تحلیل مدل‌های آمیخته خطی تعمیم‌یافته فضایی با استفاده از تقریب لاپلاس آشیانی جمع‌بسته

فاطمه حسینی[1]، امید کریمی، منور محمدکریمی

گروه آمار، دانشگاه سمنان

تاریخ دریافت: ۱۳۹۳/۳/۲۵ تاریخ پذیرش: ۱۳۹۳/۶/۲۳

چکیده: برای مدل‌بندی پاسخ‌های گسسته فضایی زمین‌آمار از مدل‌های آمیخته خطی تعمیم-یافته فضایی استفاده می‌شود و ساختار همبستگی فضایی داده‌ها از طریق متغیرهای پنهان در نظر گرفته می‌شود. از مهم‌ترین اهداف در بررسی این مدل‌ها پیش‌گویی متغیرهای پنهان و برآورد پارامترهای مدل است. در این مقاله برای تحلیل این مدل‌ها، ابتدا یک روش پیش‌گویی ارائه و سپس به بیان رهیافت بیزی و الگوریتم‌های مونت‌کارلویی پرداخته می‌شود. به دلیل پیچیدگی این مدل‌ها و استفاده از نمونه‌های مونت‌کارلویی در تحلیل بیزی، زمان محاسبات بسیار طولانی است. برای رفع این مشکل روش بیزی تقریبی با استفاده از تقریب لاپلاس آشیانی جمع‌بسته بررسی می‌شود. درنهایت یک مجموعه داده واقعی مربوط به تعداد روزهای دارای بارندگی استان سمنان در سال ۱۳۹۱، مشاهده‌شده در ایستگاه‌های هواشناسی این استان با مدل و روش‌های معرفی‌شده موردمطالعه قرار می‌گیرد.

واژه‌های کلیدی: مدل آمیخته خطی تعمیم‌یافته فضایی، الگوریتم‌های مونت‌کارلویی، رهیافت بیزی

رده‌بندی ریاضی (۲۰۱۰): ۶۲F۱۵، ۹۱B۷۲

۱- مقدمه

مدل‌های خطی تعمیم‌یافته شامل دامنه گسترده‌ای از مدل‌های آماری هستند که می‌توان آن‌ها را به مدل‌هایی با متغیرهای پاسخ گسسته تعمیم داد. در این مدل‌ها با فرض استقلال مشاهدات، با استفاده از یک تابع پیوند بین میانگین مشاهدات و متغیرهای کمکی ارتباط برقرار می‌شود. نلدر و ودربرن [۱] اولین کسانی بودند که چارچوب واحدی برای این مدل‌ها ارائه کردند. مک‌کلا و نلدر [۲] به تعمیم مدل‌های خطی برای مدل‌بندی متغیرهای پاسخ گسسته

1- آدرس الکترونیکی نویسنده مسئول مقاله: fatemeh.hoseini@semnan.ac.ir

پرداختند و به‌طور مفصل پیرامون مدل‌های خطی تعمیم‌یافته مطالعه نمودند. در حالتی که بین مشاهدات همبستگی وجود دارد، در این مدل فرض استقلال مشاهدات به استقلال شرطی تعدیل و همبستگی بین آن‌ها با اضافه کردن اثرات تصادفی از طریق متغیرهای پنهان به مدل در نظر گرفته می‌شود و در این صورت مدل خطی تعمیم‌یافته تبدیل به مدل آمیخته خطی تعمیم‌یافته می‌شود. درصورتی‌که همبستگی مذکور از نوع فضایی باشد، مدل آمیخته خطی تعمیم‌یافته به مدل آمیخته خطی تعمیم‌یافته فضایی[1] (SGLM) تبدیل می‌شود. برسلو و کلایتون [۳] از این مدل در مطالعات پزشکی استفاده کردند. دیگل و همکاران [۴] این مدل‌ها را برای تحلیل متغیرهای فضایی گسسته در یک ناحیه پیوسته به‌کار بردند.

یک مسئله مهم در مدل SGLM پیش‌گویی متغیرهای پنهان در موقعیت‌های فاقد مشاهده است، که مستلزم برآورد پارامترهای مدل و متغیرهای پنهان در موقعیت‌های دارای مشاهده‌ی پاسخ است. لازم به ذکر است که چون در این مدل توزیع پاسخ‌های فضایی گاوسی نیست و علاوه بر این به دلیل وجود متغیرهای پنهان، برخلاف مدل‌های خطی، تابع درستنمایی شکل بسته‌ای ندارد و بنابراین برآورد پارامترها و پیش‌گویی‌ها به‌راحتی و از راه‌های مستقیم امکان‌پذیر نیستند، لذا در اکثر مقالات با پذیرش فرض نرمال بودن متغیرهای پنهان به ارائه راه‌حلی برای برآورد پارامترهای مدل و متغیرهای پنهان با ماکسیمم کردن توابع درستنمایی، شبه درستنمایی تاوانیده یا درستنمایی سلسله مراتبی به روش‌های عددی پرداخته‌شده است. ازجمله مک‌کلا [۵]، با به‌کار بردن روش‌های عددی مثل ماکسیمم‌سازی امید ریاضی مونت‌کارلویی، نیوتن رافسون مونت‌کارلویی و ماکسیمم درستنمایی شبیه‌سازی شده، برآورد ماکسیمم درستنمایی پارامترها را در مدل‌های آمیخته خطی تعمیم‌یافته با اثرات تصادفی غیر فضایی به دست آورد. همچنین پن و تامسون [۶]، برآورد شبه مونت‌کارلو و لین [۷]، برآورد پارامترهای مدل آمیخته خطی تعمیم‌یافته پواسونی را با ترکیب درستنمایی نما و شبه درستنمایی تاوانیده موردمطالعه قراردادند. ناتارجان و کاس [۸]، با بسط روش‌های بیزی در تحلیل مدل‌های آمیخته خطی تعمیم‌یافته نحوه انتخاب پیشین در این مدل‌ها را مطالعه کردند.

