



Trend Removal in Spatial Statistics Using Support Vector Regression

Haddadi, S. , Etminan, J. 

Department of Statistics, University of Birjand, Birjand, Iran.

Corresponding author: J. Etminan, jetminan@birjand.ac.ir

Received: 24/12/2022 Revised: 22/6/2023 Accepted and Published Online: 26/6/2023.

Introduction

Stationarity is one of the conditions that must be investigated in the exploratory analysis of spatial data. The existence of a trend in the data is a reason for non-stationarity. As a result, the estimator of the variogram is biased and the prediction will not be accurate enough. Until now, various methods have been proposed to remove the trend, including simple linear combination, median polishing, first order difference, etc. In economics, simple linear regression and first differences are the two most popular trend removal methods. Support Vector Machine (SVM) is a supervised learning algorithm that can be used for both classification and regression problems. Support Vector Regression (SVR) is a type of SVM used for regression analysis. The goal of SVR is to find a function that approximates the relationship between the input variables and a continuous target variable while minimizing the prediction error. SVR can be an effective tool in spatial data analysis. In recent years, the use of SVR in spatial data analysis has become widespread. The purpose of this study is to investigate the detrending of data using the SVR method and its effect on prediction.

Material and Methods

First, the trend in spatial data is tried to be modeled using the SVR method. If the relationship between the input variables and the target variable is complicated so that a linear model is not suitable for fitting the data, then a nonlinear mapping can be used to transfer the input space to a space with higher dimensions, then a linear model can be fitted to the data in this space with the help of the kernel function. Residuals are obtained by removing the trend from the data. Then, a suitable variogram model is fitted to the

residuals, and prediction is made based on the universal kriging method. Finally, the trend is estimated by the SVR method and added to the prediction value. In the following, on a real data set, the results of the proposed method are compared with spline and universal kriging prediction methods through cross-validation. The minimization of the root-mean-square error (RMSE) is considered the criterion for choosing the appropriate forecasting method. The `geoR` and `gstat` packages in R software were used for analysis.

Results and Discussion

Kent and Mohammadzadeh (2000) compared the smooth spline as a detrending method and its prediction results on the calcium data (Oliveira, 2003) set with the prediction results of the universal kriging method. For this data set, the trend in the data was modeled by SVR and through three kernels: linear, polynomial and Gaussian. The results of this study showed that the proposed method has a higher accuracy than the two universal kriging and spline methods. Among the applied kernels, the results of the Gaussian kernel were better. The results of this research show that SVR can be a suitable tool in spatial data analysis.

Conclusion

In trend data analysis, simple functions such as polynomials are usually used to model the trend. In SVR, in addition to linear models, it is possible to use different types of non-linear models through kernel functions. By fitting different models to the data and comparing them, you can easily choose the most suitable model. Therefore, trend modeling using SVR is an efficient method that can improve prediction results.

Keywords: Variogram, Kriging, Spatial prediction, Support vector regression, Kernel function.

Mathematics Subject Classification (2010): 62J02, 62M20.



روندزدایی در آمار فضایی با استفاده از رگرسیون بردار پشتیبان

ساره حدادی و جواد اطمینان

گروه آمار، دانشگاه بیرجند

نویسنده مسئول: جواد اطمینان، jetminan@birjand.ac.ir

تاریخ دریافت: ۱۴۰۱/۱۰/۳ تاریخ بازنگری: ۱۴۰۲/۴/۱ تاریخ پذیرش و انتشار: ۱۴۰۲/۴/۵

چکیده: مدل‌بندی و برآوردی کارا از تابع روند در برآورد تغییرنگار و پیش‌گویی داده‌های فضایی از اهمیت بالایی برخوردار است. در این مقاله برای مدل‌بندی تابع روند از روش رگرسیون بردار پشتیبان استفاده شده است. سپس داده‌ها روندزدایی شده و برآورد تغییرنگار و پیش‌گویی انجام می‌شود. بر روی یک مجموعه داده واقعی، نتایج پیش‌گویی حاصل از روش پیشنهادی با روش پیش‌گویی اسپلاین و کریگیدن از طریق اعتبارسنجی متقابل مقایسه شده است. کمینه بودن ریشه دوم میانگین توان‌های دوم خطا معیار انتخاب روش مناسب پیش‌گویی است. نتایج پیش‌گویی برای چند موقعیت با مقادیر معلوم که بنا به دلایلی از مجموعه داده‌ها کنار گذاشته شده‌اند و برای موقعیت‌های جدید به‌دست آمده‌اند. نتایج بیانگر دقت بالای پیش‌گویی با روش پیشنهادی نسبت به کریگیدن و اسپلاین است.

واژه‌های کلیدی: تغییرنگار، کریگیدن، پیش‌گویی فضایی، رگرسیون بردار پشتیبان، تابع هسته. کد موضوع‌بندی ریاضی (۲۰۱۰): 62J02، 62M20.

۱ مقدمه

داده‌های فضایی مشاهداتی هستند که وابستگی آن‌ها ناشی از موقعیت‌شان در فضای مورد مطالعه است و این وابستگی معمولاً تابعی از فاصله مشاهدات از یکدیگر است (کرس، ۱۹۹۳). معمولاً یک میدان تصادفی مانند $\{Z(s) : s \in R^d, d \geq 1\}$ به‌عنوان مدل آماری برای داده‌های فضایی در نظر گرفته می‌شود. این میدان تصادفی



را می‌توان به صورت $Z(s) = \mu(s) + \delta(s)$ تجزیه کرد، که در آن $\mu(\cdot)$ نشان‌دهنده روند و $\delta(\cdot)$ فرایند خطا است. اگر میانگین میدان تصادفی ثابت (به موقعیت s بستگی نداشته باشد) و کواریانس فقط تابعی از اختلاف موقعیت‌ها باشد، میدان فضایی مانای مرتبه دوم است که تحقیق آن بسیار دشوار است (ویرا، ۲۰۰۰). مترن (۱۹۶۳) شرط مانای ذاتی را پیشنهاد کرد. در این نوع مانایی میانگین میدان تصادفی ثابت و واریانس اختلاف هر دو متغیر تصادفی در میدان فقط تابعی از اختلاف موقعیت‌های آن‌ها است.

