باسمه تعالی

مجموعه مقالات

دومین کارگاه آموزشی

آمار فضایی و کاربردهای آن

قطب علمی داده های ترتیبی و فضایی دانشگاه فردوسی مشهد

با همکاری

گروه های آمار دانشگاه تربیت مدرس و دانشگاه فردوسی مشهد

و انجمن آمار ایران

مشهد- ایران

۱۰ تا ۱۱ خردادماه ۱۳۹۱

پیشگفتار

با یاری خداوند متعال و همکاری انجمن آمار ایران و همت اعضای محترم هستهٔ قطب علمی داده‌های ترتیبی و فضایی دانشگاه فردوسی مشهد و اعضای کمیته علمی، دومین کارگاه آموزشی **آمار فضایی و کاربردهای آن** در روزهای ۱۰ تا ۱۱ خرداد ماه ۱۳۹۱ در دانشگاه فردوسی مشهد برگزار می‌شود. استقبال اساتید، دانشجویان و محققین در زمینه آمار فضایی در ارسال مقالات موجب گردید این کارگاه از سطح علمی مطلوبی برخوردار شود و زمینه‌های ارتباط بین محققین در زمینه آمار فضایی فراهم گردد.

دبیر دومین کارگاه آموزشی
آمار فضایی و کاربردهای آن

خرداد ۱۳۹۱

اعضای کمیته اجرایی:

۱- دکتر جعفر احمدی
۲- دکتر هادی جباری نوقابی
۳- دکتر محسن محمدزاده
۶- دکتر سید محمد حسینی

اعضای کمیته علمی و داوران:

۱- دکتر محسن محمد زاده (دبیر کارگاه)	دانشگاه تربیت مدرس
۲- دکتر امید کریمی	دانشگاه سمنان
۳- دکتر حسین باغیشنی	دانشگاه شاهرود
۴- دکتر نصرا... ایران‌پناه	دانشگاه اصفهان
۵- دکتر فاطمه حسینی	دانشگاه سمنان
۶- آقای بهزاد محمودیان	دانشگاه تربیت مدرس

اعضای کمیته اجرایی دانشجویی:

۱- آقای خیراله اخلی
۲- آقای حامد احمدزاده
۳- آقای حسینعلی محتشمی

فهرست مندرجات

عنوان صفحه

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱
مجموعه مقالات، ص ۱-۱۴

مقایسه مدل پواسن کریگینگ و مدل بیز کامل در تحلیل فضایی میزان بروز سرطان گوارش در ایران

نعیمه السادات اثماریان[۱]، امیر کاوسی[۲]، مسعود صالحی[۳]

[۱] *گروه آمار زیستی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی*
[۲] *گروه علوم پایه دانشکده سلامت، ایمنی ومحیط زیست دانشگاه علوم پزشکی شهید بهشتی*
[۳] *گروه آمار زیستی، دانشگاه علوم پزشکی تهران*

چکیده : پهنه‌بندی میزان رویداد یک بیماری یا مرگ و میر ناشی از بیماری‌های مختلف بر روی نقشه جغرافیایی یکی از موضوعات مورد علاقه متخصصان و برنامه‌ریزان امور بهداشتی بخصوص اپیدمیولوژیست‌ها می‌باشد. یکی از معروف‌ترین مدل‌های برآورد پارامترهای نقشه در سال‌های اخیر، مدل بیزکامل است. در این مقاله به‌منظور رفع نقایص مدل بیز کامل در مناطق ناهمگن فضایی، به معرفی و بررسی مدل پواسن کریگینگ که روشی نوین و کاربردی در تحلیل داده‌های فضایی است، پرداخته می‌شود. این مدل در مناطق ناهمگون فضایی هموارسازی کمتر و دقت بالاتری در برآورد میزان بیماری نسبت به مدل بیز کامل دارد. با استفاده از این دو مدل، اقدام به برآورد میزان بروز سرطان گوارش به همراه دقت (واریانس) این برآوردها و ارائه نقشه پهنه‌بندی آن‌ها در سطح نقشه ایران شده

آدرس الکترونیک مسئول مقاله: نعیمه السادات اثماریان، ns.asmarian@gmail.com
کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲H۱۱

است. داده‌های مورد استفاده در این تحقیق شامل کلیه داده‌های ثبت شده توسط اداره سرطان مرکز مدیریت بیماری‌های غیر واگیر وزارت بهداشت درمان و آموزش پزشکی در کل کشور ایران در سطح کلیه‌ی شهرستان‌ها از سال ۱۳۸۲ تا ۱۳۸۶ می‌باشد.

واژه‌های کلیدی : آ مار فضایی، زمین‌آمار، پواسن کریگینگ، بیز کامل، پهنه‌بندی بیماری، سرطان گوارش

۱ مقدمه

هم‌زمان با رشد روزافزون اطلاعات پیرامون بیماری‌ها و مرگ و میر، روش‌های متناسب برای تحلیل این نوع داده‌ها که پاسخگوی نیازهای مختلف باشد، نیز رو به گسترش است. یکی از این روش‌ها، پهنه‌بندی بیماری یا مرگ و میر است که توزیع جغرافیایی بیماری‌ها یا مرگ را در کنار دیگر عوامل خطر در نظر می‌گیرد. پهنه‌بندی بیماری[۱] یا مرگ و میر به مجموعه‌ای از روش‌های آماری اطلاق می‌شود که هدف آن‌ها به‌دست آوردن برآوردهایی دقیق از میزان‌های بروز یا شیوع بیماری‌ها یا مرگ و میر و تنظیم آن‌ها در قالب نقشه‌های جغرافیایی می‌باشد.

امروزه پهنه‌بندی و برآورد خطر بیماری‌ها مورد توجه فعالان و برنامه‌ریزان عرصه سلامت جامعه می‌باشد چرا که توزیع جغرافیایی میزان‌های بروز، شیوع و مرگ و میر نقش مهمی در تشخیص عوامل خطر و پیش‌گیری از آن‌ها را بازی می‌کند. تحلیل جغرافیایی نرخ‌های بیماری علاوه‌بر فرمول‌بندی و ارزیابی فرضیات سبب‌شناختی و اعمال مداخله در مناطقی که نیازمند توجه خاص هستند می‌تواند نقش مهمی در زمینه تخصیص منابع، امکانات و نیروی انسانی ایفا نماید.

سرطان یک مسئله اصلی سلامت عمومی در علم بهداشت است. علی‌رغم تلاش‌های زیادی که برای کاهش مرگ و میر ناشی از سرطان در سال‌های اخیر شده است، هنوز انواع سرطان دومین علت مرگ و میر در دنیا به حساب می‌آید. سرطان گوارش هم یکی از مهم‌ترین آن‌ها می‌باشد که پزشکان معتقدند عوامل محیطی و

[۱] Disease mapping

زیستی در بروز آن دخالت دارد. برآورد دقیق این سرطان‌ها برای هر منطقه به‌ویژه تهیه نقشه و پهنه‌بندی آنها روی نقشه جغرافیایی برای مدیران و تصمیم‌گیرندگان امور سلامت و بهداشت جامعه از اهمیت بالایی برخوردار است. به همین علت بررسی، تحلیل فضایی و پهنه‌بندی این سرطان روی نقشه جغرافیایی ایران موضوع مورد توجه این مقاله قرار گرفت.

در سال‌های اخیر بیشترین کاربرد برای پهنه‌بندی بیماری‌ها را مدل بیز کامل[۲] با استفاده از اطلاعات واحدهای مجاور منطقه مورد نظر به خود اختصاص داده است. به‌دلیل اینکه این روش الگوی فضایی را در مدل لحاظ نمی‌کند در مناطقی که ناهمگنی فضایی دارند از دقت خوبی برخوردار نیست. در خصوص بررسی پهنه‌بندی بیماری در مناطقی که با جمعیت‌های پراکنده و یا بیماری‌های نادر روبه‌رو هستیم باید به دنبال روشی باشیم که اغتشاش[۳] را کاهش دهد. در مدل بیز کامل برای برآورد پارامترها از روش‌های تکراری مانند روش زنجیر مارکف مونت کارلو[۴] استفاده می‌شود که احتیاج به رایانه‌های پر قدرت و محاسبات دقیق دارد که کاربرد و تفسیر را برای غیر آماری‌ها مشکل می‌سازد.

با وجود این دلایل، در این پژوهش دنبال به‌کارگیری روشی کاربردی در آمار فضایی[۵] برای تهیه نقشه‌ی با کیفیت و دقت بالا از بیماری‌ها هستیم و به‌طور خاص روش پواسن کریگینگ[۶] (نوعی درونیاب فضایی برای متغیرهای نادر مانند انواع سرطان) را که از کاراترین روش‌های آمار فضایی می‌باشد، مورد توجه قرار دادیم. لازم به توضیح است که روش کریگینگ به‌عنوان بهترین پیشگوی خطی نااریب که پیش از روش پواسن کریگینگ ارائه شده است، برای تحلیل داده‌های زمین آماری[۷] (داده‌هایی که امکان تحقق و اندازه‌گیری آن‌ها در هر نقطه از منطقه مورد مطالعه وجود دارد) به‌کار می‌رود و به‌طور وسیع مورد اقبال و کاربرد رشته‌های هواشناسی،

---

۲ Full Bayes

۳ Noise

۴ Mont Carlo Markov Chain (MCMC)

۵ Spatial Statistic

۶ Poisson kriging

۷ Geostatistical

زمین‌شناسی و معدن قرار گرفته است، اما روی داده‌های پزشکی به ویژه داده‌های سلامت و بهداشت که عمدتاً از نوع داده‌های گسسته و شمارشی می‌باشند، قابل کاربرد نیست، در حالی‌که پواسن کریگینگ با بهره‌گیری از روش کریگینگ برای تحلیل این نوع داده‌ها قابل استفاده است.

در مورد روند بهبود برآورد پارامترهای پهنه‌بندی بیماری می‌توان به تحقیقات انجام شده اشاراتی داشت. ابتدایی‌ترین نقشه را می‌توان به جان اسنو نسبت داد که در سال ۱۸۵۴ و به‌دنبال همه‌گیری وبا در شهر لندن تهیه کرد. کلایتون و کالدور برای اولین بار در سال ۱۹۸۷ مدل‌های سلسله مراتبی و استنباط بیز تجربی مرتبط با آن را برای میزان‌های مرگ و میر استاندارد شده در حالتی‌که همبستگی فضایی بین مشاهدات در نواحی همسایه نیز لحاظ می‌شود، مطرح کردند. منظور از مدل بیزی که بیشترین کاربرد را در پهنه‌بندی بیماری در سال‌های اخیرداشته است همان مدل معرفی شده توسط بیساگ، یورک و مولیه در سال ۱۹۹۱ معروف به مدل BYM است.

در تحلیل داده‌های بیماری که غالباً گسسته و شمارشی هستند می‌توان کایسر و همکارانش (۱۹۹۷) را نام برد که برای تحلیل فضایی داده‌های دارای توزیع پواسن، مدل توزیع پواسن فضایی را معرفی نمودند. الیور و همکارانش (۱۹۹۸) کوکریگینگ دوجمله‌ای را برای تهیه پهنه‌بندی خطر سرطان کودکان در غرب انگلیس به‌کار بردند. مونستیز و همکارانش (۲۰۰۶-۲۰۰۴) کوشش کردند پواسن کریگینگ را برای مدل‌سازی مشاهدات ناهمگن فضایی توسعه دهند. روش به‌کار رفته توسط مونستیز و همکارانش مشابه روش کوکریگینگ دوجمله‌ای پیشنهاد شده الیور و همکارانش بود، تنها تفاوتش این بود که از توزیع پواسن برای داده‌های قابل شمارش سرطان به‌جای دوجمله‌ای استفاده می‌شد. ابتدا گوورتس این روش را با فرض اینکه تمام واحدهای جغرافیایی اندازه یکسان دارند انجام داد ولی بعد این روش را به ناهمگنی فضایی تعمیم داد. گوورتس و همکارانش (۲۰۱۱-۲۰۰۶) مقالات زیادی در این زمینه نوشته و به بررسی انواع سرطان در کشورهای مختلف پرداخته‌اند. از جمله مقاله‌ای قابل توجه با عنوان مقایسه مدل پواسن کریگینگ و مدل بیز کامل برای پهنه‌بندی میزان بیماری (۲۰۰۸) منتشر کردند، که این دو مدل را در دو ایالت متفاوت (ایالتی در هند با ۹۲ شهرستان که واحدهای جغرافیایی دارای

شکل و اندازه‌ی یکسان و دیگری ایالتی در غرب آمریکا با ۱۱۸ شهرستان که واحدها کاملاً ناهمگنی فضایی داشتند) انجام دادند و نتایج را مقایسه کردند. آنها نتیجه گرفتند، در موقعیت‌هایی که واحدهای جغرافیایی از نظر الگوی فضایی و اندازه جمعیتی یکسان هستند، مدل بیز کامل با مدل پواسن کریگینگ تفاوت چندانی ندارد، ولی در نواحی که واحدها از نظر اندازه و الگوی فضایی متفاوتند مدل پواسن کریگینگ به‌دلیل این‌که به‌آسانی توزیع فضایی جمعیت (ناهمگنی در شکل و اندازه‌ی واحد جغرافیایی) در معرض خطر را در برآورد لحاظ می‌کند از دقت خوبی برخوردار است و هموارسازی کمتری نسبت به مدل بیز کامل اعمال می‌کند.
در این مقاله ابتدا اقدام به معرفی دو مدل و سپس برآورد میزان بروز سرطان گوارش به همراه دقت (واریانس) این برآوردها و ارائه نقشه پهنه‌بندی آن‌ها در سطح نقشه ایران با استفاده از آن‌ها شده است.

۲ روش‌ها

۱.۲ *مجموعه‌ی داده‌های سرطان*

داده‌های مورد استفاده در این تحقیق شامل کلیه داده‌های ثبت شده توسط اداره سرطان مرکز مدیریت بیماری‌های غیرواگیر وزارت بهداشت درمان و آموزش پزشکی در کل کشور ایران از سال ۱۳۸۲ تا ۱۳۸۶ می‌باشد. در این تحلیل واحد جغرافیایی کلیه‌ی شهرستان‌های کشور ایران (به‌عبارتی مجموعه‌ای از ۳۳۶ شهرستان) می‌باشد و جمعیت در معرض خطر برمبنای سال ۱۳۸۵ و میانگین جمعیت – وزن‌دار به‌ازای ۱۰۰۰۰۰ شخص ــ سال درنظر گرفته شده است. واحدهای جغرافیایی از نظر اندازه و شکل مشابه نیستند و جمعیت به‌طور یکنواخت پراکنده نیست، در مرکز و سمت شمال غرب کشور تراکم بیشتری دارد.

۲.۲ مدل پواسن کریگینگ

این مدل در سال ۲۰۰۶ توسط گوورتس برای برآورد میزان بیماری و ایجاد پهنه‌بندی بیماری به‌کار گرفته شد که به اختصار معرفی می‌شود.
فرض کنید متغیر $D(v_i)$ نشان دهنده تعداد افراد سرطانی در ناحیه جغرافیایی با مختصات v_i باشد. چون این متغیر یک متغیر شمارشی و نادر است، بنابراین توزیع $D(v_i)$ را می‌توان پواسن با پارامتر $\theta = n(v_i) \times r(v_i)$ در نظر گرفت که در آن $n(v_i)$ تعداد افراد در معرض خطر و $r(v_i)$ میزان سرطان در ناحیه v_i است، که برای یک ناحیه مشخص v_α مقدار $r(v_\alpha)$ می‌تواند به‌صورت یک ترکیب خطی از میزان سرطان یعنی $Z(v_i) = \frac{d(v_i)}{n(v_i)}$ در k همسایه مجاور به‌صورت

$$\hat{r}_{pk}(v_\alpha) = \sum_{i=1}^{k} \lambda_i(v_\alpha) z(v_i) \qquad (1)$$

برآورد شود که $d(v_i)$ تعداد افراد سرطانی به‌عنوان مقدار مشاهده شده متغیر $D(v_i)$ و $\lambda_i(v_\alpha)$ وزن ناحیه v_iام در برآورد میزان سرطان در ناحیه v_α است که توسط حل دستگاه زیر معروف به سیستم پواسن کریگینگ به‌دست می‌آیند.

$$\sum_{j=1}^{k} \lambda_j(v_\alpha) \left[C(v_i - v_j) + \delta_{ij} \frac{m^*}{n(v_i)} \right] + \mu(v_\alpha) = C(v_i - v_\alpha) \qquad i = 1, \ldots, k$$
$$\sum_{j=1}^{k} \lambda_j(v_\alpha) = 1, \qquad \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad (2)$$

درمعادله (۳)، m^* میانگین جمعیت ـ وزن‌دار است که برای تعداد کل نواحی تحت بررسی (N) به‌صورت زیر محاسبه می‌شود:

$$m^* = \frac{\sum_{\alpha=1}^{N} n(v_\alpha) z(v_\alpha)}{\sum_{\alpha=1}^{N} n(v_\alpha)}$$

و جمله $\frac{m^*}{n(v_i)}$ که واریانس خطا نام دارد، منجربه اختصاص وزن کمتر به جمعیت‌های کوچکتر می‌شود.

در معادلات (۳)، $\mu(v_\alpha)$ ضریب لاگرانژ است که از کمینه کردن واریانس برآورد با قید شرط ناااریبی برآوردگر یعنی $\sum_{j=1}^{k}\lambda_j(v_\alpha) = 1$ ایجاد می‌شود. عبارت $C(v_i - v_j)$ مربوط به کواریانس فضایی (همبستگی فضایی) بین دو ناحیه جغرافیایی i و j یعنی $C(v_i - v_j) = Cov\{Z(v_i), Z(v_j)\}$ است که معادل آن را می‌توان از تابع نیمه تغییرنگار[۸] (واریانس ناشی از فاصله مکانی) به‌صورت زیر استفاده کرد

$$\begin{aligned}\gamma(h) &= C(v_\alpha - v_\alpha) - C(v_i - v_\alpha)\\ &= C(0) - C(h)\end{aligned}$$

که در آن $h = v_i - v_\alpha$ یعنی فاصله مکانی بین دو موقعیت i و α است. نیمه تغییرنگار در عمل نامعلوم است و لازم است براساس مشاهدات برآورد شود، برآورد تجربی آن به‌صورت زیر است که به آن نیم تغییرنگار تجربی گویند.

$$\hat{\gamma}(h) = \frac{1}{2\sum_{\alpha=1}^{N(h)}\frac{n(v_\alpha)n(v_\alpha+h)}{n(v_\alpha)+n(v_\alpha+h)}}\sum_{\alpha=1}^{N(h)}\left\{\frac{n(v_\alpha)n(v_\alpha+h)}{n(v_\alpha)+n(v_\alpha+h)}[z(v_\alpha)-z(v_\alpha+h)]^2 - m^*\right\}$$

که در آن $N(h)$ تعداد جفت مناطقی است که در فاصله h از هم قرار گرفته‌اند. در روش پواسن کریگینگ برای هر نقطه‌ی برآورد شده با استفاده از فرمول (۲) واریانس آن نیز به‌صورت زیر برآورد و ارائه می‌شود.

$$\sigma^2_{pk}(v_\alpha) = C(v_\alpha - v_\alpha) - \sum_{i=1}^{k}\lambda_i(v_\alpha)C(v_i - v_\alpha) - \mu(v_\alpha). \qquad (۳)$$

۳.۲ مدل بیز کامل

این مدل در سال ۱۹۸۷ توسط کلایتون و کالدور مطرح شد و سپس توسط بیساگ و همکارانش در سال ۱۹۹۱ تکمیل و تحت عنوان مدل BYM معرفی گردید. این مدل

[۸] Semivariogram

که یکی از معروف‌ترین مدل‌های بیز سلسله مراتبی می‌باشد، به روش بیز کامل نیز معروف است که در قالب تحلیل داده‌های سرطان معرفی می‌شود.
فرض کنید که y_i متغیر مورد بررسی یک متغیر شمارشی و نادر مانند تعداد افراد سرطانی باشد در این‌صورت توزیع y_i را می‌توان پواسن با پارامتر (میانگین) $e_i\theta_i$ به‌صورت زیر درنظر گرفت

$$y_i \sim Poisson(e_i\theta_i)$$

که در آن y_i تعداد موارد بیماری مشاهده شده در منطقه i است و e_i تعداد مورد انتظار و معلوم است که از فرمول $e_i = n_i(\sum_i y_i / \sum_i n_i)$ با استفاده از استانداردسازی داخلی به‌دست می‌آید؛ که در آن n_i تعداد افراد در معرض خطر بیماری است. θ_i بردار خطر نسبی و نامعلوم است. در مدل بیز کامل که به مدل اتو رگرسیو شرطی[۹] نیز موسوم است خطر نسبی به سه مؤلفه به‌صورت زیر تقسیم می‌شود:

$$\log(\theta_i) = \mu + v_i + u_i$$

که در آن μ میانگین کلی (روند فضایی) و u_i و v_i در مدل‌سازی اثرات تصادفی هستند، که v_i ناهمگنی ناهمبسته فضایی و u_i ناهمگنی همبسته فضایی هستند. در مدل بیز کامل تخصیص توزیع پیشین از اهمیت ویژه‌ای بر خوردار است. بنابراین باید توزیع پیشین اثرات تصادفی و میانگین کلی مشخص شود. جمله v_i مؤلفه ناهمبسته که به مکان جغرافیایی وابسته نیست و فرض می‌شود دارای توزیع نرمال با میانگن صفر و با واریانس τ_v^2 است و توزیع پیشین u_i که مجاورت با همسایه‌ها را درنظر می‌گیرد برطبق اتو رگرسیو شرطی به‌صورت زیر است:

$$u_i \mid u_j, i \neq j\, \tau_u^2 \sim N\left[\frac{\sum_j w_{ij}u_i}{\sum_i w_{ij}}, \frac{\tau_u^2}{\sum_j w_{ij}}\right]$$

$$w_{ij} = \begin{cases} 1 & if\ i,j\ are\ adjacant \\ 0 & otherwise \end{cases}$$

[۹] CAR: Conditional Auto Regressive model

توزیع توأم:

$$prior(\tau_u, \tau_v) \propto e^{\frac{-\varepsilon}{۲\tau_u}} e^{\frac{-\varepsilon}{۲\tau_v}}, \qquad \tau_u, \tau_v > ۰$$

توزیع پسین کامل به‌صورت زیر است

$$\begin{aligned} P(u, v, \tau_u, \tau_v \mid y_i) &= \prod_{i=۱}^{m} \{\exp(-e_i\theta_i)(e_i\theta_i)^{y_i}/y_i!\} \\ &\times \tau^{-\frac{-n}{۲}} \exp\left\{\frac{-۱}{۲\tau_u}\sum_i\sum_j (u_i - u_j)^۲\right\} \\ &\times \exp\left\{\frac{-۱}{۲\tau_u}\sum_{i=۱}^{n} v_i^۲\right\} \times prior(\tau_u, \tau_v) \end{aligned}$$

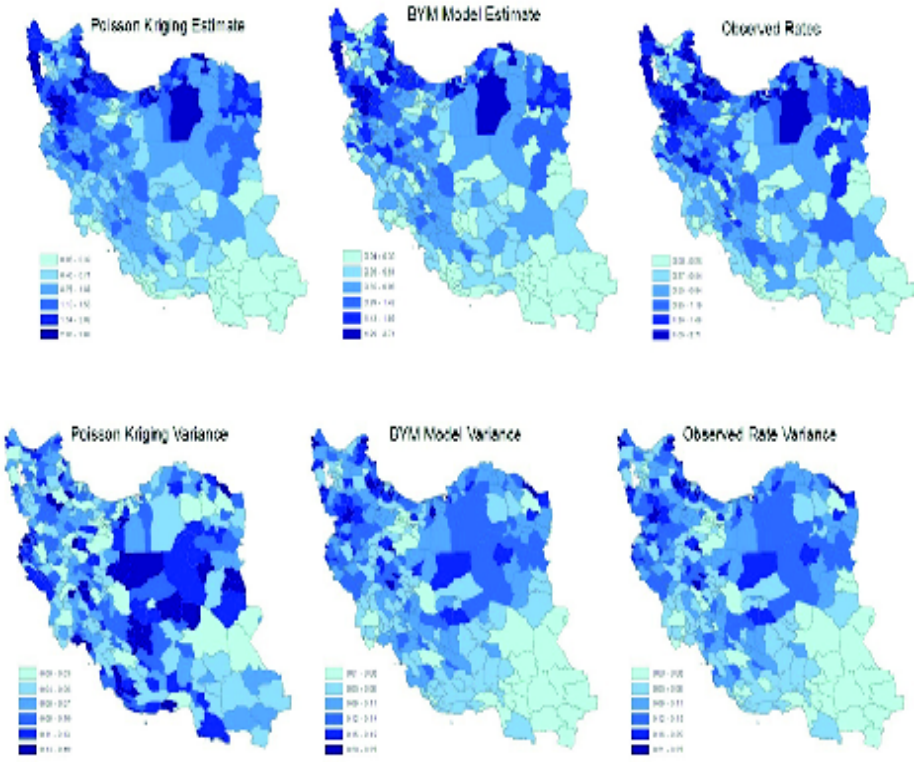
این توزیع پسین با استفاده از الگوریتم‌های زنجیر مارکف مونت کارلو نظیر نمونه گیری‌های گیبز یا متروپولیس ـ هستینگ نمونه‌گیری می‌شود. توزیع ابر پیشین درنظر گرفته شده برای پارامترهای $\tau_u^۲$ و $\tau_v^۲$ براساس نظریه کلسال و ویکفیلد برای داده‌های بیماری $\Gamma(۰/۵, ۰/۰۰۰۵)$ درنظر گرفته شده است.

۳ اجرای مدل‌ها ونتایج

مدل پواسن کریگینگ با نرم‌افزار Space Stat و با انتخاب همسایگی مجاور اجرا شد تا مقایسه دو روش در شرایط یکسان آسان شود. در این نرم‌افزار لازم است که شهرستان‌ها با مختصات جغرافیایی مشخص شوند. در این تحقیق برای بررسی همسانگردی و انتخاب مدل تغییرنگار نظری مناسب، تغیرنگار تجربی میزان بروز سرطان گوارش را در چهار جهت جغرافیایی ۰، ۴۵، ۹۰ و ۱۳۵ درجه بررسی شد و همسانگردی مورد تأیید قرار گرفت و مدل نمایی به داده‌ها برازش داده شد. مدل بیز کامل با نرم‌افزار Open BUGS با استفاده از روش MCMC، نمونه‌گیری گیبز، با یک زنجیر و ۳۰۰۰۰ تکرار انجام شد. همگرایی با آماره گلمن ـ روبین (G-R) تأیید شد. در نهایت با استفاده از نرم‌افزار Arc GIS 9.2 برآوردها بر روی نقشه در شکل (۲) نمایش داده شد. میانگین میزان بروز مشاهده شده براساس نسبت‌های بروز استاندارد شده[۱۰] (۰/۸۲۴) به روش بیز کامل (۰/۸۴۳) به روش پواسن کریگینگ

[۱۰] Standardized Incidence Ratio (SIR)

(۰/۸۹۹) برآورد شده است. میانگین واریانس میزان بروز مشاهده شده (۰/۱۰۲) به روش بیز کامل (۰/۰۹۲) به روش پواسن کریگینگ (۰/۰۷۲) برآورد شده است. بیشترین میزان بروز مشاهده شده (۲/۷۱) با واریانس (۰/۰۹۸) مربوط به شهرستان ساری و کمترین (۰/۰) با واریانس (۰/۰) مربوط به شهرستان دنا بود. بیشترین برآورد میزان بروز با مدل بیز کامل (۲/۷۰) با واریانس (۰/۰۹۴) مربوط به شهرستان ساری و کمترین (۰/۰۴۳) با واریانس (۰/۰۱۸) مربوط به شهرستان سرباز و بیشترین برآورد میزان بروز با مدل پواسن کریگینگ (۲/۸۸۱) و واریانس ۰/۰۲۰ مربوط به شهرستان ساری و کمترین (۰/۰۵۴) با واریانس (۰/۰۶۰۵) مربوط به شهرستان سرباز برآورد شده است.



شکل ۱: پهنه‌بندی برآورد میزان بروز بیماری سرطان گوارش در ایران و نقشه‌ی دقت (واریانس) با دو مدل پواسن کریگینگ و بیز کامل

بحث و نتیجه‌گیری

از آنجایی که روش زمین‌آمار از تحلیل داده‌های معدن به داده‌های پزشکی و بهداشتی با استفاده از مدل پواسن کریگینگ گسترش یافته است، لازم بود ابتدا این مدل مورد بررسی قرارگیرد و با یکی از مدل‌های پر کاربرد در پهنه‌بندی بیماری مقایسه شود، اما مدل‌های دیگری نیز هستند که در تحلیل داده‌های پزشکی به کار می‌روند که می‌توانند با مدل پواسن کریگینگ مقایسه شوند. توجه به اینکه تعداد بیماران سرطان گوارش یک متغیر نادر است و از طرفی نقشه جغرافیایی تقسیم‌بندی شهرستان‌های ایران غالباً شکل نامنظم دارند، مدل پواسن کریگینگ به علت در نظر گرفتن جمعیت و الگوی فضایی نسبت به سایر مدل‌ها مناسب‌تر به‌نظر می‌رسد. برای اینکه بتوان این دو روش را مقایسه کرد شرایط یکسانی از نظر تعداد همسایگی درنظر گرفته شد، با وجود اینکه روش پواسن کریگینگ می‌تواند با درنظر گرفتن همسایه‌هایی که فاصله‌ی بیشتری از منطقه مورد بررسی دارند هموارسازی دقیق‌تری را اعمال کند. مدل بیز کامل در مورد مناطقی که الگوی فضایی یکسان دارند نتایج یکسان با مدل پواسن کریگینگ دارد، اما همان طور که ذکر شد برآورد پارامترهای مدل بیز کامل نیاز به روش‌های شبیه‌سازی دارد که این روش‌های پیچیده، طاقت فرسا بوده و روش‌های تکراری آن احتیاج به زمان زیاد و رایانه‌های پرقدرت دارد که تفسیر نتایج آن‌ها برای غیر آماری‌ها مشکل است.

نتایج این پژوهش نشان می‌دهند مناطق شمال و شمال غربی ایران (به‌ویژه استان‌های گلستان، مازندران و اردبیل) دارای میزان بروز سرطان گوارش بیشتری نسبت به مناطق کویری و جنوبی (به‌ویژه استان‌های کرمان و سیستان و بلوچستان) است، که این نتایج را باید در عوامل بروز سرطان در این مناطق جستجو کرد. در این بررسی از دیدگاه شهرستانی، شهرستان سرباز کمترین و ساری بیشترین میزان بروز را به‌خود اختصاص دادند. متخصصان علوم بهداشت علت بروز سرطان گوارش را فرهنگ غذایی و استفاده از غذاهای ادویه‌دار، کنسروی، سیر و ترشیجات، مصرف طولانی مدت غذاهای نمک‌زده، دودی و خشک شده که حاوی مقادیر زیادی نیترات و همگی محرک دستگاه گوارش هستند، می‌دانند که این عوامل در مناطق

شمالی ایران بیشتر از سایر نواحی گزارش شده اند. برعکس در مناطق جنوب به دلیل مصرف خرما که یکی از مواد سرشار از آنتی‌اکسیدان است و می‌تواند در پیشگیری از سرطان گوارش مؤثر واقع شود مردم این منطقه کمتر به این بیماری مبتلا می‌شوند. البته برای بررسی دقیق این موضوع و سایر عوامل دخیل در بروز سرطان گوارش و نوع الگوی به‌دست آمده در پراکندگی سرطان گوارش در ایران در این تحقیق متخصصان امر باید نظر بدهند.

مراجع

Ali, M., Goovaerts, P., Nazia, N., Haq, MZ., Yunus, M., and Emch, M., (2006), Application of Poisson Kriging to the Mapping of Chlera and Dysentery Incidence in an Endemic Area of Bangladesh. *International Journal of Health Geographics.*

Best, N., Richardson, S., Thomson, A.,(2005), A Comparison of Bayesian Spatial Models for Disease Mapping. *Statistics Methods in Medical Research*, 14:35-59.

Clayton, D., Kaldor, J., (1987), Empirical Bayes Estimation of Age- standardized Relative Risks for Use in Disease Mapping. Biometrics, **43**, 671-681.

Cressie, N., (1993), Statistics For Spatial Data, Revised Edition, Wiley: New York.

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(31).

Goovaerts, P., and Gebreab, S., (2008), "How Does Poisson Kriging Compare to the Popular BYM Model for Mapping Disease Risks?" *International Journal of Health Geographics* 7(6).

Goovaerts, P., (2009), Medical Geography: A Promising Field of Application for Geostatistics. *Mathematical Geosciences*, **41**, 243-64.

Goovaerts, P., (2010), Geostatistical Analysis of Disease Data: Geostatistical Analysis of County-level Lung Cancer Mortality Rates in the Southeastern United States. Wiley.

Kalsall, JE., Wakefield, JC., (1999), Discussion on Bayesian Models for Spatially Correlated Disease and Exposure data. In Bayesian Statistics 6 Edited by: Bernardo, JM., Berger, JO., Dawid, AP. and Smith AFM. Oxford, UK, Oxford University Press:151.

Kavousi, A., Meshkani, M. and Mohammadzadeh, M., (2008),"Spatial Analysis of Relative Risk of Lip Cacer in Iran: a Bayesian approach" Environmetrics; **19**, 1-13.

Lawson A., Bigger, A., Boehning, D., Lesaffre E, Viel J.F., Clark, A., Schlattmann P. (2000), Disease mapping models: an empirical evaluation. *Statistics in Medicine*, 2217-2241.

Lawson, A., Brown, W., Vidal, R. and Carmen, L., (2003), Disease Mapping with WinBugs and MLwiN UK: Wiley & Sons Ltd, pp 1-28.

Monestiez, P., Dubroca, E., Bonnin, J., Durbec, P., and Guinet, C., (2006), Geostatistical Modeling of Spatial Distribution of Balenoptera physalus in the Northwestern Mediterranean Sea From Sparse Count Data and Heterogeneous Observation Efforts. *Ecological Modelling***193**, 615-28.

Rao J. N. K. (2003), Small Area Estimation. New Jersey: Wiley & Sons, pp: 205-210.

Shao, C.Y. , Mueller and J.Cross. (2009), Area-to-point poisson kriging analysis for lung cancer incidence in Perth areas". 18^{th} World IMACS/MODSIM Congeress, Caire, Australia 13-17 July 2009.

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱
مجموعه مقالات، ص ۱۵-۳۳

روش‌های بوت‌استرپ در تحلیل داده‌های فضایی

نصراله ایران پناه
گروه آمار، دانشگاه اصفهان

چکیده : اغلب استنباطهای آمار فضایی مبتنی بر گاوسی بودن میدان تصادفی است، که در عمل ممکن است این شرط برقرار نباشد. همچنین تعیین ساختار همبستگی فضایی داده‌ها نیز با مشکلاتی در مورد برآورد روبرو است. در این موارد روش بوت‌استرپ را می‌توان برای استنباط ناپارامتری داده‌ها به کار گرفت. افرون (۱۹۷۹) روش بوت‌استرپ را برای داده‌های مستقل ارائه کرد، که در آن می‌توان با استفاده از بازنمونه‌گیری داده‌ها، اریبی، واریانس و توزیع برآوردگرها را برآورد نمود. این روش برای داده‌های فضایی به علت وابستگی مشاهدات کاربرد ندارد و برای آنها از روشهای بوت‌استرپ بلوکی استفاده می‌شود. در روش بوت‌استرپ بلوک متحرک مشاهدات به بلوکهایی متداخل تقسیم و بازنمونه‌گیری از این بلوکها انجام می‌شود. چون در این روش حضور مشاهدات مرزی در بلوکهای بازنمونه‌گیری شده نسبت به سایر مشاهدات شانس کمتری دارند، برآوردگرهای اندازه‌های دقت اریب خواهند بود. در این مقاله، برای رفع این مشکل روش بوت‌استرپ بلوک مجزا ارائه می‌شود. چون دقت برآوردگرها شدیداً به اندازهٔ بلوکها بستگی دارد، اندازه

آدرس الکترونیک مسئول مقاله: نصراله ایران پناه، iranpanah@sci.ui.ac.ir
کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲G۰۹ و ۶۲M۳۰

بلوک بهینهٔ مجانبی در روش بوت‌استرپ بلوک مجزا برای برآورد واریانس میانگین نمونه‌ای داده‌های مشبکه‌ای تعیین و روشی تجربی برای برآورد اندازه بلوک بهینه ارائه می‌گردد. همچنین بهینگی برآورد تجربی اندازه بلوک در یک مطالعهٔ شبیه‌سازی مورد ارزیابی عددی قرار می‌گیرد.
از طرف دیگر، روشهای بوت‌استرپ بلوک فضایی در عمل با محدودیتها و نقاط ضعفی همراه هستند. از جمله برآورد اندازه بلوک بسیار مشکل و واریانس برآوردگرها عموماً کم برآورد می‌شوند. برای این منظور روش بوت‌استرپ نیم پارامتری برای برآورد اندازه‌های دقت برآوردگرها در آمار فضایی نیز ارائه می‌گردد. همچنین کارایی روشهای بوت‌استرپ نیم پارامتری، بوت‌استرپ بلوک مجزا و بلوک متحرک در یک مطالعهٔ شبیه‌سازی مورد مقایسه قرار می‌گیرند. در نهایت، روش بوت‌استرپ نیم پارامتری برای تحلیل داده‌های خاکستر ذغال سنگ در زمین شناسی مورد استفاده قرار می‌گیرد.

واژه‌های کلیدی : بوت‌استرپ بلوک مجزا، بوت‌استرپ بلوک متحرک، بوت‌استرپ نیم پارامتری، اندازه بلوک بهینه، کریگیدن.

۱ مقدمه

داده‌های فضایی مشاهداتی هستند که وابستگی آنها ناشی از موقعیتشان در فضای مورد مطالعه است و این وابستگی تابعی از فاصلهٔ مشاهدات از یکدیگر است (کرسی، ۱۹۹۳). دو ویژگی مهم داده‌های فضایی، نمایش هر داده با موقعیت آن در فضای مورد مطالعه و همبستگی فضایی این داده‌هاست. معمولاً داده‌های فضایی که از موقعیتهای مجاور جمع‌آوری می‌شوند، همبستگی بیشتری دارند و با افزایش فاصلهٔ بین موقعیت داده‌ها، همبستگی کاهش می‌یابد. از جمله مباحث مهم در تحلیل داده‌های فضایی بررسی و مطالعهٔ ساختار همبستگی فضایی و برآورد آن با استفاده از مشاهدات و همچنین پیشگویی در موقعیتهای فاقد مشاهده و تعیین خطای پیشگویی است.

