نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۳، شماره ۲، خرداد – تیر ۱۳۹۸، ص. ۳۷۷-۳۶۱

Journal of Water and Soil
Vol. 33, No. 2, May-Jun. 2019, p. 361-377

ترمیم داده‌های مفقود هواشناسی با روش‌های تکاملی و یادگیری ماشین
مطالعه موردی: بارش و دمای ماهانه درازمدت مشهد

محبوبه فرزندی[۱]- سید حسین ثنایی نژاد[۲*]- بیژن قهرمان[۳]- مجید سرمد[۴]

تاریخ دریافت: ۱۳۹۷/۰۵/۰۷
تاریخ پذیرش: ۱۳۹۸/۰۱/۲۶

چکیده

بارش و دما از مهم‌ترین متغیرهای هوا و اقلیم‌شناسی هستند. طول دوره آماری اهمیت بسزایی در دقت تحلیل این دو متغیر دارد. حجم نمونه کمتر از ۱۰۰ سال نمی‌تواند نوسانات دراز مدت را به خوبی منعکس کند. طولانی‌ترین آمار مربوط به دما و بارش ماهانه مشهد نزدیک به ۱۲۵ سال (از حدود ۱۸۹۳ الی ۲۰۱۷) است. متاسفانه این آمار مفقودی دارد. ترمیم داده‌های مفقود و افزایش دقت برآورد آن‌ها هدف این پژوهش است. ایستگاه‌هایی از کشورهای مجاور به‌عنوان ایستگاه‌های مبنا انتخاب شدند. ابتدا داده‌های مفقود با برازش ده الگوی رگرسیونی چندگانه برای بارش ماهانه (با ضرایب تعیین ۰/۶۳ تا ۰/۸۱) و شش الگو برای دمای ماهانه (۰/۹۸۶تا ۰/۹۹۳) ترمیم شدند. سپس برای کاهش خطاها، پارامترهای الگوهای رگرسیونی با روش‌های GA و ACO بهینه شدند. افزون بر این دو روش ANN و SVR نیز به‌منظور الگوسازی این داده‌ها نیز به‌کار گرفته شدند. نتایج نشان داد GA و ACO دقت برآورد داده‌های مفقود بارش را نسبت به روش‌های رگرسیونی فوق به طور چشمگیری افزایش می‌دهد. کمترین RMSE بین تمام الگوهای رگرسیونی بارش ۹/۷۹ میلی‌متر است. این معیار با روش GA به ۲/۵۶۰ میلی‌متر و با ACO به ۲/۵۵۹ کاهش می‌یابد. کمترین RMSE بین الگوهای رگرسیونی دما ۰/۹۸۶ میلی‌متر است. این معیار با روش ANN به ۰/۷۲۶ میلی‌متر و با SVR نیز به ۰/۵۵۱ کاهش می‌یابد. مقایسه ترمیم دما و بارش نشان می‌دهد که روش‌های تکاملی برای بارش و روش‌های یادگیری ماشین برای دما عملکرد بهتری دارند.

واژه‌های کلیدی: الگوریتم ژنتیک، داده مفقود، رگرسیون بردار پشتیبان، شبکه عصبی مصنوعی، کلونی مورچگان

مقدمه

داده‌های گمشده[5] (مفقودی) در آمار اهمیت ویژه‌ای دارد. تحلیل‌های کلاسیک آماری با نمونه‌های کامل (بدون مفقودی) سروکار دارد. زیرا تحلیل نمونه‌های شامل مفقودی اریب هستند. یعنی در صحت آنها تردید وجود دارد (۱). تحلیل‌های معمول آماری در آب و هواشناسی به روش‌های کلاسیک صورت می‌گیرد. بنابراین وجود داده‌های مفقود می‌توانند خللی در صحت نتایج به وجود آورند. مقدار اریبی با رشد نسبت گمشدگی افزایش می‌یابد (۱۳). حجم نمونه (طول دوره آماری) به ویژه در مناطق خشک و نیمه خشک برای تحلیل فراوانی، تحلیل سری‌های زمانی، تحلیل خشکسالی‌ها و ... باید حداقل ۱۰۰ سال باشد، زیرا داده‌هایی با طول کمتر نوسانات دراز مدت را منعکس نمی‌کنند. دوره بازگشت نیز به طول دوره آماری وابسته است. برآورد بزرگترین دوره بازگشت با دقت قابل قبول معادل یک پنجم طول داده‌ها است (۱۰). بنابراین در اختیار داشتن آمار طولانی مدت و کامل اولین نیاز تحلیل‌های قابل اعتماد در آب و هواشناسی است. این خود نیاز به روش‌های دقیق تری برای برآورد داده‌های مفقود دارد. زیرا مفقودی در داده‌های آب و هواشناسی معمول است.

روش‌های مختلفی برای ترمیم داده‌های مفقود پیشنهاد شده است. که هر یک بر اصول ریاضی خاصی بنا شده‌اند. رگرسیون به عنوان یک روش کلاسیک آماری با روش برآورد کمترین مربعات کاربرد زیادی در آب و هواشناسی دارد (۱۷). اکبال و همکاران (۹) تغییرات بارش در دریای زرد چین را در دوره آماری ۲۰۱۴-۱۹۶۰ بررسی کرده‌اند. آن‌ها روش رگرسیون خطی را برای برآورد داده‌های مفقود بارش به کار برده‌اند. طولانی‌ترین آمار دما و بارش ماهانه ایران در شهر مشهد و از سال ۱۸۹۳ (بیش از ۱۲۵ سال) قابل دسترس

۱، ۲ و ۳- به‌ترتیب دانشجوی دکتری، دانشیار و استاد گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه فردوسی مشهد
(*- نویسنده مسئول: Email: sanaei@um.ac.ir)
۴- دانشیار گروه آمار، دانشکده ریاضی، دانشگاه فردوسی مشهد
DOI: 10.22067/jsw.v33i2.74125
5- Missing data

است (۶ و ۲۰). این آمار مفقودی دارد. ترمیم ماهانه این داده‌ها توسط فرزندی و همکاران (۶) با الگوهای رگرسیونی انجام شده است. همچنین دو پژوهش در مقیاس سالانه بر روی بارش داده های فوق صورت گرفته است. ابتدا مفقودی ها ترمیم و سپس داده های کامل سالانه تحلیل شده‌اند. خلیلی و بذر افشان (۱۱) تداوم خشکسالی‌ها را با تحلیل فراوانی بارش سالانه طولانی مدت مشهد بررسی کردند. آن‌ها روش خودهمبستگی را برای ترمیم بارش سالانه انتخاب کردند. قهرمان و احمدی (۷) پانزده سال مفقودی بارش سالانه مشهد را با روش کریجینگ و برازش رگرسیون چندگانه بر میانگین‌های متحرک باران سالانه ترمیم کردند. آن‌ها فقط اطلاعات درون خود داده‌ها را به‌کار بردند.

امروزه روش‌های دیگری باعنوان کلی هوش مصنوعی شامل روش‌های یادگیری ماشین (شبکه عصبی، رگرسیون بردار پشتیبان و ...) و روش‌های تکاملی (الگوریتم ژنتیک، کلونی مورچگان و ...) برای بهینه‌سازی و کاهش خطای برآورد پیشنهاد شده است (۱۳). پریس و استفیلد (۱۵) از روش الگوریتم ژنتیک برای کالیبراسیون پارامترهای مدل درخت (MT) استفاده کردند. سپس این مدل ترکیبی را برای پیش‌بینی اجزای کیفیت جریان آب استفاده کردند. دیپاک و بیچکار یک روش ترکیبی شامل الگوریتم ژنتیک و درخت تصمیم برای بالا بردن دقت برآورد داده‌های مفقود ارائه داده اند (۴). دستورانی و همکاران داده های مفقودی ماهانه ده ایستگاه مختلف آبسنجی ایران را به چهار روش ANN، عصبی فازی، همبستگی و نسبت نرمال[1] برآورد و نتیجه گرفتند که هر چهار روش جواب قابل قبولی را می‌دهند. روش عصبی فازی نسبت به بقیه برتر و روش شبکه عصبی در رتبه دوم قرار دارد (۲). یوزگاتلیجیل و همکاران شش روش ترمیم داده‌های مفقود را برای بارش و دمای ماهانه ترکیه ارزیابی و مقایسه نمودند. روش‌ها در این تحقیق به دو دسته ساده و پیچیده تقسیم شدند. روش‌های ساده عبارتند از میانگین حسابی، نسبت نرمال (NR) و نسبت نرمال وزنی با روش همبستگی. روش‌های پیچیده شامل پرسپترون چندلایه شبکه عصبی مصنوعی (ANN)، استراتژی جاگذاری چندگانه با زنجیره مارکوف–مونت کارلو بر اساس حداکثر کردن امید ریاضی (EM-MCMC) و یک روش اصلاح شده EM-MCMC است. آنها مجموع مربعات خطا را برای انتخاب روش برتر به کار بردند. افزون براین روش تحلیل سری‌های زمانی پویای غیرخطی را با فن همبستگی بی‌بعد نیز برای وابستگی فضایی مکانی داده های جاگذاری شده به کار گرفتند. تحلیل آنها نشان داد که دو روش ANN و EM-MCMC از بقیه بهتر عمل می‌کنند (۲۳). برخی محققین برای بالا بردن دقت برآوردهای خود از روش‌های ترکیبی استفاده نموده‌اند. لیائو و همکاران (۱۲) از یک الگوریتم ابتکاری ترکیبی به نام SAEM-HMM پیشنهاد دادند که شامل مدل مخفی مارکف (HMM) الگوریتم EM و الگوریتم گرم و سرد کردن (SA) است. این الگوریتم برای غلبه بر حساسیت مدل مارکف مخفی به مقادیر اولیه از EM و برای رهایی از گیر افتادن در بهینه‌های محلی از SA استفاده می‌کند. اسعد و همکاران (۵) نیز روشی ترکیبی ارائه کردند که با استفاده از الگوریتم EM پارامترهای الگوی پیشنهادی را با دقت بیشتری برآورد می‌کند. الگوریتم پیشنهادی این مقاله VEM-DyMix نام دارد که از ترکیب الگوریتم تغییرات بیشینه‌سازی– امید ریاضی (VEM) با یک مدل مخلوط گوسی با متغیرهای پنهان با توزیع گام زدن تصادفی (DyMix) به دست می‌آید.

