نشریه هواشناسی کشاورزی
جلد ۶، شماره ۱، بهار و تابستان ۱۳۹۷، صص.۴۹-۴۱

Journal of Agricultural Meteorology
Vol. 6, No. 1, Spring &Summer 2018, pp. 41-49

ارزیابی عملکرد رهیافت‌های برنامه‌ریزی ژنتیک و ماشین بردار پشتیبان در بازسازی داده‌های گمشده بارش

مصطفی کدخداحسینی[۱]، رسول میرعباسی نجف آبادی*[۲]، حامد نوذری[۳]، علی اصغر رستمی[۴]

تاریخ دریافت: ۱۳۹۶/۰۷/۲۹
تاریخ پذیرش: ۱۳۹۷/۰۴/۳۰

چکیده

فقدان یا گسست در سری زمانی داده‌های بارش در بسیاری از ایستگاه‌های هواشناسی، یکی از محدودیت‌های اصلی در مطالعات اقلیم‌شناسی و منابع آب است. در پژوهش حاضر از دو رهیافت هوشمند برنامه‌ریزی ژنتیک و ماشین بردار پشتیبان به منظور بازسازی داده‌های بارش ماهانه چهار ایستگاه باران‌سنجی واقع در استان همدان، در دوره آماری ۱۳۷۰ تا ۱۳۸۹ استفاده شد. خلاء آماری ابتدا به کمک اطلاعات یک ایستگاه، سپس دو ایستگاه و در نهایت داده‌های سه ایستگاه، بازسازی گردید. نتایج نشان داد که با افزایش حافظه و تعداد ایستگاه‌های دخیل در مرحله آموزش، عملکرد مدل‌ها بهبود می‌یابد. همچنین رهیافت ماشین بردار پشتیبان در بازسازی داده‌های بارش ماهانه ایستگاه سرابی و مریانج به ترتیب با ریشه میانگین مربعات خطا ۱۲/۹ و ۱۱/۴ میلی‌متر و ضریب همبستگی ۰/۹۳ و ۰/۹۵ نسبت به روش برنامه‌ریزی ژنتیک با ریشه میانگین مربعات خطای ۱۳ و ۱۲/۲ میلی‌متر و ضریب همبستگی ۰/۹۳ و ۰/۹۵ از عملکرد بهتری برخوردار بوده است.

واژه‌های کلیدی: بارش ماهانه، خلاء آماری، روش‌های هوشمند، همدان

مقدمه

تنوع مکانی و زمانی مناسب بارش در حوضه‌ها از جمله موارد مهم در مدل‌سازی‌های هیدرولوژیکی، مطالعات خشکسالی، تجزیه و تحلیل و طراحی سیستم‌های منابع آب است. در دسترس بودن سری کامل داده‌های بارش در مقیاس‌های زمانی و مکانی برای مدل‌های شبیه‌سازی هیدرولوژیکی که بارش به عنوان پارامتر ورودی آن‌ها در نظر گرفته می‌شود، ضروری است (Vieux, 2001). با این حال ایستگاه‌های هواشناسی اغلب به دلیل عدم نظارت بر قرائت داده‌ها یا مشکلات فنی دارای خلاءهای آماری می‌باشند (Tardivo and Berti, 2012). بازسازی داده‌های گمشده همواره مورد توجه هیدرولوژیست‌ها و کارشناسان هواشناسی و محیط زیست بوده است (Geerts, 2004; Kashani and Dinpashoh, 2012; Li et al., 2013). برای بازسازی داده‌های گمشده در ایستگاه‌های هواشناسی روش‌های متعددی مورد استفاده قرار گرفته است. از جمله این روش‌ها استفاده از مدل‌ها و توابع آماری (Coulibaly and Evora, 2007; Ramos Calzado et al., 2008)، میانگین بلندمدت سری داده‌های بارش به جای داده‌های گمشده (Linacre, 1992)، داده‌های بارش چند روز قبل و چند روز بعد از روزی که داده از دست رفته است، در رگرسیون‌های خطی به جهت پر کردن شکاف‌های یک روزه داده‌ها (Acock and Pachepsky, 2000)، داده‌های بارش ایستگاه‌های همسایه که بر پایه فاصله هندسی بین ایستگاه‌ها و بازسازی از داده‌های نزدیک‌ترین ایستگاه همسایه (Xia et al, 1999; Vicente-Serrano et al., 2010)، معکوس وزنی فاصله (Khorsandi et al, 2011; Teegavarapua and Chandramouli, 2005)، روش‌های درون‌یابی آماری مانند کریجینگ (Jeffrey et al., 2001) و درون‌یابی نیم‌فاصله مکانی و فضایی می‌باشد. به تدریج با پیشرفت و

[۱] دانشجوی دکتری مهندسی منابع آب، گروه مهندسی آب، دانشکده کشاورزی، دانشگاه شهرکرد
[۲] استادیار گروه مهندسی آب، دانشکده کشاورزی، دانشگاه شهرکرد
(*نویسنده مسئول: mirabbasi_r@yahoo.com)
DOI: 10.22125/agmj.2018.113742
[۳] استادیار گروه مهندسی آب، دانشکده کشاورزی، دانشگاه بوعلی سینا همدان
[۴] دانشجوی کارشناسی ارشد مهندسی منابع آب، گروه مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز

توسعه مدل‌های هوشمند و ارائه دقت و عملکرد بالای این مدل‌ها در پیش‌بینی پارامترها و کاربرد در علوم مختلف، استفاده از این مدل‌ها در دهه‌ای اخیر بسیار مورد توجه قرار گرفته است (Dastorani et al., 2010; Khorsandi et al, 2011; Linacre, 1992; Golabi et al., 2013). از جمله روش‌های هوش مصنوعی که برای بازسازی داده‌ها می‌توان استفاده کرد، روش‌های ماشین بردار پشتیبان (SVM)[1] و برنامه‌ریزی ژنتیک (GP)[2] در حالت GEP می‌باشد. ماشین بردار پشتیبان یکی از روش‌های یادگیری تحت نظارت[3] است که برای دسته‌بندی و رگرسیون قابل استفاده است. این روش توسط Vapnik (1998) بر پایه تئوری یادگیری آماری[4] بنا نهاده شده است. ماشین بردار پشتیبان روشی برای طبقه‌بندی دوتایی در فضای ویژگی‌های دلخواه است و از این رو روشی مناسب برای مسائل پیش‌بینی به شمار می‌رود و در اصل یک دسته‌بندی کننده دو کلاسه است که کلاس‌ها را توسط یک مرز خطی از هم جدا می‌کند. در این روش نزدیک‌ترین نمونه‌ها به مرز تصمیم‌گیری را بردارهای پشتیبان می‌نامند. این بردارها معادله مرز تصمیم‌گیری را مشخص می‌کنند. الگوریتم‌های شبیه‌سازی هوشمند کلاسیک مانند شبکه‌های عصبی مصنوعی، معمولاً قدر مطلق خطا یا مجموع مربعات خطای داده‌های آموزشی را حداقل می‌کنند، ولی مدل‌های SVM، اصل حداقل‌سازی خطای ساختاری را به کار می‌گیرند (Ahmadi et al,. 2015). اخیراً این مدل در گستره وسیعی از مسائل هیدرولوژیکی و به ویژه پیش‌بینی بارش (Lin et al., 2010; Lin et al., 2013; Maity et al., 2010; Hong and Pai, 2010; Wu et al., 2010) استفاده شده است. تئوری برنامه‌ریزی ژنتیک برای اولین بار توسط Koza (1992) ارائه شد. در دهه‌های اخیر این مدل به عنوان یک روش قوی برای حل طیف گسترده‌ای از مشکلات مدل‌سازی بارش-رواناب (Dorado et al., 2003; Nourani et al., 2011; Whigham and Crapper, 2001)، تعیین هیدروگراف واحد (Rabunal et al., 2007)، روندیابی سیلاب (Sivapragasam et al., 2008) و پیش‌بینی تراز دریاچه‌ها (Ghorbani et al., 2010) مورد استفاده قرار گرفته است. شکل توسعه یافته برنامه‌ریزی ژنتیک، برنامه‌ریزی بیان ژن است. برای تخمین داده‌های ناقص مربوط به ارتفاع امواج در خلیج مکزیک از مدل برنامه‌ریزی ژنتیک استفاده شد (Ustoorikar and Deo, 2008). نتایج نشان داد این مدل از دقت بسیار مطلوبی در پیش‌بینی داده‌های مربوط به سری‌های زمانی برخوردار است. به منظور بازیافت داده‌های گمشده امواج در طول ساحل غربی هندوستان از مدل برنامه‌ریزی ژنتیک استفاده شد (Kalra and Deo, 2007). پژوهش Solgy (2017) با استفاده از دو مدل هوشمند برنامه‌ریزی بیان ژن و ماشین بردار پشتیبان برای پیش‌بینی بارش ماهانه شهرستان نهاوند با استفاده از داده‌های بارش، دما و رطوبت نسبی ماهانه ایستگاه وراینه در یک دوره ۳۲ ساله (۱۳۹۳-۱۳۶۲) انجام شد. نتایج نشان داد که عملکرد هر دو مدل خوب و مشابه بوده (ضریب همبستگی حدود ۰/۹۲) ولی با توجه به بررسی معیارهای ارزیابی مختلف، مدل برنامه‌ریزی بیان ژن عملکرد کمی بهتری داشته است. به طورکلی می‌توان گفت که مدل برنامه‌ریزی بیان ژن برای مدل‌سازی و پیش‌بینی بارش ماهانه ایستگاه وراینه در شهرستان نهاوند مناسب‌تر بوده است. نتایج این پژوهش نشان از عملکرد موفق این مدل در بازسازی داده‌های گمشده در مقابل سایر روش‌های آماری دارد. مطالعه Isazadeh et al., (2016) برای پیش‌بینی جریان ماهانه حوضه خرخره‌چای با استفاده از تابع کرنل چندجمله‌ای درجه چهارم (به عنوان نماینده مدل SVM) با نتایج مدل ARMA (۲،۶) به عنوان نماینده مدل‌های سری زمانی) مقایسه گردید و نشان داد که مدل SVM از کارایی بهتری نسبت به مدل‌های سری زمانی در پیش‌بینی جریان ماهانه برخوردار است. Che-Ghani et al., (2014) از مدل برنامه‌ریزی بیان ژن برای بازسازی داده‌های بارش ماهانه ایستگاه‌های باران‌سنجی استفاده کردند. نتایج نشان داد که این مدل عملکرد بالایی در بازسازی داده‌های بارش با ضریب تبیین ۰/۸۸۶ دارد. هدف از انجام این پژوهش تکمیل نمودن نواقص و اطلاعات آماری داده‌های بارش ایستگاه‌هایی می‌باشد که اساس مطالعات و پژوهش‌های دیگر است. به این منظور عملکرد ماشین بردار پشتیبان (SVM) و برنامه‌ریزی ژنتیک (GP) در بازسازی داده‌های بارش ایستگاه‌های باران‌سنجی استان همدان مقایسه شد.

[1] Support Vector Machines
[2] Genetic Programming
[3] Supervised learning
[4] Statistical Learning Theory

مواد و روش‌ها

منطقه مورد مطالعه

استان همدان با مساحت ۲۰۱۷۲ کیلومتر مربع، ۲/۱ درصد از کل مساحت کشور را در بر گرفته است. این استان بین عرض‌های °۴۹ و '۳۵ تا °۵۹ و '۳۳ شمالی و طول‌های °۳۴ و '۴۷ تا °۳۴ و '۴۹ شرقی واقع شده و دارای اقلیمی سرد می‌باشد. در این مطالعه از داده‌های بارش ماهانه ایستگاه‌های آقاجانبلاغی، سرابی، آق‌کهریز و مریانج طی سال‌های ۱۳۷۰ تا ۱۳۸۹ استفاده شد که خصوصیات آماری آن‌ها در جدول ۱ ارائه شده است. همچنین مشخصات و موقعیت جغرافیایی این ایستگاه‌ها در جدول ۲ نشان داده شده است. دلیل انتخاب این ایستگاه‌ها نزدیک بودن مسافت هندسی و بالا بودن ضریب همبستگی داده‌های بارش ماهانه بین آن‌ها می‌باشد (جدول ۳).

Table 1- Statistical characteristics of available data at studied stations

جدول ۱- خصوصیات آماری داده های موجود در ایستگاه‌های مورد مطالعه

Station	Aghajan bolaghi	Agh kahriz	Maryanej	Sarabi
Max	150	187	174	190
Min	0	0	0	0
Mean	19.5	18.25	31.5	26.75
Variance	1209.79	1160.99	1655.71	1743.51
SD	34.78	34.07	40.69	41.76
CV	178.37	186.7	129.18	156.09

Table 2 – Geographical location of studied stations and annual average precipitation

جدول ۲- موقعیت جغرافیایی ایستگاه‌های مورد مطالعه و میانگین بارش سالانه

Station	Annual rainfall (mm)	Latitude (N)	Longitude (E)
Sarabi	436	34° 58'	48° 10'
Aghajan bolaghi	332	34° 50'	48° 03'
Agh kahriz	317	34° 59'	48° 48'
Maryanj	451	34° 49'	48° 48'

Table 3- Correlation coefficient between precipitation data of studied stations

جدول۳- ضریب همبستگی میان داده‌های بارش ایستگاه‌های مورد مطالعه

Station	Sarabi	Maryanaj	Agh kahriz	Aghajan bolaghi
Aghajan bolaghi	0.88	0.88	0.88	1
Agh kahriz	0.90	0.90	1	0.88
Maryanj	0.93	1	0.90	0.88
Sarabi	1	0.93	0.90	0.88

برنامه‌ریزی ژنتیک

برنامه‌ریزی ژنتیک جزو روش‌های الگوریتم گردشی محسوب می‌شود که مبتنی بر نظریه داروین است. الگوریتم‌های یاد شده اقدام به تعریف یک تابع هدف در قالب معیارهای کمی نموده و سپس تابع یاد شده را برای مقایسه جواب‌های مختلف حل مسئله در یک فرآیندگام به گام تصحیح ساختار داده‌ها به کار می‌گیرند و در نهایت جواب مناسب را ارائه می‌نمایند. فرآیند اجرایی گام به گام برنامه‌ریزی ژنتیک به صورت مراحل زیر است.

