انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1990

دانشگاه زنجان

مجموعه مقالات

سومین سمینار آمار فضایی و کاربردهای آن

۱-۲ آبان ماه ۱۳۹۸

به نام خدا

سومین سمینار آمار فضایی و کاربردهای آن

مجموعه مقالات

۱ الی ۲ آبان ۱۳۹۸

دانشگاه زنجان

مجموعه مقالات سومین سمینار آمار فضایی و کاربردهای آن

چاپ: اول آبان ماه ۱۳۹۸

آدرس دبیرخانه: زنجان، دانشگاه زنجان، دانشکده علوم، گروه آمار

تلفن: ۰۲۴ – ۳۳۰۵۴۰۴۲

پست الکترونیک: spatial3@znu.ac.ir

آدرس سایت: spatial3.znu.ac.ir

پیش‌گفتار

سپاس بی کران خدای منان را که هستی‌مان بخشید، به طریق علم و دانش رهنمونمان شد، به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه چینی از علم و دانش را روزیمان ساخت. ضمن عرض خیر مقدم به همه شرکت کنندگان گرامی و تشکر از حضور آنان، بسیار خرسندیم که در راستای خدمت به جامعه علمی آمار کشور، اعتلای علمی این مرز و بوم و به مناسبت گرامی‌داشت روز ملی آمار و برنامه‌ریزی، سومین سمینار آمار فضایی و کاربردهای آن را با همکاری انجمن آمار ایران و قطب علمی تحلیل داده‌های وابسته فضایی_زمانی دانشگاه تربیت مدرس و به میزبانی گروه آمار دانشگاه زنجان برگزار نماییم. امید است برنامه‌های سمینار که حاصل تلاش‌های پیوسته و موثر کمیته‌های علمی و اجرایی است، مورد رضایت واقع شود. تمام تلاش کمیته برگزار کننده بر آن بوده که فرصتی ایجاد شود تا همه پژوهشگران، متخصصان و علاقه‌مندان به این شاخه از علم آمار به ارائه آخرین یافته‌های پژوهشی خود پرداخته و زمینه توسعه و استفاده جدی‌تر این شاخه از آمار کاربردی را در سایر حوزه‌های علمی و اجتماعی فراهم آورند. امید است توانسته باشیم با لطف و عنایت پروردگار و همراهی و همدلی شرکت کنندگان گرامی، این سمینار را به نحو مطلوب برگزار کرده باشیم. با این وجود اگر کمبودهایی در برگزاری سمینار بوده است آن را بر ما ببخشایید و با اعلام آنها به کمیته برگزاری، امکان برگزاری بهتر سمینارهای آتی آمار فضایی و کاربردهای آن را فراهم آورید. در پایان از همه کسانی که ما را در برگزاری سمینار یاری نموده‌اند کمال سپاس و امتنان را داریم. بخصوص از داوران گرامی، قطب علمی تحلیل داده‌های وابسته فضایی_زمانی دانشگاه تربیت مدرس و همکاران هیات علمی، دانشجویان گروه آمار و معاونت محترم پژوهشی، ریاست محترم دانشگاه و کارکنان مختلف دانشگاه زنجان که برگزاری این سمینار بدون مساعدت‌های بی‌دریغ و همه جانبه آن‌ها امکان‌پذیر نبود، قدردانی و تشکر می‌نماییم. امید است با برگزاری این سمینار گامی هر چند کوچک در راستای ارتقای دانش و تجربه در حوزه آمار فضایی کشور برداشته باشیم.

دبیر سومین سمینار آمار فضایی و کاربردهای آن

دانشگاه زنجان

۱ الی ۲ آبان ۱۳۹۸

فهرست همکاران سمینار:

انجمن آمار ایران

قطب علمی تحلیل داده‌های وابسته فضایی_زمانی

فهرست حامیان سمینار:

اعضای کمیته علمی

اعضای کمیته اجرایی

فهرست مقالات

عنوان _____ صفحه

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

مدل‌بندی داده‌های بقای فضایی با تابع مفصل زوجی کلایتون

نسرین ابراهیمی[۱]**، محسن محمدزاده**
گروه آمار، دانشگاه تربیت مدرس

چکیده: بررسی تأثیر عوامل مخاطره مختلف بر زمان بقای آزمودنی‌ها همواره مورد توجه محققن در حوزه تحلیل داده‌های بقا بوده است. هنگامی که بین زمان‌های بقا وابستگی وجود داشته باشد، این وابستگی به عنوان یک عامل مخاطره پنهان عمل می‌کند و بررسی تاثیر آن بر زمان‌های بقا از طریق مدل‌های رگرسیونی معمول بقا نظیر مدل خطرهای متناسب کاکس امکان‌پذیر نیست. یکی از روش‌های مدل‌بندی زمان‌های بقا با لحاظ کردن وابستگی بین آن‌ها، استفاده از توابع مفصل است. در این مقاله تابع بقا با استفاده از توابع بقای توأم دو به دو از طریق تابع مفصل زوجی کلایتون مدل می‌شود. تابع درستنمایی مرکب برای براورد پارامترهای وابستگی به کار می‌رود و عملکرد مدل معرفی‌شده از طریق مطالعه شبیه‌سازی مورد بررسی قرار می‌گیرد.

واژه‌های کلیدی: داده‌های بقای فضایی، تابع بقای توأم، تابع درستنمایی مرکب، مدل‌های حاشیه‌ای، تابع مفصل کلایتون.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62N01.

۱ مقدمه

هدف عمده بسیاری از مطالعات صورت گرفته در حوزه تحلیل داده‌های بقا، تعیین تاثیر عوامل مخاطره مختلف روی زمان بقا است. در بسیاری از موارد، عوامل خطر ناشناخته یا غیر قابل اندازه‌گیری نیز روی زمان بقا تاثیرگذار هستند و موجب وابستگی بین زمان‌های بقا می‌شوند که نادیده گرفتن این وابستگی در مدل، نتایج گمراه‌کننده‌ای را به همراه خواهد داشت. یکی از این عوامل ناشناخته تاثیرگذار بر داده‌های بقا، موقعیت فضایی مشاهده مورد نظر است، به‌طوری که هر داده بقا برحسب موقعیت خود در فضای مورد مطالعه، با داده‌های دیگر وابستگی فضایی دارد. اما این اثر فضایی به‌طور مستقیم قابل مشاهده یا شناسایی نیست. تحلیل و بررسی داده‌های بقای فضایی همواره مورد توجه محققین علوم زیستی در مطالعات همه‌گیرشناسی و آزمایش‌های بالینی خانوادگی بوده است. تحلیل داده‌های بقای فضایی مستلزم مدل‌بندی تابع بقای توأم این داده‌هاست. در حالت کلی دو رهیافت عمده برای مدل‌بندی داده‌های بقای فضایی با لحاظ کردن وابستگی آن‌ها، استفاده از

[۱] نام و ایمیل ارائه دهنده مقاله: نسرین ابراهیمی، nasrin.ebrahimi@modares.ac.ir

مدل‌های شکنندگی و توابع مفصل است. مدل‌های شکنندگی، وابستگی فضایی بین آزمودنی‌ها را در قالب یک اثر تصادفی پنهان به نام شکنندگی در مدل لحاظ می‌کنند. برآورد پارامترهای حاصل از این مدل، دارای تفسیرهایی به ازای یک مقدار خاص شکنندگی هستند و در مواردی که استنباط‌های درون‌خوشه‌ای مدنظر باشند مورد استفاده قرار می‌گیرند.

لی و ریان (۲۰۰۲) مدل بقای فضایی را با اضافه کردن اثرات تصادفی فضایی به مدل نیم‌پارامتری خطرهای متناسب معرفی کردند و این مدل را در برازش به داده‌های مربوط به آسم کودکان به کار بردند. لی و ریان (۲۰۰۲) یک میدان تصادفی گاوسی را به عنوان توزیع اثرات تصادفی فضایی در نظر گرفتند. مترجم و همکاران (۱۳۹۴) این مدل را برای میدان تصادفی ناگاوسی مورد مطالعه قرار دادند و شناسایی‌پذیری پارامترهای مدل را در این حالت اثبات کردند. بانرجی و دی (۲۰۰۵) از تکنیک مدل‌بندی شکنندگی در قالب مدل نیم‌پارامتری بخت‌های متناسب استفاده کردند. پن و همکاران (۲۰۱۴) مدل نیم‌پارامتری خطرهای متناسب را با شکنندگی‌های CAR برای داده‌های سانسوریده بازه‌ای به کار بردند. یکی از ابزارهای قوی برای لحاظ کردن وابستگی داده‌ها در مدل استفاده از توابع مفصل است. این توابع مدلی را برای بررسی ساختار ارتباط بین متغیرها ارائه می‌دهند که بر اساس آن تمام خصوصیات وابستگی میان متغیرها قابل تبیین است (امیدی و محمدزاده ، ۱۳۹۰). با توجه به این ویژگی، می‌توان توابع مفصل را برای مدل کردن تابع بقای توأم بر اساس توابع بقای حاشیه‌ای مفروض به کار برد. کلایتون (۱۹۷۸) اولین بار یک مدل پیوند دومتغیره را در تحلیل داده‌های بقای وابسته به کار برد. در مدل پیشنهادی او یک تابع مفصل برای مدل‌بندی زمان‌های بقای دومتغیره وابسته به کار رفته است. امیدی و محمدزاده (۱۳۹۰) و شیائو (۲۰۰۶) از توابع مفصل برای مدل‌بندی داده‌های خشکسالی استفاده کردند. لی و لین (۲۰۰۶) با فرض آنکه زمان‌های بقا به‌طور حاشیه‌ای از مدل نیم‌پارامتری خطرهای متناسب پیروی می‌کنند، از تابع مفصل گاوسی برای مدل‌بندی ساختار همبستگی فضایی داده‌های بقای زمین‌آماری با یک ساختار همبستگی خاص استفاده کردند. ژو و همکاران (۲۰۱۵) از رهیافت بیز ناپارامتری برای مدل‌بندی تابع خطر داده‌های بقای زمین‌آمار استفاده کردند و تابع مفصل گاوسی را برای مدل‌بندی ساختار همبستگی فضایی آن‌ها به کار بردند. همچنین نسخه بیزی مدل معرفی شده توسط لی و لین (۲۰۰۶) را با در نظر گرفتن پیشین تکه‌ای‌نمایی برای تابع خطر پایه، ارائه کردند. پرنن و همکاران (۲۰۱۷) روش‌شناسی توابع مفصل ارشمیدسی را برای مدل‌بندی داده‌های بقای خوشه‌بندی شده بسط دادند.

در این مقاله با در نظر گرفتن مدل حاشیه‌ای خطرهای متناسب برای زمان‌های بقای وابسته فضایی، از تابع مفصل کلایتون برای مدل‌بندی ساختار همبستگی آن‌ها استفاده می‌شود و یک روش دو مرحله‌ای برای برآورد پارامترهای مدل به کار می‌رود که در آن ابتدا پارامترهای رگرسیونی مدل به روش ماکسیمم درستنمایی جزیی برآورد می‌شوند و سپس در برآورد پارامتر ساختار همبستگی فضایی داده‌ها با استفاده از تابع درستنمایی مرکب به کار می‌روند.

۲ معرفی مدل

فرض کنید به ازای $i = ۱, \cdots, n$ متغیرهای T_i و C_i به ترتیب زمان‌های شکست و سانسور را برای آزمودنی iام نشان دهند. در عمل T_i و C_i به‌طور بالقوه مشاهده نشده‌اند و داده‌های مشاهده شده عبارتند از: $X_i = \min(T_i, C_i)$ و $\Delta_i = I(T_i \leq C_i)$. فرض کنید $Z_i(t)$ بردار متغیرهای کمکی با بعد p باشد. همچنین مسیر[۱] متغیرهای کمکی تا زمان t با $\bar{Z}_i(t) = \{Z_i(s) | ۰ \leq s \leq t\}$ نمایش داده شود و فرض کنید T_i به شرط $\bar{Z}_i(T_i)$ مستقل از C_i است. همچنین به شرط معلوم بودن $\bar{Z}_i(t)$، تابع نرخ خطر به صورت

$$\lambda(t|\bar{Z}_i(t)) = \lambda_۰(t) \exp(\beta' Z_i(t)) \qquad (۱.۲)$$

[۱] Path

از مدل خطرهای متناسب پیروی می‌کند. معادله (۱.۲) یک مدل حاشیه‌ای برای هر T_i است و از آن‌جا که ضریب رگرسیونی β برای تمام i ها ثابت در نظر گرفته شده است، دارای تفسیری در سطح جامعه است. برای مدل‌بندی ساختار وابستگی T_iها، فرض کنید که تابع توزیع توأم $F(t_i,t_j)$ با استفاده از توابع توزیع حاشیه‌ای $F(t_i)$ و $F(t_j)$ از طریق تابع مفصل کلایتون به صورت

$$F(t_i,t_j)=C(F(t_i),F(t_j))=\left(F(t_i)^{-\alpha_{ij}}+F(t_j)^{-\alpha_{ij}}-1\right)^{-1/\alpha_{ij}} \qquad (2.2)$$

حاصل شود، که در آن $\alpha_{ij}\in(0,\infty)$ پارامتر تابع مفصل کلایتون است که میزان وابستگی دو متغیر T_i و T_j را اندازه می‌گیرد. نلسن (۲۰۰۵) ارتباط بین تابع مفصل بقا و تابع مفصل را در حالت کلی به‌صورت

$$S(t_i,t_j)=S(t_i)+S(t_j)-1+C(F(t_i),F(t_j)) \qquad (3.2)$$

بیان می‌کند. بنابراین برای تابع مفصل کلایتون، تابع مفصل بقا یا به عبارت دیگر همان تابع بقای توأم به صورت

$$S(t_i,t_j)=S(t_i)+S(t_j)-1+\left(F(t_i)^{-\alpha_{ij}}+F(t_j)^{-\alpha_{ij}}-1\right)^{-1/\alpha_{ij}} \qquad (4.2)$$

است، که در آن $S(t_i)$ و $S(t_j)$ به ترتیب توابع بقای حاشیه‌ای T_i و T_j هستند که براساس مدل حاشیه‌ای خطرهای متناسب برآورد می‌شوند. بین پارامتر تابع مفصل کلایتون و ضریب همبستگی τـ کندال رابطه یک به یک به صورت $\alpha=2\tau/(1-\tau)$ وجود دارد (نلسن ، ۲۰۰۵). در این مقاله ضریب همبستگی τـ کندال بین واحدهای iام و jام، تابعی نمایی از فاصله اقلیدسی بین دو واحد به صورت

$$\tau_{ij}\equiv\varphi(d_{ij};\psi)=\exp(-\frac{d_{ij}}{\psi}) \qquad (5.2)$$

در نظر گرفته شده است، که در آن d_{ij} فاصله اقلیدسی بین موقعیت‌های فضایی واحدهای iام و jام و ψ پارامتر وابستگی فضایی مدل است. برای برآورد مدل باید ضرایب رگرسیونی مدل حاشیه‌ای، β، و پارامتر وابستگی فضایی، ψ، برآورد شوند.

۳ برآورد مدل

برای برآورد پارامترهای مدل، از یک روش دو مرحله‌ای استفاده می‌شود. ابتدا ضرایب رگرسیونی مدل حاشیه‌ای خطرهای متناسب با روش ماکسیمم درستنمایی جزئی برآورد می‌شوند. سپس برآوردهای حاصل از این مرحله در تابع درستنمایی مرکب قرار می‌گیرند و در برآورد پارامتر وابستگی فضایی ψ به کار می‌روند.

مرحله اول: برآورد ضرایب رگرسیونی

در مدل رگرسیونی خطرهای متناسب کاکس هنگام برآورد پارامتر β تابع خطر پایه به‌عنوان یک پارامتر مزاحم نامتناهی‌ـ بعد ظاهر می‌شود، بنابراین روش‌های درستنمایی معمول نمی‌توانند در برآورد پارامتر رگرسیونی به کار روند. کاکس (۱۹۷۵) استنباط مبتنی بر تابع درستنمایی جزئی را در برآورد مدل‌های شامل پارامتر مزاحم نامتناهی‌ـ بعد معرفی کرد. برآورد ماکسیمم درستنمایی جزئی پارامتر β با ماکسیمم کردن تابع درستنمایی جزئی

$$L_p=\prod_{i=1}^{n}\left[\frac{\exp(\beta' Z_i(t))}{\sum_{j\in\Re(X_i)}\exp(\beta' Z_j(t))}\right]^{\Delta_i} \qquad (1.3)$$

نسبت به β حاصل می‌شود، که در آن $\Re(X_i)$ مجموعه واحدهای در مخاطره تا زمان t است.

مرحله دوم: برآورد پارامتر وابستگی فضایی

برای برآورد پارامتر وابستگی مدل، از تابع درستنمایی مرکب استفاده شده است. تابع درستنمایی مرکب برای داده‌های بقا در حالت کلی به صورت

$$LC = \prod_{i \leq j} S(t_i, t_j)^{(1-\delta_i)(1-\delta_j)} S^{(1)}(t_i, t_j)^{\delta_i(1-\delta_j)} S^{(2)}(t_i, t_j)^{(1-\delta_i)\delta_j} f(t_i, t_j)^{\delta_i \delta_j}. \qquad (۲.۳)$$

است، که در آن $S^{(1)}(t_i, t_j) = \frac{\partial}{\partial t_i} S(t_i, t_j)$، $S^{(2)}(t_i, t_j) = \frac{\partial}{\partial t_j} S(t_i, t_j)$ و $f(t_i, t_j) = \frac{\partial^2}{\partial t_i \partial t_j} S(t_i, t_j)$ است. برای مدل معرفی‌شده، با قرار دادن برآورد پارامتر رگرسیونی حاصل از مرحله قبل در تابع درستنمایی مرکب و ماکسیمم کردن آن نسبت به ψ، برآورد ماکسیمم درستنمایی پارامتر وابستگی حاصل می‌شود.

۴ مطالعه شبیه‌سازی

برای شبیه‌ساری داده‌های بقای فضایی، ابتدا با فرض معلوم بودن پارامتر وابستگی فضایی، ۱۰۰ موقعیت فضایی از توزیع یکنواخت تولید و فاصله اقلیدسی این نقاط در قالب ماتریس فاصله محاسبه شد و با در نظر گرفتن مدل کواریوگرام نمایی به صورت

$$C(d) = \exp(-\frac{d}{\psi}) \qquad (۱.۴)$$

ماتریس کوواریانس مشاهدات به دست آمد. برای تولید داده بقا از مدل خطرهای متناسب، با استفاده از رابطه

$$S(t|Z) = \exp\left(-\Lambda_0(t)\exp(\beta' Z)\right) \qquad (۲.۴)$$

در مدل خطرهای متناسب و با توجه به قضیه تبدیل انتگرال احتمال، می‌توان نتیجه گرفت که متغیر تصادفی

$$T = \Lambda_0^{-1}\left(-\frac{\ln S}{\exp(\beta' Z)}\right) \qquad (۳.۴)$$

دارای تابع بقای (۲.۴) خواهد بود. حال با تولید ۱۰۰ مشاهده از توزیع $U(0,1)$ به عنوان زمان بقای S و در نظر گرفتن هر تابع خطر پایه دلخواهی می‌توان مشاهداتی از مدل خطرهای متناسب شبیه‌سازی کرد. در این مقاله زمان‌های بقا با فرض تابع خطر پایه وایبل با پارامترهای $\lambda = 0/01$ و $\rho = 1$ با استفاده از رابطه

$$t|z = \left(\frac{-\ln S}{\lambda \exp(\beta' Z)}\right)^{\frac{1}{\rho}}$$

شبیه‌سازی شدند. با ضرب تجزیه چولسکی ماتریس کواریانس محاسبه شده در رابطه (۱.۴) در داده‌های بقای مستقل شبیه‌سازی شده و سپس حذف روند، داده‌های بقای فضایی با ساختار کواریانس موردنظر حاصل می‌شوند. متغیر تبیینی Z از توزیع برنولی شبیه‌سازی شد و نرخ سانسوری در حدود r درصد برای داده‌ها در نظر گرفته شد. این روند شبیه‌سازی ۱۰۰ بار تکرار شد. در هر بار تکرار برآورد ماکسیمم درستنمایی پارامتر β محاسبه شد. نتایج این شبیه‌سازی در جدول ۱ مشاهده می‌شود. همان‌طور که ملاحظه می‌شود، برآوردهای حاصل از این روش برای پارامتر رگرسیونی β از دقت مناسبی برخوردار هستند. همان‌طور که انتظار می‌رود، با افزایش نرخ سانسور، دقت برآوردها تا حدودی کاهش می‌یابد. نکته قابل

جدول ۱: مقادیر واقعی پارامتر β به همراه برآورد و MSE آن

		$r = ۰/۰۵$		$r = ۰/۲$		$r = ۰/۵$	
مقدار معلوم ψ	مقدار واقعی β	برآورد β	MSE	برآورد β	MSE	برآورد β	MSE
۰/۱	۰/۵	۰/۴۸۸۵	۰/۰۴۵۲	۰/۵۲۵۴	۰/۰۵۷۷	۰/۵۲۷۲	۰/۱۰۲۸
۰/۵	۰/۵	۰/۴۹۸۳	۰/۰۳۷۵	۰/۵۳۲۶	۰/۰۴۸۶	۰/۴۹۸۲	۰/۰۹۱
۲	۰/۵	۰/۴۸۲۷	۰/۰۴۷۳	۰/۴۷۷۶	۰/۰۵۵۸	۰/۵۵۱۶	۰/۱۲۵
۰/۱	۱/۵	۱/۵۷۸۳	۰/۰۸۴۹	۱/۵۰۶۶	۰/۰۹۴۳	۱/۵۱۶۱	۰/۱۱۸۶
۰/۵	۱/۵	۱/۵۲۱۶	۰/۰۵۵۴	۱/۵۱۳۳	۰/۰۸۹۸	۱/۵۴۰۱	۰/۱۰۲
۲	۱/۵	۱/۳۸۱۱	۰/۰۹۹۶	۱/۴۰۹	۰/۱۰۱	۱/۴۴۰۵	۰/۰۸۹
۰/۱	۳	۳/۱۱۸۶	۰/۱۵۰۹	۳/۰۷۸۷	۰/۱۹۸۲	۳/۱۲۴۹	۰/۳۷۳۹
۰/۵	۳	۳/۰۱۶۲	۰/۲۹۱	۲/۹۸۴۴	۰/۲۴۹۶	۲/۹۲۶۲	۰/۳۶۰۸
۲	۳	۲/۳۰۷۲	۰/۶۵۴۳	۲/۲۹۹۸	۰/۷۱۱۸	۲/۱۷۹۸	۰/۸۷۵۶

توجه اینکه با افزایش مقدار پارامتر وابستگی، دقت برآوردهای β کاهش پیدا می‌کند. به عبارت دیگر هنگامی که وابستگی فضایی مشاهدات افزایش پیدا می‌کند، دقت برآورد حاصل از مدل حاشیه‌ای برای پارامتر رگرسیونی کاهش پیدا می‌کند.

نتایج شبیه‌سازی در حالت نامعلوم بودن پارامتر وابستگی فضایی، به دلیل وجود اشکال در برنامه محاسبه برآوردها ارائه نشده‌اند.

۵ بحث و نتیجه‌گیری

در مدل‌بندی داده‌های بقای وابسته فضایی، علاوه بر برآورد میزان تأثیر متغیرهای تبیینی مختلف بر تابع خطر، برآورد میزان وابستگی داده‌های بقا نیز مورد توجه است. در این مقاله، یک روش برآورد دو مرحله‌ای برای این منظور معرفی شد که بر اساس آن میزان تأثیر متغیرهای تبیینی بر تابع خطر، با استفاده از مدل حاشیه‌ای خطرهای متناسب برآورد می‌شود و از تابع مفصل زوجی کلایتون برای لحاظ کردن وابستگی داده‌ها و سپس برآورد پارامتر وابستگی استفاده می‌شود. این روش محدودیتی در نوع تابع کوواریانس فضایی مشاهدات ندارد. کافی است تابع درستنمایی مرکب (۲.۳) نسبت به پارامترهای تابع کوواریانس، ماکسیمم شود. نتایج مطالعه شبیه‌سازی، بیانگر عملکرد مناسب این روش در برآورد ضرایب رگرسیونی و پارامتر وابستگی بود.

مراجع

امیدی، م.، محمدزاده ، م. (۱۳۹۰)، مدل بندی درست‌نمایی و بیز تجربی خشکسالی با استفاده از تابع مفصل، *نشریه علوم دانشگاه خوارزمی*، ۱۰، ۶۵۴ـ۶۴۳.

مترجم ، ک.، محمدزاده ، م. و آبیار ، آ. (۱۳۹۴)، مدل‌بندی فضایی داده‌های بقای سانسور شده، *پژوهش های ریاضی*، ۱، ۷۰ـ۶۱.

Banerjee S. , Dey D. K. (2005), Semiparametric Proportional Odds Models for Spatially Correlated Survival Data, *Lifetime Data Analysis*, **11**, 175–191.

Clayton D. (1978), A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence, *Biometrika*, **65**, 141–151.

Cox DR. (1975), Partial Likelihood, *Biometrika*, **62**, 269-276.

Li Y. , Lin X. (2006), Semiparametric Normal Transformation Models for Spatially Correlated Survival Data, *Journal of the American Statistical Association*, **101**, 591–603.

Li Y. , Ryan L. (2002), Modeling Spatial Survival Data Using Semiparametric Frailty Models, *Biometrics*, **58**, 287–297.

Nelsen R.B. (2005), *An Introduction to Copulas*. Second Edition. Springer Series in Statistics .

Pan C., Cai B., Wang L., Lin X. (2014), Bayesian SemiParametric Model for Spatial Interval-Censored Survival Data, *Computational Statistics and Data Analysis*, **74**, 198–209.

Prenen L., Braekers R. (2017), Extending the Archimedean Copula Methodology to Model Multivariate Survival Data Grouped in Clusters of Variable Size, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 483-505.

Shiau J. T. (2006), Fitting Drought Duration and Severity with Two-Dimensional Copulas, *Water Resources Research*, **20**, 795-815.

Zhou H., Hanson T., Knapp R. (2015), Marginal Bayesian Nonparametric Model for Time to Disease Arrival of Threatened Amphibian Populations, *Biometrics*, **71**, 1101-1110.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

پیشگویی فضایی میدان تصادفی ناگاوسی با استفاده از قضیه تصویر

اسحاق الماسی[۱]**، مهدی امیدی**
گروه ریاضی، دانشگاه ایلام

چکیده: یکی از مهمترین مسائل آمار فضایی تعیین توزیع تحقق‌های میدان تصادفی به منظور پیشگویی در نقاط نامعلوم از میدان تصادفی است. فرض گاوسی بودن میدان باعث می‌شود استفاده از پیشگوهای خطی توجیه‌پذیر باشد. نامشخص بودن توزیع میدان تصادفی محققین را ملزم به استفاده از روش‌هایی می‌کند که بر اساس آن‌ها امکان پیشگویی ناگاوسی میسر شود. در این مقاله یک روش ناپارامتری را بر اساس قضیه تصویر برای پیشگویی ارایه می‌دهیم و بر اساس آن مدلی برای پیشگویی در میدان‌های ناگاوسی مطرح می‌شود. در ادامه با استفاده از مطالعه شبیه‌سازی میزان دقت و صحت این روش مورد بررسی قرار می‌گیرد.

واژه‌های کلیدی: آمار فضایی، پیشگوی ناگاوسی، قضیه تصویر، آماره‌های مرتب.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 91B72.

۱ مقدمه

در آمار فضایی پیشگویی مقدار میدان تصادفی در مکان‌های فاقد مشاهده معمولاً با فرض گاوسی بودن میدان تصادفی انجام می‌شود. وجود این فرض باعث می‌شود که استفاده از پیشگوهای خطی مانند کریکینگ (کرسی ، ۱۹۹۳) امکان‌پذیر باشد. در برخی شرایط ممکن است واقع‌گرایانه نباشد از این‌رو باید از روش‌هایی استفاده نمود که بتوان برای میدان‌های ناگاوسی پیشگویی انجام داد. یکی از روش‌هایی که تاکنون در پیشگویی میدان‌های ناگاوسی فضایی مطرح شده استفاده از تابع مفصل (امیدی و محمدزاده ، ۲۰۱۸) است، بر اساس این روش با استفاده از محاسبات پیچیده توزیع میدان تصادفی به‌دست می‌آید اما محدودیت‌هایی از قبیل تعیین توزیع حاشیه‌ای و نسبی بودن تابع توزیع به‌دست آمده در آن مطرح است. براکول و داویس (۱۹۹۱) پیش بینی در سری‌های زمانی را با استفاده از قضیه تصویر مطرح کردند، الماسی و همکاران (۲۰۱۷) با استفاده از قضیه تصویر روشی را برای پیشگویی آماره‌های ترتیبی سانسور شده معرفی کرند. در این مقاله با استفاده از روش مطرح شده توسط الماسی و همکاران (۲۰۱۷) بر اساس نزدیکترین همسایه به نقاط گمشده فضایی ترکیب خطی

[۱] نام و ایمیل ارائه دهنده مقاله: اسحاق الماسی، i.almasi@ilam.ac.ir

ناپارامتری برای پیشگویی فضایی نقاط گمشده ارائه می‌شود. در ادامه با استفاده از شبیه‌سازی دقت و صحت این روش با استفاده از معیار RMSE مورد بررسی قرار می‌گیرد.

۲ ساختار همبستگی میدان تصادفی

در آمار فضایی مقدار میدان تصادفی $Z = \{Z(s); s \in D \subset R^d\}$ در موقعیت مشخص s، بر اساس تحقق‌های میدان تصادفی $s_1, \ldots, s_n$، یعنی $z = (z(s_1), \ldots, z(s_n))$ پیشگویی می‌شود. توابع میانگین و کواریانس فضایی میدان تصادفی به‌ترتیب به‌صورت

$$\begin{aligned} \mu(s) &= E(Z(s)); \ \ (s) \in D \\ C(s, s') &= Cov(Z(s), Z(s')); \ \ (s, s') \in D \end{aligned}$$

تعریف می‌شوند. میدان تصادفی فضایی‌ـزمانی $Z(\cdot)$ مانای مرتبه دوم است هرگاه میانگین آن نسبت به (s) ثابت باشد، یعنی $\mu(s) = \mu$ و کواریانس فضایی‌ـزمانی تابعی از فاصله موقعیت‌های فضایی

$$Cov(Z(s, s')) = C(\mathbf{h_s}); \ \ (s, s') \in D$$

باشد، که در آن $\mathbf{h_s} = \mathbf{s} - \mathbf{s'}$ تأخیر فضایی نامیده می‌شود. اگر تابع کواریانس فقط تابعی از اندازه تأخیرها بوده و به جهت آنها بستگی نداشته باشد، همسانگرد نیز نامیده می‌شود.

تغییرنگار فضایی‌ـزمانی به‌صورت $۲\gamma(s, s') = Var[Z(s) - Z(s'')]$ تعریف می‌شود. چنانچه میانگین ثابت و تغییرنگار تابعی از تأخیرهای فضایی و زمانی باشد، یعنی $Var[Z(s) - Z(s')] = ۲\gamma(\mathbf{h_s})$، میدان مانای ذاتی نامیده می‌شود. تحت مانایی ضعیف تابع تغییرنگار به‌صورت

$$۲\gamma(\mathbf{h_s}) = ۲[C(۰) - C(\mathbf{h_s})]$$

با تابع کواریانس فضایی‌ـزمانی در ارتباط است. توابع کواریانس فضایی ساختار همبستگی داده‌ها فضایی را تفسیر نموده و خواص آن‌ها به گونه‌ای است که همواره اندازه همبستگی مثبت را به‌دست می‌دهند و مقدار آن مشاهداتی با فاصله نزدیک حداکثر و برای فواصل دورتر مقدار آن کم می‌شود. امیدی و محمدزاده (۲۰۱۶) روش‌هایی برای ساخت توابع کواریانس همیشه مثبت ارائه دادند که به‌عنوان مثال تابع

$$C(\mathbf{h_s}) = \sigma^{\mathbf{2}} \exp(-\alpha \mathbf{h_s}^{\beta}) \qquad (۱.۲)$$

برای σ و α مثبت و $\beta \in [۰, ۲]$ یک تابع کواریانس فضایی معتبر است.

۳ پیشگویی با قضیه تصویر

بر اساس قضیه تصویر بهترین پیش بینی خطی $Z = \{Z(s); s \in D \subset R^d\}$ بر اساس مشاهدات $s_1, \ldots, s_n$، یعنی $z = (z(s_1), \ldots, z(s_n))$ به صورت زیر

$$\hat{Z}(s_l) = \sum_{i=۰}^{n} a_i z(s_i)$$

است، که در آن ضرایب ترکیب خطی از حل همزمان معادلات

$$\sum_{i=۰}^{n} a_i E(Z(s_i)Z(s_j) = E(Z(s_l)Z(s_j), \quad j = ۰, \ldots, n$$

به‌دست می‌آیند. اگر $d_{ik} = d(z(s_i), z(s_k)), \quad i, k = ۱, \ldots, n$ فاصله اقلیدسی مشاهدات iـام و kـام و بردار $d_{(۱)}, \ldots, d_{(n)}$ مقادیر مرتب شده فواصل اقلیدسی $d_{il}, i = ۱, \ldots, n$ باشند، آنگاه بر اساس قضیه تصویر، بهترین پیشگوی خطی در مکان $z(s_l)$ به‌صورت

$$\hat{Z}(s_l) = a_۰ + a_r z(s_r) + a_k z(s_k) \qquad (۱.۳)$$

به‌دست می‌آید، که در آن $a_۰$، a_r و a_k از طریق حل همزمان معادلات

$$\begin{cases} a_۰ + a_r E(Z(s_r)) + a_k E(Z(s_k)) = E(Z(s_l)), \\ a_۰ E(Z(s_r)) + a_r E(Z(s_r)^۲) + a_s E(Z(s_r)Z(s_k)) = E(Z(s_r)Z(s_l)), \\ a_۰ E(Z(s_k) + a_r E(Z(s_r)Z(s_k)) + a_k E(Z(s_k)^۲) = E(Z(s_l)Z(s_k)). \end{cases} \qquad (۲.۳)$$

به‌دست می‌آید. با در نظر گرفتن فاصله از دو مشاهده از نزدیکترین مکان‌ها به نقطه فاقد مشاهده، ضرایب به‌صورت $a_۰ = E(Z(s_l)) - a_r E(Z(s_r)) - a_k E(Z(s_k))$ و

$$a_r = \frac{\rho(d_{(۱)}) - \rho(d_{kr})\rho(d_{(۲)})}{۱ - \rho^۲(d_{kr})}, \quad a_k = \frac{\rho(d_{(۲)}) - \rho(d_{kr})\rho(d_{(۱)})}{۱ - \rho^۲(d_{kr})}$$

برآورد می‌شوند، که در آن $\rho(d_{kr}) = Corr(z(s_r), z(s_k))$ مقدار همبستگ فضایی میان $z(s_r)$ و $z(s_k)$ است.

۴ مطالعه شبیه‌سازی

در این بخش با استفاده از تابع کواریانس (۱.۲) یک میدان تصادفی فضایی به‌ازای تعداد مکان‌های مختلف تولید می‌شود. در ادامه برای مقادیر مختلف α و β صحت و دقت برآورد رابطه (۱.۳) در ۵۰، ۱۰۰، ۵۰۰ و ۱۰۰۰ مکان با استفاده از معیار کارایی نسبی RMSE مورد بررسی قرار می‌گیرد. به منظور کنترل اثر تصادفی بودن داده‌های تولید شده در دقت برآوردگر روند اجرای برنامه ده بار تکرار شده و میانگین RMSE به‌دست آمده در ۱۰ بار مورد بررسی قرار گرفت. در جدول ۱، میانگین RMSE را در ده بار تکرار مجزای پیشگویی فضایی نقاط تولید شده آمده است. هماطور که مشاهده می‌شود نمی‌توان اظهار داشت که افزایش تعداد نقاط و مقادیر α و β تاثیری در برآوردگر داشته‌است.

بحث و نتیجه‌گیری

در این مقاله، با استفاده از قضیه تصویر یک روش ناپارامتری برای پیش گویی در میدان‌های تصادفی ناگاوسی مطرح شد. در ادامه با استفاده از مطالعه شبیه‌سازی میزان دقت و صحت این روش بر اساس معیار RMSE مورد بررسی قرار می‌گیرد. نتایج نشان داد تغییرات در مقادیر پارامترهای تابع کواریانس فضایی و همچنین افزایش تعداد مکان‌ها تاثیری بر دقت برآوردگر ندارد.

اگرچه در این مقاله برآوردگری برای پیشگویی ناگاوسی مطرح شد اما این روش بر اساس نزدیکترین دو همسایه به‌دست آمد که خود می‌تواند از نقاط ضعف برآوردگر تلقی گردد. از این‌رو به‌دست آوردن برآوردگر برای همسایه‌های بیشتر و مقایسه این روش با پیشگوی کرکینگ و تابع مفصل فضاییاز موضوعات مهم در مبحث پیشگویی فضایی است.

جدول ۱: میانگین RMSE پیشگویی‌ها

تعداد مکان‌ها	β	α ۱	۵	۱۰	۲۰
۵۰	۰/۵	۰/۸۸۸	۰/۹۴۵	۰/۹۴۱	۰/۹۳۲
	۱	۰/۹۰۹	۰/۸۰۱	۰/۸۸۷	۰/۹۳۲
	۱/۵	۱/۰۱۵	۰/۹۶۷	۰/۹۶۰	۰/۹۴۹
۱۰۰	۰/۵	۰/۸۶۸	۰/۸۹۵	۰/۹۹۲	۰/۹۴۹
	۱	۱/۰۹۳	۰/۹۰۳	۰/۸۹۲	۰/۹۵۳
	۱/۵	۰/۸۶۷	۱/۰۱۹	۰/۹۷۶	۰/۹۶۹
۵۰۰	۰/۵	۰/۹۸۷	۰/۹۱۹	۰/۹۸۴	۰/۹۹۱
	۱	۰/۹۹۵	۱/۰۷۷	۱/۰۴۳	۰/۹۴۶
	۱/۵	۱/۱۵۲	۱/۱۴۲	۱/۱۸۳	۱/۰۹۷
۱۰۰۰	۰/۵	۱/۱۴۲	۰/۹۳۵	۰/۹۷۹	۰/۹۹۷
	۱	۱/۳۳۷	۱/۲۳۳	۱۱/۰۱	۰/۹۸۵
	۱/۵	۰/۸۶۶	۱/۲۶۴	۱/۳۱۱	۱/۰۷۵

مراجع

Almasi, I., Mohammadpour, A., and Mohammadi, M. (2017), Best Linear Unbiased Interpolation of Order Statistics, *Communications in Statistics-Simulation and Computation*, **46**(5), 4161–4171.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (1992), *A First Course in Order Statistics* (Vol. 54), Siam.

Brockwell, P. J., Davis, R. A. (1991). *Time Series: Theory and Methods*, Second ed., Springer Series in Statistics, Springer-Verlag, New York.

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

Omidi, M. and Mohammadzadeh, M. (2016), A New Method to Build Spatio-Temporal Covariance Functions: Analysis of Ozone Data. *Statistical Papers*, **57**(3), 689-703.

Omidi, M. and Mohammadzadeh, M. (2018), Spatial Interpolation Using Copula for non-Gaussian Modeling of Rainfall Data, *JIRSS*, **17**(2), 165-179.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

رگرسیون توزیعی فضایی با دم‌های نیمه‌سنگین

جمیل اونق، حسین باغیشنی[۱]
گروه آمار، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود

چکیده: در این مقاله، به مدل‌بندی داده‌های فضایی مشبکه‌ای معطوف می‌شویم که در بین خود چند داده پرت دارند. برای این کار از رده جدید و بسیار منعطف مدل‌های رگرسیون توزیعی با در نظر گرفتن یک توزیع با دم نیمه‌سنگین برای پاسخ استفاده می‌کنیم. در این رده از مدل‌ها، همه پارامترهای خانواده توزیع پاسخ را می‌توان به متغیرهای تبیینی رگرسیونی و اثرات غیرخطی مثل اثر فضایی مرتبط کرد. برای مدل‌بندی اثر فضایی داده‌ها از یک میدان تصادفی مارکوفی گاوسی (GMRF) استفاده می‌کنیم و در یک مثال واقعی کارایی مدل پیشنهادی را نمایش می‌دهیم.

واژه‌های کلیدی: توزیع با دم نیمه‌سنگین، داده‌های مشبکه‌ای، رگرسیون توزیعی، میدان تصادفی مارکوفی گاوسی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 97K80.

۱ مقدمه

بسیاری از داده‌ها به‌طور ذاتی دارای تغییرپذیری مکانی هستند که به آن‌ها داده‌های فضایی می‌گوییم. امروزه با پیشرفت‌های متعدد ثبت داده‌ها، اطلاعات فضایی داده‌ها به همراه کمیت‌های مختلف مد نظر ثبت می‌شوند که ذخیره‌ای با ارزش برای تحلیل و تصمیم‌سازی‌های کلان کشورها محسوب می‌شوند. مدل‌های رگرسیونی به‌عنوان پرکاربردترین مدل‌های آماری، وظیفه ایجاد ارتباط بین متغیر پاسخ هدف با متغیرهای تبیینی رگرسیونی را برعهده دارند. دو رده پرکاربرد رگرسیونی، مدل‌های خطی تعمیم‌یافته[۱] (GLMs) (نلدر و ودبرن ، ۱۹۷۲) و مدل‌های جمعی تعمیم‌یافته[۲] (GAMs) (هیستی و تبشیرانی ، ۱۹۹۰) هستند. در هر دوی این رده از مدل‌ها، توزیع متغیر پاسخ، Y، عضوی از خانواده توزیع‌های نمایی در نظر گرفته می‌شود، که در آن پارامتر مکان توزیع متغیر پاسخ، μ، به‌عنوان تابعی از متغیرهای تبیینی رگرسیونی مدل‌بندی می‌شود. یعنی در مدل‌های GLM و GAM سایر پارامترهای توزیع، مانند مقیاس و چولگی، به‌طور مستقیم با متغیرهای تبیینی تغییر نمی‌کنند.

[1] Generalized Linear Models
[2] Generalized Additive Models

[۱] نام و ایمیل ارائه دهنده مقاله: حسین باغیشنی، hbaghishani@shahroodut.ac.ir

مدل‌های GAM با در نظر گرفتن اثرات غیرخطی (ناپارامتری) متغیرهای تبیینی به‌صورت جمعی، نسبت به مدل‌های GLM، انعطاف بیشتری دارند، اما به غیر از پارامتر مکان اجازه مدل‌بندی سایر پارامترهای توزیع پاسخ را نمی‌دهند. برای رفع این محدودیت و مدل‌بندی توام پارامترهای مکان، مقیاس و شکل متغیر پاسخ، ریگبی و استاسینوپولوس (۲۰۰۵) رده مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل[۳] (GAMLSS) را معرفی کردند. با توجه به آنکه در این رده از مدل‌ها همه پارامترها را می‌توان به‌طور مستقیم به متغیرهای تبیینی و اثرات غیرخطی (ناپارامتری) مرتبط کرد، از عنوان رگرسیون توزیعی[۴] برای آن‌ها نیز استفاده کرده‌اند. به‌عنوان نمونه می‌توان کلین و نیب (۲۰۱۶) و کلین و همکاران (۲۰۱۵) را نام برد. یکی از قابلیت‌های مدل‌های رگرسیون توزیعی امکان افزودن اثرات تصادفی به هر کدام از پارامترهای توزیع شرطی پاسخ برای پاسخ‌های وابسته است. به‌عنوان مثال، می‌توان با استفاده از نظریه میدان‌های تصادفی مارکوفی[۵] (MRF)، یک اثر تصادفی فضایی را وارد مدل کرد (رو و هلد ، ۲۰۰۵). بی‌سگ و کوپربرگ (۱۹۹۵) مدل اتورگرسیو شرطی ذاتی[۶] (ICAR) نرمال را که زیررده مهمی از MRF هستند معرفی و بررسی کردند. وود (۲۰۰۶) و رو و هلد (۲۰۰۵) مدل ICAR را در قالب رده GAM توسعه دادند. همچنین مدل ICAR توسط دی‌بستیانی و همکاران (۲۰۱۸) برای مدل‌های رگرسیون توزیعی گسترش داده شد.

پیچیدگی‌های زندگی واقعی همیشه اثبات کرده‌اند که همه چیز به دلخواه و مطابق انتظار پیش نخواهند رفت. معمولا در بین داده‌ها یک یا چند داده غیرمعمول نسبت به سایر داده‌ها حضور دارند. اگر چه حضور توزیع‌های مختلف چوله و دم سنگین در رده مدل‌های رگرسیون توزیعی، امکان مدل‌بندی داده‌های پرت را میسر می‌سازند، ولی در موقعیت‌هایی که تعداد آن‌ها نسبت به همه مشاهدات کوچک است، استفاده از توزیع‌های دم‌سنگین می‌تواند پیچیدگی بیش از اندازه وارد مساله کند و بر احتمال‌های مرکزی توزیع به‌طور نامطلوبی تاثیرگذار باشد. این واقعیت، ما را بر آن داشت تا از یک توزیع متقارن با دم نیمه‌سنگین برای مدل‌بندی داده‌های پرت در چارچوب رگرسیون توزیعی استفاده کنیم. برای این منظور، از توزیع هایپربولیک سکانت[۷] (HS) استفاده کردیم (فیشچر ، ۲۰۱۳). بنابراین، هدف اصلی ما در این مقاله، معرفی یک مدل رگرسیون توزیعی (GAMLSS) فضایی جدید برای مدل‌بندی داده‌های فضایی با دم‌های نیمه‌سنگین است.

در بخش دوم توزیع HS و چارچوب مدل‌های رگرسیون توزیعی را تشریح می‌کنیم. در بخش سوم ساخت اثر فضایی ICAR به‌عنوان یکی از اثرات تصادفی ممکن در این مدل‌ها را معرفی می‌کنیم. بخش چهارم شامل تحلیل مجموعه داده سرقت شهر کلمبوس در ایالت اوهایو آمریکا و مقایسه عملکرد سه مدل رقیب است. در پایان نیز نتیجه‌گیری خواهیم کرد.

۲ چارچوب مدل‌بندی رگرسیون توزیعی

اصالت توزیع HS به فیشر (۱۹۲۱) برمی‌گردد. ویژگی‌های این خانواده از توزیع‌ها توسط دود (۱۹۲۵) و پرک (۱۹۳۲) مورد بررسی قرار گرفتند. توزیع HS حول پارامتر مکان، که میانگین توزیع است، متقارن است. وگان (۲۰۰۲) این توزیع را به حالت چوله تعمیم داد و فیشچر و وگان (۲۰۰۲) توزیع هایپربولیک سکانت تعمیم‌یافته چوله را معرفی کردند. فیشچر (۲۰۱۳) در کتاب خود تعمیم‌های مختلف این توزیع را گردآوری کرده است و با برازش این خانواده به داده‌های متنوع، کاربردهای آن را نشان داده است.

[3] Generalized Adittive Models for Location, Scale, and shape
[4] Distributional regression
[5] Markov Random Fields
[6] Intrinsic Conditional Autoregressive
[7] Hyperbolic Secant

تابع چگالی توزیع HS به صورت زیر تعریف می‌شود:

$$f_Y(y;\mu,\sigma)=\frac{۲}{\pi\sigma}\frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{\exp\left(۲\frac{y-\mu}{\sigma}\right)+۱},\quad y\in\mathbb{R}$$

که در آن پارامتر μ میانگین توزیع و σ پارامتر مقیاس هستند. واریانس این توزیع $\frac{\pi^۲}{۴}\sigma^۲$ است و برای نمایش توزیع از نماد $Y\sim HS(\mu,\sigma)$ استفاده می‌کنیم. این توزیع یک توزیع با دم نیمه‌سنگین محسوب می‌شود، به‌طوری که دم‌های آن بین دم‌های توزیع نرمال و توزیع کوشی (به‌عنوان یک توزیع با دم سنگین) است.

رگرسیون توزیعی یک چارچوب مدل‌بندی جامع و انعطاف‌پذیر برای متغیر پاسخ فراهم کرده است. توزیع‌های متقارن، چوله و دم‌سنگین (پیوسته و گسسته) متعددی برای توزیع متغیر پاسخ در بسته‌افزار gamlss در محیط نرم‌افزار R تعبیه شده‌اند. برای مشاهده فهرست این توزیع‌ها می‌توانید به استاسینوپولوس و همکاران (۲۰۱۷) مراجعه کنید. برخی از این توزیع‌ها تا چهار پارامتر را شامل می‌شوند که می‌توان آن‌ها را با μ، σ، ν و τ نشان داد، که به ترتیب بیانگر پارامترهای مکان (مثل میانگین)، مقیاس (مثل انحراف معیار) و شکل (مثل چولگی و کشیدگی) هستند. در چارچوب مدل‌های رگرسیون توزیعی می‌توان همه پارامترهای توزیع (شرطی) پاسخ را به صورت پارامتری یا ترکیبی جمعی از توابع ناپارامتری هموار از متغیرهای تبیینی و اثرات تصادفی پنهان مدل‌بندی کرد. همچنین می‌توان آمیخته‌ای از دو بخش پارامتری و ناپارامتری را به صورت مدلی نیمه‌پارامتری در نظر گرفت.

برای شروع فرض کنید مشاهدات y_i، برای $i=۱,\ldots,n$، به‌طور مستقل (مستقل شرطی) دارای تابع (چگالی) احتمال $f(y_i|\boldsymbol{\theta}^i)$ با بردار پارامترهای $\boldsymbol{\theta}^i=(\theta_{۱i},\theta_{۲i},\theta_{۳i},\theta_{۴i})'=(\mu_i,\sigma_i,\nu_i,\tau_i)'$ باشند. بنا بر ریگبی و استاسینوپولوس (۲۰۰۵) یک مدل رگرسیون توزیعی به صورت زیر تعریف می‌شود:

$$g_k(\theta_k)=\boldsymbol{X}_k\boldsymbol{\beta}_k+\sum_{j=۱}^{J_k}h_{jk}(x_{jk}) \qquad (۱.۲)$$

که در آن $g_k(\cdot)$، برای $k=۱,۲,۳,۴$، یک تابع پیوند معلوم یکنوا است که پارامتر θ_k را به متغیرهای تبیینی مرتبط می‌کند. همچنین، برای هر k و $j=۱,۲,...J_k$، $\boldsymbol{\beta}_k=(\beta_{۱k},\beta_{۲k},\ldots,\beta_{J'_k k})'$ بردار پارامترهای J'_k بعدی، X_k ماتریس طرح معلوم $n\times J'_k$ بعدی، $h_{jk}(\cdot)$ یک تابع ناپارامتری هموار از متغیر تبیینی X_{jk} و x_{jk} ها بردارهایی با طول n از مشاهدات متغیرهای تبیینی با اثرات غیرخطی هستند. مدل (۱.۲) یک مدل نیمه‌پارامتری را در چارچوب مدل‌های GAM برای پارامتر θ_k تشکیل می‌دهد.

در مدل (۱.۲) اثرات هموار متغیرها را می‌توان به صورت یک اثر تصادفی، مشابه $h(\boldsymbol{x})=\boldsymbol{Z}\boldsymbol{\gamma}$، نمایش داد که در آن $\boldsymbol{Z}$ یک ماتریس پایه است که بر اساس مقادیر $\boldsymbol{x}$ ساخته می‌شود و $\boldsymbol{\gamma}$ برداری از ضرایب (اثرات) تصادفی است که فرض می‌شود دارای توزیع نرمال چندمتغیره است. با این نمایش، واضح است که می‌توان اثرات تصادفی را نیز به پیشگوی (۱.۲)، برای هر پارامتر، به صورت جمعی اضافه کرد. بنابراین مدل‌بندی داده‌های وابسته مثل داده‌های طولی، سری زمانی و فضایی در چارچوب مدل‌های رگرسیون توزیعی نیز ممکن می‌شود. با این مقدمه، مدل (۱.۲) را برای هر کدام از پارامترهای مکان، مقیاس و شکل می‌توان به صورت زیر نمایش داد:

$$g_۱(\boldsymbol{\theta}_k)=\boldsymbol{X}_۱\boldsymbol{\beta}_۱+\sum_{j=۱}^{J_k}\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$$

که در آن $\boldsymbol{Z}_{jk}$ ماتریس پایه معلوم $n\times q_{jk}$ بعدی وابسته به متغیرهای تبیینی هستند و $\boldsymbol{\gamma}_{jk}$ بردار تصادفی q_{jk} بعدی با پذیره $\boldsymbol{\gamma}_{jk}\sim N_{q_{jk}}\left(\boldsymbol{۰},\boldsymbol{G}_{jk}^{-۱}\right)$ است که در آن $\boldsymbol{G}_{jk}^{-۱}$ معکوس (تعمیم‌یافته) ماتریس متقارن $\boldsymbol{G}_{jk}$ بعدی $q_{jk}\times q_{jk}$

$\boldsymbol{G}_{jk}(\lambda_{jk})$ است که ممکن است به ابرپارامتر λ_{jk} وابسته باشد. اگر ماتریس $\boldsymbol{G}_{jk}$ ناویژه باشد، بردار ضرایب γ_{jk} دارای یک توزیع ناسره است که چگالی آن متناسب است با $\exp\left(-\frac{۱}{۲}\lambda_{jk}\gamma'_{jk}\boldsymbol{G}_{jk}\gamma_{jk}\right)$. به‌عنوان مثال می‌توان به اثر تصادفی ICAR (رو و هلد ، ۲۰۰۵) برای داده‌های فضایی مشبکه‌ای اشاره کرد که یک توزیع ناسره دارد. تعیین ساختارهای مختلف برای ماتریس‌های $\boldsymbol{Z}$ و $\boldsymbol{G}$ انواع مختلفی از اثرات جمعی را نتیجه می‌دهد. این اثرات شامل مولفه‌های تصادفی، مولفه‌های هموار متغیرها، مولفه‌های سری زمانی و اثرات فضایی می‌شوند. برای مولفه‌های هموار انتخاب‌های مختلفی مانند اسپلاین‌های جریمه‌ای، اسپلاین‌های درجه سوم، و چندجمله‌ای‌های موضعی وجود دارند.

چارچوب مدل‌بندی رگرسیون توزیعی این امکان را فراهم می‌کند که توزیع شرطی متغیر پاسخ (به شرط اثرات تصادفی) هر توزیع دلخواهی چه از خانواده توزیع‌های نمایی چه خارج از آن باشد. این یک ویژگی خیلی منعطف و خوب محسوب می‌شود. البته به‌عنوان یک محدودیت توزیع اثرات تصادفی باید گاوسی باشد. در این مقاله، ما با تعبیه کردن توزیع HS به چارچوب مدل‌های رگرسیون توزیعی امکان مدل‌بندی منعطف داده‌های پیوسته‌ای را فراهم کرده‌ایم که یک یا چند داده پرت را شامل می‌شوند. مدل مورد نظر ما یک مدل نیمه‌پارامتری مکان‌ـ‌مقیاس است که در ساختار مدل (۱.۲) با دو پارامتر $\theta_۱ = \mu$ و $\theta_۲ = \sigma$ صدق می‌کند. همچنین برای در نظر گرفتن وابستگی فضایی داده‌های مشبکه‌ای از یک اثر تصادفی ICAR استفاده می‌کنیم که در زیربخش بعدی آن را معرفی خواهیم کرد.

رهیافت استنباطی مد نظر برای مدل پیشنهادی، مبتنی بر درستنمایی است. با توجه به این‌که در این مدل از مولفه‌های ناپارامتری جمعی مختلف می‌توان استفاده کرد، برای جلوگیری از پیچیدگی بیش از حد، از یک چارچوب درستنمایی جریمه‌ای استفاده می‌شود. برای سادگی تشریح روش، فرض می‌کنیم داده‌ها مستقل باشند. بنابراین تابع لگاریتم درستنمایی مدل پیشنهادی به صورت

$$\ell(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=۱}^{n} \log\left(f(y_i; \mu_i, \sigma_i)\right)$$

است. تابع درستنمایی جریمه‌ای برای این مدل به صورت زیر است:

$$\ell_p = \ell(\boldsymbol{\mu}, \boldsymbol{\sigma}) - \sum_{k=۱}^{۲} \sum_{j=۱}^{J_k} \gamma'_{kj} \boldsymbol{G}_{kj}(\lambda_{kj}) \gamma_{kj}. \qquad (۲.۲)$$

مجموعه پارامترهایی که باید برآورد شوند، عبارتند از پارامترهای بخش خطی مدل، $\boldsymbol{\beta} = (\boldsymbol{\beta}_۱, \boldsymbol{\beta}_۲)'$، پارامترهای اثرات تصادفی، $\boldsymbol{\gamma} = (\gamma_{۱۱}, ..., \gamma_{J_۱۱}, \gamma_{۱۲}, ..., \gamma_{J_۲۲})'$، و ابرپارامترهای مدل، $\boldsymbol{\lambda} = (\lambda_{۱۱}, ..., \lambda_{J_۱۱}, \lambda_{۱۲}, ..., \lambda_{J_۲۲})'$. در این جا، پارامترهای خطی $\boldsymbol{\beta}$ و پارامترهای اثرات تصادفی $\boldsymbol{\gamma}$، به ازای مقادیر ثابتی برای ابرپارامترهای هموارساز $\boldsymbol{\lambda}$، با ماکسیمم کردن تابع درستنمایی جریمه‌ای ℓ_p در (۲.۲) برآورد می‌شوند. برای این کار دو روش پایه‌ای با نام‌های الگوریتم‌های RS و CG (ریگبی و استاسینوپولوس ، ۲۰۰۵) وجود دارند. هر دوی این روش‌ها از یک الگوریتم کمترین توان‌های دوم (جریمه‌ای) موزون تکراری برای محاسبه برآوردهای ماکسیمم درستنمایی پارامترهای مذکور، با مفروض بودن ابرپارامترهای هموارساز، استفاده می‌کنند. روش RS تعمیم الگوریتم به کار گرفته‌شده توسط ریگبی و استاسینوپلوس (۱۹۹۶) است. روش CG نیز تعمیم الگوریتمی است که توسط کول و گرین (۱۹۹۲) معرفی شد. در الگوریتم CG مشتقات مرتبه اول تابع درستنمایی و مقادیر امید ریاضی مشتقات مرتبه دوم تابع درستنمایی نسبت به پارامترهای توزیع یعنی μ و σ مورد نیاز هستند. البته برای مشتقات مرتبه دوم، مشابه روش بهینه‌سازی BFGS (بروید ، ۱۹۷۶)، می‌توان از تقریب‌های عددی نیز

استفاده کرد. مشتقات اول توزیع HS نسبت به μ و σ عبارتند از

$$\frac{\partial \ell(\mu,\sigma)}{\partial \mu} = -\frac{۱}{\sigma} + \frac{۲}{\sigma}\frac{\exp\left(۲\frac{y-\mu}{\sigma}\right)}{\exp\left(۲\frac{y-\mu}{\sigma}\right)+۱}$$

$$\frac{\partial \ell(\mu,\sigma)}{\partial \sigma} = -\frac{۱}{\sigma} - \frac{y-\mu}{\sigma^۲} + ۲\frac{y-\mu}{\sigma^۲}\frac{\exp\left(۲\frac{y-\mu}{\sigma}\right)}{\exp\left(۲\frac{y-\mu}{\sigma}\right)+۱}.$$

برای مشتقات دوم نیز از تقریب‌های عددی استفاده کرده‌ایم. برای برآورد بردار ابرپارامترهای $\boldsymbol{\lambda}$ دو روش موضعی[8] و فراموضعی[9] (ریگبی و استاسینوپلوس ، ۲۰۱۳) پیشنهاد شده‌اند. روش موضعی سریعتر و اجرای آن ساده‌تر است. در واقع در روش موضعی برآورد ابرپارامترها در هر مرحله از الگوریتم‌های RS یا CG انجام می‌شود، در حالی‌که در روش فراموضعی، از معیارهایی مثل اعتبارسنجی متقابل تعمیم‌یافته[10] (GCV) (وود ، ۲۰۰۶)، معیار اطلاع آکاییک (AIC) (آکاییک ، ۱۹۷۴)، معیار اطلاع بیزی شوارتز (BIC) (شوارتز ، ۱۹۷۸)، و مقدار ماکسیمم درستنمایی برای برآورد ابرپارامترها استفاده می‌شود.

برای مقایسه و انتخاب مدل برتر در بین مدل‌های نامزد می‌توان از معیارهای AIC و BIC استفاده کرد. بررسی نیکویی برازش مدل‌ها نیز توسط تحلیل باقی‌مانده‌ها انجام می‌شود.

۳ اثر تصادفی GMRF

در تحلیل داده‌های ناحیه‌ای (مشبکه‌ای) می‌توان انتظار داشت که نواحی همسایه، در ناحیه تحت مطالعه، شباهت بیشتری نسبت به نواحی غیرهمسایه داشته باشند. در این حالت، معمولا، ویژگی مارکوفی برای اثر تصادفی فضایی داده‌ها پذیرفته می‌شود؛ به این معنی که احتمال شرطی رخداد پیشامدی به شرط رخداد آن در نواحی همسایه از رخداد آن پیشامد در نواحی غیرهمسایه مستقل (شرطی) است. این یک ویژگی مارکوفی گرافی بدون جهت است که در آن هر رأس نشان‌دهنده ناحیه و وجود یال بین دو رأس نشان‌دهنده داشتن مرز مشترک (همسایگی) بین دو ناحیه است (رو و هلد ، ۲۰۰۵). برای توضیح بیشتر فرض کنید $\eth = (\nu, \epsilon)$ یک گراف بدون جهت باشد که شامل رأس‌های $\nu = (۱, ۲, ..., q)$ و مجموعه یال‌های ϵ است که به صورت زوج‌های (m,t) هستند، به‌طوری که $m,t \in \nu$. با توجه به رو و هلد (۲۰۰۵) بردار تصادفی $\boldsymbol{\gamma} = (\gamma_۱, ..., \gamma_q)^T$ به شرط گراف $\eth$ یک GMRF با بردار میانگین $\boldsymbol{\mu}$ و ماتریس دقت (معکوس ماتریس کوواریانس) متقارن $\lambda \boldsymbol{G}$ است، اگر و فقط اگر تابع چگالی آن به صورت

$$\pi(\boldsymbol{\gamma}) = \exp\left[-\frac{۱}{۲}\lambda(\boldsymbol{\gamma}-\boldsymbol{\mu})'\boldsymbol{G}(\boldsymbol{\gamma}-\boldsymbol{\mu})\right] \qquad (۱.۳)$$

باشد، که در آن $G_{mt} \neq ۰$ اگر و فقط اگر، برای هر $m \neq t$، $(m,t) \in \epsilon$. در این جا، G_{mt} مولفه سطر mام و ستون tام ماتریس $\boldsymbol{G}$ است. مدل CAR که اولین بار توسط بی‌سگ (۱۹۷۴) معرفی شد و توسط بی‌سگ و کوپربرگ (۱۹۹۵)

[8] Locally
[9] Globally
[10] Generalized Cross Validation

توسعه داده شد، حالت خاصی از یک GMRF به شکل (۱.۳) است. مدل CAR را به صورت زیر نیز می‌توان تعریف کرد:

$$\gamma_i|\boldsymbol{\gamma}_{-i} \sim N\left(\sum_j \alpha_{ij}\gamma_j, k_i\right)$$

که در آن $\boldsymbol{\gamma}_{-i} = (\gamma_1, \gamma_1, ..., \gamma_{i-1}, \gamma_{i+1}, ..., \gamma_q)$، $k_i = \frac{1}{\lambda G_{ii}}$ و $i = 1, 2, ..., q$

$$\alpha_{ij} = \begin{cases} -\dfrac{G_{ij}}{G_{ii}} & i \neq j \\ 0 & i = j \end{cases}$$

با استفاده از لم بروک (۱۹۶۴)، می‌توان نشان داد که به‌ازای $\boldsymbol{\mu} = 0$ برای تمام i و j ها، رابطه $\alpha_{ij}k_j = \alpha_{ji}k_i$ برقرار است که تضمین‌کننده متقارن بودن ماتریس $\boldsymbol{G}$ است (بی‌سگ و کوپربرگ ، ۱۹۹۵).

زیررده‌ای از GMRF نیز وجود دارد که در آن ماتریس $\boldsymbol{G}$ ناویژه است و به مدل ICAR معروف است. در واقع، این مدل یک نسخه حدی از مدل CAR است. مدل ICAR در مدل‌های خطی تعمیم‌یافته فضایی با پاسخ‌های مشبکه‌ای، برای تبیین اثرات تصادفی که دارای ساختار فضایی هستند به‌کار برده می‌شوند (بنرجی و همکاران ، ۲۰۱۴). دی‌بستیانی و همکاران (۲۰۱۸) در رده مدل‌های رگرسیون توزیعی، ماتریس $\boldsymbol{G}$ را به شکل زیر گسترش دادند. فرض کنید $\boldsymbol{W}$ یک ماتریس همسایگی متقارن باشد که در آن مولفه‌های روی قطر اصلی صفر هستند و سایر مولفه‌های (i, j) در صورتی که دو ناحیه i و j، با شرط $i \neq j$، مرز مشترک داشته باشند برابر ۱ و در غیر این‌صورت برابر صفر هستند. با این تعریف، ماتریس دقت $\boldsymbol{G}$ از رابطه $\boldsymbol{G} = \boldsymbol{D}_w - \boldsymbol{W}$ ساخته می‌شود که در آن $\boldsymbol{D}_w$ ماتریس قطری است که اعضای قطر اصلی آن مجموع سطرهای ماتریس $\boldsymbol{W}$ است. با افزودن یک اثر ICAR به پیشگوی خطی مدل (۱.۲)، می‌توان وابستگی فضایی داده‌ها را وارد مدل پیشنهادی کرد. برازش مدل نیز با استفاده از تابع درستنمایی جریمه‌ای (۲.۲) به سادگی قابل انجام است که در آن جریمه بر حسب $\boldsymbol{G}$ ساخته‌شده توسط مدل ICAR تعریف می‌شود.

۴ داده‌های سرقت شهر کلمبوس

برای نمایش کاربرد مدل پیشنهادی ما، در این بخش از مجموعه داده حجم سرقت منزل در شهر کلمبوس ایالت اوهایوی آمریکا استفاده کرده‌ایم. متغیر پاسخ حجم سرقت منزل و اتومبیل (به درصد در ۱۰۰۰ خانوار)، با نام crime، است. با توجه به متغیرهای تبیینی موجود در این مجموعه از داده‌ها، دو متغیر درآمد خانوار، با نام income، و قیمت منزل، با نام home.value، برای برازش مدل‌های رگرسیونی به‌کار گرفته شدند. با توجه به این‌که شهر کلمبوس دارای ۴۹ ناحیه تقسیم‌بندی‌شده است، ماتریس همسایگی مدل ICAR یک ماتریس متقارن با بعد ۴۹ است.

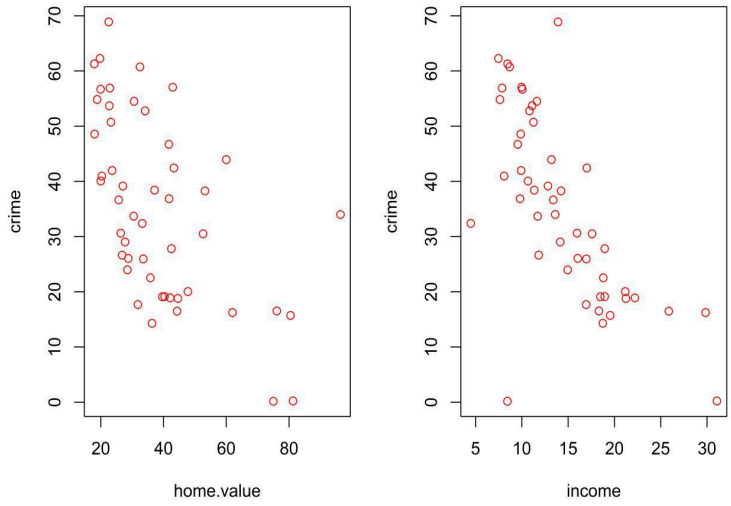
شکل ۱ نمودارهای پراکنش کناری متغیر پاسخ در مقابل متغیرهای تبیینی را نشان می‌دهد. همان‌طور که در دو نمودار پراکنش مشهود است، یک یا دو نقطه پرت حضور دارند که باید در مدل‌بندی لحاظ شوند. همچنین متغیر پاسخ به‌طور متوسط با افزایش سطوح متغیرهای تبیینی ارزش منزل مسکونی و درآمد روندی کاهشی دارد. در مقابل، با افزایش سطح درآمد تغییرات پاسخ نیز افزایش می‌یابد که می‌توان به مساله ناهم‌واریانس تا حدودی پی برد. با توجه به این تحلیل اکتشافی، از یک مدل مکان‌ـ‌مقیاس به صورت

$$\begin{cases} \mu = \beta_0 + \beta_1 income + h_1(home.value) + h_{spat}(district) \\ \sigma = \gamma_0 + \gamma_1 income \end{cases} \quad (۱.۴)$$

جدول ۱: نتایج برازش مدل‌ها برای مجموعه داده سرقت

مدل	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\nu}$	AIC	BIC
نرمال	۶۱/۱۶۹	−۱/۳۸۷	۳/۹۷۷	−۰/۱۲۹	–	۳۶۱/۶۳۹	۳۷۴/۸۸۱
HS	۵۴/۶۳۰	−۱/۳۳۳	۳/۷۲۳	−۰/۱۴۳	–	۳۵۷/۶۸۹	۳۷۰/۹۳۴
tـاستیودنت	۶۰/۲۷۹	−۱/۳۳۲	۳/۹۵۰	−۰/۱۴۵	۱/۵۰۵	۳۶۰/۷۱۱	۳۷۵/۸۴۷

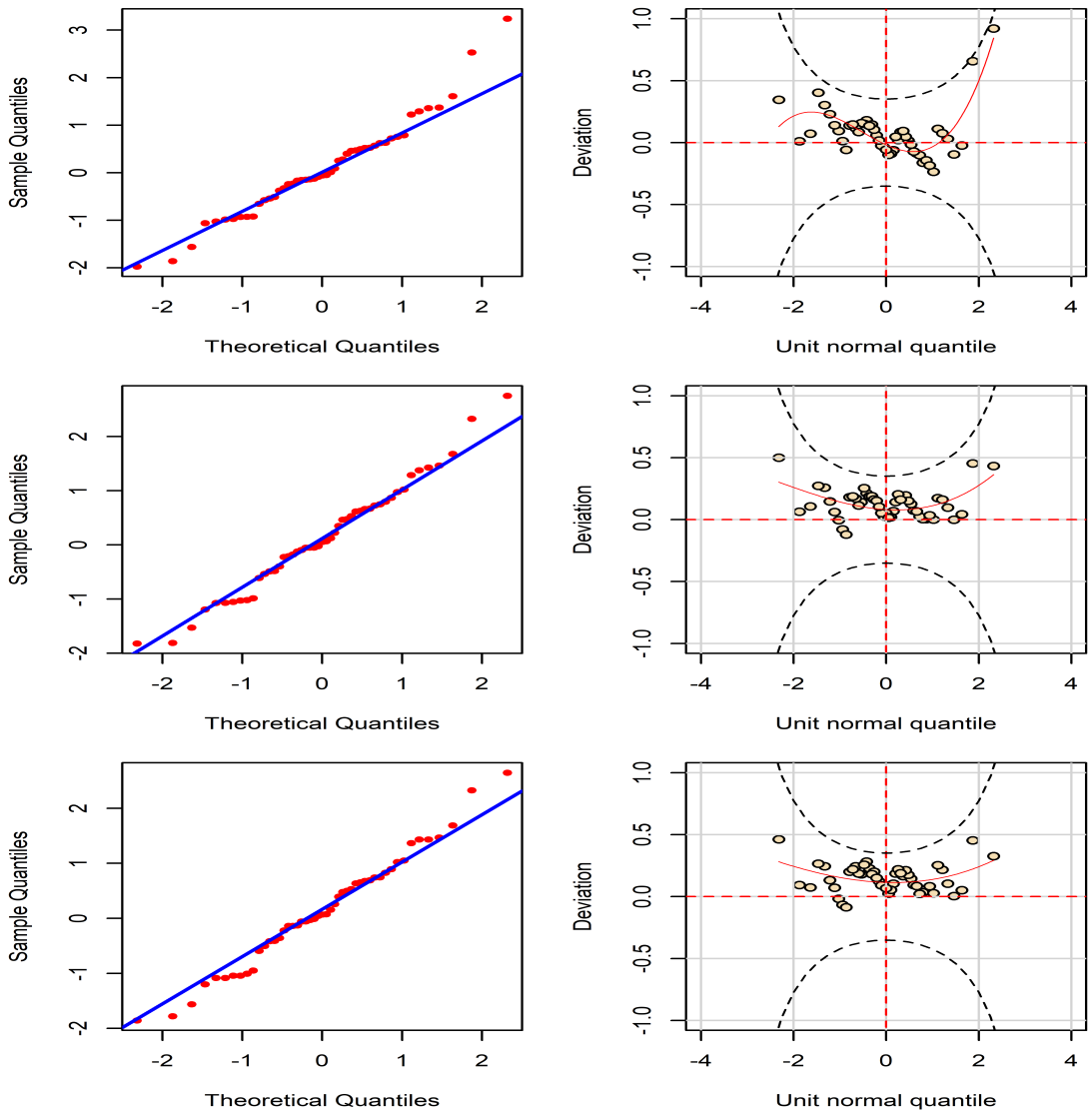
برای برازش داده‌ها استفاده کردیم. برای مقایسه عملکرد مدل HS، دو مدل رقیب مکان‌ـمقیاس نرمال (به‌عنوان یک توزیع با دم سبک) و مدل مکان‌ـمقیاس tـاستیودنت (به‌عنوان یک توزیع دم سنگین) را نیز به داده‌ها برازش دادیم. نتایج برازش هر سه مدل در جدول ۱ گزارش شده‌اند. برآوردهای ضرایب رگرسیونی در هر دو پارامتر میانگین و مقیاس برای هر سه مدل تقریبا شبیه هم هستند. البته شباهت دو مدل HS و tـاستیودنت بیشتر است. بر اساس هر دو معیار انتخاب مدل AIC و BIC، مدل برتر، مدل پیشنهادی ما است که مقادیر کوچکتری را نسبت به رقبای خود داراست. برای ارزیابی نیکویی برازش مدل‌ها، از نمودارهای چندک‌ـچندک و کرمی[11] باقی‌مانده‌های روندزدایی‌شده استفاده کردیم. برای سه مدل برازش‌شده این نمودارها در شکل ۲ نمایش داده شده‌اند. با توجه به نمودارهای مدل نرمال می‌توان نقصان برازش آن را نتیجه گرفت. در مقابل برای هر دو مدل HS و tـاستیودنت نیکویی برازش مورد تایید است.



شکل ۱: نمودارهای پراکنش کناری پاسخ و متغیرهای تبیینی برای داده‌های کلمبوس

با توجه به برتری مدل پیشنهادی ما، برآورد همه اثرات برای هر دو پارامتر میانگین و مقیاس را تنها برای مدل HS در شکل ۳ نشان داده‌ایم. جالب است که حجم سرقت در خانه‌های با ارزش کمتر یا با ارزش متوسط افزایش داشته است و برای خانه‌های گران‌قیمت به‌طور متوسط درصد سرقت کاهشی است. همچنین با افزایش درآمد خانوار، درصد سرقت کاهش می‌یابد. شکل ۴ نیز اثر فضایی برآوردشده را برای هر سه مدل نشان می‌دهد. اثر برآوردشده برای دو مدل HS و tـاستیودنت شباهت زیادی به هم دارند و تفاوت‌هایی با مدل نرمال دارا می‌باشند. بنابراین با توجه به نقصان مدل نرمال، می‌توان از نتایج حاصل از آن صرفنظر کرد. با توجه به نتایج حاصل از مدل HS می‌توان نتیجه گرفت، درصد سرقت در نواحی مرکزی و شرقی، به‌ویژه جنوب شرقی شهر کلمبوس بیشتر از سایر نواحی است. همچنین کمترین درصد سرقت مربوط

[11] Worm plot

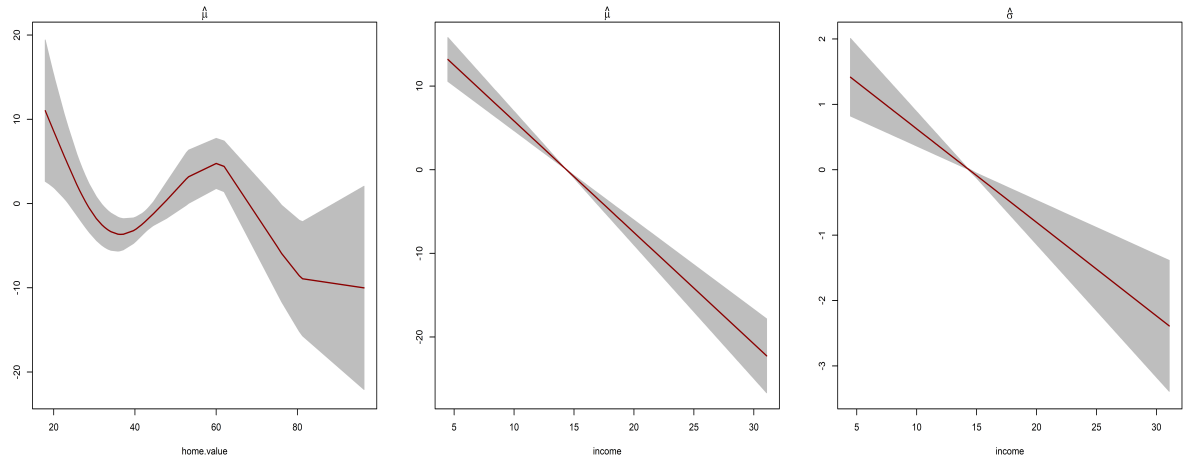


شکل ۲: نمودارهای چندک‌ـچندک و کرمی مدل‌های نرمال (بالا)، HS (وسط) و tـاستودنت (پایین)

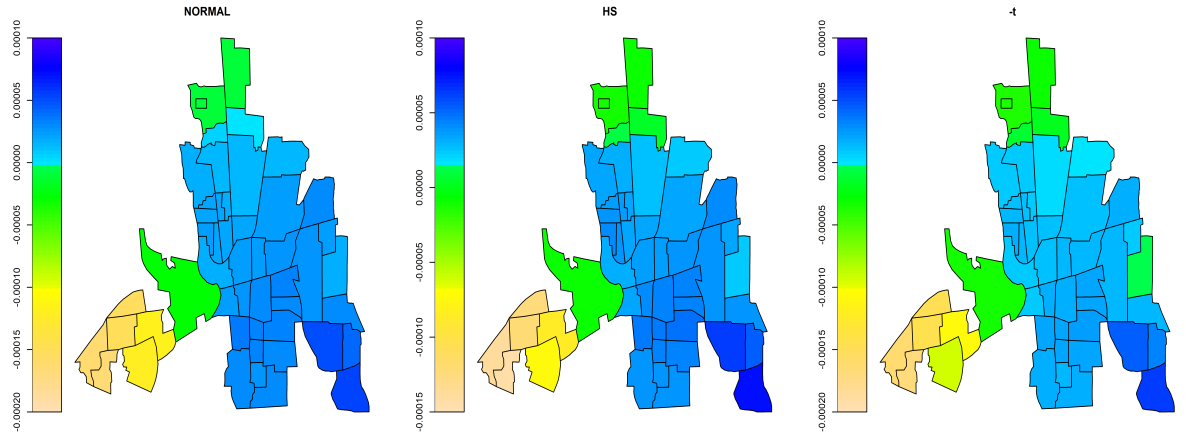
به نواحی جنوب غربی این شهر است.

بحث و نتیجه‌گیری

در این مقاله، یک مدل رگرسیون توزیعی فضایی جدید بر پایه توزیع با دم نیمه‌سنگین HS ارایه شد که می‌تواند جانشین مناسبی برای مدل نرمال در مواقعی باشد که بخش کوچکی از داده‌ها پرت محسوب می‌شوند. این مدل حتی در موقعیت‌هایی می‌تواند نسبت به مدل‌های تنومندی مثل مدل رگرسیون توزیعی tـاستیودنت برتر باشد. با این کار، شاید حضور توزیع‌های با دم نیمه‌سنگین به‌عنوان جانشین‌هایی برای مدل‌های دم‌سنگین توجیه‌پذیر باشد. به‌ویژه مدل‌بندی داده‌های فضایی که به دلیل وابستگی فضایی داده‌ها تعداد داده‌های پرت چندان زیاد نیستند، می‌تواند نشانی از اهمیت ورود این دسته از مدل‌ها باشد.



شکل ۳: منحنی‌های اثرات برآوردشده پارامترهای خطی و غیرخطی مکان (وسط و سمت چپ) و مقیاس (سمت راست) برای مدل HS



شکل ۴: اثر فضایی برآوردشده پارامتر مکان در مدل‌های نرمال (سمت چپ)، HS (وسط) و tـاستیودنت (سمت راست)

مراجع

Akaike, H. (1974), A New Look at the Statistical Model Identification, Selected Papers of Hirotugu Akaike, *Springer, New York*, 215-222.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed., Boca Raton: Chapman & Hall/CRC.

Besag, J., and Kooperberg. C. (1995), On Conditional and Intrinsic Autoregressions, *Biometrika*, **82**, 733–746.

Besag, T. (1974), Spatial Interaction and the Statistical Analysis of Lattice Systems (With Discussion), *Journal of the Royal Statistical Society, Series B*, **36**, 192–236.

Brook, D. (1964), On the Distinction Between the Conditional Probability and the Joint Probability Approaches in the Specification of Nearest-Neighbour Systems, *Biometrika*, **51**, 481–483.

Broyden, C. G. (1976), Quasi-Newton Methods and Their Application to Function Minimization, *Mathematics of Computation*, **21**, 368-381.

Cole, T. J., and Green, P. J. (1992), Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood, *Statistics in Medicine*, **11**, 1305-1319.

De Bastiani, F., Rigby, R. A., Stasinopoulous, D. M., Cysneiros, A. H., and Uribe-Opazo, M. A. (2018), Gaussian Markov Random Field Spatial Models in GAMLSS, *Journal of Applied Statistics*, **45**(1), 168-186.

Dodd, E. L. (1925), The Frequency Law of a Function of Variables with Given Frequency Laws, *Annals of Mathematics*, **27**, 12-20.

Fischer, M. J. (2013), *Generalized Hyperbolic Secant Distributions: With Applications to Finance*, Springer Science & Business Media.

Fischer, M., and Vaughan, D. C. (2002), Classes of Skewed Generalized Hyperbolic Secant Distributions, Working Paper No. 45 (Unpublished), Department of Statistics and Econometrics, FAU ErlangenNuremberg, Nuremberg.

Fisher, R. A. (1921), 014: On the ”Probable Error” of a Coefficient of Correlation Deduced From a Small Sample, *Metron*, **1**, 3-32.

Hastie, T., and Tibshirani, R. (1990), Generalized Additive Models, *Statistical Science*, **1**, 297-310.

Klein, N., and Kneib, T. (2016), Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression, *Bayesian Analysis*, **11**, 1071-1106.

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015), Bayesian Structured Additive Distributional Regression for Multivariate Responses, *Journal of the Royal Statistical Society, Series C*, **64**, 569-591.

Nelder, J. and Wedderburn, R. W. M. (1972), Generalized Linear Models, *Journal of the Royal Statistical Society: Series A (General)*, **135**, 370-384.

Perks, W. (1932), On Some Experiments in the Graduation of Mortality Statistics, *Journal of the Institute of Actuaries*, **63**, 12-57.

Rigby, R. A., and Stasinopoulos, D. M. (1996), A Semi-Parametric Additive Model for Variance Heterogeneity, *Statistics and Computing*, **6**, 57-65.

Rigby, R. A., and Stasinopoulos, D. M. (2005), Generalized Additive Models for Location, Scale and Shape, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507-554.

Rigby, R. A., Stasinopoulos, M. D., and Voudouris, V. (2013), Discussion: A Comparison of GAMLSS with Quantile Regression, *Statistical Modelling*, **13**, 335-348.

Rue, H., and Held. L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, London: Chapman and Hall/CRC.

Schwarz, G. (1978), Estimating the Dimension of a Model, *The Annals of Statistics*, **6**, 461-464.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017), *Flexible Regression and Smoothing Using GAMLSS in R*, London: Chapman and Hall/CRC.

Vaughan, D. C. (2002), The Generalized Secant Hyperbolic Distribution and Its Properties, *Communications in Statistics-Theory and Methods*, **31**, 219-238.

Wood, S. N. (2006), *Generalized Additive Models: An Introduction with R*, New York: Chapman & Hall.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

تحلیل مدل‌های سری‌زمانی شمارشی فضایی

سپیده تشرفی[1]، امید کریمی، فاطمه حسینی
گروه آمار، دانشگاه سمنان

چکیده: داده‌های شمارشی چند متغیره برحسب زمان و مکان در خیلی از زمینه‌های کاربردی مشاهده می‌شوند. در این مقاله مدل‌های خطی سری‌زمانی شمارشی پواسنی روی این نوع از داده‌ها مورد مطالعه قرار می‌گیرند و همچنین همبستگی مکانی داده‌ها به‌وسیله یک مدل سری‌زمانی شمارشی فضایی با رهیافت بیزی را مورد تحلیل قرار می‌دهیم. در انتها یک مطالعه شبیه‌سازی برای بررسی مدل‌های معرفی شده، ارائه می‌شود.

واژه‌های کلیدی: سری‌های زمانی شمارشی، فرایند پواسن فضایی، داده‌های فضایی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 62M10، 62H11.

۱ مقدمه

سری‌های زمانی مجموعه‌ای از مشاهدات هستند که بر حسب زمان مرتب شده است. ایجاد و به کارگیری مدل‌های آماری و تصادفی در قالب تحلیل سری زمانی، امروزه به کمک رایانه‌های پرسرعت، بسیار فراگیر شده و حاصل آن مدل‌هایی است که با داشتن پارامترهای بسیار انعطاف‌پذیر، می‌توانند آینده را برای هر پدیده‌ای پیش‌بینی کنند. کاربردهای تحلیل سری‌های زمانی، در زمینه‌های مختلف نظیر، کسب و کار، امور مالی، بورس، مهندسی و غیره دیده می‌شود. به بیان دیگر تحلیل سری‌زمانی، ایجاد مدلی گذشته‌نگر است تا امکان تصمیمات آینده‌نگر را فراهم سازد. معمولا در سری زمان‌ـ گسسته، داده‌ها در مقاطع مشخصی از زمان مثل ساعت، روز یا هفته و حتی سال جمع‌آوری می‌شوند. در حالتی که داده‌ها از یک فرایند شمارشی به دست آمده باشد؛ سری زماني حاصل را سری زمانی شمارشی می‌گویند (زهو و جو ، ۲۰۰۶؛ بروکول و داویس ، ۱۹۹۱). مدل‌های گوناگونی برای سری‌زمانی شمارشی وجود دارد. موردی که در مدل‌های سری زمانی شمارشی باید در نظر گرفته شود این است که وابستگی میان مشاهدات به‌طور مناسب مدل شود. یکی از این شیوه‌ها به کارگیری مدل‌های خطی تعمیم یافته است (مک‌کلاخ و نلدر ، ۱۹۸۱).

[1] نام و ایمیل ارائه دهنده مقاله: سپیده تشرفی، s.tasharofy@gmail.com

در این مقاله ابتدا مدل‌های سری زمانی برای تحلیل داده‌های شمارشی را مورد مطالعه قرار می‌دهیم، سپس همبستگی مکانی به‌وسیله تلفیق یک مدل پواسن فضایی با مدل‌های سری زمانی شمارشی مورد تحلیل قرار می‌دهیم و برخی استنباط‌های آماری مرتبط را بررسی می‌کنیم. تعدادی از این مدل‌ها را می‌توان در زمینه‌های کاربردی اقتصاد سنجی، علوم اجتماعی و فیزیک استفاده کرد. برای تحلیل رگرسیونی داده‌های شمارشی، مدل خطی معمولی مناسب نیست. زیرا که متغیر پاسخ مقادیر گسسته می‌گیرد. یک مدل مناسب، مدل رگرسیون پواسون است و یک نقطه شروع طبیعی برای گسترش آن به داده‌های شمارشی وابسته است. مدل رگرسیون پواسون در بسیاری از زمینه‌های کاربردی مورد استفاده قرار گرفته است. در واقع مدل پواسون ابزار اصلی برای مدل‌سازی داده‌های سری‌زمانی شمارشی است. با این حال، دیگر توزیع‌ها ممکن است به‌جای آن استفاده شود. بهترین آن در میان توزیع‌ها، توزیع دوجمله‌ای منفی است. صرف‌نظر از توزیع انتخاب شده عمدتا مدل‌هایی را مورد بررسی قرار خواهیم داد که در چارچوب مدل‌های خطی تعمیم یافته سری زمانی است. این کلاس مدل‌ها و قضیه‌ی ماکسیمم درستنمایی یک چارچوب سیستماتیک برای تجزیه و تحلیل داده‌های سری‌زمانی به‌صورت کمی و کیفی ارائه می‌دهند. در واقع برآورد، تشخیص، ارزیابی مدل و پیش‌بینی به روش ساده انجام می‌شود. کلاس‌های جایگزین دیگری از مدل‌های رگرسیون وجود دارد. یکی از برجسته‌ترین آن‌ها مدل اتورگرسیو صحیح مقدار است. این مدل‌ها مشابه ساختار فرآیند اتورگرسیو معمول است.

ساختار مقاله به این صورت است که در بخش ۲ مدل‌های سری‌زمانی شمارشی معرفی و مورد مطالعه قرار می‌گیرند. همچنین جزئیات مدل‌های خطی و لگاریتم خطی سری‌زمانی شمارشی بیان می‌شود. در بخش ۳ یک مدل اتورگرسیو همبسته فضایی برای داده‌های شمارشی بیان و در بخش ۴ یک مطالعه شبیه‌سازی با مدل‌های معرفی شده، مورد مقایسه قرار می‌گیرد. در انتها بحث و نتیجه‌گیری ارائه می‌شود.

۲ مدل‌های سری زمانی شمارشی

۱.۲ مدل‌سازی رگرسیون پواسون

توزیع پواسون معمولا برای مدل نرخ رویدادهای تصادفی استفاده می‌شود که در برخی از فاصله‌های زمانی ثابت رخ می‌دهد. اگر فرض کنیم λ نشان دهنده نرخ ورودی است؛ توزیع متغیر تصادفی Y نشان دهنده نرخ ورودی‌ها در یک فاصله زمانی ثابت است و دارای توزیع پواسون با تابع چگالی احتمال زیر است:

$$P[Y=y]=\frac{\exp(-\lambda)\lambda^{y}}{y!}, \qquad y=۰,۱,۲,\dots \tag{۱.۲}$$

میانگین و واریانس Y هر دو برابر λ است. در اکثر مسائل کاربردی، داده‌های شمارشی معمولا با برخی اطلاعات متغیر کمکی مشاهده می‌شود (مک‌کلاخ و نلدر ، ۱۹۸۱). به‌طور کلی فرض کنید که $X_۱,\cdots,X_p$ ، p متغیر رگرسیونی مشاهده شده همراه با متغیر پاسخ شمارشی Y با توزیع پواسون است. در آن صورت مدل رگرسیونی به صورت $\lambda=\beta_۰+\sum_{i=۱}^{p}\beta_iX_i$ است. این مدل چند مشکل برای برآورد دارد، زیرا که پارامتر λ باید مثبت باشد. یک انتخاب ساده برای مدل‌سازی رگرسیون داده‌های شمارشی، مدل لگاریتم خطی به‌صورت $log\lambda=\beta_۰+\sum_{i=۱}^{p}\beta_iX_i$ است. هر دو مدل متعلق به کلاس مدل‌های خطی تعمیم یافته‌اند که توسط نلدر و ویدربرن (۱۹۷۲) معرفی شده است. در تعریف مدل‌های سری‌زمانی شمارشی از مدل اتورگرسیو مرتبه یک ($AR(۱)$) به‌صورت

$$Y_t=b_۱Y_{t-۱}+\epsilon_t, \tag{۲.۲}$$

استفاده می شود. که در آن $|b_1| < 1$ و $\{\epsilon_t\}$ دنباله متغیرهای تصادفی نرمال با میانگین صفر و واریانس σ^2 است. این یک مدل استاندارد برای تجزیه و تحلیل سری‌های زمانی مقادیر واقعی است که مقدار فرآیند در زمان t بسته به مقدار فرایند در زمان $t-1$ به همراه یک خطای تصادفی است. برای مثال پریستی (۱۹۸۱) را ببینید.

۱.۱.۲ مدل های خطی برای سری‌های زمانی شمارشی

فرض کنید که $\{Y_t\}$ سری‌های زمانی شمارشی را نشان می‌دهد. مدل اتورگرسیو پواسون را برای فرایند پاسخ $\{Y_t\}$ به صورت

$$Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \qquad \lambda_t = d + b_t Y_{t-1}, \qquad t \geq 1, \tag{۳.۲}$$

تعریف می‌کنیم که در آن b_1 ، d و $\mathcal{F}_t = \sigma(Y_s, s \geq t)$ پارامترهای نامنفی و $\{\lambda_t\}$ به‌عنوان میانگین فرآیند Y_t با توجه به گذشته آن است. مقادیر مثبت d و b_1 ما را مطمئن می‌سازد که $\lambda_t > 0$ است زیرا که Y_t یک عدد صحیح نامنفی است. با توجه به این نکات، واضح است که مدل (۳.۲) با مولفه تصادفی توزیع پواسون و مؤلفه سیستماتیک $\eta_t = d + bY_{t-1}$ و تابع ربط همانی در چارچوب مدل های خطی تعمیم یافته قرار می‌گیرد. مدل (۳.۲) همان پویایی مدل (۲.۲) را نشان می‌دهد، بنابراین داریم:

$$Y_t = \lambda_t + (Y_t - \lambda_t) = d + b_1 Y_{t-1} + \epsilon_t \qquad t \geq 1 \tag{۴.۲}$$

که در آن $\{\epsilon_t\}$ دنباله نوفه سفید است. این یک دنباله‌ای از متغیرهای تصادفی ناهمبسته با میانگین صفر و واریانس ثابت است. حالت کلی مدل (۳.۲) با مرتبه‌ی (p, q) به‌صورت

$$Y_t|\mathcal{F}_{t-1}^{Y,\lambda} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + \sum_{i=1}^{p} a_i \lambda t - i + \sum_{j=1}^{q} b_j Y_{t-j}, t \geq \max(p, q) \tag{۵.۲}$$

تعریف می‌شود که در آن $\mathcal{F}_t^{Y,\lambda}$ یک σ میدان تشکیل شده از $\{Y_0, \ldots, Y_t, \lambda_0\}$ است. یعنی $\mathcal{F}_t^{Y,\lambda} = \sigma(Y_s, \lambda_0, s \leq t)$ و $\{\lambda_t\}$ همانند قبل یک فرآیند شدت پواسون است. مانایی مرتبه دوم با رابطه $0 < \sum_{j=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$ فراهم می‌شود. برای اطلاعات بیشتر به منابع فرلند و همکاران (۲۰۰۶) ریدبرگ و شفرد (۲۰۰۰) استریت (۲۰۰۰) هنین (۲۰۰۳) مراجعه کنید. برای توزیع پواسون، $E[Y_t|\mathcal{F}_{t-1}^{Y,\lambda}] = Var[Y_t|\mathcal{F}_{t-1}^{Y,\lambda}] = \lambda_t$ را داریم. علاوه بر این، مدل (۳.۲) را می‌توان به عنوان واریانس ناهمسانی شرطی اتورگرسیو صحیح مقدار[1] $INGARCH(1, 1)$ تعریف کرد. یعنی یک مدل واریانس ناهمسانی شرطی اتورگرسیو[2] $GARCH$ صحیح مقدار. زیرا که ساختار آن مشابه مدل $GARCH$ معمول است که به موجب آن نوسانات بر روی مقادیر گذشته‌ی خود و مجذور پاسخ رگرسیون می‌شود. با در نظر گرفتن تابع پیوند کانونی $\nu_t \equiv \log \lambda_t$ خانواده مدل‌های لگاریتم خطی اتورگرسیو به صورت

$$Y_t|\mathcal{F}_{t-1}^{Y,\nu} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + a_1 \nu_{t-1} + b_1 log(Y_{t-1} + 1), \quad t \geq 1. \tag{۶.۲}$$

حاصل می‌شود که در آن $\mathcal{F}_t^{Y,\nu}$ یک σ میدان تولید شده از $\{Y_0, ..., Y_t, \nu_0\}$ است یعنی $\mathcal{F}_t^{Y,\nu} = \sigma(Y_s, \nu_0, s \leq t)$. به‌طور کلی، پارامترهای d, a_1, b_1 می‌توانند مثبت یا منفی باشند. اما باید شرایط خاصی برقرار باشد تا یک سری زمانی مانا به‌دست آوریم.

[1] Integer-Valued Autoregressive conditional heteroskedasticity
[2] Autoregressive conditional heteroskedasticity

۲.۱.۲ تحلیل درستنمایی

در این بخش تحلیل ماکسیمم درستنمایی شرطی را برای مدل خطی (۳.۲) به طور مختصر بیان می‌کنیم. این روش برای مدل‌های (۵.۲) و (۶.۲) کاملا مشابه است. در مدل (۳.۲) پارامتر $\boldsymbol{\theta}$ را به صورت یک بردار سه بعدی با پارامترهای مجهول $\boldsymbol{\theta} = (d, a_1, b_1)'$ در نظر بگیرید. مقدار اولیه پارامتر به صورت $\boldsymbol{\theta}_0 = (d_0, a_{10}, b_{10})'$ است. سپس تابع درستنمایی شرطی برای θ بر اساس مدل (۳.۲) با توجه به مقدار اولیه λ_0 توسط رابطه‌ی زیر به‌دست می‌آید:

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n} \frac{\exp(-\lambda_t(\boldsymbol{\theta}))\lambda_t^{Y_t}(\boldsymbol{\theta})}{Y_t!}.$$

که در آن از فرض پواسون $\lambda_t(\boldsymbol{\theta}) = d + a_1\lambda_{t-1}(\boldsymbol{\theta}) + b_1 Y_{t-1}$ توسط (۳.۲) و $\lambda_t = \lambda_t(\boldsymbol{\theta}_0)$ استفاده می‌کنیم. ازاین‌رو، لگاریتم تابع درستنمایی را می‌توان به صورت

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} \left(Y_t \log(\lambda_t(\boldsymbol{\theta})) - \lambda_t(\boldsymbol{\theta})\right), \tag{۷.۲}$$

نوشت. همچنین تابع امتیاز به‌صورت زیر به‌دست می‌آید:

$$S_n(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})} = \sum_{t=1}^{n} \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1\right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{۸.۲}$$

که در آن $\partial\lambda_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ یک بردار سه بعدی با اجزای داده شده در رابطه‌ی زیر است:

$$\frac{\partial \lambda_t}{\partial d} = 1 + a_1 \frac{\partial \lambda_{t-1}}{\partial d}, \qquad \frac{\partial \lambda_t}{\partial a_1} = \lambda_{t-1} + a_1 \frac{\partial \lambda_{t-1}}{\partial a_1}, \qquad \frac{\partial \lambda_t}{\partial b_1} = Y_{t-1} + a_1 \frac{\partial \lambda_{t-1}}{\partial b_1}. \tag{۹.۲}$$

اگر جواب معادله $S_n(\theta) = 0$ ، وجود داشته باشد؛ برآوردگر ماکسیمم درستنمایی شرطی θ را به‌دست می‌دهد که توسط $\hat{\theta}$ مشخص می‌شود. علاوه‌براین، همان‌طور که در رابطه زیر مشاهده می‌شود؛ ماتریس هسین در مدل (۳.۲) با تفاوت بیشتری از معادله‌ی (۸.۲) به‌صورت

$$\begin{aligned} \mathbf{H}_n(\boldsymbol{\theta}) &= -\sum_{t=1}^{n} \frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}, \\ &= \sum_{t=1}^{n} \frac{Y_t}{\lambda_t^2(\boldsymbol{\theta})} \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)' - \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1\right) \frac{\partial^2 \lambda_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}, \end{aligned} \tag{۱۰.۲}$$

به‌دست می‌آید. ماتریس اطلاع شرطی توسط رابطه زیر به‌دست می‌آید:

$$\mathbf{G}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} Var\left[\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \middle| \mathcal{F}_{t-1}^{Y,\lambda}\right] = \sum_{t=1}^{n} \frac{1}{\lambda_t(\boldsymbol{\theta})} \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)', \tag{۱۱.۲}$$

این رابطه نقش مهمی در توزیع مجانبی برآورد ماکسیمم درست‌نمایی[3] $\hat{\boldsymbol{\theta}}$ MLE دارد. به‌طور خاص تحت شرایط نظم، می‌توان ثابت کرد که $\hat{\boldsymbol{\theta}}$ سازگار و مجانبا نرمال است. یعنی: $\sqrt{n}(\hat{\theta} - \boldsymbol{\theta}_0) \longrightarrow \mathcal{N}(0, \mathbf{G}^{-1})$، که در آن $\mathbf{G}$ به صورت زیر تعریف می‌شود:

$$\mathbf{G}(\boldsymbol{\theta}) = E\left[\frac{1}{\lambda_t}\left(\frac{\partial \lambda_t}{\partial \theta}\right)\left(\frac{\partial \lambda_t}{\partial \boldsymbol{\theta}}\right)'\right],$$

[3] Maximum likelihood estimation

که در آن $E[\cdot]$ امید ریاضی است. تمام مقادیر فوق را می‌توان محاسبه کرد و برای پیش‌بینی‌ها، فواصل اطمینان و غیره مورد استفاده قرارمی‌گیرد. اگرچه فرمول‌های بالا برای مدل خطی (۳.۲) ارائه شده‌اند. اما می‌توانند به‌طور مناسب برای مدل لگاریتم خطی (۶.۲) اصلاح شوند. در ادامه یک مدل سری‌زمانی شمارشی فضایی بیان و مورد مطالعه قرار می‌گیرد که تاثیر مکان را به این نوع مدل‌ها اضافه می‌کنیم.

۳ یک مدل اتورگرسیو همبسته فضایی برای داده‌های شمارشی

داده‌های شمارشی در آمار فضایی برخلاف سری زمانی عمدتا از طریق یک فرآیند پنهان مورد بررسی قرار می‌گیرد. به عنوان مثال، مدل $INGARCH$ که توسط فرلند و همکاران (۲۰۰۶) و هنین (۲۰۰۳) بیان شده است و در آن مدل داده‌ها پواسون است و تابعی از شمارش‌های گذشته و امیدهای گذشته‌ی آن است.
ساختار مدل به این گونه است که، $Z(s_i,t)$ را مشاهدات در مکان‌های فضایی ثابت به‌صورت $s_i = \{s_1,\dots,s_n\}$ و نقاط زمانی $t = \{1,\dots,T\}$ درنظر می‌گیریم. سپس، $Y(s_i,t)$ را به عنوان یک متغیر تصادفی پنهان در موقعیت مکانی ـ زمانی (s_i,t) در یک ناحیه فضایی به‌صورت $\{s_j : s_i$ یک همسایه فضایی $s_j\}$ تعریف می‌کنیم. توجه داشته باشید که s_i به طور معمول یک بردار در $\mathbb{R}^2$ است. فرض کنید $\mathcal{H}_{Z(s_i,t)}$ گذشته فرایند مشاهدات در موقعیت مکانی s_i تا دوره‌ی زمانی t باشند. فرآیند واریانس ناهمسانی شرطی اتورگرسیو همبسته‌ی فضایی[4] $SPINGARCH$، $[Z(s_i,t)|Y(s_i,t),\mathcal{H}_{Z(s_i,t)}]$ دارای توزیع پواسون شرطی با تابع احتمال

$$f\left(Z(s_i,t)|Y(s_i,t),\mathcal{H}_{Z(s_i,t)}\right) = \exp\left[A_i\left(Y(s_i,t),\mathcal{H}_{Z(s_i,t)}\right)Z(s_i,t) - B_i\left(Y(s_i,t),\mathcal{H}_{Z(s_i,t)}\right)C\left(Z(s_i,t)\right)\right] \quad (۱.۳)$$

است، که در آن

$$\begin{aligned}
\exp\left[A_i\left(Y(s_i,t),\mathcal{H}_{Z(s_i,t)}\right)\right] &= \exp\left[Y(s_i,t)\right] + \eta Z(s_i,t-1) + \kappa E\left[Z(s_i,t-1)|Y(s_i,t-1),\mathcal{H}_{Z(s_i,t-1)}\right] \\
B_i\left(Y(s_i,t),\mathcal{H}_{Z(s_i,t)}\right) &= \exp\left[A_i\left(Y(s_i,t),\mathcal{H}_{Z(s_i,t)}\right)\right] \\
C\left(Z(s_i,t)\right) &= -\log\left[Z(s_i,t)\right],
\end{aligned}$$

به‌طوری‌که

$$\begin{aligned}
Y(s_i,t)|y_t(N_i) &\sim N\left(\mu(s_i,t),\sigma^2\right) \\
\mu(s_i,t) &= \alpha(s_i) + \zeta\sum_{s_j\in N_i}\{y(s_j,t)-\alpha(s_j)\}.
\end{aligned} \quad (۲.۳)$$

مدل $SPINGARCH$ از ترکیب (۱.۳)، که یک زنجیره مارکوف در زمان و (۲.۳) که یک زنجیره مارکوف در فضا است، تشکیل شده است. مشاهده $Z(s_i,t)$ روی فرایند گذشته کامل و فرآیند پنهان حال $Y(s_i,t)$ با استفاده از توزیع شرطی کامل در فضا در هر نقطه‌ی زمانی، مدل‌بندی شده است. فرضیات مارکوف، شرط را به یک مرحله زمانی و همسایگی‌های فضایی کاهش می‌دهد. قرار دهید $\mathbf{A}$ یک ماتریس همسایگی[5] باشد به‌طوری‌که اگر s_i و s_j همسایه باشند؛ ورودی $(i,j) = 1$ را داریم و ζ را به $\zeta \in (\psi_{(1)}^{-1},\psi_{(n)}^{-1})$ محدود می‌کنیم که $\psi_{(k)}$ ، k امین بزرگترین مقدار ویژه‌ی $\mathbf{A}$ است. نتیجه‌ی مدل فرآیند پنهان در هر زمان t ، یک اتورگرسیو شرطی یا مدل CAR است که در کریستی و ویکل (۲۰۱۵) استفاده شده است. فرض کنید $\boldsymbol{Y}_t \sim N\left(\boldsymbol{\alpha},(I-C)^{-1}M\right)$ باشد؛ آنگاه CAR دارای توزیع توام $\boldsymbol{\alpha} = (\alpha(s_i),\dots,\alpha(s_n))$ است که C یک ماتریس $n\times n$ با ورودی‌های ζ است و اگر $s_j,t \in N(s_i)$ باشد؛ مقدار یک و در غیر این صورت برابر با ۰ خواهد بود. M یک ماتریس قطری با ورودی σ است. برای نمادگذاری راحت‌تر، $(s_i,t) \equiv \exp[A_i(.)]$ را تعریف می‌کنیم. برای

[4] Spatially Correlated Integer Autoregressive Conditionally Heteroskedastic
[5] Neighborhood matrix

$\alpha(s_i) = \alpha$ ، $i = ۱, \ldots, n$ را نظر می‌گیریم، اما این شرایط را می‌توان به‌عنوان، توابع متغیرهای کمکی فضایی مدل‌بندی کرد. برای سادگی نمادها فرض کنید $\boldsymbol{\theta}$ بردار پارامترهای $\sigma^۲$ و ζ را نشان دهد و $\Sigma_{(\boldsymbol{\theta})} \equiv (I-C)^{-۱}M$ به عنوان ماتریس کوواریانس پنهان $n \times n$ باشد.

به‌طور خلاصه مدل $SPINGARCH$ را می‌توان به‌صورت زیر نوشت:

$$\begin{aligned} Z(s_i,t)|\lambda(s_i,t) &\sim pois(\lambda(s_i,t)) \\ E\left[Z(s_i,t)|\lambda(s_i,t)\right] &= \lambda(s_i,t) \\ \boldsymbol{\lambda}_t &= \exp(\boldsymbol{Y}_t) + \eta \boldsymbol{Z}_{t-۱} + \kappa \boldsymbol{\lambda}_{t-۱} \\ \boldsymbol{Y}_t &\sim Gauss(\boldsymbol{\alpha}, (I_{n,n} - C)^{-۱}M), \end{aligned} \tag{۳.۳}$$

که در آن $\boldsymbol{\lambda}_t = (\lambda(s_۱,t), \lambda(s_۲,t), \ldots, \lambda(s_n,t))^T$ می‌باشد. که $\boldsymbol{\lambda}_t$ یک زنجیره مارکوف در فضای $(\mathbb{R}^+)^n$ است. در ادامه تحلیل بیزی مدل بیان شده، ارائه می‌شود.

۱.۳ استنباط بیزی

تجزیه و تحلیل بیزی مدل $SPINGARCH$ را می‌توان با استفاده از برنامه‌های نرم‌افزاری پیشنهادی جوزف (۲۰۱۶) انجام داد. توزیع پیشین توام پارامترهای مدل را به صورت

$$\pi(\boldsymbol{\theta}) = \pi(\mu|\kappa)\pi(\kappa)\pi(\boldsymbol{\alpha})\pi(\sigma)\pi(\zeta)$$

لحاظ کرده‌ایم، که در آن فرض استقلال را در قسمت‌های مختلف به جز μ و κ به دلیل محدودیت $\mu + \kappa < ۱$ ، در نظر می گیریم.