هدف این پژوهش افزایش دقت برآورد داده‌های مفقود در آمار بارش و دمای طولانی مدت ماهانه شهر مشهد است. الگوریتم‌های تکاملی کلونی مورچگان (ACO) و ژنتیک (GA) و روش‌های یادگیری ماشین شبکه عصبی مصنوعی (ANN) و رگرسیون بردار پشتیبان (SVR) در این راستا به کار گرفته شده و دقیق ترین ترمیم‌ها انتخاب شده‌اند. این آمار می‌تواند مبنای ارزشمندی برای مطالعات منابع آب، خشکسالی‌ها، تغییر اقلیم، گرمایش جهانی و ... باشد.

مواد و روش‌ها

این پژوهش مبادرت به ترمیم داده‌های مفقود ۱۲۵ ساله دما و بارش ماهانه مشهد با چند روش کرده است. امروزه روش‌های کارآمدتری برای رفع مشکل داده مفقود ارائه شده است که به سازوکار داده‌های مفقود بستگی دارد.

رگرسیون

تابعی است که وابستگی یک متغیر (پاسخ) به یک یا چند متغیر (پیشگو) را ارائه می‌دهد. این تابع می‌تواند خطی، غیرخطی، ساده و چندگانه باشد. رگرسیون چندگانه ابزاری سودمند در ترمیم و گسترش داده‌های مفقود است (۳). بررسی باقی‌مانده‌ها (آسیب‌شناسی[2]) الگوی رگرسیونی یکی از نقاط قوت رگرسیون است. رگرسیون امیدریاضی شرطی Y به شرط متغیرهای X_1 تا X_k مطابق رابطه $\mathbf{Y=E(Y\mid X_1=x_1,\cdots,X_k=x_k)}$ است. رابطه (۱) رگرسیون چندگانه خطی را براساس k متغیر پیشگو نشان می‌دهد. β_0 تا β_k پارامترهای الگو هستند که باید با داده‌های در دسترس برآورد شوند. u مؤلفه خطاست که از توزیع نرمال با میانگین صفر و واریانس ثابت σ^2

1- Normal ratio

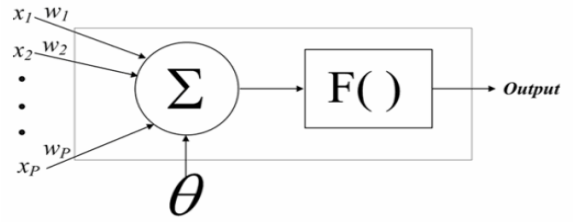
2- Diagnostic

پیروی می‌کند. σ^2 نیز باید توسط داده‌ها برآورد شود (۱۵ و ۱۶).

$$\mathbf{Y}=\boldsymbol{\beta}_{\circ}+\boldsymbol{\beta}_1\mathbf{X}_1+\boldsymbol{\beta}_2\mathbf{X}_2+\cdots+\boldsymbol{\beta}_{\mathbf{k}}\mathbf{X}_{\mathbf{k}}+\mathbf{u} \qquad (۱)$$

شبکه عصبی مصنوعی[1]

شبکه‌های عصبی مصنوعی (ANN) در واقع نوعی رگرسیون غیرخطی است. این روش برگرفته از سیستم‌های یادگیری دستگاه عصبی است که در آنها یک مجموعه پیچیده نرون‌های متصل در کار یادگیری دخیل هستند (شکل ۱). عصب‌ها (اجتماعی از نرون‌ها) اطلاعات و پیام‌های الکتروشیمیایی را منتقل می‌کنند. اگر پیام از حدی مشخص (آستانه یا سرحد) بیشتر شود، پیام منتقل می‌شود. مغز قادر است برای شناسایی الگوها و تفکیک الگوهای ناقص، آموزش ببیند. به‌علاوه، عملکرد مغز با از دست دادن برخی از نرون‌ها، از دست نخواهد رفت. نرون یک واحد تخمین‌گر ساده است که سیگنال‌ها را از طریق راه‌های ورودی رشته‌ای شکل به نام دندریت دریافت و ترکیب می‌کند. نرون‌ها دارای سوما (بدنه‌ی سلول، هسته و قسمت‌های حفاظتی)، دندریت (منطقه‌ی ورودی سلول، مجموعه‌ای از الیاف شاخه‌ای)، آکسون (ناحیه‌ی خروجی و خط انتقال نرون، رشته‌های شاخه‌ای و بلندتر از سایر اجزای سلول) و سیناپس (محل تلاقی آکسون سلول با دندریت سلول دیگر) است (۱۹).



شکل ۱- ساختار یک نرون مصنوعی

Figure 1- The structure of an artificial neuron

ANN با شناخت روابط ذاتی بین داده‌ها با فرآیند یادگیری[2] و با نرون‌ها تلاش می‌کند، نگاشتی میان فضای ورودی (لایه ورودی) و فضای مطلوب (لایه خروجی) ارائه دهد. لایه (یا لایه‌های مخفی)، اطلاعات دریافت شده از لایه ورودی را پردازش کرده و در اختیار لایه خروجی قرار می‌دهند. چندین الگوریتم آموزشی ANN وجود دارد. شبکه عصبی پس انتشار خطا[3] (RBP) و پرسپترون چندلایه (MLP) با الگوریتم‌های آموزشی مختلف که معروف‌ترین آن‌ها لونبرگ مارکوارت[4](LM) است در مطالعات هیدرولوژیکی رایج است (۱۸ و

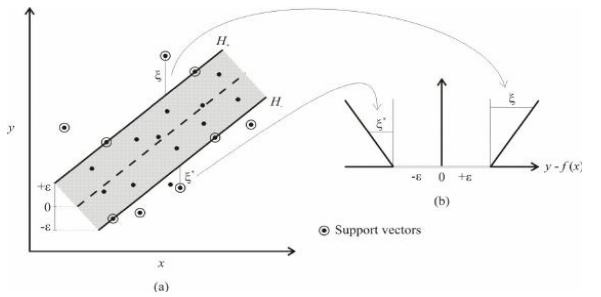
رگرسیون بردار پشتیبان[5]

رگرسیون بردار پشتیبان (SVR) مستقیماً از نظریه ماشین بردار پشتیبان (SVM) استخراج شده است (۲۱). این روش یک طبقه‌بندی کننده اطلاعات با حاشیه اطمینان نواری ایجاد می‌کند (شکل ۲). این کار کمینه کردن تابع ریسک عملی است (رابطه ۲). مجموعه‌ای از داده‌ها به شکل $(x_i,y_i)_{1<i<n}$ که $x_i\in R^n$ و $y_i\in R$ در اختیار است. SVR ساده‌ترین تابع تخمینگر را به صورت $f(x)=w^Tx+b$ در نظر می‌گیرد، به‌طوری که رابطه بین داده‌های برداری x و مقادیر خروجی y را به بهترین شکل ممکن (کمترین خطا) تخمین می‌زند.

$$R_{emp}=\frac{1}{m}\sum_{k=1}^{N}\left|y_k-w^Tx_k-b\right|_{\varepsilon} \qquad (۲)$$

رابطه (۲) تابع ریسک است. عبارت داخل سیگما در عبارت فوق تابع هزینه وپنیک نام دارد و تابعی به شکل رابطه (۳) و شکل ۳ است (۲۱).

$$\left|y-f(x)\right|_{\varepsilon}=\begin{cases}0 & \text{if } \left|y-f(x)\right|\le\varepsilon\\ \left|y-f(x)\right|-\varepsilon & \text{otherwise}\end{cases} \qquad (۳)$$



شکل ۲- نمودار تابع هزینه وپنیک

Figure 2- Vapnik's cost function graph

الگوریتم ژنتیک[6]

الگوریتم ژنتیک (GA) بر اساس نظریه تکاملی داروین بنا شده است و جواب مساله‌ای که از این طریق حل می‌شود رفته رفته بهبود می‌یابد. الگوریتم ژنتیک با یک مجموعه از جواب‌ها که از طریق کروموزوم‌ها نشان داده می‌شوند شروع می‌شود. این مجموعه

1- Artificial neural network
2- Learning
3- Back propagation
4- Levemberg-Marquart
5- Support vector Regression
6- Genetic Algorithm

جواب‌ها جمعیت[1] اولیه نام دارد. در این الگوریتم جواب‌های حاصل از یک جمعیت برای تولید جمعیت بعدی استفاده می‌شوند. در این فرآیند امید است که جمعیت جدید نسبت به جمعیت قبلی بهتر باشد. انتخاب بعضی از جواب‌ها از میان کل جواب‌ها (والدین)[2] به منظور ایجاد جواب‌های جدید یا همان فرزندان[3] بر اساس میزان برازندگی آن‌ها است. طبیعی است که جواب‌های مناسب‌تر شانس بیشتری برای تولید مجدد داشته باشند. این فرآیند تا برقراری شرط تعیین شده (مانند تعداد جمعیت‌ها یا میزان بهبود جواب) ادامه می‌یابد. دو عملگر ترکیب و جهش پایه‌ی GA است (۱۴).

