علوم و مهندسی آبیاری
جلد ۴۰، شماره‌ی٤، زمستان ۹۶، ص. ۱۹۵-۱۸۳

تحلیل منطقه‌ای سیلاب با مقایسه مدل‌های الگوریتم درخت تصمیم‌گیری M5 و رگرسیونی

حسن اسمعیلی گیساوندانی[۱]، علی محمد آخوند علی[۲*]، حیدر زارعی[۳] و مهرداد تقیان[۴]

۱- دانشجوی کارشناسی ارشد هیدرولوژی ومنابع آب دانشکده علوم و مهندسی آب دانشگاه شهید چمران اهواز.
۲*- نویسنده مسئول، استاد دانشکده علوم و مهندسی آب، دانشگاه شهید چمران اهواز.aliakh@scu.ac.ir
۳- استادیار دانشکده علوم و مهندسی آب، دانشگاه شهید چمران اهواز.
۴- استادیار، عضو هیئت علمی دانشگاه کشاورزی و منابع طبیعی رامین.

تاریخ دریافت:۹۵/۴/۲۵ تاریخ پذیرش: ۹۵/۷/۲۸

چکیده

ارزیابی فراوانی سیلاب در حوضه‌های فاقد ایستگاه‌های اندازه‌گیری، معمولاً توسط ایجاد روابط مناسب آماری (مدل‌ها)بین سیلاب و ویژگی‌های فیزیکی حوضه انجام می‌گیرد. تاکنون معادله‌های متعددی در زمینه برآورد دبی‌سیلاب در مناطق مختلف ارایه شده ولی با توجه به پیچیدگی این پدیده، روابط موجود نتوانسته‌اند دبی سیلاب طراحی را با دقت مناسب شبیه‌سازی کنند. بر همین اساس در این پژوهش علاوه بر روش‌رگرسیونی که در گذشته استفاده می‌شد از مدل درخت تصمیم‌گیری M5 استفاده شده‌است. روش‌های مختلف داده‌کاوی در علوم آب به دلیل دقت بالا آن گسترش فراوانی یافته‌است. مشخصات فیزیوگرافی حوضه توسط نرم‌افزار Arc GIS محاسبه شد. سپس کلیه پارامترهای فیزیوگرافی به همراه دوره بازگشت به عنوان داده‌های ورودی الگوریتم M5 و رگرسیون‌خطی لحاظ شد. نتایج حاصل از بررسی آماره‌های صحت‌سنجی نشان داد بر اساس ضریب همبستگی بین آمار برآورد شده و مشاهداتی، همچنین براساس معیارهای جذر میانگین مربعات خطا و میانگین مطلق خطا، الگوریتم M5 عملکرد بهتری نسبت به روش رگرسیون در برآورد دبی‌سیلاب دارد.

کلید واژه ها: تحلیل منطقه‌ای سیلاب، دبی سیلاب، درخت تصمیم، الگوریتم M5، مدل رگرسیون.

مقدمه

سیل از جمله پدیده‌هایی است که هر ساله خسارت‌های جبران‌ناپذیری به اقتصاد کشورها وارد می‌کند. لذا پژوهش در خصوص ویژگی‌های این پدیده طبیعی از اهمیت بالایی برخوردار است. آگاهی از میزان و تواتر دبی سیلابی با دوره برگشت‌های مختلف از موارد ضروری در طراحی سازه‌های آبی می‌باشد، اما اغلب در محل احداث سازه‌ها با فقدان ایستگاه اندازه‌گیری ویا کمبود آمار مواجه هستیم که در این صورت برآورد مطمئن دبی سیلابی امری الزامی است. از پارامترهای مهم در معرفی سیلاب‌ها، حداکثر دبی با دوره برگشت‌های مختلف می‌باشد که همیشه در طراحی سازه‌های آبی تخمین مناسبی از آن مورد نیازبوده و از نظر اقتصادی نقش تعیین‌کننده‌ای در ابعاد سازه‌ای ایفا می‌کند. در بسیاری از حوضه‌ها خصوصیات فیزیوگرافی نقش موثری در تولید دبی سیلاب ایفا می‌کند، این ویژگی‌ها با گذشت زمان‌های نسبتا طولانی ثابت بوده و لذا می‌توان از آنها به عنوان پارامتر مستقل برای تخمین سیلاب به خصوص در مناطقی که فاقد ایستگاه هیدرومتری[۱] هستند، استفاده نمود و این ویژگی‌ها در حوضه‌های فاقد آمار مهم‌ترین نقش را در برآوردهای هیدرولوژی دارند. تحلیل منطقه‌ای سیلاب روشی است که در حوضه‌های فاقد آمار جریان با استفاده از آمار جریان سیلاب در حوضه‌های مجهز به ایستگاهای هیدرومتری روابطی بین مقادیر جریان سیل و برخی از ویژگی‌های مورفولوژی، فیزیوگرافی، اقلیمی، زمین‌شناسی، خاک‌شناسی، پوشش‌گیاهی و نحوه کاربری اراضی حوضه‌ها برآورد می‌کند. در واقع داده‌های موجود و محدود منطقه مورد نظر به روش‌های مختلف برای تمام منطقه تعمیم می‌دهد در این راستا روش‌های گوناگونی معرفی شده‌است از جمله: روش سیل شاخص[۲] رگرسیون چند متغیره[۳]، شبکه‌های مربعی[۴] و روش هیبرید[۵] را می‌توان نام برد (علیزاده،۱۳۹۲).

از روش‌های فوق، روش رگرسیون چند متغیره بارها در مقالات و گزارش‌های ایران و جهان بیشتر به کار رفته و ارائه شده‌است. حال اینکه در دهه‌ی حاضر شاهد سیستم‌های هوشمند و کاربردش در مدل‌های هیدرولوژیکی هستیم. اصلی‌ترین مزیت نگرش سیستم‌های هوشمند نسبت به روش‌های سنتی این است که سیستم‌های هوشمند نیازی به توضیح صریح طبیعت پیچیده

1- Ungauged catchment

2- Index Flow Method(IFM)
3- Multivariate Regression Method (MRM)
4- Square Grids Method(SGM)
5- Hybrid Method(HM)

فرایندها به‌صورت ریاضی ندارد، درخت‌های تصمیم‌گیری نمونه‌ای از سیستم‌های نوین و هوشمند می‌باشند (سادهر و همکاران[1]،۲۰۰۲)که در این تحقیق مورد استفاده قرار گرفته‌است. در سطح ملی، چاوشی و اسلامیان[2] (۱۹۹۹) نشان دادند، در دوره‌های بازگشت پایین، دقت مدل هیبریدی بیش‌تر از روش رگرسیونی است. نساجیان‌زواره و همکاران[3] (۲۰۱۱) نشان دادند بهترین دبی شاخص، دبی با دوره بازگشت دو سال است. شادمانی و همکاران (۱۳۹۰) نشان دادند بهترین ساختار، شبکه عصبی مصنوعی پیشخور با دو لایه پنهان به ترتیب دارای پنج و چهار عنصر پردازشگر می باشد.