در آن صورت توزیع شرطی کامل $\boldsymbol{\theta} = (\eta, \kappa, \alpha, \zeta, \sigma^۲)^T$ به‌صورت زیر حاصل می‌شود:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{Y}) \propto \prod_{t=۱}^{T} \pi(\boldsymbol{Z}_t|\boldsymbol{\lambda}_t)\pi(\boldsymbol{\lambda}_t|\boldsymbol{\lambda}_{t-۱}, \boldsymbol{Z}_{t-۱}, \boldsymbol{\theta}, \boldsymbol{Y}_t)\pi(\boldsymbol{Y}_t|\boldsymbol{\theta})\pi(\boldsymbol{\lambda}_۰|\boldsymbol{\theta})\pi(\boldsymbol{Z}_۰|\boldsymbol{\lambda}_۰)\pi(\boldsymbol{\theta}) \tag{۴.۳}$$

و توزیع شرطی کامل Y به‌صورت

$$\pi(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta}) \propto \prod_{t=۱}^{T} \pi(\boldsymbol{Z}_t|\boldsymbol{\lambda}_t)\pi(\boldsymbol{\lambda}_t|\boldsymbol{\lambda}_{t-۱}, \boldsymbol{Z}_{t-۱}, \boldsymbol{\theta}, \boldsymbol{Y}_t)\pi(\boldsymbol{Y}_t|\boldsymbol{\theta})\pi(\boldsymbol{\lambda}_۰|\boldsymbol{\theta})\pi(\boldsymbol{Z}_۰|\boldsymbol{\lambda}_۰) \tag{۵.۳}$$

به‌دست می‌آید. برای انجام هر نوع استنباط از زنجیره مارکوف مونت کارلویی برای تولید نمونه از چگالی شرطی کامل $\boldsymbol{Y}$ استفاده می‌کنیم که نیاز به ارزیابی به‌صورت زیر دارد:

$$\begin{aligned} \log(\boldsymbol{Y}|\alpha,\sigma,\zeta) \propto{}& \frac{-T \times n}{۲}\log(۲\pi) + \frac{۱}{۲}\log|\Sigma_f^{-۱}(\boldsymbol{\theta})| \\ & -\frac{۱}{۲}(\boldsymbol{Y}-\boldsymbol{\alpha})^T\Sigma_f^{-۱}(\boldsymbol{\theta})(\boldsymbol{Y}-\boldsymbol{\alpha}) \end{aligned} \tag{۶.۳}$$

توجه داشته باشید که همان‌طور که ساختار همسایگی فرض می‌شود؛ ثابت C برای همه دوره‌های زمانی یکسان است. بنابراین می توانیم $\Sigma_f(\boldsymbol{\theta})$ رابه‌صورت $(I_{n\times T, n\times T} - I_{t,t} \otimes C)^{-۱} I_{t,t} \otimes M$ و به عنوان ماتریس کوواریانس فضا ـ زمان کامل بنویسیم. کمبود ساختار کواریانس به این معناست که تنها محاسبات $\frac{۱}{۲}(Y-\alpha)^T\Sigma_f^{-۱}(\theta)(Y-\alpha)$ است که نیاز است برای همسایگان فضایی رخ بدهد. بنابراین، دشوارترین بخش محاسبه چگالی لگاریتم، محاسبه‌ی دترمینان $\log|\Sigma_F^{-۱}(\theta)|$

است.

با استفاده از الگوریتم‌های مونت کارلویی همانند نمونه‌گیر گیبس به‌همراه روش مونت کارلویی همیلتونی یا روش‌های مشابه‌ی که در بسته نرم افزاری STAN است (جوزف ، ۲۰۱۶)، این امکان را می‌دهد تا برازش مدل نسبتاً سریع انجام شود. برای انجام ارزیابی مدل به مقادیر P پیش بینی کننده گذشته تکیه می‌کنیم. این روش بعد از نمونه‌گیری پارامترها از توزیع پیشین، مجموعه داده‌های جدید را نمونه برداری می‌کند. آمارها براساس داده‌های جدید محاسبه می‌شوند و با آمارهای مجموعه داده اصلی مقایسه می‌شوند. برای هر مجموعه داده $T_۱(Z)$ آماره‌ی فضایی موران، $T_۲(Z)$ لگاریتم نسبت واریانس به میانگین، $T_۳(Z)$ واریانس اختلاف $Z(s_i,t)-Z(s_i,t-۱)$ و $T_۴(Z)$ میانگین نمونه اتورگرسیو با یک تاخیر، را به‌دست می‌آوریم. $T_۱$ و $T_۴$ ساختار فضایی- زمانی را در داده‌ها ضبط می‌کنند؛ در حالی‌که $T_۲$ و $T_۳$ تغییرپذیری و ناهمواری در هر مکان را می‌گیرند. همان‌طور که در بخش بعد خواهیم دید هیچ یک از این آمارها برای تمایز بین $SPINGARCH$ ، $SPINGARCH$ زمان ثابت (فقط مکان) و مدل $INGARCH$ کافی نیست. در عوض، بررسی هر چهار پیشگوی گذشته معمولا برای تمایز مناسب مدل‌ها از یکدیگر است و در مورد مقادیر بزرگتر و کوچکتر از ۰/۵ باید در برازش مدل تردید کرد.

۴ مطالعه شبیه‌سازی

هدف از شبیه‌سازی‌ها، درک تأثیر مدل آماری اشتباه است و نشان دادن چگونگی بررسی پیشگوی پسینی می تواند در انجام انتخاب مدل بین $SPINGARCH$ ، $SPINGARCH$ زمان ثابت و مدل $INGARCH$ کمک کند. یک شبکه ۲۰ × ۲۰ را با ساختار چهارنزدیکترین همسایگی در نقطه زمانی برای شبیه‌سازی محدوده فضایی در نظر گرفته‌ایم. ابتدا فرض می‌کنیم مدل شبیه‌سازی داده‌ها به‌صورت زیر است:

$$\begin{aligned} Z(s_i,t) &\sim \text{Poisson}(\lambda(s_i,t)) \\ \lambda(s_i,t) &= \exp(U(s_i)) + ۰/۲Z(s_i,t-۱) + ۰/۳\lambda(s_i,t-۱) \\ U(s_i)|u(N_i) &\sim N(\mu(s_i), ۰/۵) \\ \mu(s_i) &= ۰/۲۴۵ \sum_{s_j \in N_i} u(s_j) \end{aligned} \tag{۱.۴}$$

مدل ارائه شده در (۱.۴) فرض می‌کند که تغییر امید ریاضی در هر دوره زمانی ناشی از شمارش قبلی، امید قبلی و یک عامل متغیر فضایی است که با گذشت زمان تغییر نمی‌کند. سپس، مدل $SPINGARCH$، (۳.۳) ، مدل $SPINGARCH$ زمان ثابت (فقط فضایی) و نیز مدل $INGARCH(۱,۱)$، را به داده‌های تولید شده برازش می‌کنیم. بازه باورمندی بیزی برای پارامترهای هر مدل در جدول ۱ آورده شده است. از نتایج جدول ۱ در می‌یابیم که مدل واقعی، $SPINGARCH$

جدول ۱ : بازه باورمندی ٪۹۵ برای برآورد پارامترها با استفاده از مدل ۱.۴ به عنوان مکانیزم تولید

	$SPINGARCH$	$SPINGARCH$ زمان ثابت	$INGARCH(۱,۱)$
$\alpha = ۰$	$(-۷/۹, -۶/۵)$	$(-۱/۰, ۰/۳۴)$	$(-۱/۸, -۱/۷)$
$\eta = ۰/۲$	$(۰/۰۸, ۰/۰۹)$	$(۰/۱۷, ۰/۲۱)$	$(۰/۳۸, ۰/۴۰)$
$\kappa = ۰/۳$	$(۰/۹۰, ۰/۹۲)$	$(۰/۲۳, ۰/۳۶)$	$(۰/۵۶, ۰/۵۸)$
$\sigma^۲ = ۰/۵$	$(۰/۷۹, ۱/۰۰)$	$(۰/۳۶, ۰/۶۹)$	
$\zeta = ۰/۲۴۵$	$(۰/۲۴۸, ۰/۲۴۹)$	$(۰/۲۳۸, ۰/۲۴۹)$	

زمان ثابت، پارامترهای مدل را پوشش می‌دهد در حالی‌که مدل $SPINGARCH$ و $INGARCH$ مقدار κ را بیشتر نشان

می‌دهند و اصطلاحا متورم می‌کنند. علاوه بر این، مدل $SPINGARCH$ مقادیر $\sigma^۲$ و ζ را متورم می‌کنند. در ادامه از توزیع پسین برای محاسبه پیشگوی پسینی استفاده می‌کنیم، که نتایج بررسی در جدول ۲ آمده است (نیکلاس و همکاران ، ۲۰۱۸). در جدول ۲ ملاحظه می‌شود که وقتی داده‌ها مطابق با مدل واقعی هستند؛ توزیع پیشین حاصل می‌تواند داده‌ها را به‌طوری جاگذاری کند که داری مشخصات مشابه با داده های اصلی باشند. با این حال، اگر این مدل نادرست باشد؛ این کار را نمی‌توان انجام داد. در حالی که $INGARCH$ قادر به جاگذاری هیچ یک از خصوصیات مرتبه دوم داده های اصلی نیست؛ $SPINGARCH$ به جز آماره‌ی $AR(۱)$ برای همه به خوبی عمل می‌کند. در حالی که به نظر می رسد $SPINGARCH$ و $SPINGARCH$ بدون تغییر زمان قادر به تولید خواص بسیار مشابهی در مرتبه دوم هستند؛ نیکلاس و همکاران (۲۰۱۸) نشان دادند که وقتی η نزدیک لبه فضای پارامتر باشد دیگر این مورد نیست. حال یک دامنه

جدول ۲: بررسی‌های پیشگوی پسینی با شبیه‌سازی توزیع پسین

بررسی‌های پیشگوی پسینی	$SPINGARCH$	$SPINGARCH$ زمان ثابت	$INGARCH(۱,۱)$
۰	۰/۳۲	۰/۳۷	آماره‌ی موران
۰	۰/۳۸	۰/۲۶	نسبت واریانس به میانگین
۰/۰۴	۰/۵۸	۰/۲	واریانس اختلاف
۱	۰/۵۸	۰/۹۰	$AR(۱)$

فضایی کوچکتر، ۱۰ × ۱۰ را در نظر می‌گیریم و فرض می‌کنیم مکانیزم مولد مدل کامل $SPINGARCH$ به صورت زیر است:

$$\begin{aligned} Z(s_i,t) &\sim \text{Poisson}(\lambda(s_i,t)) \\ \lambda(s_i,t) &= exp(U(s_i)) + ۰/۲Z(s_i,t-۱) + ۰/۳\lambda(s_i,t-۱) \\ Y(s_i)|y(N_i) &\sim N(\mu(s_i), ۰/۵) \\ \mu(s_i,t) &= ۰/۲۴۵ \sum_{s_j \in N_i} y(s_j,t). \end{aligned} \tag{۲.۴}$$

نتایج برازش هر سه مدل در جدول ۳ و ۴ نشان داده شده است.

جدول ۳: بازه باورمندی ۹۵٪ برای برآورد پارامترها با استفاده از مدل ۱.۴

	$SPINGARCH$	$SPINGARCH$ زمان ثابت	$INGARCH(۱,۱)$
$\alpha = ۰$	$(-۰/۰۲, ۰/۱۹)$	$(۰/۶۷, ۰/۸۳)$	$(۰/۶۶, ۰/۷۶)$
$\eta = ۰/۲$	$(۰/۱۷, ۰/۲۳)$	$(۰/۳۵, ۰/۳۷)$	$(۰/۳۵, ۰/۳۹)$
$\kappa = ۰/۳$	$(۰/۲۵, ۰/۳۳)$	$(۰/۰۱, ۰/۰۷)$	$(۰/۰۳, ۰/۰۸)$
$\sigma^۲ = ۰/۵$	$(۰/۴۳, ۰/۵۳)$	$(۰/۰۱, ۰/۰۳)$	
$\zeta = ۰/۲۴۵$	$(۰/۲۴۳, ۰/۲۴۶)$	$(۰/۰۹, ۰/۲۴۹)$	

همان‌طور که در جدول ۳ نشان داده شده است؛ اگر مکانیزم تولید داده واقعی شبیه به (۲.۴) باشد؛ تفاوت زیادی بین $SPINGARCH$ زمان ثابت و مدل $INGARCH$ وجود ندارد.

جدول ۴: بررسی‌های پیشگوی پسینی با شبیه‌سازی توزیع پسین

بررسی‌های پیشگوی پسینی	$SPINGARCH$	$SPINGARCH$ زمان ثابت	$INGARCH(۱,۱)$
آماره‌ی موران	۰/۳۸	۰	۰
نسبت واریانس به میانگین	۰/۴۳	۰	۰
واریانس اختلاف	۰/۵۶	۰	۰
$AR(۱)$	۰/۵۴	۰/۳۲	۰/۳۱

بحث و نتیجه‌گیری

در این مقاله مدل‌های سری‌زمانی شمارشی بر اساس فرایند پواسن ارائه و مورد مطالعه قرار گرفتند. همچنین ساختار مدل‌های سری‌زمانی شمارشی با در نظر گرفتن همبستگی فضایی از طریق نزدیکترین همسایه بیان شد و در یک مطالعه شبیه‌سازی نحوه تولید نمونه از این مدل‌ها ارائه گردید. به‌وسیله رهیافت بیزی، پارامترهای مدل‌ها براورد شدند و مورد تحلیل قرار گرفتند. در مطالعه شبیه‌سازی، مدل‌های INGARCH، SPINGARCH و SPINGARCH زمان ثابت را به‌وسیله پیشگوی پسینی مورد مقایسه قرار دادیم. نتایج بیانگر این بود که مدل SPINGARCH عملکرد مناسبی نسبت به بقیه مدل‌ها داشت.

مراجع

Brockwell, P. J., Davis, R.A. (1991), Time Series: Data Analysis and Theory, second ed , *Springer*, New York.

Cressie, N., Wikle, C. K. (2015). Statistics for Spatio-Temporal Data. John Wiley & Sons.

Ferland, R., Latour, A., Oraichi, D. (2006), Integer-Valued GARCH Processes, *J. Time Ser. Anal.*, **27**, 923–942.

Heinen, A. (2003), Modelling Time Series Count Data: An Autoregressive Conditional Poisson Model, *Technical Report MPRA* Paper 8113, University Library of Munich, Germany.

Joseph, M. (2016). Exact Sparse Car Models in Stan. Accessed: **2017-07-17**.

Mccillagh, P., Nelder, J.A. (1989), Generalized Linear Models, second ed. Chapman & Hall, London.

Nelder, J. A., Wedderburn, R. W. M. (1972), Generalized Linear Models, *J. R. Stat. Soc. Ser. A*, **135**, 370–384.

Nicholas J. C., Mark, S. K., Philip M. D. (2018), A Spatially Correlated Auto-regressive Model for Count Data, stat AP.

Priestley, M. B. (1981), Spectral Analysis and Time Series, Academic Press, London.

Rydberg, T. H., Shephard, N. (2000), A Modeling Framework for the Prices and Times of Trades on the New York Stock Exchange. In: Fitzgerlad, W.J., Smith, R.L., Walden, A.T., Young, P.C. (Eds.),

Nonlinear and Nonstationary Signal Processing. Isaac Newton Institute and Cambridge University Press, Cambridge, pp. 217–246.

Streett, S. (2000), Some Observation Driven Models for Time Series of Counts, Ph. D. thesis, Colorado State University, Department of Statistics.

Zhu, R., Joe, H. (2006), Modelling Count Data Time Series with Markov Processes Based on Binomial Thinning, *J. Time Ser. Anal*, **27**, 725–738.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

برآورد تغییرنگار با استفاده از روش رگرسیون بردار پشتیبان

عباس توسلی[۱]، یدا... واقعی[۱]، علیرضا ناظمی[۲]
[۱]گروه آمار، دانشکده علوم ریاضی و آمار، دانشگاه بیرجند
[۲]دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود

چکیده: برازش مدل تغییرنگار یک گام اساسی در اکثر مطالعات زمین‌آماری است. در حال حاضر، روش‌های حداکثر درستنمایی و حداقل مربعات برای برازش مدل تغییرنگار مورد استفاده قرار می‌گیرند. روش حداکثر درستنمایی نیاز به پیش‌فرض‌هایی دارد که به سختی برقرار خواهد بود. همچنین روش حداقل مربعات نیاز به انتخاب یک مدل اولیه مناسب برای تغییرنگار دارد که در صورت انتخاب نادرست مدل اولیه، بر دقت پیشگویی نهایی اثر منفی خواهد گذاشت. اشکال دیگر روشهای زمین‌آماری آن است که مدل‌های متداول و معتبر تغییرنگار رفتار هموار و افزایشی دارند؛ در حالیکه تغییرنگار داده‌ها اغلب نوسانات زیادی دارد و بهترین مدل هم به آن به خوبی برازش نمی‌شود. هدف این مطالعه استفاده از یکی از روش‌های یادگیری ماشین بنام «رگرسیون بردار پشتیبان» برای برآورد تغییرنگار و استفاده از آن در پیشگویی کریگیدن است. روش رگرسیون بردار پشتیبان می‌تواند بدون نیاز به انتخاب مدل اولیه و بصورت خودکار تغییرنگار را برآورد کند. در این مطالعه با استفاده از چندین مجموعه داده فضایی شبیه‌سازی شده، دقت روش رگرسیون بردار پشتیبان با روش حداقل مربعات در برآورد تغییرنگار و پیشگویی با کریگیدن مقایسه گردید. نتایج نشان داد که روش رگرسیون بردار پشتیبان می‌تواند رقیب مناسبی برای روش‌های زمین‌آماری در برآورد تغییرنگار باشد.

واژه‌های کلیدی: وابستگی فضایی، تغییرنگار، رگرسیون بردار پشتیبان، کریگیدن.
کد موضوع‌بندی ریاضی (۲۰۱۰): 07 – 62، 04 – 62، 62G08، 62J02، 62M45.

۱ مقدمه

داده‌هایی که در یک محدوده فضایی (یا جغرافیایی) مشاهده شده و بر اساس موقعیتشان در آن محدوده به یکدیگر وابسته‌اند داده‌های فضایی گفته می‌شود. یکی از اهداف مهم در تحلیل این داده‌ها پیشگویی فضایی در موقعیت‌های جدید است. با اینحال، دستیابی به این هدف مستلزم تعیین ساختار همبستگی فضایی موجود در داده‌هاست. تابعی که برای بیان ساختار همبستگی داده‌ها مورد استفاده قرار می‌گیرد تابع «تغییرنگار» است که اولین بار توسط ماترن (۱۹۶۳) به این اسم نامگذاری

[۱]نام و ایمیل ارائه دهنده مقاله: عباس توسلی، Tavassoli.a@Birjand.ac.ir

شد. مدلبندی دقیق تغییرنگار، گامی اساسی در اکثر مطالعات زمین‌آماری است. تغییرنگار علاوه بر توصیف رفتار فضایی متغیر مورد مطالعه، تاثیر قابل توجهی در دقت پیشگویی از طریق کریگیدن دارد (دساسیز و رنارد ، ۲۰۱۳). در حالت کلی، انتخاب مدل اولیه برای تغییرنگار شهودی بوده و ممکن است بهترین مدل تغییرنگار انتخاب نشود. در مورد مدل‌سازی تغییرنگار تاکنون مطالعات بسیاری صورت گرفته است. این مسئله هنوز هم در کاربردهای زمین‌آماری با دشواری‌هایی همراه است (الیور و وبستر ، ۲۰۱۵). در حال حاضر، دو مورد از روش‌های معمول برای مدلبندی تغییرنگار، روش حداکثر درستنمایی و روش حداقل مربعات می‌باشند (کرسی ، ۱۹۸۵، ۱۹۹۳؛ الیور و وبستر ، ۲۰۰۷) . در روش حداکثر درستنمایی، با ماکزیموم کردن تابع درستنمایی توام برای مقادیر مشاهده شده، مدل‌های تغییرنگار انتخاب و برازش داده می‌شود. با اینحال، روش حداقل مربعات از نظر محاسباتی ساده‌تر بوده و در اکثر بسته‌های نرم‌افزار زمین‌آمار مورد استفاده قرار گرفته است.

در تمامی روش‌هایی که تاکنون برای برازش مدل تغییرنگار بکار گرفته شده است، ساختار اولیه مدل بعنوان پیش‌فرض انتخاب و پذیرفته شده و سپس پارامترهای بهینه برای این مدل از پیش تعیین شده برآورد شده است. با اینحال، انتخاب یک مدل تغییرنگار بهینه از بین تعداد محدودی از ساختارهای اولیه که تاکنون معرفی شده‌اند مشکل است و بنابراین هیچگاه نمی‌توان یک مدل بهینه برای تغییرنگار به منظور کاهش خطای پیشگویی پیدا کرد. روش‌های مبتنی بر هوش مصنوعی می‌توانند بر اساس مقادیر مشاهده شده مدل بهینه را پیدا کنند. یکی از این روش‌ها استفاده از رگرسیون بردار پشتیبان [1] (SVR) است. اولین بار هوانگ و همکاران (۲۰۱۲) روش SVR را برای برازش تغییرنگار مورد استفاده قرار دادند. یوکو و همکاران (۲۰۱۵) از روشی مشابه برای پیشگویی ضخامت شکاف زغال سنگ استفاده کردند. همچنین هان و همکاران (۲۰۱۶) نشان دادند که روش مبتنی بر SVR در مقایسه با روش WLS برای مدلبندی تغییرنگار قادر است بدون در نظر گرفتن شکل مدل اولیه، به شکل دقیق‌تری به تغییرنگار تجربی برازش یابد و دقت پیشگویی کریگیدن را افزایش دهد. با توجه به اینکه روش SVR بر اصل کمینه‌سازی ریسک ساختاری استوار است دارای توانایی بالایی در تعمیم‌پذیری بوده و فرمول‌بندی آن بر مبنای داده‌های مشاهده شده ساده است.

هدف این مطالعه بررسی عملکرد روش یادگیری ماشین SVR در برازش تغییرنگار به منظور استفاده از آن در روش کریگیدن معمولی است. بدین‌منظور، در بخش ۲ توضیح مختصری در رابطه با مدلبندی تغییرنگار ارائه شده و در بخش ۳ روش SVR معرفی شده است. در بخش ۴ با در نظر گرفتن چندین مجموعه داده فضایی شبیه‌سازی شده، به مقایسه روش SVR و روش حداقل مربعات در برازش مدل تغییرنگار پرداخته شده است. مقایسه دو روش بر اساس دقت پیشگویی کریگیدن انجام گرفته است.

۲ مدلبندی زمین‌آماری تغییرنگار

برای برآورد و مدلبندی تغییرنگار نیاز به نمادها و مفروضاتی برای داده‌های فضایی داریم. فرض کنید $\{Z(s); s \in D\}$ یک میدان تصادفی باشد بطوریکه $D \subset R^d, d \geq ۱$ است. این میدان مانای ذاتی نامیده می‌شود هرگاه میانگین میدان تصادفی ثابت (مستقل از s) بوده و واریانس تفاضل بین مقادیر میدان تصادفی در دو موقعیت s و $s+h$ فقط تابعی از فاصله بین آنها باشد (محمدزاده ، ۱۳۹۸). به عبارت دیگر، هرگاه برای هر $s, s+h \in D$ دو شرط زیر برقرار باشد:

$$E\left(Z(s) - Z(s+h)\right) = ۰ \quad (۱$$

$$Var\left(Z(s) - Z(s+h)\right) = ۲\gamma(h) \quad (۲$$

[1] Support Vector Regression

تابع $\gamma(h)$ که تنها تابعی از فاصله h است نیم‌تغییرنگار و $۲\gamma(h)$ تغییرنگار نامیده می‌شود؛ ولی در این مقاله هرکجا واژه تغییرنگار را بکار می‌بریم منظور همان $\gamma(h)$ است. در روش متداول (مبتنی بر روش حداقل مربعات)، مدل‌سازی تغییرنگار شامل دو مرحله است: ابتدا محاسبه تغییرنگار تجربی و سپس انتخاب و برازش یک مدل تغییرنگار معتبر به تغییرنگار تجربی داده‌ها. ماترن (۱۹۶۲) تغییرنگار تجربی را، برای برآورد کردن تغییرنگار، بصورت زیر پیشنهاد کرد:

$$\hat{\gamma}(h) = \frac{۱}{۲|N_h|} \sum_{|N_h|} (z(s_i + h) - Z(s_i))^۲ \qquad (۱.۲)$$

بطوریکه $Z(s_i)$ و $Z(s_i + h)$ مقادیر مشاهده شده برای متغیر Z در دو موقعیت s_i و $s_i + h$ هستند. و $|N_h|$ تعداد زوج مشاهدات متمایزی است که به فاصله h از یکدیگر قرار گرفته‌اند. برآوردگر ماترن ناریب و سازگار است (میراندا و دمیراندا ، ۲۰۱۱). در صورتی که γ تنها به طول فاصله لگ‌ها (و نه به جهت آنها) بستگی داشته باشد تغییرنگار همسانگرد نامیده می‌شود.

در دومین مرحله از برآورد تغییرنگار (یعنی انتخاب یک مدل تغییرنگار معتبر و برازش آن به تغییرنگار تجربی یا به بیان دیگر، مدل بندی تغییرنگار)، هر تابع پیوسته‌ای که داری خصوصیت همیشه منفی شرطی باشد می‌تواند بعنوان یک مدل معتبر تغییرنگار در نظر گرفته شود. این خصوصیت بیان می‌کند که برای هر $\{s_i \in D \subset R^d; ۱ \leq i \leq m\}$ و برای هر دنباله $\{a_i \in R; ۱ \leq i \leq m\}$ که $\sum_{i=۱}^{m} a_i = ۰$ ، آنگاه $\sum_{i=۱}^{m} \sum_{j=۱}^{m} a_i a_j \gamma(s_i - s_j) \leq ۰$. این ویژگی، اعتبار تغییرنگار را تضمین می‌کند. پرکاربردترین توابعی که برای مدلسازی تغییرنگار بکار برده می‌شوند توابع نمایی، گاوسی و کروی هستند که مدل آنها در جدول ۱ نشان داده شده است (الیور و وبستر ، ۲۰۰۷). هر ترکیب خطی مثبت از این مدل‌های تغییرنگار نیز می‌تواند بعنوان مدل معتبر در نظر گرفته شود.

جدول ۱: مدل‌های معتبر نیم‌تغییرنگار

مدل	نیم‌تغییرنگار
خطی	$\gamma(h) = c_۰ + ch$
نمایی	$\gamma(h) = c_۰ + c\left(۱ - \exp\left(-\frac{h}{a}\right)\right),\ h \in R^d,\ d \geq ۱$
گاوسی	$\gamma(h) = c_۰ + c\left(۱ - \exp\left(-\frac{h^۲}{a^۲}\right)\right),\ h \in R^d,\ d \geq ۱$
کروی	$\gamma(h) = \begin{cases} c_۰ + c\left(\frac{۳}{۲}\frac{h}{a} - \frac{۱}{۲}\frac{h^۳}{a^۳}\right) & ۰ \leq h < a \\ c_۰ + c & h > a \end{cases}$

در این مدل‌ها، $c_۰$ اثر قطعه‌ای، $c_۰ + c$ اِزاره و a دامنه تغییرنگار را نشان می‌دهند. این پارامترها از روی مشاهدات برآورد می‌شوند. دامنه، کوچکترین فاصله‌ای است که در آن مقدار $\gamma(h)$ برابر با اِزاره می‌شود. اگر فاصله دو مشاهده بزرگتر از مقدار دامنه باشد بر یکدیگر اثری نداشته و ناهمبسته خواهند بود. همچنین اثر قطعه‌ای که مقدار حدی تغییرنگار را در نقطه $h = ۰$ نشان می‌دهد از نظر تئوری باید صفر باشد، اما در عمل بدلیل خطای نمونه‌گیری، خطای اندازه‌گیری و یا تغییرات شدید کمیت مورد بررسی در موقعیت‌های نزدیک به هم، ممکن است بزرگتر از صفر باشد (محمدزاده ، ۱۳۹۸).

پس از انتخاب مدل اولیه برای تغییرنگار، بایستی پارامترهای مدل بهینه با کمک روش‌های مینیمم‌سازی استاندارد برآورد گردد. برای این منظور زوج نقاط $(h, \hat{\gamma}(h))$ همانند زوج داده‌های رگرسیونی در نظر گرفته شده و پارامترهای مدل به گونه‌ای برآورد می‌شوند که اختلاف بین تغییرنگار تجربی و مدل تغییرنگار به کمترین مقدار ممکن برسد. بر اساس این ایده از یکی از روشهای حداقل مربعات خطا استفاده می کنیم. تاکنون چندین روش حداقل مربعات برای مدل‌بندی تغییرنگار پیشنهاد شده است؛ از جمله، روش حداقل مربعات معمولی (OLS) ، روش حداقل مربعات وزنی (WLS) و روش حداقل مربعات تعمیم‌یافته (GLS) (کرسی ، ۱۹۹۳). در تمامی این روش‌ها، برازش مدل از طریق کمینه‌سازی یک تابع هزینه انجام می‌گیرد. تابع هزینه فاصله بین مدل برآورد شده و تغییرنگار تجربی را اندازه‌گیری می‌کند. روش OLS نسبت به دیگر روش‌ها ساده‌تر است؛ از طرفی، روش GLS نیاز به برآورد دقیق ماتریس کواریانس برآوردگر تغییرنگار دارد که بدست آوردن آن حتی برای فرآیند گاوسی بسیار مشکل است. مطالعات اخیر نشان داده است که نتایج بدست آمده از روش WLS در برآورد مدل تغییرنگار نسبت به دو روش دیگر رضایت‌بخش‌تر است (دساسیز و رنارد ، ۲۰۱۳؛ الیور و وبستر ، ۲۰۱۵) و در کنار پیشنهادهایی که برای انتخاب وزن‌ها در روش WLS شده است روش کرسی (۱۹۸۵) بیشترین کاربرد را دارد.
حالت‌هایی وجود دارد که هیچیک از مدل‌های مذکور بخوبی به تغییرنگار تجربی برازش داده نمی‌شوند. اشکال روش‌های زمین‌آماری این است که مدلهای متداول (و معتبر تغییرنگار) رفتار هموار و افزایشی دارند در حالیکه تغییرنگار داده‌ها اغلب نوسانات زیادی دارد و بهترین مدل هم به خوبی به آن برازش نمی‌شود. یکی از روش‌های جایگزین برای مدل‌سازی تغییرنگار، روش SVR است. این روش نیاز به انتخاب مدل اولیه برای تغییرنگار نداشته و می‌تواند مستقیما تغییرنگار بهینه را تولید کند. در بخش بعد به معرفی این روش پرداخته شده است.

۳ روش SVR

یکی از تکنیک‌های پر کاربرد در زمینه هوش مصنوعی، ماشین بردار پشتیبان (SVM) است. SVM یکی از روشهای یادگیری با نظارت است که در ابتدا برای مسائل طبقه‌بندی طراحی شد و سپس با استفاده از تابع زیان $\epsilon - \text{insensitive}$ برای مسائل رگرسیونی توسعه داده شد. این روش در مسائل رگرسیونی به SVR شناخته می‌شود. SVR بر مبنای تئوری یادگیری آماری استوار است که از اصل کمینه‌سازي ریسک ساختاري استفاده می‌کند و در نهایت به یک جواب بهینه کلی منجر می‌شود. کمینه‌سازي ریسک ساختاری، نسبت به روش متداولِ کمینه‌سازي ریسک تجربی (مورد استفاده در الگوریتم‌هایی مانند شبکه‌هاي عصبی مصنوعی و روش‌هاي کلاسیک آماري)، برتري داشته و برخلاف روش‌هایی همانند شبکه عصبی مصنوعی، به جوابهاي موضعی همگرا نمی‌شود.

یک مدل SVR در حالت کلی شامل متغیرهای ورودی، بردارهای پشتیبان، توابع کرنل و متغیر خروجی می‌باشد. فرض کنید $\{(\boldsymbol{x}_i, y_i); i = ۱, \ldots, n\}$ مجموعه n نمونه باشد بطوریکه $\boldsymbol{x}_i \in \mathbb{R}^p$ بردار ورودی و y_i خروجی متناظر باشد. به منظور یافتن یک ابرصفحه جداکننده، SVR فضای ورودی اولیه را به یک فضای دیگری با ابعاد بالاتر بنام فضای خصوصیت از طریق نگاشت غیرخطی ϕ انتقال می‌دهد؛ بطوریکه در فضای نگاشت یافته می‌توان از یک مدل شبیه رگرسیون خطی بصورت $f(x, w) = w.\phi(x) + b$ استفاده کرد. به عبارت دیگر، متغیرهای ورودی اولیه از یک فضای با ابعاد کوچک‌تر به فضایی با ابعاد بزرگتر تصویر می‌شوند تا یک مسئله رگرسیونی غیر خطی را به یک مسئله رگرسیونی خطی تبدیل کنند. در این رابطه w بردار وزن، b بایاس و $w.\phi(x)$ ضرب داخلی دو بردار است. برای برآورد مقادیر w و b بایستی تابع ریسک $R = \frac{۱}{۲}\|w\|^۲ + \frac{C}{n}\sum_{i=۱}^{n} L_\epsilon(y_i, f(x_i))$ کمینه شود بطوریکه عبارت $\frac{۱}{۲}\|w\|^۲$ میزان پیچیدگی مدل را کنترل

می‌کند و $\frac{C}{n}\sum_{i=1}^{n} L_\epsilon\left(y_i, f(x_i)\right)$ ریسک تجربی است و توسط تابع زیان $\epsilon-$ insensitive یعنی

$$L_\epsilon\left(y, f(x)\right) = \begin{cases} 0 & |y - f(x)| < \epsilon \\ |y - f(x)| & \text{otherwise} \end{cases}$$

اندازه‌گیری می‌شود. بر اساس این تابع زیان اگر اختلاف بین مقدار پیشگویی شده و مقدار واقعی کمتر از ϵ باشد خطای پیشگویی نادیده گرفته می‌شود. برای رفع مشکل انجام عملیات در فضای با ابعاد بالا نیاز به استفاده از یک تابع کرنل است. با استفاده از تابع کرنل، مشکل چند بعدی و غیرخطی بودن نگاشت حل می‌شود. عملکرد SVR به میزان زیادی به انتخاب تابع کرنل وابسته است. از توابع کرنل شناخته شده در SVR می‌توان کرنل‌های خطی، چندجمله‌ای، گاوسین و سیگموئید را نام برد. پرکاربردترین تابع کرنل تابع گاوسی است که به آن کرنل شعاعی مبنا [۲] (RBF) نیز گفته می‌شود و بصورت $K(x,y) = e^{-\sigma\|x-y\|^2}$ است؛ بطوریکه σ ثابت مثبت بوده و در طول فرآیند آموزش تعیین می‌شود. مقادیر بهینه برای پارامترهای σ و C با استفاده از روش اعتبارسنجی متقاطع و با جستجوی شبکه‌ای در فضای پارامترها پیدا می‌شود. در مدل‌سازی تغییرنگار با روش SVR ، همانند روش حداقل مربعات از زوج نقاط $(h, \hat{\gamma}(h))$ استفاده می‌شود. به عبارت دیگر، برآورد نقطه به نقطه یا لگ‌بندی که در آن تغییرنگار تجربی مستقیما از روی مقادیر مشاهده شده و در لگ‌های مشخص محاسبه می‌شود برای هر دو روش زمین آماری و روش SVR مشترک است. تفاوت روش SVR با روش زمین‌آماری در این است که روش SVR برای برازش یک منحنی هموار به این مقادیر گسسته (جهت بدست آوردن تابعی پیوسته در تمامی لگ‌ها) نیازی به انتخاب مدل تغییرنگار معتبر ندارد و قادر است بر اساس مقادیر مشاهده شده، مدل بهینه را پیدا کند.

۴ مقایسه روش‌ها

در این بخش به مقایسه عملکرد روش SVR با روش حداقل مربعات در برآورد تغییرنگار پرداخته شده است. با توجه به اینکه حجم نمونه بر دقت برآورد تغییرنگار تاثیرگذار است از سه مجموعه داده فضایی شبیه‌سازی شده با حجم نمونه‌های ۴۰۰، ۹۰۰ و ۱۶۰۰ استفاده گردید. شبیه‌سازی داده‌های فضایی در محیط R به کمک تابع grf از بسته $geoR$ انجام گرفت، بطوریکه برای مدل تغییرنگار نظری از مدل کروی با مقادیر $c_0 = ۰/۰۱$، $c = ۴$ و $a = ۶$ برای پارامترها استفاده گردید. همچنین دامنه مقادیر x و y بعنوان موقعیت جغرافیایی مشاهدات، بین صفر تا ۳۰ انتخاب گردید. مقایسه روش‌ها بر اساس دقت پیشگویی کریگیدن انجام گرفت. بدین‌صورت که در هر یک از این سه مجموعه داده، ۳۰ درصد بصورت تصادفی بعنوان مجموعه داده تست (به منظور پیشگویی) کنار گذاشته شد و از بقیه مجموعه نقاط برای برآورد پارامترهای مدل‌ها استفاده شد. سپس با استفاده از مدل برآورد شده برای تغییرنگار، عملیات پیشگویی برای نقاط کنار گذاشته شده انجام گرفته و میزان خطای پیشگویی محاسبه گردید. به منظور استفاده از روشهای مذکور در برآورد تغییرنگار و استفاده از آنها در پیشگویی کریگیدن، مراحل زیر برای هر مجموعه داده اجرا گردید:

۱) محاسبه تغییرنگار تجربی $\hat{\gamma}$ بر اساس رابطه (۱.۲) با استفاده از داده‌های مشاهده شده (۷۰ درصد از داده‌ها).

۲) مدل بندی تغییرنگار با دو روش حداقل مربعات (با استفاده از مدل‌های اولیه جدول ۱) و روشSVR

۳) محاسبه وزن‌های کریگیدن برای هر موقعیت s_i در مجموعه داده تست و سپس محاسبه مقدار پیشگویی در آن موقعیت.

۴) محاسبه خطای پیشگویی برای هر روش (جدول ۲)

[2] Radial Basis Function (RBF)

برای مدل‌سازی تغییرنگار بر اساس روش SVR ، از پکیج $e۱۰۷۱$ در نرم‌افزار R استفاده گردید بطوریکه «فاصله لگ» بعنوان ورودی و مقادیر تغییرنگار تجربی بعنوان خروجی مدل در نظر گرفته شد. همچنین از تابع کرنل شعاعی مبنا (RBF) استفاده گردید که نتایج بهتری نسبت به توابع کرنل خطی و چندجمله‌ای در بر داشت. انتخاب مقادیر بهینه برای پارامترهای C ، ϵ و σ بر اساس روش جستجوی شبکه‌ای انجام گرفت. تغییرنگار تجربی به همراه مدل‌های برازش یافته بر اساس دو روش مورد بررسی در شکل ۱ برای هر سه مجموعه داده شبیه‌سازی شده، نشان داده شده است. مشاهده می‌شود که برای هر سه مجموعه داده، روش SVR توانسته است بدون در نظر گرفتن مدل اولیه (در مقایسه با روش حداقل مربعات)، ارتباط غیرخطی بین فاصله لگ‌ها (متغیر ورودی) و تغییرنگار تجربی (متغیر خروجی) را پیدا کند.

خطای پیشگویی بر اساس دو شاخص RMSE و $r^۲$ محاسبه گردید. بزرگ بودن مقدار $r^۲$ و کوچک بودن مقدار RMSE برای هر روش، نشان می‌دهند که دقت پیشگویی بالاتر و در نتیجه عملکرد روش مورد نظر در مدل‌بندی تغییرنگار بهتر بوده است. این معیارها بصورت زیر هستند:

$$RMSE = \sqrt{\frac{۱}{n}\sum_{i=۱}^{n}\left(Z_i - \hat{Z}_i\right)^۲}, \qquad r^۲ = \left(\frac{\sum_{i=۱}^{n}\left(Z_i - \bar{Z}\right)\left(\hat{Z}_i - \tilde{Z}\right)}{\sqrt{\sum_{i=۱}^{n}\left(Z_i - \bar{Z}\right)^۲ \sum_{i=۱}^{n}\left(\hat{Z}_i - \tilde{Z}\right)^۲}}\right)^۲$$

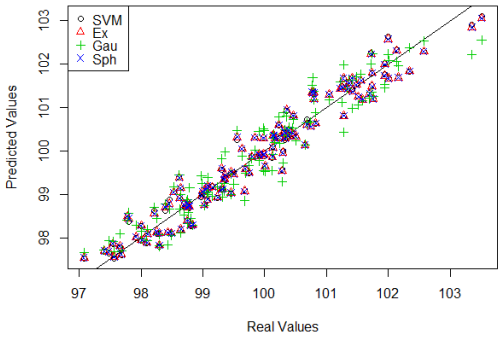
در این روابط Z_i مقادیر مشاهده شده، $\hat{Z}_i$ مقادیر پیشگویی شده، $\bar{Z}$ میانگین مقادیر مشاهده شده و $\tilde{Z}$ میانگین مقادیر پیشگویی شده است. نتایج بدست آمده برای این دو معیار، برای هر یک از روش‌های مورد بررسی در جدول ۲ گزارش شده است. با توجه به اینکه تولید داده‌های فضایی با استفاده از مدل کروی انجام گرفته است انتظار می‌رود که نتایج حاصل از روش حداقل مربعات بر اساس مدل اولیه کروی برای تغییرنگار، بهتر از بقیه باشد. با توجه به جدول ۲، برای حجم نمونه ۹۰۰ تایی، نتایج مدل کروی تا حدودی بهتر از دو مدل دیگر است. اما برای اولین مجموعه داده (با تعداد ۴۰۰ نمونه)، عملکرد روش SVR و برای سومین مجموعه داده، نتایج مدل نمایی بهتر بوده است.

جدول ۲: مقایسه روش‌های متفاوت در مدل‌بندی تغییرنگار بر اساس دو معیار $RMSE$ و $r^۲$

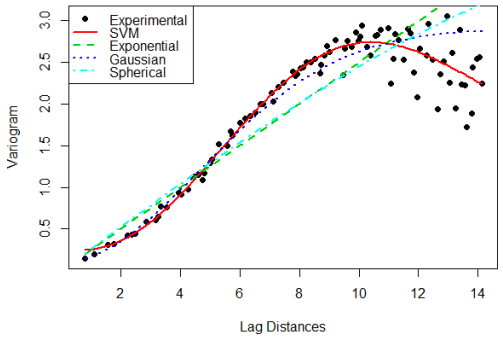
	$N = ۴۰۰$		$N = ۹۰۰$		$N = ۱۶۰۰$	
مدل‌ها	RMSE	$r^۲$	RMSE	$r^۲$	RMSE	$r^۲$
SVR	۰/۳۲۵	۰/۹۴۶	۰/۲۸۷	۰/۸۵۶	۰/۲۴۷	۰/۷۷۶
مدل نمایی	۰/۳۲۷	۰/۹۴۵	۰/۲۸۴	۰/۸۵۹	۰/۲۴۲	۰/۷۸۵
مدل گاوسی	۰/۴۲۰	۰/۹۱۰	۰/۳۳۱	۰/۸۱۰	۰/۲۵۹	۰/۷۵۳
مدل کروی	۰/۳۲۸	۰/۹۴۵	۰/۲۸۲	۰/۸۶۱	۰/۲۴۵	۰/۷۸۱

بحث و نتیجه‌گیری

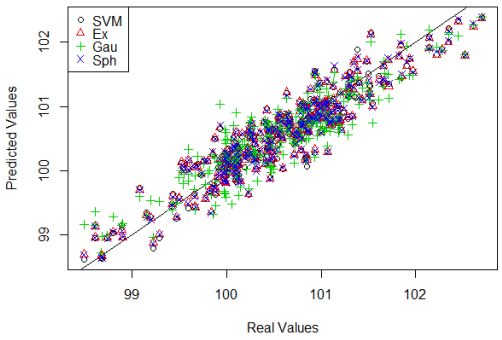
در این مطالعه، به کمک مجموعه داده‌های فضایی شبیه‌سازی شده، عملکرد روش SVR در برآورد تغییرنگار با روش حداقل مربعات بر اساس سه مدل اولیه نمایی، گاوسی و کروی مقایسه گردید. نتایج نشان داد که روش SVR قادر است بدون در نظر گرفتن مدل اولیه، به تغییرنگار تجربی بصورت دقیق برازش یافته و بر دقت نتایج پیشگویی کریگیدن تاثیرگذار باشد. بنابراین روش SVR می‌تواند رقیب مناسبی برای روش‌های کلاسیک در برآورد تغییرنگار باشد. در این مطالعه فقط از سه مجموعه داده شبیه‌سازی شده با حجم نمونه‌های متفاوت استفاده گردید. بررسی حالت‌های دیگر از جمله وجود یا عدم وجود



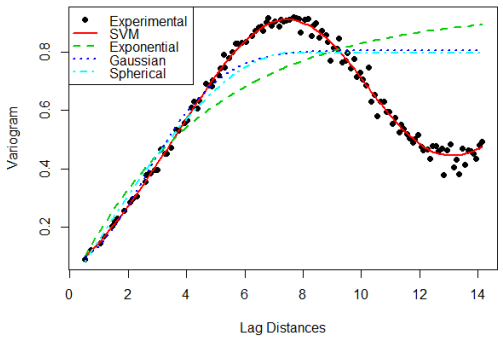
الف- نمودار پراکنش مقادیر واقعی در برابر پیشگویی ($N = ۴۰۰$)



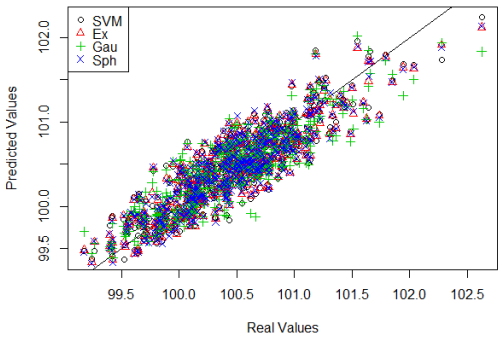
ب- تغییرنگار تجربی به همراه مدل‌های برازش شده



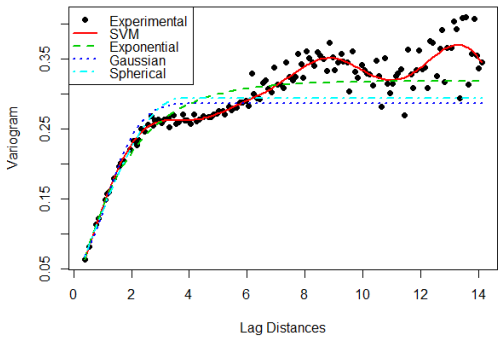
ج- نمودار پراکنش مقادیر واقعی در برابر پیشگویی ($N = ۹۰۰$)



د- تغییرنگار تجربی به همراه مدل‌های برازش شده



ه- نمودار پراکنش مقادیر واقعی در برابر پیشگویی ($N = ۱۶۰۰$)



و- تغییرنگار تجربی به همراه مدل‌های برازش شده

شکل ۱: نمودار تراکنش مقادیر واقعی در برابر پیشگویی (راست)، تغییرنگار تجربی به همراه مدل‌های برازش شده (چپ)

اثر قطعه‌ای، میزان وابستگی فضایی میان داده‌ها و یا همسانگردی یا ناهمسانگردی میدان تصادفی می‌تواند در مطالعات آینده مورد بررسی قرار گیرد.

مراجع

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

Cressie, N. (1985), Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5), 563-586.

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

Desassis, N., Renard, D. (2013), Automatic variogram modeling by iterative least squares: univariate and multivariate cases. *Mathematical geosciences*, 45(4), 453-470.

Han, C., Wang, J., Zheng, M., Wang, E., Xia, J., Li, G., Choe, S. (2016). New variogram modeling method using MGGP and SVR. *Earth Science Informatics*, 9(2), 197-213.

Huang, Z., Wang, H., Zhang, R. (2012), An improved kriging interpolation technique based on SVM and its recovery experiment in oceanic missing data. *American Journal of Computational Mathematics*, 2(01), 56.

Matheron, G. (1962), Traité de géostatistique appliquée. (Vol. 1). Editions Technip.

Matheron, G. (1963), Principles of geostatistics. *Economic geology*, 58(8), 1246-1266.

Miranda, H., de Miranda, M. S. (2011). Combining Robustness with Efficiency in the Estimation of the Variogram. *Mathematical Geosciences*, 43(3), 363-377.

Oliver, M. A., Webster, R. (2015), *Basic steps in geostatistics: the variogram and kriging* (Vol. 106). New York: Springer.

Webster, R., Oliver, M. A. (2007) *Geostatistics for environmental scientists*. John Wiley and Sons.

Youkuo, C., Yongguo, Y., Wangwen, W. (2015). Coal seam thickness prediction based on least squares support vector machines and kriging method. *Electron. J. Geotech. Eng*, 20, 167-176.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

تحلیل بیزی تقریبی مدل‌های رگرسیونی فضایی

فاطمه حسینی[1]، امید کریمی
گروه آمار، دانشگاه سمنان

چکیده: در مطالعات اقتصادی اغلب داده‌ها دارای وابستگی فضایی هستند و برای مدل‌بندی این نوع از داده‌ها معمولا از سه مدل تاخیر فضایی، دوربین فضایی و خطای فضایی به عنوان سه مدل معروف رگرسیون فضایی استفاده می‌شود. هدف اصلی در مطالعه این مدل‌ها به‌دست آوردن برآورد پارامترها و سپس پیشگویی در موقعیت‌های جدید است. برای این منظور، در این مقاله ابتدا رهیافت ماکسیمم درستنمایی بررسی شده است و در راستای بالابردن دقت برآورد پارامترها و کاهش زمان محاسبات، رهیافت بیزی تقریبی برای این سه مدل رگرسیونی فضایی بیان می‌شود. در نهایت در یک مثال واقعی به مقایسه عملکرد سه مدل و دو رهیافت پرداخته می‌شود.

واژه‌های کلیدی: مدل تاخیر فضایی، مدل دوربین فضایی، مدل خطای فضایی، رهیافت بیزی تقریبی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 91G70، 62F15، 91B72.

۱ مقدمه

معمولا برای تحلیل داده‌های اقتصادی که وابستگی مکانی دارند از مدل‌های رگرسیون فضایی استفاده می‌شود. سه مدل معروف رگرسیون فضایی که اغلب برای این نوع داده‌ها در اقتصادسنجی فضایی مورد استفاده قرار می‌گیرد. مدل‌های اتورگرسیو فضایی یا مدل تاخیر فضایی[1] (SLM)، مدل دوربین فضایی[2] (SDM) و مدل خطای فضایی[3] (SEM) هستند. در مدل خطای فضایی همبستگی فضایی در جمله خطا منظور می‌شود، در مدل اتورگرسیو فضایی اثر فضایی در متغیر پاسخ منظور می‌شود و در مدل دوربین فضایی اثر فضایی هم از طریق متغیر پاسخ و هم از طریق متغیرهای کمکی در مدل وارد می‌شود. اولین بار انسلین (۱۹۸۸) به معرفی اقتصادسنجی فضایی پرداخت و داویدسون و مکینون (۱۹۹۳)، لسج

[1] Spatial Autoregressive Model or Spatial Lag Model
[2] Spatial Durbin Model
[3] Spatial Error Model

[1] نام و ایمیل ارائه دهنده مقاله: فاطمه حسینی، fatemeh.hoseini@semnan.ac.ir

(۱۹۹۹)، لسج و پس (۲۰۰۹) و لسج و فیشر (۲۰۰۸) به طور مفصل‌تر روش‌های اقتصادسنجی فضایی را مورد مطالعه قرار دادند. انسلین (۲۰۱۰) برای داده‌ها در طول زمان طولانی و با به‌کار بردن مدل‌های رگرسیونی فضایی به بررسی و مطالعه اقتصادسنجی فضایی پرداخت. بایوند و همکاران (۲۰۱۴) به معرفی تکنیک‌های تحلیل داده‌های فضایی و کاربرد این تکنیک‌ها در مطالعات اقتصادسنجی با نرم‌افزار R پرداختند. اربیا (۲۰۱۶) و کلیجیان و پیراس (۲۰۱۷) انواع روش‌های اقتصادسنجی را به اقتصادسنجی فضایی تعمیم داده‌اند. برای تحلیل این مدل‌ها در اکثر مطالعات اشاره شده معمولا انواع روش‌های برآورد ماکسیمم درستنمایی پیشنهاد شده است. به دلیل اضافه کردن همبستگی فضایی به این مدل‌ها و اضافه شدن تعداد پارامترهای مدل برای استفاده از رهیافت ماکسیمم درستنمایی تعدادی از پارامترها معلوم فرض می‌شوند و سایر پارامترها را برآورد و مجددا با استفاده از برآوردهای به‌دست آمده پارامترهایی که معلوم فرض شدند را برآورد می‌کنند. این یکی از مشکلات اساسی در استفاده از رهیافت ماکسیم درستنمایی در این مدل‌ها است. در رهیافت بیزی برای تمام پارامترها توزیع پیشین فرض می‌شود و با استفاده از تابع درستنمایی تشکیل شده و توزیع‌های پیشین فرض شده، توزیع پسین تشکیل و با استفاده از الگوریتم‌های موجود مثل الگوریتم مونت کارلوی زنجیر مارکوفی به‌طور همزمان پارامترها برآورد می‌شوند. از طرفی اجرای الگوریتم‌های مونت کارلویی زمان‌بر است و گاهی رسیدن به همگرایی در این الگوریتم‌ها نیاز به تکرار زیاد دارد و به کندی انجام می‌شود. در این مقاله از یک رهیافت بیزی تقریبی برای تحلیل مدل‌ها استفاده می‌شود و در یک مثال روش‌ها مورد مقایسه قرار می‌گیرند. ساختار مقاله به این‌صورت است که در بخش دوم سه مدل رگرسیون فضایی معرفی می‌شوند. در بخش سه تحلیل درستنمایی و در بخش چهار تحلیل بیزی تقریبی سه مدل بیان و در بخش چهارم و پنجم برای مقایسه بین دو رهیافت درستنمایی و بیزی تقریبی به تحلیل یک مجموعه داده واقعی مربوط به قیمت مسکن در شهر تهران پرداخته می‌شود. در نهایت بحث و نتیجه‌گیری ارائه شده است.

۲ مدل‌های رگرسیونی فضایی

فرض کنید $\boldsymbol{y}' = (y_1, ..., y_n)$ بردار مشاهدات در n موقعیت فضایی $s_1, ..., s_n$ است و X ماتریس طرح با p متغیر کمکی و $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ پارامترهای رگرسیونی هستند، مدل تاخیر فضایی که در لسج و پس (۲۰۰۹) ارائه شده است، متغیر پاسخ به مجموع وزنی پاسخ‌ها در همسایگی‌هایش، متغیرهای کمکی و جمله خطا بستگی دارد و به‌صورت

$$\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{e}', \quad \boldsymbol{e}' \sim MVN(0, \sigma^2 I_n), \tag{۱.۲}$$

است، که در آن I_n یک ماتریس واحد از بعد $n \times n$ و W یک ماتریس مجاورت سطری استاندارد شده و ρ پارامتر خودهمبستگی فضایی است. مدل خطای فضایی به‌صورت

$$\begin{aligned} &\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim MVN(0, \sigma^2 (I_n - \lambda W)^{-1} (I_n - \lambda W)'^{-1}), \\ &\boldsymbol{u} = \lambda W\boldsymbol{u} + \boldsymbol{e}, \quad \boldsymbol{e} \sim MVN(0, \sigma^2 I_n), \\ &\boldsymbol{y} = X\boldsymbol{\beta} + (I_n - \lambda W)^{-1}\boldsymbol{e}, \end{aligned} \tag{۲.۲}$$

بیان می‌شود، که در آن λ پارامتر خودهمبستگی فضایی است. مدل دوربین فضایی به‌صورت

$$\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{e}', \quad \boldsymbol{e}' \sim MVN(0, \sigma^2 I_n), \tag{۳.۲}$$

تعریف می‌شود، که در آن WX متغیرهای کمکی تاخیر فضایی و γ بردار ضرایب این متغیرها هستند. در این مدل پاسخ علاوه بر این که به پاسخ‌ها در همسایگی‌اش و متغیرهای کمکی بستگی دارد به مقادیر متغیرهای کمکی همسایگی‌اش وابسته

است. در تمام مدل‌های معرفی شده، جمله خطا گاوسی با میانگین صفر و ماتریس واریانس کوواریانس به شکل اتو رگرسیو فضایی[۴] (SAR) در نظر گرفته شده است. مدل‌های SLM و SDM ساختار پیچیده‌تری در قسمت خطی مدل که مربوط به متغیرهای کمکی می‌باشد، دارند. همچنین پارامترهای خودهمبستگی فضایی ρ در بازه $(\frac{۱}{\lambda_{min}}, \frac{۱}{\lambda_{maX}})$ محدود می‌شود که در آن λ_{min} و λ_{maX}، کوچک‌ترین و بزرگ‌ترین مقدار ویژه W هستند. در این سه مدل ارتباط فضایی متغیرها به‌صورت دو به دو و به فرم عددی توسط ماتریس W مشخص می‌شودکه معمولا و نه لزوما همیشه ماتریسی متقارن در نظر گرفته می‌شود و در آن عنصر W_{ij} ارتباط فضایی مربوط به متغیر پاسخ در موقعیتiام با متغیر پاسخ در موقعیت jام است. به عنوان مثال در برخی مطالعات مربوط به قیمت مسکن فاصله مرکز منطقه iام از منطقه jام را با d_{ij} نشان می‌دهند و W_{ij} را از رابطه $\frac{۱}{d_{ij}}$ به‌دست می‌آورند به‌طوری که با افزایش فاصله دو منطقه اثر فضایی کم می‌شود. لسج و پس (۲۰۰۹) نشان دادند که می‌توان W را به‌صورت یک ماتریس سطری تصادفی ساخت، به طوری که بردار تاخیر فضایی $W\boldsymbol{y}$ شامل مقادیر ساخته شده از متوسط مشاهدات همسایه می‌باشد. منظور از ماتریس سطری تصادفی، ماتریس نامنفی است که عناصر روی هر سطر آن نرمال‌سازی شده و بنابراین مجموع هر سطر یک می‌باشد و در این‌صورت $\lambda_{max} = ۱$ و $\rho < ۱$ است.

۳ تحلیل درستنمایی مدل‌ها

برای به دست آوردن برآورد ماکسیمم درستنمایی پارامترهای مدل تاخیر فضایی، مدل (۱.۲) را می‌توان با فرض معلوم بودن ρ به‌صورت

$$\boldsymbol{y} - \rho W\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(۰, \sigma^۲ I_n),$$

نوشت. بنابراین $\boldsymbol{\epsilon} = (I - \rho W)\boldsymbol{y} - X\boldsymbol{\beta}$ و داریم

$$L(\sigma^۲, \boldsymbol{\epsilon}) = (\frac{۱}{۲\pi\sigma^۲})^{\frac{n}{۲}} \exp\{-\frac{۱}{۲\sigma^۲}\boldsymbol{\epsilon}'\boldsymbol{\epsilon}\},$$

پس تابع درستنمایی مشاهدات به‌صورت

$$L(\rho, \boldsymbol{\beta}, \sigma^۲|\boldsymbol{y}) = (\frac{۱}{۲\pi\sigma^۲})^{\frac{n}{۲}}|J| \times \exp\{-\frac{۱}{۲\sigma^۲}[(I - \rho W)\boldsymbol{y} - X\boldsymbol{\beta}]'[(I - \rho W)\boldsymbol{y} - X\boldsymbol{\beta}]\},$$

است، که در آن $|J| = |\frac{\partial \boldsymbol{\epsilon}}{\partial \boldsymbol{y}}| = |I_n - \rho W|$. با مشتق‌گیری از لگاریتم تابع درستنمایی نسبت به $\boldsymbol{\beta}$ و $\sigma^۲$ برآورد ماکسیمم درستنمایی آن‌ها به‌صورت

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (X'X)^{-۱}X'(I_n - \rho W)\boldsymbol{y}, \\
\hat{\sigma^۲} &= \frac{۱}{n}[(I_n - \rho W)\boldsymbol{y} - X\boldsymbol{\beta}]'[(I_n - \rho W)\boldsymbol{y} - X\boldsymbol{\beta}] \\
&= \frac{۱}{n}(\boldsymbol{y} - X\boldsymbol{\beta})'(I_n - \rho W)'(I_n - \rho W)(\boldsymbol{y} - X\boldsymbol{\beta}),
\end{aligned}$$

به دست می‌آیند. با جایگذاری $\hat{\boldsymbol{\beta}}$ در $\hat{\sigma}^۲$ و با تعریف

$$A = [\boldsymbol{y} - X(X'X)^{-۱}X'\boldsymbol{y} - \rho(W\boldsymbol{y} - (X'X)^{-۱}X'W\boldsymbol{y})],$$

داریم $\hat{\sigma^۲} = \frac{۱}{n}A'A$ با قرار دادن $\boldsymbol{e}_۰ = \boldsymbol{y} - X(X'X)^{-۱}X'\boldsymbol{y}$ یعنی باقیمانده رگرسیون معمولی y روی X و $\boldsymbol{e} =$ $W\boldsymbol{y} - (X'X)^{-۱}X'W\boldsymbol{y}$ باقیمانده رگرسیون معمولی $W\boldsymbol{y}$ روی X داریم

$$\hat{\sigma}^۲ = \frac{۱}{n}(\boldsymbol{e}_۰ - \rho\boldsymbol{e}_ℓ)'(\boldsymbol{e}_۰ - \rho\boldsymbol{e}_۱). \qquad (۱.۳)$$

[4] Spatial Autoregressive Model

با جایگذاری $\hat{\boldsymbol{\beta}}$ و رابطه (۱.۳) در تابع درستنمایی که اکنون تابع درستنمایی متمرکز شده [۵] نامیده می‌شود به‌صورت

$$\ell_c(\rho) = -\frac{n}{۲}\ln ۲\pi - \frac{n}{۲}\ln\frac{۱}{n}(\boldsymbol{e}_۰ - \rho\boldsymbol{e}_۱)'(\boldsymbol{e}_۰ - \rho\boldsymbol{e}_۱) + \sum_i \ln(۱ - \rho\omega_i), \qquad (۲.۳)$$

خواهد بود. پس

$$\frac{\partial \ell_c(\rho)}{\partial \rho} = \frac{n(\boldsymbol{e}'_۰\boldsymbol{e}_۱ - \rho\boldsymbol{e}'_۱\boldsymbol{e}_۱)}{\boldsymbol{e}'_۰\boldsymbol{e}_۰ - ۲\rho\boldsymbol{e}'_۰\boldsymbol{e}_۱ + \rho^۲\boldsymbol{e}'_۱\boldsymbol{e}_۱} - \frac{\sum_i \omega_i}{۱ - \rho\omega_i} = ۰,$$

به‌طوری که برآورد ρ باید در بازه‌ی $(\lambda_{min}^{-۱}, \lambda_{max}^{-۱})$ باشد. برای این منظور باری و پس (۱۹۹۷) پیشنهاد کردند که یک بردار $q \times ۱$ از مقادیری که در بازه‌ی $[\rho_{min}, \rho_{max}]$ است برای ρ در نظرگرفته شود و آنگاه

$$\begin{pmatrix} \ell(\rho_۱) \\ \vdots \\ \ell(\rho_q) \end{pmatrix} \propto \begin{pmatrix} \sum_i \ln(۱ - \rho_۱\omega_i) \\ \vdots \\ \sum_i \ln(۱ - \rho_q\omega_i) \end{pmatrix} - \frac{n}{۲}\begin{pmatrix} \ln \boldsymbol{e}'(\rho_۱)\boldsymbol{e}(\rho_۱) \\ \vdots \\ \ln \boldsymbol{e}'(\rho_q)\boldsymbol{e}(\rho_q) \end{pmatrix},$$

که در آن $\boldsymbol{e}(\rho) = \boldsymbol{e}_۰ - \rho\boldsymbol{e}_۱$ و $\rho_۱ \ldots \rho_q$ مقادیری هستند که بهینه‌سازی حول این مقادیر انجام می‌شود .با به‌کار بردن یک الگوریتم تکرار شونده و با در نظر گرفتن دقت مورد نظر یک شبکه به اندازه‌ی کافی خوب از q مقدار برای ρ، $l(\rho)$ مشخص و مقدار بهینه تعیین خواهد شد. پس از تعیین برآورد ماکسیمم درستنمایی ρ به‌صورت $\hat{\rho}$ آنگاه می‌توان $\hat{\boldsymbol{\beta}}$ و $\hat{\sigma}^۲$ را به دست آورد. برآورد درستنمایی پارامترهای دو مدل خطای فضایی و دوربین فضایی را می‌توان به‌طور مشابه به‌دست آورد.

۴ تحلیل بیزی معمولی و تقریبی مدل‌های رگرسیون فضایی

فرض کنید مدل تاخیر فضایی به صورت

$$\boldsymbol{y} = \rho_{Lag}W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{e}, \qquad (۱.۴)$$

است، به طوری‌که $\boldsymbol{e} \sim MVN(۰, \sigma^۲ I_n)$. برای $\boldsymbol{\beta}$ یک توزیع پیشین ناآگاهی بخش از نوع مبهم، یعنی یک توزیع نرمال با واریانس بزرگ و برای $\sigma^۲$ توزیع پیشین گامای وارون فرض می‌شوند. بنابراین

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^۲) &= \pi(\boldsymbol{\beta}|\sigma^۲)\pi(\sigma^۲) \\ &= N(c, \sigma^۲T)IG(a, b) \\ &= \frac{b^a}{(۲\pi)^{\frac{k}{۲}}|T|^{\frac{۱}{۲}}\Gamma(a)}(\sigma^۲)^{-(a+\frac{k}{۲}+۱)} \\ &\times \exp[\frac{-\{B-c)'T^{-۱}(B-c) + ۲b\}}{۲\sigma^۲}], \\ \pi(\sigma^۲) &= \frac{b^a}{\Gamma(a)}(\sigma^۲)^{-(a+۱)}\exp(-\frac{b}{\sigma^۲}), \quad \sigma^۲ > ۰, a, b > ۰ \end{aligned}$$

و برای پارامتر ρ_{Lag} توزیع یکنواخت به صورت $\pi(\rho_{Lag}) \sim U(\lambda_{min}^{-۱}, \lambda_{max}^{-۱})$ که در آن λ_{min} و λ_{max} کوچک‌ترین و بزرگ‌ترین مقدار ویژه W هستند.

فرض کنید تمام مشاهدات به صورت $D = \{\boldsymbol{y}, X, W\}$ و پارامترهای مدل به‌صورت $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^۲, \rho_{Lag}\}$ در نظر گرفته شوند. تابع درستنمایی برای این مدل به صورت $P(D|\boldsymbol{\beta}, \sigma^۲, \rho) = (۲\pi\sigma^۲)^{-\frac{n}{۲}}|A|\exp\{-\frac{۱}{۲\sigma^۲}(A\boldsymbol{y} - X\boldsymbol{\beta})'(A\boldsymbol{y} -$

[۵] Function Likelihood Concentrated

$\{X\boldsymbol{\beta})$ است، که در آن $A = (I_n - \rho W)$. بنابراین توزیع پسین به‌صورت

$$\pi(\boldsymbol{\beta}, \sigma^2, \rho_{Lag}|D) \quad \propto \quad \pi(D|\boldsymbol{\beta}, \sigma^2, \rho_{Lag})\pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2)\pi(\rho_{Lag}), \tag{۲.۴}$$

است، که شکل پیچیده‌ای دارد و برای به دست آوردن برآوردهای بیزی پارامترها از الگوریتم‌های $MCMC$ استفاده می‌شود که اجرای الگوریتم‌ها بسیار زمان‌بر و گاهی اوقات برای رسیدن به همگرایی نیاز به صرف زمان بسیار است. به عنوان یک پیشنهاد می‌توان از رهیافت بیزی تقریبی استفاده کرد. رو و همکاران (۲۰۰۹)، رو و هلد (۲۰۰۵) و ایدسویک و همکاران (۲۰۱۲) با فرض نرمال بودن متغیرهای پنهان فضایی در مدل‌های شامل متغیرهای پنهان گاوسی روش بیز تقریبی را به جای روش بیز معمولی و الگوریتم‌های نمونه‌گیری مونت کارلویی معرفی کرده و نشان دادند محاسبات این روش سریع‌تر و نتایجی معادل روش‌های بیزی معمولی ارائه می‌کند. ایدسویک و همکاران (۲۰۱۲) با در نظر گرفتن توزیع پیشین نرمال برای متغیرهای پنهان در مدل‌های شامل متغیر پنهان طبق قضیه زیر نشان دادند که توزیع پسین تقریبی این متغیرها متعلق به خانواده توزیع نرمال است.

قضیه ۱.۴. *اگر در مدل، متغیرهای پنهان $\boldsymbol{x}$ دارای توزیع پیشین نرمال $\boldsymbol{x} \sim N(H\boldsymbol{\beta}, \Sigma_\theta)$ و متغیر پاسخ $\boldsymbol{y}$ متعلق به خانواده نمایی $\pi(y_i|x_i) = \exp\{y_i x_i - b(x_i) + c(y_i)\}$ باشد، با خطی کردن حول یک مقدار ثابت، توزیع را می‌توان با توزیع نرمال به‌صورت*

$$\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta}) \approx N_n(\hat{\boldsymbol{\mu}}_{x|y}(\boldsymbol{y}, \boldsymbol{x}^\circ), \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^\circ)),$$

تقریب زد که در آن

$$\hat{\boldsymbol{\mu}}_x|y(\boldsymbol{y}, \boldsymbol{x}^\star) = H\boldsymbol{\beta} + \Sigma_\theta A' R^{-1}(\boldsymbol{z}(\boldsymbol{y}, \boldsymbol{x}^{obs}) - AH\boldsymbol{\beta}),$$
$$\hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^\star) = \Sigma_\theta - \Sigma'_\theta R^{-1} A\Sigma_\theta, , \quad R = A\Sigma_\theta A' + P,$$

و P یک ماتریس قطری با عناصر $P(i,i) = \frac{1}{b''(x_i)} i = 1, ..., k$ به ازای $z_i(y_i, x_i^\cdot) = [y_i - b'(x_i^\cdot) + x_i b''(x_i^\cdot)]/b(x_i^\cdot)$ می‌باشند.

برهان: مراجعه شود به ایدسویک و همکاران (۲۰۱۲).

برای تعیین مقدار ثابت و خطی‌سازی قسمت درستنمایی در آن و برازش یک تقریب نرمال به پسین $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta})$ از الگوریتم زیر استفاده می‌شود.

الف‌ـ مقدار اولیه $\boldsymbol{x}^\star$ به عنوان مثال $\boldsymbol{x}^\star = M(\boldsymbol{x}|\boldsymbol{\eta})$، انتخاب و $m = 0$ قرار داده شود.

ب‌ـ از قضیه ۱.۴ توزیع پسین تقریبی در $\boldsymbol{x}^m$، یعنی $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta}) \equiv N_n(\hat{\mu}_{x|y,\eta}(\boldsymbol{x}^{(m)}), \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{obs}))$ تعیین می‌شود.

ج‌ـ قرار دهید $\boldsymbol{x}^{(m+1)} = \hat{M}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta})$

د‌ـ $m = m + 1$ اختیار شود و تا رسیدن به همگرایی، الگوریتم از مرحله (ب) مجددا شروع شود. پس از رسیدن به همگرایی، الگوریتم متوقف و توزیع پسین تقریبی تعیین می‌شود.

برای برآورد بیزی پارامترها توزیع پسین پارامترهای مدل به‌صورت

$$\pi(\boldsymbol{\eta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})}{\pi(\boldsymbol{y})\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta})} \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})}{\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta})},$$

است، که برای هر مقدار $\boldsymbol{x}$، قابل محاسبه می‌باشد. برای مثال می‌توان از مد توزیع پسین تقریبی $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})$ برای یک بردار ثابت از پارامترها، یعنی $\boldsymbol{x} = \hat{M}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})$ استفاده کرد. با به کار بردن تقریب نرمال برازش شده به توزیع شرطی کامل $\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})$، یک چگالی تقریبی برای توزیع پسین پارامترهای مدل به‌صورت

$$\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})}{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})}|_{x=\hat{M}(x|y,\eta)}, \tag{۳.۴}$$

حاصل می‌شود، که این تقریب توسط رو و همکاران (۲۰۰۹) یک تقریب لاپلاس از چگالی پسین پارامترهای مدل نامیده شد. برای تعیین پسین تقریبی (۳.۴) از الگوریتم زیر استفاده می‌شود:

الف‌ـ مد تابع $\log \hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$ با روش بهینه‌سازی شبه‌ـنیوتن یا نیوتن رافسون تعیین و با $\boldsymbol{\eta}^*$ نشان داده می‌شود.

ب‌ـ ماتریس هسیان از رابطه $H = (\frac{\partial^2 \log \hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})}{\partial\eta^i \partial\eta^j})$[۶] محاسبه و معکوس آن به صورت $H^{-1} = V\Lambda V'$ تجزیه شود، که در آن V ماتریس بردارهای ویژه و Λ ماتریس قطری مقادیر ویژه آن است.

ج‌ـ مبدا مختصات به مد $\boldsymbol{\eta}^*$ انتقال داده شود، فرمول مختصات در مبدا $\boldsymbol{\eta}^*$ به‌صورت $\eta(t) = \boldsymbol{\eta}^* + V\Lambda^{-1}\boldsymbol{t}$ تعریف می‌شود، که در آن $\boldsymbol{t}$ مقادیر استاندارد شده هستند. به‌عنوان مثال، برای حالت دو بعدی $\boldsymbol{t} = (t_1, t_2)$، با قرار دادن مبدا مختصات $\boldsymbol{t} = (0, 0)$ در رابطه بالا $\boldsymbol{\eta}(0) = \boldsymbol{\eta}^*$ حاصل می‌شود.

د‌ـ با شروع از مبدا مختصات جدید روی هر یک از محورها نقاطی به فاصله مقادیر صحیح δ_t به گونه‌ای اختیار شوند، که شرط

$$\log \hat{\pi}(\boldsymbol{\eta}(0)|\boldsymbol{y}) - \log \hat{\pi}(\boldsymbol{\eta}(\boldsymbol{t})|\boldsymbol{y}) < \delta_t, \tag{۴.۴}$$

برقرار باشد، سپس به‌طور مشابه نقاط درون صفحات نیز تعیین شوند. معمولا در نامساوی بالا $\delta_t = 2/5$ در نظر گرفته می‌شود. با به‌کار بردن این الگوریتم η_i هایی از توزیع $\pi(\boldsymbol{\eta}|\boldsymbol{y})$ تولید می‌شوند که می‌توان از آن‌ها برای تقریب توزیع $\pi(\boldsymbol{\eta}|\boldsymbol{y})$ استفاده کرد.

برای تعمیم رهیافت بیزی تقریبی به مدل‌های رگرسیونی فضایی در مطالعات اقتصادسنجی فضایی ابتدا مدل به صورت یک مدل آمیخته خطی تعمیم‌یافته شامل متغیرهای پنهان فضایی مدل‌بندی می‌شود. در راستای کاهش زمان محاسبات متغیرهای پنهان موجود در مدل با یک میدان تصادفی مارکوفی گاوسی مدل‌بندی می‌شوند. ماتریس واریانس‌ـ کوواریانس میدان تصادفی مارکوفی گاوسی به صورت یک ماتریس دقت[۷] که عکس ماتریسی واریانس‌ـ کوواریانس میدان تصادفی گاوسی است بیان می‌شود که یک ماتریس پراکنده[۸] است و همین باعث سادگی در محاسبات و کاهش زمان محاسبات می‌شود. فرض کنید $\boldsymbol{y'} = (y_1, ..., y_n)$ بردار متغیر پاسخ در n موقعیت $s_1, ..., s_n$ باشد، به طوری‌که $\boldsymbol{y}$ می‌تواند گاوسی یا غیرگاوسی باشد و متعلق به خانواده نمایی با میانگین μ_i در نظر گرفته می‌شود. اکنون می‌توان سه مدل رگرسیون فضایی تاخیر فضایی، خطای فضایی و دوربین فضایی را به صورت مدل کلی

$$\eta_i = g(\mu_i) = z_i'\boldsymbol{\beta} + u_i, \tag{۵.۴}$$

در نظر گرفت، که در آن پیشگویی کننده η_i شامل اثرات ثابت z_i و تصادفی u_i است، به‌طوری که این مدل می‌تواند شامل اثرات غیرخطی از متغیرهای پنهان و اثرات غیرخطی از متغیرهای کمکی نیز باشد. فرض کنید بردار $\boldsymbol{\theta}_1$ ابر پارامترهای

[6] Hessian Matrix
[7] Precision Matrix
[8] Sparse Matrix

مربوط به $\boldsymbol{y}$ و $\boldsymbol{\theta}_۲$ بردار ابر پارامترهای مربوط به بردار $\boldsymbol{x}$ است، به طوری‌که بردار $\boldsymbol{x}$ شامل اثر کلی پیشگویی‌کننده خطی برای هر مشاهده، پارامترهای رگرسیونی و اثرات پنهان فضایی به‌صورت $\boldsymbol{x'} = (\eta_i, \boldsymbol{\beta'}, \boldsymbol{u'})$ باشد. بردار $\boldsymbol{x}$ به صورت یک میدان تصادفی مارکوفی گاوسی(GMRF) با ماتریس دقت $Q(\theta_۲)$ فرض می‌شود. بنابراین تابع درستنمایی مدل با در نظر گرفتن $\boldsymbol{\theta} = (\boldsymbol{\theta}_۱, \boldsymbol{\theta}_۲)$ به شکل $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i \in l} \pi(y_i|x_i, \boldsymbol{\theta})$ که در آن $۱ \leqslant l \leqslant n$، بیانگر مشاهدات به غیر از گمشده‌ها است. توزیع توام $\boldsymbol{x}$ یعنی پارامترهای مدل، اثرات پنهان و ابر پارامترها یعنی بردار $\boldsymbol{\theta}$ به‌صورت

$$\begin{aligned} \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i \in l} \pi(y_i|x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})|Q(\boldsymbol{\theta})|^{\frac{n}{۲}} \exp\{-\frac{۱}{۲}\boldsymbol{x'}Q(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i \in l} \log(\pi(y_i|x_i, \boldsymbol{\theta}))\}, \end{aligned} \qquad (۶.۴)$$

توزیع‌های حاشیه‌ای برای اثرات پنهان $\pi(x_i|\boldsymbol{y}) \propto \int \pi(x_i|\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$ و برای ابر پارامتر از $\pi(\theta_j|\boldsymbol{y}) \propto \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}$ به‌دست می‌آیند، که در آن $\boldsymbol{\theta}_{-j}$ بردار ابر پارامترها بدون θ_j است. اکنون از رهیافت بیزی تقریبی دیدیم $\pi(x|\boldsymbol{y}, \boldsymbol{\theta})$ را می‌توان با توزیع نرمال $\hat{\pi}(x|\boldsymbol{y}, \boldsymbol{\theta})$ تقریب زد و از خواص توزیع نرمال $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ به دست آورد. $\pi(\boldsymbol{\theta}|, \boldsymbol{y})$ را می‌توان از الگوریتم توضیح داده شده در بالا تعیین نمود. بنابراین تقریبی از توزیع حاشیه‌ای برای x به‌صورت

$$\hat{\pi}(x_i|\boldsymbol{y}) = \sum_k \hat{\pi}(x_i|\boldsymbol{\theta}_k, \boldsymbol{y})\hat{\pi}(\boldsymbol{\theta_k}|\boldsymbol{y})\triangle_k,$$

محاسبه می‌شود، که در آن k تعداد نمونه‌های تولید شده از $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ از الگوریتم توضیح داده شده در بالا تعیین می‌شود و $\triangle_k$ وزن تخصیص داده شده به هر یک از عناصر بردار $\boldsymbol{\theta_k}$ است که معمولا یکسان و برابر ۱ در نظر گرفته می‌شود.

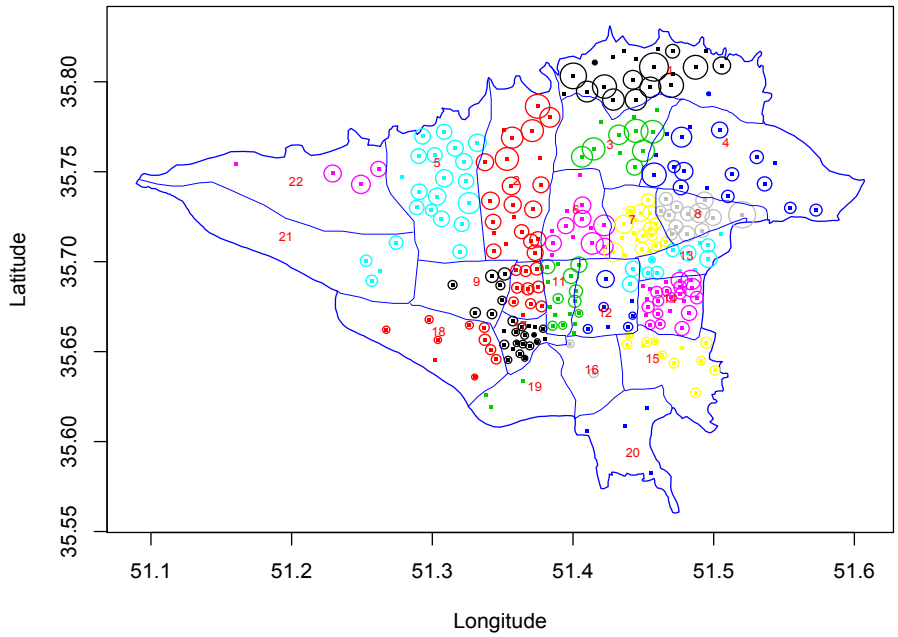
۵ داده‌های قیت مسکن شهر تهران

در این قسمت داده‌های مربوط به متوسط قیمت مسکن در مناطق مختلف تهران بررسی می‌شود. چون داده‌های قیمت مسکن دارای همبستگی از نوع فضایی (مکانی) است، با استفاده از سه مدل رگرسیون فضایی مدل‌بندی می‌شود. داده‌های قیمت مسکن برای شهر تهران به‌صورت روزانه در سامانه اطلاعات بازار املاک ایران ثبت می‌شود که این داده‌ها را می‌توان در سایت www.hmi.mrud.ir مشاهده کرد. قیمت مسکن نوساز (تا یک سال ساخت) برای محله‌های ۲۲ منطقه در شهریور ماه سال ۱۳۹۴ استخراج شد. موقعیت مناطق و دادهای در دسترس روی نقشه تهران در شکل ۱ مشخص شده است. دایره بزرگتر نشان‌دهنده قیمت مسکن بیشتر و اعداد منطقه را نشان می‌دهند. خلاصه آماره‌های مربوط به این مجموعه داده در جدول ۱ ارائه شده است. میانگین قیمت مسکن حدود چهار میلیون، ماکسیمم قیمت مسکن دوازده میلیون و مینیمم قیمت مسکن حدود یک و نیم میلیون است. نمودار بافت‌نگار، جعبه‌ای و نرمال چندک‌ـ‌چندک داده‌ها در شکل ۲ رسم شده است

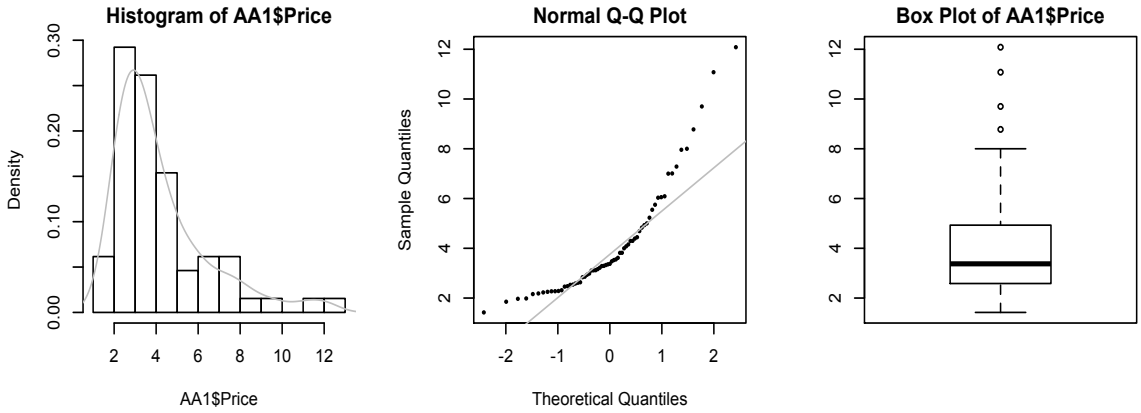
جدول ۱: آماره‌های توصیفی داده‌های مسکن تهران.

متغیر	میانگین	چارک اول	میانه	چارک سوم	مینیمم	ماکسیمم
قیمت مسکن	۱۷۷/۴	۵۸۴/۲	۳۷۶/۳	۹۳۰/۴	۴۲۳/۱	۰۸۰/۱۲

به‌دلیل عدم نرمال بودن داده‌ها از تبدیل لگاریتمی بر روی داده‌ها استفاده شد و مقدار آماره شاپیرو ویلک ۰/۰۹۶۳ به دست آمد که فرض نرمال بودن پذیرفته می‌شود. مساحت ملک به عنوان متغیر کمکی در نظر گرفته شد. داده‌ها با سه مدل رگرسیون فضایی SEM ، SLM و SDM داده‌ها با سه رهیافت ماکسیمم درستنمایی، بیزی و بیزی تقریبی تحلیل و نتایج بیزی تقریبی در جدول ۲ آورده شده است. برای تحلیل بیزی معمولی از الگوریتم گیبز و متروپلیس هاستینگز با ۵۰۰۰۰ تکرار



شکل ۱: موقعیت داده‌های متوسط قیمت مسکن شهر تهران.



شکل ۲: نمودار جعبه‌ای، چندک‌ـچندک و هیستوگرام داده‌های مسکن تهران.

استفاده شد. برازش و مقادیر مجذور میانگین توان دوم خطای نسبی[9] (RRMSE) به‌صورت

$$RRMSE = \sqrt{\frac{۱}{n}\sum_{i=۱}^{n}\frac{(y_i - \hat{y}_i)^۲}{\hat{y}_i^{۲}}},$$

برای سه مدلSLM، SDM و SEM محاسبه و در جدول ۳ ارائه شده است. نتایج جدول ۳ بیانگر برتری مدل SLM نسبت به دو مدل SEM و SDM و دقت بهتر رهیافت بیزی و بیزی تقریبی نسبت به درستنمایی و معادل بودن نتایج رهیافت بیزی معمولی با بیزی تقریبی برای این مجموعه داده است، به‌طوری‌که زمان اجرای رهیافت بیزی تقریبی کمتر از یک دقیقه

[9] Relative Root Mean Square Error

جدول ۲: برآورد بیزی تقریبی پارامترها و خلاصه آماره‌ها.

مدل	پارامتر	برآورد	انحراف معیار	چندک ۰/۰۲۵	چندک ۰/۵	چندک ۰/۹۷۵
	ρ	۰/۸۱۷۳	۰/۰۸۲۳	۰/۶۳۸۶	۰/۸۲۹۲	۰/۹۵۷۸
SLM	$\beta.$	−۰/۴۱۰۲	۰/۱۳۰۰	−۰/۶۶۴۱	−۰/۴۱۱۲	−۰/۱۳۱۷
	β_{Area}	۰/۰۰۶۹	۰/۰۰۱۱	۰/۰۰۴۶	۰/۰۰۶۹	۰/۰۰۹۳
	ρ	۰/۸۲۶۱	۲/۵۸۴	۳/۳۷۶	۴/۹۳۰	۱/۴۲۳
SDM	$\beta.$	−۰/۳۴۴۹	۰/۲۴۶۶	−۰/۸۸۹۷	−۰/۳۴۲۳	۰/۱۹۹۸
	β_{Area}	۰/۰۰۶۸	۰/۰۰۱۱	۰/۰۰۴۵	۰/۰۰۶۹	۰/۰۰۹۴
	γ	−۰/۰۰۰۷	۰/۰۰۲۶	−۰/۰۰۷۴	−۰/۰۰۰۸	۰/۰۰۶۱
	λ	۰/۹۲۰۲	۰/۰۵۸۶	۰/۷۷۳۱	۰/۹۴۱۱	۰/۹۸۹۷
SEM	$\beta.$	۰/۵۵۹۸	۰/۳۱۹۲	−۰/۹۹۰۱	۰/۵۹۶۴	۱/۸۹۷۳
	β_{Area}	۰/۰۰۶۴	۰/۰۰۱۰	۰/۰۰۴۸	۰/۰۰۶۵	۰/۰۰۸۳

جدول ۳: مقادیر مجذور میانگین مربع خطای نسبی برای سه مدل.

رهیافت	SLM	SEM	SDM
ماکسیمم درستنمایی	۰/۲۴۱۸	۰/۲۴۴۷	۰/۲۴۳۵
بیزی معمولی	۰/۲۳۶۷	۰/۲۴۰۲	۰/۲۳۸۹
بیزی تقریبی	۰/۲۳۷۱	۰/۲۳۹۷	۰/۲۳۸۸

و زمان اجرای رهیافت بیزی معمولی بیش از یک ساعت به طول انجامید.

بحث و نتیجه‌گیری

در این مقاله سه مدل رگرسیونی فضایی در اقتصادسنجی برای مشاهداتی که دارای وابستگی مکانی هستند به‌کار گرفته شد. نحوه برآورد پارامترها با استفاده از روش ماکسیمم درستنمایی و رهیافت بیزی تقریبی ارائه شد. در یک مجموعه داده واقعی به بررسی ارتباط قیمت و مساحت مسکن در شهریور ماه ۱۳۹۴ شهر تهران پرداخته شد. با توجه به عدم نرمال بودن داده‌ها از تبدیل لگاریتمی داده‌ها استفاده شد. پس از برازش مدل‌ها و مقایسات انجام شده مدل مناسب مدل SLM انتخاب و در هر سه مدل و با هر دو رهیافت بیزی تقریبی و درستنمایی وجود اثر فضایی و ارتباط مستقیم بین قیمت مسکن و مساحت پذیرفته می‌شود. با استفاده از معیار مجذور میانگین توان دوم خطای نسبی و براساس این دو مجموعه داده مشاهده شد که رهیافت بیزی تقریبی برای این نوع از مدل‌های اقتصادسنجی فضایی نسبت به رهیافت درستنمایی از عملکرد تقریبا بهتری برخوردار است و زمان محاسبات رهیافت بیزی تقریبی کمتر از رهیافت بیزی معمولی است. در مدل‌های معرفی شده برای راحتی محاسبات معمولا توزیع نرمال برای خطاها منظور می‌شود، به عنوان پیشنهاد می‌توان در مدل‌های اقتصادسنجی فضایی از توزیع‌های انعطاف‌پذیرتر مثل چوله نرمال، چوله نرمال بسته و چوله تی استفاده کرد که کلاس بزرگتری نسبت به توزیع نرمال هستند.

مراجع

Anselin, L. (1988), *Spatial Econometrics: Method and Models In: Studies in Operational Regional Science*, Springer.

Anselin, L. (2010), Thirty Years of Spatial Econometrics, *Papers in Regional Science*, **89**, 3-25.

Arbia, G. (2016), Spatial Econometrics: A Broad View, *Foundations and Trends in Econometrics*, **8**, 145-265.

Barry, R. P., Pace, R. K. (1997), Kriging with Large Data Sets using Sparse Matrix Techniques, *Communication Statistics: Computation and Simulation,* **26**, 619-629.

Bivand, R.s., Gómez-Rubiob, V., Rue, H. (2014), Approximate Bayesian Inference for Spatial Econometrics Models, *Spatial Statistics*, **9**, 146–165.

Davidson, R., Mackinnon, J. G. (1993), *Estimation and Inference in Econometrics*, Oxford University, New York.

Eidsvik, J., Finley, A. O., Banerjee, S., Rue, H. (2012), Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models, *Computational Statistics and Data Analysis.* **56**, 1362-1380.

Kelejian, H., Piras, G. (2017), *Spatial Econometrics*. London, England: Academic Press.

LeSage, J. (1999), *Spatial Econometrics*, Department of Economics University of Toledo, .

LeSage, J., Pace, R. K. (2009), Introduction to Spatial Econometrics, *Chapman and Hall/CRC.*

LeSage, J., Pace, R. K., Lam, N., Campanella, R., Liu, X. (2011), New Orlean Business Recovery in the Aftermath of Hurricane Katrina, *Journal of the Royal Statistical Society, Ser. A,* **1**74, 1007-1027.

Lesage, J., Ficher, M. (2008), Spatial Growth Regression: Model Specification, Estimation and Interpretation, *Spatial Economic Analysis*, **3**, 275–304.

Rue, H., Held, L. (2005), *Gaussian Markov Random Fields. Theory and Applications*, Chapman and Hall, London.

Rue, H. Martino, S., Chopin, N. (2009), Approximate Bayesian Inference for Latent Gaussian Model by Using Integrated Nested Laplace Approximations (with discussion), *Journal of the Royal Statistical Society.* **71**, 319–392.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

مقدمه‌ای بر تقریب لاپلاس آشیانه‌ای جمع بسته: رهیافت سریع و دقیق بیزی

سحر خلفی[۱]، علی محمدیان مصمم، علی آقامحمدی
گروه آمار، دانشگاه زنجان

چکیده: یکی از عملیات‌های مهم در استنباط بیزی محاسبه انتگرال‌هایی با بعد بالاتر است. روش‌های قدیمی مانند روش لاپلاس که با بسط مرتبه دوم تیلور از راه تجزیه انتگرال محاسبه می‌شود، وجود دارد که بسیار وقت‌گیر و اغلب مشکل همگرایی دارد. با بسط و توسعه یک روش تودرتو از یک ایده کلاسیک با روش‌های عددی جدید برای ماتریس‌های تنک در این مقاله شیوه‌ای از تقریب لاپلاس آشیانه‌ای جمع بسته ($INLA$) در انجام تقریب استنباط بیزی برای مدل‌های گاوسی پنهان مورد مطالعه قرار می‌گیرد. همچنین در این مقاله به تشریح خروجی‌های $INLA$ و سرعت و دقت بسته $R-INLA$ می‌پردازیم.

واژه‌های کلیدی: تقریب لاپلاس آشیانی جمع بسته، تقریب لاپلاس، ماتریس‌های تنک، مدل‌های گاوسی پنهان، استنباط بیزی
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30.

۱ مقدمه

شبیه‌سازی مبنی بر استنباط از طریق الگوریتم زنجیره مارکف مونت کارلو ($MCMC$) در اوایل دهه ۱۹۹۰ به جامعه آماری معرفی شد که معمولاً برای محاسبه‌ی توزیع‌های پسینی مفید است. اما وجود پارامترهای زیاد در ساختار سلسله مراتبی این مدل‌ها مخصوصاً زمانی که بعد میدان پنهان بزرگ باشد و وجود همبستگی بین عناصر میدان پنهان ویژگی‌های همگرایی و آمیختگی این الگوریم را با مشکل مواجه می‌کند. از طرفی به دلیل پیچیده بودن با مشکلی مانند طولانی بودن زمان محاسبات مواجه می‌شود که برای حل این مشکلات رو و همکاران (۲۰۱۷) روش تقریب لاپلاس آشیانه‌ای جمع بسته ($INLA$) که ترکیبی از تقریب لاپلاس و انتگرال‌های عددی است را معرفی کردند و نشان دادند که این روش در مدت زمان بسیار کوتاه و با حفظ دقت برآمد پارامترها تقریبی قابل قبول ارائه می‌دهد. در واقع $INLA$ برای برازش مدل‌های متفاوتی به ویژه داده‌های فضایی در زمینه‌های بیزی کاربرد دارد و خروجی آن شامل توزیع‌های پسینی است که می‌توان از راه میانگین، واریانس و

[۱] نام و ایمیل ارائه دهنده مقاله: سحر خلفی، s.khalafi@znu.ac.ir

چندک مورد ارزیابی قرار گیرد. نمونه‌هایی که از $INLA$ برای تجزیه و تحلیل آماری اخیراً استفاده شده عبارت هستند از: تطبیق پذیری پیشینی وزنی در رگرسیون تعمیم یافته (هلد و ساتر، ۲۰۱۶)، سرایت مالاریا در آفریقا (نور و همکاران، ۲۰۱۴)، مطالعات ارتباط بین دسترسی به مسکن و سلامت و رفاه در کلان شهرها (کانت و همکاران، ۲۰۱۶)، متاآنالیز شبکه (ساتر و همکاران، ۲۰۱۵) و غیره.

سه مولفه اصلی $INLA$ در تقریب استنباط بیزی شامل مدل‌های گاوسی پنهان (LGM)، میدان تصادفی گاوسی مارکف ($GMRF$) و تقریب لاپلاس هستند که در بخش ۲ معرفی خواهند شد. در بخش ۳ تقریب لاپلاس آشیانه‌ای جمع بسته تشریح می‌شود و در بخش ۴ با یک مثال کاربردی ساده به توصیف بیشتر این روش پرداخته می‌شود. بخش ۵ شامل بحث و نتیجه گیری است.

۲ مولفه‌های اصلی INLA

علیرغم سادگی ساختار مدل‌های گاوسی پنهان اغلب مدل‌های مورد استفاده از آمار و سایر رشته‌های کاربردی نظیر مدل‌های خطی تعمیم یافته، مدل‌های جمعی تعمیم یافته، اسپلاین‌ها، مدل‌های رگرسیون شبه پارامتری، مدل‌های فضایی و فضایی‌ـ زمانی را می‌توان در قالب مدل‌های گاوسی پنهان در نظر گرفت. مدل‌های گاوسی پنهان را با استفاده از مدل سلسله مراتبی سه مرحله‌ای زیر که در آن مشاهدات y را به طور مشروط مستقل و θ_1 را ابر پارامتر و x را متغیر پنهان می‌توان فرض کرد را به دست آورد.

$$y|x,\theta_1 \sim \prod_{i\in\tau} \pi(y_i|x_i,\theta_1)$$

که در آن میدان تصادفی گاوسی پنهان x به صورت:

$$x|\theta_2 \sim N(\mu(\theta_2), Q^{-1}(\theta_2))$$

می‌باشد و θ_2 پارامتر و Q ماتریس دقت و $\theta = (\theta_1,\theta_2)$ ابر پارامتر می‌باشد. توزیع پسین به صورت:

$$\pi(x,\theta|y) \propto \pi(\theta)\pi(x|\theta)\prod_{i\in\tau}\pi(y_i|x_i,\theta)$$

می‌باشد. از آن‌جایی‌که این توزیع اغلب فرم بسته‌ای ندارد نیازمند روش‌های عددی می‌باشیم. علاوه‌براین فرضیه‌های مهم زیر نیز در نظر گرفته می‌شوند.

۱. تعداد ابرپارامترها $|\theta|$ کم و به طور نمونه بین ۲ تا ۵ می‌باشد.

۲. توزیع میدان پنهان $x|\theta$ گاوسی می‌باشد و وقتی که بعد n بزرگ باشد (10^3 تا 10^5) به یک $GMRF$ نیاز داریم.

۳. داده‌های y به طور مشروط مستقل از x,θ هستند و اشاره دارد که هر مشاهده y_i تنها به یک مولفه از میدان پنهان x_i بستگی دارد.

x یک $GMRF$ است اگر یک میدان تصادفی گاوسی با خاصیت استقلال شرطی باشد. به این معنی که x_i, x_j به شرط سایر داده‌های باقیمانده x_{-ij} مستقل باشند. یک مثال ساده مدل اتو رگرسیو مرتبه اول به صورت زیر است:

$$x_t = \phi x_{t-1} + \varepsilon_t \; ; \quad t = 1,...,m$$

که در آن ε فرآیند اغتشاش و $|\phi| < ۱$ و اندیس t نشان دهنده‌ی زمان است. به سادگی می‌توان نشان داد که در این مدل همبستگی بین x_t, x_s، $\phi^{|s-t|}$ است و برای تمام تأخیرهای $|s-t| > ۱$، x_t, x_s داده‌های مستقل شرطی روی x_{-st} هستند. بنابراین ماتریس دقت مربوطه یک ماتریس سه قطری و در نتیجه ماتریس تنک خواهد شد که باعث سادگی محاسبات می‌شود. در حالت کلی هدف تقریب انتگرال زیر وقتی $n \to \infty$ است:

$$I_n = \int_x \exp(nf(x))dx$$

ایده اصلی این است که با استفاده از بسط تیلور $f(x)$ و تفسیر $nf(x)$ به عنوان مجموع لگ‌ـ‌درستنمایی و در نظر گرفتن x به عنوان پارامتر نامعلوم طبق قضیه حد مرکزی با درستنمایی توزیع گاوسی تقریب زده شود. این تقریب بویژه برای انتگرال‌های با ابعاد بالا بسیار مفید است. به عنوان مثال فرض کنید که علاقه‌مند به محاسبه توزیع حاشیه‌ای $\pi(\gamma_۱)$ از توزیع توأم $\pi(\gamma)$ به صورت

$$\pi(\gamma_۱) = \frac{\pi(\gamma)}{\pi(\gamma_{-۱}|\gamma_۱)} \approx \quad \frac{\pi(\gamma)}{\pi_G(\gamma_{-۱}; \mu(\gamma_۱), Q(\gamma_۱))} \mid_{\gamma_{-۱}=\mu(\gamma_۱)}$$

هستیم، که در آن $\pi(\gamma_{-۱}|\gamma_۱)$ با استفاده از توزیع گاوسی تقریب زده شده است که در مدل گاوسی پنهان بخش قبل در واقع $\gamma = (x, ۰)$ است.

۳ تقریب لاپلاس آشیانه‌ای جمع بسته

حال به معرفی $INLA$ برای حالتی که مدل گاوسی پنهان شامل پارامتر با بعد پایین و $x|\theta$ یک $GMRF$ است می‌پردازیم. همانگونه که اشاره شد هدف استنباط بیزی تقریب حاشیه‌های پسین به صورت

$$\pi(\theta_j|y)\,; j = ۱, ..., |\theta| \quad , \quad (x_i|y)\,; i = ۱, ..., n$$

است. دراین روش درستنمایی مستقل شرطی است به معنی اینکه y_i تنها به x_i, θ بستگی دارد. مدل کلی زیر را در نظر بگیرید:

$$\eta_i = g(\beta)u_{j(i)}$$

که در آن برای نمونه $i = ۱, ..., n$ ،$y_i|\eta_i \sim poisson(\exp(\eta_i))$ ،$\beta \sim N(۰, ۱)$ ،$g(\cdot)$ تابع یکنوا خوش رفتار و $u \sim N(۰, Q^{-۱})$ است. اندیس $j(i)$ به معنی نگاشتی است که بعد u ثابت بوده و به n بستگی ندارد. محاسبه حاشیه‌های پسین برای β, u_j از آنجایی که یک حاصل ضرب گاوسی و یک غیر گاوسی داریم مشکل است. لذا تقریب به زیر مجموعه‌های کوچکتر تقسیم و از تقریب لاپلاس استفاده می‌شود:

$$\pi(\beta|y) \propto \pi(\beta) \int \prod_{i=۱}^{n} \pi(y_i|\lambda_i = \exp(g(\beta)u_{j(i)})) \times \pi(u)du$$

برای هر u_j توزیع حاشیه‌ای به صورت

$$\pi(u_j|y) = \int \pi(u_j|\beta, y) \times \pi(\beta|y)d\beta$$

است و از آن‌جایی‌که β یک پارامتر تک بعدی است این محاسبات به سادگی انجام می‌گیرد.

$$\pi(u|\beta, y) \propto \prod_{i=۱}^{n} \pi(y_i|\lambda_i = \exp(g(\beta)u_{j(i)})) \times \pi(u)$$

به همین ترتیب که در اینجاست چون u_j ثابت در نظر گرفته می‌شود محاسبه $\pi(u_j|\beta,y)$ شامل انتگرال‌گیری عددی است. حال هدف محاسبه توزیع پسینی برای هر θ_j است. برای این منظور یک تقریب برای

$$\pi(\theta|y) \propto \frac{\pi(\theta)\pi(x|\theta)\pi(y|x,\theta)}{\pi(x|\theta,y)}$$

در نظر گرفته می‌شود که در آن مخرج کسر با یک تقریب گاوسی به صورت

$$\begin{aligned}\pi(x|y,\theta) \propto & \exp(-\frac{1}{2}x^TQ(\theta)x+\sum_i \log\pi(y_i|x_i,\theta))\\ =&(2\pi)^{-n/2}|P(\theta)|^{1/2}\exp(-\frac{1}{2}(x-\mu(\theta))^TP(\theta)(x-\mu(\theta))\end{aligned}$$

محاسبه می‌شود، که در آن $P(\theta)=Q(\theta)+diag(c(\theta))$، $\mu(\theta)$ مکان مد، $c(\theta)$ حاوی منفی مشتقات دوم نسبت به x_i از لگ‌ـدرستنمایی در نقطه مد است. مزیت این تقریب این است که مسئله به یک $GMRF$ خلاصه شده، در نتیجه محاسبات ساده‌تر و تقریب دقیق‌تر خواهدشد. سرانجام توزیع‌های پسینی برای میدان پنهان را می‌توان تقریب زد. همانند قبل در این حالت داریم:

$$\pi(x_i|y)=\int \pi(x_i|\theta,y)\pi(\theta|y)d\theta$$

در اینجا ما به انتگرال‌گیری روی $\pi(\theta|y)$ نیاز داریم اما هزینه محاسبات از انتگرال‌گیری عددی استاندارد از مرتبه نهایی بسته به بعد θ است. همچنین ما به تقریب $\pi(x_i|\theta,y)$ برای تمام $i=1,...,n$ با n بسیار بزرگ نیاز داریم که با استفاده از تقریب لاپلاس امکان‌پذیر است.

۴ مثال کاربردی

در این بخش با یک مثال شبیه‌سازی به معرفی بسته $R-INLA$ در نرم‌افزار R می‌پردازیم. برای این منظور فرض کنید که داده‌ها از

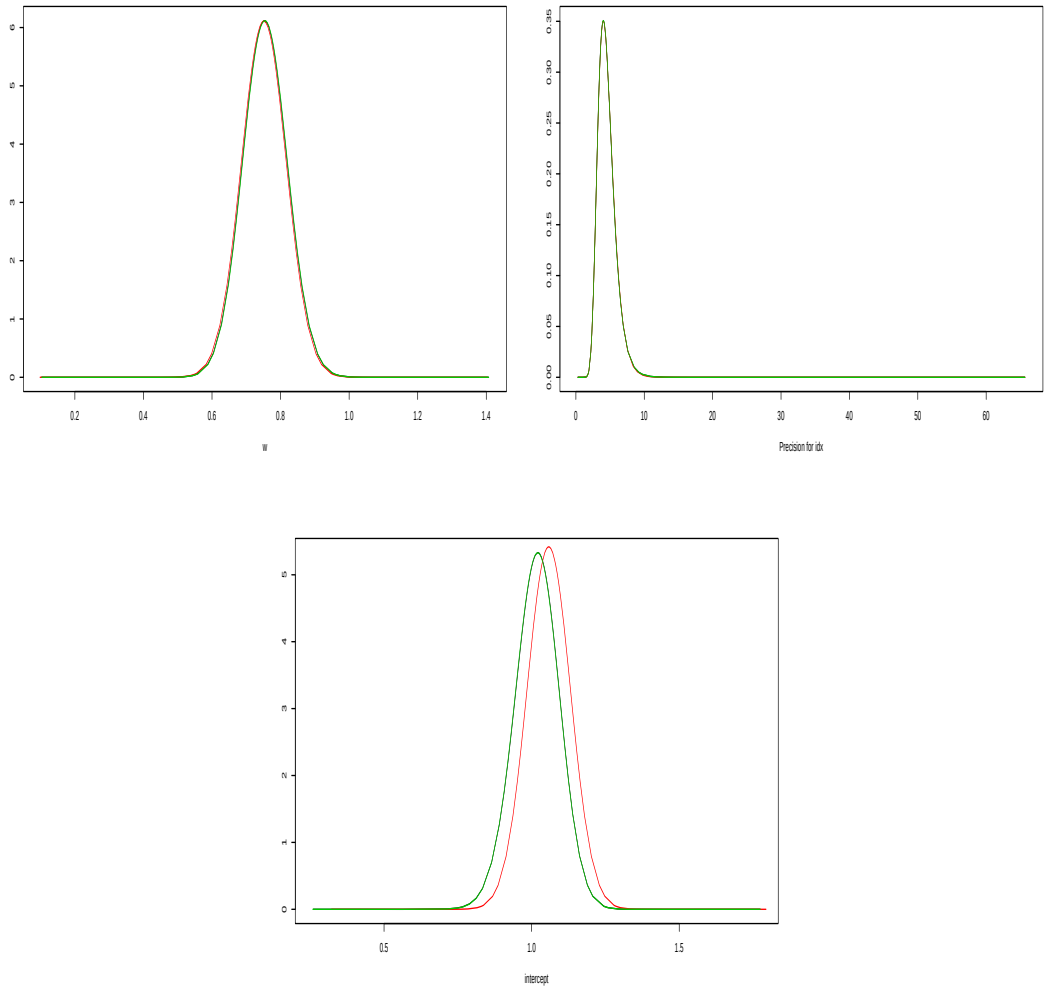
$$y|\eta \sim Poisson(\exp(\eta))$$

که در آن $\eta_i=\mu+\beta w_i+u_{j(i)}; i=1,...,n$ است، w متغیر کمکی، $u\sim N_m(0,\tau^{-1}I)$ می‌باشند، تولید شده باشند. برای تولید $n=200$ داده ما فرض کرده‌ایم که $\mu=1$، $\beta=3/4$ و $\tau=4/16$ باشد. حال با استفاده از پکیج $R-INLA$ به تحلیل بیزی داده‌های شبیه‌سازی شده می‌پردازیم.

شکل ۱ برآورد پسینی حاشیه‌ای اثرات ثابت $(intercept, w)$ و ابر پارامتر $(precison\ \ for\ \ idx)$ را با استفاده از سه روش تقریب لاپلاس (خط مشکی)، تقریب گاوسی (خط قرمز) که مانع از انتگرال‌گیری روی θ شده و انتگرال دقیق روی θ (خط سبز) نشان می‌دهد. همچنین خلاصه آماری مربوط به برآورد پارامترها در جداول ۳–۱ ارائه شده است. همان‌گونه که ملاحظه می‌شود تقریب لاپلاس بسیار نزدیک به تقریب دقیق انتگرال می‌باشد.

۵ بحث و نتیجه‌گیری

محاسبات پیچیده تحلیل بیزی یکی از مشکلات اساسی کاربران روش‌های بیزی به ویژه وقتی که اندازه نمونه و بعد پارامترها زیاد است می‌باشد. در این مقاله روش $INLA$ به عنوان روش جایگزین معرفی گردید که می‌تواند بسیار سریع‌تر و حتی دقیق‌تر از روش‌های کلاسیک نظیر $MCMC$ باشد.



شکل ۱: برآورد پسینی حاشیه‌ای اثرات ثابت و ابر پارامتر با سه روش تقریب لاپلاس (خط مشکی)، تقریب گاوسی (خط قرمز) و انتگرال دقیق روی θ (خط سبز)

جدول ۱: خلاصه آماری مربوط به برآورد پارامترها به روش تقریب لاپلاس

	میانگین	انحراف معیار	چارک اول	میانه	چارک سوم	مد
عرض از مبدأ	۱/۰۱۷	۰/۰۷۶	۰/۸۶۴	۱/۰۱۹	۱/۱۶۲	۱/۰۲۱
w	۰/۷۵۴	۰/۰۶۵	۰/۶۲۷	۰/۷۵۴	۰/۸۸۳	۰/۷۵۳
دقت	۴/۵۴۰	۱/۲۹۰	۲/۵۷۰	۴/۳۵۰	۷/۶۰۰	۴/۰۱۰

مراجع

Held, L., and Sauter, R. (2017), Adaptive Prior Weighting in Generalized Regression, *Biometrics*, **73**, 242-251.

