

技术总结

DOI: 10. 19853/j. zgjsps. 1000-4602. 2026. 15. 007

基于K-means聚类算法的小区用户用水模式研究

王志伟¹, 邱顺添¹, 王玮婕¹, 张南¹, 魏忠庆², 何新宇²,
龚珑聪², 吴晓慧²

(1. 天津大学 环境科学与工程学院, 天津 300072; 2. 福州自来水有限公司, 福建 福州 350001)

摘要: 在供水管网建模过程中, 用户用水模式设置的准确性对管网模型模拟精度的影响极大。然而, 国内目前使用的用水模式难以反映不同用户和不同时间的实际用水情况。以居民生活用水为例, 提出了一种基于K-means聚类算法的用户用水模式提取方法, 选取福州市118个小区为研究对象, 对采集到的智能水表数据进行预处理, 得到连续的用水数据, 通过提取高峰时期用水特征指标对用水曲线进行降维处理, 采用K-means算法对降维后的曲线进行聚类, 并对聚类结果进行组成分析, 得到不同季节工作日和休息日的用水模式。对比当前常用的用水模式提取方法发现, 基于K-means方法得到的曲线能更准确地反映小区不同时间的用水变化特征, 为企业后续运行管理提供重要的数据支撑。

关键词: 用水模式; K-means聚类算法; 管网模型; 特征提取

中图分类号: TU991 **文献标识码:** A **文章编号:** 1000-4602(2026)15-0047-07

Water Consumption Pattern Analysis of Residential Communities Based on K-means Clustering Algorithm

Wang Zhiwei¹, Khu Soon Thiam¹, Wang Weijie¹, Zhang Nan¹, Wei Zhongqing²,
He Xinyu², Gong Longcong², Wu Xiaohui²

(1. School of Environmental Science and Engineering, Tianjin University, Tianjin 300072, China; 2. Fuzhou Water Supply Co. Ltd., Fuzhou 350001, China)

Abstract: In the process of water supply network modeling, the accuracy of user water consumption pattern settings has a significant impact on the simulation accuracy of the network model. However, the water consumption patterns currently used in China are unable to reflect the actual water usage of different users at different times. Taking residential water use as an example, this paper proposed a method for extracting user water consumption patterns based on the K-means clustering algorithm. A total of 118 communities in Fuzhou were selected as the research subjects. The collected smart water meter data were preprocessed to obtain continuous water consumption time series. Dimensionality reduction of the water consumption curves was performed by extracting characteristic indicators of water use during peak periods. The K-means algorithm was then applied to cluster the reduced-dimension curves, and a compositional analysis of the clustering results was conducted to derive water consumption

patterns for weekdays and weekends across different seasons. Compared with conventional water consumption pattern extraction methods, the curves obtained by the K-means approach can more accurately reflect the temporal characteristics of water use variations within communities, and can provide important data support for subsequent operation and management by water utilities.

Keywords: water consumption pattern; K-means clustering algorithm; pipe network model; feature extraction

近年来,随着人们生活水平的提高和气候变化的影响,水资源需求不断增加^[1]。为了解决城市水资源供需矛盾,达到合理的水资源调度与管理,国内很多自来水公司进行了供水管网水力模型建设^[2-3]。用户用水模式的参数是否精确,对管网模型模拟计算结果的精度影响极大,对供水管网系统的安全运行至关重要。

随着智能水表的普及,供水企业积累了大量用水数据,用以校核模型参数^[4]。然而,由于用户数量庞大,很难为每个用户建立其特征用水模式,且用水具有复杂性、时变性和分布性等特点^[5-6],因此准确的用水模式建模十分困难。目前常见的做法是根据用户的用水性质,得到不同类型用户的用水模式特征^[7]。但这种方法得到的用水模式忽略了同类用户的用水差异性,不能满足供水管网模型精细化模拟需求。工商业用水活动相对固定,用水模式清晰;而生活用水受人口构成、生活习惯、季节等影响^[8],随机性更强,为所有生活用水用户赋予一条用水模式曲线并不能准确模拟用户点的用水量。

