تحلیل بیزی نرخ سرطان معده در استان گیلان با مدل اتوبتا-دوجمله‌ای فضایی

لیلا عابدین پور لیارجدمه، حسین باغیشنی[۱]، نگار اقبال

گروه آمار، دانشگاه صنعتی شاهرود

تاریخ دریافت: ۱۳۹۷/۴/۲ تاریخ پذیرش: ۱۳۹۷/۷/۱۶

چکیده: شرایط جوی و محیطی در هر منطقه جغرافیایی، زمینه را برای بروز و شیوعدل برخی بیماری‌ها مساعد می‌کنند؛ بنابراین تهیه نقشه پهنه‌بندی نرخ رویداد یک بیماری یا مرگ‌ومیر ناشی از آن بر روی نقشه جغرافیایی، در حضور عوامل مؤثر ممکن، یکی از مسائل موردتوجه پزشکان و کارشناسان امور سلامت است. با توجه به آن‌که سرطان معده شایع‌ترین نوع سرطان در استان گیلان است، در این مطالعه، با استفاده از دو مدل اتو دوجمله‌ای و اتوبتا-دوجمله‌ای بیزی فضایی به بررسی تأثیرگذاری برخی از عوامل مخاطره بر نرخ این نوع سرطان در شهرستان‌های استان گیلان می‌پردازیم. پیش‌گویی و ارائه نقشه پهنه‌بندی نرخ سرطان معده در کنار مقایسه عملکرد دو مدل پیشنهادی از دیگر اهداف انجام این مطالعه هستند. در مطالعه حاضر که از نوع کاربردی/ بوم‌شناختی است، از داده‌های ثبت‌شده توسط مرکز آموزشی درمانی رازی رشت استفاده‌شده است. داده‌های این مطالعه با استفاده از یک روش بیزی تقریبی به نام تقریب لاپلاس آشیانی جمع‌بسته (INLA) مورد تحلیل قرار گرفتند. بر اساس نتایج حاصل، می‌توان گفت مقادیر پیش‌گویی نرخ ابتلا به سرطان در اکثر شهرستان‌های استان گیلان توسط هر دو مدل پیشنهادی مشابه هستند؛ در شهرستان‌هایی که اختلاف وجود دارد، مدل اتو دوجمله‌ای نرخ بزرگ‌تری را نسبت به مدل اتوبتا-دوجمله‌ای پیش‌گویی کرده است. علت این مسئله نیز بیش برازش بودن مدل اتو دوجمله‌ای است که از توانایی آن در پیش‌گویی دقیق‌تر، می‌کاهد.

واژه‌های کلیدی: نقشه پهنه‌بندی بیماری، مدل اتو دوجمله‌ای، مدل اتوبتا-دوجمله‌ای، سرطان معده.

رده‌بندی موضوعی (۲۰۱۰): ۶۲P۱۰، ۶۲J۰۵.

۱- آدرس الکترونیکی نویسنده مسئول مقاله: hbaghishani@shahroodut.ac.ir

۱- مقدمه

همزمان با رشد روزافزون اطلاعات مربوط به بیماری‌ها، روش‌های مناسب برای تحلیل آن‌ها که پاسخ‌گوی نیازهای مختلف باشد نیز رو به گسترش است. یکی از این روش‌ها، پهنه‌بندی وضعیت بیماری یا مرگ‌ومیر حاصل از آن است که توزیع جغرافیایی بیماری‌ها یا مرگ را در کنار دیگر عوامل خطر در نظر می‌گیرد. پهنه‌بندی بیماری یا مرگ‌ومیر به مجموعه‌ای از روش‌های آماری اطلاق می‌شود که هدف آن به دست آوردن برآوردهایی دقیق از میزان بروز یا شیوع بیماری‌ها یا مرگ‌ومیر و تنظیم آن‌ها در قالب نقشه‌های جغرافیایی است [۱].

امروزه پهنه‌بندی و برآورد خطر بیماری‌ها موردتوجه فعالان و برنامه‌ریزان بخش سلامت جامعه است، چراکه توزیع جغرافیایی میزان‌های بروز، شیوع و مرگ ناشی از بیماری‌ها می‌تواند نقش مهمی در تخصیص عوامل خطر و پیش‌گیری از آن‌ها بازی کند. تحلیل جغرافیایی نرخ‌های بیماری علاوه بر مدل‌بندی و ارزیابی پذیره‌های سبب‌شناختی و اعمال مداخله در مناطقی که نیازمند توجه خاص هستند، می‌تواند نقش مهمی درزمینه‌ی تخصیص منابع، امکانات و نیروی انسانی ایفا کند.

شرایط جوی و محیطی در هر منطقه زمینه را برای بروز و شیوع برخی بیماری‌ها مساعد می‌کنند. سرطان نیز ازجمله بیماری‌های کشنده و نگران‌کننده جامعه امروزی است که انسان با آن دست‌وپنجه نرم می‌کند. عوامل عمده تأثیرگذار بر سرطان را محیطی و ژنتیکی می‌دانند و میزان بروز انواع آن در نواحی جغرافیایی مختلف، متفاوت است [۲]. در سال‌های اخیر در کشور ما این میزان افزایش پیداکرده است. به‌ویژه در استان گیلان، سرطان معده شیوع بارزی داشته است و هرسال جان صدها نفر را به خطر می‌اندازد. همچنین استان گیلان ازنظر فراوانی سرطان معده مقام اول را در کشور دارد [۳].

سرطان معده رشد بدون کنترل سلول‌های بدخیم در معده است که در آن بیش‌تر افراد تا مراحل پیشرفته بیماری، علامتی ندارند. سلول‌های سرطانی معمولاً به‌مرور رشد می‌کنند و تبدیل به تومور می‌شوند. این سرطان در استان گیلان بسیار شایع است، به‌طوری‌که به گفته یکی از مسئولان مرکز تحقیقات گوارش و کبد بیمارستان رازی رشت، هر دو روز یک مورد سرطان جدید معده تشخیص داده می‌شود و از هر سه‌نفری که بر اثر ابتلا به سرطان فوت می‌کنند، یک نفر مبتلا به سرطان معده است [۴].