Kandt, J., Chang, S. S., Yip, P., and Burdett, R. (2017), The Spatial Pattern of Premature Mortality

جدول ۲: خلاصه آماری مربوط به برآورد پارامترها به روش تقریب گاوسی

	میانگین	انحراف معیار	چارک اول	میانه	چارک سوم	مد
عرض از مبدأ	۱/۰۵۸	۰/۰۷۴	۰/۹۱۳	۱/۰۵۸	۱/۲۰۲	۱/۰۵۸
w	۰/۷۵۱	۰/۰۶۵	۰/۶۲۳	۰/۷۵۱	۰/۸۷۹	۰/۷۵۱
دقت	۴/۵۴۰	۱/۲۷۰	۲/۵۷۰	۴/۳۶۰	۷/۵۳۰	۴/۰۲۰

جدول ۳: خلاصه آماری مربوط به برآورد پارامترها به روش انتگرال دقیق

	میانگین	انحراف معیار	چارک اول	میانه	چارک سوم	مد
عرض از مبدأ	۱/۰۱۷	۰/۰۷۶	۰/۸۶۵	۱/۰۱۸	۱/۱۶۲	۱/۰۲۱
w	۰/۷۵۵	۰/۰۶۵	۰/۶۲۸	۰/۷۵۵	۰/۸۸۴	۰/۷۵۴
دقت	۴/۵۳۰	۱/۳۰۰	۲/۵۶۰	۴/۳۵۰	۷/۶۴۰	۴/۰۳۰

in Hong Kong: How Does it Relate to Public Housing? *Urban Studies*, **54**, 1211-1234.

Noor, A. M., Kinyoki, D. K., Mundia, C. W., Kabaria, C. W., Mutua, J. W., Alegana, V. A., Fall, I. S., and Snow, R. W. (2014), The Changing Risk of Plasmodium Falciparum Malaria Infection in Africa: 2000-10: A Spatial and Temporal Analysis of Transmission Intensity, *The Lancet*, **383**, 1739–1747.

Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P. and Lindgren, F.K., (2017), Bayesian Computing with INLA: A Review, *Annual Review of Statistics and Its Application*, **4**, 395-421.

Sauter, R. and Held, L. (2015), Network Meta-Analysis with Integrated Nested Laplace Approximations, *Biometrical Journal*, **57**, 1038–1050.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

میانگین‌گیری بیزی مدل‌های رگرسیونی فضایی تحت توزیع گاوسی وارون

زهرا رحیمیان آزاد[1]، افشین فلاح، زهرا زارع‌پور یکنمی
گروه آمار، دانشگاه بین‌المللی امام خمینی

چکیده: میانگین‌گیری بیزی مدل‌ها روشی توانمند در مدل‌سازی است که برخلاف روش مرسوم گزینش مدل، از اطلاعات همه یا مجموعه‌ای از بهترین مدل‌ها استفاده می‌کند. از جمله برتری‌های این روش بر سایر روش‌های مدل‌سازی می‌توان به توانایی آن در لحاظ نمودن عدم حتمیت واقعی فرایند مدل‌سازی اشاره کرد. این مقاله میانگین‌گیری بیزی مدل‌های رگرسیونی فضایی در شرایطی که متغیر پاسخ دارای توزیع گاوسی وارون است را مورد توجه قرار می‌دهد. رویکرد پیشنهادی در قالب یک مطالعه شبیه‌سازی مورد ارزیابی قرار گرفته است.

واژه‌های کلیدی: داده‌های فضایی، میانگین‌گیری بیزی مدل‌ها، توزیع گاوسی وارون.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J12، 62F15، 62H11.

۱ مقدمه

روش‌های کلاسیک مدل‌سازی آماری برپایه فرض استقلال مشاهدات توسعه یافته‌اند. این در حالی است که در بسیاری از کاربردها مشاهدات به دلیل موقعیت قرار گرفتن در فضای مورد مطالعه مستقل نیستند. تحلیل چنین داده‌هایی که داده‌های فضایی نامیده می‌شوند، با استفاده از روش‌های کلاسیک به دلیل در نظر نگرفتن وابستگی مشاهدات می‌تواند به صورت جدی گمراه‌کننده باشد. در داده‌های فضایی وابستگی مشاهدات، تابعی از از فاصله مشاهدات از یکدیگر است. به این معنی که معمولاً مشاهدات نزدیکتر وابستگی بیشتری دارند. مجموعه روش‌های آماری که وابستگی مشاهدات را در مدل‌سازی و استنباط مد نظر قرار می‌دهند، روش‌های آمار فضایی نامیده می‌شوند. امروزه مباحث آمار فضایی در علوم مختلف از جمله زمین‌شناسی (ماترون، ۱۹۶۲)، همه‌گیرشناسی (دیگل، ۱۹۹۳)، کشاورزی (اسمیت، ۱۹۳۸)، شیمی (ویتن، ۱۹۶۲) و غیره کاربرد گسترده‌ای دارند. کرسی (۱۹۹۳) یک منبع کامل درباره روش‌های مختلف آمار فضایی و کاربردهای آن است. از طرفی در شیوه مرسوم مدل‌سازی که گزینش مدل نامیده می‌شود، تحلیل‌گر براساس یک یا چند معیار، مدلی را از میان

[1] نام و ایمیل ارائه دهنده مقاله: زهرا رحیمیان آزاد، z.rahimian@edu.ikiu.ac.ir

مجموعه مدل‌های ممکن برمی‌گزینند. این شیوه مدل‌سازی از جنبه‌های مختلف مورد استفاده قرار می‌گیرد. از جمله اینکه معیار ارزیابی تابع نظر تحلیل‌گر است و از این رو ممکن است تحلیل‌گران متفاوت، مدل‌های متفاوتی را به عنوان بهترین مدل انتخاب کنند. از طرف دیگر انتخاب یک مدل باعث از دست رفتن اطلاعات سایر مدل‌ها می‌شود. میانگین‌گیری بیزی مدل‌ها، روشی نوین در مدل‌سازی است که بدون تکیه بر یک مدل منتخب از اطلاعات همه مدل‌ها استفاده می‌نماید. این روش به هر مدل احتمالی اختصاص و از میانگین همه مدل‌ها یا تعدادی از بهترین آنها استفاده می‌کند. اولین کار در این زمینه توسط لیمر (۱۹۷۸) انجام گرفت. مادیگان و رفتری (۱۹۹۴) دو روش پایه‌ای برای اجرای این شیوه مدل‌سازی ارائه کردند. رفتری (۱۹۹۵) و رفتری و همکاران (۱۹۹۳، ۱۹۹۷) میانگین‌گیری بیزی را در چارچوب مدل‌های رگرسیونی مورد بررسی قرار دادند. هوئتینگ (۱۹۹۴) و هوئتینگ و همکاران (۱۹۹۶، ۱۹۹۹) به نحوه انجام محاسبات و چگونگی تعیین مولفه‌های لازم در اجرای این روش پرداختند. در این مقاله برمبنای توزیع گاوسی وارون، نظریه لازم برای میانگین‌گیری بیزی مدل‌های رگرسیونی در شرایطی که مشاهدات دارای همبستگی فضایی هستند، توسعه داده شده است. در این راستا نحوه تعیین توزیع پیشین برای مدل‌ها و پارامترها، و چگونگی محاسبه توزیع پسین متناظر آنها، مورد بحث قرار گرفته است. در بخش ۲ روش میانگین‌گیری بیزی مدل‌ها به‌صورت اجمالی شرح داده می‌شودو در بخش سوم مدل پیشنهادی ارائه و چگونگی برازش آن مورد بحث قرار می‌گیرد. نحوه محاسبه کمیت‌های مورد نیاز برای روش میانگین‌گیری بیزی در این بخش توضیح داده می‌شود. در بخش ۴ روش پیشنهادی در قالب یک مطالعه شبیه‌سازی مورد ارزیابی قرار می‌گیرد.

۲ میانگین‌گیری بیزی مدل‌ها

فرض کنید $\mathcal{M} = \{M_۱, \ldots, M_T\}$ نشان‌دهنده فضای مدل، D مجموعه مشاهدات و Δ کمیت مورد علاقه مانند یک مشاهده در آینده یا یک پارامتر باشد. در این صورت، توزیع پسین Δ را می‌توان با استفاده از قاعده احتمال کل به‌صورت یک میانگین وزنی از توزیع‌های پیش‌بین تحت مدل‌های $M_۱, \ldots, M_T$، به‌شکل

$$\pi(\Delta|D) = \sum_{k=۱}^{T} p(\Delta|M_k, D)\pi(M_k|D), \qquad (۱.۲)$$

نوشت، که در آن وزن‌ها احتمال‌های پسین مدل‌ها هستند. برپایه رابطه (۱.۲) میانگین و واریانس پسین Δ به‌ترتیب به‌صورت

$$\begin{aligned} \mathbb{E}(\Delta|D) &= \sum_{k=۱}^{T} \mathbb{E}(\Delta|M_k, D)\pi(M_k|D) \\ &= \sum_{k=۱}^{T} \hat{\Delta}_k \pi(M_k|D), \end{aligned}$$

و

$$\begin{aligned} \mathbb{V}ar\,(\Delta|D) &= \mathbb{E}_M\left(\mathbb{V}ar(\Delta|D, M)\right) + \mathbb{V}ar_M\left(\mathbb{E}(\Delta|D, M)\right) \\ &= \sum_{k=۱}^{T} (\mathbb{V}ar\,(\Delta|D, M_k) + \hat{\Delta}_k^۲)\pi(M_k|D) - \mathbb{E}\,(\Delta|D)^۲\,, \end{aligned} \qquad (۲.۲)$$

هستند، که در آنها $\hat{\Delta}_k$ براورد بیزی کمیت Δ را تحت مدل k، $k = ۱, \ldots, T$، نشان می‌دهد (لیمر ، ۱۹۷۸). در رابطه (۲.۲) مولفه $\mathbb{V}ar_M\left(\mathbb{E}(\Delta|D, M)\right)$ عدم حتمیت بین مدل‌ها را نشان می‌دهد. در رابطه (۱.۲)، توزیع پیش‌بین کمیت مورد علاقه به‌صورت۳

$$p(\Delta|M_k, D) = \int p(\Delta|\boldsymbol{\theta}_k, M_k)\pi(\boldsymbol{\theta}_k|M_k, D)\mathrm{d}\boldsymbol{\theta}_k, \qquad (۳.۲)$$

قابل محاسبه است، که در آن $\boldsymbol{\theta}_k$ بردار پارامترهای مدل M_k را نشان می‌دهد. احتمال پسین درست بودن مدل M_k با استفاده از قاعده بیز به‌صورت

$$\pi(M_k|D) = \frac{p(D|M_k)\pi(M_k)}{\sum_{j=1}^{T} p(D|M_j)\pi(M_j)}, \qquad (۴.۲)$$

به‌دست می‌آید، که در آن

$$p(D|M_k) = \int p(D|\boldsymbol{\theta}_k, M_k)\pi(\boldsymbol{\theta}_k|M_k)\mathrm{d}\boldsymbol{\theta}_k, \qquad (۵.۲)$$

درستنمایی حاشیه‌ای مدل M_k، $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \lambda_k)$ بردار پارامترها در مدل M_k، $p(D|\boldsymbol{\theta}_k, M_k)$ تابع درستنمایی، $\pi(M_k)$ احتمال پیشین درست بودن مدل M_k و $\pi(\boldsymbol{\theta}_k|M_k)$ توزیع پیشین پارامترهای مدل M_k است. برای محاسبه $\pi(\Delta|D)$ باید مولفه‌های تشکیل دهنده آن را مشخص و جایگزین کرد. بنابراین لازم است توزیع پیشین پارامترها، احتمال پیشین و پسین هر مدل و توزیع پیش‌بین کمیت مورد علاقه را برای همه مدل‌های موجود در فضای مدل، مشخص کرد.

۳ مدل پیشنهادی

با فرض آنکه متغیر پاسخ دارای توزیع گاوسی وارون است، مدل رگرسیونی

$$y_i|\boldsymbol{x}_i^{(k)} \sim IG(\mu(\boldsymbol{x}_i^{(k)}), \lambda) \qquad \mu^{-1}(\boldsymbol{x}_i^{(k)}) = \boldsymbol{x}_i'^{(k)}\boldsymbol{\beta}^{(k)} + \phi_i^{(k)}, i = 1, \ldots, n, \qquad (۱.۳)$$

را در نظر بگیرید، که در آن $IG(\cdot,\cdot)$ توزیع گاوسی وارون را نشان می‌دهد، $\boldsymbol{\beta}^{(k)} = (\beta_1^{(k)}, \beta_2^{(k)}, \ldots, \beta_p^{(k)})'$ بردار ضرایب رگرسیونی، $\boldsymbol{x}_i^{(k)} = (x_{i1}^{(k)}, x_{i2}^{(k)}, \ldots, x_{ip}^{(k)})'$ بردار متغیرهای تبیینی و $\boldsymbol{\phi}^{(k)} = (\phi_1^{(k)}, \phi_2^{(k)}, \ldots, \phi_n^{(k)})'$ بردار اثرهای تصادفی فضایی است. اثر تصادفی فضایی $\phi_i^{(k)}$ را می‌توان به‌صورت شرطی بر اثرهای تصادفی فضایی همسایگی‌ها به‌صورت

$$\phi_i^{(k)}|\boldsymbol{\phi}_{-i}^{(k)}, \sigma^{2(k)} \sim N\left(\xi \sum_{j\neq i} \frac{w_{ij}}{w_{i+}}\phi_j^{(k)}, \frac{\sigma^{2(k)}}{w_{i+}}\right),$$

در نظر گرفت، که در آن $\boldsymbol{\phi}_{-i}^{(k)} = \{\phi_j^{(k)}; j \neq i\}$، $\sigma^{2(k)}$ واریانس مشترک اثرهای تصادفی فضایی، w_{ij} درایه (i,j) از ماتریس همسایگی است که به‌شکل

$$w_{ij} = \begin{cases} 1, & \text{اگر مشاهده‌ی } i \text{ و } j \text{ همسایه باشند.} \\ 0, & i = j \end{cases}$$

تعریف می‌شود. همچنین پارامتر ξ که در بازه $\left(\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}}\right)$ تغییر می‌کند، پارامتر هموارسازی فضایی است، که در آن $\lambda_{(1)} \geq \ldots \geq \lambda_{(n)}$ مقادیر ویژه ماتریس $\boldsymbol{D}_{\boldsymbol{W}}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}_{\boldsymbol{W}}^{-\frac{1}{2}}$ هستند، $\boldsymbol{W} = \left(\frac{w_{ij}}{w_{i+}}\right)$ ماتریس وزن و $\boldsymbol{D}_{\boldsymbol{W}}$ ماتریس قطری با درایه‌های قطری $\{w_{1+}, \ldots, w_{n+}\}$ است.

تابع درستنمایی مدل به‌صورت

$$L(\boldsymbol{\beta}^{(k)}, \lambda, \boldsymbol{\phi}^{(k)}, \sigma^{2(k)}|y_1, \ldots, y_n) = \frac{\lambda^{\frac{n}{2}}}{(2\pi y_i^3)^{\frac{n}{2}}} \exp\left\{\frac{-\lambda}{2}\sum_{i=1}^{n}\frac{(y_i - \mu_i^{(k)})^2}{y_i\mu_i^{2(k)}}\right\}$$

$$\begin{aligned} &\propto \lambda^{\frac{n}{۲}} \exp\left\{\frac{-\lambda}{۲}\left((\boldsymbol{Y}\boldsymbol{X}^{(k)}\boldsymbol{\beta}^{(k)}-\boldsymbol{۱}_n)'\boldsymbol{Y}^{-۱}(\boldsymbol{Y}\boldsymbol{X}^{(k)}\boldsymbol{\beta}^{(k)}-\boldsymbol{۱}_n)\right.\right.\\ &+ \left.\left. ۲\boldsymbol{\phi}'^{(k)}\boldsymbol{Y}\boldsymbol{X}^{(k)}\boldsymbol{\beta}^{(k)}+\boldsymbol{\phi}'^{(k)}\boldsymbol{Y}\boldsymbol{\phi}^{(k)}-۲\boldsymbol{۱}'_n\boldsymbol{\phi}^{(k)}\right)\right\}\\ &= \lambda^{\frac{n}{۲}} \exp\left\{\frac{-\lambda}{۲}\left(Q_۱(\boldsymbol{\beta}^{(k)})+g(\boldsymbol{\phi}^{(k)})\right)\right\} \end{aligned} \quad (۲.۳)$$

نوشته می‌شود که در آن $Q_۱(\boldsymbol{\beta}^{(k)})$ یک نمایش درجه دوم برحسب $\boldsymbol{\beta}^{(k)}$، $\boldsymbol{Y}$ یک ماتریس قطری با درایه‌های قطری $(y_۱,\ldots,y_n)$ و $g(\boldsymbol{\phi}^{(k)})$ تابعی برحسب $\boldsymbol{\phi}^{(k)}$ است.

۱.۳ تعیین توزیع پیشین پارامترها تحت مدل kام

لازمه تحلیل بیزی مدل (۱.۳) آن است که برای پارامترهای آن، $(\boldsymbol{\beta}^{(k)},\lambda,\boldsymbol{\phi}^{(k)},\sigma^{۲(k)})$، توزیع‌های پیشین مناسبی در نظر گرفته شود که بتوانند به خوبی باورهای پیشین تحلیل‌گر را در مدل‌سازی بازتاب دهند. از این رو، توزیع پیشین توام پارامترها را می‌توان با فرض استقلال پیشینی $(\boldsymbol{\beta}^{(k)},\lambda)$ و $(\boldsymbol{\phi}^{(k)},\sigma^{۲(k)})$ به‌شکل

$$\begin{aligned} \pi(\boldsymbol{\beta}^{(k)},\lambda,\boldsymbol{\phi}^{(k)},\sigma^{۲(k)}) &= \pi(\boldsymbol{\beta}^{(k)},\lambda)\pi(\boldsymbol{\phi}^{(k)},\sigma^{۲(k)})\\ &= \pi(\boldsymbol{\beta}^{(k)}|\lambda)\pi(\lambda)\pi(\boldsymbol{\phi}^{(k)}|\sigma^{۲(k)})\pi(\sigma^{۲(k)}), \end{aligned} \quad (۳.۳)$$

نوشت. با توجه به این که دامنه تغییرات ضرایب رگرسیونی مجموعه اعداد حقیقی است و پارامتر λ، تنها مقادیر مثبت را اختیار می‌کند، توزیع‌های پیشین تحت مدل M_k به‌صورت

$$\begin{aligned} \boldsymbol{\beta}^{(k)}|\lambda &\sim N_{p_k}(\boldsymbol{\eta}_۰,\lambda^{-۱}\boldsymbol{V}_۰) \quad \boldsymbol{\beta}^{(k)},\boldsymbol{\eta}_۰\in\mathbb{R}^{p_k}\\ \lambda &\sim Gamma\left(\frac{a_۰}{۲},\frac{b_۰}{۲}\right) \quad \lambda,a_۰,b_۰>۰, \end{aligned}$$

در نظر گرفته می‌شوند، به‌گونه‌ای که $\boldsymbol{V}_۰=diag(v_{۰۱},\ldots,v_{۰p_k})$ یک ماتریس قطری با درایه‌های قطری مثبت است و $\boldsymbol{\eta}_۰=(\eta_{۰۱},\ldots,\eta_{۰p_k})$ و $\boldsymbol{\beta}^{(k)}=(\beta_۱^{(k)},\ldots,\beta_{p_k}^{(k)})$، کمیت‌های $\boldsymbol{\eta}_۰$، $\boldsymbol{V}_۰$، $a_۰$ و $b_۰$ که با اندیس صفر متمایز شده‌اند، ابرپارامترهای مدل بیزی را نشان می‌دهند و توسط تحلیل‌گر و براساس باورهای پیشین وی درباره پارامترها تعیین می‌شوند (مشکانی و همکاران، ۲۰۱۶). بنابراین توزیع پیشین توام $(\boldsymbol{\beta}^{(k)},\lambda)$ به‌صورت

$$\begin{aligned} \pi(\boldsymbol{\beta}^{(k)},\lambda) &= \pi(\boldsymbol{\beta}^{(k)}|\lambda)\pi(\lambda)\\ &\propto \lambda^{\frac{a_۰+p_k}{۲}-۱}\exp\left\{\frac{-\lambda}{۲}b_۰\right\}\exp\left\{\frac{-\lambda}{۲}\left((\boldsymbol{\beta}^{(k)}-\boldsymbol{\eta}_۰)'\boldsymbol{V}_۰^{-۱}(\boldsymbol{\beta}^{(k)}-\boldsymbol{\eta}_۰)\right)\right\}, \end{aligned} \quad (۴.۳)$$

به‌دست می‌آید، که یک توزیع نرمال ـ گاما است. براساس لم بروک (بانرجی و همکاران، ۲۰۰۴) توزیع توام بردار اثرهای تصادفی فضایی، $\boldsymbol{\phi}^{(k)}$، به شرط معلوم بودن $\sigma^{۲(k)}$ به‌صورت

$$\boldsymbol{\phi}^{(k)}|\sigma^{۲(k)} \sim N_n\left(۰,(\boldsymbol{D_W}-\xi\boldsymbol{W})^{-۱}\sigma^{۲(k)}\right),$$

خواهد بود. از این رو اگر برای پارامتر $\sigma^{۲(k)}$ نیز توزیع گامای وارون به‌شکل

$$\frac{۱}{\sigma^{۲(k)}} \sim Gamma(f_۰,g_۰),$$

در نظر گرفته شود، که در آنها $f_۰$ و $g_۰$ ابرپارامترهای توزیع پیشین هستند، توزیع توام $(\boldsymbol{\phi}^{(k)},\sigma^{۲(k)})$ عبارت است از:

$$\begin{aligned} \pi(\boldsymbol{\phi}^{(k)},\sigma^{۲(k)}) &= \pi(\boldsymbol{\phi}^{(k)}|\sigma^{۲(k)})\pi(\sigma^{۲(k)})\\ &\propto \exp\left\{\frac{-۱}{۲\sigma^{۲(k)}}\boldsymbol{\phi}'^{(k)}(\boldsymbol{D_W}-\xi\boldsymbol{W})^{-۱}\boldsymbol{\phi}^{(k)}\right\}\left(\frac{۱}{\sigma^{۲(k)}}\right)^{f_۰+۱}\exp\left\{\frac{-g_۰}{\sigma^{۲(k)}}\right\} \end{aligned} \quad (۵.۳)$$

۲.۳ محاسبه توزیع پسین توام پارامترها تحت مدل kام

توزیع پسین توام پارامترها را می‌توان با استفاده از تابع درستنمایی (۲.۳) و توزیع پیشین توام (۴.۳) و (۵.۳)، به‌صورت

$$\begin{aligned}
\pi(\boldsymbol{\beta}^{(k)},\lambda,\boldsymbol{\phi}^{(k)},\sigma^{۲(k)}|\boldsymbol{y}) &\propto L(\boldsymbol{\beta}^{(k)},\lambda,\boldsymbol{\phi}^{(k)},\sigma^{۲(k)}|\boldsymbol{y})\pi(\boldsymbol{\beta}^{(k)},\lambda,\boldsymbol{\phi}^{(k)},\sigma^{۲(k)})\\
&= \lambda^{\frac{a_{۰}+n+p_k}{۲}-۱}\exp\left\{-\frac{\lambda}{۲}\left(b_{۰}+(\boldsymbol{\beta}^{(k)}-\boldsymbol{\eta}_{۰})'\boldsymbol{V}_{۰}^{-۱}(\boldsymbol{\beta}^{(k)}-\boldsymbol{\eta}_{۰})\right)\right\}\\
&\times \exp\left\{-\frac{\lambda}{۲}(\boldsymbol{Y}\boldsymbol{X}^{(k)}\boldsymbol{\beta}^{(k)}-\boldsymbol{۱}_n)'\boldsymbol{Y}^{-۱}(\boldsymbol{Y}\boldsymbol{X}^{(k)}\boldsymbol{\beta}^{(k)}-\boldsymbol{۱}_n)\right\}\\
&\times \exp\left\{\frac{-۱}{۲\sigma^{۲(k)}}\boldsymbol{\phi}'^{(k)}(\boldsymbol{D}_{\boldsymbol{W}}-\xi\boldsymbol{W})^{-۱}\boldsymbol{\phi}^{(k)}\right\}\left(\frac{۱}{\sigma^{۲(k)}}\right)^{f_{۰}+۱}\exp\left\{\frac{-g_{۰}}{\sigma^{۲(k)}}\right\}\\
&= \lambda^{\frac{a_{۰}+n+p_k}{۲}-۱}\exp\left\{-\frac{\lambda}{۲}\left(Q_{۲}(\boldsymbol{\beta}^{(k)})+u^{(k)}+g(\boldsymbol{\phi}^{(k)})\right)\right\}\\
&\times \left(\frac{۱}{\sigma^{۲(k)}}\right)^{f_{۰}+۱}\exp\left\{\frac{-۱}{\sigma^{۲(k)}}\left(Q_{۳}(\boldsymbol{\phi}^{(k)})\right)\right\} \qquad (۶.۳)
\end{aligned}$$

نوشت، که در آن

$$Q_{۲}(\boldsymbol{\beta}^{(k)}) = (\boldsymbol{\beta}^{(k)}-{\boldsymbol{\Sigma}^{(k)}}^{-۱}\boldsymbol{d}^{(k)})'\boldsymbol{\Sigma}^{(k)}(\boldsymbol{\beta}^{(k)}-{\boldsymbol{\Sigma}^{(k)}}^{-۱}\boldsymbol{d}^{(k)}), \qquad (۷.۳)$$

یک نمایش درجه دوم بر حسب $\boldsymbol{\beta}^{(k)}$ را نمایش می‌دهد،

$$u^{(k)} = b_{۰}+\boldsymbol{۱}_n'\boldsymbol{Y}^{-۱}\boldsymbol{۱}_n+\boldsymbol{\eta}_{۰}'\boldsymbol{V}_{۰}^{-۱}\boldsymbol{\eta}_{۰}-\boldsymbol{d}^{(k)'}{\boldsymbol{\Sigma}^{(k)}}^{-۱}\boldsymbol{d}^{(k)}, \qquad (۸.۳)$$

کمیتی وابسته به مشاهدات است و

$$g(\boldsymbol{\phi}^{(k)}) = ۲\boldsymbol{\phi}'^{(k)}\boldsymbol{Y}\boldsymbol{X}^{(k)}\boldsymbol{\beta}^{(k)}+\boldsymbol{\phi}'^{(k)}\boldsymbol{Y}\boldsymbol{\phi}^{(k)}-۲\boldsymbol{۱}_n'\boldsymbol{\phi}^{(k)}, \qquad (۹.۳)$$

$$\begin{aligned}
Q_{۳}(\boldsymbol{\phi}^{(k)}) &= \frac{\boldsymbol{\phi}'^{(k)}(\boldsymbol{D}_{\boldsymbol{W}}-\xi\boldsymbol{W})^{-۱}\boldsymbol{\phi}^{(k)}}{۲}-g_{۰}, \qquad (۱۰.۳)\\
\boldsymbol{\Sigma}^{(k)} &= (\boldsymbol{V}_{۰}^{-۱}+\boldsymbol{X}^{(k)'}\boldsymbol{Y}\boldsymbol{X}^{(k)}),\\
\boldsymbol{d}^{(k)} &= \boldsymbol{V}_{۰}^{-۱}\boldsymbol{\eta}_{۰}+n\bar{\boldsymbol{x}}^{(k)'}.
\end{aligned}$$

همان‌گونه که ملاحظه می‌شود توزیع پسین توام پارامترها (۶.۳) پیچیده بوده و دارای نمایش بسته نیست. از این رو، برای استنباط بیزی لازم است از روش‌های مونت کارلو زنجیر مارکفی برای نمونه‌گیری از توزیع پسین توام پارامترها و تقریب کمیت‌های پسینی مورد علاقه استفاده کرد. به منظور استفاده از الگوریتم گیبز برای نمونه‌گیری از توزیع پسین توام پارامترها، لازم است توزیع‌های پسین شرطی کامل پارامترها را به‌دست آورد.

۳.۳ توزیع‌های پسین شرطی کامل

برپایه‌ی توزیع پسین توام پارامترها در رابطه‌ی (۶.۳)، توزیع پسین شرطی کامل پارامترهای λ، $\sigma^{۲(k)}$، $\boldsymbol{\beta}^{(k)}$ و $\boldsymbol{\phi}^{(k)}$ را می‌توان به‌ترتیب به‌شکل

$$\lambda \mid others \sim Gamma\left(\frac{a_{۰}+p_k+n}{۲}, \frac{Q_{۲}(\boldsymbol{\beta}^{(k)})+u^{(k)}+g(\boldsymbol{\phi}^{(k)})}{۲}\right),$$

$$\begin{aligned}
\frac{۱}{\sigma^{۲(k)}} \mid others &\sim Gamma\left(f_{\cdot}, Q_{۳}(\boldsymbol{\phi}^{(k)})\right), \\
\pi(\boldsymbol{\beta}^{(k)} \mid others) &\propto \prod_{i=۱}^{n}\prod_{j=۱}^{r} \exp\left\{-\frac{\lambda}{۲}\left(Q_{۲}(\boldsymbol{\beta}^{(k)}) + g(\boldsymbol{\phi}^{(k)})\right)\right\}, \\
\pi(\boldsymbol{\phi}^{(k)} \mid others) &\propto \prod_{i=۱}^{n}\prod_{j=۱}^{r} \exp\left\{-\frac{\lambda}{۲} g(\boldsymbol{\phi}^{(k)}) - \frac{۱}{\sigma^{۲(k)}} Q_{۳}(\boldsymbol{\phi}^{(k)})\right\},
\end{aligned}$$

محاسبه نمود، که $Q_{۲}(\boldsymbol{\beta}^{(k)})$، $u^{(k)}$، $g(\boldsymbol{\phi}^{(k)})$ و $Q_{۳}(\boldsymbol{\phi}^{(k)})$ به‌ترتیب در روابط (۷.۳)، (۸.۳)، (۹.۳) و (۱۰.۳) معرفی شده‌اند. همان‌گونه که ملاحظه می‌شود این توزیع‌ها نمایش شناخته شده‌ای ندارند. به‌همین دلیل، برای نمونه‌گیری از آنها لازم است از الگوریتم متروپولیس ـ هاستینگز استفاده شود. از این‌رو، برای نمونه‌گیری از توزیع پسین توام پارامترها به تلفیقی از الگوریتم گیبز و متروپولیس ـ هاستینگز نیاز است.

۴.۳ محاسبه احتمال پسین مدل‌ها تحت مدل kام

محاسبه احتمال پسین مدل‌ها، رابطه (۴.۲)، مستلزم محاسبه درستنمایی حاشیه‌ای و احتمال پیشین مدل‌ها است. برای این منظور می‌توان درستنمایی حاشیه‌ای، رابطه (۵.۲)، را برای مدل پیشنهادی با داشتن نمونه‌ای از توزیع پسین توام پارامترها که از روش مونت کارلو زنجیر مارکفی به‌دست آمده به‌صورت

$$\begin{aligned}
p(\boldsymbol{y}|M_k) &= \int p(\boldsymbol{y}|\boldsymbol{\beta}^{(k)}, \lambda, \boldsymbol{\phi}^{(k)}, \sigma^{۲(k)}, M_k)\pi(\boldsymbol{\beta}^{(k)}, \lambda, \boldsymbol{\phi}^{(k)}, \sigma^{۲(k)}|M_k)\mathrm{d}(\boldsymbol{\beta}^{(k)}, \lambda, \boldsymbol{\phi}^{(k)}, \sigma^{۲(k)}) \\
&= \mathbb{E}_{\boldsymbol{\beta}^{(k)}, \lambda, \boldsymbol{\phi}^{(k)}, \sigma^{۲(k)}|M_k}\left(p(\boldsymbol{y}|\boldsymbol{\beta}^{(k)}, \lambda, \boldsymbol{\phi}^{(k)}, \sigma^{۲(k)}, M_k)\right) \\
&\simeq \frac{۱}{R}\sum_{r=۱}^{R} p(\boldsymbol{y}|(\boldsymbol{\beta}^{(k)}, \lambda, \boldsymbol{\phi}^{(k)}, \sigma^{۲(k)})^{(r)}, M_k) \qquad (۱۱.۳)
\end{aligned}$$

تقریب زد، که در آن $p(\boldsymbol{y}|\boldsymbol{\beta}^{(k)}, \lambda, \boldsymbol{\phi}^{(k)}, \sigma^{۲(k)}, M_k)$ تابع درستنمایی است.

برای محاسبه احتمال پسین مدل M_k لازم است احتمال پیشین درست بودن برای همه مدل‌های موجود در فضای مدل $\pi(M_k)$، $k = ۱, \ldots, T$، مشخص شوند. اگر هیچ اطلاع پیشینی در مورد مدل‌ها وجود نداشته باشد، توزیع مدل‌ها در فضای مدل، یکنواخت و به‌صورت

$$\pi(M_۱) = \cdots = \pi(M_T) = \frac{۱}{T},$$

در نظر گرفته می‌شود. در این صورت احتمال پسین مدل M_k به‌صورت

$$\begin{aligned}
\pi(M_k|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|M_k)\pi(M_k)}{\sum_{j=۱}^{T} p(\boldsymbol{y}|M_j)\pi(M_j)} \\
&= \frac{p(\boldsymbol{y}|M_k)}{\sum_{j=۱}^{T} p(\boldsymbol{y}|M_j)},
\end{aligned}$$

نوشته می‌شود.

۵.۳ پنجره اوکام

با توجه به رابطه (۱.۲)، ملاحظه می‌شود که محاسبه مجموع مورد نظر به دلیل تعداد زیاد جملات آن دشوار است. به منظور کاهش تعداد مدل‌هایی که در مجموع (۱.۲) مشارکت دارند، از راهکاری تحت عنوان پنجره اوکام که توسط

مادیگان و رفتری (۱۹۹۴) پیشنهاد شده است، استفاده می‌شود. این راهکار از دو اصل پیروی می‌کند. بنابر اصل اول مدل‌هایی که در مقایسه با پیش‌بینی صورت گرفته براساس بهترین مدل‌ها، دارای پیش‌بینی نامطلوب‌تری هستند، کنار گذاشته می‌شوند. با اجرای این اصل مدل‌هایی که در مجموعه

$$\mathcal{A}' = \left\{ M_k : \frac{\max_{M_l \in \mathcal{M}} \pi(M_l|\boldsymbol{y})}{\pi(M_k|\boldsymbol{y})} \leq c \right\}$$

قرار نمی‌گیرند، از رابطه (۱.۲) خارج می‌شوند. آستانه c توسط تحلیل‌گر تعیین می‌شود. در اصل دوم که اصل تیغ اوکام نیز نامیده‌اند، مدل‌های پیچیده‌ای که نسبت به همتایان ساده‌تر خود، کمتر از جانب داده‌ها حمایت می‌شوند نیز از مجموع کنار گذاشته می‌شوند. به عبارتی مدل‌هایی که متعلق به مجموعه

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A} \quad s.t \quad M_l \subset M_k, \frac{\pi(M_l|\boldsymbol{y})}{\pi(M_k|\boldsymbol{y})} > ۱ \right\}$$

هستند، از مجموع خارج می‌شوند. بنابراین رابطه (۱.۲)، به صورت

$$\pi(\Delta|\boldsymbol{y}) = \sum_{M_k \in \mathcal{A}} p(\Delta|M_k, \boldsymbol{y}) \pi(M_k|\boldsymbol{y})$$

بازنویسی می‌شود، که در آن $\mathcal{A} = \mathcal{A}' - \mathcal{B} \in \mathcal{M}$، مجموعه همه مدل‌هایی که در مجموع (۱.۲) قرار می‌گیرند.

۴ مطالعه شبیه‌سازی

در این بخش برای ارزیابی روش پیشنهادی از یک مطالعه شبیه‌سازی استفاده شده است. برای این منظور، یک متغیر پاسخ و هشت متغیر تبیینی در نظر گرفته شده است. در این صورت فضای مدل شامل $۲^۸ = ۲۵۶$ مدل است. مدل اشباع شده که دربرگیرنده همه متغیرهای تبیینی است، به‌صورت

$$y_i|\boldsymbol{x}_i \sim IG(\mu(\boldsymbol{x}_i), \lambda), \quad \mu^{-۱}(\boldsymbol{x}_i) = \beta_۱\boldsymbol{x}_۱ + \beta_۲\boldsymbol{x}_۲ + \cdots + \beta_۸\boldsymbol{x}_۸ + \phi_i, \quad i = ۱, \ldots, n, \qquad (۱.۴)$$

نوشته می‌شود. مقادیر متغیرهای تبیینی به‌صورت مستقل و از توزیع نرمال شبیه‌سازی شده‌اند. سپس با فرض $\lambda = ۲$ و

$$(\beta_۱, \beta_۲, \beta_۳, \beta_۴, \beta_۵, \beta_۶, \beta_۷, \beta_۸) = (۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸),$$

مقادیر متغیرهای پاسخ از رابطه (۱.۴) شبیه‌سازی شده‌اند. روش پنجره اوکام چهارده مدل را براساس مقایسه احتمال پسین مدل‌ها به عنوان بهترین مدل‌های موجود معرفی می‌کند. برای مقایسه کارایی مدل فضایی پیشنهادی با مدل کلاسیک همتای آن، از میانگین وزنی معیارهای اطلاع اکائیک، بیز و هانان ـ کوئین که به‌شکل

$$\begin{aligned} AIC &= -۲l(\hat{\boldsymbol{\beta}}) + ۲k, \\ BIC &= -۲l(\hat{\boldsymbol{\beta}}) + k\log(n), \\ HQC &= -۲l(\hat{\boldsymbol{\beta}}) + ۲k\log(\log n), \end{aligned} \qquad (۲.۴)$$

تعریف می‌شوند و با احتمال پسین متناظرشان وزن‌دار شده‌اند، استفاده شده است. در رابطه (۲.۴)، k تعداد پارامترهای مدل، n اندازه نمونه، $\hat{\boldsymbol{\beta}}$ برآورد بیز بردار ضرایب رگرسیونی تحت تابع زیان درجه دوم و $l(\cdot)$ تابع لگ‌درستنمایی را نشان می‌دهند. مقادیر به‌دست آمده برای میانگین وزنی معیارهای یاد شده برای مدل فضایی پیشنهادی و همتای کلاسیک آن در جدول ۱ نشان داده شده است. همان‌گونه که مشاهده می‌شود، مقادیر این میانگین‌ها برای روش میانگین‌گیری بیز مدل‌ها برای داده‌های فضایی از مقادیر متناظر برای میانگین‌گیری بیزی مدل‌ها کمتر هستند، که این موضوع نشان می‌دهد که مدل فضایی پیشنهادی به خوبی همبستگی‌های فضایی مشاهدات را در مدل‌بندی لحاظ می‌نماید.

جدول ۱: میانگین وزنی شاخص‌های نیکویی برازش برای مدل فضایی پیشنهادی و مدل کلاسیک.

میانگین‌گیری بیزی مدل‌ها	AIC	BIC	HQC
مدل فضایی	$-۳۳۶/۷۶۶$	$-۳۳۱/۶۹۵۶$	$-۳۳۴/۹۳۲۷$
مدل کلاسیک	$-۳۲۸/۸۸۹۲$	$-۳۲۳/۹۸۲۲$	$-۳۲۷/۱۱۵$

بحث و نتیجه‌گیری

روش‌های مرسوم مدل‌سازی آماری که برپایه مشاهدات مستقل بنا شده‌اند، در حالتی که مشاهدات دارای وابستگی فضایی هستند منجر به نتایج گمراه‌کننده می‌شوند. همچنین روش گزینش مدل با انتخاب یک مدل و مبنا قرار دادن آن در کاربردها، عدم حتمیت فرایند مدل‌سازی را به طور واقعی در نظر نمی‌گیرد. روش میانگین‌گیری بیزی مدل‌های رگرسیونی فضایی بدون از دست دادن اطلاعات مدل‌ها، از همه یا بهترین آنها بهره می‌گیرد و وابستگی فضایی داده‌ها را نیز لحاظ می‌کند.

مراجع

Banerjee, S. Carlin, B. P. and Gelfand, A. E. (2004), *hierarchical Modeling and Analysis for Spatial Data*, Boca Raton: Chapman Hall/CRC.

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

Diggle, P. J. (1993), Point Process Modeling in Environmental Epidemiology, In: Barnett, V. and Turkman, K. eds. *Statistics in Environment SPRUCE*, New York, Willey, Vol. 1.

Fairfield, S. H. (1938), An Empirical Law Describing Heteorogeneity in the Yields of Agricultural Crops, *Journal of Agricultural science*, **28**, 1-23.

Hoeting J. (1994), Accounting for Uncertainty in Linear Regression Models, Ph.D Disertation, Department of Statistics, University of Washington.

Hoeting, J., Raftery, A. and Madigan, D. (1996), A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regrassion Models, *Journal of Statistical Computation and Simulation*, **22**, 251-271.

Hoeting, J., Raftery, A. and Madigan, D. (1999), Basesian Simultaneous Variable Selection and Transformation Selection in Linear Regression Models, Technical Report 9905, Dept. Statistics, Colorado State Univ. Avilable at www. Colostate. edu.

Leamer, E. E. (1978), Specification Searches, Wiley, New York.

Madigan, D. and Raftery, A. (1994), Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window, *Journal of the American Statistical Association*, **89**, 1535-1546.

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

Meshkani, M. R. Fallah, A. and Kavousi, A. (2016), Bayesian analysis of covariance under inverse Gaussian model, *Journal of Applied Statistics*,**43**, 280-298.

Raftery, A. E. (1995), Bayesian Model Selection in Social Research (With Discussion), in Sociological Methodology 1995(P. V. Marsden, ed.), 111-195. Blakwell. Cambridge, MA.

Raftery, A. E., David, M. and Jenifer, A. H. (1993) Model Selection and Accounting for Model Uncertainty in Generalized Linear Models, Technical Report, **262**.

Raftery, A, E., David, M. and Jenifer, A. H. (1997) Bayesian model averaging in Linear regression Models, *Journal of the American Statistical Association*, **97**, 179-191.

Whitten, E. H. T. (1962), An reply to Chayes and Suzuki *Journal of Petrology*, **4**, 313-316.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

مدل‌بندی توأم بیزی داده‌های فضایی‌ـ‌زمانی با گمشدگی غیرقابل چشم‌پوشی

سمیرا زحمتکش[۱]، محسن محمدزاده
گروه آمار، دانشگاه تربیت مدرس

چکیده: اغلب داده‌های فضایی‌ـ‌زمانی در محیط‌های خارج از آزمایشگاه جمع‌آوری می‌شوند، به طوری که برخی عوامل تصادفی یا غیرتصادفی، بروز گمشدگی در داده‌ها را اجتناب ناپذیر می‌نمایند. از طرفی به دلیل وجود وابستگی بین مشاهدات، مقادیر گمشده‌ای که در همسایگی مکانی یا زمانی مشاهدات قرار دارند می‌توانند شامل اطلاعات مفیدی باشند که بازیابی این اطلاعات از دست رفته می‌تواند موجب افزایش دقت تحلیل‌ها شود. در این مقاله تحت فرض گمشدگی غیرقابل چشم‌پوشی، مدل‌بندی توأم فرایند اندازه‌گیری فضایی‌ـ‌زمانی و فرایند گمشدگی با استفاده از تکنیک مدل پارامتر مشترک پیشنهاد شده است، که در آن از طریق یک یا چند متغیر پنهان فضایی‌ـ‌زمانی مشترک، بین دو فرایند ارتباط برقرار می‌شود. استنباط‌ها از طریق رهیافت بیزی و روش تقریب لاپلاس آشیانه‌ای جمع‌بسته و بهره‌گیری از راهکار معادلات دیفرانسیل جزئی صورت گرفته است. سپس مطالعات شبیه‌سازی برای بررسی عملکرد مدل‌بندی توأم انجام شده است.

واژه‌های کلیدی: داده‌های فضایی‌ـ‌زمانی، داده‌های گمشده، مدل‌بندی توأم، مدل اتورگرسیو پویا، تقریب بیزی INLA.
کد موضوع‌بندی ریاضی (۲۰۱۰): 91B72، 91D25، 62F15.

۱ مقدمه

عموماً داده‌های فضایی‌ـ‌زمانی مانند داده‌های ثبت میزان و زمان بارندگی، دمای هوا یا آلودگی هوا در محیط‌های غیرآزمایشگاهی و خارج از کنترل محقق بدست می‌آیند. در مواردی وجود ابر، ناصاف بودن آسمان یا حتی بارش سنگین باران می‌توانند از جمله دلایل ایجاد گمشدگی باشند. از طرفی خرابی دستگاه‌ها و ابزارهای اندازه‌گیری نیز می‌توانند منجر به تولید مقادیر گمشده در داده‌ها شوند. رابین (۱۹۷۶) فرایندهای مختلف گمشدگی را از هم متمایز نمود و سازوکار گمشدگی را به سه دسته گمشدگی کاملاً تصادفی[۱] (MCAR) ، گمشدگی تصادفی[۲] (MAR) و گمشدگی غیرتصادفی[۳] (MNAR)

[1] Missing Completely At Random
[2] Missing At Random
[3] Missing Not At Random

[۱] نام و ایمیل ارائه دهنده مقاله: سمیرا زحمتکش، samira.zahmatkesh@modares.ac.ir

رده‌بندی کرد. وقتی فرایند گمشدگی مستقل از داده‌های مشاهده شده و مشاهده نشده باشد، در این صورت MCAR رخ می‌دهد،رده MAR وقتی وقوع می‌یابد که فرایند گمشدگی تنها به مقادیر مشاهده شده بستگی داشته باشد. مواردی که فرایند گمشدگی در هیچ یک از دو دسته MCAR و MAR نباشد گمشدگی MNAR رخ داده است، در این حالت گمشدگی هم به داده‌های مشاهده شده و هم به داده‌های مشاهده نشده بستگی دارد و اصطلاحاً گمشدگی غیرقابل چشم‌پوشی نامیده می‌شود.

از آنجا که از روی داده‌ها نمی‌توان به سازوکار گمشدگی پی برد، در مواردی که نتوان با قاطعیت ادعا کرد که گمشدگی تصادفی رخ داده است لازم است به منظور دستیابی به نتایج معتبر فرایند گمشدگی به همراه فرایند اندازه‌گیری مدل‌بندی و در تحلیل داده‌ها استفاده شود. در این زمینه روش‌های متعددی بر پایه سه چارچوب مختلف از تجزیه مدل توأم فرایند اندازه‌گیری و فرایند گمشدگی ارائه شده است که یکی از آن‌ها مدل پارامتر مشترک[۴] (SPM) است (دنیل و هوگان ، ۲۰۰۸). مدل پارامتر مشترک برای مدل‌بندی داده‌های فضایی گمشده توسط زحمتکش و محمدزاده (۲۰۱۸) مورد بررسی قرار گرفته است. در این مقاله مدل پارامتر مشترک را به داده‌های فضایی_زمانی تعمیم خواهیم داد، که در آن با در نظر گرفتن یک متغیر پنهان فضایی_زمانی مشترک برای توصیف ارتباط بین دو فرایند، تلاش می‌شود بخشی از اطلاعات از دست‌رفته را از طریق داده‌های مشاهده شده بازیابی نموده و برای تحلیل در داده‌های فضایی_زمانی مورد استفاده قرار دهیم. در این راستا از مدل‌های اتورگرسیو پویای مرتبه یک فضایی_زمانی برای مدل‌بندی فرایند اندازه‌گیری و فرایند گمشدگی استفاده می‌شود. همچنین در مدل‌بندی و استنباط پارامترهای مدل از روش بیزی تقریب لاپلاس آشیانه‌ای جمع بسته[۵] (INLA) که توسط رو و همکاران (۲۰۰۹) ارایه شده است، استفاده خواهد شد. برای این منظور توزیع توأم دو پاسخی (فرایند اندازه‌گیری و فرایند گمشدگی) در استفاده از INLA در نظر گرفته می‌شود. بعلاوه برای کاهش حجم محاسبات ناشی از ماتریس کواریانس چگال از روشی کارا و نوین از مدل‌بندی فضایی با استفاده از معادلات دیفرانسیل جزئی تصادفی [۶] (SPDE) استفاده می‌شود، که در آن میدان تصادفی گاوسی در فضایی مارکفی تقریب زده می‌شود و در عین حال امکان استفاده از INLA نیز وجود دارد (لینگرن و همکاران ، ۲۰۱۱).

ساختار مقاله به این‌صورت است که در بخش ۲ مدل‌های فضایی_زمانی معرفی می‌شوند در بخش ۳ مدل توأم فرایند اندازه‌گیری فضایی_زمانی و فرایند گمشدگی معرفی می‌شود. در بخش ۴ مروری کوتاه بر روش INLA و راهکار SPDE خواهیم داشت و در بخش ۵ مطالعه شبیه‌سازی برای بررسی عملکرد مدل توأم اجرا خواهد شد و در انتها به بحث و نتیجه‌گیری پرداخته می‌شود.

۲ مدل‌بندی داده‌های فضایی_زمانی در دامنه پیوسته فضایی

میدان تصادفی فضایی_زمانی $\{y(s,t),\ (s,t) \in D \subset R^2 \times R\}$ را در نظر بگیرید. فرض کنید مشاهدات در n موقعیت فضایی $s_1, \cdots, s_n$ و در T نقطه از زمان، $t = 1, \cdots, T$ جمع‌آوری شده‌اند و $y(s_i,t)$ تحققی از میدان تصادفی فضایی_زمانی در ایستگاه s_i و در زمان t باشد. مدل فضایی_زمانی به صورت

$$y(s_i,t) = \mathbf{z}(s_i,t)\beta + \xi(s_i,t) + \varepsilon(s_i,t) \qquad (۱.۲)$$

را در نظر بگیرید، که در آن $\mathbf{z}(s_i,t) = (z_1(s_i,t), \cdots, z_p(s_i,t))$ بردار مشاهدات p متغیر تبیینی در موقعیت s_i و زمان t ، $\beta = (\beta_1, \cdots, \beta_p)'$ بردار ضرایب، $\varepsilon(s_i,t)$ خطای اندازه‌گیری از توزیع $N(0, \sigma_\varepsilon^2)$ و $\xi(s_i,t)$ تحققی از یک

[4] Shared Parameter Model
[5] Integrated Nested Laplace Approximation
[6] Stochastic Partial Differential Equations

میدان تصادفی پنهان فضایی_زمانی است. به عنوان مثال در بررسی کیفیت هوا، این میدان تصادفی پنهان می‌تواند سطح واقعی مشاهده‌نشده از آلودگی هوا باشد. فرض می‌شود $\xi(s_i,t)$ در طی زمان بر اساس یک مدل پویای اتورگرسیو مرتبه اول، AR(1)، به صورت

$$\xi(s_i,t)=a\xi(s_i,t-1)+\omega(s_i,t) \qquad (۲.۲)$$

تغییر می‌کند (کملتی و همکاران ، ۲۰۱۱؛ کوچی و همکاران ، ۲۰۰۷؛ شاهو ، ۲۰۱۱) ، که در آن $|a|<1$ و $\xi(s_i,1)$ دارای توزیع $N(0,\frac{\sigma_\omega^2}{1-a^2})$ و $\omega(s_i,t)$ یک میدان تصادفی گاوسی مانا با میانگین صفر و تابع کواریانس فضایی_زمانی

$$Cov(\omega(s_i,t),\omega(s_j,t'))=\begin{cases}0 & t\neq t'\\ \sigma_\omega^2\rho(h) & t=t'\end{cases} \qquad (۳.۲)$$

است، که در آن تابع همبستگی فضایی $\rho(h)$ تنها از طریق $h=\|s_i-s_j\|\in R$ به موقعیت‌های فضایی s_i و s_j وابسته است. به این ترتیب فرض می‌شود که میدان تصادفی $\omega(s,t)$ مانای مرتبه دوم و همسانگرد است (محمدزاده ، ۱۳۹۸). تابع همبستگی فضایی $\rho(h)$ عموماً به دلیل انعطاف‌پذیری، تابع مترن به صورت

$$\rho(h)=\frac{1}{\Gamma(\nu)2^{\nu-1}}(\kappa h)^\nu \mathrm{K}_\nu(\kappa h) \qquad (۴.۲)$$

در نظر گرفته می‌شود، که در آن K_ν تابع بسل تعدیل‌یافته نوع دوم و از مرتبه $\nu>0$ است (آبرامویتز و استگان ، ۱۹۷۲)، پارامتر ν بیانگر درجه همواری فرایند و $\kappa>0$ پارامتر مقیاس مرتبط با دامنه r است. دامنه فاصله‌ای است که در آن همبستگی فضایی کوچک و نزدیک به صفر می‌شود و رابطه $r=\frac{\sqrt{8\nu}}{\kappa}$ برای آن تعریف می‌شود (لینگرن و همکاران ، ۲۰۱۱). فرض کنید $\mathbf{y}_t=(y(s_1,t),...,y(s_n,t))$ تحققی از فرایند اندازه‌گیری در زمان t و n موقعیت فضایی باشد، مدل‌های (۱.۲) و (۲.۲) را می‌توان به صورت

$$\mathbf{y}_t=\mathbf{z}_t\beta+\xi_t+\varepsilon_t \qquad \varepsilon_t\sim N(0,\sigma^2 I_n) \qquad (۵.۲)$$

$$\xi_t=a\xi_{t-1}+\omega_t \qquad \omega_t\sim N(0,R=\sigma_\omega^2\tilde{R}) \qquad (۶.۲)$$

نوشت، که در آن I_n ماتریس همانی از بعد n است و $\mathbf{z}_t=(\mathbf{z}(s_1,t)',\cdots,\mathbf{z}(s_n,t)')'$ و $\xi_t=(\xi(s_1,t),\cdots,\xi(s_n,t))'$. ξ_1 یک فرایند $AR(1)$ از توزیع مانای $N(0,\frac{1}{1-a^2}R)$، و $\tilde{R}$ ماتریس همبستگی از بعد n با عناصر $\rho(\|s_i-s_j\|)$ است. روابط (۵.۲) و (۶.۲) مدل‌های سلسله مراتبی را ارایه می‌دهند که به دلیل انعطاف‌پذیری آن‌ها در مدل‌بندی همزمان اثرات متغیرهای تبیینی و همبستگی‌های فضایی_زمانی، مکرراً در مطالعات بررسی کیفیت هوا استفاده شده‌اند.

فرض کنید $\mathbf{y}=(\mathbf{y}_1,\cdots,\mathbf{y}_T)$ بردار مشاهدات متغیر پاسخ و $\xi=(\xi_1,\cdots,\xi_T)$ تحققی از میدان تصادفی پنهان فضایی_زمانی در n موقعیت فضایی و T لحظه زمانی است. قرار می‌دهیم $\theta_1=(a,\sigma_\omega^2,r)$ و $\theta_2=(\beta,\sigma_\varepsilon^2)$، بنابراین $\theta=(\theta_1,\theta_2)$ بردار پارامترهای مدل است که باید برآورد شود. با درنظر گرفتن توزیع پیشینی مستقل $\pi(\theta)=\Pi_{i=1}^{\dim(\theta)}\pi(\theta_i)$ برای پارامترهای مدل و با فرض استقلال شرطی $\mathbf{y}$ روی ξ در زمان و اینکه میدان تصادفی پنهان از یک مدل پویای زمانی مارکفی پیروی می‌کند توزیع پسینی توأم پارامترهای مدل به صورت

$$\begin{aligned}\pi(\theta,\xi|\mathbf{y}) &\propto \pi(\mathbf{y}|\xi,\theta)\pi(\xi|\theta_1)\pi(\theta)\\ &\propto \Big(\prod_{t=1}^{T}\pi(\mathbf{y}_t|\xi_t,\theta)\Big)\Big(\pi(\xi_1|\theta_1)\prod_{t=2}^{T}\pi(\xi_t|\xi_{t-1},\theta_1)\Big)\pi(\theta)\end{aligned} \qquad (۷.۲)$$

حاصل می‌شود، که با توجه به (۱.۲) و (۲.۲) داریم

$$\begin{aligned}\pi(\theta,\xi|\mathbf{y}) &\propto (\sigma_\varepsilon^{۲})^{-\frac{nT}{۲}}\exp\big(-\frac{۱}{۲\sigma_\varepsilon^{۲}}\sum_{t=۱}^{T}(\mathbf{y}_t-\mathbf{z}_t\beta-\xi_t)'(\mathbf{y}_t-\mathbf{z}_t\beta-\xi_t)\big)\\ &\times (\frac{\sigma_\omega^{۲}}{۱-a^{۲}})^{-\frac{n}{۲}}\big|\tilde{R}\big|^{-\frac{۱}{۲}}\exp\big(-\frac{۱-a^{۲}}{۲\sigma_\omega^{۲}}\xi_۱'\tilde{R}^{-۱}\xi_۱\big)\\ &\times (\sigma_\omega^{۲})^{-\frac{n(T-۱)}{۲}}\big|\tilde{R}\big|^{-\frac{(T-۱)}{۲}}\exp\big(-\frac{۱}{۲\sigma_\omega^{۲}}\sum_{t=۲}^{T}(\xi_t-a\xi_{t-۱})'\tilde{R}^{-۱}(\xi_t-a\xi_{t-۱})\big)\\ &\times \prod_{i=۱}^{\dim(\theta)}\pi(\theta_i) \qquad (۸.۲)\end{aligned}$$

که در آن $\big|\tilde{R}\big|$ دترمینان ماتریس همبستگی تنک $\tilde{R}$ است.

۳ مدل‌بندی توأم بیزی فرایندهای گمشدگی و اندازه‌گیری تحت فرض MNAR

در داده‌های فضایی‌ـ زمانی این گمان می‌تواند وجود داشته باشد که وجود مقادیر گمشده تحت تأثیر عواملی غیرتصادفی باشند، بنابراین تحت فرض MNAR، یعنی با فرض غیرقابل چشم‌پوشی بودن گمشدگی، می‌توان مدلی توأم از فرایند اندازه‌گیری (متغیر پاسخ) و فرایندگمشدگی ساخت. در اینجا SPM برای حالت فضایی‌ـ زمانی بسط داده می‌شود.

فرض کنید فرایند گمشدگی به صورت $\mathbf{m}=(\mathbf{m}_۱,\cdots,\mathbf{m}_T)$، است، که در آن $\mathbf{m}_t=(m(s_۱,t),\cdots,m(s_d,t))$، $t=۱,\cdots,T$ از طریق تابع نشانگر

$$m(s_i,t)=\begin{cases}۱ & y(s_i,t)\text{ گمشده است}\\ ۰ & y(s_i,t)\text{ مشاهده شده است}\end{cases}$$

تعریف می‌شود، که دارای توزیع برنولی $Ber(p)$ با احتمال گمشدگی p است. در این‌صورت احتمال گمشدگی برای $y(s_i,t)$ به صورت $P(m(s_i,t)=۱|\varphi)=p(s_i,t)$ است، که در آن φ بردار پارامترهای مدل فرایند گمشدگی است. احتمال گمشدگی $p(s_i,t)$ از طریق یک رگرسیون لوژستیک به صورت

$$\text{logit}(p(s_i,t))=\alpha_۰+\alpha_۱\xi(s_i,t)+\mathbf{H}(s_i,t)\gamma \qquad (۱.۳)$$

مدل‌بندی می‌شود، که در آن $\alpha_۰$ عرض از مبدأ و بیانگر نسبت ثابتی از گمشدگی است، $\xi(s_i,t)$ تحققی از یک میدان تصادفی پنهان فضایی‌ـ زمانی در موقعیت فضایی s_i و زمان t است که در مدل متغیر پاسخ نیز حضور دارد و در واقع بین $\mathbf{y}$ و $\mathbf{m}$ ارتباط برقرار می‌کند و از طریق ضریب $\alpha_۱$ شدت و چگونگی این ارتباط توصیف می‌شود. این میدان پنهان فضایی‌ـ زمانی مشابه (۲.۲) در طی زمان بر اساس اتورگرسیو پویای مرتبه اول تغییر می‌کند. بردار q بعدی $\mathbf{H}(s_i,t)=(H_۱(s_i,t),\cdots,H_q(s_i,t))$ متغیرهای تبیینی در ارتباط با احتمال گمشدگی است و $\gamma=(\gamma_۱,...,\gamma_۲)$ بردار ضرایب مربوطه است. به این ترتیب $\varphi=(\alpha_۰,\alpha_۱,\gamma)$. اکنون با وارد کردن مدل گمشدگی در رابطه (۷.۲) به جای توزیع $\pi(\mathbf{y}|\xi,\theta)$ توزیع توأم $\pi(\mathbf{y},\mathbf{m}|\xi,\varphi,\theta)$ جایگزین خواهد شد. در مدل‌بندی توأم فرایند اندازه‌گیری و فرایند گمشدگی، ξ به عنوان میدان تصادفی پنهان فضایی‌ـ زمانی مشترک در نظر گرفته می‌شود و با فرض استقلال شرطی آن دو روی ξ توزیع توأم آن‌ها بر اساس SPM به صورت

$$\pi(\mathbf{y},\mathbf{m}|\xi,\varphi,\theta)=\pi(\mathbf{y}|\xi,\theta)\pi(\mathbf{m}|\xi,\varphi) \qquad (۲.۳)$$

تجزیه می‌شود. تابع درستنمایی داده‌های کامل به صورت

$$\begin{aligned} L(\xi,\varphi,\theta) &= \prod_{t=1}^{T} \pi(\mathbf{y}_t,\mathbf{m}_t|\xi_t,\varphi,\theta) \\ &= \prod_{t=1}^{T} \pi(\mathbf{y}_t|\xi_t,\theta) \prod_{t=1}^{T} \pi(\mathbf{m}_t|\xi_t,\varphi) \end{aligned}$$

خواهد بود. با در نظر گرفتن $\eta = (\varphi,\theta)$ به عنوان بردار پارامترهای مدل توأم، رابطه (۸.۲) به سادگی بازنویسی می‌شود که فرم بسته‌ای ندارد. برای تولید نمونه از توزیع‌های پسینی حاشیه‌ای از روش تقریب بیزی INLA استفاده می‌شود که در بخش بعد معرفی خواهد شد.

۴ تحلیل بیزی با استفاده از INLA و SPDE

مدل سلسه‌مراتبی تعریف شده در (۱.۲) و (۲.۲) عضوی از رده مدل‌های گاوسی پنهان است که از طریق INLA قابل برآورد است که روشی جایگزین برای روش‌های MCMC به منظور دستیابی سریع‌تر به تقریبی از توزیع‌های حاشیه‌ای پسینی برای متغیرهای پنهان و پارامترها است. بر اساس رو و همکاران (۲۰۰۹)، به میدان پنهان مدل توزیع GMRF تخصیص می‌دهیم. یک GMRF یک میدان تصادفی است که وابستگی فضایی داده‌ها را در دامنه‌های فضایی مختلف مدل‌بندی می‌کند (رو و هلد ، ۲۰۰۵). فرض کنید $\mathbf{x} = (x_1,...,x_n)$ از توزیع $\mathbf{x} \sim \mathbf{N}(\mu,\mathbf{Q^{-1}})$ یک GMRF با بعد n باشد، که در آن Q ماتریس معین مثبت و متقارن دقت، یعنی وارون ماتریس کواریانس است. به این ترتیب تابع چگالی $\mathbf{x}$ به صورت

$$\pi(\mathbf{x}) = (\mathbf{2}\pi)^{(-\frac{\mathbf{n}}{\mathbf{2}})}|\mathbf{Q}|^{\frac{1}{2}} \exp\left(-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{x}-\mu)'\mathbf{Q}(\mathbf{x}-\mu)\right)$$

است،که در آن Q با توجه به خاصیت مارکفی $\mathbf{x}$، از طریق ساختار همسایگی مشاهدات تعیین می‌شود. در واقع مزیت استنباط بر اساس GMRF ویژگی محاسباتی خوب و امکان بهبود استنباط بیزی از طریق INLA است (رو و همکاران ، ۲۰۰۹). یکی از روش‌های تبدیل GRF به GMRF استفاده از راهکار SPDE است که هدف آن پیدا کردن GMRF با یک همسایگی و ماتریس دقت تنک Q است به طوری‌که مشکل n بزرگ (بانرژی و همکاران ، ۲۰۰۸) را که هنگام کار با ماتریس کواریانس رخ می‌دهد برطرف می‌نماید. اساساً راهکار SPDE از یک ارایه عنصر متناهی به منظور تعریف GRF به عنوان ترکیب خطی از توابع پایه در یک توری مثلثی از دامنه D استفاده می‌کند. این کار شامل افراز D به مجموعه‌ای از مثلث‌های مجزا است که حداکثر در یک ضلع یا زاویه مشترک هستند. برای تشکیل توری مثلثی ابتدا هر یک از n ایستگاه به عنوان یک رأس مثلث در نظر گرفته می‌شوند، سپس سایر رئوس اضافه شده و مثلث‌ها ساخته می‌شوند. فرض کنید $\mathbf{x}(s) = \{x(s), s \in D \subseteq R^2\}$ یک میدان تصادفی گاوسی مانای همسانگرد با تابع کواریانس مترن (۴.۲) باشد، با فرض معلوم بودن توری مثلثی، یکSPDE ترکیب خطی از توابع پایه به صورت

$$\mathbf{x}(s) = \sum_{\ell=1}^{m} \psi_\ell(s) w_\ell \qquad (۱.۴)$$

است، که در آن m تعداد رئوس مثلث‌ها در توری، $\psi_\ell(s)$ توابع پایه و w_ℓ وزن‌هایی با توزیع گاوسی هستند. آنچه که در راهکار SPDE مورد هدف قرار می‌گیرد ارایه عنصر متناهی (۱.۴) است که ارتباطی بین GRF و GMRF ایجاد می‌کند و از طریق وزن‌های گاوسی w_ℓ یک ساختار مارکفی ساخته می‌شود (لینگرن و همکاران ، ۲۰۱۱). اکنون در مدل فضایی_زمانی مفروض، در استفاده از راهکار SPDE در هر نقطه از زمان، $t = 1,...,T$ میدان مترن ω_t از طریق یک GMRF به صورت $\tilde{\omega}_t \sim N(0, Q_s^{-1})$ ارایه می‌شود، که در آن Q_s ماتریس دقت است. ماتریس Q_s با توجه به (۳.۲) مستقل زمانی است و

بعد آن برابر با تعداد گره‌های توری مثلثی است. بنابراین معادله (۶.۲) برای $t = ۱, ..., T$ به صورت

$$\xi_t = a\xi_{t-۱} + \tilde{\omega}_t \qquad \tilde{\omega}_t \sim N(۰, Q_s^{-۱}) \tag{۲.۴}$$

نوشته می‌شود، که در آن $\xi_۱ \sim N(۰, \frac{Q_s^{-۱}}{۱-a^۲})$. به این ترتیب توزیع توأم GMRF بعدی nT بعدی $\xi = (\xi_۱', ..., \xi_T')$ به صورت $\xi \sim N(۰, Q^{-۱})$ است، که در آن $Q = Q_T \otimes Q_S$ و Q_T ماتریس دقت T بعدی فرایند اتورگرسیو زمانی مرتبه ۱ در (۳.۲) است. بعلاوه معادله (۵.۲) به صورت

$$\mathbf{y}_t = \mathbf{z}_t\beta + B\xi_t + \varepsilon_t \qquad \varepsilon_t \sim N(۰, \sigma_\varepsilon^۲ I_n) \tag{۳.۴}$$

نوشته می‌شود که در آن ماتریس $n \times m$ بعدی B مقدار ξ_t را برای مشاهده y_t انتخاب می‌کند و یک ماتریس تنک است به طوری‌که

$$y(s_i, t) = z(s_i, t)\beta + \sum_{j=۱}^{m} B_{ij}\xi_t + \varepsilon_t(s_i, t) \tag{۴.۴}$$

که در آن $B_{ij} = ۱$ اگر رأس j ام در مکان s_i باشد و در غیر اینصورت $B_{ij} = ۰$.

فرض کنید $\mathbf{x} = \{\xi, \beta\}$ بردار پارامترهای مدل با پیشین‌های مستقل هستند و β توزیع پیشین گاوسی مبهم با ماتریس دقت معلوم و ξ دارای توزیع GMRF است. بنابراین چگالی $\pi(\mathbf{x}|\theta)$ گاوسی با میانگین صفر و ماتریس دقت $Q(\theta_۱)$ با ابرپارامتر $\theta_۱ = (\sigma_\omega^۲, a, \kappa)$ است. اگر $\theta_۲ = \sigma_\varepsilon^۲$ و $\theta = (\theta_۱, \theta_۲)$، آنگاه مشاهدات $\mathbf{y} = \{\mathbf{y}_t\}$ دارای توزیع نرمال و مستقل شرطی روی $\mathbf{x}$ و θ هستند. در این‌صورت توزیع پسینی توأم به صورت

$$\pi(\mathbf{x}, \theta|\mathbf{y}) = \pi(\theta)\pi(\mathbf{x}|\theta) \prod_{t=۱}^{T} \pi(\mathbf{y}_t|\mathbf{x}, \theta) \tag{۵.۴}$$

است، که در آن $\pi(\mathbf{y}_t|\mathbf{x}, \theta)$ توزیع شرطی مشاهدات در زمان t است و بنابر (۳.۴) توزیع نرمال $N(\mathbf{z}_t\beta + B\xi_t, \sigma_\varepsilon^۲ I_n)$ است. در اینجا لازم است توزیع‌های پسینی حاشیه‌ای میدان پنهان و ابرپارامترها به صورت

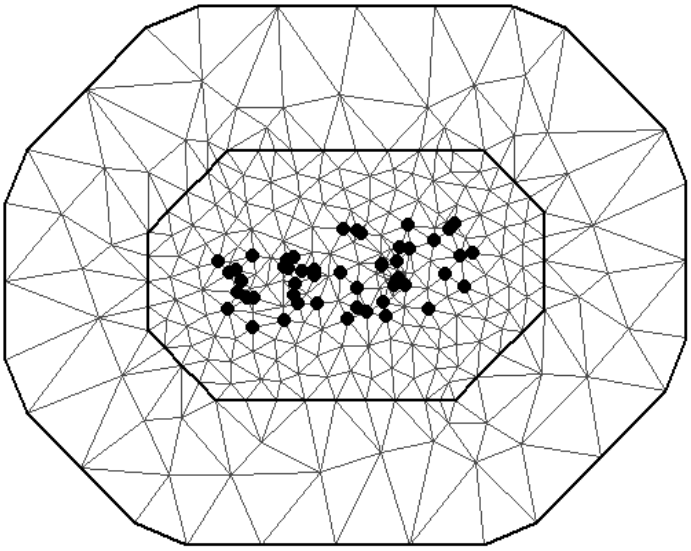
$$\pi(x_i|\mathbf{y}) = \int \pi(x_i|\theta, \mathbf{y})\pi(\theta|\mathbf{y})d\theta, \qquad i = (۱, ..., T + p)$$
$$\pi(\theta_j|\mathbf{y}) = \int \pi(\theta|\mathbf{y})d\theta_{-j}, \qquad j = ۱, ..., ۴$$

محاسبه شوند. تقریب INLA از روش‌های تقریب گاوسی و لاپلاس و انتگرال‌های عددی برای محاسبه توزیع‌های پسینی حاشیه‌ای استفاده می‌کند (رو و همکاران ، ۲۰۰۹). در مدل‌بندی توأم فرایند گمشدگی و فرایند اندازه‌گیری فضایی_زمانی، از آن‌جا که میدان تصادفی پنهان فضایی_زمانی ξ به عنوان عامل مشترک در مدل (۱.۳) در نظر گرفته شده است، راهکار SPDE که منجر به مدل‌های سلسله مراتبی (۲.۴) و (۳.۴) شد را به مدل (۱.۳) تعمیم داده و استنباط بیزی را اجرا خواهیم نمود.

۵ شبیه سازی مدل توأم فرایندهای گمشدگی و اندازه‌گیری

در این بخش از طریق شبیه‌سازی داده‌های فضایی_زمانی با فرض وجود MNAR در داده‌ها، عملکرد مدل توأم فرایند گمشدگی و فرایند اندازه‌گیری فضایی_زمانی را بررسی نموده و دقت دو مدل توأم و MAR مورد ارزیابی و مقایسه قرار می‌گیرد.

برای بررسی عملکرد مدل توأم دو فرایند گمشدگی و اندازه‌گیری ابتدا $n = ۵۰$ موقعیت فضایی به عنوان ایستگاه‌های اندازه‌گیری به طور نامنظم در دامنه فضایی مورد نظر تولید شده است. موقعیت قرارگیری این نقاط در مشبکه مثلثی ساخته شده برای راهکار SPDE در شکل ۱ نمایش داده شده است. فرایند تولید در n موقعیت فضایی و $T = ۵$ زمان انجام شده



شکل ۱: توری مثلثی ساخته شده برای موقعیت‌های فضایی مورد استفاده در راهکار SPDE

است، که در نهایت مجموعه داده شامل $N = n \times T$ مشاهده خواهد بود. مشاهدات فرایند اندازه‌گیری (متغیر پاسخ)، در موقعیت فضایی s_i و زمان t، $y(s_i, t)$، بر اساس مدل

$$\begin{aligned} y(s_i,t) &= \beta_{۰} + \xi(s_i,t) + \varepsilon(s_i,t) \\ \xi(s_i,t) &= a\xi(s_i,t-۱) - \omega(s_i,t) \end{aligned} \qquad (۱.۵)$$

تولید شده‌اند، که در آن $\varepsilon \sim N(۰, \sigma_{\varepsilon}^{۲})$ و $\omega(s_i,t)$ یک میدان تصادفی گاوسی مانا با تابع کواریانس مترن است. همچنین فرایند گمشدگی $m(s_i,t)$ از توزیع برنولی با احتمال گمشدگی $\pi(s_i,t)$ تولید شده است که در آن احتمال گمشدگی واحد i ام در زمان t به صورت

$$\mathrm{logit}(\pi(\mathrm{s_i,t})) = \alpha_0 + \alpha_1 \xi(\mathrm{s_i,t})$$

مدل‌بندی شده است، ξ به عنوان عامل مشترک با مدل پاسخ بر اساس مدل پارامتر مشترک (SPM) در نظر گرفته می‌شود. به این ترتیب $\theta = (\beta_{۰}, \sigma_{\varepsilon}, r, \sigma, a, \alpha_{۰}, \alpha_{۱})$ بردار پارامترهای مدل است که باید برآورد شوند. اجرا در $m = ۲۰۰$ تکرار انجام شده است و در نهایت از نتایج برآورد پارامترها میانگین گرفته شده است. با توجه به اینکه پارامتر $\alpha_{۱}$ شدت وابستگی فرایند اندازه‌گیری و فرایند پاسخ را در مدل‌بندی توأم بیان می‌کند، به منظور بررسی رفتار عملکرد مدل دو مقدار $\alpha_{۱} = ۰/۲, ۰/۸$ در نظر گرفته شده است، همچنین به دلیل تأثیری که احتمالاً میدان تصادفی پنهان بر روی الگوی گمشدگی دارد، برای انحراف استاندارد حاشیه‌ای میدان پنهان فضایی دو مقدار $\sigma = ۴, ۸$ در نظر گرفته شده است. شبیه‌سازی برای دو مجموعه از مقادیر اولیه پارامترهای مدل اجرا شده است. بردار پارامترها در اجرای اول و دوم به ترتیب برابر با

$$\theta_{۱} = (\beta_{۰}, \sigma_{\varepsilon}, r, \sigma, a, \alpha_{۰}, \alpha_{۱}) = (-۲, ۱, ۲۵, ۴, ۰/۶, ۰/۵, ۰/۲)$$
$$\theta_{۲} = (\beta_{۰}, \sigma_{\varepsilon}, r, \sigma, a, \alpha_{۰}, \alpha_{۱}) = (-۲, ۱, ۲۵, ۸, ۰/۸, ۰/۵, ۰/۸)$$

قرار داده شده‌اند. همان‌طور که ملاحظه می‌شود در اجرای دوم مقادیر اولیه برای پارامترهای α_1 و σ نسبت به اجرای اول افزایش یافته است. لازم به ذکر است بعد از بررسی نتایج در حضور پارامتر α_0 و مشاهده عدم معنی‌داری آن، از مدل کنار گذاشته شده است. برای ارزیابی عملکرد مدل توأم، مدل دیگری تحت عنوان مدل MAR درنظر گرفته شده است. در این مدل فرض می‌شود مکانیزم گمشدگی تصادفی است و با استفاده از یکی از روش‌های رایج، با جانهی مقادیر گمشده، داده‌ها بازسازی شده و با صرف‌نظر کردن از مدل فرایند گمشدگی، مدل اتورگرسیو پویای مرتبه ۱ (۱.۵) به داده‌ها برازش داده شده است. به عنوان نمونه، برآورد پارامترهای دو مدل توأم و MAR برای مقادیر اولیه θ_2 در جدول ۱ گزارش شده است. همانطور که ملاحظه می‌شود انحراف استاندارد و میزان اریبی در برآورد پارامترها بر اساس مدل توأم نسبت به مدل MAR کمتر است.

جدول ۱: برآوردهای پسینی پارامترها برای دو مدل توأم و MAR

		مدل						
		توأم				MAR		
پارامتر	مقدار اولیه	برآورد	انحراف استاندارد	۰/۰۲۵	۰/۹۷۵	برآورد	انحراف استاندارد	۰/۰۲۵	۹۷۵.۰
β_0	−۲	−۲/۰۹۹	۰/۸۱۰	−۴/۴۴۱	−۱/۲۵۷	−۱۲/۲۰۱	۱/۶۳۷	−۱۵/۴۰۸	−۸/۹۲۵
σ_ε^2	۱	۱/۲۰۲	$۱/۶e-۰۷$	۱/۱۸۲	۱/۲۴۱	۰/۳۰۲۱	۰/۰۰۵	۰/۰۰۱	۰/۵۱۷
r	۲۵	۲۵/۲۳۲	۱/۳۹۰	۲۲/۵۴۱	۲۷/۹۸۸	۲۴/۰۳۷	۳/۹۰۵	۱۷/۰۷۶	۳۲/۴۰۴
σ	۸	۷/۳۶۵	۰/۶۲۸	۶/۷۲۷	۹/۶۶۴	۱۰/۰۱۴	۱/۰۴۴	۸/۲۸۹	۱۲/۳۸۹
a	۰/۸	۰/۸۱۳	۰/۰۰۱	۰/۷۹۱	۰/۸۷۸	۰/۶۹۱	۰/۰۷۰	۰/۵۱۶	۰/۸۲۲
α_1	۰/۸	۰/۸۲۸	۰/۰۰۲	۰/۴۷۸	۴/۲۷۲	–	–	–	–

برای مقایسه عملکرد دو مدل توأم و MAR در دو اجرای متفاوت درصد پوشش دهی بازه باورمندی در m تکرار برای پارامترهای مدل بدست آمده است. در اجرای ۱ هر دو مدل عملکرد یکسانی دارند، اما در اجرای ۲ که دو پارامتر α_1 و σ افزایش داده شده‌اند میزان پوشش دهی بازه باورمندی در مدل MAR برای دو پارامتر β_0 و a به ترتیب ۴۹ و ۹۵ درصد بدست آمده است. این در حالی است که برای مدل توأم این معیار برای تمامی پارامترها ۱۰۰ درصد بدست آمده است. برای ارزیابی عملکرد پیشگویی این دو مدل، زمان $t = ۲$ انتخاب شده است و با درنظر گرفتن $n_p = ۱۰$ موقعیت فضایی به عنوان موقعیت‌های هدف برای پیشگویی، مشاهدات مربوط به این موقعیت‌ها به عنوان مجموعه آزمایشی از مطالعه کنار گذاشته شده و پیشگویی برای دو مجموعه از مقادیر اولیه انجام شده است. سپس معیار جذر میانگین توان‌های دوم خطای پیشگویی‌ها[۷] (RMSE) محاسبه شده است. در اجرای ۱، میزان RMSE برای مدل توأم و مدل MAR به ترتیب برابر با ۴/۹۳۳ و ۴/۳۵۷ بدست آمده و در اجرای ۲، این معیار برای مدل توأم و MAR به ترتیب برابر با ۵/۹۶۶ و ۹/۷۳۰ حاصل شده است. همانطور که ملاحظه می‌شود در اجرای ۱، دقت پیشگویی برای دو مدل تقریباً نزدیک به هم و حتی مدل MAR عملکرد نسبتاً بهتری داشته است، ولی برتری عملکرد مدل توأم در اجرای ۲ آشکار می‌شود که در آن پارامتر همبستگی دو فرایند اندازه‌گیری و گمشدگی بزرگتر و تغییرپذیری میدان تصادفی پنهان فضایی بیشتر است.

بحث و نتیجه‌گیری

در این مقاله، تحت فرض گمشدگی MNAR به مدل‌بندی داده‌های فضایی_زمانی پرداخته شد. با توجه به اینکه در برخی مطالعات فضایی_زمانی گمشدگی می‌تواند متأثر از عوامل پنهان فضایی_زمانی باشد، نمی‌توان به سادگی تحت فرض

[7] Root Mean Square Error

MAR به نتایج قابل اعتمادی دست پیدا کرد. نتایج مطالعات شبیه‌سازی دراین مقاله نشان می‌دهد که در چنین مواردی می‌توان با بکارگیری تکنیک‌هایی نظیر مدل پارامتر مشترک، دو فرایند گمشدگی و اندازه‌گیری را توأماً مدل‌بندی کرد و ارتباط عوامل پنهان فضایی‌ـ زمانی و فرایند گمشدگی را کشف و بررسی کرد و به نتایج معتبر و قابل اعتمادتری دست یافت.

مراجع

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

Abramowitz, M. and Stegun, I. (1972), *Handbook of Mathematical Functions*, Courier Dover Publications.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H., (2008), Gaussian Predictive Process Models for Large Spatial Datasets, *Journal of the Royal Statistical Society,* **70**, 825-848.

Cameletti, M., Ignaccolo, R. and Bande, S., (2011), Comparing Spatio-temporal Models for Particulate Matter in Piemonte, *Environmetrics*, DOI: 10.1002/env.1139.

Cocchi, D., Greco, F. and Trivisano, C., (2007), Hierarchical space-time Modelling of PM10 Pollution, *Atmospheric environment*, **41**, 532–542.

Daniels, M. J. and Hogan, J. W. (2008), *Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis*, Chapman and Hall.

Lindgren, F., Rue, H., and Lindstr, O., (2011), An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach, *Journal of the Royal Statistical Society*, **73**, 423-498.

Rubin, D. B. (1976), Inference and Missing Data. *Biometrika*, **63**, 581-92.

Rue, H. and Held, L., (2005), *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall.

Rue, H., Martino, S. and N. Chopin, (2009), Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations, *Journal of Statistical Planning and Inference*, **137** , 3177-3199.

Sahu, S. (2011), *Hierarchical Bayesian Models for Space-time Air Pollution Data, Handbook of Statistics* , vol 30, Elsevier Publishers, Holland.

Zahmatkesh, S. and Mohammadzadeh, M. (2018), Bayesian Prediction of Spatial Data with Non-ignorable Missingness, Submitted.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

اثر الگوی جغرافیایی بر رضایت مشتری

فاطمه زنجیران[۱۱]، کیومرث مترجم[۲]
[۱]گروه آمار، دانشگاه علم و فرهنگ
[۲]گروه آمار، دانشگاه تربیت مدرس

چکیده:

در این مطالعه، به معرفی مدلی بر مبنای اقتصادسنجی فضایی پرداخته شده است که یک شرکت را قادر می‌سازد الگوهای منطقه‌ای را در داده‌های رضایت‌مندی با استفاده از رگرسیون وزنی جغرافیایی (GWR) شناسایی کند. برای پیاده‌سازی این مدل، از داده‌های مربوط به رضایت‌مندی مشتریان از کیفیت خودرو در ایالات متحده آمریکا استفاده شد. بر مبنای داده‌ها که براساس کدپستی مشتریان است، مختصات طول و عرض جغرافیایی برای هر کدپستی به دست آورده شده است. سپس، باتوجه به فاصله اقلیدسی بین هر کدپستی از مدل GWR استفاده شد. همبستگی فضایی نیز با استفاده از آزمون موران و گری مورد بررسی قرار گرفت و نتایج حاکی از مثبت بودن همبستگی فضایی برای متغیرهای مورد پژوهش بود. سپس، عملکرد مدل براساس توانایی مدل در پیش‌بینی رضایت‌مندی کلی و پایایی برآورد پارامترهای مدل فضایی مورد ارزیابی قرار گرفت.

واژه‌های کلیدی: آمارفضایی، رضایت‌مندی مشتری، رگرسیون وزن جغرافیایی .
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J05، 62J99، 62P20.

۱ مقدمه

پراکندگی مشتریان شرکت‌ها در مناطق مختلف جغرافیایی سبب دشواری مدیریت کیفیت خدمات می‌شود؛ زیرا همبستگی فضایی بین مناطق موجب ایجاد تفاوت در رضایت‌مندی مشتریان می‌شود. در نظر گرفتن اثرات فضایی در رضایت داده‌ها می‌تواند کیفیت خدمات را افزایش دهد؛ جایی که رضایت‌مندی از کیفیت خدمات پایین است اما مشتریان به دنبال بهبود کیفیت خدمات دریافتی هستند. معمولا در مناطقی که کیفیت خدمات بسیار بالا است این مطالبه‌گری در مشتریان کمرنگ‌تر

[۱] نام و ایمیل ارائه دهنده مقاله: فاطمه زنجیران، n.zanjiran@gmail.com

است. رضایت مشتری برای موفقیت درازمدت شرکت ضروری است(آنتونی و همکاران، ۱۹۹۵) و (والاریه و همکاران، ۲۰۰۰). با این حال، به ویژه در شرکت‌هایی که خدمات را با استفاده از شبکه‌های پراکنده منطقه‌ای از رسانه‌ها (به عنوان مثال، نمایندگی‌های خودرو، شاخه‌های بانک) ارائه می‌دهند، مدیران با اجرای یک استراتژی رضایت مشتری مواجه می‌شوند. برای چنین رسانه‌های پراکنده جغرافیایی، رضایت کلی و اهمیت قرار داده شده بر کیفیت خدمات، منطقه به منطقه متفاوت است. توانایی شرکت در ارائه خدمات برتر نیز ممکن است از لحاظ جغرافیایی متفاوت باشد. به همین ترتیب، ترکیب تفاوت‌های منطقه‌ای در تصمیم گیری‌های رضایت مشتری مهم است (ویکاس و واگنر ، ۲۰۰۴). در اصطلاح بازاریابی، رضایت مشتری به عنوان اقدامی در مورد چگونگی تحقق محصول یا خدمات ارائه شده توسط سازمان از انتظارات مشتری توصیف می‌شود. این یکی از مهم‌ترین کلیدها برای اطمینان از موفقیت کسب و کار است؛ زیرا رضایت مشتری تعیین کننده رشد بازار سازمان در آینده خواهد بود. میزان رضایت از طریق سطح کیفیت محصول، کیفیت خدمات ارائه شده، مکان خرید محصول یا خدمات و قیمت محصول یا خدمات اندازه‌گیری می‌شود (ابراهیم و همکاران ، ۲۰۱۶) کیفیت خدمات به عنوان یکی از مهم‌ترین عوامل تعیین کننده موفقیت سازمان‌های خدمات در محیط رقابتی امروزه مورد توجه جدی قرار گرفته است. هرگونه کاهشی در رضایت مشتری به دلیل کیفیت پایین خدمت موجب ایجاد نگرانی‌هایی برای سازمان خدماتی است (فولرتن، ۲۰۰۳). در نتیجه بسیاری از متخصصان بازاریابی معتقدند سازمان‌های خدماتی همواره بایستی انتظارات مشتریان از کیفیت خدمات را تحت‌نظر داشته باشند (پاراسورامان و همکاران، ۱۹۸۵). تعاریف متعددی از کیفیت خدمات ارائه شده است از جمله گرانروس (۲۰۰۱) کیفیت خدمات را اندازه مغایرت بین ادراک مشتری از خدمت و انتظارات او تعریف کرده است (فسنقری و گودرزی ، ۱۳۹۲). اگر میانگین نمره رضایت‌مندی برای یک خدمت در یک منطقه خاص بسیار متفاوت باشد، این امر نشان دهنده‌ی اشکال در خدمات ارائه شده است و در نهایت منجر به زیان شرکت‌ها می‌گردد؛ و لذا باید رضایت‌مندی افراد در مناطقی که اختلاف سطح خدمات زیاد است، مورد توجه جدی قرار گیرد (برینت و فرای، ۲۰۱۹).

۲ رگرسیون وزنی جغرافیایی

این مقاله به معرفی مدلی بر مبنای اقتصادسنجی فضایی می‌پردازد که یک شرکت را قادرمی سازد الگوهای منطقه‌ای را در داده‌های رضایت‌مندی با استفاده از رگرسیون وزنی جغرافیایی [1] **GWR** شناسایی کند. برای پیاده‌سازی این مدل از داده‌های مربوط به رضایت‌مندی مشتریان از کیفیت خودرو در ایالات متحده آمریکا استفاده شده است. این رویکرد برمبنای یک مجموعه داده از ۱۶۴،۰۸۵ مشتری که ۲۱،۶۳۶ کدپستی پنج رقمی در سراسر ایالات متحده را نشان می‌دهد، اعمال شد. مکان‌هایی مورد بررسی قرار گرفته است که در آن رضایت کلی از کیفیت خدمات پایین است اما مطالبه‌گری در آن پررنگ‌تر است. استفاده از روش GWR در تحلیل داده‌های رضایت مشتری به دلیل تطبیق پذیری در پرداختن به داده‌های پراکندگی جغرافیایی یک انتخاب مناسب است. تکنیک GWR در حالتی که امکان جمع‌آوری داده‌ها در همه نواحی فراهم نیست مورد استفاده قرار می‌گیرد؛ زیرا این روش از نظر آماری، اطلاعات را برای مناطق همسایه در طول برآورد قرض می‌گیرد (ویکاس و واگنر ، ۲۰۰۴). برونسدون و فوترینگام و چارلتون (۱۹۹۶)، توصیف می‌کنند که چگونه GWR از نظر مفهومی شبیه به **رگرسیون هسته**[2] است (رست، ۱۹۸۸). تفاوت اصلی این است که در رگرسیون هسته، وزن‌گیری بر روی "فضای ویژگی" متغیرهای مستقل انجام می‌شود، در حالی که GWR در فضای جغرافیایی دوبعدی انجام می‌شود. تفاوت مهم دیگر این است که در رگرسیون هسته، متغیر وابسته از طریق یک رابطه منفرد، بسیار انعطاف‌پذیر، غیرپارامتری که برای همه مشاهدات یا مکان‌ها اعمال می‌شود، با پیش‌بینی کننده‌ها مرتبط است. در مقابل، GWR رابطه خطی بین پیش‌بینی

[1] Graphically Weighted Regression
[2] Kernel Regression

کننده‌ها و متغیر وابسته است که پارامترها را تخمین می‌زند و در نقاط مختلف متفاوت است. هدف GWR برازش یک مدل خطی است که متغیر وابسته را با متغیرهای تبیینی پس از درنظرگرفتن همبستگی مشاهدات در مکان‌های همسایه مرتبط می‌کند (ویکاس و واگنر ، ۲۰۰۴).

۱.۲ شرح مدل

مدل رگرسیون خطی عمومی را در همه‌ی مکان‌ها درنظر بگیرید: $Y = X\beta + \epsilon$ که در آن Y متغیر وابسته و X پیش‌بینی کننده و ϵ عرض از مبدا است. هدف GWR استفاده از تمام اطلاعات موجود در همه نقاط برای برآورد ضرایب رگرسیون است. GWR فرض می‌کند که ضرایب رگرسیون در نقاط مختلف، متفاوت است. ویکاس و واگنر (۲۰۰۴) تکنیک GWR از وابستگی فضایی در داده‌ها استفاده می‌کند. وابستگی فضایی به این معنی است که داده‌های موجود در مکان‌های نزدیک به کانون، اطلاعات بیشتری در مورد ارتباط بین متغیر مستقل و وابسته دارند. GWR هنگام محاسبه برآوردها برای یک موقعیت کانونی، وزن بیشتری به داده‌های مکان‌های نزدیک‌تر نسبت به داده‌های مکان‌های دورتر می‌دهد. از لحاظ آماری، وابستگی فضایی با استفاده از یک طرح وزنی در یک مدل حداقل مربعات عمومی [۳] **GLS** اجرا می‌شود، به طوری که مکان‌هایی که نزدیک به کانون هستند، وزن بیشتری در تعیین معادله رگرسیون برای محل کانونی دارند. ماتریس وزن حاوی وزن W_j برای همه نقاطی است که در محاسبات معادله رگرسیون برای محل کانونی استفاده می‌شود. با استفاده از این ماتریس وزن جغرافیایی، می‌توان برآورد حداقل مربعات وزنی را برای هر مکان j به صورت زیر به دست آورد:

$$\hat{\beta}_j = (X' W_j X)^{-1} X' W_j y \qquad (۱.۲)$$

این یک رگرسیون عمومی است که با GLS تخمین زده می‌شود، به طوری که W_j ماتریس قطری $n \times n$ است که شامل $W_{jj'}, j' = 1, ..., j$ در قطر، توسط یک تابع فروپاشی مبتنی بر فاصله تعریف شده است.

$$w_{jj'} = \exp(-d^2_{jj'}/\theta) \qquad (۲.۲)$$

به طوری که θ پارامتر انقطاع فاصله است، و djj' فاصله اقلیدسی بین مکان‌های j و j' است.

پیاده‌سازی تابع فشرده مبتنی بر فاصله، نیاز به برآورد پهنای باند مطلوب (θ) قبل از برآورد حداقل مربعات وزنی دارد. پارامتر پهنای باند ارتباط هر مشاهده همسایه را برای برآورد پارامترهای سطح منطقه‌ای β تعیین می‌کند. هنگامی که پهنای باند به اندازه کافی بزرگ است، مدل GWR به یک رگرسیون استاندارد در سراسر مناطق برمی‌گردد. مناسب‌ترین مقدار پهنای باند با استفاده از روش حداقل مربعات متقاطع که کلیولند (۱۹۷۹) پیشنهاد می‌کند، تعیین شد. اعتبارسنجی متقابل برای مقدار بهینه θ تعیین شد که به صورت تابع امتیاز زیر است:

$$\sum_{j=1}^{n} [y_j - \hat{y}_{\neq j'}(\theta)]^2 \qquad (۳.۲)$$

که در آن $\hat{y}_{\neq j}$ مقدار متناسب با y را با مشاهدات مربوط به محل کانونی j که از فرآیند کالیبراسیون حذف شده است، نشان می‌دهد. مقدار θ که این تابع امتیاز را به حداقل می‌رساند به عنوان پهنای باند برای محاسبه ماتریس وزن استفاده می‌شود. جزئیات مربوط به برآورد GWR با اعتبارسنجی متقابل GLS در کار برونسدون و فوترینگام و چارلتون (۱۹۹۶) یافت می‌شود. مدل GWR همانطور که برونسدون و فوترینگام و چارلتون (۱۹۹۸) پیشنهاد می‌کنند، این مورد را تنها با یک

[3] Generalized Least Squares

مشاهده در هر محل بررسی می‌کنند. با این حال، در این تجزیه و تحلیل، در هر کدپستی چندین پاسخ را مشاهده می‌شود. به جای جمع آوری داده‌ها در هر کدپستی، مدل در سطح پاسخ‌دهندگان تخمین زده می‌شود. این ارزیابی سطح فردی می‌تواند مشکل ساز باشد، زیرا بر اساس تعداد مشاهدات در آن، بر هر مکان تأکید می‌شود. اگرچه این امر در مورد نمونه‌گیری تصادفی ساده یا تصادفی طبقه‌ای ساده قابل قبول است، اما این احتمال وجود دارد که تخمین‌ها را غیر از این سوگیری کند. بنابراین، مفهوم وزن برابر برای هر مکان، مانند مدل GWR عمومی، با وزن هر مشاهده با معکوس اندازه نمونه در محل آن حفظ شده است. مدل GWR تغییر یافته را می‌توان به صورت زیر بیان کرد:

$$\hat{\beta}_j = (X' w_i f X)^{-1} X' w_i f y \qquad (۴.۲)$$

که در آن f یک ماتریس قطری $n \times n$ که شامل معکوس اندازه نمونه در مکان j ام است که مشاهده منحصر به‌فرد i به آن تعلق دارد و ماتریس قطری w_iشامل وزن‌های مبتنی بر فاصله بین هر مشاهده فردی و کانونی (i) است.

توجه شود که اگر ضرایب رگرسیونβ_j در سطح مکان تعریف شده باشد، تخمین‌ها برای همه مشاهدات در همان مکان یکسان است؛ زیرا تمام مشاهدات در همان مکان دارای وزن واحد هستند که ناشی از جمع شدن همه مشاهدات از همان مکان می‌شود (ویکاس و واگنر ، ۲۰۰۴).

۲.۲ چه زمانی از GWR باید استفاده کرد؟

تکنیک GWR هنگام استفاده از همبستگی فضایی در متغیرها باید مورد استفاده قرار گیرد. همبستگی مثبت و قوی نشان می‌دهد که مقادیر مناطق همسایه مشابه یکدیگر هستند، در حالی که همبستگی منفی و قوی حاکی از آن است که مقادیر مناطق همسایه با یکدیگر متفاوت هستند. بزرگی و جهت همبستگی فضایی برای یک متغیر را می‌توان با استفاده از دو آزمون آماری **موران**[۴] و **گری**[۵] اندازه‌گیری کرد. اگر مقادیر موران بزرگ‌تر از $\frac{-۱}{N-۱}$ باشد، نشان دهنده‌ی وابستگی مثبت و معکوس است. برای آزمون گری مقادیر کمتر از ۱ نشان دهنده‌ی همبستگی مثبت و مقادیر بیشتر از ۱ نشان دهنده‌ی همبستگی منفی فضایی است. برای هر دو آمار، آزمون‌های آماری معنی‌دار می‌تواند برای شناسایی همبستگی فضایی انجام شود. بر اساس این آزمون‌های معنی‌داری آماری، می‌توان تصمیم به ادامه GWR گرفت (کلیف و ارد ، ۱۹۸۱، ۱۹۷۳).

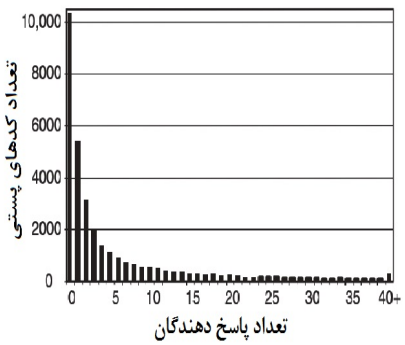
۳ داده‌ها

در این قسمت برای پیاده‌سازی مدل معرفی شده از ۱۶۴،۰۸۵ داده مربوط به مشتریانی که در نظرسنجی رضایت‌مندی از کیفیت خودرو شرکت کرده‌اند، استفاده می‌شود. از این مجموعه داده‌ها، نمونه‌ی نهایی به نحوی انتخاب می‌شود که هر کدپستی حداقل شامل ۵ مشتری باشد؛ پس ازین شرط، تعداد داده‌ها به ۱۳۲،۰۸۵ مورد کاهش می‌یابد که در مجموع در برگیرنده‌ی ۲۱،۶۳۶ کدپستی هستند. در جدول ۱ آمار توصیفی مربوط به نمونه نشان داده شده است. از ۳۱،۹۵۶ کدپستی پنج رقمی در ایالات متحده، تنها ۲۱،۶۳۶ یا ۶۷/۷٪ داده وجود دارد. به عبارت دیگر، ۳۲/۳٪ از کدپستی هیچ داده‌ای ندارند. در شکل ۱ تعداد پاسخ‌دهندگان از هر کدپستی برای کسی که اطلاعات آن در دسترس است، نشان داده شده است. به طور خلاصه، بخش بزرگی از (۳۲/۳٪) کدپستی بدون داده است؛ بیش از نیمی از کدهای پستی باقی مانده، دارای پنج عدد داده یا کمتر هستند. تجزیه و تحلیل سطح کدپستی به طور سیستماتیک تقریبا یک سوم از کدهای پستی را حذف می‌کند. در میان کدهای باقی مانده، برآوردهای سطح کدپستی ممکن است برای کسانی که دارای پنج داده یا کمتر باشند، قابل اعتماد نباشند. مهم‌تر از همه، چنین تحلیلی نمی‌تواند مزایای همبستگی فضایی را در فرآیند برآورد به کار گیرد. به همین

[4] Moran's I
[5] Geary's C

جدول ۱: شرح نمونه

	مشخصات نمونه	درصد
جنسیت	مردان	۶۴/۲
	زنان	۳۵/۸
سطح تحصیلات	دبیرستان یا کمتر	۲۶/۳
	تحصیلات دانشگاهی	۲۷/۳
	فارق التحصیل دانشگاه	۴۶/۵
وضعیت تاهل	متاهل	۷۶/۰
	مجرد	۱۹/۵
	سایر	۵/۴
سن	زیر ۳۰ سال	۹/۴
	۳۰ـ۵۹ سال	۶۸/۱
	بالای ۶۰ سال	۲۲/۵
رضایت کلی	۱	۰/۹
	۲	۰/۵
	۳	۰/۹
	۴	۲/۱
	۵	۲/۸
	۶	۸/۷
	۷	۹/۴
	۸	۲۵/۴
	۹	۱۸/۵
	۱۰	۳۰/۷



شکل ۱: فراوانی پاسخ دهندگان با کدپستی

دلیل از GWR برای حل این مشکل استفاده شد: GWR به طور آماری از مشاهدات نقاط همسایه در هنگام برآورد ضرایب برای یک موقعیت کانونی استفاده می‌کند. این ویژگی GWR مفید است، زیرا در اغلب موارد موقعیت کانونی دارای تعداد کمی از مشاهدات است. در چنین مواردی، برآوردهای بهتر را به دست می‌آوریم، زیرا مشاهدات از کدهای پستی همسایه اطلاعات اضافی را فراهم می‌کنند. استفاده از کدهای پستی همسایه نیز به شناسایی مناطقی که دارای ضرایب مشابه هستند،

جدول ۲: همبستگی فضایی در متغیرها

متغیر	Geary's C	Moran's I
رضایت کلی	۰/۰۰۸ـ	۰/۶۹
خدمات نمایندگی	۰/۰۰۷ـ	۰/۷۶
کیفیت خودرو	۰/۰۰۷ـ	۰/۷۷

کمک می‌کند و منجر به دید منظمی از الگوهای فضایی داده‌ها می‌شود.

۱.۳ متغیرها

هر مشتری (i) به سوالات زیر با استفاده از مقیاس ده نمره (۱ = "بسیار ناراضی"، ۱۰ = "بسیار راضی") پاسخ داد: "بر اساس تجربه، رضایت کلی خود را از مالکیت خودرو ($OVERALLSAT_i$)، کیفیت خودرو ($PRODQUAL_i$) و خدمات نمایندگی ($DLRSRV_i$) چگونه ارزیابی می‌کنید؟"

مشتریان نیز برای پاسخگویی، کدپستی پنج رقمی از محل اقامت فعلی خود را نشان می‌دادند. سپس، مختصات طول و عرض جغرافیایی برای مرکز هر کدپستی پنج رقمی تعیین شد. فاصله اقلیدسی بین هر کدپستی در مجموعه داده محاسبه شد. از این فاصله در برآورد مدل GWR استفاده شد. سپس، همبستگی فضایی در متغیرها تخمین زده شد. اندازه‌گیری همبستگی فضایی (موران و گری) در هر ۲۱،۶۳۶ کدپستی برای هر سه متغیر در جدول ۲ نشان داده شده‌است. موران بزرگ‌تر از ($\frac{-۱}{N-۱}$) و ($p < ۰/۰۵$) و گری کمتر از ۱ و ($p < ۰/۰۵$) است. این نشان دهنده‌ی همبستگی فضایی مثبت برای هر سه متغیر است.

۴ نتایج

با استفاده از مدل GWR رابطه نهایی به صورت زیر نوشته می‌شود؛ که در آن OVERALLSAT رضایت کلی از تجربه مالکیت، PRODQUAL کیفیت عملکرد خودرو و DLRSRV خدمات نمایندگی است:

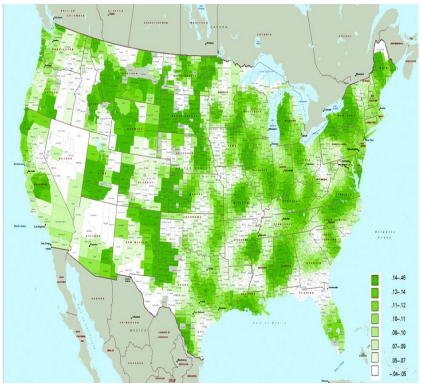
$$OVERALLSAT_i = \beta_{۰j} + \beta_{۱j} PRODQUAL_i + \beta_{۲j} DLRSRV_i + e_i \qquad (۱.۴)$$

که در آن e_i دارای توزیع نرمال است. قابل ذکر است که برای هر فرد i سه پارامتر در هر مکان j تخمین زده شد.

۱.۴ الگوهای منطقه‌ای در اهمیت خدمات نمایندگی

ضریب رگرسیون برای خدمات نمایندگی در شکل ۲ نشان داده شده است. به عنوان مثال در بخش جنوب شرقی کلرادو، اهمیت کیفیت خدمات بالاست (توسط رنگ تیره نشان داده شده)، در حالی که در جنوب غربی کشور، کیفیت خدمات پایین (توسط رنگ روشن‌تر نشان داده شده) است. از مرکز به شمال کلرادو، کم‌ترین اهمیت خدمات مشاهده می‌شود، که توسط سایه سفید نشان داده شده است. علاوه بر این، در بخش شمال شرقی کلرادو، اهمیت رضایت از خدمات به طور یکنواخت پایین است، همانطور که در مجاورت نبراسکا است.

به طور کلی، خدمات نمایندگی در بخش شرقی یا غربی کشور، به‌ویژه در بخش‌هایی از جنوب اورگان و شمال کالیفرنیا، اهمیت بیشتری دارد. سایر مناطق که خدمات نمایندگی در آن مهم است، شمال شرقی نیو مکزیکو تا کلرادو و کانزاس را شامل می‌شود. میدوست، کنتاکی، ایندیانا، و ایلینوی نیز مناطقی هستند که خدمات نمایندگی در آن‌ها از اهمیت بالا یا

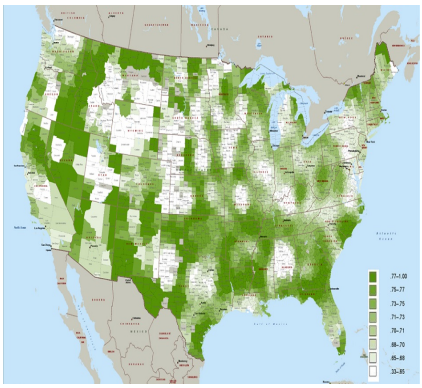


شکل ۲ : ضریب اطمینان خدمات نمایندگی

متوسطی برخوردار است. جالب است که یک منطقه بزرگ در غرب آمریکا وجود دارد ــ از جمله نوادا، آریزونا، اورگان و آیداهو ــ که خدمات نمایندگی از اهمیت کمتری برخوردار است. علاوه بر این ، کلان شهرهای مهم مختلفی در مناطقی قرارگرفته اند که دارای اهمیت متفاوتی هستند. به عنوان مثال ایندیاناپولیس، کلمبوس، اوهایو و فیلادلفیا اهمیت خدمات نمایندگی در آن‌ها زیاد است، در حالی که جکسونویل، فلوریدا، سانفرانسیسکو، سان خوزه و کالیفرنیا، از اهمیت متوسط تا کمی برخوردار است.

۲.۴ الگوهای منطقه‌ای در اهمیت کیفیت خودرو

نتایج مربوط به اهمیت کیفیت وسایل نقلیه در شکل ۳ نشان داده شده است. همان‌طور که انتظار می‌رفت، آن‌ها معکوس نتایج مربوط به خدمات نمایندگی هستند. به عنوان مثال در شمال شرقی تگزاس، اهمیت کیفیت خودرو در تعیین رضایت کلی نسبتا بالا است. این بخش از تگزاس الگوی اهمیت بالای کیفیت خودرو را با کشورهای همسایه لوئیزیانا و آرکانزاس به اشتراک می‌گذارد. با این حال، در دالاس و آستین، اهمیت نسبی کیفیت خودرو، متوسط رو به پایین است، همانطور که در شمال غربی تگزاس این حالت وجود دارد. در مقابل، اهمیت کیفیت خودرو در جنوب تگزاس بسیار زیاد است. الگوی اهمیت کیفیت خودرو برای بازارهای مختلف کلان شهر نیز متفاوت است. در حالی که فیلادلفیا، نیویورک و واشنگتن در مناطق کم اهمیتی قرار دارند، ممفیس و سانفرانسیسکو در مناطقی از اهمیت بالایی برخوردار هستند.



شکل ۳: ضریب اطمینان کیفیت خودرو

جدول ۳: اعتبارسنجی مدل با استفاده از ۲۰٪ نمونه پشتیبان (توانایی پیش‌بینی)

مدل	سطح تجزیه و تحلیل	روش برآورد	دامنه مقدار پیش‌بینی شده به طور متوسط در همه مناطق	درصد مشاهدات واقع در فاصله اطمینان ۹۵٪
۱	۱۳،۸۴۶ کدپستی	GWR	۰/۲۸	۸۷/۰۵
۲	۱۳،۸۴۶ کدپستی	GWR No	۱/۰۷	۴۰/۴۲
۳	شهرستان (تعداد = ۳۱۰۲)	GWR No	۲/۳۹	۷۲/۵۶
۴	وضعیت (تعداد = ۵۰)	GWR No	۳/۸۷	۸۳/۵۸
۵	نمونه از کدپستی (تعداد = ۱۷،۲۲۸)	GWR	۰/۶۲	۷۱/۴۱

۳.۴ عملکرد مدل

عملکرد مدل با استفاده از دو معیار ارزیابی می‌شود؛ توانایی مدل برای پیش‌بینی رضایت کلی و پایایی برآوردهای پارامتر برای رانندگان. برای ارزیابی عملکرد پیش‌بینی شده از **نمونه پشتیبان**[6] با استفاده از ۲۰ درصد مشاهدات استفاده شده است. برای ایجاد نمونه پشتیبان، کدهای پستی‌ای انتخاب شد که دارای بیش از پنج مشاهده بود تا بتوان برآوردهای رگرسیونی استاندارد برای هر کدپستی را به دست آورد؛ به همین ترتیب حداقل سه مشاهده برای برآورد مدل‌های معیار نیاز است؛ GWR می‌تواند برآورد پارامتر را حتی در یک منطقه کدپستی که داده‌ای ندارد، بر اساس داده‌ها در مناطق همسایه محاسبه کند. از کدهای پستی که شامل پنج یا بیشتر مشاهده است، به طور تصادفی ۳۲،۰۰۰ مشاهده انتخاب شد؛ کدپستی دارای مشاهدات متعدد در نمونه پشتیبان است. نمونه نهایی حاوی ۳۲،۰۰۰ مشاهده و ۱۳،۸۴۶ کدپستی را نشان می‌دهد.