برآورد ساختار همبستگی و پارامترهای آن در آمار فضایی با استفاده از روشهای

عددی و معمولاً با مشکلات محاسباتی همراه است. از طرف دیگر، اغلب استنباطهای آمار فضایی مبتنی بر گاوسی بودن میدان تصادفی است، که در عمل ممکن است این فرض برقرار نباشد و در صورت برقراری، بررسی این فرض ساده نمی‌باشد. در این موارد می‌توان از روشهای بوت‌استرپ در تحلیل داده‌های فضایی بدون نیاز به معلوم بودن توزیع مشاهدات یا برآورد ساختار همبستگی استفاده نمود. افرون (۱۹۷۹) روش بوت‌استرپ را برای داده‌های مستقل ارائه کرد، که در آن می‌توان با استفاده از بازنمونه‌گیری داده‌ها مشخصات توزیع برآوردگرها را برآورد نمود. همچنین از بازه‌های اطمینان و آزمون فرضهای بوت‌استرپ می‌توان در استنباط پارامترها استفاده نمود. این روش مستقیماً برای داده‌های وابسته مانند سریهای زمانی و داده‌های فضایی قابل استفاده نمی‌باشد. به همین دلیل روشهای بوت‌استرپ بلوکی و نیم پارامتری برای تحلیل دادهای فضایی ارائه می شود.

لیو (۱۹۸۸) روشهای بوت‌استرپ را برای داده‌های همبسته ارائه نمود. هال و همکاران (۱۹۹۵) روش بوت‌استرپ بلوکی را برای داده‌های وابسته معرفی نمودند. لاهیری و همکاران (۱۹۹۹) روش بوت‌استرپ را برای برآورد تابع توزیع تجمعی فضایی به کار بردند. سوستدت دلونا (۲۰۰۱) روش بوت‌استرپ را برای داده‌های فضایی ناهمگن ارائه نمود. اکستروم و سوستدت دلونا (۲۰۰۴) روشهای بوت‌استرپ را برای برآورد واریانس میانگین نمونه در داده‌های فضایی نامانا در میانگین ارائه نمودند.

در این مقاله، مقدمه‌ای بر آمار فضایی و روش بوت‌استرپ در فصل‌های ۲ و ۳ ارائه می‌گردد. در بخش ۴ روش بوت‌استرپ بلوک مجزا برای برآورد اندازه‌های دقت برآوردگرها معرفی می‌گردد. در بخش ۵ اندازه بلوک بهینه در روش بوت‌استرپ بلوک مجزا تعیین و روشی تجربی برای برآورد آن ارائه می‌گردد. روش بوت‌استرپ نیم پارامتری در آمار فضایی در بخش ۶ معرفی می‌گردد. در بخش ۷ کارایی روش بوت‌استرپ نیم پارامتری با روشهای بوت‌استرپ بلوک مجزا و بلوک متحرک در یک مطالعه شبیه‌سازی مورد مقایسه قرار می‌گیرند. کاربرد روش بوت‌استرپ نیم پارامتری در تحلیل داده‌های خاکستر ذغال سنگ در زمین شناسی در بخش ۸ ارائه می‌گردد. سرانجام در بخش ۹ بحث و نتیجه‌گیری آورده شده است.

۲ آمار فضایی

برای تجزیه و تحلیل داده‌های فضایی، لازم است یک مدل آماری در نظر گرفته شود. معمولاً یک میدان تصادفی به عنوان مدل آماری برای داده‌های فضایی در نظر گرفته می‌شود. میدان تصادفی مجموعه‌ای از متغیرهای تصادفی مانند $\{Z(s); s \in D \subseteq R^d; d \geq 1\}$ است، که در آن D یک مجموعه اندیس گذار است. هر میدان تصادفی را می‌توان به صورت $Z(s) = \mu(s) + \delta(s)$ تجزیه کرد، که درآن $\mu(s)$ تغییرات مقیاس بزرگ یا روند و $\delta(s)$ تغییرات مقیاس کوچک یا فرآیند خطای میدان تصادفی است. اگر توزیع توأم هر تعداد متناهی از متغیرهای تصادفی یک میدان دارای توزیع نرمال باشند، میدان تصادفی گاوسی خواهد بود. اگر میانگین میدان تصادفی ثابت و به موقعیت s بستگی نداشته باشد، یعنی $E[Z(s)] = \mu$ و واریانس عبارت $[Z(s) - Z(s+h)]$ فقط تابعی از فاصله موقعیتها باشد، یعنی $Var[Z(s) - Z(s+h)] = 2\gamma(h)$، آنگاه میدان تصادفی را مانای ذاتی می‌نامند. اگر علاوه بر مانایی در میانگین، کوواریانس بین $Z(s)$ و $Z(s+h)$ فقط تابعی از فاصله موقعیتها باشد، یعنی $Cov[Z(s), Z(s+h)] = \sigma(h)$ آنگاه میدان تصادفی را مانای مرتبه دوم می‌نامند. توابع $2\gamma(h)$ و $\sigma(h)$ که ساختار همبستگی فضایی را مشخص می‌کنند، به ترتیب تغییرنگار و همتغییرنگار نامیده می‌شوند و رابطه $\gamma(h) = \sigma(0) - \sigma(h)$ بین آنها برقرار است. توابع تغییرنگار و همتغییرنگار به ترتیب همیشه منفی شرطی و همیشه مثبت هستند. تغییرنگار و همتغییرنگار علاوه بر مقدار h، به جهت آن نیز بستگی دارند. اگر مقادیر $2\gamma(h)$ و $\sigma(h)$ در جهتهای مختلف یکسان و به جهت بستگی نداشته باشند، به عبارت دیگر فقط تابعی از اندازه فاصله h یعنی $|h| = |s_i - s_j|$ باشند، آنها (یا میدان تصادفی) را همسانگرد می‌نامند.

۱.۲ *برآورد ساختار همبستگی*

در آمار فضایی میزان همبستگی داده‌ها با تغییرنگار و همتغییرنگار اندازه‌گیری می‌شود. افزایش تغییرنگار بیانگر کاهش همبستگی داده‌ها است. با توجه به اینکه

معمولاً بعد از فاصله‌ای همبستگی داده‌های فضایی ناچیز و حتی مستقل از یکدیگر می‌شوند، تغییرنگار از این فاصله به بعد تغییر چندانی ندارد و به صورت تخت درمی‌آید، این فاصله، دامنه تغییرنگار (a) نامیده می‌شود. تغییرنگار بعد از دامنه به مقدار ثابتی گرایش می‌یابد که آستانه ($c_۰ + c$) نام دارد. پارامتر دیگر تغییرنگار که معمولاً به دلیل خطاهای اندازه‌گیری ظاهر می‌شود، اثر قطعه‌ای $c_۰ = \lim_{h \to ۰} \gamma(h)$ است.

در عمل تغییرنگار نامعلوم است و باید بر اساس داده‌های فضایی برآورد شود. یک برآورد تجربی برای تغییرنگار براساس مشاهدات $\{Z(s_۱), \ldots, Z(s_n)\}$ به صورت

$$۲\hat{\gamma}(h) = \frac{۱}{N_h} \sum_{i=۱}^{N_h} [Z(s_i + h) - Z(s_i)]^۲,$$

تعریف می‌شود، که در آن N_h تعداد زوج نقاط متمایزی است که به فاصله h از یکدیگر قرار دارند. $۲\hat{\gamma}(h)$ یک برآوردگر نااریب است، ولی مستقیماً نمی‌توان آن را در روشهای مختلف تجزیه و تحلیل داده‌های فضایی از جمله پیش‌بینی به کار برد. زیرا این بر آوردگر لزوماً خاصیت همیشه منفی بودن شرطی تغییرنگار را ندارد. برای تحلیل داده‌های فضایی از جمله پیش‌بینی، لازم است یک مدل به تغییرنگار تجربی داده‌ها برازش داده شود. برای این منظور ابتدا تغییرنگار تجربی مشاهدات در فواصل مختلف محاسبه و سپس یک مدل پارامتری مناسب به تغییرنگار تجربی برازانده شده و در نهایت پارامترهای آن فقط به صورت عددی و با یکی از روشهای OLS، WLS، ML و REML برآورد می‌گردند. مدلهای مختلفی از جمله مدلهای نمایی، کروی، خطی، گاوسی، توانی و موجی مورد استفاده قرار می‌گیرد. به عنوان مثال مدل تغییرنگار پارامتری نمایی به صورت

$$۲\gamma(h;\theta) = \begin{cases} ۰ & h = ۰, \\ c_۰ + c_۱(۱ - e^{\frac{-h}{a}}) & h \neq ۰ \end{cases} \qquad (۱)$$

است، که در آن $\theta = (c_۰, c, a)$، پارامترهای اثر قطعه‌ای، آستانه جزئی و دامنه هستند.

۲.۲ پیشگویی فضایی کریگیدن

کریگیدن یک روش پیش‌بینی کننده در آمار فضایی به ویژه در زمین آمار می‌باشد و معادل بهترین پیش‌بینی کننده نااریب خطی ($BLUP$) می‌باشد. مسئله پیش‌بینی $Z(s_0)$ بر اساس مشاهدات $Z = (Z(s_1), \ldots, Z(s_n))$ را در نظر بگیرید. فرض کنید $P(Z; s_0)$ پیش‌بینی کننده بر اساس تابع زیان درجه دوم باشد، در این صورت بهترین پیش‌بینی کننده خطی نااریب و واریانس آن به صورت

$$\begin{aligned} P(Z; s_0) &= \sigma^T \Sigma^{-1}(Z - \mu) + \mu(s_0), \\ \sigma^2(s_0) &= \sigma(0) - \sigma^T \Sigma^{-1} \sigma, \end{aligned}$$

به دست می‌آید، که در آن $\mu = (\mu(s_1), \ldots, \mu(s_n))$ ساختار میانگین، Σ یک ماتریس $n \times n$ با عناصر $\sigma(s_i, s_j)$ و $\sigma = (\sigma(s_0, s_1), \ldots, \sigma(s_0, s_n))$ است. با فرض معلوم بودن μ بهترین پیش‌بینی کننده خطی فضایی را کریگیدن ساده، μ ثابت ولی نامعلوم کریگیدن معمولی و چنانچه μ ثابت نباشد، کریگیدن عمومی نامیده می‌شود. اگر میدان تصادفی $Z(\cdot)$ مانای مرتبهٔ دوم باشد، کریگیدن معمولی و واریانس آن به صورت

$$\hat{Z}(s_0) = \lambda^T \mathcal{Z}; \qquad \lambda^T = (\sigma + Im)^T \Sigma^{-1}, \tag{۲}$$

$$\sigma^2(s_0) = \sigma(0) - \lambda^T \sigma + m; \quad m = \frac{1 - I^T \Sigma^{-1} \sigma}{I^T \Sigma^{-1} I}, \tag{۳}$$

ارائه می‌گردد، که در آنها $\sigma = (\sigma(s_0 - s_1), \ldots, \sigma(s_0 - s_n))^T$، Σ یک ماتریس $n \times n$ با مؤلفهٔ (i,j)ام $\sigma(s_i - s_j)$ و $I = (1, \ldots, 1)$ است. کریگیدن عمومی و واریانس آن به طور مشابه در کرسی (۱۹۹۳) ارائه گردیده است.

۳ روش بوت‌استرپ

افرون (۱۹۷۹) روش بوت‌استرپ را برای داده‌های مستقل ارائه نمود. فرض کنید $X_1, \ldots, X_n$ متغیرهای تصادفی مستقل و هم‌توزیع (iid) با تابع توزیع نامعلوم F_θ و $T = t(X_1, \ldots, X_n)$ برآوردگر پارامتر θ باشد. روش بوت‌استرپ برای داده‌های iid

(IIDB) بر اساس ایدهٔ بازنمونه‌گیری از داده‌ها برای برآورد اریبی، واریانس و توزیع T و بازهٔ اطمینان و آزمون فرض برای θ است. روش بوت‌استرپ را می‌توان به صورت الگوریتم زیر ارائه نمود (افرون و تیبشیرانی، ۱۹۹۳):

۱) نمونهٔ بوت‌استرپ $X_1^*, \ldots, X_n^*$ را با نمونه‌گیری تصادفی سادهٔ با جایگذاری (SRSW) از $X_1, \ldots, X_n$ به دست آورید.

۲) برآوردگر بوت‌استرپ $T^* = t(X_1^*, \ldots, X_n^*)$ را محاسبه کنید.

۳) برآورد نظری بوت‌استرپ اریبی، واریانس و توزیع T به ترتیب به صورت

$$\begin{aligned} Bias_*(T^*) &= E_*(T^*) - T \\ Var_*(T^*) &= E_*[T^* - E_*(T^*)]^2 \\ G_*(t) &= P_*(T^* \le t) \end{aligned}$$

است، که در آنها E_*، Var_* و P_* به ترتیب امید ریاضی، واریانس و احتمال شرطی بوت‌استرپ به شرط $X_1, \ldots, X_n$ است.

۴) اگر $\mathrm{Bias}_*(T^*)$، $\mathrm{Var}_*(T^*)$ و $\mathrm{G}_*(t)$ جوابهای دقیق یا شکل بسته‌ای نداشته باشند، با استفاده از شبیه‌سازی مونت کارلو و تکرار B بار مراحل ۱ و ۲ و محاسبهٔ $T_1^*, \ldots, T_B^*$، برآورد تجربی آنها را به ترتیب به صورت

$$\widehat{Bias}_*(T^*) = \overline{T^*} - T \qquad (۴)$$

$$\widehat{Var}_*(T^*) = \frac{1}{B-1} \sum_{b=1}^{B} (T_b^* - \overline{T^*})^2 \qquad (۵)$$

$$\widehat{G}_*(t) = \frac{1}{B} \sum_{b=1}^{B} I(T_b^* \le t) \qquad (۶)$$

محاسبه کنید، که در آنها $\overline{T^*} = B^{-1} \sum_{b=1}^{B} T_b^*$ و $I(T_b^* \le t)$ تابع نشانگر با مقدار یک است اگر $T_b^* \le t$ و در غیر این صورت مقدار صفر را اختیار می‌کند.

در اغلب تحلیلهای نظری بوت‌استرپ برای F_θ و T شرایطی فرض می‌شود، به طوری که با افزایش n، توزیع و گشتاورهای T^* به توزیع و گشتاورهای مجانبی T میل کند. برای آشنایی بیشتر با روش بوت‌استرپ و کاربردهای آن به ایران‌پناه و پاشا (۱۳۷۶)، کاربرد روش بوت‌استرپ در استنباط بیزی به ایران‌پناه (۱۳۷۷) مراجعه شود.

۴ روش‌های بوت‌استرپ بلوکی

روش IIDB برای داده‌های وابسته مانند سریهای زمانی و داده‌های فضایی کاربرد ندارد (سینگ ۱۹۸۱). برای این گونه داده‌ها از روشهای بوت‌استرپ بلوکی استفاده می‌شود. هال (۱۹۸۵) دو روش بر اساس بلوکی کردن مشاهدات و موقعیتها برای حالت خاص داده‌های موزاییک ارائه کرد. بولمان و کونش (۱۹۹۹)، زو و لاهیری (۲۰۰۱) و لاهیری (۲۰۰۳) نیز روش بوت‌استرپ بلوک متحرک (MBB) را برای داده‌های فضایی ارائه کردند. در این روش نمونه‌ای از مشاهدات در بلوکهای متحرک بازنمونه‌گیری می‌شود، به طوری که هر مشاهده حداقل در یکی از بلوکها قرار گیرد، اما شانس کمتر مشاهدات مرزی در مقایسه با مشاهدات مرکزی برای حضور در بلوکها موجب اریبی برآوردگرها می‌شود. برای رفع این مشکل، در این بخش روش بوت‌استرپ بلوک مجزا (SBB)، برای داده‌های مشبکه‌ای معرفی و الگوریتمی برای برآورد اندازه‌های دقت برآوردگرها ارائه می‌شود (ایران‌پناه و محمدزاده، ۱۳۸۴ و ۱۳۸۶).

۱.۴ *بوت‌استرپ بلوک مجزا*

فرض کنید مشاهدات یک میدان تصادفی مانای به طور ضعیف وابستهٔ $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbf{Z}^d\}$ در موقعیتهای $\mathcal{S}_n \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_{N_n}\}$ از مشبکهٔ $\mathbf{Z}^d$ که در داخل ناحیهٔ نمونه‌گیری $D_n \subset \mathbf{R}^d$ قرار دارند، به صورت مجموعه داده‌های $\mathcal{Z}_n = \{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{S}_n = D_n \cap \mathbf{Z}^d\}$ باشند. برای بررسی خواص مجانبی برآوردگرهای بوت‌استرپ، فرض می‌کنیم ناحیهٔ نمونه‌گیری D_n وقتی $n \to \infty$ غیر

کراندار باشد. برای این منظور، ابتدا فرض کنید D_0 یک زیر مجموعهٔ بورل $[-1/2, 1/2]^d$ شامل یک همسایگی باز مبدأ است، به طوری که برای هر دنبالهٔ مثبتی از اعداد حقیقی $a_n \to \infty$، تعداد مکعبهای مشبکهٔ مقیاس شدهٔ $a_n\mathbf{Z}^d$ که از تقاطع بستارهای $\overline{D_0}$ و $\overline{D_0^c}$ می‌باشد، وقتی $n \to \infty$ از مرتبه $O((a_n^{-1})^{d-1})$ است. سپس گیریم $\{\lambda_n\}_{n\geq 1} \subset [1, \infty)$ یک دنباله از اعداد حقیقی باشد، به طوری که وقتی $n \to \infty$ آنگاه $\lambda_n \to \infty$. حال ناحیهٔ نمونه‌گیری را به صورت

$$D_n = \lambda_n D_0 \tag{۷}$$

در نظر می‌گیریم که با متورم کردن مجموعهٔ نمونه اولیهٔ D_0 توسط عامل مقیاس‌بندی λ_n تعیین می‌شود. در این صورت اندازهٔ ناحیهٔ نمونه‌گیری برابر $|D_n| = \lambda_n^d |D_0|$ و با اندازهٔ نمونه به صورت $N_n = |D_n \cap \mathbf{Z}^d|$ در ارتباط است، که در آن نماد $|A|$ برای مجموعهٔ شمارای $A \subset \mathbf{Z}^d$، تعداد عناصر یا کاردینالیتی A و برای مجموعهٔ ناشمارای $A \subset \mathbf{R}^d$، اندازهٔ لبگ A است. اگر $\hat{\theta}_n = t_n(\mathcal{Z}_n)$ برآوردگر پارامتر θ بر اساس مشاهدات $\mathcal{Z}_n$ باشد، هدف برآورد واریانس آمارهٔ نرمال شدهٔ $\sqrt{N_n}\hat{\theta}_n$ یعنی $\sigma_n^2 = N_n \mathrm{Var}(\hat{\theta}_n)$ به روش SBB است.

برای اجرای روش SBB، ناحیهٔ نمونه‌گیری D_n باید به بلوکهای مکعبی افراز شود. برای این منظور، فرض کنید $\{\beta_n\}_{n\geq 1}$ یک دنباله از اعداد صحیح مثبت باشد، به طوری که، وقتی $n \to \infty$ آنگاه $\beta_n^{-1} + \beta_n \lambda_n^{-1} = o(1)$، یعنی β_n که آن را اندازهٔ بلوک می‌نامیم، با نرخ آهسته‌تری از عامل مقیاس‌بندی λ_n در (۷) به بینهایت میل کند. گیریم $\mathcal{K}_n = \{\mathbf{k} \in \mathbf{Z}^d : \beta_n(\mathbf{k} + \mathcal{U}) \subset D_n\}$ مجموعه اندیس بلوکهای d بعدی به صورت مجزا، مساوی و کامل به شکل $\beta_n(\mathbf{k} + \mathcal{U})$ باشد، که در ناحیهٔ نمونه‌گیری D_n قرار دارند و در آن $\mathcal{U} = (0, 1]^d$ مکعبی واحد در $\mathbf{R}^d$ است. فرض کنید $\mathcal{Z}_n(D_n) = \{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_{N_n})\}$ نمونهٔ کامل و $\mathcal{Z}_n(D_n(\mathbf{k}))$ زیر نمونهٔ قرار گرفته در بلوک $\mathbf{k}$ام یعنی

$$D_n(k) = \beta_n(k + \mathcal{U}) \cap D_n, \quad k \in \mathcal{K}_n \tag{۸}$$

باشد. با توجه به ساختار جدید ناحیهٔ نمونه‌گیری بر اساس بلوکها، اندازهٔ نمونهٔ جدید برابر $N_{1n} = |\mathcal{K}_n|\beta_n^d \leq N_n$ است. برای سادگی فرض می‌کنیم

که $N_n = N_{1n}$، یعنی ناحیهٔ نمونه‌گیری D_n با تعداد $|\mathcal{K}_n|$ بلوک به اندازهٔ β_n^d کامل می‌شود. گیریم $\mathcal{I}_n = \{\mathbf{i} \in \mathbf{Z}^d : \beta_n(\mathbf{i}+\mathcal{U}) \subset D_n\}$ مجموعه اندیس بلوکهای مجزا به اندازهٔ β_n^d در D_n با نقاط شروع $\mathbf{i} \in \mathbf{Z}^d$ باشد، در نتیجه $\mathcal{B}_{n,SBB} = \{\beta_n(\mathbf{i}+\mathcal{U}) : \mathbf{i} \in \mathcal{I}_n\}$ یک مجموعه از بلوکهای مکعبی مجزا واقع در D_n هستند. برای به دست آوردن یک نمونهٔ SBB، ابتدا برای هر $\mathbf{k} \in \mathcal{K}_n$ یک بلوک به صورت تصادفی از مجموعهٔ بلوکهای مجزای $\mathcal{B}_{n,SBB}$ و مستقل از بلوکهای دیگر انتخاب می‌شود. سپس با استفاده از مشاهدات در هر $\mathbf{k} \in \mathcal{K}_n$ بلوک بازنمونه‌گیری شده و پیوستن آنها نمونهٔ بوت‌استرپ تولید می‌گردد. به بیان دقیق‌تر، فرض کنید $\{I_{\mathbf{k}} : \mathbf{k} \in \mathcal{K}_n\}$ یک مجموعه از متغیرهای تصادفی iid با توزیع مشترک

$$P(I_1 = i) = \frac{1}{|\mathcal{I}_n|}, \quad i \in \mathcal{I}_n$$

باشد. برای هر $\mathbf{k} \in \mathcal{K}_n$، زیر نمونهٔ SBB به صورت $\mathcal{Z}_n^*(D_n(\mathbf{k})) = \mathcal{Z}_n(\beta_n(I_{\mathbf{k}}+\mathcal{U}))$ به دست آورده می‌شود. حال نمونهٔ بوت‌استرپ بلوک مجزای $\mathcal{Z}_n^*(D_n)$ از به هم پیوستن مشاهدات بلوکهای بازنمونه‌گیری شده به صورت $\{\mathcal{Z}_n^*(D_n(\mathbf{k})) : \mathbf{k} \in \mathcal{K}_n\}$ تعیین می‌گردد. سرانجام نسخهٔ بوت‌استرپ بلوک مجزای برآوردگر $\hat{\theta}_n$ به صورت $\hat{\theta}_n^* = t_n(\mathcal{Z}_n^*(D_n))$ و برآوردگر نظری واریانس σ_n^2 به صورت $\hat{\sigma}_n^2(\beta_n) = N_n \mathrm{Var}_*(\hat{\theta}_n^*)$ تعریف می‌شود، که در آن Var_* واریانس شرطی بوت‌استرپ به شرط مشاهدات $\mathcal{Z}_n$ است. اگر $\hat{\sigma}_n^2(\beta_n)$ شکل بسته‌ای نداشته باشد، با استفاده از روش شبیه‌سازی مونت کارلو و تکرار B بار مراحل قبل و محاسبهٔ $\hat{\theta}_{n,1}^*, \ldots, \hat{\theta}_{n,B}^*$، برآورد تجربی بوت‌استرپ بلوک مجزای $\hat{\sigma}_n^2(\beta_n)$ به صورت $\hat{\sigma}_n^2(\beta_n) \simeq N_n \widehat{\mathrm{Var}}_*(\hat{\theta}_n^*)$ تقریب زده می‌شود، که در آن

$$\widehat{Var}_*(\hat{\theta}_n^*) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_{n,b}^* - \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_{n,b}^*)^2.$$

ایران‌پناه و محمدزاده (۱۳۸۴) نشان دادند برآورد بوت‌استرپ $\hat{\sigma}_n^2(\beta_n) = N_n \mathrm{Var}_*(\bar{Z}_n^*)$ یک برآوردگر سازگار σ_∞^2 است، که در آن

$$\sigma_\infty^2 = \lim_{n\to\infty} N_n Var(\bar{Z}_n) = \sum_{k \in Z^d} E[Z(0)-\mu][Z(k)-\mu] \qquad (۹)$$

می‌باشد. ایران‌پناه و محمدزاده (۱۳۸۶) همچنین نشان دادند این خاصیت برای پیشگوی فضایی کریگیدن نیز برقرار است.

۲.۴ *بوت‌استرپ بلوک متحرک*

بولمان و کونش (۱۹۹۹)، زو و لاهیری (۲۰۰۱) و لاهیری (۲۰۰۳) روش MBB را برای برآورد مشخصات توزیع نمونه‌ای برآوردگرها در آمار فضایی ارائه کردند. فرض کنید $D_n(k)$ افراز ناحیهٔ نمونه‌گیری D_n داده شده در (۸) به صورت مکعبهایی به اندازهٔ β_n^d باشد. در این روش ابتدا با استفاده از مجموعه بلوکهای متداخل $\mathcal{B}_{n,MBB} = \{(j + \beta_n \mathcal{U}) : j \in \mathcal{J}_n\}$ که در آن $\mathcal{J}_n = \{j \in \mathbf{Z}^d : (j + \beta_n \mathcal{U}) \subset D_n\}$ مجموعه اندیس بلوکهای متداخل به اندازهٔ B_n در D_n با نقاط شروع $j \in \mathbf{Z}^d$ است، $K = |\mathcal{K}_n|$ متغیر تصادفی $\{J_{\mathbf{k}} : \mathbf{k} \in \mathcal{K}_n\}$ iid با توزیع مشترک

$$P(J_1 = j) = \frac{1}{|\mathcal{J}_n|}, \quad j \in \mathcal{J}_n$$

تولید می‌شود. سپس زیر نمونهٔ MBB به صورت $\mathcal{Z}_n^*(D_n(\mathbf{k})) = \mathcal{Z}_n(J_{\mathbf{k}} + \beta_n \mathcal{U})$ به دست می‌آید. حال نمونهٔ بوت‌استرپ بلوک متحرک $\mathcal{Z}_n^*(D_n)$ از به هم پیوستن مشاهدات بلوکهای بازنمونه‌گیری شده به صورت $\{\mathcal{Z}_n^*(D_n(\mathbf{k})) : \mathbf{k} \in \mathcal{K}_n\}$ تعیین می‌گردد. ادامهٔ روش مشابه روش SBB است.

ایران‌پناه و محمدزاده (۱۳۸۴) نداشتن خطا در برآورد اریبی برآوردگر میانگین نمونه‌ای به روش SBB و سازگاری برآوردگر واریانس آن را نشان دادند. آنها همچنین در یک مطالعهٔ شبیه‌سازی نشان دادند معیار MSE برای برآورد اریبی به روش SBB نسبت به روش MBB کاهش قابل ملاحظه‌ای دارد، در حالی که برآورد واریانس به هر دو روش SBB و MBB دقت تقریباً یکسانی دارند. همچنین تأثیر اندازه نمونه و خاصیت سازگاری برآوردگرهای اریبی و واریانس به هر دو روش مورد مطالعه شبیه‌سازی قرار گرفته است. علت اصلی افزایش دقت برآورد اریبی شانس یکسان ظاهر شدن مشاهدات در بلوکهای مختلف و رفع مشکل نقاط مرزی می‌باشند. ایران‌پناه و محمدزاده (۱۳۸۶) دو روش SBB و MBB را برای برآورد

اریبی و واریانس پیشگوی جایگذاری در یک مطالعهٔ شبیه‌سازی مورد مقایسه قرار دادند. آنها نشان دادند در موقعیتهای متفاوت روش SBB نسبت به روش MBB مقدار اریبی پیشگوی جایگذاری را بسیار نزدیک به مقدار واقعی برآورد می‌نماید. همچنین روش SBB واریانس برآورد شده پیشگوی جایگذاری را در موقعیتهای نزدیک مرزها با دقت بیشتری برآورد می‌کند، در حالی که روش MBB واریانس را در موقعیتهای مرکزی بهتر برآورد می‌نماید.

۵ اندازه بلوک بوت‌استرپ

دقت برآورد بوت‌استرپ بلوکی اندازه‌های دقت برآوردگرها به انتخاب اندازهٔ بلوک حساسیت دارد. تعیین اندازه بلوک بهینه برای روشهای بوت‌استرپ بلوکی سریهای زمانی توسط کونش (۱۹۸۹)، هال و همکاران (۱۹۹۵) و لاهیری (۱۹۹۹) مطالعه شده است. همچنین نوردمن و لاهیری (۲۰۰۴) اندازه بلوک بهینه را برای روشهای زیرنمونه‌گیری فضایی ارائه کردند. در این بخش، اندازه بلوک بهینه برای روش SBB به منظور برآورد واریانس میانگین نمونه‌ای داده‌های مشبکه‌ای به صورت مجانبی تعیین می‌گردد (ایران‌پناه و همکاران، ۲۰۰۹). برای این منظور، ابتدا اریبی و واریانس مجانبی برآوردگر واریانس میانگین نمونه‌ای به روش SBB تعیین می‌شود. سپس با کمینه کردن میانگین توان دوم خطای مجانبی برآوردگر مورد نظر، اندازه بلوک بهینه به صورت مجانبی محاسبه می‌گردد. اندازه بلوک بهینهٔ مجانبی وابسته به پارامترهای جامعه است، که برای برآورد آن از روش جایگذاری ناپارامتری استفاده می‌کنیم و نشان می‌دهیم این برآوردگر سازگار است. سرانجام در یک مطالعهٔ شبیه‌سازی نتایج نظری و مجانبی را مورد ارزیابی عددی قرار می‌دهیم.

۱.۵ *اندازه بلوک بهینه*

دقت برآوردگر $\hat{\sigma}_n^2 = \hat{\sigma}_n^2(\beta_n)$ به روش SBB، شدیداً به اندازه بلوک β_n حساس است. در این بخش مقدار بهینهٔ اندازه بلوک β_n مورد بررسی قرار می‌گیرد. فرض کنید تحت ساختار نمونه‌گیری روش SBB، میانگین نمونه‌ای

$\bar{Z}_n = N_n^{-1}\sum_{i=1}^{N_n} Z(\mathbf{s}_i)$ به عنوان برآوردگر میانگین میدان تصادفی، یعنی $\mu = \mathrm{E}[Z(0)]$، بر اساس مشاهدات $\mathcal{Z}_n$ باشد. اگر $\bar{Z}_n^* = N_n^{-1}\sum_{i=1}^{N_n} Z^*(\mathbf{s}_i)$ میانگین نمونهٔ بوت‌استرپ $\mathcal{Z}_n^*$ به روش SBB باشد، در این صورت برآورد بوت‌استرپ $\sigma_n^2 = N_n\mathrm{Var}(\bar{Z}_n)$ به صورت $\hat{\sigma}_n^2(\beta_n) = N_n\mathrm{Var}_*(\bar{Z}_n^*)$ در نظر گرفته می‌شود. ایران‌پناه و همکاران (۲۰۰۹) اریبی و واریانس مجانبی $\hat{\sigma}_n^2(\beta_n)$ را به صورت

$$Bias[\hat{\sigma}_n^2(\beta_n)] = -\frac{B_0}{\beta_n}(1+o(1)); \qquad B_0 = \sum_{k\in Z^d} ||k||_1\sigma(k) \quad (10)$$

$$Var[\hat{\sigma}_n^2(\beta_n)] = \frac{2\sigma_\infty^4 B_n}{N_n}(1+o(1)), \quad (11)$$

به دست آوردند، که در آنها B_0 و σ_∞^2 به ترتیب مؤلفه‌های اریبی و واریانس هستند. این مسئله نشان می دهد $\hat{\sigma}_n^2(\beta_n)$ یک برآوردگر MSE–سازگار و در نتیجه سازگار σ_n^2 است. اریبی و واریانس برآوردگر $\hat{\sigma}_n^2(\beta_n)$ بستگی به اندازه بلوک β_n دارد. افزایش اندازه بلوک β_n باعث کاهش اریبی و افزایش واریانس برآوردگر $\hat{\sigma}_n^2(\beta_n)$ می‌گردد. بهترین مقدار اندازه بلوک β_n با استفاده از کمینه کردن ترکیبی از دو مقدار اریبی و واریانس برآوردگر $\hat{\sigma}_n^2(\beta_n)$ به دست می‌آید.

قضیهٔ ۱. اندازه بلوک بهینهٔ مجانبی برای $\hat{\sigma}_n^2(\beta_n)$ برابر است با

$$\beta_n^{opt} = \left(\frac{N_n B_0^2}{d\sigma_\infty^4}\right)^{1/(d+2)}(1+o(1)).$$

برهان: مقدار β_n^{opt} با استفاده از کمینه کردن عبارت

$$\begin{aligned} MSE[\hat{\sigma}_n^2(\beta_n)] &= [Bias(\hat{\sigma}_n^2(\beta_n))]^2 + Var[\hat{\sigma}_n^2(\beta_n)] \\ &= \left(\frac{B_0^2}{\beta_n^2} + \frac{2\sigma_\infty^4\beta_n^d}{N_n}\right)(1+o(1)) \end{aligned}$$

نسبت به β_n به دست می‌آید.

اندازه بلوک بهینهٔ $\beta_n^{\rm opt}$ بستگی به دو پارامتر مؤلفهٔ اریبی B_0 و مؤلفهٔ واریانس σ_∞^2 دارد که در عمل نامعلوم هستند. برای برآورد آنها و در نتیجه برآورد $\hat{\beta}_n$ از روش جایگذاری ناپارامتری پیشنهادی لاهیری و همکاران (۲۰۰۷) استفاده می‌کنیم. این روش که برای سریهای زمانی ارائه شده است را برای داده‌های مشبکهٔ فضایی تعمیم می‌دهیم. فرض کنید اندازه‌های بلوک اولیهٔ $\beta_{n,1}$ و $\beta_{n,2}$ دنباله‌هایی از اعداد صحیح مثبت باشند. بر اساس اندازه بلوک اولیهٔ $\beta_{n,1}$، مؤلفهٔ واریانس σ_∞^2 را به صورت $\hat{\sigma}_\infty^2 = \hat{\sigma}_n^2(\beta_{n,1})$ برآورد می‌کنیم. همچنین برای مؤلفهٔ اریبی B_0 بر اساس دو برآورد واریانس بوت‌استرپ بلوک مجزا با استفاده از اندازه بلوک اولیهٔ $\beta_{n,2}$ برآورد $\hat{B}_0 = 2\beta_{n,2}[\hat{\sigma}_n^2(2\beta_{n,2}) - \hat{\sigma}_n^2(\beta_{n,2})]$ را ارائه می‌کنیم. در نتیجه اندازه بلوک بهینهٔ $\beta_n^{\rm opt}$ را با استفاده از روش جایگذاری ناپارامتری به صورت $\hat{\beta}_n = (N_n \hat{B}_0^2 / d\hat{\sigma}_\infty^4)^{1/(d+2)}$ برآورد می‌کنیم.

قضیهٔ ۲. وقتی $n \to \infty$ آنگاه

$$\frac{\hat{\beta}_n}{\beta_n^{opt}} \to_p 1.$$

برهان: با استفاده از اریبی و واریانس مجانبی $\hat{\sigma}_n^2(\beta_n)$ در روابط ۱۰ و ۱۱، $\hat{B}_0$ و $\hat{\sigma}_\infty^2$ به ترتیب برآوردگرهای MSE-سازگار B_0 و σ_∞^2 هستند. در نتیجه $\hat{\beta}_n$ یک برآوردگر سازگار $\beta_n^{\rm opt}$ است.

برآورد جایگذاری ناپارامتری $\hat{\beta}_n$ بستگی به دو اندازه بلوک اولیهٔ $\beta_{n,1}$ و $\beta_{n,2}$ دارد. با استفاده از قضیهٔ ۱ نرخ بهینهٔ اندازه بلوک اولیهٔ $\beta_{n,1}$ برای برآورد مؤلفهٔ واریانس σ_∞^2 برابر $N_n^{1/(d+2)}$ و برای $\beta_{n,2}$ به عنوان مؤلفهٔ اریبی B_0 برابر $N_n^{1/(d+4)}$ است. بنابراین انتخابهای قابل قبول برای اندازه‌های بلوک اولیه به صورت $\beta_{n,i} = C_i N_n^{1/(d+2i)}; i = 1, 2$ هستند. با بررسیهای عددی انجام گرفته مقادیر مناسب برای C_1 و C_2 در فاصلهٔ $[0/5, 2]$ با فواصل ۰/۲۵ قرار دارند، و در این فاصله مقادیر $C_1 = \{0/5, 0/75\}$ و $C_2 = 0/5$ توصیه می‌گردد.

