مجله علوم آماری، پاییز و زمستان ۱۳۹۹
جلد ۱۴، شماره ۲، ص ۳۰۷ – ۳۳۴
DOI: 10.29252/jss.14.2.307

تحلیل بیزی داده‌های شمارشی فضایی در جوامع متناهی با رهیافت معادلات دیفرانسیل جزئی تصادفی

نگار اقبال، حسین باغیشنی

گروه آمار، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود

تاریخ دریافت: ۱۳۹۸/۱۱/۱۰ تاریخ آخرین بازنگری: ۱۳۹۹/۰۳/۱۰

چکیده:

داده‌های شمارشی زمین‌آماری در جوامع متناهی در کاربردهای مختلفی، مثل مدیریت شهری و پزشکی، دیده می‌شوند. مدل معمول برای تحلیل این نوع پاسخ‌ها، مدل لوجیت–دوجمله‌ای فضایی است. در اکثر موقعیت‌های کاربردی، این نوع داده‌ها جدا از تغییرپذیری فضایی دارای بیش‌پراکندگی هستند که مدل دوجمله‌ای توانایی مدل‌بندی آن را ندارد. رهیافت جانشین در این حالت، یک مدل بتا–دوجمله‌ای است که از انعطاف لازم برای لحاظ کردن بیش‌پراکنشی موجود در داده‌ها برخوردار است. در این مقاله، ابتدا برازش مدل بتا–دوجمله‌ای فضایی برای داده‌های شمارشی زمین‌آماری با یک رهیافت بیزی ترکیبی مبتنی بر تقریب لاپلاس آشیانی جمع‌بسته و معادلات دیفرانسیل جزیی تصادفی توصیف می‌شود. سپس این مدل، در یک مطالعه موردی، برای تحلیل تعداد تصادف‌های منجر به جرح یا فوت در شهر مشهد به‌کار گرفته می‌شود. همچنین با یک مطالعه شبیه‌سازی، عملکرد مدل پیشنهادی ارزیابی می‌شود.

واژه‌های کلیدی: بتا–دوجمله‌ای فضایی، بیش‌پراکنش، رهیافت بیزی تقریبی، معادلات دیفرانسیل جزیی تصادفی، تصادفات رانندگی.

آدرس الکترونیکی نویسنده مسئول مقاله: حسین باغیشنی، hbaghishani@shahroodut.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11, 62M30, 91B72.

۱ مقدمه

تعداد رخدادهای یک پدیده (تعداد موفقیت‌ها) در یک جامعه متناهی، با همراه بودن اطلاعات مکانی رخداد آن‌ها، کمیت مورد علاقه در بسیاری از مطالعات علمی محسوب می‌شود. به‌عنوان نمونه، مطالعه روی عوامل موثر بر شیوع یک بیماری در نواحی مختلف یک کشور را می‌توان نام برد که در آن تعداد افراد مبتلا به آن بیماری از کل جمعیت ناحیه، متغیر پاسخ هدف است (بیلبی و همکاران ، ۲۰۱۴). داده‌هایی که در این مقاله مورد توجه هستند نیز تعداد تصادف‌های منجر به جرح یا فوت در نقاط جغرافیایی مختلف شهر مشهد (شامل خیابان‌ها، تقاطع یا میدان‌ها) هستند که نسبتی از کل تصادفات در آن نقاط محسوب می‌شوند. تحلیل‌گران داده، به‌طور معمول، برای مدل‌بندی این نوع از پاسخ‌های فضایی از مدل لوجیت–دوجمله‌ای استفاده می‌کنند (کریستنسن و همکاران ، ۲۰۰۶؛ دیگل و ریبرو ، ۲۰۰۶؛ استانتون و دیگل ، ۲۰۱۳). یکی از پذیره‌های ذاتی توزیع دوجمله‌ای، کوچکتر بودن واریانس توزیع از میانگین آن است. اما در عمل، در بسیاری از موقعیت‌های کاربردی واریانس پاسخ از واریانس نظری توزیع دوجمله‌ای و حتی از میانگین پاسخ بزرگتر است که از آن به بیش‌پراکنشی[1] یاد می‌شود. اگرچه با افزودن یک اثر تصادفی فضایی به مدل می‌توان بخشی از ناهمپراشی موجود در داده‌ها به واسطه وابستگی فضایی آن‌ها را مدل‌بندی کرد، ولی در بعضی از موقعیت‌ها باز هم ناهمپراشی باقی‌مانده در داده‌ها می‌تواند قابل توجه باشد و لزوم اعمال سطح دومی از مدل‌بندی برای آن احساس شود. نادیده گرفتن بیش‌پراکنشی موجود در داده‌ها (بعد از لحاظ کردن اثر ناهمپراشی فضایی) و به‌کار بردن توزیع دوجمله‌ای برای تحلیل آن‌ها، می‌تواند به استنباط‌های گمراه‌کننده، شامل برآورد اریب پارامترها و خطای معیار آن‌ها، منتهی شود. این مساله خود منجر به تشخیص نادرست عوامل تاثیرگذار بر متغیر پاسخ و پیشگویی‌های فضایی با خطای بالا خواهد شد. رهیافت استاندارد برای حل این مشکل، استفاده از ساختارهای احتمالی مرکب است که در آن‌ها فرض می‌شود احتمال موفقیت در توزیع دوجمله‌ای، یک متغیر تصادفی است (ویلیامز ، ۱۹۸۲؛ ریچاردز ، ۲۰۰۸). در این رهیافت، پذیره توزیع بتا برای احتمال موفقیت یک انتخاب متداول شده است (ناجرا–زولگا و همکاران ، ۲۰۱۹). با در نظر گرفتن این پذیره، توزیع کناری پاسخ بتا–دوجمله‌ای خواهد بود، که با داشتن دو پارامتر انعطاف زیادی نسبت به توزیع دوجمله‌ای در مدل‌بندی واریانس پاسخ دارد و به سادگی می‌تواند بیش‌پراکندگی ممکن در داده‌ها را مدل‌بندی کند.

استفاده از مدل بتا–دوجمله‌ای برای تحلیل داده‌های فضایی چندان مورد توجه محققین نبوده است. کارهای هوگز و مادن (۱۹۹۳)، باندیوپادیای و همکاران (۲۰۱۱)، عابدین‌پور و همکاران (۱۳۹۸)،

[1] Overdispersion

شواب و مارکس (۲۰۱۵)، و کولووس و همکاران (۲۰۱۶) از جمله مطالعات اندکی هستند که می‌توان به آن‌ها اشاره کرد که سه مطالعه اول بر تحلیل داده‌های فضایی مشبکه‌ای تمرکز کرده‌اند و دو مطالعه بعدی به پیشگویی فضایی کلاسیک برای داده‌های زمین‌آماری پرداخته‌اند. افزون بر این، دو مطالعه آخر با ارایه یک برآورد تعدیل‌یافته از تابع تغییرنگار برای توزیع بتا-دوجمله‌ای از معادلات کریگیدن مشابه مدل دوجمله‌ای برای پیشگویی فضایی استفاده کرده‌اند و توجهی به بررسی عوامل تاثیرگذار بر پاسخ در حضور وابستگی فضایی نداشته‌اند. برای داده‌های فضایی زمین‌آماری، با توجه به چالش تجزیه ماتریس‌های چگال، تا جایی که نویسندگان بررسی کردند، در دیدگاه بیزی مطالعه‌ای یافت نشد. بنابراین، یکی از نوآوری‌های این مقاله، استفاده از مدل بتا-دوجمله‌ای بیزی سلسله‌مراتبی برای تحلیل داده‌های فضایی زمین‌آماری است.

با توجه به پیچیدگی تابع درستنمایی شرطی مدل بتا-دوجمله‌ای فضایی و چالش محاسباتی انتگرال‌گیری لازم برای محاسبه تابع درستنمایی کناری مدل، به دلیل بعد بالای انتگرال‌گیری نسبت به اثر فضایی، روشن است که انجام استنباط مبتنی بر درستنمایی در این مدل اگر ناممکن نباشد، بسیار پرهزینه و مشکل است. بنابراین، در این مقاله، از رهیافت بیزی برای انجام استنباط در مدل پیشنهادی استفاده شده است. روش معمول برای اجرای استنباط بیزی در مدل‌های پیچیده فضایی، نمونه‌گیری مبتنی بر شبیه‌سازی مانند الگوریتم‌های مونت کارلوی زنجیر مارکوفی[1] (MCMC) است. اما این الگوریتم‌ها با این‌که بسیار پرکاربرد هستند، در مدل‌های سلسله‌مراتبی پیچیده، با مشکلات جدی از قبیل همگرایی کند و آمیختگی ضعیف زنجیر، مواجه هستند (رو و همکاران ، ۲۰۰۹). برای دوری از این مشکلات و بالا بردن سرعت تقریب توزیع پسین مدل، رو و همکاران (۲۰۰۹) یک روش تقریبی را به نام تقریب لاپلاس آشیانی جمع‌بسته[2] (INLA) معرفی کردند. این روش از نظر محاسباتی بسیار کاراست و دقت نتایج حاصل از آن با دقت نتایج الگوریتم‌های MCMC رقابت می‌کند (قلی‌زاده گزور و همکاران ، ۱۳۹۲).

سرعت بالای روش INLA در برازش مدل‌های فضایی، به دلیل گسسته بودن ناحیه تحت مطالعه است. به عبارتی این روش برای داده‌های مشبکه‌ای کاراست؛ اما وقتی ناحیه چگال باشد، کارایی روش مذکور از دست می‌رود. لیندگرن و همکاران (۲۰۱۱) روش معادلات دیفرانسیل جزیی تصادفی[3] (SPDE) را معرفی کردند که در ترکیب با روش INLA مشکل محاسباتی برازش مدل‌های فضایی چگال را مرتفع می‌کند. در این مقاله، برای برازش مدل بتا-دوجمله‌ای فضایی بر داده‌های تصادف‌های منجر به جرح یا فوت در شهر مشهد، از رهیافت ترکیبی INLA+SPDE استفاده شده است. همچنین، امکان محاسبه

[1] Markov Chain Monte Carlo
[2] Integrated Nested Laplace Approximation
[3] Stochastic Partial Differential Equations

معیارهای مختلف ارزیابی برازش مدل با روش INLA کمک کرده است تا بتوان جنبه‌های مختلف برازش، ارزیابی و انتخاب مدل فضایی پیچیده بتا-دوجمله‌ای را نیز اجرا کرد.

در بخش ۲ مدل بتا-دوجمله‌ای فضایی معرفی و نحوه مدل‌بندی بیش‌پراکنشی توصیف می‌شود. در بخش ۳ استنباط بیزی با این مدل با استفاده از رهیافت ترکیبی INLA+SPDE تشریح می‌شود. سپس، انتخاب توزیع‌های پیشین مناسب و چگونگی ارزیابی مدل نیز بحث خواهند شد. در بخش ۴ با مطالعه شبیه‌سازی عملکرد مدل پیشنهادی ارزیابی و با مدل فضایی و غیرفضایی دوجمله‌ای مقایسه می‌شود. در بخش ۵ داده‌های تعداد تصادف‌های منجر به جرح یا فوت در شهر مشهد تحلیل و نتایج آن تفسیر می‌شوند.

