Research Paper

A Monthly Water Consumption Prediction Model for Municipal Consumers

Sajjad Zarifzadeh[1]*, Fatemeh Kaveh-Yazdy[2]

1. Assistant Prof. of Computer Engineering, Computer Engineering Dept., Yazd University, Yazd, Iran
2. Senior Data Scientist, Data Science Institute, Yazd University, Yazd, Iran

Received:2020/06/10
Revised:2020/06/26
Accepted:2021/03/08
Online: 2021/09/01

Use your device to scan and read the article online

DOI:
10.30495/wej.2021.25230.2251

Keywords:
Consumption Prediction, Regression, Random Forest, Anomaly, Quantile Regression.

Abstract

Introduction: Smart water resource management is the best short-time solution for water resource shortages around the world. Predicting water demand is the primary prerequisite to being aware of the required water within a short time. Several types of features ranging from consumption history to meteorological have been used in water consumption prediction studies. In this article, we aimed at introducing a forecast model which predicts the monthly water consumption of urban consumers in Yazd city.

Methods: The proposed prediction framework uses the billing records of water consumers in Yazd city to extract consumption history. In addition, external data resources such as business calendar data, urban water production, meteorological parameters, the financial value of buildings, and in-stream pressure are collected and employed in the prediction model. This framework tracks the changes in consumption behaviors of consumers, which are grouped according to their volume of water usage to remove consumers with anomalous consumption behaviors. The cleaned grouped records of consumption are utilized in the fitting of a quantile regressor with three breakpoints to forecast the water demand of the consumers for the next month.

Findings: The results of the experiments showed that the proposed model's prediction percentage error is less than 10%. Besides, the model can recognize consumers with anomalous consumption behaviors.

Citation: Zarifzadehm S, Kaveh-Yazdy F. A Monthly Water Consumption Prediction Model for Municipal Consumers. Water Resources Engineering Journal.2022; 15(52): 94-112

***Corresponding author:** Sajjad Zarifzadeh, Ph.D.
Address: Computer Engineering Department, Yazd University, Yazd, Iran
Tell: +98 353 820 0145
Email: szarifzadeh@yazd.ac.ir

Extended Abstract

Introduction

Water shortage is becoming a major concern affecting the lives of millions of people around the world. Drought cycles directly threaten the drinking water resources, agriculture, and economy of numerous countries and implicitly target peace in various regions on the earth. Predicting the amount of drinking water, providing a perspective for future consumption, has been studied as a prerequisite for active water management. Consumption prediction studies focus on their domain and can be grouped into micro-scale and macro-scale forecasts. Macro-scale studies focus on the prediction of consumption of water on an urban scale, and micro-scale projects forecast consumer-scale consumption. Regardless of their scale, these studies have faced entirely different challenges. While macro-scale forecasts are more sensitive to numerical errors, micro-scale forecasts suffer from the massive amount of data and the curse of dimensionality. Urban water consumption prediction studies with respect to the data type they used, are categorized into four groups.

(1) prediction methods using consumption history

(2) prediction methods using consumption history and meteorological data

(3) prediction methods using consumption history, socio-economic and demographic

(4) prediction methods using a mixture of the above-mentioned data types, as well as a working calendar

Among the mentioned groups of methods, the second group is mainly used in water consumption prediction research studies. We should note that while demographic and socio-economic parameters directly affect the consumption value and can increase the accuracy of prediction methods, they have not been taken into account because collecting them is expensive and time-consuming. Furthermore, empirical projects involving socio-economic data gathering raise privacy concerns. In this article, we aimed to predict the monthly water consumed by urban consumers in the Yazd city using consumption history and meteorological data, as well as the working calendar.

Materials and Methods

In this research, we collect lean a repository of various data, including,

(1) Working calendar in addition to the date of unexpected closure

(2) Meteorological data (including 17 different variables)

(3) Consumption history of Yazd urban consumers

(4) Feed-in water provided by the water plant and its pressure

All types of data are cleaned to avoid inconsistencies, and null values are removed or imputed with appropriate values with respect to the type of data[1]. The next step includes data normalization, transformation, and feature selection. In this way, the type of a day as working days, holidays, and unexpected closures are converted to one-hot vectors. Meteorological variables are transformed and aggregated. Aggregated weather parameters are analyzed using the LassoCV method to be weighted. The higher the weight of a parameter, the more capable it is to forecast the future consumption value.

History of consumption in the six previous months and billing factors such as wastewater usage, water pressure, feed-in diameter, and type of meter (21 factors in total) are analyzed using a random forest regressor to estimate the impact of features. Accordingly, nine features with maximum impact on future consumption are selected to be utilized in the forecast model.

In the final step, eighteen methods such as OLS Regressor, Bayesian Regressor, MLP Neural Regressor, LSTM Kernel Regressor, ELM Regressor, SGD Regressor, and Quantile Regressor are run on selected features to predict the consumption value a month ahead. During our investigation, we found

[1] details of data cleaning methods are denoted in the article

that applied methods are suppressed by their limitations in problem space. To overcome this challenge, we propose a combination of an automatic segmentation method with quantile regression to achieve more accurate forecasts. The segmentation method finds the quantiles, including the data points' major proportion. Then, outliers (data points has consumption values less than the lowest quantile or higher than the highest quantiles) are removed from the data. In the last step, for every segment, the quantile regressor is run to estimate the future consumption value.

Findings

Among the tested methods, the random forest regressor, Huber regressor, and the quantile regressor show the best estimation performance (eq. lowest error rate). The common feature of all these methods is their robustness to outlier values. Our deep investigations show that we tune the random forest regressor and Huber regressor with the best values to achieve higher performance without being stuck in local optima or over-fitted. We found that the last best method that is the quantile regressor, has the ability to be optimized or manipulated to achieve higher accuracy. Thus, we combine this method with an automatic segment detection method to determine the segments in which the estimation error of the quantile regressor is minimized. This method is called piecewise quantile regression. Our finding show that boundary quantiles are not symmetrically placed, and the amount of data points that are removed from the training step ranges from 12% to 17% of the total amount of data in different segments. Our results show that the proposed method beats the random forest regressor and the Huber regressor. Furthermore, the proposed method is benefitted from the leverage of removing outliers.

Discussion

We track the error of the proposed method to find the root cause. Our investigation reveals that two main patterns are behind the errors, that are, (1) high consumption registered in the target month in contrast to very low consumption in the last three months, and (2) low consumption in comparison with a high consumption three-month history. Both patterns might be happened due to relocation between rented houses. In addition, we found that after three months, the prediction method learns the consumption pattern of the new residents and forecasts the consumption in the next month accurately. One of the most important findings of this research is the detected boundaries of consumer segments. Results confirm that urban consumers in Yazd city can be grouped into four segments with respect to their consumption history.

Conclusion

In this article, we review the methods and approaches that are utilized in forecasting micro-scale urban water consumption. We implement eighteen methods from twelve groups of methods and investigate the top three methods with minimum error. During our investigation, we discovered relocation of residents to rented houses caused the prediction errors, which came from inconsistency between the history of consumption before and after the relocation of residents. By taking into account this impact and removing the outliers, our method, which is constructed from a segmentation method and quantile regressor, achieves higher accuracy.

Ethical Considerations compliance with ethical guidelines

The cooperation of the participants in the present study was voluntary and accompanied by their consent.

Funding

No funding.

Authors' contributions

Design and conceptualization: Fatemeh Kaveh-Yazdy, Sajjad Zarifzadeh.
Methodology and data analysis: Fatemeh Kaveh-Yazdy.
Supervision: Sajjad Zarifzadeh
Writing: Fatemeh Kaveh-Yazdy, Sajjad Zarifzadeh

Conflicts of interest

The authors declared no conflict of interest.

مقاله پژوهشی

ارائه مدلی برای پیش‌بینی میزان مصرف آب ماهانه برای مشترکین خانگی

سجاد ظریف‌زاده[1*]، فاطمه کاوه یزدی[2]
۱. استادیار مهندسی کامپیوتر، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران
۲. دانش‌آموخته دکترا مهندسی کامپیوتر، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

تاریخ دریافت: ۱۴۰۰/۰۳/۲۱
تاریخ اولین بازنگری: ۱۳۹۹/۰۴/۰۶
تاریخ پذیرش: ۱۳۹۹/۱۲/۱۸
تاریخ آنلاین: ۱۴۰۰/۰۶/۱۰

از دستگاه خود برای اسکن و خواندن مقاله به صورت آنلاین استفاده کنید

DOI:
10.30495/wej.2021.25230.2251

واژه‌های کلیدی:
پیش‌بینی میزان مصرف، رگرسیون، جنگل تصادفی، ناهنجاری، رگرسیون چندک.

چکیده

مقدمه: مدیریت هوشمند منابع آب بهترین راهکار برای معضل کمبود آب در سرتاسر جهان است. پیش‌بینی میزان مصرف یک پیش‌نیاز اصلی برای اطلاع از میزان آب مورد نیاز در آینده است. انواع مختلفی از ویژگی‌ها، از سابقه مصرف تا پارامترهای هواشناسی را می‌توان برای پیش‌بینی آب مصرفی بکار گرفت. در این مقاله، به معرفی یک مدل پیش‌بینی برای میزان مصرف آب مشترکین شهری در شهر یزد خواهیم پرداخت.

روش: چارچوب پیش‌بینی پیشنهادی از رکوردهای سامانه قبوض مصرف در شهر یزد برای استخراج سوابق مصرف مشترکین بهره می‌گیرد. به علاوه، منابع اطلاعاتی دیگری مانند تقویم کاری، میزان آب تولیدی (ورودی به شبکه شهری)، پارامترهای هواشناسی، ارزش مالی املاک مشترکین، و میزان فشار جریان آب ورودی به ملک مشترکین در پیش‌بینی مورد استفاده قرار می‌گیرند. این چهارچوب تغییرات در الگوی رفتار مصرف مشترکین را تعقیب می‌کند و آنها را گروه‌بندی می‌نماید تا بتواند مواردی را که رفتار غیرمتعارف دارند از میان آنها حذف کند. گروه‌های پاک شده (بدون موارد با مصرف نامتعارف) با استفاده از یک روش تخمین مبتنی بر چندک با سه خط برش مورد تحلیل قرار گرفته و براساس آنها میزان مصرف مشترکین در ماه آتی محاسبه می‌شود.

یافته‌ها: نتایج آزمایشات نشان می‌دهند که مدل پیشنهادی با خطای کمتر از ۱۰٪ می‌تواند میزان مصرف آتی را پیش‌بینی کند. به علاوه، این روش قادر است مشترکین با الگوی مصرف نامتعارف را نیز شناسایی کند.

نتیجه‌گیری: از میان روش‌های مورد بررسی، روش‌هایی توانسته‌اند با کمترین خطا میزان مصرف را پیش‌بینی کنند که به موارد غیرمتعارف مقاوم بوده‌اند. براساس بررسی‌های صورت گرفته این موارد ریشه در جابجایی ساکنین منازل دارند و بعد از جایگزینی یک مشترک کم‌مصرف/پرمصرف با یک مشترک پرمصرف/کم‌مصرف بروز می‌کنند. با الهام از این حقیقت و حذف اولین ماه‌های تغییر الگوی مصرف از دادگان و آموزش مدل یادگیری با باقیمانده موارد، می‌توان یک الگوریتم پیش‌بینی با دقت بالا داشت که در اکثر موارد خطای بسیار کمی داشته باشد.

* **نویسنده مسئول:** سجاد ظریف زاده
نشانی: دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران.
تلفن: ۰۳۵۳۸۲۰۰۱۴۵
پست الکترونیکی: szarifzadeh@yazd.ac.ir

مقدمه

با افزایش روزافزون جمعیت جهان و بروز تغییرات آب و هوایی منجر به خشکسالی، مسئله کمبود منابع آبی بیش از پیش رخ نموده و نیاز به واکنش‌های زودهنگام پیشگیرانه را به امری ضروری بدل ساخته است. معکوس نمودن روند تغییرات آب و هوایی و یا تأمین منابع آبی امن جدید برای توده‌های جمعیتی راهکاریهایی نیستند که در کوتاه‌مدت قادر به حل مسئله‌ی کمبود آب برای ساکنین کره زمین باشند؛ بنابراین به نظر می‌رسد بهترین اقدام در این مرحله، مدیریت هوشمند منابع آبی موجود و اجرای اقدامات پیشگیرانه برای مدیریت و کاهش مصرف آب است. مدیریت هوشمند منابع آبی می‌تواند شامل اقداماتی برای کاهش مصرف، مدیریت مصرف فصلی و راه‌اندازی کمپین‌های تبلیغاتی باشد. همه‌ی این اقدامات برای اجرای موثر نیاز به دادگانی از الگوهای رفتاری کاربران دارند تا با استفاده از آنها و طراحی مدل‌هایی متناسب با رفتار کاربران بتوان اثر هر یک از اقدامات را ارزیابی نموده و در مورد زمان یا وسعت اقدامات تصمیم مقتضی را اتخاذ نمود. برای این منظور باید دادگان مصرف مشترکین را در سال‌های مختلف جمع‌آوری نموده و برای ارزیابی مورد استفاده قرار داد. ولی تحلیل این دادگان چالش‌هایی از قبیل حجیم بودن، تاریخچه‌های ذخیره‌شده با استانداردهای متنوع و قدیمی، داده‌های ناکامل و متناقض را به دنبال دارد که موانع زیادی در مسیر مدل‌سازی و پیش‌بینی رفتار کاربران ایجاد می‌کنند.

