انجمن آمار ایران

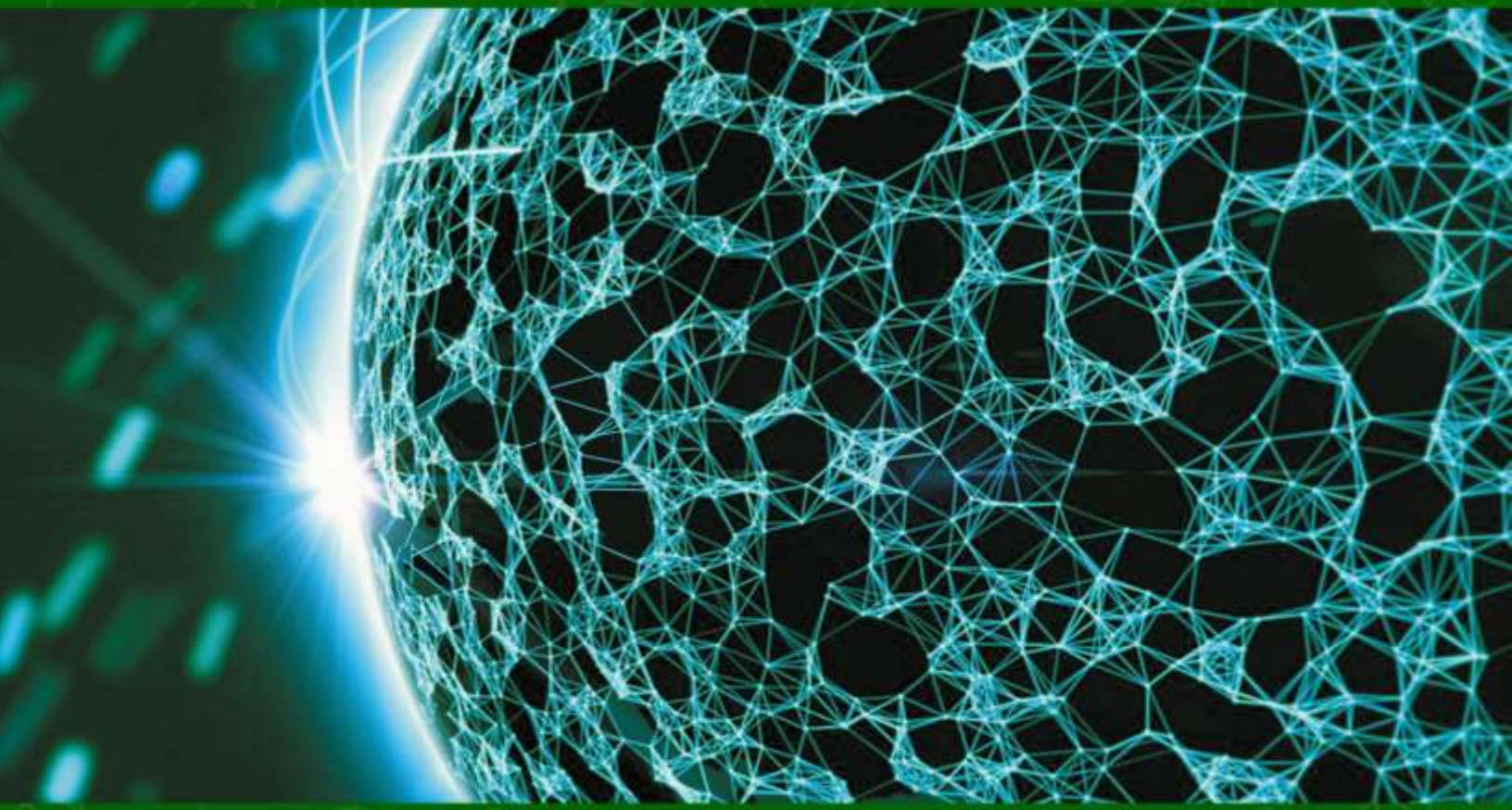
IRANIAN STATISTICAL SOCIETY 1990

دانشگاه صنعتی شاهرود

مجموعه مقاله ها

دومین سمینار آمار فضایی و کاربردهای آن

۳-۴ آبان ۱۳۹۶



بسمه تعالی

مجموعه مقاله های دومین سمینار آمار فضایی و کاربردهای آن

دانشگاه صنعتی شاهرود

۳-۴ آبان ۱۳۹۶

دانشگاه صنعتی شاهرود

مجموعه مقاله‌های دومین سمینار آمار فضایی و کاربردهای آن

تدوین: حسین باغیشنی، مینا نوروزی‌راد

صفحه‌آرایی: مهسا نادی‌فر، مینا نوروزی‌راد

طراح جلد: علیرضا صفریان

تاریخ انتشار: آبان ۱۳۹۶

به نام خداوند جان و خرد

پیش‌گفتار

حمد و سپاس خداوند بلندمرتبه یکتا را که یاریمان کرد و توفیق خدمت به جامعه علمی کشور را سرلوحه و هدفمان قرار داد تا **دومین سمینار آمار فضایی و کاربردهای آن** را با همکاری انجمن آمار ایران و دانشگاه تربیت مدرس تهران، و همچنین تلاش صادقانه همکاران و دانشجویان گروه آمار و دانشکده علوم ریاضی، و حمایت‌های مسئولین دانشگاه صنعتی شاهرود، پژوهشکده آمار ایران و معاونت اجتماعی و پیشگیری از وقوع جرم قوه قضاییه، در روزهای سوم و چهارم آبان ۱۳۹۶ در دانشگاه صنعتی شاهرود برگزار نماییم.

هدف این سمینار، بررسی، گفتگو و فراهم نمودن شرایطی برای ارایه فعالیت‌های پژوهشی در زمینه آمار فضایی و کاربردهای آن است. امیدواریم این گردهمایی، فرصتی مناسب برای دستیابی به اهداف سمینار را فراهم کند و ارتباطی نزدیک‌تر و قوی‌تر بین پژوهش‌گران و کارشناسان آمار ایجاد کند و دلیلی برای ارتقای فرهنگ آماری در جامعه باشد.

کمیته علمی سمینار از بین ۳۷ مقاله ارسال‌شده به دبیرخانه، تعداد ۳۵ مقاله را به‌عنوان سخنرانی پذیرفت و با برگزاری یک کارگاه آموزشی نیز موافقت کرد. از بین مقاله‌های پذیرفته‌شده ۲۹ مقاله به زبان فارسی و ۶ مقاله به زبان انگلیسی نوشته شده‌اند.

برگزاری این سمینار بدون مساعدت‌های بی‌دریغ همکاران گرامی در دانشگاه‌های سراسر کشور امکان‌پذیر نبود؛ همکاری این عزیزان در داوری مقاله‌ها و ارایه پیشنهادهای سازنده، کمک شایانی به هر چه بهتر شدن مطالب این مجموعه و سایر متون علمی این سمینار کرده است. از طرف کمیته‌های اجرایی و علمی سمینار از تلاش‌های بی‌شائبه و زحمات صادقانه همه مسئولین، همکاران هیأت علمی، دانشجویان گروه آمار و کارکنان قسمت‌های مختلف دانشگاه صنعتی شاهرود و همه افرادی که در برگزاری سمینار ما را یاری کردند، قدردانی می‌کنیم.

کمیته برگزاری دومین سمینار آمار فضایی و کاربردهای آن

دانشگاه صنعتی شاهرود

آبان ۱۳۹۶

همکاران و حامیان

این سمینار با مشارکت و حمایت سازمان‌ها و موسسات زیر برگزار می‌گردد. از همه افراد و سازمان‌هایی که برای برپایی این سمینار مشارکت و از آن حمایت کردند، صمیمانه قدردانی می‌کنیم.

انجمن آمار ایران

دانشگاه تربیت مدرس تهران

پژوهشکده آمار ایران

معاونت اجتماعی و پیشگیری از وقوع جرم قوه قضاییه

پایگاه استنادی علوم جهان اسلام (ISC)

کمیته برگزاری دومین سمینار آمار فضایی و کاربردهای آن

دانشگاه صنعتی شاهرود

آبان ۱۳۹۶

اعضای کمیته اجرایی

داود شاهسونی	دانشگاه صنعتی شاهرود
ابراهیم هاشمی	دانشگاه صنعتی شاهرود
احمد نزاکتی	دانشگاه صنعتی شاهرود
محمدرضا ربیعی	دانشگاه صنعتی شاهرود
محمد آرشی	دانشگاه صنعتی شاهرود
نگار اقبال	دانشگاه صنعتی شاهرود
حسین باغیشنی	دانشگاه صنعتی شاهرود
مهرداد غزنوی	دانشگاه صنعتی شاهرود
علیرضا عرب‌امیری	دانشگاه صنعتی شاهرود

اعضای کمیته علمی

محسن محمدزاده	دانشگاه تربیت مدرس تهران
حسین باغیشنی	دانشگاه صنعتی شاهرود
محمد آرشی	دانشگاه صنعتی شاهرود
امید کریمی	دانشگاه سمنان
فاطمه حسینی	دانشگاه سمنان
افشین فلاح	دانشگاه بین‌المللی امام خمینی قزوین
نصرالله ایران‌پناه	دانشگاه اصفهان

فهرست

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1390

خوشه‌بندی داده‌های فضایی بعدبالا

آمنه آبیار*، محسن محمدزاده ، کیومرث مترجم

گروه آمار، دانشگاه تربیت مدرس

چکیده: امروزه توسعه و گسترش سریع علوم در زمینه‌های مختلف و افزایش سرعت ثبت و تحلیل اطلاعات، آماردانان را با حجم عظیمی از اطلاعات پیچیده مواجه نموده است که بکارگیری روش‌های موجود را ناممکن ساخته است. همچنان که تعداد زیاد متغیرهای تبیینی می‌تواند موجب بیش‌برازشی شود، وجود تعداد زیاد آن‌ها موجب انحراف مدل‌های معمول از الگوی واقعی داده‌ها می‌گردد. یکی از راه حل‌های فراهم‌سازی امکان تحلیل داده‌های بعدبالا، تعیین گروه‌های پنهان در این داده‌ها است. برای این منظور از روش پرکاربرد و کارآمد خوشه‌بندی داده‌ها برای تحلیل تشخیصی داده‌های بعدبالا استفاده می‌شود، که معمولا مبتنی بر داده‌های نابسته است. اما اغلب با داده‌هایی سر و کار داریم که همبسته فضایی هستند. در این مقاله طی یک مطالعه شبیه‌سازی کارایی روش تحلیل تشخیصی بعدبالا در خوشه‌بندی داده‌ها مورد بررسی قرار می‌گیرد و تاثیر وجود همبستگی فضایی در داده‌ها در کارایی این روش ارزیابی می‌شود تا بتوان راه حل‌های مناسبی برای تحلیل این گونه داده‌ها ارائه نمود.

واژه‌های کلیدی: خوشه‌بندی، تحلیل تشخیصی، داده‌های فضایی بعدبالا.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 91B72، 62H30.

۱ مقدمه

در بسیاری از مطالعات برای تحلیل و بررسی یک موضوع، با انبوهی از اطلاعات در قالب متغیرهای تبیینی روبرو هستیم. امروزه تولید چنین داده‌هایی به ویژه در مطالعات پژوهشی علوم پزشکی، تجاری و کامپیوتر به‌طور چشم‌گیری افزایش یافته است. در برخی مطالعات زیستی و کامپیوتری، تعداد این متغیرهای تبیینی چند برابر حجم نمونه است. اما در مواردی به غیر از آن تعداد متغیرهای تبیینی معمولا از تعداد نمونه تجاوز نمی‌کند اما تعداد زیاد آن موجب کاهش شدید دقت مدل‌بندی داده‌ها می‌شود. بررسی و مدل‌بندی این حجم عظیم داده امکان بررسی اولیه، نحوه برخورد مناسب و پیش‌گویی را در هر حیطه‌ای فراهم می‌کند. اما مشکلات بسیاری برای تحلیل این‌گونه داده‌ها مطرح است که فرایندهای آماری حاضر قادر نیستند به طور مناسب به آنها بپردازند. به این گونه مسائل مطرح شده، مسئله بعدبالا و به چنین داده‌هایی، داده‌های بعدبالا[1] گفته می‌شود.

*ارایه‌دهنده مقاله: آمنه آبیار، a.abyar@modares.ac.ir

[1] High dimensional data

یکی از تکنیک‌های مدل‌بندی و تحلیل داده‌های بعدبالا که حجم نمونه بسیار کمتر از تعداد متغیرهای تبیینی است، کاهش تعداد متغیرهای تبیینی است. به این صورت که با شناسایی و نادیده گرفتن متغیرهای تبیینی کم اثر یا استفاده از ترکیب خطی از آن‌ها، مدل‌بندی و تحلیل داده‌ها صورت می‌پذیرد. با ظهور داده‌های بعدبالا، اهمیت براوردیابی تُنُک[۲] در مدل‌بندی رگرسیون افزایش یافت. روش‌های معمول مانند گزینش گام به گام[۳] برای این داده‌ها ناپایدار هستند (بریمن، ۱۹۹۶).

تلاش‌های دیگر انجام شده بر روش‌های منظم‌کننده[۴] متمرکز هستند که انتخاب متغیرها و براوردیابی را به طور هم‌زمان انجام می‌دهند، مانند روش لسو[۵] (تیبشیرانی، ۱۹۹۶)، SCAD (فن و لی، ۲۰۰۱) و گروه لسو[۶] (یوآن و لین، ۲۰۰۶). این فرایند همچنین در تحلیل مجموعه داده‌های بعدبالا معمول که تعداد متغیرهای تبیینی p، می‌تواند بزرگتر از تعداد مشاهدات n باشد، موثر شناخته شده است. اغلب چنین فرایندهایی براساس تنظیم کردن برخی توابع مانند تابع درستنمایی است. اگر چه، در بسیاری از مواقع، تعیین کامل مدل‌های درستنمایی بسیار دشوار است، به‌ویژه برای داده‌هایی با ساختار پیچیده مانند مشاهدات همبسته یا داده‌های سانسور شده. در موارد دیگر، مانند براورد نیرومند[۷]، به جای مدل‌بندی کل درستنمایی، تنها مدل‌بندی گشتاور اول یا چند گشتاور اول داده‌ها نیاز است. در روش‌های سنتی برای تحلیل یک زیرگروه اغلب به علت آزمون چندگانه با مشکلاتی مواجه هستیم (لاگاکوس، ۲۰۰۶)، که استفاده از فرایندهای انتخاب متغیر چندبعدی مانند لسو می‌تواند این مشکلات را کاهش بدهد. در این حالت چند عامل طبقه‌بندی می‌شوند و نیاز داریم آنها را با هم گروه‌بندی کنیم اما نمی‌توان از گروه لسو برای داده‌های همبسته استفاده کرد. زمانی که فرم تابع درستنمایی مشخص نیست، اغلب براورد از طریق معادلات براوردیابی[۸] استخراج می‌شود. فرایندهای منظم کننده زیادی برای معادلات براوردیابی که براورد تُنُک مانند روش LARS ارائه شده است، اما مشکلات بسیاری همچنان باقی است. فو (۲۰۰۳)، جانسون و همکاران (۲۰۰۸)، ونگ و همکاران (۲۰۱۲) معادلات براوردیابی جزئی را مورد مطالعه قرار دادند اما روش‌های پیشنهادی ارائه شده دارای پیچیدگی محاسباتی هستند و یا این روش‌ها نمی‌توانند براوردهای تُنُک دقیق را به دست بیاورند (ژانگ و همکاران، ۲۰۱۰). یوئکی (۲۰۰۹) روش‌های نوع‌ـ نرم‌ـ ℓ_1 تنظیم‌کننده براساس کران‌سازی هموار را توسعه داد اما به مقادیر اولیه $n^{1/2}$ پارامتر براوردهای سازگار نیاز دارد که به‌دست آوردن آنها با متغیرهای تبیینی بعدبالا دشوار است. ولفسون (۲۰۱۱) الگوریتم بوستینگ (EEBoost) را برای معادلات براوردیابی توسعه داد، اما EEBoost نمی‌تواند با معادلات تنظیم‌کننده پیچیده‌تر مانند تابع جریمه لسو گروهی مطابقت پیدا کند.

برای قسمی دیگر از مطالعات پژوهشی نیز امکان اخذ نمونه با حجم بالا وجود دارد. اما همچنان تعداد متغیرهای تبیینی نسبتا زیاد است به‌طوری‌که تاثیر به‌سزایی در دقت مدل‌بندی خواهند داشت. به همین علت تعیین گروه‌بندی داده‌های بعدبالا می‌تواند به عنوان اولین گام در مواجهه با این داده‌ها نقش پراهمیتی در تحلیل منطقی داده‌ها را ایفا کند. بلمن (۱۹۵۷) نشان داد دقت و کارایی خوشه‌بندی داده‌های بعدبالا با افزایش بعد به سرعت کاهش می‌یابد. فلوری و همکاران (۱۹۹۷) یک روش تشخیصی با تکنیک کاهش بعد را با هدف خوشه‌بندی ارائه دادند به طوری که تمام فواصل بین دو خوشه در بعد پایین‌تر نیز رخ می‌داد.

---

[۲] Sparse estimation
[۳] Stepwise selection
[۴] Regularization methods
[۵] Lasso method
[۶] Group Lasso
[۷] Robust estimation
[۸] Estimation equations

اسکات و تامپسون (۱۹۸۳) پدیدهٔ فضای خالی[۹] را مطرح ساخت. بوویرون و همکاران (۲۰۰۷) با استفاده از این تعریف، فرض امکان تعلق داشتن داده‌های بعدبالا به بعدی کمتر را مطرح و تحلیل تشخیصی بعدبالا[۱۰] (HDDA) را براساس توزیع گاوسی ارائه کردند.

در ادامه، روش تحلیل تشخیصی بعدبالا در بخش ۲ ارائه خواهد شد. در بخش ۳، طی یک مطالعه شبیه‌سازی روش تشخیصی بعدبالا برای خوشه‌بندی داده‌های بعدبالا معمولی و فضایی مورد استفاده قرار می‌گیرد. در بخش نهایی به بحث و بررسی نتایج پرداخته و پیشنهاداتی برای مطالعات آتی ارائه خواهد شد.

۲ تحلیل تشخیصی بعدبالا

هدف از تحلیل تشخیصی، تخصیص هر مشاهده x به یک گروه از k گروه است، طوری که گروه واقعی مشاهده نامعلوم است. در این بخش تحلیل تشخیصی بعدبالا (بوویرون و همکاران، ۲۰۰۷) شرح داده خواهد شد. فرض کنید توزیع‌های شرطی هر خوشه نرمال به‌صورت $N(\mu_i, \Sigma_i)$ برای $i = ۱, \ldots, k$ باشند، که در آن μ_i میانگین خوشه iام، Σ_i ماتریس کوواریانس خوشه iام و k تعداد خوشه‌ها است. همچنین فرض کنید Q_i ماتریس متعامد[۱۱] بردارهای ویژه ماتریس Σ_i و Δ_i ماتریسی به‌صورت

$$\Delta_i = \begin{pmatrix} \begin{bmatrix} a_{i1} & & 0 \\ & \ddots & \\ 0 & & a_{id_i} \end{bmatrix} & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} b_i & & 0 \\ & \ddots & \\ 0 & & b_i \end{bmatrix} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{bmatrix} a & & 0 \\ & \ddots & \\ 0 & & a \end{bmatrix}}\right\} d_i \\ \left.\vphantom{\begin{bmatrix} a & & 0 \\ & \ddots & \\ 0 & & a \end{bmatrix}}\right\} (p - d_i) \end{matrix}$$

باشد، که در آن $a_{ij} \geq b_i$ برای $j = ۱, \ldots, d_i$ و $d_i < p$ است. زیرفضای E_i توسط d_i بردار ویژه اول متناظر با a_{ij} تولید می‌شوند به طوری که $\mu_i \in E_i$ است. خارج از این زیرفضا، واریانس توسط تک پارامتر b_i مدل‌بندی شده است. به علاوه، عملگر تصویر x روی E_i و $E_i^{\perp}$ به‌صورت

$$P_i(x) = \widetilde{Q}_i \widetilde{Q}_i^{\top} (x - \mu_i) + \mu_i$$
$$P_i^{\perp}(x) = \bar{Q}_i \bar{Q}_i^{\top} (x - \mu_i) + \mu_i$$

تعریف می‌شوند، که در آن $\widetilde{Q}_i$ با d_i ستون اول Q_i ساخته شده است و بقیه ستون‌ها صفر هستند به‌طوری‌که $\bar{Q}_i = Q_i - \widetilde{Q}_i$.

قضیه ۱.۲. (بوویرون و همکاران، ۲۰۰۷): با استفاده از قانون بیز در HDDA مشاهده x به خوشهٔ C_{i^*} تخصیص می‌یابد به‌گونه‌ای که $i^* = \operatorname{argmin}_{i=۱,\ldots,k} K_i(x)$ و

$$K_i(x) = ||\mu_i - P_i(x)||^{۲}_{\mathscr{A}_i} + \frac{۱}{b_i}||x - P_i(x)||^{۲} + \sum_{j=۱}^{d_i} \log(a_{ij}) + (p - d_i)\log(b_i) - ۲\log(\pi_i)$$

که در آن $||\cdot||_{\mathscr{A}_i}$ نُرم روی E_i است به‌طوری‌که $||x||^{۲}_{\mathscr{A}_i} = x^{\top}\mathscr{A}_i x$ و $\mathscr{A}_i = \widetilde{Q}_i \Delta_i^{-۱} \widetilde{Q}_i^{\top}$.

[۹] Empty space phenomenon
[۱۰] High-Dimensional Discrimination Analysis
[۱۱] Orthogonal matrix

این قاعده براساس دو فاصله بین مشاهدات و زیرفضای E_i و فاصله تصویر مشاهده x روی زیرفضای E_i و میانگین خوشه است. در این قاعده تصمیم‌گیری، واریانس‌های a_{ij} و b_i تحت احتمال پیشین π_i ساخته می‌شوند. بنابراین احتمال پسین به‌صورت

$$P(x \in C_i|x) = \frac{۱}{\sum_{\ell=۱}^{k} \exp\left\{\frac{۱}{۲}(K_i(x) - K_\ell(x)\right\}}$$

محاسبه می‌شود.

۳ مطالعه شبیه‌سازی

در این بخش برای مقایسه نحوه کارکرد روش تشخیصی HDDA هشت مجموعه داده تولید می‌شود. هر مجموعه داده شامل ۱۰۰۰ مشاهده و به ترتیب با $p = ۱۰, ۱۵, \ldots, ۱۴۰$ تولید می‌شوند، به‌طوری‌که اولین مجموعه داده فاقد هیچ‌گونه همبستگی است و دیگر مجموعه‌ها از یک میدان تصادفی گاوسی با همبستگی‌نگارهای متفاوت تولید می‌شوند. مجموعه دادهٔ ۲ و ۳ دارای ساختار همبستگی‌نگار گاوسی به‌صورت

$$C(h) = \sigma^۲ \exp\left\{-\frac{h^۲}{a^۲}\right\}, \quad \sigma^۲, a > ۰$$

است، که در آن h لگ فضایی، $\sigma^۲$ و a به ترتیب ازاره و دامنه همبستگی‌نگار هستند. مجموعه دادهٔ ۲ و ۳ به‌ترتیب دارای $\sigma^۲$ برابر ۱ و ۴ هستند. مجموعه دادهٔ ۴ تا ۶ دارای ساختار همبستگی‌نگار نمایی به‌صورت

$$C(h) = \sigma^۲ \exp\left\{-\left|\frac{h}{a}\right|\right\}, \quad \sigma^۲, a > ۰$$

است و مجموعه دادهٔ ۴ تا ۶ به‌ترتیب دارای $\sigma^۲$ برابر ۰/۵، ۱ و ۴ هستند. مجموعه دادهٔ ۷ و ۸ دارای ساختار همبستگی‌نگار توانی به‌صورت

$$C(h) = \sigma^۲ \exp\left\{-\left|\frac{h}{a}\right|^\theta\right\}, \quad \sigma^۲, a > ۰, \ ۰ < \theta < ۲ \qquad (۱.۳)$$

است و مجموعه دادهٔ ۷ و ۸ به‌ترتیب دارای $\sigma^۲$ برابر ۱ و ۴ و $\theta = ۰/۵$ هستند. همان‌طور که در (۱.۳) ملاحظه می‌شود همبستگی‌نگار گاوسی و نمایی حالت‌های خاص همبستگی‌نگار نمایی با θ برابر ۲ و ۱ است.

برای بررسی تأثیر تعداد خوشهٔ پنهان این مجموعه داده‌ها با تعداد خوشه‌های ۳ و ۴ تولید شدند. اما برای ارزیابی میزان کارایی روش HDDA از معیار شاخص مرزی اصلاح شده[۱۲] (ARI) به‌صورت

$$\text{ARI} = \frac{\sum_{i=۱}^{K} \binom{n_i}{۲} \sum_{j=۱}^{K} \binom{n_j}{۲}}{\binom{n}{۲}}$$

استفاده شد (هوبرت و آرابی، ۱۹۸۵)، که در آن n_i تعداد داده درخوشهٔ iام و n تعداد کل مشاهدات است.

شاخص ARI مقداری بین صفر و یک را اخذ می‌کند، به‌گونه‌ای که نزدیکی مقدار این شاخص یک نشان‌دهندهٔ خوشه‌بندی مناسب‌تر است. نتایج خوشه‌بندی مجموعه داده‌های تولید شده برای تعدادخوشهٔ ۳ و ۴ به‌ترتیب در جداول ۱ و ۲ و شکل‌های ۱ و ۲ ارائه شده است. همان‌طور که ملاحظه می‌شود، افزایش تعداد خوشه تأثیر به‌سزایی در کاهش عملکرد روش HDDA داشته

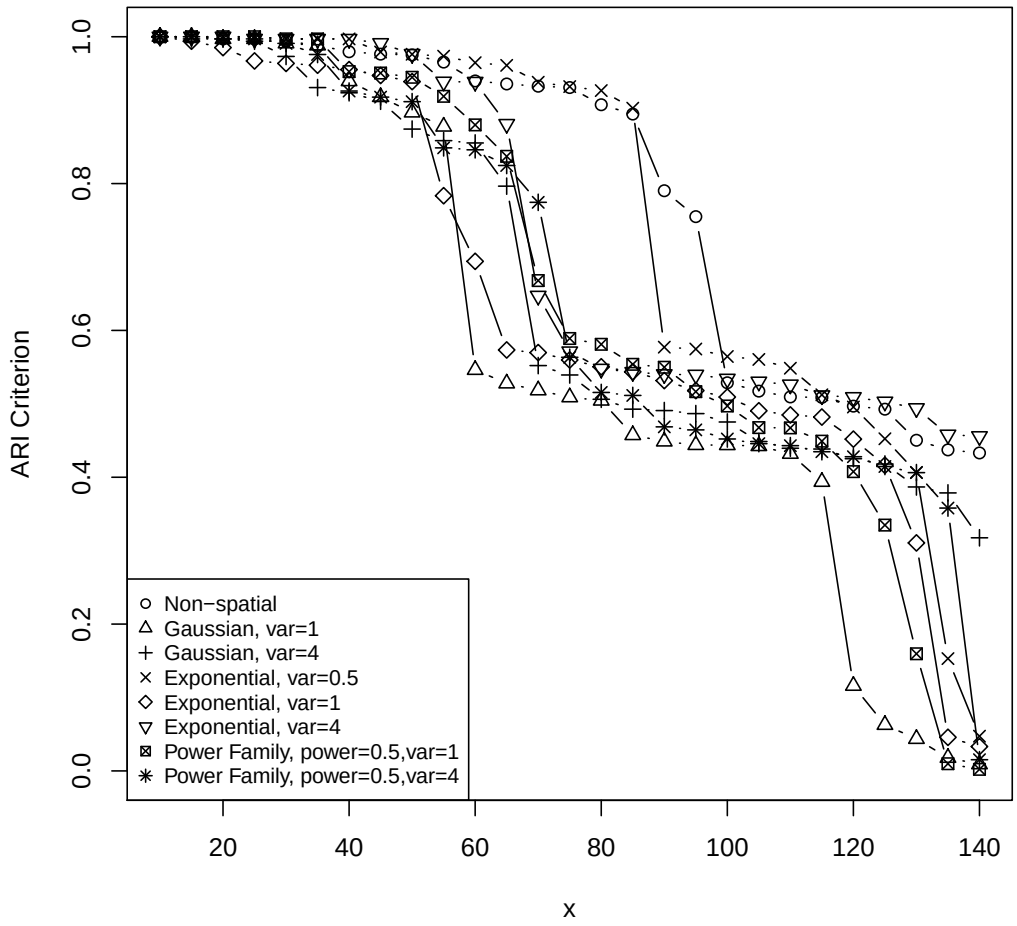
[۱۲] Adjusted Rand Index

جدول ۱: نتایج ARI در حالت ۳ خوشه

		گاوسی		نمایی			توانی	
p	غیرفضایی	$\sigma^۲=۱$	$\sigma^۲=۴$	$\sigma^۲=۰/۵$	$\sigma^۲=۱$	$\sigma^۲=۴$	$\sigma^۲=۱$	$\sigma^۲=۴$
۱۰	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹
۱۵	۰/۹۹۹	۰/۹۹۹	۰/۹۹۷	۰/۹۹۹	۰/۹۹۴	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹
۲۰	۰/۹۹۷	۰/۹۹۷	۰/۹۹۷	۰/۹۹۹	۰/۹۸۵	۰/۹۹۹	۰/۹۹۹	۰/۹۹۷
۲۵	۰/۹۹۷	۰/۹۹۷	۰/۹۹۴	۰/۹۹۹	۰/۹۶۷	۰/۹۹۷	۰/۹۹۹	۰/۹۹۷
۳۰	۰/۹۹۴	۰/۹۹۱	۰/۹۷۳	۰/۹۹۷	۰/۹۶۴	۰/۹۹۷	۰/۹۹۷	۰/۹۹۱
۳۵	۰/۹۸۸	۰/۹۸۸	۰/۹۳۱	۰/۹۹۷	۰/۹۶۱	۰/۹۹۷	۰/۹۹۷	۰/۹۷۶
۴۰	۰/۹۸۰	۰/۹۴۰	۰/۹۲۴	۰/۹۹۷	۰/۹۵۵	۰/۹۹۷	۰/۹۵۲	۰/۹۲۶
۴۵	۰/۹۷۶	۰/۹۱۸	۰/۹۱۲	۰/۹۷۹	۰/۹۴۷	۰/۹۹۱	۰/۹۵۱	۰/۹۱۸
۵۰	۰/۹۷۵	۰/۸۹۷	۰/۸۷۴	۰/۹۷۶	۰/۹۳۹	۰/۹۷۶	۰/۹۴۵	۰/۹۱۲
۵۵	۰/۹۶۶	۰/۸۷۸	۰/۸۵۸	۰/۹۷۴	۰/۷۸۳	۰/۹۳۹	۰/۹۱۹	۰/۸۴۹
۶۰	۰/۹۴۰	۰/۵۴۶	۰/۸۵۵	۰/۹۶۵	۰/۶۹۴	۰/۹۳۸	۰/۸۸۰	۰/۸۴۶
۶۵	۰/۹۳۶	۰/۵۲۸	۰/۷۹۶	۰/۹۶۱	۰/۵۷۳	۰/۸۸۱	۰/۸۳۷	۰/۸۲۵
۷۰	۰/۹۳۳	۰/۵۱۸	۰/۵۵۲	۰/۹۳۸	۰/۵۷۰	۰/۶۴۷	۰/۶۶۸	۰/۷۷۴
۷۵	۰/۹۳۱	۰/۵۰۹	۰/۵۳۹	۰/۹۳۲	۰/۵۶۰	۰/۵۷۱	۰/۵۸۹	۰/۵۶۳
۸۰	۰/۹۰۷	۰/۵۰۴	۰/۵۰۶	۰/۹۲۷	۰/۵۵۰	۰/۵۴۸	۰/۵۸۱	۰/۵۱۵
۸۵	۰/۸۹۵	۰/۴۵۷	۰/۴۹۳	۰/۹۰۳	۰/۵۴۴	۰/۵۴۳	۰/۵۵۴	۰/۵۱۱
۹۰	۰/۷۹۰	۰/۴۴۹	۰/۴۹۱	۰/۵۷۷	۰/۵۳۲	۰/۵۴۰	۰/۵۵۰	۰/۴۶۹
۹۵	۰/۷۵۵	۰/۴۴۴	۰/۴۸۷	۰/۵۷۵	۰/۵۱۸	۰/۵۴۰	۰/۵۱۷	۰/۴۶۵
۱۰۰	۰/۵۲۹	۰/۴۴۴	۰/۴۷۵	۰/۵۶۴	۰/۵۰۹	۰/۵۳۴	۰/۴۹۷	۰/۴۵۲
۱۰۵	۰/۵۱۷	۰/۴۴۲	۰/۴۴۹	۰/۵۶۰	۰/۴۹۱	۰/۵۳۰	۰/۴۶۸	۰/۴۴۶
۱۱۰	۰/۵۰۹	۰/۴۳۲	۰/۴۴۰	۰/۵۴۸	۰/۴۸۵	۰/۵۲۶	۰/۴۶۷	۰/۴۴۳
۱۱۵	۰/۵۰۹	۰/۳۹۴	۰/۴۳۸	۰/۵۱۲	۰/۴۸۲	۰/۵۱۱	۰/۴۴۹	۰/۴۳۵
۱۲۰	۰/۴۹۷	۰/۱۱۶	۰/۴۲۵	۰/۴۹۶	۰/۴۵۲	۰/۵۰۹	۰/۴۰۸	۰/۴۲۸
۱۲۵	۰/۴۹۳	۰/۰۶۳	۰/۴۱۸	۰/۴۵۲	۰/۴۱۶	۰/۵۰۳	۰/۳۳۵	۰/۴۱۵
۱۳۰	۰/۴۵۰	۰/۰۴۴	۰/۳۸۷	۰/۴۰۷	۰/۳۱۱	۰/۴۹۴	۰/۱۶۰	۰/۴۰۶
۱۳۵	۰/۴۳۷	۰/۰۱۸	۰/۳۷۹	۰/۱۵۳	۰/۰۴۶	۰/۴۵۸	۰/۰۱۰	۰/۳۵۸
۱۴۰	۰/۴۳۳	۰/۰۰۹	۰/۳۱۸	۰/۰۴۷	۰/۰۳۳	۰/۴۵۶	۰/۰۰۲	۰/۰۱۵

جدول ۲: نتایج ARI در حالت ۴ خوشه

		گاوسی		نمایی			توانی	
p	غیرفضایی	$\sigma^2=1$	$\sigma^2=4$	$\sigma^2=0/5$	$\sigma^2=1$	$\sigma^2=4$	$\sigma^2=1$	$\sigma^2=4$
۱۰	۰/۹۹۹	۰/۹۵۸	۰/۹۹۸	۰/۹۹۹	۰/۹۸۸	۰/۹۹۹	۰/۹۹۹	۰/۹۹۲
۱۵	۰/۹۹۹	۰/۷۸۹	۰/۹۷۶	۰/۹۹۳	۰/۹۷۹	۰/۷۵۵	۰/۹۸۳	۰/۴۲۹
۲۰	۰/۹۹۶	۰/۶۷۴	۰/۹۷۴	۰/۹۷۹	۰/۷۰۸	۰/۶۶۷	۰/۹۴۳	۰/۴۱۲
۲۵	۰/۹۹۲	۰/۶۴۲	۰/۶۸۴	۰/۶۶۱	۰/۶۶۳	۰/۶۵۶	۰/۶۹۰	۰/۳۵۵
۳۰	۰/۹۹۲	۰/۵۰۳	۰/۶۵۱	۰/۵۸۹	۰/۶۲۷	۰/۶۱۱	۰/۶۴۳	۰/۳۵۵
۳۵	۰/۹۸۰	۰/۴۷۲	۰/۶۳۹	۰/۵۰۷	۰/۶۰۰	۰/۵۶۶	۰/۵۵۴	۰/۳۵۳
۴۰	۰/۹۲۱	۰/۴۴۰	۰/۶۳۸	۰/۳۶۵	۰/۵۵۵	۰/۳۴۶	۰/۴۳۱	۰/۳۳۹
۴۵	۰/۹۰۹	۰/۳۹۵	۰/۶۲۲	۰/۳۴۶	۰/۴۱۲	۰/۳۴۶	۰/۴۰۲	۰/۳۳۵
۵۰	۰/۶۷۱	۰/۳۸۷	۰/۶۱۰	۰/۳۴۳	۰/۳۸۶	۰/۳۴۳	۰/۳۸۱	۰/۳۳۴
۵۵	۰/۶۶۷	۰/۳۶۷	۰/۴۹۴	۰/۳۳۵	۰/۳۵۸	۰/۳۴۱	۰/۳۴۸	۰/۳۲۱
۶۰	۰/۶۲۸	۰/۳۴۹	۰/۴۱۱	۰/۳۳۰	۰/۳۵۰	۰/۳۳۸	۰/۳۴۴	۰/۳۱۱
۶۵	۰/۴۷۹	۰/۳۳۹	۰/۳۹۴	۰/۳۲۹	۰/۳۴۳	۰/۳۲۲	۰/۳۴۳	۰/۳۰۳
۷۰	۰/۴۲۵	۰/۳۳۳	۰/۳۶۸	۰/۳۲۸	۰/۳۴۰	۰/۳۲۱	۰/۳۴۱	۰/۲۶۶
۷۵	۰/۳۸۴	۰/۳۳۳	۰/۳۳۹	۰/۳۰۳	۰/۳۳۸	۰/۲۷۹	۰/۳۳۵	۰/۲۴۳
۸۰	۰/۳۵۳	۰/۳۳۱	۰/۳۳۷	۰/۲۰۷	۰/۳۳۴	۰/۲۷۴	۰/۳۳۲	۰/۰۲۵
۸۵	۰/۳۴۵	۰/۳۳۰	۰/۳۳۱	۰/۱۷۲	۰/۳۲۹	۰/۲۶۴	۰/۳۳۲	۰/۰۲۳
۹۰	۰/۳۴۴	۰/۳۲۵	۰/۳۲۷	۰/۱۴۶	۰/۳۲۹	۰/۲۵۳	۰/۳۲۲	۰/۰۲۳
۹۵	۰/۳۳۱	۰/۳۱۷	۰/۳۱۲	۰/۰۶۳	۰/۳۲۳	۰/۱۳۵	۰/۳۰۹	۰/۰۲۱
۱۰۰	۰/۳۲۰	۰/۲۸۶	۰/۲۷۱	۰/۰۲۹	۰/۳۱۵	۰/۱۲۲	۰/۲۷۶	۰/۰۰۸
۱۰۵	۰/۳۱۴	۰/۲۶۶	۰/۲۱۴	۰/۰۱۶	۰/۲۹۰	۰/۰۶۲	۰/۱۹۱	۰/۰۰۴
۱۱۰	۰/۳۰۵	۰/۲۶۴	۰/۰۵۲	۰/۰۰۶	۰/۲۰۴	۰/۰۱۷	۰/۱۵۲	۰/۰۰۴
۱۱۵	۰/۰۳۹	۰/۰۶۶	۰/۰۲۵	۰/۰۰۴	۰/۱۶۵	۰/۰۰۰	۰/۰۷۱	۰/۰۰۳
۱۲۰	۰/۰۱۴	۰/۰۶۳	۰/۰۲۳	۰/۰۰۴	۰/۱۳۵	۰/۰۰۰	۰/۰۲۲	۰/۰۰۰
۱۲۵	۰/۰۱۱	۰/۰۲۰	۰/۰۲۱	۰/۰۰۱	۰/۱۱۰	۰/۰۰۰	۰/۰۰۸	۰/۰۰۰
۱۳۰	۰/۰۱۱	۰/۰۰۳	۰/۰۱۹	۰/۰۰۱	۰/۰۴۳	۰/۰۰۰	۰/۰۰۳	۰/۰۰۰
۱۳۵	۰/۰۰۴	۰/۰۰۰	۰/۰۰۱	۰/۰۰۰	۰/۰۳۶	۰/۰۰۰	۰/۰۰۱	۰/۰۰۰
۱۴۰	۰/۰۰۰	۰/۰۰۰	۰/۰۰۰	۰/۰۰۰	۰/۰۰۷	۰/۰۰۰	۰/۰۰۰	۰/۰۰۰

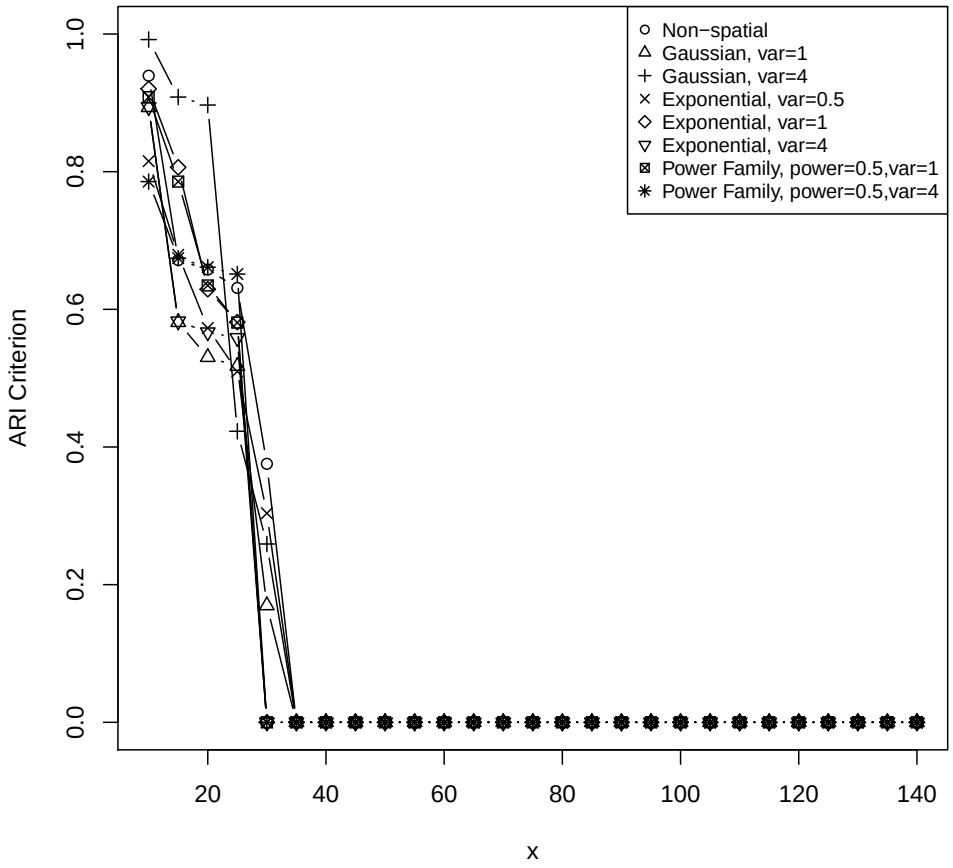


شکل ۱: روند شاخصARI با افزایش متغیرهای تبیینی با ۳ خوشه

است. به‌طور کلی افزایش ازارهٔ σ^2 در ساختارهای همبستگی گاوسی و نمایی موجب کاهش عملکرد این روش شده است. این در حالی است که براساس جدول ۱ افزایش ازارهٔ σ^2 تأثیر معکوس در عملکرد HDDA برای داده‌های تولید شده با ساختار توانی داشته است. در واقع در نگاه اول به نظر می‌رسیٔ کمتر بودن مقدار پارامتر θ از عدد یک موجب این عملکرد متفاوت شده است، وقتی که تعداد خوشه افزایش می‌یابد نتیجه کاملا متفاوت است. اما تعداد متغیرهای تبیینی موجود در هر مجموعه داده بیشترین تأثیر را در کاهش عملکرد مناسب روش HDDA داشته است.

بحث و نتیجه‌گیری

در این مقاله، عملکرد خوشه‌بندی برای داده‌های همبسته فضایی بعد بالا مورد بررسی قرار گرفت. همان‌طور که نتایج شبیه‌سازی نشان داد، افزایش وجود همبستگی فضایی موجب کاهش دقت این خوشه‌بندی که روش مناسب خوشه‌بندی داده‌های بعد بالا است، شده است.



شکل ۲: روند شاخص ARI با افزایش متغیرهای تبیینی با ۴ خوشه

در روش HDDA که با تغییر ساختار پارامترها خود شامل ۲۸ مدل متفاوت است، در نظر گرفتن فرض استقلال داده‌ها موجب محدود شدن این مدل شده است اما به طور کلی افزایش تعداد متغیرهای تبیینی در کاهش دقت تأثیر به‌سزایی داشته است. لذا برای تحقیقات آتی پیشنهاد می‌شود در این مدل یا مدل‌های مشابه، ساختار همبستگی‌ای در نظر گرفته شود، به گونه‌ای که برای داده‌های مستقل نیز کاربرد داشته باشد. رهیافت دیگر می‌تواند استفاده از روش‌هایی برای کاهش تعداد متغیرهای تبیینی و سپس استفاده از روش‌های تشخیصی است.

مراجع

Bellman, R. (1957), *Dynamic Programming*, Princeton Univercity Press, Princeton.

Bouveyron, C., Girard, S. and Schmid, C. (2007), High-Dimensional Discriminant Analysis, *Communications in Statistics - Theory and Methods*, **36**, 2607-2623.

Breiman, L. (1996), Heuristics of Instability and Stabilization in Model Selection, *Annals of Statistics*, **24**, 2350-2383.

Fan, J. and Li, R. (2001), Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

Flury, L., Boukai, B and Flury, B. (1997), The Discrimination Subspace Model, *Journal of American Statistical Association*, **92**, 758-766.

Fu, W. J. (2003), Penalized Estimating Equations, *Biometrics*, **59**, 126-132.

Hubert, L. and Arabie, P. (1985) Comparing Partitions, *Journal of Classification*, **2**, 193-218.

Johnson, B. A., Lin, D. Y. and Zeng, D. (2008), Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models, *Journal of the American Statistical Association*, **103**, 672-680.

Lagakos, S. (2006), The Challenge of Subgroup Analyses – Reporting without Distorting, *New England Journal of Medicine*, **354**, 1667-1669.

Scott, D. and Thompson, J. (1983), Probability Density Estimation in Higher Dimensions, *Fifteenth Symposium on the Interface*, Elsevier Science Publishers, 173-179.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lass, *Journal of the Royal Statistical Society*, Series B, **58**, 267-288.

Ueki, M. (2009), A Note on Automatic Variable Selection using Smooth-Threshold Estimating Equations, *Biometrika*, **96**, 1005–1011. 100

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis, *Biometrics*, **68**, 353-360.

Wolfson, J. (2011), EEBoost: A General Method for Prediction and Variable Selection Based on Estimating Equations, *Journal of the American Statistical Association*, **106**, 296-305.

Yuan, M. and Lin, Y. (2006), Model Selection and Estimation in Regression with Grouped Variables, *Journal of the Royal Statistical Society*, Series B, **68**, 49-67.

Zhang, H. H., Lu, W. and Wang, H. (2010), On Sparse Estimation for Semiparametric Linear Transformation Models, *Journal of Multivariate Analysis*, **101**, 1594-1606.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1390

مدل‌بندی نیم‌پارامتری داده‌های بقای وابسته فضایی با توابع مفصل

نسرین ابراهیمی*، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده: در برخی از مطالعات مربوط به زمان‌های بقا، علاوه بر تعیین تأثیر عوامل مخاطره مختلف روی زمان‌های بقا، بررسی وجود وابستگی‌های بالقوه بین زمان بقای آزمودنی‌های مختلف مورد توجه است. وقتی بین داده‌های بقا وابستگی وجود داشته باشد، می‌توان از توابع مفصل برای مدل‌بندی ساختار وابستگی بین آن‌ها استفاده کرد. در این مقاله داده‌های بقای فضایی با استفاده از تابع مفصل گاوسی مدل‌بندی شده و یک روش دو مرحله‌ای برای برآورد پارامترهای مدل ارائه می‌شود. سپس میزان دقت برآوردها در مطالعه‌ای شبیه‌سازی مورد ارزیابی قرار می‌گیرد.

واژه‌های کلیدی: داده‌های بقای چندمتغیره، تابع مفصل، مدل‌های حاشیه‌ای، مدل خطرهای متناسب.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62N02، 62H12.

۱ مقدمه

در بسیاری از مطالعات همه‌گیرشناسی، بررسی الگوی توأم بیماری در گروه‌هایی از افراد مورد توجه است. این گروه‌ها می‌توانند اعضای یک خانواده، افراد ساکن در یک منطقه جغرافیایی، بیماران مراجعه‌کننده به یک مرکز پزشکی و غیره باشند. تحلیل چنین داده‌هایی نیازمند استفاده از یک مدل توأم برای داده‌های بقای چندمتغیره است. در حالت کلی دو رهیافت برای لحاظ کردن وابستگی داده‌ها در چنین مدل‌هایی، استفاده از مدل‌های شکنندگی و تابع مفصل هستند. مدل‌های شکنندگی، وابستگی درون‌خوشه‌ای بین آزمودنی‌ها را از طریق لحاظ کردن یک اثر تصادفی پنهان تحت عنوان شکنندگی در مدل مدنظر قرار می‌دهند. اغلب، توزیع‌های پایدار مثبت و توزیع گاما برای اثر تصادفی در مدل شکنندگی در نظر گرفته می‌شوند. مدل‌های شکنندگی زمانی که استنباط‌های درون‌خوشه‌ای مدنظر باشند مورد استفاده قرار می‌گیرند، چرا که برآورد پارامترهای حاصل از این مدل، دارای تفسیرهایی به ازای یک مقدار خاص شکنندگی هستند. لی و ریان (۲۰۰۲) اثرات شکنندگی فضایی را با در نظر گرفتن یک میدان تصادفی گاوسی در

* **ارایه‌دهنده مقاله:** نسرین ابراهیمی، nasrin.ebrahimi@modares.ac.ir

چارچوب مدل نیم‌پارامتری خطرهای متناسب مدل کردند و این مدل بقای فضایی را در برازش به داده‌های مربوط به آسم کودکان به کار بردند. 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. هندرسون و همکاران (۲۰۰۲) ساختار فضایی داده‌های مربوط به سرطان خون را هم در قالب داده‌های زمین‌آماری و هم در قالب داده‌های ناحیه‌ای، با استفاده از مدل بقای فضایی مدل‌بندی کردند. بانرجی و دی (۲۰۰۵) تکنیک مدل‌بندی شکنندگی را در چارچوب مدل نیم‌پارامتری بخت‌های متناسب معرفی کردند. ژائو و همکاران (۲۰۰۹) اثرات تصادفی از مدل CAR را به‌عنوان شکنندگی در مدل‌های زمان شکست شتابیده، خطرهای متناسب و بخت‌های متناسب لحاظ کردند. پن و همکاران (۲۰۱۴) مدل نیم‌پارامتری خطرهای متناسب را با شکنندگی‌های CAR برای داده‌های سانسوریده بازه‌ای به کار بردند.

از آن‌جا که توابع مفصل نیز ابزاری قوی برای لحاظ کردن وابستگی داده‌ها هستند، می‌توان توابع مفصل را برای ساخت تابع بقای توأم با استفاده از 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. باردوسی (۲۰۰۶) برای اولین بار توابع مفصل را در تحلیل داده‌های وابسته فضایی به کار برد. 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. لی و همکاران (۲۰۱۵) تابع مفصل گاوسی را برای مدل‌بندی داده‌های بقای ناحیه‌ای به کار بردند.

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

۲ معرفی مدل بقای فضایی

فرض کنید متغیرهای T_{ij} و C_{ij} به ترتیب زمان‌های شکست و سانسور را برای آزمودنی jام در خوشه iام نشان دهد، که در آن $i = ۱, \cdots, m$ و $j = ۱, \cdots, n_i$. قابل توجه اینکه T_{ij} و C_{ij} به‌طور بالقوه مشاهده نشده‌اند و داده‌های مشاهده شده عبارتند از: $X_{ij} = \min(T_{ij}, C_{ij})$ و $\Delta_{ij} = I(T_{ij} \leq C_{ij})$. فرض کنید $Z_{ij}(t)$ بردار متغیرهای کمکی با بعد p باشد. همچنین مسیر[۱] متغیرهای کمکی تا زمان t با $\bar{Z}_{ij}(t) = \{Z_{ij}(s) | ۰ \leq s \leq t\}$ نمایش داده شود و فرض کنید T_{ij} به شرط $\bar{Z}_{ij}(T_{ij})$ مستقل از C_{ij} است. همچنین به شرط معلوم بودن $\bar{Z}_{ij}(t)$، تابع نرخ خطر به صورت

$$\lambda(t|\bar{Z}_{ij}(t)) = \lambda_{۰}(t) \exp\{\beta^{\top} Z_{ij}(t)\} \qquad (۱.۲)$$

[۱] Path

از مدل خطرهای متناسب پیروی می‌کند. معادله (۱.۲) یک مدل حاشیه‌ای برای هر T_{ij} 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 وابستگی T_{ij}ها، تبدیل نرمال به صورت

$$\tilde{T}_{ij} = \Phi^{-1}[1 - S(T_{ij}|\bar{Z}_{ij}(T_{ij}))] \qquad (۲.۲)$$

را در نظر بگیرید، که در آن Φ تابع توزیع تجمعی نرمال استاندارد و $S(.)$ تابع بقای متناظر با مدل خطرهای متناسب است. طبق تبدیل انتگرال احتمال، متغیر تصادفی $1 - S(T_{ij}|\bar{Z}_{ij}(T_{ij}))$ دارای توزیع $U(0,1)$ است، بنابراین $\tilde{T}_{ij}$ از توزیع نرمال استاندارد پیروی می‌کند. برای نشان دادن وابستگی $(\tilde{T}_{i1}, \cdots, \tilde{T}_{in_i})$ از ماتریس Σ_i استفاده می‌شود که درایه‌های آن براساس یک تابع کوواریانس یک پارامتری با پارامتر σ تعیین می‌شود.

۳ برآورد مدل

فرض کنید Y_{ij} نسخه بالقوه سانسورشده $\tilde{T}_{ij}$ باشد. از آن‌جا که تبدیل (۲.۲) یک به یک است، الگوی سانسور در داده‌ها حفظ می‌شود. با توجه به صفر و یک بودن مقادیر متغیر نشانگرِ رخداد Δ_{ij}، مجموع آزمودنی‌هایی که در در خوشه iام، پیشامد موردنظر را تجربه کرده‌اند عبارت است از $\Delta_i = \sum_{j=1}^{n_i} \Delta_{ij}$. همچنین بردار مشاهدات به‌گونه‌ای مرتب می‌شود که

$$\Delta_{i1} = \cdots = \Delta_{i\Delta_i} = 1$$

و Y_i به صورت $(Y_i^{\Delta_i}, Y_i^{n_i-\Delta_i})$ افراز می‌شود، که در آن $Y_i^{\Delta_i} = (Y_{i1}, \cdots, Y_{i\Delta_i})$ و $Y_i^{n_i-\Delta_i} = (Y_{i\Delta_i+1}, \cdots, Y_{in_i})$.

در ادامه برای مقادیر مختلف Δ_i سه حالت در نظر گرفته می‌شود.

الف‌ـ برای حالت $1 \leq \Delta_i \leq n_i - 1$، ماتریس Σ_i متناظر با افراز Y_i، به صورت

$$\Sigma_i = \begin{pmatrix} \Sigma_{i,11} & \Sigma_{i,12} \\ \Sigma_{i,21} & \Sigma_{i,22} \end{pmatrix}, \quad i = 1, \cdots, m$$

افراز می‌شود، که در آن $\Sigma_{i,11}$ ماتریسی با ابعاد $\Delta_i \times \Delta_i$ است. بردار $Y_i^{\Delta_i}$ دارای توزیع نرمال چندمتغیره با بردار میانگین صفر و ماتریس کوواریانس $\Sigma_{i,11}$، و $Y_i^{n_i-\Delta_i}|Y_i^{\Delta_i}$ دارای توزیع نرمال با بردار میانگین $\Sigma_{i,21}\Sigma_{i,11}^{-1}Y_i^{\Delta_i}$ و ماتریس کوواریانس $\Sigma_{i,22} - \Sigma_{i,21}\Sigma_{i,11}^{-1}\Sigma_{i,12}$ است.

از آن‌جا که تبدیل (۲.۲) یکنواست، درستنمایی داده‌های مشاهده شدهٔ X_{ij}، می‌تواند با استفاده از رابطه

$$P(X_{ij} < x) = P(Y_{ij} < y)$$

براساس داده‌های تبدیل‌یافتهٔ Y_{ij} نوشته شود، که در آن y تبدیل نرمال x است. بر این اساس، سهم خوشه iام در درستنمایی کل عبارت است از

$$L_i(\sigma, \beta, \Lambda) = \varphi^{\Delta_i}(Y_i^{\Delta_i})\tilde{\Phi}^{n_i-\Delta_i}(Y_i^{n_i-\Delta_i}|Y_i^{\Delta_i}) \prod_{j=1}^{n_i} \left[f(X_{ij}|\bar{Z}_{ij}(X_{ij}))/\varphi(Y_{ij})\right]^{\Delta_{ij}}, \quad i = 1, \cdots, m \qquad (۱.۳)$$

که در آن φ^{Δ_i} چگالی نرمال چندمتغیره، $\tilde{\Phi}^{n_i-\Delta_i}$ تابع بقای نرمال چندمتغیره، f چگالی متناظر با مدل خطرهای متناسب، و φ تابع چگالی نرمال استاندارد است.

ب- برای حالت $\Delta_i = ۰$، یعنی حالتی که تمام آزمودنی‌های خوشه iام سانسور شده‌اند، با قرار دادن $\varphi^{\Delta_i} = ۱$ و

$$\tilde{\Phi}^{n_i-\Delta_i}(Y_i^{n_i-\Delta_i}|Y_i^{\Delta_i}) = \tilde{\Phi}^{n_i-\Delta_i}(Y_i^{n_i-\Delta_i})$$

مجددا معادله (۱.۳) محاسبه می‌شود.

ج- برای حالت $\Delta_i = n_i$، یعنی حالتی که تمام آزمودنی‌های خوشه iام پیشامد را تجربه کرده‌اند، با قرار دادن $\tilde{\Phi}^{n_i-\Delta_i} = ۱$ مجددا معادله (۱.۳) محاسبه می‌شود.

حال درستنمایی کل از حاصل‌ضرب درستنمایی‌ها به‌صورت $L(\sigma,\beta,\Lambda) = \prod_{i=۱}^{m} L_i(\sigma,\beta,\Lambda)$ حاصل می‌شود.

۴ مطالعه شبیه‌سازی

برای شبیه‌سازی داده‌های بقای فضایی، ۱۰۰ مشاهده $(\tilde{T}_{۱},\cdots,\tilde{T}_{۱۰۰})$ از یک میدان تصادفی گاوسی با تابع کوواریانس نمایی به صورت

$$C(h) = \sigma^{۲}\exp\{-h/۲\} \qquad (۱.۴)$$

تولید شده است. برای این کار ابتدا ۱۰۰ موقعیت فضایی از توزیع یکنواخت شبیه‌سازی و با استفاده از ماتریس فاصلهٔ این نقاط و رابطه (۱.۴)، ماتریس کواریوگرام محاسبه می‌شود. با ضرب تجزیه چولسکی این ماتریس در ۱۰۰ مشاهده تولید شده از توزیع نرمال و سپس حذف روند، مشاهدات مورد نیاز از میدان تصادفی گاوسی تولید و با استفاده از معادله (۲.۲) مقادیر تابع بقا محاسبه می‌شوند. مدل خطرهای متناسب کاکس، با تابع خطر پایه ثابت $\lambda_۰(t) = ۱$ به عنوان مدل خطر حاشیه‌ای برای زمان‌های بقا در نظر گرفته شده و زمان‌های بقای موردنظر با استفاده از رابطه $t = -\dfrac{\ln(S(t))}{\exp\{\beta X\}}$، متغیر تبیینی X از توزیع بتا و زمان‌های سانسور با استفاده از توزیع نمایی با میانگین مقادیر t تولید شده و حدود ۳۵درصد نرخ سانسور در آن‌ها ایجاد شده است. جدول ۱ مقادیر واقعی پارامترها به همراه برآوردهای حاصل از روش معرفی‌شده و MSE آن‌ها را با ۲۰، ۵۰ و ۱۰۰ بار شبیه‌سازی نشان می‌دهد. همان‌طور که ملاحظه می‌شود برآوردهای حاصل از روش برآورد دو مرحله‌ای، دارای اریبی بسیار کمی هستند. همچنین MSE برآوردها خصوصاً در مورد پارامتر وابستگی، عملکرد مناسب این روش برآورد را نشان می‌دهد.

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 گاوسی 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 (۱.۳) نسبت به پارامترهای تابع کوواریانس، ماکسیمم شود. نتایج مطالعه شبیه‌سازی، بیانگر عملکرد مناسب این روش در برآورد ضرایب رگرسیونی و پارامتر وابستگی بود.

جدول ۱: مقادیر واقعی پارامترها به همراه برآورد و MSE آن‌ها

		$n=۲۰$		$n=۵۰$		$n=۱۰۰$	
پارامتر	مقدار واقعی	برآورد	MSE	برآورد	MSE	برآورد	MSE
β	۱	۱/۰۲۷۳	۰/۳۱۸۰	۱/۰۰۳	۰/۳۶۲۶	۰/۹۴۵۹	۰/۳۲۷۶
$\sigma^۲$	۱	۰/۹۸۱۱	۰/۰۰۸۵	۰/۹۳۲	۰/۰۲۵	۰/۹۲۷۲	۰/۰۳۱۶
β	−۱	−۰/۹۰۳۷	۰/۷۸۹۸	−۱/۱۸۳۶	۰/۶۶۱۰	−۱/۱۱۰۷	۰/۹۰۴۷
$\sigma^۲$	۱	۱/۲۴۰۱	۰/۱۵۱۸	۱/۳۱۵۴	۰/۲۲۸۴	۱/۲۲۶۷	۰/۱۴۹۹
β	۲	۱/۱۴۰۲	۰/۹۶۷۲	۱/۱۵۷۸	۱/۰۱۰۸	۱/۲۵۸۴	۰/۷۹۴۰۰
$\sigma^۲$	۳	۲/۵۸۳۴	۰/۵۴۲۵	۲/۵۵۱۸	۰/۳۳۲۸	۲/۶۳۸۸	۰/۳۱۷۳

سپاس‌گزاری

نویسندگان مقاله از پیشنهادات داوران گرامی که موجب بهبود مقاله و ارائه بهتر آن شده‌اند، کمال تشکر را دارند.

مراجع

امیدی، م.، محمدزاده ، م. (۱۳۹۰)، مدل بندی درست‌نمایی و بیز تجربی خشکسالی با استفاده از تابع مفصل، *نشریه علوم دانشگاه خوارزمی*، ۱۰، ۶۴۳-۶۵۴.

مترجم ، ک.، محمدزاده ، م. و آبیار ، آ. (۱۳۹۴)، مدل‌بندی فضایی داده‌های بقای سانسور شده، *پژوهش های ریاضی*، ۱، ۶۱-۷۰.

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

Bárdossy A. (2006). Copula-based Geostatistical Models for Groundwater Quality Parameters, *Water Resources Research*, **42**, 1-12.

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.

Henderson R., Shimakura S., and Gorst D. (2002). Modeling Spatial Variation in Leukemia Survival Data, *Journal of the American Statistical Association*, **97**, 965–972.

Li L., Hanson T., and Zhang J. (2015). Spatial Extended Hazard Model with Application to Prostate Cancer Survival, *Biometrics*, **71**, 313-322.

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

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

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

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

Zhao L., Hanson T. E., and Carlin B. P., (2009). Mixtures of Polya Trees for Flexible Spatial Frailty Survival Modelling, *Biometrika*, **96**, 263–276.

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

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶

رگرسیون فضایی برای داده‌های دومدی و چوله

جمیل اونق*، حسین باغیشنی، احمد نزاکتی

گروه آمار، دانشگاه صنعتی شاهرود

چکیده: تحلیل داده‌های فضایی دومدی که در موقعیت‌های کاربردی مختلف مشاهده می‌شوند، چندان مورد بررسی پژوهشگران قرار نگرفته است. در این مقاله با معرفی یک توزیع چندمتغیره منعطف، مدل رگرسیونی فضایی برای تحلیل این نوع داده‌ها را ارائه می‌کنیم. برای برازش مدل رهیافت درستنمایی را به‌کار خواهیم برد. با مثال‌های شبیه‌سازی و واقعی عملکرد مدل پیشنهادی را نسبت به رگرسیون فضایی معمول ارزیابی می‌کنیم.

واژه‌های کلیدی: داده‌های دومدی، رگرسیون فضایی، برآوردگر ماکسیمم درستنمایی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 62J12.

۱ مقدمه

داده‌های فضایی به داده‌هایی گفته می‌شود که بر حسب موقعیت قرار گرفتن آن‌ها، وابسته هستند. بنابراین در تحلیل رگرسیونی این نوع داده‌ها پذیره ناهمبسته بودن خطاها برقرار نیست و بین مشاهدات مجاور همبستگی قوی‌تری وجود دارد. در حالت کلی، برحسب آن‌که جمله خطا، متغیرهای تبیینی یا ضرایب رگرسیونی توابعی از موقعیت‌های فضایی باشند یا نه، انواع مختلفی از مدل‌های رگرسیون خطی برای داده‌های فضایی قابل تعریف هستند (محمدزاده ، ۱۳۹۱). در این مقاله تمرکز ما بر روی مدلی است که در آن ساختار وابستگی جمله خطا تابعی از موقعیت‌های فضایی است. معمولا در تحلیل رگرسیونی این نوع داده‌ها فرض بر این است که توزیع خطا نرمال چندمتغیره است. اما در موقعیت‌های مختلف با مواردی برخورد می‌کنیم که خطاها از توزیع نرمال تبعیت نمی‌کنند. به‌عنوان مثال در یک معدن طلا ممکن است دو ناحیه مجزا دارای رگه‌هایی با عیار طلای قابل ملاحظه و متفاوت باشند؛ در این حالت توزیع داده‌ها دومدی خواهد بود. به عنوان مثالی دیگر، در داده‌های خاک‌شناسی، توانایی تبادل کاتیوم به عنوان تابعی از موقعیت مکانی ناحیه تحت مطالعه، با فراوانی بیشتری نزدیک به دو مقدار عددی مشاهده می‌شود.

* **ارایه‌دهنده مقاله:** نام ارایه‌دهنده، email@email.ac.ir

برای تحلیل پاسخ‌های فضایی چوله، کارهای متعددی انجام شده‌اند. برای مثال به مطالعات ساهو و همکاران (۲۰۰۳)، آرنولد وال و همکاران (۲۰۰۵)، بولفرونی و لوکاس (۲۰۰۷)، کریمی و همکاران (۲۰۱۰)، حسینی و همکاران (۲۰۱۱) و حسینی و محمدزاده (۲۰۱۲) که بر پایه رده توزیع‌های پارامتری مانند چوله نرمال شکل گرفته‌اند، می‌توان اشاره کرد. اما برای پاسخ‌های دومدی، تا حدی که نویسندگان مطلع هستند، تنها مدل‌هایی معرفی شده‌اند که بر اساس فرآیندهای آمیخته[۱] پارامتری و ناپارامتری ساخته شده‌اند. به‌عنوان نمونه می‌توان فلاح و سدودی (۲۰۱۵) برای مدل آمیخته متناهی نرمال (پارامتری) و گلفاند و همکاران (۲۰۰۵) و گریفین و استیل (۲۰۰۶) برای مدل فرآیند آمیخته دیریکله (ناپارامتری) را نام برد.

ما در این مقاله، یک توزیع چندمتغیره منعطف معرفی می‌کنیم که ویژگی‌های مطلوبی دارد و از آن برای تحلیل پاسخ‌های فضایی غیرنرمال پیوسته شامل پاسخ‌های چوله و دومدی استفاده می‌کنیم. این توزیع که آن‌را نرمال دومدی‌ـ‌یک‌مدی چندمتغیره[۲] (MBUN) می‌نامیم، می‌تواند ویژگی‌های متنوع موجود در داده‌ها شامل یک‌مدی یا دومدی بودن، تقارن، چولگی یا ترکیب آن‌ها را توسط پارامترهای خود کنترل و مدل‌بندی کند. همچنین تحت مقادیر مشخصی برای پارامترهای توزیع MBUN، توزیع نرمال چندمتغیره نتیجه خواهد شد. بنابراین رده توزیع‌های نرمال چندمتغیره زیررده‌ای از رده توزیع‌های MBUN است.

در بخش ۲ به معرفی توزیع MBUN می‌پردازیم. در بخش ۳ یک مدل رگرسیون فضایی معرفی می‌کنیم که در آن جمله خطا از یک توزیع MBUN پیروی می‌کند. با این کار می‌توان ماهیت چوله و دومدی موجود در پاسخ‌ها را در مدل لحاظ کرد. ساختار وابستگی فضایی پاسخ‌ها را نیز توسط ماتریس کوواریانس توزیع جمله خطا وارد مدل می‌کنیم. همچنین در این بخش برآوردگرهای ماکسیمم درستنمایی پارامترها را محاسبه می‌کنیم. در بخش چهارم با مثال‌های شبیه‌سازی و واقعی برتری مدل جدید نسبت به نرمال را ارزیابی خواهیم کرد. در پایان نیز بحث و نتیجه‌گیری مقاله را ارایه می‌دهیم.

۲ توزیع نرمال دومدی‌ـ‌یک‌مدی چندمتغیره

توزیع‌های چندمتغیره در بسیاری از علوم مانند زیست‌شناسی، فیزیک، اقتصاد، زمین‌شناسی و علوم پزشکی کاربردهای گسترده‌ای دارند. به همین دلیل، آماردانان به دنبال تعمیم توزیع‌های یک متغیره به چندمتغیره هستند. توزیع نرمال چندمتغیره از پرکاربردترین توزیع‌های چندمتغیره است که در زمینه‌های مختلف از آن استفاده می‌شود. البته این توزیع برای داده‌هایی که چوله، دم سنگین یا چندمدی هستند، مناسب نیست. تلاش‌های مختلفی برای معرفی توزیع‌های چندمتغیره مناسب این نوع داده‌ها صورت گرفته‌اند. به‌عنوان چند نمونه، توزیع چوله نرمال چندمتغیره[۳] توسط آزالینی و وال (۱۹۹۶) ، آزالینی (۲۰۰۵) و جو و همکاران (۲۰۱۲) معرفی و مطالعه شده‌اند. اندرسون (۱۹۹۲) توزیع لاپلاس چندمتغیره را به عنوان حالت خاصی از توزیع لینیک چندمتغیره[۴] معرفی کرد و التوفت و همکاران (۲۰۰۶) درمورد توزیع لاپلاس چندمتغیره به عنوان مدل آمیخته مقیاس نرمال بحث کردند. توزیع لجستیک چندمتغیره توسط هنریک و همکاران (۱۹۷۳) معرفی شد و ین (۲۰۱۰) توزیع چندمتغیره شبه لجستیک[۵] را معرفی کرد. توزیع هایپربولیک تعمیم‌یافته چوله‌نرمال چندمتغیره[۶] توسط ویلکا و همکاران (۲۰۱۴) معرفی شد. همچنین بالاکریشنان و ریستیک (۲۰۱۶) خانواده چندمتغیره تولیدشده گاما را پیشنهاد کردند.

[۱] Mixture
[۲] Multivariate Bimodal-Unimodal Normal
[۳] Multivariate skew normal
[۴] Multivariate linnik
[۵] Multivariate semi-logistic
[۶] Multivariate skew normal generalized hyperbolic

توزیع‌های معرفی‌شده در مدل‌بندی چولگی داده‌های چندمتغیره ممکن است موفق باشند، اما برای داده‌های دومدی انتخاب‌های مناسبی نیستند. توزیع MBUN پیشنهادی ما می‌تواند ویژگی‌های متنوع موجود در داده‌ها شامل یک‌مدی یا دومدی بودن، تقارن، و چولگی را مدل‌بندی کند. توزیع نرمال چندمتغیره نیز یکی از اعضای این خانواده محسوب می‌شود. توزیع MBUN برای بردار تصادفی p متغیره $\boldsymbol{X}$ دارای تابع چگالی به شکل

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\exp\{-\frac{\boldsymbol{k}^\top \boldsymbol{k}}{2}\} \exp\left\{\left[\boldsymbol{k}^\top \left|\Sigma^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right|\right]\right\} \phi_p\left(\boldsymbol{x}, \boldsymbol{\mu}+\Sigma^{\frac{1}{2}}\boldsymbol{a}, \Sigma\right)}{\Pi_{i=1}^{p}\left(\exp\{k_i a_i\}\Phi(k_i+a_i)+\exp\{-k_i a_i\}\Phi(k_i-a_i)\right)} \tag{۱.۲}$$

است که در آن $\phi_p(\cdot, \boldsymbol{\mu}, \Sigma)$ تابع چگالی نرمال چندمتغیره با بردار میانگین $\boldsymbol{\mu}$ و ماتریس کوواریانس Σ و $\Phi(\cdot)$ تابع توزیع نرمال یک‌متغیره استاندارد است. پارامترهای این توزیع سه بردار p بعدی $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^\top$، $\boldsymbol{k} = (k_1, k_2, \ldots, k_p)^\top$ و $\boldsymbol{a} = (a_1, a_2, \ldots, a_p)^\top$ با مولفه‌های حقیقی مقدار و ماتریس $p \times p$ معین مثبت Σ است. اگر بردار تصادفی $\boldsymbol{X}$ از این توزیع پیروی کند، برای نمایش آن از نماد $\boldsymbol{X} \sim \text{MBUN}(\boldsymbol{\mu}, \Sigma, \boldsymbol{k}, \boldsymbol{a})$ استفاده می‌کنیم. همان‌طور که از شکل تابع چگالی (۱.۲) قابل مشاهده است، پارامترهای μ_i، k_i و a_i مختص متغیر iام، $i = 1, \ldots, p$، هستند. مقدار پارامتر k_i دومدی یا یک‌مدی بودن و در صورت دومدی بودن فاصله دو مد را مشخص می‌کند. پارامترهای μ_i و a_i نیز به ترتیب پارامترهای مکان و شکل متغیر متناظر هستند. برای تعیین وابستگی بین متغیرها نیز از ماتریس Σ استفاده می‌شود.

ویژگی اصلی داده‌های فضایی وجود ساختار وابستگی مکانی داده‌ها بر اساس موقعیت قرار گرفتن آن‌ها است. برای پرهیز از پیچیدگی غیرقابل کنترل، در این مقاله فرض می‌کنیم تعداد مدها، چولگی و شکل داده‌ها به موقعیت مکانی داده‌ها وابسته نیستند و ساختار وابستگی فضایی را از طریق ماتریس Σ وارد مدل می‌کنیم. البته این پذیره ساده‌سازی بیش از اندازه نیست، زیرا در عمل نیز ساختار وابستگی را معمولا از طریق پارامترهای مقیاس و مکان وارد مدل می‌کنند. بنابراین به ازای $i = 1, \ldots, p$ می‌پذیریم $k_i = k$ و $a_i = a$. در این‌صورت تابع چگالی (۱.۲) به شکل زیر تبدیل می‌شود:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\exp\left\{-\frac{pk^2}{2}\right\} \exp\left\{\left[k\boldsymbol{1}^\top \left|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right|\right]\right\} \phi_p(\boldsymbol{x}, \boldsymbol{\mu}+\Sigma^{\frac{1}{2}}a\boldsymbol{1}, \Sigma)}{\left(\exp\{ka\}\Phi(k+a)+\exp\{-ka\}\Phi(k-a)\right)^p}$$

که در آن $\boldsymbol{1}$ برداری از ۱ها با بعد p است.

برای سهولت در به‌دست آوردن برخی ویژگی‌های توزیع از نسخه استاندارد استفاده می‌کنیم که در آن $\boldsymbol{Z} = \Sigma^{-\frac{1}{2}}(\boldsymbol{X}-\boldsymbol{\mu})$. گشتاور اول و دوم متغیر jام به ترتیب

$$E[Z_j] = \frac{(a_j+k_j)\exp\{k_j a_j\}\Phi(k_j+a_j) + (a_j-k_j)\exp\{-k_j a_j\}\Phi(k_j-a_j)}{\exp\{k_j a_j\}\Phi(k_j+a_j)+\exp\{-k_j a_j\}\Phi(k_j-a_j)}$$

و

$$\begin{aligned} E\left[Z_j^2\right] &= \frac{\left(1+(a_j+k_j)^2\right)\exp\{k_j a_j\}\Phi(k_j+a_j) + \left(1+(a_j-k_j)^2\right)\exp\{-k_j a_j\}\Phi(k_j-a_j)}{\exp\{k_j a_j\}\Phi(k_j+a_j)+\exp\{-k_j a_j\}\Phi(k_j-a_j)} \\ &+ \frac{\frac{k_j}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(a_j^2+k_j^2\right)\right\}}{\exp\{k_j a_j\}\Phi(k_j+a_j)+\exp\{-k_j a_j\}\Phi(k_j-a_j)} \end{aligned}$$

می‌باشند. بردار میانگین و ماتریس کوواریانس MBUN نیز با توجه به $\boldsymbol{X} = \Sigma^{\frac{1}{2}}\boldsymbol{Z} + \boldsymbol{\mu}$ عبارتند از

$$\begin{aligned} E[\boldsymbol{X}] &= \Sigma^{\frac{1}{2}}E[\boldsymbol{Z}] + \boldsymbol{\mu} \\ \text{Var}[\boldsymbol{X}] &= \Sigma^{\frac{1}{2}}\text{Var}[\boldsymbol{Z}]\Sigma^{\frac{1}{2}}. \end{aligned}$$

با فرض $a_i = a$ و $k_i = k$، می‌توان نوشت

$$\begin{aligned} E[\boldsymbol{X}] &= h_1(k,a)\Sigma^{\frac{1}{2}}\mathbf{1} + \boldsymbol{\mu} \\ \mathrm{Var}[\boldsymbol{X}] &= h_2(k,a)\Sigma \end{aligned}$$

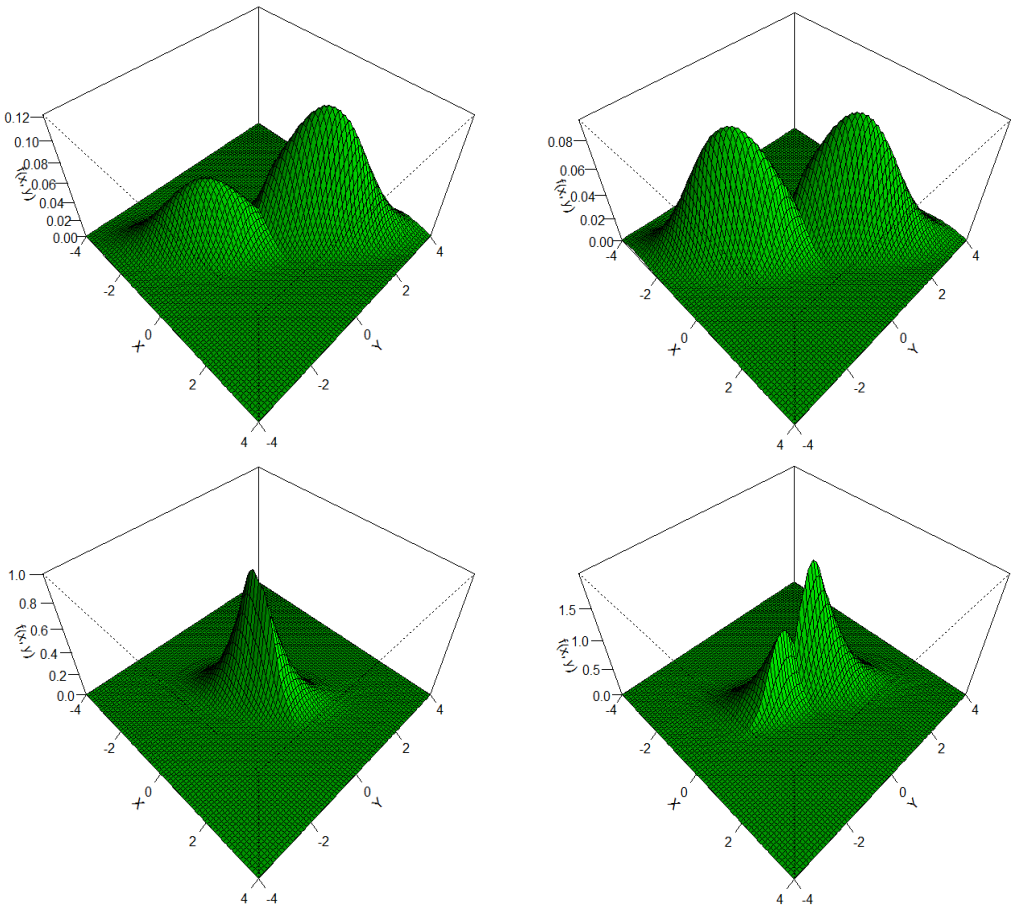
که در آن

$$h_1(k,a) = \frac{(a+k)\exp\{ka\}\Phi(k+a) + (a-k)\exp\{-ka\}\Phi(k-a)}{\exp\{ka\}\Phi(k+a) + \exp\{-ka\}\Phi(k-a)}$$

و

$$\begin{aligned} h_2(k,a) = & \frac{\left(1+(a+k)^2\right)\exp\{ka\}\Phi(k+a) + \left(1+(a-k)^2\right)\exp\{-ka\}\Phi(k-a)}{\exp\{ka\}\Phi(k+a) + \exp\{-ka\}\Phi(k-a)} \\ & + \frac{\frac{k}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(a^2+k^2\right)\right\}}{\exp\{ka\}\Phi(k+a) + \exp\{-ka\}\Phi(k-a)} - h_1(k,a)^2. \end{aligned}$$

منحنی تابع چگالی (۱.۲) با دو متغیر را در شکل ۱ برای مقادیر مختلف پارامترها می‌توانید مشاهده کنید.



شکل ۱: منحنی تابع چگالی توزیع MBUN برای $p = 2$ به‌ازای مقادیر مختلف پارامترهای توزیع

۳ رگرسیون فضایی پاسخ‌های دومدی

میدان تصادفی $\{Z(s);\ s \in D\}$ را در نظر بگیرید که در آن ناحیه D یک زیرمجموعه از فضای اقلیدسی $d \geq ۲$ بعدی R^d است. میدان تصادفی $Z(\cdot)$ را به‌صورت

$$Z(s) = \mu(s) + \delta(s) \qquad (۱.۳)$$

تجزیه می‌کنیم که در آن $\mu(\cdot)$ تغییرات مقیاس بزرگ[۷] یا روند[۸] و $\delta(\cdot)$ تغییرات مقیاس کوچک[۹] یا فرآیند خطای میدان تصادفی نامیده می‌شود. در تحلیل رگرسیون فضایی با n مشاهده از میدان تصادفی و با حضور متغیرهای تبیینی، می‌توان روند را به صورت تابع خطی از ضرایب رگرسیونی و متغیرهای تبیینی به صورت

$$\mu(s_i) = \boldsymbol{x}(s_i)^\top \boldsymbol{\beta} \qquad i = ۱, \ldots, n$$

نشان داد که در آن $\boldsymbol{\beta} = (\beta_۱, \ldots, \beta_q)^\top$ و $\boldsymbol{x}(s_i)$ بردار متغیرهای تبیینی q بعدی با درایه‌های $x_j(s_i)$، برای $j = ۱, \ldots, q$، است. برای در نظر گرفتن پارامتر عرض از مبدا در مدل، می‌توان قرار داد $x_۱(s_i) = ۱$. در مدل (۱.۳) فرض می‌کنیم

$$\boldsymbol{\delta} \sim \text{MBUN}\,(۰, \Sigma, k ۱_n, a ۱_n)$$

که در آن $\boldsymbol{\delta} = (\delta(s_۱), \ldots, \delta(s_n))^\top$ و اندیس n در ۱_n برای تاکید بر تعداد موقعیت‌های فضایی مشاهده‌شده است.

ساختار وابستگی فضایی را از طریق ماتریس مقیاس معین مثبت Σ وارد مدل می‌کنیم. همچنین فرض می‌کنیم ساختار وابستگی فضایی، مانا[۱۰] باشد. برای این کار، مشابه مدل رگرسیون با فرآیند خطای فضایی نرمال، از توابع پارامتری هم‌تغییرنگار[۱۱]

$$\sigma(h; \boldsymbol{\theta}) = \text{Cov}_{\boldsymbol{\theta}}[\delta(s), \delta(s+h)]$$

استفاده می‌کنیم. به‌عنوان مثالی از این نوع توابع، می‌توان دو مدل نمایی و گاوسی را نام برد که به ترتیب به‌صورت زیر تعریف می‌شوند (کرسی ، ۱۹۹۳) :

$$\sigma(h; \theta_۱, \theta_۲) = \theta_۱ \exp\{-\|h\|/\theta_۲\}$$

و

$$\sigma(h; \theta_۱, \theta_۲) = \theta_۱ \exp\{-\|h\|^۲/\theta_۲^۲\}$$

که در آن‌ها پارامترهای وابستگی $\theta_۱$ و $\theta_۲$ به ترتیب ازاره[۱۲] و دامنه هستند.

---

[۷] Large scale variation
[۸] Trend
[۹] Small scale variation
[۱۰] Stationary
[۱۱] Covariogram
[۱۲] Sill

۱.۳ برازش مدل: رهیافت ماکسیمم درستنمایی

مدل خطی (۱.۳) را در نظر بگیرید. تابع لگاریتم درستنمایی مدل با فرآیند خطای فضایی $\boldsymbol{\delta} \sim \text{MBUN}\left(۰, \Sigma_h^{\boldsymbol{\theta}}, k۱_n, a۱_n\right)$ به‌صورت

$$\begin{aligned}\ell\left(\boldsymbol{\zeta}\right) &= -\frac{nk^{۲}}{۲} - n\log\left(\exp\{ka\}\Phi(k+a) + \exp\{-ka\}\Phi(k-a)\right) + k۱_n^{\top}\left|\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\right)\right| - \\ &\frac{n}{۲}\log\left(۲\pi\right) - \frac{۱}{۲}\log\left(\det\left(\Sigma_h^{\boldsymbol{\theta}}\right)\right) - \frac{۱}{۲}\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}-a\Sigma_h^{\boldsymbol{\theta}^{\frac{۱}{۲}}}۱_n\right)^{\top}\Sigma_h^{\boldsymbol{\theta}^{-۱}}\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}-a\Sigma_h^{\boldsymbol{\theta}^{\frac{۱}{۲}}}۱_n\right)\end{aligned}$$

است، که در آن $\boldsymbol{\zeta} = \left(\boldsymbol{\beta}^{\top}, \boldsymbol{\theta}^{\top}, k, a\right)^{\top}$ بردار پارامترها و $\Sigma_h^{\boldsymbol{\theta}}$ ماتریس مقیاس معتبر با بردار پارامتر $\boldsymbol{\theta}$ است. بردار امتیاز این مدل

$$\boldsymbol{U}_n = \left(\left(\frac{\partial\ell\left(\boldsymbol{\zeta}\right)}{\partial\boldsymbol{\beta}}\right)^{\top}, \left(\frac{\partial\ell\left(\boldsymbol{\zeta}\right)}{\partial\boldsymbol{\theta}}\right)^{\top}, \frac{\partial\ell\left(\boldsymbol{\zeta}\right)}{\partial k}, \frac{\partial\ell\left(\boldsymbol{\zeta}\right)}{\partial a}\right)^{\top}$$

است، به‌طوری که

$$\begin{aligned}\frac{\partial\ell\left(\boldsymbol{\zeta}\right)}{\partial\boldsymbol{\beta}} &= -k\boldsymbol{X}'\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}\text{sign}\left(\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\right)\right) + \boldsymbol{X}^{\top}\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}-a\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}۱_n\right)^{\top}\Sigma_h^{\boldsymbol{\theta}^{-۱}} \\ \frac{\partial\ell\left(\boldsymbol{\zeta}\right)}{\partial\theta_j} &= -\left(k۱_n^{\top}\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}\frac{\partial\Sigma_h^{\boldsymbol{\theta}^{\frac{۱}{۲}}}}{\partial\theta_j}\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}\text{sgn}\left(\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\right)\right) + \frac{۱}{۲}\text{tr}\left(\Sigma_h^{\boldsymbol{\theta}^{-۱}}\frac{\partial\Sigma_h^{\boldsymbol{\theta}}}{\partial\theta_j}\right)\right. \\ &\left. -\frac{۱}{۲}\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\right)^{\top}\Sigma_h^{\boldsymbol{\theta}^{-۱}}\frac{\partial\Sigma_h^{\boldsymbol{\theta}}}{\partial\theta_j}\Sigma_h^{\boldsymbol{\theta}^{-۱}}\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\right) + a\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\right)^{\top}\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}\frac{\partial\Sigma_h^{\boldsymbol{\theta}^{\frac{۱}{۲}}}}{\partial\theta_j}\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}۱_n\right) \\ \frac{\partial\ell\left(\boldsymbol{\zeta}\right)}{\partial k} &= -nk - na\frac{a\exp\{ka\}\Phi\left(k+a\right) - a\exp\{-ka\}\Phi\left(k-a\right) + ۲k\exp\{ka\}\phi\left(k+a\right)}{\exp\{ka\}\Phi\left(k+a\right) + \exp\{-ka\}\Phi\left(k-a\right)} \\ &+۱_n^{\top}\left|\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}\left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\right)\right| \\ \frac{\partial\ell\left(\boldsymbol{\zeta}\right)}{\partial a} &= -na - nk\frac{\exp\{ka\}\Phi\left(k+a\right) - \exp\{-ka\}\Phi\left(k-a\right)}{\exp\{ka\}\Phi\left(k+a\right) + \exp\{-ka\}\Phi\left(k-a\right)} + \left(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\right)^{\top}\Sigma_h^{\boldsymbol{\theta}^{-\frac{۱}{۲}}}۱_n\end{aligned}$$

که در آن $\boldsymbol{X}$ ماتریس طرح حاصل از متغیرهای تبیینی با بعد $n \times q$ است و تابع $\text{sgn}(\cdot)$ تابع علامت است. بدیهی است که برآوردگرهای ماکسیمم درستنمایی پارامترهای مدل صورت بسته ندارند و با استفاده از روش‌های عددی حل دستگاه معادلات $\boldsymbol{U}_n = ۰$ برآوردها را محاسبه می‌کنیم.

۴ ارزیابی کارایی مدل پیشنهادی

در این بخش با استفاده از مثال‌های شبیه‌سازی و واقعی، عملکرد مدل پیشنهادی را برای داده‌های فضایی دومدی نسبت به مدل جانشین نرمال مقایسه می‌کنیم.

۱.۴ مطالعه شبیه‌سازی

در این مثال داده‌ها را از مدل

$$z(s_i) = x_۱(s_i) + ۲x_۲(s_i) + \delta(s_i), \quad i = ۱, \ldots, n$$

تولید شد، که در آن $\boldsymbol{\delta} = (\delta(s_1), \ldots, \delta(s_n))^\top$ توسط یک فرآیند آمیخته از دو توزیع نرمال n متغیره به صورت

$$\frac{1}{2} N_n(\boldsymbol{\mu}_1, \Sigma) + \frac{1}{2} N_n(\boldsymbol{\mu}_2, \Sigma)$$

شبیه‌سازی شد، که در آن $\boldsymbol{\mu}_1$ و $\boldsymbol{\mu}_2$ بردارهای n بعدی ثابت با مولفه‌های به ترتیب ۲ و -2 هستند و ماتریس کوواریانس Σ توسط یک مدل نمایی با پارامترهای $\theta_1 = 1/5$ و $\theta_2 = 0/5$ روی یک مشبکه منظم ساخته شده است. این نوع شبیه‌سازی فرآیند $\delta(\cdot)$ منجر به تولید داده‌های دومدی می‌شود. متغیرهای تبیینی x_1 و x_2 از توزیع نرمال استاندارد شبیه‌سازی شدند و سه حجم نمونه $n = 50, 100, 200$ نیز در نظر گرفته شدند.

برای برازش مدل، باید پارامتر مکانی مدل MBUN را برابر $z - \beta_1 \boldsymbol{x}_1 - \beta_2 \boldsymbol{x}_2$ قرار دهیم. برای ارزیابی برآوردهای مدل MBUN و مقایسه با برآوردهای حاصل از مدل نرمال (MN)، شبیه‌سازی را ۱۰۰ بار تکرار کردیم. مقادیر اریبی و MSE[13] برآوردهای ماکسیمم درستنمایی ضرایب رگرسیونی برای هر دو مدل در جدول ۱ گزارش شده‌اند. با توجه به مقادیر جدول، واضح است که مدل MBUN، بر حسب هر دو معیار اریبی و MSE، برآوردهای کاراتری نسبت به مدل MN ارایه کرده است. با مقایسه مقادیر MSE، مدل MBUN حداقل ۶۶ درصد کارایی بیشتری دارد. این وضعیت با افزایش حجم نمونه نیز بارزتر می‌شود.

دو مدل را از منظر دو معیار انتخاب مدل AIC[14] (آکاییک ، ۱۹۷۳) و BIC[15] (اسکووارز ، ۱۹۷۸) نیز مورد مقایسه قرار دادیم. شکل ۲ مقادیر این دو معیار را برای ۱۰۰ تکرار شبیه‌سازی برای دو مدل MBUN و MN نمایش می‌دهد. پراکندگی مقادیر هر دو معیار گویای توصیف بهتر داده‌ها توسط مدل MBUN است. بنابراین عدم توانایی مدل نرمال و در مقابل کارایی مدل MBUN در برازش پاسخ‌های فضایی دومدی قابل مشاهده و نتیجه‌گیری است.

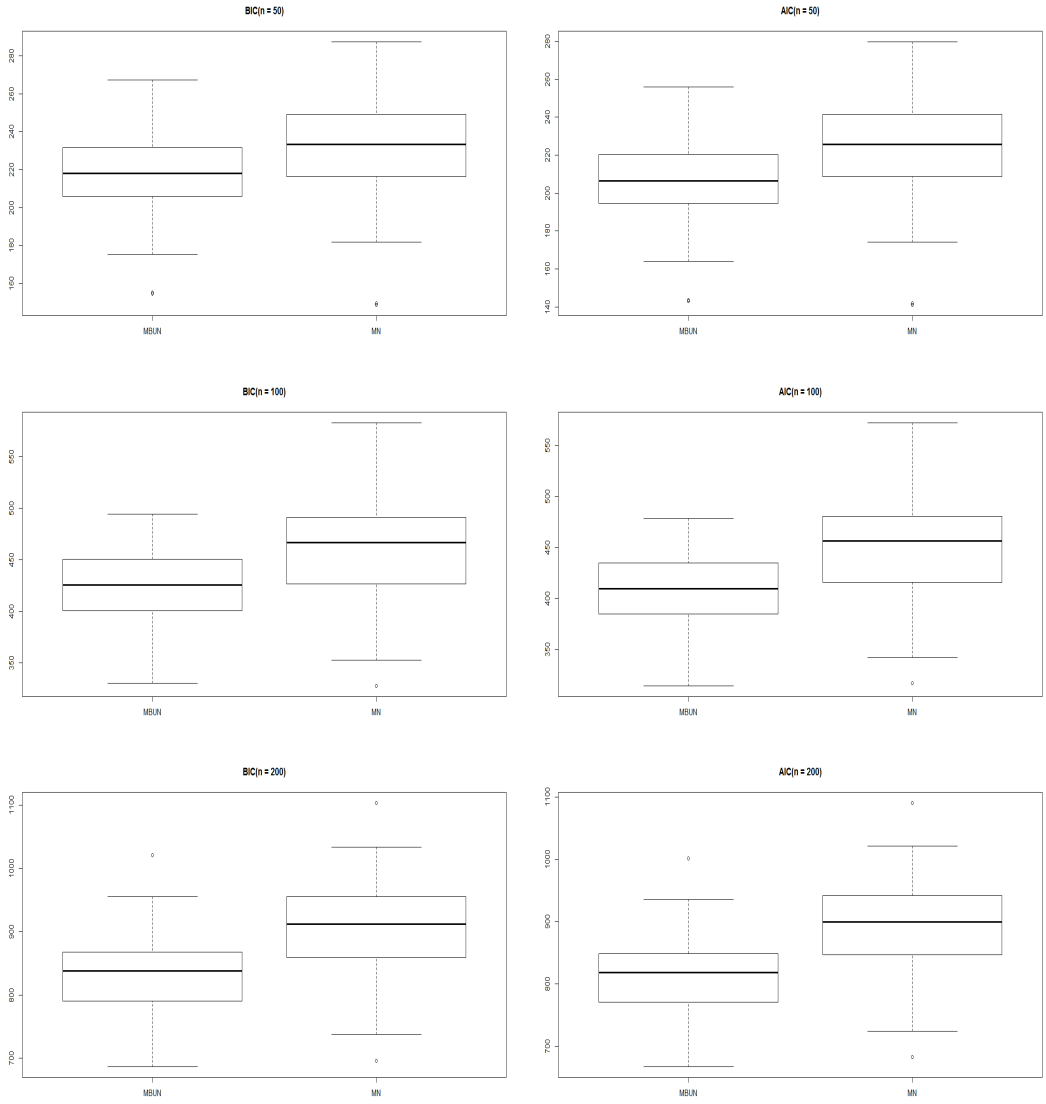
جدول ۱: مقادیر اریبی و MSE برآوردهای ضرایب رگرسیونی در مثال شبیه‌سازی

		n = ۵۰		۱۰۰		۲۰۰	
مدل		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
MBUN	اریبی	−۰/۰۰۸۴	۰/۰۴۳۷	۰/۰۱۹۲	۰/۰۲۱۶	۰/۰۰۱۲	−۰/۰۱۴۴
	MSE	۰/۰۵۰۱	۰/۰۵۷۸	۰/۰۲۲۶	۰/۰۱۸۱	۰/۰۰۹۳	۰/۰۱۳۰
MN	اریبی	−۰/۰۶۰۸	۰/۰۶۰۱	۰/۰۵۴۴	۰/۰۲۳۷	−۰/۰۱۲۴	−۰/۰۰۱۴
	MSE	۰/۱۳۰۸	۰/۱۰۱۱	۰/۰۹۸۹	۰/۰۴۲۸	۰/۰۲۵۸	۰/۰۲۲۲

۲.۴ مثال‌های واقعی

در این بخش عملکرد و کاربرد مدل MBUN را در برازش دو مثال واقعی مورد بررسی قرار می‌دهیم و از دو معیار AIC و BIC برای مقایسه با مدل نرمال استفاده می‌کنیم. اولین مثال مجموعه داده غلظت فلزات سنگین است که در نزدیکی رودخانه میوز[16]

[13] Mean Squared Error
[14] Akaike Information Criterion
[15] Bayesian Information Criterion
[16] Meuse



شکل ۲: مقادیر AIC (ستون سمت راست) و BIC (ستون سمت چپ) برای دو مدل MBUN و MN به ازای حجم‌های نمونه مختلف

نمونه‌برداری شده است و در بسته georob[۱۷] در نرم‌افزار R با نام Meuse در دسترس است. این مجموعه داده شامل ۱۴ متغیر و ۱۶۴ مشاهده است. مدل خطی که در نظر گرفتیم، به صورت

$$\log(\text{zinc}) = \beta_۱ + \beta_۲\sqrt{\text{dist}} + \beta_۳\text{cadmium} + \delta$$

است که در آن zinc غلظت عنصر روی، dist فاصله تا رودخانه و cadmium غلظت کادمیم سطح خاک است.

مجموعه داده دوم مربوط به خواص شیمیایی اندازه‌گیری‌شده خاک با ۲۵۰ مشاهده و ۲۲ متغیر است که در بسته geoR[۱۸] در نرم‌افزار R با نام soil250 موجود است. مدل مورد نظر برای این داده‌ها به صورت

$$\text{CTC} = \beta_۱ + \beta_۲\text{K} + \beta_۳\text{Ca} + \delta$$

[۱۷] https://cran.r-project.org/package=georob

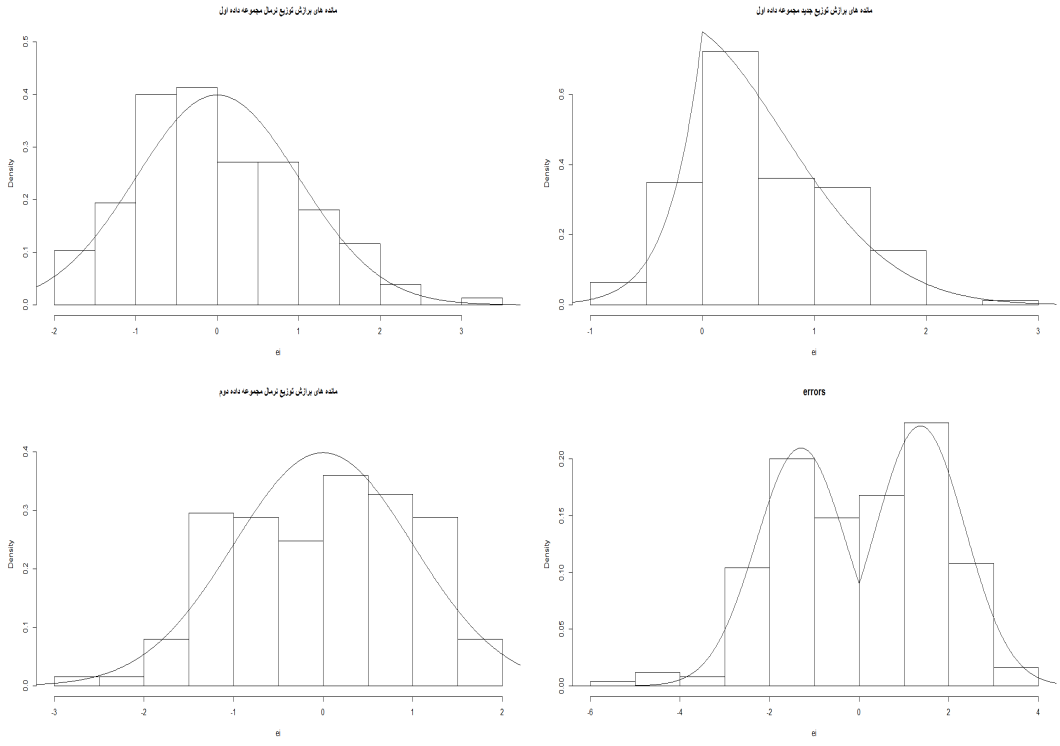
[۱۸] https://cran.r-project.org/package=geoR

است که در آن CTC توانایی تبادل کاتیوم، K مقدار پتاسیم و Ca مقدار کلسیم موجود در خاک است. برای هر دو مدل از مدل هم‌تغییرنگار نمایی استفاده کردیم. نتایج برآورد پارامترهای رگرسیونی هر دو مدل به همراه معیارهای AIC و BIC در جدول ۲ گزارش شده‌اند. با توجه به نتایج جدول، برای هر دو مثال، مدل MBUN نسبت به MN مدل برتر است.

جدول ۲: برآورد پارامترهای رگرسیونی و معیارهای انتخاب مدل دو مجموعه داده واقعی

مجموعه داده	مدل	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	AIC	BIC
Mesue	MBUN	۵/۷۵	−۱/۱۴	۰/۱۱	۱۰۵/۱۲	۱۲۶/۴۲
	MN	۶/۰۲	−۱/۱۹	۰/۱۲	۱۱۲/۰۸	۱۲۷/۳۰
soil250	MBUN	۲/۷۰	۱/۳۵	۱/۲۷	۳۱۷/۳۹	۳۴۲/۰۴
	MN	۳/۸۵	۰/۶۱	۱/۰۲	۴۵۶/۳۳	۴۷۳/۹۳

با استفاده از هیستوگرام مانده‌های استاندارد شده و منحنی‌های دو تابع چگالی برآوردشده (شکل ۳)، نیکویی برازش مدل MBUN کاملا مشهود است. افزون بر این، عدم نیکویی برازش مدل نرمال در هر دو مثال نیز واضح است؛ در مجموعه داده Meuse مانده‌های مدل نرمال چوله هستند و در مجموعه داده soil250 مانده‌های مدل نرمال دومدی هستند. بنابراین واضح است که مدل MBUN انعطاف لازم برای مدل‌بندی هر دو ویژگی چولگی و دومدی بودن داده‌ها را داراست.



شکل ۳: برازش توابع چگالی برآوردشده MBUN و MN بر مانده‌های استانداردشده دو مدل در دو مثال واقعی

۵ بحث و نتیجه‌گیری

در این مقاله، در مورد تحلیل رگرسیون خطی داده‌های فضایی چوله و دومدی بحث کردیم و توزیع چندمتغیره منعطف MBUN را معرفی کردیم. همچنین رهیافت ماکسیمم درستنمایی برای برازش مدل را ارایه دادیم. با مثال‌های شبیه‌سازی و واقعی عملکرد کارا و بهتر مدل پیشنهادی را نسبت به مدل نرمال زمانی که داده‌ها چوله یا دومدی هستند، نشان دادیم.

یکی از مباحثی که در تحلیل داده‌های زمین‌آماری مطرح و مهم است، پیشگویی فضایی در مکان‌هایی است که فاقد مشاهده هستند. این مساله برای داده‌هایی که دومدی یا چوله هستند، می‌تواند انگیزه‌ای برای پژوهش بر روی آن توسط مدل پیشنهادی ما باشد.

مراجع

محمدزاده م، (۱۳۹۴). *آمار فضایی و کاربردهای آن*، چاپ اول، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in Petrov, B.N. and Csáki, F., 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, 267–281.

Anderson, D. N. (1992). A multivariate Linnik distribution, *Statistics and probability letters*, **14**(4), 333-336.

Arellano-Valle, R. B., Bolfarine, H. and Lachos, V. H. (2005). Skew-normal linear mixed models, *Journal of Data Science*, **3**(4), 415-438.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families, *Scandinavian Journal of Statistics*, **32**(2), 159-188.

Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution, *Biometrika*, **83**(4), 715-726.

Balakrishnan, N. and Ristić, M. M. (2016). Multivariate families of gamma-generated distributions with finite or infinite support above or below the diagonal, *Journal of Multivariate Analysis*, **143**, 194-207.

Bolfarine, H., and Lachos, V. H. (2007). Skew-probit measurement error models, *Statistical Methodology*, **4**(1), 1-12.

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

Eltoft, T., Kim, T. and Lee, T. W. (2006). On the multivariate Laplace distribution, *IEEE Signal Processing Letters*, **13**(5), 300-303.

Fallah, B. and Sodoudi, S. (2015). Bimodality and regime behavior in atmosphere–ocean interactions during the recent climate change, *Dynamics of Atmospheres and Oceans*, **70**, 1-11.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing, *Journal of the American Statistical Association*, **100**(471), 1021-1035.

Griffin, J. E. and Steel, M. J. (2006). Order-based dependent Dirichlet processes, *Journal of the American Statistical Association*, **101**(473), 179-194.

Hosseini, F., Eidsvik, J. and Mohammadzadeh, M. (2011). Approximate Bayesian inference in spatial GLMM with skew normal latent variables, *Computational Statistics and Data Analysis*, **55**(4), 1791-1806.

Hosseini, F. and Mohammadzadeh, M. (2012). Bayesian prediction for spatial generalized linear mixed models with closed skew normal latent variables, *Australian and New Zealand Journal of Statistics*, **54**(1), 43-62.

Malik, H. J. and Abraham, B. (1973). Multivariate logistic distributions, *The Annals of Statistics*, **1**(3), 588-590.

Joe, H., Seshadri, V. and Arnold, B. C. (2012). Multivariate inverse Gaussian and skew-normal densities, *Statistics and Probability Letters*, **82**(12), 2244-2251.

Karimi O., Omre H. and Mohammadzadeh, M. (2010). Bayesian closed-skew Gaussian inversion of seismic AVO data for elastic material properties, *Geophysics*, **75**, 185–198.

Sahu, S. K., Dey, D. K. and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models, *Canadian Journal of Statistics*, **31**(2), 129-150.

Schwarz, G. E. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6** (2), 461–464.

Vilca, F., Balakrishnan, N. and Zeller, C. B. (2014). Multivariate skew-normal generalized hyperbolic distribution and its properties, *Journal of Multivariate Analysis*, **128**, 73-85.

Yeh, H. C. (2010). Multivariate semi-logistic distributions, *Journal of Multivariate Analysis*, **101**(4), 893-908.

دومین سمینار آمار فضایی و کاربردهای آن
۳ – ۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1990

برآورد و پیش‌گویی با مدل‌های رگرسیو ـ اتورگرسیو فضایی

حسن ایتام، یدا... واقعی*، حمیدرضا نیلی‌ثانی

گروه آمار دانشگاه بیرجند

چکیده: در سال‌های اخیر مدل‌های مختلفی برای مدل‌سازی و پیش‌گویی داده‌های فضایی مورد استفاده قرار گرفته‌اند که برخی از آن‌ها (مانند مدل رگرسیون فضایی) شامل متغیرهای مستقل‌اند و برخی دیگر (مانند مدل‌های اتورگرسیو یک‌طرفه) متغیر مستقل ندارند.

در این مقاله چند مدل مختلف فضایی را به اختصار معرفی می‌کنیم و سپس به برآورد پارامترها و نیز پیش‌گویی با مدل رگرسیو ـ اتورگرسیو می‌پردازیم و در انتها با استفاده از یک مجموعه داده فضایی نامنظم چگونگی برآورد پارامترها و پیش‌گویی داده‌ها را به کمک بسته **spdep** از نرم‌افزار **R** بیان می‌کنیم.

واژه‌های کلیدی: مدل‌های فضایی، مدل اتورگرسیو فضایی، پیش‌گویی فضایی، مدل رگرسیو ـ اتورگرسیو.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62M30.

۱ مقدمه

در برخی علوم مانند محیط زیست، جغرافیا، سنجش از دور و . . .، با داده‌هایی مواجه می‌شویم که مستقل از یکدیگر نیستند و وابستگی آن‌ها به نوعی ناشی از موقعیت و مکان قرار گرفتن داده‌ها در فضای مورد بررسی است. این گونه داده‌ها، داده‌های فضایی نامیده می‌شوند. در آمار فضایی متغیر مورد اندازه گیری ممکن است گسسته یا پیوسته باشد. فضای موقعیت یا مکان مشاهدات نیز ممکن است پیوسته یا گسسته، نقطه‌ای یا ناحیه‌ای، منظم یا نامنظم باشد. وقتی مقدار متغیر در یک ناحیه ثبت می‌شود، موقعیت ناحیه‌ای و اگر مقدار متغیر در یک نقطه ثابت با طول و عرض جغرافیایی معین اندازه گیری شود، موقعیت آن نقطه‌ای است. در موقعیت نقطه‌ای اگر نقاط در فواصل مساوی از هم قرار داشته باشند، موقعیت منظم و در غیر این صورت نامنظم‌اند. به داده‌های فضایی که موقعیت آن‌ها ناحیه‌ای یا منطقه‌ای باشد (مانند نرخ بیکاری در شهرستان‌های مختلف کشور) داده‌های مشبکه‌ای گفته می‌شود که شامل داده‌های مشبکه‌ای منظم و نامظم است.

* **ارایه‌دهنده مقاله:** یدا...واقعی، ywaghei@birjand.ac.ir

کریگیدن یک روش متداول در آمار فضایی است که برای داده‌های فضایی با موقعیت نقطه‌ای قابل استفاده است. هر چند بعضاً برای داده‌های مشبکه‌ای هم از کریگیدن استفاده شده است. ولی اساساً کریگیدن یک روش مدل‌سازی نیست و اغلب از آن به عنوان روش پیش‌گویی یاد می‌شود. برای مدل‌سازی داده‌های فضایی مدل‌های مختلفی پیشنهاد شده‌اند. برخی مدل‌ها بین مقادیر متغیر فضایی در موقعیت‌های مختلف رابطه برقرار می‌کنند که به آن‌ها مدل اتورگرسیو می‌گویند. اولین بار ویتل (۱۹۵۴) با فرض اینکه متغیرهای مورد مطالعه بر روی یک شبکه منظم $m \times n$ مشاهده شده و خطاهای هر موقعیت فقط وابسته به موقعیت‌های قبلی، بعدی، بالا و پایین باشند، وابستگی فضایی را در یک مدل اتورگرسیو فضایی[۱] (SAR) لحاظ کرد. بسج (۱۹۷۴)، اورد (۱۹۷۵)، انسلین (۱۹۸۸) همچنین هاینینگ (۱۹۹۰)، کرسی (۱۹۹۳) و والر و گاتوی (۲۰۰۴) نیز به تحلیل و مدل‌بندی داده‌های فضایی با استفاده از مدل‌های اتورگرسیو پرداخته‌اند.

یک نوع از مدل‌های اتورگرسیو فضایی بنام مدل‌های اتورگرسیو یک‌طرفه برای داده‌های مشبکه‌ای منظم هستند، در مقابل مدل‌های اتورگرسیو دیگری برای داده‌های مشبکه‌ای نامنظم معرفی شده‌اند که مشهورترین آن‌ها مدل اتورگرسیو همزمان، مدل رگرسیو ـ اتورگرسیو (یا تأخیر فضایی)، مدل خطای فضایی هستند که مدل نخست متغیر مستقل ندارد ولی در مدل دوم و سوم متغیرهای مستقل هم حضور دارند. مطالبی که در این مقاله می‌آید در راستای مدل‌سازی داده‌های مشبکه‌ای نامنظم شامل بخش‌های زیر است. در بخش بعد چند تا از مدل‌های فضایی را به اختصار معرفی می‌کنیم. در بخش سوم برآورد پارامترهای مدل رگرسیو ـ اتورگرسیو را بدست آوردیم و پیش‌گویی فضایی با این مدل را در بخش چهارم بیان کرده و در بخش پایانی با یک مثال کاربردی برآورد و پیش‌گویی مدل رگرسیو ـ اتورگرسیو را به کمک نرم‌افزار بدست آوردیم.

۲ مدل‌های فضایی

در این بخش مدل‌های رگرسیون فضایی، اتورگرسیو همزمان فضایی، رگرسیو ـ اتورگرسیو و خطای فضایی را به اختصار معرفی می‌کنیم.

۱.۲ مدل رگرسیون فضایی

به طور معمول در تحلیل مدل رگرسیونی فرض بر این است که خطاهای مدل ناهمبسته‌اند. اما در بسیاری از زمینه‌های کاربردی مانند زمین‌شناسی، علوم جغرافیا، همه‌گیر شناسی و پردازش تصاویر، با مواردی مواجه می‌شویم که فرض ناهمبسته بودن خطاها بدلیل همبستگی ناشی از مجاورت داده‌ها، برقرار نیست. به‌علاوه تحقق‌های متغیرهای مستقل یک مدل رگرسیونی ممکن است در موقعیت‌های فضایی مختلفی که مشاهده می‌شوند، متفاوت باشند. در حالت کلی، بر حسب آن‌که جمله‌ی خطا، متغیرهای توضیحی یا ضرایب رگرسیونی توابعی از موقعیت‌های فضایی باشند یا نه، انواع مختلفی از مدل‌های رگرسیون خطی برای داده‌های فضایی پیدا می‌شوند. یکی از متداول‌ترین مدل‌های فضایی که تشابه زیادی با رگرسیون خطی چند گانه دارد، مدل رگرسیون خطی فضایی است که بصورت زیر می‌باشد.

$$Z(s_j) = \sum_{i=1}^{k} \beta_i x_i(s_j) + \delta(s_j), \qquad s \in D, j = 1, 2, \ldots, n \tag{۱.۲}$$

[۱] Spatial AutoRegressive

که در آن $x_i(.) : i = ۱, \dots k$ها دسته‌ای از k متغیر مستقل غیرتصادفی، β_iها ضریب رگرسیونی و $\delta(.)$ جمله‌ی خطا با میانگین صفر و واریانس متناهی است، که معمولاً دارای همبسته فضایی است. تفاوت اساسی مدل رگرسیون خطی فضایی (۱.۲) با مدل رگرسیون خطی چندگانه در این است که در مدل رگرسیون خطی چندگانه جملات خطا از یکدیگر مستقل می‌باشند ولی در مدل (۱.۲) بدلیل وابستگی فضایی متغیرهای $Z(s_i)$، جملات خطا $\delta(s_i)$ وابسته می‌باشند، به همین دلیل برآورد پارامترهای این مدل مشکل‌تر از مدل رگرسیون خطی چندگانه است. محمدزاده (۱۳۹۴) با استفاده از یک مدل کواریانس مشخص برای وابستگی خطاها، برآورد پارامترهای مدل (۱.۲) را به دست آورد.

۲.۲ مدل اتورگرسیو همزمان فضایی

یکی ازساده‌ترین مدل‌های اتورگرسیو فضایی، مدل اتورگرسیو همزمان فضایی [۲] (SSAM) است که به صورت زیر تعریف می‌شود:

$$Z(s_i) = \mu_i + \sum_j \rho_{ij}(Z(s_j) - \mu_j) + \epsilon_j, \qquad i = ۱, \dots, n, \qquad i \neq j, \qquad \mu_i = \mu(s_i) \tag{۲.۲}$$

با قرار دادن $\rho_{ij} = \rho W_{ij}$ و $\mu_i = ۰$ معادله (۲.۲) را به فرم ماتریسی زیر بیان می‌کنیم.

$$\mathbf{z} = \rho \mathbf{W}\mathbf{z} + \varepsilon, \qquad \varepsilon \sim N(۰, \sigma^۲ \mathbf{I}) \tag{۳.۲}$$

که در آن $\mathbf{z} = (Z(s_۱), \dots, Z(s_n))$ بردار متغیرهای پاسخ فضایی، ρ ضریب مدل اتورگرسیو فضایی و $\mathbf{W}$ یک ماتریس $n \times n$ از وزن‌های فضایی است ε بردار خطای مدل است (اورد ، ۱۹۷۵).

۱.۲.۲ مدل رگرسیو ـ اتورگرسیو فضایی

با افزودن متغیرهای مستقل در مدل اتورگرسیو همزمان، مدل رگرسیو ـ آتورگرسیو فضایی [۳] به صورت زیر بیان می‌شود:

$$\mathbf{z} = \rho \mathbf{W}\mathbf{z} + \mathbf{X}\boldsymbol{\beta} + \varepsilon, \qquad \varepsilon \sim N(۰, \sigma^۲ \mathbf{I}) \tag{۴.۲}$$

که در آن $\mathbf{z} = (Z(s_۱), \dots, Z(s_n))$ بردار متغیرهاي پاسخ فضایی، ρ ضریب مدل اتورگرسیو فضایی، $\beta_{k\times ۱}$ بردار ضرایب رگرسیون و ε بردار خطای مدل اتورگرسیو فضایی و $\mathbf{X}_{n\times(k+۱)}$ ماتریس طرح متشکل از مشاهدات k متغیر مستقل است. در این مدل چون $\mathbf{z}$ از طریق عبارت $\mathbf{X}\boldsymbol{\beta}$ با متغیرهای توضیحی و از طریق عبارت $\rho\mathbf{W}\mathbf{z}$ با خودش رابطه رگرسیونی دارد به آن مدل رگرسیو ـ اتورگرسیو فضایی گفته می‌شود. برخی از آماردانان مانند گولارد و همکاران (۲۰۱۷) این مدل را مدل تأخیر فضایی [۴] (SLM) نامیده‌اند چون بین متغیرهای پاسخ رابطه‌ی اتورگرسیو برقرار است.

۳.۲ مدل خطای فضایی

در مدل‌های اتورگرسیو همزمان و رگرسیو ـ اتورگرسیو خطاها مستقل هستند و وابستگی از طریق عبارت $\rho\mathbf{W}\mathbf{z}$ به مدل القا می‌شود اما همانند مدل‌های میانگین متحرک سرهای زمانی می‌توان فرض کرد که خطاهای مدل همبسته باشند یا به عبارت دیگر بین

[۲] Spatial Simultaneous Autoregressive Model
[۳] Spatial regressive-autoregressive model
[۴] Spatial lag model

خطاهای مدل یک رابطه‌ی اتورگرسیو برقرار باشد در این صورت به آن مدل خطای فضایی[۵] (SEM) گفته می‌شود و به صورت زیر تعریف می‌شود:

$$Z(s_i) = \mu_i + \sum_j \lambda_{ij} U(s_j) + \varepsilon_j, \qquad i = ۱, \ldots, n, \qquad j \neq i, \qquad s \in D \tag{۵.۲}$$

که در آن ε_j ها و $U(s_j)$ ها، میدان‌های خطای فضایی مستقل از یکدیگر هستند.

مدل (۵.۲) را با اضافه کردن متغیرهای توضیحی و با فرض $\lambda_{ij} = \lambda W_{ij}$ می‌توان به فرم ماتریسی زیر بیان کرد.

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \qquad \mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(۰, \sigma^۲ \mathbf{I}) \tag{۶.۲}$$

در اینجا نیز λ مانند ρ ضریب مدل اتورگرسیو فضایی، $\boldsymbol{\varepsilon}$ و $\mathbf{u}$ بردارهای خطای فضایی و $\mathbf{W}_{n\times n}$ ماتریس وزن‌های فضایی هستند. لو و ژنگ (۲۰۱۱) و رسولی (۱۳۹۰) این مدل را مدل خطای فضایی نامیده‌اند ولی نام مدل رگرسیونی با خطای خودهمبسته فضایی برای آن مناسب‌تر است از طرف دیگر این مدل تشابه زیادی به مدل رگرسیون فضایی (۱.۲) دارد و می‌توان آن را حالت خاصی از مدل (۱.۲) دانست.

۳ برآورد پارامترهای مدل رگرسیو ـ اتورگرسیو

در این بخش ابتدا تابع درستنمایی مدل رگرسیون تأخیر فضایی را بدست آورده، سپس طریقه برآورد پارامترها را ارائه می‌کنیم. مدل رگرسیو ـ اتورگرسیو (۴.۲) به صورت زیر ساده می‌شود:

$$\mathbf{z} = (\mathbf{I} - \rho\mathbf{W})^{-۱}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}), \qquad \boldsymbol{\varepsilon} \sim N_n(۰, \sigma^۲\mathbf{I}) \tag{۱.۳}$$

حال امید ریاضی و ماتریس کواریانس $\mathbf{z}$ به صورت

$$E(\mathbf{z}) = (\mathbf{I} - \rho\mathbf{W})^{-۱}\mathbf{X}\boldsymbol{\beta} \tag{۲.۳}$$

$$Var(\mathbf{z}) = \mathrm{Cov}(\mathbf{z}) = \sigma^۲(\mathbf{I} - \rho\mathbf{W})^{-۱}[(\mathbf{I} - \rho\mathbf{W})^{-۱}]^\top \tag{۳.۳}$$

محاسبه می‌شوند. $\mathbf{z}$ یک ترکیب خطی ماتریسی از بردار $\boldsymbol{\varepsilon}$ است و چون $\boldsymbol{\varepsilon}$ دارای توزیع $N_n(۰, \sigma^۲\mathbf{I})$ است پس $\mathbf{z}$ نیز دارای توزیع نرمال چند متغیره با میانگین و ماتریس کواریانس به دست آمده در رابطه‌های (۲.۳) و (۳.۳) است، در نتیجه تابع درستنمایی عبارت است از:

$$\begin{aligned} L(\boldsymbol{\beta}, \rho, \sigma^۲; z) &= (۲\pi\sigma^۲)^{\frac{-n}{۲}} |(\mathbf{I} - \rho\mathbf{W})^{-۱}[(\mathbf{I} - \rho\mathbf{W})^{-۱}]^\top)^{-\frac{۱}{۲}} \\ &\quad \times \exp\left\{-\frac{۱}{۲}[\mathbf{z} - (\mathbf{I} - \rho\mathbf{W})^{-۱}\mathbf{X}\boldsymbol{\beta}]^\top V^{-۱}[\mathbf{z} - (\mathbf{I} - \rho\mathbf{W})^{-۱}\mathbf{X}\boldsymbol{\beta}]\right\} \end{aligned}$$

از آنجا که

$$\begin{aligned} |(\mathbf{I} - \rho\mathbf{W})^{-۱}[(\mathbf{I} - \rho\mathbf{W})^{-۱}]'| &= |(\mathbf{I} - \rho\mathbf{W})^{-۱}| \cdot |[(\mathbf{I} - \rho\mathbf{W})^{-۱}]^\top| \\ &= |\mathbf{I} - \rho\mathbf{W}|^{-۲} \end{aligned}$$

[۵] Spatial Error Model

دترمینان واریانس $\mathbf{z}$ عبارتست از:

$$|\mathrm{Var}(\mathbf{z})| = |\mathrm{Cov}(\mathbf{z})| = \sigma^{۲n}|(\mathbf{I} - \boldsymbol{\rho}\mathbf{W})|^{-۲} \tag{۴.۳}$$

حال با جایگذاری (۴.۳) در تابع درستنمایی و گرفتن لگاریتم از آن داریم:

$$\begin{aligned}\ell(\boldsymbol{\beta}, \rho, \sigma^{۲}; Z) &= \frac{-n}{۲}\ln(۲\pi\sigma^{۲}) - \frac{۱}{۲}\ln|(\mathbf{I} - \rho\mathbf{W})|^{-۲}) \\ &\quad -\frac{۱}{۲\sigma^{۲}}[\mathbf{z} - (\mathbf{I} - \rho\mathbf{W})^{-۱}\mathbf{X}\boldsymbol{\beta}]^{\top}(\mathbf{I} - \rho\mathbf{W})^{\top}(\mathbf{I} - \rho\mathbf{W})[\mathbf{z} - (\mathbf{I} - \rho\mathbf{W})^{-۱}\mathbf{X}\boldsymbol{\beta}]\end{aligned}$$

چون $(\mathbf{I} - \rho\mathbf{W})[\mathbf{z} - (\mathbf{I} - \rho\mathbf{W})^{-۱}\mathbf{X}\beta] = (\mathbf{I} - \rho\mathbf{W})\mathbf{z} - \mathbf{X}\boldsymbol{\beta}$، داریم:

$$\ell(\boldsymbol{\beta}, \rho, \sigma^{۲}; Z) = \frac{-n}{۲}\ln(۲\pi\sigma^{۲}) + \ln|\mathbf{I} - \rho\mathbf{W}| - \frac{۱}{۲\sigma^{۲}}[(\mathbf{I} - \rho\mathbf{W})\mathbf{z} - \mathbf{X}\boldsymbol{\beta}]^{\top}[(\mathbf{I} - \rho\mathbf{W})\mathbf{z} - \mathbf{X}\boldsymbol{\beta}]$$

اکنون با استفاده از تابع درستنمایی نیمرخ برآورد پارامترها در مراحل زیر به دست آورده می‌شوند (رسولی ، ۱۳۹۰):

مرحله اول: یک مدل رگرسیون به فرم

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta}_{۰} + \boldsymbol{\varepsilon}_{۰}, \qquad \boldsymbol{\varepsilon}_{۰} \sim N_n(۰, \sigma^{۲}\mathbf{I}) \tag{۵.۳}$$

برازش داده و $\boldsymbol{\beta}_{۰}$ به روش کمترین توان‌های دوم خطا[۶] به صورت $\hat{\boldsymbol{\beta}}_{۰} = (\mathbf{X}^{\top}\mathbf{X})^{-۱}\mathbf{X}^{\top}\mathbf{z}$ برآورد می‌شود.

مرحله دوم: یک مدل رگرسیون به فرم

$$\mathbf{W}\mathbf{z} = \mathbf{X}\boldsymbol{\beta}_{L} + \boldsymbol{\varepsilon}_{L}, \qquad \boldsymbol{\varepsilon}_{L} \sim N_n(۰, \sigma^{۲}\mathbf{I}) \tag{۶.۳}$$

برازش داده و $\boldsymbol{\beta}_{L}$ به روش کمترین توان‌های دوم خطا به صورت $\hat{\boldsymbol{\beta}}_{L} = (\mathbf{X}^{\top}\mathbf{X})^{-۱}\mathbf{X}^{\top}\mathbf{W}\mathbf{z}$ برآورد می‌شود.

مرحله سوم: مانده‌های دو مدل (۵.۳) و (۶.۳) به صورت

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}}_{۰} &= \mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}_{۰} \\ \hat{\boldsymbol{\varepsilon}}_{L} &= \mathbf{W}\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}_{L}\end{aligned} \tag{۷.۳}$$

محاسبه می‌شوند.

مرحله چهارم: با استفاده از مانده‌های (۷.۳) لگاریتم تابع درستنمایی نیمرخ را به صورت

$$\ell(\rho; \hat{\boldsymbol{\varepsilon}}_{۰}, \hat{\boldsymbol{\varepsilon}}_{L}) = c(\hat{\boldsymbol{\varepsilon}}_{۰}, \hat{\boldsymbol{\varepsilon}}_{L}) - \frac{n}{۲}\ln\left(\frac{۱}{n}(\hat{\boldsymbol{\varepsilon}}_{۰} - \rho\hat{\boldsymbol{\varepsilon}}_{L})^{\top}(\hat{\boldsymbol{\varepsilon}}_{۰} - \rho\hat{\boldsymbol{\varepsilon}}_{L})\right) + \ln|\mathbf{I} - \rho\mathbf{W}| \tag{۸.۳}$$

بازنویسی نموده و از ماکسیمم کردن آن به روش‌های بهینه‌سازی برآورد ρ حاصل می‌شود.

مرحله پنجم: با استفاده از $\hat{\rho}$، پارامترهای $\sigma^{۲}$ و $\boldsymbol{\beta}$ به صورت

$$\begin{aligned}\hat{\sigma^{۲}} &= \frac{۱}{n}(\hat{\boldsymbol{\varepsilon}}_{۰} - \rho\hat{\boldsymbol{\varepsilon}}_{L})^{\top}(\hat{\boldsymbol{\varepsilon}}_{۰} - \rho\hat{\boldsymbol{\varepsilon}}_{L}) \\ \hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}}_{۰} - \rho\hat{\boldsymbol{\beta}}_{L}\end{aligned} \tag{۹.۳}$$

[۶] Least square error

برآورد می‌شوند.

برای ماکسیمم کردن تابع (۸.۳) از روش‌های مختلفی می‌توان استفاده کرد. وقتی حجم نمونه خیلی زیاد باشد به خاطر محاسبه کردن دترمینان در عبارت دوم آن ماکسیمم کردن این تابع ممکن است پیچیده باشد. پیس و باری (۱۹۹۷) یک روش بهینه‌سازی ساده با ارزیابی تابع درستنمایی یک بردار $q \times 1$ از مقادیر ρ در فاصله $[\rho_{\min}, \rho_{\max}]$ پیشنهاد دادند. به این منظور $\ell(\rho_i)$ که از رابطه (۸.۳) به‌دست می‌آید به صورت زیر می‌نویسیم.

$$\ell(\rho_i) = c(\hat{\varepsilon}_0, \hat{\varepsilon}_L) - \frac{n}{2}\ln[\frac{1}{n}(\hat{\varepsilon}_0 - \rho_i\hat{\varepsilon}_L)^\top(\hat{\varepsilon}_0 - \rho_i\hat{\varepsilon}_L)] + \ln|\mathbf{I} - \rho_i\mathbf{W}|, \qquad i = 1, \ldots, q \qquad (۱۰.۳)$$

در این روش یک مجموعه دلخواه (معمولاً یک دنباله عددی) را برای ρ از بازه $[\rho_{\min}, \rho_{\max}]$ انتخاب کرده و مقدار $\ell(\rho_i)$ را حساب می‌کنیم و نهایتاً مقداری از ρ_i که رابطه (۱۰.۳) را ماکسیمم کند یک برآورد ماکسیمم درستنمایی تقریبی برای ρ خواهد بود.

۴ پیش‌گویی فضایی با مدل رگرسیو ـ اتورگرسیو

همان طور که در بخش دوم دیدیم، مدل رگرسیو ـ اتورگرسیو را به صورت زیر می‌نویسیم.

$$\mathbf{z} = \rho\mathbf{W}\mathbf{z} + \mathbf{X}\boldsymbol{\beta} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2\mathbf{I})$$

که می‌توان آن را صرفاً بر حسب ε ، به صورت زیر نوشت:

$$\mathbf{z} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\varepsilon \qquad (۱.۴)$$

میانگین و ماتریس کواریانس $\mathbf{z}$ نیز به صورت زیر است.

$$E(\mathbf{z}) = \boldsymbol{\mu} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}$$

و

$$\mathrm{Cov}(\mathbf{z}) = \Sigma = [(\mathbf{I} - \rho\mathbf{W})^\top(\mathbf{I} - \rho\mathbf{W})]^{-1}\sigma^2$$

اگر ρ معلوم باشد، بهترین برآوردگر خطی ناریب بردار میانگین $\boldsymbol{\mu}$ برابر است با:

$$\hat{\boldsymbol{\mu}} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}$$

در این بخش چگونگی بدست آوردن پیش‌گویی درون نمونه‌ای توضیح داده می‌شود. برای آشنایی با پیش‌گویی برون نمونه‌ای در مدل تأخیر فضایی می‌توان به گولارد و همکاران (۲۰۱۷) مراجعه کرد.

در مدل (۱.۴) وقتی که $\rho = 0$ پیش‌گوی ناریب خطی[۷] برای $\mathbf{z_s}$ با بهترین برآوردگر خطی ناریب $\boldsymbol{\mu}$ برابر است و به صورت زیر است.

$$\hat{\mathbf{z}_\mathbf{s}}^T = \mathbf{X_s}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\mu}} \qquad (۲.۴)$$

[۷] Linear unbiased predictor

در اینجا $\hat{\boldsymbol{\beta}}$ برآوردگر $\boldsymbol{\beta}$ است که با استفاده از برازش کمترین توان‌های دوم خطا در مدل رگرسیون خطی چندگانه به دست می‌آید.
در مدل تأخیر فضایی گولارد و همکاران (۲۰۱۷) این پیش‌گو را پیش‌گوی روند[۸] نامیده‌اند.
وقتی که $\rho \neq 0$ باشد پیش‌گویی به صورت زیر است:

$$\hat{\mathbf{z}_s}^{TC} = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X_s}\hat{\boldsymbol{\beta}} \tag{۳.۴}$$

که به این پیش‌گویی فضایی، پیش‌گوی تصحیح شده‌ی روند[۹] گفته می‌شود. $\hat{\rho}$ و $\hat{\boldsymbol{\beta}}$ در معادله (۳.۴) برآوردگرهای ρ و $\boldsymbol{\beta}$ هستند که دربخش سوم به ترتیب با رابطه‌های (۸.۳) و (۹.۳) بدست آوردیم.پکیج spdep در نرم‌افزار R از روش دیگری استفاده می‌کند که بوسیله هاینینگ (۱۹۹۰) بیان شده است. در این روش $\texttt{trend} = \mathbf{X}\hat{\boldsymbol{\beta}}$ و $\texttt{signal} = \hat{\rho}\mathbf{W}\mathbf{z_s}$ است و پیش‌گویی عبارت است از:

$$\hat{\mathbf{z}_s}^{TS} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W}\mathbf{z_s} \tag{۴.۴}$$

گولارد و همکاران (۲۰۱۷) پیش‌گویی که از رابطه (۴.۴) بدست می‌آید، پیش‌گوی روند ـ سیگنال ـ اغتشاش[۱۰] نام‌گذاری کرده‌اند.
اگر $\rho = 0$ باشد هر سه روش پیش‌گویی با هم برابرند و اگر $\rho = \hat{\rho}$ و $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ باشد می‌توان نوشت:

$$E(\hat{\mathbf{z}_s}^{TC}) = E(\hat{\mathbf{z}_s}^{TS}) = E(\mathbf{z_s}) \tag{۵.۴}$$

۵ مثال کاربردی

در این مثال از مجموعه داده‌ای به نام columbus استفاده کردیم که در داخل پکیج spdep نرم‌افزار R موجود است. که این مجموعه داده از ۴۹ سطر و ۲۲ ستون تشکیل شده است. هر سطر مربوط به یک منطقه و هر ستون مربوط به یک متغیر در این مناطق می‌باشد که مهمترین متغیرهای این مجموعه داده عبارتند از:

HOVAL(hosing value) : قیمت مسکن برحسب ۱۰۰۰ دلار

INC(household income) : میانگین درآمد ماهیانه خانوار

CRIME[۱۱] : سرقت‌های مسکونی و دزدی خودرو در هر هزار خانوار در این مثال CRIME به عنوان متغیر وابسته (z) و از HOVAL و INC به عنوان متغیرهای مستقل استفاده می‌کنیم.

شکل ۱ موقعیت مناطق در شهر کلمبوس و همچنین همسایگی‌های این مناطق را که با شماره‌ها مشخص شده‌اند را نشان می‌دهد.
هر یک از مناطق یک ناحیه‌ی فضایی و مناطق دارای مرز مشترک به عنوان همسایه در نظر گرفته شده‌اند. که

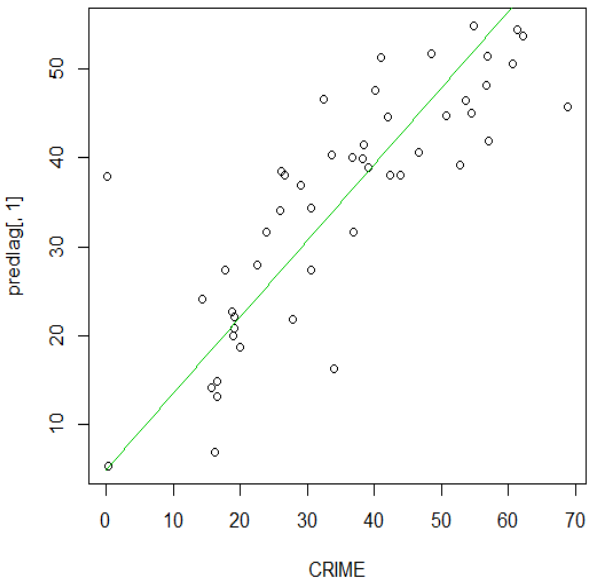
$$w_{ij} = \begin{cases} ۱ & \text{اگر نواحی } i \text{ و } j \text{ همسایه باشند} \\ ۰ & \text{اگر } i = j \\ ۰ & \text{در غیر این‌صورت} \end{cases}$$

[۸] Trend predictor
[۹] Trend corrected predictor
[۱۰] Trend signal noise predictor
[۱۱] Residential burglaries and vehicle thefts per thousand households in the neighborhood

شکل ۱: شماره مناطق در شهر کلمبوس

با تقسیم هر سطر ماتریس وزن بر مجموع درایه‌های آن سطر، ماتریس وزن استاندارد تعیین می‌شود. مقدار برآورد پارامترهای مدل رگرسیو ـ اتورگرسیو فضایی در جدول ۱ آمده است.

نمودار مقادیر واقعی در مقابل مقادیر پیش‌گویی شده با یک خط ممتد ۴۵ درجه برای مدل رگرسیو ـ اتورگرسیو در شکل ۲ نشان داده شده است. هر چه نقاط به این خط نزدیکتر باشند نشان دهنده این است که پیش‌گویی دقیق‌تر است. بنابراین به غیر از چند داده، بقیه داده‌ها پیش‌گویی مناسبی داشته‌اند. ایتام (۱۳۹۶) علاوه بر مدل رگرسیو ـ اتورگرسیو، دو مدل اتورگرسیو همزمان و خطای فضایی را هم به داده‌های مشهور columbus برازش داد و مقادیر AIC این سه مدل در جدول ۲ آمده است. بنابراین برای این داده‌ها مدل رگرسیو ـ اتورگرسیو بهتر برازش می‌شود. در عین حال برای اینکه ببینیم در حالت کلی چه مدلی مناسب‌تر است به مطالعات شبیه‌سازی بیشتری نیاز است .



شکل ۲: نمودار پیش‌گویی مدل اتورگرسیو تأخیر فضایی در برابر داده‌ها

جدول ۱: برآورد پارامترها برای مدل تأخیر فضایی

	برآورد
β_0(Intercept)	۴۵/۰۷۹۲۵۱
β_1(INC)	−۱/۰۳۱۶۱۶
β_2(HOVAL)	−۰/۲۶۵۹۲۶
ρ(Rho)	۰/۴۳۱۰۲
σ^2	۹۵/۴۹۴

جدول ۲: مقایسه سه مدل همزمان فضایی، رگرسیو- اتورگرسیو فضایی و خطای فضایی

	df	AIC
lag	۵	۳۷۴/۸۸
error	۵	۳۷۶/۷۶
simultaneous	۳	۳۹۸/۴۰

بحث و نتیجه‌گیری

مدل رگرسیو ـ اتورگرسیو فضایی یکی از مدل‌هایی است که برای داده‌های مشبکه‌ای نامنظم (و حتی منظم) قابل کاربرد است، بدیهی است چنانچه متغیر کمکی فضایی نداشته باشیم می‌توان عبارت $\mathbf{X}\boldsymbol{\beta}$ را نادیده گرفت (یا به جای آن صرفاً جمله β. مربوط به عرض از مبدأ یک مدل رگرسیونی را قرارداد) که در اینصورت مدل رگرسیو ـ اتورگرسیو به مدل اتورگرسیو همزمان فضایی تبدیل می‌شود. مدل رگرسیو ـ اتورگرسیو برای داده‌های فضایی نامنظم هستند، برای داده‌های فضایی نامنظم هم قابل کاربرد هستند. از طرفی برای داده‌های فضایی منظم هم روش کریگیدن هم مدل رگرسیو ـ اتورگرسیو را می‌توان بکار برد لذا با استفاده از مطالعات شبیه‌سازی می‌توان مقایسه دقت این مدل‌ها را بررسی کرد.

مراجع

ایتام ح، (۱۳۹۶). برآورد و پیش‌گویی با استفاده از مدل‌های اتورگرسیو فضایی، پایان‌نامه کارشناسی ارشد، دانشگاه بیرجند.

رسولی ح، (۱۳۹۰). مدل‌های اتورگرسیو فضایی و تحلیل داده‌های معاملات مسکونی شهر تهران، *علوم آماری*، **۲** ، ۲۰۲-۱۸۹.

محمدزاده م، (۱۳۹۴). *آمار فضایی و کاربردهای آن*، چاپ اول، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Klumer Academic Publishers. The Netherlands.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Iattice Systems, *Journal of Royal Statistical Society B*, **36**, 192-236.

Cressie, N. (1993). *Statistics for Spatial Data*, Johan Wilely, New York.

Goulard, M., Laurent, T., and Thomas-Agnan, C. (2017). About Predictions in Spatial Autoregressive Models: Optimal and Almost Optimal strategies, *Spatial Economic Analysis* , **17**(2-3), 304–325.

Haining, R. P. (1990), *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge, Cambridge University Press.

Lu, J. and Zhang, L. (2011). Modeling and Prediction of Tree Height-Diameter Relationships Using Spatial Autoregressive Models, *Forest Science*, **57**, 252-264.

Ord, J. K. (1975). Estimation Methods for Models of Spatial Interaction, *Journal of the American Statistical Association*, **70**, 120-126.

Pace, R. K. and Barry, R. (1997). Sparse Spatial Autoregressions, *Statistics and Probability letters*, **33**, 291-297.

Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*, John Wiley, Hoboken, N. J.

Whittile, P. (1954). On Stationary Processes in the Plane, *Biometrika*, **41**, 434-449.

دومین سمینار آمار فضایی و کاربردهای آن
۳ - ۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1390

مدل‌سازی و پیش‌بینی نرخ مرگ‌ومیر با استفاده از آمار فضایی

نصراله ایران پناه*[۱] ، هدی توسلی[۲]

[۱]گروه آمار، دانشگاه اصفهان

[۲]گروه ریاضی، دانشگاه آزاد اسلامی واحد نجف‌آباد

چکیده: بررسی روند و احتمال مرگ و میر برای برنامه‌ریزان، کارشناسان جمعیت، سازمان‌های بازنشستگی و شرکت‌های بیمه از اهمیت بالایی برخوردار است. تحقیقات نشان می‌دهد جداول مرگ‌ومیر ایستا، احتمالات مرگ‌ومیر را زیاد نشان می‌دهند. علت بیش برآوردی احتمالات مرگ‌ومیر در جداول ایستا را می‌توان به نادیده گرفتن تغییرات مرگ‌ومیر بر حسب زمان نسبت داد. جداول مرگ‌ومیر پویا با در نظر گرفتن اثر زمان در سنین مختلف بر روند مرگ‌ومیر از دقت بالاتری برخوردار هستند. بعد از مدل‌سازی داده‌ها بر حسب زمان و گروه های سنی، ساختار باقیمانده‌های وابسته را به عنوان تحققی از یک میدان تصادفی گاوسی همگن در نظر می‌گیریم. در این مقاله، برای برآورد روند مرگ‌ومیر ابتدا با استفاده از روش‌های لی ـ کارتر و پالایش ـ میانه مدلی بر حسب زمان و رده‌های سنی به داده‌های نرخ مرگ‌ومیر برازش و پارامترهای مدل را برآورد می‌کنیم. سپس پارامترهای مدل لی‌ـ کارتر به روش‌های تجزیه ویژه مقدار و مدل‌های خطی تعمیم یافته برآورد می‌گردند. در ادامه ساختار برآورد شده برای پیش‌بینی نرخ مرگ‌ومیر با استفاده از معیارهای نیکویی برازش مورد مقایسه قرار می‌گیرند. در نهایت روش‌های ارائه شده برای مدل‌سازی و پیش‌بینی زمان‌های آینده داده‌های مرگ‌ومیر استان اصفهان مورد استفاده قرار می‌گیرند.

