Statistics and Application 统计学与应用, 2022, 11(6), 1500-1504
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.116156

Hans汉斯

基于随机森林的居民用水量预测

赵　娜

广西师范大学数学与统计学院，广西　桂林

收稿日期：2022年11月27日；录用日期：2022年12月17日；发布日期：2022年12月29日

摘　要

水资源管理中至关重要的工作之一就是用水量预测工作，建立模型对未来一定时段内居民用水量进行合理预测，对于水资源规划、城市供水管网建设、供水调度优化等都具有重要指导意义。选取气象和疫情相关变量，构建随机森林模型对深圳市某小区日用水量进行预测，并结合每日用水规律对小时用水量进行预测，对本地城市居民用水量预测工作具有一定参考价值。

关键词

居民用水量，随机森林，时间序列

Prediction of Household Daily Water Consumption Based on Random Forest

Na Zhao

School of Mathematics and Statistics, Guangxi Normal University, Guilin Guangxi

Received: Nov. 27th, 2022; accepted: Dec. 17th, 2022; published: Dec. 29th, 2022

Abstract

One of the most important tasks in water resources management is water consumption prediction. The establishment of a model to reasonably predict the water consumption of residents in a certain period of time in the future has important guiding significance for water resources planning, urban water supply network construction and optimization of water supply scheduling. The random forest model was built to predict the daily water consumption of a residential district in Shenzhen by selecting the variables related to weather and epidemic, and the hourly water consumption was predicted by combining the daily water consumption rule. At the same time, it had certain reference value for the local urban residential water consumption prediction.

文章引用: 赵娜. 基于随机森林的居民用水量预测[J]. 统计学与应用, 2022, 11(6): 1500-1504.
DOI: 10.12677/sa.2022.116156

Keywords

Residential Water Consumption, Random Forest, Time Series

1. 引言

水资源是人类赖以生存的必要资源之一，当今时代，水资源短缺问题日益严重，中国人均水资源占有量仅为世界平均水平的四分之一；同时，随着城市的不断发展，供水管网也变得日益庞大而复杂，因此，对水资源的合理规划和对供水系统的优化调度变得越来越重要，其中最为基础的任务就是用水量规律分析及预测。这一任务是供水基础设施建设、供水调度规划的前提和主要依据，对于水务行业的升级和高质量发展具有重要的现实意义。

掌握用水量规律并建立合适的模型对其进行预测，可以更加精确的把控供水管网每个节点的用水需求，有助于在保证水压、水量的前提下降低供水成本减少用水行为波动对管网水质的影响，提升百姓幸福感的同时推进智慧水务建设，实现供水科学调度和低碳高效精细化管理。

已有学者运用神经网络[1]、灰色预测[2]、组合模型[3]、实证分析[4]、数据挖掘[5]等方法进行城市用水量进行影响因素分析和预测，可以发现影响居民用水量的因素主要分为人为因素和环境因素，其中人为因素相关数据不易获得，仅适合进行结构分析和趋势分析，难以用于预测居民用水量，因此本文考虑根据环境因素分析居民用水行为并进行预测。这些文献在模型的解释性、预测准确度和实时预测方面无法兼顾，而且多以城市或地区为单位进行整体分析，预测行为也以年为单位进行。对于居民小区日用水量分析和预测的相关文献相对较少，王锦涛等学者[6]探究了气候因素对居民小区日用水量的影响，并结合数据挖掘方法进行预测；但是近年来新冠肺炎疫情起起伏伏，对居民生活包括居民用水行为也产生了较大的影响[7]，周骅[8]对新疫情封控期间特大型城市用水量进行影响因素分析，发现城市供水量受疫情封控影响较大，但尚未有研究考虑疫情对居民用水行为的影响。

本文主要是通过智能水表运行历史数据分析城市居民小区用水量规律，并结合气象数据进行回归、时序建模，建立区域居民小区需水预测模型，利用历史用水数据和实时感知数据预测未来一定周期内小区用水量，以指导实际供水运行工作。该模型结合供水设施水泵和水池调蓄能力，在保障供水水质水压稳定的前提下，可实现区域蓄水平峰调控优化和泵组优化辅助决策，保障管网供水输送的水质稳定和供水设施的节能高效运行。

2. 数据描述

本文使用的数据来自 Data Fountain (简称 DF 平台)，该数据集包括多个小区数据，本文只选取其中一个小区进行研究，其他小区可参考本文方法进行分析建模，取得较好的预测效果。数据集中包括日用水量数据和小时用水量数据，另外还有气象变量和疫情相关变量的日数据。本文研究目的是预测该小区日用水量、分析每日小时用水量的规律，并据此预测小时用水需求量。

在分析和建模之前，首先对数据进行适当的处理。本文选择 2020 年 2 月 1 日至 2022 年 6 月 30 日的数据作为训练集进行分析建模，以 2022 年 7 月和 8 月的数据作为测试集来检验模型的预测效果。训练集

