**مجله پژوهش آب ایران**
**سال ششم/ شماره یازدهم/ پاییز و زمستان ۱۳۹۱ (۹-۱۶)**

# مقایسه رگرسیون درخت تصمیم، رگرسیون وزن‌دار جغرافیایی و رگرسیون معمولی در ترسیم نقشه‌های هم‌بارش

**خلیل قربانی[۱*]، ابوطالب هزارجریبی[۲]، مهدی ذاکری‌نیا[۳] و ابراهیم اسعدی اسکوئی[۴]**

**چکیده**

از روش‌های ترسیم نقشه‌های هم‌بارش، استفاده از روابط رگرسیونی بین بارش با عوامل جغرافیایی مؤثر بر تغییرات مکانی آن است که خود به چند روش مختلف امکان‌پذیر است. بر این اساس تحقیقی صورت گرفت تا مقایسه‌ای بین روش‌های رگرسیونی سراسری مانند چند جمله‌ای سراسری و رگرسیون معمولی حداقل مربعات با روش‌های رگرسیون موضعی مانند چند جمله‌ای موضعی و رگرسیون وزن‌دار جغرافیایی و همچنین روش رگرسیون درخت تصمیم انجام شود و دقت آن‌ها ارزیابی شود. در انجام این تحقیق از آمار میانگین ۲۰ ساله بارش سالانه ۱۸۵ ایستگاه هواشناسی واقع در استان گیلان و مجاورت آن استفاده و به کمک پارامترهای دیگر مانند ارتفاع از سطح دریا و موقعیت نقاط نسبت به دریا، تغییرات مکانی بارش مدل‌سازی شد. نتایج حاصل از تکنیک اعتبارسنجی تقابلی نشان داد که روش رگرسیون وزن‌دار جغرافیایی با ۸۷= $R^2$ درصد، RMSE=۱۴۷ میلی‌متر از بهترین دقت نسبت به سایر روش‌های رگرسیونی مورد مطالعه برخوردار است و می‌تواند روش مناسبی در ترسیم نقشه‌های هم‌بارش در منطقه مطالعاتی به حساب آید.

**واژه‌های کلیدی**: رگرسیون معمولی حداقل مربعات، رگرسیون وزن‌دار جغرافیایی، رگرسیون درخت تصمیم، گیلان، هم‌بارش.

**ارجاع:** قربانی خ. هزارجریبی ا. ذاکری‌نیا م. و اسعدی‌اسکوئی ا. ۱۳۹۱. مقایسه رگرسیون درخت تصمیم، رگرسیون وزن‌دار جغرافیایی و رگرسیون معمولی در ترسیم نقشه‌های هم‌بارش. مجله پژوهش آب ایران. ۶(۱۱):۹-۱۶.

۱- استادیار گروه مهندسی آب، دانشگاه علوم کشاورزی و منابع طبیعی- گرگان.
۲- استادیار گروه مهندسی آب، دانشگاه علوم کشاورزی و منابع طبیعی- گرگان.
۳- استادیار گروه مهندسی آب، دانشگاه علوم کشاورزی و منابع طبیعی- گرگان.
۴- کارشناس‌ارشد سازمان هواشناسی استان گیلان.
*** نویسنده مسئول:** ghorbani.khalil@yahoo.com
تاریخ دریافت: ۱۳۹۰/۰۶/۲۴ تاریخ پذیرش:۱۳۹۱/۰۱/۱۶

**مقدمه**

با توجه به اهمیت آگاهی از نحوه پراکنش مکانی بارش و از طرفی مشکلات و هزینه‌های اندازه‌گیری‌های میدانی، یافتن راهکار مناسب برای تخمین منطقه‌ای آن امری اجتناب‌ناپذیر است. عوامل محیطی مختلفی مانند ارتفاع و دوری از دریا می‌تواند در تغییرات مکانی بارش مؤثر باشد و با دخالت دادن آن‌ها در معادلات رگرسیونی به عنوان متغیر وابسته، شاید بتوان با دقت قابل قبولی متغیر مستقل یا بارش را پیش‌بینی کرد. اما بدین منظور روش‌های مختلفی توسط محققین مختلف بررسی شده است. خلیلی (۱۳۷۵) نشان داد که گرادیان تغییرات دمای هوا، یک بردار در فضای سه بعدی است که می‌توان تغییرات مؤلفه‌های قائم، نصف النهاری و مداری آن را به وسیله مدل‌های رگرسیونی ساده خطی بیان کرد. همچنین لادو و همکاران (۲۰۰۷) در مدل‌سازی مکانی حداکثر، حداقل و میانگین دمای هوا در ایالت سائوپائولو، دو روش کریجینگ و رگرسیون چند متغیره را ارزیابی کردند و نتیجه گرفتند روش رگرسیون چند متغیره نتایج دقیق‌تری را نسبت به روش کریجینگ ارایه می‌کند. در روش‌های رگرسیونی نحوه دخالت دادن داده‌ها باعث می‌شود تا به دو دسته سراسری[1] و موضعی[2] تقسیم شوند. در روش سراسری کلیه نقاط مشاهده‌ای در برقراری رابطه رگرسیونی دخالت دارند در حالی که در روش رگرسیون موضعی فقط از نقاط مشاهده‌ای موجود در یک محدوده، که به عنوان همسایه یا پنجره شناخته می‌شود، استفاده می‌شود.