نسبت بالایی از سرطان‌ها به‌طور مستقیم با عوامل محیطی ارتباط دارند و لازم است که عوامل مؤثر در این مورد تعیین و مشخص شوند. از هر ۹ مرگ در جهان یک مورد مربوط به سرطان است. سرطان بعد از بیماری‌های قلبی و عروقی، دومین علت مرگ در جوامع انسانی است و سومین علت مرگ بعد از بیماری‌های قلبی و عروقی و تصادفات در کشور ما است [۴]. با توجه

به رابطه بین بروز انواع سرطان و شرایط جغرافیایی منطقه‌ای، تفاوت‌های آشکاری در میزان شیوع و فراوانی هرکدام از سرطان‌ها در مناطق مختلف مشاهده می‌شوند. به‌عنوان‌مثال، در عرض جغرافیایی بالای ۳۲ درجه شیوع بیماری سرطان مری بالا است و عوامل محیطی، شرایط جغرافیایی و نوع اشتغال را با این بیماری در ارتباط دانسته‌اند [۵]؛ بنابراین انتظار می‌رود که مناطق جغرافیایی نزدیک به هم میزان بروز بیماری مشابهی داشته باشند. در نظر گرفتن این واقعیت در مدل‌بندی احتمالی نرخ بروز بیماری‌ها، توسط ساختارهای وابستگی فضایی انجام می‌شود [۶]. با این مقدمه، تهیه نقشه پیش‌گویی (پهنه‌بندی) جغرافیایی نرخ سرطان معده در سطح شهرستان‌های استان گیلان موردنظر این مطالعه است.

برای بهبود برآوردهای پهنه‌بندی بیماری‌ها تحقیقات مختلفی انجام‌شده‌اند. ابتدایی‌ترین نقشه را می‌توان به اسنو [۷] نسبت داد که در سال ۱۸۵۴ و به دنبال همه‌گیری وبا در شهر لندن تهیه کرد. کلایتون و کالدور [۸] مدل‌های سلسله‌مراتبی و استنباط بیزی تجربی مرتبط با آن را برای نسبت مرگ‌ومیر وقتی‌که همبستگی فضایی بین مشاهدات در نواحی همسایه لحاظ شود، مطرح کردند. در تحلیل داده‌های بیماری که غالباً گسسته و شمارشی هستند، می‌توان کایسر و کرسی [۹] را نام برد که برای تحلیل فضایی داده‌های دارای توزیع پواسون، مدل پواسون فضایی را معرفی کردند. لاجونی [۱۰] مدل هم‌کریگینگ دوجمله‌ای را برای پیش‌گویی نرخ فضایی بر اساس نسبت نمونه‌ای مشاهده‌شده از یک بیماری نادر، پیشنهاد داد. الیور و همکاران [۱۱]، از این مدل برای تحلیل نرخ سرطان کودکان در غرب انگلستان استفاده کردند. مونستیز و همکاران [۱۲] استفاده از مدل کریگینگ پواسون را برای داده‌های شمارشی همبسته فضایی پیشنهاد دادند. آن‌ها در مدل پیشنهادی خود از یک طرح وزن‌دار به‌منظور اختصاص دادن وزن بزرگ‌تر به مشاهدات بزرگ‌تر استفاده کردند. گورث [۱۳] نیز از این روش برای تحلیل داده‌های پزشکی با فرض این‌که تمام مناطق جغرافیایی (مثل استان، یا شهرستان) اندازه یکسان دارند، استفاده کرد و آن را به ناهمگنی فضایی تعمیم داد. گورث [۱۴] و [۱۵]، شائو و همکاران [۱۶] و کری و همکاران [۱۷] نیز به تحلیل فضایی انواع سرطان در کشورهای مختلف پرداخته‌اند.

یکی از مشکلات در خصوص نقشه پهنه‌بندی، انتخاب مدل مناسب برای پیش‌گویی و تهیه نقشه است. در این مطالعه از مدل اتوبتا-دوجمله‌ای فضایی برای این منظور استفاده کرده‌ایم. این مدل با داشتن دو پارامتر شکل برای مدل‌بندی احتمال موفقیت دوجمله‌ای از مدل اتو دوجمله‌ای منعطف‌تر است؛ بااین‌وجود تا جایی که نویسندگان اطلاع دارند، استفاده از نسخه بیزی مدل اتوبتا-دوجمله‌ای برای تحلیل داده‌های نرخ فضایی چندان موردتوجه محققان قرار نگرفته است. علت آن نیز می‌تواند پیچیدگی بیشتر مدل مذکور باشد. به‌عنوان چند نمونه انگشت‌شمار می‌توان به کارهای انجام‌شده در منابع [۱۸] و [۱۹] اشاره کرد.

۲- مدل اتوبتا-دوجمله‌ای

در مواردی که پاسخ موردعلاقه، مانند تعداد افراد مبتلا به سرطان معده در شهرستان‌های مختلف استان گیلان، یک متغیر تصادفی گسسته در یک جامعه متناهی است، مدل معمول برای تحلیل آن مبتنی بر توزیع دوجمله‌ای تعریف می‌شود که تابع احتمال آن برای متغیر پاسخ در ناحیه i ام با موقعیت s_i به‌صورت

$$P(Z(s_i)=z(s_i)\,|\,\{z(s_j):j\neq i\})=\binom{n(s_i)}{z(s_i)}\pi_i(\{z(s_j):j\neq i\})^{z(s_i)}$$
$$(1-\pi_i(\{z(s_j):j\neq i\})^{n(s_i)-z(s_i))} \qquad z(s_i)=0,1,\ldots,n(s_i)$$

است که در آن $\pi(s_i)=\pi_i$ احتمال موفقیت (مثل ابتلا به سرطان) و $1-\pi_i$ احتمال شکست در ناحیه i ام است و می‌تواند از سایر نواحی تأثیرپذیر باشد. همچنین $z(s_i)$ ها مشاهدات در نواحی مختلف هستند. امید ریاضی و واریانس پاسخ $Z(s_i)$ در این مدل به ترتیب برابر با $n_i\pi_i$ و $n_i\pi_i(1-\pi_i)$ است. برای در نظر گرفتن عوامل مؤثر بر احتمال موفقیت در قالب متغیرهای تبیینی رگرسیونی در حضور وابستگی فضایی داده‌ها، پارامتر π_i را معمولاً با استفاده از تابع پیوند لجیت به‌صورت زیر مدل‌بندی می‌کنند:

$$\text{logit}(\pi_i)=\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\eta_i=\boldsymbol{x}_i'\boldsymbol{\beta}+u_i \qquad i=1,\ldots,n$$

که در آن $\boldsymbol{x}_i'=(x_{i1},\ldots,x_{ik})$ بردار متغیرهای تبیینی و $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_k)'$ بردار پارامترهای رگرسیونی است. همچنین بردار $\boldsymbol{u}=(u_1,\ldots,u_n)'$ اثر تصادفی فضایی برای لحاظ کردن وابستگی فضایی بین n ناحیه تحت مطالعه است و با توجه به ماهیت مشبکه‌ای بودن داده‌های فضایی موردنظر این مقاله، با در نظر گرفتن فرض مارکوفی بودن فرآیند تصادفی، از یک مدل اتورگرسیو شرطی[1] (CAR) [۲۰] برای مدل‌بندی آن استفاده می‌کنیم؛ بنابراین فرض می‌کنیم

$$\boldsymbol{u}\sim N(0,\Sigma)$$

که در آن $\Sigma=(I-\rho W)^{-1}D\sigma^2$ به‌طوری‌که ρ و σ^2 ابرپارامترهای مدل CAR هستند، W ماتریس وزن‌های همسایگی نواحی و D یک ماتریس قطری برای منظور کردن واریانس ناهمگن توزیع‌های کناری است.

