https://doi.org/10.22034/aa.2026.733260
اندیشه آماری، سال ۱۴۰۳، دوره ۲۹، شماره ۲، ص ۱۲ – ۱

تحلیل متغیرهای هیدرولوژیکی و خصوصیات خاک با روش‌های رگرسیون ریج و بردار پشتیبان فضایی

مهدی روزبه[۱]* و آرش عامری[۲]

[۱] گروه آمار، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه سمنان، سمنان، ایران

[۲] گروه آمار، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه سمنان، سمنان، ایران

تاریخ دریافت: ۱۴۰۴/۰۸/۲۱
تاریخ پذیرش: ۱۴۰۴/۱۰/۱۳

چکیده

روش‌های سنتی آماری با گسترش داده‌های فضایی با ساختار مکانی–زمانی پیچیده با چالش‌های جدی مواجه شده‌اند. این داده‌ها به دلیل خودهمبستگی مکانی، ناهمسانی واریانس و وابستگی‌های جغرافیایی پیچیده، نیازمند روش‌های تخصصی هستند. در این پژوهش، رگرسیون بردار پشتیبان به‌عنوان یک روش نوین برای تحلیل و مدل‌سازی ساختار پیچیدی فضایی داده‌های زمین‌آماری مربوط به مقادیر کلسیم و منیزیم موجود در خاک معرفی می‌شود. این تحلیل بر اساس مختصات جغرافیایی مختلف (شرقی–غربی و شمالی–جنوبی)، در دو عمق ۰ تا ۲۰ و ۲۰ تا ۴۰ سانتی‌متر و در سه منطقه‌ی جغرافیایی متنوع انجام گرفته است. روش رگرسیون بردار پشتیبان با توانایی مدل‌سازی روابط غیرخطی پیچیده و در عین حال حفظ ساختار فضایی داده‌ها، امکان پیش‌بینی دقیق‌تر و واقع‌بینانه‌تر توزیع عناصر تغذیه‌ای خاک را فراهم می‌سازد. این رویکرد، با بهره‌گیری از توابع هسته، قابلیت تحلیل فضاهای ویژگی با ابعاد بالا و پیچیدگی‌های ساختاری را فراهم می‌آورد و اثربخشی خود را در برابر جملات اختلال و داده‌های پرت اثبات می‌کند. برای سنجش دقیق کارایی رگرسیون بردار پشتیبان، عملکرد آن در برابر رگرسیون ریج مورد ارزیابی مقایسه‌ای قرار می‌گیرد.

واژه‌های کلیدی: اعتبارسنجی متقابل تعمیم‌یافته، رگرسیون ریج، رگرسیون بردار پشتیبان، ماشین‌های بردار پشتیبان
رده‌بندی ریاضی: ۶۲J۰۵; ۶۲H۲۵.

۱ مقدمه

در عصر حاضر، پیشرفت‌های علمی و فناوری منجر به تولید حجم گسترده‌ای از داده‌های فضایی شده است که دارای ساختارهای پیچیده مکانی–زمانی هستند. این داده‌ها شامل اطلاعات جغرافیایی، موقعیت‌های مکانی و روابط میان پدیده‌ها در فضا می‌باشند و تحلیل آن‌ها با چالش‌هایی مانند خودهمبستگی مکانی، ناهمسانی واریانس و ساختارهای وابستگی پیچیده همراه است. از آنجا که روش‌های آماری سنتی توانایی پاسخگویی کامل به این چالش‌ها را ندارند، توسعه و به‌کارگیری روش‌های نوین برای مدل‌سازی و تحلیل داده‌های فضایی ضرورت دارد. در این راستا، رگرسیون بردار پشتیبان[۱] (SVR) به‌عنوان یک روش قدرتمند در پردازش داده‌های مکانی مطرح است که توانایی

*نویسنده مسئول مقاله، mahdi.roozbeh@semnan.ac.ir

[1] Support Vectors Regression
[2] Ridge Regression

مدل‌سازی روابط غیرخطی پیچیده در کنار حفظ ساختار فضایی داده‌ها را داراست [۹]. علاوه بر این، رگرسیون ریج[۲] نیز به‌عنوان یک روش مقایسه‌ای مورد استفاده قرار می‌گیرد. مزیت اصلی الگوریتم ماشین بردار پشتیبان در تحلیل‌های فضایی، قابلیت ترکیب اطلاعات مکانی با متغیرهای کمکی، مقاومت در برابر داده‌های پرت و اختلال، و توانایی پردازش فضاهای ویژگی پیچیده با استفاده از توابع هسته است. این ویژگی‌ها موجب می‌شود که این الگوریتم بتواند وابستگی‌های فضایی را بدون نیاز به فرضیات سخت‌گیرانه مدل‌سازی کند و در مسائل کاربردی نتایج دقیق‌تری ارائه دهد.

آمار فضایی شاخه‌ای از علم آمار است که به تحلیل داده‌هایی می‌پردازد که وابستگی مکانی یا ساختار جغرافیایی دارند. در این نوع داده‌ها، مقدار مشاهده شده در یک موقعیت نه‌تنها به ویژگی‌های همان نقطه بلکه به مقادیر نقاط مجاور نیز وابسته است. هدف اصلی آمار فضایی، شناسایی و مدل‌سازی الگوهای مکانی، بررسی روابط بین پدیده‌ها در فضا و پیش‌بینی رفتار آن‌ها در موقعیت‌های ناشناخته است. این حوزه با استفاده از ابزارهایی مانند مدل‌های خودرگرسیون فضایی، آزمون‌های وابستگی مکانی و روش‌های خوشه‌بندی فضایی، امکان درک بهتر ساختارهای جغرافیایی و تصمیم‌گیری علمی در زمینه‌هایی همچون محیط زیست، شهرسازی، سلامت عمومی و علوم زمین را فراهم می‌آورد [۱].

۲ رگرسیون ریج

رگرسیون ریج یکی از روش‌های تحلیل رگرسیونی است که برای مقابله با مشکل همخطی چندگانه[۳] در داده‌ها معرفی شده است. این روش نخستین بار توسط هورل و کنارد در سال ۱۹۷۰ ارائه شد و به‌عنوان جایگزینی برای روش کمترین توان‌های دوم معمولی (OLS)[۴] در شرایطی که متغیرهای مستقل همبستگی بالایی دارند یا زمانی که تعداد متغیرها (p) نزدیک به تعداد مشاهدات (n) باشد، به کار می‌رود. در رگرسیون ریج، با افزودن یک پارامتر تنظیم (λ) به ماتریس واریانس–کوواریانس، واریانس برآوردگرها کاهش می‌یابد [۸]. این روش با قبول اندکی اریبی در برآوردها، واریانس را به میزان قابل توجهی کاهش داده و منجر به برآوردهای پایدارتری می‌شود. همچنین، توسعه این روش به‌صورت رگرسیون ریج تعمیم‌یافته[۵]، امکان در نظر گرفتن پارامتر اریب جداگانه برای هر متغیر مستقل را فراهم می‌سازد که موجب بهبود بیشتر در دقت برآوردها می‌گردد. رگرسیون ریج به‌ویژه در تحلیل داده‌های فضایی که اغلب با چالش همخطی بین متغیرهای کمکی مواجه هستند، می‌تواند به‌عنوان یک روش کارآمد برای مدل‌سازی روابط پیچیده مورد استفاده قرار گیرد [۵].

