Sari Agriculture Sciences and Natural Resources University

Journal of Watershed Management Research, Vol. 16, Issue 2, 2025 p: 19-34

OPEN ACCESS

Research Paper

An Evaluation of Techniques for Outlier Detection and Missing Values Imputation of Hydrological Data Series in the Zarrineh-Roud Basin, Lake Urmia

Edith Eishoeei[1], Mirhassan Miryaghoubzadeh[2], Mahdi Erfanian[3], Reza Mahboobi Esfanjani[4] and Marco Mancini[5]

1- Ph.D, Department of Watershed Management Engineering, Natural Resources Faculty, Urmia, Urmia, Iran
2- Associate Professor, Department of Watershed Management Engineering, Natural Resources Faculty, Urmia University, Urmia, Iran, (Corresponding author: m.miryaghoubzadeh@urmia.ac.ir)
3- Associate Professor, Department of Watershed Management Engineering, Natural Resources Faculty, Urmia University, Urmia, Iran
4- Professor, Department of Electrical Engineering, Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran
5- Professor, Department of Civil and Environmental Engineering, Politecnico di Milano, Milan, Italy

Received:02 February, 2025 Revised:02 April, 2025 Accepted: 11 May, 2025

Extended Abstract

Background: Accurate river flow measurements are essential for effective water resource management, flood mitigation, river conservation and restoration, and stream rehabilitation. The majority of flood control and design flow strategies in river management and restoration initiatives are derived from hydrological and hydraulic analyses based on observed river flow. Hydrological investigations are fundamentally reliant on observational statistical data, which frequently contain multiple errors. Outliers, which are defined as data points deviating significantly from the norm, can introduce substantial calculation errors. Outlier detection techniques include supervised, semi-supervised, and unsupervised approaches, which may include distribution-based, clustering-based, and density-based methods. These errors can arise from computational issues, misreporting, sampling inaccuracies, and human or instrumental errors, leading to problems such as unrecorded data, incorrect values, equipment failure or loss, and the misidentification of outliers as missing data. Consequently, the estimation and assessment of these data are essential for their application in models, and to mitigate mistakes, preprocessing must be performed before their utilization. Preprocessing methods prepare data series for computations, such as classification, prediction, and estimation, and include the elimination of missing data, removal of outliers, imputation of missing values, and data normalization.

Method: This study utilized flow and rainfall data from six hydrometeorological stations and 16 rain stations to identify outliers and impute missing or incomplete hydrological values. The data, obtained from the Zarrineh-roud basin, were implemented using R software. The Zarrineh River watershed constitutes the largest watershed of Lake Urmia. Normalization tests, including the Shapiro-Wilk and Kolmogorov-Smirnov tests, were used to normalize the data, and the findings indicated that the data did not conform to a normal distribution. Subsequent to data normalization, outlier detection was executed using approaches including boxplot, z-score, histogram, chi-square, mean and standard deviation, and median techniques. Values exceeding the established maximum were removed. Missing values were imputed using K-Nearest Neighbor (KNN), Lasso regression, and Bayesian linear regression. Lasso regression is a regularization technique designed to diminish model complexity and avoid overfitting. Bayesian linear regression is a statistical analysis method that integrates linear regression with Bayesian techniques. The KNN algorithm is a sample-based method related to nonparametric models and supervised learning classification. Cross-validation was used to assess the accuracy of the imputation methods, with RMSE and R^2 serving as performance metrics.

Result: According to the results, P-values at all six study stations were less than 0.05. The cross-validation approach was used to assess the accuracy and precision of the KNN, Lasso regression, and linear Bayesian regression techniques. RMSE values near zero and R^2 values above 0.7 across all stations indicated that KNN was a robust and accurate method for missing value imputation. It provides significantly more accurate and reliable outcomes without reshaping the data series trend than Lasso regression and Bayesian linear regression. Outliers were removed from the Jan-Agha and Darreh Pandedan stations during normalization. Histogram analysis revealed skewness

[DOI: 10.61882/jwmr.2025.1310]

and outliers at the Jan-Agha, Sariqamish, and Pol-Anyan stations, indicating a heterogeneous and non-normally distributed dataset. Outliers were identified and removed following normalization. The Shapiro-Wilk and Kolmogorov-Smirnov tests yielded p-values significantly below 0.05 after normalization, confirming a normal distribution. This suggests that the normalization process and outlier removal were executed with precision, indicating the significan detection and estimation of outliers. The Rosner test established the upper limit for each data series across two successive tests, classifying values beyond this limit as outliers. The consistency of the probability density functions between the observed and imputed values using the KNN method indicates an adequate alignment of the two probability density functions. This method has proved effective in imputing the maximum, average, and minimum values relative to the other two methods at the studied stations.

Conclusion: The results of this investigation indicate that the boxplot identifies data values outside the lines as outliers, leading to a substantial number of outliers being detected compared to the other methods. Consequently, this method is considered unsuitable for outlier detection in hydrological data. KNN proved highly effective for missing data imputation compared to Lasso regression and Bayesian linear regression. This study involved normalizing the data series, calculating the values of outliers, and employing the KNN algorithm to identify incomplete or unmeasured and missing values. In datasets exhibiting little variation, KNN has high accuracy and is regarded as one of the most valuable and dependable techniques for attributing and imputing missing values. Cross-validation confirmed the performance of KNN, Lasso regression, and Bayesian linear regression. KNN achieved R^2 values above 0.7 and RMSE values close to zero. KNN outperformed the other two methods in estimating missing values in continuous and discontinuous flow data. This effectiveness is attributed to KNN's ability to identify optimal nearest neighbor values, making it suitable for accurate predictions, even during low flow periods. The precision of KNN stems from its computational simplicity and high efficacy in calculating and imputing missing values while preserving the integrity of the data series.

Keywords: Bayesian linear regression, K Nearest Neighbor, Lasso regression, Shapiro-Wilk test, Zarrineh-roud basin

How to Cite This Article: Eishoeei, E., Miryaghoubzadeh, M., Erfanian, M., Mahboobi Esfanjani, R., & Mancini, M. (2025). An Evaluation of Techniques for Outlier Detection and Missing Values Imputation of Hydrological Data Series in the Zarrineh-Roud Basin, Lake Urmia. *J Watershed Manage Res,* 16(2), 19-34. DOI: 10.61882/ jwmr.2025.1310

[DOI: 10.61882/jwmr.2025.1310]

دانشگاه علوم کشاورزی و منابع طبیعی ساری

مقاله پژوهشی

ارزیابی روش‌های تشخیص و بازسازی مقادیر پرت و گمشده در سری داده‌های هیدرولوژیکی حوزه آبخیز زرینه‌رود، دریاچه ارومیه

ادیت عیشویی[۱]، میرحسن میریعقوب‌زاده[۲]، مهدی عرفانیان[۳]، رضا محبوبی اسفنجانی[۴] و مارکو مانچینی[۵]

۱- دانش‌آموخته دکتری، گروه علوم و مهندسی آبخیزداری، دانشکده منابع طبیعی، دانشگاه ارومیه، ارومیه، ایران
۲- دانشیار، گروه علوم و مهندسی آبخیزداری، دانشکده منابع طبیعی، دانشگاه ارومیه، ارومیه، ایران، (نویسنده مسوول: m.miryaghoubzadeh@urmia.ac.ir)
۳- دانشیار، گروه علوم و مهندسی آبخیزداری، دانشکده منابع طبیعی، دانشگاه ارومیه، ارومیه، ایران
۴- استاد گروه مهندسی برق، دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی سهند، تبریز، ایران
۵- استاد، گروه مهندسی عمران، دانشکده مهندسی محیط زیست و عمران، دانشگاه پلی‌تکنیک میلان، میلان، ایتالیا

تاریخ دریافت: ۱۴۰۳/۱۱/۱۴ تاریخ ویرایش:۱۴۰۴/۰۱/۱۳ تاریخ پذیرش: ۱۴۰۴/۰۲/۲۱
صفحه: ۱۹ تا ۳۴

چکیده مبسوط

مقدمه و هدف: اندازه‌گیری‌های جریان رودخانه و داده‌های آن در مدیریت منابع آب، کنترل سیل، حفاظت و احیای رودخانه، بازسازی جریان اهمیت بسزایی دارند. اکثر طرح‌های کنترل سیل و دبی طراحی در پروژه‌های مدیریت و احیای رودخانه توسط تحلیل‌های هیدرولوژیکی و هیدرولیکی مبتنی بر دبی مشاهداتی حوزه تخمین زده می‌شوند. پایه مطالعات هیدرولوژیکی به داده‌های آماری مشاهداتی وابسته است و این داده‌ها در اغلب موارد دارای خطاهای متعدد هستند. داده پرت داده‌ای است که از نرم طبیعی فاصله گرفته است و باعث بروز خطا در محاسبات می‌شود. روش‌های تشخیص داده‌های پرت شامل روش‌های نظارت‌شده، نیمه نظارت‌شده و نظارت‌نشده هستند و برخی روش‌های مبتنی بر توزیع، مبتنی بر خوشه‌بندی و مبتنی بر چگالی را شامل می‌شوند. به‌دلیل خطای محاسباتی، مقادیر صحیح خاص، گزارش اشتباه و یا خطای نمونه‌برداری و همچنین به دلیل خطاهای انسانی و ابزاری ممکن است مواردی مانند ثبت نشدن آمار، ثبت آمار غلط، خرابی یا ازبین رفتن دستگاه‌های اندازه‌گیری یا تشخیص داده‌های پرت و حذف آن‌ها با عنوان داده‌های گم‌شده پیش آید. بنا بر این، تخمین و برآورد این داده‌ها برای استفاده در مدل‌ها ضروری است و به منظور کاهش بروز خطا باید پیش از به کارگیری آن‌ها پیش‌پردازش صورت گیرد. عملیات پیش‌پردازش، سری داده را برای محاسبات از جمله کلاسه‌بندی، پیش‌بینی و تخمین آماده می‌کند و شامل حذف داده‌های گم‌شده، حذف داده‌های پرت، بازسازی مقادیر گم‌شده، و نرمال‌سازی داده‌ها است.

مواد و روش‌ها: در این تحقیق، به‌منظور تشخیص داده‌های پرت و بازسازی داده‌های گمشده و ناقص سری زمانی داده‌های هیدرولوژی، داده‌های دبی ماهانه شش ایستگاه هیدرومتری و داده‌های اقلیمی ۱۶ ایستگاه باران‌سنجی در حوزه زرینه‌رود در نرم‌افزار R برنامه‌نویسی و مورد بررسی قرار گرفتند. حوزه آبخیز زرینه‌رود بزرگترین حوضه آبخیز دریاچه ارومیه است. به‌منظور آزمون نرمال بودن داده‌ها از آزمون شاپیرو–ویلک و کولموگروف–اسمیرنف استفاده گردید که مطابق نتایج به‌دست آمده داده‌های مورد استفاده دارای توزیع نرمال نبودند و پس از نرمال‌سازی داده‌ها محاسبات داده‌های پرت به روش‌های نمودار جعبه‌ای، z-score، هیستوگرام، مربع کای، میانگین و انحراف معیار و روش میانه انجام شد و داده‌هایی که از بالاترین مقدار مشخص‌شده بیشتر بودند حذف گردیدند. به‌منظور نسبت‌دهی و جایگذاری مقادیر گمشده از الگوریتم‌های KNN، رگرسیون لاسو و رگرسیون خطی بیزین استفاده گردید. روش رگرسیون لاسو یک روش منظم‌سازی است که هدف آن کاهش پیچیدگی مدل و جلوگیری از بیش‌برازشی است. رگرسیون خطی بیزین نوعی تحلیل آماری است که ترکیبی از روش‌های رگرسیون خطی و بیزین را استفاده می‌کند. الگوریتم KNN یکی از روش‌های مبتنی بر نمونه است که با مدل‌های ناپارامتری و طبقه‌بندی یادگیری نظارت‌شده ارتباط دارد. برای ارزیابی دقت الگوریتم‌های نسبت‌دهی داده‌های گمشده از روش Cross Validation استفاده گردید و در ادامه، جهت محاسبه دقت روش‌های تخمین از دو معیار RMSE و R^2 استفاده شد.