جدول ۳ GWR را با رگرسیون ساده با استفاده از دو شاخص عملکرد پیش‌بینی شده مقایسه می‌کند. ابتدا، با استفاده از فاصله اطمینان ۹۵٪ برای هر مکان، بهترین گزینه ممکن به عنوان نمونه انتخاب شد. این کار با استفاده از روش پیش‌بینی انجام شد که نستر (۱۹۹۶) و برونسدون و فوترینگام و چارلتون (۱۹۹۸) پیشنهاد دادند. این سنجش، توانایی مدل را برای پیش‌بینی نمرات رضایت‌مندی بر اساس کدپستی نشان می‌دهد. اندازه‌گیری دوم درصد مشاهدات در نمونه پشتیبان است که برای آنها نمره رضایت واقعی در بازه اطمینان ۹۵٪ توسط مدل پیش‌بینی شده قرار گرفته است. این اندازه‌گیری با نشان دادن درصد دفعات فاصله اطمینان، میزان رضایت واقعی را در یک نمونه پشتیبان، با تأیید تجربی از فاصله اطمینان ۹۵٪ در بین مشتریان ارائه می‌دهد. مدل ۱ از برآورد GWR در سطح کدپستی پنج رقمی استفاده می‌کند. مدل ۲ یک مدل در سطح کدپستی و بدون GWR است که با اجرای رگرسیون‌های جداگانه در هر کدپستی تخمین زده شده است. مدل ۲ حداقل از سه مشاهده در یک کدپستی استفاده می‌کند و برای سایر مشاهدات در کدپستی نمرات رضایت پیش‌بینی شده را با نمونه‌های واقعی مقایسه می‌کند. جدول ۳ نشان می‌دهد که فاصله اطمینان ۹۵٪ برای برآورد مقادیر پیش‌بینی شده با مدل GWR نسبت به مدل بدون GWR کمتر است (۰/۲۸ در مقابل ۱/۰۷). بنابراین، پیش‌بینی‌هایی که با GWR به‌دست آمده است، نسبت به پیش‌بینی‌های به دست آمده با این فرض که همه مناطق مستقل هستند، بهتر است. ستون سمت چپ جدول ۳ نشان می‌دهد که هنگام استفاده از GWR (مدل ۱)، ۸۷/۰۵٪ از مشاهدات در نمونه پشتیبان دارای نمره رضایت‌مندی هستند که در فاصله اطمینان ۹۵٪ از مقدار پیش‌بینی شده قرار می‌گیرند. بدون GWR (مدل ۲)، فقط ۴۰/۴۲ درصد از مقادیر رضایت کلی در نمونه پشتیبان در فاصله اطمینان ۹۵٪ از مقدار پیش‌بینی شده قرار می‌گیرند. بنابراین، پیش‌بینی‌های بازه‌ای که مدل GWR تولید می‌کند، احتمالا مقدار واقعی را شامل می‌شود حتی اگر فواصل کوچک‌تر باشند، که نشان‌دهنده‌ی عدم قطعیت پایین در مورد پیش‌بینی‌ها است. بازه اطمینان بر اساس رگرسیون مستقل نه تنها گسترده‌تر نبود، بلکه نشان دهنده برآورد خطای بالاتر است، همچنین پیش‌بینی رضایت در یک نمونه پشتیبان نیز از دقت کمتری برخوردار بود. برای مقایسه،

[6] Holdout Sample

همچنین مدل‌ها در سطح شهرستان (مدل ۳) و سطح دولت (مدل ۴) تخمین زده شد. همانطور که در جدول ۳ نشان داده شده است، نمره رضایت واقعی در بازه اطمینان ۹۵٪ پیش‌بینی شده برای ۷۲/۵۶٪ از پیش‌بینی‌های انجام شده با برآورد سطح شهرستان و ۸۳/۵۸٪ با مدل سطح دولت می‌باشد. در نگاه اول، به نظر می‌رسد نتایج قابل مقایسه با مواردی است که با برآورد سطح کدپستی به دست آمده است. با این حال، مشاهده می‌کنیم که میانگین دامنه مقادیر پیش‌بینی شده برای هر منطقه برای مدل در سطح شهرستان (۲/۳۹) بسیار گسترده‌تر است و حتی برای مدل سطح دولت (۳/۸۷) حتی بیشتر از مدل تخمین سطح کدپستی است. این بدان معنی است که به احتمال زیاد نمره رضایت واقعی در بازه اطمینان وسیع‌تر قرار می‌گیرد. از آنجا که عدم اطمینان بیشتر در پیش‌بینی‌ها در سطح شهرستان و ایالت وجود دارد، فواصل اطمینان بسیار بیشتر است. بنابراین، برای مدیرانی که می‌خواهند از پیش‌بینی دقیق رضایت کلی برای هر کدپستی مورد نظر اطمینان حاصل کنند، مدل سطح کدپستی با GWR بهترین عملکرد را دارد.

در نهایت، یک نمونه پشتیبان با مجموعه ی دیگری از مشاهدات انجام شد. در این حالت به‌طور تصادفی ۳۲،۰۰۰ مشاهده از مجموعه داده‌ها به دست آمد، بدون اینکه محدود به کدپستی‌های شامل ۵ یا بیشتر مشاهده باشد. این نمونه از ۳۲،۰۰۰ نقطه داده شامل ۱۷،۲۲۸ کدپستی است.۸۰٪ باقی‌مانده از داده‌ها برای تخمین مدل GWR استفاده شد. از مجموعه پارامترهایی که از این مدل به دست آمد، استفاده شد تا پیش‌بینی مشاهدات از نمونه‌های برگزیده انجام شود. نتایج این تحلیل در ردیف پنجم جدول۳ آمده است. توجه داشته باشید که نتایج این پیش‌بینی با برآوردهای حداقل مربعات معمولی قابل مقایسه نیستند، زیرا برای بسیاری از کدهای پستی، چند نظرسنجی وجود داشت تا مدل حداقل مربعات معمولی برآورد شود. محدوده مقادیر پیش‌بینی شده در حال حاضر بزرگ‌تر از زمانی است که فقط کدپستی با پنج یا بیشتر مشاهدات در دست بود (۰/۶۲ در مقابل ۰/۲۸). این به این دلیل است که کدهای پستی زیادی بود که در منطقه مرکزی هیچ مشاهداتی نداشت و بنابراین باید از طریق مدل GWR پیش‌بینی‌ها از نقاط همسایه تشکیل شود. علاوه بر این، تعداد مشاهداتی که در محدوده $\pm 1/96(\frac{\sigma}{\sqrt{n}})$ قرار دارد، ۷۱/۴۱٪ بود، در مقایسه با ۸۷/۰۵٪ برای مدل قبلی. با این حال، این عملکرد پیش‌بینی هنوز هم نسبت به مدل‌های معیار برتر است.

۴.۴ پایداری برآوردهای پارامتر

برای ارزیابی ثبات برآوردها از اهمیت خدمات نمایندگی و کیفیت خودرو، نمونه‌ها در هر کدپستی به صورت تصادفی به دو قسمت تقسیم شد. هنگامی که فقط یک مشاهده برای یک کدپستی در دسترس بود، به صورت تصادفی به یکی از نمونه‌های تقسیم شده اختصاص داده شد، و در نمونه دیگر گم شد. سپس GWR در دو طرف نمونه اعمال شد، و ضرایب کدهای پستی که هیچ داده‌ای نداشتند محاسبه شد. برای ارزیابی ثبات پارامتر، همبستگی از دو طرف و برای داده‌های کامل برای سه مجموعه از کدپستی‌ها محاسبه شد: ۱. آن‌هایی که داده‌هایشان در هر دو نمونه دردسترس بود، ۲. آن‌هایی که داده‌هایشان در یک نمونه گم شده بود، و ۳. آن‌هایی که داده‌هایشان در هر دو نمونه از دست رفته بود (یعنی برآورد پارامتر در هر دو طرف محاسبه شد). تقسیم نیمه ده بار تکرار شد و میانگین و انحراف معیار مربوط به همبستگی در ده تکرار گزارش شد. همبستگی تقسیم نیمه، که در ۳۱،۹۵۶ کدپستی (از جمله کسانی که بدون اطلاعات هستند) محاسبه شده است، ارزیابی ثبات پارامتر برای استفاده از مدل GWR را ارائه می‌دهد. هنگامی که داده‌ها در هر دو طرف تقسیم شده بودند، همبستگی بیشتر از ۰/۶۵ بود. این شواهد قوی در مورد ثبات پارامتر درنظر گرفته می شود، زیرا همبستگی بین دو برآورد و یک نمونه بزرگ (۳۱،۹۵۶) محاسبه شده است. وقتی برآورد پارامتر برای یک نمونه تصادفی (یعنی هیچ اطلاعاتی برای کدپستی خاصی دردسترس نیست) محاسبه شود، همبستگی حدود ۰/۴۰ کاهش می‌یابد. این ضعف در تقسیم نیمه متقابل انتظار می‌رفت، زیرا همبستگی‌ها در حال حاضر شامل یک برآورد پارامتر و یک مقدار محاسبه شده است. یک نتیجه قابل توجه این است که همبستگی تقسیم نیمه در زمانی که هر دو نمونه دارای مقادیر محاسبه شده بودند (ستون سمت چپ جدول ۴)

کاهش یافت.

جدول ۴: میانگین و انحراف استاندارد همبستگی تقسیم شده و نیمه (پایداری پارامتر)

	همبستگی بین دو نمونه			
پارامتر	رضایت کلی	کدپستی با مشاهده	داده‌های گمشده در یک نمونه	داده‌های گمشده در دو نمونه
عرض از مبدا	۰/۰۲ و ۰/۶۱	۰/۰۴ و ۰/۶۶	۰/۰۴ و۰.۴۱	۰/۰۸ و ۰/۴۱
خدمات نمایندگی	۰/۰۴ و ۰/۶۲	۰/۰۶ و ۰/۶۸	۰/۰۵ و ۰/۴۱	۰/۰۷ و ۰/۴۱
کیفیت خودرو	۰/۰۶ و ۰/۵۸	۰/۰۸ و ۰/۷۰	۰/۰۶ و ۰/۴۱	۰/۱۱ و ۰/۴۶

بحث و نتیجه‌گیری

در این مقاله براساس روش ،GWR نتایج (برای یک شرکت در صنعت خودرو) موارد زیر را نشان می‌دهد:

یک تنوع مکانی سیستماتیک در الگوی رضایت کلی و اهمیتی که در رانندگان آن به وجود می‌آید، وجود دارد. اگرچه الگوی خاص تغییرات منطقه‌ای بر اساس دسته مورد بررسی متفاوت است، اما وجود تغییرات مکانی منظمی باید در تحقیقات بیشتری از داده‌های رضایت گنجانیده شود. اختلافات منطقه‌ای در اهمیت رانندگان و الگوهای کلی رضایت‌مندی یک شرکت را قادر می‌سازد مناطقی که باید در آن خدمات را بهبود بخشند، شناسایی کند. این شرکت می‌تواند در اولویت بندی مناطق برای سرمایه گذاری در بهبود خدمات نمایندگی اقدام کند.

مراجع

فسنقری، ا. گودرزی، م.، (۱۳۹۲)، بررسی تاثیر ابعاد کیفیت خدمات مدل سروکوال بر رضایت‌مندی مشتریان زن باشگاه‌های زنان ، *نشریه پژوهش‌های فیزیولوژی و مدیریت در ورزش*، **۹**، ۲۱ ـ ۳۴.

Anthony, J., Zahorik. and Timothy, L., Keiningham. (1995), Return on Quality: Making Service Quality Financially Accountable, *Journal of Marketing*, **59**, 58–70.

Brint, A., Fry, J. (2019), Regional Bias When Benchmarking Services Using Customer Satisfaction Scores, *Total Quality Management and Business Excellence*, 1-15.

Bronsdon, A., Fotheringham, S., Martin, E and Charlton. (1996), Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity, *Geographical Analysis*, **28**, 281–298.

Bronsdon, A., Fotheringham, S., Martin, E and Charlton. (1998), Geographically Weighted Regression: Modeling Spatial Nonstationarity, *Journal of the Royal Statistical Society D*, **47**, 431–443.

Cliff., Andrew, D. and Ord, J.K. (1973), *Spatial Autocorrelation*, London.

Cliff., Andrew, D. and Ord, J.K. (1981), *Spatial Processes: Models and Applications*, London.

Cleveland., William, S. (1979), Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, **74**, 829–836.

Fullerton, G. (2003), Women Sport Marketing Business, **124**, 42-48.

Gronroos, C. (2001), *Service Management and Marketing*, Second Editions, Wiley, 95.

Iberahim, H., Mohd Taufik, N.K., Mohd Adzmir, A.S. and Saharuddin, H. (2016), Customer Satisfaction on Reliability and Responsiveness of Self Service Technology for Retail Banking Services, *Procedia Economics and Finance* , **37**, 13 – 20.

Nester., Marks, R. (1996), An Applied Statistician's Creed, *Applied Statistics*, **45 (4)**, 401–410.

Parasuraman, A., Valarie, A., and Berry, L. (1985), A Conceptual Model of Service Quality and its Implications for Future Research, *Journal of marketing*, **49**, 41-50.

Rust., Roland, T. (1988), Flexible Regression, *Journal of Marketing Research*, **25**, 10–24.

Valarie, A., Zeithaml, and Katherine, N., Lemon. (2000), How Customer Lifetime Value Is Reshaping Corporate Strategy, *Driving Customer Equity*, New York: The Free Press.

Vikas, M., Wagner, A., Kamakura. and Rahul, G. (2004), Geographic Patterns in Customer Service and Satisfaction: An Empirical Investigation, *Journal Of Marketing*, **68**, 48-62.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

تحلیل داده‌های فضایی–زمانی بر اساس تجزیه تاکر تانسور کوواریانس

سمیرا سعادتی[1]، محسن محمدزاده
گروه آمار، دانشگاه تربیت مدرس

چکیده:
در تحلیل داده‌های فضایی–زمانی، روش متداول برای لحاظ کردن ساختار همبستگی داده‌ها، استفاده از تابع کوواریانس فضایی–زمانی است، که معمولاً نامعلوم است و باید بر اساس مشاهدات برآورد شود. برای این منظور معمولا محدود کننده‌هایی از قبیل مانایی، همسانگردی و تفکیک‌پذیری میدان تصادفی در نظر گرفته می شود که برازش مدل‌های معتبر به تابع کوواریانس فضایی–زمانی را تسهیل می‌کنند. اما در اغلب مسائل کاربردی این فرض‌ها لزوما محقق نمی شوند. در این مقاله، به منظور رهایی از این مشکل و تسریع محاسبه پیشگویی فضایی–زمانی برای یک میدان تصادفی، یک مدل احتمالی بر اساس تجزیه تاکر تانسور کوواریانس فضایی–زمانی ارائه می‌شود، سپس نحوه کاربست روش مطرح شده در پیشگویی فضایی–زمانی داده‌های سرعت باد ایران نشان داده خواهد شد.

واژه‌های کلیدی: داده‌های فضایی–زمانی، تانسور کوواریانس، تجزیه تاکر.
کد موضوع‌بندی ریاضی (۲۰۱۰): 60Gxx، 60Hxx، 60Bxx.

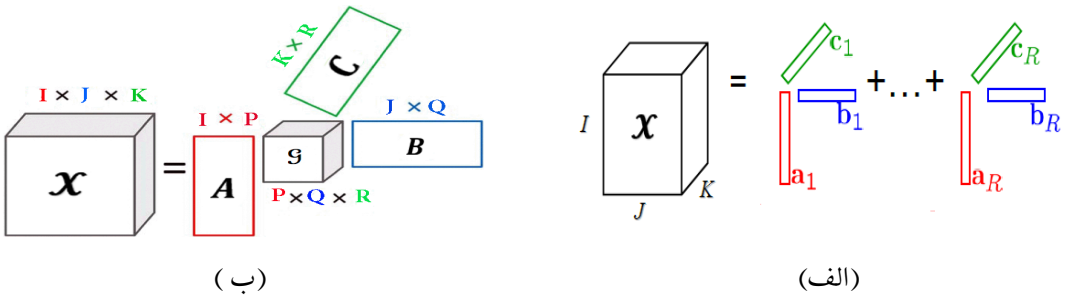
۱ مقدمه

تانسور تعمیمی از مفاهیم اسکالر، بردار و ماتریس است که ساده‌ترین حالت آن می‌تواند دارای یک عضو باشد و تانسور اسکالر نامیده می‌شود. تانسور مرتبه یک، یک بردار، تانسور مرتبه دو، یک ماتریس و تانسورهای مراتب سه و بالاتر تانسورهای مرتبه بالا نامیده می‌شوند. تانسورها از آن جهت مورد استفاده قرار می‌گیرند که باعث ایجاد نظم بین داده‌های یک مسئله و دسته‌بندی اطلاعات آن می‌شوند و بستر ریاضی مناسب و ساده‌ای را برای فرمول‌بندی و حل مسائل متعدد در مباحث گوناگون نظیر مکانیک سیالات و نسبیت عام فراهم می‌کنند. تجزیه تانسور اولین بار توسط هیچکاک (۱۹۲۷) معرفی شد، اما کمتر مورد توجه قرار گرفت تا زمانی که تاکر (۱۹۶۳) و کراسکال (۱۹۷۷) روش‌های مختلف تجزیه تانسور را مطرح کردند و باعث

[1] Canonical Polyadic
[2] Tucker decomposition

[1] نام و ایمیل ارائه دهنده مقاله: سمیرا سعادتی، s.saadati@modares.ac.ir

انجام پژوهش‌های متنوعی بر اساس تجزیه تانسور شد. دو نوع اصلی از تجزیه تانسورها، تجزیه چندتایی کانونی[۱] (CP) و تجزیه تاکر[۲] هستند.



(ب) (الف)

شکل ۱ : تجزیه‌های تانسور مرتبه سه، الف_تجزیه CP، ب_تجزیه تاکر

تجزیه CP، یک تانسور را به مجموع تانسورهای با رتبه یک تجزیه می‌کند. برای مثال، یک تانسور مرتبه سه $\mathcal{X} \in \mathbb{R}^{I\times J\times K}$ با استفاده از تجزیه CP به صورت

$$\mathcal{X} \approx \sum_{r=1}^{R} a_r \circ b_r \circ c_r = [\![A, B, C]\!] \qquad (1.1)$$

تجزیه می‌شود، که در آن R یک عدد صحیح مثبت،

$$a_r \in R^I \, , b_r \in R^J \, , c_r \in R^K, \quad r = 1, ..., R.$$

نماد $\circ$ نشان‌دهنده ضرب خارجی بردارها است. مؤلفه‌ای تانسور (۱.۱) به صورت

$$\mathcal{X}_{ijk} \approx \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \; i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \cdots, K.$$

نوشته می‌شود، که در شکل ۱ ـ الف نشان داده شده است. تجزیه تاکر نیز هر تانسور مرتبه سه را به یک تانسور هسته ضرب شده در سه ماتریس مانند شکل ۱ ـ ب به صورت

$$\mathcal{X} \approx G \times_1 A \times_2 B \times_3 C = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_p \circ b_q \circ c_r = [\![G; A, B, C]\!] \qquad (2.1)$$

تجزیه می‌کند، که در آن $A \in R^{I\times P}$، $B \in R^{J\times Q}$ و $C \in \mathbb{R}^{K\times R}$ ماتریس‌های فاکتور و $G \in \mathbb{R}^{P\times Q\times R}$ تانسور هسته است و ورودی‌های آن، سطح تعامل بین مؤلفه‌های مختلف را نشان می‌دهد. تجزیه عنصر به عنصر تاکر مرتبه سه در (۲.۱) به صورت

$$x_{ijk} \approx \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_{ip} \circ b_{jq} \circ c_{kr}, \quad i = 1, \cdots I, \; j = 1, \cdots, J, \; k = 1, \cdots, K,$$

است، که در آن P, Q و R تعداد مؤلفه‌ها (ستون‌ها) به ترتیب در ماتریس‌های فاکتور A, B و C هستند.

پیشگویی قابل اعتماد سرعت وزش باد در برخی حوزه‌ها از قبیل ایمنی هوانوردی، ایمنی سازه‌های ساختمانی، انرژی باد و انتقال آلاینده‌ها بسیار مهم است.

اغلب ابزارهای پیشگویی سرعت باد مبتنی بر فیزیک (لینچ ، ۲۰۰۸) یا داده‌های آماری (سومن و همکاران ، ۲۰۱۰) هستند. گاهی هم ترکیبی از مدل‌های فیزیکی و ابزارهای آماری هستند (سالسدو-سانز و همکاران ، ۲۰۰۹).

عصمتی و محمدزاده ۱۳۹۶ ساختار همبستگی فضایی-زمانی یک میدان تصادفی نامانا و تفکیک‌ناپذیر را با استفاده از تانسور کوواریانس فضایی-زمانی تعیین و پیشگویی فضایی-زمانی مبتنی بر تجزیه CP روی مجموعه داده‌های سرعت باد در کشور ایرلند را انجام دادند. در این مقاله، به منظور رسیدن به نتایج دقیق‌تر به بررسی ساختار همبستگی فضایی-زمانی یک میدان تصادفی نامانا و تفکیک‌ناپذیر (محمدزاده ، ۱۳۹۸) با استفاده از تجزیه تانسور کوواریانس فضایی-زمانی مبتنی بر تجزیه تاکر پرداخته خواهد شد و در مطالعه‌ای شبیه‌سازی دقت پیشگویی‌ها مبتنی بر تجزیه CP و تجزیه تاکر مورد ارزیابی و مقایسه قرار می‌گیرد. در انتها نحوه بکارگیری روش ارائه شده برای تحلیل داده‌های سرعت باد ایران نشان داده خواهد شد.

۲ معرفی مدل

فرض کنید سرعت باد در موقعیت فضایی s و لحظه زمانی t با میدان تصادفی $Z(s,t,\theta)$ نشان داده شود، که در آن θ نشان‌دهنده عدم قطعیت موجود در آن است. فرض کنید سرعت باد در موقعیت‌های فضایی $s_i : i = ۱,...,N_s$ و لحظه‌های زمانی $t_j : j = ۱,...,N_t$ اندازه‌گیری شده است و مقادیر آن در بردار

$$Z(\theta)^T = \{Z(s_۱,t_۱,\theta), Z(s_۲,t_۱,\theta),...,Z(s_{N_s},t_۱,\theta), Z(s_۱,t_۲,\theta),...,Z(s_{N_s},t_{N_t},\theta)\}^T$$

ارائه شده است. تابع کوواریانس فضایی میدان در لحظه زمانی t به صورت

$$Cov(Z(s_i,t,\theta), Z(s_j,t,\theta)) = E\{[Z(s_i,t,\theta) - \overline{Z}(s_i,t)][Z(s_j,t,\theta) - \overline{Z}(s_j,t)]\} \qquad (۱.۲)$$

تعریف می‌شود، که ماتریس کوواریانس آن به صورت تانسور مرتبه سه با اندازه $N_sN_t \times N_sN_t$ است و با C نمایش داده می‌شود. اکنون پیشگویی میدان تصادفی بر اساس تابع کوواریانس (۱.۲) طی شش مرحله ارائه می‌شود.

مرحله ۱ـ جمع‌آوری و آماده‌سازی داده‌ها: در صورت وجود روندهای فصلی یا دوره‌ای، داده‌ها روند زودوده شوند.

مرحله ۲ـ برآورد کوواریانس: بر اساس داده‌های روند زودوده ماتریس کوواریانس تجربی یا تانسور محاسبه شود.

مرحله ۳ـ تجزیه تانسور ماتریس کوواریانس: درایه‌های ماتریس تانسور کوواریانس C به صورت

$$c_{i,j,k} = E\{(Z(s_i,t_k,\theta) - \overline{Z}(t_k,\theta))(Z(s_j,t_k,\theta) - \overline{Z}(t_k,\theta))\} \qquad (۲.۲)$$

ایجاد شود.

مرحله ۴ـ تجزیه‌های تانسور CP و تاکر ماتریس کوواریانس C به صورت

$$\widehat{C}_{CP} = [\sum_{r=۱}^{R} \lambda_r \cdot a_r \circ b_r] \circ c \quad , \quad \lambda_r \in R\,,\ a,b \in R^{N_s}\,,\ c \in R^{N_t} \qquad (۳.۲)$$

$$\widehat{C}_{Tucker} = [\sum_{p=۱}^{P}\sum_{q=۱}^{Q}\sum_{r=۱}^{R} g_{pqr} \cdot a_p \circ b_q] \circ c \quad , \quad g_{pqr} \in R\,,\ a,b \in R^{N_s}\,,\ c \in R^{N_t} \qquad (۴.۲)$$

محاسبه شوند، که در آن $\widehat{C}$ تقریبی از تانسور کوواریانس C است.

مرحله ۵ـ خطای تقریب $\widehat{C}$ به صورت

$$er_C = \frac{||C - \widehat{C}||_{Fr}}{||C||_{Fr}} \qquad (۵.۲)$$

محاسبه شود، که در آن $||.||_{Fr}$ نرم فروبنیوس است که به صورت

$$||C||_{Fr} = \sqrt{\sum_{i=۱}^{I_۱}\sum_{j=۱}^{I_۲}\sum_{k=۱}^{I_۳} c_{i,j,k}^{۲}}$$

تعریف می‌شود (کلدا و بادر ، ۲۰۰۹).

مرحله ۶ـ پیش‌گویی میدان تصادفی مبتنی بر دو تجزیه CP و تاکر به ترتیب به صورت

$$Z_{CP}(\theta) = \overline{Z} + \sum_{r=۱}^{R} \sqrt{\lambda_r} a_r \sqrt{c^T}\, \eta_r \quad , \quad \lambda_r \in R\,,\ a \in R^{N_s}\,,\ c \in R^{N_t} \qquad (۶.۲)$$

$$Z_{Tucker}(\theta) = \overline{Z} + \sum_{p=۱}^{P}\sum_{q=۱}^{Q}\sum_{r=۱}^{R} \sqrt{g_{pqr}} a_p \sqrt{c^T}\, \eta_{pqr}\ ,\ g_{pqr} \in R\,,\ a \in R^{N_s},\ c \in R^{N_t} \qquad (۷.۲)$$

محاسبه شود، که در آن η_rها عضوهای یک مجموعه از متغیرهای تصادفی مستقل با میانگین صفر هستند.

۳ مطالعه شبیه‌سازی و تحلیل داده‌های واقعی

در این بخش طی یک مطالعه شبیه‌سازی عملکرد پیشگویی مبتنی بر دو تجزیه CP و تاکر براساس میانگین توان دوم خطای پیش‌گویی مورد ارزیابی و مقایسه عددی قرار می‌گیرد. در این مطالعه ۱۰ موقعیت از بین ۱۰۰ موقعیت جغرافیایی مرز ایران انتخاب شده است. ساختار همبستگی فضایی‌ـ‌زمانی تفکیک‌ناپذیر آن از مدل نایتینگ (۲۰۰۲) به صورت

$$C(h,u) = \left(\frac{\sigma^۲}{۱+u^{\lambda}}\right)\exp^{\frac{-h^{\nu}}{(۱+u^{\lambda})^{۰/۵\gamma\nu}}}, \qquad \lambda,\nu \in [۰,۲]\,,\gamma \in [۰,۱]\,.$$

در نظر گرفته شده است، که در آن h و u به ترتیب تأخیرهای فضایی و زمانی هستند. بر اساس این تابع کوواریانس و موقعیت‌های در نظر گرفته شده یک میدان تصادفی گاوسی تولید می‌شود. هر یک از موقعیت‌ها به ترتیب کنار گذاشته شده و میانگین توان دوم خطای اعتبارسنجی متقابل (CVMSE) پیش‌گویی‌ها مبتنی بر تجزیه‌های CP و تاکر به ازای مقادیر مختلف $\lambda = ۰/۱$ ، $\nu = ۰/۱, ۱, ۲$ و $\sigma^۲ = ۰/۵, ۱, ۲$ محاسبه شده است. با مقایسه مقادیر میانگین توان دوم خطای اعتبارسنجی متقابل پیش‌گویی‌ها به ازای مقادیر $\nu = ۰/۱, ۱, ۲$ همواره پیش‌گویی مبتنی بر تجزیه تاکر نسبت به تجزیه CP از دقت بیش‌تری برخوردار است.

در ادامه، داده‌های متوسط سرعت باد روزانه از سال ۲۰۰۰ تا ۲۰۱۶ در سی و یک ایستگاه بادسنجی کشور که از سایت https://www.tutiempo.net/clima قابل دسترسی است، تحلیل آماری شده و با استفاده از مدل ارائه شده اقدام به پیش‌گویی فضایی‌ـ‌زمانی متوسط هفتگی سرعت باد می‌شود. برای این منظور با حذف روند فصلی داده‌ها، تانسور کوواریانس با درایه‌های (۲.۲) بر اساس داده‌ها برآورد شده و به‌صورت‌های (۳.۲) و (۴.۲) تجزیه شده اند. سپس با استفاده از نرم افزار Mathlab مقدار $R = ۵۲$ برای رتبه تجزیه CP و رتبه $R = [p,q,r] = [۲۴, ۲۴, ۴۹]$ برای تجزیه تاکر حاصل شده است. خطای نسبی بین تانسور C و $\hat{C}$ بر طبق (۵.۲) برای تجزیه‌های تاکر و CP به ترتیب ۰/۰۲ و ۰/۱ شده است. همان‌طور که ملاحظه می‌شود، تقریب تانسور C با تجزیه تاکر خطای کمتری نسبت به تقریب حاصل از تجزیه CP دارد. برای پیشگویی میدان تصادفی از روابط (۶.۲) و (۷.۲) استفاده شده است، که در آن η_r دارای توزیع نرمال استاندارد منظور گردیده است. نتایج پیشگویی متوسط سرعت باد در پنج هفته اول سال ۲۰۱۶ همه ایستگاه‌ها در جدول ۲ ارائه شده است. همان‌طورکه ملاحظه می‌شود مقادیر میانگین توان دوم خطای اعتبارسنجی متقابل پیش‌گویی‌های مبتنی بر تجزیه تاکر کوچکتر از مقادیر متناظر حاصل از تجزیه CP هستند، بنابراین پیش‌گویی‌های مبتنی بر تجزیه تاکر از

جدول ۱: میانگین توان دوم خطای اعتبارسنجی متقابل برای پیش‌گویی‌ها براساس تجزیه‌های CP و تاکر

		تجزیه CP			تجزیه تاکر		
ν	موقعیت	$\sigma^2=۰/۵$	$\sigma^2=۱$	$\sigma^2=۲$	$\sigma^2=۰/۵$	$\sigma^2=۱$	$\sigma^2=۲$
	۱	۰/۴۵	۱/۰۵	۱/۲۳	۰/۴۲	۰/۹۶	۱/۱۳
	۲	۰/۶۰	۱/۴۴	۲/۴۴	۰/۳۷	۱/۳۶	۱/۸۰
	۳	۰/۸۷	۱/۱۳	۲/۰۳	۰/۸۴	۰/۹۷	۱/۵۸
	۴	۰/۴۷	۰/۶۴	۲/۹۰	۰/۴۲	۰/۵۳	۱/۵۲
۰/۱	۵	۰/۵۴	۱/۱۷	۳/۲۱	۰/۳۹	۰/۹۵	۳/۰۷
	۶	۰/۶۹	۱/۲۲	۲/۰۹	۰/۴۷	۰/۸۱	۱/۲۶
	۷	۰/۴۰	۰/۹۰	۰/۹۸	۰/۳۷	۰/۷۱	۰/۹۶
	۸	۰/۶۷	۰/۸۲	۲/۰۵	۰/۴۵	۰/۷۱	۱/۰۲
	۹	۰/۶۲	۱/۷۴	۳/۰۶	۰/۴۴	۱/۱۰	۱/۲۹
	۱۰	۰/۶۹	۰/۸۵	۱/۳۰	۰/۶۷	۰/۷۴	۰/۹۳
	۱	۰/۴۶	۱/۱۲	۱/۳۳	۰/۴۴	۰/۹۶	۱/۲۰
	۲	۰/۹۳	۱/۰۷	۱/۹۸	۰/۱۰	۰/۹۸	۱/۴۷
	۳	۰/۶۵	۱/۱۰	۲/۴۷	۰/۵۶	۱/۰۲	۱/۲۸
	۴	۰/۶۲	۰/۸۶	۱/۸۴	۰/۵۸	۰/۷۹	۱/۲۶
۱	۵	۰/۲۶	۰/۵۹	۱/۱۹	۰/۱۱	۰/۴۹	۰/۹۴
	۶	۰/۵۸	۱/۰۸	۱/۶۱	۰/۳۸	۰/۶۶	۱/۲۸
	۷	۰/۴۶	۱/۱۷	۲/۰۳	۰/۱۶	۰/۹۲	۱/۲۳
	۸	۰/۸۶	۲/۱۷	۲/۵۲	۰/۱۴	۱/۰۱	۱/۸۱
	۹	۰/۵۰	۱/۲۴	۳/۰۳	۰/۴۴	۰/۸۴	۱/۴۸
	۱۰	۰/۸۵	۱/۷۹	۲/۵۳	۰/۷۰	۱/۰۹	۱/۹۳
	۱	۰/۹۶	۲/۳۰	۲/۳۶	۰/۷۴	۱/۴۶	۱/۸۵
	۲	۱/۲۰	۱/۳۴	۲/۵۵	۰/۶۶	۱/۱۸	۱/۷۲
	۳	۱/۳۰	۱/۵۵	۱/۸۴	۰/۸۹	۰/۹۷	۱/۶۶
	۴	۱/۷۰	۲/۵۲	۲/۹۱	۰/۴۴	۱/۰۶	۱/۸۰
۲	۵	۲/۰۵	۲/۶۶	۳/۵۰	۱/۴۲	۱/۵۵	۱/۸۳
	۶	۱/۳۴	۱/۶۷	۳/۳۸	۰/۸۴	۱/۵۵	۲/۳۴
	۷	۱/۰۶	۱/۷۲	۲/۲۵	۰/۷۰	۱/۶۲	۱/۸۴
	۸	۱/۹۱	۲/۱۴	۲/۲۰	۱/۳۳	۱/۴۳	۱/۷۰
	۹	۱/۸۴	۲/۷۴	۳/۳۵	۰/۶۲	۱/۷۰	۲/۶۶
	۱۰	۱/۰۹	۱/۸۰	۲/۶۷	۰/۷۲	۱/۶۱	۲/۰۵

دقت بیش‌تری نسبت به تجزیه CP برخوردار هستند. باید توجه شود که تغییرپذیری زیادی که در سرعت باد پیش‌گویی شده گزارش شده است، ناشی از مدل پیش‌گویی نیست، بلکه به دلیل عدم قطعیت موجود در داده‌های واقعی است که توسط پارامتر θ در مدل (۲.۲) قابل بررسی است. به عبارت دیگر این تغییرپذیری زیاد موجود در داده‌های واقعی است که در پیش‌گویی‌ها نیز منعکس شده است.

[3] Threshold value

جدول ۲: پیش‌گویی متوسط سرعت باد پنج هفته اول سال ۲۰۱۶

ردیف	نام ایستگاه	مقدار واقعی	تجزیه CP		تجزیه تاکر	
			پیش‌گویی	CVMSE	پیش‌گویی	CVMSE
۱	آذربایجان شرقی	۲/۷۷	۲/۱۳	۱/۵۶	۲/۴۲	۰/۴۴
۲	آذربایجان غربی	۲/۱۳	۲/۴۰	۲/۱۳	۲/۲۰	۱/۲۲
۳	اردبیل	۱/۲۷	۱/۴۴	۱/۲۳	۱/۲۹	۱/۱۱
۴	اصفهان	۱/۷۳	۱/۲۱	۱/۱۴	۱/۸۳	۰/۸۵
۵	البرز	۲/۶۲	۲/۵۴	۱/۶۶	۲/۵۸	۱/۶۰
۶	ایلام	۲/۳۷	۲/۱۴	۲/۱۲	۲/۳۰	۰/۷۵
۷	بوشهر	۲/۳۵	۲/۵۶	۱/۹۹	۲/۳۰	۰/۹۱
۸	تهران	۱/۷۶	۱/۶۶	۱/۱۶	۱/۷۰	۰/۷۲
۹	چهارمحال و بختیاری	۲/۶	۲/۳۳	۲/۳۲	۲/۷۲	۱/۲۴
۱۰	خراسان جنوبی	۱/۸۱	۱/۱۴	۱/۰۹	۱/۷۶	۰/۷۸
۱۱	خراسان رضوی	۲/۷۶	۲/۸۲	۲/۴۶	۲/۸۰	۱/۴۲
۱۲	خراسان شمالی	۲/۰۲	۲/۳۲	۱/۹۱	۱/۹۶	۱/۸۴
۱۳	خوزستان	۲/۸۴	۳/۰۳	۲/۳۳	۲/۹۲	۱/۱۱
۱۴	زنجان	۲/۷۲	۲/۴۱	۱/۷۷	۲/۸۰	۰/۹۴
۱۵	سمنان	۲/۲۷	۲/۳۲	۲/۱۱	۲/۲۳	۱/۲۵
۱۶	سیستان و بلوچستان	۱/۴۲	۱/۷۵	۱/۳۳	۱/۵۰	۰/۳۲
۱۷	فارس	۲/۹۴	۳/۰۹	۲/۸۹	۲/۹۸	۱/۲۵
۱۸	قزوین	۳/۱۵	۳/۴۸	۲/۱۳	۳/۲۰	۰/۶۱
۱۹	قم	۲/۹۳	۲/۸۶	۱/۹۸	۲/۸۶	۱/۸۸
۲۰	کردستان	۲/۸۱	۲/۸۱	۱/۲۵	۲/۸۰	۰/۵۲
۲۱	کرمان	۱/۷۳	۱/۶۹	۱/۲۵	۱/۷۷	۱/۰۹
۲۲	کرمانشاه	۲/۶۳	۲/۳۲	۱/۷۹	۲/۶۵	۱/۶۷
۲۳	کهگیلویه و بویراحمد	۱/۰۴	۱/۳۷	۱/۰۵	۱/۰۵	۰/۹۷
۲۴	گلستان	۱/۲۴	۱/۷۵	۱/۶۶	۱/۵۱	۱/۵۳
۲۵	گیلان	۷/۳۸	۶/۷۵	۱/۸۶	۷/۵۴	۱/۵۶
۲۶	لرستان	۲/۴۷	۲/۸۰	۱/۴۱	۲/۶۷	۰/۸۹
۲۷	مازندران	۲/۹۷	۲/۷۳	۱/۳۰	۲/۸۱	۰/۶۳
۲۸	مرکزی	۲/۰۳	۲/۱۳	۱/۳۶	۲/۱۸	۱/۲۴
۲۹	هرمزگان	۲/۳۶	۲/۲۰	۱/۶۱	۲/۳۰	۰/۸۴
۳۰	همدان	۳/۹۸	۳/۶۷	۲/۵۹	۳/۹۳	۰/۹۵
۳۱	یزد	۲/۳۷	۲/۴۵	۱/۲۵	۲/۴۰	۱/۱۱

برای مدیریت مؤثر تولید و توزیع انرژی، فهم ماهیت احتمالاتی خروجی انرژی ضروری است. کمیت‌هایی از قبیل احتمال بیشتر شدن یا کمتر شدن توان خروجی انرژی از یک مقدار آستانه[۳] جذابیت ویژه‌ای برای بررسی خواهد داشت. با این انگیزه پیش‌گویی‌های سرعت باد را برای پیش‌گویی توان خروجی انرژی مورد استفاده قرار داده و در جدول ۳ ارائه شده است. توان خروجی انرژی باد به صورت $P = \frac{۰/۵۹۳}{۲}\rho u^۳$ برآورد می‌شود، که در آن P بیانگر توان خروجی انرژی باد بر

حسب W/m^2، ۱/۲۲۷ $= \rho$ بر حسب kg/m^3 چگالی هوا و u سرعت آنی باد در یک ایستگاه خاص بر حسب متر بر ثانیه است. برای خروجی‌های انرژی، آستانه منتخب دلخواه $120 W/m^2$ و $240 W/m^2$، احتمال‌های خروجی انرژی واقعی کمتر از ۱۲۰ و بیشتر از ۲۴۰ محاسبه می‌شود که کمترین و بیشترین آستانه یک توربین برای توان خروجی است. برآوردهای هفتگی این احتمال‌ها در جدول ۳ ارائه شده است. برای افزایش ظرفیت تولید برق بادی در ایستگاه‌هایی که توان خروجی آن‌ها زیاد است می‌توان مدیریت صحیح و مؤثری برای آن‌ها به‌کار گرفت.

جدول ۳: برآورد احتمال وقوع متوسط خروجی انرژی هفتگی در پنج هفته اول سال ۲۰۱۶

ردیف	نام ایستگاه	$P_r(۱۲۰ < Power < ۲۴۰)$	ردیف	نام ایستگاه	$P_r(۱۲۰ < Power < ۲۴۰)$
۱	آذربایجان شرقی	۰/۳۳	۱۶	سیستان و بلوچستان	۰/۵۵
۲	آذربایجان غربی	۰/۶۱	۱۷	فارس	۰/۲۵
۳	اردبیل	۰/۷۲	۱۸	قم	۰/۵
۴	اصفهان	۰/۳	۱۹	کردستان	۰/۵۰
۵	البرز	۰/۸۷	۲۰	کرمان	۰/۶۶
۶	ایلام	۰/۷۱	۲۱	کرمانشاه	۰/۹۱
۷	بوشهر	۰/۶۸	۲۲	کهگیلویه و بویراحمد	۰/۴
۸	تهران	۰/۷	۲۳	گلستان	۰/۳۳
۹	چهارمحال و بختیاری	۰/۷۸	۲۴	گیلان	۰/۶
۱۰	خراسان جنوبی	۰/۵۱	۲۵	لرستان	۰/۷۰
۱۱	خراسان رضوی	۰/۴۸	۲۶	مازندران	۰/۴
۱۲	خراسان شمالی	۰/۲۱	۲۷	مرکزی	۰/۶۸
۱۳	خوزستان	۰/۷۷	۲۸	هرمزگان	۰/۷۴
۱۴	زنجان	۰/۷۱	۲۹	همدان	۰/۲۶
۱۵	سمنان	۰/۴۱	۳۰	یزد	۰/۰۶

بحث و نتیجه‌گیری

برای پیشگویی فضایی‌ـزمانی در یک میدان تصادفی نامانا و افزایش سرعت محاسبات، از تانسور کوواریانس فضایی‌ـزمانی مبتنی بر دو تجزیه تاکر و CP مورد بررسی قرار گرفت و نشان داده شد خطاهای محاسباتی کاهش، دقت پیشگویی ها افزایش و سرعت محاسبات افزایش می‌یابد. بعلاوه نشان داده شد پیشگویی‌های مبتنی بر تجزیه تاکر از دقت بیشتری نسبت به تجزیه CP برخوردار هستند.

مراجع

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

عصمتی، م و محمدزاده، م. (۱۳۹۶)، پیش‌گویی فضایی‌ـزمانی میدان‌های تصادفی نامانا و تفکیک‌ناپذیر، *دومین سمینار آمار فضایی و کاربردهای آن*، دانشگاه صنعتی شاهرود.

Gneiting, T. (2002), Nonseparable, Stationary Covariance Functions for Space-Time Data, *Journal of the American Statistical Association*, **97**, 590–600.

Hitchcock, F. L. (1927). The Expression of a Tensor or a Polyadic as a Sum of Products, *Journal of Mathematical Physics*, **6**, 164–189.

Kolda T.G. and Badder B.W. (2009), Tensor Decompositions and Applications, *SIAM Review*, **51**, 455–500.

Kruskal, J. B. (1977), Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics, *Linear Algebra and its Applications*, **18**, 95–138.

Lynch, P. (2008), The Origins of Computer Weather Prediction and Climate Modeling, *Journal of Computational Physics*, **227**, 3431–3444.

Salcedo-Sanz S., Perez-Bellido, A.M., Ortiz-Garca, E.G., Portilla-Figueras, A., Prieto, L. and Correoso, F. (2009), Accurate Short-Term Wind Speed Prediction by Exploiting Diversity in Input Data Using Banks of Artificial Neural Networks, *Neurocomputing*, **72**, 1336–1341.

Soman,S., Zareipour, H., Malik, O. and Mandal, P. (2010), A Review of Wind Power and Wind Speed Forecasting Methods with Different Time Horizons, *North American Power Symposium (NAPS)*, Arlington, TX, 26–28 September, 1–8.

Tucker, L. R. (1963) Implications of Factor Analysis of Three-Way Matrices for Measurement of Change, in Problems in Measuring Change, C. W. Harris, ed., *University of Wisconsin Press*, 122–137.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

مقایسه کارائی آماری و محاسباتی روش‌های برآوردیابی در مدل‌های نیمه‌طیفی فضایی-زمانی

علی شهنواز[۱]، علی محمدیان مصمم[۲]، محمد حسن بهزادی[۱۳]

[۱]گروه آمار، دانشگاه آزاد اسلامی، واحد زنجان

[۲]گروه آمار، دانشگاه زنجان

[۳]گروه آمار، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات، تهران

چکیده: : در این مقاله، ابتدا مبانی نظری مدل‌سازی نیمه‌طیفی بررسی و به توصیف چند خاصیت از مدل‌های نیمه‌طیفی اخیر پرداخته می‌شود. سپس یک روش برای برآورد تابع کوواریانس فضایی-زمانی در حالت نیمه‌طیفی پیشنهاد شده است. به منظور ارزیابی عملکرد مدل‌های نیمه‌طیفی ارائه شده، دو شبیه‌سازی انجام و در هرکدام از آن‌ها روش برآورد پیشنهادی با سایر روش‌ها مقایسه شده است. روش مورد نظرعملکرد بهتری نسبت به سایر روش‌ها به‌خصوص برای مجموعه مشاهدات بزرگ دارد. سرانجام برای تحلیل داده‌های واقعی مربوط به متوسط روزانه سرعت باد در استان زنجان از روش پیشنهادی استفاده شده است.

واژه‌های کلیدی: کوواریانس فضایی-زمانی، مدل فضایی-زمانی، مدل نیمه‌طیفی، درستنمایی وایتل.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M15، 62M30.

۱ مقدمه

در آمار کلاسیک معمولاً داده‌ها ناهمبسته فرض می‌شوند ولی در عمل موارد زیادی وجود دارند که این فرض ناهمبستگی نقض شده است. به عنوان مثال در سری‌های زمانی شاموای و استافر (۲۰۱۷) مشاهدات دارای نوعی همبستگی از طریق زمان رخداد آنها هستند. همچنین در علوم مربوط به زمین آمار (چیلس و دلفنر ، ۲۰۱۲)، محیط‌زیست، هواشناسی، همه‌گیرشناسی و اقیانوس‌شناسی ملاحظه می‌شود که داده‌ها دارای نوعی همبستگی از نظر موقعیت مشاهدات هستند. به چنین

[1]Spatial Data

[۱]نام و ایمیل ارائه دهنده مقاله: علی شهنواز، shahnavaz.ali2000@yahoo.com

داده‌هایی که همبستگی آنها ناشی از موقعیت جغرافیایی است، داده‌های فضایی [۱] گفته می‌شود (کرسی و ویکل ، ۲۰۱۳). داده‌هایی که هم از نظر موقعیت فضایی و هم از نظر زمانی وابسته باشند، داده‌های فضایی‌ـ‌زمانی [۲] نامیده می‌شوند. برای تحلیل سری‌های زمانی و داده‌های فضایی معمولاً دو استراتژی وجود دارد: یکی تحلیل سری‌های زمانی و داده‌های فضایی در قلمرو مشاهدات و دیگری تحلیل سری‌های زمانی و داده‌های فضایی در قلمرو فرکانس. در دو دهه اخیر مطالعات زیادی برای تعیین ساختار همبستگی، مدل‌بندی و تحلیل داده‌ها انجام شده است. تحلیل این گونه داده‌ها به روش اول مستلزم تعیین ساختار همبستگی به ترتیب، زمانی، فضایی و فضایی‌ـ‌زمانی از طریق تابع کواریانس آنها است. این تابع نقش به سزایی در پیش‌گویی موقعیت‌های فضایی یا زمانی فاقد مشاهده دارد. تحلیل داده‌ها به روش دوم مستلزم تعیین تابع چگالی طیفی زمانی، فضایی و فضایی‌ـ‌زمانی است. توابع کواریانس و توابع چگالی طیفی زمانی و فضایی محض بسیار زیادی در متون آماری وجود دارند (چیلس و دلفنر ، ۲۰۱۲؛ کرسی ، ۱۹۹۳). ولی برای ساخت توابع کواریانس فضایی‌ـ‌زمانی مطالعات علمی کمتری انجام گرفته است. یکی از ویژگی‌های مهم تابع کواریانس، ویژگی معین مثبت [۳] است و تابعی که چنین شرطی را داشته باشد، تابع کواریانس معتبر نامیده می‌شود.

ساخت توابع کواریانس فضایی‌ـ‌زمانی معتبر تا حدودی امری مشکل و در طی چند سال گذشته یکی از موضوعات علمی آمارشناسان دنیا بوده است. به این منظور برخی فرض‌های ساده‌کننده نظیر، مانایی [۴]، همسانگردی [۵]، کاملاً متقارن [۶] و تفکیک‌پذیری [۷] مورد استفاده قرار می‌گیرد (کرسی ، ۱۹۹۳). یکی از روش‌ها برای ساخت توابع کواریانس فضایی‌ـ‌زمانی معتبر استفاده از نمایش طیفی [۸] تابع کواریانس است که بر اساس قضیه‌ی بوخنر (۱۹۵۵) تعیین می‌شود. این قضیه به درستی تحلیل در قلمرو مشاهدات را به تحلیل در قلمرو فرکانس مرتبط می‌سازد. بوخنر نشان داد که یک تابع پیوسته، معین مثبت است اگر و تنها اگر تبدیل فوریه‌ی یک اندازه نامنفی متناهی باشد. از آن‌جا که توابع کواریانس فضایی‌ـ‌زمانی هر میدان تصادفی مانای مرتبه دوم، موجود و معین مثبت است، براساس قضیه بوخنر رده توابع کواریانس فضایی‌ـ‌زمانی مانا روی $R^d \times R$ معادل رده تبدیل فوریه‌ی اندازه‌های نامنفی و متناهی روی $R^d \times R$ است، لذا با توجه به این قضیه، نمایش طیفی تابع کواریانس بر اساس تابع چگالی طیفی عبارت است از:

$$K(s,t) = \iint g(\boldsymbol{\lambda},\omega)\ \exp\{i\mathbf{s}'\boldsymbol{\lambda} + it\omega\}\ d\boldsymbol{\lambda}d\omega, \quad (\mathbf{s},t) \in R^d \times R \qquad (۱.۱)$$

که در آن s و t به ترتیب تاخیرهای فضایی و زمانی و همچنین $\boldsymbol{\lambda} \in R^d$ و $\omega \in R$ فرکانس‌های فضایی و زمانی هستند و نماد " $'$ " نشانگر ترانهاده‌ی یک بردار است. بنابراین، با اختیار کردن تابع چگالی طیفی می‌توان یک تابع کواریانس فضایی‌ـ‌زمانی معتبر بنا نهاد.

در تحلیل داده‌های فضایی‌ـ‌زمانی علاوه بر تحلیل در قلمرو مشاهدات و قلمرو فرکانس می‌توان قلمرو جدیدی در مابین این دو تعریف کرد که به اصطلاح نیمه‌طیفی [۹] نامیده می‌شود. از مهمترین مزایای این روش استفاده از انتگرال صرفاً فضایی به جای انتگرال چندگانه فضایی‌ـ‌زمانی در رابطه (۱.۱) است. کرسی و هوانگ (۱۹۹۹) با فرض انتگرال‌پذیری تابع کواریانس و به کارگیری قضیه بوخنر، روشی برای ساخت کواریانس‌های تفکیک‌ناپذیر مانا ارائه کردند که روش آنان محدود به رده‌ای از توابع می‌شود که انتگرال dـ‌گانه تبدیل فوریه آنها به صورت تحلیلی قابل حل باشد (کاملر ، ۲۰۰۷). گنتینگ (۲۰۰۲) برای غلبه بر انتگرال صرفاً فضایی فوق با استفاده از توابع کاملاً یکنوا و برنشتاین رده‌ای از مدل‌های

[2] Spatio-Temporal Data
[3] Positive Definite
[4] Stationary
[5] Lsotropic
[6] Completely Symmetric
[7] Separable
[8] Spectral Representation
[9] Half Spectral

کواریانس فضایی‌ـزمانی تفکیک‌ناپذیر ارائه کرد. علیرغم سادگی بسیار زیاد روش گنیتینگ، تابع کواریانس حاصل یک تابع کواریانس کاملاً متقارن است. علاوه بر آن کنت و همکاران (۲۰۱۱) نشان دادند که تابع کواریانس ارائه شده توسط گنیتینگ دارای مشکل گودی [۱۰] هست، به این معنی که به ازای برخی مقادیر تاخیر فضایی، تابع کواریانس زمانی مربوطه به صورت یکنوا نزولی نیست. امیدی و محمدزاده (۲۰۱۵)، امیدی و محمدزاده (۱۳۹۲) با استفاده از توابع مفصل برای حل مشکل گودی راهکاری ارائه نموده‌اند. هورل و استاین (۲۰۱۷) با استفاده از نمایش نیمه‌طیفی تابع کواریانس فضایی‌ـزمانی مدل‌های گاوسی فضایی‌ـزمانی جدیدی را ارائه کرده‌اند.

در این مقاله چند مدل نیمه‌طیفی معرفی و از بین آنها بهترین مدل برای برازش به داده‌های میانگین سرعت باد روزانه استان زنجان انتخاب و با استفاده از روش ماکسیمم درستنمایی چند متغیره وایتل، پارامترهای مدل انتخابی برآورد شده‌اند.

۲ مدل‌های نیمه‌طیفی

نمایش نیمه‌طیفی تابع کواریانس $K(\mathbf{s},t)$ را می‌توان با انتخاب ترتیب انتگرال‌گیری نسبت به $\boldsymbol{\lambda}$ یا ω به صورت نیمه‌طیفی زمانی یا نیمه‌طیفی فضایی بدست آورد. در این مقاله نمایش نیمه‌طیفی در زمان به صورت زیر مورد بحث قرار خواهد گرفت.

$$K(\mathbf{s},t)=\int_R H(\mathbf{s},\omega)\ \exp(it\omega)\ d\omega,$$

که در آن:

$$H(\mathbf{s},t)=f(\omega)\mathbb{C}(\mathbf{s}\delta(\omega))\ \exp(i\theta(\omega)\phi'\mathbf{s}), \qquad (۱.۲)$$

که در آن f تابع چگالی طیفی صرفاً زمانی، $\mathbb{C}$ تابع همبستگی معتبر در R^d با یک تابع چگالی طیفی، δ تابعی زوج و مثبت که نشان‌دهنده اثر تعامل فضایی‌ـزمانی، $\theta(\cdot)$ تابعی حقیقی و فرد و ϕ برداریکه در R^d است (استاین ، ۲۰۰۵).

۱.۲ چند خاصیت از مدل‌های نیمه‌طیفی

مدل معرفی شده در رابطه (۱.۲) در حالت کلی نامتقارن است. در این مقاله ابتدا حالت فضایی‌ـزمانی کاملاً متقارن این مدل یعنی حالتی که $\theta(\omega)=۰$ در نظر گرفته می‌شود. با این فرض، مدل نیمه‌طیفی رابطه (۱.۲) به صورت $f(\omega)\mathbb{C}(\mathbf{s}\delta(\omega))$ است. هورل و استاین (۲۰۱۷) نشان دادند که بسیاری از خواص مدل‌سازی $K(\mathbf{s},t)$ مستقل از θ یا ϕ حفظ می‌شود. خاصیت تعامل فضایی‌ـزمانی با استفاده از δ تعیین می‌شود. لازم به ذکر است وقتی δ مقدار ثابت باشد توابع کواریانس حاشیه‌ای فضایی و زمانی مستقل از هم تعیین می‌شوند. در این حالت $K(\mathbf{s},t)$ متناسب با حاصل‌ضرب توابع کواریانس فضایی و زمانی است یعنی

$$K(\mathbf{s},t)\propto K(\mathbf{s},۰)K(۰,t)$$

این مدل‌ها را مدل‌های تفکیک‌پذیر (رودریگاز و مجیا ، ۱۹۷۴) گویند که به علت عدم تطابق با واقعیت داده‌ها، مورد انتقاد هستند. همان‌طور که در بخش بعد مشاهده خواهد شد در مدل‌های نیمه‌طیفی در تابع کواریانس گودی ظاهر می‌شود (کنت و همکاران ، ۲۰۱۱). به این معنی که اگر فرض کنیم s_{here} نشان‌دهنده موقعیت جاری و t_{now} نیز زمان جاری باشد و همبستگی $Z(s_{here},t_{now})$ را با مقدار فرایند در یک همسایگی از موقعیت s_{there} در دو زمان ممکن t_{now} و $t_{yesterday}$

[10] Dimple

در نظر بگیریم، در این حالت همبستگی با $Z(s_{there}, t_{yesterday})$ یا معادل آن با $Z(s_{there}, t_{tomorrow})$ ممکن است بزرگ‌تر یا کوچک‌تر از همبستگی با $Z(s_{there}, t_{now})$ باشد.

استاین (۲۰۱۱) شرط زیر را برای تابع چگالی طیفی فضایی‌ـزمانی g، به عنوان یکی از شرایط مطلوب برای توابع کاملاً طیفی، ذکر کرده است. با برقراری این شرط، خاصیت غربالگری Effect Screening در این مدل‌ها تضمین می‌شود (ژورنل و هوجبرگتس ، ۱۹۷۸).

$$\lim_{(\boldsymbol{\lambda},\omega)\to\infty} \sup_{(u,v)<R} \left| \frac{g(\boldsymbol{\lambda}+\nu, \omega+u)}{g(\boldsymbol{\lambda},\omega)} - ۱ \right| = ۰, \tag{۲.۲}$$

که در آن R می‌تواند هر مقدار متناهی باشد. لازم به ذکر است که مدل نیمه‌طیفی رابطه (۱.۲) در حالت کلی در شرط (۲.۲) صدق نمی‌کند. فرض کنید $\mathbf{h}$ چگالی طیفی تابع کواریانس فضایی $\mathbb{C}$ باشد پس نمایش کاملاً طیفی مدل استاین (۲۰۰۵) به صورت زیر است (مصمم و کنت ، ۲۰۱۶):

$$K(\mathbf{s}, t) = \iint_{R^d R} f(\omega)\delta(\omega)^{-d}\, h\left(\frac{\boldsymbol{\lambda} - \theta(\omega)\phi}{\delta(\omega)}\right) \exp(is'\boldsymbol{\lambda} + it\omega)\, d\boldsymbol{\lambda} d\omega.$$

در حالتی که $g(\boldsymbol{\lambda}, \omega)$ یک تابع چگالی طیفی کاملاً متقارن در نظر گرفته شود خواهیم داشت:

$$g(\boldsymbol{\lambda}, \omega) = f(\omega)\delta(\omega)^{-d} h\left(\frac{\boldsymbol{\lambda}}{\delta(\omega)}\right). \tag{۳.۲}$$

بنابراین نمایش کلی g را می‌توان به صورت $g(\boldsymbol{\lambda} - \theta(\omega)\phi, \omega)$ نوشت. در قضیه ۱.۲ شرایط لازم برای اینکه $g(\boldsymbol{\lambda}, \omega)$ یا $g(\boldsymbol{\lambda} - \theta(\omega)\phi, \omega)$ در شرط (۲.۲) صدق کند بیان شده است.

قضیه ۱.۲ (هورل و استاین ، ۲۰۱۷). *فرض کنید $g(\boldsymbol{\lambda}, \omega)$ تابع چگالی طیفی تابع کواریانس فضایی‌ـ زمانی $K(\boldsymbol{s}, t)$ باشد که در آن $\boldsymbol{\lambda}$ فرکانس فضایی و ω فرکانس زمانی است.*

الف. فرض کنید $\theta(\cdot)$ به صورت موضعی کراندار باشد. چگالی طیفی $g(\boldsymbol{\lambda}, \omega)$ در (۲.۲) صدق می‌کند اگر و فقط اگر $g(\boldsymbol{\lambda} - \theta(\omega)\phi, \omega)$ در (۲.۲) صدق کند.

ب. فرض کنید $g(\boldsymbol{\lambda}, \omega)$ اکیداً مثبت باشد و در (۲.۲) صدق کند. اگر $g(\boldsymbol{\lambda} - \theta(\omega)\phi, \omega)$ در (۲.۲) صدق کند، آنگاه باید $\theta(\omega)$ به صورت موضعی کراندار باشد.

در واقع چون استراتژی‌های ساخت مدل اغلب با مدل‌های کاملاً متقارن شروع می‌شوند، به کمک این قضیه، استراتژی ساخت مدل را می‌توان به مدل‌های نامتقارن تعمیم داد. حال با این ذهنیت از قضیه ۱.۲ بر روی مدل‌های نیمه‌طیفی کاملاً متقارن تمرکز خواهد شد. هورل و استاین هورل و استاین (۲۰۱۷) نشان دادند برای اینکه $g(\boldsymbol{\lambda}, \omega)$ در شرط (۲.۲) صدق کند باید طیف h و تابع δ به صورت

$$h\left(\frac{\boldsymbol{\lambda}}{\delta(\omega)}\right) = H(\boldsymbol{\lambda}, \omega)\delta(\omega)^d \,/\, f(\omega),$$

قابل تجزیه باشند که در آن H در شرط (۲.۲) صدق می‌کند. بنابراین نمایش ساده‌تری به صورت $g(\boldsymbol{\lambda}, \omega) = p(\omega)q(\boldsymbol{\lambda}, \omega)$ را می‌توان برای g در نظر گرفت. در ادامه شرایطی را که p و q باید داشته باشند تا g در شرط (۲.۲) صدق کند ذکر خواهد شد.

محدودیت ۱.۲. *فرض کنید نمایش طیفی تابع کواریانس K شکلی به صورت $g(\boldsymbol{\lambda},\omega)=p(\omega)q(\boldsymbol{\lambda},\omega)$ داشته باشد در اینصورت دو گزاره زیر برای چگالی‌های طیفی که به صورت فوق پارامتری شده‌اند برقرار هستند:*

الف. اگر g و q در شرط (۲.۲) صدق کنند آنگاه باید p نسبت به q ثابت باشد.

ب. اگر g در شرط (۲.۲) صدق کند، آنگاه برای هر نقطه ثابت ω_0 که $p(\omega_0)$ متناهی و مثبت باشد، طیف حاشیه‌ای $g^*_{\omega_0}(\boldsymbol{\lambda})=g(\boldsymbol{\lambda},\omega_0)$ باید در شرط (۲.۲) صدق کند.

پیامد این محدودیت این است که برای تعیین مدل‌هایی که در شرط (۲.۲) صدق می‌کنند، تابع δ را نمی‌توان مستقل از f و $\mathbb{C}$ در نظر گرفت.

محدودیت ۲.۲. *فرض کنید $g(\boldsymbol{\lambda},\omega)=p(\omega)q(\boldsymbol{\lambda}/\delta(\omega))$ که در آن $p(\omega)$ برای تمام $|\omega|$ های به اندازه کافی بزرگ مثبت و متناهی است، q یک تابع انتگرال‌پذیر، پیوسته و نامنفی و $\delta(\omega)$ زوج با حد خوش تعریف وقتیکه $|\omega|\to\infty$. اگر در (۲.۲) صدق کند آنگاه $\lim_{|\omega|\to\infty}\delta(\omega)=\infty$.*

محدودیت های ۱.۲ و ۲.۲ شکل δ و h را تعیین می‌کنند.

۲.۲ رده‌ای از مدل‌های نیمه‌طیفی

در این بخش در قالب چند مثال، نمونه‌هایی از مدل‌های نیمه‌طیفی کاملاً متقارن معرفی می‌شوند.

مثال ۱.۲. مدل تفکیک‌پذیر (کرسی و هوانگ ، ۱۹۹۹) با تابع کواریانس حاشیه‌ای نمایی : فرض کنید $f(\omega)=\mathbb{C}(\mathbf{s}\delta(\omega))=\exp\{-b\parallel\mathbf{s}\parallel\}$، $\delta(\omega)=1$ است. همچنین بردار پارامترها به صورت $\theta=(a,b,\sigma^2)'$ را در نظر بگیرید که در آن $a\geq 0$ پارامتر مقیاس زمان [11]، $b\geq 0$ پارامتر مقیاس فضا [12] و σ^2 واریانس میدان تصادفی (ازاره) است. در اینصورت مدل نیمه‌طیفی متناظر با آن به صورت

$$H(\mathbf{s},\omega)=f(\omega)\mathbb{C}(\mathbf{s}\delta(\omega))=2a\sigma^2(a^2+\omega^2)^{-1}\exp\{-b\parallel\mathbf{s}\parallel\} \qquad (۴.۲)$$

خواهد بود و تابع کواریانس فضایی_زمانی مربوطه به صورت

$$K(\mathbf{s},t)=\sigma^2\exp(-a|t|-b\parallel\mathbf{s}\parallel)$$

است. با استفاده از رابطه (۳.۲) تابع چگالی طیفی فضایی_زمانی به صورت زیر خواهد بود:

$$g(\boldsymbol{\lambda},\omega)=\sigma^2\left(\frac{2a}{a^2+\omega^2}\right)\left(\frac{2b}{b^2+\|\boldsymbol{\lambda}\|^2}\right),$$

لازم به ذکر است که در این حالت $\delta(\omega)$ ثابت، تابع کواریانس و همچنین تابع چگالی طیفی تفکیک‌پذیر و $H(\mathbf{s},\omega)$ تابعی حقیقی مقدار است.

مثال ۲.۲. مدل تفکیک‌پذیر (کرسی و هوانگ ، ۱۹۹۹) با توابع کواریانس حاشیه‌ای مربع نمایی فضایی و نمایی زمانی

[11] Time Scale Parameter
[12] Space Scale Parameter

: فرض کنید $f(\omega) = ۲a\sigma^۲(a^۲+\omega^۲)^{-۱}$ تابع چگالی طیفی زمانی باشد که تابع کواریانس فضایی آن به صورت $\mathbb{C}(\mathbf{s}\delta(\omega)) = \exp\{-b \parallel \mathbf{s} \parallel^۲\}$، $\delta(\omega) = ۱$ است. همچنین بردار پارامترها به صورت $\theta = (a, b, \sigma^۲)'$ را در نظر بگیرید که در آن $a \geq ۰$ پارامتر مقیاس زمان، $b \geq ۰$ پارامتر مقیاس فضا و $\sigma^۲$ واریانس میدان تصادفی (ازاره) است. در اینصورت مدل نیمه‌طیفی متناظر با آن به فرم

$$H(\mathbf{s}, \omega) = f(\omega)\mathbb{C}(\mathbf{s}\delta(\omega)) = ۲a\sigma^۲(a^۲+\omega^۲)^{-۱}\exp\{-b \parallel \mathbf{s} \parallel^۲\}$$

خواهد بود و تابع کواریانس فضایی‌ـزمانی مربوطه به صورت

$$K(\mathbf{s}, t) = \sigma^۲\exp(-a|t| - b \parallel \mathbf{s} \parallel^۲)$$

است. با استفاده از رابطه (۳.۲) تابع چگالی طیفی فضایی‌ـزمانی به صورت زیر خواهد بود:

$$g(\boldsymbol{\lambda}, \omega) = \frac{۲a\sigma^۲\sqrt{\pi}}{b(a^۲+\omega^۲)}\exp\left(-\frac{\|\boldsymbol{\lambda}\|^۲}{۴b^۲}\right).$$

لازم به ذکر است که در این حالت $\delta(\omega)$ ثابت، تابع کواریانس و همچنین تابع چگالی طیفی تفکیک‌پذیر و $H(\mathbf{s}, \omega)$ تابعی حقیقی مقدار است.

مثال ۳.۲. مدل تفکیک‌ناپذیر کاملاً متقارن (گنتینگ ، ۲۰۰۲): فرض کنید

$$H(\mathbf{s}, \omega) = \frac{\sigma^۲}{(b\|\mathbf{s}\|)^\alpha + ۱} \times \frac{۲a\left((b\|\mathbf{s}\|)^\alpha + ۱\right)^{-\beta/۲}}{a^۲\left((b\|\mathbf{s}\|)^\alpha + ۱\right)^{-\beta} + \omega^۲}, \qquad (۵.۲)$$

که در آن بردار پارامترها $\theta = (a, b, \alpha, \beta, \sigma^۲)'$، $a \geq ۰$ پارامتر مقیاس زمان، $b \geq ۰$ پارامتر مقیاس فضا، $\sigma^۲$ واریانس میدان تصادفی (ازاره)، α پارامتر همواری Smoothness Parameter در بازه $(۰, ۲]$ و β پارامتر تعامل [۱۳] فضایی‌ـزمانی است، در اینصورت تابع کواریانس فضایی‌ـزمانی متناظر با آن به صورت زیر خواهد بود:

$$K(\mathbf{s}, \omega) = \frac{\sigma^۲}{(b\|\mathbf{s}\|)^\alpha + ۱} \times \exp\left(\frac{-a|t|}{\left((b\|\mathbf{s}\|)^\alpha + ۱\right)^{\beta/۲}}\right).$$

متاسفانه در این حالت نمایش بسته‌ای برای $g(\boldsymbol{\lambda}, \omega)$ وجود ندارد و همچنین نمی‌توان آن را به صورت مدل استاین یعنی $f(\omega)\mathbb{C}(\mathbf{s}\delta(\omega))$ نشان داد و لذا محاسبات به صورت عددی انجام می‌شود.

مثال ۴.۲. مدل تفکیک‌ناپذیر کاملاً متقارن (هورل و استاین ، ۲۰۱۷) : فرض کنید $h(\boldsymbol{\lambda}) = \phi(\alpha^۲+ \parallel \boldsymbol{\lambda} \parallel^۲)^{-(v+d/۲+۱/۲)}$ چگالی طیفی ماترن با پارامتر همواری $v + \frac{۱}{۲} > ۰$، پارامتر دامنه وارون $\alpha > ۰$ و پارامتر مقیاس $\phi > ۰$ باشد. برای یک f مفروض به منظور اطمینان از عدم نقض محدودیت ذکر شده، $\delta(\omega)$ به صورت زیر در نظر گرفته شده است:

$$\delta(\omega) = f(\omega)^{-\frac{۱}{۲}\frac{۱}{v+۱/۲}}, \qquad (۶.۲)$$

با جایگذاری رابطه (۶.۲) در (۳.۲) خواهیم داشت:

$$g(\boldsymbol{\lambda}, \omega) = \phi\left(\alpha^۲ f(\omega)^{-\frac{۱}{v+۱/۲}} + \parallel \boldsymbol{\lambda} \parallel^۲\right)^{-(v+d/۲+۱/۲)}. \qquad (۷.۲)$$

[13] Separability Parameter

اگر f به صورت $f \sim c_f \omega^{-k}$ در نظر گرفته شود برای بعضی از مقادیر $k > ۱$ و c_f که یک مقدار ثابت مثبت وابسته به f می‌باشد، آنگاه $g(\boldsymbol{\lambda}, \omega)$ تعریف شده به صورت فوق، رده‌ای از مدل‌ها را توصیف می‌کند که در شرط (۲.۲) صدق می‌کنند. نمایش نیمه‌طیفی رابطه (۷.۲) با جایگذاری تابع کواریانس ماترن در رابطه (۱.۲) به صورت

$$\begin{aligned} H(\mathbf{s}, \omega) &= f(\omega)\mathbb{C}(\mathbf{s}\delta(\omega)) \\ &= f(\omega)\frac{\phi\pi^{d/۲}}{۲^{v-۱/۲}\Gamma(v+(d+۱)/۲)\alpha^{۲v+۱}}\left(\alpha \parallel \mathbf{s} \parallel f(\omega)^{-۱/۲\frac{۱}{v+۱/۲}}\right)^{v+۱/۲} \\ &\times K_{v+۱/۲}\left(\alpha \parallel \mathbf{s} \parallel f(\omega)^{-۱/۲\frac{۱}{v+۱/۲}}\right), \end{aligned} \tag{۸.۲}$$

حاصل می‌شود که در آن $K_{v+۱/۲}$ تابع بسل مرتبه دوم است. اگر مقادیری که به ω بستگی ندارند در ϕ خلاصه کنیم، با فرض

$$\mathcal{M}_{V+۱/۲}(\mathbf{s}) = \parallel \mathbf{s} \parallel^{v+۱/۲} K_{v+۱/۲}(\parallel \mathbf{s} \parallel),$$

رابطه (۸.۲) را می‌توان به صورت زیر نوشت:

$$H(\mathbf{s}, \omega) = f(\omega)\mathbb{C}(\mathbf{s}\delta(\omega)) = \phi f(\omega)\mathcal{M}_{V+۱/۲}\left(\alpha \parallel \mathbf{s} \parallel f(\omega)^{-۱/۲\frac{۱}{v+۱/۲}}\right). \tag{۹.۲}$$

مثال ۵.۲. ماترن زمانی (هورل و استاین ، ۲۰۱۷) : فرض کنید $f(\omega) = (\beta^۲ + \omega^۲)^{k+۱/۲}$ چگالی طیفی ماترن یک بعدی باشد که برای سادگی، پارامتر مقیاس برابر یک فرض شده است. با جایگذاری آن در (۶.۲) و (۷.۲) نمایش طیفی و نیمه‌طیفی به صورت زیر خواهد بود:

$$g(\boldsymbol{\lambda}, \omega) = \phi\left(\alpha^۲(\beta^۲ + \omega^۲)^{(k+۱/۲)/(v+۱/۲)} + \parallel \boldsymbol{\lambda} \parallel^۲\right)^{-(v+d/۲+۱/۲)}.$$

$$f(\omega)\mathbb{C})(\mathbf{s}\delta(\omega)) = \phi\left((\beta^۲ + \omega^۲)^{-(k+۱/۲)}\right)\mathcal{M}_{v+۱/۲}\left(\alpha \parallel \mathbf{s} \parallel (\beta^۲ + \omega^۲)^{۱/۲\frac{k+۱/۲}{v+۱/۲}}\right). \tag{۱۰.۲}$$

لازم به ذکر است به منظور بهبود مدل‌های استفاده شده همانند گنتینگ (۲۰۰۲) و استاین (۲۰۰۵) به هر پنج مدل مورد مطالعه یک اثر قطعه‌ای [۱۴] اضافه شده است. بنابراین نمایش نیمه‌طیفی مدل‌های مورد بحث به صورت

$$f(\omega)\left[\mathbb{C}\left(\mathbf{s}\delta(\omega)\right) + \eta^۲\mathbb{I}_{\mathbf{s}=۰}\right],$$

خواهد بود که در آن $\mathbb{I}_{\mathbf{s}=۰}$ تابع نشانگر و $\eta^۲ \geq ۰$ اثر قطعه‌ای است.

۳ روش‌های برآورد پارامترها

در این بخش به معرفی چند روش برای برازش مدل‌های معرفی شده در بخش قبل پرداخته و سپس مدل‌های پیشنهادی در بخش قبل به داده‌های میانگین روزانه سرعت باد استان زنجان برازش شده است. چهار روش مختلف ماکسیمم درستنمایی (ML)، درستنمایی ترکیبی (CL)، حداقل مربعات وزنی (WLS) و درستنمایی وایتل [۱۵] (WL) برای مدل‌های معرفی شده بکار رفته است.

[14] Nugget Nugget
[15] Whittle Likelihood

برای برآورد پارامترها وقتی که توزیع مشاهدات معلوم است از روش درستنمایی ماکسیمم، استفاده می‌شود. به عنوان مثال اگر میدان تصادفی از توزیع گاوسی (ماردیا و مارشال ، ۱۹۸۴) پیروی کند می‌توان از روش درستنمایی ماکزیمم برای برآورد تغییرنگار و هم تغییر نگارها استفاده کرد.

فرض کنید $\mathbf{Z} = (Z(s_1,t_1).\cdots.Z(s_n,t_n))'$، n مشاهده از میدان تصادفی گاوسی با میانگین صفر و ماتریس کواریانس Σ باشد. در این صورت لگاریتم تابع درستنمایی به صورت

$$\ell(\boldsymbol{\theta}) \propto \log|\Sigma(\boldsymbol{\theta})| - \mathbf{Z}'\Sigma(\theta)^{-1}\mathbf{Z},$$

است که در آن $\Sigma(\theta) = Var(\mathbf{Z})$ به $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ وابسته است. برای یک میدان تصادفی گاوسی با یک تابع کواریانس پارامتری، برای محاسبه دقیق درستنمایی به محاسبه معکوس و دترمینان ماتریس کوواریانس نیاز است. هنگامی که تعداد مشاهدات زیاد است این ارزیابی مشکل خواهد بود. این واقعیت انگیزه‌ی اصلی تحقیق برای دست یافتن به تقریب درستنمایی است که نیاز به محاسبات کمتری داشته باشد.

در مواردی که توزیع مشاهدات معلوم نیست، برای برآورد پارامترهای مدل می‌توان از روش کم‌ترین توان‌های دوم وزنی استفاده کرد (کرسی ، ۱۹۸۵). در این روش فقط از عناصر روی قطر ماتریس کواریانس Σ استفاده می‌شود و بیشترین وزن به فاصله‌های نزدیک و وزن‌های کم به فاصله‌های دور اختصاص داده می‌شوند. در این روش $\boldsymbol{\theta}$ به نحوی انتخاب می‌شود که رابطه زیر مینیمم شود (کرسی ، ۱۹۹۳):

$$WLS(\boldsymbol{\theta}) = \sum_{l=1}^{L}\sum_{u=0}^{U} \mid N(h(l);u) \mid \left\{\frac{\hat{\gamma}(h(l);u)}{\hat{\gamma}(h(l);u \mid \boldsymbol{\theta})} - 1\right\}^2,$$

که در آن، $N(h(l);u) \equiv \{(i,j,t,t') : s_i - s_j \in Tol(h(l)); \mid t-t' \mid = u\}$ مجموعه زوج موقعیت‌هایی است که در فاصله (h,u) از یکدیگر قرار دارند. در اینجا $Tol(h(l))$ ناحیه تحمل [16] اطراف $h(l)$ است. روش درستنمایی ترکیبی به طور کلی یک روش شبه‌درستنمایی است که از توزیع حاشیه‌ای یا شرطی با بعد کمتری ساخته شده است. ایده اصلی درستنمایی ترکیبی در ابتدا توسط بسیج (۱۹۷۴) برای مدل‌سازی داده‌های فضایی ارائه گردید و بعداً توسط لیندسی (۱۹۸۸) به عنوان درستنمایی ترکیبی معرفی شد. در مواردی که برآورد و استنباط برای پارامترها و مدل‌های شبه‌پارامتری که ارزیابی توابع درستنمایی آنها با توجه به ساختار پیچیده داده‌ها دشوار است، محبوبیت این روش بیشتر است. دو روش برای ساخت تابع درستنمایی ترکیبی مطرح است: روش درستنمایی حاشیه‌ای ترکیبی و روش درستنمایی شرطی ترکیبی. طبق تعریف لیندسی تابع درستنمایی ترکیبی به صورت

$$CL(\theta,z) = \prod_{i=1}^{K} L_k(\theta,z),$$

است که در آن $k = 1,\cdots,K$، $L_k = (\theta,z)$ توابع درستنمایی حاشیه‌ای یا شرطی هستند.

درستنمایی وایتل (۱۹۵۴) یک تقریب برای درستنمایی سری‌های زمانی گاوسی مانا است. در یک مدل سری‌زمانی گاوسی مانا، تابع درستنمایی (بطور معمول گاوسی) توسط پارامترهای میانگین و کواریانس مربوط به مدل تعیین می‌شود و با وجود تعداد زیاد مشاهدات ممکن است ماتریس کواریانس بسیار بزرگ و محاسبات در عمل پرهزینه شود. بنابراین با توجه به مانایی، ماتریس کواریانس دارای یک ساختار نسبتاً ساده می‌شود که با استفاده از یک تقریب، محاسبات به طور قابل توجهی ساده خواهد شد. تابع درستنمایی وایتل چند متغیره به صورت

$$L(\theta) = \sum_{k=0}^{T-1}\left\{\log \mid F(\omega_k,\theta) \mid + tr\left(F(\omega_k,\theta)^{-1}I(\omega_k)\right)\right\} \qquad (۱.۳)$$

[16] Tolerance Region

است که در آن F و I به ترتیب ماتریس‌های طیفی و دوره نگار تجربی هستند.

۴ نتایج عددی

در این بخش یک مطالعه شبیه‌سازی با هدف ارزیابی عملکرد روش نیمه‌طیفی با توجه به مدل‌های کواریانس فضایی‌ـ زمانی تفکیک‌پذیر و تفکیک‌ناپذیر انجام شده است. در هرکدام از آنها روش پیشنهادی با چهار روش برآورد ML، CL، WLS و MWL و با تعداد مختلفی مشاهده مورد مقایسه قرار خواهد گرفت و سرانجام روش انتخاب شده برای داده‌های میانگین سرعت باد روزانه استان زنجان به کاررفته است. تعداد مشاهدات به گونه‌ای انتخاب شده که محاسبات مربوط به ماکسیمم درستنمایی به سادگی انجام شود. روش‌ها بر حسب میانگین مجذور خطاها (MSE) با ۱۰۰۰ تکرار از میدان تصادفی گاوسی $Z(\mathbf{s};t)$ با میانگین صفر و واریانس یک که در آن $t \in ۱,\cdots,T$ با $T = ۶۰,۹۰$ و $s_۱ \in [۰,۱۰]$ و $s_۲ \in [۰,۱۰]$ با $n = ۱۰,۲۰$ و همچنین سرعت انجام محاسبات در یک تکرار مقایسه شده‌اند.

۱.۴ شبیه‌سازی با مدل تفکیک‌پذیر

ابتدا مدل تفکیک‌پذیر با تابع کواریانس حاشیه‌ای نمایی مثال ۱.۲ در نظر گرفته شده و نتایج برازش بر اساس ۱۰۰۰ تکرار در جدول ۱ نشان داده شده است. همانطوری که از جدول ۱ ملاحظه می‌شود MSE روش ML طبق انتظار کمترین مقدار و روش برآورد MWL نسبت به CL و WLS از نظر MSE بهتر عمل می‌کند. لازم به ذکر است که با افزایش n و یا T، MSE برآوردها به کندی کاهش می‌یابد.

برای مدل تفکیک‌پذیر با تابع کواریانس حاشیه‌ای نمایی مثال ۱.۲ زمان لازم برای یک تکرار روش‌های CL، ML، WLS و MWL برحسب تعداد زمان و مکان متفاوت اندازه‌گیری و نتایج در جدول ۲ آمده است. همانطوری که از جدول ۲ ملاحظه می‌شود بجز در مورد روش WLS که به دلیل دقت کمتر آماری مورد نظر ما نیست، کارایی زمانی روش MWL به ویژه در ابعاد بالای تعداد مشاهدات نسبت به روش‌های CL و ML به مراتب بهتر است.

۲.۴ شبیه‌سازی با مدل تفکیک‌ناپذیر

در ادامه یک مدل کاملاً متقارن تفکیک‌ناپذیر استفاده شده است. کاملاً متقارن بودن معادل این شرط است که $H(\mathbf{s},\omega)$ حقیقی مقدار است. در این شبیه‌سازی از تابع کواریانس کاملاً متقارن تفکیک‌ناپذیر مثال ۳.۲ استفاده خواهد شد. در این شبیه‌سازی پارامترهای تعامل فضایی‌ـزمانی و همواری فضایی مقادیر ثابت $\beta = ۰/۵$ و $\alpha = ۱$ در نظر گرفته شده‌اند. نتایج برازش بر اساس ۱۰۰۰ تکرار در جدول ۳ آمده است. مانند مثال قبل با افزایش n و یا T، MSE برآوردها به کندی کاهش می‌یابد. با توجه به مقادیر MSE در جدول‌های ۱ و ۳ بهترین روش برای برآورد پارامترهای مدل، روش ML و پس از آن روش MWL مناسب است. همانطوری که قبلاً ذکر شد در حالتی که تعداد مشاهدات بزرگ است، روش ML دشوار و به صرفه نیست، بنابراین در این مقاله برای برآورد پارامترهای مدل برازش شده به داده‌های واقعی از روش MWL استفاده خواهد شد.

برای مدل تفکیک‌ناپذیر مثال ۳.۲ با $\beta = ۰/۵$ و $\alpha = ۱$ زمان لازم برای یک تکرار روش‌های CL، ML، WLS و MWL برحسب تعداد زمان و مکان متفاوت اندازه‌گیری و نتایج در جدول ۴ آمده است. همانطوری که از جدول ۴ ملاحظه می‌شود در این حالت نیز بجز روش WLS که به دلیل دقت کمتر آماری مورد نظر ما نیست، کارایی زمانی روش MWL به ویژه در ابعاد بالای تعداد مشاهدات نسبت به روش‌های CL و ML بهتر است.

جدول ۱: مقایسه MSE در داده‌های شبیه‌سازی شده برای مدل تفکیک‌پذیر مثال ۱.۲

پارامترها	روش برآورد	$n=۱۰$		$n=۲۰$	
		$T=۶۰$	$T=۹۰$	$T=۶۰$	$T=۹۰$
		MSE	MSE	MSE	MSE
اثر قطعه‌ای	CL	۰/۰۴۵	۰/۰۴۰	۰/۰۱۰	۰/۰۰۸
	ML	۰/۰۱۸	۰/۰۱۱	۰/۰۰۶	$۰۴e-۳$
	WLS	۰/۱۴۰	۰/۱۲۴	۰/۱۰۰	۰/۰۵۰
	MWL	۰/۰۳۲	۰/۰۲۵	۰/۰۱۰	۰/۰۰۱
مقیاس فضایی	CL	۰/۰۶۴	۰/۰۶۱	۰/۰۱۰	۰/۰۰۸
	ML	۰/۰۶۰	۰/۰۱۲	۰/۰۰۳	$۰۴e-۱$
	WLS	۰/۴۱۰	۰/۱۳۶	۰/۱۱۸	۰/۱۰۴
	MWL	۰/۰۶۴	۰/۰۵۳	۰/۰۵۷	۰/۰۰۷
مقیاس زمانی	CL	۰/۰۴۶	۰/۰۴۱	۰/۰۳۲	۰/۰۲۰
	ML	۰/۰۳۶	۰/۰۲۵	۰/۰۱۶	۰/۰۷
	WLS	۰/۱۸۵	۰/۱۴۳	۰/۱۹۷	۰/۰۷۴
	MWL	۰/۰۴۶	۰/۰۳۷	۰/۰۲۹	۰/۰۰۵
ازاره	CL	۰/۰۲۴	۰/۰۲۸	۰/۰۱۸	۰/۰۰۴
	ML	۰/۰۱۵	۰/۰۱۹	۰/۰۰۹	$۴e-۶$
	WLS	۰/۱۳۸	۰/۱۴۰	۰/۱۰۰	۰/۰۸۴
	MWL	۰/۰۲۴	۰/۰۲۱	۰/۰۱۰	۰/۰۲۰

جدول ۲: زمان لازم (برحسب ثانیه) برای تکرار روش‌های CL، ML، WLS و MWL در مدل تفکیک‌پذیر برحسب تعداد زمان و مکان متفاوت $n=۱۰,۲۰$ و $T=۶۰,۹۰$

T	n	CL	ML	WLS	MWL
۶۰	۱۰	۳۴/۹۱	۲۶۳/۸۵	۰/۳۱	۳/۳۷
۹۰	۱۰	۷۲/۰۸	۳۰۴/۸۰	۰/۶۹	۳/۹۷
۶۰	۲۰	۱۰۱/۲۸	۹۸۳/۵۲	۱/۰۸	۱۴/۸۶
۹۰	۲۰	۲۳۶/۶۸	۳۱۰۳/۹۲	۲/۶۴	۱۵/۵۳

۳.۴ تحلیل داده‌های واقعی

داده‌های مورد بررسی، مقادیر میانگین سرعت باد روزانه بر حسب متر بر ثانیه در ۸ ایستگاه هواشناسی استان زنجان مربوط به سال‌های ۲۰۰۳ تا ۲۰۱۷ است که توسط اداره هواشناسی استان زنجان اندازه‌گیری و ثبت شده‌اند. به دلیل وفور داده‌های گمشده در ایستگاه‌های مختلف، تنها از داده‌های ۵ ایستگاه که به طور کامل اطلاعات مربوط به سرعت باد طی سال‌های یاد شده را دارند، استفاده شده است. جدول ۵ طول و عرض جغرافیایی ایستگاه‌های مورد نظر را نشان می‌دهد که در آن طول و عرض جغرافیایی هر ایستگاه برحسب درجه، دقیقه و ثانیه، ارتفاع از سطح دریا برحسب متر، میانگین و انحراف معیار سرعت باد در هر ایستگاه بر حسب متر بر ثانیه ثبت شده است.

برای تحلیل فضایی‌ـ زمانی داده‌ها نخست تحلیل کاوشگرانه مشاهدات از نظر نرمال بودن، مانایی و همگنی در واریانس انجام می‌شود. نمودار بافت نگار داده‌ها در شکل ۱ سمت چپ، نشان‌دهنده عدم تقارن و انحراف از نرمال بودن توزیع

جدول ۳: مقایسه MSE در داده‌های شبیه‌سازی شده برای مدل تفکیک‌ناپذیر مثال ۳.۲ یا $\alpha = ۱$، $\beta = ۰/۵$

پارامترها	روش برآورد	$n = ۱۰$		$n = ۲۰$	
		$T = ۶۰$	$T = ۹۰$	$T = ۶۰$	$T = ۹۰$
		MSE	MSE	MSE	MSE
اثر قطعه‌ای	CL	۰/۵۹۱	۰/۴۱۴	۰/۲۲۹	۰/۰۴۱
	ML	۰/۴۶۸	۰/۲۵۳	۰/۱۰۷	۰/۰۰۵
	WLS	۰/۷۴۱	۰/۶۳۸	۰/۳۶۲	۰/۱۳۱
	MWL	۰/۵۷۳	۰/۳۱۲	۰/۱۵۶	۰/۰۱۲
مقیاس فضایی	CL	۰/۲۸۹	۰/۲۸۹	۰/۱۶۷	۰/۰۴۳
	ML	۰/۲۳۸	۰/۱۲۰	۰/۰۵۶	۰/۰۰۳
	WLS	۰/۴۲۶	۰/۴۱۸	۰/۲۰۱	۰/۱۱۴
	MWL	۰/۲۵۴	۰/۲۱۷	۰/۱۴۳	۰/۰۱۲
مقیاس زمانی	CL	۰/۳۸۴	۰/۲۶۵	۰/۱۳۶	۰/۰۱۵
	ML	۰/۳۲۸	۰/۲۵۱	۰/۰۸۷	۰/۰۰۹
	WLS	۰/۵۷۸	۰/۵۹۳	۰/۳۲۵	۰/۲۶۱
	MWL	۰/۳۳۲	۰/۲۸۵	۰/۱۰۴	۰/۰۶۳
ازاره	CL	۰/۱۹۳	۰/۱۲۷	۰/۰۷۴	۰/۰۰۸
	ML	۰/۱۵۲	۰/۰۶۸	۰/۰۰۸	۰/۰۰۱
	WLS	۰/۳۶۱	۰/۲۴۲	۰/۱۸۱	۰/۱۱۶
	MWL	۰/۱۶۱	۰/۱۰۲	۰/۰۵۴	۰/۰۱۲

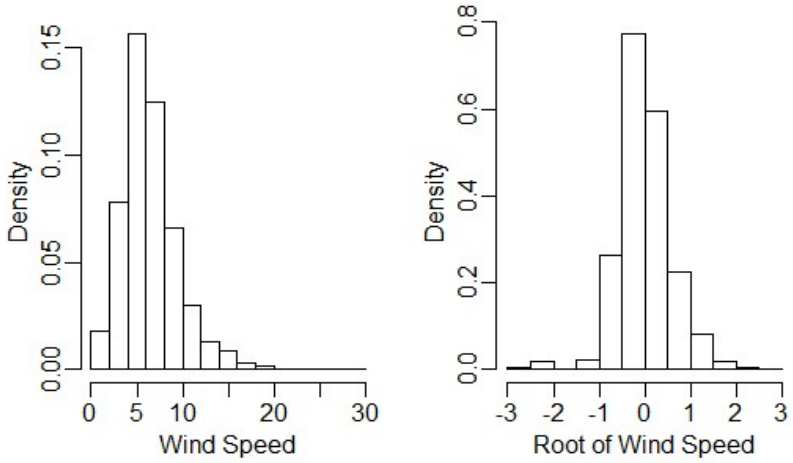
جدول ۴: زمان لازم (برحسب ثانیه) برای تکرار روش‌های CL، ML، WLS و MWL در مدل تفکیک‌ناپذیر برحسب تعداد زمان و مکان متفاوت $n = ۱۰, ۲۰$ و $T = ۶۰, ۹۰$

T	n	CL	ML	WLS	MWL
۶۰	۱۰	۱۶/۷۹	۵۷/۶۴	۰/۳۶	۱۰/۸۸
۹۰	۱۰	۵۴/۵۰	۸۷/۷۰	۰/۸۰	۳۲/۹۸
۶۰	۲۰	۹۳/۴۴	۱۶۶/۵۶	۱/۲۲	۳۸/۸۸
۹۰	۲۰	۳۷۲/۱۱	۶۶۳/۵۸	۲/۹۴	۱۰۴/۹۲

جدول ۵: شماره، نام و موقعیت جغرافیایی ایستگاه‌ها

شماره	نام ایستگاه	طول و عرض جغرافیایی		ارتفاع از سطح دریا	میانگین سرعت باد	انحراف معیار سرعت باد
		طول	عرض			
۱	ماهنشان	$۴۷d۴۱'E$	$۳۶d۴۴'N$	۱۳۵۰	۶/۳۲۱	۲/۳۶۸
۲	زنجان	$۳۶d۳۹'N$	$۴۸d۳۱'E$	۱۶۶۳	۶/۷۴۸	۲/۸۱۳
۳	خرم‌دره	$۳۶d۱۱'N$	$۴۹d۱۲'E$	۱۵۷۵	۷/۳۳	۲/۸۱۸
۴	خدابنده	$۳۶d۰۸'N$	$۴۸d۳۵'E$	۱۹۹۷	۸/۷۰۲	۳/۷۳۰
۵	ابهر	$۳۶d۵۶'N$	$۴۸d۵۶'E$	۱۵۴۰	۵/۶۸۱	۳/۰۱۹

داده‌ها است. به منظور نرمال‌سازی داده‌ها از دو تبدیل ریشه دوم و لگاریتم استفاده شده است. اما تبدیل ریشه دوم داده‌ها که نمودار آن در شکل ۱ سمت راست رسم شده، موجب تقارن و به طور تقریبی نرمال شدن توزیع داده‌ها شده است. همچنین باید فرض همگنی واریانس روی فضا و زمان برای داده‌های اصلی بررسی شود. ساهو و همکاران (۲۰۰۶) و هوانگ و همکاران (۲۰۰۷) برای بررسی فرض همگنی واریانس در داده‌های فضایی_زمانی از نمودار انحراف استاندارد نمونه در مقابل میانگین نمونه استفاده کردند. با رسم این نمودار برای متوسط‌های زمانی در شکل ۲ سمت چپ، وجود روند و انحراف از فرض همگونی واریانس مشاهده می‌شود که خوشبختانه این روند و ناهمگونی در واریانس با بکار بردن تبدیل ریشه دوم داده‌های اصلی، همانطوری که در شکل ۲ سمت راست، ملاحظه می‌شود حذف شده است. همچنین همانطوری که در شکل ۳ سمت راست، ملاحظه می‌شود در داده‌های اصلی روند فصلی وجود دارد که این روند فصلی به کمک روش برازش رگرسیون ناپارامتری با استفاده از تابع $Lowess$ در R حذف شده است. این تابع محاسبات هموارسازی $Lowess$ را با استفاده از رگرسیون چندجمله‌ای وزنی موضعی انجام می‌دهد (کلولند ، ۱۹۸۱). شکل ۳ سمت چپ، به وضوح حذف این اثر فصلی را نشان می‌دهد. نهایتاً مدل فضایی_زمانی ارائه شده روی باقیمانده‌های حاصل از حذف روند انجام گرفته است. برازش مدل‌های معرفی شده در بخش قبل به داده‌ها می‌تواند با محاسبه تابع کواریانس از طریق حل عددی



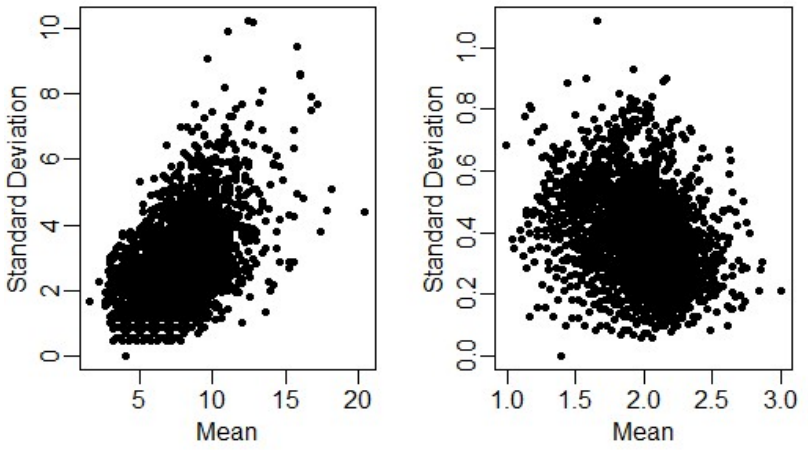
شکل ۱ : هیستوگرام داده‌های اصلی و ریشه دوم داده‌ها

انتگرال رابطه (۱.۲) انجام شود و یا اینکه به داده‌ها به عنوان سری‌زمانی چند متغیره نگاه کرد و روش‌های طیفی را بکار برد. در این مقاله از برازش‌های طیفی استفاده و در روش طیفی برای برآورد پارامترها از درستنمایی چند متغیره وایتل استفاده شده است. مدل‌های معرفی شده در بخش قبل در زمان و فضا پیوسته هستند، در حالی‌که روش درستنمایی وایتل برای سری‌های زمانی گسسته بکار برده می‌شود. به منظور رفع این مشکل تصحیح هم‌اثرسازی [1] زیر بکار رفته و برای مدل پیوسته $f(\omega)\mathbb{C}(\mathbf{s}\delta(\omega))$ از تقریب

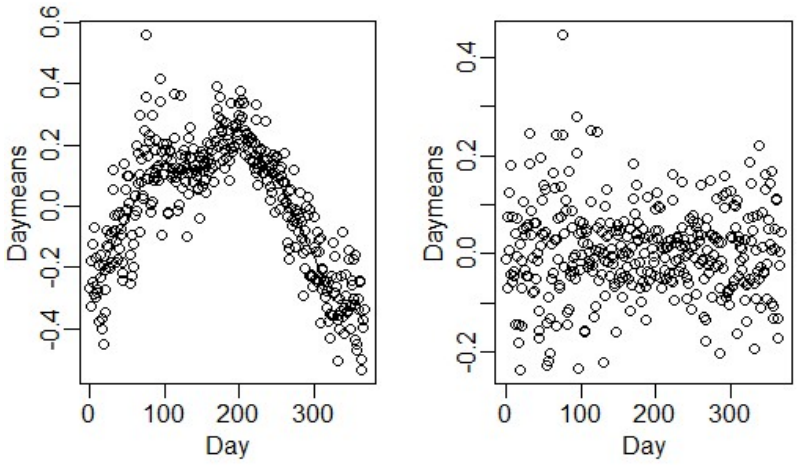
$$\sum_{j=-m}^{m} f(\omega + ۲j\pi)\mathbb{C}(\mathbf{s}\delta(\omega + ۲j\pi)), \quad -\pi < \omega < \pi, \quad m = ۵۰$$

استفاده شده است. نتایج برازش مدل‌های ذکر شده به داده‌ها و مقادیر لگاریتم درستنمایی وایتل و همچنین تقریبی از معیار اطلاع آکائیک (AIC) برای هریک از این مدل‌ها در جدول ۶ ثبت شده است. هرچند که معیار اطلاع آکائیک یکی از کاربردی‌ترین ابزارهای انتخاب مدل می‌باشد، ولی تحقیقات کمتری درمورد کارایی این ابزار در مدل‌سازی فضایی و

[1] Aliasing Correction



شکل ۲: نمودار انحراف معیار نمونه در مقابل میانگین نمونه قبل (چپ) و بعد از تبدیل ریشه دوم (راست)



شکل ۳: حذف روند فصلی در داده‌ها

فضایی—زمانی انجام شده است. هوتینگ و همکاران (۲۰۰۶) و همچنین لی و گاش (۲۰۰۹) تأثیر همبستگی فضایی روی انتخاب مدل توسط معیار AIC را در آمار فضایی مورد بررسی قرار داده‌اند و نتایج شبیه‌سازی آنها نشان داد که معیار AIC با در نظر گرفتن همبستگی فضایی بهتر از حالتی است که ساختار فضایی در نظر گرفته نشده است. از طرفی آنها نشان داده‌اند که اگر اندازه نمونه بزرگ باشد، معیار AIC پیشنهادی آنها تقریباً برابر معیار AIC در حالتی که داده‌ها مستقل هستند می‌باشد. علاوه بر این هوانگ و همکاران (۲۰۰۷) نشان دادند که AIC هنوز هم می‌تواند معیار مناسبی برای انتخاب مدل بین مدل‌های فضایی_زمانی باشد. در این مقاله نیز به پیروی از مقالات (هوتینگ و همکاران ، ۲۰۰۶) و (هوانگ و همکاران ، ۲۰۰۷) به دلیل اینکه مجموع نمونه‌های فضایی و زمانی بسیار بزرگ است، برای سادگی از AIC بدون در نظر گرفتن همبستگی داده‌ها استفاده شده است و تعداد پارامترهای موثر تقریباً برابر تعداد پارامترهای مدل در نظر گرفته شده‌اند. با مقایسه مقادیر لگاریتم درستنمایی وایتل چند متغیره و معیار اطلاع آکائیک (AIC) در جدول ۶ نتیجه می‌شود که در بین پنج مدل استفاده شده، مدل مربوط به مثال ۳.۲، برازنده‌ترین مدل برای داده‌های باد استان زنجان بوده است.

جدول ۶: مقایسه درستنمایی وایتل در مدل‌های برازش شده به داده‌های باد استان زنجان

مدل	مقدار ماکسیمم	اختلاف با مدل مثال ۳.۲	تعداد پارامترها	AIC
مثال ۱.۲	۲۵۵۲۵/۲	۷۰۰۴۵/۷	۴	−۵۱۰۴۲/۴
مثال ۲.۲	۲۵۵۴۹/۸	۷۰۰۲۱/۱	۴	−۵۱۰۹۱/۶
مثال ۳.۲	۹۵۵۷۰/۹	۰	۶	−۱۹۱۱۲۹/۸
مثال ۴.۲	۶۰۱۴۴/۲۵	۳۵۴۲۶/۷	۵	−۱۲۰۲۷۸/۵
مثال ۵.۲	۳۳۰۴۹	۶۲۵۲۱/۹	۴	−۶۶۰۹۰

با در نظر گرفتن اثر قطعه‌ای، مدل نیمه‌طیفی مثال ۳.۲ به صورت زیر در نظر گرفته شده است:

$$H(\mathbf{s},\omega)=\begin{cases}\eta^2\sigma^2\times\dfrac{2a}{a^2+\omega^2} & \|\mathbf{s}\|=0\\ \dfrac{\sigma^2}{(b\|\mathbf{s}\|)^\alpha+1}\times\dfrac{2a\left((b\|\mathbf{s}\|)^\alpha+1\right)^{\frac{-\beta}{2}}}{a^2\left((b\|\mathbf{s}\|)^\alpha+1\right)^{-\beta}+\omega^2} & \|\mathbf{s}\|\neq 0\end{cases}$$

با استفاده از تابع درستنمایی وایتل چند متغیره در رابطه (۱.۳) برای برآورد پارامترهای $H(\mathbf{s},\omega)$ یعنی $\theta=(\eta^2,\sigma^2,a,b,\alpha,\beta)'$ مقادیر

$$\hat{\eta}=0/029,\ \sigma^2=0/267,\ \hat{a}=0/0012,\ \hat{b}=1/121,\ \hat{\alpha}=1/13,\ \hat{\beta}=1/57$$

و مقدار لگاریتم درستنمایی وایتل مربوطه $MWL(\theta)=95570/9$ بدست آمده است.

۵ بحث و نتیجه گیری

در این مقاله تحلیل نیمه‌طیفی داده‌های فضایی ـ زمانی توسط مدل‌های نیمه‌طیفی مورد بررسی و مطالعه قرار گرفت. یک مطالعه شبیه‌سازی به منظور مقایسه عملکرد روش نیمه‌طیفی با روش‌های کلاسیک با توجه به مدل‌های کوواریانس فضایی — زمانی تفکیک‌پذیر و تفکیک‌ناپذیر انجام شد. برای این منظور از چهار روش برآوردیابی ماکسیمم درستنمایی، درستنمایی ترکیبی، حداقل مربعات وزنی و روش نیمه‌طیفی استفاده و روش‌ها بر اساس میانگین مجذور خطاها با ۱۰۰۰ تکرار از میدان تصادفی گاوسی با میانگین صفر و واریانس یک مقایسه شده‌اند. نتایج شبیه‌سازی گویای عملکرد بهتر روش نیمه‌طیفی نسبت به دیگر روش‌ها هم از نظر محاسباتی و هم از نظر آماری است. در پایان نیز تحلیل نیمه‌طیفی داده‌های میانگین سرعت باد روزانه استان زنجان به عنوان کاربردی از مدل پیشنهادی ارائه شد.

مراجع

امیدی، م. و محمدزاده، م. (۱۳۹۲). *تعیین ساختار همبستگی داده‌های فضایی با توابع مفصل*، نشریه علوم دانشگاه خوارزمی، ۳، شماره ۳، ۷۹۷ـ۸۰۸.

Besag, J. (1974), *Spatial Interaction and the Statistical Analysis of Lattice Systems with Discussion*, Journal of the Royal Statistical Society, Series B, 192-236.

Bochner. S. (1955), *Harmonic Analysis and The Theory of Probability*. University of California Press, Berkely and Los Angeles.

Chiles, J. P. and Delfner, P. (2012), *Geostatistics: Modeling Spatial Uncertainty*. 2nd ed., Wiley, New York.

Cleveland, W. S. (1981), LOWESS: *A program for smoothing scatterplots by robust locally weighted regression*. The American Statistician 35, 54.

Cressie, N. (1985), *Fitting variogram models by weighted least squares*. Journal of the International Association for Mathematical Geology, 17(5), 563586.

Cressie, N. (1993), *Statistics for Spatial Data*. Revised edn, New York: Wi ley.

Cressie, N. and Huang, H. C. (1999), Classes of nonseparable, spatio-temporal stationary covariance functions. Journal of the American Statistical Association, 94, 1330-1340.

Cressie, N. and Wikle, C. K. (2013), *Statistics for spatio-temporal data*, John Wiley and Sons.

Gneiting, T. (2002), *Nonseparable stationary covariance functions for space-time data*. Journal of the American Statistical Association, 97, 590-600.

Hoeting, J. A., Davis, R. A., Merton, A. A., & Thompson, S. E. (2006), *Model selection for geostatistical models. Ecological Applications*, 16, 87-98.

Horrell, M. T. and Stein, M. L. (2017), *Half-spectral space-time covariance models*. Spat. Stat., 19, 90-100.

Huang, H. C., Martinez, F., Mateu, J and Montes, F. (2007), *Model comparison and selection for stationary space-time models*. Computational statistics and data analysis, 51, 4577-4596.

Journel, A. G. and Huijbregts, C. J. (1978), *Mining geostatistics*. Academic press.

Kammler, D. W. (2007), A *first course in Fourier analysis. 2nd ed.*, Cambridge University Press,Cambridge.

Kent, J. T., Mohammadzadeh, M. and Mosammam, A. M. (2011), *The dimple in Gneiting's spatial-temporal covariance model*. Biometrika, 98, 489-494.

Lee, H., and Ghosh, S. K. (2009), *Performance of information criteria for spatial models*. Journal of statistical computation and simulation, 79(1), 93-106.

Lindsay, B. G. (1988), *Composite likelihood methods*. Contemporary Mathematics, 80, 221-39.

Mardia, K. V. and Marshall, R. J. (1984), *Maximum likelihood estimation of models for residual covariance in spatial regression*. Biometrika, 71, 135-146.

Mosammam, A. M. and Kent, J. T. (2016), *Estimation and testing for covariance-spectral spatial-*

temporal models. Environ. Ecol. Stat., 23, 43-64.

Omidi, M. and Mohammadzadeh, M., (2015), *A New Method to Build Spatio-Temporal Covariance Functions: Analysis of Ozone Data*. Statistical Papers, 57, 689–703.

Rodriguez-Iturbe, I. and Meji'a, J. M. (1974), *The design of rainfall networks in time and space*. WaterResources Research, 10, 713-728.

Sahu, S. K. G., Gelfand, A. E., Holland, D. M. and Mardia, K. (2006), *Spatio-Temporal modeling of fine particulate matter*. Journal of Agricultural, Biological and Environmental Statistics, 11, 61-86.

Shumway, R. H. and Stoffer, D. S. (2017), *Time Series Analysis and its Applications with R Examples. 4th ed.*, Springer Texts in Statistics, Springer, Cham.

Stein, M. L. (2005), *Statistical methods for regular monitoring data*. J. R. Stat. Soc. Ser. B Stat.Methodol., 67, 667-687.

Stein, M. L. (2011), *When does the screening effect not hold?*. Ann. Statist., 39, 2795-2819.

Whittle, P. (1954), *On stationary processes in the plane*. Biometrika, 41, 434-449.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

مقایسه دو تابع درستنمایی مرکب در برآورد مدل های آمیخته خطی تعمیم یافته فضایی

لیلا صالحی[1]، محسن محمدزاده
گروه آمار، دانشگاه تربیت مدرس

چکیده: هنگامی که با متغیرهای همبسته فضایی ناگاوسی سر و کار داریم، مدل‌های آمیخته خطی تعمیم‌یافته فضایی از انعطاف لازم برای مدل‌بندی داده‌های گوناگون برخوردارند. اگر حجم داده‌ها بزرگ باشد، معمولا روش‌های ماکسیمم درستنمایی در برآورد پارامترهای مدل با محاسبات سنگین و سرعت پایین همراه هستند. برای رفع این مشکل، توابع درستنمایی مرکب که از حاصل ضرب درستنمایی زیر مجموعه‌های مشاهدات به دست می‌آیند، به کار گرفته می‌شوند. در این مقاله برآورد پارامترهای مدل‌های آمیخته خطی تعمیم‌یافته فضایی با استفاده از تابع درستنمایی زوجی بررسی می‌شود. سپس تابع درستنمایی زوجی موزون و تابع درستنمایی زوجی تاوانیده برای برآورد پارامترهای این مدل‌ها را مورد استفاده قرار داده و در مطالعه‌ای شبیه‌سازی دقت برآوردها بر اساس این توابع درستنمایی مورد ارزیابی و مقایسه قرار می‌گیرد.