۱.۲ جریمه در رگرسیون ریج

رگرسیون ریج با اعمال جریمه بر ضرایب رگرسیونی، آن‌ها را کاهش و مجموع توان‌های دوم باقیمانده‌ها را به حداقل می‌رساند. این روش، که به‌عنوان رگرسیون ریج شناخته می‌شود، پارامترها را به سمت صفر هدایت کرده و واریانس برآوردها را کاهش می‌دهد. بنابراین، می‌توان تعریف این رویکرد را مطابق رابطهٔ (۱) به‌صورت زیر ارائه کرد:

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left[\sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{P} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \right] \quad (۱)$$

در رابطه (۱) $\lambda \geq 0$ به‌عنوان پارامتر جریمه تعریف می‌شود که باید به‌طور جداگانه تعیین شود و بخش دوم معادله $\lambda \sum_{j=1}^{P} \boldsymbol{\beta}_j^2$ که به‌عنوان جریمه انقباض شناخته می‌شود، زمانی که ضرایب $\beta_1, \beta_2, \ldots, \beta_p$ نزدیک به صفر هستند، کوچک می‌شود و باعث کاهش برآوردهای $\boldsymbol{\beta}_j$ به سمت صفر می‌گردد.

همچنین پارامتر تنظیم‌کننده λ به کنترل تأثیر نسبی این دو بخش بر برآوردهای ضرایب رگرسیون کمک می‌کند. زمانی که $\lambda = 0$ باشد، جریمه اثری ندارد و رگرسیون ریج مشابه برآوردهای کمترین توان‌های دوم معمولی خواهد بود. اما با افزایش λ به سمت بی‌نهایت، تأثیر جریمه انقباض بیشتر می‌شود و برآوردهای ضرایب رگرسیون ریج به سمت صفر سوق پیدا می‌کنند. همچنین می‌توان برآورد رگرسیون ریج را به شکل ماتریسی طبق رابطه (۲) نوشت:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \hat{\boldsymbol{\beta}}(k) = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \quad \lambda \geq 0 \quad (۲)$$

که $\hat{\boldsymbol{\beta}}^{Ridge}$ به‌عنوان برآوردگر ریج شناخته می‌شود. بر خلاف بعضی از روش‌های رگرسیونی، در روش رگرسیون ریج، ماتریس $\boldsymbol{X}^\top \boldsymbol{X}$ دارای وارون است [۶].

[3] Multicollinearity
[4] Ordinary Least Squares
[5] Generalized Ridge Regression

۳ ماشین‌های بردار پشتیبان و رگرسیون بردار پشتیبان

در ادامه به بررسی ماشین‌های بردار پشتیبان[۶] (SVM) و رگرسیون بردار پشتیبان می‌پردازیم.

۱.۳ ماشین‌های بردار پشتیبان با حاشیه سخت

ماشین‌های بردار پشتیبان یکی از روش‌های قدرتمند یادگیری ماشین هستند که توسط ولادیمیر وپنیک و الکسی چروننکیس در دهه ۱۹۹۰ میلادی توسعه یافتند. ماشین‌های بردار پشتیبان براساس مفاهیم ابرصفحه (مرز تصمیم‌گیری)، بردارهای پشتیبان (نقاط کلیدی نزدیک به مرز تصمیم‌گیری) و حاشیه (فاصله بین ابرصفحه و نزدیک‌ترین نقاط داده) عمل می‌کنند. هدف اصلی SVM، بیشینه‌سازی حاشیه برای دستیابی به مدلی با تعمیم‌پذیری بهتر است [۷، ۴].

هنگامی که داده‌ها به‌طور خطی جداپذیر باشند، از حاشیه سخت استفاده می‌شود. در این حالت، SVM به دنبال یافتن ابرصفحه‌ای است که دو کلاس $y_i \in \{-۱, +۱\}$ را با حداکثر فاصله از یکدیگر جدا کند، به‌طوری که هیچ نقطه‌ای درون حاشیه قرار نگیرد. این مسئله به‌صورت یک مسئله بهینه‌سازی برای مینیمم‌سازی تابع هدف موجود در رابطه (۳) فرمول‌بندی می‌شود [۲]:

$$\begin{aligned} & J(\boldsymbol{w}) = \tfrac{۱}{۲}\|\boldsymbol{w}\|^۲ \\ & \text{s.t.} \quad y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq ۱, \quad i = ۱, ۲, \ldots, n \end{aligned} \tag{۳}$$

۲.۳ ماشین‌های بردار پشتیبان با حاشیه نرم

برای داده‌هایی که به‌طور خطی جداپذیر نیستند، حاشیه نرم معرفی شده است. در این روش با معرفی متغیرهای آستانه خطا (ξ_i)، به مدل اجازه داده می‌شود تا در طبقه‌بندی برخی از نقاط نزدیک به مرز مقداری خطا وجود داشته باشد. مسئلهٔ بهینه‌سازی مرتبط با مینیمم‌سازی تابع هدف در این شرایط مطابق رابطه (۴) تعریف می‌شود و پارامترها با توجه به این محدودیت‌ها تنظیم می‌گردند [۲]:

$$\begin{aligned} \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \quad & J(\boldsymbol{w}, b, \boldsymbol{\xi}) = \tfrac{۱}{۲}\|\boldsymbol{w}\|^۲ + C\sum_{i=۱}^{n} \xi_i \\ \text{s.t.} \quad & y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq ۱ - \xi_i, \qquad i = ۱, \ldots, n \\ & \xi_i \geq ۰, \qquad i = ۱, \ldots, n \end{aligned} \tag{۴}$$

پارامتر تنظیم (C) در این میان نقش مهمی دارد، به‌طوری‌که مقادیر بزرگ‌تر C خطاهای طبقه‌بندی را کاهش داده اما حاشیه کوچک‌تری ایجاد می‌کند، در حالی که مقادیر کوچک‌تر C به مدل اجازه خطای بیشتر داده و حاشیه بزرگ‌تری فراهم می‌آورد.

۳.۳ روش‌های هسته در ماشین‌های بردار پشتیبان

برای مسائل غیرخطی، از توابع هسته برای نگاشت داده‌ها به فضایی با ابعاد بالاتر استفاده می‌شود، جایی که داده‌ها به‌صورت خطی جداپذیر می‌گردند. برخی از پرکاربردترین توابع هسته عبارتند از [۲]:

۱: هسته خطی: $K(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{x}^\top \cdot \boldsymbol{u}$

۲: هسته چندجمله‌ای: $K(\boldsymbol{x}, \boldsymbol{u}) = (a\boldsymbol{x}^\top \boldsymbol{u} + c)^q$

۳: هسته سیگموئید: $K(\boldsymbol{x}, \boldsymbol{u}) = \tanh(\beta \boldsymbol{x}^\top \boldsymbol{u} + \gamma)$

۴: هسته گاوسی (شعاعی): $K(\boldsymbol{x}, \boldsymbol{u}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{u}\|^۲}{\sigma^۲}\right)$

۴.۳ رگرسیون بردار پشتیبان

رگرسیون بردار پشتیبان تعمیمی از ماشین‌های بردار پشتیبان به مسائل رگرسیون است که به‌جای دسته‌بندی، خروجی‌های پیوسته را برآورد می‌کند. در این روش، یک ناحیه حساسیت‌ناپذیر (ε) حول تابع برآورد تعریف می‌شود که خطاهای کوچک‌تر از ε را نادیده می‌گیرد و تنها خطاهای خارج از این ناحیه جریمه می‌شوند. بدین ترتیب، بردارهای پشتیبان نقاطی هستند که خارج از این حاشیه قرار دارند و ساختار مدل را تعیین می‌کنند. فرمول‌بندی بهینه‌سازی SVR به‌شکل رابطه (۵) ارائه می‌شود:

$$\begin{aligned} \min_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \quad & \tfrac{۱}{۲}\|\mathbf{w}\|^۲ + C\sum_{i=۱}^{n}(\xi_i + \xi_i^*) \\ \text{s.t.} \quad & y_i - \mathbf{w}^\top \mathbf{x}_i - b \leq \varepsilon + \xi_i, \quad i = ۱, \ldots, n \\ & \mathbf{w}^\top \mathbf{x}_i + b - y_i \leq \varepsilon + \xi_i^*, \quad i = ۱, \ldots, n \\ & \xi_i \geq ۰,\ \xi_i^* \geq ۰, \qquad i = ۱, \ldots, n \end{aligned} \tag{۵}$$

که در آن C پارامتر تنظیم تعادل میان پیچیدگی مدل و دقت پیش‌بینی و همچنین ξ_i و ξ_i^* متغیرهای خطا برای نقاط خارج از ناحیه ε هستند. این فرمول‌بندی با استفاده از توابع زیان مختلف (خطی، درجه دوم، هوبر) و همچنین رویکرد حاشیه نرم، امکان کنترل خطا و مقابله با نقاط پرت را فراهم می‌کند.

