مقایسه روشهای درون‌یابی برای داده‌های فضایی

محسن محمدزاده: دانشگاه تربیت مدرس
رحیم صفری فارفار: وزارت علوم تحقیقات و فناوری

چکیده

گاهی در تجزیه و تحلیل داده‌ها با مشاهداتی مواجه می‌شویم که مستقل نیستند و نوعاً وابستگی آن‌ها ناشی از مکان یا موقعیت فضایی است. در آمار فضایی معمولا روش‌های مختلفی برای درون‌یابی این‌گونه داده‌ها به کار گرفته می‌شوند. در این مقاله ضمن بررسی انواع درون‌یاب‌های فضایی، میزان دقت آن‌ها با استفاده از معیار میانگین مربعات خطا مقایسه عددی شده است و کریگیدن به عنوان روشی برتر در مقایسه با سایر درون‌یاب‌ها معرفی می‌شود.

مقدمه

در بعضی مطالعات محیطی، جغرافیایی و همه‌گیرشناسی با مشاهداتی مواجه می‌شویم که مکان یا موقعیت قرار گرفتن آن‌ها در ناحیه یا فضای مورد مطالعه تأثیر به سزایی در همبستگی بین مشاهدات دارد، به گونه‌ای که همبستگی مشاهدات در موقعیت‌های نزدیک به هم زیاد است و هر قدر موقعیت مشاهدات از یکدیگر دورتر می‌شود، همبستگی بین آن‌ها نیز کاهش می‌یابد. لذا یکی از اجزای اصلی این گونه مشاهدات، که داده‌های فضایی نامیده می‌شوند، موقعیت آن‌ها در بررسی شده است. در آمار فضایی معمولا یک میدان تصادفی برای مدل‌بندی داده‌های فضایی به کار گرفته می‌شود. میدان تصادفی گردایه‌ای از متغیرهای تصادفی مانند $\{Z(t);t\in D\}$ است، که در آن مجموعه اندیس‌گذار D، زیرمجموعه‌ای از فضای اقلیدسی d بعدی R^d است.

به دلیل عدم استقلال داده‌های فضایی، لازم است ساختار همبستگی در تجزیه و تحلیل آن‌ها در نظر گرفته شود. لذا شناسایی این ساختار امری ضروری است. معمولاً در آمار فضایی از تغییرنگار که به صورت

$$2\gamma(t_1,t_2)=Var[Z(t_1)-Z(t_2)] \quad ; \quad t_1,t_2\in D$$

تعریف می‌شود، برای تعیین ساختار همبستگی بین داده‌های فضایی استفاده می‌شود. هر میدان تصادفی را می‌توان به صورت

$$Z(t)=\mu(t)-\sigma(t) \quad ; \quad t\in D \qquad (۱)$$

واژه‌های کلیدی: کریگیدن، اسپلاین، عکس مجذور فاصله، عکس توان p ام فاصله

تجزیه نمود، که در آن $\mu(t)=E[Z(t)];\ \ t\in D$ میانگین میدان تصادفی یا روند و همچنین عبارت $\sigma(t)$ خطای تصادفی میدان است. در صورتی که میدان تصادفی فاقد روند باشد، یعنی $\mu(t)=\mu$ ثابت و مستقل از موقعیت t باشد و تغییرنگار $2\gamma(t_1,t_2)$ فقط تابعی از فاصله $h=t_1-t_2$ باشد، میدان تصادفی ایستای ذاتی نامیده می‌شود که در این مقاله مورد استفاده قرار خواهد گرفت.