通过用户聚类对用户进行细分,将具有相似用水模式的用户划分为一类,有助于解释不同用户用水模式的差异。程桂等^[7]采用层次聚类分析法得到了苏州某区七大主要行业的用水模式。Abu-Bakar等^[9]通过K-means算法对家庭用户平均日用水曲线进行聚类,量化了COVID-19期间对管网供水影响最大和最小的家庭用户集群。Ioannou等^[10]通过构建用户平均日用水量、用水量标准差等用水特征对高维用水数据进行降维,从而提升聚类质量。然而,这些研究大多基于平均日用水数据进行分析,与实际用水数据存在一定差异,并未考虑季节变化、日期类型(即工作日或休息日)等因素对用水模式的影响,因此无法在聚类特征中体现出不同用户用水在时间维度上的差异。

本研究利用K-means聚类算法对多个小区每

日真实的用水数据进行聚类分析,综合考虑季节变化、工作日和节假日对小区用户用水模式的影响。首先简要介绍基于K-means聚类算法分析的基本思路和框架,然后以福州市118个小区2021年日用水数据为研究样本,提取用水特征指标,实现对小区用户用水模式的聚类分析,并通过与福州市当前的用水模式获取方法进行对比,验证了本研究获取的用水模式的科学有效性。

1 用水模式研究方法

采用K-means方法对小区日用水曲线进行聚类分析,具体流程如图1所示,主要包括数据准备和数据聚类两部分。首先,对采集到的智能水表实测数据进行数据预处理和标准化,得到用水系数数据;然后,对用水系数数据进行特征提取降维;最后,对处理后的数据进行聚类分析。

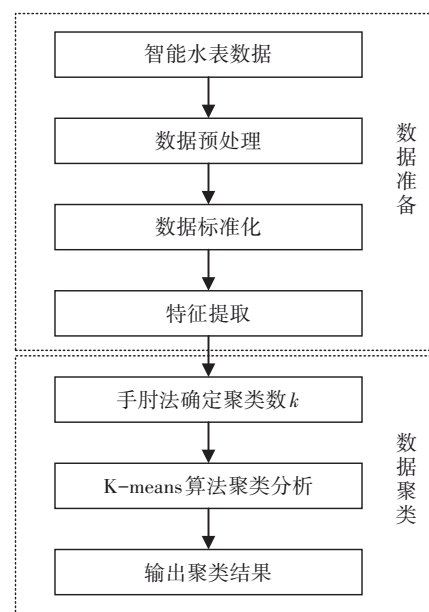


图1 基于K-means聚类的用水模式分析流程

Fig.1 Water consumption pattern analysis process based on K-means clustering

1.1 数据预处理

智能水表在用水数据采集过程中可能会出现测量装置异常,以及数据传输中断、管道检修等事件造成的用水数据异常。异常数据无法准确反映用水变化的规律,同时还会对数据分析的结果产生影响,因此需要进行数据预处理。异常数据包括畸变数据和缺失数据。畸变数据包括用水量突变数据、带毛刺的用水数据。缺失数据包括零值、空值和负值等。

畸变数据处理:对于用水曲线中骤升、骤降等畸变数据,可通过式(1)计算用水量变化率予以判断。对于得到的用水量变化率采用“ 3σ 原则”标记相应时刻的异常用水数据,对该位置数据做缺失数据处理。

$$\rho_t = \frac{|Q_t - Q_{t-1}|}{Q_t} \quad (1)$$

式中: Q_{t-1} 、 Q_t 分别为用户在 $t-1$ 、 t 时刻的用水量; ρ_t 为 t 时刻的用水量变化率。

缺失数据处理:对于用水数据缺失较少的情况采用统计学中的“前后相邻”填充方法,采用线性插值估计缺失数据;数据缺失过多的用户不宜选为样本,可删除用户数据。假设 t 时刻流量缺失,找到距离 t 时刻最近的流量值前后 t_A 和 t_B 时刻($t_A < t < t_B$),假设两个时刻的流量分别为 Q_A 和 Q_B ,则缺失数据 Q_t 按式(2)计算。

