مقایسه کارآیی مدل‌های درختی در محاسبه ضریب پراکندگی طولی آلاینده‌ها در آبراهه‌های مستقیم

حسین نزارتیان[۱]، جواد ظهیری[۲]، سید محمد کاشفی‌پور[۳]

تاریخ ارسال۱۳۹۵/۰۹/۲۱
تاریخ پذیرش۱۳۹۶/۰۳/۱۳

چکیده

مدل‌سازی پیشروی آلاینده‌ها در آبراهه‌های طبیعی یکی از مهم‌ترین مسائل محیط زیست است. ضریب پخشیدگی طولی یکی از پارامترهای اساسی در مدل‌سازی انتشار آلودگی‌ها به‌حساب می‌آید. طی پژوهش‌های صورت گرفته توسط محققان مختلف روابط متعددی جهت برآورد این ضریب ارائه شده است که اغلب این روابط به‌صورت تجربی و یا نیمه تجربی به‌دست آمده‌اند. با این وجود، نیاز به روش‌های دقیق‌تر تخمین ضریب پخشیدگی طولی همچنان احساس می‌شود. در این تحقیق جهت تخمین این ضریب، مدل‌های داده‌کاوی با توجه به اطلاعات هیدرولیکی و هندسی رودخانه‌ها توسعه یافته است. بر این اساس الگوریتم‌های درختی M5، CART و برنامه‌ریزی ژنتیک(GP) مورد استفاده قرار گرفت. جهت مقایسه کارایی مدل‌ها با معادلات موجود از پارامترهای آماری جذر میانگین مربعات خطا، میانگین خطای مطلق و نسبت اختلاف استفاده گردید. نتایج تحلیل‌های آماری نشان داد که مدل‌های داده‌کاوی می‌توانند ضریب پخشیدگی طولی را با دقت بهتر برآورد نمایند. مدل CART با وجود دقت زیاد در مرحله آموزش، در مرحله صحت سنجی از دقت کمتری برخوردار بوده است. مدل‌های M5 و GP به ترتیب دارای جذر میانگین مربعات خطای ۰/۴۱ و ۰/۴۴ و معیار دقت ۶۱٪ و ۶۲٪ بوده و در مقایسه با روابط تجربی موجود از دقت بیشتری برخوردار می‌باشند. با توجه به اختلاف ناچیز میان این دو مدل و سادگی مدل ارائه شده توسط M5، از این مدل می‌توان جهت برآورد ضریب پراکندگی طولی در رودخانه‌ها استفاده کرد.

واژه‌های کلیدی: پیشروی آلاینده‌ها ، ضریب پراکندگی طولی، CART، M5، GP

[۱] کارشناسی ارشد، گروه مهندسی آب دانشگاه کشاورزی و منابع طبیعی رامین خوزستان، ۰۹۱۶۹۵۴۸۷۶۸، Hosein.Nezaratian69@gmail.com
[۲] استادیار، گروه مهندسی آب دانشگاه کشاورزی و منابع طبیعی رامین خوزستان، ۰۹۱۶۶۵۳۱۸۹۶، Zahiri_Javad@yahoo.com (نویسنده مسئول)
[۳] استاد، گروه مهندسی آب دانشگاه شهید چمران اهواز، ۰۹۱۲۱۷۱۵۷۵۲، Kashefipour@excite.com

مقدمه

در چند دهه اخیر نحوه عملکرد آبراهه‌های طبیعی و شناسایی رفتار آن‌ها از جهات مختلفی مورد توجه قرار گرفته است که از آن جمله می‌توان به نحوه رفتار آبراهه‌ها در هنگام انتشار آلودگی اشاره کرد. این موضوع به‌منظور حفظ سلامت عمومی در مناطقی که صنایع بزرگ در کنار رودخانه‌ها قرار گرفته و رودخانه‌ها تامین کننده نیاز اصلی آبی می‌باشند، بیشتر مورد توجه قرار می‌گیرد. بدین‌جهت توانایی تخمین، شبیه سازی جریان و انتقال آلودگی و رسوب در سیستم مجاری روباز و رودخانه‌ها جهت برنامه‌ریزی منابع آب از اهمیت بالایی برخوردار است (Li and et al.,1998). با تزریق آلاینده‌های مختلف به بالادست یک آبراهه، انتقال آن به پایین‌دست توسط فرآیندهای اختلاط و انتشار[1] طولی، عرضی و عمقی رخ می‌دهد(Tayfur and Singh,2005). اولین بار Taylor(1954) پی برد که سرعت برشی و اختلاط در عرض جریان پس از گذشت زمانی خاص به تعادل رسیده اما این فرآیند در طول مسیر ادامه دارد. او با استفاده از روش فیک برای شار جرمی متلاطم معادله یک بعدی انتقال - پخش را برای کانال‌های یکنواخت ارائه کرد.

$$\frac{\partial C}{\partial t}+u\frac{\partial C}{\partial x}=K_x\frac{\partial^2 C}{\partial x^2} \qquad (۱)$$

در معادله ۱، C متوسط غلظت، u متوسط سرعت طولی، t زمان، x جهت طولی در راستای جریان و K_X ضریب پراکندگی طولی[2] است. این معادله تمام ویژگی‌های جریان را در کانال‌های آزمایشگاهی و رودخانه‌ها به‌طور کامل نشان می‌دهد (Ratherford,1994). ضریب پراکندگی معرفی شده در معادله ۱ نقش مهمی در مدل‌سازی فرآیند نشت و طراحی آبگیرها داشته و می‌توان آن را نماینده شدت اختلاط در رودخانه‌ها در نظر گرفت(Deng et al., 2002). توانایی جریان رودخانه و دیگر جریان‌های سطحی در پراکنش مواد تزریق‌شده در جهات طولی، عرضی و عمقی توسط ضرایب پراکندگی K_x، K_y، K_z بیان می‌گردد (Tayfur and Singh, 2005). با توجه به این نکته که اندکی پس از انتشار شدت اختلاط در جهت‌های عرضی و عمقی به تعادل می‌رسد، بعد طولی ضریب پراکندگی بیشتر مورد توجه محققان قرار گرفته است (Chatila,1997). دامنه تغییرات ضریب انتشار طولی در رودخانه های طبیعی با توجه به خصوصیات جریان و هندسه مقطع بسیار متغیر و پیچیده است. اگرچه روابط تجربی و تئوری بسیاری جهت تخمین مقدار ضریب پراکندگی طولی پیشنهاد شده است ولی با این وجود مطالعات بسیاری نیز بر روی این ضریب در حال انجام است (Noori et al.,2009). روابط ارائه شده جهت برآورد ضریب پراکندگی طولی اغلب به‌صورت تجربی یا نیمه تجربی بوده و در رودخانه‌هایی با شرایط متفاوت دارای دقت‌های متفاوتی هستند. تعدادی از روابط تجربی مورد استفاده جهت برآورد ضریب پخشیدگی طولی در جدول (۱) ارائه شده است. معادلات ارائه شده جهت برآورد ضریب پخشیدگی طولی از پارامترهای مختلف هیدرولیکی و هندسی آبراهه استفاده می‌کنند. در جدول (۱)، w و h به ترتیب عرض سطح مقطع جریان و عمق جریان برحسب متر، u سرعت جریان و u_* سرعت برشی جریان برحسب متر بر ثانیه و K_x ضریب پراکندگی طولی برحسب مترمربع بر ثانیه است. در سال‌های اخیر استفاده از الگوریتم‌های هوشمند در شاخه‌های مختلف مهندسی ازجمله مهندسی رودخانه رو به افزایش بوده است. الگوریتم‌های فراکاوشی توانایی تخمین پدیده‌ها و فرآیندهای پیچیده طبیعی را دارا بوده و اغلب دارای دقت بیشتری نسبت به روابط تجربی می‌باشند. Etemad-shahidi and Ghaemi (2011) جهت بررسی کارایی الگوریتم M5 در محاسبه عمق آبشستگی اطراف مجموعه پایه‌های پل از اطلاعات آزمایشگاهی و میدانی استفاده کردند. Etemad-shahidi and Taghipour (2012) با استفاده از الگوریتم M5 و با به‌کارگیری اطلاعات ۱۴۹ رودخانه موجود در سرتاسر جهان به تخمین ضریب پراکندگی طولی پرداختند. این محققین جهت تخمین این ضریب از پارامتر سینوسی بودن رودخانه‌ها نیز استفاده کردند. نتیجه این تحقیق بیان‌گر دقت و قدرت بالای این الگوریتم در تخمین ضریب انتشار طولی است. Sattar and Gharehbaghi (2015) با استفاده از الگوریتم ژنتیک و اطلاعات مربوط به ۱۵۰ رودخانه موجود