۳.۵ مطالعهٔ شبیه‌سازی

در این بخش، ابتدا اندازه بلوک بهینهٔ $\beta_n^{\rm opt}$ را تعیین نموده، سپس برآورد $\hat{\beta}_n$ به روش جایگذاری ناپارامتری را در یک مطالعهٔ شبیه‌سازی مونت کارلوی داده‌های فضایی مورد ارزیابی عددی قرار می‌دهیم. فرض کنید $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbf{N}^۲\}$ یک میدان تصادفی گاوسی مانای مرتبهٔ دو با میانگین صفر و تغییرنگار نمایی (۱) باشد. با در نظر گرفتن دو مدل با پارامترهای تغییرنگار $\theta_۱ = (۰/۵, ۰/۵, ۰/۵)$ و $\theta_۲ = (۱, ۱, ۱)$، نمونه‌ها را در یک مشبکهٔ منظم مربعی در سه ناحیه، به ازای $D_۰ = (۰, ۱]^۲$ و $\lambda_n = ۱۲, ۲۴, ۴۸$ به روش تجزیهٔ چولسکی (کرسی،۱۹۹۳) تولید می‌کنیم. اگر $\bar{Z}_n$ میانگین نمونه‌ای در مشبکه‌های مورد نظر باشد، برآورد بوت‌استرپ بلوک مجزای $\sigma_n^۲ = N_n \mathrm{Var}(\bar{Z}_n)$ به صورت $\sigma_n^۲ = N_n \mathrm{Var}_*(\bar{Z}_n^*)$ می‌باشد که دارای شکل بسته به صورت

$$\hat{\sigma}_n^۲(\beta_n) = \beta_n^۲ \mid \mathcal{K}_n \mid^{-۱} \sum_{k \in \mathcal{K}_n} (\bar{Z}_{k,n} - \bar{Z}_n)^۲,$$

است که در آن $\bar{Z}_{\mathbf{k},n}$ میانگین نمونه‌ای $\beta_n^۲$ مشاهده در بلوکهای مجزای

$$\begin{aligned} D_n(k) &= (\beta_n k_۱ - \beta_n, \beta_n k_۱] \times (\beta_n k_۲ - \beta_n, \beta_n k_۲]; \quad k = (k_۱, k_۲)^T \in \mathcal{K}_n \\ \mathcal{K}_n &= \{k \in N^۲, ۰ < k_۱, k_۲ \leq \lambda_n \beta_n^{-۱}\} \end{aligned}$$

است. اندازه‌های بلوک مجزای β_n را برای سه مقدار λ_n به ترتیب $(۲, ۳, ۴, ۶)$، $(۲, ۳, ۴, ۶, ۸, ۱۲)$ و $(۲, ۳, ۴, ۶, ۸, ۱۲, ۱۶, ۲۴)$ در نظر می‌گیریم. مقدار $\sigma_n^۲$ به ازای $\theta = \theta_۱$ برای سه مقدار λ_n به ترتیب ۱/۴۳۰، ۱/۴۶۳ و ۱/۴۸۰ و به ازای $\theta = \theta_۲$ به ترتیب ۶/۳۱۱، ۶/۸۹۰ و ۷/۱۹۳ است. مقادیر حدی $\sigma_\infty^۲$ نیز برای دو مدل به ترتیب ۱/۴۸۳ و ۷/۲۸۸ هستند. حال برای هر دو مقدار $\theta_۱$ و $\theta_۲$ و سه مقدار λ_n و همچنین اندازه‌های بلوک مجزای مورد نظر β_n مقدار $\hat{\sigma}_n^۲(\beta_n)$ را محاسبه می‌کنیم.

جدول ۱ تقریبهای اریبی نسبی $\mathrm{E}[\hat{\sigma}_n^۲(\beta_n)/\sigma_n^۲ - ۱]$، واریانس نسبی $\mathrm{E}[(\hat{\sigma}_n^۲(\beta_n) - \mathrm{E}\hat{\sigma}_n^۲(\beta_n))/\sigma_n^۲]^۲$ و میانگین توان دوم خطای نسبی

جدول ۱: برآوردهای اریبی، واریانس و میانگین توان دوم خطای نسبی برآوردگر واریانس بوت‌استرپ بلوک مجزای $\hat{\sigma}^2_n(\beta_n)$.

		$\theta_1 = (0/5, 0/5, 0/5)$			$\theta_2 = (1, 1, 1)$		
λ_n	β_n	Bias	Var	MSE	Bias	Var	MSE
	۲	-۰/۲۱۳	۰/۰۳۵	۰/۰۸۱*	-۰/۵۵۵	۰/۰۱۳	۰/۳۲۲
۱۲	۳	-۰/۱۸۴	۰/۰۸۷	۰/۱۲۱	-۰/۴۶۱	۰/۰۴۲	۰/۲۵۴*
	۴	-۰/۱۹۳	۰/۱۶۰	۰/۱۹۶	-۰/۴۰۶	۰/۰۹۰	۰/۲۵۵
	۶	-۰/۲۹۰	۰/۳۳۵	۰/۴۱۹	-۰/۴۲۲	۰/۲۲۴	۰/۴۰۲
	۲	-۰/۲۱۱	۰/۰۰۹	۰/۰۵۳	-۰/۵۷۵	۰/۰۰۳	۰/۳۳۴
	۳	-۰/۱۵۸	۰/۰۲۳	۰/۰۴۷*	-۰/۴۶۳	۰/۰۱۰	۰/۲۲۴
۲۴	۴	-۰/۱۳۲	۰/۰۴۴	۰/۰۶۱	-۰/۳۸۳	۰/۰۲۳	۰/۱۶۹
	۶	-۰/۱۲۷	۰/۰۹۸	۰/۱۱۴	-۰/۲۹۷	۰/۰۶۹	۰/۱۵۷*
	۸	-۰/۱۶۳	۰/۱۷۶	۰/۲۰۳	-۰/۲۷۴	۰/۱۳۲	۰/۲۰۷
	۱۲	-۰/۲۸۱	۰/۳۵۶	۰/۴۳۵	-۰/۳۴۴	۰/۲۸۱	۰/۴۰۰
	۲	-۰/۲۱۴	۰/۰۰۳	۰/۰۴۸	-۰/۵۸۸	۰/۰۰۱	۰/۳۴۷
	۳	-۰/۱۵۵	۰/۰۰۷	۰/۰۳۱	-۰/۴۷۵	۰/۰۰۳	۰/۲۲۸
	۴	-۰/۱۲۴	۰/۰۱۳	۰/۰۲۹*	-۰/۳۸۹	۰/۰۰۶	۰/۱۵۸
۴۸	۶	-۰/۰۹۱	۰/۰۳۳	۰/۰۴۱	-۰/۲۸۳	۰/۰۱۹	۰/۰۹۹
	۸	-۰/۰۸۸	۰/۰۵۸	۰/۰۶۶	-۰/۲۲۷	۰/۰۳۹	۰/۰۹۱*
	۱۲	-۰/۰۹۰	۰/۱۳۳	۰/۱۴۱	-۰/۱۸۶	۰/۱۰۰	۰/۱۳۵
	۱۶	-۰/۱۳۹	۰/۲۲۳	۰/۲۴۲	-۰/۱۹۶	۰/۱۸۵	۰/۲۲۴
	۲۴	-۰/۲۶۴	۰/۴۰۹	۰/۴۷۸	-۰/۲۹۴	۰/۳۶۳	۰/۴۴۹

$\mathrm{E}[\hat{\sigma}^2_n(\beta_n)/\sigma^2_n - 1]^2$ به روش مونت کارلو بر اساس ۱۰۰۰ تکرار نشان می‌دهد. همان طور که ملاحظه می‌شود برای هر دو مدل و هر سه مقدار λ_n، با افزایش اندازه بلوک β_n مقدار اریبی تقریباً کاهش و مقدار واریانس افزایش می‌یابد که با نتایج حاصل از اریبی و واریانس مجانبی $\hat{\sigma}^2_n(\beta_n)$ در روابط ۱۰ و ۱۱ تقریباً مطابقت دارد. مقادیر اریبی، واریانس و MSE غیر نسبی در مدل ۲ که دارای ساختار همبستگی قوی‌تر است در مقایسهٔ متناظر با مدل ۱ که دارای همبستگی ضعیف‌تر است، بزرگ‌تر هستند، که با نتیجه حاصل از قضیه مجانبی ۱ مطابقت دارند. با مقایسهٔ مقادیر MSE و مشاهدهٔ کمترین آنها مقادیر اندازه بلوک بهینهٔ β^{opt}_n به دست می‌آید که در مدل ۱ برای سه مقدار λ_n به ترتیب ۲، ۳ و ۴ و در مدل ۲ به ترتیب ۳، ۶ و ۸ هستند. مقایسهٔ مقادیر مختلف β^{opt}_n نشان می‌دهد با افزایش اندازه نمونهٔ $N_n = \lambda^2_n$، مقدار β^{opt}_n افزایش می‌یابد. همچنین در مدلهایی با ساختار همبستگی قوی‌تر، β^{opt}_n بزرگ‌تر است.

اکنون برآورد جایگذاری ناپارامتری اندازه بلوک $\hat{\beta}_n = (N_n \hat{B}^2_0 / d\hat{\sigma}^4_\infty)^{1/(d+2)}$ را بر اساس ساختار شبیه‌سازی مونت کارلوی داده‌های فضایی قبل مورد ارزیابی

جدول ۲: فراوانی مقادیر مختلف $\hat{\beta}_n$ بر اساس ۱۰۰۰ بار تکرار شبیه‌سازی مونت کارلو.

			$\hat{\beta}_n$									
مدل	λ_n	C_1	۱	۲	۳	۴	۵	۶	۷	۸	۹+	β_n^{opt}
۱	۱۲	۰/۵	۹۱	۲۹۲*	۲۴۱	۲۱۱	۱۲۶	۲۶	۴	۰	۰	۲
		۰/۷۵	۱۲۶	۲۷۶*	۲۱۳	۱۸۸	۱۱۸	۵۰	۱۴	۳	۲	
	۲۴	۰/۵	۱۱۸	۱۷۰	۲۷۵*	۲۳۵	۱۴۳	۳۶	۹	۰	۰	۳
		۰/۷۵	۱۰۰	۲۰۵	۲۸۸*	۲۵۶	۱۲۱	۱۴	۵	۱	۰	
	۴۸	۰/۵	۵۳	۱۳۱	۲۵۰	۲۵۴*	۲۱۱	۸۴	۷	۰	۰	۴
		۰/۷۵	۵۹	۱۴۲	۲۲۸	۲۵۳*	۲۱۳	۸۱	۱۶	۰	۰	
۲	۱۲	۰/۵	۷۴	۱۷۹	۲۱۶*	۲۰۹	۱۸۵	۹۶	۳۰	۱	۰	۳
		۰/۷۵	۱۰۱	۱۸۶	۲۸۵*	۲۱۲	۱۴۶	۵۱	۱۳	۱	۰	
	۲۴	۰/۵	۲۱	۶۰	۱۱۲	۲۰۱	۲۴۵	۲۹۴*	۶۱	۲	۰	۶
		۰/۷۵	۷	۱۰	۴۶	۱۰۲	۲۹۶	۴۵۸*	۸۱	۰	۰	
	۴۸	۰/۵	۰	۰	۴	۱۰	۵۰	۱۹۷	۲۸۶	۳۹۲*	۶۱	۸
		۰/۷۵	۰	۱	۲	۷	۵۹	۳۵۴	۴۷	۵۳۰*	۰	

عددی قرار می‌دهیم. برآوردهای دو مقدار $\hat{B}_0 = 2\beta_{n,2}[\hat{\sigma}_n^2(2\beta_{n,2}) - \hat{\sigma}_n^2(\beta_{n,2})]$ و $\hat{\sigma}_\infty^2 = \hat{\sigma}_n^2(\beta_{n,1})$ به اندازه‌های بلوک اولیهٔ $\beta_{n,1}$ و $\beta_{n,2}$ بستگی دارد، که مقادیر $\beta_{n,i} = C_i N_n^{1/(d+2i)}; i = 1, 2$ برای آنها پیشنهاد گردید. همچنین با توجه به نتایج تجربی و عددی مقادیر $C_1 = \{0/5, 0/75\}$ و $C_2 = 0/5$ را در این تحلیل مورد استفاده قرار داده‌ایم. جدول ۲ فراوانی مقادیر مختلف برآورد جایگذاری ناپارامتری اندازهٔ بلوک $\hat{\beta}_n = 1, 2, \cdots, 8, 9^+$ را به ازای ۱۰۰۰ بار تکرار شبیه‌سازی مونت کارلو بر اساس دو مدل ۱ و ۲، سه مقدار λ_n و دو مقدار مختلف C_1 نشان می‌دهد. ستون آخر جدول ۲ نیز مقادیر اندازه بلوک بهینهٔ β_n^{opt} حاصل از جدول ۱ را برای مدلهای مختلف نشان می‌دهد. به عنوان مثال برای سطر اول جدول ۲، ابتدا دو اندازه بلوک اولیهٔ پیشنهادی $\beta_{n,1} = 0/5(144)^{1/4} \simeq 2$ و $\beta_{n,2} = 0/5(144)^{1/6} \simeq 2$ محاسبه و بر اساس آنها برآورد بوت‌استرپ بلوک مجزای $\hat{\sigma}_n^2(\beta_n)$ براساس بلوکهایی به اندازه‌های ۲ و ۴ از مشاهدات شبیه‌سازی شده از ساختار مورد نظر به دست آورده شده است. سپس ۲ مقدار $\hat{B}_0 = 4[\hat{\sigma}_n^2(4) - \hat{\sigma}_n^2(2)]$ و $\hat{\sigma}_\infty^2 = \hat{\sigma}_n^2(2)$ محاسبه و در نهایت برآورد جایگذاری ناپارامتری اندازه بلوک $\hat{\beta}_n = (144\hat{B}_0^2/2\hat{\sigma}_\infty^4)^{1/4}$ به دست می‌آید که ممکن است یکی از اندازه بلوکهای ۱ تا 9^+ در جدول ۲ باشد. همان طور که ملاحظه می‌شود، مد مقادیر $\hat{\beta}_n$ با β_n^{opt} حاصل از جدول ۱ در حالتهای

مختلف برابر است که نشان می‌دهد $\hat{\beta}_n$ یک برآورد مناسب برای β_n^{opt} است.

۶ روش بوت‌استرپ نیم‌پارامتری

روشهای بوت‌استرپ بلوک فضایی در عمل با محدودیتها و نقاط ضعفی همراه هستند. از جمله تعیین اندازه بلوک بهینه بسیار مشکل و واریانس برآوردگرها عموماً کم برآورد می‌شوند. در این بخش محدودیتها و نقاط ضعف روشهای بوت‌استرپ بلوک فضایی را مورد بررسی قرار داده و روش بوت‌استرپ نیم پارامتری (SPB) را در آمار فضایی معرفی و در یک مطالعهٔ شبیه‌سازی سه روش MBB، SBB و SPB مورد مقایسه قرار می‌گیرند (ایران‌پناه و همکاران، ۲۰۱۱) و نشان داده می شود برآورد واریانس میانگین نمونه به روش SPB دقیق‌تر از دو روش MBB و SBB است. در روش SPB ابتدا با برآورد میانگین و ساختار همبستگی و حذف آنها از مدل، باقیمانده‌های ناهمبستهٔ فاقد روند به دست می‌آیند. سپس با اجرای روش IIDB برای باقیمانده‌های ناهمبسته و تبدیل عکس، نمونهٔ بوت‌استرپ تولید می‌گردد.

۱.۶ *محدودیتها و نقاط ضعف روشهای بوت‌استرپ بلوک فضایی*

روشهای بوت‌استرپ بلوک فضایی شامل بلوک متحرک و بلوک مجزا در مقایسه با روش SPB در عمل با محدودیتها و نقاط ضعف زیر همراه هستند:

۱) دقت برآوردگرهای بوت‌استرپ بلوکی بستگی به اندازهٔ بلوک دارند. از طرفی اندازه بلوک بهینه (قضیهٔ ۱ در بخش ۱.۵) وابسته به پارامترهای نامعلوم است که برآورد آنها بستگی به ساختار همبستگی فضایی داده‌ها دارند و عمدتاً پیچیده و مشکل است.

۲) تعیین اندازه بلوک بهینه برای برآوردگرهای مختلف مانند میانگین نمونه، برآوردگر پارامترهای تغییرنگار و پیشگوی جایگذاری و اندازه‌های دقت مختلف مانند اریبی و واریانس متفاوت می‌باشد.

۳) تعیین اندازه بلوک بهینه برای برآورد اندازهٔ دقت برآوردگرهایی که شکل بسته ندارند، مانند برآورد پارامترهای تغییرنگار، مقدور نمی‌باشد.

۴) دقت برآوردگرهای بوت‌استرپ بلوکی، علاوه بر اندازهٔ بلوک بستگی به شکل مجموعه نمونهٔ اولیهٔ D_0 (رابطهٔ ۷) دارند. تعیین شکل بهینهٔ D_0 برای اندازه‌های دقت و برآوردگرهای مختلف پیچیده و مشکل است.

۵) بررسی سازگاری برآورد اندازه دقت برآوردگرهای بوت‌استرپ بلوک فضایی و تعیین اندازه بلوک بهینه نیاز به شرایط آمیختن قوی میدان تصادفی دارند. شرط آمیختن قوی، وابستگی را در فاصله‌های کوچک بین موقعیتهای میدان تصادفی فراهم می‌آورد و با افزایش فاصلهٔ بین موقعیتها این وابستگی کاهش می‌یابد. از طرفی چون در روش بوت‌استرپ بلوک فضایی، بازنمونه‌گیری از بلوکها انجام می‌شود بنابراین وابستگی داخل بلوکها باید زیاد و بین بلوکها کم باشد. این مسئله با شرط آمیختن قوی میدان تصادفی تحقق می‌یابد.

۶) روشهای بوت‌استرپ بلوک فضایی برای مشاهدات واقع بر یک مشبکهٔ منظم در $\mathbf{Z}^d$ کاربرد دارند. اگر مشاهدات به طور نامنظم در $\mathbf{R}^d$ قرار داشته باشند، نتایج به دست آمده از دقت لازم برخوردار نخواهند بود.

۷) در کاربرد روش بوت‌استرپ بلوک فضایی با توجه به اندازه بلوک β_n ممکن است نتوان Kای به دست آورد به طوری که $N = K\beta_n^d$، که در آن $K \in \mathbf{N}$ تعداد بلوکهای بازنمونه‌گیری شده برای به دست آوردن یک نمونهٔ بوت‌استرپ است. به بیان دیگر، با توجه به تعیین اندازه بلوک β_n، ممکن است اندازهٔ نمونهٔ جدید برابر $N_1 = K\beta^d \leq N$ باشد، که در این صورت $N - N_1$ مشاهده از مجموعهٔ مشاهدات بدون استفاده خواهند شد.

۸) در برآورد واریانس برآوردگرها $\hat{\sigma}_n^2(\beta_n)$، فقط تغییرات بین بلوکها در نظر گرفته می‌شود و تغییرات داخل بلوکها مورد نظر قرار نمی‌گیرد. به همین دلیل معمولاً $\hat{\sigma}_n^2(\beta_n)$ یک کم برآورد σ_n^2 است.

فرض کنید $\mathcal{Z} = (Z(s_1), \ldots, Z(s_N))^T$ بردار مشاهداتی از میدان تصادفی $\{Z(s) : s \in D \subset \mathbf{R}^d\}$ باشد. گیریم میدان تصادفی $Z(\cdot)$ به صورت $Z(s) = \mu(s) + \delta(s)$ تجزیه شده باشد که در آن $\mu(\cdot)$ ساختار میانگین یا روند و $\delta(\cdot)$ ساختار خطا به صورت یک میدان تصادفی مانای با میانگین صفر و ماتریس کوواریانس معین مثبت Σ با مؤلفهٔ (i,j)ام $\sigma(s_i - s_j)$ است. تجزیهٔ چولسکی ماتریس Σ را به صورت $\Sigma = LL^T$ در نظر بگیرید، که در آن L یک ماتریس $N \times N$ پایین مثلثی است. اگر $\varepsilon = L^{-1}\delta$ آنگاه $\varepsilon = (\varepsilon(s_1), \ldots, \varepsilon(s_N))^T$ یک بردار از متغیرهای تصادفی ناهمبسته با میانگین صفر و ماتریس کوواریانس واحد از یک توزیع نامعلوم $F(\epsilon)$ است. در روش SPB، نیاز به تابع توزیع تجربی $F_N(\epsilon)$ به عنوان برآورد تابع توزیع $F(\epsilon)$ داریم. مراحل مختلف الگوریتم SPB به صورت زیر است:

مرحلهٔ ۱. برآورد و حذف روند.

ابتدا میانگین یا روند $\mu(\cdot)$ با استفاده از الگوریتم پالایش میانه (کرسی، ۱۹۹۳) برآورد و با $\hat{\mu}(\cdot)$ نمایش داده می‌شود. سپس باقیمانده‌های فاقد روند $R(s_i) = Z(s_i) - \hat{\mu}(s_i); i = 1, \cdots, N$ به دست می‌آیند.

مرحلهٔ ۲. برآورد و حذف ساختار همبستگی.

گیریم $\hat{\Sigma} = \Sigma_{\hat{\theta}}$ برآورد ساختار همبستگی باقیمانده‌های $R(s_i)$ باشد، که یک ماتریس متقارن $N \times N$ معین مثبت با مؤلفهٔ (i,j)ام برآورد جایگذاری همتغییرنگار به صورت $\hat{\sigma}(s_i - s_j) = \sigma(s_i - s_j; \hat{\theta})$ است. آنگاه $\hat{\varepsilon} = (\hat{\varepsilon}(s_1), \ldots, \hat{\varepsilon}(s_N))^T = \hat{L}^{-1}\mathcal{R}$ یک بردار از باقیمانده‌های ناهمبسته است، که در آن $\hat{L}$ یک ماتریس $N \times N$ پایین مثلثی حاصل از تجزیهٔ چولسکی $\hat{\Sigma} = \hat{L}\hat{L}^T$ و $\mathcal{R} = (R(s_1), \ldots, R(s_N))^T$ بردار باقیمانده‌ها است.

مرحلهٔ ۳. محاسبهٔ تابع توزیع تجربی $F_N(\epsilon)$.

فرض کنید $\tilde{\varepsilon} = (\tilde{\varepsilon}(s_1), \ldots, \tilde{\varepsilon}(s_N))^T$ مقادیر مرکزی شدهٔ بردار $\hat{\varepsilon}$ باشد، که در آن $\tilde{\varepsilon}(s_i) = \hat{\varepsilon}(s_i) - N^{-1}\sum_{j=1}^{N}\hat{\varepsilon}(s_j); i = 1, \cdots, N$. تابع توزیع تجربی مجموعهٔ باقیمانده‌های ناهمبستهٔ مرکزی $\{\tilde{\varepsilon}(s_1), \ldots, \tilde{\varepsilon}(s_N)\}$ به صورت

$F_N(\epsilon) = N^{-1} \sum_{i=1}^{N} I(\tilde{\varepsilon}(s_i) \leq \epsilon)$ محاسبه شوند.

مرحلهٔ ۴. بازنمونه‌گیری و نمونهٔ بوت‌استرپ.

حال الگوریتم بوت‌استرپ IID برای بردار باقیمانده‌های ناهمبستهٔ مرکزی شدهٔ $\tilde{\varepsilon}$ مورد استفاده قرار می‌گیرد. برای این منظور، ابتدا متغیرهای تصادفی مستقل و هم‌توزیع $\varepsilon^*(s_1), \ldots, \varepsilon^*(s_N)$ از توزیع مشترک $F_N(\epsilon)$ به دست می‌آید. به بیان دیگر، $\varepsilon^* = (\varepsilon^*(s_1), \ldots, \varepsilon^*(s_N))^T$ یک نمونهٔ SRSW از مجموعهٔ باقیمانده‌های ناهمبستهٔ مرکزی $\{\tilde{\varepsilon}(s_1), \ldots, \tilde{\varepsilon}(s_N)\}$ است. سپس نمونهٔ بوت‌استرپ $\mathcal{Z}^* = (Z^*(s_1), \ldots, Z^*(s_N))^T$ با استفاده از تبدیل عکس $\mathcal{Z}^* = \hat{\mu} + \hat{L}\varepsilon^*$ به دست می‌آید، که در آن $\hat{\mu} = (\hat{\mu}(s_1), \ldots, \hat{\mu}(s_N))^T$ برآورد بردار میانگین می‌باشد.

مرحلهٔ ۵. برآوردگر بوت‌استرپ.

اگر $T = t(\mathcal{Z}; \mu, \theta)$ کمیت تصادفی مورد نظر و $\hat{T} = t(\mathcal{Z}; \hat{\mu}, \hat{\theta})$ برآوردگر جایگذاری آن باشد، آنگاه نسخهٔ SPB برآوردگر $\hat{T}$ به صورت $T^* = t(\mathcal{Z}^*; \hat{\mu}, \hat{\theta})$ ارائه می‌گردد.

مرحلهٔ ۶. برآورد نظری بوت‌استرپ اندازه‌های دقت.

برآورد نظری بوت‌استرپ اریبی، واریانس و توزیع T به ترتیب به صورت $\mathrm{Bias}_*(T^*) = \mathrm{E}_*(T^*) - T$، $\mathrm{Var}_*(T^*) = \mathrm{E}_*[T^* - \mathrm{E}_*(T^*)]^2$ و $\mathrm{G}_*(t) = \mathrm{P}_*(T^* \leq t)$ است، که در آنها E_*، Var_* و P_* به ترتیب امید ریاضی، واریانس و احتمال شرطی بوت‌استرپ به شرط $Z(s_1), \ldots, Z(s_N)$ است.

مرحلهٔ ۷. برآورد تجربی بوت‌استرپ اندازه‌های دقت.

اگر $\mathrm{Bias}_*(T^*)$، $\mathrm{Var}_*(T^*)$ و $\mathrm{G}_*(t)$ جوابهای دقیق و بسته‌ای نداشته باشند، با استفاده از شبیه‌سازی مونت کارلو و تکرار B بار مراحل ۵-۱ و محاسبه $T_1^*, \ldots, T_B^*$، برآورد تجربی بوت‌استرپ اریبی، واریانس و توزیع T به ترتیب از روابط (۴) تا (۶) در مرحلهٔ ۴ روش IIDB در بخش ۳ محاسبه می‌گردند.

۳.۶ برآورد بوت‌استرپ نیم‌پارامتری اندازه‌های دقت

در ادامهٔ این بخش نشان داده می‌شود برآورد نظری اریبی و واریانس به روش SPB در مرحلهٔ ۶ برای میانگین نمونه‌ای و پیشگوهای جایگذاری کریگیدن ساده، معمولی و عام دارای شکل بسته هستند، بنابراین برای آنها نیازی به انجام مرحلهٔ ۷ نمی‌باشد.

لم ۱: فرض کنید $\mathcal{Z} = (Z(s_1), \ldots, Z(s_N))^T$ بردار مشاهداتی از میدان تصادفی $\{Z(s) : s \in D\}$ با میانگین $\mu(s) = \mathrm{E}[Z(s)]$ و ماتریس کوواریانس Σ باشد. اگر $\mathcal{Z}^* = (Z^*(s_1), \ldots, Z^*(s_N))^T$ یک نمونهٔ بوت‌استرپ به روش SPB باشد، آنگاه

$$\begin{aligned} E_*(\mathcal{Z}^*) &= \hat{\mu}, \\ Var_*(\mathcal{Z}^*) &= S_{\tilde{\varepsilon}}^2 \hat{\Sigma}, \end{aligned}$$

که در آن $S_{\tilde{\varepsilon}}^2 = N^{-1} \sum_{i=1}^{N} \tilde{\varepsilon}^2(s_i)$ واریانس نمونهٔ باقیمانده‌های ناهمبستهٔ مرکزی $\{\tilde{\varepsilon}(s_1), \ldots, \tilde{\varepsilon}(s_N)\}$ و $\hat{\Sigma} = \Sigma_{\hat{\theta}}$ برآورد جایگذاری Σ می‌باشد.

اثبات: در مرحلهٔ ۴ الگوریتم SPB، چون $\varepsilon^* = (\varepsilon^*(s_1), \ldots, \varepsilon^*(s_N))^T$ یک نمونهٔ SRSW از مجموعهٔ باقیمانده‌های ناهمبستهٔ مرکزی $\{\tilde{\varepsilon}(s_1), \ldots, \tilde{\varepsilon}(s_N)\}$ است پس

$$\begin{aligned} E_*(\varepsilon^*) &= \bar{\tilde{\varepsilon}} = 0, \\ Var_*(\varepsilon^*) &= S_{\tilde{\varepsilon}}^2 I_{N \times N}, \end{aligned}$$

در نتیجه

$$\begin{aligned} E_*(\mathcal{Z}^*) &= E_*(\hat{\mu} + \hat{L}\varepsilon^*) = \hat{\mu}, \\ Var_*(\mathcal{Z}^*) &= Var_*(\hat{\mu} + \hat{L}\varepsilon^*) = S_{\tilde{\varepsilon}}^2 \hat{\Sigma}. \end{aligned}$$

قضیهٔ ۳: اگر میانگین میدان تصادفی ثابت به صورت $\mu(s) = \mu$ باشد، آنگاه نسخهٔ SPB برای $T = \sqrt{N}(\bar{Z} - \mu)$ به صورت $T^* = \sqrt{N}(\bar{Z}^* - \bar{Z})$ است، که در آن $\bar{Z}$ و $\bar{Z}^*$ به ترتیب میانگین نمونهٔ معمولی و بوت‌استرپ می‌باشند و داریم

$$\begin{aligned} E_*(T^*) &= 0, \\ Var_*(T^*) &= S_{\tilde{\varepsilon}}^2 (N^{-1} I^T \hat{\Sigma} I). \end{aligned}$$

اثبات: با استفاده از لم ۱،

$$\begin{aligned} E_*(\bar{Z}^*) &= N^{-1}\sum_{i=1}^{N} E_*[Z^*(s_i)] \\ &= E_*[Z^*(s_i)] \\ &= \bar{Z}. \end{aligned}$$

در نتیجه، $\mathrm{E}_*(T^*) = 0$. همچنین

$$\begin{aligned} Var_*(T^*) &= NVar_*(\bar{Z}^*), \\ &= N^{-1}I^T Var_*(\mathcal{Z}^*)I, \\ &= S_{\tilde{\varepsilon}}^2(N^{-1}I^T\hat{\Sigma}I). \end{aligned}$$

قضیهٔ ۴: اگر $\hat{Z}(s_0) = \sigma^T\Sigma^{-1}(\mathcal{Z}-\mu)+\mu(s_0)$ و $\hat{\hat{Z}}(s_0) = \hat{\sigma}^T\hat{\Sigma}^{-1}(\mathcal{Z}-\hat{\mu})+\hat{\mu}(s_0)$ به ترتیب کریگیدن ساده و پیشگوی جایگذاری آن باشد، آنگاه نسخهٔ SPB پیشگوی جایگذاری $\hat{\hat{Z}}(s_0)$ به صورت $Z^*(s_0) = \hat{\sigma}^T\hat{\Sigma}^{-1}(\mathcal{Z}^* - \hat{\mu}) + \hat{\mu}(s_0)$ است و

$$\begin{aligned} Bias_*[Z^*(s_0)] &= 0, \\ Var_*[Z^*(s_0)] &= S_{\tilde{\varepsilon}}^2\hat{\sigma}^T\hat{\Sigma}^{-1}\hat{\sigma}. \end{aligned}$$

اثبات: با استفاده از لم ۱،

$$\begin{aligned} Bias_*[Z^*(s_0)] &= E_*[Z^*(s_0)] - \hat{\mu}(s_0) \\ &= E_*[\hat{\sigma}^T\hat{\Sigma}^{-1}(\mathcal{Z}^* - \hat{\mu}) + \hat{\mu}(s_0)] - \hat{\mu}(s_0) \\ &= \hat{\sigma}^T\hat{\Sigma}^{-1}[E_*(\mathcal{Z}^*) - \hat{\mu})] = 0 \\ Var_*[Z^*(s_0)] &= Var_*[\hat{\sigma}^T\hat{\Sigma}^{-1}(\mathcal{Z}^* - \hat{\mu}) + \hat{\mu}(s_0)] \\ &= Var_*(\hat{\sigma}^T\hat{\Sigma}^{-1}\mathcal{Z}^*) \end{aligned}$$

$$\begin{aligned} &= (\hat{\sigma}^T\hat{\Sigma}^{-1})Var_*(\mathcal{Z}^*)(\hat{\sigma}^T\hat{\Sigma}^{-1})^T \\ &= S_{\hat{\varepsilon}}^2\hat{\sigma}^T\hat{\Sigma}^{-1}\hat{\sigma}. \end{aligned}$$

توجه کنید که $\sigma_k^{2*}(s_0) = \hat{\sigma}(s_0,s_0) - S_{\hat{\varepsilon}}^2\hat{\sigma}^T\hat{\Sigma}^{-1}\hat{\sigma}$.

اگر $\hat{Z}(s_0) = \lambda\mathcal{Z}$ و $\hat{\hat{Z}}(s_0) = \hat{\lambda}\mathcal{Z}$ به ترتیب کریگیدن معمولی یا عام و پیشگوی جایگذاری آنها باشد، که در آنها ضریب λ از رابطه (۲) محاسبه می‌شود و $\hat{\lambda} = \lambda_{\hat{\theta}}$ برآورد جایگذاری آنها می‌باشد، آنگاه نسخهٔ SPB پیشگوی جایگذاری $\hat{\hat{Z}}(s_0)$ به صورت $Z^*(s_0) = \hat{\lambda}\mathcal{Z}^*$ است و مشابه قضیهٔ ۴ می‌توان نشان داد $\mathrm{Var}_*[Z^*(s_0)] = S_{\hat{\varepsilon}}^2\hat{\lambda}^T\hat{\Sigma}\hat{\lambda}$.

۷ مطالعه شبیه سازی

در این بخش روشهای SBB، MBB و SPB را برای برآورد واریانس میانگین نمونه $\sigma_n^2 = N\mathrm{Var}(\bar{Z}_n) = N^{-1}\mathbf{1}^T\Sigma\mathbf{1}$ مورد مقایسه قرار می‌دهیم. فرض کنید $\{Z(s) : s \in \mathbf{Z}^2\}$ یک میدان تصادفی گاوسی مانای مرتبهٔ دوم با میانگین صفر و تغییرنگار نمایی (۱) با پارامترهای $\theta_1 = (1,1,1)$ (همبستگی ضعیف) و $\theta_2 = (0,2,2)$ (همبستگی قوی) باشد. داده‌های فضایی $\mathcal{Z} = \{Z(s_1),\ldots,Z(s_N)\}$ را در یک مشبکهٔ منظم مربعی $D_n = n\times n$، برای مقادیر $n = 12,24$ در نظر می‌گیریم. برای انجام روش بوت‌استرپ بلوک فضایی، ناحیهٔ نمونه‌گیری را به صورت $D_n = [-6,6)\times[-6,6)$ و $[-12,12)\times[-12,12)$ با ثابت مقیاس‌بندی $\lambda_n = 12,24$ و مجموعه نمونهٔ اولیهٔ $D_0 = [-\frac{1}{2},\frac{1}{2})$ در نظر می‌گیریم. برای مثال، برای اندازه نمونهٔ $N = \lambda_n^2 = 144$ و اندازه بلوک $\beta_n = 2$، ناحیهٔ نمونه‌گیری D_n (۸) به $K = |\mathcal{K}_n| = 36$ زیر ناحیه یا بلوک به صورت

$$D_n(k) = [2k_1, 2k_1+2)\times[2k_2, 2k_2+2);\quad k = (k_1,k_2) \in \mathcal{K}_n$$

افراز می‌شود، که در آن $\mathcal{K}_n = \{k \in \mathbf{Z}^2, -3 \le k_1,k_2 < 3\}$. در روش SBB، بازنمونه‌گیری از بلوکهای مجزای

$$\mathcal{B}_{n,SBB}(i) = [2i_1, 2i_1+2)\times[2i_2, 2i_2+2);\quad i = (i_1,i_2) \in \mathcal{I}_n$$

انجام می‌شود، که در آن $\mathcal{I}_n = \{i \in \mathbf{Z}^۲,\ -۳ \le i_۱, i_۲ < ۳\}$ در روش MBB، بازنمونه‌گیری از بلوکهای متحرک

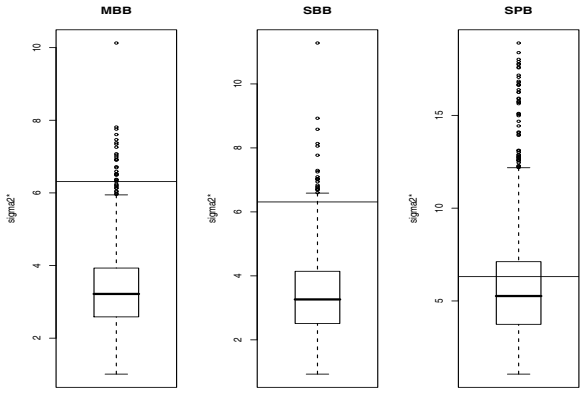
$$\mathcal{B}_{n,MBB}(j) = [j_۱, j_۱ + ۲) \times [j_۲, j_۲ + ۲);\ \ j = (j_۱, j_۲) \in \mathcal{J}_n$$

انجام می‌شود، که در آن $\mathcal{J}_n = \{j \in \mathbf{Z}^۲,\ -۶ \le j_۱, j_۲ < ۴\}$. تعداد کل بلوکهای مجزا و متحرک در این شبکه به ترتیب ۳۶ و ۱۰۰ است، که از بین آنها به روش SRSW، ۳۶ بلوک به اندازهٔ ۴ را انتخاب و از به هم پیوستن آنها یک نمونهٔ بوت‌استرپ $\mathcal{Z}^* = (Z^*(s_۱), \ldots, Z^*(s_N))$ به اندازهٔ $N = ۱۴۴$ تولید می‌شود. برآوردگر واریانس $\hat{\sigma}^۲_n(\beta_n)$ برای $\sigma^۲_n = N\mathrm{Var}(\bar{Z}_n)$ به دو روش SBB و MBB دارای شکل بسته به صورت

$$\begin{aligned}
\hat{\sigma}^۲_{n,SBB}(\beta_n) &= |\mathcal{I}_n|^{-۱} \sum_{i \in \mathcal{I}_n} \beta_n^۲ (\bar{Z}_{i,n} - \hat{\mu}_n)^۲,\ \ \hat{\mu}_n = |\mathcal{I}_n|^{-۱} \sum_{i \in \mathcal{I}_n} \bar{Z}_{i,n} = \bar{Z}_n \\
\hat{\sigma}^۲_{n,MBB}(\beta_n) &= |\mathcal{J}_n|^{-۱} \sum_{j \in \mathcal{J}_n} \beta_n^۲ (\bar{Z}_{j,n} - \hat{\mu}_n)^۲,\ \ \hat{\mu}_n = |\mathcal{J}_n|^{-۱} \sum_{j \in \mathcal{J}_n} \bar{Z}_{j,n} \ne \bar{Z}_n
\end{aligned}$$

می‌باشند، که در آن $\bar{Z}_{i,n}$ و $\bar{Z}_{j,n}$ به ترتیب میانگین نمونه $\beta^۲_n$ مشاهده در بلوکهای مجزای $\mathcal{B}_{n,SBB}(i),\ i \in \mathcal{I}_n$ و بلوکهای متحرک $\mathcal{B}_{n,MBB}(j),\ j \in \mathcal{J}_n$ می‌باشند. اندازه بلوکهای β_n را برای دو مقدار $n = ۱۲, ۲۴$ به ترتیب $(۲, ۳, ۴, ۶)$ و $(۲, ۳, ۴, ۶, ۸, ۱۲)$ در نظر می‌گیریم. مقدار $\sigma^۲_n$ برای ساختار همبستگی ضعیف $\theta_۱$ و دو مقدار n به ترتیب ۶/۳۱۱ و ۶/۸۹۰ و همچنین برای ساختار همبستگی قوی $\theta_۲$ به ترتیب ۳۲/۰۷۴ و ۴۰/۵۹۸ است. حال برای هر دو حالت ساختار همبستگی و دو مقدار n و همچنین اندازه بلوکهای مورد نظر β_n مقدار $\hat{\sigma}^۲_n(\beta_n)$ را برای دو روش SBB و MBB محاسبه می‌کنیم. برای تعیین اندازه بلوک بهینهٔ β^{opt}_n، مشابه جدول ۱ میانگین توان دوم خطای نسبی $\mathrm{E}[\hat{\sigma}^۲_n(\beta_n)/\sigma^۲_n - ۱]^۲$ برآوردگر واریانس $\hat{\sigma}^۲_n(\beta_n)$ به دو روش MBB و SBB برای مقادیر مختلف β_n بر اساس ۱۰۰۰ بار تکرار شبیه‌سازی مونت کارلو محاسبه می‌گردد. اندازه‌های بلوک بهینهٔ β^{opt}_n در دو روش MBB و SBB در جدول ۳ ارائه شده‌اند.