درخت M5 اولین بار توسط کوینلان[4](۱۹۹۲) مطرح شد و سپس این ایده توسط وانگ و وایتن[5](۱۹۹۷) در قالب روشی بهبود یافت و به M5′ نامگذاری شد.

در همین راستا بهاتاچاریا و سولماتین[6](۲۰۰۵)از دو روش شبکه عصبی و الگوریتم M5 درخت تصمیم برای بررسی رابطه سطح آب ـ دبی در یک رودخانه استفاده و نشان دادند که دقت پیش‌بینی الگوریتم M5 درخت تصمیم بسیار بالا است. همچنین اعتماد شهیدی و بنکدار[7] (۲۰۰۹) به بررسی پیش‌بینی بالاروی موج روی موج شکن سنگریزه‌ای با استفاده از الگوریتم M5′ درخت تصمیم پرداختند و با فرمولی تجربی مقایسه نمودند. نتایج نشان دادند با وجود این که پارامترهای ورودی مدل درختی و فرمول تجربی یکسان می‌باشند، ولی دقت مدل درختی به مراتب بالاتر از مدل تجربی می‌باشد. اعتماد شهیدی و محجوبی (۲۰۰۹) در پیش‌بینی ارتفاع موج در دریاچه سوپریور دو روش M5′ مدل درخت و شبکه عصبی مصنوعی را مقایسه کردند. آن‌ها در مطالعه خود سرعت باد را به عنوان متغیر ورودی و ارتفاع موج اصلی را متغیر خروجی انتخاب نمودند. نتایج حاصل نشان دادند که خطای هر دو مدل مشابه بودند، اما دقت مدل درخت بیش‌تر بوده است.

تاکنون در زمینه تحلیل منطقه‌ای سیلاب مطالعات متعددی با روش‌های متنوع انجام پذیرفته است که از جمله آن‌ها می‌توان به پژوهش انجام شده توسط پژوهشگران زیر اشاره کرد اما الگوریتم M5 کمتر مورد توجه بوده است.

هینز و استدینگر[8](۱۹۹۸)با شبیه سازی ۱۴۵حوضه در مونت‌کارلوی آمریکا به این نتیجه رسیدند که استفاده از رگرسیون غیرخطی به جواب‌های مناسب و واقعی‌تری نسبت به تحقیقات قبلی در حوضه مذکور منتهی می‌شود.

دیمیتری و همکاران[9](۲۰۰۴)با استفاده از الگوریتم M5 و همچنین شبکه عصبی به پیش‌بینی دبی رودخانه هوآی در چین پرداخت و برای هر زیر حوضه از حوضه مذکور معادله‌هایی ارایه کردند و همچنین به بررسی مقدار پیش‌بینی دبی با مقدار مشاهداتی پرداختند.

داوسون[10]و همکاران (۲۰۰۶) با بررسی شبکه عصبی مصنوعی برای مناطق فاقد داده برای حوضه‌ای در انگلستان دبی حداکثر را با دوره برگشت‌های ۱۰،۲۰و۳۰ ساله برآورد کردند ونتیجه‌گرفتند که با دوره برگشت‌های کمتر جواب‌های دقیق‌تری بدست می‌آید.

ثروتی و قنبری (۱۳۸۶) در تحقیقی به برآورد سیلاب در حوضه رودخانه وربند لارستان پرداختند. آنها در این تحقیق ابتدا با استفاده از تحلیل خوشه‌ای به تعیین مناطق همگن پرداختند که براین اساس حوضه مربوطه را به دو منطقه همگن تقسیم و سپس برای هر منطقه همگن معادله‌هایی به ازای دوره بازگشت‌های مختلف برای محاسبه دبی سیلاب ارایه کردند.

شادمانی و همکاران(۱۳۹۰) در تحقیقی به مدلسازی منطقه‌ای دبی سیلابی در استان همدان با استفاده از شبکه عصبی مصنوعی پرداختند. در این تحقیق با توجه به مراحل آموزش، اعتبارسنجی و آزمون، نتایج بدست آمده مشخص شد که در منطقه مورد مطالعه شبکه عصبی مصنوعی پیشخور با دو لایه پنهان به ترتیب دارای چهار و پنج عنصر پردازشگر بود.

رسول‌زاده و همکاران(۱۳۹۴) در نه ایستگاه همگن در مرکز استان اردبیل اقدام به ایجاد و بررسی مدل‌های مختلف تحلیل منطقه‌ای تناوب سیلاب تابعی از دوره بازگشت کردند، آنها در این تحقیق به مدل بندی به چهار صورت پرداختند و با توجه به جذر میانگین مربعات خطا کمتر مدل، مدل مساحت - شیب -طول آبراه‌ها به‌دست آمد که این مدل نسبت به مساحت، مساحت -شیب و مدل فولر در برآورد دبی سیلاب دقیق‌تر می‌باشد. آنها در مدل‌بندی‌های خود برای حصول به مدل واحد و کاستن تعداد مدل‌ها، از دوره بازگشت به عنوان عامل مستقل در مدل در نظر گرفته شد. هدف از این تحقیق تحلیل منطقه‌ای سیلاب با استفاده از روش رگرسیونی به عنوان عمومی‌ترین روش موجود و مقایسه آن با مدل‌سازی الگوریتم درخت تصمیم‌گیری M5 است. این نخستین بار است که از روش الگوریتم M5 مدل درختی برای تحلیل منطقه‌ای سیلاب استفاده می شود. در این تحقیق زمانی که به مدل‌سازی از طریق رگرسیون پرداخته می‌شود بعد از به‌دست آوردن مناطق همگن در هر منطقه دبی سیلاب را بعنوان عامل وابسته و مشخصات فیزیوگرافی به‌عنوان عامل مستقل در نظر گرفته شده‌است.

در مدل‌سازی توسط درخت تصمیم‌گیری M5 نیزهمانند مدل‌سازی رگریسونی برای رسیدن به مدل واحد و کاستن تعداد

1- Sadheer *et al.*
2 - Chavoshi and Eslamian
3 - Nassajian Zavareh *et al.*
4- Quinlan
5-Wang and Witten
6- Bhattacharya and Solomatine
7 - Etemad-Shahidi and Bonakdar
8 -Heinz ,D .F. and Stedinger.
9 -Dimitri P. *et al.*
10- Dawson

مدل‌ها، دوره بازگشت به عنوان عامل مستقل در مدل در نظر گرفته شد.(رسول‌زاده و همکاران،۱۳۹۴ و هاشمی،۱۳۸۲).

