مجله علوم آماری، بهار و تابستان ۱۳۹۲
جلد ۷، شماره ۱، ص ۱۰۳-۱۲۴

تحلیل فضایی رگرسیون جمعی ساختاری و مدل‌بندی داده‌های جرم شهر تهران با تقریب لاپلاس آشیانی جمع‌بسته

کبری قلی‌زاده گزور، محسن محمدزاده، زهرا قیومی
گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۱/۱۰/۲ تاریخ آخرین بازنگری: ۱۳۹۲/۶/۸

چکیده : در تحلیل بیزی مدل‌های رگرسیون جمعی ساختاری که قالبی انعطاف پذیر از مدل‌های آماری در زمینه‌های کاربردی دارند توزیع‌های پسینی فرم بسته‌ای ندارند و استفاده از الگوریتم‌های مونت کارلوی زنجیر مارکوفی به دلیل پیچیده بودن و تعداد زیاد پارامترهای این مدل زمان‌بر هستند. روش تقریب لاپلاس آشیانی جمع بسته می‌تواند با استفاده از تقریب‌های گاوسی و لاپلاس نیاز به شبیه‌سازی‌های سنگین را مرتفع سازد. در این مقاله نحوه لحاظ کردن همبستگی فضایی داده‌ها در مدل‌های رگرسیونی جمعی ساختاری و برآورد پارامترهای آن با تقریب لاپلاس آشیانی جمع‌بسته مورد مطالعه قرار می‌گیرند. سپس داده‌های جرم شهر تهران با این روش مدل‌بندی شده و در مطالعه‌ای شبیه‌سازی، دقت و سرعت محاسبهٔ مدل‌های حاصل از تقریب لاپلاس آشیانی جمع بسته و الگوریتم‌های مونت کارلوی زنجیر مارکوفی مورد ارزیابی و مقایسه قرار می‌گیرند.

آدرس الکترونیک مسئول مقاله: کبری قلی‌زاده، k.gholizadeh@modares.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲H۱۱

واژه‌های کلیدی : مدل رگرسیون جمعی ساختاری، تقریب لاپلاس آشیانی جمع‌بسته، الگوریتم‌های مونت کارلوی زنجیر مارکوفی.

۱ مقدمه

دگرگونی‌های عمده در حجم و پیچیدگی داده‌ها در کنار پیشرفت‌های اساسی روش‌های آماری و ارتقای تکنولوژی فناوری اطلاعات ما را به سمت استفاده از مدل‌های پیچیده اما دقیق‌تر سوق می‌دهد. معمولاً مدل‌های خطی با فرض پیوسته بودن متغیر پاسخ به کار می‌روند اما در عمل ممکن است متغیر پاسخ پیوسته نباشد به همین دلیل مدل خطی تعمیم‌یافته توسط نلدر و ودرنبرن (۱۹۷۲) معرفی شد. در هر دو مدل خطی و مدل‌های خطی تعمیم‌یافته فرض بر این است که متغیرهای تبیینی از طریق تابع پیوند اثر خطی بر متغیر پاسخ دارند در حالی که در بعضی مسائل ممکن است این اثر غیر خطی باشد. به منظور در نظر گرفتن توابعی هموار از این متغیرها، مدل جمعی تعمیم‌یافته (هیستی و تیبشرانی، ۱۹۹۰) معرفی شد. در مدل‌های خطی، خطی تعمیم‌یافته و جمعی تعمیم‌یافته فرض بر این است که متغیرهای پاسخ مستقل هستند، اما گاهی در عمل این فرض غیرواقع گرایانه است و متغیرهای پاسخ وابسته‌اند. بریسلو و کلیتون (۱۹۹۳) با اضافه کردن اثرات تصادفی و پذیرش فرض استقلال شرطی متغیر پاسخ در مدل‌های خطی تعمیم‌یافته، مدل آمیخته خطی تعمیم‌یافته را معرفی کردند. علاوه بر این لین و ژانگ (۱۹۹۹) مدل آمیخته جمعی تعمیم‌یافته را به‌عنوان تعمیمی از مدل‌های خطی تعمیم‌یافته برای داده‌هایی با متغیر پاسخ وابسته معرفی کردند، به‌طوری که این وابستگی از طریق یک متغیر تبیینی در مدل لحاظ می‌شود. فهرامیر و تاتز (۲۰۰۱) مدل کلی رگرسیون جمعی ساختاری[۱] (STAR) را معرفی کردند، که در آن متغیر پاسخ متعلق به خانواده نمایی است و متغیرهایی تبیینی با اثرات خطی و غیر خطی نیز در مدل لحاظ می‌شوند. به‌علاوه ناهمگونی متغیرهای کمکی می‌تواند در مدل مورد توجه قرار گیرد و فرض استقلال شرطی متغیر پاسخ جایگزین فرض استقلال شود. هر یک از

[۱] Structured additive regression model

مدل‌های ذکر شده زیر رده‌ای از مدل‌های رگرسیون جمعی ساختاری هستند.

برای تحلیل این مدل‌ها با رهیافت بیزی نیاز به محاسبهٔ توزیع‌های پسینی است و معمولاً الگوریتم‌های مونت کارلوی زنجیر مارکوفی[۲] (MCMC) می‌توانند روشی مفید در به‌دست آوردن آن‌ها و تحلیل بیزی مدل‌ها باشند. اما در عمل این الگوریتم‌ها به دلیل پیچیده بودن مدل‌های STAR ممکن است با مشکلی همچون طولانی بودن زمان محاسبات مواجه شوند که برای حل این مشکل رو و همکاران (۲۰۰۹) روش تقریب لاپلاس آشیانی جمع بسته[۳] (INLA) را معرفی کردند و نشان دادند این روش ضمن حفظ دقت برآورد پارامترها سرعت محاسبات را نیز افزایش می‌دهد. فانگ و همکاران (۲۰۱۰) نشان دادند روش INLA برای مدل‌های آمیخته خطی تعمیم‌یافته در حالت کلی دقیق است، اما برای داده‌های دوجمله‌ای با تعداد آزمایشات کم، دقت کمتری دارد. رز و هلد (۲۰۱۱) با به کاربردن روش INLA به بررسی حساسیت مدل‌های آمیخته خطی تعمیم‌یافته به پیشینی‌های فرض شدهٔ ابرپارمترها پرداختند و بر اساس فاصله هلینگر[۴] که شباهت بین دو توزیع احتمال را می‌سنجد، اندازه حساسیت را توسعه دادند، همچنین برای انتخاب مدل چندین تکنیک اعتبار سنجی متقابل پیشنهاد دادند. اخیراً اسکلر و هلد (۲۰۱۱) چگونگی استفاده از INLA برای استنباط بیزی انواع مدل‌های فضایی- زمانی را نشان دادند و برای داده‌های اسهال ویروسی گاو در سوئیس دو روش INLA و MCMC را از نظر دقت با هم مقایسه کردند. قیومی و همکاران (۱۳۹۱) نحوهٔ کاربست INLA را در مدل‌های گاوسی پنهان فضایی برای تحلیل داده‌های میزان ضعف بدنی کودکان در کشور زامبیا به نمایش گذاشته و دقت نتایج حاصل از دو روش INLA و الگوریتم‌های MCMC را مورد ارزیابی قرار دادند.

