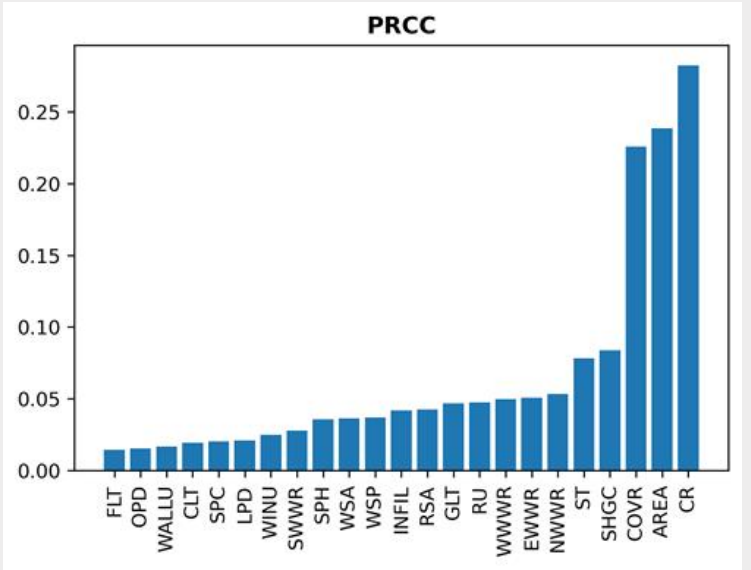


图2.4 相关系数法敏感性分析结果（上为制冷季，下为制热季）

2. 关键变量提取

负荷相关关键变量

- ✓ 确定初始变量集及其范围
- ✓ 拉丁超立方抽样得到输入变量
- ✓ 批量生成IDF文件并模拟
- 关键变量提取：随机森林

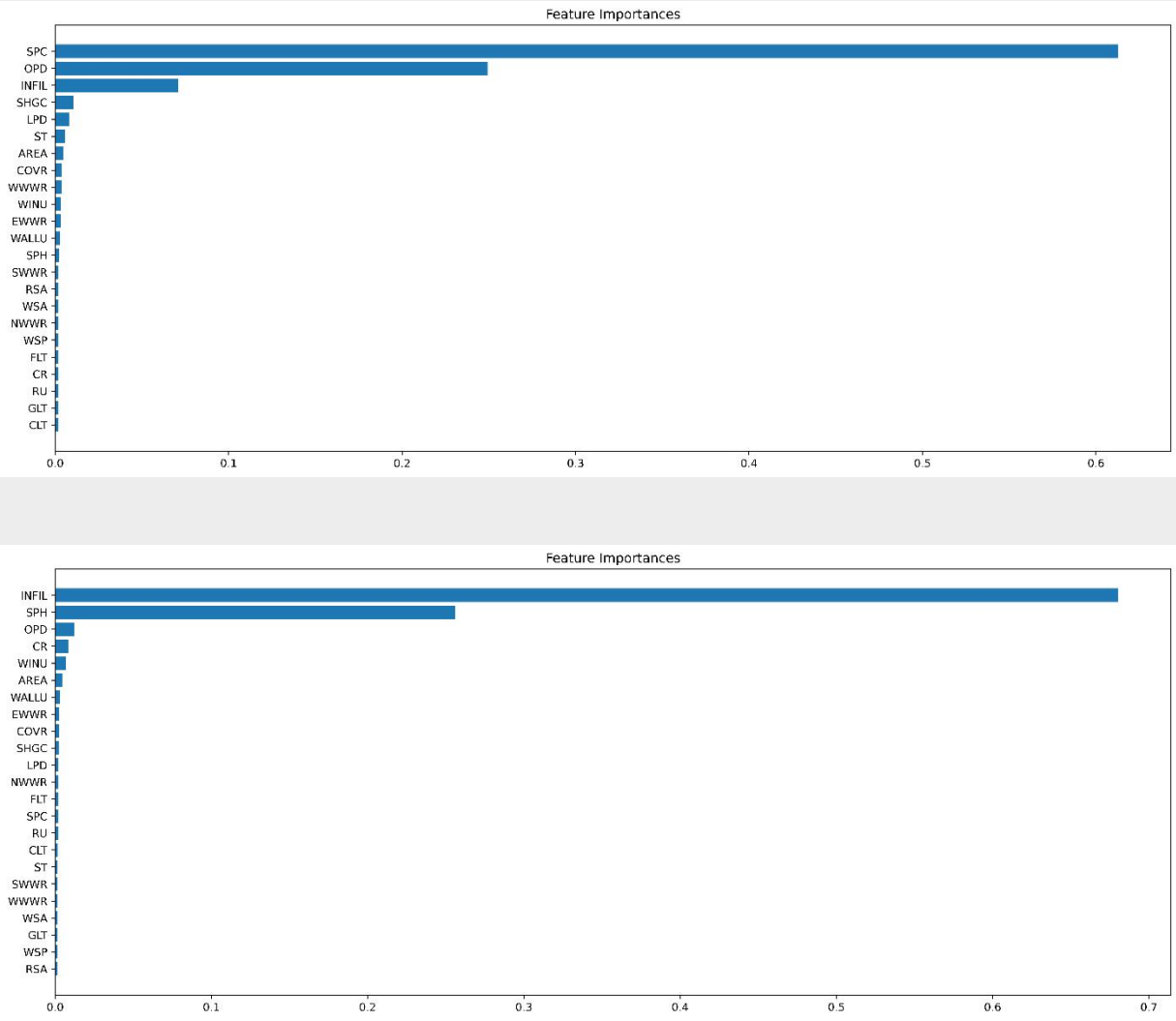


图2.4 随机森林敏感性分析结果（上为制冷季，下为制热季）

2. 关键变量提取

系统相关关键变量

- ✓ 确定初始变量集及其范围
- 拉丁超立方抽样得到输入变量
- 批量生成IDF文件并模拟
- 关键变量提取

表2.2 系统相关初始变量集

类别	变量名称	缩写	取值范围	单位
系统类型参数	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、二次泵变流量系统	-
系统运行参数	送风温差	SATD	4-10	℃
	冷冻水供水温度	CHWT	5-10	℃
	热水供水温度	HWT	50-65	℃
	风机效率	FE	0.3-0.8	
	水泵效率	PF	0.3-0.8	
	主机COP	COP	3-7	
	冷冻水供回水温差	CTD	2-7	℃
	热水供回水温差	HTD	8-15	℃
	冷却塔填料堵塞率	CFBR	0.5-1	-
	风系统过滤器堵塞率	FFBR	1-2	-

2. 关键变量提取

系统相关关键变量

- ✓ 确定初始变量集及其范围
- ✓ 拉丁超立方抽样得到输入变量
- ✓ 批量生成IDF文件并模拟
- 关键变量提取

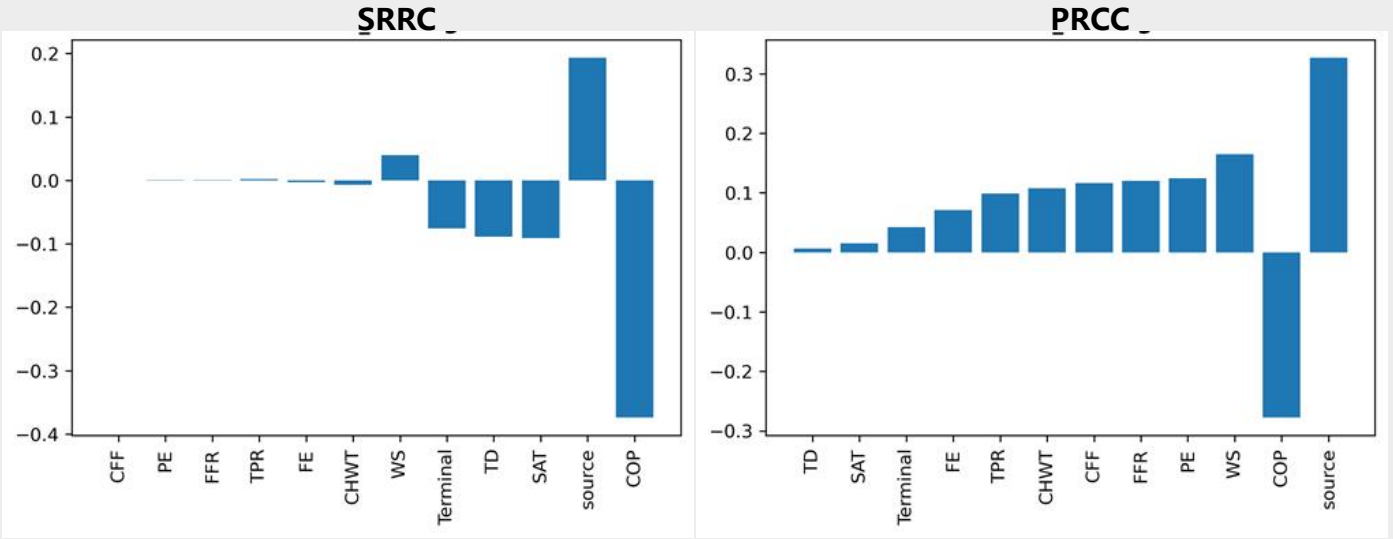


图2.5 相关系数法敏感性分析结果

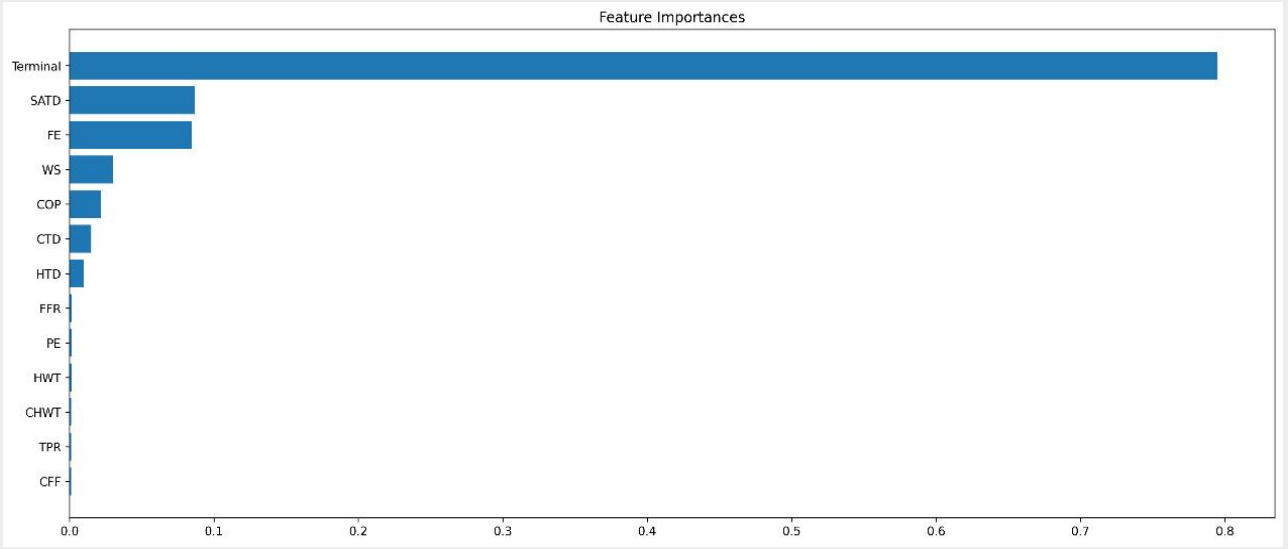


图2.6 随机森林敏感性分析结果

2. 关键变量提取

表2.3 关键变量提取结果

参数类别	参数名称	缩写	单位
建筑负荷相关变量	制冷设定温度	SPC	℃
	制热设定温度	SPH	℃
	人员密度	OPD	人/m^2
	设备照明密度	LPD	W/ m^2
	冷风渗透率	INFIL	ACH
	体形系数	CR	/
	太阳得热系数	SHGC	/
	窗墙比	WWR	/
	内遮阳开启程度	ST	/
系统相关变量	风系统类型	Terminal	/
	水系统类型	WS	/
	主机COP	COP	/
	冷冻水供回水温差	CTD	℃
	热水供回水温差	HTD	℃
	冷冻水温度	CHWT	℃
	热水温度	HWT	℃
	风机效率	FE	/
	水泵效率	PE	/

表2.4 天气参数及时间标签

参数类别	参数名称	缩写	单位
天气参数	干球温度	DryT	℃
	相对湿度	RH	/
	风速	Wind	m/s
时间标签	每年的月	month	/
	每月的日	day	/
	星期几	week	/
	是否是工作日	workday	/

综合体一般有多个空调系统（对应不同的功能分区），为了降低特征维度，对系统运行参数按照不同功能分区的面积比例进行加权获得最终建筑维度的参数值。

$$value = x_1 \times value_1 + x_2 \times value_2 \qquad (2\text{-}1)$$

3. 缺失关键变量推断

3. 缺失关键变量推断

基于遗传算法的缺失关键变量推断

- 有历史能耗的建筑
- 以缺失关键变量作为决策变量，通过遗传算法寻找对应模拟能耗与实测能耗偏差最小的缺失关键变量的取值。

基于自编码器的缺失关键变量推断

- 新建建筑等没有历史能耗的建筑
- 通过自编码器学习关键变量之间的关联关系，根据其关联关系进行缺失关键变量的推断。

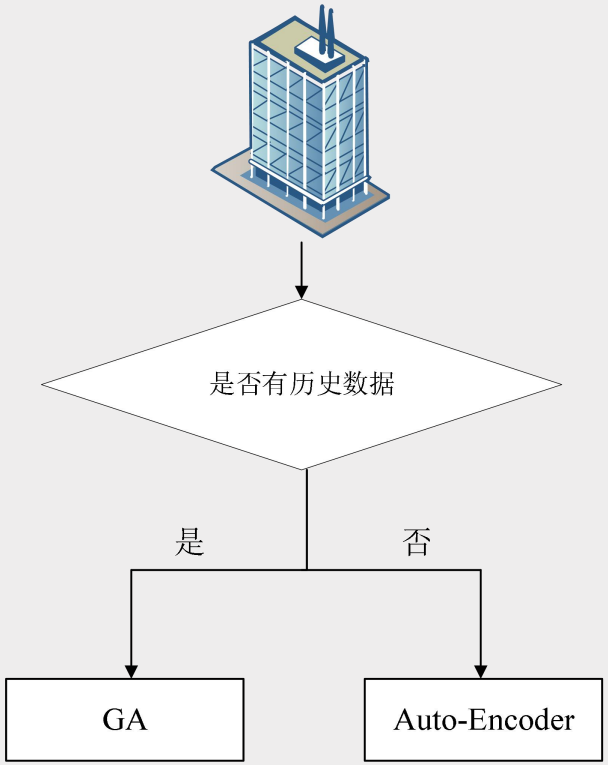


图3.1 缺失关键变量推断方法

3. 缺失关键变量推断——基于遗传算法

基于遗传算法的缺失关键变量推断

- 给定缺失参数的可行域，在可行域内生成初始种群即缺失变量推断值集合；
- 每一个个体连同其他已知的关键变量输入能耗预测模型，得到能耗预测值，计算预测值与建筑历史能耗之间的绝对值误差（即适应度）。
- 改变缺失变量推断值（交叉、变异），重复上一步。当绝对值误差收敛或者到达最大迭代次数，跳出寻优输出最终关键变量推断值。