واژه‌های کلیدی: جداول مرگ‌ومیر پویا، مدل لی ـ کارتر، تجزیه ویژه مقدار، آمار فضایی، پالایش ـ میانه.

کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲H۱۱، ۶۰G۲۵ .

۱ مقدمه

پیش‌بینی احتمال مرگ‌ومیر امروزه به طور قابل ملاحظه‌ای مورد توجه قرار گرفته است. این پیش‌بینی‌ها با توجه به افزایش امید به زندگی در سازمان‌های بیمه و تعیین حقوق بازنشستگی کاربرد دارند. تحلیل داده‌های مرگ‌ومیر با استفاده از روش‌های آمار پارامتری

*نام و آدرس پست الکترونیکی ارایه‌دهنده مقاله: نصراله ایران پناه، iranpanah@sci.ui.ac.ir

و ناپارامتری براساس جداول عمر ایستا انجام می‌گیرد، ولی این روش‌ها تغییرات مرگ‌ومیر برحسب زمان را در نظر نمی‌گیرند که این باعث بیش برآوردی پیش‌بینی‌ها و همچنین خطا در تحلیل‌ها می‌شود. برای این منظور از جداول عمر پویا در تحلیل‌ها استفاده می‌گردد که در این جداول اثر زمان بر روی مرگ‌ومیر مورد مطالعه قرار می‌گیرد.

بررسی روند مرگ‌ومیر با استفاده از مدل‌های پویا یک موضوع بسیار مهم است و توسط بسیاری از آماردانان مورد توجه قرار گرفته است. این مدل‌ها اولین بار توسط لی و کارتر (۱۹۹۲) مورد بررسی قرار گرفتند. پس از آن محققین بسیاری از جمله بروهانس و همکاران (۲۰۰۲)، پیتاکو (۲۰۰۴) و رنشاو وهابرمن (۲۰۰۶) این مدل‌ها را توسعه دادند. همچنین دبن و همکاران (۲۰۰۸a) نیز با در نظر گرفتن این مدل، نرخ مرگ‌ومیر کشور اسپانیا را مدل‌بندی و سپس این نرخ را برای سال‌های بعد پیش‌بینی نمودند. اما هیچ یک از این محققین ساختار وابسته میان داده‌ها را در نظر نگرفتند. سرانجام دبن و همکاران (۲۰۰۸b) بادر نظر گرفتن این مدل و با استفاده از آمار فضایی مدلی به این جداول اختصاص دادند. آن‌ها جداول عمر پویا ر ا مانند یك جدول دو راهه روی یک فضای همگن مشبکه‌ای در نظر می‌گیرند که در هر یک از محورهای افقی و عمودی به ترتیب سن و سال قرار دارند. داده‌ها به تغییرات مقیاس بزرگ (روند) و تغییرات مقیاس کوچک (خطا) به صورت $Z(x,t)=\mu(x,t)+\delta(x,t)$ تجزیه می‌گردند که در آن x و t به ترتیب سن و سال می‌باشد. دردر این مدل مولفه ناایستایی توسط تابع روند $\mu(x,t)$ و باقیمانده‌های $\delta(x,t)$ با میانگین صفر و ایستایی ضعیف در نظر گرفته می‌شوند. در این مقاله متغیر مورد نظر، احتمال‌های مرگ‌ومیر در گروه‌های سنی و فصل‌های مختلف به صورت دو مدل لی ـ کارتر (LC) و پالایش ـ میانه (MP) در نظر گرفته می‌شود. همچنین پارامترهای مدل LC به دو روش تجزیه مقدار ویژه (SVD) و مدل خطی تعمیم یافته (GLM) برآورد می‌گردند. روش‌های ارائه شده برای تحلیل داده‌های مرگ‌ومیر استان اصفهان از فصل بهار ۱۳۸۳ تا پاییز ۱۳۸۸ به طور فصلی در گروه‌های سنی ۵ ساله برای پیش‌بینی احتمال مرگ‌ومیر مورد استفاده قرار گرفته است.

برای این منظور در بخش ۲ مدل LC به همراه دو روش SVD و GLM ارائه می‌گردد. روش پالایش ـ میانه در بخش ۳ معرفی می‌گردد. در بخش ۴ برای تحلیل داده‌های مرگ‌ومیر استان اصفهان، ابتدا احتمال‌های مرگ‌ومیر برآورد و سپس پیش‌بینی برای زمان‌های آینده انجام می‌شود. در نهایت به بحث و نتیجه‌گیری پرداخته شده است.

۲ مدل لی ـ کارتر

لی و کارتر (۱۹۹۲) مدل LC را برای تعیین نرخ مرگ‌ومیر ارائه دادند. این مذل به صورت

$$m_{xt}=\exp\{a_x+b_xk_t+\delta_{xt}\},\quad x=۱,\ldots,n,\quad t=۱,\ldots,T$$

یا به طور معادل به صورت

$$\ln(m_{xt})=a_x+b_xk_t+\delta_{xt}$$

است، که درآن پارامترهای a_x و b_x وابسته به سن هستند و k_t شاخص مرگ‌ومیر برای هر سال یا واحد زمان می‌باشد. همچنین δ_{xt} باقیمانده‌های مدل است که فرض می‌شود تحققی از یک میدان تصادفی گوسی با میانگین صفر و واریانس $\sigma_\delta^۲$ است. برای مدل‌بندی احتمال مرگ‌ومیر q_{xt} از یک نسخه تعمیم یافته (دبن و همکاران (۲۰۰۸b)) از الگوی لوجیت q_{xt} به صورت

$$\text{logit}(q_{xt})=\ln\left(\frac{q_{xt}}{۱-q_{xt}}\right)=a_x+b_xk_t+\delta_{xt}$$

استفاده می‌گردد. قیود پیشنهادی برای یکتایی جواب‌ها به صورت $\sum_x b_x = ۱$ و $\sum_t k_t = ۰$ می‌باشند.

۱.۲ برآورد پارامترهای مدل LC به روش SVD

مراحل مختلف برآورد پارامترهای مدل LC به روش SVD به صورت زیر است:

(۱) ابتدا x از رابطه

$$\check{a}_x = \frac{۱}{T}\sum_{t=۱}^{T} \ln\left(\frac{q_{xt}}{۱ - q_{xt}}\right)$$

برآورد می‌گردد، که در آن T تعداد کل زمان‌هاست.

(۲) مقادیر b_x و k_t از عبارت اول SVD به صورت

$$\ln\left(\frac{q_{xt}}{۱ - q_{xt}}\right) - \check{a}_x = \sum_{i=۱}^{\min(n,T)} S^i U_x^i V_t^i$$

محاسبه می‌شوند، که در این معادله S^i، U_x^i و V_t^i به ترتیب مقادیر منفرد مرتب شده و بردارهای منفرد چپ و راست هستند (رنشاو وهابرمن (۲۰۰۶)). بنابراین $\hat{k}_t^{svd} = S^۱V_t^۱$ و $\hat{b}_x^{svd} = U_x^۱$. برای یکتایی جواب‌ها قیود $\sum_x b_x = ۱$ و $\sum_t k_t = ۰$ پیشنهاد می‌گردد، ولی چون $\sum_x \hat{b}_x^{svd} = c \neq ۱$، تبدیلات $\hat{b}_x = \frac{۱}{c}\hat{b}_x^{svd}$ و $\check{k}_t = c\hat{k}_t^{svd}$ در نظر گرفته می‌شود.

(۳) با برآورد پارامترها در مراحل ۱ و ۲ اختلافی بین احتمال مرگ‌ومیر واقعی و برآورد شده مشاهده می‌گردد که علت آن در نظر گرفتن الگوی لوجیت احتمال مرگ‌ومیر است. بنابراین با استفاده از روش‌های عددی و مقدار اولیه $\check{k}_t$، مقدار $\tilde{k}_t$ از رابطه

$$D_t = \sum_x E_{xt}\frac{\exp\{\check{a}_x + \hat{b}_x k_t\}}{۱ + \exp\{\check{a}_x + \hat{b}_x k_t\}}; \quad t = ۱, \ldots, T$$

برآورد می‌گردد، که در آن E_{xt} تعداد افراد در معرض مرگ‌ومیر در گروه سنی x و در زمان t و $D_t = \sum_x D_{xt}$ تعداد کل مرگ‌ها در زمان t است.

(۴) پارامترهای $\hat{a}_x$ و $\hat{k}_x t$ ، با استفاده از تبدیلات $\hat{a}_x = \tilde{a}_x + \hat{b}_x \text{mean}(\tilde{k}_t)$ و $\hat{k}_t = \tilde{k}_t - \text{mean}(\tilde{k}_t)$ برآورد می‌گردد.

(۵) در نهایت احتمال مرگ‌ومیر با عکس لوجیت به صورت

$$\hat{q}_{xt} = \text{antilogit}(\hat{a}_x + \hat{b}_x\hat{k}_t)$$

برآورد می‌گردد.

۲.۲ برآورد پارامترهای مدلLC به روش GLM

(۱) به ازای هر x، $\check{b}_x^{GLM}$ از رابطه $\ln\left(\frac{q_{xt}}{۱-q_{xt}}\right) = \text{offset}(\hat{a}_x) + b_x\hat{k}_t$ محاسبه می‌گردد، که در آن $\hat{a}_x$ و $\hat{k}_t$ از روش قبل برآورد شده‌اند.

(۲) به ازای هر t، $\check{k}_t^{GLM}$ از رابطه $\ln(\frac{q_{xt}}{1-q_{xt}}) = \text{offset}(\hat{a}_x) + \check{b}_x^{GLM} k_t$ محاسبه می‌گردد.

(۳) چون $\sum_x \check{b}_x^{GLM} = s \neq 1$ و $\sum_t \check{k}_t^{GLM} \neq 0$ می‌باشد، تبدیلات زیر انجام می‌گیرد:

$$\begin{aligned}
\hat{b}_x^{GLM} &= \frac{1}{s}\check{b}_x^{GLM} \\
\hat{k}_t^{GLM} &= s\check{k}_t^{GLM} - \text{mean}(s\check{k}_t^{GLM}) \\
\hat{a}_x^{GLM} &= \hat{a}_x + \hat{b}_x^{GLM}\text{mean}(s\check{k}_t^{GLM})
\end{aligned}$$

(۴) در نهایت احتمال مرگ‌ومیر با عکس لوجیت به صورت

$$\hat{q}_{xt} = \text{antilogit}(\hat{a}_x^{GLM} + \hat{b}_x^{GLM}\hat{k}_t^{GLM})$$

برآورد می‌گردد.

۳ روش پالایش ـ میانه

یك جدول عمر پویا می‌تواند به صورت یك جدول دو راهه، در دو جهت افقی (سن) و عمودی (زمان) در نظر گرفته شود. برای این منظور احتمال مرگ‌ومیر به سه مؤلفه

$$q_{xt} = \mu + r_x + c_t$$

تجزیه می‌شود که در آن μ اثر کل، r_x اثر سطر xام و c_t اثر ستون tام است. الگوریتم MP (کرسی (۱۹۹۳)) برای برآورد اثرهای کل، سطر و ستون به کار برده می‌شود. روش MP یک روش ناپارامتری است و نسبت به روش‌های دیگر، نسبت به داده‌های پرت حساسیت کمتری دارد. همچنین این روش اریبی کمتری نسبت به روش‌هایی که بر اساس عملگرهای میانگین هستند را دارد. برآورد پارامترها در این روش به صورت زیر می‌باشد:

برای $i = 1, 3, 5, \ldots$ تعریف می‌شود:

$$\begin{aligned}
q_{xt}^{(i)} &\equiv q_{xt}^{(i-1)} - \text{med}\{q_{xt}^{(i-1)} : t = 1, ..., T\}, \quad x = 1, ..., n+1, \quad t = 1, ..., T, \\
q_{x,T+1}^{(i)} &\equiv q_{x,T+1}^{(i-1)} + \text{med}\{q_{xt}^{(i)} : t = 1, ..., T\}, \quad x = 1, ..., n+1.
\end{aligned}$$

و برای $i = 2, 4, \ldots$ تعریف می‌شود:

$$\begin{aligned}
q_{xt}^{(i)} &\equiv q_{xt}^{(i-1)} - \text{med}\{q_{xt}^{(i-1)} : x = 1, ..., n\}, \quad x = 1, ..., n, \quad t = 1, ..., T+1, \\
q_{n+1,t}^{(i)} &\equiv q_{n+1,t}^{(i-1)} + \text{med}\{q_{xt}^{(i-1)} : x = 1, ..., n\}, \quad t = 1, ..., T+1.
\end{aligned}$$

برای شروع الگوریتم فرض می‌شود

$$q_{xt}^{(0)} = \begin{cases} q_{xt} & x = 1, \ldots, n, \quad t = 1, \ldots, T, \\ 0 & \text{o.w.} \end{cases}$$

به عبارت دیگر الگوریتم $n \times T$ داده در داخل جدول و ایجاد $n+T+۱$ مقدار صفر در خارج از جدول شروع می‌شود. با استفاده از مراحل فرد الگوریتم، میانه‌های سطرها از داده‌های هر خانه کم شده و سپس مقدار کم شده (میانه هر سطر) در خانه‌های خارج از سطرها (n تا) قرار داده می‌شود. میانه ستون‌ها نه تنها از داده‌های هر خانه کم، بلکه از ستون جمع شده سطرهای کم شده ($T+۱$ امین ستون) نیز کم می‌شود. آخرین مقدار کم شده در $(n+۱, T+۱)$ امین خانه خارج از جدول قرار می‌گیرد. الگوریتم تا رسیدن به یک همگرایی تکرار می‌شود. با فرض رسیدن به یک همگرایی، اثرات کل، سطر و ستون به ترتیب به صورت

$$\begin{aligned}
\hat{\mu} &\equiv q^{(\infty)}_{p+۱,T+۱}, \\
\hat{r}_x &\equiv q^{(\infty)}_{x,T+۱}, \quad x = ۱, ..., n, \\
\hat{c}_t &\equiv q^{(\infty)}_{n+۱,t}, \quad t = ۱, ..., T, \\
q_{xt} &= \hat{\mu} + \hat{r}_x + \hat{c}_t + q^{(\infty)}_{xt}, \quad x = ۱, ..., n, \ t = ۱, ..., T,
\end{aligned}$$

برآورد می‌شوند. در نهایت احتمال‌های مرگ‌ومیر و باقیمانده‌های مدل به ترتیب به صورت

$$\begin{aligned}
\hat{q}_{xt} &= \hat{\mu} + \hat{r}_x + \hat{c}_t; \quad x = ۱, ..., n, \ t = ۱, ..., T, \\
\hat{\delta}_{xt} &= q_{xt} - \hat{q}_{xt} = q^{(\infty)}_{xt}
\end{aligned}$$

برآورد می‌گردند، که در آن مقادیر باقی‌مانده‌ها در خانه‌های جدول $n \times T$ یعنی $q^{(\infty)}_{xt}$ در واقع همان برآورد خطاهای (باقیمانده‌های) بدون روند δ_{xt} است.

۴ تحلیل داده‌های مرگ ومیر استان اصفهان

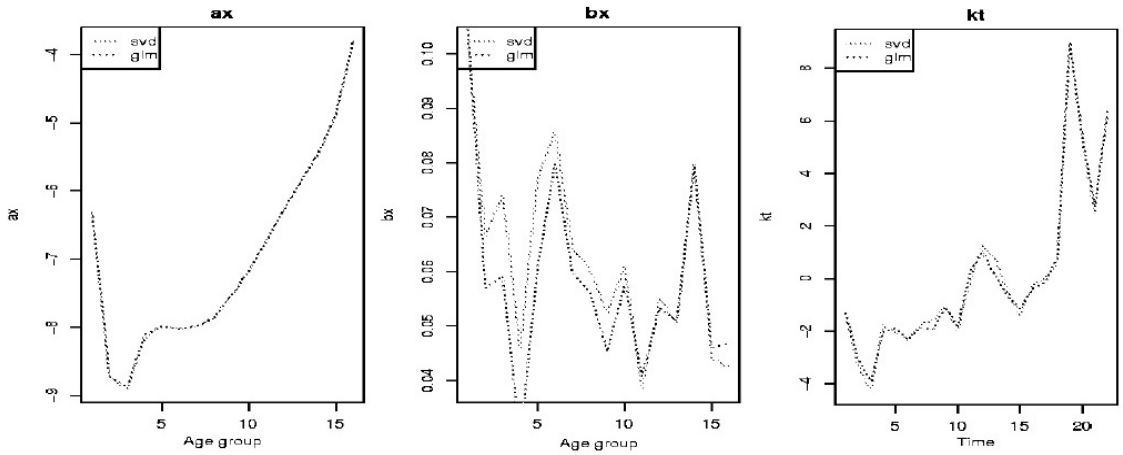
۱.۴ برآورد احتمال‌های مرگ‌ومیر

در این بخش مدل‌های ارائه شده در بخش‌های ۲ و ۳ را به داده‌های مرگ‌ومیر استان اصفهان متناظر با دوره زمانی (فصلی) ۱۳۸۸-۱۳۸۳و گروه‌های سنی ۵ ساله (۴-۰)، (۹-۵)، ... و (۷۵ و بیشتر) برازش داده می‌شود. در این مدل‌ها احتمال‌های مرگ‌ومیر q_{xt} به صورت

$$q_{xt} = \frac{\frac{۱}{۲}(d_{xt} + d_{x(t+۱)})}{P_{xt} + \frac{۱}{۲}d_{xt}} \qquad (۱.۴)$$

محاسبه می‌گردد، که در آن d_{xt} تعداد مرگ‌ومیر گروه سنی ۵ ساله x در زمان $t+۱$ و P_{xt} جمعیت گروه سنی ۵ ساله x در روز ۱۵ ماه میانی فصل t می‌باشد.

برآورد پارامترهای مدل LC $(\hat{a}_x, \hat{b}_x, \hat{k}_t)$ به دو روش SVD و GLM (بخش ۲) در شکل ۱ نشان داده شده اند. در این شکل برآورد $\hat{a}_x$ یک افزایش مرگ‌ومیر را از گروه سنی سوم یعنی ۱۵-۱۰ به بعد نشان می‌دهد. همچنین شاخص زمانی $\hat{k}_t$ یک روند صعودی را در طول زمان از خود نشان می‌دهد که با توجه به مقادیر مثبت $\hat{b}_x$، این روند بیانگر افزایش نرخ مرگ‌ومیر در طول زمان است.



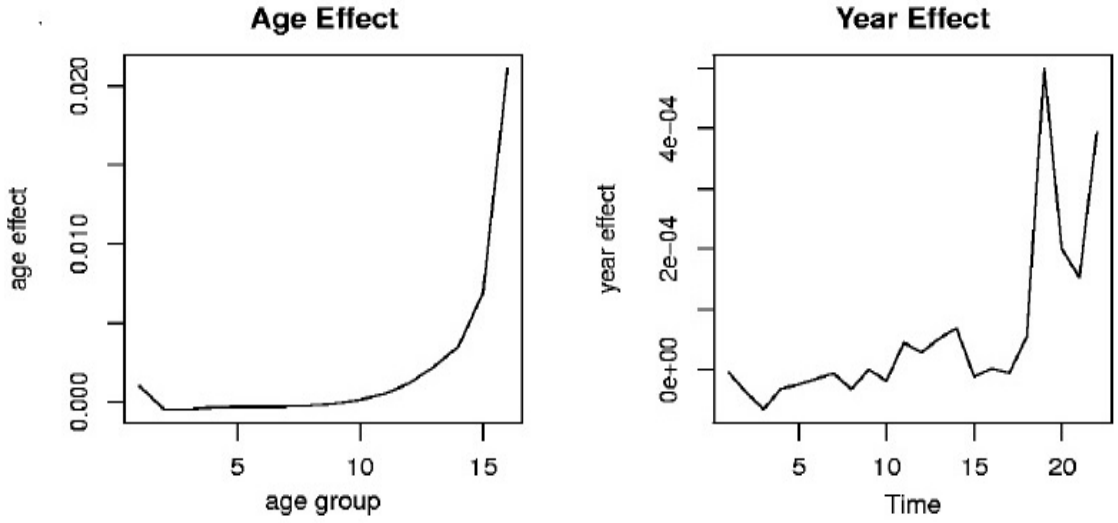
شکل ۱: برآورد پارامترهای مدل LC به دو روش SVD و GLM .

به علاوه در روش MP برآورد اثرات سطر $\hat{r}_x$ و ستون $\hat{c}_t$ یعنی گروه‌های سنی ۵ ساله و زمان به صورت فصلی در شکل ۲ نشان داده شده است. همچنین اثر کل $\hat{\mu} = ۰/۰۰۱۷۲۶$ برآورد می‌گردد. در تحلیل داده‌های مرگ‌ومیر و استفاده از مدل‌های پویا، برای مقایسه نیکویی برازش مدل‌های برازش شده LC وMP از معیارهای MSE و MAPE به صورت

$$\text{MSE} = \left\{ \frac{۱}{n} \sum_x (q_{xt} - \hat{q}_{xt})^۲ \right\}^{\frac{۱}{۲}}; \quad t = ۱, ۲, ..., ۲۴$$

$$\text{MAPE} = \frac{۱}{n} \sum_x |q_{xt} - \hat{q}_{xt}|/q_{xt}; \quad t = ۱, ۲, ..., ۲۴$$

استفاده می‌گردد، که در آن مقدار واقعی احتمال‌های مرگ‌ومیر (۱.۴) و $\hat{q}_{xt}$ برآورد احتمال‌های مرگ‌ومیر به روش‌های LC (برآورد پارامترها به روش‌های SVD و GLM در بخش ۲) و MP در بخش ۳ هستند.



شکل ۲: برآورد اثرات گروه‌های سنی و زمان به روش MP

معیارهای MSE و MAPE برای هر فصل از بهار ۱۳۸۳ تا تابستان ۱۳۸۸ با استفاده از روش‌های برآورد SVD ، GLM و MP در جدول ۱ ارائه شده است. در این جدول مدل LC در مقایسه با مدل MP به عنوان مدل بهتر در برازش احتمال‌های مرگ‌ومیر ارائه می‌گردد. زیرا این مدل مقادیر MSE و MAPE کمتری در مقایسه با مدل MP دارد. همچنین در برآورد پارامترهای مدل LC به دو روش SVD و GLM، روش SVD از دقت تقریبا بالاتری برخوردار است.

جدول ۱: معیارهای MSE و MAPE برای هر فصل با استفاده از روش‌های برآورد SVD ، GLM و MP

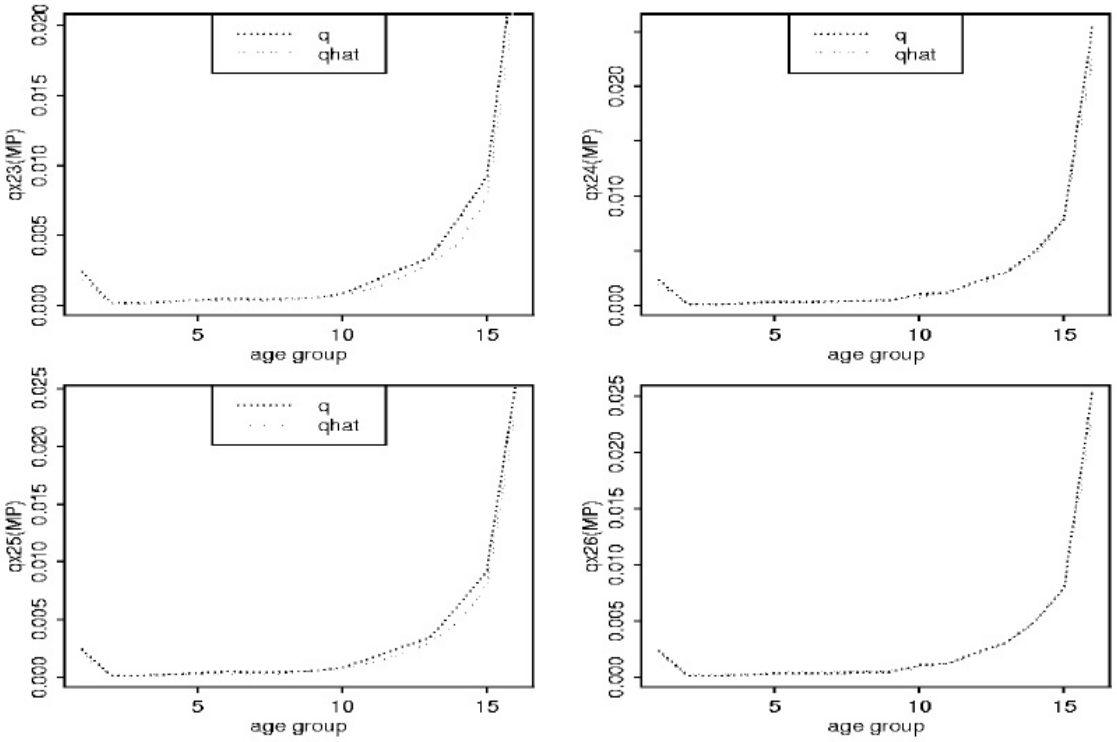
		LC (SVD)		LC (GLM)		MP	
زمان(t)	فصل ـ سال	MAPE	MSE	MAPE	MSE	MAPE	MSE
۱	بهار ۱۳۸۳	۰,۱۲۱۶۵	۰,۰۰۰۵۵	۰,۱۱۲۹۷	۰,۰۰۰۵۴	۰,۱۴۳۹۳	۰,۰۰۰۸۵
۲	تابستان ۱۳۸۳	۰,۰۹۲۰۹	۰,۰۰۰۴۰	۰,۰۷۶۹۸	۰,۰۰۰۳۹	۰,۱۵۶۵۹	۰,۰۰۱۱۲
۳	پاییز ۱۳۸۳	۰,۰۷۸۴۷	۰,۰۰۰۳۸	۰,۰۷۰۱۳	۰,۰۰۰۳۹	۰,۰۱۶۱۱۴	۰,۰۰۰۶۰
۴	زمستان ۱۳۸۳	۰,۰۶۰۵۵	۰,۰۰۰۱۱	۰,۷۳۹۴	۰,۰۰۰۲۲	۰,۰۶۶۲۳	۰,۰۰۰۲۷
۵	بهار ۱۳۸۴	۰,۳۸۷۴	۰,۰۰۰۱۱	۰,۴۴۴۳	۰,۰۰۰۱۰	۰,۰۹۱۴۱	۰,۰۰۰۵۳
۶	تابستان ۱۳۸۴	۰,۰۸۴۵۴۳	۰,۰۰۰۱۳	۰,۰۷۸۸۲	۰,۰۰۰۱۰	۰,۱۰۴۲۹	۰,۰۰۰۵۹
۷	پاییز ۱۳۸۴	۰,۰۷۷۳۳	۰,۰۰۰۱۶	۰,۰۹۱۴۷	۰,۰۰۰۲۳	۰,۰۸۱۶۵	۰,۰۰۰۲۶
۸	زمستان ۱۳۸۴	۰,۰۷۳۲۴	۰,۰۰۰۲۴	۰,۰۸۲۷۲	۰,۰۰۰۳۸	۰,۰۹۱۱۰	۰,۰۰۰۱۲
۹	بهار ۱۳۸۵	۰,۰۴۹۸۸	۰,۰۰۰۲۹	۰,۰۵۳۶۰	۰,۰۰۰۲۶	۰,۰۵۸۴۸	۰,۰۰۰۴۹
۱۰	تابستان ۱۳۸۵	۰,۰۷۶۱۱	۰,۰۰۰۱۷	۰,۰۷۰۷۴	۰,۰۰۰۱۶	۰,۰۸۳۱۴	۰,۰۰۰۳۶
۱۱	پاییز ۱۳۸۵	۰,۰۶۰۵۴	۰,۰۰۰۴۵	۰,۰۶۶۲۶	۰,۰۰۰۳۸	۰,۰۹۴۵۶	۰,۰۰۰۴۹
۱۲	زمستان ۱۳۸۵	۰,۱۲۷۰۴	۰,۰۰۰۲۹	۰,۱۱۴۸۶	۰,۰۰۰۳۲	۰,۰۹۷۴۳	۰,۰۰۰۵۹
۱۳	بهار ۱۳۸۶	۰,۱۳۵۹۵	۰,۰۰۰۳۸	۰,۱۱۹۵۷	۰,۰۰۰۲۱	۰,۱۰۹۱۰	۰,۰۰۰۱۴
۱۴	تابستان ۱۳۸۶	۰,۴۸۹۱	۰,۰۰۰۲۵	۰,۰۴۶۲۰	۰,۰۰۰۱۹	۰,۱۰۰۶۹	۰,۰۰۰۳۴
۱۵	پاییز ۱۳۸۶	۰,۸۸۴۸	۰,۰۰۰۳۹	۰,۰۹۶۱۱	۰,۰۰۰۳۷	۰,۰۶۵۵۵	۰,۰۰۰۱۵
۱۶	زمستان ۱۳۸۶	۳۰,۱۳۱۹۸	۰,۰۰۰۶۲	۵۰,۱۴۰۶۴	۰,۰۰۰۶۶	۰,۰۸۶۰۲	۰,۰۰۰۶۰
۱۷	بهار ۱۳۸۷	۰,۰۶۰۸۴	۰,۰۰۰۴۸	۰,۰۶۰۷۲	۰,۰۰۰۵۳۶	۰,۰۷۴۸۵	۰,۰۰۰۵۴
۱۸	تابستان ۱۳۸۷	۰,۲۴۰۷۷	۰,۰۰۱۶۳	۰,۲۳۰۹۰	۰,۰۰۱۵۹	۰,۲۳۵۸۰	۰,۰۰۱۸۵
۱۹	پاییز ۱۳۸۷	۰,۲۰۵۳۹	۰,۰۰۰۴۸	۰,۱۵۷۵۹	۰,۰۰۰۵۷	۰,۶۳۵۰۱	۰,۰۰۲۵۹
۲۰	زمستان ۱۳۸۷	۰,۱۷۰۶۳	۰,۰۰۰۳۰	۰,۱۵۷۳۲۵	۰,۰۰۰۴۵	۰,۳۲۳۴۴	۰,۰۰۱۳۵
۲۱	بهار ۱۳۸۸	۰,۱۵۰۰۱	۰,۰۰۰۵۱	۰,۱۳۸۰۸	۰,۰۰۰۴۰	۰,۲۱۱۰۲	۰,۰۰۱۱۷
۲۲	تابستان ۱۳۸۸	۰,۰۷۷۳۱	۰,۰۰۰۵۲	۰,۰۶۵۷۹	۰,۰۰۰۶۳	۰,۴۲۵۹۱	۰,۰۰۱۳۶

۱.۱.۴ پیش‌بینی احتمال‌های مرگ‌ومیر

با انتخاب مدل LC به عنوان مدل بهتر، پیش‌بینی‌ها را براساس این مدل و برآورد SVD انجام می‌دهیم. پیش‌بینی احتمال‌های مرگ‌ومیر با مدل LC به ازای هر گروه سنی برای زمان‌های آینده نیاز به یك تحلیل سری زمانی برای شاخص k_t دارد. بدین منظور

ابتدا با استفاده از نرم افزار R بهترین مدلی که به این سری زمانی برازش داده می‌شود ، مدل ARIMA(۰،۱،۱) می‌باشد. سپس با استفاده از مدل فوق پیش‌بینی $\hat{k}_{t+h}$ را برای زمان‌های آینده محاسبه می‌کنیم. در نهایت پیش‌بینی احتمال‌های مرگ‌ومیر در سن x و در زمان $t+h$ به صورت $\hat{q}_{x(t+h)} = antilogit(\hat{a}_x + \hat{b}_x\hat{k}_{(t+h)})$ محاسبه می‌گردد.

در شکل ۳ پیش‌بینی احتمال‌های مرگ‌ومیر $\hat{q}_{xt}$ برای گروه‌های سنی مختلف در ۴ فصل آینده $t = ۲۳, ۲۴, ۲۵, ۲۶$ یعنی از فصل پاییز ۱۳۸۸ تا تابستان ۱۳۸۹ به همراه احتمال‌های واقعی مرگ‌ومیر q_{xt} (۱.۴) ارائه شده است.



شکل ۳: مقدار واقعی q_{xt} و پیش‌بینی $\hat{q}_{xt}$ برای ۴ فصل آینده

بحث و نتیجه‌گیری

در این مقاله ضمن معرفی مدل LC برای برآورد احتمال‌های مرگ‌ومیر، برآورد پارامترهای این مدل به دو روش ارائه گردید. همچنین روش MP برای برآورد اثرات گروه‌های سنی و زمان با توجه به ساختار فضایی داده‌ها معرفی گردید. با استفاده از این مدل‌ها، احتمال‌های مرگ‌ومیر استان اصفهان برای ۱۶ گروه سنی ۵ ساله و فصل‌های بهار ۱۳۸۳ تا تابستان ۱۳۸۸ برآورد و براساس دو ملاك نیکویی برازش MSE و MAPE مدل LC با برآورد SVD برای پارامترها بهترین مدل تشخیص داده شد. با توجه به شکل‌های ۱ و ۲ افزایش نرخ مرگ‌ومیر در گروه‌های سنی بالاتر و همچنین روند رو به رشد احتمال مرگ‌ومیر در طول زمان مشاهده می‌گردد. در شکل ۳ پیش‌بینی احتمال‌های مرگ‌ومیر در گروه‌های سنی مختلف برای ۴ فصل آینده پاییز ۱۳۸۸ تا تابستان ۱۳۸۹ ارائه شده است که در مقایسه با مقادیر واقعی احتمال مرگ‌ومیر دقت بالای پیش‌بینی‌ها مشاهده گردد.

مراجع

Brouhns, N., Denuit, M., and Vermunt, J. (2002), A poisson log-bilinear regression approach to the construction of projected lifetables, *Insurance: Mathematics and Economics*, **31**, 373-393.

Cressie. N. (1993). *Statistics for Spatial Data.* revised ed.Wiley, NewYork

Debon, A., Montes, F. and Puig, F., (2008a). Modeling and forecasting mortality in Spain. *European Journal of Operational Research* **189**, 624-637.

Debon, A., Montes. F., Mateuc, J., Porcuc, E. and Bevilacquad, M., (2008b). Modelling residuals dependence in dynamic life tables: A geostatistical approach. *Computational Statistics and Data Analysis*, **52**, 3128-3147.

Lee, R. and Carter, L., (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, **87**, (419) 659–671.

Pitacco, E. (2004), Survival models in dynamic context: a survey, *Insurance: Mathematics and Economics*, **35**, 279-298.

Renshaw, A., and Haberman, S. (2006), A cohort-based extention of the Lee- Carter model for mortality reduction factors, *Insurance: Mathematics and Economics*, **38**, 556-570.

Shumway, R., and Stoffer, D. (2006), *Time series analysis and its applications with R examples*, Springer, new York.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1990

تحلیل فضایی نرخ وقوع جرم در شهر تهران

حسین باغیشنی*، محمد آرشی

گروه آمار، دانشگاه صنعتی شاهرود

چکیده: پیشگیری و کنترل جرایم شهری، یکی از مهم‌ترین عوامل تامین امنیت اجتماعی و مدیریت قابل قبول در یک شهر است. لازمه برنامه‌ریزی مفید مدیریت شهری برای رسیدن به این هدف، شناخت ماهیت، نحوه تغییرپذیری و عوامل موثر بر وقوع جرایم، به‌ویژه بر حسب نواحی جغرافیایی مختلف یک شهر، می‌باشد. بنابراین، تحلیل فضایی جرایم شهری برای متولیان مربوط به این حوزه می‌تواند بسیار راهگشا باشد. در این مقاله، برای تحلیل نرخ وقوع جرم شرارت در شهر تهران، از یک مدل فضایی رگرسیون بتا استفاده می‌کنیم. ساختار وابستگی فضایی مدل پیشنهادی را با یک تابع مفصل گاوسی وارد مدل می‌کنیم. روش استنباط نیز مبتنی بر رهیافت درستنمایی است.

واژه‌های کلیدی: شرارت، داده‌های مشبکه‌ای، مدل‌های فضایی اتورگرسیو شرطی، مفصل گاوسی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 62H11.

۱ مقدمه

رشد بزهکاری و عدم کنترل آن در هر منطقه‌ای از یک کشور، موجب ناامنی و آشفتگی مردم آن منطقه و افزایش هزینه‌های تحمیل‌شده به سازمان‌های انتظامی و قضایی برای کشف جرم، و تعقیب، دستگیری، صدور و اجرای مجازات مجرمین خواهد شد. بنابراین شناخت وضعیت بزهکاری در یک منطقه، به درستی، می‌تواند در برنامه‌ریزی مفید برای پیشگیری و وقوع آن بسیار موثر واقع شود. مکان رخداد جرم یکی از اجزای مهم ارتکاب جرم است. بنابراین شناخت مکان‌های پرطرفدار رخداد جرم، یا به عبارتی مناطق جرم‌خیز، و عوامل موثر بر انتخاب مکرر آن مناطق، می‌تواند در کنترل و پیشگیری جرایم بسیار راهگشا و مفید باشند.

میزان بزهکاری در حوزه‌های شهری بیش از سایر نقاط کشور است. شهر تهران، به‌عنوان مرکز سیاسی و تجاری کشور، دارای بالاترین نرخ جرم و جنایت در کشور است و همین مشکل یکی از مهم‌ترین مسایل اجتماعی و امنیتی این شهر محسوب می‌شود. بر اساس شماعی (۱۳۹۲)، روزانه حدود ۱۵ تا ۲۰ سرقت خانگی در تهران رخ می‌دهد. با در نظر گرفتن سایر جرایم مانند قتل،

*ارایه‌دهنده مقاله: حسین باغیشنی، hbaghishani@shahroodut.ac.ir

نزاع‌های فردی و گروهی، زورگیری، کیف‌قاپی، سرقت اتومبیل، مفاسد اجتماعی، شرارت، خرید و فروش مواد مخدر و مشروبات الکلی، می‌توان درک کرد که چه حجمی از بودجه و توان نیروهای انتظامی و قضایی صرف کنترل و تامین امنیت شهر تهران در مقابله با این جرایم می‌شود. البته هزینه‌های مادی و روانی حاصل از بازخوردهای سیاسی، اجتماعی، فرهنگی و اقتصادی در کل کشور و حتی خارج از کشور را نیز نباید نادیده گرفت. بنابراین، مطالعه و پژوهش در این زمینه حائز اهمیت و سرمایه‌گذاری است.

امروزه با توجه به پیشرفت‌های صورت‌گرفته در ثبت اطلاعات ماهواره‌ای مانند سیستم اطلاعات جغرافیایی[۱] (GIS)، و ترکیب آن‌ها با داده‌های جرم، راه برای تحلیل فضایی داده‌های جرم و جنایت هموارتر شده است. از این نوع تحلیل‌ها می‌توان برای شناسایی نواحی یا نقاط جرم‌خیز، شناسایی عوامل موثر بر جرم‌خیز بودن نواحی، و پیش‌گویی احتمالی نقاط وقوع جرم استفاده کرد. با در اختیار قرار دادن این تحلیل‌ها برای نیروهای انتظامی و قضایی، می‌توان میزان جرایم را کنترل و کاهش داد.

تا کنون، پژوهش‌های مختلفی در حوزه تحلیل فضایی جرم انجام شده‌اند. به‌عنوان نمونه می‌توان به راتکلیف (۲۰۰۴)، اکرمن و موری (۲۰۰۴)، لو (۲۰۱۲)، و شماعی (۱۳۹۲) اشاره کرد. البته در این بین توجه کمی به استفاده از مدل‌های آماری فضایی برای تحلیل داده‌های جرم شده است. از بین مراجع بالا تنها می‌توان به لو (۲۰۱۲) اشاره کرد که از یک مدل رگرسیون فضایی برای تحلیل فضایی داده‌های جرم استفاده کرده است. در این مقاله برای تحلیل داده‌های نرخ جرم شرارت در نواحی ۲۲ گانه شهر تهران، از یک مدل فضایی رگرسیون بتا استفاده می‌کنیم (فراری و سریباری ، ۲۰۰۴). مدل پیشنهادی مقاله، برای وارد کردن ساختار وابستگی فضایی، از تابع مفصل گاوسی[۲] (ماسرتو و وارین ، ۲۰۱۲) استفاده می‌کند.

در بخش ۲ مدل رگرسیون کناری[۳] با ساختار وابستگی مبتنی بر مفصل گاوسی را معرفی می‌کنیم. در بخش ۳ حالت خاص مدل بتا را در قابل رگرسیون کناری برای مدل‌بندی داده‌های نرخ، تشریح می‌کنیم. سپس در بخش ۴ داده‌های نرخ جرم شرارت را در شهر تهران، بر اساس مدل پیشنهادی، تحلیل و نتایج را ارایه می‌دهیم. در پایان نیز بحث و نتیجه‌گیری خواهیم کرد.

۲ مدل رگرسیون کناری مفصل گاوسی

در رهیافت رگرسیون کناری برای داده‌های همبسته، تعیین ساختار رگرسیونی و ساختار وابستگی بین پاسخ‌ها جدا از هم انجام می‌شود. مدل‌های رگرسیون کناری برای پاسخ‌های همبسته غیرنرمال معمولا بر اساس رهیافت معادلات برآوردیاب تعمیم‌یافته[۴] (لیانگ و زگر ، ۱۹۸۶) برازش داده می‌شوند. علی‌رغم مزایای نظری و عملی مختلف، تحلیل مبتنی بر درستنمایی مدل‌های رگرسیون کناری با پاسخ‌های غیرنرمال، خیلی کم توسعه یافته‌اند؛ به‌عنوان مثال دیگل و همکاران (۲۰۰۲) را ببینید. دلیل اصلی آن هم تعیین سخت و دشوار خانواده‌های توزیع‌های چندمتغیره برای پاسخ‌های گسسته و رسته‌ای است.

مفصل گاوسی (سانگ ، ۲۰۰۰)، یک چارچوب کلی منعطف را برای مدل‌بندی انواع پاسخ‌های همبسته فراهم کرده است. مفصل گاوسی سادگی تفسیر ضرایب در مدل‌بندی رگرسیون کناری را با انعطاف موجود در تعیین ساختار وابستگی، به خوبی ترکیب می‌کند. در کنار این مزیت، استفاده از یک مدل رگرسیون مفصل گاوسی برای پاسخ‌های همبسته غیرنرمال به دلیل حضور انتگرال‌های با بعد بالا در تابع درستنمایی، محدود است و باید از روش‌های تقریبی انتگرال استفاده کرد. ماسرتو و وارین (۲۰۱۲) یک رده از

[۱] Geographical Information System
[۲] Gaussian copula function
[۳] Marginal regression
[۴] Generalized estimating equations

مدل‌های کناری رگرسیون مفصل گاوسی[۵] را ارایه کردند و نشان دادند برای غلبه بر مشکلات عددی تقریب تابع درستنمایی مدل، می‌توان از روش‌های توسعه‌یافته برای رگرسیون پرابیت چندمتغیره[۶] (چیب و گرین‌برگ ، ۱۹۹۸) در مدل‌های مفصل گاوسی به شکلی کارا استفاده کرد. دقت داشته باشید که برای پاسخ‌های پیوسته، تابع درستنمایی مدل مفصل گاوسی صورت بسته دارد؛ اما برای پاسخ‌های گسسته و رسته‌ای تقریب‌های عددی تابع درستنمایی مورد نیاز است.

فرض کنید $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ برداری از متغیرهای پاسخ وابسته پیوسته، گسسته یا رسته‌ای و $\boldsymbol{y} = (y_1, \ldots y_n)^\top$ بردار مشاهدات متناظر باشد. منشأ وابستگی بین مشاهدات می‌تواند متفاوت باشد؛ به‌عنوان مثال می‌تواند ناشی از اندازه‌های مکرر، سری زمانی یا داده‌های فضایی باشد. تابع توزیع تجمعی کناری متغیر پاسخ تکی Y_i را با $F(\cdot|\boldsymbol{x}_i)$ نشان می‌دهیم که به یک بردار p بعدی از متغیرهای تبیینی $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ وابسته است. فرض می‌کنیم $F(\cdot|\boldsymbol{x}_i)$ بر حسب پارامتر مکانی μ_i، که معمولا میانگین توزیع شرطی پاسخ است، پارامتری شده باشد به طوری که این پارامتر مکانی، بر اساس نظریه مدل‌های خطی تعمیم‌یافته (نلدر و ودربرن ، ۱۹۷۲)، از طریق رابطه

$$g_1(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \tag{۱.۲}$$

به بردار $\boldsymbol{x}$ وابسته است، که در آن $g_1(\cdot)$ یک تابع پیوند مناسب و $\boldsymbol{\beta}$ بردار p بعدی ضرایب رگرسیونی است. این چارچوب، رده‌های متنوعی از مدل‌های معمول را از جمله مدل‌های خطی تعمیم‌یافته و رگرسیون بتا (فراری و سریباری ، ۲۰۰۴)، که مدل مورد نظر ما در این مقاله است، شامل می‌شود. اگر توزیع Y_i یک پارامتر پراکندگی نیز داشته باشد، می‌توان مدل را به حالتی که پارامتر پراکندگی نیز وابسته به بردار متغیرهای تبیینی q بعدی $\boldsymbol{z}$ باشد، بر حسب مدل رگرسیون

$$g_2(\psi_i) = \boldsymbol{z}_i^\top \boldsymbol{\gamma} \tag{۲.۲}$$

بسط داد، که در آن $g_2(\cdot)$ یک تابع پیوند مناسب برای پارامتر پراکندگی ψ و $\boldsymbol{\gamma}$ بردار q بعدی پارامترهای رگرسیونی متناظر هستند (سریباری و زیلیس ، ۲۰۱۰). در این حالت تابع توزیع به شکل $F(\cdot|\boldsymbol{x}_i, \boldsymbol{z}_i)$ نوشته می‌شود.

در مدل‌های کناری هدف اصلی، ارزیابی چگونگی تغییر توزیع پاسخ Y_i بر حسب تغییرات بردار متغیرهای تبیینی است و وابستگی بین پاسخ‌ها یک هدف جانبی و البته مهم در نظر گرفته می‌شود.

۱.۲ رگرسیون مفصل گاوسی

در یک حالت کاملا جامع، یک مدل رگرسیونی به صورت

$$Y_i = h(\boldsymbol{x}_i, \boldsymbol{z}_i, \epsilon_i; \boldsymbol{\lambda}), \quad i = 1, \ldots, n \tag{۳.۲}$$

معرفی می‌شود، که در آن $h(\cdot)$ یک تابع مناسب از متغیرهای تبیینی $\boldsymbol{x}_i$ و $\boldsymbol{z}_i$ و متغیر تصادفی پنهان ϵ_i، که معمولا به‌عنوان جمله خطا شناخته می‌شود، می‌باشد و $\boldsymbol{\lambda} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$. در مدل رگرسیونی (۳.۲)، فرض می‌شود مدل به‌جز بردار پارامترهای $\boldsymbol{\lambda}$ معلوم است. در بین نسخه‌های ممکن برای تابع $h(\cdot)$، یک انتخاب مفید به صورت

$$Y_i = F_i^{-1}(\Phi(\epsilon_i); \boldsymbol{\lambda}), \quad i = 1, \ldots, n \tag{۴.۲}$$

[۵] Gaussian copula marginal regression

[۶] Multivariate probit regression

است، که در آن ϵ_i یک متغیر نرمال استاندارد، $\Phi(\cdot)$ تابع توزیع نرمال استاندارد و $F_i(\cdot;\boldsymbol{\lambda}) = F(\cdot|\boldsymbol{x}_i, \boldsymbol{z}_i;\boldsymbol{\lambda})$ هستند. قضیه تبدیل انتگرال احتمال تضمین می‌کند توزیع کناری پاسخ Y_i که توسط مدل رگرسیونی (۴.۲) در نظر گرفته می‌شود، همان $F_i(\cdot;\boldsymbol{\lambda})$ است. یک مزیت دیگر مدل پیشنهادی استفاده از جمله خطای نرمال برای ϵ_iها است.

مدل رگرسیونی (۴.۲) همه مدل‌های پارامتری رگرسیونی ممکن را برای پاسخ‌های پیوسته و ناپیوسته در بر می‌گیرد. به‌عنوان مثال، مدل رگرسیون خطی نرمال به شکل

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \sigma\epsilon_i$$

معادل است با

$$F_i(Y_i;\boldsymbol{\lambda}_i) = \Phi\left((Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})/\sigma\right)$$

که در آن $\boldsymbol{\lambda} = (\boldsymbol{\beta}^\top, \sigma)^\top$. دقت داشته باشید نگاشت بین متغیر پاسخ Y_i و جمله خطای ϵ_i تنها برای پاسخ‌های پیوسته یک به یک است؛ در سایر موارد این نگاشت یک به چند است و در نتیجه معادله (۴.۲) معکوس‌پذیر نیست.

با توجه به آن‌که هدف از این مقاله مطالعه پاسخ‌های وابسته فضایی است، بنابراین مدل معرفی‌شده با وارد کردن ساختار وابستگی از طریق بردار جملات خطای $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ کامل می‌شود. برای این کار، فرض می‌کنیم بردار خطای $\boldsymbol{\epsilon}$ دارای توزیع نرمال چندمتغیره با بردار میانگین صفر و ماتریس همبستگی Ω است؛ یعنی

$$\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \Omega). \tag{۵.۲}$$

برای شناسایی‌پذیر شدن مدل باید بردار میانگین خطا صفر باشد و واریانس‌های هر جمله ϵ_i برابر ۱ باشند. دلیل آن هم این است که ویژگی‌های تک متغیره پاسخ‌ها توسط توزیع‌های کناری، به‌طور مجزا، مدل‌بندی می‌شوند (ماسرتو و وارین ، ۲۰۱۲). این مدل حالت خاص مشاهدات مستقل را نیز با در نظر گرفتن $\Omega = I_n$، که در آن I_n یک ماتریس یکه با بعد n است، پوشش می‌دهد.

مدل رگرسیون مفصل گاوسی به شکلی جذاب دو مولفه کناری (۴.۲) و وابستگی (۵.۲) را به‌طور مجزا مدل‌بندی می‌کند. ساختارهای مختلف وابستگی در داده‌ها شامل وابستگی‌های طولی، سری زمانی و فضایی را می‌توان با مدل‌بندی پارامتری مناسب ماتریس Ω به‌عنوان تابعی از پارامترهای $\boldsymbol{\theta}$ در نظر گرفت. اکنون کل پارامترهای مدل را با $\boldsymbol{\xi} = (\boldsymbol{\lambda}^\top, \boldsymbol{\theta}')^\top$ نشان می‌دهیم، به‌طوری که $\boldsymbol{\lambda}$ را بردار پارامترهای کناری می‌نامیم و $\boldsymbol{\theta}$ را بردار پارامترهای وابستگی.

مدل رگرسیون مفصل گاوسی معرفی‌شده در (۴.۲) و (۵.۲)، با بسیاری از مطالعات انجام‌شده در زمینه مدل‌بندی مفصل، از این منظر که در آن‌ها تمرکز و علاقه اصلی بر پارامترهای وابستگی مفصل قرار دارد و کناری‌ها مولفه‌های مزاحم محسوب می‌شوند، متفاوت است. به‌عنوان نمونه‌ای از این نوع مطالعات می‌توانید هوف (۲۰۰۷) و منابع داخل آن را ببینید.

۳ رگرسیون فضایی بتا

متغیر پاسخ مورد علاقه ما در این مقاله، نرخ ارتکاب به جرم در نواحی مختلف شهر تهران است. این متغیر پاسخ، یک متغیر پیوسته با دامنه $(۰, ۱)$ است. یک مدل رگرسیونی مناسب و منعطف برای این نوع پاسخ‌ها، مبتنی بر توزیع بتا ساخته می‌شود که اولین بار توسط فراری و سریباری (۲۰۰۴) معرفی شد. فراری و سریباری (۲۰۰۴) برای برازش مدل رگرسیونی پیشنهادی خود از رهیافت درستنمایی استفاده کردند. تحلیل بیزی این رده از مدل‌ها اولین بار توسط برانسکام و همکاران (۲۰۰۷) پیشنهاد شد.

فرض کنید متغیر پاسخ Y_i دارای توزیع بتا با پارامترهای τ_i و ϕ_i باشد، که در آن $\tau_i, \phi_i > ۰$. در این صورت، میانگین توزیع برابر $\mu_i = \frac{\tau_i}{\tau_i+\phi_i}$ خواهد بود. با قرار دادن $\psi_i = \tau_i + \phi_i$، تابع چگالی توزیع بتا به صورت زیر قابل بازنویسی است:

$$f_i(y_i|\mu_i,\psi_i) = \frac{\Gamma(\psi_i)}{\Gamma(\mu_i\psi_i)\Gamma((۱-\mu_i)\psi_i)} y_i^{\mu_i\psi_i-۱}(۱-y_i)^{(۱-\mu_i)\psi_i-۱} I_{y_i}(۰,۱), \quad ۰<\mu_i<۱,\ \psi_i>۰.$$

این نوع نمایش توزیع بتا بر حسب پارامترهای میانگین و دقت[۷]، که به نوعی پراکندگی توزیع را اندازه‌گیری می‌کند، ما را قادر می‌سازد تا از چارچوب مدل رگرسیون کناری مفصل گاوسی استفاده کنیم. برای این کار در مدل رگرسیون مفصل گاوسی، قرار می‌دهیم

$$F_i(Y_i;\boldsymbol{\lambda}_i) = \int_{۰}^{Y_i} f_i(t_i|\mu_i,\psi_i)dt_i, \quad i = ۱,\ldots,n.$$

برای مدل‌بندی μ_i و ψ_i در (۱.۲) و (۲.۲) نیز، با توجه به فضای تغییرات متناظر آن‌ها، به ترتیب از توابع پیوند لجیت و لگاریتمی استفاده می‌کنیم و قرار می‌دهیم $f(y_i|\boldsymbol{x}_i,\boldsymbol{z}_i) = f_i(y_i|\mu_i,\psi_i)$.

۱.۳ ساختار وابستگی فضایی

برای داده‌های فضایی، فرض می‌کنیم مشاهده $y_i = y(\boldsymbol{s}_i)$ در موقعیت جغرافیایی $\boldsymbol{s}_i \in D \subset \Re^d$ مشاهده شده است، که در آن D ناحیه تحت مطالعه (مثل شهر تهران) و معمولا $d \geq ۲$ است. برای وارد کردن ساختار وابستگی فضایی مشاهدات، فرض می‌کنیم جملات خطا از یک میدان تصادفی گاوسی[۸] (کرسی ، ۱۹۹۳) تولید شده‌اند. یک انتخاب منعطف برای Ω در میدان تصادفی گاوسی مورد نظر، تابع همبستگی‌نگار همسانگرد مترن[۹] با ضابطه

$$\text{corr}(\epsilon_i,\epsilon_j) = \frac{۱}{۲^{\theta_۲-۱}\Gamma(\theta_۲)}\left(\frac{\|\boldsymbol{s}_i-\boldsymbol{s}_j\|_۲}{\theta_۱}\right)^{\theta_۲} \text{B}_{\theta_۲}\left(\frac{\|\boldsymbol{s}_i-\boldsymbol{s}_j\|_۲}{\theta_۱}\right)$$

است، که در آن $\text{B}_{\theta_۲}(۰)$ تابع بسل تعدیل‌شده[۱۰] مرتبه $\theta_۲$ می‌باشد. هر دو پارامتر این تابع پارامتری باید مثبت باشند. برای اطلاعات بیشتر در مورد این تابع و سایر تابع‌های همبستگی‌نگار پارامتری به کرسی (۱۹۹۳) مراجعه کنید.

۲.۳ استنباط مبتنی بر درستنمایی

توزیع بتا یک توزیع پیوسته است. بنابراین تابع درستنمایی مدل رگرسیون فضایی بتا با مفصل گاوسی، صورت بسته دارد. در واقع در این حالت داریم (سانگ ، ۲۰۰۰)

$$L(\boldsymbol{\xi}) = \phi_n(\epsilon_۱,\ldots,\epsilon_n;\Omega)\prod_{i=۱}^{n}\frac{f(y_i|\boldsymbol{x}_i,\boldsymbol{z}_i)}{\phi(\epsilon_i)} \tag{۱.۳}$$

که در آن $\phi(۰)$ تابع چگالی نرمال استاندارد و $\phi_n(\cdot;\Omega)$ تابع چگالی توزیع نرمال n متغیره با بردار میانگین صفر و ماتریس همبستگی Ω است. برای متغیرهای پاسخ گسسته، تابع درستنمایی صورت بسته ندارد و برای محاسبه آن باید انتگرال‌های n بعدی با تابع چگالی توزیع نرمال n متغیره را حل کرد. برای این منظور، رهیافت‌های تقریبی مبتنی بر روش‌های عددی (جو ، ۱۹۹۵) و نمونه‌گیری (گنز و برتز ، ۲۰۰۲) مطرح شده‌اند.

[۷] Precision
[۸] Gaussian random field
[۹] Matern isotropic correlation function
[۱۰] Modified Bessel function

ارزیابی عدم قطعیت برآوردهای درستنمایی ماکسیمم حاصل از ماکسیمم‌سازی تابع درستنمایی (۱.۳) را می‌توان با محاسبه معکوس ماتریس فیشر مشاهده‌شده انجام داد. خطای معیار محاسبه‌شده از این طریق تا وقتی معتبر است که مدل مفصل گاوسی به درستی و به دور از تشخیص نادرست[۱۱]، تعیین شده باشد. به دلیل نگرانی همیشگی از تشخیص نادرست نوع مفصل در تحلیل داده‌ها، یک راه جانشین استفاده از خطای معیار تنومند حاصل از برآوردگرهای ساندویچی[۱۲] می‌باشد. در صورت اختلاف قابل ملاحظه بین خطای معیار مبتنی بر معکوس ماتریس فیشر و خطای معیار تنومند، می‌توان تشخیص نادرست مفصل را نتیجه گرفت.

ارزیابی نیکویی مدل

برای سنجش نیکویی مدل رگرسیون مفصل گاوسی برای پاسخ‌های پیوسته، ماسرتو و وارین (۲۰۱۲) استفاده از باقی‌مانده‌های چندکی پیشگو[۱۳] را پیشنهاد کردند که به صورت زیر تعریف می‌شوند:

$$r_i = \Phi^{-1}\left(F(y_i|y_{i-1},\ldots,y_1;\hat{\boldsymbol{\xi}})\right)$$

که در آن $\hat{\boldsymbol{\xi}}$ برآوردگر درستنمایی ماکسیمم پارامترهای مدل است. تحت درستی پذیره‌های مدل (شامل پذیره گاوسی بودن نوع مفصل)، باقی‌مانده‌های چندکی r_i تقریبا تحقق‌هایی از متغیرهای ناهمبسته نرمال استاندارد هستند و به متغیرهای تبیینی $\boldsymbol{x}_i$ و $\boldsymbol{z}_i$ وابسته نیستند. این نوع باقی‌مانده‌ها را می‌توان به شکل استانداردشده

$$r_i = \frac{\hat{\epsilon}_i - \hat{m}_i}{\hat{s}_i}$$

نوشت، که در آن $\hat{\epsilon}_i = \Phi^{-1}\left(F(y_i|\boldsymbol{x}_i,\boldsymbol{z}_i;\hat{\xi})\right)$، $\hat{m}_i = \mathrm{E}\left(\epsilon_i|\epsilon_{i-1},\ldots,\epsilon_1;\hat{\xi}\right)$ و $\hat{s}_i^2 = \mathrm{Var}\left(\epsilon_i|\epsilon_{i-1},\ldots,\epsilon_1;\hat{\xi}\right)$. بنابراین، اگر پذیره‌های مدل برای داده‌ها درست انتخاب شده باشند، باید باقی‌مانده‌های چندکی مشابه باقی‌مانده‌های نرمال استاندارد رفتار کنند. مثلا نمودار چندک‌ـ‌چندک آن‌ها حول خطی با شیب ۱ پراکنده شده باشند.

۴ تحلیل داده‌های نرخ وقوع شرارت در شهر تهران

در این بخش، اطلاعات مربوط به نرخ جرم شرارت شهر تهران را با مدل پیشنهادی در این مقاله، تحلیل می‌کنیم. در این راستا، دو هدف را دنبال می‌کنیم: ۱) تهیه یک نقشه پهنه‌بندی از نرخ وقوع این نوع جرم در سطح شهر تهران و ۲) بررسی تاثیر برخی از متغیرهای تبیینی محیطی، اجتماعی و جمعیتی بر رخداد آن.

متغیر پاسخ مورد نظر نرخ ارتکاب به جرم شرارت در ۲۲ منطقه شهرداری تهران است. در کنار این متغیر پاسخ، برخی از اطلاعات مربوط به مشخصات محیطی، اجتماعی و جمعیتی این ۲۲ ناحیه نیز وجود دارند. این اطلاعات شامل جمعیت، جمعیت شناور، نرخ بیکاری، نرخ طلاق، و سرانه فضای سبز برای هر منطقه به‌عنوان متغیرهای تبیینی، در دسترس هستند.

برای تحلیل این داده‌ها از یک مدل رگرسیون مفصل گاوسی با مولفه کناری بتا برای متغیرهای پاسخ استفاده کردیم. برای دو

[۱۱] Misspecification
[۱۲] Sandwich estimators
[۱۳] Predictive quantile residuals

جدول ۱: نتایج برازش مدل فضایی رگرسیون مفصل گاوسی با کناری بتا و رگرسیون غیرفضایی بتا برای پاسخ نرخ شرارت

	مدل مفصل گاوسی فضایی			مدل غیرفضایی		
اثر	برآورد	خطای استاندارد	مقدار p	برآورد	خطای استاندارد	مقدار p
لگاریتم سرانه فضای سبز	−۱/۰۷۸	۰/۰۹۶	$< ۲e-۱۶$	−۱/۰۹۱	۰/۰۹۹	$< ۲e-۱۶$
جمعیت شناور	۰/۰۷۱	۰/۰۵۰	۰/۱۵	−۰/۰۰۴	۰/۰۴۵	۰/۹۴
نرخ بیکاری	۰/۱۰۱	۰/۰۴۷	۰/۰۳۲	۰/۰۶۴	۰/۰۴۸	۰/۱۸۰
نرخ طلاق	−۰/۸۳۶	۰/۲۰۸	$۵/۸۳e-۰۵$	−۰/۲۶۴	۰/۲۱۲	۰/۲۱۴
لگاریتم سرانه فضای سبز (دقت)	۲/۷۴۷	۰/۴۰۳	$۹/۷۶e-۱۲$	۲/۴۲۷	۰/۶۰۰	$۵/۲۲e-۰۵$
جمعیت شناور (دقت)	۰/۹۲۷	۰/۳۱۷	۰/۰۰۳	۰/۴۹۷	۰/۲۷۱	۰/۰۶۷
نرخ بیکاری (دقت)	۰/۷۸۷	۰/۱۸۱	$۱/۳۲e-۰۵$	۰/۴۶۲	۰/۱۹۴	۰/۰۱۷

پارامتر میانگین و دقت توزیع بتا، مدل‌های رگرسیونی زیر را در نظر گرفتیم:

$$\begin{aligned} \text{logit}(\mu_i) &= \beta_۰ + \beta_۱ x_{۱i} + \beta_۲ x_{۲i} + \beta_۳ x_{۳i} + \beta_۴ x_{۴i} \\ \log(\psi_i) &= \gamma_۰ + \gamma_۱ z_{۱i} + \gamma_۲ z_{۲i} + \gamma_۳ z_{۳i}, \quad i = ۱, \ldots, ۲۲ \end{aligned}$$

که در آن $x_۱ = z_۱$ لگاریتم سرانه فضای سبز، $x_۲ = z_۲$ جمعیت شناور، $x_۳ = z_۳$ نرخ بیکاری، و $x_۴$ نرخ طلاق هستند. متغیر تبیینی لگاریتم جمعیت هر ناحیه نیز به‌عنوان یک اثر با ضریب ثابت ۱[۱۴] وارد مدل شد. در ساختار وابستگی مترن نیز $\theta_۲ = ۰/۵$ را در نظر گرفتیم که معادل با تابع همبستگی‌نگار نمایی است.

نتایج برازش مدل در جدول ۱ گزارش شده‌اند. با توجه به نتایج جدول می‌توان گفت:

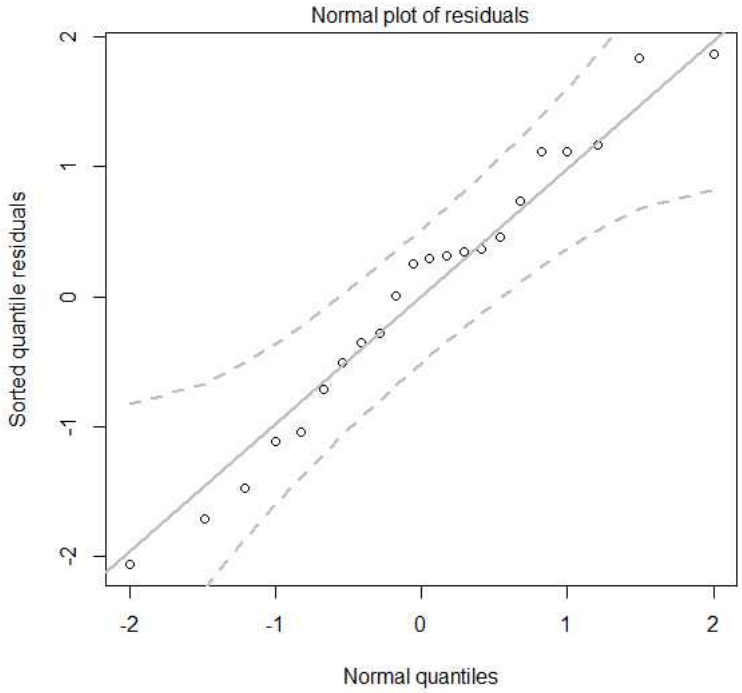
(۱) به‌جز جمعیت شناور سایر متغیرهای تبیینی بر میانگین توزیع پاسخ نرخ شرارت اثر معنی‌دار دارند. البته این متغیر تاثیر معنی‌دار بر دقت توزیع دارد.

(۲) با توجه به تفسیر پارامترها در یک مدل لجیت، بخت افزایش نرخ شرارت به کاهش آن به ازای افزایش یک واحد بر لگاریتم سرانه فضای سبز تقریبا ۳۴٪ و به ازای افزایش یک واحد بر نرخ طلاق تقریبا ۴۳٪ است. بنابراین این دو متغیر تبیینی شانس ارتکاب به شرارت را در مناطق تهران کاهش می‌دهند.

(۳) بخت افزایش نرخ شرارت به کاهش آن نیز به ازای یک واحد افزایش نرخ بیکاری نزدیک به ۱۱۱٪ است. این نتیجه به معنی افزایش شرارت در مناطقی است که نرخ بیکاری آن‌ها بیشتر است.

برای اطمینان از نیکویی برازش مدل مفصل گاوسی، نمودار باقی‌مانده‌های چندکی پیشگو را در شکل ۱ گزارش کرده‌ایم. همان‌طور که از شکل واضح است، باقی‌مانده‌های حاصل از مدل با باقی‌مانده‌های نرمال استاندارد همخوانی دارند و می‌توان به درستی مدل و پذیره‌های در نظر گرفته‌شده تکیه کرد.

[۱۴] Offset



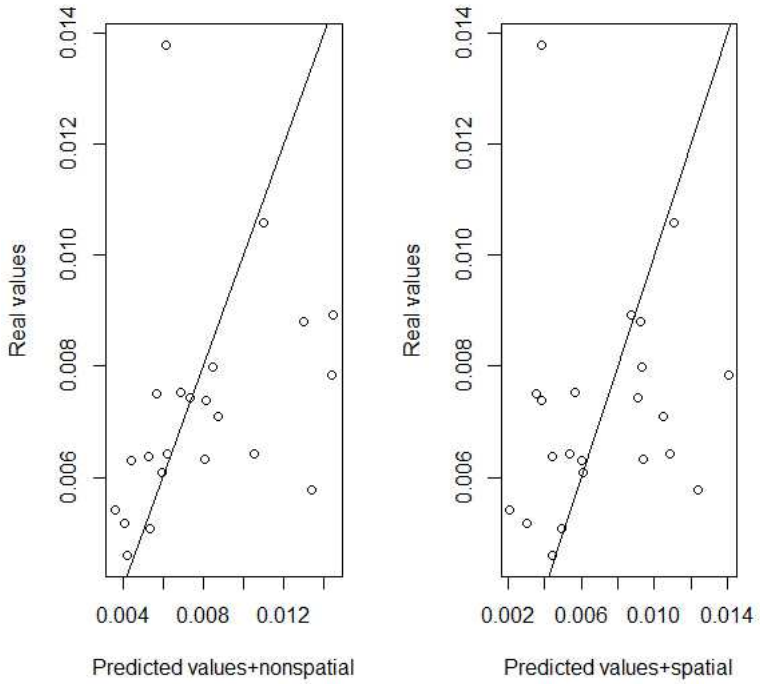
شکل ۱: نمودار باقی‌مانده‌های چندکی پیشگوی حاصل از مدل مفصل گاوسی

برای مقایسه مدل پیشنهادی با یک مدل جانشین غیرفضایی، یک مدل رگرسیون بتای معمولی را نیز به داده‌ها برازش دادیم. نتایج برازش در جدول ۱ با مدل مفصل گاوسی قابل مقایسه هستند. نتیجه برازش مدل غیرفضایی با مدل فضایی مفصل گاوسی خیلی متفاوت است. در حالت غیرفضایی تنها اثر لگاریتم سرانه فضای سبز در میانگین توزیع پاسخ معنی‌دار است. همچنین ضرایب برآوردشده مربوط به پارامتر دقت توزیع پاسخ، از نظر اندازه در دو مدل اختلاف‌های قابل ملاحظه‌ای دارند.

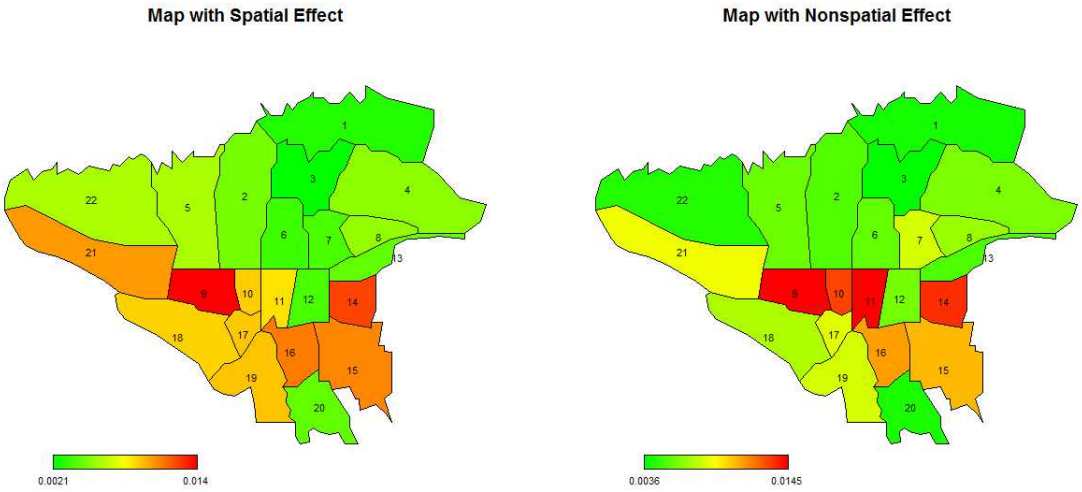
برای مقایسه دو مدل، از معیار اطلاع آکاییک[۱۵] (AIC) استفاده کردیم. مقدار این معیار برای مدل فضایی برابر ۱۹۷/۶۸− و برای مدل رگرسیون بتا برابر ۱۹۶/۹۳− است. از این منظر مدل فضایی مفصل گاوسی برتر است.

برای سنجش کیفیت برازش داده‌ها، در شکل ۲ پراکنش مقادیر برازش‌شده حاصل از دو مدل را در مقابل مقادیر واقعی نرخ جرم شرارت در ۲۲ منطقه تهران، نمایش داده‌ایم. از نمودار پراکنش مربوط به مدل فضایی معلوم است که عملکرد برازش داده‌ها توسط این مدل از عملکرد مدل رگرسیون بتای معمولی بهتر است. نقشه‌های پهنه‌بندی حاصل از دو مدل برای پیشگویی نرخ شرارت در نواحی ۲۲ گانه شهر تهران نیز در شکل ۳ نشان داده شده‌اند. نتایج پیشگویی برای دو مدل به‌ویژه در پیشگویی مناطق جرم‌خیز از نظر شرارت، با هم متفاوت هستند. به‌عنوان یک نمونه، برخلاف پیشگویی نرخ بالا در مناطق ۱۰ و ۱۱ توسط مدل معمولی رگرسیون بتا، مدل فضایی مفصل گاوسی نرخ کمتری را برای این دو منطقه پیشگویی کرده است. در مقابل، وضعیت مناطقی مثل ۱۸ و ۲۱ برای دو مدل بر عکس است؛ یعنی مدل فضایی نرخ شرارت بالاتری را نسبت به مدل معمولی برای این دو منطقه پیشگویی کرده است.

[۱۵] Akaike Information Criterion



شکل ۲: نمودار پراکنش مقادیر برازش‌شده در مقابل واقعی نرخ شرارت بر حسب دو مدل فضایی و غیرفضایی بتا



شکل ۳: نقشه‌های پهنه‌بندی نرخ شرارت شهر تهران برای دو مدل فضایی و غیرفضایی

در مجموع می‌توان بیان کرد که، با توجه به نتایج استنباط حاصل از دو مدل، استفاده از نتایج مبتنی بر مدل فضایی تاکید می‌شود.

بحث و نتیجه‌گیری

در این مقاله، نشان دادیم وارد کردن ساختار وابستگی فضایی داده‌های نرخ جرم برای اهداف تعریف‌شده شامل پیشگویی فضایی نرخ وقوع و شناخت عوامل موثر بر رخداد آن، ضروری است و نادیده گرفتن آن می‌تواند به استنباط‌های گمراه‌کننده و در نتیجه اتخاذ تصمیم‌های ضعیف و ناکارا منتهی شود. نتایج برازش مدل فضایی پیشنهادی، حکایت از دقت بیشتر نتایج حاصل از آن نسبت به رقیب غیرفضایی خود داشت.

مدل مورد نظر این مقاله به یک مفصل مشخص، یعنی مفصل گاوسی، محدود می‌شود. این انتخاب چندین مزیت، شامل تفسیرپذیری ساده نتایج، دارد که آن‌ها را از ویژگی‌های خوب توزیع نرمال چندمتغیره به ارث می‌برد. اما این نوع مفصل از محدودیت‌هایی نیز رنج می‌برد. به‌عنوان نمونه می‌توان به دو مورد اشاره کرد: ۱) ساختار وابستگی چندمتغیره بین همه پاسخ‌ها فقط توسط وابستگی‌های دومتغیره القا می‌شود؛ ۲) این مفصل نسبت به تشخیص نادرست ساختار وابستگی حساس است. با توجه به این دو مشکل، معرفی جانشین‌هایی تنومند که از وابستگی‌های توام بیش از دو متغیر به‌طور همزمان استفاده می‌کنند، از جمله پژوهش‌های جذاب در این راستا محسوب می‌شود.

به‌طور کلی، پیچیدگی محاسبه تابع درستنمایی در مدل پیشنهادی، به دلیل تجزیه ماتریس همبستگی Ω، از مرتبه $\mathcal{O}(n^۳)$ است. این پیچیدگی محاسباتی، به‌ویژه برای داده‌های فضایی حجیم، می‌تواند خیلی مهم باشد و حتی محاسبه تابع درستنمایی را غیرممکن سازد. ارایه راه‌حل‌هایی در این موارد از پیشنهادهای پژوهش جذاب آینده است.

سپاس‌گزاری

نویسندگان مقاله از معاونت اجتماعی و پیشگیری از وقوع جرم قوه قضاییه برای پشتیبانی از این تحقیق قدردانی می‌کنند.

مراجع

شماعی، ع.، قنبری، ع. و عین‌شاهی میرزا، م. (۱۳۹۲). تحلیل فضایی جرایم شهری در مناطق بیست و دو گانه کلان‌شهر تهران، *پژوهش‌های راهبردی امنیت و نظم اجتماعی*، ۶، ۱۳۰-۱۱۷.

Ackerman, W. V., and Murray, A. T. (2004). Assessing spatial patterns of crime in Lima, Ohio, *Cities*, **21**, 423-437.

Branscum, A. J., Johnson, W. O., and Thurmond, M. C. (2007). Bayesian beta regression: application to household data and genetic distance between foot-and-mouth disease viruses, *Australian and New Zealand Journal of Statistics*, **49**, 287-301.

Chib, S., and Greenberg, E. (1998). Analysis of multivariate probit models, *Biometrika*, **85**, 347-361.

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

Cribari-Neto, F., and Zeileis, A. (2010). Beta regression in R, *Journal of Statistical Software*, **34**, 1-24.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, Second Edition, Oxford University Press, Oxford.

Ferrari, S. L. P., and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportion, *Journal of Applied Statistics*, **31**, 799-815.

Genz, A., and Bretz, F. (2002). Methods for the computation of multivariate t-probabilities, *Journal of Computational and Graphical Statistics*, **11**, 950–971.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation, *Annals of Applied Statistics*, **1**, 265–283.

Joe, H. (1995). Approximation to multivariate normal rectangle probabilities based on conditional expectations, *Journal of the American Statistical Association*, **90**, 957–964.

Liang, K. L., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13-22.

Luo, X. (2012). *Spatial Patterns of Neighbourhood Crime in Canadian Cities: The Influence of Neighbourhood and City Contexts*, A thesis presented to the University of Waterloo for the degree of Master of Science in Geography, Ontario, http://etd.ohiolink.edu/.

Masarotto, G., and Varin, C. (2012). Gaussian copula marginal regression, *Electronic Journal of Statistics*, **6**, 1517-1549.

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

Ratcliffe, J. H. (2004). The hot spots matrix: a framwork for the spatio-temporal of crime reduction, *Police Practice and Research*, **5**, 23-25.

Song, P. X. K. (2000). Multivariate dispersion models generated from Gaussian copula, *Scandinavian Journal of Statistics*, **27**, 305-320.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1390

تحلیل بیزی مدل‌های نیمه‌پارامتری بقای فضایی

مینا بدخشان*، نگار اقبال، حسین باغیشنی

گروه آمار، دانشگاه صنعتی شاهرود

چکیده: با پیشرفت و گسترش ابزارهای ثبت اطلاعات مکانی و گسترش پایگاه‌های داده فضایی در سال‌های اخیر، استفاده از مدل‌های بقای فضایی نیز به‌شدت مورد توجه قرار گرفته است. دلیل آن‌هم نقش موثر و مهم اطلاعات مکانی در پیشگویی تابع بقا است. در این مقاله، دو مدل نیمه‌پارامتری بقای فضایی پرمصرف، شامل مدل‌های مخاطره‌های متناسب و بخت‌های متناسب، معرفی و برای برازش بیزی آن‌ها از الگوریتم‌های MCMC استفاده می‌شود. انواع مکانیسم‌های سانسور نیز، شامل سانسور راست، چپ و فاصله‌ای مد نظر قرار می‌گیرند. با یک مثال واقعی کاربست مدل‌های معرفی‌شده نشان داده می‌شود.

واژه‌های کلیدی: استنباط بیزی، مدل مخاطره‌های متناسب، مدل بخت‌های متناسب، همبستگی فضایی، مدل‌های شکنندگی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 60J12، 62M30.

۱ مقدمه

روش‌ها و مدل‌های آماری برای تحلیل داده‌های بقا نه فقط در پزشکی بلکه در سایر زمینه‌ها مانند مهندسی قابلیت اعتماد نیز کاربرد دارند. زمانی‌که داده‌های بقا، همبستگی فضایی داشته باشند به دلیل نقش موثری که اطلاعات جغرافیایی در پیش‌گویی بقا دارد، در نظر گرفتن این وابستگی فضایی باعث بینش دقیق‌تر و افزایش آگاهی نسبت به ماهیت و چگونگی تغییرپذیری داده‌های تحت مطالعه می‌شود؛ در صورتی‌که تغییرات فضایی معنی‌دار نادیده گرفته شوند، ممکن است منجر به استخراج نتایج نادرست و گمراه‌کننده شود.

تحلیل فضایی داده‌های بقا، به دلیل اهمیت نقشی که اطلاعات جغرافیایی در پیش‌گویی تابع بقا دارند، در سال‌های اخیر مورد توجه محققین و کاربران این حوزه قرار گرفته است. بخشی از پیشرفت‌های صورت‌گرفته در این زمینه را می‌توان به صورت زیر برشمرد: لی و رایان (۲۰۰۲) و بانرجی و همکاران (۲۰۰۳) ساختار وابستگی فضایی داده‌های بقا را در مدل‌های مخاطره‌های متناسب[1] (PH) وارد کردند و به ترتیب با انتخاب دیدگاه کلاسیک و بیزی به تحلیل آن‌ها پرداختند. بانرجی و دی (۲۰۰۵) از یک

*ارایه‌دهنده مقاله: مینا بدخشان، mina.badakhshan@yahoo.com

[1] Proportional hazards

مدل نیمه‌پارامتری بخت‌های متناسب[۲] (PO) برای مدل‌سازی داده‌های بقای وابسته فضایی استفاده کردند. لی و لین (۲۰۰۶) برای داده‌های بقا فضایی، مدل‌های نیمه‌پارامتری تبدیل نرمال را معرفی کردند و دیوا و همکاران (۲۰۰۷) ساختار همبستگی فضایی را برای داده‌های بقای فضایی در مدل‌های پارامتری لحاظ کردند. از جمله کارهای دیگر می‌توان به مطالعه بقای بیماران مبتلا به سرطان خون (هندرسون و همکاران ، ۲۰۰۲)، سرطان پروستات (وانگ و همکاران ، ۲۰۱۶؛ ژو و همکاران ، ۲۰۱۷)، مرگ و میر کودکان (ژو و همکاران ، ۲۰۱۵) و موارد فراوان دیگر اشاره کرد.

در این مقاله از دو مدل نیمه‌پارامتری PH و PO برای تحلیل داده‌های بقای فضایی استفاده می‌کنیم. برای لحاظ کردن ساختار وابستگی فضایی داده‌ها در مدل، معمولا یک متغیر پنهان فضایی به صورت یک اثر ضربی به مدل افزوده می‌شود. با توجه به وجود مکانیسم‌های مختلف سانسور در داده‌های بقا، که ویژگی بارز داده‌های بقا و وجه تمایز آن‌ها از سایر داده‌ها محسوب می‌شود، و وجود اثر تصادفی فضایی، استنباط مبتنی بر درستنمایی به دلیل پیچیدگی محاسبه تابع درستنمایی مدل گرفتار یک چالش بزرگ می‌شود که عدم کارایی محاسباتی است. بنابراین رهیافت استنباطی معمول در این مدل‌ها، رهیافت بیزی است و اجرای آن عموما توسط روش‌های مبتنی بر نمونه‌گیری مانند الگوریتم‌های مونت کارلوی زنجیر مارکوفی[۳] (MCMC) انجام می‌شود.

برای ساختار فضایی داده‌ها، هر دو نوع مشبکه‌ای و زمین‌آماری را در نظر می‌گیریم. داده‌های مشبکه‌ای فضایی، مشاهداتی هستند که در مکان‌های ناحیه‌ای دیده شده‌اند و ممکن است ناحیه‌ها منظم یا نامنظم باشند. داده‌های زمین‌آماری، داده‌هایی را شامل می‌شوند که در موقعیت‌های ثابت و مشخص در یک ناحیه پیوسته مشاهده شده‌اند. مدل‌های هر نوع با هم متفاوت هستند و هزینه محاسباتی برازش بیزی مدل‌ها، به‌ویژه برای داده‌های زمین‌آماری، می‌تواند قابل ملاحظه باشد. در ادامه، در بخش ۲ نحوه مدل‌سازی ساختار وابستگی فضایی را برای هر دو نوع مشبکه‌ای و زمین‌آماری تشریح می‌کنیم. سپس در بخش ۳ مدل‌های نیمه‌پارامتری بقای مورد نظر را معرفی می‌کنیم. در بخش ۴ مبانی و مراحل اجرای استنباط بیزی در مدل‌های پیشنهادی را ارایه می‌دهیم و در بخش ۵ نیز روش‌های معرفی‌شده را بر روی یک مثال واقعی مربوط به بیماران مبتلا به سرطان خون پیاده و نتایج را مقایسه می‌کنیم. در پایان، بحث و نتیجه‌گیری خواهیم کرد.

۲ ساختار فضایی داده‌ها

داده‌های بقایی که موقعیت جغرافیایی آن‌ها به همراه زمان رخداد پیشامدها ثبت می‌شوند، معمولا دو نوع هستند: ۱) اطلاعات مکانی افراد در سطح یک ناحیه مثل شهر، استان یا یک کشور ثبت می‌شود که به آن‌ها داده‌های مشبکه‌ای[۴] می‌گویند، یا ۲) اطلاعات مکانی موقعیت جغرافیایی فرد شامل طول و عرض جغرافیایی مثلا محل سکونت یا بستری وی ثبت می‌شود که به آن‌ها داده‌های زمین‌آماری[۵] می‌گویند. ساختار وابستگی این دو نوع داده فضایی، متفاوت هستند و برای هر کدام مدل‌های مختلفی معرفی شده‌اند (برای جزییات مدل‌ها می‌توانید به کرسی (۱۹۹۳) مراجعه کنید). البته نوع سومی از داده‌های فضایی با نام الگوهای نقطه‌ای هم وجود دارند که در مطالعات بقا کمتر مشاهده می‌شوند.

[۲] Proportional odds
[۳] Markov Chain Monte Carlo
[۴] Lattice data
[۵] Geostatistical data

۱.۲ داده‌های مشبکه‌ای

یکی از معمول‌ترین مدل‌های مشبکه‌ای برای داده‌های ناحیه‌ای[۶]، فرآیندهای اتورگرسیو شرطی ذاتی[۷] (ICAR) (بی سگ ، ۱۹۷۴) است. برای ساخت ماتریس کوواریانس یک مدل ICAR ابتدا باید ماتریس مجاورت[۸] نواحی موجود در D را تشکیل دهیم. برای این کار فرض کنید ماتریس $E = (e_{ij})$ با بعد $m \times m$ دارای درایه‌های به شکل

$$e_{ij} = \begin{cases} ۱ & i \text{ و } j \text{ همسایه باشند} \\ ۰ & \text{در غیر این‌صورت} \end{cases}$$

باشد. همچنین قرار می‌دهیم $e_{ii} = ۰$. مدل ICAR با پارامتر پراکندگی $\tau^۲$ برای $\boldsymbol{\nu}$ به کمک مجموعه‌ای از توزیع‌های شرطی به شکل

$$\nu_i | \{\nu_j\}_{j \neq i} \sim N\left(\sum_{j=۱}^{m} e_{ij}\nu_j / e_{i+}, \tau^۲ / e_{i+}\right), \quad i = ۱, ..., m$$

تعریف می‌شود، که در آن $e_{i+} = \sum_{j=۱}^{m} e_{ij}$ تعداد همسایه‌های ناحیه s_i است. این مدل ناسره است و برای شناسایی‌پذیر[۹] کردن آن از قید $\sum_{j=۱}^{m} \nu_j = ۰$ استفاده می‌شود.

۲.۲ داده‌های زمین‌آماری

برای مجموعه داده‌های زمین‌آماری، معمولا فرض می‌شود $\nu_i = \nu(s_i)$ تحققی از یک میدان تصادفی گاوسی[۱۰] (GRF) به شکل $\{\nu(s), s \in S\}$ است، به‌طوری که $\boldsymbol{\nu} = (\nu_۱, \ldots, \nu_m)$ از یک توزیع نرمال چندمتغیره به صورت $\boldsymbol{\nu} \sim N_m(۰, \tau^۲ R)$ پیروی می‌کند که در آن $\tau^۲$ میزان تغییرات فضایی در همه مکان‌ها و عضو (i, j) ماتریس R به صورت $R[i, j] = \rho(s_i, s_j)$ مدل می‌شود. تابع $\rho(۰, ۰)$ یک تابع همبستگی است که وابستگی فضایی $\nu(s)$ را کنترل می‌کند. یکی از انواع توابع همبستگی پارامتری، تابع همبستگی نمایی توانی[۱۱] با ضابطه

$$\rho(s, s') = \rho(s, s'; \phi) = \exp\{-(\phi \|s - s'\|)^\kappa\}$$

است، که در آن $\phi > ۰$ پارامتر دامنه است و کاهش وابستگی فضایی را بر حسب افزایش فاصله کنترل می‌کند. پارامتر $\kappa \in (۰, ۲]$ یک پارامتر شکل از قبل تعیین‌شده و $\|s - s'\|$ فاصله بین دو مکان s و s' است. بنابراین مدل GRF$(\tau^۲, \phi)$ به وسیله توزیع‌های شرطی زیر تعریف می‌شود:

$$\nu_i | \{\nu_j\}_{j \neq i} \sim N(-\sum_{\{j: j \neq i\}} p_{ij}\nu_j / p_{ii}, \tau^۲ / p_{ii}), \quad i = ۱, \ldots, m$$

به‌طوری‌که p_{ij} درایه (i, j) ماتریس $R^{-۱}$ است.

اگر m، یعنی تعداد داده‌های فضایی، بزرگ باشد، مشکل اصلی در مدل‌سازی این ساختار وابستگی، محاسبه معکوس ماتریس R (یا تجزیه آن) است. این مشکل به‌ویژه در استنباط بیزی که تجزیه ماتریس مورد نظر در هر مرحله نمونه‌گیری MCMC تکرار

[۶] Areal data
[۷] Intrinsic conditionally autoregressive
[۸] Adjacency matrix
[۹] Identifiability
[۱۰] Gaussian random field
[۱۱] Powered exponential

می‌شود، به شدت افزایش می‌یابد. اگر حجم نمونه m خیلی بزرگ باشد، در عمل برازش مدل غیرممکن می‌شود. در این موارد، برای کاهش هزینه‌های محاسباتی و رفع مشکل مذکور، از نسخه‌های تقریبی مانند فرآیندهای پیشگو[۱۲] (فینلی و همکاران ، ۲۰۰۹) یا تقریب کامل ـ مقیاس[۱۳] (سنگ و هوانگ ، ۲۰۱۲) استفاده می‌شود که مبتنی بر روش‌های گره[۱۴] عمل می‌کنند.

۳ مدل‌های شکنندگی نیمه‌پارامتری فضایی

فرض کنید متغیر تصادفی T زمان رخداد پیشامد مورد علاقه باشد. سه مشخصه اصلی در تحلیل بقا تابع چگالی، $f(t)$، تابع بقا[۱۵]، $S(t)$ و تابع مخاطره[۱۶]، $h(t)$، هستند. تابع بقا، احتمال رخداد پیشامد بعد از زمان t است، یعنی $S(t) = P(T > t) = ۱ - F(t)$ که در آن $F(t)$ تابع توزیع طول عمر است. تابع مخاطره نیز نرخ شکست آنی در زمان t، مشروط بر بقا تا زمان t، است که به صورت زیر تعریف می‌شود:

$$h(t) = \lim_{\Delta t \to ۰} \left\{ \frac{P(t \leq T \leq t + \Delta t)}{\Delta t \times S(t)} \right\}.$$

اکنون فرض کنید پیشامد مورد علاقه برای آزمودنی‌های مختلف در ناحیه جغرافیای $D \subset \Re^d$ رخ می‌دهد، که در آن $d \geq ۲$ و معمولا برابر ۲ است. به‌طور دقیق‌تر، فرض کنید افراد در m مکان فضایی متفاوت $s_۱, \ldots, s_m$ مشاهده شده‌اند، به‌طوری که برای هر $i = ۱, \ldots, m$، $s_i \in D$. همچنین فرض کنید t_{ij} یک زمان رخداد تصادفی مربوط به آزمودنی jام در مکان s_i و $\boldsymbol{x}_{ij}$ بردار p بعدی متغیرهای تبیینی متناظر باشد، به‌طوری که $i = ۱, \ldots, m$ و $j = ۱, \ldots, n_i$. بنابراین $n = \sum_{i=۱}^{m} n_i$ تعداد همه آزمودنی‌های تحت مطالعه است.

اگر فرض کنیم t_{ij} درون فاصله (a_{ij}, b_{ij})، $۰ \leq a_{ij} \leq b_{ij} \leq \infty$، قرار دارد، آنگاه داده‌های سانسورشده از چپ، راست و فاصله‌ای به ترتیب به صورت $(۰, b_{ij})$، (a_{ij}, ∞) و (a_{ij}, b_{ij}) هستند. برای مقادیر مشاهده‌شده (سانسورنشده) نیز $a_{ij} = b_{ij}$. بنابراین داده‌های مشاهده‌شده به صورت $\mathcal{D} = \{(a_{ij}, b_{ij}, \boldsymbol{x}_{ij}, s_i);\ i = ۱, \ldots, m, j = ۱, \ldots, n_i\}$ نمایش داده می‌شوند. معمولا برای داده‌های مشبکه‌ای فضایی، $n_i > ۱$ ولی برای داده‌های زمین‌آماری $n_i = ۱$ است.

۱.۳ مدل مخاطره‌های متناسب

مدل مخاطره‌های متناسب (کاکس ،۱۹۷۲) یکی از رایج‌ترین مدل‌های بقا برای وارد کردن متغیرهای تبیینی رگرسیونی $\boldsymbol{x}_{ij}$ و تحلیل اثرات آن‌ها روی بقا است. توابع بقا و چگالی این مدل به صورت زیر هستند:

$$\begin{aligned} S_{\boldsymbol{x}_{ij}}(t) &= S_۰(t)^{e^{\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}}} \\ f_{\boldsymbol{x}_{ij}}(t) &= \exp\{\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}\} S_۰(t)^{\exp\{\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} - ۱\}} f_۰(t) \end{aligned} \qquad (۱.۳)$$

[۱۲] Predictive processes
[۱۳] Full-scale approximation
[۱۴] Knot
[۱۵] Survival function
[۱۶] Hazard function

که در آن $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ بردار ضرایب رگرسیونی و $S_0(t)$ تابع بقای پایه با تابع چگالی $f_0(t)$ است که به ازای $\boldsymbol{x}_{ij} = 0$ به دست می‌آید. البته معمولا این مدل را بر حسب تابع مخاطره و به صورت

$$h_{\boldsymbol{x}_{ij}}(t) = h_0(t)\exp\{\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}\}$$

می‌شناسند. با توجه به تعریف $h(t) = \frac{f(t)}{S(t)}$ و رابطه (۱.۳)، به سادگی می‌توان به این شکل مدل نیز رسید.

برای مدل‌بندی تابع مخاطره پایه $h_0(t)$ از دو رهیافت پارامتری و ناپارامتری استفاده می‌شود:

- در رهیافت پارامتری، معمولا از توزیع‌های پارامتری مثل وایبل، لگ‌نرمال، و لگ‌لجستیک استفاده می‌شود.
- معمولا برای مدل‌بندی تابع مخاطره پایه از روش‌های ناپارامتری استفاده می‌شود؛ از برخی روش‌های ناپارامتری می‌توان به روش‌های تکه‌ای[۱۷]، اسپلاین‌ها و چندجمله‌ای‌ها اشاره کرد. در این مقاله از روش چندجمله‌ای برن‌اشتاین تبدیل‌یافته[۱۸] (TBP) استفاده می‌کنیم (چن و همکاران، ۲۰۱۴؛ ژو و هنسن ، ۲۰۱۷).

۲.۳ مدل بخت‌های متناسب

با همان نمادگذاری مدل PH، مدل بخت‌های متناسب دارای توابع چگالی و بقای زیر است:

$$\begin{aligned} S_{\boldsymbol{x}_{ij}}(t) &= \frac{S_0(t)\exp\{-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}\}}{1 + (\exp\{-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}\} - 1)S_0(t)} \\ f_{\boldsymbol{x}_{ij}}(t) &= \frac{f_0(t)\exp\{-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}\}}{[1 + (\exp\{-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}\} - 1)S_0(t)]^2}. \end{aligned} \qquad (۲.۳)$$

در این مدل نیز از روش ناپارامتری TBP برای مدل‌بندی تابع بقای پایه استفاده خواهیم کرد.

هر دو مدل PH و PO پذیره همگنی آزمودنی‌های تحت مطالعه را در نظر می‌گیرند. اما در عمل به‌دلیل وجود عوامل ناشناخته یا مشاهده‌نشده این پذیره برقرار نیست و نوعی ناهمگنی[۱۹] در داده‌ها به وجود می‌آید. در مدل‌های بقا برای در نظر گرفتن این ناهمگنی، از رهیافت مدل‌های شکنندگی[۲۰] استفاده می‌شود. در این حالت در رابطه (۱.۳) به مولفه $\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}$ یک مولفه شکنندگی، ν_i، اضافه می‌شود؛ بنابراین این مولفه به صورت $\boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \nu_i$ تبدیل می‌شود. همچنین در رابطه (۲.۳) به مولفه $-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}$ جمله $-\nu_i$ افزوده می‌شود و نتیجه به شکل $-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta} - \nu_i$ تبدیل می‌شود. هر ν_i اثر شکنندگی آزمودنی i را وارد مدل می‌کند. یکی از ویژگی‌های مهم مدل‌های شکنندگی بررسی سهم عوامل ناشناخته در تغییرپذیری بقای بیماران است. این سهم معمولا توسط واریانس توزیع مولفه شکنندگی اندازه‌گیری می‌شود. زمانی که داده‌ها همبستگی فضایی داشته باشند، ناهمگنی موجود در داده‌های بقا می‌تواند ناشی از ساختار وابستگی فضایی آن‌ها باشد. در این حالت مولفه شکنندگی ν یک اثر تصادفی فضایی (مشبکه‌ای یا زمین‌آماری) خواهد بود.

۴ تحلیل بیزی مدل

برای انجام استنباط بیزی در هر دو مدل PH در (۱.۳) و PO در (۲.۳)، باید توزیع‌های پیشین پارامترها و کمیت‌های نامعلوم مدل را تعیین کنیم. برای این کار ابتدا فرض کنید $\Gamma(a, b)$ توزیع گاما با میانگین a/b را نشان دهد.

[۱۷] Piecewise
[۱۸] Transformed Bernstein polynomial
[۱۹] Heterogeneity
[۲۰] Frailty models

برای پارامترهای رگرسیونی $\boldsymbol{\beta}$ از توزیع پیشین نرمال $N_p(\boldsymbol{\beta}_0, S_0)$ استفاده می‌کنیم که در آن بردار میانگین $\boldsymbol{\beta}_0$ و ماتریس واریانس S_0 ابرپارامترهای معلوم هستند. توزیع پیشین اثر تصادفی فضایی بسته به مشبکه‌ای یا زمین‌آماری بودن میدان به ترتیب عبارت است از

$$(\nu_1, ..., \nu_m)^\top | \tau \sim \text{ICAR}(\tau^2), \quad \tau^{-2} \sim \Gamma(a_\tau, b_\tau)$$

$$(\nu_1, ..., \nu_m)^\top | \tau, \phi \sim \text{GRF}(\tau^2, \phi), \quad \tau^{-2} \sim \Gamma(a_\tau, b_\tau), \quad \phi \sim \Gamma(a_\phi, b_\phi)$$

برای توزیع پیشین تابع بقای پایه از یک TBP استفاده می‌کنیم که در ادامه به اختصار معرفی می‌کنیم.

۱.۴ چندجمله‌ای برن‌اشتاین تبدیل‌یافته

در تحلیل نیمه‌پارامتری بیزی بقا، توزیع‌های پیشین ناپارامتری متنوعی برای $S_0(t)$ قابل استفاده هستند. برای مشاهده پیشنهادهای مختلف می‌توانید به مولر و همکاران (۲۰۱۵) و ژو و هنسن (۲۰۱۵) مراجعه کنید. توزیع پیشین TBP به این دلیل که در یک خانواده پارامتری مشخص مرکزی می‌شود، یکی از انتخاب‌های جذاب محسوب می‌شود. این پیشین ویژگی‌های هموار خوبی هم دارد که باعث می‌شود نمونه‌گیری از توزیع پسین حاصل کارا باشد.

برای عدد صحیح مثبت L، توزیع پیشین $\text{TBP}_L(\alpha, S_\theta(\cdot))$ به صورت زیر تعریف می‌شود:

$$S_0(t) = \sum_{j=1}^{L} w_j I(S_{\boldsymbol{\theta}}(t) | j, L-j+1), \quad \boldsymbol{w}_L \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$

که در آن $\boldsymbol{w}_L = (w_1, \ldots, w_L)^\top$ برداری از وزن‌های مثبت با توزیع دیریکله، $I(\cdot|a,b)$ تابع توزیع تجمعی بتا با پارامترهای a و b، و $\{S_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} \in \Theta\}$ یک خانواده پارامتری از توابع بقا با تکیه‌گاه $\Re^+$ است. از جمله خانواده‌های پارامتری برای $S_{\boldsymbol{\theta}}(\cdot)$ می‌توان به لگ‌لجستیک، $S_{\boldsymbol{\theta}}(t) = \{1 + (\exp\{\theta_1\}t)^{\exp\{\theta_2\}}\}^{-1}$، لگ‌نرمال، $S_{\boldsymbol{\theta}}(t) = 1 - \Phi\{(\log t + \theta_1)\exp(\theta_2)\}$، و وایبل، $S_{\boldsymbol{\theta}}(t) = 1 - \exp\{-(\exp\{\theta_1\}t)^{\exp\{\theta_2\}}\}$ اشاره کرد که در آن‌ها $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$. اگر حجم نمونه بزرگ باشد، استنباط‌های بیزی حاصل از هر سه توزیع پارامتری یکسان خواهند بود ولی برای نمونه کوچک می‌توانند تفاوت زیادی داشته باشند. یکی از ویژگی‌های این توزیع پیشین آن است که میانگین توزیع تصادفی $S_0(\cdot)$ برابر $S_{\boldsymbol{\theta}}(\cdot)$ می‌باشد؛ یعنی

$$E[S_0(t)|\alpha, \boldsymbol{\theta}] = S_{\boldsymbol{\theta}}(t).$$

پارامتر α نزدیکی وزن‌های w_j به مقدار $1/L$ را کنترل می‌کند. در واقع این پارامتر مشخص می‌کند شکل تابع پایه $S_0(\cdot)$ چقدر به حدس اولیه، یعنی $S_{\boldsymbol{\theta}}(\cdot)$، نزدیک است. مقادیر بزرگ α بیانگر اعتقاد قوی به نزدیکی $S_0(\cdot)$ به $S_{\boldsymbol{\theta}}(\cdot)$ است، به‌طوری که اگر $\alpha \longrightarrow \infty$ آنگاه با احتمال متمایل به ۱، $S_0(\cdot) \longrightarrow S_{\boldsymbol{\theta}}(\cdot)$.

۲.۴ توزیع پسین و الگوریتم MCMC

تابع درستنمایی دو مدل با پارامترهای $\boldsymbol{\zeta} = (\boldsymbol{w}_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\nu})$، با فرض مستقل بودن زمان‌های رخداد و سانسور، به صورت زیر قابل نوشتن است:

$$\mathcal{L}(\boldsymbol{\zeta}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} [S_{\boldsymbol{x}_{ij}}(a_{ij}) - S_{\boldsymbol{x}_{ij}}(b_{ij})]^{I(a_{ij}<b_{ij})} f_{\boldsymbol{x}_{ij}}(a_{ij})^{I(a_{ij}=b_{ij})}. \tag{۱.۴}$$

با ترکیب توزیع‌های پیشین معرفی‌شده برای پارامترها، $\pi(\boldsymbol{\zeta})$، که فرض می‌کنیم دارای مولفه‌های مستقل هستند، و تابع درستنمایی (۱.۴)، توزیع پسین نتیجه می‌شود:

$$\pi(\boldsymbol{\zeta}|\mathcal{D}) \propto \mathcal{L}(\boldsymbol{\zeta})\pi(\boldsymbol{\zeta}).$$

این توزیع پسین شکل بسته ندارد و برای محاسبه کمیت‌های لازم مبتنی بر آن برای استنباط، باید از روش تقریبی مبتنی بر نمونه‌گیری MCMC استفاده کرد. بر اساس ژو و هنسن (۲۰۱۷) برای اجرای الگوریتم MCMC از یک رهیافت بیزی تجربی که با یک الگوریتم متروپلیس سازوار[۲۱] (هاریو و همکاران ، ۲۰۰۱) ترکیب شده است، استفاده می‌کنیم.

با توجه به این‌که انتخاب $w_j = ۱/L$ مدل پارامتری $S_{\boldsymbol{\theta}}(t)$ را برای تابع بقای پایه $S_۰(t)$ نتیجه می‌دهد، بنابراین مدل پارامتری نقطه شروع خوبی برای توزیع پیشین TBP است. فرض کنید $\hat{\boldsymbol{\theta}}$ و $\hat{\boldsymbol{\beta}}$ برآوردهای پارامترهای $\boldsymbol{\theta}$ و $\boldsymbol{\beta}$ باشند؛ مثلا برآوردهای درستنمایی ماکسیمم پارامترها باشند. همچنین فرض کنید $\hat{V}$ و $\hat{S}$ ماتریس‌های واریانس متناظر برآوردهای پارامترها باشند. علاوه بر این، قرار می‌دهیم $\boldsymbol{z}_{L-۱} = (z_۱, \ldots, z_{L-۱})'$ به‌طوری که $z_j = \log(w_j) - \log(w_L)$. پارامترهای $\boldsymbol{\beta}$، $\boldsymbol{\theta}$، $\boldsymbol{z}_{L-۱}$، α و ϕ توسط الگوریتم‌های نمونه‌گیری متروپلیس سازوار به‌روز می‌شوند، به‌طوری که واریانس پیشنهادی اولیه آن‌ها عبارتند از $\hat{S}$ برای $\boldsymbol{\beta}$، $\hat{V}$ برای $\boldsymbol{\theta}$، $۰/۱۶I_{L-۱}$ برای $\boldsymbol{z}_{L-۱}$ و $۰/۱۶$ برای α و ϕ. هر مولفه شکنندگی فضایی، ν_i نیز توسط الگوریتم متروپلیس‌ـ هستینگز[۲۲] با واریانس پیشنهادی برابر واریانس پیشین شرطی $\nu_i|\{\nu_j\}_{j\neq i}$ به‌روز می‌شود. پارامتر $\tau^{-۲}$ نیز به کمک توزیع شرطی کامل خود که شکل بسته (گاما) دارد توسط یک مرحله گیبز به‌روز می‌شود. جزییات مراحل الگوریتم مذکور در ژو و هنسن (۲۰۱۷) در دسترس است.

برای ابرپارامترهای توزیع‌های پیشین نیز مقادیر زیر را در نظر می‌گیریم: $\boldsymbol{\beta}_۰ = ۰$، $S_۰ = ۱۰I_p$، $\boldsymbol{\theta}_۰ = \hat{\boldsymbol{\theta}}$، $V_۰ = ۱۰\hat{V}$، $a_۰ = b_۰ = ۱$، و $a_\tau = b_\tau = ۰/۰۰۱$. برای پیشین GRF در داده‌های زمین‌آماری نیز $a_\phi = ۲$ و $b_\phi = (a_\phi - ۱)/\phi_۰$ انتخاب می‌شوند، به‌طوری که مد توزیع پیشین ϕ برابر $\phi_۰$ است. در واقع $\phi_۰$ طوری انتخاب می‌شود که در تساوی

$$\rho(s', s''; \phi_۰) = ۰/۰۰۱$$

صدق کند، که در آن $\|s' - s''\| = \max_{ij} \|s_i - s_j\|$. در بعضی از متون مثل نیب و فاهرمیر (۲۰۰۷) پارامتر ϕ را مقداری ثابتی مثل $\phi_۰$ فرض کرده‌اند و برای آن پیش‌بینی تعیین نمی‌کنند.

۳.۴ مقایسه مدل و معیارهای تشخیصی

برای ارزیابی نیکویی برازش مدل از باقی‌مانده‌های کاکس و اسنل (۱۹۶۸) استفاده می‌کنیم که به صورت زیر تعریف می‌شود:

$$r(t_{ij}) = -\log\left(S_{\boldsymbol{x}_{ij}}(t_{ij})\right).$$

به شرط مفروض بودن $S_{\boldsymbol{x}_{ij}}(\cdot)$، باقی‌مانده‌های $r(t_{ij})$ دارای توزیع نمایی استاندارد هستند. اگر مدل درست باشد، تحت هر مکانیسم سانسوری، زوج‌های مرتب $(r(a_{ij}), r(b_{ij}))$ تقریبا یک نمونه تصادفی سانسورشده دلخواه از یک توزیع نمایی استاندارد هستند و بنابراین نمودار مخاطره تجمیع‌شده برآوردشده حاصل باید یک خط راست با شیب ۱ باشد. عدم قطعیت در نمودار

[۲۱] Adaptive Metropolis
[۲۲] Metropolis-Hastings

باقی‌مانده‌های کاکس و اسنل را می‌توان توسط محاسبه چند مرتبه مخاطره‌های تجمعی بر اساس نمونه‌های تولیدشده از توزیع پسین به دست آورد.

برای مقایسه مدل نیز دو معیار لگاریتم شبه‌درستنمایی کناری[۲۳] (LPML) (گیسر و اددی ،۱۹۷۹) و معیار انتخاب کیبش[۲۴] (DIC) (اسپیگل هالتر و همکاران ،۲۰۰۲) را در نظر می‌گیریم. معیار DIC که مقدار کمتر آن مدل بهتر را نشان می‌دهد، کیفیت برازش مدل را می‌سنجد؛ در مقابل LPML که مقدار بیشتر آن مدل برتر را مشخص می‌کند، بر توانایی پیشگویی مدل تاکید دارد. هر دو معیار را می‌توان از روی نمونه‌های MCMC تولیدشده محاسبه کرد.

۵ تحلیل بقای بیماران مبتلا به سرطان خون

در این بخش، کاربرد مدل‌های معرفی‌شده را بر روی مجموعه داده بیماران مبتلا به سرطان خون (هندرسون و همکاران ، ۲۰۰۲) تشریح می‌کنیم. در این مجموعه داده اطلاعات مربوط به $n = ۱۰۴۳$ بیمار شامل اطلاعات مکانی محل سکونت، سن (age)، جنسیت (sex)، تعداد گلبول‌های سفید خون (wbc)، و رتبه شهری (tpi) که مقادیر بزرگ آن نشان‌دهنده ناحیه پایین‌تر شهر است، ثبت شده‌اند. برای همه بیماران حاضر در این مجموعه داده، مختصات جغرافیایی محل سکونت آن‌ها (داده‌های زمین‌آماری) و نواحی شهری محل سکونتشان (داده‌های مشبکه‌ای) ثبت شده‌اند؛ بنابراین می‌توان هر دو ساختار وابستگی فضایی مشبکه‌ای و زمین‌آماری را برای هر دو مدل نیمه‌پارامتری مخاطره‌های متناسب و بخت‌های متناسب بر روی این داده‌ها برازش داد. یکی از اهداف این مطالعه، بررسی تاثیر تغییرات فضایی بر بقای بیماران، در حضور سایر متغیرهای تبیینی، است.

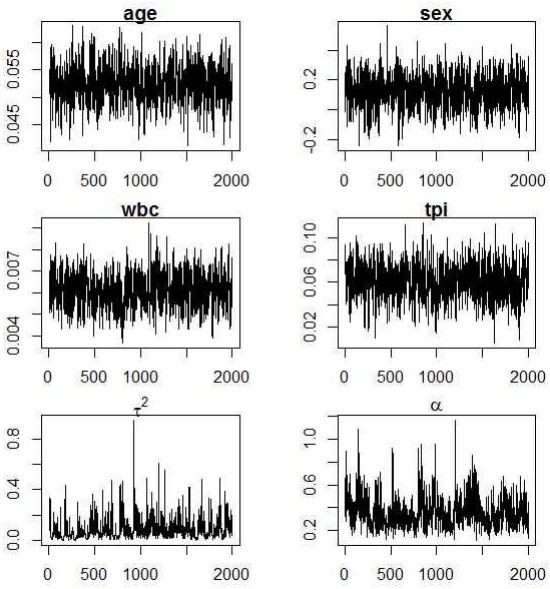
برای هر دو مدل PH و PO، در توزیع پیشین ناپارامتری TBP برای $S_۰(t)$، خانواده پارامتری وایبل را انتخاب کردیم. متغیرهای تبیینی رگرسیونی نیز قبل از ورود به تحلیل استاندارد شدند. نتایج گزارش‌شده از مدل‌های برازش‌شده در این بخش با استفاده از بسته spBayesSurv (ژو و همکاران ، ۲۰۱۷) در نرم‌افزار R به دست آمده‌اند.

۱.۵ ساختار وابستگی مشبکه‌ای: برازش دو مدل PO و PH

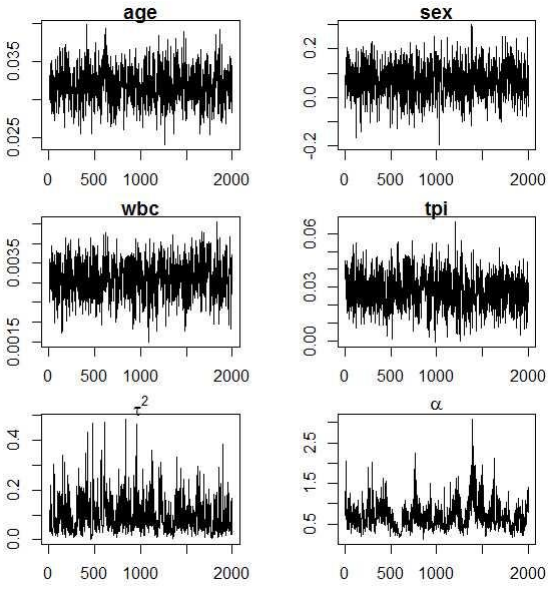
برای برازش هر دو مدل PH و PO با شکنندگی فضایی ICAR، مقدار $L = ۱۵$ را برای توزیع پیشین TBP انتخاب کردیم. سایر توزیع‌های پیشین را مشابه بخش ۲.۴ تعیین کردیم. برای اجرای الگوریتم MCMC نیز یک نمونه به حجم ۱۳ هزار تولید کردیم که با سوزاندن ۵ هزار نمونه اول و انتخاب گام ۴ برای کاهش وابستگی نمونه‌ها در نهایت ۲۰۰۰ نمونه نتیجه شد که همه استنباط‌های گزارش‌شده بر پایه این نمونه است.

نمودارهای اثر برای پارامترهای رگرسیونی و پارامترهای $\tau^۲$ و α برای دو مدل PO و PH به ترتیب در شکل‌های ۱ و ۲ نمایش داده شده‌اند. در هر دو مدل، آمیختگی زنجیر برای پارامترهای رگرسیونی مناسب است، اما آمیختگی زنجیر برای پارامتر پراکندگی مدل ICAR در هر دو مدل ضعیف است، البته در مدل PH کمی بهتر است. این نتیجه چندان هم عجیب نیست، زیرا هم از یک توزیع پیشین مبهم برای این پارامتر استفاده شده و هم تعداد نواحی، یعنی ۲۴، کوچک است.

[۲۳] Log pseduo marginal likelihood
[۲۴] Deviance information criterion

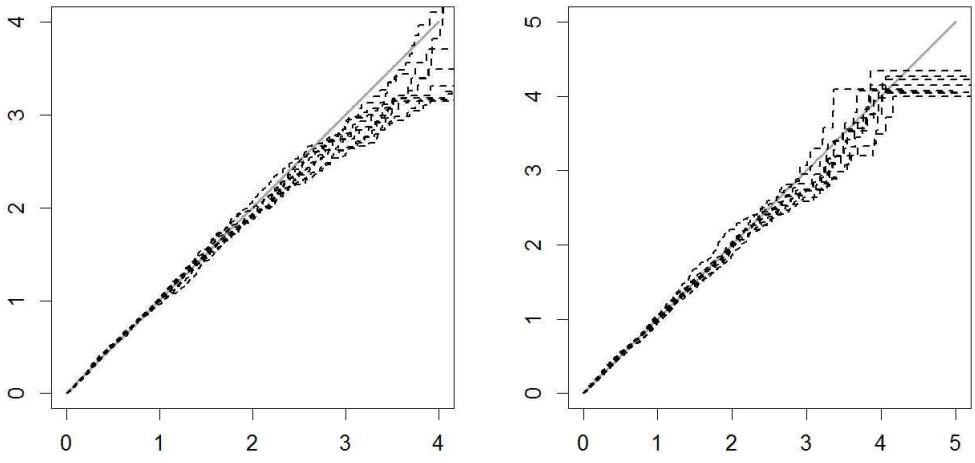


شکل ۱: نمودارهای اثر پارامترهای رگرسیونی و τ^2 و α در مدل PO با ساختار شکنندگی ICAR



شکل ۲: نمودارهای اثر پارامترهای رگرسیونی و τ^2 و α در مدل PH با ساختار شکنندگی ICAR

استنباط‌های پسینی ضرایب رگرسیونی برای هر دو مدل در جدول ۱ گزارش شده‌اند. با توجه به نتایج جدول (بر اساس فاصله‌های اعتبار ۹۵٪) می‌توان دید که متغیرهای سن، wbc و tpi عوامل خطر معنی‌دار بر بقای بیماران مبتلا به سرطان خون، بر اساس هر دو مدل، هستند. به‌عنوان مثال در مدل PO، با توجه به ضریب مثبت برآوردشده برای متغیر سن و با ثابت نگه داشتن سایر متغیرهای رگرسیونی، سن کمتر بیمار شانس مرگ او را در هر زمان دلخواهی کاهش می‌دهد؛ مثلا یک کاهش سن ۱۰ ساله شانس مرگ را به اندازه $\exp\{-۱۰ \times ۰/۰۵۲\} = ۵۹٪$ کاهش می‌دهد.



شکل ۳: نمودار باقی‌مانده‌های کاکس–اسنل برای (آ) مدل PO (ب) مدل PH با ساختار شکنندگی ICAR

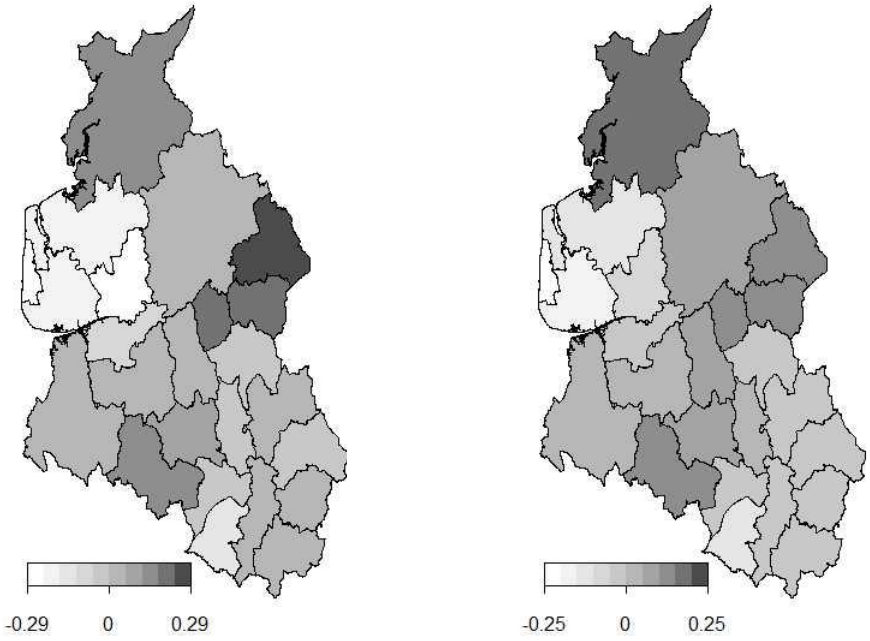
جدول ۱: آماره‌های پسینی پارامترهای دو مدل PO و PH با ساختار ICAR

	مدل PO				مدل PH			
اثر	میانگین پسینی	خطای استاندارد	کران پایین	کران بالا	میانگین پسینی	خطای استاندارد	کران پایین	کران بالا
سن	۰/۰۵۲۴	۰/۰۰۳۴	۰/۰۴۵۴	۰/۰۵۹۰	۰/۰۳۱۸	۰/۰۰۲۲	۰/۰۲۷۶۴	۰/۰۳۶۳۴
جنسیت	۰/۱۱۱۸۰۵۳	۰/۱۲۳۲	−۰/۱۰۵۴	۰/۳۳۵۱	۰/۰۶۸۹	۰/۰۶۷۵	−۰/۰۶۸۰	۰/۱۹۶۲
wbc	۰/۰۰۶۱	۰/۰۰۰۷	۰/۰۰۴۶	۰/۰۰۷۵	۰/۰۰۳۱	۰/۰۰۰۴	۰/۰۰۲۱	۰,۰۰۳۹
tpi	۰/۰۶۰۴	۰/۰۱۵۸	۰/۰۲۹۱	۰/۰۹۱۲	۰/۰۲۸۲	۰/۰۰۹۴	۰/۰۱۰۰	۰/۰۴۶۱
τ^2	۰/۰۸۴۰	۰/۰۷۸۳	۰/۰۰۴۴	۰/۲۸۳۱	۰/۰۹۱۷	۰/۰۶۱۹	۰/۰۱۶۳	۰/۲۴۲۸

برای ارزیابی نیکویی برازش دو مدل، نمودارهای کاکس–اسنل با ۱۰ باقیمانده پسینی را در شکل ۳ گزارش کرده‌ایم. با بررسی این نمودارها، می‌توان دید که نیکویی برازش هر دو مدل، به دلیل داشتن شیب نزدیک به ۱، قابل تایید است؛ اما عدم قطعیت بیشتری در مدل PO نسبت به مدل PH مشاهده می‌شود.

نقشه پهنه‌بندی میانگین‌های پسینی شکنندگی‌ها برای هر ناحیه در دو مدل در شکل ۴ نمایش داده شده است. شکنندگی بزرگ‌تر به معنی نرخ مرگ و میر بالاتر است. الگوی شکنندگی فضایی در دو مدل تقریبا مشابه هستند ولی در بعضی از نواحی مقادیر متفاوتی (با تمایل به مقادیر بزرگ‌تر برای مدل PH) برآورد شده‌اند.

برای مقایسه بین دو مدل با شکنندگی ICAR، دو معیار LPML و DIC در جدول ۲ گزارش شده‌اند. با توجه به مقادیر جدول، معیار DIC برای مدل بخت‌های متناسب کمتر و معیار LPML برای آن بیشتر است و این نشان می‌دهد که برای این مجموعه داده با شکنندگی ICAR مدل بخت‌های متناسب نسبت به مدل مخاطره‌های متناسب هم برازش بهتری دارد و هم توانایی پیشگویی بالاتر.



(ب) (آ)

شکل ۴: نمودار میانگین پسینی اثر شکنندگی‌ها برای نواحی شهری تحت مطالعه بر اساس (آ) مدل PO (ب) مدل PH

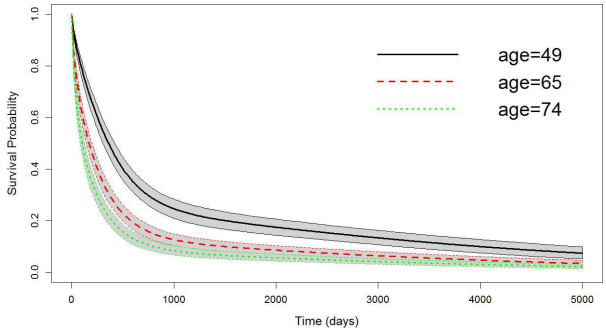
جدول ۲: معیارهای DIC و LPML برای دو مدل PO و PH با شکنندگی ICAR

	مدل PO	مدل PH
DIC	۱۱۸۵۲/۳۴	۱۱۸۹۲/۶۳
LPML	−۵۹۲۶/۳۱۲	−۵۹۴۸/۵۹۱

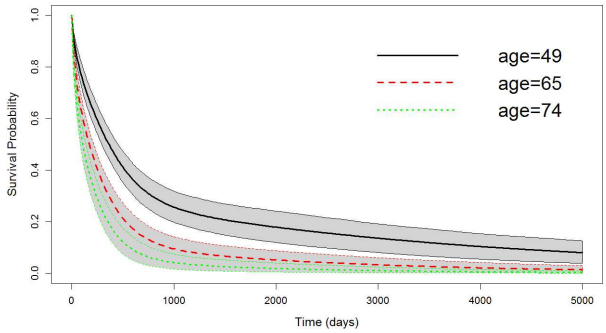
در پایان، به‌عنوان یک نمونه، منحنی‌های برآوردشده تابع بقا توسط هر دو مدل برای بیماران زن با $\text{wbc} = ۳۸/۵۹$ و $\text{tpi} = ۰/۳۳۹۸$، به ازای سن‌های مختلف، در شکل ۵ ترسیم شده‌اند. با توجه به این شکل‌های می‌توان تغییرات بقا را در سن‌های مختلف مشاهده کرد؛ احتمال بقا با بالا رفتن سن و گذر زمان کاهش می‌یابد.

۲.۵ ساختار وابستگی زمین‌آماری: برازش دو مدل PO و PH

همان‌طور که در ابتدای بخش توضیح دادیم، اطلاعات مکانی محل سکونت بیماران نیز در مجموعه داده مورد نظر ثبت شده‌اند. بنابراین در این بخش دو مدل PO و PH را برای این مجموعه داده با شکنندگی GRF نیز برازش می‌دهیم. برای این نوع شکنندگی، مختصات محل سکونت همه بیماران مجزا هستند. بنابراین برای این داده‌ها $m = ۱۰۴۳$ و $n_i = ۱$ است. برای ساختار وابستگی GRF، تابع کواریانس نمایی توانی با $\boldsymbol{\nu} = ۱$ در نظر گرفته شد. مشخصات الگوریتم MCMC مشابه حالت مشبکه‌ای انتخاب شدند.

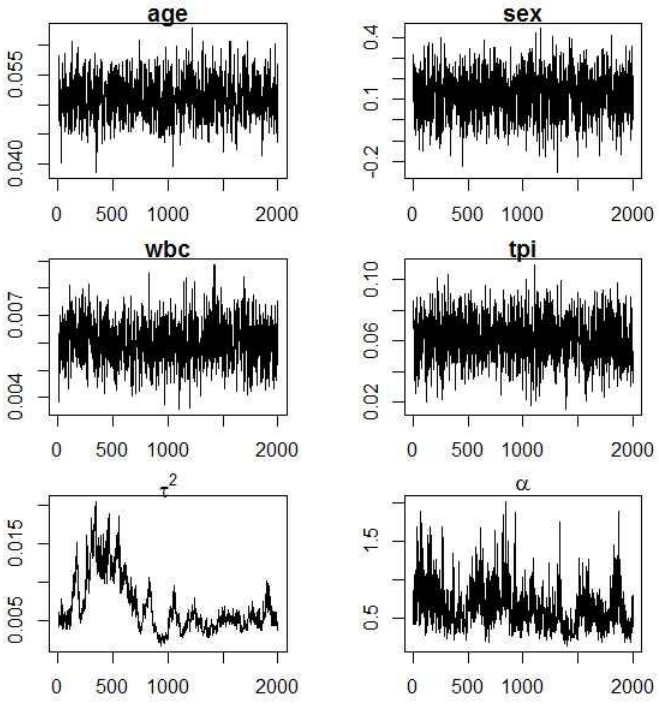


(آ)

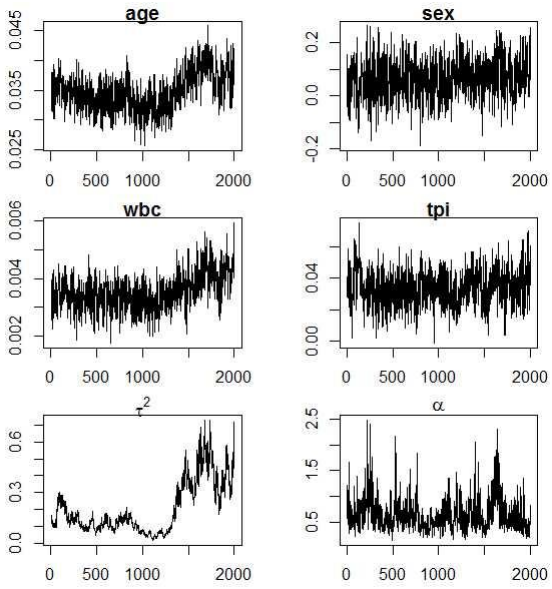


(ب)

شکل ۵: منحنی‌های برآوردشده بقا با شکنندگی ICAR برای (آ) مدل PO (ب) مدل PH



شکل ۶: نمودارهای اثر پارامترهای رگرسیونی و τ^2 و α در مدل PO با ساختار شکنندگی GRF



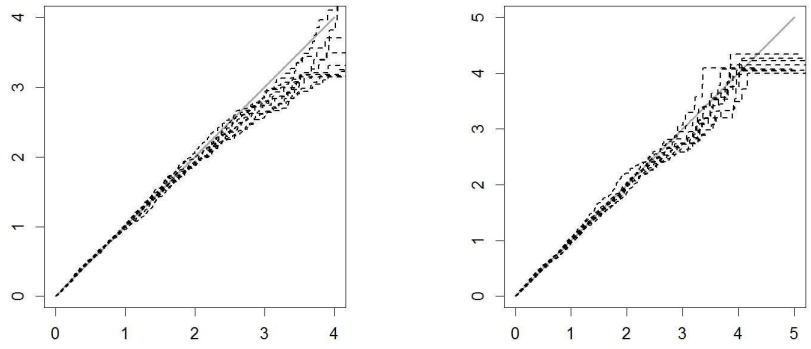
شکل ۷: نمودارهای اثر پارامترهای رگرسیونی و τ^2 و α در مدل PH با ساختار شکنندگی GRF

مشابه ساختار مشبکه‌ای، نتایج مختلف گزارش شده‌اند. نمودارهای اثر برای دو مدل در شکل‌های ۶ و ۷ نمایش داده شده‌اند. بر خلاف ساختار مشبکه‌ای، رفتار آمیختگی پارامتر τ^2 خیلی ضعیف است و نیاز به بازنگری در توزیع پیشین انتخابی برای آن، حجم نمونه مونت کارلویی و انتخاب خانواده پارامتری در توزیع پیشین TBP برای $S_0(t)$ دارد. جدول ۳ نیز استنباط‌های پسینی را برای هر دو مدل برای پارامترهای مختلف مدل نشان می‌دهد. نتایج شامل معنی‌داری پارامترهای رگرسیونی مشابه مدل با ساختار فضایی ICAR است.

جدول ۳: آماره‌های پسینی پارامترهای دو مدل PO و PH با ساختار شکنندگی GRF

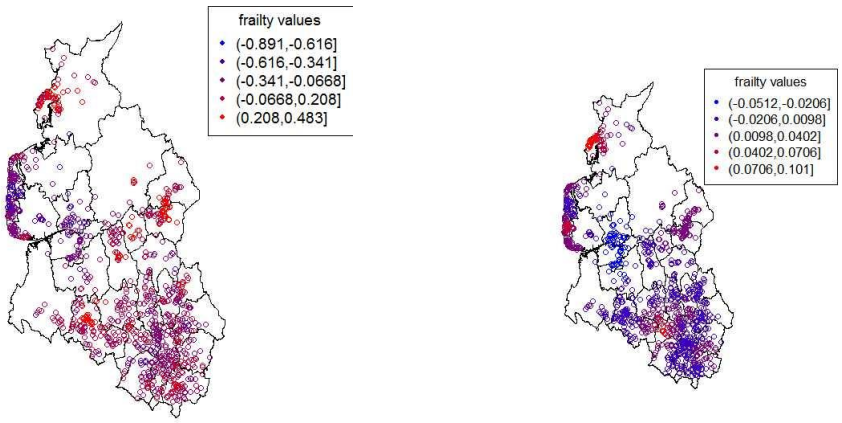
	مدل PO				مدل PH			
اثر	میانگین پسینی	خطای استاندارد	کران پایین	کران بالا	میانگینی پسین	خطای استاندارد	کران پایین	کران بالا
سن	۰/۰۵۱۲	۰/۰۰۳۴	۰/۰۴۴۹	۰/۰۵۷۶	۰/۰۳۴۷	۰/۰۰۳۱	۰/۰۲۹۰	۰/۰۴۱۴
جنسیت	۰/۱۰۷۱	۰/۱۱۹۷	−۰/۰۹۰۵	۰/۳۲۱۴	۰/۰۶۷۱	۰/۰۷۱۹	−۰/۰۶۹۷	۰/۲۰۵۸
wbc	۰/۰۰۶۰	۰/۰۰۰۷	۰/۰۰۴۵	۰/۰۰۷۵	۰/۰۰۳۵	۰/۰۰۰۶	۰/۰۰۲۴	۰/۰۰۴۸
tpi	۰/۰۶۱۲	۰/۰۱۴۸	۰/۰۳۲۲	۰/۰۸۹۱	۰/۰۳۳۴	۰/۰۱۰۸	۰/۰۱۳۸	۰/۰۵۵۷
τ^2	۰/۰۰۶۸	۰/۰۰۳۵	۰/۰۰۲۷	۰/۰۱۶۰	۰/۲۰۶۶	۰/۱۵۵۷	۰/۰۳۶۰	۰/۵۷۸۰

نمودارهای باقی‌مانده‌های کاکس–اسنل (شکل ۸) نیکویی برازش هر دو مدل را، مشابه ساختار مشبکه‌ای، تایید می‌کنند. پهنه‌بندی نقشه میانگین پسینی شکنندگی (شکل ۹) نیز نقاط با خطر بالاتر را به‌خوبی نشان می‌دهد. در این ساختار، تفاوت زیادی در تغییرپذیری فضایی خطر مرگ به دلیل سرطان خون در دو مدل PH و PO به چشم می‌آید. در واقع مدل مخاطره‌های متناسب خطر مرگ را خیلی بیشتر از مدل بخت‌های متناسب در نقاط مکانی برآورد کرده است. بنابراین استفاده از معیارهای انتخاب مدل در این ساختار فضایی حساس‌تر است.



(ب) مدل PH (آ) مدل PO

شکل ۸: نمودار باقی‌مانده‌های کاکس–اسنل برای (آ) مدل PO (ب) مدل PH با ساختار شکنندگی GRF



(ب) (آ)

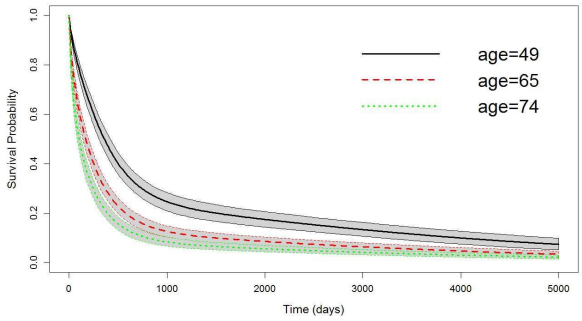
شکل ۹: نمودار میانگین پسینی اثر شکنندگی‌ها برای محل‌های سکونت بیماران بر اساس (آ) مدل PO (ب) مدل PH

جدول ۴ مقادیر معیارهای انتخاب مدل را برای دو مدل نشان می‌دهد. با مقایسه نتایج ملاحظه می‌کنید که هر دو معیار DIC و LPML مدل بخت‌های متناسب را نسبت به مدل مخاطره‌های متناسب برتر می‌دانند.

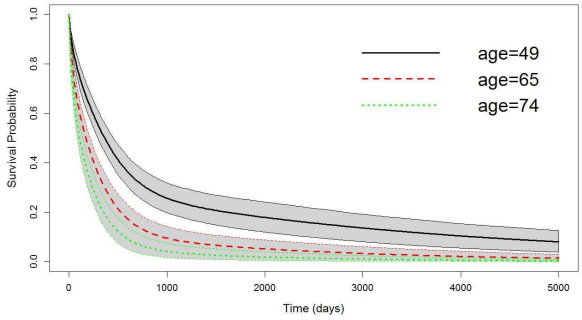
جدول ۴: معیارهای DIC و LPML برای مدل PO و PH با ساختار وابستگی GRF

	مدل PO	مدل PH
DIC	۱۱۸۶۰/۹۱	۱۱۸۶۲/۵۳
LPML	−۵۹۳۱/۰۹۷	−۵۹۴۵/۰۸۱

منحنی‌های برآوردشده تابع بقا نیز در شکل ۱۰ برای همان مقادیر wbc و tpi منتخب در ساختار مشبکه‌ای، نشان داده شده است.



(آ)



(ب)

شکل ۱۰: منحنی‌های برآوردشده بقا با شکنندگی GRF برای (آ) مدل PO (ب) مدل PH

بحث و نتیجه‌گیری

در این مقاله دو مدل پرطرفدار نیمه‌پارامتری بقا را برای مدل‌بندی داده‌های فضایی بقا مطرح کردیم. هر دو ساختار وابستگی فضایی مشبکه‌ای و زمین‌آماری را مطرح و استنباط بیزی این دو مدل را به کمک الگوریتم‌های MCMC تشریح کردیم. اهمیت مدل‌های مطرح‌شده برای مدل‌سازی داده‌های فضایی بقا، به دلیل عملی و مناسب بودن برازش و تفسیر ساده مدل، در کاربردهای واقعی قابل اشاره است. استفاده از سایر توزیع‌های پیشین ناپارامتری برای تابع بقای پایه، انتخاب متغیرهای تبیینی، و به‌کارگیری روش‌های تقریبی برای تقریب میدان تصادفی GRF از جمله مطالبی است که در این حوزه می‌توان به آن‌ها پرداخت.

مراجع

Banerjee, S., Wall, M., and Carlin, B. P.(2003). Frailty modelling for spatially correlated survival data with application to infant mortality in Minnesota, *Biostatistics*,**4**,123–142.

Banerjee, S., and Dey, D. K.(2005). Semiparametric proportional odds model for spatially correlated survival data,*Lifetime Data Analysis*, **11**, 175–191.

Besag, J.(1974).Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society*, **36**(2), 192–236.

Chen, Y., Hanson, T., and Zhang, J. (2014). Accelerated hazards model based on parametric families generalized with Bernstein polynomials, *Biometrics*, **70**(1), 192–201.

Cox, D. R., and Snell, E. J. (1968). A general definition of residuals, *Journal of the Royal Statistical Society: Series B (Methodological)*, **30**(2), 248–275.

Cox, D. R. (1972).Regression models with life tables, *Journal of the Royal Statistical Society*, **34**, 87–220.

Cressie, N. (1993). *Statistics for Spatial*, Revised Edition, Statistics for Spatial.

Diva, U. A., Banerjee, S., and Dey, D. K. (2007) Modeling spatially correlated survival data for individuals withmultiple cancers, *Statistical Modeling*, **7**(2), 1–23.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets, *Computational Statistics and Data Analysis*, **53**(8), 2873–2884.

Geisser, S., and Eddy, W. F. (1979). A Predictive approach to model selection, *Journal of theAmerican Statistical Association*, **74**(365), 153–160.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm, *Bernoulli*, **7**(2), 223–242.

Henderson, R., Shimakura, S., and Gorst, D. (2002) Modeling spatial variation in leukemia survival data, *Journal of the American Statistical Association*, **97**(460), 965–972.

Kneib, T., and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression, *Scandinavian Journal of Statistics*, **34**(1), 207–228.

Li, Y., and Ryan, L. (2002). Modeling spatial survival data using semiparametric frailty models, *Biometrics*,**58**, 287–297.

Li, Y., and Lin, X. (2006). Semiparametric normal transformation models for spatially correlated survival data, *Journal of the American Statistical Association*, **101**, 591–603.

Muller, P., Quintana, F., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*, Springer-Verlag: New York.

Sang, H., and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **74**(1), 111–132.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B*, **64(4)**, 583–639.

Wang, S., Zhang, J., Lawson, A. B. (2016). A Bayesian normal mixture accelerated failure time spatial model and its application to prostate cancer, *Statistical methods in medical research*, **25**(2), 793–806.

Zhou, H., Hanson, T., Jara, A., and Zhang, J. (2015a). Modeling county level breast cancer survival data using a covariate-adjusted frailty proportional hazards model, *The Annals of Applied Statistics*, **9**(1), 43–68.

Zhou, H., Hanson, T., and Knapp, R. (2015b). Marginal Bayesian nonparametric model for time to disease arrival of threatened amphibian populations, *Biometrics*, **71**(4), 1101–1110 .

Zhou, H., and Hanson, T. (2015). Bayesian spatial survival models, *In Nonparametric Bayesian Inference in Biostatistics*, **215**, 215–246.

Zhou, H., and Hanson, T. (2017). A unified framework for fitting Bayesian semiparametric models to arbitrarily censored spatial survival data, *arXiv preprint arXiv*,1701.06976.

Zhou, H., Hanson, T., and Zhang, J. (2017). Generalized accelerated failure time spatial frailty model for arbitrarily censored data, *Lifetime Data Analysis*, **23**(3), 495–515.

Zhou, H., Hanson, T., and Zhang, J. (2017). spBayesSurv: Fitting Bayesian Spatial Survival Models Using R, *arXiv preprint arXiv*,1705.04584.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶

مدل‌بندی فضایی‌ـزمانی داده‌های fMRI برای یافتن نقاط فعال و ارتباطات تابعی بین نواحی مغز

نسرین برومندنیا [*۱]، حمید علوی مجد[۱]، فرید زایری[۱]، احمد رضا باغستانی[۱]، سید محمد طباطبایی[۲] ، فریبرز فائقی[۳]

[۱]گروه آمار زیستی، دانشگاه علوم پزشکی شهید بهشتی

[۲]گروه انفورماتیک پزشکی، دانشگاه علوم پزشکی شهید بهشتی

[۳]گروه تکنولوژی پرتو شناسی، دانشگاه علوم پزشکی شهید بهشتی

چکیده: از چالش‌های مهم در حیطه تحلیل داده‌های fMRI، مدل‌سازی آماری حجم مغز در فضای سه بعدی است به طوری که ماهیت فضایی داده‌ها لحاظ شود. به علاوه این داده‌ها همبستگی زمانی نیز دارند و به صورت سری‌های زمانی ثبت می شوند. در سال‌های اخیر مدل‌های بیزی فضایی‌ـزمانی در این حیطه رو به گسترش هستند. این مدل‌ها با هدف تعیین نواحی فعال مغز در اثر یک محرک خاص برازش داده می‌شوند. تعیین ارتباطات بین نواحی مغز نیز هدف دیگری در این مطالعات است که با خوشه‌بندی سری‌های زمانی با رفتار مشابه در طول زمان قابل دسترسی است. به علت حجم بسیار زیاد مشاهدات و نیز ساختارهای همبستگی پیچیده خاص این نوع از داده‌ها، مدل‌های ارایه شده عموما از برخی ویژگی‌ها چشم پوشی می‌کنند تا مدل‌سازی آماری ساده‌تر گردد. در مطالعه حاضر به مدل‌سازی داده‌های fMRI پرداخته شده است به گونه‌ای که ویژگی‌های خاص داده‌ها در مدل‌سازی لحاظ گردند. مدل ارایه شده امکان لحاظ کردن همبستگی فضایی از طریق پیشین Ising و نیز خوشه‌بندی سری‌های زمانی با استفاده از پیشین فرآیند دیریکله را فراهم می‌کند. به علاوه مدل ارایه شده تابع پاسخ همودینامیکی را برای هر واکسل به طور مجزا برآورد می‌کند. پروسه براورد پارامترها بر اساس رویکرد MCMC با ترکیب الگوریم و روش متغیر اضافه انجام شده است.

واژه‌های کلیدی: داده fMRI، ساختار همبستگی فضایی سه بعدی، پیشین فضایی Ising

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 91B72.