یافته‌ها: نتایج آماری حاصل نشان می‌دهند که مقادیر p-value در هر شش ایستگاه مورد مطالعه کمتر از ۰/۰۵ بودند. به‌منظور ارزیابی صحت و دقت روش KNN از اعتبارسنجی متقابل استفاده گردید. مقادیر RMSE کمتر و نزدیک به صفر و R^2 بالاتر از ۰/۷ در تمامی ایستگاه‌ها نشان دادند که روش KNN یک روش مطمئن و دقیق در نسبت‌دهی و جایگذاری مقادیر گمشده بود و در مقایسه با روش رگرسیون لاسو و رگرسیون خطی بیزین نتایج بسیار دقیق‌تر و مطمئن‌تری را ارائه داد و موجب اختلال در روند سری داده نشد. مقادیر پرت ایستگاه‌های جان‌آقا و دره پنبه‌دان در ادامه و در نرمال‌سازی حذف گردیدند. چولگی و وجود داده پرت در روش هیستوگرام به ویژه ایستگاه‌های جان‌آقا، ساریقمیش و پل آدینان بابی‌نظمی همراه بودند و توزیع ناهمگن و غیر نرمال داشتند که پس از نرمال‌سازی، داده‌های پرت مشخص و حذف شدند. میزان p-value در هر دو آزمون شاپیرو–ویلک و کولموگروف–اسمیرنف مقادیری بسیار کمتر از ۰/۰۵ را نشان داد و گواه این مطلب است که داده‌ها در محدوده نرمال قرار دارند و نرمال‌سازی داده‌ها و حذف مقادیر پرت با دقت بالایی انجام شده است و درنتیجه محاسبه مقادیر پرت و شناسایی آن‌ها معنی‌دار است. آزمون روزنر برای هر سری داده مقدار حد بالا را در دو تست متوالی ارائه داده است و همان مقدار و مقادیر بالاتر از آن را به عنوان داده پرت در نظر می‌گیرد. نتایج مطابقت تابع چگالی احتمال مقادیر مشاهده‌ای و نسبت‌دهی شده به‌روش KNN نشان از تطابق قابل قبول دو تابع چگالی احتمال داشتند و این روش در نسبت‌دهی مقادیر حداکثر، متوسط و حداقل نسبت به دو روش دیگر در ایستگاه‌های مورد مطالعه موفق عمل کرد.

نتیجه‌گیری: با توجه به نتایج به‌دست آمده از نمودار جعبه‌ای، داده‌هایی که خارج از ساقه قرار گیرند را به عنوان داده پرت معرفی می‌کند و بر همین اساس در نمودارهای جعبه‌ای تعداد داده‌های پرت در مقایسه با سایر روش‌ها به مقدار زیادی تشخیص داده می‌شود که به‌نظر می‌رسد روش مناسبی برای تشخیص داده پرت در داده‌های هیدرولوژیکی نباشد. روش KNN در تعیین داده‌های گمشده با استفاده از داده‌های مشاهداتی متناظر، در بین دو روش دیگر بسیار موثر عمل نمود. در این مطالعه، سری داده‌ها نرمال‌سازی و سپس مقادیر داده‌های پرت در آن‌ها محاسبه گردید و برای تعیین مقادیر محاسبه نشده و گمشده از روش KNN استفاده شد. در داده‌های دارای روند تغییرات کمتر، KNN بسیار دقیق عمل می‌نماید و یکی از دقیق‌ترین و مطمئن‌ترین روش‌های نسبت‌دهی و جایگذاری داده‌های گمشده است. به منظور اعتباریابی روش KNN، رگرسیون لاسو و رگرسیون بیزین از روش اعتبارسنجی متقابل یا Cross Validation استفاده شد. با توجه به نتایج بدست آمده، الگوریتم KNN ضریب تبیین بالاتر از ۰/۷ و مقادیر RMSE نزدیک به صفر را نشان داد. روش KNN کارایی مطلوبی را در تخمین مقادیر گمشده در جریان‌های پیوسته و ناپیوسته نسبت به دو روش دیگر ارائه می‌دهد. این اثربخشی به توانایی KNN در دستیابی به مقدار بهینه نزدیکترین همسایه برمی‌گردد که آن‌را برای پیش‌بینی دقیق در شرایطی که جریان به حداقل رسیده باشد هم مناسب می‌سازد. دقت KNN به‌دلیل سادگی محاسبات و نیز اثر بالای آن در محاسبه و نسبت‌دهی داده‌های گمشده و گمشده است که در عین حال ساختار سری داده را نیز حفظ می‌کند.

واژه‌های کلیدی: حوضه زرینه رود، رگرسیون لاسو، رگرسیون خطی بیزین، شاپیرو–ویلک، نزدیکترین همسایه

[DOI: 10.61882/jwmr.2025.1310]

مقدمه

اساس مطالعات هیدرولوژی داده‌های آماری دقیق است که با توجه به وجود خطاهای گسسته و پیوسته در اغلب داده‌های هیدرولوژی مانند بارش و دبی رودخانه بازسازی، تخمین و برآورد داده‌های گمشده از اهمیت بسیار زیادی برخوردار است (Azimi-Habashi *et al.,* 2024; Naghdi *et al.,* 2010; Schafer & Graham, 2002). طبق تعریف، داده‌های پرت مقادیری هستند که خارج از محدوده طبیعی یک متغیر قرار دارند (Ben-Gal, 2005; Grubbs, 1969). وجود چنین داده‌هایی یک مشکل عمده در مطالعات تجربی در زمینه علوم تجربی است که در نتیجه عملکرد نادرست کاربران و به‌دلایل مختلفی از جمله ورود داده همراه با خطا، خطای تجهیزات در حین نمونه‌برداری و یا از دست دادن داده‌ها به‌دلیل مشکلات در ذخیره‌سازی داده رخ می‌دهد (Aryanmanesh, 2024; Boukerche *et al.,* 2020). معمولاً ساده‌ترین و متداول‌ترین رویه در بازسازی داده پرت حذف آن‌ها است. حذف داده‌های پرت از مجموع داده‌ها صورت می‌گیرد که در نتیجه مدل‌های آماری پارامتری قادر خواهند بود با داده‌های تعلیمی تطبیق داده شوند. روش‌های تشخیص داده‌های پرت شامل روش‌های نظارت‌شده، نیمه نظارت‌شده و نظارت‌نشده هستند و برخی روش‌های مبتنی بر توزیع، مبتنی بر خوشه‌بندی و مبتنی بر چگالی را شامل می‌شوند (Kiani, 2015). در این میان، داده‌های رقوم آب رودخانه‌ها به‌منظور تخمین دبی جریان از روابط تجربی استفاده می‌کنند و لازم است روش‌های محاسباتی مناسب به‌منظور افزایش دقت در محاسبات رقوم آب و تشخیص داده‌های پرت و بازسازی داده‌های گمشده مورد بررسی قرار گیرند (Bae & Ji, 2019; Boiten, 2003; Fenton & Keller, 2001; Herschy, 2008; Horner *et al.,* 2018). به‌دلیل خطای محاسباتی، مقادیر صحیح خاص، گزارش اشتباه و یا خطای نمونه‌برداری و همچنین به‌دلیل خطاهای انسانی و ابزاری ممکن است مواردی مانند ثبت نشدن آمار، ثبت آمار غلط، خرابی یا ازبین رفتن دستگاه‌های اندازه‌گیری یا تشخیص داده‌های پرت و حذف آن‌ها با عنوان داده‌های گم‌شده پیش آید. بنا بر این، تخمین و برآورد این داده‌ها برای استفاده در مدل‌ها ضروری است و به‌منظور کاهش بروز خطا باید پیش از به کارگیری آن‌ها پیش‌پردازش صورت گیرد. عملیات پیش‌پردازش، سری داده را برای محاسبات از جمله کلاسه‌بندی، پیش‌بینی و تخمین آماده می‌کند و شامل حذف داده‌های گمشده، حذف داده‌های پرت، بازسازی مقادیر گمشده، و نرمال‌سازی داده‌ها است (Bahrami, 2018; Naghdi *et al.,* 2010). برای بازسازی و تکمیل داده‌های مختلف و رفع خلا داده‌های یک ایستگاه هواشناسی و هیدرولوژیکی، روش‌های مختلفی معرفی و توسعه داده شده‌اند که از جمله این روش‌های آماری می‌توان روش ایستگاه معرف، نسبت نرمال، رگرسیون خطی و محور مختصات را نام برد که از آن‌ها به‌عنوان روش‌های کلاسیک نام برده می‌شود. از روش‌های جدید نیز می‌توان به شبکه‌های عصبی، منطق فازی و رگرسیون فازی اشاره کرد (Naghdi *et al.,* 2010). تحقیقات مختلفی درخصوص تشخیص، شناسایی و بازسازی داده‌های پرت در هیدرولوژی صورت گرفته‌اند که به برخی از آن‌ها به اختصار اشاره می‌شود. رحمدل و همکاران (Rahmdel *et al.,* 2021) به ارزیابی و همگن‌سازی سری زمانی داده‌های دما و بارش در ایستگاه‌های هواشناسی با رویکرد تحلیل اکتشافی و آزمون فرض پرداختند. در نهایت، داده‌های پرت دمای بیشینه و کمینه را محاسبه نمودند و ایستگاه‌های همگن و ناهمگن شناسایی شدند. کیانی و منتظری (Kiani, 2015) به بررسی و مرور روش‌های مختلف تشخیص و شناسایی داده‌های پرت پرداختند. در مطالعه آن‌ها، مؤلفه‌های کلیدی داده‌های پرت، الگوریتم‌های اصلی بر اساس دامنه کاربرد و نوع ناهنجاری مورد بررسی و مقایسه قرار گرفتند و انواع روش‌های پرکاربرد و جدید بررسی شدند. در مطالعه بهرامی و همکاران (Bahrami *et al*., 2018) با استفاده از داده‌های ایستگاه سینوپتیک شهر آباده به‌عنوان ورودی شبکه عصبی مصنوعی پرسپترون چندلایه، بارش پیش‌بینی شد و بهترین ساختار شبکه و بهترین روش نرمال‌سازی انتخاب شد. نتایج آن‌ها نشان دادند که مدل به حذف پارامتر بیشترین رطوبت، بیشتر از سایر پارامترها حساسیت نشان داد. در مطالعه آن‌ها، روش حداقل و حداکثر با ساختار شبکه سه لایه و ۱۳ لایه پنهان به‌عنوان بهترین روش نرمال‌سازی انتخاب شد. احمدی و همکاران (Ahmadi *et al.,* 2014) برای نرمال‌سازی داده‌های دبی جریان ماهانه و روزانه رودخانه باراندوزچای از توابع مختلف استفاده نمودند و با توجه به نتایج ضریب چولگی که مقادیر نزدیک به صفر را نشان داد، تابع لگاریتم را برای نرمال‌سازی انتخاب نمودند. آنها با استفاده از مدل‌های خطی و غیر خطی به مقایسه مدل‌ها در برآورد جریان رودخانه پرداختند. نتایج حاکی از آن بودند که مدل دوخطی در مقیاس روزانه باعث کاهش مقدار خطا شد و ضریب همبستگی را افزایش داد. نقدی و همکاران (Naghdi *et al.,* 2010) توانایی شبکه عصبی مصنوعی را در بازسازی داده‌های دبی ماهانه کارون بزرگ مورد ارزیابی قرار دادند و نتایج این روش با نتایج روش‌های دیگر ازجمله رگرسیون خطی ساده، رگرسیون خطی چندمتغیره، خودهمبستگی، نسبت نرمال و محور مختصات مورد مقایسه قرار گرفتند. در هر روش، پس از حذف داده‌های مشاهداتی مقادیر آن‌ها با روش‌های مختلف برآورد شدند و با استفاده از میانگین مجذور مربعات خطا[1] اولویت هر یک از این روش‌ها مورد بررسی قرار گرفت. در نهایت، نتایج نشان دادند که شبکه عصبی مصنوعی در مقایسه با سایر روش‌ها برتری داشت. اومر و گری (Umar & Gray, 2023) روش KNN را به‌منظور نسبت‌دهی داده‌های گمشده مورد استفاده قرار دادند. نتایج تحقیق آنها نشان دادند که روش KNN یکی از دقیق‌ترین روش‌ها جهت محاسبه و نسبت‌دهی مقادیر گمشده در داده‌های چندمتغیره بود. در پژوهش حاضر، سعی شده است که کارآیی روش‌های پلات جعبه‌ای[2]، روش میانگین و انحراف معیار[3]، روش میانه[4] روش Z-Score، تست