یک ویژگی مهم این مدل کوچک‌تر بودن واریانس از میانگین است، اما در موارد متعددی این پذیره برقرار نیست و با مسئله بیش پراکنش مواجه می‌شویم. در چنین مواردی، عدم انعطاف

1- Conditional Autoregressive

توزیع دوجمله‌ای استفاده از آن را محدود می‌کند و این مدل نمی‌تواند ماهیت داده‌ها را به‌خوبی برازش دهد. درنتیجه استفاده از آن برای مدل‌بندی داده‌ها منجر به استنباط‌های آماری نادرست در برآورد پارامترها، برآورد ساختار وابستگی و پیش‌گویی فضایی داده‌ها می‌شود.

رهیافتی جانشین برای مرتفع کردن مشکل مطرح‌شده، مدل منعطف اتوبتا-دوجمله‌ای است که در ادامه به معرفی این مدل می‌پردازیم.

۲-۱- مدل بازپارامتری‌شده

برای هر ناحیه فرض کنید

$$Z(s_i)\,|\,\pi(s_i) \sim \mathrm{Bin}(\mathrm{n}(s_i),\pi(s_i))$$

به‌طوری‌که

$$\pi(s_i) \sim \mathrm{Beta}(a(\mathrm{s}_i),\mathrm{b}(\mathrm{s}_i)).$$

بنابراین $Z(s_i)$ دارای توزیع کناری بتا-دوجمله‌ای با تکیه‌گاه $\{0,\ldots,n(s_i)\}$ و تابع احتمال

$$\mathrm{P}(Z(s_i)) = z(s_i)\,|\,\mathrm{a}(s_i),b(s_i)) = \binom{n(s_i)}{z(s_i)} \frac{\mathrm{B}(a(s_i)+\mathrm{z}(s_i),\mathrm{n}(s_i)+b(s_i)-z(s_i))}{\mathrm{B}(a(s_i),b(s_i))}$$

است. میانگین و واریانس $Z(s_i)$ به ترتیب $n(s_i)\frac{a(s_i)}{a(s_i)+b(s_i)}$ و $\frac{n(s_i)a(s_i)b(s_i)(n(s_i)+a(s_i)+b(s_i))}{(\mathrm{a}(s_i)+\mathrm{b}(s_i))^2(1+a(s_i)+b(s_i))}$ هستند. برای وارد کردن اثرات متغیرهای تبیینی رگرسیونی و اثر تصادفی فضایی به مدل، به‌طوری‌که تعبیر ضرایب بر اساس میانگین توزیع بتای مشخص‌شده قابل‌بیان باشد، از یک نسخه بازپارامتری‌شده توزیع بتا [۲۱] استفاده می‌کنیم.

در توزیع بتای تعریف‌شده برای $\pi(s_i)$ با پارامترهای مثبت $a(s_i)$ و $b(s_i)$، میانگین برابر $\mu(s_i)=\frac{a(s_i)}{a(s_i)+b(s_i)}$ است. با قرار دادن $\gamma(s_i)=a(s_i)+b(s_i)$، پارامترهای اصلی به‌صورت $a(s_i)=\gamma(s_i)\mu(s_i)$ و $b(s_i)=\gamma(s_i)(1-\mu(s_i))$ بازنویسی می‌شوند؛ بنابراین

$$\pi(s_i) \sim \mathrm{Beta}(\mu(s_i)\gamma(s_i),(1-\mu(s_i))\gamma(s_i))$$

که دارای میانگین $\mu(s_i)$ و واریانس $\sigma^2(s_i)=\frac{\mu(s_i)(1-\mu(s_i))}{\gamma(s_i)+1}$ است. درنتیجه

$$Z(s_i) \sim \text{Beta-Bin}(n(s_i), \mu(s_i)\gamma(s_i), (1-\mu(s_i)\gamma(s_i)).$$

میانگین و واریانس توزیع بتا-دوجمله‌ای بازپارامتری‌شده حاصل، عبارتند از

$$E(Z(s_i)) = n(s_i)\mu(s_i)$$

و

$$\text{Var}(Z(s_i)) = n(s_i)\mu(s_i)(1-\mu(s_i))\frac{\gamma(s_i)+n(s_i)}{\gamma(s_i)+1}.$$

لازم به ذکر است که شاخص بیش‌پراکندگی برای مدل اتوبتا-دوجمله‌ای تعریف‌شده، توسط $\frac{\gamma(s_i)+n(s_i)}{\gamma(s_i)+1}$ لحاظ می‌شود. با توجه به این شاخص، زمانی که $n(s_i)=1$ یا $\gamma(s_i)\to\infty$ واریانس کناری توزیع بتا-دوجمله‌ای به واریانس دوجمله‌ای همگرا می‌شود. در عمل پارامتر $\gamma(s_i)$ را برای تمام نواحی، ثابت γ در نظر می‌گیرند.

۳- تحلیل بیزی تقریبی مدل اتوبتا-دوجمله‌ای

فرض کنید متغیرهای پنهان $\boldsymbol{y}$ (شامل اثرات تصادفی فضایی $\boldsymbol{u}_i$، پارامترهای رگرسیونی $\boldsymbol{\beta}$ و پیش‌گوهای خطی η_i) دارای توزیع نرمال n_d بعدی با تابع چگالی $f(\boldsymbol{y}\mid\boldsymbol{\theta})$ باشد، به‌طوری‌که بردار میانگین این توزیع صفر و ماتریس دقت (معکوس ماتریس کوواریانس) آن $Q(\boldsymbol{\theta})$ با ابرپارامترهای $\boldsymbol{\theta}$ است. قابل‌توجه است که بردار $\boldsymbol{\theta}$ پارامترهای مدل CAR را نیز شامل می‌شود. با تعریف $\boldsymbol{\psi} = (\gamma, \boldsymbol{\theta}')'$ با بعد m و در نظر گرفتن توزیع پیشین برای آن با تابع چگالی $g(\boldsymbol{\psi})$، تابع چگالی پسین مدل به‌صورت

$$f(\boldsymbol{y},\boldsymbol{\psi}\mid\boldsymbol{z}) \propto \prod_{s_i}\left[\binom{n(s_i)}{z(s_i)}\frac{\text{B}(\mu(s_i)\gamma(s_i)+z(s_i), n(s_i)+1-\mu(s_i)\gamma(s_i)-z(s_i))}{\text{B}(\mu(s_i)\gamma(s_i), (1-\mu(s_i)\gamma(s_i))}\right] f(\boldsymbol{y}\mid\boldsymbol{\psi})g(\boldsymbol{\psi})$$

