建筑能耗大数据清洗与预测方法

陈旺，张云翼，林佳瑞

（清华大学土木工程系，北京 100084）

【摘　要】建筑能耗预测建筑节能调控与用能优化具有重要意义。然而，目前有关建筑能耗预测方法只针对特定的数据类型、省略数据清洗过程，且往往实现过程较为复杂，可借鉴性较差。本研究提出一个预测建筑能耗的完整流程，详细论述了流程的四个步骤：确定主要影响因素、脏数据识别与标记、数据清洗、模型构建与对比，并对比了流程中可以采用的几种清洗大数据和构建预测模型的方法。最后通过一个工程实例验证了该流程的科学性与有效性。

【关键词】建筑节能，能耗预测，数据清洗

1　背景

建筑能耗约占全球所有能耗的 30%，且由于建筑能耗产生的二氧化碳排放约占全球二氧化碳总排放量的 1/3[1]。近些年来建筑能耗相关的研究受到越来越多的关注，尽管目前建筑宏观能耗的统计和预测等方面的研究日益增长，但仍存在诸多不足，例如预测方法仅针对特定数据类型，步骤繁琐难以推广等，致使国家和政府在制定相关政策和规划时缺乏科学依据，在一定程度上阻碍了建筑节能工作的推进[2]。而建筑用能中电力是主要能源形式之一，以 2012 年全国民用建筑能源消耗为例，电力消耗占建筑能源消耗总量的 94.4%[3]。因此基于已有数据对建筑能耗做出合理的评估和预测尤为重要，可以为能源政策制定提供科学依据、推进建筑节能工作的开展。

很多学者通过确定影响建筑能耗的主要因素，并选取相应的方法构建预测模型，达到预测建筑能耗的目的，做了诸多探索。Jialin Wu 等[4]根据上海 130 栋公共建筑的能耗数据，得出建筑能耗与建筑不同功能分区的面积之间的多元线性回归方程，为预测多功能公共建筑能耗提供了一种可行的方法。李璐等[5]确定温度、湿度、风速、日照时数、天气情况和节假日 6 个因素作为影响建筑电耗的主要因素，并提出一种基于遗传算法和神经网络的公共建筑电耗预测模型，可以较为准确地预测公共建筑电耗。Zhitong Ma 等[6]确定包含天气数据和经济因素在内的 7 个因素为影响建筑能耗的主要因素，并通过 SVR（support vector regress）方法预测中国南方的建筑能耗。由以上文献可知，目前预测建筑能耗的技术手段大致分为五步：确定主要影响因素、获得并处理建筑能耗数据、构建预测模型、测试模型和模型应用。但是现有的研究中，各种预测方法只针对特定的数据类型，省略数据清洗处

【基金项目】国家重点研发计划(2017YFC0704200)，清华大学自主科研计划（2019Z02UOT），国家自然科学基金（51908323）

【作者简介】陈旺(1999−)，男，本科生。主要研究方向为数字防灾减灾。E-mail：chenwang17@mails.tsinghua.edu.cn

理过程，且往往实现过程较为复杂，可借鉴性较差。

基于以上问题，本研究提出一个预测建筑能耗的完整流程，为预测建筑能耗和推进建筑节能工作提供科学合理的依据。

2 能耗数据清洗及预测方法

基于大数据预测建筑能耗时，首先需要确定主要影响因素，即模型的影响变量。然后通过识别与标记脏数据并进行数据清洗，得到连续完整的数据集。最后基于大数据构建预测模型，并对比不同模型的精度，选择精度最高的模型预测建筑能耗，具体流程如图 1 所示。