$$Q_t = Q_A + \frac{Q_B - Q_A}{t_B - t_A}(t - t_A) \quad (2)$$

1.2 数据均值化

小区用户的用水量差异较大,而用户用水模式聚类是为了分析不同用户的用水规律,即目的在于曲线的形状而非曲线的用水量值。因此可以通过数据标准化去除用水曲线水量特征带来的影响,以突出用水趋势特征,去除无效信息。为了方便对比不同用水量用户,采用用水曲线聚类结果更为准确。用水数据标准化通过均值化方法进行处理,计算公式如下:

$$q_t = \frac{Q_t}{Q_{\text{mean}}} \quad (3)$$

式中: Q_t 为用户 t 时刻的用水量; Q_{mean} 为用户该

日平均小时用水量; q_t 为标准化后用户 t 时刻的用水系数。

1.3 特征提取

高维的用水数据存在多噪声、信息冗余和存储空间巨大等缺点,直接用于聚类则计算效率较低。可以通过提取用水特征指标保留关键特征,减少重复信息,提高后续聚类算法的效率,同时也能降低噪声数据的影响。居民生活用水曲线通常较相似,呈现“双峰型”,即“早高峰”和“晚高峰”,存在较多冗余特征,进行聚类时容易被噪声数据影响。在实际生产应用中,用户用水的高峰更受关注,通过选取“早高峰”和“晚高峰”的用水系数以及峰值出现时间这4个用水特征指标作为特征向量,可以有效减少用水数据冗余特征,具体见表1。与原24维数据相比,用4个日用水特征指标来表征日用水数据后,数据量减少了83%,实现了对高维用水数据的降维。

表1 用水特征指标

Tab.1 Water consumption feature index

用水特征指标	定义
早高峰用水系数	06:00—12:00出现的最大用水系数
早高峰时间	早高峰用水系数对应的时间
晚高峰用水系数	18:00—24:00出现的最大用水系数
晚高峰时间	晚高峰用水系数对应的时间

1.4 K-means聚类算法

K-means聚类算法是一种典型的无监督学习算法,因其效率高、易于实现等特点而被广泛应用于时间序列聚类。K-means算法需要预先确定聚类数 k ,对于给定的样本集,根据样本间距离大小将样本集划分为 k 簇;以簇内样本距离尽量小、簇间样本距离尽可能大作为约束条件,以误差平方和(SSE)作为目标函数。本研究采用“手肘法”确定最佳聚类数 k 。当 $k <$ 最佳聚类数时,随着聚类数 k 的增加,SSE下降较快;而当 $k >$ 最佳聚类数时,随着聚类数 k 的增加,SSE下降幅度逐渐变缓。SSE和 k 值二者的关系图是手肘的形状,肘部对应的聚类数即为最佳聚类数。

K-means聚类算法通过欧氏距离计算数据点之间的距离,具体见式(4)。

$$d(q_i, q_j) = \sqrt{(q_{i1} - q_{j1})^2 + (q_{i2} - q_{j2})^2 + \cdots + (q_{in} - q_{jn})^2} \quad (4)$$

式中: $q_i = (q_{i1}, q_{i2}, \cdots, q_{in})$ 和 $q_j = (q_{j1}, q_{j2}, \cdots, q_{jn})$ 是两个 n 维样本数据对象。

最终划分依据是使聚类中心的误差平方和收敛,计算公式如下:

$$SSE = \sum_{i=1}^k \sum_{q \in M_i} \|q - o_i\|^2 \quad (5)$$

式中: q 为样本数据,假设有 k 个簇 $M(M_1, M_2, \dots, M_n)$ 和 k 个聚类中心 $o(o_1, o_2, \dots, o_n)$ 。