۲ مدل بتا-دوجمله‌ای فضایی

فرض کنید $D \subset \Re^d$، $d \geq ۱$، ناحیه فضایی چگال مورد مطالعه و برای $i = ۱, \ldots, n$، $Y_i = Y(\boldsymbol{s}_i)$ متغیر پاسخ تعداد رخداد پدیده مورد علاقه (موفقیت)، در یک جامعه متناهی، در موقعیت $\boldsymbol{s}_i \in D$ باشد. مدل معمول برای تحلیل این نوع پاسخ بر مبنای توزیع دوجمله‌ای با تابع احتمال

$$\mathrm{P}(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (۱ - \pi_i)^{n_i - y_i}, \qquad y_i = ۰, ۱, \cdots, n_i \tag{۱}$$

ساخته می‌شود، که در آن $n(\boldsymbol{s}_i) = n_i$ ، $\pi(\boldsymbol{s}_i) = \pi_i$ احتمال موفقیت، $۱ - \pi_i$ احتمال شکست و $y_i = y(\boldsymbol{s}_i)$ مشاهده پاسخ در موقعیت $\boldsymbol{s}_i$ است. برای در نظر گرفتن عوامل موثر بر احتمال موفقیت در قالب متغیرهای تبیینی رگرسیونی و سایر اثرات ممکن، به‌ویژه در حضور اثر وابستگی فضایی داده‌ها، پارامتر π_i را معمولا با استفاده از یک تابع پیوند لوجیت در چارچوب یک مدل رگرسیون جمعی ساختاری[۱] (STAR) مدل‌بندی می‌کنند (فارمی‌یر و همکاران ، ۲۰۱۳). در مدل STAR پیشگوی جمعی به صورت

$$\ln\left(\frac{\pi_i}{۱ - \pi_i}\right) = \eta_i = \alpha + \sum_{j=۱}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki} + \epsilon_i, \quad i = ۱, \cdots, n_d \tag{۲}$$

نوشته می‌شود، که در آن $\{f^{(j)}(\cdot)\}$ ها توابع نامعلومی از متغیرهای تبیینی $\boldsymbol{u}$ هستند، ضرایب نامعلوم $\{\beta_k\}$ بیانگر اثرات خطی متغیرهای تبیینی $\boldsymbol{z}$ و ϵ_i مولفه‌های غیرساختارمند، مثل خطای اندازه‌گیری،

[۱] Structured Additive Regression Model

هستند. با توجه به صورت‌های متنوعی که توابع $\{f^{(j)}(\cdot)\}$ می‌توانند اختیار کنند، رده مدل‌های STAR کاربردهای وسیعی دارند. با فرض این‌که $f(\cdot)$ یکی از $f^{(j)}$ ها در (۲) است، برخی از کاربردهای مذکور عبارتند از

الف- اثرات غیرخطی (ناپارامتری) متغیرهای تبیینی: برای در نظر گرفتن اثر هموار (ناپارامتری) یک متغیر تبیینی مثل u، به‌سادگی می‌توان $f(u)$ را با اسپلاین‌های جریمه‌ای (لانگ و برزگر ، ۲۰۰۴) یا فرآیندهای قدم زدن تصادفی (رو و هلد ، ۲۰۰۵) مدل‌بندی کرد.

ب- اثرات تصادفی: در تحلیل داده‌های طولی یا خوشه‌ای، با تعریف $f(u_i) = f_i$ و قرار دادن $\{f_i\}$ به‌عنوان متغیرهای تصادفی (پنهان) گاوسی با میانگین صفر، می‌توان ساختار همبستگی بین داده‌ها را وارد مدل کرد.

ج- اثرات فضایی: اگر $u(\boldsymbol{s})$ یک متغیر تصادفی فضایی باشد، با تعریف $f(u(\boldsymbol{s})) = f_s$ می‌توان وابستگی فضایی را وارد مدل کرد. مدل تصادفی فضایی برای f_s بسته به این‌که ناحیه فضایی تحت مطالعه چگال باشد یا مشبکه‌ای، متفاوت است. برای تحلیل داده‌های ناحیه‌ای (مشبکه‌ای)، مثل پهنه‌بندی بیماری‌ها[1]، می‌توان از مدل اتورگرسیو شرطی[2] (بی‌سگ و همکاران ، ۱۹۹۱) استفاده کرد. برای تحلیل داده‌های زمین‌آماری، می‌توان از یک میدان تصادفی گاوسی[3] (GRF) بهره گرفت (دیگل و ریبیرو ، ۲۰۰۶).

در بسیاری از کاربردها، ممکن است مدل مناسب برای تحلیل داده‌ها شامل ترکیبی از مولفه‌های مختلف، مثل مولفه فضایی، اثرات تصادفی و هر دو اثر خطی و ناپارامتری برخی از متغیرهای تبیینی، باشد. یک زیررده بسیار پرطرفدار از مدل (۲)، رده مدل‌های گاوسی پنهان[4] (LGM) است که در آن فرض می‌شود α، $\{f^{(j)}(\cdot)\}$، $\{\beta_k\}$ و $\{\epsilon_i\}$ همگی متغیرهای گاوسی هستند (رو و همکاران ، ۲۰۰۹). رهیافت تقریبی INLA بر روی رده LGMs بنا شده است. مدل (۱) نیز با پذیرفتن پیشین گاوسی برای مولفه‌های ذکرشده خود در پیشگوی جمعی η_i، عضوی از رده LGMs محسوب می‌شود، که در آن امید ریاضی و واریانس پاسخ Y_i به ترتیب برابر با $n_i\pi_i$ و $n_i\pi_i(1-\pi_i)$ است. یک ویژگی مهم این مدل کوچک‌تر بودن واریانس از میانگین است؛ اما در موارد متعددی این پذیره برقرار نیست و با مساله بیش‌پراکنشی مواجه می‌شویم. در چنین مواردی، عدم انعطاف توزیع دوجمله‌ای استفاده از آن را محدود می‌کند و این مدل نمی‌تواند ماهیت داده‌ها را به خوبی برازش دهد. در نتیجه استفاده از آن برای مدل‌بندی داده‌ها منجر به استنباط‌های آماری نادرست در برآورد پارامترها، برآورد ساختار وابستگی و پیش‌گویی فضایی داده‌ها

[1] Disease mapping
[2] Conditional autoregressive
[3] Gaussian Random Field
[4] Latent Gaussian Models

می‌شود. مدل جانشین برای رفع مشکل مطرح‌شده، مدل منعطف بتا-دوجمله‌ای فضایی است. این مدل با داشتن دو پارامتر شکل برای مدل‌بندی احتمال موفقیت دوجمله‌ای از مدل دوجمله‌ای منعطف‌تر است که در ادامه معرفی می‌شود.

۲.۱ ساختار مدل

فرض کنید $Y_i|\pi_i \sim \text{bin}(n_i, \pi_i)$. برای لحاظ کردن بیش‌پراکندگی ممکن در داده‌ها، فرض کنید متغیر تصادفی π_i از توزیع بتا با پارامترهای مثبت $a_i = a(\boldsymbol{s}_i)$ و $b_i = b(\boldsymbol{s}_i)$ پیروی کند، یعنی $\pi_i \sim \text{beta}(a_i, b_i)$. انتخاب‌های مختلفی از a_i و b_i به شکل‌های گوناگونی برای چگالی π_i، شامل U-شکل، J-شکل و J-شکل معکوس، منتهی می‌شود. استفاده از توزیع بتا برای مدل‌بندی تغییرپذیری پارامتر احتمال موفقیت در توزیع دوجمله‌ای، اولین بار توسط اسکلام (۱۹۴۸) پیشنهاد شد. تعمیم آن به حجم‌های نمونه متفاوت n_i نیز اولین بار توسط ویلیامز (۱۹۷۵) در مطالعات سم‌شناسی انجام شد. با این نگاه، توزیع کناری Y_i یک توزیع بتا-دوجمله‌ای با تکیه‌گاه $\{۰, \cdots, n_i\}$ و تابع احتمال

$$\text{P}(Y_i = y_i|a_i, b_i) = \binom{n_i}{y_i} \frac{\text{Beta}(a_i + y_i, n_i + b_i - y_i)}{\text{Beta}(a_i, b_i)}$$

است، که در آن $\text{Beta}(\cdot, \cdot)$ تابع بتا است. با شرط معلوم بودن a_i و b_i، میانگین و واریانس Y_i به‌ترتیب

$$\text{E}(Y_i) = n_i \frac{a_i}{a_i + b_i}, \quad \text{Var}(Y_i) = \frac{n_i a_i b_i (n_i + a_i + b_i)}{(a_i + b_i)^۲ (۱ + a_i + b_i)}$$

هستند، که بر اساس آن‌ها، توسعه یک چارچوب رگرسیون فضایی برای مدل‌بندی پارامترهای a_i و b_i و القای یک وابستگی بین آن دو برای تشخیص پاسخ Y_i، چندان روشن نیست. برای مرتفع کردن این مشکل، به منظور وارد کردن اثرات متغیرهای تبیینی رگرسیونی و اثر تصادفی فضایی به مدل، به‌طوری که تعبیر ضرایب بر اساس میانگین توزیع بتای مشخص‌شده قابل بیان باشد، مشابه کار فراری و کریباری (۲۰۰۴)، باندیوپادیای و همکاران (۲۰۱۱) استفاده از یک نسخه بازپارامتری‌شده توزیع بتا را پیشنهاد کردند. برای این منظور با قرار دادن $\mu(\boldsymbol{s}_i) = \mu_i = \frac{a_i}{a_i + b_i}$ و $\gamma_i = a_i + b_i$، پارامترهای اصلی توزیع بتا برای π_i به‌صورت $a_i = \gamma_i \mu_i$ و $b_i = \gamma_i(۱ - \mu_i)$ بازنویسی می‌شوند. بنابراین $\pi_i \sim \text{beta}(\mu_i \gamma_i, (۱ - \mu_i)\gamma_i)$، که دارای میانگین μ_i و واریانس $\frac{\mu_i(۱-\mu_i)}{۱+\gamma_i}$ است. در نتیجه $Y_i \sim$

$\text{betabin}(n_i, \mu_i\gamma_i, (1-\mu_i)\gamma_i)$. در این مدل، با توجه به آن‌که $\mu_i \in (0,1)$، از تابع پیوند لوجیت نیز می‌توان برای پیوند μ_i به پیشگوی η_i استفاده کرد. البته سایر توابع پیوند مانند پروبیت یا لگ-لگ متمم (مک‌کالاک و نلدر ، ۱۹۸۹) را نیز می‌توان به‌عنوان جانشین به‌کار برد. میانگین و واریانس توزیع بتا-دوجمله‌ای بازپارامتری‌شده حاصل عبارتند از $\mathrm{E}(Y_i) = n_i\mu_i$ و $\mathrm{Var}(Y_i) = n_i\mu_i(1-\mu_i)\frac{n_i+\gamma_i}{1+\gamma_i}$. با مقایسه واریانس‌های دو مدل دوجمله‌ای و بتا-دوجمله‌ای، معلوم می‌شود که $\frac{n_i+\gamma_i}{1+\gamma_i} \in [1, n_i)$ پارامتر بیش‌پراکنشی مدل بتا-دوجمله‌ای است. با توجه به این شاخص، زمانی که $n_i = 1$ یا $\gamma_i \to \infty$، واریانس مدل بتا-دوجمله‌ای به واریانس مدل دوجمله‌ای همگرا می‌شود. در مثال کاربردی مورد نظر این مقاله، $n_i \geq 1$. از طرفی، در عمل پارامتر γ_i را برای تمام ناحیه تحت مطالعه ثابت γ در نظر می‌گیرند (باندیوپادیای و همکاران ، ۲۰۱۱). ویژگی این نوع مدل‌بندی، وارد کردن مستقیم بیش‌پراکنشی به مدل بدون تاثیر در میانگین پاسخ است. بنابراین مدل بتا-دوجمله‌ای برای زمانی که بعضی از مقادیر توزیع دوجمله‌ای بیشتر از مقدار معمول رخ می‌دهند، مناسب است.