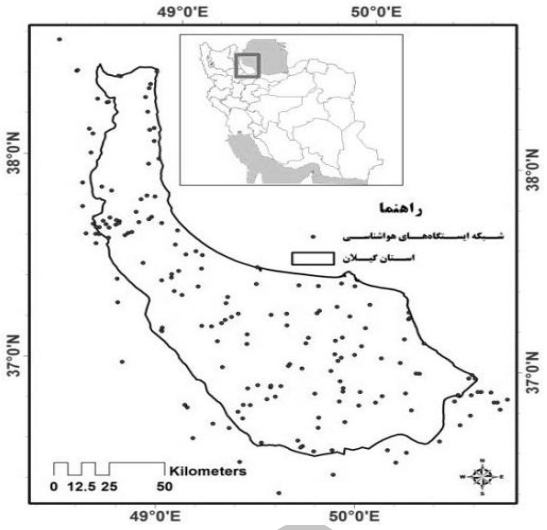
در مدل‌های رگرسیونی نیز استفاده از روش‌های موضعی باعث شده است تا جزییاتی که در روش‌های سراسری نادیده گرفته می‌شود تا حدی به تصویر کشیده شود. یکی از روش‌های رگرسیونی موضعی، روش رگرسیون وزن‌دار جغرافیایی است. در این زمینه می‌توان به مطالعات گوندوگدا و اسن (۲۰۱۰) اشاره کرد که سه روش کریجینگ، کوکریجینگ و رگرسیون وزن‌دار جغرافیایی را برای پهنه‌بندی میانگین ۲۵ ساله بارش سالانه در ترکیه بررسی کردند و با محاسبه ضریب همبستگی بین مقدار پیش‌بینی شده و مدل شده نتیجه گرفتند روش رگرسیون وزن‌دار جغرافیایی با ۸۶= $R^2$ درصد در مقایسه با روش کریجینگ و کوکریجینگ به ترتیب با ۵۱ =$R^2$ درصد و ۶۷=$R^2$ درصد کمترین مقدار خطای پیش‌بینی را به خود اختصاص می‌دهد. بوستان و آکیورک (۲۰۰۷) برای تبیین توزیع مکانی میانگین سالانه بارش و دمای هوا از روش‌های درون‌یابی کوکریجینگ و رگرسیون وزن‌دار جغرافیایی استفاده کردند و با مقایسه همبستگی مقادیر پیش‌بینی و اندازه‌گیری شده به این نتیجه رسیدند که روش رگرسیون وزن‌دار جغرافیایی با ضرایب تبیین ۹۶ و ۶۶ درصد به ترتیب برای دما و بارش نسبت به روش کوکریجینگ با ضرایب تبیین ۸۲ و ۵۴ درصد، درون‌یابی با دقت بیشتری را ارایه می‌کند. اما روش‌های دیگری نیز وجود دارد که اصطلاحاً به کشف دانش[3] می‌پردازند داده‌کاوی[4] یکی از این روش‌ها است که با آن الگوهای مفید از میان داده‌ها با حداقل دخالت کاربران شناخته می‌شوند. داده‌کاوی فرآیندی است که ابزارهای مختلف تحلیل داده را به کار می‌گیرد تا الگوها و روابط فیزیکی متغیرها در مجموعه داده‌های مختلف کشف شود (کرپرشن، ۲۰۰۵). تفاوت اصولی که بین داده‌کاوی و آمار وجود دارد این است که داده‌کاوی یک رهیافت بدون پیش فرض است، درحالی که بیشتر تکنیک‌های آماری معمول نیاز به پیش فرض دارند و آماردان‌ها در جستجوی معادلاتی برای تطبیق با پیش فرض‌ها هستند، اما الگوریتم‌های داده‌کاوی می‌توانند این معادلات را به طور خودکار از اطلاعات موجود در مجموعه داده‌ها توسعه دهند (کابنا و همکاران، ۱۹۹۸). در بین الگوریتم‌های داده‌کاوی، الگوریتم‌های درخت تصمیم روشی برای نمایش یک سری از قوانین ارایه می‌کنند که منتهی به یک رده یا مقدار می‌شوند. درخت‌های تصمیم از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و با برقراری رابطه رگرسیونی در هریک از گروه‌های مجزا، رگرسیون درخت تصمیم ساخته می‌شود که می‌تواند مقادیر کمی پیوسته را پیش‌بینی کند. در این تحقیق سعی شده است تا دقت روش‌های مختلف رگرسیونی در تخمین مکانی بارش ارزیابی شود.

**مواد و روش‌ها**

منطقه مورد مطالعه در این تحقیق استان گیلان است که از میانگین بیست ساله آمار بارش سالانه ۱۸۵ ایستگاه هواشناسی داخل و اطراف آن بین سال‌های ۱۳۶۹-۱۳۸۹ بعد از پالایش و کنترل کیفیت و بازسازی استفاده شد (شکل ۱).