واژه‌های کلیدی: مدل‌های آمیخته خطی تعمیم‌یافته فضایی، تابع درستنمایی مرکب، درستنمایی زوجی، درستنمایی زوجی تاوانیده، درستنمایی زوجی موزون.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62M30، 91B72.

۱ مقدمه

مدل‌های خطی تعمیم‌یافته برای اولین بار توسط نلدر و وددربن (۱۹۷۲) معرفی شدند و مک‌ـ کالا (۱۹۸۹) برای مدل‌بندی متغیرهای پاسخ گسسته از این مدل‌ها استفاده کردند. در این مدل‌ها با فرض استقلال مشاهدات، با استفاده از یک تابع پیوند بین میانگین مشاهدات و متغیرهای کمکی ارتباط خطی برقرار می‌شود. در حالتی که بین مشاهدات همبستگی وجود دارد، تعمیمی از مدل‌های مذکور به نام مدل‌های آمیخته خطی تعمیم‌یافته[1] (GLM) استفاده می‌شود. در این مدل‌ها فرض استقلال مشاهدات به استقلال شرطی تعدیل و همبستگی بین آنها با اضافه کردن اثرات تصادفی از طریق متغیرهای پنهان[2]

[1] Generalized Linear Mixed
[2] Latent Variables

[1] نام و ایمیل ارائه دهنده مقاله: لیلا صالحی، l.salehi@modares.ac.ir

به مدل در نظر گرفته می‌شود. در حالتی که ساختار همبستگی پاسخ‌های فضایی در قالب یک مدل مشخص باشد، با در نظر گرفتن یک میدان تصادفی تحلیل فضایی و پیشگویی مقادیر نامعلوم در موقعیت‌های مشخص با روش کریگیدن[۳] امکان‌پذیر است. اما در مواردی که ساختار همبستگی پاسخ‌های فضایی نامشخص باشد، می‌توان از مدل‌های آمیخته خطی تعمیم‌یافته فضایی[۴] (SGLM) استفاده نمود. چون در مدل‌های SGLM تابع درستنمایی برخلاف مدل‌های خطی به دلیل ناگاوسی بودن متغیر پاسخ فرم بسته‌ای ندارد، برآورد پارامترها به روش ماکسیمم درستنمایی میسر نیست. لذا در اکثر مقالات با پذیرش فرض نرمال بودن متغیرهای پنهان به ارائه راه حلی برای برآورد پارامترهای مدل و متغیرهای پنهان با ماکسیمم کردن توابع درستنمایی، شبه‌درستنمایی تاوانیده[۵] یا درستنمایی سلسله مراتبی به روش‌های عددی پرداخته شده است. از جمله مک‌ـ کولاچ (۱۹۹۷) با به کار بردن روش‌های عددی مثل ماکسیمم سازی امید ریاضی مونت کارلویی[۶] (MCEM)، برآورد پارامترهای مدل‌های GLM با اثرات تصادفی غیرفضایی را به دست آورد. محمدزاده و حسینی (۲۰۱۱) با فرض چوله بسته بودن متغیرهای پنهان الگوریتم گرادیانت EM مونت کارلو را برای برآورد ماکسیمم درستنمایی پارامترهای مدل پیشنهاد کردند. باغیشنی و محمدزاده (۲۰۱۱) و ترابی (۲۰۱۵) روش دیگری بر اساس الگوریتم همسان‌سازی داده‌ها ارائه کردند. اخیرا محققان روش‌های تقریبی در قالب رهیافت بیزی و بسامدی را مورد توجه قرار داده‌اند که نسبت به روش‌های مبتنی بر الگوریتم‌های شبیه‌سازی، بار محاسباتی کمتر و سرعت بیشتری دارند. رائو و مارتینو (۲۰۰۷) استنباط بیزی تقریبی برای فرایندهای تصادفی مارکوفی گاوسی معرفی کردند. ایدسویک و همکاران (۲۰۱۲) روش مشابه را برای مدل‌های SGLM به کار بردند.

در مطالعات اخیر روش‌های تقریبی دیگری مورد توجه قرار گرفته‌اند که مبتنی بر درستنمایی کامل مشاهدات نیستند. مزیت این روش‌ها در مقایسه با روش‌های درستنمایی تقریبی، عدم نیاز به مدل‌بندی توام تمام مشاهدات است. از جمله روش‌های شبه‌درستنمایی وددربن (۱۹۷۴) و مک‌ـ کالا (۱۹۸۳) و روش‌های مبتنی بر مشخصات جزیی درستنمایی کامل که درستنمایی مرکب[۷] نامیده می‌شود. برای اولین بار هگرتی و له له (۱۹۹۸) تابع درستنمایی مرکب زوجی[۸] را برای مشاهدات دودویی با تابع پیوند پروبیت به کار بردند. وارین و همکاران (۲۰۰۵) روش درستنمایی مرکب زوجی را برای مدل‌های SGLM استفاده کردند. آن‌ها برای ماکسیمم کردن تابع درستنمایی زوجی، یک الگوریتم جدید از نوع EM را که از مربع‌سازی عددی استفاده می‌کند، معرفی کردند. بویلاکوا و همکاران (۲۰۱۰) برای داده‌های فضایی‌ـ زمانی از تابع درستنمایی موزون استفاده کردند و نشان دادند که برآورد بر اساس این تابع، تقریب مناسبی از برآورد ماکسیمم درستنمایی است. جو و لی (۲۰۰۹) نشان دادند که موزون کردن تابع درستنمایی موجب افزایش کارایی نسبی مجانبی در برآورد در داده‌های دسته‌بندی شده می‌شود.

در این مقاله تابع درستنمایی زوجی برای مدل‌های SGLM به کار گرفته می‌شود. سپس با استفاده از یک تابع وزن، تابع درستنمایی زوجی موزون را تشکیل داده و برای برآورد پارامترهای این مدل‌ها مورد استفاده قرار می‌گیرد. همچنین برای افزایش دقت در برآورد پارامترهای مدل، تابع درستنمایی زوجی تاوانیده به کار گرفته می شود و در مطالعه‌ای شبیه‌سازی، دقت برآورد پارامترهای مدل بر اساس توابع درستنمایی زوجی، درستنمایی زوجی موزون و درستنمایی زوجی تاوانیده بر اساس ملاک میانگین توان دوم خطا(MSE) مورد ارزیابی و مقایسه قرار می‌گیرد.

---

[3] Kriging
[4] Spatial Generalized Linear Mixed
[5] Penalized Quasi Likelihood
[6] Monte Carlo Expectation Maximization
[7] Composite Likelihood
[8] Pairwise Composite Likelihood

۲ مدل‌های آمیخته خطی تعمیم‌یافته فضایی

فرض کنید $Y(s)$ متغیر پاسخ فضایی گسسته و $Z_1(s), \ldots, Z_p(s)$ متغیرهای تبیینی قابل مشاهده و $\{X(s), s \in \mathbb{R}^2\}$ یک میدان تصادفی فضایی پنهان باشد، طوری که $X(s)$ نشان‌دهنده اثر تصادفی در موقعیت s است. دیگل و همکاران (۱۹۹۸) مدل‌های SGLM را به صورت زیر تعریف کردند.

الف- $\{X(s), s \in \mathbb{R}^2\}$ یک میدان تصادفی مانای گاوسی است، به طوری که برای هر s، $E[X(s)] = 0$ و برای هر $h \in \mathbb{R}$، $Cov(X(s+h), X(s)) = C(h;\theta)$، که در آن $\theta \in \mathbb{R}^k$ بردار پارامتر همبستگی است.

ب- $\{Y(s), s \in \mathbb{R}^2\}$ به شرط $\{X(s), s \in \mathbb{R}^2\}$، مجموعه‌ای از متغیرهای تصادفی مستقل است و توزیع هر $Y(s)$ با میانگین شرطی $E[Y(s)|X(s)]$ مشخص می‌شود.

ج- برای هر تابع پیوند g و پارامترهای رگرسیونی $\beta_1, \ldots, \beta_p$ داریم $g\{E[Y(s)|X(s)]\} = \sum_{j=1}^{p} Z_j(s)\beta_j + X(s)$.

د- توزیع شرطی $[Y(s)|X(s)]$ عضو خانواده نمایی است.

۱.۲ تابع درستنمایی مرکب زوجی موزون

تابع درستنمایی مرکب از ضرب مجموعه‌ای از مولفه‌های درستنمایی به دست می‌آید، که هر مولفه درستنمایی زیرمجموعه‌ای از مشاهدات را بیان می‌کند. دو انگیزه اصلی برای استفاده از درستنمایی مرکب وجود دارد: انگیزه اول کاهش محاسبات در مدل‌بندی توزیع توام یک بردار تصادفی با بعد بالا است. انگیزه دوم استواری تحت نامعین بودن توزیع‌های با بعد بالاست. چون درستنمایی مرکب تنها به فرض‌های مربوط به چگالی‌های حاشیه‌ای و شرطی با بعد پایین نیاز دارد و آگاهی از جزییات توزیع توام ضروری نخواهد بود. استنباط بر اساس تابع درستنمایی برای داده‌های حجیم با انتگرال‌گیری و معکوس ماتریس‌های بعد بالا همراه است که ممکن است حل آن‌ها برای رایانه‌های قدرتمند نیز مشکل باشد. با استفاده از تابع درستنمایی مرکب موزون می‌توان این انتگرال‌های چندگانه را به جمع انتگرال‌های با بعد پایین تبدیل نمود. بردار تصادفی $Y = (y_1, \ldots, y_m)$ بعدی m را با تابع چگالی احتمال $f(y;\theta)$ برای پارامتر p بعدی $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$ در نظر بگیرید. فرض کنید $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ مجموعه ای از پیشامدهای شرطی یا حاشیه‌ای با تابع درستنمایی $L_k(\theta;y) \propto f(y \in \mathcal{A}_k;\theta)$ باشد. آنگاه تابع درستنمایی مرکب برای پارامتر $\boldsymbol{\theta}$ به صورت

$$L_C(\theta;y) = \prod_{k=1}^{K} L_k(\theta,y)^{w_k}$$

تعریف می شود، که در آن w_k ها وزن‌هایی غیر منفی هستند (لیندسی، ۱۹۹۸). یک نمونه از توابع درستنمایی مرکب، تابع درستنمایی مرکب زوجی موزون است. بویلاکوا و همکاران (۲۰۱۰) تابع درستنمایی موزون را برای داده‌های فضایی- زمانی به کار بردند و نشان دادند برآوردگرهای درستنمایی مرکب موزون سازگار و به طور مجانبی دارای توزیع گاوسی با واریانسی برابر معکوس اطلاع گودمب می‌باشند. آنها همچنین با مطالعات شبیه‌سازی نشان دادند که این برآورد تقریب مناسبی از برآورد ماکسیمم درستنمایی است و نسبت به برآوردهای ماکسیمم درستنمایی و درستنمایی مرکب دارای محاسبات کمتری است. جو و لی (۲۰۰۹) تابع درستنمایی مرکب موزون را برای داده‌های دسته‌بندی شده با تعداد دسته‌های متغیر به کار بردند و معیار کارایی نسبی مجانبی[9] را برای وزن‌های مختلف بررسی کردند و نشان دادند که موزون کردن تابع درستنمایی

[9] Asymptotic Relative Efficiencies

مرکب موجب افزایش کارایی نسبی مجانبی می‌شود. تابع درستنمایی مرکب زوجی موزون به صورت

$$L_W(\theta; y) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r, y_s; \theta)^{w_{rs}} \qquad (۱.۲)$$

است. بنابراین تابع لگاریتم درستنمایی زوجی موزون به صورت

$$\ell_w(\theta; y) = \sum_r \sum_{s>r} w_{rs} \log f(y_r, y_s; \theta) \qquad (۲.۲)$$

خواهد بود، که در اینجا برای مدل‌های SGLM از تابع

$$w_{i,j} = \begin{cases} 1 & ||s_i - s_j|| \leq d_s \\ 0 & \text{در غیر این‌صورت} \end{cases} \qquad (۳.۲)$$

برای موزون کردن تابع درستنمایی مرکب زوجی استفاده می شود. در اینصورت برآورد ماکسیمم درستنمایی مرکب زوجی موزون θ که تابع (۲.۲) را ماکسیمم می‌کند، تحت شرایط نظم، برابر جواب منحصر به فرد معادله $u(\theta; y) = \bigtriangledown_\theta \ell_w = 0$ است. تابع درستنمایی زوجی برای مدل‌های SGLM به صورت

$$L_W(\boldsymbol{\eta}; \mathbf{y}) = \prod_{(i,j)\in\chi} L(\eta|y_i, y_j) \propto \prod_{(i,j)\in\chi} \int\int f(y_i|x_i) f(y_j|x_j) f(x_i, x_j|\eta) dx_i dx_j \qquad (۴.۲)$$

است، که در آن χ زیرمجموعه همسایگی‌های زوجی (y_i, y_j) است.

۲.۲ تابع درستنمایی زوجی تاوانیده

روش ماکسیمم درستنمایی در برآورد پارامترها در مسائل گوناگون، گاهی با مسئله بیش برازشی، دقت پایین یا واریانس زیاد برآوردگرها همراه است. تاوانیدن تابع درستنمایی، یک راه حل برای اصلاح چنین رفتارهایی در برآوردگرها به روش ماکسیمم درستنمایی معمولی است (آزالینی و واله، ۲۰۱۳) که روش ماکسیمم درستنمایی تاوانیده نامیده می‌شود. استفاده از تقریب تابع درستنمایی به جای درستنمایی معمولی به دلیل ماهیت تقریبی آن، می‌تواند بر کاهش دقت برآوردگرها اثر بگذارد. برای داده‌های فضایی حجیم که به دست آوردن تابع درستنمایی به روش تحلیلی غیرممکن است، می‌توان از تابع درستنمایی مرکب برای برآورد پارامترها استفاده کرد. تابع درستنمایی مرکب یک تقریب از تابع درستنمایی مشاهدات است، لذا ممکن است برآوردکننده ها بر اساس این تابع دقت پایینی داشته باشند. برای افزایش دقت برآورد در مدلهای SGLM بر اساس این تابع، میتوان از تاوانیدن تابع درستنمایی مرکب استفاده کرد. در این روش، به لگاریتم درستنمایی مشاهدات یک تابع تاوان اضافه می‌شود. برای مدل‌های SGLM، تابع درستنمایی زوجی تاوانیده به صورت

$$\ell(\eta; y) = \sum_r \sum_{s>r} \log f(y_r, y_s; \eta) - \lambda J(\eta)$$

است، که در آن λ پارامتر هموارای و $J(\eta)$ تابع تاوان است که به صورت‌های گوناگون انتخاب می‌شود. برای مثال فرث (۱۹۹۳) تابع تاوان به صورت $\lambda \frac{1}{2} \log|I(\eta)|$ را برای خانواده نمایی به کار برد، که در آن $I(\eta)$ پیشین جفریز پارامتر η است. تیبشیرانی (۱۹۹۶) تابع تاوان لاسو (LASSO) به صورت $J_\lambda(\eta) = \lambda\eta$ را برای برآورد در مدل‌های خطی ارائه کرد. در این مقاله از تابع تاوان $2\eta\eta^T$ استفاده شده است (گرین ، ۱۹۹۰)، که در آن $\eta^T = (\beta_0, \beta_1, \sigma^2, \varphi)^T$ بردار پارامترهای مدل است.

۳.۲ الگوریتم EM برای ماکسیمم‌سازی تابع درستنمایی زوجی

حسینی و محمدزاده (۱۳۹۲) الگوریتم گرادیانت ماکسیمم‌سازی امیدریاضی زوجی تقریبی[۱۰] (APEMG) را برای مدل‌های SGLM با متغیر پنهان فضایی ارائه کردند. این الگوریتم بر اساس تقریب تابع چگالی شرطی متغیر پنهان ارائه شده است. ایدسویک و همکاران (۲۰۱۲) نشان دادند که اگر X دارای توزیع نرمال و توزیع Y متعلق به خانواده نمایی باشد، با خطی کردن $f(y|x)$ حول یک مقدار ثابت x^* تابع چگالی یا جرم احتمال $f(x|y,\eta)$ را می‌توان با توزیع نرمال به صورت

$$\hat{f}(x|y,\eta) \approx N_n(\hat{\mu}_{x|y,\eta}(y,x^*), \hat{\Sigma}_{x|y,\eta}(x^*)) \qquad (۵.۲)$$

تقریب زد، که در آن

$$\hat{\mu}_{x|y,\eta}(y,x^*) = H\beta + \Sigma_\theta R^{-1}(z(y,x^*) - H\beta),$$

$R = \Sigma_\theta + P$، P ماتریس قطری $n\times n$ با عناصر $Pi = 1/b''(x_i)$، $z_i(x_i^*, y_i) = [y_i - b'(x_i^*) + x_i^* b''(x_i^*)]/b''(x_i^*)$ و $\hat{\Sigma}_{x|y,\eta}(x^*) = \Sigma_\theta - \Sigma_\theta R^{-1}\Sigma_\theta$. از رابطه (۵.۲) طبق خاصیت بسته بودن حاشیه‌سازی توزیع نرمال می‌توان نوشت $(x_{ij}|y_{ij,\eta}) \approx N_2(\hat{\mu}_{ij}, \hat{\Sigma}_{ij})$، که در آن $x_{ij} = (x_i, x_j)'$، $y_{ij} = (y_i, y_j)'$ و $\hat{\mu}_{ij}$ برداری دو بعدی با مولفه‌های برگرفته از سطرهای i و j از بردار $\hat{\mu}_{x|y}$ است. همچنین $\hat{\Sigma}_{ij}$ یک ماتریس ۲ × ۲ از سطر i و ستون j از ماتریس کواریانس $\hat{\Sigma}_{x|y}$ است. در این مقاله، برای ماکسیمم کردن تابع درستنمایی زوجی موزون از الگوریتم APEM استفاده خواهد شد. گام‌های ۱ تا ۳ این الگوریتم همانند الگوریتم APEMG است و در گام ۴، برآورد پارامترها از ماکسیمم کردن تابع

$$Q(\eta|\eta^{(m)}) = \sum_{(i,j)\in\chi} w_{ij} \int\int \log\{f(x_i, x_j, y_i, y_j; \eta)\} f(x_i, x_j|y_i, y_j; \eta^{(m)}) dx_i dx_j. \qquad (۶.۲)$$

به دست می‌آید. برای محاسبه انتگرال دوگانه فوق از الگوریتم MCMC و تقریب تابع توزیع شرطی (۵.۲) استفاده می‌شود.

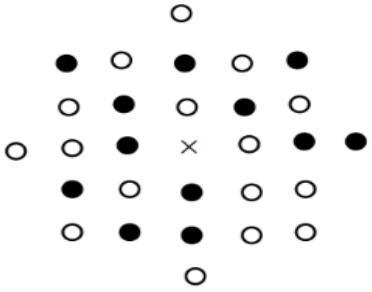
۳ مطالعه شبیه‌سازی

در این بخش تاثیر موزون کردن تابع درستنمایی زوجی بر دقت برآورد پارامترهای مدل SGLM در یک مطالعه شبیه شازی بررسی می‌شود. برای این منظور، داده‌ها از یک مشبکه منظم ۱۰ × ۱۰ با گره‌های $\{(\ell,k), \ell,k = 1,\ldots,10\}$ تولید شده‌اند. برای تولید متغیرهای پنهان فضایی X از توزیع نرمال $N_{100}(\beta_1 z, \Sigma_\theta)$، تابع کواریانس نمایی همسان‌گرد $C(h) = \sigma^2 \exp(-h/\varphi)$ و مقادیر $\sigma^2 = 1$، $\beta_0 = -1$ $\beta_1 = 0/5$ و $\varphi = 2$ در نظر گرفته شده‌اند. متغیر تبیینی در هر موقعیت $s = (\ell,k)$ به صورت $Z_{\ell,k} = \log(1+\ell)$ در نظر گرفته شده است. متغیرهای پاسخ $Y_{\ell,k}$ نیز با شرطی کردن روی متغیرهای پنهان فضایی از توزیع $Bin(100, \exp(x_{\ell,k})/[1+\exp(x_{\ell,k})])$ تولید شده‌اند. پارامترها با الگوریتم APEM برآورد شده‌اند. برای هر مشاهده ۱۲ زوج همسایه به صورت تصادفی بدون جایگذاری با همسایگی به شعاع ۳ انتخاب شده است. در شکل ۳ برآورد پارامترها با استفاده از تابع درستنمایی زوجی و درستنمایی زوجی وزنی انجام شده و دقت برآورد بر اساس ملاک MSE مورد مقایسه قرار گرفته است. برای به دست آوردن مقدار بهینه تابع وزن، مقادیر مختلف d_s در تابع (۲.۲) مورد بررسی قرار گرفته شده است. نتایج در جدول ۳ و نمودار ۳ نشان می‌دهد که برای برآورد پارامتر β_0، مقدار MSE تابع درستنمایی زوجی موزون برای همه مقادیر مختلف فاصله d_s از تابع درستنمایی زوجی معمولی کمتر است. در برآورد پارامتر β_1، تابع درستنمایی زوجی موزون با مقدار فاصله $d_s = 2/5$ دارای MSE کمتری است. برای برآورد پارامتر σ^2 تابع درستنمایی زوجی موزون با مقدار فاصله $d_s = 2/4$ بهترین عملکرد را دارد و تابع درستنمایی زوجی

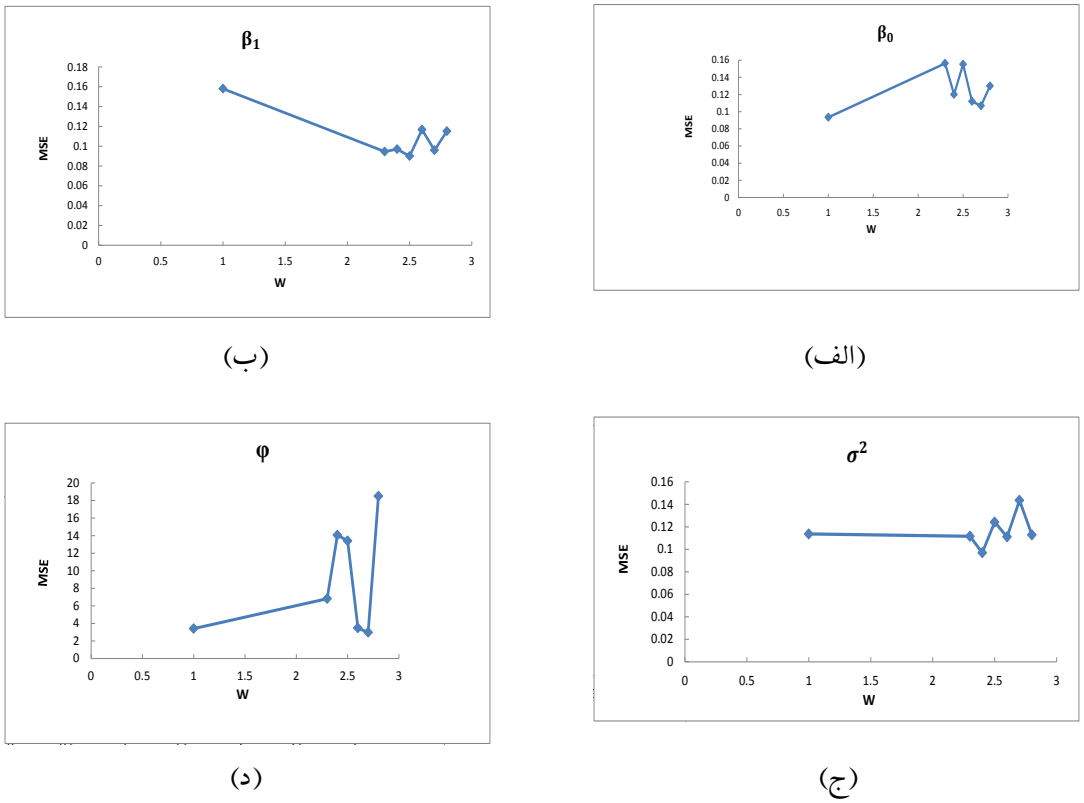
[10] Approximate Pairwise EMG

جدول ۱: برآورد پارامترهای مدلSGLM بر اساس تابع درستنمایی زوجی و درستنمایی زوجی وزنی

تابع درستنمایی	تابع وزن	پارامتر	برآورد	MSE	SE
معمولی	۱	β_0	−۰/۸۶۷۸	۰/۱۵۸۰	۰/۰۳۷۷
		β_1	۰/۵۶۵۳	۰/۰۹۳۴	۰/۰۳۰۵
		σ^2	۰/۷۱۲۵	۰/۱۱۳۷	۰/۰۱۷۷
		φ	۱/۶۶۶۶	۳/۴۱۱۱	۰/۱۸۲۶
	۲/۳	β_0	−۰/۸۶۶۱	۰/۱۵۶۳	۰/۰۳۷۴
		β_1	۰/۴۶۳۸	۰/۰۹۴۶	۰/۰۳۰۷
		σ^2	۰/۷۱۵۲	۰/۱۱۱۶	۰/۰۱۷۶
		φ	۲/۰۰۷۸	۶/۸۱۱۴	۰/۳۷۵۶
	۲/۴	β_0	−۰/۹۰۸۹	۰/۱۲۰۲	۰/۰۳۳۶
		β_1	۰/۴۷۳۶	۰/۰۹۷۰	۰/۰۳۱۲
		σ^2	۰/۷۴۹۹	۰/۰۹۷۰	۰/۰۱۸۷
		φ	۲/۳۰۵۷	۱۴/۰۵۶۴	۰/۳۷۵۶
	۲/۵	β_0	−۰/۸۸۸۱	۰/۱۵۵۲	۰/۰۳۸۰
		β_1	۰/۴۶۹۰	۰/۰۹۰۰	۰/۰۳۰۰
		σ^2	۰/۶۹۸۹	۰/۱۲۴۲	۰/۰۱۸۴
		φ	۲/۱۶۱۰	۱۳/۴۱۴۳	۰/۳۶۷۷
موزون	۲/۶	β_0	−۰/۹۲۱۹	۰/۱۱۲۰	۰/۰۳۲۷
		β_1	۰/۵۱۸۹	۰/۱۱۶۷	۰/۰۳۴۳
		σ^2	۰/۷۱۰۵	۰/۱۱۱۱	۰/۰۱۶۶
		φ	۱/۶۸۵۳	۳/۴۸۱۹	۰/۱۸۴۹
	۲/۷	β_0	−۰/۹۱۵۳	۰/۱۰۶۹	۰/۰۳۱۷
		β_1	۰/۴۷۹۱	۰/۰۹۶۰	۰/۰۳۱۱
		σ^2	۰/۶۸۴۵	۰/۱۴۳۶	۰/۰۲۱۱
		φ	۱/۹۵۸۵	۲/۹۶۸۴	۰/۱۷۳۱
	۲/۸	β_0	−۰/۹۳۴۰	۰/۱۳۰۰	۰/۰۳۵۶
		β_1	۰/۴۸۷۶	۰/۱۱۵۲	۰/۰۳۰۰
		σ^2	۰/۷۲۵۴	۰/۱۱۲۹	۰/۰۱۹۴
		φ	۱/۹۹۳۷	۱۸/۴۹۸۳	۰/۴۳۲۳



شکل ۱: نمونه‌گیری زوجی در همسایگی به شعاع ۳، موقعیت‌های با علامت ضربدر و دایره‌ای توپر یک زوج مرتب هستند



شکل ۲: نمودار MSE بر اساس توابع درستنمایی معمولی و موزون برای پارامترهای مدل

موزون یا مقدار فاصله $d_s = ۲/۷$ در بین سایر توابع عملکرد بهتری در برآورد پارامتر همبستگی φ دارد. بنابراین استفاده از تابع درستنمایی زوجی موزون در برآورد پارامترهای مدل‌های آمیخته خطی تعمیم‌یافته فضایی دارای دقت بالاتری نسبت به تابع درستنمایی زوجی معمولی است.

در ادامه، برای بررسی تاثیر تاوانیدن تابع درستنمایی زوجی بر روی دقت برآورد پارامترها بر اساس معیار MSE، تابع درستنمایی زوجی تاوانیده مورد مطالعه قرار گرفته شده است. برای برآورد پارامترها از الگوریتم APEM استفاده شده که در مرحله M آن تابع $Q(\eta|\eta^m) - \lambda J(\eta)$ ماکسیمم می‌شود. همان‌طور که در جدول ۳ ملاحظه می‌شود، مقدار کمیتهای MSE و SE در برآورد پارامترهای مدل با استفاده از تابع درستنمایی زوجی تاوانیده کمتر از مقدار این کمیت‌ها در تابع

جدول ۲: برآورد پارامترهای مدل باSGLM توابع درستنمایی زوجی و زوجی تاوانیده

تابع درستنمایی	پارامتر	برآورد	MSE	SE
زوجی تاوانیده	β_0	−۰/۸۹۰۸	۰/۱۳۷۳	۰/۰۳۵۵
	β_1	۰/۳۹۹۹	۰/۱۳۷۹	۰/۰۲۷۱
	σ^2	۰/۷۳۰۲	۰/۱۰۱۰	۰/۰۱۶۹
	φ	۱/۸۹۴۷	۲/۲۵۰۹	۰/۱۵۰۴
زوجی	β_0	−۰/۸۶۷۸	۰/۱۵۸۰	۰/۰۳۷۷
	β_1	۰/۵۶۵۳	۰/۰۹۳۴	۰/۰۳۰۵
	σ^2	۰/۷۱۲۵	۰/۱۱۳۷	۰/۰۱۷۷
	φ	۱/۶۶۶۶	۳/۴۱۱۱	۰/۱۸۲۶

درستنمایی زوجی است. برای مثال در برآورد پارامتر β_0 بر اساس تابع درستنمایی زوجی تاوانیده، مقدار MSE و SE به ترتیب ۰/۱۳۷۳ و ۰/۰۳۵۵ هستند در حالی که مقدار این کمیت‌ها در برآورد با استفاده از تابع درستنمایی زوجی، به ترتیب ۰/۱۵۸۰ و ۰/۰۳۷۷ هستند. در برآورد پارامترهای σ^2 و φ نیز استفاده از تابع درستنمایی تاوانیده دقت بالاتری نسبت به درستنمایی زوجی فراهم می کند. بنابراین می‌توان گفت که تاوانیدن تابع درستنمایی زوجی تا حدود زیادی توانسته دقت برآورد پارامترها را افزایش دهد.

بحث و نتیجه‌گیری

در این مقاله، تابع درستنمایی زوجی، درستنمایی زوجی موزون و درستنمایی زوجی تاوانیده برای مدل‌های آمیخته خطی تعمیم‌یافته فضایی به کار گرفته شد. برای ماکسیمم‌سازی این توابع، الگوریتم APEM مورد استفاده قرار گرفت. بر اساس مشاهدات شبیه‌سازی شده، دقت برآورد پارامترها بر اساس این دو تابع و بکار گیری ملاک MSE مقایسه شد. نتایج شبیه‌سازی نشان دادند که تابع درستنمایی زوجی موزون و تابع درستنمایی زوجی تاوانیده در برآورد پارامترهای مدل آمیخته خطی تعمیم‌یافته فضایی، دارای دقت بالاتری نسبت به تابع درستنمایی زوجی هستند. بعلاوه تابع درستنمایی زوجی تاوانیده در برآورد پارامترهای همبستگی و تابع درستنمایی زوجی موزون در برآورد پارامترهای رگرسیونی دارای دقت بالاتری نسبت به سایر توابع بودند. در مقایسه با تابع درستنمایی زوجی موزون با توجه مقادیر MSE دو جدول، تابع درستنمایی زوجی تاوانیده در برآورد دو پارامتر σ^2 و φ دارای دقت بالاتری است.

مراجع

حسینی، ف.، کریمی، ا.، و محمدزاده، م. (۱۳۹۲)، استنباط شبه‌درستنمایی پاسخ‌های فضایی گسسته (مطالعه موردی داده‌های بارندگی استان سمنان)، *نشریه علوم دانشگاه خوارزمی*، **۳**، ۸۰۸ـ۷۹۷.

Azzalini, A., and Arellano-Valle, R. B. (2013), Maximum Penalized Likelihood Estimation for Skew-Normal and Skew-t Distributions, *Journal of Statistical Planning and Inference*, **143**, 419-433.

Baghishani, H., and Mohammadzadeh, M. (2011), A Data Cloning Algorithm for Computing Maximum Likelihood Estimates in Spatial Generalized Linear Mixed Models, *Computational Statistics and Data Analysis*, **55**, 1748-1759.

Bevilacqua, M., Mateu, J., Porcu, E., Zhang, H., and Zini, A. (2010), Weighted Composite Likelihood-Based Tests for Space-Time Separability of Covariance Functions, *Statistics and Computing*, **20**, 283-293.

Christensen, O. F., and Waagepetersen, R. P. (2002), Bayesian Prediction of Spatial Count Data Using Generalized Linear Mixed Models, *Biometrics*, **58**, 280-286.

Diggle, P., Tawn, J. A. and Moyeed, R. A. (1998), Model-Based Geostatistic, *Royal Statistical Society, Series C*, Applied Statistics, **47**, 299-350.

Eidsvik, J., Finley, A. O., Banerjee, S., and Rue, H. (2012), Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models, *Computational Statistics and Data Analysis*, **56**, 1362-1380.

Firth, D. (1993), Bias Reduction of Maximum Likelihood Estimates, *Biometrika*, **80**, 27-38.

Green, P. J. (1990). On Use of the EM for Penalized Likelihood Estimation. *Journal of the Royal Statistical Society, Series B***(Methodological)**, 443-452.

Heagerty, P. and Lele, S. (1998), A Composite Likelihood Approach to Binary Spatial Data, *Journal of American Statistical Assosiation*, **93**, 1099-1111.

Joe, H., and Lee, Y. (2009). On Weighting of Bivariate Margins in Pairwise Likelihood, *Journal of Multivariate Analysis*, **100**, 670-685.

Lindsay, B. (1988), Composite Likelihood Methods, *Contemporary Mathematics*, **80**, 220-239.

McCulloch, C. (1997), Maximum Likelihood Algorithms for Generalized Linear Mixed Models, *Journal of the American Statistical Association*, **92**, 162-170.

McCullagh, P. (1983), Quasi-Likelihood Functions, *Annals of Statistics*, **11**, 59-67.

McCullagh, C. E., and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall. London.

Mohammadzadeh, N. and Hosseini, F. (2011), Maximum-Likelihood Estimation for Spatial GLM Models, *1st Conference on Spatial Statistics 2011*.

Nelder, J. A., and Wedderburn, R. W. M. (1972), Generalized Linear Models, *Journal of the Royal Statistical Association, Series A*, **135**, 370-384.

Rue, H., and Martino, S. (2007), Approximate Bayesian Inference for Hierarchical Gaussian Markov Random Field Models, *Journal of Statistical Planning and Inference*, **137**, 3177-3192.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B*(Methodological), 267-288.

Torabi, M. (2015), Likelihood Inference for Spatial Generalized Linear Mixed Models, *Communications in Statistics-Simulation and Computation*, **44**, 1692-1701.

Varin, C., Host, G., and Skare, O. (2005), Pairwise Likelihood Inference in Spatial Generalized Linear Mixed Models, *Computational Statistics and Data Analysis*, **49**, 1173-1191.

Wedderburn, R. (1974), Quasi-Likelihood Functions, Generalized Linear Models and the Gaussnewton Method, *Biometrica*, **61**, 973-981.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

مروری بر نمونه گیری مجموعه‌رتبه‌دار در جوامع فضایی

مسعود عظیمیان[۱]**، محمد مرادی**
گروه آمار، دانشگاه رازی کرمانشاه

چکیده: امروزه تحلیل‌های آمار فضایی به عنوان ابزاری مفید و کارامد مورد توجه بسیاری از محققین قرار گرفته است. هنگامی که هدف پژوهش بررسی جوامع فضایی است، روش‌های نمونه‌گیری کلاسیک از جمله روش نمونه‌گیری مجموعه‌رتبه‌دار، همبستگی فضایی بین واحدها را در نظر نمی‌گیرند. از این رو انتظار می‌رود که با تلفیق این روش‌ها و همبستگی فضایی بین واحدهای نمونه‌گیری، کارایی این روش‌ها افزایش یابد. در این مقاله به مروری بر کاربرد نمونه‌گیری مجموعه‌رتبه‌دار در جوامع فضایی پرداخته می‌شود. مقالاتی که در این زمینه کار شده است و نتایج آن‌ها ارائه می‌گردد.

واژه‌های کلیدی: نمونه‌گیری مجموعه‌رتبه‌دار، نمونه‌گیری فضایی، همبستگی فضایی، برآوردگر فضایی
کد موضوع‌بندی ریاضی (۲۰۱۰): 62D05، 62M30.

۱ مقدمه

نمونه‌گیری مجموعه‌رتبه‌دار برای اولین بار توسط مک‌اینتایر (۱۹۵۲) به منظور برآورد میزان محصول علوفه معرفی گردید. این طرح نمونه‌گیری در مواردی که اندازه‌گیری متغیر موردنظر بسیار هزینه‌بر، زمان‌بر و گاهاً خطرناک است اما رتبه‌بندی واحدها در مجموعه‌های کوچکتر آسان است، برآوردگرهای دقیقی در مقایسه با طرح نمونه‌گیری تصادفی ساده بدون‌جایگذاری به دست می‌دهد. در حالت کلی نحوه اجرای این طرح نمونه‌گیری به این صورت است که ابتدا یک نمونه تصادفی ساده بدون‌جایگذاری $(SRSWOR)$ [۱] با اندازه m از جامعه‌ای متناهی با اندازه N واحدی با میانگین $\bar{Y}$ ، و واریانس σ^2 استخراج و بطور تصادفی به m مجموعه هرکدام با اندازه m تقسیم می‌کنیم. سپس واحدهای هر مجموعه را صورت صعودی بر اساس یک متغیر کمکی یا هر اطلاعات کمکی دیگری مرتب کرده، از مجموعه اول، واحد متناظر با اولین آماره ترتیبی، از مجموعه دوم واحد متناظر با دومین آماره ترتیبی و ... و از مجموعه mام واحد متناظر با mامین آماره مرتب انتخاب

[1] Simple Random Sampling Without Replacement

[۱] نام و ایمیل ارائه دهنده مقاله: مسعود عظیمیان، azimian.masoud@razi.ac.ir

می‌شود. مراحل با هم را یک چرخه می‌نامند که نتیجه آن یک نمونه مجموعه‌رتبه‌دار به اندازه m است. به منظور دست‌یابی به یک نمونه با اندازه $n = rm$، می‌توان این چرخه را به تعداد r بار تکرار کرد.

هنگامی که با جوامع فضایی کار می‌شود، روش‌های نمونه‌گیری کلاسیک از جمله روش نمونه‌گیری مجموعه‌رتبه‌دار، همبستگی فضایی بین واحدها را در نظر نمی‌گیرند. با در نظر گرفتن جنبه‌ی فضایی(مکانی) واحدها، همبستگی فضایی میان واحدها آشکار می‌شود. بنابراین واحدهایی که در همسایگی هم قرار دارند تمایل به همگنی بیشتری دارند. بنابراین انتخاب یک واحد از این نقاط، انتخاب سایر نقاطی که در یک همسایگی با نقطه انتخاب شده هستند کمک زیادی به شناسایی جامعه هدف نمی‌کند. از این رو انتظار می‌رود با در نظر گرفتن اطلاعات مربوط به همبستگی فضایی واحدهای نمونه‌ای، کارایی روش نمونه‌گیری افزایش یابد. محققانی چون ساهو و همکاران (۲۰۰۶)، اربیا (۱۹۹۳)، هدایت و همکاران (۱۹۸۸) و کانکیور و ری (۲۰۰۸) این چنین موقعیت‌هایی را در نظر گرفته و روش‌هایی را برای انتخاب نمونه با استفاده از نمونه‌گیری فضایی ارائه کردند که در آن به واحدهایی که با نمونه‌های قبلی همگنی دارند و در همسایگی آن‌ها هستند احتمال کمی را برای انتخاب شدن نسبت می‌دهد. اربیا (۱۹۹۳) روش دنباله‌ای واحد وابسته(DUST) [۲] راپیشنهاد داد که یک فرآیند انتخاب نمونه برای انتخاب واحدهای فضایی از یک جامعه همبسته‌ی فضایی می‌باشد. در این روش ابتدا یک واحد به تصادف انتخاب می‌شود و باقیمانده واحدهای نمونه‌ای به ترتیب بر اساس وزنی که می‌گیرند، انتخاب می‌شوند به این ترتیب که واحدی که به واحد اولیه نزدیک‌تر است وزن کمتری می‌گیرد، لذا احتمال کمتری را برای انتخاب در نمونه نسبت به واحدهای دورتر که زودتر انتخاب شده‌اند دارد. اربیا (۱۹۹۳) به صورت تجربی نشان داد که این تکنیک به طور تقریبی ۳۰٪ از کارایی را نسبت به نمونه‌گیری تصادفی ساده افزایش می‌دهد.

برآودگرهای فضایی بر اساس همبستگی فضایی واحدهای نمونه‌گیری شده نسبت به برآوردگرهای روش‌های نمونه‌گیری سنتی در جوامع فضایی کاراتر می‌باشند. روش پیش‌بینی (رویال (۱۹۷۰)) یکی از تکنیک‌های مفید است که در آن همبستگی فضایی جوامع متناهی برای پیش‌بینی واحدهای مشاهده نشده براساس فاصله‌هایشان با واحدهای نمونه‌ای مشاهده شده به اشتراک گذاشته می‌شود. یک راه برای پیش‌بینی واحدهای نامعلوم مشاهده نشده جامعه به وسیله روش وزن‌دهی فاصله ای وارون(IDW) [۳] است که توسط دونالد (۱۹۶۸) ارائه شده است.

در این مقاله به مروری بر مقالاتی می‌پردازیم که نمونه‌گیری مجموعه‌رتبه‌دار را در جوامع فضایی مورد بررسی قرار داده‌اند. در بخش دوم مقاله ارائه شده توسط کانکیور و ری (۲۰۰۸) با عنوان، نمونه‌گیری مجموعه رتبه‌دار فضایی در جوامع همبسته فضایی، بررسی و نتایج آن ارئه می‌گردد. در بخش سوم مقاله ارائه شده توسط بیسواز و همکاران (۲۰۱۵) که به بررسی براوردگر فضایی براساس نمونه‌گیری مجموعه‌رتبه‌دار در جوامع متناهی دارای همبستگی فضایی پرداخته‌اند ارائه و بررسی می‌شود.

۲ نمونه‌گیری مجموعه‌رتبه‌دار فضایی در جوامع همبسته فضایی

نمونه‌گیری مجموعه‌رتبه‌دار فضایی (SRSS) [۴] برای اولین بار توسط کانکیور و ری (۲۰۰۸) معرفی شد. در واقع در این مقاله یک تکنیک نمونه‌گیری دومرحله‌ای معرفی می‌شود که در آن نمونه‌گیری مجموعه‌رتبه‌دار با همبستگی فضایی ترکیب می‌شوند. در مرحله اول به منظور تشکیل خوشه‌های فضایی تصادفی (RSC) [۵]، m واحد کلیدی با روش $DUST$ انتخاب می‌شوند. در مرحله دوم مجموعه‌های مرتب‌شده از هرکدام از RSCها انتخاب می‌گردند. همچنین یک برآوردگر نارایب

[2] Dependent Unit sequential Technique
[3] Inverse Distance Weighting
[4] Spatial Ranked Set Sampling
[5] Random Spatial Clustr

برای این طرح و برای میانگین جامعه براساس برآوردگر راج (۱۹۵۶) ارائه کردند.

۱.۲ نحوه اجرای طرح SRSS

فرض کنید پارامتر موردنظر میانگین متغیر Y جامعه، $\bar{Y}$ و X متغیر کمکی که همبستگی بالایی با متغیر اصلی دارد و برای تمام واحدهای جامعه معلوم است باشند. همچنین فرض کنید که این جامعه دارای N واحد همبسته مکانی باشد. به منظور به دست آوردن یک نمونه به اندازه $n = mr$ مراحل زیر باید اجرا گردد.

۱ ـ ابتدا m واحد به روش $DUST$ انتخاب می‌شود.

۲ ـ براساس مکان هرکدام از این m واحد، همه واحدهای جامعه به m خوشه فضایی تصادفی (RSC) تقسیم می‌شوند که هر واحد از جامعه براساس نزدیکی به واحد کلیدی نزدیک به آن، درون یک RSC قرار می‌گیرد.

۳ ـ درون هر RSC، r^2 واحد با روش نمونه‌گیری تصادفی ساده انتخاب کرده و طور تصادفی در r مجموعه دسته‌بندی می‌شوند.

۴ ـ واحدهای هر مجموعه براساس صفت کمکی یا براساس قضاوت چشمی به صورت صعودی مرتب می‌شوند.

۵ ـ پس از مرتب کردن، از مجموعه اول واحد متناظر با اولین اماره ترتیبی، از مجموعه دوم واحد متناظر با دومین آماره ترتیبی ... به همین ترتیب از مجموعه rام ، واحدمتناظر با بزرگترین آماره ترتیبی را از نمونه حاصل انتخاب می‌کنیم.

۶ ـ r واحد انتخاب شده از هر RSC، یک نمونه مجموعه‌رتبه‌دار را تشکیل می‌دهند.

نمونه حاصل از m خوشه فضایی تصادفی با اندازه $n = r * m$ یک نمونه مجموعه رتبه دار فضایی است.

کانکیور و ری (۲۰۰۸) یک فرمول محاسباتی برای احتمال انتخاب واحدهای نمونه‌ای را در این طرح نمونه‌گیری بدست آوردند و براساس آن براوردگر فضایی مجموعه‌رتبه‌داربرای میانگین جامعه را براساس برآوردگر راج (۱۹۵۶) معرفی کردند. نتایج شبیه سازی انجام شده توسط آنها نشان می‌دهد که طرح نمونه‌گیری مجموعه‌رتبه‌دار فضایی $(SRSS)$ ارائه شده که اطلاعات فضایی را در نمونه‌گیری بکار می‌گیرد کاراتر از نمونه‌گیری مجموعه‌رتبه‌دار معمولی و نمونه‌گیری تصادفی ساده است.

۳ روش برآورد فضایی میانگین جامعه تحت طرح نمونه‌گیری مجموعه‌رتبه‌دار

در این بخش مقاله ارائه شده توسط بیسواز و همکاران (۲۰۱۵) با عنوان روش برآورد فضایی میانگین جامعه تحت طرح نمونه‌گیری مجموعه‌رتبه‌دار معرفی می‌گردد. در این مقاله طرح نمونه‌گیری مجموعه‌رتبه‌دار تصادفی شده را برای انتخاب نمونه از یک جامعه فضایی مورد استفاده قرار داده و برآوردگر فضایی مطابق با آن را برای برآورد میانگین جامعه با استفاده از روش IDW از طریق رویکرد پیش‌بینی معرفی می‌گردد.

۱.۳ برآوردگر فضایی در حالت نمونه‌گیری مجموعه رتبه‌دار

پاتیل وهمکاران (۱۹۹۵) نشان دادند که برآوردگر طرح نمونه‌گیری مجموعه‌رتبه‌دار میانگین جامعه ، $\bar{Y}$، برابر است با:

$$\bar{Y} = \frac{1}{mr} \sum_{k=1}^{r} \sum_{i=1}^{m} y_{(i:m)k} \qquad (۱.۳)$$

که در آن $y_{(i:m)k}$، مقدار واحد مرتب شده iام در مجموعه iام از چرخه kام است. فرض کنید که $d_{(i:m)k,j}$ فاصله واحد $(i:m)k,j$ام از نمونه و هر واحد غیرنمونه‌ای U_j، $j \in \bar{s}$، باشد که در آن $\bar{s}$ مجموعه همه واحدهای غیر نمونه‌ای جامعه است که در نمونه انتخاب نشده‌اند. باتوجه به اینکه واحد نمونه‌ای $(i:m)k,j$ به صورت تصادفی انتخاب می‌شود،

$d_{(i:m)k,j}$ مقداری تصادفی است. بنابراین مقدار پیش‌بینی شده‌ی jامین واحد غیر نمونه‌ای، Y_j، $j \in \bar{s}$، با پیروی از دونالد (۱۹۶۸) از رابطه زیر بدست می‌آید.

$$y_{j,p} = \frac{\sum_{k=1}^{r}\sum_{i=1}^{m}\frac{y_{(i:m)k}}{d_{(i:m)k,j}}}{\sum_{k=1}^{r}\sum_{i=1}^{m}\frac{1}{d_{(i:m)k,j}}}, \quad (i:m)k \in s,\ j \in \bar{s} \tag{۲.۳}$$

با ترکیب کردن این مقادیر پیش بینی شده $N - mr$ واحد غیر نمونه‌ای میانگین این واحدها به صورت زیر بدست می‌آید:

$$\bar{y}_{p,RSS} = \frac{1}{N - mr}\sum_{j\in\bar{s}} y_{j,p} \tag{۳.۳}$$

بنابراین با توجه به روش پیش بینی رویال (۱۹۷۰) برآوردگر فضایی تحت طرح نمونه‌گیری مجموعه‌رتبه‌دار ارائه شده توسط بیسواز و همکاران (۲۰۱۵) برای پارامتر میانگین جامعه به صورت زیر می‌باشد:

$$\hat{\bar{Y}}_{SE,RSS} = \frac{[mr\bar{y}_{RSS} + (N - mr)\bar{y}_{p,RSS}]}{N} \tag{۴.۳}$$

نتایج شبیه سازی شده انجام شده توسط آن‌ها نشان داد که برآوردگر فضایی ارائه شده بر اساس نمونه‌گیری مجموعه‌رتبه‌دار با توجه به اینکه همبستگی فضایی واحدهای جامعه را در برآورد میانگین جامعه در نظر می‌گیردکاراتر از برآوردگر معمول نمونه‌گیری مجموعه‌رتبه‌دار و نمونه گیری تصادفی ساده است.

مراجع

Arbia, G. (1993), The Use of GIS in Spatial Statistical Surveys, *International Statistical Review*, **61**, 339-359.

Biswas, A. , Rai, A. Ahmad, T. and Sahoo, P. M. (2015), Spatial Estimation Approach Under Ranked Sst Samling from Spatial Correlation Finite Population, *Journal of the Indian Society of Agricultural Statistics*, **11**, 551-558.

Donald, S. (1968), A Two-Dimentional Interpolation Function for Irregularly-Spaced Data, *Proceedings of the 1968 ACM National Conference*, 517-524.

Hedayat, A. S., Rao, C. R. and Stufken, J. (1988), Sampling Plans Excluding Contiguous Units, *Journal of Statistical Planning and Inference*, **19**, 159-170.

Kankure, A. K. and Rai, A. (2008), Spatial Ranked Set Sampling from Spatially Correlated Population, *Journal of the Indian Society of Agricultural Statistics*, **62**, 221-230.

McIntyre, G. A. (1952), A Mmthod of Unbiased Selective Samling Using Rranked Sets, *Australian Journal of Agricultural Research*, **3**, 387-390.

Patil, G. P., A. K. Sinha and C. Taillie (1995),Finite Population Corrections for Ranked Sset Sampling, *Annals of the Institute of Statistical Mathematics*, **47**, 621-636.

Royall, R. M. (1970), On Finite Population Sampling Theory Under Certain Linear Regression Models, *Biometrika*, **57**, 377-387.

Raj, D. (1956), Some Estimators With Varying Probabilities Without Replacement, *Journal of the American Statistical Association*, **51**, 269-284.

Sahoo, P. M., Singh, R. and Rai A. (2006), Spatial Sampling Procedure for Agricultural Surveys Using Geographical Information System, *Journal of the Indian Society of Agricultural Statistics*, **60(2)**, 134-143.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

پیشگویی فضایی سه بعدی به منظور برآورد مقاومت فشاری تک محوری در یکی از مخازن نفتی ایران

فتانه فخری [۱]، حسین باغیشنی [۱]، بهزاد تخم‌چی [۲]
[۱] گروه آمار، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود
[۲] گروه اکتشاف معدن، دانشکده مهندسی نفت، معدن و ژئوفیزیک

چکیده: روش‌های مدل‌سازی مبتنی بر زمین‌آمار نقش مهمی را در صنعت اکتشاف معدن و نفت ایفا می‌کند. در این صنایع با در نظر گرفتن هزینه‌های زیادی که صرف حفاری می‌شود، شناخت ناحیه مورد مطالعه و پیش‌گویی فضایی سه‌بعدی آن می‌تواند به کم کردن هزینه‌ها و کاهش خطا کمک کند. در این مقاله تغییرنگار سه بعدی و پیش‌گویی فضایی را معرفی کردیم و برای تشریح روش از مجموعه داده ژئومکانیکی مقاومت فشاری تک محوری استفاده کردیم.

واژه‌های کلیدی: پیش‌گویی فضایی سه بعدی، زمین‌آمار، مقاومت فشاری تک محوری
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 97K80.

۱ مقدمه

دنیایی که ما در آن زندگی می‌کنیم مملو از وابستگی‌هاست، در تحقیقات نادیده گرفتن این وابستگی می‌تواند نتایج غلط و زیان‌های را بر پژوهشگر و کاربران آن عرصه متحمل سازد. در علوم مختلف از جمله نفت، زمین‌شناسی، بوم‌شناسی گیاهی و حیوانی، سنجش از دور[1]، پزشکی و ... داده‌هایی که استفاده می‌شود به موقعیت مکانی و جهت قرارگیری از هم بستگی دارد، بطوری که در اغلب موارد هر چه فاصله بین مشاهدات کمتر باشد شباهت بین آن‌ها بیشتر و در فاصله بیشتر شباهت کمتر است. تجزیه و تحلیل این نوع داده‌ها که به داده‌های فضایی شهرت دارند با روش‌های استنباطی آمار کلاسیک ناممکن است. وارد کردن ساختار وابستگی القا شده توسط موقعیت‌های مکانی داده‌ها (ساختار وابستگی فضایی) در تحلیل داده‌ها توسط روش‌های آمار فضایی ممکن شده است.از آمار فضایی در بسیاری از روش‌ها و مدل‌بندی‌های آماری برای داده‌های فضایی استفاده می‌شود(کرسی ، ۱۹۹۳) یکی از زیرشاخه‌های آمار فضایی زمین‌آمار است و به مدل‌ها و روش‌هایی اشاره دارد که از

[1] Remote Sensing

[۱] نام و ایمیل ارائه دهنده مقاله: فتانه فخری، Fatanefakhri67@gmail.com

دو مشخصه پیروی کند. اول اینکه Y_iها $i = ۱, \ldots, n$ مشاهداتی هستند از مجموعه گسسته مشتمل بر مکان نمونه‌های Z_i که در برخی از مناطق فضایی مانند A مشاهده می‌شوند، و دوم هر مقدار مشاهده شده Y_i با اندازه‌گیری مستقیم و یا با روابط آماری برآورد شده باشد(دیگل و ریبرو ، ۲۰۰۷). داده‌های زمین‌آماری در موقعیت‌های ثابت و در ناحیه‌ای پیوسته مشاهده می‌شوند و متغیر مورد بررسی می‌تواند پیوسته و یا گسسته باشد. در حقیقت زمین‌آمار از ابتدای دهه ۱۹۵۰ به بعد در صنعت معدن توسعه یافت که در حال حاضر بطور گسترده‌ای در سایر علوم زیست محیطی برای نقشه‌برداری، نظارت و مدیریت استفاده می‌شود. همچنین با استفاده از مدل‌سازی همبستگی فضایی، زمین‌آمار می‌تواند پیش‌بینی‌های نااریب با کمترین واریانس را ارائه دهد (اولیور و وستر ، ۲۰۱۵). روش‌های مدل‌سازی مبتنی بر زمین‌آمار نقش مهمی را در صنعت اکتشاف (معدن و نفت) ایفا می‌کند. چرا که در این صنایع با در نظر گرفتن هزینه‌های زیادی که صرف حفاری می‌شود، شناخت ناحیه مورد مطالعه و پیشگویی سه‌بعدی آن می‌تواند به کم کردن هزینه‌ها و کاهش خطا کمک کند. پیش‌بینی ویژگی‌های الاستیک ژئومکانیکی مخازن نفتی، به‌طور گسترده، نقش مهمی در بهینه‌سازی و کاهش خطرات و افزایش بهره‌وری حفاری در یک مخزن ایفا می‌کند (نظری استاد و همکاران ، ۲۰۱۸). هر چاه دارای طول و عرض جغرافیایی است و از طرفی هزینه‌های حفاری با افزایش عمق به صورت نمایی رشد می‌کند (عبدل‌آل ، ۲۰۱۳). مدل‌سازی سه‌بعدی متغیرهای تاثیرگذار در حفاری می‌تواند علاوه بر افزایش سرعت و کاهش هزینه، منجر به برنامه‌ریزی بهتر فرآیند حفاری و کاهش مخاطرات طبیعی گردد. با این حال، ساخت مدل‌های سه بعدی قابل اعتماد برای این ویژگی‌ها همچنان با سوال‌ها و گمانه‌هایی همراه است که پاسخ‌گویی به آن‌ها مستلزم تحقیق و توسعه مدل‌های کارا در سه بعد است.

در این پژوهش در بخش ۳ و ۴ به بررسی ناحیه مورد مطالعه می‌پردازیم و نیز تعاریف تغییرنگار و پیش‌گوی فضایی را بازگو می‌کنیم، سپس به پیشگویی سه بعدی مقاومت فشاری تک محوری می‌پردازیم و در نهایت در بخش پایانی به بحث و نتیجه‌گیری خواهم پرداخت.

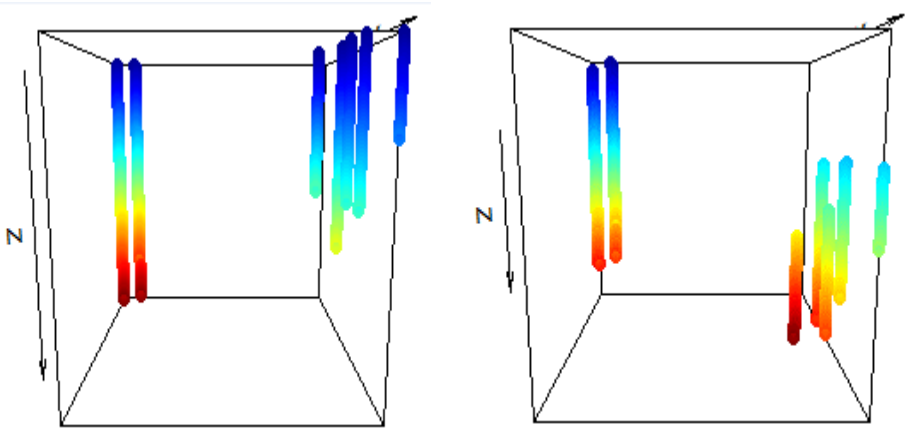
۲ ناحیه تحت مطالعه

داده‌های استفاده شده در این پژوهش مربوط به مقاومت فشاری تک محوری سنگ [۲] (UCS) است. یکی از پارامترهای مهم در ارزیابی مقاومت سنگ و ساخت مدل ژئومکانیک، مقاومت تک محوری سنگر بکر UCS است. معمولاً مقدار مقاومت فشاری از طریق آزمایش بدست می‌آید از طرفی دسترسی به مغزه حفاری مشکل است لذا محققان صنعت نفت روابطی را جهت برآورد پارامترهای مکانیکی از روی نگاره‌های پتروفیزیکی ارائه داده‌اند. در این پژوهش برای محاسبه مقدار مقاومت فشاری تک محوری از این روابط استفاده شده است. منطقه مورد مطالعه یکی از مخازن نفتی ایران شامل ۱۰ حلقه چاه با طول و عرض جغرافیایی و عمق مشخص است که مقدار UCS با استفاده از متغیرهای ثبت شده مرتبط با این متغیر بطور تقریبی محاسبه شده است. لازم به ذکر است که چاه‌ها دارای ضخامت‌های متفاوتی هستند، از طرفی این مخزن بر روی یک چین زمین‌شناسی قرار دارد پس از رابطه (۱.۲) برای یکسان‌سازی عمق چاه‌ها در همان مختصات طول و عرض جغرافیایی استفاده شد.

$$z_{(new)i} = z_{(old)i} - |\min(z_{ij}) - \min(z_{(old)i})|, \quad j = ۱ \ldots ۱۰, \quad i = ۱, \ldots, m_j \qquad (۱.۲)$$

که $z_{(old)i}$ ، $z_{(new)i}$ به ترتیب عمق حقیقی و اصلاح شده برای مکان iام در هر چاه است، $\min(z_{ij})$ کمترین عمق اندازه‌گیری شده در تمام چاه‌ها و m_j نشان دهنده تعداد مشاهدات در هر چاه است. در شکل ۱ نمودار سه بعدی چاه‌ها با

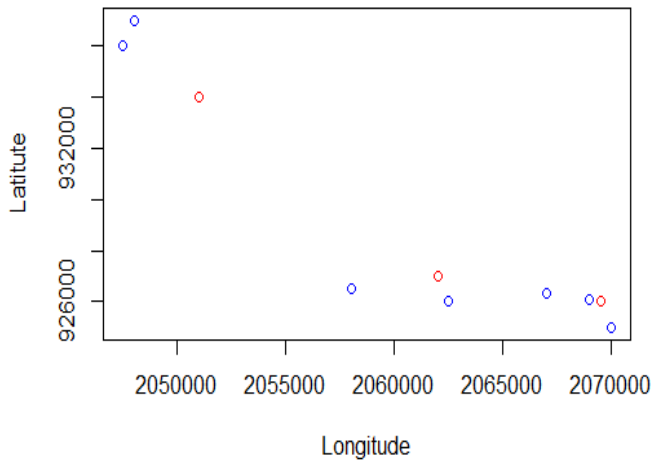
[2] Unconfined Compressive Strength
[3] Test set
[4] Cross validation



(ب) (الف)

شکل ۱ : توضیحات نمودار الف: با عمق حقیقی ب: بعد از یکسان سازی عمق

عمق حقیقی و عمق اصلاحی آورده شده است. از بین چاه‌ها سه چاه که شامل ۸۸۷۳ مشاهده است را بطور تصادفی برای مجموعه داده‌های آزمایشی[۳] انتخاب کردیم تا نتایج را بر اساس اعتبار سنجی متقابل [۴] مورد ارزیابی قرار دهیم و از بقیه چاه‌ها که ۲۸۴۵۵ مشاهده را شامل می‌شود بعنوان مجموعه آموزشی [۵] برای مدل‌سازی سه بعدی استفاده کردیم. شکل ۲ مربوط به محل قرارگیری چاه‌ها در موقعیت‌های مکانی طول و عرض جغرافیایی به تفکیک مجموعه آموزشی و آزمایش است.



شکل ۲ : محل قرارگیری چاه‌های مجموعه آموزشی(سیاه) و مجموعه آزمایشی(قرمز) در دو بعد

[5]Train set

۳ تغییرنگار

در آمار فضایی تغییرنگار نظری $۲\gamma(h)$ واریانس تفاضل بین دو مقدار مشاهده شده میدان تصادفی در دو مکان فضایی که به فاصله h از هم قرار دارند و بصورت

$$۲\gamma(h) = Var(Z(s+h) - Z(s)) \qquad (۱.۳)$$

$\gamma(h)$ تعریف می‌شود، نیم‌تغییرنگار [۶] نامیده می‌شود. تابعی برای توصیف میزان وابستگی مقادیر میدان تصادفی و یا فرآیند برآورد دقیق تغییرنگار برای پیش‌گویی قابل اعتماد توسط کریگینگ مورد نیاز است. تغییرنگار تجربی (نمونه) معمولا توسط روش‌های گشتاوری در دنباله‌ای از گام‌ها محاسبه می‌شود و یک یا چند مدل معتبر (مجاز[۷]) تغییرنگار بر روی آن برازش داده می‌شود. یک تغییرنگار ممکن است در طول یک برش یا توری در فواصل منظم یا در فواصل پراکنده نامنظم محاسبه شود. دقت تغییرنگار به اندازه نمونه، تعداد گام‌هایی که برآورد شده است و نیز به فاصله بازه نسبت به مقیاس فضایی یا متغیر، توزیع کناری ، ناهمسانگردی و روند بستگی دارد. برآوردگرهای تنومند [۸] نسبت به مقادیر دورافتاده [۹] حساس نیستند. تغییرنگار ممکن است دارای محدودیت (کراندار) باشد (برای مانای مرتبه دوم [۱۰]) و یا بدون کران (تنها در مانای ذاتی[۱۱]). پارامترهای تغییرنگار که توسط مدل‌های متعبر تغییرنگار برآورد شده‌اند خلاصه‌ای از تغییرات فضایی را که برای کریگینگ نیاز است را به ما می‌دهند. در صورت وجود ناهمسانگردی در میدان تصادفی محاسبه تغییرنگار می‌تواند حداقل در سه جهت ناهمسانگردی را تعیین کند. میانگین حداقل مربعات باقی‌مانده‌ها [۱۲] و معیار آکایک [۱۳] می‌تواند در انتخاب بهترین مدل برازش داده شده به ما کمک کند. (اولیور و وستر ، ۲۰۱۵) برآورد تغییرنگار معمولا براساس روش گشتاوری ماترون [۱۴] (MOM) به صورت

$$\hat{\gamma}(\mathbf{h}) = \frac{۱}{۲m(\mathbf{h})} \sum_{i=۱}^{m(\mathbf{h})} \{z(\mathbf{x_i}) - z(\mathbf{x_i}+\mathbf{h})\}^۲ \qquad (۲.۳)$$

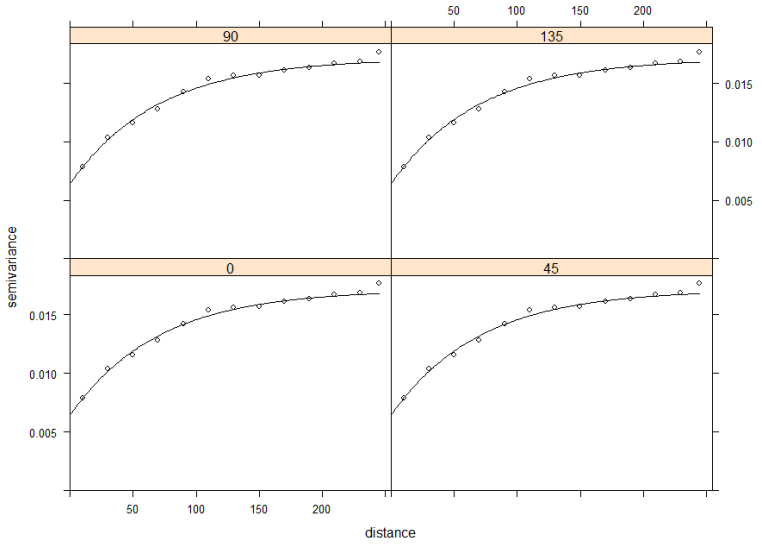
محاسبه می‌گردد. که $z(\mathbf{x_i})$ و $z(\mathbf{x_i}+\mathbf{h})$ مقادیر مشاهده شده zدر موقعیت‌های $\mathbf{x_i}$ و $\mathbf{x_i}+\mathbf{h}$ و $m(\mathbf{h})$ تعداد جفت نمونه‌هایی است که در گام $\mathbf{h}$ قرار دارد. محاسبه تغییرنگار سه‌بعدی در میدان با توری‌های منظم یا نامنظم به این شکل است که می‌توان توری را بصورت یک سری در نظر گرفت و در سه بعد مورد بررسی قرار داد. معادله (۳.۳) مربوط به تغییرنگار سه بعدی است که برای بررسی همسانگردی باید در چند جهت مختلف در طول ردیف، ستون و روی قطر یک توری محسابه شود.

$$\hat{\lambda}(p,q) = \frac{۱}{۲(m-p)(n-q)(l-r)} \sum_{i=z}^{m-p} \sum_{j=۱}^{n-q} \sum_{k=۱}^{l-r} \{z(i,j,k) - z(i+p, j+q, k+r)\}^۲ \qquad (۳.۳)$$

که p ، q و r به ترتیب اندازه گام و m ، n و l تعداد گام‌های ردیف ، ستون و بعد سوم (در اینجا عمق) هستند. نمودار سه‌بعدی نیم‌تغییرنگار تجربی داده‌های مقاومت فشاری در چهار جهت ۰، ۴۵، ۹۰ و ۱۳۵ درجه برای بررسی ناهمسانگردی

[6] Semivariogram
[7] Authorized
[8] Robust estimators
[9] Outliers
[10] Second-order stationary
[11] Intrinsic stationary
[12] Residual mean squares
[13] Akaike
[14] Matheron's method of moments

همراه با منحنی مدل پارامتری نمایی برازش داده شده در شکل ۳ آورده شده است. با توجه به شکل می‌توان دریافت که نیم‌تغییرنگار در این چهار جهت همسان‌گرد است.



شکل ۳: نمودار سه‌بعدی نیم‌تغییرنگار و منحنی مدل پارامتری نمایی برازش داده شده در چهار جهت

۴ پیش‌گویی فضایی

یکی از اهداف اصلی در تحلیل داده‌های فضایی، پیش‌گویی پاسخ در مکان‌های فاقد مشاهده است که با استفاده از روش کریگینگ انجام می‌شود (کرسی ، ۱۹۹۳). اصطلاح گریگنگ به افتخار دانیال کریگ [۱۵] (۲۰۱۳ – ۱۹۱۹) نام‌گذاری شده است، کریگ در استفاده از روش‌های آماری برای ارزیابی و تحلیل مخازن معدنی پیشگام بود و بعدها این روش‌ها توسط ماترون [۱۶] و دیگر همکارانش در دانشکده معدن پاریس فرموله، توسعه و نشر داده شد. (دیگل و جورجی ،۲۰۱۹).

کریگینگ بهترین پیش‌گوی خطی نقطه‌ای و بلوکی است، به‌طوری که بهترین در این‌جا به معنی پیش‌گو با کمترین واریانس است. پیش‌گوی کریگینگ یک میانگین وزنی است که به نقاط نزدیکتر وزن بیشتر و به نقاط دورتر وزن کمتری اختصاص می‌دهد (اولیور و وستر ، ۲۰۱۵). یکی از پذیره‌های روش کریگینگ، مانایی میدان تصادفی است. با معلوم بودن ساختار همبستگی فضایی و در نظر گرفتن تابع زیان توان دوم خطا، پیش‌گوی بهینه (کریگینگ) از می‌نیمم کردن میانگین توان دوم خطای پیش‌گو به دست می‌آید. در روش‌های کلاسیک، میدان تصادفی را گاوسی در نظر می‌گیرند. بر حسب ساختار فضایی میدان تصادفی، سه نوع کریگینگ تعریف می‌شوند: ۱) کریگینگ ساده[۱۷] که در آن فرض می‌شود میانگین میدان معلوم است؛ ۲) کریگینگ عادی[۱۸] که در آن فرض می‌شود میانگین ثابت ولی نامعلوم است؛ و ۳) کریگینگ عام[۱۹] که در آن میانگین به‌عنوان تابعی از موقعیت‌ها (و سایر متغیرهای تبیینی) در نظر گرفته می‌شود و به آن روند می‌گویند. در ادامه، به‌طور مختصر،کریگینگ ساده و عادی را معرفی می‌کنیم.

[15] Daniel G. Krige
[16] Matheron
[17] Simple Kriging
[18] Ordinary Kriging
[19] Universal Kriging

۱.۴ کریگینگ عادی

فرض کنید مقادیر متغیر پاسخ Z را در n مکان ناحیه تحت مطالعه، $D \subset \Re^d$، مثل $\{s_1, \ldots, s_n\}$، مشاهده کرده باشیم و هدف پیش‌گویی مقدار پاسخ در مکان جدید s_0، یعنی $Z(s_0)$، باشد. در کریگینگ عادی تغییرات مشاهدات پاسخ از نظر فضایی به هم وابسته و تصادفی و تنها در یک مقدار نامعلوم μ، مشترک هستند. در این حالت می‌توان میدان تصادفی گاوسی $\{Z(s) : \ s \in D\}$ را به‌صورت زیر تجزیه کرد:

$$Z(s) = \mu + \delta(s)$$

که در آن μ ثابت و نامعلوم و $\delta(\cdot)$ میدان تصادفی مانای گاوسی با میانگین صفر است. کریگینگ عادی، به‌عنوان یک میانگین وزنی از مشاهدات، به‌صورت

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

تعریف می‌شود. ضرایب λ_i وزن‌های کریگینگ هستند که سهم هر کدام از مشاهدات را (بسته به فاصله آن‌ها تا مکان جدید s_0) در پیش‌گویی پاسخ یدک می‌کشند. چنان‌چه شرط $\sum_{i=1}^{n} \lambda_i = 1$ را در نظر بگیریم، کریگینگ یک پیش‌گوی نااریب خواهد بود، یعنی $E[\hat{Z}(s_0)] = \mu$. مقادیر بهینه ضرایب $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$ از حل معادله

$$\lambda' = (\gamma + 1_n \frac{(1 - 1_n' \mathbf{\Gamma}^{-1} \gamma)}{1_n' \mathbf{\Gamma}^{-1} 1_n})' \mathbf{\Gamma}^{-1} \qquad (1.4)$$

به‌دست می‌آیند، که در آن $\mathbf{\Gamma}$ یک ماتریس $n \times n$ با درایه (i, j)

$$\gamma(s_i - s_j) = \frac{1}{2} \mathrm{Var}(Z(s_i) - Z(s_j))$$

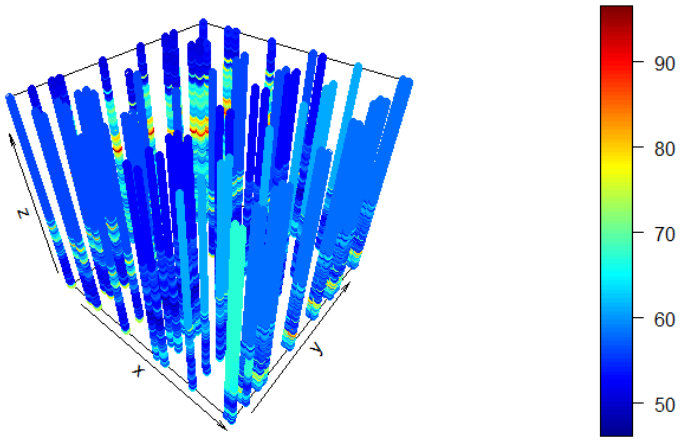
و $\gamma = (\gamma(s_0 - s_1), \gamma(s_0 - s_2), \ldots, \gamma(s_0 - s_0))'$، به‌طوری که $\gamma(\cdot)$ تابع نیم‌تغییرنگار است. همچنین 1_n برداری به طول n از ۱ ها است. در مورد چگونگی رسیدن به این ضرایب و نحوه محاسبه واریانس کریگینگ عادی می‌توانید به کرسی ، (۱۹۹۳) مراجعه کنید.

کریگینگ ساده حالت خاصی از کریگینگ عادی است که در آن فرض می‌شود، بین مشاهدات روندی وجود ندارد و میانگین معلوم و مستقل از مختصات فضایی است. در این حالت شرط نارایبی صفر بودن مجموع ضرایب لازم نیست و معادله محاسبه ضرایب کریگینگ، مشابه (۱.۴)، با کمی تغییرات جزیی نتیجه می‌شود.

۵ پیشگویی سه‌بعدی مقاومت فشاری تک محوری سازند

برای پیش‌گویی داده‌های مقاومت فشاری در سایر نقاط ناحیه تحت مطالعه در شکل ۴ با توجه به پذیره گاوسی بودن میدان، با استفاده از کریگینگ عادی و مشبکه‌بندی ناحیه تحت مطالعه، نقشه پهنه‌بندی پیش‌گویی شده مقدار مقاومت فشاری و واریانس پیش‌گویی را به‌دست آوردیم. شکل ۵ نمودار دو بعدی مقادیر پیش‌گویی و واریانس پیش‌گویی را در سه عمق نشان می‌دهند.

محور افقی شکل ۶ نشان دهنده مقادیر واقعی مجموعه داده‌های آزمایشی Z_i و محور عمودی مقادیر پیشگویی بدست آمده از کریگینگ عادی $\hat{Z}_i$ برای مجموعه آزمایشی است. برای UCS کمتر از ۸۰ توزیع پراکنش داده‌ها حول نیم‌ساز ربع



شکل ۴: پیش‌گویی سه بعدی میدان نفتی تحت مطالعه

اول و سوم $(Z_i = \hat{Z}_i)$ است و بیانگر مناسب بودن مدل و مقدار ناچیز باقی‌مانده‌ها است. با افزایش مقادیر داده‌ها، پیشگو و مقادیر اصلی از نیم‌ساز فاصله گرفته و اندازه خطا بیشتر شده است، به‌عبارتی واریانس خطا ثابت نیست ولی همچنان نیم‌ساز بین داده‌ها قرار دارد و میانگین خطا برابر با صفر است.

بحث و نتیجه‌گیری

در بسیاری از علوم از داده‌های سه‌بعدی برای تحلیل و مدل‌سازی استفاده می‌شود و بررسی مدل‌سازی و پیشگویی در سه بعد می‌تواند به روشن شدن بعضی از ابهامات کمک کند. در این پژوهش از کریگینگ عادی سه بعدی برای پیشگویی فضایی مقادیر مقاومت فشاری تک محوری استفاده شد. ابتدا به طور تصادفی ۷۰ درصد از کل داده‌ها به‌عنوان مجموعه آموزشی در نظر گرفته شد و با برازش مدل مناسب به نمودار نیم‌تغییرنگار سه بعدی داده‌ها پیشگویی ناحیه تحت مطالعه صورت گرفت و در آخر نمودار پراکنش داده‌های مجموعه آزمایشی در مقابل مقادیر پیشگویی رسم شد که می‌توان گفت مناسب بودن مدل و مقدار ناچیز باقی‌مانده‌ها را نشان می‌دهد.

مراجع

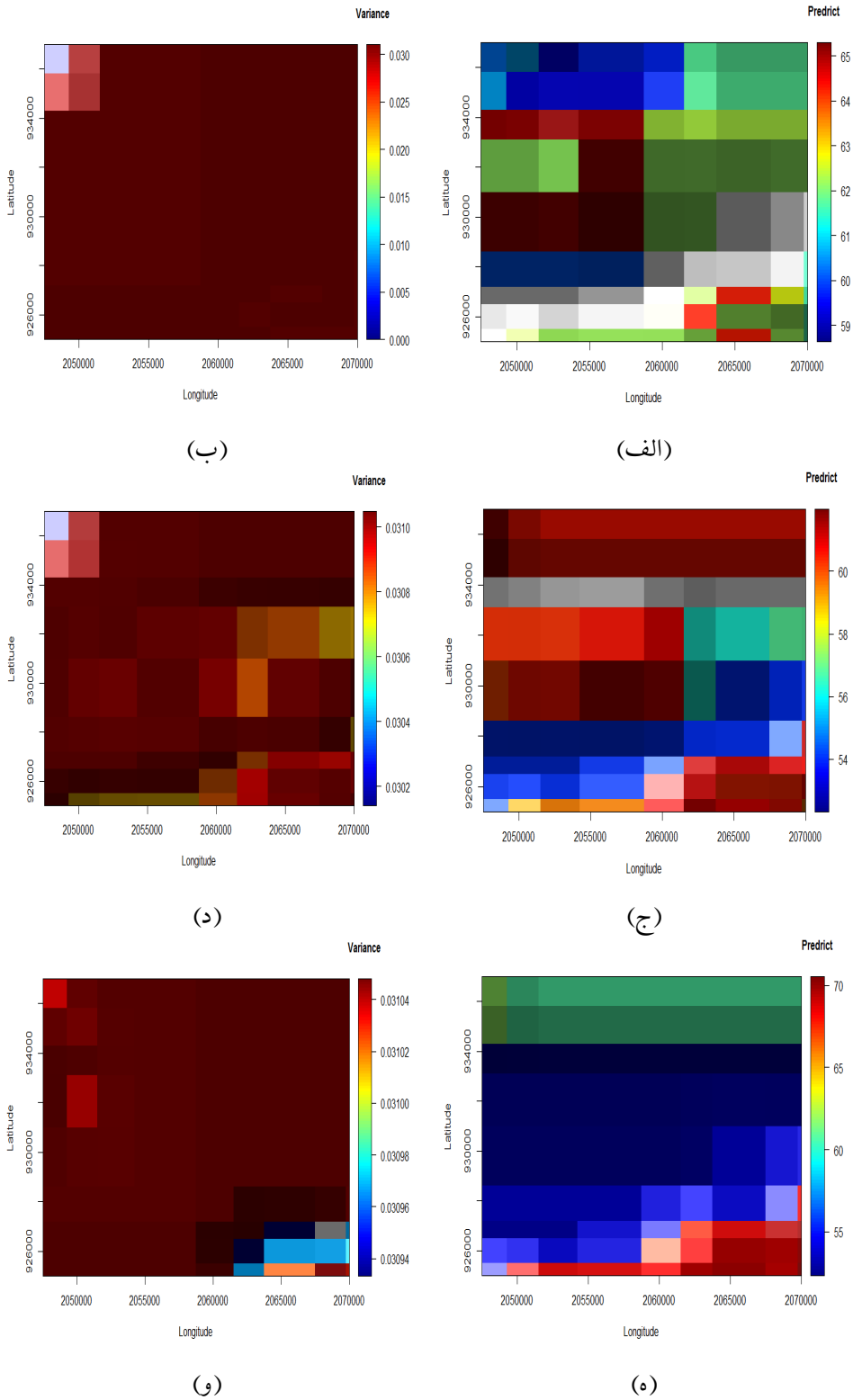
محمدزاده، م.، (۱۳۹۸)، *آمار فضایی و کاربردهای آن*، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

Abdel-Aal, H. K., Alsahlawi, M. A. (Eds.), (2013), *Petroleum Economics and Engineering*, CRC Press.

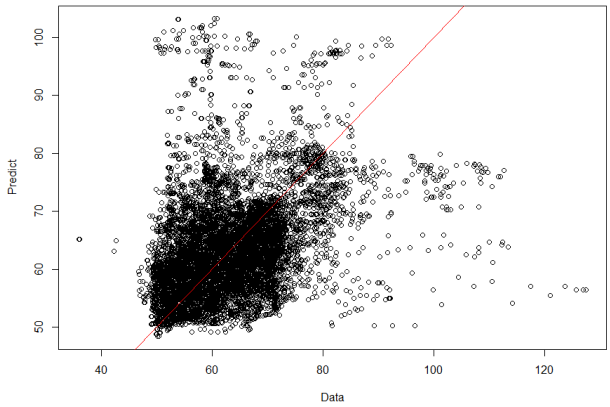
Cressie, N. (1993), *Statistics for Spatial Data*, Hoboken, NJ, USA: John Wiley and Sons, Inc.

Diggle, P. J., Ribeiro, P. J.(2007), *Model Based Geostatistics*, Springer Series in Statistics.

Diggle, P. J., Giorgi, E. (2019), *Model Based Geostatistics for Global Public Health: Methods and Applications*, Chapman and Hall/CRC.



شکل ۵: نقشه پهنه بندی پیش‌گویی مقادیر مقاومت فشاری در عمق‌های ۱۷۰۳، ۲۱۵۵ و ۲۴۸۰ متر به ترتیب در (الف)، (ج) و (ه). همچنین نقشه پهنه‌بندی واریانس پیشگویی در همین عمق‌ها به ترتیب در (ب)، (د) و (و).



شکل ۶: مقادیر پیشگویی در مقابل مقادیر واقعی داده‌های آزمایشی

Oliver, M. A., Webster, R. (2015), *Basic Steps in Geostatistics: the Variogram and Kriging* (Vol. 106), New York: Springer

Ostad, M. N., Niri, M. E., Darjani, M. (2018), 3D Modeling of Geomechanical Elastic Properties in a Carbonate-Sandstone Reservoir: A Comparative Study of Geostatistical Co-Simulation Methods, *Journal of Geophysics and Engineering*, **15**, DOI: 10.1088/1742-2140/aaa983.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

بررسی اثر موقعیت مکانی هتل‌ها بر میزان رضایت‌مندی مشتریان

زهرا فضلعلی [۱]**، کیومرث مترجم**
گروه آمار، دانشگاه تربیت مدرس

چکیده: در این مقاله اثر موقعیت مکانی هتل‌ها بر میزان رضایت‌مندی مشتریان در ایالت کالیفرنیای امریکا طی سال‌های ۲۰۰۲ تا ۲۰۱۵ بررسی شده است. هدف از انجام این مطالعه، تحقق هر چه بیشتر و بهتر خواسته‌های مشتریان وهمچنین دستیابی به موقعیت مکانی مناسب جهت ایجاد هتل بوده است. جامعه آماری این تحقیق شامل همه نظرات ثبت شده‌ای است که در مورد هتل‌های ایالات متحده در برنامه TripAdvisor.com توسط کاربران ثبت و جمع‌آوری شده است. با در نظر گرفتن اطلاعات موقعیت جغرافیایی، این مطالعه به بررسی ۲۲۰ پاسخ درباره هتل‌ها از مجموع ۷۷۰۹۸ نظری که توسط کاربران ثبت شده پرداخته است. با استفاده از تحلیل اکتشافی نشان خواهیم داد که بین رضایت مشتری و موقعیت مکانی هتل‌ها همبستگی فضایی وجود دارد. نتایج این تحقیق در حوزه‌های بازاریابی و رویکردهای مدیریتی صنعت هتل‌داری بسیار کارآمد و مورد استفاده مدیران این حوزه می‌باشد. **واژه‌های کلیدی:** آمار فضایی، رضایت‌مندی مشتری، بازاریابی .
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62M30، 91D25.

۱ مقدمه

برای نشان دادن همبستگی فضایی بین رضایت‌مندی مشتریان از هتل‌ها در سطح شهر سانفرانسیسکو از آزمون موران استفاده می‌کنیم. از آنجا که رضایت مشتری در رویکردهای مدل‌سازی تجربی معمولاً به عنوان متغیر وابسته رفتار می‌کند لذا وجود همبستگی فضایی در این متغیر باعث می‌شود استفاده از مدل‌های معمول در آمار، نتایج مطلوبی به همراه نداشته باشد. بنابراین، برای تجزیه و تحلیل اینگونه داده‌ها به یک الگوی فضایی نیاز پیدا خواهیم کرد. در بخش ۲ این مقاله، داده‌ها و رویکردهای پاکسازی آنها معرفی می‌شود. در بخش ۳ همبستگی فضایی رضایت‌مندی مشتریان، آماره آزمون موران و ماتریس‌های همسایگی معرفی می‌شوند و در بخش آخر به بحث و نتیجه‌گیری می‌پردازیم. رضایت مشتری فاکتوری مهم در موفقیت سازمان‌ها در محیط تجاری است. همه سازمان‌ها می‌کوشند رضایت مشتری را از طریق بهبود کیفیت خدمات یا محصولاتشان افزایش دهند. اما در بسیاری از مواقع این سازمان‌ها به علت کمبود منابع و تصمیم‌گیری نادرست درباره

[۱] نام و ایمیل ارائه دهنده مقاله: زهرا فضلعلی، Zahrafazlali467@gmail.com

اینکه چه مجموعه‌ای از فعالیت‌ها سبب بهترین عملکرد در بهبود رضایت مشتری می‌شود معمولاً به اهداف مشخص شده خود دست پیدا نمی‌کنند. این امر به این دلیل رخ می‌دهد که اساساً بعضی از عوامل نقش اصلی را در رضایت مشتریان یک سازمان ایفا نمی‌کنند، لذا شناسایی عوامل حیاتی و تصمیم‌ساز در رضایت مشتریان در موفقیت هر سازمان یا شرکت، نقشی اساسی و ضروری دارد . یکی از مهمترین تحولاتی که در زمینه بهبود عملکرد در دهه آخر قرن بیستم به وقوع پیوست، موضوع شناخته شدن و قابلیت اندازه‌گیری میزان رضایت مشتری به مثابه یکی از عناصر و الزامات اصلی نظام‌های مدیریتی در مؤسسات و بنگاه‌های اقتصادی کسب و کار بود. پیاده‌سازی و اجرای سیستم‌های اندازه‌گیری رضایت مشتری، یکی از مهمترین شاخص‌های بهبود عملکرد در سازمان‌های امروزی به شمار می‌رود. امروزه و در میان انبوه رسانه‌هاي اجتماعی، محتویات تولید شده توسط کاربران و اطلاعات جغرافیایی ارائه شده توسط شهروندان نوع جدیدی از جمع‌آوری داده‌ها را ایجاد کرده است. برای مثال، محتویات تولید شده توسط کاربران روش‌های جدیدی را برای درک بهتر رفتار سفر در مقیاس‌های مختلف فضایی و زمانی ارائه می‌دهد در عین حال، رضایت مشتری به عنوان یکی از شناخته شده‌ترین شاخص‌های ارزیابی مشتری است که توسط محققان مورد استفاده قرار می‌گیرد و شرکت‌های مختلف جهت ارتقاء سطح کیفی خدمات خود از آن استفاده می‌کنند. همچنین با توجه به مطالعات صورت گرفته در حوزه عملکرد محصول جدید محتویات تولید شده توسط کاربران، اطلاعات مهمی در مورد برند، محصولات و عملکرد سهام شرکت‌ها در دسترس قرار می‌دهد. طی دهه‌های گذشته، استفاده از مدل‌های فضایی به یک حوزه تحقیقاتی جدید در بازاریابی تبدیل شده است. بردلو و همکاران نشان دادند که همبستگی فضایی می‌تواند بر میزان رضایت‌مندی مشتریان از یک خدمت تاثیر زیادی داشته باشد.، آنها روشی برای درک فرایندهای تصمیم‌گیری و یافتن همبستگی فضایی در مبحث رضایت‌مندی مشتری معرفی نمودند. با استفاده از روش آنها قصد داریم مطالعه‌ای اکتشافی در خصوص اهمیت جغرافیایی میزان رضایت مشتری از صنعت هتل‌داری را انجام دهیم. این مطالعه بر اساس داده‌های به دست آمده از برنامه TripAdvisor.com در شهر سانفرانسیسکو واقع در ایالت کالیفرنیای امریکا می‌باشد.

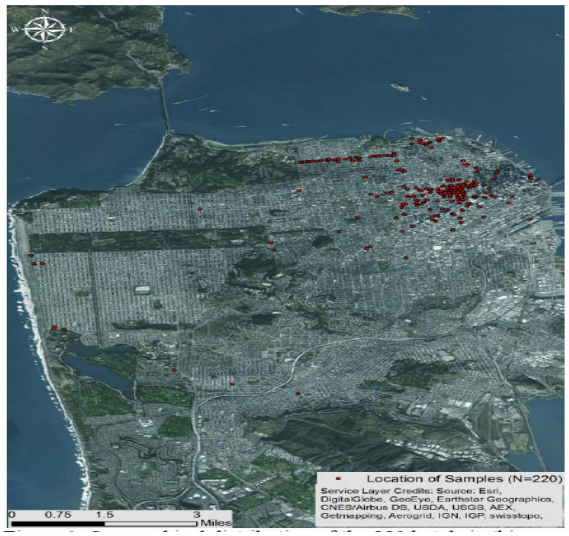
۲ معرفی داده‌ها

در این مطالعه تعداد ۱۷۲۰۵۰ نظر ثبت شده مشتریان در مورد ۲۲۳ هتل در شهر سانفرانسیسکو ایالت کالیفرنیا از تاریخ ۲۲/۰۹/۲۰۰۲ تا ۳۰/۰۳/۲۰۱۵ از طریق TripAdvisor.com جمع‌آوری شده است. این سایت نه تنها یک ابزار پیشگام در تولید محتوا توسط مشتریان است بلکه به عنوان یکی از بزرگترین سایت‌های سیر و سفر در جهان با بیش از ۶۰ میلیون کاربر و بالغ بر ۱۷۰ میلیون بازدیدکننده محسوب می‌شود که فعالیت هتل‌ها، رستوران‌ها، جاذبه‌ها و سایر کسب و کارهای مرتبط در حوزه سرگرمی را در سراسر جهان رصد می‌کند. این سایت پس از بررسی آدرس ایمیل کاربر بر اساس فعالیت‌های او در سایت‌های گردشگری و هتل‌ها اطلاعات آماری و نحوه جستجوی کاربر را جمع‌آوری می‌کند، سپس این اطلاعات در قالب یک گزارش توسط تیمی از متخصصان تحلیل داده ارزیابی می‌شود. نتایج این بررسی‌ها به مالک یا مدیر موسساتی که در فهرست این سایت قرار دارد بصورت پیام ارسال می‌شود. این ویژگی‌ها منجر به افزایش کیفیت محتویات بررسی شده در مقایسه با سایر منابع تولید محتوا و نظر کاربران می‌شود. داده‌های این مطالعه توسط یک تیم بازاریابی با هدف بررسی میزان رضایت مشتریان از خدمات هتل‌ها جمع‌آوری شده است. سپس بر اساس نظر ژانگ و همکاران هتل‌هایی که کاربری برای آنها نظری ثبت نکرده و یا نظرات کاربران ناقص ثبت شده است از مجموعه داده‌ها حذف شده‌اند. در نهایت، برای کنترل اختلافات در فرهنگ بین‌الملل و زمینه‌های سیاسی، بررسی‌های ارائه شده توسط کاربران بین‌المللی حذف شده و فقط نظرات کاربران بومی آن منطقه در دسته‌بندی‌ها قرار گرفته است. در نهایت مجموعه نهایی داده‌ها شامل ۷۷۰۹۸ نظر ارائه شده توسط کاربران ایالات متحده در مورد ۲۲۰ هتل جمع‌آوری شده است که موقعیت مکانی هتل‌ها در شکل ۱ آورده شده

است.

۱.۲ همبستگی فضایی رضایت‌مندی مشتریان

در اغلب روش‌های آماری فرض بر این است که مشاهدات تحت شرایط یکسان و به صورت مستقل از هم جمع‌آوری شده‌اند. اما اگر مشاهدات مستقل نباشند و وابستگی آنها تابعی از فاصله بین موقعیت‌های مشاهدات باشد، به‌گونه‌ای که مشاهدات نزدیک به هم وابسته‌تر و مشاهدات دورتر از هم وابستگی کمتری داشته باشند، پس مشاهدات مربوطه، داده‌های فضایی نامیده می‌شوند. محمدزاده مجموعه روش‌های تحلیل داده‌های فضایی را آمار فضایی نامیده است. برای تحلیل داده‌های فضایی لازم است یک مدل آماری در نظر گرفته شود. در آمار فضایی به‌طور معمول یک میدان تصادفی به‌عنوان مدل آماری داده‌های فضایی در نظر گرفته می‌شود.



شکل ۱: نظرات کاربران در خصوص کیفیت خدمات هتل‌ها بر اساس توزیع جغرافیایی ۲۲۰ هتل در سانفرانسیسکو

میدان تصادفی مجموعه‌ای از متغیرهای تصادفی مانند $\{Z(s); s \in D\}$ است، که در آن مجموعه اندیس‌گذار D زیر مجموعه‌ای از فضای اقلیدسی d بعدی، $d \geq ۱$، از R^d است.

برای میدان تصادفی $Z(\cdot)$، میانگین در موقعیت s و کوواریانس در موقعیت‌های $s_۱$ و $s_۲$ به‌ترتیب به صورت

$$\begin{aligned} \mu(s) &= E[Z(s)],\ s \in D \\ C(s_۱, s_۲) &= Cov[Z(s_۱), Z(s_۲)],\ s_۱, s_۲ \in D \end{aligned}$$

تعریف می‌شوند. برای $s_۱ = s_۲ = s$، واریانس میدان تصادفی $Z(\cdot)$ در مکان s به صورت

$$Var[Z(s)] = E[Z(s) - \mu(s)]^۲$$

حاصل می‌شود.

واریانس و کوواریانس در آمار کلاسیک به ترتیب میزان پراکندگی و شباهت یا همبستگی داده‌ها را نشان می‌دهند. در آمار فضایی نیز مفاهیمی تعریف شده‌اند که مشابه واریانس و کوواریانس بیانگر ساختار همبستگی فضایی داده‌ها باشند.

[1] Variogram

جدول ۱: آمار توصیفی نظرات مشتریان

	تعداد نمونه	میانگین	انحراف معیار	مینیمم	ماکسیمم
جهت‌گیری مشتری	۲۲۰	۲۲/۰	۰۹/۰	ـ۳۷/۰	۵/۰
رضایت مشتری	۲۲۰	۶۱/۳	۷۵/۰	۱	۵

واریانس تفاضل میدان تصادفی در دو موقعیت s و $s+h$ را تغییرنگار[1] می‌نامند که به‌صورت

$$۲\gamma(h) = Var(Z(s+h) - Z(s))$$

تعریف می‌شود. تابع $\gamma(h)$ نیز نیم‌تغییرنگار[2] نامیده می‌شود. کوواریانس $Z(s)$ و $Z(s+h)$، یعنی

$$C(h) = Cov(Z(s), Z(s+h))$$

هم‌تغییرنگار[3] است و با شرط $C(۰) > ۰$ عبارت

$$\rho(s, s+h) = \rho(h) = \frac{C(h)}{C(۰)}$$

همبستگی‌نگار[4] نامیده می‌شود. برای تعیین ساختار همبستگی فضایی داده‌ها می‌توان از تغییرنگار، هم‌تغییرنگار یا همبستگی‌نگار استفاده کرد. برآورد تجربی یا کلاسیک هم‌تغییرنگار به‌صورت

$$\hat{C}(h) = \frac{۱}{N_h} \sum_{N(h)} (Z(s_i) - \bar{Z})(Z(s_j) - \bar{Z})$$

تعریف می‌شود، که در آن N_h تعداد زوج نقاطی است که فاصله بین آنها برابر h است.

در داده‌های مربوط به میزان رضایت‌مندی مشتریان نیز در برخی مواقع همبستگی فضایی وجود دارد. در مطالعه حاضر از تجزیه و تحلیل احساسات مشتری (NLP) که روشی بر اساس یادگیری ماشین است استفاده می‌شود به این صورت که بر اساس نظر مشتریان و همچنین جهت‌گیری مشتری که به عنوان یک متغیر کمی تعریف می‌شود عددی پیوسته در بازه ۱− تا ۱ نسبت داده می‌شود که در آن نقطه صفر وضعیت خنثی را نشان می‌دهد. در عین حال، جهت‌گیری مثبت نشان‌دهنده احساس مثبت و جهت‌گیری منفی احساس منفی مشتری را بیان می‌کند. علاوه بر این، میزان رضایت‌مندی به صورت یک مقیاس لیکرت پنج نقطه‌ای نیز اندازه‌گیری می‌شود، که در آن عدد یک نشان‌دهنده کمترین میزان رضایت و عدد پنج به معنای بیشترین میزان رضایت است. آمارهای توصیفی مربوط به این دو متغیر اندازه‌گیری شده در جدول ۱ نشان داده شده است. گستره متغیر جهت‌گیری این مطالعه از ۰/۳۷− تا ۰/۵ با مقدار میانگین ۰/۲۲ است؛ در حالی که نمره عددی رضایت در مقیاس لیکرت از ۱ تا ۵ تغییر می‌کند که مقدار میانگین ۳/۶۱ است. اطلاعات بدست آمده نشان‌دهنده ارتباط بسیار بالایی بین جهت‌گیری ضمنی و رتبه عددی میزان رضایت اندازه‌گیری شده است که توسط ضریب همبستگی گشتاوری پیرسون محاسبه شده و مقدار آن $R = ۰/۸۶$ است.

در حالت استقلال مشاهدات استفاده از روش حداقل مربعات به عنوان یک روش برای برآورد پارمترهای رگرسیونی شناخته می‌شود. در این حالت به منظور حفظ ویژگی و اعتبار نتایج آماری مربوط به جمعیت تشکیل‌دهنده، لازم است فرضیات زیر در نظر گرفته شود.

[2] Semi variogram
[3] Covariogram
[4] Correlogram

۱ـ میانگین خطا صفر است.
۲ـ خطاها غیرهمبسته با واریانس ثابت هستند.
۳ـ خطا دارای توزیع نرمال است.
با این حال، در بسیاری از مواقع در مطالعات بازاریابی، مقدار مشاهده شده در یک مکان اغلب به مقادیر در مکان‌های همسایه وابسته است. یعنی میزان رضایت‌مندی مشتریان از خدمات هتل‌ها تا حد زیادی وابسته به محلی است که هتل در آن قرار دارد و عموما به خاطر مسایل اقتصادی و اجتماعی هتل‌هایی که نزدیک به هم در یک محل خاص قرار می‌گیرند کیفیت خدماتشان بسیار به هم شباهت دارد. لذا در این گونه بررسی‌های رضایت‌سنجی، داده‌ها مستقل نبوده و دارای همبستگی فضایی خواهند بود در ادامه با معرفی ماتریس‌های مختلف نحوه در نظر گرفتن همسایگی ناحیه‌ها برای لحاظ کردن همبستگی فضایی مورد بررسی قرار می‌گیرد.

نحوه ورود همبستگی فضایی در هنگام تحلیل داده‌های ناحیه‌ای از طریق ماتریس همسایگی یا مجاورت[۵] در مدل منظور می‌شود. اما برای وارد کردن اثر فضایی در مدل، ماتریس‌های همسایگی مستقیماً استفاده نمی‌شوند بلکه ابتدا ماتریس وزن فضایی براساس ماتریس همسایگی ساخته شده، سپس از ماتریس وزن برای ورود اثرات فضایی به مدل استفاده می‌شود. در ادامه انواع ماتریس‌های وزن فضایی معرفی خواهد شد. به‌طور کلی ماتریس وزن براساس ماتریس همسایگی ساخته می‌شود که میزان نزدیکی یا به عبارت دیگر تأثیرگذاری دو موقعیت یا ناحیه بر یکدیگر را نشان می‌دهد. با استاندارد کردن سطری ماتریس همسایگی، ماتریس وزن حاصل می‌شود به‌طوری که جمع هر سطر ماتریس وزن برابر یک است. در ادامه نمونه‌هایی از ماتریس همسایگی معرفی می‌شوند.

ماتریس وزن مرز مشترک: برای ساخت این ماتریس ابتدا مؤلفه‌های قطر اصلی ماتریس برابر صفر در نظر گرفته می‌شوند. سپس برای نواحی‌ای که مرز مشترک دارند عدد یک و نواحی‌ای که مرز مشترک ندارند عدد صفر منظور می‌شود. آنگاه تمامی مؤلفه‌های هر سطر بر مجموع مؤلفه‌های همان سطر تقسیم می‌شوند تا ماتریس وزن حاصل شود. به‌عبارت دیگر عضو ijام ماتریس همسایگی به‌صورت

$$C_{ij} = \begin{cases} ۱, & j \in N_۱(i) \\ ۰, & j \notin N_۱(i) \end{cases} \qquad (۱.۲)$$

تعریف می‌شود، که در آن $N_۱(i)$ مجموعه نواحی مجاور ناحیه iام است و $j \in N_۱(i)$ به معنی مجاور بودن ناحیه j با ناحیه i است. به‌طور مثال برای نواحی یک مشبکه ۳ در ۳ به‌صورت ماتریس‌های همسایگی C و وزن W به ترتیب عبارتند

۳	۲	۱
۶	۵	۴
۹	۸	۷

[5] Adjacency matrix

از

$$C=\begin{pmatrix} ۰ & ۱ & ۰ & ۱ & ۰ & ۰ & ۰ & ۰ & ۰ \\ ۱ & ۰ & ۱ & ۰ & ۱ & ۰ & ۰ & ۰ & ۰ \\ ۰ & ۱ & ۰ & ۰ & ۰ & ۱ & ۰ & ۰ & ۰ \\ ۱ & ۰ & ۰ & ۰ & ۱ & ۰ & ۱ & ۰ & ۰ \\ ۰ & ۱ & ۰ & ۱ & ۰ & ۱ & ۰ & ۱ & ۰ \\ ۰ & ۰ & ۱ & ۰ & ۱ & ۰ & ۰ & ۰ & ۱ \\ ۰ & ۰ & ۰ & ۱ & ۰ & ۰ & ۰ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۰ & ۱ & ۰ & ۱ & ۰ & ۱ \\ ۰ & ۰ & ۰ & ۰ & ۰ & ۱ & ۰ & ۱ & ۰ \end{pmatrix} \quad W=\begin{pmatrix} ۰ & \frac{۱}{۲} & ۰ & \frac{۱}{۲} & ۰ & ۰ & ۰ & ۰ & ۰ \\ \frac{۱}{۳} & ۰ & \frac{۱}{۳} & ۰ & \frac{۱}{۳} & ۰ & ۰ & ۰ & ۰ \\ ۰ & \frac{۱}{۲} & ۰ & ۰ & ۰ & \frac{۱}{۲} & ۰ & ۰ & ۰ \\ \frac{۱}{۳} & ۰ & ۰ & ۰ & \frac{۱}{۳} & ۰ & \frac{۱}{۳} & ۰ & ۰ \\ ۰ & \frac{۱}{۴} & ۰ & \frac{۱}{۴} & ۰ & \frac{۱}{۴} & ۰ & \frac{۱}{۴} & ۰ \\ ۰ & ۰ & \frac{۱}{۳} & ۰ & \frac{۱}{۳} & ۰ & ۰ & ۰ & \frac{۱}{۳} \\ ۰ & ۰ & ۰ & \frac{۱}{۲} & ۰ & ۰ & ۰ & \frac{۱}{۲} & ۰ \\ ۰ & ۰ & ۰ & ۰ & \frac{۱}{۳} & ۰ & \frac{۱}{۳} & ۰ & \frac{۱}{۳} \\ ۰ & ۰ & ۰ & ۰ & ۰ & \frac{۱}{۲} & ۰ & \frac{۱}{۲} & ۰ \end{pmatrix}$$

اگر علاوه بر همسایگی‌های با مرز مشترک، همسایگی‌های با k گام فاصله از هر ناحیه نیز به‌عنوان همسایه ناحیه iام درنظر گرفته شوند، عناصر ماتریس همسایگی k ناحیه نزدیک به صورت

$$C_{ij}=\begin{cases} ۱, & j\in N_k(i) \\ ۰, & j\notin N_k(i) \end{cases}$$

تعریف می‌شوند، که در آن $N_k(i)$ نشان‌دهنده k ناحیه مجاور ناحیه iام است.

همچنین می‌توان ماتریس همسایگی را براساس همبستگی میان نقاط به صورت

$$C_{ij}=r_{ij}^{-b}$$

بیان کرد، که در آن b ضریب اصطکاک فاصله[۶] و r_{ij} همبستگی میان دو ناحیه i و j است. ضریب اصطکاک b معمولاً برابر ۱ در نظر گرفته می‌شود. کلیف و ارد از این ساختار برای ساخت ماتریس وزن فضایی استفاده کردند، به همین علت ماتریس وزن حاصل از این ماتریس همسایگی، کلیف و ارد[۷] نامیده می‌شود. یکی دیگر از ماتریس‌های همسایگی برحسب همبستگی میان نواحی به صورت

$$C_{ij}=\begin{cases} ۱, & r_{ij}\le r_{۰} \\ ۰, & r_{ij}>r_{۰} \end{cases}$$

است، که در آن $r_{۰}$ یک آستانه برای میزان همبستگی است. ماتریس وزن براساس همبستگی شبه عمومی[۸] نیز به صورت

$$C_{ij}=\exp(-\frac{r_{ij}}{\bar{r}})$$

تعریف می‌شود. یکی دیگر از راه‌های شناسایی میزان همبستگی فضایی موقعیت‌ها، فاصله میان موقعیت‌ها است. ماتریس همسایگی بر اساس حد فاصل میان نواحی به صورت

$$C_{ij}=\begin{cases} ۱, & ۰\le d_{ij}\le d \\ ۰, & i=j \text{ یا } d_{ij}>d \end{cases}$$

[6] Distance friction coefficient
[7] Cliff-Ord
[8] Quasi-global correlation

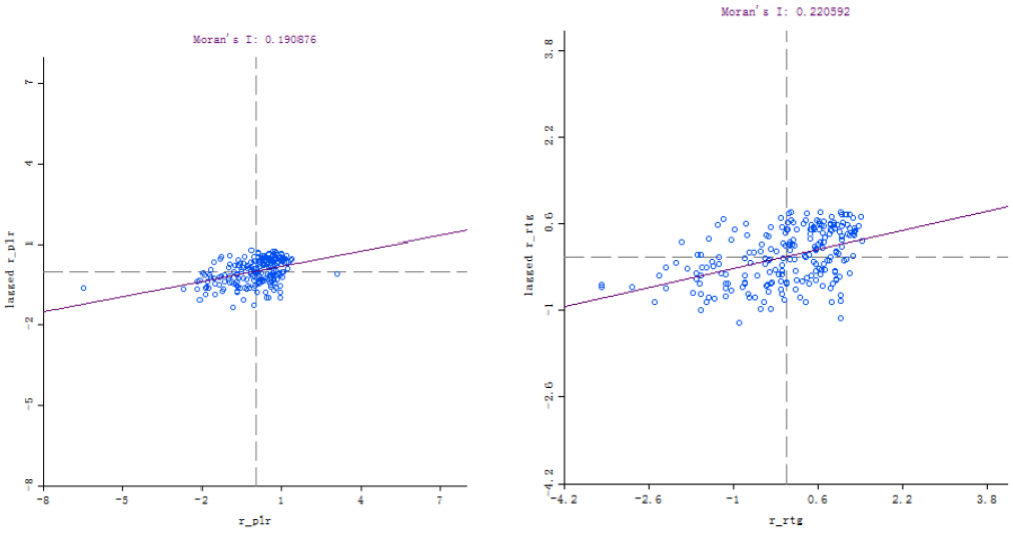
تعریف می‌شود، که در آن d_{ij} فاصله بین دو ناحیه i و j است و d آستانه است. چند نمونه از ماتریس‌های وزن که براساس فاصله بین نواحی حاصل می‌شوند به صورت

$$C_{ij} = \begin{cases} ۰, & i = j \\ d_{ij}^{-\alpha}, & i \neq j \end{cases} \quad (۲.۲)$$

$$C_{ij} = \begin{cases} ۰, & i = j \\ \exp(-\beta d_{ij}^{\alpha}), & i \neq j \end{cases} \quad (۳.۲)$$

$$C_{ij} = \begin{cases} [۱ - (\frac{d_{ij}}{d})^{k}]^{k}, & ۰ \leq d_{ij} \leq d \\ ۰, & i = j \text{ یا } d_{ij} > d \end{cases} \quad (۴.۲)$$

هستند، که در آن‌ها α، β و k اعدادی مثبت هستند. برای ماتریس همسایگی (۲.۲) مقدار α معمولاً برابر ۱ یا ۲ در نظر گرفته می‌شود.



