پیش بینی جریان رودخانه با استفاده از روش K - نزدیک‌ترین همسایگی

کبری مومنی [۱]، محمد رضا یزدانی [۲]، علی اصغر ذوالفقاری [۲]، محمد رحیمی [۲]، سیدحسن کابلی [۲]
۱-دانشجوی کارشناسی ارشد بیابان زدایی دانشگاه سمنان، ۲- استادیاران دانشکده کویرشناسیدانشگاه سمنان

چکیده
دستیابی به روش‌های مطمئن پیش بینی جریان رودخانه‌ها به منظور برنامه ریزی در بهره برداری به موقع از منابع آب از اهمیت روزافزونی برخوردار است. با توجه به اهمیت پیش بینی جریان رودخانه در مدیریت منابع آب روش‌ های مختلفی برای مدل کردن جریان رودخانه‌ها بکار برده می‌شوند. در این تحقیق از روش غیر پارامتریک KNN برای پیش بینی جریان رودخانه‌های چناران و درکش در استان خراسان شمالی استفاده شده است. برای دستیابی به بهترین پیش بینی ۶ سناریو با استفاده از روش KNN مورد آزمون قرار گرفته است که با توجه به کمترین مجذور میانگین مربعات خطا (RMSE) و بیش‌ترین کارایی مدل (EF) و ضریب همبستگی (R)، بهترین سناریو برای پیش بینی استخراج گردید.

واژه‌های کلیدی: K - نزدیک‌ترین همسایگی، جریان رودخانه، سناریو.

مقدمه
افزایش تقاضا برای آب در مناطق مختلف، نیاز به مدیریت بهینه منابع آب را بیش از پیش نشان می‌دهد. پیش بینی دقیق و به موقع منابع آب در دسترس را می‌توان نکته کلیدی در برنامه ریزی، مدیریت و بهره برداری بهینه از منابع آب قلمداد کرد. بر این پایه دستیابی به روش‌های مطمئن پیش بینی جریان رودخانه‌ها به منظور برنامه ریزی در بهره برداری به موقع از منابع آب از اهمیت روزافزونی برخوردار است (فتاحی، ۱۳۹۱). جهت طراحی بهینه و قابل اعتماد سیستم هایی مانند مخازن سدها و برنامه ریزی کارآمد در مدیریت منابع آب، تخمین صحیح آبدهی رودخانه اهمیت زیادی دارد (سیواکومار و همکاران، ۲۰۰۲). پیش بینی جریان رودخانه‌ها با توجه به اهمیت آن در طراحی تأسیسات آبی، آبگیری از رودخانه‌ها، برنامه ریزی بهره برداری از مخازن سدها، کنترل فرسایش و رسوب رودخانه‌ها و غیره از دیرباز مورد توجه مهندسان آب بوده است. از سوی دیگر با توجه به محدودیت منابع آب شیرین قابل استحصال، پیش بینی هر چه دقیق‌تر دبی جریان و تغییرات آن در طول رودخانه از ارکان اساسی برنامه ریزی و مدیریت منابع آب‌های سطحی است. تاکنون روابط و الگوهای گوناگون و پیچیده ای برای پیش بینی میزان آبدهی رودخانه‌ها ارائه شده است (دانند مهر، ۱۳۸۹).

یکی از روش‌های که برای تولید داده‌های آب و هوایی مورد استفاده قرار می‌گیرد، روش KNN[۱۰۵] است که برای تولید مصنوعی داده‌های آب و هوایی استفاده شده است. الگوریتم K- نزدیک‌ترین همسایگی، یکی از مهم‌ترین و توسعه یافته ترین رویکردهای غیر پارامتریک می‌باشد که در بسیاری از پژوهش‌های نوین به کار گرفته شده است (جلالی و همایی، ۱۳۹۰).