پژوهش حاضر با هدف بررسی الگوهای رفتاری مصرف مشترکین خانگی در شهر یزد و پیش‌بینی میزان مصرف یک ماه آتی این مشترکین صورت گرفته و دادگان لازم برای اجرای آن توسط شرکت آب و فاضلاب شهری یزد فراهم گردیده است. دادگان خام از محل پایگاه‌های داده‌ی متنوع این شرکت به همراه دادگان تأمین شده از منابع بیرونی (مانند دادگان هواشناسی) در اختیار ما قرار داده شده است و عملیات لازم برای آماده‌سازی، پیش‌پردازش و در نهایت ارائه‌ی مدل پیش‌بینی بر روی آنها صورت گرفته است که مراحل آن در بخش‌های آتی به اجمال بیان خواهند شد. این پژوهش از چند جهت بر نمونه‌های قبلی خود برتری دارد که عبارتند از:

- ارائه یک مدل پیش‌بینی جدید از ترکیب مدل‌های تخمین‌زننده‌ی قطعه‌ای و تخمین‌زننده‌ی چندک برای بیش از هفتاد هزار مشترک خانگی منحصر به فرد که این حجم از داده در کارهای قبلی سابقه نداشته است.
- ارائه راهکاری مؤثر با بار محاسباتی کم و خطای کمتر از ۱۰٪ که برای استفاده در مقیاس عملی مناسب است.
- ارائه دسته‌بندی‌هایی از مشترکین آب شهری بر اساس مشابهت در الگوی مصرف آنها

نتایج حاصل از این پژوهش فرای جایگاه علمی آن دارای ارزش عملی است؛ زیرا می‌توان از نتایج آن در طراحی و پیاده‌سازی سیستم‌های مدیریت و پیش‌بینی مصرف آب، اصلاح فرایند محاسبه‌ی صورتحساب و تعرفه‌گذاری بهره برد.

در ادامه‌ی این مقاله، به تعدادی از مهم‌ترین تحقیقات در حوزه‌ی پیش‌بینی میزان مصرف آب مشترکین شهری اشاره خواهد شد. پژوهش‌های مرتبط با این حوزه را می‌توان بر اساس گستردگی به دو دسته تقسیم نمود: (۱) پژوهش‌هایی که مجموع میزان مصرف آب در یک محدوده‌ی بزرگ (به طور اخص یک شهر) را پیش‌بینی می‌کنند و (۲) پژوهش‌هایی که سعی می‌کنند میزان مصرف تک‌تک مشترکین را پیش‌بینی نمایند. تقریباً می‌توان گفت بیش از نود درصد پژوهش‌های انجام شده در سال‌های اخیر به مسئله‌ی اوّل پرداخته‌اند. حل این مسئله به دلیل وجود تنها یک سری زمانی از کل میزان مصرف یک منطقه در مقایسه با مسئله‌ی دوم که دربردارنده‌ی تعداد بسیار زیادی سری زمانی مستقل است، ساده‌تر می‌باشد. به علاوه، فرایند پیش‌بینی در مسئله‌ی اوّل به صورت دوره‌ای (حداکثر روزانه) و برای تنها یک سری زمانی اجرا می‌شود و در مقایسه با مسئله‌ی دوم که نیاز به اجرای متعدد در مقیاس بالا و با بارِ محاسباتی قابل توجه دارد، محدودیت کمتری برای استفاده از روش‌های پیچیده، ولی دقیق دارند. سومین چالش مسئله‌ی دوم به تعیین میزان خطای قابل قبول باز می‌گردد. در مسئله‌ی اوّل، میزان خالص مصرف هر روز یک شهر در مقایسه با میزان مصرف خالص یک مشترک عدد بسیار بزرگتر است و در نتیجه پیش‌بینی در مسئله‌ی اوّل یک پیش‌بینی بزرگ–مقیاس[1] است، در حالی که مسئله‌ی دوم یک پیش‌بینی خرد–مقیاس[2] است. به طور کلی، رسیدن به دقت نسبی مناسب در مسائل خرد–مقیاس پیچیده‌تر و پرهزینه‌تر از مسائل بزرگ–مقیاس است. یکی دیگر از چالش‌های مسائل خرد–مقیاس، نیاز به جمع‌آوری و پردازش تعداد نمونه‌های بیشتر است و به همین دلیل در بسیاری از موارد، محققین با کمبود نمونه‌ی لازم برای آموزش مدل‌های با دقت مشابه در اینگونه مسائل مواجه هستند.

پیش از پرداختن به تحقیقات منتشر شده در حوزه‌ی پیش‌بینی میزان مصرف آب به ارائه دو دسته‌بندی خواهیم پرداخت. اوّل، دسته‌بندی از دید داده‌های مورد استفاده و دوم، از دید روش‌های بکار گرفته شده در این عرصه. از دید داده‌ای این تحقیقات به چهار دسته تقسیم می‌شوند که عبارتند از:

- روش‌های مبتنی بر استفاده از اطلاعات مصرف (به تنهایی)
- روش‌های مبتنی بر استفاده از دادگان هواشناسی در کنار مصرف
- روش‌های مبتنی بر استفاده از دادگان اقتصادی–اجتماعی[3] و اجتماعی–جمعیت‌شناسی[4] در کنار مصرف
- روش‌های مبتنی بر استفاده از ترکیب همه انواع دادگان قبلی با تقویم کاری

تعداد مواردی که تنها از اطلاعات مصرف برای پیش‌بینی‌های خود استفاده کرده‌اند در مقایسه با سایر انواع دادگان کمتر است و این نمونه‌ها بیشتر به تحلیل‌های سری زمانی حافظه‌دار و فصلی رجوع نموده‌اند، مانند (۱، ۲، ۳، ۴، ۵). در ایران نیز پژوهش‌هایی در این حوزه به اجرا

[1] Macro-scale forecast
[2] Micro-scale forecast
[3] Socio-economic
[4] Socio-demographic

درآمده‌اند که تنها با اتکا بر سابقه‌ی مصرف مشترکین پیش‌بینی‌های مورد نظر خود را ارائه می‌نمایند، مانند (۶، ۷، ۸، ۹، ۱۰).

در دسته‌ی دوم، پژوهش‌هایی قرار می‌گیرند که همزمان با اطلاعات مصرف، از اطلاعات هواشناسی نیز برای تعیین میزان مصرف استفاده می‌کنند (ر.ک. ۱۱، ۱۲، ۱۳، ۱۴). از نمونه های بین‌المللی نیز می‌توان به مواردی، مانند (۱۵، ۱۶، ۱۷) ، اشاره نمود. مهمترین ویژگی این پژوهش‌ها توانایی آنها در ارائه‌ی پیش‌بینی‌های دقیق برای بازه‌های زمانی طولانی‌تر (نسبت به گروه اوّل) است.

دسته سوم، از دادگان اقتصادی–اجتماعی و اجتماعی–جمعیت‌شناسی برای مدل‌سازی الگوهای مصرف فرد–محور[1] استفاده می‌کنند. این دسته از پژوهش‌ها با این استنتاج که افراد با جایگاه‌های اجتماعی متفاوت، مقدار معینی آب را در طول روز مصرف می‌کنند، معتقد هستند استفاده از اطلاعات اجتماعی و جمعیتی می‌تواند در کنار اطلاعات مصرف، آنها را در پیش‌بینی‌ها موفق‌تر گرداند (۱۸، ۱۹). در ایران نیز مواردی از این دست به اجرا در آمده است، مانند (۲۰)، که طی آن اطلاعات دقیقی مانند قیمت ملک مشترک، درآمد سالانه، وسعت ملک، زیربنای مسکونی و تعداد شیرهای آب موجود در ملک شهری برای پیش‌بینی دقیق میزان مصرف هر خانوار جمع‌آوری گردیده است. امّا باید خاطرنشان نمود جمع‌آوری اطلاعات اقتصادی و ملکی با جزئیات فوق، به علت نیاز به هزینه‌ی بالا و نقض حقوق شهروندی و عدم امکان اجبار همه‌ی شهروندان به در اختیار قراردادن این اطلاعات در مقیاس شهری میسر نیست. آخرین دسته، پژوهش‌هایی را در برمی‌گیرد که از آمیختن اطلاعات مصرف و هواشناسی با اطلاعات تقویم کاری، مانند (۲۱) و اطلاعات مصرف با اطلاعات جمعیت‌شناسی و تقویم کاری، مانند (۲۲)، برای بالابردن دقت پیش‌بینی‌ها استفاده می‌کنند.

پس از بررسی پژوهش‌ها بر اساس نوع دادگان، نوبت به بررسی این پژوهش‌ها با توجه به روش‌های مورد استفاده می‌رسد. پژوهش‌های موجود را می‌توان بر اساس نوع روش مورد استفاده به چهار دسته تقسیم نمود:

- روش‌های مبتنی بر استفاده از تخمین‌زننده[2]
- روش‌های مبتنی بر استفاده از شبکه‌های عصبی مصنوعی
- روش‌های مبتنی بر مدل‌سازی میانگین متحرک سری‌های زمانی
- سایر روش‌ها

بزرگترین دسته از روش‌های مورد استفاده برای پیش‌بینی، از انواع تحلیل‌های تخمین‌زننده برای پیش‌بینی مقادیر آتی بهره گرفته‌اند که از آن جمله می‌توان به تحلیل تخمین‌زننده خطی بیزین[3] (۱۵)، تخمین‌زننده خطی مبتنی بر پله[4] (۱۹) و تخمین‌زننده چندمتغیره[5] (۱۸) و (۲۳) اشاره کرد.

دسته‌ی دوم از پژوهش‌ها بیشتر متکی بر استفاده از شبکه‌های عصبی هستند که عمدتاً از شبکه‌های عصبی چندلایه پرسپترون[6] استفاده کرده‌اند، مانند (۸، ۹، ۱۳، ۲۴، ۲۵). امّا از میان این موارد، تقوایی و همکاران (۲۰) از مدل رگرسیون خطی در کنار شبکه‌های عصبی چندلایه پرسپترون و فتحی (۹) از شبکه‌های عصبی بازگشتی[7] در کنار آنها بهره برده‌اند. حتی در مواردی از شبکه‌های متنوع‌تری نیز بهره گرفته شده است که از آن جمله می‌توان به شبکه‌های FFNN، CCNN، GRNN (۴)، شبکه‌های WBNN، WNN، BNN (۵) و شبکه‌های ELM (۱۶) اشاره نمود. بر اساس یافته‌های این تحقیقات، به نظر می‌رسد شبکه‌هایی که توانایی خلاصه‌سازی و نمونه‌گیری بهتری دارند مانند شبکه‌های WBNN و WNN نتایج بهتری را به دنبال دارند. علاوه بر این موارد، فلورس و همکاران (۲۶) به جای استفاده از یک شبکه‌ی عصبی مصنوعی، مجموعه‌ای از این شبکه‌ها را به کار بردند که ساختار، متغیرهای ساختاری و وزن‌های آنها با استفاده از الگوریتم ژنتیک تعیین شده‌اند.

دسته سوم مجموعه‌ی پژوهش‌هایی را تشکیل می‌دهند که از روش‌های تجزیه و تحلیل سری‌های زمانی و به طور اخص مدل میانگین متحرک خود همبسته یکپارچه فصلی[8] استفاده می‌کنند. به عنوان نمونه آرام و عاقلی کهنه‌شهری (۱۱) و موسوی و کاووسی کلاشمی (۷) در پژوهش‌های خود از روش SARIMA در کنار شبکه‌های عصبی بهره گرفته‌اند، در حالی که تابش و همکاران (۱۰) از ترکیب این روش با رگرسیون و یزدانی و همکاران (۶) از روش SARIMA به تنهایی بهره برده‌اند. نمونه‌هایی مانند (۲۱ و ۲۲) از مدل میانگین متحرک خود همبسته یکپارچه[9] استفاده می‌کنند که با استفاده از مقادیر قبلی سری‌های زمانی دارای رفتار خودهمبسته‌ی تکرارشونده برای پیش‌بینی بهره می‌برند. برای استفاده از این روش‌ها لزومی برای درنظرگرفتن رفتار فصلی همانند نمونه‌های مشابهی که از SARIMA بهره می‌گیرند، وجود ندارد. به علاوه، استخراج پارامترهای این مدل‌ها نسبت به مدل SARIMA ساده‌تر و کم هزینه‌تر است. روش‌های مبتنی بر مدل‌سازی فازی و الگوریتم ژنتیک (۱۷)، مدل‌سازی مبتنی بر عامل هوشمند[10] (۱) و تحلیل اجزای مستقل[11] (۲) از دیگر روش‌هایی هستند که در پژوهش‌های پیش‌بینی مصرف آب شهری به کار گرفته شده‌اند.

در پایان، باید این بخش را بدین صورت جمع‌بندی نمود که از میان روش‌های مورد استفاده، روش‌های مبتنی بر تحلیل تخمین‌زننده و استفاده از شبکه‌های عصبی مصنوعی بیش از سایر روش‌ها مورد توجه قرار گرفته‌اند. دلیل این امر را باید در اهمیت مقادیر قبلی مصرف جست. علاوه بر اهمیت مقادیر قبلی، متغیرهای هواشناسی و متغیرهای جمعیت‌شناسی در جایگاه‌های دوم و سوم اهمیت قرار دارند. هر چند اطلاعاتی مانند تقویم کاری و اطلاعات اقتصادی و مشخصات ملک مشترکین نیز در جایگاه‌های بعدی مورد استفاده قرار گرفته‌اند ولی

[1] Individual consumption models
[2] Regression
[3] Bayesian Linear Regression (BLR)
[4] Step-wise Linear Regression (SLR)
[5] Multivariate Regression
[6] Multi-Layer Perceptron (MLP)
[7] Recurrent Neural Network (RNN)
[8] Seasonal Auto-Regressive Integrated Moving Average Model (SARIMA)
[9] Auto-Regressive Integrated Moving Average Model (ARIMA)
[10] Smart agent modeling
[11] Independent Component Analysis (ICA)

میزان تأثیر آنها به اندازه موارد سه‌گانه‌ی اوّل نبوده است. به علاوه، جمع‌آوری اطلاعات مصرف، هواشناسی و متغیرهای جمعیت‌شناسی بسیار ساده‌تر و کم هزینه‌تر هستند که به همین دلیل، استفاده از این دادگان به نسبت گسترده‌تر می‌باشد.