در این مقاله برای تحلیل بیزی مدل‌های STAR و یافتن چگالی‌های پسینی کناری از روش INLA استفاده می‌شود. برای این منظور در بخش ۲ مدل‌های رگرسیونی جمعی ساختاری معرفی می‌شوند. سپس در بخش ۳ میدان تصادفی

---

۲ Markov Chain Monte Carlo

۳ Integrated Nested Laplace Approximation

۴ Hellinger distance

مارکوفی گاوسی و خصوصیات آن به اختصار بیان می‌شوند. روش INLA و ویژگی‌های آن در بخش ۴ ارائه می‌شوند. در بخش ۵ ملاک‌هایی برای ارزیابی و مقایسهٔ مدل‌ها بیان شده و در بخش ۶ نحوهٔ کاربست روش INLA برای مدل‌بندی و تحلیل داده‌های جرم مناطق ۲۲گانه شهر تهران ارائه شده و در انتها به بحث و نتیجه‌گیری پرداخته می‌شود.

۲ مدل‌های رگرسیون جمعی ساختاری

مدل‌های STAR قالبی انعطاف پذیر برای مدل‌بندی اثرات غیر خطی متغیرهای تبیینی شامل مدل‌های خطی تعمیم‌یافته و مدل‌های جمعی تعمیم‌یافته هستند. توزیع متغیر پاسخ y_i، $i = ۱,\ldots,n$ متعلق به خانوادهٔ نمایی با تابع چگالی یا جرم احتمال به‌صورت

$$\begin{aligned} f(y_i \mid \boldsymbol{u}_i, \boldsymbol{\beta}) &= \exp\{y_i(\alpha + \sum_{j=۱}^{n_f} \boldsymbol{f}^{(j)}(u_{ji}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki}) \\ &- a(\alpha + \sum_{j=۱}^{n_f} \boldsymbol{f}^{(j)}(u_{ji}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki}) + C(y_i)\}, \qquad (۱) \end{aligned}$$

است، که در آن $\{\beta_k\}$ اثرات ثابت خطی از مولفه‌های متغیرهای تبیینی $\boldsymbol{z}_i = (z_{۱i},\ldots,z_{n_\beta i})$ و $\{\boldsymbol{f}^{(j)}(\cdot)\}$ توابعی از متغیرهای تبیینی $\boldsymbol{u}_i = (u_{۱i},\ldots,u_{n_f i})$ هستند. در این مدل $\eta_i = \alpha + \sum_{j=۱}^{n_f} \boldsymbol{f}^{(j)}(u_{ji}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki}$ پارامترهای کانونی در خانوادهٔ نمایی محسوب می‌شوند. اگر $\mu_i = E[y_i \mid \boldsymbol{u}_i, \boldsymbol{\beta}] = a'(\eta_i)$، آن‌گاه $g(\mu_i) = \eta_i = \alpha + \sum_{j=۱}^{n_f} \boldsymbol{f}^{(j)}(u_{ji}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki}$، که در آن $g(\cdot)$ تابع پیوند است. از آنجا که توابع $\{\boldsymbol{f}^{(j)}(\cdot)\}$ می‌توانند شامل اثرات غیر خطی مانند روندهای زمانی، اثرات فصلی، شیب‌ها و عرض از مبدأهای تصادفی در مدل باشند، در ادامه همبستگی فضایی داده‌ها از طریق این توابع در مدل STAR منظور خواهد شد.

فرض کنید میدان تصادفی پنهان $\boldsymbol{x} = \{\alpha, f^{(۱)},\ldots,f^{(n_f)}, \beta_۱,\ldots,\beta_{n_\beta}, \eta_i\}$ دارای توزیع نرمال چندمتغیره با بردار میانگین صفر و ماتریس دقت $\boldsymbol{Q}_{\boldsymbol{\theta}_۱} = \boldsymbol{\Sigma}_{\theta_۱}^{-۱}$ باشد، که در آن $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_۱}$ ماتریس کواریانس است که به پارامتر $\boldsymbol{\theta}_۱$ بستگی دارد، به این مدل که زیر رده‌ای از مدل‌های STAR است مدل گاوسی پنهان گویند. معمولا ناحیه

مورد مطالعه چگال را می‌توان با استفاده از معادلات دیفرانسیل تصادفی جزئی[۵] به گونه‌ای مشبکه‌ای نمود که روش INLA قابل اجرا باشد (لیندگرین و همکاان، ۲۰۱۱). اما در اینجا با فرض آن که $\mathcal{I}$ یک مشبکه با n_d گره و بردار متغیر پاسخ $\{y_i; i \in \mathcal{I}\} : \boldsymbol{y}$ دارای توزیع $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_2)$ باشد، که درایه‌های آن به شرط $\boldsymbol{x}$ و $\boldsymbol{\theta}_2$ مستقل شرطی‌اند، آن‌گاه چگالی پسینی میدان تصادفی پنهان $\boldsymbol{x}$ و $\boldsymbol{\theta}$ به صورت

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \theta_1) \prod_{i \in \mathcal{I}} \pi(y_i \mid x_i, \theta_2) \\ &\propto \pi(\boldsymbol{\theta})|\boldsymbol{Q}_{\theta_1}|^{\frac{1}{2}} \exp\{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}_{\theta_1}\boldsymbol{x} + \sum_{i \in \mathcal{I}} \log\{\pi(y_i \mid x_i, \theta_2)\}\},\end{aligned}$$

خواهد بود، که در آن $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ بردار m بعدی ابر پارامترهای مدل است.

اغلب مدل‌های گاوسی پنهان دارای دو ویژگی اساسی هستند. اول این‌که میدان تصادفی پنهان $\boldsymbol{x}$، که اغلب دارای بعدی بزرگی است، در ویژگی‌های مارکوفی دو به دو، مارکوفی موضعی و مارکوفی فراموضعی (رو و هلد، ۲۰۰۵) صدق می‌کنند، بنابراین میدان تصادفی پنهان یک میدان تصادفی مارکوفی گاوسی[۶] (GMRF) با یک ماتریس دقت تنک $\boldsymbol{Q}_{\boldsymbol{\theta}}$ است. به همین دلیل می‌توان از روش‌های عددی مربوط به ماتریس‌های تنک که بسیار سریع‌تر از روش‌های محاسباتی ماتریس‌های چگال هستند استفاد ه کرد. ویژگی دوم این است که تعداد ابرپارامترها مقداری کوچک مانند $m \leq 6$ است که شرایط لازم برای استنباط‌های سریع را فراهم می‌سازند (ادزویک و همکاران، ۲۰۰۹).

۳ میدان تصادفی مارکوفی گاوسی

میدان تصادفی مارکوفی گاوسی براساس گراف قابل تعریف است. گراف $\mathcal{G}$ مجموعه‌ای از راس‌ها است که توسط خانواده‌ای از یال‌ها به هم وصل شده‌اند و به صورت زوج مرتب $(\mathcal{V}, \mathcal{E})$ نشان داده می‌شود، که در آن $\mathcal{V}$ مجموعه‌ای متناهی و غیر تهی از رئوس و $\mathcal{E}$ یال‌های آن است. اگر راس‌ها به صورت $\mathcal{V} = \{1, \ldots, n\}$ باشند

[۵] Stochastic partial differential equation

[۶] Gaussian Markov random field

گراف را نشاندار گویند.

فرض کنید بردار تصادفی $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ دارای توزیع نرمال چندمتغیرهٔ با میانگین $\boldsymbol{\mu}$ و ماتریس کواریانس $\boldsymbol{\Sigma}$ باشد. به‌علاوه $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ گرافی نشان‌دار باشد، که در آن $\mathcal{E}$ شامل همهٔ زوج‌های $\{i, j\}$ است به طوری که راس‌های i و j هیچ یال مشترکی نداشته باشند، اگر و تنها اگر $x_i \perp x_j | \boldsymbol{x}_{-ij}$، در این‌صورت $\boldsymbol{x}$ یک میدان تصادفی مارکوفی گاوسی نسبت به گراف $\mathcal{G}$ نامیده می‌شود.