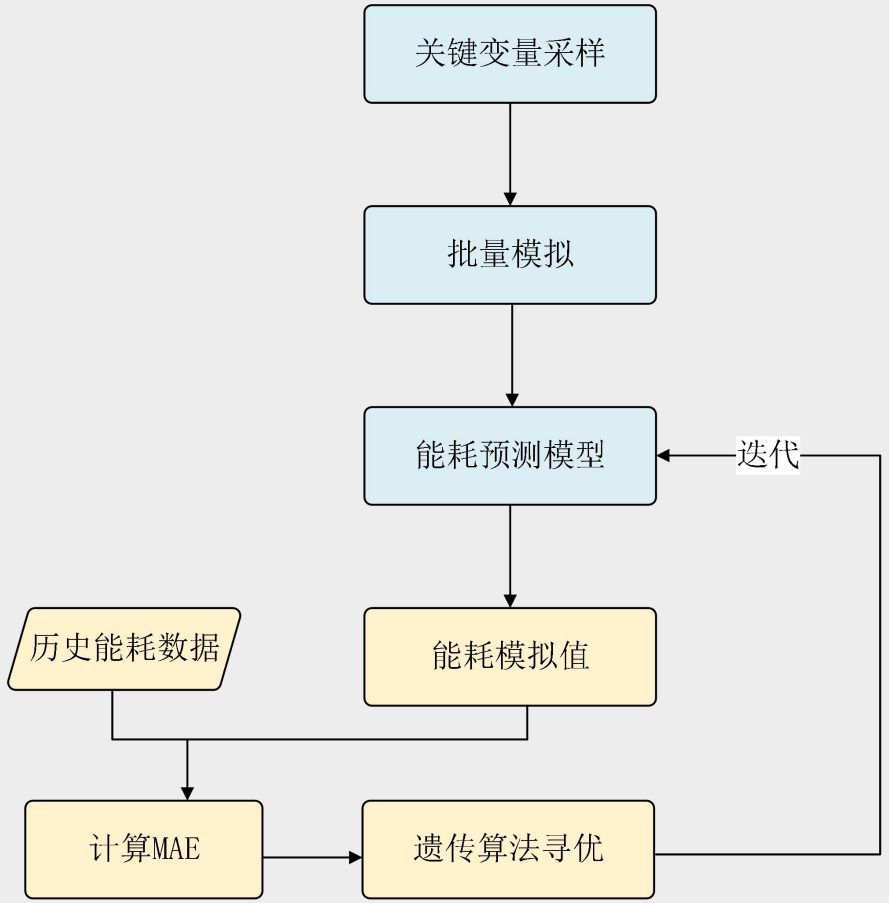


图3.3 基于遗传算法的缺失关键变量推断技术路线

3. 缺失关键变量推断——基于遗传算法

单变量缺失推断验证（80栋建筑）

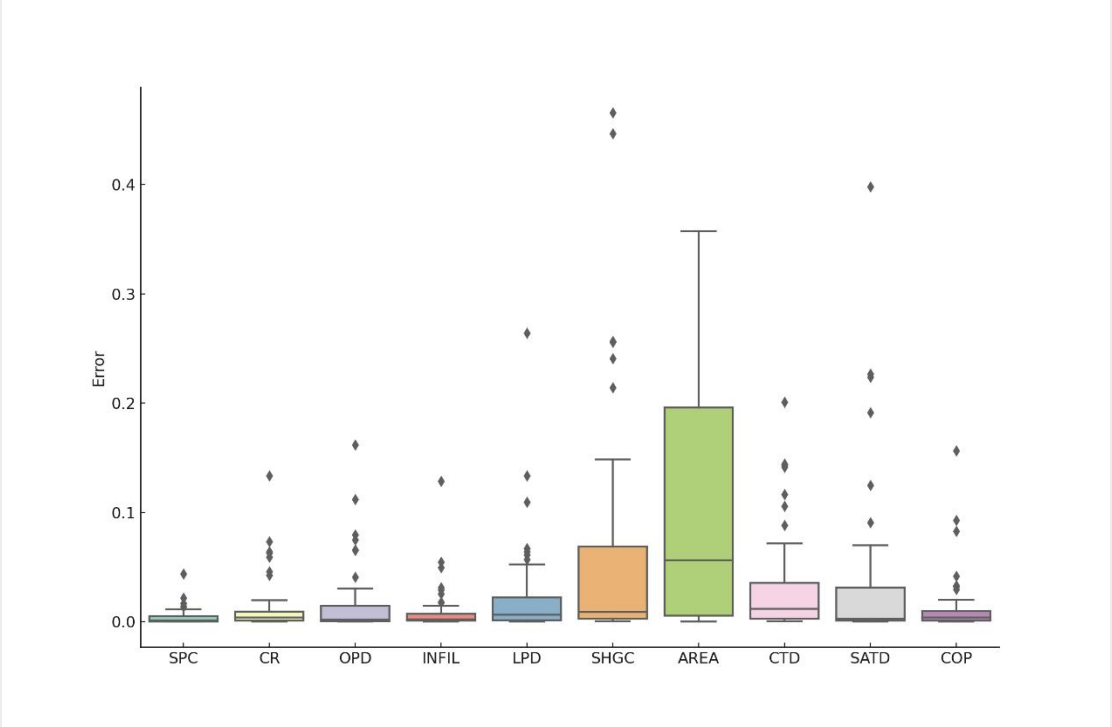


图3.3 基于遗传算法的单变量推断误差

- 建筑面积以外所有变量的误差均值在10%以内。
- 与建筑能耗的相关性越强的参数，推断的准确性也越高。

表3.1 单变量推断误差百分比

缺失变量	符号	平均误差	最小误差	最大误差
制冷设定温度	SPC	0.0038	5.54E-6	0.044
体形系数	CR	0.0219	7.91E-6	0.135
人员密度	OPD	0.017	3.65E-6	0.162
冷风渗透率	INFIL	0.010	6.35E-7	0.128
照明密度	LPD	0.031	4.12E-5	0.264
窗户太阳得热系数	SHGC	0.052	2.38E-4	0.466
建筑面积	AREA	0.121	9.67E-5	0.357
冷冻水供回水温差	CTD	0.045	2.02E-4	0.201
送风温差	SATD	0.035	1.52E-6	0.398
冷机/热泵COP	COP	0.014	7.63E-5	0.156

3. 缺失关键变量推断——基于遗传算法

多变量缺失推断验证（80栋建筑）

当缺失变量的个数继续增加，变量推断误差也在不断增大。

当多个变量缺失时，每个缺失的变量对建筑能耗有着不同的影响，那么就会出现不同变量值的组合使建筑呈现相似的能耗表现，推断结果也就很有可能偏离真实值。

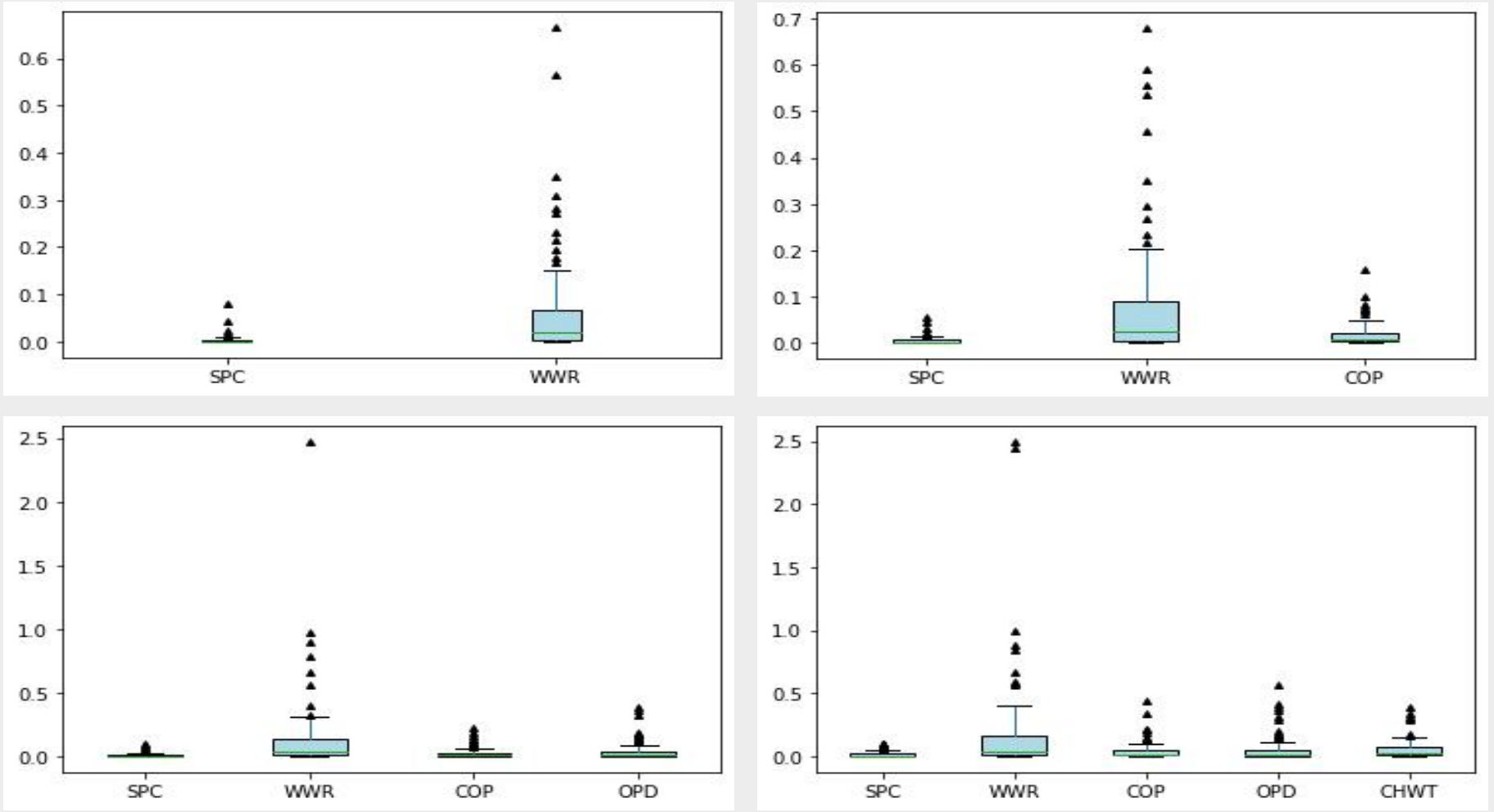


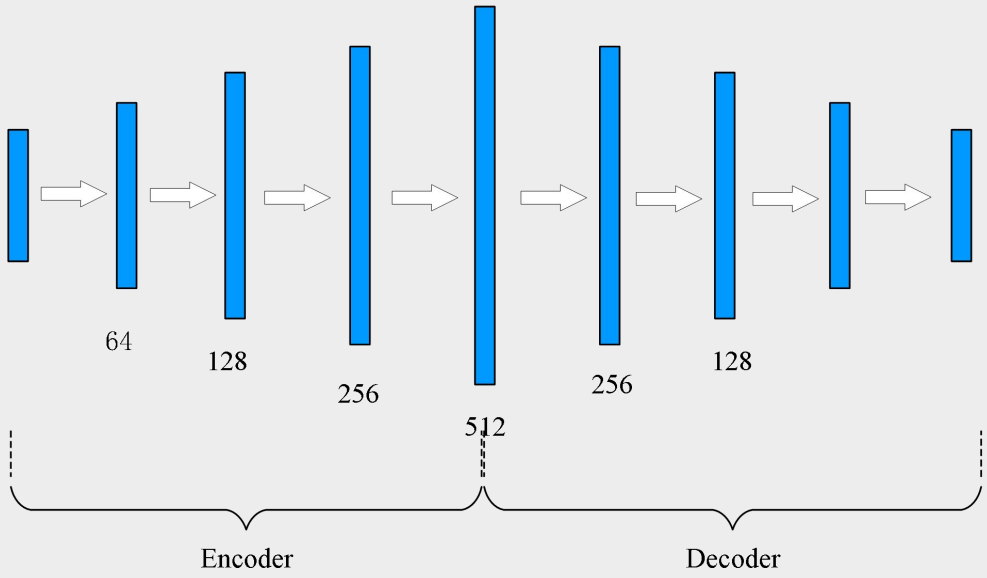
图3.4 基于遗传算法的多变量推断误差

3. 缺失关键变量推断——基于自编码器

基于自编码器的缺失关键变量推断

利用降噪自编码器挖掘关键变量组合内部的关系，重构缺失的关键变量。

编码器通过逐渐增加层级，学习输入数据的抽象和高级特征，较深的层可以捕获更抽象和高级的特征。解码器逐渐减少层级逐渐减少的解码器层可以有助于去除输入中的噪音。



单变量缺失推断验证

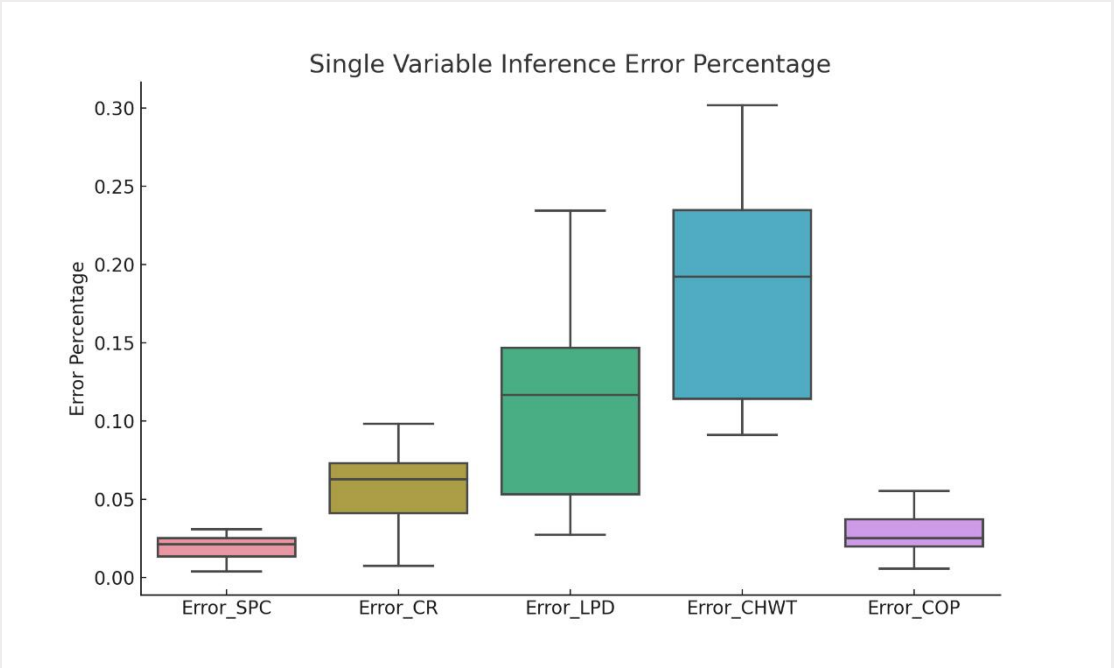


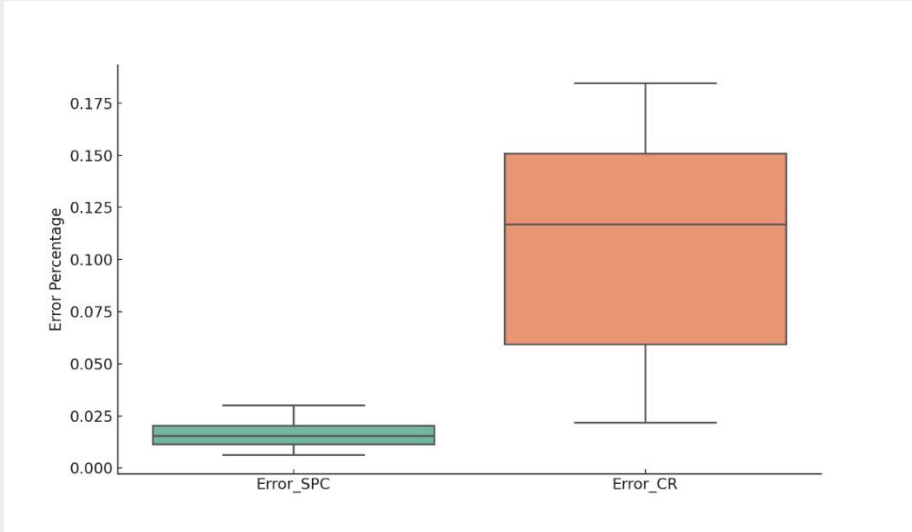
图3.5 自编码器单变量缺失推断误差百分比

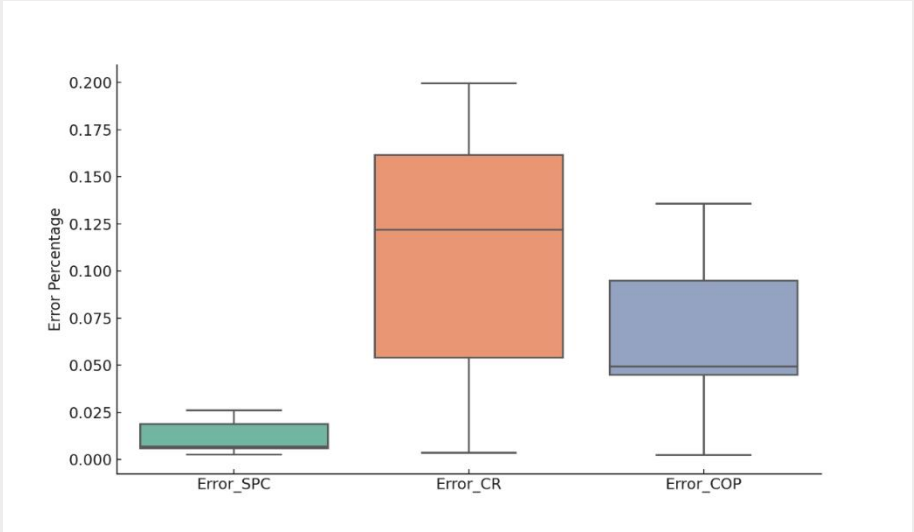
单变量缺失时，推断误差在30%以下，冷冻水供水温度的推断平均误差最大，制冷设定温度和主机性能系数的平均推断误差最小。