۱- تولید یک جمعیت اولیه از فرمول‌ها که این فرمول‌ها از ترکیب تصادفی مجموعه توابع و ترمینال‌ها ایجاد می‌شوند. ترمینال‌ها همان متغییرهای مستقل که شامل پارامترهای موثر در محاسبه داده‌های گمشده می‌باشد. انتخاب توابع مناسب در پژوهش‌های مختلف متفاوت بوده اما تقریباً در همه آن‌ها از چهار عملگر اصلی و توابع مثلثاتی به عنوان عملگرهای فرعی استفاده شده است که در این مطالعه از عملگرهای اصلی (جمع، تفریق، ضرب و تقسیم) و مثلثاتی جهت توابع استفاده شده است. سپس ساختار کروموزوم‌ها شامل طول سر و تعداد ژن‌ها از طریق آزمون و خطا انتخاب می‌گردد.

۲- هر یک از افراد جمعیت مذکور با استفاده از توابع برازش مورد ارزیابی قرار می‌گیرند.

۳- تولید یک جمعیت جدید از فرمول‌ها، که مراحل زیر برای تولید یک جمعیت جدید دنبال می‌شود:

الف) یکی از عمل های ژنتیکی تلاقی، جهش و تولید مثل انتخاب می‌شود (این سه عمل ژنتیکی، مهم‌ترین عمل‌های ژنتیکی مورد استفاده در برنامه‌ریزی ژنتیک می‌باشند. عمل‌های دیگری مثل اصلاح ساختار و غیره نیز با احتمال کمتر مورد استفاده قرار می‌گیرند)، ب) تعداد مناسبی از افراد جمعیت حاضر انتخاب می‌شوند (انتخاب فرد یا افرادی از جمعیت مذکور به صورت احتمالاتی می‌باشد که در این انتخاب احتمالاتی منفردهای با برازش بهتر به منفردهای نامرغوب ترجیح داده می‌شوند و آن معنی نیست که حتماً منفردهای نامرغوب حذف می‌شوند)، ج) از عمل ژنتیکی انتخاب شده برای تولید فرزند (فرمول جدید) استفاده می‌شود، فرزند (فرمول جدید) تولید شده در یک جمعیت جدید وارد می‌شود و ه) مدل مورد نظر با استفاده از تابع برازش مورد ارزیابی واقع می‌شود.

۴- گام سوم تا نیل به حداکثر تعداد تولید، تکرار خواهد شد.

۵- معیار پایان و ارائه نتایج اجرای برنامه (مثل، تعداد تولید جمعیت جدید، تعیین یک مقدار مشخص برای برازش فرمول‌ها که اگر میزان برازش برابر یا بیشتر از آن مقدار شد، اجرا متوقف شود). طرح کلی گام‌های اجرایی برنامه‌ریزی

ژنتیک در شکل ۱ نشان داده شده است. همچنین جدول ۴ مشخصات مدل برنامه‌ریزی ژنتیک استفاده شده در بازسازی داده‌های گمشده بارش ماهانه را نشان می‌دهد.

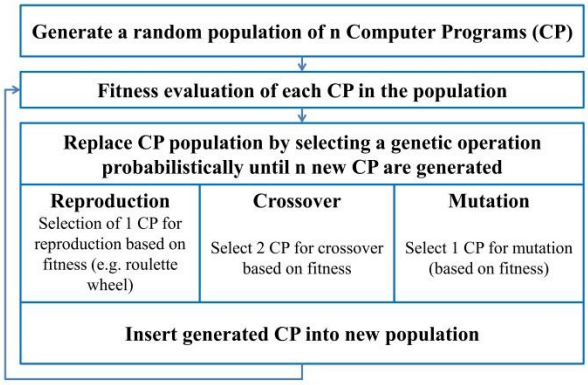

Figure 1- General description of the steps taken to implement genetic programming (Sette and Boullart, 2001)

شکل ۱- شکل کلی گام‌های اجرایی برنامه ریزی ژنتیک (Sette (and Boullart, 2001

Table 4- Specifications of genetic programming used to reconstruct missing monthly rainfall data

جدول ۴- مشخصات برنامه‌ریزی ژنتیک به کار رفته برای بازسازی داده‌های گمشده بارش ماهانه

parameter	Value
Head Size	8
Chromosomes	30
Number of Genes	3
Mutation Rate	0.044
Inverstion Rate	0.1
One-Point Recombination Rate	0.3
Two- Point Recombination Rate	0.3
Gene Recombination Rate	0.1
IS Transpositation Rate	0.1
RIS Transpositation Rate	0.1
Gene Transpositation Rate	0.1
Fitness Function Error Type	RMSE
Linking Function	+

ماشین بردار پشتیبان

در یک مدل رگرسیونی SVM لازم است وابستگی تابعی متغیر وابسته y به مجموعه‌ای از متغیرهای مستقل x تخمین زده شود. فرض بر این است که مانند دیگر مسائل رگرسیونی، معادله بین متغیرهای وابسته و مستقل توسط یک تابع معین f به علاوه یک مقدار اضافی نویز[1] مشخص می‌شود که شکل ریاضی آن در معادله ۱ نشان داده شده است.

$$Y=f(x) + Noise \qquad (۱)$$

بنابراین، موضوع اصلی پیدا کردن فرم تابع f است که بتواند به صورت صحیح، موارد جدیدی را که SVM تاکنون تجربه نکرده است پیش‌بینی کند. این تابع به وسیله آموزش مدل SVM بر روی یک مجموعه داده به عنوان مجموعه آموزش که شامل فرآیندی به منظور بهینه‌سازی دائمی تابع خطا است، قابل دسترسی است. بر مبنای تعریف این تابع خطا، دو نمونه از مدل‌های SVM شناخته شده است که عبارتند از: الف) مدل‌های رگرسیونی SVM نوع اول که مدل‌های SVM-n نیز نامیده می‌شوند و ب) مدل‌های رگرسیونی SVM نوع دوم که با نام SVM-e شناخته شده هستند. در این مطالعه، SVM-ε به دلیل کاربرد گسترده آن در مسائل رگرسیونی مورد استفاده قرار گرفت. ماشین‌های بردار پشتیبان برای حل مسائل غیرخطی، ابعاد مسأله را از طریق توابع کرنل تغییر می‌دهند. انتخاب کرنل برای SVM به حجم داده‌های آموزشی و ابعاد بردار ویژگی بستگی دارد؛ به عبارت دیگر، باید با توجه به این پارامترها تابع کرنلی را انتخاب کرد که توانایی آموزش برای ورودی‌های مسأله را داشته باشد. در عمل چهار نوع کرنل خطی، چندجمله‌ای، تانژانت هیپربولیک و گوسی (RBF) به کار گرفته می‌شوند. در جدول ۵ معادلات برخی از کرنل‌های رایج ارائه شده‌اند. درنهایت، تابع تصمیم رگرسیون بردار پشتیبان غیرخطی، به صورت معادله ۲ خواهد بود که کنترل‌کننده میزان نوسان تابع گوسی و همچنین کنترل‌کننده نتایج پیش‌بینی و تعمیم‌دهنده مدل SVM است (Yu et al., 2006).