مواد وروش‌ها

منطقه مورد مطالعه

تحقیق حاضر در حوضه آبخیز رود کرخه انجام گرفت. این حوضه به وسعت حدود ۵۱ هزار کیلومتر مربع، بین۴۶درجه و ۵۷ دقیقه تا ۴۹ درجه و ۱۰ دقیقه طول شرقی و ۳۱ درجه و ۴۸ دقیقه تا ۳۴ درجه و ۵۸ دقیقه عرض شمالی واقع شده و شامل استانهای همدان، کرمانشاه، کردستان، ایلام، لرستان و خوزستان است (شکل ۱). کرخه رودخانه‌ای است که در جنوب غربی ایران در استان خوزستان جریان دارد. این رودخانه از مناطق میانی و جنوب غربی رشته کوه‌های زاگرس در نواحی غرب و شمال غرب کشور سرچشمه گرفته و پس از طی مسافتی در حدود ۹۰۰ کیلومتر در امتداد شمال به جنوب، سرانجام در مرز مشترک ایران و عراق به مرداب هورالعظیم می‌رسد. این رودخانه پس از رودخانه‌های کارون و دز سومین رودخانه بزرگ ایران از نقطه نظر آبدهی محسوب می‌شود. رودخانه کرخه از شمال به سوی جنوب جریان دارد و در ۴۰ کیلومتری شمال اهواز مسیر آن تغییر کرده و وارد عراق می‌شود. سرشاخه‌های اصلی تشکیل دهنده رودخانه کرخه، رودخانه‌های سیمره، کشکان، قره‌سو، گاماسیاب و چرداول هستند ویکی از مشخصه‌های طبیعی رودخانه کرخه احتمال وقوع سیلاب و خطرات ناشی از آن است(بی‌نام، ۱۳۸۴). که اهمیت مطالعه رفتار جریان این حوزه را نشان می‌دهد.

مدل درختی M5

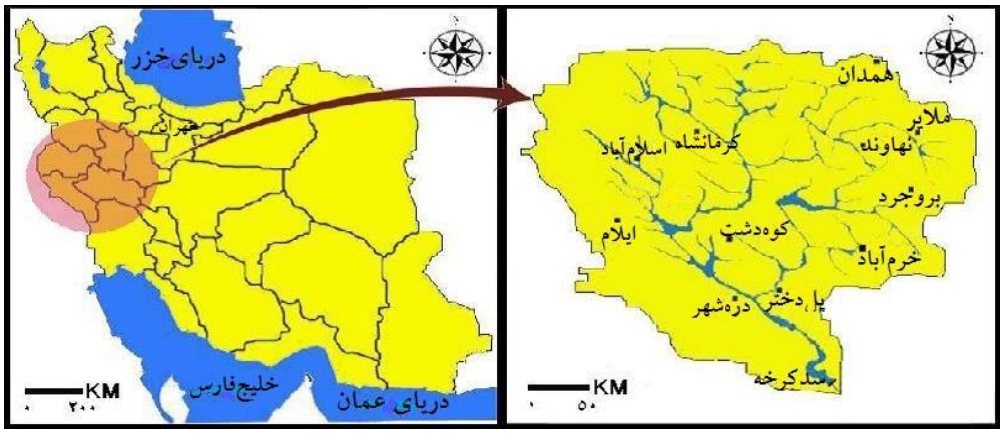
مدل درختی M_5، توسعه، ایده و مفهوم درختان طبقه بندی و رگرسیونی است که با یک ساختاردرختی وارونه که شامل یک گره ریشه در بالاترین قسمت درخت، که به گره‌های دیگر و برگ‌ها منشعب می‌شود به صورت نمایشی و در غالب قوانین اگر- آنگاه نشان داده می‌شود. این مدل قادر به استخراج دانش به شکل روابط ریاضی از مجموعه داده‌ها است.

ایده‌ای که برای ساخت این مدل به کار می‌رود بر این اساس است که یک مسئله مدل سازی چند متغیره را با تقسیم آن به چند زیر مسئله کوچک‌تر و ترکیب نتایج آن تحلیل می‌کند. برای این منظور، فضای مسئله به زیر دامنه‌هایی تقسیم شده و برای هر زیر دامنه یک مدل رگرسیون خطی چند متغیره برازش داده می‌شود، در واقع با این روش یک مجموعه از مدل‌ها خواهیم داشت که هر کدام فقط برای یک قسمت از دامنه مسئله کاربرد دارند. بنابراین یک مدل خطی تکه‌ای به منظور تقریب ارتباط بین متغیرهای ورودی و خروجی فرمول بندی می‌شود. این الگوریتم در مقایسه با دیگر گزینه‌های مشابه مانند رگرسیون نواری شناخته شده‌تر است. این الگوریتم جداسازی‌های ممکن را در فضای چند متغیره انجام داده و به طور خودکار مدل‌هایی برای این دامنه‌ها می‌سازد، در نتیجه یک درخت مرتبه‌ای با جداسازی قوانین در گره‌های داخلی، و خروجی آن، در برگ‌ها خواهیم داشت. اساس مدل درختی روش تقسیم و غلبه صفات برای نمونه‌هایی است که به یک گره می رسند. در ابتدا مدل درختی با تقسیم کردن فضای مساله به صورت برگشتی یک درخت رگرسیونی می‌سازد. در این الگوریتم برای ایجاد شاخه در یک گره تقسیم از پارامتر انحراف معیار مقادیر متغیر هدف به عنوان یک معیار اندازه‌گیری خطا در آن گره استفاده می‌شود، و آزمونی برای انجام عملیات تقسیم در گره مذکور انجام شده، سپس صفتی که موجب کاهش بیشتر انحراف معیار گردد به عنوان صفتی که روی آن شاخه زده شود، انتخاب می‌شود.

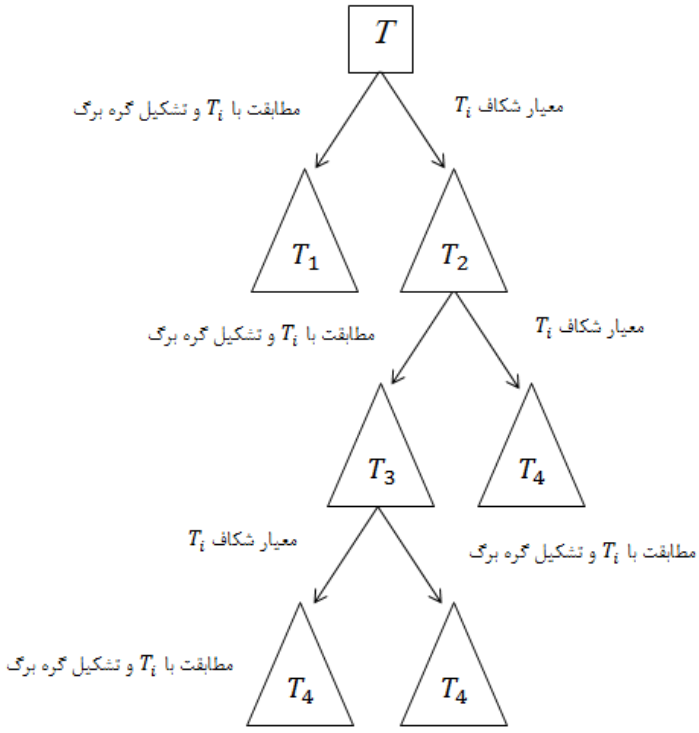
کاهش انحراف معیار استاندارد با استفاده از رابطه زیر محاسبه می‌شود.