مواد و روش‌ها

پیش از پرداختن به معرفی روش پیشنهادی به توصیف داده‌هایی می‌پردازیم که با همکاری شرکت آب و فاضلاب شهری یزد آماده شده و در این پژوهش مورد استفاده قرار گرفته است. این دادگان در گروه‌های زیر قابل دسته‌بندی می‌باشد:

- توصیف جایگاه اجتماعی–اقتصادی بخشی از مشترکین
- دادگان روزهای کاری و تعطیلات در سال‌های گذشته
- دادگان هواشناسی
- دادگان مصرف، تولید و فشار آب

در ادامه‌ی این نوشتار، هر یک از این گروه‌ها به صورت اجمالی معرفی شده و ویژگی‌های آماری و عددی آنها بررسی می‌گردند.

دادگان اجتماعی–اقتصادی

شرکت آب و فاضلاب برای توصیف جایگاه اجتماعی–اقتصادی مشترکین خود از پارامتری به نام مرغوبیت مکانی بهره می‌برد که در ادبیات سازمانی مصطلح در این حوزه f–مرغوبیت مکانی نامیده می‌شود. این پارامتر، به هر یک از املاک شهری براساس متراژ، محله و ساختمان عددی صحیح نسبت می‌دهد که مقادیر بالاتر در آن به معنای مرغوبیت بالاتر و در نتیجه طبقه‌ی اجتماعی–اقتصادی مرفه‌تر است. در این راستا اطلاعات مرغوبیت ملکی ۱۵۶۶۸۳ مشترک شهری از سامانه‌ی اطلاعات مکانی شرکت آب و فاضلاب یزد استخراج شده است که دارای ۱۰۲۲ مقدار منحصر به فرد صحیح می‌باشد، یعنی به عبارت دقیقتر، مرغوبیت املاک مشترکین شهری در یزد در ۱۰۲۲ سطح مجزا دسته‌بندی می‌گردد.

دادگان تقویمی

دادگان روزهای کاری و تعطیلات یک جدول دو ستونی است که در آن به ازای تمام روزهای مابین ۱۳۸۰/۱/۱ تا ۱۳۹۶/۱۲/۲۹، نوع روز شامل روزکاری، جمعه و روز تعطیلِ غیرجمعه ثبت شده است. در مواردی که یک مناسبت تعطیل در روز جمعه واقع شده است، برای این روز برچسب جمعه در نظر گرفته شده است.

دادگان هواشناسی

دادگان هواشناسی در تحقیقات مشابه در تخمین و پیش‌بینی میزان مصرف آب، برق و گاز از اهمیت بسزایی برخوردار هستند؛ زیرا به تبع تغییرات در درجه حرارت و شرایط جوی و همچنین نوع استفاده از هر یک از این منابع، الگوهای مصرف می‌توانند تغییر کنند. به عنوان نمونه، در فصول گرم استفاده از آب به عنوان عامل خنک‌کننده بیشتر می‌شود و یا در فصول سرد، استفاده از برق و گاز به عنوان عامل گرم‌کننده (هر چند که در فصول گرم، استفاده از برق نیز به عنوان تامین کننده‌ی انرژی سیستم‌های خنک‌کننده روند افزایشی دارد). دادگان هواشناسی تأمین شده در این پروژه یکی از دادگان ارزشمندی است که برای این پروژه تامین شده است. این دادگان مشتمل بر ۶۴۹۵ سطر است که هر سطر ویژگی‌های زیر را در خود جای داده است:

- تاریخ
- دما (سانتیگراد)
- رطوبت (درصد)
- بارندگی (میلیمتر)
- تبخیر (میلیمتر)
- ساعت آفتابی (ساعت)
- سرعت باد حداکثر (متر بر ثانیه)
- سمت باد حداکثر (درجه)
- حداقل دید افقی (متر)
- گزارشات همراه با پدیده گرد و خاک و شن (تعداد)
- تعداد گزارشات همراه با پدیده گرد وخاک (تعداد)

هر سطر از این دادگان با تاریخ یک روز آغاز می‌شود که نخستین روز این دادگان ۲۱ مارس ۲۰۰۱ میلادی برابر با ۱ فروردین ۱۳۸۰ هجری شمسی و آخرین روز ۳۱ دسامبر ۲۰۱۸ میلادی برابر با ۱۰ دی ۱۳۹۷ است. از بین ویژگی‌های موجود، دما و میزان رطوبت برای هر روز به ترتیب دارای سه و دو مقدار هستند، بدین صورت که برای هر روز کمینه، بیشینه و میانگین دما به همراه کمینه و بیشینه رطوبت ثبت شده است. سایر ویژگی‌های موجود در این دادگان تک مقداری هستند.

دادگان مصرف، تولید و فشار آب

دادگان مصرف، تولید و فشار آب در واقع مشخص‌ترین تصاویر را از الگوی مصرف آب از مبدا تا مقصد نهایی در شهر یزد نشان می‌دهند. این دادگان در سه مجموعه‌ی جداگانه تولید شده‌اند که در این بخش هر یک از آنها به تفکیک معرفی خواهند شد. در واقع، هدف اصلی در این پروژه معرفی روشی برای پیش‌بینی میزان مصرف هر یک از مشترکین است و به همین دلیل این دادگان بدنه‌ی اصلی تحلیل‌های داده‌ای آتی را تشکیل می‌دهند.

دادگان مصرف آب از پایگاه داده‌ی شرکت آب و فاضلاب شهری یزد استخراج شده و شامل اطلاعات صورتحساب‌های ۷۴,۹۰۸ مشترک منحصر به فرد در بازه‌ی بین ۱۲ دی ماه ۱۳۹۰ (مطابق با یکم ژانویه ۲۰۱۳) تا ۲۹ اسفند ۱۳۹۶ (مطابق با بیستم مارس ۲۰۱۸) به صورت خام است. این مجموعه رکورد با اعمال شروط خاصی از مجموعه‌ی تمام رکوردهای موجود در بازه زمانی اعلام شده استخراج شده‌اند تا احتمال بروز اختلال به دنبال برخی از ناهنجاری‌ها مانند کنتورهای خراب را به حداقل برسانند. این شروط عبارتند از:

- عدم وجود سابقه‌ی تعویض کنتور در بازه‌ی زمانی مذکور
- مصرف حداقل ۱۰ مترمکعب در حداقل یک دوره‌ی با طول ۳۰ روز یا بیشتر در یک سال

مجموعه رکوردهای انتخاب شده برای این مشترکین شامل ۵۳۲،۸۷۱ رکورد صورتحساب است که هر یک از آنها ۴۶ ستون دارند که اطلاعات هر صورتحساب و بخشی از اطلاعات کلی مشترک مانند کد قطر آب، تعداد خانوار و موارد مشابه را شامل می‌شوند. باید خاطرنشان نمود که بازه‌های صدور صورتحساب کاملاً متنوع هستند و از صورتحساب‌های کمتر از ده روز (به دلیل درخواست تسویه حساب) تا شش ماهه را شامل می‌شوند ولی به طور متوسط بیش از هفتاد درصد صورتحساب‌ها، طولی بین ۵۵ تا ۶۴ روز دارند.

دادگان تولید آب روزانه مقادیر تولید آب سطحی، غیرسطحی و مجموع تولید آب روزانه شهر یزد را برای تمام روزهای بین یکم فروردین ۱۳۹۰ تا ۱۸ اسفند ۱۳۹۶ شامل می‌شود. تولید آب سطحی در شهر یزد از محل خط لوله‌ی آب انتقالی از شهر استان اصفهان و تولید غیرسطحی از محل چاه‌های عمیق در محدوده‌ی شهر یزد و مناطق اطراف آن تأمین می‌شود و مجموع این مقادیر در واقع تعیین کننده‌ی مقدار آبی است که به صورت روزانه به شبکه‌ی آب شهری در شهر یزد تزریق می‌گردد.

دادگان فشار آب با استفاده از ترازیابی شبکه‌ی سراسری آب در شهر یزد و با انطباق دادگان آن با نقشه‌های موجود در سیستم اطلاعات مکانی شرکت آب و فاضلاب شهری یزد به دست آمده است و در واقع مشخص می‌کند هر مشترک در کدام تراز فشار در شبکه قرار می‌گیرد و انشعاب این مشترک به صورت عمومی دارای چه فشاری است. این دادگان، مقادیر فشار آب برای ۱۷۳۲۳۲ مشترک شهری در یزد را در برمی‌گیرد و مقادیر فشار مندرج در این داده در بازه‌ی [۳/۵۵۰۶، ۰/۳۲۶۹] متغیر هستند. بخش عمده مقادیر در بازه‌ی [۳/۰، ۱/۵] قرار می‌گیرند ولی ۱۹۹۰ مشترک فشار ورودی کمتر از ۱/۵ و ۲۱۹۴ مشترک فشار ورودی بیش از ۳ دارند.

آماده‌سازی و انتخاب دادگان

از میان دادگان موجود، تنها دادگان مصرف هستند که بسامد نمونه‌گیری معین و برابر ندارند و هر یک از مقادیر صورتحساب‌ها برای بازه‌ای دلخواه صادر شده است. داده‌های مصرف ماهانه مشترکین معمولاً هر دو ماه یکبار جمع‌آوری می‌شوند ولی در صورت بروز مشکلاتی مانند شکستگی، قطعی، نشت و یا تخلیه‌ی منزل برای مشترکین صورتحساب‌هایی با طول کمتر یا بیشتر از دو ماه نیز صادر می‌شود. با توجه به داده‌های واقعی بازه مابین دو صورتحساب متوالی ممکن است بین ۷ تا ۱۰۰ روز متغیر باشد. این انعطاف در صدور صورتحساب، فرایند آموزش و تدوین یک مدل یکپارچه که بتواند پیش‌بینی دقیقی با طول بازه‌های متغیر ارائه دهد را با چالشی بزرگ مواجه می‌کند و بار محاسباتی قابل توجهی را به مدل تحمیل می‌کند. به علاوه، هدف از اجرای این پژوهش، پیش‌بینی مقادیر مصرف ماهانه مشترکین شهری برای بازه یک ماهه‌ی آتی است. بر اساس دلایل بالا دادگان اولیه نیز در بدو امر به دادگان ماهانه تبدیل شده‌اند. برای این منظور و با توجه به نمونه‌هایی مانند (۱۵) از روش درون‌یابی اسپلاین مرتبه یک بهره گرفته شده و مقادیر مصرف مشترکین در انتهای تمام ماه‌های میلادی محاسبه می‌گردد.

در گام بعد باید سایر دادگان را نیز برای استفاده در این چارچوب هماهنگ نمود و مواردی که دارای بسامد روزانه هستند را به دادگانی با بسامد ماهانه تبدیل نمود. این عملیات برای دادگان روزهای کاری و تعطیلات به سهولت انجام می‌گردد و می‌توان برای تمام ماه‌های سال، تعداد روزهای جمعه، تعطیلِ غیر جمعه و روزهای کاری را از مجموع تعداد این روزها در آن ماه محاسبه نموده و به صورت سه ویژگی جداگانه برای هر ماه مورد استفاده قرار داد. دادگان هواشناسی جمع‌آوری شده برای این پژوهش نیز باید برای استفاده به صورت ماهانه مناسب‌سازی شوند. یکی از ساده‌ترین روش‌هایی که ممکن است برای تجمیع این داده‌ها به نظر برسد استفاده از میانگین است، امّا رسی‌فقیهی و همکاران (۱۵)، گتو و همکاران (۲۷) و همچنین اسلامیان و همکاران (۲۳) نشان داده‌اند که پارامترهای بیان کننده‌ی وضع هوا مانند دما و رطوبت به صورت یکنواخت بر میزان مصرف آب موثر نیستند و در واقع میزان مصرف آب ناشی از افزایش درجه حرارت/کاهش رطوبت زمانی که درجه حرارت به مقداری بالاتر/پایینتر از یک خطبرش[1] برسد به صورت خطی افزایش می‌یابد. با این اوصاف، اگر از میانگین برای تجمیع دادگان روزانه و تبدیل آنها به ماهانه بهره گرفته شود، در واقع ارزش نقاط با دما/رطوبت بالاتر یا پایینتر از میانگین در افزایش مصرف از بین می‌رود. برای جلوگیری از این مشکل از روشی که در تحقیقات مربوط به مصرف انرژی متداول‌تر است و از آن به عنوان نرمال‌سازی آب و هوایی نام برده می‌شود، بهره گرفته خواهد شد (۲۸).

