تحقیقات منابع آب ایران
Iran-Water Resources Research

سال سیزدهم، شماره ۴، زمستان ۱۳۹۶
Volume 13, No. 4, Winter 2018 (IR-WRR)
۱۷۹-۱۸۳

یادداشت فنی

پیش‌بینی بارش ماهانه با مدل درختی M5 و مقایسه آن با روش‌های کلاسیک آماری (مطالعه موردی: ایستگاه سینوپتیک ارومیه)

شبنم وکیلی[۱*]

چکیده

در این تحقیق به منظور تخمین داده‌های بارش ماهانه ایستگاه ارومیه که از سال ۲۰۰۶ تا ۲۰۰۷ مفقود فرض شده است از روش‌های آماری کلاسیک و مدل درختی M5 با استفاده از نرم‌افزار Weka و به کارگیری ایستگاه‌های مهاباد، خوی، سلماس، تکاب و ماکو استفاده شده است. در بین ایستگاه‌های مورد مطالعه، ایستگاه مهاباد با r=0.90 بیشترین همبستگی را با ایستگاه ارومیه داشت. ۲۶ سناریو از آمار ده ساله ایستگاه‌های مجاور در تخمین بارش ماهانه ایستگاه شاهد (ارومیه) به نرم‌افزار Weka معرفی شده است که از بین سناریوها، سناریویی که شامل سه ایستگاه مهاباد، ماکو و تکاب با RMSE=9.64، r=0.9 و MAE=7.19 بود به دلیل کم بودن پارامترهای ورودی به مدل به عنوان ساده‌ترین و دقیق‌ترین سناریو به مدل تعریف گردید. از بین روشهای کلاسیک ، روش تخمین‌گر منفرد (SIB) بهترین روش با بیشترین ضریب همبستگی و کمترین خطا (MAE=7.07، RMSE=10.51 و r=0.90) انتخاب شده است. مدل درختی M5 در برآورد داده‌ها با MAE=7.29، RMSE=9.94 و r=0.91 بهترین عملکرد را داشته است و به دلیل ارائه روابط خطی ساده و قابل فهم به عنوان روشی جایگزین و کاربردی در محاسبه داده‌های بارش ماهانه مورد توجه قرار می‌گیرد.

کلمات کلیدی: روش‌های آماری کلاسیک، مدل درختیM5، نرم‌افزار Weka، ضریب همبستگی.

تاریخ دریافت مقاله: ۹۵/۱۰/۲
تاریخ پذیرش مقاله: ۹۶/۲/۶

۱- مدرس دانشگاه آزاد تبریز
*- نویسنده مسئول
بحث و مناظره (Discussion) در مورد این مقاله تا پایان خرداد ۱۳۹۷ امکانپذیر است.

Technical Note

Forecasting of Monthly Precipitation Using M5 Model Tree and Classic Statistical Methods (Case Study: Oroumieh Synoptic Station)

*Sh. Vakili[1] **

Abstract

This study is carried out to estimate monthly rainfall data of Oroumieh station which are assumed to be lost from 2006 to 2007. This is performed using classic statistical methods and M5 model tree employing the software Weka based on data from Mahabad, Khoy, Salmas, Makoo, and Tekab stations. Among the studied stations, Mahabad station (R=0.90) had the highest correlation with Oroumieh station. From the 26 scenarios which were introduced to Weka software for 10 year data of the nearby stations, the one which included three stations of Mahabad, Makoo and Tekab with MAE=7.19, R=0.90, and RMSE=9.64 was defined as the simplest and most accurate scenario due to the less input parameters to the model. Among the classical methods, the single best estimator (SIB) method has been selected as the best method with the highest correlation coefficient and the lowest error (R=0.90, RMSE=10.51, and MAE=7.07). M5 model tree had the best performance in estimating data (R=0.91, RMSE=9.94, and MAE=7.29) and was considered as an alternative and applied method in the calculation of monthly precipitation data due to simple linear and comprehensible relationships.