یکی از روش‌هایی که پیوسته در تجزیه و تحلیل داده‌های فضایی به کار گرفته می‌شود، درون‌یابی آن‌ها براساس مشاهدات است که عموماً به منظور تخمین مقدار متغیر پاسخ در موقعیت‌های جدید و تهیه نقشه‌های جغرافیایی از متغیر بررسی شده به کار می‌رود. در آمار فضایی روش‌های مختلفی برای درون‌یابی به کار گرفته می‌شود. اولین بار برایان و وایز(۱۹۶۸) روش‌های عکس مجذور فاصله و عکس توان p ام فاصله را برای درون‌یابی داده‌های فضایی مورد استفاده قراردادند. سپس واهبا (۱۹۹۰) روش اسپلاین را برای درون‌یابی داده‌ها به کار گرفت[۸]. پس از آن کرسی (۱۹۹۳)[۱] کریگیدن را که اولین بار ماترون (۱۹۷۱)[٤] معرفی کرد، به عنوان بهترین تخمین‌گر خطی نااریب مورد استفاده قرار داد. در این مقاله روش‌های مختلف درون‌یابی در بخش دوم مورد بررسی قرار می‌گیرند. سپس در بخش سوم دقت درون‌یاب‌های مختلف براساس معیار میانگین مجذور خطاها و برای حجم‌های متفاوت نمونه ارزیابی می‌شوند. سرانجام برتری دقت روش کریگیدن برای درون‌یابی داده‌های فضایی نشان داده می‌شود.

درون‌یابهای فضایی

فرض کنید مشاهدات $Z=(Z(t_1),Z(t_2),\cdots,Z(t_n))^T$ در موقعیتهای $t_1,t_2,\cdots,t_n$ برای میدان تصادفی $\{Z(t);t\in D\}$ دراختیار باشند. دراین بخش روش‌های کریگیدن، اسپلاین، عکس مجذور فاصله وعکس توان p ام فاصله برای درون‌یابی میدان تصادفی در یک موقعیت مشخص مانند t_0 بررسی می‌شوند.

کریگیدن

فرض کنید به ازای هر t، میانگین میدان تصادفی مقدار ثابت و نامعلوم $\mu\in R$ و $\sigma(t)$ نیز یک میدان تصادفی ایستای ذاتی با میانگین صفر و تغییرنگار معلوم $2\gamma(h)$ باشد. بهترین تخمین‌گر خطی نااریب $\hat{Z}(t_0)=\sum_{i=1}^{n}\lambda_i Z(t_i)$، براساس مشاهدات $Z=(Z(t_1),Z(t_2),\cdots,Z(t_n))$ کریگیدن معمولی نامیده می‌شود، که در آن ضرایب $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_n)$ به گونه‌ای انتخاب می‌شوند که میانگین مربعات خطای

$$\sigma_e^2=E(\hat{Z}(t_0)-Z(t_0))^2 \qquad (۲)$$

با شرط $\sum_{i=1}^{n}\lambda_i = 1$ ، که تضمین کننده ناریبی آن است، کمین گردد. برای این منظور عبارت

$$E[Z(t_0)-\sum_{i=1}^{n}\lambda_i Z(t_i)]^2 - 2m(\sum\lambda_i - 1) \qquad (۳)$$

برحسب ضرایب $\lambda_1,\lambda_2,\cdots,\lambda_n$ و m ، که در آن m ضریب لاگرانژ است، کمین می‌شود و بردار ضرایب $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_n)$ به صورت

$$\Gamma^T = (\gamma + \underset{\sim}{1}\frac{(1-\underset{\sim}{1}^T\Gamma^{-1}\gamma)}{\underset{\sim}{1}^T\Gamma^{-1}\underset{\sim}{1}})^T\Gamma^{-1}$$

حاصل می‌گردد، که در آن $\underset{\sim}{1}$ یک بردار $n\times 1$ با عناصر واحد، $\gamma=(\gamma(t_0-t_1),\cdots,\gamma(t_0-t_n))^T$ و Γ یک ماتریس $n\times n$ با (i,j) امین عنصر $\gamma(t_i-t_j)$ است.