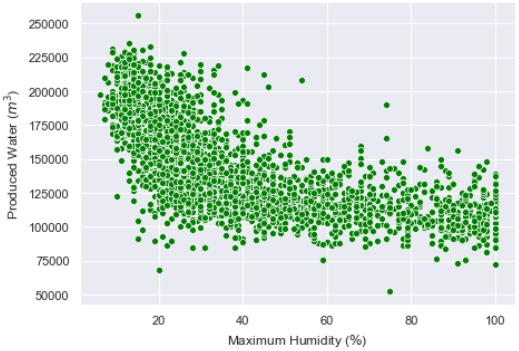
در این روش با توجه به رابطه‌ی بین مصرف و پارامترهای هواشناسی، برای آنها خطوطبرشی تعیین می‌گردد و مقادیر بالاتر/پایینتر از این خطوطبرش تجمیع می‌شوند (مانند ۱۵، ۲۳، ۲۹، ۳۰، ۳۱، ۳۲، ۳۳، ۳۴، ۳۵). به ازای هر یک از پارامترهای هواشناسی که جنبه‌ی شمارشی نداشته باشند (مثلاً دما، در مقابل تعداد روزهای بارانی که شمارشی است) می‌توان یک یا دو خطبرش تعریف نمود و سپس با تجمیع مقادیر اختلاف بین مقدار آن پارامتر با خطبرش در بسامد روزانه به یک مقدار واحد در مقیاس ماهانه رسید. در صورتی‌که بین آن پارامتر و مقادیر خاصی از آن که بالاتر از یک خطبرش خاص هستند، رابطه برقرار باشد، مقدار پارامتر از خطبرش کسر می‌شود و در صورتی که روند عکس باشد یک خطبرش برای مقادیر پایین‌تر در نظر گرفته می‌شود. روابط (۱) و (۲) نحوه‌ی محاسبه‌ی مقادیر تجمیع شده را برای هر دو نوع خطبرش نشان می‌دهند.

$$F^H = \sum_{j=1}^{n} \Delta_j^H \quad \Delta_j^H = \begin{cases} v_j - T^H & if\ v_j \geq T^H \\ 0 & v_j < T^H \end{cases} \tag{1}$$

$$F^L = \sum_{j=1}^{n} \Delta_j^L \quad \Delta_j^L = \begin{cases} T^L - v_j & if\ v_j \leq T^L \\ 0 & f v_j > T^L \end{cases} \tag{2}$$

[1] Threshold

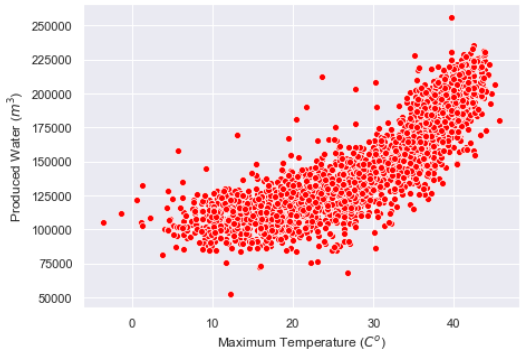
مقادیر T^H و T^L به ترتیب خطبرش‌های بالا و پایین هستند و v_j مقدار ویژگی v مربوط به j–امین نمونه در دادگان را مشخص می‌کند. باید خاطرنشان نمود که مقادیر این پارامترها بر روی دادگان روزانه در طول یک ماه محاسبه شده و در نهایت مجموع آنها به عنوان مقدار پارامتر مربوطه برای آن ماه ثبت می‌گردد. در نخستین گام پس از معرفی این روش، رابطه‌ی بین میزان تولید آب روزانه شهری را با یکی از پارامترهای هواشناسی بررسی نموده و بر اساس آن به توصیف فرایند محاسباتی بالا خواهیم پرداخت.



شکل ۱ – نمودار نقطه‌ای نشان‌دهنده‌ی رابطه بین مقادیر تولید آب روزانه و بیشینه‌ی رطوبت.

شکل ۱ نشان دهنده‌ی رابطه‌ی بین مقادیر تولید آب روزانه و بیشینه‌ی رطوبت می‌باشد. برای تولید این نمودار از دادگان مجموع تولید آب در بازه‌ی مابین سال‌های ۱۳۹۰ تا ۱۳۹۶ شمسی و دادگان هواشناسی در همین بازه‌ی زمانی بهره گرفته شده است. می‌توان دید برای مقادیر رطوبت بالای ۵۰ درصد روند تولید آب روزانه، کاهشی با شیب کم و برای مقادیر کمتر از ۴۰ درصد نیز روند کاهشی ولی با شیب متفاوت و زیادتر است. در بازه‌ی بین ۴۰ تا ۵۰ درصد تقریباً نمودار به حالت افقی در می‌آید[1]. در این حالت عدد ۵۰ به عنوان مقدار T^H بیشینه‌ی رطوبت در نظر گرفته می‌شود، زیرا برای تمام مقادیر بیشینه‌ی رطوبت بالاتر از آن می‌توان به یک رابطه خطی با شیب معین بین رطوبت و تولید آب روزانه دست یافت. از طرفی برای مقادیر بیشینه‌ی رطوبت کمتر از ۴۰٪ نیز می‌توان رابطه‌ی خطی با شیب متفاوت بین پارامترهای رطوبت و تولید آب روزانه یافت و به همین دلیل عدد ۴۰ می‌تواند نقش مقدار T^L بیشینه‌ی رطوبت را ایفا کند. در این صورت پس از استخراج مقادیر بیشینه‌ی رطوبت، برای تمام روزهایی که مقدار رطوبت آنها از ۵۰٪ بالاتر باشد، مقدار رطوبت از عدد ۵۰ کسر شده و مجموع اعداد حاصل برای تمام روزهای بازه‌ی یک ماه به عنوان مقدار ویژگی تجمیع شده‌ی T^H برای بیشینه‌ی رطوبت ذخیره می‌شوند. همینطور برای روزهایی از یک ماه که مقدار بیشینه‌ی رطوبت در آنها کمتر از ۴۰٪ باشند، مقدار بیشینه‌ی رطوبت از عدد ۴۰ کسر شده و مجموع مقادیر حاصل، ویژگی تجمیع شده‌ی T^L برای بیشینه‌ی رطوبت در آن ماه را تشکیل می‌دهند. باید به این نکته اشاره کرد که معمولاً وجود یکی از خطبرش‌ها و محاسبه‌ی یک ویژگی تجمیع شده برای اکثر پارامترها کفایت می‌کند و اجباری به استفاده از هر دو خطبرش T^H و T^L نیست.

برای مشاهده‌ی نمونه پارامتری که تنها به یک خط برش نیاز دارد، باید به مطالعه رابطه‌ی بین مقادیر بیشینه‌ی دما و تولید آب روزانه پرداخت. شکل ۲ نمودار نقطه‌ای نشان‌دهنده‌ی رابطه بین مقادیر بیشینه‌ی دما (محور افقی) و مقادیر کلی تولید آب روزانه (محور عمودی) را نشان می‌دهد که تغییرات کلی در الگوی مصرف آب در مقیاس شهری را با توجه به تغییرات دما آشکار می‌کند.



شکل ۲ – نمودار نقطه‌ای نشان‌دهنده‌ی رابطه بین مقادیر تولید آب روزانه و بیشینه‌ی دما.

آنچه از این تصویر به نظر می‌رسد، آن است که تا قبل از ۱۰ درجه سانتیگراد تغییرات میزان آب تولیدی رابطه‌ی افزایشی با بیشینه‌ی دما ندارند ولی در بازه‌ی ۱۰ تا ۲۰ درجه سانتیگراد این رابطه با شیب مثبت آغاز می‌شود. مشابه این بازه‌ی دمایی در پژوهش‌های مشابه نیز گزارش شده است و به عنوان مثال، اسلامیان و همکاران (۲۳) دمای ۱۰ درجه سانتیگراد را برای حداقل بیشینه‌ی دمای موثر بر میزان مصرف در شهر بروسار در کبک کانادا، گَتو و همکاران (۲۷) دمای ۱۵/۵۳ درجه سانتیگراد را برای شهر ملبورن در استرالیا و تیواری و آداموسکی دمای ۱۲ درجه سانتیگراد را برای شهر مونترال در کانادا بدست آورده‌اند. در این مسئله با توجه به وجود پارامترهای کمینه، بیشینه و میانگین دما، باید مقدار خطوط برش برای هر یک از این موارد به صورت جداگانه تعیین گردند.

برای تعیین خطوط برش، می‌توان از سه روش استفاده نمود: (۱) تعیین خودکار این نقاط توسط برخی از روش‌ها (مانند ۱۵، ۳۶ و ۳۷)، (۲) تغییر مقادیر خطبرش و بررسی اثر پارامتر تجمیع شده در تخمین‌زننده برای تخمین مقدار آب مصرفی با استفاده از تخمین‌زننده‌ی قطعه‌ای[2] (۳۸، ۳۹) ، (۳) تعیین خطوط برش با استفاده از نمودار نقطه‌ای به صورت چشمی (۲۳). در این مقاله از روش دوم برای تعیین خطوطبرش بالا و پایین بهره گرفته می‌شود. برای این منظور برای تمام متغیرهای آب و هوایی غیرشمارشی (مانند دما)، خطوط برش بالا و پایین در بازه‌ی بین مقادیر بیشینه و کمینه تعیین می‌گردند. برای تعیین خطوط برش برای هر متغیر انتخاب شده، ابتدا تمام نقاط بین بیشینه و

[1] فرایند تعیین نقاط ۴۰٪ و ۵۰٪ با استفاده از تخمین‌زننده‌ی قطعه‌ای در ادامه توضیح داده خواهند شد.

[2] Piecewise regression

کمینه به نوبت به عنوان خطبرش بالا مورد آزمون قرار می‌گیرند و برای همه‌ی آنها فرایند تجمیع معرفی شده در رابطه‌ی (۱) به اجرا در آمده و سپس با استفاده از میزان خطای تخمین‌زننده‌ای که با استفاده از این

قابل مشاهده است. از میان ویژگی‌ها دو ویژگی "تعداد گزارش گرد و خاک" و "تعداد گزارش گرد و خاک و شن" که شمارشی هستند، بدون اعمال خطبرش در طول هر ماه شمارش شده و مورد استفاده قرار می‌گیرند.

پس از محاسبه‌ی ویژگی‌های تجمیع شده برای هر ماه، با استفاده از یکی از روش‌های انتخاب ویژگی، باید فهرستی از ویژگی‌های مرتبط را برای پیش‌بینی الگوی کلی رفتار مصرف آب کاربران مورد استفاده قرار داد که برای این منظور در ادامه از روش LassoCV (۴۰) بهره گرفته می‌شود. این روش یک تابع هدف به صورت مندرج در رابطه (۳) دارد و طی آن تمام ویژگی‌های ورودی را در قالب ماتریس X و پارامتر بهینه‌سازی (یعنی α دلخواه) دریافت می‌کند و تلاش می‌کند با کمینه نمودن تابع هدف، مقدار مورد نظر (یعنی y) را تخمین بزند.

$$\min_{\omega} \frac{1}{2n_{sample}} \|X\omega - y\|_2^2 + \alpha\|\omega\|_1 \qquad (3)$$

با استفاده از این روش و بر اساس ضرائب پارامترهای هواشناسی برای پیش‌بینی مقادیر مصرف کلی (که تحت تأثیر وضعیت هوا و عوامل جوی است)، موثرترین‌ها انتخاب شده و برای استفاده در مدل‌های پیش‌بینی آماده‌سازی می‌شوند. نتایج میزان تأثیر هر یک از پارامترهای تجمیع شده در جدول ۲ قابل مشاهده است. بیشترین تأثیر از بین این موارد به میزان تبخیر و تعداد ساعات آفتابی مربوط می‌شود. امّا باید به یک نکته مهم توجه نمود: میزان همبستگی تبخیر، ساعات آفتابی و کمینه‌ی دما با میزان آب مصرفی به ترتیب برابر ۰/۷۶، ۰/۹۳ و ۰/۹۶ است.

بنابراین با توجه به خطر هم‌راستایی در صورت انتخاب پارامترهای تبخیر و ساعات آفتابی و مقدار همبستگی قابل توجه کمینه‌ی دما، به نظر می‌رسد بهترین انتخاب صرفنظر کردن از پارامترهای تبخیر و ساعات آفتابی به نفع کمینه‌ی دما است.

ویژگی تجمیع شده مقدار تولید آب روزانه را تخمین می‌کند می‌توان بهترین نقاط برش را تعیین کرد. روال مشابه برای خطبرش پایین نیز اجرا می‌شود. این دو روال برای تمام متغیرها به اجرا در آمده که نتایج آن به صورت خلاصه در

در نهایت و با توجه به مقادیر میزان تأثیر هر یک از ویژگی‌ها و حذف ویژگی‌های هم‌راستا (موصوف در بالا)، فهرست پارامترهای آب و هوایی انتخاب شده شامل {کمینه‌ی دما با برش T^H، بیشینه رطوبت با برش T^H، بیشینه رطوبت با برش T^L، سرعت باد با برش T^H، جهت باد با برش T^H، جهت باد با برش T^L} است. در مرحله‌ی بعد، دادگان صورتحساب‌های مصرف مورد بررسی قرار می‌گیرند. ویژگی‌های موجود در جدول صورتحساب مشترکین به سه دسته تقسیم می‌شوند:

- ویژگی‌های ثابت غیرمقداری برای هر دوره صورتحساب، مانند شناسه اطلاعات صدور صورتحساب
- ویژگی‌های ثابت مشترک، مانند مرغوبیت ملکی و فشار آب
- ویژگی‌های متغیر محاسباتی هر دوره صورتحساب، مانند میزان مصرف و تعداد روز

با توجه به اینکه دادگان مصرف با استفاده از اسپلاین به قالب ماهانه تبدیل شده‌اند، مجموعه‌ی رکوردهای قبلی (بعد از اعمال اسپلاین) به ۳,۷۴۶,۰۴۲ رکورد افزایش یافته‌اند. به دنبال اعمال اسپلاین، ویژگی‌هایی مانند تعداد روز و مبلغ صورتحساب نیز باید به صورت ماهانه محاسبه شوند و مقادیر جدید برای صورتحساب‌های ماهانه منظور گردند.