[۶] Support Vector Machines

۴ پیاده‌سازی و تحلیل نتایج

در این پژوهش از مجموعه‌داده *camg* موجود در بسته *geoR* استفاده شده است. این داده شامل اندازه‌گیری مقادیر کلسیم و منیزیم خاک در دو عمق ۰ تا ۲۰ و ۲۰ تا ۴۰ سانتی‌متر در ۱۷۸ موقعیت مکانی مختلف است. عناصر کلسیم و منیزیم از عناصر مغذی ضروری برای رشد گیاهان هستند و بررسی توزیع فضایی آن‌ها به درک بهتر حاصل‌خیزی خاک کمک می‌کند. در هر موقعیت، مختصات جغرافیایی (شرقی–غربی و شمالی–جنوبی)، ارتفاع زمین و شماره منطقه ثبت شده است. منطقه مورد مطالعه به سه بخش تقسیم می‌شود: منطقه اول که معمولاً در فصل بارندگی دچار سیلاب شده و دارای خاک طبیعی است، منطقه دوم که در گذشته کوددهی شده و عمدتاً شالیزار برنج است و منطقه سوم که به‌تازگی کوددهی شده و بیشتر برای آزمایش‌های کشاورزی به‌کار می‌رود. ساختار داده شامل یک *Data frame* با ۱۰ متغیر شامل مختصات مکانی[۷]، ارتفاع[۸]، شماره منطقه[۹]، مقادیر کلسیم، منیزیم و ظرفیت تبادل کاتیونی در دو عمق خاک ذکر شده است [۳]. این مجموعه داده برای تحلیل‌های زمین‌آماری با استفاده از نیم‌تغییرنگار[۱۰] و کریگینگ معمولی[۱۱] در بسته *geoR* به‌کار می‌رود و امکان بررسی وابستگی فضایی و پیش‌بینی مقادیر کلسیم و منیزیم در نقاط نمونه‌برداری‌نشده را فراهم می‌کند. پس از تقسیم داده‌ها به نسبت ۷۰ به ۳۰ برای آموزش و آزمون، عملکرد مدل‌های رگرسیون بردار پشتیبان (با چهار هسته مختلف) و رگرسیون ریج با استفاده از معیارهای R^2، *MSE*، *RMSE* و *MAPE* ارزیابی شد. پارامتر بهینه مدل ریج برای مناطق مختلف و متغیرهای ca_{020} و ca_{2040} با اعتبارسنجی متقابل تعیین و در جدول ۱ ارائه شده است.

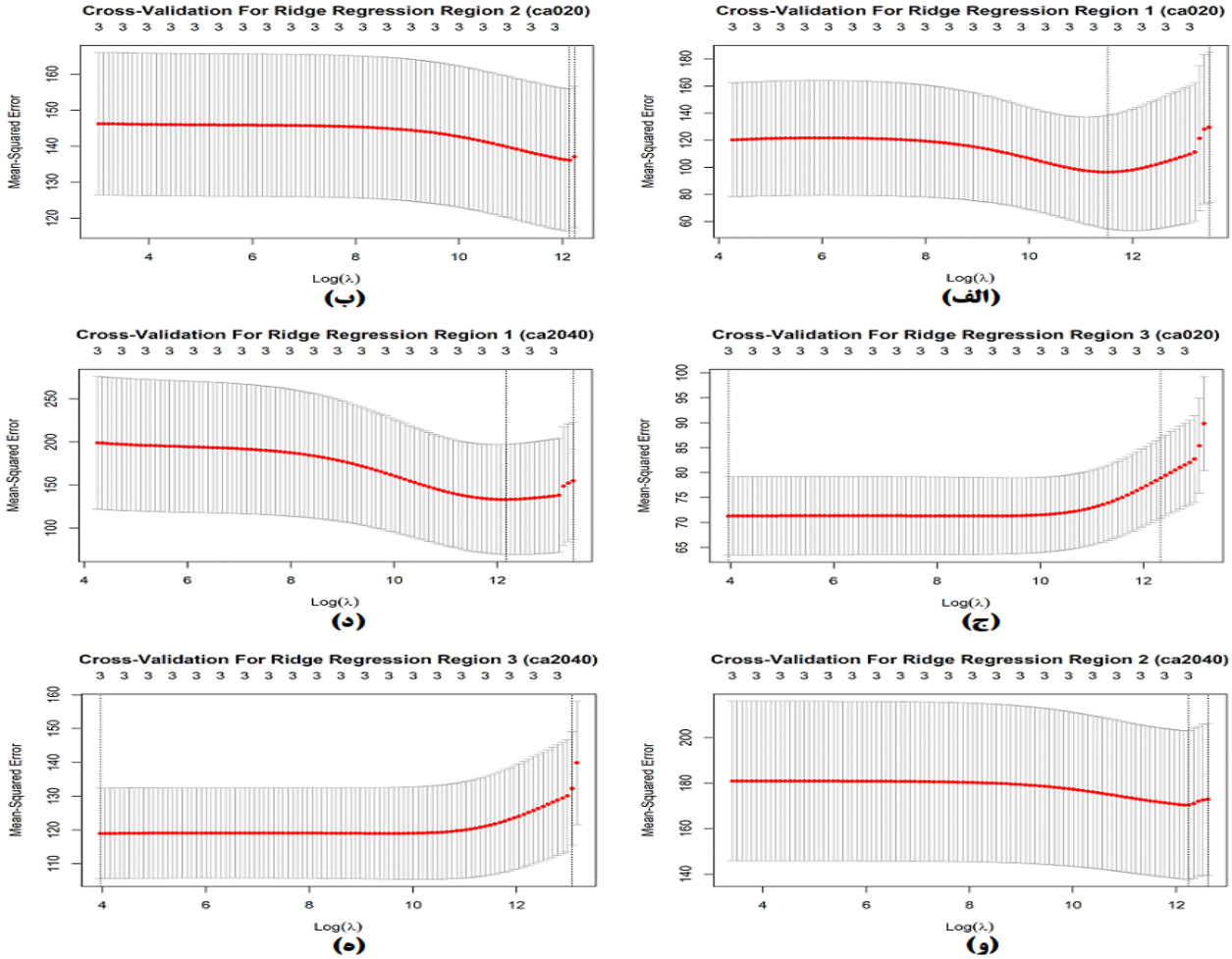
جدول ۱: مقادیر بهینه پارامتر ریج (λ)

منطقه	کلسیم در لایه‌های مختلف	
	ca_{020}	ca_{2040}
منطقه ۱	۱۰۱۱۶۹٫۲	۱۹۱۷۹۸٫۱
منطقه ۲	۱۸۷۵۲۳٫۸	۲۰۷۱۷۷٫۱
منطقه ۳	۵۲٫۶۰۲۶۱	۵۲٫۳۹۵۳۴

شکل ۱ نتایج اعتبارسنجی متقابل برای رگرسیون ریج را در سه منطقه و دو متغیر نشان می‌دهد. در هر نمودار، محور افقی لگاریتم پارامتر ریج λ و محور عمودی میانگین توان‌های دوم خطا (MSE) است. همان‌طور که منحنی‌های قرمز نشان می‌دهند، با افزایش λ ابتدا مقدار خطا کاهش یافته و پس از نقطه‌ای خاص دوباره افزایش می‌یابد که بیانگر وجود مقدار بهینه‌ای از λ برای هر منطقه است. به‌طور کلی، نتایج جدول ۱ و شکل ۱ نشان می‌دهند که تنظیم مناسب λ نقش مهمی در بهبود عملکرد مدل ریج دارد و حساسیت مدل نسبت به مقدار این پارامتر در مناطق مختلف متفاوت است.