$$f(x_i) = \sum_{i=1}^{1} (-\partial_i - \partial_i^*)K(x_i, x_j) + b \qquad (۲)$$

که ∂_i و ∂_i^* ضرایب لاگرانژ جهت بهینه‌سازی و حداکثرسازی تابع، $K(x_i, x_j)$ تابع کرنل موردنظر و b از ضرایب مدل SVM رگرسیونی می باشد.

Table 5- Common kernel functions in supporting vector machines (Hamel, 2009)

جدول ۵- توابع کرنل رایج در ماشین های بردار پشتیبان (Hamel, 2009)

Type of function	Kernel function
Linear	$K(x_i, x_j) = x_i^T . X_j$
Polynomial	$K(x_i, x_j) = (\gamma x_i^T . X_j + C)^d$
Hyperbolic tangent	$K(x_i, x_j) = \tanh(\gamma x_i^T . X_j + C)$
RBF	$K(x_i, x_j) = \exp(-\gamma \lvert x_i, x_j \rvert^2)$

معیارهای ارزیابی مدل‌ها

به منظور مقایسه بارش ماهانه مشاهده‌ای و تخمین زده شده توسط مدل‌های ماشین بردار پشتیبان (SVM) و

[1] Noise

برنامه‌ریزی ژنتیک (GP)، از معیارهای ضریب همبستگی (r) و ریشه میانگین مربعات خطا (RMSE[1]) استفاده شد

$$r = \sqrt{1 - \frac{\sum_{i=1}^{n}(R_{(obs)i} - R_{(pre)i})^2}{\sum_{i=1}^{n}(R_{(obs)i} - R_m)}} \quad (۳)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_{(pre)i} - R_{(obs)i})^2} \quad (۴)$$

که $R_{(pre)i}$ بارش تخمینی در زمان i، $R_{(obs)i}$ بارش مشاهده شده در همان زمان و R_m میانگین مقادیر مشاهداتی می‌باشند. بر این اساس هر چه مقدار RMSE در تکرارهای مختلف شبیه‌سازی مدل‌ها کمتر و مقدار همبستگی r میان داده‌های مشاهداتی و شبیه‌سازی شده نزدیکتر به یک باشد دقت و عملکرد بالاتری را نشان خواهد داد.

بحث و نتایج

در مطالعه حاضر، برای مدل‌سازی داده‌های بارش ماهانه ایستگاه باران‌سنجی سرابی و مریانج با استفاده از مدل‌های ماشین بردار پشتیبان و برنامه‌ریزی ژنتیک، داده‌های ۱۴ سال برای آموزش و ۵ سال (از سال ۷۹-۷۴) به عنوان داده‌های تست انتخاب شده‌اند. به طور کلی، برای آموزش مدل‌ها از ۷۵ درصد داده‌ها و برای صحت سنجی از ۲۵ درصد داده‌ها استفاده شد (Ahmadi et al., 2015). برای آموزش مدل‌ها شش الگو در نظر گرفته شده است (جدول ۶).

Table 6- Pattern used in training of models
جدول ۶- الگوهای مورد استفاده در آموزش مدل‌ها

field	Pattern
1	$R_{Sarabi}=f(R_{Aghajan\ bolaghi})$
2	$R_{Sarabi}=f(R_{Aghajan\ bolaghi}, R_{Aghakhriz})$
3	$R_{Sarabi}=f(R_{Maryanj}, RAghajan_{\ bolaghi}, R_{Aghakhriz})$
4	$R_{Maryanj}=f(R_{Aghajan\ bolaghi})$
5	$R_{Maryanj}=f(R_{Aghajan\ bolaghi}, R_{Aghkahriz})$
6	$R_{Maryanj}=f(R_{Sarabi}, R_{Aghajan\ bolaghi}, R_{Aghkahriz})$

همانطور که این الگوها نشان می‌دهند در سه الگو ایستگاه هدف سرابی و در سه الگوی دیگر ایستگاه هدف مریانج می‌باشد. به منظور آموزش مدل و بازسازی داده‌های گمشده در هر کدام از ایستگاه‌های هدف سه مرحله آموزش در نظر گرفته شده به طوری که آموزش با استفاده از یک، دو و سه ایستگاه مبنا انجام گردید و تابع ورودی به مدل‌ها در هر مرحله با اضافه شدن یک ایستگاه مورد بررسی قرار گرفت. پس از مدل‌سازی، داده‌های بازسازی‌شده بارش ماهانه ایستگاه‌های سرابی و مریانج با داده‌های مشاهده‌ای مورد مقایسه قرار گرفتند. به این منظور از شاخص‌های آماری میانگین مربعات خطا (RMSE) و ضریب همبستگی (r) استفاده شد. جدول ۷ نتایج دقت مدل‌سازی داده‌های بارش ایستگاه سرابی و مریانج را برای الگوهای مورد بررسی نشان می‌دهد. با توجه به این جدول زمانی‌که از داده‌های سه ایستگاه برای آموزش مدل‌ها مورد استفاده قرار می‌گیرد، برنامه‌ریزی ژنتیک و ماشین بردار پشتیبان نتایج بهتری از خود نشان می‌دهند. علاوه بر این، SVM عملکرد بهتری نسبت به GP دارد، که این نتیجه با تغییر در ایستگاه هدف (الگوی ۶) نیز به دست آمده است که مقدار مجذور میانگین مربعات خطایی با مقدار ۱۱/۴۳ میلی‌متر را داشته است. اگرچه هر دو مدل عملکرد نزدیکی را داشته اند، اما زمانی‌که از داده‌های دو ایستگاه برای آموزش مدل‌ها استفاده شد مدل GP عملکرد بهتری نسبت به مدل SVM نشان داد. بنابراین زمانی‌که تعداد ایستگاه‌های اطراف ایستگاهی که داده‌های گمشده دارد کمتر باشد، می‌توان از مدل GP برای بازسازی داده‌ها استفاده کرد. به طور کلی برای هر دو مدل ماشین بردار پشتیبان و برنامه‌ریزی ژنتیک هر چه تعداد ایستگاه‌های مجاور ایستگاهی که داده گمشده دارد بیشتر باشد، مقادیر بازسازی شده بارش در حدود ۸ میلی‌متر دقت بازسازی بالاتری را داشته است. به منظور ارزیابی و بررسی عملکرد مدل‌های SVM و GP مقادیر مشاهده شده در مقابل مقادیر بازسازی شده توسط مدل‌ها در شکل‌های ۲ تا ۵ ترسیم گردید. شکل ۲ و ۴ به ترتیب نتایج شبیه‌سازی داده‌های بارش ایستگاه سرابی و مریانج، با مدل‌های SVM و GP را نشان می‌دهند. همانطور که این شکل‌ها نشان می‌دهند با افزایش تعداد ایستگاه‌های ورودی برای آموزش مدل‌ها، داده‌های تخمینی به داده‌های مشاهده‌ای نزدیک‌تر و دقت بازسازی داده‌ها بیشتر شده است. شکل‌های ۳ و ۵ ضریب همبستگی بین داده‌های مشاهده‌ای و بازسازی شده برای الگوهای مختلف را نشان می‌دهد. همانطور که این شکل‌ها نشان می‌دهد با اضافه شدن داده‌های یک ایستگاه دیگر به آموزش مدل‌ها، خطای بازسازی به میزان قابل توجهی کاهش یافته است. شکل‌های ۳ و ۵ نشان می‌دهند، زمانی که از داده‌های یک و سه ایستگاه برای بازسازی داده‌های ایستگاه سرابی و مریانج استفاده شده است مدل SVM عملکرد بهتری داشته است. اما زمانی که از داده‌های دو ایستگاه برای آموزش مدل‌ها استفاده شد هر دو مدل تقریبا عملکرد یکسانی از خود نشان دادند.