شکل ۲: نتایج آزمون موران

پیش از تصمیم‌گیری در مورد وابستگی فضایی، لازم است وجود وابستگی فضایی را تایید کنیم، با توجه به آنکه مسئله مهم برای تحلیل رضایت مشتری تشخیص وجود خود همبستگی فضایی[9] است. برای این منظور معمولاً از آزمون موران[10] استفاده می‌شود. فرض صفر این آزمون عدم وجود خود همبستگی فضایی در داده‌ها است. آماره این آزمون براساس وزن‌های بین نواحی به صورت

$$I = \frac{n \sum_{i=۱}^{n} \sum_{j=۱}^{n} w_{ij}(t_i - \bar{t})(t_j - \bar{t})}{(\sum_{i=۱}^{n} \sum_{j=۱}^{n} w_{ij})(\sum_{i=۱}^{n}(t_i - \bar{t})^{۲})}$$

است، که در آن n تعداد داده‌ها، w_{ij} وزن میان دو ناحیه i و j و t_i مقدار ویژگی مورد بررسی در ناحیه i است. وقتی

[9] Spatial autocorrelation
[10] Moran's I test

$|I^*| > z_{۱-\alpha/۲}$، فرض وجود خود همبستگی فضایی داده‌ها پذیرفته می‌شود، که در آن

$$I^* = \frac{I - E(I)}{\sqrt{Var(I)}}, \quad P(Z \leq ۱ - \alpha/۲) = z_{۱-\alpha/۲}$$

که در آن Z متغیر تصادفی با توزیع نرمال استاندارد است. با توجه به فرض تأثیرگذاری موقعیت قرارگیری هتل‌ها بر رضایت مشتریان از آزمون موران انجام شد. با درنظر گرفتن ماتریس مجاورت (۱.۲) فرض صفر آزمون موران در سطح معنی‌داری ۰۵/۰ با pـ مقدار کمتر از ۰/۰۰۱ رد می‌شود و فرضیه وجود خود همبستگی فضایی در داده‌ها تایید می‌گردد.

بحث و نتیجه‌گیری

نتایج موران برای هر دو متغیر مورد بررسی یعنی جهت‌گیری مشتری و رضایت‌مندی مشتری در شکل ۲ رسم شده است. یک همبستگی فضایی مثبت برای رضایت مشتریان هتل‌ها در منطقه مورد مطالعه وجود دارد. بر اساس آماره موران، هتل‌هایی که جهت‌گیری مشتری بالایی دارند، از نظر فضایی در گروه هتل‌های با رضایت بالای مشتری قرار می‌گیرند. بنابراین، رویکرد مکان‌یابی هتل بسیار مهم است. مانند هر شرکت دیگری هتل‌ها نیز برای به حداکثر رساندن میزان سود خالص خود در رقابت با سایرین با استفاده از تحلیل‌های مکان‌یابی و اقدامات پیش‌بینی شده رقبا مکان خود را تعیین می‌کنند. نکته مهم، چانگ (۲۰۰۷) از این یافته‌ها هنگام استفاده از رضایت مشتری به عنوان متغیر وابسته برای تجزیه و تحلیل تجربی پیش‌آمدها و پیامدها، نشان می‌دهد که وجود وابستگی فضایی به طور بالقوه فرض نااریب بودن برآورد پارامترهای مدل رگرسیونی را نقض می‌کند. لذا برای دستیابی به نتایج مطلوب‌تر لازم است با وارد کردن همبستگی فضایی و استفاده از رویکرد فضایی تحلیل بهتری از میزان رضایت‌مندی مشتریان از خدمات هتل‌ها صورت گیرد. رضایت مشتری توسط محققان به طور گسترده مطالعه شده و مورد استفاده شرکت‌ها قرار گرفته است. گسترش محتویات تولید شده توسط کاربران و اطلاعات موقعیت جغرافیایی به مشتریان و شرکت‌ها در راستای رسیدن به کیفیت مطلوب‌تر کمک می‌کند. در این مقاله، با توجه به اندازه‌گیری آماره موران، نشان دادیم که هتل‌هایی با رتبه بالا در یک دسته قرار می‌گیرند و مکان‌یابی برای آنها اهمیت دارد. در مطالعات آتی قصد داریم با تعیین یک ماتریس همسایگی مناسب و استفاده از مدل‌های اتورگرسیون شرطی به پیش‌گویی میزان رضایت‌مندی مشتریان در یک مکان خاص که قرار است هتلی در آنجا احداث شود بپردازیم. و در نهایت با استفاده از این روش بهترین نقطه مکانی برای احداث هتل به سرمایه‌گذاران این صنعت معرفی شود.

تقدیر و تشکر

از استادان ارجمندم جناب آقای دکتر محسن محمدزاده،آقاي دکترموسي گلعلي‌زاده،آقاي دکتر کيومرث مترجم بسيار سپاسگزارم چرا که بدون راهنمایی‌ها وآموزش‌هاي ايشان تهيه اين مقاله بسيار مشکل می‌نمود.

مراجع

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

Alcácer, J., Chung, W. (2007), *Location Strategies and Knowledge Spillovers*, Management Science **53**, 760-776.

Bradlow(2005), *Spatial models in marketing*, Marketing Letters **16**, 267-278.

Cliff, A. D. and Ord, J. K. (1973), *Spatial Autocorrelation*, Pion Limited, London.

Moran, P. A. P. (1950), *Notes on Continuous Stochastic Phenomena*, **37**, 17-23.

Wang(2015), *Crowdsourcing the Landscape of Cannabis (Marijuana) of the Contiguous United States*, Environment and Planning A 10.1177/0308518x15598541.

Zhang(2015), *Social Learning in Networks of Friends versus Strangers*, Marketing Science, **34**, 573-589.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

تحلیل کواریانس گاوسی وارون برای داده‌های فضایی

زهرا زارع‌پور یکنمی، افشین فلاح[1]
گروه آمار، دانشگاه بین‌المللی امام خمینی

چکیده: نرمال بودن توزیع متغیر پاسخ و مستقل بودن مشاهدات در هر تیمار، از جمله فرض‌های پایه‌ای تحلیل کواریانس به‌شیوه‌ی سنتی هستند، به‌گونه‌ای که برقرار نبودن این فرض‌ها اعتبار نتایج حاصل از این روش را مخدوش می‌کند. در این مقاله برای تحلیل کواریانس در جوامعی که مشاهدات مثبت و دارای وابستگی فضایی هستند، مدلی پیشنهاد شده است. چگونگی برازش مدل پیشنهادی از دیدگاه بیزی مورد بحث قرار گرفته و از یک مطالعه‌ی شبیه‌سازی برای ارزیابی آن استفاده شده است.

واژه‌های کلیدی: تحلیل کواریانس، گاوسی وارون، داده‌های فضایی، رهیافت بیزی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J10، 62H11.

۱ مقدمه

رگرسیون و تحلیل واریانس از جمله مهم‌ترین روش‌های مرسوم برای بیان روابط بین متغیرها هستند. اما در برخی کاربردها هیچ یک از این روش‌های مدل‌سازی به‌تنهایی نمی‌تواند پاسخ مناسب به پرسش‌های پژوهش‌گر را فراهم آورند. تحلیل کواریانس ابزار نیرومندی است که از ترکیب دو روش رگرسیون و تحلیل واریانس ساخته می‌شود و از مزایای هر دو روش استفاده می‌کند. تحلیل کواریانس نخستین بار توسط فیشر (۱۹۳۲) مطرح شده است. کاکس و مک کالک (۱۹۸۲) برخی از جنبه‌های تحلیل کواریانس در زیست‌سنجی را مورد بررسی قرار دادند. روش‌های بیزی و غیربیزی در رابطه با مدل‌های تحلیل کواریانس تحت شرایط ناهم‌واریانسی، توسط آناندا (۱۹۹۸) مورد مطالعه قرار گرفته است. میلر و چپمن (۲۰۰۱) پذیره‌های زیربنایی تحلیل کواریانس را مورد بحث قرار داده‌اند. از جمله مهم‌ترین پذیره‌های زیربنایی تحلیل کواریانس مرسوم نرمال بودن توزیع متغیر پاسخ و مستقل بودن مشاهدات مربوط به تیمارهای گوناگون است. این در حالی است که در بسیاری از کاربردها، این پذیره‌ها برقرار نیستند. در این راستا، مشکانی و همکاران (۲۰۱۴) تحلیل کواریانس گاوسی وارون را برای استفاده در شرایطی که مشاهدات پاسخ نامنفی و چوله به راست هستند، توسعه داده‌اند. فلاح (۲۰۱۶) ، تحلیل

[1] نام و ایمیل ارائه دهنده مقاله: افشین فلاح، afshin.ikiu@gmail.com

کواریانس گاوسی آمیخته‌ی متناهی را برای شرایطی که جامعه تحت مطالعه ناهمگن بوده و از زیرجامعه‌هایی با ویژگی‌های متفاوت تشکیل شده، مورد بررسی قرار داده است. از طرفی در برخی از کاربردها مشاهدات مستقل نیستند و به‌دلیل قرار گرفتن در موقعیت‌های فضایی خاص خود دارای وابستگی فضایی هستند. در این شرایط استفاده از نظریه‌ی مرسوم تحلیل کواریانس به‌دلیل لحاظ نکردن وابستگی‌های فضایی بین مشاهدات به نتایج گمراه کننده می‌انجامد.

در این مقاله، تحلیل کواریانس گاوسی وارون برای داده‌های فضایی از دیدگاه بیزی مورد بررسی قرار گرفته و مدلی برای آن پیشنهاد شده است. در بخش ۲، چارچوب نظری مدل پیشنهادی شرح داده شده است. در بخش ۳، نحوه برازش مدل از طریق رهیافت بیزی مورد بحث قرار گرفته است. در بخش ۴، برای ارزیابی مدل پیشنهادی از یک مطالعه شبیه‌سازی استفاده شده است.

۲ مدل پیشنهادی

به پیروی از مشکانی و همکاران (۲۰۱۴)، مدل تحلیل کواریانس را می‌توان به‌صورت

$$\boldsymbol{\mu}^* = (\mu_{11}^{-1}, \ldots, \mu_{tr}^{-1}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} = \boldsymbol{W}\boldsymbol{\theta},$$

در نظر گرفت، که در آن $\boldsymbol{\mu}$ بردار میانگین مشاهدات پاسخ، $\boldsymbol{X}$ ماتریس طرح، $\boldsymbol{Z}$ ماتریس متغیرهای کمکی، $\boldsymbol{\beta}$ بردار ضرایب تیماری، γ بردار ضرایب رگرسیونی، $\boldsymbol{W} = [\boldsymbol{X}|\boldsymbol{Z}]$ ماتریس مشاهدات و $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ بردار ضرایب مدل را نشان می‌دهند. فرض کنید متغیرهای پاسخ در مدل تحلیل کواریانس دارای توزیع گاوسی وارون به‌صورت

$$y_{ij}|\boldsymbol{X}, \boldsymbol{Z} \sim f_{IG}(y_{ij}|\mu_{ij}, \lambda), \quad i = 1, \ldots, t; j = 1, \ldots, r,$$

هستند، که در آن

$$f_{IG}(y_{ij} \mid \mu_{ij}, \lambda) = \sqrt{\frac{\lambda}{2\pi y_{ij}^3}} \exp\left\{-\frac{\lambda(y_{ij} - \mu_{ij})^2}{2y_{ij}\mu_{ij}^2}\right\}, \quad y_{ij}, \mu_{ij}, \lambda > 0,$$

تابع چگالی توزیع گاوسی وارون، μ_{ij} میانگین متغیر پاسخ برای مشاهده مربوط به تکرار jـام از تیمار iـام و λ پارامتر مقیاس را نشان می‌دهند. در مدل پیشنهادی اثر وابستگی‌های فضایی موجود بین مشاهدات پاسخ در هر تیمار، به‌وسیله‌ی بردار اثرهای تصادفی فضایی $\boldsymbol{\phi} = (\phi_{11}, \ldots, \phi_{tr})'$ در تابع پیوند مدل، لحاظ می‌شوند. فرض کنید بردار مشاهدات پاسخ در موقعیت‌های مکانی $(s_{11}, \ldots, s_{tr})$ به‌صورت $\{y_{ij} = y(s_{ij}); D \subset \mathbb{R}^d, d \geq 2\}$ تعریف می‌شوند. در این صورت، تابع پیوند مدل تحلیل کواریانس گاوسی وارون فضایی به‌صورت

$$\mu_{ij}^{-1} = \boldsymbol{w}_{ij}\boldsymbol{\theta} + \phi_{ij} \tag{۱.۲}$$

خواهد بود، که در آن $\boldsymbol{w}_{ij} = (\boldsymbol{x}_i, z_{ij})$ سطر ijـام ماتریس $\boldsymbol{W}$ و ϕ_{ij} اثر تصادفی فضایی متناظر با مشاهده مربوط به تکرار jـام از تیمار iـام است. اثر تصادفی فضایی ϕ_{ij} را می‌توان به‌صورت شرطی بر اثرهای تصادفی فضایی مشاهدات همسایه در تیمار iـام به‌صورت

$$\phi_{ij}|\boldsymbol{\phi}_{i(-j)}, \sigma^2 \sim N\left(\xi \sum_{l \neq j} \frac{a_{ijl}}{a_{ij+}}\phi_{il}, \frac{\sigma^2}{a_{ij+}}\right), \tag{۲.۲}$$

در نظر گرفت، که در آن $\boldsymbol{\phi}_{i(-j)} = \{\phi_{il}; l \neq j\}$، $\sigma^۲$ واریانس مشترک اثرهای تصادفی فضایی، a_{ijl} درایه (j,l) از ماتریس همسایگی است که به‌صورت

$$a_{ijl} = \begin{cases} ۱, & \text{اگر مشاهده‌ی } i \text{ در تکرارهای } j \text{ و } l \text{ همسایه باشند.} \\ ۰, & j = l \end{cases} \tag{۳.۲}$$

تعریف می‌شود، $a_{ijl} = \sum_{l \neq j} a_{ijl}$ و ξ پارامتر هموارسازی فضایی است. پارامتر هموارسازی در بازه‌ی $\xi \in \left(\frac{۱}{\lambda_{(۱)}}, \frac{۱}{\lambda_{(n)}}\right)$ تغییر می‌کند که درآن $\lambda_{(۱)} \geq \ldots \geq \lambda_{(n)}$ مقادیر ویژه ماتریس $\boldsymbol{M} = (m_{ijl}) = \left(\frac{a_{ijl}}{a_{ij+}}\right)$ هستند، $\boldsymbol{D}_M^{-\frac{۱}{۲}} \boldsymbol{M} \boldsymbol{D}_M^{-\frac{۱}{۲}}$ ماتریس وزن و $\boldsymbol{D}_M$ ماتریس قطری با درایه‌های قطری $\{m_{۱۱+}, \ldots, m_{tr+}\}$ است (کاووسی و همکاران، ۲۰۰۹). با توجه به اینکه همبستگی‌های فضایی تنها بین مشاهدات متناظر با تکرارهای گوناگون در هر تیمار وجود دارند، ماتریس همسایگی در مدل پیشنهادی، ماتریسی بلوکی و به‌صورت

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_۱ & \ldots & ۰ \\ \vdots & \ddots & \vdots \\ ۰ & \ldots & \boldsymbol{A}_t \end{bmatrix}, \tag{۴.۲}$$

در نظر گرفته می‌شود، که در آن ماتریس $\boldsymbol{A}_i$ مشاهدات همسایه تیمار iـام، به‌ازای $i = ۱, \ldots, t$، را نشان می‌دهند.

۳ براورد پارامترهای مدل

توابع درستنمایی و تابع لگ‌درستنمایی متناظر با مدل تحلیل کواریانس پیشنهادی را می‌توان به‌ترتیب به‌صورت

$$\begin{aligned} L(\boldsymbol{\Psi}|\boldsymbol{y}, \boldsymbol{W}) &= \prod_{i=۱}^{t} \prod_{j=۱}^{r} \left(f_{IG}(y_{ij}|\mu_{ij}, \lambda)\right., \\ \ell(\boldsymbol{\Psi}|\boldsymbol{y}, \boldsymbol{W}) &= \sum_{i=۱}^{t} \sum_{j=۱}^{r} \left(f_{IG}(y_{ij}|\mu_{ij}, \lambda)\right., \end{aligned}$$

نوشت. این توابع را می‌توان با استفاده از نمادهای برداری و ماتریسی، به‌ترتیب به‌صورت

$$\begin{aligned} \ell(\boldsymbol{\Psi}|\boldsymbol{y}, \boldsymbol{W}) &= \sum_{i=۱}^{t} \sum_{j=۱}^{r} \log(\lambda) - \sum_{i=۱}^{t} \sum_{j=۱}^{r} \log(۲\pi y_{ij}^{۳}) - \frac{\lambda}{۲} \left(Q_۱(\boldsymbol{\theta}_۱) + Q_۵(\boldsymbol{\theta}, \boldsymbol{\phi})\right), \\ L(\boldsymbol{\Psi}|\boldsymbol{y}, \boldsymbol{W}) &= \prod_{i=۱}^{t} \prod_{j=۱}^{r} \left(\frac{\lambda}{۲\pi y_{ij}^{۳}}\right) \exp\left\{\frac{\lambda}{۲} \left(Q_۱(\boldsymbol{\theta}) + Q_۴(\boldsymbol{\theta}, \boldsymbol{\phi})\right)\right\}, \end{aligned} \tag{۱.۳}$$

بازنویسی نمود، که در آن

$$\begin{aligned} Q_۱(\boldsymbol{\theta}) &= \boldsymbol{\theta}' \left(\sum_{i=۱}^{t} \sum_{j=۱}^{r} y_{ij} \boldsymbol{w}_{ij}' \boldsymbol{w}_{ij}\right) \boldsymbol{\theta}_k - ۲ \left(\sum_{i=۱}^{t} \sum_{j=۱}^{r} \boldsymbol{w}_{ij}\right) \boldsymbol{\theta} + \sum_{i=۱}^{t} \sum_{j=۱}^{r} y_{ij}^{-۱} \\ &= (\boldsymbol{Y}\boldsymbol{W}\boldsymbol{\theta} - ۱_n)' \boldsymbol{Y}^{-۱} (\boldsymbol{Y}\boldsymbol{W}\boldsymbol{\theta} - ۱_n), \\ Q_۴(\boldsymbol{\theta}, \boldsymbol{\phi}) &= \sum_{i=۱}^{t} \sum_{j=۱}^{r} y_{ij} \phi_{ij}^{۲} + ۲ \sum_{i=۱}^{t} \sum_{j=۱}^{r} y_{ij} \phi_{ij} \boldsymbol{w}_{ij} \boldsymbol{\theta} - ۲ \sum_{i=۱}^{t} \sum_{j=۱}^{r} \phi_{ij} \\ &= \boldsymbol{\phi}' \boldsymbol{Y} \boldsymbol{\phi} + ۲ \boldsymbol{\phi}' \boldsymbol{Y} \boldsymbol{W} \boldsymbol{\theta} - ۲ ۱_n' \boldsymbol{\phi}. \end{aligned}$$

در رویکرد بیزی لازم است ابتدا توزیع‌های پیشین مناسبی برای پارامترهای مدل در نظر گرفته شود. برای این منظور، با فرض استقلال پیشینی پارامترهای $(\boldsymbol{\theta}, \lambda)$ و $(\boldsymbol{\phi}, \sigma^2)$، توزیع پیشین توأم پارامترها را می‌توان به‌صورت

$$\begin{aligned} \pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2, \lambda) &= \pi(\boldsymbol{\theta}, \lambda)\pi(\boldsymbol{\phi}, \sigma^2) \\ &= \pi(\boldsymbol{\theta}|\lambda)\pi(\lambda)\pi(\boldsymbol{\phi}|\sigma^2)\pi(\sigma^2), \end{aligned} \tag{۲.۳}$$

نوشت. برای پارامترهای $\boldsymbol{\theta}$ و λ به پیروی از مشکانی و همکاران (۲۰۱۶) توزیع‌های پیشین مزدوج به‌شکل

$$\begin{aligned} \boldsymbol{\theta} \mid \lambda &\sim N_p(\boldsymbol{\eta}_0, \lambda^{-1}\boldsymbol{\Delta}_0) \quad \boldsymbol{\theta}, \boldsymbol{\eta}_0 \in \mathbb{R}^p, \\ \lambda &\sim G(\frac{a_0}{2}, \frac{b_0}{2}) \quad \lambda, a_0, b_0 > 0, \end{aligned} \tag{۳.۳}$$

در نظر گرفته شده‌اند، که در آن‌ها $\boldsymbol{\Delta}_0$ یک ماتریس قطری با درایه‌های قطری مثبت $\boldsymbol{\delta}_0 = (\delta_{01}, \ldots, \delta_{0p})$ است، $p = t + q$ و $\boldsymbol{\eta}_0 = (\eta_{01}, \ldots, \eta_{0p})'$. از این‌رو، توزیع پیشین توأم $(\boldsymbol{\theta}, \lambda)$ به‌صورت

$$\begin{aligned} \pi(\boldsymbol{\theta}, \lambda) &\propto \pi(\boldsymbol{\theta} \mid \lambda)\pi(\lambda) \\ &\propto \lambda^{\frac{(a_0+p)}{2}-1} \exp\left\{-\frac{\lambda}{2}\left(b_0 + (\boldsymbol{\theta} - \boldsymbol{\eta}_0)'\boldsymbol{\Delta}_{0k}^{-1}(\boldsymbol{\theta} - \boldsymbol{\eta}_0)\right)\right\}, \end{aligned} \tag{۴.۳}$$

محاسبه می‌شود و یک توزیع گاما ـ نرمال است. برای بردار اثرهای تصادفی $\boldsymbol{\phi} = (\phi_{11}, \ldots, \phi_{tr})'$ و پارامتر واریانس σ^2، به‌ترتیب توزیع‌های پیشین نرمال چندمتغیره و گامای وارون به‌صورت

$$\begin{aligned} \boldsymbol{\phi}|\sigma^2 &\sim N_n\left(0, (\boldsymbol{D_M} - \xi\boldsymbol{M})^{-1}\sigma^2\right), \\ \frac{1}{\sigma^2} &\sim G(f_0, g_0), \end{aligned} \tag{۵.۳}$$

لحاظ می‌شوند. چگالی‌های متناظر با این توزیع‌های پیشین را می‌توان به‌شکل متناسب

$$\begin{aligned} \pi(\boldsymbol{\phi}|\sigma^2) &\propto \exp\left\{\frac{-1}{2\sigma^2}\boldsymbol{\phi}'(\boldsymbol{D_M} - \xi\boldsymbol{M})^{-1}\boldsymbol{\phi}\right\}, \\ \pi(\sigma^2) &\propto \sigma^{2-(f_0+1)} \exp\left\{\frac{-g_0}{\sigma^2}\right\}, \end{aligned} \tag{۶.۳}$$

نوشت. در رابطه‌های (۳.۳) و (۵.۳) از اندیس صفر برای نشان دادن ابرپارامترهای مدل بیزی که می‌بایست توسط تحلیل‌گر تعیین شوند، استفاده شده است. با توجه به توزیع‌های پیشین (۲.۳)، (۴.۳) و (۵.۳)، تابع درستنمایی (۱.۳)، توزیع پسین توأم پارامترها به‌صورت

$$\begin{aligned} \pi(\boldsymbol{\Psi}|\boldsymbol{y}, \boldsymbol{W}) &= L(\boldsymbol{\Psi}|\boldsymbol{y}, \boldsymbol{W}) \times \pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2, \lambda) \\ &\propto \prod_{i=1}^{t}\prod_{j=1}^{r} \lambda^{\frac{(a_0+p+n)}{2}-1} \exp\left\{-\frac{\lambda}{2}\left(Q_2(\boldsymbol{\theta}) + Q_3(b, \boldsymbol{\eta}_0, \boldsymbol{\Delta}_0) + Q_4(\boldsymbol{\theta}, \boldsymbol{\phi})\right)\right\} \\ &\times \sigma^{2-(f_0+1)} \exp\left\{-\frac{1}{\sigma^2}Q_5(\boldsymbol{\phi})\right\}, \end{aligned} \tag{۷.۳}$$

به‌دست می‌آید، که در آن

$$\begin{aligned} Q_2(\boldsymbol{\theta}) &= (\boldsymbol{\theta} - \boldsymbol{\Sigma}^{-1}\boldsymbol{d})'\boldsymbol{\Sigma}(\boldsymbol{\theta} - \boldsymbol{\Sigma}^{-1}\boldsymbol{d}), \\ Q_3(b, \boldsymbol{\eta}_0, \boldsymbol{\Delta}_0) &= b_0 + \boldsymbol{\eta}_0'\boldsymbol{\Delta}_0^{-1}\boldsymbol{\eta}_0 + \boldsymbol{1}_n'\boldsymbol{Y}^{-1}\boldsymbol{1}_n - \boldsymbol{d}'\boldsymbol{\Sigma}^{-1}\boldsymbol{d}, \end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Sigma} &= (\boldsymbol{\Delta}_{\cdot}^{-1} + \boldsymbol{W}'\boldsymbol{Y}\boldsymbol{W}),\\
\boldsymbol{d} &= \boldsymbol{\Delta}_{\cdot}^{-1}\boldsymbol{\eta}_{\cdot} + \boldsymbol{W}'\mathbf{1}_n,\\
Q_5(\boldsymbol{\phi}) &= \frac{\boldsymbol{\phi}'(\boldsymbol{D_M} - \xi\boldsymbol{M})^{-1}\boldsymbol{\phi}}{2} - g_{\cdot\cdot}
\end{aligned}$$

با توجه به پیچیدگی توزیع پسین توأم پارامترها در رابطه‌ی (۷.۳) به‌دست آوردن نمایش بسته برای کمیت‌های پسینی شدنی نیست. از این‌رو، در ادامه از روش‌های مونت کارلو زنجیر مارکفی برای نمونه‌گیری از این توزیع پسین و تقریب کمیت‌های پسینی مورد علاقه استفاده می‌شود.

۱.۳ توزیع‌های پسین شرطی کامل

در این بخش برای نمونه‌گیری از توزیع پسین، یک الگوریتم گیبز توسعه داده می‌شود. برای این منظور لازم است توزیع‌های پسین شرطی کامل متناظر پارامترهای مدل محاسبه شوند. برپایه‌ی توزیع پسین توأم پارامترها در رابطه‌ی (۷.۳)، توزیع پسین شرطی کامل پارامتر λ، σ^2، $\boldsymbol{\theta}$ و $\boldsymbol{\phi}$ به‌ترتیب به‌صورت

$$\begin{aligned}
\lambda \mid others &\sim G\left(\frac{a_{\cdot} + p + n}{2}, \frac{Q_2(\boldsymbol{\theta}) + Q_3(b, \boldsymbol{\eta}_{\cdot}, \boldsymbol{\Delta}_{\cdot}) + Q_4(\boldsymbol{\theta}, \boldsymbol{\phi})}{2}\right)\\
\frac{1}{\sigma^2} \mid others &\sim G\left(f_{\cdot}, Q_5(\boldsymbol{\phi})\right)\\
\pi(\boldsymbol{\theta} \mid others) &\propto \prod_{i=1}^{t}\prod_{j=1}^{r} \exp\left\{-\frac{\lambda}{2}\left(Q_2(\boldsymbol{\theta}) + Q_4(\boldsymbol{\theta}, \boldsymbol{\phi})\right)\right\},\\
\pi(\boldsymbol{\phi} \mid others) &\propto \prod_{i=1}^{t}\prod_{j=1}^{r} \exp\left\{-\frac{\lambda}{2}Q_4(\boldsymbol{\theta}, \boldsymbol{\phi}) - \frac{1}{\sigma^2}Q_5(\boldsymbol{\phi})\right\},
\end{aligned}$$

به‌دست می‌آیند. همان‌گونه که ملاحظه می‌شود این توزیع‌ها نمایش بسته و شناخته شده‌ای ندارند. به‌همین دلیل، برای نمونه‌گیری از آنها لازم است از الگوریتم متروپولیس ـ هاستینگز استفاده شود. از این‌رو، برای نمونه‌گیری از توزیع پسین توأم پارامترها به یک الگوریتم گیبز با تکرارهای آشیانه‌ای متروپولیس ـ هاستینگز نیاز است. مراحل کلی الگوریتم گیبز برای نمونه‌گیری از توزیع پسین (۷.۳) به شرح زیر است:

۱. مقادیر اولیه‌ی $\boldsymbol{\Psi}^{(0)} = \left\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}^{(0)}, \lambda^{(0)}, \sigma^{2(0)}\right\}$ را برای پارامترهای مدل در نظر بگیرید.

۲. در تکرار $(m+1)$ ـ ام الگوریتم، برپایه‌ی مقادیر حاصل از تکرار (m) ـ ام، پارامترهای مدل را به‌صورت زیر به‌روز کنید:

$$\begin{aligned}
\lambda^{(m+1)} \mid others &\sim G\left(\frac{a_{\cdot} + p + n}{2}\frac{Q_2(\boldsymbol{\theta}^{(m)}) + Q_3(b, \boldsymbol{\eta}_{\cdot}, \boldsymbol{\Delta}_{\cdot}) + Q_4(\boldsymbol{\theta}^{(m)}, \boldsymbol{\phi}^{(m)})}{2}\right),\\
\frac{1}{\sigma^{2(m+1)}} \mid others &\sim G\left(f_{\cdot}, Q_6(\boldsymbol{\phi}^{(m)})\right),\\
\pi(\boldsymbol{\theta}^{(m+1)} \mid others) &\propto \prod_{i=1}^{t}\prod_{j=1}^{r} \exp\left\{-\frac{\lambda^{(m+1)}}{2}\left(Q_2(\boldsymbol{\theta}^{(m)}) + Q_4(\boldsymbol{\theta}^{(m)}, \boldsymbol{\phi}^{(m)})\right)\right\},\\
\pi(\boldsymbol{\phi}_k^{(m+1)} \mid others) &\propto \prod_{i=1}^{t}\prod_{j=1}^{r} \exp\left\{-\frac{\lambda^{(m+1)}}{2}\left(Q_4(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\phi}^{(m)})\right) - \frac{1}{\sigma^{2(m+1)}}Q_5(\boldsymbol{\phi}^{(m)})\right\}.
\end{aligned}$$

۳. گام ۲ الگوریتم را تا همگرا شدن زنجیر مارکف به توزیع مانای متناظر خود تکرار کنید.

۴ مطالعه‌ی شبیه‌سازی

در این بخش برای ارزیابی مدل پیشنهادی از یک مطالعه‌ی شبیه‌سازی استفاده شده است. برای این منظور یک مدل تحلیل کواریانس با یک متغیر کمکی و یک عامل دو سطحی، به‌ازای $\boldsymbol{\theta} = (\beta_{\cdot} = ۴, \beta_{۱} = ۱, \gamma_{۱} = ۲)$ و $\lambda = ۲$ در نظر گرفته شده است. متغیرهای پاسخ y_{ij}، $i = ۱, ۲$ و $j = ۱, \ldots, r$، از توزیع (۱.۲) با تابع پیوند (۱.۲) شبیه‌سازی شده‌اند. مقادیر ریشه دوم میانگین توان دوم خطای برآوردگرهای بیزی پارامترهای مدل پیشنهادی محاسبه و در جدول ۱ ارائه شده‌اند. تمامی محاسبات ۴۰۰۰ بار تکرار شده است. همان‌گونه که ملاحظه می‌شود مدل پیشنهادی به‌خوبی وابستگی‌های فضایی مشاهدات را نشان می‌دهد.

جدول ۱: مقادیر ریشه دوم میانگین توان دوم خطا براوردگرهای ضرایب مدل تحلیل کواریانس گاوسی وارون برای داده‌های فضایی و غیرفضایی در رهیافت بیزی.

		اندازه‌ی نمونه		
پارامتر	مدل تحلیل کواریانس	۲۰	۴۰	۸۰
$\beta_{\cdot}$	مدل فضایی پیشنهادی	۱/۴۱۰۰	۰/۸۶۶۸	۱/۷۳۵۲
	مدل کلاسیک	۳/۹۲۹۳	۱/۰۱۵۵	۲/۰۰۵۴
$\beta_{۱}$	مدل فضایی پیشنهادی	۱/۵۳۳۹	۰/۹۳۱۳	۱/۹۵۷۶
	مدل کلاسیک	۳/۸۶۷۵	۱/۰۱۲۳	۲/۰۰۰۶
γ	مدل فضایی پیشنهادی	۱/۴۹۲۳	۰/۹۳۲۱	۱/۹۸۴۸
	مدل کلاسیک	۳/۹۸۰۴	۱/۰۱۳۴	۲/۰۰۴۲

بحث و نتیجه‌گیری

تحلیل کواریانس یکی از روش‌های بسیار پرکاربرد مدل‌سازی و تلفیقی از تحلیل رگرسیونی و تحلیل واریانس است. اعتبار نتایج حاصل از این روش، به میزان درستی فرضیات زیر بنایی آن مانند مستقل بودن مشاهدات در هر تیمار و نرمال بودن توزیع مشاهدات پاسخ وابسته است. در شرایطی که این فرضیات برقرار نباشند، نتایج حاصل از تحلیل کواریانس می‌تواند به‌صورت قابل‌توجه‌ای گمراه‌کننده باشد. هنگامی که مشاهدات پاسخ مثبت و دارای همبستگی فضایی هستند، مدل تحلیل کواریانس مبتنی بر توزیع گاوسی وارون برای داده‌های فضایی که در این مقاله پیشنهاد شده است، می‌تواند به‌عنوان جایگزینی انعطاف‌پذیر برای مدل مرسوم تحلیل کواریانس به‌کار گرفته شود.

مراجع

Ananda, M. A. (1998), Bayesian and Non-Bayesian Solutions to Analysis of Covariance Models Under Heteroscedasticity, *Statistics for Spatial Data*, **86**(1), 177–192.

Cox, D. R. and McCullagh, P. (1982), Biometrics Invited Paper with Discussion. Some Aspects of Analysis of Covariance, *Biometrics*, **38**, 541–561.

Fallah, A. (2016), Gaussian Mixture Analysis of Covariance, *Journal of Statistical Computation and Simulation*, **86**(16), 3158–3174.

Fisher, R. A. (1932), *Statistical Methods for Research Workers*, 4th Edition, Oliver and Boyed, Edinburgh.

Kavousi, A., Meshkani, M. R. and Mohammadzadeh, M. (2009), Spatial Analysis of Relative Risk of Lip Cancer in Iran: a Bayesian Approach, *Environmetrics*, **20**, 347–359.

Meshkani, M. R, Fallah, A. and Kavousi, A. (2014), Analysis of Covariance Under Inverse Gaussian Model, *Journal of Applied Statistics*, **41**(6), 1189-1202.

Meshkani M. R., Fallah A. and Kavousi A. (2016), Bayesian Analysis of Covariance Under Inverse Gaussian Model, *Journal of Applied Statistics*, **43**, 280–298.

Miller, G. A. and Chapman, J. P. (2001), Misunderstanding Analysis of Covariance, *Journal of the Abnormal Psychol*, **110**, 40–48.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

مدل‌بندی بیزی داده‌های نرخ فضایی با مدل رگرسیون بتا فضایی

کبری قلی‌زاده[۱]، محسن محمدزاده
گروه آمار، دانشگاه تربیت مدرس

چکیده: مدل‌های رگرسیونی بتای فضایی ابزاری مناسب برای مدل‌بندی داده‌های نرخ وابسته فضایی است. گاهی در تحلیل این مدل‌ها علاوه بر مدل‌بندی پارامتر میانگین نیاز است پارامتر دقت نیز مدل‌بندی شود، اما به دلیل پیچیده شدن این مدل، استفاده از روش‌های معمول و الگوریتم‌های نمونه‌گیری مونت کارلوی زنجیر مارکوفی موجب افزایش زمان محاسبات می‌شود. در این مقاله از روش تقریب لاپلاس آشیانی جمع‌بسته برای تحلیل مدل رگرسیون بتا فضایی استفاده می‌شود. سپس در یک مطالعه شبیه‌سازی کارایی مدل ارائه شده مورد ارزیابی قرار گرفته و نحوه کاربست آن در یک مثال واقعی نشان داده شده است.

واژه‌های کلیدی: مدل رگرسیونی بتا فضایی، تقریب لاپلاس آشیانی جمع‌بسته.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62M30، 91B72.

۱ مقدمه

در مدل‌بندی متغیرهای نرخ که در بازه $(۰, ۱)$ توزیع شده‌اند، ممکن است استفاده از مدل رگرسیونی با فرض نرمال بودن متغیر پاسخ، پیشگویی‌هایی برای متغیر پاسخ ارائه دهد که خارج از محدوده $(۰, ۱)$ باشند. از طرفی استفاده از روش تبدیل برای حل این مساله ممکن است باعث غیرقابل تفسیر شدن پارامترهای حاصل بر حسب متغیر پاسخ واقعی باشد. بنابراین فراری و کریباری (۲۰۰۴) مدل رگرسیونی بتا را برای مدل‌بندی متغیرهایی که دارای توزیع بتا هستند، معرفی کردند. فرض کنید متغیر تصادفی Y دارای توزیع بتا است بطوری‌که $E(Y) = \mu$ و $Var(Y) = \frac{\mu(۱-\mu)}{۱+\varphi}$. در این صورت، چگالی احتمال توزیع بتا را می‌توان برحسب پارامترهای میانگین μ و دقت φ به‌صورت

$$f(y|\varphi,\mu) = \frac{\Gamma(\varphi)}{\Gamma(\mu\varphi)\Gamma((۱-\mu)\varphi)} y^{\mu\varphi-۱} y^{(۱-\mu)\varphi-۱}, \quad ۰ < y < ۱, ۰ < \mu < ۱, \varphi > ۰,$$

[۱] نام و ایمیل ارائه دهنده مقاله: کبری قلی‌زاده، k.gholizadeh@modares.ac.ir

بازپارامتری کرد. در مدل رگرسیونی بتا با ثابت در نظر گرفتن پارامتر دقت φ، تابعی از میانگین متغیر پاسخ با استفاده از متغیرهای تبیینی موردنظر به‌صورت $g(\boldsymbol{\mu}) = \mathbf{Z}'\boldsymbol{\beta}$ مدل می‌شود، که در آن $\mathbf{Z}$ مشاهدات متغیرهای تبیینی، $\boldsymbol{\beta}$ بردار اثرات ثابت و $g(\cdot)$ تابعی یکنواخت از μ است که تابع پیوند نامیده می‌شود. گاهی ممکن است پارامتر دقت بر حسب مشاهدات ثابت نباشد، اسمیتسون و ورکالین (۲۰۰۶) مدل‌بندی پارامتر φ را همزمان با مدل‌بندی پارامتر میانگین پیشنهاد کردند. برای تضمین مثبت بودن واریانس، برای پارامتر دقت تابع پیوند لگاریتم در نظر گرفته می‌شود. در این صورت مدل رگرسیونی بتا به‌صورت

$$h(\varphi) = \ln(\varphi) = \mathbf{W}'\boldsymbol{\gamma},$$

خواهد بود، که در آن $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_k)$ بردار اثرات ثابت و $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_k)$ ماتریس مشاهدات متغیرهای تبیینی با k بعد است.

در بعضی مسائل کاربردی،ممکن است داده‌ها، وابسته فضایی باشند. روش معمول برای لحاظ کردن وابستگی فضایی در مدل‌بندی داده‌های فضایی استفاده از تابع کواریانس فضایی است. در صورت قرار گرفتن موقعیت‌های فضایی مشاهدات بر روی مشبکه می‌توان ازمدل‌های قدم زدن تصادفی [1] این همبستگی را در تحلیل داده‌های فضایی لحاظ کرد. در این روش با استفاده از ویژگی مارکوفی و ماتریس همسایگی ساختار کواریانس قابل تعریف است. یک روش برای بیان ویژگی مارکوفی استفاده از گراف است. گراف G مجموعه‌ای از رأس‌هاست که توسط خانواده‌ای از یال‌ها به هم وصل شده‌اند و به صورت زوج مرتب (V, E) نشان داده می‌شود، که در آن V مجموعه‌ای متناهی و ناتهی از رئوس و E یال‌های آن است. مدل‌های قدم زدن تصادفی یک میدان تصادفی مارکوفی گاوسی [2] (GMRF) است. با توجه به تعریف گراف، بردار تصادفی $\mathbf{X} = (X_1, \cdots, X_n)$ یک GMRF تحت گراف $G = (V, E)$ با میانگین $\boldsymbol{\mu}$ و ماتریس دقت $\mathbf{Q}_{n\times n} > 0$ است، اگر و تنها اگر تابع چگالی آن به صورت

$$\pi(\mathbf{x}) = (2\pi)^{-\frac{1}{2}}|\mathbf{Q}|^{\frac{1}{2}}\exp\{\frac{-1}{2}((\mathbf{x}-\boldsymbol{\mu})^T\mathbf{Q}((\mathbf{x}-\boldsymbol{\mu})\},$$

باشد، به گونه‌ای که $Q_{ij} \neq 0$ اگر و تنها اگر $i, j \in E$ (رو و هلد، ۲۰۰۵).

گامرمن و چیپدا (۲۰۱۳) با وارد کردن اثرات تصادفی مدل رگرسیون بتا فضایی را مطرح کردند. گاهی در تحلیل بیزی این مدل‌ها توزیع پسینی فرم بسته‌ای ندارند و معمولا از الگوریتم‌های MCMC استفاده می‌شود. مدل‌بندی همزمان دو پارامتر میانگین و دقت، همچنین وجود همبستگی فضایی در مدل، موجب کاهش کارایی این الگوریتم‌ها و افزایش زمان محاسبات می‌شود. برای حل این مساله می‌توان در تحلیل بیزی مدل رگرسیونی بتا از روش تقریب لاپلاس آشیانی جمع‌بسته [3] (INLA) استفاده کرد. ارائه تقریب‌های دقیق چگالی پسینی و افزایش سرعت محاسبات از مزیت‌های روش INLA نسبت به الگوریتم‌های MCMC است (رو و همکاران ، ۲۰۰۹).

در این مقاله برای تحلیل بیزی مدل رگرسیونی بتا فضایی از روش INLA استفاده می‌شود. برای این منظور در بخش ۲ روش INLA در تحلیل بیزی مدل رگرسیونی بتا ارائه می‌شود. سپس در بخش ۳ با اجرای شبیه‌سازی لزوم مدل‌بندی همزمان پارامتر میانگین و دقت بررسی می‌شود. در بخش ۴ نحوه استفاده مدل رگرسیونی بتا در تحلیل داده‌های فضایی با مثالی کاربردی تشریح می‌شود.

[1] Random Walk
[2] Gaussian Markov Random Field
[3] Integrated Nested Laplace Approximation

۲ برآورد مدل رگرسیونی بتا با INLA

فرض کنید $Y_i \sim Beta(\mu_i\varphi_i, \varphi_i(1-\varphi_i))$ ، $i = 1, \cdots, n$ بطوری‌که هر دو پارامتر میانگین μ و دقت φ برای همه مشاهدات ثابت نیستند و به‌صورت

$$\text{logit}(\mu_i) = \mathbf{z}_i'\boldsymbol{\beta} + \boldsymbol{\nu}_i, \quad \ln(\varphi_i) = \mathbf{z}_i'\boldsymbol{\gamma}, \quad i = 1, \cdots, n$$

مدل‌بندی شوند، که در آن $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$ بردار ضرایب رگرسیونی پارامتر میانگین و $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_k)$ بردار ضرایب رگرسیونی پارامتر دقت، $\mathbf{z}_i$ بردار متغیرهای تبیینی برای میانگین و دقت و $\boldsymbol{\nu}_i$ بردار اثرات تصادفی در مدل میانگین است. در تحلیل بیزی این مدل با روش INLA بردار اثرات ثابت مدل دقت به عنوان ابرپارامتر در نظر گرفته می‌شود. برای بردارهای $\boldsymbol{\beta}$ و $\boldsymbol{\nu}$ پیشینی گاوسی در نظر گرفته و فرض می‌شود $\mathbf{x} = \{\boldsymbol{\nu}, \boldsymbol{\beta}\}$ مشخص‌کننده بردار همه متغیرهای پنهان است. توزیع چگالی میدان تصادفی $\mathbf{x}$ گاوسی با میانگین صفر و ماتریس دقت $Q(\boldsymbol{\theta}_1)$ است که در آن $\boldsymbol{\theta}_1$ بردار ابرپارامتر با بعد m است. با فرض آنکه مولفه‌های میدان پنهان $\mathbf{x}$ به شرط $\boldsymbol{\theta}_1$ مستقل هستند،میدان تصادفی پنهان $\mathbf{x}$ یک GMRF با ماتریس دقت تنک $Q(\boldsymbol{\theta}_1)$ است.

فرض کنید متغیرهای پاسخ $\{Y_i : i = 1, \ldots, n\}$ به شرط $\mathbf{x}$ و $\boldsymbol{\gamma}$ مستقل هستند و توزیع آن با $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\gamma})$ نشان داده می‌شود. برای راحتی $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\theta}_1^T)$ با بعد $dim(\boldsymbol{\theta}) = m + k$ در نظر گرفته می‌شود. در این صورت، چگالی پسینی $\mathbf{x}$ و $\boldsymbol{\theta}$ به‌صورت

$$\begin{aligned}
\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) &\propto \pi(\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}_1)\prod_{i=1}^{n}\pi(y_i|\mathbf{x}_i, \boldsymbol{\gamma}) \\
&\propto \pi(\boldsymbol{\theta})|Q(\boldsymbol{\theta}_1)|^{\frac{1}{2}}\exp\{-\frac{1}{2}\mathbf{x}^TQ(\boldsymbol{\theta}_1)\mathbf{x} + \sum_{i=1}^{n}\log\{\pi(y_i|\mathbf{x}_i, \boldsymbol{\gamma}\}\},
\end{aligned}$$

خواهد بود. در روش محاسباتی INLA، هدف اصلی توزیع پسینی کناری $\pi(x_i|\mathbf{y})$ و $\pi(\theta_j|\mathbf{y})$ است که می‌توانند به‌صورت

$$\begin{aligned}
\pi(x_i|\mathbf{y}) &= \int \pi(x_i|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \\
\pi(\theta_j|\mathbf{y}) &= \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j},
\end{aligned}$$

بازنویسی شوند. برای برآورد این پسینی‌ها به روش INLA مراحل زیر انجام می‌شود:

(i) محاسبه$\pi(\boldsymbol{\theta}|\mathbf{y})$ ، که از طریق آن می‌توان همه پسینی‌های کناری $\pi(\theta_j|\mathbf{y})$ را بدست آورد.

(ii) محاسبه $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$، که برای محاسبه پسینی کناری $\pi(x_i|\mathbf{y})$ مورد نیاز است.

روش INLA با توجه به مفروضات مدل، تقریب‌های عددی بر اساس روش تقریب لاپلاس برای برآورد پسینی‌های مورد نظر ارائه می‌دهد (ترنی و کادان ، ۱۹۸۶) .

مرحله (i) شامل محاسبه تقریب پسینی توام ابرپارامترها به صورت

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\pi(\mathbf{x},\boldsymbol{\theta}|\mathbf{y})}{\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\\ &= \frac{\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\pi(\mathbf{x},\boldsymbol{\theta})}{\pi(\mathbf{y})}\frac{۱}{\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\\ &= \frac{\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})}\frac{۱}{\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\\ &\propto \frac{\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\\ &\approx \frac{\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\widetilde{\pi}(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})},\end{aligned}$$

است، که در آن $\widetilde{\pi}(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ تقریب گاوسی چگالی شرطی کامل $\mathbf{x}$، و $\mathbf{x}^*(\boldsymbol{\theta})$ مد این چگالی است (رو و هلد ، ۲۰۰۵) . در مرحله (ii)، توزیع‌های پسینی شرطی $\pi(x_i|\mathbf{y},\boldsymbol{\theta})$ می‌توانند با استفاده از تقریب لاپلاس به صورت

$$\widetilde{\pi}_{LA}(x_i|\boldsymbol{\theta},\mathbf{y}) \propto \frac{\pi(\mathbf{x},\boldsymbol{\theta},\mathbf{y})}{\widetilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y})}|_{\mathbf{x}_{-i}=\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})},$$

برآورد شوند، که در آن $\widetilde{\pi}_{GG}$ تقریب گاوسی $\pi_G(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y})$ و $\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})$ مد آن است. بعد از انجام دو مرحله، چگالی پسینی کناری $\pi(x_i|\mathbf{y})$ به‌صورت

$$\widetilde{\pi}(x_i|\mathbf{y}) = \sum_k \widetilde{\pi}(x_i|\boldsymbol{\theta}_k,\mathbf{y})\widetilde{\pi}(\boldsymbol{\theta}_k|\mathbf{y})\Delta_k,$$

بدست می‌آید، که در آن Δ_k وزن‌های متناظر با نقاط انتگرال‌گیری $\boldsymbol{\theta}_k$ است.

در این مقاله برای مقایسه برازش و پیچیدگی مدل‌ها از دو ملاک انحراف اطلاع[۴] (DIC) و ملاک اطلاع واتانابه آکایک[۵] (WAIC) استفاده می‌شود. فرض کنید $\pi(y|\theta)$ چگالی پسینی است، که در آن θ پارامتر مدل است. ملاک DIC به‌صورت

$$\begin{aligned}\text{DIC} &= -۲\log\pi(y|\hat{\theta}) + ۲(\log\pi(y|\hat{\theta}) - \text{E}(\log(\pi(y|\theta)))\\ &= -۲\log\pi(y|\hat{\theta}) + ۲p_{\text{DIC}},\end{aligned}$$

تعریف می‌شود، که در آن p_{DIC} تعداد پارامترهای فعال مدل است (اشپیگلالتر و همکاران ، ۲۰۰۲) . ملاک WAIC که پیچیدگی و برازش مدل را اندازه‌گیری می‌کند به صورت

$$\text{WAIC} = -۲\log\pi(y|\hat{\theta}) + ۲p_{\text{WAIC}},$$

محاسبه می‌شود، که در آن $p_{\text{WAIC}} = ۲\sum_{i=۱}^{n}(\log\text{E}(\pi(y|\hat{\theta})) - \text{E}(\log(\pi(y|\hat{\theta})))$ است (واتانابه ، ۲۰۱۰). برای مقایسه مدل‌ها، از نظر اعتبار پیشگویی y_i می‌توان از ملاک پیشگویی شرطی مولفه‌ها[۶] (CPO) به صورت $CPO_i = \pi(y_i|y_{-i})$ استفاده کرد، که در آن y_{-i} همه نمونه بجز y_i است. در اینجا از ملاک

$$\text{LS} = -\frac{۱}{n}\sum_{i=۱}^{n}\log CPO_i$$

[4] Deviance Information Criterion
[5] Watanabe-Akaike Information Criterion
[6] Conditional Predictive Ordinates

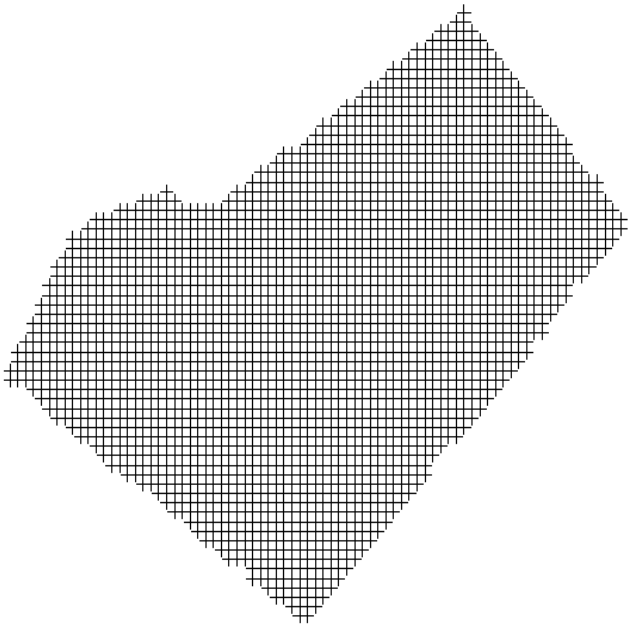
نیز برای مقایسه مدل‌ها استفاده می‌شود، که نمره لگاریتم اعتبارسنجی متقابل[۷] (LS) نامیده می‌شود (نایتینگ و همکاران ، ۲۰۰۷).

۳ مطالعه شبیه‌سازی

در این مدل‌بندی پارامتر دقت در مدل‌های رگرسیونی بتا با واریانس‌های مختلف مورد ارزیابی عددی قرار می‌گیرد. مجموعه داده از توزیع بتا $Y_i \sim Beta(\mu_{st}, \varphi_t)$، بر روی مشبکه‌ای با ۲۵۷۴ , $s = ۱, \cdots$ موقعیت و ۴ , $t = ۱, \cdots$ تکرار تولید می‌شود، که در آن پارامتر میانگین و دقت به صورت

$$\text{logit}(\mu_{st}) = i_t'\boldsymbol{\beta} + \beta_x x_۱ + f_s(s), \quad \log(\varphi_t) = i_t'\boldsymbol{\gamma},$$

مدل شده‌اند. قابل ذکر است داده‌ها بر روی مشبکه‌ای مشابه موقعیت‌های داده واقعی بخش بعد که نمای کلی آن در شکل ۱ نشان داده شده است، شبیه‌سازی شده‌اند.



شکل ۱ : مشبکه موقعیت‌ها برای شبیه‌سازی داده‌ها.

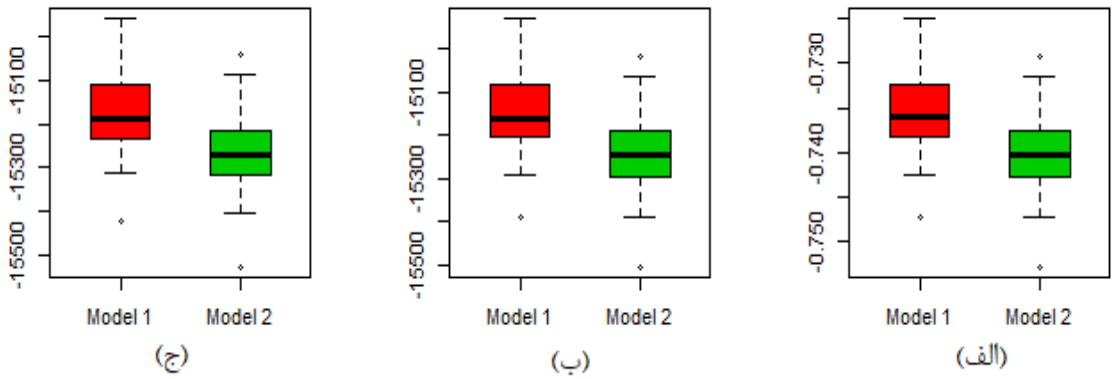
برای محاسبه پارامترهای μ_{st} وφ_t، مقادیر واقعی $\boldsymbol{\beta} = (\beta_۱, \beta_۲, \beta_۳, \beta_۴) = (۰, ۱, ۱/۵, ۲)$ ، $\beta_x = ۲$ و $\boldsymbol{\gamma} = (\gamma_۱, \gamma_۲, \gamma_۳, \gamma_۴) = (۲۵, ۲۰, ۳۰, ۳۵)$ در نظر گرفته شده‌اند. اثر تصادفی فضایی $f_s(s)$ از مدل قدم زدن تصادفی مرتبه اول روی مشبکه نمونه گرفته شده است. متغیر تبیینی x از توزیع نرمال با میانگین ۰ و واریانس ۰/۴ تولید شده است. سپس با محاسبه مقادیر μ_{st} و φ_t، از توزیع $Beta(\mu_{st}, \varphi_t)$ نمونه Y_i را تولید نموده و این کار ۱۰۰ بار تکرار می‌شود. در هر تکرار دو مدل، اولی با واریانس یکسان در تکرارها (مدل ۱) و دومی با واریانس متفاوت در تکرارها (مدل ۲) برازش و برآورد پارامترها با استفاده از INLA محاسبه می‌شود. به عبارت دیگر، در هر دو مدل میانگین یکسان است ولی دقت در مدل ۱ به‌صورت $\log(\varphi_t) = \gamma$ و در مدل ۲ به صورت $\log(\varphi_t) = i_t'\boldsymbol{\gamma}$ در نظر گرفته شده است. میانگین و انحراف استاندارد برآورد پارامترها در این ۱۰۰ تکرار در جدول ۱ نشان داده شده است. همان طور که در جدول ۱ ملاحظه می‌شود پارامترها

[7] Cross Validated Logarithmic Score

در مدل ۲ در بسیاری موارد با دقت بیشتری نسبت به مدل ۱ برآورد شده‌اند. برای مقایسه بهتر این دو مدل، معیارهای DIC، WAIC و LS در این شبیه‌سازی برای ۱۰۰ تکرار محاسبه شده است. نمودار جعبه‌ای این معیارها در شکل ۲ نمایش داده شده است. همانطور که ملاحظه می‌شود مقادیر سه معیار برای مدل ۱ بیشتر است، یعنی عملکرد مدل ۲ بهتر از مدل ۱ است.

جدول ۱: میانگین پارامترهای برآورد شده مدل‌ها در ۱۰۰ تکرار.

مدل	پارامتر	مقدار واقعی	میانگین برآورد	انحراف استاندارد	چندک	
					۰/۰۲۵	۰/۹۷۵
۱	β_x	۱	۰/۸۷۵	۰/۰۶۸	۰/۷۴۲	۱/۰۰۸
	β_1	۰	−۰/۰۰۱	۰/۰۱۲	−۰/۰۲۵	۰/۰۲۱
	β_2	۱	۰/۹۷۵	۰/۰۱۳	۰/۹۳۲	۰/۹۸۳
	β_3	۱/۵	۱/۳۹۸	۰/۰۱۴	۱/۳۶۱	۱/۴۱۷
	β_4	۲	۱/۸۱۲	۰/۰۱۶	۱/۷۸۱	۱/۸۴۳
۲	β_x	۱	۰/۸۷۹	۰/۰۶۸	۰/۷۴۵	۰/۰۱۳
	β_1	۰	−۰/۰۰۲	۰/۰۱۲	−۰/۰۲۶	۰/۰۲۲
	β_2	۱۱	۰/۹۴۱	۰/۰۱۴	۰/۹۱۴	۰/۹۶۹
	β_3	۱/۵	۱/۴۰۴	۰/۰۱۴	۱/۳۷۷	۱/۴۳۲
	β_4	۲	۱/۸۵۵	۰/۰۱۶	۱/۸۲۴	۱/۸۵۵

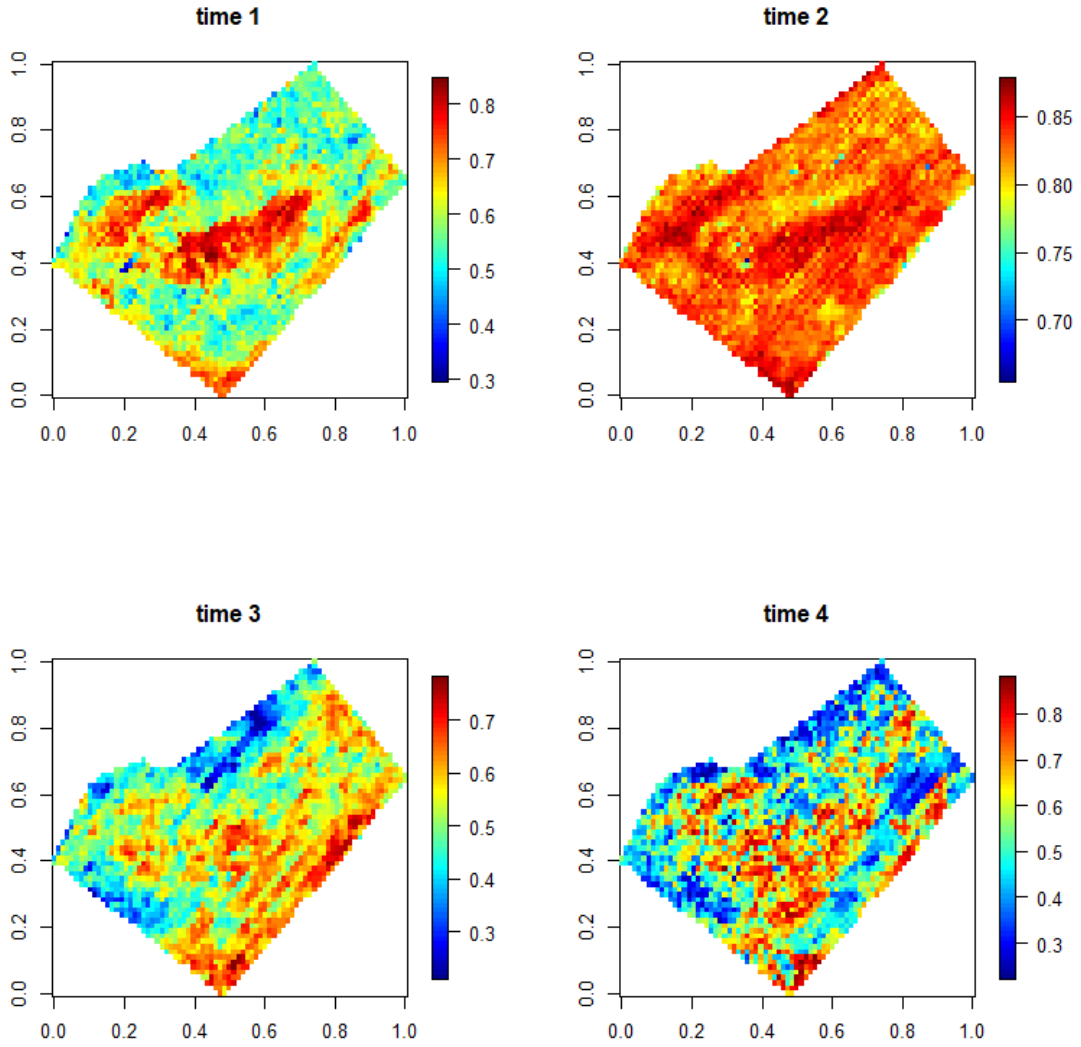


شکل ۲: معیارهای ارزیابی دو مدل الف- LS ب- WAIC ج- DIC.

۴ مدل‌بندی داده‌های رشد گیاه و ویژگی‌های خاک

داده‌های مورد مطالعه در این بخش در مورد استفاده از تکنولوژی حسگرها برای تحلیل مکانی و زمانی روابط خاک و گیاه در اجرای شیوه‌های کارآمد مدیریت کشاورزی است (پولیس و همکاران ، ۲۰۱۹). هدف مدلبندی روابط بین رشد گیاه

و ویژگی خاک است که توسط دو حسگر در مزرعه یونجه به وسعت هفت هکتار در پالومونته در ایتالیا جنوبی اندازه‌گیری می‌شود. متغیرهای پاسخ و تبیینی به ترتیب شاخص پوشش گیاهی [۸] (NDVI) و مقاومت الکتریکی خاک [۹].است(ER) موقعیت مکانی شامل مشبکه مربعی به تعداد ۲۵۷۴ موقعیت است. اندازه‌گیری های میدانی NDVI چهار مرتبه در فصول مختلف و ER فقط یک بار در سه لایه به عمق ۰/۵، ۱ و ۲ متر صورت گرفته است. شکل ۳ پهنه‌بندی NDVI در چهار زمان را نشان می‌دهد. همان‌طور که ملاحظه می‌شود NDVI دارای بالاترین مقدار و کمترین پراکندگی در مرتبه زمانی دوم است. در شکل ۴ پهنه‌بندی ER در سه عمق نمایش داده شده است. با توجه به این نمودار ER دارای تنوع سیستماتیک قوی در طول سه عمق نیست بنابراین از میانگین آن در این سه لایه در مدل‌بندی استفاده خواهد شد.

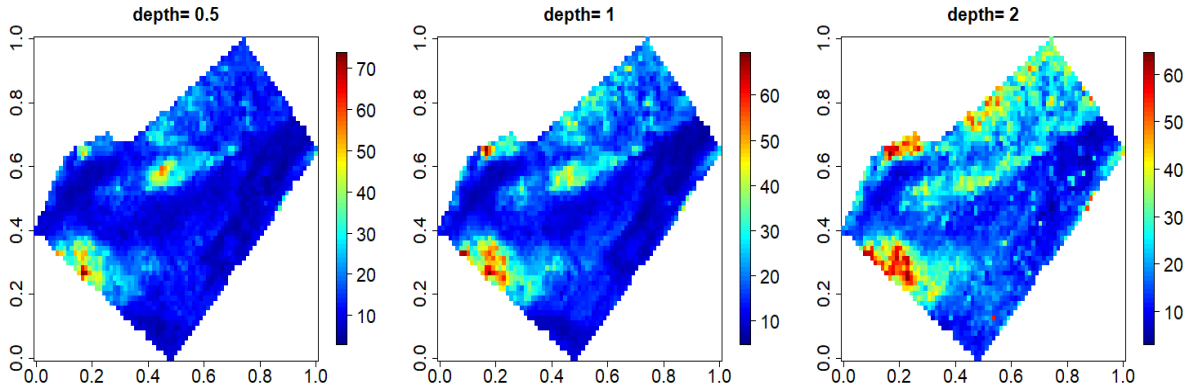


شکل ۳: پهنه‌بندی اندازه‌گیری‌های میدانی در چهار زمان مختلف.

برای بررسی ارتباط بین NDVI و ER مدل دو مدل در نظر گرفته می‌شود، در مدل اول تنها پارامتر میانگین به‌صورت

$$\text{logit}(\boldsymbol{\mu}_{st}) = \beta^{\mu}_{\cdot} + \boldsymbol{i}'_t\boldsymbol{\beta}^{\mu}_{1} + f^{\mu}_{1}(ER_s) + f^{\mu}_{2}(s), \qquad (۱.۴)$$

[8] The Normalized Density Vegetation Index
[9] The Soil Electrical Resistivity



شکل ۴: پهنه‌بندی میزان مقاومت الکتریکی خاک در سه عمق مختلف.

جدول ۲: اندازه معیارهای ارزیابی در دو مدل.

	DIC	WAIC	LS
مدل ۱	−۳۳۰۶۳/۴۴	−۳۲۸۰۶/۷۱	−۱/۵۵۹۱
مدل ۲	−۳۷۸۳۴/۲۲	−۳۷۹۵۰/۲۶	−۱/۸۱۸۶

مدل می‌شود، که در آن $\boldsymbol{i}'_t$ متغیر تبینی ظاهری زمان با ۴ سطح، $f^{\mu}_1(\cdot)$ تابع هموار غیرخطی از متغیر ER پنهان و $f^{\mu}_2(\cdot)$ تابع هموار غیرخطی از موقعیت‌ها برای برآورد اثر فضایی است. مدل پیشینی قدم‌زدن تصادفی مرتبه اول برای $f^{\mu}_1(\cdot)$ و مدل پیشینی قدم‌زدن تصادفی دو بعدی برای $f^{\mu}_2(\cdot)$ در نظر گرفته شده‌اند. پارامتر دقت در ۴ تکرار ثابت در نظر گرفته شده است. همانطور که در شکل ۳ ملاحظه می‌شود، واریانس NDVI در چهار تکرار متفاوت است. بنابراین علاوه بر مدل کردن میانگین به‌صورت (۱.۴) پارامتر دقت نیز به‌صورت

$$\ln(\varphi_{st}) = \boldsymbol{i}'_t\gamma, \qquad (۲.۴)$$

مدل می‌شود، که در آن $\boldsymbol{i}'_t$ متغیر ظاهری ۴ سطحی زمان است. برای مقایسه این دو مدل معیارهای DIC و WAIC و LS محاسبه و در جدول ۲ نشان داده شده‌اند. همان‌طور که ملاحظه می‌شود با توجه به هر سه معیار عملکرد مدل ۲ از مدل ۱ بهتر است.

بحث و نتیجه‌گیری

در این مقاله مدل رگرسیونی بتا فضایی با مدل‌بندی هم‌زمان پارامتر میانگین و دقت ارائه شد. همان‌طور که در بخش شبیه‌سازی مشاهده شد در صورت ثابت نبودن پارامتر دقت بر حسب متغیر تبیینی نیاز است که این پارامتر علاوه بر پارامتر میانگین مدل شود. در بخش داده واقعی دو مدل برای تحلیل داده‌های پوشش گیاهی درنظر گرفته شد. مدل میانگین هر دو مدل یکسان بود اما پارامتر دقت در ۴ تکرار در مدل اول ثابت و در مدل دوم متفاوت در نظر گرفته شد. در نهایت مشاهده شد که مدل با واریانس متفاوت براساس معیارهای در نظر گرفته شده مدلی بهتر است که این نتیجه با توجه به تحلیل اکتشافی داده‌ها نتیجه منطقی است.

مراجع

Ferrari S. and Cribari F. (2004), Beta Regression for Modelling Rates and Proportions, *Journal of Applied Statistics*, **31**, 799-815.

Gamerman D. and Cepeda-Cuervo E. (2013), Generalized Spatial Dispersion Models, *Technical report*, Universidade Federal do Rio de Janeiro.

Gneiting, T. and Raftery, A. E., (2007), Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association*, **102**, 359–378.

Rue H., Martino S. and Chopin N., (2009), Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion), *Journal of the Royal Statistical Society*, **71**, 319-392.

Rue H. and Held I. (2005), *Gaussian Markov Random Field: Theory and Application*. vol 104 of Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Smithson, M. and Verkuilen, J. (2006), A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables, *Psychological Methods*, **11**, 54-71.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), Bayesian Measures of Model Complexity and Fit, *Journal of the Royal Statistical Society: Series B*, **64**, 583– 639.

Tierney, L. and Kadane, J. (1986), Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association*, **81**, 82-86.

Pollice, A., Lasinio, G., Rossi, R., Amato, M., Kneib, T., and Lang, S. (2019), Bayesian Measurement Error Correction in Structured Additive Distributional Regression with an Application to the Analysis of Sensor Data on Soil-Plant Variability, *Stochastic Environmental Research and Risk Assessment*, **33**, 747-763.

Watanabe, S. (2010), Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory, *Journal of Machine Learning Research*, **11**, 3571–3594.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

ارتقای الگوریتم چندهدفه ازدحام ذرات برای نمونه‌گیری بهینه فضایی

الهه لطفیان[۱]**، محسن محمدزاده**
گروه آمار، دانشگاه تربیت مدرس

چکیده: تاکنون مطالعات زیادی در رابطه با بهینه‌سازی تک‌هدفه نمونه‌گیری فضایی انجام شده است، ولی در بسیاری از مسائل واقعی، بیش از یک هدف برای نمونه‌گیری فضایی حائز اهمیت است. با توجه به خلا مطالعاتی موجود، استفاده از روش‌های بهینه‌سازی چندهدفه به عنوان راهکاری مناسب برای حل مسائل بهینه‌سازی نمونه‌گیری فضایی با چند هدف پیشنهاد شده است. در این مقاله، یک مسئله بهینه‌سازی دوهدفه برای نمونه‌برداری از رسوبات ساحلی رودخانه میوز جهت ارزیابی میزان آلودگی آن به فلزات سنگین در نظر گرفته شده است. توابع هدف مورد مطالعه، واریانس میانگین نمونه و مسافت طی‌شده بین موقعیت‌های نمونه هستند. برای این منظور، الگوریتم چندهدفه ازدحام ذرات برای نمونه‌گیری فضایی، ارتقاء داده شده و همبستگی فضایی داده‌ها در الگوریتم لحاظ شده است. همچنین، حساسیت مقادیر توابع هدف نسبت به اندازه نمونه و عملکرد الگوریتم بر اساس معیارهای ارزیابی مورد بررسی قرار گرفته است. نتایج حاصل بیانگر عملکرد مناسب الگوریتم چندهدفه ازدحام ذرات ارتقا یافته در بهینه‌سازی این مسئله از نمونه‌گیری فضایی است.

واژه‌های کلیدی: نمونه‌گیری فضایی، بهینه‌سازی چندهدفه، الگوریتم چندهدفه ازدحام ذرات.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 90C29.

۱ مقدمه

یکی از ارکان اساسی جمع‌آوری اطلاعات برای دست‌یابی به نتایج آماری معتبر، تعیین طرح نمونه بهینه است. در مواجهه با داده‌هایی که همبستگی آن‌ها ناشی از موقعیت قرارگیری در یک فضای معین است و به اصطلاح همبسته فضایی هستند، نمی‌توان از رویکردهای مرسوم نمونه‌گیری استفاده کرد (محمدزاده ، ۱۳۹۸). در این حالت باید همبستگی فضایی داده‌ها در طراحی نمونه‌گیری لحاظ شود. طیف گسترده‌ای از مطالعات کاربردی از جمله بررسی خاک، محیط زیست، هواشناسی و غیره با داده‌های فضایی سروکار دارند که نیازمند استفاده از روش‌های نمونه‌گیری فضایی هستند. انتخاب موقعیت‌های

[۱] نام و ایمیل ارائه دهنده مقاله: الهه لطفیان، e.lotfian@modares.ac.ir

نمونه‌گیری در ناحیه مورد مطالعه به منظور بهینه‌سازی برحسب معیارهای ارزیابی، نمونه‌گیری فضایی بهینه نامیده می‌شود. با توجه به ماهیت مسئله، معیارهای مختلفی برای ارزیابی و اخذ نمونه‌های فضایی بهینه تعریف شده است. واریانس پیشگو و آنتروپی پیشگو از جمله معیارهای مورد توجه هستند. همچنین معیارهای دیگری مانند هزینه و زمان نیز بر نحوه نمونه‌گیری اثر گذارند. از آنجا که این معیارها لزوما هم‌راستا نیستند، برای بهینه‌سازی هم‌زمان برحسب چند معیار، نمی‌توان از روش‌های مرسوم بهینه‌سازی تک‌هدفه استفاده کرد. استفاده از روش‌های بهینه‌سازی چندهدفه می‌تواند راهکاری مناسب برای حل این مشکل باشد. اگر چه مطالعات بسیاری در زمینه بهینه‌سازی تک‌هدفه نمونه‌گیری فضایی صورت گرفته، اما توجه زیادی به بهینه‌سازی چندهدفه نشده است. مک براتنی و همکاران (۱۹۸۱) تحت فرض تغییرنگار معلوم، با هدف مینیمم‌کردن واریانس کریگیدن طرح نمونه بهینه را ارائه دادند. مک براتنی و وبستر (۱۹۸۳) به‌جای مینیمم‌کردن واریانس کریگیدن، مینیمم‌کردن واریانس کریگیدن بلوکی را مورد بررسی قرار دادند. ساکس و شیلر (۱۹۸۸) چندین الگوریتم مبتنی بر نوردیدن برای بهینه‌سازی طرح‌های نمونه فضایی ارائه کردند. گرونگین و استین (۱۹۹۸) روش نوردیدن شبیه‌سازی شده فضایی (SSA)[1] را برای نمونه‌گیری فضایی معرفی کردند. لارک (۲۰۰۲) از این روش با هدف برآورد بهینه تغییرنگار برای دست‌یابی به طرح نمونه استفاده کرد. مولر و همکاران (۲۰۰۴) با استفاده از شبیه‌سازی زنجیر مارکوفی ناهمگن و الگوریتم SSA به طرح نمونه بهینه دست یافتند. فیونتس و همکاران (۲۰۰۷) با استفاده از یک معیار بهینگی مبتنی بر ماکسیمم آنتروپی و الگوریتم نوردیدن شبیه‌سازی شده فضایی طرح نمونه بهینه را به دست آوردند. بیداربخت و همکاران (۱۳۹۳) یک طرح نمونه‌گیری فضایی بهینه برای پیشگویی میزان بارندگی در استان خوزستان با استفاده از معیار آنتروپی تعیین کردند. گودز و همکاران (۲۰۱۱) با استفاده از الگوریتم SSA و الگوریتم ژنتیک هیبریدی با هدف مینیمم کردن میانگین واریانس پیشگو در داده‌های زمین آماری به طرح نمونه بهینه دست یافتند. مارچانت و همکاران (۲۰۱۳) از روش SSA برای طراحی نمونه‌گیری چندمرحله‌ای خاک‌های آلوده شهری استفاده کردند. وانگ و همکاران (۲۰۱۴) این روش را برای برآورد ناحیه‌ای محصول ناخالص اولیه بکار بردند. همچنین، این روش توسط گومزکاردناس و همکاران (۲۰۱۵) برای نمونه‌گیری فضایی بهینه در طول آزمایش غیرمخرب سازه‌های بتنی مورد استفاده قرار گرفت. در این مطالعات مسئله مورد بررسی به‌صورت تک‌هدفه بیان شده است. اما در بسیاری از مسائل کاربردی، بیش از یک هدف برای نمونه‌گیری اهمیت دارد و گاهی این اهداف در تقابل با یکدیگر هستند. یکی از رویکردهای مرسوم برای حل این مسائل، استفاده از روش‌های اسکالرسازی[2] است که توابع هدف را به یک تابع هدف تبدیل می‌کنند و به این ترتیب، مسئله با استفاده از روش‌های بهینه‌سازی تک‌هدفه قابل حل است (گرکو و همکاران، ۲۰۰۵). گر چه این روش‌ها به علت سادگی محاسباتی، به‌طور گسترده در ادبیات موضوع استفاده شده‌اند، اما در بسیاری از مواقع قادر نیستند توزیع مناسبی از جواب‌های بهینه را مشخص کنند و تنها زمانی که مرز پارتو محدب است می‌توانند به خوبی مرز را پوشش دهند (ارگوت و همکاران ، ۲۰۰۵). به منظور اجتناب از این مشکلات، الگوریتم‌های بهینه‌سازی چندهدفه مورد توجه قرار گرفته‌اند. لارک (۲۰۱۶) الگوریتم بهینه‌سازی چندهدفه مبتنی بر نوردیدن شبیه‌سازی شده (AMOSA)[3] را برای نمونه‌گیری فضایی توسیع داد. لطفیان و محمدزاده (۱۳۹۷) این الگوریتم را برای دست یافتن به جواب‌های بهینه استوار ارتقا دادند.

هدف اصلی این مقاله، ارتقای الگوریتم بهینه‌سازی چندهدفه ازدحام ذرات (MOPSO)[4] در نمونه‌گیری فضایی و بررسی عملکرد آن است. برای این منظور، در بخش ۲ برخی مفاهیم اولیه در زمینه بهینه‌سازی چندهدفه، الگوریتم MOPSO و معیارهای ارزیابی عملکرد الگوریتم ارائه می‌شود. در بخش ۳، مطالعه موردی و نتایج حاصل از بکارگیری الگوریتم چندهدفه ازدحام ذرات ارتقا یافته برای نمونه‌گیری فضایی تشریح شده است. در نهایت، به بحث و نتیجه‌گیری پرداخته

[1] Spatial Simulated Annealing
[2] Scalarization
[3] Archived Multiobjective Simulated Annealing
[4] Multi-objective Particle Swarm Optimization

می‌شود.

۲ بهینه‌سازی چندهدفه

یک مسئله بهینه‌سازی چندهدفه در حالت کلی به‌صورت

$$\begin{aligned} &\min f(x) = (f_۱(x), \dots, f_m(x)), \\ &g_j(x) \le ۰, \quad j \in J = \{۱, \dots, l\}, \\ &x \in X \subseteq \mathbb{R}^n, \end{aligned} \tag{۱.۲}$$

بیان می‌شود که در آن $f_i, g_j : \mathbb{R}^n \to \mathbb{R}, (i \in I := \{۱, \dots, m\}, j \in J)$ و $m > ۱$ هستند. مجموعه ناتهی

$$S := \{x \in \mathbb{R}^n : x \in X,\ g_j(x) \le ۰, j \in J\},$$

مجموعه شدنی و $\mathbb{R}^m$، فضای هدف نامیده می‌شوند. به دلیل وجود چند تابع هدف، بعد فضای تصویر بیشتر از یک بوده و اعمال یک ترتیب کلی استاندارد روی آن ناممکن است (ارگوت و همکاران ، ۲۰۰۵). با توجه به این چالش، تعریف کردن جواب بهینه به فرم معمول برای مسائل بهینه‌سازی چندهدفه میسر نیست. بدین منظور، مفهوم جواب بهینه پارتو [۵] مورد توجه قرار می‌گیرد.

تعریف ۱.۲. در مسئله مینیمم‌سازی (۱.۲)، جواب شدنی $\hat{x} \in X$ بهینه پارتو (نامغلوب) نامیده می‌شود هرگاه جواب $x \in X$ وجود نداشته باشد که برای $k \in \{۱, \cdots, m\}$، $f_k(x) \le f_k(\hat{x})$ و برای iای، $f_i(x) < f_i(\hat{x})$.

روش‌های گوناگونی برای یافتن جواب‌های پارتو وجود دارد. اکثر این روش‌ها، توابع هدف را تبدیل به یک تابع هدف می‌کنند که در اصطلاح به روش‌های اسکالرسازی معروفند. این روش‌ها به علت سادگی در حل، در ادبیات موضوع به‌طور مکرر استفاده می‌شوند، ولی دارای اشکالاتی نیز می‌باشند. به عنوان مثال، اگر چه روش جمع وزنی در بسیاری از مسائل بهینه‌سازی چندهدفه برای تولید مرز پارتو مورد استفاده قرار می‌گیرد، ولی تنها زمانی که مرز پارتو محدب است می‌تواند از تمامی بخش‌های مرز نماینده داشته باشد. علاوه براین، حتی برای مرز پارتو محدب نیز مجموعه وزن‌های مساوی، مجموعه پارتویی با توزیع مساوی تولید نمی‌کنند و قادر نیستند توزیع مناسبی از جواب‌های پارتو را مشخص کنند. بنابراین، استفاده از این روش‌ها ممکن است با مشکلاتی روبرو شود و نتایج مبتنی بر آن‌ها از دقت کافی برخوردار نباشند. با توجه به اشکالات ذکر شده در رابطه با روش‌های اسکالرسازی و همچنین محدودیت‌های موجود در طراحی مدل به‌صورت تک‌هدفه، استفاده از الگوریتم‌های بهینه‌سازی چندهدفه مرسوم شده است. الگوریتم MOPSO از جمله الگوریتم‌های پرکاربرد در حل بسیاری از مسائل بهینه‌سازی چندهدفه است (ژو و همکاران ، ۲۰۱۱). در ادامه، روش‌های انتخاب بولتزمن [۶] و چرخ رولت [۷] معرفی می‌شوند که در روند الگوریتم مورد استفاده قرار می‌گیرند (کوئلو و همکاران ، ۲۰۰۴).

۱.۲ روش‌های انتخاب مبتنی بر شایستگی

۱.۱.۲ روش بولتزمن

فرض کنید شرط‌های زیر برای احتمال انتخاب p_i برقرار باشند:

[5] Pareto
[6] Boltzmann
[7] Roulette Wheel

۱. $۰ \leq p_i \leq ۱$

۲. $\sum_{i=۱}^{n} p_i = ۱$.

۳. اگر $n_i \leq n_j$ آنگاه $p_i \geq p_j$ (اصل شایسته‌سالاری).

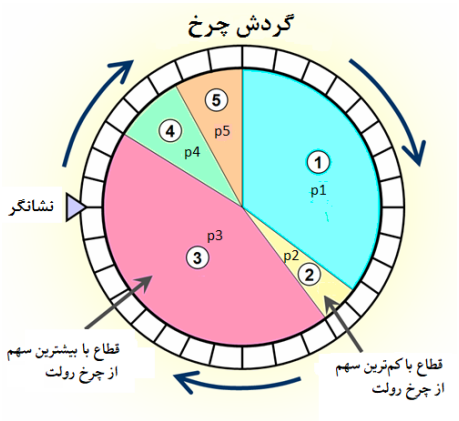
از اصل شایسته‌سالاری می‌توان نتیجه گرفت که ناحیه‌های با تعداد اعضای بیشتر احتمال کمتری برای انتخاب دارند. به‌عبارتی، در این روش به هر ناحیه، احتمال p_i به صورت

$$p_i = \frac{e^{-\beta n_i}}{\sum_j e^{-\beta n_j}}$$

نسبت داده می‌شود، که در آن n_i تعداد اعضای خانه iام و β پارامتر فشار انتخاب[۸] است.

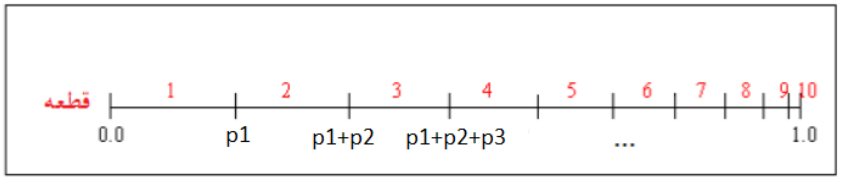
۲.۱.۲ روش چرخ رولت

در چرخ رولت، برای هر عضو، یک قطاع از چرخ اختصاص داده می‌شود. قطاع‌ها در اندازه‌های مختلف و متناسب با میزان اهمیت عضو متناظر هستند. بزرگ‌ترین قطاع به مهم‌ترین عضو و کوچک‌ترین قطاع به کم اهمیت‌ترین عضو تخصیص می‌یابد. چرخ چرخانده می‌شود و عضو مرتبط با قطاع برگزیده انتخاب می‌شود. همان‌طور که در شکل ۱ ملاحظه می‌شود، احتمال توقف نشانگر در قطاع iام برابر p_i است (پاروز ۲۰۱۴). برای سادگی بیشتر، در شکل ۲ ، دایره به صورت پاره خط



شکل ۱ : چرخ رولت

نشان داده شده است. با تولید یک عدد تصادفی r در بازه $[۰, ۱]$، یک قطاع انتخاب می‌شود. به‌طوری‌که، اگر $r \geq ۰$ و



شکل ۲ : پاره‌خط متناظر با چرخ رولت

[8] Selection Pressure Parameter

$r \leq p_1$، قطاع اول، اگر $r > p_1$ و $r \leq p_1 + p_2$، قطاع دوم و روند به همین شکل برای انتخاب سایر قطاع‌ها ادامه پیدا می‌کند.

۲.۲ الگوریتم بهینه‌سازی چندهدفه ازدحام ذرات

الگوریتم MOPSO ابتدا توسط کوئلو و همکاران (۲۰۰۴) معرفی شد. در این الگوریتم، یک آرشیو[9] خارجی برای ذخیره پاسخ‌های نامغلوبی که تاکنون تولید شده‌اند، به کار می‌رود. آرشیو، شامل دو بخش مهم کنترل‌کننده آرشیو و مشبکه‌بندی است و مهم‌ترین هدف آن نگهداری از جواب‌های نامغلوب یافت‌شده طی فرآیند جست‌وجو می‌باشد. کنترل‌کننده آرشیو تعیین می‌کند که آیا پاسخ جدیدی باید به آرشیو اضافه شود یا نه. فرآیند تصمیم‌گیری به این صورت است که پاسخ‌های نامغلوبی که در هر تکرار الگوریتم به دست می‌آیند، با اعضای آرشیو که در ابتدا تهی است مقایسه می‌شوند. اگر آرشیو خالی باشد، در این صورت پاسخ‌های فعلی قابل قبول‌اند. اگر پاسخ جدید توسط عضوی از آرشیو مغلوب شود، این پاسخ حذف و اگر هیچ یك از اعضای آرشیو، پاسخ جدید را مغلوب نکند، این پاسخ در آرشیو ذخیره می‌شود. سرانجام، اگر جمعیت آرشیو به ظرفیت ماکسیمم خود برسد، برای ایجاد مرزهای پارتو با توزیع مناسب، از روش مشبکه‌بندی تطبیقی استفاده می‌شود. در این رویکرد، فضای توابع هدف به چند ناحیه تقسیم می‌شود، اگر عضوی از آرشیو خارج از مرزهای فعلی مشبکه قرار گیرد، مشبکه باید بار دیگر محاسبه شده و هر عضو آن دوباره موقعیت‌دهی شود. سرعت و موقعیت هر ذره در این الگوریتم، به ترتیب با استفاده از رابطه‌های

$$VEL[i] = W \times VEL[i] + R_1(PBESTS[i] - POP[i]) + R_2(REP[h] - POP[i]),$$
$$POP[i] = POP[i] + VEL[i].$$

محاسبه و به‌روز می‌شوند، که در آن W وزن اینرسی، R_1 و R_2 اعداد تصادفی در بازه $[0, 1]$ و $PBESTS[i]$ بهترین موقعیت ذره iام هستند. $REP[h]$ مقداری است که از آرشیو به دست آمده و اندیس h به این صورت انتخاب می‌شود که پس از مشبکه‌بندی و محاسبه تعداد اعضای خانه‌های جدول، برای هر خانه به کمك روش بولتزمن، احتمالی در نظر گرفته می‌شود، به طوری که خانه‌های با تعداد اعضای بیشتر احتمال کمتری برای انتخاب دارند. پس از تعیین احتمال انتخاب، یك خانه به کمك روش چرخ رولت مشخص شده و در نهایت یکی از اعضای آن به تصادف انتخاب می‌شود. روند کلی الگوریتم به صورت زیر است:

۱. ایجاد جمعیت اولیه (POP)،

۲. مقداردهی اولیه به سرعت هر ذره،

۳. ارزیابی هر ذره از جمعیت،

۴. جداکردن اعضای نامغلوب جمعیت و ذخیره آن‌ها در آرشیو،

۵. مشبکه‌بندی فضای هدف کشف شده،

۶. انتخاب رهبر توسط هر ذره از میان اعضای آرشیو،

۷. به‌روز کردن بهترین موقعیت هر یك از ذرات،

[9] Archive

۸. اضافه کردن اعضای نامغلوب جمعیت فعلی به آرشیو،

۹. حذف کردن اعضای مغلوب آرشیو،

۱۰. اگر تعداد اعضای آرشیو بیش از ظرفیت تعیین شده باشد، اعضای اضافی نیز حذف می‌شوند (اندازه آرشیو محدود است)،

۱۱. اگر شرایط خاتمه محقق نشده باشد، به مرحله ۵ برمی‌گردیم و در غیر این صورت پایان.

۳.۲ معیارهای ارزیابی عملکرد الگوریتم

برای مقایسه جواب‌های بهینه بدست آمده توسط الگوریتم از لحاظ پوشش و همگرایی مرز پارتو، از معیارهایی مانند میانگین فاصله ایده‌ال [۱۰] (MID) ، متر فاصله[۱۱] (SM) ، متر پراکندگی[۱۲] (DM) و گسترش جواب‌های نامغلوب[۱۳] (SNS) استفاده می‌شود (مقصودلو و همکاران ، ۲۰۱۶).

میانگین فاصله ایده‌ال (MID): برای اندازه‌گیری فاصله جواب‌های مرزهای پارتو از یک نقطه ایده‌ال، از این معیار استفاده می‌شود. این معیار به صورت

$$MID = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{f_{1i}-f_{1best}}{f_{1total}^{\max}-f_{1total}^{\min}}\right)^2+\left(\frac{f_{2i}-f_{2best}}{f_{2total}^{\max}-f_{2total}^{\min}}\right)^2\right]^{1/2},$$

تعریف می‌شود، که در آن n تعداد جواب‌های نامغلوب، $f_{1total}^{\max}$ و $f_{2total}^{\max}$ به ترتیب بزرگ‌ترین مقادیر توابع هدف اول و دوم در بین تمام جواب‌های نامغلوب، $f_{1total}^{\min}$ و $f_{2total}^{\min}$ به ترتیب کوچک‌ترین مقادیر توابع هدف اول و دوم در بین تمام جواب‌های نامغلوب و f_{1best} و f_{2best} به ترتیب بهترین مقادیر توابع هدف اول و دوم در بین تمام جواب‌های نامغلوب هستند.

متر فاصله (SM): برای محاسبه یکنواختی جواب‌های نامغلوب از این معیار استفاده می‌شود که با استفاده از فرمول

$$SM = \frac{\sum_{i=1}^{n} \mid \bar{d} - d_i \mid}{(n-1)\bar{d}},$$

محاسبه می‌شود، که در آن n تعداد جواب‌های نامغلوب به دست آمده توسط الگوریتم، d_i فاصله اقلیدسی بین دو جواب نامغلوب متوالی مرز پارتو و $\bar{d}$ میانگین این فاصله‌ها است. الگوریتم با مقادیر کوچک‌تر MID و SM عملکرد بهتری دارد.

متر پراکندگی (DM): این معیار برای محاسبه پراکندگی جواب‌های پارتو تعریف شده است که از رابطه

$$DM = \left[\sum_{i=1}^{m}(\min f_i - \max f_i)^2\right]^{1/2},$$

محاسبه می‌شود، که در آن $\min f_i$ و $\max f_i$ به ترتیب کم‌ترین و بیش‌ترین مقدار هر تابع هدف در بین تمام جواب‌های نامغلوب هستند.

گسترش جواب‌های نامغلوب (SNS): از این معیار برای محاسبه پراکندگی و تنوع جواب‌های پارتو استفاده می‌شود. این

[10] Mean Ideal Distance
[11] Spacing Metric
[12] Diversification Metric
[13] Spread of Non-dominance Solution

معیار از رابطه

$$SNS = \left[\frac{\sum_{i=1}^{n}(MID - C_i)^2}{n-1}\right]^{1/2},$$

به‌دست می‌آید، که در آن $C_i = \left[f_{1i}^2 + f_{2i}^2\right]^{1/2}$. الگوریتم با مقادیر بزرگتر DM و SNS دارای عملکرد بهتری است.

۳ مطالعه موردی

آلودگی زیست‌بوم‌ها به فلزات سنگین، یکی از مهم‌ترین مسائل محیط زیست است که زندگی گیاهان، جانوران و به ویژه انسان‌ها را تهدید می‌کند. در حال حاضر، بسیاری از رودخانه‌ها در معرض آلودگی فلزات سنگین ناشی از منابع طبیعی همچون هوازدگی و فرسایش سنگ و همچنین منابع انسانی از جمله پساب‌های شهری، صنعتی و کشاورزی قرار دارند. رسوبات به عنوان مخزن آلاینده‌ها، شاخص مهمی برای تعیین آلودگی آب محسوب می‌شوند. این مطالعه، درصدد تعیین مکان‌های مناسب نمونه‌برداری در طول سواحل رودخانه میوز، به منظور ارزیابی هر چه دقیق‌تر غلظت فلزات سنگین موجود در رسوبات ساحلی این رودخانه است. در این پژوهش، منطقه مورد مطالعه برای داده‌های میوز در نظر گرفته شده است که اولین بار توسط بارو و مک دونال (۱۹۹۸) ارائه شد. منطقه مورد نظر شامل یک دشت سیلابی در طول رودخانه میوز واقع در هلند به طول جغرافیایی ۵۲ درجه و ۰ دقیقه و ۵۲۸۸۵۲/۱۳ ثانیه شمالی و به عرض جغرافیایی ۴ درجه و ۴۶ دقیقه و ۵۲۲۸۰۰/۲۲ ثانیه شرقی است. در این مقاله، به منظور تعیین موقعیت‌های مناسب نمونه‌برداری، واریانس میانگین نمونه و مسافت طی‌شده برای نمونه‌برداری به عنوان توابع هدف مسئله فرض می‌شوند. برآورد مدل محور واریانس میانگین نمونه، به‌عنوان تابع هدف اول به صورت

$$\sigma_m^2 = \frac{2}{n}\sum_{i=1}^{n}\bar{\gamma}(\mathbf{x}_i, \mathcal{B}) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(\mathbf{x}_i - \mathbf{x}_j) - \bar{\gamma}(\mathcal{B}, \mathcal{B}),$$

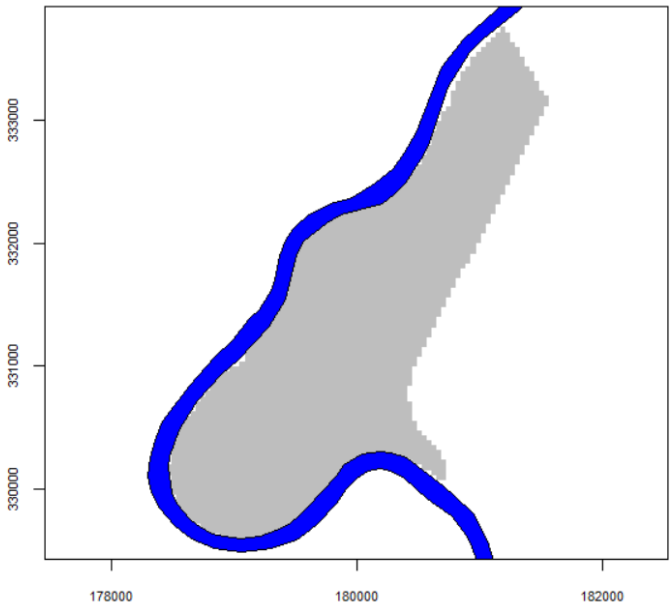
محاسبه می‌شود، که در آن $\mathbf{x}_i$ بردار موقعیت iامین نمونه، $\gamma(\mathbf{h})$ نیم‌تغییرنگار و $\mathcal{B}$ دامنه نمونه‌گیری است (وبستر و الیور ۱۹۹۰). همچنین، $\bar{\gamma}(\mathbf{x}_i, \mathcal{B})$ و $\bar{\gamma}(\mathcal{B}, \mathcal{B})$ به ترتیب، به صورت

$$\bar{\gamma}(\mathbf{x}_i, \mathcal{B}) = \int_{\mathbf{x}_k \in \mathcal{B}} \gamma(\mathbf{x}_i - \mathbf{x}_k) d\mathbf{x}_k,$$
$$\bar{\gamma}(\mathcal{B}, \mathcal{B}) = \int_{\mathbf{x}_k \in \mathcal{B}} \int_{\mathbf{x}_\ell \in \mathcal{B}} \gamma(\mathbf{x}_k - \mathbf{x}_\ell) d\mathbf{x}_\ell d\mathbf{x}_k,$$

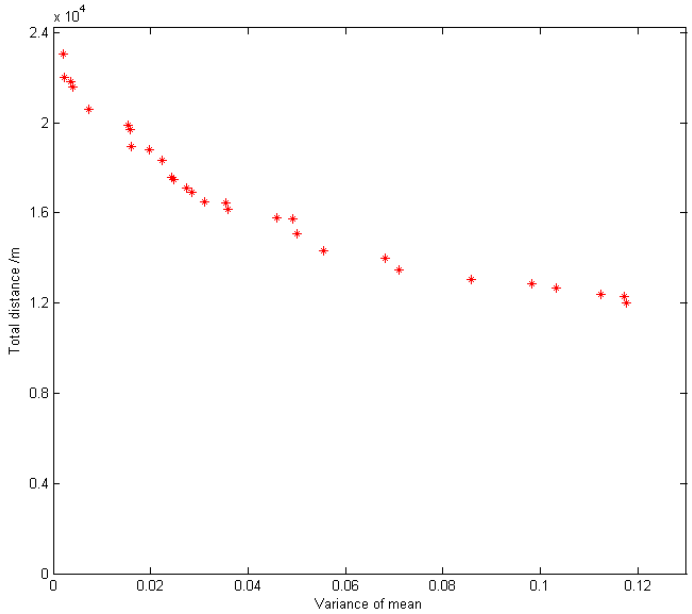
محاسبه می‌شوند. تابع هدف دوم، مسافت طی‌شده بین موقعیت‌های نمونه است که برای مینیمم‌سازی آن باید کوتاه‌ترین مسیر بین نقاط نمونه را به‌دست آورد. برای این منظور، از روش فروشنده دورگرد [14] (TSP) استفاده می‌شود (گوتین و پونن ، ۲۰۰۶). تابع هدف اول توسط طرح نمونه‌ای مینیمم می‌شود که بهترین پوشش فضایی را در سطح منطقه مورد مطالعه ایجاد کند، در حالی که چنین طرح نمونه‌ای تابع هدف دوم را ماکسیمم می‌کند و این بیانگر تقابل بین دو تابع هدف است. در این مطالعه، برای یافتن طرح نمونه بهینه، الگوریتم چندهدفه ازدحام ذرات برای داده‌های فضایی ارتقا یافته است. با توجه به ماهیت فضایی داده‌ها، ساختار همبستگی فضایی در تابع هدف واریانس میانگین نمونه لحاظ می‌شود. برای داده‌های میوز، متغیر مورد نمونه‌گیری دارای همبستگی فضایی با تغییرنگار کروی با دامنه ۸۹۷ متر، اثر قطعه‌ای ۰۵۱/۰ و ازاره ۵۹۱/۰ است و ۱۵۵ موقعیت برای نمونه‌برداری از خاک سطحی منطقه به منظور بررسی غلظت فلزات سنگین انتخاب می‌شود (پیسما،

[14] Travelling Salesman Problem

۲۰۱۹). در ادامه، نتایج حاصل از اجرای الگوریتم برای مسئله مطرح شده با ۲۰۰۰ تکرار ارائه می‌شود. برنامه‌ها روی یک کامپیوتر با واحد پردازش مرکزی (CPU)، ۳۰/۲ GHz با پنج هسته، حافظه دستیابی تصادفی (RAM)، GB۰۰/۸ و نرم‌افزار MATLAB b۲۰۱۳R اجرا شده‌اند.



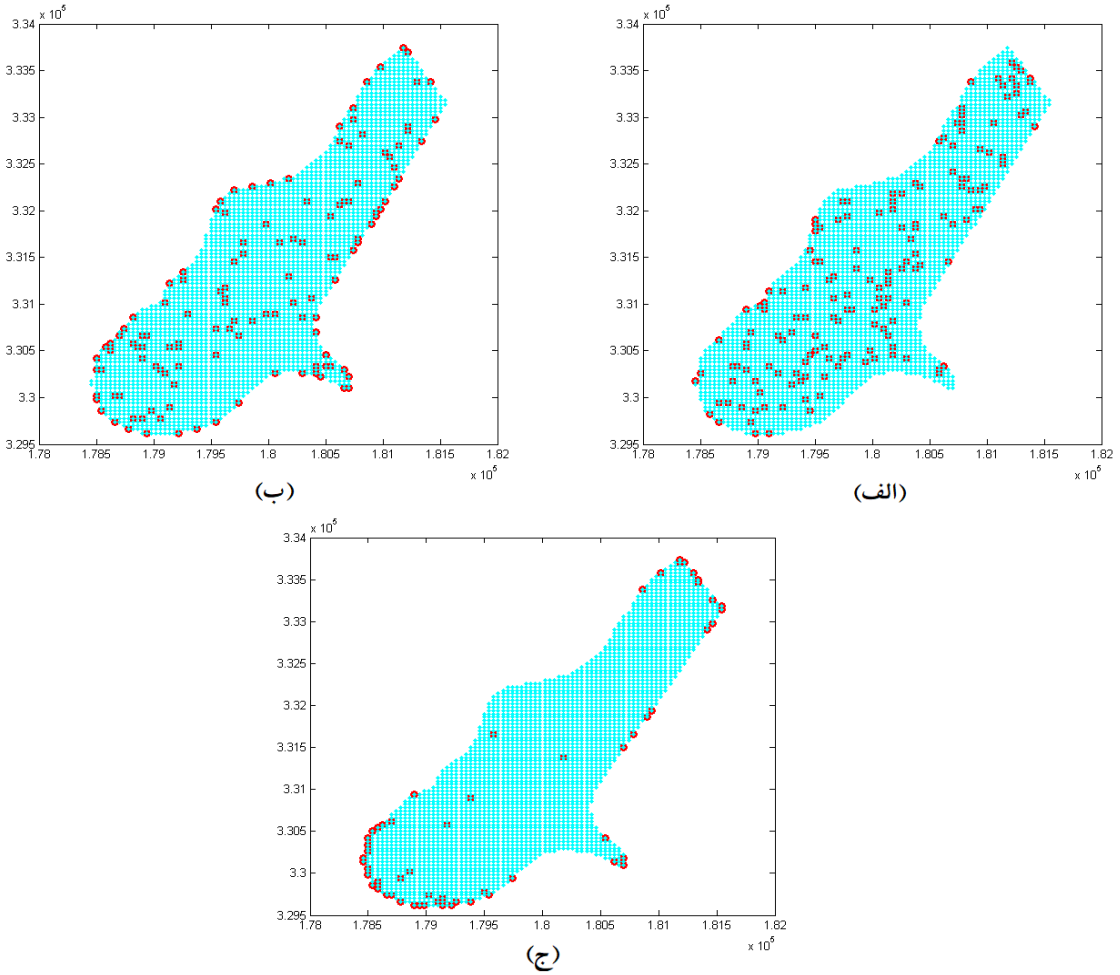
شکل ۳: منطقه مورد مطالعه در طول رودخانه میوز



شکل ۴: مرز پارتو تقریبی به‌دست آمده توسط الگوریتم

شکل‌های ۳ و ۴ به ترتیب منطقه مورد مطالعه و مرز پارتو تقریبی بدست آمده توسط الگوریتم را نشان می‌دهند. همان‌طور که از شکل ۴ ملاحظه می‌شود، مرز پارتو به‌دست آمده توسط الگوریتم، پراکندگی نسبتا خوبی دارد. شکل ۵ نیز

برخی از موقعیت‌های نمونه‌برداری به‌دست آمده توسط الگوریتم را نشان می‌دهد. در جدول ۱ مقادیر توابع هدف در برخی



شکل ۵: برخی از موقعیت‌های نمونه‌برداری به‌دست آمده توسط الگوریتم

از طرح‌های نمونه محاسبه شده است. همان‌طور که ملاحظه می‌شود، با افزایش مقدار واریانس، مسافت طی‌شده کاهش می‌یابد و برعکس، که این خود گویای تقابل بین دو تابع هدف است.

جدول ۱: مقادیر توابع هدف در برخی از طرح‌های نمونه (با اندازه نمونه ۱۵۵)

طرح نمونه	(الف)	(ب)	(ج)
واریانس میانگین نمونه	۰۰۲۲/۰	۰۱۱۹/۰	۱۰۹۸/۰
مسافت طی‌شده (متر)	۲۳۵۳۳	۲۱۴۳۹	۱۳۲۷۲

در جدول ۲ مقادیر هر یک از معیارهای ارزیابی عملکرد الگوریتم ارائه شده است. ملاحظه می‌شود معیارهای MID و SM مقادیر کم و معیارهای DM و SNS مقادیر بالایی دارند که نشان‌دهنده پوشش و پراکندگی مناسب مرز پارتو تقریبی به‌دست آمده و عملکرد مناسب الگوریتم است.

جدول ۲: مقادیر معیارهای ارزیابی عملکرد الگوریتم (با اندازه نمونه ۱۵۵)

الگوریتم	SM	MID	DM	SNS
$MOPSO$	۸۴۶۴/۰	۶۵۴۹/۰	۱۰۲۶۱	۱۸۹۱۷

جدول ۳ ماکسیمم و مینیمم مقدار توابع هدف را برای اندازه نمونه‌های مختلف نشان می‌دهد. همان‌طور که انتظار می‌رود، با افزایش اندازه نمونه، واریانس میانگین نمونه کم‌تر و مسافت طی‌شده بیشتر می‌شود.

جدول ۳: ماکسیمم و مینیمم مقدار توابع هدف برای اندازه نمونه‌های متفاوت

اندازه نمونه	واریانس میانگین نمونه		مسافت طی‌شده (متر)	
	مینیمم	ماکسیمم	مینیمم	ماکسیمم
۵۰	۰/۰۰۴۸	۰/۱۲۰۹	۱۰۵۷۸	۱۴۹۲۱
۱۰۰	۰/۰۰۳۳	۰/۱۱۰۲	۱۱۷۰۴	۱۹۳۹۴
۱۵۵	۰/۰۰۲۲	۰/۱۰۹۸	۱۳۲۷۲	۲۳۵۳۳

بحث و نتیجه‌گیری

یک مسئله بهینه‌سازی دوهدفه برای نمونه‌برداری از رسوبات ساحلی رودخانه میوز جهت ارزیابی میزان آلودگی آن به فلزات سنگین در نظر گرفته شد. واریانس میانگین نمونه و مسافت طی‌شده بین موقعیت‌های نمونه به عنوان توابع هدف مسئله فرض شدند. در این راستا، الگوریتم MOPSO برای یافتن طرح نمونه بهینه فضایی ارتقا داده شد. برای این منظور ساختار همبستگی فضایی داده‌ها در الگوریتم لحاظ شد. نتایج حاصل نشان‌دهنده عملکرد مناسب الگوریتم MOPSO ارتقا یافته در این مسئله از نمونه‌گیری فضایی است.

مراجع

بیداربخت، ح.، محمدزاده، م.، محمدی، الف. و مرید، س. (۱۳۹۳)، طراحی نمونه‌گیری فضایی بهینه با ملاک آنتروپی، *دومین کارگاه آموزشی اندازه‌های اطلاعات و کاربردهای آن*، ۸ـ۹ بهمن ۱۳۹۳، مشهد، ایران.

لطفیان، الف و محمدزاده، م.، (۱۳۹۷)، نمونه‌گیری فضایی با استفاده از بهینه‌سازی چندهدفه استوار، *چهاردهمین کنفرانس آمار ایران*، شاهرود، ایران، ۵۲۷ـ۵۳۴.

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

Burrough, P. A., McDonnell, R. A. (1998), *Principles of Geographical Information Systems*, 2nd Edition. Oxford University Press.

Coello, C. A. C., Pulido, G. T., and Lechuga, M. S. (2004), Handling Multiple Objectives with Particle Swarm Optimization, *IEEE Transactions on evolutionary computation*, **8**, 256-279.

Ehrgott, M., Figueira, J., and Greco, S. (Eds.). (2005), *Multiple Criteria Decision Analysis: State of*

the Art Surveys, Springer.

Fuentes, M., Chaudhuri, A., and Holland, D. M. (2007), Bayesian Entropy for Spatial Sampling Design of Environmental Data, *Journal Environmental Ecological Statistics*, **14**, 323-340.

Gomez-Cardenas, C., Sbartai, Z. M., Balayssac, J. P., Garnier, V., and Breysse, D. (2015), New Optimization Algorithm for Optimal Spatial Sampling During Non-Destructive Testing of Concrete Structures, *Engineering Structures*, **88**, 92-99.

Greco, S., Ehrgott, M., and Figueira, J.R. (Eds). (2005), *Multiple Criteria Decision Analysis*, State of the Art Surveys, Springer, Boston.

Groenigen, J. W., and Stein, A. (1998), Constrained Optimization of Spatial Sampling Using Continuous Simulated Annealing, *Journal of Environmental Quality*, **27**, 1078-1086.

Guedes, L. P. C., Riberio JR, P. J., Piedade, S. M. D., and UribeOpazo, M. A. (2011), Optimization of Spatial Sample Configurations using Hybrid Genetic Algorithm and Simulated Annealing, *Chilean Journal of Statistics*, **2**, 39-50.

Gutin G., and Punnen A. P. (2006), *The Traveling Salesman Problem and its Variations*, Springer Science and Business Media.

Lark, R. M. (2002), Optimized Spatial Sampling of Soil for Estimation of the Variogram by Maximum Likelihood, *Geoderma*, **105**, 49-80.

Lark, R. M. (2016), Multi-objective Optimization of Spatial Sampling, *Spatial Statistics*, **18**, 412-430.

Maghsoudlou, H., Afshar-Nadjafi, B., and Niaki, S. T. A. (2016), A Multi-objective Invasive Weeds Optimization Algorithm for Solving Multi-skill Multi-mode Resource Constrained Project Scheduling Problem, *Computers & Chemical Engineering*, **88**, 157-169.