بر اساس نتایج حاصل از ارزیابی مدل‌ها در سه منطقه مختلف موجود در جداول ۲ تا ۴، می‌توان دریافت که عملکرد بهینه مدل‌های رگرسیون به شدت تحت تأثیر شرایط خاص هر منطقه قرار دارد. در منطقه ۱، مدل بردار پشتیبان با هسته شعاعی با کسب ضریب تعیین ۰٫۸۶ برای متغیر ca_{020} به‌عنوان بهترین مدل شناسایی شد. در منطقه ۲، همین مدل با ضریب تعیین ۰٫۴۲ برای متغیر ca_{020} و مدل رگرسیون ریج با کمترین مقادیر خطا برای متغیر ca_{2040} مناسب‌ترین نتایج را ارائه دادند. در منطقه ۳، مدل‌های هسته شعاعی و چندجمله‌ای به‌ترتیب با ضرایب تعیین ۰٫۵۱ و ۰٫۳۷ برای متغیرهای ca_{020} و ca_{2040} بهترین عملکرد را از خود نشان دادند. در مقابل، مدل مبتنی بر هسته سیگموئید در تمامی مناطق و برای هر دو متغیر ضعیف‌ترین نتایج را ثبت کرد. این یافته‌ها به وضوح نشان می‌دهد که انتخاب مدل بهینه برای پیش‌بینی محتوای کلسیم خاک نیازمند در نظر گرفتن همزمان موقعیت جغرافیایی و عمق نمونه‌برداری است. نتایج این مطالعه نشان می‌دهد که تغییرپذیری مکانی خصوصیات خاک تأثیر قابل توجهی بر دقت مدل‌های پیش‌بینی دارد. به‌طور مشخص، مدل بردار پشتیبان با هسته شعاعی به دلیل توانایی بالا در مدل‌سازی روابط غیرخطی پیچیده بین متغیرهای مؤثر بر محتوای کلسیم خاک، در اکثر مناطق عملکرد مطلوبی از خود نشان داده است. همچنین مشاهده شد که با افزایش عمق نمونه‌برداری، دقت پیش‌بینی کلیه مدل‌ها کاهش می‌یابد که می‌تواند ناشی از پیچیدگی بیشتر روابط بین متغیرها در لایه‌های زیرین خاک باشد. تفاوت در عملکرد مدل‌ها در مناطق مختلف را می‌توان به عواملی همچون تغییرات در بافت خاک، شرایط اقلیمی، کاربری اراضی و مدیریت خاک در هر منطقه نسبت داد. این تنوع در نتایج بر اهمیت توسعه مدل‌های منطقه‌ای و عدم تعمیم یک مدل واحد به تمامی مناطق تأکید دارد. علاوه بر این، مقایسه معیارهای ارزیابی مختلف شامل R^2، MSE و $MAPE$ نشان داد که هر یک از این معیارها می‌توانند جنبه‌های متفاوتی از عملکرد مدل را مورد سنجش قرار دهند.

[7] East & North
[8] Elevation
[9] Region
[10] Semivariogram
[11] Ordinary Kriging



شکل ۱: نمودار اعتبارسنجی متقابل برای تعیین پارامتر بهینه λ در مناطق مختلف: الف) منطقه ۱ در $ca020$، ب) منطقه ۲ در $ca020$، ج) منطقه ۳ در $ca020$، د) منطقه ۱ در $ca2040$، و) منطقه ۲ در $ca2040$، ه) منطقه ۳ در $ca2040$.

جدول ۲: مقایسه مدل‌های رگرسیون بر اساس معیارهای ارزیابی در منطقه ۱

مدل رگرسیون	معیارها							
	ca۰۲۰ در منطقه ۱				ca۲۰۴۰ در منطقه ۱			
	R^2	MSE	RMSE	MAPE	R^2	MSE	RMSE	MAPE
بردار پشتیبان با هسته خطی	۰٫۸۲	۵۴٫۳۸	۷٫۳۷	۰٫۲۳	۰٫۰۰۲	۱۶۲٫۸۴	۱۲٫۷۶	۰٫۴۸
بردار پشتیبان با هسته چندجمله‌ای	۰٫۱۹	۷۲٫۲۰	۸٫۵۰	۰٫۲۶	۰٫۰۴	۱۰۶٫۱۵	۱۰٫۳۰	۰٫۳۹
بردار پشتیبان با هسته شعاعی	۰٫۸۶	۳۰٫۶۹	۵٫۵۴	۰٫۱۶	۰٫۲۲	۸۹٫۴۲	۹٫۴۶	۰٫۳۷
بردار پشتیبان با هسته سیگموئید	۰٫۷۷	۴۸٫۳۸	۶٫۹۵	۰٫۲۱	۰٫۰۰۰۷	۱۰۵٫۱۲	۱۰٫۲۵	۰٫۴۰
رگرسیون ریج	۰٫۴۶	۸۱٫۹۶	۹٫۰۵	۰٫۲۷	۰٫۰۰۹	۱۴۸٫۰۴	۱۲٫۱۷	۰٫۴۷

جدول ۳: مقایسه مدل‌های رگرسیون بر اساس معیارهای ارزیابی در منطقه ۲

مدل رگرسیون	معیارها							
	ca۰۲۰ در منطقه ۲				ca۲۰۴۰ در منطقه ۲			
	R^2	MSE	RMSE	MAPE	R^2	MSE	RMSE	MAPE
بردار پشتیبان با هسته خطی	۰/۱۷	۳۷/۴۳	۶/۱۲	۰/۱۳	۰/۰۴	۱۳۶/۲۵	۱۱/۶۷	۰/۳۱
بردار پشتیبان با هسته چندجمله‌ای	۰/۰۰۰۲	۴۸/۹۶	۷/۰۰	۰/۱۴	۰/۰۷	۱۷۹/۲۷	۱۳/۳۹	۰/۳۶
بردار پشتیبان با هسته شعاعی	۰/۴۲	۳۵/۶۶	۵/۹۷	۰/۱۳	۰/۱۱	۱۴۴/۲۸	۱۲/۰۱	۰/۳۳
بردار پشتیبان با هسته سیگموئید	۰/۰۳	۱۰۶/۹۴	۱۰/۳۴	۰/۲۰	۰/۰۰۴	۱۱۳/۸۴	۱۰/۶۷	۰/۲۶
رگرسیون ریج	۰/۰۲	۴۴/۶۳	۶/۶۸	۰/۱۳	۰/۰۲	۱۰۳/۹۳	۱۰/۱۹	۰/۳۰