[1] RMSE
[2] Box Plot
[3] Mean Standard Deviation
[4] Median Absolute Deviation

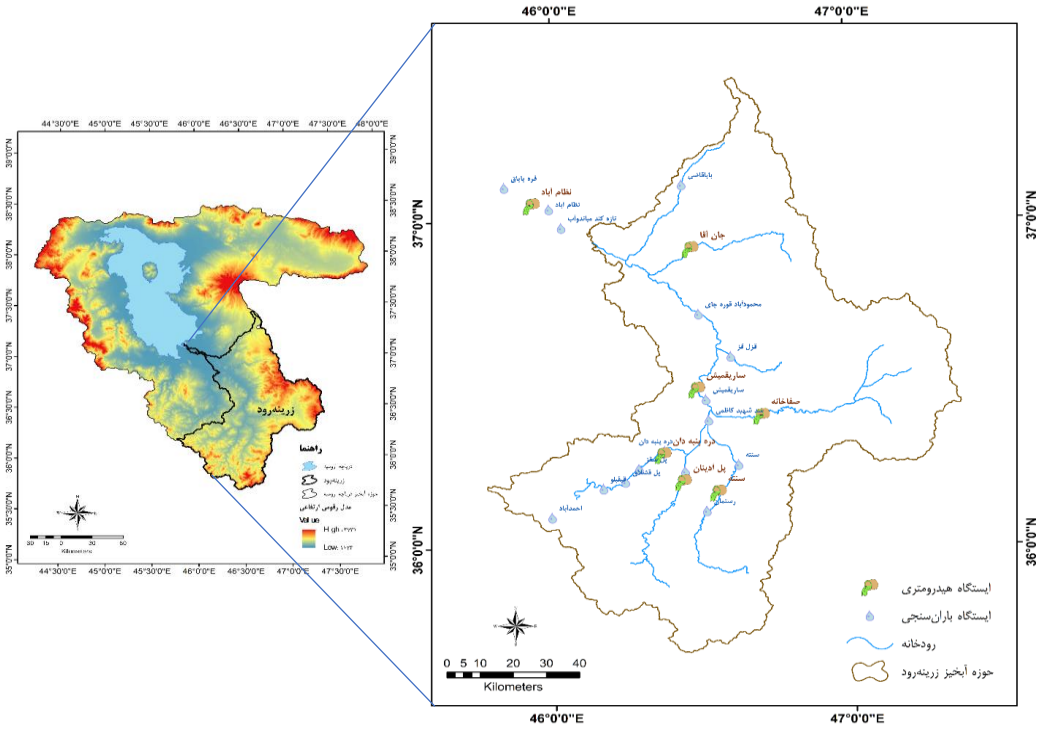
[DOI: 10.61882/jwmr.2025.1310]

نرمال‌سازی[1] و روش‌های آماری گرابز[2]، دیکسون[3]، روزنر[4] و مربع کای[5] در ایستگاه‌های هیدرومتری حوضه آبخیز زرینه‌رود مورد ارزیابی قرار گیرد و مناسب‌ترین روش تشخیص داده‌های پرت شناسایی و داده‌های گمشده در سری داده جریان رودخانه مورد بازسازی قرار گیرد.

مواد و روش‌ها

منطقه مورد مطالعه

حوزه آبخیز زرینه‌رود، بزرگترین حوضه آبخیز از حوضه آبریز دریاچه ارومیه است که در موقعیت جغرافیایی ۴۷ °۴۵ تا ۲۰ °۴۷ طول شرقی و ۴۱ °۳۵ تا ۲۷ °۳۷ عرض شمالی واقع شده است و مساحت آن حدود ۱۲۰۲۵ کیلومترمربع است. شهرهای میاندوآب، شاهین‌دژ، تکاب و سقز از کانون‌های شهری این حوضه هستند. سد شهید کاظمی زرینه‌رود تنها سد اصلی قابل بهره‌برداری حوضه است که از آن برای مصارف کشاورزی و شرب استفاده می‌گردد (شکل ۱).



شکل ۱- موقعیت جغرافیایی و ایستگاه‌های باران‌سنجی و هیدرومتری در حوزه آبخیز زرینه‌رود

Figure 1. Geographical location of the Zarrineh-Roud basin and rain gauges in Iran

ایستگاه‌های مورد مطالعه

در این تحقیق، به منظور تشخیص داده‌های پرت و بازسازی داده‌های گمشده و ناقص سری زمانی، داده های هیدرولوژی از داده‌های دبی ماهانه شش ایستگاه هیدرومتری و داده‌های اقلیمی ۱۶ ایستگاه باران‌سنجی در حوزه زرینه‌رود در نرم‌افزار R برنامه‌نویسی و مورد بررسی قرار گرفتند. نقشه ایستگاه‌ها و موقعیت قرارگیری آن‌ها در منطقه مورد مطالعه در شکل ۱ آورده شده‌اند. از داده‌های بارش ایستگاه‌های باران‌سنجی جهت محاسبه دبی استفاده شد. ابتدا داده‌های پرت ایستگاه‌های مذکور بررسی و مشخص شد. در قدم بعدی، داده‌های پرت حذف شده با روش KNN ، رگرسیون لاسو و رگرسیون خطی بیزین نسبت‌دهی گردیدند. به منظور نسبت‌دهی و جایگذاری داده‌ها از آمار ایستگاه‌های باران‌سنجی (جدول ۲) استفاده گردید.

[1] Normal Test
[2] Grubs
[3] Dixon
[4] Rosner
[5] Chi Square

[DOI: 10.61882/jwmr.2025.1310]

جدول ۱- ایستگاه‌های هیدرومتری مورد مطالعه

Table 1. Specification of discharge stations

نام ایستگاه Name of the Station	جان‌آقا Jan Agha	سنته Senteh	ساریقمیش Sarigamish	پل آدینان Pol Adinan	دره پنبه‌دان Dare Panbedan	نظام‌آباد Nezam Abad
ارتفاع به متر Altitude (m)	1410	1434	1380	1460	1470	1283

جدول ۲- ایستگاه‌های باران‌سنجی مورد مطالعه در بازسازی نواقص آماری

Table 2. Characteristics of rain gauge stations studied in missing value imputation

کد Code	PCP1	PCP2	PCP3	PCP4	PCP5	PCP6	PCP7	PCP8
نام ایستگاه Name of the Station	پل سقز Pol saghez	ساریقمیش Sarigamish	رستمان Rostaman	قزل قز Gezel gez	قبقبلو Gabgablou	دره‌پنبه‌دان Dareh Panbedan	پل	قره پاپاق Gara Papagh
ارتفاع به متر Altitude (m)	1480	1380	1900	1650	1500	1470	1460	1290
کد Code	PCP9	PCP10	PCP11	PCP12	PCP13	PCP14	PCP15	PCP16
نام ایستگاه Name of the station	تازه‌کند میاندوآب Tazekand Miandoab	سد شهید کاظمی Sad Shahid-Kazemi	نظام آباد Nezam Abad	سنته Senteh	محمودآباد قوره‌چای Mahmood Abad goorechay	پل قشلاق Pol geshlagh	احمدآباد سقز Ahmadabad sagez	باباقاضی Babaghazi
ارتفاع به متر Altitude (m)	1290	1437	1283	1434	1500	1434	1686	1584

تشخیص داده‌های پرت

به‌منظور تشخیص داده‌های پرت در سری زمانی داده‌های هیدرولوژیکی، از نرم‌افزار R و پکیج‌های VIM، DMwR2، mvoutlier و graphics استفاده شد.

روش مبتنی بر چگالی

روش مبتنی بر چگالی ابتدا توسط (Breunig *et al*., 2000)، معرفی شد. این روش توزیع چگالی در داده ورودی را تخمین زده، سپس داده‌های پرت را بر اساس داده‌های قرار گرفته در مناطق دارای چگالی کمتر بر اساس چگالی نزدیکترین همسایه تخمین می‌زند. داده‌ای که در کمترین فاصله با کمترین تراکم قرار گرفته باشد، به عنوان داده پرت و داده‌ای که در بخش با تراکم بالا قرار گرفته باشد، داده نرمال در نظر گرفته می‌شود.

روش باکس پلات یا نمودار جعبه‌ای (Box Plot)

ساده‌ترین روش برای تخمین و تشخیص داده‌های پرت، ارائه نمودار جعبه‌ای یا همان باکس پلات است. نمودار جعبه‌ای، داده‌های عددی را با توجه به توزیع آن‌ها در چارک‌های مختلف نشان می‌دهد. فاصله بین قسمت‌های مجزا در جعبه بازتاب گستردگی داده‌ها است. در یک نمودار جعبه‌ای، انتهای خطوط نمودار به احتمالات مختلف از جمله مقادیر حداقل و حداکثر داده‌ها، انحراف معیار حد بالا و پایین در میانگین داده‌ها، ۹٪ و ۹۱٪ داده‌ها، کمترین مقدار بین $Q_1 - 1.5\ IQR$ در چارک پایین و بالاترین مقدار از $Q_3 + 1.5\ IQR$ در چارک بالا اشاره دارد (Suri *et al*., 2019).

روش میانگین و انحراف معیار

در این روش، در صورت وجود داده پرت در سری داده، مقدار میانگین به‌شدت به یک سو متمایل خواهد بود. هر چه مقدار انحراف معیار کمتر باشد، نشان‌دهنده نزدیک بودن داده‌ها به مقدار میانگین است و در صورتی‌که مقدار انحراف معیار عدد بالاتری باشد، داده‌ها در محدوده بزرگتری گسترش یافته‌اند. در این روش، میانگین و انحراف معیار برای داده‌ها محاسبه می‌شوند و در صورت بالا بودن انحراف معیار در هر داده، آن داده به‌عنوان داده پرت در نظر گرفته می‌شود (Dave & Varma, 2014).