اگرمیانگین $\mu(t)$ برحسب t ثابت نباشد، با فرض آن‌که یک ترکیب خطی نامعلوم مانند $\mu(t)=\sum_{j=1}^{p+1}\phi_{j-1}(t)\beta_j$ از توابع معلوم $\{\phi_0(t),\cdots,\phi_p(t)\}$ باشد، تخمین مقدار $Z(t_0)$ برحسب مشاهدات، $\hat{Z}(t_0)=\sum_{i=1}^{n}\lambda_i Z(t_i)$ کریگیدن عمومی نامیده می‌شود و برای محاسبه آن ضرایب رابطه خطی به گونه‌ای تعیین می‌شوند که عبارت (۲) را کمین نمایند. بدین منظور باید رابطه

$$E[Z(t_0)-\sum\lambda_i Z(t_i)]^2 - 2\sum_{j=1}^{p+1}m_{j-1}\{\sum\lambda_i\phi_{j-1}(t_i)-\phi_{j-1}(t_0)\} \qquad (۴)$$

باید برحسب ضرایب کمین گردد. در این صورت ضرایب بهین به صورت

$$\lambda_u = \Gamma_u^{-1}\gamma_u$$

حاصل می‌شوند که در آن Γ_u یک ماتریس متقارن $(n+p+1)\times(n+p+1)$ است.

$$\lambda_u = (\lambda_1,\cdots,\lambda_n,m_0,\cdots,m_p)^T$$
$$\gamma_u = (\gamma(t_0-t_1),\cdots,\gamma(t_0-t_n),1,\phi_1(t_0),\cdots,\phi_p(t_0))^T$$

اسپلاین

فرض کنید مشاهدات $Z=(Z(t_1),\cdots,Z(t_n))$ در موقعیت‌های $t_1,\cdots,t_n$ در فضای اقلیدسی R^d در دسترس باشند و بخواهیم مدل $Z=g(t)+e$ ، که در آن g تابعی نامعلوم است، با استفاده از روش حداقل مربعات به داده‌ها برازش داده شود. اگر نقاط به وسیله خط راست به هم متصل شوند، در این صورت مجموع مربعات خطاها صفر خواهد شد. حال اگر فرض شود تابع g دارای مشتق مرتبه دوم پیوسته است و از تمامی

نقاط (t_i, Z_i) می‌گذرد، در این حالت نیز مجموع مربعات باقیمانده‌ها صفر می‌شود، اما منحنی مواج و دارای نوسانات زیادی خواهد بود. از طرفی مشتق مرتبه دوم g نسبت به t، اندازه ناهمواری منحنی در نقاطی که منحنی تغییر جهت می‌دهد را نشان می‌دهد، که می‌تواند مثبت یا منفی باشد. مجموع مربعات این مشتق‌ها در تمام نقاطی که منحنی تغییر جهت می‌دهد، میزان ناهمواری منحنی را مشخص می‌نماید. برای اندازه‌گیری میزان ناهمواری منحنی g می‌توان از رابطه

$$J_m^d(g) = \sum \int \cdots \int \frac{m!}{\alpha_1! \cdots \alpha_d!} \left(\frac{\partial^m g(t)}{\partial t_1^{\alpha_1} \cdots \partial t_d^{\alpha_d}}\right)^2 dt_1 \cdots dt_d$$

استفاده نمود، که در آن $\alpha_1, \cdots, \alpha_d$ مقادیر نامنفی هستند به طوری که $\sum_{i=1}^{d} \alpha_i = m$ و m مرتبه $\hat{g}_\lambda$ است.

برای برآورد یک منحنی هموار g لازم است علاوه بر عامل مجموع مربعات باقیمانده‌ها، میزان ناهمواری آن را نیز در نظر گرفت. برای این منظور مجموع مربعات جریمه‌ای به صورت

$$S(g, \lambda) = \sum_{i=1}^{n} (z_i - g(t_i))^2 + \lambda J_m^d(g) \qquad (٦)$$

تعریف می‌شود، که در آن $\lambda > 0$ پارامتر همواری نامیده می‌شود. برآورد g_λ که مجموع مربعات جریمه‌ای $S(g, \lambda)$ را مینیمم می‌کند و روی کلاس تمام توابع دوبار مشتق‌پذیر g تعریف می‌شود، اسپلاین همواری نامیده می‌شود. واهبا (۱۹۹۰)[۸] نشان داد، کمین کننده مجموع مربعات جریمه‌ای (٦) به صورت

$$g_\lambda(t) = \sum_{j=1}^{\upsilon} a_j \phi_j(t) + \sum_{i=1}^{n} b_i \sigma_\alpha(|t - t_i|)$$