وجود روند در بین مشاهدات باعث نامانایی میدان تصادفی می‌شود (ویرا و همکاران، ۱۹۸۳). فوئنتس (۲۰۰۶)، جان و جنتون (۲۰۱۲) و پریستلی و سوبا راتو (۱۹۶۹) روش‌های شناسایی مانایی در داده‌های فضایی را مورد مطالعه قرار داده‌اند. در صورت وجود نامانایی در میانگین، برای برازش تغییرنگار و اجرای پیش‌گویی باید روندزدایی از داده‌ها انجام گیرد. یک روش متداول روندزدایی در نظر گرفتن روند به صورت یک ترکیب خطی $\mu(s) = \sum_{j=0}^p \beta_j \phi_j(s)$ است که در آن $(\phi_0(s), \dots, \phi_p(s))$ بردار تابع‌هایی معلوم بر حسب موقعیت s هستند. بردار ضرایب نامعلوم $\beta = (\beta_0, \dots, \beta_p)$ به یکی از روش‌های کمترین توان‌های دوم برآورد می‌شود (محمدزاده، ۱۳۹۸). روش میانه‌پرداخت روشی ساده برای روندزدایی است، که توسط کرسی (۱۹۹۳) استفاده شده است که به وسیله بعضی نرم‌افزارهای آماری مانند مینی‌تب قابل انجام است. در علم اقتصاد روندزدایی به دو روش رگرسیون خطی ساده و تفاضلی مرتبه اول بسیار پر طرفدار است (سارجنت، ۱۹۶۸). نحوه کاهش سطح روند متناسب با روش کمترین توان‌های دوم توسط ویرا و همکاران (۲۰۱۰) مطرح شد. ماشین بردار پشتیبان، روشی یادگیری با نظارت است که از آن برای طبقه‌بندی و تحلیل رگرسیونی استفاده می‌شود. رگرسیون بردار پشتیبان می‌تواند ابزاری کارآمد در تحلیل داده‌های فضایی به ویژه پیش‌گویی باشد. سیلوا و بویورت (۲۰۲۰) و هوانگ و شیم (۲۰۱۷) به ترتیب با به کارگیری رگرسیون بردار پشتیبان با وزن محلی و رگرسیون بردار پشتیبان کمترین توان‌های دوم موزون جغرافیایی روشی جدید برای پیش‌گویی داده‌های فضایی پیشنهاد کردند. هوانگ و همکاران (۲۰۱۲) با به کارگیری رگرسیون بردار پشتیبان کمترین توان‌های دوم در برازش تغییرنگار به بهبود درونیایی کریگیدن در بازیابی داده‌های گمشده اقیانوسی پرداختند. هان و همکاران (۲۰۱۶) برای برازش تغییرنگار، مدل‌های جدیدی را به کمک رگرسیون بردار پشتیبان و برنامه‌ریزی ژنتیکی چند ژنی پیشنهاد دادند. توسلی و همکاران (۱۳۹۸) دقت روش رگرسیون بردار پشتیبان با روش کمترین توان‌های دوم را در برآورد تغییرنگار و پیش‌گویی با کریگیدن مقایسه کردند.

مدل کردن تغییرنگار به روش رگرسیون بردار پشتیبان و استفاده از آن در پیش‌گویی و مدل کردن داده‌ها به روش رگرسیون بردار پشتیبان و انجام پیش‌گویی براساس آن دو رویکرد عمده در تحقیقات صورت گرفته در استفاده از رگرسیون بردار پشتیبان در تحلیل داده‌های فضایی است. در این مطالعه ابتدا روند در داده‌های فضایی با استفاده از روش رگرسیون بردار پشتیبان مدل می‌شود و با حذف آن از داده‌ها، مانده‌ها (داده‌های روندزوده) به دست می‌آید. سپس یک مدل تغییرنگار مناسب به مانده‌ها برازش داده شده و پیش‌گویی به روش کریگیدن عادی بر اساس آن انجام می‌شود. در نهایت برآورد روند به روش رگرسیون بردار پشتیبان به مقدار پیش‌گویی اضافه می‌گردد.

در بخش ۲ ساختار همبستگی داده‌های فضایی و مفهوم پیش‌گویی آورده شده‌اند. در بخش ۳ رگرسیون بردار پشتیبان خطی و ناخطی معرفی و بررسی می‌شود. در بخش ۴ ایده روندزدایی داده‌های فضایی به کمک رگرسیون

بردار پشتیبان بحث شده است. در بخش ۵ برای ارزیابی روش پیشنهادی با سایر روش‌ها، یک مجموعه داده واقعی تحلیل شده و نتایج پیش‌گویی به‌دست آمده از روش‌های مختلف از طریق ریشه دوم میانگین توان دوم خطای روش اعتبار سنجی متقابل^۱ مقایسه می‌شوند. در انتها بحث و نتیجه‌گیری آورده شده است.

۲ ساختار همبستگی در داده‌های فضایی و پیش‌گویی

برای بیان ساختار همبستگی در داده‌ها از تابع تغییرنگار استفاده می‌شود که در یک میدان تصادفی، تحت شرایط مانای ذاتی در دو موقعیت s و $s+h$ به‌صورت $\gamma(h) = \text{Var}(Z(s+h) - Z(s))$ تعریف می‌شود (کرسبی، ۱۹۹۳).
 مترن (۱۹۶۲) تغییرنگار تجربی برای برآورد کردن تغییرنگار را به‌صورت

$$\hat{\gamma}(h) = \frac{1}{|N_h|} \sum_{N_h} ((Z(s_i) - Z(s_j))^2 \quad (۱)$$

پیشنهاد کرد، که در آن $N(h) = \{(s_i, s_j) | d(s_i, s_j) = h\}$ و $|N_h|$ تعداد اعضای این مجموعه است. نااریبی و سازگاری از ویژگی‌های این برآوردگر است (میراندا و میراندا، ۲۰۱۱). در جدول ۱ برخی از پرکاربردترین توابع نیم‌تغییرنگار آورده شده است. برای تعیین پارامترهای مدل منتخب، زوج نقاط $(h, \hat{\gamma}(h))$ مانند زوج داده رگرسیونی در نظر گرفته شده و برآورد پارامترهای مدل به‌گونه‌ای به‌دست می‌آید که کمترین اختلاف بین تغییرنگار

جدول ۱. توابع نیم‌تغییرنگار گاوسی، نمایی، کروی و مترن.		
مدل	تابع نیم‌تغییرنگار $\gamma(h)$	توضیحات
گاوسی	$c_0 + c(1 - e^{-\frac{\ h\ ^2}{a^2}})$	$h \in R^d, d \geq 1$
نمایی	$c_0 + c(1 - e^{-\frac{\ h\ }{a}})$	$h \in R^d, d \geq 1$
کروی	$\begin{cases} c_0 + c(\frac{3}{4}\frac{\ h\ }{a} - \frac{1}{4}\frac{\ h\ ^3}{a^3}) & 0 < \ h\ \leq a \\ c_0 + c & \ h\ > a \end{cases}$	$h \in R^d, d = 1, 2, 3$
مترن	$c_0 + c(1 - \frac{1}{\sqrt{\nu-1}\Gamma(\nu)}(\frac{\ h\ }{a})^\nu K_\nu(\frac{\ h\ }{a}))^*$	$h \in R^d, d \geq 1$

* K_ν تابع بسل تعدیل‌یافته نوع دوم از مرتبه ν و ν پارامتر هموارساز است (آبراموویتز و استیگان، ۱۹۷۰).