[1] Dispersion
[2] Longitudinal dispersion coefficient

در کشورهای آمریکا، کانادا، اروپا و نیوزلند به تخمین ضریب پراکندگی طولی پرداختند که نتیجه آن گویای برتری و دقت بالای آن نسبت به مدل‌های تجربی بوده است. ظهیری (۱۳۹۴) جهت تخمین عمق آبشستگی اطراف پایه‌های پل از دو الگوریتم M5 و CART استفاده کرد که نتیجه آن بیانگر کارایی بهتر و ساده‌تر الگوریتم M5 بوده است. Haghiabi (2016) به پیش‌بینی ضریب پراکندگی طولی با استفاده از رگرسیون چند متغیره اسپلاین[1](MARS) پرداخت که نتایج به‌دست آمده نشان‌دهنده توانایی مدل اسپلاین در پیش‌بینی این ضریب می‌باشد. هدف از تحقیق پیش رو مقایسه سه الگوریتم درختی M5، GP وCART در پیش‌بینی ضریب پراکندگی طولی و مقایسه نتایج مدل‌های ارائه شده با روابط تجربی است.

جدول(۱): روابط تجربی ارائه‌شده توسط محققین

محقق	رابطه ارائه‌شده
Elder (1959)	$K_x= 5.93hu_*$
Fischer (1967)	$K_x=0.011(u^2w^2/hu_*)$
McQuivey and Keefer (1974)	$K_x=0.58(h/u_*)^2uw$
Liu (1977)	$K_x=0.18(u/u_*)^{0.5}(w/h)^2hu_*$
Iwasa and Aya (1991)	$K_x=2(w/h)^{1.5}hu_*$
Li et al. (1998)	$K_x=0.55(wu_*/h^2)$
Seo and Cheong (1998)	$K_x=5.92(u/u_*)^{1.43}(w/h)^{0.62}hu_*$
Koussis and Rodriguez-Mirasol (1998)	$K_x=0.6(w/h)^2hu_*$
Li et al. (1998)	$K_x=5.92(u/u_*)^{1.2}(w/h)^{1.3}hu_*$
Kashefipour and Falconer (2002)	$K_x=10.612(u/u_*)hu$
Rajeev and Dutta (2009)	$K_x=2(u/u_*)^{0.96}(w/h)^{1.25}hu_*$

مواد و روش‌ها

در تحقیق حاضر جهت مدل‌سازی ضریب پراکندگی طولی از داده‌های به‌کار رفته در تحقیق Etemad-shahidi and Taghipour (2012) استفاده شده است. تعداد کل این داده‌ها ۱۴۹سری است که این اطلاعات از رودخانه‌ها و آبراهه‌های سراسر جهان برداشت شده‌اند. این مجموعه شامل اطلاعات هندسی و هیدرولیکی رودخانه‌ها ازجمله عرض سطح آب برحسب متر، عمق جریان برحسب متر، سرعت جریان و سرعت برشی برحسب متر بر ثانیه می‌باشند. مشخصات آماری داده‌های مورد استفاده برای آموزش و صحت سنجی مدل‌ها در جدول (۲) ارائه شده است.

ضریب پراکندگی طولی از پارامترهای متعدد هندسی و هیدرولیکی آبراهه تأثیر می‌پذیرد که موارد تأثیرگذار بر روی این ضریب را می‌توان به‌صورت تابع زیر نوشت:

$$K_x = f_1(\rho,\mu,w,h,u,u_*,s_f,s_n) \qquad (۲)$$

در این رابطه ρ چگالی مایع، μ لزوجت دینامیکی، w عرض سطح مقطع، h عمق جریان، u سرعت جریان، u_* سرعت برشی جریان، S_f شیب طولی بستر و S_n ضریب سینوسی آبراهه است. اغلب پارامترها دارای بعد بوده که جهت جلوگیری از به وجود آمدن ناهمگنی ابعادی در دو طرف معادله از تئوری باکینگهام استفاده گردید (Seo and Cheong,1998):

$$\frac{K_X}{hu_*} = f_2(\rho\frac{uh}{\mu},\frac{w}{h},\frac{u}{u_*},s_f,s_n) \qquad (۳)$$

جریان در رودخانه‌ها و آبراهه‌ها غالباً به‌صورت متلاطم بوده و می‌توان از عدد رینولدز (ρ uh/μ) صرف‌نظر نمود. علاوه بر این اندازه گیری ضریب سینوسی دشوار و پیچیده بوده و در اغلب رودخانه‌ها این پارامتر اندازه‌گیری نمی‌شود. بر این اساس می‌توان عوامل مؤثر بی‌بعد بر ضریب پراکندگی طولی را به شکل زیر نوشت (Seo and Cheong,1998;Seo and Beak, 2004):

$$\frac{K_X}{hu_*} = f_2(\frac{w}{h},\frac{u}{u_*}) \qquad (۴)$$

[1] Multivariate Adaptive Regression Splines

جدول(۲): اطلاعات مورد استفاده جهت آموزش و صحت سنجی مدل‌ها

متغیر آماری	عرض(متر)	عمق(متر)	سرعت (متر بر ثانیه)	سرعت برشی (متر بر ثانیه)	ضریب انتشار طولی (مترمربع بر ثانیه)
حداکثر	۲۵۳/۶	۸/۲	۱/۷۳	۰/۵۵	۱۴۸۶/۵
حداقل	۱/۴	۰/۱۴	۰/۰۲۹	۰/۰۰۱۶	۰/۲
میانگین	۴۹/۵۸	۱/۳۵	۰/۴۷	۰/۰۸۴	۸۳/۲۹
انحراف معیار	۴۸/۲۸	۱/۳۲	۰/۳۱	۰/۰۷۲	۱۸۰/۹۵