روش میانه

روش میانه مشابه میانگین عمل کرده، مقدار فاصله از مرکز داده‌ها را محاسبه می‌کند و این مقدار نسبت به حضور داده‌های پرت حساس‌تر است و یکی از روش‌های تشخیص داده‌های پرت با استفاده از میانه تعیین نقطه شکست است (Donoho & Huber, 1983). میزان انحراف مطلق از میانه یا MAD[1] با رابطه شماره ۱ محاسبه می‌شود.

$$MAD = b\ M_i\ (|x_i - M_j\ (x_j)|) \qquad (۱)$$

که در آن: x_j مشاهدات اصلی در سری داده، M_i میانه داده‌ها و b عدد ثابت هستند که به‌صورت پیش‌فرض عدد ثابت ۱/۴۸۲۶ با فرض نرمال بودن داده‌ها و بدون درنظر گرفتن ناهمگنی محاسبه می‌گردد (Dave & Varma, 2014).

روش HBOS

روش هیستوگرام یک روش ساده در تشخیص داده‌های پرت محسوب می‌گردد. روش مبتنی بر هیستوگرام توسط (Goldstein & Dengel, 2012)، ارائه شد و به‌نام الگوریتم HBOS[2] نیز نامیده می‌شود (Smiti, 2020). روش هیستوگرام یک روش نظارت‌نشده در تشخیص داده‌های پرت به‌حساب می‌آید. این روش مقدار هر داده را به صورت مجزا بر اساس فراوانی و توزیع آماری آن مجموعه داده بررسی نموده، مقادیر غیر نرمال را بر اساس هیستوگرام سری داده شناسایی و به عنوان داده پرت معرفی می‌نماید.

روش Rosner

هدف از آزمون Rosner بررسی نرمال بودن پیش از انجام تحلیل‌های آماری بر روی داده‌ها است و سپس آزمون‌های مرتبط با بازسازی و یا حذف داده‌های ناهمگن و یا پرت انجام

[1] Median Absolute Deviation

[2] Histogram-Based Outlier Score

[DOI: 10.61882/jwmr.2025.1310]

می‌گردد. روزنر معادله شماره ۲ را جهت تشخیص داده‌های پرت معرفی نمود.

$$T_1 = \frac{max|X_i - \bar{X}|}{\hat{\sigma}} \quad (۲)$$

که در آن: $\bar{X}$ و $\hat{\sigma}$ به ترتیب میانگین و انحراف معیار کل داده‌ها هستند. ابتدا داده پرت مربوط به $max|X_i - \bar{X}|$ حذف و T_2 بر اساس باقی داده‌ها محاسبه می‌شود و این عمل تا محاسبه T_k و تا زمانی‌که داده پرت باقی نماند ادامه می‌یابد (Cohn *et al.*, 2013).

روش آزمون مربع کای

آزمون مربع کای به‌منظور تعیین تفاوت‌ها میان متغیرها در سری داده‌ها مورد استفاده قرار می‌گیرد و احتمال فرض صفر را در مشاهدات ارزیابی می‌کند. آزمون آماری مربع کای از توزیع پیرسون تبعیت می‌نماید. در این آزمون، پیرسون محدوده‌ای را بین x_j و M تقسیم می‌کند که در این محدوده توزیع N_i و پارامترهای n طبق رابطه شماره ۳ محاسبه خواهد شد:

$$P_i = P(X_j\ fall\ in\ E_i) = \int_{E_i} d_F(x) \quad (۳)$$

توزیع پیرسون زمانی به مربع کای می‌رسد که ابتدا مقادیر N_i-np_i در تعداد نمونه‌های زیاد تقریبا از توزیع نرمال چندمتغیره پیروی نماید و اگر M-1 سلول در نظر گرفته شود، این توزیع تک‌متغیره نخواهد بود. دوم این که اگر $Y = (Y_1, \dots, Y_p)$ از توزیع نرمال p متغیره پیروی نماید، فرم مربع توزیع پیرسون به‌عنوان تابعی از Y خواهد بود. بنابراین، اگر ($Y = (N_1\text{-}np_1$ $N_{M-1}\text{-}np_{M-1}$) باشد، مربع آن طبق رابطه شماره ۴ محاسبه خواهد شد.

$$x^2 = \sum_{i=1}^{M} \frac{N_i - np_i)^2}{np_i} \quad (۴)$$

که در آن: فرم مربع $(Y-\mu)'\sum^{-1}(Y-\mu)$ در توان تابع چگالی توزیع $x^2_{(p)}$ به عنوان تابعی از Y است (D'Agostino, 1986).

نرمال‌سازی داده‌ها

توزیع نرمال یکی از متداول‌ترین و مهمترین توزیع‌های احتمالی پیوسته به‌شمار می‌رود که گاها بنام توزیع گوسین نیز نامیده می‌شود و برای آزمودن خطای محاسباتی به‌کار برده می‌شود. احتمال دارد که مقادیر داده‌های مورد مطالعه با محدوده خود یا دامنه‌ای که در آن قرار دارند، متفاوت باشند و اگر این مقدار متفاوت بسیار بزرگ باشد، به همان نسبت اثر بالایی در تابع هزینه خواهد داشت؛ این مشکل با نرمال‌سازی داده‌ها رفع می‌شود (Nazeri Tahrudi, 2014; Tourian *et al.*, 2017). در این مطالعه، آزمون Z-score جهت ارزیابی و نرمال‌سازی داده‌ها مورد استفاده قرار گرفت. پیش از تشخیص داده‌های پرت، ابتدا باید نرمال‌سازی انجام شود و به این منظور از آزمون‌های مختلف جهت نرمال‌سازی استفاده شد. آزمون‌های کلموگوروف–اسمیرنوف و آزمون شاپیرو–ویلک به‌منظور نرمال‌سازی داده‌ها مورد استفاده قرار گرفتند. با اجرای نرمال‌سازی، اثر داده‌های پرت کاهش می‌یابد چراکه داده‌های پرت می‌توانند تأثیر زیادی بر مدل‌ها داشته باشند و باعث ایجاد خطاهای جدی شوند (Montgomery & Runger, 2019).

رابطه آزمون کلموگوروف–اسمیرنف به‌صورت رابطه ۵ است (Ordooni *et al.*, 2021):

$$D_{n_1,n_2} = max_x|F_{X,n_1(x)} - F_{Y,n_2}| \quad (۵)$$

به‌منظور محاسبه بخش آماره کلموگوروف–اسمیرنوف بایدمقدار حداکثر قدر مطلق $|F_{X,n_1(x)} - F_{Y,n_2}|$ به‌دست آید. در واقع، محاسبه این آزمون بر اساس حداکثر اختلاف توابع است.

بازسازی داده‌های گمشده

به‌منظور بازسازی داده‌های گمشده یا گمشده از سه روش KNN[1]، Lasso[2] و Bayesian Linear Regression استفاده گردید. تکنیک‌های مورد استفاده موجب کاهش اثر خطی در مقادیر پیش‌بینی می‌شوند و از بیش‌برازش در مقادیر سری زمانی جلوگیری می‌کنند.

الگوریتم KNN

الگوریتم KNN یکی از روش‌های پرکاربرد مبتنی بر نمونه است که با مدل‌های ناپارامتری و طبقه‌بندی یادگیری نظارت شده ارتباط دارد و مقادیر نقاط گمشده یا محاسبه نشده در داده‌ها را باتوجه به فاصله یا شباهت آن‌ها با داده‌های متناظر مشاهداتی سایر نقاط به‌دست می‌آورد (Umar & Gray, 2023). در اجرای این الگوریتم، فاصله بین متغیرهای وابسته و مقدار هدف، سه پارامتر اهمیت بالایی دارند از جمله این پارامترها تعداد همسایه، اندازه فاصله و وزن‌دهی یا عدم وزن‌دهی (Holmström & Fransson, 2003; Maanavi & Roozbeh, 2021; Shataee *et al.*, 2012). به‌طور کلی، از روش KNN به‌منظور تخمین تابع چگالی توزیع داده و طبقه‌بندی داده‌های آزمایشی بر اساس الگوی تعلیمی استفاده می‌شود. هدف این روش، طبقه‌بندی و تخمین ویژگی‌های سری داده گمشده بر اساس مقادیر مشابه و فاصله سری داده گمشده از مقادیر مشابه مشاهداتی است. قدم اول در این روش، تعیین فاصله بین داده‌های گمشده و داده‌های مشاهداتی است و عموماً از بین روش‌های تعیین فاصله روش اقلیدسی بیشتر مورد استفاده قرار می‌گیرد که طبق فرمول شماره ۶ محاسبه می‌شود. فاصله اقلیدسی در محاسبه نزدیکترین همسایه بر پایه اختلاف فاصله دو نقطه براساس مثلث و برپایه قضیه فیثاغورس محاسبه می‌شود.

$$d,(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (۶)$$

که در آن X نمونه‌های تعلیمی و Y نمونه‌های آموزش با همان تعداد پارامتر هستند. پس از تعیین فاصله بین داده‌ها، نمونه‌ها به‌ترتیب صعودی از کمترین فاصله که همان حداکثر تشابه بین داده‌ها است، تا بیشترین فاصله قرار می‌گیرند. میزان کارایی این روش به‌دقت در انتخاب مشابه‌ترین داده‌ها بستگی دارد (Poursalehi *et al.*, 2019; Rajabi Jaghargh *et al.*, 2024). در روش KNN به‌منظور بازسازی داده از ایستگاه‌های باران‌سنجی و مقادیر بارش متناظر به‌منظور محاسبه و تشخیص دبی استفاده شد. در روش KNN برای محاسبه تشابه و نزدیکی بین هدف و داده‌های تعلیمی، فاصله اقلیدسی مورد استفاده قرار می‌گیرد که فاصله بین هر داده و مقدار هدف را تخمین می‌زند.

[1] K Nearest Neighbor

[2] Least Absolute Shrinkage and Selection Operator

[DOI: 10.61882/jwmr.2025.1310]

تعداد K داده دارای کمترین فاصله نسبت به مقدار هدف به عنوان نزدیکترین همسایه درنظر گرفته می‌شوند.