خواهد شد که در آن $\boldsymbol{z} = (z(s_1),\ldots,z(s_n))'$ بردار مشاهدات است. این توزیع پسین صورت بسته ندارد. در چنین مواردی معمولاً از نمونه‌گیری مونت‌کارلویی برای تقریب توزیع استفاده می‌شود. وجود همبستگی در میدان تصادفی فضایی، معمولاً موجب کاهش کارایی این الگوریتم‌ها، افزایش زمان محاسبات و همگرایی کند الگوریتم می‌شود. برای مرتفع کردن این مشکلات، رو و همکاران [۲۲] روش تقریب لاپلاس آشیانه‌ای جمع‌بسته[1] (INLA) را معرفی

1- Integrated nested Laplace approximation

کردند که در استنباط مدل موردنظر ما جانشین مناسبی برای الگوریتم‌های MCMC است. در این مقاله برای برازش مدل اتوبتا-دوجمله‌ای بیزی از این رهیافت تقریبی استفاده می‌کنیم.

۳-۱- رهیافت تقریبی INLA

روش INLA توسط رو و همکاران هنگامی‌که توزیع پسین مدل صورت بسته نداشته باشد، معرفی شد. در روش INLA انتگرال‌گیری عددی و تقریب لاپلاس به طریقی کارا ترکیب می‌شوند، به‌طوری‌که محاسباتی سریع و با دقت قابل‌قبول جایگزین شبیه‌سازی‌های سنگین MCMC می‌شوند و از مشکلات روش‌های MCMC بر حذر هستند [۲۲].

برای تحلیل بیزی مدل اتوبتا-دوجمله‌ای به روش INLA، لازم است توزیع‌های پسین کناری متغیرهای پنهان y_i و ابرپارامترهای ψ_j به‌صورت

$$f(y_i|\mathbf{z}) = \int f(y_i|\boldsymbol{\psi},\mathbf{z}) f(\boldsymbol{\psi}|\mathbf{z})\,d\boldsymbol{\psi} \qquad i = 1,\ldots,n_d \quad (1)$$

$$f(\psi_j|\mathbf{z}) = \int f(\boldsymbol{\psi}|\mathbf{z})\,d\boldsymbol{\psi}_{-j} \qquad j = 1,\ldots,m \quad (2)$$

محاسبه شوند که در آن $\boldsymbol{\psi}_{-j}$ بردار حاصل از حذف درایه j ام $\boldsymbol{\psi}$ است. روش INLA تقریب‌هایی برای چگالی‌های پسین (۱) و (۲) به‌صورت

$$\tilde{f}(y_i|\mathbf{z}) = \int \tilde{f}(y_i,\mathbf{z},\boldsymbol{\psi})\tilde{f}(\boldsymbol{\psi}|\mathbf{z})\,d\boldsymbol{\psi}, \qquad i = 1,\ldots,n_d$$

$$\tilde{f}(\psi_j|\mathbf{z}) = \int \tilde{f}(\boldsymbol{\psi}|\mathbf{z})\,d\boldsymbol{\psi}_{-j} \qquad j = 1,\ldots,m$$

ارائه می‌کند که با ترکیب تقریب‌های لاپلاس و انتگرال‌گیری‌های عددی به دست می‌آیند [۲۲]. این روش بسیار سریع بوده و تقریب‌هایی دقیق را جایگزین شبیه‌سازی‌های سنگین MCMC می‌کند. تقریب‌های $\tilde{f}(y_i|\mathbf{z})$ و $\tilde{f}(\psi_j|\mathbf{z})$ در سه‌گام به شرح زیر محاسبه می‌شوند:

۱. توزیع پسینی کناری $f(\boldsymbol{\psi}|\mathbf{z})$ به‌صورت

$$\tilde{f}(\boldsymbol{\psi}|\mathbf{z}) = \frac{f(\mathbf{y},\boldsymbol{\psi},\mathbf{z})}{\tilde{f}_G(\mathbf{y}|\boldsymbol{\psi},\mathbf{z})}\Big|_{\mathbf{y}=\mathbf{y}^*(\boldsymbol{\psi})}$$

تقریب زده می‌شود که در آن $\tilde{f}_G(\mathbf{y}|\boldsymbol{\psi},\mathbf{z})$ تقریب گاوسی [۲۳] برای توزیع شرطی کامل $\mathbf{y}$ و $\mathbf{y}^*$ مد توزیع شرطی کامل $\mathbf{y}$ است.

۲. محاسبه $f(y_i,\boldsymbol{\psi},\boldsymbol{z})$ که رو و همکاران [۲۲] سه تقریب گاوسی، لاپلاس و لاپلاس ساده‌شده برای محاسبه آن پیشنهاد دادند. ساده‌ترین روش، تقریب گاوسی و دقیق‌ترین روش، تقریب لاپلاس است. روش تقریب لاپلاس ساده‌شده نیز با از دست دادن کمی از دقت تقریب، ازلحاظ محاسباتی آسان‌تر و سریع‌تر از تقریب لاپلاس عمل می‌کند.

۳. گام سوم ترکیب دو گام پیشین و استفاده از روش انتگرال‌گیری عددی برای رسیدن به هدف است. توزیع پسین کناری به‌صورت

$$\tilde{\pi}(y_i\mid\boldsymbol{z})=\sum_{i=1}^{k}\tilde{\pi}(y_i\mid\boldsymbol{\psi}_k,\boldsymbol{z})\times\tilde{f}(\boldsymbol{\psi}_k\mid\boldsymbol{z})\times\Delta_k$$

به دست می‌آید که در آن k تعداد $\boldsymbol{\psi}$های انتخاب‌شده از تکیه‌گاه و Δ_kها وزن‌های این نقاط هستند. برای جزییات بیشتر در مورد نحوه عملکرد روش INLA می‌توانید به رو و همکاران [۲۲] مراجعه کنید.

۴- تحلیل داده‌ها

مجموعه داده موردنظر، شامل اطلاعات ثبت‌شده بیماران سرطانی در مرکز آموزشی درمانی رازی رشت از اول خرداد ۱۳۹۱ تا پایان اسفند ۱۳۹۵ است. تعداد کل بیماران مبتلا به سرطان معده در این مجموعه داده ۶۴۴ مورد هستند که ۶۲ درصد آن‌ها مرد و ۳۸ درصد زن هستند. متغیر پاسخ، تعداد افراد بیمار در شهرستان‌های مختلف استان گیلان و متغیرهای تبیینی شامل متوسط سن، درصد جنسیتی، نسبت تأهل، نرخ بقای بیماران (نسبت بیماران فوت‌شده)، نرخ بیکاری، سرانه فضای سبز و مدت اقامت بیماران (تعداد روزهای بستری) هستند. تعداد شهرستان‌های استان گیلان ۱۶ شهرستان است. با توجه به این‌که متغیر پاسخ شمارشی و گسسته است و مشاهدات به موقعیت جغرافیایی که در آن قرار دارند وابسته هستند، بنابراین داده‌ها ماهیت فضایی و از نوع مشبکه‌ای دارند.