1- Global
2- Local

3 - Knowledge discovery
4 -Data mining




**شکل ۱- موقعیت ایستگاه‌ها و منطقه مطالعاتی**

**روش‌های رگرسیونی ترسیم نقشه‌های هم‌بارش**

**رگرسیون معمولی حداقل مربعات**[1]

هرگاه به کمک یک تابع، مقدار متغیر وابسته‌ای چون y توسط یک یا چند متغیر دیگری به عنوان متغیرهای مستقل تعیین شود یک تابع رگرسیونی بین متغیر وابسته و متغیرهای مستقل برقرار می‌شود. فرم ریاضی تابع خطی ساده به صورت زیر است :

$$Y = \alpha + \beta x \qquad (۱)$$

که در آن مقادیر α و β ضرایب ثابت هستند. ضریب α که عرض از مبدأ نامیده می‌شود، مقدار Y به ازاء X مساوی صفر را نشان می‌دهد. ضریب β که نمایانگر شیب خط است، میزان تغییرات Y را به ازای یک واحد تغییر در X مشخص می‌کند. آمارگران بهترین برازش را عبارت از خطی می‌دانند که مجموع مربعات خطا ($\sum e_i^2$)، کمترین مقدار ممکن را داشته باشد. خطا عبارت است از فاصله عمودی بین مقدار واقعی مشاهده شده و مقداری که برای آن از خط برازش داده شده به دست می‌آید. برای هر مجموعه‌ای از مشاهدات آماری، خطوط مختلف دارای مجموع مربعات خطای متفاوتی خواهند بود. بهترین خط برازش داده شده خطی است که در آن $\sum e_i^2$ دارای کمترین مقدار باشد. این خط به نام خط حداقل مربعات نامیده می‌شود و این روش، روش رگرسیون معمولی حداقل مربعات (OLS) نامیده می‌شود. ضرایب α و β مربوط به خط حداقل مربعات به صورت زیر به دست می‌آیند.

$$\hat{\beta} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2} \quad \text{و} \quad \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \qquad (۲)$$

**روش‌های چندجمله‌ای سراسری و موضعی**

درون‌یابی به روش چند جمله‌ای سراسری یکی از روش‌های درون‌یابی قطعی است که برای سطوحی که تغییرات پیوسته و آرامی دارند مناسب است. در این روش یک سطح هموار با استفاده از یک تابع ریاضی (یک چند جمله‌ای با درجات مختلف) به تمام نقاط نمونه ورودی برازش داده می‌شود. اما درون‌یابی به روش چند جمله‌ای موضعی از قابلیت بیشتری نسبت به درون‌یابی به روش چند جمله‌ای سراسری برخوردار است و بر خلاف آن، تعدادی تابع چند جمله‌ای (با درجات مختلف) به نقاط واقع در یک همسایگی تعیین شده برازش می‌دهد و این باعث می‌شود سطوح با دقت بیشتری ارایه شود.

**روش رگرسیون وزن‌دار جغرافیایی**[2] (GWR)

روش رگرسیون سراسری معمولی، یک رابطه ثابت بین متغیرهای مکانی برای مدل‌سازی منطقه‌ای فرض می‌کند. مدل‌های رگرسیون معمول مانند روش حداقل مربعات معمولی، غیرایستایی مکانی متغیرها را به حساب نمی‌آورند. فایده عمده GWR در مقابل مدل‌های رگرسیون معمولی، توانایی آن در بررسی کردن غیرایستایی مکانی است (پراپاستین و کاپاس، ۲۰۰۸). غیرایستایی مکانی نشان می‌دهد که اندازه‌گیری یا تخمین روابط بین متغیرها از محلی به محل دیگر فرق می‌کند (منیس، ۲۰۰۶). روش GWR یک تکنیک رگرسیون موضعی است که به طور معنی‌داری دقت رگرسیون معمولی را برای استفاده در داده‌های مکانی بهبود داده است. GWR بر مشکل غیرایستا در مدل‌سازی رگرسیونی با جداسازی موضعی آماره‌های سراسری و محاسبه روابط بین متغیرهای موضعی برای هر نقطه به صورت جداگانه غلبه می‌کند. پارامترهای موضعی تخمین زده شده می‌توانند در محل‌های نقاط رگرسیونی ترسیم شوند. بر خلاف مدل‌های رگرسیون معمول که یک معادله رگرسیونی برای توصیف روابط کلی بین متغیرها برقرار می‌کنند، GWR اطلاعات مکان‌یابی تولید می‌کند که تغییرات مکانی بین روابط متغیرها را بیان می‌کند. بنابراین نقشه‌های تولید شده از این تحلیل‌ها نقش کلیدی در توصیف و تفسیر غیرایستایی مکانی بین متغیرها بازی می‌کند (منیس، ۲۰۰۶). جزئیات

1- OLS: Ordinary Least Square

2- Geographically Weighted Regression (GWR)

کامل GWR به وسیله فادرینگهام و همکاران (۲۰۰۲) ارایه شده است. یک مدل رگرسیونی چند متغیره خطی می‌تواند به صورت زیر نوشته شود:

$$y_i = \beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \ldots + \beta_{kn}x_{kn} + \varepsilon_i \quad (۳)$$

$$y_i = \beta_{0i} + \sum_k \beta_k x_{ik} + \varepsilon_i \quad (۴)$$

در اینجا $Y_i$ بارش درون‌یابی شده در موقعیت i، مقدار $\beta_o$ عرض از مبدأ، $\beta_{ik}$ برابر است با $k^{th}$ پارامتر موضعی در $i^{th}$ موقعیت، $X_{iK}$ نشان دهنده $k^{th}$ متغیر مستقل در $i^{th}$ موقعیت و n بیانگر موقعیت قبلی است. در GWR وزن اختصاص داده شده به هر یک از مشاهدات بر اساس یک تابع تنزل فاصله در مرکز مشاهده i است. مدل رگرسیون وزنی جغرافیایی GWR موقعیت مکانی نمونه‌ها را در نظر می‌گیرد و این امکان را می‌دهد تا پارامترهای تخمین‌زده شده به صورت موضعی تغییر کند، یک مدل GWR می‌تواند به صورت زیر نوشته شود:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \quad (۵)$$