تجربی و مدل منتخب تغییرنگار وجود داشته باشد. بر اساس این ایده روش‌های کمترین توان‌های دوم عادی^۲ (OLS)، کمترین توان‌های دوم تعمیم‌یافته^۳ (GLS) و کمترین توان‌های دوم موزون^۴ (WLS) تعریف شده‌اند.

^۱Cross validation

^۲Ordinary least square

^۳Generalized least square

^۴Weighted least square

پیش‌گویی معمولاً یکی از مهم‌ترین اهداف در یک مطالعه آماری است. به‌طور خلاصه پیش‌گویی تلاشی برای برآورد مقدار مشخصه مورد نظر در یک فرد از جامعه بر اساس مشاهدات به‌دست آمده (از طریق یک نمونه تصادفی) از n فرد از جامعه است. کریگیدن یکی از پرکاربردترین روش‌های تحلیل داده‌های فضایی برای پیش‌گویی در موقعیت‌های فاقد مشاهده است که اولین بار توسط مترن (۱۹۶۳) معرفی شد. فرض کنید پیش‌گویی $Z(s_0)$ بر اساس مشاهدات $Z(s_1), \dots, Z(s_n)$ در موقعیت‌های $s_1, \dots, s_n$ مورد نظر باشد. در روش کریگیدن، یک ترکیب خطی از مشاهدات به‌صورت $Z(s_0) = \sum_{i=1}^n \lambda_i Z(s_i) + k$ را به‌عنوان پیش‌گوی $Z(s_0)$ در نظر گرفته و ضرایب λ_i و k به نحوی برآورد می‌شود که این پیش‌گو ناریب بوده و میانگین توان‌های دوم خطای آن کمترین باشد. بسته به این که میانگین میدان تصادفی معلوم، ثابت نامعلوم یا تابعی از موقعیت‌ها باشد، کریگیدن به سه نوع ساده، عادی و عام تفکیک می‌شود. در واقع کریگیدن معادل بهترین پیشگوی خطی ناریب^۱ (BLUP) است (محمدزاده، ۱۳۹۸).

۳ رگرسیون بردار پشتیبان

در سال‌های اخیر انواع مدل‌های رگرسیون بردار پشتیبان^۲ (SVR) مورد توجه علوم مختلف قرار گرفته‌اند. عملکرد خوب آن‌ها در مواجهه با داده‌های پرت و در مطالعات دارای حجم نمونه کوچک از دلایل محبوبیت این مدل‌ها است. ویژگی مهم دیگر، کارایی زیاد در فضاهایی با ابعاد بالا، حتی بیشتر از حجم نمونه است. صرفه‌جویی در حافظه از طریق تصمیم‌گیری صرفاً بر اساس بردارهای پشتیبان، قابلیت تعمیم بالا و شیوه عملکرد خوب در مسائل ناخطی موجب گسترش به کارگیری رگرسیون بردار پشتیبان در علوم مختلف از جمله در تحلیل داده‌های فضایی در برازش تغییرنگار و انواع مختلف پیش‌گویی فضایی شده است. رگرسیون بردار پشتیبان از اصل کمینه‌سازی ریسک ساختاری استفاده می‌کند و در نهایت به یک جواب بهینه کلی منجر می‌شود. این مدل مبتنی بر برازش دو ابرصفحه (در حالت ساده، دو خط موازی) به مجموعه داده‌ها، به‌گونه‌ای است که ابرصفحه‌ها آنقدر از یکدیگر دور باشند که تمام داده‌ها را شامل شوند و از طرفی کمترین تفاوت ممکن بین مقادیر مشاهده شده و پیش‌گویی شده متغیر پاسخ وجود داشته باشد (آیزمن، ۲۰۰۸). در این روش، مدل رگرسیونی برابر ابرصفحه دقیقاً وسط دو ابرصفحه حاشیه‌ای و موازی با آن‌ها است. داده‌هایی که روی ابرصفحه‌های حاشیه‌ای قرار دارند، بردارهای پشتیبان نامیده می‌شوند.

۳.۱ رگرسیون بردار پشتیبان خطی

فرض کنید مجموعه‌ای شامل n نمونه باشد، که در آن $x_i \in R^p$ بردار ورودی و y_i خروجی متناظر است. هدف برازش مدل رگرسیونی خطی $f(x) = w'x + b + \varepsilon$ به داده‌ها است، که در آن b عرض از مبدأ، w بردار پارامترهای مدل، x بردار متغیرهای تبیینی و $w'x$ ضرب داخلی دو بردار را نشان می‌دهد. با در نظر گرفتن حداکثر خطای مطلق e ، باید فاصله بین دو ابرصفحه حاشیه‌ای $y = w'x + b + e$

¹Best linear unbiased predictor

²Support vector regression

$y = \mathbf{w}'\mathbf{x} + b - e$ یعنی $\frac{ye}{\|\mathbf{w}\|}$ را به گونه ای ماکسیم نمود که تمام y_i ها در بین دو ابرصفحه قرار گیرند. مسئله بهینه سازی برای تعیین پارامترهای مدل، به صورت

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|^2 / 2, \quad \begin{cases} y_i - \mathbf{w}'\mathbf{x}_i - b \leq e \\ \mathbf{w}'\mathbf{x}_i + b - y_i \leq e, \quad i = 1, \dots, n \end{cases}$$

است، که در آن مقدار ثابت e توسط کاربر اختیار می شود. یک روش اصلاحی برای وقتی که نتوان ابرصفحه های حاشیه ای را به نحوی تعیین کرد که تمام نقاط را در بگیرند تعریف مقداری جریمه (متغیر مثبت کاهش) برای نقاطی است که خطای برآورد آن ها از e تجاوز می کند. در این صورت مسئله بهینه سازی به صورت

$$\min_{\mathbf{w}, b, \zeta_i, \zeta_i^*} \|\mathbf{w}\|^2 / 2 + C \sum_{i=1}^n (\zeta_i + \zeta_i^*), \quad \begin{cases} y_i - \mathbf{w}'\mathbf{x}_i - b \leq e + \zeta_i \\ \mathbf{w}'\mathbf{x}_i + b - y_i \leq e + \zeta_i^* \\ \zeta_i \geq 0, \zeta_i^* \geq 0, \quad i = 1, \dots, n \end{cases} \quad (2)$$