中共有 881 行数据，其中 18 行包含缺失值，由于缺失比例较低，对含删失值的行直接进行删除。另外对其中一些自变量进行适当的变换，具体见表 1。除数据集原始变量外，考虑用水量是时间序列数据，将前一日用水量作为一个自变量来反映时间序列趋势。同时考虑时间序列的季节效应加入年份、月份两个变量，另外结合实际生活场景加入假期与否、周内日期序数变量。

Table 1. Table of original independent variables
表 1. 原始自变量表

变量名		变量解释	处理
气象变量	Fx	风向(度)	换算为八风向，从北到西北分别记为 1 至 8
	Fs	风速(0.1 米每秒)	换算为风力等级
	U	相对湿度(百分比)	以十为单位划分区间
	R	降水(0.1 毫米)	降水取值范围较小，转化为有雨和无雨两种
	T	气温(0.1 摄氏度)	
	V	能见度(米)	
	P	气压(百帕)	
疫情变量	Zz	当前重症人数	缺失值以 0 填补
	Wz	当前危重人数	缺失值以 0 填补
	xzqz	新增确诊人数	
	xzcy	新增出院人数	
	xzsw	新增死亡人数	
	glzl	当前隔离治疗人数	
	yxgc	当前医学观察人数	

3. 理论概述

随机森林模型由多棵分类决策树模型组合而成，各个树模型独立运作，最终进行汇总得出结果。因此，随机森林方法显著提高了预测精度，是当前最好的算法之一。

具体而言，随机森林通过自助采样技术学习训练多个决策树并进行汇总预测。对于回归问题，从原始训练集中有放回地随机抽取 k 个样本，对 k 个样本分别进行训练生成 k 个决策树模型，最后依据简单平均法将 k 个决策树的结果进行组合形成结果[9]。

随机森林模型中有两个至关重要的参数(mtry 和 ntree)会影响模型的准确性，其中，mtry 是在决策树生成过程中内部节点分裂对应的变量个数，该数值在随机森林模型形成过程中始终不变，一般选择使得均方误差最小的值；ntree 是随机森林中决策树的数量，虽然决策树越多结果越准确，但决策树过多会增加计算量，所以通常的做法是根据模型内误差变化曲线选择误差稳定时对应的值。

本文首先利用随机森林方法建立模型预测小区日用水量，然后根据日用水量的小时变化规律计算小时用水量的预测值。预测效果的评价标准选用均方根误差(*RMSE*)和平均绝对百分误差(*MAPE*)，其中，

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2},\ MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i-\hat{y}_i\right|}{y_i}.$$

4. 模型分析

对训练集数据建立随机森林模型，共有 20 个自变量，最小均方误差对应的 mtry 为 16，模型内误差

稳定时对应的 ntree 为 200，构造相应的随机森林模型，其方差解释率为 61.37%，即 20 个解释变量对响应变量(日用水量)有关方差的整体解释率为 61.37%，尚可接受。

自变量重要性得分图(见图 1)中，“%IncMSE”即 increase in mean squared error，通过对每个预测变量随机赋值计算模型误差变化，若该变量重要，那么其值被随机替换后模型预测的误差会增大，反映了各解释变量对模型预测效果的影响。“IncNodePurity”即 increase in node purity，以残差平方和度量了每个变量对决策树每个节点上观测值的异质性的影响，反映了各解释变量对模型拟合效果的影响。从拟合的角度来看(IncNodePurity)，对响应变量影响最大的解释变量为前一日用水量、当前重症人数、当前危重人数、月份、年份；从预测的角度来看(%IncMSE)，对预测效果影响最大的解释变量为前一日用水量、气温、气压、月份、当前隔离治疗人数；整体来看，对日用水量影响较大的解释变量为前一日用水量、月份、当前重症人数、当前危重人数、气温、年份、当前隔离治疗人数，而新增死亡人数、风速、风向、新增确诊人数对日用水量的影响都相对较小。