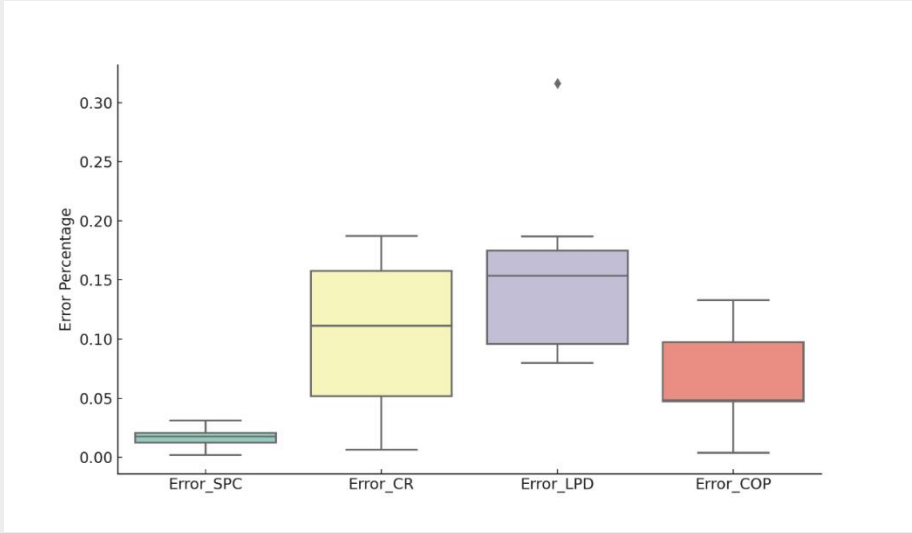
3. 缺失关键变量推断——基于自编码器

多单变量缺失推断验证

多变量缺失时推断误差均在30%以下，随着缺失变量的增加，推断误差增加，且箱线图箱子高度增加，这表明推断结果的不确定性有所增加。







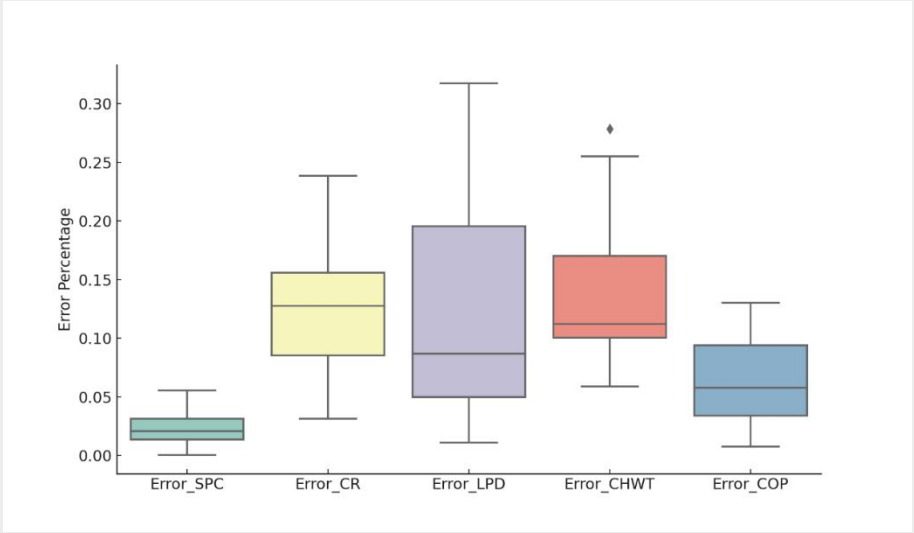


图3.6 自编码器多变量缺失推断误差百分比

4. 数据增强

4. 数据增强

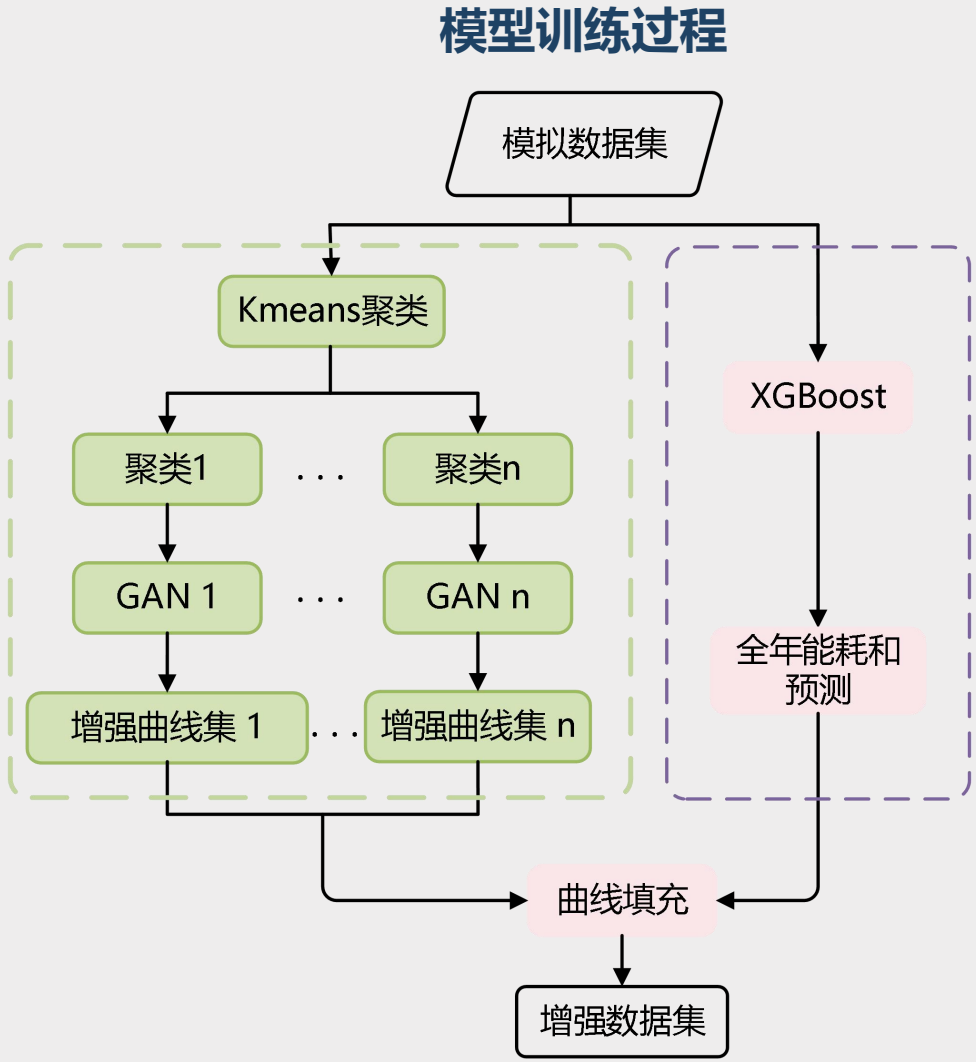


图4.1基于CGAN的数据增强技术路线图

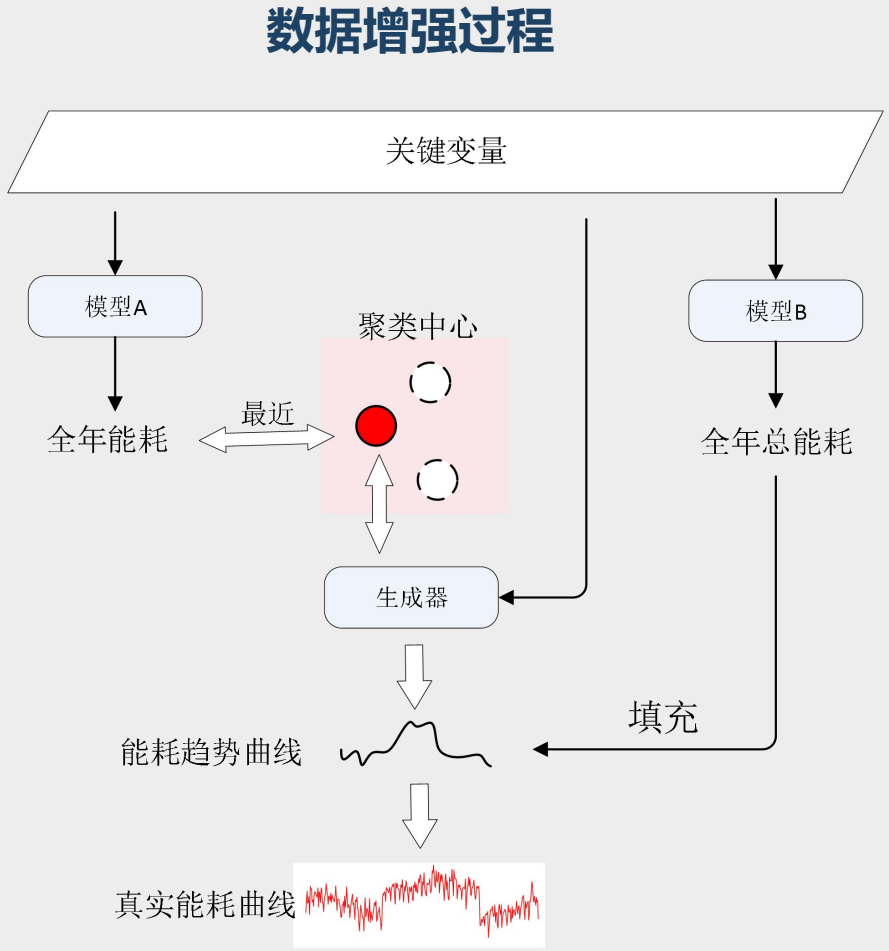


图4.2 具体数据增强流程，模型A为第三章基于模拟数据集所训练的全年能耗预测模型，模型B为全年总能耗预测模型。

4. 数据增强

模型训练过程

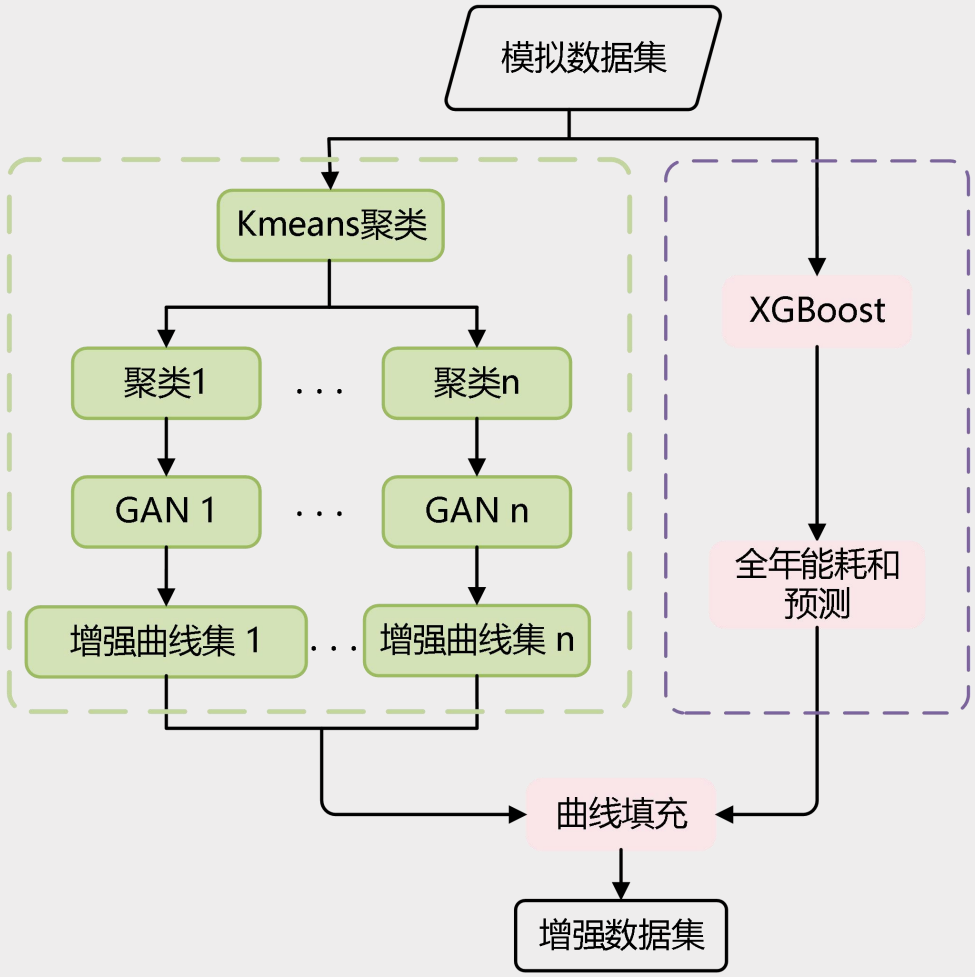
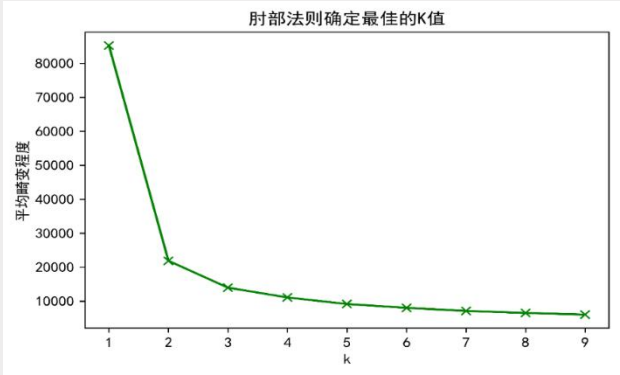
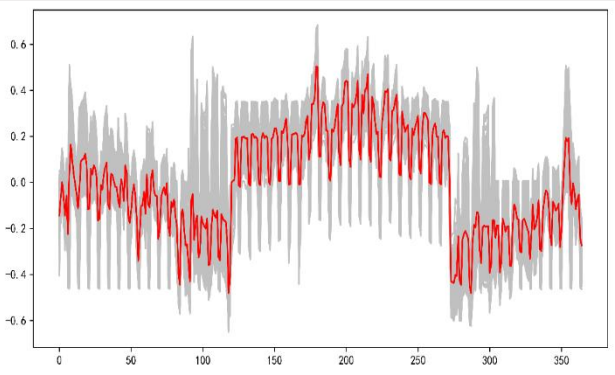
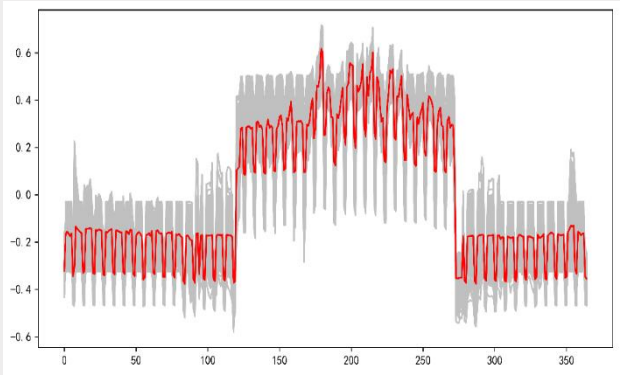