$$SDR = sd(T) - \sum_i \frac{T_i}{T} \times sd(Ti) \qquad (۱)$$

T در رابطه (۱) شامل نمونه‌هایی است که به گره رسیده‌اند. Ti : مجموعه‌هایی است که از تقسیم کردن گره بر اساس صفت انتخابی به دست آمده‌اند. sd انحراف معیار داده‌است. پس از ایجاد درخت، یک مدل رگرسیون خطی چندگانه ساخته می‌شود. این مدل بر اساس داده‌های وابسته به آن گره و تمامی صفاتی که در زیردرخت با ریشه آن گره مورد استفاده قرار گرفته‌اند، ایجاد می‌شود. در مرحله بعد مدل‌های رگرسیون خطی، با کنار گذاشتن صفاتی که حذف آن‌ها باعث کاهش میانگین خطا می‌شود، ساده‌سازی می‌شوند. سپس هر زیردرخت برای انجام عملیات هرس بررسی می‌شود. اگر خطای تخمین زده شده برای مدل خطی در ریشه زیردرخت، کوچک‌تر یا مساوی میانگین خطای درخت باشد، زیردرخت هرس می‌شود. ناپیوستگی زیادی بین مدل‌های خطی مجاور در برگ‌های درخت هرس شده ایجاد می‌شود که در شکل (۲) نمایش داده شده است. مدل نهایی در یک برگ در فرایند هموارسازی در M5 از ترکیب کردن مدل به دست آمده که در آن برگ با مدل‌های موجود در مسیر ریشه تا برگ مربوط به دست می‌آید. آزمایش‌های انجام شده توسط ونگ و وایتن(۱۹۹۷)نشان داده است که هموارسازی به میزان زیادی، دقت پیش‌گویی‌ها را بهبود می‌بخشد.



شکل ۱- موقعیت حوضه‌ی آبخیز کرخه در ایران



شکل ۲- نمایش مدل درخت تصمیم

روش انجام کار

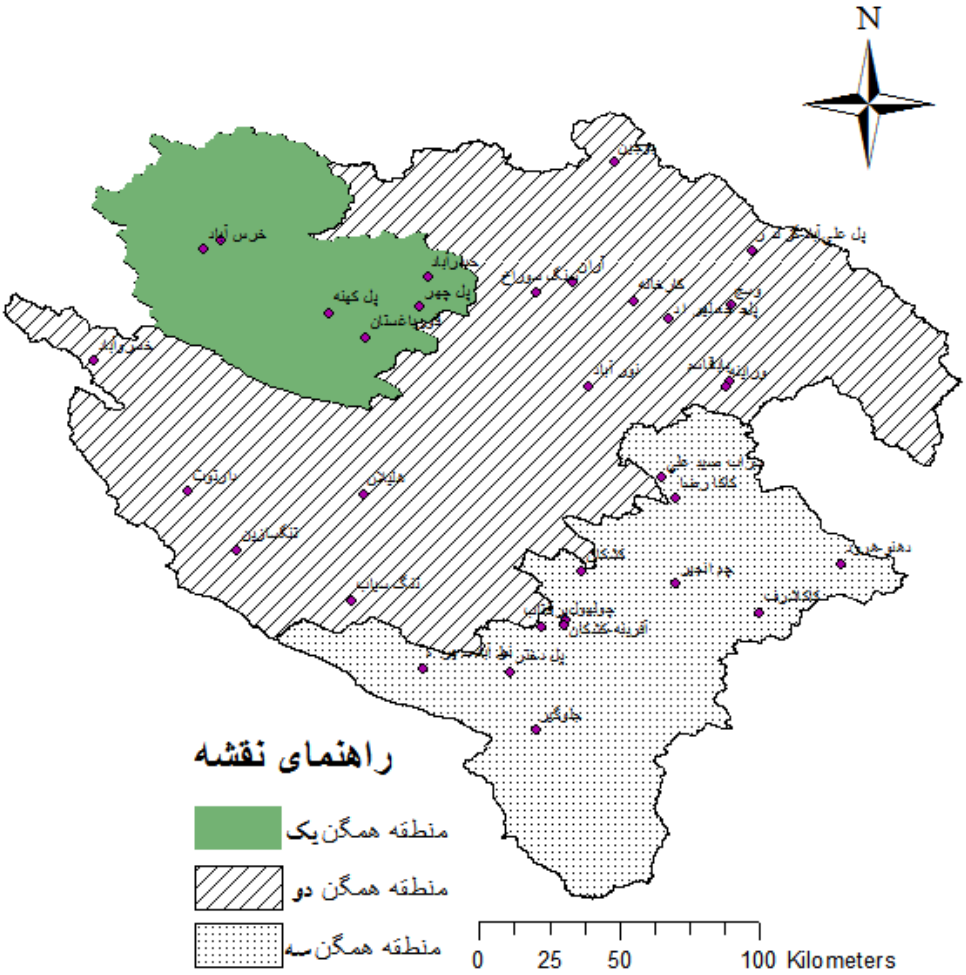
برای انجام تحقیق حاضر، ابتدا داده‌های حداکثر دبی لحظه‌ای سیلاب سالانه کلیه ایستگاه‌های موجود در منطقه از سازمان تحقیقات منابع آب ایران و سازمان آب و منطقه‌ی استان‌های موجود در حوضه‌ی کرخه اخذ و مورد بازبینی قرار گرفت. جمعاً ۸۹ ایستگاه هیدرومتری در حوضه آبخیز مطالعاتی موجود بود. تعدادی از ایستگاه‌ها دارای کفایت آماری لازم نبود و تعدادی از ایستگاه‌ها نیز در آزمون ران تست[1] برای همگن نبودن داده‌ها کنار گذاشته شده‌اند. لذا به عنوان مطلوب‌ترین حالت ممکن تعداد ۳۲ ایستگاه با دوره آماری از سال آبی ۶۱-۱۳۶۰ تا ۹۱-۱۳۹۰ وارد محاسبات گردیدند، از ایستگاه‌های همگن موجود، ۲۷ ایستگاه برای واسنجی (ایجاد مدل) و پنج ایستگاه برای صحت سنجی مدل های ایجاد شده، مورد استفاده قرار گرفت (ایستگاه‌های شماره ۲۸ الی ۳۲در جدول ۱)