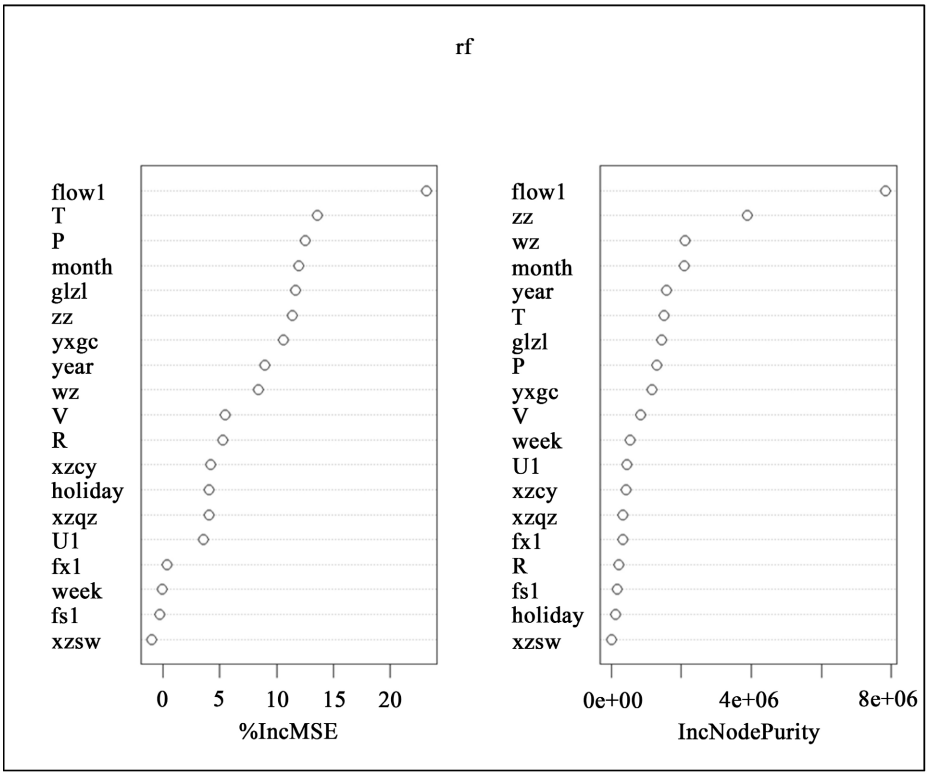

Figure 1. Score chart of the importance of independent variables
图 1. 自变量重要性得分图

接下来用构造出的随机森林对测试集进行预测，结果显示均方根误差为 81.99，平均绝对百分误差为 4.66%，预测效果较好。

分析训练集对应的小时用水量数据发现，每日小时数据变化趋势基本一致，每天晚高峰都出现在 22 时，工作日全天早高峰出现在早上 8 时，周末全天早高峰出现在 10 时，每天各小时用水量占全天用水量的比例基本不变。因此我们用训练集数据计算每天各小时用水量占比，并用随机森林模型得到的测试集日用水量预测值乘以该比例计算测试集中对应的所有小时用水量预测值，最后得出均方根误差和平均绝对百分误差分别为 9.2%和 15%，较日用水量预测效果有所下降，但尚可接受。

5. 结论

本文模型中将时间序列分析和多元回归结合起来建立的随机森林模型在预测居民日用水量时具有较高的准确度，且有很好的解释性。在此基础上对小时用水量预测的误差也尚可接受。通过模型可以发现，居民日用水量与前一日用水量相关性较强，具有较强的时间序列趋势，气象因素对深圳居民用水量影响较小，相对而言，疫情相关因素影响较大，这种情况可能是由于深圳市气象因素较为稳定导致。因此，对于深圳市供水系统，在气象发生一定变化时，供水量无需做出较大变化，但疫情形势发生明显变化时，建议及时调整供水量。

参考文献

[1] 李红冲, 陈勋俊, 刘国强, 等. 城镇居民生活用水量预测及用水需求量影响因素分析[J]. 水电能源科学, 2022, 40(9): 52-55.

[2] 吴永强, 李明凯, 唐中楠, 等. 基于灰色动态模型群的衡水市居民年用水量预测[J]. 环境工程技术学报, 2022, 12(1): 267-274.

[3] 王梓涵, 于忠清. 基于 TCK-LSTM-ATT 模型的城市用水量预测[J]. 青岛大学学报(自然科学版), 2022, 35(1): 53-59.

[4] 赵卫华. 居民家庭用水量影响因素的实证分析——基于北京市居民用水行为的调查数据考察[J]. 干旱区资源与环境, 2015, 29(4): 137-142. https://doi.org/10.13448/j.cnki. jalre.2015.130

[5] 李贺. 基于自组织数据挖掘方法的区域用水量预警分析[D]: [硕士学位论文]. 北京: 华北电力大学(北京), 2016.

[6] 王锦涛, 吴永强, 王书盛, 等. 基于数据挖掘的小区用水量影响因素研究[J]. 河北建筑工程学院学报, 2022, 40(2): 131-136.

[7] 李豪, 赵和松, 赵红莉, 等. 新冠疫情对区域用水量及社会用水行为的影响研究综述[J/OL]. 水利水电技术(中英文): 1-9. http://kns.cnki.net/kcms/detail/10.1746.TV.20220629.1630.002.html, 2022-11-26.

[8] 周骅. 新冠疫情封控管理期间特大型城市用水量影响分析[J]. 净水技术, 2022, 41(9): 137-142+149. https://doi.org/10.15890/j.cnki.jsjs.2022.09.019

[9] 刘超. 回归分析——方法、数据与 R 的应用[M]. 北京: 高等教育出版社, 2019.