جدول ۳- ایستگاه‌های متناظر باران‌سنجی و هیدرومتری در محاسبه KNN

Table 3. Related rain gauge and discharge stations in KNN calculation

ایستگاه‌های هیدرومتری Discharge Station	ایستگاه‌های باران‌سنجی متناظر Related rain gauge stations			
نظام‌آباد (Nezam Abad)	تازه‌کند میاندوآب Tazekand Miandoab	نظام‌آباد Nezam Abad		
دره پنبه‌دان (Dareh Panbedan)	قبقبلو Gabgablou	دره پنبه‌دان Dareh Panbedan	قره پاپاق Gara Papagh	پل قشلاق Pol geshlagh
پل آدینان (Pol Adinan)	قبقبلو Gabgablou	دره پنبه‌دان Dareh Panbedan	قره پاپاق Gara Papagh	پل قشلاق Pol geshlagh
ساریقمیش (Sarigamish)	پل سقز Pol Saghez	ساریقمیش Sarigamish	احمدآباد سقز Ahmadabad sagez	
سنته (Senteh)	رستمان Rostaman	سنته Senteh	پل قشلاق Pol geshlagh	باباقاضی Babaghazi
جان آقا (Jan Agha)	قزل قز Gezel gez	سد شهید کاظمی Sad Shahid-Kazemi	محمودآباد قوره‌چای Mahmood Abad goorechay	

رگرسیون لاسو[1]

روش رگرسیون لاسو یک روش منظم‌سازی است که هدف آن کاهش پیچیدگی مدل و جلوگیری از بیش‌برازشی است. در رگرسیون لاسو به جای محاسبه مربعات مقادیر، مقدار دقیق آن‌ها وارد محاسبه می‌شود. این نوع از منظم‌سازی، ضرایب را به صفر نزدیک می‌کند. از این‌رو، برخی تخمین‌گرها در ارزیابی خروجی مدل عملکرد خوبی ندارند، اما لاسو نه تنها در کاهش بیش‌برازش موثر عمل می‌کند، بلکه در انتخاب نوع ویژگی مدل نیز کمک می‌کند.

(۹)

$$Lasso\ regresion = min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2\right\}$$

استفاده از رگرسیون لاسو به محقق کمک می‌کند تا روش مناسب برای مدل‌سازی متغیر پاسخ با استفاده از کمترین و بهترین تعداد متغیرهای مستقل را پیدا نماید. از جمله دلایل استفاده از رگرسیون لاسو عبارت‌اند از: هم‌خطی درون مدل وجود داشته باشد، مدل دچار بیش برازشی باشد، و مدل دارای واریانس خیلی بالا باشد. رگرسیون لاسو با قرار دادن یک پارامتر تنظیمی عوامل ناهنجاری را کاهش می‌دهد و با توجه به این که در رگرسیون لاسو امکان فرض ضرایب مقادیر صفر نیز وجود دارد، می‌تواند یکی از روش‌های انتخاب متغیر تلقی شود.

رگرسیون خطی بیزین[2]

رگرسیون خطی بیزین نوعی تحلیل آماری است که ترکیبی از روش‌های رگرسیون خطی و بیزین را استفاده می‌کند. این روش به طور خاص در شرایطی مفید است که اطلاعات قبلی در دسترس باشد و خواسته شود داده‌ها با داده‌های جدید ترکیب شوند و تخمین‌های دقیق‌تر به دست آیند. در رگرسیون خطی بیزین، فرض می‌شود که ضریب‌ها دارای توزیع احتمالاتی هستند. به عبارت دیگر، به جای این که یک مقدار ثابت برای هر ضریب وجود داشته باشد، یک توزیع احتمال برای هر ضریب تعریف می‌شود. این توزیع‌های احتمالاتی با استفاده از اطلاعات قبلی و داده‌های جدید به‌روزرسانی می‌شوند تا تخمین بهتری از ضرایب بدست آید. با استفاده از قاعده بیز، توزیع پسین ضرایب به‌صورت زیر محاسبه می‌شود:

$$p(\beta|X,y) \propto (y|x,\beta,\sigma^2).p(\beta) \qquad (۱۰)$$

معیارهای ارزیابی

در این مطالعه، به منظور ارزیابی و بررسی صحت نتایج حاصل به روش اعتبارسنجی متقابل[3] از دو معیار ارزیابی RMSE[4] (جذر میانگین مربعات خطا) و ضریب همبستگی R^2[5] استفاده گردید. RMSE اختلاف بین دو سری داده و R^2 همبستگی میان دو سری داده را نشان می‌دهد. معادلات ۱۱ و ۱۲ به ترتیب نحوه محاسبات معیارهای ارزیابی را نشان می‌دهند (Bahrami, 2018).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2} \qquad (۱۱)$$

که در آن n نشان دهنده تعداد کل داده‌های مشاهداتی، X_i مقادیر واقعی و Y_i مقادیر تخمینی هستند.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \hat{y})^2} \qquad (۱۲)$$

رابطه بین R^2 و RMSE به شکل زیر است (رابطه ۱۳).

$$R^2 = 1 - \left(\frac{RMSE^2}{\sigma_y^2}\right) \qquad (۱۳)$$

که در آن σ_y^2 واریانس کل داده‌های مقادیر مشاهده‌ای است.

نتایج و بحث

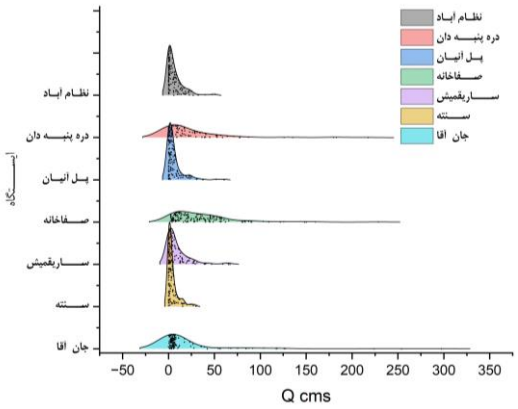
توزیع داده‌ها در ایستگاه‌های مورد مطالعه در شکل ۲ و نمودار جعبه‌ای داده‌ها در شکل شماره ۳ نشان داده شده‌اند. شکل ۲ بیانگر وجود چولگی در داده‌های ایستگاه‌های دره‌پنبه‌دان و جان‌آقا است و ایستگاه‌های نظام آباد، پل‌آدینان، سنته و ساریقمیش دارای چولگی نسبتا زیادی هستند. بر اساس نمودارهای جعبه‌ای این مطالعه، داده‌های ایستگاه‌های جان‌آقا و دره‌پنبه‌دان دارای مقادیر پرت با حدود بالاتری در مقایسه با

[1] Lasso Regression
[2] Bayesian Linear Regression
[3] Cross Validation
[4] Root Mean Square Error
[5] R-Squared

[DOI: 10.61882/jwmr.2025.1310]

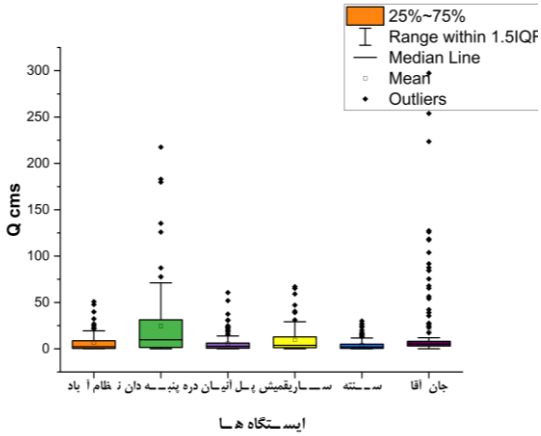
سایر ایستگاه‌ها هستند. ایستگاه‌های پل آدینان و سنته باتوجه به اینکه داده‌های خارج از ساقه داشتند اما درصد داده‌های خارج از چارک کمتر هستند و مقادیر نمایش داده شده به صورت داده پرت مورد شناسایی قرار گرفته‌اند. مقادیر پرت ایستگاه‌های جان‌آقا و دره پنبه‌دان در ادامه و در نرمال‌سازی حذف گردیدند. چولگی و وجود داده پرت در روش هیستوگرام به‌ویژه ایستگاه‌های جان‌آقا، ساریقمیش و پل آدینان بابی‌نظمی همراه است و توزیع ناهمگن و غیر نرمال دارند که پس از نرمال‌سازی، داده‌های پرت مشخص و حذف شدند. میزان p-value در هر دو آزمون شاپیرو–ویلک و کولموگروف–اسمیرنف مقادیری بسیار کمتر از ۰/۰۵ را نشان داده است و گواه این مطلب است که داده‌ها ابتدا در محدوده نرمال نبودند و با نرمال‌سازی در محدوده نرمال قرار گرفتند و نرمال‌سازی داده‌ها و حذف مقادیر پرت با دقت بالایی انجام شد. آزمون روزنر برای هر سری داده مقدار حد بالا را در دو تست متوالی ارائه داده است و همان مقدار و مقادیر بالاتر از آن را به عنوان داده پرت در نظر می‌گیرد.

به‌منظور اعتباریابی روش KNN، رگرسیون لاسو و رگرسیون بیزین از روش اعتبارسنجی متقابل یا Cross Validation استفاده شد که نتایج ارزیابی آن در جدول ۶ ارائه شده‌اند. اشکال ۷، ۸ و ۹ نیز نتایج نسبت‌دهی داده‌های گمشده را که به روش اعتبارسنجی متقابل انجام شده است، نشان می‌دهند. در این روش، ابتدا از قسمت‌های مختلف هیدروگراف که شامل دبی حداقل و دبی حداکثر و مقادیر متوسط دبی را داشت، به تعداد ۱۰ مشاهده حذف و به روش رگرسیون لاسو، رگرسیون بیزین و روش نزدیکترین همسایه، نسبت‌دهی انجام شد. با توجه به نتایج به‌دست آمده، الگوریتم KNN ضریب تبیین بالاتر از ۰/۷ و مقادیر RMSE نزدیک به صفر را نشان داده است. نتایج بیانگر تطابق خوب دبی مشاهده‌ای و دبی بازسازی شده توسط الگوریتم KNN در کلیه ایستگاه‌های مورد مطالعه هستند. در رگرسیون لاسو و رگرسیون خطی بیزین (شکل‌های ۷، ۸ و ۹)، مقادیر منفی نسبت‌دهی شده‌اند که عملا وجود مقادیر منفی در جریان رودخانه امکان‌پذیر نیست و این مساله بیانگر دقت پایین این دو روش در محاسبه و نسبت‌دهی داده‌های گمشده است. وجود مقادیر منفی باعث برهم‌خوردگی نظم توزیع داده شده است (اشکال ۱۰، ۱۱ و ۱۲). میزان پایین R^2 تا ۰/۰۰۰۱ و مقادیر بالای RMSE تا حدود ۵۵ در هر دو روش رگرسیون لاسو و رگرسیون خطی بیزین نشان‌دهنده عدم اطمینان از به کارگیری این روش‌ها در نسبت‌دهی داده‌های گمشده است (جدول ۷). در روش‌های رگرسیون لاسو و رگرسیون خطی بیزین (جدول ۷)، مقادیر R^2 و RMSE قابل قبول نیستند و در مقایسه با KNN روش‌های مطمئنی نمی‌توانند تلقی گردند. به منظور بررسی و انتخاب مناسبترین روش نسبت‌دهی داده‌های گمشده اقدام به بررسی تابع چگالی احتمال داده‌های مشاهده‌ای و تابع چگالی احتمال سری داده نسبت‌دهی شده گردید (شکل ۱۰، ۱۱ و ۱۲). نتایج مقایسه تابع چگالی احتمال به روش رگرسیون لاسو (شکل ۱۰) نشان از عدم تطابق مقادیر حداکثر و حداقل در سری داده هیدرولوژیکی دارند و این روش می‌تواند موجب برهم خوردگی تابع چگالی احتمال گردد. همچنین، نتایج مقایسه تابع چگالی احتمال به روش رگرسیون بیزین (شکل ۱۱) بیانگر عدم تطابق حداکثر و متوسط در سری داده هستند و بنا بر این، این روش می‌تواند موجب برهم خوردگی تابع چگالی احتمال حاصل از نسبت‌دهی گردد. نتایج مطابقت تابع چگالی احتمال مقادیر مشاهده‌ای و نسبت‌دهی شده به روش KNN (شکل ۱۲) نشان از تطابق قابل قبول دو تابع چگالی احتمال دارند و این روش در نسبت‌دهی مقادیر حداکثر، متوسط و حداقل نسبت به دو روش دیگر در ایستگاه‌های مورد مطالعه موفق عمل کرده است. بنابراین، می‌توان نسبت به به‌کارگیری سری داده نسبت داده شده داده‌های گمشده در مدل‌سازی‌های بعدی استفاده نمود.