است، که در آن بردارهای $a = (a_1, \cdots, a_\upsilon)$ و $b = (b_1, \cdots, b_n)$ از حل دستگاه معادلات

$$\begin{cases} \Sigma_\lambda b + Ua = z \\ U^T b = 0 \end{cases} \qquad (٧)$$

به دست می‌آیند. که در آن $\Sigma_\lambda = \Sigma + n\lambda I$ ،Σ یک ماتریس $n \times n$ با عناصر $\sigma_\alpha(|t_i - t_j|)$ و U نیز یک ماتریس $n \times \upsilon$ با $\upsilon = \begin{pmatrix} d+m-1 \\ d \end{pmatrix}$ است، که به صورت $U = (\phi_j(t_i))$ تعریف می‌شود. برای حل دستگاه معادلات (٧)، فرض کنید تجزیه QR ماتریس U به صورت $U = (Q_1 : Q_2)\begin{pmatrix} R \\ O \end{pmatrix} = Q_1 R$ باشد، که در آن Q_1 یک ماتریس $n \times \upsilon$، Q_2 یک ماتریس $n \times (n - \upsilon)$، $Q = (Q_1 : Q_2)$ یک ماتریس متعامد و R یک ماتریس $\upsilon \times \upsilon$ مثلث بالایی است. در این صورت a و b را می‌توان از روابط زیر به دست آورد.

$$b = Q_2 (Q_2^T \Sigma_\lambda Q_2)^{-1} Q_2^T z$$
$$U_a = (I - \Sigma_\lambda Q_2 (Q_2^T \Sigma_\lambda Q_2)^{-1} Q_2^T) z$$

در روش اسپلاین مقدار پارامتر هموار ی λ نقش بسیار مهمی در برآورد منحنی g ایفا می‌کند. به طوری که اگر λ بزرگ باشد، مولفه اصلی در تابع $S(g,\lambda)$ جمله جریمه ناهمواری است و لذا کمین کننده g دارای انحنای کمتری خواهد بود. وقتی λ به بینهایت میل کند، منحنی $\hat{g}$ همان برازش منحنی یا صفحه رگرسیون خواهد شد. از سوی دیگر اگر λ نسبتا کوچک باشد، مجموع مربعات باقیمانده‌ها سهم اصلی را در $S(g,\lambda)$ دارا بوده و برآورد منحنی $\hat{g}$ تا حد زیادی روند داده‌ها را دنبال می‌کند. به همین سبب تعیین مقدار مناسب λ برای هر مجموعه داده‌ها مسئله‌ای حائز اهمیت است. به همین منظور روش‌های مختلف، از جمله روش اعتبار متقابل و اعتبار متقابل تعمیم یافته برای تعیین مقدار λ استفاده می‌شود. برای بررسی این روش‌ها می‌توان به اوبانک (۱۹۸۸)، واهبا (۱۹۹۰) و راسنبلات (۱۹۹۱) مراجعه نمود. محمدزاده (۱۹۹۸) الگوریتمی مناسب برای تعیین مقدار پارامتر همواری براساس مشاهدات معرفی کرد و کنت و محمدزاده (۲۰۰۰) نیز شیوه بهینه‌سازی ملاک اعتبار متقابل تعمیم‌یافته را برای به دست آوردن پارامتر همواری ارائه کردند.

تخمین‌گر عکس مجذور فاصله

تخمین‌گر عکس مجذور فاصله به صورت

$$\hat{Z}(t_0) = \sum_{i=1}^{n} w_i Z(t_i) \qquad (۸)$$

تعریف می‌شود، که در آن

$$w_i = w_i^{IDS} = \frac{(\frac{r_s - r_i}{r_i})^2}{\sum_{i=1}^{n} (\frac{r_s - r_i}{r_i})^2} \qquad (۹)$$

و r_i فاصله بین محل تخمین و محل iامین عضو نمونه و r_s شعاع همسایگی مورد بررسی است. تخمین‌گر عکس توان p ام فاصله نیز به همان فرم (۸) محاسبه می‌شود، با این تفاوت که در آن وزن‌ها به صورت

$$w_i^p = \frac{(\frac{1}{r_i})^p}{\sum_{i=1}^{n} (\frac{1}{r_i})^p} \qquad (۱۰)$$

هستند. برای مطالعه بیشتر این روش‌ها می‌توان به دیوید(۱۹۷۷)[۲] و اسوان و ساندی لاند(۱۹۹۵) مراجعه کرد.