قضیه ۱ (رو و هلد، ۲۰۰۵): اگر $\boldsymbol{x}$ دارای توزیع نرمال چندمتغیره با میانگین $\boldsymbol{\mu}$ و ماتریس دقت $\boldsymbol{Q} > \mathbf{0}$ باشد، آن‌گاه درایهٔ ij ام ماتریس $\boldsymbol{Q}$، یعنی Q_{ij} برابر صفر است، اگر و تنها اگر $x_i \perp x_j \mid \boldsymbol{x}_{-ij}$.

با توجه به قضیهٔ ۱ درایه‌های غیر صفر ماتریس $\boldsymbol{Q}$ گراف $\mathcal{G}$ را مشخص می‌کنند و بر اساس آن‌ها می‌توان استقلال شرطی x_i و x_j را بررسی نمود. بنابراین در اینجا از ماتریس دقت به جای ماتریس کواریانس استفاده می‌شود. بنابراین، بردار تصادفی $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in R^n$ یک میدان تصادفی مارکوفی گاوسی تحت گراف نشاندار $G = (\mathrm{V}, \mathcal{E})$ با میانگین $\boldsymbol{\mu}$ و ماتریس دقت $\boldsymbol{Q}_{n\times n} > \mathbf{0}$ است اگر و تنها اگر تابع چگالی آن به‌صورت

$$\pi(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{Q}|^{\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{\mu})\},$$

باشد، به‌گونه‌ای که درایهٔ (ij)ام ماتریس دقت، یعنی $\boldsymbol{Q}_{ij}$، مخالف صفر است اگر و تنها اگر $\{i, j\} \in \mathcal{E}$. در بیشتر موارد $\boldsymbol{Q}$ ماتریسی تنک است و تنها $\mathcal{O}(n)$ درایه از n^2 درایهٔ ماتریس $\boldsymbol{Q}$ غیر صفر است (رو و مارتینو، ۲۰۰۹)، که این به علت ویژگی مارکوفی میدان تصادفی است. تنک بودن این ماتریس باعث افزایش سرعت در محاسبات و الگوریتم‌های تکراری مانند نمونه‌گیری از میدان تصادفی مارکوفی می‌شود.

اگر ماتریس دقت $\boldsymbol{Q}$ نیمه معین مثبت متقارن با مرتبهٔ $n - k > 0$ باشد، آن‌گاه $\boldsymbol{x}$ یک GMRF ناسره از مرتبهٔ $n - k$ با پارامترهای $(\boldsymbol{\mu}, \boldsymbol{Q})$ نامیده می‌شود چنانچه

چگالی آن به صورت

$$\pi(\boldsymbol{x}) = (۲\pi)^{\frac{-(n-k)}{۲}} (|\boldsymbol{Q}|^*)^{\frac{۱}{۲}} \exp\{-\frac{۱}{۲}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\},$$

باشد، که در آن نماد $|\cdot|^*$ بیانگر دترمینان تعمیم‌یافته است (رو و هلد، ۲۰۰۵). نوع خاصی از GMRF، میدان تصادفی مارکوفی گاوسی ذاتی[۷] (IGMRF) است که ناسره هستند، یعنی ماتریس دقتشان پررتبه نیست. یک میدان تصادفی مارکوفی گاوسی ذاتی از مرتبهٔ k، یک GMRF ناسره با رتبهٔ n-k با ویژگی $\boldsymbol{Q}\boldsymbol{S}_{k-۱} = \beta ۰$ است، که در آن $\boldsymbol{S}_{k-۱}$ ماتریس طرح چندجمله‌ای است (رو و هلد، ۲۰۰۵). این نوع از میدان‌ها کاربرد وسیعی در مدل‌بندی اثرات هموار فضایی دارند که از جمله می‌توان به مدل‌های قدم زدن تصادفی اشاره کرد. میدان تصادفی مارکوفی گاوسی ذاتی مرتبهٔ اول، $\boldsymbol{x}$ یک مدل قدم زدن تصادفی مرتبهٔ اول [۸](RW1) نامیده می‌شود هرگاه نموهای مرتبهٔ اول آن مستقل و دارای توزیع نرمال به‌صورت

$$\Delta x_i = x_{i+۱} - x_i \overset{iid}{\sim} \mathcal{N}(۰, \kappa^{-۱}), \qquad i = ۱, \cdots, n-۱,$$

باشند. چنانچه مشاهدات برای این مدل بر یک مشبکه واقع شده باشند همان مدل بسیج[۹] (بسیج، ۱۹۷۴) است.

۴ تقریب لاپلاس آشیانی جمع‌بسته

برای تحلیل بیزی مدل‌های گاوسی پنهان لازم است توزیع‌های پسینی کناری متغیرهای پنهان و ابرپارامترها به‌صورت

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{y},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \qquad i = ۱, \ldots, n_d, \tag{۲}$$

$$\pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}, \qquad j = ۱, \ldots, \ell, \tag{۳}$$

محاسبه شوند، که در آن $\boldsymbol{\theta}_{-j}$ بردار حاصل از حذف درایهٔ jام $\boldsymbol{\theta}$ است. تقریب لاپلاس آشیانی جمع‌بسته تقریب‌هایی برای چگالی‌های پسینی کناری (۲) و (۳)

[۷] Intrinsic GMRF

[۸] First-order Random Walk

[۹] Besag model

به‌صورت

$$\widetilde{\pi}(x_i|\boldsymbol{y}) = \int \widetilde{\pi}(x_i|\boldsymbol{y},\boldsymbol{\theta})\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \quad i = ۱,\ldots,n_d, \qquad (۴)$$

$$\widetilde{\pi}(\theta_j|\boldsymbol{y}) = \int \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}, \quad j = ۱,\ldots,\ell, \qquad (۵)$$

فراهم می‌کند. برای محاسبه تقریبی این توزیع‌ها، روش INLA با استفاده از تبدیل‌های لاپلاس و انتگرال‌گیری عددی محاسباتی سریع و تقریبی دقیق را جایگزین شبیه‌سازی‌های سنگین الگوریتم‌های MCMC می‌کند. تقریب‌های $\widetilde{\pi}(\theta_j|\boldsymbol{y})$ و $\widetilde{\pi}(x_i|\boldsymbol{y})$ در سه گام به شرح زیر محاسبه می‌شوند.