در تحقیق راجاگوپلان (۱۹۹۹) روش نزدیک‌ترین K- همسایگی (KNN) با مدل سری زمانی پارامتری مقایسه شد که برتری روش K - نزدیک‌ترین همسایه به اثبات رسید. در مطالعه قربانی (۱۳۹۳) مدل KNN نسبت به دو روش ماشین بردار پشتیبان (SVM)[۱۰۶] و درخت تصمیم (M۵) در ریزمقیاس نمایی بارش بر اساس داده‌های مدل گردش عمومی NCEP در ایستگاه سینوپتیک کرمانشاه کارایی بهتری داشته است. عزمی و عراقی نژاد (۱۳۹۰) برای پیش بینی جریان رودخانه در محدوده حوضه بالادست زاینده رود از روش رگرسیون K- نزدیک‌ترین همسایگی استفاده کردند. نتایج آن‌ها نشان داد که KNN بهبود یافته دارای عملکرد بهتری بوده است.

مواد و روش‌ها
در این تحقیق به منظور پیش بینی جریان رودخانه توسط مدل K- نزدیکترین همسایگی، از آمار دو ایستگاه هیدرومتری و ۱۳ ایستگاه بارانسنجی و ۲ ایستگاه تبخیرسنجی در استان خراسان شمالی با طول دوره ی آماری مشترک ۳۰ سال (۱۳۶۰-۱۳۸۹) استفاده شده است. ۷۰ درصد از داده‌ها (۲۱ سال) جهت آموزش مدل و ۳۰ در صد باقی مانده (۹ سال) برای آزمون مدل بکار گرفته شد. همچنین برای دستیابی به بهترین پیش بینی از دبی، سناریو های مختلفی طراحی و مورد آزمون قرار گرفت که در جداول ۱ نشان داده شده است.

جدول ۱- سناریو های مورد آزمون برای رودخانه چناران

[۱۰۵] - K-Nearest Neighbor
[۱۰۶] - Support Vector Machine

ساختار	تابع
S_1	$Q= f(P_1, P_2, P_3, P_4, P_5, P_6, P_7, T_1, T_2, T_3, T_4)$
S_2	$Q= f(P_1, P_2, P_3, P_4, P_5, P_6, P_7)$
S_3	$Q= f(P_1, P_2, P_3, P_4, P_5, P_6, P_7, Q_{(t-1)})$
S_4	$Q= f(P_{1(t-1)}, P_{2(t-1)}, P_{3(t-1)}, P_{4(t-1)}, P_{5(t-1)}, P_{6(t-1)}, P_{7(t-1)})$
S_5	$Q= f(SPI_1, SPI_2, SPI_3, SPI_4, SPI_5, SPI_6, SPI_7, T_1, T_2, T_3, T_4)$
S_6	$Q= f(SPI_1, SPI_2, SPI_3, SPI_4, SPI_5, SPI_6, SPI_7, Q_{(t-1)})$

K- نزدیک‌ترین همسایگی (KNN)

یکی از روش‌های مدل سازی در داده کاوی، الگوریتم K- نزدیک‌ترین همسایگی است. این الگوریتم جزء روش‌های یادگیری نظارت شده است که در پیش بینی استفاده می‌شود. در این تکنیک، تعداد بهینه ای از نمونه‌های موجود در یک مجموعه که دارای شبیه‌ترین ویژگی‌ها به نمونه هدف باشند (K) در نظر گرفته می‌شود و سپس جایابی و طبقه بندی عنصر هدف در این مجموعه از داده‌ها، با تعیین فاصله‌ها و سپس وزن دهی عناصر آموزش دهنده صورت می‌پذیرد.

در یک بانک داده مرجع دبی، نزدیک‌ترین (مشابه‌ترین) دبی‌ها به دبی مورد آزمون انتخاب و این دبی‌ها با توجه به میزان مشابهت به دبی مورد آزمون وزن دهی می‌شوند. به این ترتیب که در ابتدا فواصل اقلیدسی بین دبی مورد آزمون و دبی‌های مرجع محاسبه شده، سپس تعداد K عدد از نزدیک‌ترین دبی‌ها بر اساس فاصله به دست آمده، وزن دهی می‌شوند. در انتها با توجه به وزن هر یک از دبی در مجموعه ای با K عدد دبی (نزدیک‌ترین همسایه)، برآوردی از داده‌های هدف که مورد آزمون واقع شده‌اند، صورت می‌پذیرد (نمس و همکاران، ۲۰۰۶). در این تحقیق فاصله اقلیدسی بین نمونه‌های هدف با هریک از نمونه‌ها در بانک داده مرجع از رابطه زیر بدست آمد:

$$d_i = \sqrt{\sum_{j=1}^{x} \Delta a_{ij}^{\,2}} \qquad (۱)$$

که در آن، di نشان دهنده فاصله i امین دبی از بانک داده مرجع تا دبی هدف می‌باشد. Δa_{ij} : اختلاف فاصله دبی هدف، از i امین دبی و j امین متغیر موجود در بانک داده مرجع می‌باشد و x: تعداد متغیر ورودی نمونه‌های دبی (مانند بارندگی و دما) است. در اینجا منظور از فاصله، میزان شباهت و نزدیکی دو نمونه مورد نظر می‌باشد. به طوری که هرچه مقدار به دست آمده از d کوچک‌تر باشد شباهت دو نمونه مورد نظر بیش‌تر است.

یکی از مشکلات موجود در استفاده از مدل KNN این است که یک واحد اختلاف در یک متغیر ممکن است که به اندازه یک واحد اختلاف در متغیر دیگر اهمیت نداشته باشد. به عنوان مثال در این تحقیق تغییر یک واحد در داده‌های دما به عنوان یک متغیر اهمیت کمتری نسبت به تغییر یک واحد در داده‌های دبی دارد. به همین منظور قبل از هر گونه محاسباتی، ابتدا داده‌های مورد استفاده در تحقیق با استفاده از رابطه ۲ استاندارد شد:

$$a_{ij(temp)} = \frac{[(a_{ij}) - \bar{a}_j]}{\sigma(a_j)} \qquad (۲)$$

که در آن، $a_{ij(temp)}$: مقدار عددی موقت استاندارد شده j امین صفت برای هریک از نمونه‌ها ، a_{ij}: نماینده j امین متغیر از i امین دبی است. $\bar{a}_j$ و $\sigma(a_j)$ به ترتیب میانگین و انحراف معیار مقادیر مشاهده ای از j امین متغیر در بانک داده‌های مرجع می‌باشد.

سپس مقادیر نهایی تبدیل شده j امین متغیر از i امین دبی $a_{ij(trans)}$ با توجه به دامنه کمترین تا بیشترین مقدار $a_{ij(temp)}$ از رابطه ۳ بدست آمد و بجای مقادیر اعداد واقعی به عنوان داده‌های ورودی مورد استفاده قرار گرفت (نمس و همکاران، ۲۰۰۶).

$$a_{ij(trans)} = \frac{\{max[range(a_{j=1(temp)}), \ldots range(a_{j=x(temp)})]\}}{range(a_{j(temp)})} \qquad (۳)$$

در مرحله بعد، وزن دهی نمونه دبی‌های آموزش دهنده (K) بر اساس فاصله‌های به دست آمده بین دبی هدف و هر یک از دبی‌های تعیین شده در K طبق رابطه‌های ۴ صورت گرفت.

$$W_i = \frac{d_{i(rel)}}{\sum_{i=1}^{K} d_{i(rel)}} \quad (۴)$$

که در آن، Wi: وزن i امین نمونه از دبی‌های آموزش دهنده (K) است و di(rel) فاصله نسبی هدف و نمونه i ام می‌باشد که از رابطه ۵ بدست آمده است:

$$d_{i(rel)} = \left[\frac{\sum_{i=1}^{k} d_i}{d_i}\right]^{P} \quad (۵)$$

در تحقیق نمس و همکاران (۲۰۰۶) کمترین مقدار خطای تخمین در P های نزدیک به ۱ بدست آمده است که روش KNN به مقادیر مختلف P، حساسیت اندکی داشته است. در این تحقیق بهترین مقدار P از تکنیک ارزیابی تقاطعی بدست آمد.