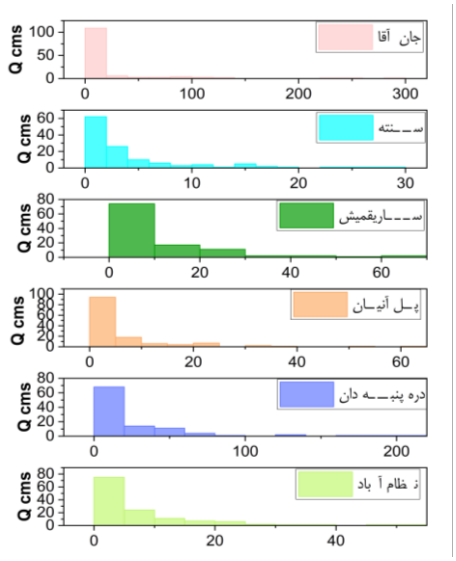


شکل ۲- توزیع داده‌ها در ایستگاه‌های مورد مطالعه
Figure 2. Probability of distribution at the studied stations

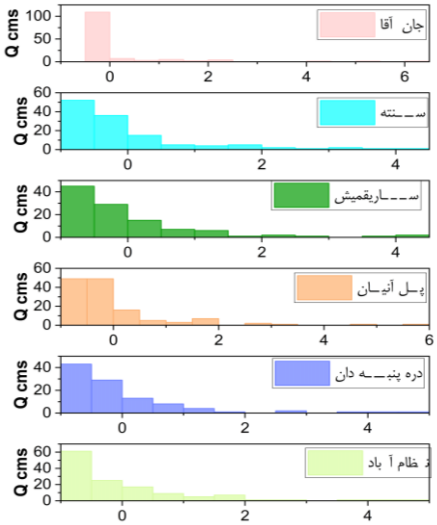
[DOI: 10.61882/jwmr.2025.1310]



شکل ۳- نمودارهای جعبه‌ای در ایستگاه‌های مورد مطالعه
Figure 3. Box plots at the studied stations

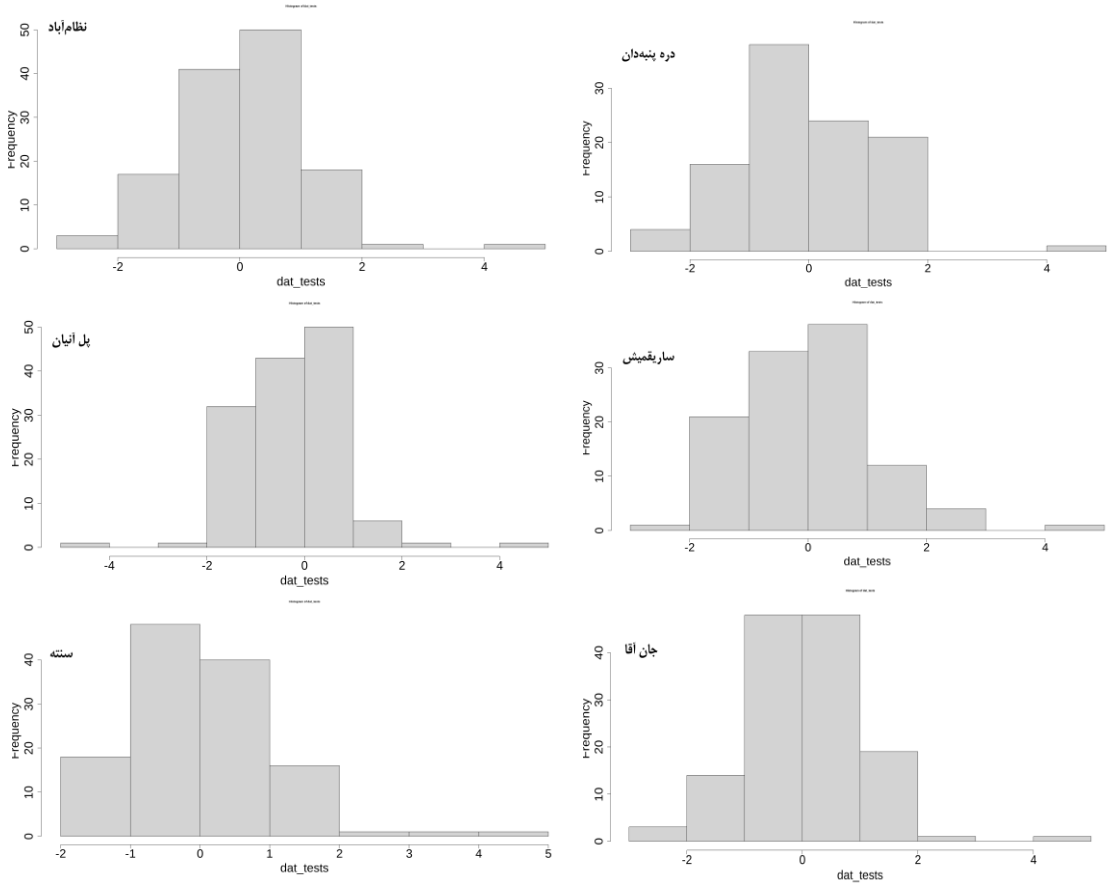


شکل ۴- نمودار هیستوگرام توزیع داده‌ها به تفکیک ایستگاه‌های مورد مطالعه
Figure 4. The histogram plot at the stations



شکل ۵- نمودارهای هیستوگرام Z-score به تفکیک ایستگاه‌های مورد مطالعه
Figure 5. Z-Score histogram plots at the stations

[DOI: 10.61882/jwmr.2025.1310]



شکل ۶- هیستوگرام توزیع داده‌ها پس از نرمال‌سازی داده‌ها
Figure 6. Histogram plots of data after normalization

جدول ۴- نتایج آزمون‌های آماری و برازش داده در ایستگاه‌های مورد مطالعه
Table 4. Results of statistical tests in study area stations

جان‌آقا Jan Agha	سنته Senteh	ساریقمیش Sarigamish	پل آدینان Pol Adinan	دره پنبه‌دان Dare Panbedan	نظام آباد Nezam Abad	
20.73	4.23	9.55	6.01	24.62	6.76	(میانگین) Mean
45.88	5.97	13.43	9.57	39.24	9.40	(انحراف معیار) Std
-4.98	-5.52	-10.64	-7.36	-30.8	-7.05	(حداقل) Min
15.16	9.48	18.05	12.64	50.24	11.59	(حداکثر) Max
3.35	1.97	3.73	2.64	9.72	2.27	(میانه) Med
Q-Q plot						
0.45	0.7	0.7	0.62	0.62	0.708	(حداقل) Min
3.38	0.62	0.62	0.56	0.59	0.67	(چارک اول) 1st quartile
0.34	0.37	0.43	0.35	0.37	0.47	(میانه) Median
0	0	0	0	0	0	(میانگین) Mean
0.27	0.12	0.24	0.02	0.15	0.205	(چارک سوم) 3rd quartile
6.03	4.31	4.27	5.70	4.91	4.69	(حداکثر) Max
-1.70	-3.08	-5.97	-4.1	-17.61	-4.02	(باند پایین) Lower band
11.88	7.04	13.38	9.38	37.05	8.56	(باند بالا) Upper band
Summary of outliers (خلاصه داده‌های پرت)						
2	3	3	3	2	2	(حداقل) Min
16.5	26.25	19.5	17	19	30	(چارک اول) 1st quartile
53	58.5	63	63	56.5	75.5	(میانه) Median
58.59	57.82	55.04	60.6	50.82	68.71	(میانگین) Mean
104.5	86.5	85	85	73.75	99.75	(چارک سوم) 3rd quartile
126	118	109	129	100	129	(حداکثر) Max

[DOI: 10.61882/jwmr.2025.1310]

جدول ۵- تشخیص نرمال بودن داده‌ها و تعیین داده‌های ناهمگن در سری داده

Table 5. Normality test results in study area stations

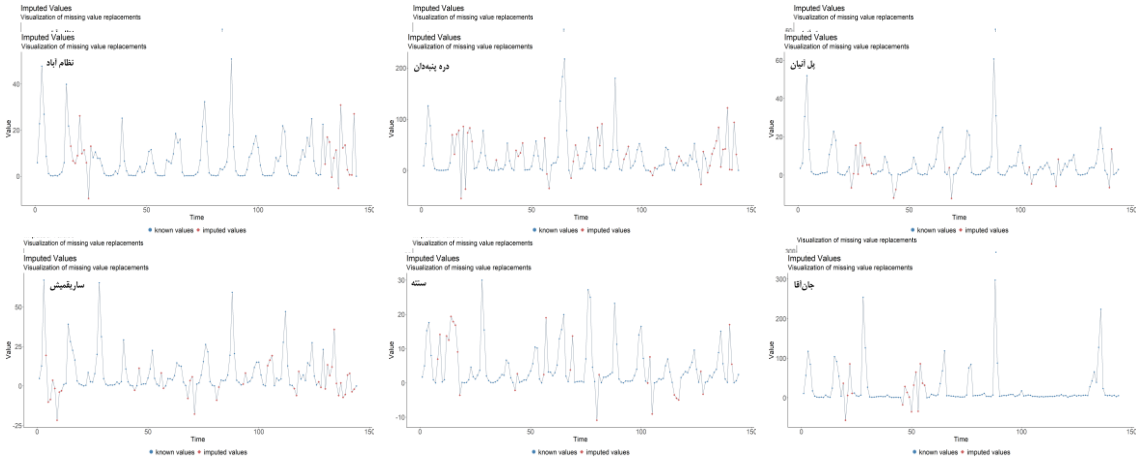
جان‌آقا Jan Agha	سنته Senteh	ساریقمیش Sarigamish	پل آدینان Pol Adinan	دره پنبه‌دان Dare Panbedan	نظام آباد Nezam Abad	
Rosner test 1						
20.73	4.23	9.55	6.01	24.62	6.76	(میانگین) Mean
45.88	5.97	13.43	9.57	39.24	9.40	(انحراف معیار) Std
297.41	30	66.97	60.66	217.51	50.89	(مقدار) Value
6.02	4.31	4.27	5.7	4.91	4.69	R i+1
3.47	3.45	3.41	3.48	3.39	3.46	Lambda i+1
true	true	true	true	true	true	(داده پرت) Outlier
Rosner test 2						
18.63	4.02	9.02	5.60	22.73	6.41	(میانگین) Mean
39.15	5.52	12.29	8.34	34.40	8.58	(انحراف معیار) Std
253.85	27.18	65.27	51.91	182.96	47.75	(مقدار) Value
6.007	4.19	4.57	5.55	4.65	4.81	R i+1
3.47	3.45	3.41	3.47	3.39	3.46	Lambda i+1
true	true	true	true	true	true	(داده پرت) Outlier
Shapiro-Wilk test						
0.45	0.69	0.69	0.62	0.63	0.71	W
0.001	0.001	0.001	0.001	0.001	0.001	P-value
Kolmogorov-Smirnov						
0.38	0.23	0.23	0.26	0.26	0.23	D
0.001	0.001	0.001	0.001	0.001	0.001	P-value
Chi square Highest values						
36.35	18.60	18.22	32.96	24.15	22.03	X
0.001	0.001	0.001	0.001	0.001	0.001	P-value
Chi square Lowest values						
0.20	0.50	0.50	0.39	0.39	0.50	X
0.65	0.47	0.47	0.53	0.53	0.47	P-value