Marchant, B. P., McBratney, A. B., Lark, R. M., and Minasny, B. (2013), Optimized Multi-Phase Sampling for Soil Remediation Surveys, *Spatial Statistics*, **4**, 1–13.

McBratney, A. B., and Webster, R. (1983), Optimal Interpolation and Isarithmic Mapping of Soil Properties, *European Journal of Soil Science*, **34**, 137-162.

McBratney, A. B., Webster, R., and Burgess, T. M. (1981), The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalized Variables-I: Theory and Method, *Computers and Geosciences*, **7**, 331-334.

Muller, P., Sanso, B., and De lorio, M. (2004), Optimal Bayesian Design by Inhomogeneous Markov Chain Simulation, *Journal of the American Statistical Association*, **99**, 788-798.

Parvez, R. (2014), Selection in Evolutionary Algorithm. https://www.slideshare.net/riyadparvez/selection-in-evolutionary-algorithm. Accessed 10 May 2018.

Pebesma, E. (2019). The Meuse Data Set: A Brief Tutorial for the gstat R Package.

Sacks, J., and Schiller, S. (1988), Spatial Designs. In *Statistical Decision Theory and Related Topics IV*, Gupta, S. S., Berger, J. O. (eds), Springer, New York, 385-399.

Wang, J. H., Ge, Y., Heuvelink, G. B. M., and Zhou, C. H. (2014), Spatial Sampling Design for Estimating Regional GPP with Spatial Heterogeneities, *IEEE Geoscience Remote Sensing Letters*, **11**, 539-543.

Webster, R., and Oliver, M. A. (1990), *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press, Oxford.

Zhou, A., Qu, B. Y., Li, H., Zhao, S. Z., Suganthan, P. N., and Zhang, Q. (2011), Multiobjective Evolutionary Algorithms: A Survey of the State of the Art, *Swarm and Evolutionary Computation*, **1**, 32-49.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

مدل‌های توام نقشه‌بندی بیماری‌ها

یدالله محرابی[۱]، امیر کاوسی [۱]، بهزاد مهکی [۲]، پریسا پورنظری[۱۲]
[۱]گروه اپیدمیولوژی، دانشکده بهداشت و ایمنی، دانشگاه علوم پزشکی شهید بهشتی
[۲]گروه آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی کرمانشاه

چکیده: نقشه بندی بیماری شامل مجموعه ای از آماری برای تهیه‌ی نقشه ی برآوردهای میزان های بیماری ها است. بیشتر مطالعات نقشه بندی بیماری ها بر مدل بندی یک بیماری واحد متمرکز بوده است اما بیماری های زیادی وجود دارند که عوامل خطر مشترکی داشته، در عمل همبسته بوده و استفاده از مدل‌هایی که اطلاعات مربوط به چند بیماری مرتبط را ادغام می‌کنند، می‌تواند مفید باشد. در سال های اخیر چندین مدل برای تحلیل توأم چند بیماری معرفی شده و مطالعات متعددی انجام گرفته که در آن از مدل های نقشه‌بندی چندمتغیره ی بیماری ها استفاده شده است. با این وجود تاکنون به صورت جامع، روش های تحلیل توأم نقشه‌بندی بیماری ها مورد بررسی و مقایسه قرار نگرفته‌اند. در مطالعه حاضر مهم ترین مدل های نقشه بندی چندمتغیره ی بیماری ها معرفی گردیده و مبانی تئوری، کاربردها، محاسن و معایب هر یک به اختصار بیان می‌شوند.

واژه‌های کلیدی: نقشه بندی بیماری ها، مدل های توام، مدل های چند متغیره.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62H30 .

۱ مقدمه

نقشه بندی بیماری شامل مجموعه ای از روش های آماری است که به تهیه نقشه های دقیق براساس برآوردهایی از میزان‌های بیماری ها و در نهایت برآورد توزیع جغرافیایی بیماری مورد نظر منجر می شوند. اهداف اصلی نقشه بندی بیماری عبارتند از: ۱. تشریح تغییرات مکانی یا زمانی در میزان های بیماری یا خطر به منظور تدوین فرضیات علت‌شناسی ۲. ارزیابی خطر و تعیین نواحی دارای خطر بالا که نیازمند اعمال مداخله هستند. ۳. استفاده از آن‌ها به منظور تخصیص بهتر منابع و خدمات ۴. ساخت اطلس بیماری ها (لاوسون و همکاران، ۲۰۰۳، ۲۰۰۰).

[۱]نام و ایمیل ارائه دهنده مقاله: پریسا پورنظری، p.pournazari93@gmail.com

در سالیان اخیر تلاش های فراوانی برای نقشه بندی میزان های بیماری ها صورت گرفته است. یکی از مهم ترین مشکلات در این خصوص، انتخاب معیار مناسب برای نقشه بندی است. نقشه بندی خام میزان های بیماری ها ممکن است کاملا گمراه کننده باشد. به منظور حل این مشکل، معمولا از یکی از معیارهای خطرات نسبی در هر ناحیه که با نسبت مرگ ومیر یا نسبت بروز استانداردشده ی اختصاصی ناحیه SMR[1] یا SIR[2] اندازه گیری می شود، استفاده می گردد. گرچه استفاده از این برآوردها در نقشه بندی بیماری ها بسیار رایج است اما نقاط ضعف بسیاری در استفاده از این شاخص ها وجود دارد که برای حل آن می توان از هموارسازی های آماری برای غلبه بر این مشکلات استفاده کرد (لاوسون و همکاران، ۲۰۰۰). تاکنون چندین مدل برای نقشه بندی تک متغیره ی بیماری ها مطرح شده اند که مهم ترین آن عبارتند از:

۱. مدل های هموارسازی که در جهت هموارسازی نوسانات بر پایه توابعی از داده ها در نواحی مورد نظر تلاش می کنند. رایج ترین روش در این کلاس، روش هموارسازی کرنل نادارایا واتسون است (سیمونوف، ۲۰۱۲). ۲. مدل های بیز خطی که مبتنی بر توابع خطی از SMR یا SIR هستند (مارشال، ۱۹۹۱). ۳. مدل های بیز کامل BYM [3] که از مهم ترین مدل های نقشه بندی بیماری ها هستند. در این مدل پارامتر خطر نسبی بیماری به سه مؤلفه روند، تغییرات همبسته مکانی و تغییرات ناهمبسته تجزیه می شود (بساگ و همکاران، ۱۹۹۱). ۴. مدل های بیز تجربی که مشابه روش های بیزی بوده اما توزیع پیشین خطرات نسبی از داده های مشاهده شده برآورد می شود. این روش ها بینابین روش حداکثر درست نمایی و روش های بیز کامل بوده و فرض می کنند که خطرات نسبی از توزیع های خاصی پیروی می کنند که با مجموعه ای از پارامترها مشخص می شوند. این پارامترها را می توان از داده ها با حداکثر نمودن درستنمایی حاشیه ای برآورد کرد (خوشکار و همکاران، ۲۰۱۵). ۶. مدل های آمیخته که کاملا از مدل های پیشین که در آن ها پراکندگی و تغییرات در خطر با مخلوطی از چند جزء مدل بندی می شود متفاوت است (روهنینگ، ۱۹۹۹). ۷.مدل های چندسطحی که به داده هایی برازش می یابند که دارای سطوح خوشه بندی در ساختار خود هستند. در نقشه‌بندی بیماری ها و کاربردهای جغرافیایی، این سطوح، سطوح متفاوت جغرافیایی در نظر گرفته می شوند (رستاقی و همکاران، ۲۰۱۵).

با وجود پیشرفت های سریع روش های آماری و محاسباتی، گسترش تحلیل های نقشه بندی بیماری ها اغلب بر مدل بندی یک بیماری واحد متمرکز بوده است حال آن که بیماری های زیادی هستند که عوامل خطر مشترکی دارند و اگر الگوهای مشترکی از پراکندگی جغرافیایی بیماری های مرتبط در تحلیل توأم مشخص گردد آنگاه نسبت به حالت تحلیل جداگانه، شواهد متقاعدکننده تری از خوشه بندی واقعی در خطر به دست می آید. اگر داده های چند بیماری یا رویداد بهداشتی که الگوهای جغرافیایی مشترکی دارند در دسترس باشد مدل بندی توأم خطرات می تواند منجر به بهبود برآوردها نسبت به تحلیل جداگانه ی هر رویداد گردد. حتی درصورتی که یک رویداد به عنوان رویداد هدف مدنظر باشد، ورود داده های دیگر رویداد های مرتبط در دسترس در تحلیل مفید است (نورهلد و همکاران، ۲۰۰۱).

در سال های اخیر، توجه به تحلیل توأم چند بیماری که بالقوه مرتبط هستند افزایش یافته، در این راستا، چندین مدل کلاسیک و بیزی، معرف شده و مطالعات موردی متعددی انجام گرفته است که در آن از کابردها و مدل های نقشه بندی چندمتغیره ی بیماری ها استفاده شده است با این وجود جز در چند مطالعه اندک و به صورت محدود، روش های متعدد تحلیل توأم نقشه های بیماری ها با هم مقایسه نشده و شرایط کلی آن ها مورد بررسی قرار نگرفته است. با توجه به اهمیت موضوع و کاربرد رو به گسترش نقشه بندی بیماری ها، در این مقاله مهم ترین مدل های نقشه بندی توأم و چندمتغیره ی بیماری ها معرفی گردیده و مبانی تئوری، کاربردها، محاسن و معایب هر یک به اختصار بیان می گردد.

[1] Standardized Mortality/Morbidity Ratio
[2] Standardized Incidence Ratio
[3] Besag, York and Mollie

۲ مدل های نقشه بندی چندمتغیره ی بیماری ها

۱.۲ مدل های رگرسیون اکولوژیک

این مدل ها توسط برناردینلی و همکاران معرفی شده اند که در آن میزان های یک بیماری ثانویه به عنوان جانشین عوامل خطر در مدل رگرسیونی وارد می شود. با استفاده از این رویکرد، یک چارچوب نقشه بندی چندمتغیره ی بیزی بیماری ها برای بررسی های اکولوژیک و مکانی الگوهای میزان های بیماری و ارتباطات بین بیماری ها و مرگ و میر با عوامل خطر اجتماعی، اقتصادی و محیطی به دست می آید. در این روش، ابتدا مدل BYM برای داشتن عوامل خطر در سطوح مکانی x_i بسط داده می شود.

$$\log(\theta_i) = \alpha + u_i + \nu_i + x_i\beta$$

که در آن، θ معرف خطر نسبی، u_i و ν_i به ترتیب بیانگر اثرات ناهمگنی همبسته مکانی و ناهمگنی ناهمبسته مکانی در ناحیه ی i و β بیانگر اثر عوامل خطر است. هدف از ورود اطلاعات مواجهه x_i ، تشریح پراکندگی مکانی است. سپس رویکرد رگرسیون اکولوژیک با در نظر گرفتن خطای نمونه گیری و خطای اندازه گیری عامل خطر در مدل، تعدیل می شود بدین ترتیب که اگر z_i نشان دهنده ی عامل خطر واقعی اما مشاهده نشده باشد و از آن به جای x_i استفاده می کنیم لگاریتم خطر نسبی عبارت خواهد بود از:

$$\log(\theta_{i۱}) = \alpha + u_i + \nu_i + z_i\gamma$$

z_i را به عنوان جانشین مواجهه به بیماری ارتباط می دهیم. این کار را می توان با فرض کردن مدل خطای اندازه گیری $\log(\theta_{i۲})|z_i \sim N(z_i, \omega)$ انجام داد که $\log(\theta_{i۲})$ لگاریتم خطر نسبی بیماری جانشین و ω دقت خطای اندازه گیری است. این مدل با فرض $O_{i۲}|\log(\theta_{i۲}) \sim poisson(E_{i۲}\theta_{i۲})$ تکمیل می شود که $O_{i۲}$ و $E_{i۲}$ به ترتیب تعداد موارد مشاهده و مورد انتظار بیماری می باشند (حداد و همکاران، ۲۰۱۵).

۲.۲ مدل های چندسطحی

لیلاند و همکاران، تحلیل مکانی توأم دو بیماری را براساس مدل چندسطحی ارائه نموده اند که در آن اجزاء نامعلوم واریانس از طریق روش های درست نمایی برآورد می شوند. در این مدل برآورد میزان های رویدادها در نواحی کوچک از طریق مدل بندی مکانی داده ها و افزودن اطلاعات بیشتر به هر ناحیه بهبود می‌یابد. استفاده از مدل های مکانی برای پیش بینی همزمان بیش از یک رویداد مد نظر قرار می گیرد. این کار با در نظر گرفتن مدل مکانی به شکل چندسطحی و به دنبال آن دربرگرفتن ساختار چندمتغیره ی داده ها انجام می شود. برآوردها نیز با استفاده از حداقل مربعات تعمیم یافته ی مکرر به دست می آیند (لانگفورد و. همکاران، ۱۹۹۹؛ لیلند و همکاران، ۲۰۰۰).

۳.۲ مدل های نرمال دو متغیره (بیزی و غیربیزی)

چندین مدل با استفاده از توزیع های نرمال برای نقشه بندی توأم دو بیماری معرفی شده اند که برخی از آن ها بیزی و برخی دیگر غیربیزی هستند. مولی و دسوزا (مولایی، ۱۹۹۰) رویکردی برای تحلیل توأم تعداد موارد ناحیه ای در نواحی، مبتنی بر پیشین نرمال دومتغیره برای پارامترهای لگاریتم خطر نسبی متناظر در هر ناحیه اتخاذ کردند اما این مدل از همبستگی مکانی ممکن خطرات نسبی در نواحی، که توسط لنگفورد و همکاران (لانگفورد و. همکاران، ۱۹۹۹) در نظر گرفته شده است، چشم پوشی می کند. دیگر محققین توزیع اتورگرسیو شرطی CAR[۴] را برای دو بیماری بسط داده اند یا مستقیما از توزیع

[4] Conditional Autoregressive

های چندمتغیره ی اتورگرسیو شرطی استفاده کرده اند. کیم و همکاران نیز جایگزینی بیزی را برای تحلیل مکانی توأم بیماری ها پیشنهاد نموده اند که محدود به مدل بندی دو بیماری است (کیم و همکاران، ۲۰۰۱).

۴.۲ مدل های میدان های تصادفی مارکف نرمال چندمتغیره

گلفند و وناتسو مدل بندی توأم با استفاده از میدان‌های تصادفی مارکف نرمال چندمتغیره را توصیه کرده اند. مدل های میدان های تصادفی نرمال با جایگزینی توزیع شرطی نرمال چندمتغیره به جای توزیع نرمال تک متغیره برای بردار لگاریتم خطرات نسبی در ناحیه i به مدل های میدان های تصادفی نرمال چندمتغیره بسط داده شده و مدل BYM چندمتغیره ساخته شده است (جین و همکاران، ۲۰۰۵؛ گلفند و همکاران، ۲۰۰۳).

۵.۲ مدل های جزء مشتر SC

نور_هلد و بست مدل جزء مشترک را برای تحلیل توأم مکانی دو یا چند بیماری معرفی کردند. از این مدل برای ارزیابی همبستگی مکانی بین دو بیماری، استفاده می شود. ایده کلیدی این مدل جداکردن و تقسیم فضای پایه ی خطر به جزء مشترک دو بیماری و جزء اختصاصی هر کدام از بیماری ها است. جزء مشترک به عنوان جانشینی برای عوامل خطر مشاهده نشده که ساختار مکانی را نشان می دهند و برای هر دو بیماری مشترکند، به کار می رود. جزء اختصاصی هر بیماری نیز تنها معرف و جانشین عوامل خطر پراکنده در سطح مکانی است که مختص همان بیماری است. همچنین وزن نسبی هر جزء مشترک برای بیماری های مرتبط در مدل وارد می شود. مدل به دست آمده قابلیت تعیین بزرگی پراکندگی را از طریق الگوهای جغرافیایی در فضای بیماری ها فراهم می آورد. تاکنون چندین شکل از مدل SC [۵] معرفی شده است که در ادامه مهم ترین آن‌ها معرفی می گردد (نورهلد و همکاران، ۲۰۰۱).

۱.۵.۲ مدل SC متقارن برای دو بیماری

به منظور مدل بندی توأم دو بیماری در ناحیه ی ام i $(i = ۱, \ldots, n)$ با موارد مشاهده شده O_{ik} و تعداد موارد مورد انتظار E_{ik} ، مدل

$$O_{ik} \sim Poisson(E_{ik}\theta_{ik}) \quad k = ۱, ۲$$

به کار می رود و تجزیه ی زیر برای لگاریتم خطرات نسبی در نظر گرفته می شود:

$$\log(\theta_{i۱}) = \alpha_۱ + \lambda_۱\delta + \phi_{i۱}$$
$$\log(\theta_{i۲}) = \alpha_۲ + \lambda_۲\delta + \phi_{i۲}$$

که در آن $\theta_{i۱}$ و $\theta_{i۲}$ به ترتیب معرف خطر نسبی دو بیماری در ناحیه ی ام i، α_k معرف عرض از مبدأ اختصاصی بیماری ام k و λ_i معرف عامل خطر مشترک دو بیماری و $\phi_{i۱}$ و $\phi_{i۲}$ نیز به ترتیب عوامل خطر اختصاصی بیماری نخست و دوم در ناحیه ی ام i هستند. δ نیز به منظور فراهم آوردن شیب خط متفاوت ارتباط عامل مشترک با بیماری ام k به مدل وارد شده است (نورهلد و همکاران، ۲۰۰۱).

[5] Shared Component

۲.۵.۲ مدل SC نامتقارن برای دو بیماری

در این مدل یک جزء مشترک مرتبط با هر دو بیماری مرتبط λ_i و یک جزء اختصاصی ϕ_{i1} و مرتبط با تنها یکی از آن ها وجود دارد.

$$\log(\theta_{i1}) = \alpha_1 + \lambda_1\delta + \phi_{i1}$$
$$\log(\theta_{i2}) = \alpha_2 + \lambda_1/\delta$$

این فرمول بندی تا اندازه ای از فرمول بندی متقارن برای مدل SC متفاوت است. مدل متقارن دارای ویژگی جالب توجه تقارن است اما معمولا عدم وجود فرض پیشین قوی درباره ی پارامترهای دقت مشکلاتی را در خصوص قابلیت شناسایی عوامل خطر ایجاد می کند. لگاریتم خطر نسبی مشترک، اشتراکات دو بیماری در هر ناحیه را اندازه گیری می کند (هلد و همکاران، ۲۰۰۵).

۳.۵.۲ مدل SC چندمتغیره

تعمیم مدل SC به بیش از دو بیماری سرراست و آسان است. چندین عامل خطر می توانند برای چند بیماری یا تنها برای یک بیماری به اشتراک گذاشته شوند. فرض می‌شود لگاریتم خطر نسبی برای ناحیه i و بیماری k دارای میانگین برابر با جمع عرض از مبدأ و جمع وزنی عوامل خطر مربوطه است. برای هر عامل خطر تصادفی مشترک یا اختصاصی با توزیع نرمال فرض می شود که بردار $[\log(\delta_{l,1}), \ldots, \log(\delta_{l,n_l})]'$ دارای توزیع نرمال چندمتغیره با میانگین صفر و ماتریس واریانس کواریانس $\sum$ است و n_l بیانگر تعداد بیماری های مرتبط با عامل خطر است (هلد و همکاران، ۲۰۰۵؛ دنینگ و همکاران، ۲۰۰۸). مهکی و همکاران (۲۰۱۱) تحلیل مکانی ۷ نوع سرطان با ۵ عامل خطر مشترک را در ۳۰ استان ایران با استفاده از این مدل به شکل زیر انجام داده‌اند:

$$\begin{aligned}
\log(\theta_{i1}) &= \alpha_1 + \lambda_{1i}\delta_{1,1} + \lambda_{2i}\delta_{2,1} + \lambda_{3i}\delta_{3,1} + \lambda_{4i}\delta_{4,l} + \varepsilon_{i1}\\
\log(\theta_{i2}) &= \alpha_2 + \lambda_{1i}\delta_{1,2} + \lambda_{3i}\delta_{3,2} + \lambda_{4i}\delta_{4,2} + \varepsilon_{i2}\\
\log(\theta_{i3}) &= \alpha_3 + \lambda_{1i}\delta_{1,3} + \lambda_{4i}\delta_{4,3} + \varepsilon_{i3}\\
\log(\theta_{i4}) &= \alpha_4 + \lambda_{2i}\delta_{2,2} + \lambda_{4i}\delta_{4,4} + \lambda_{5i}\delta_{5,1} + \varepsilon_{i4}\\
\log(\theta_{i5}) &= \alpha_5 + \lambda_{1i}\delta_{1,4} + \lambda_{4i}\delta_{4,5} + \varepsilon_{i5}\\
\log(\theta_{i6}) &= \alpha_6 + \lambda_{2i}\delta_{2,3} + \lambda_{4i}\delta_{4,6} + \varepsilon_{i,6}\\
\log(\theta_{i7}) &= \alpha_7 + \lambda_{2i}\delta_{2,4} + \lambda_{4i}\delta_{4,7} + \lambda_{5i}\delta_{5,2} + \varepsilon_{i,7}
\end{aligned}$$

که در آن $\theta_{i1}, \ldots, \theta_{i7}$ به ترتیب معرف خطرات نسبی سرطان های مری، معده، مثانه، روده بزرگ، ریه، پروستات و پستان است. $\lambda_{1i}, \ldots, \lambda_{5i}$ نیز پنج جزء مشترک به ترتیب جانشین عوامل خطر مصرف سیگار، اضافه وزن یا چاقی، معرف مصرف کم میوه و سبزیجات، شاخص توسعه انسانی و فعالیت فیزیکی کم هستند. پارامترهای مقیاس نامعلوم δ برای هر جزء، امکان فراهم آوردن شیب خطر متفاوت برای بیماری ها مرتبط با آن جزء را فراهم می آورند. ε_{ik} نیز اثرات ناهمگنی اختصاصی بیماری ها بوده که بیش پراکندگی پواسون تشریح نشده در مدل را در نظر می گیرد (مهکی و همکاران، ۲۰۱۱).

۶.۲ مدل مرگ و میر نسبی PM

این مدل هنگامی استفاده می شود که تنها تعداد مرگ ناشی از بیماری ها در دسترس بوده و تعداد جمعیت در معرض خطر نامشخص باشد. در این حالت می توان توزیع تعداد مرگ ها برای بیماری مورد نظر نسبت به کل مرگ ها را به دست آورد. به منظور فرمول بندی این مدل، ابتدا فرض می شود $O_{ik} \sim poisson(\lambda_{ik})$ که اندیس $i = ۱, \ldots, n$ معرف نواحی مورد مطالعه و اندیس $k = ۱, ۲$ بیانگر دو بیماری مورد نظر می باشند. با این فرض که $\log(\lambda_{ki}) = \alpha_{۰k} + \alpha_{۱k} x_i$ به گونه ای که مدل فاقد اثرات تصادفی است و x_i ها نمایانگر عوامل خطر در سطوح ناحیه ای مرتبط با دو بیماری از طریق خطرات نسبی مکانی $\exp(\alpha_{۱۱})$ $\exp(\alpha_{۱۲})$ هستند. اگر مجموع تعداد موارد دو بیماری در ناحیه i را با $M_i = O_{i۱} + O_{i۲}$ تعریف کنیم آنگاه $O_{i۱}|M_i = m_i \sim Bin(m_i, \gamma_i)$ که:

$$\gamma_i = \frac{\exp\{(\alpha_{۰۱} - \alpha_{۰۲}) + (\alpha_{۱۱} - \alpha_{۱۲}) x_i\}}{۱ + \exp\{(\alpha_{۰۱} - \alpha_{۰۲}) + (\alpha_{۱۱} - \alpha_{۱۲}) x_i\}}$$

بنابراین می توان داده ها را با مدل لجستیک به فرم $logit(\gamma_i) = \beta_{۰} + \beta_{۱} x_i$ مدل بندی نمود که پارامتر $\beta_{۱}$ را می‌توان به عنوان تفاوت $\alpha_{۱۱} - \alpha_{۲۱}$ تفسیر نمود. بدین ترتیب استنباط می تواند براساس تفاوت در لگاریتم خطرات نسبی بدون داشتن جامعه ی در معرض خطر انجام شود. همچنین توجه شود که هنگامی که بیماری دوم مرتبط با عامل خطر x نیست داریم $\beta_{۱} = \alpha_{۱۱}$ (دابنی و همکاران، ۲۰۰۵).

۷.۲ مدل لجستیک چندگانه PL

این مدل به عنوان بسط مدل PM برای زمانی که بیش از دو بیماری مورد تحلیل قرار می گیرد معرفی شده است که در آن یک بیماری به عنوان گروه پایه و مرجع تعریف شده و برای هر پیشگو، فرمول و دستور مدل SC به کار برده می شود. فرض می شود که $O_{ij} = (O_{ij۱}, \ldots, O_{ijk}, \ldots, O_{ijK})'$ از توزیع $multi(m_{ig}, \pi_{ij} = (\pi_{ij۱}, \ldots, \pi_{ijk}, \ldots, \pi_{ijK}))'$ پیروی می‌کند که $m_{ij} = \sum_{k=۱}^{K} O_{ijk}$ و $\sum_{k=۱}^{K} \pi_{ijk} = ۱$. مدل PL را به عنوان بسط بیش از دو گروه مدل PM در نظر می گیریم که هر گروه با احتمالی به شکل $\pi_{ijk} = \phi_{ijk} / \sum_{r=۱}^{K} \phi_{ijr}$ مدل بندی می شود و هر لگاریتم شانس، به عرض ازمبدأ اختصاصی بیماری (α_k^* معرف تفاوت کلی بین بیماری امk و بیماری مرجع امK)، عبارت ساختاریافته ی زمان با سن و بیماری (α_{jk}^* معرف تفاوت کلی بین بیماری امk و بیماری مرجع امK)، اثرات ساختاریافته و ساختارنیافته ی مکانی (u_{ik}^* و ν_{ik}^* به ترتیب معرف تفاوت بین عبارات ساختاریافته و ساختارنیافته ی فضایی بین بیماری امk و بیماری مرجع) تجزیه می شود:

$$\log(\phi_{ijk}) = \alpha_k^* + \alpha_{jk}^* + u_{ik}^* + \nu_{ik}^*$$

در نهایت، از مدل SC به منظور تجزیه عبارات ساختاریافته ی مکانی u_{ik}^* به اثرات مشترک و اختصاصی بیماری استفاده می‌شود.

$$u_{i۲}^* = ua_i \delta_{۲} + up_{i۲} \qquad u_{i۱}^* = ua_i \delta_{۱} + up_{i۱}$$

که در آن ua_i جزء ساختاریافته ی مکانی مشترک و $up_{i۱}$ و $up_{i۲}$ اجزاء ساختاریافته ی مکانی اختصاصی بیماری هستند که هدف نهایی تمرکز بر برآورد اثرات این اجزاء به عنوان متغیرهای پنهان بیانگر عوامل خطر اختصاصی بیماری است (درسی ، ۲۰۰۷؛ نصرازادانی و همکاران، ۲۰۱۸)

۸.۲ مدل های مکانی_زمانی

بسیاری از کاربردهای نقشه بندی بیماری ها مانند تعیین پراکندگی مکانی خطر بیماری ها یا بررسی خوشه بندی، معمولا محدود به یک دوره ی زمانی هستند اما در پژوهش های پزشکی و بهداشتی، بسیاری از داده ها در چندین دوره ی زمانی و به مدت چند سال گردآوری می شوند. در این صورت می توان تحلیل نقشه های بیماری ها که شامل بعد زمانی است را در نظر گرفت. در سالیان اخیر، پیشرفت قابل توجهی در مدل های آماری و تکنیک های تحلیل داده های مکانی_زمانی رخ داده و مدل های مکانی به مدل های مختلف مکانی_زمانی میزان های بیماری ها بسط داده شده است (ریچاردسون و همکاران، ۲۰۰۶؛ احمدی پناه و همکاران، ۲۰۱۸؛ مهکی و همکاران، ۲۰۱۸؛ راعی و همکاران، ۲۰۱۸).

همانند تحلیل های مکانی، در تحلیل های مکانی_زمانی نیز مدل های توأم و چندمتغیره نسبت به مدل های تک متغیره برتری ها و محاسن متعددی مانند بهبود دقت برآوردها الگوی بیماری ها و توانایی تعیین الگوهای مشترک بین بیماری ها و در نتیجه تقویت نتایج مربوطه دارند. به همین دلیل بسط مدل های مکانی_زمانی تک متغیره به مدل های مکانی_زمانی چندمتغیره مورد توجه قرار گرفته است. تاکنون مطالعات اندکی ایده ی نقشه بندی مکانی_زمانی بیماری ها را با تحلیل توأم دو یا بیش از دو بیماری ترکیب کرده و از مدل¬های بیزی سلسله مراتبی، بیز کامل، رویکرد تحلیل عاملی، مدل های مکانی_زمانی SC و .. استفاده نموده اند (احمدی پناه و همکاران، ۲۰۱۸؛ مهکی و همکاران، ۲۰۱۸؛ راعی و همکاران، ۲۰۱۸؛ تی زالا و همکاران، ۲۰۰۳).

۱.۸.۲ مدل بیز کامل مکانی_زمانی چندمتغیره

در این مدل O_{ijk} تعداد موارد بروز بیماری امk در ناحیه ی امi و دوره ی زمانی امj و θ_{ijk} برآورد خطر نسبی (SMR) در نظر گرفته می شود، فرض می شود O_{ijk} ها از توزیع پوآسون با میانگین $E_{ijk}\theta_{ijk}$ پیروی می کنند که E_{ijk} ها بیانگر مقادیر مورد انتظار بیماری ها هستند. مدل لگ خطی برای میزان های بروز بیماری ها به کار گرفته شده و با وارد نمودن اثرات پراکندگی مکانی و زمانی، این مدل تغییرپذیری در θ_{ijk} به دلیل نواحی، سال ها و بیماری ها را در نظر می گیرد. روند مکانی پنهان برای معرفی ارتباط بین نواحی همسایه و روند زمانی پنهان برای مشخص نمودن همبستگی در طول زمان انتخاب می شود. ارتباط مکانی_زمانی مورد نظر بر اساس عوامل خطر مشخص می گردد. بدین ترتیب مدل عبارت است از:

$$\begin{aligned} O_{ijk} &\sim Poisson(E_{ijk}\theta_{ijk}) \\ \log(\theta_{ijk}) &= \alpha_k + \delta_k\lambda_i + \psi_k t_j + \varepsilon_{ijk} \end{aligned}$$

که در آن K معرف بیماری ها و α_k عرض از مبدأ اختصاصی بیماری امk است. λ_i نشان دهنده ی اثر تصادفی مکانی در ناحیه ی i و δ_k معرف پارامتر مقیاس مکانی برای بیماری امk است. t_j بیانگر اثر تصادفی زمانی در دوره ی زمانی j و ψ_k معرف پارامتر مقیاس زمانی برای بیماری امk است. ε_{ijk} پراکندگی در نظر گرفته نشده در مدل است (اوسن و همکاران، ۲۰۰۸).

۹.۲ مدل های مکانی_زمانی SC دو متغیره

می توان بعد زمانی را از طریق اجزاء مشترک و اختصاصی زمانی به مدل SC وارد نمود. در ساده ترین شکل، مدل مکانی_زمانی فاقد عبارات اثر متقابل بین زمان و مکان و ناهمگنی است:

$$\mu_{ij۱} = \lambda_i\delta + t_j\psi, \quad \mu_{ij۲} = \frac{\lambda_i}{\delta} + \frac{t_j}{\psi} + \phi_i + \gamma_j$$

که در آن λ_i معرف الگوی مکانی مشترک، ϕ_i معرف تفاوت الگوی مکانی مشترک دو بیماری، t_j معرف روند زمانی مشترک و γ_j معرف تفاوت روند زمانی دو بیماری است. ضرائب δ و ψ معرف اثر نسبی سهم عبارات مشترک در خطر بیماری اول در مقایسه با بیماری دوم است. می توان اثرات متقابل مکانی_ زمانی مشترک بین دو بیماری و ناهمگنی های اختصاصی بیماری ها را به مدل اضافه نمود (احمدی پناه و همکاران، ۲۰۱۸؛ مهکی و همکاران، ۲۰۱۸؛ راعی و همکاران، ۲۰۱۸).

۱۰.۲ مدل تحلیل عاملی مکانی_زمانی چند متغیره

رویکرد تحلیل عاملی برای در نظر گرفتن تحلیل عاملی پراکندگی مکانی_زمانی در میزان های اختصاصی نواحی چند بیماری در طول زمان بسط داده شده است. در این رویکرد چارچوب مدل سازی سلسله مراتبی بیزی اتخاذ گردیده و فرمول بندی های پیشین متفاوت برای عوامل مکانی و زمانی پنهان منعکس کننده ی الگوی مشترک خطرات در نظر گرفته می شود. چارچوبی پیشنهاد می شود که به آسانی می تواند هر تعداد بیماری مورد نظر را وارد آن نمود ضمن آن که همبستگی زمانی نیز در نظر گرفته می شود. این کار با فرمول‌بندی مدل در یک چارچوب مدل بندی سلسله مراتبی انجام می گیرد که در آن در گام نخست از درست نمایی پوآسون برای تعداد مشاهده شده ی بیماری ها استفاده می شود و در گام دوم عوامل پنهان معرفی می گردند (تی زالا و همکاران، ۲۰۰۳).

۳ مقایسه ی مدل ها

۱.۳ مقایسه مدل های چند متغیره و تک متغیره

تحلیل توأم دو یا چند بیماری دارای محاسن متعددی نسبت به تحلیل تک متغیره است. این محاسن شامل بهبود ثبات برآوردها برای نواحی جداگانه، توانایی تشخیص الگوهای مشترک و اختصاصی بیماری ها و خطرات که در تحلیل جداگانه کمتر آشکار است، مشخص کردن خوشه های توأم مربوط به عوامل خطر مشترک بیماری ها، توانایی تعیین وابستگی مکانی بین چند بیماری و ارتباط آن ها با عوامل خطر و افزایش دقت برآورد خطرات بیماری ها است. به علاوه انتظار می رود مدل بندی توأم تغییرپذیری در برآورد خطرات را کاهش دهد. کاهش تغییرپذیری در بیماری های دارای شیوع و بروز بالاتر، کمتر است اما در بیماری های نادر، میزان کاهش چشمگیر است. همچنین مدل بندی توأم بهبودی قابل ملاحظه ای را بر اساس معیار DIC [۶] ، نسبت به مدل های تک متغیره نتیجه می دهد. این بهبودی به دلیل بهبود برازش مدل با تعداد کمتر پارامترهای مؤثر مور نیاز در مدل بندی توأم است (هلد و همکاران، ۲۰۰۵؛ دنینگ و همکاران، ۲۰۰۸؛ دابنی و همکاران، ۲۰۰۵).

۲.۳ مقایسه مدل های توأم و رگرسیون اکولوژیک

رویکرد رگرسیون اکولوژیک در کنار مدل های چند سطحی، اولین گام ها در نقشه بندی معرفی تحلیل مکانی توأم دو بیماری بوده است. در ابتدا پیشنهاد می شد که از میزان های بروز دیگر بیماری ها به عنوان اندازه های مواجهه ی جانشین در رگرسیون استفاده شود با این حال مدل بندی توأم که به طور همزمان تغییرات مکانی در خطر دو یا چند بیماری مرتبط با هم را مدل بندی می کند، رویکردی ساده تر، پرتوان تر و طبیعی تر برای تشخیص الگوهای جغرافیایی در خطر به نظر می رسد زیرا بیماری ها به عنوان متغیرهای پاسخ وارد مطالعه شده و قابل ارجاع و انتساب به عوامل خطر پنهان می باشند (هلد و همکاران، ۲۰۰۵).

[6] Deviance Information Criterion

۳.۳ مقایسه مدل های چندمتغیره ی بیزی و غیر بیزی

میزان های به دست آمده از مدل های چندمتغیره ی بیزی دقیق تر بوده در حالی که در روش های غیربیزی، فواصل اطمینان بسیار بزرگ است. همچنین اجرای روش زنجیر مارکف مونت کارلو ساده‌تر بوده و برآوردهای قابل اعتمادتری را ارائه می‌دهند (اسونسا و همکاران، ۲۰۰۴؛ ماندا و همکاران، ۲۰۰۷).

۴.۳ مقایسه مدل های SC و نرمال چند متغیره MVN

بر اساس شاخص ،DIC مدل SC بهتر از مدل MVN است. این مسئله به دلیل پیچیدگی مؤثر بالاتر (وجود پارامترهای مؤثر بیشتر) مدل MVN است که با برازش بهتر جبران نمی گردد. در مقابل، گرچه مدل SC از منظر تعداد پارامترهای اثرات تصادفی پیچیده تر است اما به نظر می رسد اطلاعات را از راه مناسب تری اقتباس کند و بنابراین دارای تعداد مؤثر پارامترهای کمتر است. مدل SC توانایی بیشتری در تشخیص نواحی دارای خطر بالاتر دارد. در حالت وجود دو بیماری، مدل SC اندکی بهتر از مدل MVN عمل می کند. این مسئله ممکن است به دلیل فرض بسیار محدودکننده ثابت بودن همبستگی بین دو بیماری در تمام نواحی در مدل MVN باشد درحالی که مدل SC امکان متفاوت بودن خطر در نواحی مطالعه را از طریق پارامتر مقیاس فراهم می آورد (مک نب، ۲۰۰۹).

۵.۳ مقایسه مدل های SC و PM

برای هر دو مدل می توان کووریت های در سطح ناحیه ای را به پیشگوی خطی اضافه نمود. مدل PM از این لحاظ با مدل SC مشترک است که روندهای مشابه و غیرمشابه را کاملا آشکار می کند (دابنی و همکاران، ۲۰۰۵).

۶.۳ مقایسه مدل هایSC و PL

مدل SC برای نقشه بندی توآم بیماری ها طبیعی تر و منعطف تر از مدل PL است زیرا همه عامل خطر به عنوان اجزاء مشترک یا اختصاصی مدل در نظر گرفته می شوند. با این حال مدل PL دارای محاسن متعددی است. از آن جمله می توان به امکان تحلیل داده های مرگ و میر بدون دسترسی به اطلاعات جامعه ی در معرض خطر، امکان در نظر گرفتن پراکندگی در برآوردهای اثر سن در مدل، حذف میدان تصادفی نرمال مربوط به عبارات مشترک برای همه ی بیماری ها از مدل با استفاده از یک بیماری خاص به عنوان گروه مرجع و کاهش انحراف معیار برآوردهای عبارات ناهمگنی همبسته ی مکانی اختصاصی بیماری اشاره نمود. از سوی دیگر می توان از این نکته که انقباض اجزاء خوشه بندی در مدل PL نسبت به مدل SC بیشتر است به عنوان ایراد مدل PL یاد نمود (درسی ، ۲۰۰۷).

۷.۳ مقایسه مدل های مکانی چندمتغیره و مکانی ـ زمانی چندمتغیره

تحلیل های مکانی ـ زمانی برتری هایی نسبت به تحلیل مکانی محض دارند زیرا امکان مطالعه ی هم‌زمان ثبات روندها در طول زمان و برجسته نمودن روندهای غیرمعمولی را فراهم می آورند. توانایی تعیین الگوهای مکانی خطرات بیماری ها که در طول زمان ثابت بوده یا افزایش می یابند، شواهد متقاعدکننده تری نسبت به تحلیل الگوهای مکانی در یک مقطع زمانی خاص ارائه می دهد و می تواند راهنمای مهمی برای تشریح علل بیماری و کمک به ایجاد خط مشی بهداشت محیطی فراهم آورد (اوسن و همکاران، ۲۰۰۸).

۴ بحث و نتیجه‌گیری

تاکنون مدل های گوناگونی برای نقشه بندی چندمتغیره ی بیماری ها مطرح شده است که هرکدام دارای مبانی تئوری، مزایا و معایب و نقاط قوت و ضعف خاص خود بوده و در مواردی خاص نسبت به دیگر مدل ها برتری دارند. در سال های اخیر مدل هایی معرفی شده اند که افزون بر تحلیل پراکندگی بیماری ها یا مرگ و میر ناشی از آن، الگوی پراکندگی مکانی عوامل خطر بیماری ها و نقش و اهمیت این عوامل برای بیماری های مرتبط را نیز آشکار کنند. استفاده از این مدل ها به ویژه مدل SC گسترش روزافزونی پیدا کرده و برتری آن نسبت به مدل هایی که تنها الگوی پراکندگی جغرافیایی بیماری ها را ارائه می دهند، نشان داده شده است. همچنین با توجه به اهمیت و کاربرد فراوان تحلیل روند زمانی میزان های بیماری ها، چند مدل نقشه بندی چندمتغیره ی مکانی_زمانی بیماری ها معرفی شده اند که نسبت به مدل های فاقد بعد زمانی دارای محاسن متعددی می باشند و تمرکز بر مطالعه و مقایسه ی این مدل ها نیز می تواند مفید و حائز اهمیت باشد. در این مقاله مجموعه ای از مهم ترین مدل های نقشه بندی چندمتغیره ی بیماری ها تشریح و مورد مقایسه قرار گرفته و تفاوت ها و شباهت های آن ها مورد بررسی قرار گرفت اما از آنجا که در سالیان اخیر تلاش های فراوانی در جهت گسترش و بسط بیش از پیش و روزافزون این مدل ها به عمل آمده است، لزوم مقایسه ی دامنه ی گسترده تری از مدل ها، هم از منظر تئوری و هم از منظر نتایج در یک مطالعه شبیه سازی جامع ضروری به نظر می رسد.

مراجع

Ahmadipanah, M. V., Hassanzadeh, A. and Mahaki, B. (2018), Bivariate Spatiotemporal Shared Component Modeling: Mapping of Relative Death Risk Due to Colorectal and Stomach Cancers in Iran Provinces.

Assunção, R. M. and Castro, M. S. (2004), Multiple Cancer Sites Incidence Rates Estimation Using a Multivariate Bayesian Model, *International journal of epidemiology*, **33**, 508-16.

Besag, J., York, J. and Mollie, A., (1991), Bayesian Image Restoration with Two Applications In Spatial Statistics (with discussion), *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Böhning, D.,(1999), Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others, CRC Press; Taylor & Francis Group LLC, Florida, USA.

Dabney, A. R. and Wakefield, J. C.(2005), Issues in the Mapping of Two Diseases, *Statistical Methods In Medical Research*, **14**, 83-112.

Dreassi, E. (2007), Polytomous Disease Mapping to Detect Uncommon Risk Factors for Related Diseases, *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **49**, 520-9.

Downing, A., Forman, D., Gilthorpe, M. S., Edwards, K. L. and Manda, S. O. (2008), Joint Disease Mapping Using Six Cancers in the Yorkshire Region of England, *International Journal of Health Geographics*, **7**, 41.

Gelfand, A. E. and Vounatsou, P. (2003), Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis, *Biostatistics*, **4**, 11-5.

Haddad, K. A., Jafari, K. T. and Mahaki, B. (2015), Investigating the Incidence of Prostate Cancer in Iran 2005–2008 Using Bayesian Spatial Ecological Regression Models, *Asian Pacific Journal of Cancer Prevention*, **16**, 5917 21.

Held, L., Natário, I., Fenton, S. E., Rue, H. and Becker, N. (2005), Towards Joint Disease Mapping, *Statistical Methods in Medical Rresearch*, **14**, 61-82.

Jin, X., Carlin, B. P. and Banerjee, S.(2005), Generalized Hierarchical Multivariate CAR Models For Areal Data, *Biometrics*, **61**, 950-61.

Khoshkar, A. H., Koshki, T. J. and Mahaki, B., (2015), Comparison of Bayesian Spatial Ecological Regression Models For Investigating the Incidence of Breast Cancer in Iran, 2005, *C*, **16**, 5669-73.

Kim, H., Sun, D. and Tsutakawa, R. K. A. (2001), Bivariate Bayes Method for Improving the Estimates of Mortality Rates with a Twofold Conditional Autoregressive Model, *Journal of the American Statistical Association*, **96**, 1506-21.

Knorr-Held, L. and Best, N. G. (2001),A Shared Component Model for Detecting Joint and Selective Clustering of Two Diseases, *Journal of the Royal Statistical Society: Series A*, **164**, 73–85.

Langford, I. H., Leyland, A. H., Rasbash, J. and Goldstein, H. (1999), Multilevel Modelling of the Geographical Distributions of Diseases, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48**, 253-68.

Lawson, A. B., Browne, W. J. and Rodeiro ,C. L. (2003), Disease Mapping with WinBUGS and MLwiN, John Wiley & Sons.

Lawson, A. B., Biggeri, A, B., Boehning, D., Lesaffre, E., Viel, J. F. Clark, A. Schlattmann, P. and Divino, F. (2000), Disease Mapping Models: An Empirical Evaluation, Disease Mapping Collaborative Group, *Statistics in Medicine*, **19**, 2217-2241.

Leyland AH, Langford IH, Rasbash J, Goldstein H,(2000 Sep 15), Multivariate spatial models for event data.*Statistics in Medicine*, **19**, 2469-78.

MacNab, Y. C. (2009), Bayesian Multivariate Disease Mapping and Ecological Regression with Errors in Covariates: Bayesian Estimation of DALYs and 'Preventable'DALYs, *Statistics in Medicine*, **28**, 1369-85.

Mahaki, B., Mehrabi, Y., Kavousi, A., Akbari, M. E., Waldhoer, T., Schmid, V. J. and Yaseri, M. (2011), Multivariate Disease Mapping of Seven Prevalent Cancers in Iran using Shared Component Model *Asian Pacific Journal of Cancer Prevention*, **12**, 2353.

Mahaki, B., Mehrabi, Y., Kavousi, A. and Schmid, V. J. (2018), Joint Spatio-Temporal Shared Component Model with an Application in Iran Cancer Data, *Asian Pacific Journal of Cancer Prevention*, **19**, 1553-1560.

Manda, S. O. and Leyland, A. (2070), An Empirical Comparison of Maximum Likelihood and Bayesian Estimation Methods for Multivariate Disease Mapping: Theory and Methods, *South African Statistical Journal*, **41**, 1-21.

Marshall, R. J. (1991), Mapping Disease and Mortality Rates Using Empirical Bayes Estimators, *Applied Statistics*, **1**, 283-94.

Mollie, A. (1990), Representation Geographiqu des Taux de Mortalire: Modelisation Spatiele et Methodes Bayesiennes, PhD thesis, Universite Paris 6, France.

Nasrazadani, M., Maracy, M. R., Dreassi, E. and Mahaki, B. (2018), Mapping of Stomach, Colorectal, and Bladder Cancers in Iran, 2004–2009: Applying Bayesian polytomous Logit Model.

Raei, M., Johann, S. V. and Mahaki, B. (2018), Bivariate Spatiotemporal Disease Mapping of Cancer of the Breast and Cervix Uteri Among Iranian Women, it Geospatial Health, **13**, 164-71.

Rastaghi, S., Jafari, K. T. and Mahaki, B.,(2015), Application of Bayesian Multilevel Space Time Models to Study Relative Risk of Esophageal Cancer in Iran 2005 2007 at a County Level, *Asian Pacific Journal of Cancer Prevention*, **16**, 5787 92.

Richardson, S., Abellan, J. J. and Best, N. (2006), Bayesian Spatio-Temporal Analysis of Joint Patterns of Male and Female Lung Cancer Risks in Yorkshire (UK), *Statistical Methods in Medical research*, **15**, 385-407.

Simonoff, J. S.,(2012), Smoothing Methods in Statistics, Springer Science & Business Media.

Tzala, E. H., Marshall, C. and Best, N. (2003), Multivariate Analysis Of Spatial And Temporal Variation In Cancer Mortality In Greece: Isee-398. *Epidemiology*, **14**, S79-80.

Oleson, J. J., Smith, B. J. and Kim, H. (2008), Joint Spatio-Temporal Modeling of Low Incidence Cancers Sharing Common Risk Factors, *Journal of Data Science*, **6**, 105-23.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲–۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

مدل‌سازی داده‌های فضایی ناایستا

علی محمدیان مصمم، پریا عیوضی[۱]
گروه آمار، دانشگاه زنجان

چکیده: در آمار فضایی اغلب برای سادگی فرض می‌شود که داده‌های موجود ایستا می‌باشند. لذا روش‌های مدل‌سازی داده‌ها و همچنین پیش‌بینی مشاهدات، مثلاً کریگینگ، اغلب بر اساس این فرض بنا شده‌اند. ولی فرض ایستایی داده‌های فضایی در عمل به دلیل تغییرپذیری داده‌ها به‌ویژه در مقیاس بزرگ دامنه فضایی برقرار نیست و بنابراین روش‌های موجود و کریگینگ کلاسیک برای این داده‌ها دقیق نخواهد بود. در این مقاله یک روش جدید برای پیش‌بینی داده‌های فضایی در حالت ناایستایی مورد مطالعه قرار می‌گیرد. در این روش یک فضای ترکیبی جدید تعریف شده که این امر موجب می‌شود از مدل‌های فضایی–زمانی ایستا نظیر مدل تفکیک‌پذیر ضربی–جمعی و مدل تفکیک‌ناپذیر نایتینگ استفاده گردد.

واژه‌های کلیدی: آمار فضایی، کریگینگ، فرآیند ناایستا، مدل ضربی–جمعی، مدل نایتینگ
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30.

۱ مقدمه

در آمار فضایی اغلب برای سادگی فرض می‌شود که داده‌های موجود ایستا باشند لذا روش‌های مدل‌سازی داده‌ها و همچنین پیش‌بینی مشاهدات، مثلاً کریگینگ، بر اساس این فرض بنا شده‌اند. ولی فرض ایستایی داده‌های فضایی در عمل به دلیل تغییرپذیری داده‌ها به‌ویژه در مقیاس بزرگ دامنه فضایی برقرار نیست و بنابراین روش‌های موجود و کریگینگ کلاسیک (کرسی، ۱۹۹۳) برای این داده‌ها دقیق نخواهد بود. زمین‌آمار (چیلز و دلفینر، ۲۰۰۹) امکان ترکیب انواع مختلف داده را فراهم می‌کند و از طریق استفاده از متغیرهای توضیحی (بیچامپ و همکاران، ۲۰۱۷) مناسب، فقدان ایستایی را در نظر می‌گیرد. مدل‌سازی روند خارجی یک روش مناسب برای انجام این کار است.

در این مدل یک ارتباط قطعی خطی با ضرایب مجهول بین متغیر برای برآورد و متغیر کمکی معرفی می‌شود و از آنجا که ضرایب روند، شناخته شده نیستند، این رابطه می‌تواند به‌طور محلی با استفاده از یک همسایگی متحرک تنظیم شود اما اگر روند دارای مقادیر بسیار پایینی باشد، برخی تخمین‌های غیردقیق ممکن است رخ دهد. هدف این مقاله بررسی راه‌های

[۱] نام و ایمیل ارائه دهنده مقاله: پریا عیوضی، eivazi.paria@znu.ac.ir

مختلف در نظر گرفتن متغیر توضیحی بدون ایجاد فرض یک رابطه خطی بین داده‌ها و متغیر توضیحی است. در اینجا از متغیر توضیحی $\nu \in V$ به‌عنوان یک بعد جدید فضای ترکیبی $D \times V$ و $D \subseteq R^2$ استفاده می‌شود. فضای ترکیبی که در قبل معرفی شد برای ایجاد مجموعه‌ای از برآوردگرهای زمین‌آماری از جمله کریگینگ عادی، کریگینگ با روند خارجی، هم‌کریگینگ با یک مدل متریک یا ضربی–جمعی (دیاگو و همکاران، ۲۰۰۱، ۲۰۰۳) برای برازش مدل تغییرنگار استفاده می‌شود. با استفاده از تکنیک اعتبارسنجی متقابل [1] عملکرد این برآوردگرها ارزیابی و مزایا و معایب آن‌ها مشخص می‌شود.

۲ روش کار

داده‌های فضایی Z(s) جمع‌آوری شده از شبکه پایش کیفیت هوا را در نظر بگیرید. در کیفیت هوا، حتی اگر فرآیند اساسی مشاهدات جمع‌آوری شده از شبکه پایش ناایستا باشند، Z(s) گاهی به‌عنوان یک تابع تصادفی ایستای مرتبه دوم تلقی می‌شود و کریگینگ عادی برای پیش‌بینی داده‌ها در یک شبکه منظم با فرض ایستایی استفاده می‌شود. برآوردگر $Z_{\cdot}^{OK}$ که از ترکیب خطی $\Sigma_\alpha \lambda_\alpha Z_\alpha$ و وزن λ_α به‌دست آمده است، محدودیت‌های ناارریبی و بهینگی را حل می‌کند و به سیستم کریگینگ عادی معروف است. برای پرداختن به مساله ناایستایی، یک روش رایج مبتنی بر معرفی یک روند خطی $\mu(s) = \Sigma_l a_l f_l(s)$ به‌عنوان بخش قطعی Z(s) است که مشاهدات را به خوبی توضیح می‌دهد:

$$Z(s) = \mu(s) + R(s).$$

با فرض اینکه ناایستایی فرآیند در بخش قطعی محاسبه می‌شود، مدل تغییرنگار (یا کواریانس) برای بخش تصادفی R(s) موردنیاز است. در این چارچوب وزن‌های λ_α از حل سیستم کریگینگ با روند خارجی حاصل می‌شوند. باقیمانده R_α برای برآورد تغییرنگار نمونه مورداستفاده قرار می‌گیرد که به‌عنوان تفاوت بین مشاهدات واقعی Z_α و پیش‌بینی ارائه‌شده توسط روند برآوردشده بدون در نظر گرفتن مکان s_α محاسبه می‌شود. در این حالت برآورد انجام‌شده بر روی باقیمانده‌های R_α تمایل به انحراف کمتری دارند. مدل‌سازی IRF-k [2] روش دقیق برازش غیرمستقیم تغییرنگار است. ایده اصلی اینست که متغیر توضیحی $\nu \in V$ معمولا به‌عنوان یک روند در زمین‌آمار دیده می‌شود و می‌توان آن را به‌عنوان یک مختصات فضایی معرفی کرد. در فضای ترکیبی $R^2 \times V$، $Z(s, \nu)$ به‌عنوان یک تابع تصادفی حقیقی‌مقدار تعریف می‌شود که با فرض اینکه ایستای مرتبه دوم است ν را مولفه جدیدی از تابع کواریانس در نظر می‌گیرد. تابع کواریانس به‌صورت زیر ارائه می‌شود:

$$C_{SV}(\|h_S\|, |h_V|) = Cov(Z(s + \|h_S\|, \nu + |h_V|), Z(s, \nu)).$$

۳ تعیین مدل تغییرنگار

در فضای ترکیبی $D \times V$، از فاصله اقلیدسی کلاسیک $\|h_{S\times V}\| = \sqrt{\|h_S\|^2 + (k.|h_V|)^2}$ برای محاسبه تغییرنگار استفاده می‌شود.

$$\gamma_{SV}(\|h_s\|, |h_V|) = \gamma_m(\|h_{S\times V}\|)$$

در اینجا k پارامتر ناهمسانگردی مقیاس‌بندی بین R^2 و V است. بنابراین پس از تطبیق $\|h_s\|$ و $|h_V|$ و $\|h_{S\times V}\|$ از طریق k، تمام فواصل به‌طور مساوی در مدل موسوم به متریک γ_m بررسی می‌شوند. دراین‌صورت تغییرنگارهای حاشیه‌ای $\gamma_{SV}(\|h_S\|, ۰)$ و $\gamma_{SV}(۰, |h_V|)$ ویژگی‌های اساسی یکسانی دارند و درصورتی‌که مدل محدود باشد، آستانه یکسانی دارند.

[1] Cross-validation
[2] Intrinsic Random Fields of order k

به‌طور خاص، اگر آستانه γ_m به‌وسیله یکی از تغییرنگارهای حاشیه در فواصل کوچک به‌دست آید امکان اینکه γ_{SV} مقادیر بزرگتر از آستانه داشته باشد وجود ندارد، حتی اگر تغییرنگار ناایستای تصادفی نمونه مقادیر بزرگتر را نشان دهد.

استفاده از مدل جمعی‌ـ متریک به‌عنوان ترکیبی از مدل‌های فضایی، کمکی و متریک که لزوماً ساختار یکسانی را نشان نمی‌دهند می‌تواند به‌عنوان یک راهکار درنظرگرفته شود هرچند تطبیق داده‌ها با چنین ترکیبی دشوار خواهد بود. اگر کواریانس ناایستای تصادفی به‌صورت تفکیک‌پذیر فرض شود می‌تواند به‌عنوان کواریانس‌های صرفا فضایی و کمکی نمایش داده شود. اما به‌جای در نظر گرفتن مدل‌هایی با ویژگی‌های محدود می‌توان از خانواده‌های تفکیک‌ناپذیر (نایتینگ، ۲۰۰۲) عمومی استفاده کرد که در میان آن‌ها مدل ضربی‌ـجمعی یک روش تفکیک‌ناپذیر انعطاف‌پذیر ارائه می‌دهد و قادر به تعامل با ناهمسانگردی‌های مختلف در فضای $D \times V$ می‌باشد. در اینجا یک روش یکسان برای جایگزین کردن کواریانس صرفا زمانی $C_{T(h_T)}$ توسط یک کواریانس صرفا کمکی $C_{V(h_V)}$ استفاده می‌شود:

$$C_{SV}(\|h_S\|, |h_V|) = k_1 C_S(h_S) C_V(h_V) + k_2 C_S(\|h_S\|) + k_3 C_V(|h_V|).$$

همچنین می‌توان از مدل نایتینگ به‌عنوان یک مدل دیگر از خانواده‌های تفکیک‌ناپذیر عمومی استفاده کرد که به‌صورت

$$C(\|h_S\|, |h_T|) = (\phi_T(|h_T|) + 1)^{-\frac{\delta}{2}} C_S(\|h_S\|(\phi_T(|h_T|) + 1)^{-\frac{1}{2}})$$

ارائه می‌شود، که در آن ϕ_T یک تابع دلخواه، C_S تابع کواریانس فضایی و $\delta > 0$ است. مدل ضربی‌ـجمعی به دو دلیل نسبت به مدل نایتینگ برتری دارد:

۱ـ تغییرنگار نمونه اغلب داده‌ها شباهت بیشتری به رفتار تعریف‌شده توسط مدل ضربی‌ـجمعی دارد.

۲ـ به‌نظر می‌رسد کواریانس ناایستای تصادفی نمونه اغلب به‌صورت تفکیک‌پذیر با پارامتر k نزدیک به صفر پس از برازش روش بر روی یک مجموعه داده بزرگ شامل شود.

۴ تحلیل داده‌ها

در اینجا به ارزیابی مجموعه برآوردگرهای کریگینگ در یک مطالعه موردی برای میانگین روزانه دی‌اکسید نیتروژن در ۱۹ ژانویه ۲۰۱۱ می‌پردازیم. مطالعه انجام‌شده بر روی یک حوزه متمرکز در فرانسه است که در آن ایستگاه‌های مرتبط با ترافیک و صنعت مورد استفاده قرار نمی‌گیرند. برآوردگرهای مورداستفاده در این مطالعه با نماد Z_{vxx}^{kxx} ارائه می‌شود که در آن kxx برای نوع کریگینگ است و مقادیر آن را در حالت OK[3]، KED[4]، OCK[5] و $CKED$[6] می‌گیرد، همچنین vxx برای مدل تغییرنگار استفاده می‌شود یعنی زمانی‌که هیچ مقداری داده نشود یک تغییرنگار فضایی با فاصله اقلیدسی $\|h_S\|$ است درحالیکه در فضای ترکیبی، اندیس‌های m و psm به‌ترتیب برای مدل‌های متریک و ضربی‌ـجمعی مورد استفاده قرار می‌گیرد.ساختار مورداستفاده برای برازش تغییرنگار، مدل‌های معروف تغییرنگار هستند که عموماً در تحلیل کیفیت هوا استفاده می‌شوند، که شامل مدل کروی و نمایی برای تغییرنگار ایستا و مدل خطی برای در نظر گرفتن مولفه ذاتی درصورت نیاز، استفاده می‌شود. پیش‌بینی‌ها بر اساس فضای کلاسیک و فضای ترکیبی انجام گرفته است. برای مقایسه آن‌ها از معیار $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ استفاده شده است. نتایج پیش‌بینی‌ها در جدول ۱ ارائه شده است. باتوجه به جدول ۱ روش Z_{psm}^{KED}، دارای RMSE کمتری می‌باشد.

[3] ordinary kriging
[4] kriging with external drift
[5] ordinary cokriging
[6] cokriging with external drift

جدول ۱: جذر میانگین خطای پیش‌بینی برای روش‌های مختلف در حالت کلاسیک و فضای ترکیبی

نوع فضا	روش پیش‌بینی	RMSE
بدون استفاده از CHIMERE	Z^{OK}	۱۲/۸۴
	$Z^{KED}, f=(\nu)$	۱۱/۰۳
	Z^{OK}_{psm}	۱۰/۶۵
	Z^{OK}_{m}	۱۱/۱۵
با استفاده از CHIMERE	$Z^{KED}, f=(\varphi)$	۱۰/۹۷
	$Z^{KED}, f=(\varphi,\nu)$	۱۰/۵۶
	$Z^{KED}_{psm}, f=(\varphi)$	۹/۷۸
	$Z^{KED}_{m}, f=(\varphi)$	۱۱/۵۵
فضای ترکیبی	Z^{OCK}	۱۳/۴۷
	$Z^{CKED}, f=(\nu)$	۱۱/۱۸
	Z^{OCK}_{psm}	۹/۹
	Z^{OCK}_{m}	۱۰/۶۵

بحث و نتیجه‌گیری

تعریف ناایستایی با در نظر گرفتن متغیر توضیحی به‌عنوان یک بعد از فضای ترکیبی می‌تواند نتایج کلاسیک با روند خارجی را بهبود بخشد. مطالعات روی داده‌های NO_2 نشان می‌دهد بهترین نتیجه با برازش مدل تغییرنگار ضربی‌ـجمعی روی باقیمانده‌های رگرسیون بین مشاهدات و متغیرهای توصیفی CHIMERE به‌دست آمده‌اند. در این مقاله برای سادگی مدل ضربی‌ـجمعی با رویکرد متریک مورداستفاده قرار گرفته است هرچند ممکن است که همیشه تضمین‌کننده بهترین مدل نباشد.

مراجع

Beauchamp, M., de Fouquet, C., and Malherbe, L. (2017), Dealing with Non-Stationarity Through Explanatory Variables in Kriging-based Air Quality Maps, *Spatial Statistics*, **22**, 18-46.

Chiles, J. P., and Delfiner, P. (2009), *Geostatistics: Modeling Spatial Uncertainty*, Vol. 497, John Wiley & Sons.

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

De Iaco, S., Myers, D. E., and Posa, D. (2001), Space–time Analysis Using a General Product–Sum Model, *Statistics & Probability Letters*, **52**, 21-28.

De Iaco, S., Myers, D. E., and Posa, D. (2003), The Linear Coregionalization Model and the Product–Sum Space–Time Variogram, *Mathematical Geology*, **35**, 25-38.

Gneiting, T. (2002), Nonseparable, Stationary Covariance Functions for Space–Time Data, *Journal of the American Statistical Association*, **97**, 590-600.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

تحلیل داده‌های فضایی-زمانی در حضور مقادیر گمشده

علی محمدیان مصمم، **زهرا مقیمی**[1]
گروه آمار، دانشگاه زنجان

چکیده: در بسیاری از تحقیقات با مواردی برخورد می‌کنیم که برای بعضی از متغیرها، پاسخ یا جوابی وجود ندارد یا اینکه پاسخ نامعلوم بوده و قادر به استخراج آن نیستیم. به این مقادیر، مقادیر نامعلوم یا داده‌های گمشده می‌گوییم. در داده‌های فضایی-زمانی با توجه به وجود وابستگی بین مشاهدات برحسب موقعیت فضایی و زمان مشاهده هر داده، مقادیر گمشده‌ای که در فواصل فضایی یا زمانی نزدیک‌تر نسبت به مشاهدات قرار دارند می‌توانند شامل اطلاعات مفیدی باشند که استفاده از آن‌ها می‌تواند منجر به نتایج دقیق‌تری شود. بنابراین لازم است وجود گمشدگی در داده‌ها مورد توجه و بررسی قرار گیرد. برای رسیدگی به مشکل داده‌های گمشده در این مقاله یک الگوریتم جدید مورد مطالعه قرار می‌گیرد. روش پیشنهادی هر مقدار گمشده را جداگانه براساس نقاط داده شده در یک محدوده فضایی-زمانی در اطراف آن نقطه پیش‌بینی می‌کند. پیش‌بینی مقادیر گمشده و تخمین عدم قطعیت پیش‌بینی مربوطه براساس روش‌های مرتب‌سازی و رگرسیون چندکی انجام گرفته است. در پایان روش پیشنهادی روی داده‌های حاصل از سنجش از راه دور مربوط به پوشش گیاهی منطقه‌ای از آلاسکا مورد استفاده قرار گرفته است.

واژه‌های کلیدی: داده‌های فضایی-زمانی، داده‌های گمشده، الگوریتم `gap-fill`، شاخص پوشش گیاهی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30.

۱ مقدمه

سنجش از دور[1] یک تکنولوژی جدید برای مطالعه‌ی طیف گسترده‌ای از فرآیندهای سطح زمین از یک فاصله دور از زمین است. در این مقاله داده‌های پوشش گیاهی با استفاده از این تکنولوژی مورد تحلیل قرار می‌گیرد. این داده‌ها با توجه به اینکه در مقیاس مکانی و در طول زمان اندازه‌گیری می‌شوند، معمولاً دارای همبستگی فضایی-زمانی می‌باشند. یک مشکل اصلی در تحلیل چنین داده‌هایی وجود مقادیر زیادی داده‌های گمشده[2] می‌باشد. به عنوان مثال، در اندازه‌گیری پوشش گیاهی سطح

[1] Remote Sensing
[2] Missing Values

[1] نام و ایمیل ارائه دهنده مقاله: زهرا مقیمی، zahra_moghimi@znu.ac.ir

زمین توسط ماهواره‌ها به دلیل برخی موانع نظیر وجود ابر در مسیر ثبت پوشش گیاهی با حجم عظیمی از مقادیر گمشده مواجه می‌شویم. فقدان داده‌ها، معمولاً یکی از مسائل اصلی در حوزه‌ی جمع‌آوری داده می‌باشد، برای همین روش‌های مختلفی در برخورد با داده‌های گمشده ارائه گردیده است که برای جلوگیری از مشکلاتی مانند کمی حجم نمونه و یا عدم حضور واحدهایی از نمونه در مجموعه داده‌ها باید با استفاده از روش‌های مناسب جایگذاری شوند نظیر حذف داده گمشده از تحلیل، جایگذاری داده گمشده با مقادیر مناسب مانند میانگین، استفاده از برآوردهای رگرسیونی، روش داده افزایی، روش جانهی چندگانه که روشی مبتنی بر میانگین‌های شبیه‌سازی است (راگونتان، ۲۰۱۵). رابین (۱۹۷۶) انواع مختلف داده‌های گمشده را به سه دسته کاملاً تصادفی، تصادفی و غیر تصادفی طبقه‌بندی کرده است. دو استراتژی در این روش‌ها مشاهده می‌شود. استراتژی اول در برخورد با داده های گمشده استفاده از روش‌های آماری است که نسبت به حضور مقادیر گمشده مقاوم هستند. استراتژی دوم این است که مقادیر گمشده قبل از هرگونه تحلیل آماری مورد پیش‌بینی قرار گیرند هرچند کمتر مقاله‌ای برای داده‌های گمشده فضایی_زمانی می‌توان یافت. براساس استراتژی دوم در این مقاله یک الگوریتم جدید برای تخمین و جایگزینی داده‌های گمشده فضایی_زمانی مورد مطالعه قرار می‌گیرد. ادامه ساختار مقاله به این صورت می‌باشد که در بخش ۲ به معرفی الگوریتم gap-fill می‌پردازیم. در بخش ۳ به تحلیل داده‌های پوشش گیاهی با استفاده از الگوریتم مورد مطالعه پرداخته می‌شود. سرانجام بحث و نتیجه‌گیری در بخش ۴ ارائه شده است.

۲ معرفی الگوریتم gap-fill

یکی از روش‌های پیش‌بینی مقادیر گمشده فضایی_زمانی در مجموعه داده‌های سنجش از راه دور استفاده از الگوریتم gap-fill است. این الگوریتم مشابه روش ارائه شده در وایس و همکاران (۲۰۱۴) است. با این تفاوت که روش جدید دارای کاربرد ساده‌تر بوده و علاوه بر این روش ما در انتخاب زیرمجموعه‌های همسایگی مقادیرگمشده باعث افزایش دقت می‌شود. الگوریتم gap-fill برای داده‌های ماهواره‌ای که عمدتاً دارای آرایه‌های چهار بعدی می‌باشند مناسب است. این چهار بعد به عنوان مثال می‌توانند شامل دو بعد فضایی برای طول و عرض جغرافیایی و دو بعد زمانی برای فصل و سال اندازه‌گیری‌ها باشد. از طرفی معمولاًداده‌های ماهواره‌ای داده‌های با اندازه بزرگ بوده لذا این الگوریتم باید دارای سرعت محاسباتی بالایی نیز باشد.

الگوریتم gap-fill (گربر و همکاران، ۲۰۱۸):

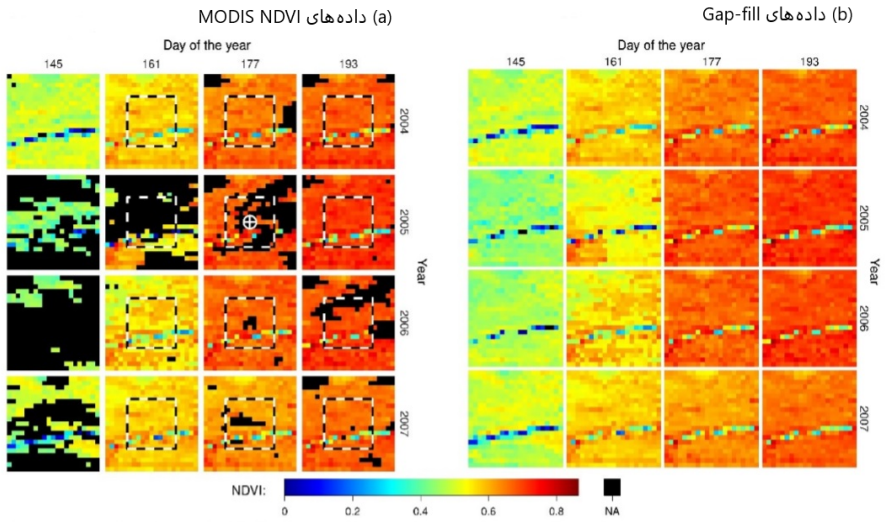
گام ۱: استخراج زیر مجموعه‌ای از داده‌ها

گام ۲: رتبه‌بندی تصاویر ماهواره‌ای

گام ۳: پیش‌بینی مقادیر گمشده براساس زیرمجموعه انتخابی با استفاده از رگرسیون چندکی[3].

در ادامه گام‌های الگوریتم gap-fill برای یک داده‌ی گمشده تشریح می‌گردد و برای تمام مقادیر گمشده قابل تکرار است. درمرحله اول یک زیرمجموعه‌ای از داده‌های مورد نیاز پیش‌بینی از همسایگی مقدار گمشده انتخاب می‌شود. این زیرمجموعه خود شامل چهار بعد بوده که شامل دو بعد همسایگی فضایی و دو بعد همسایگی زمانی می‌باشد. توجه کنید که اندازه همسایگی‌ها طوری انتخاب می‌شود که فرآیند کلی برای مجموعه مشاهدات بهینه باشد. این کار معمولاً با استفاده از روش اعتبار سنجی متقابل انجام می‌گیرد. معیار انتخاب یک زیرمجموعه خوب این است که اولاً حداقل ۵ تصویر غیرگمشده در آن موجود باشد. ثانیاً اگر مقدار گمشده‌ای در زیرمجموعه انتخابی است حداقل ۲۵ تصویر گمشده در آن موجود باشد. در

[3] Quantile Regression



شکل ۱: داده‌های اصلی و داده‌های حاصل از الگوریتم gap-fill

مرحله دوم، تصاویر ماهواره‌ای براساس شاخص‌های پوشش گیاهی رتبه‌بندی می‌شوند. در مرحله سوم مقادیر گمشده براساس زیرمجموعه انتخابی با استفاده از رگرسیون چندکی خطی (کونیکر، ۲۰۰۵) پیش‌بینی می‌شوند. الگوریتم فوق دربسته آماری gap-fill در نرم‌افزار R موجود می‌باشد. از ویژگی مهم دیگر این الگوریتم این است که با محاسبات موازی می‌توان سرعت پردازش داده‌ها را افزایش داد. ما الگوریتم فوق را روی داده‌های MODIS NDVI بکار بردیم. توجه داشته باشید که روش پیشنهادی را می‌توان برای طیف وسیعی از داده‌های سنجش از راه دور اعمال کرد.

۳ تحلیل داده‌های پوشش گیاهی

شاخص نرمال شده تفاوت پوشش گیاهی[۴] که به اختصار NDVI نامیده می شود، شاخص گرافیکی ساده‌ای است که در تحلیل‌ها و اندازه‌گیری‌های سنجش از دور و ارزیابی وجود یا عدم وجود پوشش گیاهی یک منطقه کاربرد دارد. دامنه تغییرات این شاخص بین ۱+ و ۱ـ است. برای بررسی پوشش گیاهی هر منطقه‌ای می‌توان از روش‌های مختلفی استفاده کرد. اما بکارگیری تصاویر ماهواره‌ای می‌تواند تغییرات زمانی و مکانی پوشش گیاهی هر منطقه‌ای را با دقت بالاتری فراهم کند. در این زمینه پایگاه داده پوشش گیاهی سنجده MODIS یکی ازپایگاه داده‌های مهم است. یک مشخصه از داده‌های MODIS این است که مقدار ساختار وابستگی‌های فضایی_زمانی را از خود نشان می‌دهد. داده‌های مورد مطالعه در این مقاله دارای اندازه فضایی ۲۱ × ۲۱ پیکسل و شامل ۱۶ تصویر با ۴ شاخص فصلی (روزهای ۱۹۳، ۱۷۷، ۱۶۱، ۱۴۵ از هر سال) به صورت شبکه‌های منظم، مشاهده شده در طی ۴ سال (۲۰۰۷ـ۲۰۰۴) که آرایه داده در مجموع ۷۰۵۶=۱۶ × ۲۱ × ۲۱ عنصر دارد. خروجی الگوریتم بر روی داده‌ها در شکل ۱ ارائه شده است. شکل۱ (a) داده‌های اصلی پوشش گیاهی را نشان می‌دهد.این شکل شامل ۲۱ × ۲۱ پیکسل فضایی می‌باشد که در ۱۶ نقطه در ۴ روز متفاوت در یک سال است. داده‌های رسم شده شامل ۱۶۰۳ مقدار گمشده می‌باشد که بصورت بلوک‌های سیاه رنگ نمایش داده شده‌اند.

برای پیش‌بینی ۱۷۷امین روز سال ۲۰۰۵ که با علامت سفید ⊕ رنگ نمایش داده شده است زیر مجموعه انتخابی با مستطیل خط چین مورد استفاده قرار گرفته است. در شکل ۱ (b) خروجی gap-fill رسم شده است. نمودار فوق نشان

[4] Normalized Difference Vegetation Index

دهنده‌ی کارایی بسیار خوب الگوریتم مورد مطالعه است. مطالعات بعدی با استفاده از معیارهای مناسب آماری برای ارزیابی بهتر کارایی روش ارائه شده می‌باشد.

۴ بحث و نتیجه‌گیری

الگوریتم gap-fill یک الگوریتم انعطاف‌پذیر برای پیش‌بینی مقادیر گمشده داده‌های شاخص پوشش گیاهی می‌باشد. این الگوریتم به دلیل استفاده از محاسبات موازی و روش‌های پیش‌بینی آماری دارای سرعت و دقت بهتری نسبت به الگوریتم‌های موجود در این زمینه می‌باشد. از مزایای دیگر این الگوریتم این است که نیازی به فرض نرمال بودن داده‌ها ندارد. علاوه بر این، این الگوریتم دارای کارایی بالایی در حضور مقادیر گمشده با حجم زیاد می‌باشد. مهمترین مشکل آن می‌تواند برای حالتی باشد که داده‌ها روی شبکه منظم قرار نگرفته باشند.

مراجع

Gerber, F., de Jong, R., Schaepman, M. E., Schaepman-Strub, G., & Furrer, R. (2018), Predicting Missing Values in Spatio-Temporal Remote Sensing Data, *IEEE Transactions on Geoscience and Remote Sensing*, **56**(5), 2841-2853.

Koenker, R. (2005), *Quantile Regression (Econometric Society Monographs)*, Cambridge: Cambridge University Press.

Raghunathan, T. (2015), *Missing Data Analysis in Practice*, Chapman and Hall/CRC.

Rubin, D. B. (1976), Inference and Missing Data. *Biometrika*, **63**(3), 581-592.

Weiss, D. J., Atkinson, P. M., Bhatt, S., Mappin, B., Hay, S. I., & Gething, P. W. (2014), An Effective Approach for Gap-filling Continental Scale Remotely Sensed Time-series, *ISPRS Journal of Photogrammetry and Remote Sensing*, **98**, 106-118.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

تحلیل بیزی داده‌های گسسته فضایی-زمانی

علی محمدیان مصمم، الناز عباسی[1]
گروه آمار، دانشگاه زنجان

چکیده: در تحلیل بیزی داده‌ها گاهی با داده‌های گسسته‌ای مواجه می‌شویم که به دلیل ناگاوسی بودن توزیع متغیر پاسخ و وجود تعداد زیادی متغیر پنهان در مدل تحت بررسی شکل بسته‌ای برای توزیع پسینی وجود ندارد. از طرفی ممکن است متغیرهای پاسخ در طول زمان به یکدیگر وابسته باشند. در این شرایط به دلیل عدم استقلال داده‌ها یا ناگاوسی بودن توزیع داده‌ها استفاده از روش‌های مونت‌کارلوی زنجیر مارکف ($MCMC$) با چالش‌هایی نظیر، وجود پارامترهای زیاد در ساختار سلسله مراتبی، محاسبات سنگین، شبیه‌سازی گسترده، طولانی و زمان‌بر بودن محاسبات به‌ویژه زمانی که بعد میدان تصادفی بزرگ است و عدم همگرایی توزیع پسینی مواجه هستیم. برای حل این مشکلات (رو و همکاران ، ۲۰۰۹) روش تقریب لاپلاس آشیانی جمع بسته ($INLA$) را پیشنهاد دادند. این روش ضمن حفظ دقت برآوردها و پیش‌گویی‌های روش $MCMC$ سرعت محاسبات قابل قبولی دارد. در این مقاله در یک مطالعه موردی به تحلیل داده‌های جرم و جنایت کشور کانادا با رهیافت $INLA$ می‌پردازیم.

واژه‌های کلیدی: تقریب لاپلاس آشیانی جمع بسته، آمار بیزی، آمار فضایی-زمانی
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30.

۱ مقدمه

تحلیل بیزی داده‌های جرم و جنایت معمولاً به صورت استنباط بیزی الگوهای فضایی محض یا الگوهای زمانی محض انجام می‌گیرد. اگر مشاهدات به‌گونه‌ای باشند که مشاهدات نزدیک به هم وابسته‌تر و مشاهدات دورتر وابستگی کمتری داشته باشند، این‌گونه مشاهدات داده‌های فضایی نامیده می‌شوند. بدیهی است که تحلیل آماری داده‌های فضایی با روش‌های آماری معمول مقدور نیست، زیرا شرط اساسی استقلال داده‌ها تحقق پیدا نمی‌کند. با این حال درمورد داده‌های جرم و جنایت چنین تحلیل‌های بیزی صرفاً فضایی یا زمانی به دلیل اینکه اثر متقابل فضا و زمان را در نظر نمی‌گیرند مدل‌های مناسبی نیستند. در این مقاله سعی شده است که با استفاده از مدل‌های بیزی فضایی-زمانی به تحلیل و تفسیر این اثر نیز پرداخته شود.

[1] نام و ایمیل ارائه دهنده مقاله: الناز عباسی، Elnaz.Abbasi@Znu.ac.ir

مشاهداتی که هم از نظر موقعیت فضایی و هم از نظر موقعیت زمانی وابسته باشند، داده‌های فضایی_زمانی نامیده می‌شوند. در آمار فضایی_زمانی به طور معمول یک میدان تصادفی به عنوان مدل آماری داده‌های فضایی_زمانی در نظر گرفته می‌شود. میدان تصادفی فضایی_زمانی مجموعه‌ای از متغیرهای تصادفی مانند $Z(\cdot,\cdot)=\{Z(s,t);(s,t)\in D\times T\}$ است، که در آن s موقعیت فضایی در مجموعه $D\subseteq R^d$، $d\geq 1$ و t لحظه زمانی در مجموعه $T\subseteq R$ است.

روش مرسوم برای محاسبات بیزی استفاده از روش‌های مونت‌کارلوی زنجیر مارکف $(MCMC)$ مشکلات فراوانی به دلایلی نظیر وابستگی شدید بین مشاهدات و بعد زیاد ابرپارامترها در محاسبات توزیع‌های پسینی را تجربه کرده است. لذا هدف این مقاله استفاده از روش تقریب لاپلاس آشیانی جمع بسته $(INLA)$ برای تحلیل چنین داده‌هایی است. روش $INLA$ اولین بار توسط (رو و همکاران ، ۲۰۰۹) ارائه شد و سپس توسط (مارتین و همکاران ، ۲۰۱۳) و (رو و همکارن ، ۲۰۱۷) توسعه پیدا کرد.

ادامه ساختار مقاله به صورت زیر می‌باشد. در بخش ۲ روش $INLA$ معرفی می‌شود. در بخش ۳ داده‌های تماس به پلیس برای گزارش جرم و جنایت مدل بندی می‌شوند، پارامترهای مدل برآورده شده و با استفاده از معیار بیزی اطلاع انحراف DIC بهترین مدل انتخاب می‌شود.

۲ تقریب لاپلاس آشیانی جمع بسته

همانگونه که در مقدمه ملاحظه گردید روش‌های عددی مرسوم در آمار بیزی نظیر روش $MCMC$ دارای مشکلات اساسی به ویژه سرعت همگرایی و مدت زمان لازم برای محاسبات عددی می‌باشند. در این بخش به معرفی روش جایگزین، $INLA$، می‌پردازیم. $INLA$ یک رهیافت جدید برای استنباط تقریبی در مدل‌های گاوسی پنهان (LGM) است. مطالعات نشان می‌دهند که برای چنین مدل‌هایی اغلب اوقات روش $INLA$ سریع‌تر و دقیق‌تر از روش $MCMC$ می‌باشد. چهارچوب روش $INLA$ نیازمند سه مؤلفه کلیدی است: مدل‌های گاوسی پنهان، میدان تصادفی مارکف گاوسی $(GMRF)$ (لینگرین و همکاران ، ۲۰۱۱) و تقریب لاپلاس می‌باشد. اغلب مدل‌های بیزی در رگرسیون، آمار فضایی و آمار فضایی_زمانی دارای ساختار گاوسی پنهان می‌باشند. علاوه بر این در کاربست $INLA$ نیازمند یک میدان تصادفی مارکف گاوسی می‌باشیم. $GMRF$ یک میدان تصادفی گاوسی با خاصیت استقلال شرطی است یعنی x و x' به شرط سایر مؤلفه‌ها مستقل هستند اگر و تنها اگر i و jامین مؤلفه ماتریس دقت Q صفر هستند.

برای توصیف مفاهیم فوق مدل سلسله مراتبی سه مرحله‌ای زیر را در نظر بگیرید، فرض کنید که در مرحله اول متغیر پاسخ $y=(y_1,\cdots,y_n)'$ به‌طور شرطی مستقل و دارای توزیع نمایی خاصی باشد

$$y|x,\theta_1\sim\Pi_{i=1}^{n}P(y_i|x_i;\theta_1). \qquad (۱.۲)$$

در مرحله دوم x را یک میدان تصادفی گاوسی پنهان با تابع چگالی

$$P(x|\theta_2)\propto|Q_{\theta_2}|_+^{1/2}\exp(-1/2x'Q_{\theta_2}x) \qquad (۲.۲)$$

در نظر بگیرید که در آن ماتریس دقت Q_{θ_2} یک ماتریس معین مثبت با پارامتر θ_2 و $|Q_{\theta_2}|_+$ حاصل‌ضرب مقادیر ویژه غیر صفر آن می‌باشد و وارون Q_{θ_2} همان ماتریس واریانس کواریانس می‌باشد. در مرحله نهایی فرض کنید که پارامتر $\theta=(\theta_1,\theta_2)$ دارای یک توزیع پیشینی $\pi(\theta)$ است. در نتیجه پسینی پارامترهای مدل سلسله مراتبی به صورت

$$\begin{aligned}\pi(\eta,\theta|y)&\propto\pi(\theta)\pi(\eta|\theta_2)\Pi_{i=1}^{n}\pi(y_i|\eta_i;\theta_1)\\&\propto\pi(\theta)|Q_{\theta_2}|_+^{1/2}\exp\{-1/2\eta'Q_{\theta_2}\eta+\Sigma_{i=1}^{n}\log P(y_i|\eta_{i;\theta_1})\}\end{aligned} \qquad (۳.۲)$$

می‌باشد. توزیع‌های پسینی حاشیه‌ای برای متغیرهای پنهان و ابر پارامترها به صورت

$$\begin{aligned}\pi(\eta_i|y) &= \int \pi(\eta_i|\theta,y)\pi(\theta|y)d\theta \\ \pi(\theta_i|y) &= \int \pi(\theta|y)d\theta_{-i}\end{aligned} \tag{۴.۲}$$

به‌دست می‌آیند. (رو و همکارن ، ۲۰۱۷) یک تقریب لاپلاس برای توزیع‌های حاشیه‌ای پسینی ارائه کردند. با استفاده از روش تقریب لاپلاس در $INLA$ توزیع‌های پسینی هر یک از پارامترهای نامعلوم به‌دست می‌آید.

۳ مدل بندی تماس برای خدمات پلیس

منطقه آنتاریو در ایالت واترلو شامل شهرهای واترلو، کیچینر، کمبریج و ۴ منطقه روستایی می‌باشد. جمعیت این منطقه در سال ۲۰۱۱، ۵۰۶۱۰۷ نفر بوده که در ۷۵۵ ناحیه‌ی سرشماری توزیع شده است. در کل ۲۹۰۲۷ تماس به پلیس برای گزارش جرم و جنایت در طول ۹۰۶۰ واحد فضا_ زمان ثبت شده است (لوان و همکارن ، ۲۰۱۶). این داده‌ها برای سال ۲۰۱۱ در ۱۲ دوره ۲ ساعته طبقه‌بندی شده‌اند. در این مقاله برای مدل‌سازی فضایی_زمانی بیزی از مدل سلسله مراتبی سه مرحله‌ای و رهیافت $INLA$ استفاده می‌شود. در مرحله اول چون تعداد تماس‌های جرم و جنایت کم می‌باشد لذا تعداد تماس‌ها برای خدمات پلیس در منطقه i، $i = ۱,\cdots,۷۵۵$ و دوره زمانی ۱۲ ، $t = ۱,\cdots,۱۲$، دارای توزیع پواسن به صورت

$$y_{it}|\mu_{it} \sim POI(\mu_{it}) \quad i = ۱,\cdots,۷۵۵, \quad t = ۱,\cdots,۱۲ \tag{۱.۳}$$

فرض شده است. میانگین توزیع تعداد تماس‌ها را به صورت

$$\log(\mu_{it}) = \log(E_{it}) + \alpha + u_i + s_i + \gamma_t + \phi_t + \psi_{it} \tag{۲.۳}$$

در نظر بگیرید که در آن E_{it} تعداد مورد انتظار تماس‌ها بوده و ریسک نسبی به ریسک کلی یا اثر ثابت (α)، اثرات تصادفی فضایی $(u_i + s_i)$، اثرات تصادفی زمانی $(\gamma_t + \phi_t)$ و اثر تصادفی فضایی_ زمانی (ψ_{it}) تجزیه شده است.
در مرحله دوم فرض کنید

$$\begin{aligned}u_i &\sim Normal(۰.\sigma_u^۲), \quad i = ۱,\cdots,۷۵۵ \\ s_i &\sim ICAR(W,\sigma_S^۲) \\ \gamma_t &\sim Normal(۰,\sigma_\gamma^۲), \quad t = ۱,\cdots,۱۲ \\ \phi_t &\sim ICAR(P,\sigma_\phi^۲)\end{aligned} \tag{۳.۳}$$

که در آن $\sigma_u^۲$، $\sigma_S^۲$، $\sigma_\gamma^۲$ و $\sigma_\phi^۲$ ابر پارامترها می‌باشند. همانگونه که ملاحظه می‌گردد اثرات تصادفی فضایی و اثرات تصادفی زمانی به دو مؤلفه بدون ساختار و ساختاریافته تجزیه شده‌اند که مدل‌های ساختاریافته شامل مدل $ICAR$ می‌باشند که به ترتیب همبستگی فضایی و زمانی این اثرات را مدل بندی می‌کنند. $ICAR$ یک توزیع پیشینی مرسوم در آمار فضایی برای پارامترهای اثرات تصادفی می‌باشد که در آن میانگین مورد انتظار s_i برابر میانگین همسایگی آن می‌باشد.
برای مرحله سوم تحلیل بیزی سلسله مراتبی توزیع پیشینی تمام ابر پارامترها $Gamma(۰/۵, ۰/۰۰۰۵)$ در نظر می‌گیریم.
برای تحلیل بیزی سلسله مراتبی از روش $INLA$ استفاده می‌کنیم. چهار مدل متفاوت برای تحلیل بیزی داده‌ها در نظر گرفته شده است: ۱) ψ_{it}ها مستقل و هم توزیع هستند.

۲) ψ_{it}ها دارای همبستگی فضایی و ناهمبستگی زمانی هستند.
۳) ψ_{it}ها دارای ناهمبستگی فضایی و همبستگی زمانی هستند.
۴) ψ_{it}ها دارای همبستگی فضایی و زمانی هستند.

۱.۳ انتخاب مدل

مدل‌سازی در نرم‌افزار R با روش $INLA$ با ۴ مدل فوق انجام شده است. ورودی‌های تابع $INLA$ و نقشه‌ها با نرم‌افزار GIS تهیه شده‌اند. DIC برای انتخاب مدل بهتر از معیار DIC استفاده می‌شود. DIC معیار بیزی برای نیکویی برازش و تعیین پیچیدگی مدل است. (اشپیگل و همکاران ، ۲۰۰۳) پیشنهاد دادند که امید ریاضی پسین آماره انحراف $\bar{D}$ و تعداد پارامترهای مؤثر در مدل p_D اطلاعات مدل را خلاصه می‌کنند. بنابراین معیار اطلاع انحراف DIC به صورت $DIC = \bar{D} + p_D$ است: مدلی که کمترین DIC را نسبت به بقیه مدل‌ها داشته باشد مدل بهتری است. مقادیر DIC و تعداد پارامترهای مؤثر برای ۴ مدل فضایی‌ـزمانی و تعامل فضاـزمان در جدول ۱ آورده شده است. طبق جدول ۱.۳ مدل اول با کمترین DIC بهترین مدل انتخاب شده است.

جدول ۱: مقایسه DIC مدل‌ها

تعامل فضاـزمان	پارامترهای تعامل	اطلاع انحراف	پارامتر موثر
مدل اول	γ_t و u_i	۵۴۰۸۹	۴۹۸۰
مدل دوم	ϕ_t و u_i	۷۲۷۹۱/۲۸	۷۵۸
مدل سوم	γ_t و s_i	۵۴۴۹۶	۴۸۰۰/۷۲
مدل چهارم	ϕ_t و s_i	۷۲۷۹۱/۵۲	۷۵۸/۴۴

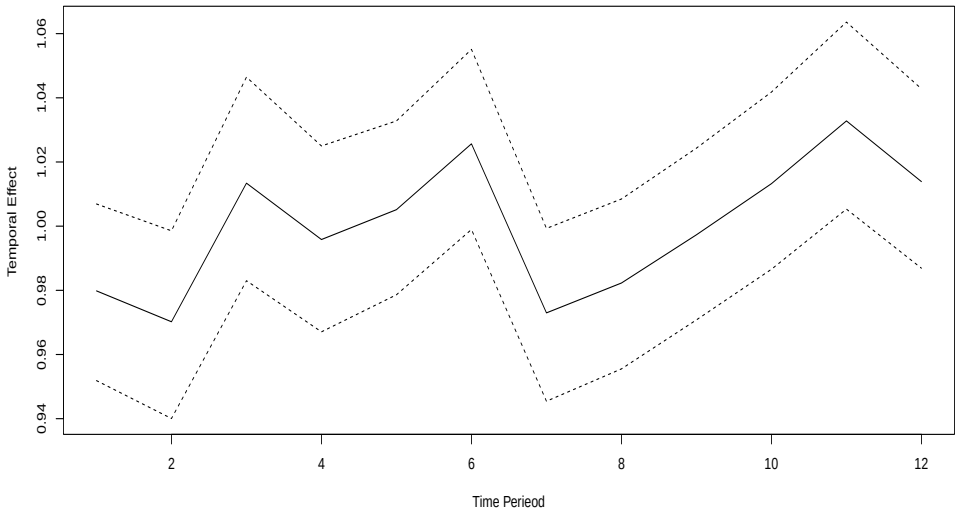
۲.۳ برآورد پارامترهای مدل

برآورد پارامترهای مدل اول با روش $INLA$ در جدول ۲ آورده شده است. نمودار ۱ نشان‌دهنده‌ی اثرات تصادفی زمانی

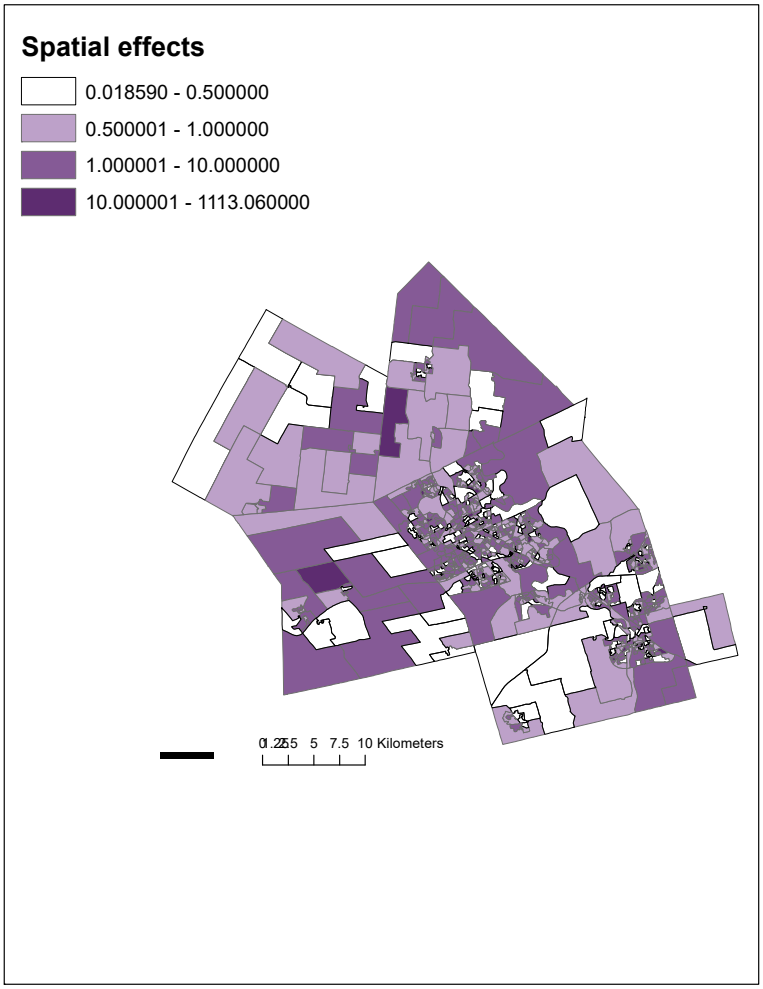
جدول ۲: برآورد پارامترهای مدل اول

پارامترهای مدل	میانگین	انحراف پسینی	چارک اول	میانه	چارک سوم	مد
σ_u^2	۰/۹۱۹	۰/۴۸۹	۰/۸۲۷	۰/۹۱۷	۱/۰۱۹	۰/۹۱۴
σ_γ^2	۱۷۸۹/۲۹۰	۹۸۷/۱۵۱	۵۴۴/۹۷۶	۱۵۷۴/۷۰۸	۴۲۸۹/۶۶۲	۱۲۰۲/۱۶۲
σ_ψ^2	۱۳/۰۳۴	۰/۳۵۵	۱۲/۳۱۵	۱۳/۰۲۸	۱۳/۷۱۳	۱۳/۰۵۴

است. براساس این نمودار بیشترین تأثیرات اصلی زمانی در دوره زمانی بین ۵۹ : ۱۱ – ۰۰ : ۱۰ و ۵۹ : ۲۱ – ۰۰ : ۲۰ است. بنابراین تخصیص منابع انسانی پلیس، باید در این دوره زمانی بیشتر باشد. شکل ۲ نشان‌دهنده‌ی اثرات تصادفی فضایی است. مناطقی با اثرات فضایی بالا، به‌طور مداوم تعداد زیادی تماس خدمات دارند و مناطقی هستند که منابع پلیس باید برای تمام دوره‌های زمانی به کار گرفته شود. در حالت کلی اثرات اصلی فضایی در مراکز منطقه مورد مطالعه بیشترین و در مناطق اطراف روستایی به‌ویژه در شمال غربی کمترین میزان است.



شکل ۱: اثرات تصادفی زمانی همراه با فاصله باورمندی ۹۵ درصد



شکل ۲: اثرات تصادفی فضایی

۴ بحث و نتیجه‌گیری

در این مقاله، در یک مطالعه موردی به تحلیل بیزی داده‌های جرم و جنایت کشور کانادا با رهیافت $INLA$ پرداخته شد. برای این منظور چهار مدل مختلف به داده‌ها برازش و بهترین مدل با استفاده از معیار DIC انتخاب گردید.

مراجع

Luan, H., Quick, M., and Law, J. (2016), Analyzing l Local Spatio-Temporal Patterns of Police Calls-for-Service Using Bayesian Integrated Nested Laplace Approximation, *International Journal of Geo-Information*, **5**(9), .162

Lindgren, F., Rue, H., and Lindström, J. (2011), An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Eifferential Equation Approach, *Journal of the Royal Statistical Society: Series B*, **73**(4), 423-498.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013), *Bayesian computing with INLA: new features.* Computational Statistics and Data Analysis, **67**, .68-83

Rue, H., Martino, S., and Chopin, N. (2009), Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations, *Journal of the Royal Statistical Society: Series B*, **71**(2), .319-392

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017), Bayesian Computing with INLA: A Review, *Annual Review of Statistics and Its Application*, **4**, 395-421.

Spiegelhalter, D., Best, N. G., Carlin, B. P., and van der Linde, A. (2003), Bayesian Measures of Model Complexity and Fit, *Quality Control and Applied Statistics*, **48**(4), .431-432

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

پایش کیفیت آب‌های سطحی با استفاده از تکنیک‌های آمار فضایی: مطالعه موردی استان زنجان

سرور موسوی[1]، یونس خسروی، عباسعلی زمان، مهسا دادشی
گروه علوم محیط زیست، دانشگاه زنجان

چکیده: آب‌های سطحی جاري یا رودخانه‌ها از مهم‌ترین منابع آب هستند که در تامین آب مورد نیاز فعالیت‌های مختلف مانند کشاورزي، صنعت، شرب و تولید برق نقش مهمی دارند. برخی کشورها منابع آب سطحی خود را براساس توسعه پایدار برنامه‌ریزی می‌کنند. یکی از نیازمندي‌های مهم در برنامه‌ریزی و توسعه منابع آب و حفاظت و کنترل آن‌ها آگاهی از کیفیت منابع آب است. کیفیت و آلودگی آب‌های سطحی تحت تاثیر تغییرات طبیعی و انسانی قرار دارند. در این مطالعه ارزیابی کیفیت آب‌های سطحی استان زنجان از نظر آلاینده ترکیب بیکربنات برای ۲۰ سال نمونه‌برداری شد. برای درک بهتر فرآیندهای شیمیایی و بررسی کیفیت آب از روش‌های آماری فضایی هم‌چون شاخص‌های موران محلی و لکه‌های داغ و نرم‌افزار $ArcGIS$ استفاده شده است. این مطالعه با هدف شناسایی عامل‌های موثر بر میزان و پراکنش بیکربنات انجام گرفته است. نتایج نشان داد که هیچ لکه‌ی داغی در روش لکه‌های داغ وجود ندارد. اما روش موران محلی حاکی از خوشه‌بندی بالا ـ بالا و پایین ـ پایین در مقادیر بیکربنات است. که احتمال می‌رود متاثر از نوع زمین‌شناسی و کاربری اراضی منطقه مورد مطالعه باشد. به منظور مدیریت بهتر محیط، می‌توان ایستگاه‌های مهم منبع آلودگی در آب‌های سطحی استان زنجان را شناسایی کرد. شاخص‌های موران محلی و لکه‌های داغ ابزاری مفید برای شناسایی نقاط مهم آلودگی در آب‌های سطحی استان زنجان است.

واژه‌های کلیدی: آب‌های سطحی، آمار فضایی، شاخص موران محلی، شاخص لکه‌های داغ
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30.

۱ مقدمه

مشکل کمبود آب و کیفیت آن از دیدگاه سازمان ملل متحد پس از مسئله جمعیت به عنوان دومین مشکل اصلي جهان شناخته شده است. با ورود به هزاره سوم و همزمان با افزایش جمعیت به ویژه در کشورهای در حال توسعه، تقاضا

[1] نام و ایمیل ارائه دهنده مقاله: سرور موسوی، s.musavi@Znu.ac.ir

برای آب به منظور تأمین نیازهای جمعیتی افزایش قابل ملاحظه‌ای یافته است و سبب شده تا استفاده بهینه از آب و حوضه‌های آبی موجود در جهان با تأمل و برنامه‌ریزی بهتری حرکت کند (محمدجانی و یزدانیان، ۲۰۱۴) . امروزه افزایش آلاینده‌های شیمیایی شهری و صنعتی و روش‌های نوین کشاورزی تهدیدی جدی برای محیط‌زیست به حساب می‌آیند. برای ارزیابی توسعه‌ی یک منطقه، کیفیت آب از مهم‌ترین مولفه‌ها است. کیفیت آب براساس متغیرهای فیزیکی و شیمیایی تعریف می‌شود. ترکیب شیمیایی آب سطحی، به عنوان منبع آبی برای مصرف انسانی و حیوانی، آبیاری و صنعتی را شامل می‌شود. بنابراین هدف، تعریف کیفیت آب نیست، بلکه استفاده مطلوب از آب در جامعه است (دشتی برمکی و همکاران، ۱۳۹۳) (فرچیچی و همکارن، ۲۰۱۸) . یکی از نیازمندی‌های مهم در برنامه‌ریزی و توسعه منابع آب، حفاظت و کنترل آن‌ها و آگاهی از کیفیت منابع آب است. موردهایی هم‌چون انتشار مواد مغذی کشاورزی و مواد زاید شهری، افزایش تقاضای آب، استانداردهای سطح بالای زندگی و کاهش منابع قابل قبول آب سبب آلودگی آب و ایجاد وضعیت نامناسب محیط‌زیستی در سراسر جهان شده است. از این رو درک بهتر تغییرات آلودگی سیستم‌های آبی ضرورت بیش‌تری یافته است (فریادی و همکاران، ۱۳۹۱) . رودخانه‌ها جز کوچکی از منابع آب‌های جهان می‌باشند که به عنوان یکی از منابع اساسی تامین آب به حساب می‌آیند (لی و همکاران، ۲۰۱۸) . امروزه تحلیل‌های آمار فضایی به عنوان ابزارهای جدید و کارآمد در علوم مختلف مورد توجه محققین قرار گرفته‌اند. توجه به عدم استقلال مشاهدات و وابسته بودن آن‌ها به یکدیگر در فضای مورد مطالعه و استفاده مستقیم از فضا، محیط، همسایگی، جهت‌گیری و روابط فضایی در محاسبات سبب برتری یافتن این روش نسبت به آمار کلاسیک شده است. در مطالعه‌ی حاضر روابط مکانی داده‌های آب‌های سطحی استان زنجان از نظر مقدار بیکربنات با استفاده از آمار فضایی بررسی شده است. به عنوان یکی از مهم‌ترین هدف‌های آمار فضایی برای تحلیل الگوهای فضایی و درک وابستگی‌های فضایی است که با استفاده از شاخص‌های موران محلی و لکه‌های داغ صورت می‌گیرد. (بحری و خسروی، ۱۳۹۶).

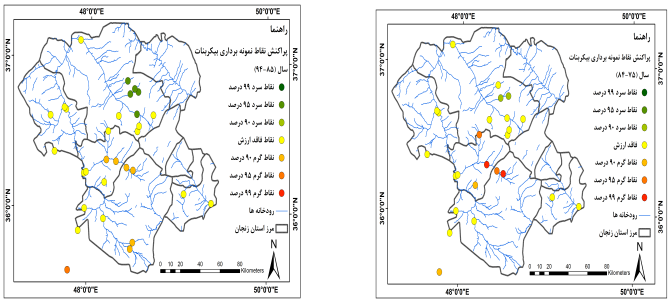
۲ مواد و روش‌ها

استان زنجان با وسعتی بیش از ۲۲ هزار کیلومتر مربع در منطقه شمال غربی کشور بین ۳۵ درجه و ۳۳ دقیقه الی ۳۷ درجه و ۱۵ دقیقه و ۳۷ درجه عرض شمالی ۴۷ درجه و ۲۶ دقیقه طول شرقی از نصف‌النهار قرار دارد. در این مطالعه از رودهای اصلی استان زنجان شامل زنجان‌رود، سجاس‌رود، رود شورچای، رودایجرود، رود قزل‌اوزن، رود ابهررود و رود خارود و علاوه بر رودهای ذکر شده از رودهای فرعی دیگری که در استان زنجان جریان دارند نمونه‌برداری شده است. پارامتر مورد مطالعه بیکربنات است که یکی از ترکیب‌های شیمیایی تاثیرگذار بر سلامت آب شرب است. مقدار بیکربنات کمتر از $۹۰ppm$ برای کشاورزی مناسب است و بیش از آن آبیاری را دچار مشکل می‌کند. بنابراین از ایستگاه‌های مختلف استان زنجان در طی ۲۰ سال نمونه‌برداری صورت گرفته است که به دو گروه ۱۰ ساله اول (از سال ۱۳۷۵ تا ۱۳۸۴) و ۱۰ ساله دوم (از سال ۱۳۸۵ تا ۱۳۹۴) تفکیک شده است. داده‌های طبقه‌بندی شده با استفاده از شاخص موران محلی $(LocalMoran)$ و شاخص لکه‌های داغ $(Hotspot)$ و نرم‌افزار $ArcGIS$ مورد بررسی قرار گرفته شد.

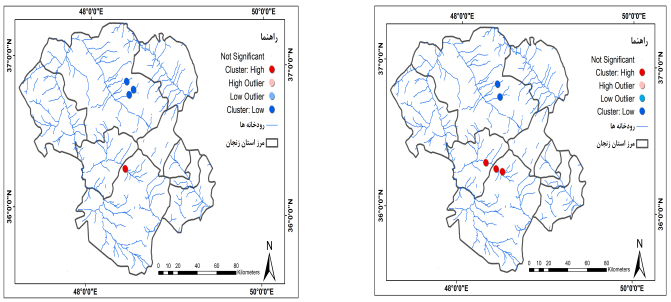
۳ بحث و نتیجه‌گیری

براساس نتایج بدست آمده، عدد خروجی شاخص لکه‌های داغ و عدد Z بدست آمده برای بیکربنات در ده سال اول (۰/۳۱) و برای ده سال دوم (۰/۷۳ـ) تایید می‌کند که هیچ‌گونه خوشه‌بندی در مقدار بیکربنات در آب‌های سطحی وجود ندارد. خروجی شاخص موران محلی نشان داد با عدد Z برابر با ۲/۱۷ در سطح اطمینان ۹۵% برای ده سال اول (۷۵ـ۸۴) خوشه‌بندی در مقدارهای بالاـبالا را تشکیل داده است. هم‌چنین عددZ برابر با ۲/۶۸ در سطح اطمینان ۹۹% برای ده

سال دوم (۸۵ـ۹۴) نشان داد که خوشه‌بندی در مقادیر بالاـ بالا تشکیل شده است. از نظر شاخص لکه‌های داغ در ده سال اول در سطح اطمینان ۹۰% نقاط سردی در رودخانه زنجان‌رود شهر زنجان مشاهده می‌شود و نقاط گرم در سطح اطمینان ۹۵% و ۹۹% نیز در رودخانه سجاس‌رود شهر ایجرود مشاهده می‌شود. هم‌چنین در ده سال دوم نیز نقاط سرد در سطح اطمینان ۹۵% در رودخانه زنجان‌رود شهر زنجان و نقاط گرم در سطح اطمینان ۹۰% در رودخانه سجاس‌رود شهر ایجرود مشاهده می‌شود. در ده سال اول و دوم از لحاظ شاخص موران محلی خوشه‌بندی بالاـ بالا در رودخانه ایجرود و خوشه‌بندی پایینـ پایین در رودخانه زنجان‌رود مشاهده می‌شود. با توجه به نقشه‌ها می‌توان گفت که احتمال رخداد این نتایج به دلیل نوع زمین‌شناسی و کاربری اراضی منطقه است. هر دو لکه‌ی مشاهده شده در یک ساختار زمین‌شناسی از نوع آهکی قرار دارند بنابراین می‌توان زمین شیمی را عامل تاثیرگذار تشکیل این لکه‌ها دانست که در شکل ۳ مشاهده می‌شود.



شکل ۱: پراکنش نقاط داغ ترکیب بیکربنات در آب‌های های سطحی استان زنجان



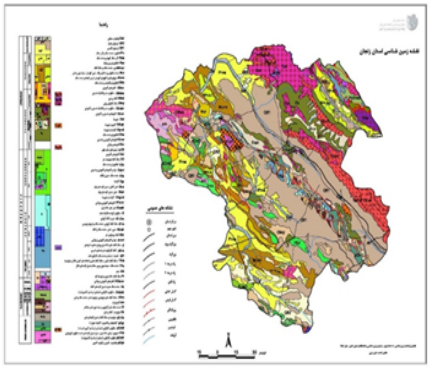
شکل ۲: وضعیت خوشه‌بندی بندی ترکیب بیکربنات در آب‌های سطحی استان زنجان

مراجع

دشتی برمکی، م.، رضایی، م.، صابری نصر، ا. (۱۳۹۳)، *ارزیابی شاخص کیفیت آب زیرزمینی (GQI) درآبخوان لنجانات با استفاده سیستم اطلاعات جغرافیایی*، نشریهٔ زمین شناسی مهندسی، جلد هشتم، شمارهٔ ۲ تابستان.