Keywords: Classic Statistical Methods, M5 Model Tree, Weka Software, Correlation Coefficient.

Received: December 22, 2016
Accepted: April 26, 2017

1- Lecturer, Tabriz Azad University, Tabriz, Iran. Email: Vakili_c@yahoo.com
*- Corresponding Author

۱- مقدمه

اساس و پایه مطالعات هیدرولوژی داده‌های آماری مورد اعتماد و مورد قبول می‌باشد. با استفاده از خلاءهای آماری گسسته و پیوسته در اغلب داده‌های هیدرولوژی مانند بارش، به دلیل عدم ثبت آمار، آمار غلط و خرابی یا از بین رفتن دستگاه‌های اندازه‌گیری، تخمین و برآورد این داده‌ها ضروری می‌باشد. بدین منظور امروزه سامانه‌های هوشمند همچون روش شبکه عصبی مصنوعی و برنامه‌ریزی ژنتیک در مدل‌سازی فرایندهای هیدرولوژیکی و مهندسی آب مورد استفاده قرار گرفته‌اند. در سالهای اخیر نیز مدل‌های درختی بعنوان یکی از تکنیکهای داده کاوی در پیش‌بینی پارامترهای هیدرولوژیکی و هواشناسی مورد توجه قرار گرفته‌اند. از جمله تحقیقات انجام یافته با مدل درختی M5 می‌توان به مطالعات Emamifar et al. (2013) مربوط به مدل‌سازی میانگین دمای روزانه با استفاده از مدل درختی M5، مطاله Ditthakit and Chinnarasi (2012) جهت تخمین ضریب تشتک تبخیر با استفاده از مدل درختی M5، مطالعه Etemad-shahidi and Mahjoobi (2009) در برآورد ارتفاع موج در دریاچه و مقایسه دو روش شبکه عصبی و مدل درختی M5، Sattari et al. (2013)جهت برآورد دقیق مقدار تبخیر و تعرق مرجع روزانه و تحقیقات Rahimikhoob et al. (2013) در تبدیل تبخیر تشتک به تبخیر و تعرق مرجع در یک ناحیه آب و هوایی خشک با استفاده از روش‌های مدل درختی و روشهای مرسوم اشاره کرد. در این تحقیق جهت برآورد داده‌های بارش ماهانه ایستگاه ارومیه که فرض بر این است که این داده‌ها از سال ۲۰۰۶ تا ۲۰۰۷ مفقود می‌باشند از روش‌های آماری کلاسیک و مدل درختی با استفاده از نرم‌افزار Weka بهره گرفته شده است .