[1] Root Mean Square Error

Table 7- Performance evaluation of different combinations of SVM and GP models in reconstructing monthly rainfall data of Sarabi and Maryanaj stations

جدول ۷- ارزیابی عملکرد ترکیب‌های مختلف مدل‌های SVM و GP در بازسازی داده‌های بارش ماهانه ایستگاه‌های سرابی و مریانج

Pattern	Base station	Target station	SVM		GP	
			r	RMSE (mm)	r	RMSE (mm)
1	Aghajan bolaghi	Sarabi	0.91	18	0.90	20
2	Agha kahriz and Aghajan bolaghi	Sarabi	0.93	14.34	0.92	14
3	Aghajan bolaghi, Agh kahriz and Maryanj	Sarabi	0.93	12.88	0.93	13
4	Aghajan bolaghi	Maryanaj	0.94	21.1	0.89	21.15
5	Aghajan bolaghi and Agh kahriz	Maryanaj	0.94	13.47	0.94	12.41
6	Aghajan bolaghi, Agh kahriz and Sarabi	Maryanaj	0.95	11.43	0.95	12.21

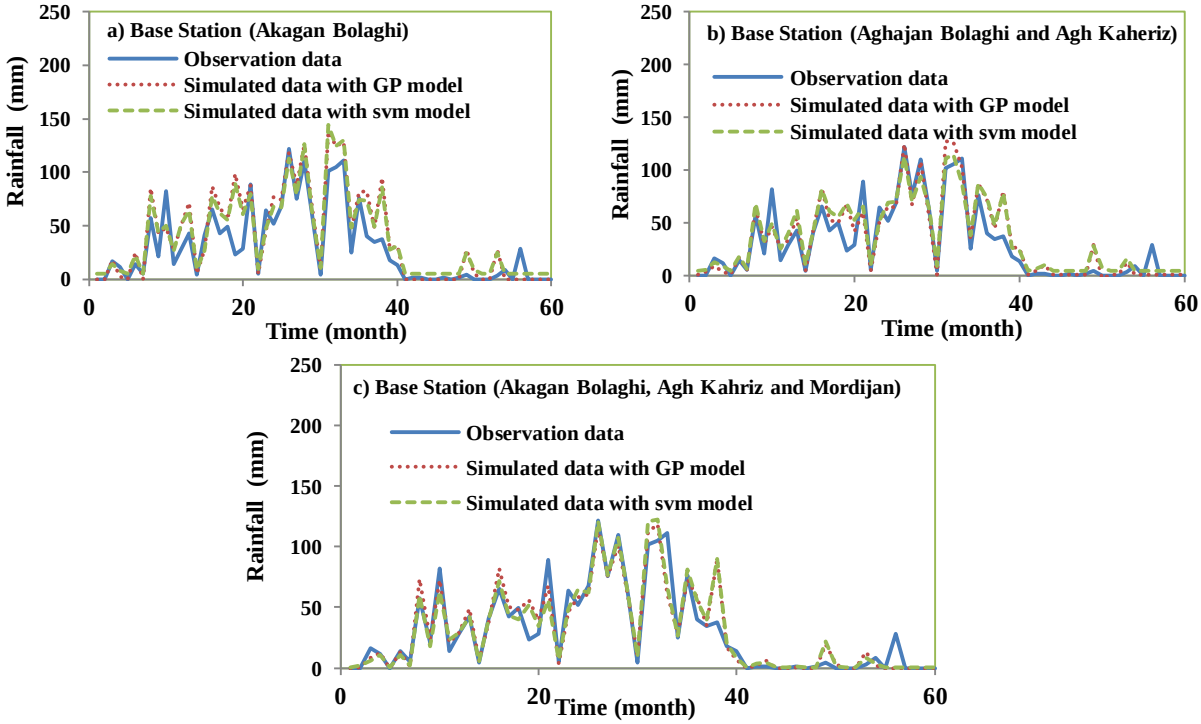

Figure 2- Comparison of observed and reconstructed data with SVM and GP models (Target station: Sarabi)

شکل ۲- مقایسه داده‌های مشاهده‌ای و بازسازی شده با مدل‌های SVM و GP (ایستگاه هدف: سرابی)