در اینجا $(u_i,v_i)$ مختصات i امین موقعیت را نشان می‌دهد، $\beta_0\ (u_i,v_i)$ و $\beta_k\ (u_i,v_i)$ پارامترهای تخمین زده شده برای iامین موقعیت هستند که مقادیر آن‌ها با موقعیت تغییر می‌کند. $x_{ik}$ و $\varepsilon_i$ به ترتیب متغیرهای مستقل و میزان خطا در موقعیت i هستند.

پارامترهای مدل رگرسیون چند متغیره خطی می‌تواند بر اساس حداقل مربعات معمولی به صورت ماتریس زیر تخمین زده شود:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (۶)$$

در اینجا X ماتریس تشکیل شده به وسیله مقادیر $X_T$ و Y بردار تشکیل شده به وسیله مقادیر متغیر Y است. برای GWR، پارامترها با استفاده از یک تابع وزنی به صورت زیر تخمین زده می‌شود:

$$\hat{\beta}(u_i, v_i) = (X^T W((u_i, v_i)X)^{-1} X^T W((u_i, v_i))Y \quad (۷)$$

در اینجا $W(u_i,v_i)$ وزن‌هایی انتخاب شده‌اند به طوری که آن‌هایی که به نقطه تحت مطالعه نزدیک‌تر هستند نسبت به نقاط دورتر تأثیر بیشتری بر نتایج داشته باشند.

**طبقه‌بندی و رگرسیون درخت تصمیم**

درخت‌های تصمیم روشی برای نمایش یک سری از قوانین هستند که منتهی به یک رده یا مقدار می‌شوند. درخت‌های تصمیم از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و هدف در این فرآیند افزایش فاصله بین گروه‌ها در هر جداسازی است. ساختار یک سیستم درختی شامل ریشه، گره‌های داخلی و برگ می‌باشد. ساختار برگ در طبقه‌بندی داده‌های ناشناخته استفاده می‌شود. برای ارزیابی صفتی که توسعه درخت با کمک آن انجام می‌شود از روش ناخالصی نسبت افزایش استفاده می‌شود (کوئینلان، ۱۹۹۳). برگ‌های درخت برچسب‌های کلاسی که اقلام داده در آن گروه‌بندی شده‌اند را می‌سازند. فن طبقه‌بندی تصمیم درختی در دو فاز اجراء می‌شود: ساختن درخت و هرس کردن درخت. درخت از بالا به پایین ساخته می‌شود و در طول این فاز است که درخت به صورت بازگشتی همه اقلام داده‌های متعلق به برچسب کلاس یکسان را افراز می‌کند (هانسن و همکاران، ۱۹۶۶). اما هرس کردن درخت با روش از پایین به بالا دقت پیش‌بینی و طبقه‌بندی را بهبود می‌بخشد. برای این کار سعی می‌شود میزان برازش بیش از حد را در مرحله آموزشی به حداقل رساند (آموزش بیش از حد به این معنی است که در مرحله آموزش، نویزها و جزییات موجود در داده‌ها نیز در یادگیری منظور شوند) (مهتا و همکاران، ۱۹۹۶). مدل‌های مختلفی از درخت تصمیم در داده‌کاوی وجود دارد، الگوریتم C&RTree (طبقه‌بندی و رگرسیون درختی) نمونه‌ای از آن‌ها است که توسط بریمن و همکاران (۱۹۸۴) معرفی شد. این الگوریتم ابزار درخت تصمیم نیرومندی را ایجاد می‌کند که می‌تواند به آسانی از میان مجموعه داده‌های بزرگ الگوها و روابط را جستجو کند. الگوریتم C&RT یک فرآیند افراز بازگشتی دودویی می‌باشد که گره‌های والدین را دقیقاً به دو گره فرزند منشعب می‌کند و به طور بازگشتی منشعب کردن را تا زمانی‌که توسعه دیگری نتواند ساخته شود ادامه می‌دهد. توسعات با پرسیدن یک سؤال با جواب بلی یا خیر تعیین می‌شوند. داده‌ها به دو زیرمجموعه افراز می‌شوند به طوری که رکوردهای درون هر زیر مجموعه نسبت به زیر مجموعه قبلی همگن‌تر باشند. این الگوریتم قادر به پردازش صفت‌های خاص با مقادیر پیوسته و گسسته است و از ضریب جینی[1] که شبیه به معیار بهره اطلاعاتی است برای افراز کلاس‌ها استفاده می‌کند.

