一、分析方法概要

本队采用数据驱动的方式进行目标建筑的能耗预测。考虑到目标建筑无任何能耗数据进行算法反馈和校正，且能耗数据集数量有限，本队建立的预测方法主要基于**邻近时域加权平均思想**。本次分析主要采用 R 和 Python 语言实现，分析方法示意图如图 1 所示，核心内容如下：

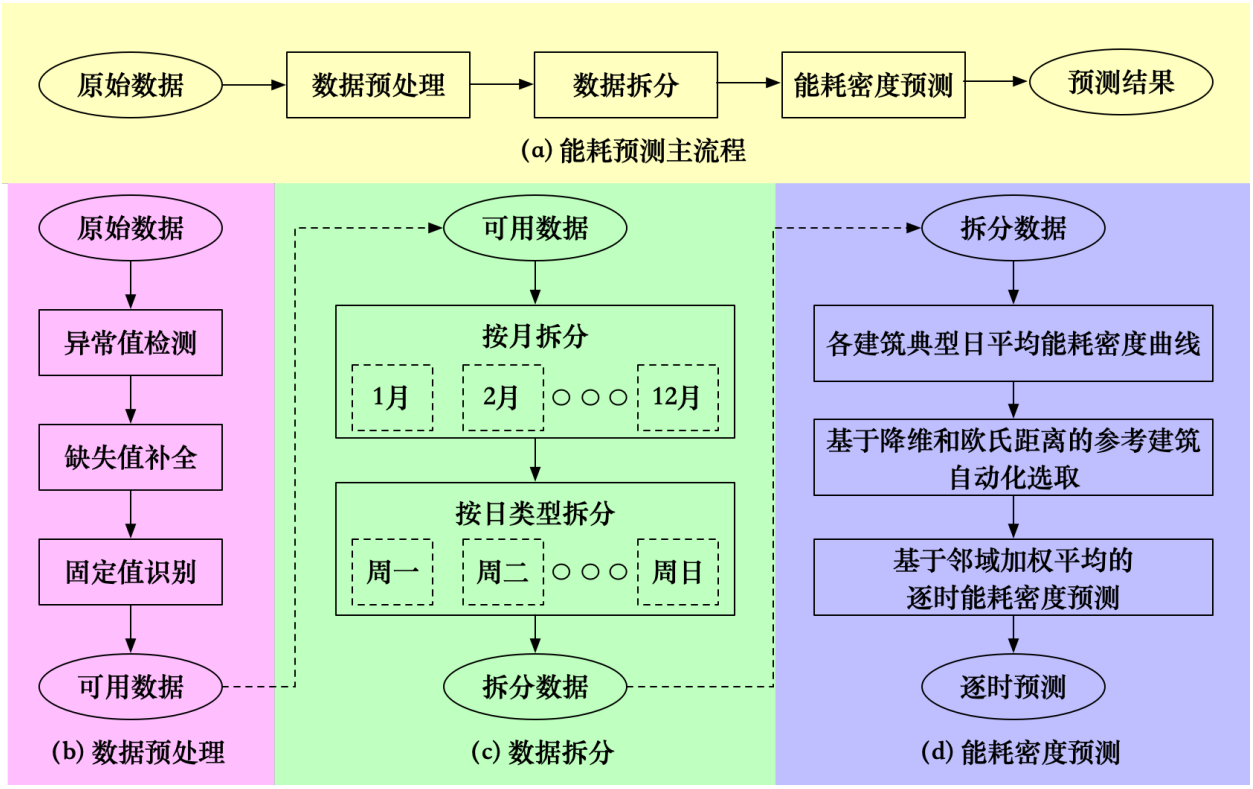

图 1 分析方法流程示意图

1. 建筑能耗数据预处理

从能耗密度的角度对 20 栋建筑进行能耗数据预处理，主要包含异常值检测、缺失值补全和静态固定值识别三类任务。

- 其中异常值检测采用直接限制法，自动将超过 99%分位数的能耗值设定为缺失值。
- 针对缺失值补全问题，采用该建筑相同月份、相同日类型、相同时刻的能耗均值作为补全值。