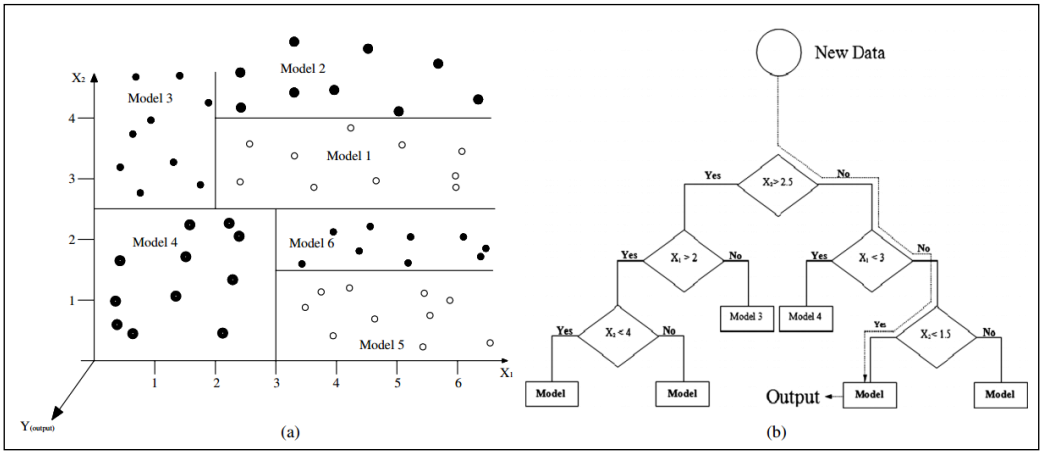
در این تحقیق از سه مدل داده‌کاوی نوین استفاده شده است. الگوریتم M5 اولین بار توسط Quinlan (1992)ابداع شد و پس از آن توسط Wang and Witten (1997)توسعه و بهبود یافت. این مدل شاخه‌های خود را به‌صورت دوتایی و تنها بر اساس یک متغیر ایجاد می‌کند، بدین گونه که بر اساس شرطی که در هر گره تعریف می‌شود، اطلاعات در آن گره به دو قسمت تقسیم می‌شود (جباری و صمدی،۱۳۹۲). در روش M5 فضای مسئله به زیر دامنه‌هایی تقسیم شده و برای هر زیر دامنه یک مدل رگرسیون خطی چند متغیره برازش داده می‌شود. این الگوریتم جداسازی‌های ممکن را در فضای چند متغیره انجام داده و به‌طور خودکار مدل‌هایی را برای هرکدام از دامنه‌ها ایجاد می‌کند (Quinlan,1992). در این الگوریتم از پارامتر انحراف معیار مقادیر متغیر هدف به‌عنوان معیار اندازه‌گیری خطا در هر گره و ایجاد شاخه در آن گره استفاده می‌شود. بدین صورت که صفتی که موجب کاهش بیشتر انحراف معیار در هر گره می‌شود به‌عنوان صفت مورد نظر جهت ایجاد شاخه انتخاب می‌شود. کاهش انحراف استاندارد [1](SDR) که به‌عنوان تابع خطا در الگوریتم M5به کار می‌رود به‌صورت زیر است:

$$SDR = sd(T) - \sum \left| \frac{Ti}{T} \right| \times sd(T) \qquad (۵)$$

در رابطه بالا T شامل نمونه‌هایی است که به گره مورد نظر رسیده‌اند و Ti شامل نمونه‌های است که از تقسیم گره مورد نظر بر اساس صفت انتخاب شده به‌دست آمده‌اند. Sd نیز معرف انحراف معیار است.

الگوریتم M5 تمامی حالت‌های مختلف جهت ایجاد شاخه بر اساس صفت خاص را بررسی کرده و درنهایت حالتی را انتخاب می‌کند که بتواند بیشتر از حالت‌های دیگر تابع خطای فوق را افزایش دهد. پس از تکمیل الگوریتم درختی برای نمونه‌های موجود در هر گره داخلی، یک مدل رگرسیون خطی چند متغیره برازش داده می‌شود. شکل ۱ نحوه تقسیم فضای مسئله به اجزای کوچک‌تر و کاربرد مدل‌های خطی را برای هرکدام از زیر دامنه‌ها با توجه به ساختار درختی نشان می‌دهد (Etemad-Shahidi and Taghipour,2012).

[1] Standard Deviation Reduction



شکل(۱): تقسیم فضای مسئله و ارائه مدل‌های خطی به ازای هرکدام از فضاها توسط الگوریتم M5

برنامه‌ریزی ژنتیک[1] (GP) از زیرشاخه‌های الگوریتم ژنتیک و از روش‌های الگوریتم گردشی است که نخستین بار توسط Koza (1992) معرفی شد. برنامه‌ریزی ژنتیک از ساختار درختی برای بهینه‌سازی مسئله استفاده می‌کند، درحالی‌که الگوریتم ژنتیک بر روی رشته های بیتی کار می کند. در این روش هیچ ساختار و ارتباطی میان متغیرهای ورودی و خروجی وجود ندارد و ساختار مدل و ضرایب بهینه نیز طی فرآیند بهینه‌سازی به دست می‌آیند. ساختار درختی از مجموعه توابع (عملگرهای ریاضی مورد استفاده در فرمول‌ها) و ترمینال‌ها (متغیرهای مسئله و اعداد ثابت) ایجاد می‌شود (Koza, 1992). سه عمل ژنتیکی تلاقی، جهش و تولیدمثل ازجمله مهم‌ترین عمل‌های ژنتیکی در برنامه‌ریزی ژنتیک می‌باشند. عمل‌های دیگر از قبیل اصلاح ساختار با احتمال کمتری بکار گرفته می‌شوند.

الگوریتم CART که اولین بار توسط Breiman et al.(1984) معرفی شد، به‌صورت یک درخت مرتبه‌ای دودویی است که فضای مسئله را به قسمت‌های جزء تقسیم می‌کند. در این روش داده‌ها به‌صورت خام و دست‌نخورده برای الگوریتم معرفی شده و هیچ‌گونه فیلتری بروی داده‌ها صورت نمی‌گیرد. در این مدل درختان بدون وجود عامل متوقف کننده به حداکثر رشد خود رسیده سپس اصلاح می‌شوند. هدف مدل CART ایجاد یک درخت اصلاح شده نیست بلکه تولید یک سری درختان اصلاح شده تودرتو است که در آن همه درختان بهینه هستند. در درخت CART مجموعه‌ای از سؤالات به شکل x<=d? است که در آن x یک متغیر مستقل و d یک مقدار ثابت بوده و جواب هر سؤال آری یا نه است. در این روش در گره انتهایی مقدار عددی به‌عنوان نتیجه مورد نظر جهت آن گره ارائه می‌شود (Breiman et al.,1984). معیارهای متفاوتی جهت ایجاد شاخه و تولید درخت‌ها در روش CART وجود دارد که در تحقیق پیش رو از معیار انحراف حداقل مربعات[2] (LSD) استفاده شده است. این معیار به‌صورت زیر تعریف می‌شود:

$$SS(t)=\sum_{i=1}^{N_t}\left(y_i(t)-\bar{y}(t)\right)^2 \qquad (۶)$$

در رابطه ۶، N_t تعداد رکوردها، $y_i(t)$ مقدار متغیر هدف و $\bar{y}(t)$ میانگین مقادیر متغیر هدف در گره مورد نظر است. در الگوریتم CART یک متغیر ورودی زمانی به‌عنوان بهترین صفت برای ایجاد شاخه در گره t مورد استفاده قرار می‌گیرد که تابع زیر بیشینه شود.

$$Q(X,t)=SS(t)-[SS(t_R)+SS(t_L)] \qquad (۷$$

(

[1] Genetic Programming

[2] Least-Squared Deviation

در معادله ۷، $SS(t)$ مجموع مربعات خطا و $SS(t_R)$ و $SS(t_L)$ مجموع مربعات خطا به ترتیب در شاخه سمت راست و چپ گره t است.

مدل M5 تنها قادر به شبیه‌سازی مدل‌های خطی بوده که این امر با اساس معادلات متداول ضریب انتشار طولی در تناقض است. بر همین اساس کلیه داده‌ها به‌صورت لگاریتم طبیعی به مدل M5 معرفی شدند و پس از ساخت مدل خطی از حالت لگاریتمی به توانی تبدیل شدند. برای مدل‌های GP و CART هم در ابتدا از لگاریتم طبیعی اعداد استفاده شد و پس از ساخت مدل، خروجی‌ها از حالت لگاریتمی به حالت توانی تبدیل شدند تا شرایط کاملاً یکسانی برای هر سه مدل داده‌کاوی فراهم شده باشد. در تحقیق حاضر از مجموع اطلاعات در دسترس ۸۰٪ (۱۲۰ سری از داده‌ها) جهت آموزش مدل‌ها و ۲۰٪ باقیمانده (۲۹ سری از داده‌ها) جهت صحت‌سنجی مورد استفاده قرار گرفتند. برای مدل‌سازی M5 ، GP وCART به ترتیب از برنامه‌های WEKA 3.7، GeneXproTools 5.0 و STATISTICA استفاده شد. جهت مقایسه بهتر کارایی مدل‌های ارائه شده از آنالیزهای آماری خطای متوسط مطلق[1] (MAE) ،جذر میانگین مربعات خطا[2] (RMSE) و ضریب تبیین[3] (R^2) استفاده گردید. علاوه بر معیارهای ذکرشده در این تحقیق از نسبت اختلاف (DR) نیز استفاده شده است. این نسبت توسط White et al.(1973) پیشنهاد شده است و از معیارهای قدرتمند آماری به‌حساب می‌آید.