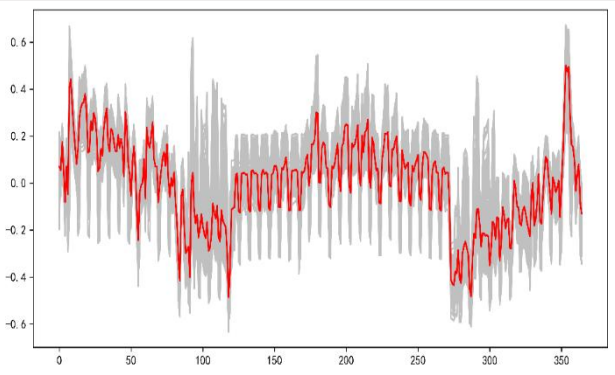
图4.1基于CGAN的数据增强技术路线图

Kmeans聚类







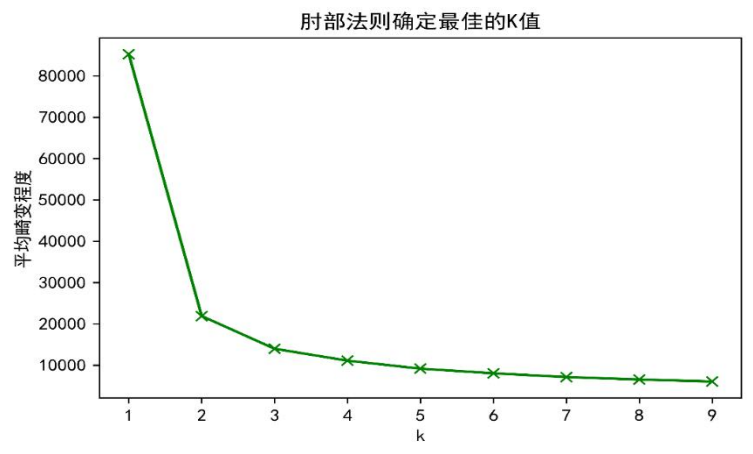


- 目的：保证生成对抗网络模型的性能以及最终得到的增强数据集的平衡性
- 肘部法则判断聚类数为3

4. 数据增强

Kmeans聚类

- 一种基于欧式距离度量的数据划分方法
- 目的：保证生成对抗网络模型的性能以及最终得到的增强数据集的平衡性
- 肘部法则判断聚类数为3



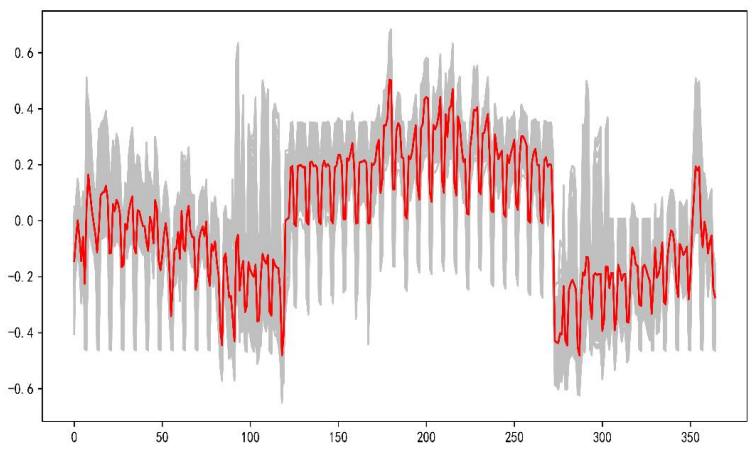
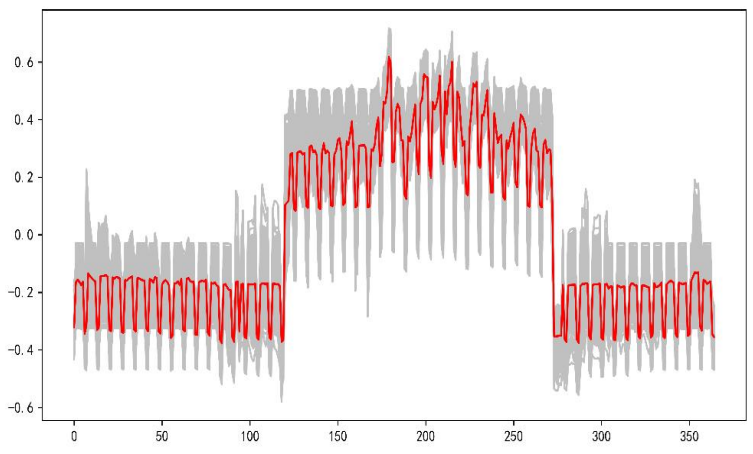


图4.3 肘部法则



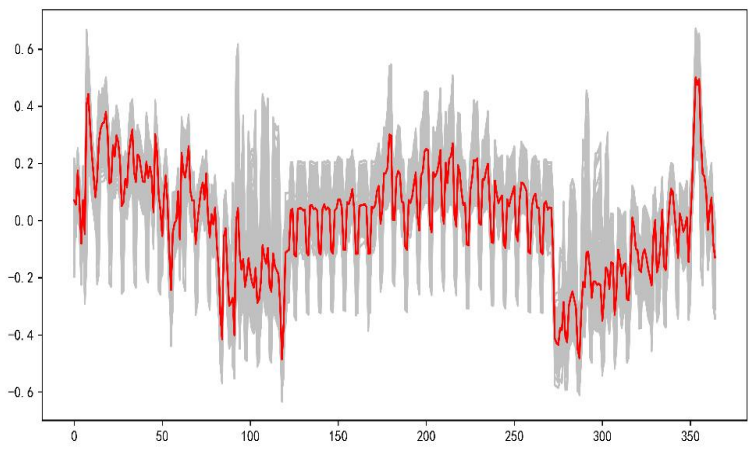


图4.4 Kmeans聚类结果

4. 数据增强

条件生成对抗网络模型的建立

条件生成对抗网络通过在生成器和鉴别器中引入额外的标签进行训练，这就使得生成的数据不仅仅是从数据分布中随机抽取的，而是在**额外条件约束**下生成的。

- 数据处理：独热编码、-1到1归一化
- 建筑静态参数、天气参数和时间标签被当作额外的条件输入到生成器中
- 生成器的输入包括随机噪声，建筑静态参数、天气参数、时间标签，输出为建筑全年能耗数据
- 鉴别器的输入为建筑全年能耗数据，输出为对该数据是否真实的判断，1为真，0为假。

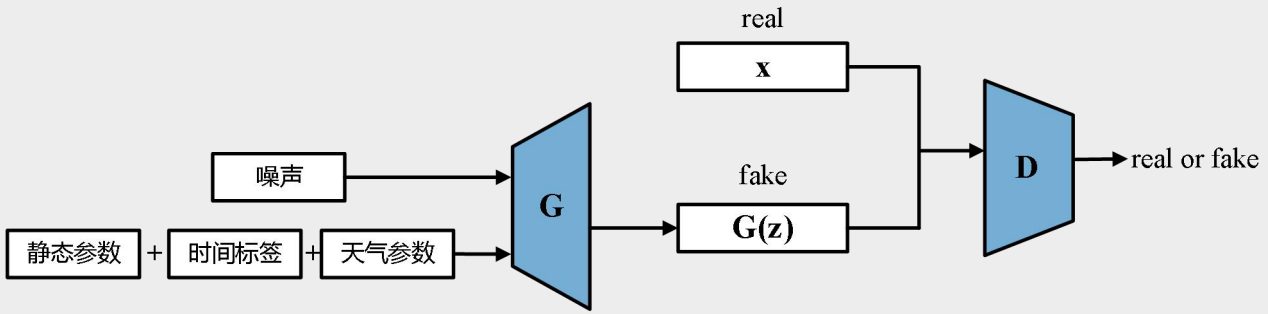
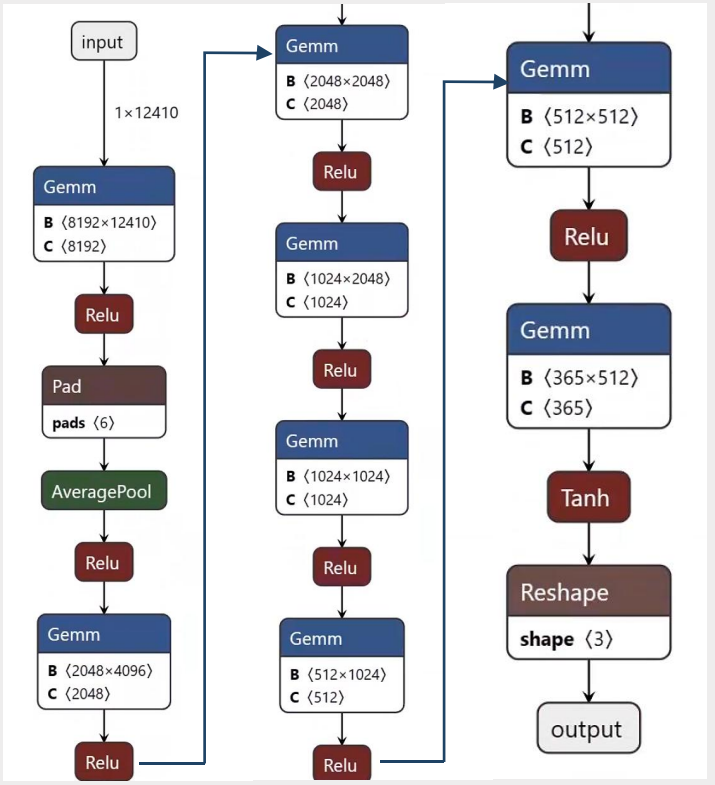
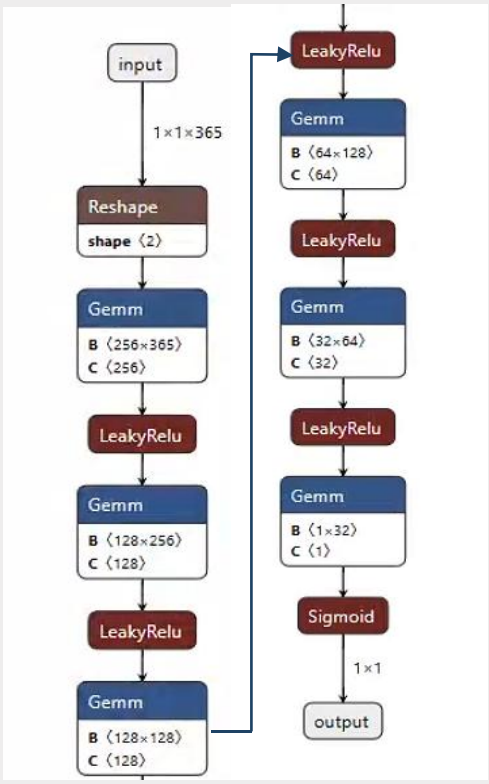


图4.5 条件生成对抗网络原理图





4. 数据增强

从收敛曲线中也可以看出训练过程也是生成器和鉴别器博弈学习的过程。

从图4.7中可以看出，数据增强在应对模拟数据中连续极度相似趋势时，可以一定程度上通过深度学习挖掘能耗曲线的内部规律来对能耗数据重新生成。

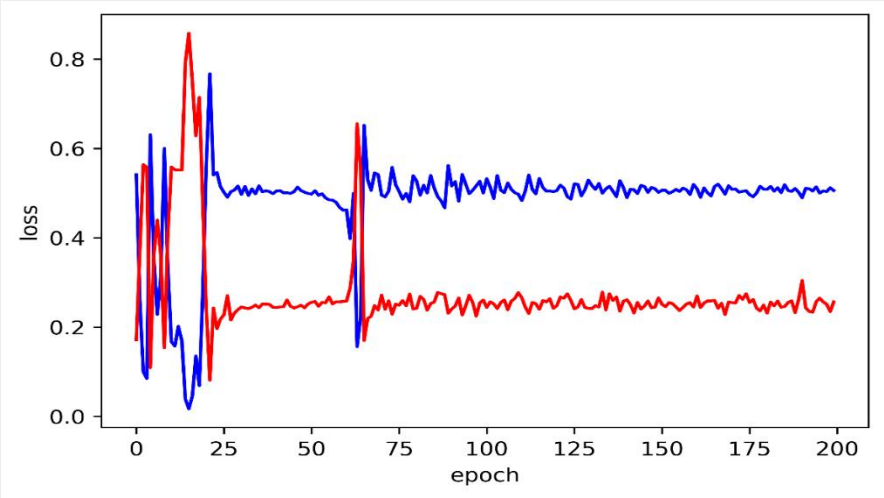
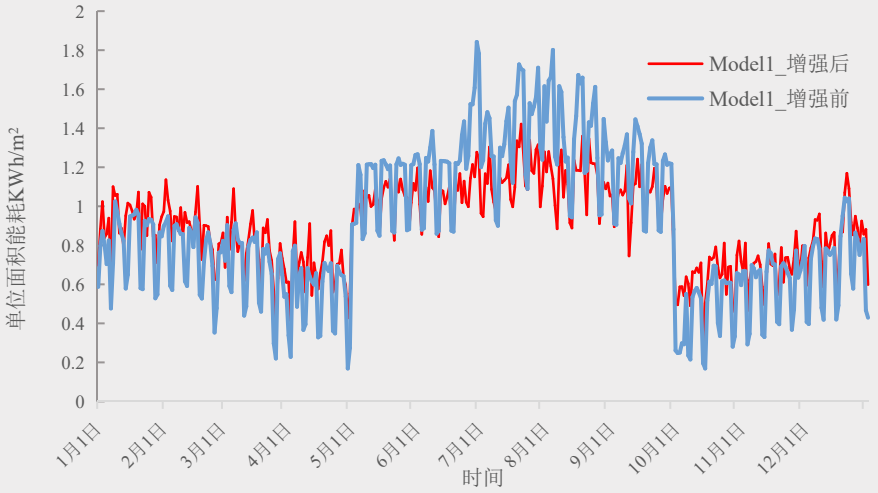
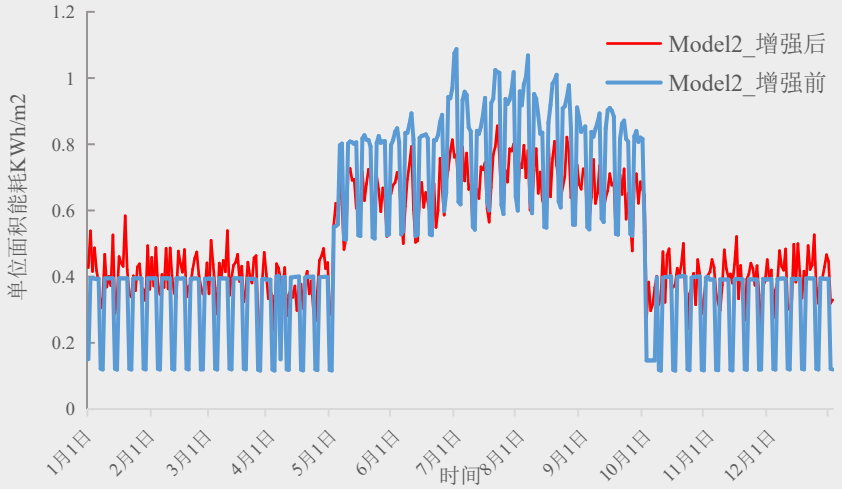