- 针对静态固定值识别，采用数据可视化方法对每栋建筑能耗进行分析，如该建筑在特定月份存在大量连续的静态固定值，则将该建筑该月份数据剔除。

2. 基于邻近时域加权平均的能耗预测方法

- 第一，对数据集按照特定月份、特定日类型（即周一至周日）、特定分项能耗进行拆分。
- 第二，建立基于数据降维和欧式距离的建筑典型日能耗相似度量化方法，自动选取典型日能耗曲线较为正常的若干栋建筑作为参考建筑，筛除掉明显的离群建筑。
- 第三，考虑相邻时间间隔的能耗时序关系，对参考建筑的能耗密度数据进行邻近时域加权平均。比如时间窗设定为 3 小时，则目标建筑 2017 年 1 月 1 日凌晨 2 点的空调能耗预测值为自动选取的若干栋参考建筑在 2017 年 1 月 1 日凌晨 1 点、凌晨 2 点和凌晨 3 点的加权平均能耗密度值。该能耗密度预测值乘以目标建筑面积即得到能耗预测值。

二、分析方法详述

1. 建筑能耗数据预处理

本部分内容涉及三类子任务，即异常值检测、缺失值补全和静态固定值识别，具体细节如下：

1) 首先，对能耗密度（即能耗值除以建筑面积）数据进行异常值检测。采用直接限制法识别能耗密度过大的数据，使用 99%分位数作为临界值，任何大于该数值的能耗密度值将被设定为缺失值。
2) 其次，对缺失值进行补全处理，采用该建筑相同月份、相同日类型、相同时刻的能耗密度均值作为补全值。比如，假设 1 号建筑在 2017 年 1 月 1 日（星期五）上午 10 点的空调能耗密度值缺失，则计算该建筑在 2017 年 1 月份的所有星期五上午 10 点的非缺失数据的平均值，以此进行缺失值补全。
3) 最后，识别能耗密度数据中的连续静态固定值。采用数据可视化的方法对每栋建筑按照照明和空调能耗进行分析，如某一建筑在特定月份存在大量连续的静态固定值，则后续分析不适用该建筑该月份的能耗密度数据。图 2 展现了 9 号建筑在 1 月份的能耗

数据，其中包含大量静态固定值，因此该部分数据应予以剔除。本次分析剔除的建筑数据总结如表 1 所示。

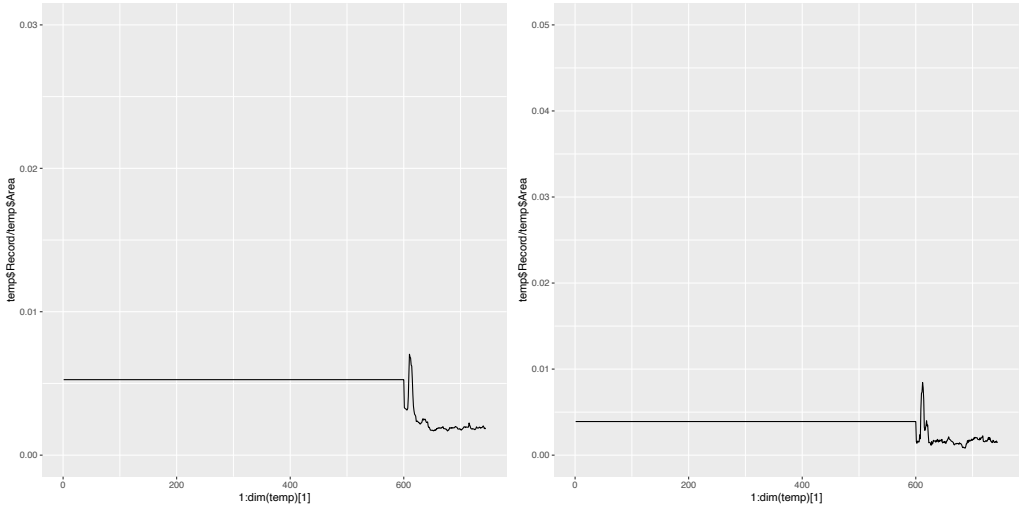

图 2 静态固定值范例：9 号建筑 1 月份的照明及空调密度数据

能耗类型	建筑序号	月份
照明能耗	2	1、4、11、
	9	1
	10	2、4、6、9、10
空调能耗	2	1、4、11
	9	1
	10	2、6、9、10、11