تمام متغیرهای تبیینی در مدل‌بندی بررسی شدند، اما تنها دو متغیر تبیینی مدت اقامت بیماران و سرانه فضای سبز معنی‌دار شدند که نتایج نهایی و پیش‌گویی تنها با حضور آن‌ها به دست آمدند؛ بنابراین برای تحلیل داده‌های سرطانی با استفاده از هر دو مدل پیشنهادی، پیش‌گوی خطی را به‌صورت

$$\eta_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+u_i\qquad i=1,\cdots,16$$

در نظر گرفتیم که در آن $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2)'$ بردار ضرایب رگرسیونی، x_1 نرخ بستری بیماران در بیمارستان و x_2 نرخ سرانه فضای سبز شهرستان‌های استان است. افزون بر این

$\boldsymbol{u} = (u_{۱},...,u_{۱۶})'$ اثر تصادفی فضایی CAR برای وارد کردن وابستگی‌های مکانی شهرستان‌های استان گیلان است. در مدل CAR در نظر گرفته‌شده برای تحلیل این داده‌ها از یک مدل ذاتی استفاده کردیم که در آن فرض می‌شود $\rho = ۱$ است و تنها ابرپارامتر وابستگی فضایی این مدل، پارامتر واریانس $\sigma^{۲}$ است. البته در روش INLA به‌جای پارامتر $\sigma^{۲}$ و تعریف توزیع پیشین برای آن، از عکس آن‌که پارامتر دقت نامیده می‌شود، استفاده و توزیع پیشین برای آن تعریف می‌شود.

برای تکمیل مدل‌بندی بیزی، باید توزیع‌های پیشین پارامترهای رگرسیونی $\boldsymbol{\beta}$ و ابرپارامترهای دقت $\theta = \sigma^{-۲}$ و بیش‌پراکندگی مدل بتا-دوجمله‌ای γ تعیین شوند. درواقع بردار ابرپارامترهای مدل بتا-دوجمله‌ای فضایی عبارت است از $\boldsymbol{\psi} = (\gamma, \theta)'$. برای این کار، در هر دو مدل اتو دوجمله‌ای و اتوبتا-دوجمله‌ای، از توزیع‌های پیشین استاندارد روش INLA استفاده کردیم؛ برای پارامترهای $\boldsymbol{\beta}$ توزیع‌های پیشین نرمال با میانگین صفر و واریانس بزرگ ۱۰۰۰ تعیین شدند و برای پارامتر دقت مدل CAR توزیع پیشین گاما با پارامترهای شکل ۱ و مقیاس ۱۰۰۰ انتخاب شد. همچنین برای پارامتر γ در مدل اتوبتا-دوجمله‌ای از توزیع پیشین نرمال با میانگین صفر و واریانس $\frac{۱}{۰/۴}$ برای تبدیل $\log(\gamma^{-۱})$ استفاده شد.

۵- نتایج تحلیل

هر دو مدل اتو دوجمله‌ای و اتوبتا-دوجمله‌ای با روش INLA و به کمک بسته R-INLA در نرم‌افزار R، بر روی‌داده‌ها برازش داده شدند. نتایج برازش هر دو مدل در جدول ۱ گزارش‌شده‌اند. همچنین شکل ۱ نمودارهای پراکنش مقادیر واقعی نسبت بیماران مبتلا به سرطان معده را در مقابل مقادیر برازش‌شده از هر دو مدل، نمایش می‌دهد. با توجه به نتایج جدول و نمودارهای پراکنش شکل ۱، می‌توان گفت

- هر دو متغیر تبیینی معنی‌دار هستند.
- برآوردهای پارامترها ازنظر اندازه اثر به هم نزدیک هستند، ولی دقت برآوردها (برحسب انحراف معیارهای برآورد شده و فاصله‌های اعتبار ۸۰ درصد) اختلاف قابل ملاحظه‌ای باهم دارند. با توجه به عدم انعطاف لازم در مسئله مدل‌بندی بیش پراکنش یا کم پراکنش موجود در داده‌ها توسط مدل اتو دوجمله‌ای، انحراف معیارهای کوچک در این مدل می‌تواند نشانه‌ای از کم برآورد کردن دقت واقعی برآوردگرها باشد. از این منظر، مدل اتوبتا-دوجمله‌ای توانایی ارائه توصیف بهتری از مدل را دارد. البته مسئله نیاز به بررسی بیش‌تر دارد.
- طول مدت اقامت بیماران بر نرخ سرطان معده در شهرستان‌های استان گیلان، تأثیر مستقیم (مثبت) دارد. با توجه به تفسیر پارامترها در مدل‌های لجیت، برآورد ۰/۶۲ برای اثر طول مدت

بستری در مدل اتوبتا-دوجمله‌ای، به این معنی است که اگر طول روزهای بستری بیماران ساکن در شهرستانی یک روز افزایش یابد، با توجه به مقدار $\exp(۰/۶۲) = ۱/۸۶$ بخت ابتلا به سرطان معده در آن شهرستان ۸۶ درصد بیشتر است.

- در مقابل، نرخ سرانه فضای سبز در هر دو مدل تأثیر منفی دارد؛ برآورد ۰/۶۱ - برای سرانه فضای سبز در مدل اتوبتا-دوجمله‌ای به این معنی است که اگر نرخ سرانه فضای سبز در شهرستانی یک واحد افزایش یابد، با توجه به مقدار $\exp(-۰/۶۱) = ۰/۵۴$ بخت ابتلا به سرطان معده در آن شهرستان ۴۶ درصد کمتر است.
- با توجه به نمودارهای پراکنش مقادیر واقعی در مقابل برازش‌شده نسبت ابتلا به سرطان در شکل ۱، به‌سختی می‌توان برازش بهتر نسبی مدل اتو دوجمله‌ای را در مقابل مدل اتوبتا-دوجمله‌ای مشاهده کرد. البته این برازش بهتر می‌تواند منبعی برای بیش برازش مدل دوجمله‌ای باشد که در بحث پیش‌گویی فضایی منتج به عملکرد ضعیف خواهد شد. برای بررسی این موضوع، از دو معیار پهنای پیش‌گویی شرطی[1] (CPO) و تبدیل انتگرال احتمال[2] (PIT) که هر دو امتیازاتی از معیار اعتبارسنجی متقابل LOO[3] هستند، استفاده کردیم. این دو معیار به ترتیب به‌صورت زیر تعریف می‌شوند:

$$\mathrm{CPO}_i = f\,(z_i \mid \boldsymbol{z}_{-i})$$
$$\mathrm{PIT}_i = \Pr(z_i^{new} < z_i \mid \boldsymbol{z}_{-i}), \quad i = ۱,\ldots,n$$

که در آن $\boldsymbol{z}_{-i}$ بردار مشاهدات بدون مشاهده iام است. درواقع CPO احتمال پسینی مشاهده مقدار z_i را زمانی که مدل بر اساس همه مشاهدات به‌جز مشاهده z_i برازش داده‌شده است، نشان می‌دهد. مقدار بزرگ آن برازش بهتر مدل به مشاهده iام را نشان می‌دهد و مقدار کوچک آن بیانگر پرت یا مؤثر بودن این مشاهده است. درواقع اختلاف در مقادیر CPO برای مشاهدات نشان از عدم پشتیبانی آن‌ها توسط مدل و درنتیجه ضعف مدل برای برازش داده‌ها است.