بهینه‌سازی گروه مورچه‌ها[4]

الگوریتم کلونی مورچگان یا ACO از رفتار مورچگان طبیعی اقتباس شده است. مورچه‌ها مسیرهای رسیدن به غذا را به تصادف انتخاب می‌کنند. هر مورچه ماده شیمیایی (به نام فرومون) را در هنگام حرکت به عنوان اثر روی مسیر به‌جا می‌گذارد. تعداد ترددهای زیاد و ایجاد فرومون بیشتر منجر به ایجاد مسیر بهینه می‌شود. روش ACO از این خاصیت استفاده کرده و راه حل بهتری برای مساله با محاسباتی عددی بر مبنای علم احتمالات بهترین مسیر را در یک تابع پیدا می‌کند. هر مورچه که ماده شیمیایی را در مسیر خود به جا می‌گذارد با احتمال رابطه (۴) می تواند مورچه بعدی را در این مسیر قرار دهد.

$$P_A(t+1)=\frac{(c+n_A(t))^{\alpha}}{(c+n_A(t))^{\alpha}+(c+n_B)^{\alpha}}=1-P_B(t+1) \qquad (۴)$$

$\mathbf{n_A(t)}$ و $n_B(t)$ تعداد مورچه‌هایی که در زمان t در مسیر A و Bقرار دارند. **c** درجه جذب برای یک مسیر ناشناخته هر چه **c** بزرگتر باشد به معنی مقدار فرومون بیشتر برای عدم انتخاب مسیر تصادفی است. α اریبی (انحراف) به سمت فرومون به جا مانده در روند تصمیم‌گیری است (۱۴).

نتایج و بحث

بارش ماهانه ایستگاه‌های عشق آباد (R_{Ash})، سرخس (R_{Ser})، کوشکا (R_{Kus})، بایرام علی (R_{Bai})، کرکی (R_{Ker}) و رپتک (R_{Rep}) از کشور ترکمنستان به عنوان متغیر توضیحی (پیشگو) برای برآورد و ترمیم بارش‌های ماهانه مفقودی مشهد (R_{Mas}) انتخاب شدند. همچنین برآورد و ترمیم دمای ماهانه مشهد (T_{Mas}) با دماهای ماهانه عشق آباد (T_{Ash})، بایرامعلی (T_{Bai})، گودان (T_{Gud})، سرخس (T_{Ser}) و تجن(T_{Ted}) از ترکمنستان برای ترمیم دمای ماهانه ایستگاه مشهد انتخاب شدند. سه متغیر فاصله، همبستگی و وجود داده در ماه‌های مفقود در انتخاب این ایستگاه ها مؤثر بودند. مشخصات ایستگاه‌های برگزیده و سه عامل فوق در جدول ۱ آمده است. آزمون t در کلیه موارد (با p-value < 2.2e-16) همبستگی بارش ماهانه شش ایستگاه برگزیده با ایستگاه پاسخ را تأیید می‌کند.

الگوهای رگرسیونی

ترمیم بارش ماهانه مشهد به کمک ایستگاه های ذکر شده در جدول ۱ با ده الگوی رگرسیونی خطی و تمام لگاریتمی به شرح زیر انجام شده است. این کار با برنامه نویسی در محیط R-Studio صورت گرفت. الگوهای برازشی به ترتیب در الگوهای (P۱) تا (P۱۰) آمده است (متغیرها در بخش قبل تعریف شده‌اند).

$$R_{Mas}=\beta_0+\beta_1R_{Ash}+\beta_2R_{Kus}+\beta_3R_{Ser} \qquad (P۱)$$

$$\log(R_{Mas}+1)=\beta_0+\beta_1\log(R_{Ash}+1)+\beta_2\log(R_{Kus}+1)+\beta_3\log(R_{Ser}+1) \qquad (P۲)$$

$$R_{Mas}=\beta_0+\beta_1R_{Kus}+\beta_2R_{Rep}+\beta_3R_{Ser} \qquad (P۳)$$

$$\log(R_{Mas}+1)=\beta_0+\beta_1\log(R_{Kus}+1)+\beta_2\log(R_{Rep}+1)+\beta_3\log(R_{Ser}+1) \qquad (P۴)$$

$$R_{Mas}=\beta_0+\beta_1R_{Bai}+\beta_2R_{Kus}+\beta_3R_{Ser} \qquad (P۵)$$

$$\log(R_{Mas}+1)=\beta_0+\beta_1\log(R_{Bai}+1)+\beta_2\log(R_{Kus}+1)+\beta_3\log(R_{Ser}+1) \qquad (P۶)$$

$$R_{Mas}=\beta_0+\beta_1R_{Kus}+\beta_2R_{Ker}+\beta_3R_{Kus} \qquad (P۷)$$

$$\log(R_{Mas}+1)=\beta_0+\beta_1\log(R_{Bai}+1)+\beta_2\log(R_{Ker}+1)+\beta_3\log(R_{Kus}+1) \qquad (P۸)$$

$$R_{Mas}=\beta_0+\beta_1R_{Ash}+\beta_2R_{Bai}+\beta_3R_{Ser} \qquad (P۹)$$

$$\log(R_{Mas}+1)=\beta_0+\beta_1\log(R_{Ash}+1)+\beta_2\log(R_{Bai}+1)+\beta_3\log(R_{Ser}+1) \qquad (P۱۰)$$

ترمیم دمای ماهانه مشهد به کمک ایستگاه‌های مبنا (جدول ۱) با شش الگوی رگرسیونی غیرخطی و نیمه لگاریتمی به شرح زیر انجام شده است. الگوهای دما به ترتیب در الگوهای (T۱) تا (T۶) آمده است.

$$T_{Mas}=\beta_0+\beta_1T_{Ash}^2+\beta_2T_{Bai}^{0.2}+\beta_3\log(T_{Gud}+5)+\beta_4\exp(T_{Ted}) \qquad (T۱)$$

$$T_{Mas}=\beta_0+\beta_1T_{Ash}^2+\beta_2\log(T_{Bai}+6)+\beta_3\log(T_{Gud}+5) \qquad (T۲)$$

1- Population
2-Parents
3- Offspring
4- Ant Colony optimization

جدول ۱- ایستگاه‌های منتخب هواشناسی ترکمنستان برای ترمیم داده‌های مفقودی ماهانه مشهد

Table 1– The selected stations of Turkmenistan for imputation monthly missing data of Mashhad

ایستگاه‌های مستقل Independent stations		سرخس Serahs	عشق آباد Ashghabad	کوشکا Kushka	بایرام علی BairamAli	رپتک Repetek	کرکی Kerki	تجن Tedzen	گودان Gudan
فاصله تا مشهد (مایل) Distance to Mashhad (mil)		91.48	137.43	166.68	168.85	252.28	326.46	91.28	116.01
فاصله تا مشهد (کیلومتر) Distance to Mashhad (km)		147.2	221.1	268.2	271.7	405.9	525.3	146.9	186.7
همبستگی با بارش مشهد Correlation with Mashhad precipitation	r	0.81	0.70	0.70	0.68	0.68	0.64	-	-
	آماره t	41.41	32.08	30.68	30.95	24.80	27.55	-	-
همبستگی با دمای مشهد Correlation with Mashhad temperature	r	0.995	0.990	-	0.992	-	-	0.991	0.986
	آماره t	255.08	226.19	-	256.92	-	-	225.22	154.67

جدول ۲- ضرایب الگوهای P۱ تا P۵ بارش مشهد، عامل تورم واریانس (VIF)، آماره دوربین واتسون، خطا و قدرت الگو

Table 2- The Mashhad rainfall Coefficients for p1 to p5 patterns, VIF[1], Durbin-Watson Statistics, RMSE and Pattern power

	ضرایب Coefficients	برآورد ضریب Coefficient estimate	خطای معیار Standard error	آماره t	هم‌خطی (VIF)	ضریب تعیین (R^2_{adj})	جذر مربع خطا (RMSE)	آماره F	دوربین واتسون (D-W)
الگوی (P۱)	β_0	2.05	0.53	3.87	-	0.78	9.79	968	1.76
	β_1	0.29	0.03	11.34	1.8				
	β_2	0.11	0.02	5.92	2.5				
	β_3	0.68	0.03	19.74	3.1				
الگوی (P۲)	β_0	0.27	0.04	6.27	-	0.81	13.56	1165	1.77
	β_1	0.34	0.03	12.96	2.4				
	β_2	0.17	0.03	5.57	5.1				
	β_3	0.46	0.04	12.59	5.7				
الگوی (P۳)	β_0	3.84	0.57	6.76	-	0.75	12.00	677	1.61
	β_1	0.11	0.02	4.63	2.8				
	β_2	0.22	0.05	4.73	2.3				
	β_3	0.73	0.04	18.46	3.2				
الگوی (P۴)	β_0	0.54	0.04	13.09	-	0.79	14.31	824	1.60
	β_1	0.19	0.04	4.50	6.8				
	β_2	0.19	0.04	4.40	4.7				
	β_3	0.52	0.05	11.16	7.2				
الگوی (P۵)	β_0	4.15	0.53	7.82	-	0.74	12.23	801	1.65
	β_1	0.096	0.038	2.50	2.8				
	β_2	0.14	0.02	6.67	2.5				
	β_3	0.78	0.042	18.57	3.9				

1- Variance of influence factor

$$T_{Mas} = \beta_0 + \beta_1 T_{Ash}^2 + \beta_2 T_{Bai}^{0.2} + \beta_3 T_{Gud}^{0.3} \quad (T۳)$$

$$T_{Mas} = \beta_0 + \beta_1 T_{Ash}^2 + \beta_2 T_{Bai}^{0.2} + \beta_3 \log(T_{Gud} + 5) \quad (T۴)$$

$$T_{Mas} = \beta_0 + \beta_1 T_{Ash} + \beta_3 \log(T_{Gud} + 6) \quad (T۵)$$

$$T_{Mas} = \beta_0 + \beta_1 T_{Ser} + \beta_3 \log(T_{Gud} + 6) \quad (T۶)$$

نتایج برای الگوهای (P۱) تا (P۱۰) در جداول (۲ و ۳) و الگوهای (T۱) تا (T۶) در جدول ۴ آمده است.