گام ۱: برای توزیع شرطی کامل

$$\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \propto \exp\{-\frac{۱}{۲}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in\mathcal{I}} \log\pi(y_i|x_i,\boldsymbol{\theta})\},$$

با استفاده از بسط تیلور $\log\pi(y_i|x_i,\boldsymbol{\theta})$ حول مد توزیع شرطی کامل $\boldsymbol{x}$، یعنی $\boldsymbol{x}^*(\theta) = (x^*_۱(\boldsymbol{\theta}),\ldots,x^*_{nd}(\boldsymbol{\theta}))$، تقریب گاوسی[۱۰] به صورت

$$\begin{aligned}\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) &\propto \exp\{-\frac{۱}{۲}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in\mathcal{I}} g_i(x_i)\}\\ &\propto \exp\{-\frac{۱}{۲}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in\mathcal{I}}(a_i + b_i x_i - \frac{۱}{۲}c_i x_i^۲)\}\\ &\propto \exp\{-\frac{۱}{۲}\boldsymbol{x}^T(\boldsymbol{Q} + diag(\boldsymbol{c}))\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\},\end{aligned}$$

به‌دست آورده می‌شود، که در آن $c_i = -g''(x_i^*(\boldsymbol{\theta}))$ و $b_i = g_i'(x_i^*(\boldsymbol{\theta})) + x_i^*(\boldsymbol{\theta})c_i$ است و a_i از تناسب حذف می‌شود (رو و هلد، ۲۰۰۵). سپس تقریب لاپلاس چگالی پسینی $\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}$، با بسط مخرج کسر حول $\boldsymbol{x} = \boldsymbol{x}^*(\boldsymbol{\theta})$ به صورت

$$\widetilde{\pi}_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})}{\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\,|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})},$$

محاسبه می‌شود.

گام ۲: تقریب لاپلاس چگالی شرطی $\pi(x_i|\boldsymbol{\theta},y) = \frac{\pi(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})}{\pi(\boldsymbol{x}_{-i}|x_i,\boldsymbol{\theta},\boldsymbol{y})}$ با بسط مخرج

[۱۰] Gaussian approximation

کسر حول $\boldsymbol{x}_{-i} = \boldsymbol{x}^*_{-i}(x_i, \boldsymbol{\theta})$ به‌صورت

$$\widetilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\widetilde{\pi}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})}|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^*_{-i}(x_i,\boldsymbol{\theta})},$$

محاسبه می‌شود، که در آن $\boldsymbol{x}^*_{-i}(\boldsymbol{x}_i, \theta)$ مد توزیع $\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ و $\widetilde{\pi}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ تقریب گاوسی $\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ است.

گام ۳: انتگرال‌های (۴) و (۵) به‌صورت جمع متناهی

$$\widetilde{\pi}(x_i|\boldsymbol{y}) = \sum_{k=1}^{n_{p1}} \widetilde{\pi}(x_i|\boldsymbol{\theta}_k, \boldsymbol{y})\widetilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y})\Delta_k, \quad i = 1, \ldots, n_d, \qquad (۶)$$

$$\widetilde{\pi}(\theta_j|\boldsymbol{y}) = \sum_{q=1}^{n_{p2}} \widetilde{\pi}(\boldsymbol{\theta}_q|\boldsymbol{y})\Delta_{jq}, \quad j = 1, \ldots, \ell, \qquad (۷)$$

تقریب زده می‌شوند، که در آن n_{p1} و n_{p2} تعداد $\boldsymbol{\theta}$های انتخاب شده از روی تکیه‌گاه و وزن‌های Δ_k و Δ_{jq} به روش طرح مرکب مرکزی (رو و مارتینو، ۲۰۰۹) به‌صورت

$$\Delta_k = \{(n_{p1})(f_0^2 - 1)[1 + \exp(-\frac{mf_0^2}{2})]\}^{-1},$$

$$\Delta_j q = \{(n_{p2})(f_0^2 - 1)[1 + \exp(-\frac{(m-1)h_0^2}{2})]\}^{-1},$$

محاسبه می‌شوند، که در آن m بُعد ابرپارامتر $\boldsymbol{\theta}$ و $f_0, h_0 > 1$ مقدارهایی ثابت هستند.

۵ ارزیابی مدل‌ها

ملاک اطلاع انحراف[۱۱] (DIC) اندازه‌ای از پیچیدگی و برازش مدل است که برای مقایسهٔ مدل‌های پیچیده استفاده می‌شود (اشپیگل‌هالتر و همکاران، ۲۰۰۲). با تعریف ملاک انحراف بیزی بر اساس لگاریتم درستنمایی به صورت $D(X, \theta) = -2\sum_{i\in\mathcal{I}} \log\{\pi(y_i|x_i, \theta)\} + c$ که در آن c مقداری ثابت است (دمپستر، ۱۹۷۴)، ملاک DIC به صورت $DIC = \overline{D} + p_D$ تعریف می‌شود، که در آن $\overline{D} = E_{\theta|y}(D)$ میانگین پسین انحراف‌ها و p_D تفاضل میانگین انحراف‌ها و انحراف میانگین‌ها است، که میزان پیچیدگی مدل را بیان می‌کند.

[۱۱] Deviance Information Criterion

برای مقایسه مدل‌ها، از نظر اعتبار پیشگویی y_i می‌توان از ملاک پیشگویی شرطی مؤلفه‌ها[۱۲] (CPO) به‌صورت $CPO_i = \pi(y_i|\boldsymbol{y}_{-i})$ استفاده کرد (پتیت، ۱۹۹۰). مقادیر صفر و خیلی کوچک CPO بیانگر پرت بودن y_i متناظر است. در این مقاله از منهای میانگین لگاریتم CPO_iها برای مقایسهٔ مدل‌ها استفاده می‌شود، که ملاک نمرهٔ لگاریتمی اعتبار سنجی متقابل[۱۳](LogScore) (گنتینگ و رفتری، ۲۰۰۷) نامیده می‌شود. با توجه به اینکه مقادیر بزرگ CPO_iها گویای پیشگویی برتر مدل هستند، بنابراین هرچه مقدار LogScore کوچک‌تر باشد مدل برای پیشگویی مناسب‌تر است.

برای مقایسهٔ تقریب توزیع‌های پسینی پارامترها با توزیع اصلی از ملاک فاصله کولبک–لیب‌لر[۱۴] (KLD) استفاده می‌شود. KLD اندازهٔ فاصلهٔ بین دو توزیع واقعی و توزیع تقریب زده شده به دست می‌دهد بنابراین هرچه این مقدار کوچکتر باشد تقریب دقیق‌تر است (کولبک و لیبلر، ۱۹۵۱) .

۶ مدل‌بندی نرخ جرم در شهر تهران

در این بخش داده‌های جرم «نزاع و درگیری» و«سرقت» مناطق ۲۲گانه شهر تهران با استفاده از مدل‌های STAR و برآورد آن با INLA مورد تحلیل قرار می‌گیرند. داده‌ها شامل نرخ جرم مناطق به عنوان متغیر پاسخ و نرخ عوامل اجتماعی مانند نرخ طلاق، نرخ جرم مواد مخدر، مهاجرت، جمعیت شناور و سرانهٔ فضای سبز در هر منطقه به‌عنوان متغیرهای تبیینی هستند. فرض بتا بودن توزیع متغیر پاسخ برای هر دو جرم «نزاع و درگیری» و «سرقت»، با به‌ترتیب p– با مقدارهای ۰/۴۳ و ۰/۳۴ آزمون کولموگروف–اسمیرنف مورد تایید قرار گرفت. فرض وجود همبستگی فضایی با به کاربردن آزمون I– موران (موران، ۱۹۵۰) و باتوجه به این نکته که فرض صفر در این آزمون عدم وجود همبستگی فضایی است، با p– مقدار ۰/۰۴ و $۳/۲۵ \times ۱۰^{-۴}$ به‌ترتیب برای نرخ جرم نزاع و درگیری و سرقت مورد تایید قرار گرفت. بنابراین همبستگی فضایی داده‌ها نیز در مدل STAR لحاظ می‌شود.