1 -Run test

علوم و مهندسی آبیاری (مجله ی علمی- پژوهشی)، جلد ۴۰، شماره‌ی۴ زمستان ۹۶

جدول ۱- نوع توزیع آماری برای هر ایستگاه

شماره ایستگاه	نام ایستگاه	نوع توزیع	شماره ایستگاه	نام ایستگاه	نوع توزیع	شماره ایستگاه	نام ایستگاه	نوع توزیع
۱	سنگ‌سوراخ	لوگ‌نرمال‌دوپارامتری	۱۲	دارتوت	پیرسون نوع سه	۲۳	پلدخترکشکان	لوگ‌نرمال دوپارامتری
۲	کارخانه	لوگ‌نرمال سه‌پارامتری	۱۳	تنگ سازین	لوگ‌نرمال‌دوپارامتری	۲۴	جلوگیر	لوگ‌نرمال سه‌پارامتری
۳	بوجین	لوگ‌نرمال سه‌پارامتری	۱۴	تنگ سیاب	لوگ‌نرمال دوپارامتری	۲۵	پل علی آباد	لوگ‌نرمال دوپارامتری
۴	آران	لوگ‌نرمال دوپارامتری	۱۵	دهنو-هرود	لوگ‌نرمال سه‌پارامتری	۲۶	باباقاسم	لوگ‌نرمال‌سه‌پارامتری
۵	پل چهر	گامبل	۱۶	کاکا رضا	لوگ‌نرمال دوپارامتری	۲۷	پل‌حاج‌علیمراد	پیرسون نوع سه
۶	خرس آباد	پیرسون نوع سه	۱۷	سراب صید علی	پیرسون نوع سه	۲۸	کانال وراینه	لوگ‌نرمال سه‌پارامتری
۷	دوآب مرک	پیرسون نوع سه	۱۸	دو آب ویسیان	لوگ‌نرمال‌دوپارامتری	۲۹	کاکاشرف	لوگ‌نرمال سه‌پارامتری
۸	قورباغستان	لوگ‌نرمال‌دوپارامتری	۱۹	چم انجیر	پیرسون نوع سه	۳۰	حیدر آباد	پیرسون نوع سه
۹	نور آباد	پیرسون نوع سه	۲۰	آفرینه-کشکان	لوگ نرمال سه‌پارامتری	۳۱	وسج	پیرسون نوع سه
۱۰	هلیلان	لوگ نرمال دوپارامتری	۲۱	آفرینه-چولهول	پیرسون نوع سه	۳۲	پل کهنه	پیرسون نوع سه
۱۱	خسروآباد	لوگ پیرسون نوع سه	۲۲	برآفتاب	پیرسون نوع سه			



شکل ۳-نقشه مناطق سه‌گانه و موقعیت ایستگاه‌های مورد استفاده در آزمون همگنی

جدول ۲- ملاک‌های آماری برای سنجش کارایی دو مدل ایجاد شده

معیار آماری	ضریب همبستگی (%)	جذر میانگین مربعات خطا (m^3)	میانگین مطلق خطا (m^3)
درخت M5	۹۶	۰/۴۲۲	۰/۳۱۶
رگرسیون خطی	۸۶	۰/۵۱۲	۰/۳۹۱

تحلیل آمار سیلاب با دوره بازگشت‌های مختلف

در این پژوهش، آمارهای حداکثر لحظه‌ای سیلاب سالیانه مد نظر قرار گرفته و برای برآورد دبی سیلاب با دوره های بازگشت مختلف و انتخاب بهترین توزیع آماری از نرم افزار SMADA استفاده گردید. بهترین توزیع، برای هر ایستگاه با استفاده از آزمون نیکویی برازش انتخاب گردید سپس با استفاده از بهترین توزیع برای هر ۳۲ ایستگاه هیدرومتری ذکر شده، دبی سیلاب با دوره بازگشت های مختلف محاسبه گردید(جدول۱).

سپس، مناطق همگن[1] با استفاده از روش لانگبین مشخص شد (شکل۳) و در ادامه مشخصات فیزیوگرافی هر یک از زیر حوضه‌های ایستگاه‌های هیدرومتری توسط نرم‌افزارArc GIS به‌دست آمد.

1 - Homogeneity region

آماده سازی داده‌های ورودی

در این پژوهش از دو روش مدل رگرسیونی، و روش الگوریتم M5 مدل درختی برای به‌دست آوردن رابطه بین دبی با دوره بازگشت مختلف و مشخصات فیزیوگرافی استفاده شد. مدل‌های برآورد دبی با دوره برگشت‌های مختلف را می‌توان در دو حالت ارائه کرد، یکی با در نظر گرفتن مناطق همگن و دیگری بدون در نظر گرفتن مناطق همگن، در این پژوهش هنگام مدل‌بندی از روش رگرسیونی مناطق همگن مدنظر قرار گرفت و در هر یک از مناطق همگن سه‌گانه مدلی به‌دست آمد ولی در مدل بندی به صورت هوشمند درخت‌تصمیم‌گیری از آنجاکه هرچه تعداد داده‌ها برای آموزش شبکه بیشتر باشد جواب دقیق‌تر است برای این مدل هوشمند از کل منطقه و بدون در نظر گرفتن مناطق همگن استفاده می‌شود (شفیعی و همکاران،۱۳۸۶ ، سلیمانی و یوسفی،۱۳۸۰).

به‌علت اینکه وارد کردن داده‌ها به صورت خام و با واحد اندازه‌گیری متفاوت باعث کاهش سرعت و دقت شبکه هوشمند می‌شود از روش نرمال‌سازی داده‌ها استفاده شده است که این کار مانع کوچک شدن بیش ازحد وزن‌ها و سبب جلوگیری از اشباع زود هنگام نرون‌ها می‌گردد(راد ومانیا[1]،۲۰۰۴). فرمول های مختلفی برای نرمال‌سازی در تحقیقات مختلف پژوهشگران ارایه شده است (نورانی و کوماسی[2]،۲۰۱۳). در این پژوهش یکی از ساده‌ترین راه‌های نرمال‌سازی یعنی Ln گرفتن از داده‌ها مورد استفاده قرار گرفت. از آنجا که یکی از اهداف این پژوهش مقایسه شبکه هوشمند با روش رگرسیونی در پیش‌بینی دبی سیلاب می‌باشد پس لازم است شرایط مدل‌سازی در هر دو روش یکسان باشد از این‌رو در مدل سازی رگرسیونی نیز ابتدا از داده‌های خام مانند مدل هوشمند Ln گرفته و مدل‌سازی با داده‌های نرمال شده انجام گرفت.

در این پژوهش برای مدلسازی با استفاده از الگوریتم درخت تصمیم‌گیری M5 از نرم افزار وکا[3] که توسط پژوهشگران دانشگاه ویکاتو تهیه شده استفاده گردیده است.

شاخص‌های آماری

معمولا سه شاخص آماری ضریب همبستگی (r)[4]، جذر میانگین مربعات خطا (RMSE)[5] و میانگین مطلق خطا (MAE)[6] برای ارزیابی الگو استفاده می‌شود. روابط (۲) تا (۴) آن‌ها را نشان می‌دهد:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2(Y_i - \bar{Y})^2}} \quad (۲)$$

$$RMSE = \sqrt{\frac{1}{n}\sum[Y_i - X_i]^2} \quad (۳)$$

$$MAE = \frac{\sum_{i=1}^{n} |X_i - Y_i|}{n} \quad (۴)$$

که در آنها، Xi و Yi به ترتیب مربوط به مقادیر اندازه‌گیری شده و برآورد شده می‌باشد. $\bar{X}$ و $\bar{Y}$ به ترتیب به میانگین مقادیر مشاهداتی و میانگین مقادیر برآوردی هستند و n تعداد داده‌ها را نشان می‌دهد.