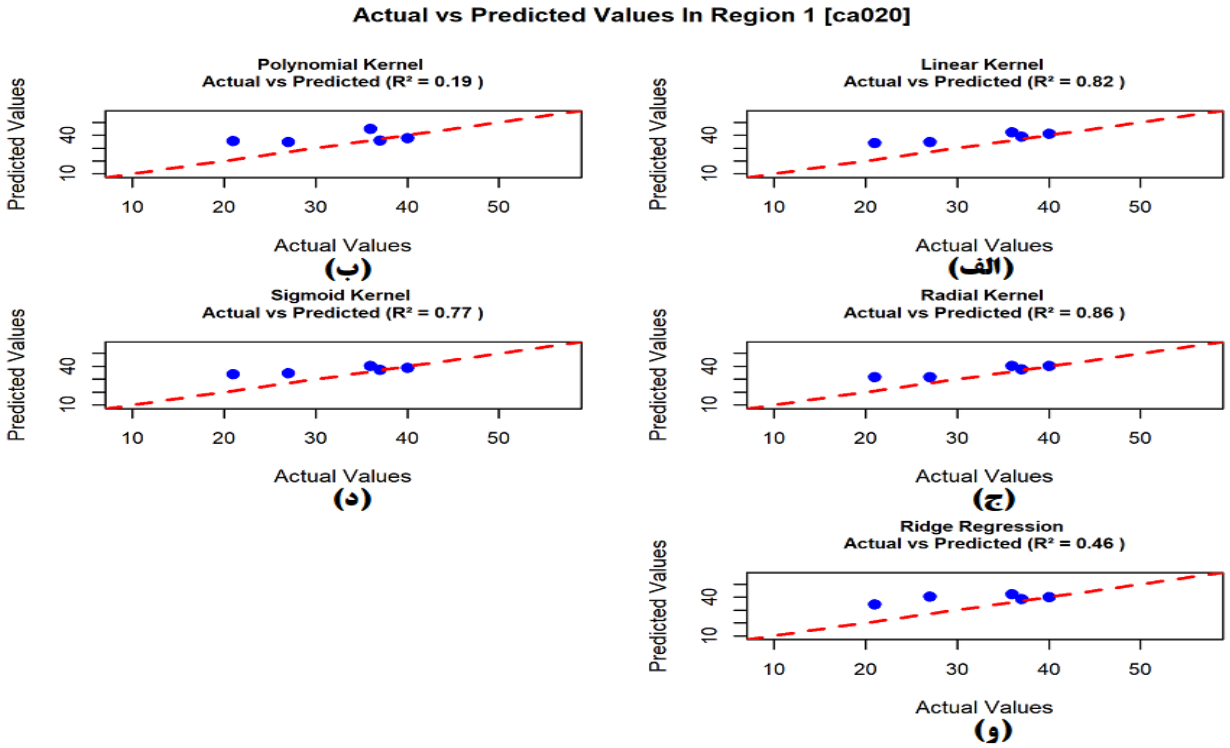
جدول ۴: مقایسه مدل‌های رگرسیون بر اساس معیارهای ارزیابی در منطقه ۳

مدل رگرسیون	معیارها							
	ca۰۲۰ در منطقه ۳				ca۲۰۴۰ در منطقه ۳			
	R^2	MSE	RMSE	MAPE	R^2	MSE	RMSE	MAPE
بردار پشتیبان با هسته خطی	۰/۲۸	۵۹/۴۹	۷/۷۱	۰/۱۱	۰/۲۳	۱۱۶/۴۸	۱۰/۷۹	۰/۱۸
بردار پشتیبان با هسته چندجمله‌ای	۰/۴۶	۴۴/۴۷	۶/۶۷	۰/۱۰	۰/۳۷	۹۵/۵۸	۹/۷۸	۰/۱۷
بردار پشتیبان با هسته شعاعی	۰/۵۱	۴۴/۳۶	۶/۶۶	۰/۰۹	۰/۳۶	۹۹/۷۶	۹/۹۹	۰/۱۷۲
بردار پشتیبان با هسته سیگموئید	۰/۰۹	۵۵۷/۷۳	۲۳/۶۲	۰/۳۲	۰/۰۲	۷۵۵/۴۵	۲۷/۴۸	۰/۴۳
رگرسیون ریج	۰/۲۰	۶۵/۸۴	۸/۱۱	۰/۱۲	۰/۲۰	۱۱۹/۴۹	۱۰/۹۳	۰/۱۹

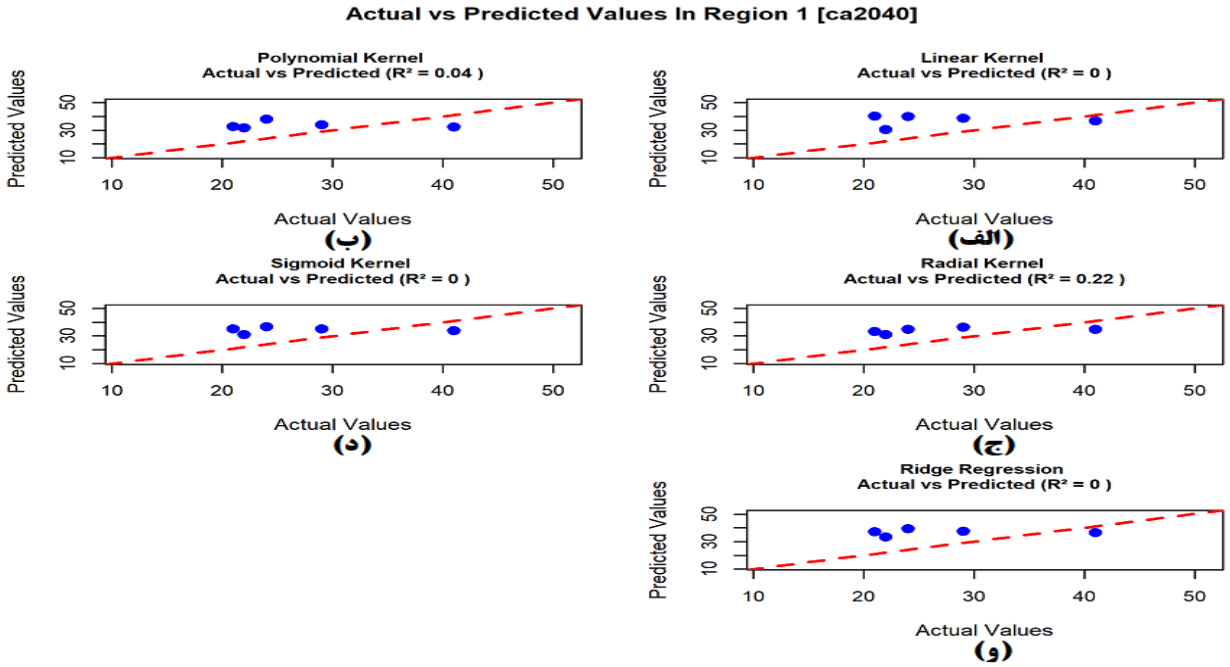
برای ارزیابی دقیق عملکرد مدل‌های رگرسیون در پیش‌بینی محتوای کلسیم خاک، نمودارهای مقادیر واقعی در مقابل مقادیر پیش‌بینی شده در سه منطقه و برای دو عمق نمونه‌برداری (ca۰۲۰ و ca۲۰۴۰) در مجموعه نمودارهای موجود در شکل‌های ۲ تا ۷ مورد تحلیل قرار گرفت. در مجموعه نمودارهای موجود در شکل ۲ مربوط به منطقه ۱ و برای متغیر ca۰۲۰، مدل بردار پشتیبان با هسته شعاعی با ضریب تعیین ۰/۸۶ بهترین عملکرد را نشان داد که بیانگر تطابق بسیار مطلوب بین مقادیر واقعی و پیش‌بینی شده و توزیع فشرده نقاط حول خط $Y = X$ است. پس از آن، مدل‌های با هسته خطی (۰/۸۲) و سیگموئید (۰/۷۷) در رده‌های بعدی قرار گرفتند. در مقابل، مدل رگرسیون ریج (۰/۴۶) و مدل با هسته چندجمله‌ای (۰/۱۹) عملکرد ضعیف‌تری داشتند. برای متغیر ca۲۰۴۰ در همین منطقه (شکل ۳)، اگرچه مقادیر ضریب تعیین به‌طور کلی کاهش یافت، مدل هسته شعاعی (۰/۲۲) و چندجمله‌ای (۰/۰۴) بهتر از سایر مدل‌ها عمل کردند، درحالی که مدل‌های خطی، سیگموئید و ریج تقریباً فاقد توان پیش‌بینی معنادار بودند.