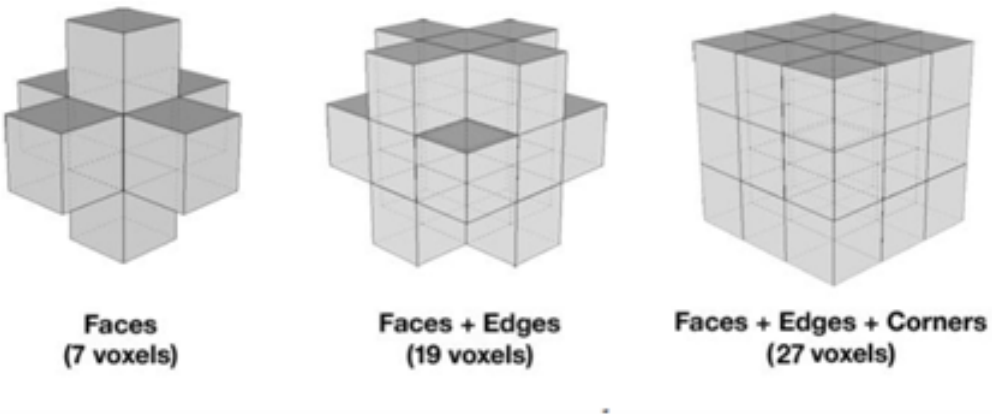
---

[*] **ارایه‌دهنده مقاله:** نسرین برومندنیا، borumand.n@gmail.com

۱ مقدمه

در یک مطالعه fMRI فرد درون دستگاه اسکنر بدون حرکت دراز کشیده و در حالی که محرک خاصی را انجام می‌دهد یا در حال استراحت است فعالیت مغزی وی ثبت می‌شود. در این فرآیند در طول زمان از حجم مغز با فاصله‌های زمانی کوتاه، چند ثانیه، تصاویری سه بعدی گرفته می‌شود که میزان اختلاف بین سطح خون اکسیژن دار به خون بدون اکسیژن را در هر ناحیه از مغز نشان می‌دهد که BOLD[۱] نامیده می‌شود (پولدارک ، ۲۰۱۱). حجم مغز متشکل از واحدهای کوچکی به نام واکسل در نظر گرفته می‌شود. در هر واکسل یک سری زمانی از پاسخ BOLD اندازه گیری می‌شود که در واقع نشان دهنده سیر زمانی فعالیت مغزی آن واکسل می‌باشد. با در نظر گرفتن تعداد اسکن‌ها و با توجه به این نکته که حجم مغز از صدها هزار واکسل تشکیل شده است، می‌توان گفت وقتی یک آزمایش fMRI بر روی یک فرد انجام می‌شود داده‌های حاصل از آن، مجموعه عظیمی از صدها هزار سری زمانی از پاسخ BOLD خواهد بود که از نواحی مختلف مغز بدست آمده است. با توجه به آنچه گفته شد پاسخ‌های BOLD هر واکسل به صورت زمانی به هم وابسته‌اند.

به علاوه ویژگی کلیدی دیگری که در مورد داده‌های fMRI مطرح می‌شود همبستگی مکانی داده‌ها است. چرا که انتظار می‌رود پاسخ در هر واکسل خاص، مشابه پاسخ‌ها در واکسل‌های همسایه آن باشد (لی و همکاران ، ۲۰۱۴). به علاوه همبستگی در میان واکسل‌ها لزوما با زیاد شدن فاصله آنها از بین نمی رود. بنابراین همبستگی مکانی بین واکسل‌ها در تحلیل داده‌های fMRI باید لحاظ شود. نمایی از همبستگی سه بعدی واکسل‌ها در شکل ۱ نشان داده شده است.



شکل ۱: همبستگی واکسل‌ها در فضای ۳ بعدی.

در مطالعات fMRI، ماتریس طرح X وابسته به زمان است و مقادیر آن طبق طراحی مطالعه صفر و یک خواهد بود . در مواقعی که X مقدار ۱ را دارد (زمان‌های اعمال محرک) در واقع متغیر پاسخ نسبت به زمان شروع اعمال محرک، با تاخیر دریافت می‌شود. چنین پاسخی به عنوان تابع پاسخ همودینامیکی یا به طور مختصر HRF شناخته می‌شود. شکل این تابع در بین افراد مختلف، واکسل‌ها، نواحی مختلف مغز و به ازای محرک‌های گوناگون تغییر پذیری اساسی دارد که لازم است در مدل‌سازی لحاظ شود (هاندورکر، اولینگر و اسپوزیتو ، ۲۰۰۴؛ لو و همکاران ، ۲۰۰۷؛ مارلک و همکاران ، ۲۰۰۳). فرم کلی تابع HRF در شکل ۲ نشان داده شده است. گستره‌ای از مدله اطوری طراحی شده‌اند که بتوانند این تاخیر بین اعمال محرک و دریافت پاسخ را به

[۱] Blood Oxygen Level Dependent

حساب آورند. در این راستا با توجه به شکل تابع، می‌توان از پیچش تابع HRF با تابع اعمال محرک برای این کار استفاده کرد (ژانگ و همکاران ، ۲۰۱۵). علاوه بر موارد فوق، موضوع دیگری که در داده‌های fMRI مطرح می‌شود مساله انتخاب واکسل‌های فعال یا مساله انتخاب متغیر است. به عبارت دیگر مساله انتخاب متغیر معادل تعیین ضرایبی از مدل رگرسیونی ۱ است که غیر صفر هستند (جرج و مک لوچ ، ۱۹۹۷). با مروری بر مطالعات پیشین در این حیطه این نکته مشخص است که به منظور ساده سازی مدل‌سازی عموما برخی از ویژگی‌های مذکور در مدل نادیده گرفته می‌شوند. به عنوان مثال لی و همکاران و پنی و همکاران مدل هایی را به کار گرفتند که هر دو همبستگی زمانی و مکانی را شامل می‌شدند، اما تابع HRF در مدل‌های آنها از قبل تعیین شده بود (لی و همکاران ، ۲۰۱۴؛ پنی، تروخیلو و فریستون ، ۲۰۰۵). برتری مدل ژانگ و همکاران (۲۰۱۴) نسبت به دو مدل مذکور، برآورد تابع HRF به‌طور جداگانه برای هر واکسل است. به علاوه ویژگی دیگری که این مدل را متمایز از سایر مدل‌های موجود می‌کند، خوشه‌بندی واکسل‌های فعال بر اساس مشابهت عملکرد آنها در طول زمان است. البته نکته حائز اهمیت در مدل ارایه شده توسط ژانگ و همکاران (۲۰۱۶) لحاظ کردن همبستگی مکانی در دو بعد است. در واقع مدل مذکور برای یک حجم مغز برازش داده نشد، بلکه مدل‌سازی در هر قطعه به طور جداگانه انجام گرفته است.

ولریش و همکاران (۲۰۰۴) ساختار اتورگرسیو بردار زمان_مکان جدانشدنی‌ای روی خطاهای مدل قرار دادند و نیز تابع HRF را نیز برآورد کردند. کروش، مونتس دیاز و گامرمن (۲۰۱۰)؛ گوسل، آئر و همکاران (۲۰۰۱)؛ اسمیت و همکاران (۲۰۰۳)؛ فلاندین و پنی (۲۰۰۷)؛ هریسون، پنی و فریسون (۲۰۰۳)؛ هریسون و همکاران (۲۰۰۸)؛ اسمیت و همکاران (۲۰۰۳) همگی از پیشین‌های فضایی بر روی پارامترهای مدل استفاده کرده‌اند. اما این نقص را داشته‌اند که ساختار خطای مستقل در نظر گرفته‌اند. کروش، مونتس دیاز و گامرمن (۲۰۱۰)؛ گوسل، آئر و همکاران (۲۰۰۱) هم چنین برآورد تابع HRF را در نظر گرفته‌اند.

در اجرای هر مطالعه fMRI اهداف گوناگونی ممکن است مطرح باشد که مطالعه جهت پاسخ به آن هدف طراحی می‌شود. در مطالعه حاضر به مدل‌سازی آماری از نقطه نظر تعیین نقاط فعال مغز بر اثر یک محرک خاص و نیز تعیین ارتباطات تابعی بین نواحی مغز پرداخته می‌شود. تلاش بر این است که مدل‌سازی به گونه‌ای انجام شود که ویژگی‌های ذکر شده در بالا در مدل لحاظ گردند. قابل ذکر است که هدف از مطالعات ارتباطات تابعی، یافتن نواحی از مغز است که در حال دریافت یک محرک خاص فعالیت مشابه نشان می‌دهند (فریسون ، ۲۰۱۱). در ادامه مقاله حاضر به این صورت تنظیم شده است: بخش دوم به معرفی مدل ارایه شده می‌پردازد. نتایج حاصل از برازش مدل ارایه شده به داده‌های مربوط به محرک شنوایی در بخش ۳ آمده است. جمع بندی نکات و بحث در بخش ۵ ارایه شداند.

۲ روش کار

در این بخش سری‌های زمانی fMRI به گونه‌ای مدل بندی شد که دو هدف از مطالعه fMRI تامین شود؛ هدف اول تعیین نواحی فعال مغز و به طور همزمان هدف دوم ارتباطات تابعی بین نواحی بوده است. به عبارت دیگر این مدل نواحی از مغز را که توسط یک محرک خاص فعال می‌شوند شناسایی کرده و هم چنین سری‌های زمانی با ویژگی‌های مشابه در طول زمان خوشه‌بندی می‌شوند.

مدل خطی تعمیم‌یافته در سطح واحدهای نمونه، به صورت زیر تعریف می‌شود:

$$y_v = X_v\beta_v + \varepsilon_v, \varepsilon_v \; N_{T_v}(\cdot, \sigma_v^2\Lambda_v)$$

که در آن Y_v سری زمانی پاسخ‌های ثبت شده برای نمونه‌ی v ام که $v = 1, 2, \ldots$ و N تعداد کل نمونه‌ها است، ماتریس X_v

طرح به ابعاد $T \times P$، که P تعداد محرک‌های مورد آزمایش و T طول سری زمانی (یعنی تعداد اسکن‌ها در طول زمان) است و حاصل آمیختن تابع HRF با ماتریس طرح است. $\int_{\cdot}^{t} x(s)h_v(t-S)ds$. به این صورت که هر ستون از ماتریس طرح با تابع HRF آمیخته می‌شود. در این مطالعه تابع پواسون $h_v(t) = (-\lambda_v)\frac{\lambda_v^t}{t!}$ به عنوان تابع RFH در نظر گرفته شد که برای هر واکسل پارامتر آن به طور مجزا برآورد گردید به علاوه $\beta_v = (\beta_{v,1}, \ldots, \beta_{v,p})$ بردار ضرایب رگرسیونی به ابعاد $P \times 1$ است و ε_v بردار خطای $T \times 1$ در مدل رگرسیونی است.

در مطالعات fMRI تعیین وضعیت فعال شدن یا نشدن واحدهای نمونه، یا واکسل‌ها، در قبال اعمال محرک مربوطه مد نظر است. بردار متغیرهای نشانگر $\gamma_v = (v_{v,1}, \ldots, \gamma_{v,p})^\top$ دنباله‌ای از صفر و یک هاست که به ترتیب نشان دهنده پذیرفته شدن و یارد شدن فرض آماری $\beta_{v,j} = 0$ برای هر نمونه است. بنابراین وقتی $\gamma_{v,j} = 0$ باشد ضریب $\beta_{v,j}$ برابر با صفر خواهد شد و اگر $\gamma_{v,j} = 1$ باشد ضریب رگرسیونی مربوطه غیر صفر است. در نتیجه بازنویسی مدل قبل به این صورت است:

$$y_v = X_v(\gamma_v)\beta_v(\gamma_v) + \varepsilon_v, \ \varepsilon_v \sim N_{T_v}(0, \sigma_v^2 \Sigma_{a_v})$$

عبارت خطای ε_v نویزهای تصادفی را لحاظ می‌کند. در مدل ارایه شده ε_v به صورت پروسه طولانی مدت مدل بندی شد که تابع خودهمبستگی آن دارای یک پارامتر تاخیر است، به عبارت دیگر

$$\gamma(h) \sim Ch^{-a}$$

که در آن $\alpha_v \in (0, 1)$ پارامتر حافظه طولانی مدت است و C مقدار ثابتی است (فریسون ، ۲۰۱۱) در ادامه تبدیل موجک گسسته بر دو طرف مدل فوق اعمال گردید. پس از اعمال تبدیل موجک گسسته بر دو طرف مدل، مدل در حیطه موجک به این صورت خواهد بود:

$$y_{vw} = X_{vw}\beta + \varepsilon_{vw}, \quad \varepsilon_{vw} \sim N(0, \Sigma_{\varepsilon_{vw}})$$

که $X_{vw} = WX_v$، $Y_{vw} = WY_v$ و $\varepsilon_{vw} = W\varepsilon_v$. ماتریس W با ابعاد $N \times N$ شامل مقادیر حقیقی است و متعامد است. پس از اعمال تبدیل موجک گسسته ساختار ماتریس کواریانس برای Y_w به صورت $\Sigma_{\varepsilon_{vw}} = \sigma_v^2 \Sigma_{a_v}$ تبدیل می‌شود، که ماتریسی قطری است و اعضای روی قطر اصلی آن $\sigma_v^2 \sigma_{mn}^2$ هستند (فریسون ، ۲۰۱۱؛ کو و ونوچی ، ۲۰۰۶؛ ژانگ و همکاران ، ۲۰۱۶).

بر اساس رویکرد ژانگ و همکاران، عبارت واریانس برای ماتریس کواریانس ضرایت موجک به صورت $\sigma_v^2 \sigma_{mn}^2 = \sigma_v^2 (2^{(a_v)})^m$ خواهد بودکه σ_v^2 واریانس و $\alpha_v \in (0, 1)$ پارامتر حافظه طولانی مدت است (ژانگ و همکاران ، ۲۰۱۶).

دیدگاه ارایه شده در این پژوهش، چندین ویژگی داده‌های fMRI را در ساختار مدل‌سازی لحاظ می‌کند. این ویژگی‌ها از طریق انتخاب توزیع‌های پیشین مناسب در مدل‌سازی لحاظ می‌گردند. توزیع پیشین مناسب برای ضرایب رگرسیونی $\beta_v(\gamma_v)$ به صورت توزیع پیشین q-prior Zellner's فرض شد که فرم آن به صورت ارایه شده در زیر است.

$$\beta_v(\gamma_v)|y_v, \sigma_v^2, \Sigma_{a_v}^{-1}, \gamma_v \sim N\left(\hat{\beta}_v(\gamma_v), T_v\sigma_v^2 \left[X_v^\top(\gamma_v)\Sigma_{a_v}^{-1}X_v^\top(\gamma_v)\right]^{-1}\right)$$

که در آن

$$\beta_v(\gamma_v) = \left[X_v^\top(\gamma_v)\Sigma_{a_v}^{-1}X_v^\top(\gamma_v)\right]^{-1} X_v^\top(\gamma_v)\Sigma_{a_v}^{-1}, y_v$$

توزیع پیشین مذکور بر اساس داده‌ها است ودر نتیجه می‌تواند بار محاسباتی را به میزان قابل توجهی کاهش دهد، که این امر در مطالعات fMRI اهمیتی زیادی دارد (زلنر ، ۱۹۸۶).

در نتیجه توزیع‌های پسین حاشیه‌ای با انتگرال‌گیری بر اساس پارامتر $\beta_v(\gamma_v)$ انجام گردید.از آنجایی که یکی از اهداف اصلی این پژوهش، تعیین پارامترهای تابع HRF مختص هر واکسل است، بنابراین پارامتر λ_v بر اساس داده‌ها برای هر واکسل برآورد گردید. طبق پیشنهاد کویرس و همکاران، توزیع یکنواخت (u_1, u_2) به عنوان توزیع پیشین برای پارامتر تاخیر λ_v در نظر گرفته شد

$$\lambda_v \sim U(u_1, u_2), \quad v = 1, \ldots, V$$

که در آن $u_1 = 0$ و $u_2 = T$ انتخاب‌های مناسبی هستند اگر اطلاعات پیشینی در مورد تاخیر پاسخ در وجود نداشته باشد (کروش، مونتس دیاز و گامرمن ، ۲۰۱۰).

در بخش همبستگی مکانی واحدهای نمونه، از توزیع پیشین فضایی Ising استفاده شد که دارای دو بخش است و علاوه بر اطلاعات فضایی، اطلاعات آناتومیکی را نیز لحاظ می‌کند (فنگ، لیانگ و تیرتی ، ۲۰۱۴).

همبستگی فضایی سه بعدی ذاتی بین واحدهای نمونه از طریق این مدل به خوبی لحاظ می‌گردد. فرم کلی مدل به صورت زیر است:

$$\pi(\gamma_{jv}|\theta_j) \propto C(\theta_{j0}\sum_{v=1}^{N}\gamma_{v_j} + \theta_{j1}\sum_{v\sim k}W_{v,k}I(\gamma_{v_j} = \gamma_{k_j}))$$

که $v \sim k$ نشان دهنده همسایه بودن دو واکسل است. پارامتر θ_{j1} نشان دهنده میزان شباهت همسایه‌ها به همدیگر است و در فاصله $(0, 1)$ قرار دارد. پارامتر θ_{j0} که field external نیز نامیده می‌شود میزان تمایل مجموعه به فعال بودن یا غیر فعال بودن را نشان می‌دهد. این پارامتر در فاصله $[-1, 1]$ است. پیشین برای پارامتر $\theta_j = (\theta_{j0}, \theta_{j1})$ به صورت زیر خواهد بود:

$$\pi(\theta) \propto \prod_{j=1}^{p} I(0 \leq \theta_j \leq \theta_{\max})$$

همبستگی فضایی سه بعدی ذاتی بین واحدهای نمونه از طریق این مدل به خوبی لحاظ می‌گردد. قابل ذکر است که عبارت $C(\theta_{j0}, \theta_{j1})^{-1}$ ثابت نرمال سازی مدل است و معمولا در طول محاسبات و برآورد پارامترها باعث مشکل می‌شود. برای برطرف کردن این مشکل از روش متغیر اضافه در برآورد مدل استفاده شد (مولر و همکاران ، ۲۰۰۶).

خوشه‌بندی سری‌های زمانی با استفاده از پیشین پروسه دیریکله بر روی پارامترهای ماتریس خطا انجام شد. چرا که یکی از اهداف مدل‌سازی خوشه‌بندی سری‌های زمانی بر اساس رفتار مشابه در طول زمان بود. پروسه دیریکله یکی از مهم ترین و پرکاربردترین روش‌های استفاده شده برای استنباط بیزی ناپارامتری در حیطه خوشه‌بندی است (فرگوسن ، ۱۹۷۳).

اگر پاسخ های $y_1, y_2, \ldots, y_n$ از آمیخته‌ای از توزیع‌های $F(\theta)$ به دست آمده باشد که توزیع‌های آمیخته روی θ با G نشان داده می‌شود. پیشین‌های DP با پارامتر تمرکز α و توزیع پایه G_0 می‌تواند به عنوان توزیع‌های آمیخته در نظر گرفته شود، بنابراین

$$y_i|\theta_i \sim F(\theta_i)$$

$$\theta_i|G \sim G$$

$$G \sim \text{DP}(G_0, \alpha)$$

در پژوهش حاضر یک پیشین دیریکله به فرم زیر روی پارامترهای ساختار خطای حافظه طولانی، σ_v^2 و α مدت به فرم زیر اعمال شد:

$$(\alpha_v, \sigma_v^2)|G \sim G$$

$$G|\eta, G_0 \sim \text{DP}(\eta, G_0)$$

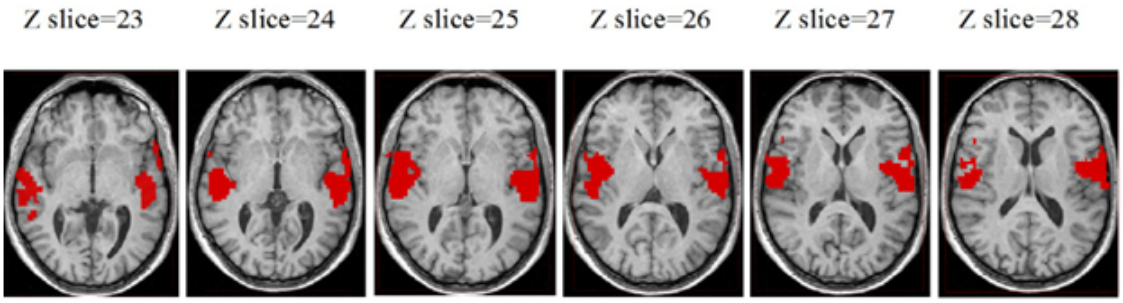
$$G_0 = Beta(b_0, b_1) \times IG(a_0, a_1)$$

که تابع پایه به صورت حاصل‌ضرب توزیع‌های بتا و معکوس گاما برای پارامترهای α و σ_v^2 است. قابل ذکر است با توجه به اینکه اثر پیشین روی η با افزایش حجم نمونه کاهش می‌یابد مقدار آن را از قبل ثابت در نظر می‌گیرند.

با در نظر گرفتن پیشین‌های مذکور و انجام استنباط پسین از طریق الگوریتم MCMC، برآورد پارامترهای مدل به دست آمد. پس از برازش مدل نتایج حاصل که همان شدت فعالیت در نقاط مختلف مغز و نیز خوشه‌بندی نواحی مغز بر اساس مشابهت رفتار سری‌های زمانی است، به صورت تصاویر مغزی ارایه شدند.

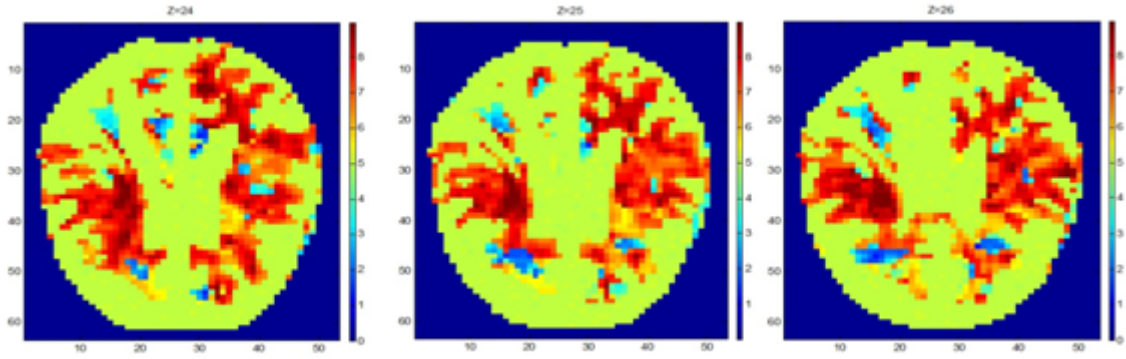
۳ نتایج

در این بخش از داده مرجع نرم افزار SPM استفاده شد، .http://www.fil.ion.ucl.ac.uk/spm طراحی مطالعه به صورت بلوکی و با یک محرک شنوایی بوده است. مطالعه به این صورت بوده است که ۴۲ ثانیه محرک شنوایی به فرد اعمال شده و ۴۲ ثانیه قطع شده است (در کل ۱۶ بلوک ۴۲ ثانیه‌ای). تعداد تصاویر مغزی ۸۴ عکس با فاصله ۷ ثانیه (طول سری زمانی ۸۴ است) و تعداد نمونه‌ها برابر ۲۶۲۱۴۴ (ابعاد تصاویر (۶۴ * ۶۴ * ۶۴ است. پس از استخراج سیگنال‌های سری زمانی از قالب تصاویر و هم چنین آماده سازی داده‌ها با انجام مراحل پیش پردازش، مدل مکانی_زمانی سه بعدی به تابع HRF ثابت (پواسون با پارامتر ۵) به داده‌ها برازش داده شد. نتایج حاصل برای گزارش، مجددا به صورت تصاویر ارایه گردید. قابل ذکر است بخشی از مغز که در دو طرف سر و مرتبط به محرک شنوایی است (لوب گیجگاهی) وارد تحلیل شد و در نتیجه نتایج در تصاویر مربوط به همان بخش‌ها گزارش شده است.



شکل ۲: نواحی فعال مغز بر اثر محرک شنوایی در داده‌های auditory برای چند برش مغزی.

در شکل ۲، چند برش افقی از مغز دیده می‌شود که در آن نواحی‌ای که مدل برای آنها مقدار پارامتر $\gamma_v = ۱$ برآورد کرده است با رنگ متفاوت مشخص شده‌اند. در واقع این نواحی رنگی، نقاط فعال شده بر اثر شنیدن محرک شنوایی بوده است. به عبارت دیگر نواحی رنگی و اکسل‌هایی از فضای مغز را نشان می‌دهد که مقدار پارامتر γ_v برای آنها ۱ برآورد شده است. همان‌طور که قبلا نیز گفته شد برآورد پارامتر $\gamma_v = ۱$ نشان دهنده فعال بودن آن واکسل است. شکل ۳ خوشه‌بندی نواحی مغز را در سه برش افقی نشان می‌دهد. در این تصویر نواحی که رنگ مشابه دارند نشان می‌دهند که سری‌های زمانی واکسل‌های آنها در طول زمان رفتار مشابه داشته‌اند.



شکل ۳: خوشه‌بندی نواحی مغز بر اساس مشابهت رفتار آنها در طول زمان که ارتباطات تابعی را نشان می‌دهد.

۴ بحث و نتیجه‌گیری

مدل سازی آماری داده‌های fMRI مساله‌ای نوین و رو به رشدی در سال‌های اخیر بوده است. همان طور که قبلا نیز اشاره شد این نوع داده ساختارهای همبستگی پیچیده‌ای دارد. به علاوه حجم بسیار زیاد مشاهدات و نیز ویژگی‌های خاصی که مختص این نوع داده است، نظیر تابع پاسخ همودینامیکی، سبب شده است که ملاحظات خاصی در مدل‌سازی آماری این داده‌ها مطرح شود (لی و همکاران ، ۲۰۱۴).

بنابراین لازم است که مدل‌سازی آماری به گونه‌ای انجام شود که تمامی این ویژگی‌ها تا حد ممکن لحاظ شود. طبیعت داده‌های fMRI به گونه‌ای است که و اکسل‌ها در فضایی سه بعدی نسبت به هم قرار دارند. بنابراین یکی از چالش‌های مهم در این حیطه مدل‌سازی تمام حجم مغز در فضای سه بعدی است. در نتیجه تعیین نقاط فعال مغز بر اثر یک محرک خاص و نیز ارتباطات بین نواحی مغز به صورتی که همبستگی سه بعدی لحاظ شود اهمیت زیادی دارد (ژانگ و همکاران ، ۲۰۱۵). در پژوهش حاضر یک مدل بیزی ناپارامتری ارایه داده شد که نواحی فعال مغز را تعیین کرده و به طور هم زمان ارتباطات تابعی بین نواحی را مشخص می‌کند. مدل ارایه شده بر اساس رویکرد بیزی است که می‌تواند تمامی ویژگی‌های ذکر شده در بالا را لحاظ کند. در مجموع می‌توان عملکرد مدل را چنین خلاصه کرد که برای هر واکسل فعال یا غیر فعال بودن آن بر اثر محرک تعیین می‌شود و هم چنین سری‌های زمانی پاسخ واکسل‌ها در یک فضای سه بعدی با استفاده از پیشین فرایند دیریکله روی ساختار خطای مدل خوشه‌بندی می‌شوند. به این صورت هدف تعیین ارتباط بین نواحی مغز توسط مدل انجام می‌شود. در طول پروسه برآورد پارامترها با روش MCMC تبدیل موجک گسسته از نوع Daubechies 4 به کار گرفته شد. تبدیل موجک گسسته برای سفید سازی داده‌های همبسته استفاده می‌شود(مک گیل و تاسول ، ۱۹۹۳؛ میر ، ۲۰۰۳؛ بولمور و همکاران ، ۲۰۰۴؛ وینک ، ۲۰۰۴).

در اکثر مطالعات تابع پاسخ همودینامیکی قبل از مدل‌سازی ثابت در نظر گرفته می‌شود ولی در مدل‌سازی مقاله حاضر به صورت مجزا برای هر واکسل برآورد گردید. در تعدادی از پژوهش‌های دیگر تابع پاسخ همودینامیکی بر اساس واکسل‌ها به طور مجزا برآورد شده است اما معمولا ساختار همبستگی‌ها به طور کامل لحاظ نشده‌اند. به علاوه رویکرد متغیر اضافه و الگوریتم Neal در پروسه MCMC بار محاسباتی ناشی از به کار بردن پیشین Ising در فضای سه بعدی را تا حدی کاهش داده است (نیل ، ۲۰۰۰).

هم چنین با استفاده از رویکرد متغیر اضافه نیازی به تقریب زدن ثابت‌ها نبوده و در طول فرآیند برآورد می‌شوند. این مدل در سناریوهای گوناگون با استفاده از داده‌های شبیه سازی ارزیابی شده است (نتایج در اینجا ارایه نشده است). در مجموع مدل عملکرد

خوبی در تعیین نقاط فعال و نیز خوشه‌بندی سری‌های زمانی داشته است. مدل هم چنین برای داده‌های واقعی مربوط به محرک شنوایی به کار گرفته شده و با موفقیت توانست نقاط فعال در کورتکس شنوایی را تعیین کند.

دیدگاه ارایه شده در مقاله حاضر می‌تواند گسترش یافته و اهداف دیگری را نیز دنبال کند. به عنوان مثال می‌توان پارامترهای مربوط به فعالیت را به گونه‌ای برآورد کرد که در طول مراحل تصویر برداری به طور وابسته به زمان تغییر کنند. به علاوه خوشه‌بندی واکسل‌ها می‌تواند با اعمال فرآید دیریکله روی بخش فضایی مدل انجام گیرد تا ناحیه بندی مغز بر اساس رفتار فضایی مشابه واکسل‌ها خوشه‌بندی شوند. تعمیم مدل برای استفاده در مورد چند نمونه به جای یک نفر زمینه دیگری برای گسترش پژوهش حاضر است. اگرچه بار محاسباتی ممکن است مشکل ساز باشد. البته رویکردهای دیگری در الگوریتم برآورد پرامترها ممکن است بتوانند مشکل بار محاسباتی را تقلیل دهند.

مراجع

Bullmore, E., Fadili, J., Maxim, V., Şendur, L., Whitcher, B., Suckling, J., Brammer, M., and Breakspear, M. (2004). Wavelets and functional magnetic resonance imaging of the human brain, *Neuroimage*, **23**, S234-49.

Feng, D., Liang, D., Tierney, L. (2014). A unified Bayesian hierarchical model for MRI tissue classification, *Statistics in medicine*, **33**(8),1349-68.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, 209-30.

Flandin, G., and Penny, W. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors, *Neuroimage*, **340**, 1108–1125.

Friston, K. J. (2011). Functional and effective connectivity, a review, *Brain Connectivity*, **10**, 13–36.

George, E. I., McCulloch, R. E. (1997). Approaches for Bayesian variable selection, *Statistica Sinica*, **70**, 339–373.

Gössl, C., Auer, D., and Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging, *Biometrics*, **570**, 554–562.

Handwerker, D. A. , Ollinger, J. M., and D'Esposito M. (2004). Variation of BOLDhemodynamic responses across subjects and brain regions and their effects on statisticalanal yses, *Neuroimage*, **21**, 1639–51.

Harrison, L. , Penny, W. , Daunizeau, J., and Friston, K. J. (2008). Diffusion-based spatial priors for functional magnetic resonance images, *Neuroimage*, **41**(2), 408–23.

Harrison, L., Penny, W., and Friston, K, J. (2003). Multivariate autoregressive modeling of fMRI time series, *Neuroimage*, **190**, 1477–1491.

Ko, K., and Vannucci, M. (2006). Bayesian wavelet analysis of autoregressive fractionally integrated moving-average processes, *Journal of Statistical Planning and Inference*, **136**(10), 3415-34.

Lee, K. J., Jones, G. L., Caffo, B. S., and Bassett S. S. (2014). Spatial Bayesian Variable Selection Models on Functional Magnetic Resonance Imaging Time-Series Data, *Bayesian Analysis*, **9**(3), 699–732.

Lu, Y., Grova, C., Kobayashi, E., Dubeau, F., and Gotman J. (2007). Using voxel-specific hemodynamic response function in EEG-fMRI data analysis, An estimation and detection model, *Neuroimage*, **34**(1), 195–203.

Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., and Poline, J. B. (2003). Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information, *Human Brain Mapping*. **19**(1), 1–17.

Meyer, F. G. (2003). Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series, *IEEE Trans Med Imaging*, **22**(3), 315–22.

McGill, K., and Taswell, C. (1993). Wavelet transform algorithms for finite-duration discrete-time signals, In *Proceedings of the International Conference on Wavelets and Applications*, Toulouse France, 221-224.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants, *Biometrika*, **93**(2), 451-8.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models, *Journal of computational and graphical statistics*, **9**(2), 249-65.

Penny, W., Trujillo-Barreto, N., and Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors, *Neuroimage*, **240**, 350–362.

Poldrack, R. A. (2011). *Handbook of Functional MRI Data Analysis*, Cambridge University Press.

Quirós, A., Montes Diez, R., and Gamerman, D. (2010). Bayesian spatiotemporal model of fMRI data, *Neuroimage*, **49**(1), 442–56.

Smith, M., Putz, B., Auer, D., and Fahrmeirc L. (2003). Assessing brain activity through spatial Bayesian variable selection, *Neuroimage*, **20**, 802–15.

Smith, M. , Utz, B. P. , Auer, D., and Fahrmeir, L. (2003). Assessing brain activity through spatial Bayesian variable selection, *Neuroimage*, **20**(2), 802–815.

Wink, A. M. (2004). *Wavelet-based methods for the analysis of fMRI time series*, University Library Groningen.

Woolrich M. W., Jenkinson, M., Brady, J. M., and Smith S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *Med Imaging, IEEE Trans*, **23**(2), 213–31.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Bayesian inference and decision techniques, *Essays in Honor of Bruno De Finetti*, **6**, 233-43.

Zhang, L., Guindani, M., and Vannucci, M. (2015). Bayesian models for functional magnetic resonance imaging data analysis, *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**, 21–41.

Zhang, L., Guindani, M., Versace, F., and Vannucci M. (2014). A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses, *Neuroimage*, **95**, 162–75.

Zhang, L., Guindani, M., Versace, F., Engelmann, J. M., and Vannucci, M. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data, *The Annals of Applied Statistics*, **10**(2), 638-66.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶

معرفی مدل‌های فضایی زمانی در تحلیل داده‌های fMRI

نسرین برومندنیا[۱]، حمید علوی مجد*[۱]، فرید زایری[۱]، احمد رضا باغستانی[۱]، سید محمد طباطبایی[۲] ، فریبرز فائقی[۳]

[۱]گروه آمار زیستی، دانشگاه علوم پزشکی شهید بهشتی

[۲]گروه انفورماتیک پزشکی، دانشگاه علوم پزشکی شهید بهشتی

[۳]گروه تکنولوژی پرتو شناسی، دانشگاه علوم پزشکی شهید بهشتی

چکیده: تصویربرداری عملکردی تشدید مغناطیسی، fMRI، نام نوعی روش در ام ار ای است که فعال شدن مغز بر اثر اعمال محرک‌ها را به طور غیر مستقیم اندازه می‌گیرد. داده‌های fMRI دارای ساختار پیچیده همبستگی زمانی و مکانی سه بعدی است. به علاوه رفتار پاسخ همودینامیکی در این مطالعات اهمیت ویژه‌ای دارد. مطالعات fMRI در سال‌های اخیر بسیار مورد توجه قرار گرفته و گسترش یافته است. از طرف دیگر روش‌های آماری نقشی اساسی در تحلیل این نوع داده‌ها ایفا می‌کنند. زیرا داده‌های fMRI دارای ویژگی‌های خاصی هستند که در نظر گرفتن این ویژگی‌ها در مدل‌سازی آماری اهمیت به سزایی دارد. ساختار پیچیده داده‌ها، حجم بسیار زیاد آن‌ها و نیز تعداد قابل توجه پارامترهای مدل در این نوع مطالعات، باعث شده که رویکردهای بیزی سهم قابل توجهی از روش‌های آماری حیطه fMRI را شامل شوند. با توجه به این که این نوع رویکردهای تحلیل داده‌ها هنوز در مجلات علمی فارسی جایگاه واقعی خود را نیافته است، در این مقاله تلاش بر این است که مروری بر رویکردهای مختلف بیزی آمار در تحلیل این نوع داده‌ها ارایه شود. در راستای این هدف ابتدا به بیان انواع مطالعات fMRI و ساختار داده‌ها پرداخته و در ادامه رویکردهای آماری موجود در هر زمینه معرفی می‌گردد.

واژه‌های کلیدی: مطالعه fMRI، تابع پاسخ همودینامیکی، همبستگی زمانی و فضایی سه بعدی، رویکردهای آمار بیزی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 91B72.

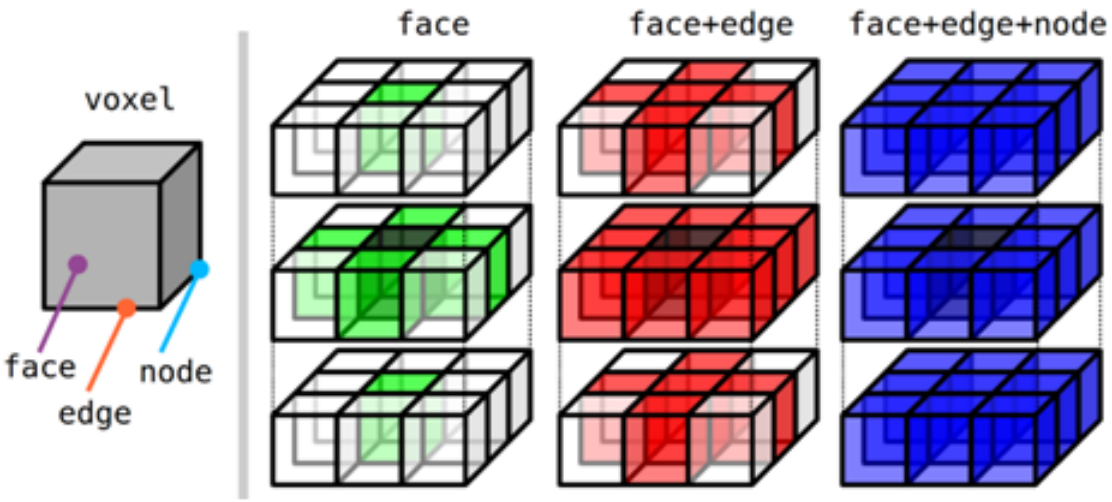
---

*ارایه‌دهنده مقاله: حمید علوی مجد، alavimajd@gmail.com

۱ معرفی داده‌های fMRI و ویژگی‌های آنها

نحوه اجرای یک مطالعه fMRI به این صورت است که تصاویری متناوب از مغز انسان در حال فعالیت و نیز در حال استراحت گرفته می‌شود. زمانی که یک محرک[۱] خاص بر فرد اعمال می‌شود برخی نواحی مغز در پاسخ به آن فعال شده و در نتیجه سطح اکسیژن خون در آن ناحیه افزایش می‌یابد. هر عامل بیرونی مثل شنیدن یک موسیقی، دیدن یک تصویر، حرکت یک عضو از بدن، بیاد آوردن یک تصویر، استراحت و ... یک محرک نامیده می‌شود. متغیری که در این گونه مطالعات ثبت می‌شود میزان اختلاف خون اکسیژن‌دار و خون فاقد اکسیژن در همه نواحی مغز هست که متغیری کمی است و با نام BOLD[۲] شناخته می شود (پولدارک ، ۲۰۱۱).

ثبت پاسخ از طریق گرفتن تصاویر مغزی صورت می گیرد. اسکن‌های سه بعدی از حجم مغز در طول زمان گرفته می‌شود، در حالی که فرد داخل دستگاه اسکنر MRI دراز کشیده است و محرک‌هابر وی اعمال می‌شود. اسکن‌ها هر چند ثانیه تکرار می‌شوند. یک آزمایش fMRI حتی اگر تنها بر روی یک فرد انجام گیرد ده‌ها و یا صدها اسکن به دست می‌دهد. در واقع هر اسکن یک حجم سه بعدی است که این حجم از هزاران مکعب کوچک به نام واکسل[۳] تشکیل شده است. این مکعب‌های کوچک که حاصل تقاطع قطعات برش[۴] مغزی در سه جهت است، در واقع واحدهای نمونه برای مدل‌سازی آماری هستند.برش‌های مغز در سه جهت، از یک گوش به گوش دیگر، از عقب سر به جلوی سر و نیز از گردن به بالا در نظر گرفته می‌شوند. مختصات فضایی هر واکسل با توجه به این که حاصل تقاطع برش‌های شماره چند است، مشخص می‌شود. در نتیجه واحدهای نمونه در همسایگی هم در فضایی سه بعدی قرار دارند. تجربه نشان داده‌هاست که واکسل‌های همسایه رفتارهای مشابهی از نظر تغییرات متغیر پاسخ نشان می‌دهند. در نتیجه در نظر گرفتن این همبستگی فضایی سه بعدی در مدل‌سازی آماری ضروری است.می‌توان در شکل زیر انواع مختلفی از همسایگی‌ها را ملاحظه نمود. نحوه قرار گرفتن واکسل‌ها در فضای سه بعدی نشان می‌دهد که اگر تحلیل آماری در سطح واکسل، فقط واکسل‌های همسایه در یک سطح دو بعدی را شامل شود، همبستگی مکانی به طور کامل لحاظ نخواهد شد.



شکل ۱: نمایی از همسایگی واکسل‌ها در فضای سه بعدی که در مدل‌سازی آماری باید لحاظ شوند.

[۱] Task
[۲] Blood Oxygen Level Dependent
[۳] Voxel
[۴] Slice

همان طور که اشاره شد از فرد چندین اسکن در طول زمان ثبت می‌شود که این نشان دهنده ماهیت طولی بودن متغیر پاسخ است. در واقع میزان پاسخ در هر واحد نمونه به صورت یک سری زمانی ثبت می‌شود. در نتیجه داده‌های حاصل سری‌های زمانی خواهند بود که به صورت مکانی در فضایی سه بعدی همبستگی دارند.

هر حجم مغز عکسبرداری شده در حدود چندین هزار واکسل است.در هر آزمایش بیش از صدها بار از حجم یک مغز در طول زمان عکس‌برداری می‌شود. هر آزمایش fMRI معمولا روی چندین نفر (به طور مثال ۱۰ تا ۴۰ نفر) تکرار می‌شود. البته تحقیقاتی که در حیطه fMRI اجرا می‌شود قابلیت اجرا بر روی یک نفر و یا چند نفر را دارند. باید توجه شود که مدل‌سازی آماری برای مطالعات چند نمونه ای متفاوت از مدل‌سازی برای مطالعه تک‌نمونه‌ای است. جمع‌بندی مطالب فوق نشان دهنده این مطلب است که در مطالعات fMRI مجموعه‌ای از صدها هزار سری زمانی باید در مدل‌سازی آماری وارد شوند که به طور فضایی نیز وابستگی دارند.

مطالعات fMRI از نظر نحوه طراحی به دو دسته تقسیم می‌شوند. دسته اول که طرح‌های بلوکی نام دارد به گونه‌ای طراحی می‌شوند که محرک‌هابه طور منظم و با فواصل زمانی مساوی به فرد اعمال می‌گردد. در این طرح دورهزمانی اعمال همه محرک‌ها یکسان است. در طرح‌های دسته دوم که به عنوان طرح‌های وابسته به پیامد شناخته شده اند، محرک‌ها به طور تصادفی و نامنظم و البته دوره‌های نامساوی روی فرد تحت آزمایش اعمال می‌شود. ماتریس طرح مرتبط با هر یک از طرح‌های مذکور باید در مدل‌سازی مربوطه لحاظ گردد.

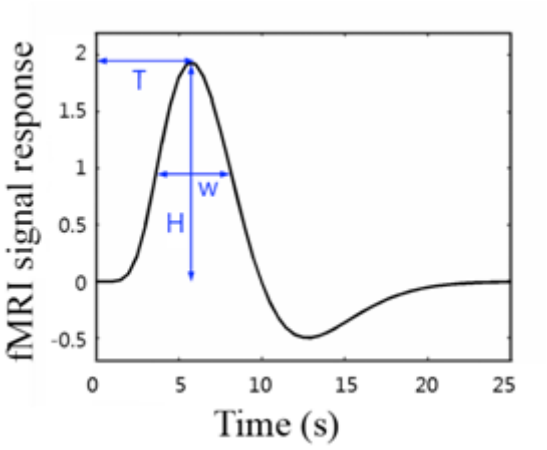
در اینجا لازم است ویژگی کلیدی دیگری از داده‌های fMRI که در مدل‌سازی آماری باید مدنظر قرار گیرد، بیان شود. هنگامی که تغییر در میزان BOLD به علت یک محرک بیرونی اندازه‌گیری می‌شود، در واقع سیگنال‌های پاسخ با تاخیر دریافت می‌شوند. چنین پاسخی به عنوان تابع پاسخ همودینامیکی[۵] یا به طور مختصر HRF شناخته می‌شود. به بیان ساده‌تر می‌توان گفت سیگنال‌های پاسخ پس از دریافت محرک بلافاصله افزایش نمی‌یابند و پس از قطع آن نیز به سرعت به مقدار پایه اولیه باز نمی‌گردند. در واقع پروسهتغییر میزان پاسخ در هر واکسل دوره‌ای حدود ۲۰ ثانیه به طول می انجامد (پولدارک و همکاران ، ۲۰۱۱). بنابراین در هر نقطه از زمان، پاسخ ثبت شده در واقع برآیند پاسخ‌های زمان‌های قبلی است. شکل کلی تابع HRF در تصویر شماره ۲ نشان داده شده است. اگرچه ثابت فرض کردن پارامترهای تابع HRF در اکثر مدل‌سازی‌های آماری رایج است اما در واقعیت فرم تابعی آن به ازای محرک‌های گوناگون، در افراد گوناگون و نواحی مختلف مغز تغییر می‌کند. بنابراین در نظر گرفتن یک فرم واحد برای همه اعضای نمونه فرضیه درستی به نظر نمی‌رسد (هندورکر و همکاران ، ۲۰۰۴؛ ماررلک و همکاران ، ۲۰۰۳؛ اشبرنر و همکاران ، ۲۰۰۳).

دیدگاه‌های گوناگونی برای لحاظ کردن فرم تابع پاسخ در مدل‌سازی اماری وجود دارد. اما رایج ترین آنها این است که پس از پیچش[۶] ماتریس طرح با تابع HRF، آن را وارد مدل آماری می کنند. به عبارت دیگر هر ستون از ماتریس طرح X_v برای واکسل vامبه این صورت مدل‌بندی می‌شود:

$$\int_{\cdot}^{t} x(s)h_v(t-S)ds$$

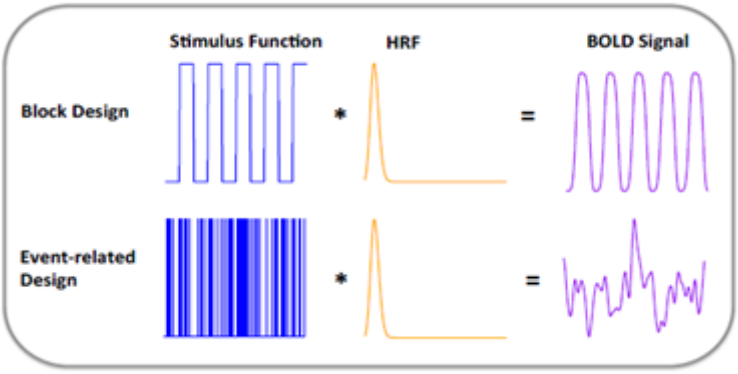
که در آن $h_v(t)$ تابع HRF است و $x(s)$ تابع محرک وابسته به زمان (برای هر محرک خاص) می‌باشد که متناظر با برداری از مقادیر صفر و ۱ است. مقادیر این بردار در زمان‌هایی که محرک خاص اعمال شود ۱ و در غیر این صورت صفر است. نمایی گرافیکی از

[۵] Hemodynamic Response Function
[۶] Convolution



شکل ۲: نحوه تغییر فرم تابع پاسخ همودینامیکی در اثر محرک

مدل‌سازی برای داده‌های fMRI در شکل شماره ۳ نشان داده شده است که به ازای نوع طراحی مطالعه، طرح بلوکی و طرح وابسته به پیشامد، تفکیک شده است. علاوه بر همبستگی زمانی و فضایی و مبحث HRF، موضوع دیگری که در داده‌های fMRI مطرح



شکل ۳: مدل بندی مقادیر پاسخ BOLD به صورت ترکیبی از ماتریس طرح آزمایش و تابع HRF

می‌شود مساله انتخاب واکسل‌های فعال است. انتخاب واکسل فعال در این مطالعات با عنوان مساله انتخاب متغیر[۷] شناخته می‌شود (جرج و مک کولوچ ، ۱۹۹۷). قابل ذکر است که این عنوان با مفهوم رایج انتخاب متغیر در مطالعات آماری، متفاوت است. چرا که در مطالعات fMRI مساله انتخاب متغیر معادل تعیین واکسل‌هایی است که ضریب مدل رگرسیونی برای آن‌ها غیر صفر است. پس از مروری بر مقدمات لازم، باید توجه کرد که روش‌های آماری نقش مهمی در تحلیل داده‌های fMRI دارند (بومن ، ۲۰۱۴؛ لازار ، ۲۰۰۸؛ لیندکوئیست ، ۲۰۰۸).

دلیل این اهمیت فراوان، حجم عظیم داده‌ها و نیز ساختار پیچیده همبستگی‌های مکانی و زمانی آنها است. دیدگاه‌های اولیه برای تحلیل این گونه داده‌ها در سطح واکسل انجام می‌گرفته است. به عنوان مثال محاسبه آزمون تی و تحلیل واریانس در هر واکسل و یا برازش یک مدل GLM در هر واکسل به طور جداگانه صورت می‌گرفته است. با این قبیل تحلیل‌ها، ویژگی‌های خاص داده‌های

[۷] Variable selection problem

fMRI نادیده گرفته می‌شود. بنابراین ارایهتحلیل در سطح واکسل مناسب نیست. به عنوان مثال آماره‌های آزمون تی در بین همه واکسل‌ها مستقل از هم نیستند. زیرا بین واکسل‌های همسایه وابستگی مکانی سه بعدی وجود دارد.

رویکردهای آماری بیزی برای تحلیل داده‌های fMRI تاکنون از جنبه‌های گوناگون نظیر نحوه لحاظ کردن همبستگی مکانی‌ـ زمانی و مساله انتخاب متغیر، توسعه یافته اند. قابلیت انعطاف پذیری بالای روش‌های بیزی در مدل‌سازی، این امکان را فراهم می‌آورد که مدل‌های بیز قادر باشند همبستگی مکانی و زمانی داده‌های fMRI را لحاظ کنند (ولریش ، ۲۰۱۲). نتایج موفقیت آمیزی بکارگیری مدل‌های بیزی سلسله مراتبی در تحلیل داده‌های fMRI قابل توجه است. در مقاله حاضر نیز رویکردهای بیزی در حیطه آمار در fMRI مرور می‌شوند.

۲ اهداف یک مطالعه fMRI

در اجرای هر مطالعه fMRI یکی از اهداف سه گانه زیر مطرح است که مطالعه جهت پاسخ به آن سوال طرح‌ریزی می‌شود. قابل ذکر است که در ادامه تحلیل‌های آماری مربوطه بر طبق اهداف مورد نظر از مطالعه fMRI دسته‌بندی می‌شوند.

۱.۲ هدف اول: یافتن نواحی فعال مغز

در این گونه مطالعات fMRI، در حالی که فرد مجموعه ای از محرک‌ها را انجام می‌دهد تمام مغز وی در چند نقطه زمانی اسکن می‌شود. هدف یافتن نواحی از مغز است که با دریافت محرک‌های خاص فعال می‌شود. مدل‌هایی که در این دسته قرار دارند به مدل‌های همبستگی مکان‌ـ زمان شناخته می‌شوند که برای یافتن نواحی فعال مغز کاربرد دارند. این مدل‌ها به صورت خطی، غیر خطی و نیز مدل‌های آمیخته هستند (ژانگ و همکاران ، ۲۰۱۵). در ادامه صرفا جهت آشنایی اولیه با مدل‌سازی در سطح واکسل یک مدل خطی تعمیم یافته در حالت کلی ارایه می شود. رایج‌ترین دیدگاه اماری برای یک سری زمانی از پاسخ‌های BOLD مدل خطی تعمیم یافته (GLM) است که اولین بار در داده‌های fMRI توسط فریستون به کار گرفته شده است (فریستون و همکاران ، ۱۹۹۴). مدل خطی تعمیم‌یافته در سطح واحدهای نمونه، به صورت زیر تعریف می‌شود:

$$y_v = X_v\beta_v + \varepsilon_v, \varepsilon_v \; N_{T_v}(\cdot, sv_v^{\cdot}\Lambda_v)$$

که در آن Y_v سری زمانی پاسخ‌های ثبت شده برای نمونه‌ی v ام که $v = ۱, ۲, \ldots$ و N تعداد کل نمونه‌ها است، X_v ماتریس طرح به ابعاد $T \times P$، که P تعداد محرک‌های مورد آزمایش و T طول سری زمانی (یعنی تعداد اسکن‌ها در طول زمان) است و حاصل آمیختن تابع HRF با ماتریس طرح است. $\beta_v = (\beta_{v,۱}, \ldots, \beta_{v,p})$ بردار ضرایب رگرسیونی به ابعاد $P \times ۱$ است و ε_v بردار خطای $T \times ۱$ در مدل رگرسیونی است.

در مطالعات fMRI تعیین وضعیت فعال شدن یا نشدن واحدهای نمونه، یا واکسل‌ها، در قبال اعمال محرک مربوطه مد نظر است. در رویکردهای بیزی، برای در نظر گرفتن مساله انتخاب متغیر، کلاسی از توزیع‌های پیشین بر روی ضرایب رگرسیونی اتخاذ شده اند. بردار متغیرهای نشانگر $\gamma_v = (v_{v,۱}, \ldots, \gamma_{v,p})^\top$ را در نظر بگیرید که دنباله‌ای از صفر و یک‌هاست که به ترتیب نشان دهنده پذیرفته شدن و یا رد شدن فرض آماری $\beta_{v,j} = ۰$ برای هر نمونه است. بنابراین وقتی $\gamma_{v,j} = ۰$ باشد ضریب $\beta_{v,j}$ برابر با صفر خواهد شد واگر $\gamma_{v,j} = ۱$ باشد ضریب رگرسیونی مربوطه غیر صفر است. در نتیجه بازنویسی مدل قبل به این صورت است:

$$y_v = X_v(\gamma_v)\beta_v(\gamma_v) + \varepsilon_v, \;\; \varepsilon_v \sim N_{T_v}(\cdot, \sigma_v^{\cdot}\Lambda_v)$$

در کل می‌توان گفت این مدل‌ها پاسخ‌های BOLD را به عنوان ترکیبی از الگوی ماتریس طرح محرک‌هابا تابع HRF مدل‌بندی می‌کنند.

همان طور که قبلا اشاره شد برآورد تعداد بسیار زیادی پارامتر در این مدل‌سازی مدنظر است. لذا توجه ما به رویکردهای بیزی خواهد بود. در رویکردهای بیزی با قرار دادن پیشین‌های مناسب در مدل فوق، ویژگی‌های مد نظر داده‌های fMRI در مدل‌سازی وارد می‌شود. برای لحاظ کردن همبستگی زمانی بسیاری از مدل‌ها این نوع از همبستگی را از طریق مدل‌سازی مستقیم عبارات خطا در مدل لحاظ می‌کنند. رایج ترین دیدگاه‌ها استفاده از ساختار خطای خودهمبستگی از مراتب گوناگون و نیز ساختار خطای حافظه طولانی مدت به همراه تبدیل موجک است. به علاوه وارد کردن تابع HRF نیز بخشی از همبستگی زمانی داده‌ها را لحاظ می‌کند. شکل HRF می‌تواند به صورت توابعی چون پواسون[۸]، کانونیک[۹] و معکوس لوجیت[۱۰] در نظر گرفته شود (ژانگ و همکاران ، ۲۰۱۴؛ ورسلی و همکاران ، ۲۰۰۲؛ لیندکوئیست و واگر ، ۲۰۰۷). همچنین پارامترهای این توابع می‌توانند از قبل ثابت در نظر گرفته شوند و یا بر اساس داده‌ها برای هر واحد نمونه مجزا برآورد شوند (ژانگ و همکاران ، ۲۰۱۴؛ کروش و همکاران ، ۲۰۱۰؛ ولریش و همکاران ، ۲۰۰۴).

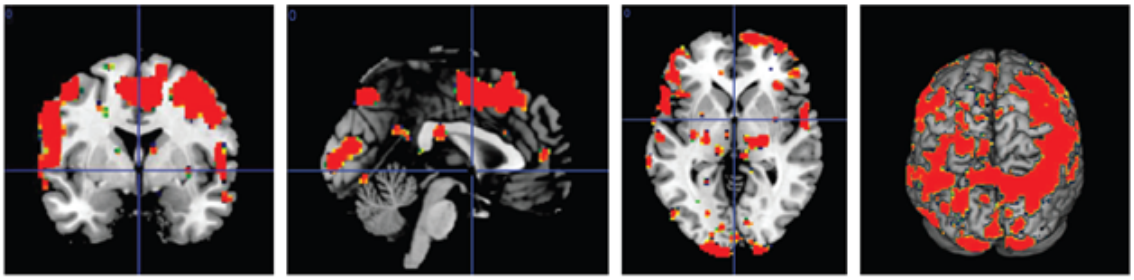
در بخش همبستگی مکانی، پیشین‌هایی که برای لحاظ کردن همبستگی مکانی کاربرد دارند شامل توزیع‌های زیر هستند: توزیع فضایی Ising که پیشین مناسب برای ضرایب فعال بودن در نظر گرفته می‌شود و هم چنین توزیع‌های slab and spike و GMRF[۱۱] می‌توانند به عنوان توزیع پیشین برای ضرایب رگرسیون درنظر گرفته شوند (لی و همکاران ، ۲۰۱۴؛ براون و همکاران ، ۱۹۹۸؛ کروش و همکاران ، ۲۰۱۰). پیشین‌های دیگری نظیر پیشین فضایی diffusion-based نیز هستند که بر روی ضرایب رگرسیونی به کار گرفته شده اند. برخی دیگر ساختار خود همبستگی شرطی را به کار برده‌اند (هریسون و گیرین ، ۲۰۱۰). به علاوه پیشین‌های SSBF[۱۲] نیز در این زمینه کاربرد داشته‌اند(فلاندین و پنی ، ۲۰۰۷).

حجم زیاد داده‌ها، تعداد بسیار زیاد پرامترها و زمان بر بودن پروسه‌های برآورد باعث می‌شود که انتخاب الگوریتم برآوردی اهمیت یابد. روش‌های MCMC Component-wize و نیز Bayes Variational از رویکردهای پر کاربرد در مدل‌سازی‌های بیزی fMRI هستند که البته هر کدام مزیت و معایبی خود را دارند. در واقع با توجه به هدف مطالعه، برآورد پارامترهای شدت فعالیت هر واکسل (در واقع ضریب رگرسیونی) و نیز احتمال فعال بودن واکسل مربوطه مد نظر است. در مورد تعیین وضعیت فعال بودن واکسل می‌توان به این نکته اشاره کرد که نقاط برش تعیین شده ای وجود دارند که اگر احتمال فعالیت برای واکسل مربوطه بیشتر از آن بود واکسل به عنوان فعال تشخیص داده شود. مقادیر گوناگونی برای نقطه برش وجود دارند که از جمله مقدار ۰/۸۷۲۲ را می‌توان نام برد(اسمیت و فاهرمیر ، ۲۰۰۷) .

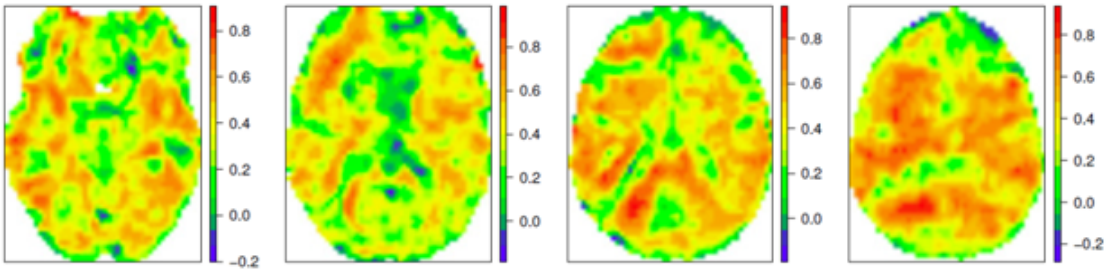
نحوه گزارش خروجی در این مطالعات از طریق ارایه برآورد پارامترهای مدل در قالب تصاویر است. به عنوان نمونه برآورد ضریب رگرسیونی، یا همان میزان شدت فعالیت، برای هر واکسل به صورت شدت رنگ گزارش می‌شود. نقاطی که رنگ روشن‌تر داشته باشند مناطقی هستند که فعالیت داشته اند. برای گزارش نقاط فعال نیز پس از اعمال نقطه برش بر روی احتمالات فعال بودن، وضعیت‌های فعال و غیر فعال با دو رنگ در قالب تصویر برای واکسل‌ها نشان داده می‌شود. به عنوان مثال نمونه ای از خروجی‌های مربوطه در شکل شماره ۴ ارایه شده‌اند که نقشه فعالیت مغزی برای چند برش از مغز پس از گذاشتن نقطه برش در احتمالات فعال

[۸] Poisson
[۹] Canonical
[۱۰] Inverse logit
[۱۱] Gaussian Markov Random Field
[۱۲] Sparse Spatial Basis Function

بودن، گزارش شده است. در شکل شماره ۵ نیز برآورد پارامتر خودهمبستگی مربوط به هر واکسل درچند برش افقی مغز به صورت شدت رنگ در قالب تصاویر آمده است.



شکل ۴: نقشه فعالیت مغزی در اثر محرک مد نظر مطالعه، بر اساس نقطه برش ۰/۸۷۲۲ (لی و همکاران ، ۲۰۱۴)



شکل ۵: برآورد پارامتر خود همبستگی برای چند برش افقی از مغز (بزنر ، ۲۰۱۷)

نقشه‌های فضایی که از فعالیت نواحی مغز و یا پارامترهای برآورد شده در استنباط پسین به دست می آیند با عنوان نقشه‌های احتمالات پسین[۱۳] شناخته می شوند (ژانگ و همکاران ، ۲۰۱۵).

۲.۲ هدف دوم: یافتن ارتباطات بین نواحی مغز

این مطالعات با هدف بررسی این که چطور نواحی مغز با هم ارتباط دارند و چطور اطلاعات بین آنها منتقل می‌شود، انجام می‌شوند. مدل‌های آماری در این حوزه بر اساس دو نوع ارتباط تابعی و موثر، به دو گروه تقسیم شده اند (فریستون ، ۱۹۹۴).

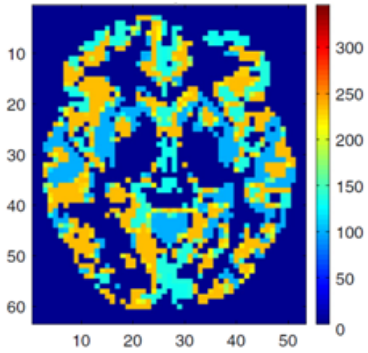
دسته اول مطالعات ارتباطی از نوع ارتباطات تابعی[۱۴] است. هدف از مطالعات ارتباطات تابعی، یافتن نواحی از مغز است که در حال دریافت یک محرک خاص فعالیت مشابه نشان می‌دهند. به بیان ساده می‌توان گفت ارتباطات تابعی در مطالعات fMRI تقریبا با استفاده از ضرایب همبستگی تعیین می‌شوند. این همبستگی‌ها بر اساس مقادیر کواریانس پاسخ بین نواحی مغز است. ساده ترین راه برای جمع‌بندی الگوهای همبستگی بین نواحی، گزارش بردارهای مقادیر ویژه و یا مولفه‌های اصلی است. برای دست‌یابی به هدف ارتباطات تابعی، روش‌های خوشه بندی و هم چنین روش‌های چند متغیره برای کاهش بعد، مورد استفاده قرار گرفته اند.

[۱۳] Posterior probability maps
[۱۴] Functional connectivity

از جمله این روش‌ها می‌توان به روش‌های تحلیل مولفه‌های اصلی[۱۵]، تحلیل مولفه‌های مستقل[۱۶] و روش گرافیکی لاسو[۱۷] نام برد (آندرسون و همکاران ، ۱۹۹۹؛ کالهون و همکاران ، ۲۰۰۱؛ مک نون و همکاران ، ۱۹۹۸؛ کریین و همکاران ، ۲۰۱۲؛ واروکو و همکاران ، ۲۰۱۰).

به علاوه مدل‌های رگرسیونی که می‌توانند به خوشه‌بندی واکسل‌ها بر اساس شباهت رفتار سری زمانی آنها بپردازند نیز در این حیطه قرار می‌گیرند. به عنوان مثال با کاربرد فرایند دیریکله[۱۸] به عنوان پیشین در یک مدل مکان_زمان با رویکرد بیزی، امکان خوشه‌بندی سری‌های زمانی فراهم می‌آید (ژانگ و همکاران ، ۲۰۱۴).

خروجی حاصل از این مدل بندی‌های آماری نیز مجددا در قالب تصاویر ارایه می‌شوند. به عنوان مثال در شکل ۶ نمونه‌ای از خروجی مربوط به ارتباطات تابعی در یک مطالعه fMRI گزارش شده است. در این تصاویر مناطقی از مغز که در طول زمان در مقابل محرک مد نظر عملکرد یکسان داشته‌اند و یا به عبارت دیگر یک خوشه را تشکیل داشته اند، هم رنگ نشان داده شده اند.

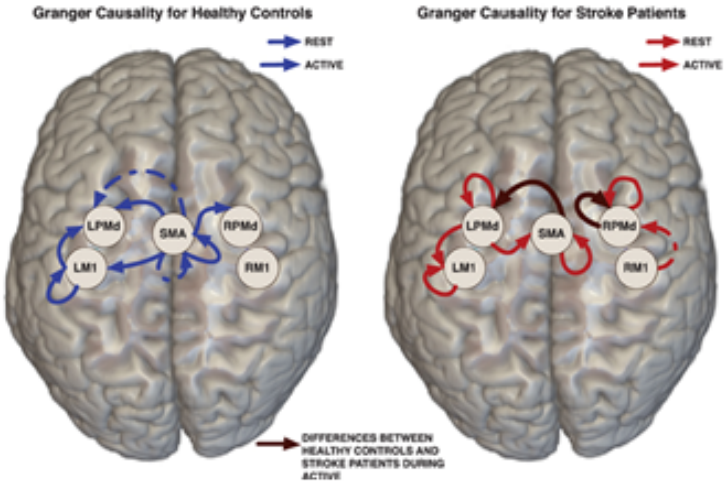


شکل ۶: ارتباطات تابعی: خوشه بندی نواحی که سری زمانی آن‌ها عملکرد یکسان داشته‌اند (ژانگ و همکاران ، ۲۰۱۴)

دسته دوم مطالعات ارتباطی، ارتباطات موثر[۱۹] نام دارند. مطالعاتی که به بررسی ارتباطات موثر می‌پردازند اثر یک ناحیه از مغز روی ناحیه دیگر، در حین دریافت محرک خاص، را مورد بررسی قرار می‌دهند.تاکنون تغییرات در ارتباطات موثر بین نواحی مغز در بیماری‌های گوناگون بررسی شده است. به عنوان مثال مطالعات نشان داده‌اند که ارتباط بین نواحی مختلف مغز در بیمارانی نظیر اتیسم، پارکینگسون و HIV با افراد سالم متفاوت است. با توجه به اهمیت ارتباطات موثر بین نواحی مغز از نظر بالینی، توجه به روش‌های آماری در این زمینه ضروری است. از جمله روش‌های آماری که در زمینه‌های ارتباطات موثر کاربرد دارند می‌توان از روش‌هایی مانند مدل‌سازی معادلات ساختاری[۲۰] (SEM) ،مدل سازی علیتی دینامیکی[۲۱]، GC[۲۲]، مدل‌های خودهمبستگی برداری[۲۳] و شبکه‌های بیزی[۲۴] نام برد(بوچل و فریستون ، ۱۹۹۷؛ مکل نوتش و گونازلز لیما ، ۱۹۹۴؛ فریستون و همکاران ، ۲۰۰۳؛ هریسون و همکاران ، ۲۰۰۳؛ ژنگ و راجاپاکسه ، ۲۰۰۶؛ گوبل و همکاران ، ۲۰۰۳) .

[۱۵] Principal components
[۱۶] Independent components analysis
[۱۷] Graphical Lasso
[۱۸] Dirichlet process
[۱۹] Effective connectivity
[۲۰] Structural Equation Modeling
[۲۱] Dynamic causal modeling
[۲۲] Granger causality
[۲۳] Vector auto-regressive models
[۲۴] Bayesian networks

نتایج مطالعات ارتباطی می‌توانند به صورت ارتباطات بین نواحی مغز بر روی تصویر نشان داده شوند. به عنوان مثال در شکل ۷ نمونه ای از این نوع ارتباطات موثر در افراد سالم و افراد بیمار نشان داده شده است. تفاوت ایجاد شده در نوع ارتباطات موثر بین نواحی مغز در افراد بیمار قابل مشاهده است. برای توضیح بیشتر و آشنایی با حیطه ارتباطات تابعی و موثر مقالات مروری کاملی در این زمینه موجود است (قریسون و همکاران ، ۲۰۱۱؛ استفان و روئربک ، ۲۰۱۲).



شکل ۷: ارتباطات موثر: تفاوت در ارتباطات موثر بین نواحی مغز در افراد سالم (چپ) و بیماران ضربه دیده (راست) (گوروشتیتا و همکاران ، ۲۰۱۳)

۳.۲ هدف سوم:پیش گویی و طبقه بندی

دسته سوم از مدل‌های آماری در زمینه مطالعات fMRI، به هدف پیشگویی فعالیت مغزی یک فرد و یا پیشگویی پاسخ رفتاری و یا بالینی فرد اختصاص دارند. به عنوان مثال پیشبینی میزان درد یک نفر که به طور مستقیم قابل اندازه‌گیری نبوده و می‌توان آن را از فعالیت مغزی وی تعیین کرد. در برخی دیگر از این مطالعات ممکن است پیش بینی فعالیت مغز پس از دریافت داروی خاص مورد نظر محقق باشد تا بتواند در موارد بعدی برای هر بیمار داروی مختص او را تجویز کند.

مدل‌سازی‌های آماری در این حیطه کمتر از دو بخش قبل گسترش یافته اند. در جهت پیش بینی فعالیت مغزی بر اثر مصرف یک دارو مدل‌های پیشنهادی به صورت مدل‌های خطی تعمیم یافته بوده‌اند که در یک رویکرد دو مرحله ای به مدل بندی همزمان اسکن‌های قبل از دریافت دارو و اسکن‌های پس از دریافت دارو پرداخته اند (درادو و همکاران ، ۲۰۱۳؛ گو و همکاران ، ۲۰۰۸). برای رسیدن به هدف پیش بینی پاسخ رفتاری یا بالینی، مدل‌های رگرسیون خطی غالبا مورد استفاده قرار گرفته اند. به عنوان مثال مدل رگرسیون بیزی پراکنده چند کلاسه [۲۵] را می‌توان نام برد (میچل و همکاران ، ۲۰۱۱). مدل رگرسیون با استفاده از توزیع‌های پیشین لاپلاس نیز در این زمینه به کار گرفته شده‌اند (فان گرون و همکاران ، ۲۰۱۰).

[۲۵] Multiclasssparse Bayesian regression

۳ بحث و نتیجه گیری

در سال‌های اخیر پیشرفت چشم گیر مطالعات fMRI و اهمیت آنها در تشخیص و درمان بیمارها، نقش مهم روش‌های آماری را در تحلیل این نوع مطالعات را مشخص می‌کند. از طرفی خلا موجود در مقالات آماری فارسی، اهمیت پرداختن به این موضوع را روشن می‌سازد. همان طور که اشاره شد یک مطالعه fMRI از نظر داده‌های آماری، ساختاری خاص و متفاوت با سایر مطالعات دارد. وجود همبستگی‌های توام زمانی و مکانی سه بعدی و نیز ماهیت خاص تابع پاسخ همودینامیکی، باعث شده است که روش‌های آماری در قالبی متفاوت با دیگر مطالعات، اینجا کاربرد داشته باشند. به عبارت دیگر حیطه‌آمار در fMRI، انواع مدل‌سازی‌ها و رویکردهای مختلف آمار را شامل می‌شود، منتها به صورتی که منطبق با ساختار خاص این نوع داده‌ها شوند.

ذکر این نکته نیز در اینجا مفید است که تحلیل‌های متا آنالیز در مطالعات fMRI اهمیت و کاربرد زیادی دارند. مطالعات تک نمونه ای در fMRI با هدفهای کمکی قبل از جراحی، کاربرد زیادی دارند. به علاوه علم پزشکی به سمت پیش بینی‌های فردی و درمان‌های مختص به فرد می رود. بنابراین به علت وجود مطالعات تک نمونه‌ای، تحلیل متاآنالیز در این شاخه کاربرد فراوانی دارد.

برخی از نرم افزارهایی که در زمینه تحلیل آماری داده‌های fMRI کاربرد دارند SPM ، FSL و Matlab هستند که آزمون‌های رایج مانند آزمون تی و تحلیل واریانس در کنار مدل‌های رگرسیونی در این نرم افزارها گنجانده شده اند. ولی نکته قابل توجه این است که ورودی و خروجی این نرم افزارها تصاویر مغزی هستند و تحلیل‌های آماری که توسط نرم افزار اجرا می‌شوند همانند دیگر نرم افزارها نمود ظاهری ندارند.شاید این امر سبب شده است که عموما این تحلیل‌ها توسط محققین آماری ناشناخته باقی مانده است. به علاوه در سایر شاخه‌های آمار، نرم افزارهای گوناگونی معرفی شده‌اند که حتی مدل‌های پیچیده را پوشش می‌دهند، اما در حیطه آمار در fMRI این گونه نیست.

علاوه بر تصاویر مغزی fMRI، انواع دیگری از داده‌های مغزی در دهه اخیر در دنیای پزشکی مورد توجه قرار گرفته اند. داده‌هایی که ساختار متفاوت و مختص به خود را دارند و زمینه را برای ارایه نوآوری‌های مدل‌سازی آماری فراهم می‌کنند. از جمله می‌توان داده‌های EEG و DTI را نام برد که ساختار آن ها و ویژگی‌هایی متفاوت با داده‌های fMRI دارند. قابل ذکر است که این داده‌ها به صورت مدل‌های توام [۲۶]

با داده‌های fMRI نیز مدل بندی می‌شوند. نکته مهم دیگر این است که مدل‌سازی‌های توام داده‌های fMRI و داده‌های ژنتیکی نیز از اهمیت زیادی برخوردارند. به طوری که پروژه‌های جهانی برای جمع آوری داده‌های ژنتیکی و fMRI راه اندازی شده‌اند و داده‌های آنها از طریق اینترنت https://openfmri.org/ در اختیار همه محققان قرار گرفته است. از جمله پروژه‌های Connectome Human [۲۷] و My Connectome [۲۸] را می‌توان در این زمینه معرفی کرد. پرداختن به مطالعات در این سطح گسترده و اشتراک بانک داده‌های جهانی در این حیطه، اهمیت تحقیق و مدل‌سازی‌های جدید آماری که منطبق با ساختار داده‌های جدید باشند، را روشن می‌سازد. بدیهی است که با توجه به جدید بودن موضوع، زمینه برای پیشرفت‌های مدل‌سازی و تحقیقات آماری در این حیطه از علوم زیستی فراهم است.

[۲۶] Joint modelling

[۲۷] http://www.humanconnectomeproject.org/

[۲۸] http://myconnectome.org/wp/

مراجع

Andersen, A. H., Gash, D. M., and Avison, M. J. (1999). Principal component analysis of the dynamic response measured by fMRI: a generalized linear systems framework. *Magnetic Resonance Imaging*, **170**, 795–815.

Ashburner, J., Friston, K. J., and Penny, W., (2003). *Human Brain Function*, 2nd Edition, Academic Press.

Bezener, M. A. (2017). *Bayesian Spatiotemporal Modeling using Hierarchical Spatial Priors with Applications to Functional Magnetic Resonance Imaging* (Doctoral dissertation, university of minnesota).

Bowman, F. D. (2014). Brain imaging analysis, *Annual review of statistics and its application*, 1:61.

Brown, P. .J, Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B*, **600**, 627–641.

Büchel, C., and Friston, K. J. (1997). Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, **70**, 768–778.

Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J .J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, **140**, 140–151.

Cribben, I., Haraldsdottir, R., Atlas, L. Y., Wager, T. D. , and Lindquist, M. A. (2012). Dynamic connectivity regression: determining state-related changes in brain connectivity. *Neuroimage*, **61**, 907–920.

Derado, G., Bowman, F. D., and Zhang ,L. (2013). Predicting brain activity using a Bayesian spatial model. *Statistical Methods in Medical Research*, **220**, 382–397.

Flandin, G., and Penny, W. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *Neuroimage*, **340**, 1108–1125.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis, *Human Brain Mapping*, **2**, 56–78.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain connectivity*, **1**(1), 13-36.

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, **190**, 1273–1302.

Friston, K. J., Jezzard, P., and Turner, R. (1994). Analysis of functional MRI time-series, *Human Brain Mapping*, **1**(2), 153–71.

Goebel, R., Roebroeck, A., Kim, D., and Formisano E. (2003). Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. *Magnetic Resonance Imaging*, **210**, 1251–1261.

George, E. I., and McCulloch, R. E. (1997). for Bayesian variable selection, *Statistica Sinica*, **70**, 339–373.

Gorrostieta, C., Fiecas, M., Ombao, H., Burke, E., and Cramer, S. (2013). Hierarchical vector auto-regressive models and their applications to multi-subject effective connectivity. *Frontiers in computational neuroscience*, **7**, 159.

Guo, Y., Bowman, F. D., and Kilts C. (2008). Predicting the brain response to treatment using a Bayesian hierarchical model with application to a study of schizophrenia. *Hum Brain Mapp*, **290**, 1092–1109.

Handwerker, D. A. , Ollinger, J. M., and D'Esposito, M. (2004). Variation of BOLDhemodynamic responses across subjects and brain regions and their effects on statisticalanal yses, *Neuroimage*, **21**, 1639–51.

Harrison, L. M., and Green, G. G. R. (2010). A Bayesian spatiotemporal model for very large data sets. *Neuroimage*, Elsevier Inc., **50**(3):1126–41.

Harrison, L., Penny, W., and Friston, K. J. (2003). Multivariate autoregressive modeling of fMRI time series. *Neuroimage*, **190**, 1477–1491.

Lazar, N. (2008). *The statistical analysis of functional MRI data*, Springer Science and Business Media.

Lee, K. J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014). Spatial Bayesian Variable Selection Models on Functional Magnetic Resonance Imaging Time-Series Data. *Bayesian Analysis*, **9**(3), 699–732.

Lindquist, M. A. (2008). The statistical analysis of fMRI data, *Statistical Science*, **23**(4), 439-64.

Lindquist, M. A., and Tor Wager, D. (2007). Validity and power in hemodynamic response modeling: a comparison study and a new approach, *Human Brain Mapping*, **28**, 764–784.

Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., and Poline, J. B. (2003). Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information, *Human Brain Mapping*, **19**(1), 1–17.

McKeown, M., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A., and Sejnowski, T. (1998). Analysis of fMRI data by blind separation into independent spatial components *Human Brain Mapping*, **6**, 160–188.

Mclntosh, A., and Gonzalez-Lima, F. (1994). Structural equation modeling and its application to network analysis in functional brain imaging. *Human Brain Mapping*, **20**, 2–22.

Michel, V., Eger, E., Keribin, C., Thirion, B. (2011). Multiclass sparse Bayesian regression for fMRI-based prediction. *Journal of Biomedical Imaging*, **2011**, 2.

Poldrack, R. A. (2011). *Handbook of Functional MRI Data Analysis*, Cambridge University Press.

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of functional MRI data analysis*, Cambridge University Press.

Quirós, A., Diez, R. M., and Gamerman, D. (2010). Bayesian spatiotemporal model of fMRI data. *Neuroimage*, **49**(1), 442–56.

Quirós, A., Diez, R. M., and Wilson, S. P. (2010). Bayesian spatiotemporal model of fMRI data using transfer functions, *Neuroimage* [Internet]. Elsevier Inc., **52**(3), 995–1004.

Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging, *Journal of the American Statistical Association*, **102**, 417–31.

Smith, A. M., Lewis, B. K., Ruttimann, U. E., Frank, Q. Y., Sinnwell, T. M., Yang, Y., Duyn, J. H., and Frank, J. A. (1999). Investigation of low frequency drift in fMRI signal, *Neuroimage*, **9**(5), 526-33.

Stephan, K. E., and Roebroeck, A. (2012). A short history of causal modeling of fMRI data. *Neuroimage*. Elsevier Inc., **62**(2), 856–63.

Van Gerven, M. A. , Cseke, B., De Lange, F. P., and Heskes, T. (2010). Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*. **50**(1), 150-61.

Varoquaux, G., Gramfort, A., Poline, J. B., Thirion, B., Zemel, R. and Shawe-Taylor, J. (2010). Brain covariance selection: better individual functional connectivity models using population prior. In: Zemel R, Shawe-Taylor J, eds, *Advances in Neural Information Processing System*.

Woolrich, M. W. (2012). Bayesian inference in FMRI. *Neuroimage*. Elsevier Inc., **62**(2), 801–10.

Woolrich, M. W., Jenkinson, M., Brady, J. M., and Smith SM. (2004). Fully Bayesian spatio-temporal modeling for fMRI data, *IEEE Trans Med Imaging*, **23**, 213–31.

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G.H., Morales, F., and Evans, A.C. (2002). A general statistical analysis for fMRI data, *Neuroimage*, **15**(1), 1-5.

Zhang, L., Guindani, M., and Vannucci, M. (2015). Bayesian models for functional magnetic resonance imaging data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**, 21–41.

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014). A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses, *Neuroimage*, **95**, 162–75.

Zheng, X.,and Rajapakse, J. C. (2006). Learning functional structure from fMR images. *Neuroimage*, **310**, 1601–1613.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1990

مقایسه عملکرد کریگیدن با شبکه‌های عصبی مصنوعی در پیش‌گویی فضایی

عباس توسلی*، یدالله واقعی

گروه آمار، دانشگاه بیرجند

چکیده: هدف از این مقاله بررسی کاربرد شبکه‌های عصبی مصنوعی[1] (ANN) برای پیشگویی فضایی است. یک ویژگی مهم در مدلهای شبکه عصبی، توانایی آنها در یادگیری از داده‌ها است، حتی زمانی که کاربر اطلاعات زیادی در مورد مجموعه داده‌ی مورد نظر ندارد. از سوی دیگر، یکی از مهمترین روشهای زمین‌آماری برای پیشگویی فضایی، یعنی کریگیدن، با اینکه عملکرد بسیار قابل قبولی دارد، اما برای برازش مدل همبستگی (نیم‌تغییرنگار) تا حدی به تخصص نیاز دارد. در این مقاله با استفاده از داده‌های فضایی شبیه‌سازی شده توانایی پیشگویی این دو روش را مقایسه خواهیم کرد. نتایج نشان می دهد که عملکرد روش زمین‌آماری کریگیدن نسبت به روش شبکه عصبی بهتر می‌باشد. با اینحال هرچند که دقت پیشگویی شبکه عصبی برای مجموعه داده‌های شبیه‌سازی شده بهتر از کریگیدن نیست، اما نقشه پهنه‌بندی داده‌ها نشان می‌دهد که شبکه عصبی می‌تواند رقیب خوبی برای پیشگوی کریگیدن باشد.

واژه‌های کلیدی: رگرسیون_ کریگیدن، شبکه عصبی_ کریگیدن،پیش‌گویی فضایی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62G08، 62J02، 62M45.

۱ مقدمه

پیشگویی فضایی در رشته‌هایی مانند زمین‌شناسی، هیدرولوژی، هواشناسی، معدن و غیره بسیار مفید است. این مسئله با استفاده از روشهایی مانند کریگیدن، وزن‌دهی معکوس فاصله، چند جمله‌ای‌های درونیاب، اسپیلانها، و غیره مورد بررسی قرار گرفته است. علاوه بر این، طیف وسیعی از مطالعات به مقایسه این روشها برای یافتن مناسب‌ترین روش در زمینه‌های مختلف منجر شده است. از بین این روش‌ها، کریگیدن که مبتنی بر مدلی پیوسته از تغییرات فضایی تصادفی است گزینه‌ای مناسب برای پیش‌گویی خواهد بود (وب استر و الیور ، ۲۰۰۷). در بین روش‌های کریگیدن، کریگیدن ساده (SK) ، کریگیدن معمولی (OK)، کریگیدن عام (UK)،

*ارایه‌دهنده مقاله: عباس توسلی، tavassoli.a@birjand.ac.ir

[1] Artificial Neural Network

کریگیدن با روند بیرونی (KED)، رگرسیون‌ـ کریگیدن (RK) و کوکریگیدن (CK) به‌طور گسترده استفاده شده است (کرسی ۱۹۹۸، زیمرمن و همکاران ۱۹۹۹،لی و همکاران ۲۰۱۰).

هرچند که کریگیدن نتایج قابل قبولی ارائه می‌دهد، اما نیاز به یک مدل برای همبستگی مکانی بین نمونه‌ها بصورت یک تغییرنگار دارد و مدلسازی تغییرنگار نیز نیاز به برخی از اطلاعات تخصصی از رویه‌های زمین‌آماری و آشنایی با مجموعه داده‌ها دارد. بنابراین، این فرآیند مدل‌سازی ممکن است مانعی برای افراد غیر متخصص در حوزه زمین‌آمار باشد.

در سال‌های اخیر شبکه‌های عصبی مصنوعی در شاخه‌های مختلف مهندسی و آمار بطور گسترده مورد استفاده قرار گرفته است. در مهندسی برای تشخیص الگو و طبقه‌بندی داده‌ها و در علوم و بویژه در مباحث آماری برای پیش‌گویی رگرسیونی و سریهای زمانی و حتی داده‌های فضایی از شبکه‌های عصبی مصنوعی استفاده شده است (پالیوال ، ۲۰۰۹). مدل‌های شبکه عصبی (NN) توانایی یادگیری از داده‌ها، حتی زمانی که کاربر اطلاعات زیادی در مورد مجموعه داده خاص نمی‌داند را دارد. برای مروری کلی از این تکنیک‌ها، مشکلات محاسباتی و تفاسیر آماری آن می‌توانید به بیشاپ (۱۹۹۵) و ریپلی (۲۰۰۷) رجوع کنید. شبکه عصبی مصنوعی بر خلاف روش‌های آمار فضایی به مفاهیم، روش‌ها و مدل‌های آمار فضایی (برقراری مانایی، برآورد و مدل‌سازی تغییرنگار و ...) نیاز ندارد. از طرفی، شبکه عصبی ساختارها و پارامترهای مختلفی دارد که می‌تواند بر سرعت و دقت یادگیری و در نتیجه بر دقت پیش‌گوییها تاثیرگذار باشد. با این‌حال، در بعضی مطالعات نشان داده شده است که شبکه‌های عصبی مصنوعی در مسائل پیش‌گویی فضایی عملکرد خوبی ندارند (نوتیپیلوا ۲۰۱۴، گوموس و سن ۲۰۱۳).

استفاده از شبکه عصبی در مواردی که با داده‌های وابسته سر وکار داریم، با توجه به مشکلاتی که در تشخیص و مدل‌بندی ساختار وابستگی داده‌ها به وجود می‌آید از اهمیت ویژه‌ای برخوردار است. تاکنون تکنیک‌های متفاوتی از طراحی و آموزش شبکه توصیف شده‌اند تا پیش‌گویی‌هایی از فرآیندهای فضایی و زمانی که تا حد ممکن دقیق هستند به‌دست آیند که از میان آنها می‌توان به کوایک و همکاران (۲۰۰۱) اشاره کرد.

از بین روش‌های شبکه عصبی مصنوعی، مدل‌های ترکیبی که مبتنی بر استفاده‌ی همزمان از ANN و روش‌های زمین‌آمار است نیز برای مدل‌بندی روندهای فضایی غیرخطی و کریگیدن باقیمانده بسط داده شده‌اند. برای جزییات بیشتر می‌توانید مطالعات کانوسکی و همکاران (۱۹۹۶)؛ یه و همکاران (۲۰۱۳)؛ چانگ (۲۰۱۲) را ببینید.

در این مقاله به کمک یک مطالعه شبیه‌سازی دقت چهار روش پیش‌گویی شامل روش کریگیدن عام، روش ترکیبی رگرسیون‌ـ کریگیدن، شبکه عصبی و روش شبکه عصبی‌ـ کریگیدن با یکدیگر مقایسه شده‌اند.

۲ پیش‌گویی فضایی با کریگیدن

کریگیدن یک روش پیش‌گویی در آمار فضایی و شامل مجموعه‌ای از فنون رگرسیون خطی است که وابستگی فضایی بین نمونه‌ها را در نظر می‌گیرد. این وابستگی جزء ماهیت ذاتی میدان تصادفی است. این روش هم برای داده‌های فضایی با موقعیت نقطه‌ای منظم و هم نامنظم قابل استفاده است. کریگیدن بر حسب نوع داده‌ها به دو صورت نقطه‌ای و بلوکی تقسیم می‌شود. در کریگیدن نقطه‌ای موقعیت مشاهدات نقاطی در D (یک زیرمجموعه از فضای اقلیدسی d بعدی) هستند و پیش‌گویی نیز در نقاطی فاقد مشاهده در D صورت می‌گیرد. اما گاهی داده‌ها در موقعیت‌های نقطه‌ای مشاهده نشده‌اند، بلکه در بلوک‌هایی از D که مختصات مرکز آنها معلوم است قرار دارند. کریگیدن بلوکی مقدار متوسط در بلوکهای فاقد مشاهده را پیش‌گویی می‌کند . فرض کنید پیش‌گویی $Z(s_0)$

بر اساس مشاهدات $Z(s_1),\dots,Z(s_n)$ در موقعیت‌های $s_1,\dots,s_n$ مورد نظر باشد. در روش کریگیدن یک ترکیب خطی از مشاهدات به‌صورت $\hat{Z}(s_0)=\sum_{i=1}^{n} w_i Z(s_i)+k$ را به عنوان پیش‌گوی $Z(s_0)$ در نظر گرفته و ضرایب w_i به نحوی برآورد می‌شوند که این پیش‌گو ناریب بوده و میانگین توان‌های دوم خطای آن کمترین باشد. در این روش معمولا به مشاهدات نزدیکتر وزن بیشتر و به مشاهدات دورتر وزن کمتری داده می‌شود. به‌علاوه روش کریکیدن دارای این ویژگی است که در هر موقعیت فضایی واریانس پیشگو را می‌توان حساب کرد. بسته به اینکه میانگین میدان تصادفی $Z(s)$ معلوم، ثابت نامعلوم یا تابعی از موقعیت‌ها باشد، کریگیدن به انواع کریگیدن ساده، عادی و عام تفکیک می‌شود.

۱.۲ کریگیدن عادی

کریگیدن عادی[۲] توسط ماترون (۱۹۷۱) برای پیش‌گویی مقدار میدان تصادفی در موقعیت مشخص s_0 تحت شرایط زیر معرفی شده است:

- میانگین میدان تصادفی ثابت (و نامعلوم) و در نتیجه میدان تصادفی به‌صورت $Z(s)=\mu+\delta(s)$ می‌باشد که در آن $\mu\in R$ نامعلوم و $\delta(s)$ میدان تصادفی مانای ذاتی با میانگین صفر است.
- پیش‌گویی در نقطه s_0 یک تابع خطی از مشاهدات به شکل $\hat{Z}(s_0)=\sum_{i=1}^{n} w_i Z(s_i)$ با قید $\sum_{i=1}^{n} w_i=1$ (که ناریبی پیشگو را تضمین می‌کند) است که در آن $Z(s_i)$ متغیرهای فضایی و w_iها ضرایب کریگیدن می‌باشند. این ضرایب به‌گونه‌ای به‌دست می‌آیند که میانگین توان دوم خطای پیشگو تحت قید ناریبی $\sum_{i=1}^{n} w_i=1$ مینیمم شود.

ضرایب بهینه به‌صورت زیر به‌دست می‌آیند (کرسی ، ۱۹۹۳).

$$\mathbf{w}^{\top}=\left(\gamma+\mathbf{1}_n\frac{\left(1-\mathbf{1}_n^{\top}\Gamma^{-1}\gamma\right)}{\mathbf{1}_n^{\top}\Gamma^{-1}\mathbf{1}_n}\right)^{\top}\Gamma^{-1} \tag{۱.۲}$$

که در این رابطه $\gamma=\left(\gamma\left(s_0-s_1\right),\dots,\gamma\left(s_0-s_n\right)\right)^{\top}$ و Γ یک ماتریس $n\times n$ است که عنصر (i,j) آن برابر $\gamma\left(s_i-s_j\right)=\frac{1}{2}\mathrm{Var}\left(Z(s_i)-Z(s_j)\right)$ است. لازم به ذکر است که در این رابطه با تغییر موقعیت s_0 بردار γ تغییر کرده و در نتیجه بردار ضرایب w نیز تغییر می‌کند.

۲.۲ کریگیدن عام

در کریگیدن عادی فرض بر این است که میانگین میدان تصادفی ثابت یعنی مستقل از موقعیت s است. ولی حالت‌های زیادی وجود دارند که میانگین میدان تصادفی ثابت نیست. یعنی داده‌های فضایی دارای روند هستند. کریگیدن عام[۳] توسط ماترون (۱۹۶۹) به عنوان پیش‌گوی ناریب برای وقتی که روند، یک ترکیب خطی با ضرایب مجهول از توابع معلوم (مانند مدل‌های رگرسیونی بر حسب موقعیت‌های x و y) است معرفی شد. در این روش نیز مانند کریگیدن عادی ضرایب w به گونه‌ای تعیین می‌شوند که اولا پیش‌گویی $\hat{Z}(s_0)=\sum_{i=1}^{n} w_i Z(s_i)$ ناریب و ثانیا دارای کمترین واریانس باشد. با این تفاوت که در اینجا $\mu(s)$ به‌صورت یک مدل رگرسیونی بر حسب مولفه‌های s مدلسازی می‌شوند. ضرایب به‌صورت زیر برآورد می‌شوند (کرسی ، ۱۹۹۳ صفحه ۱۵۳):

$$\mathbf{w}^{\top}=\left(\gamma+X\left(X^{\top}\Gamma^{-1}X\right)'\left(x_0-X^{\top}\Gamma^{-1}\gamma\right)\right)^{\top}\Gamma^{-1} \tag{۲.۲}$$

[۲] Ordinary Kriging (OK)
[۳] Universal Kriging (UK)

در اینجا X ماتریسی شبیه ماتریس طرح رگرسیون است که با توجه به روند ساخته می‌شود.

۳.۲ رگرسیون‌ـ کریگیدن

اگر داده‌های فضایی دارای روند باشند و برای مدلسازی روند، مقادیر یک یا چند متغیر کمکی را داشته باشیم بجای استفاده از کریگیدن عام از روشهای کوکریگیدن، کریگیدن با روند خارجی (KED) و رگرسیون‌ـ کریگیدن (RK) استفاده می‌شود. RK ترکیبی از رگرسیون و کریگیدن می‌باشد. در این روش، روند موجود در داده‌ها با یک مدل رگرسیونی بین متغیر اصلی و متغیرهای کمکی محاسبه شده و سپس باقیمانده‌های این مدل رگرسیونی (جزء بدون روند) با کمک یک روش زمین‌آماری درونیابی و به مقادیر روند اضافه می‌شود. بعنوان مثال رابطه بین متغیر اصلی و بقیه متغیرهای کمکی به کمک رگرسیون خطی چندگانه محاسبه شده و سپس با اعمال روش کریگیدن بر باقیمانده‌ها و جمع مقادیر دو روش، نتیجه نهایی حاصل می‌شود. یکی از مزایای RK نسبت به KED، برآورد مجزای روند از پیش‌گویی فضایی باقیمانده‌هاست، که اجازه استفاده از مدل‌های پیچیده رگرسیونی را (بر خلاف KED که فقط می‌تواند تکنیک‌های ساده خطی را استفاده کند) به ما می‌دهد. علاوه بر این دو قست مدل می‌تواند به‌طور مجزا تفسیر شود.

بر اساس پیشنهاد ماترون (۱۹۶۹) مقدار متغیر اصلی در موقعیت s می‌تواند به‌صورت مجموع دو مولفه قطعی (μ) و تصادفی (δ) مدل‌بندی شود:

$$Z(s) = \mu(s) + \delta(s)$$

در RK هر دو مولفه قطعی و تصادفیِ تغییرات فضایی می‌تواند به‌صورت جداگانه برآورد شده و نتیجه نهایی از مجموع این دو مقدار به‌دست آید(هنگل و همکاران ، ۲۰۰۳):

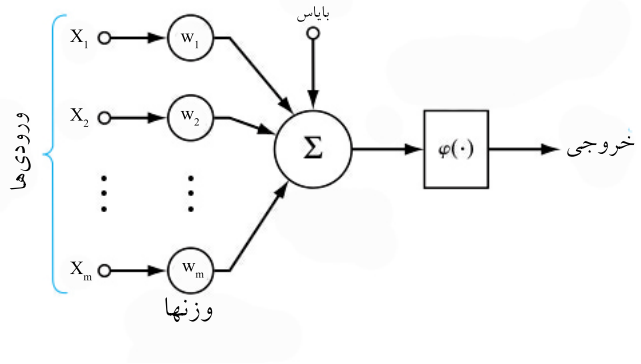
$$\begin{aligned} \hat{Z}_{RK}(s_0) &= \hat{\mu}(s_0) + \hat{\delta}(s_0) \\ &= \sum_{k=0}^{p} \hat{\beta}_k q_k(s_0) + \sum_{i=1}^{n} w_i(s_0) e(s_i); \quad q_0(s_0) = 1 \end{aligned} \tag{۳.۲}$$

به‌طوری‌که $\hat{Z}_{RK}(s_0)$ مقدار پیش‌گویی شده در مکان s_0 است. $\hat{\mu}(s_0)$ و $\hat{\delta}(s_0)$ به ترتیب برآورد قسمت قطعی (روند) و تصادفی مدل، $\hat{\beta}_k$ ضرایب برآورد شده مدل روند، $q_k(s_0)$ مقدار پیش‌گوها در مکان s_0، $w_i(s_0)$ وزن‌های کریگیدن و $e(s_i)$ باقیمانده‌های رگرسیونی در مکان s_i هستند. در این مدل، ضرایب β_k با روش حداقل مربعات تعمیم‌یافته (GLS) برآورد می‌شوند (کرسی ، ۱۹۹۳).

۳ شبکه عصبی مصنوعی

یک شبکه عصبی مصنوعی ایده‌ای است برای پردازش اطلاعات که از سیستم عصبی زیستی الهام گرفته شده است. این سیستم از شمار زیادی عناصر پردازشی به نام نرون (واحد) تشکیل شده است که برای حل یک مسأله با هم هماهنگ عمل می‌کنند. در شکل ۱ یک نرون مصنوعی با تعداد m ورودی نشان داده شده است. ورودی‌ها ابتدا در وزن‌های مربوط به خود ضرب می‌شوند که این وزن‌ها حکم سیناپس‌ها در یک نرون طبیعی را دارند. سپس این ورودی‌های وزن‌دار شده با هم جمع شده و با یک مقدار آستانه مقایسه می‌شوند.

اجزای یک شبکه عصبی عبارتند از: ورودی‌ها، وزن‌ها، تابع جمع، تابع تبدیل و خروجی (پاسخ مسئله). ورودی‌ها می‌توانند خروجی سایر لایه‌ها بوده و یا آنکه به حالت خام در اولین لایه باشد. وزن‌ها که در حقیقت پارامترهای شبکه هستند میزان تأثیر ورودی بر خروجی را در هر نرون مشخص می‌کنند. تنظیم وزن‌ها از طریق یک الگوریتم یادگیری[۴] انجام می‌گیرد. تابع تبدیل عضوی ضروری در شبکه‌های عصبی محسوب می‌گردد. انواع و اقسام متفاوتی از توابع تبدیل وجود دارد که بنا به ماهیت مسئله کاربرد دارند. این تابع توسط طراح مسئله انتخاب می‌گردد و بر اساس انتخاب الگوریتم یادگیری، پارامترهای مسئله (وزن‌ها) تنظیم می‌گردد. این فرآیند که طی آن وزن‌های شبکه برآورد می‌شوند را آموزش شبکه می‌نامند.



شکل ۱: نمونه‌ای از یک نرون با m ورودی

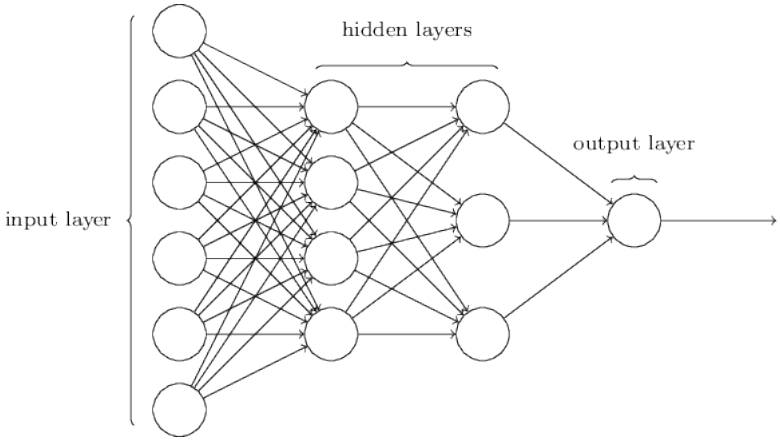
۱.۳ شبکه پرسپترون

پرسپترون ساده‌ترین شکل شبکه عصبی است که برای طبقه‌بندی الگوهایی که *به‌صورت خطی جداشدنی*[۵] هستند مورد استفاده قرار می‌گیرد. این شبکه یک تک نرون با وزن‌های سیناپسی قابل تنظیم را شامل می‌شود، به‌طوریکه الگوریتم مورد استفاده برای تنظیم پارامترهای آن، اولین بار توسط رزنبلات (۱۹۵۸؛ ۱۹۶۲) در قالب یک فرآیند یادگیری مطرح و همگرایی آن (معروف به قضیه همگرایی پرسپترون) اثبات شد. تعمیم این شبکه، شبکه پرسپترون چندلایه[۶] (MLP) می‌باشد که قادر است مسائل پیچیده و متفاوتی را توسط یک الگوریتم پرکاربرد به نام الگوریتم پس‌انتشار خطا[۷] حل کند. این شبکه از تعداد زیادی واحد پردازشگر که در چندین لایه (شامل یک لایه ورودی، یک لایه خروجی و چندین لایه مخفی) سازماندهی شده‌اند تشکیل یافته است. شکل ۲ نمونه‌ای از یک شبکه MLP با دو لایه مخفی را نشان می‌دهد.

با در نظر گرفتن شبکه‌ای با تعداد n واحد در لایه ورودی، یک لایه پنهان شامل m واحد و تنها یک خروجی، مدل MLP به‌صورت زیر فرمول‌بندی می‌شود:

$$y = O(x) = \psi\left(\sum_{j=1}^{m} w_j \phi\left(\sum_{i=1}^{n} w_{ij} x_i + w_{0j}\right) + w_0\right)$$

[۴] Learning algorithm
[۵] Linearly separable
[۶] Multilayer perceptron
[۷] Error back-propagation algorithm



شکل ۲: نمونه‌ای از یک شبکه MLP با دو لایه مخفی و ۶ ورودی

در این فرمول، ϕ تابع فعال‌سازی[۸] برای واحدهای لایه مخفی است که معمولا از توابع زیگموید و تانژانت هایپربولیک استفاده می‌شود. همچنین ψ تابع فعال‌سازی برای لایه خروجی است که برای مسایل رگرسیونی معمولا از توابع خطی یا زیگموید استفاده می‌شود. همچنین $w_{\cdot}$ مقدار ثابت برای واحد خروجی، $w_{\cdot j}$ مقدار ثابت برای واحد jام در لایه مخفی، w_{ij} وزن شبکه بین ورودی iام و واحد jام در لایه مخفی، w_j وزن شبکه بین واحد jام در لایه مخفی و خروجی شبکه و در نهایت x_i مقدار ورودی iام است.

انتخاب مدل برای شبکه‌های MLP بگونه‌ای خواهد بود که حداکثر ظرفیت تعمیم‌پذیری به‌دست آید. بدین‌منظور مجموعه داده‌ی مورد نظر به دو مجموعه داده آموزشی[۹] و مجموعه داده‌ی آزمایشی[۱۰] تقسیم می‌شود، به‌طوریکه از مجموعه داده‌ی آموزشی به منظور آموزش و ساخت شبکه و از مجموعه داده آزمایشی برای بررسی عملکرد شبکه استفاده می‌شود. از بین تکنیک‌های فراوانی که برای آموزش شبکه عصبی پیشنهاد شده است، در این مقاله از روشی که توسط فورس و هاگان (۱۹۹۷) اجرا شده، استفاده گردیده است.

برای آشنایی بیشتر با جنبه‌های ریاضی، آماری و محاسباتی شبکه‌های عصبی MLP می‌توانید به هایکین (۱۹۹۸) و بیشاپ (۱۹۹۵) رجوع کنید.

۴ مطالعه شبیه‌سازی

به منظور مقایسه عملکرد شبکه عصبی با روش‌های زمین‌آمار، چهار روش را مقایسه خواهیم کرد که عبارتند از: کریگیدن عام (OK) ، رگرسیون_کریگیدن (RK)، شبکه عصبی (NN) و شبکه عصبی_کریگیدن (NNK) به‌طوری‌که NN استفاده‌ی مستقیم از شبکه عصبی و NNK ترکیب شبکه با کریگیدن (همانند روش RK) می‌باشد. لازم به ذکر است که روش به‌کار رفته برای NNK مشابه RK می‌باشد با این تفاوت که برای محاسبه $\hat{\mu}(s)$ در رابطه ۳.۲ به‌جای مدل رگرسیونی از شبکه عصبی استفاده شده است.

برای مقایسه روش‌های مذکور از داده‌های فضایی شبیه‌سازی شده استفاده گردید که در محیط R به کمک تابع grf[۱۱] از بسته

[۸] Activation function
[۹] Train dataset
[۱۰] Test dataset
[۱۱] Gaussian random field

geoR و در یک شبکه منظم ۱۲ × ۱۲ شبیه‌سازی شدند. بدین منظور سه مدل متفاوت در نظر گرفته شد. برای مدل همبستگی داده‌ها از مدل کروی استفاده گردید. پارامترهای مربوطه در جدول ۱ نشان داده شده است.

جدول ۱: پارامترهای مدل نیم‌تغییرنگار کروی برای داده‌های شبیه‌سازی شده

مدل	آستانه	دامنه	اثر قطعه‌ای
مدل اول	۱۰	۵	۰/۰۰۱
مدل دوم	۵	۱۰	۰/۰۰۱
مدل سوم	۱	۱۵	۰/۰۰۱

این سه مدل از نظر مقدار آستانه[۱۲] و دامنه تاثیر[۱۳] با یکدیگر تفاوت دارند، به‌طوریکه اولین مدل دارای بیشترین مقدار آستانه و کمترین دامنه و مدل سوم دارای کمترین آستانه و بیشترین دامنه است. هم‌چنین به منظور اضافه کردن روند از رابطه زیر استفاده شد.

$$\text{trend} = ۰/۲x - ۰/۰۶y + ۰/۰۳x^۲ - ۰/۰۴y^۲ + ۰/۰۵xy$$

بازه در نظر گرفته شده برای مختصات x و y بازه صفر تا ده می‌باشد.

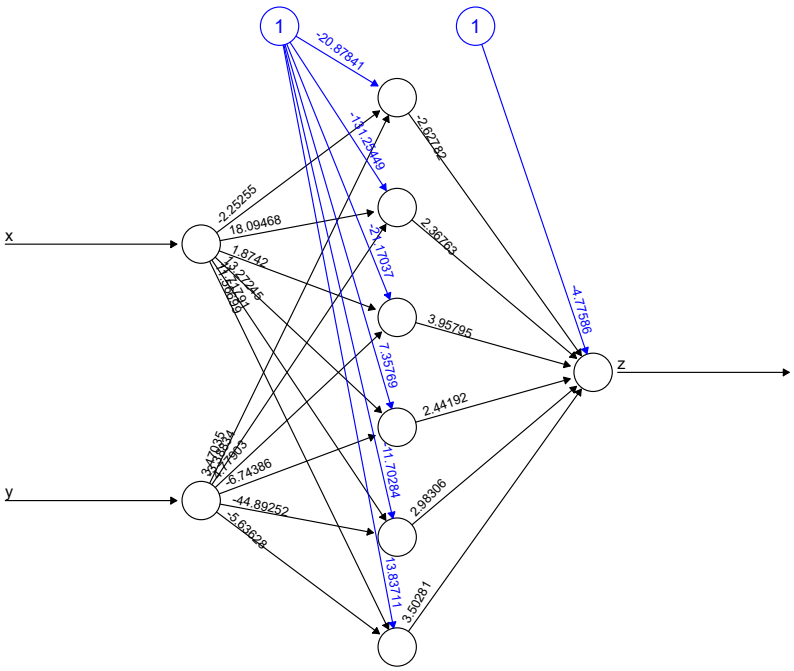
۱.۴ تعیین ساختار شبکه عصبی

برای دو روش NN و NNK از یک شبکه MLP با یک لایه مخفی، دو ورودی (مختصات x و y) و یک خروجی استفاده گردید. به منظور تعیین تعداد واحدهای موجود در لایه مخفی (H)، برای هر یک از سه مدل مذکور، شبکه‌های زیادی را با مقادیر مختلف برای H آموزش داده و شبکه‌ای را انتخاب می‌کنیم که خطای کمتری داشته باشد. بدین منظور، برای هر یک از سه مدل تعداد ۱۰۰ مجموعه داده شبیه‌سازی شده و هر مجموعه داده به‌صورت تصادفی به دو مجموعه داده آموزشی و آزمایشی تقسیم شد؛ به طوری‌که ۷۰ درصد داده‌ها به‌عنوان مجموعه داده آموزشی و بقیه به‌عنوان مجموعه داده آزمایش در نظر گرفته شد. سپس برای هر یک از مجموعه داده‌های آموزشی یک شبکه MLP آموزش داده شد و مقدار جذر میانگین مربعات خطا (RMSE) بر روی مجموعه داده آزمایشی محاسبه گردید. در نهایت متوسط مقادیر RMSE به‌دست آمده از صد تکرار به عنوان خطای شبکه در نظر گرفته شد که نتایج آن در جدول ۲ و همچنین شکل ۴ نشان داده شده است. این فرآیند برای مقادیر مختلف H بین ۱ تا ۲۵ تکرار گردید. بهترین حالت برای مدل اول تا سوم به ترتیب شبکه‌ای با ۱۱، ۶ و ۷ واحد در لایه مخفی است. لازم به ذکر است که انتخاب تعداد واحدهای لایه مخفی به گونه‌ای بوده است که از پیچیدگی شبکه جلوگیری شود. نمونه‌ای از این شبکه که دارای ۵ واحد در لایه مخفی است در شکل ۳ نشان داده شده است.

۵ بررسی دقت مقایسه روش‌های پیش‌گویی

برای مقایسه دقت پیش‌گویی چهار روش UK، RK، NN و NNK، از هر یک از سه مدل نیم تغییرنگار مذکور در بخش ۴ یک مجموعه داده فضایی منظم ۱۴۴تایی شبیه‌سازی شد. نمودار سه بعدی داده‌های شبیه‌سازی شده در شکل ۵ نشان داده شده است.

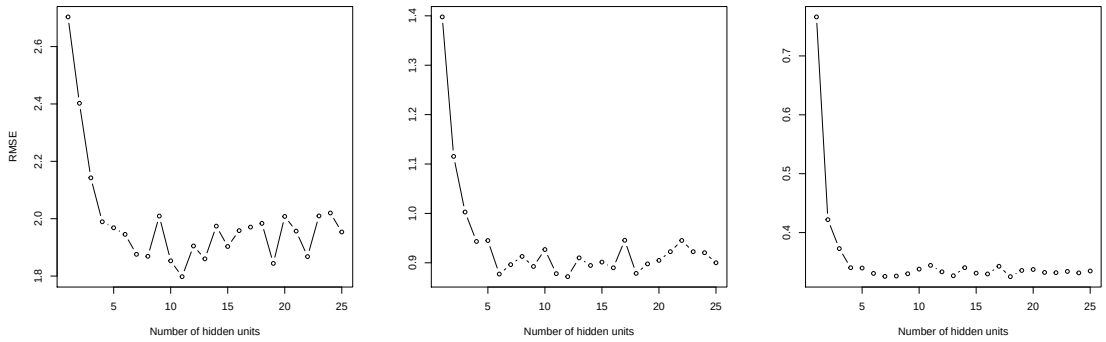
[۱۲] Sill
[۱۳] Range



شکل ۳: نمونه‌ای از شبکه مورد استفاده برای دو روش NN و NNK

جدول ۲: میزان خطای شبکه برای مقادیر مختلف تعداد نرون‌های لایه مخفی (H) برای هر یک از سه مدل

H	model1	model2	model3	H	model1	model2	model3
۱	۲/۷۰۳۴	۱/۳۹۷۸	۰/۷۶۶۲	۱۴	۱/۹۷۴۱	۰/۸۹۴۶	۰/۳۴۰۴
۲	۲/۴۰۲۴	۱/۱۱۵۴	۰/۴۲۱۹	۱۵	۱/۹۰۳۱	۰/۹۰۱۴	۰/۳۳۱۱
۳	۲/۱۴۲۴	۱/۰۰۲۸	۰/۳۷۲۸	۱۶	۱/۹۵۸۴	۰/۸۸۹۹	۰/۳۲۹۷
۴	۱/۹۸۹۴	۰/۹۴۳۳	۰/۳۴۰۳	۱۷	۱/۹۷۰۸	۰/۹۴۵۶	۰/۳۴۲۸
۵	۱/۹۶۸۴	۰/۹۴۵۰	۰/۳۳۹۸	۱۸	۱/۹۸۳۷	۰/۸۷۸۶	۰/۳۲۵۲
۶	۱/۹۴۵۴	۰/۸۷۷۲	۰/۳۳۰۷	۱۹	۱/۸۴۴۲	۰/۸۹۷۷	۰/۳۳۵۷
۷	۱/۸۷۵۹	۰/۸۹۶۳	۰/۳۲۵۶	۲۰	۲/۰۰۷۹	۰/۹۰۴۹	۰/۳۳۷۴
۸	۱/۸۶۹۰	۰/۹۱۲۹	۰/۳۲۶۲	۲۱	۱/۹۵۶۸	۰/۹۲۲۵	۰/۳۳۲۳
۹	۲/۰۰۹۲	۰/۸۹۲۶	۰/۳۳۰۱	۲۲	۱/۸۶۷۹	۰/۹۴۵۲	۰/۳۳۱۹
۱۰	۱/۸۵۳۱	۰/۹۲۶۸	۰/۳۳۷۹	۲۳	۲/۰۰۹۷	۰/۹۲۲۵	۰/۳۳۴۱
۱۱	۱/۷۹۸۰	۰/۸۷۸۰	۰/۳۴۴۳	۲۴	۲/۰۱۹۹	۰/۹۲۰۴	۰/۳۳۱۵
۱۲	۱/۹۰۴۹	۰/۸۷۲۰	۰/۳۳۳۴	۲۵	۱/۹۵۳۷	۰/۹۰۰۱	۰/۳۳۴۸
۱۳	۱/۸۶۰۱	۰/۹۱۰۰	۰/۳۲۶۷				



شکل ۴: روند کاهش مقادیر RMSE در اثر افزایش تعداد واحدهای لایه مخفی برای مدل اول (چپ) تا سوم (راست)

کمترین و بیشترین مقدار برای داده‌های شبیه‌سازی شده از مدل اول به ترتیب برابر ۶/۴۴− و ۱۱/۱۹، برای مدل دوم برابر با ۳/۶۶− و ۹/۳۹ و برای مدل سوم برابر با ۴/۱۴− و ۷/۹۱ می‌باشد. به منظور یکسان نگه داشتن شرایط برای هر چهار روش، از روش اعتبارسنجی متقاطع به‌صورت جداسازی تکی[۱۴] (LOOCV) استفاده شد. در این روش، پیش‌گویی مقدار یک نقطه داده با کنار گذاشتن مقدار مشاهده شده در آن نقطه و استفاده از بقیه مقادیر مشاهده شده برای ساخت مدل (و یا شبکه) انجام می‌گیرد. و این فرآیند برای تمامی نقاط مجموعه داده تکرار می‌شود.

عملکرد هر مدل به کمک دو معیار r^2 و RMSE مورد ارزیابی قرار گرفت به‌طوری‌که بزرگ بودن مقدار r^2 و کوچک بودن مقدار RMSE نشان می‌دهند که کارایی مدل بالاست.

$$r^2 = \left(\frac{\sum_{i=1}^{N} (z_i - \bar{z})(\hat{z}_i - \tilde{z})}{\sqrt{\sum_{i=1}^{N} (z_i - \bar{z})^2 \sum_{i=1}^{N} (\hat{z}_i - \tilde{z})^2}} \right)^2$$

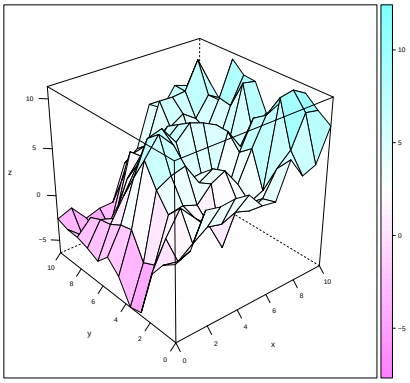
$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (z_i - \hat{z}_i)^2}$$

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. جدول ۳ مقادیر اندازه‌های عملکرد را برای هر یک از چهار روش به تفکیک برای هر مجموعه داده با دقت چهار رقم اعشار نشان می‌دهد.

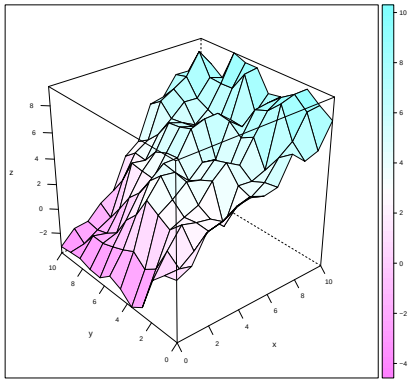
با توجه به نتایج به‌دست آمده مشاهده می‌شود که خطای پیش‌گویی هر چهار روش (مقدار RMSE) برای داده‌های شبیه‌سازی شده از مدل اول تا مدل سوم سیر نزولی داشته است (شکل ۶)؛ بطوری‌که خطای ایجاد شده برای مدل دوم حدودا نصف مدل اول و برای مدل سوم حدودا یک سوم مدل دوم است. از طرفی، برای داده‌های شبیه‌سازی شده از هر یک از سه مدل اختلاف مقادیر RMSE (و همچنین r^2) بین چهار روش مورد نظر بسیار اندک بوده است و همچنین اختلاف پیش‌گویی چهار روش برای داده‌های شبیه‌سازی شده از مدل اول تا مدل سوم سیر نزولی دارد. با اینحال برای هر سه مدل، نتایج روش UK از سه روش دیگر بهتر است. همچنین روش ترکیبی NNK نیز در هر سه حالت بهتر از NN بوده است. در نهایت مقایسه دو روش RK و NNK نشان می‌دهد که مقدار RMSE برای روش RK در مدل دوم بیشتر از NNK و در مدل اول و دوم کمتر از NNK است.

[۱۴] Leave one out cross validation

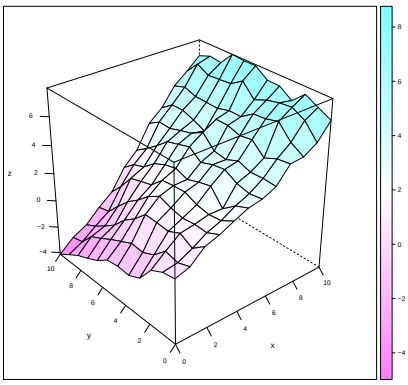
(الف) مدل اول



(ب) مدل دوم



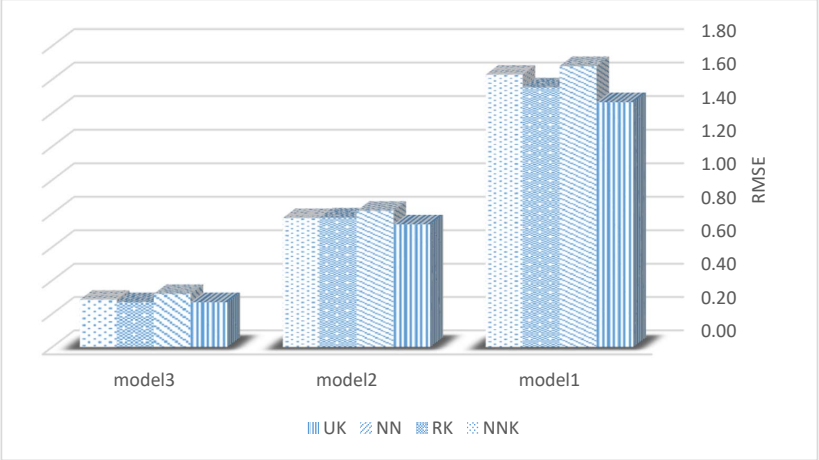
(ج) مدل سوم



شکل ۵: نمودار سه بعدی مجموعه داده‌های شبیه‌سازی شده

جدول ۳: مقایسه روش‌ها بر اساس دو معیار RMSE و r^2

مدل	روش	RMSE	r^2
مدل اول	UK	۱/۴۶۳۹	۰/۸۶۱۱
	RK	۱/۵۵۲۲	۰/۸۴۳۸
	NN	۱/۶۷۹۲	۰/۸۱۸۲
	NNK	۱/۶۲۸۴	۰/۸۲۹۳
مدل دوم	UK	۰/۷۳۵۴	۰/۹۵۲۳
	RK	۰/۷۷۶۳	۰/۹۴۶۸
	NN	۰/۸۱۶۱	۰/۹۴۱۲
	NNK	۰/۷۷۳۳	۰/۹۴۷۳
مدل سوم	UK	۰/۲۶۹۳	۰/۹۹۱۵
	RK	۰/۲۶۹۴	۰/۹۹۱۵
	NN	۰/۳۱۸۹	۰/۹۸۸۳
	NNK	۰/۲۸۵۲	۰/۹۹۰۶

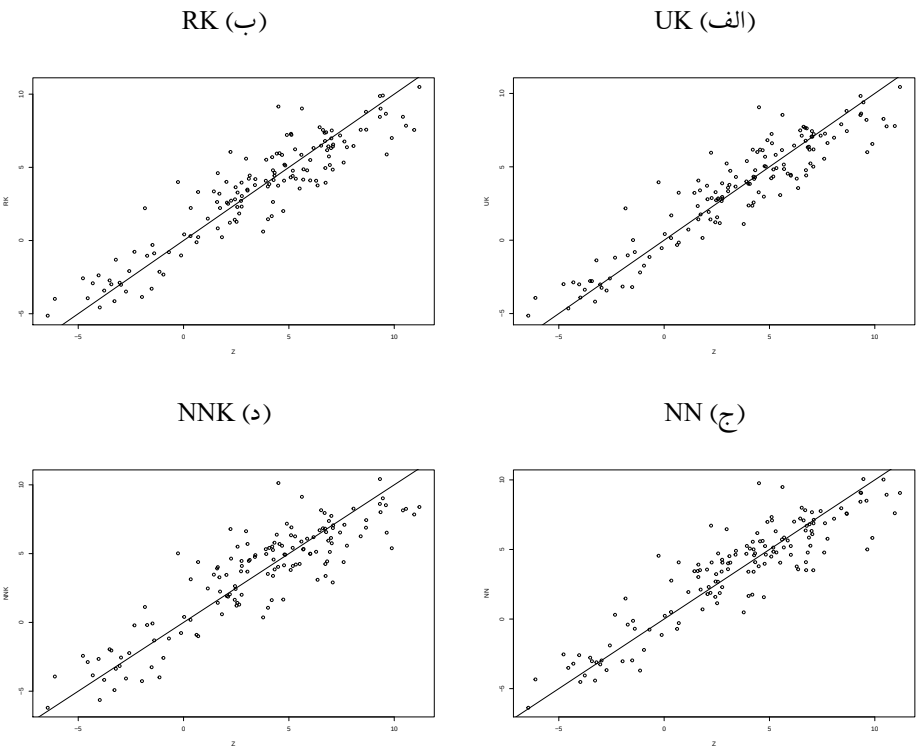


شکل ۶: مقایسه روش‌ها بر اساس معیار RMSE

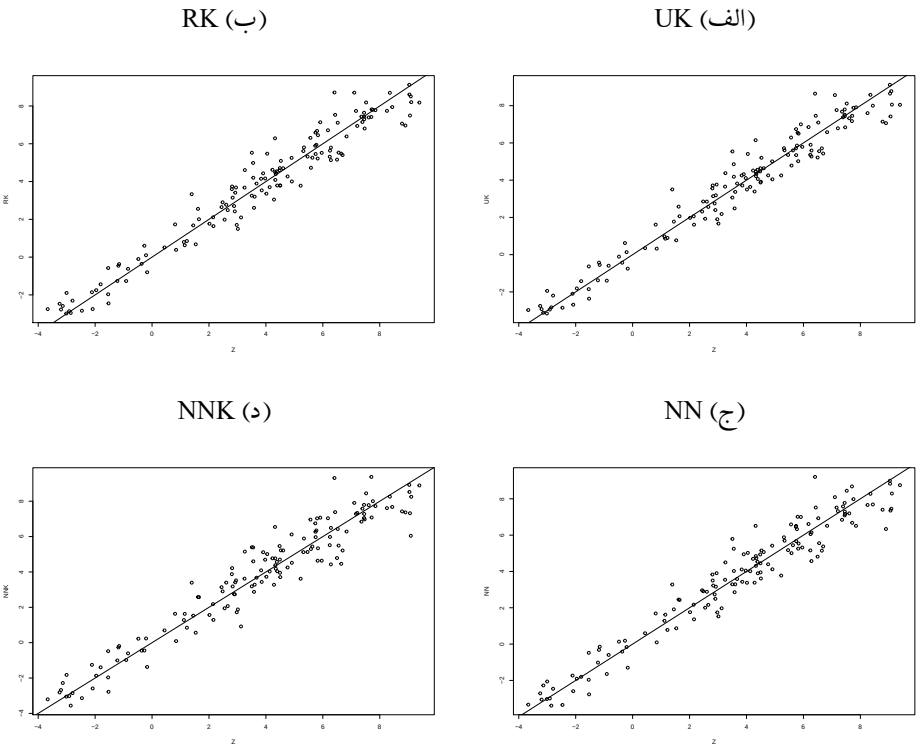
در شکل‌های ۷ تا ۹ نمودار مقادیر واقعی در برابر مقادیر پیش‌گویی شده، به همراه نمودار سه بعدی برای هر یک از چهار روش نشان داده شده است.

بحث و نتیجه‌گیری

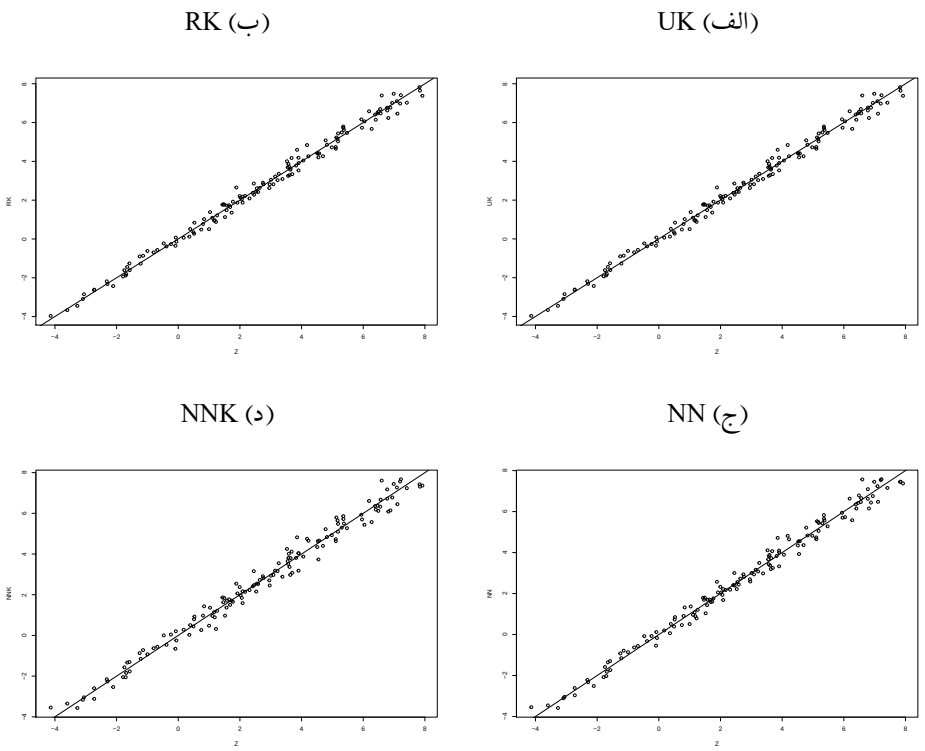
در این مقاله پیش‌گویی‌های دو روش کریگیدن و شبکه عصبی مورد مقایسه قرار گرفت. مثال‌های ارائه شده نشان داد که پیش‌گویی‌هایی که توسط روش‌های کریگیدن و رگرسیون‌ـ کریگیدن انجام گرفته است دقیق‌تر از پیش‌گویی‌های شبکه عصبی بر اساس معیار جذر میانگین مربعات خطا می‌باشد. با اینحال برای مدل‌های در نظر گرفته شده در این مقاله اختلاف مشاهده شده بین پیش‌گویی‌های چهار روش مورد نظر ناچیز است. از طرفی با کاهش تغییرپذیری داده‌ها از مدل اول تا مدل سوم، دقت پیش‌گویی‌های هر چهار روش افزایش یافته و مقادیر آنها به یکدیگر نزدیک‌تر شده است. البته بایستی توجه نمود که در اینجا داده‌ها از یک مدل خاص شبیه‌سازی شده‌اند و همان مدل‌ها هم برای پیش‌گویی‌شان توسط کریگیدن به‌کار رفته است، اما برای تنظیم پارامترهای شبکه هیچ‌گونه پیش‌فرضی استفاده نشده است. همچنین ممکن است بتوان با تغییر ساختار و تنظیم دقیق‌تر پارامترهای شبکه به نتایج بهتری دست یافت. با اینحال، هرچند که دقت پیش‌گویی‌های شبکه عصبی بهتر از کریگیدن نیست، اما نمودار سه بعدی داده‌های پیش‌گویی شده نشان می‌دهد که شبکه عصبی باز هم می‌تواند به‌عنوان یک روش پیش‌گویی برای داده‌های فضایی مورد استفاده قرار گیرد. از طرف دیگر، یکی از مزایای استفاده از شبکه عصبی اتوماتیک بودن تمامی محاسبات است، به‌طوریکه نیاز به دخالت زیادی از جانب کاربر ندارد. در روش کریگیدن بایستی مدل وابستگی فضایی با روش‌هایی (مثلا به‌صورت بصری و از روی نمودار تغییرنگار) یافت شود تا بعد بتوان از مدل یافت شده در فرآیند پیش‌گویی استفاده کرد. به‌عبارت دیگر، اگر مدل ساختار فضایی به درستی انتخاب نشود پیش‌گویی مناسبی نخواهیم داشت. اگر تعداد داده‌ها و تعداد عملیات زیاد باشد (مانند مه‌داده‌ها و داده کاوی) چنین کاری سودمند نخواهد بود و بهتر است تمامی مراحل به‌صورت خودکار انجام گیرد.



شکل ۷: نمودار مقادیر واقعی در برابر مقادیر پیش‌گویی شده از چهار روش برای مجموعه داده شبیه‌سازی شده از مدل اول



شکل ۸: نمودار مقادیر واقعی در برابر مقادیر پیش‌گویی شده از چهار روش برای مجموعه داده شبیه‌سازی شده از مدل دوم



شکل ۹: نمودار مقادیر واقعی در برابر مقادیر پیش‌گویی شده از چهار روش برای مجموعه داده شبیه‌سازی شده از مدل سوم

مراجع

محمدزاده م، (۱۳۹۴). *آمار فضایی و کاربردهای آن*، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Chung C., Chiang Y., and Chang E. (2012). A spatial neural fuzzy network for estimating pan evaporation at ungauged sites. *Hydrology and Earth System Sciences*, **16**(1), 255–266.

Cressie, N. (1993). *Statistics for Spatial Data*. Hoboken, NJ, USA: John Wiley and Sons, Inc.

Cressie, N. (1988). Spatial prediction and ordinary kriging. *Mathematical Geology*, **20**(4), 405–421

Foresee F. D., and Hagan M. T. (1997). Gauss-Newton approximation to Bayesian learning. In *Neural networks, 1997., international conference on (Vol. 3, pp. 1930-1935). IEEE*.

Gumus K., and Sen A. (2013). Comparison of spatial interpolation methods and multilayer neural networks for different point distributions on a digital elevation model. *Geodetski Vestnik*, 523–543. Retrieved from

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. 2nd ed. Upper Saddle River, N.J.: Prentice Hall,

Hengl T., Heuvelink G., and Stein A. (2003). Comparison of kriging with external drift and regression-kriging. *ITC*.

Kanevsky, M., Arutyunyan R., Bolshov L., Demyanov V., Maignan M. (1996). Artificial Neural Networks and Spatial Estimation of Chernobyl Fallout, *Geoinformatics*, **7**, 5–11.

Koike K., Matsuda S., and Gu B. (2001). Evaluation of interpolation accuracy of neural kriging with application to temperature-distribution analysis. *Mathematical Geology*, **33**(4), 421–448.

MacKay, D. J. (1992). A practical Bayesian framework for back-propagation networks. *Neural computation*, **4**(3), 448-472.

Matheron, G. (1969). Le krigeage universel (Universal kriging), *cahiers du Centre de Morphologie Mathematique*. Fontainebleau, France.

Matheron, G. (1971). *The theory of regionalized variables and its applications*, École National Supérieure des Mines.

Nevtipilova, V. (2014). Testing Artificial Neural Network (ANN) for Spatial Interpolation, *Journal of GEOsciences*, **3**(2), 1–9.

Paliwal, M., and Kumar, U. (2009). Neural networks and statistical techniques: A review of applications. *Expert systems with applications*, **36**(1), 2-17.

Ripley, B. (2007). *Pattern recognition and neural networks*. Cambridge university press.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, **65**(6), 65-386.

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*, Washington DC: Spartan.

Seo, Y., Yeo, W., Lee, S., and Jee H. (2010). Application of KED Method for Estimation of Spatial Distribution of Probability Rainfall. *Journal of Korea Water Resources Association*, **43**(8), 757–767.

Webster, R., and Oliver, M. (2007). *Geo-statistics for Environmental Scientists*. Wiley.

Yeh, I., Huang, K., and Kuo Y. (2013). Spatial interpolation using MLP–RBFN hybrid networks. *International Journal of Geographical Information Science*, **27**(10), 1884–1901.

Zimmerman, D., Pavlik, C., and Ruggles, A. (1999). An experimental comparison of ordinary and universal kriging and distance weighting. *Mathematical Geology*. **31**(4), 375-390.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶

پیش‌بینی بر اساس مدل‌های اتورگرسیو-میانگین متحرک فضایی-زمانی (STARMA)

سیمین جاوید [*۱]، علیرضا قدسی [۱]، کاظم علی آبادی [۲]، محمد بلبلیان قالیباف [۱]

[۱] گروه آمار، دانشگاه حکیم سبزواری

[۲] پژوهشکده جغرافیا، دانشگاه حکیم سبزواری

چکیده: مدل‌های اتورگرسیو-میانگین متحرک فضایی-زمانی (STARMA) دارای محبوبیت گسترده در بسیاری از زمینه‌ها از جمله تصویربرداری، حمل و نقل، کسب و کار، جرم شناسی و ... می‌باشند. این مدل‌ها ابزار مفیدی در مدل‌بندی سری‌های زمانی‌ای می‌باشند که در ارتباط با موقعیت‌های فضایی متفاوت هستند (که سری‌های فضایی-زمانی نامیده می‌شوند). ما در این مقاله با معرفی مدل‌های STARMA به بیان نحوه تعیین مرتبه مدل با استفاده از تابع‌های خودهمبستگی و خودهمبستگی جزئی می‌پردازیم. پارامترهای این مدل‌ها به روش درستنمایی ماکزیمم برآورد می‌شوند و سپس درستی مدل برازش داده شده با بررسی باقی‌مانده‌های مدل مورد ارزیابی قرار می‌گیرد. از امید ریاضی شرطی مدل حاصل به شرط مقادیر مشاهده شده برای پیش‌بینی فضایی-زمانی استفاده می‌گردد. مراحل فوق را برای حالت خاص STARMA(۱, ۲; ۱, ۱) با استفاده از بسته starma در نرم‌افزار R با مطالعه شبیه سازی بررسی می‌کنیم. در نهایت به عنوان کاربردی از این مدل، داده‌های جرم گزارش شده در شهر بوستون از ایالت ماساچوست آمریکا را مدل‌بندی می‌کنیم.

واژه‌های کلیدی: مدل اتورگرسیو-میانگین متحرک فضایی-زمانی، پیش‌بینی، ایستایی، همسایگی، جرم شناسی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62M10، 62M20، 62M30.

---

[*] **ارایه‌دهنده مقاله:** سیمین جاوید، siminjavid95@gmail.com

۱ مقدمه

اغلب داده‌های فضایی در بررسی‌های علوم محیطی مانند هواشناسی، محیط زیست، زمین شناسی و اقیانوس شناسی در طول زمان نیز به یکدیگر وابسته اند. مشاهداتی که هم از نظر موقعیت فضایی و هم از نظر زمانی وابسته باشند، داده‌های فضایی_زمانی [۱] نامیده می‌شوند (شکل۱). تحلیل این‌گونه داده‌ها مستلزم تعیین ساختار هبستگی فضایی_زمانی آن‌ها از طریق تابع کوواریانس است (محمدزاده ، ۱۳۹۴).

در سال‌های اخیر از خانواده مدل‌های اتورگرسیو_میانگین متحرک فضایی_زمانی (STARMA)[۲] در مدل‌بندی سری‌های زمانی که در ارتباط با موقعیت‌های فضایی متفاوت هستند، استفاده زیادی شده است . مدل STARMA توسط مارتین و اپن (۱۹۷۵) و پیفایفر و دویچ (۱۹۸۰) پیشنهاد شد. پیفایفر و دویچ (۱۹۸۰) کاربردی از مدل فضایی_زمانی را برای مدل کردن داده‌های جرم گزارش شده شهر بوستون (ایالت ماساچوست، آمریکا) ارائه دادند، که ما در این مقاله نتایج این تحقیق را در بخش آخر بیان خواهیم نمود. پیس و همکاران (۱۹۹۸) از مدل اتورگرسیو فضایی_زمانی (STAR)[۳] برای پیش‌بینی قیمت خانه تحت تاثیر داده‌های فضایی و زمانی در قیمت املاک و مستغلات استفاده کردند. آن‌ها با استفاده از داده‌ها بر روی قیمت مسکن نشان دادند که مزایای قابل توجهی با استفاده از مدل‌بندی فضایی و وابستگی زمانی می‌تواند به‌دست آید. علاوه بر این کاماریاناکیس و پراستاکوس (۲۰۰۵) روش اتورگرسیو_میانگین متحرک تلفیق شده فضایی_زمانی (STARIMA)[۴] را برای نشان دادن مدل‌های جریان ترافیک به‌کار بردند. این آزمایش در مرکز شهر آتن نشان داد که مدل STARIMA می‌تواند برای پیش‌بینی کوتاه مدت از فرایندهای فضایی_زمانی ایستا برای جریان ترافیک و ارزیابی اثر تغییرات آن بر روی قیمت‌های مختلف شهر مورد استفاده قرار گیرد. کریسپو و همکاران (۲۰۰۷) یک سیستم پیش‌بینی دنباله تصاویر را به‌کار بردند که تصاویر محتمل‌تری برای سری‌های داده شده با استفاده از روش‌های مبتنی بر مدل STAR به‌دست می‌آیند که مقایسه آن‌ها با تصاویر واقعی نشان دهنده این است که پیش‌بینی بسیار موفقیت آمیز بوده است. چنگ و همکاران (۲۰۰۸) نشان دادند زمانی که فرایندهای فضایی_زمانی ناایستا هستند، پیش‌بینی با استفاده از مدل ANN+STARMA نتایج دقیق‌تری نسبت به مدل STARMA دارد.

همه این مطالعات نشان می‌دهد که مدل‌های STARMA و STARIMA می‌توانند کاربردهای خوبی در مدل‌بندی فرایندهای فضایی_زمانی داشته باشند.

۲ همسایگی در فضا و زمان

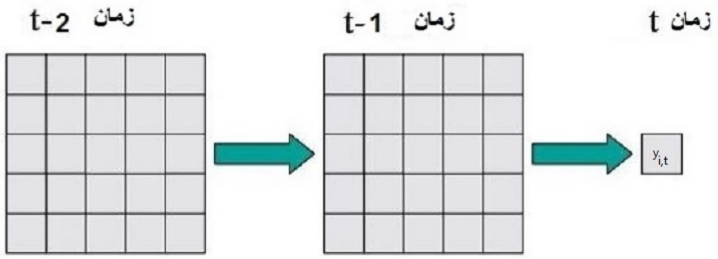
همسایگی مرتبه اول: نقاطی هستند که نزدیک ترین فاصله را به نقطه تحت بررسی دارند.

همسایگی مرتبه دوم: نقاطی هستند که باید دورتر از همسایگی مرتبه اول باشند، اما در عین حال نزدیک تر از سومین همسایگی می‌باشند.

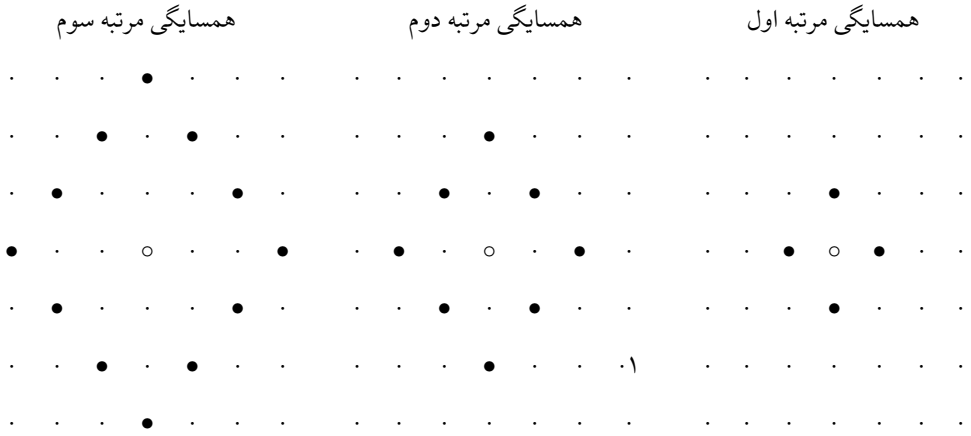
همسایگی‌های مرتبه بالاتر به‌طور مشابه تعریف می‌شوند. برای توضیح بیشتر می‌توان به شکل‌های ۱، ۲ و ۳ مراجعه نمود (پیفایفر و دویچ ، ۱۹۸۰؛ کریسپو و همکاران ، ۲۰۰۷؛ لی ، ۲۰۰۵). همان‌طور که در شکل‌های ۲ و ۳ دیده می‌شود دو نوع

[۱] Spatio-Temporal data
[۲] Spatio-Temporal autoregressive- moving average
[۳] Spatio-Temporal Autoregressive
[۴] Spatio-Temporal Autoregressive- Integrated Moving Average

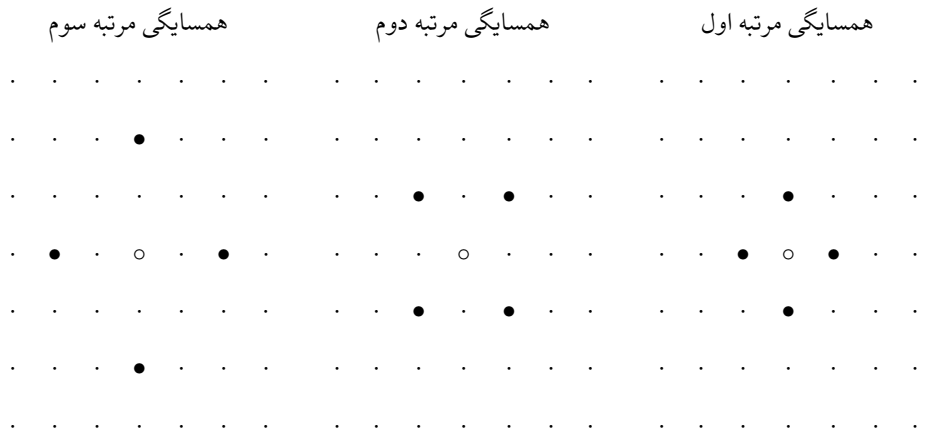
همسایگی در مقالات ذکر شده است که در بسته starmaدر نرم افزار R از نوع اول (شکل ۲) استفاده می‌شود.