**معیارهای ارزیابی روش‌های رگرسیونی در ترسیم هم‌بارش**

برای ارزیابی دقت و اعتبار روش‌های درون‌یابی از تکنیک اعتبارسنجی تقابلی (حذفی) استفاده شد. این تکنیک بر

1-Gini

این اساس استوار است که هر بار یک ایستگاه حذف شده و عمل درون‌یابی با دیگر ایستگاه‌ها انجام می‌شود. مقدار درون‌یابی شده در محل ایستگاه حذف شده به عنوان مقدار برآورد شده برای آن ایستگاه در نظر گرفته می‌شود. پس با برگرداندن ایستگاه حذف شده در جایگاه اصلی خود، ایستگاه دیگری حذف شده و به ترتیب این عمل برای هر یک از ایستگاه‌ها انجام می‌شود. در نهایت با توجه به مقادیر مشاهده شده ($Z(x_i)$) و برآورد شده ($\hat{Z}(x_i)$)، میانگین خطای مطلق[1] و ریشه میانگین مربعات خطا[2] در هر روش از روابط زیر محاسبه می‌شود:

$$MAE = \sum_{i=1}^{n} \left| (Z(x_i) - \hat{Z}(x_i)) \right| / n \qquad (۸)$$

$$RMSE = \left( \sum_{i=1}^{n} (Z(x_i) - \hat{Z}(x_i))^2 / n \right)^{1/2} \qquad (۹)$$

## نتایج و بحث

نتایج روش‌های مختلف درون‌یابی مورد مطالعه در این تحقیق با استفاده از تکنیک اعتبارسنجی تقابلی ارزیابی و مقایسه شد. مقدار خطای درون‌یابی (اختلاف بین مقادیر مشاهده شده و مقادیر مدل‌سازی شده) محاسبه شد تا برای محاسبه معیارهای ارزیابی خطا شامل ریشه مجذور مربعات خطا و میانگین خطای مطلق استفاده شود. همچنین با برقراری رابطه رگرسیونی بین مقادیر مشاهده شده و مقادیر مدل‌سازی شده، ضرایب همبستگی بین این دودسته متغیر نیز محاسبه شدند.

**رگرسیون معمولی حداقل مربعات**

میانگین بارش سالانه به عنوان متغیر وابسته در نظر گرفته شد و با درنظر گرفتن متغیرهای ارتفاع ، فاصله از دریا، طول و عرض جغرافیایی به ترتیب با نمادهای Z، D، X و Y روابط رگرسیونی یک تا چند متغیره آزمایش شد. نتایج این آزمون‌ها در جدول ۱ نشان می‌دهد که بهترین رابطه رگرسیونی رابطه OLS-ZXYD است که با $R^2$=۵۴ درصد و RMSE=۲۷۵ میلی‌متر کمترین مقدار خطا را در بین رگرسیون‌های معمولی حدقل مربعات به خود اختصاص داده است.

**روش درون‌یابی چندجمله ای سراسری و موضعی**

در روش سراسری با برقراری یک رابطه چند جمله‌ای بین مختصات طول و عرض جغرافیایی و ارتفاع ایستگاه‌ها با بارش، مقدار بارش را در کل منطقه مطالعاتی درون‌یابی می‌شود. نتایج این بررسی نشان می‌دهد که با افزایش درجه چندجمله‌ای از یک به سه، مقدار RMSE که خطای پیش‌بینی را نشان می‌دهد کاهش یافته و ضریب همبستگی رابطه چند جمله‌ای افزایش می‌یابد (جدول ۲). اما در روش درون‌یابی چند جمله‌ای موضعی بر خلاف روش درون‌یابی چندجمله‌ای سراسری از چند رابطه رگرسیونی بین مختصات طول و عرض جغرافیایی و ارتفاع ایستگاه‌ها با بارش آن‌ها استفاده شد تا مقدار بارش در کل منطقه مطالعاتی درون‌یابی شود. این عمل باعث افزایش دقت درون‌یابی شده است که نشان دهنده وجود روند موضعی تغییرات بارش در منطقه است.

**جدول ۱- نتایج ارزیابی خطا و ضرایب تبیین مدل‌های مختلف رگرسیون معمولی حداقل مربعات خطا**

| علامت مدل | معیارهای ارزیابی خطا | | ضریب تبیین |
|---|---|---|---|
| | RMSE | MAE | $R^2$ |
| OLS-Z | ۳۴۱ | ۲۶۹ | ۰/۳۰ |
| OLS-D | ۳۱۶ | ۲۵۱ | ۰/۴۰ |
| OLS-ZXY | ۳۰۹ | ۲۴۵ | ۰/۴۲ |
| OLS- ZXYD | ۲۷۵ | ۲۲۱ | ۰/۵۴ |
| OLS-ZYD | ۲۷۸ | ۲۲۴ | ۰/۵۳ |
| OLS-ZD | ۲۷۹ | ۲۲۴ | ۰/۵۳ |
| OLS-ZY | ۳۲۸ | ۲۶۳ | ۰/۳۵ |