图4.6 CGAN训练收敛曲线





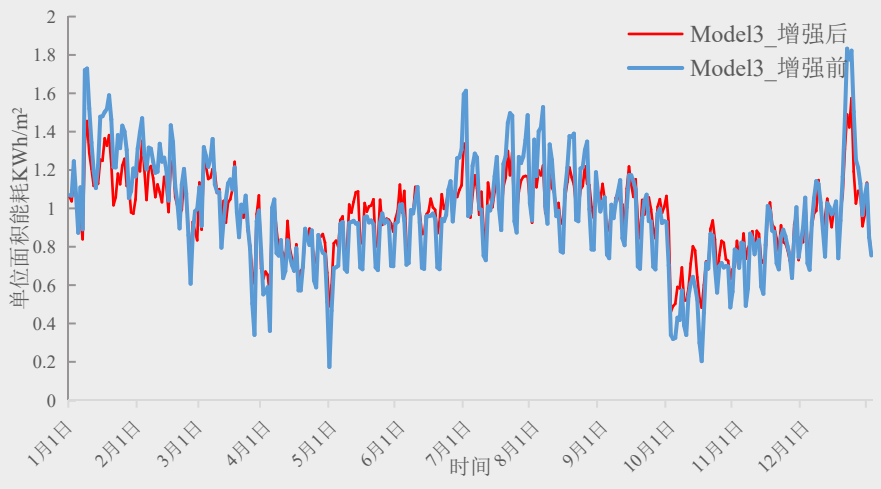


图4.7 三类曲线数据增强效果

5. 基于迁移学习的建筑能耗预测模型

5.基于迁移学习的建筑能耗预测模型

基于增强数据集的预训练模型建立

□ 数据集：增强数据集

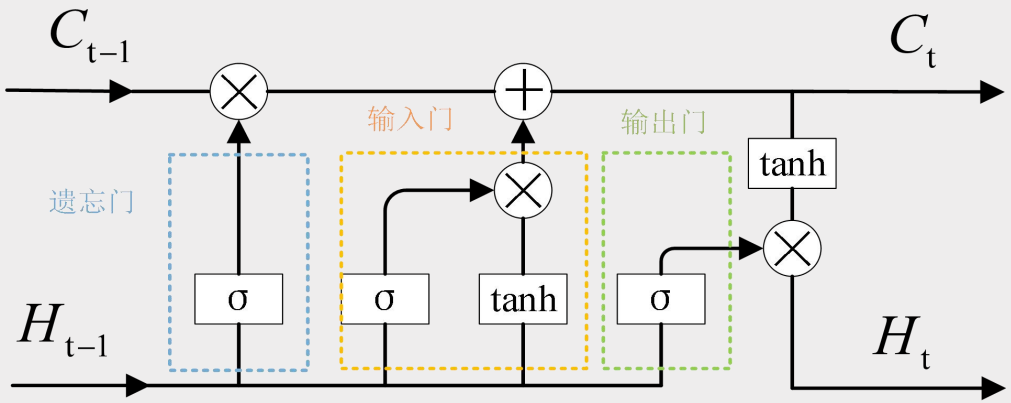
- 第四章生成的包含6000栋建筑的增强数据集
- 数据集划分：训练集、验证集、测试集的划分比例为6：2：2

□ 算法：长短时记忆网络

- 解决梯度消失问题
- 捕捉长时依赖关系
- 具有时间敏感性

□ 超参优化

- 超参优化工具：Ray.tune



tanh tanh函数　　σ sigmoid函数

参数	隐层大小	堆叠LSTM层数	drop out	学习率	批次大小	权重衰减
搜索空间	[26,24,20,16,12,10,8,6,5]	[1,2,3]	[0.1,0.2,0.3,0.4,0.5]	Loguniform (1e-5,1e-2)	[16,32,64,128]	[0.0002,0.0004,0.0006]
寻优结果	20	2	0.2	0.00944	16	0.0002

5.基于迁移学习的建筑能耗预测模型

基于增强数据集的预训练模型建立

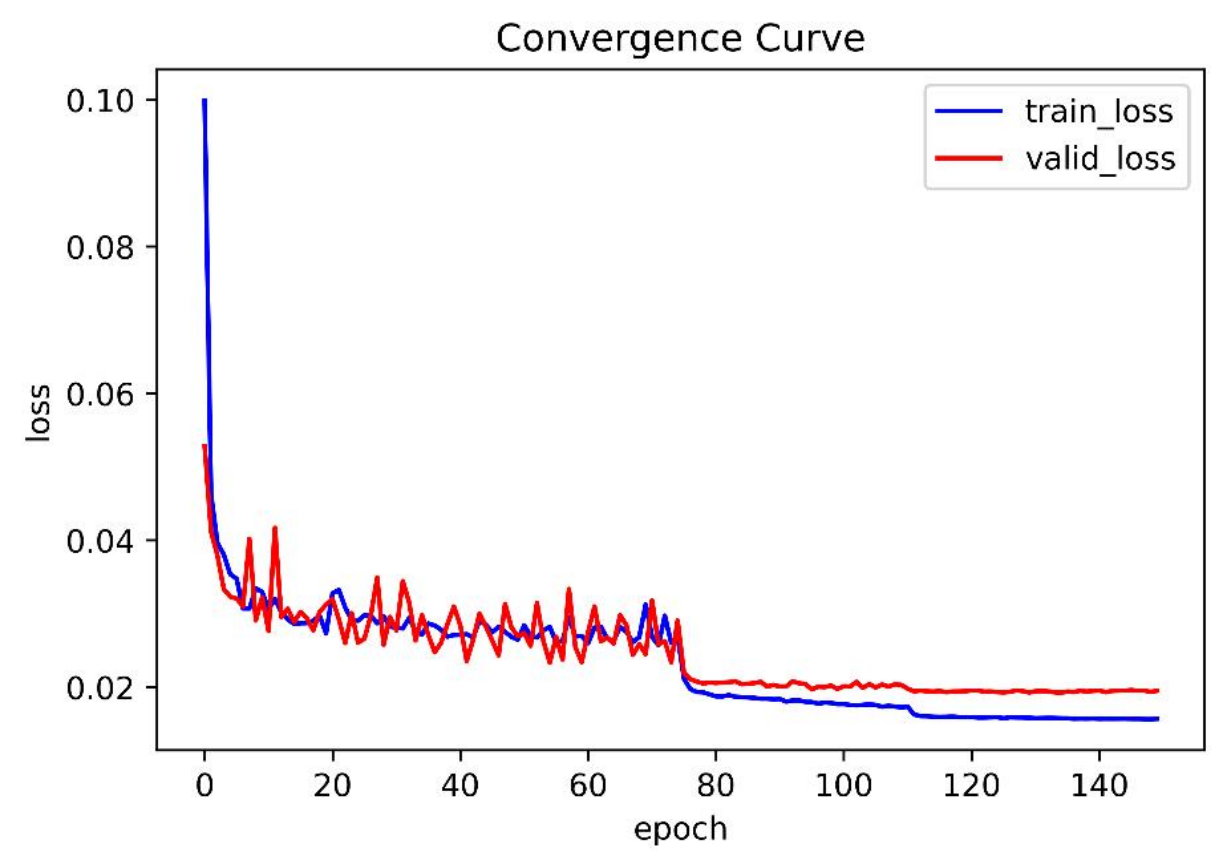
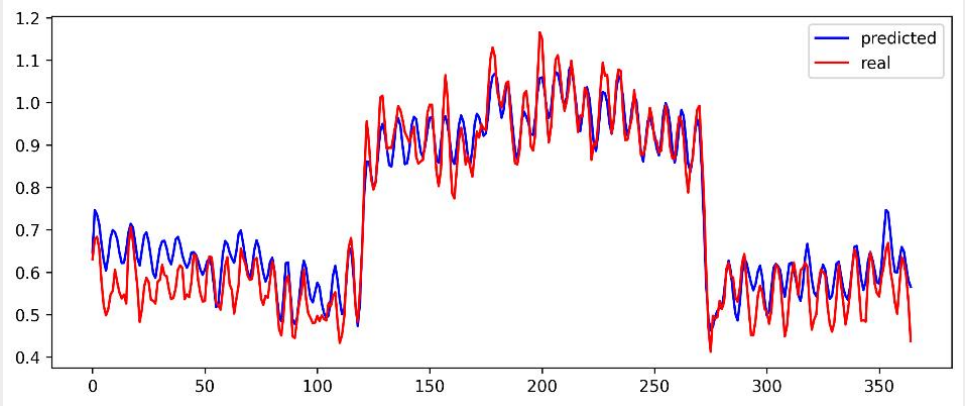


图5.3 预训练模型训练收敛曲线



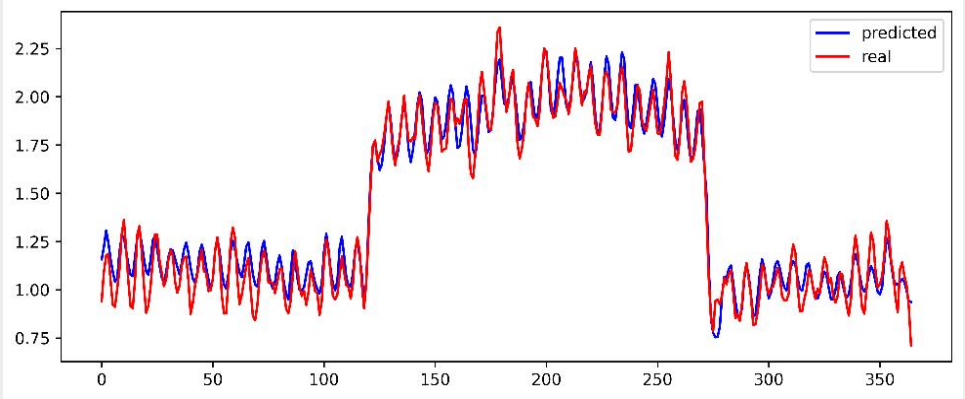


图5.4 预训练模型预测结果示例

5.基于迁移学习的建筑能耗预测模型

基于真实数据集的迁移模型的建立

数据集介绍

本研究所使用的建筑能耗数据源自上海市某能耗监测平台。能耗检测平台所提供的数据主要包括建筑基本参数（包括建筑编号、地上层数、地下层数、建筑面积、建筑类型、空调系统形式），设备清单列表，气象参数，度日数，分项计量能耗（动力、照明、空调和其他能耗）等。本研究所使用的数据集涵盖112栋建筑综合体。

图5.5 能耗监测平台页面

5.基于迁移学习的建筑能耗预测模型

基于真实数据集的迁移模型的建立

数据集

- 按照7：3划分训练集和测试集
- 存在特征缺失的建筑进行缺失值推断

迁移策略

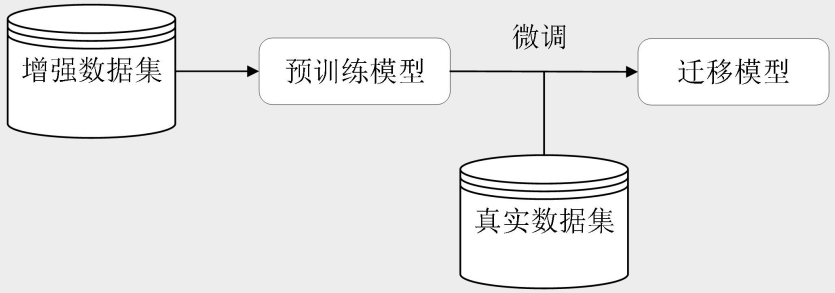
- 基于模型的迁移

模型设置

- Adam优化器
- 绝对误差损失函数（L1Loss）
- 学习率衰减策略

参数调优

- 手动调整（考虑到数据集的规模以及模型迁移所涉及的有限超参数）



超参	学习率	批次大小	权重衰减系数
调优值	0.0005	16	0.0003

模型评价指标

平均绝对误差（MAE）：预测值与实际值之间绝对偏差的平均值

$$MAE = \frac{1}{n}|y_i - f(x_i)|$$

5.基于迁移学习的建筑能耗预测模型

基于真实数据集的迁移模型的建立

口 结果

- 模型在训练集上的平均绝对误差（MAE）为0.04836，
- 在测试集上的MAE为0.05077

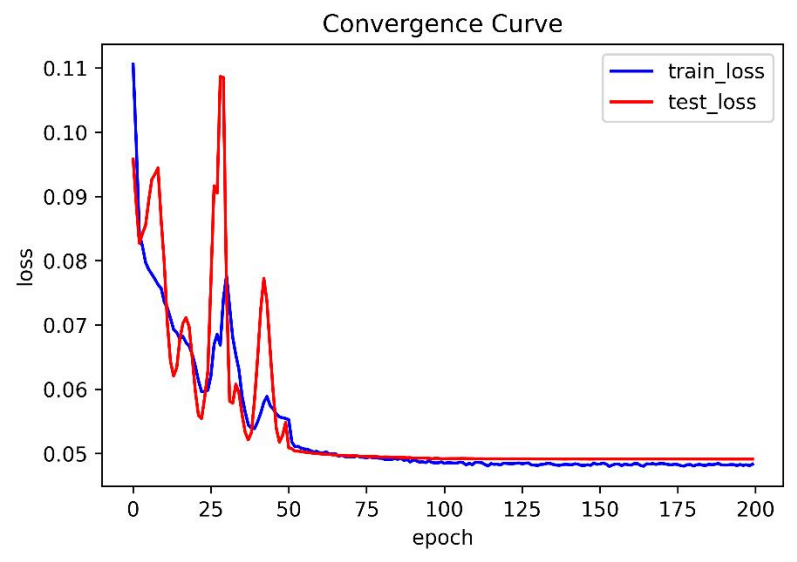


图5.6 迁移模型收敛曲线

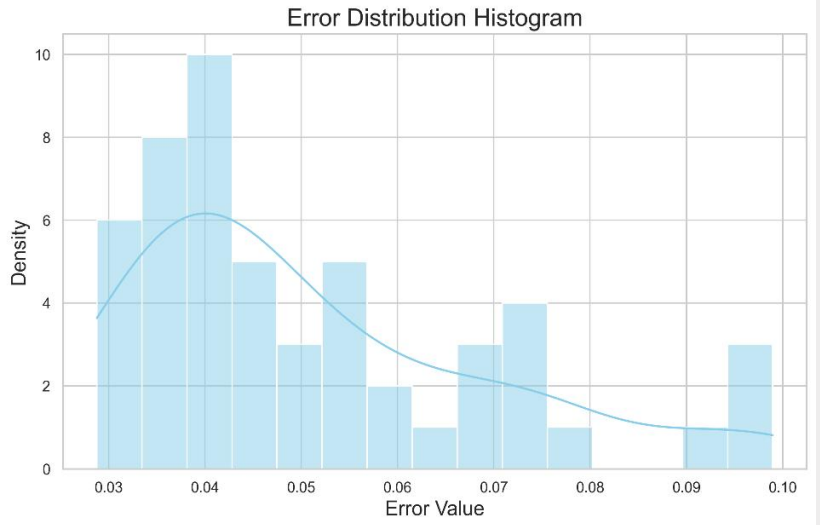


图5.7 迁移模型测试集误差分布图

6. 模型验证

6. 模型验证——迁移模型预测精度验证

表6.1 迁移模型精度验证5栋建筑基本概况

建筑编号	建筑面积（m²）	建筑类型	建筑业态
建筑1	106759	综合体	1至7层为商场，集中式全空气系统；8至31层为办公区，风机盘管系统。
建筑2	79822	综合体	1至7层为商场，集中式全空气系统；8至31层为办公区，风机盘管系统。
建筑3	96485	综合体	总共20层，风机盘管系统。
建筑4	134332	综合体	1至7层为商场，集中式全空气系统；8至16层为办公区，分体式空调或VRV的局部式机组系统。
建筑5	254000	综合体	1至9层为商场，集中式全空气系统；10至35层为办公区，风机盘管系统。

表6.2迁移模型在五栋建筑上的预测结果

建筑编号	MAE	RMSE
建筑1	0.03597	0.04661
建筑2	0.03005	0.03932
建筑3	0.02176	0.02877
建筑4	0.02314	0.03580
建筑5	0.02704	0.03482