ژانگ [۹] یک الگوریتم مونت‌کارلویی با عنوان الگوریتم گرادیانت ماکسیمم‌سازی امید ریاضی برای به‌دست آوردن برآورد ماکسیمم درستنمایی پارامترهای مدل ارائه نمود. کریستن‌سن و همکارانش مقالات زیادی در راستای حل مشکل برآورد پارامترها و پیش‌گویی در مدل‌های SGLM معرفی نمودند که اکثر مقالات ایشان بر پایه استنباط بیزی و روش‌های مونت‌کارلویی می‌باشد. کریستن‌سن و همکاران [۱۰] و کریستن‌سن و وگپترسون [۱۱] با استفاده از رهیافت بیزی و روش‌های مونت‌کارلویی به تحلیل مدل‌های SGLM پرداختند و نشان دادند که

1- Spatial Generalized Linear Mixed Model

استفاده از الگوریتم لانژون-هستینگس، برای شبیه‌سازی از توزیع پسین متغیرهای پنهان گاوسی مفید می‌باشد. کریستن‌سن [۱۲] با روش ماکسیمم درستنمایی و الگوریتم مونت‌کارلو، پارامترها و پیش‌گویی بهینه را در مدل‌های SGLM با فرض نرمال بودن متغیرهای پنهان به دست آورد. کریستن‌سن و همکاران [۱۳] با معرفی روش‌های مونت‌کارلویی استوار به تحلیل مدل‌های SGLM با متغیرهای پنهان نرمال پرداختند و اینسورث و دین [۱۴] برآوردهای بیزی را با برآوردهای شبه درستنمایی تاوانیده در این مدل‌ها مقایسه کردند.

باغیشنی و محمدزاده [۱۵]، الگوریتم عددی همسانه‌سازی داده‌ها را برای محاسبه برآورد درستنمایی معرفی کردند. حسینی [۱۶]، حسینی و همکاران [۱۷] و حسینی [۱۸] الگوریتم-های مفیدی برای پیش‌گویی و به‌دست آوردن برآوردهای درستنمایی معرفی نمودند.

روش‌های فوق، روش‌هایی مفید برای به دست آوردن برآوردها و تحلیل بیزی پارامترهاست اما مشکل بزرگی مانند طولانی بودن زمان محاسبات را دارد. برای حل این مشکل رو و همکاران [۱۹] روش تقریب لاپلاس آشیانی جمع‌بسته[1] (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 [۲۰] از تقریب معرفی‌شده توسط رو و مارتینو [۲۱] استفاده کردند و به تحلیل بیزی تقریبی مدل‌های SGLM پرداختند. حسینی و همکاران [۲۲] و حسینی و محمدزاده [۲۳] با به‌کار بردن توزیع چوله نرمال بسته برای متغیرهای پنهان تحلیل بیزی و بیزی تقریبی این مدل‌ها را بیان نمودند. قلی‌زاده و همکاران [۲۴] از تقریب لاپلاس آشیانی جمع‌بسته 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 قراردادند.

در این مقاله ابتدا مدل SGLM معرفی می‌شود، سپس برآورد پارامترهای مدل و متغیرهای پنهان در موقعیت‌های دارای مشاهده پاسخ و پیش‌گویی متغیرهای پنهان در موقعیت‌های فاقد مشاهده پاسخ، با دو رهیافت بیزی معمولی و تقریبی ارائه می‌شود. درنهایت به مطالعه مجموعه داده‌های تعداد روزهای بارندگی در استان سمنان پرداخته می‌شود.

۲- مدل

فرض کنید $\mathbf{x}=(x_1,\ldots,x_n)'$ بردار متغیرهای پنهان فضایی در n موقعیت $\{s_1,\ldots,s_n\}$ با چگالی $\pi(\mathbf{x}\,|\,\boldsymbol{\eta})=N_n(H\boldsymbol{\beta},\Sigma_{\boldsymbol{\theta}})$ از یک میدان تصادفی گاوسی باشد، که در آن

1- Integrated Nested Laplace Approximation

$\boldsymbol{\eta}=(\boldsymbol{\beta}',\boldsymbol{\theta}')'$ بردار پارامترهای مدل، H ماتریس $n\times(p+1)$ متغیرهای تبیینی، $\boldsymbol{\beta}'=(\beta_0,\beta_1,\ldots,\beta_p)$ بردار پارامترهای رگرسیونی و $\boldsymbol{\theta}$ بردار پارامترهای همبستگی فضایی مدل هستند. همچنین فرض کنید $\mathbf{y}=(y_1,\ldots,y_k)'$ بردار متغیرهای پاسخ گسسته از یک میدان تصادفی فضایی ناگاوسی در موقعیت‌های $\{s_1,\ldots,s_k\}$ باشد. با فرض استقلال شرطی روی متغیرهای پنهان، $\pi(\mathbf{y}\mid\mathbf{x})$ متعلق به خانواده نمایی با تابع چگالی $\pi(\mathbf{y}_i\mid x_i)=\exp\{y_i x_i-b(x_i)+c(y_i)\}$ است، که در آن $b(\cdot)$ و $c(\cdot)$ توابع معلوم‌اند. اکنون با توجه به خانواده توزیع متغیر پاسخ و تعریف یک تابع پیوند مشخص مانند $g(\cdot)$ مدل به‌صورت $E(y_i\mid x_i)=g^{-1}(x_i)$ تعریف می‌شود و مؤلفه‌های مدل به فرم

$$\begin{aligned}\pi(\mathbf{x}\mid\mathbf{y},\boldsymbol{\eta}) &\propto \pi(\mathbf{y}\mid\mathbf{x}^{obs})\pi(\mathbf{x}\mid\boldsymbol{\eta})\\ &\propto |\Sigma_\theta|^{-1/2}\exp\left\{\sum_{i=1}^{k}[y_i x_i-b(x_i)+c(y_i)]-\frac{1}{2}(\mathbf{x}-H\boldsymbol{\beta})'\Sigma_\theta^{-1}(\mathbf{x}-H\boldsymbol{\beta})\right\},\end{aligned}\qquad(۱)$$

می‌باشد.