**جدول ۲- نتایج ارزیابی خطا و ضرایب تبیین روش‌های رگرسیونی چندجمله‌ای سراسری و موضعی**

| مدل | علامت مدل | معیارهای ارزیابی | | ضریب تبیین |
|---|---|---|---|---|
| | | RMSE | MAE | $R^2$ |
| چندجمله‌ای سراسری با درجات ۱، ۲ و ۳ | GPI1 | ۳۵۸/۸ | ۲۹۵/۲ | ۰/۲۳ |
| | GPI2 | ۲۶۱/۴ | ۲۱۸/۱ | ۰/۵۹ |
| | GPI3 | ۲۵۱/۶ | ۲۰۴/۳ | ۰/۶۲ |
| چندجمله‌ای موضعی با درجات ۱، ۲ و ۳ | LPI1 | ۱۹۳/۹ | ۱۴۷/۸ | ۰/۷۷ |
| | LPI2 | ۲۲۰/۶ | ۱۳۷/۷ | ۰/۷۱ |
| | LPI3 | ۲۰۷/۹ | ۱۶۰/۹ | ۰/۷۴ |

**رگرسیون وزن‌دار جغرافیایی**

روش رگرسیون وزن‌دار جغرافیایی یک روش آماری است که برای مطالعه الگوهای موضعی سازگار شده است. روش GWR وزن‌های نسبی بیشتری به مشاهدات نزدیک‌تر و وزن کمتر یا صفر به آن‌هایی که در دوردست هستند، اختصاص می‌دهد. به عبارت دیگر، GWR فقط از مشاهداتی که از لحاظ جغرافیایی

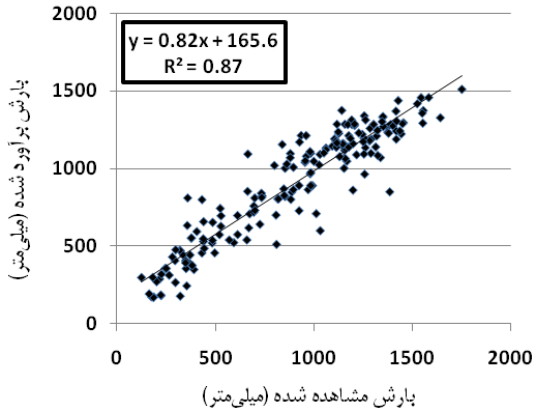
1- Mean Absolute Error (MAE)
2- Root Mean Square Error (RMSE)

نزدیک هستند برای تخمین ضرایب موضعی استفاده می‌کند. این شیوه وزن‌دهی بر اساس این تفکر است که استفاده از مشاهدات نزدیک از لحاظ جغرافیایی بهترین روش برای تخمین ضرایب موضعی است. روش GWR نه تنها اثرات موقعیت خود متغیرها را روی متغیر مستقل بلکه اثرات موقعیت‌های همسایگی را نیز در نظر می‌گیرد. این موضوع باعث شده است تا دقت این روش نسبت به دیگر روش‌های درون‌یابی برای پهنه‌بندی مقدار بارش در منطقه مطالعاتی به طور معنی‌داری افزایش یابد به طوری‌که باعث کاهش RMSE تا میزان ۱۴۷ میلی‌متر و افزایش ضریب تبیین تا حدود ۸۷ درصد شود. پارامترهای ارتفاع (Z)، فاصله از دریا (D)، زاویه آزیموت سمت دریا (A) وارد معادلات رگرسیونی شدند تا نتایج آن‌ها در تخمین مقدار بارش در منطقه مطالعاتی ارزیابی شود.

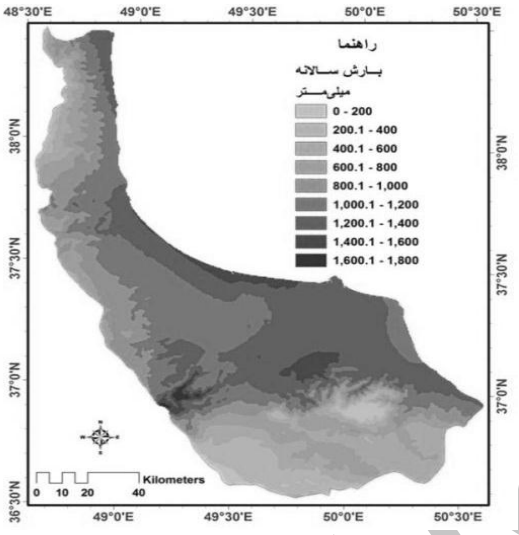
نتایج این بررسی نشان داد (جدول ۳) که دو پارامتر ارتفاع و فاصله از دریا از عوامل مؤثر در تغییرات مکانی بارش در منطقه هستند به طوری‌که مدل رگرسیونی وزن‌دار جغرافیایی حاصل از آن‌ها (GWR-ZD) کمترین مقدار خطا با RMSE=۱۴۷ میلی‌متر و بیشترین همبستگی با ضریب همبستگی ۹۳ درصد (معادل ضریب تبیین ۸۷ درصد) را دارد (شکل‌های ۲ و ۳). دخالت دادن پارامتر زاویه آزیموت سمت دریا نه تنها باعث کاهش خطای پیش‌بینی بلکه باعث افزایش خطای پیش‌بینی شده است (مدل‌های GWR-ZA و GWR-ZDA) ولی پارامتر ارتفاع به تنهایی ۸۶ درصد واریانس بارش را در منطقه توجیه می‌کند و مدل GWR-Z نیز بعد از مدل GWR-ZD کمترین مقدار خطا را با RMSE=۱۵۰ میلی‌متر دارد.