جدول ۶- بخشی از نتایج اعتباریابی به‌روش Cross Validation در الگوریتم‌های نسبت‌دهی KNN، رگرسیون لاسو و رگرسیون بیزین

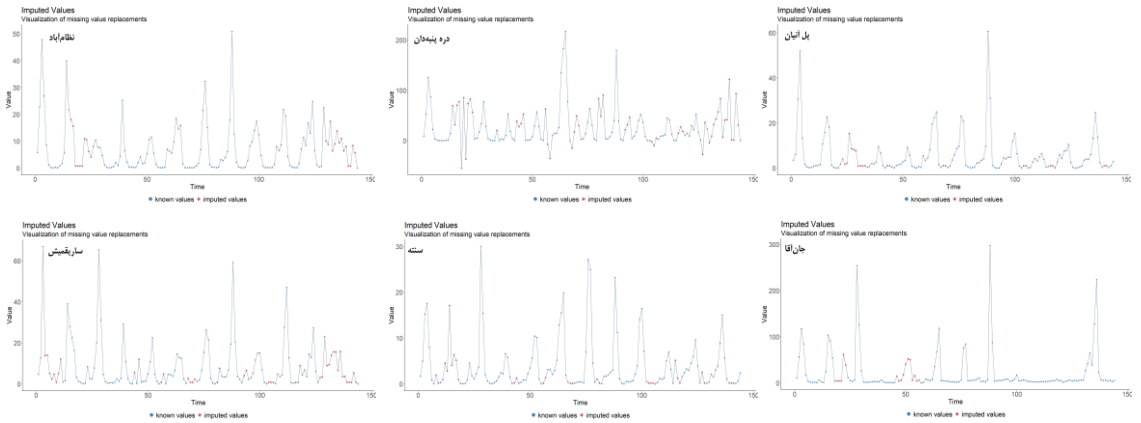
Table 6. Part of validation results with the cross validation method in imputing algorithms of KNN, Lasso Regression, and Bayesian Regression

پس از اعمال KNN (After KNN)						پیش از اجرای KNN (Before KNN)					
ردیف Series	نظام آباد Nezam Abad	دره پنبه‌دان Dare Panbedan	پل آدینان Pol Adinan	ساریقمیش Sarigamish	سنته Senteh	جان‌آقا Jan Agha	نظام آباد Nezam Abad	دره پنبه‌دان Dare Panbedan	پل آدینان Pol Adinan	ساریقمیش Sarigamish	سنته Senteh	جان‌آقا Jan Agha
1	1.28	10.52	1.11	2.59	1.20	6.16	NA	NA	NA	NA	NA	6.16
2	5.07	7.11	1.84	5.14	1.65	4.34	NA	NA	NA	NA	NA	4.34
3	1.61	2.06	0.29	1.26	0.06	6.06	NA	NA	0.29	NA	0.06	6.06
4	2.51	3.07	1.10	1.85	0.61	3.06	NA	NA	1.10	NA	0.61	3.06
5	5.09	9.42	2.94	4.09	2.41	5.41	NA	NA	2.94	NA	2.41	5.41
پس از اعمال رگرسیون لاسو (After Lasso regression)						پیش از اعمال رگرسیون لاسو (before Lasso regression)					
ردیف Series	نظام آباد Nezam Abad	دره پنبه‌دان Dare Panbedan	پل آدینان Pol Adinan	ساریقمیش Sarigamish	سنته Senteh	جان‌آقا Jan Agha	نظام آباد Nezam Abad	دره پنبه‌دان Dare Panbedan	پل آدینان Pol Adinan	ساریقمیش Sarigamish	سنته Senteh	جان‌آقا Jan Agha
1	17.59	82.04	-5.75	7.03	-5.57	-11.89	NA	NA	NA	NA	NA	NA
2	-0.19	-6.12	21.88	4.64	5.30	-14.15	NA	NA	NA	NA	NA	NA
3	-5.53	75.72	0.29	-2.68	0.06	10.15	NA	NA	0.29	NA	0.06	NA
4	1.51	1.13	1.10	30.03	0.61	6.58	NA	NA	1.10	NA	0.61	NA
5	8.57	33.01	2.94	11.07	2.41	76.26	NA	NA	2.94	NA	2.41	NA
پس از اعمال رگرسیون بیزین (After Bayesian regression)						پیش از اجرای رگرسیون بیزین (Before Bayesian regression)					
ردیف Series	نظام آباد Nezam Abad	دره پنبه‌دان Dare Panbedan	پل آدینان Pol Adinan	ساریقمیش Sarigamish	سنته Senteh	جان‌آقا Jan Agha	نظام آباد Nezam Abad	دره پنبه‌دان Dare Panbedan	پل آدینان Pol Adinan	ساریقمیش Sarigamish	سنته Senteh	جان‌آقا Jan Agha
1	13.04	69.38	-6.6	19.38	13.7	36.46	NA	NA	NA	NA	NA	NA
2	6.64	31.57	0.72	-10.18	12.49	-56.9	NA	NA	NA	NA	NA	NA
3	5.61	70.46	0.29	-8.56	0.06	5.7	NA	NA	0.29	NA	0.06	NA
4	8.93	77.92	1.10	3.58	0.61	85.68	NA	NA	1.10	NA	0.61	NA
5	26.16	-54.76	2.94	-1.58	2.41	10.98	NA	NA	2.94	NA	2.41	NA

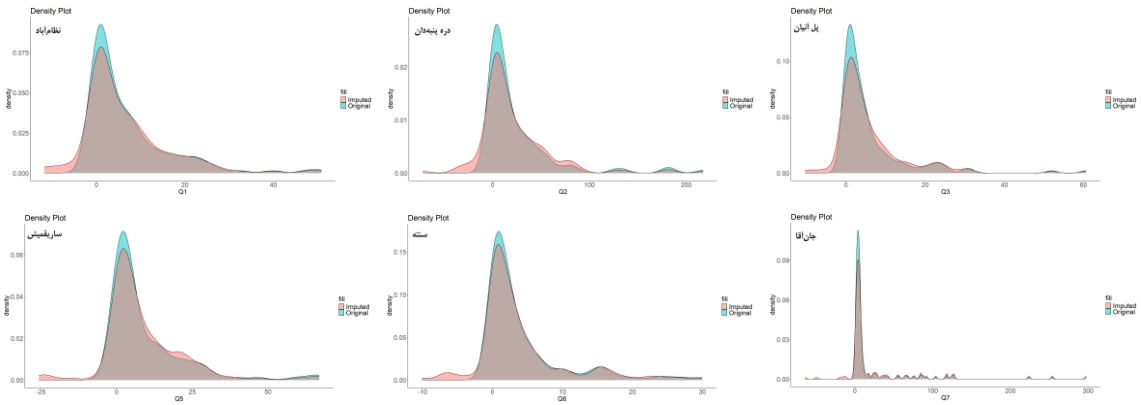
[DOI: 10.61882/jwmr.2025.1310]



شکل ۸- نسبت‌دهی داده‌های گمشده به‌روش رگرسیون بیزین
Fig 8. Imputation of missing values using the Bayesian Regression method

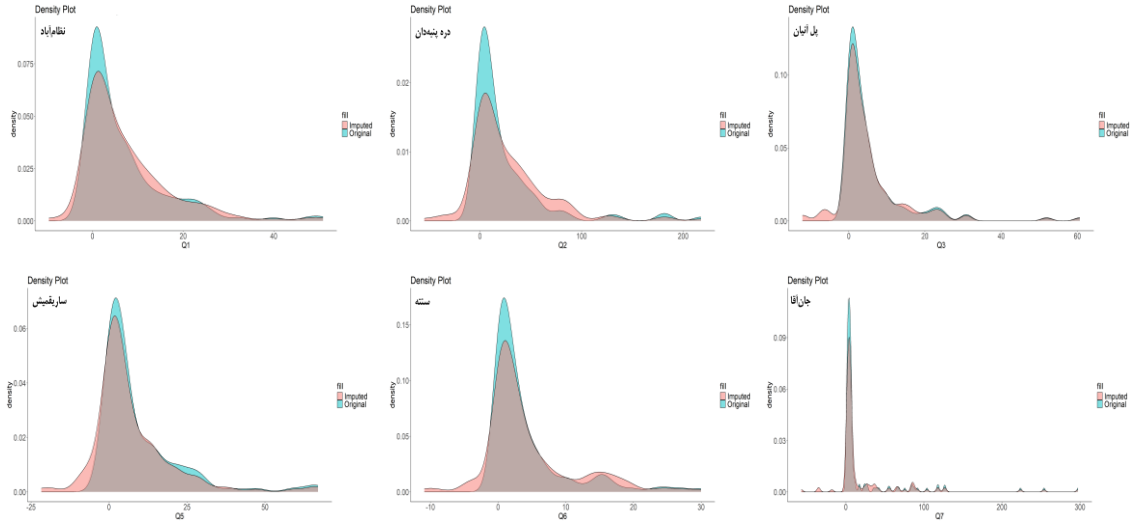


شکل ۹- نسبت‌دهی داده‌های گمشده به‌روش KNN
Fig 9. Imputation of missing values using the KNN method



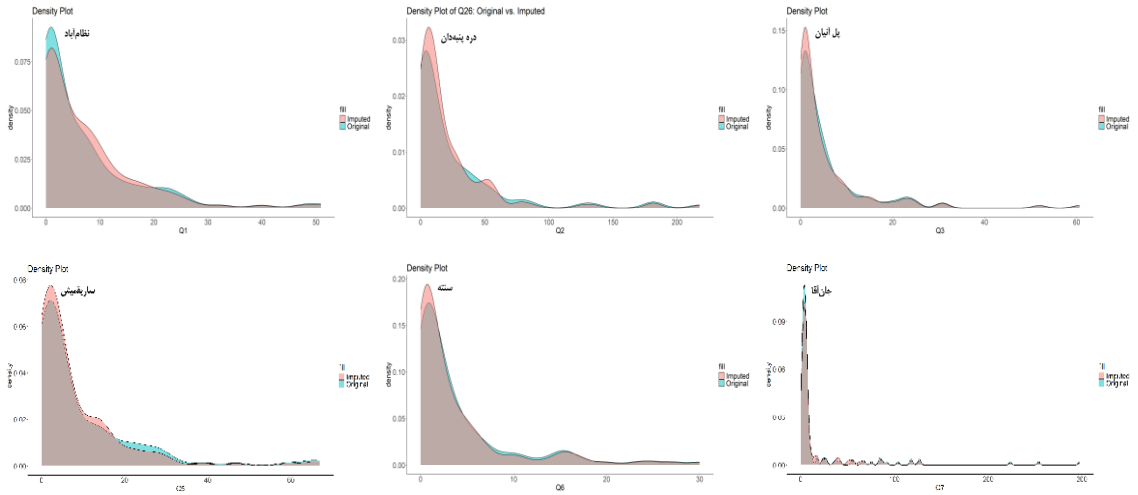
شکل ۱۰- تابع چگالی احتمال مقایسه داده‌های اصلی و نسبت‌دهی شده به روش رگرسیون لاسو
Fig 10. Probability distribution of actual values and imputed values using the Lasso Regression method