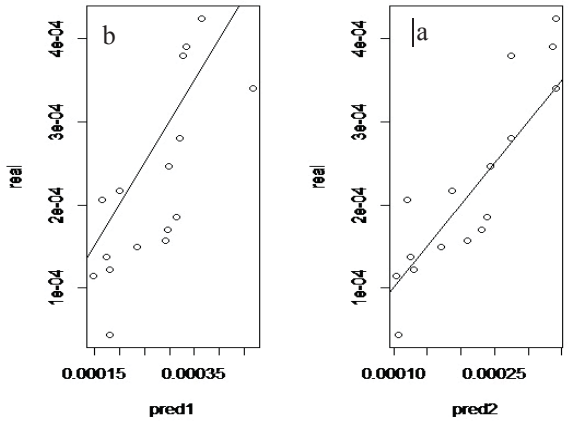
به‌طور مشابه، مقادیر PIT نیز برای یک مدل مناسب باید به‌طور یکنواخت توزیع‌شده باشند. عدم پیروی آن‌ها از توزیع یکنواخت به معنی ضعف قدرت پیش‌گویی مدل است. شکل‌های ۲ و ۳ به ترتیب نمودارهای مقادیر CPO و PIT را برای هر دو مدل نمایش می‌دهند. با توجه به این دو شکل واضح است که توزیع این مقادیر برای مدل اتوبتا-دوجمله‌ای نزدیک‌تر به توزیع یکنواخت است؛ درحالی‌که برای مدل اتو دوجمله‌ای فاصله زیادی از توزیع یکنواخت مشاهده می‌شود؛

1- Conditional Predictive Ordinate
2- Probability Integral Transform
3- Leave-one-Out

بنابراین عملکرد پیش‌گویی مدل اتوبتا-دوجمله‌ای همان‌طور که انتظار داشتیم برتر از مدل اتو دوجمله‌ای است.

جدول (۱): نتایج برازش دو مدل اتوبتا-دوجمله‌ای و اتو دوجمله‌ای برای داده‌های سرطان معده

مدل	اثر	برآورد اثر	انحراف معیار	ناحیه اعتبار ۰/۸۰
اتوبتا-دوجمله‌ای	ضریب ثابت	-۸/۳۷	۰/۲۳	(-۸/۷۰, − ۷/۹۴)
	نرخ مدت اقامت	۰/۶۲	۰/۳۸	(۰/۱۳, ۱/۱۰)
	نرخ فضای سبز	-۰/۶۱	۰/۴۰	(-۱/۱۳ , -۰/۱۰)
اتو دوجمله‌ای	ضریب ثابت	-۸/۵۴	۰/۱۷	(-۸/۶۹, -۸/۴۱)
	نرخ مدت اقامت	۰/۷۳	۰/۱۸	(۰/۵۲, ۰/۹۵)
	نرخ فضای سبز	-۰/۶۹	۰/۱۸	(-۰/۹۲, -۰/۴۶)

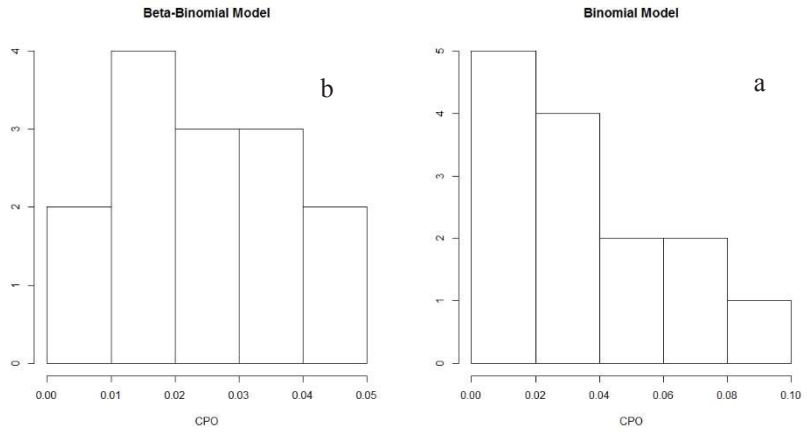


شکل (۱): نمودار پراکنش مقادیر برازش‌شده در مقابل مقادیر واقعی نرخ سرطان معده در دو مدل اتو دوجمله‌ای (a) و اتوبتا-دوجمله‌ای (b)

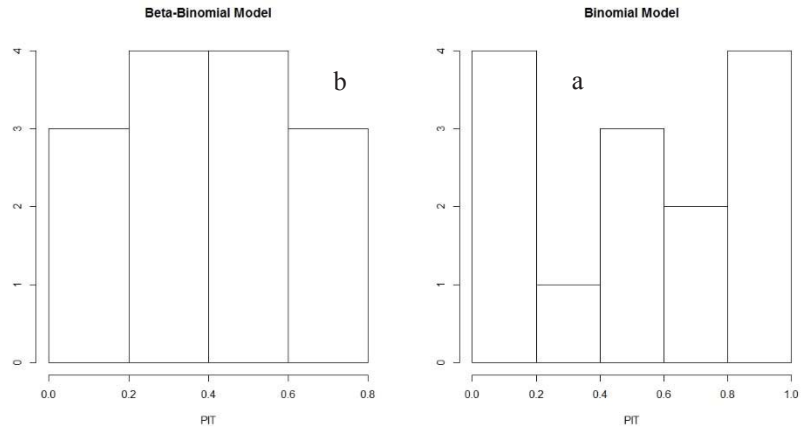
با توجه به آن‌که حجم نمونه داده‌ها کوچک هستند و استفاده از معیارهای اعتبارسنجی متقابل برای برآورد عملکرد پیش‌گویی مدل‌ها توصیه نمی‌شود (ایساکسون و همکاران [۲۴] و ورکویکس [۲۵]، از رهیافت مجموعه‌های آموزشی و آزمون برای برآوردی از عملکرد پیش‌گویی نیز استفاده کردیم. برای این منظور به‌طور تصادفی مقادیر پاسخ دو شهرستان آستانه اشرفیه و لنگرود را به‌عنوان مجموعه آزمون کنار گذاشتیم و با بقیه داده‌ها دو مدل را برازش دادیم. سپس از روی مدل‌های برازش‌شده این دو مقدار را پیش‌گویی کردیم. مقادیر میانگین توان دوم خطای