برای به دست آوردن نسخهٔ SPB برآوردگر واریانس $\hat{\sigma}^۲_n$ برای $\sigma^۲_n = N\mathrm{Var}(\bar{Z}_n)$ مراحل ۲ تا ۵ روش SPB در بخش ۲.۶ را در نظر می‌گیریم. ابتدا با استفاده از



شکل ۱: نمودارهای جعبه‌ای ۱۰۰۰ تکرار شبیه‌سازی برای برآورد بوت‌استرپ واریانس میانگین نمونه به سه روش در حالت $n = ۱۲$ و $\theta = \theta_۱$.

برآورد جایگذاری هم‌تغییرنگار $\hat{\sigma}(h;\theta) = \sigma(h;\hat{\theta})$ که در آن $\hat{\theta} = (\hat{c}_۰, \hat{c}_۱, \hat{a})$ برآورد پارامترهای هم‌تغییرنگار نمایی است، برآورد جایگذاری ماتریس کوواریانس $\hat{\Sigma} = \Sigma_{\hat{\theta}}$ را به دست می‌آوریم. اگر $\hat{L}$ تجزیهٔ چولسکی ماتریس $\hat{\Sigma}$ باشد، آنگاه $\hat{\varepsilon} = \hat{L}^{-۱}\mathcal{Z}$ یک بردار از مقادیر ناهمبسته خواهد بود. حال بردار بوت‌استرپ $\varepsilon^* = (\varepsilon^*(s_۱), \ldots, \varepsilon^*(s_N))^T$ را به روش SRSW از $\{\tilde{\varepsilon}(s_۱), \ldots, \tilde{\varepsilon}(s_N)\}$ به دست می‌آوریم، که در آن $\tilde{\varepsilon}(\cdot)$ مقدار مرکزی $\hat{\varepsilon}(\cdot)$ است. سپس با استفاده از تبدیل عکس $\mathcal{Z}^* = \hat{L}\varepsilon^*$ نمونهٔ SPB به صورت $\mathcal{Z}^* = (Z^*(s_۱), \ldots, Z^*(s_N))^T$ حاصل می‌گردد. سرانجام برآورد SPB برای $\sigma_n^۲ = N\mathrm{Var}(\bar{Z}_n)$ با شکل بسته به صورت

$$\hat{\sigma}^۲_{n,SPB} = S^۲_{\tilde{\varepsilon}}(N^{-۱}\mathbf{۱}^T\hat{\Sigma}\mathbf{۱}), \quad S^۲_{\tilde{\varepsilon}} = N^{-۱}\sum_{i=۱}^{N}\tilde{\varepsilon}^۲(s_i)$$

محاسبه می‌گردد.

شکل ۱ نمودار جعبه‌ای برآوردگر واریانس $\hat{\sigma}_n^۲$ برای $\sigma_n^۲ = N\mathrm{Var}(\bar{Z}_n)$ به سه روش MBB، SBB و SPB به ازای $n = ۱۲$ و $\theta = \theta_۱$ بر اساس ۱۰۰۰ بار تکرار شبیه‌سازی مونت کارلو را نشان می‌دهد. در این شکل مقدار واقعی $\sigma_n^۲ = ۶/۳۱۱$ به صورت خط افقی در نمودارهای جعبه‌ای نشان داده شده است. شکل ۱ نشان می دهد برآورد به روشهای MBB و SBB برای $\sigma_n^۲$ کم برآورد هستند، در حالی که

برآورد به روش SPB از اریبی بسیار کمتری و در نتیجه از دقت بالاتری برخوردار است.

جدول ۳: برآورد اریبی، واریانس و میانگین توان دوم خطای نسبی برآوردگر واریانس $\hat{\sigma}^2_n(\beta_n)$ برای $\sigma^2_n = N\mathrm{Var}(\bar{Z}_n)$ به سه روش MBB، SBB و SPB با استفاده از اندازه بلوک بهینهٔ β_n^{opt}.

θ		$\theta_1 = (1,1,1)$				$\theta_2 = (0,2,2)$			
روش	n	β_n^{opt}	Bias	Var	MSE	β_n^{opt}	Bias	Var	MSE
MBB	۱۲	۳	−۰/۴۷۱	۰/۰۳۳	۰/۲۵۴	۴	−۰/۷۳۷	۰/۰۲۲	۰/۵۶۵
SBB		۳	−۰/۴۵۲	۰/۰۴۱	۰/۲۴۶	۶	−۰/۶۳۷	۰/۰۸۴	۰/۴۹۰
SPB		–	۰/۰۵۹	۰/۲۳۹	۰/۲۴۲	–	−۰/۰۵۶	۰/۳۵۴	۰/۳۵۷
MBB	۲۴	۶	−۰/۳۱۷	۰/۰۵۲	۰/۱۵۳	۸	−۰/۵۶۳	۰/۰۵۲	۰/۳۶۹
SBB		۶	−۰/۲۹۷	۰/۰۶۸	۰/۱۵۶	۸	−۰/۵۰۵	۰/۰۶۴	۰/۳۱۹
SPB		–	۰/۰۰۹	۰/۱۴۵	۰/۱۴۵	–	۰/۰۲۹	۰/۱۹۹	۰/۲۰۰

جدول ۳ برآورد اریبی، واریانس و میانگین توان دوم خطای نسبی برآوردگر واریانس $\hat{\sigma}^2_n(\beta_n)$ برای $\sigma^2_n = N\mathrm{Var}(\bar{Z}_n)$ به سه روش MBB، SBB و SPB با استفاده از اندازه بلوک بهینهٔ β_n^{opt} به ازای $n = ۱۲, ۲۴$ و $\theta = \theta_1, \theta_2$ بر اساس ۱۰۰۰ بار تکرار شبیه‌سازی مونت کارلو را نشان می‌دهد. مقایسهٔ مقادیر MSE در این جدول نشان می‌دهد که روش SPB نسبت به روشهای MBB و SBB از دقت بالاتری برای برآورد σ^2_n به ویژه برای nهای بزرگ و ساختار هم‌بستگی قوی برخوردار هستند.

ایران‌پناه و همکاران (۲۰۱۱) در چند مطالعهٔ شبیه‌سازی دقت بالاتر روش SPB را نسبت به روشهای MBB، SBB، IIDB و روش جایگذاری برای برآورد واریانس میانگین نمونه‌ای، برآورد GLS میانگین و پیشگوی جایگذاری نشان داده‌اند.

۸ مثال کاربردی

در این بخش روش SPB برای تحلیل داده‌های خاکستر ذغال سنگ در شهر گرین پنسیلوانیای آمریکا(کرسی، ۱۹۹۳) مورد استفاده قرار گرفته است. داده‌ها در شبکه‌ای با فاصله ۲۵۰۰ فوت به صورت $\{Z(x,y) : x = ۱,\ldots,۱۶;\ \ y = ۱,\ldots,۲۳\}$ به اندازهٔ $N = ۲۰۷$ موقعیت قرار دارند. این داده‌ها همسانگرد ولی دارای روند

می‌باشند. الگوریتم SPB برای این داده‌های خاکستر ذغال سنگ به صورت مراحل زیر انجام می‌شود:

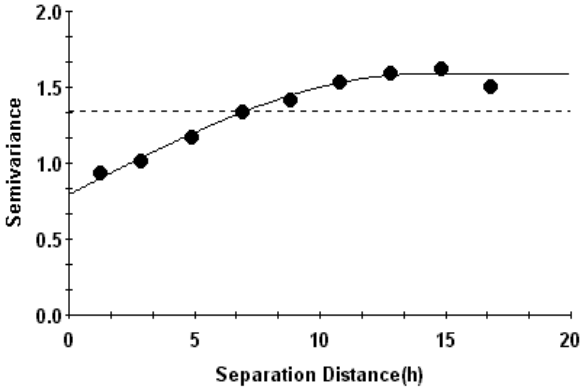
۱) برآورد و حذف روند:
ابتدا روند موجود در داده‌ها در قالب تابع میانگین $\mu(\cdot)$ با استفاده از روش پالایش میانه به صورت

$$\hat{\mu}(s_i) = \hat{a} + \hat{r}_k + \hat{c}_l; \quad s_i = (x_l, y_k),\ k = ۱, \ldots, ۲۳, l = ۱, \ldots, ۱۶,$$
$$\hat{a} = ۹/۸۲۹; \quad \hat{r}_۱ = ۰/۰۹۹, \ldots, \hat{r}_{۲۳} = ۰/۰۰۰;$$
$$\hat{c}_۱ = ۰/۷۷۹, \ldots, \hat{c}_{۱۶} = -۰/۴۲۱,$$
$$R(s_i) = Z(s_i) - \hat{\mu}(s_i); \quad i = ۱, \ldots, ۲۰۷.$$

برآورد و حذف می‌شود. باقیمانده‌های $R(s_i)$ بدون روند ولی همبسته می‌باشند.

۲) برآورد ساختار همبستگی:
برای برآورد ساختار همبستگی در قالب ماتریس $\hat{\Sigma} = \Sigma_{\hat{\theta}}$، ابتدا تغییرنگار تجربی داده‌های خاکستر ذغال سنگ را رسم و با برازش مدلهای مختلف تغییرنگار پارامتری، مدل کروی با پارامترهای $\hat{\theta} = (\hat{c}_۰, \hat{c}_۱, \hat{a}) = (۰/۷۹۳, ۰/۷۹۴, ۱۳/۹۵)$ برازش داده می‌شود. نمودار نیم تغییرنگار تجربی و مدل کروی برازش یافته برای داده‌های خاکستر ذغال سنگ در شکل ۳ نشان داده شده است.

۳) حذف ساختار همبستگی:
برای حذف ساختار همبستگی ابتدا تجزیهٔ چولسکی ماتریس $\hat{\Sigma}$ به صورت $\hat{\Sigma} = \hat{L}\hat{L}^T$ به دست می‌آید. سپس باقیمانده‌های برآورد شده که ساختار همبستگی در آنها حذف گردیده است به صورت $\hat{\varepsilon} = (\hat{\varepsilon}_۱, \ldots, \hat{\varepsilon}_N) = \hat{L}^{-۱}(\mathcal{Z} - \hat{\mu})$ به دست می‌آیند. سرانجام باقیمانده‌های ناهمبسته به صورت $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - N^{-۱}\sum_{j=۱}^{N} \hat{\varepsilon}_j; \quad i = ۱, \ldots, N$ مرکزی



شکل ۲: نیم تغییرنگار تجربی و برازش مدل کروی برای داده‌های خاکستر ذغال سنگ قبل از حذف ساختار همبستگی.

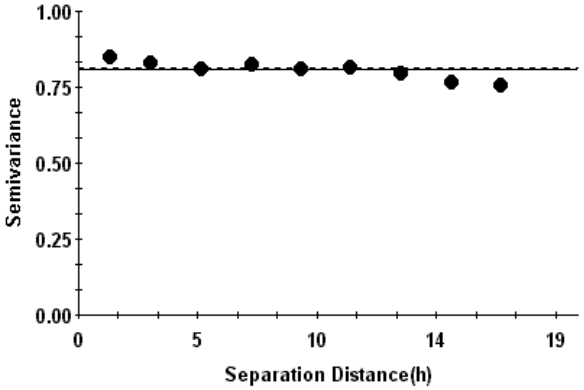
می‌شوند تا میانگین آنها صفر گردد. نمودار نیم تغییرنگار تجربی و برازش مدل خطی برای باقیمانده‌های ناهمبستهٔ مرکزی شده در شکل ۵ بیانگر ناهمبسته بودن باقیمانده‌های مرکزی است، زیرا نیم تغییرنگار تجربی به صورت تخت در آمده و نیم تغییرنگار برازش یافته نیز به صورت مدل خطی با پارامترهای $\hat{\theta} = (\hat{c}_0, \hat{c}_1, \hat{a}) = (۰/۸۱, ۰/۰۰, ۱۶/۷۷)$ است.

۴) نمونهٔ بوت‌استرپ:

اکنون می‌توان روش IIDB را برای بردار باقیمانده‌های ناهمبستهٔ مرکزی $(\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_N)$ مورد استفاده قرار داد. برای این منظور ابتدا $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_N^*)$ به روش SRSW از بردار $(\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_N)$ به دست می‌آید. سپس با تبدیل عکس $\mathcal{Z}^* = \hat{\mu} + \hat{L}\varepsilon^*$، نمونهٔ بوت‌استرپ $\mathcal{Z}^* = (Z^*(s_1), \ldots, Z^*(s_N))$ به دست می‌آید.

۵) برآورد اندازه‌های دقت :

سرانجام اریبی، واریانس و توزیع هر آمارهٔ $T = t(\mathcal{Z})$ می‌تواند با اریبی، واریانس و توزیع آماره بوت‌استرپ $T^* = t(\mathcal{Z}^*)$ تقریب شود. اریبی،

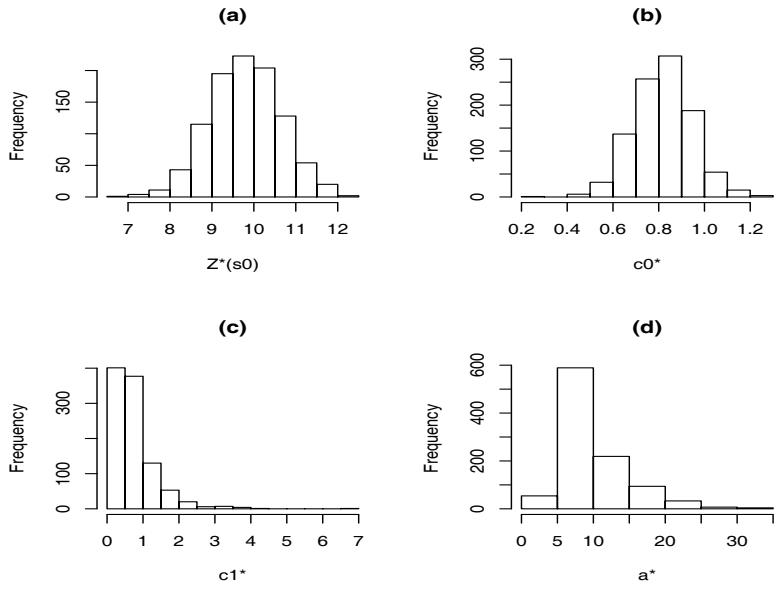


شکل ۳: نیم تغییرنگار تجربی و برازش مدل خطی برای داده‌های خاکستر ذغال سنگ بعد از حذف ساختار همبستگی.

واریانس و توزیع T^* به روش شبیه‌سازی مونت کارلو و تکرار B بار مراحل قبل و محاسبه $T_1^*, \ldots, T_B^*$ به ترتیب از روابط ۴ تا ۶ در مرحلهٔ ۴ روش IIDB در بخش ۳ برآورد می‌شوند. جدول ۴ مقادیر برآوردهای اریبی و واریانس به روش SPB را برای پیشگوی جایگذاری $\hat{\hat{Z}}(5, 6)$ و برآورد پارامترهای تغییرنگار بر اساس $B = 1000$ بار تکرار بوت‌استرپ برای داده‌های خاکستر ذغال سنگ نشان می‌دهد. شکل ۲ بافتنگارهای پیشگوی جایگذاری و برآورد پارامترهای تغییرنگار را بر اساس $B = 1000$ بار تکرار بوت‌استرپ برای داده‌های خاکستر ذغال سنگ نشان می‌دهد.

جدول ۴: برآوردهای اریبی و واریانس به روش SPB برای پیشگوی جایگذاری و برآورد پارامترهای تغییرنگار برای داده‌های خاکستر ذغال سنگ.

Var_*	Bias_*	T_i^*
۰/۷۰۶	-۰/۹۰۱	$Z^*(s_0)$
۰/۰۱۷	۰/۰۰۲	c_0^*
۰/۰۳۷	۰/۰۶۶	c_1^*
۲۱/۶۰۲	-۵/۸۲۹	a^*



شکل ۴: بافتنگارهای (a) پیشگوی جایگذاری و برآوردگرهای پارامتر تغییرنگار (b)اثر قطعه‌ای، (c) آستانه جزیی و (d) دامنه برای داده‌های خاکستر ذغال سنگ.

۹ بحث و نتیجه‌گیری

یکی از روشهای بوت‌استرپ مورد استفاده در داده‌های فضایی بلوک متحرک می‌باشد، که در آن مشاهدات را در بلوکهایی متحرک در نظر گرفته و از آنها بازنمونه‌گیری می‌شود. در این روش مشاهدات مرزی ناحیهٔ نمونه‌گیری نسبت به مشاهدات مرکزی شانس کمتری برای حضور در بلوکها را دارند و این مسئله باعث اریبی در برآوردگرها می‌شود. در این مقاله برای رفع این مشکل، روش بوت‌استرپ بلوک مجزا ارائه گردید، که در آن مشاهدات به بلوکهای مجزا افراز و بازنمونه‌گیری از این بلوکها انجام می‌شود. در روش بوت‌استرپ بلوک مجزا برای استنباط داده‌های فضایی، دقت برآوردگرها به اندازه بلوک بستگی دارد. در ادامهٔ این مقاله، اندازه بلوک بهینه برای برآورد بوت‌استرپ واریانس میانگین نمونه‌ای داده‌های مشبکه‌ای به صورت مجانبی تعیین و برآورد آن به روش جایگذاری ناپارامتری ارائه گردید. همچنین در یک مطالعهٔ شبیه‌سازی نتایج به دست آمده مورد ارزیابی عددی

قرار گرفت.

روشهای بوت‌استرپ بلوک فضایی در عمل با محدودیتها و نقاط ضعفی همراه هستند. از جمله تعیین اندازه بلوک بهینه بسیار مشکل و واریانس برآوردگرها عموماً کم برآورد می‌شوند. در این مقاله ضمن ارائهٔ محدودیتها و نقاط ضعف روشهای بوت‌استرپ بلوک فضایی، روش بوت‌استرپ نیم پارامتری در آمار فضایی معرفی گردید. در یک مطالعهٔ شبیه‌سازی دقت بالاتر روش بوت‌استرپ نیم پارامتری نسبت به دو روش بوت‌استرپ بلوک مجزا و بلوک متحرک برای برآورد واریانس میانگین نمونه‌ای نشان داده شد.

مراجع

ایران‌پناه، ن.، (۱۳۷۷). الگوریتم بوت‌استرپ بیزی، اندیشهٔ آماری، جلد **۳**، شمارهٔ ۲، ۶۷-۶۴.

ایران‌پناه، ن. و پاشا، ع. (۱۳۷۶). آشنایی با الگوریتم بوت استرپ. اندیشهٔ آماری، سال دوم، شمارهٔ ۱، ۴۶-۳۳.

ایران‌پناه، ن.، محمدزاده، م. (۱۳۸۴)، روش بوت استرپ بلوک مجزا در آمار فضایی، نشریهٔ علوم دانشگاه تربیت معلم، جلد **۵**، شماره ۴، ۶۶۶-۶۵۳.

ایران‌پناه، ن.، محمدزاده، م. (۱۳۸۶)، برآورد اندازه‌های دقت کریگیدن به روش خودگردانی بلوکی فضایی، مجلهٔ علوم دانشگاه تهران، جلد **۳۳**، شماره ۳، ۲۴-۱۹.

Buhlmann, P. and Kunsch, H. R. (1999), Comments on "Prediction of Spatial Cumulative Distribution Functions Using Subsampling", *Journal of the American Statistical Association*, **94**, 97-99.

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

Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife., *Annals of Statistics*, **7**, 1-26.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, London.

Ekstrom, E. and Sjostedt-DeLuna, S. (2004), Subsampling Methods to Estimate the Variance of Sample Means Based on Nonstationary Spatial Data with Varying Expected Values, *Journal of the American Statistical Association*, **99**, 82-95.

Hall, P., Horowitz, J.L. and Jing, B.Y. (1995), On the Blocking Rules for the Bootstrap with Dependent Data, *Biometrika*. **82**, 561-574.

Iranpanah, N., Mohammadzadeh, M. and Vahidi Asl, M.G. (2009), Optimal Block Size in Seperate Block Bootstrap to Estimate the Variance of Sample Mean for Lattice Data., *Journal of Science Tehran UniversityIslamic Repoblic of Iran*, **20**(4), 355-364.

Iranpanah, N., Mohammadzadeh, M., and Taylor C.C. (2011), Comparison Between Block and Semi-Parametric Bootstrap Methods for Spatial Data Analysis., *Computational Statistics and Data Analysis*, **55**, 578-587.

Hall, P. (1985), Resampling a Coverage Pattern, *Sthochastic Processes and their Application*, **20**, 231-246.

Kunsch, H.R. (1989), The jackknife and bootstrap for general stationary observation, *The Annals of Statistics*, **17**, 1217-1241.

Lahiri, S.N. (1999), Theoretical Comparisons of Block Bootstrap Methods, *The Annals of Statistics*, **27**, 386-404.

Lahiri, S.N. (2003), *Resampling Methods for Dependent Data.*, Springer-Verlag, New York.

Lahiri, S.N., Kaiser, M.S., Cressie, N. and Hsu, N.J. (1999), Prediction of Spatial Cumulative Distribution Function using Subsampling, *Journal of the American Statistical Association*, **94**, 86-97.

Lahiri, S.N., Furukawa, K. and Lee, Y-D. (2007), A Nonparametric Plug-in Rule for Selecting Optimal Block Lenghts for Block Bootstrap Methods, *Statistical Methodology*, **4**, 292-321.

Liu, R.Y. (1988), Bootstrap Procedures under some Non-i.i.d. Models, *The Annuals of Statistics*, **16**, 1696-1708.

Nordman, D.J. and Lahiri, S.N. (2004), On Optimal Spatial Subsample Size for Variance Estimation, *The Annals of Statistics*, **32**, 1981-2027.

Singh, K. (1981), On the asumptotic accuracy of the Efron's Bootstrap., *Annals of Statistics*, **9**, 1187-1195.

Sjostedt-DeLuna, S. (2001), Resampling Non-Homogeneous Spatial Data with Smoothly Varying Mean Values, *Statistics and Probability Lettrs*, **53**, 373-379.

Zhu, J. and Lahiri, S.N. (2001), Weak Convergence of Blockwise Bootstrapped Empirical Processes for Stationary Random Fields with Statistical Applications., Preprint, Department of Statistics, Iowa State University, Ames, IA.

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱
مجموعه مقالات، ص ۳۵-۵۸

استنباط مبتنی بر درستنمایی در مدل‌های فضایی با پاسخ گسسته: رهیافت همسانه‌سازی داده‌ها

حسین باغیشنی [۱]، محسن محمدزاده [۲]

[۱] *دانشگاه صنعتی شاهرود، گروه ریاضی کاربردی*

[۲] *دانشگاه تربیت مدرس تهران، گروه آمار*

چکیده : در موقعیت‌های کاربردی متعددی با داده‌هایی مواجه می‌شویم که گسسته بوده و به دلیل موقعیت مکانی قرارگیری آن‌ها نسبت به هم در ناحیه تحت مطالعه، نوعی همبستگی فضایی بین آن‌ها وجود دارد. در چنین مواردی، مدل‌های آمیخته خطی تعمیم‌یافته فضایی، انتخابی معمول و مناسب هستند. تحلیل مبتنی بر درستنمایی در این گونه مدل‌ها، به دلیل وجود انتگرال‌های با بعد بالا، پیچیده، طاقت‌فرسا و گاهی نشدنی است. همسانه‌سازی داده‌ها، روشی جدید برای برآورد پارامترهای یک مدل سلسله‌مراتبی است که استنباط‌های بسامدی معتبری فراهم می‌آورد.

در این مقاله به استنباط مبتنی بر درستنمایی در مدل‌های آمیخته خطی تعمیم‌یافته فضایی با روش همسانه‌سازی داده‌ها پرداخته می‌شود. با استفاده از یک مطالعه شبیه‌سازی و یک مثال واقعی در خصوص تعداد تصادفات رانندگی در شهر مشهد،

آدرس الکترونیک مسئول مقاله: حسین باغیشنی، hbaghishani@shahroodut.ac.ir

کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲M۳۰، ۶۲H۱۱

نحوه استفاده و کارایی روش همسانه‌سازی داده‌ها برای تحلیل این مدل‌ها نشان داده می‌شود.

واژه‌های کلیدی : ر وش‌های نمونه‌گیری مونت کارلوی زنجیر مارکوفی، مدل آمیخته خطی تعمیم‌یافته فضایی، همسانه‌سازی داده‌ها.

۱ مقدمه

دگرگونی‌های عمده در حجم و پیچیدگی تحلیل داده‌ها در کنار پیشرفت‌های اساسی در مدل‌های آماری و روش‌های نوین محاسباتی، استفاده از مدل‌های آماری دقیق‌تر اما پیچیده‌تر را به همراه داشته‌اند. مدل‌های آمیخته خطی تعمیم‌یافته[۱] (GLMMs)، به عنوان تعمیمی از مدل‌های خطی تعمیم‌یافته[۲] (GLMs) (مک‌کلا و نلدر، ۱۹۸۹)، یکی از انواع این مدل‌ها محسوب می‌شوند، که ابزار مفیدی را برای تحلیل داده‌های همبسته ناگاوسی (گسسته) مانند داده‌های طولی و خوشه‌ای، فراهم کرده‌اند (بریسلو و کلیتون، ۱۹۹۳). در این رده از مدل‌ها، همبستگی داده‌ها با افزودن عامل اثرات تصادفی به مدل لحاظ می‌گردد. علاقه‌مندی به تحلیل این مدل‌ها به دلیل پهنه وسیع کاربرد آن‌ها، به سرعت رو به فزونی است.

دیگل و همکاران (۱۹۹۸) یک GLMM را در حالتی که ساختار همبستگی داده‌ها از نوع فضایی است، یعنی همبستگی آن‌ها ناشی از موقعیت مکانی داده‌ها در ناحیه تحت مطالعه است، به یک GLMM فضایی[۳] (SGLMM) تعمیم دادند، که در آن همبستگی فضایی داده‌ها از طریق یک میدان تصادفی لحاظ می‌شود و معمولاً با فرض گاوسی بودن این میدان تصادفی، تحلیل می‌شود.

برازش مدل‌های آمیخته خطی تعمیم‌یافته فضایی، موضوع مورد علاقه بسیاری از محققین در سال‌های اخیر بوده است. در بسیاری از مسایل کاربردی، حضور مجموعه داده‌های حجیم همراه با ساختارهای همبستگی پیچیده، تحلیل SGLMMs

---

۱ Generalized Linear Mixed Models

۲ Generalized Linear Models

۳ Spatial GLMM

را با مشکل مواجه می‌کند. روش‌های گوناگونی برای مرتفع ساختن این مشکل پیشنهاد شده‌اند. این روش‌ها شامل هر دو رهیافت بیزی و بسامدی هستند. پیچیدگی محاسبات در روش‌های بسامدی امکان استفاده از استنباط‌های مبتنی بر درستنمایی را ناممکن می‌سازد. محاسبه برآوردگر ماکسیمم درستنمایی[۴] (MLE) در این مدل‌ها، حل عددی انتگرال‌های با بعد بالا را شامل می‌شود. بنابراین انتگرال‌های رام‌نشدنی[۵] عمده‌ترین مشکل بر سر راه استنباط مبتنی بر درستنمایی در این رده از مدل‌ها هستند.

به دلیل پیشرفت‌های محاسباتی الگوریتم‌های مونت کارلوی زنجیر مارکوفی[۶] (MCMC)، معمول‌ترین رهیافت برای استنباط در SGLMMs، رهیافت بیزی است. اما با وجود انجام‌پذیر بودن استنباط‌های بیزی، همواره دو مشکل اساسی انتخاب پیشین‌های مناسب، به ویژه برای پارامترهای ساختار همبستگی، (فونگ و همکاران، ۲۰۱۰) و وابسته بودن دقت استنباط‌ها به این انتخاب‌ها، مطرح هستند.

یک روش جانشین برای محاسبه MLE پارامترها در این رده از مدل‌ها، روش همسانه‌سازی داده‌ها[۷] (DC) است که اولین بار توسط له‌له و همکاران (۲۰۰۷) در مطالعات بوم‌شناختی[۸]، به کار گرفته شد. روش DC، روشی ساده برای محاسبه MLE با بهره‌گیری از الگوریتم‌های MCMC است. این روش چارچوب بیزی را فقط به عنوان ابزار محاسبه MLE به کار می‌گیرد و برآوردهای حاصل از آن در مقایسه با استنباط‌های بیزی، نسبت به انتخاب توزیع‌های پیشین، پایا[۹] هستند. علاوه بر این، مراحل برآورد پارامترهای مدل شامل محاسبه ساده مقادیر میانگین و واریانس نمونه‌ای است که با الگوریتم MCMC تولید می‌شود و نیازی به ماکسیمم‌سازی عددی و مشتق‌گیری از یک تابع پیچیده ندارد. هدف این مقاله، سازوار کردن روش DC برای تحلیل مبتنی بر درستنمایی یک مدل SGLM است.

---

۴ Maximum Likelihood Estimator

۵ Intractable Integrals

۶ Markov Chain Monte Carlo

۷ Data Cloning

۸ Ecological Studies

۹ Invariant

برای این منظور، الگوریتمی مبتنی بر روش DC معرفی می‌شود.

در بخش ۲، مدل‌های SGLM، مشکلات استنباط در آن‌ها و راه حل‌های موجود معرفی می‌شوند. در بخش ۳ ابتدا روش DC تشریح می‌شود، سپس برخی از ویژگی‌های مجانبی برآوردگرهای مبتنی بر DC ارایه می‌گردد. برازش SGLMMs به کمک الگوریتم DC در بخش ۴ صورت می‌پذیرد. در بخش ۵، کارایی این روش در یک مطالعه شبیه‌سازی مورد ارزیابی قرار گرفته و نحوه به کارگیری آن در یک مثال واقعی شامل تعداد تصادفات رانندگی در شهر مشهد، نشان داده می‌شود. در پایان نیز به بحث و نتیجه‌گیری پرداخته خواهد شد.

۲ مدل‌های آمیخته خطی تعمیم‌یافته فضایی

فرض کنید $D \subseteq \Re^d$، $d \geq 1$، ناحیه فضایی مورد علاقه باشد و به ازای $i = 1, \ldots, n$، $Y(\mathbf{s}_i)$ متغیر پاسخ و $\mathbf{x}(\mathbf{s}_i)$ بردار p بعدی از متغیرهای تبیینی برای اثرات ثابت در موقعیت $\mathbf{s}_i \in D$ باشند. یک SGLMM به صورت

$$E\left(Y(s_i)|u(s_i)\right) = g^{-1}\left(x'(s_i)\boldsymbol{\beta} + u(s_i)\right), \quad i = 1, \ldots, n;$$

تعریف می‌شود، که در آن:

الف) $g(\cdot)$ تابع پیوند مشتق‌پذیر و معکوس‌پذیر با دامنه اعداد حقیقی و $\boldsymbol{\beta}$ بردار p بعدی پارامترهای رگرسیونی هستند.

ب) $\{U(\mathbf{s}) : \mathbf{s} \in D\}$ یک میدان تصادفی مانای مرتبه دوم گاوسی با میانگین صفر و تابع کوواریانس $Cov\left(U(\mathbf{s}), U(\mathbf{s}')\right) = C(\mathbf{s} - \mathbf{s}'; \boldsymbol{\theta})$ است، که در آن $C(\cdot; \boldsymbol{\theta})$ یک تابع معین مثبت و $\boldsymbol{\theta}$ بردار پارامترهای وابستگی فضایی است و $\boldsymbol{u} = (u(\mathbf{s}_1), \ldots, u(\mathbf{s}_n))$ تحققی از این میدان تصادفی است.

ج) $Y(\cdot)$ یک میدان تصادفی مستقل شرطی است، یعنی به شرط معلوم بودن $\boldsymbol{u}$، متغیرهای $Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)$ مستقل هستند.

د) توزیع $Y_i = Y(\mathbf{s}_i)$ به شرط $u_i = u(\mathbf{s}_i)$ دارای تابع چگالی نمایی به صورت

$$f(y_i|u_i;\beta,\tau) = exp\left\{\frac{1}{\tau}[a(\mu_i)y_i - b(\mu_i)]\right\}c(\frac{1}{\tau}, y_i), \qquad (1)$$

است، که در آن $\mu_i = \mathrm{E}(Y_i|u_i)$، τ پارامتر پراکندگی و توابع $a(\cdot)$، $b(\cdot)$ و $c(\cdot)$ معلوم هستند. چنانچه τ معلوم باشد، که در این مقاله این گونه فرض می‌شود، تابع چگالی (۱) متعلق به خانواده توزیع‌های نمایی است.

بنابراین تابع درستنمایی حاشیه‌ای SGLMM به صورت

$$L(\beta,\theta;y) \propto \int \prod_{i=1}^{n} f(y_i|u_i;\beta)\phi_n(\boldsymbol{u};0,\Sigma_{\boldsymbol{\theta}})d\boldsymbol{u} \qquad (2)$$

است، که در آن $\phi_n(\cdot;0,\Sigma_{\boldsymbol{\theta}})$ تابع چگالی گاوسی n متغیره با بردار میانگین صفر و ماتریس کوواریانس $n \times n$، $\Sigma_{\boldsymbol{\theta}} = (C(\mathbf{s}_i - \mathbf{s}_j;\boldsymbol{\theta}))$ است.

محاسبه تابع درستنمایی حاشیه‌ای (۲) مستلزم حل انتگرالی با بعدی برابر تعداد مشاهدات است و برای داده‌های حجیم، محاسبه آن کاری طاقت‌فرسا خواهد بود. بنابراین روش‌های تقریب این انتگرال، موضوع اساسی و مورد علاقه در استنباط بسامدی این رده از مدل‌ها است. تاکنون روش‌هایی مانند شبه‌درستنمایی تاوانیده[۱۰] (بریسلو و کلیتون، ۱۹۹۳)، مربع‌بندی گاوس–هرمیت سازوار[۱۱] (پینیرو و بیتس، ۱۹۹۵؛ لسافره و اسپیسنز، ۲۰۰۱)، EM مونت کارلویی[۱۲]، نیوتون–رافسون مونت کارلویی[۱۳] (مک‌کالاک، ۱۹۹۷؛ بوس و هابرت، ۱۹۹۹) و ماکسیمم درستنمایی شبیه‌سازی‌شده[۱۴] (گیر و تامپسون، ۱۹۹۲؛ کریستنسن، ۲۰۰۴) پیشنهاد شده‌اند، که تنها برای رده محدودی از مدل‌ها، از جمله مدل‌های با تابع درستنمایی هموار، کارآمد هستند (مک‌کالاک، ۱۹۹۷).

در مقابل، پس از انقلاب روش‌های نمونه‌گیری MCMC، معمول‌ترین رهیافت برای تحلیل این گونه مدل‌ها، بر رهیافت بیزی پایه‌ریزی شد. دیگل و همکاران

[۱۰] Penalized Quasi-Likelihood

[۱۱] Adaptive Gauss-Hermite Quadrature

[۱۲] Monte Carlo EM

[۱۳] Monte Carlo Newton-Raphson

[۱۴] Simulated Maximum Likelihood

(۱۹۹۸) با استفاده از یک الگوریتم متروپولیس-هستینگز[۱۵] به تحلیل SGLMMs پرداختند. یک الگوریتم MCMC کاراتر تحت عنوان الگوریتم لانگوین-هستینگز[۱۶] برای تحلیل SGLMMs توسط کریستنسن و واگاپترسن (۲۰۰۲) ارائه شد. هرچند در سال‌های اخیر، تحقیقات گسترده‌ای در مورد روش‌های بیزی برای تحلیل SGLMMs انجام شده‌اند، که از جمله آن‌ها می‌توان به کریستنسن و همکاران (۲۰۰۶)، زائو و همکاران (۲۰۰۶)، فونگ و همکاران (۲۰۱۰) و حسینی و همکاران (۲۰۱۱) اشاره کرد، اما با وجود محبوبیت استنباط‌های بیزی، همواره دو مشکل اساسی انتخاب پیشین‌های مناسب و وابستگی دقت استنباط‌های مدل به این انتخاب‌ها، مطرح هستند.