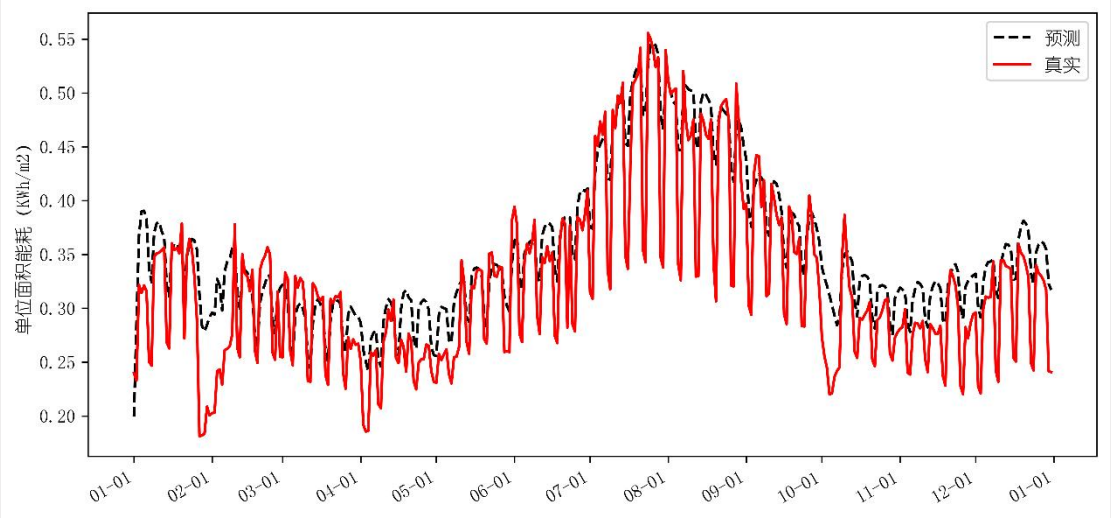
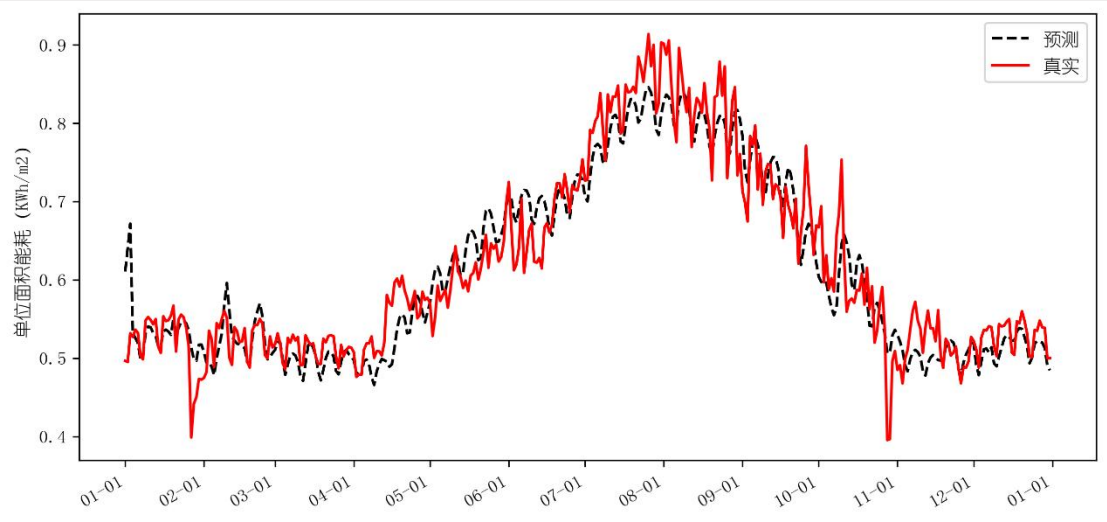
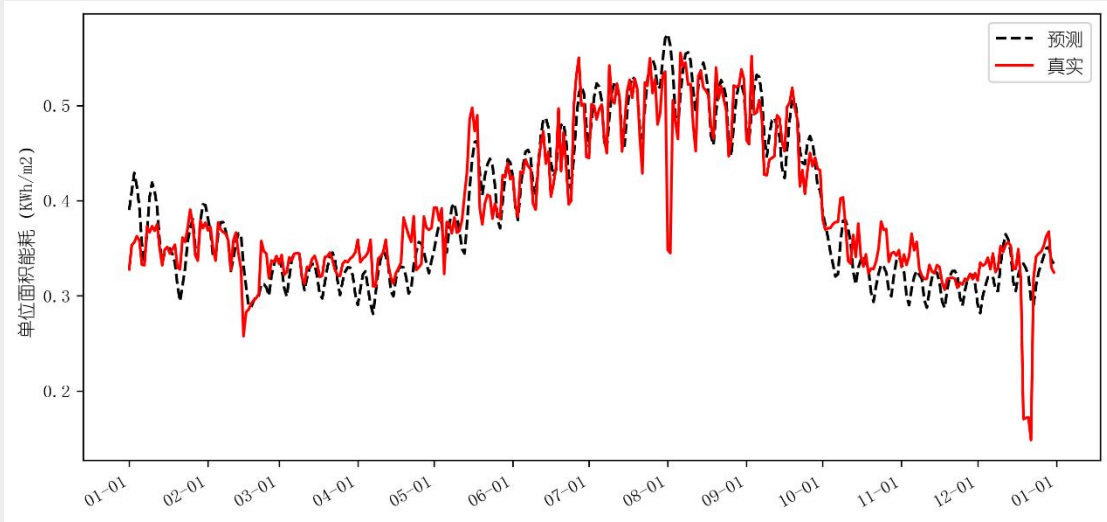
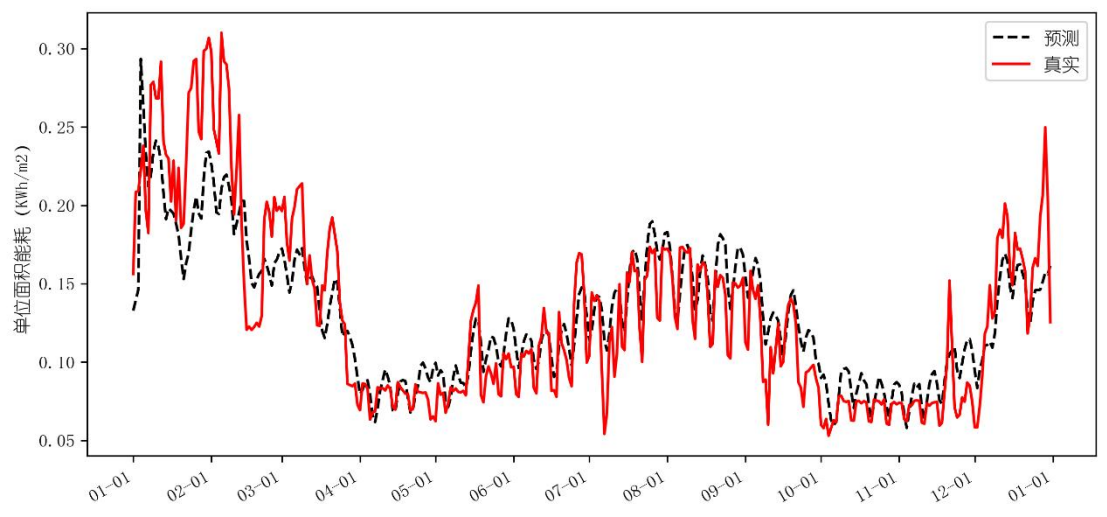


图6.1 迁移模型在建筑1上的预测结果

6.模型验证——迁移模型预测精度验证







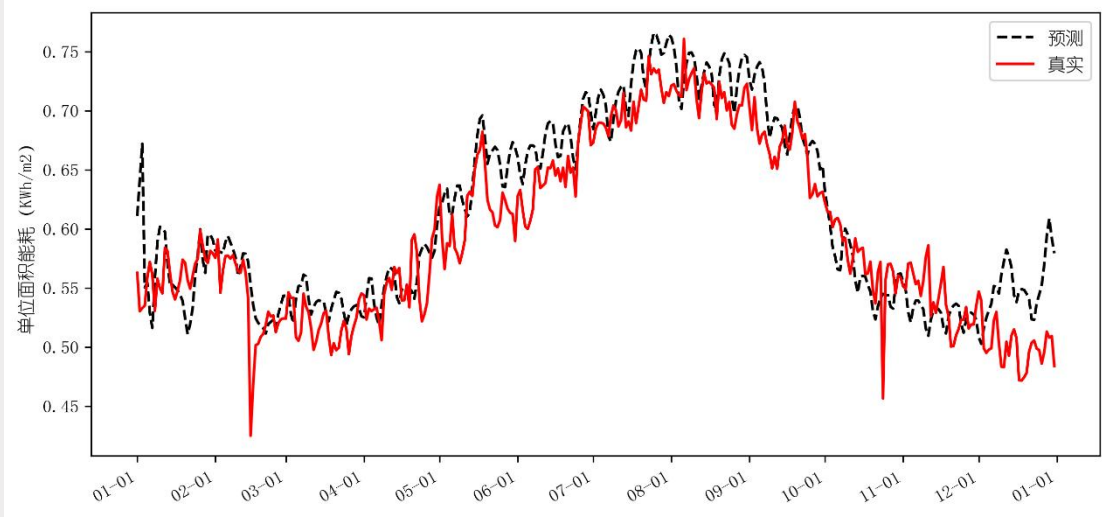


图6.2 迁移模型在建筑2、3、4、5上的预测结果

6. 模型验证——迁移模型和基础模型的对比验证

基础模型的建立

基础模型所用算法与迁移模型相同，均为长短时记忆网络；模型输入与迁移模型相同，包括建筑静态参数、天气参数以及时间标签，输出为全年逐日单位面积能耗。与迁移模型不同的是，基础模型的全程训练仅使用真实数据集。

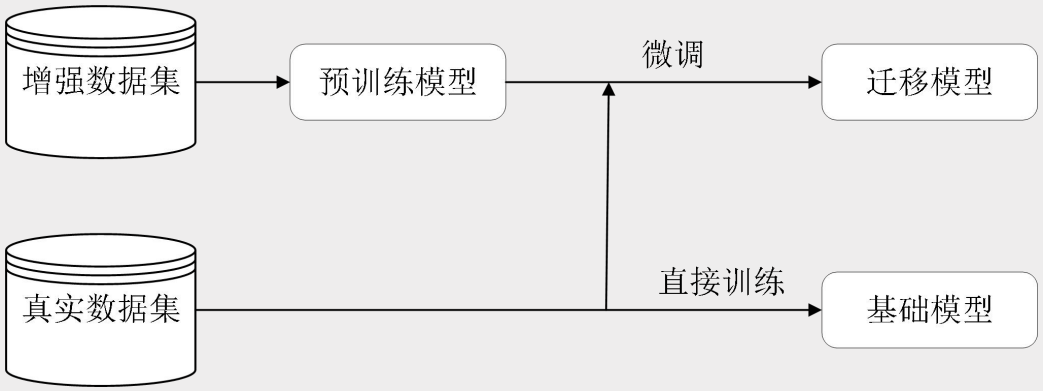


表6.3 基础模型和迁移模型预测MAE对比

模型	训练集	验证集	测试集
基础模型的平均MAE	0.03877	0.05917	0.05579
迁移模型的平均MAE	0.04836	/	0.05077

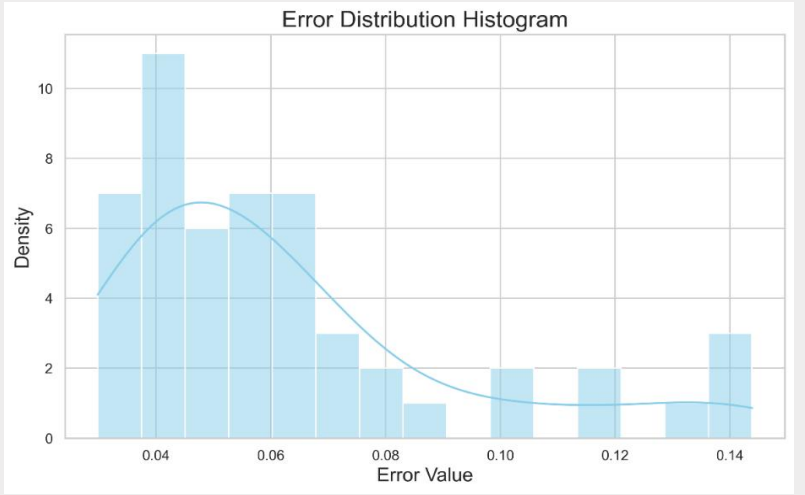


图6.3 基础模型测试集误差分布图

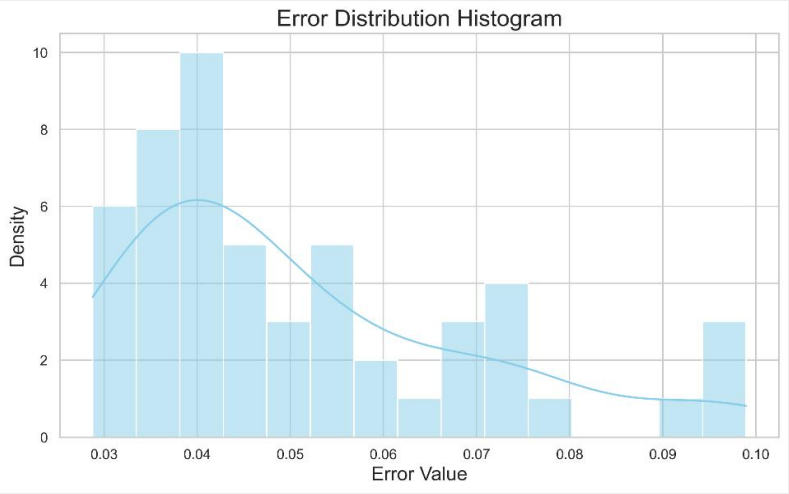
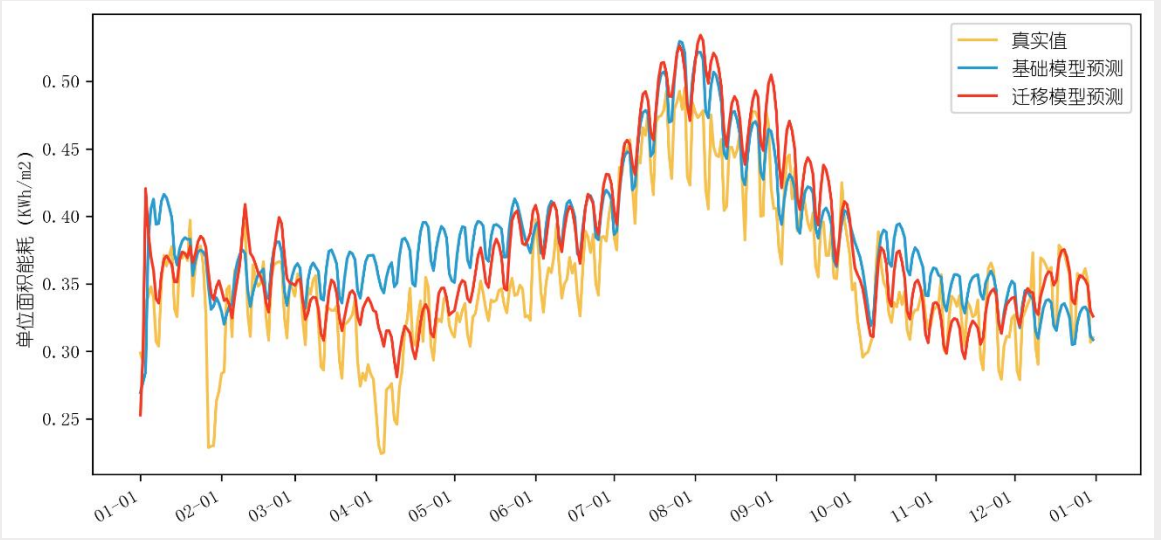