K-means 算法的具体步骤:①从数据集中选取 k 个样本作为每个簇的初始聚类中心;②计算所有样本数据与每个聚类中心的距离,将其分到距离最小的簇中;③计算每个簇中样本数据的平均值代替原有聚类中心点;④使用SSE对聚类效果进行评估。重复步骤②、③,直到每个聚类中心在某个精度范围内不变化或者达到最大迭代次数,迭代结束。

2 案例分析

2.1 样本数据及数据预处理

研究数据来自福州市自来水公司SCADA系统数据库,从中选取福州市2021年平均日用水量排名前300的小区监控水表日用水数据,收集2021年1月1日—12月31日的用水数据,时间间隔为1 h。实际用水数据可能因测量设备故障或数据传输异常等情况,出现异常值或缺失值等无效数据。为保证结果的有效性,首先对用水数据进行了清洗:对于总体数据缺失量超过20%的小区用水数据,剔除该小区的全部数据;通过式(1)计算用水量变化率,判断是否为异常值,并剔除异常值;对于单日用水数据中缺失数据少于5个点(缺失率小于20%)的情况,使用插值法进行填补;对于单日用水数据中缺失数据超过5个点(缺失率大于20%)的情况,剔除该日数据。尽量保证用原始真实数据进行分析。最终,因筛选条件相对严格,仅118个小区用水数据满足后续分析要求,其中随机选取106个小区用于聚类分析,12个小区用于模拟新用户验证用水模式,占总数据量的10.2%。为了分析不同小区用户用水规律的相似性,通过数据标准化方法,得到小区用水变化系数数据。之后通过1.3节提出的方法进行特征提取。

2.2 K-means 聚类算法分析

根据用水特性指标对小区日用水曲线进行聚类。采用“手肘法”确定最佳聚类数,计算聚类数从1增至10时对应的数据SSE结果,如图2所示。可以看出,当聚类数由1增至4时,SSE快速下降;当聚

类数超过4时,SSE的下降幅度趋于缓慢。这表明当聚类数为4时,聚类效果最好。因此样本的最佳聚类数为4。

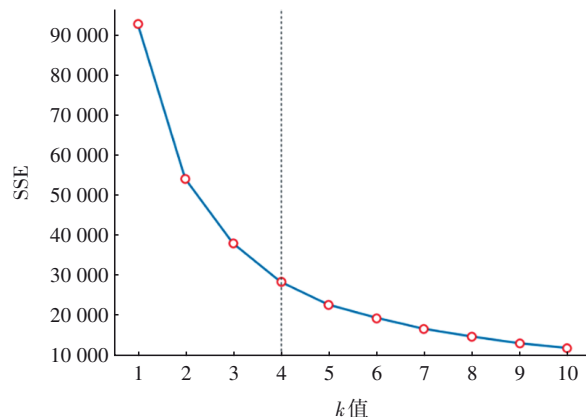
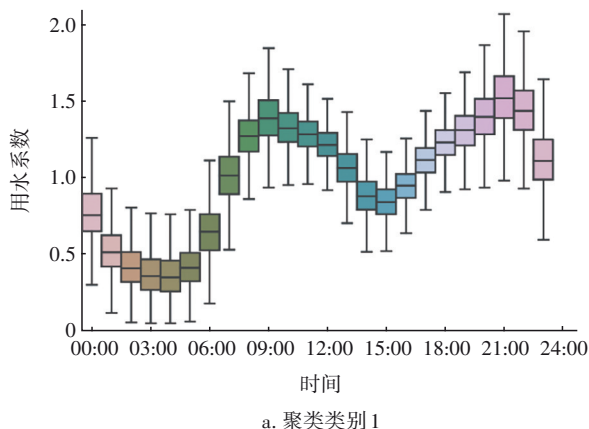