۳ تحلیل بیزی مدل بتا-دوجمله‌ای

فرض کنید μ_i دارای ساختار (۲) و بردار $\boldsymbol{x}$ شامل متغیرهای گاوسی α، $\{f^{(j)}(\cdot)\}$، $\{\beta_k\}$ و $\{\epsilon_i\}$ با بعد n باشد. در این صورت $\pi(\boldsymbol{x}|\boldsymbol{\theta}_1)$ تابع چگالی توزیع گاوسی n متغیره با بردار میانگین صفر و ماتریس دقت $Q(\boldsymbol{\theta}_1)$ است، که در آن $\boldsymbol{\theta}_1$ بردار ابرپارامترها است. فرض کنید مشاهدات پاسخ، به شرط معلوم بودن $\boldsymbol{x}$ و γ، مستقل شرطی باشند و از مدل بتا-دوجمله‌ای فضایی $\text{betabin}(n_i, \mu_i\gamma_i, (1-\mu_i)\gamma_i)$ پیروی کنند. با در نظر گرفتن $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \gamma)^T$ با بعد m، تابع درستنمایی مدل بتا-دوجمله‌ای فضایی، به شرط متغیر تصادفی پنهان $\boldsymbol{x}$ (که شامل اثر فضایی نیز هست)، به صورت

$$\mathrm{L}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n_d} \Big[\binom{n_i}{y_i} \frac{\big[\prod_{k=0}^{y_i-1}(\gamma\mu_i + k)\big]\big[\prod_{k=0}^{n_i-y_i-1}(\gamma(1-\mu_i)+k)\big]}{\prod_{k=0}^{n_i-1}(\gamma+k)} \Big] \quad (۳)$$

است، که در آن برای هر r حقیقی مثبت، $\prod_{k=1}^{-1}(r+k) \equiv 1$. واضح است که تابع درستنمایی (۳) به خانواده نمایی تعلق ندارد و محاسبه آن پیچیده است. از طرفی استفاده از نسخه‌های تقریبی برآوردهای

ماکسیمم درستنمایی، مانند برآوردهای شبه درستنمایی (مک‌کالاک و نلدر ، ۱۹۸۹)، به شدت پیچیده و ناکاراست (باندیوپادیای و همکاران ، ۲۰۱۱). این پیچیدگی، گرایش به استفاده از رهیافت استنباط بیزی برای برازش مدل بتا-دوجمله‌ای فضایی را در قالب رده LGMs به روشنی بیان می‌کند.

۳.۱ کارایی روش INLA در برازش مدل پیشنهادی

اگر $\pi(\boldsymbol{\theta})$ توزیع پیشین ابرپارامترها باشد، توزیع پسین توام برای مدل بتا-دوجمله‌ای فضایی به‌صورت $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{L}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y})$ حاصل می‌شود. تقریب گاوسی چگالی شرطی کامل

$$\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}) \propto \exp\Big\{-\frac{1}{2}\boldsymbol{x}^T Q(\boldsymbol{\theta}_1)\boldsymbol{x} + \sum_{i=1}^{n_d} g_i(x_i)\Big\} \qquad (۴)$$

نقش مهمی در INLA دارد، که در آن $g_i(x_i) = \log \pi(y_i|x_i,\boldsymbol{\theta})$ تابع لگاریتم درستنمایی مدل بتا-دوجمله‌ای برای مشاهده i ام در (۳) است. بسط تیلور مرتبه دوم $g_i(x_i)$ حول مقدار اولیه $\zeta^{(0)}$ بصورت

$$g_i(x_i) \approx g_i(\zeta^{(0)}) + c_i x_i - \frac{1}{2} d_i x_i^2$$

است، که در آن ضرایب c_i و d_i به $\zeta^{(0)}$ وابسته هستند. یک تقریب گاوسی برای (۴) با به کارگیری ماتریس دقت $Q + \mathrm{diag}(\boldsymbol{d})$ و مد محاسبه‌شده توسط معادله $\{Q + \mathrm{diag}(\boldsymbol{d})\}\boldsymbol{\zeta}^{(1)} = \boldsymbol{c}$ به دست می‌آید، که در آن بردارهای $\boldsymbol{c}$ و $\boldsymbol{d}$ شامل مقادیر c_i ها و d_i ها هستند و منظور از $\mathrm{diag}(\boldsymbol{d})$ ماتریس قطری ساخته‌شده توسط بردار $\boldsymbol{d}$ است. این فرآیند تا همگرایی به یک توزیع گاوسی با بردار میانگین $\boldsymbol{\zeta}^*$ و ماتریس دقت $Q^* = Q + \mathrm{diag}(\boldsymbol{d}^*)$ ادامه می‌یابد، که در آن $\boldsymbol{d}^* = \boldsymbol{d}(\boldsymbol{\zeta}^*)$. در روش INLA این تقریب را با $\pi_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ نشان می‌دهند. برای انجام استنباط در مدل، باید توابع چگالی پسین کناری

$$\begin{aligned}
\pi(x_i|\boldsymbol{y}) &= \int \pi(x_i|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \quad i = 1,\ldots,n \\
\pi(\theta_j|\boldsymbol{y}) &= \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}, \quad j = 1,\ldots,m
\end{aligned}$$

را تقریبی به‌دست آورد، که در آن $\boldsymbol{\theta}_{-j}$ بردار ابرپارمترها بدون مولفه j ام آن است. توزیع پسین بصورت

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})}{\pi_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} \qquad (۵)$$

تقریب زده می‌شود، که در آن $\boldsymbol{x}^*(\boldsymbol{\theta})$ مد چگالی شرطی کامل $\boldsymbol{x}$ برای $\boldsymbol{\theta}$ مفروض است. تقریب (۵) همان تقریب لاپلاس برای یک توزیع پسین کناری است. این تقریب برای زمانی که چگالی شرطی کامل $\boldsymbol{x}$ گاوسی است، دقیق است. اکنون پرسش اصلی برای مدل پیشنهادی بتا-دوجمله‌ای، که عضو خانواده توزیع‌های نمایی هم نیست، اطمینان از امکان اجرای کارای روش INLA است. برای رسیدن به این اطمینان باید چگالی شرطی (۴)، حداقل در یک همسایگی حول مد خود، لگ-مقعر[1] و به یک چگالی گاوسی نزدیک باشد (مارتینز و رو ، ۲۰۱۴). با توجه به ساختار (۴)، کافی است لگاریتم تابع درستنمایی مدل بتا-دوجمله‌ای مقعر باشد. بنابراین باید نشان داده شود $\frac{d^2}{d\eta_i^2}\log\pi(y_i|\eta_i) \leq 0$. لوین و ریدز (۱۹۷۷) نشان دادند تابع درستنمایی یک مدل چندجمله‌ای مرکب که در حالت خاص، با داشتن تنها دو رده، همان مدل بتا-دوجمله‌ای است، بر حسب a_i و b_i لگ-مقعر است. بنابراین به سادگی می‌توان دریافت که تابع درستنمایی (۳) بر حسب μ_i لگ-مقعر است. با توجه به این‌که تابع پیوند لجیت $\eta_i = g(\mu_i) = \ln(\frac{\mu_i}{1-\mu_i})$ یک تابع پیوسته و مشتق‌پذیر خوش‌تعریف است، لگ-مقعر بودن تابع درستنمایی مدل پیشنهادی بر حسب η_i قابل دریافت است. این نتیجه یک اطمینان نسبی از معتبر و کارا بودن نتایج حاصل از تقریب INLA در مدل بتا-دوجمله‌ای فضایی ایجاد می‌کند. کارایی تقریب INLA به سه ویژگی پایه‌ای دیگر برای مدل‌هایی که این روش برای برازش و استنباط بیزی در آن‌ها به‌کار می‌رود، وابسته است:

الف- پذیره گاوسی بودن میدان پنهان $\boldsymbol{x}$. این پذیره توزیعی تاثیری غیرقابل چشم‌پوشی بر توزیع پسین توام مدل دارد.

ب- ویژگی مارکوفی یا همان استقلال شرطی میدان تصادفی پنهان $\boldsymbol{x}$، که اغلب دارای بعد بزرگی در مقیاس‌هایی مثل $n = 10^2$ تا $n = 10^5$ است. در این حالت GRF تعریف‌شده روی $\boldsymbol{x}$ یک میدان تصادفی مارکوفی گاوسی[2] (GMRF) با یک ماتریس دقت تنک است. بنابراین برای محاسبات ماتریسی، شامل تجزیه ماتریس دقت، می‌توان از روش‌های عددی کارا برای ماتریس‌های تنک استفاده کرد که بسیار سریع‌تر از محاسبات ماتریس‌های چگال هستند (رو و هلد ، ۲۰۰۵). کارایی محاسباتی روش INLA بیشتر مرهون برقراری همین ویژگی مارکوفی است که منتهی به یک ماتریس دقت تنک می‌شود.

[1] Log-concave
[2] Gaussian Markov Random Field

ج- کوچک بودن بعد ابرپارامترهای $\boldsymbol{\theta}$ مثل $m \leq ۶$.

معمولا هر سه این ویژگی‌ها برای برازش و انجام یک استنباط کارا، از نظر محاسباتی، لازم هستند. با کوچک بودن بعد بردار ابرپارامترها، برقراری ویژگی مارکوفی میدان تصادفی پنهان برای بسیاری از کاربردها در رده LGMs منطقی است؛ مثلا برای داده‌های مشبکه‌ای ناحیه‌ای، مارکوفی بودن میدان تصادفی پنهان کاملا پذیرفتنی است. اما برای داده‌های فضایی زمین‌آماری که ناحیه تحت مطالعه برای آن‌ها چگال است، این ویژگی به سختی برقرار است و ماتریس دقت حاصل تنک نیست. در این حالت، روش INLA کارایی محاسباتی خود را از دست می‌دهد. لیندگرن و همکاران (۲۰۱۱) برای کارا کردن روش INLA در استفاده و تحلیل میدان‌های تصادفی چگال، یک روش مبتنی بر معادلات دیفرانسیل جزیی تصادفی معرفی کردند که در آن یک تقریب GMRF برای یک GRF چگال با کمک مثلث‌بندی ناحیه تحت مطالعه و بر اساس روش عناصر متناهی[1] فراهم می‌شود. از تقریب GMRF حاصل، ماتریس دقت تنکی به‌دست می‌آید که کارایی محاسباتی روش INLA را در پی خواهد داشت. در واقع، با این رهیافت، با وجود این‌که تمام مدل‌بندی‌ها در یک ناحیه پیوسته چگال صورت می‌گیرد، ولی تمام محاسبات از طریق یک GMRF و با یک ماتریس دقت تنک انجام می‌شود. بنابراین استفاده همزمان از رهیافت SPDE و روش INLA راه حل مناسبی برای غلبه بر پیچیدگی محاسباتی در تحلیل داده‌های زمین‌آماری است.