نتایج و بحث

همانطور که قبلا اشاره شد از مجموع ۳۲ ایستگاه موجود در حوضه، ۲۷ ایستگاه برای واسنجی (ایجاد مدل) و ۵ ایستگاه برای صحت سنجی مدل‌های ایجاد شده، مورد استفاده قرارگردید. از این رو برای مقایسه بین مدل درختیM5 و مدل رگرسیونی از ۵ ایستگاه مورد نظر استفاده گردید.

تحلیل آماری

مقایسه مقادیر واقعی با برآورد شده از طریق محاسبه معیارهای پراکندگی شامل میانگین خطای مطلق، جذر میانگین مربعات خطا و همبستگی بین آن‌ها صورت می‌گیرد (و بستر و الویر[7]، ۲۰۰۱)

میزان ضریب همبستگی بالای مدل الگوریتم M5 درخت تصمیم انجام شده و پایین بودن جذر میانگین مربعات خطا و میانگین خطای مطلق نسبت به رگرسیون خطی در جدول (۲) نشان دهنده کارایی مدل درختی ایجاد شده می‌باشد.

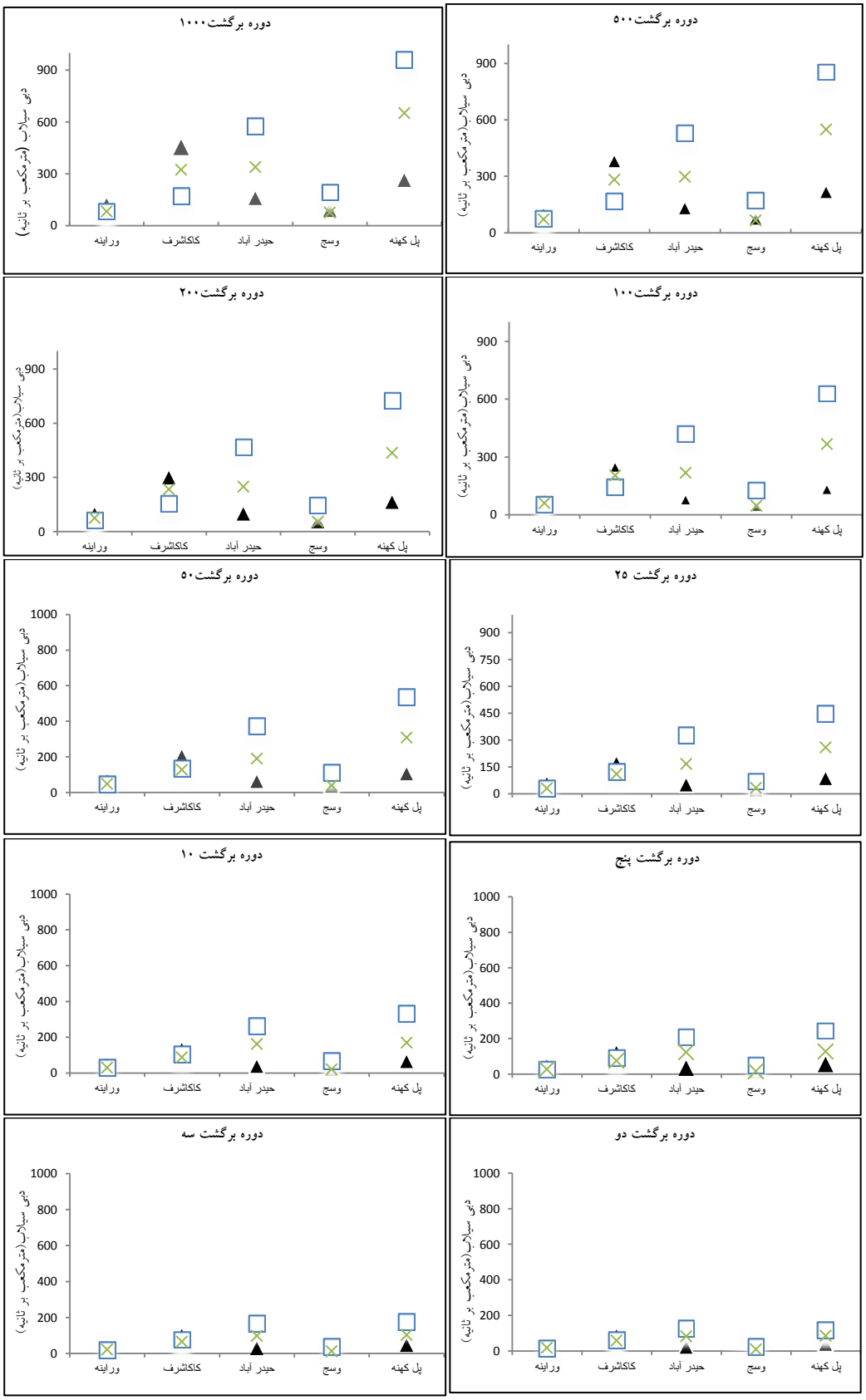
نتایج حاصل از مدل‌سازی انجام گرفته توسط مدل درخت با بهره‌گیری از داده‌های آزمون مطابق با قوانین زیر ارائه شده است.

همان‌گونه که مشخص است عملکرد درخت در مدل‌سازی بسیار خوب بوده به طوری که داده‌های پیش‌بینی شده و مشاهده شده به خوبی بر هم منطبق شده‌اند و اختلاف کمی با یکدیگر دارند.

با مشاهده شکل(۴) می توان به وضوح مطابقت خوب مدل درختی M5 را با دبی سیلاب واقعی رو متوجه شد، از این رو در حوضه آبریز کرخه برای محاسبه‌ی دبی سیلاب می‌بایست از مدل درختیM5 استفاده گردد.

با دقت در نمودار رسم شده برای مدل رگرسیونی می‌توان نتیجه گرفت که با افزایش دوره بازگشت مقدار دبی محاسبه شده از مقدار واقعی فاصله می‌گیرد. این نتیجه با نتایج چاوشی و اسلامیان(۱۹۹۹) و نساجیان زواره و همکاران (۲۰۱۱) همخوانی دارد

1- Riad and Mania
2- Nourani and Komasi
3 -Weka
4- Correlation Coefficient
5- Root Mean Square Error
6 -Mean Absolute Error
7 -Webster and Oliver



شکل ٤-مقایسه سیلاب واقعی و تخمین زده شده از روش رگرسیون و درخت تصمیم گیری M5

نتیجه‌گیری

با توجه به مطالعات تحلیل منطقه‌ای سیلاب در حوضه‌های مختلف، نیاز به استفاده از ابزار سیستم‌های هوشمند به منظور بالا بردن دقت برآورد تخمین دبی سیلاب احساس شد.

در این مطالعه پارامترهای موثر بر دبی سیلاب، بررسی شد و از سیستم هوشمند درخت تصمیم‌گیری M5 و مدل رگرسیون به منظور تخمین دبی‌سیلاب استفاده شد. پس از به‌دست آوردن دبی سیلاب محاسباتی با دو ابزار فوق، به مقایسه آنها با دبی سیلاب واقعی پرداخته‌شد.