图2 SSE确定的最佳聚类数

Fig.2 Optimal cluster count based on SSE

通过K-means算法进行聚类分析,聚类类别1、2、3、4的频数分别为8 704、6 079、17 826、3 403,占比分别为24.2%、16.9%、49.5%、9.4%。可以看出,聚类类别3的占比最高,聚类类别4占比最低。降维后的用水曲线更为规律清晰,降维前聚类时间为1.74 s,而降维后的聚类时间为0.41 s,不到降维前的1/4。通过提取用水特征指标对用水数据降维,能够很好地提升聚类效率。

聚类类别1、2、3、4的结果如图3所示。由于是生活用水,4类曲线的形状均为双峰型,包括“早高峰”和“晚高峰”。从聚类结果可以看出,小区用水模式各类型的“早高峰”时间为07:00—10:00,“晚高峰”时间为19:00—22:00,这与福州地区居民的生活习惯相符。各类用水模式的主要区别在于用水峰值系数、早晚高峰峰值时间和峰谷的幅度。



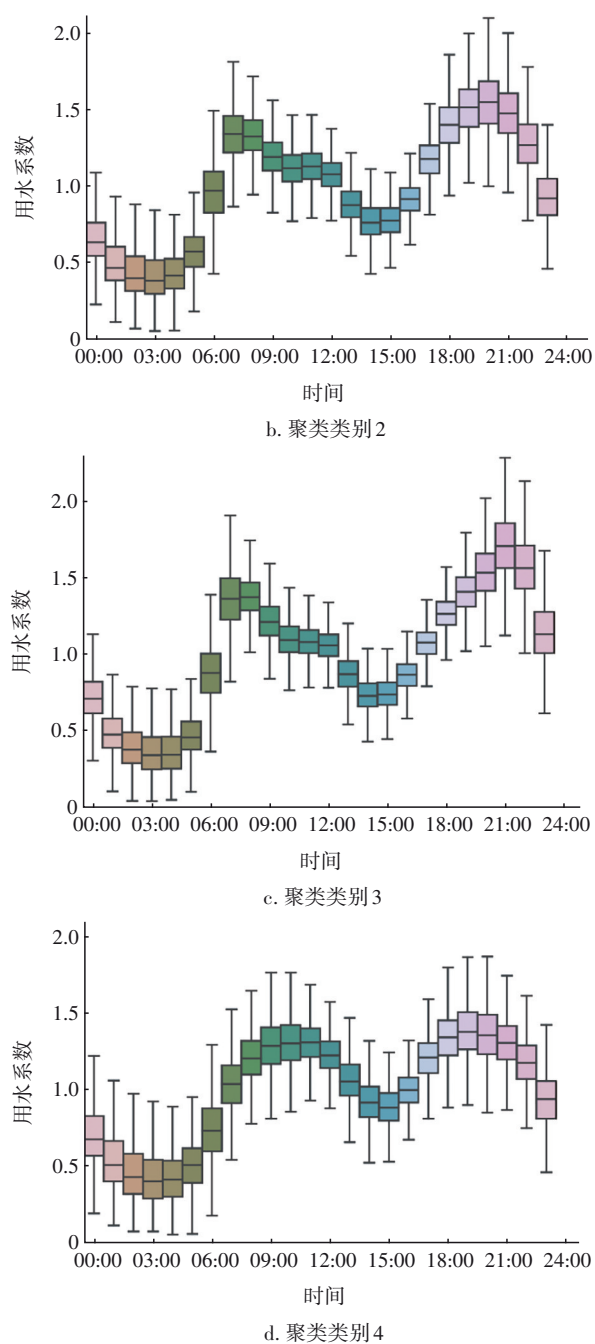


图3 聚类结果

Fig.3 Clustering results

2.3 聚类结果分析

2.3.1 聚类特征

在本研究中,使用实际特征(如用水性质、季节、是否为工作日等)作为样本标签,对聚类结果进行解释分析。研究对象是住宅小区,用水性质一致,均为居民生活用水,故更为关注时间上的差异。一方面,生活用水在不同季节往往表现出不同的规律;另一方面,日期类型对居民用水时间也有很大