در منطقه ۲ و برای متغیر ca۰۲۰ (شکل ۴)، مدل با هسته شعاعی با ضریب تعیین ۰/۴۲ بهترین عملکرد را نشان داد و پس از آن مدل با هسته خطی با مقدار ۰/۱۷ در رتبه دوم قرار گرفت. سایر مدل‌ها از جمله سیگموئید (۰/۰۳)، ریج (۰/۰۲) و چندجمله‌ای (۰) عملکرد قابل توجهی نداشتند. همچنین برای متغیر ca۲۰۴۰ در همین منطقه (شکل ۵)، مدل‌های خطی (۰/۰۴) و چندجمله‌ای (۰/۰۷) نسبت به سایر مدل‌ها نتایج بهتری ارائه دادند، هرچند دقت کلی پایین ارزیابی شد.

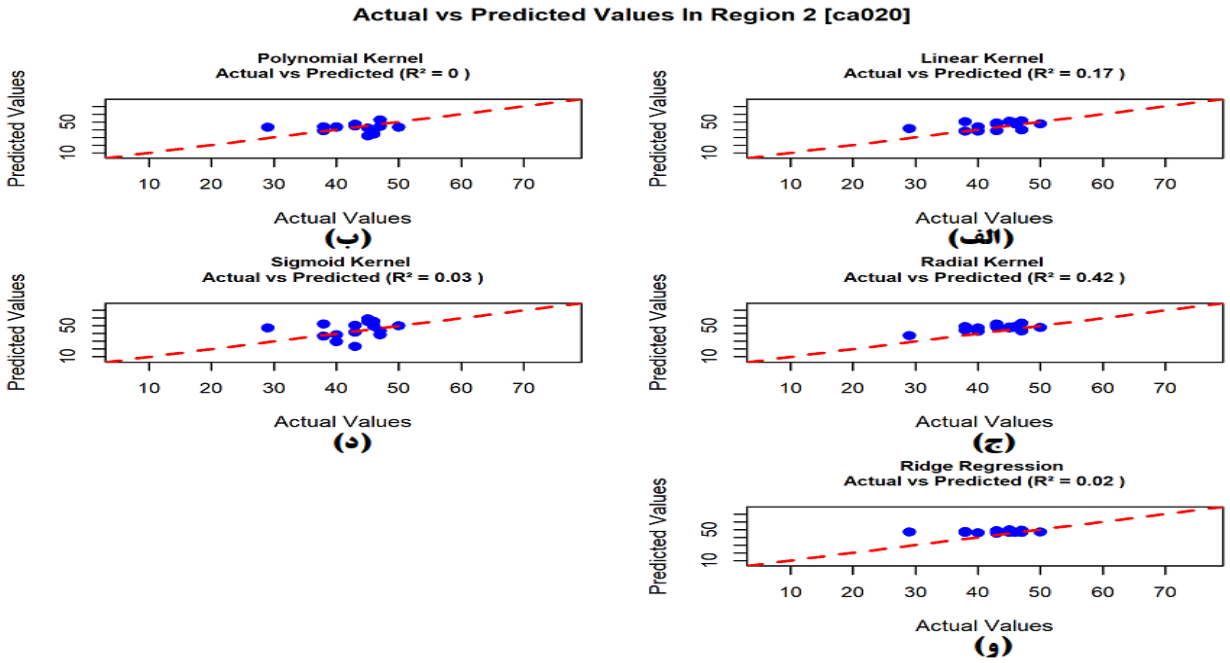
در منطقه ۳ برای متغیر ca۰۲۰ (شکل ۶)، مدل با هسته شعاعی با ضریب تعیین ۰/۵۱ بهترین عملکرد را داشت و پس از آن مدل‌های چندجمله‌ای (۰/۴۶) و خطی (۰/۲۸) قرار گرفتند، مدل‌های ریج (۰/۲) و سیگموئید (۰/۰۹) عملکرد ضعیفی داشتند. برای متغیر ca۲۰۴۰ (شکل ۷) نیز مدل چندجمله‌ای (۰/۳۷) برتر بود و مدل‌های شعاعی (۰/۳۶) و خطی (۰/۲۳) در رتبه‌های بعدی قرار گرفتند، در حالی که ریج (۰/۲) و سیگموئید (۰/۰۲) نتایج مطلوبی ارائه نکردند.



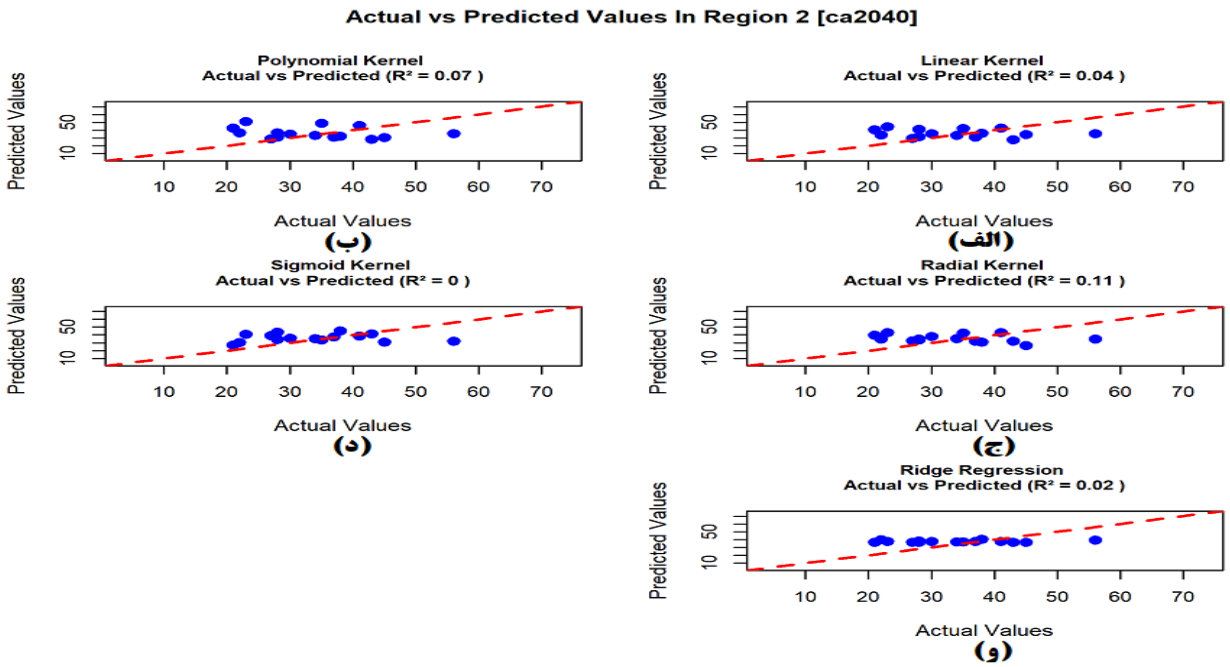
شکل ۲: نمودار مقادیر واقعی در برابر مقادیر پیش‌بینی شده برای چهار هسته بردار پشتیبان و رگرسیون ریج در منطقه ۱ در $ca020$: الف) هسته خطی، ب) هسته چندجمله‌ای، ج) هسته شعاعی، د) هسته سیگموئید، و) رگرسیون ریج.