تغییر می کند، که در آن ζ_i و ζ_i^* مقدار تجاوز از e و ثابت C میزان جریمه ای است که توسط کاربر مشخص می شود (انجم، ۲۰۰۴). دوگان مسئله (۲) با استفاده از ضرایب لاگرانژ به صورت

$$\begin{aligned} L_p = & \|\mathbf{w}\|^2 / 2 + C \sum_{i=1}^n (\zeta_i + \zeta_i^*) - \sum_{i=1}^n \alpha_i (e + \zeta_i - y_i + \mathbf{w}'\mathbf{x}_i + b) \\ & - \sum_{i=1}^n \alpha_i^* (e + \zeta_i^* + y_i - \mathbf{w}'\mathbf{x}_i - b) - \sum_{i=1}^n \eta_i^* \zeta_i^* - \sum_{i=1}^n \eta_i \zeta_i \end{aligned} \quad (3)$$

به دست می آید، که در آن $\alpha_i^*, \alpha_i, \eta_i^*, \eta_i$ ضرایب لاگرانژ هستند. بنا به مبانی بهینه سازی، مسئله به

$$\begin{aligned} \min \frac{1}{2} \sum_{i,j=1}^n (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(\mathbf{x}_i' \mathbf{x}_j) + e \sum_{i=1}^n (\alpha_i^* + \alpha_i) - \sum_{i=1}^n y_i (\alpha_i^* - \alpha_i), \\ \sum_{i=1}^n (\alpha_i^* - \alpha_i) = 0, \quad \alpha_i^*, \alpha_i \in [0, C] \end{aligned}$$

تبدیل می‌شود. در نهایت نشان داده می‌شود که متغیر پاسخ به صورت

$$\hat{f}(x) = w'x + b = \sum_{i=1}^n (\alpha_i^* - \alpha_i) x_i' x + b$$

برآورد می‌شود. $S = \{i | 0 < \alpha_i^* + \alpha_i < C\}$ مجموعه بردارهای پشتیبان تعریف می‌شود و

$$b = \frac{1}{|S_{sv}|} \sum_{i \in S} (y_i - w'x_i - \text{sign}(\alpha_i^* - \alpha_i)e)$$

که در آن $|S_{sv}|$ تعداد بردارهای پشتیبان و sign تابع علامت است.

۳.۲ رگرسیون بردار پشتیبان ناخطی

اگر رابطه بین بردار ورودی و متغیر خروجی پیچیده باشد به گونه‌ای که مدل خطی برای برازش به داده‌ها مناسب نباشد می‌توان با انتقال فضای ورودی اولیه از طریق نگاشت ناخطی $\zeta : \chi \rightarrow \zeta$ به فضایی با ابعاد بالاتر به نام فضای ویژگی، یک مدل خطی به داده‌ها در این فضا برازش داد. برای این منظور از تابع هسته استفاده می‌شود. تابع هسته $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ ضرب داخلی در فضای ویژگی است. از مهمترین ویژگی‌های استفاده از تابع هسته، عدم نیاز به تعیین صورت صریحی برای $\phi(x)$ در فضای ویژگی است بلکه فقط محاسبه ضرب داخلی مشاهدات در فضای ویژگی $\langle \phi(x_i), \phi(x_j) \rangle$ کافی است. در صورتی که $\zeta \subset \mathbf{D}$ مجموعه داده شامل $x_i, i = 1, \dots, n$ باشد، مقایسه زوجی نقاط $\mathbf{D}$ از طریق مقادیر ماتریس مربعی $K = [K_{ij} = K(x_i, x_j)]_{n \times n}$ صورت می‌گیرد. برای محاسبه تابع هسته از همان نقاط فضای ورودی استفاده می‌شود بنابراین تعیین تابع هسته به مراتب کار ساده‌تری نسبت به تعیین ϕ مناسب و انجام ضرب داخلی در فضای ویژگی است. برای به کارگیری هسته فقط کافی است که نوع تابع هسته مناسب انتخاب شود (روحی و همکاران، ۱۴۰۲). هسته خطی، چندجمله‌ای، گاوسی و حلقوی متداول‌ترین توابع هسته هستند که در SVR استفاده می‌شوند و در جدول ۲ آورده شده‌اند. تعیین مقادیر بهینه پارامترهای هسته از طریق روش اعتبارسنجی متقابل و با جستجوی شبکه‌ای در فضای پارامتر انجام می‌شود (آیزمن، ۲۰۰۸؛ اتهم، ۲۰۰۴).

مسئله بهینه‌سازی در رگرسیون بردار پشتیبان ناخطی، پیشگویی متغیر پاسخ و برآورد بردار ضرایب w و ثابت b به صورت زیر به دست می‌آید.

$$\min \frac{1}{2} \sum_{i,j=1}^n (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) + e \sum_{i=1}^n (\alpha_i^* + \alpha_i) - \sum_{i=1}^n y_i (\alpha_i^* - \alpha_i),$$

$$\sum_{i=1}^n (\alpha_i^* - \alpha_i) = 0, \quad \alpha_i^*, \alpha_i \in [0, C],$$

جدول ۰۲. توابع هسته متداول.

تابع هسته	$K(\mathbf{x}_i, \mathbf{x}_j)$
خطی	$\mathbf{x}_i' \mathbf{x}_j$
چندجمله‌ای	$(c + \gamma \mathbf{x}_i' \mathbf{x}_j)^p$
گاوسی	$e^{-\gamma \ \mathbf{x}_i - \mathbf{x}_j\ ^2}$
حلقوی	$\tanh(c + \gamma \mathbf{x}_i' \mathbf{x}_j)$

$$\hat{f}(x) = \mathbf{w}'\phi(x) + b = \sum_{i=1}^n (\alpha_i^* - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) + b,$$

$$\mathbf{w} = \sum_{i=1}^n (\alpha_i^* - \alpha_i) \phi(\mathbf{x}_i), \quad (۴)$$

$$b = \frac{1}{|S_{sv}|} \sum_{i \in S} (y_i - \sum_{i=1}^n (\alpha_i^* - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) - \text{sign}(\alpha_i^* - \alpha_i) e). \quad (۵)$$