[DOI: 10.61882/jwmr.2025.1310]



شکل ۱۱- تابع چگالی احتمال داده‌های اصلی و نسبت‌دهی شده به روش رگرسیون بیزین

Fig 11. Probability distribution of actual values and imputed values using the Bayesian Regression method



شکل ۱۲- تابع چگالی احتمال داده‌های اصلی و نسبت‌دهی شده به روش KNN

Fig 12. Probability distribution of actual values and imputed values using the KNN method

جدول ۷- نتایج آماری محاسبه داده‌های گمشده به روش‌های KNN، رگرسیون لاسو و رگرسیون بیزین

Table 7. Statistical results of computing missing values by KNN, Lasso Regression, and Bayesian Regression

ایستگاه	نظام‌آباد (Nezam Ababd)		دره پنبه‌دان (Dare Panbedan)		پل آدینان (Pol Adinan)		ساریقمیش (Sarigamish)		سنته (Senteh)		جان‌آقا (Jan Agha)	
روش	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
Lasso (لاسو)	0.07	8.74	0.01	44.11	0.3	10.75	0.08	18.73	0.5	7.51	0.6	31.7
Bayesian (بیزین)	0.0001	11.04	0.04	55.12	0.24	10.68	0.5	15.5	0.14	10.39	0.1	49.09
KNN	0.9	1.72	0.63	25.8	0.99	0.98	0.91	4.07	0.75	5.51	0.9	1.72

[DOI: 10.61882/jwmr.2025.1310]

نتیجه‌گیری کلی

با توجه به نتایج به‌دست آمده، نمودار جعبه‌ای داده‌هایی را که خارج از ساقه قرار گیرند به‌عنوان داده پرت معرفی می‌کند و بر همین اساس در نمودارهای جعبه‌ای تعداد داده‌های پرت در مقایسه با سایر روش‌ها به مقدار زیادی تشخیص داده می‌شود که به‌نظر می‌رسد روش مناسبی برای تشخیص داده پرت در داده‌های هیدرولوژیکی نباشد. روش KNN در تعیین داده‌های گمشده با استفاده از داده‌های مشاهداتی متناظر، در بین دو روش دیگر بسیار مؤثر عمل نموده است و در نتیجه، نیاز به نرمال‌سازی دارند. در این مطالعه، سری داده‌ها نرمال‌سازی و سپس مقادیر داده‌های پرت در آن‌ها محاسبه گردید و جهت تعیین مقادیر محاسبه نشده و گمشده از روش KNN استفاده شد. یکی از معایب این روش این است که دقت آن در مناطقی که داده‌های یکنواختی ندارند کمتر از حد مورد انتظار بوده است. در داده‌های دارای روند تغییرات کمتر، KNN بسیار دقیق عمل می‌نماید و یکی از دقیق‌ترین و مطمئن‌ترین روش‌های نسبت‌دهی و جایگذاری داده‌های گمشده است. روش KNN کارایی مطلوبی را در تخمین مقادیر گمشده در جریان‌های پیوسته و ناپیوسته نسبت به دو روش دیگر ارائه می‌دهد. این اثربخشی به توانایی KNN در دستیابی به مقدار بهینه نزدیکترین همسایه برمی‌گردد که آن را برای پیش‌بینی دقیق در شرایطی که جریان به حداقل رسیده باشد هم مناسب می‌سازد. دقت KNN به‌دلیل سادگی محاسبات و نیز اثر بالای آن در محاسبه و نسبت‌دهی داده‌های گمشده و گمشده است که در عین حال ساختار سری داده را نیز حفظ می‌کند. در موارد خاص، این روش به داده پرت و وجود نویز حساس است که ممکن است بر اثربخشی آن تاثیر منفی بگذارد. وجود ناپایداری در سری داده و وجود تغییرات در آن و همین طور نبود داده در مابین سری داده‌های پیوسته باعث کاهش دقت رگرسیون لاسو می‌شود. این نوع رگرسیون برای داده‌های کاملا پیوسته که با هدف ایجاد پایداری در مدل مورد استفاده قرار می‌گیرند نتایج خوبی می‌تواند ارائه نماید اما در سری داده جریان رودخانه دقت پایینی دارد که با نتایج (Li *et al.,* 2020; Maanavi & Roozbeh, 2021) تطابق دارد.

References

Ahmadi, F., Dinpajoh, Y., & Fard, A. F. (2014). Comparing linear and nonlinear time series models in river flow forecasting (case study: Baranduz-chai river). *Irrigation Sciences and Engineering*, *37*(1), 93-105. [In Persian]

Aryanmanesh J, N. H., Mahmoodi P, Khosravi P. (2024). Reconstruction of Missing Daily Streamflow Data using the MissForest Algorithm in Southern Baluchestan Basin, Iran. *Journal of Watershed Management Research*, *15*(2), 49-64. [In Persian]

Azimi-Habashi, S., Miryaghoubzadeh, M., Erfanian, M., & Javan, K. (2024). Projection of Future Climatic Variables based on CMIP5 and CMIP6 Models in the Gedarchay Catchment (West Azarbaijan). *Journal of Watershed Management Research*, *15*(2), 1-16. https://doi.org/doi:10.61186/jwmr.15.2.1. [In Persian]

Bae, I., & Ji, U. (2019). Outlier detection and smoothing process for water level data measured by ultrasonic sensor in stream flows. *Water*, *11*(5), 951. https://doi.org/(doi.org/10.3390/w11050951

Bahrami, M., Amiri, M.J., Rezaei Maharlouyi, F., & Ghaffari, K. (2018). Determining the effect of data preprocessing on the performance of artificial neural networks for predicting monthly precipitation in Abadeh County. *Eco-Hydrology*, *4*(1), 29-37. [In Persian]

Ben-Gal, I. (2005). Outlier detection. *Data Mining and Knowledge Discovery Handbook*, 131-146. https://doi.org/doi.org/10.1007/0-387-25465-X_7

Boiten, W. (2003). *Hydrometry: IHE Delft lecture note series*. CRC press. https://doi.org/doi.org/10.1201/9780203971093

Boukerche, A., Zheng, L., & Alfandi, O. (2020). Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)*, *53*(3), 1-37. https://doi.org/doi.org/10.1145/3381028

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD. *International Conference on Management of Data*.

Cohn, T. A., England, J., Berenbrock, C., Mason, R., Stedinger, J., & Lamontagne, J. (2013). A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. *Water Resources Research*, *49*(8), 5047-5058. https://doi.org/doi.org/10.1002/wrcr.20392

D'Agostino, R. B. (1986). *Goodness-of-fit-techniques* (Vol. 68). CRC press.

Dave, D., & Varma, T. (2014). A review of various statistical methods for outlier detection. *International Journal of Computer Science & Engineering Technology (IJCSET)*, *5*(2), 137-140.

Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, *157184*.

Fenton, J. D., & Keller, R. J. (2001). The calculation of streamflow from measurements of stage.

Goldstein, M., & Dengel, A. (2012). Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012:Poster and Demo Track*, *1*, 59-63.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1-21.

Herschy, R. W. (2008). *Streamflow Measurement*. CRC press.

[DOI: 10.61882/jwmr.2025.1310]

Holmström, H., & Fransson, J. E. (2003). Combining remotely sensed optical and radar data in k NN-estimation of forest variables. *Forest Science*, *49*(3), 409-418. https://doi.org/doi.org/10.1093/forestscience/49.3.409

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H., & Pierrefeu, G. (2018). Impact of stage measurement errors on streamflow uncertainty. *Water Resources Research*, *54*(3), 1952-1976. https://doi.org/doi.org/10.1002/2017WR022039

Kiani, R. a. M., M. . (2015). A review of outlier detection methods. *International Conference on Research in Science and Technology. 14 December 2015, Kualalumpur, Malaysia*. [In Persian]

Li, Q., Fisher, K., Meng, W., Fang, B., Welsh, E., Haura, E. B., Koomen, J. M., Eschrich, S. A., Fridley, B. L., & Chen, Y. A. (2020). GMSimpute: a generalized two-step Lasso approach to impute missing values in label-free mass spectrum analysis. *Bioinformatics*, *36*(1), 257-263. https://doi.org/doi.org/10.1093/bioinformatics/btz488

Maanavi, M., & Roozbeh, M. (2021). Regression Analysis Methods for High-dimensional Data. *Andishe _ye Amari*, *25*(1), 69-90. [In Persian]

Montgomery, D. C., & Runger, G. C. (2019). *Applied Statistics and Probability For Engineers*. John wiley & sons.

Naghdi, R., Shayannezhad, M., & Sadati, N. S. (2010). Comparison of different methods for estimating of monthly discharge missing data in Grand Karoon River Basin. [In Persian]

Nazeri Tahrudi, M. (2014). Compared to the normal mechanism becomes the normal monthly rainfall data from different regions of Iran. *Water and Soil*, *28*(2), 365-372. [In Persian]

Ordooni, M., Memarian, H., Akbari, M., & Pourreza, M. (2021). Evaluation and Comparison of GPM Satellite Precipitation Data with Meteorological Station using Kolmogorov-Smirnov Test. *Iranian Journal of Rainwater Catchment Systems*, *9*(2), 11-24. [In Persian]

Poursalehi, F., Shahidi, A., & Khashei Siuki, A. (2019). Comparison of decision tree m5 and k-nearest neighborhood algorithm models in the prediction of monthly precipitation (case study: birjand synoptic station). *Iranian Journal of Irrigation & Drainage*, *13*(5), 1283-1293. [In Persian]

Rahmdel, M., Mohamadian, A., Javanshiri, Z., & Sanaeinejad, S. (2021). Exploratory analysis and in-homogeneity study of temperature and rainfall series of meteorological stations in Iran (period 1989-2018). [In Persian]

Rajabi Jaghargh, M., Mousavi Baygi, S. M., Araghi, S. A., & Jabari Noghabi, H. (2024). Calibration of ERA5 daily precipitation using MLP, D-Tree, and KNN algorithms in Razavi Khorasan province. *Iranian Journal of Rainwater Catchment Systems*, *12*(1), 129-147. [In Persian]

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147.

Shataee, S., Kalbi, S., Fallah, A., & Pelz, D. (2012). Forest attribute imputation using machine-learning methods and ASTER data: comparison of k-NN, SVR and random forest regression algorithms. *International Journal of Remote Sensing*, *33*(19), 6254-6280. https://doi.org/doi.org/10.1080/01431161.2012.682661

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, 100306. https://doi.org/doi.org/10.1016/j.cosrev.2020.100306

Suri, N. M. R., Murty, M. N., & Athithan, G. (2019). *Outlier detection: Techniques and Applications*. Springer. https://doi.org/doi.org/10.1007/978-3-030-05127-3

Tourian, M., Schwatke, C., & Sneeuw, N. (2017). River discharge estimation at daily resolution from satellite altimetry over an entire river basin. *Journal of Hydrology*, *546*, 230-247. https://doi.org/doi.org/10.1016/j.jhydrol.2017.01.009

Umar, N., & Gray, A. (2023). Comparing single and multiple imputation approaches for missing values in univariate and multivariate water level data. *Water*, *15*(8), 1519. https://doi.org/doi.org/10.3390/w15081519

[DOI: 10.61882/jwmr.2025.1310]