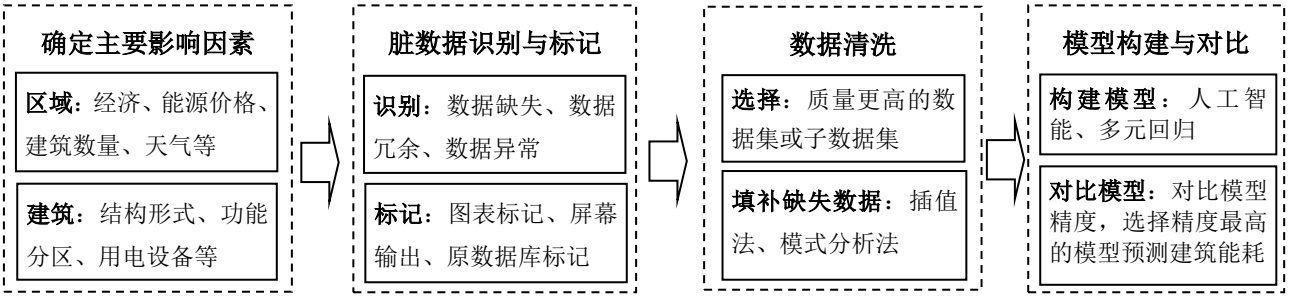


图 1 建筑能耗预测流程

2.1 确定主要影响因素

影响建筑能耗的因素有很多，例如建筑区域内的经济、能源价格、建筑数量、天气情况等，具体建筑的结构形式、用电设备、居住者的用能习惯等都会影响建筑能耗。而很多数据（如经济、用电设备参数等）获取难度较大，且对模型的精度影响较小，可以通过替代变量（如时间序列、空间序列等）间接考虑这些因素的影响。此外，并非数据越多、考虑的影响因素越多，模型的精度就越高，而是根据需求确定主要影响因素，从而构建更加精准的预测模型。

2.2 脏数据识别与标记

详实、准确的数据是建筑能耗预测的基础，如果不能保证数据质量，将无法得到任何可靠的结论。因此在清洗和处理建筑能耗数据前，应该识别和标记脏数据，从而在整体上把握数据质量，为进一步的数据处理做准备。建筑能耗数据大多通过实地测量获得，由于测量仪器较多、测量周期较长，仪器读数出现异常的概率很大；且在整合数据时，也经常出现数据遗漏或重复的情况，导致建筑能耗数据中包含较多的缺失数据、冗余数据以及异常数据。而建筑能耗数据通常按一定规律分布，如按楼层分布、等时间间隔分布等，故很容易通过程序化的方法来识别脏数据，并将脏数据的类型、位置、数量等可视化，从而直观展示原始数据的质量和分布规律。常见的脏数据可视化方法如下：

(1) **图表标记。**即用图表的方式展现数据分布规律，并标记脏数据（如将脏数据全部置零），便于从整体上把握数据质量。

(2) **屏幕输出。**即程序执行过程中，直接在屏幕上输出脏数据的类型和位置，便于精准定位脏数据及数据清洗。

(3) **原数据库标记。**即在原始数据库的相应位置标记脏数据，便于在数据共享或数据重复利用时根据需求直接进行数据清洗。

2.3 数据清洗

根据 2.2 中的可视化结果可以判断原始数据质量，从而选择数据质量更高的数据集或子数据集进行清洗，以便构建更加精确的预测模型。在对选择的数据集进行清洗时，首先删除冗余和异常数据，然后再填补缺失数据。常用的数据填补方式有插值法和模式分析法两种。插值法计算简单，在缺失数值不多的情况下可以快速得到结果。常用的插值法有拉格朗日插值、分段线性插值、三次样条插值等，其中样条插值可以通过低阶多项式实现较小的插值误差，得到了广泛应用。模式分析法是对历史各年同期能耗数据规律进行分析，并与缺失值前后的用能数据进行对比，找到用能模式相同的历史数据，利用加权平均的方法进行填补[7]。这种方法对历史数据的要求较高，若历史同期也存在数据缺失，则可以考虑利用同一时间不同区域的用能数据进行模式关联和分析。

2.4 模型构建与对比

构建建筑能耗预测模型的方法大致分为人工智能和多元回归两种[2]。多元回归法计算简便，节约计算资源，能够快速获得对预测值的估计，而缺点在于预测结果依赖于人为设定的回归方程的形式。人工智能法推广能力好，可以应用到各类预测问题中，且预测结果精度更高。但是缺点在于模型的复杂度较高，因此计算资源需求较大，而且对样本的数量也有较高的要求，很容易过拟合，反而无法得到正确的预测结果。

构建预测模型后，通过模型预测一定时间范围内的建筑能耗，并将预测结果与已知数据比较，检验模型的精度。对比不同模型的精度可以确定最佳的数据清洗方式和预测模型，并将精度最高的预测模型用于实际预测建筑能耗。

3 工程实例

本研究选取湖南长沙向日葵广场的能耗数据，该建筑的总面积为 149386m^2，数据形式为间隔 15min 的建筑总用电量表头值，原始数据如图 2 所示。

3.1 确定主要影响因素

公共建筑电耗与天气、节假日、当地经济、建筑功能分区等因素有关，但是准确获取当地经济和建筑功能分区等数据难度较大。此外天气数据虽然在构建模型时容易获取，但是在使用模型进行预测时，天气作为影响变量仍需要通过预测才能得到，会增大预测结果的误差。故考虑时间序列和节假日两个主要影响因素，即将前一天的用电量和当天是否是节假日作为影响变量。