۳ همسانه‌سازی داده‌ها

همسانه‌سازی داده‌ها، ترفندی محاسباتی است که می‌تواند به عنوان روشی جانشین برای استنباط مبتنی بر درستنمایی در مدل‌های پیچیده‌ای مانند SGLMMs به کار گرفته شود. این روش از الگوریتم MCMC برای تولید نمونه از یک توزیع به طور مصنوعی ساخته‌شده[۱۷]، تحت عنوان توزیع مبتنی بر همسانه‌سازی داده‌ها[۱۸]، برای محاسبه MLE پارامترها و برآورد واریانس آن‌ها استفاده می‌کند. در همسانه‌سازی داده‌ها بردار nk بعدی $\mathbf{y}^{(k)} = (\mathbf{y}, \ldots, \mathbf{y})$ از تکرار k بار بردار n بعدی مشاهدات $\mathbf{y}$، تشکیل می‌شود. به همین روش مقادیر متغیرهای تبیینی نیز کپی می‌شوند. همچنین از k تحقق بردار اثرات تصادفی $\boldsymbol{u}$، با استفاده از توزیع احتمالی آن‌ها، بردار $\boldsymbol{u}_{(k)} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$ تولید می‌شود. در نمودار ۱ روند همسانه‌سازی داده‌ها نمایش داده شده است. همان طور که ملاحظه می‌شود، بردارهای $\mathbf{y}$ و متغیرهای تبیینی $\mathbf{x}$ عیناً تکرار شده‌اند ولی هر یک از بردارهای $\boldsymbol{u}_1$ تا $\boldsymbol{u}_k$ به طور مستقل از توزیع متناظر خود تولید شده‌اند. اگرچه در واقعیت k تکرار مستقل یک آزمایش مشابه به

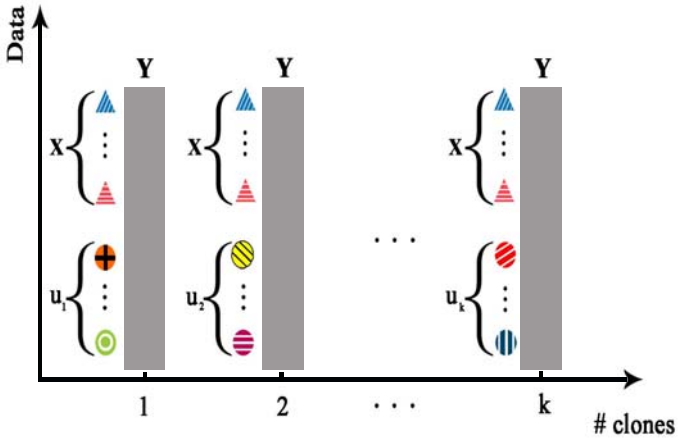
[۱۵] Metropolis-Hastings Algorithm

[۱۶] Langevin-Hastings

[۱۷] Artificially Constructed Distribution

[۱۸] DC-based Distribution

مجموعه داده یکسانی نمی‌انجامد و به طور نظری احتمال چنین رخدادی صفر است، اما این عمل توسط برنامه‌های رایانه‌ای تولید مقادیر تصادفی، قابل انجام است



شکل ۱: ساختار داده‌های همسانه‌سازی شده

توزیع مبتنی بر DC توسط مجموعه داده همسانه‌سازی شده ساخته می‌شود. از آنجا که ایده پشت این روش مبتنی بر ویژگی‌های مجانبی توزیع‌های پسین می‌باشد، تعداد نسخه‌ها، k، باید به اندازه کافی بزرگ اختیار شود به طوری که میانگین نمونه‌ای و ماتریس واریانس مقیاس‌بندی شده به MLE و برآورد واریانس متناظرشان همگرا شوند. برای مقیاس‌بندی ماتریس واریانس از ضریب k^{-1} استفاده می‌شود.

با قرار دادن $\psi = (\beta, \theta)$ و استفاده از مجموعه داده همسانه‌سازی شده جدید، توزیع مبتنی بر DC به صورت

$$\pi^{(k)}(\psi, \boldsymbol{u}|y) \propto \pi^{(k)}(\boldsymbol{u}|y, \psi)\pi^{(k)}(\psi|y), \qquad (۳)$$

ساخته می‌شود. هرچند (۳) شبیه یک توزیع پسین بیزی به نظر می‌رسد، اما ساختار آن مبتنی بر دو تابع است که در واقع تابع درستنمایی و توزیع پیشین نیستند ولی در این مقاله از آن‌ها به عنوان درستنمایی و پیشین یاد می‌شود. بر اساس خواص

مجانبی توزیع حاشیه‌ای مبتنی بر DC پارامترها، $\pi^{(k)}(\boldsymbol{\psi}|\mathbf{y})$، باغیشنی و محمدزاده (۲۰۱۱) نشان دادند، وقتی $k \longrightarrow \infty$، آنگاه

$$\begin{aligned} E^{(k)}(\boldsymbol{\psi}|y) &\longrightarrow \hat{\psi}, \\ V^{(k)}(\boldsymbol{\psi}|y) &\longrightarrow k^{-1} V\left(\hat{\psi}\right), \end{aligned}$$

که در آن $\mathrm{E}^{(k)}(\boldsymbol{\psi}|\mathbf{y})$ و $V^{(k)}(\boldsymbol{\psi}|\mathbf{y})$ به ترتیب امید ریاضی و واریانس توزیع حاشیه‌ای مبتنی بر DC پارامترهای مدل و $\hat{\psi}$ برآوردهای ML پارامترهای $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ است. یکی از مهم‌ترین مزایای روش DC، پایایی نتایج به دست آمده نسبت به انتخاب توزیع‌های پیشین است (له‌له و همکاران، ۲۰۰۷).

۱.۳ ویژگی‌های مجانبی برآوردگرهای مبتنی بر DC

ایده همسانه‌سازی داده‌ها بر رفتار مجانبی توزیع‌های پسین شکل گرفته است. به عبارت دیگر در همسانه‌سازی داده‌ها، کپی کردن بردار داده‌ها به تعداد k بار، در واقع افزایش حجم نمونه است به طوری که شرایط نظری مجانبی توزیع پسین برقرار شود. در این زیربخش به اختصار ویژگی‌های مجانبی برآوردگرهای حاصل از روش DC، در قالب چند قضیه مطرح می‌شوند.

فرض کنید $\boldsymbol{\psi}$ یک بردار q بعدی در فضای پارامتر $\boldsymbol{\Psi} \in \Re^q$ و $\hat{\psi}$ برآوردگر ML متناظر آن باشد. همچنین فرض کنید تابع لگاریتم درستنمایی حاشیه‌ای $\log L_n(\boldsymbol{\psi}|\mathbf{y}) = \ell_n(\boldsymbol{\psi})$، به طور پیوسته دو بار مشتق‌پذیر باشد. علاوه بر این، فرض کنید $\ell_n^{(k)}(\boldsymbol{\psi}) = \log[L_n(\boldsymbol{\psi}|\mathbf{y})]^k$ تابع لگاریتم درستنمایی متناظر با داده‌های همسانه‌سازی شده باشد.

توزیع مجانبی: قضیه ۱ در زیر، نرمال مجانبی بودن برآوردگرهای مبتنی بر DC را نشان می‌دهد.

قضیه ۱ (باغیشنی و محمدزاده، ۲۰۱۱) اگر $\pi^{(k)}(\boldsymbol{\psi}|\mathbf{y})$ تابع چگالی مبتنی بر DC باشد، آنگاه تحت شرایط نظم واکر (۱۹۶۹) برای بعد q، چنانچه $k \longrightarrow \infty$،

$$\sqrt{k}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})|\mathbf{y}^{(k)} \xrightarrow{D} N_q(\circ, \Gamma),$$

که در آن $\Gamma = I^{-1}$ و $I = (\nu_{ij}) = (-\frac{\partial^2 \ell_n(\boldsymbol{\psi})}{\partial\psi_i \partial\psi_j}|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}})$.

پایایی: لهله و همکاران (۲۰۰۷) نشان دادند برای مقادیر بزرگ k، برآوردگرهای حاصل از روش DC نسبت به انتخاب توزیع‌های پیشین، پایا هستند. این ویژگی بارز روش DC، برتری محسوسی را نسبت به استنباط بیزی فراهم می‌آورد که وابستگی آن به توزیع پیشین، بیشترین انتقادها را به همراه دارد. این ویژگی، جدای از ویژگی‌های توزیع پیشین شامل دامنه، کران‌داری و غیره به دست می‌آید. تنها لازم است پیشین در تکیه‌گاه خود مثبت باشد. در واقع، پایایی روش DC و همگرایی توزیع مبتنی بر DC پارامترها به توزیع نرمال با میانگینی برابر MLE و واریانسی برابر $1/k$ واریانس MLE، ایده اساسی استفاده از آن برای استنباط‌های مبتنی بر درستنمایی است. قضیه زیر هر دو ویژگی همگرایی میانگین به MLE و پایایی روش DC را بیان می‌کند.

قضیه ۲ (لهله و همکاران، ۲۰۰۷) فرض کنید $\pi(\boldsymbol{\psi})$ همه جا روی $\boldsymbol{\Psi}$ مثبت باشد و $k \longrightarrow \infty$. آنگاه توزیع مبتنی بر DC، $\pi^{(k)}(\boldsymbol{\psi}|\mathbf{y})$، مستقل از $\pi(\boldsymbol{\psi})$ به یک جرم دیراک[۱۹] واقع در MLE پارامتر $\boldsymbol{\psi}$ میل می‌کند.

سازگاری: قضیه ۳ نیز سازگاری برآوردگرهای مبتنی بر DC را بیان می‌کند.

قضیه ۳ برآوردگرهای مبتنی بر DC، $\tilde{\boldsymbol{\psi}}_{DC}$، برای $\boldsymbol{\psi}$ سازگار هستند.

[۱۹] Dirac Mass

برهان برای نمایش سازگاری برآوردگرهای مبتنی بر DC باید نشان داد وقتی $k \longrightarrow \infty$، آنگاه برای $\epsilon > 0$ دلخواه

$$P\left[|\tilde{\psi}_{DC} - \psi_0| > \epsilon\right] \longrightarrow 0, \qquad (۴)$$

که در آن ψ_0 مقدار واقعی پارامتر است. برای این منظور می‌توان نوشت

$$P\left[|\tilde{\psi}_{DC} - \psi_0| > \epsilon\right] \leq P\left[|\tilde{\psi}_{DC} - \hat{\psi}| > \epsilon/2\right] + P\left[|\hat{\psi} - \psi_0| > \epsilon/2\right]. \qquad (۵)$$

بر اساس قضیه ۲، یعنی همگرایی برآوردگرهای مبتنی بر DC به MLE، و سازگاری MLE، طرف راست رابطه (۵) به صفر میل می‌کند. بنابراین رابطه (۴) برقرار است.

۴ الگوریتم DC برای برازش SGLMMs

ایده روش DC برای برازش یک SGLMM، تولید نمونه‌هایی از توزیع توأم مبتنی بر DC پارامترها و k تحقق $\boldsymbol{u}_{(k)} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$ از میدان تصادفی گاوسی است. بنا بر قضیه ۲، اگر $k \longrightarrow \infty$، آنگاه چگالی حاشیه‌ای مبتنی بر DC پارامترها، یعنی $\pi^{(k)}(\psi|\mathbf{y})$، حول MLE پارامترها متمرکز خواهند شد. به عبارت دیگر میانگین توزیع حاشیه‌ای مبتنی بر DC پارامترهای مدل، پس از انتگرال‌گیری بر حسب $\boldsymbol{u}_{(k)}$، برابر MLE پارامترهای مدل و k برابر واریانس این توزیع، برابر واریانس مجانبی MLE هستند.

فرض کنید $\mathbf{y}^{(k)}$ بردار همسانه‌سازی شده داده‌ها و $c(\boldsymbol{h};\boldsymbol{\theta})$ تابع کوواریانس میدان تصادفی گاوسی باشد، که به دو پارامتر واریانس و دامنه $\boldsymbol{\theta} = (\sigma^2, \phi)$ وابسته است. به دلیل سادگی، از الگوریتم متروپولیس-هستینگز برای نمونه‌گیری از توزیع حاشیه‌ای مبتنی بر DC استفاده کرده و الگوریتم DC را به شرح زیر ارایه می‌کنیم:

گام اول : انتخاب مقادیر اولیه $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{u}^{(0)})$، قرار دادن $\boldsymbol{u}_{(k)}^{(0)}$ برابر k نسخه یکسان $\boldsymbol{u}^{(0)}$ و $\ell = 0$.

گام دوم : تولید مقدار پیشنهادی $\boldsymbol{\theta}^* = (\sigma^{*2}, \phi^*)$ در دو مرحله:

۱) تولید σ^{*2} از چگالی پیشنهادی[۲۰] $q(\sigma^{*2}|\sigma^2)$.

[۲۰] Proposal Density

۲) تولید ϕ^* از چگالی پیشنهادی $q(\phi^*|\phi)$.

گام سوم : تولید k مقدار $\boldsymbol{u}^*_{(k)} = (\boldsymbol{u}^*_1, \ldots, \boldsymbol{u}^*_k)$ از میدان تصادفی گاوسی با میانگین صفر و پارامترهای همبستگی $\boldsymbol{\theta}^*$.

گام چهارم : تولید مقدار پیشنهادی β^* از چگالی پیشنهادی $q(\beta^*|\beta)$.

گام پنجم : پذیرش مقادیر تولیدشده $\boldsymbol{\psi}^* = (\boldsymbol{\beta}^*, \sigma^{*۲}, \phi^*)$ با احتمال

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}^*) = \min\{۱, \frac{\{\prod_{m=۱}^{k} f(y|\boldsymbol{u}^*_m, \boldsymbol{\beta}^*)\}\pi(\boldsymbol{\psi}^*)q(\boldsymbol{\beta}|\boldsymbol{\beta}^*)q(\sigma^۲|\sigma^{*۲})q(\phi|\phi^*)}{\{\prod_{m=۱}^{k} f(y|\boldsymbol{u}_m, \boldsymbol{\beta})\}\pi(\boldsymbol{\psi})q(\boldsymbol{\beta}^*|\boldsymbol{\beta})q(\sigma^{*۲}|\sigma)q(\phi^*|\phi)}\},$$

و قرار دادن $\boldsymbol{\psi}^{(\ell+۱)} = \boldsymbol{\psi}^*$ و $\boldsymbol{u}^{(\ell+۱)}_{(k)} = \boldsymbol{u}^*_{(k)}$، یا رد $\boldsymbol{\psi}^*$ و قرار دادن $\boldsymbol{\psi}^{(\ell+۱)} = \boldsymbol{\psi}^{(\ell)}$ و $\boldsymbol{u}^{(\ell+۱)}_{(k)} = \boldsymbol{u}^{(\ell)}_{(k)}$.

گام ششم : افزایش یک واحدی ℓ و تکرار گام‌های دوم تا پنجم مادامی که زنجیر به توزیع مانای خود همگرا شود.

بر اساس نظریه زنجیرهای مارکوفی، انتظار می‌رود نمونه تولید شده توسط الگوریتم DC، بعد از حذف نمونه‌های دوره داغیدن[۲۱]، از توزیع حاشیه‌ای مبتنی بر DC پارامترها به دست آمده باشد. بنابراین وقتی هدف برازش مدل است، چون k تحقق $\boldsymbol{u}_j$، به ازای $j = ۱, \ldots, k$، در هر گام الگوریتم، نادیده گرفته می‌شوند، می‌توان دقت نمونه‌ها را با زیاد کردن طول نمونه افزایش داد.

۵ مثال‌ها

در این بخش، ابتدا کارایی الگوریتم پیشنهاد شده در بخش ۴ توسط یک مطالعه شبیه‌سازی مورد ارزیابی قرار می‌گیرد. سپس یک مثال واقعی شامل تعداد تصادفات رانندگی، تحلیل می‌شود.

[۲۱] Burn-in Period

۱.۵ مطالعه شبیه‌سازی

کارایی روش DC در برآورد SGLMMs، در قالب دو مثال شبیه‌سازی با پاسخ‌های دودویی و شمارشی، در این قسمت بررسی می‌شود. در مثال اول، یک مجموعه داده دودویی ناهمبسته به حجم $n = ۹۰۰$ بر روی یک مشبکه[۲۲] ۳۰×۳۰ از موقعیت‌های با فواصل یکسان، تولید شده است. این مثال، یک مدل ساده خطی تعمیم‌یافته را دربر می‌گیرد که MLE پارامترهای مدل با روش کمترین توان‌های دوم بازموزون تکراری[۲۳] (IRLS) به سادگی قابل محاسبه‌اند. از آنجا که تابع درستنمایی مدل به طور دقیق قابل محاسبه است، این مثال آزمونی برای سنجش کارایی الگوریتم DC محسوب می‌شود. در مثال دوم، دو مجموعه داده شمارشی همبسته فضایی بر روی دو مشبکه ۱۰×۱۰ و ۲۵×۲۵ تولید شده‌اند.

مثال ۱ (داده‌های دودویی ناهمبسته) داده‌های دودویی ناهمبسته بر روی یک مشبکه منظم ۳۰×۳۰ از مدل

$$\begin{aligned} f(y_i|\boldsymbol{\beta}) &= \left(\frac{exp(\eta_i)}{۱ + exp(\eta_i)}\right)^{y_i} \left(۱ - \frac{exp(\eta_i)}{۱ + exp(\eta_i)}\right)^{۱-y_i}, \\ \eta_i &= \ln(\frac{p_i}{۱ - p_i}) = \beta_۰ + \beta_۱ x_i, \ \ i = ۱, \ldots, n, \end{aligned}$$

تولید شدند، که در آن $p_i = E(Y_i|\boldsymbol{\beta}) = P(Y_i = ۱|\boldsymbol{\beta})$، $(\beta_۰, \beta_۱) = (-۲, ۰/۷۵)$ و برای $i = ۱, \ldots, ۹۰۰$، $x_i = i/۹۰۰$ در نظر گرفته شده‌اند.

برای داده‌های شبیه‌سازی شده، مقادیر دقیق MLE با روش IRLS محاسبه شده‌اند. برای رهایی از گیر افتادن در یک ماکسیمم موضعی و بررسی پایایی روش DC، از سه مجموعه توزیع پیشین متفاوت استفاده شده است. مجموعه اول شامل توزیع $U(-۴, ۴)$ برای هر دو پارامتر است، مجموعه دوم توزیع‌های $N(۰, ۱۰)$ و $N(-۹, ۹)$ را به ترتیب برای $\beta_۰$ و $\beta_۱$ در بر می‌گیرد و مجموعه سوم نیز شامل توزیع $N(۰, ۱۰)$ برای دو پارامتر است. همچنین $k = ۲۰۰$ لحاظ شده است. مقادیر MLE پارامترها نیز به عنوان مقادیر اولیه پارامترها در الگوریتم DC مورد

---

[۲۲] Grid

[۲۳] Iterativley Reweighted Least Squares

استفاده قرار گرفته‌اند. برای هر مجموعه پیشین، ۵۰۰۰ نمونه، بعد از حذف ۱۰۰۰ نمونه دوره داغیدن، از توزیع مبتنی بر DC تولید شده‌اند. برآورد پارامترها و خطای استاندارد آن‌ها در جدول ۱ گزارش شده‌اند. همان طور که ملاحظه می‌شود برای سه پیشین مختلف، برآوردها تا دو رقم اعشار یکسان و به MLE نزدیک هستند. خطای استاندارد آن‌ها نیز، با اندکی اختلاف که می‌تواند به دلیل استفاده از توزیع‌های پیشین متفاوت و خطای نمونه‌گیری به وجود آمده باشند، با هم یکی هستند.

جدول ۱: برآورد پارامترهای مدل خطی تعمیم‌یافته و خطای استاندارد متناظرشان برای داده‌های دودویی ناهمبسته با روش‌های IRLS و DC به ازای $k = ۲۰۰$

پارامتر	مقدار واقعی	MLE	$HDC_۱$	$HDC_۲$	$HDC_۳$
$\beta_۰$	-۲	-۱/۹۶ (۰/۱۸۹)	-۲/۰۹ (۰/۱۶۷)	-۲/۰۹ (۰/۱۸۴)	-۲/۰۹ (۰/۱۷۵)
$\beta_۱$	۰/۷۵	۰/۸۱ (۰/۳۰۷)	۰/۶۷ (۰/۲۶۶)	۰/۶۷ (۰/۲۹۳)	۰/۶۷ (۰/۲۸۰)

مثال ۲ (داده‌های شمارشی همبسته فضایی) دو مجموعه داده شمارشی فضایی بر روی دو مشبکه منظم کوچک و متوسط ۱۰ × ۱۰ و ۲۵ × ۲۵ از مدل پواسون به صورت

$$\begin{aligned} f(y_i|\beta) &= exp(y_i\eta_i - exp\{\eta_i\} - \ln(y_i!)), \\ \eta_i &= \ln(\mu_i) = \beta d_{۱i} + u_i, \quad i = ۱, \ldots, n, \end{aligned}$$

تولید شده‌اند، که در آن $d_{۱i}$ مولفه اول موقعیت مکانی iام، یعنی $\mathbf{s}_i = (d_{۱i}, d_{۲i})$، و $u_i = u(\mathbf{s}_i)$، به ازای $i = ۱, \ldots, n$، تحققی از میدان تصادفی گاوسی با میانگین صفر و تابع کوواریانس همسانگرد[۲۴] نمایی به فرم $C(\mathbf{s} - \mathbf{s}'; \boldsymbol{\theta}) = \sigma^۲ \exp(-\frac{\|\mathbf{s}-\mathbf{s}'\|}{\phi})$ است، که در آن $\boldsymbol{\theta} = (\sigma^۲, \phi)$ و $\|\mathbf{s} - \mathbf{s}'\|$ نشان‌دهنده فاصله اقلیدسی بین دو موقعیت فضایی $\mathbf{s}$ و $\mathbf{s}'$ است. برای تولید مقادیر شمارشی بزرگ، $(\beta, \sigma^۲, \phi) = (۰/۵, ۱/۲۵, ۳)$ انتخاب شده‌اند. برای رهایی از گیر کردن در نقاط ماکسیمم موضعی و بیان خاصیت پایایی روش DC نسبت به انتخاب توزیع‌های

[۲۴] Isotropic

پیشین، از سه مجموعه پیشین متفاوت استفاده شده است. توزیع‌های پیشین برای هر پارامتر مستقل از یکدیگر در نظر گرفته شده‌اند. مجموعه اول شامل $N(۰, ۹)$، $U(۰/۱, ۴)$ و $U(۰/۱, ۵)$ به ترتیب برای β، $\sigma^۲$ و ϕ است. مجموعه دوم، توزیع‌های $U(-۶, ۶)$، $LN(۰, ۳)$[۲۵]، $LN(۰, ۳)$ و مجموعه سوم، توزیع‌های $N(۰, ۲۰)$، $IG(۱, ۴)$[۲۶]، $G(۲, ۳)$[۲۷] را به ترتیب برای پارامترهای β، $\sigma^۲$ و ϕ شامل می‌شوند.

جدول ۲: برآوردهای مبتنی بر DC در مدل پواسونی برای داده‌های شمارشی فضایی با مجموعه توزیع‌های پیشین اول در دو طرح مشبکه‌ای ۱۰ × ۱۰ و ۲۵ × ۲۵

		مشبکه ۱۰ × ۱۰			مشبکه ۲۵ × ۲۵		
	پارامتر	β	$\sigma^۲$	ϕ	β	$\sigma^۲$	ϕ
k	مقدار واقعی	۰/۵	۱/۲۵	۳	۰/۵	۱/۲۵	۳
۱۰۰	Est.	۰/۳۹۷	۱/۵۷۵	۳/۷۱۰	۰/۴۳۲	۱/۵۵۵	۳/۷۳۵
	SE	۰/۰۲۷	۰/۱۹۲	۰/۱۹۱	۰/۰۰۲	۰/۱۸۰	۰/۱۸۰
۲۰۰	Est.	۰/۴۰۰	۱/۵۹۷	۳/۷۳۸	۰/۴۳۴	۱/۵۶۲	۳/۷۸۱
	SE	۰/۰۳۸	۰/۲۷۴	۰/۲۶۶	۰/۰۰۴	۰/۲۶۰	۰/۲۵۸
۴۰۰	Est.	۰/۳۹۹	۱/۵۵۹	۳/۶۹۴	۰/۴۲۹	۱/۵۷۹	۳/۷۹۳
	SE	۰/۰۵۲	۰/۳۸۹	۰/۳۸۹	۰/۰۰۵	۰/۳۷۵	۰/۳۳۲
۸۰۰	Est.	۰/۴۰۰	۱/۵۶۳	۳/۷۲۳	۰/۴۲۹	۱/۵۴۶	۳/۷۲۸
	SE	۰/۰۶۷	۰/۴۷۷	۰/۴۶۰	۰/۰۰۵	۰/۴۶۱	۰/۴۵۴

برای اجرای الگوریتم از یک نمونه‌گیر متروپولیس-هستینگز قدم زدن تصادفی[۲۸] با چگالی‌های نامزد $q(\beta^*|\beta) = N(\beta^{(t-۱)}, G)$، $q(\sigma^{*۲}|\sigma^۲) = IG(\sigma^{(t-۱)۲}, ۲)$ و $q(\phi^*|\phi) = LN(\phi^{(t-۱)}, ۲)$ استفاده شده است، و $\boldsymbol{\delta}^{(t-۱)}$ در گام جاری الگوریتم، مقادیر پارامترهای $\boldsymbol{\delta}$ در گام قبلی است. در چگالی نامزد $q(\beta^*|\beta)$، G واریانس مقیاس‌بندی شده یک مدل خطی تعمیم‌یافته ساده است، که پیشگوی خطی η_i به ازای $i = ۱, \ldots, n$، مقادیر $\frac{۱}{k}\sum_{j=۱}^{k} \boldsymbol{u}_i^{(j)}$ را نیز دارد. مقادیر اولیه $(\boldsymbol{\delta}^{(۰)}, \boldsymbol{u}^{(۰)})$ نیز بر اساس رهیافت پیشنهاد شده وارین و همکاران (۲۰۰۵) انتخاب شده‌اند.

[۲۵] Log Normal
[۲۶] Inverse Gamma
[۲۷] Gamma
[۲۸] Random Walk

جدول ۳: برآوردهای مبتنی بر DC در مدل پواسونی برای داده‌های شمارشی فضایی با مجموعه توزیع‌های پیشین دوم در دو طرح مشبکه‌ای ۱۰ × ۱۰ و ۲۵ × ۲۵

		مشبکه ۱۰ × ۱۰			مشبکه ۲۵ × ۲۵		
	پارامتر	β	σ^2	ϕ	β	σ^2	ϕ
k	مقدار واقعی	۰/۵	۱/۲۵	۳	۰/۵	۱/۲۵	۳
۱۰۰	Est.	۰/۳۹۲	۰/۶۵۲	۳/۸۰	۰/۴۳۰	۰/۶۸۸	۳/۷۶۵
	SE	۰/۰۲۸	۰/۲۳۵	۰/۱۷۷	۰/۰۰۳	۰/۲۲۱	۰/۱۹۰
۲۰۰	Est.	۰/۳۹۹	۰/۶۴۷	۳/۶۹۵	۰/۴۳۲	۰/۶۳۰	۳/۸۰۴
	SE	۰/۰۳۸	۰/۳۳۸	۰/۲۹۷	۰/۰۰۴	۰/۳۱۲	۰/۲۵۵
۴۰۰	Est.	۰/۳۹۱	۰/۶۵۴	۳/۷۸۲	۰/۴۳۲	۰/۶۵۱	۳/۷۸۴
	SE	۰/۰۵۶	۰/۴۸۲	۰/۳۷۷	۰/۰۰۵	۰/۴۴۰	۰/۳۷۳
۸۰۰	Est.	۰/۳۹۲	۰/۶۸۶	۳/۸۰۳	۰/۴۳۰	۰/۶۳۶	۳/۷۸۹
	SE	۰/۰۶۸	۰/۵۷۷	۰/۴۴۲	۰/۰۰۵	۰/۵۴۸	۰/۴۶۳

جدول ۴: برآوردهای مبتنی بر DC در مدل پواسونی برای داده‌های شمارشی فضایی با مجموعه توزیع‌های پیشین سوم در دو طرح مشبکه‌ای ۱۰ × ۱۰ و ۲۵ × ۲۵

		مشبکه ۱۰ × ۱۰			مشبکه ۲۵ × ۲۵		
	پارامتر	β	σ^2	ϕ	β	σ^2	ϕ
k	مقدار واقعی	۰/۵	۱/۲۵	۳	۰/۵	۱/۲۵	۳
۱۰۰	Est.	۰/۳۹۴	۰/۶۷۰	۳/۷۹۵	۰/۴۳۰	۰/۶۸۸	۳/۷۶۴
	SE	۰/۰۲۷	۰/۲۳۴	۰/۱۷۹	۰/۰۰۳	۰/۲۲۱	۰/۱۹۰
۲۰۰	Est.	۰/۳۹۱	۰/۶۷۲	۳/۷۴۹	۰/۴۳۲	۰/۶۲۹	۳/۸۰۳
	SE	۰/۰۴۱	۰/۳۳۳	۰/۲۷۰	۰/۰۰۴	۰/۳۲۲	۰/۲۵۶
۴۰۰	Est.	۰/۳۹۴	۰/۷۰۶	۳/۷۷۸	۰/۴۳۳	۰/۶۵۰	۳/۷۸۴
	SE	۰/۰۶۰	۰/۴۸۸	۰/۳۸۵	۰/۰۰۵	۰/۴۴۷	۰/۳۷۶
۸۰۰	Est.	۰/۴۰۰	۰/۶۴۲	۳/۸۱۶	۰/۴۲۵	۰/۶۳۹	۳/۷۸۸
	SE	۰/۰۶۷	۰/۵۷۸	۰/۴۱۵	۰/۰۰۵	۰/۵۴۰	۰/۴۵۳

جدول‌های ۲ تا ۴ نتایج شبیه‌سازی را به ترتیب برای سه مجموعه پیشین اول تا سوم نمایش می‌دهند. مقادیر جداول، آماره‌های نمونه‌ای محاسبه شده از نمونه‌های به حجم ۲۰۰۰ می‌باشند. نمونه‌ها از یک زنجیر به طول ۱۰۰۰۰ با دوره داغیدن ۱۰۰۰ و طول دوره انتخاب ۵ به دست آمده‌اند. در شبیه‌سازی‌ها از چهار مقدار مختلف $(۱۰۰, ۲۰۰, ۴۰۰, ۸۰۰)$ برای k استفاده شده است تا اثر آن نیز بر روی نتایج مورد بررسی قرار گیرد. در هر سه جدول، نتایج تقریباً یکسان هستند. بر اساس مقادیر برآورد شده حاصل از روش DC، می‌توان نکته‌های زیر را به عنوان نتیجه ارایه کرد:

الف) برآورد پارامترها در مجموعه داده با حجم متوسط، مشبکه ۲۵×۲۵، برای دو پارامتر ϕ و β تقریباً نزدیک به مقدار واقعی هستند. برآوردهای مشابهی برای مجموعه داده با حجم کم، ۱۰×۱۰، به دست آمده‌اند، با این تفاوت که برآوردهای حاصل در مجموعه داده با حجم متوسط دارای انحراف کمتری از مقادیر واقعی هستند. اما میزان انحراف از مقدار واقعی برای پارامتر $\sigma^۲$ بیش از حد انتظار است. علت آن را می‌توان به عدم وجود برآوردگر سازگار برای $\sigma^۲$ در مدل SGLMM (وارین و همکاران، ۲۰۰۵) مرتبط کرد. به علاوه مشهود است که برآورد پارامترها با افزایش k نوسان قابل ملاحظه‌ای ندارند.

ب) دقت برآوردگرهای مبتنی بر DC هم به حجم نمونه n و هم به k وابسته است. سرعت همگرایی خطای استاندارد برآوردها، SE، به مقدار برآورد بهینه، با افزایش k، برای مجموعه داده با حجم متوسط بیشتر از مجموعه داده با حجم نمونه کم است و خطاهای استاندارد با افزایش k روند صعودی خود را دنبال می‌کنند. بنابراین در مواردی که حجم داده‌ها کم و حتی متوسط است، خطاهای استاندارد گرایش به کم‌برآورد شدن دارند. در نتیجه به منظور کسب برآوردهای قابل اعتماد، باید k به اندازه کافی بزرگ اختیار شود. همانطور که در جداول ۲ تا ۴ ملاحظه می‌شود، در برخی موارد مقدار واقعی در فاصله اطمینان پارامتر قرار نمی‌گیرد، که این مساله به دلیل کم‌برآورد شدن خطاهای استاندارد و همچنین کم دقت بودن استنباط‌های مجانبی نوع والد

اتفاق می‌افتد. در این موارد برآورد خطاهای استاندارد برآوردها با استفاده از روش‌هایی مانند بوت‌استرپ[۲۹] توصیه می‌شود.

ج) نرخ پذیرش الگوریتم DC با افزایش حجم نمونه n افزایش می‌یابد ولی برای مقادیر مختلف k یکسان است. این نتیجه در جدول ۵ برای مجموعه پیشین اول در مدل شبیه‌سازی شده تشریح شده است. اعداد جدول، نرخ‌های پذیرش نمونه در الگوریتم DC برای چهار k مختلف و دو حجم نمونه ۱۰۰ و ۶۲۵ می‌باشند. نتایج برای سایر پیشین‌ها یکی است.

جدول ۵: نرخ پذیرش الگوریتم DC برای مجموعه پیشین اول در مدل پواسون برای داده‌های شمارشی فضایی

	k			
n	۱۰۰	۲۰۰	۴۰۰	۸۰۰
۱۰۰	۰/۲۸۳	۰/۲۷۲	۰/۲۷۵	۰/۲۷۶
۶۲۵	۰/۴۶۲	۰/۴۶۲	۰/۴۷۱	۰/۴۶۳

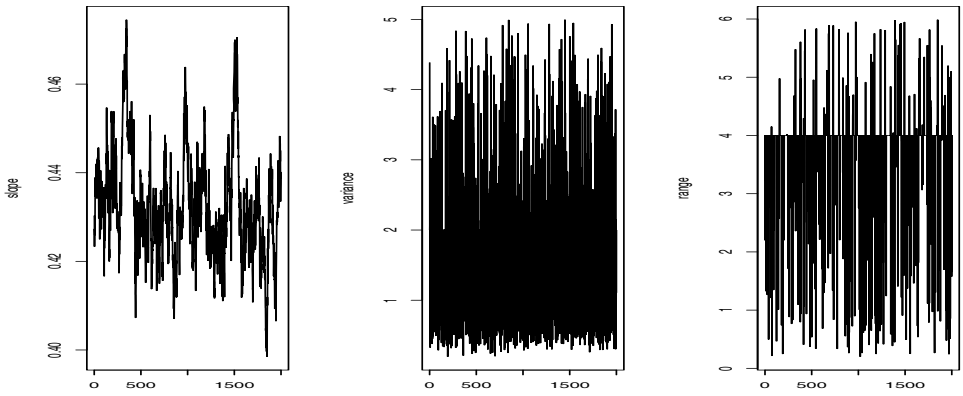
د) برای مطالعه رفتار آمیختگی الگوریتم DC، از نمودارهای اثری[۳۰] نمونه‌های تولید شده پارامترها استفاده شده است. شکل ۲ نمودارهای اثری را برای مدل شبیه‌سازی شده با مجموعه پیشین سوم، $k = ۴۰۰$ و $n = ۶۲۵$ نشان می‌دهد. نمودارها بیانگر آمیختگی خوب الگوریتم DC هستند.

۲.۵ *مثال کاربردی: تعداد تصادفات رانندگی*

ایران در فهرست کشورهای با نرخ تصادفات رانندگی بسیار بالا قرار دارد. دلایل مختلف محیطی و انسانی در بروز این مشکل سهیم هستند. تجربه رانندگان و اطلاع آن‌ها از قوانین ترافیکی، دو عامل انسانی و محل رخداد تصادف اعم از خیابان، بزرگراه و غیره، یک عامل محیطی تأثیرگذار محسوب می‌شوند.

[۲۹] Bootstrap

[۳۰] Trace Plots



شکل ۲: نمودارهای اثری نمونه‌های تولیدشده پارامترهای مدل پواسون برای داده‌های شمارشی فضایی با $k = ۴۰۰$ و $n = ۶۲۵$

داده‌های این مثال، تعداد تصادفات رانندگی در ۳۴۶ موقعیت مکانی شهر مشهد طی سال ۱۳۸۵ است. متغیر پاسخ یک متغیر دودویی اندازه‌گیری شده در موقعیت‌های مکانی است که نشان دهنده تصادف منجر به خرابی وسیله نقلیه (۰) یا جراحت (مرگ) (۱) سرنشینان می‌باشد. به علاوه برای هر موقعیت مکانی، دو متغیر تبیینی نوع گواهینامه راننده[۳۱] (DLT)، شامل دو سطح پایه ۱ و پایه ۲، و نوع مکان تصادف، شامل چهار سطح خیابان (Str)، چهارراه (Ints)، بزرگراه (Hw) و سایر مکان‌ها، ثبت شده‌اند. با توجه به آن که رانندگان دارای گواهینامه پایه ۱ از رانندگان دارای گواهینامه پایه ۲ باتجربه‌تر هستند و انواع تصادفات در مکان‌های مختلف وقوع تصادف، تفاوت دارند، در این مطالعه اثر این دو متغیر تبیینی بر انواع تصادف، مورد بررسی قرار می‌گیرد. برای در نظر گرفتن اثر مکان تصادف از متغیرهای نشانگر استفاده شده است.

شکل ۳ نمودار پراکنش داده‌ها را روی نقشه شهر مشهد نمایش می‌دهد. برای تحلیل داده‌ها از مدل رگرسیون لوژستیک فضایی با تابع پیوند لوجیت و ۵ پارامتر رگرسیونی با $\eta_i = \beta_0 + \beta_1 DLT_i + \beta_2 Str_i + \beta_3 Ints_i + \beta_4 Hw_i + u_i$ استفاده شده است، که در آن $u_i; i = ۱, \ldots, ۳۴۶$ تحققی از یک میدان تصادفی گاوسی با

[۳۱] Driver's License Type

شکل ۳: نمودار پراکنش داده‌های تصادفات رانندگی در شهر مشهد، دایره‌های روشن بیانگر تصادفات خسارتی و ستاره‌های سیاه بیانگر تصادفات جرحی یا فوتی هستند.

میانگین صفر و تابع کوواریانس مانای $C(\boldsymbol{h};\boldsymbol{\theta})$ است. مقادیر اولیه برای پارامترهای رگرسیونی، پس از برازش یک مدل با اثرات ثابت، عبارتند از $\beta_0^{(0)} = -۳/۸۵۹$، $\beta_1^{(0)} = ۱/۰۵۸$، $\beta_2^{(0)} = ۰/۷۸۱$، $\beta_3^{(0)} = ۰/۳۸۷$ و $\beta_4^{(0)} = ۰/۴۵۳$. همچنین با محاسبه هم‌تغییرنگار تجربی مانده‌های تبدیل‌یافته حاصل از مدل، فرم تابع کوواریانس برازش شده مناسب برای آن، یک تابع کوواریانس نمایی است. مقادیر اولیه پارامترهای وابستگی فضایی نیز عبارتند از $\sigma^{2(0)} = ۰/۶۴۹$ و $\phi^{(0)} = ۰/۲۶۶$ کیلومتر. از پیشین‌های $N(۰, ۹)$ برای پارامترهای رگرسیونی و $U(۰/۱, ۴)$ برای پارامترهای وابستگی فضایی، استفاده شده‌اند. جدول ۶ نتایج حاصل از اجرای الگوریتم DC بر روی داده‌های تصادفات را نشان می‌دهد که از یک نمونه به حجم ۳۰۰۰ با دوره داغیدن ۲۰۰۰ و طول دوره انتخاب ۱۰ به دست آمده‌اند. مقدار k نیز برابر ۴۰۰ اختیار شده است.