---

[۱۲] Conditional Predictive Ordinates

[۱۳] Cross Validated Logarithmic Score

[۱۴] Kullback Leibler Distance

جرم نزاع و درگیری: فرض کنید y_i، نرخ نزاع و درگیری در هر منطقه، دارای توزیع بتا به‌صورت

$$y_i|\mu_i \sim \mathcal{B}(\alpha_i, \beta_i), \qquad i = ۱, \ldots, ۲۲, \tag{۸}$$

است، که در آن $\mu_i = \frac{\alpha_i}{\alpha_i+\beta_i}$. با فرض این‌که $\phi = \alpha_i + \beta_i$ مقداری ثابت است، میانگین و واریانس y_i به صورت

$$E(y_i) = \mu_i, \qquad Var(y_i) = \frac{\mu_i(۱ - \mu_i)}{۱ + \phi},$$

به‌دست آورده می‌شوند، که در آن ϕ به‌عنوان پارامتر دقت شناخته شده است. در اینجا برای مرتبط کردن میانگین با پیشگوی جمعی ساختاری از تابع پیوند لجیت به‌صورت $\mu_i = \frac{\exp(\eta_i)}{۱+\exp(\eta_i)}$ استفاده شده است. برای بررسی تاثیر متغیرهای تبیینی معرفی شده بر نرخ نزاع و درگیری، سه مدل مختلف با روش INLA تحلیل و مورد مقایسه قرار می‌گیرند.

مدل ۱: اولین مدل برای η_i به‌صورت

$$\eta_i = \boldsymbol{\beta}_۱ \boldsymbol{Z}^T + f_S(s_i) + f_R(s_i) \qquad i = ۱, \ldots, ۲۲,$$

در نظر گرفته شده است، که در آن $\boldsymbol{\beta}_۱ = (\beta_۰, \beta_D, \beta_N, \beta_G)$ بردار ضرایب اثرات ثابت به‌ترتیب مربوط به عرض از مبدا، نرخ مواد مخدر، سرانهٔ فضای سبز و نرخ طلاق است، همچنین $f_R(\cdot)$ و $f_S(\cdot)$ توابعی نامعلوم برای لحاظ کردن به‌ترتیب اثر تصادفی و همبستگی فضایی بین مناطق هستند. در مدل ۱ اثر همهٔ متغیرهای تبیینی $\boldsymbol{Z}$ خطی در نظر گرفته شده است، برای هر یک از ضرایب آن‌ها مطابق معمول منابع توزیع پیشینی نرمال با میانگین صفر و واریانس نسبتا بزرگ ۱۰۰۰ اختیار شده است. به دلیل عدم اطلاع از نوع ساختار همبستگی فضایی مناطق، این پدیده به‌صورت متغیری پنهان با مدل پیشینی بسیج با دقت نامعلوم τ_S در نظر گرفته شده است، تا بتوان ساختار همبستگی فضایی همسایگی‌ها را با استفاده از گراف در مدل لحاظ نمود. همین‌طور توزیع پیشینی نرمال با میانگین صفر و دقت نامعلوم τ_R برای اثرات تصادفی به جز ساختار همبستگی فضایی مناطق فرض می‌شود. میدان تصادفی

پنهان در این مدل $\boldsymbol{x} = \{f_S(\cdot), f_R(\cdot), \beta_0, \beta_D, \beta_N, \beta_G, \eta_i\}$ و بردار ابر پارامترها $\boldsymbol{\theta} = (\phi, \tau_S, \tau_R)^T$ است. برای هر یک از درایه‌های بردار $\boldsymbol{\theta}$ توزیع پیشینی گامای مبهم به ترتیب با پارامترهای $(۱, ۱۰^{-۶})$، $(۱, ۱۰^{-۷})$ و $(۰/۱, ۱۰^{-۵})$ در نظر گرفته شده است.

برآورد پارامترهای این مدل برای اثرهای ثابت به همراه انحراف استاندارد برآورد، صدک‌های ۰/۰۲۵ و ۰/۹۷۵ و ملاک KLD برای هر یک از کناری‌های پسینی در جدول ۱ ارائه شده است. باتوجه به برآورد ضرایب ملاحظه می‌شود که

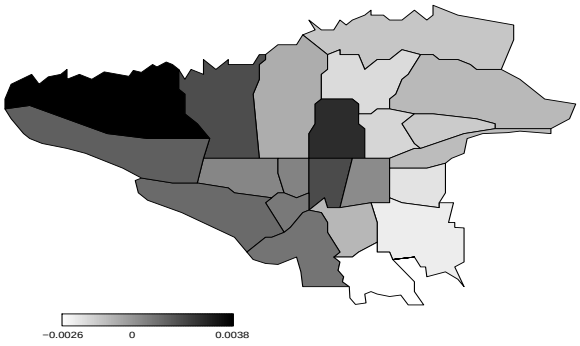
جدول ۱: برآورد ضرایب رگرسیونی و ملاک‌های ارزیابی مدل ۱ نرخ جرم نزاع و درگیری

		انحراف	صدک‌ها		
ضریب	برآورد	استاندارد	۰/۰۲۵	۰/۹۷۵	KLD
عرض از مبدا	-۳/۱۲۷	۰/۴۰۷	-۳/۹۰۷	-۲/۳۰۲	$۱۰^{-۳۱}$
نرخ طلاق	۲۹/۰۹۰	۱۸/۱۳۱	-۷/۷۶۷	۶۳/۶۰۷	$< ۱۰^{-۳۱}$
نرخ جرم مواد مخدر	۲۰/۹۵۷	۹/۹۷۲	۰/۰۹۴	۳۹/۴۹۵	$< ۱۰^{-۳۱}$
سرانهٔ فضای سبز	۰/۰۳۲	۰/۰۶۱	-۰/۰۹۲	۰/۱۴۸	$< ۱۰^{-۳۱}$

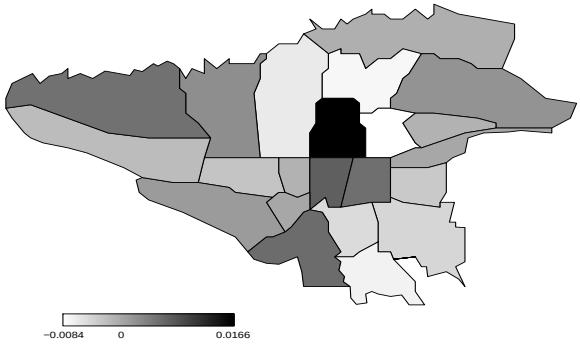
طلاق بیش از مواد مخدر و سرانهٔ سبز در بروز جرم اثرگذار است. به‌علاوه مقادیر بسیار کوچک KLD می‌تواند بیانگر فاصله کم توزیع پسینی برآورد شده از توزیع واقعی باشد. شکل‌های ۱ (الف) و ۱ (ب) به‌ترتیب اثر فضایی و تصادفی برآوردشده مناطق را نمایش می‌دهند. همان‌طور که در شکل ۱(الف) ملاحظه می‌شود مناطق مجاور هم دارای اثرات مشابهی بر وقوع جرم هستند که گویای فضایی بودن این اثر است.