ویژگی‌های ثابت مشترکین نیز در کنار هر رکورد درج می‌شوند. از این رکوردها ویژگی‌هایی از قبیل "شناسه اطلاعات صدور صورتحساب"، "شناسه قبض"، "شناسه پرداخت" و "شماره سطر هر مشترک" نیز حذف می‌شوند. در نهایت، هر رکورد مشتمل بر ویژگی‌های {شماره اشتراک، تعداد واحد، تعداد خانوار، کد قطر آب، کد مانع، کد کاربری مصرف، مصرف، بدهی فاضلاب، بدهی قبلی صورتحساب، کد قطر فاضلاب، روز، مبلغ قبض، اقساط فاضلاب، بدهی اقساط فاضلاب،

جدول ۱ - مقادیرTH و TL تعیین شده برای هر یک از متغیرهای هواشناسی با استفاده از رگرسیون قطعه‌ای.

#	متغیر	T^H	T^L	#	متغیر	T^H	T^L
۱	کمینه دما (C^o)	۸	--	۸	ساعات آفتابی (h)	۸	--
۲	میانگین دما (C^o)	۱۰	--	۹	سرعت باد (m/s)	۱۵	۵
۳	بیشینه دما (C^o)	۱۵	--	۱۰	جهت باد (^{o})	۱۵۰	۲۵۰
۴	کمینه رطوبت (٪)	۳۰	۱۷	۱۱	دید افقی (m)	۱۰۰۰۰	--
۵	بیشینه رطوبت (٪)	۵۰	۴۰	۱۲	تعداد گزارش گرد و خاک و شن	--	--
۶	میزان بارش (mm)	۴	--	۱۳	تعداد گزارش گرد و خاک	--	--
۷	تبخیر (mm)	۱۵	--				

جدول ۲ - رتبه‌ی و میزان تاثیر هر یک از ویژگی‌های هواشناسی در تخمین مقدار مصرف آب در کل محدوده‌ی شهری یزد با استفاده از روش LassoCV.

#	ویژگی	برش	میزان تأثیر	رتبه	#	ویژگی	برش	میزان تأثیر	رتبه
۱	کمینه دما (C^o)	T^H	۲۷۲۰/۷۰	۴	۱۰	ساعات آفتابی (h)	T^H	۳۱۷۹/۵۱	۳
۲	میانگین دما (C^o)	T^H	۰/۰۰۰۰	۱۱	۱۱	سرعت باد (m/s)	T^H	۳۶۶/۱۸	۶
۳	بیشینه دما (C^o)	T^H	۰/۰۰۰۰	۱۱	۱۲	سرعت باد (m/s)	T^L	۰/۰۰۰۰	۱۱
۴	کمینه رطوبت (٪)	T^H	۰/۰۰۰۰	۱۱	۱۳	جهت باد (^{o})	T^H	۳۵۸/۶۳	۷
۵	کمینه رطوبت (٪)	T^L	۰/۰۰۰۰	۱۱	۱۴	جهت باد (^{o})	T^L	۴۸۷/۸۱	۵
۶	بیشینه رطوبت (٪)	T^H	-۴۱/۸۳	۹	۱۵	دید افقی (m)	T^H	۱/۲۳	۱۰
۷	بیشینه رطوبت (٪)	T^L	۲۵۲/۴۹	۸	۱۶	تعداد گزارش گرد و خاک و شن	--	۰/۰۰۰۰	۱۱
۸	میزان بارش (mm)	T^H	-۱۱۲۴۶/۲۵	۱	۱۷	تعداد گزارش گرد و خاک	--	۰/۰۰۰۰	۱۱
۹	تبخیر(mm)	T^H	۵۹۷۲/۱۹	۲					

ظرفیت مقطوع فاضلاب، ظرفیت مقطوع آب، میانگین مصرف بازرسی، مبلغ فروش خدمات، فشار، مرغوبیت مکانی} خواهد بود. ویژگی‌هایی مانند کد مانع که ترتیبی هستند با استفاده از متغیرهای ساختگی[1] نمایش داده شده‌اند. با توجه به اینکه در برخی از ویژگی‌ها مانند کد قطر فاضلاب (به علت عدم برقرای اشتراک فاضلاب) برای تعداد نسبتاً زیادی از مشترکین مقادیر صفر درج شده است و به علاوه، به علت تنوع ویژگی‌های مورد استفاده و احتمال بروز الگوهای پیچیده در رفتار مشترکین، برای انتخاب ویژگی از یک تخمین‌زننده‌ی جنگل تصادفی[2] بهره گرفته شده است. این تخمین زننده، برای تخمین از ۱۰۰ درخت بهره می‌برد و استفاده از نمونه‌گیری بوت‌استرپ[3] (۴۱) نیز برای آن فعال شده است. نتایج میزان تأثیر این ویژگی‌ها در **Error! Reference source not found.** قابل مشاهده است. در نهایت براساس میزان تأثیر پارامترها، نُه ویژگی با بالاترین تأثیر شامل {تعداد واحد، روز، مصرف یک ماه قبل، مصرف دو ماه قبل، کد قطر فاضلاب، مصرف سه ماه قبل، قطر آب، فشار، مرغوبیت مکانی} انتخاب شده‌اند.

لازم به ذکر است که ویژگی روز به صورت تجمیع شده نشان‌دهنده‌ی تعداد روزهای مربوط به هر رکورد (در طول دوره یک‌ماهه مربوط به آن رکورد) است و روزهای تعطیل، جمعه و کاری در آنها مجزا نشده‌اند. زیرا در این بخش هدف تنها بررسی میزان اثر تعداد روز بر مصرف بوده است. در پایان باید متذکر شد که دادگان تولید آب روزانه به صورت مستقیم در پیش‌بینی مصرف ماه آینده مورد استفاده قرار نگرفته‌اند و از آنها تنها به عنوان ویژگی تصمیم‌یار در تجمیع و انتخاب دادگان هواشناسی بهره گرفته شده است. به علاوه، از میان سوابق مصرف ماه‌های گذشته، سوابق سه ماه قبل از ماه هدف دارای اثر محسوس بر روی مقادیر مصرف ماه هدف هستند که در جریان انتخاب ویژگی انتخاب شده و مورد استفاده قرار می‌گیرند. البته از کل سوابق مشترکین در هنگام ساخت مدل انتخاب ویژگی بهره گرفته شده است، ولی با توجه به میزان تأثیر، پیش‌بینی مقدار مصرف در ماه هدف در مدل نهایی تنها بر اساس سابقه سه ماهه می‌باشد.

پیش‌بینی میزان مصرف ماهانه

این بخش، فرایند پیش‌بینی مصرف آب را معرفی می‌کند. در نهایت با نمایش نتایج روش‌های مورد استفاده در پژوهش‌های قبلی و تحلیل آنها، مراحل را تا پیشنهاد یک روش مناسب برای حل مسئله‌ی فوق که ابعادی بسیار بزرگتر در مقایسه با نمونه‌های قبلی دارد ادامه می‌دهیم. امّا پیش از آن، رویه و معیارهای ارزیابی موردنیاز برای مقایسه این روش‌ها در زیربخش آتی معرفی خواهند شد و سپس با استفاده از آنها به تحلیل روش‌ها پرداخته خواهد شد.

رویه و معیارهای ارزیابی

با توجه به نوع مسئله مورد بحث که به تخمین مقدار یک پارامتر حقیقی می‌پردازد، متداول‌ترین معیارهای ارزیابی عبارتند از:

- میانگین قدرمطلق خطا[4]
- جذر میانگین مربعات خطا[5]
- میانگین قدرمطلق درصد خطا[6]

دو معیار اوّل در تمام مقالات و پژوهش‌های مشابه مورد بهره‌برداری قرار گرفته‌اند، ولی در این گزارش با توجه به حساسیت و دقت موردنیاز برای تخمین مقادیر، علاوه بر این دو معیار، معیار آخر نیز افزوده شده است که می‌تواند نشان‌دهنده‌ی مقیاس خطای پیش‌بینی در مقابل میزان واقعی مصرف آب باشد. نحوه‌ی محاسبه هر یک از این سه معیار در روابط (۳) تا (۶) قابل مشاهده است.

[1] Dummy variables
[2] Random Foreset Regressor (RFR)
[3] Bootstrap
[4] Mean Absolute Error (MAE)
[5] Root Mean Square Error (RMSE)
[6] Mean Absolute Percentage Error (MAPE)

جدول ۳: رتبه‌ی و میزان تاثیر هر یک از ویژگی‌های جدول مصرف به علاوه فشار آب و مرغوبیت ملکی در تخمین مقدار مصرف آب در کل محدوده‌ی شهری یزد با استفاده از روش تخمین‌زننده‌ی جنگل تصادفی .

#	ویژگی (نام فارسی)	قدرمطلق میزان تأثیر	رتبه	#	ویژگی	قدرمطلق میزان تأثیر	رتبه
۱	اقساط فاضلاب	۰/۰۰۰۰۰۴	۱۹	۱۲	کد قطر فاضلاب	۰/۳۹۷۴۸	۵
۲	بدهی اقساط فاضلاب	۰/۰۰۰۰۰۱	۲۱	۱۳	کد کاربری مصرف	۰/۰۰۶۹۱۷	۱۵
۳	بدهی فاضلاب	۰/۰۰۰۰۰۲	۲۰	۱۴	مرغوبیت مکانی	۰/۱۲۳۹۶۹	۹
۴	بدهی قبلی صورتحساب	۰/۰۰۰۰۰۶	۱۸	۱۵	مصرف ۱ ماه قبل	۰/۵۹۳۹۴۹	۳
۵	تعداد خانوار	۰/۰۰۰۱۹۷	۱۷	۱۶	مصرف ۲ ماه قبل	۰/۵۳۹۴۴۴	۴
۶	تعداد واحد	۱۱/۸۲۹۹۹	۱	۱۷	مصرف ۳ ماه قبل	۰/۲۷۷۱۵۹	۶
۷	روز	۱/۹۴۷۳۹۷	۲	۱۸	مصرف ۴ ماه قبل	۰/۰۷۹۸۷۴	۱۰
۸	ظرفیت مقطوع آب	۰/۰۰۹۲۶۵	۱۶	۱۹	مصرف ۵ ماه قبل	۰/۰۷۶۹۶۱	۱۲
۹	ظرفیت مقطوع فاضلاب	۰/۰۶۷۷۳۲	۱۳	۲۰	مصرف ۶ ماه قبل	۰/۰۶۰۴۳۱	۱۴
۱۰	فشار	۰/۱۳۶۴۶۲	۸	۲۱	میانگین مصرف بازرسی	۰/۰۷۹۸۳۴	۱۱
۱۱	کد قطر آب	۰/۲۳۸۳۰۹	۷	۲۲			

$$MAE = \frac{\sum_{i=1}^{n}|A_i - P_i|}{n} \quad (4)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(A_i - P_i)^2}{n}} \quad (5)$$

$$MAPE = 100 \times \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right| \quad (6)$$

در همه‌ی این روابط، متغیرهای A_i و P_i به ترتیب مقدار مصرف واقعی و مقدار مصرف پیش‌بینی شده هستند و در نهایت، متغیر n تعداد

قابل مشاهده است. برای مواردی مانند شبکه‌های عصبی انواع پیکربندی‌ها انتخاب و مورد آزمون قرار گرفته و در نهایت بهترین موارد قابل مشاهده است.

با توجه به توضیحات در بخش مقدمه، از بین مدل‌های تخمین‌زننده و شبکه‌های عصبی به عنوان دو مجموعه‌ی پرکاربرد در به روال معرفی شده در مقالات مرجع عمل شده است. برای انتخاب پارامتر مدل‌های تخمین‌زننده (بدون مرجع) از روش آزمون و خطای مقادیر مختلف و انتخاب بهترین پارامتر با کمترین خطا بهره گرفته شده است.

در میان این مدل‌ها، بهترین نتایج به ترتیب متعلق به مدل‌های تخمین‌زننده‌ی جنگل تصادفی[2]، تخمین‌زننده‌ی چندک[3] و تخمین زننده‌ی هوبر[4] هستند. وجه مشترک همه‌ی این مدل‌ها توانایی آنها در مقاومت برابر موارد ناهنجار و یا مواردی است که با الگوهایی متفاوت منعکس شده نیانجامیده است.

رکوردهای مورد استفاده را نشان می‌دهد. فرایند آزمایش‌ها بر اساس روش اعتبارسنجی k-بخشی[1] که در آنها k برابر ۱۰ است به اجرا در می‌آیند. برای روش‌های مبتنی بر شبکه‌ی عصبی مجموعه‌ی ارزیابی از محل یک پنجم مجموعه‌ی آموزش انتخاب می‌شوند.

روش‌های مورد مقایسه

از مجموعه‌ی پژوهش‌هایی که در بخش ابتدایی این مقاله به آنها اشاره شد، از هر خانواده از روش‌ها تعدادی انتخاب شده و سپس بنا به مقتضیات روش، هر یک آموزش داده شده‌اند که خلاصه‌ی آن در

در این جدول آورده شده‌اند. با آزمایش هر یک از مدل‌های پیشنهادی، مقادیر خطای آنها تعیین شده که در

این حوزه، مدل‌ها با بهترین کارایی انتخاب شده و در این پژوهش برای مقایسه بکار گرفته شده‌اند. در تعیین پارامتر مدل‌های دارای ارجاع در

با سایرین به پیش‌بینی می‌پردازند. بهترین نتیجه در این میان به روش جنگل تصادفی اختصاص دارد. این روش می‌تواند با دسته‌بندی موارد ناهنجار یا متفاوت و آموزش یک یا چند درخت ویژه‌ی آنها، دقت خود را بالا ببرد؛ به عبارت بهتر، این روش موارد ناهنجار را به رسمیت شناخته و تلاش می‌کند برای آنها نیز مدل ویژه‌ای ایجاد کند. این روش با ۱۰۰ درخت، بهترین نتیجه را ارائه نموده است و تلاش‌ها برای بهبود آن با استفاده از افزایش تعداد درختان مورد استفاده و پارامترهای عددی (بدون بیش‌برازش[5]) به نتیجه‌ای بهتر از آنچه در

[1] K-fold cross validation
[2] Random Forest Regressor
[3] Quantile Regressor
[4] Huber Regressor
[5] Overfitting

جدول ٤: فهرست روش‌های انتخابی همراه با ویژگی‌های اجمالی.