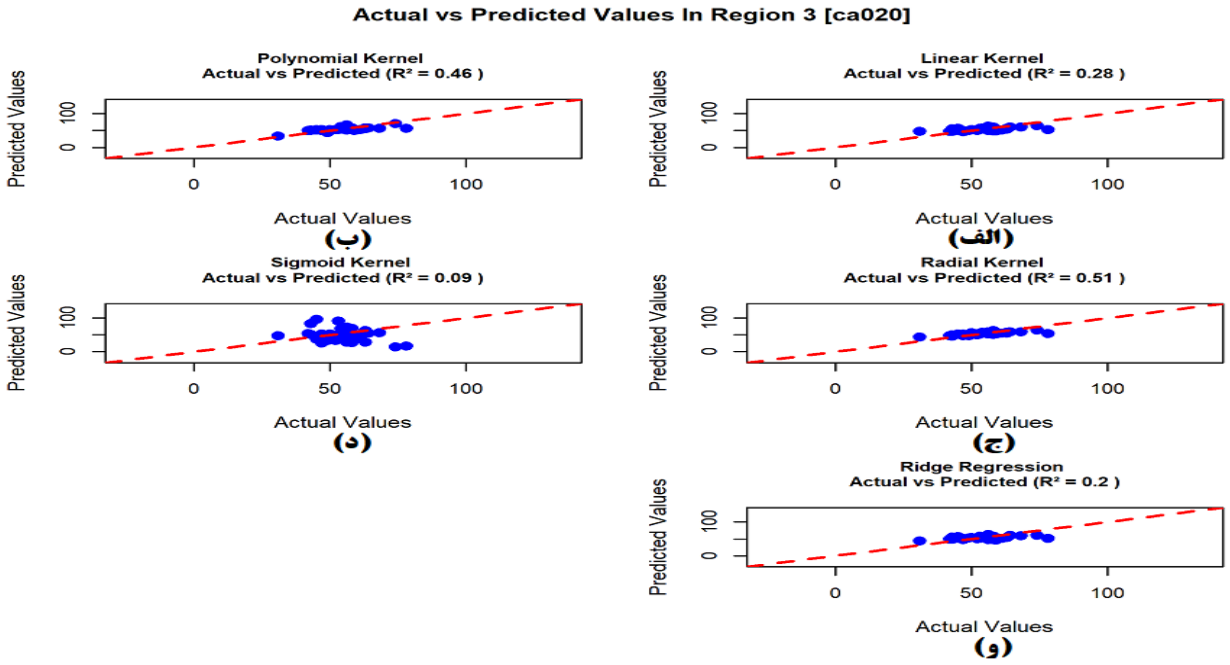
شکل ۳: نمودار مقادیر واقعی در برابر مقادیر پیش‌بینی شده برای چهار هسته بردار پشتیبان و رگرسیون ریج در منطقه ۱ در $ca2040$: الف) هسته خطی، ب) هسته چندجمله‌ای، ج) هسته شعاعی، د) هسته سیگموئید، و) رگرسیون ریج.



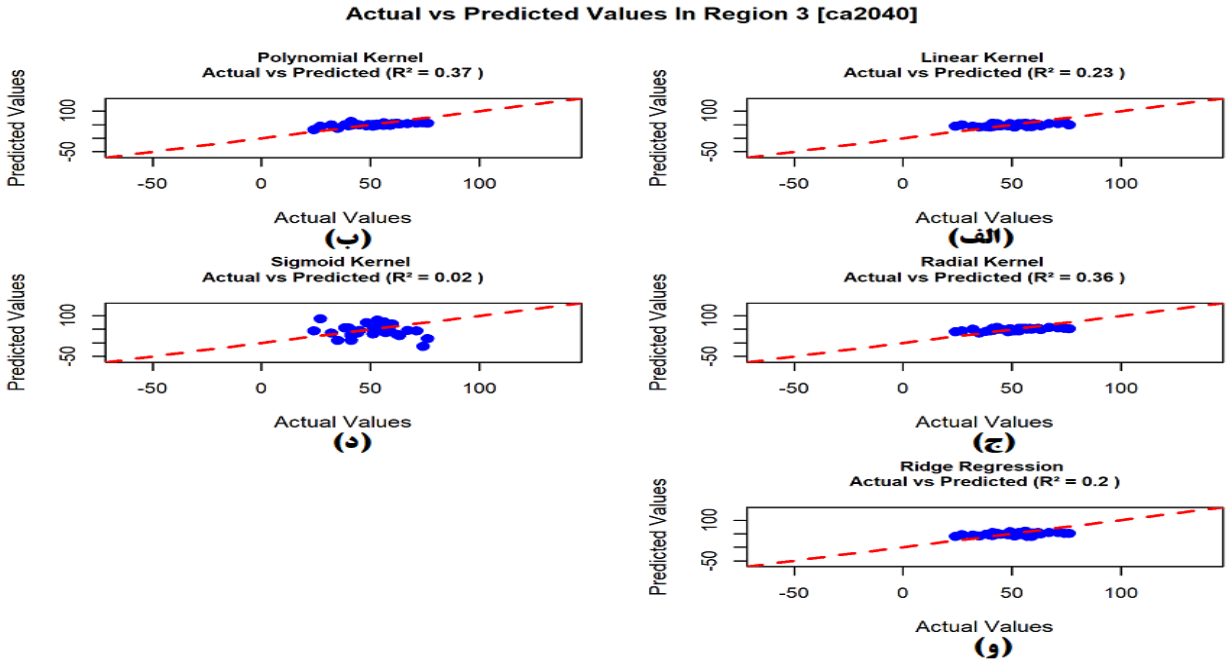
شکل ۴: نمودار مقادیر واقعی در برابر مقادیر پیش‌بینی شده برای چهار هسته بردار پشتیبان و رگرسیون ریج در منطقه ۲ در $ca۰۲۰$: الف) هسته خطی، ب) هسته چندجمله‌ای، ج) هسته شعاعی، د) هسته سیگموئید، و) رگرسیون ریج.



شکل ۵: نمودار مقادیر واقعی در برابر مقادیر پیش‌بینی شده برای چهار هسته بردار پشتیبان و رگرسیون ریج در منطقه ۲ در $ca۲۰۴۰$: الف) هسته خطی، ب) هسته چندجمله‌ای، ج) هسته شعاعی، د) هسته سیگموئید، و) رگرسیون ریج.



شکل ۶: نمودار مقادیر واقعی در برابر مقادیر پیش‌بینی شده برای چهار هسته بردار پشتیبان و رگرسیون ریج در منطقه ۳ در $ca020$: الف) هسته خطی، ب) هسته چندجمله‌ای، ج) هسته شعاعی، د) هسته سیگموئید، و) رگرسیون ریج.



شکل ۷: نمودار مقادیر واقعی در برابر مقادیر پیش‌بینی شده برای چهار هسته بردار پشتیبان و رگرسیون ریج در منطقه ۳ در $ca2040$: الف) هسته خطی، ب) هسته چندجمله‌ای، ج) هسته شعاعی، د) هسته سیگموئید، و) رگرسیون ریج.