تجربه راننده شامل مهارت در رانندگی و آگاهی از قوانین رانندگی است. همان طور که ملاحظه می‌شود، تأثیر نوع گواهینامه رانندگی که بیانگر تجربه رانندگان است، بر وقوع و نوع تصادف به وضوح غیرقابل انکار است. ضریب مثبت و

معنی‌دار تجربه رانندگان، β_1، نشان می‌دهد که رانندگان کم‌تجربه بیشتر درگیر تصادفات منجر به جرح یا فوت هستند. همچنین با توجه به معنی‌داری بالای پارامترهای β_2، β_3 و β_4، می‌توان گفت تصادفاتی که در خیابان‌ها یا بزرگراه‌ها روی می‌دهند، با احتمال بیشتری منجر به جرح یا فوت می‌شوند. معنی‌داری دو پارامتر وابستگی فضایی، σ^2 و ϕ، بیانگر وجود وابستگی نتیجه تصادف در موقعیت‌های مکانی نزدیک به هم است. همچنین در نظر گرفتن وابستگی فضایی در تحلیل این داده‌ها، کارایی نتایج به دست آمده را افزایش می‌دهد.

جدول ۶: استنباط مبتنی بر DC برای مدل‌بندی فضایی داده‌های تصادفات رانندگی مشهد

پارامتر	برآورد	خطای معیار	آماره t	p-مقدار
β_0	-۵/۵۵۲	۰/۵۵۴	-۰/۰۲۲	۰/۴۹۱
β_1	۱/۵۳۴	۰/۲۶۵	۵/۷۸۷	۰/۰۰۰
β_2	۱/۳۳۹	۰/۲۴۸	۵/۳۹۹	۰/۰۰۰
β_3	۰/۷۷۵	۰/۲۷۰	۲/۸۷۰	۰/۰۰۲
β_4	۰/۹۹۹	۰/۳۲۳	۳/۰۹۳	۰/۰۰۱
σ^2	۱/۴۰۴	۰/۵۰۸	۲/۷۶۴	۰/۰۰۳
ϕ	۰/۷۹۰	۰/۲۹۱	۲/۷۱۵	۰/۰۰۴

بحث و نتیجه‌گیری

اگرچه DC روشی مناسب برای استنباط آماری در رده SGLMMs است، اما از محدودیت‌ها و معایبی رنج می‌برد که در باغیشنی و محمدزاده (۲۰۱۱) به آن‌ها اشاره شده است. روش DC وجوه مشترکی با برخی روش‌های پیشنهادی سایر محققین نیز دارد. روش‌های پس‌خوران پیشین[۳۲] (رابرت، ۱۹۹۳)، نوردیدن شبیه‌سازی‌شده[۳۳] (بروکس و مورگان، ۱۹۹۵؛ گیر و تامپسون، ۱۹۹۵)، و الگوریتم SAME[۳۴] (دوچت و همکاران، ۲۰۰۲؛ گایتون و یائو، ۲۰۰۳)، از جمله آن‌ها

[۳۲] Prior Feedback

[۳۳] Simulated Annealing

[۳۴] State-Augmentation for Marginal Estimation

محسوب می‌شوند.

معمولاً پیچیدگی اجرای روش‌های بسامدی برای SGLMMs، منجر به استفاده از رهیافت بیزی می‌شود. اما روش‌های استنباط بسامدی مانند محاسبه MLE، فاصله‌های اطمینان و آزمون‌های فرضیه در مدل‌های فضایی پیچیده، به کمک روش DC قابل انجام است. بنابراین می‌توان چنین نتیجه گرفت که ملاک انتخاب رهیافت‌های بیزی و بسامدی برای تحلیل این رده از مدل‌ها، مانند گذشته، امکان انجام آن‌ها نیست، بلکه دیدگاه‌های فلسفی محقق تعیین‌کننده انتخاب یکی از دو رهیافت خواهد بود.

مراجع

Baghishani, H. and Mohammadzadeh, M. (2011), A Data Cloning Algorithm for Computing Maximum Likelihood Estimates in Spatial Generalized Linear Mixed Models, *Computational Statistics and Data Analysis*, **55**, 1748-1759.

Booth, J. G. and Hobert, J. P. (1999), Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm, *Journal of the Royal Statistical Society, Series B*, **61**, 265-285.

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

Brooks, S. P. and Morgan, B. J. T. (1995), Optimization Using Simulated Annealing, *Statistician*, **44**, 241-257.

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

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

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

Diggle, P., Tawn, J. A. and Moyeed, R. A. (1998), Model-Based Geostatistic, *Applied Statistics*, **47**, 299-350.

Doucet, A., Godsill, S. J. and Robert, C. P. (2002), Marginal Maximum A Posteriori Estimation Using Markov Chain Monte Carlo, *Statistics and Computing*, **12**, 77-84.

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

Gaetan, C. and Yao, J. F. (2003), A Multiple-Imputation Metropolis Version of the EM Algorithm, *Biometrika*, **90**, 643-654.

Geyer, C. J. and Thompson, E. A. (1995), Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference, *Journal of the American Statistical Association*, **90**, 909-920.

Geyer, C. J. and Thompson, E. A. (1992), Constrained Monte Carlo Maximum Likelihood for Dependent Data (with Discussion), *Journal of the Royal Statistical Society, Series B*, **54**, 657-699.

Hosseini, F., Eidsvik, J. and Mohammadzadeh, M. (2011), Approximate Bayesian Inference in Spatial Generalized Linear Mixed Models with

Skew Normal Latent Variables, *Computational Statistics and Data Analysis*, **55**, 1791-1806.

Lele, S. R., Dennis, B. and Lutscher, F. (2007), Data Cloning: Easy Maximum Likelihood Estimation for Complex Ecological Models Using Bayesian Markov Chain Monte Carlo Methods, *Ecology Letters*, **10**, 551-563.

Lesaffre, E. and Speissens, B. (2001), On the Effect of the Number of Quadrature Points in a Logistic Random-Effects Model: An Example, *Applied Statistics*, **50**, 325-335.

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

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Second edition, Chapman and Hall/CRC, London.

Pinheiro, J. C. and Bates, D. M. (1995), Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model, *Journal of Computational and Graphical Statistics*, **4**, 12-35.

Robert, C. P. (1993), Prior Feedback: Bayesian Tools for Maximum Likelihood Estimation, *Journal of Computational Statistics*, **8**, 279-294.

Varin, C., Gudmund, H. and Skare, O. (2005), Pairwise Likelihood Inference in Spatial Generalized Linear Mixed Models, *Computational Statistics and Data Analysis*, **49**, 1173-1191.

Walker, A. M. (1969), On the Asymptotic Behaviour of Posterior Distributions, *Journal of the Royal Statistical Society, Series B*, **31**, 80-88.

Zhao, Y., Staudenmayer, J., Coull, B. A. and Wand, M. P. (2006), General Design Bayesian Generalized Linear Mixed Models with Applications to Spatial Statistics, *Statistical Science*, **21**, 35-51.

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱
مجموعه مقالات، ص ۵۹-۸۲

مدل‌های آمیخته خطی تعمیم‌یافته فضایی با متغیرهای پنهان چوله نرمال بسته

فاطمه حسینی[۱]، **محسن محمدزاده**[۲]

[۱] *دانشگاه سمنان، گروه آمار*

[۲] *دانشگاه تربیت مدرس، گروه آمار*

چکیده : مدل‌های آمیخته خطی تعمیم‌یافته فضایی برای مدل‌بندی پاسخ‌های فضایی گسسته به‌کار می‌روند، که در آن‌ها ساختار همبستگی فضایی داده‌ها از طریق متغیرهای پنهان با توزیع نرمال در نظر گرفته می‌شود. هر چند فرض نرمال بودن توزیع متغیرهای پنهان موجب سهولت محاسبات می‌شود، اما در عمل به‌دلیل غیرقابل مشاهده بودن متغیرهای پنهان، بررسی نرمال بودن این متغیرها مقدور نیست و پذیرش ناصحیح این فرض می‌تواند روی دقت برآورد پارامترها و پیشگوها تأثیر سوء داشته باشد. در این مقاله استفاده از خانواده توزیع چوله‌نرمال بسته که شامل خانوادهٔ نرمال و چوله‌نرمال است، برای توزیع متغیرهای پنهان پیشنهاد شده و پیشگویی فضایی بیزی متغیرهای پنهان چوله ارائه و در مطالعه‌ای شبیه‌سازی مورد ارزیابی قرار گرفته است. طولانی بودن زمان محاسبه برآوردها و پیشگویی‌های بیزی

آدرس الکترونیک مسئول مقاله: فاطمه حسینی، fatemeh.hoseini@profs.semnan.ac.ir

کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲F۱۱

انگیزه‌ای برای معرفی روش بیز تقریبی گردید که از سرعت بالاتری برخوردار است. این روش در یک مجموعه داده واقعی به‌کار گرفته شده و دقت پیشگویی مدل با دو فرض نرمال و چوله‌نرمال بودن متغیرهای پنهان بررسی و عملکرد دو روش بیز تقریبی و معمولی مورد ارزیابی و مقایسه قرار گرفته است.

واژه‌های کلیدی : توزیع چوله‌نرمال بسته، مدل آمیخته خطی تعمیم‌یافته، متغیر پنهان، استنباط بیز تقریبی.

۱ مقدمه

مک‌کلاح و نلدر (۱۹۸۹) مدل‌های خطی تعمیم‌یافته[۱](GLM) را برای مدل‌بندی متغیرهای پاسخ گسسته پیشنهاد دادند. در این مدل‌ها با فرض استقلال مشاهدات، با استفاده از یک تابع پیوند بین میانگین مشاهدات و متغیرهای کمکی ارتباط خطی برقرار می‌شود. در حالتی که بین مشاهدات همبستگی وجود دارد، تعمیمی از مدل‌های مذکور به‌نام مدل‌های آمیخته خطی تعمیم‌یافته[۲](GLMM) استفاده می‌شود. در این مدل‌ها 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. اگر همبستگی از نوع فضایی باشد، معمولاً مدل آمیخته خطی تعمیم‌یافته فضایی[۳](SGLMM) به‌کار گرفته می‌شود. یک مسئله مهم در مدل‌های SGLM پیشگویی متغیرهای پنهان در موقعیت‌های فاقد مشاهده می‌باشد، که مستلزم برآورد پارامترهای مدل و متغیرهای پنهان در موقعیت‌های دارای مشاهده پاسخ می‌باشد. چون در مدل‌های آمیخته خطی تعمیم‌یافته، تابع درستنمایی برخلاف مدل‌های خطی به‌دلیل ناگاوسی بودن متغیر پاسخ و وجود متغیرهای پنهان فرم بسته‌ای ندارد، برآورد پارامترها به‌راحتی امکان‌پذیر نیست. زانگ (۲۰۰۲)، ورین و همکاران (۲۰۰۵)، زو و پترسون (۲۰۰۷) و باغیشنی و محمدزاده (۲۰۱۱) با در نظر گرفتن

---

۱ Generalized Linear Models

۲ Generalized Linear Mixed Models

۳ Spatial Generalized Linear Mixed Model

توزیع نرمال برای متغیرهای پنهان به مطالعه مدل‌های SGLM با رهیافت ماکسیمم درستنمایی پرداخته‌اند.

در تحقیقاتی مانند دیگل و همکاران (۱۹۹۸)، کریستنسن و همکاران (۲۰۰۰)، کریستن و وگپترسون (۲۰۰۲)، کریستن (۲۰۰۴) و کریستنسن و همکاران (۲۰۰۶) مدل‌های SGLM با متغیرهای پنهان فضایی نرمال با رهیافت بیزی و الگوریتم‌های MCMC مورد بررسی قرار گرفته است. در اکثر مطالعات پیرامون مدل‌های SGLM، با فرض نرمال بودن متغیرهای پنهان اقدام به حل مسئله شده است. اما در عمل به‌دلیل غیر قابل مشاهده بودن متغیرهای پنهان در مدل‌های SGLM، بررسی نرمال بودن آن‌ها مقدور نیست و پذیرش ناصحیح این فرض می‌تواند بر روی دقت برآورد پارمترها و پیشگوها تأثیر سوء داشته باشد.

کیم و مالیک (۲۰۰۴) تحلیل داده‌های فضایی پیوسته را برای یک میدان تصادفی چوله گاوسی[۴](SGR) مورد بررسی قرار دادند و پیشگویی فضایی بیزی را در یک مثال کاربردی به‌کار گرفتند. کریمی و محمدزاده (۲۰۰۷، ۲۰۰۹، ۲۰۱۰) و کریمی و همکاران (۲۰۱۰) تحلیل داده‌های فضایی را برای یک میدان تصادفی چوله گاوسی بسته[۵](CSG) در مسائل کاربردی مورد مطالعه قرار دادند.

با توجه به این که خانواده توزیع‌های چوله‌نرمال[۶]، (آزالینی، ۱۹۸۵) و توزیع چوله‌نرمال بسته[۷](CSN)، (دامینگوس و همکاران، ۲۰۰۳) از خانواده توزیع‌های نرمال بزرگتر و از انعطاف‌پذیری بیشتری برخوردارند و شامل این خانواده می‌باشند، انتظار می‌رود پذیرش فرض چوله‌نرمال بودن توزیع متغیرهای پنهان در مدل‌های SGLM به واقعیت نزدیک‌تر از پذیرش فرض نرمال بودن آن‌ها باشد. حسینی و همکاران (۲۰۰۹)، محمدزاده و حسینی (۲۰۱۱)، حسینی و همکاران (۲۰۱۱)، کریمی و همکاران (۲۰۱۱) و حسینی و محمدزاده (۲۰۱۲) برای اجتناب از پذیرش فرض نرمال بودن، استفاده از خانواده توزیع چوله‌نرمال بسته برای مدل‌بندی متغیرهای پنهان فضایی را پیشنهاد کردند و با در نظر گرفتن توزیع چوله‌نرمال بسته

---

۴ Skew Gaussian Random Field

۵ Closed Skew Gaussian Random Field

۶ Skew Normal

۷ Closed Skew Normal

برای متغیرهای پنهان، به تحلیل این مدل‌ها پرداخته‌اند. در این مقاله ابتدا تعریفی از مدل‌های SGLM با متغیرهای پنهان چوله‌نرمال بسته و نحوه پیشگویی متغیرهای پنهان ارائه شده است. سپس با رهیافت بیزی به تحلیل این مدل‌ها پرداخته شده است، که به‌دلیل طولانی بودن زمان محاسبات تحلیل بیز تقریبی مدل‌ها ارائه شده است.

۲ مدل‌های SGLM با متغیرهای پنهان چوله‌نرمال بسته

فرض کنید $\boldsymbol{x} = (x_1, \cdots, x_n)'$ بردار متغیرهای پنهان فضایی در n موقعیت $\{s_1, \cdots, s_n\}$ دارای توزیع چوله‌نرمال بسته به فرم $CSN_{n,q}(Z\boldsymbol{\beta}, \Sigma_\theta, D, \boldsymbol{\nu}, \Delta)$ باشد، که در آن $\boldsymbol{\eta} = (\boldsymbol{\beta}', \boldsymbol{\theta}, D, \boldsymbol{\nu}, \Delta)'$ پارامترهای مدل، Z ماتریس $n \times (p+1)$ متغیرهای کمکی، $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)'$ بردار پارامترهای رگرسیونی، D ماتریس $q \times n$ پارامترهای چولگی و $\boldsymbol{\theta}$ بردار پارامترهای همبستگی فضایی مدل باشند. بنابراین چگالی متغیرهای پنهان به‌صورت

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\eta}) &= \frac{k}{(2\pi)^{\frac{n}{2}}|\Sigma_\theta|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(\boldsymbol{x} - Z\boldsymbol{\beta})'\Sigma_\theta^{-1}(\boldsymbol{x} - Z\boldsymbol{\beta})\right\} \\ &\times \quad \Phi_q(D(\boldsymbol{x} - Z\boldsymbol{\beta}); \boldsymbol{\nu}, \Delta), \end{aligned} \tag{1}$$

است، که در آن $k = \Phi_q^{-1}(\mathbf{0}; \boldsymbol{\nu}, \Delta + D\Sigma_\theta D')$. اکنون فرض کنید مشاهدات در موقعیت‌های فضایی $\{s_1, \cdots, s_k\}$ در اختیار باشند و هدف پیشگویی متغیرهای پنهان در موقعیت‌های فاقد مشاهده $\{s_{k+1}, \cdots, s_n\}$ باشد. متغیرهای پنهان در k موقعیت مشاهده شده را به‌صورت $\boldsymbol{x}^{obs} = A\boldsymbol{x}$ نمایش می‌دهیم، که در آن $A = [I_{k\times k}|\mathbf{0}_{k\times n-k}]$ است. در این‌صورت بردار $\boldsymbol{x}$ را می‌توان به‌صورت $(\boldsymbol{x}^{obs'}, \boldsymbol{x}^{pred'})$ تجزیه کرد، که در آن $\boldsymbol{x}^{pred}$ بردار متغیرهای پنهان در $n-k$ موقعیت انتخاب شده برای پیشگویی است. فرض کنید $\boldsymbol{y} = (y_1, \cdots, y_k)$ بردار متغیرهای پاسخ فضایی گسسته در k موقعیت $\{s_1, \cdots, s_k\}$ باشد. طبق مک‌کلاح و نلدر (۱۹۸۹) فرض می‌شود که با فرض استقلال شرطی این متغیرها روی متغیرهای پنهان، $f(\boldsymbol{y}|\boldsymbol{x})$ متعلق به خانواده نمایی با تابع چگالی

$$f(y_i|x_i) = exp\{y_i x_i - b(x_i) + c(y_i)\}, \quad i = 1, \ldots, k, \tag{2}$$

است، که در آن $b(\cdot)$ و $c(\cdot)$ توابعی معلوم هستند. میانگین شرطی $\mathrm{E}(y_i|x_i)$ و x_i با یک تابع پیوند معلوم g، به‌صورت $\mathrm{E}(y_i|x_i) = g^{-1}(x_i)$ در ارتباط می‌باشند.
بنابراین مولفه‌های مدل SGLM به‌صورت

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\eta}) =& f(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\eta}) \\ \propto& |\Sigma_\theta|^{-1/2} \Phi_q(D(\boldsymbol{x} - Z\boldsymbol{\beta}); \boldsymbol{\nu}, \Delta) \times \Phi_q^{-1}(\mathbf{o}; \boldsymbol{\nu}, \Delta + D\Sigma_\theta D') \\ &\times exp\{\sum_{i=1}^{k}[y_i x_i - b(x_i) + c(y_i)] - \frac{1}{2}(\boldsymbol{x} - Z\boldsymbol{\beta})'\Sigma_\theta^{-1}(\boldsymbol{x} - Z\boldsymbol{\beta})\}. \end{aligned}$$

خلاصه می‌شوند.

۱.۲ پیشگویی در مدل‌های SGLM با متغیرهای پنهان‌چوله نرمال بسته

یک مسئله مهم در این مدل‌ها پیشگویی متغیرهای پنهان می‌باشد. حسینی و محمدزاده (۲۰۱۱) با فرض آن که متغیرهای پنهان مدل SGLM دارای توزیع چوله‌نرمال بسته و پارامترهای مدل معلوم هستند، پیشگویی به روش مینیمم متوسط توان‌های دوم خطا[۸] (MMSE) برای این متغیرها را به‌دست آوردند.

قضیه ۱ فرض کنید توزیع بردار متغیرهای پنهان فضایی متعلق به خانواده چوله‌نرمال بسته به فرم (۱) باشد و با شرط روی متغیرهای پنهان فضایی، توزیع متغیرهای مستقل شرطی پاسخ $\{y_i,\ i = 1, \ldots, k\}$، متعلق به خانواده نمایی به فرم (۲) باشد. در این‌صورت

$$E(\boldsymbol{x}^{pred}|\boldsymbol{y}) = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(E(\boldsymbol{x}^{obs}|\boldsymbol{y}) - \boldsymbol{\mu}_1) + \Sigma_{22.1}D_2' E(\psi|\boldsymbol{y}) \qquad (۳)$$

پیشگوی MMSE متغیرهای پنهان در $n-k$ موقعیت $(s_{k+1}, \ldots, s_n)$ است.

برهان با در نظر گرفتن $\boldsymbol{x}' = (\boldsymbol{x}^{obs\prime}, \boldsymbol{x}^{pred\prime})$، که در آن $\boldsymbol{x}^{obs} = (x_1, \ldots, x_k)'$ و $\boldsymbol{x}^{pred} = (x_{k+1}, \ldots, x_n)'$ و با تفکیک‌بندی

$$Z\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix},\ \Sigma_\theta = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},\ D = [D_1\ \ D_2],$$

[۸] Minimum Mean squared Error

و بنا بر خاصیت بسته بودن توزیع چوله‌نرمال بسته نسبت به شرطی کردن و دامینگوس و همکاران (۲۰۰۳، قضیه ۱۶) داریم

$$\begin{aligned}\boldsymbol{x}^{pred}|\boldsymbol{x}^{obs} \sim CSN_{1,q}(&\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}^{obs} - \boldsymbol{\mu}_1), \Sigma_{22.1},\\ &D_2, \boldsymbol{\nu} - D^*(\boldsymbol{x}^{obs} - \boldsymbol{\mu}_1), \Delta),\end{aligned}$$

که در آن $D^* = D_1 + D_2\Sigma_{21}\Sigma_{11}^{-1}$ و $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. بنابراین از تعریف امید ریاضی توزیع چوله‌نرمال بسته می‌توان نوشت

$$\mathrm{E}(\boldsymbol{x}^{pred}|\boldsymbol{x}^{obs}) = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}^{obs} - \boldsymbol{\mu}_1) + \Sigma_{22.1}D_2'\psi$$

که در آن $\psi = \frac{\Phi_q^*(r;\boldsymbol{\nu}-D^*(\boldsymbol{x}^{obs}-\boldsymbol{\mu}_1),\Delta+D_2\Sigma_{22.1}D_2')}{\Phi_q(r;\boldsymbol{\nu}-D^*(\boldsymbol{x}^{obs}-\boldsymbol{\mu}_1),\Delta+D_2\Sigma_{22.1}D_2')}|_{r=0}$. چون متغیرهای پاسخ $\boldsymbol{y} = \{y_i;\ i = 1,\ldots,k\}$ فقط وابسته به بردار $\boldsymbol{x}^{obs}$ هستند، داریم

$$f(\boldsymbol{x}^{obs}, \boldsymbol{x}^{pred}, \boldsymbol{y}) = \quad = \quad f(\boldsymbol{y}, \boldsymbol{x}^{obs})f(\boldsymbol{x}^{pred}|\boldsymbol{x}^{obs})$$

از طرفی $f(\boldsymbol{x}^{pred}, \boldsymbol{x}^{obs}, \boldsymbol{y}) = f(\boldsymbol{x}^{pred}|\boldsymbol{x}^{obs}, \boldsymbol{y})f(\boldsymbol{x}^{obs}, \boldsymbol{y})$، بنابراین

$$\mathrm{E}(\boldsymbol{x}^{pred}|\boldsymbol{x}^{obs}) = \mathrm{E}(\boldsymbol{x}^{pred}|\boldsymbol{x}^{obs}, \boldsymbol{y})$$

و می‌دانیم $\mathrm{E}(\boldsymbol{x}^{pred}|\boldsymbol{y}) = \mathrm{E}(\mathrm{E}(\boldsymbol{x}^{pred}|\boldsymbol{x}^{obs}, \boldsymbol{y})|\boldsymbol{y})$. بنابراین

$$\begin{aligned}E(\boldsymbol{x}^{pred}|\boldsymbol{y}) &= E(E(\boldsymbol{x}^{pred}|\boldsymbol{x}^{obs})|\boldsymbol{y})\\ &= \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(E(\boldsymbol{x}^{obs}|\boldsymbol{y}) - \boldsymbol{\mu}_1) + \Sigma_{22.1}D_2'E(\psi|\boldsymbol{y}).\end{aligned}$$ ■

اما توزیع $(\boldsymbol{x}^{obs}|\boldsymbol{y})$ دارای فرم بسته‌ای نیست و می‌توان نمونه‌های مونت کارلوی $\boldsymbol{x}^{obs(1)}, \ldots, \boldsymbol{x}^{obs(K)}$ را از توزیع شرطی $(\boldsymbol{y})$ تولید $f_{x^{obs}|y}(\cdot|\boldsymbol{y})$ و با الگوریتم متروپولیس-هستینگز $\mathrm{E}(\boldsymbol{x}^{obs}|\boldsymbol{y}) \approx \frac{1}{K}\sum_{m=1}^{K}\boldsymbol{x}^{obs(m)}$ و $\mathrm{E}(\boldsymbol{\psi}|\boldsymbol{y}) \approx \frac{1}{K}\sum_{m=1}^{K}\boldsymbol{\psi}^{(m)}$ حاصل می‌شوند که در آن $\boldsymbol{\psi}^{(m)}$ مقدار $\boldsymbol{\psi}$ متناظر با $\boldsymbol{x}^{obs}$ است. با اختیار تابع $h(\boldsymbol{x}^{obs}) = f(\boldsymbol{x}^{obs}|\boldsymbol{\eta})$ به‌عنوان توزیع نامزد، احتمال پذیرش یک مقدار جدید $\boldsymbol{x}^{obs*}$ به شرط مقدار قبلی $\boldsymbol{x}^{obs}$ به‌صورت $\min\left\{\prod_{i=1}^{k}\frac{f(y_i|x_i^*)}{f(y_i|x_i)}, 1\right\}$

خلاصه می‌شود. باید توجه کرد که $\boldsymbol{x}^{obs} = (x_1, \ldots, x_k)'$ مولفه‌های چوله‌نرمال بسته وابسته‌اند. لذا برای تولید یک مقدار جدید برای ℓامین مولفه، به طوری که مولفه‌های دیگر تغییر نکنند، باید از توزیع شرطی x_ℓ به شرط $\boldsymbol{x}_{-\ell}^{obs}$ استفاده کرد، که در آن $\boldsymbol{x}_{-\ell}^{obs} = \{x_1, \ldots, x_\ell, x_{\ell+1}, \ldots, x_k\}$ می‌باشد. برای به‌دست آوردن توزیع شرطی $x_\ell|\boldsymbol{x}_{-\ell}^{obs}$، می‌دانیم $\boldsymbol{x}^{obs} = A\boldsymbol{x}$، که در آن $A = [I_{k\times k}|\mathbf{0}_{k\times n-k}]$. اکنون بنابر خاصیت بسته بودن توزیع CSN تحت کناری‌سازی که در آن $\Delta^* = \Delta + D_2\Sigma_{22.1}D_2'$ می‌باشد. برای هر $1 \le \ell \le k$ داریم

$$\begin{pmatrix} x_\ell \\ \boldsymbol{x}_{-\ell}^{obs} \end{pmatrix} = A_\ell \boldsymbol{x}^{obs} \sim CSN_{k,q}(A_\ell\boldsymbol{\mu}_1, A_\ell\Sigma_{11}A_\ell', D^*A_\ell', \boldsymbol{\nu}, \Delta^*), \quad \ell = 1, \cdots, k,$$

که در آن A_ℓ یک ماتریس همانی است که ℓامین سطر آن به سطر اول منتقل شده است، بنابراین $A_\ell A_\ell' = I$. اکنون با در نظر گرفتن

$$A_\ell\boldsymbol{\mu}_1 = \begin{pmatrix} \mu_{1.1} \\ \mu_{1.2} \end{pmatrix}, \; A_\ell\Sigma_{11}A_\ell' = \begin{pmatrix} \Sigma_{1.11} & \Sigma_{1.12} \\ \Sigma_{1.21} & \Sigma_{1.22} \end{pmatrix}, \; D^*A_\ell' = [D_{1.1} \;\; D_{1.2}],$$

و با توجه به خاصیت کناری‌سازی توزیع CSN، توزیع $x_\ell|\boldsymbol{x}_{-\ell}^{obs}$ به‌صورت

$$\begin{aligned} CSN_{1,q}(&\mu_{1.1} + \Sigma_{1.12}\Sigma_{1.22}^{-1}(\boldsymbol{x}_{-\ell}^{obs} - \boldsymbol{\mu}_{1.2}), \Sigma_{11.2}, \\ &D_{1.1}, \boldsymbol{\nu} - D^\circ(\boldsymbol{x}_{-\ell}^{obs} - \boldsymbol{\mu}_{1.2}), \Delta), \end{aligned} \qquad (۴)$$

می‌شود، که در آن $\Sigma_{11.2} = \Sigma_{1.11} - \Sigma_{1.12}\Sigma_{1.22}^{-1}\Sigma_{1.21}$ و $D^\circ = D_{1.2} + D_{1.1}\Sigma_{1.12}\Sigma_{1.22}^{-1}$. برای تولید نمونه تصادفی از توزیع الگوریتم متروپولیس–هستینگز تک مولفه‌ای با گام‌های زیر $f(\boldsymbol{x}^{obs}|\boldsymbol{y}, \boldsymbol{\eta})$ اجرا می‌شود:

گام ۱– مقادیر اولیه $\boldsymbol{x}^{obs} = (0, \ldots, 0)$ و $m = 0$ اختیار شوند.

گام ۲– مقدار x_ℓ^* از توزیع چوله‌نرمال بسته (۴) و مقدار U از توزیع یکنواخت $(0, 1)$ تولید شوند. اگر رابطه $U < \min\left\{\frac{f(y_\ell|x_\ell^*)}{f(y_\ell|x_\ell)}, 1\right\}$ برقرار باشد، $\boldsymbol{x}^{obs^{(m)}} = (x_1, \ldots, x_{\ell-1}, x_\ell^*, x_{\ell+1}, \ldots, x_k)$، در غیر این‌صورت $\boldsymbol{x}^{obs^{(m)}}$ بدون تغییر بماند. این مرحله تا تکمیل همه مولفه‌های بردار $\boldsymbol{x}^{obs^{(m)}}$ تکرار شود.

گام ۳– قرار دهید $m = m + 1$ و مرحله گام (۲) تکرار شود.

۳ پیشگویی بیزی در مدل‌های SGLM با متغیرهای پنهان چوله‌نرمال بسته

در عمل اغلب پارامترهای مدل نامعلوم هستند. در این قسمت ابتدا نحوه برآورد پارامترها با رهیافت بیزی ارائه می‌شود، سپس پیشگوی بیزی متغیرهای پنهان چوله‌نرمال بسته به‌دست آورده می‌شود. فرض کنید $\boldsymbol{\nu}$ و Δ معلوم هستند و یک مدل پارامتری به‌صورت $C_\theta(s_i,s_j)=\sigma^2\exp(-||s_i-s_j||/\varphi)$ برای کوواریانس فضایی در نظر گرفته شده باشد، که در آن $||\cdot||$ نرم اقلیدسی، σ و φ به‌ترتیب پارامترهای مقیاس و همبستگی فضایی هستند. همچنین برای سادگی تفسیر چولگی و کاهش بعد پارامتر چولگی در مدل، دو حالت برای ماتریس D در نظر گرفته می‌شود. در حالت اول فرض می‌شود چولگی میدان تصادفی در سراسر ناحیه فضایی U یکسان است، که در این‌صورت می‌توان ماتریس چولگی را به‌صورت $D=\lambda J$ تعریف کرد، که در آن λ پارامتر چولگی و J ماتریس $q\times(n+1)$ بعدی با عناصر یک است. در حالت دوم ناحیه فضایی U به ℓ ناحیه، هر یک با چولگی ثابت تقسیم‌بندی می‌شود، که در این‌صورت می‌توان ماتریس چولگی را به‌صورت $D=(\lambda_1 J_1,\cdots,\lambda_\ell J_\ell)$ تعریف کرد، که در آن $\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_\ell)'$ برداری از پارامترهای چولگی و J_iها ماتریس‌های $q\times n_i$ بعدی با درایه‌های یک و $\sum_{i=1}^{\ell} n_i=n+1$ هستند.

برای سره بودن توزیع پسینی، برای پارامترهای مدل، توزیع‌های پیشینی سره در نظر گرفته می‌شود. توزیع‌های پیشینی متداول برای $\boldsymbol{\beta}$، $N(\boldsymbol{a},B)$ و برای σ و φ، به ترتیب $IG(\alpha,\tau)$ و $\Gamma(\gamma,\omega)$ هستند. اکنون با در نظر گرفتن یک توزیع پیشینی چند متغیره دلخواه برای پارامترهای چولگی و با فرض استقلال، توزیع پیشینی توأم

$$\pi(\boldsymbol{\eta})=\pi(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda})=\pi(\boldsymbol{\beta})\pi(\boldsymbol{\theta})\pi(\boldsymbol{\lambda}),$$

حاصل می‌شود. لذا توزیع پسینی متناسب با

$$\begin{aligned}\pi(\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{y}) &= \pi(\boldsymbol{y}|\boldsymbol{x})\phi_n(\boldsymbol{x};Z\boldsymbol{\beta},\Sigma_\theta)\Phi_q(D(\boldsymbol{x}-Z\boldsymbol{\beta});\boldsymbol{\nu},\Delta)\phi_p(\boldsymbol{\beta};\boldsymbol{a},B)\\ &\quad\times\frac{\tau^\alpha\omega^\gamma}{\Gamma(\gamma)\Gamma(\alpha)}(\sigma^{-\alpha-1}\varphi^{\gamma-1})exp\{-\omega\varphi-\frac{\tau}{\sigma}\}\pi(\boldsymbol{\lambda}),\qquad(۵)\end{aligned}$$

خواهد شد، که دارای فرم پیچیده‌ای است. بنابراین از روش‌های MCMC برای شبیه‌سازی از توزیع پسینی استفاده می‌شود. برای به‌کارگیری الگوریتم نمونه‌گیری گیبز[۹] (گیلکس وهمکاران، ۱۹۹۶) توزیع‌های شرطی کامل پارامترها مورد نیاز است. توزیع شرطی کامل $\boldsymbol{\beta}$ به‌صورت

$$\begin{aligned}\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{x},\sigma,\varphi,\boldsymbol{\lambda}) \quad \propto \quad & exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)^\top \Sigma_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)\right\} \\ & \times\Phi_q\left(-DZ(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta);\boldsymbol{\nu}-D\boldsymbol{x}+DZ\boldsymbol{\mu}_\beta,\Delta\right).\end{aligned}$$

خواهد شد و دارای فرم مشخصی از توزیع CSN به‌صورت

$$[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{x},\sigma,\varphi,\boldsymbol{\lambda}] \sim CSN_{p,q}(\boldsymbol{\mu}_\beta,\Sigma_\beta,-DZ,\boldsymbol{\nu}_\beta,\Delta)$$

است، که در آن $\boldsymbol{\mu}_\beta = \Sigma_\beta(B^{-1}\boldsymbol{a}+Z'\Sigma_\theta^{-1}\boldsymbol{x})$، $\Sigma_\beta = (Z'\Sigma_\theta^{-1}Z+B^{-1})^{-1}$. و $\boldsymbol{\nu}_\beta = \boldsymbol{\nu}-D(\boldsymbol{x}-Z\boldsymbol{\mu}_\beta)$ بنابراین تولید نمونه از توزیع شرطی کامل $\boldsymbol{\beta}$ به‌راحتی امکان‌پذیر است. برای سایر پارامترها داریم

$$\begin{aligned}\pi(\sigma|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta},\varphi,\boldsymbol{\lambda}) \quad &\propto \quad \pi(\boldsymbol{x}|\boldsymbol{\beta},\sigma,\varphi,\boldsymbol{\lambda})\pi(\sigma) \\ &= \quad IG(\alpha,\tau)\phi_n(\boldsymbol{x};Z\boldsymbol{\beta},\Sigma_\theta)\Phi_q(D(\boldsymbol{x}-Z\boldsymbol{\beta}), \\ \pi(\varphi|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta},\sigma,\boldsymbol{\lambda}) \quad &\propto \quad \pi(\boldsymbol{x}|\boldsymbol{\beta},\sigma,\varphi,\boldsymbol{\lambda})\pi(\varphi) \\ &= \quad \Gamma(\gamma,\omega)\phi_n(\boldsymbol{x};Z\boldsymbol{\beta},\Sigma_\theta)\Phi_q(D(\boldsymbol{x}-Z\boldsymbol{\beta}), \\ \pi(\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta},\sigma,\varphi) \quad &\propto \quad \pi(\boldsymbol{x}|\boldsymbol{\beta},\sigma,\varphi,\boldsymbol{\lambda})\pi(\boldsymbol{\lambda}) \\ &= \quad \phi_n(\boldsymbol{x};Z\boldsymbol{\beta},\Sigma_\theta)\Phi_q(D(\boldsymbol{x}-Z\boldsymbol{\beta})\pi(\boldsymbol{\lambda}).\end{aligned}$$

همان‌طور که ملاحظه می‌شود این توزیع‌های شرطی کامل دارای فرم خاصی از توزیع‌های شناخته شده نیستند. بنابراین برای تولید نمونه از الگوریتم متروپولیس-هستینگز استفاده می‌شود. توزیع‌های پیشنهادی برای اجرای الگوریتم متروپولیس-هستینگز برای σ، φ و $\boldsymbol{\lambda}$ به‌ترتیب گامای معکوس، گاما و نرمال در نظر گرفته می‌شود. توزیع شرطی کامل برای هر مولفه بردار متغیرهای پنهان نیز به‌صورت

$$\pi(x_k|\boldsymbol{x}_{-k},\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\lambda}) \quad \propto \quad \pi(\boldsymbol{y}|\boldsymbol{x})\pi(x_k|\boldsymbol{x}_{-k},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda}),\ k=1,\ldots,n.$$

[۹] Gibbs Smpling

است. در نهایت برای پیشگویی بیزی متغیرهای پنهان، توزیع پیشگو به‌صورت

$$\pi(x_\circ|\boldsymbol{y}) = \int \pi(x_\circ|\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda})\pi(\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{y})\; d\boldsymbol{x}\; d\boldsymbol{\beta}\; d\boldsymbol{\theta}\; d\boldsymbol{\lambda}, \qquad (۶)$$

است. همچنین اگر پیشگویی $y_\circ$ در موقعیت $s_\circ$ مد نظر باشد، می‌توان از توزیع پیشگوی بیزی $y_\circ$ نیز به‌صورت زیر استفاده کرد.