۲- مواد و روش ها

داده‌کاوی فرآیندی است که از ابزارهای تحلیلی گوناگونی برای کشف الگوها و روابط بین داده‌ها استفاده می‌کند که ممکن است برای اعتبار بخشیدن به پیش‌بینی استفاده شود (Berry et al., 1997). روش‌های کلاسیک جهت پیدا کردن داده‌های گم شده متمرکز روی مقیاس‌های زمانی ماهانه و سالانه می‌باشند که روند بارش واقعی را نشان می‌دهند. این روش‌ها شامل روش میانگین‌گیری ساده (AA^1)، نسبت نرمال (NR^2)، بهترین تخمین‌گر منفرد (SIB^3)، روش نسبت‌ها (UK^4) و روش تحلیل رگرسیون (REG) می‌باشند. الگوریتم مدل درختی M5 به وسیله کویلن در سال ۱۹۹۲ توسعه یافته است (Quinlan, 1992). مدل درختی ترکیب یک درخت تصمیم‌گیری معمولی با امکان‌پذیری توابع رگرسیون خطی در سطوح می‌باشد. ساختار مدل درختی شبیه درخت‌های تصمیم می‌باشد (Dimitri et al., 2002). در این مطالعه جهت بهره‌گیری از مدل درختی M5 از نرم‌افزار Weka استفاده شده است. جهت مقایسه نتایج مدل M5 با روشهای آماری و کلاسیک شامل روش میانگین‌گیری ساده، نسبت نرمال، بهترین تخمین‌گر منفرد، روش نسبت‌ها و روش تحلیل رگرسیون از شاخص‌های آماری مانند ضریب هبستگی (r) و ریشه میانگین مربعات خطا (RMSE) و خطای مطلق میانگین (MAE) استفاده شده است. جهت مقایسه نتایج مدل M5 باروشهای آماری و کلاسیک از آمار ده ساله ایستگاه‌های هواشناسی استان آذربایجان‌غربی شامل ارومیه، تکاب، خوی، مهاباد، سلماس و ماکو از شاخص‌های آماری مانند ضریب هبستگی (r)، ریشه میانگین مربعات خطا (RMSE) و خطای مطلق میانگین (MAE) با فرض مفقود بودن آمار دو سال ۲۰۰۶ و ۲۰۰۷ در ایستگاه ارومیه استفاده شده است. در این مطالعه ابتدا از آمار ده ساله تمامی ایستگاهها جهت برآورد روشهای کلاسیک آماری طبق روابط مربوط به هر یک از آن‌ها استفاده شده است، سپس به برآورد شاخص‌های آماری پرداخته شده است. جهت وارد کردن داده‌های ایستگاه‌ها به نرم‌افزار Weka از سناریوهای شامل دو ایستگاه، سه ایستگاه، چهار ایستگاه و پنج ایستگاه استفاده گردید. داده های هر کدام از سناریوها وارد مدل شد و رابطه ایی خطی برای هر کدام ارائه گردید. بعد از برآورد شاخص آماری سناریوها، روشهای کلاسیک و مدل M5 و سناریوها جهت ارائه نتایج با یکدیگر مقایسه شدند.

۳- تحلیل نتایج

۳-۱- بررسی نتایج روشهای آماری کلاسیک

ابتدا داده‌های مفقود ایستگاه شاهد با استفاده از روش‌های آماری کلاسیک مورد بررسی قرار گرفته است. برای هر روش مدلی خطی برازش داده شده است و مقدار ضریب همبستگی بین داده‌های مشاهداتی و محاسباتی حاصل از روش مورد بررسی محاسبه شده است. نتایج حاصل از این مقایسه در شکل‌های ۱ تا ۵ نشان داده شده است.

با توجه به نمودارها مشاهده می‌شود که روش SIB با ۰/۹۱ و روش REG با ضریب همبستگی ۰/۸۶ نسبت به روش‌های دیگر بهترین نتایج را ارائه داده‌اند. با توجه به کم بودن نمونه مورد بررسی ضریب همبستگی قابل قبولی برای تمامی روش‌ها حاصل شده است.

۳-۲- نتایج مدل درختی M5

جهت تعیین همبستگی بین ایستگاه‌های منطقه، ضریب همبستگی ایستگاه‌ها برآورد شده و در جدول ۱ ارائه شده است.