رهیافت SPDE

فرض کنید $f(\boldsymbol{s}) = \{f(\boldsymbol{s}), \ \boldsymbol{s} \in D \subseteq R^d\}$ یک GRF مانا با ساختار کوواریانس فضایی مترن (استاین ، ۲۰۰۵) به صورت

$$C(\boldsymbol{h}) = \frac{\sigma^۲}{۲^{\nu-۱}\Gamma(\nu)}(\kappa\|\boldsymbol{h}\|)^{\nu}K_{\nu}(\kappa\|\boldsymbol{h}\|) \qquad (۶)$$

باشد، که در آن $\|\boldsymbol{h}\|$ فاصله اقلیدسی بین دو موقعیت $\boldsymbol{s}$ و $\boldsymbol{s}'$ است، به‌طوری که $\boldsymbol{h} = \boldsymbol{s} - \boldsymbol{s}'$، $\Gamma(\cdot)$ تابع گاما و $K_{\nu}(\cdot)$ تابع بسل اصلاح‌شده نوع دوم از مرتبه ν است. برای تابع کوواریانس مترن، $\sigma^۲$ واریانس کناری است و ν درجه هموار‌ی تابع را کنترل می‌کند که معمولا مقداری ثابت است. در رهیافت INLA+SPDE، برای $d = ۲$، لیندگرن و همکارانش ν را برابر مقدار ثابت ۱ در نظر گرفتند. افزون بر این، $\kappa > ۰$ پارامتر مقیاس با رابطه $r = \sqrt{۸\nu}/\kappa$ برای پارامتر دامنه تجربی r است که بنا بر پیشنهاد

[1] Finite elements

لیندگرن و همکاران (۲۰۱۱) آن را طوری تعیین می‌کنند که به ازای هر ν همبستگی فضایی نزدیک به ۰٫۱ باشد. به ازای یک مقدار ثابت برای ν، مقدار بزرگ‌تر r همبستگی فضایی قوی‌تر را نشان می‌دهد.

روش SPDE پیشنهادی لیندگرن و همکاران (۲۰۱۱)، یک GMRF با ماتریس دقت تنک Q پیدا می‌کند به‌طوری که به بهترین صورت نشان‌دهنده میدان تصادفی مترن باشد. دلیل انتخاب ساختار کوواریانسی مترن، نتیجه گزارش‌شده توسط ویتل (۱۹۶۳) است. ویتل ثابت کرد که میدان‌های تصادفی مترن تنها پاسخ مانای یک معادله دیفرانسیل جزیی تصادفی هستند. به‌طور دقیق‌تر، میدان تصادفی $f(\boldsymbol{s})$ با تابع کوواریانس مترن (۶) یک پاسخ مانا برای معادله دیفرانسیل تصادفی بصورت

$$(\kappa^2 - \triangle)^{\frac{\zeta}{2}}(\tau f(\boldsymbol{s})) = \Omega(\boldsymbol{s}), \quad \boldsymbol{s} \in D \subseteq \Re^d \tag{۷}$$

است، که در آن $\triangle$ عملگر دیفرانسیلی مرتبه دوم یا همان عملگر لاپلاس است، τ واریانس را کنترل می‌کند و فرآیند $\Omega(\boldsymbol{s})$ یک نوفه سفید گاوسی فضایی با واریانس برابر یک است. رابطه دو پارامتر ν و σ^2 در تابع کوواریانس مترن با معادله دیفرانسیل تصادفی (۷) با دو برابری

$$\zeta = \nu + d/2, \quad \sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\zeta)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}$$

مشخص می‌شود. برای به‌دست آوردن یک GMRF باید یک مقدار صحیح برای ζ انتخاب شود. با در نظر گرفتن $\nu = 1$، برای $d = 2$، مقادیر $\zeta = 2$ و $\sigma^2 = 1/(4\pi\kappa^2\tau^2)$ نتیجه می‌شوند. از روی این نتیجه، واضح است که هر دو پارامتر κ و τ بر روی واریانس کناری میدان تصادفی $f(\boldsymbol{s})$ تاثیر دارند.

می‌توان تقریبی برای پاسخ معادله دیفرانسیل (۷)، با به‌کارگیری روش عناصر متناهی به‌دست آورد که بر روی یک شبکه مثلث‌بندی از ناحیه تحت مطالعه D تعریف شده است. روش عناصر متناهی بر اساس مثلث‌بندی یک روش عددی برای حل تقریبی معادلات دیفرانسیل جزیی است. به‌طور جزیی‌تر، ناحیه D به مجموعه‌ای از مثلث‌ها تقسیم می‌شود که در بیشتر از یک یال یا زاویه مشترک متقاطع نیستند. رأس‌های مثلث‌بندی اولیه در n_d موقعیت مکانی مشاهده‌شده قرار می‌گیرند و سپس سایر رأس‌ها برای دستیابی به یک مثلث‌بندی مناسب برای پیش‌گویی فضایی اضافه می‌شوند تا تعداد کل رأس‌ها به n برسد. نمایش تقریب عناصر متناهی معادله (۷) به‌صورت $f(\boldsymbol{s}) = \sum_{k=1}^{n} \psi_k(\boldsymbol{s})\omega_k$ است، که در آن $\{\omega_k\}$ وزن‌های تصادفی با توزیع گاوسی هستند و به شکلی انتخاب می‌شوند که توزیع $f(\boldsymbol{s})$ تقریب خوبی از توزیع پاسخ معادله (۷) را در ناحیه D نتیجه دهد. همچنین، توابع $\{\psi_k\}$ توابع پایه تکه‌ای خطی هستند، به‌طوری که

در راس k برابر یک و در سایر رئوس برابر صفر هستند. این توابع تکه‌ای خطی یک ساختار مارکوفی را به میدان القا می‌کنند که به یک ماتریس دقت تنک منتهی می‌شود. لازم به تاکید است که دقت نتایج تقریب به نحوه مثلث‌بندی ناحیه وابسته است.

۳.۲ انتخاب توزیع‌های پیشین

برای تکمیل مدل بتا-دوجمله‌ای فضایی و انجام استنباط، در یک چارچوب بیزی، باید برای ابرپارامترهای مدل، یعنی ۱) پارامتر بیش‌پراکندگی مدل بتا-دوجمله‌ای، γ، ۲) پارامتر انحراف معیار کناری، σ، و ۳) پارامتر دامنه تجربی، r، توزیع پیشین مناسب انتخاب کرد. ابرپارامترهای توزیع پیشین گاوسی برای ضرایب رگرسیونی $\boldsymbol{\beta}$ نیز باید تعیین شوند. دلیل استفاده از پارامترهای σ و r به‌جای τ و κ، ترجیح رهیافت SPDE معرفی‌شده توسط لیندگرن و همکارانش است (لیندگرن و رو ، ۲۰۱۵). ابتدا، فرض کنید مولفه‌های فضای پارامتر مدل بتا-دوجمله‌ای (و به‌طور مشابه مدل دوجمله‌ای) از هم مستقل باشند؛ یعنی

$$\pi(\boldsymbol{\beta}, \gamma, \sigma, r) = \pi(\boldsymbol{\beta})\pi(\gamma)\pi(\sigma)\pi(r).$$

برای بردار ضرایب رگرسیونی با توزیع پیشین گاوسی، مولفه‌های بردار مستقل با میانگین صفر و واریانس بزرگ (برابر ۱۰۰۰) در نظر گرفته می‌شوند. این انتخاب بر پایه تعیین یک توزیع پیشین ناآگاهی‌بخش است که از نبودن اطلاعات پیشین در تحلیل مثال کاربردی مورد نظر این مقاله، نشأت گرفته است. توزیع پیشین پارامتر γ در مدل بتا-دوجمله‌ای بر اساس تبدیلی از آن تعریف می‌شود؛ به‌طور دقیق، برای $\log(\frac{۱}{\gamma})$ یک توزیع پیشین نرمال با میانگین صفر و واریانس ۲٫۵ در نظر گرفته می‌شود که انتخاب پیش‌فرض بسته R-INLA است. مقدار ۲٫۵ برای واریانس توزیع پیشین، بر اساس شبیه‌سازی‌های صورت‌گرفته توسط تیم توسعه‌دهنده INLA، به این دلیل انتخاب شده است که یک توزیع آگاهی‌بخش ضعیف نتیجه دهد.

برای تعیین توزیع پیشین دو پارامتر ساختار وابستگی فضایی مترن، از رهیافت جدیدی استفاده شده است که توسط سیمپسون و همکاران (۲۰۱۷) معرفی و پیشین با پیچیدگی تاوانیده[1] (PC) نام‌گذاری شده است. پیشین‌های PC ویژگی‌های جذابی را ارایه کرده‌اند. برخی از آن‌ها عبارتند از ۱) پایایی نسبت به بازپارامتری شدن، ۲) ارتباط با پیشین‌های جفریز، ۳) دارا بودن ویژگی‌های استواری و ۴) پشتیبانی از اصل پنجره اوکام[2]. پیشین‌های PC مورد استفاده در این مقاله، برای پارامترهای وابستگی میدان مترن،

[1] Penalized Compelxity
[2] Occam's razor

اولین بار توسط فوگلستاد و همکاران (۲۰۱۹) معرفی شدند به‌طوری که توزیع پیشین PC برای σ با تعیین پارامترهای $\sigma_0 > 0$ و $0 < \alpha_1 < 1$ به شکلی تعریف می‌شود که $\mathrm{P}(\sigma > \sigma_0) = \alpha_1$. به‌طور مشابه، برای پارامتر r باید $r_0 > 0$ و $0 < \alpha_2 < 1$ طوری تعیین شوند که $\mathrm{P}(r < r_0) = \alpha_2$. باید توجه داشت که در این تعریف‌ها $\mathrm{P}(\cdot)$ تابع احتمال تعریف‌شده بر اساس توزیع پیشین PC مورد نظر است. مقادیر σ_0، r_0، α_1 و α_2 توسط کاربر طوری تعیین می‌شوند که شدت باور آن‌ها نسبت به همگرایی مدل به یک مدل ساده بدون وابستگی فضایی را منعکس کنند (کراینسکی و همکاران ، ۲۰۱۸).

۳.۳ ارزیابی و انتخاب مدل

برای انتخاب مدل، از معیار اطلاع انحراف[1] (DIC) استفاده می‌شود (اشپیگل‌هالتر و همکاران ، ۲۰۰۲)، که نیکویی برازش را به همراه پیچیدگی آن در یک چارچوب بیزی منعکس می‌کند. این معیار به‌صورت $\mathrm{DIC} = \bar{D} + p_D$ تعریف می‌شود، که در آن $\bar{D} = \mathrm{E}(D(\boldsymbol{\varphi})|\boldsymbol{y})$ میانگین پسینی آماره انحراف، $\boldsymbol{\varphi}$ مجموعه پارامترهای مدل، و p_D تعداد پارامترهای موثر مدل است. اشپیگل‌هالتر و همکاران (۲۰۰۲) کمیتی برای تقریب p_D به‌صورت $p_D = \bar{D} - D(\hat{\boldsymbol{\varphi}})$ معرفی کردند که در آن $\hat{\boldsymbol{\varphi}}$ برآوردی از بردار پارامترها مانند میانگین یا میانه پسینی است. هر چه مقدار DIC برای یک مدل کوچکتر باشد، آن مدل تقریب بهتری از مدل واقعی است که داده‌ها از آن تولید شده‌اند.