مقایسه عددی

در این بخش با استفاده از تکنیک شبیه‌سازی داده‌های فضایی، تخمین‌گرهای ارائه‌شده ارزیابی می‌شوند. برای شبیه‌سازی داده‌های فضایی روش‌های متعددی وجود دارد. در این مقاله روش طیفی که کرسی(۱۹۹۳) ارائه کرده است، مورد استفاده قرارمی‌گیرد. شبیه‌سازی داده‌های فضایی به کمک نرم‌افزار کامپیوتری S-plus و براساس توزیع نرمال برای ۲۰۰ بار تکرار در سه حجم نمونه ۵۰، ۸۰ و ۱۰۰ انجام گرفته است. برای هر بار تکرارمجموع مربعات خطا محاسبه و میانگین آن‌ها به عنوان معیار مقایسه روش‌های مختلف واندازه‌های متفاوت نمونه مورد استفاده قرارگرفته است. مقادیر محاسبه شده برای دو تخمین‌گر کریگی و اسپلاین برای سه حجم نمونه مختلف در جدول ۱ آورده شده است. همان طور که ملاحظه می‌شود برای هر دو روش کریگی و اسپلاین با افزایش حجم نمونه مقدار MSE کاهش می‌یابد، اما تخمین‌گر کریگی نسبت به تخمین‌گر اسپلاین میانگین مربعات خطای کم‌تری دارد.

جدول ۱- میانگین مربعات خطاهای تخمین‌گرهای کریگی و اسپلاین

n	۱۰۰	۸۰	۵۰
اسپلاین	۰/۰۲۹	۰/۰۳۶	۰/۰۴۷
کریگیدن	۰/۰۲۱	۰/۰۲۸	۰/۰۳۰

برای داده‌های شبیه‌سازی شده میانگین مربعات خطای تخمین‌گرهای عکس مجذور فاصله و عکس توان p ام فاصله به ازای مقادیر مختلف r و p محاسبه و در جداول ۲ و۳ آورده شده‌اند.

مقادیر میانگین مربعات خطای تخمین‌گرهای عکس مجذور فاصله برای ۸ مقدار متفاوت r محاسبه شده‌اند. با افزایش مقدار r برای مقادیر کمتر از ۳، مقدار MSE کاهش یافته و پس از آن سیر صعودی داشته است. کمترین مقدار MSE به ازای r =۳ برای سه حجم نمونه ۵۰، ۸۰ و ۱۰۰ به ترتیب برابر ۰/۵۵۹، ۰/۲۲۵٦ و ۰/۲۰۲۳ به دست آمده‌اند، که با افزایش حجم نمونه مقدار آن کاهش نشان می‌دهد. برای تخمین‌گر عکس توان p ام فاصله نیز نتایجی مشابه عکس مجذور فاصله حاصل گردیده است. مقدار میانگین مربعات خطای تخمین‌گرهای عکس توان p ام فاصله نیز برای ۸ مقدارمختلف p محاسبه شده است. با افزایش مقدار میانگین مربعات خطای تخمین‌گرها برای $p \leq ۲$ کاهش و پس از آن افزایش یافته است. با مقایسه نتایج مندرج در جدول‌های ۱ تا ۳ تخمین‌گر کریگی برای تمام مقادیر نسبت به تخمین‌گرهای دیگر از MSE کمتری برخوردار است.