۳- پیش‌گویی مدل

در یک مدل SGLM هدف پیش‌گویی متغیرهای پنهان در موقعیت‌های فاقد مشاهده پاسخ $\{s_1,\ldots,s_k\}$ و برآورد متغیرهای پنهان در موقعیت‌های دارای مشاهده پاسخ $\{s_{k+1},\ldots,s_n\}$ است. برای این منظور بردار متغیرهای پنهان به‌صورت $\mathbf{x}=(\mathbf{x}^{obs'},\mathbf{x}^{pred'})'$ تجزیه می‌شود. که $\mathbf{x}^{obs}=A\mathbf{x}$، متعلق به k موقعیت دارای مشاهده پاسخ است، که در آن $A=[I_{k\times k}\mid 0_{k\times n-k}]$ و $\mathbf{x}^{pred}$ بردار متغیرهای پنهان در $n-k$ موقعیت انتخاب‌شده برای پیش‌گویی است. حسینی [۱۸] قضیه زیر را با فرض معلوم بودن پارامترهای مدل برای محاسبه پیش‌گوی مینیمم میانگین مربع خطا متغیرهای پنهان در $n-k$ موقعیت ارائه نمود.

قضیه ۱: فرض کنید توزیع متغیرهای پنهان فضایی $\mathbf{x}=(\mathbf{x}^{obs'},\mathbf{x}^{pred'})'$ متعلق به خانواده نرمال $N_n(H\boldsymbol{\beta},\Sigma_{\boldsymbol{\theta}})$ باشد، به‌طوری‌که $H\boldsymbol{\beta}=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\Sigma_\theta=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$ و با شرط روی متغیرهای پنهان فضایی، توزیع متغیرهای مستقل شرطی پاسخ متعلق به خانواده نمایی باشد، در این صورت $E(\mathbf{x}^{pred}\mid\mathbf{y})=\boldsymbol{\mu}_2+\Sigma_{21}\Sigma_{11}^{-1}(E(\mathbf{x}^{obs}\mid\mathbf{y})-\boldsymbol{\mu}_1)$ پیش‌گوی مینیمم میانگین مربع خطا متغیرهای پنهان در $n-k$ موقعیت $j=k+1,\ldots,n$ ام می‌باشد.

اما در عمل اغلب پارامترهای مدل نامعلوم هستند و به دلیل وجود متغیرهای پنهان فضایی و ناگاوسی بودن متغیرهای پاسخ، در این مدل‌ها محاسبه برآورد پارامترهای مدل به‌راحتی

قابل‌محاسبه نیستند. یک روش معمول برای برآورد پارامترها، رهیافت ماکسیمم درستنمایی است.

فرض کنید $(\mathbf{y},\mathbf{x})$ شامل بردار متغیرهای پاسخ گسسته فضایی و متغیرهای پنهان باشد، آنگاه تابع درستنمایی به‌صورت $L(\boldsymbol{\eta}\,|\,\mathbf{y})=\int\prod_{i=1}^{k}\pi(y_i\,|\,x_i)\pi(\mathbf{x}\,|\,\boldsymbol{\eta})d\mathbf{x}$ است و تابع درستنمایی کامل به‌صورت

$$L_c(\boldsymbol{\eta}\,|\,\mathbf{y},\mathbf{x})=\prod_{i=1}^{k}\pi(y_i\,|\,x_i)\pi(\mathbf{x}\,|\,\boldsymbol{\eta}), \qquad (۲)$$

خواهد شد که شکل بسته‌ای ندارد و برای به‌دست آوردن برآورد ماکسیمم درستنمایی باید از الگوریتم‌های عددی استفاده کرد. علاوه بر این چون در عمل گاهی اوقات گردآوری نمونه با حجم زیاد امکان‌پذیر نیست، در این صورت نتایج تحلیل کلاسیک از دقت کمی برخوردار است و همچنین در تحلیل‌های کلاسیک از اطلاعات پیشینی در خصوص پارامترها استفاده نمی‌شود و لذا نتایجی با دقت کم‌تری نسبت به تحلیل‌های بیزی ارائه می‌کند در بخش بعد تحلیل مدل-های SGLM با رهیافت بیزی ارائه می‌شود.

۴- تحلیل بیزی مدل

در این قسمت تحلیل بیزی مدل‌های SGLM روی یک ناحیه پیوسته فضایی موردبررسی قرار می‌گیرد. یک مدل پارامتری به‌صورت $C_\theta(s_i,s_j)=\sigma^2\exp\{-\|s_i-s_j\|/\varphi\}$ برای کوواریانس فضایی در نظر بگیرید. که در آن $\|\cdot\|$ نرم اقلیدسی و σ و φ به ترتیب پارامترهای مقیاس و همبستگی فضایی هستند. برای سره بودن توزیع پسین، برای پارامترهای مدل پیشین‌های سره در نظر گرفته می‌شود. پیشین متداول برای β نرمال و برای σ و φ به ترتیب گامای معکوس و گاما در نظر گرفته می‌شود. اکنون با فرض استقلال پیشین‌ها، توزیع پیشین توأم $\pi(\boldsymbol{\beta},\sigma,\varphi)=\pi(\boldsymbol{\beta})\pi(\sigma)\pi(\varphi)$ حاصل می‌شود. لذا توزیع پسین

$$\pi(\mathbf{x},\boldsymbol{\beta},\sigma,\varphi\,|\,\mathbf{y})\propto\pi(\mathbf{y}\,|\,\mathbf{x})\pi(\mathbf{x}\,|\,\boldsymbol{\beta},\sigma,\varphi)\pi(\boldsymbol{\beta})\pi(\sigma)\pi(\varphi)$$

خواهد شد که دارای شکل پیچیده‌ای است. بنابراین برای تولید نمونه و شبیه‌سازی از توزیع پسین از روش‌های MCMC استفاده می‌شود. فرض کنید $\boldsymbol{\beta}\sim N(a,B)$ آنگاه توزیع شرطی کامل $\boldsymbol{\beta}$

$$\begin{aligned}\pi(\boldsymbol{\beta}\mid \mathbf{y},\mathbf{x},\sigma,\varphi) &\propto \pi(x\mid \boldsymbol{\beta},\sigma,\varphi)\pi(\boldsymbol{\beta})\\ &\propto \exp\{-\frac{1}{2}[\boldsymbol{\beta}'(H'\Sigma_\theta^{-1}H+B^{-1})\boldsymbol{\beta}-2\boldsymbol{\beta}'(B^{-1}a+H'\Sigma_\theta^{-1}\mathbf{x})]\} \qquad (۳)\\ &\propto \exp\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)'\Sigma_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)\},\end{aligned}$$