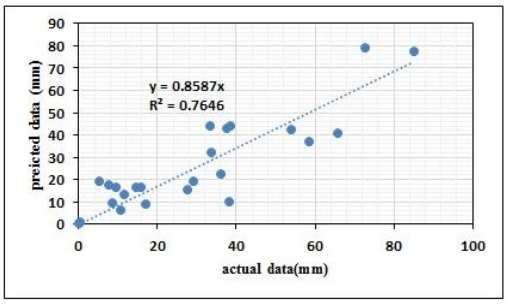

Fig.1- Comparison of the actual and predicted data by AA

شکل ۱- مقایسه مقادیر واقعی و محاسباتی با روش AA

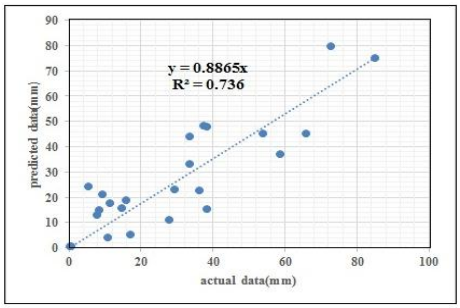

Fig.2- Comparison of the actual and predicted data by NR

شکل ۲- مقایسه مقادیر واقعی و محاسباتی با روش NR

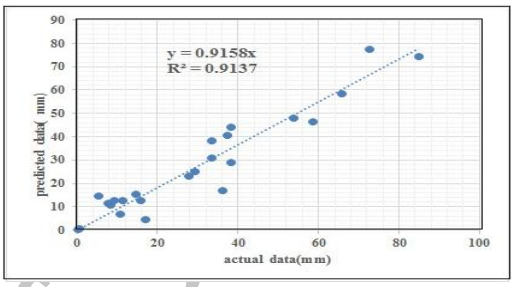

Fig.3- Comparison of the actual and predicted data by SIB

شکل ۳- مقایسه مقادیر واقعی و محاسباتی با روش SIB

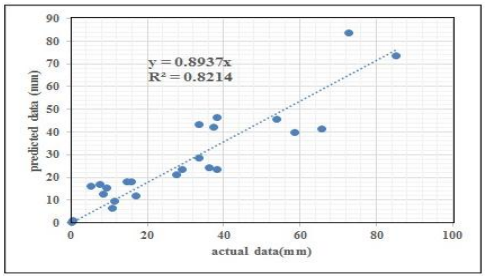

Fig.4- Comparison of the actual and predicted data by UK

شکل ۴- مقایسه مقادیر واقعی و محاسباتی با روش UK

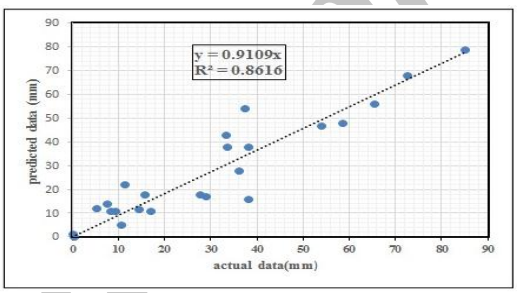

Fig.5- Comparison of the actual and predicted data by REG

شکل ۵- مقایسه مقادیر واقعی و محاسباتی با روش REG

Table 1- Correlation of stations

جدول ۱- همبستگی ایستگاه‌های منطقه مورد مطالعه

Prameter	orumieh	salmas	tekab	khoy	mahabad	makoo
orumieh	1	0.3248	0.7108	0.7852	0.9017	0.8682
salmas	0.3248	1	0.21	0.2252	0.4536	0.436
tekab	0.7108	0.21	1	0.7568	0.6004	0.6036
khoy	0.7852	0.2252	0.7568	1	0.766	0.6886
mahabad	0.9017	0.4536	0.6004	0.766	1	0.9576
makoo	0.8682	0.436	0.6036	0.6886	0.9576	1

با مقایسه داده‌های بارش ماهانه ایستگاه‌ها و تعیین همبستگی بین آنها، از بین ایستگاه‌های مورد مطالعه، ایستگاه مهاباد با ضریب همبستگی ۰/۹۰ بیشترین همبستگی را با ایستگاه ارومیه و ایستگاه سلماس با ضریب (۰/۳۲) کمترین همبستگی با ایستگاه ارومیه را داشته‌اند. در این تحقیق، ۲۶ سناریو از ترکیب متفاوت داده‌های ورودی با در نظرگرفتن یک سناریو شامل ۵ ایستگاه، ۵ سناریو شامل ۴ ایستگاه، ۱۰ سناریو شامل ۳ ایستگاه و ۱۰ سناریو با ۲ ایستگاه، وارد مدل درختی شده است که این سناریوها و نتایج مربوط به هر سناریو بر اساس سه آماره r، RMSE و MAE در جدول ۲ نشان داده شده است. در شکل ۶ تا ۹ با برازش مدل خطی بین داده های بارش ماهانه سناریوهای ذکر شده در جدول با مقادیر واقعی همبستگی مقادیر محاسباتی و واقعی مورد بررسی قرار گرفته است.