اگرچه DIC معیاری برای سنجش نسبی نیکویی برازش در بین مدل‌های رقیب است، اما اطلاعی از کفایت مدل فراهم نمی‌کند. به همین دلیل، از دو معیار دیگر برای ارزیابی عملکرد مدل‌ها از منظر توانایی پیشگویی آن‌ها استفاده می‌شود: اول، معیار اطلاع واتانابه-آکاییک[2] (WAIC) (واتانابه ، ۲۰۱۰) و دوم، معیار لگاریتم درستنمایی‌نمای کناری[3] (LPML) که بر اساس آماره مختصات پیشگویی شرطی[4] (CPO) تعریف می‌شود (پتیت ، ۱۹۹۰). معیار WAIC اندازه‌ای از اعتبارسنجی متقابل با هدف سنجش و مقایسه توان پیشگویی مدل‌ها فراهم می‌کند و استفاده از آن توسط گلمن و همکاران (۲۰۱۴) توصیه شده است. مقادیر کوچکتر این معیار نشان از عملکرد پیشگویی بهتر مدل دارد. معیار LPML نیز به صورت $\mathrm{LPML} = \sum_i \log(\mathrm{CPO}_i)$ تعریف می‌شود، که در آن برای هر $i = 1, \ldots, n_d$، مقدار $\mathrm{CPO}_i = \pi(y_i|\boldsymbol{y}_{-i})$ احتمال پسینی مشاهده y_i را وقتی مدل روی تمام مشاهدات بدون مشاهده i

[1] Deviance Information Criterion
[2] Watanabe-Akaike Information Criterion
[3] Log Pseudo-Marginal Likelihood
[4] Conditional Predictive Ordinate

ام برازش شده است، نشان می‌دهد. با این تفسیر، مقادیر بزرگ‌تر معیار LPML پشتیبانی بهتر مدل از داده‌های مشاهده‌شده را نشان می‌دهند.

۴ مطالعه شبیه‌سازی

به منظور ارزیابی مدل بتا-دوجمله‌ای فضایی، یک مطالعه شبیه‌سازی انجام شد. با این فرض که $u(\boldsymbol{s})$، $\boldsymbol{s} \in [۰, ۱] \times [۰, ۱]$، یک GRF با میانگین صفر و ماتریس کوواریانس مترن (۶) باشد و با استفاده از نمایش SPDE در (۷)، یک تحقق به حجم $n_d = ۱۰۰$ از این میدان با مقادیر واقعی $\nu = ۱$، $\sigma = ۱$ و $r = ۰/۳$ تولید شد. مشاهدات متغیر پاسخ با استفاده از مدل سلسله‌مراتبی بصورت

$$\begin{aligned}
(u(\boldsymbol{s}_۱), \ldots, u(\boldsymbol{s}_{n_d})) &\sim \mathrm{N}(۰, \Sigma) \\
\mathrm{logit}(\mu(\boldsymbol{s}_i)) &= \beta_۰ + \beta_۱ z(\boldsymbol{s}_i) + u(\boldsymbol{s}_i) \\
\pi(\boldsymbol{s}_i) &\sim \mathrm{beta}(\gamma\mu(\boldsymbol{s}_i), \gamma(۱ - \mu(\boldsymbol{s}_i))) \\
Y_i(\boldsymbol{s}_i) &\sim \mathrm{bin}(n(\boldsymbol{s}_i), \pi(\boldsymbol{s}_i))
\end{aligned}$$

تولید شدند، که در آن ماتریس کوواریانس Σ از تقریب SPDE میدان مترن (لیندگرن و همکاران ، ۲۰۱۱) حاصل شده است. مقادیر متغیر تبیینی z از یک توزیع نرمال استاندارد و حجم‌های نمونه $n(\boldsymbol{s}_i)$ از مجموعه $\{۱, \ldots, ۱۰\}$ با احتمال‌های برابر تولید شدند. همچنین $(\beta_۰, \beta_۱) = (-۰/۵, ۰/۷)$ به عنوان مقادیر واقعی ضرایب رگرسیونی انتخاب شدند.

برای ارزیابی اثر بیش‌پراکنشی در داده‌ها، دو طرح شبیه‌سازی مختلف بر حسب پارامتر γ در نظر گرفته شدند:

طرح اول- مدل با بیش‌پراکنشی بزرگ که در آن مقدار واقعی برای پارامتر γ برابر ۳ تعیین شد. با این انتخاب، پارامتر بیش‌پراکنشی $\frac{n_i+\gamma}{۱+\gamma}$ می‌تواند در بازه $(۱, ۱۳/۲۵)$ تغییر کند.

طرح دوم- مدل نزدیک به دوجمله‌ای که در آن مقدار واقعی برای پارامتر γ برابر ۵۰ انتخاب شد. در این حالت پارامتر بیش‌پراکنشی در بازه $(۱, ۱/۱۸)$ تغییر می‌کند که مقدار کوچکی از بیش‌پراکنشی در پاسخ را در نظر می‌گیرد.

برای محاسبه معیارهای ارزیابی مدل، ۲۰۰ مجموعه داده برای هر طرح شبیه‌سازی شدند. به هر مجموعه داده، مشابه کار باندیوپادیای و همکاران (۲۰۱۱)، چهار مدل مختلف برازش داده شدند: اول) مدل فضایی بتا-دوجمله‌ای که با نام SpaBB نشان داده شده است، دوم) مدل فضایی دوجمله‌ای با نام اختصاری SpaBin، سوم) مدل غیرفضایی بتا-دوجمله‌ای با نام اختصاری NonSpaBB و چهارم) مدل غیرفضایی دوجمله‌ای با نام اختصاری NonSpaBin. توزیع‌های پیشین ابرپارامترهای مدل بر اساس بخش انتخاب توزیع‌های پیشین تعیین شدند. مقادیر احتمال‌های دمی توزیع‌های پیشین PC برابر $\alpha_1 = \alpha_2 = ۰/۰۵$ و $r_0 = ۰/۱$ و $\sigma_0 = ۰/۰۵$ انتخاب شدند. برازش مدل‌های پیشنهادی به داده‌ها با کدنویسی در محیط نرم‌افزار R و به کمک بسته نرم‌افزاری R-INLA انجام شده است.

برای هر مدل، معیارهای DIC و LPML محاسبه شدند که به ترتیب نمودار جعبه‌ای آن‌ها در شکل‌های ۱ و ۲ نمایش داده شده‌اند. برای برآوردهای پارامترهای مدل، قدر مطلق اریبی نسبی، میانگین توان دوم خطا (MSE) و توان تجربی ضرایب رگرسیونی محاسبه و در جدول ۱ گزارش شده‌اند. توان تجربی ضرایب رگرسیونی نسبتی از مجموعه‌های شبیه‌سازی‌شده است که در آن‌ها ناحیه باورمندی ۹۵٪ ضرایب رگرسیونی، صفر را شامل نمی‌شوند. با توجه به نتایج جدول ۱، اریبی نسبی برآوردگرها در مدل

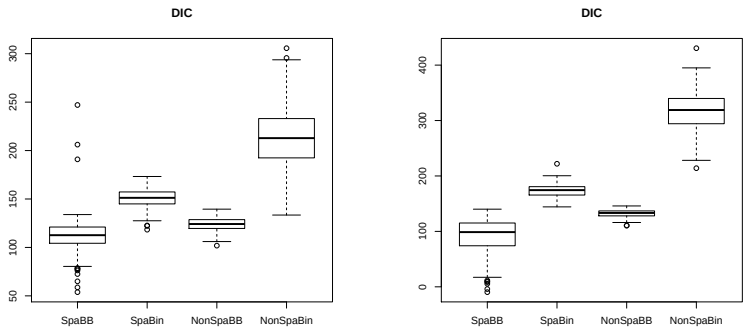
جدول ۱. مقادیر قدر مطلق اریبی نسبی، MSE و توان تجربی پارامترهای مدل‌ها در مطالعه شبیه‌سازی

طرح	مدل	اریبی نسبی (MSE)			توان تجربی	
		β_0	β_1	γ	β_0	β_1
اول	SpaBB	(۰/۸۹۸)۰/۰۵۷	(۰/۸۳۶)۰/۰۵۱	(۱/۹۰۳)۰/۰۱۱	۰/۰۹۲	۰/۹۷۴
	SpaBin	(۱/۰۷۰)۰/۳۱۳	(۱/۰۴۷)۰/۳۰۵		۰/۴۲۶	۰/۹۷۴
	NonSpaBB	(۰/۸۴۰)۰/۱۶۸	(۰/۷۸۷)۰/۱۶۰	(۱/۲۳۶)۰/۳۷۴	۰/۶۴۱	۰/۹۶۹
	NonSpaBin	(۰/۸۴۹)۰/۱۵۶	(۰/۸۰۰)۰/۱۵۴		۰/۸۰۵	۱/۰۰۰
دوم	SpaBB	(۰/۹۱۹)۰/۰۰۲	(۰/۸۳۹)۰/۰۴۲	(۳۱/۳۹۰)۰/۵۹۴	۰/۱۱۱	۱/۰۰۰
	SpaBin	(۰/۹۵۴)۰/۰۶۶	(۰/۸۷۳)۰/۰۲۸		۰/۱۴۱	۱/۰۰۰
	NonSpaBB	(۰/۸۳۷)۰/۱۸۶	(۰/۷۸۷)۰/۱۴۹	(۴۴/۱۵۸)۰/۸۸۰	۰/۶۹۳	۱/۰۰۰
	NonSpaBin	(۰/۸۴۱)۰/۱۸۲	(۰/۷۹۳)۰/۱۳۵		۰/۷۹۹	۱/۰۰۰

SpaBB، در هر دو نوع طرح شبیه‌سازی، به‌جز یک مورد در طرح نوع دوم، از سایر مدل‌ها به‌طور قابل ملاحظه‌ای کمتر است. بر اساس معیار MSE نیز مدل SpaBB بر مدل SpaBin برتری دارد، ولی نسبت به مدل‌های غیرفضایی این برتری، در حد ناچیزی، عکس است. البته در هر دو طرح شبیه‌سازی، مدل بتا-دوجمله‌ای نسبت به مدل دوجمله‌ای برتر است. مقدار بزرگ‌تر MSE برای برآوردگر γ در طرح اول، برای مدل بتا-دوجمله‌ای فضایی نسبت به نسخه غیرفضایی آن، با توجه به کوچک بودن اریبی نسبی برآوردگر، احتمالا به دلیل وجود چند برآورد بزرگ برای γ در ۲۰۰ تکرار شبیه‌سازی در مدل فضایی است. توان تجربی برای ضریب متغیر تبیینی، β_1، در هر دو طرح شبیه‌سازی و برای همه مدل‌ها تقریبا برابر و

نزدیک به ۱ است. اما توان پایین ضریب ثابت در مدل فضایی بتا-دوجمله‌ای و همین‌طور در هر دو مدل فضایی نسبت به مدل‌های غیرفضایی نتیجه جالب توجهی است. این مساله می‌تواند نشان‌دهنده مشکل شناسایی‌ناپذیر بودن ضریب ثابت رگرسیونی در مدل فضایی و مخلوط شدن این اثر با اثر فضایی باشد.