影响。通过分析各分类中所包含的主要特征,寻找用户的用水规律,发现其共同点,并对各类别曲线的实际意义进行分析。

2.3.2 用水模式分析

在聚类类别1中,用水量从05:00开始快速增加,09:00—10:00为“早高峰”,高峰系数为1.40,之后用水量逐渐下降,15:00到达用水低谷,15:00—21:00用水量逐渐增加,21:00—22:00到达全天用水最高峰,即“晚高峰”,用水系数为1.53,之后用水量逐渐减少,01:00—05:00为全天用水低谷。这类曲线占总样本的24.2%。该类别在春、夏、秋、冬的占比分别为26%、26%、24%、24%,四季占比相近。其中工作日占比为25%,休息日占比为75%(见图4),因此主要为休息日用水,用水“早高峰”出现在09:00之后也符合休息日的用水特点。故类别1可归类为春、夏、秋、冬四季休息日的用水模式。

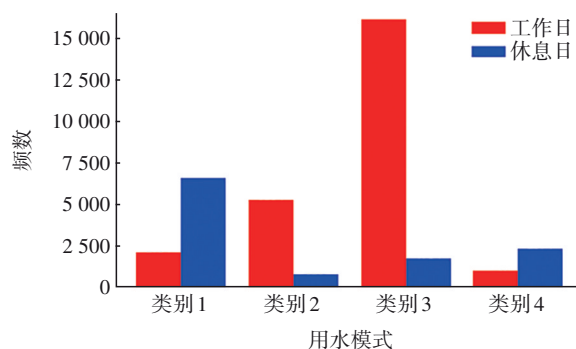


图4 不同类别中工作日和休息日的数量

Fig.4 Number of weekdays and weekends in different clusters

在聚类类别2中,用水“早高峰”出现在07:00—08:00,用水系数为1.35,“晚高峰”出现在20:00—21:00,用水系数为1.56。这类用水曲线占比为16.9%。工作日用水曲线占比为86%,因此主要为工作日用水,用水“早高峰”出现在07:00也符合工作日的用水特点。在春、夏、秋、冬的占比分别为22%、21%、24%、33%,均在20%以上,占比差异不大,故类别2可归类为四季工作日的用水模式。

在聚类类别3中,用水“早高峰”出现在08:00—09:00,用水系数为1.39,“晚高峰”为21:00—22:00,用水系数为1.72。这类用水曲线占比为49.5%。工作日用水曲线占比为90%。在季节特征上,这类曲线中冬季占比最小,春、夏、秋季占比相近,分别为27%、29%、29%。故类别3可归类为春、夏、秋季

工作日用水模式。

在聚类类别4中,用水“早高峰”出现在10:00—11:00,用水系数为1.33,“晚高峰”出现在19:00—20:00,用水系数为1.41,全天用水情况相较其他三类较为平稳,用水波动较小。这类用水曲线占比为9.4%。该类别中休息日占比为70%。季节特征上,冬季占比最大(52%),其他三季占比较小(春、夏、秋占比分别为18%、16%、14%),故聚类类别4可归为冬季休息日用水模式。

尽管所有用水曲线均为生活用水曲线,但在不同的用水日,其用水规律具有明显差异。在本研究中,福州市生活用水用户春、夏、秋三季的用水规律较为相似,而冬季的用水规律则有所不同。相比其他三季,冬季用水更为平缓,尤其是在冬季的休息日,用水峰值较低。这可能是因为冬季气温较低,居民出行减少,有更多的时间待在室内,用水时间分散,用水分布更均匀,用水峰值较低。同时还发现,工作日的“早高峰”用水时间基本在07:00—09:00;而在休息日,由于需要早起的居民减少,用水“早高峰”推迟到09:00—11:00,且用水高峰系数存在一定差异。在对小区用水建模时,不同条件下推荐的用水模式见表2。