Table 2- Resulted MAE, RMSE and R for the best scenarios out of 26 scenarios

جدول ۲- نتایج آماره‌های MAE و RMSE و r بهترین سناریوها از بین ۲۶ سناریو

SCENARIOU	r	MAE	RMSE
MAHABAD, MAKOO, KHOY, TEKAB,SALMAS	0.8879	7.3959	10.044
MAHABAD, MAKOO, KHOY, MAKOO	0.9182	7.2925	9.9439
MAHABAD, TEKAB,MAKOO	0.902	7.1945	9.6481
MAHABAD, TEKAB	0.8912	7.0245	9.8481

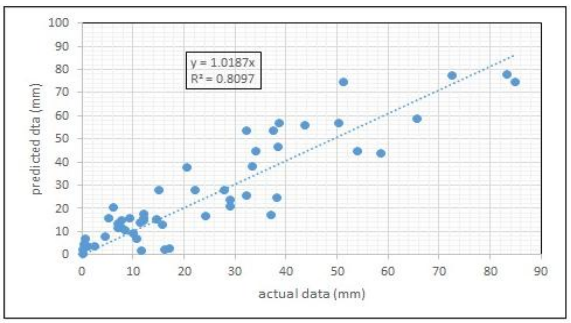

Fig. 6- Precipitation results from Scenario 1 including five stations compared with actual values

شکل۶- مقایسه مقادیر بارش حاصل از سناریوی ۱ با پنج ایستگاه با مقادیر واقعی

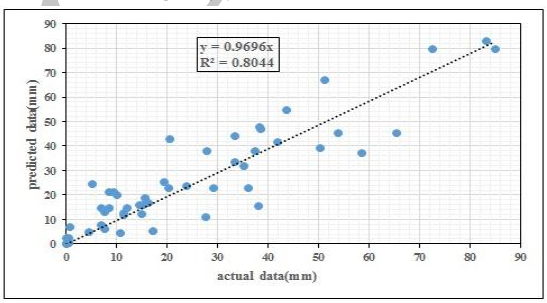

Fig. 7- Precipitation results from Scenario 2 including four stations compared with actual values

شکل ۷- مقایسه مقادیر بارش حاصل از سناریوی ۲ با چهار ایستگاه با مقادیر واقعی

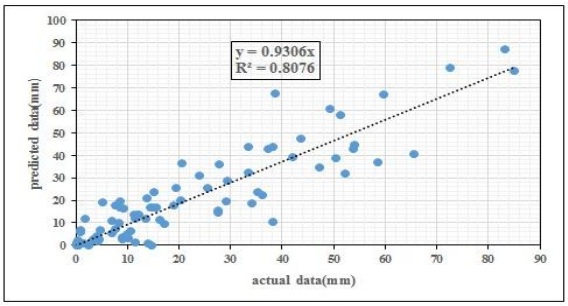

Fig. 8- Precipitation results from Scenario 3 including three stations compared with actual values

شکل ۸- مقایسه مقادیر بارش حاصل از سناریوی ۷ با سه ایستگاه با مقادیر واقعی

۳-۳- مقایسه نتایج مدل درختی M5 و روش‌های کلاسیک

در جدول ۳ روشهای کلاسیک و آماری با مدل درختی با ارائه مقادیر ضریب همبستگی و میانگین خطای مربعات و خطای مطلق حاصل از روشهای مذکور مقایسه گردیده است. در این مقایسه، ملاحظه می‌شود که مدل درختی نتایج قابل قبولی در مقایسه با روشهای کلاسیک ارائه داده است. در شکل ۱۰ مقادیر بارش ماهانه محاسباتی حاصل از روشهای کلاسیک و مدل درختی M5 با مقادیر بارش مشاهداتی ایستگاه شاهد (ارومیه) مقایسه شده است.