#	مدل	احتمالی	نوع	توضیحات و سایر پارامترها
۱	OLS Regression (16)(45)	☒	Numerical	--
۲	Bayesian Regression (16)(45)	☑	Stochastical	--
۳	MLP Neural Regressor (16)(45)	☑	ANN	Activation:Relu | weight_opt: Adam | L=(30,5,1)
۴	MLP Neural Regressor (16)(45)	☑	ANN	Activation:Relu | weight_opt: Adam | L=(30,4,1)
۵	MLP Neural Regressor (16)(45)	☑	ANN	Activation:Relu | weight_opt: Adam | L=(30,4,2,1)
۶	MLP Neural Regressor (16)(45)	☑	ANN	Activation:Relu | weight_opt: Adam | L=(30,5,2,1)
۷	MLP Neural Regressor (16)(45)	☑	ANN	Activation: tanh | weight_opt: Adam | L=(30,5,1)
۸	MLP Neural Regressor (16)(45)	☑	ANN	Activation: logistic | weight_opt: Adam | L=(30,5,1)
۹	LSTM Kernel Regressor (45)	☑	ANN	L=(30, 5,1)
۱۰	ELM Regressor (16)(45)	☑	ANN	Activation: tanh | Regressor: LinearRegression | L=(30,4,1)
۱۱	ELM Regressor (16)(45)	☑	ANN	Activation: tanh | Regressor: SGDRegressor | L=(30,4,1)
۱۲	SGD Regressor (23)	☑	ANN	penalty=L^2 | tolerance=1e-2
۱۳	LassoCV	☑	Stochastical	Alpha= automatic
۱۴	ElasticNet	☒	Numerical	--
۱۵	Huber Regressor	☒	Numerical	alpha=0.0001 | epsilon=1.1
۱۶	Random Forest Regressor (45)	☑	Stochastical	Estimator=100 | BTstrap: True
۱۷	RANSAC Regressor	☑	Stochastical	--
۱۸	Quantile Regressor	☒	Numerical	Q=0.5

در نقطه‌ی مقابل جنگل تصادفی، دو روش دیگر نه تنها موارد ناهنجار را به رسمیت نمی‌شناسند بلکه آنها را از دامنه حذف نموده و یا برای آنها پنالتی بزرگتری را در نظر می‌گیرند. روش هوبر از یک تابع خطا با عنوان هوبر بهره می‌گیرد که برای اختلاف‌های کم بین مقادیر پیش‌بینی شده و مقادیر واقعی از امتیاز پنالتی کمتر و برای موارد با اختلاف بالاتر، پنالتی‌های بزرگتر در نظر می‌گیرد و با توجه به اینکه هدف کمینه نمودن میزان پنالتی است، موارد با اختلاف زیاد از مدل کنار گذاشته می‌شوند (۴۲).

این روش آموزش چند مرحله‌ای و تکراری است و یکی از مهم‌ترین مواردی که باید به آن توجه نمود درصد نمونه‌های ناهنجاری است که به این مدل وارد می‌شود. زیرا در صورتی که درصد این نمونه‌ها زیاد باشند و یا میزان ناهنجاری (تفاوت آنها با نمونه‌های معمول و در نتیجه اختلاف مقادیر آنها با مقادیر خروجی مدل) بیشتر باشد، احتمال بروز اشکال برای مدل بالاتر خواهد رفت (۴۳). در این راستا، انواع مقادیر اپسیلون برای این مدل مورد آزمون قرار گرفت. اپسیلون پارامتری است که میزان حساسیت این مدل را نسبت به موارد ناهنجار تعیین می‌کند و هر چه پایین‌تر باشد، مقاومت مدل در برابر موارد ناهنجار بیشتر است و به دنبال آن تعداد نمونه‌هایی که از فرایند آموزش کنار گذاشته می‌شوند هم بیشتر خواهد بود.

کمینه‌ی مقدار اپسیلون برابر ۱/۰ است که بیشترین مقاومت را ایجاد می‌کند، ولی با تعیین این مقدار، میزان خطاهای خروجی از مقدار خطاهای بدست آمده برای مقدار اپسیلون ۱/۱ به عنوان بهترین گزینه کمتر نیست و بنابراین با استفاده از این روش نمی‌توان به نتیجه‌ای بهتر دست یافت. روش بعدی که به عنوان آخرین گزینه با کمترین میزان خطا باید مورد بررسی قرار بگیرد، روش تخمین‌زننده‌ی چندک[1] است. روش تخمین‌زننده‌ی چندک یک تفاوت عمده با روش‌های تخمین‌زننده‌ی پیشین دارد. بیشتر روش‌هایی که تلاش می‌کنند خطای مینیمم مربعات یا خطاهای مشابه را حداقل نمایند در واقع تلاش می‌کنند خطی را به عنوان خط تخمین‌زننده محاسبه کنند که به میانگین مقادیر در هر نقطه نزدیک‌تر است.

این روش‌ها در صورتی که موارد ناهنجار با مقادیر خیلی بزرگ یا خیلی کوچک در دادگان موجود باشند بسیار ضعیف عمل می‌کند؛ زیرا این مقادیر می‌توانند مانند وزنه‌هایی باعث تغییر در شیب خط تخمینی شوند. امّا روش تخمین‌زننده‌ی چندک چنان که از نام آن نیز مشخص

[1] Quantile Regressor

است، توجه خود را معطوف به مقادیر یکی از چندک‌های تعیین شده می‌کند.

یعنی تلاش می‌کند خط تخمینی از محل چندک تعیین شده، عبور کند. با بررسی میزان خطای تخمین‌زننده‌های چندک مختلف مشخص گردید، بیشترین اختلاف در میزان خطا بین چندک‌های پایینی و همینطور چندک‌های بالایی بروز می‌کنند.

بدین معنا که دادگانی که در چندک‌های پایین (کمتر از ۱۰٪) و بالا (بیشتر از ۹۰٪) واقع می‌شوند بیشترین خطا را به مسئله تحمیل می‌کنند. بنابراین در نخستین گام، این رکوردها از سایرین جدا شده و مورد بررسی قرار گرفتند. نتایج نشان داد این رکوردها عمدتاً رکوردهایی هستند که سابقه‌ی مصرف آنها نسبت به مصرف ماه آتی (یعنی ماه هدف برای پیش‌بینی) اختلاف بیشتری دارد. به عنوان نمونه، در یک مورد برای یک مشترک در مدت شش ماه مصرف صفر و در ماه هفتم مقدار مصرف ۲۱/۴ مترمکعب ثبت گردیده است. در این حالت، میزان اختلاف بین مقادیر ثبت شده برای ماه هفتم در مقایسه با سه ماهه‌ی گذشته که به عنوان سابقه‌ی مصرف مورد استفاده قرار می‌گیرد می‌تواند به خطای قابل توجهی در مدل منجر شود.

بدین معنا که از مشترکینی با میانگین مصرف سه ماهه‌ی مشخص انتظار می‌رود در ماه هدف هم میزان مصرفی متناسب با میانگین سه ماهه داشته باشند، نه خیلی کمتر و نه خیلی بیشتر. به علاوه مشخص گردید رفتار مشترکین هم کاملاً یکدست نیست و مشترکین کم مصرف و پرمصرف الگوی رفتاری متفاوتی دارند. با توجه به نکات بدست آمده در این بررسی یک چارچوب پیش‌بینی طراحی گردید که با غلبه بر این مشکلات، خطای پیش‌بینی را کاهش می‌دهد. جزئیات طراحی و پیاده‌سازی این چارچوب در بخش آتی معرفی خواهد شد.

روش پیشنهادی

با نظر به نتایج به دست آمده از مرحله‌ی بررسی رفتار مشترکین و نتایج مدل‌های به کار گرفته شده، به نظر می‌رسد استفاده از یک مدل تخمین‌زننده‌ی چندک که در آن گروه‌های مشترکین با بازه‌ی مصرف متفاوت تفکیک شده باشند بهترین انتخاب خواهد بود.

جدول ۵: فهرست روش‌های انتخابی همراه با ویژگی‌های آماری و میزان خطای آنها در تخمین میزان مصرف.

#	مدل	MAE	RMSE	MAPE (٪)	توضیحات و سایر پارامترها
۱	OLS[1] Regression	۲/۰۷۸	۵/۳۵۸	۱۱/۷۴۰	--
۲	Bayesian Regression	۲/۰۷۷	۵/۳۵۷	۱۱/۷۳۹	--
۳	MLP[2] Neural Regressor	۲/۰۵۳	۴/۸۱۷	۱۱/۷۹۹	Activation:Relu | weight_opt: Adam | L=(30,5,1)
۴	MLP Neural Regressor	۲/۱۰۶	۵/۳۰۰	۱۲/۶۶۳	Activation:Relu | weight_opt: Adam | L=(30,4,1)
۵	MLP Neural Regressor	۲/۴۵۲	۵/۴۷۸	۱۲/۷۸۶	Activation:Relu | weight_opt: Adam | L=(30,4,2,1)
۶	MLP Neural Regressor	۵۸/۵۱۴	۳۴/۴۷۵	۱۰/۰۸۰	Activation:Relu | weight_opt: Adam | L=(30,5,2,1)
۷	MLP Neural Regressor	۸/۶۸۷	۱۹/۶۲۱	۵۳/۲۵۸	Activation: tanh | weight_opt: Adam | L=(30,5,1)
۸	MLP Neural Regressor	۷/۷۱۲	۳۳/۲۰۰	۴۴/۷۲۸	Activation: logistic | weight_opt: Adam | L=(30,5,1)
۹	LSTM Kernel Regressor	۲/۰۳	۴/۷۸	۱۱/۴۰	L=(30, 5,1)
۱۰	ELM Regressor	۱۰/۱۰۷	۳۳/۷۱۴	۵۸/۶۷۲	Activation: tanh | Regressor: LinearRegression | L=(30,4,1)
۱۱	ELM Regressor	۱۰/۰۶۵	۳۴/۳۰۴	۵۷/۶۲۳	Activation: tanh | Regressor: SGDRegressor | L=(30,4,1)
۱۲	SGD Regressor	۲/۱۶۴	۵/۵۷۴	۱۲/۳۷۵	penalty=L2 | tolerance=1e-2
۱۳	LassoCV	۲/۱۰۹	۵/۴۶۷	۱۱/۹۶۸	Alpha= automatic
۱۴	ElasticNet	۲/۲۲۴	۵/۵۹۱	۱۲/۳۳۳	--
۱۵	Huber Regressor	۱/۹۴۰	۵/۳۷۸	۱۰/۳۷۹	alpha=0.0001 | epsilon=1.1
۱۶	Random Forest Regressor	۱/۶۰۱	۴/۸۶۴	۹/۲۸۰	Estimator=100 | BTstrap: True
۱۷	RANSAC Regressor	۲/۱۴۶	۶/۳۰۸	۱۱/۲۶۲	--
۱۸	Quantile Regressor	۱/۹۳۷	۵/۴۰۵	۱۰/۲۹۰	Q=0.5

[1] Ordinary Least Squares (OLS)

[2] Multi-Layer Perceptron

بر همین اساس چارچوبی طراحی گردید که از محاسن روش تخمین‌زننده‌ی چندک و دسته‌بندی مشترکین بر اساس مصرف به صورت توأم بهره ببرد. مراحل طراحی روش پیشنهادی به صورت خلاصه به شرح زیر است:

- تعیین نقاط برش با استفاده از تخمین‌زننده‌ی قطعه‌ای
- تعیین چندک‌های مرز ناهنجاری
- حذف داده‌های بالاتر/پایینتر از چندک‌های مرز ناهنجاری
- آموزش مجدد مدل بر روی داده‌های باقیمانده در هر یک از بازه‌ها

با استفاده از روش استخراج نقاط برش (۴۴)، محل‌های تغییرات الگوی مصرف مشترکین مشخص گردیده‌اند که به ترتیب برابر ۱۰، ۱۰۰ و ۱۰۰۰ مترمکعب بوده‌اند که بازه‌های چهارگانه‌ی [۰,۱۰]، [۱۰,۱۰۰]، [۱۰۰,۱۰۰۰] و [۱۰۰۰,۲۰۰۰] را تشکیل می‌دهند. این مقادیر برش با استفاده از آزمون تمام مقادیر از صفر تا ۲۰۰۰ و به صورت خودکار به دست آمده‌اند، ولی وجود نظم در بین مقادیر به معنای وجود یک توزیع نمایی نیست. بعد از تعیین نقاط برش باید به سراغ تعیین آن دسته از مشترکین برویم که میزان مصرف ماه هدف در آنها با میانگین مصرف سه ماهه‌ی آنها همخوانی ندارد. ابتدا باید مشاهده نمود عرفِ جامعه‌ی آماری مشترکین از چه الگویی پیروی می‌کند و سپس رکوردهایی را که از این عرف کلی پیروی نمی‌کنند در جریان آموزش کنار گذاشت تا مانع از بروز خطا به دلیل اثر آنها بر کل جامعه گردد. برای این منظور در هر یک از مجموعه رکوردهای متعلق به بازه‌های چهارگانه و برای تمام مقادیر میانگین مصرف سه ماهه، تعداد محدودی از چندک‌های بالا و پایین از نظر میزان مصرف در ماه هدف به عنوان چندک‌های خارج از عرف تعیین می‌شوند.