مدل ۲: شکل ۲ نحوهٔ پراکندگی نرخ طلاق در شهر تهران را نشان می‌دهد. همان‌طور که ملاحظه می‌شود نرخ طلاق در مناطق ۲۲گانه دارای اثر فضایی هستند، زیرا مناطقی که در مجاورت هم قرار دارند رنگ‌هایی مشابه دارند یعنی نرخ طلاق در مناطق همسایه به یکدیگر وابسته‌اند. به همین دلیل برای بررسی فرض غیر خطی بودن اثر این متغیر بر نرخ جرم مدل دوم به‌صورت

$$\eta_i = \beta_۲ \boldsymbol{Z}^T + f_S(s_i) + f_R(s_i) + f_D(d_i) \qquad i = ۱, \ldots, ۲۲,$$

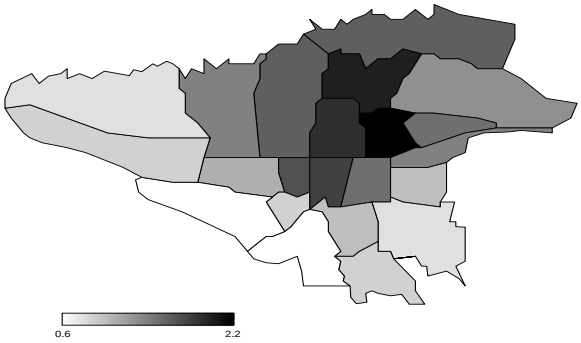


(الف)



(ب)

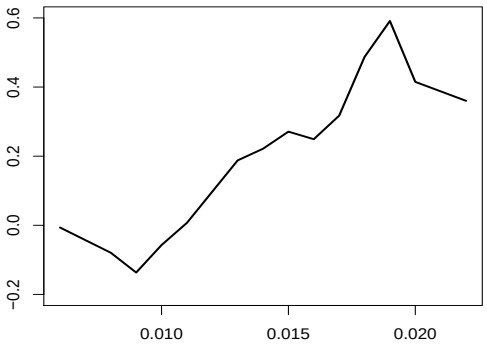
شکل ۱: پهنه‌بندی اثرات (الف) فضایی و (ب) تصادفی مناطق ۲۲گانه شهر تهران



شکل ۲: پهنه‌بندی نرخ طلاق مناطق ۲۲گانه شهر تهران

در نظر گرفته شده است، که در آن $\boldsymbol{\beta}_۲ = (\beta_۰, \beta_N, \beta_G)$ است. از آنجا که تعداد مناطق محدود است و با توجه به شکل ۲ به نظر می‌رسد نرخ طلاق در هر منطقه تنها به همسایگی‌های مرتبه یک وابسته است، به‌علاوه نرخ طلاق به‌صورت متغیر تصادفی پنهان با مدل RW1 و دقت نامعلوم τ_D فرض شده است. برای ابرپارامتر τ_D نیز توزیع پیشینی گامای مبهم با پارامترهای (۰/۰۰۰۱, ۰/۰۰۰۴) درنظر گرفته شده است. شکل ۳ به خوبی اثر غیر خطی طلاق بر نزاع و درگیری را در مدل ۲ نشان می‌دهد.

همان طور که در جدول ۴ ملاحظه می‌شود ملاک DIC برای مدل ۲ نسبت به مدل ۱ کاهش زیادی یافته است که بیانگر بهبود چشم‌گیر مدل به دلیل غیرخطی در نظر گرفتن متغیر طلاق در مدل ۲ است.



شکل ۳: نمودار تاثیر نرخ طلاق (محور افقی) بر نرخ نزاع و درگیری (محور عمودی)

با توجه به برآورد ضرایب در جدول ۲، سرانهٔ فضای سبز در این مدل مثبت اما کوچک و برابر ۰/۰۵۱ است، که به نظر می‌رسد تاثیر چندانی بر مدل نداشته باشد.

مدل ۳:

با حذف اثر سرانهٔ فضای سبز، مدل جدید به‌صورت

$$\eta_i = \beta_۰ + \beta_N Z_i + f_S(s_i) + f_R(s_i) + f_D(d_i) \quad i = ۱, \ldots, ۲۲,$$

در نظر گرفته شده و برآورد ضرایب آن در جدول ۳ ارائه گردیده است. با توجه به جدول ۴، بعد از حذف متغیر سرانهٔ سبز ، مقادیر p_D، DIC و LogScore افزایش

جدول ۲: برآورد ضرایب و ملاک‌های ارزیابی مدل ۲ نرخ جرم نزاع و درگیری

		انحراف	صدک‌ها		
ضریب	برآورد	استاندارد	۰/۰۲۵	۰/۹۷۵	KLD
عرض از مبدا	−۲/۸۷۶	۰/۱۰۴	−۳/۰۷۹	−۲/۶۶۳	$< ۱۰^{-۱۶}$
نرخ جرم مواد مخدر	۲۴/۴۴۱	۱۵/۱۸۴	۲۴/۵۱۵	۳۳/۲۵۴	$< ۱۰^{-۱۶}$
سرانهٔ فضای سبز	۰/۰۵۱	۰/۰۲۹	−۰/۰۰۷	۰/۱۰۸	$< ۱۰^{-۱۶}$

یافته است. بنابراین مدل ۲ بر مدل ۳ نیز ارجحیت داده می‌شود.

جدول ۳: برآورد ضرایب رگرسیونی و ملاک‌های ارزیابی مدل ۳ نرخ جرم نزاع و درگیری

		انحراف	صدک‌ها		
ضریب	برآورد	استاندارد	۰/۰۲۵	۰/۹۷۵	KLD
عرض از مبدا	−۲/۷۶۵	۰/۰۸۸	−۲/۹۳۲	−۲/۵۸۲	$< ۱۰^{-۳۱}$
نرخ جرم مواد مخدر	۲۵/۶۴۸	۴/۹۷۲	۱۵/۳۱۱	۳۵/۰۰۰	$< ۱۰^{-۳۱}$

جدول ۴: ملاک‌های ارزیابی مدل‌های نرخ نزاع و درگیری

مدل	DIC	p_D	LogScore
۱	−۱۰۱/۰۶۴	۳/۸۱۶	−۲/۳۹۴
۲	**−۱۲۷/۰۸۸**	**۱۰/۱۴۲**	**−۳/۱۸۷**
۳	−۱۲۴/۳۰۱	۱۰/۳۸۳	−۳/۱۰۴

جرم سرقت: با فرض این‌که y_i، نرخ سرقت در هر منطقه، دارای توزیع بتا به‌صورت (۸) است، برای بررسی تاثیر متغیرهای تبیینی معرفی شده بر نرخ سرقت، سه مدل مختلف با روش INLA تحلیل و مورد مقایسه قرار می‌گیرند.

مدل ۱: با درنظر گرفتن متغیرهای تبیینی جرم مواد مخدر، جمعیت شناور، مهاجرت و طلاق به‌صورت خطی و لحاظ همبستگی فضایی، مدل ۱ به‌صورت

$$\eta_i = \boldsymbol{\beta Z}^T + f_S(s_i) + f_R(s_i) \qquad i = ۱, \ldots, ۲۲,$$

در نظر گرفته می‌شود، که در آن $\boldsymbol{\beta} = (\beta_0, \beta_N, \beta_P, \beta_D, \beta_{Im})$ بردار ضرایب اثرات ثابت به‌ترتیب مربوط به عرض از مبدا، جرم مواد مخدر، نرخ جمعیت شناور، نرخ مهاجرت و نرخ طلاق است. برای هر یک از ضرایب توزیع پیشینی نرمال با میانگین صفر و واریانس ۱۰۰۰۰ اختیار شده است. همچنین $f_R(\cdot)$ و $f_S(\cdot)$ توابعی نامعلوم از ساختار غیرفضایی و فضایی بین مناطق هستند، که برای همبستگی فضایی مدل پیشینی بسیج با دقت نامعلوم τ_S و توزیع پیشینی نرمال با میانگین صفر و دقت نامعلوم τ_R برای اثر غیر فضایی مناطق در نظر گرفته شده است.