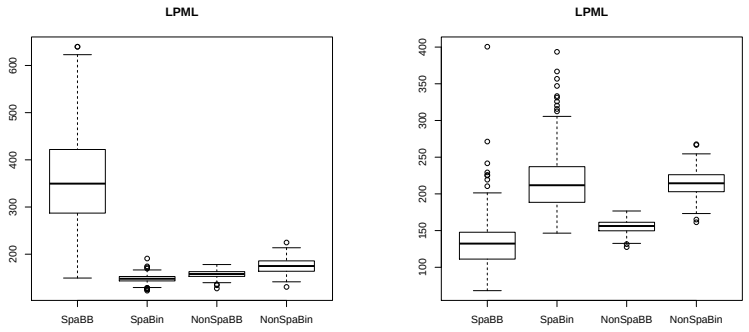
هر دو شکل ۱ و ۲ نشان می‌دهند که مدل بتا-دوجمله‌ای فضایی برای طرح شبیه‌سازی اول از منظر معیار انتخاب مدل DIC و معیار ارزیابی توان پیشگویی مدل LPML نسبت به هر سه مدل دیگر کاملا برتر است. از طرف دیگر، برای طرح شبیه‌سازی دوم، که مدل بتا-دوجمله‌ای فضایی نزدیک به مدل فضایی دوجمله‌ای است، معیار DIC باز هم برتری مدل SpaBB را نسبت به سایر مدل‌ها نشان می‌دهد، ولی از منظر معیار LPML مدل دوجمله‌ای فضایی برتری دارد. البته قابل ذکر است که این نتیجه با توجه به ساده‌تر بودن مدل دوجمله‌ای و وجود ساختار وابستگی فضایی در داده‌ها برای هر دو معیار منطقی است.



(ب) (الف)

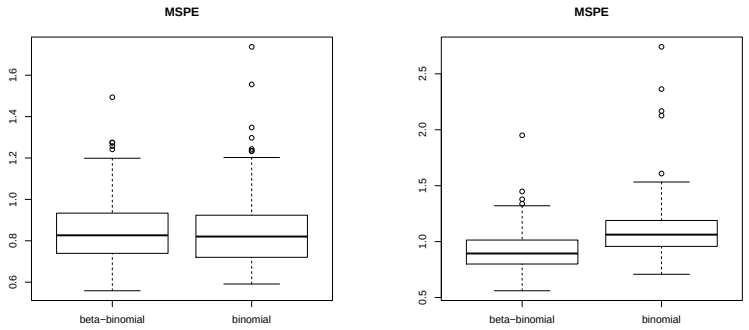
شکل ۱. مقادیر DIC برای دو مدل فضایی دوجمله‌ای و بتا-دوجمله‌ای در طرح اول (الف) و دوم (ب)

اگر چه معیارهای انتخاب و ارزیابی مدل به طور نسبی برتری مدل بتا-دوجمله‌ای فضایی را، در هر دو طرح شبیه‌سازی، نشان می‌دهند، با این حال برای کاووش بیشتر، مقادیر میانگین توان دوم خطای پیشگویی (MSPE) برای تمام مجموعه‌های داده محاسبه و نمودار جعبه‌ای آن‌ها برای دو مدل فضایی در شکل ۳ نمایش داده شده‌اند. برای حالتی که بیش‌پراکنشی در داده‌ها به اندازه بزرگ وجود دارد (طرح اول)، مقادیر MSPE برای مدل SpaBB کوچکتر از مدل SpaBin هستند. اما مساله برای طرح دوم که در آن بیش‌پراکنشی ناچیزی در داده‌ها وجود دارد و مدل بتا-دوجمله‌ای تقریبا همان مدل دوجمله‌ای است، کمی برعکس است و مقادیر MSPE برای مدل دوجمله‌ای تمایل بیشتری به کوچکتر بودن دارند. نتایج این مطالعه نشان می‌دهند وقتی بیش‌پراکنشی در داده‌های شمارشی متناهی با ساختار وابستگی فضایی



(ب) (الف)

شکل ۲. مقادیر معیار LPML برای دو مدل فضایی دوجمله‌ای و بتا-دوجمله‌ای در طرح اول (الف) و دوم (ب)



(ب) (الف)

شکل ۳. مقادیر MSPE برای دو مدل فضایی دوجمله‌ای و بتا-دوجمله‌ای در طرح اول (الف) و دوم (ب)

وجود دارد، مدل بتا-دوجمله‌ای فضایی برتری محسوسی نسبت به مدل معمول دوجمله‌ای فضایی دارد.

۵ تحلیل داده‌های تصادف‌های منجر به جرح یا فوت

داده‌ها مربوط به تصادف‌های منجر به جرح یا فوت در شهرستان مشهد در سال ۱۳۸۵ هستند که توسط سازمان ترافیک شهرداری مشهد در اختیار نویسندگان مقاله قرار داده شده است. دو هدف از تحلیل این داده‌ها مد نظر بودند: اول، شناسایی عوامل تاثیرگذار بر نرخ تصادف‌هایی که منجر به جرح یا فوت می‌شوند

و دوم، تهیه نقشه پهنه‌بندی از اثر فضایی حاکم بر نرخ این نوع تصادف‌ها برای شهر مشهد تا بر اساس آن بتوان نقاط پرخطر را شناسایی کرد. مجموعه داده‌ها شامل ۷۰۰ نقطه مکانی مختلف در سرتاسر شهر مشهد است که علاوه بر تعداد کل تصادف‌ها در آن نقاط و تعداد تصادف‌های منجر به جرح یا فوت، به‌عنوان متغیر پاسخ، شامل چند متغیر تبیینی نیز می‌باشد. این متغیرها عبارتند از: متوسط سن راننده‌هایی که در یک مکان مشخص تصادف کرده‌اند (age)، نوع محل تصادف (loc) که یک متغیر رسته‌ای با سه سطح خیابان، تقاطع و میدان است، نرخ تصادف‌هایی که در آخر هفته رخ داده‌اند (weekend)، و نرخ روشنایی (light)، که درصد تصادف‌ها در روز را نشان می‌دهد. برای وارد کردن متغیر رسته‌ای loc از دو متغیر ظاهری استفاده شد که مقادیر آن‌ها طوری تعریف شدند که خیابان را به‌عنوان رسته پایه محل تصادف برای مقایسه با دو رسته دیگر، مشخص کنند. چون متغیر تبیینی age یک متغیر پیوسته است، برای وارد کردن انعطاف بیشتر به مدل، اثر آن به صورت غیرخطی (ناپارامتری) وارد مدل شد. در رهیافت INLA، اثرات ناپارامتری متغیرهای تبیینی معمولا با فرآیند قدم زدن تصادفی مدل‌بندی می‌شوند. در این‌جا نیز از یک فرآیند قدم زدن تصادفی مرتبه اول با پارامتر دقت ϱ استفاده شد. توزیع پیشینی که برای ابرپارامتر ϱ انتخاب شد، یک پیشین PC است که بر اساس انحراف معیار فرآیند، $\sigma_a = \varrho^{-1/2}$، تعریف می‌شود. با معلوم بودن $\sigma_{\circ a}$ و α_a، این پیشین طوری تعریف می‌شود که $\mathrm{P}(\sigma_a > \sigma_{\circ a}) = \alpha_a$. بنابراین از این توزیع پیشین به‌سادگی می‌توان برای استخراج و اعمال باور پیشینی نسبت به این ابرپارامتر بهره برد. ضابطه واقعی توزیع پیشین روی ϱ به صورت

$$\pi(\varrho) = \frac{\lambda}{2}\varrho^{-3/2}\exp(-\lambda\varrho^{-1/2}), \quad \varrho > 0$$

است (سیمپسون و همکاران، ۲۰۱۷)، که در آن $\lambda = -\log(\alpha_a)/\sigma_{\circ a}$. توسعه‌دهندگان این نوع پیشین برای پارامتر دقت، مقدار $\alpha_a = 0/01$ را پیشنهاد کرده‌اند. همچنین آن‌ها برای مدل‌های لوجیت استفاده از مقدار $\sigma_{\circ a} = 0/5$ را توصیه کرده‌اند. در تحلیل این داده‌ها نیز از همین مقادیر استفاده شده است. با وارد کردن همه متغیرهای تبیینی رگرسیونی، چهار مدل رقیب برای برازش داده‌ها فهرست شدند:

مدل ۱: مدل رگرسیونی دوجمله‌ای بدون اثر فضایی که در آن پیشگوی STAR به صورت

$$\mathrm{logit}(\mu_i) = \eta_i = \alpha + \beta_1 \mathrm{loc}_{1i} + \beta_2 \mathrm{loc}_{2i} + \beta_3 \mathrm{weekend}_i + \beta_4 \mathrm{light}_i + f_{rw}(age_i)$$

است، که در آن $\text{loc}_۱$ و $\text{loc}_۲$ دو متغیر ظاهری هستند که برای وارد کردن اثر متغیر رسته‌ای محل وقوع تصادف تعریف شده‌اند.

مدل ۲: مدل رگرسیونی بتا-دوجمله‌ای بدون اثر فضایی با همان پیشگوی مدل ۱.

مدل ۳: مدل رگرسیونی دوجمله‌ای با اضافه شدن اثر تصادفی فضایی به پیشگوی مدل ۱، یعنی

$$\eta_i = \alpha + \beta_۱\text{loc}_{۱i} + \beta_۲\text{loc}_{۲i} + \beta_۳\text{weekend}_i + \beta_۴\text{light}_i + f_{rw}(age_i) + f(\boldsymbol{s}_i)$$

که در آن $f(\boldsymbol{s})$ با یک رهیافت SPDE که در بخش ۳ تشریح شد، تقریب زده شده است.

مدل ۴: مدل رگرسیونی بتا-دوجمله‌ای با پیشگوی مدل ۳.

مقادیر ابرپارامترهای توزیع‌های پیشین PC پارامترهای میدان تصادفی متزن به صورت $\sigma_۰ = ۱$، $\alpha_۱ = ۰/۰۱$، $r_۰ = ۰/۰۵$ و $\alpha_۲ = ۰/۰۱$ تعیین شدند. جدول ۲ نتایج معیارهای نیکویی برازش و کفایت مدل‌ها را بر اساس معیارهای معرفی‌شده در این مقاله، نشان می‌دهد. با کمی افزایش پیچیدگی در مدل بتا-دوجمله‌ای نسبت به مدل دوجمله‌ای و افزودن اثر تصادفی فضایی با رهیافت SPDE، بهبود قابل توجه در برازش مدل بتا-دوجمله‌ای فضایی (یعنی مدل ۴) نسبت به سایر مدل‌ها، بر اساس دو معیار DIC و LPML، نتیجه شده است. با توجه به مقادیر تقریبا یکسان معیار WAIC برای هر چهار مدل، نمی‌توان در مورد آن نظر مشخصی داد.