باید به این نکته توجه نمود که هر چه تعداد چندک‌های انتخاب شده برای حذف بیشتر باشد، رکوردهای بیشتری نیز حذف خواهند شد و بالتبع از جامعه‌ی پشتیبانی تابع تخمین‌زننده و دقت آن خواهند کاست. به همین دلیل از هر دو سوی جامعه، حداکثر ۱۰ چندک انتخاب می‌شوند که البته ممکن است برای برخی بازه‌های چهارگانه کمتر از ۱۰ چندک نیز باشد.

برای درک فرایند تعیین چندک‌ها فرض کنید، اگر برای مشترکینی که میزان میانگین مصرف سه ماهه‌ی آنها بین صفر تا یک مترمکعب است، چندکِ ۱ درصد برابر ۰/۰۲ و چندکِ ۹۹ درصد برابر ۱/۵ باشد، یعنی پایین‌ترین یک درصد از این جامعه در ماه هدف میزان مصرفی کمتر از ۰/۰۲ مترمکعب و بالاترین یک درصد از همین جامعه میزان مصرفی بیش از ۱/۵ مترمکعب دارند.

برای تعیین مرز ناهنجاری، پس از استخراج چندک‌ها باید همه جفت‌های ممکن از آنها را به نوبت مورد آزمون قرار داد. این فرایند با حذف تمام رکوردهای خارج از چندک‌ها اجرا می‌شود و سپس با استفاده از میزان خطای تابع تخمین‌زننده، بهترین جفت از چندک‌های بالا و پایین برای هر یک از بازه‌های چهارگانه انتخاب می‌شوند. با احتساب وجود چهار بازه، ده چندک بالا و ده چندک در پایین، مجموعاً ۴۰۰ حالت مختلف باید مورد آزمون قرار گیرند. فرایند استخراج چندک‌ها و سپس انتخاب بهترین جفت چندک‌ها بر روی هر یک از بازه‌ها در قالب شبه‌کد توابع ۱ و ۲ به ترتیب در شکل ۳ و شکل ۴ قابل مشاهده هستند (این شبه‌کدها در بخش ضمیمه به صورت جزئی‌تر توصیف شده‌اند) و فهرست این چندک‌ها برای بازه‌های مختلف در

- کوچک بودن اختلاف بین میانگین قدرمطلق خطای روش پیشنهادی با جنگل تصادفی
- به رسمیت نشناختن موارد ناهنجار و حذف موارد تغییرات ناگهانی در الگوی مصرف
- عملکرد بهتر در میانگین مربعات خطا و میانگین قدرمطلق درصد خطا

نتایج ارزیابی‌ها به وضوح برتری روش پیشنهادی بر روش‌های مورد آزمون را نشان می‌دهند. طبق جدول ۶ درصد خطا در هر چهار بازه کمتر از ۱۰٪ بوده است. با توجه به اینکه با حذف دادگان چندک‌های بالا و پایین، دقت روش تخمین‌زننده‌ی چندک نیز افزایش یافت، می‌توان نتیجه گرفت مهم‌ترین علت خطای تخمین‌زننده‌های مختلف وجود این رکوردها بوده‌اند.

جدول ٦: چندک‌های بالا و پایین برای تعیین مرز ناهنجاری در هر یک از بازه‌های تصمیم‌گیری.

نمایش داده شده‌اند. در کنار چندک‌ها در این جدول، کسری از جمعیت هر گروه که حذف شده‌اند نیز مشخص شده است که بیشترین کسر مربوط به بازه‌ی اول با حدود ۱۶/۸ درصد و کمترین مربوط به بازه‌ی دوم با ۱۲ درصد بوده است. بعد از تعیین چندک‌ها و حذف رکوردهای مربوطه، یک بار دیگر تخمین‌زننده آموزش داده شده و میزان خطای آن به صورت کلی و بر روی هر یک از چهار بازه اندازه‌گیری می‌شود (ر.ک. **Error! Reference source not found.**).

نتایج حاکی از آن است که میزان خطای چارچوب پیشنهادی برای همه‌ی معیارها، از سایر روش‌ها به جز روش مبتنی بر تخمین‌زننده‌ی جنگل تصادفی کمتر است و در مقایسه با جنگل تصادفی نیز تنها در معیار میانگین قدرمطلق خطاها این روش به اندازه‌ی جنگل تصادفی خوب عمل نکرده است. ولی باید به سه نکته توجه نمود:

جدول ٦: چندک‌های بالا و پایین برای تعیین مرز ناهنجاری در هر یک از بازه‌های تصمیم‌گیری.

#	بازه	چندک پایین (٪)	چندک بالا (٪)	رکوردهای حذف شده (٪)
۱	[0, 10]	۱۰	۹۰	۱۶/۸۳
۲	[10, 100]	۶	۹۴	۱۲/۰۰
۳	[100, 1000]	۶	۹۳	۱۳/۵۷

۴	[1000, 50000]	۵	۹۴	۱۲/۴۱

Function 1: Generate Quantiles Table.

Input: Breakpoint Lower Bounds, L ={0,10,100,1000}
Breakpoint Upper Bounds, U ={10,100,1000, +∞}
Lower Bound Quantiles, Q_L ={0.01, 0.02, …, 0.1}
Upper Bound Quantiles, Q_U ={0.90, 0.91, …, 0.99}
Records DataFrame, *Data*
Output: Quantile Boundary Table, *q_table*

```
1  function quantile_table_generation (.)
2    q_table=[ ]
3    for  l, u in zip(L,U) :
4      for i in range(l,u) :
5        for q1 in QL :
6          for q2 in QU :
7            slc_rec=
             Data[(Data['Mean_Consumption'] ≥ i) &
             (Data['Mean_Consumption ']< (i+1) )]
8            qL= np.quantile(slc_rec['Consumption'],q1)
9            qU= np.quantile(slc_rec['Consumption'],q2)
10           q_table.append([i, i+1, q1, q2, qL, qU])
11   return q_table
```

شکل ۳: شبه‌کد تابع تولیدکننده‌ی جدول چندک‌ها.

این رکوردها به دنبال بروز دو حالت تولید می‌شوند که عبارتند از:

- افزایش مصرف ماه هدف به دنبال مصرف کم در سه ماه قبل
- کاهش مصرف ماه هدف به دنبال مصرف بالا در سه ماه قبل

این دو حالت معمولاً به دنبال تغییر در ترکیب ساکنین املاک شهری بروز می‌کند ولی با توجه به اینکه اطلاعات مصرف ماه‌های قبل به مرور بروز می‌شوند، با گذشت زمان الگوهای مصرف ساکنین جدید فرصت بروز می‌یابند و با دخیل شدن این سوابق، دقت روش مجدداً افزایش می‌یابد.

مهمترین نتیجه‌ی دیگری که از تحلیل ساختار روش پیشنهادی بدست می‌آید، دسته‌بندی مشترکین براساس الگوی مصرف آنها است. وجود خطوط برش در بازه‌های مصرف ۱۰، ۱۰۰ و ۱۰۰۰ مترمکعب نشان می‌دهند که می‌توان مشترکین را براساس سابقه‌ی مصرف سه ماه‌ی گذشته‌ی آنها در چهاردسته با مرزهای فوق دسته‌بندی نمود. مهمترین کاربرد این دسته‌بندی استفاده از آن در سیاست‌گذاری مالی و تعیین تعرفه‌ی گروه‌های مصرفی مختلف است.

بحث و نتیجه‌گیری

در این مقاله، مسئله‌ی پیش‌بینی مصرف مشترکین شهری در شهر یزد مورد بررسی قرار گرفته است. برای حل این مسئله از دادگان مصرف مشترکین، مستخرج از صورتحساب‌های آنها و دادگان دیگری از قبیل دادگان هواشناسی، تقویم کاری و تولید آب روزانه در کل شبکه‌ی آب شهری یزد بهره گرفته شده است. بخش اوّل این مقاله به معرفی فرایندهای استخراج و پیش‌پردازش دادگان پرداخته و بخش دوم به ارائه‌ی یک راهکار عملی برای این مسئله اختصاص یافته است. در فرایند طراحی روش پیشنهادی ابتدا نتایج بدست آمده از روش‌های مورد استفاده در تحقیقات مشابه مورد تحلیل قرار گرفته و با مشخص شدن دلیل بروز خطا و موارد ناهنجار در رکوردها، روش پیشنهادی با خاصیت غلبه بر این مشکل طراحی گردیده است.

Function 2: Select Best Quantiles.

Input: Breakpoint Lower Bounds, L ={0,10,100,1000}
Breakpoint Upper Bounds, U ={10,100,1000, +∞}
Lower Bound Quantiles, Q_L ={0.01, 0.02, …, 0.1}
Upper Bound Quantiles, Q_U ={0.90, 0.91, …, 0.99}
Records DataFrame, *Data*
Quantile Boundary Table, *q_table*
Output: Best Quantiles Table *best_q*

```
1   function best_quantile_selection (.)
2     best_q=[ ]
3     for l, u in zip(L,U) :
4       for q1 in QL :
5         for q2 in QU :
6           dp=Data[(Data['Mean_Consumption'] ≥ l)
7            & (Data['Mean_Consumption ']<u)]
8           for row in range(q_table.shape[0]):
9             lx, ux, q, qq, ly, uy=q_table[row,:]
10            if q1==q and q2==qq:
11              data_part.drop(dp[
12              (dp['Mean_Consumption'] ≥ lx)
13              & (dp['Mean_Consumption']<ux)
14              & (dp['Consumption'] < ly)
15              & (dp['Consumption']≥uy)].index,
16              inplace=True)
17          error =Evaluate(Regressor(dp,'Consumption')
18          E.append([error, q1, q2])
19      best_q.append(l, u, E[Argmin (E[:,0]),1:2])
20    return best_q
```

شکل ۴: شبه‌کد تابع انتخاب بهترین چندک‌های بالا و پایین.

روش پیشنهادی با ترکیب دو روش تخمین‌زننده‌ی چندک و تخمین‌زننده‌ی قطعه‌ای تدوین شده است و تنها براساس چندک‌های برگزیده به پالایش مقادیر مصرف پرداخته و برای هر بخش از مشترکین با مقادیر مصرف متفاوت، تابع تخمین جدیدی را ارائه می‌دهد. این روش در ارزیابی‌ها به دقت بالاتری نسبت به سایر روش‌ها دست یافته و با خطای کمتر از ۱۰٪ توانسته خود را به عنوان نامزد قابل اعتمادی برای استفاده‌ی عملی در سیستم‌های امور مشترکین در این حوزه مطرح کند. این تحقیق در مقایسه با موارد مشابه در داخل و خارج از ایران دارای برتری دیگری نیز هست که به ابعاد مسئله باز می‌گردد. بر اساس اطلاعات نویسندگان در مقطع حاضر، تعداد مشترکین و رکوردهای مورد استفاده برای آموزش مدل پیشنهادی در این مقاله بیش از تمام نمونه‌های مشابه داخلی و جهانی است. در موارد مشابه، به ندرت تعداد مشترکین مورد بررسی به بیش از ۵۰۰ مورد می‌رسد در حالی که در این تحقیق، اطلاعات بیش از ۷۶۰۰۰ مشترک مورد استفاده قرار گرفته‌اند. دستاورد دیگری که این چارچوب در اختیار می‌گذارد، ارائه گروه‌بندی‌هایی برای مشترکین براساس سابقه‌ی مصرف آنها است که می‌تواند در جریان تدوین سیاست‌های تعرفه‌گذاری

مورد استفاده قرار بگیرد. در پایان باید خاطرنشان نمود که این چارچوب علاوه بر یک پژوهش علمی، یک دستاورد عملی در مقیاس سیستم‌های مبتنی بر پردازش کلان‌داده‌ها است که زمینه‌ی استفاده‌ی عملی از چارچوب‌های یادگیری ماشین را در مدیریت منابع آبی کشور فراهم خواهد کرد.

پیشنهادها

در این تحقیق با توجه به محدودیت منابع و ضروریات طرح پژوهشی تعریف شده از سوی شرکت آب و فاضلاب شهر یزد که باید دادگان لازم برای تحلیل را فراهم می‌نموده است، مواردی مانند فشار لحظه‌ای درب ملک، کیفیت آب، تاریخ دقیق اعمال سیاست‌های محاسبه آب‌بها، تبلیغات شهری برای صرفه‌جویی و تخفیفات برای مصرف کمتر در نظر گرفته نشده است. پیشنهاد می‌گردد برای بررسی اثر پارامترهایی مانند فشار آب درب ملک و مواردی که می‌توانند به تفکیک هر مشترک متفاوت باشند از اطلاعات سابقه‌ی مصرف ماهانه مشترکین بهره گرفته شود؛ ولی برای متغیرهای سراسری مانند نرخ آب‌بها، تعرفه تخفیف کاهش مصرف و تغییر کیفیت آب که بر کل یک محدوده‌ی شهری اثرگذار هستند از تغییرات مصرف کلی آب در محدوده‌ی شهری (در این پژوهش، آب تولیدی روزانه در شهر یزد) استفاده نمود. به عنوان نمونه، در شهر یزد که قطع آب انتقالی از اصفهان به وضوح می‌تواند بر طعم آب آشامیدنی موثر باشد (در صورت قطع این جریان، آب آشامیدنی شهری از محل چاه‌های عمیق تأمین می‌گردد)، پارامتر کیفیت آب می‌تواند به عنوان یک عامل مهم در تخمین میزان مصرف کلی در مقیاس شهری مورد استفاده قرار بگیرد. به علاوه باید به مسئله خطای اندازه‌گیری نیز پرداخت و با استفاده از روش‌های مقاوم به خطا نیز به بررسی آن پرداخت. کاوه یزدی و ظریفزاده (۴۶) نشان داده‌اند میزان آب اندازه‌گیری شده در برخی از کنتورهای معیوب می‌تواند به صورت خطی کاهش یابد که این خطا به صورت گسترده و با گذشت زمان می‌تواند بر دقت کلی روش اثرگذار باشد.