آزمون‌های آسیب‌شناسی الگوها (تحلیل باقیمانده‌ها) انجام شد. به دلیل حجم بالا از آوردن نمودارها صرف نظر می‌شود و فقط به نتایج حاصل اکتفا می‌کنیم.

استقلال باقیمانده‌ها: آماره آزمون دوربین واتسون در محدوده ناهمبسته بودن مقادیر جدول دوربین واتسون (۲/۲۶- ۱/۶۰) قرار دارد. بنابراین استقلال باقیمانده‌ها تأیید می‌شود (جداول ۲ تا ۴).

پایایی واریانس: مقدار آماره کای-دو در آزمون پایایی واریانس (ncvtest) نشان می‌دهد واریانس باقیمانده ها برای الگوهای دمایی و بارشی پایا است (با مقادیر احتمال نزدیک به صفر). همچنین نمودار پراکنش تثبیت واریانس نیز آن را تأیید می‌کند (مقادیر پیش‌بینی شده در برابر باقیمانده‌های استاندارد شده به صورت ایده‌آل و مستطیلی توزیع شده‌اند).

داده پرت: مقدار کم آماره میانگین فاصله کوک در تمام الگوها (۰/۰۰۱۲ تا ۰/۰۰۲۴ برای الگوهای بارشی و ۰/۰۰۱۶ تا ۰/۰۰۳۸ برای الگوهای دمایی) نبود داده پرت را نشان می‌دهد.

نرمال بودن باقیمانده‌ها: بررسی نرمال بودن باقیمانده‌ها با آزمون شاپیرو و نمودار چندکی انجام شد. نتایج نشان داد که باقیمانده‌ها در الگوهای دمایی به صورت نرمال توزیع شده‌اند. اما در الگوهای بارشی کمی انحراف از نرمال بودن باقیمانده‌ها مشاهده شد. با توجه به اینکه اندازه نمونه به میزان کافی بزرگ و سایر فروض کلاسیک برقرار، انحراف از فرض نرمال بودن معمولا کم‌اهمیت و پیامدهای آن ناچیز است (۲۲) لذا می‌توان از آن چشم‌پوشی کرد.

جدول ۳- ضرایب الگوهای P۶ تا P۱۰ بارش مشهد، عامل تورم واریانس (VIF)، آماره دوربین واتسون، جذر مربع خطا و قدرت الگو

Table 3- Coefficients of Mashhad rainfall for p6 to p10 patterns, VIF, Durbin-Watson Statistics, RMSE and Pattern power

	ضرایب Coefficients	برآورد ضریب Coefficient estimate	خطای معیار Standard error	آماره t	همخطی (VIF)	ضریب تعیین (R^2_{adj})	جذر مربع خطا (RMSE)	آماره F	دوربین واتسون (D-W)
الگوی (P۶)	β_0	0.54	0.037	14.68	-	0.79	14.80	1075	1.67
	β_1	0.30	0.041	7.39	6.3				
	β_2	0.19	0.032	5.82	5.4				
	β_3	0.42	0.045	9.34	8.2				
الگوی (P۷)	β_0	4.91	0.59	8.33	-	0.63	13.48	497	1.60
	β_1	0.41	0.039	10.55	2.4				
	β_2	0.22	0.041	5.39	3.1				
	β_3	0.26	0.022	11.62	2.3				
الگوی (P۸)	β_0	0.57	0.039	14.56	-	0.75	15.84	893	1.70
	β_1	0.47	0.037	12.46	4.7				
	β_2	0.11	0.038	2.89	5.5				
	β_3	0.29	0.031	9.23	4.8				
الگوی (P۹)	β_0	2.83	0.55	5.15	-	0.74	11.48	806	0.71
	β_1	0.28	0.028	9.89	2.01				
	β_2	0.04	0.039	0.98	2.9				
	β_3	0.78	0.036	0.33	2.9				
الگوی (P۱۰)	β_0	0.29	0.042	6.84	-	0.81	12.32	1226	1.78
	β_1	0.21	0.041	5.19	6.9				
	β_2	0.30	0.027	11.22	1.8				
	β_3	0.47	0.037	12.74	6.2				

جدول ٤- ضرایب الگوهای T١ تا T٦ دمای مشهد ، عامل تورم واریانس (VIF)، آماره دوربین واتسون، جذر میانگین مربع خطا و قدرت الگو

Table 4- Mashhad temperature Coefficients of T1 to T6, VIF, Durbin-Watson Statistics, RMSE and Pattern power

	ضرایب Coefficients	برآورد ضریب Coefficient estimate	خطای معیار Standard error	آماره t	همخطی (VIF)	ضریب تعیین (R^2_{adj})	جذر مربع خطا (RMSE)	آماره F	دوربین واتسون (D-W)
الگوی (T١)	β_0	10	0.046	-35.25	–	0.9877	0.90	1.3×10^4	1.66
	β_1	0.012	0.00025	47.77	5.7				
	β_2	10.38	0.47	22.12	10.0				
	β_3	3.12	0.197	16.25	9.6				
	β_4	8.6×10^{-15}	3.9×10^{-15}	-2.12	1.4				
الگوی (T٢)	β_0	-1.89	0.35	-5.40	–	0.9864	1.00	1.7×10^4	1.68
	β_1	0.74	0.01	71.23	7.0				
	β_2	-0.46	0.18	-2.52	8.3				
	β_3	1.89	0.14	13.00	8.7				
الگوی (T٣)	β_0	-14.74	0.57	-25.78	–	0.9868	0.90	1.6×10^4	1.61
	β_1	0.02	0.0003	47.39	5.3				
	β_2	10.25	0.53	19.19	10				
	β_3	3.7	0.25	14.69	10				
الگوی (T٤)	β_0	-16.40	0.45	-36.1	-	0.9874	0.89	1.7×10^4	1.63
	β_1	0.12	0.0003	52.51	9.5				
	β_2	10.63	0.47	22.78	10				
	β_3	3.03	0.19	15.90	4.6				
الگوی (T٥)	β_0	-4.55	0.021	-15.86	-	0.9931	0.93	3.7×10^4	1.64
	β_1	0.79	0.0096	18.55	8.7				
	β_2	1.81	0.16	11.17	8.7				
الگوی (T٦)	β_0	-4.87	0.31	-15.6	-	0.9881	0.71	2.8×10^4	1.66
	β_1	0.69	0.0097	70.50	7.5				
	β_2	2.74	0.16	16.15	7.5				

بهینه‌سازی الگوها با الگوریتم ژنتیک و کلونی موچگان

برآورد پارامترهای رگرسیونی به روش کمترین مربعات خطا انجام می‌شود. چون این الگوها برازش خوبی بر داده‌ها دارند صورت کلی الگوها را پذیرفته و ضرایب الگوها با دیدگاه دیگری (GA و ACO) برآورد می‌شوند. با توجه به اینکه نتایج الگوهای بارشی مشابه هستند پنج الگوی اصلی (الگوهای خطی) انتخاب شد. پارامترهای این پنج الگو و ۶ الگوی دما (جداول ۲ تا ۴) به کمک GA) و الگوریتم ACO بهینه‌سازی می‌شوند. شبیه‌سازی این الگوریتم‌ها در محیط نرم‌افزار متلب انجام شده است. پیش فرض‌های GA عبارتند از: محدوده تغییرات ضرایب ۲۰- تا ۲۰ (براساس تحلیل پایلوت). بیشترین تکرار الگوریتم از ۲۰۰ تا ۳۰۰۰ برای هر الگو. ورودی‌های اولیه مشترک برای الگوها مطابق جدول ۵ است. بهینه‌سازی با ACO نیز نیازمند مقادیر اولیه است. پارامترهای ثابت در همه الگوها در جدول ۶ آمده است. تعداد جمعیت اولیه را ۱۰ در نظر می‌گیریم. مقدار وزن ها و احتمالات به دست آمده مطابق جدول ۷ است.