۴ روندزدایی داده‌های فضایی به روش رگرسیون بردار پشتیبان

برای روندزدایی به کمک رگرسیون بردار پشتیبان دو رویکرد عمده دنبال می‌شود: الف- تعیین مدل روند به کمک رگرسیون بردار پشتیبان، ب- حذف روند از داده‌ها برای برآورد تغییرنگار. رایج‌ترین روش مدل کردن روند، در نظر گرفتن تابعی خطی مانند $\mu(s) = \sum_{j=0}^p \beta_j \phi_j(s)$ است، که در آن $(\phi_0(s), \dots, \phi_p(s))$ برداری از توابع معلوم بر حسب موقعیت s هستند و بردار ضرایب نامعلوم $\beta = (\beta_0, \dots, \beta_p)$ می‌تواند به یکی از روش‌های کمترین توان‌های دوم برآورد شود. از آن جا که ملاکی برای تعیین بردار $(\phi_0(s), \dots, \phi_p(s))$ وجود ندارد، ابتدا مجموعه داده‌ها ترسیم شده و سپس توابعی که برای برازش به مجموعه داده‌ها مناسب به نظر برسند که معمولاً توابع ساده مانند چندجمله‌ای‌ها هستند، انتخاب می‌شوند. بنابراین استفاده از رگرسیون بردار پشتیبان در مدل کردن روند باید پیشنهاد مناسبی باشد. اگر برازش مدل خطی برای مجموعه داده‌ها مناسب نباشد استفاده از توابع هسته مختلف امکان به کارگیری مدل‌های مختلف ناخطی را به کاربر می‌دهد. با برازش مدل‌های مختلف به داده‌ها و مقایسه نتایج می‌توان مناسب‌ترین مدل را تعیین کرد. پس از برآورد روند، با کاستن $\hat{\mu}(s)$ از $Z(s)$ ،

$$\hat{\delta}(s) = Z(s) - \hat{\mu}(s). \quad (۶)$$

به‌دست می‌آید. تغییرنگار تجربی داده‌های روندزوده براساس (۱) محاسبه و نمودار آن رسم می‌شود. از بین مدل‌های معتبر تغییرنگار، مناسب‌ترین مدل برای برازش به تغییرنگار تجربی انتخاب و پارامترهای آن برآورد می‌شود. برای

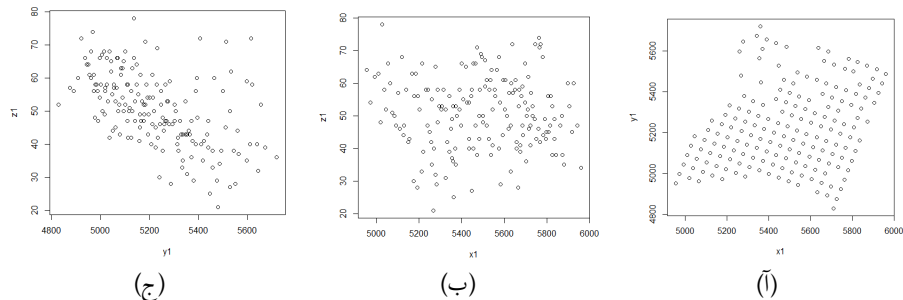
پیش‌گویی در یک موقعیت جدید ابتدا جمله خطا $\delta(s_{new})$ به روش کریگیدن ساده برآورد و سپس به آن برآورد روند به صورت $\hat{Z}(s_{new}) = \hat{\mu}(s_{new}) + \hat{\delta}(s_{new})$ اضافه می‌شود. برای ارزیابی مدل پیشنهادی از روش اعتبارسنجی متقابل استفاده می‌شود. در این روش داده‌ها به دو دسته آموزشی و آزمایشی تفکیک می‌شوند. از داده‌های آموزشی برای تعیین مدل و برآورد پارامترهای آن و از داده‌های آزمایشی برای بررسی کارایی مدل استفاده می‌شود. در اعتبارسنجی متقابل با k لایه ابتدا داده‌ها به k زیرمجموعه افراز می‌شوند. یکی از آن‌ها برای داده‌های آزمایشی و $k-1$ تای دیگر برای آموزشی به‌کار می‌رود. با k بار تکرار این کار، همه داده‌ها یک بار برای آزمایش و $k-1$ بار برای آموزش استفاده می‌شوند. در نهایت از طریق ریشه دوم میانگین توان‌های دوم خطا،

$$CV\text{-}RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Z(s_i) - \hat{Z}(s_{-i}))^2}$$

معیاری برای ارزیابی کارایی و مقایسه مدل‌ها حاصل می‌شود.

۵ تحلیل داده‌های کلسیم خاک

برای ارزیابی روندزدایی به روش رگرسیون بردار پشتیبان و تأثیر آن بر پیش‌گویی از مجموعه داده کلسیم خاک شامل ۱۷۸ مشاهده که از لایه‌های صفر تا بیست سانتی‌متری زمین برگرفته شده (اولیویرا، ۲۰۰۳) استفاده می‌شود. این داده‌ها از طریق بسته geoR در نرم‌افزار R در دسترس هستند. کنت و محمدزاده (۲۰۰۰) به روش اسپلین همواری از داده‌ها روندزدایی و نتایج پیش‌گویی را با روش کریگیدن عام مقایسه کردند. شکل ۱ (آ) نمای کلی از داده‌ها را نشان می‌دهد. مقادیر مشاهده شده در برابر طول جغرافیایی x و عرض جغرافیایی y به‌ترتیب در (ب) و (ج) نشان داده شده‌اند. ملاحظه می‌شود داده‌ها در جهت محور x روند نظام‌مندی ندارند، اما در جهت محور y دارای روندی کاهشی‌اند. به دلیل شناسایی پنج داده دورافتاده، کنت و محمدزاده (۲۰۰۰) سه درصد بالایی و پایینی داده‌ها را حذف و مدل روند را برای داده‌های پیراسته تعیین کردند. برای مقایسه نتایج روش پیشنهادی با نتایج آن‌ها، ما نیز به همان شیوه داده‌ها را پیراسته می‌کنیم. برای مدل کردن روند به روش رگرسیون بردار پشتیبان از بسته e1071 در نرم‌افزار R استفاده می‌شود. از آن جا که بُعد متغیرهای مستقل (ورودی) دو است مدل رگرسیونی به‌صورت، $f(x) = w'\phi(x) + b + \varepsilon$ ، تعریف می‌شود که در آن $\phi(x)$ تصویر طول و عرض جغرافیایی در فضای ویژگی است. سه تابع هسته خطی، چندجمله‌ای و گاوسی در نظر می‌گیریم. مشخصات و نتایج هسته‌ها برای مجموعه داده‌های اصلی و پیراسته در جدول ۳ نشان داده شده‌اند. جریمه (C) و مقدار مجاز برای خطا (e) برای هر سه هسته به‌ترتیب ۱ و ۰/۱ در نظر گرفته شده است. پارامتر گاما در هسته‌های چندجمله‌ای و گاوسی ۰/۵ و در هسته چندجمله‌ای ثابت c برابر ۰ و درجه آن ۳ تعیین شده است. در دو ستون آخر جدول ۳ تعداد بردارهای پشتیبان برای هسته‌های مختلف داده‌های اصلی و پیراسته شده آورده شده، بیشترین تعداد آن‌ها مربوط به مدل چندجمله‌ای است.