表 1-存在连续静态固定值的建筑能耗密度数据总结

2. 基于邻近时域加权平均的能耗预测方法

建筑运行具有明显的周期性，因此在计算目标建筑能耗密度时应充分考虑时间变量的影响。在本次分析中，我们选取了月份和日类型（即周一至周日）作为时间变量，通过时间变量交叉组合细化分析维度。由此，我们将目标建筑的能耗预测问题分解为 168 个子问题（即 12 个月份乘以 7 个日类型乘以 2 个能耗类型）。对于每一个子任务，全部采用邻近时域加权平均的方法进行能耗预测，该方法的关键问题在于：第一，如何选取可靠的参

考建筑用于加权平均计算；第二，如何定义邻近时域，并且如何设定最优权重进行加权平均。针对以上两个关键问题，本文的解决方案如下：

1） 参考建筑的自动化选取

对于每一类子问题（即某一特定月份、日类型和能耗类型的组合），首先建立 20 栋建筑的典型日能耗密度曲线。比如，对于 1 号建筑，它在 1 月份星期一的典型日空调能耗密度曲线为其在一月份所有星期一空调能耗密度曲线的逐时平均值，即一个长度为 24 的向量。

其次，为了自动化识别离群建筑，我们建立了基于数据降维和欧氏距离的相似度量化方法：第一，将日能耗密度曲线分成 3 个时间段，即 0-7 点、8-18 点、19-23 点；第二、针对各个时间段，计算能耗密度的均值作为特征数值。由此，可以将原始 24 维向量转化为 3 维特征向量，以避免欧氏距离计算中可能面临的“维度灾难”问题[1,2]；第三，基于 20 栋建筑的 3 维特征向量，计算特征向量均值，并以此为依据计算各建筑到特征向量均值的欧氏距离；第四，将20组欧式距离由大到小进行排列，剔除数值最大的5栋建筑，并结合表 1 的结果选取不超过 15 栋建筑作为参考建筑。

图 3 展示了该方法的参考建筑自动化选取效果，其中左侧为 20 栋建筑在 7 月份星期一的空调能耗密度曲线，右侧为方法自动选取的 15 栋建筑能耗密度曲线。可见该方法具有明显效果，可以自动剔除空调能耗模式明显不同的建筑。

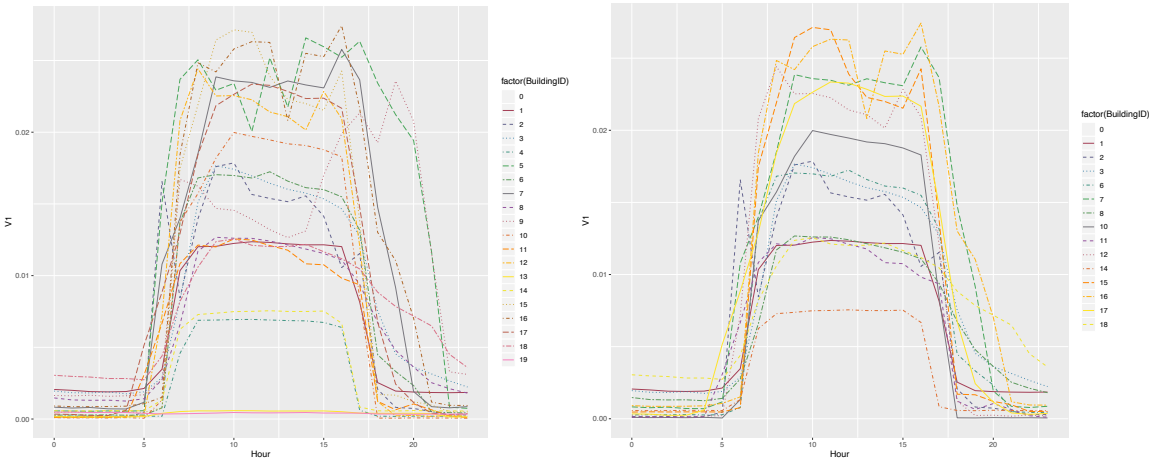

图 3 针对 7 月份星期一空调能耗的参考建筑选取结果

2） 邻近时域加权平均法

目标建筑在特定时刻的照明和空调能耗密度可以根据自动选取的参考建筑数据计算而得。考虑到建筑能耗可能具有突变性（如空调能耗在早晚通常存在突然增加或减少的情况），本文采用邻近时域加权平均的方法进行数据平滑，以提升预测的稳定性。该方法的基本思想是选取预测点前后 1 个小时为邻近时域，对时间窗内参考建筑的能耗密度进行加权平均计算，三个时间点的权重设定值分别为 0.15、0.7、0.15。