شکل ۱: وابستگی فضایی_زمانی مشاهدات



شکل ۲: همسایگی فضایی



شکل ۳: همسایگی فضایی

۳ عملگر تأخیر فضایی

فرض کنید $y_{i,t}$ مشاهده در ناحیه $(i = ۱, \ldots, n)$ و در زمان $(t = ۱, \ldots, T)$ و $\boldsymbol{y}_t$ بردار مشاهدات $(y_{۱,t}, \ldots, y_{n,t})$ باشد. عملگر تأخیر فضایی را با L نشان می‌دهیم و به‌صورت زیر تعریف می شود

$$L^{۰}\boldsymbol{y}_t = W^{(۰)}\boldsymbol{y}_t = I\boldsymbol{y}_t, \qquad L^{s}\boldsymbol{y}_t = W^{(s)}\boldsymbol{y}_t, \qquad \forall s > ۰,$$

که در آن $W^{(s)}$ یک ماتریس $n \times n$ از وزن‌های s امین تأخیر فضایی $(w_{ij}^{(s)})$ است که به روش‌های مختلفی تعیین می‌شود که در بسته STARMA از نرم‌افزار R از روش زیر استفاده می‌شود

$$w_{ij}^{(s)} = \begin{cases} \frac{۱}{\text{تعداد همسایه‌های مرتبه s سلول i}} & \text{اگر سلول } j \text{ در همسایگی مرتبه } s \text{ سلول } i \text{ قرار داشته باشد} \\ ۰ & \text{سایر موارد} \end{cases}.$$

۴ مدل‌های اتورگرسیو‌ـ میانگین متحرک فضایی‌ـ زمانی STARMA$(l, p; m, q)$

مدل اتورگرسیو‌ـ میانگین متحرک فضایی‌ـ زمانی STARMA$(l, p; m, q)$ به صورت زیر تعریف می‌شود (اپرسون ، ۲۰۰۰؛ دای و بیلارد ، ۱۹۹۸؛ لی ، ۲۰۰۴، ۲۰۰۵):

$$\boldsymbol{y}_t = \sum_{s=۰}^{l}\sum_{k=۱}^{p}\alpha_{s,k}L^{s}\boldsymbol{y}_{t-k} - \sum_{s=۰}^{m}\sum_{k=۱}^{q}\beta_{s,k}L^{s}\boldsymbol{\varepsilon}_{t-k} + \boldsymbol{\varepsilon}_t \qquad (۱.۴)$$

که در آن $E[\boldsymbol{\varepsilon}_t] = ۰$، $E[\boldsymbol{\varepsilon}_t\boldsymbol{y}_{t-k}^{\top}] = ۰$ برای $k > ۰$ و $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_{t-k}^{\top}] = \sigma^{۲}I$ در صورتی‌که $k = ۰$ و در سایر موارد برابر صفر است. l، مرتبه فضایی اتورگرسیو؛ p، مرتبه زمانی اتورگرسیو؛ m، مرتبه فضایی میانگین متحرک؛ q، مرتبه زمانی میانگین متحرک؛ $\alpha_{s,k}$، ضریب اتورگرسیو در تأخیر زمانی k و تأخیر فضایی s؛ $\beta_{s,k}$، ضریب میانگین متحرک در تأخیر زمانی k و تأخیر فضایی s؛ L^s: ماتریس $n \times n$ از وزن‌ها برای مرتبه فضایی s می‌باشند.

از مدل (۱.۴) با حذف جمله دوم سمت راست، مدل STAR(l, p) و با حذف جمله اول سمت راست، مدل STMA(m, q) به‌دست می‌آید.

۵ تابع خودهمبستگی فضایی‌ـ زمانی

تعریف ۱.۵. *فرض کنید* $\{Y_{i,t}; i = ۱, ۲, \cdots, n \quad و \quad t = ۱, ۲, \cdots, T\}$ *سری فضایی‌ـ زمانی باشد. تابع خودهمبستگی فضایی‌ـ زمانی در تأخیر فضایی* s *و تأخیر زمانی* k *به‌صورت زیر بیان می‌شود (ACF)* [۵]

$$\rho_{s,k} = \frac{\gamma_{s,k}}{(\sigma_{۰}^{۲}\sigma_{s}^{۲})^{\frac{۱}{۲}}} = \frac{E[(Y_{i,t} - \mu)(L^{s}Y_{i,t-k} - \mu)]}{\{E[(Y_{i,t} - \mu)^{۲}]E[(L^{s}Y_{i,t} - \mu)^{۲}]\}^{\frac{۱}{۲}}},$$

که در آن $\mu = E[Y_{i,t}]$. *برآورد ACF از رابطه زیر به‌دست می‌آید (جنکینز و واتس ، ۱۹۶۸)*

$$r_{s,k} = \frac{\sum_{t=k+۱}^{T}\sum_{i=۱}^{n}(y_{i,t} - \bar{y})(L^{s}y_{i,t-k} - \bar{y})}{\left\{\left[\sum_{t=۱}^{T}\sum_{i=۱}^{n}(y_{i,t} - \bar{y})^{۲}\right]\left[\sum_{t=۱}^{T}\sum_{i=۱}^{n}(L^{s}y_{i,t} - \bar{y})^{۲}\right]\right\}^{\frac{۱}{۲}}} \qquad (۱.۵)$$

که در آن $s = ۰, ۱, \cdots$، $k = ۱, ۲, \cdots$ و $\bar{y} = \frac{۱}{nT}\sum_{t=۱}^{T}\sum_{i=۱}^{n}y_{i,t}$

[۵] Auto Correlation Function

$$
\begin{array}{cc}
 & \\
 & \\
j=0 & h=0 \\
 & h=0 \\
 & h=1 \\
j=1 & h=2 \\
 & \vdots \\
 & h=l \\
 & \vdots \\
 & \vdots \\
 & h=0 \\
 & h=1 \\
j=p & \vdots \\
 & h=l
\end{array}
\begin{array}{c}
\overbrace{\phantom{r_{0,0,0,0}}}^{k=0} \\
s=0 \\
\hline
r_{0,0,0,0} \\
r_{0,1,0,0} \\
r_{1,1,0,0} \\
r_{2,1,0,0} \\
\vdots \\
r_{l,1,0,0} \\
\vdots \\
\vdots \\
r_{0,p,0,0} \\
r_{1,p,0,0} \\
\vdots \\
r_{l,p,0,0}
\end{array}
\begin{array}{c}
\overbrace{\phantom{r_{0,0,0,1}\ r_{0,0,1,1}\ r_{0,0,2,1}\ \cdots\ r_{0,0,l,1}}}^{k=1} \\
s=0 \quad s=1 \quad s=2 \quad \cdots \quad s=l \\
\hline
r_{0,0,0,1} \quad r_{0,0,1,1} \quad r_{0,0,2,1} \quad \cdots \quad r_{0,0,l,1} \\
r_{0,1,0,1} \quad r_{0,1,1,1} \quad r_{0,1,2,1} \quad \cdots \quad r_{0,1,l,1} \\
r_{1,1,0,1} \quad r_{1,1,1,1} \quad r_{1,1,2,1} \quad \cdots \quad r_{1,1,l,1} \\
r_{2,1,0,1} \quad r_{2,1,1,1} \quad r_{2,1,2,1} \quad \cdots \quad r_{2,1,l,1} \\
\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
r_{l,1,0,1} \quad r_{l,1,1,1} \quad r_{l,1,2,1} \quad \cdots \quad r_{l,1,l,1} \\
\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
r_{0,p,0,1} \quad r_{0,p,1,1} \quad r_{0,p,2,1} \quad \cdots \quad r_{0,p,l,1} \\
r_{1,p,0,1} \quad r_{1,p,1,1} \quad r_{1,p,2,1} \quad \cdots \quad r_{1,p,l,1} \\
\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
r_{l,p,0,1} \quad r_{l,p,1,1} \quad r_{l,p,2,1} \quad \cdots \quad r_{l,p,l,1}
\end{array}
\begin{array}{c}
\overbrace{\phantom{r_{0,0,0,p}\ r_{0,0,1,p}\ \cdots\ r_{0,0,l,p}}}^{k=p} \\
s=0 \quad s=1 \quad \cdots \quad s=l \\
\hline
r_{0,0,0,p} \quad r_{0,0,1,p} \quad \cdots \quad r_{0,0,l,p} \\
r_{0,1,0,p} \quad r_{0,1,1,p} \quad \cdots \quad r_{0,1,l,p} \\
r_{1,1,0,p} \quad r_{1,1,1,p} \quad \cdots \quad r_{1,1,l,p} \\
r_{2,1,0,p} \quad r_{2,1,1,p} \quad \cdots \quad r_{2,1,l,p} \\
\vdots \quad \vdots \quad \vdots \quad \vdots \\
r_{l,1,0,p} \quad r_{l,1,1,p} \quad \cdots \quad r_{l,1,l,p} \\
\vdots \quad \vdots \quad \vdots \quad \vdots \\
\vdots \quad \vdots \quad \vdots \quad \vdots \\
r_{0,p,0,p} \quad r_{0,p,1,p} \quad \cdots \quad r_{0,p,l,p} \\
r_{1,p,0,p} \quad r_{1,p,1,p} \quad \cdots \quad r_{1,p,l,p} \\
\vdots \quad \vdots \quad \vdots \quad \vdots \\
r_{l,p,0,p} \quad r_{l,p,1,p} \quad \cdots \quad r_{l,p,l,p}
\end{array}
$$

شکل ۴: ساختار ماتریس خودهمبستگی فضایی‌ـ زمانی R_y

۶ تابع خودهمبستگی جزئی فضایی‌ـ زمانی

منظور از همبستگی بین $Y_{i,t}$ و $L^sY_{i,t-k}$ بعد از حذف وابستگی خطی مشترک $(L^sY_{i,t-1}, L^sY_{i,t-2}, \cdots, L^{s-1}Y_{i,t-k+1})$، همبستگی شرطی زیر است و معمولا در تحلیل سری‌های فضایی‌ـ زمانی، تابع خودهمبستگی جزئی (PACF) [۶] نامیده می‌شود.

$$\phi_{s,s;k,k} = \mathrm{Corr}(Y_{i,t}, L^sY_{i,t-k} | L^sY_{i,t-1}, L^sY_{t-2}, ..., L^sY_{i,t-k+1}).$$

برای برآورد خودهمبستگی‌های جزئی، ما به ماتریس خودهمبستگی‌ها بین همه متغیرهای تأخیری جفت شده نیاز داریم، $L^sy_{i,t-k}$ و $L^hy_{i,t-j}$ $(j,k = 1, 2, \ldots$ و $h,s = 0, 1, 2, \ldots)$ این همبستگی‌ها را با نماد توسعه یافته $r_{h,j,s,k}$ نشان می‌دهیم که به‌صورت زیر تعریف می‌شود

$$r_{h,j,s,k} = \frac{\sum_{t=\nu+1}^{T}\sum_{i=1}^{n}(L^hy_{i,t-j} - \bar{y})(L^sy_{i,t-k} - \bar{y})}{\left[\sum_{t=1}^{T}\sum_{i=1}^{n}(L^hy_{i,t} - \bar{y})^2\right]^{\frac{1}{2}}\left[\sum_{t=1}^{T}\sum_{i=1}^{n}(L^sy_{i,t} - \bar{y})^2\right]^{\frac{1}{2}}} \qquad (۱.۶)$$

که در آن $\nu = \max(j,k)$ است. همبستگی جزئی بین $y_{i,t}$ و $L^sy_{i,t-k}$ از رابطه

$$\psi_{0,0,s,k} = \frac{-c_{0,0,s,k}}{(c_{0,0,0,0}c_{s,k,s,k})^{\frac{1}{2}}}$$

به‌دست می‌آید که $c_{0,0,s,k}$ کوفاکتور $r_{0,0,s,k}$ در دترمینان ماتریس خودهمبستگی فضایی‌ـ زمانی $|R_y|$ (شکل ۴) است.

[۶] Partial Auto Correlation Function

۷ فرایندهای فضایی_زمانی ناایستا

اگر نمودار همبستگی‌نگار تخمین زده شده به سرعت کاهش نیابد، آنگاه فرایند فضایی_زمانی ناایستا در نظر گرفته می‌شود. فرایندهای ناایستا را می‌توان با استفاده از عملگرهای تفاضلی فضایی و زمانی مرتبه اول به صورت $\nabla_T y_{i,t} = y_{i,t} - y_{i,t-1}$، $\nabla_S y_{i,t} = y_{i,t} - \sum_{j \in J_1} w_{ij} y_{i,t}$ و $\nabla_S \nabla_T y_{i,t} = \nabla_T y_{i,t} - \sum_{j \in J_1} w_{ij} \nabla_T y_{i,t}$ تبدیل به فرایند ایستا نمود.

۸ تعیین مرتبه مدل

برای تشخیص فرایندهای اتورگرسیو، میانگین متحرک و اتورگرسیو_میانگین متحرک در فضا و زمان، $r_{s,k}$ ها و $\phi_{s,s;k,k}$ ها را به‌دست می‌آوریم و نمودار آن‌ها را رسم می‌کنیم. به نمودار $r_{s,k}$ در مقابل s,k خودهمبستگی نگار[۷] و به نمودار $\phi_{s,s;k,k}$ در مقابل s,k خودهمبستگی نگار جزئی گوییم. مرتبه‌های l و p با استفاده از خودهمبستگی نگارجزئی و مرتبه‌های m و q با استفاده از خودهمبستگی نگار با توجه به جدول ۱ تعیین می‌شوند.

جدول ۱: الگوهای خودهمبستگی و خودهمبستگی جزئی برای مدل‌های STARMA

مدل	$r_{s,k}$	$\phi_{s,s;k,k}$
STAR(l,p)	$r_{s,k} \searrow ۰$ وقتی $s,k \to \infty$	$\phi_{s,s;k,k} = ۰$ برای $s > l, k > p$
STMA(m,q)	$r_{s,k} = ۰$ برای $s > l, k > p$	$\phi_{s,s;k,k} \searrow ۰$ وقتی $s,k \to \infty$
STARMA(l,p,m,q)	$r_{s,k} \searrow ۰$ وقتی $s > m-l, k > q-p$	$\phi_{s,s;k,k} \searrow ۰$ وقتی $s > l-m, k > p-q$

۹ برآورد ضرایب مدل STARMA

با فرض این‌که خطاهای ε_t دارای توزیع نرمال با میانگین صفر و ماتریس واریانس_کوواریانس برابر با $\sigma^۲ I_{nT}$ باشند، تابع درستنمایی ماکزیمم مدل به‌صورت زیر است

$$f(\boldsymbol{\varepsilon}|\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^۲) = (۲\pi)^{-\frac{nT}{۲}} |\sigma^۲ I_{nT}|^{-\frac{۱}{۲}} \exp\left\{-\frac{۱}{۲\sigma^۲}\boldsymbol{\varepsilon}' I \boldsymbol{\varepsilon}\right\}$$
$$= (۲\pi)^{-\frac{nT}{۲}} (\sigma^۲)^{-\frac{nT}{۲}} \exp\left\{-\frac{S(\boldsymbol{\alpha},\boldsymbol{\beta})}{۲\sigma^۲}\right\}$$

که $\boldsymbol{\alpha} = [\alpha_{۰,۱},\ldots,\alpha_{۰,p},\ldots,\alpha_{l,۱},\ldots,\alpha_{,p}]^\top$، $\boldsymbol{\beta} = [\beta_{۰,۱},\ldots,\beta_{۰,q},\ldots,\beta_{m,۱},...,\beta_{m,q}]^\top$، $S(\boldsymbol{\alpha},\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\top I \boldsymbol{\varepsilon} = \sum_{i=۱}^{n}\sum_{t=۱}^{T} \varepsilon_{i,t}^۲$ و $\boldsymbol{\varepsilon} = [\varepsilon_{۱,۱},...,\varepsilon_{۱,T},...,\varepsilon_{n,۱},...\varepsilon_{n,T}]^\top$. برآورد پارامترها با ماکزیمم کردن تابع فوق به‌دست می‌آیند.

۱۰ بررسی تشخیصی

اگر مدل برازش داده شده به داده‌ها مناسب باشد باقی‌مانده‌ها باید فرایند تصادفی محض دارای توزیع نرمال با میانگین صفر و ماتریس واریانس_کوواریانسی برابر $\sigma^۲ I_n$ باشد یعنی کوواریانس و همچنین خودهمبستگی بین باقی‌مانده‌ها باید صفر باشد. یک روش

[۷] Autocorrelogram

آزمون صفر بودن همبستگی‌ها استفاده از خودهمبستگی‌های باقی‌مانده‌های مدل با استفاده از فاصله اطمینان $\pm ۲\sqrt{\mathrm{Var}(r_{s,k})}$ است (پیفایفر و دویچ ، ۱۹۸۱) که در آن

$$\mathrm{Var}(r_{s,k}) \approx \frac{۱}{n(T-k)}.$$

پارامترهای برآورد شده نیز باید از لحاظ معنی‌داری آماری آزمون شوند. فرض کنید $\delta = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ و $\hat{\delta}$ برآورد کمترین مربعات بردار تمام پارامترها و $\hat{\delta^*}$ برآورد کمترین مربعات بردار پارامترها تحت فرضیه صفر $H_۰ : \delta_K = ۰$ باشند. آماره آزمون به‌صورت زیر است

$$\frac{(nT-K)[S_*(\hat{\delta^*}) - S_*(\hat{\delta})]}{S_*(\hat{\delta})} \qquad (۱.۱۰)$$

که دارای توزیع حدی $F_{۱,nT-K}$ تحت فرض صفر است. هر پارامتر برآورد شده‌ای که ثابت شود از لحاظ آماری معنی‌دار نیست باید از مدل حذف شود و مدل ساده‌تر در نظر گرفته شود و مرحله برآورد باید دوباره تکرار شود.

۱۱ پیش‌بینی برای مدل STARMA در حالت ایستا

در رابطه (۱.۴) اگر t را به $t+h$ تبدیل کنیم خواهیم داشت

$$\boldsymbol{y}_{t+h} = \sum_{s=۰}^{l}\sum_{k=۱}^{p} \alpha_{s,k} L^s \boldsymbol{y}_{t+h-k} - \sum_{s=۰}^{m}\sum_{k=۱}^{q} \beta_{s,k} L^s \boldsymbol{\varepsilon}_{t+h-k} + \boldsymbol{\varepsilon}_{t+h}$$

اگر از طرفین رابطه فوق امید شرطی بگیریم، معادله تفاضلی برای محاسبه پیش‌بینی‌ها به‌صورت زیر است

$$\begin{aligned}
\hat{\boldsymbol{y}}_t(h) = {} & \alpha_{۰,۱}\hat{\boldsymbol{y}}_t(h-۱) + \cdots + \alpha_{۰,p}\hat{\boldsymbol{y}}_t(h-p) + \cdots + \alpha_{l,۱}L^l\hat{\boldsymbol{y}}_t(h-۱) + \cdots + \alpha_{l,p}L^l\hat{\boldsymbol{y}}_t(h-p) \\
& - \beta_{۰,۱}\mathrm{E}(\boldsymbol{\varepsilon}_{t+h-۱}|\boldsymbol{y}_t, \boldsymbol{y}_{t-۱}, \cdots, \boldsymbol{y}_۱) - \cdots - \beta_{۰,q}E(\boldsymbol{\varepsilon}_{t+h-q}|\boldsymbol{y}_t, \boldsymbol{y}_{t-۱}, \cdots, \boldsymbol{y}_۱) - \cdots \\
& - \beta_{m,۱}L^m E(\boldsymbol{\varepsilon}_{t+h-۱}|\boldsymbol{y}_t, \boldsymbol{y}_{t-۱}, \cdots, \boldsymbol{y}_۱) - \cdots - \beta_{m,q}L^m E(\boldsymbol{\varepsilon}_{t+h-q}|\boldsymbol{y}_t, \boldsymbol{y}_{t-۱}, \cdots, \boldsymbol{y}_۱) \\
& + E(\boldsymbol{\varepsilon}_{t+h}|\boldsymbol{y}_t, \boldsymbol{y}_{t-۱}, \cdots, \boldsymbol{y}_۱)
\end{aligned}$$

که در آن

$$E(\boldsymbol{\varepsilon}_{t+j}|\boldsymbol{y}_t, \boldsymbol{y}_{t-۱}, \cdots, \boldsymbol{y}_۱) = \begin{cases} ۰ & j \geq ۱ \\ \boldsymbol{\varepsilon}_{t+j} & j \leq ۰ \end{cases} \qquad (۱.۱۱)$$

که برای $j \leq ۰$ ، $\boldsymbol{\varepsilon}_{t+j}$ از رابطه زیر محاسبه می‌شود

$$\boldsymbol{\varepsilon}_{t+j} = \boldsymbol{y}_{t+j} - \sum_{s=۰}^{l}\sum_{k=۱}^{p} \alpha_{s,k} L^s \boldsymbol{y}_{t+j-k} + \sum_{s=۰}^{m}\sum_{k=۱}^{q} \beta_{s,k} L^s \boldsymbol{\varepsilon}_{t+j-k}$$

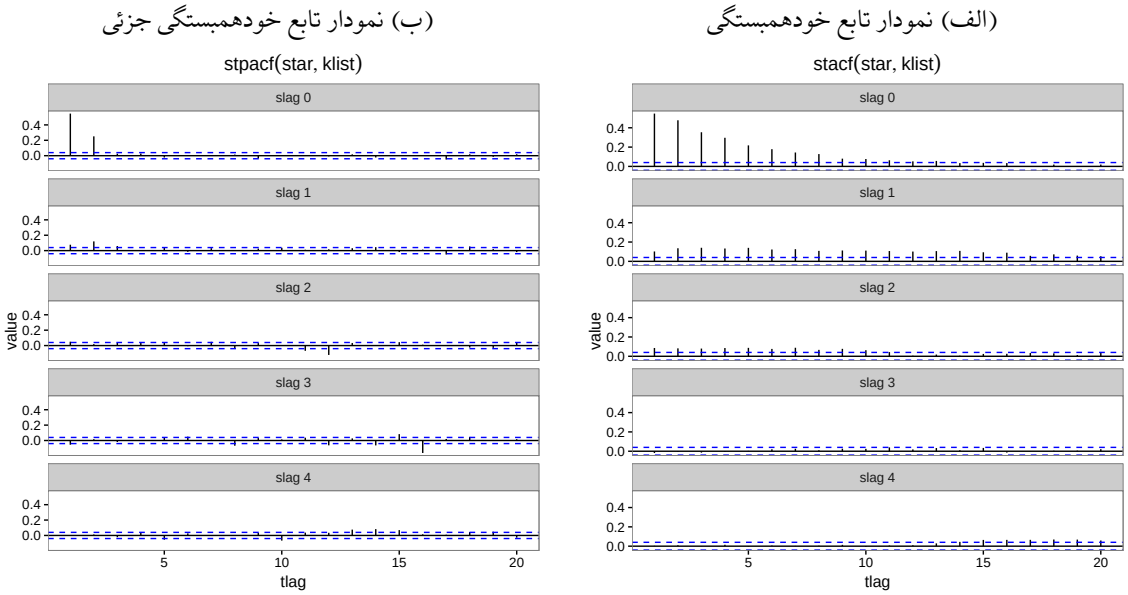
برای $۱ \leq t \leq T$ به جای $\boldsymbol{y}_t$ مقدار مشاهده شده را قرار می‌دهیم.

۱۲ شبیه‌سازی

در این بخش فرایند STARMA$(۱, ۲; ۱, ۱)$ یعنی

$$\boldsymbol{y}_t = \alpha_{۰,۱}\boldsymbol{y}_{t-۱} + \alpha_{۰,۲}\boldsymbol{y}_{t-۲} + \alpha_{۱,۱}L^۱\boldsymbol{y}_{t-۱} + \alpha_{۱,۲}L^۱\boldsymbol{y}_{t-۲} - \beta_{۰,۱}\boldsymbol{\varepsilon}_{t-۱} - \beta_{۱,۱}L^۱\boldsymbol{\varepsilon}_{t-۱} + \boldsymbol{\varepsilon}_t. \qquad (۱.۱۲)$$

با فرض $\alpha_{۰,۱} = ۰/۴$ ، $\alpha_{۰,۲} = ۰/۲۵$ ، $\alpha_{۱,۱} = ۰/۲۵$ ، $\alpha_{۱,۲} = ۰$ ، $\beta_{۰,۱} = ۰$ ، و $\beta_{۱,۱} = ۰/۳$ را در یک ناحیه‌ی مشبکه‌ای منظم ۵×۵ و ۲۰۱ نقطه زمانی شبیه‌سازی و سپس به منظور ایستایی، داده‌های ۱۰۰ زمان اول را حذف می‌کنیم و داده‌های زمان ۲۰۱ را نیز برای ارزیابی مدل کنار می‌گذاریم. ابتدا داده‌ها را مرکزی می‌نماییم. با توجه به نمودار خودهمبستگی نگار (شکل ۵، الف) و نمودار خودهمبستگی نگار جزئی (شکل ۵، ب) مرتبه مدل برابر $p = ۲، l = ۱ و q = ۱، m = ۱$ تعیین می‌شود. پس از برآورد پارامترهای مدل، چون مقدار احتمال p-value در آزمون معنی‌داری ضرایب مدل برای ضرایب $\alpha_{۱,۲}$ و $\beta_{۰,۱}$ بیشتر از ۰/۰۵ می‌باشد، این ضرایب را صفر قرار می‌دهیم و سایر ضرایب را دوباره برآورد می‌نماییم که نتایج در جدول ۲ آمده است.



شکل ۵: نمودارهای همبستگی

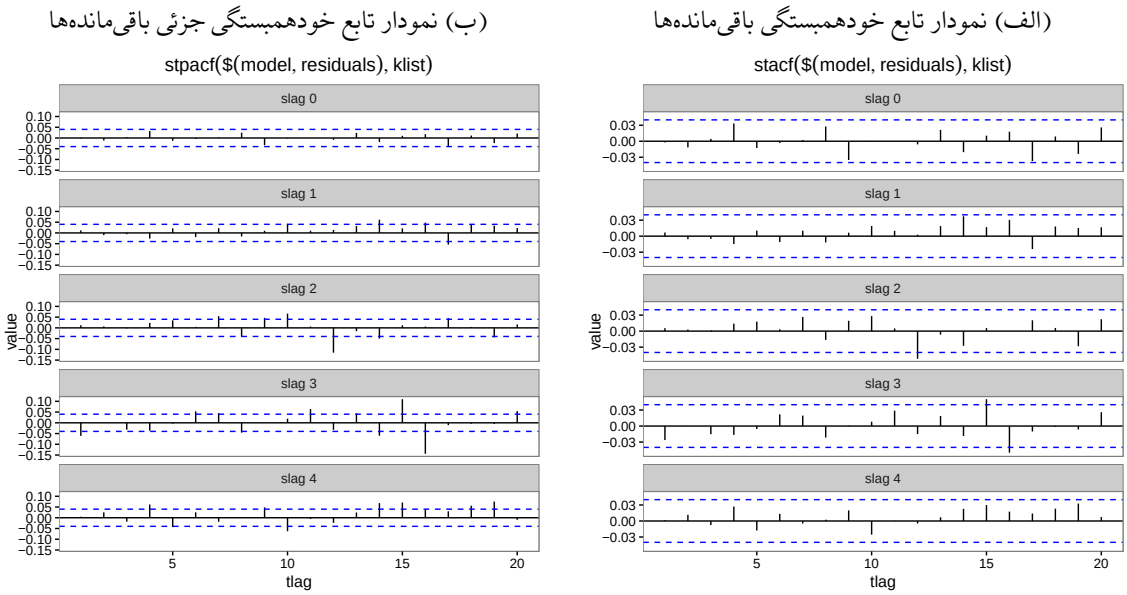
همان‌طور که مشاهده می‌شود مقادیر برآورد شده ضرایب تقریبا برابر مقادیر واقعی است. این موضوع و نمودار خودهمبستگی (شکل ۶، الف) و نمودار خودهمبستگی جزئی باقی‌مانده‌ها (شکل ۶، ب) نشان دهنده مناسب بودن مدل به‌دست آمده است.

با توجه به رابطه (۱.۱۲) پیش‌بینی یک مرحله بعد به‌صورت زیر خواهد بود

$$\hat{\boldsymbol{y}}_t(۱) = \alpha_{۰,۱}\boldsymbol{y}_t + \alpha_{۰,۲}\boldsymbol{y}_{t-۱} + \alpha_{۱,۱}L^{۱}\boldsymbol{y}_t + \alpha_{۱,۲}L^{۱}\boldsymbol{y}_{t-۱} - \beta_{۰,۱}\boldsymbol{\varepsilon}_t - \beta_{۱,۱}L^{۱}\boldsymbol{\varepsilon}_t.$$

جدول ۲: برآورد پارامترها

	برآورد	میانگین مربعات خطا	t-value	p-value
$\alpha_{۰,۱}$	۰/۳۹۸۷	۰/۰۱۹۷	۲۰/۱۸۴۲	$۲/۲ \times ۱۰^{-۱۶}$
$\alpha_{۱,۱}$	۰/۲۲۴۴	۰/۰۴۴۵	۵/۰۳۹۳	$۵/۰۰۸ \times ۱۰^{-۷}$
$\alpha_{۰,۲}$	۰/۲۴۳۸	۰/۰۱۹۶	۱۲/۴۲۶۵	$۲/۲ \times ۱۰^{-۱۶}$
$\beta_{۱,۱}$	۰/۲۷۰۰	۰/۰۵۷۰	−۴/۷۳۵۴	$۲/۳۰۸ \times ۱۰^{-۶}$



شکل ۶: نمودارهای همبستگی باقی‌مانده‌ها

با فرض $\alpha_{۱,۲} = ۰$ و $\beta_{۰,۱} = ۰$ و استفاده از مقادیر برآورد شده سایر پارامترها (داده شده در جدول ۲) پیش‌بینی‌های یک مرحله بعد با استفاده از رابطه زیر به‌دست می‌آید

$$\hat{\boldsymbol{y}}_t(۱) = ۰/۴\boldsymbol{y}_t + ۰/۲۵\boldsymbol{y}_{t-۱} + ۰/۲۵L^۱\boldsymbol{y}_t - ۰/۳L^۱\varepsilon_t$$

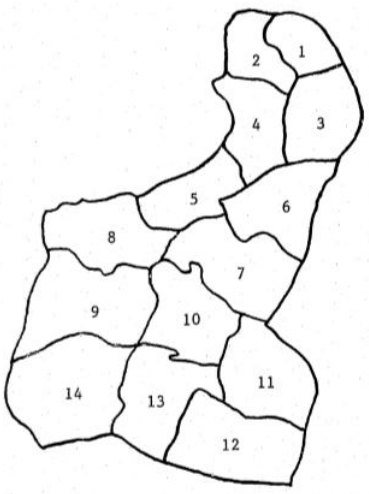
حال با فرض $t = ۱۰۱$ مقادیر آینده را با استفاده از مدل فوق پیش‌بینی می‌کنیم. مقدار میانگین مربعات خطای پیش‌بینی برابر ۰/۹۱۲۰۲۹۷ به‌دست آمد.

۱۳ کاربرد مدل فضایی‌ـ‌زمانی STARMA در جرم شناسی

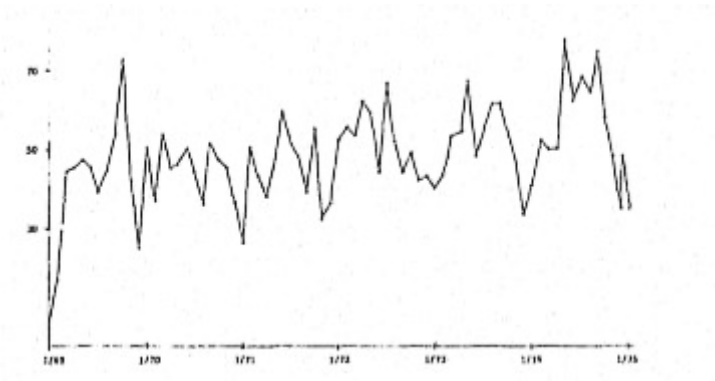
برای نشان دادن کاربردی از مدل‌های فضاـ‌زمان STARMA، تعداد بازداشتی‌های ناشی از ضرب و شتم ماهانه ۱۴ ناحیه از شمال شرق شهر بوستون از ایالت ماساچوست آمریکا (شکل ۷) در سال‌های ۱۹۶۹ الی ۱۹۷۴ را در نظر می‌گیریم.

مجموع داده‌ها در $n = ۱۴$ موقعیت و $T = ۷۲$ دوره زمانی مشاهده شده در شکل ۸ نشان داده شده است.

بررسی اولیه داده‌ها نشان می‌دهد که خودهمبستگی ها با سرعت بسیار آهسته نزول می‌کنند، بنابراین داده‌ها ناایستا بوده و نیاز به تفاضلی کردن مرتبه اول دارند. مقادیر خودهمبستگی و خوهمبستگی جزئی فضاـ‌زمان سری تفاضلی شده در جدول ۳ ارائه شده است. چون خودهمبستگی‌های جزئی سیر نزولی دارند و خودهمبستگی‌ها هم در بعد فضا و هم در بعد زمان بعد از تأخیر ۱ قطع می‌شوند، مدل مناسب برای سری تفاضلی شده STMA$(۰, ۱)$ است. از آن‌جا که داده‌ها یک‌بار تفاضلی شده‌اند، این مدل را می‌توان STARIMA$(۰, ۰; ۱; ۰, ۱)$ نامید، که یک فرایند اتورگرسیوـ‌میانگین متحرک تلفیق شده فضایی‌ـ‌زمانی است و ۱ اضافی برای تفاضلی کردن مرتبه اول است. این مدل به‌صورت $\boldsymbol{y}_t - \boldsymbol{y}_{t-۱} = -\beta_{۰,۱}\varepsilon_{t-۱} + \varepsilon_t$ است. پارامتر این مدل $\hat{\beta}_{۰,۱}$، ۰/۸۰۳ برآورد شد و مجموع مربعات باقی‌مانده‌ها ۶۵۰۴/۶۷۹ به‌دست آمد. واریانس برآورد شده $\hat{\beta}_{۰,۱}$ برابر با ۰/۰۰۰۳۶۸ و بنابراین



شکل ۷: ۱۴ ناحیه شمال شرق بوستون



شکل ۸: مجموع تعداد بازداشتی‌های ناشی از ضرب و شتم در شمال شرق بوستون در سال‌های ۱۹۶۹ تا ۱۹۷۴

فاصله اطمینان تقریبی ۹۵٪ برای $\beta_{\cdot,1}$ (۰/۷۶۴, ۰/۸۴۱) است. جدول ۴ تابع‌های همبستگی فضاـ زمان را برای باقی‌مانده‌های این مدل نشان می‌دهد. مقادیر نسبتا بزرگ برای $r_{1,1}$ (در مقایسه با واریانس تقریبی $(۱۴ \times ۷۰)^{-۱}$) و $\phi_{۱,۱;۱,۱}$ در جدول ۴ مدل جدید، STARIMA$(۰, ۰; ۱; ۱, ۱)$ را پیشنهاد می‌دهد. برای مدل جدید، $\hat{\beta}_{\cdot,1}$ و $\hat{\beta}_{1,1}$ به‌ترتیب ۰/۸۱۲، ۰/۰۹۲− و مجموع مربعات باقی‌مانده‌ها ۶۴۵۷/۲۶۶ به‌دست آمد. همبستگی باقی‌مانده‌های این مدل در جدول ۵ ارائه شده است. یادآوری می‌کنیم هر دو خودهمبستگی و خودهمبستگی جزئی فضایی مرتبه اول کاهش یافته است و به‌طور کلی به‌نظر می‌رسد که ساختاری در باقی‌مانده‌ها وجود ندارد به استثنای $r_{1,6}$ که در بازه $(-۲(۱۴ \times ۷۰)^{-۱} = -۰/۰۸۹, +۲(۱۴ \times ۷۰)^{-۱} = ۰/۰۸۹)$ قرار نمی‌گیرد، که تحقیقات بیشتر در رابطه با شکل‌های فصلی STARMA احتمال دارد مفید باشد. آزمون معنی‌داری پارامتر $\beta_{1,1}$ از طریق رابطه (۱.۱۰) انجام می‌شود. در این‌جا $n = ۱۴$، $T = ۷۱$، $K = ۲$، $S(\hat{\delta}^*) = ۶۵۰۴/۶۷۹$ و $S(\hat{\delta}) = ۶۴۵۷۰۲۶۶$ می‌باشد. با ترکیب این‌ها از طریق رابطه (۱.۱۰) مقدار تقریبی $F_{1,۹۹۲}$ برابر ۷/۳ به‌دست می‌آید. مقدار بحرانی F در $\alpha = ۰/۰۱$ با ۱ و ۹۹۲ درجه آزادی تقریبا ۶/۷ است. بنابراین، $\beta_{1,1}$ با ۹۹٪ اطمینان معنی‌دار است.

جدول ۳: تابع‌های همبستگی فضایی_زمانی سری تفاضلی شده

(الف) خودهمبستگی‌های فضایی_زمانی $(r_{s,k})$

تأخیر فضایی / تأخیر زمانی	۰	۱	۲	۳
۱	−۰/۴۸۴	۰/۰۰۷	۰/۰۴۱	۰/۰۱۳
۲	۰/۰۲۳	−۰/۰۳۸	۰/۰۱۲	−۰/۰۲۷
۳	−۰/۰۱۷	۰/۰۰۴	−۰/۰۴۵	۰/۰۴۹
۴	۰/۰۲۶	۰/۰۱۳	۰/۰۱۹	−۰/۰۳۸
۵	−۰/۰۵۶	۰/۰۳۹	۰/۰۰۵	۰/۰۱۷
۶	۰/۰۴۹	−۰/۰۷۴	۰/۰۴۵	−۰/۰۲۱
۷	−۰/۰۰۳	۰/۰۱۵	−۰/۰۹۱	−۰/۰۱۸
۸	−۰/۰۳۲	۰/۰۱۵	۰/۰۵۳	۰/۰۳۷

(ب)خودهمبستگی‌های جزئی فضایی_زمانی $(\phi_{s,s;k,k})$

تأخیر فضایی / تأخیر زمانی	۰	۱	۲	۳
۱	−۰/۴۸۴	۰/۰۶۸	۰/۰۰۷	۰/۰۲۰
۲	−۰/۲۸۱	۰/۰۱۵	۰/۰۳۴	−۰/۰۱۱
۳	−۰/۱۹۶	۰/۰۶۸	−۰/۰۲۶	۰/۰۷۰
۴	−۰/۱۱۱	۰/۰۱۵	−۰/۰۲۵	۰/۰۲۸
۵	−۰/۱۴۰	۰/۰۰۴	−۰/۰۲۱	۰/۰۳۰
۶	−۰/۰۹۲	۰/۰۲۵	۰/۰۷۱	−۰/۰۰۷
۷	−۰/۰۴۸	۰/۱۳۸	−۰/۰۳۶	−۰/۰۵۵
۸	−۰/۰۷۳	۰/۰۰۸	−۰/۰۱۱	۰/۰۰۸

چون خودهمبستگی‌ها و خودهمبستگی‌های جزئی فضا_زمان عدم وجود ساختار مشخصی را نشان می‌دهند (با توجه به استثنای ذکر شده) و از آن‌جایی که تمام پارامترها از لحاظ آماری معنی‌دار هستند مدل

$$\boldsymbol{y}_t - \boldsymbol{y}_{t-۱} = -۰/۸۱۲\varepsilon_{t-۱} + ۰/۰۹۲L^{(۱)}\varepsilon_{t-۱} + \varepsilon_t$$

پذیرفته می‌شود. این مدل در حال حاضر آماده است تا به عنوان یک تابع پیش‌بینی برای تعداد بازداشت‌های ضرب و شتم در ۱۴ ناحیه شمال شرق بوستون به‌کار گرفته شود.

جدول ۴: تابع‌های خودهمبستگی فضایی‌ـ زمانی باقی‌مانده‌ها از مدل $\boldsymbol{y}_t - \boldsymbol{y}_{t-۱} = -۰/۸۰۳\varepsilon_{t-۱} + \varepsilon_t$

(الف) خودهمبستگی‌های فضایی‌ـ زمانی $(r_{s,k})$

تأخیر فضایی	۰	۱	۲	۳
تأخیر زمانی				
۱	۰/۰۲۴	۰/۰۸۴	۰/۰۵۱	۰/۰۵۸
۲	۰/۰۳۱	۰/۰۱۶	۰/۰۲۳	۰/۰۲۷
۳	−۰/۰۰۷	۰/۰۲۷	−۰/۰۲۶	۰/۰۴۶
۴	−۰/۰۰۲	۰/۰۳۸	۰/۰۲۱	۰/۰۱۶
۵	−۰/۰۵۶	۰/۰۲۴	۰/۰۲۹	−۰/۰۱۴
۶	۰/۰۰۰	−۰/۰۶۱	۰/۰۲۶	−۰/۰۴۱
۷	−۰/۰۳۱	−۰/۰۰۹	−۰/۰۶۲	−۰/۰۲۶
۸	−۰/۰۴۶	۰/۰۱۱	۰/۰۳۰	۰/۰۲۱

(ب) خودهمبستگی‌های جزئی فضایی‌ـ زمانی $(\phi_{s,s;k,k})$

تأخیر فضایی	۰	۱	۲	۳
تأخیر زمانی				
۱	۰/۰۲۴	۰/۱۳۳	۰/۰۴۲	۰/۰۵۵
۲	۰/۰۱۹	−۰/۰۰۷	۰/۰۱۶	۰/۰۱۰
۳	−۰/۰۱۳	۰/۰۲۸	−۰/۰۴۲	۰/۰۴۵
۴	−۰/۰۰۹	۰/۰۵۸	۰/۰۱۵	−۰/۰۴۴
۵	−۰/۰۶۰	۰/۰۴۷	۰/۰۳۱	−۰/۰۳۵
۶	−۰/۰۰۳	−۰/۰۹۴	۰/۰۳۹	−۰/۰۵۰
۷	−۰/۰۲۲	۰/۰۰۳	−۰/۰۶۴	−۰/۰۲۱
۸	−۰/۰۴۳	۰/۰۶۴	۰/۰۳۹	۰/۰۴۲

بحث و نتیجه‌گیری

در این مقاله، به معرفی مدل‌های اتورگرسیو‌ـ میانگین متحرک فضایی‌ـ زمانی (STARMA) و استفاده از این مدل برای پیش‌بینی سری در آینده پرداختیم. روش پیشنهاد شده با استفاده از شبیه‌سازی مدل STARMA$(۱, ۲; ۱, ۱)$ مورد بررسی قرار گرفت. نتایج نشان داد که روش پیشنهادی توانایی خوبی در پیش‌بینی فرایندهای فضایی‌ـ زمانی را دارد. هم‌چنین با مدل‌بندی داده‌های جرم گزارش شده در شهر بوستون از ایالت ماساچوست آمریکا، مفید بودن مدل STARMA را نشان دادیم.

جدول ۵: تابع‌های خودهمبستگی فضایی‌ـ زمانی باقی‌مانده‌ها از مدل $\boldsymbol{y}_t - \boldsymbol{y}_{t-۱} = -۰/۸۱۲\varepsilon_{t-۱} + ۰/۰۹۲L^{(۱)}\varepsilon_{t-۱} + \varepsilon_t$

(الف) خودهمبستگی‌های فضایی‌ـ زمانی $(r_{s,k})$

تأخیر فضایی	۰	۱	۲	۳
تأخیر زمانی				
۱	۰/۰۲۳	۰/۰۴۵	۰/۰۳۶	۰/۰۴۶
۲	۰/۰۲۳	۰/۰۶۲	۰/۰۱۴	۰/۰۰۸
۳	−۰/۰۱۰	−۰/۰۱۴	−۰/۰۴۳	۰/۰۴۱
۴	−۰/۰۰۴	۰/۰۲۹	۰/۰۱۵	−۰/۰۴۶
۵	−۰/۰۵۳	۰/۰۲۰	۰/۰۳۴	−۰/۰۳۶
۶	۰/۰۰۷	−۰/۱۲۰	۰/۰۴۲	−۰/۰۵۰
۷	−۰/۰۲۳	−۰/۰۲۳	−۰/۰۶۲	−۰/۰۲۳
۸	−۰/۰۴۴	۰/۰۳۶	۰/۰۴۱	۰/۰۴۰

(ب) خودهمبستگی‌های جزئی فضایی‌ـ زمانی $(\phi_{s,s;k,k})$

تأخیر فضایی	۰	۱	۲	۳
تأخیر زمانی				
۱	۰/۰۲۳	۰/۰۳۰	۰/۰۳۶	۰/۰۴۲
۲	۰/۰۲۸	−۰/۰۲۶	۰/۰۱۰	۰/۰۱۱
۳	−۰/۰۰۸	−۰/۰۰۹	−۰/۰۳۶	۰/۰۳۰
۴	−۰/۰۰۱	۰/۰۱۶	۰/۰۱۴	−۰/۰۲۹
۵	−۰/۰۵۵	۰/۰۰۳	۰/۰۲۷	−۰/۰۲۸
۶	−۰/۰۰۳	−۰/۰۷۴	۰/۰۲۶	−۰/۰۵۱
۷	−۰/۰۳۲	−۰/۰۲۲	−۰/۰۵۸	−۰/۰۳۴
۸	−۰/۰۴۵	۰/۰۰۴	۰/۰۳۴	۰/۰۱۵

مراجع

محمدزاده م، (۱۳۹۴). *آمار فضایی و کاربردهای آن*، چاپ دوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

Cheng, T., Wang, J. Q., Li, X., and Zhang, W. (2008). A hybrid approach to model nonstationary space-time series, *Remote sensing and spatial information sciences*. Vol. XXXVII. Part B2.

Crespo, J. L., Zorrilla, M., Bernardos, P., and Mora, E. (2007). A new image prediction model based on spatio-temporal techniques, *The Visual Computer*, **23**(6), 419- 431.

Dai, Y., and Billard, L. (1998). A space□time bilinear model and its identification, *Journal of Time Series Analysis*, **19**(6), 657-679.

Epperson, B. K. (2000). Spatial and space–time correlations in ecological models, *Ecological modelling*, **132**(1), 63-76.

Jenkins, G. M., and Watts, D. G. (1968). *Spectral analysis and its applications*. Holden- Day, Michigan.

Kamarianakis, Y., and Prastacos, P. (2005). Space–time modeling of traffic flow, *Computers & Geosciences*, **31**(2), 119-133.

Lee, C. Y. (2004). *An Integrated Environment for Analyzing STARMA Models*, Department of Foresty, Michigan: Michigan State University, (December). (http://fried. for. msu. edu/ieast-manual/iemanual.pdf).

Lee, C. Y. (2005). *Space-time modeling and application to emerging infectious diseases*, Doctorate Dissertation, Supervisor: B.K. Epperson, Michigan State University, Department of Foresty.

Martin, R. L., and Oeppen, J. E. (1975). The identification of regional forecasting models using space: time correlation functions, *Transactions of the Institute of British Geographers*, **66**, 95-118.

Pace, R. K., Barry, R., Clapp, J. M., and Rodriquez, M. (1998). Spatiotemporal autoregressive models of neighborhood effects, *The Journal of Real Estate Finance and Economics*, **17**(1), 15-33.

Pfeifer, P. E., and Deutrch, S. J. (1980). A three-stage iterative procedure for space-time modeling phillip, *Technometrics*, **22**(1), 35-47.

Pfeifer, P. E., and Deutsch, S. J. (1981). Variance of the sample space-time autocorrelation function, *Journal of the Royal Statistical Society. Series B (Methodological)*, **43**(1), 28- 33.

دومین سمینار آمار فضایی و کاربردهای آن
۳ - ۴ آبان ۱۳۹۶

یک الگوریتم تقریبی جدید برای تحلیل درستنمایی مدل‌های آمیخته خطی تعمیم‌یافته فضایی با متغیرهای پنهان چوله نرمال

فاطمه حسینی*، امید کریمی

گروه آمار، دانشگاه سمنان

چکیده: در بیشتر مطالعات پیرامون مدل‌های آمیخته خطی تعمیم یافته فضایی به صورت یک فرض استاندارد فرض می‌شود متغیرهای پنهان فضایی دارای توزیع نرمال هستند. اگرچه فرض نرمال بودن متغیرهای پنهان فضایی باعث سهولت در محاسبات می‌گردد اما تخطی از این فرض به راحتی باعث نامعتبر کردن محاسبات آماری خواهد شد. در این مقاله توزیع چوله نرمال بسته که انعطاف‌پذیرتر و شامل توزیع نرمال است برای متغیرهای پنهان فضایی در نظر گرفته می‌شود. تابع درستنمایی مدل‌های آمیخته خطی تعمیم‌یافته فضایی شامل انتگرال‌هایی با بعد بالا است به طوری که این تابع شکل بسته‌ای ندارد و به راحتی قابل ارزیابی نیست. در این مقاله براساس یک رهیافت تقریبی و الگوریتم گرادیانت بیشینه‌سازی امیدریاضی یک الگوریتم جدید برای تحلیل مدل‌های آمیخته خطی تعمیم‌یافته فضایی با متغیرهای پنهان چوله نرمال بسته معرفی خواهد شد.

واژه‌های کلیدی: مدل‌های آمیخته خطی تعمیم‌یافته فضایی، چوله نرمال بسته، رهیافت درستنمایی تقریبی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 91B72، 62J12.

۱ مقدمه

مدل‌های خطی تعمیم‌یافته[۱] (GLM's) توسط مک‌کلاح و نلدر (۱۹۸۹) معرفی شده‌است. در این مدل‌ها فرض بر استقلال مشاهدات است و توسط یک تابع پیوند مشتق‌پذیر یکنوا بین متغیرهای کمکی و میانگین پاسخ ارتباط برقرار می‌شود. مدل‌های آمیخته خطی تعمیم‌یافته[۲] (GLMM's) تعمیمی از مدل‌های GLM's هستند که توسط بریسلو و کلیتون (۱۹۹۳) به‌کار گرفته شد. در این مدل‌ها استقلال مشاهدات به استقلال شرطی روی یک مجموعه از متغیرهای پنهان تبدیل می‌شود و همبستگی موجود بین داده‌ها با افزودن

*ارایه‌دهنده مقاله: فاطمه حسینی، fatemeh.hoseini@semnan.ac.ir

[۱] Generalized Linear Models

[۲] Generalized Linear Mixed Models

متغیرهای پنهان به مدل در نظر گرفته می‌شود. اگر این همبستگی از نوع مکانی باشد، مدل آمیخته خطی تعمیم‌یافته حاصل مدل آمیخته خطی تعمیم‌یافته فضایی[۳] (SGLMM's) نامیده می‌شود. این مدل‌ها اغلب برای مدل‌بندی مشاهدات پاسخ فضایی گسسته و یا رسته‌ای استفاده می‌گردد، به‌طوری که همبستگی فضایی داده‌ها توسط متغیرهای پنهان فضایی در مدل منظور می‌شود، (برای جزئیات بیشتر به دیگل و همکاران (۱۹۹۸) مراجعه شود).

از اهداف اصلی در SGLMM's برآورد پارامترهای مدل، برآورد متغیرهای پنهان در موقعیت‌های دارای مشاهده پاسخ و پیش‌گویی متغیرهای پنهان در موقعیت‌های مورد علاقه و فاقد مشاهده پاسخ می‌باشند. یک فرض استاندارد در بیشتر مطالعات پیرامون این مدل‌ها و در راستای راحتی محاسبات، در نظر گرفتن توزیع نرمال برای متغیرهای پنهان است. در صورتی که می‌توان برای افزایش دقت در محاسبات از توزیع‌های انعطاف‌پذیرتر و البته شامل توزیع نرمال مثل توزیع‌های چوله نرمال، چوله نرمال بسته و چوله تی استفاده کرد. توزیع چوله نرمال یک متغیره اولین بار توسط آزالینی (۱۹۸۵) و توزیع چوله نرمال چندمتغیره و خواص آن توسط آزالینی و کپیتانیو (۱۹۹۹) و آزالینی (۱۹۹۶) معرفی شده‌اند. توزیع چوله نرمال بسته[۴] (CSN) توسط آزالینی و کپیتانیو (۲۰۰۳)، آزالینی (۲۰۰۴a) و آزالینی و کپیتانیو (۲۰۰۴b) معرفی شده‌است. به‌کار گرفتن توزیع چوله نرمال بسته برای متغیرهای پنهان در SGLMM's اولین بار در حسینی و همکاران (۲۰۱۱) و حسینی و محمدزاده (۲۰۱۲) پیشنهاد و مورد بررسی قرار گرفت. خانواده توزیع چوله نرمال بسته از خانواده توزیع چوله نرمال و نرمال انعطاف‌پذیرتر است و خواص خوبی مشابه توزیع نرمال مثل بسته بودن نسبت به ترکیبات خطی، شرطی‌سازی و حاشیه‌سازی دارد. برای جزئیات بیشتر و کاربردهای این خانواده از توزیع‌ها می‌توان به کریمی و محمدزاده (۲۰۱۱) ، کریمی و محمدزاده (۲۰۱۲) ،کریمی و همکاران (۲۰۱۰) و ریمستاد و امره (۲۰۱۴) مراجعه نمود.

به دلیل وجود متغیرهای پنهان در SGLMM's و گسسته بودن مشاهدات معمولا تابع درستنمایی در این مدل‌ها شکل بسته‌ای ندارد و اغلب از الگوریتم‌های عددی برای به‌دست آوردن برآوردهای ماکسیمم درستنمایی استفاده می‌شود. از جمله مطالعات که از الگوریتم‌های عددی برای محاسبه برآوردهای ماکسیمم درستنمایی در SGLMM's استفاده کرده‌اند، می‌توان به ژانگ (۲۰۰۲)؛ باغیشنی و همکاران (۲۰۱۲)؛ ترابی (۲۰۱۳، ۲۰۱۵)؛ باغیشنی و همکاران (۲۰۱۱) و ایوانگلو و همکاران (۲۰۱۱). اشاره نمود. رو و مارتینو (۲۰۰۷)، رو و همکاران (۲۰۰۹) و ایدسویک و همکاران (۲۰۰۹) روش‌های تقریبی را به جای الگوریتم‌های مونت کارلویی در تحلیل بیزی SGLMM's معرفی نمودند.

در این مقاله مدل‌های SGLM با متغیرهای پنهان چوله نرمال بسته برای مدل‌بندی پاسخ‌های فضایی گسسته به کار گرفته می‌شود و براساس یک رهیافت تقریبی و الگوریتم گرادیانت بیشینه‌سازی امیدریاضی[۵] (EMG) یک الگوریتم جدید برای به‌دست آوردن برآوردهای ماکسیمم درستنمایی پارامترهای مدل در نظر گرفته شده معرفی می‌گردد. بخش‌بندی مقاله به این‌صورت است که در بخش دوم مروری بر توزیع چوله نرمال بسته و خواص آن می‌شود. در بخش سوم SGLMM's با متغیرهای پنهان چوله نرمال بسته بیان می‌گردد. در بخش چهار به رهیافت درستنمایی تقریبی پیشنهادی پرداخته می‌شود و در بخش پنج مدل و رهیافت پیشنهادی برروی یک مجموعه داده واقعی پیاده‌سازی می‌گردد. در نهایت نتایج و پیشنهادات در بخش شش ارائه شده‌است.

[۳] Spatial Generalized Linear Mixed Models
[۴] Closed Skew Normal
[۵] Expectation Maximization Gradient

۲ توزیع چوله نرمال بسته

فرض کنید $\boldsymbol{x}$ یک بردار تصادفی n_dبعدی با توزیع چوله نرمال بسته باشد، در این‌صورت تابع چگالی آن به‌صورت

$$f_{n_d,q}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma,D,\boldsymbol{\nu},\Delta) = k\phi_n(\boldsymbol{x};\boldsymbol{\mu},\Sigma)\,\Phi_q(D(\boldsymbol{x}-\boldsymbol{\mu});\boldsymbol{\nu},\Delta), \tag{۱.۲}$$

می‌باشد، که در آن $k = \Phi_q^{-1}(\boldsymbol{0};\boldsymbol{\nu},\Delta+D\Sigma D^\top)$ و $\Phi_q(\cdot;\boldsymbol{\nu},\Delta)$ یک تابع توزیع تجمعی نرمال q بعدی با میانگین $\boldsymbol{\nu}$ و ماتریس کوواریانس Δ است. D یک ماتریس $q\times n$ می‌باشد که عناصرش پارامترهای چولگی هستند. $\phi_n(\cdot;\boldsymbol{\mu},\Sigma)$ یک تابع چگالی نرمال n متغیره با میانگین $\boldsymbol{\mu}\in R^n$ و ماتریس کوواریانس Σ است. در این‌صورت از نماد $\mathrm{CSN}_{n,q}(\boldsymbol{\mu},\Sigma,D,\boldsymbol{\nu},\Delta)$ برای نشان دادن توزیع چوله نرمال بسته استفاده می‌گردد. اگر D ماتریس صفر باشد تابع چگالی (۱.۲) به یک تابع چگالی نرمال n متغیره تبدیل خواهد شد. اگر $q=1$، $\boldsymbol{\nu}=0$، $\Delta=1$ و $D=\boldsymbol{\lambda}^\top\Sigma^{-\frac{1}{2}}$ تابع چگالی (۱.۲) به تابع چگالی چوله نرمال معرفی شده توسط آزالینی و کپیتانیو (۱۹۹۹) تبدیل می‌شود. این توزیع دارای خواصی به شرح زیر است که برای توضیحات بیشتر می‌توان به دامینگوس و همکاران (۲۰۰۳) و گنزالس و همکاران (۲۰۰۴b) مراجعه کرد.

امیدریاضی توزیع چوله نرمال بسته به‌صورت

$$E(\boldsymbol{x}) = \boldsymbol{\mu} + \Sigma D^\top\boldsymbol{\psi}, \tag{۲.۲}$$

به‌دست آورده شده‌است، که در آن $\boldsymbol{\psi} = \Phi_q^*(\boldsymbol{r},\boldsymbol{\nu},\Delta+D\Sigma D^\top)/\Phi_q(\boldsymbol{r},\boldsymbol{\nu},\Delta+D\Sigma D^\top)|_{\boldsymbol{r}=0}$، و برای هر ماتریس معین مثبت Ω، $\Phi_q^*(\boldsymbol{r};\boldsymbol{\nu},\Omega) = [\nabla_r\Phi_q(\boldsymbol{r};\boldsymbol{\nu},\Omega)]^\top$ به طوری‌که $\nabla_r = (\partial/\partial r_1,\cdots,\partial/\partial r_q)^\top$. واریانس X به‌صورت

$$\mathrm{Var}(\boldsymbol{x}) = \Sigma + \Sigma D^\top\boldsymbol{\xi}D\Sigma - \Sigma D^\top\boldsymbol{\psi}\boldsymbol{\psi}^\top D\Sigma, \tag{۳.۲}$$

محاسبه می‌گردد، که در آن

$$\boldsymbol{\xi} = \Phi_q^{**}(\boldsymbol{r},\boldsymbol{\nu},\Delta+D\Sigma D^\top)/\Phi_q(\boldsymbol{r},\boldsymbol{\nu},\Delta+D\Sigma D^\top)|_{\boldsymbol{r}=0},\quad \Phi_q^{**}(\boldsymbol{r};\boldsymbol{\nu},\Omega) = [\nabla_r\nabla_r^\top\Phi_q(\boldsymbol{r};\boldsymbol{\nu},\Omega)]^\top.$$

توزیع چوله نرمال بسته تحت ترکیبات خطی بسته است. فرض کنید A یک ماتریس ثابت $r\times n(r\le n)$ باشد، به‌طوری که $AA^\top$ نامنفرد است و $\boldsymbol{b}\in R^r$ در این‌صورت

$$A\boldsymbol{x}+\boldsymbol{b} \sim \mathrm{CSN}_{r,q}(A\boldsymbol{\mu}+\boldsymbol{b}, A\Sigma A^\top, D\Sigma A^\top(A\Sigma A^\top)^{-1},\boldsymbol{\nu},\Delta_1). \tag{۴.۲}$$

که در آن $\Delta_1 = \Delta + D\Sigma D^\top - D\Sigma A^\top(A\Sigma A^\top)^{-1}A\Sigma D^\top$. این توزیع تحت حاشیه‌سازی بسته است. فرض کنید بردار $\boldsymbol{x}$ به‌صورت $\boldsymbol{x} = (\boldsymbol{x}_1^\top,\boldsymbol{x}_2^\top)^\top$ افراز شده اسا که در آن $\boldsymbol{x}_1$ بردار k بعدی و $\boldsymbol{x}_2$ بردار $n-k$ بعدی است، در این‌صورت

$$\boldsymbol{x}_1 \sim \mathrm{CSN}_{k,q}(\boldsymbol{\mu}_1,\Sigma_{11},D^*,\boldsymbol{\nu},\Delta^*), \tag{۵.۲}$$

که در آن $D^* = D_1 + D_2\Sigma_{21}\Sigma_{11}^{-1}$، $\Delta^* = \Delta + D_2\Sigma_{22\cdot1}D_2^\top$، $\Sigma_{22\cdot1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ به طوری‌که $\boldsymbol{\mu}_1$، Σ_{11}، Σ_{22}، Σ_{12}، Σ_{21}، D_1 and D_2 از افراز $\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}$, $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$ به دست می‌آیند و $D = (D_1,\ D_2)$.
توزیع چوله نرمال بسته تحت شرطی کردن بسته است. توزیع $\boldsymbol{x}_2$ به شرط $\boldsymbol{x}_1$ به‌صورت

$$(\boldsymbol{x}_2|\boldsymbol{x}_1) \sim \mathrm{CSN}_{n-k,q}(\boldsymbol{\mu}_2+\Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1),\Sigma_{22\cdot1},D_2,\boldsymbol{\nu}-D^*(\boldsymbol{x}_1-\boldsymbol{\mu}_1),\Delta), \tag{۶.۲}$$

به‌دست می‌آید.

۱.۲ SGLMM's چوله نرمال بسته

فرض کنید $\boldsymbol{x} = (x_1, \cdots, x_n)^\top$ متغیرهای پنهان فضایی در n موقعیت $\{s_1, \cdots, s_n\}$ با چگالی

$$\text{CSN}_{n,1}(H\boldsymbol{\beta}, \Sigma_\theta, \boldsymbol{\lambda}^\top \Sigma_\theta^{-\frac{1}{2}}, 0, 1),$$

باشند. از (۱.۲) داریم

$$f(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{2}{(2\pi)^{n/2}|\Sigma_\theta|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - H\boldsymbol{\beta})^\top \Sigma_\theta^{-1}(\boldsymbol{x} - H\boldsymbol{\beta})\right) \cdot \Phi(\boldsymbol{\lambda}^\top \Sigma_\theta^{-\frac{1}{2}}(\boldsymbol{x} - H\boldsymbol{\beta})). \quad (7.2)$$

که در آن $H\boldsymbol{\beta}$ شامل ماتریس متغیرهای کمکی H با بعد $n \times (p+1)$ و $p+1$ پارامتر رگرسیونی $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^\top$ می‌باشد. ماتریس $\Sigma_{\boldsymbol{\theta}}$ یک ماتریس معین مثبت $n \times n$ است، که معمولا با دو پارامتر مقیاس و همبستگی به‌صورت $\boldsymbol{\theta} = (\sigma, \varphi)^\top$ ساختار آن معلوم می‌گردد. $\boldsymbol{\lambda}$ پارامتر چولگی است که به شکل $\boldsymbol{\lambda} = \lambda_0 1$ فرض شده است. پس بردار پارامترهای مدل را می‌توان به‌صورت $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \sigma, \varphi, \lambda_0)^\top$ در نظر گرفت. فرض کنید $\{s_1, \cdots, s_k\}$ موقعیت‌های دارای مشاهده و $\{s_{k+1}, \cdots, s_n\}$ موقعیت‌های فاقد مشاهده باشند. متغیرهای پنهان فضایی در k موقعیت دارای مشاهده را با $\boldsymbol{x}^{obs} = A\boldsymbol{x}$ نشان می‌دهیم، که در آن $A = [I_{k\times k}|\mathbf{0}_{k\times n-k}]$ به عبارت دیگر بردار متغیرهای پنهان را به دو بردار $\boldsymbol{x} = (\boldsymbol{x}^{obs\top}, \boldsymbol{x}^{pred\top})^\top$ افراز می‌کنیم. $\boldsymbol{x}^{\boldsymbol{pred}}$ متغیرهای پنهان در موقعیت‌های مورد نظر برای پیش‌گویی می‌باشند. همچنین فرض کنید $\boldsymbol{y}^\top = (y_1, \cdots, y_k)$ بردار مشاهدات متغیر پاسخ فضایی در موقعیت‌های $\{s_1, \cdots, s_k\}$ باشند و طبق مک‌کلاچ و نلدر (۱۹۸۹) فرض شده‌است که به طور شرطی روی متغیرهای پنهان مستقل و متعلق به خانواده‌ی نمایی است، بنابراین داریم

$$f(y_i|x_i) = \exp\{y_i x_i - b(x_i) + c(y_i)\}, \quad i = 1, \ldots, k, \quad (8.2)$$

که در آن $b(\cdot)$ و $c(\cdot)$ توابع معلومند. در این‌صورت $E(y_i|x_i)$ و x_i توسط تابع پیوند معلوم g به شکل $E(y_i|x_i) = g^{-1}(x_i)$ در ارتباط می‌باشند.

۳ برآورد ماکسیمم درستنمایی تقریبی مدل

بردار پارامترهای مدل را به‌صورت $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \sigma, \varphi, \lambda_0)^\top$ در نظر بگیرید، از (۷.۲) و (۸.۲) داریم

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\eta}) &= f(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\eta}) \\ &= \frac{2}{(2\pi)^{n/2}|\Sigma_\theta|^{1/2}} \Phi(\boldsymbol{\lambda}^\top \Sigma_\theta^{-\frac{1}{2}}(\boldsymbol{x} - H\boldsymbol{\beta})) \\ &\times \exp\left(\sum_{i=1}^{k}[y_i x_i - b(x_i) + c(y_i)] - \frac{1}{2}(\boldsymbol{x} - H\boldsymbol{\beta})^\top \Sigma_\theta^{-1}(\boldsymbol{x} - H\boldsymbol{\beta})\right). \end{aligned} \quad (1.3)$$

و بنابراین تابع درستنمایی به‌صورت

$$L(\boldsymbol{\eta}|\boldsymbol{y}) = \int \prod_{i=1}^{k} f(y_i|x_i) f(\boldsymbol{x}|\boldsymbol{\eta}) d\boldsymbol{x}. \quad (2.3)$$

به‌دست می‌آید. بعد انتگرال (۲.۳) برابر تعداد متغیرهای پنهان مدل است و لذا این تابع شکل بسته‌ای ندارد.

برای محاسبه برآورد ماکسیمم درستنمایی مدل‌هایی که شامل متغیرهای پنهان هستند و تابع درستنمایی پیچیده‌ای دارند استفاده از الگوریتم بیشینه‌سازی امیدریاضی[۶] (EM) پیشنهاد شده است (دمستر و همکاران ، ۱۹۷۷). در برخی مسئله‌ها همگرایی این

[۶] Expectation Maximization

الگوریتم معمولا به راحتی اتفاق نمی‌افتد و به همین منظور لانجی (۱۹۹۵) الگوریتم گرادیانت بیشینه‌سازی امیدریاضی را معرفی کرد و نشان داد در مدل‌های شامل متغیرهای پنهان و شامل پارامترهای زیاد استفاده از این الگوریتم برای محاسبه‌ی برآوردهای ماکسیمم درستنمایی بهتر است و همگرایی‌ها سریع‌تر حاصل می‌شود. در این بخش با استفاده از قضیه‌ی زیر و الگوریتم EMG یک الگوریتم جدید برای به‌دست آوردن برآوردهای ماکسیمم درستنمایی در SGLMM's با متغیرهای پنهان فضایی چوله نرمال بسته ارائه می‌گردد.

قضیه ۱.۳. (حسینی و همکاران ، ۲۰۱۱) *فرض کنید متغیرهای پنهان فضایی در SGLMM's دارای توزیع چوله نرمال بسته*

$$\boldsymbol{x}|\boldsymbol{\eta} \sim \mathrm{CSN}_{n,1}(H\boldsymbol{\beta}, \Sigma_\theta, \boldsymbol{\lambda}^\top \Sigma_\theta^{-\frac{1}{2}}, 0, 1)$$

باشد و $\boldsymbol{x}$ به‌صورت $\boldsymbol{x} = (\boldsymbol{x}^{obs\top}, \boldsymbol{x}^{pred\top})^\top$ افراز شده باشد. همچنین فرض کنید متغیرهای پاسخ گسسته فضایی متعلق به خانواده‌ی نمایی k باشند. با در نظر گرفتن $\pi(y_i|x_i) = \exp\{y_i x_i - b(x_i)\},\ i = 1, \cdots k$ و با خطی‌سازی قسمت درستنمایی $f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x}|\boldsymbol{\eta})$ حول یک مقدار ثابت $\boldsymbol{x}$ داریم $A = [I_{k\times k}|\mathbf{0}_{k\times n-k}]$ و $\boldsymbol{x}^{obs} = A\boldsymbol{x}$.

$$\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta} \approx \mathrm{CSN}_{n,1}(\boldsymbol{\mu}_{x|y,\eta}, \Sigma_{x|y,\eta}, D_{x|y,\eta}, \nu_{x|y,\eta}, 1), \tag{۳.۳}$$

که در آن

$$\boldsymbol{\mu}_{x|y,\eta} = H\boldsymbol{\beta} + \Sigma_\theta A^\top (A\Sigma_\theta A^\top + P)^{-1}(z(\boldsymbol{y}, \boldsymbol{x}^{obs}) - AH\boldsymbol{\beta}),$$

و $z_i(y_i, x_i) = [y_i - b'(x_i) + x_i b''(x_i)]/b''(x_i)$ خطی‌سازی قسمت درستنمایی در مقدار ثابت از $\boldsymbol{x}$ است. $i = 1, \ldots, k$، ماتریس $R = A\Sigma_\theta A^\top + P$ و $P = P(\boldsymbol{x})$ یک ماتریس قطری با عناصر روی قطر $P(i,i) = 1/b''(x_i)$ می‌باشد. به‌علاوه

$$\begin{aligned}
\Sigma_{x|y,\eta} &= \Sigma_\theta - \Sigma_\theta A^\top (A\Sigma_\theta A^\top + P)^{-1} A\Sigma_\theta, \\
D_{x|y,\eta} &= \boldsymbol{\lambda}^\top \Sigma_\theta^{-\frac{1}{2}}, \\
\nu_{x|y,\eta} &= -\boldsymbol{\lambda}^\top \Sigma_\theta^{-\frac{1}{2}} \Sigma_\theta A^\top (A\Sigma_\theta A^\top + P)^{-1}(z(\boldsymbol{y}, \boldsymbol{x}^{obs}) - AH\boldsymbol{\beta}) \\
&= \boldsymbol{\lambda}^\top \Sigma_\theta^{-\frac{1}{2}} (H\boldsymbol{\beta} - \boldsymbol{\mu}_{x|y,\eta}).
\end{aligned}$$

اکنون یک الگوریتم برای محاسبه‌ی برآوردهای ماکسیمم درستنمایی مدل پیشنهادی به شرح زیر می‌باشد:

۱ـ مقدار اولیه پارامترها را $\boldsymbol{\eta}^{(0)}$ قرار دهید، به‌طوری که $L(\boldsymbol{\eta}^{(0)}|\boldsymbol{y}) > 0$ و قرار دهید $m = 0$.

۲ـ مقدار اولیه متغیرهای پنهان را $\boldsymbol{x}^{(0)}$ قرار دهید، برای مثال فرض کنید $\boldsymbol{x}^{(0)} = E(\boldsymbol{x}|\boldsymbol{\eta})$ و قرار دهید $d = 0$.

۳ـ از قضیه‌ی ۱.۳ عبارت

$$\hat{f}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta}) = \mathrm{CSN}_{n,1}(\hat{\boldsymbol{\mu}}_{x|y,\eta}(\boldsymbol{x}^{(d)}), \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{(d)}), D_{x|y,\eta}, \hat{\boldsymbol{\nu}}_{x|y,\eta}(\boldsymbol{x}^{(d)}), 1).$$

را محاسبه کنید.

۴ـ از رابطه‌ی (۲.۲) قرار دهید $\boldsymbol{x}^{(d+1)} = \hat{E}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta}) = \hat{\boldsymbol{\mu}}_{x|y,\eta}(\boldsymbol{x}^{(d)}) + \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{(d)}) D_{x|y,\eta}^\top \hat{\psi}$ که در آن

$$\hat{\psi} = \psi(0, \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{(d)}), D_{x|y,\eta}, \hat{\nu}_{x|y,\eta}(\boldsymbol{x}^{(d)}), 1)$$

۵ـ قرار دهید $d = d + ۱$ به مرحله سوم برگردید و تا رسیدن به همگرایی ادامه دهید.

۶ـ $\boldsymbol{\eta}^{(m+۱)}$ را طوری انتخاب کنید

$$\begin{aligned}\boldsymbol{\eta}^{(m+۱)} &= \boldsymbol{\eta}^{(m)} - \left[\frac{\partial^۲}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}\hat{E}\Big\{(\ln f(\boldsymbol{x}|\boldsymbol{\eta}))|\boldsymbol{y}\Big\}\right]^{-۱}_{\boldsymbol{\eta}=\boldsymbol{\eta}^{(m)}}\left[\frac{\partial}{\partial\boldsymbol{\eta}}\hat{E}\Big\{(\ln f(\boldsymbol{x}|\boldsymbol{\eta}))|\boldsymbol{y}\Big\}\right]_{\boldsymbol{\eta}=\boldsymbol{\eta}^{(m)}}\\ &= \boldsymbol{\eta}^{(m)} - \left[\frac{\partial^۲}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}\int \ln f(\boldsymbol{x}|\boldsymbol{\eta})\hat{f}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})d\boldsymbol{x}\right]^{-۱}_{\boldsymbol{\eta}=\boldsymbol{\eta}^{(m)}}\left[\frac{\partial}{\partial\boldsymbol{\eta}}\int \ln f(\boldsymbol{x}|\boldsymbol{\eta})\hat{f}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})d\boldsymbol{x}\right]_{\boldsymbol{\eta}=\boldsymbol{\eta}^{(m)}}\end{aligned} \tag{۴.۳}$$

۷ـ قرار دهید $m = m + ۱$ و به مرحله ۶ برگردید و تا رسیدن به همگرایی ادامه دهید.

امید ریاضی‌های شرطی در معادله‌ی (۴.۳) قابل محاسبه هستند. از (۷.۲) داریم

$$\begin{aligned}\hat{E}\left[\{\ln f(\boldsymbol{x}|\boldsymbol{\eta})\}|\boldsymbol{y}\right] &= -\frac{۱}{۲}\ln|\Sigma_{\boldsymbol{\theta}}| - \frac{n}{۲}\ln ۲\pi + \ln ۲\\ &\quad -\frac{۱}{۲}E\left\{(\boldsymbol{x}-H\boldsymbol{\beta})^\top\Sigma_{\boldsymbol{\theta}}^{-۱}(\boldsymbol{x}-H\boldsymbol{\beta})|\boldsymbol{y}\right\}\\ &\quad +E\left\{\ln\Phi(\boldsymbol{\lambda}^\top\Sigma_\theta^{-\frac{۱}{۲}}(\boldsymbol{x}-H\boldsymbol{\beta}))|\boldsymbol{y}\right\}.\end{aligned} \tag{۵.۳}$$

برای به‌دست آوردن $E\left\{\ln\Phi(\boldsymbol{\lambda}^\top\Sigma_\theta^{-\frac{۱}{۲}}(\boldsymbol{x}-H\boldsymbol{\beta}))|\boldsymbol{y}\right\}$ در معادله (۵.۳) از روش‌های عددی استفاده می‌شود. داریم

$$E\left\{\ln\Phi(\boldsymbol{\lambda}^\top\Sigma_\theta^{-\frac{۱}{۲}}(\boldsymbol{x}-H\boldsymbol{\beta}))|\boldsymbol{y}\right\} = \int \ln\Phi(\boldsymbol{\lambda}^\top\Sigma_\theta^{-\frac{۱}{۲}}(\boldsymbol{x}-H\boldsymbol{\beta}))\hat{f}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})d\boldsymbol{x}.$$

این انتگرال نیز با استفاده از روش انتگرال‌گیری مونت کارلو قابل حل است. ابتدا نمونه‌های مونت کارلوی $\boldsymbol{x}^{(۱)}\cdots\boldsymbol{x}^{(N)}$ از $\hat{f}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})$ تولید می‌شوند، سپس

$$E\left\{\ln\Phi(\boldsymbol{\lambda}^\top\Sigma_\theta^{-\frac{۱}{۲}}(\boldsymbol{x}-H\boldsymbol{\beta}))|\boldsymbol{y}\right\} = \frac{۱}{N}\sum_{\ell=۱}^{N}\ln\Phi(\boldsymbol{\lambda}^\top\Sigma_\theta^{-\frac{۱}{۲}}(\boldsymbol{x}^{(\ell)}-H\boldsymbol{\beta})). \tag{۶.۳}$$

با در نظر گرفتن $\boldsymbol{w} = [(\boldsymbol{x}-H\boldsymbol{\beta})^\top|\boldsymbol{y}]$ از (۴.۲) و (۳.۳) بردار $\boldsymbol{w}$ دارای توزیع چوله نرمال بسته به‌صورت

$$\mathrm{CSN}_{n,۱}(\hat{\boldsymbol{\mu}}_w, \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^\star), \boldsymbol{\lambda}^\top\Sigma_\theta^{-\frac{۱}{۲}}, \hat{\nu}_{x|y,\eta}(\boldsymbol{x}^\star), ۱),$$

است، که در آن $\hat{\boldsymbol{\mu}}_w = (\hat{\boldsymbol{\mu}}_{x|y,\eta}(\boldsymbol{y},\boldsymbol{x}^\star) - H\boldsymbol{\beta})$.

قضیه ۲.۳. (رنچر ، ۲۰۰۸) *فرض کنید $\boldsymbol{w}$ یک بردار تصادفی با میانگین $\boldsymbol{\mu}$ و ماتریس کوواریانس Σ باشد. اگر A یک ماتریس متقارن ثابت باشد آن‌گاه*

$$E(\boldsymbol{w}^\top A\boldsymbol{w}) = \boldsymbol{\mu}^\top A\boldsymbol{\mu} + \mathrm{tr}(A\Sigma).$$

بنابراین چون $\left\{(\boldsymbol{x}-H\boldsymbol{\beta})^\top\Sigma_{\boldsymbol{\theta}}^{-۱}(\boldsymbol{x}-H\boldsymbol{\beta})|\boldsymbol{y}\right\}$ در (۵.۳) یک فرم درجه دو به شکل $\boldsymbol{w}^\top\Sigma_{\boldsymbol{\theta}}^{-۱}\boldsymbol{w}$ می‌باشد، با استفاده از قضیه‌ی (۲.۳) داریم

$$E(\boldsymbol{w}^\top\Sigma_{\boldsymbol{\theta}}^{-۱}\boldsymbol{w}) = E(\boldsymbol{w})^\top\Sigma_{\boldsymbol{\theta}}^{-۱}E(\boldsymbol{w}) + \mathrm{tr}(\Sigma_{\boldsymbol{\theta}}^{-۱}\mathrm{V}(\boldsymbol{w})) \tag{۷.۳}$$

که در آن $E(\boldsymbol{w})$ و $\mathrm{Var}(\boldsymbol{w})$ با استفاده از (۳.۲) و (۲.۲) به‌صورت

$$E(\boldsymbol{w}) = \hat{\boldsymbol{\mu}}_w + \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^\star)\Sigma_\theta^{-\frac{۱}{۲}}\boldsymbol{\lambda}\psi$$

$$\begin{aligned}\mathrm{Var}(\boldsymbol{w}) &= \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{\star}) + \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{\star})\Sigma_{\theta}^{-\frac{1}{2}}\boldsymbol{\lambda}\psi\psi^{\top}\lambda^{\top}\Sigma_{\theta}^{-\frac{1}{2}}\hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{\star}) \\ &- \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{\star})\Sigma_{\theta}^{-\frac{1}{2}}\boldsymbol{\lambda}\,\xi\lambda^{\top}\Sigma_{\theta}^{-\frac{1}{2}}\hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{\star}).\end{aligned}$$

محاسبه می‌شوند. اکنون با جایگذاری (۶.۳) و (۷.۳) در (۵.۳) امیدریاضی‌های شرطی در (۴.۳) به‌دست آورده می‌شوند. بعد از مشخص شدن برآوردهای ماکسیمم درستنمایی پارامترها می‌توان متغیرهای پنهان را در موقعیت‌های مورد علاقه برای پیش‌گویی به‌دست آورد. برای این‌منظور از تقریب چوله نرمال بسته چگالی شرطی $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})$ استفاده می‌گردد و پیش‌گوی مینیمم مربع خطای تقریبی به‌صورت زیر محاسبه می‌شود.

ابتدا فرض کنید $\boldsymbol{x}|\boldsymbol{\eta} \sim \mathrm{CSN}_{n,1}(H\boldsymbol{\beta}, \Sigma_{\theta}, \boldsymbol{\lambda}^{\top}\Sigma_{\theta}^{-\frac{1}{2}}, 0, 1)$ و $\boldsymbol{x}$ را به شکل $\boldsymbol{x} = (\boldsymbol{x}^{obs\top}, \boldsymbol{x}^{pred\top})^{\top}$ افراز کنید. از قضیه (۱.۳) داریم

$$\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta} \sim \mathrm{CSN}_{n,1}(\mu_{x|y,\eta}, \Sigma_{x|y,\eta}, D_{x|y,\eta}, \nu_{x|y,\eta}, \Delta_{x|y,\eta}),$$

که در آن از (۴.۲) $\mu_{x|y,\eta} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$، $\Sigma_{x|y,\eta} = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$، $D_{x|y,\eta} = [D_1 \quad D_2]$، $\nu_{x|y,\eta} = \nu_{x|y,\eta}$ و $\Delta_{x|y,\eta} = 1$ می‌باشند. از (۵.۲)

$$\boldsymbol{x}^{pred}|\boldsymbol{y},\boldsymbol{\eta} \sim \mathrm{CSN}_{n-k,1}(\mu_2, \Sigma_{22}, D_2 + D_1\Sigma_{12}\Sigma_{22}^{-1}, \nu_{x|y,\eta}, 1 + D_1\Sigma_{11\cdot 2}D_1^{\top})$$

که در آن $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ است. اکنون با به‌کار بردن رابطه‌ی (۲.۲) داریم

$$E(x^{pred}\boldsymbol{y}) = \boldsymbol{\mu}_2 + \Sigma_{22}(D_2 + D_1\Sigma_{12}\Sigma_{22}^{-1})^{\top}\psi, \tag{۸.۳}$$

که در آن

$$\psi = \frac{\Phi^{*}(0, \nu_{x|y,\eta}, 1 + D_1\Sigma_{11\cdot 2}D_1^{\top} + (D_2 + D_1\Sigma_{12}\Sigma_{22}^{-1})\Sigma_{22}(D_2 + D_1\Sigma_{12}\Sigma_{22}^{-1})^{\top})}{\Phi(0, \nu_{x|y,\eta}, 1 + D_1\Sigma_{11\cdot 2}D_1^{\top} + (D_2 + D_1\Sigma_{12}\Sigma_{22}^{-1})\Sigma_{22}(D_2 + D_1\Sigma_{12}\Sigma_{22}^{-1})^{\top})}.$$

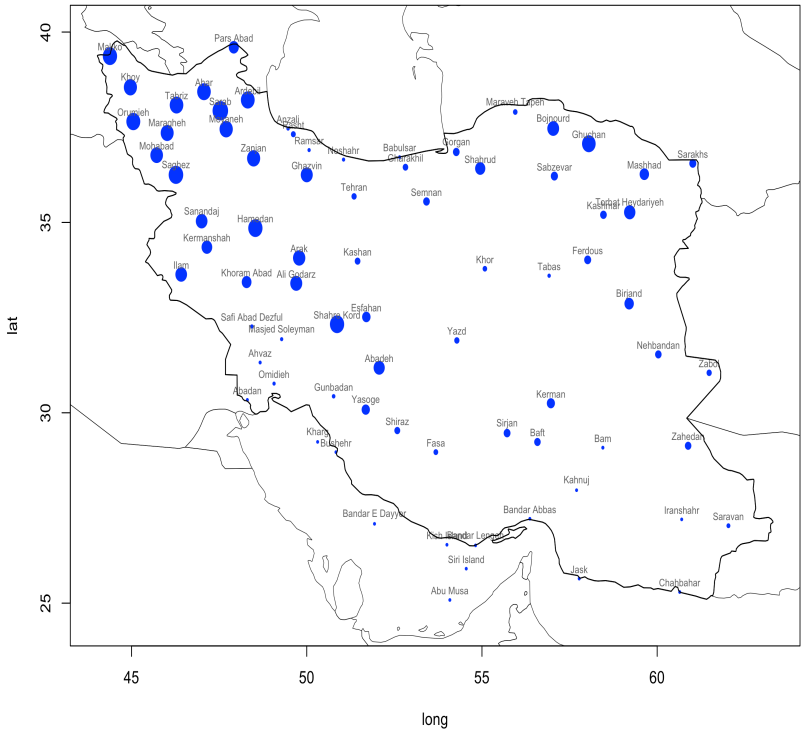
رابطه‌ی (۸.۳) یک پیش‌گوی مینیمم مربع خطای تقریبی می‌باشد.

۱.۳ داده‌های یخبندان ایران

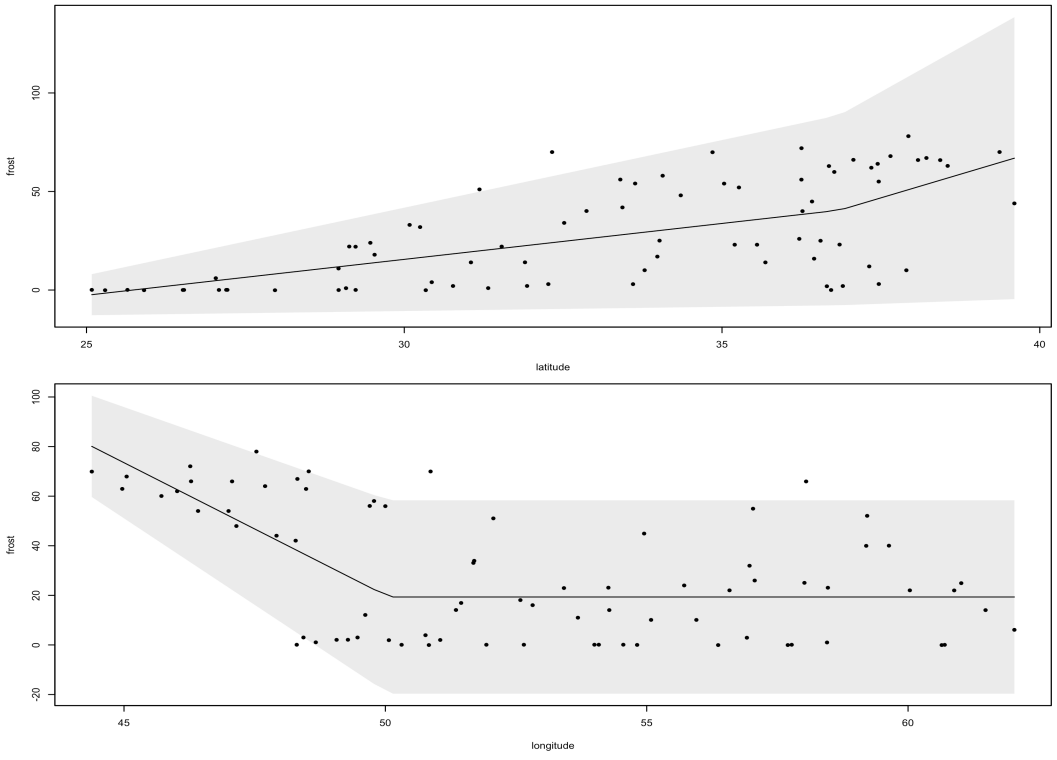
به دلیل گستردگی کشور ایران از انواع اقلیم‌های آب و هوایی برخوردار است. به‌خصوص دو رشته کوه زاگرس و البرز تاثیر بسزایی برروی اقلیم کشور دارند. در این مقاله تعداد روزهای دارای یخبندان در کشور مورد مطالعه قرار گرفت. داده‌های در دسترس تعداد روزهای دارای یخبندان (یعنی تعداد روزهایی که مینیمم دما کمتر از $0°C$ ($32F$, $273/15K$) است) در ۷۷ ایستگاه هواشناسی سینوپتیک می‌باشند که از اول ژانویه تا پایان مارس ۲۰۱۶ جمع‌آوری شده‌اند. این داده‌ها از سایت *http://www.ogimet.com* استخراج گردیده است. در شکل ۱ داده‌های جمع‌آوری شده برروی نقشه ایران مشخص شده است. برای مدل‌بندی داده‌ها از SGLMM's با متغیرهای پنهان چوله نرمال استفاده گردید. تعداد روزهای یخبندان متغیرهای دوجمله‌ای به‌صورت

$$f(y_i|x_i) = \exp\{y_i x_i - u_i \log(1 + \exp\{x_i\})\}, \quad u_i = 180, \quad i = 1, \cdots, 77$$

فرض شدند. یک ساختار همبستگی نمایی برای $\Sigma_{\boldsymbol{\theta}}$ با دو پارامتر $\boldsymbol{\theta} = (\sigma^2, \varphi)^{\top}$ در نظر گرفته شد. شکل ۲ نمودار پراکنش داده‌ها در مقابل طول و عرض جغرافیایی را نشان می‌دهد. به دلیل وجود روند بیشتر و واضح تر در عرض جغرافیایی، مقادیر استاندارد

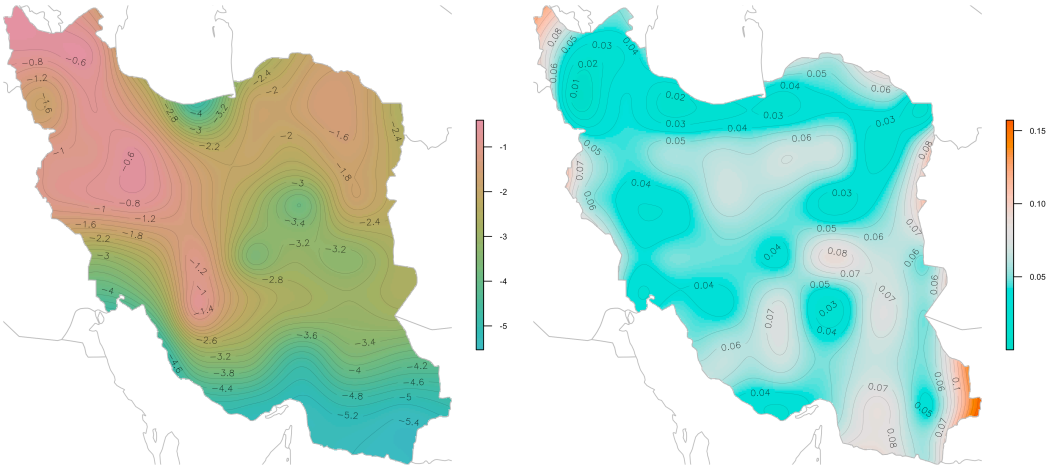


شکل ۱: داده‌های یخبندان ایران در ۷۷ ایستگاه هواشناسی. اندازه دایره با تعداد روزهای دارای یخبندان متناسب است.



شکل ۲: نمودارهای پراکنش داده‌ها در مقابل طول و عرض جغرافیایی

شده‌ی عرض جغرافیایی به عنوان متغیر کمکی وارد مدل گردید. با استفاده از الگوریتم تقریبی معرفی شده برآوردهای ماکسیمم درستنمایی پارامترها به‌صورت $(\hat{\beta}_1, \hat{\sigma}^2, \hat{\varphi}, \hat{\lambda}_0)^\top = (۰/۵۷۳۵, ۱/۷۸۳۲, ۲۹/۹۹۶, ۲/۹۹۸۳)$ محاسبه شد. پس از به دست آوردن پارامترها نقشه پیش‌گویی متغیرهای پنهان فضایی رسم گردید، شکل ۳ـ چپ. این نقشه همبستگی مناطق مختلف را از نظر شدت یخبندان نشان می‌دهد. همان‌طور که ملاحظه می‌گردد شدت یخبندان در نواحی اطراف رشته کوه‌های البرز و زاگرس بیشتر از سایر مناطق است. شکل ۳ـ راست نقشه انحراف‌های استاندارد پیش‌گویی می‌باشد.



شکل ۳: چپ‌ـ نقشه پیش‌گویی، راست‌ـ انحراف استاندارد پیش‌گویی متغیرهای پنهان فضایی

بحث و نتیجه‌گیری

اغلب در مدل‌های آمیخته خطی تعمیم‌یافته فضایی فرض می‌شود که متغیرهای پنهان دارای توزیع نرمال هستند. در این مقاله توزیع چوله نرمال بسته برای مدل‌بندی متغیرهای پنهان استفاده گردید و یک رهیافت تقریبی درستنمایی با استفاده از یک الگورتم جدید برای برآورد پارامترها و سپس یک روش پیش‌گویی تقریبی معرفی گردید. مدل و رهیافت معرفی شده برروی یک مجموعه داده‌ی واقعی پیاده‌سازی گردید و نقشه همبستگی شدت یخبندان سال ۲۰۱۶ در نواحی مختلف ایران به دست آورده شد. در راستای کاهش زمان محاسبات استفاده از درستنمایی مرکب به جای درستنمایی معمولی پیشنهاد می‌گردد.

مراجع

Azzalini, A., (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, **12**(2), 171–178.

Azzalini, A., and Capitanio, A., (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(3), 579–602.

Azzalini, A., and Dalla Valle, A., (1996). The multivariate skew-normal distribution. *Biometrika*, **83**(4), 715.

Baghishani, H., Rue, H. and Mohammadzadeh, M., (2012). On a Hybrid Data Cloning Method and Its Application in Generalized Linear Mixed Models, *Statistics and Computing*, **22**, 597-613.

Baghishani, H. and Rue, 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.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9-25.

Diggle, P., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistic, *Journal of the Royal Statistical Society, Series C. Applied Statistics*, **47**, 299-350.

Dempster, A. P., Laird, N. M. and Rubin, D. B., (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, **39**, 1–38, 1977.

Dominguez-Molina, J.A., Gonzalez-Farias, G. and Gupta, A.K. (2003). The Multivariate Closed Skew Normal Distribution. *Technical Report 03-12, Department of Mathematics and Statistics, Bowling Green State University*.

Eidsvik, J., Martino, S. and Rue, H. (2009). Approximate Bayesian inference in spatial generalized linear mixed models, *Scandinavian Journal of Statistics*, **36**, 1-22.

Evangelou, E., Zhu, Z. and Smith, R. L., (2011). Estimation and prediction for spatial generalized linear mixed models using high order Laplace approximation. *Journal of Statistical Planning and Inference*, **141**(11), 3564-3577.

Gonzalez-Farias, G., Dom´ınguez-Molina, J., Gupta, A., (2004a). Additive properties of skew normal random vectors. *Journal of statistical planning and inference*, **126**(2), 521–534.

Gonzalez-Farias, G., Dom´ınguez-Molina, J., Gupta, A., (2004b). *The closed skew-normal distribution. In: Genton, M. G. (Ed.), Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality.* Chapman and Hall / CRC, Boca Raton, FL, pp. 25–42.

Hosseini, F., Eidsvik, J. and Mohammadzadeh, M. (2011). Approximate Bayesian Inference in Spatial GLMM with Skew Normal Latent Variables, *Computational Statistics and Data Analysis*, **55**, 1791-1806.

Hosseini, F. and Mohammadzadeh, M. (2012). Bayesian Prediction for Spatial GLMM's with Closed Skew Normal Latent Variables, *Australian and New Zealand Journal of Statistics*, **54**, 43-62.

Hosseini, F, (2016). A new algorithm for estimating the parameters of the spatial generalized linear mixed models. *Environmental and Ecological Statistics*, **23**, 205-217.

Karimi, O., Mohammadzadeh, M., (2011). Bayesian spatial prediction for dis- crete closed skew Gaussian random field. *Mathematical Geosciences*, **43**, 565–582.

Karimi, O., Mohammadzadeh, M., (2012). Bayesian spatial regression mod- els with closed skew normal correlated errors and missing observations. *Statistical Papers*, **53**, 205–218.

Karimi, O., Omre, H., Mohammadzadeh, M., (2010). Bayesian closed-skew gaussian inversion of seismic AVO data for elastic material properties. *Geophysics*, **75**, 1-11.

Lange, K., (1995). A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **57**, 425-37.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial statistics. *Biometrika*, **71**, 135–146.

McCulloch, C. E., (1997). Maximum likelihood algorithms for generalized linear mixed models.*Journal of the American Statistical Association,* **92**(437), 162–170.

McCullagh, P. and Nelder, J. A. (1989). *Generalized l Linear Models,* Chapman and Hall, London.

Rencher, A. C. and Schaalje, G. B. (2008). *Linear Models in Statistics*, Wiley.

Rimstad, K., and Omre, H. (2014), Skew-Gaussian random fields, *Spatial Statistics*, **10**, 43-62.

Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of Statistical Planning and Inference*, **137**, 3177-3192.

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**, 319-392.

Torabi, M., (2013). Likelihood inference in generalized linear mixed measurement error models, *Computational Statistics and Data Analysis*, **57**, 549-557.

Torabi, M., (2015). Likelihood inference for spatial generalized linear mixed models. *Communications in Statistics- Simulation and Computation*, **44**, 1692-1701.

Wei, G. C. G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation. *Journal of the American Statistical Association*, **85**, 699 - 704.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models, *Biometrics,* **58**, 129-136.

دومین سمینار آمار فضایی و کاربردهای آن
۳ - ۴ آبان ۱۳۹۶

تحلیل مدل اتورگرسیو فضایی در حضور داده‌های گمشده

سمیرا زحمتکش [*]، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده: گاهی مجموعه داده‌ها، به عنوان تحقق‌های یک میدان تصادفی فضایی، شامل مقادیر گمشده هستند. مقادیر مشاهده شده و گمشده فضایی که در همسایگی قرار دارند می‌توانند حاوی اطلاعات مفیدی برای بهبود تحلیل‌های فضایی باشند. در این صورت می‌توان از مدل‌های اتورگرسیو برای مدل‌بندی داده‌ها استفاده کرد، اما برآورد ماکسیمم درستنمایی پارامترهای مدل متجر به محاسبات زمان‌بر و ماکسیمم‌های موضعی می‌شود. در این مقاله روش جایگزین «ماکسیمم درستنمایی برداری‌شده» معرفی و نحوه تحلیل مدل‌ها در حضور مقادیر گمشده مورد بررسی قرار می‌گیرد.

واژه‌های کلیدی: داده‌های گمشده، داده‌های فضایی، مدل اتورگرسیو، برآورد ماکسیمم درستنمایی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62M20، 62M30، 91B72.

۱ مقدمه

وجود مقادیر گمشده در داده‌های جمع‌آوری شده در زمینه‌های مختلف علوم امری طبیعی است. در داده‌های فضایی، مشاهداتی که در فواصل نزدیکتر به هم قرار دارند می‌توانند دارای ویژگی‌هایی باشند که مستقیماً بر روی مقادیر برآورد شده و سایر استنباط‌ها تاثیر بگذارند. به این ترتیب مقادیر گمشده‌ای که در همسایگی سایر مشاهدات قرار دارند می‌توانند شامل اطلاعات مفیدی باشند که برای بهبود نتایج حاصل از رگرسیون فضایی قابل استفاده‌اند. لیسیج (۲۰۰۴) از سه نوع مدل برای مدل‌بندی داده‌های فضایی و پیشگویی استفاده کرد که عبارتند از مدل اتورگرسیو فضایی[۱]، (SAM) مدل خطای فضایی[۲] (SEM) و مدل داربین فضایی[۳] (SDM) که در این راستا از اطلاعات مقادیر گمشده و مشاهده شده و میانگین شرطی داده‌های گمشده به شرط داده‌های مشاهده شده استفاده می‌کند. برآوردیابی پارامترهای این مدل‌های فضایی عموماً شامل عملیات محاسباتی پیچیده مانند استخراج مقادیر

[*] **ارایه‌دهنده مقاله:** سمیرا زحمتکش، samira.zahmatkesh@modares.ac.ir

[۱] Spatial Autoregressive Model
[۲] Spatial Error Model
[۳] Spatial Durbin Model

ویژه، معکوس کردن و محاسبه دترمینان ماتریس کواریانس است که به میزان قابل توجهی با زیاد شدن مشاهدات افزایش می‌یابد. در روش ماکسیمم درستنمایی نیاز به روش‌های بهینه‌سازی غیرخطی با استفاده از مشتقات و سایر محاسبات تقریبی است که خود یک مساله محسوب می‌شود. متاسفانه چنین مسایلی سبب می‌شوند که کاربر به یک مقدار بهینه فراگیر دست پیدا نکند و حتی از خطاهایی که در حین بهینه‌سازی رخ می‌دهد بی‌خبر باشد. ریپلی (۱۹۸۸) مثالی را مطرح نمود که در آن چندین مقدار بهینه برای یک پارامتر در یک مساله درستنمایی حاصل می‌شد. در واقع یک برآوردگر ایده‌آل باید قابلیت به کار گرفتن داده‌هایی با حجم بالا و استنباط سریع داشته باشد و نباید متکی به الگوریتم‌های بهینه‌سازی غیرخطی باشند که تنها یک مقدار موضعی و نه فراگیر را به عنوان مقدار بهینه در اختیار می‌گذارند. چون در مدل‌های اتورگرسیو فضایی تنها مشاهدات همسایه بر مشاهده‌ای مشخص تاثیر می‌گذارند و بسیاری از مشاهدات بر یکدیگر تاثیری ندارند، در واقع این موضوع باعث ایجاد یک ماتریس وزن تنک خواهد شد که به معنای شیوع عناصر صفر در ماتریس وزن است. پیس و باری (۱۹۹۷) روشی برای محاسبه سریع برآوردها وقتی متغیر وابسته از یک فرایند اتورگرسیو فضایی تبعیت می‌کند ارایه دادند که در آن با استفاده از ویژگی تنکی ماتریس وزن و بازنویسی محاسبات مربوط به روش ماکسیمم درستنمایی می‌توان برآوردها را با هزینه کمتری بدست آورد. در این مقاله با مدل‌های اتورگرسیو فضایی برای مدل‌بندی داده‌های فضایی گمشده سروکار داریم به طوری‌که در محاسبه برآورد پارامترهای مدل و پیشگویی مقادیر گمشده با روش ماکسیمم درستنمایی به نتایج قابل قبولی نخواهیم رسید. لذا برای رفع این مشکل روش ماکسیمم درستنمایی برداری معرفی و ارزیاب و نحوه کاربرد آن نشان داده می‌شود. ابتدا در بخش ۲ مدل‌بندی داده‌های فضایی گمشده به کمک مدل‌های اتورگرسیو مورد بحث قرار می‌گیرد. سپس در بخش ۳ عملیات محاسباتی آن معرفی و روش ماکسیمم درستنمایی برداری شرح داده می‌شود. در بخش ۴ از طریق مثال کاربردی عملکرد روش برآورد پارامترهای مدل اتورگرسیو فضایی مورد ارزیابی قرار می‌گیرد.

۲ مدل‌های اتورگرسیو برای داده‌های گمشده فضایی

در یک میدان تصادفی فضایی، مدل‌های اتورگرسیو فضایی، پیشگویی متغیر پاسخ را بر اساس رگرسیون معمولی $Y = X\beta + \varepsilon$ از طریق میانگین وزنی مقادیر در همسایگی مشاهدات یعنی WY و با در نظر گرفتن ضریب اتورگرسیو فضایی در مدل تصحیح می‌کنند. در اینجا به معرفی مدل اتورگرسیو و ارایه تابع درستنمایی برای آن می‌پردازیم. مدل اتورگرسیو فضایی به صورت

$$Y = \rho W_1 Y + X\beta + U, \quad U = \lambda W_2 U + \varepsilon \tag{۱.۲}$$

تعریف می‌شود، که در آن $Y = (y_1, ..., y_n)^\top$، $\varepsilon \sim N(0, \sigma^2 I)$ بردار متغیر پاسخ، X ماتریس طرح متشکل از k متغیر تبیینی، ρ و λ ضرایب مدل اتورگرسیو، β بردار $k \times 1$ ضرایب رگرسیونی و W_1 و W_2 ماتریس‌های وزن فضایی با بعد $n \times n$ هستند. ماتریس وزن بر اساس مختصات جغرافیایی و همسایگی موقعیت داده‌ها تعیین می‌شود برای دو مشاهده همسایه i و j، w_{ij} مقداری مثبت و برای مشاهدات غیر همسایه صفر است و همچنین $w_{ii} = 0$ یعنی وزن همسایگی هر مشاهده با خودش صفر است. حالت‌های مختلفی از مدل اتورگرسیو فضایی وجود دارد که عبارتند از: مدل اتورگرسیو همزمان، مدل تاخیر فضایی و مدل خطافضایی. با قرار دادن $X = 0$ و $W_2 = 0$ در (۱.۲) مدل اتورگرسیو همزمان به صورت

$$Y = \rho W_1 Y + \varepsilon$$

حاصل می‌شود. در این مدل تغییرات Y بر اساس ترکیب خطی از همسایگی‌هاست. اگر در رابطه (۱.۲)، $W_۲ = ۰$ قرار داده شود مدل تاخیر فضایی به صورت

$$Y = \rho W_۱ Y + X\beta + \varepsilon \qquad (۲.۲)$$

به‌دست می‌آید. چنان‌چه در آن $W_۱ = ۰$ قرار داده شود مدل خطا فضایی به صورت

$$Y = X\beta + U, \quad U = \lambda W_۲ U + \varepsilon$$

حاصل می‌شود که در آن خطاها خودهمبسته فضایی هستند. در این‌جا تابع درستنمایی برای مدل تاخیر فضایی که در (۲.۲) ارایه شده است تعیین و تحلیل‌ها بر اساس آن انجام می‌شود. با فرض آنکه در n موقعیت فضایی تعداد n_{obs} مشاهده و n_{mis} داده گمشده داشته باشیم، مدل (۲.۲) را می‌توان به دو بخش مربوط به داده‌های مشاهده شده و داده‌های گمشده به صورت

$$\begin{pmatrix} Y_{obs} \\ Y_{mis} \end{pmatrix} = \rho \begin{pmatrix} w_{oo} & w_{om} \\ w_{mo} & w_{mm} \end{pmatrix} \begin{pmatrix} Y_{obs} \\ Y_{mis} \end{pmatrix} + \begin{pmatrix} X_{obs} \\ X_{mis} \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_{obs} \\ \varepsilon_{mis} \end{pmatrix}$$

تجزیه کرد، که در آن ماتریس w_{oo} از بعد $n_{obs} \times n_{obs}$، ماتریس w_{om} از بعد $n_{obs} \times n_{mis}$ و به طور مشابه ماتریس‌های w_{mo} و w_{mm} تعریف می‌شوند. می‌توان نشان داد ماتریس کواریانس متغیر پاسخ به صورت

$$\Gamma = \sigma^۲[(I - \rho W)(I - \rho W)']^{-۱}$$

با وارون

$$\Omega = \Gamma^{-۱} = (\frac{۱}{\sigma^۲}) \left[(I - \rho W)(I - \rho W)^\top\right]$$

است که ماتریس دقت نامیده می‌شود. تابع درستنمایی این مدل به شرط داده‌های مشاهده شده Y_{obs} به صورت

$$L(\beta, \sigma^۲, \rho | Y_{obs}) = \int f(Y|\beta, \sigma^۲, \rho) dY_{mis}$$

است، که در آن

$$\begin{aligned} f(Y|\beta, \sigma^۲, \rho) &= (۲\pi\sigma^۲)^{-\frac{n}{۲}} |I - \rho W| \\ &\quad \times \exp\left\{-(\frac{۱}{۲\sigma^۲})((I - \rho W)Y - X\beta)^\top \Omega ((I - \rho W)Y - X\beta\right\}. \end{aligned}$$

در این مدل میانگین شرطی به صورت

$$E(Y_{mis}|Y_{obs}) = \mu_{mis} + \Gamma_{mo}\Omega_{oo}(Y_{obs} - \mu_{obs})$$

است، که در آن

$$\begin{aligned} \mu_{mis} &= B_{mm}X_{mis}\beta + B_{mo}X_{obs}\beta \\ \mu_{obs} &= B_{oo}X_{obs}\beta + B_{om}X_{mis}\beta \end{aligned}$$

$$B^{-1} = \begin{pmatrix} I_{obs} - \rho w_{oo} & -\rho w_{om} \\ -\rho w_{mo} & I_{mis} - \rho w_{mm} \end{pmatrix}$$

B_{oo}، B_{om}، B_{mo} و B_{mm} اجزای ماتریس B هستند که مانند ماتریس وزن W تجزیه شده‌اند. همانطور که ملاحظه می‌شود ماتریس B تنها تابعی از پارامتر ρ است. با جانهی $E(Y_{mis}|Y_{obs})$ به جای Y_{mis} در مجموعه داده‌های نمونه، لگاریتم تابع درستنمایی داده‌های کامل بر حسب پارامترهای مدل ماکسیمم و مدل برآورد می‌شود.

۳ برآورد نظری مدل و پیشگویی مقادیر گمشده

روش ارایه شده در بخش ۲ برای برآورد ماکسیمم درستنمایی مدل به لحاظ تئوری به راحتی قابل بیان است، ولی محاسبات ماتریس‌های Γ_{mo} و Ω_{oo} که هرکدام شاید از بعد بسیار بالایی باشند، پیچیده و دشوار خواهد بود و چون این ماتریس‌ها تابعی از پارامتر ρ هستند در هر بار عملیات ماکسیمم کردن تابع درستنمایی برای برآورد پارامترهای β، σ^2 و ρ باید محاسبه شوند. همچنین در این ماکسیمم کردن تابع درستنمایی لگاریتم دترمینان جمله $|I - \rho W|$ باید محاسبه شود. برای رفع این مشکل از روش ماکسیمم درستنمایی برداری که توسط پیس و باری (۱۹۹۷) و باری و پیس (۱۹۹۹) ارایه شده است، استفاده خواهد شد. لگاریتم تابع درستنمایی برای مدل (۱.۲) به صورت

$$\ell(\beta, \rho, \sigma^2) = c + \ln|I - \rho W| - \frac{n}{2}\ln(\text{SSE}) \qquad (۱.۳)$$

قابل باز نویسی است، که در آن c مقداری ثابت و SSE مجموع توان دوم خطاها و به صورت

$$\text{SSE} = (Y - \rho WY - X\beta_\rho)^\top (Y - \rho WY - X\beta_\rho)$$

است. مقدار بهینه β بر حسب ρ به صورت

$$\beta_\rho = (X^\top X)^{-1} X^\top (I - \rho W)Y = \beta_o - \rho\beta_d$$

قابل محاسبه است، که در آن $\beta_d = (X^\top X)^{-1} X^\top WY$ و $\beta_o = (X^\top X)^{-1} X^\top Y$. به این ترتیب SSE به صورت

$$\begin{aligned} \text{SSE} &= (e_o - \rho e_d)^\top (e_o - \rho e_d) \\ &= {e_o}^\top e_o - 2\rho {e_d}^\top e_o + \rho^2 {e_d}^\top e_d \end{aligned} \qquad (۲.۳)$$

بازنویسی می‌شود که در آن e_o مانده‌های حاصل از رگرسیون کمترین توان‌های دوم معمولی[۴] (OLS)، متغیر پاسخ Y روی X را نشان می‌دهد و e_d مانده‌های حاصل از رگرسیون OLS، عبارت WY روی X است. با جاگذاری SSE از رابطه (۲.۳) در رابطه (۱.۳) و صرف نظر کردن از مقدار ثابت، تابع لگاریتم درستنمایی به صورت

$$\ell(\beta, \rho, \sigma^2) = \ln|I - \rho W| - \frac{n}{2}\ln({e_o}^\top e_o - 2\rho {e_d}^\top e_o + \rho^2 {e_d}^\top e_d) \qquad (۳.۳)$$

[۴] Ordinary Least Square

حاصل می‌شود. اکنون با انتخاب برداری تصادفی به طول m از مقادیر ρ در بازه $[۰, ۱)$ به صورت $\rho_v = (\rho_۱, ..., \rho_m)$، لگاریتم تابع درستنمایی (۳.۳) به صورت

$$\begin{bmatrix} \ln(\beta, \rho_۱, \sigma^۲) \\ \ln(\beta, \rho_۲, \sigma^۲) \\ \vdots \\ \ln(\beta, \rho_m, \sigma^۲) \end{bmatrix} \propto \begin{bmatrix} \ln|I_n - \rho_۱ W| \\ \ln|I_n - \rho_۲ W| \\ \vdots \\ \ln|I_n - \rho_m W| \end{bmatrix} - \frac{n}{۲} \begin{bmatrix} \ln(\Phi(\rho_۱)) \\ \ln(\Phi(\rho_۲)) \\ \vdots \\ \ln(\Phi(\rho_m)) \end{bmatrix} \tag{۴.۳}$$

محاسبه می‌شود، که در آن $\Phi(\rho_i) = e_o{}^\top e_o - ۲\rho_i e_d{}^\top e_o + \rho_i{}^۲ e_d{}^\top e_d$. با معلوم بودن مقادیر اسکالر $e_o{}^\top e_o$، $e_d{}^\top e_o$ و $\rho^۲ e_d{}^\top e_d$ و بردار لگاریتم دترمینان مقادیری که تابعی از ρ_v هستند، محاسبه تابع لگاریتم درستنمایی (۴.۳) بسیار ساده می‌شود. عنصری از ρ_v که بزرگترین مقدار لگاریتم درستنمایی را نتیجه دهد با ρ_{ml} نشان می‌هیم. از دیدگاه محاسباتی برداری کردن مساله از تحمیل هزینه‌های استفاده از بهینه‌گرهای غیرخطی، که معمولاً به همراه تکرار هستند، جلوگیری می‌کند. در محیط‌های محاسباتی چنین عملیات‌های همراه با تکرار، کارایی را به طور چشمگیری کاهش می‌دهند. البته قابل ذکر است که با استفاده از تعداد متناهی مقادیر ρ در انتخاب ρ_{ml}، در مقایسه با روش ماکسیمم‌سازی از طریق بهینه‌گیرهای غیرخطی، به میزان اندکی از دقت برآورد کاسته می‌شود ولی محاسبه تابع لگاریتم درستنمایی بر مشبکه‌ای شامل مقادیر ρ بر میزان استواری برآوردگر می‌افزاید.

وقتی که در داده‌های فضایی گمشدگی وجود داشته باشد جواب مساله ماکسیمم‌سازی با تکیه بر روشی که لیسیج (۲۰۰۴) ارایه کرد، با جانهی $E(Y_{mis}|Y_{obs})$ و ساختن بردار Y تکمیل شده بدست می‌آید، به طوری‌که بردار Y تکمیل شده در عبارت برداری (۴.۳) قابل استفاده است تا اینکه یک ماکسیمم‌سازی تک متغیره نسبت به پارامتر ρ صورت گیرد. با معلوم بودن ρ_{ml} مقادیر جدید $\beta = \beta_o - \rho_{ml}\beta_d$ و $\sigma^۲ = \frac{۱}{n_{obs}}(e_o^{obs} - \rho_{ml}e_d^{obs})^\top(e_o^{obs} - \rho_{ml}e_d^{obs})$ قابل محاسبه‌اند. بنابراین از این مقادیر می‌توان برای شکل دادن مقداری جدید برای $E(Y_{mis}|Y_{obs})$ استفاده کرد تا بردار Y بازسازی‌شده جدیدی نتیجه شود. به این ترتیب تکرار این فرایند برآورد ماکسیمم درستنمایی محاسبه می‌شود که هم به محاسبه کارآمد $E(Y_{mis}|Y_{obs})$ منجر می‌شود و هم به روش ماکسیمم درستنمایی برداری شده متکی است که برای برآورد مدل‌های اتورگرسیو فضایی قابل استفاده است. توجه نمایید که $\sigma^۲$ تنها با استفاده از جمله‌های خطای نمونه مشاهده شده یعنی e_o^{obs} و e_d^{obs} تعیین می‌شود(لیتل و رابین ، ۲۰۰۲) .

۴ مثال کاربردی

در این مقاله از اطلاعات جمع آوری شده مربوط به انتخابات ریاست جمهوری سال ۱۹۸۰ در ۵۰ ایالت آمریکا استفاده شده است. مرکز هر ایالت به عنوان مختصات فضایی آن ایالت درنظر گرفته شده است به این ترتیب یک ماتریس وزن با ابعاد ۵۰×۵۰ خواهیم داشت. چهار متغیری که به عنوان متغیرهای مستقل درنظر گرفته شده‌اند عبارت‌اند از $x_۱$ جمعیت واجد شرایط رای دادن (بالای ۱۹ سال) در هر ایالت، $x_۲$ جمعیت با تحصیلات آکادمیک در هر ایالت، $x_۳$ جمعیت واجد شرایط در هر ایالت که مالکیت واحد مسکونی دارند و $x_۴$ میانگین درآمد سرانه افراد واجد شرایط در هر ایالت به طوری‌که لگاریتم این متغیرها به عنوان متغیرهای کمکی در نظر گرفته شده است. همچنین لگاریتم نسبت کل آراء به جمعیت واجد شرایط در هر ایالت به عنوان متغیر پاسخ در نظر گرفته می‌شود. لازم به ذکر است به دلیل عدم دستیابی به تعداد کل آراء در برخی ایالت‌ها در مقادیر متغیر پاسخ قسمتی از اطلاعات از دست رفته است و به این ترتیب در متغیر پاسخ گمشدگی وجود دارد. به منظور ارزیابی عملکرد روش ارایه شده در این مقاله دو

جدول ۱: برآورد OLS و ماکسیمم درستنمایی برداری برای داده‌های انتخابات

.

مدل				
	رگرسیون خطی		تاخیر فضایی	
پارامتر	برآورد	آماره آزمون t	برآورد	آماره آزمون t
β_0	۱/۰۶۵	۴/۲۴۱	۶/۸۹۹	۳/۱۴۸
β_1	−۰/۶۲۶	−۳/۵۱۵	−۰/۶۲۶	−۴/۵۷۳
β_2	۰/۸۶۲	۸/۹۲۸	۰/۴۵۳	۸/۶۳۰
β_3	۰/۰۵۲	۰/۲۷۸	۰/۱۷۰	۱/۵۶۰
β_4	−۰/۲۸۳	−۲/۱۸۷	−۰/۰۴۸	−۰/۴۷۴
β_5	–	–	−۵/۳۰۸	−۲/۹۹۹
β_6	–	–	۱/۰۹۳	۲/۵۲۶
β_7	–	–	۵/۲۳۶	۳/۸۲۳
β_8	–	–	−۰/۸۵۰	−۰/۶۴۴

مدل یکی به روش OLS و دیگری به روش ماکسیمم درستنمایی برداری به داده‌ها برازش داده شده است. مدل رگرسیون خطی

$$y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + \beta_4 \ln(x_4) + \varepsilon$$

از طریق روش OLS و همچنین مدل

$$\begin{aligned} y \quad = \quad & \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + \beta_4 \ln(x_4) + W \ln(x_1)\beta_5 \\ & + W \ln(x_2)\beta_6 + W \ln(x_3)\beta_7 + W \ln(x_4)\beta_8 + \rho W y + \varepsilon \end{aligned}$$

به روش ماکسیمم درستنمایی بردرای به داده‌ها برازش داده شده است. همان‌طور که مشاهده می‌شود مدل دوم، مدل اول را دربردارد و بر اساس مدل تاخیر فضایی ساخته شده است با این تفاوت که علاوه بر تاخیرهای فضایی متغیر وابسته تاخیر فضایی متغیرهای مستقل نیز به مدل اضافه شده است. pـمقدار برای آزمون موران برابر با $۷/۴۰۹ \exp\{-۷\}$ است که نشان‌دهنده وجود وابستگی فضایی بین داده‌هاست. همچنین پارامتر ρ که شدت وابستگی فضایی را بیان می‌کند در مدل تاخیر فضایی برابر با ۰/۴۸۲۸ برآورد شده است که وجود این وابستگی فضایی را تایید می‌کند. برآورد پارامترهای مدل و مقدار آماره آزمون t برای معنی‌داری ضرایب رگرسیونی دو مدل رگرسیون خطی و مدل تاخیر فضایی در جدول ۱ ارایه شده است. ضریب تعیین تعدیل‌یافته برای مدل تاخیر فضایی که از طریق روش ماکسیمم درستنمایی برداری برازش داده شده است برابر با ۰/۹۳۴۶ است که ۴۷ درصد بیشتر است از مقدار آن برای مدل رگرسیون خطی که از طریق روش OLS برازش داده شده است. مقدار این کمیت برای مدل رگرسیون خطی برابر با ۰/۶۳۵۶ است. مقدار SSE در روش OLS برابر با ۰/۲۵۴۵ و برای روش ماکسیمم درستنمایی برداری برابر با ۰/۰۶۴۷ است. همچنین مقدار RMSE برای روش OLS برابر با ۰/۱۳۹۷ و برای روش ماکسیمم درستنمایی برداری برابر با ۰/۰۴۲۸ است که

مقادیر این کمیت‌ها حاکی از آن است که در مدل تاخیر فضایی که به روش ماکسیمم درستنمایی برداری برازش داده شده است با به کارگیری اطلاعات فضایی موجود در داده‌ها منجر به خطای کمتری شده است. در هر دومدل ضریب متغیر جمعیت واجد شرایط مقداری منفی برآورد شده است که این رخداد به دلیل در نظر گرفتن متغیر پاسخ به صورت تابع لگاریتمی نسبت آراء به جمعیت واجد شرایط امری طبیعی است. بعلاوه در مدل تاخیر فضایی ضریب متغیر جمعیت با تحصیلات آکادمیک بیشتر برآورد شده است همچنین اثر دو متغیر جمعیت دارای مالکیت واحد مسکونی و درآمد سرانه در هر دو مدل معنی‌دار نیست این در حالیست که اثر متغیر وابسته فضایی جمعیت دارای مالکیت واحد مسکونی در مدل تاخیر فضایی معنی‌دار و ضریب آن نسبتاً بزرگ و مثبت برآورد شده است. همچنین متغیر وابسته فضایی جمعیت با تحصیلات آکادمیک در مدل تاخیر فضایی معنی‌دار است و نادیده گرفتن این متغیر در مدل رگرسیون خطی نشان از وجود ضعف‌های احتمالی در این مدل است.

بحث و نتیجه‌گیری

در این مقاله نشان داده شد که چنانچه بین مشاهدات وابستگی فضایی وجود داشته باشد این وابستگی فضایی را می‌توان از طریق مدل‌های اتورگرسیو فضایی مدل‌بندی کرد و همچنین در صورت وجود مقادیر گمشده می‌توان با تجزیه بردار پاسخ، ماتریس طرح و ماتریس وزن و با برآورد مقادیر گمشده از طریق میانگین شرطی داده‌های گمشده به شرط داده‌های مشاهده‌شده و سپس برآورد پارامترهای مدل از طریق روش ماکسیمم درستنمایی برداری به استباط قابل قبولی دست یافت.

مراجع

Barry, R., and Pace, R. K. (1999). A Monte Carlo Estimator of the Log Determinant of Large Sparse Matrices, *Linear Algebra and its Application.* **289**, 41-45.

Lesage, J. P. (2004). Models for Spatially Dependent Missing Data, *Journal of Real Estate Finance an Economics.* **29**, 233-254.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.

Pace, R. K., and Barry R. (1997). Quick Computation of Spatial Autoregressive Estimators, *Geographical Analysis.* **29**, 232-246.

Ripley, B. (1988). *Statistical Inference for Spatial Processes*, Cambridge: Cambridge University Press.

دومین سمینار آمار فضایی و کاربردهای آن
۳ - ۴ آبان ۱۳۹۶

نمایش نیمه طیفی توابع کوواریانس فضایی ـ زمانی

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

[۱] گروه آمار، دانشگاه آزاداسلامی، واحدعلوم وتحقیقات

[۲] گروه آمار، دانشگاه زنجان

چکیده: در این مقاله کلاسی از مدل‌های کوواریانس فضایی ـ زمانی به نام مدل‌های نیمه طیفی معرفی خواهد شد. نمایش نیمه طیفی یک تابع کوواریانس فضایی ـ زمانی شکل خاصی از نمایش طیفی تابع کوواریانس می‌باشد. نشان خواهیم داد که مدل‌های نیمه طیفی خواص مطلوبی برای مدل‌سازی داده‌های فضایی ـ زمانی دارند، همچنین محدودیت‌ها و شرایط لازم برای آنکه مدل‌های ذکرشده دارای خواص مطلوب مورد نظر باشند ذکر خواهد شد.

واژه‌های کلیدی: کوواریانس فضایی ـ زمانی، تابع چگالی طیفی، مدل نیمه طیفی .

کد موضوع‌بندی ریاضی (۲۰۱۰): 62M15، 62M30.

۱ مقدمه

معمولاً در آمار کلاسیک فرض می‌شود که داده‌ها ناهمبسته هستند ولی در عمل موارد زیادی وجود دارند که این ناهمبستگی نقض شده است. به عنوان مثال در داده‌های سری‌های زمانی، مشاهدات دارای نوعی همبستگی از طریق زمان رخداد آن‌ها می‌باشد. همچنین در علوم مربوط به زمین آمار، محیط زیست، هواشناسی،همه‌گیرشناسی و اقیانوس‌شناسی ملاحظه می‌گردد که مشاهدات دارای نوعی همبستگی از نظر موقعیت مشاهدات می‌باشند. به چنین داده‌هایی که همبستگی آنها ناشی از موقعیت جغرافیایی داده است، داده‌های فضایی[۱] گفته می‌شود.

داده‌هایی که هم از نظر موقعیت فضایی و هم از نظر زمانی وابسته باشند، داده‌های فضایی ـ زمانی[۲] نامیده می‌شوند. برای تحلیل سری‌های زمانی و داده‌های فضایی معمولاً دو استراتژی وجود دارد: یکی تحلیل داده‌های سری‌های زمانی و داده‌های فضایی در قلمرو مشاهدات و دیگری تحلیل داده‌های سری‌های زمانی و داده‌های فضایی در قلمرو فرکانس. در دو دهه اخیر بررسی‌های

[*] **ارایه‌دهنده مقاله:** علی شهنواز، shahnavaz_ali2000@gmail.com

[۱] Spatial data

[۲] Spatio-temporal data

زیادی برای تعیین ساختار همبستگی، مدل‌بندی و تحلیل داده‌ها صورت پذیرفته است. تحلیل این گونه دادهها به روش اول مستلزم تعیین ساختار همبستگی به ترتیب زمانی، فضایی و فضایی_زمانی از طریق تابع کوواریانس زمانی، فضایی و فضایی_زمانی 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 فضایی_زمانی می‌باشد. توابع کوواریانس و تابع چگالی طیفی زمانی و فضایی محض بسیار زیادی در متون آماری وجود دارد ولی مطالعات علمی کمتری برای ساخت توابع کوواریانس فضایی_زمانی انجام گرفته است. یکی از ویژگی‌های مهم تابع کوواریانس، ویژگی معین مثبت[۳] باشد و تابعی که چنین شرطی داشته باشد را تابع کوواریانس معتبر می‌نامند. ساخت توابع کوواریانس فضایی_زمانی معتبر تا حدودی امری مشکل است و برای چند سال گذشته یکی از موضوعات علمی آمارشناسان دنیا بوده است. به این منظور برخی فرض‌های ساده کننده نظیر، ایستایی[۴]، همسانگردی[۵]، کاملاً متقارن[۶] و تفکیک‌پذیری [۷] در نظر گرفته شده است.

یکی از روش‌ها برای ساخت توابع کوواریانس فضایی_زمانی معتبر استفاده از نمایش طیفی[۸] تابع کوواریانس است که بر اساس قضیه بوخنر (۱۹۵۵) تعیین می‌شود. این قضیه به خوبی 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 فوریه یک اندازه نامنفی متناهی باشد. از آنجا که توابع کوواریانس فضایی_زمانی هر میدان تصادفی ایستای مرتبه دوم، موجود و معین مثبت است، براساس قضیه بوخنر رده توابع کوواریانس فضایی_زمانی مانا روی $R^d \times R$ معادل رده تبدیل فوریه‌ای اندازه‌های نامنفی و متناهی روی $R^d \times R$ است، لذا با توجه به این قضیه، نمایش طیفی تابع کوواریانس براساس تابع چگالی طیفی عبارت است از:

$$K(s,t) = \int\int g(\lambda,\omega)\exp\{is'\lambda + it\omega\}d\lambda d\omega. \quad (s,t) \in R^d \times R \qquad (۱.۱)$$

بنابراین با اختیار کردن تابع چگالی طیفی $g(\lambda,\omega)$ می‌توان یک تابع کوواریانس فضایی_زمانی معتبر کاملاً طیفی را بنا نهاد. در تحلیل داده‌های فضایی_زمانی علاوه بر تحلیل در قلمرو مشاهدات و قلمرو فرکانس می‌توان قلمرو جدیدی در مابین این دو قلمرو تعریف کرد که به اصطلاح نیمه طیفی[۹] نامیده می‌شود. از مهمترین مزایای این روش این است که به جای انتگرال چندگانه فضایی_زمانی رابطه (۱.۱) از انتگرال صرفاً فضایی استفاده می‌گردد.

کرسی و هانگ (۱۹۹۹) با فرض انتگرال‌پذیری تابع کوواریانس و به‌کارگیری قضیه باکنر، روشی برای ساخت کوواریانس‌های تفکیک‌ناپذیر ایستا ارائه کردند که روش آنان محدود به رده‌ای از توابع می‌شود که انتگرال d_گانه تبدیل فوریه آن‌ها به صورت تحلیلی قابل حل باشد. برای غلبه بر انتگرال صرفاً فضایی (۱.۱)، گنتینگ (۲۰۰۲) با استفاده از توابع کاملاً یکنوا و برنشتاین[۱۰] رده‌ای از مدل‌های کوواریانس فضایی_زمانی تفکیک‌ناپذیر ارائه کرد. علی‌رغم سادگی بسیار زیاد روش گنیتینگ، تابع کوواریانس حاصل یک تابع کوواریانس کاملاً متقارن می‌باشد. علاوه برآن کنت و همکاران (۲۰۱۱) نشان دادند که تابع کوواریانس ارائه شده توسط گنیتینگ دارای مشکل گودی[۱۱] می‌باشد. به معنی که به‌ازای برخی مقادیر تاخیر فضایی، تابع کوواریانس زمانی مربوطه

[۳] Positive definite
[۴] Stationary
[۵] Isotropic
[۶] Completely symmetric
[۷] Separable
[۸] Spectral representation
[۹] Half spectral
[۱۰] Bernestine
[۱۱] Dimple

بصورت یکنوا نزولی نمی‌باشد. هورل و استاین (۲۰۱۷) با استفاده از نمایش شبه طیفی تابع کوواریانس فضایی–زمانی مدل‌های گوسی فضایی–زمانی جدیدی را ارائه کرده‌اند.

۲ مدل‌های نیمه طیفی

نمایش نیمه طیفی تابع کواریانس طیفی $K(S,t)$ را می‌توان با انتخاب ترتیب انتگرال‌گیری نسبت به λ یا ω به‌صورت نیمه طیفی زمانی یا نیمه طیفی فضایی به‌دست آورد. در این مقاله فرم نیمه طیفی در زمان به شکل زیر مورد بحث قرار خواهد گرفت.

$$K(s,t)=\int_R H(s,\omega)\exp\{t\omega\}d\omega, \tag{۱.۲}$$

که در آن

$$H(s,\omega)=f(w)\mathbb{C}(s\delta(\omega))\exp\{i\theta(\omega)\phi^\top s\}, \tag{۲.۲}$$

که در آن f تابع چگالی طیفی صرفاً زمانی، $\mathbb{C}$ تابع همبستگی معتبر در R^d با تابع چگالی طیفی، δ تابعی زوج ومثبت که اثر تعامل فضایی–زمانی، $\theta(.)$ تابع حقیقی و فرد و ϕ بردار یکه که در R^d می‌باشد.

در این مقاله ابتدا حالت تقارن فضایی–زمانی یعنی $\theta(\omega)=۰$ در نظر گرفته می‌شود، یعنی فرم نیمه‌طیفی به شکل $f(\omega)\mathbb{C}(s\delta(\omega))$.

نشان خواهیم داد که بسیاری از خواص مدل‌سازی $K(s,t)$ مستقل از θ و ϕ حفظ می‌شود. خواص اثر تعامل فضایی–زمانی با استفاده از δ تعیین می‌شود، وقتی δ مقدار ثابت باشد توابع کوواریانس کناری فضایی و زمانی بطور مستقل از هم تعیین می‌شوند. در این حالت $K(s,t)$ متناسب با حاصلضرب توابع کوواریانس فضایی و زمانی می باشد یعنی

$$K(s,t)\propto K(s,۰)K(۰,t).$$

این مدل‌ها را مدل‌های تفکیک‌پذیر گویند که به علت غیرطبیعی بودن، مورد انتقاد هستند. استاین (۲۰۱۵) شرط زیر را به عنوان یکی از شرایط مطلوب برای توابع کاملا طیفی ذکر کرده است:

$$\lim_{\|(\lambda,\omega)\|\to\infty}\sup_{\|(u,v)\|<R}\left|\frac{g(\lambda+v,\omega+u)}{g(\lambda,\omega)}-۱\right|=۰. \tag{۳.۲}$$

۱.۲ محدودیت بر روی مدل‌های نیمه طیفی

مدل نیمه طیفی ذکر شده در رابطه (۲.۲) در حالت کلی در شرط (۳.۲) صدق نمی‌کند. فرض کنید h چگالی طیفی تابع کواریانس $\mathbb{C}$ باشد پس فرم کاملاً طیفی K به‌صورت زیر می‌باشد:

$$K(s,t)=\int_R\int_{R^d} f(\omega)\delta(\omega)^{-d}h\left(\frac{\lambda-\theta(\omega)\phi}{\delta(\omega)}\right)\exp\{is'\lambda+it\omega\}d\lambda d\omega. \tag{۴.۲}$$

پس $g(\lambda,\omega)$ یعنی تابع چگالی طیفی کاملاً متقارن را می‌توان به‌صورت زیر نوشت:

$$g(\lambda,\omega)=f(\omega)\delta(\omega)^{-d}h\left(\frac{\lambda}{\delta(\omega)}\right) \tag{۵.۲}$$

در حالت کلی چگالی طیفی k را می‌توان به‌صورت $g(\lambda-\theta(\omega)\phi,\omega)$ نوشت.

قضیه ۱.۲. *فرض کنید $g(\lambda,\omega)$ چگالی طیفی یک تابع کوواریانس فضایی_ زمانی باشد. همچمین فرض کنید $\theta(.)$ یک تابع فرد و ϕ بردار یکه در R^d باشد. فرض کنید $\theta(.)$ به‌صورت موضعی کراندار باشد، در این صورت $g(\lambda,\omega)$ در شرط (۳.۲) صدق می‌کند اگر و تنها اگر $g(\lambda-\theta(\omega)\phi,\omega)$ در شرط (۳.۲) صدق کند.*

در واقع چون استراتژی ساخت مدل‌ها اغلب با مدل های کاملاً متقارن شروع می‌شوند، به کمک این قضیه، استراتژی‌های ساخت مدل را می‌توان به مدل‌های نامتقارن تعمیم داد. حال با ذهنیتی که از قضیه داریم بر روی مدل‌های نیمه طیفی کاملاً متقارن تمرکز خواهیم داشت.

برای اینکه $g(\lambda,\omega)$ در شرط (۳.۲) صدق کند نشان خواهیم داد که طیف h و تابع δ باید به‌صورت قابل تجزیه باشند:

$$h\left(\frac{\lambda}{\delta(\omega)}\right)=\frac{H(\lambda,\omega)\delta(\omega)^d}{f(\omega)}, \tag{۶.۲}$$

که در آن H در شرط (۳.۲) صدق می‌کند. بنابراین فرم ساده‌تر زیر را می‌توان برای g در نظر گرفت:

$g(\lambda,\omega)=p(\omega)q(\lambda,\omega)$ در ادامه شرایطی را که p و q باید داشته باشند تا g در شرط (۳.۲) صدق کند ذکر خواهد شد.

۲.۲ محدودیت

فرض کنید نمایش طیفی تابع کوواریانس k فرمی به‌صورت $g(\lambda,\omega)=p(\omega)q(\lambda,\omega)$ داشته باشد. در این صورت اگر f و g در شرط (۳.۲) صدق کنند آنگاه باید p نسبت به ω ثابت باشد.

پیامد این محدودیت این است برای تعیین مدل‌هایی که در شرط (۳.۲) صدق می‌کنند، تابع δ را نمی‌توان مستقل از f و $\mathbb{C}$ در نظر گرفت.

۳ کلاس جدیدی از مدل‌ها

در این بخش کلاس جدیدی از مدل‌ها معرفی خواهد شد که در آن $\theta(\omega)=۰$ در نظر گرفته می‌شود. لازم به ذکر است تا زمانی که $\theta(\omega)$ به‌صورت موضعی کراندار باشد، صفر یا غیرصفر بودن آن تأثیر در اینکه مدل ساخته شده در شرط (۳.۲) صدق کند ندارد.

فرض کنید $h(\lambda)=\phi(\alpha^۲+\parallel\lambda\parallel^۲)^{-(v+\frac{d}{۲}+\frac{۱}{۲})}$ چگالی طیفی ماترن با پارامتر همواری $v+\frac{۱}{۲}>۰$، $\alpha>۰$ پارامتر دامنه وارون و $\phi>۰$ پارامتر مقیاس باشد. برای یک f داده شده به‌منظور اطمینان از عدم نقض محدودیت ذکر شده $\delta(\omega)$ را به‌صورت زیر در نظر می‌گیریم:

$$\delta(\omega)=f(\omega)^{-\frac{۱}{۲}}\frac{۱}{v+۱/۲}, \tag{۱.۳}$$

$$g(\lambda,\omega)=\phi\left(\alpha^۲f(\omega)^{-\frac{۱}{v+۱/۲}}+\parallel\lambda\parallel^۲\right)^{-(v+d/۲+۱/۲)} \tag{۲.۳}$$

اگر f که خود یک تابع چگالی طیفی است به‌صورت $f\sim c_f\omega^{-k}$ در نظر گرفته شود، برای بعضی از مقادیر $k>۰$ و c_f که یک مقدار ثابت مثبت وابسته به f می‌باشد، آنگاه $g(\lambda,\omega)$ تعریف شده به‌صورت فوق، کلاسی از مدل‌ها را توصیف می‌کند که در شرط (۳.۲) صدق می‌کنند.

فرم نیمه طیفی رابطه (۲.۳) با جایگذاری تابع کواریانس ماترن رابطه (۲.۲) به‌صورت زیر حاصل می‌شود:

$$f(\omega)\mathbb{C}(s\delta(\omega)) = f(\omega)\frac{\phi\pi^{d/2}}{2^{v-1/2}\Gamma(v+\frac{(d+1)}{2})\alpha^{2v+1}} \times \left(\alpha \parallel s \parallel f(\omega)^{-\frac{1}{2}\frac{1}{v+1/2}}\right)^{v+\frac{1}{2}} K_{v+\frac{1}{2}}\left(\alpha \parallel s \parallel f(\omega)^{-\frac{1}{2}\frac{1}{v+1/2}}\right) \quad (3.3)$$

که در آن $K_{v+\frac{1}{2}}$ تابع بسل مرتبه دوم می‌باشد. اگر در نظر بگیریم

$$\mathcal{M}_{v+\frac{1}{2}}(s) = \parallel s \parallel^{v+\frac{1}{2}} K_{v+\frac{1}{2}}(\parallel s \parallel),$$

در این صورت رابطه (۲.۳) را می‌توان به‌صورت زیر نوشت:

$$f(\omega)\mathbb{C}(s\delta(\omega)) = \phi f(\omega)\mathcal{M}_{v+\frac{1}{2}}\left(\alpha \parallel s \parallel f(\omega)^{-\frac{1}{2}\frac{1}{v+1/2}}\right). \quad (4.3)$$

مثال ۱.۳. *فرض کنید $f(\omega) = (\beta^2+\omega^2)^{-(k+\frac{1}{2})}$ چگالی طیفی ماترن یک بعدی باشد که برای سادگی کار پارامتر مقیاس برابر یک فرض شده است. با جایگذاری این f در (۲.۳) و (۴.۳) خواهیم داشت:*

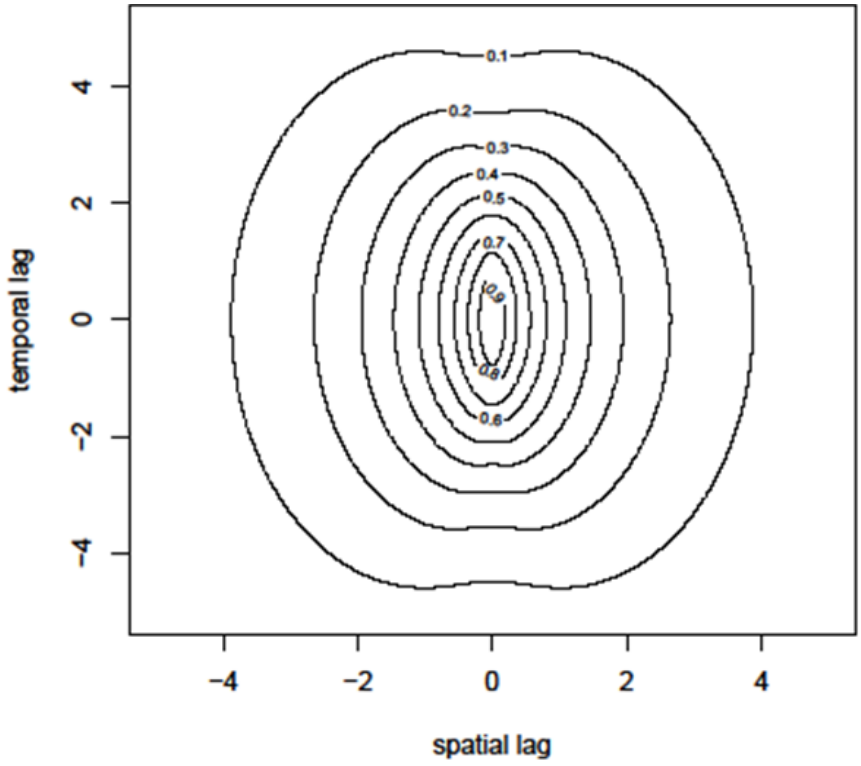
$$g(\lambda,\omega) = \phi\left(\alpha^2(\beta^2+\omega^2)^{\frac{k+\frac{1}{2}}{v+1/2}} + \parallel \lambda \parallel^2\right)^{(v+\frac{d}{2}+\frac{1}{2})},$$

$$f(\omega)\mathbb{C}(s\delta(\omega)) = \phi(\beta^2+\omega^2)^{k+1/2}\mathcal{M}_{\frac{1}{v+1/2}}\left(\alpha \parallel s \parallel (\beta^2+\omega^2)^{\frac{1}{2}+\frac{k+1/2}{v+1/2}}\right). \quad (5.3)$$

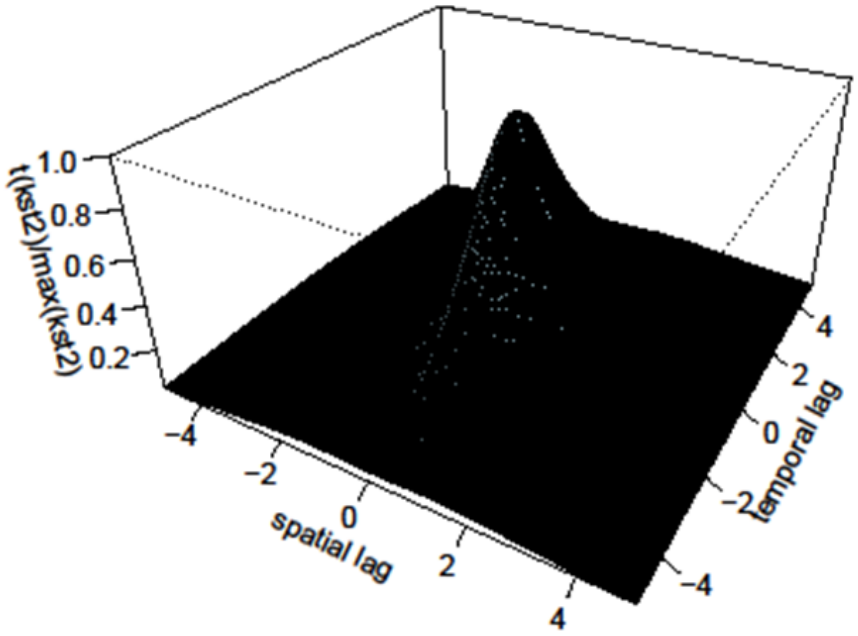
در شکل ۱ نمودار تراز تابع کواریانس مثال فوق به‌ازای $v = \frac{1}{8}$، $\alpha = 0/5$، $k = 2$ و $\beta = 1$ رسم شده است. در شکل۲ نمودار تابع کوواریانس فضایی_ زمانی رسم شده است.