$$MAE = \frac{1}{N}\sum|D_{Ri}| \quad (۸)$$

$$RMSE = \sqrt{\frac{1}{N}\sum\left(D_{R_i}\right)^2} \quad (۹)$$

$$R^2 = 1 - \frac{\sum(K_m - K_C)^2}{\sum\left(K_m - \overline{K_m}\right)} \quad (۱۰)$$

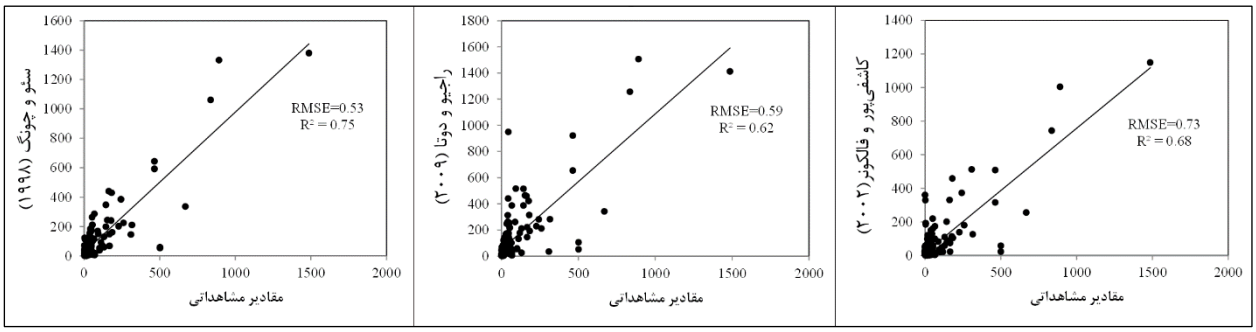
$$D_R = log\,\frac{K_c}{K_m} \quad (۱۱)$$

در روابط فوق K_m ضریب انتشار طولی مشاهداتی و K_c ضریب انتشار طولی محاسباتی است.

نتایج و بحث

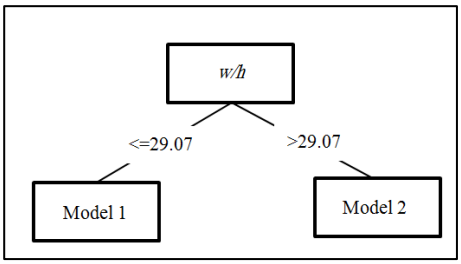
پس از بررسی‌های آماری صورت گرفته بر روی روابط تجربی، سه رابطه Seo and Cheong (1998)، Kashefipour and Falconer (2002) و Rajeev and Dutta (2009) به‌عنوان بهترین روابط در پیش‌بینی ضریب پراکندگی طولی در این تحقیق انتخاب و برای مقایسه با سه مدل داده‌کاوی به کار گرفته شدند. شکل ۲ عملکرد سه رابطه تجربی نام‌برده را برای تخمین K_x نشان می‌دهد.

[1] Mean-Absolute Error
[2] Root Mean Square Error
[3] Determination coefficient



شکل(۲): عملکرد روابط تجربی انتخاب شده در تخمین ضریب پخشیدگی طولی

پس از معرفی داده‌های مشاهده‌ای به‌صورت لگاریتم طبیعی به M5، مدل مورد آموزش قرار گرفت. دو معادله ارائه شده توسط این مدل، روابط خطی بوده که توسط الگوریتم درختی برای هرکدام از حالت‌ها توسعه یافته است. نمودار درختی مدل پیشنهادی M5 در شکل ۳ ارائه شده است.



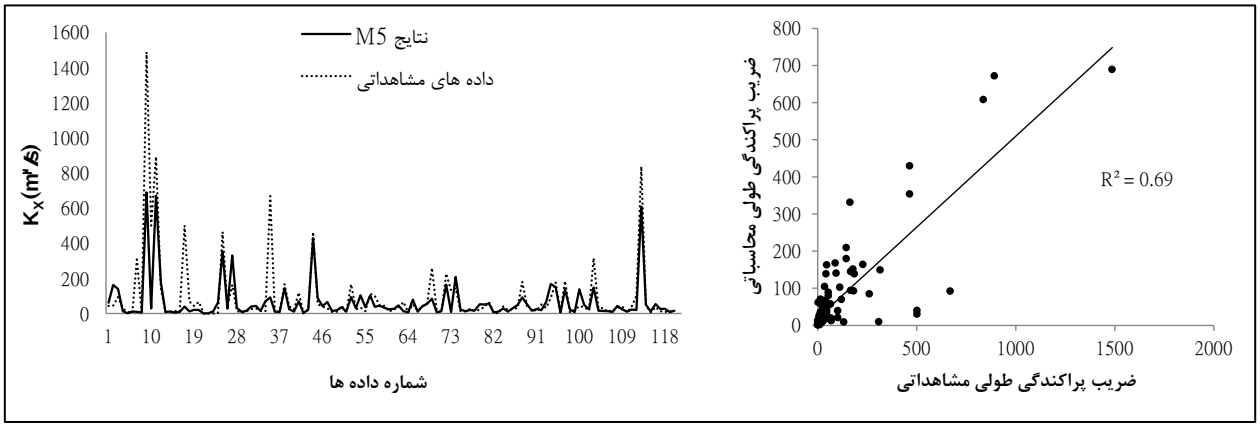
شکل(۳): ساختار درختی M5 جهت برآورد ضریب پراکندگی طولی در رودخانه‌ها

روابط ارائه شده توسط مدل M5 پس از تبدیل از حالت لگاریتمی، به‌صورت معادلات ۱۲ و ۱۳ می‌باشند.

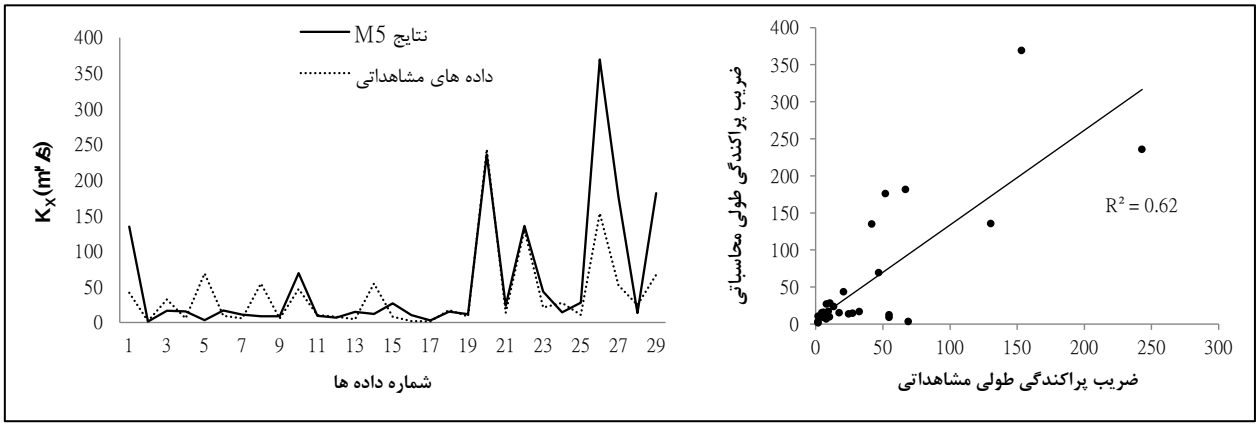
$$Model\ 1: IF\left(\frac{w}{h}\right)\le 29.07 \Rightarrow \left(\frac{K_x}{hu_*}\right)=12.06\left(\frac{w}{h}\right)^{0.8}\left(\frac{u}{u_*}\right)^{0.15} \quad (۱۲)$$

$$Model\ 2: IF\left(\frac{w}{h}\right)> 29.07 \Rightarrow \left(\frac{K_x}{hu_*}\right)=13.73\left(\frac{w}{h}\right)^{0.62}\left(\frac{u}{u_*}\right)^{0.91} \quad (۱۳)$$