است که در آن $\Sigma_\beta=(H\ \Sigma_\theta^{-1}\ H+\ B^{-1})$ و $\boldsymbol{\mu}_\beta=\Sigma_\beta(B^{-1}a+H'\Sigma_\theta^{-1}\mathbf{x})$ هستند. بنابراین تولید نمونه از توزیع شرطی کامل $\boldsymbol{\beta}$ به‌راحتی امکان‌پذیر است. یک مزیت دیگر در نظر گرفتن پیشین نرمال برای $\boldsymbol{\beta}$ کمک به کاهش تعداد پارامترهای مدل است به‌طوری‌که اگر فرض کنیم $\boldsymbol{\beta}\sim N(a,B)$ و $\mathbf{x}\mid\boldsymbol{\beta},\sigma,\varphi\sim N_n(H\boldsymbol{\beta},\Sigma_{\boldsymbol{\theta}})$ آنگاه به‌راحتی می‌توان نشان داد

$$\mathbf{x}\mid\sigma,\varphi\sim N_n(\boldsymbol{\mu}_x,\Sigma_x),$$

که در آن $\boldsymbol{\mu}_x=Ha$ و $\Sigma_x=(\Sigma_\theta+HBH')$. توزیع شرطی کامل برای سایر پارامترها

$$\begin{aligned}\pi(\sigma\mid \mathbf{y},\mathbf{x},\boldsymbol{\beta},\varphi) &\propto \pi(x\mid\boldsymbol{\beta},\sigma,\varphi)\pi(\sigma)\\ &= IG(\alpha,\tau)\phi_n(\mathbf{x};H\boldsymbol{\beta},\Sigma_\theta),\\ \pi(\varphi\mid \mathbf{y},\mathbf{x},\boldsymbol{\beta},\sigma) &\propto \pi(x\mid\boldsymbol{\beta},\sigma,\varphi)\pi(\varphi)\\ &= \Gamma(\gamma,\omega)\phi_n(\mathbf{x};H\boldsymbol{\beta},\Sigma_\theta),\end{aligned} \qquad (۴)$$

که دارای شکل خاصی از توزیع‌های شناخته‌شده نیستند و برای تولید نمونه از الگوریتم متروپلیس- هستینگس استفاده می‌شود. توزیع شرطی کامل برای هر مؤلفه بردار متغیرهای پنهان نیز به‌صورت

$$\pi(x_k\mid\mathbf{x}_{-k},\mathbf{y},\boldsymbol{\beta},\sigma,\varphi)\propto\pi(x_k\mid\mathbf{x}_{-k},\boldsymbol{\beta},\sigma,\varphi)\pi(\mathbf{y}\mid\mathbf{x})$$

می‌باشد. برای پیش‌گویی بیزی متغیرهای پنهان، توزیع پیش‌گو به‌صورت

$$\pi(x_\circ\mid\mathbf{y})=\int\pi(x_\circ\mid\mathbf{x},\boldsymbol{\beta},\sigma,\varphi)\pi(\mathbf{x},\boldsymbol{\beta},\sigma,\varphi\mid\mathbf{y})d\mathbf{x}d\boldsymbol{\beta}d\sigma d\varphi$$

است که در آن $\pi(x_\circ\mid\mathbf{x},\boldsymbol{\beta},\sigma,\varphi)$ طبق خواص توزیع نرمال دارای توزیع نرمال است. توزیع پیش‌گوی بیزی $y_\circ$ نیز به‌صورت $\pi(y_\circ\mid\mathbf{y})=\int\pi(y_\circ\mid x_\circ)\pi(x_\circ\mid\mathbf{y})dx_\circ$ است.

۴-۱- تحلیل بیزی مدل با تقریب لاپلاس آشیانی جمع‌بسته (INLA)

استفاده از الگوریتم‌های مونت‌کارلوی زنجیر مارکوفی به دلیل پیچیده بودن مدل‌های SGLM زمان‌بر هستند. ایدسویک و همکاران [۲۰] روش بیزی تقریبی را به‌جای روش بیزی معمولی و الگوریتم‌های نمونه‌گیری مونت‌کارلویی با استفاده از INLA معرفی کرده و نشان دادند

محاسبات این روش سریع‌تر و نتایجی معادل روش‌های بیزی معمولی ارائه می‌کند. مسئله اصلی در مدل‌های بیزی SGLM برای برآورد پارامترها و پیش‌گویی‌ها، محاسبه دو توزیع پسین $\pi(\mathbf{x}|\mathbf{y},\boldsymbol{\eta})$ و $\pi(\boldsymbol{\eta}|\mathbf{y})$ است. اگر $\mathbf{x}$ دارای توزیع نرمال و $\mathbf{y}$ متعلق به خانواده نمایی باشد، با خطی کردن $\pi(\mathbf{y}|\mathbf{x}^{obs})$ حول یک مقدار ثابت $\mathbf{x}^{\circ}$، ایدسویک و همکاران [۲۰] نشان دادند توزیع $\pi(\mathbf{x}|\mathbf{y},\boldsymbol{\eta})$ را می‌توان با توزیع نرمال به‌صورت

$$\hat{\pi}(\mathbf{x}|\mathbf{y},\boldsymbol{\eta}) \approx N_n(\hat{\mu}_{x|y,\eta}(\mathbf{x}^{\circ}), \hat{\Sigma}_{x|y,\eta}(\mathbf{x}^{\circ})), \qquad (۵)$$