پس از وزن دهی نمونه‌ها، برآورد مقادیر دبی در نمونه‌های هدف از طریق محاسبه میانگین وزنی دبی در نمونه دبی‌های K با استفاده از رابطه ۶ صورت گرفت.

$$P_i = \sum_{i=1}^{K} w_i o_i \quad (۶)$$

که در آن، Pi: مقدار دبی پیش بینی شده دبی هدف، Wi: وزن مربوط به i امین دبی در K و Oi: مقدار دبی اندازه گیری شده i امین دبی در K است.

در آخرین مرحله، جهت ارزیابی و مقایسه سناریو های مختلف در مدل از معیارهای مجذور میانگین مربعات خطا (RMSE)[۱۰۷]، کارایی مدل (EF)[۱۰۸] و ضریب همبستگی (R)استفاده شده است.

نتایج و بحث

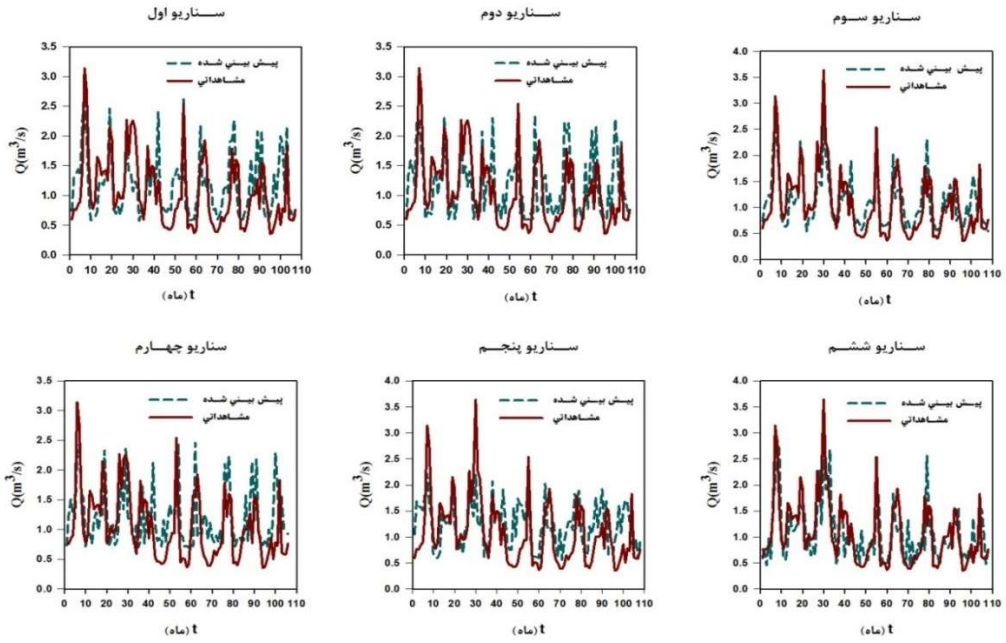
بهینه سازی هم زمان P و K به گونه ای انجام شد که به ازای هر مقدار P از ۱/۰ تا ۲ (با فاصله‌های ۰.۱) K های مختلف از ۱ تا ۳۰ در نظر گرفته شده و مقدار RMSE آن‌ها بدست آمده است. در این تحقیق با توجه به تعداد نمونه‌های بانک مرجع، مقدار خطای برآورد در دامنه ای از مقادیر مختلف K محاسبه شد و بر اساس روش ارزیابی تقاطعی، بهترین تعداد برای K، زمانی که خطای تخمین‌ها (RMSE) به کم‌ترین مقدار خود رسید تعیین گردید. نتایج تحقیق نشان داد، حساسیت روش KNN به مقادیر P اندک می‌باشد و تغییر مقدار P باعث تغییر بسیار اندکی در RMSE شده است. بنابرین مقدار P در این تحقیق ۱ در نظر گرفته شد.

نتایج بدست آمده برای پیش بینی جریان رودخانه‌های چناران و درکش در جدول (۲) و شکل‌های (۱ و ۲) زیر نشان داده شده است.