جدول ۵- مفروضات و ورودی‌های اولیه برای اجرای الگوریتم ژنتیک

Table 5- Assumptions and initial inputs for implementing the genetic algorithm

روش Method	پارامتر μ Mutation parameter	پارامتر گاما Gamma	تعداد والدین Number of parents	احتمال (درصد انتخاب داده) μ mutation rate	درصد تقاطع Crossover rate	تعداد جمعیت اولیه Population size	بیشترین تکرار Maximum iteration
چرخ رولت Roulette wheel	0.02	0.05	2	0.3	0.8	30	1000 to 3000

جدول ٦- پارامترهای ثابت در همه الگوهای الگوریتم کلونی مورچگان

Table 6- Fixed Parameters in All Patterns of the Ant Colony Algorithm

نرخ فاصله-خطا (زیتا) Deviation-distance ratio	فاکتور شدت q Intensification factor	اندازه نمونه n Sample size	بیشترین تکرار Maximum iteration	تعداد جمعیت اولیه Initial population count
1	0.5	50	100	10

جدول ۷- مقدار وزن‌ها و احتمالات به دست آمده در همه الگوهای الگوریتم کلونی مورچگان

Table 7- Value of the weights and probabilities obtained in all patterns of the Ant Colony Algorithm

وزن‌ها (w) Weights	0.0798	0.0782	0.0737	0.0666	0.0579	0.0484	0.0388	0.0299	0.0222	0.0158
احتمالات (p) Probabilities	0.1560	0.1529	0.1440	0.1303	0.1133	0.0946	0.0759	0.0586	0.0434	0.0309

نتایج شامل ضرایب بهینه شده و معیار خطا (RMSE) برای الگوهای دمایی مطابق جدول ۸ و برای الگوهای بارشی مطابق جدول ۹ است. همان طور که در جدول ۹ آمده است خطای برآورد (RMSE) در همه الگوها کمتر از ۳ است. مقدار RMSE نشان می‌دهد که این روش‌ها در برآورد بارش ماهانه بسیار موفق عمل کرده و خطای الگوها به طور چشمگیری کاهش یافته است. الگوی (P۱) بارش پس از بهینه‌سازی به کمک GA و ACO با RMSE = 2.56 کمترین خطا را دارد (جدول ۹). لذا داده‌های مفقود ماهانه بارش طولانی مدت مشهد با یکی از این الگوها تکمیل و جاگذاری می‌شود. نمودار سری زمانی بارش طولانی مدت سالانه مشهد پس از تکمیل و ترمیم با الگوی GA در شکل ۸ آمده است. مقادیر جاگذاری شده با رنگ قرمز نشان داده شده است. تغییر محسوسی در کاهش میزان خطای الگوهای دما پس از بهینه‌سازی با این دو روش (GA و ACO) مشاهده نمی‌شود (جدول ۸).

برآورد داده‌های مفقود دما و بارش ماهانه مشهد به روش شبکه عصبی مصنوعی

مدل MLP از روش ANN به منظور مدل‌سازی داده‌های بارش و دمای ماهانه مشهد انتخاب و اجرا شده است. بسته نرم‌افزاری nftool در نرم‌افزار متلب ۲۰۱۷ برای اینکار استفاده شده است. ۷۰٪ داده‌های کامل و معلوم برای آموزش شبکه ۱۵٪ برای اعتبارسنجی و ۱۵٪ برای آزمایش به طور تصادفی انتخاب شد.

بارش ماهانه عشق آباد، کوشکا و سرخس به منظور مدل‌سازی بارش ماهانه مشهد و دمای ماهانه ایستگاه‌های عشق آباد، گودان، سرخس و تجن نیز برای مدل‌سازی دمای ماهانه مشهد با روش ANN انتخاب شد. تعداد نرون‌های پنهان در لایه ورودی برای دما ۵ و برای بارش ۳ انتخاب شد (با تکرار و خطا) بهترین تابع فعالیت تابع سیگموئید (tanh) و الگوریتم آموزش LM است. تقسیم بندی داده‌ها به صورت تصادفی است. میزان کارایی با MSE سنجیده می‌شود. به الگویتم اجازه می‌دهیم تا ۵۰۰ بار تکرار شود. کارایی (میزان خطا) و گرادیان اعدادی نزدیک صفر و تعداد شکست۲۰ یعنی اجازه می‌دهیم ۲۰ بار شکست بخورد. هر کدام از موارد فوق صورت گیرد باعث توقف فرآیند می‌شود. نتیجه در تکرار ۳۶ و خطای ۰/۴۵۸ حاصل شده و الگوریتم خاتمه یافت.

رسم کل داده‌های مشاهده شده و پیش‌بینی شده توسط شبکه داده‌ها در شکل‌های ۳ و ۴ به همراه میزان خطا و ضریب تعیین رگرسیونی آمده است. نحوه توزیع خطاها در شکل ۵ آمده است. توزیع میزان خطا در بخش‌های آموزش، آزمایش و اعتبارسنجی با رنگ‌های مجزا مشخص شده است. توزیع خطاها تقریبا نرمال و میانگین نزدیک به صفر است. خطای MSE برای مدل‌سازی دما در تکرار هشتم به کمترین مقدار خود یعنی ۰/۴۹۸ می‌رسد. در مورد بارش تکرار ۴۳ام با مقدار ۸۴/۷ برای MSE کمینه است (RMSE =9.2). این مقادیر کمینه خطا در بخش اعتبارسنجی می‌باشد و برای کل داده‌ها کمی افزایش می‌یابد (اشکال ۳ و ۴). یعنی خطای برآورد بارش در روش ANN نسبت به رگرسیون تفاوت چندانی ندارد اما در مورد دما عملکرد خوبی داشته است (RMSE=0.725).

جدول ۸- ضرایب اصلاح شده الگوهای دما با الگوریتم ژنتیک و الگوریتم کلونی مورچگان

Table 8- Improved coefficients of temperature patterns with genetic algorithm and Ant colony algorithm.

	ضرایب Coefficients	برآورد ضرایب با GA Coefficient estimate by GA	بیشترین تکرار Max iteration	NFE	$RMSE_{GA}$	برآورد ضرایبACO Coefficient estimate by ACO	$RMSE_{ACO}$
الگوی (T۱)	β_0	-2.263	500	16530	1.057	-10.137	0.976
	β_1	0.016				0.014	
	β_2	-4.735				-0.071	
	β_3	7.035				7.119	
	β_4	-0.308				-0.222	
الگوی (T۲)	β_0	-15.53	200	6630	1.054	-13.979	1.004
	β_1	0.016				0.018	
	β_2	3.006				2.007	
	β_3	6.287				6.287	
الگوی (T۳)	β_0	-13.773	500	16530	1.061	-37.382	0.874
	β_1	0.0123				0.009	
	β_2	-5.146				18.706	
	β_3	14.320				5.993	
الگوی (T۴)	β_0	-17.655	200	6630	0.986	-18.592	0.976
	β_1	0.018				0.017	
	β_2	5.469				4.320	
	β_3	6.304				7.386	
الگوی (T۵)	β_0	-2.913	200	6630	0.856	-5.235	0.854
	β_1	0.758				0.681	
	β_2	1.589				2.926	
الگوی (T۶)	β_0	-5.47	200	4420	1.744	-4.439	0.744
	β_1	0.757				0.790	
	β_2	2.364				1.764	

الگوریتم رگرسیون بردار پشتیبان (SVR)

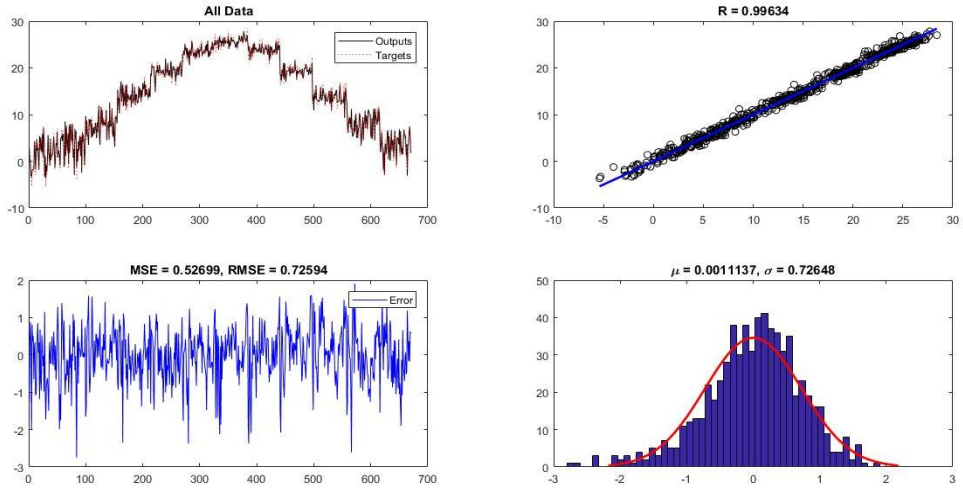
SVR به عنوان یک روش یادگیری ماشین به‌منظور ترمیم بارش و دمای ماهانه مشهد برازش داده شد. نتایج حاصل از برازش الگوی SVR به دمای ماهانه ایستگاه مشهد در برابر پنج ایستگاه مبنا در جدول ۱۰ آمده است. روش تابع کرنل[1] رادیال[2] برای الگوریتم SVR انتخاب شد. پارامترهای ورودی هزینه و اپسیلون در مرحله اول به طور پیش فرض به ترتیب ۱ و ۰/۱ را در نظر گرفته شد. بازه ای از اعداد حقیقی در مراحل بعدی برای هر کدام ازاین دو پارامتر آزمایش و بهترین خروجی انتخاب شد. بازه به کار رفته در مرحله دوم برای هزینه اعداد 2^{0}، 2^{1}، ... ، 2^{5} و برای اپسیلون اعداد ۰، ۰/۱، ۰/۲، ... ، ۱ را دو به دو آزمون می‌کند (جمعا ۶۶ آزمون). بهترین پارامترهای انتخاب شده در جدول ۱۰ آمده است. پارامتر هزینه در مرحله سوم در بازه $2^{0.5}$ تا 2^{8} و اپسیلون بین صفر تا ۰/۲ با فاصله‌های ۰/۰۱ تغییر می‌کند (۳۳۶ آزمون). با توجه به بهترین پارامترهای مراحل قبل، در مرحله چهارم هزینه 2^{0}، 2^{1}، 2^{3} و اپسیلون ۰/۰۱ تا ۰/۱۵ با فاصله‌های ۰/۰۱ انتخاب شدند (۴۵ تکرار).