در مدل درختی M5 و مدل رگرسیونی از مشخصات فیزیوگرافی حوضه (مساحت، محیط، طول آبراهه اصلی، ضریب گراویلیوس، شیب متوسط حوضه، ارتفاع متوسط حوضه) و دوره‌بازگشت‌های مختلف بعنوان ورودی‌های مدل استفاده شده و هدف نهایی به‌دست آوردن متغیر وابسته (دبی سیلاب) می‌باشد.

همان‌طور که نتایج به‌دست آمده در شکل(۴) نشان می‌دهد، مدل درختی M5 به عنوان یکی از ابزارهای مناسب و قدرتمند برای پیش‌بینی دبی سیلاب در حوضه کرخه معرفی شد که در صحت سنجی مدل(شکل ۴) بیشترین مطابقت را با دبی سیلاب واقعی نسبت به روش رگرسیون داشت.

قابل ذکر است مدل‌های رگرسیونی با توجه به نمودارهای شکل(۴)در زیر حوضه‌هایی که دبی‌سیلابشان در دوره بازگشت‌های مختلف کم باشد، از مطابقت خوبی با دبی سیلاب واقعی برخوردار است. ولی به طور کلی می‌توان نتیجه‌گیری نمود که از نظر دقت پیش‌بینی، مدل درخت تصمیم‌گیری M5 نسبت به مدل رگرسیونی در تمام دوره بازگشت‌ها و همچنین در تمام دبی سیلاب‌ها از مطابقت بالاتری با دبی سیلاب واقعی برخوردار بوده و می‌تواند به عنوان بهترین ابزار برای پیش‌بینی دبی سیلاب در دوره‌بازگشت‌های مختلف در حوضه آبریز کرخه معرفی گردد.

منابع

۱-البرزی، م. ۱۳۸۱. آشنایی با شبکه های عصبی مصنوعی، انتشارات دانشگاه صنعتی امیر کبیر .۱۳۷ صفحه.

۲-بی‌نام. ۱۳۸۴. اطلس ملی منابع آب ایران. انتشارات معاونت استفاده و مدیریت منابع آب ایران.

۳-ثروتی، م. ع. قنبری، ۱۳۸۶. برآورد سیلاب در حوضه رودخانه وربند لارستان. فصلنامه جغرافیایی سرزمین، علمی .۱۴(۴) ۷۴-۵۵.

۴-رسول‌زاده، ع، آذرتاج، الف. و پ. فرضی. ۱۳۹۴. ایجاد و بررسی مدل های مختلف تحلیل منطقه ای تناوب سیلاب تابعی از دوره بازگشت(مطالعه موردی : استان اردبیل). نشریه پژوهش های حفاظت آب و خاک، ۲۳(۴): ۲۶۸-۲۶۱.

۵-سلیمانی، ک و ا. یوسفی. ۱۳۸۰. بررسی نقش عوامل فیزیوگرافیک حوضه بر دبی های حداکثر در زیر حوضه‌های گرگان رود. مجله علوم کشاورزی و منابع طبیعی. (۴)۱۷۱-۱۶۴.

۶-شادمانی، م، معروفی، ص. و ک. محمدی. ۱۳۹۰. مدلسازی منطقه‌ای دبی سیلابی در استان همدان با استفاده از شبکه عصبی مصنوعی. مجله علوم کشاورزی و منابع طبیعی گرگان ، ۱۸(۴). ۲۱-۴۲

۷-شقیعی، م. شیرزاد، م. ن. نیک نیا،۱۳۸۶. تحلیل منطقه ای سیلاب توسط شبکه های عصبی مصنوعی(مطالعه موردی: حوضه ماسال استان گیلان). دومین کنفرانس مدیریت منابع آب، گیلان.

۸-علیزاده، الف. ۱۳۹۲. هیدرولوژی کاربردی. انتشارات دانشگاه فردوسی مشهد.

۹-هاشمی،۱۳۸۲. هیدرولوژی مهندسی. انتشارات شعرا، چاپ اول.

10.-Aziz.K, Rahman. A , Fang.G,Shrestha.S, (2014) Application of artificial neural networks in regional flood frequency analysis: a case study for Australia, 28: 541-554

11-Bhattacharya, B. and D.P. Solomatine. 2005. Neural networks and M5 model trees in modelling water level–discharge relationship Journal of Neurocomputing, 63(1): 407-412.

12-Boughton, W.C.1984. Flood freuency characteristics of some Arizona watersheds. Water Resources Bulletin 20(5): 761-769.

13- Chavoshi, S. and Eslamian, S.S., 1999. Regional flood frequency analysis in Zayandeh-Roud watershed using the Hybrid method. JWSS-Isfahan University of Technology, 3(3), pp.1-12.

14-Chiari, F et al. 2000; Predidtion of the Hydrologic Behavior of watershed using artificial neural network and Geographic information system. IEEE.1:382-386.

15-Dawson, C.W. Abrahart, R.J. Shamseldin, A.Y. and Wilby, R.L. 2006. Flood estimation at ungauged sites using artificial neural networks. Journal of Hydrology. 319:4. p 391-409.

16-Dayhoff, J.E. 1990. Neural Network Principles. Prentice-Hall International. U.S.A. 197 pp.

17-Dibike,Y.B. Solomatine,D. P. 2001. River flow forecasting using artificial neural networks. Physics and Chemistry of the Earth, [Journal]. - [s.l.] : Hydrology, Oceans and Atmosphere,. - 1 : Vol. 26. - pp. 1-7.

18- Dimitri, P. Solomatine and Yunpeng Xue. 2004." M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China", Journal of Hydrologic Engineering, Vol. 9, No. 6, P 491-591

19-Etemad-Shahidi, A. and L. Bonakdar. 2009. Design of rubble-mound breakwaters using M5′ machine learning method. Journal of Applied Ocean Research. 31(3): 197-201.

20-Etemad-Shahidi, A. and J. Mahjoobi. 2009. Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior. Journal of Ocean Engineering. 36(15): 1175-1181.

21-Fausett, L. 1994. Fundamentals of neural networks architectures algorithms and applications. Prentice-Hall Inc. New Jersey. 476 pp.

22-Heinz, D.F. and J.R. Stedinger . 1998. Using regional regression within index flood procedures and an empirical Bayesian estimator. Journal of Hydrology No.210. P 128-145.

23-Mehmed, K. 2003. Data Mining: Concepts, Models, Methods, and Algorithms. Journal of IEEE Computer Society, IEEE Press.

24-Mitchell, T.M. 1997. Machine Learning McGrow-Hill.

25-Nassajian Zavareh M.H., Vafakhah, M., and Telvari, A.R. 2011. Regional Flood Frequency Analysis in the Part of Large Central Watershed of Iran. Watershed Management Science and Engineering.

26-Nourani V. M. Komasi. 2013. A geomorphology-based ANFIS model for multi-station modeling of rainfall-runoff process. Journal of Hydrology, p. 41-55.

27-Kurtulus, B. and M. Razack. 2010. Modeling daily discharge responses of a large karstic aquifer using soft computing methods: artificial neural network and neurofuzzy. Journal of Hydrology, 381: 101-111.