میدان تصادفی پنهان و بردار ابر پارامترهای مدل ۱ به ترتیب $\boldsymbol{x} = \{f_S(\cdot), f_R(\cdot), \beta_0, \beta_N, \beta_P, \beta_D, \beta_{Im}, \eta_i\}$ و $\boldsymbol{\theta} = (\phi, \tau_S, \tau_R)^T$ هستند. برای هر یک از درایه‌های بردار $\boldsymbol{\theta}$ توزیع پیشینی گامای مبهم به‌ترتیب با پارامترهای $(0.01, 10^{-7})$، $(10^{-6}, 10^{-4})$ و $(10^{-7}, 10^{-3})$ فرض شده است. برآورد ضرایب، انحراف استاندارد و صدک‌های ۰/۰۲۵ و ۰/۹۷۵ و ملاک KLD برای این مدل در جدول ۵ ارائه شده‌اند. با توجه به مقدار آن‌ها تاثیر طلاق بیش از تاثیر مواد مخدر، جمعیت شناور و مهاجرت بر سرقت است. به‌علاوه مقادیر بسیار کوچک KLD می‌تواند بیانگر فاصله کم توزیع پسینی برآورد شده از توزیع واقعی باشد.

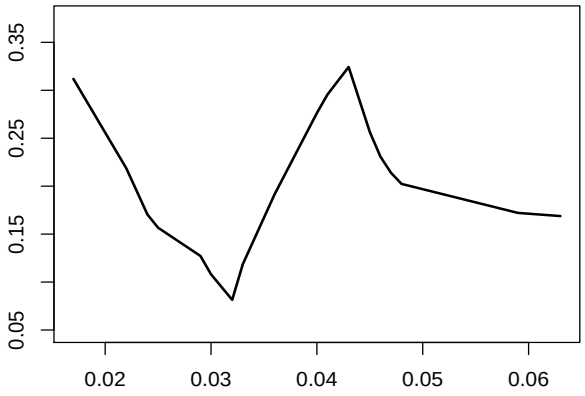
جدول ۵: برآورد ضرایب رگرسیونی و ملاک‌های ارزیابی مدل ۱ نرخ جرم سرقت

		انحراف	صدک‌ها		
ضریب	برآورد	استاندارد	۰/۰۲۵	۰/۹۷۵	KLD
عرض از مبدا	-۳/۴۲۹	۰/۴۲۳	-۴/۲۷۴	-۲/۶۰۰	$< 10^{-31}$
نرخ طلاق	۵۱/۸۹۱	۱۵/۷۳۹	۲۰/۱۰۹	۸۲/۲۸۷	$< 10^{-31}$
نرخ جمعیت شناور	۱/۱۴۸	۸/۹۱۹	-۱۶/۶۶۵	۱۸/۵۷۶	$< 10^{-31}$
نرخ جرم مواد مخدر	۱۹/۹۷۵	۱۱/۳۶۸	-۲/۸۸۴	۴۲/۰۳۶	$< 10^{-31}$
نرخ مهاجرت	-۰/۲۴۱	۰/۳۹۶	-۱/۰۷۳	-۰/۴۹۸	$< 10^{-31}$

مدل ۲: برای بررسی تاثیر غیر خطی نرخ جمعیت شناور، این بار آن را به‌صورت پنهان وارد مدل کرده و مدل ۲ به‌صورت

$$\eta_i = \boldsymbol{\beta}_۲ \boldsymbol{Z}^T + f_S(s_i) + f_R(s_i) + f_P(p_i) \qquad i = ۱, \ldots, ۲۲,$$

در نظر گرفته شده است، که در آن $\boldsymbol{\beta}_۲ = (\beta_۰, \beta_D, \beta_N, \beta_{Im})$ و $f_P(\cdot)$ تابعی نامعلوم از نرخ جمعیت شناور است. برای نرخ جمعیت شناور مدل پیشینی RW1 با دقت τ_P و برای τ_P با انجام تحلیل حساسیت توزیع پیشینی گامای مبهم با پارامترهای $(۱۰^{-۳}, ۱۰^{-۴})$ اختیار شده است. همان طور که در جدول ۷ ملاحظه می‌شود مقدار ملاک DIC مدل ۱ کاهش چشم‌گیری داشته است که بیانگر بهبود مدل ۲ نسبت به مدل ۱ است. شکل ۴ اثر غیر خطی نرخ جمعیت شناور بر جرم سرقت در مدل ۲ را نشان می‌دهد. با توجه به برآورد ضرایب مدل ۲ در جدول ۶، نرخ مهاجرت در مقایسه با نرخ طلاق و مواد مخدر تاثیر کمی بر مدل دارد.



شکل ۴: اثر برآورد شده جمعیت شناور

مدل ۳: با حذف متغیر تبیینی نرخ مهاجرت از مدل ۲، مدل ۳ به‌صورت

$$\eta_i = \boldsymbol{\beta}_۳ \boldsymbol{Z}^T + f_S(s_i) + f_R(s_i) + f_P(p_i) \qquad i = ۱, \ldots, ۲۲,$$

در نظر گرفته شده است، که در آن $\boldsymbol{\beta}_۳ = (\beta_۰, \beta_N, \beta_D)$. با توجه به ملاک‌های ارزیابی مدل در جدول ۷، مقدار p_D و LogScore تقریباً با مدل قبل اختلاف چندانی ندارد، اما مقدار DIC افزایش یافته است و مدل نسبت به مدل دوم بهبود نیافت.

جدول ۶: برآورد ضرایب رگرسیونی و ملاک‌های ارزیابی مدل ۲ نرخ جرم سرقت

		انحراف	صدک‌ها		
ضریب	برآورد	استاندارد	۰/۰۲۵	۰/۹۷۵	KLD
عرض از مبدا	−۳/۴۵۸	۰/۲۵۲	−۳/۹۵۵	−۲/۹۵۸	$< ۱۰^{-۳۱}$
نرخ طلاق	۵۶/۱۰۳	۱۵/۷۳۰	۲۴/۵۴۲	۸۶/۷۱۶	$< ۱۰^{-۳۲}$
نرخ جرم مواد مخدر	۱۹/۷۶۸	۹/۶۹۶	۰/۲۷۰	۳۸/۵۹۶	$< ۱۰^{-۱۶}$
نرخ مهاجرت	−۰/۱۹۹	۰/۳۸۲	−۰/۹۹۲	۰/۵۲۳	$< ۱۰^{-۴۵}$

بنابراین از بین سه مدل معرفی شده مدل ۲ برتر است. با وجود این‌که این مدل با مقدار p_D، ۵/۱۷۱ از مدل ۱ پیچیده‌تر است اما مقادیر DIC و LogScore آن نسبت به مدل‌های دیگر کمتر است، بنابراین بر دو مدل دیگر ارجحیت دارد.