جدول ۲- میانگین مربعات خطای تخمین‌گر عکس مجذور فاصله

r	۱۰۰	۸۰	۵۰
۱	۰/۵۷۶	۰/۶۶۸	۰/۹۰۹
۲	۰/۲۲۳	۰/۳۶۰	۰/۸۹۵
۳	۰/۲۰۲	۰/۲۲۶	۰/۵۵۹
٤	۰/۳٤۳	۰/٤۵۹	۰/۶۸۱
۵	۰/۹۷٤	۱/۶۳۰	۱/۷۵٤
۶	۱/۷۵۶	۲/۷٤۰	۲/۹۸۰
۷	۲/۶۲۶	۲/۹۷۳	۳/۲۳۳
۸	۳/۰۰۱	۳/۰۳۶	۳/۳۵۶

جدول ۳- میانگین مربعات خطای تخمین‌گر عکس توان p ام فاصله

r	۱۰۰	۸۰	۵۰
۱	۰/٤۶۳	۰/۵۷۳	۰/۶۲۷
۲	۰/۲۰۲	۰/۲۱۳	۰/۲۲۲
۳	۰/۹۶۲	۰/۱۱۳	۰/۱۳۶
٤	۰/۲۰۷	۰/۲۵۷	۰/۱۶۸
۵	۰/۲۹٤	۰/۳۶۶	۰/۱۹۳
۶	۰/۸۶٤	۰/۹۲۱	۱/۶۱۳
۷	۱/۱۰۳	۱/۲۳٤	۱/۸٤۶
۸	۱/۵٤۳	۱/۸۲۶	۲/۱۳۱

نتیجه‌گیری

برآوردگرهای عکس فاصله نسبت به شعاع همسایگی و توان فاصله به کار رفته در وزن‌ها بسیار حساس می‌باشد. با کاهش شعاع همسایگی و افزایش توان فاصله، دقت برآورد افزایش می‌یابد. کریگیدن معمولی با به کار بردن مدل‌های تغییرنگار برازش داده شده، نسبت به تغییرنگار تجربی بکار گرفته شده در درون‌یابی دارای ثبات است. دقت کریگیدن با افزایش تعداد همسایگی‌ها بدون توجه به نوع داده‌ها، بهبود می‌یابد. یکی از نقاط ضعف تخمین‌گرهای عکس فاصله، عدم وجود روشی خاص برای تعیین مقادیر مناسب r و p است. تنها براساس نوع داده‌ها و مساله مورد بررسی و به صورت تجربی می‌توان مقادیر آن‌ها را تعیین نمود. لذا این نقیصه دامنه استفاده از این روش‌ها را در عمل کاهش می‌دهد. برای رهایی از این نقص استفاده از روش‌های اسپلاین و

کریگیدن برای درون‌یابی داده‌های فضایی توصیه می‌شوند. اما نتایج این تحقیق نشان دهنده آن است که کریگیدن به عنوان روشی برتر برای درون‌یابی داده‌های فضایی عمل می‌کند.

منابع

1. N. Cressie, *Statistics for Spatial Data*, Revised edition, John Wiley, New York(1993).
2. M. David, *Geostatistical Ore Reserve Estimation*, Elsevier Scintific Publishing Co.(1977) 364 .
3. R.L. Eubank, *Splines Smoothing and Nonparametric Regression*, Marcel Dekker, New York (1988).
4. B. Matron, The Theory of Regionalized Variables and its application, *Morphologie Mathematique*, No. 5, Fontainebleau, France(1971).
5. J.T. Kent, and M. Mohammadzadeh, Global Optimization of the Generalized Cross-Validation, *Statistics and Computing, Vol. 10,* (2000) *231-236.*
6. M. Mohammadzadeh, An Algorithm to Find the Smoothing Parameter in Smoothing Splines, Proceeding of the 4th Iranian Statistical Conference, Shahid Beheshti University, Tehran, Iran(1998).
7. M. Rossenblatt, *Stochastic Curve Estimation*, NSF- CBMS Regional Conference Series in Probability and Statistics, Vol 3, Institute of Mathematical Statistics, California(1991).
8. G. Wahba, *Spline Models for Observational Data*, Philadelphia: SIAM(1990).