تقریب زد، که در آن $\hat{\mu}_{x|y,\eta}(\mathbf{x}^{\circ}) = H\boldsymbol{\beta} + \Sigma_\theta A' R^{-1}(z(\mathbf{y},\mathbf{x}^{\circ}) - AH\boldsymbol{\beta})$، $R = A\Sigma_\theta A' + P$، $\hat{\Sigma}_{x|y,\eta}(\mathbf{x}^{\circ}) = \Sigma_\theta - \Sigma_\theta A' R^{-1} A\Sigma_\theta$. همچنین P یک ماتریس قطری $k \times k$ با عناصر $P(i,i) = 1/b''(x_i)$ است و داریم

$$z_i(y_i, x_i^{\circ}) = [y_i - b'(x_i^{\circ}) + x_i b''(x_i^{\circ})] / b''(x_i^{\circ}) .$$

یک تقریب لاپلاس اصلاح‌شده نیز برای توزیع‌های حاشیه‌ای پسین متغیرهای پنهان توسط ایدسویک و همکاران [۲۰] معرفی‌شده است که به‌صورت

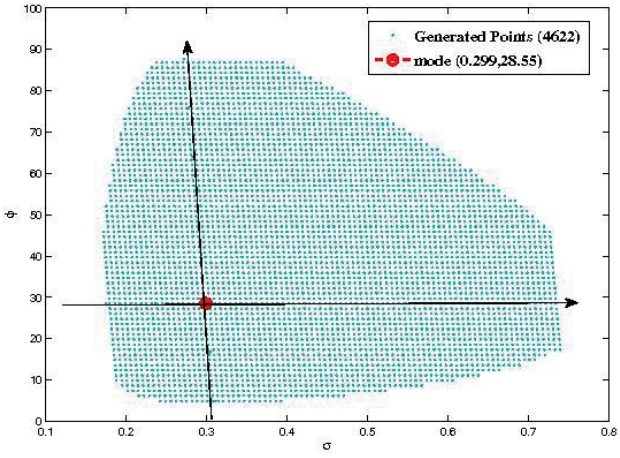
$$\tilde{\pi}_{LA}(x_j|\mathbf{y},\boldsymbol{\eta}) \propto \frac{\pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}|\boldsymbol{\eta})}{\hat{\pi}(\mathbf{x}_{-j}|x_j,\mathbf{y},\boldsymbol{\eta})}\Big|_{x=\hat{E}(\mathbf{x}_{-j}|x_j,\mathbf{y},\boldsymbol{\eta})}, \qquad (۶)$$

می‌باشد، بنا به خواص توزیع نرمال $\hat{\pi}(\mathbf{x}_{-j}|x_j,\mathbf{y},\boldsymbol{\eta})$ یک توزیع تقریبی نرمال است. پس از تعیین این توزیع پسین تقریبی می‌توان تقریب لاپلاس توزیع پسین پارامترها را از رابطه

$$\hat{\pi}_{LA}(\boldsymbol{\eta}|\mathbf{y}) \propto \frac{\pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})}{\hat{\pi}(\mathbf{x}|\mathbf{y},\boldsymbol{\eta})}\Big|_{x=\hat{E}(x|y,\eta)}, \qquad (۷)$$

تعیین نمود. برای تعیین تقریب (۷) مد تابع $\log\hat{\pi}(\boldsymbol{\eta}|\mathbf{y})$ با روش‌های بهینه‌سازی به‌صورت $\boldsymbol{\eta}^*$ تعیین می‌شود. ماتریس هسین $H = \frac{\partial^2 \log\hat{\pi}(\boldsymbol{\eta}|\mathbf{y})}{\partial\eta^i\partial\eta^j}$ محاسبه و عکس آن به‌صورت $H^{-1} = V\Lambda V'$ تجزیه می‌شود، که در آن V ماتریس بردارهای ویژه و Λ ماریس قطری مقادیر ویژه آن است. مبدأ مختصات را به مد $\boldsymbol{\eta}^*$ انتقال داده، فرمول مختصات در مبدأ $\boldsymbol{\eta}^*$ به‌صورت $\boldsymbol{\eta}(t) = \boldsymbol{\eta}^* + V\Lambda^{-1}t$ تعریف می‌شود که در آن t مقادیر استانداردشده هستند. با شروع از مبدأ مختصات جدید روی هر یک از محورها نقاطی به فاصله مقادیر صحیح δ_t به‌گونه‌ای اختیار می‌شوند که شرط $\log\hat{\pi}(\boldsymbol{\eta}(\circ)|\mathbf{y}) - \log\hat{\pi}(\boldsymbol{\eta}(t)|\mathbf{y}) < \delta_t$ برقرار باشد. سپس به‌طور مشابه نقاط درون صفحات نیز تعیین می‌شوند. با به‌کار بردن این الگوریتم η_iهایی از توزیع $\hat{\pi}(\boldsymbol{\eta}|\mathbf{y})$ تولید می‌شوند و می‌توان تقریبی از توزیع را به دست آورد، (برای جزئیات

بیشتر به رو و همکاران [۱۹] مراجعه شود). به‌عنوان مثال برای مجموعه داده استفاده‌شده در این مقاله از روش توضیح داده‌شده تعداد ۴۶۲۲ نقطه تولید و در شکل ۱ رسم شده است.



شکل ۱: نقاط تولیدشده برای توزیع پسین تقریبی پارامترها

اکنون با داشتن نمونه‌های تولیدشده برای توزیع پسین تقریبی $\hat{\pi}_{LA}(\boldsymbol{\eta}\,|\,\mathbf{y})$ و مشخص شدن توزیع پسین تقریبی $\tilde{\pi}_{LA}(\mathbf{x}_j\,|\,\mathbf{y},\boldsymbol{\eta})$ توزیع تقریبی پیش‌گو برای متغیرهای پنهان به‌صورت

$$\begin{aligned}\hat{\pi}(x_j\,|\,\mathbf{y}) &= \int \hat{\pi}(x_j\,|\,\mathbf{y},\boldsymbol{\eta}_\ell)\hat{\pi}(\boldsymbol{\eta}\,|\,\mathbf{y}) \\ &= \sum_{\ell} \tilde{\pi}_{LA}(x_j\,|\,\mathbf{y},\boldsymbol{\eta}_\ell)\hat{\pi}_{LA}(\boldsymbol{\eta}_\ell\,|\,\mathbf{y}),\ \ j=1,\ldots,n,\end{aligned} \tag{۸}$$