以图 3 为例，当计算目标建筑在 2017 年 7 月 10 日（星期一）上午 10 点的空调能耗密度时，选取{0、1、2、3、6、7、8、10、11、12、14、15、17、17、18}号建筑在 2017 年 7 月 10 日（星期一）上午 9 点、10 点和 11 点的平均空调能耗密度为基础，并按照权重进行加权，得到目标建筑该时刻的空调密度预测值。该密度值与目标建筑面积的乘积即为该时刻目标建筑的空调能耗预测值。

三、创新性及心得思考

1. 方法创新性

1） 本方法充分考虑建筑运行的周期特性，将整体预测问题按照特定月份、特定日类型和能耗类型分解为$12 \times 7 \times 2 = 168$个子问题，以提升预测精细化水平；
2） 建立基于特征工程和欧式距离的参考建筑自动选取方法，以数据驱动的方式自动剔除能耗模式明显不同的建筑，进而提升预测精度；
3） 建立了邻近时域加权平均的能耗密度预测方法，充分考虑建筑设备系统运行中的突变和动态变化，通过邻近时域加权平均的方法对预测值进行平滑过滤，以降低预测误差。

2. 心得思考

在本次比赛中，本队以能耗密度为基础，采用平均法进行能耗预测，先后在以下三方面进行了尝试和改进，心得思考如下：

1） 在能耗密度预测方法中，尝试了简单平均法和邻近时域加权平均法，结果显示邻近时域加权平均法可以提升预测精度。

2) 针对邻近时域加权平均法的权重进行了尝试，时间窗锁定为 3 小时，分别尝试了[0.20,0.60,0.20]、[0.15,0.70,0.15]、[0.10，0.80，0.10]的权重组合，结果表明[0.15,0.70,0.15]的表现最好，[0.20,0.60,0.20]的表现最差。一方面，该方法的主要思想是通过数据平滑降低某一栋建筑能耗突变对最终预测结果的影响。另一方面，过大的平滑效果不利于跟踪建筑能耗的变化趋势动态变化，因此在实际预测中应当选取适中的权重组合。
3) 本此分析通过时间变量（即月份和日类型）的组合将整体问题分解为 168 个子问题（即 12 个月份乘以 7 个日类型乘以 2 个能耗类型）。此外，我们尝试将问题按照每一天进一步细化，即 730 个子问题（365 天乘以 2 个能耗类型），所得能耗精度略低，其原因是根据每日能耗密度进行参考建筑选取具有一定巧合性，自动选取的参考建筑可能并非最优组合。

四、R 语言代码详解

针对本文整体分析方法的实施关键点，可参考以下信息：

1. 经过数据预处理的文件命名为 ED01E10006_data_clean.csv。
2. 基于以上 csv 文件，可执行 R 语言脚本 ED01E10006_Rscript.R 生成预测结果，其中：
   a) 20 栋建筑的典型日能耗密度特征提取，见 ED01E10006_Rscript.R 第 73-80 行；
   b) 基于欧式距离的参考建筑自动化选取，见 ED01E10006_Rscript.R 第 90-100 行；
   c) 邻域加权平均预测方法，见 ED01E10006_Rscript.R 第 221 和 239 行。

参考文献

1. Ferhatosmanoglu H, Tuncel E, Agrawal D, et al. High dimensional nearest neighbor searching[J]. Information Systems, 2006, 31(6): 512.
2. Beyer K, Goldstein J, Ramakrishnan R, et al. When is “nearest neighbor” meaningful?[C]. International conference on database theory. Springer, Berlin, Heidelberg, 1999: 217-235.