فریادی، س.، شاهدی، ک.، نباتپور، م. (۱۳۹۱)، *مطالعه پارامترهای کیفیت آب رودخانه تجن با استفاده از تکنیک‌های آماری چند متغیره* ، پژوهشنامه مدیریت حوزه آبخیز سال سوم، شماره ۶ پاییز و زمستان.

بحری، ع.، خسروی، ی.، (۱۳۹۶)، *مروری بر آمار فضایی و کاربردهای آن در محیط¬زیست* ، چهارمین کنفرانس بین¬المللی برنامه¬ریزی و مدیریت محیط زیست، خرداد ماه.



شکل ۳: زمین شناسی استان زنجان

Ferchichi, H., Hamouda, M. B., Farhat, B., and Mammou, A. B. (2018), Assessment of Groundwater Salinity Using GIS and Multivariate Statistics in a Coastal Mediterranean Aquifer, *International Journal of Environmental Science and Technology*, 1-20.

Li, P., Tian, R., and Liu, R. (2018), Solute Geochemistry and Multivariate Analysis of Water Quality in the Guohua Phosphorite Mine, Guizhou Province, China, *Exposure and Health*, 1-4.

Mohammadjani, E., and Yazdanian, N. (2014), Analysis of the Water Crisis and its Management Requirements, *Journal Ravand*, 21, 117-144.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

تحلیل زمانی و فضایی_زمانی داده‌های ریسک سلامت در تصادفات جاده‌ای استان قم

حسین والی[1]، امید کریمی، فاطمه حسینی
گروه آمار، دانشگاه سمنان

چکیده: تلفات جاده‌ای یکی از عوامل مهم مرگ غیر طبیعی در ایران و کشورهای درحال توسعه می‌باشد. جلوگیری از بروز تلفات جاده‌ای برای کاهش آمار بالای کشته‌ها امری ضروری به نظر می‌رسد. تعداد مرگ ومیر، تعداد مصدومین و تعداد تصادفات جاده‌ای از مولفه‌های مهم ریسک سلامت در تصادفات جاده‌ای به شمار می‌رود. هدف از این مقاله مدل‌سازی نرخ مرگ و میر و مجروحین در محورهای برون شهری استان قم ازسال ۱۳۸۸ الی ۱۳۹۵ بر اساس مدل‌های آماری مناسب است. داده‌ها برحسب مکان و زمان وقوع تصادف جمع‌آوری شده‌اند و بعد از پالایش و دسته‌بندی به صورت چهار محور اصلی تردد استان قم (محورشمال ـ محورجنوب ـ محورشرق ـ محورغرب) در هشت سال به‌طور ماهیانه خلاصه شده‌اند. با استفاده از مدل‌های خطی و خطی تعمیم یافته، متغیرهای کمکی تاثیرگذار را به‌وسیله یک تحلیل اکتشافی روی داده‌ها شناسایی کردیم. با توجه به تاثیر زمان و مکان در داده‌ها از مدل‌های سری‌زمانی چندمتغیره و مدل‌های فضایی_زمانی، برای تحلیل و بررسی آنها استفاده می‌کنیم. در نهایت مدل مناسب برای تحلیل و پیش‌بینی‌های آتی معرفی می‌شود.

واژه‌های کلیدی: داده‌های ریسک سلامت تصادفات استان قم، سری زمانی چندمتغیره، مدل‌های فضایی_زمانی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 62M10، 62H11.

۱ مقدمه

پرداختن به داده‌های نرخ مرگ و میر و مصدومین در تصادفات کشور دارای اهمیت زیادی است. از آنجایی که براساس برآوردهای آماری صورت پذیرفته در طول دو دهه اخیر بیش از ۴۶۰ هزار نفر از هموطنان در کشور جان خود را بر اثر سوانح رانندگی از دست داده‌اند و در عین حال در این ۲۰ سال چیزی نزدیک به ۴/۵ میلیون نفر نیز در جریان تصادفات رانندگی زخمی شده‌اند (والی ، ۱۳۹۲). در این مقاله به تحلیل داده‌های ریسک سلامت درتصادفات جاده‌ای استان قم طی سال های ۱۳۸۸ الی ۱۳۹۵ می‌پردازیم. مولفه‌های اصلی و موجود برای بررسی ریسک سلامت تصادفات جاده‌ای استان شامل،

[1] نام و ایمیل ارائه دهنده مقاله: حسین والی، h.vali.4715@gmail.com

تعداد تصادفات، تعداد مجروحین، تعداد مرگ ومیرها و نوع وسیله‌نقلیه (سواری، باربری، اتوبوس و موتورسیکلت) برحسب چهارمحور (شمال، جنوب، شرق، غرب) به صورت ماهیانه است. با توجه به تاثیر زمان در میزان تلفات در تصادفات و حتی تعداد تصادفات، لزوم استفاده از مدل‌هایی که اثر زمان را در نظر بگیرند از اهمیت بالایی برخوردار است. از این‌رو می‌توان به مدل‌های سری زمانی اشاره کرد که برای تحلیل داده‌هایی به‌کار می‌روند که برحسب زمان مرتب شده باشند. ابتدا مدل‌های سری‌زمانی را به تفکیک محورها را مورد مطالعه قرار می‌دهیم، سپس همه محورها را با هم در قالب مدل‌های سری‌زمانی چندمتغیره (نیرومند ، ۱۳۹۱) به کار می‌گیریم. موارد زیادی وجود دارند که مشاهدات مستقل نبوده و برحسب موقعیت قرارگرفتن خود در فضای مورد بررسی به یکدیگر وابسته هستند. اگر این وابستگی تابعی از فاصله بین موقعیت‌های مشاهدات باشد، به گونه‌ای که مشاهدات نزدیک به هم وابسته‌تر و مشاهدات دورتر از هم وابستگی کمتری داشته باشند. اینگونه مشاهدات داده‌های فضایی [1] نامیده می‌شوند. به‌طور معمول داده‌های فضایی علاوه بر مقادیر مربوط به متغیر یا متغیرهای مورد بررسی، اطلاع مربوط به مکان مشاهده را نیز شامل می‌شوند. بنابراین منطقی به نظر می‌رسد که مشاهده مربوط به یک مکان به نوعی به مشاهدات سایر مکان‌ها وابسته باشد. داده‌های زمانی نیز دارای چنین خصوصیتی هستند. با این تفاوت که زمان یک جریان یک طرفه و همواره روبه جلو حرکت می‌کند، یعنی از گذشته شروع و از طریق حال به آینده جریان دارد، در حالی که در داده‌های فضایی همبستگی در همه جهات منظور می‌شود. در داده‌های زمانی به دلیل جریان تک بعدی زمان، مشاهدات به طور متوالی به ترتیب زمان رخداد جمع‌آوری و در نتیجه وابستگی به صورت پیاپی بروز می‌کند. اما در داده‌های فضایی، مکان مشاهدات در فضای چند بعدی واقع شده و ممکن است به طور منظم و متوالی روی یک خط قرار نداشته باشند. در این صورت مفهوم ترتیب موقعیت مشاهدات برای داده‌های فضایی معنایی ندارد. بدیهی است که تحلیل آماری داده‌های فضایی با روش‌های آماری معمولی مقدور نیست، زیرا شرط اساسی که همان استقلال داده‌ها است، تحقق پیدا نمی‌کند. از این رو شاخه آمار فضایی برای تحلیل اینگونه مشاهدات شکل گرفته و با فراهم آمدن فنون مختلف درحال توسعه است. علم بشری در طول رشد و شکوفایی خود قدم‌های موثری در این رابطه برداشته است و در این فرآیند، مدل‌سازی را به عنوان یکی از مهمترین راهکارها برای غلبه بر نادانسته‌های خود انتخاب کرده است. یکی از مهمترین روش‌های مدل‌سازی، زمین آمار است که شامل مجموعه‌ای از ابزارهای تکنیکی برای حل مسائل محیطی و ارزیابی‌های زمانی و مکانی پدیده‌های محیطی_فیزیکی مثل عوامل شکل‌گیری و وقوع تصادفات جاده‌ای می‌باشد. برای حل اینگونه مسائل، هر چند می‌توان از روش‌های آمار کلاسیک استفاده کرد، اما مزیت زمین آمار نسبت به آمار کلاسیک در توجه به موقعیت، آرایش و همبستگی بین مشاهدات دقت کافی ندارند. این در حالی است که زمین آمار توجه ویژه‌ای به همبستگی و ساختار مکانی داده‌ها دارد (محمدزاده ، ۱۳۹۸؛ کرسی ، ۱۳۹۳).

در ادامه بیان می‌کنیم که بعد زمان و مکان داده‌ها به‌وسیله مدل‌های فضایی_زمانی و سری‌زمانی چندمتغیره (برحسب محورها) چگونه مورد مطالعه قرار می‌گیرد. در بخش بعدی مدل سری‌زمانی چندمتغیره را ارائه می‌کنیم. سپس در بخش ۳ یک تحلیل اکتشافی از داده‌ها بیان می‌شود، دربخش ۴ مدل فضایی_زمانی را روی داده‌ها به‌کار می‌گیریم و در نهایت یک بحث و نتیجه‌گیری ارائه می‌شود.

۲ مدل‌های سری‌زمانی چندمتغیره

در بسیاری از مطالعات تجربی داده‌های سری‌زمانی مشاهدات مربوط به چندین متغیر را شامل می‌شود. برای مثال، در مطالعهٔ چگونگی فروش، متغیرها، می‌توانند حجم فروش، قیمتها، توان فروش و هزینهٔ تبلیغات را در بر داشته باشد. در الگوی تابع تبدیل رابطهٔ ویژهٔ بین متغیرها، یعنی یک رابطهٔ ورودی_خروجی بین یک متغیر خروجی و یک یا چند متغیر ورودی را مطالعه

[1] spatial data

می‌کنیم. با وجود این در بسیاری از زمینه‌های کاربردی یک الگوی تابع تبدیل از نوع خروجی، ممکن است مناسب نباشد. در این بخش یک ردهٔ کلی الگوی سری‌زمانی برداری را برای بیان روابط میان چندمتغیر سری‌زمانی معرفی می‌کنیم.

۱.۲ توابع ماتریس کوواریانس و همبستگی

فرض کنید $\mathbf{X}_t = [X_{۱,t}, X_{۲,t}, \cdots, X_{m,t}]'$ و $t = ۰, \pm۱, \pm۲, \cdots$ فرآیند برداری با مقدار حقیقی تواما ایستای m ـ بعدی باشد که میانگین $E(X_{i,t}) = \mu_i$ برای هر $i = ۱, ۲, \cdots, m$ ثابت و کوواریانس متقابل بین $X_{i,t}$ و $X_{j,s}$ برای تمام $i = ۰, ۱, ۲, \cdots, m$ و $j = ۰, ۱, ۲, \cdots, m$ توابع فقط از اختلاف زمان $(s-t)$ است. بنابراین بردار میانگین به‌صورت

$$E(\mathbf{X}_t) = \boldsymbol{\mu} = \begin{bmatrix} \mu_۱ \\ \mu_۲ \\ \vdots \\ \mu_m \end{bmatrix}, \tag{۱.۲}$$

و ماتریس کوواریانس به‌صورت

$$\begin{aligned}
\Gamma(k) &= \mathrm{Cov}(X_t, X_{t+k}) = E\big((X_t - \mu)(X_{t+k} - \mu)'\big), \\
&= E\begin{bmatrix} X_{۱,t} - \mu_۱ \\ X_{۲,t} - \mu_۲ \\ \vdots \\ X_{m,t} - \mu_m \end{bmatrix} [X_{۱,t+k} - \mu_۱, X_{۲,t+k} - \mu_۲, \cdots, X_{m,t+k} - \mu_m], \\
&= \begin{bmatrix} \gamma_{۱۱}(k) & \gamma_{۱۲}(k) & \cdots & \gamma_{۱m}(k) \\ \gamma_{۲۱}(k) & \gamma_{۲۲}(k) & \cdots & \gamma_{۲m}(k) \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{m۱} & \gamma_{m۲} & \cdots & \gamma_{mm} \end{bmatrix} = \mathrm{Cov}(X_{t-k}, X_t),
\end{aligned} \tag{۲.۲}$$

است که به ازای $k = ۰, \pm۱, \pm۲, \cdots$ و $i = ۱, ۲, \cdots, m$ داریم:

$$\begin{aligned}
\gamma_{ij}(k) &= E[(X_{i,t} - \mu_i)(X_{j,t+k} - \mu_j)], \\
&= E[(X_{i,t-k} - \mu_i)(X_{j,t} - \mu_j)].
\end{aligned}$$

به عنوان تابعی از k، $\Gamma(k)$ تابع ماتریس کوواریانس بردار فرآیند X_t نامیده می‌شود. برای $i = j$، $\gamma_{i,j}$ تابع کوواریانس برای مؤلفهٔ i ام فرآیند یعنی $X_{i,t}$ و $X_{j,t}$ بوده و برای $i \neq j$، $\gamma_{i,j}$ تابع کوواریانس متقابل بین $X_{i,t}$ و $X_{j,t}$ است. به سهولت می‌توان دید که $\Gamma(۰)$ ماتریس واریانس‌ـ کوواریانس فرآیند است. باید توجه کنیم که در فرآیند تواماً ایستا نتیجه می‌شود که هر مؤلفهٔ یک متغیر فرایند ایستاست. با وجود این بردار فرآیندهای ایستای یک متغیری، الزاماً یک فرآیند تواماً ایستای برداری نیست.

تابع ماتریس همبستگی برای فرآیند برداری به‌صورت

$$\rho(k) = D^{-\frac{۱}{۲}}\Gamma(k)D^{-\frac{۱}{۲}} = \big[\rho_{ij}(k)\big], \tag{۳.۲}$$

برای $i = ۱, \cdots, m$ و $j = ۱, \cdots, m$ تعریف می‌شود، که D ماتریس قطری است که در آن عنصر قطری iام واریانس فرآیند iام است، یعنی:

$$D = \text{diag}\left[\gamma_{۱۱}(0), \gamma_{۲۲}(0), \cdots, \gamma_{mm}(0)\right].$$

واضح است که عنصر قطری iام $\rho(k)$ تابع خودهمبستگی مولفه iام سری یعنی $\mathbf{X}_{i.t}$ است. در حالی که عنصر غیر قطری (i,j) $\rho(k)$، یعنی:

$$\rho_{ij}(k) = \frac{\gamma_{ij}(k)}{[\gamma_{ij}(0)\gamma_{ij}(0)]^{\frac{۱}{۲}}}, \tag{۴.۲}$$

است، که تابع خودهمبستگی متقابل بین $X_{i.t}$ و $X_{j.t}$ را نشان می دهد.

۲.۲ الگوهای اتورگرسیو میانگین متحرک برداری ناایستا

در تحلیل سری‌های زمانی، مشاهده سری‌هایی که رفتاری ناایستا را نشان می‌دهند، بسیار متداول است. یک راه مفید تبدیل سری‌های ناایستا به ایستا، تفاضلی کردن است. برای مثال، در سری‌های زمانی یک متغیری یک سری ناایستای $\mathbf{X}_t$ به یک سری ایستا $(۱ - B)^d\mathbf{X}_t$، $d > ۰$ تبدیل می شود، به‌طوری که می‌توان نوشت:

$$\phi_p(B)(1 - B)^d X_t = \theta_q(B)Z_t, \tag{۵.۲}$$

که در آن $\phi_p(B)$ یک عملگر اتورگرسیو [۲] (AR) ایستاست. چنین به نظر می‌رسد که تعمیم (۵.۲) فرآیند برداری:

$$\Phi_p(B)(\mathrm{I} - \mathrm{I}B)^d\mathbf{X}_t = \Theta_q(B)\mathbf{Z}_t, \tag{۶.۲}$$

است. حال از این تعمیم، نتیجه می‌شود که تمام مؤلفه‌های سری به دفعات مساوی تفاضلی شده است. روشن است که این محدودیت لازم نبوده و ناخواسته است. برای انعطاف پذیری بیشتر، فرض می کنیم که $\mathbf{X}_t$ ایستا نباشد ولی آن را با به کار بردن عملگر تفاضلی زیر می‌توان به ایستا تبدیل نمود:

$$D(B)\mathbf{X}_t, \tag{۷.۲}$$

که در آن

$$D(B) = \begin{bmatrix} (۱ - B)^{d_۱} & \cdots & \cdots & \cdots \\ \cdots & (۱ - B)^{d_۲} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & (۱ - B)^{d_m} \end{bmatrix}, \tag{۸.۲}$$

و $(d_۱, \cdots, d_m)$ مجموعه‌ای از اعدادصحیح نامنفی است. لذا الگوی [۳] (ARMA) برداری ناایستا به صورت

$$\Phi_p D(B)\mathbf{X}_t = \Theta_q(B)\mathbf{Z}_t, \tag{۹.۲}$$

[2] Autoregressive
[3] Autoregressive Moving Average

را داریم، که ریشه‌های صفر $|\Phi_p(B)|$ و$|\Theta(B)|$ خارج دایره واحدند. تفاضلی کردن سری‌های‌زمانی برداری بسیار پیچیده است و باید با دقت انجام شود. زیاد تفاضلی کردن هر مؤلفه سری، ممکن است منتهی به پیچیدگی‌هایی در برازش الگو شود. ثابت می‌شود که به طور یکسان تفاضلی کردن پیشنهاد شده برای یک فرآیند برداری، ممکن است نتیجه‌اش یک نمایش وارون‌پذیر نباشد. زیرا یک اریبی بزرگ در براورد پارامترها، ممکن است از یک الگوی وارون‌پذیر ناشی شود، که از زیاد تفاضلی کردن باید پرهیز شود (نیرومند (۱۳۹۱)). از این رو، تعمیم زیر برای الگوی ARMA برداری:

$$\Phi_p(B)\mathbf{X}_t = \Theta_q(B)\mathbf{Z}_t, \qquad (۱۰.۲)$$

پیشنهاد شده است (نیرومند ، ۱۳۹۱)، که صفرهای چندجمله‌ای دترمینانی $\Phi_p(B)$ رو یا خارج دایره واحدند. در ادامه مدل های ARMA (۹.۲) را در حالت ناایستا بررسی می کنیم.

۳.۲ مدل اتو رگرسیو میانگین متحرک تلفیق شده

مدل اتو رگرسیو میانگین متحرک تلفیق شده [۴] (ARIMA) که فرآیند ایستای حاصل از یک سری ناایستای همگن تفاضلی است. به‌طوری که سری تفاضلی شدهٔ $(۱-B)^d X_t$ از یک فرآیند ایستای $ARMA(p,q)$ پیروی می‌کند. بنابراین داریم:

$$\phi_p(B)(۱-B)^d X_t = \theta_۰ + \theta_q(B)Z(t), \qquad (۱۱.۲)$$

که در آن عملگر AR ایستای $\phi_p(B) = (۱ - \phi_۱ B - \cdots - \phi_p B^p)$ و عملگر MA وارون‌پذیر

$$\theta_q(B) = (۱ - \theta_۱ B - \cdots - \theta_q B^q),$$

دارای ریشهٔ مشترکی نیستند. الگوی ناایستای همگن حاصل در (۱۱.۲) را یک الگوی اتو رگرسیو تلفیق شده با مرتبهٔ (p,d,q) می‌نامند که به صورت $ARIMA(p,d,q)$ نشان داده می‌شود. وقتی $p = ۰$ الگوی $ARIMA(p,d,q)$ را یک الگوی میانگین متحرک تلفیق شده از مرتبهٔ (d,q) می‌نامند که به صورت الگوی $IMA(d,q)$ نشان داده می‌شود. $q = ۰$، فرآیند یک الگوی برداری AR(p) به‌صورت

$$\mathbf{X}_t = \Phi_۱\mathbf{X}_{t-۱} - \cdots - \Phi_p\mathbf{X}_{t-p} + \mathbf{Z}_t, \qquad (۱۲.۲)$$

می‌شود. فرآیند ایستاست هرگاه ریشه‌های چندجمله‌ای دترمینانی $\Phi_q(B)$ خارج دایره واحد باشند. برای اطلاعات بیشتر در مورد مدل‌های سری‌زمانی چندمتغیره (نحوه برازش مدل و پیشگویی) به نیرومند (۱۳۹۱) مراجعه شود. در ادامه به تحلیل اکتشافی داده‌ها می‌پردازیم و این مدل‌ها را روی آنها به‌کار می‌گیریم.

۳ تحلیل اکتشافی داده‌ها

در این بخش با استفاده از مدل‌های خطی و خطی تعمیم‌یافته به بررسی هر یک از متغیرها روی متغیر پاسخ می‌پردازیم. با این رویکرد که متغیرهای تاثیرگذار شناسایی شوند. داده‌ها برحسب مکان و زمان وقوع تصادف براساس اطلاعات حاصل از آمارهای رسمی و ثبتی پلیس راه و پزشکی قانونی استان قم جمع‌آوری شده‌اند. این داده‌ها شامل متغیرهای سال (year)، ماه (months)، موقعیت محورها (location)، تعداد تصادفات (Accidents)، تعداد مصدومین (Injured)، تعداد مرگ و میر(Fatalities)، تعداد وسیله نقلیه سواری‌ها در تصادف (Riding)، تعداد موتور سیکلت‌ها در تصادف (motorcycle)،

[۴] Autoregressive integrated Moving Average

عابر پیاده (pedestrian)، وسیله نقلیه باری (load)، اتوبوس (Bus)، عرض جغرافیایی محور تصادف (lat)، طول جغرافیایی محور تصادف (long) و نرخ مرگ و میر به تعداد تصادف (Rate) است.

در حالت اول متغیر تعداد مرگ و میرها را به عنوان پاسخ و بقیه متغیرها به عنوان متغیر کمکی، مستقل یا کمی در نظر گرفته‌ایم به این صورت است که متغیرهای سال، ماه و موقعیت محور به عنوان متغیرهای کیفی طبقه‌بندی شده را وارد مدل کرده‌ایم و متغیرهای تعداد تصادفات، تعداد مصدومان و ... به عنوان متغیرهای کمی در مدل لحاظ کرده‌ایم. خلاصه برازش مدل این بود که متغیرهای وسیله نقلیه سواری، باربری، اتوبوس و عابرپیاده درسطح ۵٪ معنی‌دار هستند. البته با ملاک اطلاع آکائیک (AIC) مدل نهایی به‌صورت

$$\log E(Fatalities) \sim Injured + Riding + motorcycle + pedestrian + load + Bus + \varepsilon_t$$

حاصل شد. هر چند شرایط مدل با بررسی خطاها برقرار نشد و لذا مدل فوق ناکارآمد می‌باشد. در حالت دوم متغیر نرخ مرگ و میر بر تعداد تصادفات را به عنوان متغیر پاسخ در نظر گرفتیم و همانند مدل قبلی متغیرهای سال، ماه و موقعیت محور به عنوان متغیر کیفی وارد مدل کردیم. خلاصه نتایج برازش مدل به این صورت بود که متغیرهای محور، ماه، انواع وسیله نقلیه و عابر پیاده در سطح ۵٪ معنی‌دار هستند. البته در این حالت نیز شرایط مدل با بررسی باقیمانده‌ها مورد تایید نیست. بنابراین با اطلاعاتی که کسب کردیم به این نتیجه رسیدیم که محورها را به تفکیک مورد بررسی قرار دهیم، لذا در حالت سوم به این مورد پرداختیم که تحلیل سری زمانی به تفکیک محورها با در نظرگرفتن متغیر تعداد مرگ و میر به عنوان متغیر پاسخ و متغیرهای کمکی تعداد تصادفات، تعداد مصدومین، انواع وسیله‌های نقلیه انجام دهیم. در واقع یک مدل خطی تعمیم‌یافته با خطاهای سری زمانی $ARMA$ به داده‌ها برازش دهیم. مدل کلی (M Full) بدون تفکیک محورها به صورت

$$\log E(Fatalities) \sim Acciident + Injured + Riding + motorcycle + pedestrian + load + Bus + \varepsilon_t$$

است، که در آن خطای ε_t را به صورت یک مدل سری زمانی $ARMA$ در نظر می‌گیریم. با توجه به اینکه داده‌ها در چهار محور (شمال، جنوب، شرق، غرب) استان ثبت شده است برای هر محور مدل به طور جداگانه برازش می‌شود. با نمادهای، مدل محور شمال (M North)، مدل محور جنوب (M South)، مدل محور غرب (M West) و مدل محور شرق (M East) نمایش می‌دهیم. نتایج در جدول ۱ خلاصه شده است: در جدول ۱ مقادیر برآورد ضرایب مدل (Est) و pـ مقدار آنها

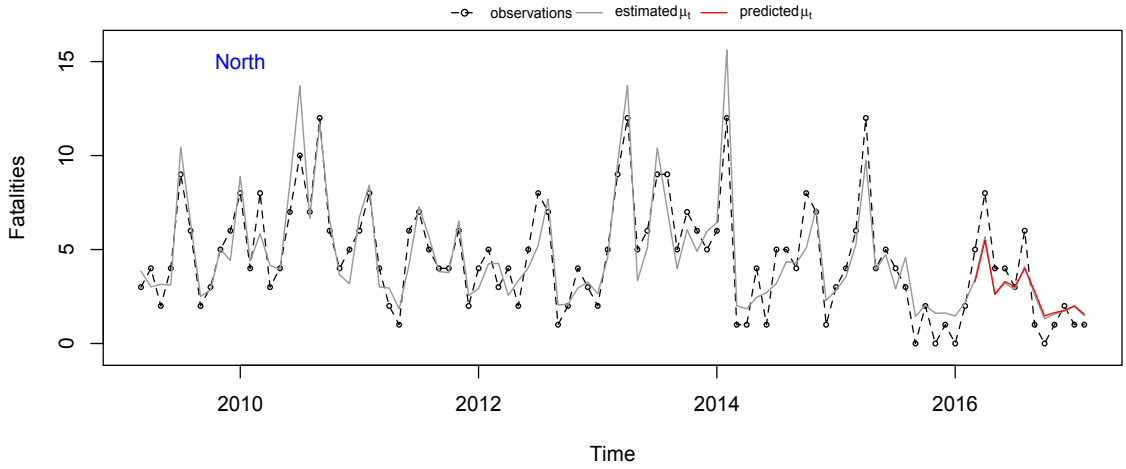
جدول ۱: نتایج برآورد پارامترهای مدل و pـ مقدار برای مدل کلی و به تفکیک چهار محور جاده‌ای قم

	M . Full		M . North		M . South		M . West		M . East	
	Est	p	Est	p	Est	p	Est	p	Est	p
$Intercept$	۰/۶۸۹۱	۰/۰۰	۰/۱۰۱۷	۰/۷۹	۰/۷۸۷۹	۰/۰۲	۰/۳۶۶	۰/۲۰	۰/۱۰۷	۰/۷۹
$Accident$	−۰/۰۰۱۵	۰/۵۳	۰/۰۰۴۶	۰/۴۶	۰/۰۰۴۸	۰/۴۹	−۰/۰۰۱	۰/۸۶	۰/۰۰۲۹	۰/۷۷
$Injured$	۰/۰۰۸۵	۰/۱۲	۰/۱۱۸	۰/۲۵	۰/۰۰۰۸	۰/۹۴	۰/۰۱۱	۰/۳۵	۰/۰۰۹۱	۰/۵۰
$Riding$	۰/۱۸۵۶	۰/۰۰	۰/۲۲۶۶	۰/۰۰	۰/۱۴۱۶	۰/۰۰	۰/۲۶۳	۰/۰۰۷	۰/۲۹۶۲	۰/۰۰
$Motor$	۰/۱۳۳۲	۰/۱۵۸	۰/۱۴۳۹	۰/۴۶	۰/۰۵۳۸	۰/۷۰	۱/۱۸۷	۰/۰۲	۰/۰۹۱۴	۰/۷۱
$Pedes$	۰/۲۸۹۴	۰/۰۰۶	۰/۰۴۶۹	۰/۸۳	۰/۱۵۴۹	۰/۴۲	۰/۲۱۲	۰/۶۱	۰/۰۱	۰/۲۱۶
$Load$	۰/۱۷۰۴	۰/۰۰	۰/۱۳۷۹	۰/۱۲	۰/۱۷۶۷	۰/۰۱	۰/۸۴۱	۰/۳۶	۰/۰۱۵۴	۰/۰۱۵
Bus	۰/۲۳۷۲	۰/۰۰۲	۰/۰۱۸۲	۰/۹۳	۰/۰۷۹۶	۰/۵۹	۰/۱۷۱	۰/۲۹	۰/۳۲۳۶	۰/۱۶
MSE	۲/۸۰۱۳۱۱		۱/۵۶۲۸		۲/۰۸۴		۱/۲۶۶۵		۱/۲۲۲۰	
AIC	۱۴۵۵/۹۴۸		۳۰۶/۴۶		۳۹۴/۱۴۱		۳۵۹/۰۰۳۵		۳۴۱/۵۸	

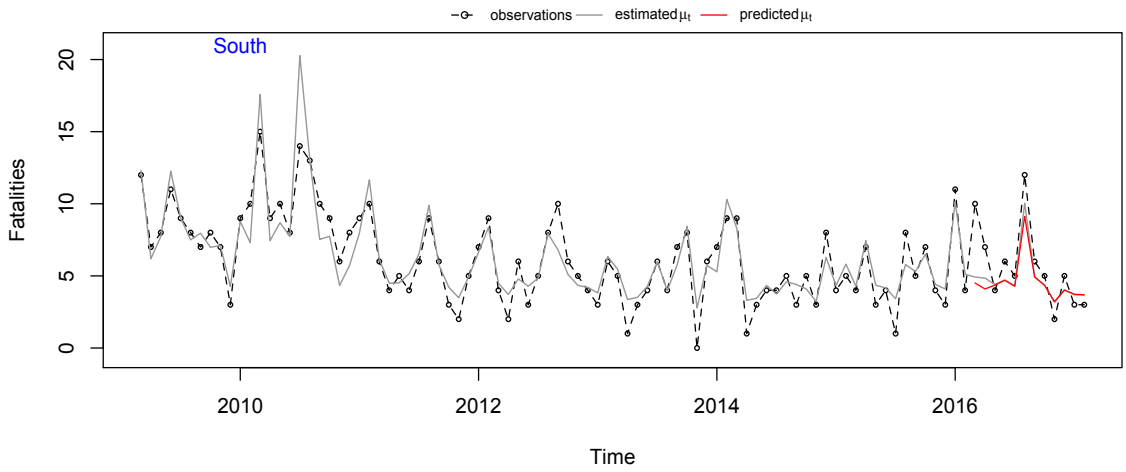
(p) برای مدل کلی و به تفکیک چهار محور ارائه شده است و همچنین مقدار MSE و AIC برای مقایسه بین مدل‌ها نیز

[5] Mean Square Error

محاسبه شده‌اند. می‌توان نتیجه گرفت که وقتی مدل کلی را بدون توجه به چهار محور اجرا کردیم هم مقدار MSE [۵] و هم AIC [۶] آن نسبت به حالت تفکیک شده خیلی بیشتر است یعنی اینکه بدون توجه اطلاع در مورد چهار محور دقت مدل پایین می‌آید. با توجه به نتایج برازش چهار محور در جدول ۱، یک نتیجه‌گیری واحد در مورد متغیرهای مورد نظر در مدل نمی‌توان گرفت. اما براساس مقادیر MSE و AIC می‌توان گفت که محور شرق MSE کمتر و محور شمال AIC کمتری دارند. شکل‌های ۱ تا ۴ نمودار سری زمانی مشاهدات، مقادیر برازش شده و مقادیر پیش‌بین برای سال ۲۰۱۶ را نشان می‌دهند. که بیانگر این است چگونه مدل به مشاهدات به خوبی برازش شده‌اند و همچنین پیش‌بینی آینده نیز با واقعیت تا حدودی هم‌خوانی دارد.

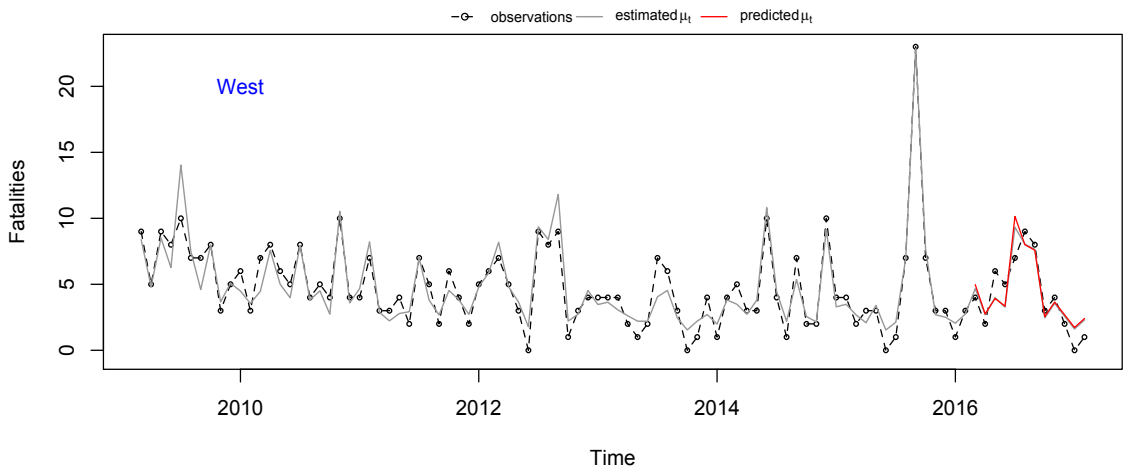


شکل ۱: سری زمانی تعداد مرگ و میرهای تصادفات (شمال): مقادیر برازش‌شده (خط‌چین)، مقادیر پیش‌بینی آینده (خط قرمز)

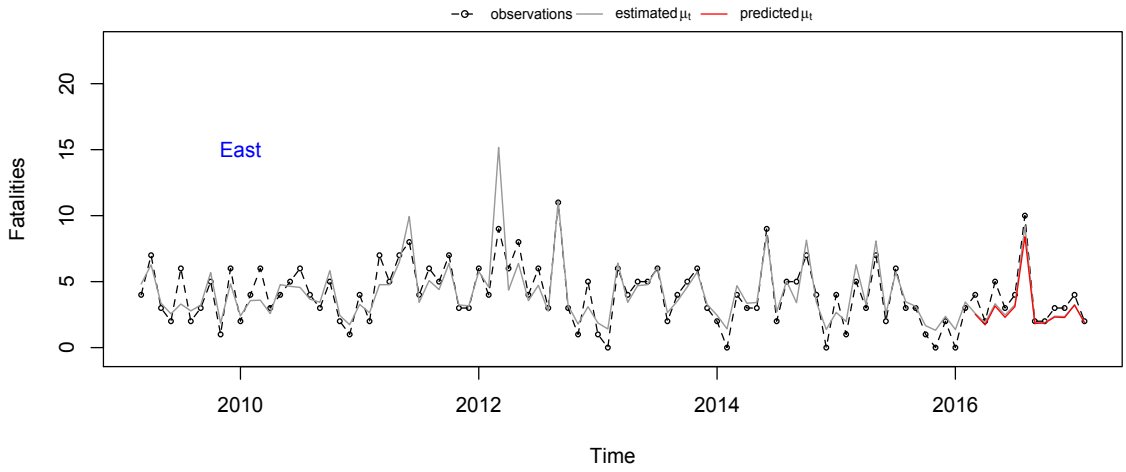


شکل ۲: سری زمانی تعداد مرگ و میرهای تصادفات (جنوب): مقادیر برازش‌شده (خط‌چین) و مقادیر پیش‌بینی آینده (خط قرمز)

[6] Akaike information criterion



شکل ۳: سری زمانی تعداد مرگ و میرهای تصادفات (غرب): مقادیر برازش‌شده (خط‌چین) و مقادیر پیش‌بینی آینده (خط قرمز)



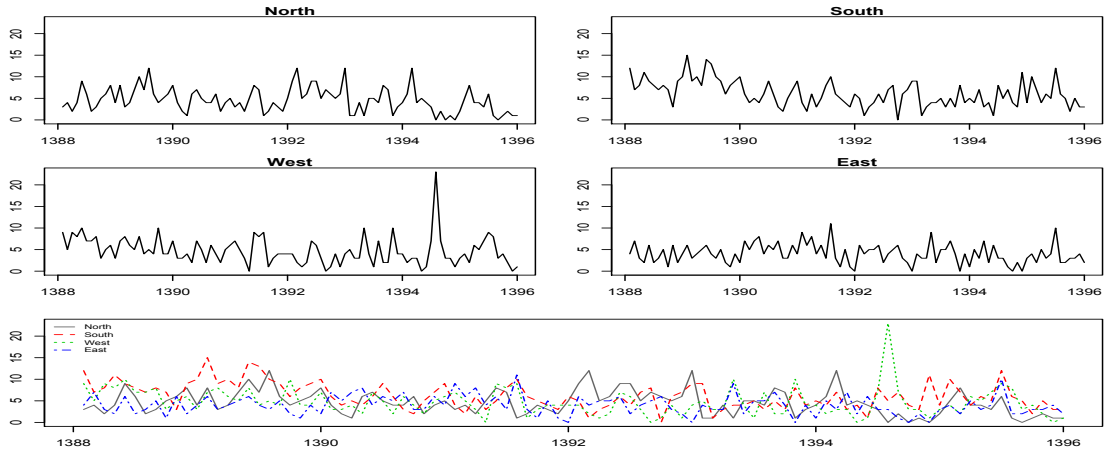
شکل ۴: سری زمانی تعداد مرگ و میرهای تصادفات (شرق): مقادیر برازش‌شده (خط‌چین) و مقادیر پیش‌بینی آینده (خط قرمز)

۱.۳ تحلیل سری زمانی چند متغیره داده‌های مرگ و میر تصادفات جاده‌ای استان قم

در این بخش به تحلیل سری‌زمانی چند متغیره روی داده‌های تعداد مرگ و میر تصادفات جاده‌ای می‌پردازیم. در اینجا هر محور را به عنوان یک متغیر پاسخ در نظر گرفته‌ایم،بنابراین یک متغیر پاسخ (چهار متغیره) برداری به صورت

$$Z_t = \begin{bmatrix} z_{۱t} \\ z_{۲t} \\ z_{۳t} \\ z_{۴t} \end{bmatrix}, \qquad t = ۱, ۲, \ldots, ۱۲ * ۸$$

داریم. نمودار سری زمانی تعداد مرگ و میرها برای چهار محور در شکل ۵ ارائه شده است. همانطور که از نمودار برداشت می‌شود.تعداد مرگ ومیرها در محور جنوب روند کاهشی دارد و در محور شرق تغییرات کمتر می‌باشد. تعداد مرگ و میرها



شکل ۵: نمودار سری زمانی تعداد مرگ و میرهای تصادفات در چهار محور شمال، جنوب، غرب و شرق

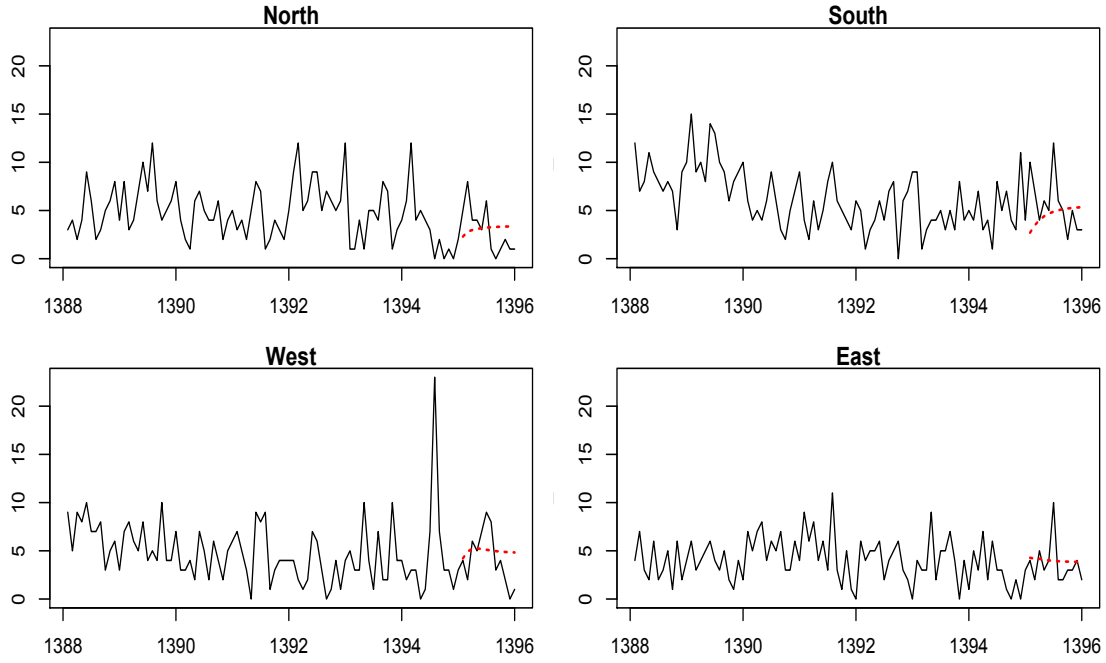
در محور جنوب از سال ۱۳۸۸ تا سال ۱۳۹۶ یک روند کاهشی مشاهده می‌شود و در محور شرق تغییرات کمتری نسبت به محورهای دیگر به نظر می‌رسد. بیشترین کشته‌ها در تصادفات مربوط به محور غرب استان در سال ۹۴ مشاهده می‌شود. مدل‌های سری زمانی چندمتغیره VARMA(p, q)) [۷] با مرتبه‌های متفاوت روی داده‌ها بوسیله نرم‌افزار R با بسته MTS پیاده‌سازی کردیم. نتایج حاصل با در نظرگرفتن معیار $RMSE$ و AIC انتخاب مدل‌ها در جدول ۲ خلاصه شده‌اند. با

جدول ۲: مقادیر AIC و $RMSE$ برای مدل $VARMA$ با مرتبه‌های مختلف

$Model$	AIC	$RMSE$
$VARMA(۰,۱)$	۷/۹۸۵۶	۲/۶۵۸۸
$VARMA(۱,۰)$	۷/۹۵۶۵	۲/۶۵۰۳
$VARMA(۱,۱)$	۷/۸۶۱۸	۲/۵۳۸۴
$VARMA(۲,۱)$	۸/۱۵۸۲	۲/۵۱۸۴
$VARMA(۱,۲)$	۸/۱۰۱۸	۲/۵۱۵۷
$VARMA(۲,۲)$	۸/۰۳۹۹	۲/۳۷۷۰

توجه به مقادیر $RMSE$ و AIC مدل $VARMA(۱,۱)$ را انتخاب کردیم چون هم AIC کمتری دارد و هم خلاصه‌تر از مدل $VARMA(۲,۲)$ از نظر تعداد پارامتر می‌باشد. شکل ۶ نمودار سری زمانی داده‌ها و مقادیر پیش‌بینی برای مدل $VARRMA(۱,۱)$ در ۱۲ ماه سال ۱۳۹۵ را نمایش می‌دهد. با توجه به شکل می‌توان نتیجه گرفت که پیش‌بینی به خوبی صورت نگرفته است. برای پیش‌بینی مناسب، بهتر است متغیرهای کمکی تعداد تصادفات، تعداد مصدومین و وسیله نقلیه سواری در مدل لحاظ شود. برای اینکار ابتدا حذف روند به وسیله یک مدل رگرسیونی چند متغیره خطی انجام داده‌ایم سپس مدل $VARMA(p,q)$ را برای داده‌های بدون روند برازش کرده‌ایم و نتایج $RMSE$ و AIC انتخاب مدل‌ها در جدول ۳ خلاصه شده‌اند. لازم به ذکر است که مدل‌های $VARMA$ با مرتبه همزمان p و q یعنی $VARMA(۱,۱)$، $VARMA(۲,۱)$ و $VARMA(۲,۲)$ از نظر محاسباتی همراه با سه متغیر کمکی قابل انجام نبود یعنی برآورد ماکسیمم

[7] Vector Autoregressive integrated Moving Average

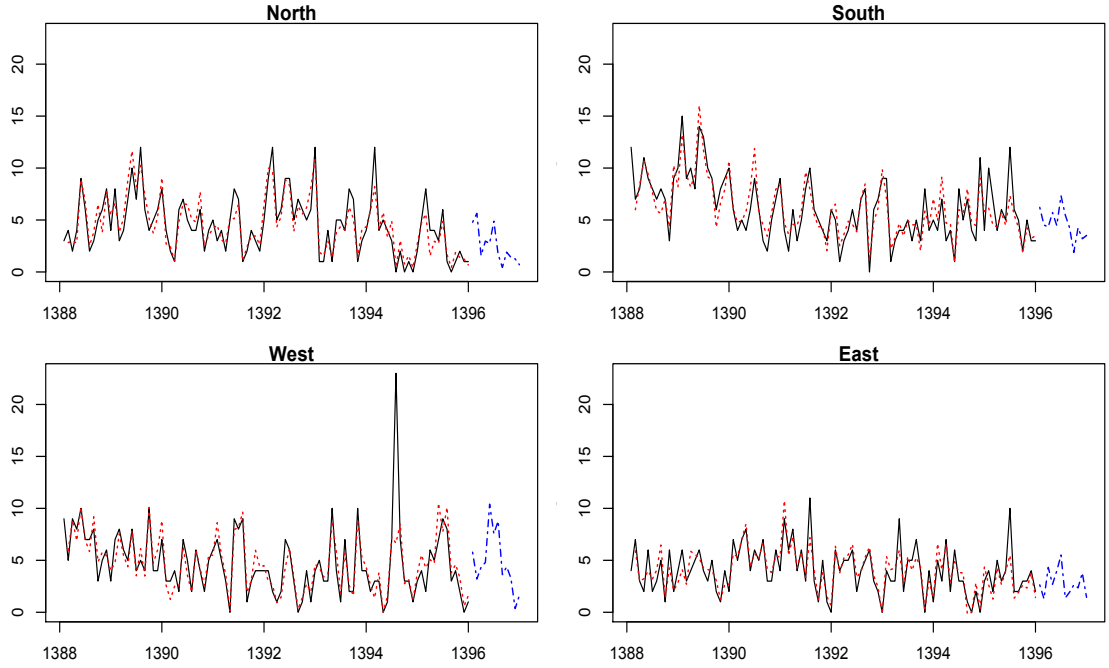


شکل ۶: نمودار سری داده‌های مرگ و میر به همراه مقادیر پیش‌بینی در ۱۲ ماه سال ۱۳۹۶ (خط‌چین)

جدول ۳: مقادیر AIC و $RMSE$ برای مدل بدون روند $VARMA(p,q)$

$Model$	AIC	$RMSE$
$VARMA(۰,۱)$	۲/۷۰۰۲	۱/۴۲۶۸
$VARMA(۱,۰)$	۲/۷۰۰۱	۱/۴۲۴۷
$VARMA(۰,۲)$	۲/۸۹۲۲	۱/۴۰۴۶
$VARMA(۲,۰)$	۲/۹۱۱۷	۱/۴۰۵۴

درستنمایی پارامترها به روش عددی قابل محاسبه نیست، زیرا تعداد پارامترها خیلی زیاد است. اما با دو متغیر کمکی،تعداد تصادفات و تعداد مصدومان قابل انجام بود ولی مقادیر AIC آنها بیشتر از عدد هفت شد و عملاً قابل رقابت با مدل‌های جدول ۳ نیستند، بنابراین مناسب‌ترین مدل با توجه به نتایج جدول ۳ مدل $VARMA(۱,۰)$ است. شکل ۷ نمودار سری زمانی داده‌ها به همراه مقادیر برازش شده سری از مدل $VARMA(۱,۰)$ و مقادیر پیش‌بینی برای ۱۲ ماه سال ۱۳۹۶ را نمایش می‌دهد و حاکی از برازش مناسب $VARMA(۱,۰)$ بدون روند نسبت به مدل $VARMA(۱,۱)$ شکل ۶ است. اکنون که مناسب‌ترین مدل ازبین مدل‌ها و روش‌های پیشنهادی مشخص گردید،می‌توان براساس پیش‌بینی برای تعدادمرگ ومیرماه‌های آتی بادقت بیشتری انجام داد البته لازم به ذکراست همه این پیش‌بینی‌هافقط زمانی قابل اتکاهستندکه همان شرایطی که برای سال‌های گذشته بوده است نیزبرای سال‌های آتی برقرارباشد. دربخش بعدموقعیت محورها رابه‌طورمستقیم به‌وسیله مدل‌های فضایی_زمانی درنظرگرفته‌ایم و نتایج را با هم مورد مقایسه قرار داده‌ایم.



شکل ۷: نمودار سری زمانی مشاهدات به همراه مقادیر برازش شده(خط چین) و مقادیر پیش‌بینی سال ۱۳۹۶

۴ مدل‌های فضایی ـ زمانی

فرض کنید $z_\ell(s_i,t)$ مقادیر مشاهدات پاسخ در موقعیت s_i، $i=1,\ldots,n$ و ماه t ام در سال ℓ ام برای $\ell=1,\ldots,r$ و $t=1,\ldots,T_\ell$ باشد. در این‌جا دو اندیس زمان به‌صورت ماه درون سال مشخص شده است که می‌تواند به‌صورت روز درون سال باشد. قرار دهید $\boldsymbol{z}_{\ell t}=(z_\ell(s_1,t),\ldots,z_\ell(s_n,t))'$ و فرض کنید $x_{\ell j}(s,t)$ متغیر کمکی jام در ماه tام سال ℓام باشد و قرار دهید $X_{\ell t}$ ماتریس $n\times p$ از متغیرهای که شامل عرض از مبدأ است. گلفند (۲۰۱۲) و گلفند و همکاران (۲۰۰۵) مدل‌های فضایی ـ زمانی بیزی را به‌صورت سلسله‌مراتبی در سه مرحله بیان کرد: ۱ـ مشاهدات به شرط یک فرآیند تصادفی و پارامترها $(\boldsymbol{z}|\boldsymbol{u},\boldsymbol{\theta})$، ۲ـ فرآیند تصادفی به شرط پارامتر $(\boldsymbol{u}|\boldsymbol{\theta})$. ۳ـ پارامترها $(\boldsymbol{\theta})$. در مرحله اول مدل به‌صورت

$$\boldsymbol{z}_{\ell t}=\boldsymbol{u}_{\ell t}+\boldsymbol{\epsilon}_{\ell t},\quad \ell=1,\ldots,r,\ t=1,\ldots,T,$$

در نظر گرفته می‌شود، که در آن $\boldsymbol{\epsilon}_{\ell t}=(\varepsilon_\ell(s_1,t),\ldots,\varepsilon_\ell(s_n,t))'$ بردار خطاهای مدل با توزیع $N(0,\sigma_\varepsilon^2 I_n)$ است و σ_ε^2 اثر قطعه‌ای یا واریانس خطای تصادفی است. $\boldsymbol{u}_{\ell t}=(u_\ell(s_1,t),\ldots,u_\ell(s_n,t))'$ مقادیر واقعی متناظر با $z_\ell(s_i,t)$ که شامل مولفه‌های $\boldsymbol{u}_{\ell t}=X_{\ell t}\boldsymbol{\beta}+\boldsymbol{\eta}_{\ell t}$ است، که در آن $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)'$ بردار پارامترهای ضرایب رگرسیونی، $\boldsymbol{\eta}_{\ell t}=(\eta_\ell(s_1,t),\ldots,\eta_\ell(s_n,t))'$ اثر تصادفی فضایی ـ زمانی مدل است که دارای توزیع $N(0,\Sigma_\eta)$ و مستقل از زمان می‌باشد، که در آن $\Sigma_\eta=\sigma_\eta^2 S_\eta$، σ_η^2 پارامتر واریانس فضایی و S_η ماتریس همبستگی فضایی است. در این‌جا از تابع همبستگی ماترن (ماترن ، ۱۹۸۶) به‌صورت

$$K(s_i,s_j;\varphi,\nu)=\frac{1}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}\parallel s_i-s_j\parallel\varphi)^\nu K_\nu(2\sqrt{\nu}\parallel s_i-s_j\parallel\varphi)$$

استفاده شده است، که در آن K_ν تابع بسل اصلاح شده نوع دوم با مرتبه ν و $\parallel s_i-s_j\parallel$ فاصله بین موقعیت‌های s_i و s_j است. φ پارامتر دامنه فضایی و ν پارامتر هموارساز میدان تصادفی است، (کرسی ، ۱۳۹۳؛ کرسی و ویکل ، ۲۰۱۱).

مدل سلسله‌مراتبی ذکر شده را مدل فرآیند گاوسی (GP) مستقل می‌گویند. فرض کنید $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_\eta^2, \varphi, \nu)$ بردار پارامترهای مدل، $z_{\ell t}^{obs}$ بردار داده‌های مشاهده شده و $z_{\ell t}^*$ بردار داده‌های گمشده باشد. با در نظر گرفتن توزیع‌های پیشین معمول برای پارامترهای مدل ؛$\pi(\boldsymbol{\theta})$، لگاریتم توزیع پسین توأم پارامترها و داده‌های گمشده برای مدل GP به‌صورت

$$\begin{aligned}\log \pi(\boldsymbol{\theta}, \boldsymbol{u}_{\ell t}, z_{\ell t}^* | z_{\ell t}^{obs}) \propto & -\frac{N}{2}\log\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_{\ell=1}^{r}\sum_{t=1}^{T_\ell}(\boldsymbol{z}_{\ell t}-\boldsymbol{u}_{\ell t})'(\boldsymbol{z}_{\ell t}-\boldsymbol{u}_{\ell t}) \\ & - \frac{\sum_{\ell=1}^{r} T_\ell}{2}\log \mid \sigma_\eta^2 S_\eta \mid \\ & - \frac{1}{2\sigma_\eta^2}\sum_{\ell=1}^{r}\sum_{t=1}^{T_\ell}(\boldsymbol{u}_{\ell t}-\boldsymbol{X}_{\ell t}\boldsymbol{\beta})' S_\eta^{-1}(\boldsymbol{u}_{\ell t}-\boldsymbol{X}_{\ell t}\boldsymbol{\beta}) + \log\pi(\boldsymbol{\theta})\end{aligned}$$

به‌دست می‌آید. مدل سلسله مراتبی اتوررگرسیو (AR) به‌صورت

$$\begin{aligned}\boldsymbol{z}_{\ell t} &= \boldsymbol{u}_{\ell t} + \varepsilon_{\ell t}, \\ \boldsymbol{u}_{\ell t} &= \rho \boldsymbol{u}_{\ell,t-1} + X_{\ell t}\boldsymbol{\beta} + \boldsymbol{\eta}_{\ell t},\end{aligned}$$

معرفی می‌شوند، که در آن $\varepsilon_{\ell t}$ و $\boldsymbol{\eta}_{\ell t}$ همانند مدل GP نرمال فرض می‌شوند و ρ پارامتر همبستگی زمانی بین $(-1, 1)$ است. اگر $\rho = 0$ باشد، مدل GP حاصل می‌شود. در این مدل مقادیر شروع $\boldsymbol{u}_{\ell 0}$ برای هر $\ell = 1, \ldots, r$ از توزیع $N(\boldsymbol{\mu}_\ell, \sigma_\ell^2 S_0)$ مشخص می‌شود، که در آن S_0 ماتریس همبستگی فضایی حاصل از تابع همبستگی ماترن با پارامترهای φ و ν همانند $\boldsymbol{\eta}_{\ell t}$ است. بنابراین همه پارامترهای مدل به‌صورت $\boldsymbol{\theta} = (\boldsymbol{\beta}, \rho, \sigma_\varepsilon^2, \sigma_\eta^2\varphi, \nu, \boldsymbol{\mu}_\ell, \sigma_\ell^2)$ می‌باشند. با در نظر گرفتن پیشین معمول $\pi(\boldsymbol{\theta})$ برای پارامترها و با لگاریتم گرفتن توزیع پسین توأم پارامترها و مقادیر گمشده به‌صورت

$$\begin{aligned}\log \pi(\boldsymbol{\theta}, \boldsymbol{u}_{\ell t}, z_{\ell t}^* | z_{\ell t}^{obs}) \propto & -\frac{N}{2}\log\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_{\ell=1}^{r}\sum_{t=1}^{T_\ell}(\boldsymbol{z}_{\ell t}-\boldsymbol{u}_{\ell t})'(\boldsymbol{z}_{\ell t}-\boldsymbol{u}_{\ell t}) \\ & - \frac{1}{2}\log \mid \sigma_\ell^2 S_0 \mid - \frac{1}{2\sigma_\eta^2}\sum_{\ell=1}^{r}\sum_{t=1}^{T_\ell}(\boldsymbol{u}_{\ell 0}-\boldsymbol{X}_{\ell t}\boldsymbol{\beta})' S_\eta^{-1}(\boldsymbol{u}_{\ell 0}-\boldsymbol{X}_{\ell t}\boldsymbol{\beta}) \\ & - \frac{\sum_{\ell=1}^{r} T_\ell}{2}\log \mid \sigma_\eta^2 S_\eta \mid \\ & - \frac{1}{2\sigma_\eta^2}\sum_{\ell=1}^{r}\sum_{t=1}^{T_\ell}(\boldsymbol{u}_{\ell t}-\boldsymbol{\mu}_\ell)' S_0^{-1}(\boldsymbol{u}_{\ell t}-\boldsymbol{\mu}_\ell) + \log\pi(\boldsymbol{\theta})\end{aligned}$$

به‌دست می‌آید (برای اطلاعات بیشتر به مراجع فینلی و همکاران (۲۰۰۹)، فینلی و همکاران (۲۰۱۲) مراجعه شود). در ادامه این دو مدل را روی داده‌های تلفات جاده ای استان قم برازش و مورد مقایسه قرار می‌دهیم.

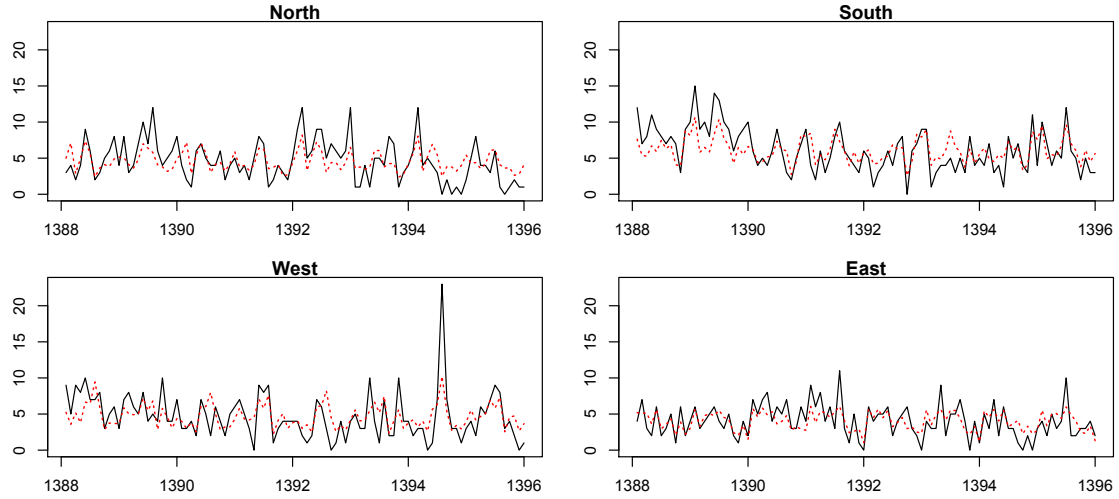
۱.۴ نتیجه تحلیل فضایی زمانی

لازم به ذکر است که موقعیت مکانی در مدل فضایی‌ـ زمانی برای چهار محور در دسترس بود، لذا از نظر مکانی تعداد کمی موقعیت جغرافیایی وجود دارد و با توجه به اینکه از نظر زمانی ۹۶ زمان است، بنابراین از نظر مکانی این مدل کارایی لازم را نسبت به زمان ندارد و این در مقدار $RMSE$ برای مدل فضایی ـ زمانی و مدل‌های سری زمانی چند متغیره ارائه شده در بخش ۳ کاملاً مشهود است. اگر بتوان در محورهای بیشتری با تفکیک ریزتری از نظر طول و عرض جغرافیایی مشاهدات را جمع‌آوری کرد، حتماً مدل‌های فضایی‌ـزمانی ارائه شده در این بخش هم از نظر زمان اجرا محاسبات و هم قابل انجام بودن مدل مناسب‌تر از مدل‌های سری زمانی چند متغیره است.

جدول ۴: نتایج برازش مدل‌های فضایی زمانی AR و GP

	مدل GP			مدل AR		
	برآورد	انحراف معیار	فاصله اطمینان	برآورد	انحراف معیار	فاصله اطمینان
$Intercept$	$-$۰/۸۳۹۳	۰/۶۳۷۳	$(-$۲/۰۶, ۰/۴۴$)$	$-$۰/۸۲۵۱	۰/۶۴۱۴	$(-$۲/۰۷۵, ۰/۴۳۰$)$
$Accident$	۰/۰۲۰۹	۰/۰۱۰۴	$($۰/۰۰۰۵, ۰/۰۴۰$)$	۰/۰۲۱۳	۰/۰۱۰۵	$($۰/۰۰۰۸, ۰/۰۴۱$)$
$Injured$	۰/۰۲۲	۰/۰۲۰۶	$(-$۰/۰۱۸, ۰/۰۶۲$)$	۰/۰۲۱۹	۰/۰۲۰۱	$(-$۰/۰۱۸۱, ۰/۰۶۱$)$
$Riding$	۱/۲۶۵۹	۰/۰۴۳۶	$($۱/۱۷۸, ۱/۳۵$)$	۱/۲۷	۰/۰۴۴۶	$($۱/۱۸۶, ۱/۳۶۰$)$
σ^2_ε	۰/۰۱۰۱	۰/۰۰۰۸	$($۰/۰۰۸۶, ۰/۰۱۱۸$)$	۰/۰۱۰۲	۰/۰۰۰۸	$($۰/۰۰۸۶, ۰/۰۱۱۸$)$
σ^2_η	۵/۸۵۶۸	۰/۴۲۱۵	$($۵/۰۹۴, ۶/۷۴$)$	۵/۸۸۲	۰/۴۳۷	$($۵/۰۹۲, ۶/۸۰۴$)$
φ	۰/۰۱	۰/۰۰۰۱	$($۰/۰۱, ۰/۰۲$)$	۰/۰۱۱	۰/۰۰۰۱	$($۰/۰۱, ۰/۲۱$)$
ρ	$-$	$-$	$-$	$-$۰/۰۱۲۵	۰/۰۱۶۴	$(-$۰/۰۴۵, ۰/۰۱۸$)$
$RMSE$	۲/۱۳۲۱۹۸			۲/۱۲۱۸۴۲		

شکل ۸ نمودار سری زمانی که همان مرگ و میر به همراه برازش‌های مدل فضایی زمانی AR برای چهار محور استان قم را نمایش می‌دهد که خوبی برازش مدل AR به داده‌های واقعی را نشان می‌دهد.



شکل ۸: نمودار سری زمانی داده‌های مرگ و یر به همراه برازش مدل AR (خطچین) برای چهار محور استان قم

در اینجا متغیر پاسخ را تعداد مرگ و میر و متغیرهای کمکی را تعداد تصادفات، تعداد مجروحین و تعداد وسیله نقلیه سواری در نظر گرفته‌ایم. نتایج برازش مدل برای ۱۰۰۰۰ تکرار با مقادیر ذخیره یا سوخته ۵۰۰۰ در جدول ۴ خلاصه شده‌اند. با توجه به نتایج $RMSE$ جدول، مدل‌های فضایی زمانی ذکر شده تفاوت چندانی ندارند. با توجه به نتایج فاصله اطمینان ۹۵٪ برای جدول ۴، در بین متغیرهای کمکی هر دو مدل معنی‌داری دو متغیر کمکی تعداد تصادفات و وسیله نقلیه سواری را نشان می‌دهند (چون عدد صفر داخل فاصله اطمینان قرار ندارد).

۵ بحث و نتیجه‌گیری

با توجه به ماهیت داده‌های مورد مطالعه که به زمان و مکان وابسته هستند، در این مقاله به تحلیل زمانی و فضایی‌ـ زمانی داده‌ها پرداختیم. در یک تحلیل اکتشافی بر اساس مدل‌های خطی و خطی تعمیم یافته، متغیرهای کمکی تاثیر گذار بر نرخ

مرگ و میر تصادفات شناسایی شدند. سپس براساس نتایج حاصل از مدل‌های سری‌زمانی چندمتغیره، بهترین مدل براساس ملاک‌های MSE و AIC برای)VARMA(۱،۱ است و در مدل‌های فضایی_زمانی بهترین مدل، مدلAR رتبه اول انتخاب گردید. در مقایسه کلی بین مدل‌های سری‌زمانی چندمتغیره و مدل‌های فضایی_زمانی، مدل‌های فضایی_زمانی عملکرد اجرایی محاسباتی مناسب‌تری برای این داده‌ها داشتند.

مراجع

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

والی، ح. (۱۳۹۲). تحقیق و بررسی علت وقوع حادثه درحوزه ۳۰کیلومتری شهرها، انتشارات یاقوت.

ویلیام دبلیو.اس.وی. ترجمه نیرومند ح ع. ،(۱۳۹۱) تحلیل سری‌های زمانی، روش‌های یک متغیری و چندمتغیری، انتشارات دانشگاه فردوسی مشهد.

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

Cressie, NAC, and Wikle CK., (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, New York.

Finley AO., Sang H., BANERJEE S.,and Gelfand A., (2009), Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Anlaysis*, 53, 2873–2884.

FINLEY, A., Banerjee, S., and Gelfand, A., (2012), Bayesian Dynamic Modeling for Large Space-Time Datasets Using Gaussian Predictive Processes. *Journal of Geographical Systems*, 14(1), 29-47.

Gelfand, A., Banerjee, S., and Gamerman, D., (2005). Univariate and Multivariate Dynamic Spatial Modelling. *Environmetrics*, 16, 465-479.

Gelfand, A.E., (2012), Hierarchical Modeling for Spatial Data Problems. *Spatial Statistics*, 1, 30-39.

Matérn B., (1986), *Spatial Variation*. 2nd edition. Springer-Verlag, Berlin.

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸
3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

مروری بر طرح‌های نمونه‌گیری فضایی ساخته شده از دنباله هالتون

حسین ویسی‌پور[1]، محمد مرادی
گروه آمار، دانشگاه رازی

چکیده: در این مقاله به مروری بر کاربرد طرح‌های نمونه‌گیری پذیرش متعادل، نمونه‌گیری پذیرش متعادل اصلاح‌شده و نمونه‌گیری افرازشده تکراری هالتون می‌پردازیم. در این طرح‌ها، واحدهای نمونه با استفاده از دنباله هالتون استخراج می‌شوند. این طرح‌ها در جوامع پیوسته و گسسته چند بعدی مورد استفاده قرار می‌گیرند و واحدهای نمونه به وسیله این روش‌ها به هر دو صورت بااحتمال برابر و احتمال نابرابر قابل استخراج می‌باشند.

واژه‌های کلیدی: احتمال شمول، برآوردگرهورویتز_تامپسون، به‌پخشی، دنباله هالتون، طرح نمونه‌گیری متعادل فضایی، همبستگی فضایی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62N01.

۱ مقدمه

در بررسی جوامع طبیعی و فضایی اطلاع از موقعیت واحدهای جامعه بسیار مهم است. در چنین حالتی واحدها بر اساس موقعیت جغرافیایی اندیس گذاری می‌شوند. در بررسی این جوامع واحدهای جامعه بر اساس مختصاتشان انتخاب می‌شوند. طرح های نمونه گیری که واحدهای نمونه را بوسیله انتخاب موقعیتشان استخراج می کنند را طرح‌های نمونه‌گیری فضایی می‌گویند. در جوامع فضایی معمولا واحدهای نزدیک به هم تحت تاثیر همدیگر و عوامل مشترک همبسته می باشند. در نظر گرفتن این وابستگی از طریق انتخاب مکان‌های واحد جامعه به صورت نمونه می‌باشد (استیونس و اولسن ۲۰۰۴). استفاده از اطلاعات مکانی واحدهای نمونه باعث می‌شود که استنباط‌های دقیق‌تری در مورد کل فضای جامعه به‌دست آید (مک دونالد ۲۰۱۴) . در صورت وجود همبستگی فضایی، هنگامی که یک واحد از نقاط همسایه انتخاب شود، انتخاب سایر نقاط همسایه کمک زیادی به شناسایی جامعه هدف نمی‌کند (هدایت، رائو واستوفکن ۱۹۸۸). . بنابراین در چنین حالتی

[1] نام و ایمیل ارائه دهنده مقاله: حسین ویسی‌پور، h.veisipour@gmail.com

طرح های نمونه گیری نیاز است که شانس انتخاب همزمان واحدهای نزدیک به هم کم و شانس انتخاب واحدهای همزمان واحدهای دورتر زیاد باشد. و طرح‌های در اولویت قرار می‌گیرند که واحدهای انتخاب شده در نمونه به خوبی در فضای جامعه پخش شوند. که استیونس و اولسن (۲۰۰۴) این طرح‌ها را طرح‌های فضایی متعادل نامیدند. معروفترین طرح‌های نمونه‌گری فضایی متعادل عبارتند از روش موزاییک بندی تصادفی (RTS) [۱] که توسط اورتون و استهمن (۱۹۹۳) معرفی گردید که یک طرح طبقه‌بندی تصادفی دو مرحله می‌باشد. در مرحله اول یک موزاییک‌بندی با شبکه منظم به‌طور تصادفی بر روی دامنه واقع می‌شود و سپس در مرحله دوم یک نقطه تصادفی درون هرسلول موزاییک انتخاب می‌گردد. استیونس و اولسن (۲۰۰۴) طرح موزاییک‌بندی تصادفی تعمیم‌یافته ($GRTS$) [۲] را معرفی کردند که فضای دوبعدی را به یک‌بعدی تبدیل می‌کند و سپس از روش سیستماتیک یک بعدی با احتمال نابرابر استفاده کرده و واحدهای نمونه را انتخاب کردند. که انتخاب نمونه‌ها به‌صورت احتمال نابرابر و متناسب با طول نیز می‌تواند انجام گردد. باندسون وتربورن (۲۰۰۸) نمونه‌گیری پواسن همبسته (CPS) [۳] را برای انتخاب با احتمال نابرابر معرفی کردند. که به‌صورت مرحله‌ای اجراشده و در هر مرحله احتمال شمول اصلاح می‌گردد. گرفستروم (۲۰۱۲) نمونه‌گیری پواسن همبسته فضایی (SCPS) [۴] را با ادغام کردن یک اندازه از فاصله‌ی واحدها در انتخاب کردن واحدها معرفی کرد که تعمیم روش(CPS) بود. گرفستروم و همکاران (۲۰۱۲) دو روش محلی محوری (LPM۱, [۵]LPM۲) را معرفی کرند، این روش‌ها تعمیم روش محوری (PM) می‌باشند که توسط دویل و تایلی (۲۰۱۲) معرفی شده است. این طرح‌ها برای انتخاب واحدهای نمونه πps می‌باشند که نمونه در k مرحله انتخاب می‌گردد، در هر مرحله احتمال شمول برای هر دو واحد اصلاح می‌گردد و تا زمانی که احتمال شمول به یک یا صفر برسد این مراحل ادامه دارد. سه تا از جدیدترین روشهای نمونه‌گیری که از دنباله هالتون ساخته شده اند عبارتند از نمونه گیری فضایی متعادل پذیرفتنی (BAS) [۶] رابرتسون و همکاران (۲۰۱۳)، نمونه گیری فضایی متعادل پذیرفتنی اصلاح شده رابرتسون و همکاران (۲۰۱۷) و نمونه گیری فضایی افزازشده تکراری هالتون (HIP) [۷] که رابرتسون و همکاران (۲۰۱۸) آن را معرفی کردند.

در این مقاله به مروری بر مقالاتی می‌پردازیم که طرح‌های BAS، BAS اصلاح شده و HIP را معرفی کرده‌اند. در بخش دوم دنباله هالتون و خواص آن برگرفته از مقاله‌های هالتون (۱۹۶۰) و پرایس و پرایس (۲۰۱۲) را بیان می‌کنیم. در بخش سوم مقاله‌های ارائه شده توسط رابرتسون و همکاران (۲۰۱۳، ۲۰۱۷ و ۲۰۱۸) که به معرفی طرح‌های نمونه گیری BAS، BAS اصلاح شده و HIP و خواص آنها پرداخته اند را بررسی و نتایج را بیان می‌کنیم.

۲ دنباله هالتون

هالتون (۱۹۶۰) بردار d بعدی $\{x_k\}_{k=۰}^{\infty}$ در $[۰,۱]^d$ را معرفی کرد که هر مولفه آن یک دنباله از اعداد شبه تصادفی است که بطور یکنواخت در بازه $[۰,۱]$ توزیع می‌شود. مولفه i ام از هر نقطه دنباله به پایه bi مرتبط است که $b_۲ = ۳, b_۱ = ۲$ و b_J برابر با j امین عدد اول طبیعی است. مقدار مولفه i از نقطه k ام این دنباله به صورت

$$x_k^{(i)} = \phi_{b_i}(k) = \sum_{j=۰}^{\inf} \{\lfloor \frac{k}{b_i^j} \rfloor \mod b_i\} \frac{۱}{b_i^{j+۱}} \qquad (۱.۲)$$

[1] Random Tessellation Stratified
[2] Generalized Random Tessellation Stratified
[3] Correlated Poisson Sampling
[4] Spatially Correlated Poisson Sampling
[5] Local Pivotal Method
[6] Balanced Acceptance Sampling
[7] Halton Iterative Partitioning

محاسبه می‌شود، که در آن $\lfloor x \rfloor$ تابع جز صحیح است که برابر بزرگترین عدد صحیح کمتر یا مساوی x می‌باشد و $Mod\ k$ b تابع باقیمانده است که باقیمانده تقسیم عدد k بر b را می‌دهد. به عنوان مثال، جمله سوم دنباله هالتون دوبعدی برابر $x_۳ = (\phi_۲(۳), \phi_۳(۳)) = (۳/۴, ۱/۹)$ است.

۱.۲ خواص دنباله هالتون

خاصیت ۱. هالتون (۱۹۶۰) و پرایس و پرایس (۲۰۱۲) ثابت کردند برای هر عدد صحیح غیر منفی J_i و $B = \prod_{i=۱}^{d} b_i^{J_i}$، از تعداد B نقطه متوالی از دنباله هالتون با پایه‌های b_i دقیقا یک نقطه در هر خانه تعریف شده زیر قرار می‌گیرد.

$$\prod_{i=۱}^{d} [m_i b_i^{-J_i}, (m_i + ۱) b_i^{-J_i}) \qquad (۲.۲)$$

که برای هر i، $b_i^{J_i} ... ۲, ۱, ۰ = m_i$ می‌باشد، که نشان می‌دهد دنباله هالتوان به طور یکنواخت در بازه $[۰, ۱]^d$ پخش میی‌گردد.

رابرتسون و همکاران (۲۰۱۸) این خانه ها را جعبه های هالتون نامیدند و این حعبه ها را با اندیس‌های $\{۰, ۱, ..., B - ۱\}$ نامگذاری کردند. برای مطالعه بیشتر مقاله رابرتسون و همکاران (۲۰۱۸) را ببینید.

خاصیت ۲. پرایس و پرایس (۲۰۱۲) ثابت کردند، اگر X_k در یک خانه هالتون خاص قرار گیرد در آن صورت مقدار تابع باقیمانده k نسبت به B یکی از مقدارهای $\{۰, ۱, ..., B - ۱\}$ است، بنابراین نقاط $X_k k, X_{k+B}, X_{k+۲B,...}$ از دنباله هالتون در همان جعبه خاص قرار می‌گیرند که نشان می‌دهد دنباله هالتون یک دنبای شبه دوه‌ای با دوره تناوب B می‌باشد.

۳ طرح‌های نمونه‌گیری فضایی

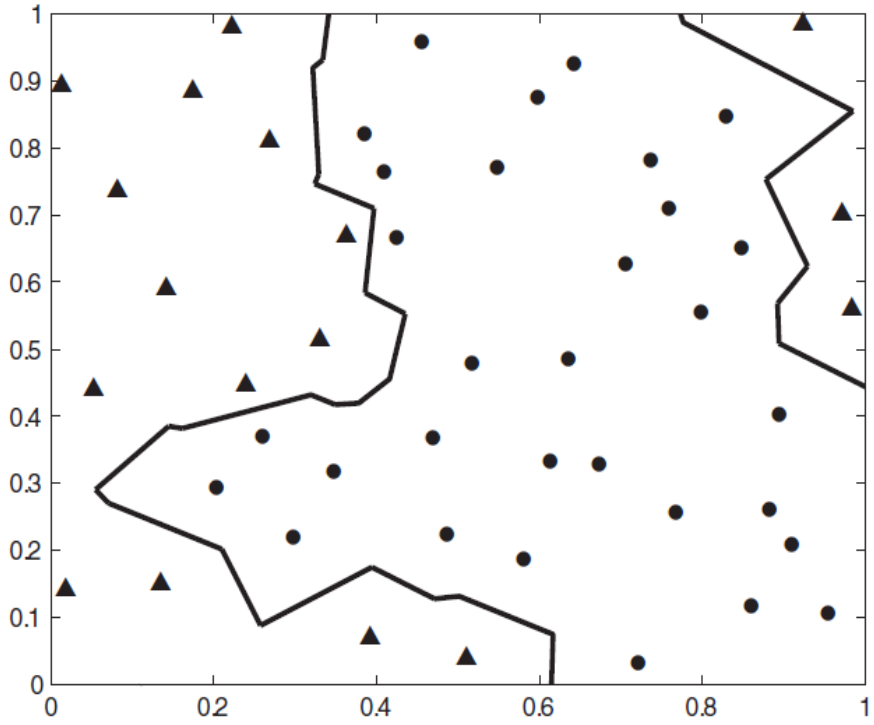
۱.۳ طرح نمونه‌گیری پذیرش متعادل

طرح نمونه‌گیری BAS توسط رابرتسون و همکاران (۲۰۱۳) معرفی گردید. که از دنباله شبه تصادفی هالتون با شروع تصادفی وانگ و هیکرنل (۲۰۰۰) برای به‌دست آوردن یک نمونه n تایی BAS استفاده کردند. یک نمونه ۳۰تای که به روش BAS انتخاب شده است در شکل ۱ نشان داده شده است. نقاط داخل ناحیه مورد بررسی با نماد ● و نقاط خارج ناحیه با نماد ▲ نمایش داده شده‌اند. تعداد ۴۷ نقطه متوالی از دنباله هالتون با شروع تصادفی در انتخاب نمونه استفاده شده‌اند، که ۱۷ نقطه آن جهت انتخاب واحد در نمونه پذیرفته نشده‌اند. آنها بیان کردند که این طرح دارای خواص یک طرح خوب مانند سادگی اجرا، به‌پخشی، قابل استفاده در فضای پیوسته و گسته با ابعاد بالا، توانایی تغییر در اندازه نمونه در هر مرحله از اجرای طرح، قابل استفاده در جوامع با شکل نامنظم و نیز با احتمال نابرار قابل استفاده می‌باشد.

با شبیه سازی واریانس برآوردگر هورویتز–تامپسون میانگین چامعه $\hat{\mu}_{HT} = \frac{۱}{n} \sum_{i=۱}^{n} \frac{y_i}{\pi_i}$، طرح BAS با طرح‌های نمونه‌گیری تصادفی ساده (SRS)[8] $GRTS$، LPM۱، LPM۲، $StrataSRS$[9] و $SPCS$ مشاهده کردند که این واریانس در نمونه BAS کمتر از نمونه‌های به‌دست آمده از طرح‌های دیگر است. همچنین با شبیه سازی میزان به‌پخشی طرح BAS را با طرح‌های SRS و $GRTS$ مشاهده کردند که به‌پخشی این طرح بهتراست.

[8] Simple Random Sampling
[9] Stratified Simple Random Sampling

یکی از عیب های این طرح این است که وقتی در انتخاب واحدهای نمونه، واحدهای بی پاسخی یا جعبه‌های بدون واحد وجود داشته باشد، احتمال شمول واحدهای به دست آمده از نمونه‌های BAS با احتمال شمول تعیین شده واحدها قبل از نمونه گیری یکسان نیست. این عیب اریبی و واریانس برآوردکننده را افزایش می‌دهد.



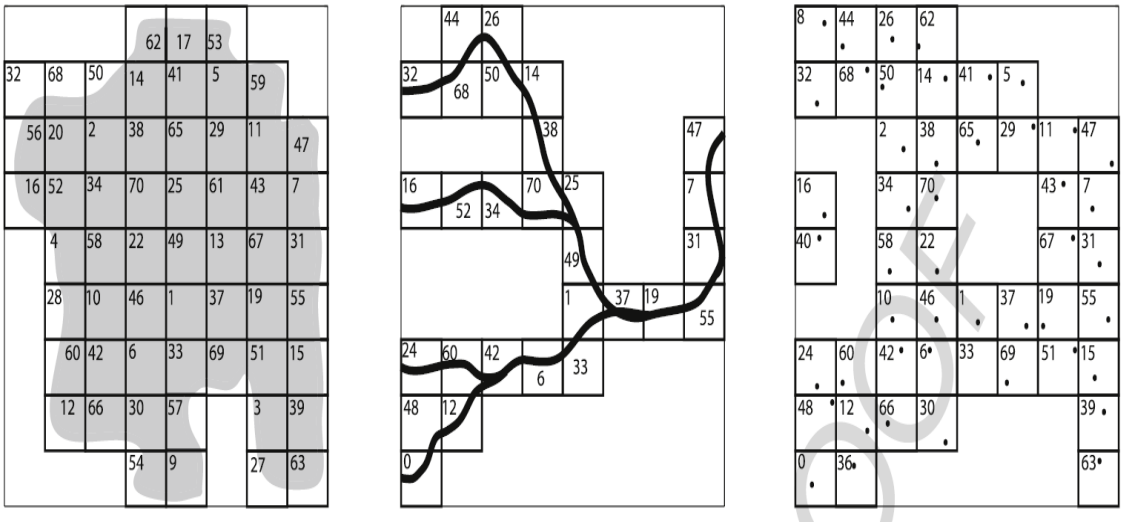
شکل ۱: نقاط داخل ناحیه مورد بررسی (●) و نقاط خارج ناحیه (▲).

۲.۳ طرح نمونه‌گیری پذیرش متعادل اصلاح شده

رابرتسون و همکاران (۲۰۱۷) عیب طرح نمونه‌گیری BAS را اصلاح کردند. آنها دو روش برای بر طرف کردن این عیب پیشنهاد کردند، که روش اول برای نمونه‌گیری بااحتمال برابر و نابرابر و روش دوم برای نمونه‌گیری بااحتمال برابر استفاده می‌شود.

۱. اگر اولین نقطه دنباله هالتون با شروع تصادفی منجر به انتخاب واحد نمونه نشود در آن صورت این دنباله را نادیده می‌گیرند و سپس از یک دنباله دیگر هالتون با شروع تصادفی استفاده می‌کند.

۲.چون ممکن است تعدادزیادی دنباله منجر به انتخاب نشوند، لذا در صورت امکان از چارچوب هالتون استفاده می‌کند که در این حالت ابتدا تعداد $B = \prod_{i=1}^{d} b_i^{J_i}$، جعبه هالتون را درنظر می‌گیرند و سپس در جوامع گسسته با تو جه به مکان واحدهای جامعه به هر واحد جامعه و در جوامع پیوسته با توجه به موقعیت ناحیه‌های افراز شده به هر ناحیه یک جعبه هالتون اختصاص می‌دهند. در ادامه به روش نمونه گیری تصادفی یک عدد از مجموعه $\{0, 1, ..., B\}$ انتخاب و متناسب با عدد انتخاب شده، n جعبه هالتون متوالی به عنوان مکان های واحدهای نمونه مشخص می‌گردند(شکل ۲). شماره خانه‌های چارچوب هالتون برابر با باقیمانده تقسیم اندیس نقاط دنباله هالتون بر ۷۲ است.



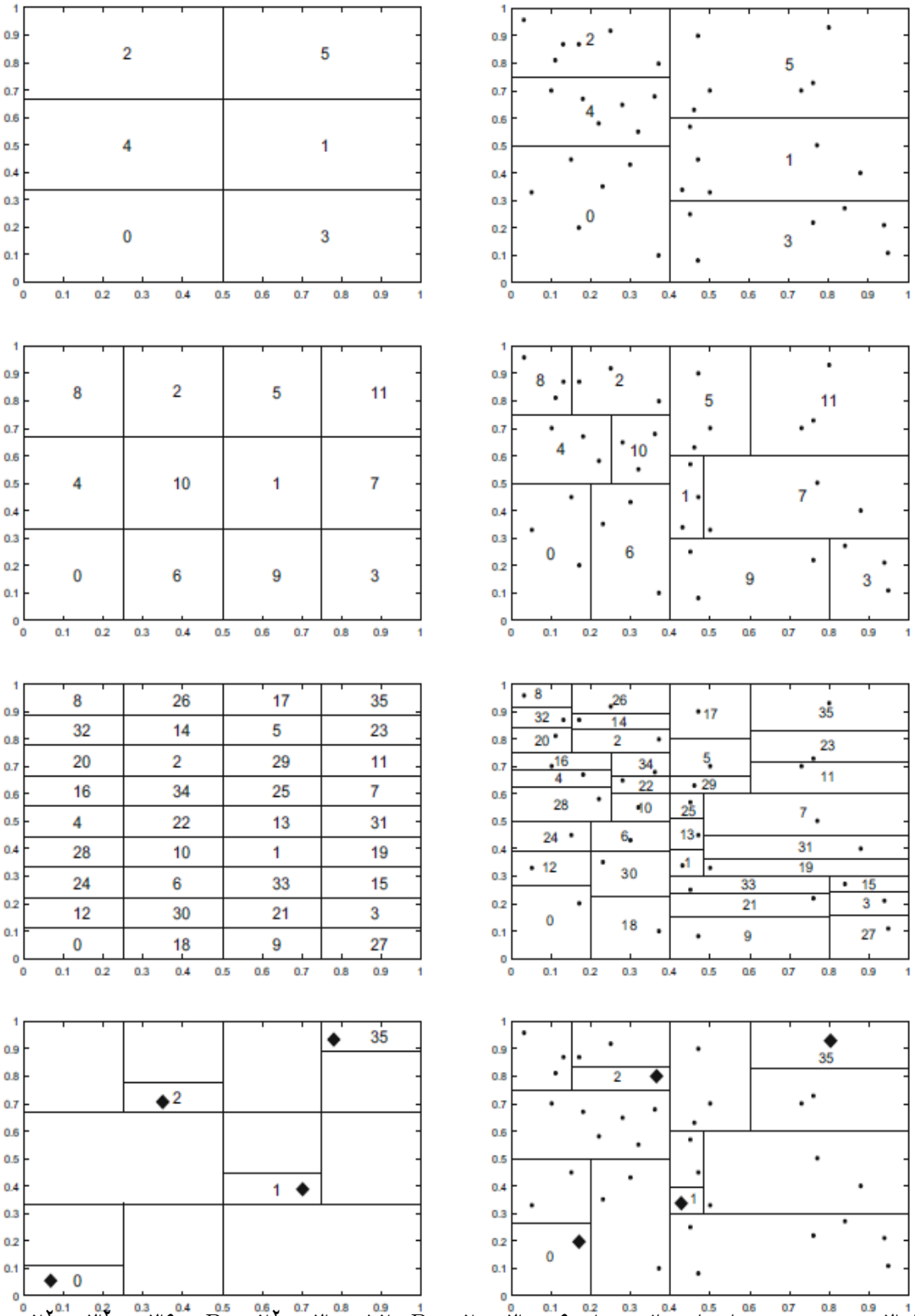
شکل ۲: چارچوب هالتون برای ناحیه مورد بررسی پیوسته (سایه دار)، خطی و نقطه‌ای با $b_1 = ۲, b_2 = ۳$ و $J = (۳, ۲)$.

۳.۳ طرح نمونه‌گیری افزاز شده تکراری هالتون

طرح نمونه‌گیری HIP توسط رابرتسون و همکاران (۲۰۱۳) معرفی گردید. که از دنباله شبه تصادفی هالتون با شروع تصادفی (وانگ و هیکرنل ۲۰۰۰) و یک افراز تکراری هالتون از جامعه با تعداد ناحیه‌های برابر با B برای به‌دست آوردن یک نمونه n تایی HIP استفاده کردند. شکل ۳ سمت چپ، خانه‌های هالتون برای $B = ۲ \times ۳ = ۶$، $B = ۲^۲ \times ۳ = ۱۲$ و $B = ۲^۲ \times ۳^۲ = ۳۶$ را نشان می‌دهد. نقطه X_K از دنباله هالتون در خانه هالتون با شماره برابر با باقیمانده تقسیم K بر B قرار می‌گیرد. بعنوان مثال $X_{۵۹}$ برای $B = ۶$ در خانه شماره ۵، برای $B = ۱۲$ در خانه شماره ۱۱ و برای $B = ۳۶$ در خانه شماره ۲۳ قرار می‌گیرد. شکل سمت راست، افرازهای تو در تو برای یک جامعه نقطه‌ای با احتمال برابر و با $N = ۳۶$ را نشان می‌دهد. به افرازهای انجام شده شماره منتاظر با شماره خانه‌های هالتون اختصاص یافته است. ردیف آخر سمت چپ یک نمونه ۴ تایی HIP از یک ناحیه پیوسته با احتمال برابر را نشان می‌دهد که $K = ۳۵$ و واحدهای نمونه در خانه های با شماره‌های $S_{۳۵} = \{۳۵, ۰, ۱, ۲\}$ قرار می‌گیرند، و شکل سمت راست این نمونه ۴ تایی را در جامعه نقطه‌ای نشان می‌دهد.

این طرح دارای اکثر خاصیت‌های یک طرح فضایی متعادل خوب مانند سادگی اجرا، به‌پخشی، قابل استفاده در فضای پیوسته و گسته با ابعاد بالا، توانایی تغییر در اندازه نمونه در هر مرحله از اجرای طرح، قابل استفاده در جوامع با شکل نامنظم، قابل استفاده با احتمال نابرار و نیر در صورت وجود واحدهای بی پاسخصی احتمال شمول به دست آمده از نمونه‌ها با احتمال شمول قبل از نمونه گیری خیلی نزدیک به‌هم می‌باشند.

در سه جامعه مصنوعی، شبیه‌سازی واریانس برآوردگر هورویتز–تامپسون نمونه‌های این طرح با طرح‌های نمونه‌گیری SRS، $GRTS$ و $LPM۲$ انجام گرفته و مشاهده گردید که در اکثر مواقع این واریانس، در طرح HIP کمتر از طرح های دیگر است. و نیز با شبیه‌سازی میزان به‌پخشی نسبی این طرح با طرح های $LPM۱$، $LPM۲$، $GRTS$ و $GRTS(۲n)$ مورد مقایسه قرار گرفته‌است. نتایج شبیه سازی نشان می‌دهد، این طرح و طرح $LPM۲$ به‌پخشی بهتری دارند.



شکل ۳: سمت چپ: خانه‌های هالتون برای $B = ۲ \times ۳ = ۶$، $B = ۲^۲ \times ۳ = ۱۲$ و $B = ۲^۲ \times ۳^۲ = ۳۶$ و سمت راست، افرازهای تو در تو برای یک جامعه نقطه‌ای با احتمال برابر و با $N = ۳۶$.

مراجع

Bondesson, L., Thorburn, D. (2008), A List Sequential Sampling Method Suitable for Real-Time Sampling, *Canadian Journal of Statistic*, **35**, 466-483.

Deville, JC., Till´e, Y. (1998), Unequal Probability Sampling without Replacement Through a Splitting Method, *Biometrika*, **85**, 89-101.

Grafström, A. (2012), Spatially Correlated Poisson Sampling, *Journal of Statistical Planning and Inference*, **142**, 139-147.

Grafström, A., Lundström, NLP., Schelin, L. (2012), Spatially Balanced Sampling through the Pivotal Method , *Biometrics*, **68**, 514-520.

Halton, J. H. (1993), On the Efficiency of Certain Quasi-Random Sequences of Points in Evaluating Multi-Dimensional Integrals, *Numerische, Mathematik*, **2**, 339-359.

Hedayat, A. S., Rao, C. R., Stufken, J. (1988), Sampling Plans Excluding Contiguous Units, *Journal of Statistical Planning and Inference*, **19**, 159-170.

McDonald T. (2014), Sampling Designs for Environmental Monitoring, *In: Manly BFJ, Navarro Alberto JA (eds) Introduction to Ecological Sampling. CRC Press, Taylor and Francis Group, Boca Raton*

Overton, W. S., Stehman, S. V. (1993), Properties of Designs for Sampling Continuous Spatial Resources From a Triangular Grid, *Communications in Statistics, Part A Theory and Methods*, **22** , 2641-2660.

Pric, C. J., Pric, C. J. (2012), Recycling Primes in Halton Sequences: An Optimization Perspective, *Advanced Modeling and Optimization*, **14**, 17-29.

Robertson, B. L., Brown, J. A., McDonald, T., Jaksons, P. (2013), BAS: Balanced Acceptance Sampling of Natural Resources, *Biometrics*, **3**, 776-784.

Robertson, BL., McDonald, T., Price, C. J., Brown, JA. (2017), A Modification of Malanced Acceptance Sampling. *Statistics & Probability Letters*, **129**, 107-112.

Robertson, B. L., McDonald, T., Price, C. J., Brown, J. A. (2018), Halton Iterative Partitioning: Spatially Balanced Sampling via Partitioning, *Environmental and Ecological Statistics*, **25**, 305-323.

Stevens, DL, Jr,. Olsen, AR. (2004), Spatially Balanced Sampling of Natural Resources, *Journal of the American Statistical Association*, **99**, 262-278.

Wang, X., Hickernell, F. J. (2000), Randomized Halton sequences, *Mathematical and Computer Modelling*, **32**, 887-899.

3rd Seminar on
Spatial Statistics and Its Applications
University of Zanjan
23-24October 2019

سومین سمینار آمار فضایی و کاربردهای آن
دانشگاه زنجان ۲-۱ آبان ماه ۱۳۹۸

استفاده از رتبه‌ها در تحلیل ممیزی فضایی داده‌های چند متغیره

فائزه یزدانی[۱]**، مجید عظیم محسنی، مهناز خلفی**
گروه آمار، دانشگاه گلستان

چکیده: در این مقاله به روش ناپارامتری تحلیل ممیزی فضایی داده‌های چندمتغیره پرداخته می‌شود. ابتدا برای داده‌های چندمتغیره از روش مقیاس‌بندی چندبعدی رتبه‌ای یا غیر متریک مختصات فضایی تخصیص داده می‌شود. سپس به روش تحلیل شباهت رتبه‌ای به نحوه تفکیک گروه‌ها از لحاظ داده‌های چندمتغیره پرداخته می‌شود.بر اساس یک مجموعه داده واقعی کارایی روش مورد ارزیابی قرار می‌گیرد.

واژه‌های کلیدی: داده‌های فضایی، مقیاس‌بندی چندبعدی، تحلیل شباهت.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62H30 .

۱ مقدمه

تحلیل ممیزی به ارائه مرزبندی میان چند گروه از لحاظ مشاهدات چندمتغیره می‌پردازد. اما به دلیل بعد بالای مشاهدات، نمایش گرافیکی این مرزبندی‌ها عملاً امکان‌پذیر نمی‌باشد. روش مقیاس‌بندی چندبعدی یک روش کارا برای ارائه مختصات فضایی در ابعاد پایین برای داده‌های چندمتغیره فراهم می‌کند. علاوه بر آن این امکان را فراهم می‌سازد که بتوان از تحلیل‌های فضایی برای مشاهدات چندمتغیره استفاده کرد. روش مقیاس‌بندی چندبعدی معمولاً به دو روش متریک و غیر متریک صورت می‌گیرد. در روش متریک از ماتریس فاصله یا عدم شباهت داده‌ها در یافتن مختصات فضایی کمک گرفته می‌شود. وجود مشاهدات دور افتاده در مقادیر فاصله ممکن است انحراف زیادی را در نتایج ایجاد کند.به همین دلیل توابع یکنوا از فاصله که حساسیت کمتری به مشاهدات دور افتاده دارند مورد استفاده قرار می‌گیرد. مرسوم‌ترین تابع یکنوا رتبه‌ها می‌باشد که مبنای بسیاری از مطالعات ناپارامتری می‌باشد. استفاده از رتبه‌ها منجربه ارائه روش غیر متریک برای مقیاس‌بندی چندبعدی گردیده است. برای مطالعه بیشتر در خصوص مقیاس‌بندی چندبعدی می‌توان به لژاندر و لژاندر (۲۰۱۲) ، آگراوال و همکاران (۲۰۰۷) ، کاکسون (۱۹۸۲) ، کودی و اسپرینگر (۲۰۱۸) و گرفلمن و همکاران (۲۰۱۸) مراجعه

[۱] نام و ایمیل ارائه دهنده مقاله: فائزه یزدانی، yazdanifaeze@yahoo.com

کرد. چه داده‌ها به طور فضایی ثبت شده باشند و چه با مقیاس‌بندی چندبعدی به آن‌ها مختصات فضایی تخصیص داده شود، می‌توان از آنالیز شباهت فضایی رتبه‌ای به نحوه تفکیک گروه‌ها پرداخت. به‌طور دقیق‌تر می‌توان برحسب فواصل مختلف به تفکیک گروه‌ها پرداخت و آزمون آماری تفکیک گروه‌ها را انجام داد. برای مطالعه بیشتر در خصوص آنالیز شباهت فضایی رتبه‌ای می‌توان به لژاندر و لژاندر (۲۰۱۲) ، کلارک (۱۹۹۳) و باتیگیگ و همکاران (۲۰۱۴) مراجعه کرد. در بخش دوم به‌طور خلاصه به روش‌های به کار رفته در تحلیل ممیزی فضایی رتبه‌ای پرداخته می‌شود. در بخش سوم بر اساس یک سری داده واقعی به کارایی روش‌های رتبه‌ای در تفکیک گروه‌ها پرداخته می‌شود.

۲ روش‌ها

مقیاس‌بندی چندبعدی یک روش برای تخصیص مختصات فضایی در فضای با ابعاد کمتر به داده‌های چندمتغیره می‌باشد. الگوریتم این روش به دو صورت متریک و غیر متریک انجام می‌گیرد. در این روش ابتدا فاصله داده‌ها با یک معیار فاصله مانند فاصله اقلیدسی محاسبه می‌شود و ماتریس فاصله یا عدم شباهت $D = [d_{ij}]$ نقطه آغازین روش مقیاس‌بندی چندبعدی می‌شود.

فرض کنید $\mathbf{x}_1, ..., \mathbf{x}_n$ مختصات فرضی داده‌های چندمتغیره در فضای p بعدی باشد. به‌طور معمول $p = 2$ یا $p = 3$ درنظر گرفته می‌شود. هدف مقیاس‌بندی چندبعدی متریک یافتن مختصات $\mathbf{x}_1, ..., \mathbf{x}_n$ به‌طوری است که داشته باشیم:

$$d_{ij} \leq d_{kl} \Longleftrightarrow \|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \|\mathbf{x}_k - \mathbf{x}_l\|^2$$

برای یافتن این مختصات معمولاً از مینیمم‌سازی توابع استرس استفاده می‌شود که نمونه‌ای از آن به‌صورت

$$Stress = \left(\frac{\sum_i \sum_j (d_{ij} - \|\mathbf{x}_i - \mathbf{x}_j\|^2)}{\sum_{i,j} d_{ij}^2}\right)^{\frac{1}{2}}$$

است. در مقیاس‌بندی چندبعدی غیر متریک به جای استفاده از d_{ij} از تابعی یکنوا از آن استفاده می‌شود که به نوعی این مقادیر را اصلاح کند. به‌عنوان مثال بتواند مشاهدات دور افتاده را کنترل کند. به‌طور معمول در این مقیاس‌بندی چندبعدی از رتبه فواصل به‌جای خود فواصل استفاده می‌شود. اگر داشته باشیم $r_{ij} = rank(d_{ij})$ در این صورت مختصات داده‌ها به‌گونه‌ای تعیین می‌شود که داشته باشیم

$$r_{ij} \leq r_{kl} \Longleftrightarrow \|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \|\mathbf{x}_k - \mathbf{x}_l\|^2$$

این روش طی دو گام انجام می‌شود:

گام ۱ : یک مقدار اولیه برای مختصات درنظر گرفته شود.

گام ۲ : مقادیر گام اول آنقدر تغییر پیدا کند که تابع استرس به کم‌ترین مقدار خود برسد.

تحلیل شباهت یک آزمون ناپارامتری برای تحلیل ممیزی فضایی محسوب می‌شود. آماره این آزمون بر اساس شباهت درون گروه‌ها و بین گروه‌ها ساخته می‌شود. به‌طور دقیق‌تر این آماره ثبات گروه‌بندی را در فواصل مختلف مورد آزمون قرار می‌دهد. این آماره بر اساس رتبه‌ها شکل گرفته است به‌همین دلیل آزمون فرض آن به‌صورت ناپارامتری می‌باشد. آماره آزمون تحلیل شباهت R در فواصل مختلف به‌صورت

$$R(h) = \frac{r_B(h) - r_W(h)}{\frac{1}{4}n(n-1)} \qquad (۱.۲)$$

ساخته می‌شود، که در آن $r_B(h)$ میانگین شباهت‌های رتبه‌ای بین گروه ها، $r_W(h)$ میانگین شباهت‌های رتبه‌ای درون گروه‌ها و n تعداد کل نمونه‌های مورد بررسی است. آماره R مقادیر خود را در بازه $(-۱, ۱)$ اختیار می‌کند که اعداد مثبت نشان‌دهنده شباهت بیشتر درون گروه‌ها است و مقادیر نزدیک به صفر نشان می‌دهد که هیچ تفاوتی بین گروه ها و درون گروه‌ها از نظر شباهت وجود ندارد. مقادیر منفی R نیز نشان‌دهنده شباهت بیشتر بین گروه‌ها نسبت به درون گروه‌ها است.

محاسبه مقدار معنی‌داری به روش جایگشتی انجام می‌گیرد. الگوریتم محاسبه مقدار معنی‌داری از روش جایگشتی به‌صورت زیر است:

گام ۱: مشاهدات در گروه‌ها به‌طور تصادفی جابه‌جا شود.

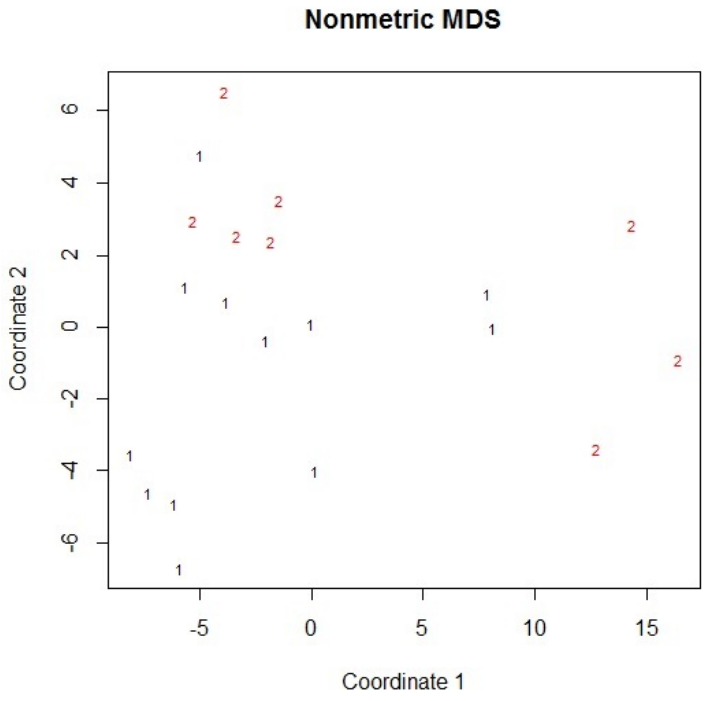
گام ۲: آماره R برای مشاهدات گام اول محاسبه شود.

گام ۳: گام اول و گام دوم B بار تکرار می‌شود مقدار آماره محاسبه و $R_۱, ..., R_B$ نام‌گذاری شوند.

گام ۴: در صد R_i که از R کوچکتر هستند محاسبه می‌شود و به عنوان مقدار معنی‌داری درنظر گرفته شود.

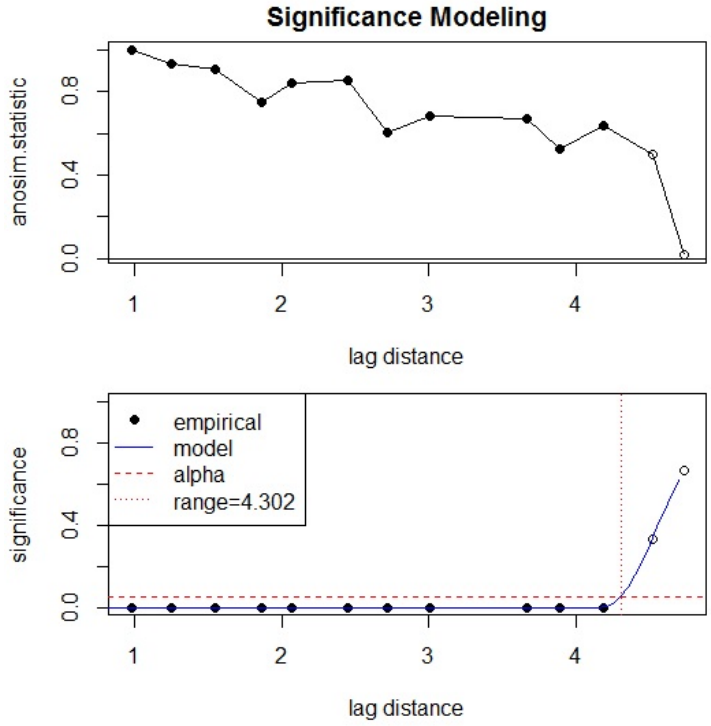
۳ نتایج عددی

در این بخش روش‌های ذکر شده در بخش دوم را برای یک مجموعه داده واقعی با دو گروه به کار می‌بریم. از دو گروه با وزن کم (کد ۱) و وزن زیاد (کد۲) ۱۱ فاکتور خونی، PH، تعداد رنگدانه کراتنین، ضریب اصلاح شده کراتنین، فسفات، کلسیم، فسفر، کراتنین، بور، کلرید، کولین، مس ثبت شده است. در این آزمایش ۱۲ نفر با وزن کم و ۸ نفر با وزن زیاد انتخاب شده‌اند. شکل ۱ آرایش فضایی مشاهدات چندمتغیره را در فضای دو بعدی نشان می‌دهد. تجمع مشاهدات با کد



شکل ۱: مختصات فضایی مشاهدات فاکتورهای خونی دو گروه با وزن کم (کد۱) و وزن زیاد (کد۲) با روش مقیاس بندی چند بعدی رتبه‌ای

یکسان در کنار هم نشان‌دهنده کارایی روش مقیاس‌بندی رتبه‌ای چندبعدی می‌باشد.



شکل ۲: نتایج حاصل از تحلیل شباهت فضایی: نمودار بالا مقدار آماره R در فواصل مختلف. نمودار پایین مقادیر معنی داری برحسب فاصله به همراه نمودار گوسی برازش داده شده

شکل ۲ نتایج حاصل از آنالیز شباهت داده‌ها برای دو گروه افراد با وزن زیاد و کم را نشان می‌دهد. بر اساس نمودار قسمت بالا مشخص می‌شود که در فواصل کمتر میزان شباهت درون گروه‌ها بالا و اختلاف بین گروه‌ها زیاد می‌باشد. هرچه فاصله بیشتر شده است عدم شباهت بین گروه‌ها کاهش پیدا کرده است. نمودار پایین مقادیر معنی‌داری در فواصل مختلف را براساس آزمون جایگشتی نشان می‌دهد. براساس این نمودار می‌توان قضاوت کرد که تا فاصله $h = ۴/۳۰۲$ اختلاف معنی‌داری بین گروه‌ها وجود دارد و پس از آن این فاصله دیگر اختلاف بین گروه‌ها معنی‌دار نیست. با توجه به این که این نمودار به‌صورت غیرنزولی افزایش می‌یابد می‌توان به آن یک تابع گوسی تجمعی به‌صورت

$$y^s(h) = C_h(۱ - e^{C_w(h-C_c)^۲}) \qquad (۱.۳)$$

برازش داد، که در آن C_h ارتفاع منحنی، C_w عرض یا شیب منحنی و C_c مرکز منحنی است. در شکل ۲ قسمت پایین منحنی گاوسی برازش داده شده به مقادیر معنی‌داری نیز ترسیم شده است. پارامترهای برآورد شده در این مدل $Ch = ۱/۰۰۰۵۳۴, Cc = ۴/۱۸۳۱۷۳, Cw = ۳/۶۵۴۰۷۹$ هستند.

بحث و نتیجه‌گیری

در این مقاله به یک تحلیل ممیزی فضایی برای داده‌های چندمتغیره پرداخته شد. روش مقیاس‌بندی چندبعدی این امکان را فراهم ساخت که نه‌تنها بتوان به داده‌های کلاسیک چندگانه آرایش فضایی داد، بلکه بتوان از آنالیز شباهت فضایی برای این

داده‌ها استفاده کرد. کلیه مراحل این تحلیل ممیزی فضایی مبتنی بر رتبه‌ها ارائه گردید. نتایج عددی که بر اساس مجموعه‌ای از داده‌های واقعی ارائه گردید نشان داد که استفاده از رتبه‌ها به‌خوبی، توانایی تفکیک گروه‌ها را دارد. نتایج عددی این مقاله براساس تحلیل ممیزی دو گروه از افراد ارائه گردید، اما این روش را به‌راحتی می‌توان برای چند گروه از افراد نیز به‌کار گرفت.

مراجع

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., and Belongie, S. (2007), Generalized Non-Metric Multidimensional Scaling, *In Artificial Intelligence and Statistics*, 11-18.

Buttigieg, Pier Luigi; Ramette, Alban (2014), A Guide to Statistical Analysis in Microbial Ecology: A Community-Focused, Living Review of Multivariate Data Analyses, *FEMS Microbiology Ecology*, *90*, 543–550.

Clarke, K. R. (1993), Non-Parametric Multivariate Analyses of Changes in Community Structure, *Australian Journal of Ecology*, **18**, 117-143.

Coxon, A. P. M. (1982), The User's Guide to Multidimensional Scaling, *London: Heinemann Educational Books*.

Ding, C. S. (2018), Fundamentals of Applied Multidimensional Scaling For Educational and Psychological Research, *Springer: Cham.*

Graffelman, J., Femenia, I. G., de Cid, R., and Barcelo-i-Vidal, C. (2018), Multidimensional Scaling and Relatedness Research, *bioRxiv*, 297879.

Legendre, P., and Legendre, L. F. (2012), *Numerical Ecology*, Elsevier, Vol. 24.