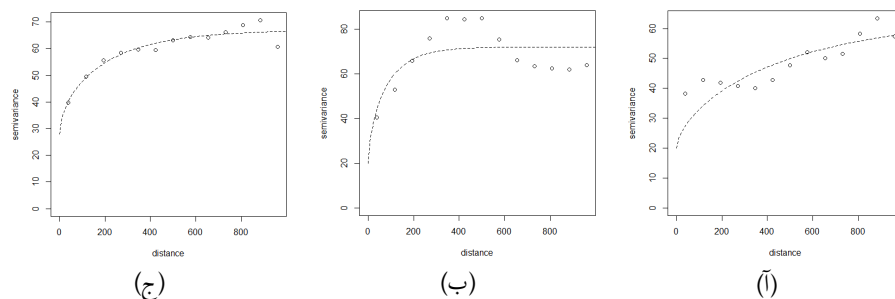
شکل ۱. آ: موقعیت مشاهدات کلسیم ، مقادیر مشاهده شده در برابر ب: طول جغرافیایی ج: عرض جغرافیایی.

در مرحله دوم بنا به (۶) با کم کردن تابع روند برآورد شده از داده‌ها، داده‌های روندزدوده به دست می‌آید. در

جدول ۳. مشخصات و نتایج توابع هسته.

بردار پشتیبان		پارامترهای هسته			هسته
اصلی	پیراسته	درجه (p)	مقدار ثابت (c)	گاما	
۱۵۹	۱۵۱	-	-	-	خطی
۱۶۷	۱۵۴	۳	۰	۰/۵	چندجمله‌ای
۱۵۸	۱۴۰	-	-	۰/۵	گاوسی

مرحله سوم تغییرنگار تجربی برای داده‌های روندزدوده محاسبه شده و مدل مترن به عنوان مدل مناسب به تغییرنگار تجربی برازش داده می‌شود. در شکل ۲ برازش نیم‌تغییرنگار به روش کمترین توان‌های دوم عادی (OLS) نشان داده شده‌اند. پارامترهای مدل منتخب شامل دامنه، ازاره، اثرقطعه‌ای و ضریب کاپا است که برای هر سه نوع تابع هسته



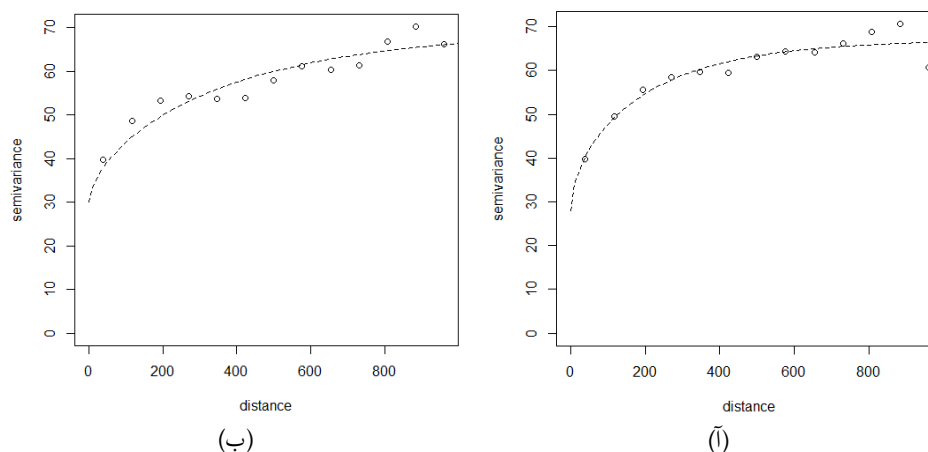
شکل ۲. برازش نیم‌تغییرنگار به داده‌های روندزدوده با روش کمترین توان‌های دوم عادی آ: هسته گاوسی ب: هسته چندجمله‌ای ج: هسته خطی.

در روش رگرسیون بردار پشتیبان (به اختصار ر.ب.پ) در جدول ۴ تعیین شده‌اند.

جدول ۴. برآورد پارامترهای تغییرنگار.

برآورد پارامتر				روش
K	c_0	$c + c_0$	a	
۰/۳۰	۲۰	۵۱/۴۹۳۳	۲۹۰/۳۰۴۲	ر.ب.پ. با هسته خطی
۰/۳۰	۲۰	۵۲/۰۲۱۱	۱۱۴/۰۸۰۰	ر.ب.پ. با هسته چندجمله‌ای
۰/۳۰	۲۰	۴۴/۴۸۸۶	۶۹۱/۸۰۰۰	ر.ب.پ. با هسته گاوسی
۰/۲۷	۲۸	۳۹/۰۶۶۲	۲۹۳/۷۶۴۴	کریگیدن عام
۰/۳۰	۳۰	۳۹/۶۳۷۲	۵۱۱/۰۰۷۹	اسپلین

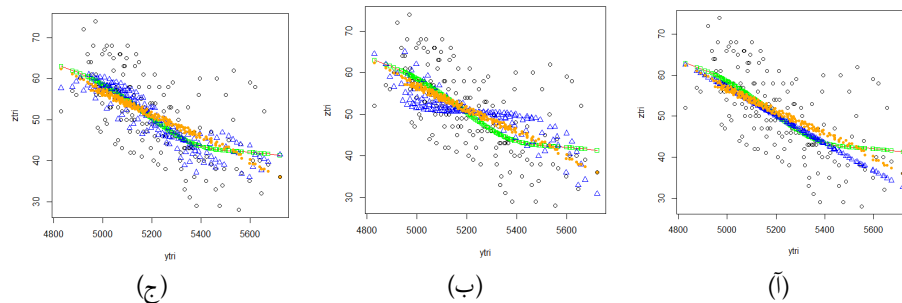
در نهایت پیش‌گویی برای هر موقعیت مشاهده شده، برای داده‌های حذف شده و موقعیت جدید انجام می‌شود. به منظور مقایسه نتایج پیش‌گویی به روش رگرسیون بردار پشتیبان با روش‌های اسپلین و کریگیدن عام، مراحل انجام شده در روش رگرسیون بردار پشتیبان برای این دو روش نیز تکرار می‌شود. در روش اسپلین، تابع هموار به کمک دستور smooth.spline و در روش کریگیدن عام یک تابع چندجمله‌ای مرتبه اول به کمک بسته geoR در نرم‌افزار R برای برآورد روند به مجموعه داده‌ها برازش می‌شود. نمودار نیم‌تغییرنگارهای برازش داده شده به داده‌های روندزدوده در شکل ۳ نشان داده شده‌اند.