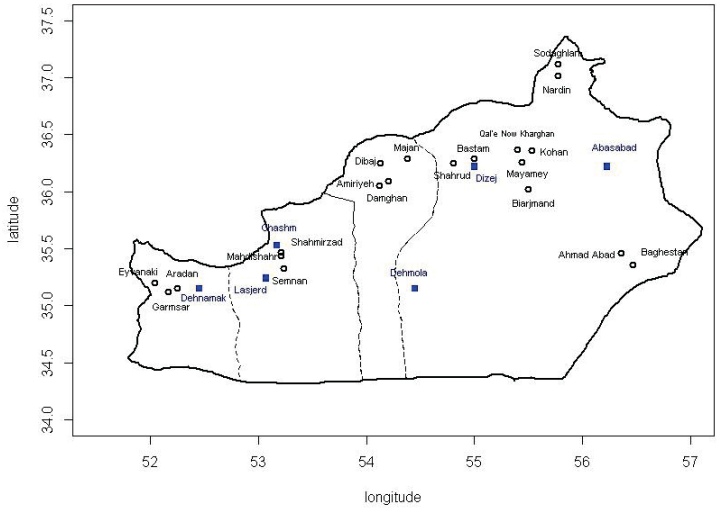
حاصل می‌شود، که در آن ℓ تعداد نقاط تولیدشده از الگوریتم تعیین توزیع پسین لاپلاس تقریبی پارامترها می‌باشد و بنابراین با توجه به رابطه (۸) این روش محاسبه توزیع تقریبی پیش‌گو به روش INLA معروف شده است. در عمل پیش‌گویی در موقعیت‌های $j=k+1,\ldots,n$ مدنظر می‌باشد. همچنین می‌توان توزیع پیشگوی y_j در موقعیت فاقد مشاهده را به‌صورت

$$\pi(y_j\,|\,\mathbf{y}) = \int \pi(y_j\,|\,x_j)\pi(x_j\,|\,\mathbf{y})dx_j,\ j=k+1,\ldots,n$$

به‌دست آورد که در آن به‌جای $\pi(x_j\,|\,\mathbf{y})$ از تقریب مستقیم یا اصلاح‌شده آن استفاده می‌شود.

۵- بررسی داده‌های بارندگی استان سمنان

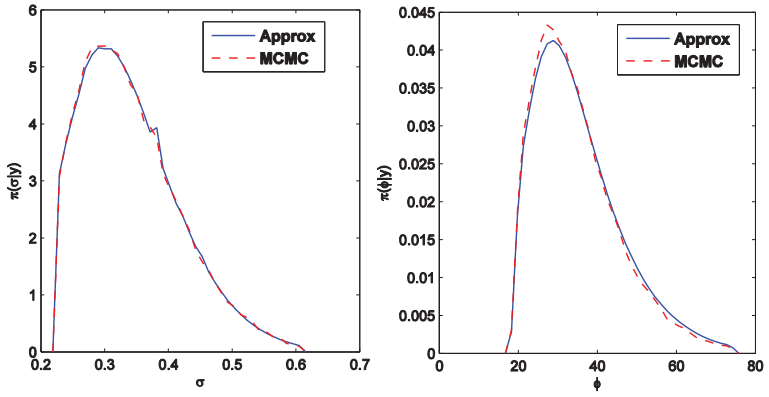
در این بخش مدل SGLM به داده‌های تعداد روزهای دارای بارندگی در استان سمنان برازش داده می‌شود و مورد تحلیل آماری قرار می‌گیرد. متغیر پاسخ تعداد روزهای دارای بارندگی است، که طی پنج ماه از اول آبان تا پایان اسفند ۱۳۹۱ در ۲۰ ایستگاه هواشناسی استان سمنان مشاهده‌شده‌اند. در شکل ۲ موقعیت ۲۰ ایستگاهی که تعداد روزهای دارای بارندگی در آن‌ها ثبت‌شده است و موقعیت‌هایی که به‌منظور پیش‌گویی انتخاب‌شده‌اند بر روی نقشه استان سمنان مشخص‌شده است. فرض می‌شود متغیرهای پاسخ دوجمله‌ای شرطی مستقل هستند و متغیرهای پنهان x نیز دارای توزیع نرمال به شکل $N(\beta_\circ\ \Sigma_\theta$ باشند. برای ساختار همبستگی فضایی تابع کوواریانس نمایی همسان‌گرد در نظر گرفته‌شده است. برای پارامتر مقیاس σ، پیشین گامای معکوس با پارامترهای (۰/۵ و ۵)، برای پارامتر همبستگی پیشین گاما با پارامترهای ۸ و۵ منظور شده است.



شکل ۲: موقعیت جغرافیایی ایستگاه‌های هواشناسی روی نقشه استان سمنان

برای این مجموعه داده‌ها متغیر کمکی وجود ندارد و برای جمله ثابت $\beta_\circ$ پیشین نرمال با میانگین ۱ و واریانس ۵ در نظر گرفته شد که به دلیل عدم اهمیت این پارامتر و با به‌کار بردن رابطه (۴)، از مدل حذف می‌شود. شکل ۳ نشان‌دهنده توزیع‌های پسین حاشیه‌ای پارامترهاست، که با ۱۰۰۰۰ تکرار از روش‌های MCMC و به‌کار بردن الگوریتم متروپولیس-هستینگس و از روش بیزی تقریبی با به‌کار بردن ۴۶۲۲ نقطه از توزیع تقریبی پسین و الگوریتم توضیح داده‌شده حاصل‌شده‌اند. خط ممتد روش بیزی تقریبی و خط‌چین روش‌های

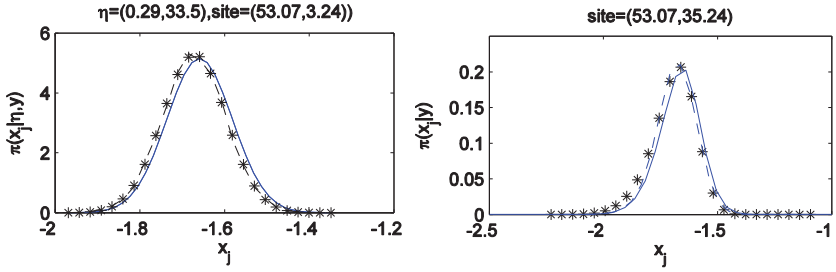
MCMC هستند، مشاهده می‌شود که دو روش بسیار شبیه به‌هم هستند. برآورد بیزی پارامترها به‌صورت $(\hat{\sigma},\hat{\varphi})=(0/29,33/5)$ به‌دست‌آمده است. اکنون با جایگذاری برآوردهای بیزی پارامترها می‌توان در نقاط جدید پیش‌گویی انجام داد.