با توجه به شکل ۳، پارامتر مورد استفاده جهت تشکیل مدل درختی، نسبت عرض آبراهه به عمق جریان (w/h) می‌باشد. از لحاظ فیزیکی باریک یا عریض بودن رودخانه تاثیر زیادی بر روی میزان پخشیدگی طولی داشته و $(w/h)\le 29.07$ می‌تواند نماینده رودخانه‌های باریک و $(w/h)> 29.07$ نشان‌دهنده رودخانه‌های عریض باشد. با توجه به معادلات ۱۲ و ۱۳، در آبراهه‌هایی که نسبت عرض به عمق آن‌ها کمتر از ۲۹/۰۷ باشد، ترم (w/h) تأثیر بیشتری نسبت به ترم (u/u*) در پیش‌بینی ضریب پراکندگی طولی خواهد داشت. در صورتی‌که در آبراهه‌های عریض که نسبت عرض به عمق آن‌ها بیش از ۲۹/۰۷ است، شاهد افزایش تأثیر ترم (u/u*) نسبت به (w/h)، در تخمین ضریب پراکندگی خواهیم بود. در رودخانه‌های باریک با (w/h) کوچک تأثیر تنش برشی ناچیز است ولی با افزایش (w/h) تأثیر تنش برشی بر روی ضریب پخشیدگی افزایش می‌یابد(Papadimitrakis and Orphanos, 2004). علاوه بر این Rutherford(1994) نیز در مطالعه‌ای به این نتیجه رسیده است که در رودخانه‌های عریض نسبت به رودخانه‌های باریک، پارامتر سرعت تأثیر بیشتری بر روی ضریب پخشیدگی دارد. شکل‌های ۴ و ۵ نمایش‌دهنده نحوه عملکرد مدل M5 در مرحله آموزش و صحت سنجی برای تخمین ضریب پراکندگی طولی می‌باشند.



شکل(۴): عملکرد مدل M5 در مرحله آموزش برای تخمین ضریب پراکندگی طولی

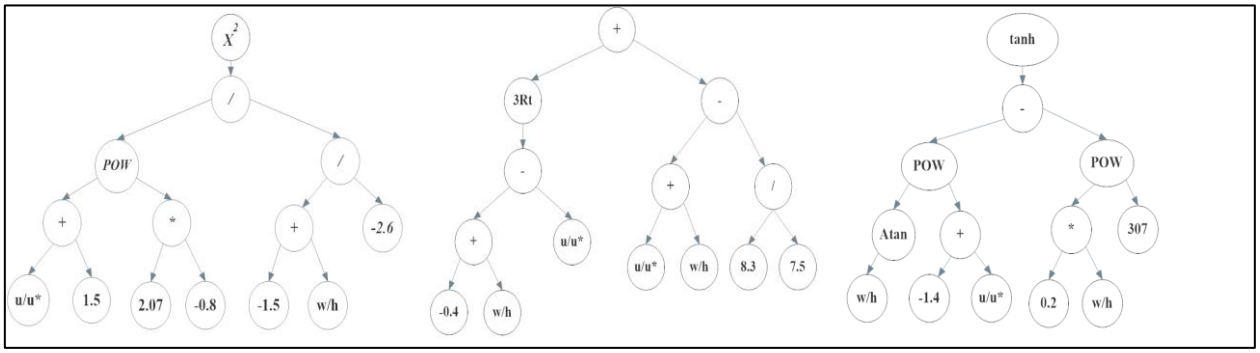


شکل(۵): عملکرد مدل M5 در مرحله صحت سنجی برای تخمین ضریب پراکندگی طولی

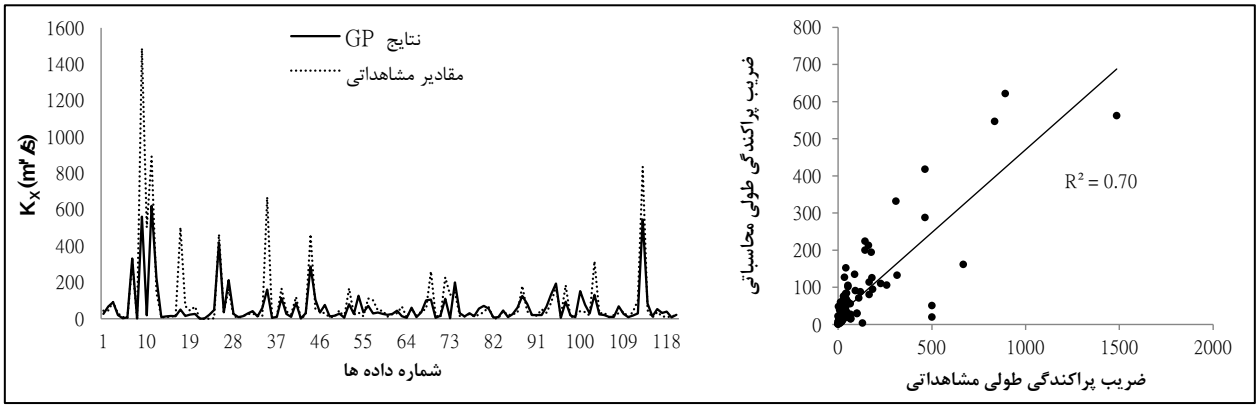
جهت مقایسه دقیق‌تر کارایی الگوریتم‌های درختی، داده‌های ورودی در مراحل آموزش و صحت سنجی برای تمامی مدل‌ها یکسان معرفی شدند. علاوه بر آن داده‌های ورودی در ابتدا برای تمامی مدل‌ها به‌صورت لگاریتم طبیعی بوده و پس از اجرای مدل‌ها، نتایج به مقادیر واقعی تبدیل شدند.

در مدل GP ابتدا از چهار عملگر اصلی جمع، تفریق، ضرب و تقسیم استفاده شد، و با افزودن عملگرهای sqrt، power، tanh و $\tan^{-1}$ نتایج بهتری حاصل گردید. معادله ۱۳ ضمن پیچیدگی فراوان، به‌عنوان بهترین و دقیق‌ترین مدل ارائه شده توسط مدل GP انتخاب شد. شکل ۶ نشان‌دهنده ساختار درختی مدل GP جهت برآورد ضریب پراکندگی طولی بوده که حاصل جمع این سه درخت، مدل نهایی را ارائه می‌دهد. شکل‌های ۷ و ۸ نمایش‌دهنده نحوه عملکرد مدل GP در مرحله آموزش و صحت سنجی برای تخمین ضریب پراکندگی طولی است.