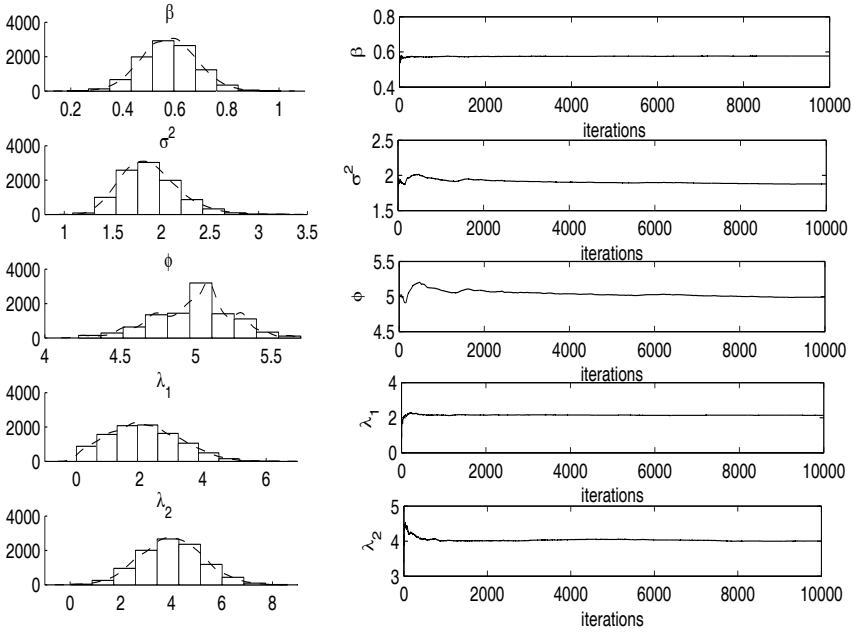
$$\pi(y_\circ|\boldsymbol{y}) = \int_{x_\circ} \pi(y_\circ|x_\circ)\pi(x_\circ|\boldsymbol{y})dx_\circ. \qquad (۷)$$

۴ مطالعه شبیه‌سازی

در $n = ۱۰۰$ موقعیت روی یک مشبکه منظم ۱۰×۱۰، با موقعیت‌های $\{(\ell,k);\ell,k = ۱,\ldots,۱۰\}$ از توزیع $CSN_{۱۰۰,۲}(\beta z,\Sigma_\theta,D,\boldsymbol{\nu},\Delta)$، داده‌های متغیر پنهان $\boldsymbol{x}$ تولید شده‌اند. در این شبیه‌سازی فرض شده است $R_۱$ و $R_۲$ افرازهایی از ناحیه فضایی U هستند که توزیع داده‌ها دارای پارامترهای چولگی متفاوت به ترتیب با مقادیر $\lambda_۲ = ۴$ و $\lambda_۱ = ۲$ هستند. برای ساختار همبستگی فضایی، تابع کواریانس همسانگرد نمایی با $\boldsymbol{\theta} = (\sigma,\varphi) = (۲,۵)$ در نظر گرفته و $\beta = ۰/۵$، $\boldsymbol{\nu} = (۰,۰)'$ و $\Delta = I_۲$ فرض شده است. بردار متغیر کمکی z در (ℓ,k)'امین موقعیت به‌صورت $z_{\ell,k} = \log(۱+\ell)$ در نظر گرفته شد. در ناحیه U، ۱۰۰ موقعیت وجود دارد که موقعیت‌های $s_۱,\ldots,s_{۲۰}$ در ناحیه $R_۱$ و بقیه موقعیت‌های $s_{۲۱},\ldots,s_{۱۰۰}$ در ناحیه $R_۲$ قرار دارند. در این‌صورت ماتریس چولگی D را می‌توان به‌صورت

$$D = \begin{pmatrix} \lambda_۱\mathbf{1}'_{۲۰} & \mathbf{0}'_{۸۰} \\ \mathbf{0}'_{۲۰} & \lambda_۲\mathbf{1}'_{۸۰} \end{pmatrix}_{۲\times ۱۰۰},$$

افراز کرد، که در آن $\mathbf{1}_k$ بردار k بعدی با درایه‌های یک و $\mathbf{0}_k$ بردار k بعدی با درایه‌های صفر است. به‌شرط متغیر پنهان تولید شده، پاسخ‌های گسسته y_j به ازای $j = ۱,\ldots,n$ از توزیع دوجمله‌ای $y_j \sim Bin(u_j,p_j)$, استخراج شده‌اند، که در آن $p_j = \exp(x_j)/(۱+\exp(x_j))$ و $u_j = ۱۰۰$ می‌باشند. برای برآورد بیزی



شکل ۱: (راست): نمودار همگرایی میانگین نمونه‌های تولید شده از توزیع‌های شرطی کامل پارامترها در مقابل تکرار، (چپ): هیستوگرام نمونه‌های تولید شده.

پارامترهای مدل، توزیع‌های پیشینی

$$\beta \sim N(\mathbf{0}, ۱۰),\ \sigma^{۲} \sim IG(۲, ۵),\ \theta \sim \Gamma(۳, ۵),\ \boldsymbol{\lambda} = (\lambda_۱, \lambda_۲)' \sim N_۲(\mathbf{0}, ۵I).$$

منظور شده‌اند. با توجه به بررسی حساسیت پیشین، توزیع‌های پیشینی با واریانس بالا اتخاذ گردیده‌اند. سپس الگوریتم‌های MCMC با ۱۰۰۰۰ تکرار اجرا و نمودارهای همگرایی پارامترها در شکل ۱ (راست) رسم شده‌اند که نشانگر همگرایی الگوریتم‌ها هستند. نمودار هیستوگرام نمونه‌های تولید شده از توزیع‌های شرطی کامل پارامترها در شکل ۱ (چپ) توزیع پسینی کناری پارامترها را نشان می‌دهد.

رهیافت بیزی برای مدل‌های SGLM نرمال و CSN اجرا و نتایج در جدول ۱ خلاصه شده‌اند. برای محاسبه متوسط برآوردهای بیزی پارامترها و MSEها، ۱۰۰ مجموعه داده با مفروضات اولیه شبیه‌سازی تولید شده است. همان‌طور که ملاحظه

می‌شود متوسط برآوردهای بیزی پارامترها در مدل CSN نسبت به مدل نرمال به مقادیر واقعی نزدیک‌تر و از MSEهای کوچک‌تری برخوردارند. برای بررسی دقت پیشگوی بیزی، معیار CVMSE برای دو مدل SGLM با متغیرهای پنهان چوله‌نرمال بسته و نرمال محاسبه و به‌ترتیب مقادیر ۰/۰۰۱ و ۰/۰۱۰۳ به‌دست آمد، که بیان‌گر دقیق‌تر بودن پیشگویی بیزی در مدل SGLM باتوزیع پیشینی CSN برای متغیر پنهان است.

جدول ۱: متوسط برآورد بیزی پارامترها و MSE برای دو مدل SGLM با متغیرهای پنهان چوله‌نرمال بسته و نرمال.

		چوله‌نرمال بسته		نرمال	
پارامتر	مقدار واقعی	برآورد	MSE	برآورد	MSE
β	۰/۵	۰/۵۴۰۱	۰/۰۰۵۷	۰/۵۶۳۵	۰/۰۶۳۵
σ^2	۲	۱/۷۸۲۹	۰/۰۶۷۴	۳/۳۱۹۲	۱/۷۵۱۴
φ	۵	۴/۹۵۷۶	۰/۰۱۲۰	۵/۲۲۸۱	۰/۰۹۳۶
λ_1	۲	۲/۱۲۹۵	۰/۰۱۶۹	-----	-----
λ_2	۴	۳/۹۹۴۷	۰/۰۰۰۴	-----	-----

۵ پیشگوی بیز تقریبی در مدل SGLM با متغیرهای پنهان چوله‌نرمال بسته

در رهیافت بیزی نیاز به نمونه‌های مونت کارلویی و اجرای الگوریتم‌های تکرارشونده است که دارای محاسبات زمان‌بر هستند. ایدسویک و همکاران (۲۰۰۹) با در نظر گرفتن توزیع پیشینی نرمال برای متغیرهای پنهان در مدل‌های SGLM، نشان دادند توزیع پسینی تقریبی این متغیرها متعلق به خانواده توزیع نرمال است. حسینی و همکاران (۲۰۰۹، ۲۰۱۱) با تعمیم روش بیز تقریبی ایشان به مدل‌های SGLM با متغیرهای پنهان چوله‌نرمال، نشان دادند توزیع پسینی این متغیرها نیز به‌طور تقریبی CSN خواهد بود.

قضیه ۲ (حسینی و همکاران، ۲۰۱۱) اگر در مدل SGLM متغیرهای پنهان دارای توزیع پیشینی $(\boldsymbol{x}|\boldsymbol{\eta}) \sim SN(H\boldsymbol{\beta}, \Sigma_\theta, \boldsymbol{\lambda})$ باشند، آن‌گاه توزیع تقریبی پسینی آن‌ها

به‌صورت

$$(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta}) \approx CSN_{n,1}(\boldsymbol{\mu}_{x|y,\eta}, \Sigma_{x|y,\eta}, D_{x|y,\eta}, \boldsymbol{\nu}_{x|y,\eta}, 1), \qquad (8)$$

است، که در آن

$$\begin{aligned}
\boldsymbol{\mu}_{x|y,\eta} &= H\boldsymbol{\beta} + \Sigma_\theta A' R^{-1}(z(\boldsymbol{y}, \boldsymbol{x}^{obs}) - AH\boldsymbol{\beta}),\\
\Sigma_{x|y,\eta} &= \Sigma_\theta - \Sigma_\theta A' R^{-1} A \Sigma_\theta,\\
D_{x|y,\eta} &= \boldsymbol{\lambda}' \Sigma_\theta^{-\frac{1}{2}}\\
\boldsymbol{\nu}_{x|y,\eta} &= \boldsymbol{\lambda}' \Sigma_\theta^{-\frac{1}{2}}(H\boldsymbol{\beta} - \boldsymbol{\mu}_{x|y,\eta}),
\end{aligned}$$

به طوری که $R = A\Sigma_\theta A' + P$، P ماتریس قطری با درایه‌های $P_{ii} = 1/b''(x_i)$، $i = 1, \ldots, k$ و $z_i(y_i, x_i) = [y_i - b'(x_i) + x_i b''(x_i)]/b''(x_i)$ هستند.
بنابراین برای تعیین توزیع پسینی تقریبی متغیرهای پنهان ابتدا مقدار اولیهٔ $\boldsymbol{x}^{(0)}$، به‌عنوان مثال $\boldsymbol{x}^{(0)} = E(\boldsymbol{x}|\boldsymbol{\eta})$، انتخاب و $m = 0$ قرار داده می‌شود. سپس از قضیه ۲، توزیع پسینی تقریبی برای $\boldsymbol{x}^{(m)}$، به‌صورت

$$\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta}) \equiv CSN_{n,1}(\hat{\boldsymbol{\mu}}_{x|y,\eta}(\boldsymbol{x}^{(m)}), \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{(m)}), D_{x|y,\eta}, \hat{\boldsymbol{\nu}}_{x|y,\eta}(\boldsymbol{x}^{(m)}), 1)$$

تعیین می‌شود. اکنون قرار دهید

$$\boldsymbol{x}^{(m+1)} = \hat{\mathrm{E}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta}) = \hat{\boldsymbol{\mu}}_{x|y,\eta}(\boldsymbol{x}^{(m)}) + \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{(m)}) D'_{x|y,\eta} \hat{\psi},$$

که در آن $\hat{\psi} = \psi(0, \hat{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{(m)}), D_{x|y,\eta}, \hat{\nu}_{x|y,\eta}(\boldsymbol{x}^{(m)}), 1)$ است. با قرار دادن $m = m + 1$ و تکرار مراحل الگوریتم تا رسیدن به همگرایی، توزیع پسینی تقریبی تعیین می‌شود.

۱.۵ *برآورد بیز تقریبی پارامترهای مدل*

با به‌کار بردن تقریب CSN برازش شده به توزیع شرطی کامل $\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})$، یک چگالی تقریبی برای توزیع پسینی پارامترهای مدل به‌صورت

$$\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})}{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})}\Big|_{\boldsymbol{x}=\hat{E}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})}. \qquad (9)$$

حاصل می‌شود، که با جایگذاری تقریب به دست آورده شده برای توزیع پسینی متغیرهای پنهان به‌صورت $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})$ و قرار دادن آن در مخرج کسر (۹)، می‌توان تقریب $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$ را برای توزیع پسینی پارامترها به دست آورد. برای تعیین این تقریب، حسینی و همکاران (۲۰۱۱) الگوریتمی پیشنهاد دادند که به‌طور خلاصه به این‌صورت است که ابتدا مد تابع $\log\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$ تعیین $(\boldsymbol{\eta}^*)$ و سپس ماتریس هسین $H=\left(\frac{\partial^2\log\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})}{\partial\eta^{(i)}\partial\eta^{(j)}}\right)$ محاسبه و عکس آن به‌صورت $H^{-1}=V\Lambda V'$ تجزیه شود، که در آن V ماتریس بردارهای ویژه و Λ ماتریس قطری مقادیر ویژه آن است. مبدأ مختصات به مد $\boldsymbol{\eta}^*$ منتقل که فرمول مختصات در مبدا $\boldsymbol{\eta}^*$ به‌صورت

$$\boldsymbol{\eta}(t)=\boldsymbol{\eta}^*+V\Lambda^{\frac{1}{2}}t, \qquad (۱۰)$$

تعریف می‌شود، که در آن t مقادیر استاندارد شده هستند. به‌عنوان مثال برای حالت دو بعدی $t=(t_1,t_2)$، با قرار دادن مبدأ مختصات $t=(0,0)$، در رابطه (۱۰)، عبارت $\boldsymbol{\eta}(0)=\boldsymbol{\eta}^*$ حاصل می‌شود. با شروع از مبدأ مختصات جدید روی هر یک از محورها نقاطی به فاصله مقادیر صحیح δ_t به گونه‌ای اختیار شوند، که شرط

$$\log\hat{\pi}(\eta(0)|\boldsymbol{y})-\log\hat{\pi}(\eta(t)|\boldsymbol{y})<\delta_\pi, \qquad (۱۱)$$

برقرار باشد. سپس به‌طور مشابه نقاط درون صفحات نیز تعیین می‌شوند. معمولاً در نامساوی (۱۱)، مقدار $\delta_\pi=2/5$ در نظر گرفته می‌شود. با به‌کار بردن این الگوریتم η_iهایی از توزیع $\pi(\boldsymbol{\eta}|\boldsymbol{y})$ تولید می‌شوند که می‌توان از آن‌ها برای تقریب توزیع $\pi(\boldsymbol{\eta}|\boldsymbol{y})$ استفاده کرد. این الگوریتم برای مدل‌هایی که بعد بردار پارامتری کمی دارند دقیق‌تر است. چون در عمل پارامتر β_0 از بردار پارامترهای رگرسیونی کمتر مورد توجه است و در راستای کاهش تعداد پارامترهای مدل طبق قضیه زیر می‌توان آن را با کناری‌سازی از بردار پارامترهای مدل حذف و سایر پارامترهای مورد علاقه رگرسیونی را در بردار پارامترهای مدل نگه داشت.

قضیه ۳ (حسینی و همکاران، ۲۰۱۱) اگر در مدل SGLM توزیع $\pi(\boldsymbol{x}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda})$ به فرم (۱) و پارامتر $\boldsymbol{\beta}$ دارای توزیع پیشینی $N(\boldsymbol{a},B)$ باشد، آن‌گاه

$$(\boldsymbol{x}|\boldsymbol{\theta},\lambda_0)\sim CSN_{n,1}(\boldsymbol{\mu}_x,\Sigma_x,D_x,\boldsymbol{\nu}_x,\Delta_x), \qquad (۱۲)$$

که در آن $\boldsymbol{\mu}_x = H\boldsymbol{a}$، $\Sigma_x = (\Sigma_\theta + HBH')$، $D_x = \boldsymbol{\lambda}'\Sigma_\theta^{\frac{1}{2}}\Sigma_x^{-1}$، $\boldsymbol{\nu}_x = 0$ و $\Delta_x = 1 + \boldsymbol{\lambda}'\boldsymbol{\lambda} - \boldsymbol{\lambda}'\Sigma_\theta^{\frac{1}{2}}\Sigma_x^{-1}\Sigma_\theta^{\frac{1}{2}}\boldsymbol{\lambda}$.

بنابراین با کناری‌سازی می‌توان پارامتر β_0 را از بردار پارامترهای رگرسیونی حذف کرد.

۲.۵ پیشگویی فضایی بیز تقریبی

یکی از اهداف مدل SGLM، پیشگویی متغیرهای پنهان $\boldsymbol{x}^{pred}$، در موقعیت‌های فاقد مشاهده است. تقریب CSN برای توزیع پسینی متغیرهای پنهان، نقش کلیدی در به‌دست آوردن توزیع پیشگوی بیزی دارد. با توجه به تقریب (۸) برای $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})$ و بسته بودن توزیع CSN نسبت به کناری‌سازی، $\hat{\pi}(x_j|\boldsymbol{y},\boldsymbol{\eta})$ متعلق به خانواده توزیع‌های CSN است. بنابراین توزیع تقریبی پیشگو برای متغیرهای پنهان به‌صورت

$$\hat{\pi}(x_j|\boldsymbol{y}) = \sum_{\ell} \hat{\pi}(x_j|\boldsymbol{y},\boldsymbol{\eta}_\ell) \times \hat{\pi}(\boldsymbol{\eta}_\ell|\boldsymbol{y}), \quad j = 1,\cdots,n,$$

حاصل می‌شود، که در آن ℓ تعداد نقاط تولید شده از الگوریتم توزیع پسینی تقریبی پارامترها می‌باشد. در عمل پیشگویی در موقعیت‌های $j = k+1,\ldots,n$ مدنظر می‌باشد.

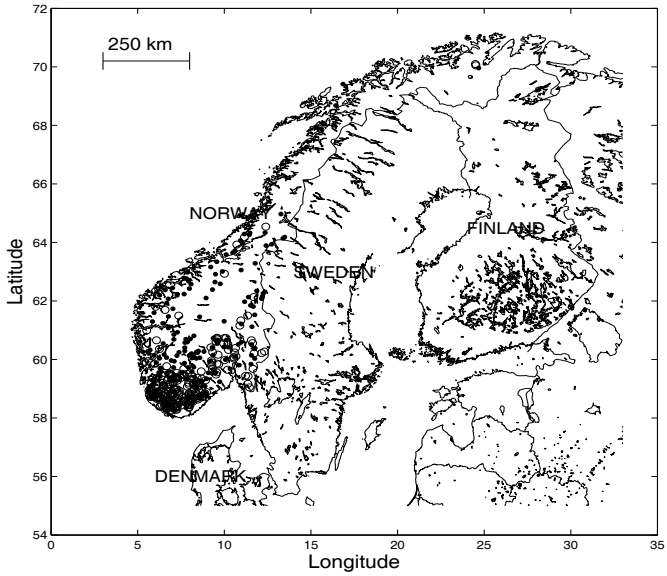
همچنین می‌توان توزیع پیشگوی y_j در موقعیت‌های فاقد مشاهده، را به‌صورت

$$\pi(y_j|\boldsymbol{y}) = \int_{x_j} \pi(y_j|x_j)\pi(x_j|\boldsymbol{y}), \quad j = k+1,\ldots,ndx_j. \qquad (۱۳)$$

به‌دست آورد، که با یک مجموع روی مقادیر x_j تقریب زده می‌شود.

۳.۵ داده‌های آلودگی دریاچه

اسیدسازی دریاچه فرآیندی است که در آن میانگین سالانه تغییرات اسیدی آن، یعنی مقدار طبیعی PH به مقدار کمتر از ۵/۶ برسد. به عبارت دیگر، آب طبیعی دارای PH=7 است که درصورت سکون ممکن است از طریق دی اکسید کربن اتمسفر



شکل ۲: موقعیت داده‌های آلودگی ماهی‌های قزل‌آلا در دریاچه‌های نروژ، دایره‌های توپر مکان ماهی‌های آلوده و دایره‌های توخالی مکان ماهی‌های آلوده نشده هستند.

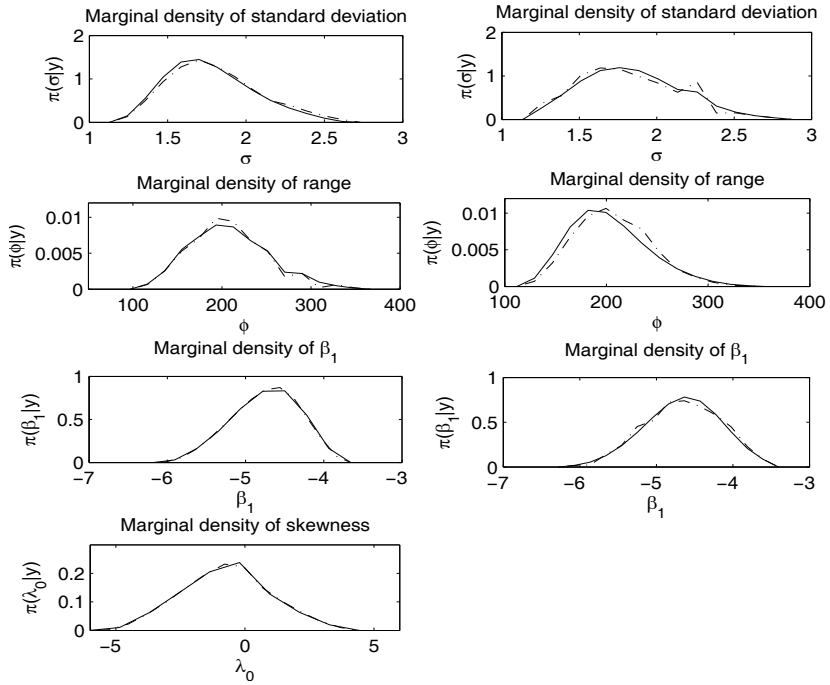
اسیدی بشود و مقدار PH آن تغییر کند و زمانی که این مقدار به ۵/۶ می‌رسد، آب در حالت اسیدی شده است. این تغییر در PH بر روی بعضی از گیاهان و جانوران دریایی به خصوص ماهی‌ها تاثیرگذار است. برای مثال، ماهی قزل‌آلا به اسیدسازی دریاچه بسیار حساس است. ورین و همکاران (۲۰۰۵)، در ۵۴۲ موقعیت واقع در دریاچه‌های نروژ، ماهی‌های قزل‌آلایی که توسط ماهیگیران جمع‌آوری شده بود، مورد مطالعه قرار دادند. آن‌ها برای تحلیل این داده‌ها از مدل SGLM با متغیرهای پنهان نرمال استفاده کردند. متغیر پاسخ برنولی است، که وضعیت آلودگی ماهی‌های آن موقعیت را نشان می‌دهد. یعنی اگر اسید سازی دریاچه باعث آلودگی ماهی‌ها شده باشد، متغیر پاسخ مقدار ۱ و در غیر این‌صورت مقدار ۰ را اختیار می‌کند. در شکل ۲ موقعیت داده‌ها نشان داده شده است. ورین و همکاران (۲۰۰۵)، مجموعه‌ای ۴۰۰تایی از ۵۴۲ داده را به تصادف انتخاب نموده و برای تحلیل در نظر گرفتند و از ۱۴۲ داده باقیمانده برای اعتبار سنجی متقابل استفاده کردند. در این مطالعه مدل SGLM با متغیرهای پنهان چوله‌نرمال برای داده‌ها

به‌کار گرفته شده است. برای بررسی و مقایسه‌های دقیق‌تر ۱۹ مجموعه داده ۱۴۲تایی دیگر از بین ۵۴۲ مجموعه داده اولیه انتخاب شده‌اند و با ۲۰ مجموعه داده ۱۴۲ تایی که در دسترس می‌باشد پیشگویی به روش اعتبار سنجی متقابل انجام شده است. ظرفیت خنثی سازی اسید[۱۰] (ANC) به‌عنوان متغیر کمکی در نظر گرفته شده است. مشاهدات پاسخ تحقق‌های متغیر برنولی مستقل شرطی با توزیع $\pi(y_i|x_i) = \exp\{y_i x_i - \log(1+\exp(x_i))\}$ در نظر گرفته شده است. توزیع متغیرهای پنهان x چوله‌نرمال به فرم $SN(\beta_0 + \beta_1 h, \Sigma_\theta, \boldsymbol{\lambda})$ فرض شده است، که در آن β_0 پارامتر سطح، β_1 اثر متغیر کمکی ANC، $\boldsymbol{\lambda} = \lambda_0 \mathbf{1}$ پارامتر چولگی و Σ_θ ماتریس کواریانس با ساختار همبستگی فضایی همسانگرد نمایی با پارامترهای (φ, σ) است. برای β_0 توزیع پیشینی $N(0/5, 1)$ فرض شد که با استفاده از رابطه (۱۲) از پارامترهای مدل حذف گردید. برای سایر پارامترها توزیع‌های پیشینی به‌صورت

$$\beta_1 \sim N(-2, 5),\ \lambda_0 \sim N(0, 5),\ \sigma \sim IG(2, 1),\ \varphi \sim \Gamma(250, 2500)$$

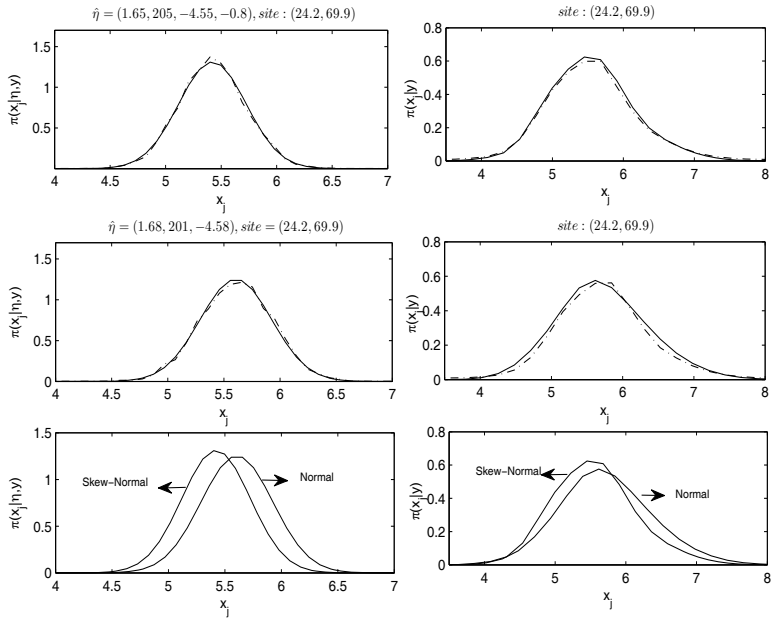
در نظر گرفته شده است. اکنون با به‌کار بردن مدل SGLM چوله‌نرمال و استنباط بیز تقریبی به روش عددی توزیع‌های کناری پسینی به‌دست آورده شدند. در شکل ۳ (چپ) توزیع‌های کناری پسینی برای $\boldsymbol{\eta} = (\sigma, \varphi, \beta_1, \lambda_0)$ رسم شده است. خطوط ممتد با روش بیز تقریبی و خطوط نقطه‌چین با روش MCMC به‌دست آمده‌اند. همان طور که ملاحظه می‌شود تفاوت قابل ملاحظه‌ای بین نتایج این دو روش وجود ندارد. نرخ پذیرش برای نمونه‌های MCMC حدود ۹۰% می‌باشد. همچنین روش بیز تقریبی برای مدل SGLM با متغیرهای پنهان نرمال استفاده شد و توزیع‌های کناری پسینی برای $\boldsymbol{\eta} = (\sigma, \varphi, \beta_1)$ در شکل ۳ (راست) نشان داده شده است. در این حالت نیز نتایج روش MCMC و بیز تقریبی بسیار مشابه هم هستند. در شکل ۴ (چپ) نمودار توزیع پیشگو برای یک موقعیت فاقد مشاهده دلخواه با طول و عرض جغرافیایی (۲۴/۲, ۶۹/۹) رسم شده است. برای هر دو مدل SGLM با متغیرهای پنهان نرمال و چوله‌نرمال، نمودار توزیع شرطی پیشگو با جایگذاری مد توزیع

[۱۰] Acid Neutralizing Capacity



شکل ۳: نمودار توزیع‌های کناری پسینی پارامترها با روش بیز تقریبی (خطوط ممتد) و روش $MCMC$ (خط‌چین) برای مدل (راست) نرمال و (چپ) چوله‌نرمال.

پسینی تقریبی به‌جای بردار پارامتر $\boldsymbol{\eta}$ رسم شده‌اند. نمودار مشخص شده با خطوط ممتد از روش بیز تقریبی و خطوط خط چین از روش MCMC به‌دست آمده‌اند. مد توزیع پیشگو برای مدل چوله‌نرمال حدود ۵/۶ و برای مدل نرمال حدود ۵/۴ می‌باشند. همچنین شکل ۴ (راست) نشان دهنده کناری‌های توزیع پیشگو دو مدل که حاصل از دو روش بیز تقریبی $\hat{\pi}(x_j|\boldsymbol{y})$ و روش MCMC، $\pi^{MCMC}(x_j|\boldsymbol{y})$ می‌باشد. زمان محاسبات به روش بیز تقریبی حدود ۵۰ ثانیه و MCMC حدود ۲۰ روز به‌طول انجامیده است. در انتها دو مدل SGLM با متغیرهای پنهان چوله‌نرمال و نرمال با روش اعتبارسنجی متقابل مورد مقایسه قرار گرفته است. بدین منظور ۲۰ مجموعه داده ۱۴۲تایی به‌طور تصادفی از مجموعه داده اصلی استخراج شده است.



شکل ۴: (راست) چگالی کناری پیشگو $\pi(x_j|\boldsymbol{y})$ حاصل از بیز تقریبی مستقیم (خطوط ممتد) و روش $MCMC$ (خط‌چین) برای مدل چوله‌نرمال (بالا)، مدل نرمال (وسط) و هر دو مدل (پایین). (چپ) چگالی شرطی $\hat{\pi}(x_j|\boldsymbol{y}, \hat{\boldsymbol{\eta}})$ حاصل از بیز تقریبی مستقیم (خطوط ممتد) و روش $MCMC$ (خط‌چین) برای مدل چوله‌نرمال (بالا)، مدل نرمال (وسط) و هر دو مدل (پایین).

برای هر موقعیت j در هر مجموعه داده دسته‌بندی به‌صورت

$$\hat{y}_j = \begin{cases} ۰ & \hat{\pi}(۰|\boldsymbol{y}) > \hat{\pi}(۱|\boldsymbol{y}) \\ ۱ & \hat{\pi}(۱|\boldsymbol{y}) > \hat{\pi}(۰|\boldsymbol{y}) \end{cases} \qquad (۱۴)$$

انجام شده است. با این دسته‌بندی چهار وضعیت پیشگویی $(y_j = ۰, \hat{y}_j = ۰)$، $(y_j = ۱, \hat{y}_j = ۱)$، $(y_j = ۰, \hat{y}_j = ۱)$ و $(y_j = ۱, \hat{y}_j = ۰)$ وجود دارد. جدول ۲ نشان می‌دهد که تعداد پیشگویی‌های غلط برای مدل SGLM با متغیرهای پنهان نرمال بیشتر از مدل SGLM با متغیرهای پنهان چوله نرمال است. برای اطمینان بیشتر

جدول ۲: نتایج اعتبارسنجی متقابل برای دو مدل نرمال و چوله‌نرمال.

مدل	نتیجه	نرمال		مجموع
		درست	نادرست	
چوله نرمال	درست	۲۳۹۹	۸۹	۲۴۸۸
	نادرست	۳۱	۳۲۱	۳۵۲
	مجموع	۲۴۳۰	۴۱۰	۲۸۴۰

آزمون مک‌نمار[۱۱] (مک‌نمار، ۱۹۴۷) انجام و مقدار آماره آن به‌صورت

$$M = \frac{(|۸۹ - ۳۱| - ۰/۵)^۲}{۸۹ + ۳۱} = ۲۷/۵$$

حاصل شد. با توجه به آن که $M \sim \chi^۲_۱$، آزمون در سطح ۰/۰۵ معنی‌دار شده، یعنی دو مدل متفاوت عمل کرده‌اند و نتایج جدول ۲ بیان‌گر آن است که پیشگویی مدل SGLM با متغیرهای پنهان چوله‌نرمال به‌طور معنی‌داری بهتر از مدل SGLM با متغیرهای پنهان نرمال می‌باشد.

بحث و نتیجه‌گیری

وجود متغیرهای پنهان فضایی در مدل‌های SGLM و نامعلوم بودن توزیع واقعی آنها، روی برآورد پارامترهای مدل و دقت پیشگویی تأثیر گذار می‌باشد. لذا در عمل فرض نرمال بودن متغیرهای پنهان فضایی گاهی فرض نادرست و گمراه کننده است. برای افزایش دقت برآورد پارامترها و پیشگوها، استفاده از توزیع‌های چوله‌نرمال و چوله‌نرمال بسته که کلاس بزرگ‌تر و انعطاف‌پذیرتری از کلاس توزیع نرمال هستند، برای متغیرهای پنهان پیشنهاد گردید. پیشگوی MMSE برای متغیرهای پنهان در موقعیت‌های فاقد مشاهده معرفی شد و پیشگوی فضایی بیزی برای متغیرهای پنهان چوله ارائه و برآورد بیزی پارامترها و متغیرهای پنهان در موقعیت‌های دارای مشاهده توسط الگوریتم‌های MCMC محاسبه شدند. از آن‌جا که اجرای الگوریتم‌های MCMC برای این مدل‌ها بسیار طولانی و زمان‌بر هستند، روش بیز تقریبی ارائه گردید. در مطالعه شبیه‌سازی و مثال کاربردی نشان داده شد که روش بیز تقریبی

[۱۱] McNemar

معادل روش بیز معمولی و الگوریتم‌های MCMC عمل می‌کند با این تفاوت که روش تقریبی ارائه شده نیاز به چند ثانیه اجرای کامپیوتری نیاز دارند در صورتی که روش‌های MCMC نیاز به چندین روز اجرای کامپیوتری دارند.

در مطالعه شبیه‌سازی برتری مدل SGLM با متغیرهای پنهان چوله نسبت به مدل مذکور با متغیرهای پنهان نرمال نشان داده شد. به‌علاوه نتایج به دست آمده بیانگر این مطلب است که داده‌های فضایی اطلاعات زیادی در مورد پارامترهای رگرسیونی و واریانس و دامنه همبستگی دارند، اما در مورد پارامتر چولگی اطلاع زیادی در اختیار نمی‌دهند. با تحلیل‌های حساسیت و استفاده از توزیع‌های پیشینی متفاوت برای پارامتر چولگی، نشان داده شد که انتخاب توزیع‌های پیشینی متفاوت برای این پارامتر تاثیر چندانی بر پارامترهای دیگر مدل و همچنین پیشگویی در نقاط جدید ندارد. تحلیل بیز تقریبی روی مجموعه داده آلودگی دریاچه‌های نروژ بیان‌گر آن است که که روش بیز تقریبی معادل روش بیز معمولی عمل می‌کند.

چون توزیع t–چوله دارای دم‌های پهن‌تری نسبت به توزیع چوله نرمال و چوله‌نرمال بسته می‌باشد، در بعضی مسائل استفاده از توزیع t–چوله به‌عنوان توزیع متغیرهای پنهان و تعمیم روش‌های تحلیلی ارائه شده برای این خانواده، ممکن است نتایج بهتر و قابل اعتمادتری ارائه نماید.

مراجع

Azzalini, A. (1985), A Class of Distributions which Includes the Normal Ones, *Scandinavian Journal of Statistics,* **12**, 171-178.

Baghishani, H. and Mohammadzadeh, M. (2011), A Data Cloning Algorithm for Computing Maximum Likelihood Estimates in Spatial Generalized Linear Mixed Models, *Computational Statistics and Data Analysis*, **55**, 1748-1759.

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

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

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

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.

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.

Dominguez-Molina, J., Gonzalez-Farias, G., and Gupta, A. (2003), The Multivariate Closed Skew Normal Distribution. *Technical Report 03-12,* Department of Mathematics and Statistics, Bowling Green State University.

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

Hosseini, F., Eidsvik, J., and Mohammadzadeh, M. (2009), Bayesian Inference in Spatial Models with Skew Normal Latent Variables, *Sym-*

posium Workshop on Markov Chain-Monte Carlo, Mathematics Institute Warwick, UK.

Hosseini, F., Eidsvik, J., and Mohammadzadeh, M. (2011), Approximate Bayesian Inference in Spatial GLMM with Skew Normal Latent Variables, *Computational Statistics and Data Analysis*, **55**, 1791-1806.

Hosseini, F., and Mohammadzadeh, M. (2012), Bayesian Prediction for Spatial GLMM with Closed Skew Normal Latent Variables, In Press in *Australian and New Zealand Journal of Statistics.*

Karimi, O.,Hosseini, F., and Mohammadzadeh, M. (2011), Pairwise Likelihood in Spatial GLMM with Skew Normal Latent Variables, *Proceedings of the 15th Conference of the International Association for Mathematical Geology*, Salzburg, September 5-9, 61-67.

Karimi, O., and Mohammadzadeh, M. (2007), Bayesian Spatial Prediction for Closed Skew Gaussian Random Field. *Proceedings of the 12th Conference of the International Association for Mathematical Geology*, China, August 26-31, 684-687.

Karimi, O., and Mohammadzadeh, M. (2009), Bayesian Spatial Prediction for Discrete Closed Skew Gaussian Random Field. *Mathematical Geosciences,* **43** 565-583.

Karimi, O., and Mohammadzadeh, M. (2010), Bayesian Spatial Regression Models with CSN Correlated Errors and Missing Observations. *Statistical Paper,* **53**, 205-218.

Karimi, O., Omre, H., and Mohammadzadeh, M., (2010), Bayesian Closed Skew Gaussian Inversion of Seismic AVO Data for Elastic Material Properties, *Geophysics,* **75**, R1-R11.

Kim, H. M., and Mallick, B. K. (2004), A Bayesian Prediction Using the Skew Gaussian Distribution. *Journal of Statististical Planning and Inference,* **120**, 85-101.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall, London,.

McNemar Q. (1947), Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages, *Psychometrika,* **12**, 153-157.

Mohammadzadeh, M., Hosseini, F. (2011), Maximum-Likelihood Estimation for Spatial GLM Models with Closed-Skew Normal Latent Variables, *Procedia Environmental Sciences,* **3** ,63-68.