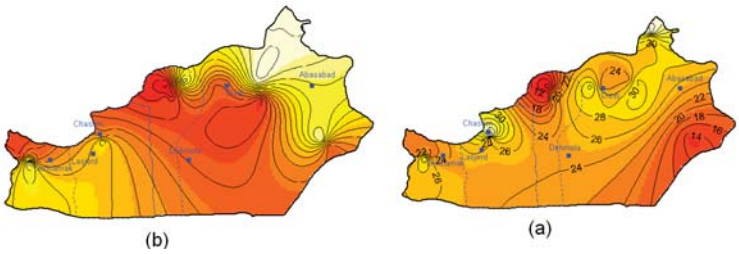
شکل ۳: نمودار توزیع‌های حاشیه‌ای پسین پارامترها با روش بیزی تقریبی (خطوط ممتد) و روش MCMC (خط‌چین)

شکل ۴(چپ) نمودار توزیع پیش‌گو برای یک موقعیت فاقد مشاهده دلخواه با طول و عرض جغرافیایی $(53/07,35/24)$ رسم شده است. برای مدل SGLM با متغیرهای پنهان نرمال، نمودار توزیع شرطی پیش‌گو با جایگذاری مد پسین تقریبی به‌جای بردار پارامتر رسم شده‌اند. نمودار مشخص‌شده با خطوط ممتد از روش بیزی تقریبی، خطوط مشخص‌شده با * از روش بیزی تقریبی اصلاح‌شده و خطوط خط‌چین از روش MCMC به‌دست‌آمده‌اند. همچنین شکل ۴ (راست) نشان‌دهنده چگالی حاشیه‌ای پیش‌گو $\pi(x_j\,|\,\mathbf{y})$ است که حاصل از سه روش بیزی تقریبی، بیزی تقریبی اصلاح‌شده و روش MCMC، می‌باشد و تشابه هر سه روش در این شکل مشخص است. همچنین برای بررسی بیشتر معادل بودن سه روش، مقادیر پیش‌گویی در شش موقعیت مشخص‌شده در شکل ۲ محاسبه و مقدار میانگین مربع انحرافات و میانگین قدرمطلق انحرافات مقادیر پیش‌گویی روش MCMC از روش تقریبی اصلاح‌شده به ترتیب ۰/۰۰۰۰۴۵۲ و ۰/۰۰۵۵۷ و میانگین مربع انحرافات و میانگین قدرمطلق انحرافات مقادیر پیش‌گویی روش MCMC از روش تقریبی مستقیم به ترتیب ۰/۰۰۲۴۶ و ۰/۰۳۹۱۸ به‌دست آمد. این مقادیر بیان‌گر نزدیک بودن روش تقریبی اصلاح‌شده به روش MCMC نسبت به روش تقریبی مستقیم است. شکل ۵ (چپ) نقشه پیش‌گویی متغیر پنهان به‌صورت لایه‌هایی در ۱۵ سطح، روی نقشه سمنان و پیش‌گویی تعداد روزهای دارای بارندگی، در ۶ ایستگاه مذکور و کل نقشه می‌باشد. بیشترین بارندگی در این استان در مناطق شمالی استان می‌باشد که به دلیل

همسایگی با استان‌های شمالی است. پیش‌گویی تعداد روزهای دارای بارندگی در ۶ ایستگاه هواشناسی موردنظر، لاسجرد، ده نمک، ده‌ملا، چاشم، دیزج و عباس‌آباد نیز در نمودار سمت راست مشخص‌شده است. زمان اجرای برنامه‌ها برای روش بیزی تقریبی در حدود ۵۰ ثانیه، برای بیزی تقریبی اصلاح‌شده حدود ۴ دقیقه و برای روش‌های MCMC حدود یک روز به طول انجامیده است.



شکل ۴: (چپ) چگالی شرطی $\hat{\pi}(x_j \mid \mathbf{y}, \boldsymbol{\eta})$ حاصل از بیزی تقریبی مستقیم (خطوط ممتد)، روش تقریبی اصلاح‌شده (*) و روش MCMC (خط‌چین). (راست) چگالی حاشیه‌ای پیش‌گو $\pi(x_j \mid \mathbf{y})$ حاصل از بیزی تقریبی مستقیم (خطوط ممتد)، روش تقریبی اصلاح‌شده (*) و روش MCMC (خط‌چین).



شکل ۵: نقشه پیش‌گویی متغیرهای پنهان (چپ)، نقشه پیش‌گوی بیزی تعداد روزهای دارای بارندگی (راست).