图6.4 迁移模型测试集误差分布图

6. 模型验证——迁移模型和基础模型的对比验证

表6.4 建筑基本概况

建筑编号	建筑面积（m2）	建筑类型	建筑业态
建筑a	79245	综合体	1至7层商场，集中式全空气系统；8至24层为办公楼，风机盘管系统。
建筑b	33986	综合体	1至6层为商场，集中式全空气系统；7至30层为办公区，风机盘管系统。
建筑c	56475	综合体	1至6层为商场，集中式全空气系统；7至25层为办公区，风机盘管系统。
建筑d	90000	综合体	1至7层商场，集中式全空气系统；8至16为办公区，多联机系统。
建筑e	43680	综合体	1至6层为商场，集中式全空气系统；7至24层为办公区，风机盘管系统。



迁移模型和基础模型在建筑a上的预测结果比对

表6.5 迁移模型和基础模型在5栋建筑上的预测误差对比

建筑编号	迁移模型		基础模型	
	MAE	RMSE	MAE	RMSE
建筑a	0.02687	0.03416	0.03400	0.04172
建筑b	0.07812	0.09638	0.1190	0.1408
建筑c	0.03376	0.04247	0.05107	0.06213
建筑d	0.03041	0.03933	0.05078	0.06326
建筑e	0.04277	0.05305	0.06526	0.07512

6. 模型验证

表6.4 建筑基本概况

建筑编号	建筑面积（m2）	建筑类型	建筑业态
建筑a	79245	综合体	1至7层商场，集中式全空气系统；8至24层为办公楼，风机盘管系统。
建筑b	33986	综合体	1至6层为商场，集中式全空气系统；7至30层为办公区，风机盘管系统。
建筑c	56475	综合体	1至6层为商场，集中式全空气系统；7至25层为办公区，风机盘管系统。
建筑d	90000	综合体	1至7层商场，集中式全空气系统；8至16为办公区，多联机系统。
建筑e	43680	综合体	1至6层为商场，集中式全空气系统；7至24层为办公区，风机盘管系统。

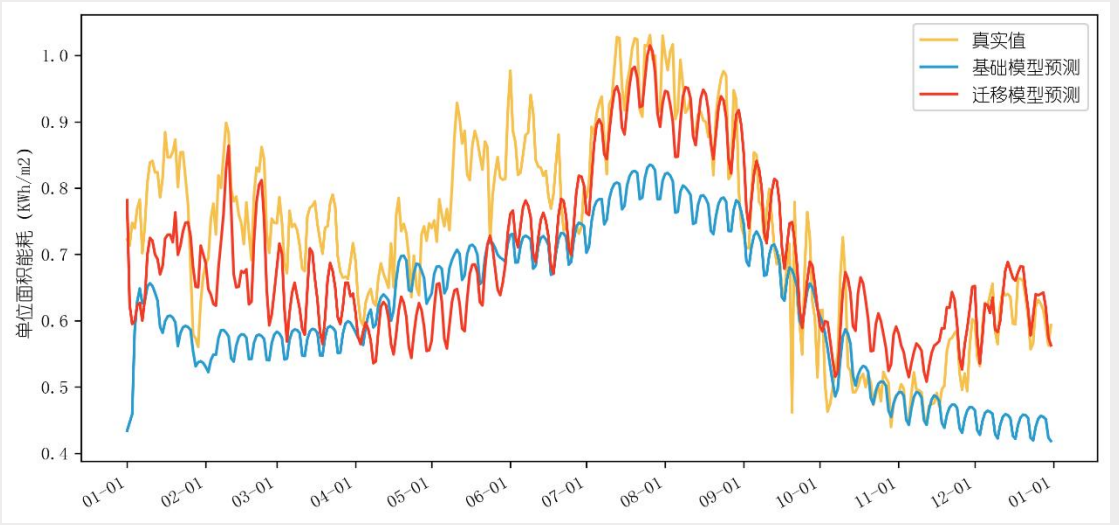


表6.5 迁移模型和基础模型在5栋建筑上的预测误差对比

建筑编号	迁移模型		基础模型	
	MAE	RMSE	MAE	RMSE
建筑a	0.02687	0.03416	0.03400	0.04172
建筑b	0.07812	0.09638	0.1190	0.1408
建筑c	0.03376	0.04247	0.05107	0.06213
建筑d	0.03041	0.03933	0.05078	0.06326
建筑e	0.04277	0.05305	0.06526	0.07512

6. 模型验证

表6.4 建筑基本概况

建筑编号	建筑面积（m2）	建筑类型	建筑业态
建筑a	79245	综合体	1至7层商场，集中式全空气系统；8至24层为办公楼，风机盘管系统。
建筑b	33986	综合体	1至6层为商场，集中式全空气系统；7至30层为办公区，风机盘管系统。
建筑c	56475	综合体	1至6层为商场，集中式全空气系统；7至25层为办公区，风机盘管系统。
建筑d	90000	综合体	1至7层商场，集中式全空气系统；8至16为办公区，多联机系统。
建筑e	43680	综合体	1至6层为商场，集中式全空气系统；7至24层为办公区，风机盘管系统。

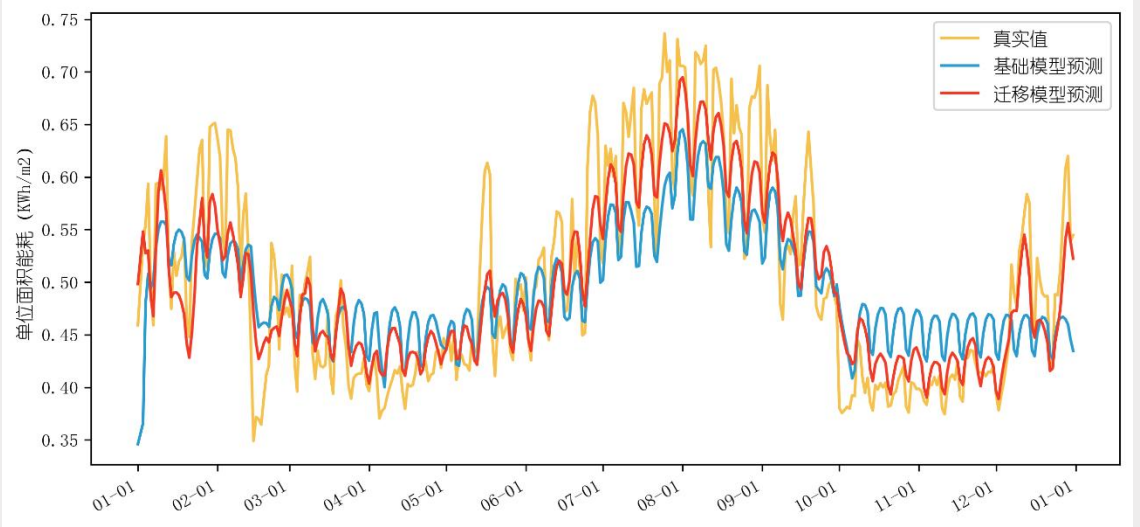


表6.5 迁移模型和基础模型在5栋建筑上的预测误差对比

建筑编号	迁移模型		基础模型	
	MAE	RMSE	MAE	RMSE
建筑a	0.02687	0.03416	0.03400	0.04172
建筑b	0.07812	0.09638	0.1190	0.1408
建筑c	0.03376	0.04247	0.05107	0.06213
建筑d	0.03041	0.03933	0.05078	0.06326
建筑e	0.04277	0.05305	0.06526	0.07512

6. 模型验证

表6.4 建筑基本概况

建筑编号	建筑面积（m2）	建筑类型	建筑业态
建筑a	79245	综合体	1至7层商场，集中式全空气系统；8至24层为办公楼，风机盘管系统。
建筑b	33986	综合体	1至6层为商场，集中式全空气系统；7至30层为办公区，风机盘管系统。
建筑c	56475	综合体	1至6层为商场，集中式全空气系统；7至25层为办公区，风机盘管系统。
建筑d	90000	综合体	1至7层商场，集中式全空气系统；8至16为办公区，多联机系统。
建筑e	43680	综合体	1至6层为商场，集中式全空气系统；7至24层为办公区，风机盘管系统。

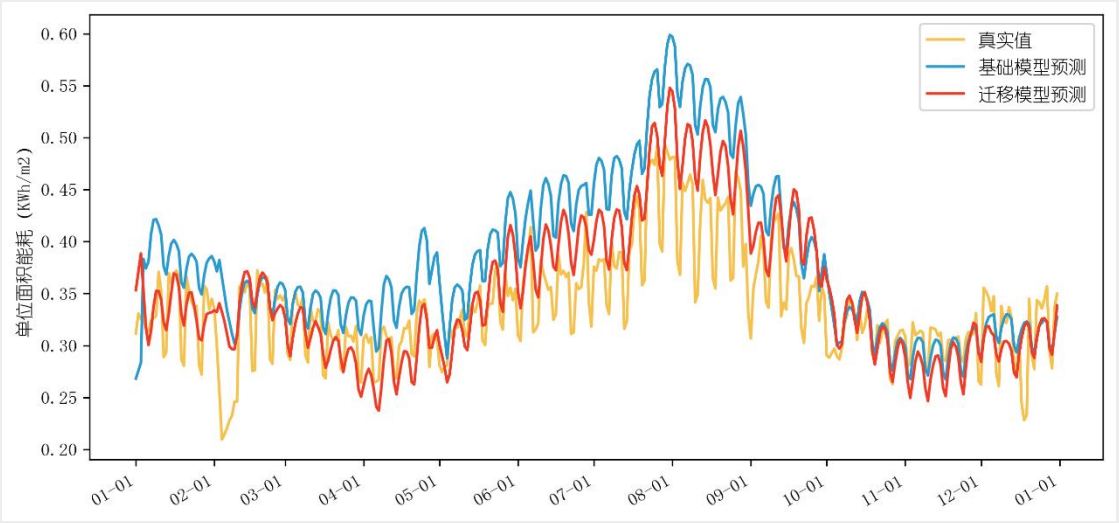


表6.5 迁移模型和基础模型在5栋建筑上的预测误差对比

建筑编号	迁移模型		基础模型	
	MAE	RMSE	MAE	RMSE
建筑a	0.02687	0.03416	0.03400	0.04172
建筑b	0.07812	0.09638	0.1190	0.1408
建筑c	0.03376	0.04247	0.05107	0.06213
建筑d	0.03041	0.03933	0.05078	0.06326
建筑e	0.04277	0.05305	0.06526	0.07512

6. 模型验证

表6.4 建筑基本概况

建筑编号	建筑面积（m2）	建筑类型	建筑业态
建筑a	79245	综合体	1至7层商场，集中式全空气系统；8至24层为办公楼，风机盘管系统。
建筑b	33986	综合体	1至6层为商场，集中式全空气系统；7至30层为办公区，风机盘管系统。
建筑c	56475	综合体	1至6层为商场，集中式全空气系统；7至25层为办公区，风机盘管系统。
建筑d	90000	综合体	1至7层商场，集中式全空气系统；8至16为办公区，多联机系统。
建筑e	43680	综合体	1至6层为商场，集中式全空气系统；7至24层为办公区，风机盘管系统。

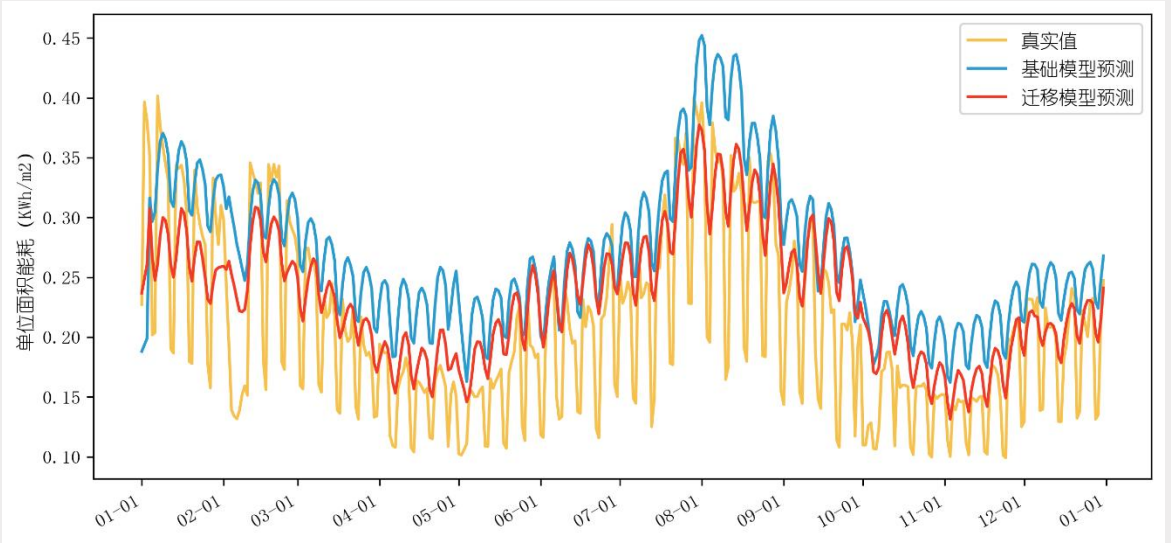


表6.5 迁移模型和基础模型在5栋建筑上的预测误差对比

建筑编号	迁移模型		基础模型	
	MAE	RMSE	MAE	RMSE
建筑a	0.02687	0.03416	0.03400	0.04172
建筑b	0.07812	0.09638	0.1190	0.1408
建筑c	0.03376	0.04247	0.05107	0.06213
建筑d	0.03041	0.03933	0.05078	0.06326
建筑e	0.04277	0.05305	0.06526	0.07512

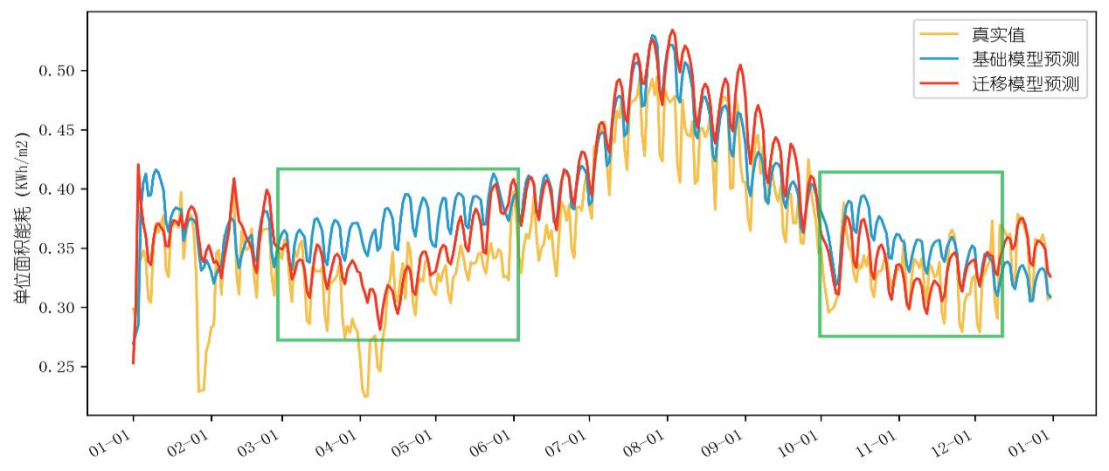
6. 模型验证

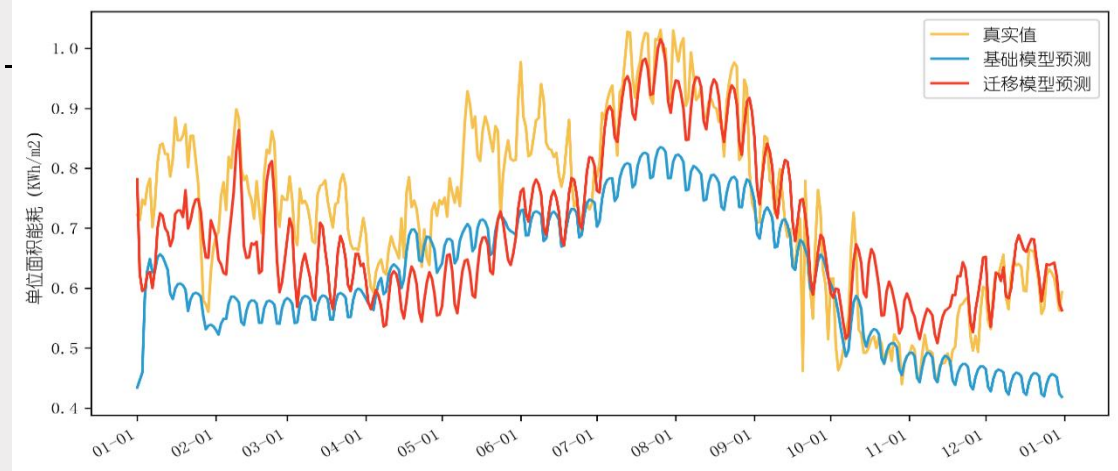
表6.6 建筑a逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1月	0.03320	0.04916	0.04784	0.06365	0.01464	0.01449
2月	0.02121	0.02869	0.01513	0.01833	-0.00608	-0.01036
3月	0.02244	0.02764	0.04152	0.04824	0.01908	0.0206
4月	0.02794	0.03689	0.07108	0.07636	0.04314	0.03947
5月	0.03226	0.03542	0.05038	0.05167	0.01812	0.01625
6月	0.03274	0.03469	0.03230	0.03519	-0.00044	0.0005
7月	0.02600	0.03011	0.02133	0.02616	-0.00467	-0.00395
8月	0.04377	0.05100	0.03056	0.03586	-0.01321	-0.01514
9月	0.03198	0.03679	0.01825	0.02137	-0.01373	-0.01542
10月	0.01916	0.02259	0.03722	0.04016	0.01806	0.01757
11月	0.01572	0.01887	0.02236	0.02560	0.00664	0.00673