表2 不同条件下用水模式建议

Tab.2 Recommended water consumption patterns under different conditions

项目	用水模式
春、夏、秋季	工作日
	休息日
冬季	工作日
	休息日

2.3.3 曲线描述能力分析

为验证所得用水模式的准确性,与福州市当前用水模式获取方法进行比较。目前,福州市利用历史用水数据进行统计,以平均日用水曲线作为用水模式。用预留的12个小区进行验证分析,以这些小区全年用水实测数据进行用水模拟。A、B、C、D、E、F、G、H、I、J、K、L小区的日均用水量分别为230、420、460、470、500、590、660、680、770、950、1 100、1 800 m³/d,时均用水量分别为10、18、19、20、21、25、28、28、32、40、46、75 m³/h。

用户 t 时刻用水量按式(6)计算。

$$Q_t = \frac{Q_{\text{总}}}{24} \times q_t \quad (6)$$

式中: Q_t 为模拟用户 t 时刻用水量; $Q_{\text{总}}$ 为用户该日总用水量; q_t 为对应用水模式 t 时刻用水系数。

方法一:根据上述K-means聚类获得的4类典型用水模式,依据4条用水模式曲线,通过日总用水量和用水模式进行每日用水量曲线模拟,进而得到每个小区全年的用水曲线;方法二:采用全部小区平均日用水模式曲线作为用水模式,通过日总用水量和平均日用水模式进行每日用水量曲线模拟,进而得到每个小区全年的用水曲线。通过两种方式所得用水曲线分别与实测用水曲线进行比较,以均方误差(MSE)、平均绝对误差(MAE)和平均绝对百分比误差(MAPE)定量评估模拟效果。上述3个评价指标越接近0,表明模拟值与实测值越接近,预测效果越好。12个小区的用水模拟结果见表3。可以看出,方法一的MSE更小,说明通过所得用水模式模拟用水时结果更稳定,误差波动更小;相比方法二,方法一的MAPE下降1.0%~2.3%,说明方法一的预测结果更准确。可见,总体上K-means聚类后的用水模式能有效减小预测误差,模拟精度更高,更贴近真实用水情况。

表3 12个小区的用水模拟结果

Tab.3 Water consumption simulation results for 12 residential communities

小区	MSE/%		MAE/m ³		MAPE/%	
	方法一	方法二	方法一	方法二	方法一	方法二
A	3.1	3.7	2.3	2.8	12.7	14.6
B	3.2	3.4	2.3	2.5	18.5	20.8
C	3.3	3.7	2.6	3.0	15.0	17.3
D	2.8	3.0	2.1	2.4	12.4	14.4
E	4.1	4.2	2.7	2.9	14.4	15.7
F	3.2	3.5	2.4	2.7	12.4	13.5
G	4.7	5.0	3.4	3.7	13.9	15.6
H	4.2	4.6	2.7	3.1	11.5	13.5
I	4.0	4.4	2.8	3.1	10.2	11.3
J	6.8	7.3	5.4	5.8	15.5	16.8
K	6.5	6.9	5.0	5.3	12.9	14.4
L	11.1	12.4	7.7	9.2	10.9	13.1

分别在春、夏、秋季以及冬季随机选取一个工作日和一个休息日数据绘制模拟曲线,具体如图5所示。可以看出,两种方法均能在一定程度上反映用水趋势,但聚类后得到的用水模式能更好地模拟