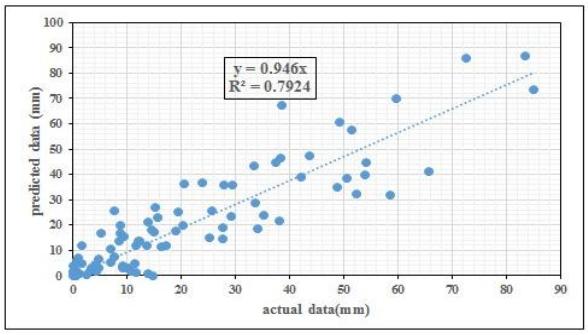

Fig. 9- Precipitation results from Scenario 4 including two stations compared with actual values

شکل ۹- مقایسه مقادیر بارش حاصل از سناریوی ۲۱ با دو ایستگاه با مقادیر واقعی

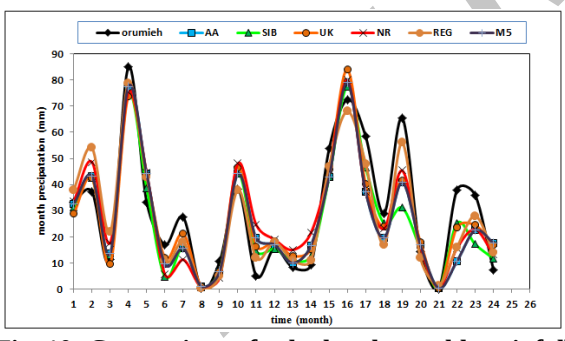

Fig. 10- Comparison of calculated monthly rainfall data of Oroumieh station with classic methods and M5 model tree

شکل ۱۰- مقادیر بارش ماهانه محاسباتی ایستگاه ارومیه با روشهای کلاسیک و مدل درختی M5

۴- نتیجه‌گیری

در این مطالعه، از داده‌های بارش ماهانه ایستگاه‌های سینوپتیک استان آذربایجان غربی شامل ایستگاههای ارومیه، مهاباد، ماکو، تکاب، سلماس و خوی استفاده شده است.

Table 3- Calculated Statistical values of classic methods and M5 model tree

جدول ۳- مقادیر آماره‌های محاسبه شده روشهای کلاسیک و مدل M5

Method	AA	NR	SIB	UK	REG	M5
RMSE	15.9952	16.3822	10.5128	13.5166	10.4234	9.9439
MAE	17.4133	19.2591	7.0720	15.6466	7.0214	7.2925
R	0.8609	0.8759	0.9049	0.8914	0.9026	0.9182