شکل ۳. نمودار نیم‌تغییرنگار برازش داده شده به روش کمترین توان‌های دوم عادی آ: کریگیدن عام ب: اسپلین.

در دو روش اسپلین و کریگیدن عام به تغییرنگار تجربی مدل‌های پارامتری کواریانس مختلف برازش داده شده، مدل مترن نسبت به سایر مدل‌ها دارای مجموع توان دوم کمتری است. برآورد پارامترهای این مدل در روش‌های

کریگیدن عام و اسپلاین در جدول ۴ آورده شده‌اند. در شکل ۳ نتایج پیش‌گویی برای داده‌های پیراسته اسپلاین با رنگ سبز، کریگیدن عام با رنگ نارنجی و رگرسیون بردار پشتیبان با هسته‌های متفاوت با رنگ آبی نشان داده شده‌اند. همان‌طور که در شکل ۴ مشاهده می‌شود نتایج پیش‌گویی کریگیدن عام، اسپلاین و رگرسیون بردار پشتیبان



شکل ۴. نتایج پیش‌گویی داده‌های پیراسته روش اسپلاین، کریگیدن عام و رگرسیون بردار پشتیبان با هسته آ: خطی ب: چندجمله‌ای ج: گاوسی.

با هسته‌های متفاوت به‌طور تقریبی در یک راستا قرار دارند. نتایج پیش‌گویی سه روش به روش اعتبارسنجی متقابل و برازش تغییرنگار به روش OLS با یکدیگر مقایسه می‌شوند. ملاک ارزیابی دقت پیش‌گویی، کمینه بودن ریشه دوم میانگین توان‌های دوم خطا است. از نتایج به‌دست آمده در جدول ۵ مشاهده می‌شود که به‌طور کلی خطای پیش‌گویی روش رگرسیون بردار پشتیبان (با هسته خطی، چندجمله‌ای و گاوسی) نسبت به اسپلاین و کریگیدن عام کمتر است و هسته گاوسی بیشترین دقت پیش‌گویی را دارد. در ادامه برای ۱۲ داده حذف شده در ابتدا، شامل ۳ درصد بالا و پایین داده‌ها، که موقعیت و مقدار آن‌ها مشخص است پیش‌گویی به سه روش انجام می‌شود. نتایج آورده شده در جدول ۵ نشان می‌دهد که ریشه دوم میانگین توان‌های دوم خطای روش رگرسیون ماشین بردار پشتیبان با هسته گاوسی کمینه است. مجدداً مشاهده می‌شود روش رگرسیون بردار پشتیبان نسبت به دو روش اسپلاین و کریگیدن عام عملکرد بهتری دارد.

در نهایت پیش‌گویی برای چند موقعیت جدید انجام می‌شود. از بین ۶ موقعیت، مقدار دو موقعیت معلوم در نظر گرفته شده است. در جدول ۶ موقعیت جغرافیایی مشاهدات (بر حسب درجه و دقیقه) و نتایج پیش‌گویی آورده شده است. مطابق انتظار مقدار پیش‌گویی در نقاط دارای مشاهده به‌طور دقیق با مقدار واقعی برابر و واریانس آن صفر است. نتایج پیش‌گویی چهار موقعیت جدید در روش‌های مختلف بسیار به یکدیگر نزدیک است. دامنه تغییرات نتایج پیش‌گویی روش‌های مختلف در دو موقعیت جدید دوم و چهارم کمتر از ۱ است. بیشترین دامنه تغییرات پیش‌گویی برابر با ۱/۳۸ مربوط به موقعیت پنجم است. بنابراین روش رگرسیون بردار پشتیبان می‌تواند یک رقیب خوب در مقابل پیش‌گویی به روش کریگیدن عام و اسپلاین باشد.

براساس شکل ۱ در داده‌های کلسیم روندی نظام‌مندی بر حسب طول جغرافیایی دیده نمی‌شود. ولی در تمامی

جدول ۵. نتایج پیش‌گویی موقعیت‌ها بر اساس اعتبار سنجی متقابل و داده‌های حذفی.

ریشه دوم میانگین توان‌های دوم خطا		
اعتبار سنجی متقابل	داده‌های حذفی	
۷/۰۱۴	۱۸/۹۹۷	کریگیدن عام
۶/۹۸۱	۱۸/۷۳۶	اسپالین
۶/۹۴۳	۱۸/۲۸۴	ر.ب.پ. با هسته خطی
۶/۸۹۷	۱۸/۲۸۸	ر.ب.پ. با هسته چندجمله‌ای
۶/۸۲۳	۱۶/۳۶۲	ر.ب.پ. با هسته گاوسی

جدول ۶. پیش‌گویی موقعیت‌های جدید.

رگرسیون بردار پشتیبان با هسته						
موقعیت	طول	عرض	خطی	چندجمله‌ای	گاوسی	کریگیدن عام اسپالین
معلوم	۴۹۵۷	۴۹۵۱	۶۴	۶۴	۶۴	۶۴
نامعلوم	۵۰۰۰	۱۵۰۱۸	۵۷/۵۱	۵۷/۷۷	۵۸/۱۹	۵۷/۶۲
نامعلوم	۵۵۹۲	۵۴۳۰	۳۸/۴۵	۳۸/۹۵	۳۸/۴۸	۳۹/۶۸
نامعلوم	۵۶۰۰	۵۴۱۰	۳۹/۶۱	۴۰/۲۰	۳۹/۵۳	۴۰/۵۲
نامعلوم	۵۸۱۰	۵۱۵۰	۴۷/۳۷	۴۷/۴۷	۴۶/۶۰	۴۷/۹۹
معلوم	۵۹۶۱	۵۴۸۷	۳۴	۳۴	۳۴	۳۴

نتایج به دست آمده، به استثنای روش اسپالین که فقط از مختصات عرض جغرافیایی در آن استفاده شده است، مختصات طول و عرض جغرافیایی در نظر گرفته شده است. به منظور بررسی تاثیر حذف طول جغرافیایی بر نتایج، برای روش رگرسیون بردار پشتیبان با همان هسته‌ها و کریگیدن عام، تمام مراحل قبل با در نظر گرفتن فقط عرض جغرافیایی تکرار می‌کنیم. نتایج حاصل در جدول ۷ ارائه شده است. مقایسه جدول‌های ۵ و ۷ نشان می‌دهد که حذف طول جغرافیایی در نتایج روش کریگیدن عام تاثیری نداشته است. در روش رگرسیون بردار پشتیبان با هسته‌های چندجمله‌ای و گاوسی استفاده از طول و عرض جغرافیایی به نتایج کمی بهتر منجر می‌شود اما تفاوت بسیار ناچیز است. اما با هسته خطی نتایج براساس فقط عرض جغرافیایی بهتر است که این تفاوت نیز قابل ملاحظه نیست. بنابراین می‌توان تحلیل داده‌ها را فقط براساس عرض جغرافیایی انجام داد.