بحث و نتیجه‌گیری

مدل های نیمه طیفی فضایی_زمانی معرفی شده با استفاده از فرم $f(w)\mathbb{C}(s\delta(\omega))\exp\{i\theta(\omega)\phi^\top s\}$ تفکیک‌ناپذیر هستند و در شرط (۳.۲) و محدودیت (۱.۱) صدق می‌کنند. مدل ذکر شده کاملاً متقارن می باشد با این حال قضیه ۱.۲ نشان داد، تا زمانی که $\theta(\omega)$ به‌صورت موضعی کراندار باشد، تعمیم به حالت نامتقارن کار ساده‌ای می‌باشد. هورل و استاین (۲۰۱۷) با برازش این مدل به داده های باد ایرلند نشان دادند که این مدل ها از انعطاف بیشتری برخوردار هستند و نتایج معقول‌تری نسبت به مدل‌های تفکیک‌پذیر و تفکیک‌ناپذیر معرفی شده توسط گنتینگ (۲۰۰۲) دارند.



شکل ۱: نمودار تراز تابع کوواریانس مثال(۱.۳)



شکل ۲: نمودار تابع کوواریانس مثال(۱.۳)

مراجع

محمدزاده م، (۱۳۹۴). *آمار فضایی و کاربردهای آن*، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،

Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability*, University of California Press, Berkeley and Los Angeles.

Cressie, N. and Huang, H. C. (1999), Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions, *Journal of the American Statistical Association*, **94**, 1330-1340.

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

Horrell, M. T., and Stein, M. L. (2017), Half-spectral Space-Time Covariance Models, *Journal of Spatial Statistics*, V19.

Kent, J., Mohammadzadeh, M. and Mosammam, A. (2011), The dimple in gneiting's spatialtemporal covariance model, *Biometrika* **98**, 489–494.

Stein, M. (2015), When does the screening effect not hold?, *Spatial Statistics* **11**, 65–80.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶

تحلیل بیزی تقریبی الگوی نقطه‌ای فضایی زمین‌لرزه‌های شمال غرب ایران با فرآیندهای کاکس لگ‌گاوسی

فاطمه شیاسی*، حسین باغیشنی، نگار اقبال

گروه آمار، دانشگاه صنعتی شاهرود

چکیده: در این مقاله، مدل‌بندی بیزی داده‌های الگوی نقطه‌ای فضایی زمین‌لرزه‌های ناحیه شمال غرب ایران مرتبط با گسل شمال تبریز و برآورد تابع شدت الگوی مورد نظر، مورد توجه است. ابزار معمول برای تحلیل بیزی الگوهای نقطه‌ای فضایی، به دلیل بسته نبودن شکل توزیع پسین متناظر، الگوریتم‌های MCMC هستند. از آن‌جا که مجموعه داده مورد نظر حجیم است، استفاده از الگوریتم‌های مذکور کند و زمان‌بر خواهد بود و تشخیص همگرایی و آمیختگی خوب زنجیر نیز چالش‌برانگیز است. با توجه به کاربردهای وسیع و انعطاف مناسب فرآیندهای کاکس لگ‌گاوسی و امکان برازش آن توسط روش تقریبی بیزی INLA، از این فرآیند برای تحلیل داده‌ها استفاده می‌کنیم. روش INLA در مقایسه با الگوریتم‌های MCMC بسیار سریع بوده و از مشکلات همگرایی و آمیختگی برحذر است.

واژه‌های کلیدی: داده‌های زمین‌لرزه، فرآیند کاکس لگ‌گاوسی، تقریب لاپلاس آشیانه‌ای، فرآیندهای نقطه‌ای، تابع شدت.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30.

۱ مقدمه

داده‌های الگوی نقطه‌ای فضایی، کاربردهای متنوعی در دنیای واقعی دارند. به‌عنوان نمونه، می‌توان شناسایی نقاط زلزله‌خیز و بررسی گونه‌های مختلف گیاهان را نام برد. در تحلیل این نوع از داده‌ها ویژگی اصلی مورد علاقه، تابع شدت است. برای مدل‌بندی این تابع، رهیافت‌های متفاوتی مانند فرآیندهای نقطه‌ای دوجمله‌ای، پواسن و کاکس توسط آماردان‌های مختلف معرفی شده‌اند. برای اولین بار کاکس (۱۹۵۵) فرآیند پواسونی را مطرح کرد که در آن اندازه شدت یک اندازه تصادفی بود. فرآیند پیشنهادی وی به فرآیند کاکس معروف شد. این فرآیند زیررده‌ای از رده انعطاف‌پذیر مدل‌های فرآیند نقطه‌ای فضایی است (، مولر و واگه پترسن ۲۰۰۷). بعد از معرفی فرآیندهای کاکس، راتبن و کرسی (۱۹۹۴) تعریفی از فرآیندهای لگ‌گاوسی ارائه دادند، اما توجه خود ر تنها به مواردی

*ارایه‌دهنده مقاله: فاطمه شیاسی، shiasi.fateme92@gmail.com

معطوف کردند که تابع شدت در آن‌ها ثابت است. مولر و همکاران (۱۹۹۸) فرآیند کاکسی را معرفی کردند که در آن لگاریتم میدان هادی، یک میدان تصادفی گاوسی است و استفاده از آن منجر به تعریف دقیق‌تری از فرآیند کاکس لگ‌گاوسی[۱] (LGC) شد. با توجه به ماهیت داده‌های الگوی نقطه‌ای فضایی و انعطاف‌پذیری خوب فرآیندهای LGC، استفاده از این فرآیندها طرفداران زیادی پیدا کرده است.

فرآیندهای کاکس لگ‌گاوسی، به دلیل سادگی تعبیه کردن متغیرهای کمکی درون خود، بسیار منعطف عمل می‌کنند. استنباط کلاسیک در این فرآیندها نیازمند محاسبات پیچیده برای به‌دست آوردن تقریبی از تابع درست‌نمایی است. به همین دلیل، رهیافت معمول محققان برای برازش تابع شدت با فرآیندهای LGC، بیزی است. با توجه به صریح نبودن شکل توزیع پسین در این فرآیندها ابزار متداول برای برازش مدل، الگوریتم‌های مونت کارلوی زنجیر مارکوفی[۲] (MCMC) است. اجرای این الگوریتم‌ها نیاز به مهارت‌های برنامه‌نویسی دارد و از لحاظ همگرایی و زمان محاسبات، آلوده به مشکلات جدی است (مولر و واگه پترسن ۲۰۰۷).

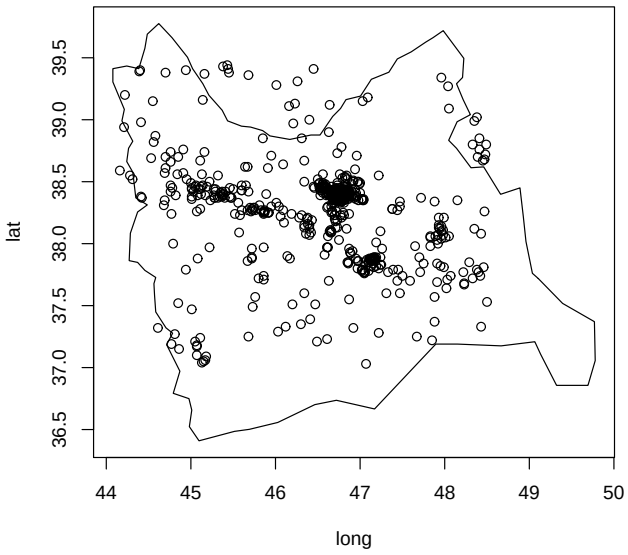
با توجه به این‌که فرآیندهای LGC زیررده‌ای از مدل‌های گاوسی پنهان[۳] (LGMs) محسوب می‌شوند، می‌توان از یک روش بیزی تقریبی با نام تقریب لاپلاس آشیانه‌ای جمع‌بسته[۴] (INLA) برای استنباط در آن‌ها استفاده کرد. این روش توسط رو و همکاران (۲۰۰۹) معرفی شده است و می‌تواند در استنباط مدل‌های گاوسی پنهان، رهیافت جانشین الگوریتم‌های سنگین MCMC باشد. روش INLA در مقایسه با روش‌های نمونه‌گیری MCMC بسیار سریع است و نتایج آن از نظر دقت برابری می‌کند. اسکلر و هلد (۲۰۱۱) و قیومی و همکاران (۱۳۹۱) در مطالعات موردی خود به‌طور جداگانه نتایج حاصل از دو روش INLA و الگوریتم‌های MCMC را مورد ارزیابی قرار داده‌اند.

هدف از این مقاله، مطالعه الگوی حاکم بر نقاط زلزله‌خیز شمال غرب ایران مرتبط با گسل تبریز است. با توجه به مقدمه ذکرشده، از فرآیندهای LGC برای مدل‌بندی داده‌های مربوط به زمین‌لرزه‌های ثبت‌شده در این ناحیه در سال ۲۰۱۴ میلادی استفاده کردیم. برای مدل‌بندی تابع شدت الگوی نقطه‌ای زمین‌لرزه‌های ناحیه مورد نظر، با استفاده از فرآیندهای LGC، نیازمند یافتن چگالی‌های پسین کناری با استفاده از روش INLA هستیم. برای این منظور، ابتدا در بخش ۲ داده‌ها را تشریح می‌کنیم. سپس در بخش ۳ فرآیندهای لگ‌گاوسی را معرفی می‌کنیم. در بخش ۴ مقدمه‌ای بر دیدگاه استنباطی و روش INLA را ارایه می‌کنیم. در بخش ۵ نیز نتایج تحلیل داده‌ها را بیان می‌کنیم. در پایان به بحث و نتیجه‌گیری می‌پردازیم.

۲ معرفی داده‌ها

زمین‌لرزه یا زلزله، لرزش و جنبش زمین است که به علت آزاد شدن انرژی ناشی از گسیختگی سریع در گسل‌های پوسته زمین در مدتی کوتاه روی می‌دهد. داده‌های مورد نظر این مقاله، داده‌های لرزه‌ای ثبت‌شده در ناحیه شمال غرب ایران مرتبط با گسل شمال تبریز و گسلش‌های آن است. این داده‌ها از سایت پژوهشکده بین‌المللی زلزله برای دوره زمانی یک ساله در سال ۲۰۱۴ که توسط دستگاه لرزه‌نگار ثبت شده‌اند، دریافت شده است. داده‌های لرزه‌ای و نقاط مرزی ناحیه در یک چهارگوش با رئوس (طول جغرافیایی۴۸/۷۶، عرض جغرافیایی۳۹/۲۲)، (طول جغرافیایی۴۴/۴۷، عرض جغرافیایی۳۹/۲۲)، (طول جغرافیایی ۴۸/۷۶، عرض جغرافیایی۳۶/۲۵) و (طول جغرافیایی ۴۸/۷۶، عرض جغرافیایی۳۹/۲۲) قرار دارند (شکل ۱).

[۱] Log Gaussian Cox
[۲] Markov Chain Monte Carlo
[۳] Latent Gaussian Models
[۴] Integrated Nested Laplace Approximation



شکل ۱: نمایش ۵۸۳ موقعیت لرزه‌ای شمال غرب ایران طی سال ۲۰۱۴

مجموعه داده لرزه‌ای شامل موقعیت‌های جغرافیایی نقاط زلزله با طول و عرض مشخص، تاریخ و زمان دقیق وقوع لرزه‌ها و اندازه زمین‌لرزه (اندازه‌گیری‌شده به واحد ریشتر) است. موقعیت مورد مطالعه یکی از مناطق زلزله‌خیز ایران است و بر اساس نقشه تهیه‌شده توسط زارع و همکاران (۲۰۱۶) (شکل ۲) جزو مناطق پرخطر لرزه‌ای دسته‌بندی شده است.

۳ فرآیندهای کاکس لگ‌گاوسی

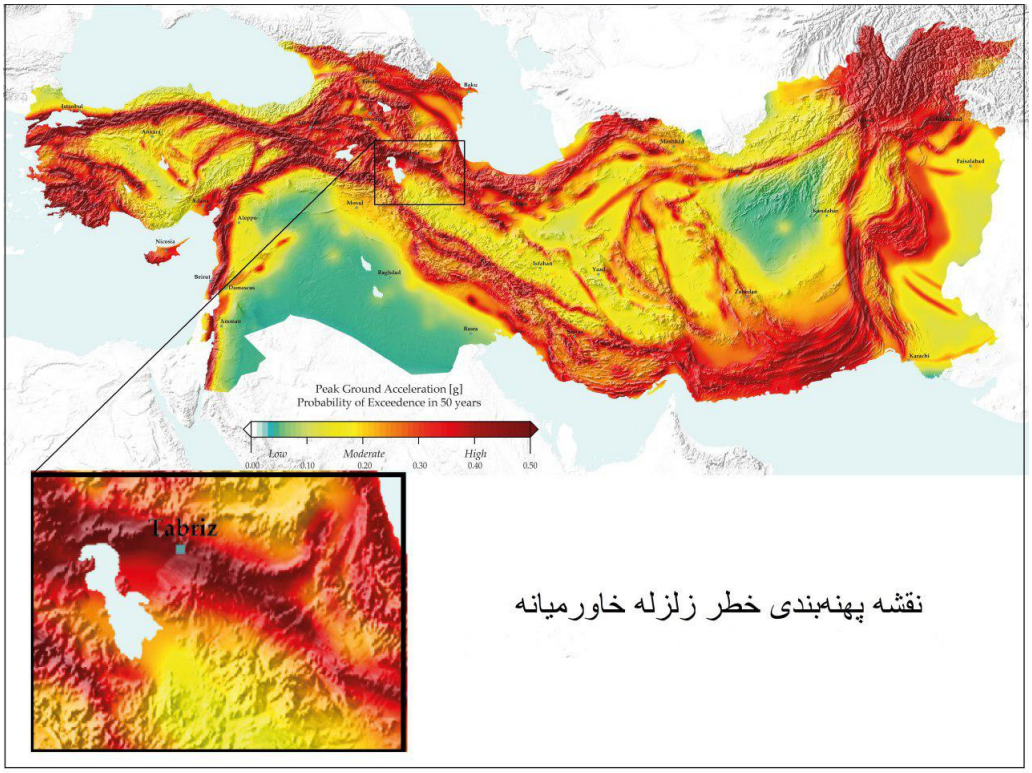
فرآیند LGC، فرآیند کاکسی است که لگاریتم تابع شدت آن یک فرآیند گاوسی است. مولر و همکاران (۱۹۹۸) فرآیند کاکسی را معرفی کردند که در آن لگاریتم میدان هادی یک میدان گاوسی است. چون یک میدان گاوسی مقادیر منفی را اختیار می‌کند، نمی‌توان به‌طور مستقیم از آن به‌عنوان میدان هادی یک فرآیند کاکس استفاده کرد. برای رفع این مشکل، کافی است از تبدیل نمایی به روی میدان گاوسی استفاده کرد تا یک میدان تصادفی نامنفی به دست آید.

به‌طور دقیق‌تر، فرآیند کاکس X با میدان هادی $\Lambda(s) = exp(\zeta(s))$ را یک فرآیند LGC بر S گویند، هرگاه

$$\zeta = \{\zeta(s); s \in S \subset R^d\}$$

یک میدان تصادفی گاوسی با تابع میانگین $m(s)$ و تابع کواریانس $c(s, v)$ باشد. این فرآیند به‌طور قریب به یقین، انتگرال‌پذیر موضعی است. شایان ذکر است که مانایی و همسانگردی[۵] یک فرآیند LGC از مانایی و همسانگردی میدان گاوسی متناظر خود نتیجه می‌شود.

[۵] Isotropy



شکل ۲: نقشه پهنه‌بندی خطر زلزله در ناحیه خاورمیانه (گزارش‌شده توسط زارع و همکاران، ۲۰۱۶)

تابع احتمال پوچی[۶] در یک فرآیند LGC به صورت

$$\nu(B) = E\Big[\exp\Big\{-\int_B \exp(\zeta(s))ds\Big\}\Big] \quad B \subseteq S$$

تعریف می‌شود، که در آن متغیر تصادفی $\zeta(s)$ برای هر $s \in S$ دارای توزیع نرمال است؛ اما توزیع انتگرال تصادفی

$$\int_B \exp\{\zeta(s)\}ds$$

شکل بسته‌ای ندارد و نمی‌توان آن را تعیین کرد. برای هر $s, v \in S$، تابع همبستگی زوجی[۷] و تابع شدت[۸] به ترتیب به صورت زیر تعریف می‌شوند:

$$\begin{aligned} g(s,v) &= \exp\{c(s,v)\} \\ \rho(s) &= \exp\{m(s) + \frac{c(s,v)}{۲}\} \end{aligned}$$

با توجه به دو تابع همبستگی زوجی و شدت، می‌توان نتیجه گرفت که توزیع یک فرآیند LGC توسط این دو تابع تعیین می‌شود. معمولا برای تعیین تابع همبستگی زوجی $g(\cdot,\cdot)$، یا معادل آن تابع کواریانس $c(\cdot,\cdot)$، و تابع شدت $\rho(\cdot)$ از مدل‌های پارامتری استفاده می‌شود. بنابراین برای تعیین یک فرآیند کاکس لگ‌گاوسی کافی است بردار پارامترهای تابع همبستگی زوجی، تابع شدت و همچنین میدان تصادفی ζ برآورد شوند.

[۶] Void probability function
[۷] Pair correlation function
[۸] Intensity function

۴ استنباط بیزی تقریبی با روش INLA

در رویکرد بیزی فرض می‌شود پارامترهای تابع همبستگی زوجی و تابع شدت تحقق‌هایی از توزیع‌های احتمالی با نام توزیع‌های پیشین هستند. با این نگاه می‌توان میدان تصادفی ζ را نیز به‌عنوان یک پارامتر با توزیع پیشین گاوسی در نظر گرفت. در این رویکرد تمام استنباط‌ها شامل برآورد پارامترها و پیش‌گویی میدان تصادفی ζ مبتنی بر توزیع پسین مدل ساخته می‌شوند. بنابراین با توجه به این‌که میدان تصادفی ζ یک میدان پنهان به حساب می‌آید و پیش‌بینی آن بر اساس مشاهدات X با کمیت

$$E\Big(\zeta(s)\Big|X\Big) \quad s \subseteq S$$

انجام می‌شود، استفاده از رویکرد بیزی برای استنباط آماری فرآیندهای LGC مناسب‌تر است.

فرآیند کاکس لگ‌گاوسی عضوی از رده مدل‌های گاوسی پنهان است. در این رده از مدل‌ها فرض می‌شود متغیرهای پاسخ y_i، برای $i = ۱, \ldots, n$، به شرط میدان پنهان ζ با ابرپارامترهای $\boldsymbol{\theta} = (\theta_۱, \ldots, \theta_J)'$، مستقل شرطی هستند. رو و همکاران (۲۰۰۹) نشان دادند اگر ζ یک ماتریس دقت تنک[۹] داشته باشد و تعداد ابرپارامترهای آن کوچک باشد (کمتر از ۷)، آنگاه استنباط مبتنی بر روش INLA می‌تواند کارا باشد.

هدف اصلی روش INLA به‌دست آوردن تقریبی از توزیع‌های پسین کناری میدان پنهان، $\pi(\zeta_i|y)$، و ابرپارامترها، $\pi(\theta_i|y)$، به صورت زیر است:

$$\begin{aligned} \pi(\zeta_i|\boldsymbol{y}) &= \int \pi(\zeta_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \quad i = ۱, \ldots, n_d \qquad (۱.۴) \\ \pi(\theta_i|\boldsymbol{y}) &= \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j} \quad j = ۱, \cdots, J (۲.۴) \end{aligned}$$

که در آن $\boldsymbol{\theta}_{-j}$ بردار حاصل از حذف درایه jام $\boldsymbol{\theta}$ و n_d بعد میدان ζ است. فرمول آشیانه‌ای استفاده‌شده برای محاسبه $\pi(\zeta_i|\boldsymbol{y})$، به‌وسیله تقریب $\pi(\zeta_i|\boldsymbol{\theta}, \boldsymbol{y})$ و $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ انجام می‌پذیرد و سپس با استفاده از تقریب‌های عددی و یک‌پارچه کردن $\boldsymbol{\theta}$ محاسبه می‌شود. به‌طور کلی روش INLA با استفاده از تبدیل‌های لاپلاس و انتگرال‌گیری عددی محاسباتی سریع، تقریبی دقیق را جایگزین الگوریتم‌های MCMC می‌کند. توزیع پسین کناری در (۱.۴) و (۲.۴) با استفاده از تقریب لاپلاس به صورت

$$\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\zeta, \boldsymbol{\theta}, \boldsymbol{y})}{\widetilde{\pi}_G(\zeta|\boldsymbol{\theta}, \boldsymbol{y})}\Bigg|_{\zeta=\zeta^*(\boldsymbol{\theta})}$$

محاسبه می‌شود، که در آن $\widetilde{\pi}_G(\zeta|\boldsymbol{\theta}, \boldsymbol{y})$ تقریب گاوسی $\pi(\zeta_i|\zeta, \boldsymbol{\theta}, \boldsymbol{y})$ و $\zeta^*(\boldsymbol{\theta})$ مد توزیع $\pi(\zeta_i|\zeta, \boldsymbol{\theta}, \boldsymbol{y})$ است. برای جزییات بیشتر در مورد این روش تقریبی به رو و همکاران (۲۰۰۹) مراجعه کنید.

۵ مدل‌بندی الگوی نقاط زلزله‌خیز ناحیه شمال غرب ایران

همان‌طور که بیان شد فرآیند LGC عضویی از رده مدل‌های گاوسی پنهان است و مدل‌های گاوسی پنهان نیز زیررده‌ای بزرگ و انعطاف‌پذیر از رده مدل‌های رگرسیون جمعی ساختاری[۱۰] (STAR) هستند و بنابراین روش INLA برای برازش آن‌ها می‌تواند به‌کار گرفته شود و کارا هم باشد.

[۹] Sparse precision matrix

[۱۰] Structured additive regression model

فرض می‌کنیم X الگوی نقطه‌ای مورد نظر باشد. نقاط زلزله در پنجره مشاهدات S، مشخص‌شده در شکل‌های ۱ و ۲، پراکنده شده‌اند. برای برازش مدل فرآیند LGC، پنجره مشاهدات را با گسسته‌سازی به یک مشبکه با $N = n_{row} \times n_{col}$ سلول $\{s_{ij}\}$ تقسیم می‌کنیم، به‌طوری که مساحت هر سلول برابر $|s_{ij}|$، برای $i = ۱, \ldots, n_{row}$ و $j = ۱, \ldots, n_{col}$، می‌باشد. بنابراین نقاط درون الگو را می‌توان به شکل $\{\xi_{ijk_{ij}}\}$ نمایش داد، به‌طوری که $k_{ij} = ۱, \ldots, y_{ij}$ و y_{ij} تعداد نقاط مشاهده‌شده در سلول s_{ij} است. تعداد نقاط مشاهده‌شده در سلول s_{ij} به شرط مفروض بودن تابع شدت $\Lambda(s_{ij}) = \exp\{\zeta(s_{ij})\} = \exp\{\zeta_{ij}\}$ از توزیع پواسون پیروی می‌کند؛ به عبارت دیگر

$$y_{ij}|\zeta_{ij} \sim Po(|s_{ij}| \exp\{(\zeta_{ij})\}).$$

می‌توان ζ_{ij} را به صورت زیر مدل‌بندی کرد:

$$\zeta_{ij} = \beta_{۰} + f(z_c(s_{ij})) + f_{spat}(s_{ij}) + u_{ij} \qquad (۱.۵)$$

که در آن $\beta_{۰}$ پارامتر عرض از مبدا است، $z_c(s_{ij})$ یک متغیر تبیینی است که برای داده‌های ما اندازه زمین‌لرزه (به واحد ریشتر) است و چون ممکن است رابطه بین متغیر تبیینی و لگاریتم تابع شدت خطی نباشد، ارتباط دقیق آن‌ها را با تابع هموار $f(z_c(s_{ij}))$ نشان می‌دهیم. اثر فضایی تابع شدت با $f_{spat}(s_{ij})$ وارد مدل می‌شود و u_{ij} نیز جمله خطای ناهمبسته مدل است که فرض می‌کنیم دارای توزیع گاوسی با میانگین صفر است. بنابراین داده‌های زلزله را با یک الگوی فرآیند نقطه‌ای نشاندار[۱۱] (استفاده از متغیر کمکی در نقاط زلزله) مدل‌بندی و تحلیل می‌کنیم.

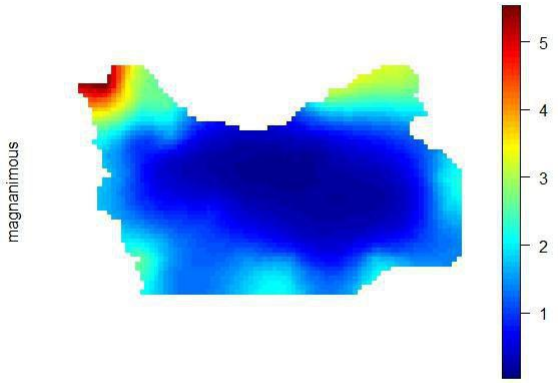
برای داده‌های زلزله شمال غرب ایران، مدل‌بندی توابع $f(z_c(s_{ij}))$ و $f_{spat}(s_{ij})$ توسط رهیافت معادلات دیفرانسیل جزیی تصادفی[۱۲] که توسط لیندگرین و همکاران (۲۰۱۵) معرفی شد، انجام شده است. البته روش‌های جانشین دیگری نیز مانند استفاده از فرآیندهای قدم زدن تصادفی[۱۳] مرتبه اول و دوم (ایلیان و همکاران، ۲۰۱۲) برای مدل‌بندی این توابع هموار پیشنهاد شده‌اند.

شکل ۴ نتیجه برآورد لگاریتم تابع شدت مدل فرآیند LGC را برای ناحیه مورد مطالعه نشان می‌دهد. شکل ۳ نیز برآورد اثر هموار فضایی متغیر نشاندار الگوی نقطه‌ای را نمایش می‌دهد. شکل ۳ اثر متقابل اندازه زمین‌لرزه در نقاط ثبت‌شده را در مقایسه با فراوانی رخداد زلزله به‌خوبی نمایان می‌کند. از روی این دو شکل، به روشنی می‌توان ملاحظه کرد که بیشترین تعداد زلزله در میانه ناحیه مورد مطالعه و بیشتر متمایل به شرق رخ داده‌اند. از طرف دیگر اندازه زمین‌لرزه در این بخش ناحیه چندان نگران‌کننده نبوده است؛ در واقع بیشتر از ۲ ریشتر نیست. بنابراین به نظر می‌رسد با این‌که این ناحیه نزدیک به گسل تبریز زلزله‌خیز است، اما اندازه زلزله‌ها نگران‌کننده نیست. شاید افزودن سایر اطلاعات کمکی به‌عنوان متغیرهای کمکی مانند عمق زلزله و اطلاعات سایر سال‌ها، به برازش بهتر الگو و تفسیر دقیق‌تر کمک کند.

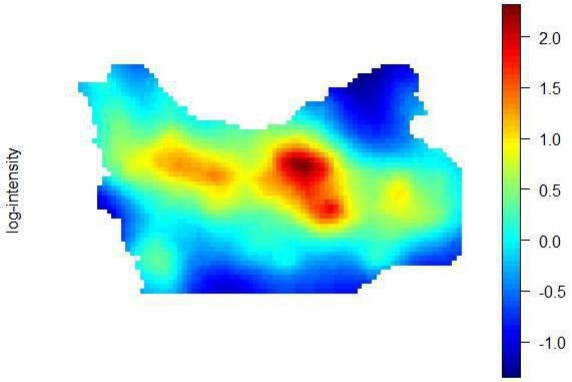
بحث و نتیجه‌گیری

در این مقاله، فرآیندهای LGC به‌عنوان فرآیندهای منعطف برای مدل‌بندی الگوهای نقطه‌ای فضایی معرفی شدند. از تحلیل بیزی این فرآیندها استقبال بیشتری شده است. اما استنباط بیزی معمول در این فرآیندها به کمک الگوریتم‌های MCMC انجام می‌شود.

[۱۱] Marked spatial point process
[۱۲] Stochastic partial differential equations
[۱۳] Random walk processes



شکل ۳: نقشه پهنه‌بندی اثر متغیر تبیینی شدت زمین‌لرزه در ناحیه شمال غرب ایران



شکل ۴: نقشه پهنه‌بندی لگاریتم تابع شدت در منطقه زلزله‌خیز شمال غرب ایران

روش تقریبی INLA جانشین جدید کارایی است که در کنار سرعت بالا، از مشکلات همیشگی همراه الگوریتم‌های MCMC، شامل همگرایی کند و آمیختگی ضعیف، دور است. با استفاده از این روش تقریبی، مدل‌بندی نقاط زلزله‌خیز ناحیه شمال غرب ایران با یک فرآیند LGC نشاندار را اجرا و نتایج رو تفسیر کردیم.

تحلیل بیزی فضایی_زمانی این داده‌ها با استفاده از فرآیندهای LGC به کمک روش تقریبی INLA می‌تواند به‌عنوان گام بعدی مطرح باشد.

مراجع

قیومی ز.، ک.، محمدزاده، م. (۱۳۹۱)، تحلیل مدل‌های گاوسی پنهان فضایی با تقریب لاپلاس آشیانی ترکیبی، مجموعه مقالات دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۹۳-۱۰۶.

Cox, D. R. (1955). Some Statistical Methods Connected with Series of Events, *Royal Statistical Society*, **17**(2), 129–164.

Illian, J. B., Sorbye, H. and Rue, H. (2012). A Toolbox for Fitting complex Spatial Point Process Models using Integrated Laplace Transformation (INLA), *The Annals of Applied Statistics*, **6**(4), 1499–1530.

Lindgren, F. and Rue, H. (2015), Bayesian Spatial Modelling with R-INLA, *Journal of Statistical Software*, **63**(29), 1–25.

Moller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998). Log Gaussian Cox Processes, *Scandinavian Journal of Statistics*, **25**, 451–482.

Moller, J. and Waagepetersen, R. P. (2007). Modern Statistics for Spatial Point Processes, *Scandinavian Journal of Statistics*, **34**, 643–711.

Schrodle, B., Held, L., Riebler, A. and Danuser, J. (2011). Using INLA for The Evaluation of Veterinary Surveillance Data form Switzerland, *Royal Statistical Society*, **60**, 261-279.

Rathbun, S. L., and Cressie, N. (1994). A Space-time Survival Point Process for a Longleaf Pine Forest in Southern Georgia, *Journal of the American Statistical Association*, **89**, 1164–1174.

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 B*, **71**, 319–392.

دومین سمینار آمار فضایی و کاربردهای آن
۳ - ۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1990

پیش‌گویی فضایی پاسخ‌های نرخ: مقایسه مدل‌های کریگینگ بتا_دوجمله‌ای و هم‌کریگینگ دوجمله‌ای

لیلا عابدین پور لیارجدمه *، حسین باغیشنی، نگار اقبال

گروه آمار، دانشگاه صنعتی شاهرود

چکیده: مدلی که به‌طور معمول برای برازش پاسخ‌های شمارشی فضایی در جوامع متناهی مورد استفاده قرار می‌گیرد، مبتنی بر توزیع دوجمله‌ای است. در مباحث کاربردی، به دلیل وجود بیش‌پراکنش معمول در داده‌ها، استفاده از این پذیره توزیعی چندان مناسب نیست و می‌تواند به استنباط ناکارا در برآورد پارامترها و پیش‌گویی فضایی منتهی شود. به منظور رفع نقایص مدل دوجمله‌ای، یک رهیافت جانشین برای مدل‌سازی نسبت‌های همبسته فضایی استفاده از توزیع بتا_دوجمله‌ای است. کریگینگ بتا_دوجمله‌ای در مقایسه با هم‌کریگینگ دوجمله‌ای ساده‌تر است و پیش‌گویی‌های دقیق‌تری را تحت شرایط مختلف ارایه می‌دهد. در این مقاله عملکرد این دو مدل را در پیشگویی فضایی پاسخ‌های نرخ، با استفاده از یک مطالعه شبیه‌سازی، مقایسه و تشریح می‌کنیم.

واژه‌های کلیدی: بیش‌پراکنش، تغییرنگار، کریگینگ بتا_دوجمله‌ای، هم‌کریگینگ دوجمله‌ای، میدان تصادفی فضایی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62M30، 91B72.

۱ مقدمه

پاسخ‌های شمارشی فضایی در جوامع متناهی در بسیاری از تحقیقات کاربردی در زمینه‌هایی مانند پزشکی، اقتصاد، زیست‌شناسی و علوم اجتماعی مورد توجه محققان قرار دارند. داده‌های فضایی مشاهداتی هستند که وابستگی آن‌ها ناشی از موقعیت جغرافیایی آن‌ها در فضای مورد مطالعه است. در بسیاری از موارد ماهیت پاسخ، تعدادی از افراد جامعه با یک ویژگی خاص است. به‌عنوان نمونه، نسبت افراد مبتلا به سرطان معده می‌تواند پدیده مورد علاقه مطالعه باشد. در این موارد مدل متداول مبتنی بر توزیع دوجمله‌ای تعریف می‌شود. ویژگی مهم این توزیع، کوچک‌تر بودن واریانس از میانگین است. اما در موقعیت‌های متعددی، این پذیره برقرار نیست و با مساله بیش‌پراکنش مواجه می‌شویم. بنابراین، در چنین شرایطی توزیع دوجمله‌ای نمی‌تواند ماهیت داده‌ها را به خوبی برازش دهد و استفاده از آن برای مدل‌بندی داده‌ها منجر به استنباط‌های آماری ناکارا در برآورد پارامترها، برآورد ساختار وابستگی

* **ارائه‌دهنده مقاله:** لیلا عابدین پور لیارجدمه، leila.abedinpour@yahoo.com

و پیش‌گویی فضایی داده‌ها می‌شود. با این مقدمه، مدل بتا_دوجمله‌ای، به دلیل وجود دو پارامتر شکل در توزیع بتا، جانشینی بهتر و منعطف‌تر برای مدل دوجمله‌ای است. این توزیع قادر است تمام حالت‌های بیش‌پراکنش، کم‌پراکنش و پراکنش متعادل را برای داده‌ها در نظر بگیرد.

مدل بتا_دوجمله‌ای اولین بار به صورت رسمی توسط اسکلام (۱۹۴۸) ارایه شد. اما مبانی آن ابتدا توسط پیرسن (۱۹۲۵) مطرح شد. ویلیامز (۱۹۷۵) داده‌های مربوط به تاثیر دارویی خاص بر روی حیوانات آزمایشگاهی را مورد بررسی قرار داد. در آزمایش وی متغیر پاسخ، تعداد توله‌هایی بود که در دوره شیرخوارگی زنده مانده‌اند. توزیع دوجمله‌ای با پارامتر ثابت p گزینه معمول برای این نوع داده‌ها بود. اما نکته قابل توجه در آزمایش او آن بود که نسبت بقای توله‌ها در یک زایمان، از یک زایمان به دیگری تغییر می‌کرد. به عبارت دیگر داده‌ها بیش‌پراکندگی داشتند و مدل دوجمله‌ای نمی‌توانست آن‌ها را به‌خوبی مدل‌بندی کند. ویلیامز از توزیع بتا_دوجمله‌ای به‌عنوان یک مدل انعطاف‌پذیر برای تغییرپذیری بین نمونه‌ها استفاده کرد. پل (۱۹۸۲) در تحلیل نسبت‌های جنین تاثیرپذیر در آزمایش‌های تراتولوژیکی دریافت که مدل بتا_دوجمله‌ای نسبت به مدل دوجمله‌ای بهتر عمل می‌کند. تارون (۱۹۸۲) نیز برای بسیاری از انواع غده‌ها به این نتیجه رسید که نرخ کنترل عمر غده‌ها بسیار تغییرپذیرتر از آن است که پذیره توزیع دوجمله‌ای برای آن‌ها مناسب باشد و توزیع بتا_دوجمله‌ای را برای نرخ عمر غده‌ها به‌کار برد.

تاکنون مدل‌های مختلفی برای تحلیل داده‌های فضایی غیر نرمال معرفی شده‌اند. نخستین مدل هم‌کریگینگ دوجمله‌ای، توسط لاجونی (۱۹۹۱) برای پیش‌گویی متغیر پاسخ نرخ فضایی بر اساس نسبت نمونه‌ای مشاهده‌شده از یک بیماری نادر، پیشنهاد شد. الیور و همکاران (۱۹۹۸) از این مدل برای تحلیل نرخ سرطان کودکان استفاده کردند. مونستیز و همکاران (۲۰۰۶) استفاده از مدل کریگینگ پواسون را برای داده‌های شمارشی همبسته فضایی پیشنهاد دادند. آن‌ها در مدل پیشنهادی خود از یک طرح وزن‌دار به منظور اختصاص دادن وزن بیش‌تر به مشاهدات بزرگ‌تر استفاده کردند. مدل هم‌کریگینگ دوجمله‌ای فاقد وزن است، بنابراین ممکن است به مکان‌هایی با اندازه نمونه کوچک اهمیت بیش‌تری اختصاص داده شود.

در مواردی که ساختار ناهمگن موجود در داده‌های شمارشی فضایی موجب بیش‌پراکندگی می‌شود، عدم انعطاف توزیع دوجمله‌ای استفاده از آن را محدود می‌کند. برای مرتفع کردن این مساله از رهیافت جانشین آن مبتنی بر توزیع بتا_دوجمله‌ای استفاده می‌کنیم (شواب و مارکس ، ۲۰۱۵). انعطاف‌پذیری توزیع بتا_دوجمله‌ای باعث شده است که این توزیع در برازش مجموعه داده‌های بیش‌پراکنده سهم به‌سزایی را ایفا کند. در بخش ۲ مشخصات مدل بتا_دوجمله‌ای را معرفی می‌کنیم. سپس برآورد تغییرنگار و معادلات کریگینگ مدل بتا_دوجمله‌ای را ارایه می‌دهیم. در بخش ۳ به مقایسه دو مدل کریگینگ بتا_دوجمله‌ای و هم‌کریگینگ دوجمله‌ای می‌پردازیم. در بخش ۴ کارایی دو مدل را در توانایی پیش‌گویی و برآورد پارامترها، در یک مطالعه شبیه‌سازی، مورد ارزیابی قرار می‌دهیم. در پایان نیز بحث و نتیجه‌گیری ارایه می‌شود.

۲ مدل بتا_دوجمله‌ای فضایی

فرض کنید میدان تصادفی $\{Z = Z(s);\ s \in D \subset \Re^2\}$ که در m مکان s_i، $i = 1, 2, \cdots m$، مشاهده شده است، دارای توزیع شرطی $Z_i|Y_i \sim bin(n_i, Y_i)$ باشد، به‌طوری که احتمال موفقیت Y_i دارای توزیع بتا با پارامترهای α و β و ساختار وابستگی Σ_Y است. در واقع، در این حالت میدان تصادفی Y_i همبسته فضایی و مدل زیربنایی فرآیند دوجمله‌ای محسوب می‌شود. به عبارت

دیگر، وابستگی فضایی فقط در سطح Y_i وجود دارد. علاوه بر این، فرض کنید میدان تصادفی Y_i مانای مرتبه دوم و همسانگرد[۱] باشد. یعنی میانگین توزیع بتا مقدار ثابت π و تابع کوواریانس فقط تابعی از فاصله مشاهدات به‌صورت $C_Y(||s_i - s_j||)$ است. افزون بر این، تابع تغییرنگار[۲] فضایی مدل را با $\gamma_Y(||s_i - s_j||)$ نشان می‌دهیم. بنابراین می‌توان نوشت

$$E(Y_i) = \frac{\alpha}{\alpha+\beta} = \pi \qquad \text{Var}(Y_i) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \sigma_Y^2.$$

با توجه به این ویژگی‌ها متغیر تصادفی Z_i دارای توزیع کناری بتا_دوجمله‌ای با تابع احتمال

$$P(Z_i = z_i) = \frac{\text{Beta}(\alpha + z_i, \beta + n_i - z_i)}{\text{Beta}(\alpha,\beta)}\binom{n_i}{z_i} \qquad z_i = 0, \cdots, m$$

است. با استفاده از قاعده مکرر امید ریاضی و واریانس شرطی، امید و واریانس توزیع بتا_دوجمله‌ای به‌صورت زیر نتیجه می‌شوند:

$$\begin{aligned} E(Z_i) &= E[E(Z_i|Y_i)] = n_i E(Y_i) = n_i\pi \\ \text{Var}(Z_i) &= \text{Var}[E(Z_i|Y_i)] + E[\text{Var}(Z_i|Y_i)] = n_i\sigma_Y^2(n_i-1) + n_i\pi(1-\pi). \end{aligned}$$

این روابط در حقیقت تاثیر میدان تصادفی پنهان بتا را با وارد کردن پارامترهای شکل α و β در میانگین و واریانس مدل بتا_دوجمله‌ای نسبت به دوجمله‌ای نشان می‌دهند. میانگین و واریانس توزیع دوجمله‌ای به‌ترتیب $n_i\pi$ و $n_i\pi(1-\pi)$ است. بنابراین عبارت $n_i\sigma_Y^2(n_i-1)$ در واریانس توزیع بتا_دوجمله‌ای به‌عنوان عامل افزاینده واریانس توزیع دوجمله‌ای عمل کرده و مساله بیش‌برازش داده‌ها را لحاظ می‌کند.

۱.۲ تغییرنگار فرآیند بتا_دوجمله‌ای

به راحتی می‌توان نشان داد تغییرنگار تجربی نسبت‌های نمونه‌ای بتا_دوجمله‌ای، یعنی γ_Z، و تغییرنگار فرآیند احتمالی بتای پنهان، یعنی γ_Y، در تساوی زیر صدق می‌کنند:

$$\gamma_Y(s_i - s_j) = \gamma_Z(s_i - s_j) - \frac{1}{2}\left(\frac{n_i+n_j}{n_i n_j}\right)\left(\pi_Y(1-\pi_Y) - \sigma_Y^2\right). \qquad (۱.۲)$$

اکنون با در نظر گرفتن m مشاهده $Z_1, \ldots, Z_m$، از توزیع دوجمله‌ای با اندازه نمونه n_i در مکان s_i یک برآوردگر تعدیل‌یافته برای تغییرنگار با استفاده از رابطه (۱.۲) به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned} \hat{\gamma}_Y(h) = \frac{1}{N(h)}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{n_i n_j}{n_i+n_j}\Bigg[&\frac{1}{2}\left(\frac{Z_i}{n_i} - \frac{Z_j}{n_j}\right)^2 - \\ &\frac{1}{2}\left(\frac{n_i+n_j}{n_i n_j}\right)\left(\hat{\pi}(1-\hat{\pi}) - \hat{\sigma}_Y^2\right)\Bigg] I_d(i,j) \sim h \end{aligned}$$

که در آن منظور از $I_d(i,j) \sim h$ تابع نشانگر برای نقاط i و j ای است که در فاصله h از هم قرار دارند. عدد $N(h)$ نیز تعداد نقاط همسایه در فاصله h است که بر حسب وزن‌های $\frac{n_i n_j}{n_i+n_j}$ برای همگن کردن تمام مکان‌های فضایی به‌صورت زیر تعریف می‌شود:

$$N(h) = \sum_{i,j}\frac{n_i n_j}{n_i+n_j} I_d(i,j) \sim h.$$

[۱] Isotropic
[۲] Variogram

به‌عبارت دیگر این وزن تغییر تعداد مشاهدات در هر مکان فضایی را نشان می‌دهد. اگر همه مکان‌های فضایی اندازه نمونه یکسانی را داشته باشند، یعنی برای $i \neq j$ $n_i = n_j$، آنگاه به تمام نقاط مکانی وزن ثابت n داده می‌شود. مولفه تصحیح اریبی، یعنی $\left(\frac{n_i+n_j}{n_i n_j}\right)(\hat{\pi}(۱-\hat{\pi})-\sigma_Y^۲)$، به‌طور مستقیم از رابطه (۱.۲) تغییرنگار به‌دست می‌آید. همچنین $\hat{\pi}$ و $\hat{\sigma}_Y^۲$ به‌ترتیب برآوردهای میانگین و واریانس توزیع بتای پنهان را نمایش می‌دهند. برآوردگر تغییرنگار با استفاده از پارامترهای شکل توزیع بتای پنهان به‌صورت زیر تبدیل می‌شود:

$$\hat{\gamma}_Y(h) = \frac{۱}{۲N(h)} \sum_{i=۱}^{m} \sum_{j=۱}^{m} \left[\frac{n_i n_j}{n_i + n_j} \left(\frac{Z_i}{n_i} - \frac{Z_j}{n_j} \right)^۲ - \left(\frac{\hat{\alpha}\hat{\beta}}{\left(\hat{\alpha}+\hat{\beta}\right)\left(\hat{\alpha}+\hat{\beta}+۱\right)} \right) \right] I_d(i,j) \sim h$$

۲.۲ پیش‌گویی فضایی فرآیند بتا_دوجمله‌ای: معادلات کریگینگ

برای هر مکان فضایی جدید $s_۰$ مقدار پیش‌گویی فرآیند بتای پنهان، $Y_۰ = Y(s_۰)$، می‌تواند بر حسب یک ترکیب خطی از نسبت‌های نمونه‌ای دوجمله‌ای $\frac{Z_i}{n_i}$ که در مکان‌های s_i مشاهده شده‌اند با وزن‌های λ_i، $i = ۱, \ldots, m$، پیش‌گویی شود. یعنی

$$Y_۰ = \sum_{i=۱}^{m} \lambda_i \frac{Z_i}{n_i}.$$

وزن‌های λ_i باید به‌گونه‌ای انتخاب شوند که پیش‌گوی $Y_۰$ یک پیش‌گوی نااریب با کم‌ترین واریانس[۳] (BLUP) باشد. برای اطمینان از ناریبی پیشگو باید قید $\sum_{i=۱}^{m} \lambda_i = ۱$ را روی ضرایب اعمال کنیم. با در نظر گرفتن این قید، دستگاه معادلات کریگینگ بتا_دوجمله‌ای به‌صورت زیر نتیجه می‌شوند:

$$\frac{\lambda_i}{n_i}\left[\pi(۱-\pi) - \sigma_Y^۲\right] + \sum_{j=۱}^{m} \lambda_j C_{ij} + \mu = C_{۰i} \qquad i \in \{۱, ۲, \cdots, n\}$$

که در آن C_{ij} کوواریانس بین مکان‌های فضایی s_i و s_j، $C_{۰i}$ کوواریانس بین مکان‌های s_i با مکان $s_۰$ و μ ضریب لاگرانژ است. با حل این معادلات، تحت معلوم بودن تابع کوواریانس، ضرایب کریگینگ λ_i برآورد می‌شوند و بر حسب آن می‌توان مقدار پیش‌گویی در نقطه $s_۰$ را به‌دست آورد.

۳ مقایسه مدل کریگینگ بتا_دوجمله‌ای و هم‌کریگینگ دوجمله‌ای

لاجونی (۱۹۹۱) یک روش مشابه را برای برآورد احتمال زیربنایی بر اساس نسبت نمونه‌ای مشاهده‌شده به نام هم‌کریگینگ دوجمله‌ای معرفی کرد. در این مدل نسبت‌های نمونه‌ای به‌عنوان متغیرهای کمکی برای میدان تصادفی پنهان مدل‌بندی می‌شوند. دلیل نام‌گذاری هم‌کریگینگ نیز همین واقعیت است. برآوردگر تغییرنگار هم‌کریگینگ دوجمله‌ای با استفاده از نمادگذاری تغییرنگار بتا_دوجمله‌ای به صورت زیر است:

$$\hat{\gamma}_Y(h) = \hat{\gamma}_Z(h) - \frac{۱}{۲}\left(\hat{\pi}(۱-\hat{\pi}) - \hat{\sigma}_Y^۲\right)\frac{n_i + n_j}{n_i n_j}.$$

شکل کلی برآوردگر تغییرنگار مشابه مدل کریگینگ بتا_دوجمله‌ای است. با این وجود، برخی تفاوت‌های کلیدی در هر دو تغییرنگار و همچنین پذیره‌های دو مدل وجود دارند. به‌طور خاص این تفاوت‌ها عبارتند از

[۳] Best Linear Unbiased Predictor

(۱) در مدل کریگینگ بتا_دوجمله‌ای به صراحت فرض می‌شود میدان تصادفی پنهان، همبسته فضایی است و از توزیع بتا پیروی می‌کند. اما در مدل هم‌کریگینگ دوجمله‌ای هیچ پذیره توزیعی بر روی مدل زیربنایی وجود ندارد.

(۲) در مدل کریگینگ بتا_دوجمله‌ای یک شرط اضافی بر روی پارامترهای برآوردشده میانگین و واریانس، یعنی $\hat{\pi}$ و $\hat{\sigma}_Y^2$، گذاشته می‌شود تا میدان تصادفی پنهان را بین $[۰, ۱]$ محدود کند. در مقابل هم‌کریگینگ دوجمله‌ای چنین شرطی را لازم ندارد و به‌جای آن از میانگین و واریانس نسبت‌های نمونه‌ای برای برآورد $\hat{\pi}$ و $\hat{\sigma}_Y^2$ استفاده می‌شود.

(۳) عبارت تصحیح اریبی در هم‌کریگینگ دوجمله‌ای مقداری ثابت است.

(۴) در مدل کریگینگ بتا_دوجمله‌ای، تغییرنگار برآوردشده نسبت‌های نمونه‌ای به هر یک از جفت مکان‌ها با توجه به اختلاف واریانس‌ها وزن اختصاص می‌دهد، به‌طوری که با استفاده از این وزن به اختلاف بین جفت مکان‌های فضایی با اندازه نمونه بزرگ‌تر اهمیت بیش‌تری داده می‌شود. در مدل هم‌کریگینگ دوجمله‌ای هیچ گونه وزنی به تغییرنگار نسبت‌های نمونه‌ای اختصاص داده نمی‌شود. این عامل ممکن است سبب شود به نقاطی با اندازه نمونه کوچک اهمیت بیش‌تری داده شود.

با در نظر گرفتن این اختلاف‌ها، انتظار داریم مدل کریگینگ بتا_دوجمله‌ای، برآوردهای دقیق‌تری برای پارامترها و خطای پیش‌بینی کوچک‌تری تولید کند.

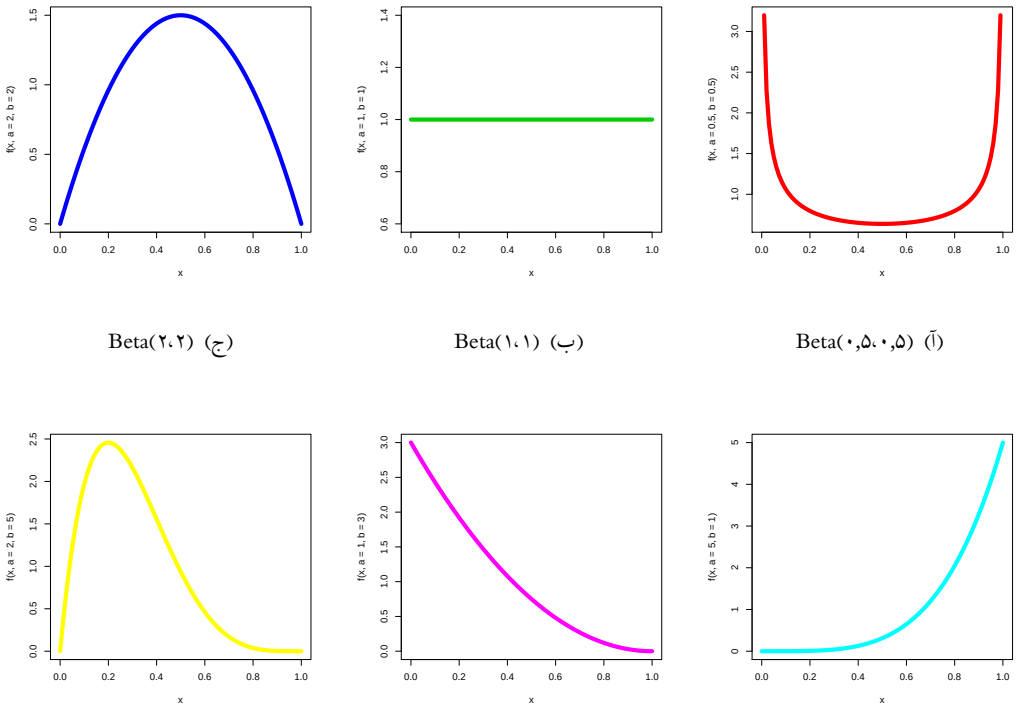
۴ ارزیابی مدل با مطالعه شبیه‌سازی

برای ارزیابی عملکرد مدل بتا_دوجمله‌ای پیشنهادی در برآورد پارامترها و پیش‌گویی فضایی، از یک مطالعه شبیه‌سازی تقریبا جامع استفاده کردیم. به منظور نمایش انعطاف توزیع‌های بتا که شامل توزیع‌های متقارن و چوله می‌باشد، در این مطالعه شش توزیع بتای مختلف را انتخاب کردیم. توابع چگالی احتمال توزیع‌های انتخاب‌شده در شکل ۱ نشان داده شده‌اند. برای شبیه‌سازی ۵۴ سناریوی مختلف تولید داده را تعریف کردیم؛ برای هر توزیع بتا سه اندازه نمونه در نظر گرفتیم: محدوده بین ۲:۵ برای اندازه نمونه کوچک، محدوده بین ۱۰:۲۰ برای اندازه نمونه متوسط و محدوده بین ۵۰:۱۰۰ برای اندازه نمونه بزرگ. پارامتر دامنه فضایی را نیز سه مقدار ۵، ۱۰ و ۱۵ در نظر گرفتیم. پاسخ‌ها و فرآیند بتای همبسته فضایی را با استفاده از روش تبدیل معکوس تولید کردیم. الگوریتم تبدیل معکوس برای تولید داده‌های بتا_دوجمله‌ای همبسته فضایی شامل مراحل زیر است:

(۱) ابتدا میدان تصادفی نرمال $X(s)$ با ماتریس کوواریانس فضایی Σ_X و تابع کوواریانس فضایی γ_X را شبیه‌سازی می‌کنیم. در این مطالعه شبیه‌سازی نقاط مکانی، گره‌های یک مشبکه فضایی ۲۰ × ۲۰ در نظر گرفته شدند. برای تابع کوواریانس فضایی نیز یک مدل کروی با پارامترهای واقعی اثر قطعه‌ای $\tau_X^2 = ۰$، دامنه فضایی $\phi_X = ۵, ۱۰, ۱۵$ و ازاره فضایی $\sigma_X^2 = ۱$ در نظر گرفته شد. بنابراین

$$X(s) \sim N_m(\mu_X = ۰, \Sigma_X)$$

$$\gamma_X(h) = \begin{cases} \tau_X^2 + \sigma_X^2 \left(\frac{۳h}{۲\phi_X} - \frac{h^۳}{۲\phi_x^۳} \right) & ۰ < h < \phi_X \\ \tau_X^2 + \sigma_X^2 & h \geq \phi_X \end{cases}$$



(آ) Beta(۰٫۵،۰٫۵) (ب) Beta(۱،۱) (ج) Beta(۲،۲)

(د) Beta(۵،۱) (ه) Beta(۱،۳) (و) Beta(۲،۵)

شکل ۱: توابع چگالی توزیع بتا برای مقادیر مختلف پارامترهای توزیع

(۲) برای هر نقطه مکانی در مشبکه فضایی، تابع توزیع تجمعی توزیع نرمال استاندارد، یعنی $\Phi_X(X_i)$، را برای تبدیل میدان تصادفی نرمال به میدان تصادفی بتا محاسبه می‌کنیم. اگر تابع چندک توزیع بتا $Y_i \sim \text{Beta}(\alpha, \beta)$ را با $\Xi_Y^{-1}(\cdot)$ نشان دهیم، آن‌گاه

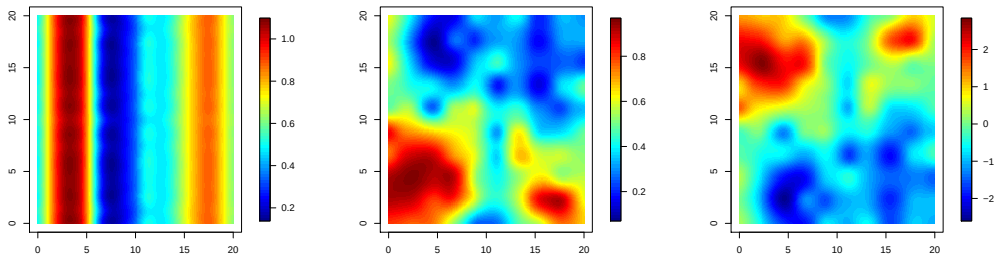
$$\Xi_Y^{-1}(\Phi(X_i)) = Y_i.$$

چون هر دو توزیع پیوسته هستند، تبدیل انجام‌شده یک به یک خواهد بود. توزیع بتا دارای ساختار وابستگی Σ_Y است، که این ساختار هم از یک مدل کروی با اثر قطعه‌ای $\tau_Y^2 = 0$ پیروی می‌کند. پارامتر دامنه نیز مشابه مدل میدان تصادفی نرمال است. فقط پارامتر ازاره فضایی برای میدان تصادفی بتا برابر با واریانس توزیع بتا یعنی

$$\sigma_Y^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

خواهد بود.

(۳) یک زیرمجموعه m تایی از مکان‌های فضایی با اندازه نمونه تصادفی n_i را به‌صورت تصادفی انتخاب می‌کنیم. سپس در هر مکان نمونه‌گیری، نمونه تصادفی دوجمله‌ای $Z_i|Y_i \sim Bin(n_i, Y_i)$ تولید می‌شود. بنابراین متغیر تصادفی Z_i دارای توزیع کناری بتا_دوجمله‌ای همبسته فضایی است. شکل ۲ فرآیند تولید داده‌های بتا_دوجمله‌ای همبسته فضایی را نمایش می‌دهد.



(آ) میدان تصادفی نرمال (ب) میدان تصادفی بتا (ج) میدان حاصل از نسبت‌های دوجمله‌ای

شکل ۲: میدان‌های شبیه‌سازی‌شده نرمال و بتا با روش تبدیل معکوس

۱.۴ روش‌های مختلف برآورد تغییرنگار

در متون آمار فضایی، برای برآورد تغییرنگار سه استراتژی معمول برای تعیین تعداد نقاط همسایگی و تابع زیان به‌صورت زیر پیشنهاد شده‌اند:

(۱)

$$N(h) = \sum_{i,j} I_d(i,j) \sim h \qquad L_1(\theta) = \sum_k N(h)\left[\gamma_Y(k) - \gamma_\theta(k)\right]$$

(۲)

$$N^*(h) = \sum_{i,j} \frac{n_i n_j}{n_i + n_j} I_d\left(i,j\right) \sim h \qquad L_2(\theta) = \sum_k N^*(h)\left[\gamma_Y(k) - \gamma_\theta(k)\right]$$

(۳)

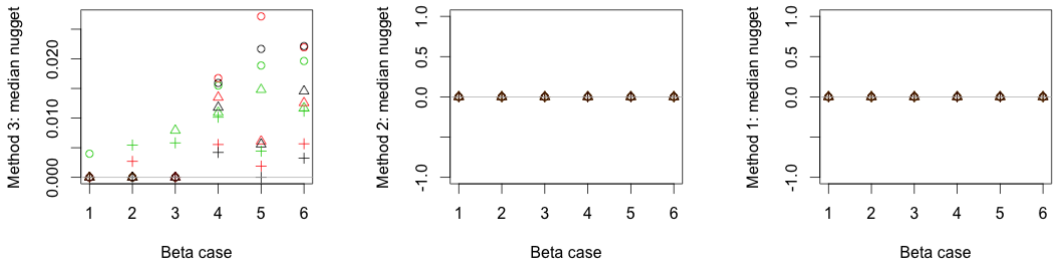
$$N(h) = \sum_{i,j} I_d(i,j) \sim h \qquad L_3(\theta) = \sum_k \left[\frac{\gamma^*(k) - \gamma_\theta(k)}{\gamma_\theta(k)}\right]^2$$

توابع زیان دو استراتژی اول و دوم یکسان هستند و تنها اختلاف آن‌ها در تعداد مشاهدات موجود برای هر فاصله h می‌باشد. استراتژی اول فاقد وزن است و استراتژی دوم از وزن تعداد مشاهدات یعنی $\frac{n_i n_j}{n_i+n_j}$ برای هر (i,j) استفاده می‌کند.

برای هر یک از ۵۴ سناریو شامل شش توزیع بتا، سه اندازه نمونه مختلف و سه دامنه فضایی، ۱۰۰ مجموعه داده شبیه‌سازی کردیم. سپس برای تمام مجموعه‌های تولیدشده هر سه استراتژی برآورد پارامتر را اجرا کردیم. شکل ۳ برآورد اثر قطعه‌ای[۴] را با استفاده از سه استرانژی نشان می‌دهد. در این شکل دامنه ۵ با ◯، ۱۰ با △ و ۱۵ با + نشان داده شده‌اند. همچنین اندازه نمونه کوچک با رنگ سبز، اندازه نمونه متوسط با رنگ قرمز و اندازه نمونه بزرگ با رنگ سیاه نمایش داده شده‌اند. با توجه به نتایج شبیه‌سازی‌ها، استراتژی سوم نسبت به دو استراتژی دیگر بدترین عملکرد را نشان می‌دهد. در این روش اثر قطعه‌ای برای توزیع‌های متقارن به‌خوبی برآورد شده است، اما تقریبا برای تمام توزیع‌های نامتقارن بیش‌برآورد شده است. کرسی (۱۹۹۳) استفاده از استراتژی سوم را برای توزیع‌های گاوسی پیشنهاد داد؛ بنابراین عملکرد نامناسب این روش برای داده‌های دوجمله‌ای شمارشی کاملا طبیعی است.

[۴] Nugget effect

برآوردهای میانه و انحراف معیار برای سایر پارامترهای تغییرنگار را با استفاده از هر سه استراتژی مورد بررسی قرار دادیم. به دلیل حجیم بودن خروجی‌ها از گزارش نتایج آن‌ها خودداری و تنها نتیجه‌گیری نهایی را گزارش می‌کنیم: ۱) انحراف معیار اثر قطعه‌ای برای روش دوم کمی کم‌تر است، ۲) برآورد میانه ازاره[۵] فضایی در روش اول و دوم یکسان است، و ۳) روش دوم دامنه فضایی را کمی بهتر برآورد می‌کند. با این نتیجه‌گیری، برای برآورد پارامترهای تغییرنگار و استفاده از آن‌ها در پیش‌گویی از روش دوم استفاده کردیم. برای روش هم‌کریگینگ دوجمله‌ای نیز استراتژی دوم انتخاب شد.



(آ) میانه اثرقطعه‌ای با استفاده از استراتژی اول (ب) میانه اثرقطعه‌ای با استفاده از استراتژی دوم (ج) میانه اثرقطعه‌ای با استفاده از استراتژی سوم

شکل ۳: برآورد میانه اثر قطعه‌ای تغییرنگار بتا_دوجمله‌ای با استراتژی‌های مختلف

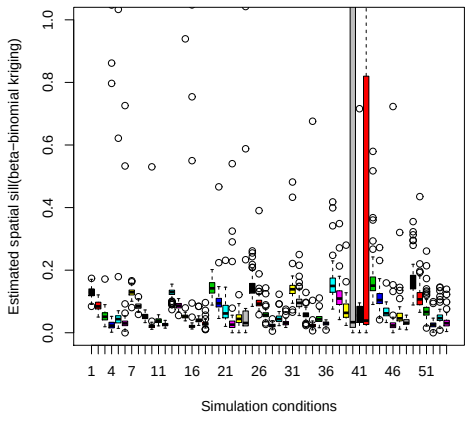
۲.۴ برآورد پارامترهای ساختار وابستگی فضایی

برای ارزیابی عملکرد برآورد پارامترها، برای هر سناریو ۱۰۰ مجموعه داده شبیه‌سازی کردیم. همان‌طور که در زیربخش قبلی بیان شد، با توجه به نتایج حاصل از شبیه‌سازی، برای برآورد پارامترها از تابع زیان $L_2(\cdot)$ و تعریف تعداد نقاط همسایه $N^*(h)$، در استراتژی دوم، استفاده کردیم. برآوردهای پارامترهای تغییرنگار برای هر دو مدل بتا_دوجمله‌ای و دوجمله‌ای از می‌نیمم کردن تابع زیان $L_2(\theta)$، که در تغییرنگار تجربی و تغییرنگار واقعی (کروی) مقداردهی شده است، به‌دست آمده‌اند. نمودارهای جعبه‌ای برآوردهای پارامترهای اثر قطعه‌ای، ازاره و دامنه برای مدل بتا_دوجمله‌ای در شکل ۴ و برای مدل دوجمله‌ای در شکل ۵ نشان داده شده‌اند. در هر دو شکل، ۱۸ سناریوی اول مربوط به دامنه ۵، ۱۸ سناریو دوم مربوط به دامنه ۱۰ و ۱۸ سناریو سوم مربوط به دامنه ۱۵ هستند. همچنین ۶ سناریو اول هر دامنه فضایی، اندازه نمونه کوچک، ۶ سناریو دوم اندازه نمونه متوسط و ۶ سناریو سوم اندازه نمونه بزرگ را نشان می‌دهند. پارامترهای واقعی دامنه فضایی به ترتیب برای این سناریوها ۵، ۱۰، و ۱۵ می‌باشند. اثر قطعه‌ای واقعی صفر است که تقریبا در تمام شبیه‌سازی‌ها به درستی برآورد شده است. اثر قطعه‌ای مدل دوجمله‌ای حتی کوچک‌تر از اثر قطعه‌ای کریگینگ بتا_دوجمله‌ای برآورد شده است. همچنین با توجه به دو نمودار جعبه‌ای مربوط به برآورد این پارامتر در هر دو مدل، می‌توان گفت اندازه نمونه در برآورد اثر قطعه‌ای تاثیری ندارد. در واقع الگوی قابل تشخیصی در شکل معلوم نیست که این نشان می‌دهد ۱) اصلاح تغییرنگار بتا_دوجمله‌ای، برآورد پارامترهای فرآیند بتای پنهان را بهبود بخشیده است؛ ۲) اصلاحی مشابه در تغییرنگار دوجمله‌ای نیز بهبودی مشابه را در پی داشته است.

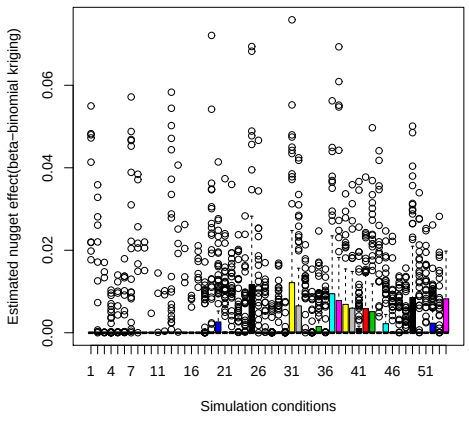
[۵] Sill

نمودارهای جعبه‌ای برآورد پارامتر ازاره در هر دو مدل، چوله به راست هستند و در مدل بتا_دوجمله‌ای تعداد نقاط پرت کم‌تری نسبت به اثر قطعه‌ای وجود دارد. با افزایش دامنه فضایی نقاط پرت بیش‌تر هم می‌شوند. واریانس برآورد این پارامتر در مدل دوجمله‌ای خیلی بزرگ‌تر از مدل بتا_دوجمله‌ای است. مقادیر میانه ازاره فضایی تقریبا همیشه کوچک‌تر از تغییرپذیری میدان زیربنایی بتا است. در حقیقت تفاوت بین میانه ازاره و ازاره فضایی واقعی شدید است. این نشان می‌دهد که هم‌کریگینگ دوجمله‌ای ممکن است با همگرایی در تقابل باشد، زیرا در بسیاری از موارد برآوردهای ازاره برای داده‌ها بسیار غیرمنطقی هستند. در بعضی سناریوها، بیش از ۲۵ درصد شبیه‌سازی‌ها مقدار برآوردشده ازاره کاملا بی‌معنی دارند. از طرفی، برای مدل بتا_دوجمله‌ای نیز برآوردهای ازاره بزرگ‌تر از یک وجود دارند که با توجه به کراندار بودن توزیع بتا در فاصله صفر و یک طبیعی نیست و نشان می‌دهد همگرایی برای بعضی از توزیع‌ها ممکن است مشکل‌ساز شود.

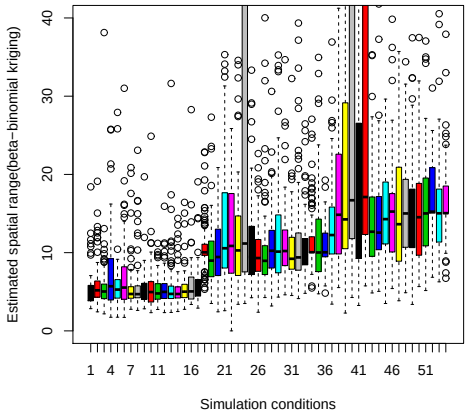
برای پارامتر دامنه فضایی، در مدل بتا_دوجمله‌ای، هر سه مقدار در نظر گرفته‌شده به خوبی بازیابی شده‌اند. روندی تقریبا مشابه برای مدل دوجمله‌ای برقرار است اما مشابه ازاره، واریانس برآوردها نسبت به مدل بتا_دوجمله‌ای خیلی بزرگ‌تر است.



(ب) ازاره فضایی

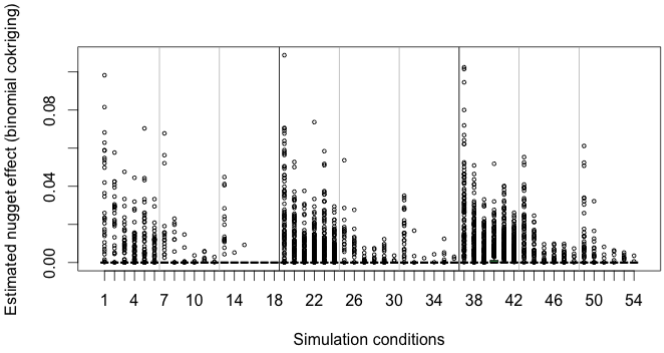


(آ) اثر قطعه‌ای

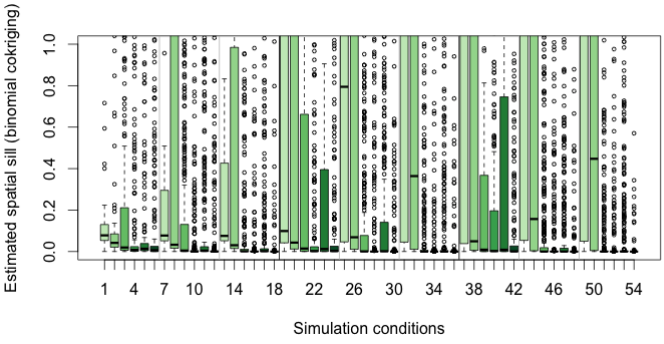


(ج) دامنه فضایی

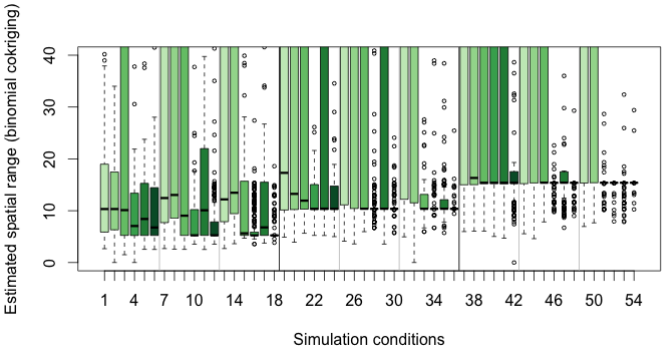
شکل ۴: برآورد پارامترهای تغییرنگار بتا_دوجمله‌ای برای ۵۴ سناریو بر اساس ۱۰۰ مجموعه داده شبیه‌سازی‌شده



(آ) اثر قطعه‌ای



(ب) ازاره فضایی



(ج) دامنه فضایی

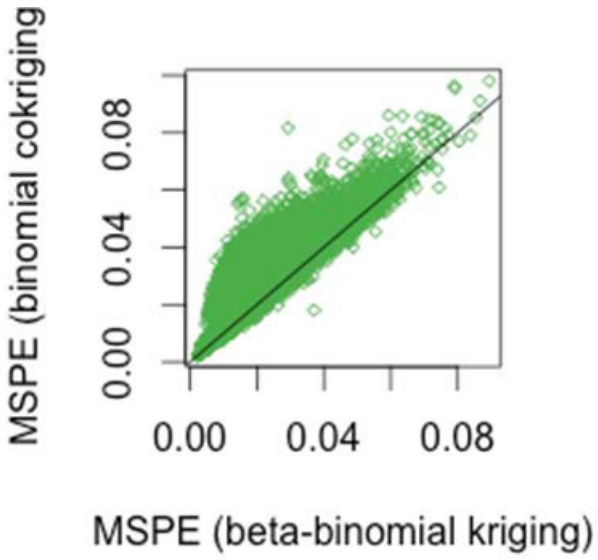
شکل ۵: برآورد پارامترهای تغییرنگار دوجمله‌ای برای ۵۴ سناریو بر اساس ۱۰۰ مجموعه داده شبیه‌سازی‌شده

با توجه به برآوردهای پارامترها در مدل هم‌کریگینگ دوجمله‌ای، می‌توان گفت که نگرانی جدی در مورد استفاده از این مدل برای پیش‌گویی فضایی وجود دارد. به نظر می‌رسد در این مدل، بهترین سناریوی استفاده از آن سناریویی با اندازه نمونه بزرگ، دامنه فضایی نسبتا بزرگ و توزیع زیربنایی بتا با تغییرپذیری کم است.

۳.۴ پیش‌گویی فضایی

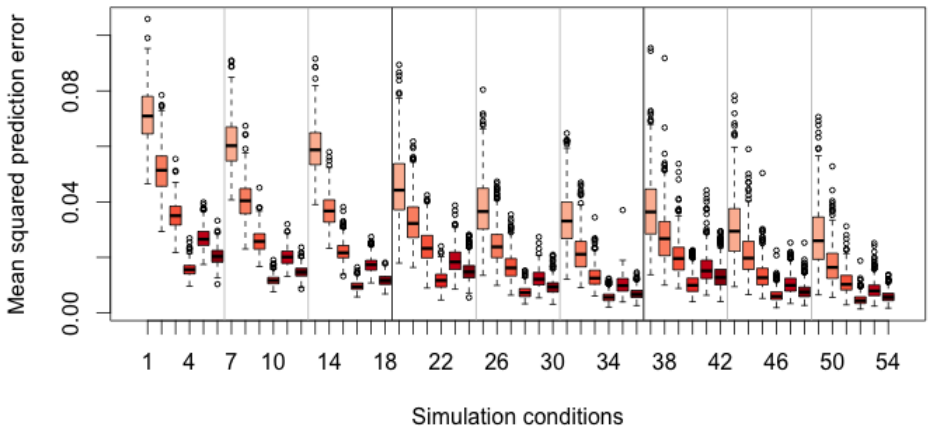
برای بررسی عملکرد پیش‌گویی دو روش معرفی‌شده، ۱۰۰ مجموعه داده را (به ازای هر سناریوی شبیه‌سازی) بر روی یک مشبکه ۵۰ × ۵۰ شامل ۲۵۰۰ گره تولید کردیم و با استفاده از دو روش کریگینگ بتا_دوجمله‌ای و هم‌کریگینگ دوجمله‌ای (با پارامترهای برآوردشده تغییرنگار) مقادیر نسبت‌های دوجمله‌ای را، برای هر مجموعه داده، پیش‌گویی کردیم. سپس میانگین توان دوم خطای پیش‌گویی[۶] (MSPE) را برای دو مدل بتا_دوجمله‌ای و دوجمله‌ای محاسبه کردیم. شکل ۶ نمودار پراکنش مقادیر MSPE دو روش را نشان می‌دهد. این شکل مقادیر MSPE تجمیع‌شده همه سناریوها را با هم نشان می‌دهد. از این شکل واضح است که مدل کریگینگ بتا_دوجمله‌ای، بر حسب معیار MSPE، عملکرد کاملا برتری نسبت به مدل هم‌کریگینگ دوجمله‌ای دارد. در واقع، در ۷۹/۴ درصد شبیه‌سازی‌ها، کریگینگ بتا_دوجمله‌ای MSPE کم‌تری از هم‌کریگینگ دوجمله‌ای تولید کرده است. در مواردی هم که هم‌کریگینگ دوجمله‌ای MSPE کوچک‌تری دارد، اختلاف نسبتا ناچیز است.

برای بررسی تاثیر حجم نمونه، و اندازه دامنه فضایی بر قدرت پیش‌گویی، مقادیر MSPE را برای دو مدل به تفکیک هر ۵۴ سناریو در شکل ۷ گزارش کرده‌ایم. با توجه به این دو شکل اختلاف توانایی دو روش، به‌ویژه در اندازه نمونه کوچک، مشهود است. ولی دامنه تاثیر چندانی ندارد. بنابراین می‌توان نتیجه‌گیری کرد که برای اندازه‌های نمونه کوچک و متوسط دقت کریگینگ بتا_دوجمله‌ای بیشتر است و باید از آن استفاده کرد.

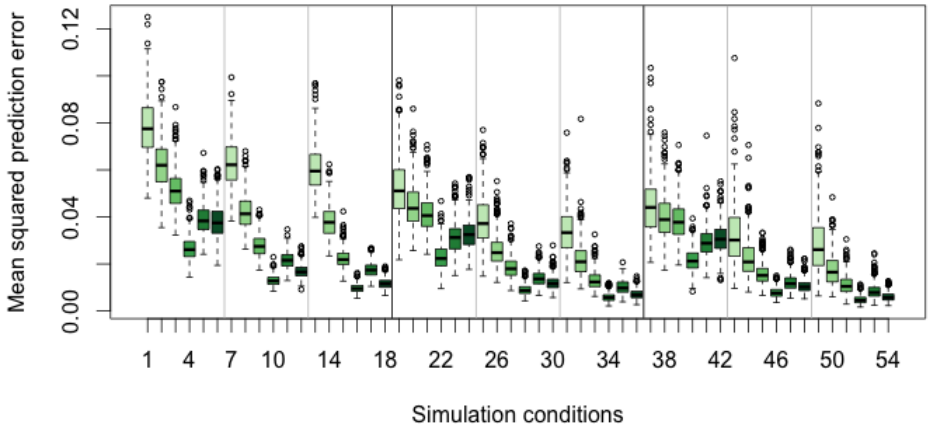


شکل ۶: مقایسه مقادیر MSPE برای دو روش کریگینگ بتا_دوجمله‌ای و هم‌کریگینگ دوجمله‌ای

[۶] Mean Squared Prediction Errors



(آ) میانگین توان دوم خطای پیش‌گویی مدل بتا_دوجمله‌ای



(ب) میانگین توان دوم خطای پیش‌گویی مدل دوجمله‌ای

شکل ۷: میانگین توان دوم خطای پیش‌گویی دو مدل به ازای ۵۴ سناریوی شبیه‌سازی

بحث و نتیجه‌گیری

در این مقاله برای رفع نقص‌های مدل دوجمله‌ای، به معرفی و بررسی عملکرد مدل کریگینگ بتا_دوجمله‌ای به‌عنوان جانشینی برای مدل‌بندی پاسخ‌های نرخ همبسته فضایی پرداختیم. با استفاده از یک مطالعه شبیه‌سازی، دقت این مدل در برآورد پارامترها و پیش‌گویی فضایی پاسخ‌های نرخ را نسبت به هم‌کریگینگ دوجمله‌ای، به‌ویژه برای اندازه‌های نمونه نسبتا کوچک، نشان دادیم. مدل پیشنهادی می‌تواند جانشینی ساده برای مدل‌های پیچیده‌تر فضایی، مانند مدل‌های آمیخته خطی تعمیم‌یافته و مدل‌های سلسله‌مراتبی بیزی، در اندازه نمونه کوچک باشد.

مراجع

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

Lajaunie, C. (1991). Local risk estimation for a rare noncontagious disease based on observed frequencies, *Note N-36/91/G, Centre de Géostatisque, Ecole des Mines de Paris.*

Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J. P., and Guinet, C. (2006). Geostatistical modelling of spatial distribution of Balaenoptera physalus in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts, *Ecological Modelling*, **193**, 615-628.

Oliver, M. A., Webster, R., Lajaunie, C., Muir, K.R., Parkes, S.E., Cameron, A. H., Stevens, M. C., and Mann, J. R. (1998). Binomial cokriging for estimating and mapping the risk of childhood cancer, *IMA Journal of Mathematics Appliead in Medicine and Biology*, **15**, 279-297.

Paul, S. R. (1982). Analysis of proportions of affected fetuses in teratology experiments, *Ecological Modelling*, **38**, 361-370.

Pearson, E. S. (1925). Bayes' theorem, examined in the light of experimental sampling, *Biometrika*, **17**, 388-442.

Schwab, A., and Marx, D. B. (2015). Beta-binomial kriging: an improved model for spatial rates, *Procedia Environmental Sciencos*, **27**, 30-37.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as a variable between the sets of trials, *Journal of the Royal Statistical Society, B*, **10**, 257–261.

Tarone, R. E. (1982). The use of historical control information in testing for a trend in proportions, *Ecological Modelling*, **34**, 69-76.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproductionand teratogenicity, *Biometrics*, **31**, 949- 952.

دومین سمینار آمار فضایی و کاربردهای آن
۳ – ۴ آبان ۱۳۹۶

تجزیه تانسور تابع کوواریانس میدان تصادفی فضایی–زمانی

میثم عصمتی*، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده: در تحلیل داده‌های فضایی–زمانی، متداول‌ترین روش برای لحاظ کردن ساختار همبستگی فضایی–زمانی داده‌ها، استفاده از تابع کوواریانس است، که نامعلوم و باید بر اساس مشاهدات برآورد شود. این روش نیازمند فرض‌های محدود کننده‌ای از قبیل فرض‌های مانایی و تفکیک‌پذیری میدان تصادفی می‌باشد. پذیرش این فرض‌ها برازش مدل‌های معتبر به تابع کوواریانس فضایی–زمانی را تسهیل می‌کند. اما ضرورتاً این فرض‌ها در مسائل کاربردی محقق نیستند. در این مقاله برای تعیین ساختار همبستگی فضایی–زمانی در میدان تصادفی نامانا و تفکیک‌ناپذیر، یک چارچوب احتمالی مبتنی بر تجزیه تانسور تابع کوواریانس میدان تصادفی فضایی–زمانی پیشنهاد می‌شود، سپس نحوه کاربست روش مطرح شده در پیش‌گویی فضایی–زمانی داده‌های سرعت باد نشان داده می‌شود.

واژه‌های کلیدی: داده‌های فضایی–زمانی، کوواریانس فضایی–زمانی، تجزیه تانسور.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 91B72.

۱ مقدمه

تانسور، آرایه‌ای چند بعدی از اعداد است که یک سری اعداد به‌طور خاص در یک جدول نامحسوس مرتب شده‌اند، که هر کدام از این اعداد سیستم مختصات خود را دارد. کلمه تانسور توسط ویلیام همیلتون (۱۸۴۶) برای تعمیم اعداد مختلط معرفی شد. حساب تانسور اولین بار توسط گرگریوریتچی–کورباسترو (۱۸۹۲) تحت عنوان حساب دیفرانسیل مطلق ارائه شد. در واقع ادامه مطالعات برنهارت ریمان و الوین برونو کریستوفل و دیگران در حساب دیفرانسیل مطلق بود. در واقع تانسور تعمیمی از مفاهیم اسکالر، بردار و ماتریس است.

تانسورها از آن جهت مورد استفاده قرار می‌گیرند که باعث ایجاد نظمی بین داده‌های یک مسئله و دسته‌بندی اطلاعات آن می‌شوند و بستر ریاضی مناسب و ساده‌ای را جهت فرمول‌بندی و حل مسائل متعدد در مباحث گوناگون فیزیک نظیر مکانیک سیالات و نسبیت عام فراهم می‌کنند. تجزیه تانسور توسط هیچکاک (۱۹۲۷) معرفی شد. اما در آن زمان کمتر مورد توجه قرار گرفت تا زمانی که

* **ارایه‌دهنده مقاله:** میثم عصمتی، m.esmati@modares.ac.ir

توکر (۱۹۶۳)، کارول و چانگ (۱۹۷۰)، هارشمن (۱۹۷۰) و کراسکال (۱۹۷۷) روش‌های مختلف تجزیه تانسور را تعمیم دادند و باعث انجام پژوهش‌های بر اساس تجزیه تانسور شد.

در چند دهه اخیر محققان علاقمند به گسترش تجزیه تانسور در سایر زمینه های علمی شدند. برای مثال در جبر خطی عددی لانگویل و استورت (۲۰۰۴) در آنالیز عددی، کرومس‌کیج و همکاران (۲۰۰۷) در داده کاوی آکار و همکاران (۲۰۰۶)، در تحلیل گراف بادر و کلدا (۲۰۰۷) و برای تحلیل کاربردی داده‌های چندطرفه کروننبرگ (۲۰۰۸) از تجزیه تانسور بهره بردند. امروزه مدل‌های تانسور مرتبه بالا را تجزیه چندطرفه می‌نامند. تجزیه چندتایی کانونی[۱] (CP) تانسورهای با مراتب بالا برای نشان دادن و تحلیل داده‌های چند بعدی بسیار مفید است. چندین بسته محاسباتی در نرم‌افزار MATLAB برای کار با تانسورها فراهم آورده شده است (بادر و کلدا ، ۲۰۰۷).

پیش‌گویی قابل اعتماد سرعت وزش باد در برخی حوزه‌ها از قبیل ایمنی هوانوردی، ایمنی سازه‌های ساختمانی و پل‌ها، انرژی باد و انتقال آلاینده‌ها بسیار مهم است. برای مثال در صنعت انرژی باد وضعیت سرعت باد کم ممکن است منجر به کاهش تولید انرژی و وضعیت باد شدید ممکن است به تعطیلی کارخانه‌ها منجر شود. در هوانوردی وضعیت باد شدید ممکن است عملیات پرواز را لغو کند. زیرا اغلب بادهای شدید از قبیل طوفان خطر سانحه هوایی را به همراه دارد. بنابراین تحلیل و پیشگویی فضایی و زمانی سرعت باد براساس اطلاعات در دسترس و حوزه کاربردی مورد توجه 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). با این وجود این ابزارهای پیش‌گویی باید عدم قطعیت‌های ناشی از منابع مختلف از قبیل دانش ناکامل مربوط به قوانین فیزیکی از روند و شرایط مرزی، داده ناکافی، خطاهای اندازه‌گیری و تقریب‌ها که شامل مدل‌سازی آماری است را مورد بررسی قرار دهد.

در تحلیل سری زمانی تغییرات زمانی سرعت باد مانند فرآیند تصادفی مانا و مدل‌های اتورگرسیو از قبیل میانگین متحرک اتورگرسیو[۲] (ARMA) یا انواع آن که برای پیشگویی استفاده می‌شود مدل‌بندی شده است. غالباً هموارسازی فضایی به منظور لحاظ کردن ویژگی‌های منطقه انجام می‌شود. در مدل‌بندی فضایی_زمانی هم تغییرات فضایی و هم تغییرات زمانی سرعت باد در یک مدل لحاظ شده و پیشگویی بر طبق آن صورت می‌گیرد.

هسلت و رفتری (۱۹۸۹) یک ابزار پیش‌گویی فضایی_زمانی برای مجموعه داده‌های باد کشور ایرلند که شامل متوسط سرعت باد روزانه اندازه‌گیری شده در ۱۱ ایستگاه بین سال‌های ۱۹۶۱ تا ۱۹۷۸ ارائه دادند. آن‌ها وابستگی فضایی را با استفاده از کریگیدن کرسی و ویکل (۲۰۱۱) و وابستگی زمانی را با استفاده از مدل‌های میانگین متحرک جمع بسته اتورگرسیو[۳] (ARIMA) مدل‌بندی کردند. در این مقاله پیشگویی براساس مدل حاصل از تجزیه تانسور ماتریس کواریانس فضایی_زمانی ارائه می‌شود. سپس دقت پیش‌گویی‌ها بر اساس ملاک اعتبارسنجی متقابل مورد ارزیابی قرار می‌گیرد. آنگاه نحوه کاربست این مدل در تحلیل داده‌های سرعت باد نشان داده خواهد شد. در بخش نهایی نیز به بحث و نتیجه‌گیری پرداخته خواهد شد.

[۱] Canonical Decomposition Parallel Factor
[۲] Autoregressive Moving Average
[۳] Autoregressive Integrated Moving Average

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

$$Z(\theta)^\top = \{Z(s_۱,t_۱,\theta), Z(s_۲,t_۱,\theta), ..., Z(s_{N_s},t_۱,\theta), Z(s_۱,t_۲,\theta), ..., Z(s_{N_s},t_{N_t},\theta)\}^\top \quad (۱.۲)$$

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

$$\text{Cov}(Z(s_i,t,\theta), Z(s_j,t,\theta)) = E\{[Z(s_i,t,\theta) - \overline{Z}(s_i,t)][Z(s_j,t,\theta) - \overline{Z}(s_j,t)]\} \quad (۲.۲)$$

تعریف می‌شود، که ماتریس کوواریانس آن به صورت تانسور مرتبه سوم با اندازه $N_sN_t \times N_sN_t$ است و با C نمایش داده می‌شود. اکنون پیش‌گویی میدان تصادفی بر اساس تابع کوواریانس (۲.۲) طی چهار مرحله ارائه می‌شود.

مرحله ۱ ـ جمع‌آوری و آماده‌سازی داده‌ها: وقتی داده‌ها شامل اندازه‌گیری در دوره زمانی یک سال یا بیشتر باشند لازم است اثرات فصلی یا دوره‌ای از داده‌ها حذف شود.

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

مرحله ۳ـ تجزیه تانسور ماتریس کوواریانس: یک میدان تصادفی با استفاده از حالت غالب این تجزیه تولید می‌شود. درایه‌های ماتریس تانسور کوواریانس C به صورت

$$c_{i,j,k} = E\{(Z(s_i,t_k,\theta) - \overline{Z}(s_i,t_k,\theta))(Z(s_j,t_k,\theta) - \overline{Z}(s_j,t_k,\theta))\} \quad (۳.۲)$$

ایجاد می‌شود. برای استفاده از تجزیه تانسور در شبیه‌سازی فرآیندهای فضایی‌ـ‌زمانی، روشی توسط گاش و سوریوانشی (۲۰۱۵) مطرح شده است، به‌طوری که ابتدا ماتریس کوواریانس C به صورت

$$\widehat{C} = \left[\sum_{r=۱}^{R} \lambda_r . v_r^{(۱)} \circ v_r^{(۲)}\right] \circ v^{(۳)} \qquad \lambda_r \in R \quad v_r^{(۱)}, v_r^{(۲)} \in R^{N_s} \quad v^{(۳)} \in R^{N_t} \quad (۴.۲)$$

تجزیه می‌شود، که در آن $\{v_r^{(۱)}\}_{r=۱}^{R}$ بردارهایی متعامد، $\circ$ نماد ضرب خارجی بردارها و $\widehat{C}$ تقریبی از تانسور کوواریانس C است. خطای این تقریب به صورت

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

تعریف می‌شود، که در آن $||.||_{Fr}$ نرم فروبنیوس است و برای تانسور کوواریانس C به صورت

$$||C||_{Fr} = \sqrt{\sum_{i=۱}^{I_۱}\sum_{j=۱}^{I_۲}\sum_{k=۱}^{I_۳} c_{i,j,k}^۲} \quad (۶.۲)$$

تعریف می‌شود (کلدا و بادر ، ۲۰۰۹). برای محاسبه تجزیه (۴.۲) و مینیمم کردن خطا به روش کم‌ترین توان‌های دوم مقابل استفاده می‌شود.

دی لاسور و همکاران (۲۰۰) نشان داد تقارن کوواریانس فضایی در هر لحظه زمانی متضمن تساوی $v_r^{(1)} = v_r^{(2)}$ به ازای همه rها است. بنابراین تحقق میدان تصادفی به صورت

$$Z(\theta) = \overline{Z} + \sum_{r=1}^{R} \sqrt{\lambda_r} v_r^{(1)} \sqrt{v^{(3)^T}} \eta_r \quad v_r^{(1)} \in R^{N_s}, v^{(3)} \in R^{N_t} \qquad (۷.۲)$$

قابل تولید است، که در آن نیز η_rها عضوهای یک مجموعه از متغیرهای تصادفی مستقل با میانگین صفر هستند.

۳ تحلیل داده‌های سرعت باد

در این بخش مدل معرفی شده در بخش ۲ برای تحلیل داده‌های سرعت باد مورد استفاده قرار می‌گیرد. این مجموعه داده‌ها که توسط هسلت و رفتری (۱۹۸۹) و نایتینگ (۲۰۰۲) نیز مورد استفاده قرار گرفته‌اند، شامل مقادیر متوسط سرعت باد روزانه بر حسب گره دریایی[۴] اندازه‌گیری شده در یازده ایستگاه بادسنجی طی دوره زمانی سال ۱۹۶۱ تا ۱۹۷۸ در کشور ایرلند هستند و در سایت آماری http://lib.stat.cmu.edu/datasets/ موجود و قابل دسترسی است. متوسط سرعت باد روزانه بر روی ایستگاه‌ها به همراه مختصات جغرافیایی ایستگاه‌ها در جدول ۱ ارائه شده است. در این مقاله از این مجموعه داده برای پیش‌گویی فضایی_زمانی استفاده می‌شود. در پیش‌گویی فضایی_زمانی متوسط سرعت باد، هفتگی پیش‌گویی می‌شود، که درستی این پیش‌گویی برای چند هفته آتی مفید است. سپس برای نشان دادن قابلیت و آزمون درستی این مدل دو حالت متمایز در نظر گرفته شده است.

۱. پیش‌گویی در زمان: با فرض در اختیار داشتن مقادیر سرعت باد همه یازده ایستگاه در مدت هفده سال، سرعت باد برای همه این ایستگاه‌ها در هیجدهمین سال پیش‌گویی می‌شود.

۲. پیش‌گویی در مکان و زمان: با فرض در اختیار داشتن مقادیر سرعت باد ده ایستگاه در هفده سال نخست و هم‌چنین مقادیر ایستگاه یازدهم فقط در هفدهمین سال، آن‌گاه سرعت باد برای یازدهمین ایستگاه در هیجدهمین سال پیش‌گویی می‌شود.

موقعیت جغرافیایی ایستگاه‌های بادسنجی در کشور ایرلند و نمودار جعبه‌ای مربوط به متوسط سرعت باد روزانه هر یک از ایستگاه‌ها در شکل ۱ نمایش داده شده‌اند. همان‌طور که در شکل ۱ ملاحظه می‌شود ایستگاه‌های شماره هشت و نه به‌ترتیب دارای بیشترین و کمترین میزان متوسط سرعت باد هستند. نکته قابل توجه در این نمودار، چولگی توزیع متوسط سرعت باد روزانه در ایستگاه‌ها است، که باید مورد بررسی قرار گیرد. برای تحلیل فضایی_زمانی داده‌های متوسط سرعت باد روزانه، ابتدا ماهیت اولیه مشاهدات از نظر همگنی واریانس، نرمال بودن و مانایی در فضا و زمان مورد بررسی قرار می‌گیرد.

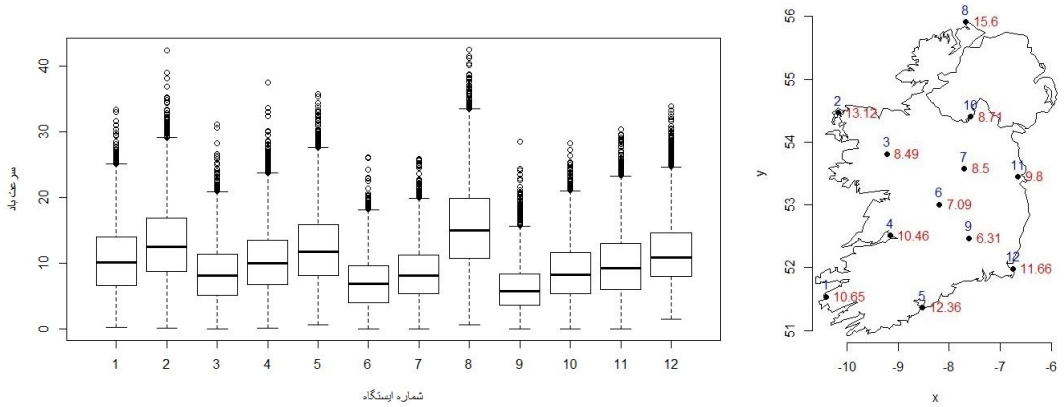
برای بررسی مانایی از حیث کوواریانس نمودار H_پراکنش داده‌ها در شکل ۲ ارائه شده است. همان طور که ملاحظه می‌شود نقاط اطراف نیمساز به طور نامتقارن پراکنده شده‌اند، بنابراین مانایی برای داده‌های متوسط سرعت باد روزانه برقرار نیست. البته دلیل آن می‌تواند وجود روند در داده‌ها باشد. برای تشخیص نامانایی از حیث میانگین، از نمودار $Z(s_i, t_j)$ها نیز در جهت‌های مختلف شرقی_غربی و شمالی_جنوبی استفاده شده است که با توجه به نمودارهای الف و ب در شکل ۳، نقاط به صورت تصادفی پراکنده و روند مشخصی نداشتند. این امر بیانگر ثابت بودن میانگین متوسط سرعت باد روزانه در جهت‌های مختلف است.

رفتار نمودارهای سری زمانی متوسط سرعت باد روزانه در ایستگاه دوبلین در طول زمان، که در شکل ۴ رسم شده، می‌تواند بیانگر وجود اثر فصلی در داده‌ها باشد. بنابراین لازم است ابتدا اثر فصلی از داده‌ها حذف، و تغییرات باقیمانده‌ها به صورت

[۴] 1 Knot=0.5148 m/s

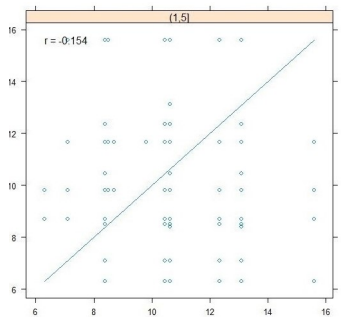
جدول ۱: نام و موقعیت ایستگاه‌های بادسنجی و متوسط سرعت باد روزانه بر روی ایستگاه‌ها در ایرلند

ردیف	ایستگاه	درجه و دقیقه		متر		متوسط سرعت
		طول	عرض	طول	عرض	بر حسب (گره دریایی)
۱	والنتیا	۱۰°۱۵′	۵۱°۵۶′	۵۸۵۹۳۸	۵۷۵۴۳۶۱	۱۰/۶۵
۲	بل‌مولت	۱۰°۰۰′	۵۴°۱۴′	۵۶۵۱۸۱	۶۰۰۹۹۴۴	۱۳/۱۲
۳	کلارموریس	۸°۵۹′	۵۳°۴۳′	۴۹۸۹۰۰	۵۹۵۱۹۹۸	۸/۴۹
۴	شانون	۸°۵۵′	۵۲°۴۲′	۴۹۴۳۶۸	۵۸۳۸۹۰۲	۱۰/۴۶
۵	روچزپوینت	۸°۱۵′	۵۱°۴۸′	۴۴۸۲۸۳	۵۷۳۹۰۶۰	۱۲/۳۶
۶	بر	۷°۵۳′	۵۳°۰۵′	۴۲۵۲۰۵	۵۸۸۲۱۲۳	۷/۱
۷	مولینگار	۷°۲۲′	۵۳°۳۲′	۳۹۱۷۴۶	۵۹۳۲۸۴۳	۸/۵
۸	مالین‌هد	۷°۲۰′	۵۵°۲۲′	۳۹۴۳۶۵	۶۱۳۶۸۵۹	۱۵/۶
۹	کیل‌کنی	۷°۱۶′	۵۲°۴۰′	۳۸۲۷۸۶	۵۸۳۶۶۰۰	۶/۳۶
۱۰	کلونس	۷°۱۴′	۵۴°۱۱′	۳۸۴۷۱۰	۶۰۰۵۳۶۱	۸/۷
۱۱	دوبلین	۶°۱۵′	۵۳°۲۶′	۳۱۷۳۱۹	۵۹۲۳۹۹۹	۹/۸

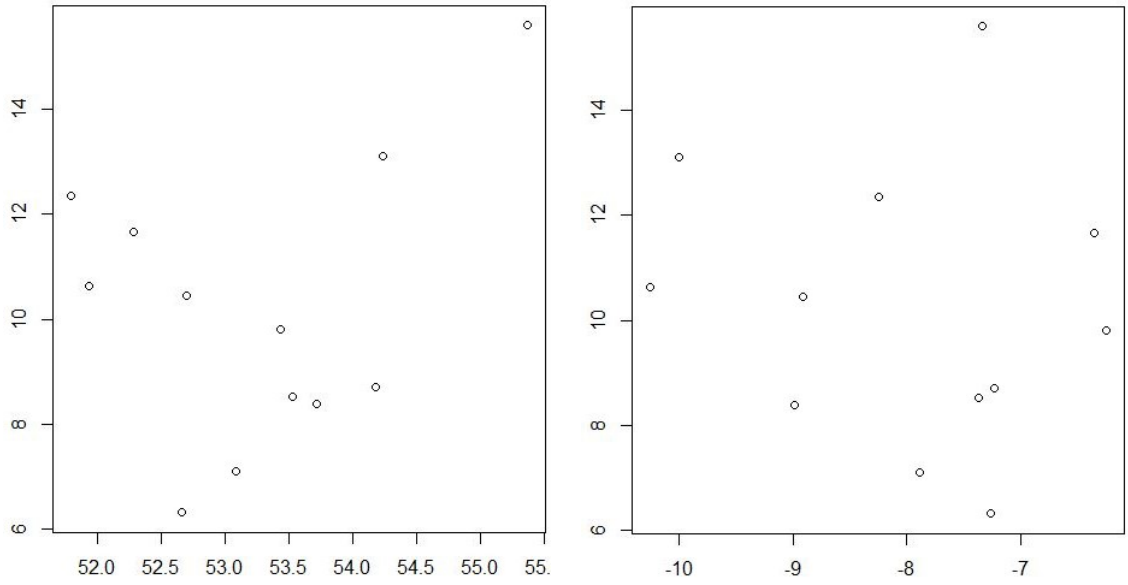


(ب) (الف)

شکل ۱: موقعیت ایستگاه‌های بادسنجی در کشور ایرلند و نمودار جعبه‌ای متوسط سرعت باد روزانه ایستگاه‌ها.



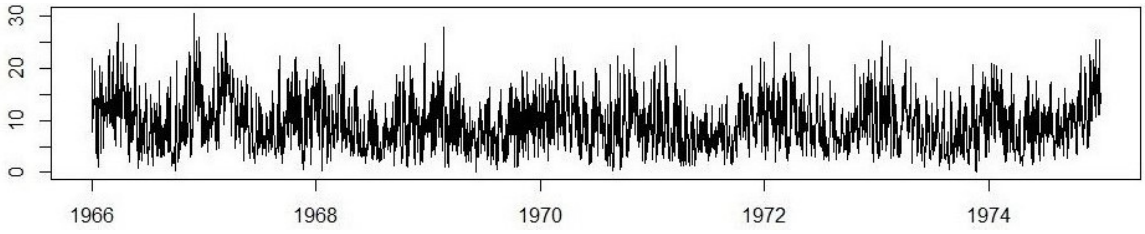
شکل ۲: نمودار Hـ پراکنش برای داده‌های متوسط سرعت باد روزانه.



(ب) (الف)

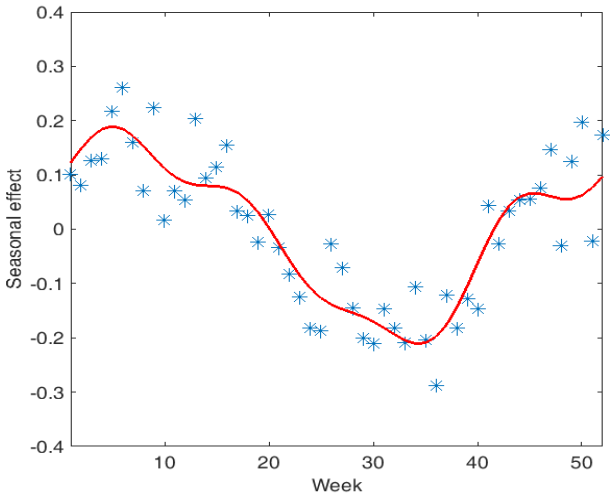
شکل ۳: الف: نمودار $Z(x,y)$ در مقابل x، ب: نمودار $Z(x,y)$ در مقابل y.

آماری مدل‌بندی شود. برای شناسایی اثر فصلی، نمودار پراکنش متوسط سرعت باد هفتگی همه ایستگاه‌ها در هفده سال اول رسم می‌شود. همان‌طور که در شکل ۵ ملاحظه می‌شود الگوی مشخصی برای متوسط سرعت باد در طول یک سال قابل مشاهده است. در این نمودار ستاره‌ها نشان دهنده متوسط سرعت باد هفتگی همه ایستگاه‌ها در هفده سال اول و منحنی رسم شده بیانگر اثر فصلی برآورد شده است. برای بررسی فرض همگنی واریانس در داده‌های فضایی_زمانی از نمودار انحراف معیار در مقابل میانگین

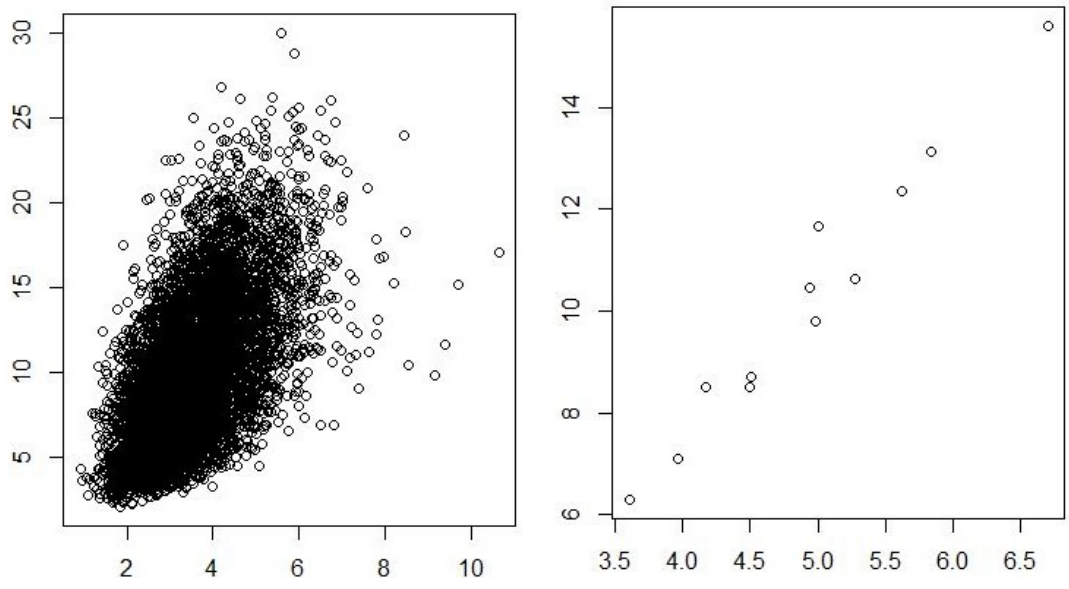


شکل ۴: نمودارهای سری زمانی متوسط سرعت باد روزانه در ایستگاه دوبلین.

استفاده شده است. همان‌طور که در شکل ۶ ملاحظه می‌شود، انحراف معیار متوسط‌های فضایی و متوسط‌های زمانی با افزایش میانگین، زیاد می‌شود، بنابراین واریانس داده‌ها نمی‌تواند همگن باشد. لازم است با انتخاب تبدیل مناسب برای داده‌ها، واریانس را همگن نمود. به‌طور معمول در مبحث سرعت باد برای همگن کردن واریانس از دو تبدیل لگاریتمی و ریشه دوم استفاده می‌شود (میرینگ و همکاران ، ۱۹۹۸). وقتی تبدیل لگاریتمی روی داده‌ها اعمال می‌شود، نمودار انحراف معیار در مقابل میانگین الگوی مشخصی دارد. اما وقتی تبدیل ریشه دوم روی داده‌ها اعمال می‌شود، همان‌طور که در شکل ۷ ملاحظه می‌شود هیچ الگوی مشخصی قابل رؤیت نیست. بنابراین همگنی واریانس بر روی فضا و زمان برای داده‌های تبدیل یافته برقرار است. بنابراین تبدیل مناسب برای همگن کردن واریانس تبدیل ریشه دوم است.



شکل ۵: اثر فصلی متوسط سرعت باد هفتگی (⋆) و مدل برازش شده (خط ممتد).

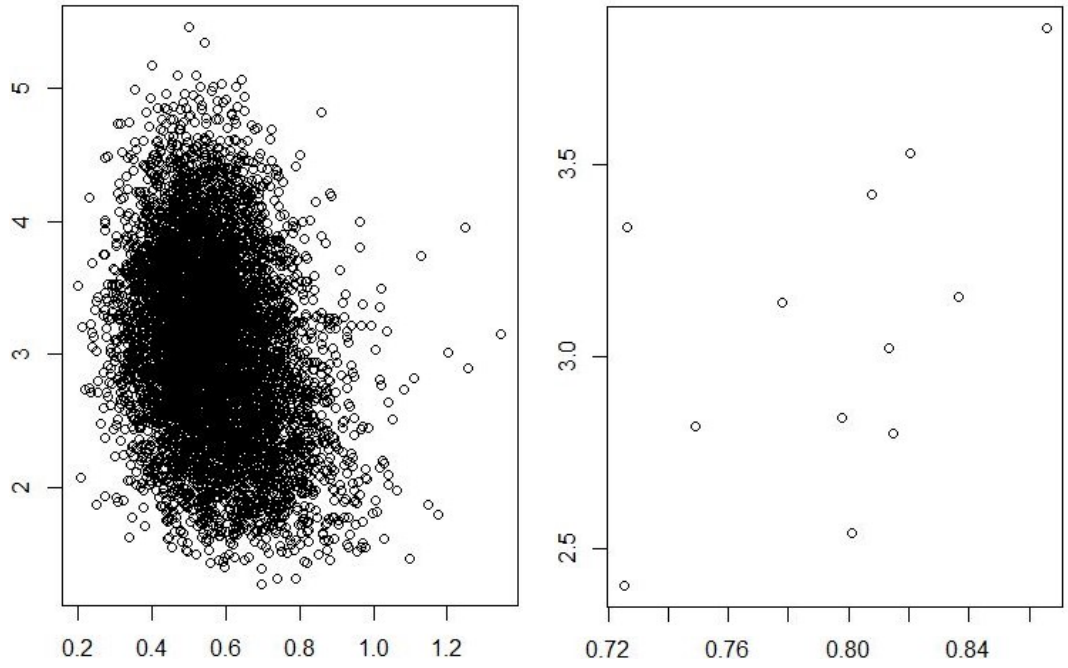


(ب) (الف)

شکل ۶: نمودار انحراف معیار در مقابل میانگین متوسط سرعت باد روزانه، الف- موقعیت ایستگاه‌ها و ب- لحظه‌های زمانی .

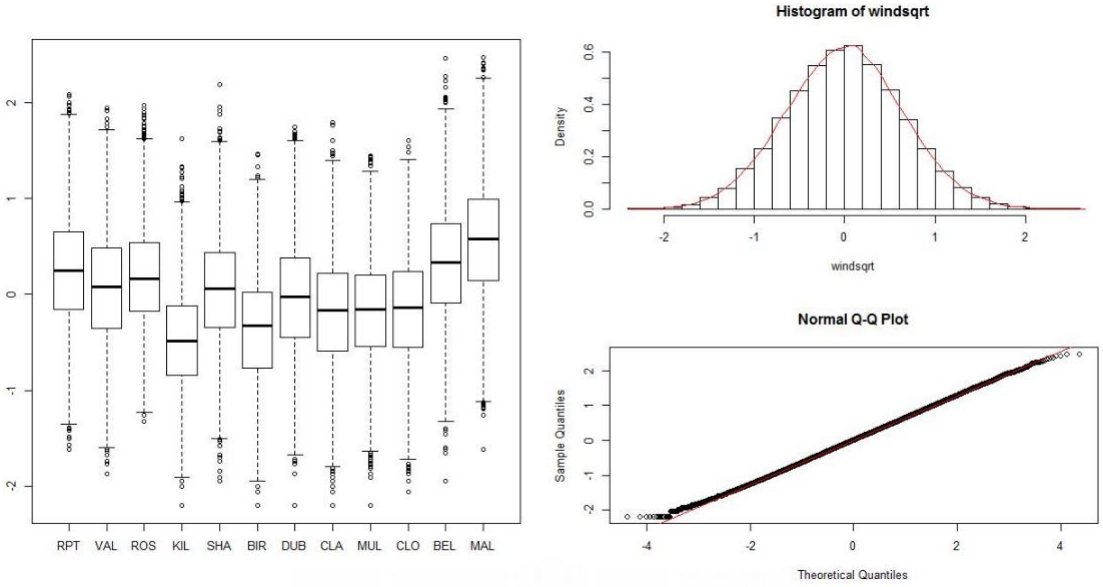
در هسلت و رفتری (۱۹۸۹) نشان داده شده است که گرفتن تبدیل ریشه دوم و حذف اثر فصلی منجر به حاصل شدن توزیع احتمالی تقریبا گاوسی می‌شود. بنابراین در اینجا این عملیات به خوبی صورت گرفته و مقدارهای به دست آمده تحت عنوان اندازه سرعت[۵] معرفی شده است. نمودارهای بافت‌نگار، Q-Q و جعبه‌ای که در شکل ۸ نمایش داده شده است، بیانگر این امر می‌باشد. بنابراین در زیر بخش بعد مدل‌بندی آماری، تحلیل و پیشگویی براساس اندازه سرعت انجام می‌شوند.

[۵] Velocity measure



(ب) (الف)

شکل ۷: نمودار انحراف معیار در مقابل میانگین ریشه دوم متوسط سرعت باد روزانه، الف_ موقعیت ایستگاه‌ها و ب_ لحظه‌های زمانی

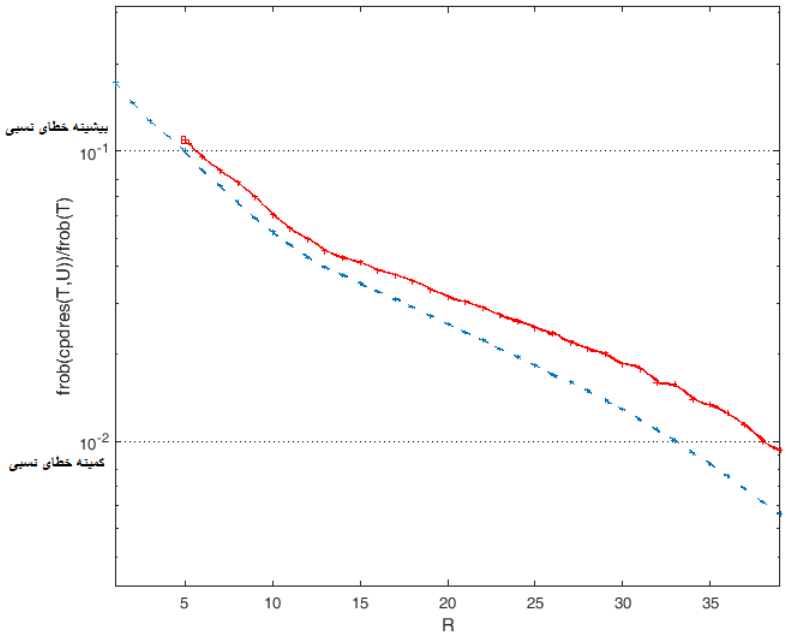


شکل ۸: نمودارهای بافت‌نگار، Q-Q و جعبه‌ای ریشه دوم متوسط سرعت باد روزانه

۱.۳ پیش‌گویی فضایی_زمانی

در این بخش میانگین هفتگی متوسط سرعت باد در پنج هفته آتی برای سال هیجدهم بدونه هیچ فرض محدود کننده مانایی مکانی و زمانی پیش‌گویی خواهد شد. برای حذف ماهیت فصلی داده‌ها، مجموعه‌ای بر اساس تبدیل‌های سینوس و کسینوس انتخاب

می‌شود که نخستین همساز دارای دوره زمانی ۵۲ هفته است. سپس ماتریس تانسور کوواریانس با درایه‌های (۳.۲) بر اساس داده‌ها برآورد شده و به‌صورت (۴.۲) تجزیه می‌شود. یکی از شناسه‌های مهم برای انجام این تجزیه تعیین تعداد رتبه تانسور است. برای این منظور مقادیر مختلف تانسور به ازای رتبه‌های مختلف در شکل ۹ رسم شده است. از کوچکترین عدد صحیح برای کران پایین متناسب با خطا شروع می‌شود که از بیشینه خطای واقعی کمتر است. کران پایین بر اساس خطای برشی[۶] مقادیر تکینی چندخطی تانسورها است. تعداد عبارات رتبه‌ـ یک تا زمانی که خطای نسبی مجانبی کمتر از خطای نسبی واقعی است افزایش می‌یابد و در آخر مناسب‌ترین مقدار برای رتبه تانسور تعیین می‌شود. در این‌جا به ازای $R = ۵$ یک تقریب همگرایی ارائه شده است که بر اساس آن خطای نسبی بین تانسور C و $\hat{C}$ بر طبق (۵.۲) برای پیشگویی هفتگی ۰/۱ است. در شکل ۹ خطای نسبی تجزیه تانسور را در مقابل تعداد رتبه‌ـ یک‌ها نشان می‌دهد، که در آن خط ممتد خطای نسبی تجزیه تانسور و خط چین بیانگر کران پایین خطاها است. برای تولید مقادیر میدان تصادفی (۷.۲) استفاده شده است، که در آن η_r دارای توزیع نرمال استاندارد منظور گردیده است.



شکل ۹: نمودار تعداد رتبه‌ـ یک‌های تانسور متوسط سرعت باد هفتگی.

الف‌ـ پیش‌گویی در زمان: میانگین هفتگی متوسط سرعت باد در پنج هفته آتی برای هر ایستگاه در هجدهمین سال پیش‌گویی شده و پس از به توان دوم رساندن نتایج و اضافه کردن میانگین و اثر فصلی نتایج پیشگویی سرعت باد در جدول ۲ ارائه شده است. شماره ایستگاه‌ها در ستون اول، نام ایستگاه‌ها در ستون دوم، مقادیر مشاهده شده واقعی در ستون سوم، مقدار پیشگویی در ستون چهارم، میانگین توان دوم خطای پیشگویی در ستون بعدی و کران‌های بالایی و پایینی بازه پیش‌گویی ۹۵ درصدی مونت گمری و رانگیر (۲۰۰۷) در دو ستون آخر ارائه شده است. با توجه به مقادیر میانگین توان دوم خطا که تفاوت مقدار پیش‌گویی شده توسط مدل را از مقدار واقعی نشان می‌دهد، سرعت‌های پیش‌گویی شده برای ایستگاه‌های شماره ۸ و ۹ که به‌ترتیب دارای بیشترین و کمترین مقدار متوسط سرعت باد بودند، مقدار میانگین توان دوم خطا معنی‌دار شده است، اما برای سایر ایستگاه‌ها به خوبی پیشگویی شده و در داخل بازه پیش‌گویی ۹۵ درصدی قرار می‌گیرند.

[۶] Truncation error

جدول ۲: پیش‌گویی در زمان متوسط هفتگی سرعت باد در پنج هفته آتی

ردیف	نام ایستگاه	مشاهده	پیش‌گویی	میانگین توان دوم خطا	بازه پیش‌گویی ۹۵ درصدی	
					کران پایین	کران بالا
۱	والنتیا	۵/۸۶	۵/۶۹	۰/۵۹	۵/۱۳	۶/۲۶
۲	بل مولت	۷/۸۶	۷/۳۸	۰/۶۶	۶/۸۹	۷/۸۸
۳	کلارموریس	۵/۴۵	۵/۲۲	۰/۴۳	۴/۷۶	۵/۶۹
۴	شانون	۶/۱۲	۶/۲۶	۰/۶۹	۵/۶۳	۶/۸۷
۵	روچز پوینت	۷/۳۶	۷/۱۴	۰/۵۴	۶/۰۵	۸/۲۲
۶	بر	۴/۰۲	۴/۲۸	۰/۴۶	۳/۷۹	۴/۷۴
۷	مولینگار	۵/۰۸	۴/۹۴	۰/۲۶	۴/۵۶	۵/۴۴
۸	مالین	۱۰/۵۸	۸/۵۵	۴/۹۹	۷/۸۵	۹/۲۴
۹	کیل کنی	۳/۴۵	۴/۶۸	۲/۰۴	۴/۱۴	۵/۲۳
۱۰	کلونس	۴/۸۲	۵	۰/۳۷	۴/۵۶	۵/۴۴
۱۱	دوبلین	۶/۳۴	۶/۱	۰/۵۱	۵/۵۸	۶/۶

ب_ پیش‌گویی در مکان و زمان: میانگین هفتگی سرعت باد پیش‌گویی شده در ایستگاه یازدهم زمانی که فقط داده‌های سال هفدهم این ایستگاه موجود است با مقادیر مشاهده شده در جدول ۳ مقایسه می‌شود. همان‌طورکه ملاحظه می‌شود متوسط سرعت باد هفتگی برای همه ایستگاه‌ها به خوبی پیش‌گویی شده است. به طوری که مقادیر مشاهده شده برای تعداد زیادی از ایستگاه‌ها در داخل بازه پیش‌گویی ۹۵ درصدی قرار می‌گیرد.

جدول ۳: پیش‌گویی در مکان و زمان متوسط هفتگی سرعت باد در پنج هفته آتی

ردیف	نام ایستگاه	مشاهده	پیش‌گویی	میانگین توان دوم خطا	بازه پیش‌گویی ۹۵ درصد	
					کران پایین	کران بالا
۱	والنتیا	۵/۸۶	۵/۳۷	۰/۸۱	۴/۸	۵/۹۴
۲	بل مولت	۷/۸۶	۷/۲	۰/۹۳	۶/۵۱	۷/۵
۳	کلارموریس	۵/۴۵	۴/۸۶	۰/۷۳	۴/۳۹	۵/۳۲
۴	شانون	۶/۱۲	۵/۷۷	۰/۷	۵/۱۵	۶/۳۹
۵	روچز پوینت	۷/۳۶	۶/۷۳	۰/۸۷	۵/۶۴	۷/۸۱
۶	بر	۴/۰۲	۴/۲۶	۰/۴۶	۳/۷۸	۴/۷۴
۷	مولینگار	۵/۰۸	۴/۸۴	۰/۳	۴/۴۶	۵/۲۱
۸	مالین	۱۰/۵۸	۹/۶۹	۱/۶۵	۸/۹۹	۱۰/۳۸
۹	کیل کنی	۳/۴۵	۴/۲۶	۱/۱۷	۳/۷۱	۴/۸
۱۰	کلونس	۴/۸۲	۴/۹	۰/۳۴	۴/۴۷	۵/۳۴
۱۱	دوبلین	۶/۳۴	۵/۸۵	۰/۶۹	۵/۳۴	۶/۳۵

بحث و نتیجه‌گیری

در این مقاله مدل‌بندی ساختار همبستگی فضایی_زمانی یک میدان تصادفی با استفاده از تانسور کوواریانس و پیش‌گویی فضایی_زمانی مبتنی بر تجزیه تانسور مورد بررسی قرار گرفته است. به طور معمول ساختار همبستگی با استفاده از مدل‌های مختلف تابع کوواریانس فضایی زمانی مدل‌بندی می‌شود، که این روش نیازمند فرض‌های محدود کننده و در بسیاری موارد تعداد پارامترهای مدل تابع کوواریانس زیاد و باعث پیچیدگی محاسبات در برازش مدل می‌شود. در حقیقت هر چه مدل تعداد پارامتر کمتری داشته باشد احتمال بروز خطاهای محاسباتی در آن کمتر خواهد بود. بنابراین برای پیشگویی فضایی_زمانی در یک میدان تصادفی نامانا و افزایش سرعت محاسبات، از تجزیه تانسور کوواریانس فضایی_زمانی مورد بررسی قرار گرفت که نیازمند فرض‌های محدود کننده نیست و تعداد پارامترهای آن کمتر، خطاهای محاسباتی کاهش و دقت پیشگویی افزایش یافته است.

مراجع

Acar, E., Camtepe, S. A. and Yener, B. (2006), Collective Sampling and Analysis of High Order Tensors for Chatroom Communications, *in ISI 2006: Proceedings of the IEEE International Conference on Intelligence and Security Informatics, Lecture Notes in Journal of Computational Science 3975, Springer*, 213–224.

Bader, B. W. and Kolda, T. G. (2007), *MATLAB Tensor Toolbox*, Version 2.2. Available at http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox.

Carroll, J. D. and Chang, J. J. (1970), Analysis of Individual Differences in Multidimensional scaling via an N-way Generalization of Eckart-Young Decomposition, *Psychometrika*, **35**, 283–319.

Cressie N. and Wikle C. K. (2011), *Statistics for Spatio-Temporal Data*, John Wiley Sons Inc, Hoboken.

De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000), A Multilinear Singular Value Decomposition, *SIAM Journal on Matrix Analysis and Applications*, **21**, 1258–1278.

Ghosh, D. and Suryawanshi و A. (2015), Approximation of Spatio-Temporal Random Processes Using Tensor Decomposition, *Communications in Computational Physics*, **16**(1), 75 -95.

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

Harshman, R. A. (1970), Foundations of The PARAFAC Procedure: Models and Conditions for an ‘‘Explanatory’’ Multi-Modal Factor Analysis, *UCLA working papers in phonetics*, http://publish.uwo.ca/*harshman/wpppfac0.pdf, **16**, 1–84.

Hamilton, W. R. (1846), On Quaternions, or on a New System of Imaginaries in Algebra, *Journal of Science*, 1844–1850.

Haslett, J. and Raftery, A. E. (1989), Space-Time Modelling with Long-Memory Dependence: Assessing Ireland's Wind-Power Resource, *Journal of Applied Statistics*, **38**, 1–50.

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

Khoromskij, B. N. and Khoromskaia, V. (2007), Low Rank Tucker-Type Tensor Approximation to Classical Potentials, *Central European Journal of Mathematics*, **5**, 523–550.

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

Kroonenberg, P. M. (2008), *Applied Multiway Data Analysis*, Wiley, New York.

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.

Langville, A. N. and Stewart, W. J. (2004), A Kronecker Product Approximate Precondition for SANs, *Numerical Linear Algebra with Applications*, **11**, 723–752.

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

Meiring, W., Guttorp, P. and Sampson, P. D. (1998), Space-Time Estimation of Grid-Cell Hourly Ozone Levels for Assessment of a Deterministic Model, *Environmental Ecological Statistics*, **5**, 197-222.

Montgomery, D. C. and Runger, G. C. (2007), *Applied Statistics and Probability for Engineers*, Wiley, New York.

Ricci-Curbastro, G. (1892), Absolute Differential Calculus, *Bulletin des Sciences Mathématiques*, **16**, 167-189.

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. 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.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1390

نقشه‌بندی بروز بیماری افسردگی در استان همدان با استفاده از مدل فضایی بیز کامل

بهناز علافچی[۱]، پریسا چمن‌پرا[۲]، *محمد ابراهیم غفاری[۱]، علی قعله‌ای‌ها[۳]

[۱]گروه آمار زیستی، دانشگاه علوم پزشکی همدان

[۲]مرکز توسعه پژوهشهای بالینی، دانشگاه علوم پزشکی شیراز

[۳]گروه روانپزشکی مرکز تحقیقات اختلالات رفتاری و سوء مصرف مواد دانشگاه علوم پزشکی همدان

چکیده:

مقدمه: افسردگی یکی از شایع‌ترین اختلالات روانی و چهارمین علت اصلی ابتلا به سایر بیماری‌ها در افراد افسرده محسوب می‌شود. افسردگی در بدترین حالت می‌تواند منجر به خودکشی گردد. با توجه به اهمیت این بیماری و شیوع بالای آن انجام مطالعاتی در خصوص تشخیص مناطق پر خطر، امری ضروری است. یکی از روش‌های مناسب در تجزیه و تحلیل هر مجموعه از داده‌ها، تولید و بررسی نمودارهاست که در اپیدمیولوژی مکانی به آن‌ها نقشه بندی بیماری گویند. نقشه‌بندی بیماری شامل مجموعه‌ای از روش‌های آماری است که با تهیه نقشه‌های دقیق بر اساس شیوع، بروز و مرگ‌ومیر بیماری منجر به برآورد توزیع جغرافیایی بیماری می‌شود. در این پژوهش نقشه‌بندی بروز افسردگی در استان همدان بررسی شده است.

روش‌ها: در این مطالعه، داده‌ها موارد جدید بیماری افسردگی در استان همدان طی سال‌های ۱۳۸۷ تا ۱۳۹۵ به تفکیک شهرستان‌های این استان مورد تجزیه و تحلیل قرار گرفت. مدل فضایی بیز کامل (BYM) برای تعیین الگوی تغییرات مکانی خطر نسبی (RR) این بیماری استفاده شد.

نتایج: بالاترین نرخ بروز افسردگی به ترتیب مربوط به شهرستانهای همدان، بهار و اسدآباد بود.

بحث:به دلیل اینکه در تمامی نقشه‌های ترسیم شده به تفکیک سال‌های مختلف، بالاترین نرخ بروز افسردگی مربوط به شهرستان همدان بود، توصیه می‌شود، مسئولین سلامت مداخلات لازم را جهت کاهش خطر این بیماری در این شهرستان در پیش گیرند.

واژه‌های کلیدی: مدل‌های فضایی، بیز، نقشه‌بندی،افسردگی، همدان .

کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲pxx.

---

*ارایه‌دهنده مقاله: محمد ابراهیم غفاری، m.gh19@yahoo.com

۱ مقدمه

طبق تعریف سازمان جهانی بهداشت (درگاه سازمان جهانی بهداشت، ۲۰۱۷) سلامتی مجموعه‌ای از رفاه، آسایش کامل جسمانی، روانی و اجتماعی است که هیچ‌کدام بر دیگری برتری ندارد. طبق آمار WHO، ۵۲ میلیون نفر از مردم جهان در سنین مختلف از بیماری‌های شدید روانی رنج می‌برند و ۲۵۰ میلیون نفر بیماری خفیف روانی دارند، در ایران نیز این آمار از سایر کشورها کمتر نیست. اختلالات روان‌پزشکی نوعی بیماری است که افراد از مشکلات عاطفی و احساسی رنج می‌برند و حالات ذهنی غیر طبیعی دارند. چنین تصور می‌شود که این اختلالات در نتیجه بروز اشکال در نحوه عملکرد و فعالیت مغز ایجاد می‌شوند. اختلالات روان‌پزشکی تمامی جنبه‌های زندگی فرد را در بر می‌گیرند و می‌توانند زندگی اجتماعی و اقتصادی را تحت تاثیر قرار دهند (ساتو و همکاران ، ۲۰۰۶). پیشگیری از اختلالات روانی می‌تواند به طور عظیمی از عمل خودکشی پیشگیری نماید (فخاری و همکاران ، ۲۰۰۷). افسردگی یکی از شایع‌ترین اختلالات روان‌پزشکی و چهارمین علت اصلی بیماری محسوب می‌شود (چو و همکاران، ۲۰۰۹). بیش از ۳۰۰ میلیون نفردر تمام رده‌های سنی در سراسر جهان از افسردگی رنج می‌برند. افسردگی اصلی‌ترین علت ناتوانی در جهان و عامل اصلی شکل‌گیری سایر بیماری‌هاست (تیچن و یورک، ۲۰۱۳). هم‌ابتلایی بالایی بین افسردگی با سردرد (مرسکی و همکاران، ۱۹۸۵؛ ناداوکا و همکاران، ۱۹۹۷) و نیز سلامت روانی با سکته مغزیو بیماری‌های قلبی و عروقی وجود دارد (می و همکاران، ۲۰۰۲؛ راسول و همکاران، ۲۰۰۲).

شیوع افسردگی در زنان بیش از مردان است(درگاه سازمان جهانی بهداشت، ۲۰۱۷). افسردگی بر جنبه‌های هیجانی، رفتاری، شناختی و فیزیکی فرد تأثیر می‌گذارد و باعث بی‌انگیزگی، عدم علاقه به کارهاو عدم لذت، کاهش میل جنسی، احساس بی فایدگی و تهی بودن و ناامیدی، باورهای بدبینانه، فکر کردن به خودکشی، ناتوانی در تمرکز و تصمیم‌گیری، و اختلالات خواب می‌شود(کاردونا و همکاران، ۲۰۱۵). افسردگی به یک مسئله جهانی تبدیل شده و بار سنگینی بر سیستم مراقبت‌های بهداشتی و جامعه اعمال می‌نماید. در سالمندان نیز افسردگی موجب کاهش فعالیت‌های مراقبت از خود و کیفیت زندگی و کاهش عملکرد جسمی می‌شود. افسردگی بر تعامل اجتماعی افراد اثرات منفی به جای گذاشته و می‌تواند باعث ایجاد چالش‌های اقتصادی شود (هارادا و همکاران، ۲۰۱۲).

نقشه‌بندی یک روش منحصر به فرد و مؤثر به منظور نشان‌دادن توزیع پدیده‌ها در فضا و مکان است. از نقشه‌ها به منظور ثبت مشاهدات در یک شکل مختصرتر و یا برای اهدافی چون آنالیز فرضیاتی برای بررسی وجود یا عدم وجود ارتباط بین پدیده‌ها استفاده می‌شود. نقشه‌بندی بیماری ابزاری مناسب به منظور تعیین گروه‌های بیماری، شناسایی و نظارت بر بیماری‌های همه‌گیر، ارائه‌ی اطلاعات پایه در الگوی سلامت و نمای تغییرات الگوی بیماری در طول زمان است. همچنین دست اندرکاران نظام سلامت در مواردی که با محدودیت در تخصیص منابع روبرو هستند از این نقشه‌ها به منظور تشخیص نواحی نیازمندتر به تخصیص منابع استفاده می‌کنند. تأثیر محیط بر روی سلامت با استفاده از نمایش نقشه‌های ساخته شده برای بیماری بهتر درک می‌شود .محیط و شرایط آب و هوایی تأثیر بسزایي در میزان سلامت رواني افـراد از جمله ابتلا به بیماری افسردگیدارد که این امر بحث جغرافياي پزشكي را مطرح می‌کند (غیاث و همکاران، ۲۰۱۱).

کوک و همکاران (۲۰۱۲) بروز بیماری سل در هلند را نقشه‌بندی کردند. اندرسون و آلن (۲۰۱۱) در مطالعه‌ای به نقشه‌بندی شیوع بیماری کرم قلاب‌دار در آمیرکا پرداختند. از آنجایی که تا کنون مطالعه‌ای در زمینه نقشه‌بندی بیماری افسردگی انجام نشده است، هدف مطالعه حاضر نقشه‌بندی موارد بروز بیماری افسردگی در استان همدان است.

۲ مواد و روش‌ها

در این مطالعه از موارد جدید مشاهده شده افسردگی در استان همدان طی سالهای ۱۳۸۷ تا ۱۳۹۵ که توسط اطلاعات جمع آوری شده از بیمارستان سینا همدان استان همدان استخراج شده است، استفاده کردیم. براساس اطلاعات جمعیتی به دست آمده از این مرکز کل افراد در معرض خطر بیماری (معادل جمعیت استان) در سال ۱۳۹۰، ۱۷۵۸۲۶۸ نفر بوده است. در بسیاری از مطالعات متغیر پاسخ تعداد رخدادهای نادر مانند موارد جدید سرطان در یک جامعه در مدت زمانی خاص می باشد. در اینگونه حالات فرض می شود که متغیر پاسخ دارای توزیع پواسن است (اگرستی ، ۱۹۹۱).

فرض کنید O_i نشانگر تعداد موارد جدید بیماری افسردگی در شهرستان i ام و E_i نشانگر تعداد موارد مورد انتظار آن باشد به‌طوری که E_i از حاصلضرب جمعیت هر شهرستان (n_i) در بروز استانی آن (تعداد کل موارد جدید مشاهده شده آن بیماری تقسیم بر جمعیت کل استان) به دست می‌آید.

$$E_i = \frac{\sum_{i=1}^{9} O_i}{\sum_{i=1}^{9} n_i} \times n_i$$

O_i از توزیع پواسن با میانگین μ_i پیروی می‌کند به‌طوری که μ_i تابعی از تعداد موارد مورد انتظار (E_i) و خطر نسبی بیماری R_i در شهرستان iام $(\mu_i = E_i.R_i)$ می‌باشد. در اینجا R_i پارامتر مجهول مدل است. مجهول مدل است.

در مطالعه حاضر از مدل رایج بسیج و همکاران (۱۹۹۱) (BYM) برای لحاظ کردن اطلاعات مربوط به همسایه‌های مجاور هر شهرستان استفاده شده است. در این مدل لگاریتم خطر نسبی، $\log(R_i)$، به مجموع عرض از مبدا و دو مؤلفه اثرات تصادفی ساختاریافته و ساختارنایافته تجزیه می‌شود $(\log(R_i) = a + u_i + \nu_i)$. اثر تصادفی ساختارنایافته (اثر ناهمگنی بعلت ناهمبستگی)می باشد و مؤلفه‌ایست که عواملی که بین شهرستان ها تغییر میکند را مدلبندی می‌کند. این اثر از توزیع نرمال با میانگین صفر و واریانس τ_ν^2 پیروی می‌کند. اثر تصادفی ساختاریافته (اثر ناهمگنی به علت همبستگی) می‌باشد که برای هر شهرستان وابستگی مکانی فرض می‌کند. برای در نظر گرفتن این وابستگی، از مدل پیشنهاد شده توسط بسیج و همکاران (۱۹۹۱) استفاده کردیم که در آن به شهرستان‌های مجاور شهرستان مورد نظر وزن یک و برای شهرستان‌های غیرمجاور وزن صفر درنظر می‌گیرد. این مؤلفه توسط توزیع پیشین نرمال اتورگرسیوشرطی مدل‌بندی میشود به‌طوری که فرض می‌کند مؤلفه ساختاریافته مکانی مخصوص هر شهرستان بشرط معلوم بودن اثرات ساختاریافته دیگر شهرستان‌ها دارای توزیع نرمال با میانگینی برابر با متوسط همسایگی‌های شهرستان موردنظر و واریانسی برابر با نسبت معکوسی از تعداد شهرستان‌های مجاور هر شهرستان می‌باشد. لذا هرچه تعداد شهرستان‌های مجاور زیاد باشد دقت اثر ساختاریافته شهرستان مورد نظر بیشتر است.

$$[u_i|u_j, i \neq j, \tau_\nu^2] \sim N\left(\bar{u_i}, \tau_i^2\right)$$

$$\bar{u_i} = \frac{1}{\sum_j W_{ij}} \sum_j u_j W_{ij}$$

$$\tau_i^2 = \frac{\tau_u^2}{\sum_j W_{ij}}$$

$$W_{ij} = \begin{cases} 1 & \text{if} \quad i,j \quad \text{adjacent} \\ 0 & \text{otherwise} \end{cases}$$

در مدل‌بندی بیزی به تمامی پارامترهای مجهول، چه اثرات ثابت و چه اثرات تصادفی، توزیع پیشین اختصاص داده می‌شود. عرض از مبدا α، از توزیع $N(۰, ۰/۰۰۱)$و پارامترهای دقت τ_u و τ_ν از توزیع فوق پیشین گاما $\Gamma(۰/۰۱, ۰/۰۱)$ پیروی می‌کنند. در نهایت با استفاده از نرم افزار OpenBUGS مدل را به داده‌ها برازش دادیم. سپس از آزمون تشخیصی گلمن رابین[۱] برای بررسی همگرایی استفاده کردیم.