1- SVM-Kernel
2- Radial

جدول ۹- جدول ضرایب اصلاح شده پنج الگوی انتخابی بارش با الگوریتم ژنتیک و الگوریتم کلونی مورچگان

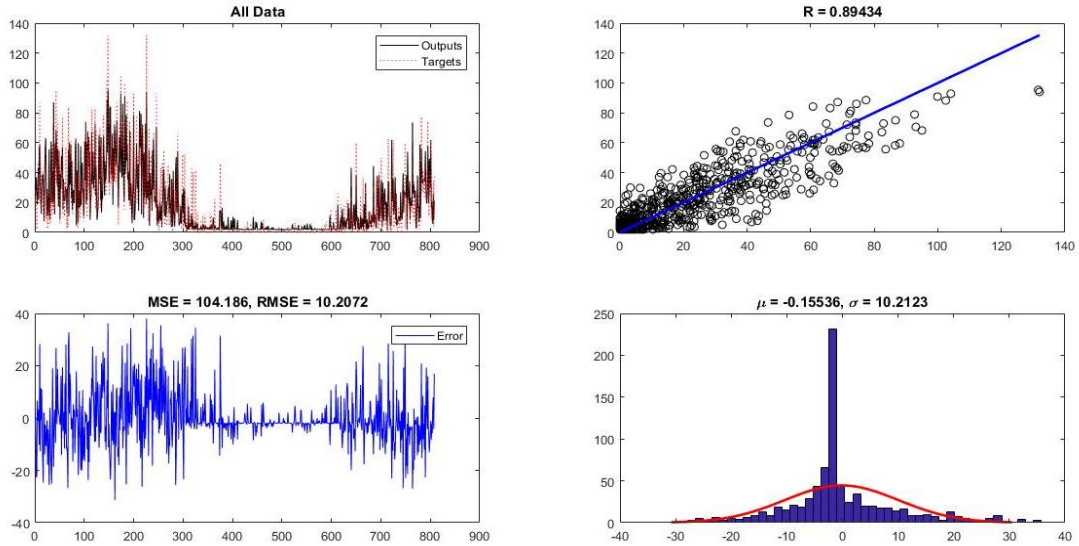
Table 9- Improved coefficients of five selected rainfall patterns with genetic algorithm and Ant colony algorithm

	ضرایب Coefficients	برآورد ضرایب با GA Coefficient estimate by GA	بیشترین تکرار Max iteration	NFE	$RMSE_{GA}$	برآورد ضرایبACO Coefficient estimate by ACO	$RMSE_{ACO}$
الگوی (P۱)	β_0	-0.0026	1000	23030	2.560	-2.768×10^{-7}	2.559
	β_1	0.276				0.262	
	β_2	0.123				0.137	
	β_3	0.665				0.654	
الگوی (P۳)	β_0	0.122	2000	66030	2.628	4.618×10^{-10}	2.628
	β_1	0.101				0.119	
	β_2	0.325				0.312	
	β_3	0.717				0.702	
الگوی (P۵)	β_0	0.00028	3000	99030	2.673	2.304×10^{-6}	2.673
	β_1	0.116				0.146	
	β_2	0.148				0.168	
	β_3	0.794				0.726	
الگوی (P۷)	β_0	0.00065	3000	99030	2.925	1.753×10^{-6}	2.925
	β_1	0.485				0.443	
	β_2	0.235				0.286	
	β_3	0.255				0.238	
الگوی (P۹)	β_0	0.00055	3000	99030	2.649	6.251×10^{-9}	2.648
	β_1	0.280				0.296	
	β_2	0.092				0.032	
	β_3	0.775				0.820	



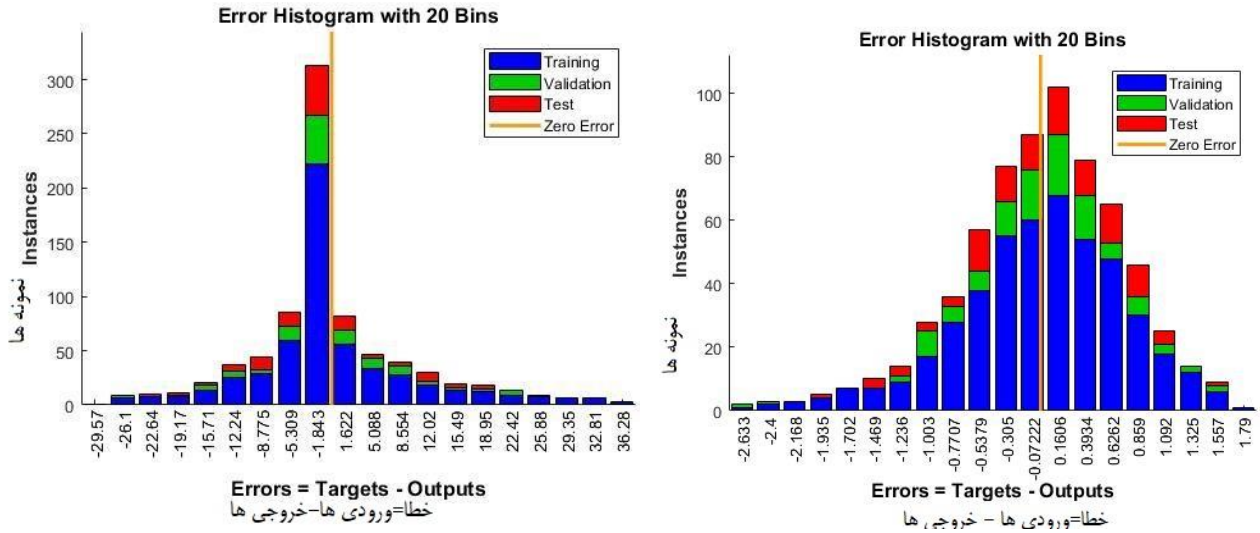
شکل ۳ -رسم کل دمای ماهانه مشاهده‌ای مشهد در برابر داده‌های پیش‌بینی شده توسط شبکه عصبی، ضریب تعیین، نمودار سری زمانی مقادیر خطا و توزیع μ

Figure 3- A graph of the observed monthly temperature of Mashhad against predicted data by network, determination coefficient, time series of error values and μ distribution



شکل ٤- رسم کل بارش ماهانه مشاهده‌ای مشهد در برابر داده‌های پیش‌بینی شده توسط شبکه عصبی، ضریب تعیین، نمودار سری زمانی مقادیر خطا و توزیع μ

Figure 4- A graph of the observed monthly rainfall of Mashhad against predicted data by network, determination coefficient, time series of error values and μ distribution



شکل ٥- توزیع خطاهای مدل‌سازی بارش مشهد (راست) و دمای مشهد (چپ) در بخش‌های آموزش (آبی)، اعتبارسنجی (سبز) و آزمایش (قرمز)

Figure 5- Distribution of modeling errors in Mashhad precipitation (right) and Mashhad temperature (left) in education (blue), validation (green) and experiment (red)

میزان خطا (MSE) در هر مرحله با نمودار پله رنگ نمایش داده می شود. نمودار مراحل ۳ و ۴ جدول ۱۰ در شکل ۶ آمده است. محور افقی مقدار اپسیلون و محور عمودی هزینه را نشان می‌دهد. قسمت‌های تیره تر خطای کمتری دارد. همانگونه که مشاهده می‌شود این روش دقت برآورد دما را در به طور قابل ملاحظه‌ای بالا می‌برد (RMSE=0.561).

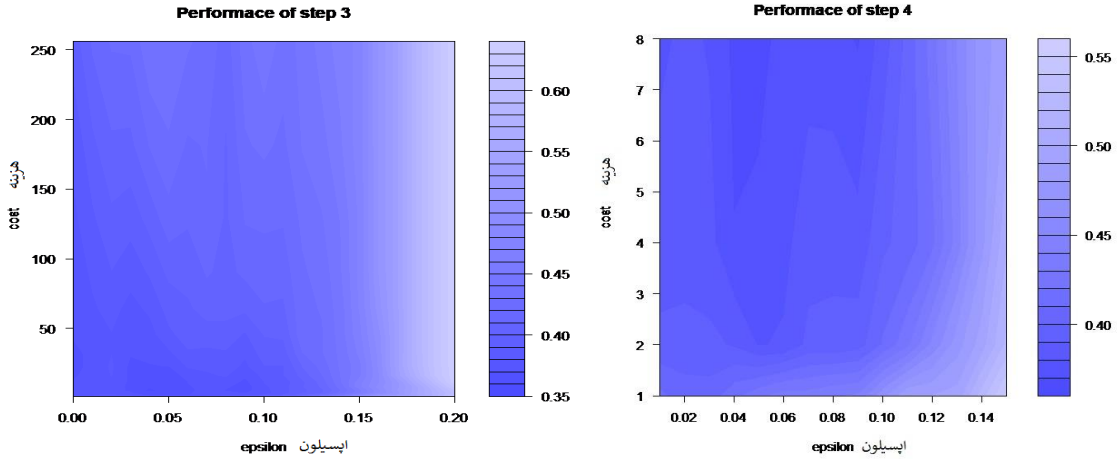
با توجه به عملکرد خوب این روش در افزایش دقت برآورد دمای ماهانه ایستگاه مشهد جاگذاری[1] مفقودی‌های دما به روش بهینه‌سازی با رگرسیون بردار پشتیبان انجام و سری زمانی دراز مدت آن در شکل ۹ آمده است. مفقودی‌های برآورد شده با رنگ مجزا نمایش داده شده است.