表6.7 建筑b逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1月	0.10750	0.11270	0.19471	0.22243	0.08721	0.10973
2月	0.08265	0.09147	0.21566	0.22078	0.13301	0.12931
3月	0.08807	0.09448	0.15196	0.15637	0.06389	0.06189
4月	0.09192	0.10475	0.05145	0.05938	-0.04047	-0.04537
5月	0.1756	0.18634	0.13271	0.14326	-0.04289	-0.04308
6月	0.08799	0.10516	0.10319	0.12456	0.0152	0.0194
7月	0.03583	0.04321	0.15625	0.16096	0.12042	0.11775
8月	0.04514	0.05768	0.13790	0.14517	0.09276	0.08749
9月	0.06045	0.07996	0.05647	0.07411	-0.00398	-0.00585
10月	0.08409	0.09367	0.03429	0.04994	-0.0498	-0.04373
11月	0.06232	0.06587	0.04212	0.05980	-0.0202	-0.00607





6. 模型验证

表6.8 建筑c逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1月	0.04780	0.06757	0.06395	0.09118	0.01615	0.02361
2月	0.05441	0.06296	0.06125	0.07132	0.00684	0.00836
3月	0.02272	0.02491	0.03421	0.03911	0.01149	0.0142
4月	0.02418	0.02881	0.04011	0.04470	0.01593	0.01589
5月	0.03211	0.04461	0.03465	0.04702	0.00254	0.00241
6月	0.03980	0.05031	0.03918	0.05746	-0.00062	0.00715
7月	0.04009	0.04737	0.07519	0.08611	0.0351	0.03874
8月	0.04326	0.05052	0.06980	0.07941	0.02654	0.02889
9月	0.03581	0.04311	0.03564	0.04492	-0.00017	0.00181
10月	0.03345	0.03865	0.05774	0.06091	0.02429	0.02226
11月	0.01355	0.01554	0.04691	0.04887	0.03336	0.03333
12月	0.02789	0.03503	0.02578	0.03437	-0.00211	0.00066

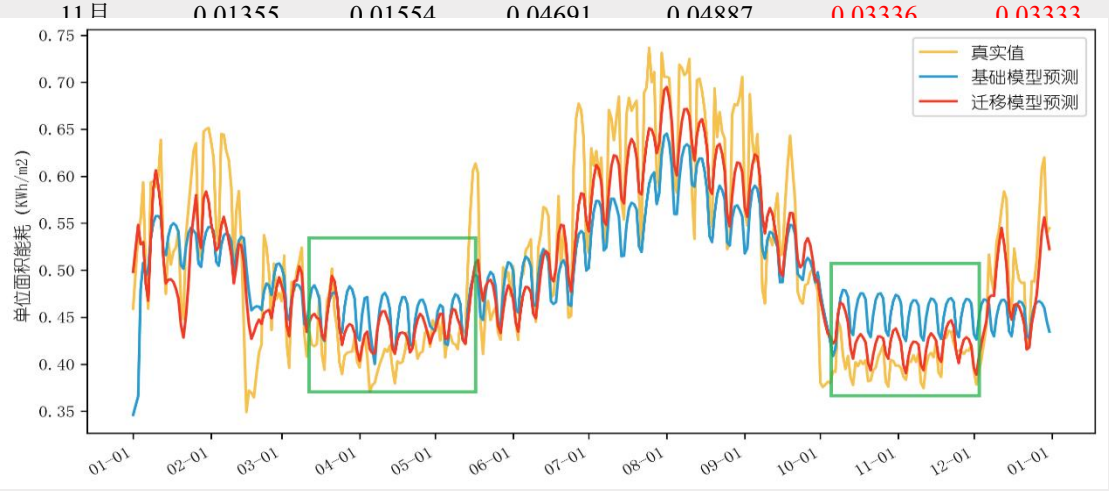
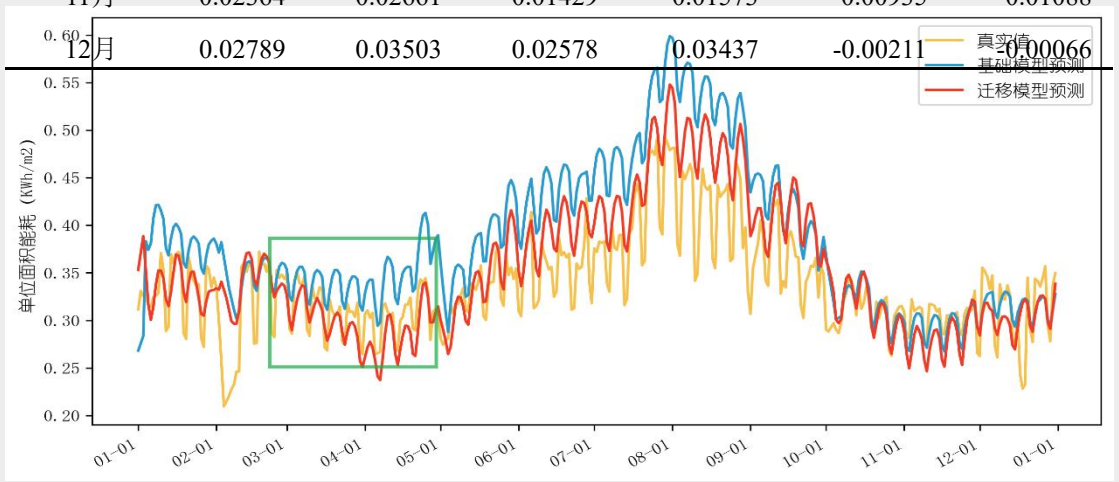


表6.9 建筑d逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1月	0.03427	0.07138	0.05654	0.06767	0.02227	-0.00371
2月	0.03820	0.05136	0.04618	0.06384	0.00798	0.01248
3月	0.01497	0.01840	0.02726	0.03118	0.01229	0.01278
4月	0.02144	0.02572	0.04796	0.05086	0.02652	0.02514
5月	0.01947	0.02571	0.04764	0.05372	0.02817	0.02801
6月	0.03961	0.04485	0.07764	0.08073	0.03803	0.03588
7月	0.03628	0.04227	0.08873	0.09163	0.05245	0.04936
8月	0.06363	0.06937	0.1145	0.1178	0.05087	0.04843
9月	0.04330	0.05041	0.05400	0.05904	0.0107	0.00863
10月	0.01649	0.02452	0.01614	0.02370	-0.00035	-0.00082
11月	0.02364	0.02661	0.01429	0.01573	-0.00935	-0.01088
12月	0.02789	0.03503	0.02578	0.03437	-0.00211	0.00066



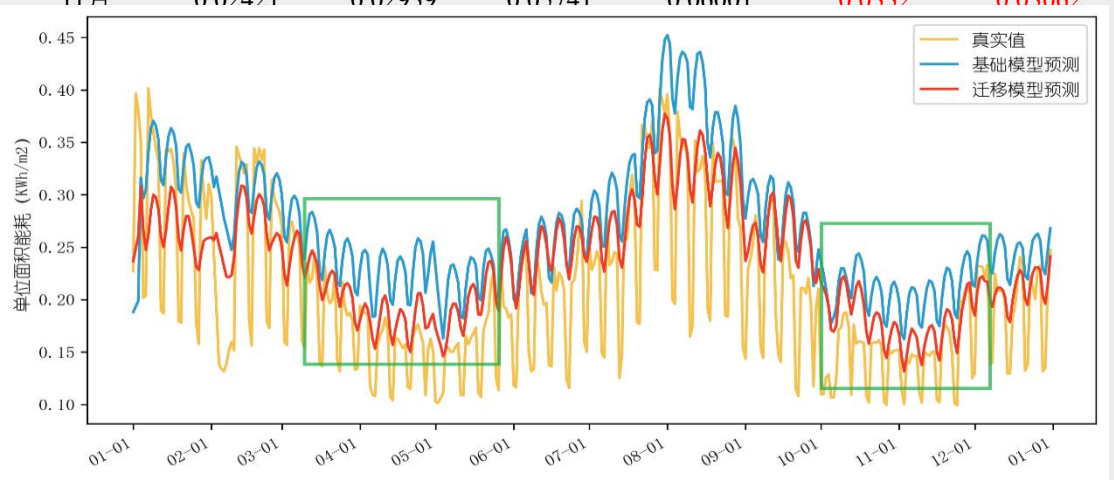
6. 模型验证

迁移模型相比基础模型性能的提升，主要来源于过渡季。

- 对于建筑a，迁移模型在3至5月和10月、11月的预测误差低于基础模型；
- 对于建筑c，迁移模型在3至5月和10月、11月的MAE和RMSE比基础模型低了30%左右；
- 对于建筑d，迁移模型在3至5月预测误差小于基础模型，虽前者在10月、11月预测误差高于后者，但MAE和RMSE差距均在0.001左右；
- 对于建筑e，迁移模型在3至5月和10月、11月的MAE和RMSE均比基础模型低了30%以上。
- 对于建筑b，迁移模型在4、5月份和10、11月份的预测误差反而比基础模型稍大

表6.10 建筑e逐月预测误差

月份	迁移模型		基础模型		差值	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1月	0.06449	0.09378	0.06327	0.07932	-0.00122	-0.01446
2月	0.05927	0.06619	0.06476	0.08374	0.00549	0.01755
3月	0.03176	0.03921	0.05608	0.06173	0.02432	0.02252
4月	0.02873	0.03357	0.07475	0.07620	0.04602	0.04263
5月	0.04548	0.05021	0.06730	0.07167	0.02182	0.02146
6月	0.05230	0.06204	0.05777	0.06487	0.00547	0.00283
7月	0.04386	0.05423	0.06100	0.07215	0.01714	0.01792
8月	0.04997	0.06563	0.09513	0.10976	0.04516	0.04413
9月	0.05117	0.06397	0.06754	0.07630	0.01637	0.01233
10月	0.04514	0.05107	0.06784	0.07091	0.0227	0.01984
11月	0.02421	0.02939	0.05741	0.06001	0.0332	0.03062

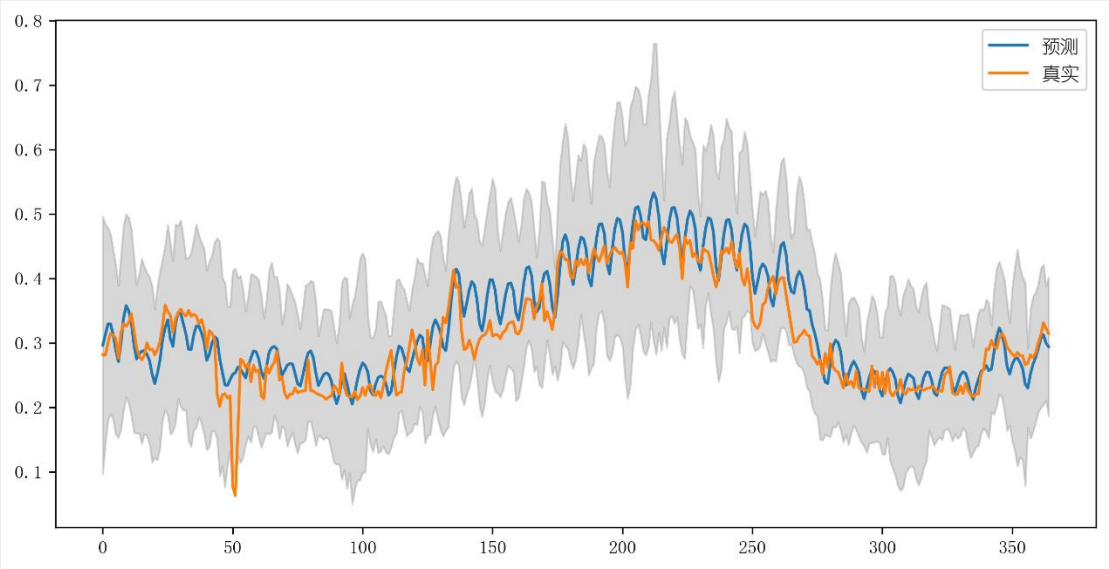


6. 模型验证

迁移模型预测不确定性

本研究采用如下方法衡量建筑能耗预测的不确定性：

- 首先计算模型测试集样本误差的标准差，标准差代表了模型预测的不确定性水平，即模型的预测偏离真实值的平均程度；
- 其次确定置信水平，本文选择置信水平为95%；将标准差乘以所选置信水平对应的临界值，以计算置信区间的宽度；
- 最后，将计算得到的置信区间宽度添加到模型的预测结果中。



7. 总结与展望

7. 总结与展望

主要贡献

- 提出了基于自编码器的缺失关键变量推断算法。利用降噪自编码器挖掘关键变量之间的关联关系，进而进行缺失关键变量的推断，解决了现实中建筑信息缺失无法预测能耗这一常见问题，实现了综合体建筑关键变量缺失情景下的能耗预测。
- 提出了基于条件生成对抗网络的数据增强方法，利用聚类来保证了生成对抗网络模型的性能。
- 利用迁移学习融合了增强数据和真实数据
- 实现了综合体建筑的无历史能耗情景下的能耗预测。

与课题组之前研究的不同

- 缺失关键变量推断部分，提出了基于自编码器的缺失关键变量推断算法
- 提出了基于条件生成对抗网络的数据增强方法，利用聚类来保证了生成对抗网络模型的性能。
- 数据融合部分，利用迁移学习来融合增强数据和真实数据，
- 实现了综合体建筑的无历史能耗情景下的能耗预测。

7. 总结与展望

展望

- 建筑空调系统的运行是一个非常复杂的问题，系统的连接方式及控制逻辑难以参数化表示，故本文的初始变量集仅根据专家知识选择，无法囊括所有建筑能耗影响因素。此外提取得到关键变量后，未对关键变量的相关性做进一步研究。
- 本文的预测模型无法准确预测突发情况下的建筑能耗。本文在进行初始变量集的选取时，并未包括一些非常规因素，比如建筑由于一些不可控因素临时停业、疫情关闭等，因此该模型无法准确预测这种突发情况下的能耗。
- 由于真实数据缺失，本研究在第三章基于遗传算法的关键变量推断中所使用的能耗预测模型是利用模拟数据集训练得到的，这可能会对推断结果产生了负面影响。

恳请各位老师批评指正

答辩人：夏壮　　指导老师：许鹏教授

二〇二四年三月