**جدول ۳- نتایج ارزیابی خطا و ضرایب تبیین رگرسیون وزن‌دار جغرافیایی**

| علامت مدل | معیارهای ارزیابی خطا | | ضریب تبیین |
|---|---|---|---|
| | RMSE | MAE | $R^2$ |
| GWR-Z | ۱۵۰/۲ | ۱۱۵/۳ | ۰/۸۶ |
| GWR-ZD | ۱۴۷/۱ | ۱۱۳/۱ | ۰/۸۷ |
| GWR-ZA | ۲۱۳/۵ | ۱۷۲/۹ | ۰/۷۲ |
| GWR-ZDA | ۱۷۷/۷ | ۱۳۷/۶ | ۰/۸۱ |




**شکل ۲- رابطه بین مقادیر مشاهده شده و برآورد شده بارش با مدل رگرسیون وزن‌دار جغرافیایی GWR-ZXY**



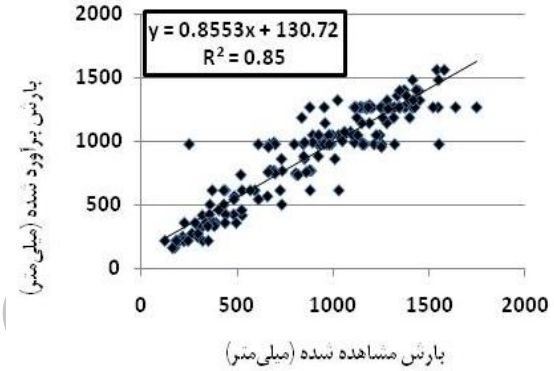

**شکل ۳- نقشه هم‌بارش سالانه ترسیم شده با مدل رگرسیون وزن‌دار جغرافیایی**

### رگرسیون درخت تصمیم

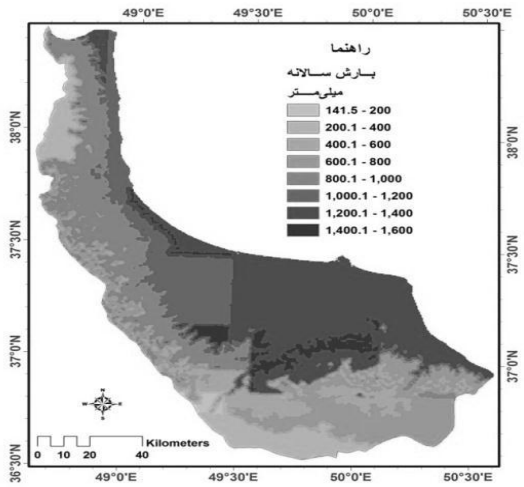
رگرسیون درخت تصمیم با افراز داده‌ها به کلاس‌های مختلف و در نهایت برقراری روابط رگرسیونی در هر کلاس، متغیر وابسته را پیش‌بینی می‌کند. نتایج حاصل از اجرای این مدل با متغیرهای ورودی مختلف (جدول ۴) نشان می‌دهد که مدل C&RT-XYZ بهترین دقت را در بین مدل‌های رگرسیون درخت تصمیم، داراست (شکل‌های ۴ و ۵). ورود متغیر طول جغرافیایی در مدل C&RT-ZXYD افزایش قابل توجهی را در دقت مدل ندارد ولی برعکس دو متغیر ارتفاع و دوری از دریا بیشترین تأثیر را در دقت مدل دارند به طوری که هر یک از آن‌ها به تنهایی ضریب تبیینی بیش از ۴۶ درصد را به همراه دارند و استفاده از هر دوی آن‌ها باعث افزایش ضریب تبیین تا ۷۵ درصد و کاهش RMSE می‌شود.

**جدول ۴- نتایج ارزیابی خطا و ضرایب تبیین رگرسیون‌های درخت تصمیم**

| علامت مدل | معیارهای ارزیابی خطا | | ضریب تبیین |
|---|---|---|---|
| | RMSE | MAE | $R^2$ |
| C&RT-Z | ۲۷۶ | ۲۰۰ | ۵۴ |
| C&RT-X | ۳۵۸ | ۲۹۱ | ۲۲/۵ |
| C&RT-Y | ۲۸۴ | ۲۲۴ | ۵۱ |
| C&RT-D | ۲۹۸ | ۲۲۷ | ۴۶ |
| C&RT-ZY | ۱۸۵ | ۱۲۹ | ۷۹/۲ |
| C&RT-ZD | ۱۹۰ | ۱۴۸ | ۷۸/۲ |
| C&RT-ZXY | ۱۵۵ | ۱۰۹ | ۸۵/۵ |
| C&RT-ZXYD | ۱۵۵ | ۱۱۶ | ۸۵/۴ |
| C&RT-ZYD | ۱۵۹ | ۱۲۰ | ۸۴/۷ |




**شکل ۴- رابطه بین مقادیر مشاهده شده و برآورد شده بارش با مدل رگرسیون درخت تصمیم C&RT-ZXY**




**شکل ۵- نقشه هم‌بارش سالانه ترسیم شده با مدل رگرسیون درخت تصمیم C&RT-ZXY**

رگرسیون درخت تصمیم بعد از رگرسیون وزن‌دار جغرافیایی کم‌ترین مقدار خطا را به خود اختصاص داده است. مدل‌سازی به کمک داده‌کاوی مستلزم وجود داده‌های زیادی است که از توزیع مناسبی برخوردار باشند. در این تحقیق تغییرات مکانی بارش دامنه‌ای در حدود ۱۲۴ تا ۱۷۵۰ میلی‌متر را داراست.