پیش‌گویی، MSPE، برای دو مدل اتو دوجمله‌ای و اتوبتا-دوجمله‌ای به ترتیب برابر ۰/۰۱۸ و ۰/۰۰۴ به دست آمدند. درواقع مدل اتوبتا-دوجمله‌ای بیش از چهار برابر مدل اتو دوجمله‌ای در پیش‌گویی تواناتر و دقیق‌تر عمل کرده است. همان‌طور که اشاره کردیم، دلیل این نتیجه بیش برازش بودن مدل اتو دوجمله‌ای در مقابل مدل اتوبتا-دوجمله‌ای است.

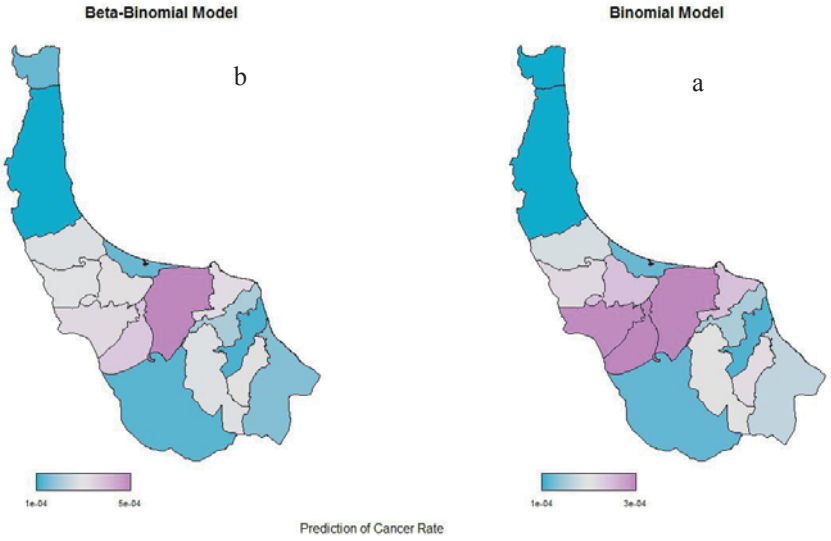
نقشه‌های پهنه‌بندی حاصل از دو مدل برای پیش‌گویی نرخ سرطان معده در استان گیلان در شکل ۴ نشان داده‌شده‌اند. در این شکل مقادیر پیش‌گویی مدل اتوبتا-دوجمله‌ای در فاصله (۰/۰۰۰۱۴،۰/۰۰۰۵۴) و برای مدل اتو دوجمله‌ای در فاصله (۰/۰۰۰۱۴،۰/۰۰۰۳۴) تغییر می‌کنند. همان‌طور که از نقشه‌های پهنه‌بندی مشخص است، مقادیر پیش‌گویی نرخ ابتلا به سرطان در اغلب شهرستان‌های استان گیلان توسط هر دو مدل مشابه هستند؛ در شهرستان‌هایی که اختلاف وجود دارد، مدل اتو دوجمله‌ای نرخ بزرگ‌تری را نسبت به مدل اتوبتا-دوجمله‌ای پیش‌گویی کرده است. ریشه ممکن این مسئله را نیز در بیش برازش بودن مدل اتو دوجمله‌ای مطرح کردیم؛ بنابراین، با توجه به توانایی بیش‌تر مدل اتوبتا-دوجمله‌ای در پیش‌گویی، بر استفاده از نتایج مبتنی بر مدل اتوبتا-دوجمله‌ای هم در توصیف عوامل مؤثر بر افزایش یا کاهش نرخ سرطان و هم پیش‌گویی، تأکید می‌کنیم.



شکل (۲): نمودارهای هیستوگرام CPO برای مدل‌های اتو دوجمله‌ای (a) و اتوبتا-دوجمله‌ای (b)



شکل (۳): نمودارهای هیستوگرام PIT برای مدل‌های اتو دوجمله‌ای (a) و اتوبتا-دوجمله‌ای (b)



شکل (۴): نقشه‌های پهنه‌بندی نرخ سرطان معده برای مدل‌های اتو دوجمله‌ای (a) و اتوبتا-دوجمله‌ای (b)

۶- بحث و نتیجه‌گیری

هدف اصلی این تحقیق، پیش‌گویی فضایی نرخ سرطان معده در استان گیلان، بررسی تأثیر متغیرهای تبیینی و ارائه نقشه پهنه‌بندی بیماری در سطح استان بود که می‌تواند به‌عنوان یک راهنما برای برنامه‌ریزان بهداشتی در پیشگیری از یک‌سو و زمینه‌ای برای ایده‌های مطالعات

علت‌شناسی بیماری از سوی دیگر باشد. علاوه بر این، با توجه به این‌که سرطان معده در استان گیلان بیش‌ترین فراوانی را در کل ایران دارد و شرایط محیطی و اقلیمی در هر منطقه زمینه را برای بروز و شیوع سرطان معده مساعد می‌کند، توجه به مطالعه محیطی برای پیش‌گویی نرخ سرطان، بررسی تأثیر متغیرهای تبیینی و نقشه پهنه‌بندی آن موردتوجه این مقاله قرار گرفت. اگرچه میزان بروز این بیماری در بعضی از جوامع پیشرفته به دلیل مداخلات مناسب مانند آموزش بهداشت و درزمینه‌ی تغذیه‌شناسی و کنترل رفتارهای مستعد کننده در حال کاهش است، اما در کشورهای درحال‌توسعه به علت افزایش سن، فرهنگ نامناسب، تغذیه و عدم کنترل رفتارهای نامناسب مانند استعمال دخانیات و الکل، در حال افزایش است. ازآنجایی‌که به نظر می‌رسد مناطق نزدیک به هم ازلحاظ موقعیت جغرافیایی، نرخ بیماری یا مرگ‌ومیر مشابهی داشته باشند، مناسب است که الگوی فضایی در پیش‌گویی و تحلیل آن‌ها منظور شود. در نظر گرفتن وابستگی فضایی موجب دقیق‌تر شدن برآوردها و پیش‌گویی می‌شود. در این مقاله با در نظر گرفتن وابستگی فضایی بین شهرستان‌های استان گیلان، نرخ سرطان معده را با استفاده از دو مدل اتو دوجمله‌ای و اتوبتا-دوجمله‌ای تحلیل کردیم. نتایج این بررسی نشان می‌دهد که عامل محیطی فضای سبز با این بیماری ارتباط منفی دارد، به‌طوری‌که با افزایش آن بخت ابتلا به سرطان کاهش می‌یابد. از طرف دیگر، نرخ مدت بستری دارای رابطه مثبت با این بیماری است. با توجه به نقشه پهنه‌بندی، مشاهده می‌شود که هر دو مدل در اکثر شهرستان‌های استان نرخ مشابهی را پیش‌گویی کرده‌اند؛ اما در شهرستان‌هایی که اختلاف وجود دارد، مدل اتو دوجمله‌ای نرخ بزرگ‌تری را برای سرطان پیش‌گویی کرده است و علت آن نیز بیش برازش بودن این مدل است و نباید به آن استناد کرد. نتایج پیش‌گویی نرخ بالا یا پایین سرطان را باید در عوامل بروز سرطان در این شهرستان‌ها جستجو کرد.