1- Imputation

جدول ۱۰- جدول ضرایب اصلاح شده الگوهای دما با الگوریتم رگرسیون بردار پشتیبان

Table 10- Temperature correction coefficients with support vector regression algorithm

مرحله Step	هزینه Best cost	اپسیلون Best epsilon	گاما Gama	وزن‌ها (W) Weights (W)					ضریب b	تعداد بردارهای پشتیبان Number of support vectors	RMSE
				TAsh	TBai	TGau	TSer	TTed			
اول 1	1	0.1	0.2	5.32	7.03	7.21	6.79	6.39	0.176	84	0.629
دوم 2	4	0.1	0.2	5.39	8.01	7.94	7.65	6.82	0.179	74	0.592
سوم 3	5.66	0.04	0.2	5.27	8.41	7.72	8.61	7.85	0.129	288	0.567
چهارم 4	8	0.05	0.2	5.09	8.58	8.19	8.71	7.73	0.187	241	0.561



شکل ٦- نمایش نموداری میزان خطای مراحل ۳ و ٤ الگوریتم SVR برای دمای مشهد

Figure 6- Graph showing the error rate of steps 3 and 4 of the SVR algorithm for temperature of Mashhad

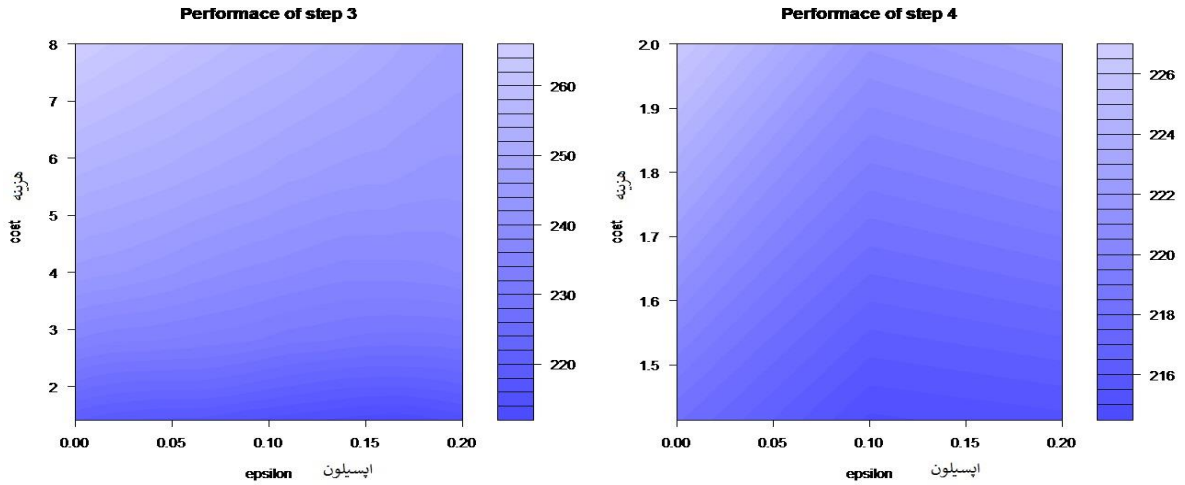
برازش الگوریتم SVR به داده‌های بارش مشهد

ترمیم بارش با الگوریتم SVR مشابه دما انجام می‌شود. پارامتر هزینه و اپسیلون در مراحل اول و دوم مشابه دما انتخاب می‌شود. ضرایب و پارامترها در این مرحله تغییر نکرد، لذا بازه‌های منتخب را با توجه به شکل خروجی تغییر می‌دهیم. به طوری‌که در مرحله سوم پارامتر هزینه در بازه $2^{0.5}$ تا 2^{3} با تأخیرهای ۰/۵ برای توان و اپسیلون ۰ تا ۰/۲ با تأخیرهای ۰/۰۱ در نظر گرفته و آزمون می‌شود. در این مرحله با کاهش اندکی در خطا (RMSE) روبرو می‌شویم (۱۱/۸). بهترین پارامترها برای هزینه و اپسیلون به ترتیب ۱/۴ و ۰/۱۶ است (جدول ۱۱ و شکل ۷).

جدول ۱۱- جدول ضرایب اصلاح شده الگوهای بارش با الگوریتم رگرسیون بردار پشتیبان

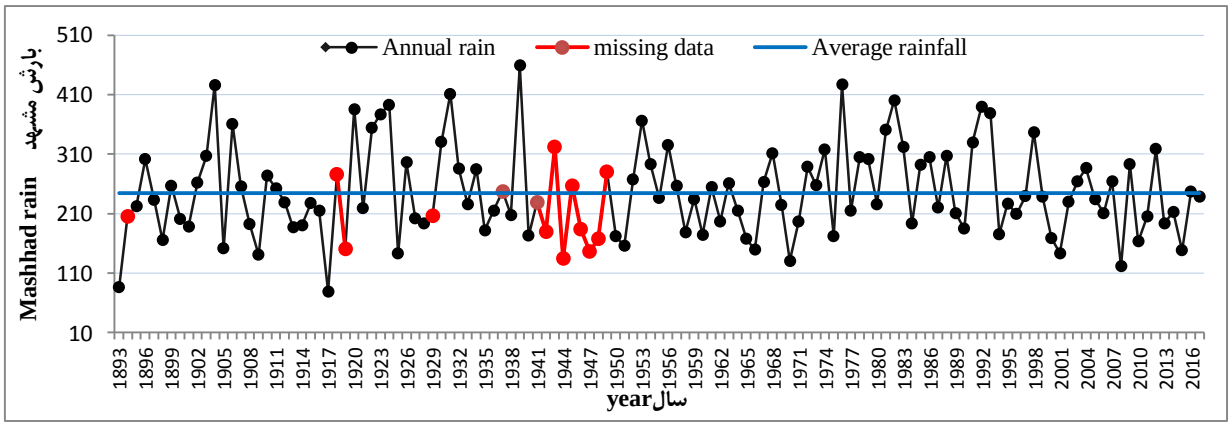
Table 11- Corrected coefficients of rainfall patterns with support vector regression algorithm

مرحله Step	هزینه Best cost	اپسیلون Best epsilon	گاما Gama	وزن‌ها (W) Weights (W)						ضریب b	تعداد بردارهای پشتیبان Number of support vectors	RMSE
				RAsh	RBai	RKer	RKus	RRep	RSer			
اول 1	1	0.1	0.167	13.97	14.16	7.87	13.30	12.07	19.04	-1.24	381	12.18
دوم 2	1	0.1	0.167	13.97	14.16	7.87	13.30	12.07	19.04	-1.24	381	12.18
سوم 3	1.4	0.16	0.167	16.23	16.11	8.17	14.30	12.33	20.84	-1.32	328	11.89
چهارم 4	1.4	0.1	0.167	16.36	16.33	8.12	14.12	12.60	20.28	-1.31	382	11.89



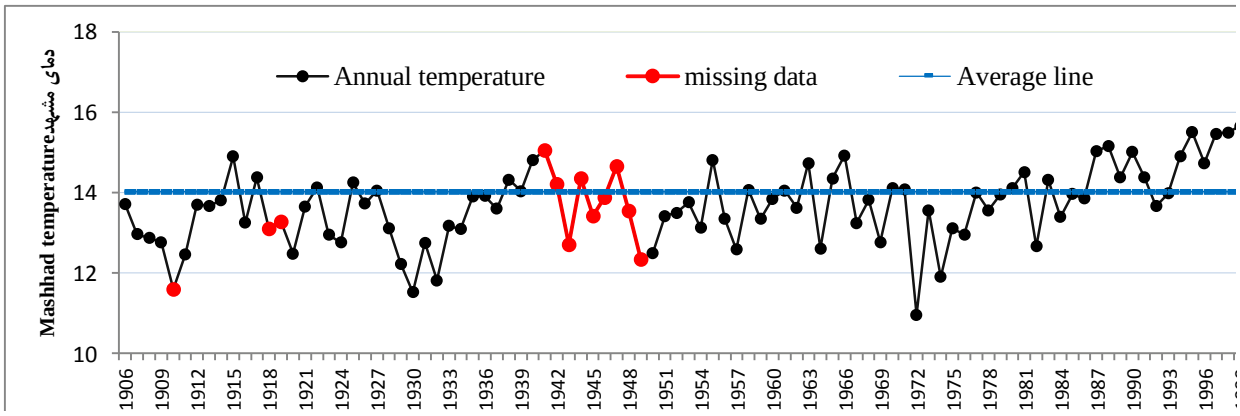
شکل ۷- نمایش نموداری میزان خطای مراحل ۳ و ٤ الگوریتم SVR برای بارش مشهد

Figure 7- Graph showing the error rate of steps 3 and 4 of the SVR algorithm for precipitation of Mashhad



شکل ۸- سری زمانی ۱۲۵ ساله بارش مشهد پس از ترمیم و تکمیل داده‌های ماهانه با روش بهینه‌سازی ژنتیک الگوریتم

Figure 8- The 125-year time series of Mashhad precipitation after the restoration and completion of monthly data using Genetic Optimization Algorithm



شکل ۹- سری زمانی ۱۱۲ ساله دمای مشهد پس از ترمیم و تکمیل داده‌های ماهانه با روش رگرسیون بردار پشتیبان

Figure 9- The 112-year time series of Mashhad temperature after the restoration and completion of monthly data using Support vector regression

نتیجه‌گیری

ترمیم و برآورد مفقودی‌های دما و بارش ماهانه طولانی مدت مشهد هدف این مقاله است. انتساب مقادیر مفقود متغیرهای هواشناسی با کمترین خطای ممکن همواره از اهمیت بالایی برخوردار بوده است. ایستگاه‌هایی از کشورهای مجاور به‌عنوان ایستگاه‌های مبنا انتخاب شدند. ترمیم اولیه داده‌های بارش با برازش ده الگوی رگرسیونی چندگانه (با ضرایب‌تعیین ۰/۶۳ تا ۰/۸۱) و شش الگو برای دمای ماهانه (۰/۹۸۶ تا ۰/۹۹۳) انجام شد. ضرایب الگوهای رگرسیونی با روش‌های GA و ACO به منظور بالا بردن دقت برآوردها، بهینه شدند. روش ANN و SVR نیز برای الگوسازی این داده‌ها نیز به کار گرفته شدند.