جدول ۷. نتایج پیش‌گویی موقعیت‌ها بر اساس اعتبار سنجی متقابل و برای داده‌های حذفی بر اساس عرض جغرافیایی داده‌ها.

ریشه دوم میانگین توان‌های دوم خطا		
اعتبار سنجی متقابل	داده‌های حذفی	
۷/۰۱۴	۱۸/۹۹۷	کریگیدن عام
۶/۹۸۱	۱۸/۷۳۶	اسپلاین
۶/۹۲۸	۱۷/۵۵۰	ر.ب.پ. با هسته خطی
۷/۰۲۳	۱۸/۳۲۵	ر.ب.پ. با هسته چندجمله‌ای
۶/۸۵۴	۱۷/۳۲۱	ر.ب.پ. با هسته گاوسی

بحث و نتیجه‌گیری

از موضوعاتی که در تحلیل اکتشافی داده‌های فضایی مورد بررسی قرار می‌گیرد مانایی داده‌ها است که دست کم باید شرایط مانای ذاتی برقرار باشد. یک دلیل عدم مانایی وجود روند در داده‌ها است. برای از بین بردن اثر نامطلوب روند بر برآورد تغییرنگار و پیش‌گویی باید ابتدا روند کشف و مدل شود که رگرسیون بردار پشتیبان می‌تواند ابزاری مفید و کارآمد برای این منظور باشد. امکان به کارگیری هسته‌های متنوع از مزایای این روش در شناسایی مدل مناسب برای روند است. تحلیل یک مجموعه داده واقعی نشان داد که پیش‌گویی به روش پیشنهادی دارای دقت بالاتری نسبت به دو روش کریگیدن عام و اسپلاین است. همچنین روش پیشنهادی در پیش‌گویی مقدار موقعیت‌های حذف شده نیز دارای عملکرد بهتری بود.

تقدیر و تشکر

از داوران محترم جهت صرف وقت، مطالعه دقیق و ارائه نظرات ارزنده که باعث ارتقای کیفیت مقاله شدند کمال قدردانی و تشکر می‌شود. همچنین از رهنمودهای ارزنده و ویرایش ادبی سردبیر محترم و هیئت تحریریه مجله علوم آماری سپاسگزاریم.

مراجع

توسلی، ع.، واقعی، ی. و ناظمی، ع. (۱۳۹۸)، برآورد تغییرنگار با استفاده از روش رگرسیون بردار پشتیبان، سومین سمینار آمار فضایی و کاربردهای آن، زنجان، ۳۲-۳۹.

روحی، آ.، جهادی، ف.، روزبه، م. و زال‌زاده، س. (۱۴۰۲)، تحلیل داده‌های با بعد بالا با استفاده از رگرسیون بردار پشتیبان تعمیم یافته، رگرسیون تابعی، رگرسیون ستیغی و لاسو، مجله علوم آماری، ۱۷(۱)، ۸۱-۱۰۲.

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

Abramowitz, M. and Stegun, I. (1970), *Handbook of Mathematical Functions*, Dover, New York.

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley and Sons, New York.

Ethem, A. (2004), *Introduction to Machine Learning*, The MIT Press, Cambridge.

Fuentes, M. (2006), Testing for Separability of a Spatial Temporal Covariance Functions, *Journal of Statistical Planning and Inference*, **136**, 447-466.

Han, C., Wang, J., Zheng, M., Wang, E., Xia, J., Li, G., and Choe, S. (2016), New Variogram Modeling Method Using MGGP and SVR, *Earth Sci Inform*, **9**, 197-213.

Huang, Z., Wang, H. and Zhang, R. (2012), An Improved Kriging Interpolation Technique Based on SVM and Its Recovery Experiment in Oceanic Missing Data. *American Journal of Computational Mathematics*, **20(01)**, 56.

Hwang, C. and Shim, J. (2017), Geographically Weighted Least Squares-Support Vector Machine, *Journal of the Korean Data and Information Science Society*, **28(1)**, 227-235.

Izenman, A. J. (2008), *Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning*, Springer-Verlag, New York.

Jun, M. and Genton, M. (2012), A Test for Stationarity of Spatio-Temporal Random Fields on Planar and Spherical Domains, *Statistica Sinica*, **22**, 1737-1764.

Kent, J.T. and Mohammadzadeh, M. (2000), Global Optimization of the Generalized Cross-Validation Criterion, *Statistics and Computing*, **10**, 231-236.

Matheron, G. (1962), Traite de Geostatistique Appliquee, Tome I. *Memoires du Bureau de Recherches Geologiques et Minières*, **14**, Editions: Technip, Paris.

Matheron, G. (1963), Principles of Geostatistics, *Economic Geology*, **58**, 1246-1266.

- Miranda, H. and de Miranda, M.S. (2011), Combining Robustness with Efficiency in the Estimation of the Variogram, *Mathematical Geosciences*, **43**(3), 363-377.
- Oliveira, M.C.N. (2003), Métodos de Estimação de Parâmetros Em Modelos Geostatísticos Com Diferentes Estruturas de Covariâncias, *Uma aplicação ao Teor de Cálculo no Solo, Tese de Doutorado, ESALQ/USP/Brasil*.
- Priestley, M. B. and Subba Rao, T. (1969), A Test for Non-Stationarity of a Time Series, *Journal of the Royal Statistical Society, Series A*, **31**, 140-149.
- Sarjent, T.J. (1968), Interest Rate in the Nineteen-Fifties, *Review of Economics and Statistics*, **50**, 164-172.
- Silva, C. and Boisvert, J. (2020), Locally Weighted Support Vector Regression for Spatial Predictions, *CCG Annual Report 22*.
- Vieira, S. R. (2000), Geoestatística Em Estudos de Variabilidade Espacial Do Solo, *In: Novais, R. F.; Alvarez, V. H.; Schaefer, G. R. (Ed.). Tópicos em Ciência do solo. Viosa: Sociedade Brasileira de Ciência do solo*, **1**, 1-54.
- Vieira, S. R., Hatfield, T. L., Nielsen, D. R. and Biggar, J. W. (1983), Geostatistical Theory and Application to Variability of Some Agronomical Properties, *Hilgardia*, **51**, 1-75.
- Vieira, S. R., de Carvalho, J. R. P., Ceddia, M. A. and Gonzalez, A. P. (2010), Detrending Non-Stationary Data for Geostatistical Applications, *Bragantia*, **69**, 1-8.