28-Quinlan, J.R. 1992. Learning with continuous classe. Paper presented at the Proceedings of the 5th Australian joint Conference on Artificial Intelligence, Hobart, Tasmania. 343-348.

29- Riad, S., and Mania, J. 2004. Rainfall Runoff Model Using an Artificial Neural Network Approach, Mathematical and Computer Modeling, 40: 839-846

30- Ross, T.J. 1995. Fuzzy logic with engineering application. McGraw Hill Inc. USA. p:585.

31-Sadheer.K.P,Gosain.A.K,Ramassastri.K.S. 2002. A data algorithm for consrusting artificial neural network rainfall-runoff models. Journal of Hydrology. 128(16):1325-1330.

32- Tabari, H., Marofi, S., and Savziparvar, A. 2010. "Estimation of daily pan evaporation using artificial neural networks.

33-Wang, Y. and I.H. Witten. 1997. Induction of model trees for predicting continuous lasses. In Proceedings of the Poster Papers of the European Conference on Machine Learning, University of Economics. Faculty of Informatics and Statistics, Prague.

34-Webster, R. and M.A. Oliver. 2001. Geostatistics for Environmental Scientists. John Wiley and Sons, New York.

35-Zare Abyaneh, H. and M. Bayat Varkeshi. 2011. Evaluation of artificial intelligent and empirical models in estimation of annual runoff. Journal of Water and Soil. 25(2): 365-379.

36-Zhang, D. and J.P. Tsai. 2007. Advances in Machine Learning Applications in Software Engineering, Idea Group Inc.

Shahid Chamran University of Ahvaz
Irrigation Sciences and Engineering **Vol. 40, No. 4, Winter 2018, p. 183-195**

EXTENDED ABSTRACT

Regional Flood Analysis Via Comparison of The M5 Decision Tree Algorithm and Regression Models

H. Esmaeili Gisavandani[1] ,A. M. Akhond Ali[2*] ,H. Zarei[3] and M.Taghian[4]

1-MSc of Hydrology and Water resource Engineering Department of Shahid Chamran University, Ahvaz, Iran.
2*- Corresponding Author, Professor, Faculty Member of Hydrology and Water Resource engineering Department of Shahid Chamran University, Ahvaz, Iran. (*aliakh@scu.ac.ir*)
3- Assistant professor, faculty member of hydrology and water resource engineering Department of Shahid Chamran University, Ahvaz, Iran.
4- Assistant professor, faculty member of Agriculture and Natural Resources University, Ahvaz, Iran.

Received: Accepted:

Keywords*:* Regional flood analysis, Flood, Decision tree, M5 Algorithm, Regression model.

Introduction

Developing of techniques for regional flood frequency estimation in ungauged sites is one of the foremost goals of contemporary hydrology. The flood frequency evaluation for ungauged catchments is usually approached by deriving suitable statistical relationships (models) between flood statistics and basins characteristics. Already, several equations have been presented to estimate the flood frequency in different areas such as Karkheh basin. However, due to the complexity of this phenomenon, the relationships have not been capable to simulate the flood frequency with desired accuracy. Accordingly, in this study, in addition to the regression method has been used in the previous studies, the ANN and ANFIS models are applied. In fact, these are a type of black box models without any knowledge of processes within the system, in which inputs are converted into outputs (or output). This situation indicates that this type of new models is actually similar to the regression relations, however, there is further flexibility in adjusting the weights and thus can be used as a replacement to multivariate regressions.

Materials and Methods

In this research, four methods of linear regression model, M5 decision tree algorithm were used to obtain the relationship between high flow with various return periods and the physiographic characteristics. The model of the M5 tree is the development, the idea and concept of trees, classification and regression that is with a reverse tree structure which includes a root node in the highest part of the tree, that splits into other nodes and leaves, it is graphically displayed and in the form of rules if- then. This model is able to extract knowledge in the form of mathematical relations from the data set.

The ideas used to build this model are based on this analyzes a multivariate modeling problem by dividing it into several smaller sub problems, and combining its results. For this purpose, the problem space is subdivided into sub-domains and for each sub-domain of a multivariate linear regression model was fitted, in fact, with this method, a set of models, each of which will be used for only part of the problem Therefore, a linear model is formulated to approximate the relationship between input and output variables. This algorithm is the most

known than other options such as regression splines. This algorithm performs possible separations in a multivariate space and automatically generates models for these domains, the basis of the tree model is the method of dividing and overcoming traits for samples that reaches a node. Initially, the tree model, by partitioning the problem space, turns the tree into a regression. In this algorithm, to create a branch in a node, the standard deviation of the target variable values is used as a measurement criterion for the error in that node. And a test for doing division operations in the node, then the attribute that causes further deviation reduction is selected as the attribute to be applied to that branch.

Reduced standard deviation is calculated using the following equation.

$$SDR=sd(t)\sum\frac{|T_i|}{|T|}\times sd(T_i) \quad (1)$$

In the above equation T contains examples that have reached the target group, T_i is the number of data which are obtained by dividing into the desired node based on the selected attribute, sd the standard deviation is the sample that reaches the desired node.

Flow estimation models with different return periods can be presented in two modes, one by considering homogeneous regions and the other without regard to homogeneous regions, in this research, preparing the regression method model, case study are considered as Homogeneous regions. However, in the modeling of M5 decision tree algorithm the higher the number of data for training the network is, the more accurate the answer for artificial intelligence models for the whole region becomes , regardless of homogeneous regions.

Due to the fact that the data input in raw and different measuring units reduces the speed and accuracy of the smart grid, a method of data normalization has been used that prevents weighing too little and prevents early saturation of neurons. Different formulas for normalizing are presented in various researches of the researchers. In this research, one of the easiest ways to normalize (Ln) was used since one of the objectives of this study is to compare intelligent networks to predict high flood. Therefore, it is necessary that the modeling conditions be the same in all four methods. Therefore, in regression modeling take Ln from data such as intelligent ways and modeling was done with normalized data.

The study area, including 32 hydrometric stations, is located in the west of Iran. In this study, 27 of the stations for calibration and 5 of the stations for validation were used. To approach a unique model, return period was taken into account as the independent factor.

Results

the M5 Tree Algorithm model is also implemented. Application of decision tree models in water resources, due to its high accuracy, has been highly developed. In this research, physiographic characteristics of the basin were calculated by the Arc GIS, then, all of the physiographic parameters as well as return periods were considered as input data for M5 algorithm and linear regression.

Conclusion

The results indicated that The M5 algorithm, in comparison to the regression method, has a better performance to estimate flood based on correlation coefficient between estimated and observational data and also according to RMSE and MAE criteria the M_5 Model has got full match wit the actual high flow. Therefore, in the Karkhe catchment area, the M_5 model should be used to the high flood discharge. Carefully plotted on the graph for the regression model It can be concluded that with the increase in the return period, the calculated flow rate differs from the actual value. This result is in line with the results of Chavushi, Boroujeni, Islami, and Nasajian Zavare et al.