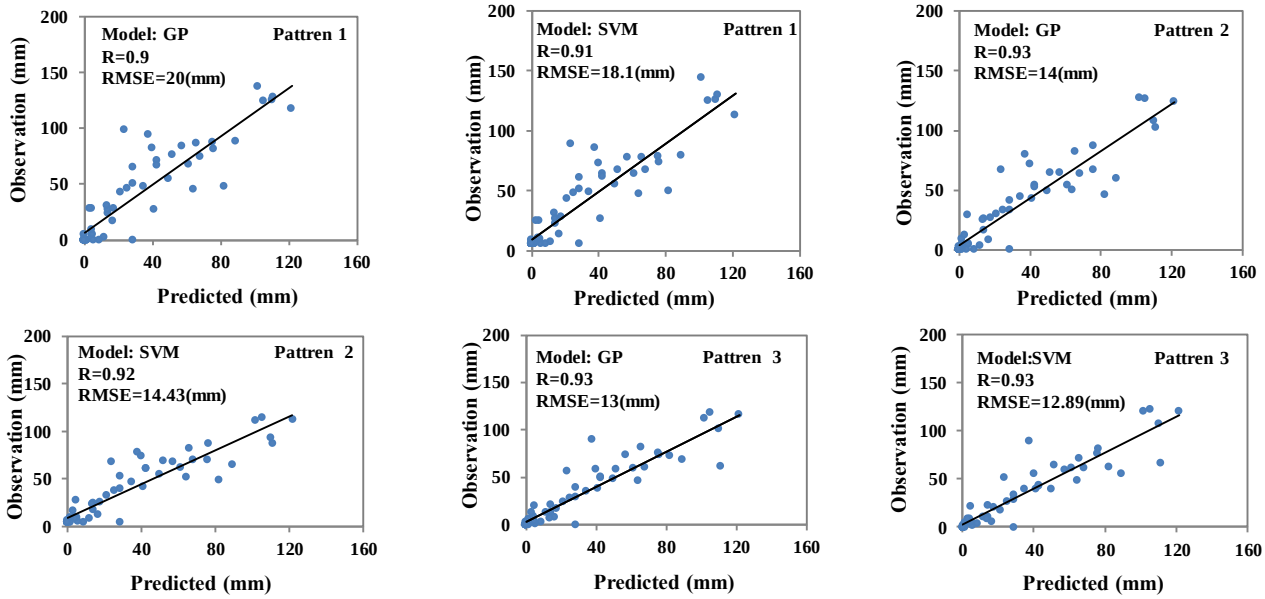

Figure 3- Comparison of correlation coefficient between observed and reconstructed rainfall data (Target station: Sarabi)

شکل ۳- مقایسه ضریب همبستگی بین داده‌های مشاهده‌ای و بازسازی شده (ایستگاه هدف: سرابی)

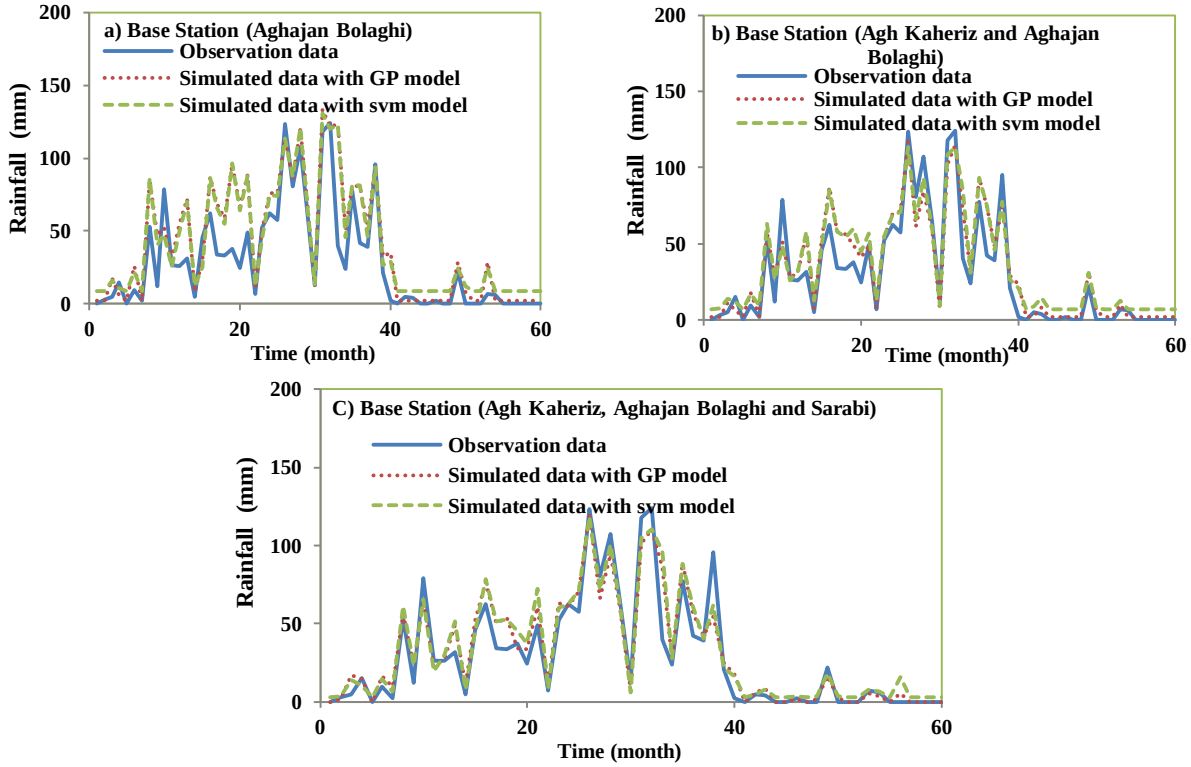

Figure 4- Comparison of observed and reconstructed data with Support Vector Machine and genetic programming (Target station: Maryanaj)

شکل ۴- مقایسه داده‌های مشاهده‌ای و بازسازی شده با ماشین بردار پشتیبان و برنامه‌ریزی ژنتیک (ایستگاه هدف: مریانج)