بر اساس جدول پارامترهای بهینه مدل SVR (جدول ۵)، می‌توان مشاهده کرد که هر منطقه دارای مشخصات منحصر به فردی است که منجر به تنظیمات پارامتری متفاوتی شده است. در منطقه ۱، مدل برای لایه سطحی خاک (۰-۲۰ سانتی‌متر) با مقادیر $(C, \varepsilon) = (۱, ۰/۶۶)$ تنظیم شده که نشان دهنده یک مدل نسبتاً ساده با تحمل خطای متوسط

جدول ۵: پارامترهای بهینه در مدل SVR

منطقه	پارامترهای بهینه			
	ca۰۲۰		ca۲۰۴۰	
	C	ε	C	ε
منطقه ۱	۱/۰	۰/۶۶	۱۰/۰	۰/۱۶
منطقه ۲	۰/۰۱	۰/۹۸	۰/۰۱	۰/۸۲
منطقه ۳	۱/۰	۰/۰۲	۱/۰	۰/۰

است، اما در لایه زیرین خاک (۲۰-۴۰ سانتی‌متر) با مقادیر $(C, \varepsilon) =$ $(۱۰, ۰/۱۶)$ به مدلی با پیچیدگی بیشتر و حساسیت بالاتر به خطا تغییر می‌یابد. این امر حاکی از آن است که داده‌های مربوط به کلسیم در لایه زیرین خاک این منطقه نیاز به دقت پیش‌بینی بیشتری دارند. منطقه ۲ در هر دو لایه خاک دارای مقدار C بسیار پایین (۰/۰۱) است که نشان دهنده ساده‌سازی شدید مدل و اعمال جریمه سنگین برای خطاها می‌باشد. همچنین مقادیر بالای ε (۰/۹۸ و ۰/۸۲) بیانگر تحمل خطای گسترده است که احتمالاً به دلیل تغییرپذیری زیاد مقدار کلسیم یا وجود اختلال در داده‌های این منطقه می‌باشد. در منطقه ۳، مدل برای لایه سطحی خاک با مقادیر $(C, \varepsilon) = (۱, ۰/۰۲)$ تنظیم شده که حساسیت بسیار بالایی به خطاهای کوچک دارد، و در لایه زیرین این حساسیت تا حد ε صفر افزایش می‌یابد که نشان دهنده نیاز به دقت کامل در پیش‌بینی‌های مربوط به کلسیم است. این تنظیمات نشان می‌دهد که مدل SVR توانایی بالایی در تطبیق با شرایط مختلف لایه‌های خاک دارد و می‌تواند برای هر منطقه و عمق خاک به‌صورت بهینه تنظیم شود.

۵ نتیجه‌گیری

رشد حجم و پیچیدگی داده‌های فضایی، ضرورت بهره‌گیری از روش‌های نوین مدل‌سازی را آشکار ساخته است. در این مقاله، عملکرد رگرسیون بردار پشتیبان با هسته‌های مختلف و رگرسیون ریج برای پیش‌بینی محتوای کلسیم خاک در سه منطقه و دو لایه عمقی بررسی شد. نتایج نشان دادند که کارایی مدل‌ها به شدت به شرایط مکانی و عمق نمونه‌برداری وابسته است. مدل SVR با هسته شعاعی در اغلب موارد بهترین عملکرد را ارائه داد، در حالی‌که هسته سیگموئید ضعیف‌ترین نتایج را ثبت کرد. رگرسیون ریج نیز در برخی شرایط (مانند منطقه ۲ و متغیر $ca۲۰۴۰$) توانست دقت بالاتری نسبت به مدل‌های غیرخطی داشته باشد. تحلیل پارامترهای بهینه نیز نشان داد که هر منطقه نیازمند تنظیمات متفاوتی است. در منطقه ۱ مدل برای لایه زیرین با C بالا و ε پایین سخت‌گیرانه‌تر تنظیم شد، در حالی‌که منطقه ۲ به دلیل اختلال و تغییرپذیری زیاد داده‌ها با C پایین و ε بالا به مدلی ساده‌تر گرایش یافت.

در منطقه ۳، حساسیت شدید به خطاهای کوچک در هر دو لایه مشاهده شد که بیانگر نیاز به دقت بالا در پیش‌بینی است. به‌طور کلی، نتایج بیانگر آن است که انتخاب مدل بهینه و تنظیم پارامترها باید متناسب با ویژگی‌های مکانی و عمقی داده‌ها صورت گیرد. استفاده از روش‌های مبتنی بر یادگیری ماشین مانند SVR این امکان را فراهم می‌کند که بدون نیاز به فرضیات سخت‌گیرانه، ساختارهای فضایی پیچیده مدل‌سازی شوند و پیش‌بینی‌های دقیق‌تری به‌دست آید. این یافته‌ها می‌تواند در مدیریت منابع خاک و بهینه‌سازی تصمیم‌گیری‌های کشاورزی و محیط زیستی مورد استفاده قرار گیرد.

تقدیر و تشکر

نویسندگان مقاله کمال قدردانی و تشکر را از پیشنهادات ارزنده داوران، سردبیر و ویراستار محترم مجله که باعث ارائه بهتر و افزایش سطح کیفی مقاله شده است، دارند.

مراجع

[۱] محمدزاده، م.، (۱۳۹۸)، آمار فضایی و کاربردهای آن، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،

[2] Awad, M., and Khanna, R. (2015), *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, First Edition*, Apress Media, New York.

[3] Capeche, C.L., Macedo, J.R., Manzatto, H.R.H., and Silva, E.F. (1997), *Caracterização pedológica da fazenda Angra -*

PESAGRO/RIO - Estação experimental de Campos (RJ), In: Congresso BRASILEIRO de Ciência do Solo, 26., Rio de Janeiro, Embrapa/SBCS.

[4] Gutierrez, D. D. (2015), *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R, First Edition*, Technics Publications, NJ.

[5] Izenman, A.J. (2008), *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer Science and Business Media, New York.

[6] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R*, Springer Science and Business Media, New York.

[7] Lauer, F., and Bloch, G. (2008), Incorporating Prior Knowledge in Support Vector Regression, *Machine Learning*, **70**, 89–118.

[8] Roozbeh, M. (2025), Optimal ridge estimation in the restricted logistic semiparametric regression models using generalized cross-validation, *Journal of Applied Statistics*, 1–20.

[9] Roozbeh, M., Rouhi, A., Mohamed, N.A., and Jahadi, F. (2023), Generalized support vector regression and symmetry functional regression approaches to model the high-dimensional data, *Symmetry*, **15**, 1262.

Andisheye Amari, 2024, Vol. 29, No. 2, pp. 1 – 12
https://doi.org/10.22034/aa.2026.733260

Analysis of Hydrological Variables and Soil Properties Using Spatial Ridge and Support Vector Regression Methods

Mahdi Roozbeh[1]*, **Arash Ameri**[2]

[1] *Department of Statistics, Faculty of Mathematics, Statistics and Computer Science, Semnan University, Semnan, Iran*

[2] *Department of Statistics, Faculty of Mathematics, Statistics and Computer Science, Semnan University, Semnan, Iran*

Received: 2025/11/12

Accepted: 2026/01/03

Abstract

Traditional statistical methods have faced serious challenges due to the expansion of spatial data with complex spatiotemporal structure. These data require specialized methods due to spatial autocorrelation, variance heterogeneity, and complex geographical dependencies. In this study, support vector regression is introduced as a novel approach for analyzing and modeling the complex spatial structure of geostatistical data related to soil calcium and magnesium contents. This analysis is performed based on different geographical coordinates (east–west and north–south), at two depths of 0–20 cm and 20–40 cm, and across three distinct geographical regions. The support vector regression method, with its capability to model complex nonlinear relationships while preserving the spatial structure of the data, allows for more accurate and realistic prediction of the nutrient elements' distribution in soil. This approach, utilizing kernel functions, enables the analysis of high-dimensional feature spaces and structural complexities, proving its effectiveness against noise and outliers. To accurately measure the efficiency of support vector regression, its performance is compared against ridge regression.

Keywords: Generalized Cross-validation, Ridge Regression, Support Vector Machines, Support Vector Regression

Mathematics Subject Classification: $62J05; 62H25$.

*Corresponding author, *mahdi.roozbeh@semnan.ac.ir*