۳ نتایج و یافته‌ها

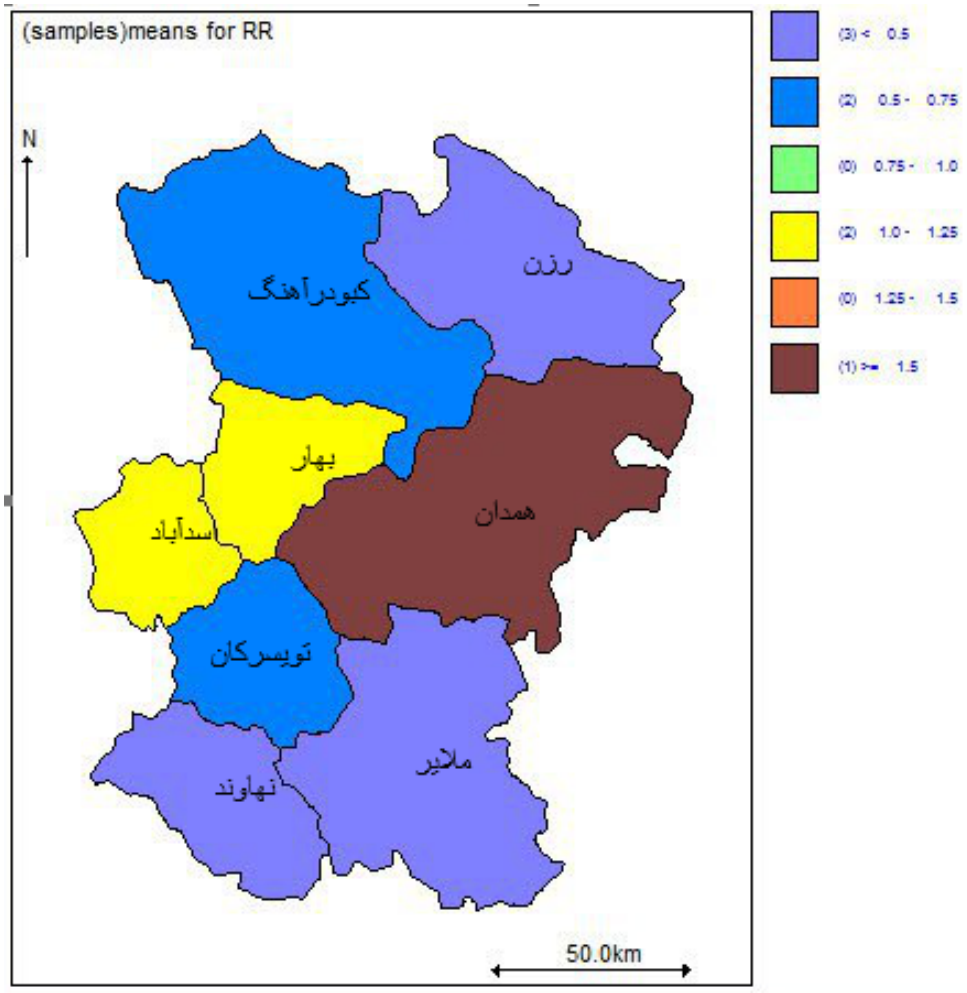
تعداد کل موارد جدید افراد مبتلا به بیماری افسردگی بین سال‌های ۱۳۸۷ و ۱۳۹۵ در استان همدان۳۴۰۲ مورد است که بیشترین تعداد مربوط به شهرستان همدان (۲۲۹۰) و کمترین مربوط به شهرستان رزن (۹۱) است. سایر اطلاعات مربوط به شهرستان‌ها در جدول ۱ نمایش داده شده است. شکل ۱ و شکل ۲ توزیع جغرافیایی خطر نسبی (RR) ابتلا به بیماری افسردگی در شهرستان‌های استان همدان را نمایش می‌دهند که مقادیر به ۶ خوشه تقسیم‌بندی شده است؛

$$RR < ۰/۵$$
$$۰/۵ < RR < ۰/۷۵$$
$$۰/۷۵ < RR < ۱/۰$$
$$۱/۰ < RR < ۱/۲۵$$
$$۱/۲۵ < RR < ۱/۵$$
$$RR > ۱/۵$$

جدول ۱: موارد مشاهده شده، جمعیت در معرض خطر و موارد مورد انتظار برای بیماری افسردگی در استان همدان در سال ۱۳۹۵

شهرستان	موارد مشاهده شده	جمعیت در معرض خطر	موارد مورد انتظار
همدان	۲۲۹۰	۶۹۴۳۰۶	۱۳۴۳/۳۸
بهار	۲۴۵	۱۲۳۸۶۹	۲۳۹/۶۷
تویسرکان	۱۰۲	۱۰۳۱۷۶	۲۰۰/۸۱
رزن	۹۱	۱۱۶۴۳۷	۲۲۵/۲۹
کبودرآهنگ	۱۶۷	۱۴۳۱۷۱	۲۷۷/۰۷
ملایر	۱۷۲	۲۸۷۹۸۲	۵۵۷/۲
نهاوند	۱۱۵	۱۸۱۷۱۱	۳۵۱/۵۹
اسدآباد	۲۲۰	۱۰۷۰۰۶	۲۰۷/۰۴

[۱] Gelman-Rubin



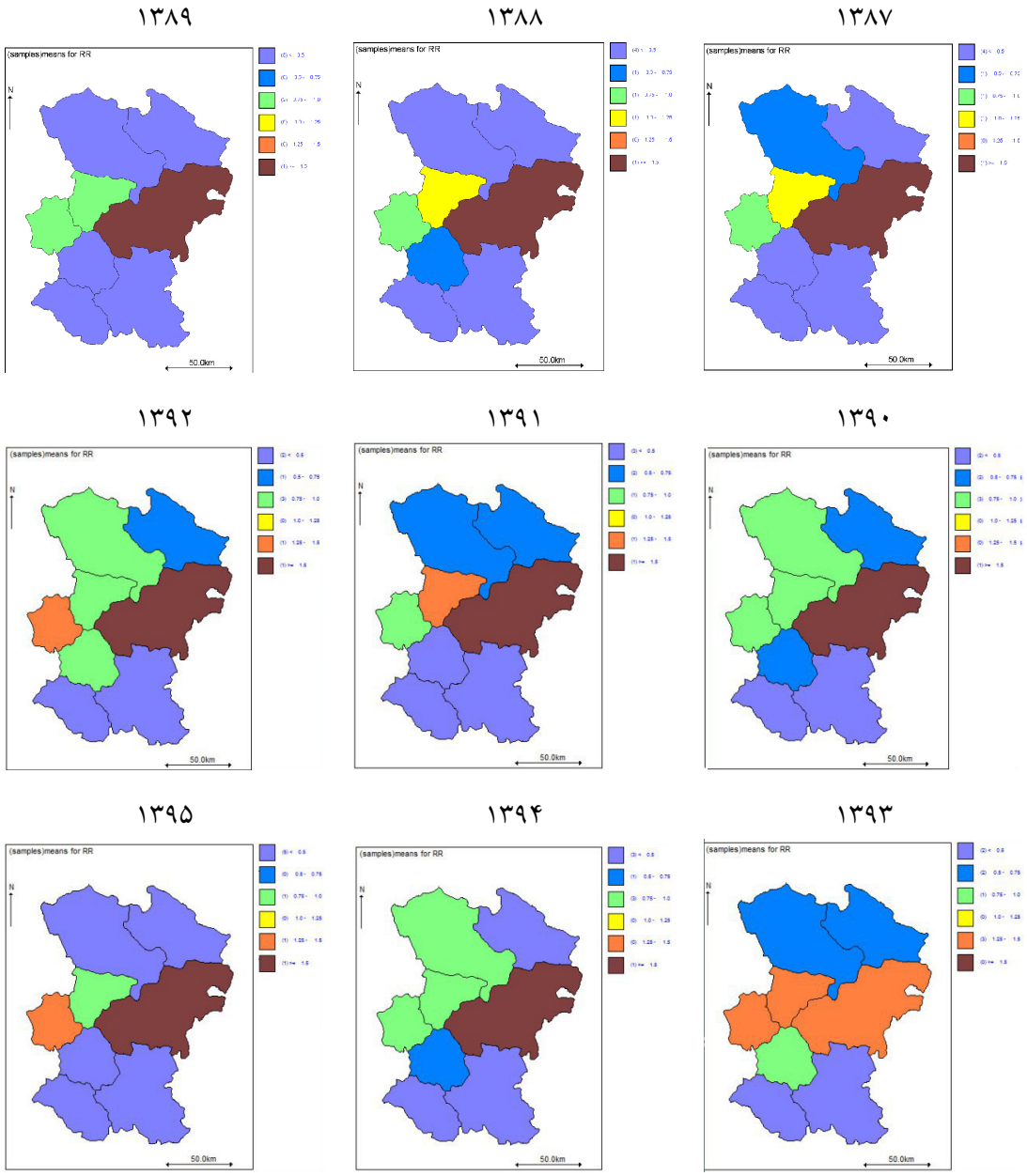
شکل ۱: نقشه پراکندگی نرخ خطر ابتلا به بیماری افسردگی در استان همدان

۴ بحث و نتیجه‌گیری

در این پژوهش برای مدل‌بندی خطر نسبی از یک مدل رگرسیون پواسن با اثرات تصادفی استفاده شده است و به منظور برآورد پارامترهای به کار رفته در مدل رویکرد بیز تجربی مورد استفاده قرار گرفته است. نقشه‌بندی بیماری دارای سابقه‌ای طولانی در اپیدمیولوژی به عنوان بخشی از نظریه کلاسیک وقوع بیماری به علت دلایل سه گانه زمان، مکان و شخص است. تجزیه و تحلیل توزیع جغرافیایی بروز بیماری نقش مهمی در مطالعات اپیدمیولوژیکی و بهداشت عمومی دارد. هدف از نقشه‌بندی بیماری ایجاد نقشه توزیع جغرافیایی بروز بیماری است (بت و همکاران، ۲۰۱۳).

از جمله کاربردهای نقشه بندی می توان به برآورد بار بیماری، مقایسه نرخ بیماری برای مقاصد علت شناسی در مناطق مختلف جغرافیایی، استفاده به عنوان ابزاری تخصیص منابع و یا خوشه‌بندی مناطق بر اساس خطر ابتلا به بیماری و سپس استفاده از تجزیه وتحلیل خوشه‌ای اشاره کرد (ماری و همکاران، ۱۹۹۶). تا کنون در مطالعات بسیاری به نقشه‌بندی انواع بیماری‌ها پرداخته شده است. همپتون و همکاران (۲۰۱۱) به نقشه‌بندی بیماری ایدز در کارولینای شمالی پرداختند و پردهان و همکاران (۲۰۱۲) به نقشه‌بندی پراکندگی بیماری‌های قابل انتقال دستگاه گوارش در منطقه شمال غربی کانادا پرداختند. در پژوهش حاضر به نقشه‌بندی

نرخ خطر ابتلا به بیماری افسردگی در استان همدان پرداخته شده است. با توجه به نتایج به دست آمده از مدل‌بندی خطر نسبی و ترسیم نقشه بیماری برای بیماری افسردگی ملاحظه شد که این شاخص در برخی از شهرستان‌های استان همدان بالا بوده است و این مسأله بیانگر اهمیت توجه هر چه بیشتر به این بیماری می‌باشد. مسلما با برنامه‌ریزی برای پیشگیری و کنترل این بیماری می‌توان گام بزرگی در جهت ارتقاء سطح سلامت و بهبود کیفیت زندگی در جامعه برداشت.



شکل ۲: نقشه پراکندگی نرخ خطر ابتلا به بیماری افسردگی در استان همدان در سال‌های مختلف

مراجع

Agresti, A.(2007). *An introduction to categorical data analysis*, 2nd edn. Hoboken. NJ: Wiley-Interscience.

Anderson, A. L., and Allen, T. (2011). Mapping Historic Hookworm Disease prevalence in the southern Us, comparing percent prevalence with percent soil Drainage Type Using GIs. *Infectious Diseases: Research and Treatment*,1-4.

Besag, J., York, J., and Mollié, A.(1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*,**43**,1-20.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., and Moyes, C. L. (2013). The global distribution and burden of dengue. *Nature*,**496**,(7446):504.

Chu, I-H, Buckworth, J., Kirby, T. E., and Emery, C. F.(2009). Effect of exercise intensity on depressive symptoms in women. *Mental Health and Physical Activity*, **2**, 37-43.

Cardona, D., Segura, A., Segura, Á., and Garzón, M. O.(2015). Contextual effects associated with depression risk variability in the elderly, Antioquia, Colombia, 2012. *Biomédica*, **35** ,73 - 80.

Cook, V., Shah, L., Gardy, J., and Bourgeois, A.(2012). Recommendations on modern contact investigation methods for enhancing tuberculosis control. *The International Journal of Tuberculosis and Lung Disease*, **16**,297-305.

Sato, R., Kawanishi, C., Yamada, T., Hasegawa, H., Ikeda, H., and Kato, D.(2006). Knowledge and attitude towards suicide among medical students in Japan: Preliminary study. *Psychiatry and clinical neurosciences*, **60**,558-562.

Ghias, M., Seidaiy, E., Taghdicy, A., Rozbahani, R., Poorsafa, P., and Barghi, H. (2011). The Impact of Weather on Some Psychological and Behavioral Disorders among 4-7-year-old Children in Rural Area of Isfahan Province According to Medical Geography. *Journal of Isfahan Medical School*,**28**,1-7 .

Hampton, K. H. , Serre, M. L. , Gesink, D. C., Pilcher, C. D., Miller, W. C. (2011). Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping. *International journal of health geographics*, **10**,1-17.

Harada, N., Takeshita, J., Ahmed, I., Chen, R., Petrovitch, H., Ross, G. W., (2012). Does cultural assimilationinfluence prevalence and presentation of depressive symptoms in older Japanese American men? The Honolulu-Asia aging study. *The American Journal of Geriatric Psychiatry*, **20**, 337-345.

May, M., McCarron, P., Stansfeld, S., Ben-Shlomo, Y., Gallacher, J., Yarnell, J.(2002). Does psychological distress predict the risk of ischemic stroke and transient ischemic attack? *Stroke*,**33**,7-12.

Merskey, H., Brown, J., Brown, A., Malhotra, L., Morrison, D., and Ripley C.(1985). Psychological normality and abnormality in persistent headache patients. *Pain*, **23**, 35-47.

Murray, C. J., Lopez, A. D., and Organization, W. H.(1996). The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: summary.

Nadaoka, T., Kanda, H., Oiji, A., Morioka, Y,, Kashiwakura, M., Totsuka, S.(1997). Headache and stress in a group of nurses and government administrators in Japan. *Headache: The Journal of Head and Face Pain*, **37**, 386-391.

Organization, W. H. (1990). The introduction of a mental health component into primary health care.

Pardhan-Ali, A., Berke, O., Wilson, J., Edge, V. L., Furgal, C., Reid-Smith R., (2012). A spatial and temporal analysis of notifiable gastrointestinal illness in the Northwest Territories, Canada, 1991-2008. *International journal of health geographics*, **11**,1-10.

Rasul, F., Stansfeld, S. A., Hart, C., Gillis, C., Smith, G. D.(2002). Common mental disorder and physical illness in the Renfrew and Paisley (MIDSPAN) study. *Journal of psychosomatic research*, **53**,1163-1170.

Fakhari, A., Ranjbar, F., Dadashzadeh, H., and Moghaddas, F. (2007). An epidemiological survey of mental disorders among adults in the north, west area of Tabriz, Iran. *Pakistan Journal of Medical Sciences*, **23**, 54-58.

Teychenne, M., and York, R.(2013). Physical activity, sedentary behavior, and postnatal depressive symptoms: a review. *American journal of preventive medicin*, **45**, 217-227.

WHO. Depression and Other Common Mental Disorders http://www.who.int/mental_health/management/depression/prevalence_global_health_ estimates/en/2017

دومین سمینار آمار فضایی و کاربردهای آن
۳ - ۴ آبان ۱۳۹۶

کریگیدن بتادوجمله‌ای برای مدل‌بندی نسبت‌های فضایی و تحلیل میزان طلاق در استان‌های ایران

علی فتح‌خانی*، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده: در بسیاری از علوم کاربردی متغیرهای پاسخی مانند نرخ‌ها در بازه (۰،۱) قرار می‌گیرند. مدل رگرسیون استاندارد به دلیل برقرار نبودن فرض‌های اساسی، برای مدل‌بندی این‌گونه داده‌ها مناسب نیست. در سال‌های اخیر مدل رگرسیون بتا برای مدل‌بندی این گونه مشاهدات معرفی شده است. توزیع بتا یک توزیع انعطاف‌پذیر در بازه (۰،۱) است که فرم‌های متقارن و چوله را در بر می‌گیرد. مدل رگرسیون بتا براساس این فرض است که متغیر پاسخ از توزیع بتا پیروی می‌کند. در مدل رگرسیون بتا فرم بازپارامتریده این توزیع که شامل دو پارامتر میانگین و دقت است، برای مدل‌بندی متغیر پاسخ به‌کار گرفته می‌شود. در این مقاله مدلی برای پیش‌گویی نسبت‌های همبسته فضایی تحت عنوان کریگیدن بتا دوجمله‌ای ارائه می‌شود که پیش‌گویی‌های با دقت بالاتری ارائه می‌کند. در نهایت نحوه کاربست دو مدل کریگیدن بتا دوجمله‌ای و رگرسیون بتا در برازش به داده‌های نرخ طلاق در استان‌های کشور نشان داده می‌شود. سپس دقت پیش‌گویی‌های حاصل از مدل‌های کریگیدن بتا دوجمله‌ای و رگرسیون بتا ارزیابی و مورد مقایسه قرار می‌گیرد.

واژه‌های کلیدی: نسبت‌های فضایی، کریگیدن بتا دوجمله‌ای، مدل رگرسیون بتا، طلاق.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 91B72، 62H30.

۱ مقدمه

مشاهدات برای بسیاری از کمیت‌های تصادفی در زمینه‌های زمین‌شناسی، بوم‌شناسی، علوم محیط‌زیست و ... وابسته هستند و نوعاً این وابستگی تابعی از موقعیت مکانی آن‌ها است. در این‌گونه موارد، نیاز به روش‌ها و مدل‌های آماری جدیدی است که بتوانند ساختار همبستگی فضایی داده‌ها را در نظر گرفته و به تحلیل آن‌ها بپردازد. کریگیدن یکی از پرکاربردترین روش‌های تحلیل داده‌های فضایی برای پیش‌گویی در موقعیت‌های فاقد مشاهده است که اولین بار توسط ماترون (۱۹۶۳) معرفی شد. فرض گاوسی بودن داده‌های فضایی، موجب برتری پیش‌گوی کریگیدن در مقایسه با پیش‌گوهای دیگر می‌شود. به همین خاطر پیش‌گویی فضایی به‌طور

* **ارایه‌دهنده مقاله:** فتح‌خانی، ع.، ali.fathkhani.dugs@gmail.com

معمول براساس فرض گاوسی بودن میدان تصادفی انجام می‌شود، اما در عمل با موارد زیادی مواجه می‌شویم که این فرض غیر واقع‌گرایانه است یا بررسی صحت آن مقدور نیست. لازم به ذکر است که نسبت‌های فضایی نیز اغلب ناگاوسی هستند.

پائولینو (۲۰۰۱) برای اولین بار پیشنهاد کرد برای مدل‌بندی متغیرهای پاسخ پیوسته در بازه (۱و۰)، توزیع بتا به متغیر پاسخ اختصاص داده شود. او نشان داد که این فرض با لحاظ دشواری‌های محاسباتی، برآوردهای کاراتری برای پارامترها به‌دست می‌دهد. از آنجا که چگالی بتا بسته به مقادیر متفاوت پارامترهایش، شکل‌های کاملاً متفاوتی را به خود اختصاص می‌دهد از انعطاف‌پذیری بالایی در مدل‌بندی رگرسیونی برخوردار است. این انعطاف‌پذیری مشوق استفاده از این توزیع در محدوده گسترده‌ای از کاربردها است. به منظور به‌دست آوردن ساختار رگرسیونی برای میانگین پاسخ همراه با پارامتر دقت، فراری و کریباری (۲۰۰۴) توزیع بتای بازپارامتریده[1] را تحت عنوان رگرسیون بتا[2] ارائه نمودند. سپیدا و گامرمن (۲۰۰۵)، مدل رگرسیون بتا را با متغیر در نظر گرفتن پارامتر دقت بررسی کردند. مدل رگرسیون بتا فضایی نیز توسط کلهری و محمدزاده (۲۰۱۷) برای تحلیل نسبت‌های وابسته فضایی ارائه شده است.

مونستیز و همکاران (۲۰۰۶) یک مدل خاص کریگیدن فضایی از داده‌های شمارشی را ارائه کردند، که در آن به طور ضمنی فرض شده است که فرآیند فضایی مشاهده شده دارای یک توزیع پواسون با متغیر تصادفی گاوسی پنهان است. با استنتاج رابطه بین تغییرنگار گاوسی پنهان و تغییرنگار پواسون می‌توان آن مشکلات را حذف کرده و با روش‌های تناسب کریگیدن یک مدل بیزی قابل درک را ارائه کرد. مدل هم‌کریگیدن دوجمله‌ای توسط الیور و همکاران (۱۹۹۸) برای پیش‌گویی نسبت فضایی براساس نسبت‌های نمونه مشاهده شده از یک بیماری نادر معرفی شده‌اند، با این حال در آن فرض توزیع بتا برای مشاهدات در نظر گرفته نشده است. اما مدل کریگیدن بتا دوجمله‌ای، مشابه مدل هم‌کریگیدن دوجمله‌ای مبتنی بر توزیع بتا است که توسط شواب و مارکس (۲۰۱۵) برای تحلیل نسبت‌های فضایی معرفی شده و در عین حال ایرادات مدل هم‌کریگیدن دوجمله‌ای را ندارد.

در بخش ۲ مدل رگرسیون بتا فضایی معرفی می شود. در بخش ۳ مدلی برای پیش‌گویی نسبت‌های همبسته فضایی تحت عنوان کریگیدن بتا دوجمله‌ای ارائه می‌شود. در بخش ۴ نحوه کاربست دو مدل کریگیدن بتا دوجمله‌ای و رگرسیون بتا در برازش به داده‌های نرخ طلاق در استان‌های کشور نشان داده می‌شود. سپس دقت پیش‌گویی‌های حاصل از مدل‌های کریگیدن بتا دوجمله‌ای و رگرسیون بتا ارزیابی و مورد مقایسه قرار می‌گیرد. در انتها بحث و نتیجه‌گیری ارائه خواهد شد.

۲ رگرسیون بتا فضایی

معمولا مقادیر نسبت‌ها و نرخ‌ها در بازه (۱و۰) قرار دارند و در بسیاری از زمینه‌های کاربردی به عنوان تحقق‌های متغیر پاسخ در نظر گرفته می‌شوند. در سال‌های اخیر مدل رگرسیون بتا برای مدل‌بندی این‌گونه مشاهدات معرفی شده است. از آنجا که توزیع بتا با تغییر مقادیر پارامترهایش به فرم‌های متفاوتی ظاهر می‌شود که در بردارنده فرم‌های نامتقارن نیز هست، مدل رگرسیون بتا در مقایسه با مدل‌های رگرسیون خطی از انعطاف‌پذیری بیشتری برخوردار است. . با توجه به این‌که در مدل‌های رگرسیونی، الگوی رفتار میانگین متغیر پاسخ مشروط بر متغیرهای تبیینی مورد بررسی قرار می‌گیرد، فراری و کریباری (۲۰۰۴) توزیع بتای بازپارامتریده را برای مطالعه نسبت‌ها پیشنهاد کردند. آن‌ها پارامترهای توزیع بتا را به گونه‌ای بازنویسی کردند که مدل رگرسیونی براساس میانگین

[1] Parameterization
[2] Beta Regression

متغیر پاسخ معین شود. بنابراین توزیع $Beta\left(\alpha,\beta\right)$ را با فرض $\mu=\frac{\alpha}{\alpha+\beta}$ و $\phi=\alpha+\beta$ به صورت

$$\pi\left(y,\mu,\phi\right)=\frac{\Gamma\left(\phi\right)}{\Gamma\left(\mu\phi\right)\Gamma\left(\left(1-\mu\right)\phi\right)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}\quad 0<y<1 \tag{۱.۲}$$

در نظر گرفتند، که در آن $0<\mu<1$ و $\phi>0$. در نتیجه $E\left(Y\right)=\mu$ و $\mathrm{Var}\left(Y\right)=\frac{\mu(1-\mu)}{1+\phi}$، که در آن ϕ پارامتر دقت نامیده می‌شود.

با فرض آن‌که متغیرهای تصادفی مستقل $y_1,\ldots,y_n$ از توزیع بتای $Beta\left(\mu_i\phi,\left(1-\mu_i\right)\phi\right)$، $i=1,\ldots,n$ پیروی کنند، که در آن ϕ مقداری نامعلوم و ثابت است، آنگاه مدل رگرسیون بتا به صورت $g\left(\mu_i\right)=\sum_{j=0}^{k}x_{ij}\beta_j$ معرفی می‌شود، که در آن $x_{i0},\ldots,x_{ik}$ متغیرهای تبیینی و $\beta=\left(\beta_0,\ldots,\beta_k\right)^\top$ بردار ضرایب رگرسیونی هستند. اما در حالاتی که پارامتر دقت ϕ ثابت نباشد، شرط همگنی واریانس برقرار نیست و لازم است که علاوه بر میانگین، با استفاده از توابع پیوند مناسب مدلی برای پارامتر دقت نیز در نظر گرفته شود. سپیدا و گامرمن (۲۰۰۵)، مدل رگرسیون بتا را با مدل‌بندی همزمان میانگین و پارامتر دقت توسعه دادند. در این مطالعات لوجیت میانگین و لگاریتم پارامتر دقت با در نظر گرفتن الگویی خطی به متغیرهای تبیینی ربط داده شده‌اند. اگر $y_1,\ldots,y_n$ متغیرهای تصادفی مستقل از توزیع بتای $Beta\left(\mu_i\phi_i,\left(1-\mu_i\right)\phi_i\right)$، $i=1,\ldots,n$ باشند، آنگاه مدل رگرسیون بتا را می‌توان به صورت $h\left(\phi_i\right)=\sum_{l=0}^{m}z_{il}\delta_l$ در نظر گرفت که در آن $z_{i0},\ldots,z_{il}$ متغیرهای تبیینی و $\delta=\left(\delta_0,\ldots,\delta_m\right)^\top$ بردار ضرایب رگرسیونی هستند. این مدل می‌تواند تابعی از متغیرهای تبیینی باشد که لزوما با متغیرهای تبیینی مدل میانگین یکسان نیستند. کریباری و زیلیس (۲۰۱۰)، بسته betareg در نرم‌افزار R را برای محاسبات رگرسیون بتا، با فرض ثابت یا متغیر بودن پارامتر دقت، ارائه نمودند. سپیدا و همکاران (۲۰۱۲) مدل‌بندی همزمان پارامترهای میانگین و دقت را برای داده‌های همبسته فضایی مورد مطالعه قرار دادند. آن‌ها اثر فضایی را در مدل رگرسیون بتا با استفاده از حاصل ضرب ماتریس وزن در متغیر پاسخ به عنوان یک متغیر تبیینی وارد کردند. مدل آن‌ها به صورت

$$\begin{aligned}\mathrm{logit}\left(\mu\right) &= X\beta+\rho WY\\ \log\left(\phi\right) &= Z\delta+\lambda WY\end{aligned}$$

ارائه شد، که در آن ρ و λ ضرایب همبستگی و W ماتریس وزن فضایی و Y بردار متغیر پاسخ است. کلهری و محمدزاده (۱۳۹۵) نیز برآورد بیزی پارامترهای رگرسیونی مدل معرفی شده توسط سپیدا و همکاران (۲۰۱۲) را طی مطالعه‌ای شبیه‌سازی به دست آورده‌اند. همچنین کلهری و محمدزاده (۲۰۱۷) مدل رگرسیون بتا فضایی را با در نظر گرفتن اثرات تصادفی مورد مطالعه قرار داده‌اند.

۳ کریگیدن بتا دوجمله‌ای

فرض کنید در موقعیت s_i درون دامنه فضایی D، میدان تصادفی $Z_i=Z(s_i)$ با توزیع شرطی $Z_i|Y_i\sim Binomail\left(n,Y_i\right)$ مشاهده شده باشد، که در آن Y_i احتمال موفقیت در مکان s_i به طور فضایی همبسته است. با فرض معلوم بودن احتمال پنهان Y_i، متغیرهای دوجمله‌ای حاشیه‌ای فضایی Z_i ناهمبسته خواهند بود. علاوه بر این فرض کنید $Y_i\sim Beta\left(\alpha,\beta\right)$، با ساختار فضایی همبسته Σ_Y باشد. در این صورت $E\left(Y_i\right)=\frac{\alpha}{\alpha+\beta}=\pi$ و $\mathrm{Var}\left(Y_i\right)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}=\sigma_Y^2$. چون میانگین توزیع بتا، π ثابت و تابع کوواریانس تنها به فاصله بین دو مشاهده وابسته است، بنابراین هر دو تابع کوواریانس $C_Y\left(|s_i-s_j|\right)$ و تغییرنگار

$\gamma_Y(|s_i - s_j|)$ که تنها به $|s_i - s_j|$، یعنی فاصله بین موقعیت‌های فضایی s_i و s_j بستگی دارند، همسانگرد هستند. از آنجا که $E(Z_i|Y_i) = n_i Y_i$ و $\text{Var}(Z_i|Y_i) = nY_i(۱ - Y_i)$ داریم:

$$
\begin{aligned}
&E(Z_i) = E[E(Z_i|Y_i)] = n_i\pi,\\
&\text{Var}(Z_i) = \text{Var}[E(Z_i|Y_i)] + E[Var(Z_i|Y_i)] = n_i\sigma_Y^۲(n_i - ۱) + n_i\pi(۱ - \pi),\\
&E(Z_iZ_j|Y) = \text{Cov}Z_i, Z_j|Y) + E(Z_i|Y_i)E(Z_j|Y_j) = \delta_{ij}n_iY_i(۱ - Y_i) + n_in_jY_iY_j.
\end{aligned}
$$

که در آن δ_{ij}، دلتای کرونکر[۳] است، به طوری که برای $i = j$ مقدار یک و برای $i \neq j$ مقدار صفر را اختیار می کند.

اکنون می توان نشان داد نیم‌تغییرنگار نسبت‌های دوجمله‌ای Z_i/n_i با نیم تغییرنگار میدان تصادفی پنهان Y_i به صورت

$$\gamma_Z(s_i - s_j) = \frac{۱}{۲}\left(\frac{n_i + n_j}{n_in_j}\right)\left(\pi(۱ - \pi) - \sigma_Y^۲\right) + \gamma_Y(s_i - s_j) \tag{۱.۳}$$

رابطه دارند و می توان آن را به صورت

$$\gamma_Y^*(h) = \frac{۱}{۲N(h)}\sum_{i=۱}^{I}\sum_{j=۱}^{J}\left[\frac{n_in_j}{n_i + n_j}\left(\frac{Z_i}{n_i} - \frac{Z_j}{n_j}\right)^۲ - \left(\hat{\pi}(۱ - \hat{\pi}) - \hat{\sigma}_Y^۲\right)\right] \tag{۲.۳}$$

برآورد نمود، که در آن $N(h) = \sum_{i,j}\frac{n_in_j}{n_i+n_j}$.

با کریگیدن بتا دوجمله‌ای، می توان مقدار متغیر تصادفی پنهان $Y_{\cdot}^*$ را به صورت ترکیب خطی از نسبت‌های نمونه دوجمله‌ای مشاهده شده Z_i/n_i در موقعیت s_i با وزن λ_i به فرم

$$Y_{\cdot}^* = \sum_{i=۱}^{n}\lambda_i\frac{z_i}{n_i} \tag{۳.۳}$$

پیشگویی نمود. برای این منظور لازم است وزن‌های λ_i در هر موقعیت فضایی $s_{\cdot}$ برآورد شوند، طوری که $Y_{\cdot}^*$ ناریب و دارای کم‌ترین واریانس باشد. بنابراین لازم است دستگاه معادلات

$$\begin{cases}\frac{\lambda_i}{n_i}\left[\pi(۱ - \pi) - \sigma_Y^۲\right] + \sum_{j=۱}^{n}\lambda_jC_{ij} + \mu = C_{\cdot i} & i = ۱, \ldots, n\\ \sum_{i=۱}^{n}\lambda_i = ۱\end{cases} \tag{۴.۳}$$

حل شود، که در آن C_{ij} تابع کوواریانس در موقعیت‌های s_i و s_j و μ ضریب لاگرانژ چندگانه است. برآورد پارامترهای کریگیدن بتا دوجمله‌ای باید برای برآورد تابع کوواریانس و پارامترهای توزیع بتا، یعنی π_Y و $\sigma_Y^۲$ استفاده شوند.

۴ تحلیل داده‌های طلاق استان‌های کشور

طلاق به معنای پایان قانونی ازدواج و جدا شدن همسران از یکدیگر است که در پی آن حقوق و تکالیف متقابلی از میان می‌رود که بین زوجین هنگام ازدواج وجود داشته است. طلاق معمولاً وقتی اتفاق می‌افتد که استحکام رابطه زناشویی از بین می‌رود و میان زوجین اختلاف و درگیری به وجود می‌آید. معمولاً زنان بیش از مردان در معرض آسیب‌های روانی، اجتماعی و اقتصادی طلاق قرار می‌گیرند که می‌توان به تن دادن به ازدواج موقت، سرخوردگی، خودکشی، اعتیاد، انحرافات جنسی و ... اشاره کرد. براساس آمار

[۳] Kronecker

ثبت احوال ایران در سال ۱۳۹۴، ۱۶۳ هزار و ۷۶۵ طلاق به وقوع پیوسته است، یعنی روزانه ۴۴۹ زوج از یکدیگر جدا شده‌اند که ۲۱۴۵۱ مورد کمتر از یک سال با یکدیگر زندگی کرده‌اند. این آمار تکان‌دهنده موجب شده است که کارشناسان آمار طلاق در ایران از طلاق به عنوان زلزله خاموش تعبیر کنند که هر روز در حال تکان دادن بنیان جامعه است. ایران سومین کشور جهان از لحاظ داشتن جمعیت جوان است. میانگین سنی جمعیت کشور براساس نتایج سرشماری سال ۱۳۹۵ مرکز آمار ایران حدود ۳۱/۱ سال است که نشان از جوانی جمعیت کشور دارد.

در این بخش تأثیر عوامل کلیدی اقتصادی مثل بیکاری و عوامل دیگری همچون میزان شهرنشینی و میزان تحصیلات عالی بر طلاق مورد مطالعه قرار می‌گیرند. آمار ارائه شده توسط سازمان ثبت احوال ایران نشان می‌دهد که میزان طلاق در استان تهران نسبت به سایر استان‌ها بیشتر است. پایتخت به عنوان مرکز سیاسی کشور، بیشترین امکانات عمرانی، تولیدی، تسهیلات اجتماعی و رفاهی را در خود جای داده است و همچنین نقش این استان در امور آموزش عالی، بهداشت و درمان و صنعت، وجود فرصت‌های شغلی با سایر استان‌ها متفاوت است. با توجه به این‌که فرهنگ نواحی مختلف می‌تواند بر طلاق اثر داشته باشد، اثر فاصله جغرافیایی را به عنوان یک متغیر موثر بر طلاق مورد بررسی قرار می‌دهیم. داده‌های این مطالعه به استان‌های ایران و سال ۱۳۹۴ مربوط است که از پایگاه ثبت احوال و مرکز آمار ایران تهیه شده است. تعداد طلاق ثبت شده در سال ۱۳۹۴ بر مبنای مدت زمان گذشته از ازدواج که کمتر از یک سال است و نیز تعداد کل طلاق‌های ثبت شده در سال ۱۳۹۴ به تفکیک استان‌ها، متغیرهای مورد نظر این مطالعه هستند.

یکی از شاخص‌های مهم ازدواج‌های ناموفق تعداد مواردی است که کمتر از یک سال منجر به طلاق می‌شوند. در این بخش داده‌های میزان طلاق سال ۱۳۹۴ حاصل از تقسیم تعداد طلاق کمتر از یک سال هر استان به تعداد کل طلاق همان استان تحلیل و دو مدل رگرسیون بتا فضایی و کریگیدن بتا دوجمله‌ای به آنها برازش داده می‌شود. سپس براساس مدل‌های برازانده شده میزان طلاق در استان‌های کشور پیش‌گویی و ارزیابی می‌شود.

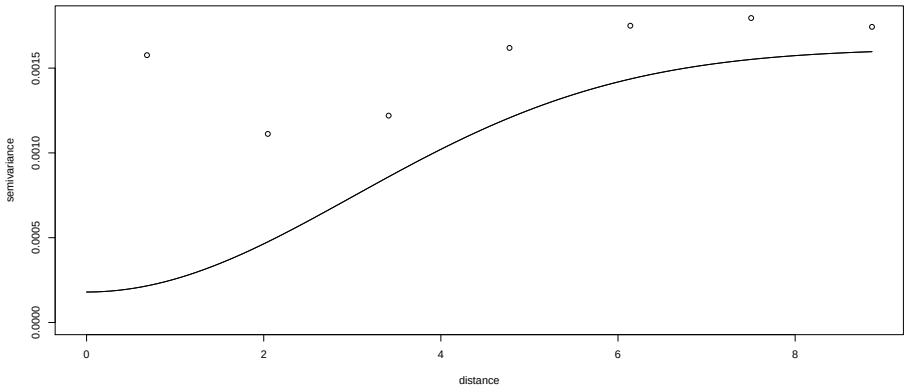
برای تحلیل داده‌ها با کریگیدن بتا دوجمله‌ای، ابتدا لازم است ساختار همبستگی فضایی داده‌ها تعیین شود. لذا به منظور بررسی همسانگردی و انتخاب مدل تغییرنگار نظری مناسب، تغییرنگار تجربی میزان طلاق در چهار جهت جغرافیایی ۰، ۴۵، ۹۰ و ۱۳۵ درجه بررسی شد و همسانگردی مورد تأیید قرار گرفت. برای برازش یک مدل نیم‌تغییرنگار معتبر به مشاهدات میزان طلاق، مدل‌های مختلف نیم‌تغییرنگار به داده‌ها برازش داده شده و از بین مدل‌های مختلف، مدل گاوسی

$$\gamma(h) = c_0 + c\left(1 - e^{\frac{-|h|^2}{a^2}}\right), \qquad h \in R^d, \qquad d \geq 1,$$

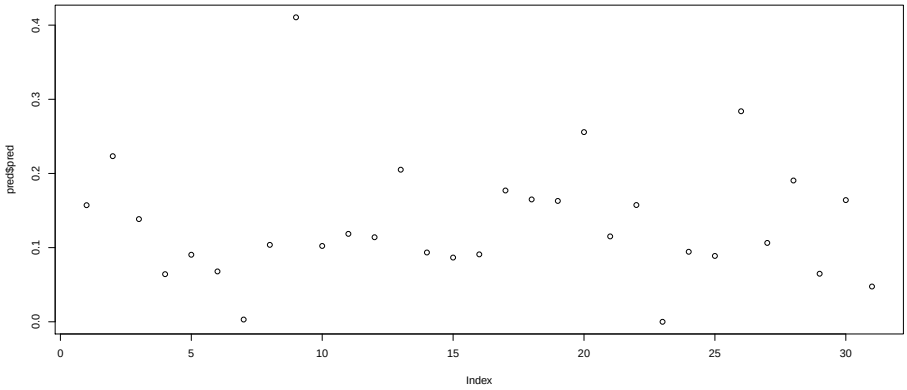
که در مقایسه با سایر مدل‌ها دارای کمترین مقدار مجموع توان دوم خطا است، انتخاب و پارامترهای آن به صورت $c = 0/0012$، $a = 4/26$ و $c_0 = 0/000018$ برآورد گردید که نمودار آن در شکل ۱ رسم شده است.

پیش‌گویی فضایی برای داده‌های میزان طلاق استان‌ها با استفاده از کریگیدن بتا دوجمله‌ای محاسبه و در شکل ۲ ارائه شده است. ارزیابی آن‌ها نیز بر اساس میانگین توان‌های دوم خطا انجام می‌شود. همانطور که در شکل ۳ ملاحظه می‌شود میانگین توان دوم خطاهای پیش‌گویی‌ها بیانگر خطای نزدیک به صفر پیش‌گویی‌ها است.

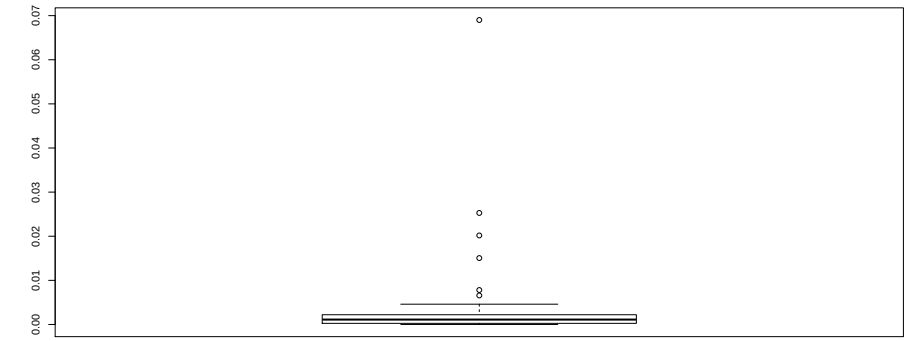
برای تحلیل میزان طلاق با استفاده از مدل رگرسیون بتای فضایی، عواملی مانند میزان شهرنشینی، بیکاری و میزان تحصیلات عالی به عنوان متغیرهای تبیینی برای میزان اثرگذاری بر میزان طلاق در نظر گرفته شده‌اند. علاوه بر عوامل ذکر شده، فاصله اقلیدسی بین مراکز استان‌ها و تهران نیز به عنوان متغیر تبیین کننده اثر فضایی در نظر گرفته می‌شود. در واقع اثر فضایی بیانگر تاثیر موقعیت



شکل ۱: نمودار برآوردگر نیم‌تغییرنگار برازش شده به داده‌های میزان طلاق



شکل ۲: نمودار مقادیر پیش‌گویی شده



شکل ۳: نمودار جعبه‌ای توان دوم خطای پیش‌گویی‌ها

جغرافیایی بر میزان طلاق است. با فرض همگن بودن واریانس پاسخ، میزان طلاق به صورت $\text{logit}(\mu) = X\beta$ و $\log(\phi) = Z\delta$ مدل‌بندی می‌شود. پس از برازش مدل‌هایی با متغیرهای تبیینی مختلف، مدل نهایی در جدول ۱ ارائه شده است. ضرایب β_0، β_1 و β_2 به ترتیب معرف عرض از مبدأ، ضریب نرخ بیکاری و ضریب میزان شهرنشینی مدل میانگین و ضرایب δ_0 و δ_1 نیز بیانگر عرض از مبدأ و ضریب نرخ بیکاری مدل پارامتر دقت می‌باشند.

جدول ۱: برآورد پارامترهای مدل‌بندی همزمان میانگین و دقت

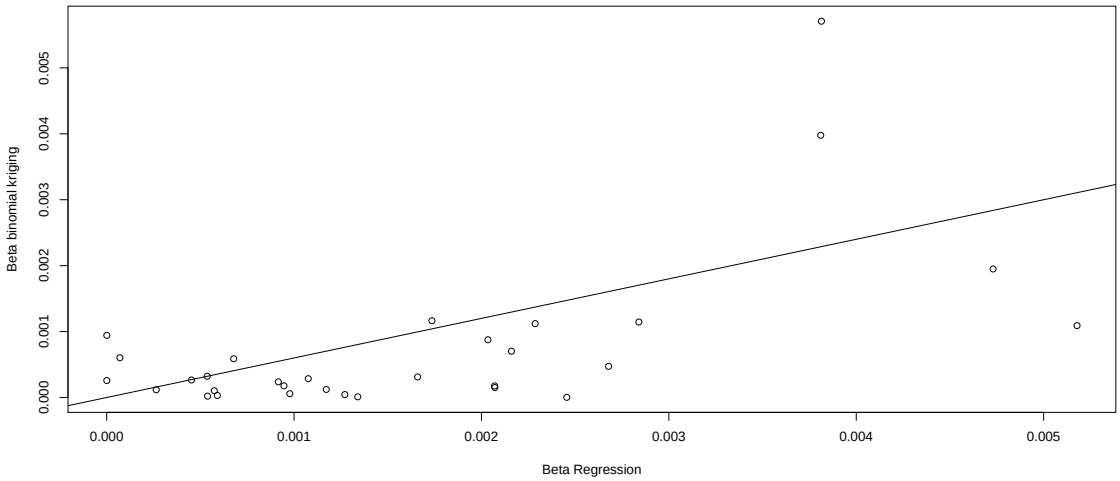
مدل	ضرایب	برآورد	انحراف استاندارد
میانگین	β_0	-۲/۱۶۵۸۳	۰/۴۶۳۰۵
	β_1	۰/۰۱۵۹۸	۰/۰۲۳۳۱
	β_2	۰/۲۲۲۰۱	۰/۴۹۵۱۶
دقت	δ_0	۴/۳۶۰۹۵۳	۱/۱۵۰۷۲۷
	δ_1	-۰/۰۰۵۸۸۴	۰/۰۹۷۸۰۸

همان‌طور که ملاحظه می‌شود نرخ بیکاری و میزان شهرنشینی بر میزان طلاق اثر مثبت و عوامل فاصله از تهران و میزان تحصیلات عالی اثر منفی داشته‌اند. از طرفی می‌بینیم که با افزایش نرخ بیکاری پارامتر دقت کاهش می‌یابد، به عبارت دیگر واریانس میزان طلاق در کشور ثابت نیست و با افزایش نرخ بیکاری افزایش می‌یابد.

برای مقایسه عملکرد پیشگویی بر اساس دو مدل کریگیدن بتا دوجمله‌ای و رگرسیون بتا فضایی میانگین توان دوم پیش‌گویی خطاهای به دست آمده از این دو مدل در مقابل هم در شکل ۴ رسم شده است. در این شکل محور عمودی نمایانگر میانگین توان دوم پیش‌گویی خطاهای به دست آمده از مدل کریگیدن بتا دوجمله‌ای و محور افقی نشان دهنده میانگین توان دوم پیش‌گویی خطاهای مدل رگرسیون بتا فضایی هستند. همان‌طور که ملاحظه می‌شود خطاهای پیش‌گویی مدل کریگیدن بتا دوجمله‌ای عموما از خطاهای مدل رگرسیون بتا فضایی کمتر هستند. از طرفی، مقدار میانگین توان دوم خطای پیش‌گوی دو مدل کریگیدن بتا دوجمله‌ای و رگرسیون بتا به ترتیب برابر ۰/۰۰۰۷۴ و ۰/۰۰۱۶۴ است که همین مطلب نشان از دقت بیش‌تر مدل کریگیدن بتا دوجمله‌ای نسبت به مدل رگرسیون بتا دارد.

بحث و نتیجه‌گیری

در این مطالعه داده‌های طلاق را با دو مدل کریگیدن بتا دوجمله‌ای و مدل رگرسیون بتا فضایی مورد تحلیل و بررسی قرار دادیم. نتایج به دست آمده از هر مدل را به طور وضوح بیان کردیم و دیدیم که مدل کریگیدن بتا دوجمله‌ای، خطاهای پیش‌گویی کمتری نسبت به مدل رگرسیون بتا فضایی دارد. به علاوه نتایج برازش رگرسیون بتا فضایی به داده‌های میزان طلاق بیانگر آن است که افزایش بیکاری مطلوبیت ماندن در زندگی را کاهش می‌دهد. برخلاف نرخ بیکاری و میزان شهرنشینی که باعث افزایش میزان طلاق کمتر از یک سال می‌شوند، میزان فاصله از تهران باعث کاهش مقدار طلاق و به خصوص طلاق کمتر از یک سال می‌شود، یعنی هرچه فاصله از تهران بیش‌تر باشد، میزان طلاق کم‌تر می‌شود. همچنین در مورد میزان تحصیلات عالی نیز می‌توان گفت که افزایش آن باعث کاهش طلاق می‌شود.



شکل ۴: مقایسه میانگین توان دوم پیش‌گویی خطاهای دو مدل

سپاس‌گزاری

نویسندگان از معاونت اجتماعی و پیشگیری از وقوع جرم قوه قضاییه به خاطر پشتیبانی از موضوع این تحقیق کمال تشکر را دارند.

مراجع

کلهری، ل. و محمدزاده، م. (۱۳۹۵)، مدل رگرسیون بتا فضایی و مدل‌بندی میزان طلاق استان‌ها، *سیزدهمین کنفرانس آمار ایران، کرمان، ایران*، ۶۸۶-۶۸۱.

Cepeda, E. D., and Gamerman, D. (2005), Bayesian Methodology for Modeling Parameters in the Two Parameter Exponential Family, *Revista Estadística*, **57**, 168-169.

Cepeda-Cuervo, E., Urdinola, B. P., and Rodr¡guez, D., (2012), Double Generalized Spatial Econometric Models, *Communications in Statistics-Simulation and Computation*, **41**, 671-685.

Cribari, F., and Zeileis, A. (2010), Beta Regression in R, *Journal of Statistical Software*, **34**, 1-24.

Ferrari, S. and Cribari, F. (2004), Beta Regression for Modelling Rates and Proportions, *Journal of Applied Statistics*, **31**, 799-815.

Kalhori, L., and Mohammadzadeh, M., (2017), Spatial Beta Regression Model with Random Effect, *Journal of Statistical Research of Iran*, **13**, 215-230.

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

Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J. P., and Guinet, C. (2006). Geostatistical Modelling of Spatial Distribution of Balaenoptera Physalus in the Northwestern Mediterranean Sea from Sparse Count Data and Heterogeneous Observation Efforts. *Ecological Modelling*, **193**, 615-628.

Oliver, M. A., Webster, R., Lajaunie, C., Muir, K. R., Parkes, S. E., Cameron, A. H., and Mann, J. R. (1998), Binomial Cokriging for Estimating and Mapping the Risk of Childhood Cancer. *Mathematical Medicine and Biology*, **15**, 279-297.

Paolino, P. (2001), Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables.*Political Analysis*, **9**, 325-346.

Schwab, A. D. and Marx, D. B. (2015), Beta-Binomial Kriging: An Improved Model for Spatial Rates. *Procedia Environmental Sciences*, **27**, 30-37.

دومین سمینار آمار فضایی و کاربردهای آن
۳ - ۴ آبان ۱۳۹۶

برآورد پارامترهای تغییرنگار فضایی ـ زمانی : یک روش دامنه فرکانس

فاطمه قرنجیک*، مهناز خلفی، مجید عظیم محسنی

گروه آمار، دانشگاه گلستان

چکیده: در روش‌های ساخت و برآورد پارامترهای توابع معتبر کواریانس فضایی ـ زمانی در دهه‌های اخیر پیشرفت‌های مهمی به‌وجود آمده است. این مقاله به یک روش جدید برای پارامترهای تابع کواریانس فضایی ـ زمانی در دامنه فرکانس پرداخته است. همچنین این روش بر روی داده‌های شاخص خشکسالی ثبت شده از ایستگاه‌های استان گلستان به‌کار گرفته خواهد شد.

واژه‌های کلیدی: کواریانس فضایی ـ زمانی، دوره‌نگار متقابل، تغییرنگار فرکانس

کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 62M15، 62M10، 62P12.

۱ مقدمه

داده‌های حاصل از فرایند فضایی ـ زمانی که در بسیاری از زمینه‌های علمی از قبیل اپیدمیولوژی، علوم زیست محیطی، کشاورزی و زمین شناسی و غیره کاربرد دارد به داده‌های فضایی ـ زمانی اطلاق شده که اطلاعات آن‌ها وابستگی به فضا و زمان دارد که با تغییر این دو متغیر اطلاعات آنها نیز تغییر می‌کند. داده‌های فضایی ـ زمانی در داشتن نقشه‌ای جامع و مدل‌سازی محیط کمک بسیاری به مهندسان و کارشناسان می‌کند. از موارد دیگر به‌کارگیری فرایند فضایی ـ زمانی می‌توان به پردازش صوت، هواشناسی، زمین شناسی، مهندسی پزشکی، پزشکی و ترافیک شهری اشاره کرد. که در این موارد گفته شده داده‌ها وابستگی شدیدی به فضا و زمان دارند و برای بیان این همبستگی فضایی ـ زمانی تعریف توابعی حائز اهمیت است.

تغییرات فضایی ـ زمانی را با تغییرنگار و همبستگی فضایی ـ زمانی را می‌توان از طریق تابع کواریانس بررسی کرد. تقریبا در سه دهه گذشته موضوع داده‌های فضایی و داده‌های فضایی ـ زمانی تا حد زیادی گسترش یافته است. یکی از ابزار اصلی در پیش بینی فضایی، تابع کواریانس است که با توجه به این تابع می‌توان پیش‌بینی‌های بهینه از فضاهای نمونه برداری نشده بدست آورد. در عمل تابع کواریانس ناشناخته است و نیاز به برآورد براساس داده‌های نمونه دارد، بنابراین دانش زمین آمار به بررسی توابع کواریانس

* **ارایه‌دهنده مقاله:** فاطمه قرنجیک، f.gharanjik@stu.gu.ac.ir

و تغییرنگار و برآورد آن می‌پردازد. از طرفی گنجاندن بعد زمان این توابع و برآوردکردن آن‌ها را پیچیده می‌کند به‌طوری که منابع در برآورد تابع کواریانس و تغییرنگار فضایی ـ زمانی فرایندهای تصادفی فضایی ـ زمانی بسیار گسترده نمی‌باشد. چندین متون قبلی در زمین آمار برآورد توابع تغییرنگار و کواریانس فضایی را در نظر گرفته است از جمله، کریس (۱۹۹۳)، چیلز و دلفینر (۱۹۹۹)، کریس و هوانگ (۱۹۹۹)، سریواستاوا (۱۹۸۹) و استاین (۱۹۹۹). بررسی ویژگی های برآورد تابع کواریانس فضایی ـ زمانی توسط لی و همکاران (۲۰۰۷)، استاین (۲۰۰۵) صورت گرفته است.

به‌جهت کاهش پیچیدگی در دانش زمین آمار فرض‌های متعددی برای توابع تغییرنگار و کواریانس در نظر گرفته می‌شود یک فرض متداول بر روی توابع تغییرنگار و یا کواریانس همسانگردی آن می‌باشد بدین منظور که این توابع به جهت وابسته نیستند. فرض دیگر برای توابع کواریانس فضایی ـ زمانی تفکیک پذیری آن از لحاظ فضا و زمان می‌باشد (شرمن (۲۰۱۰)).

تحولات مهم در مدل‌های کواریانس فضایی ـ زمانی توسط محققینی مانند برآون و همکاران (۲۰۰۰)، کریس و هوانگ (۱۹۹۹)، نایتینگ (۲۰۰۲) و ما (۲۰۰۳) صورت گرفته است. نایتینگ (۲۰۰۲) و نایتینگ و همکاران (۲۰۰۷) یک کلاس تفکیک پذیر از توابع کواریانس براساس تعمیم ایده‌های کریس و هوانگ (۱۹۹۹) پیشنهاد دادند.

سوبا رآو و همکاران (۲۰۱۴) و سوبا رآو و تردیک (۲۰۱۵) برای برآورد پارامترهای تابع کواریانس فضایی ـ زمانی در دامنه فرکانس، تغییرنگار فرکانس (FV) را تعریف کرده و رفتار حدی این برآوردگرها در دامنه فرکانس توسط آن‌ها بررسی شده است.

در این مقاله ابتدا در بخش دوم، به تعریف فرایند فضایی ـ زمانی و ویژگی آن اشاره می‌شود، در ادامه در بخش سوم روش دستیابی به برآوردگرها در دامنه فرکانس ارائه می‌شود و در بخش پایانی با استفاده از داده‌های واقعی، این روش برآورد برای پارامترهای دو مدل از توابع تغییرنگار فضایی ـ زمانی به‌کار گرفته می‌شود .

۲ فرایند فضایی ـ زمانی

فرایند فضایی ـ زمانی حقیقی، مجموعه ای از متغیرهای تصادفی به فرم

$$\{Z(\mathbf{s},t) : \mathbf{s} \in D_s \subset \mathbb{R}^d, t \in D_t \subset \mathbb{R}\}$$

است. به‌طوری که مجموعه‌ی D_s دامنه‌ی فضایی مورد نظر و مجموعه‌ی D_t هم دامنه‌ی زمانی، بنابراین اندیس $\mathbf{s}$ یک موقعیت فضایی و اندیس t بیانگر زمان می‌باشد.

برای فرایند فضایی ـ زمانی تابع تغییرنگار که نشان‌دهنده‌ی تغییرات فرایند در دامنه ی فضا و زمان می‌باشد و تابع کواریانس که بیانگر همبستگی فضایی و زمانی است به‌صورت زیر تعریف می‌شوند:

$$\begin{aligned} \text{var}(Z(s_1,t_1) - Z(s_2,t_2)) &= 2\gamma(t_1,t_2,s_1,s_2), \qquad s_1,s_2 \in D_s, t_1,t_2 \in D_t \\ C(s_1,s_2,t_1,t_2) &= \text{Cov}(Z(s_1,t_1), Z(s_2,t_2)) \\ &= E(Z(s_1,t_1)Z(s_2,t_2)) - \mu(s_1,t_1)\mu(s_2,t_2) \end{aligned}$$

بنابراین کواریانس فرایند فضایی ـ زمانی تابعی ازفضا و زمان می باشد.

همچنین تابع کواریانس برای فرایند فضایی $Z(s)$ در دو فضای مختلف s_1 و s_2 به‌صورت زیر تعریف می‌شود:

$$C(s_1,s_2) = \text{Cov}(Z(s_1), Z(s_2))$$

۱.۲ مانای مرتبه دوم فرایند فضایی

فرایند فضایی $\{Z(s)\}$ مانای ضعیف می‌باشد اگر

$$E[Z(s)] = \mu \qquad \forall s \in D_s \quad (۱)$$

$\text{Cov}\,(Z(s), Z(s')) = \text{Cov}\,(Z(s+h), Z(s'+h))$ (۲) ، برای جابجایی‌های به اندازه h

۲.۲ مانای مرتبه دوم فرایند فضایی ـ زمانی

فرایند فضایی ـ زمانی $\{Z(s,t)\}$ مانای ضعیف می‌باشد اگر:

$$E[Z(s,t)] = \mu \qquad \forall s \in D_s, \forall t \in D_t \quad (۱)$$

$Cov[(Z(s+h,t+u), Z(s,t)] = Cov[Z(h,u), Z(۰,۰)] = C(h,u)$ (۲) ، برای جابجایی‌های به اندازه h و تاخیر زمانی u

۳ برآورد پارامترهای تغییرنگار در دامنه فرکانس

فرض کنید $\{Z(\mathbf{s},t) : \mathbf{s} \in D_s \subset \mathbb{R}^d, t \in D_t \subset \mathbb{R}\}$ یک فرایند فضایی ـ زمانی باشد و سری زمانی $Y_{ij}(t)$ که تفاضل این فرایند در فضای j از فضای iـام است به‌صورت زیر تعریف شود:

$$Y_{ij}(t) = Z(\mathbf{s_i}, t) - Z(\mathbf{s_j}, t), \qquad t = ۱, ۲, \ldots, n \qquad (۱.۳)$$

مطابق معمول تبدیل فوریه و دوره‌نگار در فرکانس های فوریه $\{\omega_k = ۲\pi k/n, k = ۰, ۱, \ldots, [n/۲]\}$ برای سری زمانی $Y_{ij}(t)$ به‌ترتیب به‌صورت زیر نوشته می‌شود:

$$\begin{aligned} J_{\mathbf{s_i s_j}}(\omega_k) &= \frac{۱}{\sqrt{۲\pi n}} \sum_{t=۱}^{n} Y_{ij}(t) \exp\{-it\omega_k\} \\ I_{\mathbf{s_i,s_j}} &= \mid J_{\mathbf{s_i s_j}}(\omega_k) \mid^۲ = \frac{۱}{۲\pi} \sum_{u=-(n-۱)}^{n-۱} \widehat{C}_{y,ij}(u) e^{-iu\omega_k} \end{aligned}$$

به‌طوری که $\widehat{C}_{y,ij}(u)$، اتوکواریانس نمونه‌ای سری مانای $\{Y_{ij}(t), i \neq j\}$ که تابعی از تاخیر زمانی u است به‌صورت زیر محاسبه می‌شود:

$$\widehat{C}_{y,ij}(u) = \frac{۱}{n} \sum_{t=۱}^{n-|u|} (Y_{ij}(t+u) - \overline{Y})(Y_{ij}(t) - \overline{Y}), \qquad \mid u \mid \leq n - ۱$$

و $\overline{Y} = \frac{۱}{n} \sum_{t=۱}^{n} Y_{ij}(t)$ میانگین نمونه سری زمانی $\{Y_{ij}(t)\}$ است.

با استفاده از تعریف (۱.۳) تبدیل فوریه سری زمانی $\{Y_{ij}(t)\}$ به‌صورت زیر نتیجه می‌شود :

$$\begin{aligned} J_{\mathbf{s_i s_j}}(\omega_k) &= \frac{۱}{\sqrt{۲\pi n}} \sum_{t=۱}^{n} Y_{ij}(t) \exp\{-it\omega_k\} \\ &= \frac{۱}{\sqrt{۲\pi n}} \sum_{t=۱}^{n} [Z(\mathbf{s_i}, t) - Z(\mathbf{s_j}, t)] \exp\{-it\omega_k\} \\ &= \frac{۱}{\sqrt{۲\pi n}} \sum_{t=۱}^{n} Z(\mathbf{s_i}, t) e^{-it\omega_k} - \frac{۱}{\sqrt{۲\pi n}} \sum_{t=۱}^{n} Z(\mathbf{s_j}, t) \exp\{-it\omega_k\} \\ &= J_{\mathbf{s_i}}(\omega_k) - J_{\mathbf{s_j}}(\omega_k) \end{aligned} \tag{۲.۳}$$

به‌طوری که $J_{\mathbf{s_i}}(\omega_k)$ و $J_{\mathbf{s_j}}(\omega_k)$ به‌ترتیب تبدیلات فوریه سری زمانی $\{Z(\mathbf{s_i}, t)\}$ و $\{Z(\mathbf{s_j}, t)\}$ می‌باشند.

با کمک رابطه‌ی (۲.۳) دوره‌نگار سری زمانی $\{Y_{ij}(t)\}$ به شکل زیر حاصل می شود:

$$\begin{aligned} I_{\mathbf{s_i},\mathbf{s_j}}(\omega_k) &= \mid J_{\mathbf{s_i s_j}}(\omega_k) \mid^{۲} \\ &= \mid J_{\mathbf{s_i}}(\omega_k) - J_{\mathbf{s_j}}(\omega_k) \mid^{۲} \\ &= (J_{\mathbf{s_i}}(\omega_k) - J_{\mathbf{s_j}}(\omega_k))(\overline{J_{\mathbf{s_i}}(\omega_k) - J_{\mathbf{s_j}}(\omega_k)}) \\ &= (J_{\mathbf{s_i}}(\omega_k) - J_{\mathbf{s_j}}(\omega_k))(\overline{J_{\mathbf{s_i}}(\omega_k)} - \overline{J_{\mathbf{s_i}}(\omega_k)}) \\ &= J_{\mathbf{s_i}}(\omega_k)\overline{J_{\mathbf{s_i}}(\omega_k)} - J_{\mathbf{s_i}}(\omega_k)\overline{J_{\mathbf{s_j}}(\omega_k)} - J_{\mathbf{s_j}}(\omega_k)\overline{J_{\mathbf{s_i}}(\omega_k)} + J_{\mathbf{s_j}}(\omega_k)\overline{J_{\mathbf{s_j}}(\omega_k)} \\ &= \mid J_{\mathbf{s_i}}(\omega_k) \mid^{۲} + \mid J_{\mathbf{s_j}}(\omega_k) \mid^{۲} - ۲Re[J_{\mathbf{s_i}}(\omega_k)\overline{J_{\mathbf{s_j}}(\omega_k)}] \end{aligned}$$

از طرفی $J_{\mathbf{s_i}}(\omega_k)\overline{J_{\mathbf{s_j}}(\omega_k)}$ برابر است با :

$$\begin{aligned} J_{\mathbf{s_i}}(\omega_k)\overline{J_{\mathbf{s_j}}(\omega_k)} &= \left(\frac{۱}{\sqrt{۲\pi n}} \sum_{t=۱}^{n} Z(\mathbf{s_i}, t) \exp\{-it\omega_k\} \right) \times \left(\frac{۱}{\sqrt{۲\pi n}} \sum_{t=۱}^{n} Z(\mathbf{s_j}, t) \exp\{it\omega_k\} \right) \\ &= \frac{۱}{۲\pi n} \left(\sum_{t=۱}^{n} Z(\mathbf{s_i}, t) \exp\{-it\omega_k\} \right) \times \left(\sum_{t=۱}^{n} Z(\mathbf{s_j}, t) \exp\{it\omega_k\} \right) = I_{\mathbf{s_i s_j}}(\omega_k) \end{aligned}$$

که در نهایت دوره‌نگار سری زمانی $\{Y_{ij}(t)\}$ به‌صورت زیر به‌دست می‌آید:

$$I_{\mathbf{s_i},\mathbf{s_j}}(\omega_k) = I_{\mathbf{s_i}}(\omega_k) + I_{\mathbf{s_j}}(\omega_k) - ۲Re[I_{\mathbf{s_i s_j}}(\omega_k)] \tag{۳.۳}$$

به طوری که $I_{\mathbf{s_i}}(.)$، $I_{\mathbf{s_j}}(.)$ و $I_{\mathbf{s_i s_j}}(.)$ به ترتیب دوره‌نگارهای مربوط به سری‌های زمانی $\{Z(\mathbf{s_i}, t)\}$ و $\{Z(\mathbf{s_j}, t)\}$ و دوره‌نگار متقابل آنها می‌باشند.

با امید ریاضی گرفتن از طرفین رابطه‌ی (۳.۳) نتیجه می‌شود:

$$E[I_{\mathbf{s_i},\mathbf{s_j}}(\omega_k)] = E[I_{\mathbf{s_i}}(\omega_k)] + E[I_{\mathbf{s_j}}(\omega_k)] - ۲ReE[J_{\mathbf{s_i}}(\omega_k)\overline{J_{\mathbf{s_j}}(\omega_k)}].$$

اگر تابع $g_{\mathbf{s_i},\mathbf{s_j}}(\omega, \boldsymbol{\theta})$ که به بردار پارامتر $\boldsymbol{\theta}$ بستگی دارد، تابع چگالی طیفی سری زمانی $\{Y_{ij}(t), i \neq j\}$ و $f_{\|\mathbf{h}\|}(\omega_k, \boldsymbol{\theta})$ تابع چگالی طیفی متقابل سری زمانی $\{Z(\mathbf{s_i}, t)\}$ و $\{Z(\mathbf{s_j}, t)\}$ باشند، از آن جایی که دوره‌نگار و دوره‌نگار متقابل به‌طور مجانبی برای تابع چگالی طیفی نااریب است (برکول و دیویس ، ۱۹۹۱) رابطه‌ی فوق به صورت زیر بیان می‌شود:

$$g_{\mathbf{s_i},\mathbf{s_j}}(\omega_k, \boldsymbol{\theta}) = ۲f(\omega_k, \boldsymbol{\theta}) - ۲f_{\|\mathbf{h}\|}(\omega_k, \boldsymbol{\theta})$$

که در آن $f(\omega_k,\boldsymbol{\theta})$ تابع چگالی طیفی فرایند مانای فضایی در زمان ثابت t، $\{Z(\mathbf{s_i},t);i=۱,۲,\ldots,m\}$ است و $f_{\mathbf{h}}(\omega_k,\boldsymbol{\theta})$ تابع چگالی طیفی متقابل سری‌های زمانی $\{Z(\mathbf{s_i},t);i=۱,۲,\ldots,m\}$ و $\{Z(\mathbf{s_j},t);j=۱,۲,\ldots,m\}$ می‌باشند که به صورت زیر بیان می‌شود:

$$f_{\mathbf{h}}(\omega_k,\boldsymbol{\theta})=\frac{۱}{۲\pi}\sum_{u=-\infty}^{\infty}C(\mathbf{s_i}-\mathbf{s_j},u)\exp\{-iu\omega_k\}$$

با در نظر گرفتن فرض همسانی کواریانس $C(\mathbf{h},u)$ ، تابع چگالی طیفی متقابل فوق به صورت

$$f_{\mathbf{h}}(\omega_k,\boldsymbol{\theta})=\frac{۱}{۲\pi}\sum_{u=-\infty}^{\infty}C(\parallel\mathbf{s_i}-\mathbf{s_j}\parallel,u)\exp\{-iu\omega_k\}=f_{\|\mathbf{h}\|}(\omega_k,\boldsymbol{\theta})$$

بیان می‌شود.

با استفاده از مطالب بالا سوبا رآو و همکاران (۲۰۱۴) تابع نیم تغییرنگار را در دامنه فرکانس به‌صورت زیر تعریف کردند

$$\begin{aligned}\frac{۱}{۲}g_{\|\mathbf{h}\|}(\omega_k,\boldsymbol{\theta})&=f(\omega_k,\boldsymbol{\theta})-f_{\|\mathbf{h}\|}(\omega_k,\boldsymbol{\theta})\\&=f(\omega_k,\boldsymbol{\theta})[۱-f_{\|\mathbf{h}\|}(\omega_k,\boldsymbol{\theta})/f(\omega_k,\boldsymbol{\theta})]\\&=f(\omega_k,\boldsymbol{\theta})[۱-W_{\|\mathbf{h}\|}(\omega_k,\boldsymbol{\theta})]\end{aligned}$$

به طوری که

$$W_{\|\mathbf{h}\|}(\omega_k,\boldsymbol{\theta})=f_{\|\mathbf{h}\|(\omega_k,\boldsymbol{\theta})}/f(\omega_k,\boldsymbol{\theta})$$

از آن‌جایی که $f(\omega_k,\boldsymbol{\theta})$ و همچنین $f_{\|\mathbf{h}\|}(\omega_k,\boldsymbol{\theta})$ اطلاعات تغییرنگار را به دامنه‌ی فرکانس منتقل می کنند، تابع $\frac{۱}{۲}g_{\|h\|}(۰,\boldsymbol{\theta})$ به عنوان نیم تغییرنگار در دامنه فرکانس نامگذاری شده است.

با فرض اینکه بردار $\mathbf{J}_{\|\mathbf{h}\|}=[J_{\|\mathbf{h}\|}(\omega_۱),J_{\|\mathbf{h}\|}(\omega_۲),\ldots,J_{\|\mathbf{h}\|}(\omega_M)]^\top$، که بردار تبدیلات فوریه سری زمانی $Y_{ij}(t)$ برای دو فضای i و j به فاصله $\parallel\mathbf{h}\parallel$ است، دارای توزیع نرمال مختلط چند متغیره با بردار میانگین صفر و ماتریس واریانس کواریانس Σ قطری با عناصر $[g_{\|\mathbf{h}\|}(\omega_۱,\boldsymbol{\theta}),g_{\|\mathbf{h}\|}(\omega_۲,\boldsymbol{\theta}),\ldots,g_{\|\mathbf{h}\|}(\omega_M,\boldsymbol{\theta})]^\top$ می‌باشد (بریلینجر ، ۲۰۰۱)، تابع درستنمایی و لگاریتم آن مطابق معمول تشکیل می‌شود.

$$\begin{aligned}L(\boldsymbol{\theta};\mathbf{J}_{\|\mathbf{h}\|}(\omega_۱),\ldots,\mathbf{J}_{\|\mathbf{h}\|}(\omega_M))&=(۲\pi\mid\Sigma\mid)^{-M/۲}exp(\frac{-۱}{۲}\mathbf{J}_{\|\mathbf{h}\|}\Sigma^{-۱}\mathbf{J}^*_{\|\mathbf{h}\|})\\\log L(\boldsymbol{\theta};\mathbf{J}_{\|\mathbf{h}\|}(\omega_۱),\ldots,\mathbf{J}_{\|\mathbf{h}\|}(\omega_M))&\propto-\frac{M}{۲}ln\mid\Sigma\mid-\frac{۱}{۲}\mathbf{J}_{\|\mathbf{h}\|}(\Sigma)^{-۱}\mathbf{J}^*_{\|\mathbf{h}\|}\\&=-\frac{۱}{۲}\sum_{k=۱}^{M}\ln\left(g_{\mathbf{s_i},\mathbf{s_j}}(\omega_k,\theta)\right)-\frac{۱}{۲}(\frac{\mathbf{J}_{\|\mathbf{h}\|}(\omega_۱)\overline{\mathbf{J}_{\|\mathbf{h}\|}(\omega_۱)}}{g_{\|\mathbf{h}\|}(\omega_۱)}+\\&\quad\ldots+\frac{\mathbf{J}_{\|\mathbf{h}\|}(\omega_M)\overline{\mathbf{J}_{\|\mathbf{h}\|}(\omega_M)}}{g_{\|\mathbf{h}\|}(\omega_M)})\\&=-\frac{۱}{۲}\left(\sum_{k=۱}^{M}lng_{\mathbf{s_i},\mathbf{s_j}}(\omega_k,a)+\sum_{k=۱}^{M}\frac{I_{\mathbf{s_i},\mathbf{s_j}}(\omega_k)}{g_{\mathbf{s_i},\mathbf{s_j}}(\omega_k)}\right)\end{aligned}\qquad(۴.۳)$$

سوبا رآو و همکاران (۲۰۱۴) تابع درستنمایی در دامنه فرکانس را براساس تمام مکان‌هایی که فاصله آن‌ها به اندازه $\parallel h\parallel$ است را با توجه به مطالب بالا به‌صورت زیر معرفی کردند

$$Q_{n,N(\mathbf{h})}(\boldsymbol{\theta})=\frac{۱}{\mid N(\mathbf{h})\mid}\sum_{(\mathbf{s_i},\mathbf{s_j})\in N(\mathbf{h})}\sum_{k=۱}^{M}\left[\ln(g_{\mathbf{s_i},\mathbf{s_j}}(\omega_k,\boldsymbol{\theta}))+\frac{I_{\mathbf{s_i},\mathbf{s_j}(\omega_k)}}{g_{\mathbf{s_i},\mathbf{s_j}}(\omega_k,\boldsymbol{\theta})}\right]$$

که در اینجا $N(\mathbf{h}) = \{\mathbf{s_i}, \mathbf{s_j} : \| \mathbf{s_i} - \mathbf{s_j} \| = \| \mathbf{h} \|\}$ می‌باشد.

از آن‌جایی که $Q_{n,N(\mathbf{h})}(\boldsymbol{\theta})$ بستگی به تاخیر فضایی $\mathbf{h}$ دارد $Q_n(\boldsymbol{\theta})$ به صورت زیر در نظر گرفته شده است:

$$Q_n(\boldsymbol{\theta}) = \frac{1}{L} \sum_{l=1}^{L} Q_{n,N(h_l)}(\boldsymbol{\theta})$$

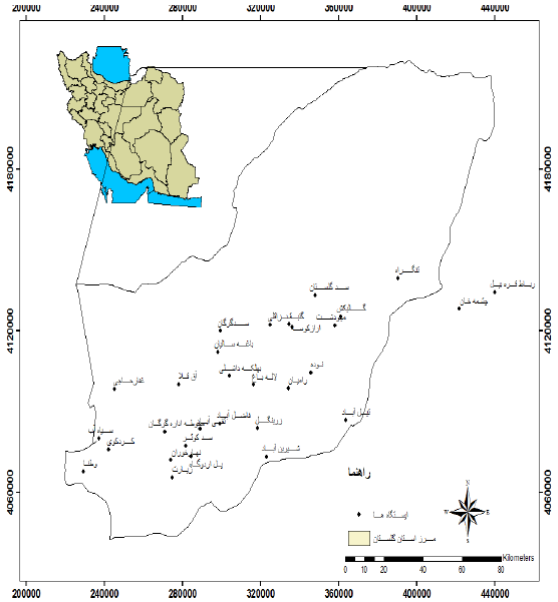
به عبارتی $Q_n(\boldsymbol{\theta})$ میانگین $Q_{n,N(\mathbf{h})}(\boldsymbol{\theta})$ها می‌باشد.

با توجه به رابطه‌ی (۴.۳) با مینیمم کردن تابع $Q_n(\boldsymbol{\theta})$ برآورد پارامترها حاصل می‌شود.

۴ برآورد پارامترهای تغییرنگار در دامنه فرکانس براساس داده های واقعی

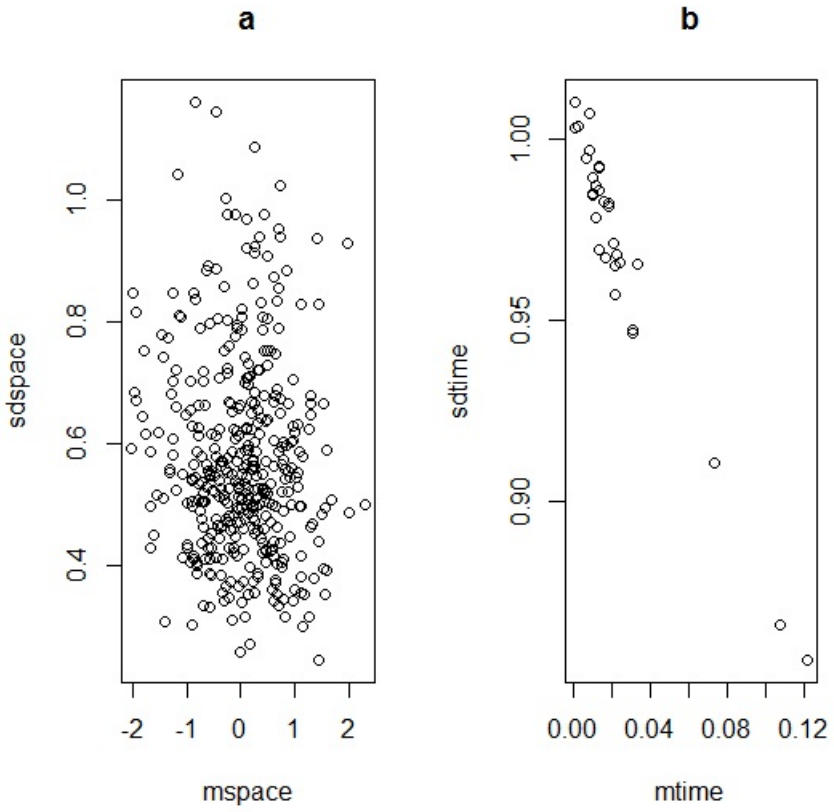
خشکسالی پدیده ای آرام و خزنده است که با آهنگ بسیار کند یک منطقه یا یک حوزه آبخیز را فرا گرفته و به یک بلای طبیعی تبدیل می‌گردد. یکی از خطرناک ترین و مخرب ترین اثرات خشکسالی ضایعاتی است که بر محیط زیست، منابع طبیعی، زیستگاه‌ها و اکوسیستم‌ها وارد می‌گردد. به هر حال کشور ایران و از جمله گلستان از وقوع خشکسالی مستثنی نبوده و همانند سایر کشورها و مناطق جهان درگذشته خشکسالی‌هایی در آن به وقوع پیوسته و در آینده نیز در آن رخ خواهد داد. شاخص SPI در حال حاضر به طور گسترده ای در امور تحقیقاتی و اجرایی در سراسر جهان جهت پایش خشکسالی استفاده می‌شود و شاخص RDI که بر مبنای بارندگی و تبخیر و تعرق محاسبه می‌شود و نسبت به SPI به متغیرها و تغییرات آب و هوایی حساستر است، در حال فراگیر شدن می‌باشد.

در این بخش ابتدا به طور توصیفی ایستایی فرایند فضایی ـ زمانی شاخص SPI سه ماهه بررسی شده است، (شکل ۲ و شکل ۳). سپس با استفاده از احتمال معناداری (P-value) از بین مدل‌های تغییرنگار، مدل جمعی ـ ضربی برای فرایند SPI سه ماهه در نظر گرفته شد و در پایان تغییرنگار تجربی فرایند فضایی ـ زمانی شاخص SPI سه ماهه در شکل ۴ نمایش داده شده است.



شکل ۱: موقعیت جغرافیایی استان گلستان در کشور

برای بررسی ایستایی سری فضایی ـ زمانی شاخص SPI، ابتدا نمودارهای توصیفی میانگین مکانی (زمانی) در مقابل انحراف استاندارد مکانی (زمانی) در شکل ۲، a (۲، b) رسم شده است.



شکل ۲: نمودار پراکنش میانگین مکانی در مقابل انحراف استاندارد.(a)میانگین مکانی در مقابل انحراف استاندارد مکانی، (b) میانگین زمانی در مقابل انحراف استاندارد زمانی

با مشاهده رفتار تصادفی نقاط در شکل ۲، a بررسی دقیق‌تر شرط ایستایی از طریق رسم نمودار خودهمبستگی نمونه‌ای سری زمانی میانگین مکانی صورت می‌گیرد، شکل ۳.

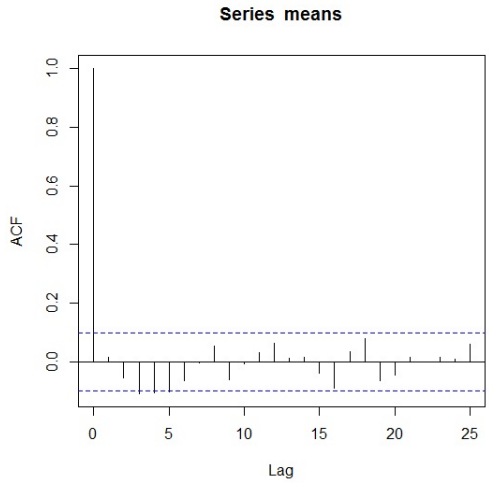
با توجه به اینکه مقادیر خودهمبستگی نمونه‌ای در تمام تاخیرها داخل باند می‌باشد شرط ایستایی قابل تایید است.

پس از بررسی شرط ایستایی، در گام بعدی قبل از برآورد پارامترهای تغییرنگار در دامنه فرکانس، نمودار تغییرنگار فضایی ـ زمانی تجربی شاخص SPI سه ماهه و برش های آن به ترتیب در شکل های ۴ و ۵ رسم شده است.

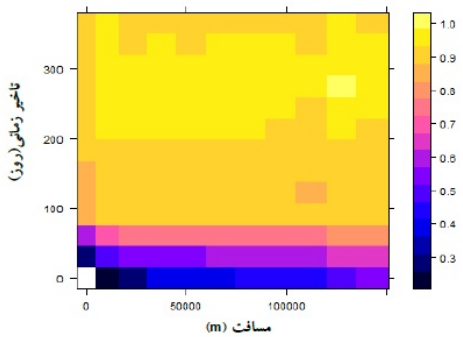
پس از بررسی فرایند SPI سه ماهه در دامنه زمان، پارامترهای تغییرنگار این فرایند در دامنه طیف توسط روش ارائه شده در این مقاله به عنوان بخشی از پایان‌نامه‌ی نویسنده اول این مقاله برآورد خواهند شد.

سپاس‌گزاری

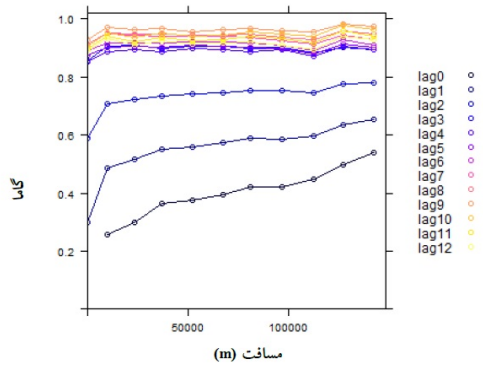
از جناب آقای دکتر محسن حسینعلی زاده و خانم اکرم عباسی در تهیه داده‌های شاخص خشکسالی SPI تشکر و قدردانی می‌شود.



شکل ۳: نمودار خودهمبستگی نمونه‌ای سری زمانی میانگین مکانی



شکل ۴: تغییرنگار فضایی ـ زمانی خشکسالی براساس شاخص SPI در بازه ی زمانی ۳ ماهه



شکل ۵: برش‌های تغییرنگار فضایی ـ زمانی خشکسالی برای هر تاخیر زمانی براساس شاخص SPI در بازه ی زمانی ۳ ماهه

مراجع

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

Brown, P. E., Roberts, G. O., Karesen, K. F., and Tonellato, S. (2000). Blur-generated Non-separable Space–time Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(4), 847–860.

Brillinger, D. R. (2001) *Time Series: Data Analysis and Theory*. Philadelphia: Society for Industrial Mathematics.

Chiles, J. P. and Delfiner, P. (1999). *Geostatistics; Modeling Spatial Uncertainty*, John Wiley and Sons, Inc.: New York.

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

Cressie, N. and Huang, H. C. (1999). Classes of Nonseparable, Spatiotemporal Stationary Covariance Functions, *Journal of the American Statistical Association*, **94**, 1330–1340.

Gneiting, T. (2002). Nonseparable, Stationary Covariance Functions for Space–time Data. *Journal of the American Statistical Association*, **97**(458), 590–600.

Gneiting T, Genton M. G., and Guttorp, P. (2007). Geostatistical Space–time Models, Stationarity, Separability, and Full Symmetry. *Monographs On Statistics and Applied Probability*, **107**: 151–175.

Isaaks, E. H. and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*, Oxford University Press, Oxford.

Li, B., Genton, M. G., and Sherman, M. (2007). A Nonparametric Assessment of Properties of Space-time Covariance Functions, *Journal of the American Statistical Association*, **102**, 736–744.

Ma, C. (2003). Families of Spatio-temporal Stationary Covariance Models. *Journal of Statistical Planning and Inference*, **116**(2), 489–501.

Sherman, M. (2010). *Spatial Statistics and Spatio-temporal Data: Covariance Functions and Directional Properties*, John Wiley and Sons, New York. .

Stein, M. L. (1999). *Interpolation of Spatial Data*, Springer, New York.

Stein, M. L. (2005). Space-time Covariance Functions, *Journal of the American Statistical Association*, **100**, 310–321.

Subba Rao, T. , Das, S. and Boshnakov, G. (2014). A Frequency Domain Approach For the Estimation of Parameters of Spatio-temporal Random Processes, *Journal of Time Series Analysis*,**35**, 357–377.

Subba Rao, T. and Terdik, G. Y. (2015). *A Space-time Covariance Function for Spatiotemporal Random Processes and Spatio-temporal Prediction (Kriging)*. ArXiv e-prints, 311.1981v2 [math.ST].

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶

مدل‌بندی داده‌های جرم با رگرسیون بتای فضایی با استفاده از تقریب لاپلاس آشیانی جمع بسته

کبری قلی‌زاده*، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده: معمولاً مدل‌های رگرسیونی بتای فضایی برای مدل‌بندی داده‌های وابسته فضایی نرخ (نسبت) که متأثر از متغیرهای تبیینی هستند مورد استفاده قرار می‌گیرند. گاهی در استنباط بیزی این مدل‌ها توزیع‌های پسینی فرم بسته‌ای ندارند و الگوریتم‌های مونت‌کارلوی زنجیر مارکوفی برای حل انتگرال‌های مربوط، ممکن است به دلیل تعدد پارامترها زمان‌بر و حتی با مشکل ناهمگرایی مواجه باشند. استفاده از استنباط بیز تقریبی می‌تواند راهکاری برای به دست آوردن توزیع‌های پسینی در این مدل‌ها باشد. در این مقاله از روش لاپلاس آشیانی جمع‌بسته برای تحلیل مدل رگرسیونی بتای فضایی استفاده می‌شود. بعلاوه مزایا و معایب این روش با روش‌های MCMC از لحاظ زمان و دقت مورد مقایسه قرار می‌گیرد. در انتها نحوه کاربست این مدل برای تعیین ارتباط داده‌های جرم با عوامل اجتماعی در شهر تهران نشان داده می‌شود.

واژه‌های کلیدی: مدل رگرسیونی بتای فضایی، تقریب لاپلاس آشیانی جمع‌بسته.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62M30، 91B72.

۱ مقدمه

در تحلیل متغیرهایی که با هم در ارتباط هستند معمولاً از مدل‌های رگرسیونی با فرض نرمال بودن متغیر پاسخ استفاده می‌شود. اما در بسیاری از مسائل کاربردی که با داده‌هایی مانند نرخ یا نسبت یک پیشامد که در بازه باز $(۰, ۱)$ توزیع‌شده‌اند سروکار داریم، این مدل‌ها ممکن است پیشگویی‌هایی برای متغیر پاسخ ارائه دهند که خارج از محدوده $(۰, ۱)$ باشند. یک‌راه حل ممکن استفاده از تبدیلات مناسب است به‌طوری‌که داده‌های تبدیل یافته روی خط اعداد حقیقی قرار گیرد و سپس میانگین متغیر تبدیل‌یافته برحسب

* **ارایه‌دهنده مقاله:** کبری قلی‌زاده، k.gholizadeh@modares.ac.ir

متغیرهای تبیینی موردنظر مدل‌بندی شود. اما در صورت استفاده از این روش، پارامترهای حاصل برحسب متغیر پاسخ واقعی به‌راحتی قابل تفسیر نخواهد بود(ویتینگوف و همکاران ، ۲۰۰۵). بعلاوه در بعضی از مطالعات توزیع داده‌های نرخ نامتقارن هستند و استفاده از توزیع‌های متقارن ازجمله توزیع نرمال منجر به نتایج نامعتبر می‌شوند. برای حل این مسئله، فراری و کریباری (۲۰۰۴) مدل رگرسیونی بتا[۱] را معرفی کردند. ممکن است داده‌ها وابسته فضایی باشند گامرمن و چیپدا (۲۰۱۳) با وارد کردن اثرات تصادفی مدل رگرسیونی بتای فضایی را مطرح کردند.

گاهی پارامترهای مدل رگرسیونی بتا به دلیل ماهیت مسئله متغیر هستند. در این موارد استفاده از رهیافت بیزی راه‌کاری الزامی است. علاوه بر این رهیافت بیزی می‌تواند محاسبات پیچیده برای برآورد مدل با رهیافت بسامدی را نیز میسر کند. اما تحلیل بیزی این مدلها عموماً نیازمند محاسبه انتگرال‌هایی چندگانه است. استفاده از الگوریتم‌های MCMC می‌تواند راه‌کاری برای این مسئله باشد. برنسکام و همکاران (۲۰۰۷) مدل رگرسیونی بتا با رهیافت بیزی را با الگوریتم‌های MCMC تحلیل کردند. اما در عمل این الگوریتم‌ها گاهی با مشکلاتی همچون طولانی بودن زمان محاسبات و ناهمگرایی مواجه می‌شوند. در این صورت تقریب انتگرال‌ها می‌تواند راهکاری مناسب برای حل این مشکل باشد. رو و همکاران (۲۰۰۹) در برآورد پارامترهای مدل خطی آمیخته تعمیم‌یافته با فرض گاوسی بودن میدان تصادفی روش تقریب لاپلاس آشیانی جمع‌بسته [۲] (INLA) را معرفی کردند و نشان دادند این روش ضمن حفظ دقت برآورد پارامترها سرعت محاسبات را نیز افزایش می‌دهد. در روش INLA ضمن استفاده از تقریب‌های گاوسی و لاپلاس از روش‌های عددی نیز در تقریب انتگرال‌ها استفاده می‌شود که موجب افزایش بیشتر سرعت محاسبات می‌شود.

در این مقاله برای تحلیل بیزی مدل رگرسیونی بتا و یافتن پسینی کناری از روش INLA استفاده می‌شود. برای این منظور در بخش ۲ میدان تصادفی مارکوفی گاوسی و در بخش ۳ مدل رگرسیونی بتای فضایی معرفی می‌شوند. تحلیل بیزی این مدل‌ها در بخش ۳ و روش INLA در بخش ۴ ارائه می‌شوند. در بخش ۵ با مطالعه‌ای شبیه‌سازی دقت و سرعت روش INLA ارزیابی و با روش MCMC مقایسه می‌شود. سپس با عنایت به توجه معاونت اجتماعی و پیشگیری از وقوع جرم قوه قضاییه به بررسی اهمیت آمار فضایی درمدل‌بندی و تحلیل داده‌های مربوط به جرم جنایت در کشور، با استفاده از مدل ارائه شده در این مقاله ارتباط بین داده‌های جرم با عوامل اجتماعی در شهر تهران تعیین می شود. در انتها بحث و نتیجه‌گیری ارائه خواهد شد.

۲ میدان تصادفی مارکوفی گاوسی

روش معمول برای لحاظ کردن همبستگی فضایی در تحلیل داده‌های فضایی استفاده از تابع کواریانس فضایی است. همچنین اگر موقعیت‌های فضایی مشاهدات بر یک مشبکه قرار داشته باشند، می توان با استفاده از مدل های اتورگرسیو شرطی[۳](CAR) همبستگی فضایی را از طریق یک ساختار کواریانس که تابعی از ماتریس همسایگی است در تحلیل داده‌ها لحاظ کرد. فرض کنید $\mathbf{Z} = (Z_1, \cdots, Z_n)$ برداری تصادفی بر یک مشبکه نامنظم با موقعیت‌های فضایی $\mathcal{I} = \{1, \cdots, n\}$ است و موقعیت‌های دارای مرز مشترک، همسایه یکدیگر محسوب می‌شوند. اگر n_i نماد تعداد همسایه‌های موقعیت i باشد، آن گاه مدل CAR با استفاده از چگالی‌های شرطی کامل به‌صورت

$$Z_i|(\mathbf{Z}_{-i}, \kappa) \sim N(\frac{1}{n_i}\sum_{j\in N(i)} Z_j, \frac{1}{n_i\kappa})$$

[۱] Beta regression model

[۲] Integrated Nested Laplace Approximation

[۳] Conditional Auto Regresive

تعریف می‌شود، که در آن $N(i)$ تمام همسایه‌های موقعیت i است. مدل CAR در حقیقت یک میدان تصادفی مارکوفی گاوسی[۴] (GMRF) است که مدل بسیج[۵] نیز نامیده می‌شود. یک GMRF میدان تصادفی گاوسی $\mathbf{Z} = (Z_۱, \cdots, Z_n) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ با ویژگی مارکوفی است. این ویژگی به‌راحتی در ماتریس دقت $\mathbf{Q} = (Q_{ij}) = \boldsymbol{\Sigma}^{-۱}$ قابل کدگذاری است، به‌طوری که $Q_{ij} = ۰$ اگر و تنها اگر Z_i و Z_j به شرط $\mathbf{Z}_{-ij}$ مستقل باشند، که در آن $\mathbf{Z}_{-ij}$ همه مولفه‌های $\mathbf{Z}$ به جز i و j است.

ویژگی مارکوفی با استفاده از گراف قابل تعریف است. گراف G مجموعه‌ای از راس‌هاست 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 غیر تهی از رئوس و E یال‌های آن است. با توجه به تعریف گراف، بردار تصادفی $\mathbf{Z} = (Z_۱, \cdots, Z_n)$ یک GMRF تحت گراف $G = (V, E)$ با میانگین $\boldsymbol{\mu}$ و ماتریس دقت $\mathbf{Q}_{n\times n} > ۰$ است اگر و تنها اگر تابع چگالی آن به صورت

$$\pi(\mathbf{Z}) = (۲\pi)^{-\frac{۱}{۲}} |\mathbf{Q}|^{\frac{۱}{۲}} \exp\{\frac{-۱}{۲}((\mathbf{Z}-\boldsymbol{\mu})^\top \mathbf{Q}((\mathbf{Z}-\boldsymbol{\mu})\}$$

باشد، به گونه‌ای که $Q_{ij} \neq ۰$ اگر و تنها اگر $i, j \in E$.

۳ مدل رگرسیونی بتای فضایی

متغیر تصادفی Y دارای توزیع بتا $\beta(p, q)$ است اگر تابع چگالی احتمال آن به‌صورت

$$f(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q} y^{p-۱} Y^{۱-q} I_{(۰,۱)}(y)$$

باشد، که در آن $p, q > ۰$، $\Gamma(\cdot)$ تابع گاما و $I_{(۰,۱)}(y)$ تابع نشانگر در بازه باز $(۰, ۱)$ است. میانگین و واریانس توزیع بتا برابر

$$\begin{aligned} \mu &= \frac{p}{p+q} \\ \sigma^۲ &= \frac{pq}{(p+q)^۲(p+q+۱)} \end{aligned}$$

است. با بازپارامتری کردن توزیع بتا به عنوان تابعی از μ و $\phi = p + q$ توزیع بتا می‌تواند به‌صورت

$$f(y|\phi,\mu) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱} Y^{(۱-\mu)\phi-۱}$$

باز نویسی شود .(جرجنسن ، ۱۹۷۷) در این صورت توزیع Y به‌صورت $\beta(\mu\phi, (۱-\mu)\phi)$ با میانگین μو واریانس $\sigma^۲ = \frac{\mu(۱-\mu)}{\phi+۱}$ است، از آنجا که ϕ رابطه معکوس با واریانس دارد پارامتر دقت توزیع نامیده می‌شود. در ادامه فرم بازپارامتری شده توزیع بتا به صورت $Y \sim \text{RB}(\mu, \phi)$ نمایش داده می‌شود.

فرض کنید Y_i متغیر پاسخ و $Y_i \sim \text{RB}(\mu_i, \phi)$، $i = ۱, \cdots, n$ باشد. در مدل رگرسیونی بتا با ثابت در نظر گرفتن پارامتر دقت ϕ، تابعی از میانگین متغیر پاسخ برحسب از متغیرهای موردنظر به‌صورت $g(\mu_i) = \sum_{j=۱}^{k} x_{ij}\beta_j$ مدل می‌شود، که در آن $g(\cdot)$ تابع پیوند یکنواخت، $x_{i۱}, \cdots, x_{ik}$ متغیرهای تبیینی و $\boldsymbol{\beta} = (\beta_۱, \cdots, \beta_k)$ بردار ضرایب رگرسیونی است. در مواردی که داده‌ها وابسته باشند با افزودن اثرات تصادفی u_i و پذیرش فرض استقلال شرطی $Y_i|u_i$، رابطه

$$g(\mu_i) = \sum_{j=۱}^{k} x_{ij}\beta_j + u_i = \eta_i \qquad (۱.۳)$$

[۴] Gaussian Markov Random Field

[۵] Besag

مدل رگرسیونی بتا با اثرات تصادفی نامیده می‌شود. حال اگر اثرات تصادفی u_i وابسته فضایی باشند مدل رگرسیونی بتای فضایی نام دارد (کلهری و محمدزاده ، ۲۰۱۶).

۴ تحلیل بیزی مدل رگرسیون بتای فضایی

اگر در (۱.۳) پارامترهای مدل و اثرات تصادفی $\mathbf{z} = \{\boldsymbol{\beta}, \mathbf{U}\}$ یک میدان تصادفی در نظر گرفته شود، در این صورت تابع درستنمایی به‌صورت

$$\mathcal{L}(\mathbf{Z}, \boldsymbol{\theta} \mid \mathbf{Y}) = \prod_{i \in \mathcal{I}} \frac{\Gamma(\phi)}{\Gamma(g^{-1}(\eta_i)\phi)\Gamma((1 - g^{-1}(\eta_i))\phi)} \times y_i^{g^{-1}(\eta_i)\phi - 1}(1 - y_i)^{(1-g^{-1}(\eta_i))\phi}, \qquad (۱.۴)$$

خواهد بود، که در آن $\boldsymbol{\theta} = (\alpha, \phi)$ بردار ابرپارامتر است. چنانچه $\boldsymbol{\theta}$ دارای توزیع پیشینی $\pi(\boldsymbol{\theta})$ باشد، توزیع پسینی به‌صورت

$$\pi(\mathbf{Z}, \boldsymbol{\theta}|\mathbf{Y}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{Z}|\alpha)\mathcal{L}(\mathbf{Z}, \phi \mid \mathbf{Y}) \qquad (۲.۴)$$

حاصل می‌شود، که با روش‌های MCMC می‌توان برآورد بیزی پارامترها را به‌دست آورد. اما از آنجا که این روش شامل نمونه‌گیری‌های تکراری از شرطی‌های کامل $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ و $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x})$ است، با پیچیده شدن مدل ممکن است محاسبات آن زمان‌بر و گاهی با عدم همگرایی زنجیره تولید شده مواجه شود. امروزه سرعت محاسبات برای کاربران مدل‌های دقیق اما پیچیده از اهمیت بالایی برخورد دار است. رو و همکاران (۲۰۰۹) تقریبی بیزی با نام تقریب لاپلاس آشیانی جمع‌بسته را معرفی کردند که ضمن حفظ دقت محاسبات به دلیل استفاده از تقریب‌های گاوسی، لاپلاس و روش‌های محاسبات عددی جایگزینی مناسب برای الگوریتم‌های MCMC است. قلی‌زاده و همکاران (۱۳۹۲) از روش INLA در مدل رگرسیونی جمعی ساختاری استفاده کرده و در یک مطالعه شبیه‌سازی بر روی داده‌های جرم نشان دادند دقت روش INLA اختلاف ناچیزی با روش MCMC دارد، در حالی که محاسبات INLA در حدود شش برابر سریع‌تر از MCMC انجام شده است.

۵ تقریب لاپلاس آشیانی جمع‌بسته

در تحلیل بیزی به روش INLA، محاسبه پسینی کناری $\tilde{\pi}(Z_i|\mathbf{Y})$ مورد نظر است که می توان آن را به صورت

$$\pi(Z_i|\mathbf{Y}) = \int \pi(Z_i|\mathbf{Y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\boldsymbol{\theta} \qquad i = 1, \cdots, n \qquad (۱.۵)$$

نوشت. در روش INLA از یک ساختار آشیانی برای تقریب معادله (۱.۵) استفاده می‌شود، که شامل سه گام است. اول، توزیع پسینی کناری $\pi(\boldsymbol{\theta}|\mathbf{Y})$ به صورت

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) \propto \left.\frac{\pi(Z_i, \mathbf{Y}, \boldsymbol{\theta})}{\tilde{\pi}_G(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta})}\right|_{\mathbf{Z}=\mathbf{Z}^*(\boldsymbol{\theta})} \qquad (۲.۵)$$

تقریب زده می‌شود، که در آن $\tilde{\pi}_G(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta})$ تقریب گاوسی (رو و هلد ، ۲۰۰۵) برای توزیع شرطی کامل $\mathbf{Z}$ و $\mathbf{Z}^*(\boldsymbol{\theta})$ مد توزیع شرطی کامل $\mathbf{Z}$ است. گام دوم ، محاسبه $\pi(Z_i|\mathbf{Y}, \boldsymbol{\theta})$ است که رو و همکاران (۲۰۰۹) سه تقریب گاوسی، لاپلاس و لاپلاس ساده شده را برای تقریب پیشنهاد دادند. ساده‌ترین روش تقریب گاوسی و دقیق‌ترین روش تقریب لاپلاس است. در نهایت، روش

تقریب لاپلاس ساده‌شده با کاهش جزئی در دقت از لحاظ محاسباتی سهل‌تر از تقریب لاپلاس است. گام سوم، ترکیب دو گام پیشین و استفاده از روش انتگرال گیری عددی برای رسیدن به هدف است. توزیع پسینی $\pi(Z_i|\mathbf{y})$ به صورت

$$\tilde{\pi}(Z_i|\mathbf{y}) = \sum_k \tilde{\pi}(Z_i|\mathbf{Y}, \boldsymbol{\theta}_k)\tilde{\pi}(\boldsymbol{\theta}_k|\mathbf{Y})\ \triangle_k$$

به‌دست می‌آید، که درآن k تعداد $\boldsymbol{\theta}$ های انتخاب شده از تکیه‌گاه و $\triangle_k$ وزن‌های این نقاط هستند (رو و همکاران ، ۲۰۰۹).

۶ مطالعه شبیه‌سازی

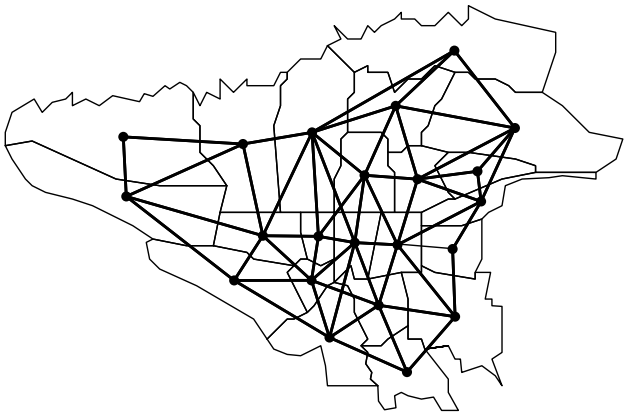
در این بخش برای مقایسه دو روش MCMC و INLA در برآورد پارامترهای مدل رگرسیون فضایی بتا، از این مدل مجموعه داده‌ای به حجم ۲۲ شبیه سازی و با دو روش اجرا می شود. متغیر پاسخ Y_i با توزیع بازپارمتری‌شده بتا به‌صورت

$$Y_i|\mu_i \sim \text{RB}(\mu_i, \phi), \qquad i = ۱, \cdots, ۲۲$$

را در نظر بگیرید که میانگین آن با تابع پیوند لوجیت[۶] به صورت

$$\eta_i = g(\eta_i) = \frac{\exp\{\mu_i\}}{۱ + \exp\{\mu_i\}} = \beta_۰ + \beta_۱ x_{۱i} + u_i, \qquad i = ۱, \cdots, ۲۲ \tag{۱.۶}$$

مدل‌بندی شده است. برای اجرای شبیه‌سازی، مقادیر $\boldsymbol{\beta} = (\beta_۰, \beta_۱) = (۱, ۳)$ و $\varphi = ۲۰$ در نظر گرفته شده و برای β توزیع پیشینی ناآگاهی بخش $N(۰, ۱۰۰)$، برای ϕ توزیع پیشینی گامای $\Gamma(۱, ۰/۱)$ و برای اثرات تصادفی مدل بسیج روی گراف شهر تهران (شکل ۱) در نظر گرفته شده است. ابتدا از توزیع نرمال استاندارد برای متغیر تبینی $x_۱$ و از مدل بسیج برای اثرات تصادفی $\mathbf{U}$ نمونه گرفته شده است. سپس با محاسبه مقادیر μ_i، از توزیع $RB(\mu_i, ۲۰)$ نمونه y_i را تولید نموده و این کار ۲۰۰ بار تکرار می‌شود. در هر تکرار، پارامترهای مدل (۱.۶) با دو روش INLA و MCMC برآورد می‌شود. روش MCMC بر اساس نمونه‌ای تصادفی به حجم ۱۰۰۰ با ۵۰۰۰ تکرار، مرحله داغیدن ۱۰۰۰ و تاخیر ۴۰ انجام شده است. میانگین و انحراف معیار پارامترهای مدل به همراه مقادیر MSE در جدول ۱ ارائه شده است.



شکل ۱: مناطق ۲۲گانه و گراف شهر تهران

[۶] Logit

جدول ۱: برآورد پارامترهای مدل (۱.۶) و ملاک MSE

		MCMC			INLA		
			انحراف			انحراف	
پارامتر	مقدار واقعی	برآورد	استاندارد	MSE	برآورد	استاندارد	MSE
β_0	۱	۰/۹۳۰	۰/۲۱۱	۰/۰۵۴	۰/۹۷۲	۰/۱۶۷	۰/۰۰۱
β_1	۳	۲/۵۳۶	۰/۴۰۲	۰/۵۲۲	۲/۸۴۵	۰/۳۴۹	۰/۰۲۴
ϕ	۲۰	۱۶/۳۱۳	۹/۴۵۳	۳۹/۴۲۶	۱۴,۱۵۲	۵/۶۸۰	۳۴/۱۹۴
u_1	۰/۴۷۵	۰/۱۸۴	۰/۵۹۷	۰/۱۸۴	۰/۱۸۱	۰/۳۷۴	۰/۰۸۷
u_2	۰/۰۶۶	۰/۱۰۸	۰/۵۴۳	۰/۰۶۰	۰/۰۰۳	۰/۲۹۷	۰/۰۰۴
u_3	۱/۰۱۶	۰/۳۱۴	۰/۶۱۴	۰/۵۶۲	۰/۱۰۴	۰/۳۳۸	۰/۸۳۲
u_4	۰/۴۰۳	۰/۰۴۵	۰/۶۳۳	۰/۱۷۳	−۰/۰۰۹	۰/۳۴۶	۰/۱۷۰
u_5	−۰/۱۲۴	−۰/۰۵۱	۰/۵۴۷	۰/۱۰۱	−۰/۰۷۰	۰/۰۳۲۳	۰/۰۰۳
u_6	۰/۶۷۸	۰/۲۰۲	۰/۵۹۷	۰/۳۳۳	۰/۰۳۰	۰/۳۲۲	۰/۴۲۰
u_7	−۰/۶۴۹	−۰/۳۶۳	۰/۵۶۱	۰/۱۴۶	−۰/۱۰۵	۰/۳۰۵	۰/۲۹۶
u_8	۰/۲۷۲	۰/۱۱۳	۰/۵۹۱	۰/۱۳۲	۰/۰۱۴	۰/۳۴۳	۰/۰۶۶
u_9	−۰/۳۳۶	−۰/۰۹۴	۰/۵۳۱	۰/۱۳۳	−۰/۱۰۲	۰/۳۰۹	۰/۰۷۰
u_{10}	−۰/۶۲۲	−۰/۳۶۱	۰/۵۵۰	۰/۱۴۹	−۰/۱۹۵	۰/۳۳۸	۰/۱۸۲
u_{11}	۰/۳۳۷	۰/۱۴۹	۰/۵۲۲	۰/۱۰۹	۰/۰۲۱	۰/۲۹۱	۰/۱۰۰
u_{12}	−۱/۰۹۹	−۰/۵۷۲	۰/۵۳۶	۰/۳۶۶	−۰/۲۴۶	۰/۳۳۸	۰/۷۲۷
u_{13}	−۰/۲۱۹	−۰/۳۴۴	۰/۵۵۴	۰/۰۹۷	−۰/۱۰۹	۰/۲۹۱	۰/۰۱۲
u_{14}	−۰/۰۵۸	−۰/۲۷۳	۰/۸۵۹	۰/۱۴۰	−۰/۰۲۶	۰/۳۳۸	۰/۰۰۱
u_{15}	−۰/۳۳۳	−۰/۰۰۴	۰/۷۵۸	۰/۱۴۸	۰/۰۸۶	۰/۳۱۲	۰/۱۷۶
u_{16}	۱/۰۳۳	۰/۳۴۵	۰/۵۶۶	۰/۵۶۲	۰/۱۲۳	۰/۴۴۷	۰/۸۲۹
u_{17}	۰/۵۶۳	−۰/۱۹۴	۰/۵۸۳	۰/۲۱۷	۰/۰۲۶	۰/۴۰۱	۰/۲۸۸
u_{18}	−۰/۵۵۵	−۰/۱۸۶	۰/۷۷۳	۰/۱۷۴	−۰/۱۰۸	۰/۳۱۳	۰/۲۰۰
u_{19}	۰/۳۸۷	۰/۲۹۸	۰/۵۳۹	۰/۱۰۱	۰/۱۵۵	۰/۴۲۰	۰/۰۵۴
u_{20}	۰/۶۵۰	۰/۴۵۱	۰/۶۷۳	۰/۱۹۷	۰/۱۹۳	۰/۳۸۸	۰/۲۰۹
u_{21}	۰/۰۷۲	۰/۰۰۰	۰/۶۸۳	۰/۰۸۵	−۰/۰۸۰	۰/۳۸۶	۰/۰۲۳
u_{22}	۰/۱۰۵	−۰/۱۸۳	۰/۷۰۹	۰/۲۲۰	۰/۰۰۴	۰/۴۱۴	۰/۰۱۰

همان طور که ملاحظه می‌شود در برآورد ضرایب رگرسیونی، روشINLA از اریبی کمتری نسبت به روش MCMC برخوردار است، ضمن اینکه انحراف استاندارد و MSE در روش INLA کمتر است. در برآورد پارامتر دقت روش MCMC دارای اریبی کمتری است اما انحراف استاندارد و MSE روش INLA کوچکتر است که می‌توان نتیجه گرفت روش INLA قابل اعتمادتر است. در برآورد اثرات تصادفی در اکثر مواقع روش INLA با توجه به معیار MSE عمل کرد بهتری از MCMC داشته است.

زمان انجام محاسبات با روش INLA، دقیقا ۲/۶۷ ثانیه ولی با روش MCMC حدود ۱۰ ساعت به طول انجامید، یعنی محاسبات INLA حدود ۲۰۰ برابر سریع‌تر از روش MCMC انجام شده است.

۷ مدل‌بندی نرخ جرم مواد مخدر در شهر تهران

در این بخش داده‌های جرم «مواد مخدر » مناطق ۲۲ گانه شهر تهران با استفاده از مدل رگرسیونی بتای فضایی با روش INLA مورد تحلیل قرار می‌گیرد. داده‌ها شامل نرخ جرم مواد مخدر مناطق به عنوان متغیر پاسخ و عوامل اجتماعی مانند نرخ طلاق، نرخ بیکاری و نرخ باسوادی به عنوان متغیرهای تبیینی هستند. جرم مواد مخدر به همه جرم‌هایی اتلاق می‌شود که در ارتباط با مواد مخدر رخ می‌دهد. فرض بتا بودن توزیع متغیر پاسخ با pـ مقدار ۰/۴۷ مورد تایید قرار گرفت. برای بررسی تاثیر متغیرهای تبیینی معرفی شده بر جرم مواد مخدر مدل

$$\eta_i = \boldsymbol{\beta}\mathbf{Z} + u_i, \qquad i = ۱, \cdots, ۲۲ \tag{۱.۷}$$

در نظر گرفته شده است، که در آن $\boldsymbol{\beta} = (\beta_۰, \beta_U, \beta_D, \beta_L)$ بردار ضرایب به ترتیب مربوط به عرض از مبدا، نرخ بیکاری، نرخ طلاق و نرخ باسوادی است، همچنین u_i برای لحاظ کردن همبستگی فضایی بین مناطق است. برای هر یک از ضرایب ثابت توزیع پیشنی ناآگاهی بخش نرمال با میانگین صفر و واریانس نسبتا بزرگ ۱۰۰۰ و برای اثر تصادفی فضایی، مدل پیشینی بسیج با پارامتر τ_s اختیار شده است. در این مدل میدان تصادفی $\mathbf{Z} = \{\boldsymbol{\beta}_۱, \mathbf{U}\}$ و بردار ابرپارامتر $\boldsymbol{\theta} = (\phi, \tau_s)$ است. برای هر یک از درایه های ابرپارامتر $\boldsymbol{\theta}$ توزیع پیشینی گامای $G(۰/۱, ۰/۱)$ در نظر گرفته شده است. برآورد پارامترهای این مدل برای اثرهای ثابت به همراه انحراف استاندارد آن‌ها در جدول ۲ ارائه شده است. با توجه به بردار ضرایب ملاحظه می‌شود که نرخ باسوادی تاثیر منفی بر وقوع این جرم دارد یعنی هرچه تعداد باسوادی بیشتر شود جرم مواد مخدر کمتر رخ می‌دهد. نرخ بیکاری و نرخ طلاق هر دو تاثیر مثبت بر وقوع جرم دارند یعنی با افزایش بیکاری و طلاق در جامعه جرم مواد مخدر بیشتر اتفاق می‌افتد.

جدول ۲: برآورد ضرایب رگرسیونی مدل (۱.۷)

ضرایب	برآورد	انحراف استاندارد
عرض از مبدا	۰/۷−	۳/۵۷
نرخ بیکاری	۰/۲۶	۰/۸۳
نرخ طلاق	۰/۷۳	۰/۵۷
باسوادین	۹/۰۶−	۵/۸۵

بحث و نتیجه‌گیری

مدل‌های رگرسیون فضایی بتا که در مدل‌بندی داده‌های نرخ فضایی نرخ پرکاربرد هستند معرفی شدند. در تحلیل بیزی این مدل‌ها، زمانی که مدل با افزایش اندازه نمونه و افزودن اثرات تصادفی پیچیده‌تر می‌شود، توزیع‌های پسینی فرم بسته‌ای ندارند و محاسبات با روش‌های MCMC زمان‌بر هستند. اما محاسبات با استفاده از تقریب لاپلاس آشیانی جمع بسته به دلیل استفاده از تقریب‌های گاوسی و لاپلاس و روش‌های عددی سرعت می‌یابد. در مطالعه شبیه‌سازی نشان داده شد روش INLA در عین حالی که سریع‌تر از روش MCMC عمل می‌کند، نتایج دقیقی به‌دست می‌دهد که اختلاف ناچیزی با نتایج حاصل از روش MCMC دارند. در بررسی نرخ جرم مواد مخدر شهر تهران مشخص شد عوامل اجتماعی مانند نرخ طلاق و نرخ بیکاری تاثیر مثبت و نرخ باسوادی تاثیر منفی بر وقوع این جرم دارد.

سپاس‌گزاری

نویسندگان از معاونت اجتماعی و پیشگیری از وقوع جرم قوه قضاییه به خاطر پشتیبانی از موضوع این تحقیق کمال تشکر را دارند.

مراجع

قلی‌زاده ک، قیومی ز، محمدزاده م، (۱۳۹۲). تحلیل فضایی رگرسیون جمعی ساختاری و مدل‌بندی داده‌های جرم شهر تهران با تقریب لاپلاس آشیانی جمع بسته، *مجله علوم آماری*، **۱**، ۱۲۴-۱۰۳.

Branscum, A. J., Johnson W. O. and Thurmond, M. (2007). Bayesian Beta Regression: Applications to Houshold Expenditure Data and Genetic Distance Between Food and Mouth Dieseas Viruses.*Australian and New Zealand Journal of Statistics*, **49**, 287-301.

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.

Jørgensen, B. (1997). *The theory of Dispersion Models*, Chapman and Hall, London.

Kalhori, L., and Mohammadzadeh, M. (2016). Spatial Beta Regression Model with Random Effect, *Journal of Statistical Research of Iran*, **13**, 215-230.

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.

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.

Vittinghoff, E., Gilidden D., Shiboski S. and McCulloch C. (2005).*Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*, Springer, New York.

دومین سمینار آمار فضایی و کاربردهای آن
۳ – ۴ آبان ۱۳۹۶

تحلیل بیزی و بیزی تقریبی برای داده‌های قیمت ملک تهران

امید کریمی*، فاطمه حسینی

گروه آمار، دانشگاه سمنان

چکیده: داده‌های قیمت ملک شهر تهران نسبت به مناطق و محله‌های مختلف دارای همبستگی مکانی هستند. همچنین قیمت ملک فروخته شده در ماه‌های مختلف سال متفاوتند، بنابراین همبستگی از نوع زمانی نیز دارند. برای تحلیل این نوع داده‌ها از مدل های فضایی ـ زمانی استفاده می‌شود. در این مقاله مدل‌های مختلف فضایی ـ زمانی برای داده‌های قیمت ملک فروخته شده در نظر گرفته شده است و از هیافت بیزی برای برازش مدل‌ها و پیشگویی در موقعیت‌های جدید و پیش‌بینی در زمان‌های آینده قیمت ملک شهر تهران استفاده شده است. با توجه به افزایش بعد مکان و زمان ممکن است خیلی از مدل های فضایی زمانی از کارایی و به خصوص سرعت محاسباتی مناسبی برخوردار نباشند، لذا یک راهکار مناسب استفاده از رهیافت بیزی تقریبی به وسیله میدان تصادفی مارکوفی گاوسی است. بهترین مدل با ملاک میانگین مجذور خطای پیشگویی‌ها و مقایسه زمان محاسبات معرفی می‌گردد سپس نقشه پیشگویی قیمت ملک تهران ارائه می‌شود.

واژه‌های کلیدی: مدل‌های فضایی ـ زمانی، رهیافت بیزی، رهیافت بیزی تقریبی

کد موضوع‌بندی ریاضی (۲۰۱۰): 91B72، 62M10، 62F15.

۱ مقدمه

حجم بالایی از داده‌های فضایی ـ زمانی به‌طور روزانه و ماهانه جمع‌آوری می‌شوند. به‌عنوان مثال قیمت املاک فروخته شده شهر تهران به طور روزانه در سامانه بازار املاک ایران ثبت می‌گردد. وجود بعد فضایی و زمانی به‌طور توأم و افزایش آن‌ها پیچیدگی‌های اساسی بر تحلیل و استنباط داده‌ها اضافه می‌کند. به‌ویژه برای مدل‌های سلسله‌مراتبی بیزی پیچیدگی محاسبات به سرعت افزایش می‌یابد. لزوم مدل‌هایی کارا و سریع برای این‌گونه از داده‌ها از اهمیت خاصی برخوردار است. کارهای مختلفی در این زمینه انجام شده است که به عنوان مثال کتاب اخیر کرسی و ویکل (۲۰۱۱) و منابع معرفی شده توسط آنها اشاره کرد. مدل‌های فضایی ـ زمانی پویا به‌صورت سلسله‌مراتبی توسط ساهو و همکاران (۲۰۰۷) معرفی گردید. هرچند این مدل‌ها برای تحلیل داده‌های حجیم در سطح وسیعی از ناحیه مورد مطالعه مناسب نبودند. مشکل اصلی در معکوس کردن ماتریس کوواریانس فضایی و تکرار آن در

* **ارایه‌کننده مقاله:** امید کریمی omid.karimi@semnan.ac.ir

الگوریتم‌های برازش مدل بود (بانرجی و همکاران ، ۲۰۰۴). کارهای متعددی در این زمینه صورت گرفته است، به‌عنوان مثال کرسی (۱۹۹۳) از کریگیدن محلی استفاده کرد.

یک زیرنمونه‌گیری از بعد فضایی بالا به‌وسیله روش پنجره متحرک توسط هارتمن و هوسجر (۲۰۰۸) پیشنهاد شد و تکنیک‌هایی براساس کارایی محاسبات با استفاده از الگوریتم‌های محاسباتی ماتریس پراکنده[۱] معرفی کردند که برای این منظور از تقریب‌های میدان تصادفی مارکوفی گاوسی رو و هلد (۲۰۰۶) بهره گرفتند. بانرجی و همکاران (۲۰۰۸) یک روش تقریبی ساده با استفاده از فرآیند پیشگوی گاوسی[۲] (GPP) برای مدل بیزی داده‌های فضایی چند متغیره پیشنهاد کردند و روش آن‌ها توسط فینلی و همکاران (۲۰۰۹) و ساهو و باکار (۲۰۱۲) بسط داده شد و همچنین ساهو و باکار (۲۰۱۱) مدل‌های اتورگرسیو بیزی را برای داده‌ها واقعی ازون به‌کار گرفتند.

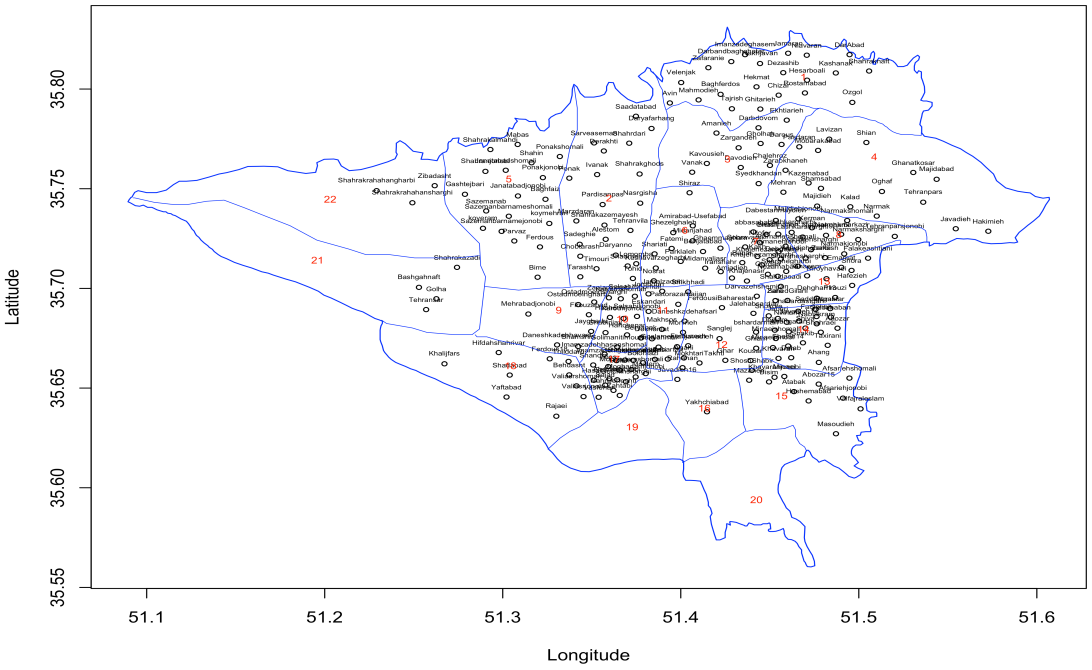
به دلیل داشتن مشکلاتی مانند طولانی بودن زمان محاسبات و گاهی عدم همگرایی در مدل‌های فضایی پیچیده سلسله‌مراتبی، رو و مارتینو (۲۰۰۷) روش تقریب لاپلاس آشیانی جمع‌بسته[۳] (INLA) را برای تحلیل مدل‌های پیچیده فضایی معرفی کردند و نشان دادند این روش ضمن حفظ دقت برآورد پارامترها سرعت محاسبات را نیز افزایش می‌دهد. ایدسویک و همکاران (۲۰۰۹) از تقریب معرفی شده توسط رو و مارتینو (۲۰۰۷) و رو و همکاران (۲۰۰۹) استفاده کردند و به تحلیل بیزی تقریبی مدل‌های آمیخته خطی تعمیم‌یافته فضایی پراختند. حسینی و همکاران (۲۰۱۱) و حسینی و محمدزاده (۲۰۱۲) با به‌کار بردن توزیع چوله نرمال بسته برای متغیرهای پنهان تحلیل بیزی و بیزی تقریبی این مدل‌ها را با الگوریتم‌های مونت کارلوی زنجیر مارکوفی و رهیافت INLA بیان و مقایسه نمودند. قلی‌زاده و همکاران (۲۰۱۳) از تقریب INLA استفاده نمودند و به تحلیل بیزی تقریبی مدل‌های رگرسیون جمعی ساختاری پرداختند و داده‌های جرم شهر تهران را با این مدل و روش مورد تحلیل قرار دادند. کاملتی و همکاران (۲۰۱۳) با درنظر گرفتن میدان تصادفی مارکوفی گاوسی[۴] (GMRF) و رهیافت بیز تقریبی معرفی شده توسط رو و همکاران (۲۰۰۹) به مطالعهٔ داده‌های فضایی_ زمانی پرداخته‌اند.

در این مقاله روش‌های ارائه شده روی داده‌های قیمت ملک شهر تهران به‌کار گرفته می‌شود و با رهیافت بیزی و بیزی تقریبی تحلیل و مورد مقایسه قرار می‌گیرند و روش محاسباتی سریع برای داده‌های فضایی ـ زمانی حجیم براساس روش‌های محاسباتی بیزی سلسله‌مراتبی روی داده‌های قیمت ملک شهر تهران در سال ۱۳۹۴ ارائه می‌گردد. یک تحلیل اکتشافی روی داده‌های قیمت مسکن در بخش ۲ فراهم می‌گردد و در بخش ۳ مدل‌های سلسله‌مراتبی بیزی برای داده‌ها معرفی می‌شوند. در بخش ۴ میدان تصادفی مارکوفی گاوسی معرفی می شود و رهیافت معادلات دیفرانسیل جزیی تصادفی SPDE در بخش ۵ بیان می شود. سپس برازش مدل و پیشگویی فضایی و پیش‌بینی در زمان‌های آینده روی داده‌های قیمت مسکن در بخش ۶ انجام می‌گردد و در انتها بحث و نتیجه‌گیری بیان می‌گردد.

۲ تحلیل اکتشافی داده‌های قیمت مسکن

داده‌های قیمت مسکن تهران به‌صورت روزانه در سامانه اطلاعات بازار املاک ایران ثبت می‌گردد. این داده‌ها را می‌توان از سایت www.hmi.mrud.ir مشاهده کرد. قیمت ملک نوساز (تا یک سال ساخت) برای محله‌های ۲۲ منطقه به‌صورت روزانه استخراج

[۱] Sparse matrix
[۲] Gaussian Prediction Process
[۳] Integrated Nested Laplace Approximation
[۴] Gaussian Markov Random Field



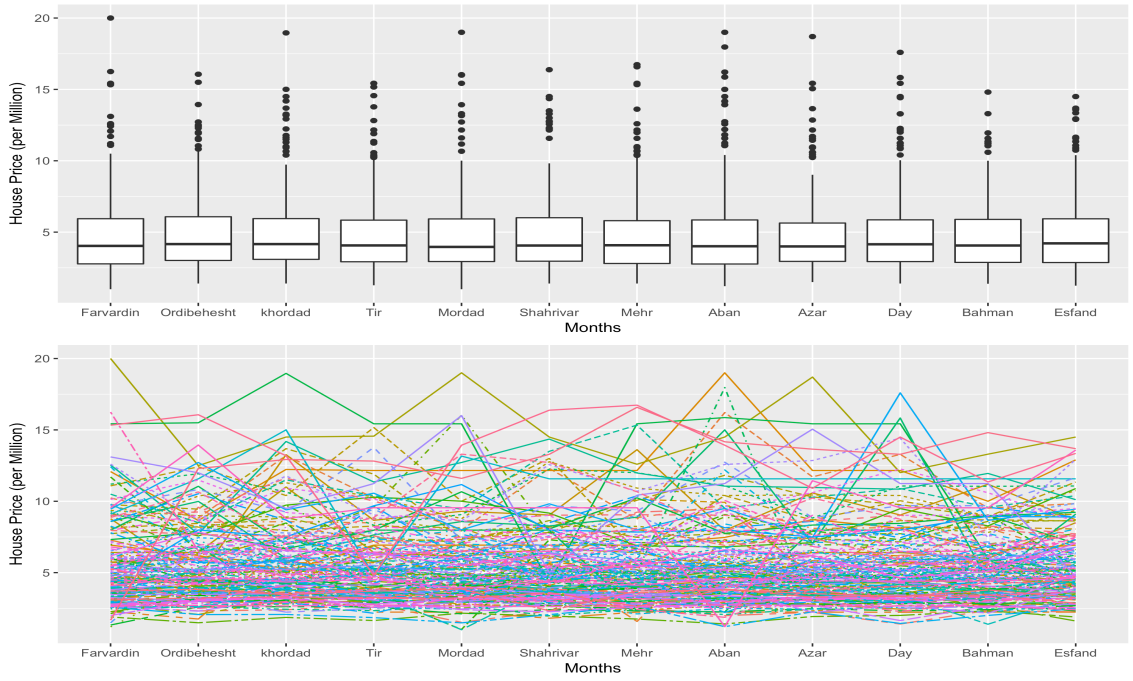
شکل ۱: موقعیت مشاهدات روی نقشه تهران.

گردید. با توجه به این‌که به‌صورت روزانه داده‌های گمشده زیادی برای هر موقعیت (محله‌ها) وجود دارد، متوسط ماهیانه قیمت ملک نوساز را برای هر محله شهر تهران واقع در مناطق ۲۲ گانه محاسبه گردید به‌طوری که ۳۲۱۶ مشاهده (یعنی ۲۶۸ محله در ۱۲ ماه) به‌دست آمد. موقعیت‌های محله‌ها روی نقشه تهران در شکل ۱ نشان داده‌اند. شکل ۲ بالا نمودار جعبه‌ای قیمت مسکن برحسب ماه‌های سال و شکل ۲ (پایین) نمودار سری زمانی داده‌ها در ۲۶۸ موقعیت را نشان می‌دهد و حدود ۳۱ درصد داده‌ها گمشده‌اند.

هدف پیش‌بینی در ماه‌هایی آینده و کل ناحیه شهر تهران براساس بهترین مدل فضایی ـ زمانی می‌باشد. با توجه به نمودار احتمال نرمال (بالا ـ چپ) که در شکل ۳ مشاهده می‌شود، داده‌های قیمت مسکن چوله به راست هستند و با یک تبدیل لگاریتمی داده‌ها متقارن‌تر و به توزیع نرمال نزدیک‌تر می‌شوند. (شکل ۳، بالا ـ راست). نمودار پراکنش برحسب طول و عرض جغرافیایی در شکل۳ پایین رسم شده‌اند که یک روند خطی با شیب مثبت در جهت جنوب به شمال (عرض جغرافیایی) مشاهده می‌شود. بنابراین در مدل‌های فضایی ـ زمانی که در بخش بعد ارائه خواهند شد، از طول و عرض جغرافیایی به‌عنوان متغیر کمکی استفاده می‌شود.

۳ مدل های سلسله‌مراتبی بیزی

فرض کنید $z_\ell(s_i,t)$ مقادیر مشاهدات پاسخ در موقعیت s_i، $i = ۱,\ldots,n$ و ماه tام در سال ℓام برای $\ell = ۱,\ldots,r$ و $t = ۱,\ldots,T_\ell$ باشد. در این‌جا دو اندیس زمان به‌صورت ماه درون سال مشخص شده است که می‌تواند به‌صورت روز درون سال باشد. قرار دهید $z_{\ell t} = (z_\ell(s_۱,t),\ldots,z_\ell(s_n,t))^\top$ و فرض کنید $x_{\ell j}(s,t)$ متغیر کمکی jام در ماه tام سال ℓام باشد و قرار



شکل ۲: نمودارهای جعبه‌ای و سری زمانی داده‌های قیمت مسکن برحسب میلیون در ۲۶۸ محله برای ۱۲ ماه از سال ۱۳۹۴.

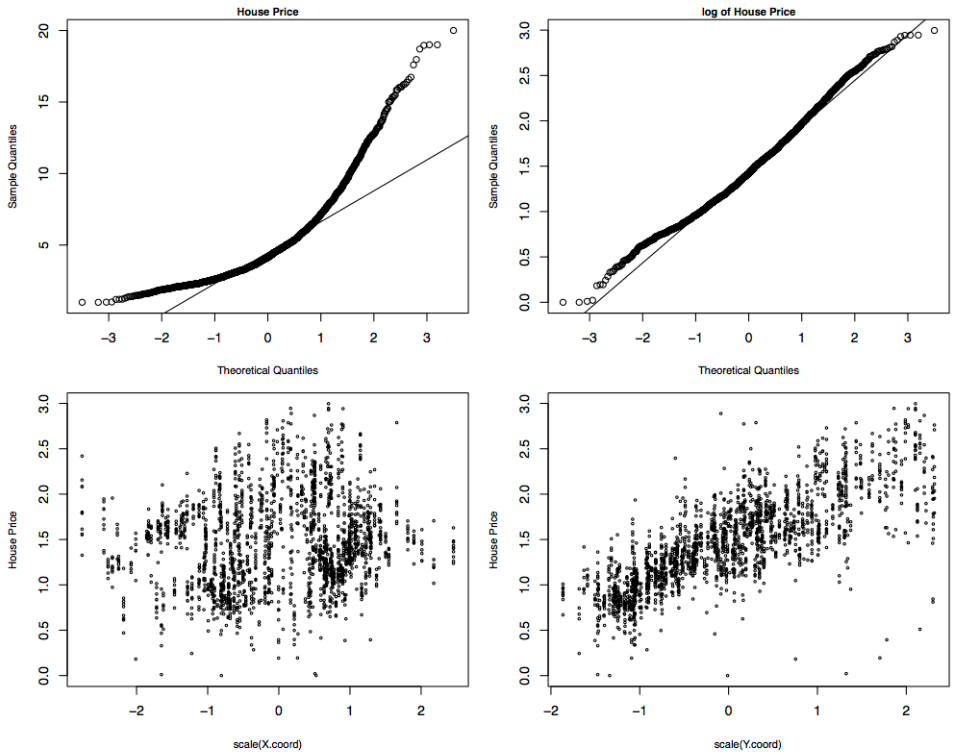
دهید $X_{\ell t}$ ماتریس $n \times p$ از متغیرهای که شامل عرض از مبدأ است. گلفند (۲۰۱۲) مدل‌های فضایی ـ زمانی بیزی را به‌صورت سلسله‌مراتبی در سه مرحله بیان کرد: (۱) مشاهدات به شرط یک فرآیند تصادفی و پارامترها $(z|\boldsymbol{u},\boldsymbol{\theta})$، (۲) فرآیند تصادفی به شرط پارامتر $(\boldsymbol{u}|\boldsymbol{\theta})$ و (۳) پارامترها $(\boldsymbol{\theta})$. در مرحله اول مدل به‌صورت

$$\boldsymbol{z}_{\ell t} = \boldsymbol{u}_{\ell t} + \boldsymbol{\varepsilon}_{\ell t}, \quad \ell = ۱, \ldots, r, \ t = ۱, \ldots, T,$$

در نظر گرفته می‌شود، که در آن $\boldsymbol{\varepsilon}_{\ell t} = (\varepsilon_\ell(s_۱,t), \ldots, \varepsilon_\ell(s_n,t))^\top$ بردار خطاهای مدل با توزیع $N(۰, \sigma_\varepsilon^۲ I_n)$ است و $\sigma_\varepsilon^۲$ اثرقطعه‌ای یا واریانس خطای تصادفی است. $\boldsymbol{u}_{\ell t} = (u_\ell(s_۱,t), \ldots, u_\ell(s_n,t))^\top$ 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 $\boldsymbol{\beta} = (\beta_۱, \beta_۲, \ldots, \beta_p)^\top$ بردار پارامترهای ضرایب رگرسیونی، $\boldsymbol{\eta}_{\ell t} = (\eta_\ell(s_۱,t), \ldots, \eta_\ell(s_n,t))^\top$ اثر تصادفی فضایی ـ زمانی مدل است که دارای توزیع $N(۰, \Sigma_n)$ و مستقل از زمان می‌باشد، که در آن $\Sigma_\eta = \sigma_\eta^۲ S_\eta$، $\sigma_\eta^۲$ پارامتر واریانس فضایی و S_η ماتریس همبستگی فضایی است. در این‌جا از تابع همبستگی ماترن (ماترن ، ۱۹۸۶) به‌صورت

$$K(s_i, s_j; \varphi, \nu) = \frac{۱}{۲^{\nu-۱}\Gamma(\nu)}(۲\sqrt{\nu} \parallel s_i - s_j \parallel \varphi)^\nu K_\nu(۲\sqrt{\nu} \parallel s_i - s_j \parallel \varphi)$$

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 ، ۱۹۹۳). مدل سلسله‌مراتبی ذکر شده را مدل فرآیند گاوسی (GP) مستقل می‌گویند. فرض کنید $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\varepsilon^۲, \sigma_\eta^۲, \varphi, \nu)$ بردار پارامترهای مدل، $\boldsymbol{z}_{\ell t}^{obs}$ بردار داده‌های مشاهده شده و $\boldsymbol{z}_{\ell t}^*$ بردار داده‌های گمشده باشد. با در نظر گرفتن توزیع‌های پیشین معمول برای پارامترهای مدل ؛$\pi(\boldsymbol{\theta})$، لگاریتم توزیع پسین توأم



شکل ۳: (بالا چپ) نمودار احتمال نرمال قیمت مسکن، (بالا راست) نمودار احتمال نرمال لگاریتم قیمت مسکن، (پایین چپ) نمودار پراکنش داده‌ها در مقابل طول جغرافیایی و (پایین چپ) نمودار پراکنش داده‌ها در مقابل عرض جغرافیایی .

پارامترها و داده‌های گمشده برای مدل GP به‌صورت

$$\log\pi(\boldsymbol{\theta},\boldsymbol{u}_{\ell t},\boldsymbol{z}^{*}_{\ell t}|\boldsymbol{z}^{obs}_{\ell t}) \propto -\frac{N}{2}\log\sigma^{2}_{\varepsilon}-\frac{1}{2\sigma^{2}_{\varepsilon}}\sum_{\ell=1}^{r}\sum_{t=1}^{T_\ell}(\boldsymbol{z}_{\ell t}-\boldsymbol{u}_{\ell t})^{\top}(\boldsymbol{z}_{\ell t}-\boldsymbol{u}_{\ell t})-\frac{\sum_{\ell=1}^{r}T_\ell}{2}\log\mid\sigma^{2}_{\eta}S_{\eta}\mid$$
$$-\frac{1}{2\sigma^{2}_{\eta}}\sum_{\ell=1}^{r}\sum_{t=1}^{T_\ell}(\boldsymbol{u}_{\ell t}-\boldsymbol{X}_{\ell t}\boldsymbol{\beta})^{\top}S^{-1}_{\eta}(\boldsymbol{u}_{\ell t}-\boldsymbol{X}_{\ell t}\boldsymbol{\beta})+\log\pi(\boldsymbol{\theta})$$

به‌دست می‌آید. ساهو و باکار (۲۰۱۲) مدل سلسله‌مراتبی اتوررگرسیو[۵] (AR) به‌صورت

$$\boldsymbol{z}_{\ell t}=\boldsymbol{u}_{\ell t}+\boldsymbol{\varepsilon}_{\ell t},$$
$$\boldsymbol{u}_{\ell t}=\rho\boldsymbol{u}_{\ell,t-1}+X_{\ell t}\boldsymbol{\beta}+\boldsymbol{\eta}_{\ell t},$$

معرفی کردند، 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 $N(\boldsymbol{\mu}_{\ell},\sigma^{2}_{\ell}S_{0})$ 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 $\boldsymbol{\theta}=(\boldsymbol{\beta},\rho,\sigma^{2}_{\varepsilon},\sigma^{2}_{\eta}\varphi,\nu,\boldsymbol{\mu}_{\ell},\sigma^{2}_{\ell})$ می‌باشند. با در نظر گرفتن پیشین معمول $\pi(\boldsymbol{\theta})$ برای پارامترها

[۵] Autoregressive

و با لگاریتم گرفتن توزیع پسین توأم پارامترها و مقادیر گمشده به‌صورت

$$\begin{aligned}\log\pi(\boldsymbol{\theta},\boldsymbol{u}_{\ell t},\boldsymbol{z}^{*}_{\ell t}|\boldsymbol{z}^{obs}_{\ell t}) \propto & -\frac{N}{۲}\log\sigma^{۲}_{\varepsilon}-\frac{۱}{۲\sigma^{۲}_{\varepsilon}}\sum_{\ell=۱}^{r}\sum_{t=۱}^{T_\ell}(\boldsymbol{z}_{\ell t}-\boldsymbol{u}_{\ell t})^{\top}(\boldsymbol{z}_{\ell t}-\boldsymbol{u}_{\ell t})-\frac{۱}{۲}\log\mid\sigma^{۲}_{\ell}S_{\cdot}\mid\\ & -\frac{۱}{۲\sigma^{۲}_{\eta}}\sum_{\ell=۱}^{r}\sum_{t=۱}^{T_\ell}(\boldsymbol{u}_{\ell\cdot}-\boldsymbol{X}_{\ell t}\boldsymbol{\beta})^{\top}S^{-۱}_{\eta}(\boldsymbol{u}_{\ell\cdot}-\boldsymbol{X}_{\ell t}\boldsymbol{\beta})-\frac{\sum_{\ell=۱}^{r}T_\ell}{۲}\log\mid\sigma^{۲}_{\eta}S_{\eta}\mid\\ & -\frac{۱}{۲\sigma^{۲}_{\eta}}\sum_{\ell=۱}^{r}\sum_{t=۱}^{T_\ell}(\boldsymbol{u}_{\ell t}-\boldsymbol{\mu}_{\ell})^{\top}S^{-۱}_{\cdot}(\boldsymbol{u}_{\ell t}-\boldsymbol{\mu}_{\ell})+\log\pi(\boldsymbol{\theta}),\end{aligned}$$

به‌دست می‌آید. فینلی و همکاران (۲۰۱۲) مدل پویای فضایی ـ زمانی را به‌صورت

$$\begin{aligned}&z(s,t)=\boldsymbol{x}^{\top}(s,t)\boldsymbol{\beta}_t+u_t(s,t)+\varepsilon(s,t),\ \varepsilon(s,t)\sim N(۰,\tau^{۲}_{t}),\\&\boldsymbol{\beta}_t=\boldsymbol{\beta}_{t-۱}+\boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t\sim N(۰,\Sigma_{\eta}),\\&u(s,t)=u(s,t-۱)+w(s,t),\ w(s,t)\sim \mathrm{GP}(۰,C_t(۰,\theta_t)),\ t=۱,\ldots,T,\end{aligned}$$

ارائه کردند، که در آن $\mathrm{GP}(۰,\sigma^{۲}_{t}\rho_t(۰,\varphi_t))$ یک فرآیند گاوسی با تابع همبستگی فضایی $\rho_t(۰,\varphi_t)$ است و $\sigma^{۲}_{t}$ و φ_t پارامترهای تابع کوواریانس فضایی هستند. همچنین فرض می‌شود $u(s,۰)=۰$ و $\boldsymbol{\beta}_{۰}\sim N(\boldsymbol{m}_{۰},\Sigma_{۰})$ است.

به دلیل پیچیدگی توزیع پسین و عدم وجود شکل بسته و وجود تعداد زیادی متغیر پنهان و پارامتر معمولاً استفاده از رهیافت بیزی معمولی و نمونه‌های مونت کارلویی بسیار وقت‌گیر و گاهی اوقات همگرایی‌ها به راحتی امکان‌پذیر نیست. استفاده از روش بیزی تقریبی برای مدل‌های فضایی‌ـ زمانی، به نمونه‌گیری‌های وقت‌گیر مونت کارلویی نیاز نداشته و به‌طور قابل ملاحظه‌ای سرعت محاسبات را افزایش می‌دهد. در این رهیافت از تقریب INLA که توسط رو و همکاران (۲۰۰۹) معرفی شده است استفاده می‌شود، که محاسباتی سریع و تقریبی دقیق از توزیع‌های پسین را جایگزین شبیه‌سازی‌های سنگین الگوریتم‌های مونت کارلویی می‌کند. همچنین در راستای راحتی و کوتاه کردن محاسبات، بردار متغیرهای پنهان در این مدل‌ها به‌صورت یک میدان تصادفی مارکوفی گاوسی درنظر گرفته می‌شود. در بخش بعد به تعریف میدان تصادفی مارکوفی گاوسی و نحوه تولید این میدان با رهیافت SPDE پرداخته می‌شود.

۴ میدان تصادفی مارکوفی گاوسی

قبل از تعریف میدان تصادفی مارکوفی گاوسی نیاز به تعریف گراف است. مجموعه‌ای از رأس‌ها که توسط تعدادی یال به هم وصل شده‌اند گراف نامیده می‌شود. به عبارت دیگر یک گراف به صورت زوج $\Im=(\nu,\varepsilon)$ تعریف می‌شود، که در آن ν یک مجموعهٔ متناهی ناتهی از عناصری به نام رئوس است و ε مجموعه‌ای متناهی از زوج‌های نامرتب و نامساوی از عناصر متمایز ν به‌صورت $\varepsilon\subseteq\{\{i,j\}:i,j\in\nu,i\neq j\}$ موسوم به یال‌ها است، به‌طوری که رأس‌ها را به هم وصل می‌کنند. در صورتی که رأس‌های یک گراف با مجموعه‌ای از نمادها مانند $\nu=\{۱,\cdots,n\}$ مشخص شود، گراف نشان‌دار نامیده می‌شود.