هدف تخمین داده‌های بارش ماهانه مفقود سالهای ۲۰۰۶ تا ۲۰۰۷ ایستگاه شاهد (ارومیه) با استفاده از روشهای کلاسیک و مدل درختی M5می‌باشد. جهت تخمین این داده‌ها ۲۶ سناریو شامل سناریوهای ۵ ایستگاهی، ۴ ایستگاهی، ۳ ایستگاهی و ۲ ایستگاهی به عنوان پارامتر ورودی به مدل M5 معرفی گردید. همانطور که نتایج جدول ۲ نشان می‌دهد، سناریوی شامل ۴ ایستگاه و سناریوی شامل ۳ ایستگاه به عنوان سناریوهای قابل قبول مورد توجه واقع شده‌اند. همچنانکه نتایج نشان می‌دهند سناریوی ۲ با چهار پارامتر ورودی، بهترین نتیجه را می‌دهد (RMSE=9.94, r =0.91 و MAE=7.29) از طرفی سناریوی سه ایستگاهه با تعداد پارامترهای ورودی کمتر، یعنی سه پارامتر، نتیجه قابل قبولی را ارائه می‌دهد. براساس نتایج حاصل از جدول ۲ (RMSE=9.64، r=0.90 و MAE=7.19) می‌توان دریافت گرچه سناریوی چهار پارامتری، از دقت بالایی برخوردار است، منتها تعداد پارامترهای بکار رفته در آن زیاد بوده و شبکه نسبتا پیچیده می‌باشد. قطعاً سناریویی که در آن به پارامترهای کمتر نیاز بوده و دقت بالایی داشته باشد از نظر کاربردی مطلوب تر خواهد بود. از این‌رو سناریویی با سه پارامتر ورودی (ایستگاه‌های مهاباد، تکاب و ماکو)، به عنوان بهترین سناریو جهت ورود به مدل درختی پذیرفته می‌شود. نتایج مشابهی در تحقیق ستاری و نهرین در سال (۲۰۱۳) در برآورد مقدار تبخیر و تعرق مرجع بدست آمده است که سناریویی که تعداد پارامتر کمتر ولی نتایج قابل قبولی ارائه داده است در مقایسه با سناریویی که تعداد پارامتر ورودی بیشتر و نتیجه بهتری داشته است، به دلیل سادگی مدل و نتیجه مطلوب ترجیح داده شده است (Sattari et al., 2013). از بین روشهای کلاسیک نیز روش بهترین تخمین‌گر منفرد (SIB) و روش REG نتایج قابل قبولی ارائه داده‌اند. با مقایسه نتایج روشهای کلاسیک و مدل درختی M5 مشاهده می‌شود که مدل درختی M5 قابلیت مناسبی در پیش‌بینی مقادیر بارش ماهانه داشته و با توجه به اینکه این مدل یک الگوریتم پیش‌بینی بوده و گره‌های درخت با خواص بیشینه خطاهای مورد انتظاری که به عنوان تابعی از انحراف استاندارد پارامترهای خروجی می‌باشد انتخاب می‌شوند، روابط خطی ارائه شده توسط این مدل، متغیر وابسته هدف را مدل می‌کنند. بنابراین، با توجه به ساده و کارآمد بودن این مدل که دقت قابل قبولی نیز در این مطالعه داشته است می‌توان کاربرد این روش را در مباحث مرتبط توصیه نمود.

پی‌نوشت‌ها

1- Simple Arithmetic Averaging
2- Normal Ratio Method
3- Single Best Estimator
4- UK Traditional Method

۵- مراجع

Berry M, Linoff G (1997) Data mining techniques:for marketing, sales, and customer support. New York John Wiley and Sons

Solomatine DP and Xue Y (2004) Pr M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai river in China. Journal of Hydrologic Engineering, 9(6), Doi.org/10. 1061/(ASCE)1084-0699(2004)9:6(491)

Ditthakit P, Chinnarasri C (2012) Estimation of pan coefficient using M5 model tree. American Journal of Environmental Sciences 8(2):95-103

Emamifar S, Rahimikhoob A, Noroozi A A (2013) Daily mean air temperature estimation from MODIS land surface temperature products based on M5 model tree. International Journal of Climatology Int. J. Climatol. 33:3174–3181

Etemad-shahidi A, Mahjoobi J (2009) Comparis on between M5 model tree and neural networks for prediction of significant wave height in lake Superior. Ocean Engineering 36:1175–1181

Quinlan JR (1992) Learning with continuous classes. Proceedings of Australian Joint Conference on Artificial Intelligence. World Scientific Press: Singapore, 343–348

Rahimikhoob A, Asadi M, Mashal M (2013) A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. Water Resour Manage 27:4815–4826

Sattari M , Nahrein F, Azimi V (2013) Forecasting of daily evapotranspiration using artificial neural network model and the model tree M5 case study: Bonab station. Iranian Journal of lrrigation and Drainage 1(7):104-113 (In Persian)