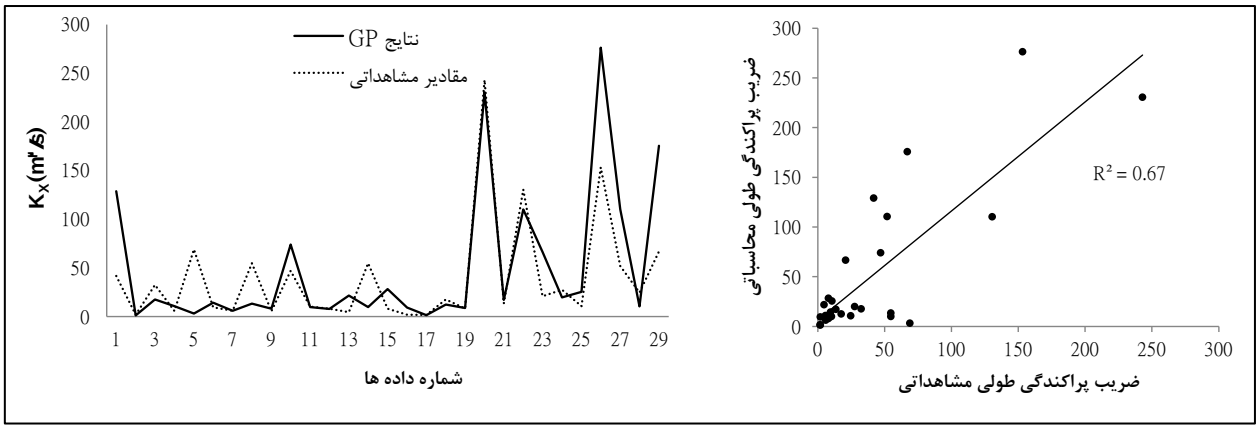
$$\left(\frac{K_x}{hu_*}\right)=\exp\left(\left(\frac{\left(\frac{u}{u_*}+1.5\right)^{-1.6}}{\frac{\left(-1.5+\frac{w}{h}\right)}{-2.6}}\right)^2+\left(\sqrt[3]{\left(-0.4+\frac{w}{h}\right)-\frac{u}{u_*}}+\frac{w}{h}+\frac{u}{u_*}+20.75\right)+\left(\tanh\left(\tan^{-1}\left(\frac{w}{h}\right)\right)^{-1.4+\frac{u}{u_*}}-\left(\left(0.2*\frac{w}{h}\right)^{307}\right)\right)\right) \quad (۱۳$$



شکل(۶): ساختار درختی GP جهت برآورد ضریب پراکندگی طولی



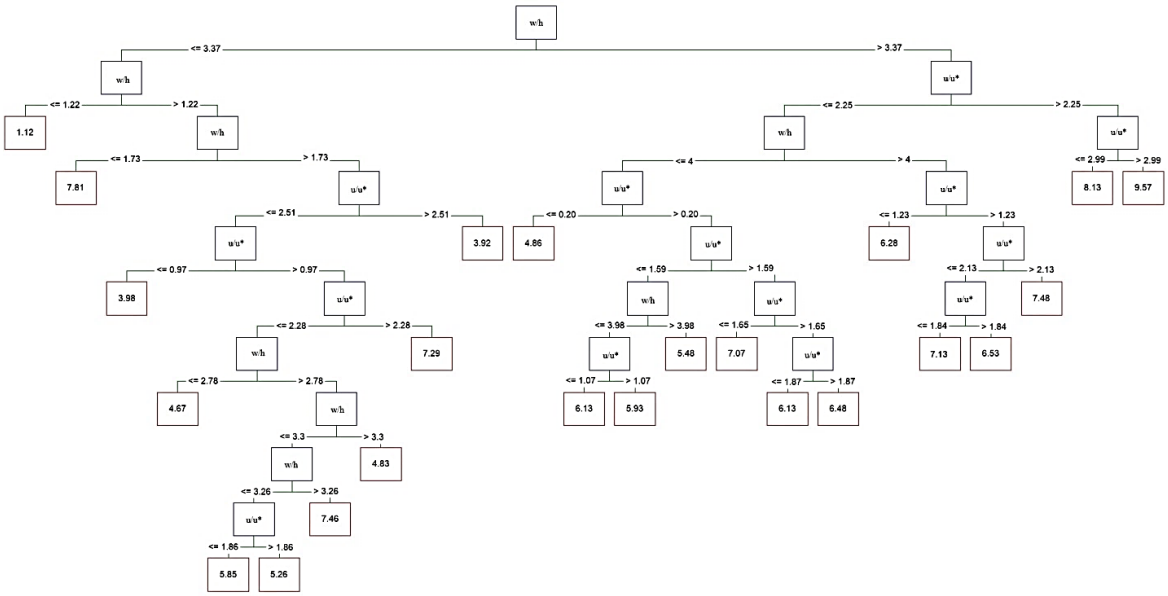
شکل(۷): عملکرد مدل GP در مرحله آموزش برای تخمین ضریب پراکندگی طولی



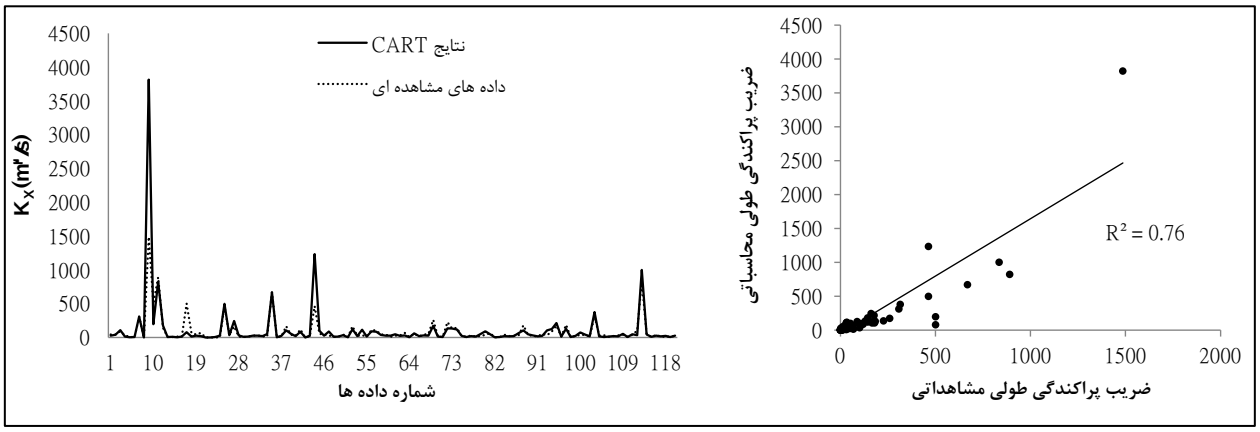
شکل(۸): عملکرد مدل GP در مرحله صحت سنجی برای تخمین ضریب پراکندگی طولی

خروجی مدل درختی CART بر اساس داده‌های ورودی در شکل ۹ نشان داده شده است. همچنان که در این شکل مشاهده می‌شود، در اولین گره، ترم (w/h) به عنوان بهترین صفت جهت ایجاد شاخه انتخاب گردید که این امر نشان دهنده تأثیر شکل هندسی آبراهه بر ضریب پراکندگی طولی است. همان‌گونه که در شکل ۹ مشاهده می‌شود، الگوریتم CART از گره های متعددی تشکیل شده و در انتهای هر گره یک کمیت عددی به عنوان خروجی ارائه می دهد (Avg). این کمیت بیانگر میانگین ضریب پراکندگی طولی مربوط به اطلاعات آن گره است. تعداد زیاد گره ها در این الگوریتم باعث مشکل شدن کاربرد آن در سایر مدل‌ها می‌گردد. همچنین اشکال

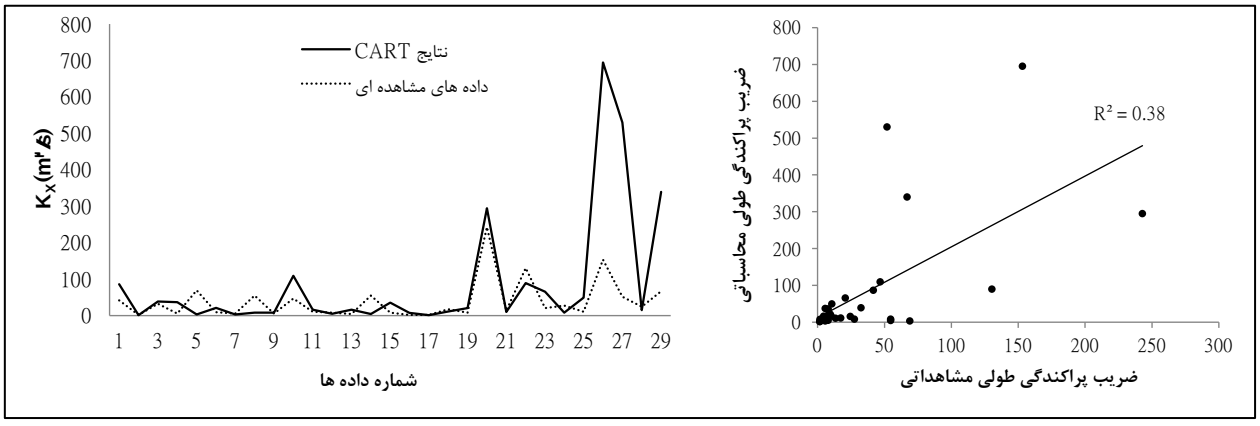
۱۰ و ۱۱ بیانگر نحوه عملکرد مدل CART در مراحل آموزش و صحت سنجی است.