فرض کنید بردار تصادفی $\boldsymbol{u}=(u_۱,\cdots,u_n)^{\top}$ دارای توزیع نرمال چندمتغیره با بردار میانگین $\boldsymbol{\mu}$ و ماتریس کوواریانس معین مثبت $Q^{-۱}$ باشد، یعنی

$$\pi(\boldsymbol{u})=(۲\pi)^{-\frac{n}{۲}}\mid Q\mid^{\frac{۱}{۲}}\exp(-\frac{۱}{۲}(\boldsymbol{u}-\boldsymbol{\mu})^{\top}Q(\boldsymbol{u}-\boldsymbol{\mu}),$$

و $\Im = (\nu, \varepsilon)$ یک گراف باشد، که در آن ν مجموعه‌ای متناهی و غیرتهی از رئوس است به‌طوری که $\nu = \{1, \cdots, n\}$ یعنی $\Im$ یک گراف نشان‌دار است و ε شامل همه زوج‌های $\{i, j\}$است که رأس‌های i و j هیچ یال مشترکی نداشته باشند، اگر و تنها اگر $u_i \perp u_j \mid \boldsymbol{u}_{-ij}$، یعنی دو مولفه u_i و u_j به شرط بردار $\boldsymbol{x}$ به جز مولفه i و j آن مستقل باشند، در این صورت $\boldsymbol{u}$ یک میدان تصادفی مارکوفی گاوسی GMRF نسبت به گراف $\Im$ نامیده می‌شود (رو و هلد ، ۲۰۰۶).

درنظر گرفتن $\boldsymbol{u}$ به صورت یک GMRF منجر به این می‌شود که توزیع شرطی کامل هر مولفه از $\boldsymbol{u}$ به‌صورت توزیع شرطی آن مولفه به شرط سایر مولفه‌ها شود، یعنی $\pi(u_i|\boldsymbol{u}_{-i})$ که در آن $\boldsymbol{u}_{-i}$ بردار $\boldsymbol{u}$ به جز مولفه u_i است و از خاصیت مارکوفی توزیع شرطی کامل x_i، $i = ۱, \cdots, n$ تنها به تعداد کمی از مولفه های $\boldsymbol{u}$ بستگی دارد. اگر این مجموعه از مولفه‌ها را با δ_i مشخص شود که تشکیل‌دهنده مجموعه‌ای از همسایگی های مولفه ی iام است، آنگاه $\pi(x_i|\boldsymbol{u}_{-i}) = \pi(u_i|\boldsymbol{u}_{\delta x_i})$، به‌طوری که فرض می‌شود جملات x_i و $\boldsymbol{x}_{-\{i,\delta_i\}}$ مستقل هستند. برای هر $i = ۱, \cdots, n$ این رابطهٔ استقلال شرطی می‌تواند به‌صورت $x_i \perp \boldsymbol{x}_{-\{i,\delta_i\}} | \boldsymbol{x}_{\delta_i}$ نوشته شود. در یک SPDE به‌وسیله‌ی یک ماتریس دقت که ساختار تُنُک دارد، تعریف می‌شود و استفاده از روش‌های عددی محاسباتی موثر را امکان‌پذیر می‌سازد و دلیل اصلی برای استفاده از یک GMRF در مدل‌های پیچیده سلسله‌مراتبی فضایی این است که تابع کوواریانس فضایی_ زمانی و ماتریس کوواریانس غیرتُنُک مربوط به میدان تصادفی گاوسی، به‌ترتیب توسط یک ساختار همسایگی و یک ماتریس دقت پراکنده جایگزین می‌شوند.

۵ رهیافت SPDE برای پیدا کردن یک GMRF

فرض کنید $U(s) = \{U(s), s \in D \subseteq R^۲\}$ یک میدان تصادفی گاوسی مانای مرتبه دوم و همسانگرد با تابع کوواریانس ماترن با پارامتر هموار ν و پارامتر مقیاس κ باشد. همچنین فرض کنید تحققی از فرایند $U(s_i)$ در d موقعیت $s_۱, \cdots, s_d$مشاهده شده باشد. روش SPDE یک GMRF با همسایگی فضایی و ماتریس دقت تُنُک Q پیدا می‌کند که به بهترین وجه نشان‌دهنده میدان ماترن باشد. در زیربخش ۱.۵ نحوه تبدیل یک میدان ماترن به یک میدان تصادفی مارکوفی گاوسی بیان می‌شود . در این روش میدان تصادفی ماترن به‌صورت یک ترکیب خطی از توابع تکه‌ای خطی که بر یک مثلث‌بندی از دامنهٔ D تعریف شده است، بیان می‌شود، دامنه D به مجموعه‌ای از مثلث‌ها تقسیم‌بندی می‌شود که در بیشتر از یک یال مشترک یا زاویه مشترک متقاطع نیستند. رئوس مثلث‌بندی اولیه در موقعیت‌های $s_۱, \cdots, s_d$ قرار می‌گیرد و سپس رئوس دیگر به منظور رسیدن به یک مثلث‌بندی مناسب برای پیش‌بینی فضایی اهداف اضافه می‌شوند.

۱.۵ معادلات دیفرانسیل جزئی

ویتل (۱۹۶۳) به مطالعه معادلات دیفرانسیل جزئی پرداخت و ثابت کرد که میدان‌های تصادفی ماترن تنها جواب‌های پایدار برای یک معادله دیفرانسیل جزئی هستند. یک حل پایدار[۶] برای معادله دیفرانسیل جزئی تصادفی

$$(\kappa^۲ - \nabla^۲)^{\frac{\alpha}{۲}} x(\boldsymbol{s}) = W(\boldsymbol{s}),\ \boldsymbol{s} \in \boldsymbol{R}^d,\ \alpha = \nu + \frac{d}{۲},\ \kappa > ۰,\ \nu > ۰ \qquad (۱.۵)$$

[۶] Stationary solution

است، که در آن W میدان تصادفی فضایی نوفه سفید و ∇^2 عملگر دیفرانسیلی مرتبه دوم یا همان عملگر لاپلاس است. $(\kappa^2 - \nabla^2)^{\frac{\alpha}{2}}$ یک عملگر شبه- دیفرانسیل[۷] می‌باشد، به‌طوری که با به‌کار بردن تعریف تبدیل فوریه یک عملگر لاپلاس کسری[۸] به‌صورت

$$\{F(\kappa^۲ - \nabla^۲)^{\frac{\alpha}{۲}}\phi\}(\boldsymbol{K}) = (\kappa^۲ + \|\boldsymbol{K}\|^۲)^{\frac{\alpha}{۲}}(F\phi)(\boldsymbol{K})$$

درنظر گرفته می‌شود، که در آن ϕ تابعی در R^d است. لیندگرن و همکاران (۲۰۱۱) برای ساختن یک میدان تصادفی مارکوفی گاوسی که بیانگر یک میدان تصادفی ماترن روی یک شبکه مثلث‌بندی باشد، در مرحله اول یک فرم ضعیف تصادفی[۹] معادله دیفرانسیل جزئی (۱.۵) را درنظر گرفتند. با تعریف ضرب داخلی به‌صورت $\langle f, g\rangle = \int f(\boldsymbol{u})g(\boldsymbol{u})d\boldsymbol{u}$، برای هر مجموعه متناهی مناسب از توابع تست[۱۰] $\{\phi_j(\boldsymbol{s}), j = ۱, \cdots, m\}$، با داشتن شرط

$$\{\langle\phi_j, (\kappa^۲ - \nabla^۲)^{\frac{\alpha}{۲}}x\rangle, j = ۱, \cdots, n\} \stackrel{d}{=} \{\langle\phi_j, W\rangle, j = ۱, \cdots, n\}$$

حل ضعیف تصادفی به‌دست آوردند، که در آن d نشان‌دهنده هم‌توزیعی است. در مرحله بعد برای نمایش المان محدود[۱۱] حل معادله دیفرانسیل جزئی، یک ترکیب خطی از توابع پایه ای $\psi_l(s)$ به شکل

$$x(s) = \sum_{l=۱}^{n} \psi_l(s)\omega_l \qquad (۲.۵)$$

ایجاد نمودند. روش المان محدود روشی عددی برای حل تقریبی معادلات دیفرانسیل جزئی است و مثلث‌بندی یک راه‌حل معمول برای حل معادلات دیفرانسیل جزئی با استفاده از روش المان محدود می‌باشد، به طوری که دقت جواب‌ها بستگی به نحوه مثلث‌بندی دارد. در (۲.۵) n، تعداد کل رئوس، ω_l وزن هایی هستند که دارای توزیع گاوسی هستند و توابع $\psi_l(s)$ خطی تکه‌ای روی هر مثلث می‌باشند، به‌طوری که $\psi_l(s)$ در رأس l برابر یک و در سایر رئوس صفر است.

معمولاً در مطالعات فضایی یک ناحیه فضایی مثل $D \subseteq \mathbb{R}^d$ یعنی $d = ۲$ مورد نظر است. دامنه D به مجموعه‌ای از مثلث‌ها تقسیم‌بندی می‌شود که در بیشتر از یک یال مشترک یا زاویه مشترک متقاطع نیستند. در عمل رئوس مثلث‌بندی اولیه در موقعیت‌های مشاهده شده قرار می‌گیرد و سپس رئوس دیگر به منظور رسیدن به یک مثلث‌بندی مناسب برای پیش‌گویی فضایی اضافه می‌شوند. مثلث‌بندی دلونی[۱۲] نوعی مثلث‌بندی است که در آن دایره محیطی هر مثلث شامل هیچ رأسی از رئوس مثلث‌های دیگر نباشد (شرط دایره محیطی) و واضح است که این نوع مثلث‌بندی یکتا خواهد بود. تاکنون الگوریتم‌های مختلفی جهت بهبود زمانی ایجاد مثلث‌بندی و همچنین مثلث‌بندی دلونی مطرح شده است. این الگوریتم‌ها دارای پیچیدگی زمانی متفاوتی هستند. در این مقاله برای مثلث‌بندی از بسته INLA استفاده شده است، که در این بسته مثلث‌بندی دلونی و راه‌حل هیلی و دایلن (۲۰۰۶) به‌کار گرفته شده است. برای مطالعه بیش‌تر به هیلی و دایلن (۲۰۰۶)، دی برگ و همکاران (۲۰۰۸) و لیندگرن و همکاران (۲۰۱۱) مراجعه شود.

در بسته INLA برای $d = ۲$ معادله دیفرانسیل جزئی (۱.۵) به‌صورت عمومی

$$(\kappa^۲(\boldsymbol{s}) - \nabla^۲)^{\frac{\alpha}{۲}}(\tau(\boldsymbol{s})x(\boldsymbol{s})) = W(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{R}^۲,\ \alpha = \nu + ۱,\ \nu > ۰$$

[۷] Pseudo-differential operator
[۸] Fractional Laplacian
[۹] Stochastic weak formulation
[۱۰] Test functions
[۱۱] Finite element method
[۱۲] Delaunay triangulation

درنظر گرفته می‌شود که میدان‌های ناایستا را شامل شود، یعنی واریانس $W(\boldsymbol{s})$ و پارامتر κ ثابت نباشند. برای سادگی واریانس $W(\boldsymbol{s})$ را ثابت فرض می‌کنند و یک پارامتر مقیاس دیگر $\tau(\boldsymbol{s})$ برای $x(\boldsymbol{s})$ درنظر می‌گیرند، به‌طوری که اگر یکی یا هر دو پارامتر مقیاس τ و κ ثابت نباشند، میدان ناایستا خواهد بود و در نهایت از رابطه $\sigma_{\omega}^{۲}=\frac{۱}{۴\pi\kappa^{۲}\tau^{۲}}$ می‌شود.

لیندگرن و همکاران (۲۰۱۱) نشان دادند وزن‌های گاوسی $\{\omega_l\}$در (۲.۵) یک ساختار مارکوفی دارند و این وزن‌ها میدان تصادفی گاوسی ماترن را به یک میدان تصادفی مارکوفی گاوسی تبدیل می‌کنند. همچنین لیندگرن و همکاران (۲۰۱۱) سه ماتریس $n\times n$ به‌صورت

$$C_{ij}=\langle\psi_i,\psi_j\rangle, G_{ij}=\langle\nabla\psi_i,\nabla\psi_j\rangle, (\boldsymbol{K}_{\kappa^{۲}})_{ij}=\kappa^{۲}C_{ij}+G_{ij}$$

درنظر گرفتند و نشان دادند که ماتریس دقت Q میدان تصادفی مارکوفی گاوسی $\boldsymbol{\omega}\sim N(۰,Q^{-۱})$ تابعی از κ^2 برای هر $\alpha=\nu+۱$ و $\nu=۰,۱,\cdots$می‌باشد. این تبدیل یک نگاشت صریح و روشن از پارامترهای تابع کوواریانس میدان تصادفی گاوسی (κ,ν) به مولفه‌های ماتریس دقت Q میدان تصادفی مارکوفی گاوسی $\boldsymbol{\omega}$ با یک هزینه محاسباتی از مرتبه $O(n)$ برای هر مثلث‌بندی را ایجاد می‌کند. لیندگرن و همکاران (۲۰۱۱) برای $\alpha=۲$ و $\phi_\kappa=\psi_\kappa$ نشان داد ماتریس دقت میدان تصادفی مارکوفی گاوسی ω_l، تابعی از $\kappa^{۲}$ به‌صورت $K_{\kappa^{۲}}C^{-۱}K_{\kappa^{۲}}$ می‌شود. ماتریس‌های C و G به راحتی قابل محاسبه هستند و ماتریس $C^{-۱}$یک ماتریس تُنُک است. بعد از اینکه میدان تصادفی گاوسی، به وسیله روش SPDE به یک GMRF تبدیل شد. مدل فضایی‌ـزمانی بازنویسی می شود و پارامترهای مدل به وسیله رهیافت بیزی تقریبی رو و همکاران (۲۰۰۹) برآورد می شود. برای جزییات بیشتر روش INLA به رو و همکاران (۲۰۰۹) مراجعه شود. این رهیافت را در ادامه روی داده های قیمت ملک فروخته شده صورت گرفته است.

۶ تحلیل داده‌های قیمت ملک

سه مدل سلسله‌مراتبی بیزی GP، AR و DYN با در نظر گرفتن پیشین‌های معمول برای پارامترهای مدل ساهو و باکار (۲۰۱۲)؛ گلفند (۲۰۰۵) روی داده‌های قیمت ملک برازش شد. با توجه به تحلیل اکتشافی صورت گرفته در بخش ۲، طول و عرض جغرافیایی به‌عنوان متغیرهای کمکی و از لگاریتم قیمت مسکن به‌عنوان متغیر پاسخ استفاده شده است. نتایج برازش مدل برای ۱۰۰۰۰ تکرار با مقادیر سوخته ۵۰۰۰ در جدول ۱ خلاصه شده‌اند.

با توجه به نتایج جدول ۱ متغیر کمکی طول جغرافیایی تاثیر معنی‌داری در سطح ۵ درصد روی قیمت مسکن ندارد، اما عرض جغرافیایی معنی‌دار است. با توجه به ملاک مجذور میانگین توان دوم خطای پیش‌بینی و زمان اجرای مدل AR سلسله‌مراتبی برای داده‌های قیمت مسکن مناسب‌تر است. پارامترهای مدل DYN به زمان وابسته است و به همین دلیل در جدول ۱ فقط برای ماه فروردین ارائه شده است و این باعث طولانی شدن زمان محاسبات می‌شود. پیشگویی رفتار قیمت مسکن برای ۳ محله شهر تهران براساس سه مدل در شکل ۴ ارائه شده است که نشان می‌دهد مدل AR حتی با وجود داده گمشده نسبتا زیاد بهتر از دو مدل دیگر برای این داده‌ها عمل می‌کند. یکی از مشکلات مدل DYN این است که برای پیش‌بینی در زمان‌های آینده که در هیچ موقعیتی مشاهده وجود ندارد عملکرد بدی را نشان می‌دهد که در شکل ۴ به‌وضوح دیده می شود.

در ادامه به وسیله رهیافت SPDE، میدان گاوسی را به یک میدان تصادفی مارکوفی تبدیل می کنیم و با روش بیزی تقریبی رو و همکاران (۲۰۰۹) سه مدل GP، AR مرتبه یک و قدم زدن تصادفی مرتبه اول RW1 را به داده ها برازش می دهیم. مثلث

جدول ۱: نتایج برازش بیزی سه مدل فضایی زمانی روی داده‌های قیمت مسکن.

	مدل GP		مدل AR		مدل DYN
پارامتر	برآورد (انحراف معیار) [فاصله اطمینان ٪۹۵]	پارامتر	برآورد (انحراف معیار) [فاصله اطمینان ٪۹۵]	پارامتر	برآورد (انحراف معیار) [فاصله اطمینان ٪۹۵]
β_0	۱/۴۵۰۰ (۰/۰۰۵۸) [۱/۴۳۸۵, ۱/۴۶۱۳]	β_{0t}	۰/۸۴۳۳ (۰/۰۲۴۱) [۰/۷۹۵۵, ۰/۸۹۰۶]	β_{0t}	۱/۴۳۴۴ (۰/۳۴۳۹) [۰/۵۸۵۳, ۲/۰۱۱۳]
β_1	۰/۲۶۱۱ (۰/۰۰۵۸) [۰/۲۴۹۵, ۰/۲۷۲۶]	β_{1t}	۰/۱۵۱۱ (۰/۰۰۶۷) [۰/۱۳۸۳, ۰/۱۶۴۶]	β_{1t}	۰/۲۳۶۱ (۰/۱۹۷۴) [−۰/۱۶۵۴, ۰/۶۱۰۴]
β_2	۰/۰۰۳۹ (۰/۰۰۵۷) [−۰/۰۰۷۳, ۰/۰۱۵۱]	β_{2t}	۰/۰۰۲۱ (۰/۰۰۵۳) [−۰/۰۰۸۴, ۰/۰۱۲۶]	β_{2t}	۰/۱۳۷۴ (۰/۱۸۹۰) [−۰/۱۹۵۲, ۰/۵۳۲۷]
σ_ϵ^2	۰/۱۱۶۲ (۰/۰۰۶۷) [۰/۱۰۳۴, ۰/۱۳۰۱]	τ_t^2	۰/۰۸۹۲ (۰/۰۰۴۸) [۰/۰۸۰۲, ۰/۰۹۸۹]	τ_t^2	۰/۳۸۴۶ (۰/۰۴۲۷) [۰/۳۱۳۱, ۰/۴۸۰۴]
σ_η^2	۰/۱۰۲۴ (۰/۰۰۲۶) [۰/۰۹۷۴, ۰/۱۰۷۴]	σ_t^2	۰/۰۸۴۳ (۰/۰۰۲۱) [۰/۰۸۰۲, ۰/۰۸۸۶]	σ_t^2	۰/۶۳۸۹ (۰/۰۸۰۶) [۰/۴۹۴۵, ۰/۸۱۲۲]
ϕ	۱/۱۲۴۷ (۰/۸۳۷۸) [۰/۱۱۰۵, ۳/۶۳۶۴]	ϕ_t	۱/۹۱۹۵ (۱/۲۳۶۷) [۰/۳۸۸۵, ۵/۱۴۶۶]	ϕ_t	۰/۰۰۰۱ ($۳/۷ \times ۱۰^{-۶}$) [۰/۰۰۰۱۱, ۰/۰۰۰۱۲]
ρ	–	–	۰/۴۲۰۲ (۰/۰۱۶۲) [۰/۳۸۸۶, ۰/۴۵۲۱]	–	–
زمان اجرا	۴۰ دقیقه و ۳۸ ثانیه		۵۹ دقیقه و ۱۷ ثانیه		یک ساعت و ۴۷ دقیقه
RMSE	۱/۲۵۲۸		۱/۱۰۱۷		۰/۹۸۹۷

بندی ۲۶۸ محله تهران که در راس ها قرار دارند، در شکل ۵ رسم شده است. لازم به ذکر است که تحلیل SPDE-INLA به وسیله بسته نرم افزاری INLA نسخه ۲۰۱۷ تحت نرم افزار R انجام شده است. نتایج مدل های مختلف رهیافت SPDE-INLA روی داده های قیمت مسکن در جدول ۲ ارایه شده است. همانطور که ملاحظه می شود مدل GP در رهیافت SPDE-INLA دارای MSE کمتری است، حتی نسبت به رهیافت بیز معمولی برازش بهتری ارایه می دهد. از نظر زمان اجرا نیز این مدل زمان اجرای کمتری دارد. در مقایسه با بیز معمولی زمان محاسبات در رهیافت SPDE-INLA خیلی پایین تر است. نتایج جدول ۲ براساس $\alpha = ۱/۵$ است که همان تابع همبستگی نمایی در معادله ماترن است. در این جا نیز یک تحلیل روی مقادیر مختلف α صورت گرفته است که خلاصه نتایج در جدول ۲ آورده شده است. نتایج نشان می دهد که هر چه α کمتر می شود برازش مدل بهتر می گردد که یک رابطه عکس را به نمایش می گذارد.

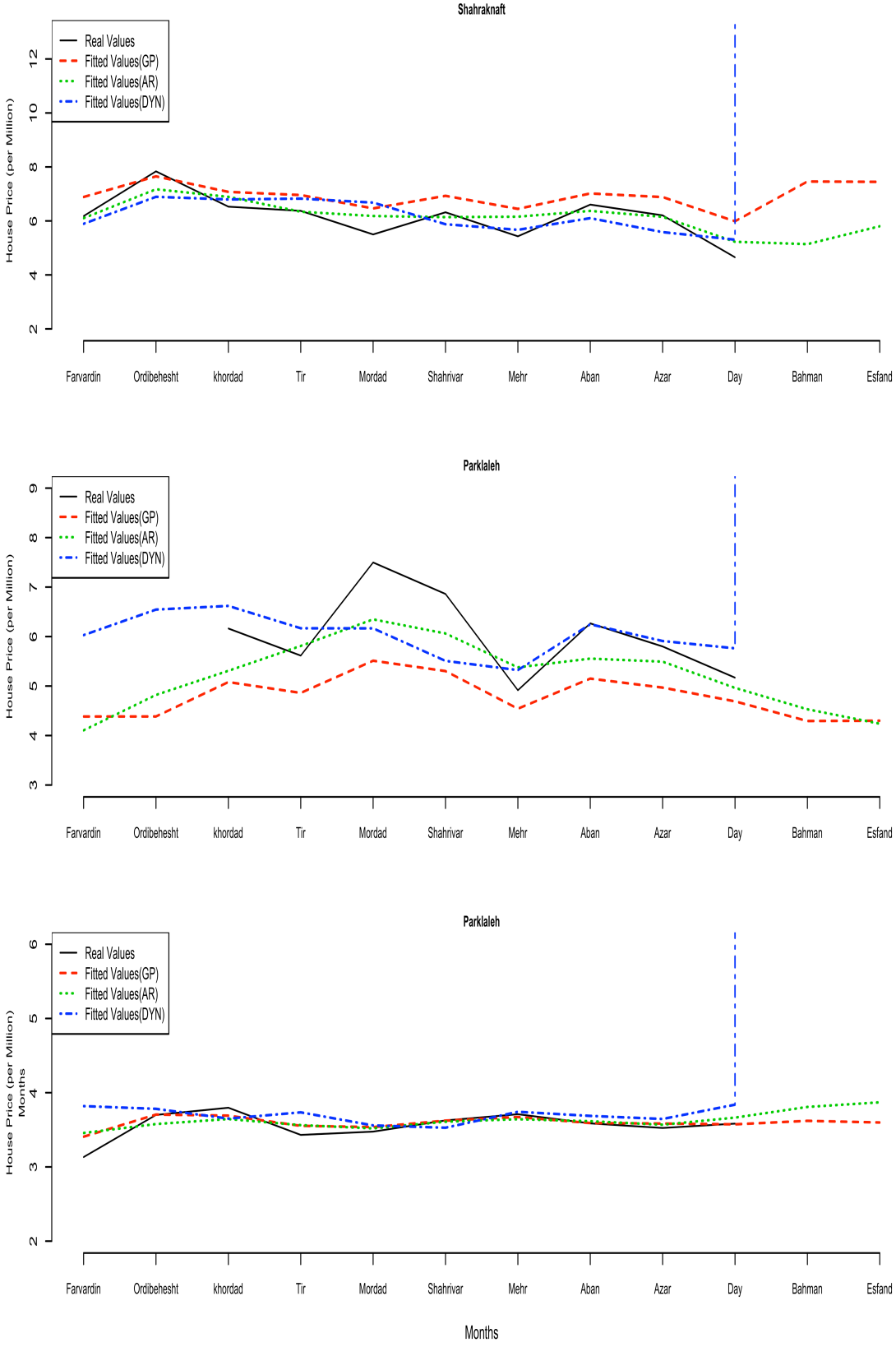
جدول ۲: نتایج MSE و زمان اجرا در رهیافت SPDE-INLA برای $\alpha = ۱/۵$.

مدل	GP	AR1	RW1
MSE	۰/۲۸۲۷	۱/۲۹۸۲	۱/۱۱۸۴
زمان اجرا (ثانیه)	۲۲/۰۲	۳۰۸/۱۵	۱۷۰۶/۱۵

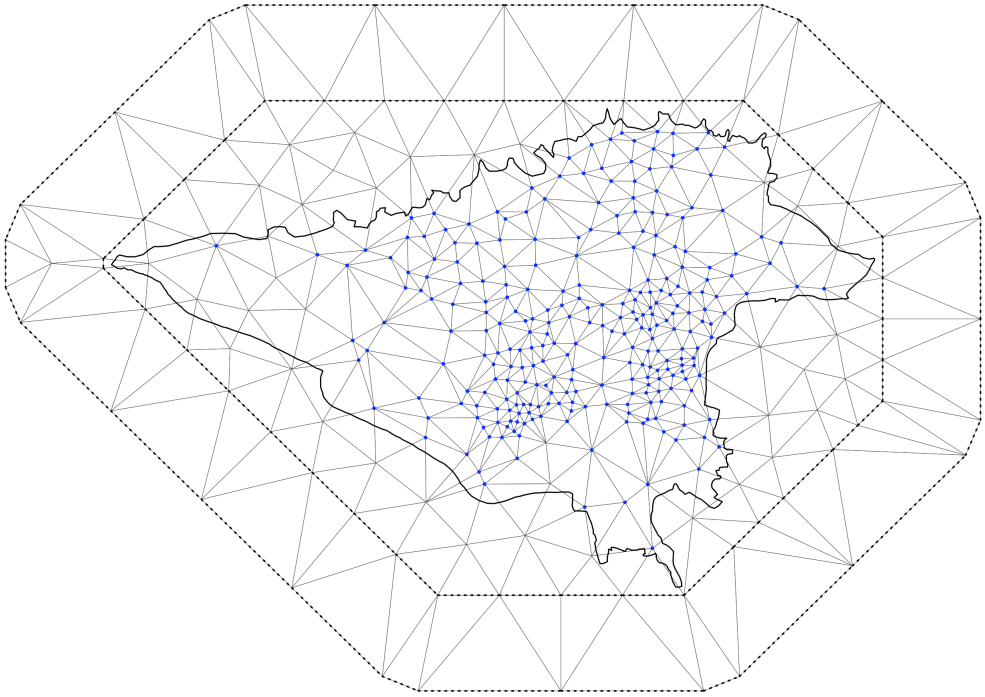
جدول ۳: تغییرات MSE مدل GP برای مقادیر مختلف α در رهیافت SPDE-INLA .

α	۲	۱/۵	۱	۰/۵	۰/۲	۰/۱	۰/۰۵
MSE	۰/۷۵۵۳	۰/۶۱۵۴	۰/۲۸۲۷	۰/۲۷۸۷	۰/۲۶۸۲	۰/۲۷۷۹	۰/۲۹۱۲
زمان اجرا (ثانیه)	۷۵/۵۶	۵۳/۵۷	۱۶/۸۵	۲۰/۷۸	۴۴/۵۴	۵۱/۴۷	۱۰۱/۳۰

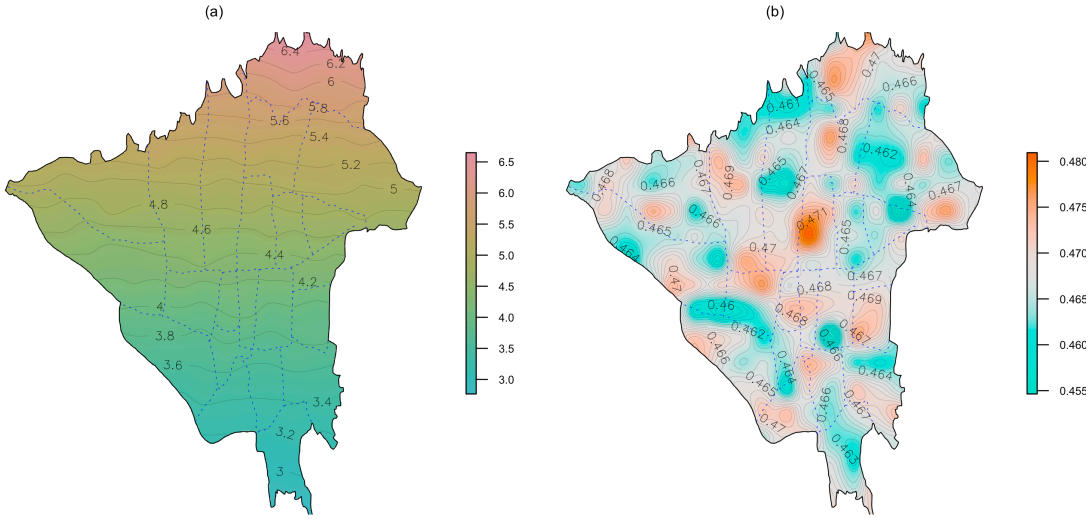
در انتها نقشه پیشگویی و خطای پیشگویی قیمت مسکن براساس مدل AR برای ماه‌های بهمن و اسفند در شکل ۶ ارائه شده است، که نشان می‌دهد تغییر چندانی در این دو ماه دیده نمی‌شود و حتی با مقایسه با ماه فروردین این تغییر خیلی محسوس نیست.



شکل ۴: برازش سه مدل برای سه محله متفاوت شهر تهران.



شکل ۵: مثلث بندی ۲۶۸ محله تهران.



شکل ۶: (چپ) نقشه پیشگویی و (راست) انحراف معیار پیشگویی قیمت ملک در شهر تهران.

بحث و نتیجه‌گیری

در این مقاله سه مدل فضایی ـ زمانی برای داده‌های حجیم قیمت ملک در شهر تهران به‌کار گرفته شد. با استفاده از روش بیزی سلسله‌مراتبی و بیزی تقریبی برآورد پارامترهای مدل ارائه گردید. سپس براساس ملاک RMSE در رهیافت بیزی معمولی مدل AR نسبت به مدل‌های دیگر برازش نسبتا مناسب‌تری به داده‌های قیمت ملک داشت هر چند مدل DYN به نظر RMSE کمتری نسبت به این مدل داشت اما از نظر زمان اجرا و پیش‌بینی در زمان‌های آینده که در هیچ موقعیتی مشاهده‌ای وجود ندارد از عملکرد خوبی برخوردار نبود و پارامترهای این مدل با افزایش بعد زمان زیادتر می‌شوند. برای داده‌های حجیم‌تر مثلا برای داده‌های قیمت ملک نوساز در سال ۱۳۹۳ که در حدود ۳۵۰ محله و به‌صورت روزانه ثبت شده است، پارامترهای مدل زیاد و مدل در عمل ناکارآمد خواهد بود. همچنین از رهیافت بیزی تقریبی با تبدیل میدان تصادفی گاوسی به یک GMRF با روش SPDE مدل های مختلف فضایی زمانی مورد ارزیابی قرار گرفت و به این نتیجه رسیدم که علاوه بر کاهش معنی داری در زمان محاسبات دقت برازش مدل نیز بهتر شد. در نهایت با استفاده از مدل AR نقشه پیشگویی قیمت مسکن در ماه اسفند برای کل ناحیه شهر تهران رسم شد. به‌عنوان پیشنهاد می‌توان رهیافت بیز تقریبی INLA که توسط رو و همکاران (۲۰۰۹) معرفی شده است را برای این داده‌ها بدون استفاده از تبدیل لگاریتمی در یک مدل چوله به کار گرفت و با مدل‌های دیگر مقایسه کرد.که می‌توان با توزیع‌های چوله مدل‌بندی کرد، البته با وجود زمان از پیچیدگی‌های خاصی برخوردار است.

مراجع

Banerjee S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society Series: B*, **70**, 825–848.

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

Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2013). Spatio-Temporal Modeling of Particulate Matter Concentration through the SPDE Approach, *Advances in Statistical Analysis*, **97**, 109-131.

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

Cressie N. A. C., and Wikle, C. K., (2011). *Statistics for Spatio-Temporal Data*. John Wiley and Sons, New York.

De Berg, M., Cheong, O., van Kreveld, M., and Overmars, M. (2008), *Computational Geometry: Algorithms and Applications* ,Springer-Verlag Berlin Heidelberg.

Eidsvik, J., Martino, ARTINO, S. and Rue, H. (2009), Approximate Bayesian Inference in Spatial Generalized Linear Mixed Models. *Scandinavian Journal of Statistics*, **36**: 1–22.

Finley, A. O., 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.

Lindgren, F., Rue, H., and Lindstrom, J. (2011), An explicit link between Gaussian Fields and Gaussian Markov Random Fields: the Stochastic Partial Differential Equation Approach (with discussion)

Hartman L.,and Hössjer O., (2008), Fast kriging of large data sets with Gaussian Markov random fields. *Computational Statistics and Data Analysis*, **52**, 2331–2349.

Hjelle, O. and Daehlen, M. (2006), *Triangulations and Applications*,Springer.

Matérn, B., (1986), *Spatial Variation*. 2nd edition. Springer-Verlag, Berlin.

Hosseini, F., Eidsvik, J., and Mohammadzadeh, M. (2011), Approximate Bayesian Inference in Spatial GLMM with Skew Normal Latent Variables, *Computational Statistics and Data Analysis*, **55**, 1791-1806.

Hosseini, F., and Mohammadzadeh, M. (2011), Bayesian Prediction for Spatial GLMM's with Closed Skew Normal Latent Variables , *Australian and New Zealand Journal of Statistics*, **54**, 43-62.

Gholizadeh, K., Mohammadzadeh, M. and Ghayyomi, Z. (2013), Spatial Analysis of Structured Additive Regression and Modeling of Crime Data in Tehran City using Integrated Nested Laplace Approximation, *Journal of Statistical Science*, **7**, 103-124.

Rue H., and Held L., (2006), *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC: Boca Raton.

Rue, H., Martino, S. (2007), Approximate Bayesian Inference for Hierarchical Gaussian Markov Random fields models, *Journal of Statistical Planning and Inference*, **137**, 3177-3199.

Rue, H., Martino, S., and Chopin, N., (2009), Approximate Bayesian Inference for latent Gaussian models by using integrated Laplace approximation, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 319-392.

Sahu S. K., and Bakar K. S., (2011), A Comparison of Bayesian Models for Daily Ozone Concentration Levels.*Statistical Methodology*, **9**, 144-157.

Sahu S. K., and Bakar, K. S., (2012). Hierarchical Bayesian Autoregressive Models for Large Space Time Data with Applications to Ozone Concentration Modelling. *Applied Stochastic Models in Business and Industry*, **28**, 395-415.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2007). High-Resolution Space-Time Ozone Modeling for Assessing Trends. *Journal of the American Statistical Association*, **102**, 1221-1234.

Whittle, P. (1963), Stochastic Processes in Several Dimensions, *Bulletin of the International Statistical Institute*, **40**, 974-994.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1390

مدل‌بندی همزمان پارامترهای میانگین و دقت در مدل آمیخته خطی تعمیم‌یافته بتا فضایی

لیدا کلهری ندرآبادی*[۱]، محسن محمدزاده [۲]

[۱] پژوهشکده‌ی آمار

[۲] گروه آمار، دانشگاه تربیت مدرس

چکیده: مدل رگرسیون استاندارد برای مدل‌بندی متغیرهای پاسخ پیوسته مورد استفاده قرار می‌گیرد. در عمل با مواردی مواجه می‌شویم که متغیر پاسخ در بازه (۰, ۱) قرار می‌گیرد. مدل رگرسیون بتا برای مدل بندی این‌گونه مشاهدات معرفی شده است. در این مقاله مدل‌بندی همزمان پارامترهای میانگین و دقت توزیع بتا با در نظر گرفتن ساختار همبستگی فضایی در مدل میانگین انجام می‌شود و برآورد پارامترها با رهیافت بیزی به‌دست می‌آیند. همچنین عملکرد این مدل با مدل گاوسی مورد مقایسه قرار می‌گیرد. کاربردی از این مدل برای تحلیل داده‌های آمارگیری هزینه و درآمد در شهر تهران ارائه می‌شود.

واژه‌های کلیدی: رگرسیون بتا، پارامتر دقت، رهیافت بیزی، آمارگیری هزینه و درآمد

کد موضوع‌بندی ریاضی (۲۰۱۰): 62F15، 62M30.

۱ مقدمه

در بسیاری از مطالعات کاربردی با متغیرهای پاسخی مواجه می‌شویم که مقادیر آن‌ها در یک بازه محدود قرار می‌گیرد که در این مورد می‌توان به نرخ‌ها و نسبت‌ها اشاره کرد. از آنجایی‌که یکی از اهداف استفاده از مدل‌های رگرسیونی ساختن الگوهایی مفید برای پیشگویی مناسب است، وقتی حوزه مقادیر متغیر پاسخ بازه (۰, ۱) باشد، استفاده از این مدل‌ها ممکن است منجر به پیشگویی مقادیری خارج از این بازه شود. پائولینو (۲۰۰۱) با توجه به تکیه‌گاه توزیع بتا پیشنهاد کرد برای مدل‌بندی متغیرهای پاسخ از جنس نسبت، فرض شود که متغیر پاسخ از توزیع بتا پیروی می‌کند. فراری و کریباری (۲۰۰۴) با الهام گرفتن از اصول مدل‌های خطی تعمیم‌یافته مک‌کلاخ و نلدر (۱۹۸۹)، توزیع بتای بازپارامتریده[۱] را برای مطالعه نسبت‌ها پیشنهاد کردند. آن‌ها پارامترهای

*ارایه‌دهنده مقاله: لیدا کلهری، kalhori@srtc.ac.ir

[۱] Reparametrized beta distribution

توزیع بتا را به‌گونه‌ای بازنویسی کردند که مدل رگرسیونی بر اساس میانگین متغیر پاسخ معین شود. مدل‌های خطی تعمیم‌یافته نلدر و ودربرن (۱۹۷۲) رده استانداردی از مدل‌ها است که با فرض همگنی[۲] واریانس پاسخ به داده‌ها برازش داده می‌شوند. چولگی[۳] و ناهمگنی متغیر پاسخ یکی از مشکلاتی است که گاهی در مدل‌بندی داده‌های واقعی با آن مواجه می‌شویم. پس از معرفی خانواده توزیع‌های نمایی دو پارامتری توسط دی و همکاران (۱۹۹۷) مدل‌بندی همزمان پارامترهای میانگین و واریانس مورد توجه قرار گرفت. از آنجا که توزیع بتا در بردارنده فرم‌های چوله و متقارن است و حوزه مقادیر آن بازه $(۰, ۱)$ است، اگر داده‌ها کران‌دار باشند، مدل‌بندی همزمان پارامترهای توزیع بتای بازپارامتریده می‌تواند راه حلی برای مواجه با این مشکل باشد. مدل‌بندی همزمان پارامترهای میانگین و دقت توزیع بتا با به‌کارگیری رهیافت بیزی و در نظر گرفتن توزیع نرمال به عنوان توزیع پیشین برای ضرایب رگرسیونی، توسط سپیدا (۲۰۰۱) و سپیدا و گامرمن (۲۰۰۵) انجام شد. توسعه این مدل‌ها با در نظر گرفتن ساختار رگرسیونی غیر خطی توسط سپیدا و آچکار (۲۰۱۰) انجام شد. علاوه بر آن، مدل‌بندی همزمان پارامترهای توزیع‌های نرمال، گاما و بتا با در نظر گرفتن اثرهای تصادفی توسط زیمپریچ (۲۰۱۰)، فیگیورا و همکاران (۲۰۱۳) و سپیدا و همکاران (۲۰۱۴) با استفاده از رهیافت بیزی انجام شد. اگر همبستگی متغیر پاسخ ناشی از موقعیت قرار گرفتن واحدهای نمونه باشد، لازم است همبستگی فضایی آن‌ها در مدل‌بندی پارامترهای میانگین و دقت مد نظر قرار بگیرد. در بخش ۲ مدل رگرسیون بتا و مدل آمیخته خطی تعمیم‌یافته بتا فضایی با فرض ثابت بودن پارامتر دقت و همچنین در نظر گرفتن پارامتر دقت ساختارمند[۴] معرفی می‌شود و نحوه مدل‌بندی همزمان پارامترهای میانگین و دقت توزیع بتای بازپارامتریده با در نظر گرفتن اثر فضایی در مدل‌بندی میانگین ارائه می‌شود. در بخش ۳ برآورد پارامترهای مدل با رهیافت بیزی به‌دست می‌آیند و مدل‌های بتا و گاوسی برای مدل‌بندی داده‌های همبسته فضایی محدود در بازه $(۰, ۱)$ مورد مقایسه قرار می‌گیرند. در بخش ۴ نحوه کاربست مدل برای تحلیل سهم درآمد خانوار از یارانه در شهر تهران که اطلاعات آن از طرح آمارگیری هزینه و درآمد به‌دست آمده است، ارائه می‌شود. در پایان بحث و نتیجه‌گیری ارائه می‌شود.

۲ مدل آمیخته خطی تعمیم‌یافته بتا فضایی

فراری و کریباری (۲۰۰۴) توزیع بتای بازپازامتریده را برای مطالعه نسبت‌ها پیشنهاد کردند. آن‌ها پارامترهای توزیع بتا را به‌گونه‌ای بازنویسی کردند که مدل رگرسیونی بر اساس میانگین متغیر پاسخ معین شود. بنابراین توزیع $Beta(a, b)$ را با فرض $\mu = \frac{a}{a+b}$ و $\phi = a + b$ به صورت

$$f(y, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱} \qquad ۰ < y < ۱ \qquad (۱.۲)$$

در نظر گرفتند، که در آن $۰ < \mu < ۱$ و $\phi > ۰$. برای این توزیع $E(Y) = \mu$ ، $\text{Var}(Y) = \frac{\mu(۱-\mu)}{۱+\phi}$، که در آن ϕ پارامتر دقت نامیده می‌شود. در توزیع بتای بازپارامتریده، وقتی پارامتر دقت ثابت باشد، شرط همگنی واریانس برقرار است و می‌توان مانند مدل‌های خطی تعمیم‌یافته عمل کرد. اما در حالاتی که پارامتر دقت متغیر باشد، شرط همگنی واریانس برقرار نیست و لازم است علاوه بر میانگین، با استفاده از توابع ربط مناسب مدلی برای پارامتر دقت نیز در نظر گرفت. این مدل می‌تواند تابعی از متغیرهای تبیینی باشد که لزوما با متغیرهای تبیینی مدل میانگین یکسان نیستند.

[۲] Homogenity
[۳] Skewness
[۴] Structured Precision Parameter

سپیدا و گامرمن (۲۰۰۰، ۲۰۰۵) و اسمیتسون و ورکولین (۲۰۰۶) مدل رگرسیون بتا را با مدل‌بندی همزمان میانگین و دقت توسعه دادند. اگر $Y_1 \ldots, Y_n$ نمونه‌ای تصادفی از توزیع بتا باشند به طوری که $Y_i \sim Beta(\mu_i\phi_i, (1-\mu_i)\phi_i)$، آنگاه مدل رگرسیون بتا به‌صورت

$$g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}, \quad h(\phi_i) = \boldsymbol{z}_i\boldsymbol{\delta}, \quad i = 1, \ldots, n \tag{۲.۲}$$

تعریف می‌شود، که در آن $\boldsymbol{x}_i = (x_{i0}, \ldots, x_{ip})$ و $\boldsymbol{z}_i = (z_{i0}, \ldots, z_{im})$ به ترتیب بردار متغیرهای تبیینی در مدل میانگین و دقت، $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$ و $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_m)^\top$ بردار ضرایب رگرسیونی، و $g(\cdot)$ و $h(\cdot)$ توابع ربط هستند. وقتی متغیرهای پاسخ همبسته فضایی باشند، می‌توان یک مؤلفه تصادفی در مدل‌بندی داده‌ها وارد کرد که ساختار همبستگی فضایی داده‌ها از طریق آن در مدل لحاظ شود (دیگل و همکاران، ۱۹۹۸). فرض کنید $\mathbf{y}(s) = (y(s_1), \ldots, y(s_n))^\top$ تحقق‌های متغیرهای پاسخ $Y(s) = (Y(s_1), \ldots, Y(s_n))^\top$ در n موقعیت $s_1, \ldots, s_n$ باشند. هدف مدل‌بندی متغیرهای تصادفی همبسته فضایی $Y_1, \ldots, Y_n$ از توزیع بتا است که همبستگی فضایی آن‌ها از طریق یک میدان تصادفی گاوسی[۵] (GRF) در نظر گرفته شود. فرض کنید $\boldsymbol{\tau}(s) = (\tau(s_1), \ldots, \tau(s_n))^\top$ یک میدان تصادفی مانای مرتبه دوم گاوسی با میانگین صفر باشد، و متغیرهای تصادفی $Y(s)$ به شرط مؤلفه تصادفی $\boldsymbol{\tau}(s)$ مستقل و دارای توزیع بتا باشند. به عبارت دیگر

$$[Y(s) \mid \boldsymbol{\tau}(s)] \sim Beta(\mu(s)\phi(s), (1-\mu(s))\phi(s))$$

که توزیع آن از جایگذاری $\mu(s)$ و $\phi(s)$ در (۱.۲) به‌دست می‌آید. بنابراین مدل آمیخته خطی تعمیم‌یافته بتا فضایی با پارامتر دقت ثابت به‌صورت

$$g(E(Y_i \mid \tau_i)) = \boldsymbol{x}_i\boldsymbol{\beta} + \tau_i \qquad i = 1, \ldots, n, \tag{۳.۲}$$

تعریف می‌شود، که در آن $g(\cdot)$ تابع پیوند یک به یک و مشتق‌پذیر با دامنه اعداد حقیقی، $\boldsymbol{x}_i = (x_{i0}, \ldots, x_{ip})$ بردار متغیرهای تبیینی، $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$ بردار ضرایب رگرسیونی، Y_i نمایشگر متغیر پاسخ فضایی $Y(s_i)$، و τ_i نمایشگر اثر تصادفی $\tau(s_i)$ است که ساختار همبستگی فضایی داده‌ها را در مدل وارد می‌کند. فرایند تصادفی $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^\top$ از توزیع نرمال چند متغیره به‌صورت $\boldsymbol{\tau}(s) \sim \mathrm{N_n}(\mathbf{0}, \Sigma_\tau)$ پیروی می‌کند، که در آن

$$(\Sigma_\tau)_{ij} = \mathrm{Cov}(\tau(s_i), \tau(s_j)) = \sigma^2\mathrm{Cor}(\tau(s_i), \tau(s_j)),$$

مؤلفه‌های Σ_τ تابعی از فاصله بین موقعیت‌ها هستند به‌گونه‌ای که تشریح کننده ساختار کوواریانس فضایی باشند. تابع پیوند لوجیت برای مدل‌بندی میانگین در نظر گرفته می‌شود. بنابراین (۳.۲) به‌صورت

$$\mathrm{logit}(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta} + \tau_i, \quad i = 1, \ldots, n. \tag{۴.۲}$$

بازنویسی می‌شود. در نتیجه مدل سلسله مراتبی به صورت

$$\begin{aligned} [Y_i \mid \boldsymbol{\beta}, \phi, \tau_i] &\sim Beta(\mu_i\phi, (1-\mu_i)\phi), \quad i = 1, \ldots, n, \\ [\boldsymbol{\tau} \mid \boldsymbol{\upsilon}] &\sim N_n(\mathbf{0}, \Sigma_\tau), \end{aligned} \tag{۵.۲}$$

[۵] Gaussian Random Field

تعریف می‌شود، که در آن $\boldsymbol{\upsilon}$ بردار پارامترهای مربوط به ساختار کوواریانس فضایی، Σ_τ، است. کلهری و محمدزاده (۲۰۱۶) برآورد پارامترهای یک مدل آمیخته خطی تعمیم‌یافته بتا فضایی با پارامتر دقت ثابت را با رهیافت بیزی به‌دست آوردند. برآورد بیزی پارامترهای مدل برای نمونه‌های با اندازه کوچک در مطالعه محمدزاده و کلهری (۲۰۱۷) به‌دست آورده شده است. اگر پارامتر دقت ثابت نباشد که معادل ناهمگن بودن واریانس است، مدل رگرسیون بتا فضایی به‌صورت

$$\text{logit}(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta} + \tau_i, \quad \log(\phi_i) = \boldsymbol{z}_i\boldsymbol{\delta}, \qquad i = ۱, \ldots, n, \tag{۶.۲}$$

بیان می‌شود، که در آن

$$\begin{aligned} [Y_i \mid (\boldsymbol{\beta}, \boldsymbol{\delta}, \tau_i)] &\sim Beta(\mu_i\phi_i, (۱-\mu_i)\phi_i), \\ [\boldsymbol{\tau} \mid \boldsymbol{\upsilon}] &\sim N_n(۰, \Sigma_\tau). \end{aligned} \tag{۷.۲}$$

برآورد پارامترهای مدل با در نظر گرفتن توزیع‌های پیشین مناسب برای آن‌ها با رهیافت بیزی انجام می‌شود. با در نظر گرفتن توزیع‌های پیشین دلخواه $\pi(\boldsymbol{\beta})$، $\pi(\boldsymbol{\delta})$، $\pi(\sigma^{۲}_{۰})$ و $\pi(\psi^{-۱})$ توزیع پسین توأم متناسب با

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\delta}, , \sigma^{۲}_{۰}, \psi^{-۱}, \boldsymbol{\tau}) \propto \prod_{i=۱}^{n} f(y_i \mid \tau_i, \boldsymbol{\beta}, \boldsymbol{\delta}) \prod_{i=۱}^{n} \pi(\tau_i \mid \boldsymbol{\beta}, \sigma^{۲}_{۰}, \psi^{-۱}) \\ \times \pi(\boldsymbol{\beta})\pi(\boldsymbol{\delta})\pi(\sigma^{۲}_{۰})\pi(\psi^{-۱}) \end{aligned}$$

به‌دست می‌آید که دارای فرم بسته نیست، بنابراین برآورد پارامترهای مدل با نمونه‌گیری از توزیع‌های شرطی کامل انجام می‌شود. روش نمونه‌گیری گیبز با استفاده از بسته R2WinBUGS (استوآرتز و همکاران، ۲۰۰۵) انجام می‌شود.

۳ برآورد بیزی پارامترهای مدل و ارزیابی آن‌ها

در این بخش مدل‌بندی همزمان پارامترهای میانگین و دقت توزیع بتا با درنظر گرفتن اثر فضایی در مدل میانگین انجام می‌شود. علاوه برآن با فرض آن که متغیر پاسخ در بازه (۱و۰) دارای توزیع نرمال باشد، مدل‌بندی همزمان پارامترهای توزیع نرمال نیز انجام می‌شود و عملکرد دو مدل در برآورد پارامترها مورد مقایسه قرار می‌گیرد. مدل رگرسیون بتا فضایی معرفی شده در (۶.۲) را به‌صورت

$$\text{logit}(\mu_i) = \beta_{۰} + \beta_{۱}x_{۱} + \beta_{۲}x_{۲} + \tau_i, \qquad \log(\phi_i) = \delta_{۰} + \delta_{۱}z_{۱}, \qquad i = ۱, \ldots, n,$$

در نظر بگیرید. همچنین با فرض آن که $Y \sim N_n(\mu, \sigma^{۲}I_n)$ مدل

$$\mu_i = \beta_{۰} + \beta_{۱}x_{۱} + \beta_{۲}x_{۲} + \tau_i, \qquad \log(\sigma^{-۲}) = \delta_{۰} + \delta_{۱}z_{۱}, \qquad i = ۱, \ldots, n,$$

را در نظر بگیرید. در مطالعه شبیه‌سازی یک مشبکه منظم ۱۵ × ۱۵ با موقعیت‌های $\{s_{ij}, i, j = ۱, \ldots ۱۵\}$ در نظر گرفته می‌شود. تولید داده‌ها با در نظر گرفتن مقادیر $\boldsymbol{\beta} = (\beta_{۰}, \beta_{۱}, \beta_{۲})^\top = (-۱, ۲, -۱/۵)^\top$ و $\boldsymbol{\delta} = (\delta_{۰}, \delta_{۱})^\top = (۴, -۰/۴)^\top$، و با به‌کار بردن تابع کوواریانس همسانگرد نمایی $(\Sigma)_{ij} = \sigma^{۲}_{۰}\exp\{- \mid s_i - s_j \mid \psi^{-۱}\}$ انجام شده است که در آن $\sigma^{۲}_{۰} = ۰/۵$ و $\psi^{-۱} = ۰/۱$. متغیرهای تبیینی از توزیع یکنواخت تولید شده‌اند. به شرط متغیر پنهان، پاسخ‌ها از توزیع $Beta(\mu_i\phi_i, (۱-\mu_i)\phi_i)$ تولید شده‌اند که در آن $\mu_i = \frac{\exp\{\boldsymbol{x}_i\boldsymbol{\beta}+\tau_i\}}{۱+\exp\{\boldsymbol{x}_i\boldsymbol{\beta}+\tau_i\}}$ و $\phi_i = \exp\{\boldsymbol{z}_i\boldsymbol{\delta}$ است. برای برآورد بیزی پارامترهای مدل توزیع‌های پیشین

متداول به‌صورت $\beta_i \sim N(۰, ۱۰۰)$، $\delta_i \sim N(۰, ۱۰۰)$، $\sigma_\varepsilon^۲ \sim \text{IG}(۰/۰۱, ۰/۰۱)$ و $\psi^{-۱} \sim U(۰/۰۱, ۲۰)$ در نظر گرفته می‌شوند. مقادیر پارامترهای توزیع‌های پیشین به‌گونه‌ای انتخاب شده‌اند که توزیع‌های پیشین ناآگاهی بخش[۶] باشند. برای نمونه با اندازه ۲۲۵، با تولید ۵۰۰۰۰۰ نمونه از توزیع‌های شرطی کامل و در نظر گرفتن دوره داغیدن ۲۵۰۰۰۰ و انتخاب نمونه‌ها با فاصله ۱۰۰، برآورد پارامترها بر اساس نمونه‌ای با اندازه ۲۵۰۰ به‌دست آمده‌اند. تعداد نمونه از توزیع‌های شرطی کامل و فاصله نمونه‌گیری مورد نیاز برای برآورد پارامترهای مدل، با انجام شبیه‌سازی‌های مختلف و بررسی رفتار زنجیره‌های مارکوف تعیین شد. آزمون تشخیصی همگرایی آماره گویک[۷] (گویک، ۱۹۹۲) فرض همگرایی را رد نمی‌کند. همچنین معیار تحلیل نسبی چند متغیره[۸] MPRF با مقدار ۱/۰۲ که کوچک‌تر از مقدار ۱/۲ است، فرض همگرایی را تأیید می‌کند. با در نظر گرفتن این مفروضات ۵۰ مجموعه داده شبیه‌سازی شده و برآورد بیزی پارامترها به همراه اریبی نسبی (RB) و جذر میانگین توان دوم خطا (RMSE) از روابط

$$\text{RB}(\theta) = \frac{۱}{۵۰}\sum_{i=۱}^{۵۰}(\widehat{\theta}_i/\theta - ۱), \qquad \text{RMSE}(\theta) = \left\{\frac{۱}{۵۰}\sum_{i=۱}^{۵۰}(\widehat{\theta}_i - \theta)^۲\right\}^{۱/۲} \tag{۱.۳}$$

محاسبه شده است، که در آن‌ها $\boldsymbol{\theta} = (\beta_۰, \beta_۱, \beta_۲, \delta_۰, \delta_۱, \sigma_\varepsilon^۲, \psi^{-۱})^\top$ و $\hat{\theta}_i$ برآورد بیزی θ_i در iامین نمونه است. با مقایسه

جدول ۱: برآورد پارامترهای مدل همزمان میانگین و دقت برای مشبکه با اندازه ۲۲۵

		مدل بتا فضایی			مدل گاوسی		
پارامتر	مقدار واقعی	برآورد	اریبی نسبی	RMSE	برآورد	اریبی نسبی	RMSE
$\beta_۰$	-۱	$-۰/۷۰۱$	$-۰/۲۹۹$	۰/۳۸۷	۰/۴۷۶	$-۱/۴۷۶$	۱/۴۷۶
$\beta_۱$	۲	۲/۰۱۳	۰/۰۰۷	۰/۰۸۴	۰/۰۳۳	$-۰/۹۸۴$	۱/۹۶۷
$\beta_۲$	$-۱/۵$	$-۱/۵۲۵$	۰/۰۱۶	۰/۰۸۴	$-۰/۰۱۵$	$-۰/۹۹۰$	۱/۴۸۵
$\delta_۰$	۴	۴/۰۱۳	۰/۰۰۳	۰/۲۸۳	۱۴/۱۶۵	۲/۵۴۱	۱۰/۱۷۰
$\delta_۱$	$-۰/۴$	$-۰/۳۹۰$	$-۰/۰۲۶$	۰/۴۰۶	۴/۹۴۳	$-۱۳/۳۵۹$	۵/۳۵۰
$\sigma_\varepsilon^۲$	۰/۵	۰/۵۱۰	۰/۰۲۰	۰/۲۲۴	۰/۰۱۲	$-۰/۹۷۵$	۰/۴۸۸
$\psi^{-۱}$	۰/۱	۰/۱۸۸	۰/۸۸۰	۰/۱۲۱	۱۰/۷۹۳	۱۰۶/۹۳۱	۱۰/۶۹۷

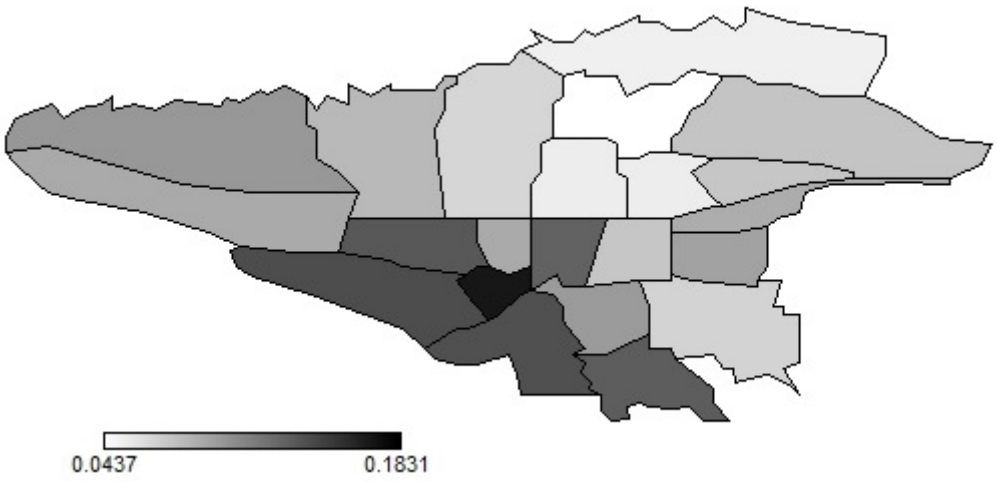
مقادیر RMSE در جدول ۱ ملاحظه می‌شود استفاده از مدل رگرسیون آمیخته خطی تعمیم‌یافته بتا فضایی منجر به بهبود در برآورد پارامترها شده است. بنابراین لازم است برای مدل‌بندی داده‌های همبسته فضایی محدود در بازه $(۰, ۱)$ از مدل بتا فضایی به جای مدل گاوسی استفاده شود.

۴ تحلیل سهم درآمد خانوار از یارانه

حدود پنجاه سال است که آمارگیری از هزینه و درآمد خانوارهای کشور همه ساله توسط مرکز آمار ایران اجرا و نتایج آن استخراج و منتشر شده است. برای شناسایی متغیرهای مؤثر بر سهم درآمد خانوار از یارانه، که مقداری در بازه $(۰, ۱)$ است، می‌توان از

[۶] Noninformative
[۷] Geweke's Statistics
[۸] Multivariate Proportional Reduction Factor

مدل رگرسیون بتا استفاده کرد. یک نمونه تصادفی به اندازه ۲۰۳ از خانوارهای شهر تهران در سال ۱۳۹۰ انتخاب نموده و سهم درآمد خانوار از یارانه را به‌عنوان متغیر پاسخ در نظر می‌گیریم. شکل ۱ پهنه‌بندی فضایی میانگین متغیر پاسخ در شهر تهران را نشان می‌دهد. همان‌طور که ملاحظه می‌شود سهم درآمد خانوارها از یارانه در مناطق هم‌جوار مشابه است.



شکل ۱: میانگین سهم درآمد خانوار از یارانه در مناطق شهر تهران

بر اساس آزمون موران با pـ مقدار نزدیک به صفر فرض همبستگی فضایی متغیر پاسخ معنی‌دار است، بنابراین مدل آمیخته خطی تعمیم یافته بتا فضایی را برای تحلیل داده‌ها مورد استفاده قرار می‌دهیم. متغیرهای تبیینی که اطلاعات آن‌ها از طرح هزینه و درآمد خانوار در دسترس است و می‌توانند بر سهم درآمد خانوار از یارانه اثر داشته باشند شامل دهک وزنی درآمد (DI)، بعد خانوار (HS)، مساحت واحد مسکونی (AHU)، و دهک وزنی هزینه[۹] (DE) هستند. برای تمام ضرایب رگرسیونی توزیع پیشین نرمال به‌صورت $\boldsymbol{\beta} \sim N_5(۰, ۱۰^۲ I_5)$ و $\boldsymbol{\delta} \sim N_5(۰, ۱۰^۲ I_5)$، و برای پارامترهای اثر فضایی پیشین‌های $\sigma^۲ \sim \text{IG}(۰/۰۱, ۰/۰۱)$ و $\varphi^{-۱} \sim U(۰/۰۱, ۲۰)$ در نظر گرفته شده است. برآورد پارامترها پس از داغیدن ۲۵۰۰۰۰ نمونه از ۵۰۰۰۰۰ نمونه توزیع پسین و انتخاب یک نمونه از هر ۱۰۰ نمونه باقی‌مانده انجام شده است. با حذف متغیرهای تبیینی که اثر آن‌ها معنی‌دار نیست، مدل نهایی همزمان پارامترهای میانگین و دقت به‌صورت

$$\log\left(\frac{\mu_i}{۱-\mu_i}\right) = \beta_۰ + \beta_۱ \text{DI}_i + \beta_۲ \text{HS}_i + \beta_۳ \text{AHU}_i + \tau_i,$$

$$\log(\phi_i) = \delta_۰ + \delta_۱ \text{DI}_i + \delta_۲ \text{HS}_i \quad i = ۱, \ldots, ۲۰۳,$$

است. برآورد پارامترهای مدل به همراه بازه باور ۹۵٪ در جدول ۲ آمده است. همان‌طور که ملاحظه می‌شود متغیرهای تبیینی دهک درآمدی خانوار، بعد خانوار و مساحت واحد مسکونی در مدل میانگین معنی‌دار هستند. با افزایش دهک درآمدی خانوار و مساحت واحد مسکونی، سهم درآمد خانوار از یارانه کاهش و با افزایش جمعیت خانوار سهم درآمد خانوار از یارانه افزایش می یابد. در واقع در خانوارهایی با درآمدهای بیشتر و واحدهای مسکونی بزرگ‌تر، سهم درآمد خانوار از یارانه در مقایسه با خانوارهای کم درآمد کمتر است. از آن‌جا که با افزایش بعد خانوار سهم درآمد خانوار از یارانه افزایش می‌یابد می‌توان سهم درآمد خانوار از یارانه به ازای هر عضو خانوار (سرانه درآمد خانوار از یارانه) را به‌عنوان معیاری برای تشخیص وضعیت رفاه خانوار در نظر گرفت. به‌علاوه پارامترهای

[۹] Decile of Expenditure

مربوط به اثر فضایی معنی‌دار شده‌اند، یعنی محل زندگی خانوار به لحاظ جغرافیایی بر سهم درآمد خانوار از یارانه مؤثر است. در نتیجه محل زندگی خانوار به همراه مساحت واحد مسکونی و سرانه درآمد خانوار از یارانه می‌تواند برای تشخیص خانوارهایی که نیازمند دریافت یارانه نیستند مورد استفاده قرار گیرند.

جدول ۲: برآورد پارامترهای مدل سهم درآمد خانوار از یارانه در شهر تهران

منبع	پارامتر	برآورد	sd	بازه باور ۹۵٪
عرض از مبدأ	β_0	−۲/۲۲۴	۰/۰۵۹	(−۲/۱۰۷, −۲/۳۵۱)
دهک وزنی درآمد	β_1	−۰/۱۴۰	۰/۰۴۲	(−۰/۰۶۳, −۰/۲۳۰)
بعد خانوار	β_2	۰/۴۰۱	۰/۰۵۸	(۰/۳۷۲, ۰/۵۲۷)
مساحت واحد مسکونی	β_3	−۰/۲۱۳	۰/۰۴۸	(−۰/۲۱۹, −۰/۰۷۱)
عرض از مبدأ	δ_0	۴/۱۴۴	۰/۱۱۳	(۴/۰۴۵, ۴/۴۴۱)
دهک وزنی درآمد	δ_1	۰/۳۸۶	۰/۱۱۶	(۰/۳۸۰, ۰/۸۳۰)
بعد خانوار	δ_2	−۰/۶۲۱	۰/۱۳۷	(−۰/۹۷۳, −۰/۵۰۳)
واریانس فضایی	σ_0^{-2}	۱۰/۱۱۰	۶/۱۹۵	(۱/۸۲۰, ۲۴/۴۹۰)
دامنه فضایی	ψ^{-1}	۱/۹۸۷	۱/۹۳۷	(۰/۱۷۷, ۷/۴۰۷)

دهک درآمدی خانوار و بعد خانوار متغیرهای معنی‌دار در مدل پارامتر دقت هستند. با افزایش دهک درآمدی خانوار پارامتر دقت افزایش و در نتیجه واریانس پاسخ کاهش می‌یابد. از طرف دیگر با افزایش بعد خانوار، پارامتر دقت کاهش و درنتیجه واریانس متغیر پاسخ افزایش می‌یابد. بنابراین در خانوارهای پر درآمد کم جمعیت، تفاوت سهم درآمد خانوار از یارانه در مقایسه با خانوارهای کم درآمد پرجمعیت کمتر است.

بحث و نتیجه‌گیری

در این مقاله، مدل‌بندی همزمان پارامترهای میانگین و دقت توزیع بتای بازپارامتریده با رهیافت بیزی انجام شد. این مدل ابزاری برای مطالعه داده‌های چوله با واریانس ناهمگن فراهم می‌کند. مقایسه مدل آمیخته خطی تعمیم‌یافته بتا با اثرات فضایی و مدل گاوسی با اثرات فضایی برای مدل‌بندی داده‌های محدود در بازه (۱و۰) که همبستگی فضایی دارند، نشان داد که استفاده از مدل آمیخته خطی تعمیم‌یافته بتا با کاهش RMSE منجر به بهبود در برآورد پارامترها می‌شود.

مطالعه سهم درآمد خانوارها از یارانه در شهر تهران به‌عنوان یک کاربرد از مدل‌بندی همزمان پارامترهای میانگین و دقت با اثر فضایی، مورد بررسی قرار گرفت. نتایج نشان داد که سهم درآمد خانوار از یارانه با دهک درآمدی و متراژ واحد مسکونی رابطه مستقیم و با بعد خانوار رابطه عکس دارد. همچنین محل سکونت خانوار نیز بر متغیر پاسخ اثر دارد. تغییرپذیری متغیر پاسخ در میان خانوارهای کم درآمد پرجمعیت در مقایسه با خانوارهای پردرآمد کم جمعیت، بیشتر است. در واقع خانوارهای کم درآمد با جمعیت بالا از ثبات اقتصادی کمتری برخوردارند و دریافت یارانه در وضعیت رفاه آن‌ها نقش مؤثری دارد.

مراجع

Aitkin, M., (1987). Modelling Variance Heterogeneity in Normal Regression Using GLIM, *Journal of the Royal Statistical Society*, **36**, 332-339.

Cepeda-Cuervo, E. (2001). *Variability Modeling in Generalized Linear Models*, Unpublished Ph. D. Thesis, Mathematics Institute, Universidade Federal do Rio de Janeiro.

Cepeda-Cuervo, E. and Gamerman, D. (2000). Bayesian Modeling of Variance Heterogeneity in Normal Regression Models, *Brazilian Journal of Probability and Statistics*, **14**, 207-221.

Cepeda-Cuervo, E. and Gamerman, D. (2005). Bayesian Methodology for Modeling Parameters in the Two Parameter Exponential Family, *Revista Estadistica*, **57**, 93-105.

Cepeda-Cuervo, E., and Achcar, J. A. (2010). Heteroscedastic Nonlinear Regression Models, *Communications in Statistics—Simulation and Computation*, **39**, 405–419.

Cepeda-Cuervo, E., Migon, H. S., Garrido, L., and Achcar, J. A. (2014). Generalized Linear Models with Random Effects in the Two-Parameter Exponential Family, *Journal of Statistical Computation and Simulation*, **84**, 513-525.

Dey, D. K., Gelfand, A. E., and Peng, F. (1997). Overdispersed Generalized Linear Models, *Journal of Statistical Planning and Inference*, **64**, 93-107.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model Based Geostatistics, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**, 299-350.

Ferrari, S., and Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions, *Journal of Applied Statistics*, **31**. 799-815.

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B., and Ferrari, S. L. (2013). Mixed Beta Regression: A Bayesian Perspective, *Computational Statistics and Data Analysis*, **61**, 137-147.

Geweke, J., (1991). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments, **196** , Minneapolis: Federal Reserve Bank of Minneapolis, Research Department.

Kalhori, L., and Mohammadzadeh, M. (2016). Spatial Beta Regression Model with Random Effect, *Journal of Statistical Research of Iran*, **13**, 215-230.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models Second Edition*, London: Chapman and Hall.

Mohammadzadeh, M., and Kalhori, L. (2017). Bayesian Analysis Of Spatial Beta Generalized Linear Mixed Models, International Association for Mathematical Geosciences (IAMG 2017), Australia, Fremantle, 2-9 September 2017.

Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalized Linear Models, *Journal of the Royal Statiatical Association, Ser. A*, **135**, 370-384.

Paolino, P., (2001). Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables, *Political Analysis*, **9**, 325-346.

Smithson, M., and Verkuilen, J. (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables, *Psychological Methods*, **11**, 54-71.

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R, *Journal of Statistical Software*, **12**, 1-16.

Zimprich, D., (2010). Modeling Change in Skewed Variables Using Mixed Beta Regression Models, *Research in Human Development*, **7**, 9-26.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶

مدل بقای فضایی با اثرات تصادفی چوله گاوسی بسته

کیومرث مترجم*، محسن محمدزاده ، آمنه آبیار

گروه آمار، دانشگاه تربیت مدرس

چکیده: اثرات تصادفی یا همان مؤلفه شکنندگی در مدل‌های بقا به منظور ورود عوامل خطر ناشناخته در مدل بقا مورد استفاده قرار می‌گیرند. اما در بسیاری از مواقع همبستگی بین زمان‌های بقا از نوع فضایی است. در این حالت معمولاً یک میدان تصادفی گاوسی برای اثرات تصادفی فضایی در نظر گرفته می‌شود که با وارد کردن آن در مدل بقا، مدل شکنندگی بقای فضایی حاصل می‌شود. اما فرض گاوسی بودن اثرات تصادفی فضایی گاهی در عمل مطابق با واقعیت نیستند. در این مقاله با در نظر گرفتن یک میدان تصادفی چوله گاوسی بسته برای اثرات تصادفی در مدل بقای فضایی انحراف از فرض گاوسی بودن اثرات تصادفی بر برآورد پارامترهای مدل بقای فضایی مورد بررسی قرار می‌گیرد.

واژه‌های کلیدی: داده بقای فضایی، شکنندگی، مدل بقای فضایی، میدان تصادفی چوله گاوسی بسته.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62N86، 62N01.

۱ مقدمه

تحلیل بقا دارای سابقه طولانی در مطالعات پزشکی و قابلیت اعتماد در مهندسی است (کاکس و اوکس، ۱۹۸۴). معمولاً در مدل‌های بقا فرض می‌شود که زمان شکست واحدهای آزمایشی تحت مطالعه مستقل از هم هستند در حالی که در بسیاری از مواقع این فرض برقرار نبوده و ممکن است این زمان‌های شکست به‌صورت فضایی همبسته باشند.

محققانی مانند بگری و همکاران (۲۰۰۰) و تیوریچک و مادن (۲۰۰۲) و لیکستاین و همکارن (۲۰۰۲) و رامزی و همکاران (۲۰۰۳) در تحقیقات خود نشان دادند که در صورت وجود همبستگی فضایی در داده‌های بقا، عدم توجه به آن در مدل‌سازی و تحلیل داده‌های بقا می‌تواند منجر به نتایج نادرست و گمراه کننده‌ای شود.

معمولاً اثرات تصادفی به‌صورت یک مؤلفه پنهان در داده‌های بقا نهفته هستند که می‌توان با تشخیص وجود همبستگی فضایی و وارد کردن آن در مدل از طریق یک مدل بقای فضایی به نتایجی منطبق با واقعیت دست یافت. وارد کردن اثرات فضایی در مدل‌بندی داده‌های بقا سابقه‌ای کمتر از دو دهه دارد. اولین کوشش‌ها برای معرفی مدل‌های بقای فضایی را می‌توان در کارهای

* **ارائه‌دهنده مقاله:** کیومرث مترجم، k.motarjem@modares.ac.ir

لی و ریان (۲۰۰۲)، بانرجی و کارلین (۲۰۰۳)، بانرجی و همکاران (۲۰۰۴) و دیوا و همکاران (۲۰۰۷) مشاهده کرد. بیشتر مدل‌های بقای فضایی معرفی شده توسط محققین برای حالاتی است که داده‌های بقا به‌صورت مشبکه‌ای باشند. در واقع همبستگی فضایی بین نواحی که داده‌های بقا در آن قرار دارند برقرار باشد اما در حالاتی که داه‌های بقا به‌صورت زمین آماری باشند، تحلیل مدل‌های بقا به‌واسطه پیچیدگی در توابع درستنمایی و برآورد پارامترها مشکلات زیادی را به همراه دارد. مترجم و همکاران (۲۰۱۷) یک مدل بقای فضایی معرفی کردند که با استفاده از آن در حالتی که داده‌های بقا از نوع زمین آماری باشند نیز می‌توان اثرات تصادفی فضایی را در مدل وارد و پارامترهای مدل بقا را برآورد کرد. در این مدل یک میدان تصادفی گاوسی برای ورود اثرات تصادفی فضایی در نظر گرفته شده است. اما به دلیل وجود چولگی در داده‌های بقا ممکن است در نظر گرفتن فرض گاوسی بودن اثرات تصادفی در حالت کلی برقرار نباشد. لذا در این مقاله با در نظر گرفتن یک میدان تصادفی چوله گاوسی بسته[1] (CSN) برای اثرات تصادفی به بحث و بررسی میزان انحراف از فرض گاوسی بودن اثرات تصادفی پرداخته می‌شود. همچنین میزان تأثیر این انحراف بر برآورد پارامترهای مدل، به‌ویژه پارامترهای رگرسیونی که مورد علاقه محققان بقا در تفسیر زمان بقا است، مورد ارزیابی قرار می‌گیرد. در بخش ۲ به اختصار یک میدان تصادفی چوله گاوسی بسته معرفی می‌شود. سپس لم ارائه شده در مترجم و همکاران (۱۳۹۴) برای تولید داده‌های بقای فضایی با اثرات تصادفی گاوسی برای شبیه‌سازی از یک مدل بقا فضایی با اثرات تصادفی چوله گاوسی بسته تعمیم داده می‌شود. در بخش ۳ تأثیر انحراف از فرض گاوسی بودن اثرات تصادفی بر برآورد پارامترهای مدل بقای فضایی مورد ارزیابی و مقایسه قرار می‌گیرد. در انتها به بحث و بررسی نتایج پرداخته و پیشنهاداتی برای مطالعات آتی ارائه خواهد شد.

۲ میدان تصادفی چوله گاوسی بسته

گنزالس و همکاران (۲۰۰۴) توزیع چندمتغیره CSN که تعمیمی از توزیع نرمال چندمتغیره است را معرفی کردند. بردار nـبعدی Y دارای توزیع چوله نرمال بسته چندمتغیره است و به‌صورت $\mathrm{CSN}_{n+m}(\mu, \Sigma, D, \gamma, \Delta)$ نمایش داده می‌شود، هرگاه دارای تابع چگالی به‌صورت

$$C_m \phi_n(Y; \mu, \Sigma) \Phi_m(D^\top (Y - \mu); \gamma, \Delta) \qquad (۱.۲)$$
$$C_m^{-1} = \Phi_m(۰, \gamma, \Delta + D^\top \Sigma D)$$

باشد، که در آن $\mu \in R^n$، $\gamma \in R^m$ میانگین و $\Sigma \in R^{n\times n}$ و $\Delta \in R^{m\times m}$ ماتریس واریانس کوواریانس هستند و $D \in R^{n\times m}$ ماتریس چولگی و $\phi(Y; \mu, \Sigma)$ و $\Phi_n(Y; \mu, \Sigma)$ توابع چگالی و توزیع نرمال nـمتغیره با میانگین μ و ماتریس کوواریانس Σ هستند.

در توزیع چندمتغیره CSN با تابع چگالی (۱.۲) اگر $D = ۰$ ماتریس صفر باشد، تابع چگالی نرمال چندمتغیره و به‌ازای $m = ۱$، توزیع چوله نرمال چندمتغیره آزالینی و دالاواله (۱۹۹۶) را نتیجه خواهد داد. در این‌صورت بردار Y دارای توزیع $CSN_{n,۱}(\mu, \Sigma, \alpha, ۰, ۱)$ خواهد بود، که در آن α یک بردار با طول n و شامل ضرائب چولگی است. این توزیع اولین توزیع چندمتغیره چوله نرمال است که توسط آزالینی (۱۹۸۵)و آزالینی (۱۹۸۶) و آزالینی و دالاواله (۱۹۹۶) معرفی گردید.

[1] Closed skew normal random field

توزیع چوله نرمال چندمتغیره (۱.۲) از ترکیب دو بردار تصادفی U توزیع نرمال mمتغیره و Z توزیع نرمال nمتغیره به‌صورت

$$\begin{pmatrix} U \\ Z \end{pmatrix} \sim N_{m+n}\left(\begin{pmatrix} \gamma \\ 0 \end{pmatrix}, \begin{pmatrix} \Delta + D^\top \Sigma D & -D^\top \Sigma \\ -\Sigma D & \Sigma \end{pmatrix}\right) \tag{۲.۲}$$

ایجاد می‌شود. با در نظر گرفتن ترکیب بالا می‌توان به‌سادگی نشان داد که متغیر تصادفی $\mu + [Z|U \leq 0]$ دارای توزیع $\mathrm{CSN}_{n,m}(\mu, \Sigma, D, \gamma, \Delta)$ خواهد بود. توجه شود که نماد $U \leq 0$ معادل $U_i \leq 0$ برای تمام مقادیر $i = 1, \ldots, m$ است. همچنین

$$[Z|U] \sim N_n(-\Sigma D(\Delta + D^\top \Sigma D)^{-1}(U - \nu), \Sigma - \Sigma D(\Delta + D^\top \Sigma D)^{(-1)} D^\top \Sigma) \tag{۳.۲}$$

با انتخاب مقادیر برای n، m، μ، ν، Δ، Σ و D و استفاده از ساختار فوق می‌توان خانواده‌ای از توزیع چوله نرمال بسته معرفی کرد (دومینگز و همکاران، ۲۰۰۳). از توزیع CSN براساس (۲.۲) با یک الگوریتم دو مرحله‌ای می‌توان مقادیر تصادفی تولید کرد:

گام ۱. یک نمونه تصادفی از توزیع $U \overset{d}{=} N_m(N, \Delta + D^\top \Sigma D)$، به‌طوری‌که $U \leq 0$ تولید شود.

گام ۲. یک نمونه تصادفی از $[Z|U]$ براساس (۳.۲) تولید شود.

با انجام این دو گام و استفاده از $\mu + [Z|U \leq 0]$ یک نمونه تصادفی از توزیع $\mathrm{CSN}_{n,m}(\mu, \Sigma, D, \nu, \Delta)$ تولید می‌شود.

۱.۲ میدان تصادفی چوله گاوسی

فرض کنید $Z(s)$، $s \in R$ یک میدان تصادفی گاوسی مانا با میانگین صفر و تغییرنگار

$$2\gamma Z(h) = \mathrm{Var}(Z(s+h) - Z(s)); \quad h \in R^2$$

باشد، که در آن $\mathrm{Var}(Z(s)) = \sigma^2$ است (کرسی، ۱۹۹۳). ماتریس کوواریانس بردار $Z = (Z(s_1), \ldots, Z(s_n))^\top$ براساس تابع کوواریانس $C_Z(h) = \sigma^2 - \gamma_Z(h)$ ساخته و با نماد Σ نمایش داده می‌شود. برای پیوند دادن این ساختار فضایی با توزیع چوله نرمال براساس (۲.۲)، میدان تصادفی چوله گاوسی $Y(s)$ به‌صورت

$$Y(s) \overset{d}{=} \mu + [Z(s)|U \leq 0] \tag{۴.۲}$$

تعریف می‌شود، به‌طوری‌که برای هر بردار nبعدی $Z = (Z(s_1), \ldots, Z(s_n))^\top$ از میدان $Z(\cdot)$ داریم:

$$Y \overset{d}{=} \mu + [Z|U \leq 0]$$

که در آن $U \sim N_m(0, \Delta + D^{top}\Sigma D)$ و Z براساس (۲.۲) توزیع شده است.

در عمل مشاهدات تحقیقی از بردار $(Y(s_1), \ldots, Y(s_n))^\top$ است و از U و Z مشاهده‌ای در دست نیست. اما بردار Y را می‌توان از مجموع دو فرایند مستقل به‌صورت

$$Y = \begin{pmatrix} U \\ V \end{pmatrix} \sim N_{m+n}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Delta + D^\top \Sigma D & 0 \\ 0 & I_n \end{pmatrix}\right)$$

بیان کرد، که در آن I_n یک ماتریس همانی از اندازه n است. در این حالت بردار Z را می‌توان به‌صورت $Z = -FU + G^{\frac{1}{2}}V$ نوشت، که در آن $F = \Sigma D(\Delta + D^\top \Sigma D)^{-1}$ و $G = \Sigma - \Sigma D(\Delta + D^\top \Sigma D)^{-1} D^\top \Sigma$.

با توجه به (۴.۲) و استقلال بین U و V داریم:

$$Y \overset{d}{=} \mu - F[U|U \leq 0] + G^{\frac{1}{2}}V \tag{۵.۲}$$

برای پرهیز از پیچیدگی و حجم زیاد محاسبات یک حالت ساده از میدان تصادفی CSN شبیه‌سازی می‌شود. انتخاب مقادیر مناسب برای m، D و Δ تأثیر زیادی بر حجم این محاسبات خواهد داشت. برای تولید داده از میدان تصادفی ارائه شده در (۵.۲)، $D = \delta A$، قرار داده می‌شود، که در آن مقدار ثابت δ نشان دهندهٔ میزان چولگی و A یک ماتریس غیر صفر با عناصری مستقل از δ است. در این‌صورت رابطه (۵.۲) به‌صورت

$$Y \overset{d}{=} \mu - \delta\Sigma A(\Delta + \delta^2 A^\top \Sigma A)^{-1}[U|U \leq 0] + (\Sigma - \delta^2 \Sigma A(\Delta + \delta^2 A^\top \Sigma A)^{-1} A^\top \Sigma)^{\frac{1}{2}} V \tag{۶.۲}$$

است (آلارد و ناوی، ۲۰۰۷). واضح است که درعبارت (۶.۲) در صورتی که $\delta = 0$ باشد، Y از U مستقل خواهد بود و در نتیجه Y بردار گاوسی با میانگین μ و ماتریس واریانس Σ خواهد بود. با قرار دادن $A^\top \Sigma A = \Sigma_m$ و $\Delta = \Sigma_m$، عبارت (۶.۲) به‌صورت

$$Y \overset{d}{=} \mu - \frac{\delta}{1+\delta^2}\Sigma A \Sigma_m^{-1}[U|U \leq 0] + (I - \frac{\delta^2}{1+\delta^2}\Sigma A \Sigma_m^{-1} A^\top \Sigma)^{\frac{1}{2}} V \tag{۷.۲}$$

ساده می‌شود. به‌ازای $n = m$ و $A = I_n$، عبارت (۷.۲) به‌صورت

$$\begin{aligned} Y &\overset{d}{=} \mu - \frac{\delta}{1+\delta^2}[U|U \leq 0] + \frac{1}{\sqrt{1+\delta^2}}\Sigma^{\frac{1}{2}} V \\ Y &\overset{d}{=} \mu + \frac{\delta}{1+\delta^2}[U|U \geq 0] + \frac{1}{\sqrt{1+\delta^2}}\Sigma^{\frac{1}{2}} V \end{aligned} \tag{۸.۲}$$

خواهد شد. در این مقاله با استفاده از (۸.۲) به شبیه‌سازی یک میدان تصادفی چوله گاوسی بسته و در ادامه تولید داده‌های بقای فضایی پرداخته می‌شود.

۳ مدل بقای فضایی

اگر مقادیر زمان‌های بقا به موقعیت قرارگیری آنها در فضای مورد مطالعه وابسته باشند اینگونه داده‌ها را داده‌های بقای فضایی می‌نامیم. برای تولید داده‌های بقای فضایی با اثرات تصادفی چوله گاوسی و استفاده از رابطه (۸.۲) در بخش قبل ابتدا مدل بقای فضایی زمین آماری (مترجم و همکاران، ۲۰۱۷) به اختصار معرفی می‌شود. این مدل با در نظر گرفتن یک میدان تصادفی گاوسی برای اثرات تصادفی و تعمیم مدل کاکس (کاکس، ۱۹۷۲) مدل بقای فضایی زمین آماری به‌صورت

$$h(t|X, Z) = h_0(t) \exp\{\beta^\top X + Z(s)\}$$

تعریف می‌شود، که در آن $h_0(\cdot)$ تابع خطر پایه، β بردار پارامترهای رگرسیونی و $Z(\cdot)$ یک میدان تصادفی گاوسی و s موقعیت داده بقای فضایی است.

برای تولید داده بقای فضایی با اثرات تصادفی چوله گاوسی روشی که توسط مترجم و همکاران، (۱۳۹۴) معرفی شد، تعمیم داده می‌شود.

لم ۱.۳. فرض کنید تابع خطر پایه $h_0(\cdot)$ در مدل بقا به‌صورت پارامتری و معلوم باشد. از اینرو با توجه به فرم تابع بقا برای مدل مخاطرات متناسب داریم:

$$S(t|X) = \exp\left\{-H_0(t)\exp\{\beta^\top X + Y(s)\}\right\}$$

که در آن $H_0(t) = \int_0^t h_0(u)du$ تابع خطر پایه تجمعی، β بردار پارامترهای رگرسیونی $Y(\cdot)$ یک میدان تصادفی چوله گاوسی بسته و S تابع بقا است.

گزاره ۲.۳. *تابع بقا به‌صورت $S(t|X) = 1 - F(t|X)$ تعریف می‌شود، که در آن F تابع توزیع تجمعی است. همچنین می‌دانیم $F(\cdot)$ دارای توزیع یکنواخت پیوسته روی بازه $[0, 1]$ است. لذا داریم:*

$$U = \exp\left\{-H_0(t)\exp\{\beta^\top X + Y(s)\}\right\} \qquad (۱.۳)$$

با توجه به آنکه $U \sim U[0, 1]$، داریم:

$$\begin{aligned} \log(U) &= -H_0(t)\exp\{\beta^\top X + Y(s)\} \\ H_0(t) &= \frac{-\log(U)}{\exp\{\beta^\top X + Y(s))\}} \Rightarrow T = H_0^{-1}\left[\frac{-\log(U)}{\exp\{\beta^\top X + Y(s)\}}\right] \end{aligned}$$

با در نظر گرفتن تابع خطر پایه نمایی با پارامتر $\lambda > 0$ که در این مطالعه از آن استفاده شده است، داریم:

$$T = -\frac{\log U}{\lambda\exp\{\beta^\top X + Y(s)\}}$$

با قرار دادن $Y(\cdot)$ از رابطه (۸.۲) و تعیین مقادیر λ و β، داده‌های بقای فضایی با اثرات تصادفی چوله گاوسی تولید می‌شوند.

در بخش بعدی با مشخص کردن مقادیر پارامترهای رابطه (۱.۳) به تولید و شبیه‌سازی داده‌های بقا با اثرات تصادفی چوله گاوسی می‌پردازیم.

۴ مطالعه شبیه‌سازی

در این مطالعه برای شبیه‌سازی داده‌های بقای فضایی با اثرات تصادفی چوله گاوسی بسته با استفاده از لم ۱.۳ در مجموع ۴۰۰ داده بقای فضایی در یک مشبکه منظم ۲۰ × ۲۰ تولید می‌شود. برای شبیه‌سازی این داده‌ها از تابع خطر پایه با توزیع نمایی و پارامتر $\lambda = 1$، بردار متغیرهای تبیینی از توزیع نرمال استاندارد و ضریب رگرسیونی $\beta = 1$ استفاده شده است. $Y(\cdot)$ تحققی از یک میدان تصادفی چوله گاوسی بسته با مقادیر $\sigma^2 = 2$، $a = 1$ و $\delta = 5$ است که در آن σ^2، a و δ به ترتیب واریانس، دامنه و میزان چولگی میدان است. مجموعه داده‌ها با درصدهای سانسور ۲۰ و ۸۰ و تکرارهای ۱۰۰ و ۵۰۰ تولید شد. پس از تولید داده‌های بقا با اثرات تصادفی چوله گاوسی بسته مدل‌های کاکس، شکنندگی و بقای فضایی به این داده‌ها برازش داده شد. در مرحله بعد با در نظر گرفتن $\delta = 0$ عملاً میدان تصادفی $Y(\cdot)$ تبدیل به یک میدان تصادفی گاوسی می‌شود و با ثابت نگه داشتن سایر پارامترها مجدداً شبیه‌سازی را تکرار شد تا داده‌های بقای فضایی با اثرات تصادفی گاوسی تولید شود.

جدول ۱: برآورد پارامترهای مدل کاکس در برازش به داده‌های بقای فضایی با اثرات تصادفی گاوسی و چوله گاوسی

	گاوسی						چوله گاوسی بسته					
درصد	$H=۱۰۰$			$H=۵۰۰$			$H=۱۰۰$			$H=۵۰۰$		
سانسور	β	MAPB	MSE	β	MAPB	MSE	β	MAPB	MSE	β	MAPB	MSE
۲۰٪	۰/۶۲	۳۷/۲۶	۰/۰۸۷	۰/۶۹	۳۱/۴۴	۰/۰۷۹	۰/۵۷	۴۲/۱۷	۰/۰۹۶	۰,۶۳	۳۷/۱۵	۰/۰۸۴
۸۰٪	۰/۵۷	۴۲/۱۱	۰/۰۹۶	۰/۶۴	۳۳/۹۹	۰/۰۸۳	۰/۵۲۰	۴۸/۱۶	۰/۱۰۴	۰,۵۸	۴۲/۱۹	۰/۰۹۴

جدول ۲: برآورد پارامترهای مدل شکنندگی در برازش به داده‌های بقای فضایی با اثرات تصادفی گاوسی و چوله گاوسی

	گاوسی						چوله گاوسی بسته					
درصد	$H=۱۰۰$			$H=۵۰۰$			$H=۱۰۰$			$H=۵۰۰$		
سانسور	β	MAPB	MSE	β	MAPB	MSE	β	MAPB	MSE	β	MAPB	MSE
۲۰٪	۰/۶۷	۳۲/۳۵	۰/۰۸۳	۰/۷۴	۲۶/۷۵	۰/۰۷۷	۰/۶۱	۳۹/۴۷	۰/۰۹۳	۰/۶۸	۳۲/۴۵	۰/۰۸۹
۸۰٪	۰/۶۱	۳۸/۲۱	۰/۰۸۹	۰/۶۷	۳۳/۱۳	۰/۰۸۵	۰/۵۶	۴۴/۱۷	۰/۱۰۱	۰/۶۱	۳۹/۱۳	۰/۰۹۸

برای ارزیابی عملکرد مدل‌های مختلف در برازش به داده‌های بقا از معیار درصد قدرمطلق اریبی[۲] (MAPB) به‌صورت

$$\text{MAPB} = \frac{۱}{H}\sum_{i=۱}^{H}\left|\frac{\hat{\theta}_i - \theta}{\theta}\right| \times ۱۰۰$$

استفاده شد، که در آن، $\theta = (\sigma^۲, a, \beta, \delta)$ بردار پارامترهای مدل بقای فضایی و H تعداد تکرارها است.

جدول ۳: برآورد پارامترهای مدل بقای فضایی در برازش به داده‌های بقای فضایی با اثرات تصادفی گاوسی

درصد		$H=۱۰۰$			$H=۵۰۰$		
سانسور	پارامتر	براورد	MAPB	MSE	براورد	MAPB	MSE
	β	۰/۹۶	۳/۲۱	۰/۰۱۲	۰/۹۸	۱/۹۸	۰/۰۰۷
۲۰٪	a	۰/۹۵	۵/۳۲	۰/۰۰۸	۰/۹۷	۳/۲۱	۰/۰۰۴
	$\sigma^۲$	۱/۹۷	۲/۱۳	۰/۰۰۴	۱/۹۸	۱/۱۹	۰/۰۰۱
	β	۰/۹۲	۷/۱۲	۰/۰۱۶	۰/۹۴	۵/۱۹	۰/۰۰۹
۸۰٪	a	۰/۹۰	۹/۱۸	۰/۰۱۵	۰/۹۳	۶/۱۲	۰/۰۰۱
	$\sigma^۲$	۱/۹۱	۷/۱۱	۰/۰۱۱	۱/۹۵	۴/۱۴	۰/۰۰۹

بحث و نتیجه‌گیری

همان‌طور که در جدول ۱ ملاحظه می‌شود، مدل کاکس در برازش به داده‌های بقای فضایی با اثرات تصادفی گاوسی و چوله گاوسی نتایج بسیار نامناسبی را ارائه می‌کند. این نشان می‌دهد در صورت وجود اثرات تصادفی گاوسی یا چوله گاوسی در داده‌های بقای فضایی، استفاده از مدل کاکس نتایج بسیار دور از واقعیت را ارائه می‌کند. لذا بررسی وجود اثرات تصادفی هنگام استفاده از مدل کاکس توسط کاربران بسیار مهم و ضروری به نظر می‌رسد. نتایج جدول ۲ نشان می‌دهد علی رغم وجود یک مؤلفه تصادفی در

[۲] Mean Absolute Percentage Bias

جدول ۴: برآورد پارامترهای مدل بقای فضایی در برازش به داده‌های بقای فضایی با اثرات تصادفی چوله گاوسی بسته

درصد		$H=۱۰۰$			$H=۵۰۰$		
سانسور	پارامتر	براورد	MAPB	MSE	براورد	MAPB	MSE
	β	۰/۸۵	۱۴/۳۸	۰/۰۳۲	۰/۸۸	۱۲/۷۸	۰/۰۲۵
۲۰٪	a	۰/۸۱	۱۹/۶۴	۰/۰۱۹	۰/۸۳	۱۶/۹۲	۰/۰۱۶
	$\sigma^۲$	۱/۷۱	۲۹/۱۶	۰/۰۳۹	۱/۷۷	۲۳/۱۶	۰/۰۲۳
	β	۰/۷۹	۲۱/۱۳	۰/۰۴۳	۰/۸۳	۱۷/۲۳	۰/۰۳۱
۸۰٪	a	۰/۷۷	۲۴/۱۶	۰/۰۳۹	۰/۸۱	۱۹/۶۴	۰/۰۲۴
	$\sigma^۲$	۱/۶۵	۳۵/۱۸	۰/۰۵۵	۱/۶۹	۲۹/۱۲	۰/۰۴۲

مدل شکنندگی، این مدل نیز نتوانسته برازش خوبی را به داده‌ها ارائه دهد و عملاً استفاده از آن در حضور اثرات تصادفی فضایی انتخاب خوبی نخواهد بود. درجدول ۳ همان‌طور که مشاهده می‌شود برازش مدل بقای فضایی به داده‌های بقای فضایی با اثرات گاوسی نتایج بسیار مطلوبی را به همراه دارد. اما مقایسه نتایج جداول ۳ و ۴ نشان می‌دهد مدل بقا فضایی در حضور اثرات تصادفی چوله گاوسی عملکرد چندان مطلوبی ندارد. لذا می‌توان مدل بقای فضایی ارائه شده را برای حالتی که اثرات تصادفی از یک میدان تصادفی چوله گاوسی باشند تعمیم داد و برآورد پارامترهای مدل را بهبود بخشید.

مراجع

مترجم، ک.، محمدزاده، م. و آبیار، آ. (۱۳۹۴)، مدل‌های شکنندگی و خطرهای متناسب برای تحلیل داده‌های بقای فضایی، *مجله مدل‌سازی پیشرفته ریاضی*، **۴**، ۱۲۳-۱۰۱.

Allard, D. and Naveau. P. (2007), A New Spatial Skew-Normal Random Field Model, *Communications in Statistics* , **36**, 1821-1834.

Azzalini, A. (1985), A Class of Distributions which Includes the Normal Ones, *Scandinavian Journal of Statistics*, **12**, 171-178.

Azzalini, A. (1986), Further Results on A Class of Distributions which Includes the Normal Ones, *Statistica*, **46**, 199-208.

Azzalini, A. and Dalla Valle, A. (1996), The Multivariate Skew-normal Distribution, *Biometrika*, **83**, 715-726.

Banerjee, S. and Carlin, B. P. (2003), Semi–parametric Spatiotemporal Analysis, *Environmetrics*, **14**, 523-535.

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

Biggeri, A., Marchi, M., Lagazio, C., Martuzzi, M. and B¨ohning, D. (2000), Nonparametric Maximum Likelihood Estimators for Disease Mapping, *Statistics in Medicine*, **19**, 2539-2554.

Cox, D. R. (1972), Regression Models and Life-Tables, *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.

Cox, D. R. and Oakes, D. (1984), *Analysis of Survival Data*, Chapman and Hall, London.

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

Diva, U., Banerjee, S. and Dey, D. K. (2007), Modeling Spatially Correlated Survival Data for Individuals with Multiple Cancers, *Statistical Modelling*, **7**, 191–213.

Dominguez-Molina, J., Gonzalez- Farias, G. and Gupta, A. (2003), The Multivariate Closed Skew Normal Distribution, *Technical Report*, 03-12, Department of Mathematics and Statistics, Bowling Green State University.

Gonzalez-Farias, G., Dominguez-Molina, J. and Gupta, A. (2004), The Closed Skewnormal Distribution, In: Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality, Genton, M. G., Ed., Chapman and Hall, Boca Raton.

Motarjem, K., Mohammadzadeh, M. and Abyar, A. (2017), Geostatistical Survival Model with Gaussian Random Effect, *Statistical Papers*, DOI: 10.1007/s00362-017- 0922-8.

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

Lichstein, J. W., Simons, T. R., Shriner, S. A. and Franzreb, K. E. (2002), Spatial Autocorrelation and Autoregressive Models in Ecology, *Ecological Monographs*, **72**, 445-463.

Ramsay, T., Burnett, R. and Krewski, D. (2003), Exploring Bias in a Generalized Additive Model for Spatial Air Pollution Data, *Environmental Health Perspectives*, **111**, 1283-1288.

Turechek, W. W. and Madden, L. V. (2002), A Generalized Linear Modeling Approach for Characterizing Disease Incidence in Spatial Hierarchy, *Phytopathology*, **93**, 458-466.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶

برآورد و پیش‌گویی استوار با تاثیر کراندار در برآورد کوچک ناحیه‌ای فضایی

انور محمدی*، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده: روش‌های برآورد کوچک‌ناحیه‌ای در سال‌های اخیر بسیار مورد توجه قرار گرفته‌اند. هدف اصلی این مدل‌ها بکارگیری اطلاعات کمکی برای ارتقاء برآوردهای مستقیم است. مشابهت فضایی بین نواحی همجوار یکی از این دسته اطلاعات مفید است و مدلهای کوچک ناحیه‌ای فضایی به هدف بهره‌گیری از این اطلاعات توسعه یافته‌اند. بعلاوه با توجه به حجم کم نمونه در این مدل‌ها، نتایج به راحتی تحت تاثیر داده‌های دورافتاده قرار می‌گیرند. از این رو اخیراً مطالعات مختلفی درخصوص روش‌های استوار در این حوزه انجام شده است. روش‌های ارائه شده تاکنون متمرکز بر کنترل اثر مشاهدات دورافتاده در متغیر پاسخ بوده‌اند. اما اثر مقادیر دورافتاده در متغیرهای توضیحی همچنان توسط این روش‌ها قابل کنترل نیست. در این مقاله روش‌های استواری برای برآورد کوچک ناحیه‌ای فضایی ارائه می‌شود که قادرند تاثیر داده‌های دورافتاده در متغیر پاسخ و متغیرهای توضیحی را بطور همزمان کنترل کنند. عملکرد این روش‌ها در مطالعه‌های شبیه‌سازی مورد ارزیابی قرار گرفته است.

واژه‌های کلیدی: برآورد کوچک‌ناحیه‌ای، برآورد کوچک‌ناحیه‌ای فضایی، برآورد کوچک‌ناحیه‌ای استوار، جی‌ام‌ـ برآوردگرها، نقاط اهرمی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 91B72. 91D25.

۱ مقدمه

در آمارگیری‌های سراسری در سطح کشورها علاوه بر نتایج کلی، سازمان‌های آماری و تصمیم‌گیران و سیاست‌گذاران اقتصادی و اجتماعی علاقه‌مندند که پیش‌گویی‌هایی در سطح ناحیه‌های کوچکتر نیز در اختیار داشته باشند. در عمل برآورد مستقیم با استفاده از روش‌های معمول به علت حجم اندک (یا نبود) نمونه در این ناحیه‌ها امکان‌پذیر نیست. اما در روش‌های موسوم به برآورد کوچک ناحیه‌ای سعی می‌شود تا با استفاده از اطلاعات کمکی، برآوردهای آماری در سطح نواحی کوچک ارائه گردد. از این نظر روش‌های

* **ارایه‌دهنده مقاله:** محمدی، انور، anvar.mohammadi@modares.ac.ir

برآورد کوچک‌ناحیه‌ای در سال‌های اخیر بسیار مورد توجه پژوهشگران قرار داشته است. رائو (۲۰۰۳) جیانگ و لاهیری (۲۰۰۶) پففرمن (۲۰۱۳) و رائو و مولینا (۲۰۱۵) تاریخچه‌ای از شکل‌گیری روش‌های برآورد کوچک‌ناحیه‌ای را شرح داده و روش‌ها و مطالعات موجود را دسته‌بندی کرده‌اند.

روش‌های آماری عموماً برپایه برخی پیش‌فرض‌ها ساخته می‌شوند. به این ترتیب نتایج آنها زمانی که پیش‌فرض‌های مربوطه برقرار نباشند، قابل استفاده نخواهد بود. بعلاوه حضور داده‌های دورافتاده نیز می‌تواند تاثیر نامناسبی در برآوردهای آماری داشته باشد. از این نظر روش‌های آماری استوار برای پرهیز از تاثیر داده‌های دورافتاده یا عدم‌برقراری پیش‌فرض‌ها توسعه یافته‌اند. یکی از مسائلی که در سال‌های اخیر در برآورد کوچک‌ناحیه‌ای بسیار مورد توجه پژوهشگران قرار گرفته است، نحوه رویارویی با تاثیر داده‌های دورافتاده است. چمبرز و زاویدیس (۲۰۰۶) با تعمیم رگرسیون ام‌ـ‌چندک، روشی را برای برآورد استوار کوچک ناحیه‌ای پیشنهاد نمودند که در آن چندک‌های توزیعی از متغیر موردعلاقه مدل می‌شوند. اما سینها و رائو (۲۰۰۹) با تمرکز بر مدل‌های آمیخته خطی و با الگوگیری از روش‌های استواری که فلنر (۱۹۸۶) هاگینز (۱۹۹۳) و ریچاردسون و ولش (۱۹۹۵) برای این مدل‌ها ارائه کرده بودند، روشی استوار برای برآورد کوچک‌ناحیه‌ای در مدل‌های سطح واحد آماری ارائه نمودند. کار سینها و رائو در سال‌های بعد توسط پژوهشگران زیادی مورد توجه قرار گرفت و توسعه یافت. این روش توسط اشمید و مونیش (۲۰۱۴) به مدل کوچک‌ناحیه‌ای فضایی در سطح واحد آماری توسعه داده شد. در ادامه رائو و مولینا (۲۰۱۵) روش آنها را برای برآورد در مدل‌های کوچک‌ناحیه‌ای در سطح ناحیه توسعه دادند و وارنهولز (۲۰۱۶) آن را به مدل‌های کوچک‌ناحیه‌ای فضایی‌ـ‌زمانی در سطح ناحیه تعمیم و نیز الگوریتم‌های محاسباتی موردنیاز را بهبود بخشید.

در روش‌های استوار ارائه شده که بر پایه روش سینها و رائو توسعه یافته‌اند، از رهیافت ام‌ـ‌برآوردگرها استفاده می‌شود. این دسته از روش‌های استوار بر روی داده‌های دورافتاده در متغیر پاسخ تمرکز دارند. به عبارتی داده دورافتاده در این روش‌ها آن داده‌ای است که منجر به خطای بزرگی گردد. اما علاوه بر داده دورافتاده در متغیر پاسخ، لازم است تا اثر مقادیر نامتعارف در متغیرهای توضیحی نیز کنترل گردد. در برآورد مدل‌های رگرسیونی، کلاسی از برآوردگرها موسوم به ام‌ـ‌برآوردگرهای تعمیم یافته (به اختصار جی‌ام‌ـ‌برآوردگرها) برای این هدف توسعه یافته‌اند که در این مقاله از این رهیافت در برآورد کوچک‌ناحیه‌ای بهره خواهیم برد.

در بخش دوم این مقاله مدل‌های کوچک ناحیه‌ای و تعمیم آنها برای استفاده از اطلاعات همبستگی فضایی معرفی شده و روش‌های کنونی برآورد و پیش‌گویی استوار در آنها مرور شده است. در بخش سوم جی‌ام‌ـ‌برآوردگرهای رگرسیونی معرفی شده‌اند و در بخش چهارم روش استوار پیشنهادی برای برآورد کوچک‌ناحیه‌ای با استفاده از جی‌ام‌ـ‌برآوردگرها ارائه شده و الگوریتم‌های محاسباتی برای روش پیشنهادی تشریح شده است. در بخش پنجم یک مطالعه شبیه‌سازی برای ارزیابی و مقایسه روش پیشنهادی انجام شده است.

۲ برآورد استوار در مدل‌های کوچک ناحیه‌ای‌ـ روش‌های موجود

مدل‌های مورد استفاده در برآورد کوچک ناحیه‌ای در دو دسته کلی مدل‌های در سطح ناحیه و مدل‌های در سطح واحد آماری قابل دسته‌بندی هستند. در ادامه مدل‌های در سطح ناحیه معرفی و روش استوار برآورد و پیش‌گویی در آنها مرور می‌گردد.

۱.۲ مدل کوچک ناحیه ای در سطح ناحیه

در برآورد کوچک‌ناحیه‌ای، اگر اطلاعات کمکی برای هر واحد آماری در دسترس باشد، مدل آماری مورد استفاده مدل در سطح واحد آماری[۱] نامیده می‌شود. در مقابل اگر اطلاعات کمکی تنها در سطح ناحیه‌ها در دسترس باشند، از مدل در سطح ناحیه [۲] که توسط فی و هریوت (۱۹۷۹) معرفی شد، استفاده می‌شود. فرم کلی مدل فی‌ـهریوت را می‌توان به صورت

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i + e_i, \quad i = 1, ..., D \tag{۱.۲}$$

نوشت که در آن y_i برآورد مستقیم پارامتر مورد علاقه، $\mathbf{x}_i$ بردار مقادیر متغیرهای کمکی، e_i جمله خطا و u_i اثر تصادفی خاص ناحیه iاُم هستند. همچنین D تعداد ناحیه‌ها و $\boldsymbol{\beta}$ بردار ضرائب رگرسیونی است. در این مدل $\{e_i\}$ و $\{u_i\}$ مستقل و دارای توزیع نرمال با میانگین صفر و واریانس‌های σ_e^2 (معلوم) و σ_u^2 فرض می‌شوند (رائو، ۲۰۰۳). مدل فوق را با تعریف،$\mathbf{y}_{n\times 1} = (y_i)$، $\mathbf{u}_{n\times 1} = (u_i)$، $\mathbf{X}_{n\times p} = (x_i{}^\top)$ و $\mathbf{e}_{n\times 1} = (e_i)$ و $\mathbf{Z}_{n\times n} = \mathbf{I}_n$ می‌توان به صورت ماتریسی به فرم زیر نمایش داد

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{۲.۲}$$

در مدل فوق، $\mathbf{y}$ دارای توزیع نرمال با میانگین $\mathbf{X}\boldsymbol{\beta}$ و واریانس

$$\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \mathbf{Z}\mathrm{Var}(\mathbf{u})\mathbf{Z}^\top + \mathrm{Var}(\mathbf{e}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^\top + \mathbf{R}$$

می‌باشد که تابعی از پارامتر نامعلوم $\theta = \sigma_u^2$ است. همانگونه که می‌بینید که واریانس بردارهای تصادفی $\mathbf{u}$ و $\mathbf{e}$ مطابق نمادگذاری رایج با $\mathbf{G}$ و $\mathbf{R}$ نشان داده شده‌اند. در صورت معلوم بودن پارامترهای واریانس θ، بهترین برآوردگر خطی نااریب (BLUE) برای $\boldsymbol{\beta}$ و بهترین پیشگوی خطی نااریب (BLUP) برای $\mathbf{u}$ به صورت زیر قابل محاسبه‌اند:

$$\hat{\mathbf{u}}^{\mathrm{BLUP}} = \mathbf{G}\mathbf{Z}^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{۳.۲}$$

$$\hat{\boldsymbol{\beta}}^{\mathrm{BLUE}} = (\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{y} \tag{۴.۲}$$

هدف نهایی در برآورد کوچک ناحیه‌ای پیش‌گویی کمیت مورد علاقه است. به طور مثال اگر هدف ما پیش‌گویی میانگین ناحیه iام باشد، بهترین پیش‌گوی خطی نااریب (BLUP) به صورت زیر حاصل می‌شود:

$$\hat{\mu}_i^{\mathrm{BLUP}} = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}^{\mathrm{BLUE}} + \mathbf{Z}\hat{u}_i^{\mathrm{BLUP}} \tag{۵.۲}$$

با توجه به اینکه در عمل θ نامعلوم است و باید برآورد گردد، با برآورد و جایگذاری آن در روابط (۳.۲) و (۴.۲) به ترتیب بهترین برآوردگر و پیش‌گوی تجربی برای $\boldsymbol{\beta}$ و $\mathbf{u}$ حاصل می‌شوند. و در نهایت با جایگذاری آنها در رابطه (۵.۲) بهترین پیش‌گوی خطی نااریب تجربی (EBLUP) به‌دست می‌آید.

[۱] Unit level
[۲] Area level

۲.۲ مدل کوچک ناحیه ای فضایی در سطح ناحیه

در یک آمارگیری نمونه‌ای در مقیاس کلان، عموماً برآورد مستقیم برای ناحیه‌های کوچک‌تر به علت اندازه‌ی کوچک نمونه معتبر نیست. برآورد کوچک ناحیه‌ای به عنوان راه حل این مشکل، به دنبال استفاده از اطلاعات کمکی و ارائه روش‌های برآورد غیرمستقیم برای پارامترهای کوچک‌ناحیه‌ها است. یک منبع مفید اطلاعات کمکی برای برآورد غیرمستقیم کوچک ناحیه‌ای، اطلاعات مربوط به وابستگی فضایی متغیر مورد مطالعه در ناحیه‌های مجاور است. کرسی (۱۹۹۳) داده‌های فضایی را در سه گروه عمده «داده‌های زمین‌آماری[۳]»، «داده‌های مشبکه‌ای[۴]» و «الگوهای نقطه‌ای[۵]» دسته‌بندی نمود. اگر در آمارگیری‌های مرتبط با برآورد کوچک ناحیه‌ای، داده‌ها در سطح ناحیه گردآوری شده و پارامتر مورد علاقه به ناحیه‌ها اختصاص یابد، نوع داده‌های فضایی مشبکه‌ای خواهد بود.

لحاظ نمودن وابستگی فضایی در مدل‌های برآورد کوچک ناحیه‌ای، ابتدا توسط سالواتی (۲۰۰۴) مطرح شد. وی مدل‌های اتورگرسیو همزمان (SAR) را برای این هدف به کار برد و بهترین پیشگوی نااریب خطی تجربی[۶] (EBLUP) را برای مدل فضایی پیشنهادی خود به دست آورد. به طور همزمان سینگ و همکاران (۲۰۰۵) نیز مدل FH را برای استفاده از وابستگی فضایی تعمیم داده و بطور مشابه از مدل‌های SAR برای این هدف استفاده کردند. مدل ارائه شده برای استفاده از مشابهت فضایی بین نواحی مجاور تعمیم مستقیمی از مدل (۱.۲) فی‌ـ هریوت است. مطابق پیشنهاد پراتسی و سالواتی (۲۰۰۸) مدل (۱.۲) به صورت زیر تغییر می‌یابد:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + v_i + e_i, \quad i = 1, ..., D \tag{۶.۲}$$

که در آن برای v_i به عنوان اثر تصادفی ناحیه iاُم به علت وجود همبستگی فضایی بین نواحی، می‌توان یک مدل SAR مرتبه اول به صورت $v_i = \sum_{j \neq i} s_{ij} v_i + u_i$ در نظر گرفت. در این رابطه $0 < \rho < 1$ ضریب همبستگی فضایی و s_{ij} وزن‌های شباهت فضایی و درایه‌های ماتریس $\mathbf{S}$ هستند که حاصل تبدیل استاندارد شده سطری ماتریس مشابهت $\mathbf{S}^*$ است. درایه‌های ماتریس $\mathbf{S}^*$ برای دو ناحیه هم‌جوار برابر ۱ و برای سایر حالات صفر است. به این ترتیب مدل در نظر گرفته شده برای اثرات تصادفی همبسته فضایی را می‌توان به صورت $\mathbf{v} = (\mathbf{I} - \rho\mathbf{S})^{-1}\mathbf{u}$ نوشت. در نتیجه نمایش ماتریسی آن بفرم زیر خواهد بود:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{S})^{-1}\mathbf{u} + \mathbf{e} \tag{۷.۲}$$

که با تعریف جدید $\mathbf{Z} = (\mathbf{I} - \rho\mathbf{S})^{-1}$ می‌توان آن را به فرم کلی رابطه (۲.۲) یعنی $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ نشان داد. با این تعریف تمام مراحل برآورد و پیش‌گویی انجام شده برای مدل فی‌ـ هریوت برای این مدل نیز قابل انجام است.

۳.۲ برآورد استوار در مدل‌های کوچک ناحیه‌ای

برآوردگرها و پیش‌گوهای ارائه شده در بخش قبل تحت تاثیر داده‌های دورافتاده قرار می‌گیرند. این چالش در مورد تمام روش‌های آماری وجود دارد و روش‌های استوار برای مقابله با این مشکل بسیار توسعه یافته‌اند. تحلیل استوار در برآورد کوچک ناحیه‌ای در چند سال اخیر بسیار مورد توجه قرار گرفته است. در برخی از تلاش‌ها پیش‌فرض نرمال بودن در مدل با توزیع‌های دم کلفت‌تر جایگزین شده است. داتا و لاهیری (۱۹۹۵) توزیع کوشی بل و هوانگ (۲۰۰۶) توزیع t و فراز و مورات (۲۰۱۲) و دیالو

[۳] Geostatistical Data
[۴] Lattice Data
[۵] Point Patterns
[۶] Emprical Best Linear Unbiased Predictor

(۲۰۱۴) توزیع چوله‌نرمال را برای این هدف به کار بردند. البته به نوعی این روش‌ها را نمی‌توان در دایره روش‌های استوار گنجاند، چون در یک روش استوار پیش‌فرض‌های اصلی پابرجا می‌مانند و تلاش می‌شود روش‌هایی پیشنهاد گردد که بتوانند در مقابل تخطی از آنها استوار بمانند.

مدل‌های آمیخته خطی چارچوب رایج برای مدل‌بندی در برآورد کوچک ناحیه‌ای هستند. از این نظر سینها و رائو (۲۰۰۹) با الگوگیری از روش‌های استوار ارائه شده برای تحلیل مدل‌های آمیخته خطی، روشی را برای برآورد و پیش‌گویی استوار در مدل‌های کوچک ناحیه‌ای پیشنهاد دادند. پیشنهاد آنها بسیار مورد توجه محققین قرار گرفت و محور اصلی مطالعات آتی در برآورد کوچک ناحیه‌ای قرار گرفت. اشمید و مونیش (۲۰۱۴) روش پیشنهادی آنها را برای مدل کوچک‌ناحیه‌ای فضایی در سطح واحد آماری توسعه دادند. اما چون این روش برای مدل‌های کوچک‌ناحیه‌ای در سطح واحد آماری ارائه شده بود، در عمل برای مدل‌های در سطح ناحیه قابل کاربرد نبود. رائو و مولینا (۲۰۱۵) روش آنها را برای برآورد در مدل‌های کوچک‌ناحیه‌ای در سطح ناحیه توسعه دادند و وارنهولز (۲۰۱۶) الگوریتم‌های محاسباتی موردنیاز را بهبود بخشید. در ادامه به علت قالب مشترک، برآورد و پیش‌گویی استوار به روش سینها و رائو (۲۰۰۹) تشریح می‌گردد.

برای برآورد و پیش‌گویی استوار در مدل‌های آمیخته خطی، فلنر (۱۹۸۶) روشی را ارائه نمود که در آن معادلات برآورد با نسخه‌هایی استوار جایگزین می‌شدند. هاگینز (۱۹۹۳) برآورد با استفاده از استوارسازی معادلات حاصل از درستنمایی را پیشنهاد نمود. با الگوگیری از روش‌های فوق، سینها و رائو (۲۰۰۹) برای برآورد استوار $\boldsymbol{\beta}$ و θ در مدل (۲.۲)، ابتدا معادلات همزمان برآورد را با مشتق‌گیری از تابع لگ‌درستنمایی همزمان و به صورت زیر به دست آوردند:

$$\mathbf{X}^\top \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = 0 \tag{۸.۲}$$

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \mathrm{tr}(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}) = 0 \tag{۹.۲}$$

بدیهی است که تاثیر داده‌های دورافتاده در معادلات فوق در جمله‌ی $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ می‌باشد. استوارسازی معادلات برآورد فوق با جایگزینی جمله مذکور با معادلی است که تاثیرپذیری آن از داده‌های دورافتاده کمتر باشد. چنین ایده‌ای با رهیافت ام‌ـ برآورد در مدل‌های رگرسیونی قابل تحقق است. هاگینز (۱۹۹۳) در برآورد استوار مدل‌های آمیخته خطی جایگزینی $\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ را با معادل استوار شده آن یعنی $\mathbf{V}^{-\frac{1}{2}}\psi(\mathbf{V}^{-\frac{1}{2}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}))$ پیشنهاد نمود. در این رابطه $\psi(.)$ یک تابع تاثیر است که اثر مقادیر بزرگ $\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ را که مقادیر دورافتاده هستند، کاهش می‌دهد. سینها و رائو (۲۰۰۹) به هدف ساده‌سازی پیشنهاد دیگری را برای استوارسازی معادله (۸.۲) به صورت

$$\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{U}^{\frac{1}{2}}\psi(\mathbf{U}^{-\frac{1}{2}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})) = 0 \tag{۱۰.۲}$$

مطرح کردند، که در آن $\mathbf{U} = \mathrm{diag}(\mathbf{V})$ ماتریسی قطری و هم‌اندازه با $\mathbf{V}$ است که تنها عناصر قطر اصلی آن را شامل می‌شود. توابع تاثیر مختلفی در تحلیل استوار مدل‌های آمیخته خطی به کار برده شده‌اند. سینها و رائو (۲۰۰۹) و چندین محقق دیگر استفاده از تابع $\psi(.)$ هابر را که به صورت $\psi(z) = z.\min(1, \frac{b}{|z|})$ تعریف می‌شود، پیشنهاد داده‌اند. ثابت b در این تابع معمولاً برابر ۱/۳۴ فرض می‌شود. برخی دیگر از محققین مانند هاگینز (۱۹۹۳) استفاده از تابع دو وزنی توکی را توصیه کرده‌اند.

به همین ترتیب با فرض $\mathbf{r} = \mathbf{U}^{-\frac{1}{2}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ معادله (۹.۲) طبق پیشنهاد سینها و رائو (۲۰۰۹) به صورت

$$\psi(\mathbf{r})^\top \mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{U}^{\frac{1}{2}}\psi(\mathbf{r}) - k\mathrm{tr}(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}) = 0 \tag{۱۱.۲}$$

استوار می‌شود، که در این رابطه ضریب k به صورت $k = \mathrm{E}(\psi^{۲}(z))$ بدست می‌آید که z دارای توزیع نرمال استاندارد است. توضیحات بیشتر در این خصوص را می‌توانید در ریچاردسون و ولش (۱۹۹۵) ببینید.

فرض کنید $\hat{\theta}^M$ و $\hat{\boldsymbol{\beta}}^M$ برآوردهای استوار حاصل از معادلات (۱۰.۲) و (۱۱.۲) باشند. هرچند در مقالات دیگر نمادهای دیگری برای نشان دادن برآوردها و پیش‌گویی‌های استوار استفاده شده اند، اما در این مقاله از بالانویس نمادین M که اشاره به استفاده از رویکرد ام‌ـ برآورد دارد، استفاده می‌کنیم. پیشنهاد سینها و رائو برای پیش‌گویی استوار $\mathbf{u}$ استفاده از نسخه استوار شده معادله برآورد پیشنهادی فلنر (۱۹۸۶) به صورت

$$\mathbf{Z}^{\top}\mathbf{R}^{-\frac{۱}{۲}}\psi(\mathbf{R}^{-\frac{۱}{۲}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}))-\mathbf{G}^{-\frac{۱}{۲}}\psi(\mathbf{G}^{-\frac{۱}{۲}}\mathbf{u})=۰ \qquad (۱۲.۲)$$

می‌باشد. و در نهایت و با فرض اینکه $\hat{\mathbf{u}}^M$ پیش‌گویی استوار $\mathbf{u}$ باشد، پیش‌گویی نهایی میانگین ناحیه مورد نظر را می‌توان به صورت زیر محاسبه نمود:

$$\hat{\mu}_i^{\text{M-REBLUP}} = \mathbf{x}_i^{\top}\hat{\boldsymbol{\beta}}^M + \mathbf{Z}\hat{u}_i^M \qquad (۱۳.۲)$$

سینها و رائو (۲۰۰۹) برای بدست آوردن برآوردها و پیش‌گویی‌های استوار θ و $\boldsymbol{\beta}$ و $\mathbf{u}$ از معادلات استوار شده فوق، از الگوریتم نیوتن‌ـرافسون استفاده کردند. شوچ (۲۰۱۲) و چترچی (۲۰۱۲) به مشکلات عددی این الگوریتم اشاره و آن را فاقد کارایی لازم دانستند. شوچ (۲۰۱۲) استفاده از از الگوریتم کمترین توانهای دوم بازموزون تکراری [۷] (IRWLS) را برای برآورد $\boldsymbol{\beta}$ پیشنهاد داد. چترچی (۲۰۱۲) الگوریتم تکراری توابع نقطه‌ثابت را برای برآورد پارامترهای واریانس پیشنهاد نمود. وارنهولز (۲۰۱۶) برای برآورد $\boldsymbol{\beta}$ و پیش‌گویی $\mathbf{u}$ در مدل‌های در سطح ناحیه الگوریتمی را با بهره‌گیری از ایده نمایش شبه‌خطی معادلات برآورد چمبرز و همکاران (۲۰۱۱) پیشنهاد نمود که برای مدل‌های در سطح ناحیه کارایی دارد.

۳ جی‌ام‌ـ برآوردگرها

نتایج حاصل از روش‌های استوار ارائه شده برای برآورد کوچک ناحیه‌ای در مقابل داده‌های دورافتاده در متغیر پاسخ استوارند. اما همچنان اثر مقادیر دورافتاده و غیرمعمول در متغیرهای توضیحی کنترل نشده‌اند. در روش‌های مبتنی بر رویکرد ام‌ـ برآورد، تابع تاثیر وارد شده در معادلات برآورد تاثیر مقادیر دورافتاده در متغیر پاسخ را کنترل می‌کند یا اصطلاحاً کراندار می‌کند. اما دسته دیگری از روش‌ها استوار ارائه شده‌اند که علاوه بر متغیر پاسخ تاثیر داده‌های دورافتاده در متغیرهای تبیینی را نیز تحت کنترل در می‌آورند و از این نظر روش‌های با تاثیر کراندار گفته می‌شوند (ریچاردسون و ولش ، ۱۹۹۵). جی‌ام برآوردگرها شناخته شده ترین روش برآورد با تاثیر کراندار هستند که توسط ریچاردسون (۱۹۹۷) برای مدل‌های آمیخته خطی توسعه داده شده‌اند.

روش‌های رگرسیونی استوار نخستین بار توسط هابر (۱۹۷۳) معرفی شدند. در این روش‌ها که تعمیمی از روش کمترین توان‌های دوم هستند، به جای کمینه‌سازی توان‌های دوم خطا، تابع جایگزینی از آنها مانند $\rho(.)$ را کمینه می‌کنند. این تابع به گونه‌ای انتخاب می‌شود که تاثیر مشاهدات دورافتاده را کنترل کند. این کلاس از برآوردگرها به ام‌ـ برآوردگر مشهور شدند. بطور مثال تابع ρ در برآورد کمترین توان‌های دوم به صورت $\rho(z) = \sum z^{۲}$ می‌باشد. یکی از تابع‌های ρ شناخته شده، تابع هابر است که به

[۷] Iteratively reweighted least squares

صورت

$$\rho(z)=\begin{cases}\frac{1}{2}z^2 & z<0\\ b|z|-\frac{1}{2}b^2 & z\geq 0\end{cases} \qquad (۱.۳)$$

تعریف می‌شود. تابع تاثیر یا تابع ψ که در بخش قبل دیدیم، مشتق تابع ρ است. ام‌ـ برآوردگر رگرسیونی مطابق تعریف هابر (۱۹۷۳) آن مقداری از $\hat{\beta}$ است که عبارت $\sum\rho(\frac{r_i(\hat{\beta})}{\hat{\sigma}})$ را کمینه کند. با مشتق گیری از این رابطه و برابر قرار دادن آن با صفر معادله برآورد استوار رگرسیونی به صورت $\sum x_i\psi(\frac{r_i(\hat{\beta})}{\hat{\sigma}})=0$ حاصل می‌شود. مباحث بیشتر در خصوص ام‌ـ برآوردگرها و سایر برآوردگرهای استوار رگرسیونی را می‌توانید در همپل و همکاران (۱۹۸۶) و روسیو و لروی (۱۹۸۷) ببینید.

در روش‌های رگرسیونی استوار علاوه بر داده‌های دورافتاده در متغیر پاسخ، مقادیر غیرمعمول و بزرگ در متغیرهای تبیینی، که اصطلاحاً نقاط اهرمی گفته می‌شوند، نیز میتوانند باعث نتایج نادرست شوند. رهیافت جدیدی که اثر نقاط اهرمی را نیز کنترل می‌کند، تعمیمی از ام‌ـ برآوردهاست که بعدها به ام‌ـ برآورد تعمیم‌یافته یا اصطلاحاً جی‌ام‌ـ برآورد مشهور شد. محققین پیشنهادات مختلفی برای توسعه معادلات ام‌ـ برآورد در این راستا ارائه دادند که در همپل و همکاران (۱۹۸۶) گردآوری و تشریح شده است. در تمام این روش‌ها وزن‌هایی به معادله ام‌ـ برآورد افزوده می‌شود که تاثیر نقاط اهرمی را نیز کنترل می‌کنند. یکی از این پیشنهادات توسط مالوز (۱۹۷۵) ارائه شد که در آن معادلات برآورد رگرسیونی، به صورت

$$\sum w_i x_i\psi(\frac{r_i(\hat{\beta})}{\hat{\sigma}})=0$$

تغییر می‌یابند. برای وزن‌ها، پیشنهادهای مختلفی مطرح شده است که در این مقاله از وزن‌های پیشنهادی سیمپسون و همکاران (۱۹۹۲) استفاده شده است.

با توجه به تاثیری که مقادیر دورافتاده در متغیرهای توضیحی می‌تواند بر برآورد و پیش‌گویی در مدل‌های آماری بگذارد، استفاده از جی‌ام‌ـ برآوردگرها در مدل‌های مختلف آماری بسیار مورد توجه قرار گرفته است. ریچاردسون (۱۹۹۷) جی‌ام برآوردگرها را برای برآورد استوار با تاثیر کراندار در مدل‌های آمیخته خطی توسعه داد و ولش و ریچاردسون (۱۹۹۷) چارچوب کلی استفاده از جی‌ام‌ـ برآوردگرها را ارائه نمودند.

۴ تحلیل استوار با تاثیر کراندار در برآورد کوچک ناحیه‌ای فضایی

با توجه به توضیحات فوق و دقت در معادلات برآورد استوار کنونی در مدل‌های کوچک ناحیه‌ای، می‌بینیم که این برآوردگرها و پیشگوها برپایه منطق امبرآورد ساخته شده‌اند. در این بخش کلیه معادلات برآورد ارائه شده توسط سینها و رائو (۲۰۰۹) به معادلات منطبق با رویکرد جی‌امبرآورد تعمیم داده می‌شود. برای این هدف، معادلات (۱۰.۲) و (۱۱.۲) را با معادلات زیر جایگزین می‌کنیم:

$$\mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{U}^{\frac{1}{2}}\psi(\mathbf{U}^{-\frac{1}{2}}\mathbf{W}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}))=0 \qquad (۱.۴)$$

(۲.۴)

$$\psi\left(\mathbf{U}^{-\frac{1}{2}}\mathbf{W}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)^\top\mathbf{W}\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{U}^{\frac{1}{2}}\mathbf{W}\psi\left(\mathbf{U}^{-\frac{1}{2}}\mathbf{W}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)-\mathrm{tr}(K\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta})=0$$

الگوی مورد استفاده در این روابط جی‌ام‌ـ برآوردگرهای مالوز است که توسط ولش و ریچاردسون (۱۹۹۷) به صورت مشابه مورد استفاده قرار گرفته است. پس از برآورد پارامترهای اثرات ثابت و واریانس، پیش‌گویی استوار اثرات تصادفی $\mathbf{u}$ نیز با استفاده از

جی‌ام‌ـ برآوردگرها قابل انجام است.

بطور مشابه الگوی جی‌ام‌ـ برآوردگرها را می‌توان جایگزین ام‌ـ برآوردگرها در برابری (۱۲.۲) نمود:

$$\mathbf{Z}^\top \mathbf{W}\mathbf{R}^{-\frac{1}{2}}\psi(\mathbf{R}^{-\frac{1}{2}}\mathbf{W}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})) - \mathbf{G}^{-\frac{1}{2}}\psi(\mathbf{G}^{-\frac{1}{2}}\mathbf{u}) = 0 \qquad (۳.۴)$$

توجه داشته باشید که در برابری فوق، تاثر داده‌های دورافتاده در $\mathbf{X}$ تنها در بخش اول معادله است ودر نتیجه الگوی جی‌ام‌ـ برآورد تنها در این بخش اعمال گردیده است. در نهایت با یافتن برآوردهای استوار $\hat{\theta}^{GM}$ و $\hat{\boldsymbol{\beta}}^{GM}$ و نیز پیش‌گویی استوار $\hat{\mathbf{u}}^{GM}$، پیش‌گویی استوار کمیت مورد علاقه به صورت زیر حاصل می‌شود:

$$\hat{\mu}_i^{\text{GM-REBLUP}} = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}^{GM} + \mathbf{Z}\hat{u}_i^{GM} \qquad (۴.۴)$$

۱.۴ الگوریتم‌های محاسباتی برای روش پیشنهادی

استفاده از رهیافت جی‌ام‌ـ برآورد، سبب تغییراتی در معادلات برآورد نسبت به ام‌ـ برآوردها می‌گردد. با وجود این تغییرات همچنان استفاده از الگوریتم‌های قبلی در محاسبه ام‌ـ برآوردها امکان‌پذیر است. سینها و رائو (۲۰۰۹) از الگوریتم تکراری نیوتن‌ـ رافسون را برای حل معادلات برآورد استفاده کردند، که اولاً برای مدل‌های در سطح ناحیه قابل استفاده نبود و ثانیاً در مطالعات بعدی به کارایی آن ایرادات جدی وارد شد.

چترچی (۲۰۱۲) با برشمردن معایب الگوریتم نیوتن‌ـ رافسون، استفاده از الگوریتم توابع نقطه ثابت را برای برآورد پارامترهای واریانس θ پیشنهاد داد. وارنهولز (۲۰۱۶) الگوریتم جدیدی را برای برآورد استوار و پیش‌گویی استوار در مدل‌های کوچک ناحیه‌ای ارائه نمود که در آن از ایده نمایش شبه‌خطی معادلات برآورد چمبرز و همکاران (۲۰۱۱) استفاده شده بود. در این مقاله از چارچوب الگوریتم پیشنهادی وارنهولز (۲۰۱۶) استفاده می‌کنیم. برای این هدف نخست نمایش شبه‌خطی معادلات برآورد جدید را بدست آورده و سپس روش برآورد هر جزء به ترتیب می‌آید.

۱.۱.۴ نمایش شبه‌خطی

ایده نمایش شبه‌خطی چمبرز و همکاران (۲۰۱۱) در واقع راهکاری برای محاسبه میانگین مربع خطاهای پیش‌گویی (MSPE) بود. در این روش پیش‌گویی نهایی ارائه شده برای کمیت مورد علاقه به صورت تابعی خطی از بردار مشاهدات نمایش داده می‌شود. وارنهولز (۲۰۱۶) این فرم نمایش را برای ساختن الگوریتم‌های برآورد برای $\boldsymbol{\beta}$ و $\mathbf{u}$ پیشنهاد داد. نمایش شبه‌خطی پیشگوی نهایی استوار را می‌توان به صورت

$$\hat{\mu}_i^{\text{GM-REBLUP}} = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}^{GM} + \mathbf{Z}\hat{u}_i^{GM} = (\mathbf{x}_i^\top \mathbf{A} + \mathbf{Z}_i^\top \mathbf{B}(\mathbf{I}-\mathbf{X}\mathbf{A})\mathbf{y} = \tilde{\mathbf{c}}_i\mathbf{y} \qquad (۵.۴)$$

که در واقع با جایگزینی روابط $\hat{\boldsymbol{\beta}}^{GM} = \mathbf{A}\mathbf{y}$ و $\hat{u}^{GM} = \mathbf{B}(\mathbf{I}-\mathbf{X}\mathbf{A})\mathbf{y} = \mathbf{B}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{GM})$ حاصل شده اند. با محور قرار دادن این روابط، ماتریس‌های $\mathbf{A}$ و $\mathbf{B}$ بطور مستقیم با استفاده از معادلات برآورد (۱.۴) و (۳.۴) قابل محاسبه هستند. برای محاسبه $\mathbf{A}$ ، ابتدا برای سادگی محاسبات و کنار گذاشتن تابع $\psi(.)$، رابطه (۱.۴) را به صورت زیر می‌نویسیم:

$$\mathbf{X}^\top \mathbf{W}\mathbf{V}^{-1}\mathbf{U}^{\frac{1}{2}}\mathbf{D}\mathbf{U}^{-\frac{1}{2}}\mathbf{W}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = 0 \qquad (۶.۴)$$

که در آن $\mathbf{D}$ ماتریسی قطری است که عناصر قطر اصلی آن به صورت زیر تعریف می‌شوند:

$$(\mathbf{D})_{ii} = \psi\left(U_i^{-\frac{1}{2}} w_i^{-1}(y_i - \mathbf{x}_i\boldsymbol{\beta})\right)\left(U_i^{-\frac{1}{2}} w_i^{-1}(y_i - \mathbf{x}_i\boldsymbol{\beta})\right)^{-1}$$

به این ترتیب با توجه به اینکه $\mathbf{U}^{\frac{1}{2}}$ ، $\mathbf{U}^{-\frac{1}{2}}$ و $\mathbf{D}$ همگی ماتریس‌های قطری هستند، (۶.۴) را می‌توان نوشت:

$$\begin{aligned}
&\mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{D}\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\
\Rightarrow\quad &\mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{D}\mathbf{W}^{-1}\mathbf{y} = \mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{D}\mathbf{W}^{-1}\mathbf{X}\boldsymbol{\beta} \\
\Rightarrow\quad &\boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{D}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{D}\mathbf{W}^{-1}\mathbf{y} \\
\Rightarrow\quad &\mathbf{A} = (\mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{D}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{D}\mathbf{W}^{-1} \qquad (۷.۴)
\end{aligned}$$

برای محاسبه $\mathbf{B}$ از معادله (۳.۴)، مشابه آنچه در بخش قبل انجام دادیم ماتریس‌های وزن را به صورت زیر تعریف می‌کنیم:

$$(\mathbf{D}_1)_{ii} = \psi\left(R_i^{-\frac{1}{2}} w_i^{-1}(y_i - \mathbf{x}_i\boldsymbol{\beta} - Z_i\mathbf{u})\right)\left(R_i^{-\frac{1}{2}} w_i^{-1}(y_i - \mathbf{x}_i\boldsymbol{\beta} - Z_i\mathbf{u})\right)^{-1}$$

$$(\mathbf{D}_2)_{ii} = \psi\left(G_i^{-\frac{1}{2}}\mathbf{u}\right)\left(G_i^{-\frac{1}{2}}\mathbf{u}\right)^{-1}$$

حال (۳.۴) را می‌توان نوشت:

$$\mathbf{Z}^\top\mathbf{W}\mathbf{R}^{-\frac{1}{2}}\mathbf{D}_1\mathbf{R}^{-\frac{1}{2}}\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) - \mathbf{G}^{-\frac{1}{2}}\mathbf{D}_2\mathbf{G}^{-\frac{1}{2}}\mathbf{u} = 0$$

که با توجه به قطری بودن $\mathbf{R}^{-\frac{1}{2}}$، $\mathbf{D}_1$، $\mathbf{G}^{-\frac{1}{2}}$ و $\mathbf{D}_2$ می‌توان نوشت:

$$\begin{aligned}
&\mathbf{Z}^\top\mathbf{W}\mathbf{R}^{-1}\mathbf{D}_1\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) - \mathbf{G}^{-1}\mathbf{D}_2\mathbf{u} = 0 \\
\Rightarrow\quad &\mathbf{Z}^\top\mathbf{W}\mathbf{R}^{-1}\mathbf{D}_1\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{Z}^\top\mathbf{W}\mathbf{R}^{-1}\mathbf{D}_1\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1}\mathbf{D}_2)\mathbf{u} \\
\Rightarrow\quad &\mathbf{u} = (\mathbf{Z}^\top\mathbf{W}\mathbf{R}^{-1}\mathbf{D}_1\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1}\mathbf{D}_2)^{-1}\mathbf{Z}^\top\mathbf{W}\mathbf{R}^{-1}\mathbf{D}_1\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
\Rightarrow\quad &\mathbf{B} = (\mathbf{Z}^\top\mathbf{W}\mathbf{R}^{-1}\mathbf{D}_1\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1}\mathbf{D}_2)^{-1}\mathbf{Z}^\top\mathbf{W}\mathbf{R}^{-1}\mathbf{D}_1\mathbf{W} \qquad (۸.۴)
\end{aligned}$$

۲.۱.۴ برآورد و پیش‌گویی نهایی

با دقت در رابطه به دست آمده برای $\mathbf{A}$ در (۷.۴) مشاهده می‌شود که $\mathbf{A}$ خود تابعی از پارامتر $\boldsymbol{\beta}$ است (ماتریس $\mathbf{D}$ تابعی از $\boldsymbol{\beta}$ می‌باشد.) لذا برآورد $\boldsymbol{\beta}$ را می‌توان نوشت:

$$\hat{\boldsymbol{\beta}} = \mathbf{A}(\hat{\boldsymbol{\beta}})\mathbf{y} = (\mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{D}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{V}^{-1}\mathbf{D}\mathbf{W}^{-1}\mathbf{y}$$

در واقع نمایش فوق یک تابع نقطه‌ثابت[۸] از است. از این نظر وارنهولز (۲۰۱۶) استفاده از یک الگوریتم تکراری IRWLS را با استفاده از نمایش فوق پیشنهاد داد. به این ترتیب می‌توان از رابطه تکراری $\hat{\boldsymbol{\beta}}^{(m+1)} = \mathbf{A}(\hat{\boldsymbol{\beta}}^{(m)})\mathbf{y}$ برای برآورد استفاده نمود.

فرآیند مورد اجرا برای پیش‌گویی $\mathbf{u}$ مشابه بخش قبل به یک تابع نقطه ثابت از $\mathbf{u}$ منجر می‌شود. برابری (۸.۴) نشان می‌دهد که B خود تابعی از $\mathbf{u}$ است. لذا می‌توان برای پیش‌گویی $\mathbf{u}$ از رابطه تکراری $\hat{\mathbf{u}}^{(m+1)} = \mathbf{B}(\hat{\mathbf{u}}^{(m)})\mathbf{y}$ استفاده کرد. به عنوان مقدار اولیه در الگوریتم پیش‌گویی می‌توان از نسخه استوار شده برآوردگر معمول آن بهره برد.

[۸] Fixed Point Function

برای برآورد θ نیز با افزودن $\theta^{-1}\theta$ به داخل trace در برابری (۲.۴) داریم:

$$\psi\left(\mathbf{U}^{-\frac{1}{2}}\mathbf{W}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)^{\top}\mathbf{W}\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{U}^{\frac{1}{2}}\mathbf{W}\psi\left(\mathbf{U}^{-\frac{1}{2}}\mathbf{W}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)=\text{tr}\left(K\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}(\theta\theta^{-1})\right)$$

$$\begin{aligned}\Rightarrow\theta &= \text{tr}\left(K\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}(\theta^{-1})\right)^{-1}\\ &\quad\psi\left(\mathbf{U}^{-\frac{1}{2}}\mathbf{W}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)^{\top}\mathbf{W}\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{U}^{\frac{1}{2}}\mathbf{W}\psi\left(\mathbf{U}^{-\frac{1}{2}}\mathbf{W}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)\end{aligned}$$

به این ترتیب یک معادله بازگشتی به فرم $\theta=c(\theta)$ حاصل می‌شود.

۵ مطالعه شبیه‌سازی

مدل مورد استفاده برای تولید داده‌های شبیه‌سازی، مدلی با یک متغیر کمکی به صورت زیر می‌باشد

$$y_i=\beta_0+\beta_1 x_i+v_i+e_i,\quad i=1,...,D \tag{۱.۵}$$

در این مدل D تعداد ناحیه‌ها برابر ۲۵، مقدار ضرایب مدل برابر ۱۰۰ و ۱۰ و واریانس‌های σ_e^2 و σ_u^2 هر دو برابر ۲۵ فرض شده‌اند. در حالتی که داده‌ها فاقد آلودگی باشند x_iها از توزیع نرمال با میانگین ۱۰ و واریانس ۱۰ تولید می‌شوند. و برای سناریوهای آلودگی درصدی از این مقادیر با مقدار ثابت ۲۰۰ جمع می‌شوند تا گویای یک مقدار دورافتاده باشند. به همین ترتیب e_iها از توزیع نرمال با میانگین صفر و واریانس σ_e^2 تولید می‌شوند. برای آلوده ساختن مقادیر y_i نیز مقدار ثابت ۵۰۰ به جمله خطای مدل افزوده می‌شود. مقادیر v_i نیز پس از تولید مقادیر u_i از توزیع نرمال با میانگین صفر و واریانس σ_u^2 و محاسبه ماتریس S از رابطه $\mathbf{v}=(\mathbf{I}-\rho\mathbf{S})^{-1}\mathbf{u}$ تولید می‌شوند. مقدار ρ نیز برابر ۰/۵ فرض می‌شود.