3.2 脏数据识别与标记

由于原始数据等时间间隔分布，故很容易程序化识别脏数据，并将脏数据全部置零，

然后通过图的方式直观显示数据分布规律及脏数据的数量，如图 3 所示。

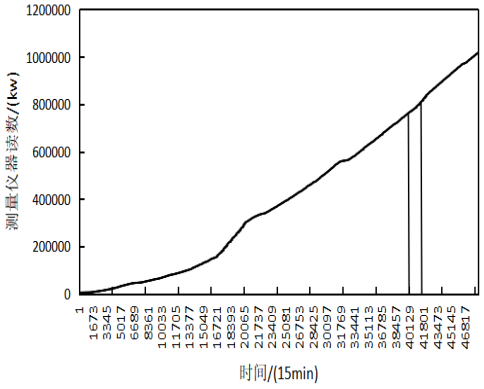


图 2 原始数据

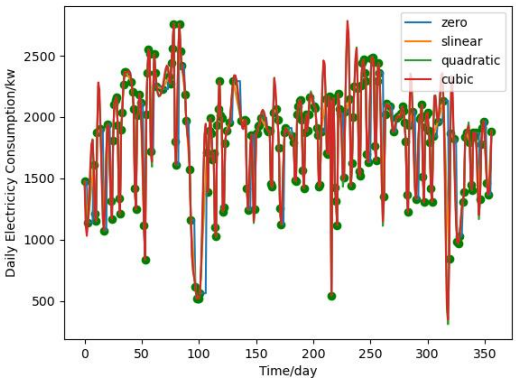


图 4 数据清洗后建筑每日耗电量

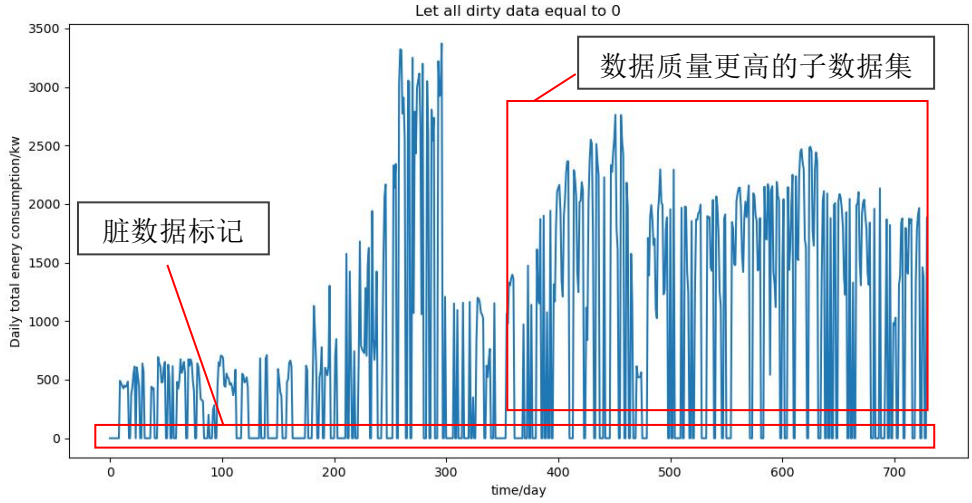


图 3 数据清洗前建筑每日耗电量

3.3 数据清洗

由图 3 可以看出，前后建筑用能的规律不一致，可能是局部建筑功能或建筑能源结构中途发生改变所致，因此在构建预测模型时不考虑前半部分的电耗数据，而是选择后半部分数据质量更高的子数据集。数据清洗时，首先删除异常数据和冗余数据，然后填补缺失数据。由于历史数据较少，采用模式分析法填补数据误差较大，且缺失数据不多，故用插值法填补数据。分别通过样条插值中的梯度插值（zero）、线性插值（slinear）、二次（quadratic）和三次 B 样条曲线插值（cubic）填补缺失数据，得到清洗后的数据集，结果如图 4 所示。