شکل(۹): ساختار درختی CART جهت برآورد ضریب پراکندگی طولی



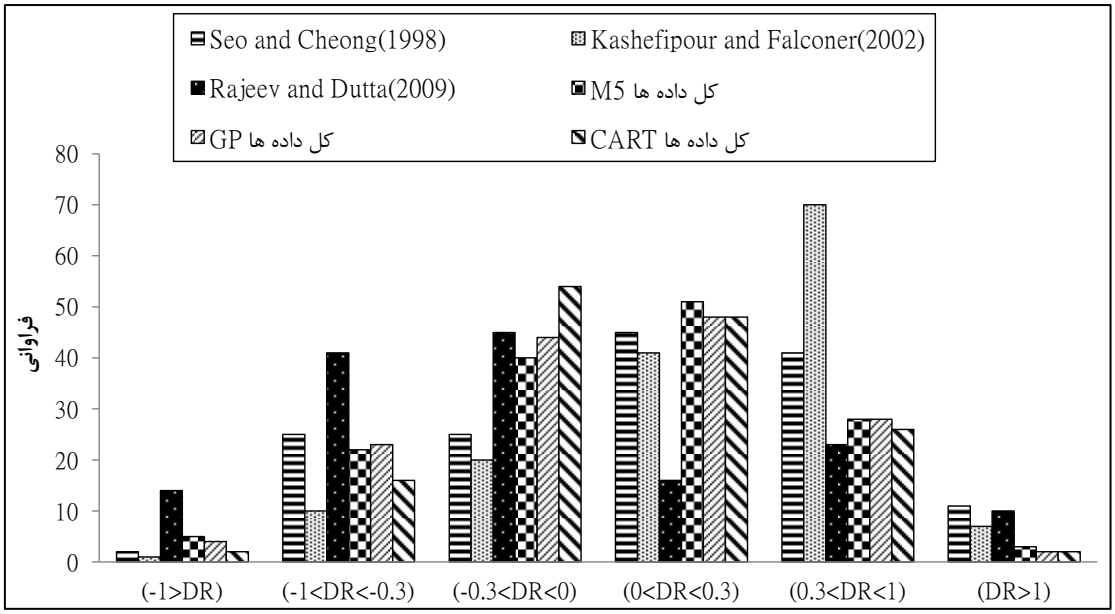
شکل(۱۰): عملکرد مدل CART در مرحله آموزش برای تخمین ضریب پراکندگی طولی



شکل(۱۱): عملکرد مدل CART در مرحله صحت سنجی برای تخمین ضریب پراکندگی طولی

جهت بررسی دقت مدل‌ها از شاخص‌های آماری متنوعی استفاده گردید که یکی از مهم‌ترین شاخص‌ها، DR است. چنانچه DR برابر صفر باشد نشان‌دهنده تطابق کامل نتایج مدل‌ها و مقادیر واقعی است. در غیر این صورت چنانچه (DR>۰) باشد مدل ضریب پراکندگی طولی را بیشتر از مقدار واقعی[1] برآورد کرده است در غیر این صورت (DR<۰)، ضریب پراکندگی طولی کمتر از مقدار واقعی محاسبه شده است. در این تحقیق درصد مقادیر با DR بین ۰/۳- و ۰/۳ به‌عنوان دقت[2] هر مدل در نظر گرفته شده است که مطابق دقت تعریف شده توسط Seo and Cheong (1998) است. هیستوگرام فراوانی DR برای سه روش تجربی و مدل‌های درختی در شکل ۱۲ نشان داده شده است. علاوه بر این نتایج تحلیل‌های آماری صورت گرفته در جدول ۳ ارائه شده است.

[1] Overestimation
[2] Accuracy



شکل(۱۲): مقایسه نسبت اختلاف(DR) مدل‌ها

جدول(۳): بررسی کارایی مدل‌های هوشمند و تجربی در تخمین ضریب پخشیدگی طولی

مدل		DR >1 %	0.3<DR<1 %	0<DR<0.3 %	-0.3<DR<0 %	-1<DR<-0.3 %	DR<-1 %	Accuracy %	MAE	RMSE	R^2
Seo and Cheong(1998)		۷/۳	۲۷/۵	۳۰/۲	۱۶/۷	۱۶/۷	۱/۳	۴۶/۹	۰/۴۳	۰/۵۳	۰/۷۵
Kashefipour and Falconer(2002)		۶/۷	۱۵/۴	۱۰/۷	۳۰/۲	۲۷/۵	۹/۳	۴۰/۹	۰/۵۳	۰/۷۳	۰/۶۸
Rajeev and Dutta(2009)		۴/۶	۴۶/۹	۲۷/۵	۱۳/۴	۶/۷	۰/۶	۴۰/۹	۰/۴۷	۰/۵۹	۰/۶۲
M5	کل داده‌ها	۲	۱۸/۷	۳۴/۲	۲۶/۸	۱۴/۷	۳/۳	۶۱	۰/۳۲	۰/۴۴	۰/۶۵
	صحت سنجی	۰	۳۴/۴	۲۷/۵	۲۷/۵	۶/۸	۳/۴	۵۵	۰/۳۵	۰/۴۴	۰/۶۲
GP	کل داده‌ها	۱/۳	۱۸/۷	۳۲/۲	۲۹/۵	۱۵/۴	۲/۶	۶۱/۷	۰/۲۹	۰/۴۱	۰/۶۸
	صحت سنجی	۰	۲۷/۵	۲۷/۵	۳۱	۱۰/۳	۳/۴	۵۸/۵	۰/۳۱	۰/۴۳	۰/۶۷
CART	کل داده‌ها	۱/۳	۱۷/۴	۳۲/۲	۳۶/۲	۱۰/۷	۱/۳	۶۸/۴	۰/۲۵	۰/۳۶	۰/۷۴
	صحت سنجی	۳/۴	۴۱/۳	۱۳/۷	۲۷/۵	۶/۸	۶/۸	۴۱,۲	۰/۴۶	۰/۵۶	۰/۳۸

با توجه به شکل ۱۲ و جدول ۳، دقت مدل‌های درختی به کار گرفته شده همگی بیشتر از ۶۰ درصد بوده است. این درحالی است که رابطه تجربی Seo and Cheong (1998) با توجه به دارا بودن شاخص‌های آماری مطلوب‌تر نسبت به سایر روابط تجربی، تنها دارای دقتی معادل ۰/۴۶ است که این امر بیانگر دقت بیشتر مدل‌های درختی نسبت به روابط تجربی در تخمین ضریب پراکندگی طولی است. در میان مدل‌های درختی، مدل CART برای کل داده‌ها با دارا بودن ضریب تبیین ۰/۷۴ و جذر میانگین مربعات خطا ۰/۳۶، از دقت بیشتری نسبت به دیگر مدل‌ها برخوردار است. این در حالی است که دقت این مدل در مرحله صحت‌سنجی بسیار پایین بوده که نشان‌دهنده ضعف این مدل در تعمیم نتایج است. نزدیک بودن نتایج تحلیل آماری دو مدل پیشنهادی GP و M5 در مراحل آموزش و صحت‌سنجی بیانگر آموزش صحیح این دو الگوریتم است. دقت دو مدل GP و M5 بسیار نزدیک به‌هم بوده، اما پیچیدگی فراوان مدل GP و سادگی و قابل درک بودن مدل M5 نشان‌دهنده برتری این مدل در تخمین ضریب پخشیدگی طولی است. در مدل M5، جذر میانگین مربعات خطا که در تحقیقات مربوط به ضریب پراکندگی طولی کاربرد فراوانی دارد، نسبت به رابطه Seo and Cheong (1998) از ۰/۵۳ به ۰/۴۴ کاهش یافته که این امر نشان‌دهنده کارایی این الگوریتم در تخمین ضریب پراکندگی طولی است.