نتایج نشان داد روش‌های بهینه‌سازی GA و ACO مقدار خطای RMSE را برای برآورد بارش ماهانه مشهد به طور قابل ملاحظه‌ای نسبت به روش رگرسیونی کاهش می‌دهد (جداول ۳ و ۸). خطای الگوی (P۱) در بهترین حالت از ۹/۷۹ با روش رگرسیونی به ۲/۵۶ میلیمتر با GA و ACO کاهش می‌یابد. کمترین معیار خطای (RMSE) بین الگوهای رگرسیونی دمای مشهد ۰/۹۸۶ میلی‌متر بود. این معیار با روش ANN به ۰/۷۲۶ میلی‌متر و با SVR نیز به ۰/۵۵۱ کاهش یافت. مقایسه نتایج الگوهای ترمیم دما و بارش نشان می‌دهد که روش‌های تکاملی فوق برای برآورد بارش و روش‌های یادگیری ماشین برای برآورد دما عملکرد بهتری دارند. لذا بارش ماهانه ۱۲۵ سال مشهد با روش GA ترمیم و ارائه شدند (شکل ۸). همچنین SVR برای ترمیم دمای ماهانه ۱۱۲ سال مشهد انتخاب شد (شکل ۹). این آمار طولانی مدت می تواند مبنای ارزشمندی برای تحقیقات آب و هواشناسی باشد.

سپاسگزاری

از آقای مهندس حجت رضایی پژند برای راهنمایی‌های ایشان در این پژوهش سپاسگزارم.

منابع

1- Arghami N.R., Sanjari N., and Bozorgnia A. 2010. Elementary Survey Sampling, Mashhad University Pub, pp 435.
2- Dastorani M.T., Moghadamnia A., Piri J., and Rico-Ramirez M. 2010. Application of ANN and ANFIS models for reconstructing missing flow data. Environmental Monitoring and Assessment 166(1-4): 421-434.
3- Dingman S.L. 2002. Physical Hydrology, Second Edition, PRENTICE HALL.
4- Dipak V.P., and Bichkar R.S. 2010. Multiple Imputation of Missing Data with Genetic Algorithm based Techniques, IJCA Special Issue on Evolutionary Computation for Optimization Techniques.
5- El Assaad H., Samé A., Govaert G., and Aknin P. 2016. A variational Expectation–Maximization algorithm for temporal data clustering, Computational Statistics and Data Analysis 103: 206–228.
6- Farzandi M., Rezaee-Pazhand H., and Sanaeinejad H. 2014. Restoration and development of 127 years of monthly temperature in Mashhad, Journal of Climate Research 5(17&18): 123-111. (In Persian with English abstract)
7- Ghahraman B., and Ahmadi F. 2007. Application of Geo statistics in Time series: Mashhad Annual Rainfall, Iran-Watershed Management Science & Engineering 1(1): 7-15.
8- Golabi M.R., Akhond-Ali A.M., and Radmanesh F. 2013. Comparison of performance of different artificial neural network algorithms, Journal of Applied Geosciences Research 30: 131-169. (In Persian)
9- Iqbal M., Wen J., Wang Sh., Tian Hu., and Adnan M. 2018. Variations of precipitation characteristics during the period 1960-2014 in the Source Region of the Yellow River, China. Journal of Arid Land 10(3): 388-401.
10- Jacob D., Reed D.W., and Robson A.J. 1999. Choosing a pooling group. Flood Estimation Handbook. Vol. 3. Institute of Hydrology, Wallingford, UK.
11- Khalili A., and Bazrafshan J. 2008. Evaluation of drought duration risk using annual secular precipitation data in ancient stations of Iran, Journal of Geophysical, Volume 2, Number 2. (In Persian with English abstract)
12- Liao W., Li D., and Cui Sh. 2018. A heuristic optimization algorithm for HMM based on SA and EM in machinery diagnosis, J Intell Manuf © Springer Science 29(8): 1845-1857.
13- Little R., JA Rubin D., B. 2002. Statistical analysis with missing data. John Wiley & Sons, 408 pages.
14- Motia Ghader H., Lotfi Sh., and Seyyedsafelan M.M. 2010. A Review of Some Intelligent Optimization Techniques, Shabestar Branch of Islamic Azad University Publishing, 215 pp.
15- Preis A., and Ostfeld A. 2008. A coupled model tree–genetic algorithm scheme for flow and water quality predictions in watersheds. Journal of Hydrology (Elsevier) 349: 364–375.
16- Ranhao S., Baiping Z., and Jing T. 2008. A Multivariate Regression Model for Predicting Precipitation in the Daqing Mountains, Mountain Research and Development 28(3): 318-325.
17- Rezaee-Pazhand H., and Bozorgnia A. 2002. Nonlinear Regression Analysis with application, Mashhad University Pub, 400 pp. (In Persian)

18- Safavi S.A.A., Pour Jafarian N., and Safavi S.A. 2014. Optimization based on Meta-heuristic algorithms, Publishing Institute of Academic Publishers. (In Persian with English abstract)
19- Shanmuganathan S., and Samarasinghe S. 2016. Artificial Neural Network Modelling, Springer International Publishing Switzerland, 468 pages.
20- Smithsonian Institution. (1927, 1934, 1947): World weather records, 1910-1920., 1921-1930., 1931-1940., Smithson. Miss C. Collect. 79,90,105. (Publication2913.,3216.,3803)
21- Smola A., and Vishwanathan S.V.N. 2008. Introduction to Machine Learning. Cambridge university press 234 pages.
22- Souri A. 2017. Advanced econometric studies, c. 2, with the use of stata12 and eviews8, Culture Publishing, 1022 pages. (In Persian)
23- Yozgatligil C., Aslan S., Iyigun C., and Batmaz I. 2013. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data, Theory Apply Climatology 112: 143–167.

Journal of Water and Soil
Vol. 33, No. 2, May-Jun. 2019, p. 361-377

نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۳، شماره ۲، خرداد – تیر ۱۳۹۸، ص. ۳۷۷-۳۶۱

Imputation of Missing Meteorological Data with Evolutionary and Machine Learning Methods
Case Study: Long-term Monthly Precipitation and Temperature of Mashhad

M. Farzandi[1]- H. Sanaeinejad[2*]- B. Ghahraman[3]- M. Sarmad[4]

Received: 29-07-2018
Accepted: 15-04-2019

Introduction: Temperature and precipitation are two of the main variables in meteorology and climatology. These are basic inputs in water resource management. The length of the statistical period plays a pivotal role in the accurate analysis of these variables. Observation data at Iran's first synoptic station from 1330 (1951) is available at the Iranian Meteorological Organization website The historical monthly precipitation and temperature of five stations in Iran is available since 1880 with missing data. These data measured by the Embassy of the United States and Britain from the Qajar period and recorded in World Weather records books. These synoptic stations include Mashhad, Isfahan, Tehran, Bushehr, and Jask. The monthly missing data were predominantly recorded during World War II (1941-1949). Unfortunately, these data have missing. Therefore, the accuracy of simulating these variables is very important. The current research aimed to predict the missing values of monthly temperature and precipitation in Mashhad station. The stations in the neighboring countries were selected due to the distance to Mashhad, relationship, and completeness of data since 1880, as the predictive variables. Monthly precipitation of Ashgabat from Tajikistan and Sarakhs, Kooshkah, Bayram Ali, Kerki and Repetek from Turkmenistan were selected as an independent variable in the making of Missing Rainfall in Mashhad. Also, the temperature of Ashgabat, Bayram Ali, Gudan, Sarakhs, and Tajan were selected to restore the monthly temperature of the Mashhad station. This research has fitted ten multiple regression models to monthly rainfall of Mashhad station and has fitted 6 multiple regression to the monthly temperature of Mashhad. then the parameters of these patterns are optimized by genetic and Ant Colony algorithm. Also, the Artificial Neural Network (MLP) model and Support vector regression have been selected and implemented in order to simulate monthly precipitation and temperature data of Mashhad.

Materials and Methods: In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). Genetic algorithm (GA) is a metaheuristic inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms (EA). Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover, and selection. Ant colony optimization algorithm (ACO) is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. This algorithm is a member of the ant colony algorithms family, in swarm intelligence methods, and it constitutes some metaheuristic optimizations. Artificial neural networks are one of the main tools used in machine learning. As the "neural" part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize. In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt

1, 2 and 3- Ph.D. Student, Associate Professor and Professor Department of Water Engineering, College of Agriculture, Ferdowsi University of Mashhad, Respectively
(*- Corresponding Author Email: sanaei@um.ac.ir)
4- Associate Professor, Department of Statistic, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

scaling exist to use SVM in a probabilistic classification setting).

Results and Discussion: At the first stage, several multiple regressions were fitted to monthly precipitation (with coefficients ranging from 0.63 to 0.81) and six patterns for monthly temperature (0.986-0.993). Afterward, GA and ACO were applied to improve the accuracy of the selected regression models by optimizing their parameters. At the next stage, ANN and SVR were used to estimate the monthly missing values separately. Finally, the results of the previous stages were compared using the root mean square error (RMSE), and the optimal models were applied to determine the missing values of monthly temperature and precipitation of Mashhad. The results showed that the Genetic Algorithm and Ant Colony increase the accuracy of the estimation of missing rainfall data significantly more than the previous methods. The lowest error criterion (RMSE) between regression patterns is 9.8 millimeters. By genetic algorithm, this criterion is reduced to 2.56 mm, and by ant colony algorithm to 2.559.

Conclusion: Comparison of the above methods in restoration temperature and precipitation shows that evolutionary methods (GA and ACO) are the best for estimating the missing monthly precipitation and machine learning methods (ANN and SVR) are the best to imputation missing data of monthly temperature.

Keywords: Ant colony, Artificial neural network, Genetic algorithm, Missing data, Support vector regression