3.4 模型构建与对比

由于建筑电耗数据量并不大，采用人工智能的方式构建模型容易出现过拟合的现象，

反而导致预测精度不高，故考虑通过多元回归的方式构建模型。以前一天的用电量和当天是否是节假日为影响变量，通过多元线性回归的方式构建预测模型，回归方程如下：

$$E = a + bE_0 + cD \tag{1}$$

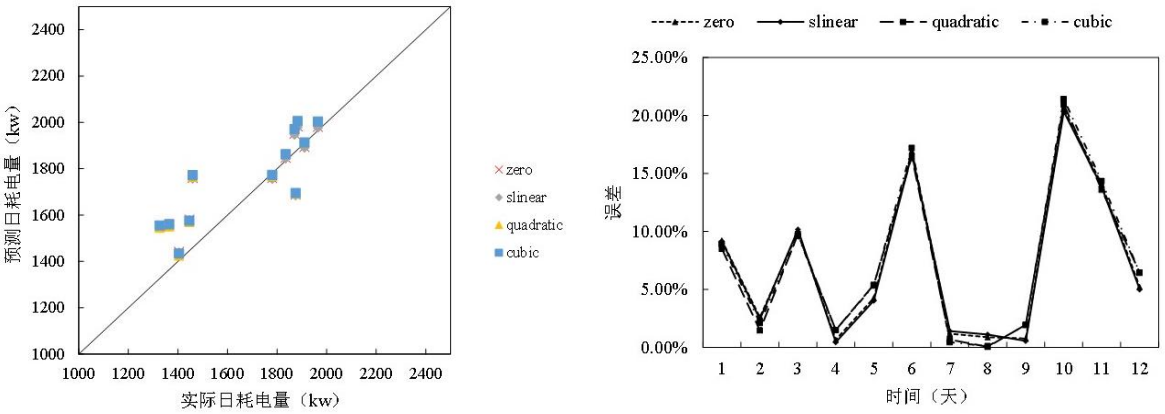
式中，E 为待预测的建筑电耗；E_0 为前一天的用电量；D 表示当天是否是节假日，若是，D=1，否则 D=0；a、b、c 为常数。

四种不同的插值方式得到的数据集的回归结果如表 1 所示。四种回归结果总体相近，其中阶梯插值得到的数据集回归精度最低，线性插值得到的数据集回归精度略高于其它三种插值方式。

表 1　多元线性回归结果

插值方法	常数项 (a)	前一天用电量系数(b)	节假日系数 (c)	R^2
阶梯插值	756.8	0.646	-336.7	0.607
线性插值	752.5	0.648	-336.3	0.683
二次B样条曲线插值	778.4	0.641	-359.1	0.682
三次B样条曲线插值	769.7	0.645	-351.7	0.680

选取最后 12 天的建筑电耗数据为测试集，检验四个模型的预测精度，检验结果如图 5 所示。四个模型的预测精度大致相近，其中线性插值得到的数据集回归的模型预测精度略高于其它三种方式。四个模型 12 天的预测结果的平均误差均在 7%左右，预测精度较高。



(a)实际和预测日耗电量对比　　(b)预测误差

图 5 模型精度检验结果

由以上计算和分析可知，多元线性回归是一种简单有效的建筑能耗预测模型的构建方法，既能快速建模，又能保证一定的精度，使模型满足预测要求。在进行数据清洗时，通过可视化的方式展现数据的分布规律并标记脏数据，可以准确把握数据质量，便于选取数

据质量更高的数据集或子数据集进行后期的清洗和模型构建。通过案例分析与对比，采用线性插值的方式填补数据集中缺失的数据，计算量较小，且能保证较高的精度。

4 结论

本研究提出了一个预测建筑能耗的完整流程，即首先根据可获得的数据确定建筑能耗的主要影响因素，然后标记并识别脏数据并进行数据清洗，最后构建预测模型并对比不同模型的精度，选择精度最高的模型预测建筑能耗。本研究对比了预测建筑能耗流程中可以采用的几种数据清洗方法以及模型构建方法，为预测建筑能耗时清洗数据以及构建预测模型提供了科学依据。工程实例验证结果表明，根据数据特点选择合理的数据清洗方式和模型构建方式，可以得到更好的预测效果。

建筑能耗预测一直都是研究的热点，但是基于预测结果对建筑节能工作的开展提出科学系统的建议却很少受到关注。之后的研究可以从优化能源价格、能源结构、建筑区域分布、建筑用电设备等角度出发，基于现有的建筑能耗预测方法及预测结果，探究推进建筑节能工作的可行方案，为国家和政府在制定相关政策和规划时提供科学依据。

参 考 文 献

[1]International Energy Agency.Tracking clean energy progress 2016[EB/OL].
http://www.iea.org/publications/freepublications/publication/TrackingCleanEnergyProgress2016.pdf,2016.

[2]侯静.我国城镇公共建筑能耗预测及能效提升路径研究[D].北京:北京交通大学,2017.

[3]刘海柱,丁洪涛,曾狄.2012 年民用建筑能耗统计数据分析报告[J].建设科技,2013,18:34-37.

[4] Jialin Wu, Zhiwei Lian, Zhuling Zheng, et al. A method to evaluate building energy consumption based on energy use index of different functional sectors[J]. Sustainable Cities and Society, 2020,53: 1-6.

[5]李璐,于军琪,杨益.基于 GA-BP 神经网络的大型公共建筑能耗预测研究[J].中外建筑,2014,3:112-114.

[6]Zhitong Ma,Cantao Ye,Weibin Ma.Support vector regression for predicting building energy consumption in southern China[J].Energy Procedia,2019,1:3433-3438.

[7]周璇, 崔少伟, 周裕东. 办公建筑逐时能耗异常数据在线插补方法[J]. 建筑科学, 2018, 34(6): 82-90.