ملاحظات اخلاقی پیروی از اصول اخلاق پژوهش

همکاری مشارکت‌کنندگان در تحقیق حاضر به صورت داوطلبانه و با رضایت آنان بوده است.

حامی مالی

این تحقیق با حمایت مالی و داده‌ای شرکت آب و فاضلاب شهری یزد به اجرا در آمده است.

مشارکت نویسندگان

طراحی و ایده‌پردازی: فاطمه کاوه یزدی، سجاد ظریف‌زاده؛
روش‌شناسی و تحلیل داده‌ها: فاطمه کاوه یزدی، سجاد ظریف‌زاده
نظارت: سجاد ظریف‌زاده
نگارش: فاطمه کاوه یزدی، سجاد ظریف‌زاده

تعارض منافع

بنابر اظهار نویسندگان، مقاله حاضر فاقد هرگونه تعارض منافع بوده است.

ضمیمه الف

شبه‌کد مندرج در شکل ۳ تابع تولیدکننده‌ی جدول چندک‌ها را توصیف می‌کند. این تابع با دریافت فهرست چندک‌های بالا و پایین، حدود بازه‌ها و مجموعه رکوردهای مصرف، فرایند استخراج چندک‌ها را به نوبت بر روی هر یک از بازه‌های چهارگانه اجرا می‌کند. برای این منظور، تمام مقادیر صحیح بین حدود بالا و پایین هر بازه مورد پیمایش قرار گرفته و رکوردهایی که میانگین مصرف سه ماهه‌ی آنها بین هر دو مقدار صحیح در بازه‌ی مذکور قرار می‌گیرند انتخاب می‌شوند. در گام بعد، مقدار چندک متناظر با هر یک از مقادیر ۰/۰۱ تا ۰/۱۰ به عنوان چندک پایین و ۰/۹ تا ۰/۹۹ به عنوان چندک بالا استخراج می‌شوند. در پایان، جدولی شامل حدود جزئی (برای تعیین بازه میانگین سه ماهه)، چندک‌های بالا و پایین و مقادیر مصرف ماه هدف متناظر با چندک‌های بالا و پایین تولید می‌شود.

شکل ۴ تابعی را نشان می‌دهد که با دریافت فهرست چندک‌ها افزون بر دادگان مشابه تابع تولیدکننده‌ی جدول چندک‌ها، همه‌ی رکوردهایی که میانگین آنها در بازه‌ی تعیین شده در جدول چندک‌ها قرار گرفته ولی مقادیر مصرف ماه هدف در آنها از میزان متناظر با چندک بالا بیشتر یا از مقدار متناظر با چندک پایین کمتر باشد را از رکوردها حذف می‌کند. پس از حذف رکوردهای با رفتار متمایز از عرفِ آماری جامعه، مجدداً تخمین‌زننده‌ی چندک مورد استفاده با داده‌های جدید آموزش داده شده و ارزیابی می‌شود. میزان خطای همه‌ی حالات در یک جدول ذخیره شده و در پایان هر دور بررسیِ تمام مقادیر چندک‌ها، بهترین چندک‌ها (متناظر با کمترین خطا) انتخاب شده و در جدولی که به این منظور در نظر گرفته شده است قرار داده می‌شوند. جدول حاصل از اجرای این تابع، نه تنها مقادیر چندک‌ها بر روی هر یک از بازه‌ها را ذخیره نموده، بلکه میزان خطای تخمین‌زننده با شرایط مذکور را نیز معین می‌کند.

References

1. Alvi, M. S. Q., Mahmood, I., Javed, F., Malik, A. W., and Sarjoughian, H., 2018. Dynamic behavioural modeling, simulation and analysis of household water consumption in an urban area: a hybrid approach, in Proceedings of the 2018 Winter Simulation Conference: 2411–2422.
2. Haque, M. M., de Souza, A. and Rahman, A., 2017. Water Demand Modelling Using Independent Component Regression

Technique, Water Resour. Manag., 31(1): 299–312.
3. Arandia, E., Ba, A., Eck, B. and McKenna, S., 2016. Tailoring Seasonal Time Series Models to Forecast Short-Term Water Demand, J. Water Resour. Plan. Manag., 142(3): 4015067.
4. Firat, M., Turan, M. E. and Yurdusev, M. A., 2010. Comparative analysis of neural network techniques for predicting water consumption time series, J. Hydrol., 384 (1-2): 46–51.
5. Tiwari, M. K., and Adamowski, J., 2013. Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models, Water Resour. Res., 49 (10): 6486–6507.
6. Yazdani, S., Abedi, S., and Abedi, S., 2014. Forecasting Models for Domestic and Agricultural Water Consumptions in Tehran Province (Case Study: Amirkabir Dam), Iran. J. Agric. Econ. Dev. Res., 45 (1): 41–48. [In Persian]
7. Mousavi, S. N. and Kavoosi-Kalashami, M., 2016, Evaluation of Seasonal, ANN, and Hybrid Models in Modeling Urban Water Consumption A Case Study of Rash City, J. Water Wastewater, 27(4): 93–98. [In Persian]
8. Mirdehghan, S., Saadatjoo, F., Mirjalili, M., 2014. Water consumption forecast and effective features: Yazd City, The first national computer engineering symposium, Tehran, Iran.
9. Ejlali, R. G., Ghasedi-Fathabadi, M., 2014, Water consumption forecast using artificial neural networks: Soufian City, National symposium of design and computation in civil engineering and architecture using high-tech methods, Maragheh, East Azarnaijan, Iran. [In Persian]
10. Tabesh, M., Dini, M. Khoshkholgh, A. J. and Zahraie, B., 2008. Estimation of Tehran Daily Water Demand Using Time Series Analysis, Iran-Water Resour. Res., 4 (2): 57–65.
11. Aram A.-R. and Agheli, L., 2012. A Hybrid Model for Forecasting Daily Urban Water Demand, J. Quant. Econ., 9 (1): 1–17.
12. Abdi-dehkordi, M., Meftah-halqi, M., Kahe, M., 2014, Determining meteorological parameters affecting urban water consumption using fuzzy clustering algorithm (Case study: Gorgan city), Sol. and Wat. Res. Consrv., 68(2): 293-301.
13. Gooshe, S., Yazdanpanah, M. J., Tabesh, M., 2008. Forecasting Tehran's Short-time Water Consumption using Artificial Neural Networks, University of Tehran School of Engineering Journal, 41 (1): 11-24. [In Persian]
14. Karimi, D., 2004. Applications of Fuzzy Logic in Forecasting Tehran's Short-time Water Consumption, Tarbiat Moddares University. [In Persian].
15. Rasifaghihi, N., Li, S. S. and Haghighat, F., 2020. Forecast of urban water consumption under the impact of climate change, Sustain. Cities Soc., 52: 101848.
16. Mouatadid, S. and Adamowski, J., 2017. Using extreme learning machines for short-term urban water demand forecasting, Urban Water J., 14 (6): 630–638.
17. Papageorgiou, E. I., Poczeta, K. and Laspidou, C., 2016. Hybrid model for water demand prediction based on fuzzy cognitive maps and artificial neural networks, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1523–1530.
18. Villarín, M. C., 2019. Methodology based on fine spatial scale and preliminary clustering to improve multivariate linear regression analysis of domestic water consumption, Appl. Geogr., 103: 22–39.
19. Dias, T. F., Kalbusch, A. and Henning, E., 2018. Factors influencing water consumption in buildings in southern Brazil, J. Clean. Prod., 184: 160–167.
20. Taghvaei, A., Pourjafar, M. R., Hossein-Abadi, M., and Riyahi-Medvar, H., 2011. Developing an Expert System Model for Predicting Annual Urban Residential Water Demand Using Artificial Neural Network (Case Study: Ilam City), Arman. Archit. Urban Dev., 4(6): 63–74.
21. Sardinha-Lourenço, A., Andrade-Campos, A., Antunes, A. and Oliveira, M. S., 2018. Increased performance in the short-term water demand forecasting through the use of a parallel adaptive

weighting strategy, J. Hydrol., 558: 392–404.
22. Ebrahim-Banihabib, M. and Mousavi-Mirkalaei, P., 2019. Extended linear and non-linear auto-regressive models for forecasting the urban water consumption of a fast-growing city in an arid region, Sustain. Cities Soc., 48: 101585.
23. Eslamian, S. A., Li, S. S. and Haghighat, F., 2016, A new multiple regression model for predictions of urban water use, Sustain. Cities Soc., 27: 419–429.
24. Goli Ejlali, R., 2018. Hybrid Artificial Neural Network-Geostatistics Model for Urban Water Consumption Prediction. A Case Study: Osku City, J. Water Wastewater, 29(5): 98–111. [In Persian]
25. Tabesh, M. and Dini, M., 2010. Forecasting Daily Urban Water Demand Using Artificial Neural Networks, A Case Study of Tehran Urban Water, J. Water Wastewater, 21 (1): 84–95.
26. Flores, J. J., López Farías, R. , Puig, V. and Rodriguez Rangel, H., 2017. Short-term demand forecast using a bank of neural network models trained using genetic algorithms for the optimal management of drinking water networks, Journal of Hydroinformatics, 19: 1–16.
27. Gato, S., Jayasuriya, N. and Roberts, P., 2007. Temperature and rainfall thresholds for base use urban water demand modelling, J. Hydrol., 337(3): 364–376.
28. Beheshti, S. , Sahebalam, A. and Nidoy, E., 2019. Structure dependent weather normalization, Energy Sci. Eng., 7(2): 338–353.
29. Donevska K. and Panov, A., 2019. Climate change impact on water supply demands: case study of the city of Skopje, Water Supply, vol. 19(7): 2172–2178.
30. Vonk, E., Cirkel, D. G. and Blokker, M., 2019. Estimating peak daily water demand under different climate change and vacation scenarios, Water (Switzerland), 11(9): 1874-1882.
31. Colace, F., De Santo, M., Greco, L. and Napoletano, P., 2015. Weighted Word Pairs for query expansion, Inf. Process. Manag., 51(1): 179–193.
32. Sadiq W. A. and Karney, B., 2005, Modeling Water Demand Considering Impact of Climate Change – a Toronto Case Study, J. Water Manag. Model, 13(1): 1-11.
33. Ouyang, Y., Wentz, E. A., Ruddell, B. L. and Harlan, S. L., 2014. A Multi-Scale Analysis of Single-Family Residential Water Use in the Phoenix Metropolitan Area, JAWRA J. Am. Water Resour. Assoc., 50(2): 448–467.
34. Taylor, B. A., 2012. Predicting normalised monthly patterns of domestic external water demand using rainfall and temperature data, Water Sci. Technol. Water Supply, 12(2): 168–178.
35. Zhang, D., Ni, G., Cong, Z. , Chen, T. and Zhang, T., 2014. Statistical interpretation of the daily variation of urban water consumption in Beijing, China, Hydrol. Sci. J., 59(1): 181–192.
36. Tiwari, M. K. and Adamowski, J. F., 2015. Medium-Term Urban Water Demand Forecasting with Limited Data Using an Ensemble Wavelet–Bootstrap Machine-Learning Approach, J. Water Resour. Plan. Manag., 141(2): 04014053.
37. Yasar, A., Bilgili, M. and Simsek, E., 2012. Water Demand Forecasting Based on Stepwise Multiple Nonlinear Regression Analysis, Arab. J. Sci. Eng., 37(8): 2333–2341.
38. Toms J. D., and Lesperance, M. L., 2003. Piecewise Regression: A Tool For Identifying Ecological Thresholds, Ecology, 84(8): 2034–2041.
39. Li, H. , Deng, X. , Kim, D.-Y., and Smith, E. P., 2014. Modeling maximum daily temperature using a varying coefficient regression model, Water Resour. Res., 50(4): 3073–3087.
40. Community-Contributors, scikit-learn 0.22.2 Documentation: LassoCV, 2020, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html#sklearn.linear_model.LassoCV.
41. Community-Contributors, scikit-learn 0.22.2 Documentation: RandomForestRegressor, 2020.
42. Hastie, T., Tibshirani, R. and Friedman, J., 2009. The elements of statistical

learning: data mining, inference and prediction, 2nd ed. Springer.
43. Sun, Q., Zhou, W.-X. and Fan, J., 2018. Adaptive Huber Regression, J. Am. Stat. Assoc., 2018(1): 1–24.
44. Wagner, A. K., Soumerai, S. B., Zhang, F. and Ross-Degnan, D., 2002. Segmented regression analysis of interrupted time series studies in medication use research, J. Clin. Pharm. Ther., 27 (4): 299–309.
45. Vijai, P., and Sivakumar, P. B., 2018. Performance comparison of techniques for water demand forecasting, Procedia Comput. Sci., 143(1): 258–266.
46. Kaveh-Yazdy, F. and Zarifzadeh, S., 2021. Water Meter Replacement Recommendation for Municipal Water Distribution Networks using Ensemble Outlier Detection Methods, J. AI Data Min., 9(4): 425–438.