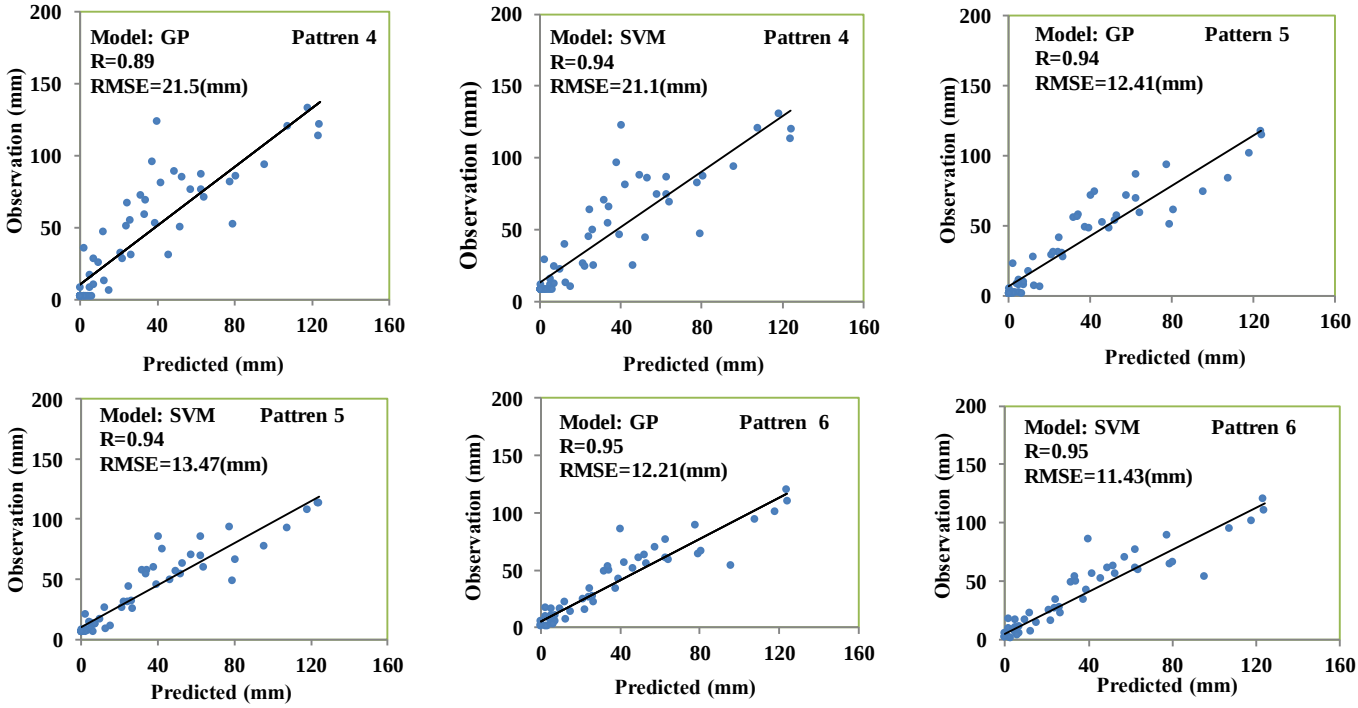

Figure 3- Comparison of correlation coefficient between observed and reconstructed rainfall data (Target station: Sarabi)

شکل ۳- مقایسه ضریب همبستگی بین داده‌های مشاهده‌ای و بازسازی شده (ایستگاه هدف: سرابی)

نتیجه‌گیری

در این پژوهش، عملکرد مدل‌های ماشین بردار پشتیبان و برنامه‌ریزی ژنتیک در بازسازی مقادیر بارش ماهانه ایستگاه باران‌سنجی سرابی و مریانج در استان همدان در طی سال‌های ۱۳۷۰ تا ۱۳۸۹ مورد بررسی قرار گرفت. در بخش آموزش شبکه دو حالت و برای هر حالت

سه سناریو در نظر گرفته شد. در سناریوی اول از آمار ایستگاه آقاجانبلاغی که یکی از ایستگاه‌های مجاور ایستگاه سرابی و مریانج است و دارای آمار و اطلاعات کاملی می باشد برای آموزش مدل ها استفاده شد. در سناریوی دوم از داده های ایستگاه آق‌کهریز و آقاجانبلاغی و در آخرین سناریو برای بازسازی داده‌های ایستگاه سرابی از ایستگاه‌های آق‌کهریز، آقاجانبلاغی و مریانج و برای بازسازی داده‌های ایستگاه مریانج از ایستگاه‌های آق‌کهریز، آقاجانبلاغی و سرابی برای آموزش مدل‌ها استفاده شد. پس از کنترل نتایج داده‌های ایستگاه‌های سرابی و مریانج بازسازی شدند. نتایج این مطالعه نشان می‌دهد هنگامی‌که از داده‌های یک و سه ایستگاه برای آموزش مدل‌ها استفاده گردید، مدل ماشین بردار پشتیبان عملکرد بهتری از خود نشان داده است و زمانی‌که از داده‌های دو ایستگاه برای آموزش مدل‌ها استفاده شد، مدل‌ها عملکرد نزدیکی باهم دارند بنابراین بطور کلی مدل ماشین بردار پشتیبان دقت بالاتری در بازسازی داده‌های بارش دارد. در بازسازی داده‌های بارش ماهانه، بهترین عملکرد در استفاده از داده‌های سه ایستگاه برای آموزش مدل‌های SVM و GP به دست آمده که ضریب همبستگی و میانگین مربعات خطا برای این الگو در بازسازی داده‌های ایستگاه سرابی به ترتیب ۰/۹۳ و ۱۳ میلیمتر برای مدل GP و ۰/۹۳ و ۱۲/۸۸ میلیمتر برای مدل SVM و در بازسازی داده‌های ایستگاه مریانج ۰/۹۵ و ۱۲/۲۱ میلیمتر برای مدل GP و ۰/۹۵ و ۱۱/۴۳ میلیمتر برای مدل SVM می‌باشد. گرچه هر دو مدل در سه حالت مختلف آموزش مدل‌ها، عملکرد تقریباً مشابهی در بازسازی داده‌های بارش ماهانه ایستگاه‌های سرابی و مریانج از خود نشان دادند، اما نتایج نشان داد که هرچه تعداد ایستگاه‌های ورودی برای آموزش مدل‌ها بیشتر شود، عملکرد مدل‌ها افزایش می‌یابد.

منابع

Ahmadi, F., Radmanesh, F., Abadi, R. M. N. 2015. Comparison between Genetic Programming and Support Vector Machine methods for daily river flow forecasting (Case Study: Barandoozchay River). Journal of Water and Soil, 28(6), 1162-1171. (In Farsi)

Acock, M. C., Pachepsky, Y. A. 2000. Estimating missing weather data for agricultural simulations using group method of data handling. Journal of Applied Meteorology, 39(7): 1176–1184.

Che-Ghani, N., Abu Hasan, Z., Liang, L. 2014. Estimation of missing rainfall data using GEP: Case Study of Raja River, Alor Setar, Kedah. Lecture Notes Artificul Inteligence, Article ID 716398: 1-5.

Coulibaly, P., Evora, N. D. 2007. Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology, 341(1–2): 27–41.

Dastorani, T. M., Moghadamnia, A., Piri, J., Ramirez, M. R. 2010. Application of ANN and ANFIS models for reconstructing missing flow data. Environmental Monitoring and Assessment, 166: 421-434.

Dorado, J., Rabunal, J. R., Pazos, Rivero, A., Santos, D., Puertas, J. 2003. Prediction and modeling of the rainfall–runoff transformation of a typical urban basin using ANN and GP. Applied Artificial Intelligence, 17: 329–343.

Geerts, B. 2003. Empirical estimation of the monthly-mean daily temperature range. Theoretical and Applied Climatology, 74(3–4): 145–165.

Golabi, M., Akhondi, A. Radmanesh, F. 2013. Comparison of performance of different artificial neural network algorithms in seasonal modeling Case study; Selected stations in Khuzestan province. Applied Geosciences Research, 30: 169-151. (In Farsi)

Ghorbani, M., Khatibi, R., Aytek, A., Makarynskyy, O. 2010. Sea water level forecasting using genetic programming and artificial neural networks. Computer and Geoscience, 36 (5): 620- 627.

Hamel, L. 2009. Knowledge Discovery with Support Vector Machines, Hoboken, N.J. John Wiley.

Hong, W. C., Pai, P. F. 2007. Potential assessment of the support vector regression technique in rainfall forecasting. Water Resources Managment, 21(2): 495–513.