[۱۰۷]- Root Mean Square Error
[۱۰۸]- Efficiency of Model

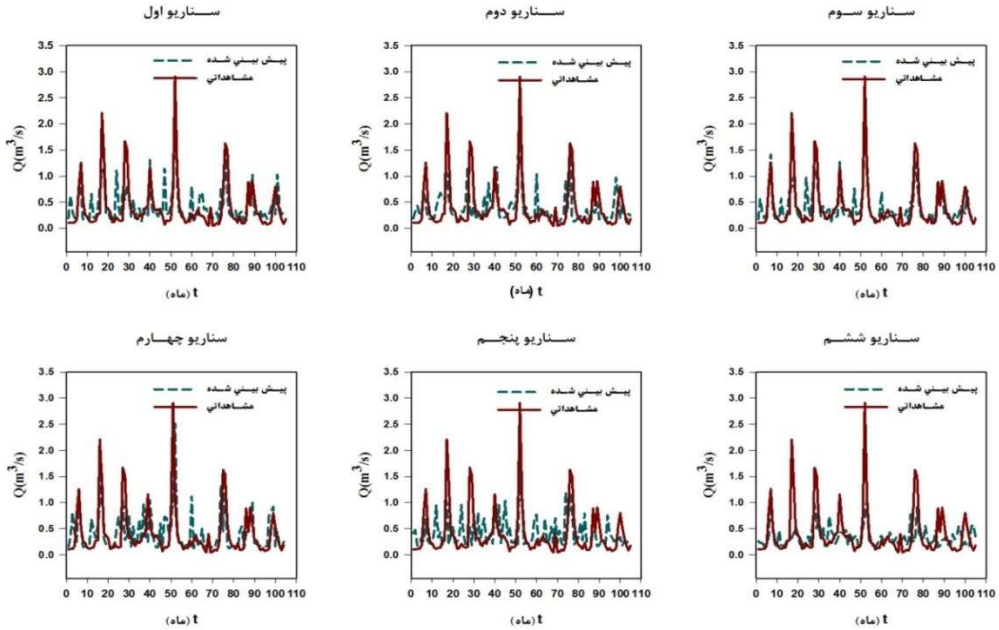
جدول ۲- نتایج سناریو های مختلف برای پیش بینی جریان رودخانه های چناران و درکش

رودخانه چناران					رودخانه درکش				
سناریو	K value	RMSE (m^3/s)	EF	R	سناریو	K value	RMSE (m^3/s)	EF	R
۱	۲۵	۵۰/۰	۳۶/۰	۶۵	۱	۱۳	۶۷/۰	۳۵/۰	۶۴/۰
۲	۲۵	۵۴/۰	۲۶/۰	۵۹/۰	۲	۹	۷۲/۰	۲۷/۰	۵۵/۰
۳	۱۳	۳۸/۰	۶۳/۰	۷۹/۰	۳	۱۴	۶۳/۰	۴۴/۰	۷۴/۰
۴	۲۷	۵۹/۰	۱۱/۰	۴۷/۰	۴	۱۱	۷۲/۰	۲۷/۰	۵۳/۰
۵	۲۶	۵۶/۰	۲۱/۰	۵۴/۰	۵	۱۴	۷۵/۰	۱۹/۰	۴۶/۰
۶	۱۳	۴۷/۰	۴۴/۰	۶۹/۰	۶	۱۴	۷۸/۰	۱۴/۰	۴۵/۰



شکل ۱- نمودارهای نتایج مدل KNN برای سناریو های مختلف در پیش بینی جریان رودخانه چناران

پس از بررسی ساختارهای مختلف ورودی برای پیش بینی دبی رودخانه چناران و محاسبه آماره‌های EF ،RMSE و R، با توجه به جدول ۲ مشخص گردید که سناریو سوم بهترین پیش بینی را برای دبی داشته است. بنابرین می‌توان گفت جریان رودخانه چناران در یک ماه مشخص بیشتر به بارندگی در همان ماه و دبی در یک ماه قبل بستگی داشته است.