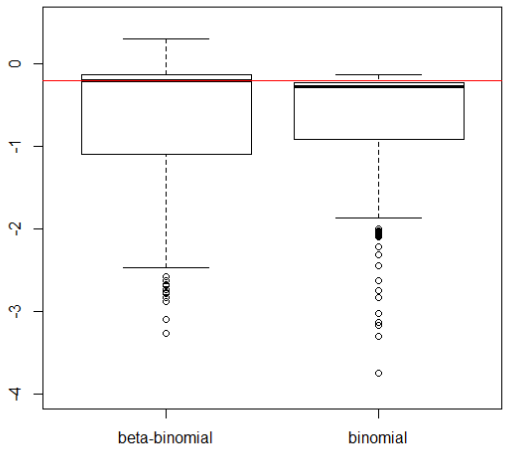
جدول ۲. مقایسه مدل‌ها با استفاده از معیارهای DIC، WAIC و LPML

مدل	DIC	WAIC	LPML
۱	۷۴۱٫۱۴۰	۸۴۵٫۶۴۱	−۴۲۲٫۸۲۶
۲	۷۰۵٫۵۹۳	۸۴۵٫۲۶۱	−۴۳۸٫۶۰۴
۳	۷۴۲٫۱۷۶	۸۴۶٫۷۸۵	−۴۲۳٫۴۹۷
۴	۶۷۹٫۱۹۷	۸۴۶٫۳۹۳	−۴۰۸٫۴۶۸

به‌طور دقیق‌تر، اختلاف مقادیر دو معیار DIC و LPML بین مدل‌های ۳ و ۴ به اندازه کافی بزرگ هست که بتوان برتری مدل بتا-دوجمله‌ای را با قدرت نتیجه گرفت. همان‌طور که از روی مقدار LPML مشاهده می‌شود، اضافه کردن اثر تصادفی فضایی به مدل بتا-دوجمله‌ای، عملکرد اعتبارسنجی متقابل از نوع خارج کردن هر بار یک مشاهده (که به LOO[1] معروف است) را بهبود بخشیده است. برای ارزیابی جزئی‌تر اعتبار مدل بر حسب عملکرد پیشگویی، از نمودار جعبه‌ای مقادیر آماره $\log(\text{CPO})$ برای دو

[1] Leave-one-out

مدل دوجمله‌ای فضایی و بتا-دوجمله‌ای فضایی که در شکل ۴ مشاهده می‌شود، استفاده شده است. مقادیر بزرگ‌تر پشتیبانی قوی‌تر داده‌ها از مدل را نشان می‌دهند. در شکل ۴ میانه مقادیر $\log(\text{CPO})$ برای مدل بتا-دوجمله‌ای فضایی با خط افقی نشان داده شده است. نتیجه حاکی از برتری مدل بتا-دوجمله‌ای است. میانه $\log(\text{CPO})$ برای دو مدل دوجمله‌ای و بتا-دوجمله‌ای به ترتیب برابر ۰٫۲۸۰− و ۰٫۲۰۶− هستند که به اندازه کافی از هم دور هستند.



شکل ۴. نمودار جعبه‌ای مقادیر $\log(\text{CPO})$ برای دو مدل دوجمله‌ای و بتا-دوجمله‌ای.

در نهایت، مدل بتا-دوجمله‌ای مدل برتر منتخب است که نتایج برازش و استنباط بر اساس آن استخراج شده‌اند. جدول ۳ برآوردهای میانگین پسینی و ناحیه باورمندی ٪۹۵ پارامترهای مدل را نشان می‌دهد. اثر برآوردشده متغیر تبیینی متوسط سن نیز در شکل ۵ ترسیم شده است. با توجه به نتایج گزارش‌شده می‌توان یافته‌های زیر را بیان کرد:

۱- دامنه تجربی وابستگی فضایی مشاهدات نزدیک به ۷۲۰ متر برآورد شده است که می‌تواند تا ۲٫۵ کیلومتر نیز باشد. با توجه به ساختار شهری خیابان‌ها و تقاطع شهری مثل مشهد، این عدد برای وابستگی فضایی مشاهدات و تاثیرگذاری نتیجه تصادف معقول است.

۲- برآورد پسینی انحراف معیار میدان تصادفی فضایی برابر ۰٫۱۲۶، با ناحیه باورمندی (۰٫۰۰۳, ۰٫۳۱۰)، است که به وضوح از صفر فاصله دارد. این برآورد، وجود معنی‌دار تغییرپذیری فضایی را در داده‌ها نشان

می‌دهد.

۳- برآورد γ برابر ۱۴٫۳۲۹ به‌دست آمده است که مقدار $\frac{\gamma+n_i}{\gamma+1} \in (1, 1.587)$ را برای پارامتر بیش‌پراکنش، به ازای تعداد تصادف‌های مشاهده‌شده در نقاط مکانی که از ۱ تا ۱۰ تصادف ثبت شده‌اند، نشان می‌دهد. این دامنه مشاهده‌شده برای پارامتر بیش‌پراکنش، حاکی از وجود نسبتا ملایم بیش‌پراکنشی در داده‌ها است.

۴- اثر برآوردشده برای سن (شکل ۵) نشان می‌دهد که این متغیر بر روی نرخ تصادفات منجر به جرح یا فوت تاثیر معنی‌داری ندارد. جدا از این نتیجه کلی، تا تقریبا سن ۳۸ سالگی، نرخ تصادف منجر به جرح یا فوت افزایشی است و پس از آن کاهشی، به‌طوری که از سن ۶۰ سالگی به بعد اثر آن صفر می‌شود. دلیل نوار باورمندی پهن، به‌ویژه برای گروه‌های سنی زیر ۳۰ سال و بالای ۵۰ سال، تعداد مشاهدات کم برای این رده‌های سنی نسبت به رده سنی بین ۳۰ تا ۵۰ سال است.

۵- برآوردهای پارامترهای ثابت را می‌توان بر حسب (به‌طور متوسط) افزایش یا کاهش در بخت[1] احتمال داشتن یک تصادف منجر به جرح یا فوت، به ازای افزایش یک واحدی در متغیرهای تبیینی پیوسته یا تغییر رسته نسبت به رسته پایه برای متغیر رسته‌ای محل تصادف، بعد از تعدیل اثر فضایی، تعبیر کرد. به‌عنوان مثال، با توجه به معنی‌داری نرخ روشنایی روز، نسبت بخت این متغیر برابر $e^{0.597} = 1.82$ با ناحیه باورمندی (۱٫۱۹, ۲٫۷۸) است. بنابراین، می‌توان گفت (به‌طور متوسط) بخت داشتن احتمال تصادف منجر به جرح یا فوت به ازای یک واحد افزایش در نرخ روشنایی روز، با در نظر گرفتن اثر فضایی، ۸۲٪ افزایش خواهد داشت. به عبارت دیگر، این نوع تصادف در روشنایی روز شانس بیشتری برای وقوع دارد.

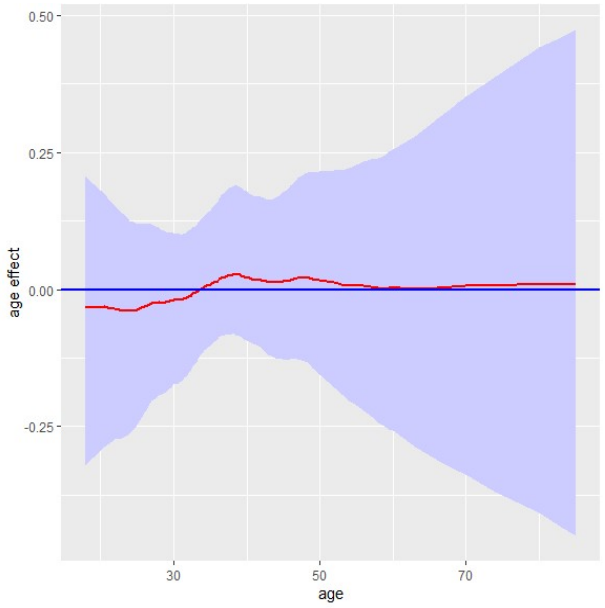
۶- با توجه به ضریب منفی اثرات برآوردشده برای محل وقوع تصادف با نوع تقاطع و میدان، می‌توان گفت که بخت وقوع این نوع تصادف در خیابان‌ها بیشتر از تقاطع یا میدان‌ها است. این نتیجه به‌طور منطقی نیز درست است، زیرا معمولا در تقاطع و میدان‌ها سرعت رانندگی پایین بوده و احتمال وقوع حوادث جرحی یا فوتی کمتر از خیابان‌ها است. البته لازم به توجه است که محل وقوع تصادف نسبت بالایی از مشاهدات، خیابان است که می‌تواند این نتیجه را تحت تاثیر قرار دهد.

نقشه پهنه‌بندی اثر فضایی بر اساس مدل برازش‌شده منتخب، به همراه برآورد انحراف معیار متناظر آن در شکل ۶ نمایش داده شده‌اند. با توجه به نقشه پهنه‌بندی اثر فضایی، کاملا مشخص است که در بخش‌های مرکزی متمایل به جنوب شرقی شهر، شانس داشتن تصادف‌های منجر به جرح یا فوت افزایش می‌یابد. در مقابل، در بخش‌های شمال غربی شهر این شانس کاهش می‌یابد. البته تردد در بخش‌های جنوب شرقی نیز بیشتر است و نسبت موقعیت‌های مکانی تصادف در این ناحیه نیز بیشتر از شمال غربی است. به‌عنوان

[1] Odds

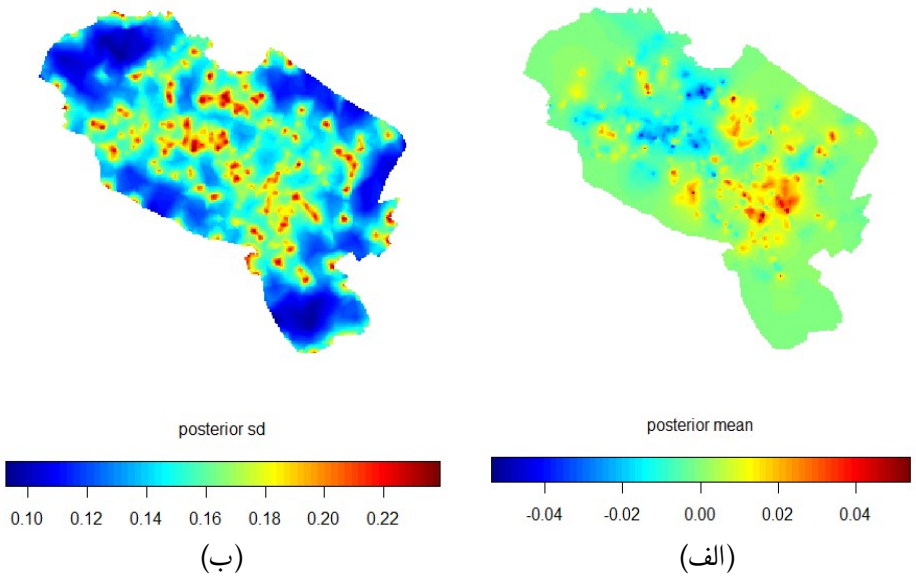
جدول ۳. برآوردهای پسینی و نواحی باورمندی ٪۹۵ برای مدل بتا-دوجمله‌ای فضایی

پارامتر	میانگین	ناحیه باورمندی ٪۹۵
ضریب ثابت	−۱٫۹۰۸	(−۲٫۲۹۳, −۱٫۵۴۵)
محل تصادف (رده پایه: خیابان)		
تقاطع	−۰٫۳۹۷	(−۱٫۴۰۳, ۰٫۵۵۴)
میدان	−۰٫۰۹۳	(−۶۲٫۰۴۷, ۶۱٫۸۶۱)
نرخ تعطیلات آخر هفته	۰٫۳۱۴	(−۰٫۱۰۷, ۰٫۷۳۱)
نرخ روشنایی	۰٫۵۹۷	(۰٫۱۷۵, ۱٫۰۲۲)
γ	۱۴٫۳۲۹	(۲٫۸۰۵, ۳۳٫۴۴۰)
r	۰٫۷۱۷	(۰٫۰۰۲, ۲٫۵۴۷)
σ	۰٫۱۲۶	(۰٫۰۰۳, ۰٫۳۱۰)



شکل ۵. اثر برآوردشده متغیر تبیینی متوسط سن رانندگان به همراه نوار باورمندی ٪۹۵

نتیجه‌گیری، نقاط پرخطر برای وقوع تصادف‌های خطرناک بر روی نقشه مشخص هستند که می‌تواند مورد توجه مدیریت ترافیک شهری مشهد برای کنترل کاراتر و پیشگیری بیشتر قرار گیرد.