۶- بحث و نتیجه‌گیری

در این مقاله دو رهیافت بیزی معمولی و تقریبی برای تحلیل مدل‌های آمیخته خطی تعمیم-یافته فضایی ارائه شد و از آن‌جا که اجرای الگوریتم‌های مونت‌کارلویی برای مدل‌ها بسیار طولانی و زمان‌بر هستند، روش بیزی تقریبی بررسی شد. در یک مثال کاربردی بارندگی در استان سمنان نشان داده شد که روش بیزی تقریبی معادل روش بیزی معمولی عمل می‌کند با

این تفاوت که روش تقریبی ارائه‌شده نیاز به چند ثانیه و روش تقریبی اصلاح‌شده نیاز به چند دقیقه اجرای کامپیوتری دارند، درصورتی‌که روش‌های MCMC نیاز به‌صرف زمان طولانی‌تری برای محاسبات کامپیوتری دارند. حجم نمونه داده‌های موردبررسی در این مقاله کوچک است درصورتی‌که در بسیاری از مسائل با مجموعه داده‌ها با حجم بالا سروکار داریم که استفاده از استنباط‌های معمول کلاسیک و بیزی بسیار پیچیده و زمان‌بر است و روش بیزی تقریبی لاپلاس آشیانی جمع‌بسته به دلیل استفاده از تقریب‌های گاوسی و لاپلاس بسیار مفید به نظر می‌رسد. از ایرادات اصلی روش INLA، کاهش دقت نتایج با افزایش پارامترها می‌باشد. یکی از محدودیت‌های مدل‌های SGLM این است که متغیرهای پنهان به‌صورت میدان تصادفی گاوسی پنهان مدل‌بندی می‌شوند و به دلیل پنهان بودن، توزیع دقیق آن‌ها مشخص نیست و فرض نرمال بودن بر روی نتایج تأثیرگذار است. تاکنون مطالعاتی در خصوص به‌کار بردن کلاس چوله نرمال به‌عنوان توزیع متغیرهای پنهان انجام‌شده است که تا حدودی نتایج بهتری در بعضی موارد حاصل‌شده است. به‌عنوان پیشنهاد می‌توان متغیرهای پنهان را با کلاس بزرگ‌تری از توزیع‌ها مثل کلاس توزیع چوله t مدل‌بندی نمود.

مراجع

[1] Nelder, J.A., and Wedderburn, R.W.M. (1972). Generalized Linear Mixed Models, *Journal of the Royal Statistical Association*,Ser. A, **135**, 370-384.

[2] McCulloch, P. and Nelder, J.A. (1989), *Generalized Linear Models*, London, Chapman and Hall.

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

[4] 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.

[5] McCulloch, C. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models, *Journal of the American Statistical Association*, **92**, 162-170.

[6] Pan, J. and Thompson, R. (2007). Quasi-Monte Carlo Estimation in Generalized Linear Mixed Models, *Computational Statistics and Data Analysis*,**51**, 5765-5775.

[7] Lin, X. (2007). Estimation Using Penalized Quasi-Likelihood and Quasi-Pseudo-Likelihood in Poisson Mixed Models, *Lifetime Data Anal*, **13**, 533-544.

]8] Natarjan, R. and Kass, R.E. (2000). Refrence Bayesian methods for Generalized Linear Mixed Models, *Journal of the American Statistical Assoociation*, **95**, 227-237.

[9] Zhang, H. (2002). On Estimation and Prediction for Spatial Generalized Linear Mixed Models, *Biometrics*, **58**, 129-136.

[10] Christensen, O.F., Moller, J. and Waagepetersen R.P. (2000). Analysis of Spatial Data Using Generalized Linear Mixed Models and Langevin-Type Markov Chain Monte Carlo, Research Report R-00-2009, *Department of Mathematical Sciences, Aalborg University*.

[11] Christensen, O.F. and Waagepetersen R.P. (2002). Bayesian Prediction of Spatial Count Data Using Generalized Linear Mixed models, *Biometrics*, **58**, 280-286.

[12] Christensen, O.F. (2004). Monte Carlo Maximum Likelihood in Model-Based Geostatistics, *Journal of Computational and Graphical Statistics*, **13**, 702-718.

[13] Christensen, O.F., Roberts, G.O. and Skold, M. (2006). Robust MCMC Methods for Spatial Generalized Linear Mixed Models*Journal of Computational and Graphical Statistics*, **15**, 1-17.

[14] Ainsworth, L.M. and Dean, C.B. (2006). Approximate Inference for Disease Mapping, *Computational Statistics and Data Analysis*, **50**, 2552-2570.

[15] Baghishani, H. and Mohammadzade, 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.

[16] Hosseini, F. (2014). An efficient Algorithm to Analysis of Categorical Spatial Data, *proceedings of 12th Iranian Statistical Conference, Razi university, Iran*, 83-89.

[17] Hosseini, F. and Mohammadzadeh, M. (2013). Estimation of Spatial Generalized Linear Mixed Models with Closed Skew Normal Latent Variables, *Journal of Science Kharazmi University*, **12**, 305-312.

[18] Hosseini, F. and Mohammadzadeh, M. and Karimi, O. (2014). Pseudo-likelihood Inference for Discrete Spatial Response (A Case Study of the Semnan rainfall data), *Journal of Science Kharazmi University*, **13**, 797-808.

[19] Rue, H. and Martino, S. (2007). Approximate Bayesian Inference for Hierarchical Gaussian Markov Random fields models, *Journal of Statistical Planning and Inference*, **137**, 3177-3199.

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

[21] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian Inference for models, *Journal of Statistical Planning and Inference*, **137**, 3177-3199.

[22].Hosseini, J., 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.

[23] Hosseini, F. and Mohammadzadeh, M. (2012). Bayesian Prediction for Spatial GLMM's with Closed Skew Normal Latent Variables, *Australian & New Zealand Journal of Statistics*, **54,** 43-62

[24] 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.

Inference of Spatial Generalized Linear Mixed Models using Integrated Laplace Nested Approximation

Fatemeh Hosseini, Omid Karimi, Monavar Mohammad Karimi

Department of Statistics, University of Semnan, Semnan, Iran.

Abstract

Spatial generalized linear mixed models are used for modeling geostatistical discrete spatial responses and spatial correlation of the data is considered via latent variables. The most important interest in these models is estimation of the parameters and prediction of the latent variables. In this paper, first, a prediction method is presented. Then a Bayesian approach and MCMC algorithms are proposed. Since these models are complicated and Monte Carlo sampling is used in the Bayesian inference of these models, computation time is long. In order to resolve this problem, the Approximate Bayesian methods are considered. Finally, the proposed methods are applied to a case study on rainfall data observed in the weather stations of Semnan in 1391.

Keywords: Spatial generalized linear mixed model, Monte Carlo algorithm, Approximation Baysian approach.

Mathematics Subject Classification (2000): 91B72,62F15 .