چهار سناریوی مختلف برای مقایسه عملکرد برآوردگرها و پیشگوی پیشنهادی در نظر گرفته می‌شود. در سناریوی نخست داده‌های پاک مورد استفاده قرار می‌گیرند، اما در ۳ سناریوی دیگر درصدی از مقادیر با مقادیر دورافتاده مطابق روش بیان شده، جایگزین می‌شوند. به این ترتیب که در سناریوی دوم تنها داده دورافتاده در متغیر پاسخ ایجاد می‌شود. در مقابل در سناریو سوم مقادیر متغیر توضیحی آلوده می‌شوند. و نهایتاً در چهارمین سناریو آلودگی در هر دو بخش متغیر پاسخ و متغیر توضیحی وارد می‌شود. درصد آلودگی در این مطالعه ۲۰ درصد در نظر گرفته شده است.

فرآیند طراحی شده فوق ۱۰۰ بار تکرار شده و در هر بار برآورد و پیش‌گویی‌ها به سه روش بهترین پیشگوی خطی تجربی (EBLUP)، روش استوار برمبنای ام‌ـبرآورد (M-REBLUP) و نهایتاً روش استوار پیشنهادی برپایه جی‌ام‌ـبرآورد (GM-REBLUP) انجام می‌شود. شاخص مورد استفاده برای ارزیابی دقت برآوردگرها و پیشگوها میانگین قدر مطلق اریبی برآورد یا پیش‌گویی است.

در جدول ۱ نتایج مربوط به برآورد پارامترهای رگرسیونی آمده است. در سناریوی اول مطابق انتظاری که از روش‌های استوار داریم، کیفیت برآورد تا حد زیادی به برآوردگرهای EBLUP نزدیک است. به عبارتی حتی اگر احتمال حضور داده دورافتاده کم باشد، استفاده از روش‌های استوار نتایج قابل اتکایی دارد. اما در سناریوهای بعدی به وضوح برآوردگرهای EBLUP غیر قابل استفاده می‌باشد. در سناریو دوم هر دو روش‌های M-REBLUP و GM-REBLUP نتایج قابل دفاعی ارائه داده‌اند و تفاوت چشمگیری در کیفیت برآوردگرهای حاصل نداشته‌اند. اما در سناریوی سوم نتایج کاملاً گویای کنترل موفق اثر داده‌های دورافتاده در

جدول ۱: نتایج شبیه‌سازی: میانگین قدر مطلق اریبی برآورد پارامترهای رگرسیونی برای سه روش

سناریو	پارامتر برآوردی	روش برآورد		
		EBLUP	M-REBLUP	GM-REBLUP
داده های پاک	β_0	۶/۰۲۳	۶/۶۰۵	۶/۵۸۹
	β_1	۱/۱۱۸	۱/۱۴۶	۱/۱۸۴
داده دورافتاده در Y	β_0	۴۲/۱۶۵	۹/۲۳۸	۹/۳۳۴
	β_1	۴/۴۵۱	۱/۹۲۵	۱/۸۹۶
داده دورافتاده در X	β_0	۵۳/۸۱۷	۳۵/۳۰۱	۱۰/۳۹۴
	β_1	۵/۶۰۸	۴/۶۸۴	۱/۹۶۷
داده دورافتاده در X و Y	β_0	۷۰/۳۲۹	۴۱/۲۲۸	۱۰/۸۳۵
	β_1	۶/۲۵۵	۴/۸۹۰	۲/۳۳۸

جدول ۲: نتایج شبیه‌سازی: میانگین قدر مطلق اریبی برآورد ضریب همبستگی فضایی برای سه روش

سناریو	پارامتر برآوردی	روش برآورد		
		EBLUP	M-REBLUP	GM-REBLUP
داده های پاک	ρ	۰/۰۹۲	۰/۰۹۷	۰/۰۹۶
داده دورافتاده در Y	ρ	۰/۲۷۹	۹/۱۲۳	۰/۱۲۹
داده دورافتاده در X	ρ	۰/۲۲۵	۰/۱۹۶	۰/۱۳۶
داده دورافتاده در X و Y	ρ	۰/۳۶۲	۰/۲۳۷	۰/۱۶۷

روش GM-REBLUP است. در این بخش برآورد به روش M-REBLUP نتوانسته است اثر داده‌های دورافتاده در متغیر توضیحی را کنترل نماید. در سناریوی چهارم که منشاء داده‌های دورافتاده در هر دوی متغیرهای پاسخ و توضیحی می‌باشد، دوباره برآوردهای حاصل از روش GM-REBLUP به خوبی عمل کرده و نشان داده‌اند توان کنترل هر دو نوع داده دورافتاده را دارا هستند.

بطور مشابه نتایج برآورد ضریب همبستگی فضایی در جدول ۲ عملکرد خوب برآورد حاصل به روش GM-REBLUP را در هر سه سناریوی حضور داده‌های دورافتاده نشان می‌دهد. اما برآورد M-REBLUP در سناریوهایی که داده‌های دورافتاده در متغیر توضیحی داشته‌ایم، عملکرد خوبی نداشته است. در نهایت روش‌های استوار باید قادر باشند میانگین کمیت مورد علاقه در ناحیه‌ها را به خوبی پیش‌گویی کنند. جدول ۳ نتایج حاصل از مقایسه پیش‌گویی‌های انجام شده برای هر ۲۵ ناحیه را برای سه روش مورد بررسی نشان می‌دهد. شاخص میانگین قدر مطلق اریبی پیش‌گویی‌ها برای تمام ناحیه‌ها محاسبه و میانگین‌گیری شده است. همانگونه که مشاهده می‌کنید نتایج مطابق انتظار و مشابه برآوردهای انجام شده می‌باشد. نتایج پیش‌گویی حاصل از EBLUP تنها در سناریوی داده‌های پاک عملکرد خوبی را نشان می‌دهد و در سایر سناریوها ناچار به استفاده از روش‌های استوار هستیم. عملکرد پیش‌گویی‌های حاصل از M-REBLUP نیز تنها برای داده‌های دورافتاده در متغیر پاسخ جوابگو است. اما پیش‌گوی استوار GM-REBLUP برای تمام سناریوها عملکرد مناسبی داشته است.

جدول ۳: نتایج شبیه‌سازی: میانگین کل قدر مطلق اریبی پیش‌گویی ناحیه‌ها برای سه روش

سناریو	روش برآورد		
	EBLUP	M-REBLUP	GM-REBLUP
داده های پاک	۲۳/۳۵	۲۴/۸۶	۲۴/۷۴
داده دورافتاده در Y	۱۸۹/۳۱	۳۶/۵۵	۳۶/۷۲
داده دورافتاده در X	۱۶۱/۷۵	۱۲۰/۲۳	۴۳/۶۰
داده دورافتاده در X و Y	۲۳۶/۱۱	۱۶۹/۷۱	۴۸/۸۳

بحث و نتیجه‌گیری

روش‌های برآورد کوچک ناحیه‌ای در سال‌های اخیر توجه زیادی را به خود جلب نموده و پیشرفت‌های چشمگیری نیز داشته است. شاید علت اصلی این توجه و پیشرفت، نیاز رو به رشد به ارائه برآوردهای دقیق در سطح کوچک‌ناحیه‌ها باشد. در روزگاران پیشین، کاربران عمدتاً به برآوردهای کشوری و برآوردهای مربوط به ناحیه‌های جغرافیایی عمده قانع بودند (سینها و رائو ، ۲۰۰۹). اما امروزه با وضعیت کاملاً متفاوتی روبرو هستیم و برنامه‌ریزان و سیاست‌گذاران برای تصمیم‌گیری به برآوردهایی در سطح ناحیه‌های کوچکتر نیاز دارند. اما چالشی که همواره کاربران این روش‌ها را نگران کرده است، تاثیر داده‌های دورافتاده است. بررسی اثرات داده‌های دورافتاده و یافتن روش‌هایی که در برابر این اثرات «استوار» باشند، از دیرباز مورد توجه آماردانان بوده است. هابر (۱۹۸۱) با تدوین و بنیان‌گذاری مفاهیم پایه آمار استوار در کتاب خود، نقش بسزایی در توسعه این زمینه از آمار داشت. از آن تاریخ در عموم روش‌های آماری، شاهد تلاش محققان برای یافتن روش‌های استوار در مقابل داده‌های دورافتاده یا سایر پیش‌فرض‌ها بوده‌ایم. در برآورد کوچک ناحیه‌ای فعالیت‌های سینها و رائو (۲۰۰۹) و چمبرز و زاویدیس (۲۰۰۶) اولین گام‌ها در ارائه روش‌های استوار تحلیل مدل‌های کوچک ناحیه‌ای بوده است. به دنبال سینها و رائو (۲۰۰۹) محققین بسیاری روش پیشنهادی آنها را برای برآورد استوار در مدل کوچک ناحیه‌ای توسعه دادند.

در این مقاله رویکردی جدیدی در برآورد استوار در مدل کوچک ناحیه‌ای به ویژه مدل‌های کوچک ناحیه‌ای فضایی ارائه شد. آنچه این رویکرد جدید به مسائل قبلی می‌افزاید، کنترل اثرات مقادیر دورافتاده در متغیر(های) توضیحی است. داده‌های کمکی که در قالب متغیرهای توضیحی وارد مدل‌های کوچک ناحیه‌ای می‌شوند، منشاء اصلی ارتقاء کیفیت برآوردگرها هستند. از این نظر کنترل اثرات داده‌های دورافتاده از آنها ضرورت خود را نشان می‌دهد. ابزار اصلی مورد استفاده برای این هدف جی‌ـ‌ام برآوردگرها بودند که قبلاً در مدل‌های خطی و مدل‌های آمیخته خطی بکار گرفته شده بودند. نتایج مطالعات شبیه‌سازی به خوبی نشان می‌دهد که بکارگیری جی‌ام‌ـ برآوردگرها سبب نتایج مناسبی از نظر کنترل اثر مشاهدات دورافتاده شده است.

برای مطالعات آتی توسعه این رویکرد برای مدل‌های کوچک ناحیه‌ای در سطح واحد آماری می‌تواند مورد توجه قرار گیرد. بعلاوه دسته‌ای از مدل‌های کوچک ناحیه‌ای به دنبال بهره‌گیری از اطلاعات کمکی زمان‌های قبل هستند که به مدل‌های کوچک ناحیه‌ای زمانی مشهورند. بستر ارائه شده در این مقاله به سادگی برای این مدل‌ها نیز قابل تعمیم است. همچنین نمایش شبه‌ـ‌خطی ارائه شده را می‌توان برای برآورد خطای پیش‌گویی برآورد می‌توان مورد استفاده قرار داد که قدم بعدی موردنظر نویسندگان این مقاله خواهد بود.

مراجع

Anselin, L. (1992), *patial Econometrics, Method and Models*, Kluwer Academic Publishers, Boston.

Bell, W. R. and Huang, E. T. (2006), Using the t-Distribution to Deal with Outliers in Small Area Estimation, *Proceeding of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health*

Breckling, J. and Chambers, R. (1988), M-quantiles, *Biometrika*, **75**, 761-771.

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014), Outlier Robust Small Area Estimation, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* , **76**, 47-69.

Chambers, R., Chandra, H. and Tzavidis, N. (2011) on bias-robust mean squared error estimation for pseudolinear small area estimators, *Survey Meyhodology*, **37**, 153-170.

Chambers, R. and Tzavidis, N. (2006), M-quantile Models for Small Area Estimation, *Biometrika*, **93**, 255-268.

Chatrchi, G. (2012), *Robust Estimation of Variance Components in Small Area Estimation*, MA Thesis, School of Mathematics ans Statistics, Carlton University, Ottawa, Canada.

Cressie, N. (1991), Small-area prediction of undercount using the general linear model. *In: Proceedings of the statistic symposium 90: measurement and improvement of data quality . Statistics Canada, Ottawa*, 93-105.

Cressie, N. (1993). *Statistics for Spatial*, Revised Edition, Wiley, New York.

Datta, G. S. and Lahiri, P. (1995), Robust Hierarchical Bayes Estimation of Small Area Characteristics in the presence of Covariates and Outliers, *Journal of the Multivariate Analysis*, **54**, 310-328.

Diallo, M. S. (2014), *Small Area Estimation under Skew-Normal Nested Error Models*, PhD Thesis, Ottawa-Carleton Institute for Mathematics and Statistics.

Fay, R. and Herriot, R. (1979), Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data, *Journal of the American Statistical Association*, **74**, 269-277.

Fellner, W. H. (1986). Robust Estimation of Variance Components, *Technometrics*, **28**, 51-60.

Ferraz, V. R. S. and Mourab, F. A. S. (2012). Small area estimation using skew normal models, *Computational Statistics and Data Analysis*, **56**, 2864-2874.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

Hill, R. W. (1997), *Robust Regression when there are Outliers in the Carriers*, PhD Thesis, Harvard University, Cambridge.

Huber, P. J. (1973), Robust Regression: Asyrnptotics, Conjectures, and Monte Carlo, *the Annals of Statistics*, **1**, 799-821.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

Huggins, R. M. (1993). A Robust Approach to the Analysis of Repeated Measures, *Biometrics* , **49**, 715-720.

Jiang, J. and Lahiri, P. (2006) Mixed Model Prediction and Small Area Estimation, *Test* , **15**, 1-96.

Mallows, C. L. (1975), *On Some Topics in Robustness* , technical memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

Molina, I., Salvati, N. and Pratesi, M. (2009), Bootstrap for Estimating the MSE of the Spatial EBLUP, *Computational Statistics*, **24**, 441-458.

Petrucci, A. and Salvati, N., (2006), Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment, *Agricultural, Biological, and Environmental Statistics*, **11**, 169-182.

Pfeffermann, D. (2013), New Important Developments in Small Area Estimation , *Statistical Science*, **28**, 40-68.

Pratesi, M. and Salvati, N. (2008), Small Area Estimation: the EBLUP Estimator Based on Spatially Correlated Random Area Effects , *Statistical methods and applications* , **17**, 113-141.

Rao, J. N. K. (2003) , *Small Area Estimation*, New York: Wiley.

Rao, J. N. K. and Molina, I. (2015), *Small Area Estimation. 2nd Edition* , New York: Wiley.

Richardson, A. M. (1997) Bounded Influence Estimation in mixed linear models , *Journal of American Statistical Association* , **92**, 154-161.

Richardson, A. M. and Welsh, A. H. (1995) Robust restricted maximum likelihood in mixed linear models , *Biometrics*, **51**, 1429-1439.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection* , New York: Wiley.

Salvati, N. (2004), Small Area Estimation by Spatial Models: The Spatial Empirical Best Linear Unbiased Prediction (Spatial EBLUP) , *Working Paper 2004/03*, University of Florence.

Simpson, D.G., Ruppert, D. and Carroll, R.J. (1992) On One-Step GM Estimates and Stability of Inferences in Linear Regression , *Journal of the American Statistical Association*, **418**, 439-450.

Singh, B. B., Shukla, G. K. and Kundu, D. (2005), Spatio-Temporal Models in Small Area Estimation, *Survey Methodology* , **31**, 183-195.

Sinha, S. K., Rao, J. N. K. (2009), Robust Small Area Estimation, *The Canadaian Journal of Statistics*, **37**, 381-399.

Shmid, T. N. and Munnich, R. T. (2014) Robust Spatial Small Area Estimation , *Statistical Papers*, **55**, 653-670.

Shmid, T. N., Tzavidis, N., Munnich, R. T. and Chambers, R. (2016) Robust Spatial Small Area Estimation, *Statistical Papers* , **55**, 653-670.

Schoch, T. (2012) Robust Unit-Level Small Area Estimation: A Fast Algorithm for Large Datasets, *Austrian Journal of Statistics* , **41**, 243-265.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010), Robust Estimation of Small Area Means and Quantiles, *Australian and New Zealand Journal of Statistics*, **52**, 167-186.

Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2008), M-quantile Models with Application to Poverty Mapping , *Statistical Methods and Applications*, **17**, 393-411.

Warnholz, S. (2016) , *Small Area Estimation using Robust Extensions to Area Level Models*, PhD Thesis, Freie University, Berlin.

Welsh, A. H., Richardson, A. M. (1997), 13 Approaches to the robust estimation of mixed models, *Handbook of Statistics* , **15**, 343-384.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶

آزمون ایستایی داده‌های فضایی

علی محمدیان مصمم، الهه شریفی*

گروه آمار، دانشگاه زنجان

چکیده: گاهی در مطالعات محیطی با مشاهداتی سروکار داریم که مستقل از یکدیگر نیستند و نوعاً وابستگی آن‌ها ناشی از موقعیت و مکان قرار گرفتن مشاهدات در فضای مورد مطالعه می‌باشد، این گونه مشاهدات داده‌های فضایی نامیده می‌شوند. به دلیل وجود ساختار و همبستگی فضایی بین چنین داده‌هایی، روش‌های معمول آمار برای تحلیل آنها قابل استفاده نمی‌باشد و لازم است به نحوی همبستگی داده‌ها در تحلیل آن‌ها لحاظ گردد. برای تحلیل این داده‌ها معمولاً فرض‌های ساده کننده‌ای نظیر ایستایی، همسانگردی و ... در عمل در نظر گرفته می‌شود. بنابراین بررسی صحت این فرض‌های ساده کننده‌ی مدل در داده‌های تحت بررسی امری ضروری می‌باشد. هدف ما در این مقاله، مطالعه‌ی فرض ایستایی این نوع داده‌ها با استفاده از خواص تبدیلات فوریه می‌باشد.

واژه‌های کلیدی: تبدیل فوریه، فرآیندهای ناایستا، فرآیندهای ایستا، تابع چگالی طیفی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11

۱ مقدمه

داده‌های فضایی مشاهداتی هستند که وابستگی آن‌ها ناشی از موقعیتشان در فضای مورد مطالعه است و این وابستگی تابعی از فاصله‌ی مشاهدات از یکدیگر است (کرسی، ۱۹۹۳) . دو ویژگی مهم داده‌های فضایی، نمایش هر داده با موقعیت آن در فضای مورد مطالعه و همبستگی فضایی این داده‌ها است. معمولاً داده‌های فضایی که از موقعیت‌های مجاور جمع‌آوری می‌شوند، همبستگی بیشتری دارند و با افزایش فاصله‌ی بین موقعیت داده‌ها این همبستگی کاهش می‌یابد. برای تجزیه و تحلیل داده‌های فضایی، لازم است یک مدل آماری در نظر گرفته شود. معمولاً یک میدان تصادفی مانند $\{Z(\mathbf{s}); \mathbf{s} \in \mathbb{R}^d, d \geq ۱\}$ که مجموعه‌ای از متغیرهای تصادفی است، به عنوان مدل آماری برای داده‌های فضایی در نظر گرفته می‌شود. اگر میانگین میدان تصادفی ثابت و به موقعیت $\mathbf{s}$ بستگی نداشته باشد و کوواریانس بین دو سری فضایی فقط تابعی از فاصله‌ی موقعیت ها باشد در این صورت سری فضایی ایستای مرتبه‌ی می‌باشد. در مقالات متعدد و مختلفی نظیر جان و جنتون (۲۰۱۲)، فنتس (۲۰۰۶) و پریستلی و سوبا رائو (۱۹۶۹) ایستایی داده‌های فضایی مورد مطالعه قرار گرفته است.

* **ارایه‌دهنده مقاله:** الهه شریفی، E_sharifi@znu.ac.ir

۲ آزمون ایستایی فضایی

در این بخش می‌خواهیم ایستایی میدان تصادفی فضایی را بر اساس خواص تبدیلات فوریه مورد آزمون قرار دهیم. فرض کنید $\{Z(\mathbf{s}); \mathbf{s} \in \mathbb{R}^d, d \geq ۱\}$ میدان تصادفی ایستای مرتبه‌ی دوم باشد به عبارت دیگر، داریم

$$E(Z(\mathbf{s})) = \mu \quad \text{Cov}\left(Z(\mathbf{s}), Z(\mathbf{s}+\boldsymbol{\nu})\right) = c(\boldsymbol{\nu})$$

با توجه به این مطلب که تبدیل فوریه‌ی گسسته‌ی یک فرآیند تصادفی ایستا تقریباً ناهمبسته و تبدیل فوریه‌ی گسسته‌ی یک فرآیند تصادفی ناایستا همبسته می‌باشد، ابتدا تبدیل فوریه‌ی فرآیند و همبستگی بین این تبدیلات را به دست آورده و با استفاده از آن آماره‌ی آزمون مناسب معرفی می‌شود.

فرض کنید $\{Z(\mathbf{s}); \mathbf{s} \in \mathbb{R}^d, d \geq ۱\}$ یک فرآیند تصادفی ایستا باشد که فقط در تعداد متناهی از مکان‌های فضایی نامنظم مشاهده شود، یعنی مشاهدات $\{(\mathbf{s}_j, Z(\mathbf{s}_j)); j = ۱, ۲, \ldots, n\}$ به صورت $\{\mathbf{s}_j\}_{j=۱}^{n}$ در ناحیه‌ی $\left[-\frac{\lambda}{۲}, \frac{\lambda}{۲}\right]^d$ باشند. وقتی مکان‌ها در فاصله‌ی نامنظم هستند، یعنی در یک صفحه‌ی هم‌فاصله روی $\left[-\frac{\lambda}{۲}, \frac{\lambda}{۲}\right]^d$ نیستند، هیچ راه یکتایی برای تعریف تبدیل فوریه وجود ندارد. با این حال یک تبدیل فوریه‌ی مناسب برای داده‌های نمونه‌گیری شده‌ی نامنظم که خاصیت «تقریباً ناهمبسته» را حفظ کرده است، به صورت زیر تعریف می‌شود:

$$J_n(\boldsymbol{\Omega}) = \frac{\lambda^{d/۲}}{n} \sum_{j=۱}^{n} Z(\mathbf{s}_j) \exp(i\mathbf{s}_j^\top \boldsymbol{\Omega}), \quad \boldsymbol{\Omega} \in \mathbb{R}^d.$$

این تبدیل فوریه برای اولین بار توسط ماتسودا و یاجیما (۲۰۰۹) و بندیوپدیای و لاهیری (۲۰۰۹) ارائه شده است. توجه داشته باشید عامل $\frac{\lambda^{d/۲}}{n}$ تضمین می‌کند که واریانس $J_n(\boldsymbol{\Omega})$ وقتی $\lambda \longrightarrow \infty$، ناتباهیده است. در معادله‌ی بالا $\boldsymbol{\Omega}$ نشان دهنده‌ی فرکانس فضایی است. تحت شرایطی که مکان‌های $\{\mathbf{s}_j\}$ متغیرهای تصادفی مستقل باشند و روی $\left[-\frac{\lambda}{۲}, \frac{\lambda}{۲}\right]^d$ به طور یکنواخت توزیع شده باشند و $\{Z(\mathbf{s}); \mathbf{s} \in \mathbb{R}^d, d \geq ۱\}$ یک فرآیند ایستای مرتبه‌ی چهارم باشد، بندیوپدیای وسوبا رائو (۲۰۱۶) نشان دادند که تبدیل فوریه $\{J_n(\boldsymbol{\Omega}_\mathbf{k})\}$ در مختصات $\boldsymbol{\Omega}_\mathbf{k} = ۲\pi\left(\frac{k_۱}{\lambda}, ..., \frac{k_d}{\lambda}\right)^\top$، $\mathbf{k} = (k_۱, k_۲, ..., k_d) \in \mathbb{Z}^d$، متغیرهای تصادفی تقریباً ناهمبسته هستند. همچنین آنها نشان دادند که اگر میدان‌های تصادفی فضایی بطور نامنظم نمونه‌گیری شوند، فرکانس‌های "تقریباً ناهمبسته" روی اندازه‌ی دامنه‌ی فضایی وابسته اتفاق می‌افتد.

تابع چگالی طیفی فرآیند ایستای فضایی به صورت زیر تعریف می‌شود:

$$f(\boldsymbol{\Omega}) = \int_{\mathbb{R}^d} c(\boldsymbol{\nu}) \exp\left\{-i\mathbf{s}^\top \boldsymbol{\Omega}\right\} d\boldsymbol{\Omega}.$$

در رابطه‌ی بالا، چون فرآیند فضایی روی $\mathbb{R}^d$ تعریف شده است، چگالی طیفی $f(\boldsymbol{\Omega})$ نیز روی $\mathbb{R}^d$ تعریف می‌شود؛ همچنین در این رابطه $\boldsymbol{\nu}$ نشان دهنده‌ی تأخیر فضایی می‌باشد. همچنین طبق قضیه‌ی دیگری در بندیوپدیای وسوبا رائو (۲۰۱۶) مشاهده می‌کنیم که در حالت ایستایی وقتی $\lambda \longrightarrow \infty$ (حوزه‌ی فضایی رشد می‌کند) تبدیلات فوریه‌ی گسسته، بیشتر ناهمبسته می‌شوند. علاوه بر این مشاهده می‌شود که وقتی $\lambda \longrightarrow \infty$ و $\frac{\lambda^d}{n} \longrightarrow ۰$ آنگاه واریانس تبدیلات فوریه به چگالی طیفی همگرا می‌شود. به عنوان مثال وقتی حوزه‌ی فضایی رشد می‌کند، تعداد مشاهدات باید در حوزه‌ی فضایی متراکم‌تر شود.

با توجه به اینکه فرآیند فضایی ایستای مرتبه‌ی دوم باشد یا نباشد، کوواریانس بین تبدیلات فوریه می‌تواند به شیوه‌های بسیار متفاوت عمل کند. در اصل اگر میدان تصادفی فضایی ایستای مرتبه‌ی دوم باشد، آنگاه تبدیل فوریه‌ی $J_n(\boldsymbol{\Omega})$ روی فرکانس‌های $\boldsymbol{\Omega}_\mathbf{k} = ۲\pi\mathbf{k}/\lambda$ تقریباً ناهمبسته است. توجه داشته باشید که در مورد ناایستایی وضعیت متفاوت است.

۱.۲ آماره‌ی آزمون

در ادامه برای فرض H_0 : فرآیند ایستای مرتبه‌ی دوم می‌باشد، در مقابل فرض H_A : فرآیند ایستای مرتبه‌ی دوم نمی‌باشد، آماره‌ی آزمونی را به صورت زیر تعریف می‌کنیم.

ابتدا برای اندازه‌گیری میزان همبستگی بین تبدیلات فوریه، کوواریانس وزنی بین تبدیلات فوریه (استاندارد نشده) با تأخیر $\mathbf{r}$ به صورت زیر تعریف می‌شود:

$$\hat{A}_\lambda(g;\mathbf{r}) = \frac{1}{\lambda^d}\sum_{k_1,\ldots,k_d=-a}^{a} g(\boldsymbol{\Omega}_\mathbf{k}) J_n(\boldsymbol{\Omega}_\mathbf{k})\overline{J_n(\boldsymbol{\Omega}_{\mathbf{k}+\mathbf{r}})} - \left[\frac{1}{n}\sum_{k_1,\ldots,k_d=-a}^{a} g(\boldsymbol{\Omega}_\mathbf{k}) \times \frac{1}{n}\sum_{j=1}^{n} Z(\mathbf{s}_j)^2 \exp(-i\mathbf{s}_j^\top \boldsymbol{\Omega}_\mathbf{r})\right]. \tag{۱.۲}$$

در رابطه‌ی بالا $\mathbf{r} \neq 0$ و $\mathbf{r} = (r_1, r_2, \ldots, r_d)^\top \in \mathbb{Z}^d$ ($\mathbf{k}$ و $\mathbf{k}+\mathbf{r}$ شبیه هم تعریف شده اند). همچنین $g : \mathbb{R}^d \to \mathbb{R}$ تابع پیوسته‌ی لیپشیتز [1] می‌باشد، که به صورت $g(\boldsymbol{\Omega}) = \sum_{j=1}^{L} \exp\{i\boldsymbol{\nu}_j^\top \boldsymbol{\Omega}\}$ تعریف شده است . در رابطه‌ی (۱.۲)، $\hat{A}_\lambda(g;\mathbf{r})$ عددی مختلط می‌باشد و قسمت‌های حقیقی و موهومی آن مجانبا دارای توزیع نرمال به صورت زیر می‌باشند:

$$c_{a,\lambda}^{-1/2}\lambda^{d/2}\left[\Re\hat{A}_\lambda(g;\mathbf{r}_1), \Im\hat{A}_\lambda(g;\mathbf{r}_1), \ldots, \Re\hat{A}_\lambda(g;\mathbf{r}_m), \Im\hat{A}_\lambda(g;\mathbf{r}_m)\right]^\top \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_{2m}). \tag{۲.۲}$$

در رابطه‌ی (۲.۲) وقتی $\lambda, a, n \to \infty$ آنگاه $\lambda^d / \left[n \log^{2d} a\right] \to 0$ ، همچنین I_{2m} نشان‌دهنده‌ی ماتریس همانی $2m$ ـ بُعدی می‌باشد.

در معادله‌ی بالا $c_{a,\lambda}^{-1/2}\lambda^{d/2}$ واریانس تقریبی قسمت حقیقی $\Re\hat{A}_\lambda(g;\mathbf{r})$ وقسمت موهومی $\Im\hat{A}_\lambda(g;\mathbf{r})$ می‌باشد. سپس برای اندازه‌گیری همبستگی بین تبدیلات فوریه، آماره‌ی آزمون برای قسمت‌های حقیقی و موهومی با استفاده از روش نمونه‌های متعامد [2] به شکل زیر تعریف می‌شود:

$$\lambda^{d/2}\frac{\Re\hat{A}_\lambda(g;\mathbf{r})}{c_{a,\lambda}^{-1/2}\lambda^{d/2}}. \tag{۳.۲}$$

و

$$\lambda^{d/2}\frac{\Im\hat{A}_\lambda(g;\mathbf{r})}{c_{a,\lambda}^{-1/2}\lambda^{d/2}}. \tag{۴.۲}$$

در روابط بالا $c_{\lambda,1}$ پارامتر نامعلوم بوده که نیاز به برآورد دارد. وقتی $\mathbf{r}$ در همسایگی صفر قرار داشته باشد، ضرایب $\hat{A}_\lambda(g;\mathbf{r})$ تقریباً ناهمبسته بوده و دارای واریانس مشابه می‌باشند. برای برآورد $c_{\lambda,1}$ مجموعه‌ی $\mathcal{S} \in \mathbb{Z}^d/\{0\}$ را که نسبت به صفر بسته است و در شرط $\mathcal{S} \cap \mathcal{S}' = \emptyset$ صدق می‌کند تعریف کرده و سپس $c_{\lambda,1}$ به صورت زیر برآورد می‌شود:

$$\hat{c}_{a,\lambda}(\mathcal{S}') = \frac{\lambda^d}{(2|\mathcal{S}'|-1)}\sum_{\mathbf{r}\in\mathcal{S}'}\left([\Re\hat{A}_\lambda(g;\mathbf{r}) - \bar{a}]^2 + [\Im\hat{A}_\lambda(g;\mathbf{r}) - \bar{a}]^2\right).$$

که در آن $\bar{a} = \frac{1}{2|\mathcal{S}'|}\sum_{\mathbf{r}\in\mathcal{S}'}[\Re\hat{A}_\lambda(g;\mathbf{r}) + \Im\hat{A}_\lambda(g;\mathbf{r})]$ می‌باشد. با جای‌گذاری $\hat{c}_\lambda(\mathcal{S}')$ به جای $c_{\lambda,1}$ در روابط (۳.۲) و (۴.۲) و با استفاده از رابطه‌ی (۲.۲) داریم:

$$\mathcal{T}_{\mathcal{S},1} = \lambda^{d/2}\frac{\Re\hat{A}_\lambda(g;\mathbf{r})}{\sqrt{\hat{c}_{a,\lambda}(\mathcal{S}')}} \xrightarrow{\mathcal{D}} \frac{Z_{1,\mathbf{r}}}{\sqrt{\frac{1}{2|\mathcal{S}'|-1}\chi^2_{2|\mathcal{S}'|-1}}} \sim t_{2|\mathcal{S}'|-1}.$$

[1] Lipschitz
[2] Orthogonal samples

و

$$\mathcal{T}_{\mathcal{S},۲}=\lambda^{d/۲}\frac{\Im\hat{A}_{\lambda}(g;\mathbf{r})}{\sqrt{\hat{c}_{a,\lambda}(\mathcal{S}')}}\xrightarrow{\mathcal{D}}\frac{Z_{۲,\mathbf{r}}}{\sqrt{\frac{۱}{۲|\mathcal{S}'|-۱}\chi^{۲}_{۲|\mathcal{S}'|-۱}}}\sim t_{۲|\mathcal{S}'|-۱}.$$

برای $\mathbf{r}\in\mathcal{S}$ وقتی $n\to\infty$ و $\lambda\to\infty$ آنگاه $\lambda^d/n\to ۰$. در روابط بالا $\{Z_{۱,\mathbf{r}},Z_{۲,\mathbf{r}};\mathbf{r}\in\mathcal{S}\}$ نشان دهنده‌ی متغیرهای تصادفی مستقل و هم توزیع نرمال استاندار بوده و $\chi^{۲}_{۲|\mathcal{S}'|-۱}$ متغیر تصادفی توزیع کای اسکور با $۲|\mathcal{S}'|-۱$ درجه‌ی آزادی می‌باشد. t_q نیز توزیع t با q درجه‌ی آزادی را نشان می‌دهد.

حال می‌توان ایستایی فرآیند $Z(\mathbf{s})$ را در سطح $\alpha\times ۱۰۰٪$ ، $\alpha\in(۰,۱)$ آزمون کرد. بنابراین فرض صفر (فرآیند ایستای مرتبه‌ی دوم می‌باشد.(رد می‌شود، هرگاه رابطه‌ی $\mathcal{T}_{\mathcal{S},۱}\geqslant t_{۲|\mathcal{S}'|-۱}$برقرار باشد.

بحث و نتیجه‌گیری

در این مقاله یک آماره‌ی آزمون برای فرض ایستایی فضایی فرآیند فضایی در قلمرو فرکانس ارائه شده است که فرض بسیار اساسی در تحلیل‌ها و پیش‌بینی‌های آمار فضایی می‌باشد، زیرا به دلیل وجود ساختار و همبستگی بین داده‌های فضایی، روش‌های معمول آمار برای این داده‌ها قابل استفاده نمی‌باشد و لازم است همبستگی داده‌ها در تحلیل آنها لحاظ گردد. بنابراین بررسی صحت فرض‌های ساده کننده‌ی مدل از جمله ایستایی در داده‌های تحت بررسی امری ضروری می‌باشد که در این مقاله به آن پرداخته شد. هدف بعدی ما کاربرد این آزمون بر روی داده‌های واقعی و شبیه سازی شده می‌باشد.

مراجع

محمدزاده م، (۱۳۹۴). *آمار فضایی و کاربردهای آن*، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

Bandyopadhyay, S., Jentsch, C., and Subba Rao, S. (2016). A spectral domain test for stationarity of spatio-temporal data, *Journal of Time Series Analysis*, **38**, 326-351.

Bandyopadhyay, S., and Lahiri, S. N. (2009). Asymptotic properties of Discrete Fourier Transforms for spatial data, *Snakhya, Series A*, **71**, 221-259.

Bandyopadhyay, S., and Subba Rao, S. (2016). A test for stationarity irregularly spaced spatial data, *Journal of the royal Statistical Society, series B*, **79**, 95-123.

Cressie, N. (1993). *Statistics for Spatial data*, John Wiley and Sons, Inc.

Fuentes, M. (2006). Testing for separability of a spatial temporal covariance functions, *Journal of Statistical Planning and Inference*, **136**, 447-466.

Jentsch, C., and Subba Rao, S. (2015). A test for second order stationarity of a multivariate time series, *Journal of Econometrics*, **185**, 124-161.

Jun, M., and Genton, M. (2012). A test for stationarity of spatio-temporal random fields on planar and spherical domains, *Statistica Sinica*, **22**, 1737–1764.

Matsuda, Y., and Yajima, Y., (2009). Fourier analysis of irregularly spaced data on R , *Journal of the Royal Statistical Sosiety, series B*, **71**, 191-217.

Priestley, M. B., and Subba Rao, T. (1969). A test for non-stationarity of a time series, *Journal of the Royal Statistical Society, Series A*, **31**, 140–149.

دومین سمینار آمار فضایی و کاربردهای آن
۳-۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1390

بررسی روش‌های خوشه‌بندی داده‌های فضایی

سیده سمیه موسوی*، عادل محمدپور

گروه آمار، دانشگاه صنعتی امیرکبیر

چکیده: خوشه‌بندی فضایی یک دید کلی و جامعی نسبت به توزیع مشاهدات در ناحیه موردمطالعه فراهم می‌کند. بنابراین خوشه‌بندی فضایی گام مهمی در جهت فرایند تصمیم‌گیری در زمینه‌های کاربردی مختلف و مهم مانند میزان جرم و جنایت، میزان امنیت عمومی، میزان و نوع بیماری در مناطق مختلف، مشکلات زیست‌محیطی، میزان بهداشت عمومی، الگوهای آب و هوا و تسهیلات حمل‌ونقل خاص می‌باشد. روش‌ها و الگوریتم‌های متعددی برای خوشه‌بندی داده‌ها معرفی شده است که هر یک از آن‌ها مزایا و محدودیت‌هایی برای پردازش داده‌های چندبعدی مانند داده‌های فضایی دارند. در این مقاله بر اساس روش تخصیص داده‌ها به درون خوشه‌ها، الگوریتم‌های خوشه‌بندی به چهار دسته‌ی افرازی، سلسله‌مراتبی، چگالی‌پایه و شبکه‌پایه تقسیم شده است. همچنین الگوریتم‌های خوشه‌بندی فضایی تحت چهار دسته فوق از نظر معایب و مزایا مورد مطالعه قرار گرفته‌اند.

هدف از این مقاله، انجام یک مطالعه مروری از الگوریتم‌های خوشه‌بندی موجود است که می‌توانند داده‌های فضایی را پردازش کنند. آن الگوریتم خوشه‌بندی که انتخاب می‌شود باید الزامات کاربردی برای مطالعه انجام‌شده را داشته باشد. همچنین این الگوریتم باید در پردازش داده‌های دورافتاده و نوفه که در مجموعه داده‌های جغرافیایی به‌طور ذاتی وجود دارند، کارا باشد. نشان می‌دهیم که برای پیاده‌سازی الگوریتم‌های خوشه‌بندی قدیمی‌تر روی داده‌های فضایی، تطابق قابل توجهی به منظور انجام عملکردشان ایجاد شده و با اطلاع از مشکلات ذاتی در خوشه‌بندی داده‌های فضایی بزرگ و با ابعاد بالا، این الگوریتم‌ها هم از نظر سرعت و هم کارآیی بهبود یافته‌اند.

واژه‌های کلیدی: داده‌های فضایی، خوشه‌بندی فضایی، الگوریتم‌های خوشه‌بندی .

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11 ، 62H30.

* **ارایه‌دهنده مقاله:** سیده سمیه موسوی، s.s.mousavi@aut.ac.ir

۱ مقدمه

در بسیاری مواقع با داده‌هایی سروکار داریم که مستقل از هم نبوده و برحسب موقعیت قرارگیری‌شان در فضای موردمطالعه به یکدیگر وابسته‌اند. اگر این وابستگی تابعی از فاصله‌ی بین موقعیت‌های مشاهدات باشد، به طوری‌که مشاهدات نزدیک به‌هم وابسته‌تر و مشاهدات دورتر وابستگی کمتری داشته باشند، این‌گونه مشاهدات داده‌های فضایی نامیده می‌شوند. به‌طورمعمول داده‌های فضایی علاوه بر مقادیر مربوط به متغیر یا متغیرهای موردبررسی، اطلاعات مربوط به مکان مشاهده را نیز شامل می‌شوند، به همین دلیل منطقی به نظر می‌رسد که مشاهده مربوط به یک مکان به‌نوعی به مشاهدات مکان‌های دیگر وابسته باشد؛ بنابراین همبستگی فضایی در داده‌های فضایی وجود دارد (محمدزاده ، ۱۳۹۴). ازجمله این داده‌ها می‌توان به داده‌های جغرافیایی (نقشه)، داده‌های تصویربرداری پزشکی و ماهواره‌ای، داده‌های مربوط به میزان یا شیوع یک بیماری در مناطق مختلف، داده‌های غلظت گازهای آلاینده هوا در میدان‌های مختلف یک شهر، داده‌های میزان آلودگی آب رودخانه‌ها اشاره کرد.

یکی از روش‌های اساسی برای جستجوی الگوها در داده‌های فضایی، استفاده از فنون خوشه‌بندی است. به‌طورکلی هدف از خوشه‌بندی فضایی[۱]، گروه‌بندی مشاهدات فضایی مشابه به زیرمجموعه‌های معنی‌دار است. در این گروه‌بندی، اشیا شبیه به هم ازنظر صفات فضایی بر اساس برخی ملاک‌های تشابه، در داخل گروه‌های یکسان قرار می‌گیرند (هان و همکاران ، ۲۰۰۱). کاربردهای وسیع خوشه‌بندی توجه بسیاری از محققان را به خود جلب کرده و به‌عنوان یک زمینه‌ی تحقیقاتی در طی دهه‌های اخیر موردمطالعه قرارگرفته است. تاکنون بررسی‌های جامع درباره خوشه‌بندی داده‌های فضایی صورت گرفته است که ازجمله‌ی آن‌ها می‌توان به مقالات هان و همکاران (۲۰۰۱)، کولچ (۲۰۰۱)، چاندرا و آنوردها (۲۰۱۱) و وارجس و یونیکریشنان (۲۰۱۳) اشاره کرد. خوشه‌بندی فضایی در زمینه‌های بسیاری مانند کشف اطلاعات جغرافیایی، حسگرها، کشف بیماری‌های نادر، ستاره‌شناسی، نقشه‌برداری‌های ماهواره‌ای زمین، شناسایی سرزمین‌های قابل‌استفاده مشابه، ادغام نواحی دارای الگوی آب و هوایی مشابه و بررسی میزان جرم و جنایت در مناطق مختلف اهمیت بسزایی دارد.

روش‌ها و الگوریتم‌های متعددی برای خوشه‌بندی داده‌ها وجود دارد که هر یک از آن‌ها مزایا و محدودیت‌هایی برای خوشه‌بندی داده‌های فضایی دارند. اگرچه شباهت‌هایی بین خوشه‌بندی فضایی و غیرفضایی در داده‌گان های بزرگ و خصوصاً داده‌های فضایی وجود دارد، اما نیازمندی‌های خاص در داده‌های فضایی باعث ایجاد نیازهای ویژه‌ای برای الگوریتم‌های خوشه‌بندی فضایی می‌شود؛ بنابراین برای انجام خوشه‌بندی مناسب داده‌های فضایی با اندازه زیاد، باید الگوریتمی بیابیم که معیارهای زیر را دارا باشد:

(۱) برای داده‌های دارای تعداد زیادی متغیر (داده‌های با ابعاد بالاتر) قابل کاربرد باشد.

(۲) با توجه به اندازه زیاد داده‌ها، مقیاس‌پذیر و کارا باشد. یعنی زمان اجرای مناسب برای اندازه زیاد داده‌ها را داشته و خوشه‌های دارای کیفیت بالا ایجاد کند.

(۳) قادر به شناسایی اَشکال نامنظم و ایجاد خوشه‌های دارای اَشکال دلخواه مانند حفره‌دار، مقعر و تودرتو باشد.

(۴) به نوفه و داده‌های دورافتاده حساس نباشد.

(۵) الگوریتم به رتبه ورودی حساس نباشد؛ یعنی نتایج خوشه‌بندی باید مستقل از رتبه داده‌ها باشد.

[۱] Clustering spatial

(۶) الگوریتم نیاز به دانش قبلی از داده‌ها یا تعداد خوشه‌هایی که باید ایجاد شود، نداشته باشد و از کاربر هیچ اطلاعاتی درباره داده‌های ورودی درخواست نکند.

در این مقاله یک مطالعه مروری از الگوریتم‌های خوشه‌بندی موجود که برای خوشه‌بندی داده‌های چندبعدی فضایی مورد استفاده قرار می‌گیرند، ارائه کرده و نقاط ضعف و قوت هریک را بیان می‌کنیم. همچنین نشان می‌دهیم که برای پیاده‌سازی الگوریتم‌های خوشه‌بندی قدیمی‌تر روی داده‌های فضایی، با توجه به خواص ذاتی این نوع داده‌ها، باید این الگوریتم‌ها از نظر سرعت و کارآیی بهبود یابند.

در این مقاله مروری، الگوریتم‌های خوشه‌بندی مرسوم برای داده‌های فضایی به چهار دسته‌ی اصلی افرازی، سلسله‌مراتبی، چگالی‌پایه و شبکه‌پایه تقسیم شده است که در بخش ۲ و زیربخش‌های آن به مرور این روش‌ها و برخی از الگوریتم‌های مطرح شده، مزایا و معایب هر یک از آن‌ها می‌پردازیم. در بخش ۳ به بحث و نتیجه‌گیری از این کار می‌پردازیم.

۲ روش‌های خوشه‌بندی فضایی

در زمینه‌ی خوشه‌بندی فضایی اشیاء یک ناحیه بر اساس نزدیکی مکانی‌شان نسبت به یکدیگر، مطالعات زیادی صورت گرفته است. طی دهه‌های گذشته، غالباً از الگوریتم‌های خوشه‌بندی غیرفضایی شناخته‌شده در حوزه فضایی استفاده‌شده است و برای به دست آوردن الگوریتم‌های فضایی به‌منظور خوشه‌بندی داده‌های فضایی، با توجه به ماهیت ذاتی ویژه این نوع داده‌ها، الگوریتم‌های غیرفضایی ازنظر سرعت اجرا و کارایی بهبود یافتند (رادیک و هورنسبای ، ۲۰۰۱؛ هان و همکاران ، ۲۰۰۱؛ تائو و همکاران ، ۲۰۰۴؛ وانگ و همیلتون ، ۲۰۰۵).

به علت کاربردهای وسیع خوشه‌بندی فضایی در زمینه‌های متعدد، این شاخه به‌عنوان یک موضوع کاربردی در داده‌کاوی[۲] مطرح است و روش‌های خوشه‌بندی مقیاس‌پذیری ایجاد شده است. این روش‌های خوشه‌بندی فضایی را می‌توان به چهار دسته‌ی روش‌های افرازی، سلسله‌مراتبی، چگالی‌پایه و شبکه‌پایه تقسیم کرد. در این بخش هر یک از این روش‌ها را معرفی و برخی الگوریتم‌های آن‌ها را به همراه مزایا و محدودیت‌هایشان بیان می‌کنیم.

۱.۲ روش‌های افرازی

از مهم‌ترین و قدیمی‌ترین روش‌های خوشه‌بندی بکار برده شده برای داده‌های فضایی، روش‌های افرازی[۳] است. فرض کنید مجموعه D شامل n مشاهده در یک فضای d بعدی و پارامتر ورودی k نشانگر تعداد خوشه ها باشد. یک الگوریتم افراز، به نحوی این n مشاهده را به k خوشه مجزا تقسیم می‌کند که یک معیار افراز عینی بهینه شود. به عبارت دیگر خوشه‌ها به نحوی ایجاد می‌شوند که انحراف کلی هر مشاهده از مرکز خوشه‌اش یا از یک توزیع خوشه کمینه باشد.

در این بخش سه روش افراز را معرفی می‌کنیم: روش‌های k-means، k-medoid و EM. این سه روش، الگوریتم‌های متفاوتی برای نمایش خوشه‌ها دارند. الگوریتم k-means مرکز هندسی (یا میانگین) مشاهدات درون خوشه‌ها را به‌عنوان مرکز خوشه در نظر می‌گیرد، درحالی‌که الگوریتم k-medoid مرکزی‌ترین مشاهده درون خوشه را به عنوان مرکز خوشه استفاده می‌کند. برخلاف دو روش فوق، الگوریتم خوشه‌بندی EM یک توزیع شامل یک میانگین و یک ماتریس کوواریانس را برای هر خوشه در نظر می‌گیرد.

[۲] Mining data
[۳] Methods partitioning

این الگوریتم به‌جای اینکه هر مشاهده را به یک خوشه خاصی اختصاص دهد، هر مشاهده را به خوشه مطابق با یک احتمال عضویت که از توزیع هر خوشه محاسبه می‌شود اختصاص می‌دهد. در این روش، هر مشاهده احتمال معینی از متعلق بودن به هر خوشه دارد.

باوجود تفاوت این سه روش در نمایش خوشه‌ها، ملاحظه می‌کنیم که الگوریتم‌هایشان روش کلی یکسانی را برای یافتن خوشه‌ها به کار می‌برند. در واقع این سه الگوریتم سعی در یافتن k خوشه یا توزیع دارند به طوری‌که یک معیار هدف را بهینه کنند. ابتدا k مرکز بهینه یا k توزیع بهینه پیدا می‌شوند و سپس عضویت n مشاهده به داخل k خوشه به طور خودکار تعیین می‌شود.

اغلب روش‌های خوشه‌بندی افرازی، بر اساس فاصله بین مشاهدات خوشه‌بندی انجام می‌دهند. چنین روش‌هایی فقط می‌توانند خوشه‌های کروی شکل را بیابند و در یافتن خوشه‌های دارای شکل دلخواه با مشکل مواجه می‌شوند. بنالراین ایراد اصلی الگوریتم‌های بر مبنای افراز، نیاز آن‌ها به مشخص بودن تعداد خوشه ها (پارامتر k) و ناتوانی‌شان در یافتن خوشه‌های با شکل‌های دلخواه است. در ادامه به شرح سه الگوریتم فوق خواهیم پرداخت.

۱.۱.۲ روش k-means

الگوریتم k-means مقدار میانگین مشاهدات درون یک خوشه را به‌عنوان مرکز آن خوشه در نظر می‌گیرد و معیار هدف مورداستفاده در این الگوریتم به‌صورت زیر تعریف می‌شود:

$$E = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathbf{C}_j} \|\mathbf{x}_i - \mu_j\|^2; \qquad i = 1, \ldots, n, \; k < n \tag{۱.۲}$$

که در آن $\mathbf{C}_j$ خوشه jام، بردار $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^\top$ مقدار مشاهده iام و بردار μ_j میانگین خوشه‌ی jام است (برکهین، ۲۰۰۲). در ابتدا به طور کاملاً تصادفی k نقطه به عنوان مراکز اولیه k خوشه انتخاب می‌شوند و فاصله بین آن نقاط با کلیه نقاط دیگر محاسبه شده و نزدیکترین مشاهدات به این k نقطه اولیه مشخص و تشکیل k خوشه می‌دهند. در مرحله دوم مجدداً میانگین‌گیری همه مشاهدات درون خوشه‌های قبلی انجام شده و خوشه‌های جدید بر اساس نزدیکترین مشاهدات نسبت به میانگین‌های جدید شکل می‌گیرند. این فرایند تا جایی ادامه می‌یابد که تابع معیار E بعد از تکرار تغییر نکند. البته به جای فاصله اقلیدسی در (۱.۲) می‌توان از متر یا فاصله‌ی دیگر استفاده کرد.

الگوریتم k-means در پردازش مجموعه داده‌های بزرگ نسبتاً مقیاس‌پذیر و کارا است اما علاوه بر ایراد اساسی الگوریتم‌های بر مبنای افراز، این الگوریتم به داده‌های دورافتاده و نوفه بسیار حساس است زیرا تعداد کمی از چنین داده‌هایی به‌طور قابل‌توجهی مقدار میانگین را تحت تأثیر قرار می‌دهند. ضمنا همبستگی مشاهدات فضایی را در نظر نمی‌گیرد.

۲.۱.۲ روش k-medoid

روش k-medoid مرکزی‌ترین مشاهده درون هر خوشه را به‌عنوان مرکز خوشه در نظر می‌گیرد. در مرحله اول این الگوریتم k مشاهده را به‌طور تصادفی به‌عنوان مرکز k خوشه انتخاب می‌شود. مشابه روش k-means در این روش نیز هر مشاهده به نزدیک‌ترین مرکز اختصاص داده می‌شود. الگوریتم‌های PAM، CLARA و CLARANS از روش k-medoid برای خوشه‌بندی استفاده می‌کنند. الگوریتم PAM همه k مرکز خوشه را تکرار می‌کند و هر یک از آن‌ها را با یکی از $n-k$ مشاهده دیگر جایگزین می‌کند. تا وقتی‌که تابع توان دوم خطای E در (۱.۲) کاهش یابد، این جایگزینی صورت می‌گیرد و باعث وقوع تکرار بعدی الگوریتم می‌شود. اگر هیچ جایگزینی پیدا نشود آنگاه هیچ تغییری در تابع E رخ نمی‌دهد و الگوریتم با مقدار بهینه موضعی به پایان می‌رسد. ایراد اصلی

این الگوریتم عدم کارایی برای مجموعه داده‌های حجیم است که در آن صورت از الگوریتم‌های CLARA و CLARANS استفاده می‌شود. مزیت روش k-medoid نسبت به روش k-means کاهش حساسیت به داده‌های دورافتاده و نوفه است. در این روش نیز همبستگی داده‌های فضایی نادیده گرفته می‌شود.

۳.۱.۲ روش EM (خوشه‌بندی احتمالاتی)

در رویکرد احتمالاتی، داده‌ها به عنوان یک نمونه مستقل از یک مدل احتمالاتی آمیخته در نظر گرفته می‌شوند و فرض اصلی این است که داده‌ها به صورت تصادفی از یک مدل j با احتمال τ_j، $j = ۱, \ldots, k$ انتخاب می‌شوند. محدوده اطراف میانگین هر توزیع یک خوشه طبیعی[۴] است. بنابراین خوشه‌بندی احتمالاتی، هر خوشه را با استفاده از یک توزیع پارامتری بیان می‌کند.

فرض کنید هر مشاهده $\mathbf{x}$ متعلق به تنها یک خوشه است، در این صورت احتمال مشاهده $\mathbf{x}$ مرتبط با مدل j ام را با $P(\mathbf{x}|\mathbf{C}_j)$ نشان می‌دهیم و درستنمایی کلی داده‌ها به صورت زیر بیان می‌شود:

$$L(\mathbf{X}|\mathbf{C}) = \prod_{i=۱}^{n} \sum_{j=۱}^{k} \tau_j P(\mathbf{x}_i|\mathbf{C}_j) \qquad (۲.۲)$$

که در آن $\mathbf{X} = \{\mathbf{x}_۱, \ldots, \mathbf{x}_n\}$ بیانگر بردار مشاهدات و $\mathbf{C} = \bigcup_{j=۱}^{k} \mathbf{C}_j$ است.

لگاریتم درستنمایی (۲.۲) به عنوان یک تابع هدف عمل می کند که به روش EM[۵] منجر می شود (میچل ، ۱۹۹۷). الگوریتم EM یک روش بهینه‌سازی تکراری دو مرحله‌ای است. در هر تکرار از این الگوریتم، به ترتیب در مرحله E یعنی محاسبه امید ریاضی، احتمالات محاسبه می‌شوند و در مرحله M یعنی بیشینه کردن تابع درستنمایی (که امید ریاضی مرحله اول در آن جایگزین شده)، تقریبی از یک مدل آمیخته به دست می‌آید. درواقع این الگوریتم به دنبال یافتن پارامترهای مدل آمیخته‌ای است که لگاریتم تابع درستنمایی را ماکسیمم می‌کند. در بیان دقیق‌تر، الگوریتم EM به دنبال برآورد برچسب برای هر یک از مشاهدات در یک دنباله از متغیرهای پنهان است. جهت پیاده‌سازی این الگوریتم به مقادیر آغازین پارامترهای هر زیرجامعه نیاز است. اگر مقادیر اولیه برای این پارامترها به درستی صورت نگیرد، ممکن است الگوریتم همگرا نشود یا زمان همگرایی آن طولانی شود. این فرایند تا زمان همگرایی لگاریتم درستنمایی ادامه می یابد. یعنی وقتی افزایش در لگاریتم درستنمایی بین دو تکرار متوالی ناچیز باشد، این الگوریتم خاتمه می‌یابد.

۲.۲ روش‌های سلسله‌مراتبی

خوشه‌بندی سلسله‌مراتبی[۶] الگوریتمی است که خوشه‌ها را بر اساس معیار فاصله پیدا می‌کند و در نهایت فاصله بین اشیا در خوشه‌های یکسان کمتر از فاصله اشیا نسبت به اشیا بیرون خوشه است. فنون خوشه‌بندی سلسله‌مراتبی ناشی از یک سری ادغام‌های متوالی (یا تجمیعی) و یک سری تقسیمات پی‌درپی (یا تقسیمی) است. در روش تجمیعی ابتدا هر مشاهده به‌عنوان گروه جداگانه در نظر گرفته می‌شود. به‌طور متوالی مشاهدات باهم ادغام می‌شوند یا بر اساس برخی اندازه‌ها مانند فاصله بین مراکز دو گروه، گروه‌بندی می‌شوند و این فرایند تا جایی ادامه دارد که همه گروه‌ها باهم ادغام شده و در بالاترین سطح سلسله‌مراتب فقط یک گروه حاصل شود و یا اینکه یک شرط پایانی برقرار شود. روش تقسیمی با یک خوشه که شامل همه مشاهدات است، شروع می‌کند. در هر

[۴] Cluster natural
[۵] Expectation-Maximization
[۶] Clustering hierarchical

تکرار متوالی یک خوشه مطابق برخی اندازه‌ها به خوشه‌های کوچک‌تر تقسیم می‌شود تا اینکه سرانجام هر مشاهده در یک خوشه جداگانه قرار گیرد و یا اینکه یک شرط پایانی خاصی برقرار شود مانند اینکه تعداد مورد نظری از خوشه‌ها به دست آید یا فاصله بین دو نزدیک‌ترین خوشه بیشتر از فاصله معینی باشد.

نتایج خوشه‌بندی‌های سلسله‌مراتبی تجمیعی و تقسیمی را می‌توان در یک نمودار دو بعدی با عنوان دارنگاشت[۷] نشان داد که ادغام‌ها و تقسیمات ایجاد شده را نشان می‌دهد. همچنین در انجام خوشه‌بندی سلسله‌مراتبی موارد زیر باید مشخص شوند:

- یک اندازه برای تشابه یا عدم تشابه بین اشیا، مانند فاصله اقلیدسی
- یک معیار برای ادغام خوشه ها در مراحل متوالی مانند تک اتصالی[۸]

برخلاف الگوریتم خوشه‌بندی بر مبنای افراز که برای بهبود کیفیت خوشه‌بندی، مشاهدات را از یک خوشه به خوشه دیگر اختصاص می‌دهد، در بیشتر الگوریتم‌های ابتدایی خوشه‌بندی سلسله‌مراتبی عضویت یک مشاهده به هر خوشه‌ای تثبیت می‌شود یعنی هر مشاهده به هر خوشه‌ای اختصاص یابد تا پایان خوشه‌بندی در همان خوشه باقی می‌ماند و امکان جابه‌جایی مشاهدات وجود ندارد و ایراد اساسی این نوع روش‌های خوشه‌بندی محسوب می‌شود. ازجمله الگوریتم‌های سلسله‌مراتبی می‌توان به الگوریتم‌های AGNES، DIANA، BIRCH، CURE و CHAMELEON اشاره کرد که در ادامه هر یک از آن‌ها را توضیح می‌دهیم.

۱.۲.۲ الگوریتم‌های AGNES و DIANA

AGNES یک الگوریتم سلسله مراتبی تجمیعی و DIANA یک الگوریتم سلسله مراتبی تقسیمی است. در هرکدام از این دو روش می‌توان شرط پایانی الگوریتم را رسیدن به تعداد مشخصی خوشه یا اینکه فاصله بین دو نزدیک‌ترین خوشه بیشتر از فاصله معینی باشد، در نظر گرفت. باوجود سادگی این دو الگوریتم، غالباً با مشکلاتی راجع به انتخاب نقاط ادغام یا تقسیم مواجه می‌شویم. اینکه چه نقاطی را در هر مرحله برای ادغام یا تقسیم انتخاب می‌کنیم مهم است زیرا هر بار که گروهی از مشاهدات ادغام یا تقسیم می‌شوند فرایند در مرحله بعد، روی خوشه‌های ایجادشده مرحله قبل خودش انجام می‌شود. بنابراین اگر در هر مرحله‌ای برای اینکه کدام خوشه‌ها باهم ادغام شوند یا کدام خوشه‌ها تقسیم گردند انتخاب درستی نداشته باشیم، خوشه‌های با کیفیت پایین حاصل خواهد شد.

۲.۲.۲ الگوریتم BIRCH

الگوریتم BIRCH خوشه‌بندی سلسله‌مراتبی چند مرحله‌ای با کمک درخت مشخصه خوشه‌بندی[۹] است. روش اجرای این الگوریتم، متراکم کردن مشاهدات به‌صورت زیر خوشه‌های کوچک و سپس انجام خوشه‌بندی با این زیر خوشه‌ها است که به علت متراکم‌سازی، تعداد زیر خوشه‌ها خیلی کمتر از تعداد مشاهدات می‌شود. در الگوریتم BIRCH هر زیرخوشه کوچک توسط مشخصه خوشه‌بندی (CF) مشخص می‌شود که شامل خلاصه‌ای از اطلاعات هر خوشه است. فرض کنید $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$، نشانگر مجموعه مشاهدات d بعدی در یک زیرخوشه باشد، در این صورت CF به فرم بردار سه‌جزئی زیر نمایش داده می‌شود:

$$\text{CF} = (m,\, LS,\, SS)\,; \quad m < n$$

[۷] Dendogram
[۸] Linkage single
[۹] Tree feature clustering

که در آن n تعداد کل مشاهدات، m تعداد مشاهدات درون زیرخوشه، $\mathbf{x}_l$ مشاهده lام درون زیر خوشه، $LS = \sum_{l=1}^{m} \mathbf{x}_l$ و $SS = \sum_{l=1}^{m} \mathbf{x}_l^2$ می‌باشد.

این CFها در یک درخت مشخصه خوشه‌بندی (CF-tree) ذخیره می‌شوند که خلاصه‌ای از هر خوشه‌اند و دارای دو پارامتر به نام های فاکتور شاخه‌ای[۱۰] (B) و مقدار آستانه[۱۱] (T) می‌باشند. الگوریتم خوشه‌بندی BIRCH در ۴ مرحله اجرا می‌شود. در مرحله اول، CF اولیه از مجموعه داده‌ها بر اساس فاکتور شاخه‌ای B و مقدار آستانه T ساخته می‌شود. در مرحله ۲ که یک مرحله اختیاری است، اندازه درخت CF اولیه کاهش می‌یابد تا درخت CF کوچک‌تر به دست آید. در مرحله ۳ خوشه‌بندی کلی داده‌ها از درخت CF اولیه یا درخت کوچک‌تر مرحله ۲، انجام می‌شود. همچنین می‌توان در مرحله ۳، ارزیابی خوشه‌های مناسب را انجام داد. اگر نیاز به بهبود کیفیت خوشه‌ها باشد باید مرحله ۴ الگوریتم انجام شود. در این مرحله همه مجموعه نقطه‌داده‌ها خوانده شده و تعدادی از آن‌ها بر اساس یک تابع فاصله انتخاب می‌شوند. نقطه‌داده‌های انتخابی در گره‌های درخت CF ذخیره می‌شوند. نقاطی که نزدیک هم قرارگرفته‌اند برای قرار گرفتن در خوشه‌هایی موردبررسی قرار می‌گیرند و نقاطی که به‌طور پراکنده از هم قرارگرفته‌اند به‌عنوان داده‌های دورافتاده تلقی شده و از فرایند خوشه‌بندی کنار گذاشته می‌شوند.

این الگوریتم دو مشکل اساسی عدم مقیاس‌پذیری و عدم برگشت‌پذیری به مرحله قبل که در روش‌های خوشه‌بندی تجمیعی وجود دارد را برطرف می‌نماید و با کمک ساختارهای موجود در آن، سرعت مناسب و قابلیت مقیاس‌پذیری خوبی در مواجه با پایگاه داده‌های بزرگ از خود نشان می‌دهد.

۳.۲.۲ الگوریتم CURE

الگوریتم CURE یک روش خوشه‌بندی سلسله‌مراتبی تجمیعی است که دو ایده اصلی برای به دست آوردن خوشه‌های با کیفیت بالا بکار می‌برد: اول اینکه به‌جای استفاده از مرکز تنها یا مشاهده، تعداد ثابتی از مشاهدات خوب پراکنده‌شده برای بیان هر خوشه انتخاب شوند. دوم اینکه مشاهدات انتخاب‌شده نسبت به مراکز خوشه‌هایشان توسط کسر معینی به نام فاکتور انقباض[۱۲] منقبض شوند. فاکتور انقباض با α نشان داده می‌شود و مقدار آن بین ۰ و ۱ تغییر می‌کند.

در واقع این الگوریتم ابتدا یک نمونه تصادفی از مجموعه داده موجود انتخاب می‌کند و این نمونه داده را به گروه‌هایی تقسیم کرده و خوشه‌های جزئی با افراز مشاهدات نمونه‌ای به دست می‌آیند و برخی نقاط نماینده هر یک از این گروه‌ها را شناسایی می‌کند.

در مرحله اول، یک مجموعه‌ای از نقاط که در مجموعه داده موجود به‌طور گسترده پراکنده شده‌اند را بررسی می‌کند. در مرحله بعدی نقاط پراکنده‌شده‌ی انتخابی با یک مقدار مشخصی از فاکتور انقباض α، به مرکز خوشه منتقل می‌شوند. یک نتیجه از این فرایند، ایجاد خوشه‌هایی با شکل‌های تصادفی از مجموعه داده‌ها، شناسایی و حذف داده‌های دورافتاده است. در مرحله بعد الگوریتم، نقاط نماینده خوشه‌ها ازنظر نزدیکی به یک مقدار آستانه بررسی می‌شوند و خوشه‌هایی که مجاور یکدیگرند (نزدیک‌ترین خوشه‌ها) باهم ادغام شده و مجموعه خوشه‌های جدید ایجاد می‌شود. درنهایت از مجموعه خوشه‌های جزئی، خوشه‌های پالایش‌شده ایجاد می‌شوند. به‌کارگیری فاکتور انقباض توسط الگوریتم CURE باعث غلبه بر محدودیت‌های روش‌های بر مبنای مرکز و بر مبنای همه مشاهدات می‌شود و اثرات بد داده‌های دورافتاده تا حد زیادی کاهش می‌یابد.

[۱۰] Factor branching
[۱۱] Threshold
[۱۲] Factor shrinking

۴.۲.۲ الگوریتم CHAMELEON

الگوریتم CHAMELEON یک الگوریتم خوشه‌بندی سلسله‌مراتبی است که با کمک مدل‌سازی پویا شباهت میان زوج خوشه‌ها را تعیین می‌کند. در این الگوریتم شباهت خوشه بر اساس چگونگی اتصال مناسب اشیا درون خوشه و نزدیکی خوشه‌ها ارزیابی می‌شود. بنابراین دو خوشه‌ای ادغام خواهند شد که دارای اتصالات قوی در درون خود و به یکدیگر بسیار نزدیک باشند.

ایراد اصلی روش‌های سلسله مراتبی تجمیعی وابستگی آن‌ها به یک مدل ایستای تعیین‌شده توسط کاربر است که الگوریتم CHAMELEON بر این محدودیت غلبه کرده و به صورت خودکار و بر اساس خصوصیات داخلی خوشه‌های ادغام شونده عمل می‌کند. این الگوریتم از یک نمودار تُنُک[۱۳] استفاده می‌کند که در آن راس‌ها و وزن یال‌ها به ترتیب بیانگر مشاهدات و شباهت‌های بین مشاهدات هستند. این نمودار برای مجموعه داده‌های بزرگ کارا است. این الگوریتم شباهت بین هر جفت خوشه C_i و C_j را با محاسبه $RI(C_i, C_j)$ که نشانگر به‌هم‌پیوستگی نسبی و $RC(C_i, C_j)$ که بیانگر نزدیکی نسبی دو خوشه C_i و C_j است، تعیین می‌کند. این دو کمیت به‌صورت زیر بیان می‌شوند:

$$RI\left(C_i, C_j\right) = \frac{\left|EC_{\{C_i,C_j\}}\right|}{\frac{\left|EC_{C_i}\right|+\left|EC_{C_j}\right|}{۲}} \tag{۳.۲}$$

که در آن $|EC_{C_i,C_j}|$ برش یال خوشه‌ی شامل C_i و C_j ، $|EC_{C_i}|$ (یا $|EC_{C_j}|$) حداقل مجموع وزن یال‌هایی است که خوشه C_i (یا C_j) را به دو قسمت تقریبا هم اندازه تقسیم می‌کند. همچنین

$$RC\left(C_i, C_j\right) = \frac{\overline{S}_{EC_{\{C_i,C_j\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\overline{S}_{EC_{C_i}} + \frac{|C_i|}{|C_i|+|C_j|}\overline{S}_{EC_{C_j}}} \tag{۴.۲}$$

که در آن $\overline{S}_{EC_{\{C_i,C_j\}}}$ میانگین وزن یال‌هایی است که خوشه C_i را به خوشه C_j متصل می‌کند و $\overline{S}_{EC_{C_i}}$ (یا $\overline{S}_{EC_{C_j}}$) میانگین وزن یال‌هایی در خوشه C_i (یا C_j) است که با حذف آن‌ها خوشه مزبور به دو قسمت تقریباً مساوی شکسته می‌شود.

هرگاه مقدار هر دو کمیت برای جفت خوشه‌ای بالا باشد، الگوریتم تجمیعی CHAMELEON آن دو خوشه را باهم ادغام می‌کند. در مرحله اول، مدل‌بندی پویا از مشاهدات با خوشه‌بندی آن‌ها به زیرخوشه‌ها انجام می‌شود. در مرحله دوم، یک چارچوب مدل‌بندی پویا روی مشاهدات بکار می‌رود تا به روش سلسله‌مراتبی زیرخوشه‌ها باهم ادغام شوند و خوشه‌های با کیفیت بهتری به دست آید. این چارچوب مدل‌بندی پویا را به دو روش متفاوت می‌توان بکار برد. در روش اول، مقادیر به‌هم‌پیوستگی نسبی و نزدیکی نسبی بین یک جفت خوشه بررسی می‌شود که از مقدار آستانه مشخص‌شده توسط کاربر بیشتر نباشد. برای این منظور، باید این دو پارامتر در روابط زیر صدق کنند:

$$۱)\ \ RI\left(C_i, C_j\right) \geq T_{RI},$$

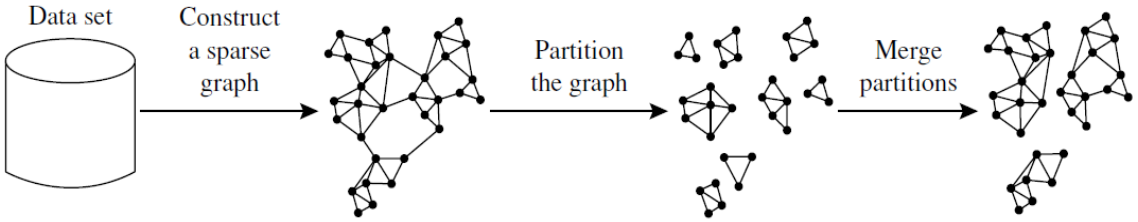
$$۲)\ \ RC\left(C_i, C_j\right) \geq T_{RC}$$

در روش دوم، الگوریتم CHAMELEON جفت خوشه‌ایی را انتخاب می‌کند که تابع زیر را ماکسیمم کند:

$$RI\left(C_i, C_j\right) * RC(C_i, C_j)^{\alpha}$$

که در آن α پارامتر مشخص‌شده توسط کاربر است که مقادیر بین ۰ و ۱ را می‌گیرد. فرآیند الگوریتم CHAMELEON در شکل ۱ نمایش داده شده است.

[۱۳] Graph sparse



شکل ۱: فرآیند الگوریتم CHAMELEON(هان و همکاران ، ۲۰۰۱)

۳.۲ روش‌های چگالی‌پایه

روش‌های دیگری برای خوشه‌بندی بر مبنای مفهوم چگالی ایجاد شده است که خوشه‌ها را به‌عنوان نواحی با چگالی بالا از مشاهدات در فضای داده‌ای در نظر می‌گیرند که به‌وسیله نواحی کم چگال از یکدیگر جدا شده‌اند. روش‌های چگالی پایه می‌توانند داده‌های دورافتاده و نوفه را شناسایی و همچنین خوشه‌های با هر شکل دلخواه را پیدا کنند. از جمله الگوریتم‌های چگالی‌پایه که برای کشف خوشه‌ها بکار میروند می‌توان به DBSCAN، OPTICS و DENCLU اشاره کرد. در این بخش این سه الگوریتم روش چگالی‌پایه را به‌تفصیل بیان خواهیم کرد.

۱.۳.۲ الگوریتم DBSCAN

چگالی مشاهده‌ای مانند o را می‌توان با کمک تعداد مشاهدات نزدیک به آن اندازه‌گیری کرد. الگوریتم DBSCAN یک روش خوشه‌بندی چگالی‌پایه به دنبال مشاهدات دارای همسایگی‌های متراکم است و این مشاهدات را مشاهدات هسته[۱۴] می‌نامد. پس از آن با اتصال مشاهدات هسته و همسایه‌های آن‌ها مناطق متراکم به دست آمده به عنوان خوشه‌ها در نظر گرفته می‌شوند. روش کار این الگوریتم به این صورت است که با کمک مقدار ε که توسط کاربر مشخص می‌شود، یک شعاع همسایگی برای هر یک از مشاهدات تعیین می‌گردد. شعاع همسایگی ε برای مشاهده o فضای دایره شکل به مرکز o و شعاع ε است. برای تعیین چگالی همسایگی یک مشاهده، این الگوریتم پارامتر دیگری به نام MinPts معرفی می‌کند که این پارامتر نیز توسط کاربر مشخص می‌شود. چنانچه در شعاع همسایگی ε بتوان حداقل به تعداد MinPts مشاهده پیدا کرد، آن مشاهده به عنوان هسته شناخته می‌شود. بر اساس یک تابع فاصله، شکل همسایگی نقاط به دست می‌آید و فاصله نقاط p و q با $dist(p,q)$ نشان داده می‌شود. به طور کلی خوشه‌بندی حاصل از الگوریتم DBSCAN از قواعد زیر پیروی می‌کند:

(۱) یک مشاهده می‌تواند متعلق به یک خوشه معین باشد اگر و تنها اگر درون εـ همسایگی مشاهده هسته‌ای آن خوشه باشد.

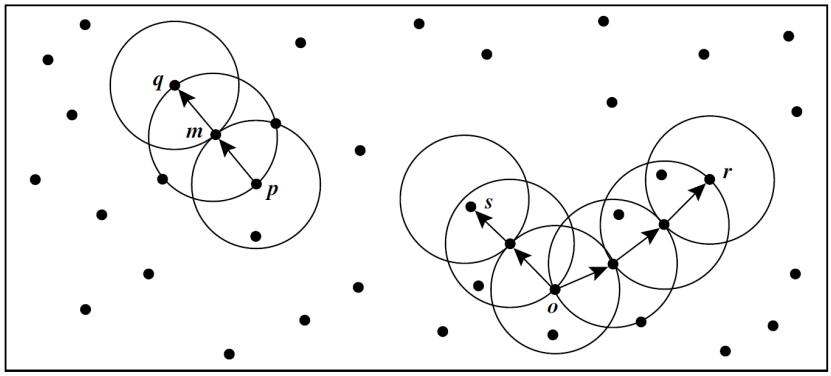
(۲) اگر یک مشاهده هسته‌ای مانند p درون εـ همسایگی مشاهده هسته‌ای دیگری مانند q قرار گیرد باید هر دو متعلق به خوشه یکسانی باشند.

(۳) اگر یک مشاهده غیرهسته‌ای q درون εـ همسایگی برخی مشاهدات هسته‌ای $p_1, \ldots, p_i$، $i > 0$ قرار گیرد، باید متعلق به خوشه حداقل یکی از مشاهدات هسته‌ای $p_1, \ldots, p_i$ باشد.

[۱۴] Object Core

(۴) یک مشاهده غیرهسته‌ای که درون هیچ εـ همسایگی مشاهدات هسته‌ای ارزیابی‌شده قرار نگیرد، به عنوان نوفه شناخته می‌شود.

بنابراین الگوریتم DBSCAN برای کشف خوشه‌ها در داده‌ها، εـ همسایگی همه نقاط در مجموعه داده‌ها را بررسی می‌کند. اگر εـ همسایگی نقطه p بیشتر از MinPts را شامل شود آنگاه یک خوشه جدید با در نظر گرفتن p به‌عنوان مشاهده هسته‌ای آن ایجاد می‌شود. مشاهداتی که در εـ همسایگی p قرار دارند به این خوشه جدید اضافه می‌شوند. فرایند وقتی خاتمه می‌یابد که هیچ نقطه جدیدی به خوشه‌ای اضافه نشود. شکل ۲ روش کار الگوریتم DBSCAN را نشان می‌دهد.



شکل ۲: فرآیند الگوریتم DBSCAN (هان و همکاران ، ۲۰۰۱)

۲.۳.۲ الگوریتم OPTICS

علیرغم اینکه الگوریتم DBSCAN قادر است خوشه‌های دارای شکل‌های دلخواه را در داده‌های حاوی نوفه شناسایی کند، اما به مقادیر ورودی ε و MinPts حساس است. برای کشف خوشه‌ها با در نظر گرفتن خوشه‌های قابل‌قبول، کاربر باید الگوریتم را چندین بار اجرا کند تا مقادیر پارامترهای متفاوت را در هر اجرا بررسی کند. بنابراین زمان زیادی را باید صرف نماییم تا مقادیر مناسبی برای این دو پارامتر بیابیم. برای غلبه بر این مشکل، یک روش مبتنی بر مفهوم نظم‌دهی خوشه بر مبنای چگالی به نام OPTICS پیشنهاد شده است که یک تعمیم از الگوریتم DBSCAN می‌باشد. روش OPTICS یک الگوریتم خوشه‌بندی چگالی‌پایه است که یک خوشه‌بندی ضمنی در مجموعه داده موجود شناسایی می‌کند. برخلاف روش DBSCAN که برای شناسایی خوشه به دو پارامتر بستگی داشت، روش OPTICS از چند پارامتر استفاده می‌کند. این الگوریتم به‌جای ارائه نتیجه‌ی خوشه‌بندی برای یک جفت مقدار پارامتر، یک نظم‌دهی از نقطه داده‌ها انجام می‌دهد به‌نحوی‌که نتیجه خوشه‌بندی برای کمترین مقدار ε و مقدار متناظر MinPts به‌سادگی محاسبه شود. این الگوریتم از روش‌های مصورسازی برای مجموعه داده‌های چندبعدی بزرگ استفاده می‌کند. همچنین از تکنیک‌های خودکار برای تشخیص شروع و پایان ساختارهای خوشه‌ای بهره می‌گیرد تا الگوریتم را شروع کند و بعد آن‌ها را باهم گروه‌بندی کند تا یک مجموعه از خوشه‌های تودرتو ایجاد کند. اما ایراد اصلی هر دو الگوریتم DBSCAN و OPTICS این است که برای داده‌های بُعد بالا چندان کارا نیستند.

۳.۳.۲ الگوریتم DENCLUE

یک روش کارا برای خوشه‌بندی داده‌های با بعد بالا الگوریتم DENCLUE است. این الگوریتم یکی از روش‌های خوشه‌بندی چگالی‌پایه است که بر اساس توابع توزیع چگالی کار خود را انجام می‌دهد. در حوزه آمار و احتمال، برآورد چگالی به فرآیند تخمین یک تابع چگالی احتمال یک متغیر تصادفی بر اساس مجموعه‌ای از مشاهدات آن متغیر گفته می‌شود. در زمینه خوشه‌بندی چگالی‌پایه این تابع یک توزیع واقعی از همه مشاهداتی است که باید تحلیل شوند. مجموعه داده‌ها به عنوان یک نمونه تصادفی از یک جامعه محسوب می‌شوند. برای این منظور می‌توان از برآورد چگالی کِرنل استفاده کرد که یک رویکرد برآورد چگالی بدون پارامتر (با کمک آماره‌ها) تلقی می‌شود.

تعریف ۱.۲. *فرض کنید $x_1, x_2, \ldots, x_n$ نمونه‌ی تصادفی و مستقل از یک توزیع احتمال f باشد. تقریب چگالی کِرنل از این تابع چگالی احتمال برابر است با:*

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{۵.۲}$$

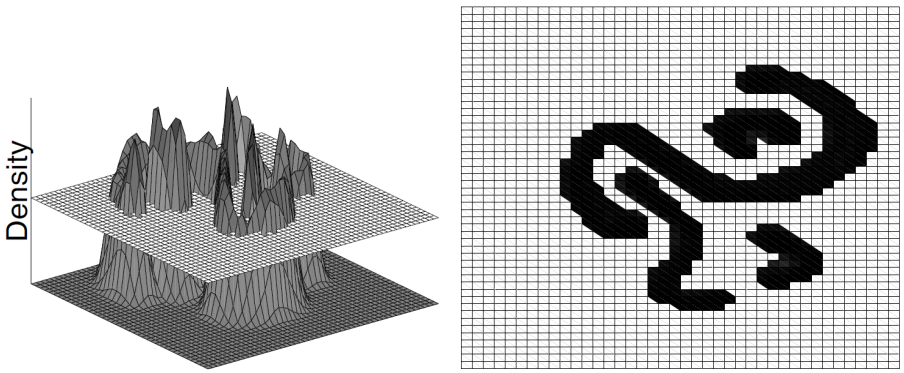
که در آن $K(\cdot)$ یک تابع کِرنل است و h نیز به عنوان پارامتر هموارسازی استفاده می‌شود.

یک کرنل را می‌توان به عنوان یک تابع تاثیر در نظر گرفت که تاثیر یک نقطه‌ی نمونه را بر همسایه‌های آن مدل‌سازی می‌کند.

یکی از کِرنل‌های پرکاربرد تابع گاوسی استاندارد است که به صورت زیر بیان می‌شود:

$$K(\frac{x - x_i}{h}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x - x_i)^2}{2h^2}\right\}$$

الگوریتم DENCLUE بهتر و کاراتر از الگوریتم DBSCAN عمل می‌کند و در مواجهه با نوفه عملکرد ثابتی دارد. اما این الگوریتم نیز نیاز به انتخاب دقیق پارامترهای خوشه‌بندی دارد که تأثیر قابل‌توجهی بر کیفیت نتایج خوشه‌بندی دارند. شکل ۳ روش کار الگوریتم DBSCAN را نشان می‌دهد.



شکل ۳: نمونه‌ای از خوشه‌بندی چگالی‌پایه با استفاده از الگوریتم DENCLUE (هان و همکاران ، ۲۰۰۱)

۴.۳.۲ خوشه‌بندی بر اساس ابَرگراف

الگوریتم خوشه‌بندی بر مبنای ابَرگراف، یک الگوریتم چگالی‌پایه است که با استفاده از نموداری به نام وُرُنویی[۱۵] عمل می‌کند. روش کار این الگوریتم به این صورت است که ابتدا از مجموعه مشاهدات فضایی، نمودار وُرُنویی ایجاد می‌شود. سپس مثلث‌های دلونِی[۱۶] از این نمودار ساخته می‌شوند که هر یک از این مثلث‌ها متشکل از سه‌نقطه (بیانگر سه مشاهده) هستند. اگر طول ترکیبی یک مثلث بسیار بیشتر از مثلث‌های دیگر باشد آنگاه میزان نزدیکی در میان سه مشاهده کم است و درنتیجه آن مثلث حذف می‌شود. اصولاً یک ابرگراف نزدیکی و شباهت بین مشاهدات همسایه را نشان می‌دهد و از نمودار منظم به دست می‌آید. یک نمودار ابرگراف را با نماد $G(V,E)$ نمایش می‌دهند که در آن رئوس V نشانگر مشاهدات و مجموعه یال‌های E بیانگر میزان نزدیکی و شباهت مشاهدات به هم هستند. هر یک از رأس‌ها در ابرگراف ممکن است بیش از یک یال داشته باشند که به میزان نزدیکی که آن با بقیه مشاهدات درون خوشه دارد، بستگی دارد. وزن یال‌های ابرگراف درجه تشابه میان مشاهدات را بیان می‌کند. وزن یال در ابرگراف از طول کلی یال‌های مثلث متناظرش تعیین می‌شود و از تابع زیر محاسبه می‌شود:

$$W\left(e_H\right)=\frac{\sum_{e\in T_H}\text{length}\left(e\right)}{\sum_{e\in T_H}\left(\text{length}\left(e\right)\right)^{۲}} \qquad (۶.۲)$$

این الگوریتم خوشه‌بندی در دو مرحله اجرا می‌شود. در مرحله اول یک مجموعه از خوشه‌های اولیه مشخص می‌شوند و در مرحله دوم که مرحله خوشه‌بندی سلسله‌مراتبی الگوریتم است، از بین خوشه‌های اولیه مرحله اول آن‌هایی که بهم نزدیک‌ترند داخل خوشه‌های یکسان گروه‌بندی شده و خوشه‌های جدیدی ایجاد می‌شوند. فرایند تا زمانی ادامه می‌یابد که یک مجموعه از خوشه‌ها با بیشترین شباهت درون خوشه‌ای ایجاد شود.

به طورکلی، ابرگراف‌ها و مثلث‌های دِلونی در الگوریتم‌های خوشه‌بندی فضایی مورد استفاده قرار می‌گیرند. به عنوان مثال یانگ و کویی (۲۰۱۰) یک الگوریتم خوشه‌بندی فضایی استوار به نام NSCABDT بر اساس مثلث‌بندی دِلونی پیشنهاد داده‌اند که نمودار دِلونی برای تعیین همسایگی‌ها بر اساس مفهوم همسایگی و روابط فضایی بکار می‌رود. تشخیص اشکال دلخواه از خوشه‌ها، دقت و صحت در شناسایی خوشه‌ها، عدم نیاز به اطلاعات پیشین از خوشه‌ها، زمان اجرای مناسب و عملکرد با کفایت در مواجهه با داده‌های پرت، چند مزیت مهم این الگوریتم است

۴.۲ روش‌های شبکه‌پایه

روش‌های خوشه‌بندی زیربخش‌های قبل داده محور هستند. به عبارت دیگر آن روش‌ها با توجه به توزیع داده‌ها، آن‌ها را گروه‌بندی می‌کنند. با توجه به بخش قبلی، کارایی روش‌های چگالی پایه مانند DBSCAN و OPTICS با بالا رفتن تعداد ابعاد مشاهدات، کاهش می‌یابد. برای افزایش کارایی، روش خوشه‌بندی شبکه‌پایه که از رویکرد فضا محور بهره می‌گیرد معرفی شده است. در این روش‌ها، فضای مورد نظر به تعداد متناهی سلول‌ها مستقل از توزیع مشاهدات ورودی افراز می‌شود و در آن‌ها یک ساختمان شبکه‌ای (توری شکل) با چندین درجه دقت و وضوح بکار می‌رود. کلیه عملیات مربوط به خوشه‌بندی بر روی این ساختار شبکه‌ای انجام می‌شود. یکی از مزیت‌های اصلی این روش‌ها سرعت بالای پردازش است. به‌عنوان مثال‌هایی از خوشه‌بندی بر مبنای شبکه، می‌توان به الگوریتم STING که اطلاعات آماری ذخیره‌شده در هر سلول شبکه را استخراج می‌کند، الگوریتم CLIQUE که روشی بر مبنای

[۱۵] Voronoi
[۱۶] Delaunay

چگالی و شبکه برای خوشه‌بندی داده‌های بعد بالا معرفی می‌کند، الگوریتم WaveCluster که خوشه‌بندی مشاهدات را با استفاده از روش تبدیل موجک انجام می‌دهد، اشاره کرد.

۱.۴.۲ الگوریتم STING

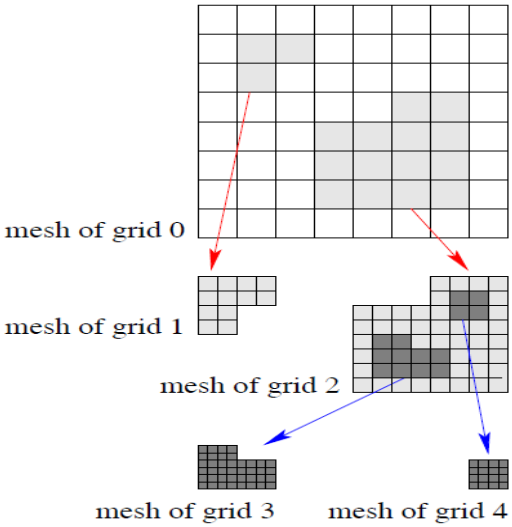
الگوریتم STING یک روش خوشه‌بندی شبکه‌پایه است که با توجه به‌وضوح ساختار داده‌ها، ناحیه فضایی را به سلول‌های مستطیل شکل تقسیم می‌کند. معمولاً چندین سطح از این سلول‌های مستطیلی متناظر با سطوح مختلف وضوح وجود دارد و این سلول‌ها یک ساختار سلسله مراتبی ایجاد می‌کنند. هر سلول در یک سطح بالا به تعدادی سلول در سطح پایین‌تر افراز می‌شود. اطلاعات آماری راجع به متغیرهای هر سلول شبکه (مانند میانگین، ماکسیمم و مینیمم) از قبل محاسبه شده و ذخیره می‌شوند. پارامترهای آماری سلول‌های سطح بالاتر به‌راحتی از پارامترهای سلول‌های سطح پایین‌تر محاسبه می‌شوند. این پارامترها شامل شمارش مقادیر، میانگین، انحراف استاندارد، مینیمم، ماکسیمم و نوع توزیعی که مقدار متغیر در سلول دارد مانند نرمال، یکنواخت، نمایی یا توزیع نامعلوم، می‌باشد. داده‌ها از پایین‌ترین سطح که همان سطح دارای بالاترین وضوح است، وارد ساختار می‌شوند. پارامترهای میانگین، انحراف استاندارد، مینیمم و ماکزیمم مربوط به پایین‌ترین سطح، مستقیماً از داده‌ها محاسبه می‌شوند. اگر نوع توزیع از قبل معلوم باشد مقدار توزیع توسط کاربر تعیین می‌شود. همچنین می‌توان با آزمون فرض مانند آزمون χ^2 آن را به دست آورد. نوع توزیع سلول سطح بالاتر را می‌توان بر اساس اکثریت توزیع‌های متناظر سلول‌های سطح پایین‌تر در رابطه با یک فرایند فیلترسازی آستانه محاسبه کرد. اگر توزیع سلول‌های سطوح پایینتر متفاوت از همدیگر باشند و آزمون آستانه رد شود، نوع توزیع سلول سطح بالا نامعلوم در نظر گرفته می‌شود.

برای انجام خوشه‌بندی در چنین ساختار داده‌ای، باید ابتدا یک سطح توزیعی را به‌عنوان پارامتر ورودی در نظر بگیریم. با استفاده از این پارامتر، برای یافتن نواحی دارای چگالی مناسب، روش شبکه‌پایه از بالا به پایین بکار می‌رود. ابتدا یک سطح درون ساختار سلسله‌مراتبی ازآنچه فرایند شروع‌شده، تعیین می‌شود. معمولاً، این سطح شامل تعداد کمی سلول است. برای هر سلول در سطح اخیر، فاصله اطمینان اینکه سلول مرتبط با نتیجه خوشه‌بندی خواهد بود، محاسبه می‌کنیم. به نظر می‌رسد سلول‌هایی که این فاصله اطمینان را قطع نمی‌کنند غیر مرتبط‌اند و از بررسی بیشتر کنار گذاشته می‌شوند و از فرایند حذف می‌شوند. برای پردازش لایه بعدی تنها سلول‌های باقیمانده در نظر گرفته می‌شوند. این فرایند تا جایی تکرار می‌شود که سطح پایینی به دست آید. در شکل ۴ سه لایه متوالی یک ساختار STING نشان داده شده است که هر سلول در مرحله بعدی به سلول‌های ریزتر تقسیم می‌شود.

۲.۴.۲ الگوریتم CLIQUE

الگوریتم CLIQUE دو روش خوشه‌بندی چگالی‌پایه و شبکه‌پایه را باهم ادغام می‌کند. برخلاف الگوریتم‌های خوشه‌بندی دیگر که توصیف شد، این روش قادر به کشف خوشه‌هایی در زیرفضای داده‌هاست و برای داده های با بُعد بالا و پراکندگی زیاد که در فضای بُعد کامل خوشه ها ایجاد نمی‌شوند، مناسب است. الگوریتم CLIQUE هر بُعد را به بازه‌هایی افراز می‌کند و بدین ترتیب کل فضای داده‌ها به درون سلول‌های مستطیلی شکل غیر مشترک تقسیم می‌شود. این الگوریتم از یک حد آستانه چگالی برای شناسایی سلول‌های متراکم و تُنُک استفاده می‌کند. یک سلول هنگامی متراکم است که تعداد مشاهدات درون آن بیشتر از حد آستانه چگالی (تراکم) باشد.

الگوریتم کار خود را با یک سلول متراکم دلخواه آغاز می‌کند، منطقه ماکسیمالی که سلول را پوشش می‌دهد را پیدا می‌کند و



شکل ۴: نمونه‌ای از خوشه‌بندی شبکه‌پایه با استفاده از الگوریتم STIN (لیائو و همکاران ، ۲۰۰۴)

پس از آن به سراغ سلول‌های متراکم دیگری می‌رود که هنوز پوشش داده نشده‌اند. هرگاه تمام سلول‌های متراکم پوشش داده شوند کار خود را به پایان می‌رساند. الگوریتم CLIQUE به صورت خودکار به دنبال زیرفضاهایی است که در آن خوشه‌هایی با چگالی بالا یافت می‌شود. این الگوریتم به ترتیب ورودی داده‌ها حساس نیست و همچنین هیچ فرضی درباره توزیع داده‌ها لحاظ نمی‌کند و در مواجهه با افزایش ابعاد داده‌ها نیز از مقیاس‌پذیری خوبی برخوردار است.

۳.۴.۲ الگوریتم WaveCluster

الگوریتم WaveCluster جزو الگوریتم‌های خوشه‌بندی شبکه‌پایه محسوب می‌شود که ابتدا با اِعمال ساختار شبکه‌ای چندبعدی به فضای داده‌ها، باعث خلاصه‌سازی داده‌ها می‌شود. سپس از تبدیل موجک برای تغییر فضای ویژگی اصلی استفاده می‌کند. درنهایت نواحی چگال در فضای تبدیل‌یافته را می‌یابد. در این روش هر سلول شبکه، اطلاعات گروهی از مشاهدات داخل سلول را خلاصه می‌کنند. معمولاً این خلاصه اطلاعات برای استفاده از تبدیل موجک و تحلیل خوشه‌ای بعدی، درون حافظه اصلی قرار می‌گیرد. تبدیل موجک یک تکنیک پردازش سیگنال است که در آن یک سیگنال به زیرباندهای فرکانس‌های مختلف تجزیه می‌شود. مدل‌های موجی را می‌توان برای سیگنالهای n بُعدی به صورت n بار تکرار یک تبدیل موجک یک بعدی بکار برد. این الگوریتم از یک تابع کِرنل مناسب در یک فضای تبدیل یافته استفاده می‌کند. سپس با جستجو برای نواحی چگال در دامنه جدید، خوشه ها را تعیین می‌کند.

روش WaveCluster برای مجموعه داده‌های فضایی با بُعد بالا کارا است، خوشه های با شکل‌های دلخواه و با کیفیت بالا را کشف می‌کند، داده‌های پرت را به درستی شناسایی و رفتار می‌کند، به رتبه ورودی حساس است و نیاز به پارامترهای بیرونی و یک دانش قبلی برای تشکیل خوشه‌های خوب ندارد. در مطالعات تجربی، از نظر کارایی وکیفیت خوشه‌بندی، این روش بهتر از روش‌های BIRCH، CLARANS و DBSCAN عمل کرده است.

۳ بحث و نتیجه‌گیری

در این مقاله الگوریتم‌های خوشه‌بندی داده‌های چندبعدی مانند داده‌های فضایی و الگوریتم‌هایی که برای داده‌های جغرافیایی مفید هستند، از نظر کارایی برای چنین داده‌هایی، شکل و کیفیت خوشه‌های خروجی و سرعت اجرا موردمطالعه قرار گرفتند. بیان کردیم که به طور کلی می‌توان این الگوریتم‌ها به چهار دسته‌ی افرازی، سلسله‌مراتبی، چگالی‌پایه و شبکه‌پایه تقسیم کرد.

روش‌های افرازی مانند k-means، k-medoids و EM خوشه‌های کروی شکل و ازنظر اندازه مشابه را پیدا می‌کنند. این روش‌ها برای کاربردهایی که به‌طور عینی نمی‌توان خوشه طبیعی یافت اما مجموع فاصله‌ی مشاهدات از مراکز خوشه‌ها مینیمم می‌شود، مناسب‌تر می‌باشند.

برخلاف الگوریتم خوشه‌بندی بر مبنای افراز که برای بهبود کیفیت خوشه‌بندی، مشاهدات را از یک خوشه به خوشه دیگر اختصاص می‌دهند، در الگوریتم‌های خوشه‌بندی سلسله‌مراتبی عضویت یک مشاهده به هر خوشه‌ای تثبیت می‌شود یعنی هر مشاهده به هر خوشه‌ای اختصاص یابد تا پایان خوشه‌بندی در همان خوشه باقی می‌ماند که یکی از ایرادات اصلی این نوع روش‌ها است. با وجود روش خیلی ساده برای ادغام یا تقسیم خوشه‌ها در الگوریتم‌های خوشه‌بندی سلسله‌مراتبی، الگوریتم‌هایی مانند AGNES و DIANA احتمال ایجاد خوشه‌های نادرست دارند. الگوریتم خوشه‌بندی سلسله مراتبی BIRCH با ایجاد درخت‌های CF برای مجموعه داده‌ها، خوشه‌بندی انجام می‌دهد. این الگوریتم برای اجرا، به مقدار آستانه T بستگی دارد و در مواجهه با نقاط دورافتاده با یک مقدار آستانه بهینه، قابل‌اجرا است. برای یافتن خوشه‌های دارای کیفیت بهتر می‌توان از الگوریتم BIRCH استفاده کرد که این الگوریتم ابتدا روی داده‌ها خوشه‌بندی سلسله‌مراتبی انجام می‌دهد تا برای بهبود کیفیت خوشه‌بندی، داده‌ها را نسبت به مشخصه‌های خوشه‌بندی متراکم کند. BIRCH برای پایگاه داده‌های بزرگ کنترل‌پذیر است.

CURE یکی دیگر از الگوریتم‌های خوشه‌بندی سلسله‌مراتبی است که به یک پارامتر با نام فاکتور انقباض وابسته است. پارامتر فاکتور باید یک مقدار بهینه برای خوشه‌بندی کارا از مشاهدات داشته باشد. بعلاوه اینکه فاکتور انقباض اثرات بد داده‌های دورافتاده را به‌طور قابل توجهی کاهش می‌دهد. الگوریتم CURE می‌تواند خوشه‌های دارای شکل‌های نامنظم را حتی در ابعاد بالاتر شناسایی کند. CHAMELEON یک الگوریتم سلسله‌مراتبی تجمیعی است. پارامترهای این الگوریتم بهم پیوستگی نسبی و نزدیکی نسبی بین خوشه‌ها است که در برابر یک مقدار آستانه بهینه بررسی می‌شوند و خوشه‌های دارای کیفیت خوب تولید می‌کند. اگرچه در الگوریتم‌های CURE و CHAMELEON تقسیم و ادغام همچنان برگشت‌ناپذیر است اما به دلیل استفاده از روش بهتر برای تقسیم یا ادغام، خطاهای کمتری ایجاد می‌شوند. هر دو روش در بهبود کیفیت خوشه‌بندی موفق بوده‌اند.

روش‌های خوشه‌بندی چگالی‌پایه به‌جای استفاده از فاصله برای تشخیص عضویت مشاهدات، الگوریتم‌هایی مانند DBSCAN را که چگالی نقاط درون یک ناحیه معین را در نظر می‌گیرند، برای کشف خوشه‌ها بکار می‌برد. برای تعریف چگالی پیرامون ناحیه یک مشاهده، DBSCAN به مقادیر دو پارامتر ε و MinPts نیاز دارد. DBSCAN برای خوشه‌بندی داده‌های چندبعدی دارای نوفه قابل‌اجرا است. محدودیت این روش این است که کاربر باید به هر دو پارامتر مقدار درست دهد. برای غلبه به این مشکل، الگوریتم دیگری به نام OPTICS معرفی شده است که برخلاف DBSCAN از تعداد زیادی پارامتر فاصله برای پردازش و گروه‌بندی مشاهدات به داخل خوشه‌ها استفاده می‌کند. با توجه به یک مشاهده هسته، الگوریتم OPTICS یک نظم‌دهی کارا برای خوشه‌بندی مجموعه‌ای از مشاهدات منظم شده با مقدار قابل‌دسترس و مقادیر هسته ایجاد می‌کند. OPTICS داده‌های چندبعدی را با یک تکنیک مصورسازی پردازش می‌کند و به روش خودکار، مجموعه‌ای از خوشه‌های تودرتو تعیین می‌کند. در واقع یک رتبه‌بندی

خوشه‌ای افزایشی برای دامنه وسیعی از مجموعه‌های پارامتری محاسبه می‌کند و به کاربران امکان تحلیل خوشه‌ای تکراری را می‌دهد. شرط اولیه هر دو الگوریتم DBSCAN و OPTICS باعث کاهش کارایی آن‌ها برای خوشه‌بندی داده‌های بُعد بالا و داده‌های فضایی می‌شود. این مسئله به‌وسیله الگویتم DENCLUE بررسی‌شده است. این الگوریتم چگالی کلی نقاط را به‌صورت مجموع توابع تأثیر نقطه داده‌ها مدل‌بندی می‌کند و یک ساختار شبکه‌ای کارا برای انجام محاسبه بکار می‌گیرد. این الگوریتم نیز نیاز به انتخاب دقیق پارامترهای خوشه‌بندی دارد. یک الگوریتم خوشه‌بندی بر اساس اَبَرگراف معرفی شد که الگوریتم چگالی‌پایه است. ابرگراف با استفاده از ایجاد مثلث‌های دلونی روی مجموعه داده‌های فضایی ساخته می‌شود.

روش‌های خوشه‌بندی شبکه‌پایه برای افزایش کارایی خوشه‌بندی، نواحی چگال فضای خوشه‌بندی را به تعداد محدودی از سلول‌ها تقسیم کرده و به شناسایی سلول‌هایی که شامل بیشتر از تعداد نقاط چگال هستند، می‌پردازند. سپس خوشه‌ها با اتصال این سلول‌های چگال ایجاد می‌شوند. روش‌های خوشه‌بندی شبکه‌پایه شامل الگوریتم‌های STING، CLIQUE و WaveCluster هستند. STING اطلاعات آماری موجود در سلول‌های شبکه را استخراج می‌کند. الگوریتم CLIQUE که روشی بر مبنای چگالی و شبکه برای خوشه‌بندی داده‌های بُعد بالا معرفی می‌کند و خوشه‌ها را در زیرفضاها کشف می‌کند. این الگوریتم برای خوشه‌بندی داده‌های با ابعاد بالا کارا است. روش خوشه‌بندی WaveCluster نیز با استفاده از تبدیل موجک مشاهدات را خوشه‌بندی می‌کند. در WaveCluster نشان داده شد که پارامترهای بیرونی و یک دانش قبلی برای تشکیل خوشه‌های خوب مورد نیاز نیست. مزیت اصلی روش‌های شبکه‌پایه سرعت بالای فرایند است که معمولاً مستقل از تعداد مشاهدات بوده و فقط به تعداد سلول‌های پایین‌ترین سطح بستگی دارد که خیلی کمتر از تعداد داده‌ها می‌باشد. همچنین معمولاً یک روش شبکه‌پایه نسبت به روش چگالی‌پایه خصوصاً در فضای بعد بالا کاراتر است.

در حال حاضر افراد بسیاری با انجام روش‌ها و طرح الگوریتم‌های خوشه‌بندی مختلف و جدید در صدد بهبود هر دو معیار اثر بخشی و کارایی در ابعاد بالاتر هستند. محققان همچنان به منظور ارتقاء و تغییر الگوریتم‌های قدیمی‌تر می‌پردازند و در تلاش برای توسعه الگوریتم‌هایی هستند که می‌توانند برای پایگاه داده‌های بزرگ و دارای ابعاد بالاتر کارا باشند. شش نیاز منحصر به فرد برای الگوریتم های خوشه‌بندی پایگاه داده‌های فضایی هنوز هم معیار موثر برای اندازه‌گیری الگوریتم‌های جدید و به روز هستند و در آینده الگوریتم های بیشتری توسعه می‌یابند که با این معیارها مطابقت بیشتری داشته باشند.

مراجع

محمدزاده م، (۱۳۹۴). *آمار فضایی و کاربردهای آن*، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،

Berkhin, P. (2002). Survey of Clustering Data Mining Techniques, Technical Report, *Accrue Software, Inc* .

Chandra, E. and Anuradha, V. P. (2011). A Survey on Clustering Algorithms for Data in Spatial Database Management Systems, *International Journal of Computer Applications*, **24**(9), 19–26.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*, Third Edition, Elsevier.

Han, J., Kamber, M., and Tung, A. K. H. (2001). Spatial Clustering Methods in Data Mining: A Survey, *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS*, 1–29.

Kolatch, E. (2001). Clustering Algorithms for Spatial Databases: A survey, *Department of Computer Science. University of Maryland, College Park., http//www.cs.umd.edu/ kolatch/index.html*, 1–22.

Liao, W., Liu Y. and Choudhary A. (2004). A Grid-based Clustering Algorithm using Adaptive Mesh Refinement, *Appears in the 7th Workshop on Mining Scientific and Engineering Datasets* .

Mitchell, T. (1997). *Machine Learning*, McGraw-Hill, New York.

Roddick, J. F. and Hornsby, K. (Eds) (2001). Temporal, Spatial and Spatio-Temporal Data Mining, *First International Workshop TSDM 2000 Lyon*, France,September 12.

Tao, Y., Zhang, J., Papadias, D. and Mamoulis N. (2004). An Efficient Cost Model for Optimization of Nearest Neighbor Search in Low and Medium Dimensional Spaces, *IEEE Transactions on Knowledge and Data Engineering* , **16**(10), 1169–1184.

Varghese, B. M. and Unnikrishnan, A. (2013). Spatial Clustering Algorithms: An Overview, *Asian Journal of Computer Science and Information Technology*, **3**(1), 1–8.

Wang, X. and Hamilton, H. J. (2005). Clustering Spatial Data in The Presence of Obstacles, *International Journal on Artificial Intelligence Tools*, **14**, 177–198.

Yang, X. and Cui, W. (2010). A Novel Spatial Clustering Algorithm Based on Delaunay Triangulation, *Journal of Software Engineering and Applications*, **3**(2), 141–149.

دومین سمینار آمار فضایی و کاربردهای آن
۳ – ۴ آبان ۱۳۹۶
دانشگاه صنعتی شاهرود
انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1390

برآورد ML نواحی کوچک برای پاسخ‌های گامای شمارشی فضایی با روش HDC

مهسا نادی‌فر*[۱]، حسین باغیشنی[۱] ، افشین فلاح[۲]

[۱]گروه آمار، دانشگاه صنعتی شاهرود

[۲]گروه آمار، دانشگاه بین‌المللی امام خمینی قزوین

چکیده: در کاربردهای نواحی کوچک، پاسخ‌های شمارشی به فراوانی مشاهده می‌شوند. چالش بیش‌پراکنش یا کم‌پراکنش موجود در داده‌های شمارشی، همیشه امکان استفاده از توزیع پواسون برای مدل‌بندی آن‌ها را دچار خدشه می‌کند. به منظور لحاظ کردن عدم پراکنش متعادل در این نوع پاسخ‌ها، یک مدل رگرسیونی فضایی جدید را برای داده‌های مشبکه‌ای در نواحی کوچک، معرفی می‌کنیم. زیربنای این مدل بر پایه توزیع گامای شمارشی شکل گرفته است. با توجه به پیچیدگی محاسبه تابع درستنمایی مدل، برای استنباط مبتنی بر درستنمایی از یک روش تقریبی ترکیبی معروف به HDC استفاده می‌کنیم. کارایی مدل پیشنهادی را در یک مطالعه شبیه‌سازی مورد ارزیابی قرار می‌دهیم و کاربرد آن را در تحلیل یک مجموعه داده واقعی مربوط به مرگ و میر حاصل از سرطان حنجره در آلمان، نمایش می‌دهیم.

واژه‌های کلیدی: نواحی کوچک، برآورد ML، گامای شمارشی، همسانه‌سازی داده‌ها، HDC، INLA.

کد موضوع‌بندی ریاضی (۲۰۱۰): 06 – 62، 62J12، 62F15، 62H11.

۱ مقدمه

برآورد نواحی کوچک از دیرباز مورد توجه دولت‌ها و نهادهای مختلف تصمیم‌گیرنده قرار گرفته است. روش‌های برآورد نواحی کوچک شامل حل مباحثی هستند که با روش‌های معمول آماری نمی‌توان پاسخ مناسبی برای آن‌ها فراهم کرد. روش‌های نمونه‌گیری برای استفاده در نواحی بزرگ با تعداد مشاهدات کافی قابل استفاده هستند. اما افزایش تقاضا برای برآوردهایی از نواحی‌ای که به اندازه کافی نمونه ندارند و عدم وجود برآوردگرهای قابل اعتماد در این‌گونه موارد، محققان را بر آن داشته‌اند روش‌هایی معرفی کنند که بتوان با نمونه کم نیز برآوردهای قابل اعتمادی به‌دست آورد. برای مثال، فرض کنید می‌خواهیم نرخ بیکاری استان‌های کشور را

*ارایه‌دهنده مقاله: مهسا نادی‌فر، m.nadifar@shahroodut.ac.ir

به‌دست آوریم. طبق روش‌های نمونه‌گیری می‌دانیم با بهینه کردن اندازه نمونه برای استان‌های کشور، می‌توان برآوردهای مستقیم را به‌دست آورد که از دقت مناسبی برخوردار هستند. حال اگر بخواهیم برای شهرستان‌ها نیز نرخ بیکاری را برآورد کنیم، دیگر نمونه به اندازه کافی بزرگ نیست که بتوان برآورد کارایی به‌دست آورد. برای این منظور یا باید به حجم نمونه افزود یا روش‌هایی را به کار گرفت که دقت لازم را تضمین کنند. افزایش حجم نمونه، به‌ویژه زمانی که تعداد ناحیه‌ها بزرگ باشد، مستلزم افزایش هزینه و زمان است که در عمل ممکن است خارج از ظرفیت و هزینه در نظر گرفته‌شده باشد. بنابراین افزایش نمونه در این شرایط عملا غیر ممکن است. از طرفی سرشماری‌ها نیز نمی‌توانند راه‌هایی کامل برای تولید برآوردهای نواحی کوچک باشند. چون سرشماری هر ده سال یک بار انجام می‌شود و نمی‌توان برآوردگرهای رضایت‌بخشی برای سال‌هایی که سرشماری انجام نمی‌شود، به‌دست آورد. از این رو، روش‌های نواحی کوچک مورد استفاده قرار می‌گیرند. هم‌چنین برآوردهای نواحی کوچک باعث روی‌گردانی از برآوردهای مستقیم کلاسیک و حرکت به سوی استفاده از برآوردهای مدل‌وابسته[۱] نامستقیم شده‌اند (رائو ، ۲۰۰۳). مدل‌وابسته به این معنی است که در به‌دست آوردن برآوردگر از مدل آماری‌ای استفاده شود که برای برآورد کردن پارامترهای نواحی کوچک از اطلاعات دیگر نواحی یا دیگر دوره‌های زمانی یا هر دوی آن‌ها استفاده شود.

در واقع همه روش‌های نواحی کوچک از داده‌های کمکی در دسترس در سطح ناحیه‌های کوچک، از قبیل داده‌های ثبتی یا داده‌های حاصل از سرشماری، برای تشکیل متغیرهای پیش‌گو به قصد استفاده در یک مدل آماری استفاده می‌کنند (شکوهی ، ۱۳۹۱). این وام‌گیری اطلاعات را می‌توان از طریق اضافه کردن اثرهای تصادفی هر ناحیه در مدل در نظر گرفت. بنابراین به‌کارگیری مدل‌های آمیخته خطی[۲] (LMM) (سیرل و همکاران ، ۱۹۹۲) یا مدل‌های آمیخته خطی تعمیم‌یافته[۳] (GLMM) (مک‌کلاک و سیرل ، ۲۰۰۱) در برآورد نواحی کوچک ضرورت می‌یابد. در رهیافت بسامدی، پارامترهای یک مدل آمیخته خطی یا آمیخته خطی تعمیم‌یافته را می‌توان با روش‌های مختلفی از جمله ماکسیمم درستنمایی یا ماکسیمم درستنمایی مقید[۴] به‌دست آورد. هم‌چنین به‌طور مستقیم پیش‌گویی نواحی کوچک تحت مدل‌های خطی آمیخته را می‌توان با استفاده از بهترین پیش‌گوی نااریب خطی[۵] (BLUP) محاسبه کرد. البته محاسبه خطای پیش‌گویی و هم‌چنین فاصله پیش‌گویی متناظر دشوار و مستلزم محاسبه انتگرال‌هایی با بعد بالا است. در مقابل، روش‌های برآوردیابی در رهیافت بیزی با وجود پیشرفت‌های محاسباتی از سختی محاسباتی کم‌تری برخوردار هستند. اما دقت برآوردیابی به شدت به توزیع‌های پیشین انتخابی وابسته است. انتخاب پیشین‌های ناآگاهی‌بخش یا مبهم انتخابی معمول می‌باشند. اما به دلیل عدم وجود تعریف مشخص و یکتا برای پیشین‌های ناآگاهی‌بخش، پیشنهادهای مختلفی برای انتخاب این‌گونه پیشین‌ها وجود دارند. افزون بر این، انتخاب پیشین به شدت بر روی مقادیر پیش‌گویی تاثیر دارد. بنابراین کاربست روش‌های بیزی مستلزم دقت فراوان است.

روش همسانه‌سازی داده‌ها[۶] اولین بار توسط لهله و همکاران (۲۰۰۷) مطرح شد. در این روش برآورد ماکسیمم درستنمایی را می‌توان با استفاده از الگوریتم‌های MCMC محاسبه کرد. از ویژگی‌های این روش، حساس نبودن نتایج استنباط نسبت به انتخاب توزیع پیشین و عدم نیاز به محاسبه انتگرال‌های با بعد بالا است. از آن‌جایی که الگوریتم‌های MCMC را به‌سادگی می‌توان با نرم‌افزارهای استاندارد آماری اجرا کرد، روش همسانه‌سازی داده‌ها به‌راحتی قابل انجام است. باغیشنی و محمدزاده (۲۰۱۱) روش

[۱] Model-dependent
[۲] Linear Mixed Model
[۳] Generalized Linear Mixed Model
[۴] Restricted Maximum Likelihood
[۵] Best Linear Unbiased Predictor
[۶] Data cloning

DC را برای مدل‌های آمیخته خطی تعمیم‌یافته فضایی[۷] (SGLMM) به کار بستند. اما چون الگوریتم‌های MCMC از سرعت محاسباتی پایین و قدرت همگرایی کمی برخوردار هستند، با افزایش حجم نمونه توسط روش همسانه‌سازی داده‌ها این مشکلات در این روش جدی‌تر نیز می‌شوند. باغیشنی و محمدزاده (۲۰۱۲) با ترکیب دو روش DC و تقریب لاپلاس آشیانه‌ای تجمیع‌شده[۸] (INLA) که توسط رو و همکاران (۲۰۰۹) معرفی شد، الگوریتم جدیدی با نام HDC[۹] معرفی کردند که مشکلات همگرایی و سرعت پایین الگوریتم‌های MCMC و DC را مرتفع می‌کند. روش INLA تقریبی دقیق از نمونه‌گیری MCMC است و در عین حال از مشکلات همگرایی و سرعت پایین محاسبات برحذر است.

در برآورد نواحی کوچک گاهی هدف بررسی متغیرهایی است که مربوط به تعداد رخداد یک پیشامد است. در این‌گونه موارد با مشاهدات شمارشی مواجه هستیم. در این مقاله، هدف برآورد نواحی کوچک با متغیر پاسخ گامای شمارشی[۱۰] (GC) با در نظر گرفتن اثرهای تصادفی فضایی مشبکه‌ای است. از آن‌جا که محاسبه برآوردهای ماکسیمم درستنمایی با مشکلاتی از قبیل حل انتگرال‌هایی با بعد بالا و ماکسیمم‌سازی توابع پیچیده مواجه است، رهیافت انتخابی ما رهیافت ترکیبی همسانه‌سازی داده‌ها، HDC، است.

۲ برآورد نواحی کوچک تحت توزیع گامای شمارشی فضایی

توزیع GC یک توزیع گسسته است که از پذیرفتن توزیع گاما برای فواصل زمانی بین وقوع پیشامدها در یک فرآیند شمارشی، برای تعداد رخدادها نتیجه می‌شود. فرض کنید $\{\tau_k\}$، $k \in \mathbb{N}$، دنباله زمانی مورد انتظار بین پیشامدهای $(k-1)$ام و kام باشد، به‌طوری که τ_kها متغیرهایی مستقل و هم‌توزیع با توزیع $Gamma(\alpha,\gamma)$، با میانگین $E(\tau)=\alpha/\gamma$ و واریانس $\mathrm{Var}(\tau)=\alpha/\gamma^2$، هستند. می‌توان نشان داد اگر Y_T نشان‌دهنده تعداد پیشامدها در فاصله زمانی $(0,T)$ باشد، آن‌گاه متغیر تصادفی Y_T دارای توزیع گامای شمارشی با پارامترهای α و γ، $Y_T \sim \mathrm{GC}(\alpha,\gamma)$، است (وینکلمن ، ۱۹۹۵؛ زیوانی و همکاران ، ۲۰۱۴). تابع احتمال Y_T به صورت

$$P(Y_T=y)=G(y\alpha,\gamma T)-G((y+1)\alpha,\gamma T),\quad y=0,1,2,\ldots \tag{۱.۲}$$

به‌دست می‌آید، که در آن $G(n\alpha,\gamma T)=\frac{1}{\Gamma(n\alpha)}\int_0^{\gamma T}x^{n\alpha-1}e^{-x}dx$. میانگین این توزیع نیز به صورت

$$E(Y_T)=\sum_{i=1}^{\infty}G(\alpha i,\gamma T) \tag{۲.۲}$$

نتیجه می‌شود (وینکلمن ، ۱۹۹۵).

برای معرفی مدل گامای شمارشی، در برآورد نواحی کوچک، فرض کنید Y_{ij}، برای $i=1,\ldots,m$ و $j=1,\ldots,n_i$، متغیر پاسخ برای jامین واحد از iامین ناحیه است و از توزیع گامای شمارشی پیروی می‌کند. اثرات تصادفی را از طریق پیشگوی خطی به مدل وارد می‌کنیم. در این‌جا فرض می‌کنیم اثرات تصادفی دارای وابستگی فضایی از نوع مشبکه‌ای هستند و از یک مدل خودبرگشت

[۷] Spatial Generalized Linear Mixed Models
[۸] Integrated Nested Laplace Approximation
[۹] Hybrid Data Cloning
[۱۰] Gamma Count

شرطی[۱۱] (CAR) پیروی می‌کنند. بنابراین مدل گامای شمارشی با در نظر گرفتن اثر تصادفی فضایی، به‌صورت

$$\begin{aligned} Y_{ij} &\sim GC(\alpha, \alpha\gamma_{ij}) \\ \log(\gamma_{ij}) &= \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + u_i \end{aligned} \qquad (۳.۲)$$

تعریف می‌شود، که در آن

$$u_i|\boldsymbol{u}_{-i} \sim N(\rho \sum_{\ell:\ell\neq i} \frac{w_{i\ell}}{w_{i+}} u_\ell, \frac{\sigma_u^۲}{w_{i+}}), i = ۱, \ldots, n,\ \ell = ۱, \ldots, m$$

به‌طوری که $\boldsymbol{u}_{-i} = \{u_\ell : \ell \neq i\}$، $w_{i+} = \sum_{i\neq\ell} w_{i\ell}$ و

$$w_{ij} = \begin{cases} ۱ & \text{اگر نواحی } i \text{ و } j \text{ همسایه باشند} \\ ۰ & i = j \text{ اگر} \\ ۰ & \text{در غیر این‌صورت} \end{cases}$$

تابع درستنمایی متناظر با (۳.۲) به‌صورت

$$\begin{aligned} L(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{X}) &= \int_{\Re^m} \prod_{i=۱}^{m} f(\boldsymbol{y}_i, u_i|\phi) d\boldsymbol{u} \\ &= \int_{\Re^m} \prod_{i=۱}^{m} GC\big(\boldsymbol{y}_i\alpha, \alpha\exp(\boldsymbol{x}_i^{\top}\boldsymbol{\beta} + u_i)\big) N_m(\boldsymbol{u}; ۰, \sigma_u^۲\boldsymbol{A}^{-۱}) d\boldsymbol{u} \end{aligned} \qquad (۴.۲)$$

حاصل می‌شود، که در آن $\boldsymbol{\phi} = \{\alpha, \boldsymbol{\beta}, \sigma_u^۲\}$ بردار پارامترها، $A = (D - W)$ و D ماتریس قطری از w_{i+}، $i = ۱, \ldots, m$ و W ماتریس متشکل از w_{ij} هستند. به دلیل وجود انتگرال چندبعدی در (۴.۲) و بعد معمولا بالای میدان تصادفی، یعنی m، محاسبه تابع درستنمایی (۴.۲) یک چالش جدی بر سر راه استنباط مبتنی بر درستنمایی در این مدل است. رهیافت همسانه‌سازی داده‌ها با استفاده از روش HDC، که در بخش بعدی تشریح می‌شود، جانشینی کارا برای انجام استنباط مبتنی بر درستنمایی است.

۳ برآورد نواحی کوچک با استفاده از روش HDC

رویکرد ترکیبی همسانه‌سازی داده‌ها (باغیشنی و محمدزاده ، ۲۰۱۲) بر اساس ایده بیزی و استفاده از روش INLA عمل می‌کند. این روش قابل استفاده در اغلب شرایطی است که مساله را بتوان به‌صورت یک مدل بیزی فرمول‌بندی کرد. مشابه دیدگاه بیزی، روش ترکیبی همسانه‌سازی داده‌ها با اجتناب از انتگرال‌گیری‌های عددی با بعد بزرگ، نیازی به ماکسیمم کردن و مشتق‌گیری از توابع ندارد. این روش تنها بر اساس محاسبه میانگین و واریانس استوار است. نکته قابل توجه آن است که اگر چه این روش از فرمول‌بندی بیزی و الگوریتم‌های محاسباتی تقریبی بیزی INLA استفاده می‌کند، اما برآورد ماکسیمم درستنمایی را نتیجه می‌دهد و بنابراین استنباط در این روش بر مبنای دیدگاه درستنمایی است. همچنین استنباط در این روش، به انتخاب توزیع پیشین بستگی ندارد. در ادامه تشریحی خلاصه از روش ترکیبی همسانه‌سازی داده‌ها را بیان می‌کنیم.

برای نوشتن مدل بیزی به‌جای درستنمایی داده‌های مشاهده‌شده، درستنمایی k نسخه مستقل از مشاهدات در نظر گرفته می‌شود که در آن k به‌اندازه کافی بزرگ انتخاب می‌شود. سپس با در نظر گرفتن توزیع‌های پیشین برای پارامترهای نامعلوم، تقریبی از تابع

[۱۱] Conditional Auto Regressive

توزیع پسین متناظر با درستنمایی k نسخه‌ای از داده‌ها و توزیع‌های پیشین تعیین‌شده، با استفاده از روش بیزی تقریبی INLA به‌دست می‌آید. اگر k به‌اندازه کافی بزرگ باشد، تحت شرایط نظم، توزیع پسین متناظر با k نسخه از داده‌ها دارای توزیع مجانبی نرمال با بردار میانگین برابر برآوردگر ماکسیمم درستنمایی و ماتریس واریانس برابر ماتریس اطلاع فیشر تقسیم بر k است (لهله و همکاران ، ۲۰۰۷). بنابراین در این حالت میانگین توزیع پسین به‌دست آمده، برابر با برآورد ماکسیمم درستنمایی و k برابر واریانس پسین به‌دست آمده، با واریانس مجانبی برآورد ماکسیمم درستنمایی برابر است (باغیشنی و محمدزاده ، ۲۰۱۱).

با توجه به هدف مساله یعنی برآورد پارامترهای مدل، $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}, \sigma_u^2)$، و پیش‌گویی اثرهای تصادفی، $\boldsymbol{u} = (u_1, \ldots, u_m)$، الگوریتم HDC برای مدل مذکور به شرح زیر است:

۱. با در نظر گرفتن یک مقدار مناسب برای k، مشاهدات و متغیرهای تبیینی را k بار عینا تکرار کنید. مشاهدات کپی‌شده را با $\boldsymbol{y}^{(k)} = (\boldsymbol{y}, \ldots, \boldsymbol{y})$ و $\boldsymbol{X}^{(k)} = (\boldsymbol{X}, \ldots, \boldsymbol{X})$ نشان می‌دهیم.

۲. تابع درستنمایی بر اساس مجموعه داده جدید را به‌دست آورید:

$$\begin{aligned} L(\boldsymbol{\phi}|\boldsymbol{y}^{(k)}, \boldsymbol{X}^{(k)}) &= \int_{\Re^m} \prod_{i=1}^{m} GC\big(y_i^{(k)}\alpha, \alpha \exp(\boldsymbol{x}_i^{(k)\top}\boldsymbol{\beta} + u_i + \varepsilon_{ij})\big) N_m(\boldsymbol{u}; \cdot, \sigma_u^2 \boldsymbol{A}^{-1}) d\boldsymbol{u} \\ &= \left(L(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{X})\right)^k \end{aligned} \tag{۱.۳}$$

۳. توزیع پیشین برای پارامترهای نامعلوم را به‌صورت $\pi(\boldsymbol{\phi})$ در نظر بگیرید.

۴. توزیع پسین متناظر با (۱.۳) و $\pi(\boldsymbol{\phi})$ را به‌صورت

$$\pi_k(\boldsymbol{\phi}|\boldsymbol{y}^{(k)}, \boldsymbol{X}^{(k)}) = \frac{\left(L(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{X})\right)^k \pi(\boldsymbol{\phi})}{C\left(\boldsymbol{y}^{(k)}\right)} \tag{۲.۳}$$

محاسبه کنید، که در آن $C\left(\boldsymbol{y}^{(k)}\right)$ ثابت نرمال‌ساز متناظر با توزیع پسین همسانه‌سازی‌شده است.

توزیع پسین حاصل از همسانه‌سازی داده‌ها در (۲.۳) فاقد صورت بسته است. یکی از راهکارهای معمول، استفاده از الگوریتم‌های MCMC است. اما انجام استنباط مبتنی بر MCMC دشوار و از کارایی ضعیفی برخوردار است. به‌ویژه مشکلات سرعت و تشخیص همگرایی برای داده‌های همسانه‌سازی‌شده شدت بیشتری دارد. اما روش HDC که توسط باغیشنی و محمدزاده (۲۰۱۲) معرفی شد، از آن جایی که از روش INLA برای تقریب توزیع پسین کمک می‌گیرد، از مشکلات تشخیص همگرایی و زمان طولانی محاسبات الگوریتم‌های MCMC دور است. از این رو، عملکرد محاسباتی روش کارای HDC نسبت به DC که با MCMC کار می‌کند، بسیار سریع‌تر است. همچنین دقت برآوردگرهای حاصل از HDC معادل برآوردگرهای DC است. روش HDC با استفاده از بسته R-INLA در نرم‌افزار R قابل اجرا است. برای اطلاعات بیشتر می‌توانید به پایگاه اطلاعاتی این روش، http://www.r-inla.org، مراجعه کنید.

۴ مطالعه شبیه‌سازی

در این بخش، برآورد ML نواحی کوچک مدل GC فضایی در چارچوب یک مطالعه شبیه‌سازی با استفاده از رهیافت HDC، مورد ارزیابی قرار گرفته و با مدل‌های رگرسیون پواسون و دوجمله‌ای منفی مقایسه می‌شود. برای این منظور، یک مجموعه داده به تعداد

نواحی $m = ۳۱$ و حجم هر ناحیه $n_i = ۳$ از مدل

$$Y_{ij}|\boldsymbol{x}_{ij}; \alpha, \beta_۰, \beta_۱, \beta_۲, u_i \sim \text{GC}\left(\alpha, \alpha \exp(\beta_۰ + \beta_۱ x_{۱ij} + \beta_۲ x_{۲ij} + u_i)\right), \qquad (۱.۴)$$
$$j = ۱, \ldots, n_i, \; i = ۱, \ldots, m$$

شبیه‌سازی شد. مقادیر متغیر تبیینی $\boldsymbol{x}_{ij}$ از توزیع یکنواخت $(۰, ۱)$ تولید شدند. مقادیر واقعی ضرایب رگرسیونی نیز برابر $(\beta_۰, \beta_۱, \beta_۲) = (۱, -۱, ۰/۷۵)$ در نظر گرفته شدند. میدان تصادفی را نیز یک توزیع چندمتغیره نرمال به صورت

$$N_m(\boldsymbol{u}; ۰, \sigma_u^۲ \boldsymbol{A}^{-۱})$$

با $\sigma_u^۲ = ۱$ در نظر گرفتیم. از آن‌جایی که پارامتر α پراکندگی مدل رگرسیون GC در شرایط مختلف را کنترل می‌کند، سه مقدار مختلف $(۰/۲, ۱, ۵)$ برای آن در نظر گرفته شد. توجه داشته باشید که به ازای $\alpha = ۱$ مدل GC همان مدل پواسون خواهد بود. هم‌چنین به‌صورت تجربی، مقدار $k = ۲۰$ را در نظر گرفتیم. پارامترهای مدل GC و مدل‌های رقیب با روش HDC برآورد شدند. به‌منظور ارزیابی مدل‌های مورد بحث، معیار ریشه میانگین توان دوم خطا[۱۲] (RMSE) برآوردگرهای مختلف محاسبه شدند. به‌منظور لحاظ نمودن عدم قطعیت حاکم بر شبیه‌سازی نمونه‌های تصادفی، تمامی محاسبات ۵۰۰ بار تکرار شدند. برآورد پارامترها و مقادیر RMSE در جدول ۱ گزارش شده‌اند. همان‌طور که مشاهده می‌شود، مدل رگرسیون GC در حالتی که مشاهدات کم‌پراکنده هستند، برازش بهتری نسبت به مدل‌های پواسون و دوجمله‌ای منفی دارد؛ وقتی $\alpha = ۱$ است مدل رگرسیون GC تقریبا عملکرد یکسانی با مدل رگرسیون پواسون دارد که تاییدی بر یکسان بودن دو مدل رگرسیون GC و پواسون برای $\alpha = ۱$ است. اما در حالتی که مشاهدات دچار بیش‌پراکنش هستند، $\alpha = ۰/۲$، مدل رگرسیون GC در مقایسه با مدل‌های رقیب تقریبا یکسان عمل کرده و در برخی موارد مغلوب شده است؛ ریشه این مساله را شاید تداخل ناهمگنی حاصل از بیش‌پراکندگی مشاهدات و وابستگی فضایی موجود در داده‌ها بتوان تفسیر کرد.

۵ مطالعه کاربردی: داده‌های کم‌پراکنده سرطان حنجره

مجموعه داده مورد مطالعه مربوط به تعداد مرگ و میر بیماران مبتلا به سرطان حنجره در ۵۴۴ ناحیه در آلمان طی سال‌های ۱۹۸۶ تا ۱۹۹۰ است. این مجموعه داده در سایت r-inla قابل دسترس است. در این مجموعه داده تنها یک متغیر تبیینی مربوط به میزان مصرف دخانیات حاضر است. فرض می‌کنیم پاسخ‌ها به صورت شرطی مستقل هستند و با توجه به ماهیت شمارشی آن‌ها، برای استفاده در این مطالعه مناسب هستند. پیشگوی خطی را با η_i، $i = ۱, \ldots, n$، نشان داده و به صورت

$$\eta_i = \beta_۰ + \beta_۱ x + f_§(S_i)$$

در نظر می‌گیریم که در آن $f_§(\cdot)$ اثر فضایی است و فرض می‌کنیم از مدل CAR پیروی می‌کند و x متغیر تبیینی میزان مصرف دخانیات است. به مشاهدات سه مدل رگرسیونی GC، پواسون و NB را برازش دادیم که برآوردها و فواصل HPD در جدول ۲ گزارش شده‌اند. با توجه به نتایج فاصله‌های HPD، مدل GC بر خلاف دو مدل پواسون و دوجمله‌ای منفی اثر میزان مصرف دخانیات را، در حضور وابستگی فضایی نواحی، بر نرخ مرگ و میر معنی‌دار نمی‌داند. البته با توجه به اندازه اثر برآوردشده که در هر سه مدل تقریبا یکسان به دست آمده‌اند، می‌توان گفت هر سه مدل به شکلی اثر این متغیر را ناچیز می‌دانند.

[۱۲] Root Mean Squared Error

جدول ۱: مقادیر ریشه میانگین توان دوم خطا و برآوردهای (داخل پرانتز) پارامترهای مدل

α	مدل	پارامتر				
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\sigma}_u^2$
۰/۲	GC	۰/۵۷۲۱	۰/۵۷۶۹	۰/۵۷۷۸	۰/۰۹۱۲	۰/۷۳۲۹
		(۱/۰۰۷۰)	(−۱/۱۶۶۶)	(۰/۹۰۴۸)	(۰/۲۶۳۶)	(۰/۳۷۲۱)
	Poisson	۰/۵۲۳۵	۰/۴۷۰۲	۰/۴۲۵۳	—	۰/۷۶۲۷
		(۱/۴۵۳۵)	(−۰/۸۷۳۴)	(۰/۶۳۰۱)	—	(۰/۲۷۹۶)
	NB	۰/۵۸۱۳	۰/۴۷۲۶	۰/۴۱۲۵	—	۰/۷۹۶۷
		(۱/۴۸۳۹)	(−۰/۸۵۰۲)	(۰/۶۹۲۳)		(۰/۲۶۵۹)
۱	GC	۰/۲۹۳۳	۰/۲۵۵۶	۰/۲۶۵۸	۰/۲۹۲۷	۰/۸۴۱۴
		(۰/۹۷۴۱)	(−۰/۹۷۷۹)	(۰/۷۴۸۶)	(۱/۱۷۲۱)	(۰/۱۶۷۱)
	Poisson	۰/۲۸۵۸	۰/۲۵۶۹	۰/۲۶۳۷	—	۰/۸۴۰۴
		(۰/۹۷۱۹)	(−۰/۹۸۳۹)	(۰/۷۵۰۴)	—	(۰/۱۶۸۶)
	NB	۶۴/۴۵۶۷	۱۰۴/۱۹۰۱	۱۱۳/۲۰۲۳	—	۰/۸۴۴۶
		(۴/۵۶۰۳)	(−۱/۶۷۹۷)	(−۳/۱۱۹۸)	—	(۰/۱۶۳۸)
۵	GC	۰/۲۷۵۹	۰/۱۳۶۶	۰/۱۳۷۴	۱/۹۸۳۷	۰/۹۲۱۶
		(۰/۹۷۹۴)	(۰/۹۶۲۹)	(۰/۷۳۹۶)	(۵/۵۸۶۸)	(۰/۰۶۲۵)
	Poisson	۰/۳۱۹۴	۰/۲۵۷۳	۰/۲۴۵۲	—	۰/۹۴۶۶
		(۰/۸۲۶۸)	(−۱/۱۱۵۶)	(۰/۸۴۵۳)	—	(۰/۰۴۵۵)
	NB	۱۱۲/۸۵۷۱	۱۶۷/۰۵۱۹	۱۸۴/۳۵۷۱	—	۰/۹۴۸۸
		(۲۷/۸۶۳۷)	(۱۱/۱۵۳۱)	(۳۹/۹۷۴۹)	—	(۰/۰۴۳۸)

جدول ۲: مقادیر برآوردشده پارامترها و فاصله HPD آن‌ها

مدل	α	β_0	β_1
GC	۱/۲۵۸	۱/۸۰۵۱	۰/۰۰۶۴
	(۱/۲۷۴۰, ۱/۲۴۲۳)	(۲/۲۸۸۴, ۱/۳۲۰۲)	(۰/۰۱۶۰, −۰/۰۰۳۳)
پواسون	—	۱/۷۱۶۶	۰/۰۰۸۲
	—	(۲/۰۱۸۹, ۱/۴۱۳۳)	(۰/۰۱۴۳, ۰/۰۰۲۲)
NB	—	۱/۶۶۴۶	۰/۰۰۸۹
	—	(۲/۲۳۶۱, ۰/۳۶۴۲)	(۰/۰۱۶۴ × ۱۰$^{-۵}$, ۶/۴۹۸)

برای سنجش قدرت پیشگویی مدل‌ها، اطلاعات ۱۰ ناحیه را به صورت تصادفی کنار گذاشته و بر اساس مدل برازش‌شده با سایر مشاهدات نواحی، آن‌ها را پیشگویی و معیار میانگین توان دوم خطای پیش‌گویی[۱۳] (MSPE) را برای مقادیر پیش‌گویی‌شده محاسبه کردیم؛ نتیجه در جدول ۳ گزارش شده است. با توجه به مقادیر جدول ۳، مدل رگرسیون GC از قدرت پیشگویی بیشتری نسبت به مدل‌های پواسون و NB برخوردار است. هم‌چنین نمودارهای پراکنش مقادیر برازش‌شده و پیش‌گویی‌شده در مقابل مقادیر واقعی به ترتیب در شکل‌های ۱ و ۲ نمایش داده شده‌اند. مقایسه عملکرد پیش‌گویی سه مدل، به‌طور جزیی‌تر، از روی این شکل‌ها قابل انجام است.

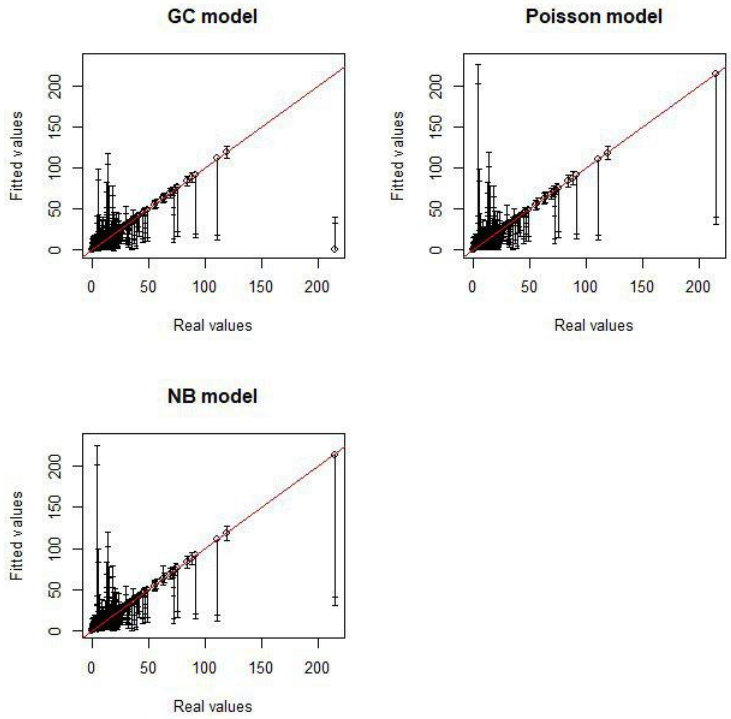
جدول ۳: مقدار MSPE برای مشاهدات پیشگویی‌شده

	مدل		
MSPE	GC	پواسون	NB
میانگین	۱۷/۶۴۸۲	۲۴/۳۹۰۸	۲۴/۲۳۲۲
میانه	۵/۵۴۱۹	۶/۶۳۰۶	۶/۵۹۴۴

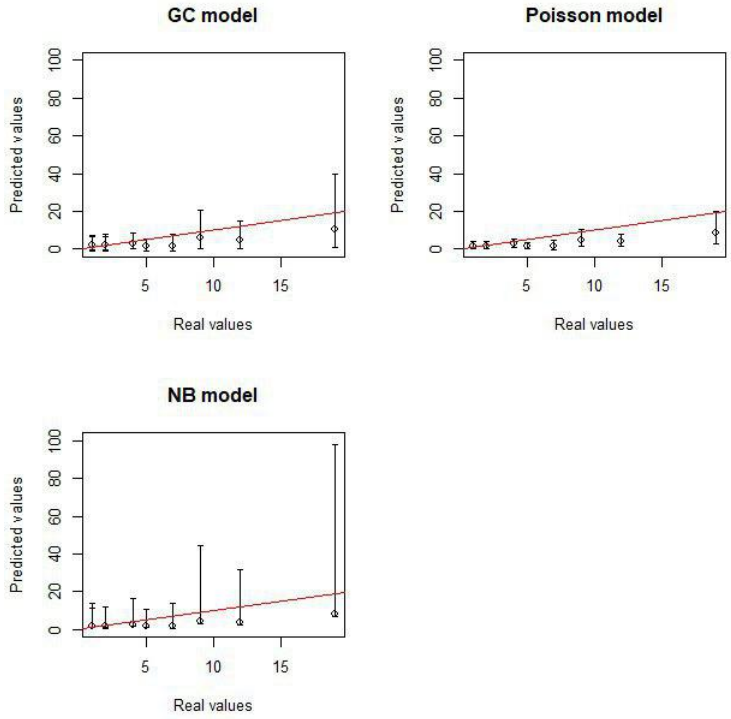
بحث و نتیجه‌گیری

مدل رگرسیون گامای شمارشی فضایی در برآورد نواحی کوچک داده‌های شمارشی‌ای که دارای وابستگی فضایی از نوع مشبکه‌ای هستند، یک مدل جدید جانشین برای مواجهه با مشکل کم‌پراکنش یا بیش‌پراکنش است. از آن‌جا که تحلیل مبتنی بر درستنمایی داده‌های شمارشی فضایی مشبکه‌ای، پیچیده و دشوار است، در این مقاله با استفاده از رهیافت HDC به استنباط ML با کارایی محاسباتی بالا پرداختیم. با توجه به پیچیدگی مدل GC، برازش مدل رگرسیون GC فضایی با استفاده از روش DC، که از الگوریتم‌های MCMC استفاده می‌کند، با مشکلاتی همچون سرعت محاسبات و همگرایی الگوریتم گریبان‌گیر است. بنابراین، روش HDC که ترکیبی از دو روش DC و INLA است، جایگزین مناسبی برای استنباط مبتنی بر درستنمایی مدل پیشنهادی است. نتایج حاصل از مطالعه شبیه‌سازی و مثال کاربردی، نیز حکایت از برتری مدل GC نسبت به مدل‌های رقیب پواسون و دوجمله‌ای منفی دارند.

[۱۳] Mean Squared Prediction Error



شکل ۱: نمودار پراکنش مقادیر برازش شده در مقابل مقادیر واقعی.



شکل ۲: نمودار پراکنش مقادیر پیش‌گویی شده در مقابل مقادیر واقعی.

مراجع

شکوهی ف. (۱۳۹۱). *استنباط درستنمایی در برآورد کوچک‌ناحیه‌ای با ترکیب داده‌های مقطعی و سری‌های زمانی و استفاده از روش‌های ناپارامتری*، رساله دکتری، گروه آمار، دانشگاه شهید بهشتی.

Baghishani, H., and Mohammadzadeh, M. (2011). data cloning algorithm for computing maximum likelihood estimates in spatial generalized linear mixed models, *Computational Statistics and Data Analysis*, **55**, 1748-1759.

Baghishani, H., Rue, H., and Mohammadzadeh M. (2012). On a hybrid data dloning method and its application in generalized linear mixed models, *Statistics and Computing*, **22**, 597-616.

Lele, S.R., Dennis, B., and Lutscher, F., (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, **10**, 551–563.

McCulloch, C. E., and Searle, S. R. (2001). *Generalized Linear and Mixed Models,* John Wiley and Sons, New York.

Rao, J. N. K., (2003). *Small Area Estimation,* John Wiley and Sons, New York.

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**, 319-392.

Searle, S. R., Casella, G., McCulloch, C. E., (1992). *Variance Components,* John Wiley and Sons, New York.

Winkelmann, R. (1995), Duration Dependence and Dispersion in Count-Data Models, *Journal of Business and Economic Statistics,* **13**, 467 - 474.

Zeviani, W. M., Ribeiro, Jr, P. J., Bonat, W. H., Shimakura, S. E., and Muniz, J. A. (2014). The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics* **41**, 2616-2626.