شکل ۶. نقشه پهنه‌بندی برآورد اثر فضایی (الف) و انحراف معیار آن (ب) برای داده‌های تصادفات

بحث و نتیجه‌گیری

تبعات ناخوشایند ناشی از تصادف‌هایی که منجر به جرح یا فوت برخی از سرنشینان یا اشخاص ثالث می‌شوند، برای افراد درگیر و کل جامعه روشن و بدیهی است. بنابراین، شناخت عوامل تاثیرگذار بر افزایش نرخ این نوع حوادث، برای تصمیم‌سازان و برنامه‌ریزان شهری، یک فرآیند حیاتی برای کاهش این تبعات است. این مساله برای شهرستان مشهد، به‌عنوان دومین شهر بزرگ کشور، با توجه به جاذبه مذهبی حضور مرقد مبارک ثامن‌الحجج، حضرت رضا (ع)، برای زائران و مسافران و مسوولان شهری مشهد به‌طور جدی‌تری مطرح است. به همین دلیل، در این مقاله با هدف بررسی عوامل ممکن تاثیرگذار بر نرخ رخداد این نوع تصادف‌ها و تهیه نقشه پهنه‌بندی نقاط پرخطر، مدل‌بندی تعداد تصادف‌های منجر به جرح یا فوت در سرتاسر شهر مشهد مد نظر قرار گرفت.

با توجه به پیچیدگی مدل بتا-دوجمله‌ای فضایی پیشنهادی، در کنار چگال بودن ناحیه تحت مطالعه، از یک چارچوب بیزی تقریبی مبتنی بر روش INLA و ترکیب آن با رهیافت SPDE برای برازش مدل و انجام استنباط بیزی استفاده شد. نتایج حاصل از مطالعه شبیه‌سازی و مثال واقعی، بر اساس معیارهای نیکویی برازش و ارزیابی کفایت مدل، نشان از برتری مدل پیشنهادی نسبت به مدل معمول لوجیت-

دوجمله‌ای دارد. نقشه پهنه‌بندی نتیجه‌شده برای داده‌های تصادف نیز نقاط پرخطر را برای تصمیم‌سازان شهر مشهد نمایش می‌دهد. در عمل، در اغلب کاربردهای واقعی، بیش‌پراکنشی واقعیتی است که دیده می‌شود. از منظر نویسندگان واضح است که در حالتی که مساله بیش‌پراکنشی در داده‌ها وجود نداشته باشد، مدل دوجمله‌ای که نسبت به بتا-دوجمله‌ای ساده‌تر است، باید عملکرد کاراتری هم در برازش و هم پیشگویی فضایی داشته باشد. این نتیجه در مطالعه شبیه‌سازی به روشنی دیده شد.

تقدیر و تشکر

از همکاری معاونت حمل و نقل و ترافیک شهر مشهد برای در اختیار قرار دادن داده‌های تصادفات رانندگی تشکر می‌کنیم. همچنین از داوران و ویراستار محترم مجله برای نظرات سازنده آن‌ها که موجب ارتقای مقاله شدند، قدردانی می‌کنیم.

مراجع

عابدین‌پور، ل.، باغیشنی، ح. و اقبال، ن. (۱۳۹۸)، تحلیل بیزی سرطان معده در استان گیلان با مدل اتوبتا-دوجمله‌ای فضایی، *مجله مدل‌سازی پیشرفته ریاضی*، ۹، ۲۷-۴۳.

قلی‌زاده‌گزور، ک.، محمدزاده، م و قیومی، ز. (۱۳۹۲)، تحلیل فضایی رگرسیون جمعی ساختاری و مدل‌بندی داده‌های جرم شهر تهران با تقریب لاپلاس آشیانی جمع‌بسته، *مجله علوم آماری*، ۷، ۱۰۳-۱۲۴.

Bandyopadhyay, D., Reich, B. J. and Slate, E. H. (2011), A Spatial Beta-Binomial Model for Clustered Count Data on Dental Caries, *Statistical Methods in Medical Research*, **20**, 85-102.

Besag, J., York, J. and Mollie, A. (1991), Bayesian Image Restoration with Two Applications in Spatial Statistics (With Discussion), *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.

Bielby, J., Donnelly, C. A., Pope, L. C., Burke, T. and Woodroffe, R. (2014), Badger Responses to Small-Scale Culling May Compromise Targeted Con-

trol of Bovine Tuberculosis, *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 9193-9198.

Christensen, O. F., Roberts, G. O. and Skold, M. (2006), Robust Markov Chain Monte Carlo Methods for Spatial Generalized Linear Mixed Models, *Journal of Computational and Graphical Statistics*, **15**, 1-17.

Diggle, P. J. and Ribeiro, P. J. (2006), *Model-based Geostatistics*, Springer, New York.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013), *Regression Models, Methods and Applications*, Springer, Berlin.

Ferrari, S. and Cribari-Neto, F. (2004), Beta Regression for Modelling Rates and Proportions, *Journal of Applied Statistics*, **31**, 799-815.

Fuglstad, G., Simpson, D. P., Lingren, F. and Rue, H. (2019), Constructing Priors that Penalize the Complexity of Gaussian Random Fields, *Journal of the American Statistical Association*, **114**, 445-452.

Gelman, A., Hwang, J. and Vehtari, A. (2014), Understanding Predictive Information Criteria for Bayesian Models, *Statistics and Computing*, **24**, 997-1016.

Hinde, J. and Demetrio, C. (1998), Overdispersion: Models, and Estimation, *Computational Statistics and Data Analysis*, **27**, 151-170.

Hughes, G. and Madden, L. V. (1993), Using the Beta-Binomial Distribution to Describe Aggegated Patterns of Disease Incidence, *Phytopathology*, **83**, 759-763.

Kolovos, A., Smith, L. M., Schwab-McCoy, A., Gengler, S. and Yu, H. L. (2016), Emerging Patterns in Multi-Sourced Data Modeling Uncertainty, *Spatial Statistics*, **18**, 300-317.

Krainski, E. T., Gomez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D. P., Lindgren, F. and Rue, H. (2018), *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*, Chapman and Hall/CRC, New York.

Lang, S. and Brezger, A. (2004), Bayesian P-Splines, *Journal of Computational and Graphical Statistics*, **13**, 183-212.

Levin, B. and Reeds, J. (1977), Compound Multinomial Likelihood Functions are Unimodal: Proof of a Conjecture of I. J. Good, *The Annals of Statistics*, **5**, 79-87.

Lindgren, F. and Rue, H. (2015), Bayesian Spatial Modelling with R-INLA, *Journal of Statistical Software*, **63**, DOI: 10.18637/jss.v063.i19.

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), *Journal of the Royal Statistical Society, Series B*, **73**, 423-498.

Martins, T. G. and Rue, H. (2014), Extending Integrated Nested Laplace Approximation to a Class of Near-Gaussian Latent Models, *Scandinavian Journal of Statistics*, **41**, 893-912.

McCullach, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman, London.

Najera-Zuloaga, J., Lee, D. J. and Arostegui, I. (2019), A Beta-Binomial Mixed-Effects Model Approach for Analysing Longitudinal Discrete and Bounded Outcomes, *Biometrical Journal*, **61**, 600-615.

Pettit, L. I. (1990), The Conditional Predictive Ordinate for the Normal Distribution, *Journal of the Royal Statistical Society, Series B*, **52**, 175-184.

Richards, S. A. (2008), Dealing with Overdispersed Count Data in Applied Ecology, *Journal of Applied Ecology*, **45**, 218-227.

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

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations (With Discussion), *Journal of the Royal Statistical Society, Series B*, **71**, 319-392.

Schwab, A. D. and Marx, D. B. (2015), Beta-Binomial Kriging: An Improved Models for Spatial Rates, *Procedia Environmental Sciences*, **27**, 30-37.

Simpson, D. P., Rue, H., Riebler, A., Martins, T. G. and Sorbye, S. H. (2017), Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors, *Statistical Science*, **32**, 1-28.

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, Series B*, **10**, 257-261.

Spiegelhalter, D., Best, N., Carlin, B. and Van Der Linde, A. (2002), Bayesian Measures of Model Complexity and Fit, *Journal of the Royal Statistical Society, Series B*, **64**, 583-639.

Stanton, M. C. and Diggle, P. J. (2013), Geostatistical Analysis of Binomial Data: Generalised Linear or Transformed Gaussian Modelling?, *Environmetrics*, **24**, 158-171.

Stein, M. (2005), Space-Time Covariance Functions, *Journal of American Statistical Association*, **100**, 310-321.

Watanabe, S. (2010), Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory, *Journal of Machine Learning Research*, **11**, 3571-3594.

Whittle, P. (1963), Stochastic Processes in Several Dimensions, *Bulletin of the International Statistical Institute*, **40**, 974-994.

Williams, D. A. (1982), Extra-Binomial Variation in Logistic Linear Models, *Applied Statistics*, **31**, 144-148.

Williams, D. A. (1975), The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity, *Biometrics*, **31**, 949-952.

Journal of Statistical Sciences, Autumn and Winter, 2020
Vol. 14, No. 2, pp 307-334
DOI: 10.29252/jss.13.2.307

Bayesian Analysis of Spatial Count Data in Finite Populations Using Stochastic Partial Differential Equations

Eghbal, N., Baghishani, H.

Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran

Abstract: Geostatistical spatial count data in finite populations can be seen in many applications, such as urban management and medicine. The traditional model for analyzing these data is the spatial logit-binomial model. In the most applied situations, these data have overdispersion alongside the spatial variability. The binomial model is not the appropriate candidate to account for the overdispersion. The proper alternative is a beta-binomial model that has sufficient flexibility to account for the extra variability due to the possible overdispersion of counts. In this paper, we describe a Bayesian spatial beta-binomial for geostatistical count data by using a combination of the integrated nested Laplace approximation and the stochastic partial differential equations methods. We apply the methodology for analyzing the number of people injured/killed in car crashes in Mashhad, Iran. We further evaluate the performance of the model using a simulation study.

Keywords: Spatial beta-binomial, Overdispersion, Approximate Bayesian approach, Stochastic partial differential equations, Car crashes.

Mathematics Subject Classification (2010): 62H11, 62M30, 91B72.