سپاس‌گزاری

از همکاری معاونت محترم درمان دانشگاه علوم پزشکی گیلان و مرکز آموزشی درمانی رازی رشت برای در اختیار قرار دادن داده‌های مربوط به سرطان معده تشکر و قدردانی می‌کنیم. همچنین از داوران محترم که با نظرات سازنده خود به بهتر شدن مقاله کمک کردند، قدردانی می‌کنیم.

منابع

[1] Rao, J. N. (2015). *Small-Area Estimation*. John Wiley and Sons, Ltd.

[2] Wilkinson, D., and Tanser, F. (1999). GIS/GPS to document increased access to community-based treatment for tuberculosis in Africa. *The Lancet*, **354**(9176), 394-395.

[۳] رمضانی، بهمن؛ حنیفی، اعظم (۱۳۸۷). شناخت پراکندگی جغرافیایی شیوع سرطان معده در استان گیلان. *فصلنامه علوم و تکنولوژی محیط‌زیست*، دوره‌ی ۱۳، شماره‌ی ۲، ص۹۲-۷۹.

[۴] رضوانی، محمود و همکاران ۱۳۷۴، طرح ثبت سرطان در استان گیلان، معاونت بهداشتی استان گیلان، ص ۱.

[5] Ali, M., Rasool, S., Park, J. K., Saeed, S., Ochiai, R. L., Nizami, Q., and Bhutta, Z. (2004). Use of satellite imagery in constructing a household GIS database for health studies in Karachi, Pakistan. *International Journal of Health Geographics*, **3**(1), 20.

[6] Cressie, N. (1993). *Statistics for Spatial Data: Revised Edition*. John Wiley and Sons.

[7] Snow, J. (1854). The cholera near Golden-square, and at Deptford. *Medical Times and Gazette*, **9**, 321-322.

[8] Clayton, D., and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671-681.

[9] Kaiser, M. S., and Cressie, N. (1997). Modeling Poisson variables with positive spatial dependence. *Statistics and Probability Letters*, **35**(4), 423-432.

[10] Lajaunie, C. (1991). *Local risk estimation for a rare noncontagious disease based on observed frequencies*. Note N-36/91/G, Centre de Géostatisque, Ecole des Mines de Paris.

[11] Oliver, M. A., Webster, R., Lajaunie, C., Muir, K. R., Parkes, S. E., Cameron, A. H., and Mann, J. R. (1998). Binomial cokriging for estimating and mapping the risk of childhood cancer. *Mathematical Medicine and Biology: A Journal of the IMA*, **15**(3), 279-297.

[12] Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J. P., and Guinet, C. (2006). Geostatistical modelling of spatial distribution of Balaenoptera physalus in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. *Ecological Modelling*, **193**(3-4), 615-628.

[13] Goovaerts P. (2010). Geostatistical Analysis of County-Level Lung Cancer Mortality Rates in the Southeastern United States, *Geographical analysis*, **42**(1), 32-52.

[14] Goovaerts, P. (2005). Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *International Journal of Health Geographics*, **4**(1), 31.

[15] Goovaerts, P. (2009). Medical geography: a promising field of application for geostatistics. *Mathematical Geosciences*, **41**(3), 243.

[16] Shao, C., Mueller, U., and Cross, J. (2009). Area-to-point Poisson kriging analysis for lung cancer in Perth areas. *Proceedings of the 18th World IMACS/MODSIM Congress*, Jul 13-17, Caire, Australia.

[17] Kerry, R., Goovaerts, P., Smit, I., and Ingram, P. R. (2010). A Comparison of Indicator and Poisson Kriging of Herbivore Species Abundance in Kruger National Park. *South Africa [Online]*.

[18] 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**(2), 85-102.

[19] Harrison, X. A. (2015). A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology and evolution, *PeerJ*, **3**, e1114.

[20] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B (Methodology),* **36**(2), 192-236.

[21] Ferrari, S., and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799-815.

[22] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B (Methodology)*, **71**(2), 319-392.

[23] Rue, H., and Held, L. (2005). *Gaussian Markov random fields: theory and applications,* CRC press, London.

[24] Isaksson, A., Wallman, M., Göransson, H., and Gustafsson, M. G. (2008). Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, **29**(14), 1960-1965.

[25] Varoquaux, G. (2017). Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage*, **180**, 68-77.

Bayesian Analysis of Gastric Cancer Rate in Gilan Province by Using the Auto-Beta Binomial Model

Leila Abedinpour Liiarjadmeh, Hossein Baghishani, Negar Eghbal

Departmant of Statistics, Shahrood University of Technology, Shahrood, Iran

Abstract

The climatic and environmental conditions in each region contribute to the outbreak of certain diseases. Therefore, providing a map of the event rate of a disease or mortality from various diseases on a geographic area is one issue of concern for physicians and health experts. Considering that gastric cancer is the most common cancer in Gilan province, Iran, in this paper, we study the impact of some risk factors on the rate of this cancer for the cities of Gilan province by using two auto-binomial and auto-beta-binomial Bayesian spatial models. The other purposes of this study are providing the gastric cancer rate prediction map and comparing the performance of the two proposed models. We used a dataset from the Razi Educational Center of Rasht in which the data were collected for sixteen cities of Gilan during the period of 2012-2017. We fitted the proposed models for these data by using an approximate Bayesian approach, called the integrated nested Laplace approximation (INLA). Based on the results, it was found that prediction of the rates of cancer in most of the cities of Gilan are similar by using of both models; in cities where there is a difference, the auto-binomial model predicts a higher rate than the auto-beta-binomial model. The reason for this is also that the auto-binomial model is over-fitted, which reduces its ability to predict.

***Keywords*:** Zoning map, Binomial model, Beta-binomial model, Gastric cancer.

Mathematics Subject Classification (2010): 62P10, 62J05.