نتیجه‌گیری

هدف از این تحقیق برآورد ضریب پخشیدگی طولی در آبراه‌ها با استفاده از مدل‌های داده‌کاوی M5، GP و CART می‌باشد. طبق نتایج به‌دست آمده، روابط تجربی از دقت پایین‌تری در مقایسه با مدل‌های داده‌کاوی برخوردارند. در میان مدل‌های داده‌کاوی مورد استفاده، مدل CART در مرحله آموزش و صحت سنجی به ترتیب دارای ضریب تبیین ۰/۷۶و ۰/۳۸ بوده که خود بیانگر قدرت بالای این مدل در مرحله آموزش و ضعف آن در مرحله صحت سنجی است. دقت مدل CART برای کل داده‌ها و داده‌های صحت سنجی به ترتیب ۶۸/۴ و ۴۱/۲ است. بررسی نتایج سایر مدل‌ها نشان می‌دهد که مدل پیشنهادی CART در مرحله صحت سنجی ضعیف ظاهر شده است. در مقابل نتایج آماری دو مدل GP و M5 نزدیک به هم بوده اما رابطه پیشنهادی توسط مدل GP به‌مراتب پیچیده‌تر از دو رابطه پیشنهادی M5 است. بر اساس آنالیزهای صورت گرفته مدل پیشنهادی M5 دارای دقت ۶۱ و ۵۵ درصد به ترتیب برای کل داده‌ها و داده‌های صحت‌سنجی است که این میزان دقت به طرز چشم‌گیری بیشتر از دقت ارائه شده توسط روابط تجربی است. بر این اساس می‌توان از مدل M5 با توجه به‌دقت بالا و سادگی معادلات ارائه شده و قابل درک بودن آن جهت برآورد ضریب پراکندگی طولی استفاده کرد.

منابع

ظهیری، ج. ۱۳۹۴. کاربرد مدل‌های ناپارامتریک CART و M5 در محاسبه عمق آبشستگی اطراف پایه‌های پل. فصلنامه علمی پژوهشی مهندسی آبیاری و آب ، شماره ۲۰، ص ۵۰-۳۵.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. 1984. Classification and Regression Trees. Chapman & Hall/CRC Press, Boca Raton, FL.

Chatila, G. J. 1997. Modeling of pollutant transfer in compound open channels. PhD Dissertation, University of Ottawa, Ontario, Canada.

Deng, Z.Q., Bengtsson, L., Singh, V. P., et al. 2002. Longitudinal dispersion coefficient in single-channel streams. Journal of Hydraulic Engineering. 128 (10): 901-916.

Etemad-Shahidi, A and M. Taghipour. 2012. Predicting longitudinal dispersion coefficient in natural streams using M5′ model tree. Journal of Hydraulic Engineering, 138(6): 542-554.

Etemad-Shahidi, A and N. Ghaemi. 2011. Model tree approach for prediction of pile groups scour due to waves. Ocean Engineering, 38: 1522–1527.

Haghiabi, A. H. 2016. Prediction of longitudinal dispersion coefficient using multivariate adaptive regression splines. Journal of Earth System Science, 125: 985-995.

Kashefipour, M.S., Falconer, R.A. 2002. Longitudinal dispersion coefficients in natural channels. Water Res. 36 (6): 1596–1608.

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

Li, Z.H., Huang, J., Li, J. 1998. Preliminary study on longitudinal dispersion coefficient for the gorges reservoir. In: Proceedings of the Seventh International Symposium Environmental Hydraulics, 16e18 December, Hong Kong, China.

Noori, R., Karbassi, A., Farokhnia, A and Dehghani, M. 2009. Predicting the longitudinal dispersion coefficient using support vector machine and adaptive Neuro-Fuzzy inference system techniques; Environmental Engineering Science. 26(10):1503–1510.

Papadimitrakis, I., and Orphanos, I. 2004. Longitudinal dispersion characteristics of rivers and natural streams in Greece. Water, Air, & Soil Pollution, 4(4–5): 289–305.

Quinlan, J. R. 1992. Learning with continuous classes. Proc., 5th Australian Joint Conf. on Artificial Intelligence, World Scientific, Singapore, 343–348.

Rutherford, J.C. 1994. River Mixing. John Wiley, Chichester, U. K.

Sattar, A. M. A and Gharabaghi B .2015. Gene expression models for prediction of longitudinal dispersion coefficient in streams; Journal of Hydrology. 524: 587–596.

Seo, I and Baek, K .2004 .Estimation of the longitudinal dispersion coefficient using the velocity profile in natural streams; Journal of Hydraulic Engineering, 130(3): 227–236.

Seo, I and Cheong, T. 1998 .Predicting longitudinal dispersion coefficient in natural streams; Journal of Hydraulic Engineering, 124(1): 25–32.

Tayfur, G., Singh, V.P. 2005. Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. Journal of Hydraulic Engineering, 131 (11): 991-1000.

Wang, Y., and Witten, I. H. 1997. Induction of model trees for predicting continuous classes. Proc. of the Poster Papers of the European Conf. on Machine Learning, Univ. of Economics, Faculty of Informatics and Statistics, Prague, Czech Republic.

Investigation of Tree Models Performance for Estimation of Longitudinal Dispersion Coefficient in Straight River

Hosein Nezaratian[1], Javad Zahiri[2], Seyed Mahmood Kashefipour[3]

Abstract

Modeling pollution transmission in rivers is an important subject in environmental studies. Longitudinal dispersion coefficient is one of the key factors in the modelling of lateral dispersion of pollutants. Several researchers have attempted to estimate this coefficient using empirical and semi-empirical methods. However, robust models that can accurately estimate longitudinal dispersion coefficient in river streams are still required. In this study, data driven models were developed using the hydraulic and geometric parameters of rivers. The classification and regression tree (CART), M5 and genetic programming (GP) were used for this purpose. The models performances were then compared quantitatively with those of existing ones using accuracy parameters such as root mean square error (RMSE), mean absolute error (MAE) and discrepancy ratio (DR). The results illustrated that data driven models outperform the existing formulae in term of accuracy. CART model outperform other models in training step, but its performance decrease for testing data. M5 and GP models have RMSE of 0.41 and 0.44 and accuracy of 61% and 62%, respectively. According to small difference between M5 and GP performances, and simple structure of M5 algorithm, this model can be used for estimating longitudinal dispersion coefficient in streams.

Keywords: Pollution Transmission, Longitudinal Dispersion Coefficient, CART, GP, M5

[1] M. Sc., Water structures, Khuzestan Ramin Agriculture and Natural Resources University, Hosein.nezaratian69@gmail.com
[2] Assistant Professor, Department of Water Engineering, Khuzestan Ramin Agriculture and Natural Resources University, Zahiri_Javad@yahoo.com. (Corresponding author)
[3] Professor, Department of Water Engineering, Shahid Chamran University of Ahvaz, Kashefipour@excite.com