小区用户用水量的真实值,模拟误差更小。尤其是在早晚用水高峰,采用K-means聚类所得的用水模式能更好地贴合实测用水曲线,误差更小。

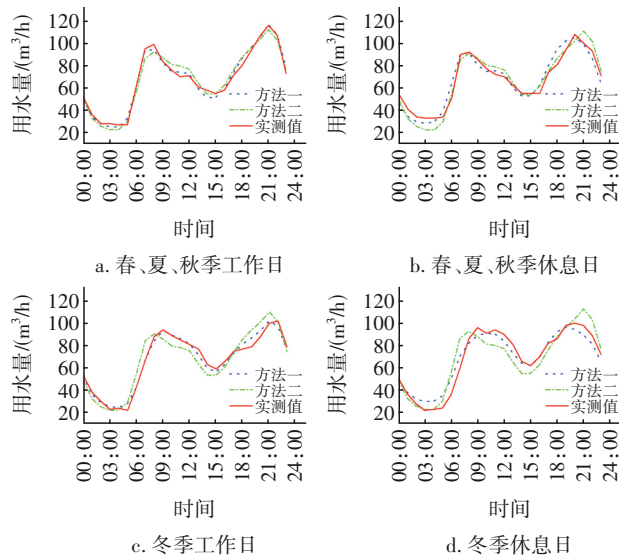


图5 用水量模拟曲线

Fig.5 Simulation curve of water consumption

3 结论

基于K-means聚类的用水模式提取方法,对居民生活用水模式进行了聚类分析,结果表明,居民生活用水模式随季节变化较为明显,冬季的用水规律与其他三个季节差异较大。在工作日和休息日,居民生活用水“早高峰”出现时间相差较大,用水高峰系数也存在差异。聚类的方法能有效提取小区用户的用水模式,所得用水模式相比平均日用水模式能更精确地模拟小区实际用水变化,可为企业后续运行管理提供重要的数据支撑。

参考文献:

- [1] 程小文. 21世纪我国城市用水变化及趋势分析[J]. 中国给水排水, 2022, 38(14): 47-51.
Cheng Xiaowen. Analysis on the change and trend of urban water consumption in China in the 21st century [J]. China Water & Wastewater, 2022, 38(14): 47-51 (in Chinese).
- [2] 李冠民. 天津水司的水力模型建设与应用[J]. 中国给水排水, 2015, 31(3): 72-74.
Li Guanmin. Construction and application of hydraulic model for water supply in Tianjin [J]. China Water & Wastewater, 2015, 31(3): 72-74 (in Chinese).

- [3] 高泽晨. 基于水力模型的城市供水主干管网联通工程研究[J]. 中国给水排水, 2023, 39(11): 35-41.
Gao Zechen. Drinking water distribution main network interconnection project based on hydraulic model [J]. China Water & Wastewater, 2023, 39(11): 35-41 (in Chinese).
- [4] Du K, Long T Y, Wang J H, et al. Inversion model of water distribution systems for nodal demand calibration [J]. Journal of Water Resources Planning and Management, 2015, 141(9): 04015002.
- [5] Costa D F, Soares A K. Using smart meters data to model customer water consumption and evaluate customer flow patterns [J]. Urban Water Journal, 2023, 20(1): 89-99.
- [6] Abu-Bakar H H, Williams L, Hallett S H. Contextualising household water consumption patterns in England: a socio-economic and socio-demographic narrative [J]. Cleaner and Responsible Consumption, 2023, 8: 100104.
- [7] 程桂, 李翠梅, 郜阔. 城市行业用水量变化与用水模式实例研究[J]. 深圳大学学报(理工版), 2016, 33(1): 49-54.
Cheng Gui, Li Cuimei, Gao Kuo. A case study of water consumption change and water-use pattern for city industries [J]. Journal of Shenzhen University (Science and Engineering), 2016, 33(1): 49-54 (in Chinese).
- [8] Abu-Bakar H H, Williams L, Hallett S H. A review of household water demand management and consumption measurement [J]. Journal of Cleaner Production, 2021, 292: 125872.
- [9] Abu-Bakar H H, Williams L, Hallett S H. Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England [J]. npj Clean Water, 2021, 4(1): 13.
- [10] Ioannou A E, Creaco E F, Lapidou C S. Exploring the effectiveness of clustering algorithms for capturing water consumption behavior at household level [J]. Sustainability, 2021, 13(5): 2603.

作者简介:王志伟(1998—),男,江西上饶人,大学本科,主要研究方向为供水管网优化。

E-mail:tjuwzw@126.com

收稿日期:2025-02-21

修回日期:2025-05-10

(编辑:任莹莹)