Isazadeh, M., Ahmadzadeh, H., Ghorbani, M. A. 2016. Assessment of kernel functions performance in river flow estimation using support vector machine. Journal of Water and Soil Conservation, 23(3): 89- 69.

Jeffrey, S. J., Carter, J. O., Moodie, K. B., Beswick, A. R. 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environment Modell Softwere, 16(4): 309–330.

Kalra, R., Deo, M. C. 2007. Genetic programming to retrieve missing information in wave records along the west coast of India. Applied Ocean Research, 29(3): 99-111.

Kashani, M. H., Dinpashoh, Y. 2012. Evaluation of efficiency of different estimation methods for missing climatological data. Stochastic

Environmental Research and Risk Assessment A, 26: 59–71.

Khorsandi, Z., Mahdavi, M., Salajeghe, A., Eslamian, S. S. 2011. Neural network application for monthly precipitation data reconstruction. Journal of Environmental Hydrology, 19: 1-12.

Koza, J. R. 1992. Genetic programming: On the programming of computers by means of natural selection. Cambridge, MA: MIT Press.

Li, X., Li, L., Wang, X., Jiang, F. 2013. Reconstruction of hydrometeorological time series and its uncertainties for the Kaidu River Basin using multiple data sources. Theoretical and Applied Climatology, 113: 45–62.

Lin, G. F., Chen, G. R., Huang, P. Y. 2010. Effective typhoon characteristics and their effects on hourly reservoir inflow forecasting. Advance Water Resource, 33(8): 887– 898.

Lin, G. F., Chou, Y. C., Wu, M. C. 2013. Typhoon flood forecasting using integrated two-stage support vector machine approach. Journal of Hydrology, 486: 334– 342.

Linacre, E. 1992. Climate Data and Resources – A Reference and Guide, Routledge. Lon and NY.

Maity, R., Bhagwat, P. P., Bhatnagar, A. 2010. Potential of support vector regression for prediction of monthly streamflow using endogenous property. Hydrological Processes, 24(7): 917–923.

Nourani, V., Kisi, O., Komasi, M. 2011. Two hybrid artificial intelligence approaches for modeling rainfall–runoff process. Jouranl of Hydrology, 402: 41– 59.

Rabunal, J. R., Puertas, J., Suarez, J., Rivero, D. 2007. Determination of the unit hydrograph of a typical urban basin using genetic programming and artificial neural networks. Hydrological Processes, 21: 476– 485.

Ramos-Calzado, P., Gomez-Camacho, J., Perez-Bernal, F., Pita-Lopez, M.F. 2008. A novel approach to precipitation series completion in climatological datasets: application to Andalusia. International of Journal Climatology, 28(11): 1525–1534.

Sette, S., Boullart, L. 2001. Genetic programming: principles and applications. Engineering Applications of Artificul Intelligence, 14: 727– 736.

Solgi, A., Zarei, H., Shehnidarabi, M., Alidadis, A. 2017. Monthly precipitation forecast using gene expression and backup vector machine programming models. Journal of Applied Geosciences Research, 50: 91-103:

Sivapragasam, C., Maheswaran, R., Veena, V. 2008. Genetic programming approach for flood routing in natural channels. Hydrolgy Processes, 22: 623–628.

Tardivo, G., Berti, A. 2012. A dynamic method for gap filling in daily temperature datasets. Journal of Applied Meteorology and Climate, 51: 1079–1086.

Teegavarapua, R. S. V., Chandramouli, V. 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing. Journal of Hydrology, 312: 191-206.

Ustoorikar, K., Deo, M. C. 2008. Filling up gaps in wave data with genetic programming. Marine Structure, 21:177-195.

Vapnik, V. N. 1998. Statistical Learning Theory. Wiley, NY, 740 p.

Vicente-Serrano, S. M., Beguería, S., López-Moreno, J. I., García-Vera, M. I., Stepanek, P. 2010. A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. International of Journal Climatology, 30(8): 1146-1163.

Vieux, B. E. 2001. Distributed Hydrologic Modeling using GIS. In: Distributed Hydrologic Modeling Using GIS. Water, Science and Technology Library, 38: 217-238.

Whigham, P. A, Crapper, P. F. 2001. Modelling rainfall runoff using genetic programming. Math Computing, 33: 707–721.

Wu, J., Liu, M., Jin, L. 2010. Least square support vector machine ensemble for daily rainfall forecasting based on linear and nonlinear regression. Advances in Neural Network Research and Applications. Lecture Notes of Electrnic Engineer, 67(1): 55–64.

Xia, Y. L., Fabian, P., Stohl, A., Winterhalter, M. 1999. Forest climatology: Estimation of missing values for Bavaria, Germany. Agriculture Forest Meteorology, 96 (1–3): 131–144.

Yu, P. S., Chen, S. T., Chang, I. F. 2006. Support vector regression for real-time flood stage forecasting. Journal of Hydrology, 328: 704-716.

Journal of Agricultural Meteorology
Vol. 6, No. 1, Spring &Summer 2018, pp. 41-49

نشریه هواشناسی کشاورزی
جلد ۶، شماره ۱، بهار و تابستان ۱۳۹۷، صص.۴۹-۴۱

Performance evaluation of the genetic programming and support vector machine models in reconstruction of missing precipitation data

M. Kadkhodahosseini[1], R. Mirabbasi-Najafabadi[2], H. Nozari[3], A. Rostami[4]

Received: 21/10/2017
Accepted: 21/07/2018

Abstract
Incomplete rainfall datasets with missing gaps is a major challenge in climatology and water resource studies. In the present study, two intelligent models, namely Genetic Programing (GP) and Support Vector Machines (SVM) were used to reconstruct the monthly rainfall data of four rain-gauges located in Hamedan province, Iran during the period of 1992 to 2011. The incomplete rainfall data was reconstructed first by using the data of one, two and three stations respectively. The results showed that increasing the memory and the number of stations involved in the training phase, will improve the performance of the models. In reconstruction of monthly precipitation data of Sarabi and Maryanj stations, the Support Vector Machine method showed better performance with RMSE of 12.9 mm and 11.4 mm, and correlation coefficients (r) of 0.93 and 0.95, respectively. The corresponding values of RMSE for GP approach were 13 mm and 12.21 mm, which indicated the superior performance of SVM.

Keywords: Rainfall, Missing data, Intelligent methods, Hamedan

[1] Ph. D. Student of Water Resources Engineering, Department of Water Engineering, College of Agriculture, Shahrekord University, Shahrekord, Iran
[2] Assistant Professor, Department of Water Engineering, College of Agriculture, Shahrekord University, Shahrekord, Iran
(*Corresponding Author Email Address: mirabbasi_r@yahoo.com)
DOI: 10.22125/agmj.2018.113742
[3] Assistant Professor, Department of Water Engineering, College of Agriculture, Buali Sina University, Hamedan, Iran
[4] M. Sc. Student of Water Resources Engineering, Department of Water Engineering, College of Agriculture, Tabriz University, Tabriz, Iran