جدول ۷: ملاک‌های ارزیابی مدل‌های سرقت

مدل	DIC	p_D	LogScore
۱	−۱۰۱/۹۵	۵/۷۶۵	−۲/۴۹۹
۲	**−۱۱۰/۰۴**	**۵/۱۷۱**	**−۲/۶۵۴**
۳	−۱۰۴/۲۴	۷/۵۲۱	−۲/۵۶۱

۷ مطالعه شبیه‌سازی

در این بخش برای مقایسهٔ اختلاف سرعت محاسبات با دو روش INLA و MCMC، پارامترهای مدل ۲ که از بین سه مدل معرفی شده به عنوان مدل برتر برای مدل‌بندی نرخ جرم نزاع و درگیری انتخاب شد را یک بار به روش INLA و بار دیگر به روش MCMC برآورد و از نظر زمان محاسبات با هم مقایسه عددی می‌شوند. می‌توان برای ارزیابی نظری سرعت و دقت روش INLA براساس تعداد پارامترهای موثر و نرخ خطای تقریب‌ها به بخش ۴ رو و مارتینو (۲۰۰۹) مراجعه کرد. در این شبیه‌سازی، برای متغیرهای تبیینی با اثر خطی در مدل از مشخصه‌های جرم شهر تهران در موقعیت‌های مناطق ۲۲ گانه شهر تهران استفاده می‌شود. برای تولید نمونه از میدان‌های تصادفی مارکوفی، پس از تعیین ماتریس دقت در هر موقعیت یک

مقدار برای متغیر پاسخ شبیه‌سازی شده است. این کار ۱۰۰۰۰ بار تکرار شده و در هر تکرار، پارامترهای مدل (۹) با دو روش INLA و MCMC برآورد گردیده و اختلاف دو مدل براساس ملاک میانگین قدر مطلق اختلاف[۱۵] به‌صورت

$$ADM = \frac{۱}{۱۰۰۰۰} \sum_{j=۱}^{۱۰۰۰۰} \sum_{i=۱}^{۲۲} \frac{|\hat{\eta}_{Iij} - \hat{\eta}_{Mij}|}{۲۲},$$

ارزیابی می‌شود، که در آن $\hat{\eta}_{Iij}$ و $\hat{\eta}_{Mij}$ برآوردهای میانگین متغیر پاسخ به‌ترتیب با روش‌های INLA و MCMC در موقعیت i و تکرار jام هستند. لازم به ذکر است، روش INLA با بستهٔ نرم افزاری INLA و روش MCMC با بستهٔ نرم افزاری R2BayesX، که به ترتیب از پایگاه‌های www.r-inla.org و cran.r-project.org قابل دسترس هستند، در محیط R نسخهٔ ۲.۱۵.۱ ۱ اجرا شده‌اند. روش MCMC براساس نمونه‌ای تصادفی به حجم ۱۰۰۰ با ۱۲۰۰۰ تکرار، مرحلهٔ داغیدن ۲۰۰۰ و تاخیر دهم انجام شده است. ملاک محاسبه شده در ۱۰۰۰۰ بار شبیه‌سازی برابر $۵/۴۱ \times ۱۰^{-۱۰}$ است که بیانگر اختلاف ناچیز نتایج دو روش است، در حالی که زمان انجام محاسبات در رایانه (تحت سیستم عامل ۶۴ بیت با حافظهٔ ۴ گیگابایت) با روش INLA، ۷ ساعت و با الگوریتم MCMC، ۴۷ ساعت به طول انجامید، یعنی محاسبات INLA حدود شش برابر سریع‌تر از روش MCMC انجام شده است.

۸ بحث و نتیجه‌گیری

مدل‌های رگرسیون جمعی ساختاری که مدل‌های پر کاربرد خطی، خطی تعمیم‌یافته، آمیخته خطی تعمیم‌یافته، جمعی، جمعی تعمیم‌یافته، آمیخته جمعی تعمیم‌یافته زیر رده‌ای از این مدل‌ها هستند معرفی شدند. در تحلیل بیزی این مدل‌ها، توزیع‌های پسینی فرم بسته‌ای ندارند و محاسبات با روش‌های MCMC زمان‌بر هستند. اما محاسبات با استفاده از تقریب لاپلاس آشیانی جمع بسته به‌دلیل استفاده از تقریب‌های گاوسی، لاپلاس و لاپلاس ساده شده سرعت می‌یابد. در مطالعه نرخ جرم شهر تهران، برای هر یک از جرم‌های سرقت و نزاع و درگیری سه مدل با

[۱۵] Absolute deviance mean

روش INLA برازش داده و از بین آن‌ها، مدل برتر انتخاب شد. نتایج حاصل از مدل‌بندی نرخ نزاع و درگیری حاکی از آن است که سرانهٔ فضای سبز و جرم مواد مخدر به‌صورت خطی و نرخ طلاق به‌صورت غیر خطی بر نرخ نزاع و درگیری در مناطق ۲۲گانه شهر تهران تأثیر گذار هستند. همچنین مدل‌بندی نرخ جرم سرقت بیان کنندهٔ تاثیر خطی نرخ جرم مواد مخدر، نرخ طلاق و مهاجرت و تأثیر غیر خطی نرخ جمعیت شناور بر نرخ جرم سرقت است که گاهی تأثیر آن افزاینده و گاهی کاهنده است. بعلاوه 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.

تقدیر و تشکر

نویسندگان از پیشنهادات ارزنده داوران گرامی مجله که باعث ارائه بهتر و بهبود مقاله شده است، کمال تشکر را دارند. از حمایت و همراهی پژوهشکده آمار برای در اختیار قرار دادن بخشی از داده‌ها و همچنین پلیس پیشگیری ناجا به خاطر در اختیار قرار دادن داده‌های جرم شهر تهران قدردانی می‌شود. از حمایت قطب علمی داده‌های ترتیبی و فضایی دانشگاه فردوسی مشهد نیز تشکر می‌شود.

مراجع

قیومی، ز.، قلی‌زاده، 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.

Besag, J. (1974), Spatial Interaction and the Statistical Analysis of Lattice System (with discussion), *Journal of the Royal Statistical Society*, B, **36**, 192-225.

Breslow, N. E. and Clayton, D. G. (1993), Approximate Inference in Generalized Linear Mixed Models, *Journal of the American Statistical Association*, **88** , 9-25.

Eidsvik, J., Martino, S. and Rue, H. (2009), Approximate Bayesian Inference in Spatial Generalized Linear Mixed Models, *Scandinavian Journal of Statistics*, **36**, 1-22.

Fahrmeir, L. and Tutz, G. (2001), *Multivariate Statistical Modeling Based on Generalized Linear Models*, 2nd edn. Springer, Berlin.

Fong Y., Rue, H. and Wakefield, J. (2010), Bayesian Inference for Generalized Linear Mixed Models, *Biostatistics*, **11**, 397-412.

Gneiting, T. and Raftery, A. (2007), Strictly Proper Scoring Rules, Prediction and Estimation, *Journal of American Statistical Association*, B, **102**, 359-378.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models, Volume 43 of Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Kullback, S. and Leibler, R. A. (1951), On Information and Sufficiency, *Annals of Mathematical Statistics*, **22**, 79-86.

Lin, X. and Zhang, D. (1999), Inference for Generalized Additive Mixed Models By Using Smoothing Splines, *Journal of Royal Statistical Society*, **61**, 381-400.

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

Moran, P. (1950), Notes on Continuous Stochastic Phenomena, *Biometrika*, **37**, 17-23.

Nelder, J. and Wedderburn, R. (1972), Generalized Linear Models, *Journal of The Royal Statistical Society*, A, **135**, 370-384.

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

Roos, M. and Held, L. (2011), Sensitivity Analysis in Bayesian Generalized Linear Mixed Models for Binary Data, *Journal of Bayesian Analysis*, **6**, 259-278.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications, Vol 104 of Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

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

Schrödle, B., Held, L., Riebler, A. and Danuser, J. (2011), Using INLA for the Evaluation of Veterinary Surveillance Data from Switzerland, *Journal of the Royal Statistical Society*, **60**, 261-279.

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