مطالعات زیادی در زمینه ارزیابی روش‌های مختلف ترسیم نقشه‌های هم‌مقدار انجام شده است که بیشتر به مقایسه روش‌های درون‌یابی قطعی و زمین آمار پرداخته‌اند. با وجود شباهت‌هایی که در نتایج مطالعات مختلف به دست آمده، اختلافات زیادی نیز در آن‌ها به چشم می‌خورد. اما در مورد روش رگرسیون وزن‌دار جغرافیایی در پهنه‌بندی و ترسیم نقشه‌های هم‌مقدار مطالعات کمی صورت گرفته است که می‌توان به مطالعات گوندوگدا و اسن (۲۰۱۰) در ترکیه اشاره کرد. آن‌ها روش رگرسیون وزن‌دار جغرافیایی را با دیگر روش‌های درون‌یابی را مقایسه کردند و نتیجه گرفتند روش رگرسیون وزن‌دار جغرافیایی نسبت به روش‌های کریجینگ و کوکریجینگ در میان‌یابی بارش نتایج بهتری را به همراه دارد در این تحقیق نیز برتری روش رگرسیون وزن‌دار جغرافیایی نسبت به روش رگرسیون درخت تصمیم و دیگر روش‌های رگرسیونی به اثبات رسید.

## نتیجه‌گیری

رگرسیون وزن‌دار جغرافیایی یک روش آماری است که برای مطالعه الگوهای موضعی سازگار شده است و در مناطقی که از الگوهای تغییرات مکانی مختلفی از بارش پیروی می‌کنند روش مناسبی برای پهنه‌بندی و تخمین داده‌های مکانی به حساب می‌آید. نتایجی که از این تحقیق به دست آمد نیز نشان داد که این روش برتری معنی‌داری نسبت به دیگر روش‌های رگرسیونی مورد مطالعه در این تحقیق دارد به طوری که با دخالت دادن دو پارامتر ارتفاع و فاصله از دریا به کمک این روش می‌توان با مقدار خطایی برابر RMSE=۱۴۷ میلی‌متر بارش را در منطقه مطالعاتی پهنه‌بندی کرد. بعد از روش رگرسیون وزن‌دار جغرافیایی، روش رگرسیون درخت تصمیم دقت بالاتری را نشان می‌دهد. بر این اساس می‌توان نتیجه گرفت که روش رگرسیون وزن‌دار جغرافیایی می‌تواند ابزاری مناسب برای توجیه تغییرات موضعی بارش و مناسب‌ترین روش در ترسیم نقشه‌های هم‌بارش سالانه در استان گیلان به حساب آورده شود.

## منابع


۱. خلیلی ع. ۱۳۷۵. تغییرات سه بعدی میانگین‌های سالانه درازمدت دمای هوا در گستره ایران. مجله نیوار. ۳۲:۱۳-۲۴.

2. Bostan A. P. and Akyurek Z. 2007. Exploring the mean annual precipitation and temperature values over Turkey by using environmental variables. International Society for Photogrammetry and Remote Sensing: Visualization and Exploration of Geospatial Data. 6 pp.
3. Breiman L. Freidman J. H. Olshen R. A. and Stone C. J. 1984. Classification and Regression Trees: Chapman and Hall/CRC, New York. 358 pp.
4. Cabena P. H. Stadler R. Verhees J. and Zanasi A. 1998. Discovering Data Mining: From Concept to Implementation, IBM, New Jersey. 195 pp.
5. Fotheringham A. S. Brunsdon C. and Charlton M. E. 2002. Geographically weighted regression. Chichester. John Wiley & Sons. 268 pp.
6. Gundogdu I. and Esen O. 2010. The importance of secondary variables for mapping of meteorological data. 3rd international conference on cartography and GIS, Nessebar, Bulgaria. 10 pp.
7. Hansen M. Dubayah R. and DeFries R. 1996. Classification trees: an alternative to traditional land cover classifiers: International Journal of Remote Sensing. 17:1075-1081.
8. Lado L. Sparovek G. Torrado P. Neto D. and Vázquez F. 2007. Modeling air temperature for the state of Sao Paul, Brazil. Scientia Agricola journal. 64(5):460-467.
9. Mehta M. Agrawal R. and Rissanen J. 1996. SLIQ: A fast scalable classifier for data mining: Proceeding of the Fifth International Conference on Extending Database Technology, Avignon, France. 1057:18-32.
10. Mennis J. 2006. Mapping the Results of Geographically Weighted Regression. The Cartographic Journal. 43(2):171-179.
11. Propastin P. and Kappas M. 2008. Reducing uncertainty in modeling the NDVI–precipitation relationship: a comparative study using global and local regression techniques. GISci Remote Sens. 45:47-67
12. Quinlan J. R. 1993. C45: Programs for Machine Learning: Morgan Kaufmann Publishers Inc, San Francisco. 302 pp.
13. Tobler W. R. 1970. A computer movie simulating urban growth in the Detroit region, Economic Geography. 46:234-240.
14. Corporation T. C. 2005. Introduction to data mining and knowledge discovery. Third edition. 36 pp.