شکل ۲- نمودارهای نتایج مدل KNN برای سناریو های مختلف در پیش بینی جریان رودخانه درکش

نتایج مربوط به شکل ۲ و جدول ۲ نشان می‌دهد که در سناریو سوم، مجذور میانگین مربعات خطا (RMSE) ۶۳/۰، کارایی مدل (EF) ۴۴/۰ و ضریب همبستگی (R) ۷۴ درصد بوده است. بنابرین سناریو سوم بهترین پیش بینی را برای جریان رودخانه درکش داشته است. در نتیجه جریان رودخانه درکش در هر ماه به بارندگی همان ماه و دبی ماه قبل وابستگی بیشتری داشته است.

منابع

جلالی، و. و همایی، م. ۱۳۹۰. ارائه مدلی غیرپارامتریک با استفاده از تکنیک k- نزدیک‌ترین همسایه در برآورد جرم مخصوص ظاهری خاک. مجله علوم آب و خاک-علوم و فنون کشاورزی و منابع طبیعی. جلد ۱۵ شماره ۵۶ صفحات ۱۸۱-۱۹۱.

داننده مهر، ع. مجدزاده طباطبائی، م. ۱۳۸۹. بررسی تأثیر توالی دبی روزانه در پیش بینی جریان رودخانه‌ها با استفاده از برنامه ریزی ژنتیک. نشریه آب و خاک (علوم و صنایع کشاورزی). جلد ۲۴، شماره ۲، صفحه های ۳۲۵-۳۳۳.

عزمی،م. و عراقی، ش. ۱۳۹۰. توسعه روش رگرسیون K- نزدیکترین همسایگی در پیش بینی جریان رودخانه. نشریه آب و فاضلاب، شماره ۲، صفحه های ۱۰۸-۱۱۹.

فتاحی، ا. دلاور، م. نوحی، ک. ۱۳۹۱. پیش بینی جریان رودخانه های کارون شمالی با استفاده از شبکه عصبی مصنوعی. فصلنامه تحقیقات جغرافیایی، جلد ۲۷. صفحه های ۵۱-۷۷.

قربانی، خ. ۱۳۹۳. ارزیابی مدل های داده کاوی در ریز مقیاس نمایی بارش بر اساس داده‌های مدل گردش عمومی NCEP (مطالعه موردی: ایستگاه سینوپتیک کرمانشاه). مجله پژوهش آب ایران. سال ۸، شماره ۱۵، صفحه های ۱۷۷-۱۸۶.

El-Shafie A, RedaTaha M and Noureldin A, ۲۰۰۷. A neuro-fuzzy model for inflow forecasting of the Nile river at Aswan high dam. Water Resources Management ۲۱: ۵۳۳-۵۵.

Nemes, A., Rawls, W.J., and Pachepsky, Ya.A. ۲۰۰۶. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. ۷۰: ۳۲۷-۳۳۶.

Rajagopalan, B. Lall, U. ۱۹۹۹. A k-nearest-neighbor simulator for daily precipitation and other weather variables. WATER RESOURCES RESEARCH, VOL. ۳۵, NO. ۱۰, PAGES ۳۰۸۹-۳۱۰۱.

Sivakumar B, Jayawardena AW and Fernando TMKG, ۲۰۰۲. River flow forecasting: use of phase-space reconstruction and artificial neural networks approaches. Journal of Hydrology ۲۶۵: ۲۲۵-۲۴۵.

Zahoor J. Abrar M. Bashir Sh. And Mirza A. ۲۰۰۹. Seasonal to inter-annual climate prediction using data mining KNN technique. Wireless Networks, Information Processing and Systems, Communications in Computer and Information Science. ۲۰:۴۰-۵۱.

Abstract

Access to reliable methods to predict river flow is very important to timely utilization of water resources planning. Given the importance of predictability of river flow in water resources management, different methods is used to river flow modeling. In this study, non-parametric methods of K Nearest neighbor (KNN) was used to predict river flow in two basins of Chenaran and Darkesh in the Northern Khorasan province. To achieve the best forecast, ۶ scenarios were analyzed using the KNN method. The best scenario was selected according to the lowest root mean square error (RMSE), the highest efficiency of model (EF) and correlation coefficient (R) of the best scenarios to prediction river flow.