Varin, C., Host, G., and Skare, O. (2005), Pairwise LikelihoodInference in Spatial Generalized Linear Mixed Models, *Computational Statistics and Data Analysis,* **49**, 1173-1191.

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

Zhu, H., Gu, M., and Peterson, B. (2007), Maximum Likelihood from Spatial Random Effects Models via the Stochastic Approximation Expectation Maximization Algorithm, *Statistics and Computing*, **17**, 163-177.

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱

مجموعه مقالات، ص ۸۳-۹۱

تحلیل بیزی مدل‌های پروبیت فضایی برای متغیر پاسخ دودویی

حمیدرضا رسولی، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده : معمولاً تحلیل مدل‌های رگرسیون پروبیت برای مدل‌بندی متغیرهای پاسخ دودویی با فرض استقلال خطاها صورت می‌گیرد. اما در عمل با موارد زیادی مانند داده‌های فضایی مواجه می‌شویم که مشاهدات دودویی از لحاظ موقعیت قرار گرفتن در فضای مورد مطالعه به یکدیگر وابسته‌اند. بنابراین خطاهای مدل نیز همبسته خواهند بود و لازم است این همبستگی فضایی در مدل‌بندی داده‌ها لحاظ شود. در این مقاله مدل پروبیت فضایی برای تحلیل داده‌های دودویی بیان شده و پارامترهای آن با فرض آن که بین متغیرهای پنهان یا خطاهای مدل رابطهٔ اتورگرسیو فضایی برقرار باشد، برآورد شده است. در انتها نحوه کاربست مدل در مثالی کاربردی نشان داده شده است.

واژه‌های کلیدی : مدل پروبیت، اتورگرسیو فضایی، متغیر پنهان.

آدرس الکترونیک مسئول مقاله: حمیدرضا رسولی، hr_rasoli64@yahoo.com

کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲F۱۵، ۶۲H۱۱

۱ مقدمه

در تحلیل رگرسیون چنانچه خطاها مستقل و متغیر پاسخ گسسته، شمارشی یا رسته‌ای باشد از مدل‌های خطی تعمیم‌یافته[۱] که توسط نلدر و ودربرن (۱۹۷۲) ارائه گردیده‌اند، استفاده می‌شود. در این گونه مدل‌ها، تابعی از میانگین متغیر تصادفی پاسخ با متغیرهای تبیینی رابطه خطی برقرار می‌کند، که تابع پیوند[۲] نامیده می‌شود. چنانچه این تابع، توزیع تجمعی نرمال استاندارد باشد، مدل حاصل را مدل پروبیت[۳] می‌نامند. مدل پروبیت برای تعیین ارتباط متغیرهای پاسخ دودویی و متغیرهای تبیینی تحت فرض استقلال خطاها مورد استفاده قرار می‌گیرد (کاکس و اسنل، ۱۹۸۹). در زمینه‌های کاربردی مانند ژئوفیزیک، کشاورزی، علوم اقتصادی و اجتماعی، پزشکی و بازسازی تصاویر، داده‌هایی وجود دارند که از لحاظ موقعیت قرار گرفتن در فضای مورد مطالعه به یکدیگر وابسته‌اند. چنانچه بین مشاهدات وابستگی فضایی وجود داشته باشد چندین شیوه برای برآورد پارامترهای مدل پیشنهاد شده است. مکمیلن (۱۹۹۲) الگوریتم EM را پیشنهاد کرده است. لسج (۲۰۰۰) رهیافت بیزی را با فرض آنکه بین متغیرهای پنهان رابطه اتورگرسیو برقرار باشد برای برآورد پارامترها در مدل پروبیت فضایی به‌کار گرفت. برون و ویجوربرگ (۲۰۰۴) نمونه‌گیری از نقاط مهم بازگشتی را برای برآورد مدل پیشنهاد کردند. لسج و پس (۲۰۰۹) از رهیافت بیزی برای برآورد پارامترها استفاده کردند و توزیع نرمال چندمتغیره بریده شده را برای نمونه‌گیری از متغیر پنهان پیشنهاد دادند. در این مقاله مدل‌های اتورگرسیو تأخیرفضایی و خطافضایی معرفی می‌شوند. سپس مدل پروبیت برای تحلیل مشاهدات دودویی خودهمبسته فضایی معرفی می‌شود و با رهیافت بیزی پارامترهای مدل برآورد می‌شوند. در انتها مدل‌های ارائه شده در تحلیل داده‌های معاملات مسکونی شهر تهران به‌کار گرفته خواهد شد.

---

۱ Generalized linear models

۲ Link function

۳ Probit model

۱.۱ مدل اتورگرسیو فضایی

برای مشاهدات فضایی می‌توان از مدل اتورگرسیو فضایی به‌صورت

$$Y = \rho W_1 Y + X\beta + U, \;\; U = \lambda W_2 U + \epsilon, \;\; \epsilon \sim N(0, \sigma^2 I), \qquad (1)$$

استفاده نمود، که در آن Y بردار $n \times 1$ متغیرهای پاسخ، $X_{n\times k}$ ماتریس طرح متشکل از متغیرهای تبیینی، ρ و λ ضرایب مدل اتورگرسیو، β بردار $k \times 1$ ضرایب رگرسیونی، U خطاهایی هستند که بین آن‌ها وابستگی فضایی وجود دارد، W_1 و W_2 ماتریس‌های وزن فضایی با بعد $n \times n$ است. ماتریس وزن فضایی را می‌توان براساس طول و عرض جغرافیایی یا مجاورت تعیین کرد. در ماتریس مجاورت عناصر w_{ij} برای مشاهدات i و j که همسایگی با یکدیگر دارند مقدار مثبت و برای مشاهدات غیر همسایه مقدار صفر اختیار می‌کند. به علاوه $w_{ii} = 0$، یعنی وزن همسایگی هر مشاهده با خودش صفر در نظر گرفته می‌شود. در ماتریس مجاورت می‌توان وزن‌ها را به‌صورت استاندارد شده که از تقسیم وزن‌ها بر مجموع هر سطر به‌دست می‌آید، استفاده کرد. در این‌صورت مجموع عناصر هر سطر برابر یک است. وزن‌های ماتریس فاصله معمولاً براساس معکوس فاصله نقاط از یکدیگر به‌صورت $w_{ij} = d_{ij}^{-\alpha}$ تعیین می‌شوند، که در آن $\alpha > 0$ و d_{ij} اندازه فاصله دو زوج مشاهده (x_i, y_i) و (x_j, y_j) می‌باشد. به‌عنوان مثال می‌توان فاصله بین نقاط را فاصله اقلیدسی در نظر گرفت. اگر در مدل (۱)، $W_2 = 0$ قرار داده شود، یک مدل آمیخته رگرسیون اتورگرسیو فضایی به‌صورت

$$Y = \rho W_1 Y + X\beta + \epsilon, \;\; \epsilon \sim N(0, \sigma^2 I),$$

به‌دست می‌آید، که به آن مدل تأخیرفضایی[۴] نیز گفته می‌شود و اگر در مدل (۱) قرار داده شود $W_1 = 0$، مدل خطافضایی عبارت است از:

$$Y = X\beta + \lambda W_2 U + \epsilon, \;\; \epsilon \sim N(0, \sigma^2),$$

---

[۴] Spatial Lag Model

۲ مدل رگرسیون پروبیت فضایی

فرض کنید شانس حضور یک شخص یا وقوع یک پیشامد براساس مقادیر متغیرهای تبیینی، $x_1, \cdots, x_k$ مورد نظر باشد. اگر فردی از نمونه، ویژگی مورد نظر را داشته باشد متغیر پاسخ y مقدار ۱ و در غیر این‌صورت مقدار ۰ را اختیار کند، آنگاه لازم است شانس آن‌که متغیر y برابر ۱ باشد، تخمین زده شود. فرض کنید ساختار مدل متغیر پنهان در پروبیت فضایی به‌صورت

$$Y^* = \rho W Y^* + X\beta + \epsilon, \;\; \epsilon \sim N(0, \sigma^2 I),$$

است، که به‌صورت

$$Y^* = (I - \rho W)^{-1} X\beta + U, \;\; U = (I - \rho W)^{-1}\epsilon,$$

نیز قابل بیان است. متغیر تصادفی y_i را به‌صورت

$$y_i = \begin{cases} 1 & y_i^* > 0 \\ 0 & y_i^* \leq 0 \end{cases}, \qquad (۲)$$

تعریف نمایید. با توجه به (۲)، داریم

$$\begin{aligned} P(y_i = 1|X) = P(y_i^* > 0|X) &= P(u_i < \frac{[(I - \rho W)^{-1} X\beta]_i}{\sigma_i}) \\ &= \Phi\{\frac{[(I - \rho W)^{-1} X\beta]_i}{\sigma_i}\}. \qquad (۳) \end{aligned}$$

که در آن $\Phi(\cdot)$ تابع توزیع تجمعی نرمال می‌باشد و $U = (u_1, \cdots, u_k)$ دارای توزیع نرمال nمتغیره با میانگین صفر و ماتریس کواریانس $\Sigma_\rho = [(I - \rho W)^T (I - \rho W)]^{-1}$ می‌باشد. به دلیل وابستگی u_iها محاسبه (۳) نیاز به حل انتگرال $n - 1$ بعدی دارد. مدل پروبیت با فرض وابسته بودن خطاها به‌صورت

$$Y = X\beta + U, \;\; U = \lambda W U + \epsilon, \; \epsilon \sim N(0, \sigma^2),$$

با $U = (I - \lambda W)^{-1}\epsilon$ و احتمال کناری

$$P(y_i = 1|x_i) = P(y_i^* > 0|x_i) = P(u_i < \frac{x_i^T\beta}{\sigma_i}) = \Phi(\frac{x_i^T\beta}{\sigma_i}),$$

است، که در آن x_i، i امین سطر ماتریس X، U دارای توزیع نرمال چندمتغیره با میانگین صفر و ماتریس کواریانس $\Sigma_\lambda = [(I-\lambda W)^T(I-\lambda W]^{-1}$ و $\Phi(\cdot)$، تابع توزیع تجمعی i امین نرمال کناری می‌باشد، که برای محاسبه آن نیز نیاز به حل انتگرال $n-1$ بعدی است.

۳ تحلیل بیزی مدل رگرسیون پروبیت فضایی

فرض کنید فرم متغیرهای پنهان به‌صورت

$$Y^* = \rho W_1 Y^* + X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

باشد. متغیر y_i را به‌صورت

$$y_i = \begin{cases} 1 & y_i^* > 0 \\ 0 & y_i^* \leq 0 \end{cases}$$

در نظر بگیرید. تابع چگالی متغیر پنهان Y^* به‌صورت

$$f(Y^*|\rho, \beta, \sigma^2) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n |I-\rho W| exp(\frac{-1}{\sigma^2}\epsilon^T\epsilon) \qquad (۴)$$

است، که در آن $\epsilon = (I-\rho W)Y^* - X\beta$. توزیع توأم پسین به‌صورت

$$\pi(\rho, \beta, \sigma|Y^*) \propto L(Y^*|\rho, \beta, \sigma^2)\pi(\rho, \beta, \sigma) \qquad (۵)$$

است. چنانچه توزیع پیشین توأم به فرم ساده $\pi(\rho, \beta, \sigma) = \frac{1}{\sigma}$ در نظر گرفته شود، توزیع توأم پسین به‌صورت

$$\pi(\rho, \beta, \sigma|Y^*) \propto |I-\rho W| \sigma^{-(n+1)} exp(\frac{-1}{2\sigma^2}\epsilon^T\epsilon) \qquad (۶)$$

خواهد شد که فرم پیچیده‌ای دارد. بنابراین لازم است از الگوریتم‌های MCMC برای برآورد پارامترها استفاده شود. برای به‌کار بردن الگوریتم نمونه‌گیری گیبز، با در نظر گرفتن توزیع پیشین $\pi(\sigma) \sim IG(a, b)$، توزیع شرطی کامل σ به‌صورت

$$\pi(\sigma|Y^*, \rho, \beta) \propto \sigma^{-(\frac{n}{2}+a+1)} exp(\frac{-1}{\sigma^2}(\frac{\epsilon^T\epsilon}{2} + b)) \qquad (۷)$$

حاصل می‌شود، که توزیع معکوس گاما $IG(\frac{n}{۲}+a, \frac{\epsilon^T\epsilon}{۲}+b)$ است و می‌توان از آن نمونه تولید کرد. توجه شود که مشروط بر ρ عبارت $|I-\rho W|$ به‌عنوان یک نسبت ثابت در نظر گرفته شده است. با فرض توزیع پیشین مزدوج نرمال برای β توزیع شرطی کامل آن به‌صورت

$$\pi(\beta|Y^*,\rho,\sigma) \sim N[\widetilde{\beta}, \sigma_\epsilon^۲(X^TC^TCX)^{-۱}] \qquad (۸)$$

خواهد شد، که در آن $C=I_n$ و $\widetilde{\beta}=(X^TX)^{-۱}X^T(I-\rho W)Y^*$. براساس توزیع پیشین یکنواخت برای ρ توزیع شرطی کامل آن به‌صورت

$$\begin{aligned}\pi(\rho|\beta,\sigma,Y^*) &= \frac{\pi(\rho,\beta,\sigma|Y^*)}{\pi(\beta,\sigma|\rho)} \\ &\propto |I-\rho W|\sigma^{-(n+۱)}exp(\frac{-۱}{\sigma^۲}\epsilon^T\epsilon) \qquad (۹)\end{aligned}$$

است، که فرم پیچیده‌ای دارد و برای نمونه‌گیری از آن از الگوریتم متروپلیس - هستینگس استفاده می‌شود. لسج و پس (۲۰۰۹) برای نمونه‌گیری از متغیرهای پنهان با توزیع نرمال چندمتغیره بریده شده به‌صورت

$$\begin{aligned}Y^*|\rho,\beta,\sigma &\sim TMVN((I-\rho W)^{-۱}X\beta, \sigma^۲[(I-\rho W)^T(I-\rho W)]^{-۱}) \\ &= \begin{cases} \text{اگر } y_i=۱ & \text{نرمال بریده شده از چپ در صفر} \\ \text{اگر } y_i=۰ & \text{نرمال بریده شده از راست در صفر} \end{cases} \qquad (۱۰)\end{aligned}$$

استفاده کردند. با داشتن توزیع‌های شرطی کامل الگوریتم زیر قابل اجرا است.

۱) مقادیر اولیه $y_۰^*, \rho_۰, \beta_۰, \sigma_۰$ در نظر گرفته شود.

۲) $\sigma_۱$ از توزیع (۷) تولید شود.

۳) $\beta_۱$ از توزیع (۸) تولید شود.

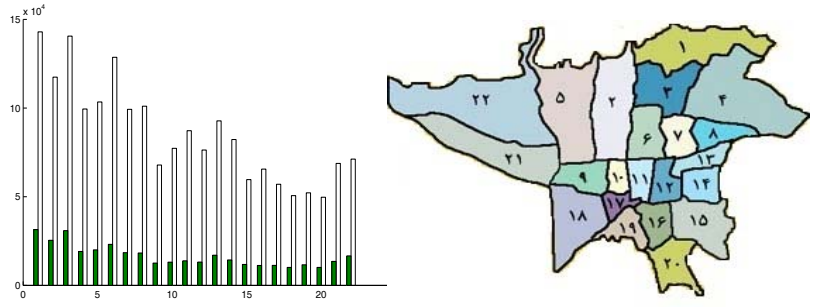
۴) $\rho_۱$ از توزیع (۹) با الگوریتم متروپلیس هستینگس تولید شود.

۵) با استفاده از $\sigma_۱, \beta_۱$ و $\rho_۱$ از توزیع (۱۰) مقدار y^* تولید شود.

حال با تکرار مراحل نمونه‌گیری گیبز و نمونه‌های به‌دست آمده برآورد پارامترها و تحلیل بیزی آن‌ها میسر است.

مثال ۱ : داده‌های این مثال شامل قیمت خرید و فروش و اجارهٔ به علاوه سه درصد ودیعه پرداختی بابت یک متر مربع زیربنای واحدهای مسکونی شهر تهران

است که از طریق بنگاه‌های معاملات ملکی در سامانه اطلاعات مدیریت معاملات املاک و مستغلات کشور ثبت شده‌اند. متغیر پاسخ به این‌صورت انتخاب شده‌اند که اگر متوسط قیمت خرید و فروش هر مترمربع زیربنای مسکونی معامله شده در سال ۱۳۸۹ در مناطق ۲۲ گانه شهر تهران شکل ۱- الف از چارک سوم داده‌ها بیشتر بود مقدار یک و در غیر این‌صورت صفر می‌باشد. متوسط اجاره ماهانه به‌علاوه سه درصد ودیعه پرداختی بابت اجاره یک متر مربع زیربنای مسکونی بنگاه‌های به‌عنوان متغیر تبیینی در نظر گرفته شده است. شکل ۱- ب نمودار میانگین قیمت زیربنا و اجاره‌بها در مناطق در مناطق ۲۲ گانه شهر تهران را نشان می‌دهد. هر یک از مناطق یک ناحیه فضایی را تشکیل می‌دهند. مناطق دارای مرز مشترک به‌عنوان همسایه در نظر گرفته شده و ماتریس وزن فضایی به‌صورت مجاورت استاندارد شده در نظر گرفته شده است. مدل‌های پروبیت فضایی با فرض این که بین متغیر پنهان مدل تأخیرفضایی، خطافضایی و رگرسیون خطی ساده برقرار باشد، به داده‌ها برازش داده شده است.



(ب) (الف)

شکل ۱: الف: مناطق ۲۲ گانه شهر تهران و ب: میانگین قیمت زیربنا (مستطیل تو پر) و اجاره (مستطیل تو خالی) در مناطق ۲۲ گانه شهر تهران

برای تحلیل بیزی از ۵۰۰۰۰ بار تکرار الگوریتم، مرحله داغیدن ۲۵۰۰۰ و طول انتخاب پنجم نمونه‌ای به حجم ۵۰۰۰ حاصل شده است. با در نظر گرفتن

توزیع‌های پیشین ناآگاهی‌بخش برای پارامترهای مدل به‌صورت

$$\beta_0, \beta_1 \sim N_2(0, 10^{12} I_2), \quad \sigma^2 \sim IG(0/001, 0/001), \quad \rho \sim U(-1, 1)$$

برآورد بیزی پارامترها به‌دست آورده شده است. احتمال معنی‌دار بودن آن‌ها مشابه لسج و پس (۲۰۰۹) با استفاده از توزیع مجانبی آماره‌هایی که از نمونه‌های پسین به‌دست می‌آید به همراه ملاک BIC که مولنبرگ (۲۰۰۵) برای مقایسه مدل‌های لوژستیک و پروبیت فضایی به‌کار برد، محاسبه شده است. با فرض آنکه بین متغیرهای پنهان مدل‌های تأخیرفضایی، خطافضایی و رگرسیون خطی ساده برقرار باشد نتایج در جدول ۱ ارائه شده‌اند. همان‌طور که ملاحظه می‌شود مقدار ملاک BIC برای مدل رگرسیون خطی ساده برابر ۷۶۶/۲- و برای مدل تأخیرفضایی برابر ۷۷۰/۶۵ است. در حالی که مدل خطافضایی دارای کمترین مقدار این ملاک، یعنی ۷۶۵/۴- می‌باشد و ضریب اتورگرسیو فضایی آن نیز در سطح ۰/۰۵ معنی‌دار است. بنابراین مدل خطافضایی کاراتر از دو مدل برازانده شدهٔ دیگر است.

جدول ۱: برآورد و احتمال معنی‌داری پارامترهای مدل‌ها

	مدل					
	تأخیرفضایی		خطافضایی		رگرسیون خطی ساده	
پارامتر	برآورد	احتمال	برآورد	احتمال	برآورد	احتمال
β_0	۲۲۳۵/۷۴-	۰/۰۳۶	۷۰۴/۰۱-	۰/۷۴	۱۵۲۲/۲-	۰/۲۸
β_1	۰/۱۹	۰/۰۰۰	۰/۱۸	۰/۰۰۰	۰/۲۲	۰/۰۰۰
ρ	۰/۲۴	۰/۲۸	–	–	–	–
λ	–	–	۰/۴۸	۰/۰۲۱	–	–
BIC	۷۷۰/۶۵-		۷۶۵/۴-		۷۶۶/۲-	

بحث و نتیجه‌گیری

برای مدل‌بندی داده‌های دودویی فضایی مدل رگرسیون پروبیت فضایی معرفی گردید. لحاظ کردن وابستگی فضایی داده‌ها باعث افزایش کارایی و معتبر بودن تحلیل‌ها شد. مزیت تحلیل داده‌های فضایی دودویی با استفاده از این مدل‌ها از تعیین ساختار همبستگی داده‌ها در قالب توابع تغییرنگار یا هم‌تغییرنگار بی‌نیاز می‌سازد. زیرا ساختار همبستگی داده‌ها از طریق مدل اتورگرسیو در تحلیل داده‌ها منظور می‌گردد.

مراجع

Beron, K. J. and Vijverberg, W. P. M., (2004), *Probit in a Spatial Context: A Monte Carlo Analysis,* Berlin: Springer-Verlag.

Cox, D. R. and Snell, E. J. (1989), *Analysis of Binary Data,* 2nd Ed, London: Champan and Hall.

Lesage, J. P. (2000), Bayesian Estimation of Limited Dependent Variable Spatial Autoregressive Models, *Geographical Analysis,* **32**, 19-35.

Lesage, J. P. and Pace R. K. (2009), *Introduction to Spatial Econometrics*, CRC Press Taylor and Francis Group New York.

McMillen, D. P, (1992), Probit with Spatial Autocorrelation, *Journal of Regional Science*, **32**, 335-48.

Molenberghs, G. (2005), *Models for Discrite Longitudinal Data*, Springer Science, New York.

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

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱
مجموعه مقالات، ص ۹۳-۱۰۶

تحلیل مدل‌های گاوسی پنهان فضایی با تقریب لاپلاس آشیانی ترکیبی

زهرا قیومی، کبری قلی‌زاده گزور، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده : در این مقاله تحلیل بیزی تقریبی رده‌ای از مدل‌های رگرسیونی جمعی‌ساختاری تحت عنوان مدل‌های گاوسی پنهان فضایی مورد نظر است. گاهی در تحلیل این‌گونه مدل‌ها توزیع‌های پسینی یا شرطی کامل فرم بسته‌ای ندارند و معمولاً از الگوریتم‌های نمونه‌گیری مونت کارلوی زنجیر مارکوفی استفاده می‌شود. وجود همبستگی در میدان تصادفی پنهان معمولاً موجب کاهش کارایی این الگوریتم‌ها ، افزایش زمان محاسبات و ناهمگرایی الگوریتم می‌شود. برای حل این مشکل روش تقریب لاپلاس آشیانی ترکیبی استفاده می‌شود که در آن روش‌های انتگرال‌گیری عددی و تقریب لاپلاس به طریقی کارا ترکیب می‌شود به طوری که محاسباتی سریع و تقریبی دقیق جایگزین شبیه‌سازی‌های سنگین می‌گردد و ملاک‌های مناسبی نیز برای ارزیابی و مقایسهٔ مدل‌ها ارائه می‌شود.

آدرس الکترونیک مسئول مقاله: زهرا قیومی، z.ghayyomi@modares.ac.ir
کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲H۱۱

واژه‌های کلیدی : تقریب لاپلاس آشیانی ترکیبی، مدل رگرسیونی جمعی‌ساختاری، میدان تصادفی مارکوفی گاوسی، مدل گاوسی پنهان.

۱ مقدمه

تنوع زمینه‌های کاربردی، تحول در حجم و پیچیدگی تحلیل داده‌ها در کنار پیشرفت‌های اساسی در مدل‌های آماری، روش‌های نوین محاسباتی و قابلیت رایانه‌های پیشرفته استفاده از مدل‌های آماری پیچیده‌تر اما دقیق‌تر را به همراه داشته‌اند. مدل‌های رگرسیونی جمعی‌ساختاری[۱] که توسط فهرامیر و تاتز (۲۰۰۱) معرفی شدند ردهٔ گسترده‌ای از مدل‌ها شامل مدل‌های خطی (تعمیم‌یافته)، مدل‌های جمعی (تعمیم‌یافته)، مدل‌های هموار، مدل‌های رگرسیون نیمه‌پارامتری، مدل‌های فضایی و فضایی-زمانی هستند. معمولاً در این مدل‌ها فرض می‌شود توزیع متغیر پاسخ عضوی از خانوادهٔ نمایی است که میانگین آن‌ها با یک پیشگوی جمعی‌ساختاری از طریق یک تابع پیوند مرتبط می‌شود. مدل‌های گاوسی پنهان که در زمینه‌های مختلف کاربرد دارند رده‌ای گسترده از مدل‌های رگرسیون جمعی‌ساختاری هستند و عبارتند از: الف- مدل‌های رگرسیونی، شامل مدل‌های خطی تعمیم‌یافته (دی و همکاران، ۲۰۰۰)، مدل‌های اسپلاین تاوانیده (لانگ و برزگر،۲۰۰۴)، مدل‌های قدم زدن تصادفی (رو و هلد، ۲۰۰۵)، فرایندهای گاوسی (چو و قهرمانی، ۲۰۰۵). ب- مدل‌های پویا (وست و هریسون، ۱۹۹۷)، مدل‌هایی هستند که همبستگی زمانی مشاهدات از طریق یک متغیر تبیینی وارد مدل می‌شود. ج- مدل‌های فضایی یا فضایی-زمانی، مدل‌هایی هستند که همبستگی‌های فضایی یا فضایی-زمانی را از طریق یک متغیر تبیینی وارد مدل می‌کنند (بنرجی و همکاران، ۲۰۰۸).

در این مقاله برای تحلیل بیزی مدل‌های گاوسی پنهان فضایی و یافتن چگالی‌های پسینی کناری علاوه بر الگوریتم‌های مونت کارلوی زنجیر مارکوفی[۲] (MCMC) از

---

۱ Structured Additive Regression Models

۲ Markov Chain Monte Carlo

روش تقریب لاپلاس آشیانی ترکیبی[۳] (INLA) (رو و همکاران، ۲۰۰۹) استفاده می‌شود و علاوه بر ارائهٔ تقریب‌های دقیق از چگالی‌های پسینی مزیت‌های روش INLA نسبت به الگوریتم‌های MCMC مطرح می‌گردد. برای این منظور در بخش ۲ مدل‌های رگرسیونی جمعی‌ساختاری و گاوسی پنهان معرفی می‌شوند. سپس در بخش ۳ میدان تصادفی 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 نحوه کاربست روش INLA در مثالی کاربردی تشریح و دقت نتایج حاصل با نتایج الگوریتم‌های MCMC مقایسه می‌شوند.

۲ مدل‌های آماری

مدل‌های رگرسیونی جمعی‌ساختاری قالبی انعطاف پذیر برای مدل‌بندی اثرات غیر خطی متغیرهای تبیینی شامل مدل‌های خطی تعمیم‌یافته، مدل‌های جمعی تعمیم‌یافته، مدل‌های فضایی و مدل‌های فضایی-زمانی هستند. در این مدل‌ها توزیع متغیر پاسخ y_i متعلق به خانواده نمایی است به طوری که میانگین $\mu_i = E(y_i)$ با یک پیشگوی جمعی‌ساختاری مانند

$$\eta_i = \alpha + \sum_{j=1}^{n_f} \boldsymbol{f}^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i, \qquad i = 1, \ldots, n, \qquad (1)$$

از طریق تابع پیوند $g(\cdot)$ به‌صورت $g(\mu_i) = \eta_i$ مرتبط است، که در آن $\{\boldsymbol{f}^{(j)}(\cdot)\}$ توابعی نامعلوم از مولفه‌های متغیرهای تبیینی $\boldsymbol{u}_i = (u_{1i}, \ldots, u_{n_f i})$، $\{\beta_k\}$ اثرات ثابت خطی[۴] از مولفه‌های متغیرهای تبیینی $\boldsymbol{z}_i = (z_{1i}, \ldots, z_{n_\beta i})$ و ϵ_i عبارت خطا است.

مدل‌های گاوسی پنهان رده‌ای از مدل‌های جمعی‌ساختاری با پیشگوی خطی به صورت (۱) هستند، که در آن برای α، $\{\boldsymbol{f}^{(j)}(\cdot)\}$ و $\{\beta_k\}$ پیشین‌های گاوسی و برای خطای ϵ_i توزیع نرمال با میانگین صفر در نظر گرفته می‌شود. فرض کنید میدان

[۳] Integrated Nested Laplace Approximation

[۴] Fixed Linear Effect

پنهان $\boldsymbol{x} = (\alpha, f^{(1)}, \ldots, f^{(n_f)}, \beta_1, \ldots, \beta_{n_\beta})$ دارای توزیع نرمال با بردار میانگین صفر و ماتریس دقت $\boldsymbol{Q}_{\theta_1}$ باشد، که به پارامتر θ_1 بستگی دارد. با فرض آن که $\mathcal{I}$ یک مشبکه با n_d گره و بردار متغیر پاسخ $\boldsymbol{y} : \{y_i; i \in \mathcal{I}\}$ دارای توزیع $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_2)$ باشد، که درایه‌های آن به شرط $\boldsymbol{x}$ و θ_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 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$ بردار ℓ بعدی ابر پارامترهای مدل است.

۳ میدان تصادفی مارکوفی گاوسی

میدان تصادفی مارکوفی گاوسی [۵] (GMRF) براساس گراف قابل تعریف است. گراف G مجموعه‌ای از راس‌ها است که توسط خانواده‌ای از یال‌ها به هم وصل شده‌اند و به صورت زوج مرتب (V, E) نشان داده می‌شود، که در آن V مجموعه‌ای متناهی و غیر تهی از رئوس و E یال‌های آن است. اگر راس‌ها به صورت $V = \{1, \ldots, n\}$ باشند گراف را نشاندار گویند.

بردار تصادفی $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in R^n$ یک میدان تصادفی مارکوفی گاوسی تحت گراف نشاندار $G = (V, E)$ با میانگین $\boldsymbol{\mu}$ و ماتریس دقت $\boldsymbol{Q}_{n \times n} > \mathbf{0}$ است اگر و تنها اگر تابع چگالی آن به‌صورت

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

باشد، به‌گونه‌ای که درایهٔ (ij)ام ماتریس دقت، یعنی $\boldsymbol{Q}_{ij}$، مخالف صفر است اگر و تنها اگر $\{i, j\} \in E$ باشد.

اگر ماتریس دقت $\boldsymbol{Q}$ نیمه معین مثبت متقارن با مرتبهٔ $n - k > 0$ باشد، آن‌گاه $\boldsymbol{x}$ یک

[۵] Gaussian Markov Random Field

GMRF ناسره از مرتبهٔ $n-k$ با پارامترهای $(\boldsymbol{\mu}, \boldsymbol{Q})$ نامیده می‌شود اگر 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[۶] (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 $\boldsymbol{Q}\boldsymbol{S}_{k-۱} = \mathbf{o}$ است، که در آن $\boldsymbol{S}_{k-۱}$ ماتریس طرح چندجمله‌ای است (رو و هلد، ۲۰۰۵).

این نوع از مدل‌ها 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 دوم $\boldsymbol{x}$ یک مدل قدم زدن تصادفی مرتبهٔ دوم[۷](RW2) نامیده می‌شود هرگاه تفاضل‌های پیشرو مرتبهٔ دوم آن مستقل و دارای توزیع نرمال باشند، یعنی

$$\Delta^۲ x_i = x_i - ۲x_{i+۱} + x_{i+۲} \overset{iid}{\sim} N(۰, \kappa^{-۱}), \quad i = ۱, \cdots, n-۲.$$

۴ تقریب لاپلاس آشیانی ترکیبی

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

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

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

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

[۶] Intrinsic GMRF

[۷] Second-order Random Walk

به‌صورت

$$\begin{aligned}\widetilde{\pi}(x_i|\boldsymbol{y}) &= \int \widetilde{\pi}(x_i|\boldsymbol{y},\boldsymbol{\theta})\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \quad i = ۱,\ldots,n_d,\\ \widetilde{\pi}(\theta_j|\boldsymbol{y}) &= \int \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}, \quad j = ۱,\ldots,\ell\end{aligned}$$

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. توزیع‌های پسینی کناری (۲) و (۳) با محاسبهٔ تقریب‌های $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ و $\widetilde{\pi}(x_i|\boldsymbol{y},\boldsymbol{\theta})$ و استفاده از انتگرال‌های عددی قابل حصول هستند، که در سه گام به شرح زیر محاسبه می‌شوند.

گام اول: با استفاده از بسط تیلور توزیع شرطی کامل

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

حول مد توزیع شرطی کامل $\boldsymbol{x}$ یعنی $\boldsymbol{x}^*(\theta) = (x_۱^*(\boldsymbol{\theta}),\ldots,x_{nd}^*(\boldsymbol{\theta}))$، تقریب گاوسی[۸] آن به صورت

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

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 $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ به صورت

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

[۸] Gaussian Approximation

محاسبه می‌شود.

گام دوم: تقریب لاپلاس چگالی شرطی $\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})$ به صورت

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

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، که از تقریب گاوسی $\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ حاصل می‌شود.

گام سوم: برای محاسبهٔ تقریب (۲) و (۳) دو گام اول و دوم با استفاده از انتگرال‌های عددی ترکیب می‌شوند.

لازم به ذکر است که چگالی‌های پسینی این روش توسط نرم افزار INLA در محیط R قابل محاسبه است، که از پایگاه www.r-inla.org قابل دسترس است.

۵ ملاک ارزیابی مدل‌ها

ملاک اطلاع کیبش[۹](DIC) اندازه‌ای از پیچیدگی و برازش مدل است که برای مقایسهٔ مدل‌های پیچیده استفاده می‌شود (اشپیگل‌هالتر، ۱۹۹۸). ملاک انحراف بیزی بر اساس لگاریتم درستنمایی توزیع پسین به صورت

$$D(X,\theta) = -۲\sum_{i\in\mathcal{I}}\log\{\pi(y_i|x_i,\theta)\} + c$$

تعریف می‌شود که در آن c مقداری ثابت است (دمپستر، ۱۹۷۴).

ملاک اطلاع کیبش به صورت

$$DIC = \overline{D} + P_{_D}$$

قابل تعریف است، که در آن $\overline{D} = E_{\theta|y}(D)$ میانگین پسین انحراف‌ها و $P_{_D}$ تعداد پارامترهای فعال مدل، یعنی تفاضل بین میانگین انحراف‌ها و انحراف میانگین‌ها است.

---

[۹] Deviance Information Criterion

چگالی پیش‌گوی y_i به شرط سایر مشاهدات $\pi(y_i|\boldsymbol{y}_{-i})$ اندازهٔ پیشگویی[۱۰] است، که برای ارزیابی مدل و تشخیص مشاهدات نامتعارف یا دورفتاده به کار می‌رود. لازم به ذکر است که حذف y_i از مجموعه داده‌ها بر چگالی‌های کناری x_i و $\boldsymbol{\theta}$ به صورت

$$\pi(x_i|\boldsymbol{y}_{-i},\boldsymbol{\theta}) \propto \frac{\pi(x_i|\boldsymbol{y})}{\pi(y_i|x_i,\boldsymbol{\theta})}$$
$$\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-i}) \propto \frac{\pi(\boldsymbol{\theta}|\boldsymbol{y})}{\pi(x_i|\boldsymbol{y}_{-i},\boldsymbol{\theta})}$$

تاثیر می‌گذارد. مقدار خیلی کوچک $\pi(y_i|\boldsymbol{y}_{-i})$ بیانگر نامتعارف بودن مشاهدهٔ y_i است.

۶ مثال کاربردی

در این بخش سه مدل فضایی مختلف برای مجموعه دادهٔ زامبیا (کاندالا، ۲۰۰۱) با روش تقریب لاپلاس آشیانی ترکیبی و الگوریتم MCMC تحلیل و مورد مقایسه قرار می‌گیرند. این مجموعه شامل $n_d = ۴۸۴۷$ دادهٔ مربوط به میزان ضعف بدنی کودکان در ۵۷ منطقه از زامبیا است که توسط یونیسف به صورت

$$Z_i = \frac{AI_i - MAI}{\sigma}, \qquad i = ۱,\ldots,n_d$$

اندازه‌گیری و ارائه شده‌اند، که در آن AI اندازه بدن کودکان، MAI میانه و σ انحراف معیار جامعه است. با فرض این که Z_i متغیر تصادفی گاوسی مستقل شرطی با میانگین η_i و دقت τ_z است، تاثیر شش متغیر تبیینی حجم بدن مادر، سن کودک (برحسب ماه)، موقعیت محل زندگی، وضعیت تحصیلی مادر، جنس کودک و نوع محل زندگی بر میزان رنجوری کودکان بررسی می‌شود.

سه مدل مختلف برای پارامتر میانگین η_i در نظر گرفته می‌شود. اولین مدل به صورت

$$\eta_i = \mu + Z_i^{\tau}\beta + f_s(s_i) + f_u(s_i), \qquad i = ۱,\ldots,n_d \qquad (۴)$$

[۱۰] Predictive Measure

است، که در آن $\boldsymbol{\beta} = (\beta_{age}, \beta_{edu۱}, \beta_{edu۲}, \beta_{tpr}, \beta_{sex}, \beta_{bmi})$ بردار اثرات ثابت است. در این مدل اثر هر شش متغیر تبیینی خطی در نظر گرفته شده است. به علاوه اثر تصادفی بین مناطق، $f_u(s_i)$ دارای ساختاری غیرفضایی، مستقل و نرمال با میانگین صفر و دقت τ_u است. اما عبارت $f_s(s_i)$ با ساختاری فضایی تاثیر موقعیت‌ها را در مدل به‌صورت هموار لحاظ می‌کند. به منظور بیان اثر همواری، $f_s(s_i)$ به عنوان یک میدان تصادفی مارکوفی گاوسی ذاتی با دقت τ_s مدل‌بندی می‌شود. برای شناسایی پذیر بودن میانگین μ باید مجموع نموها صفر باشد که قید خطی پنهان نامیده می‌شود (رو و هلد، ۲۰۰۵). میدان گاوسی پنهان و بردار ابر پارامترها برای این مدل به ترتیب به صورت $\boldsymbol{x} = \{\mu, \boldsymbol{\beta}, f_s(.), f_u(.)\}$ و $\boldsymbol{\theta} = \{\tau_z, \tau_u.\tau_s\}$ هستند. برای هر یک از درایه‌های بردار $\boldsymbol{\theta}$ به طور مستقل توزیع پیشینی گامای مبهم در نظر گرفته شده است. برای مشخص کردن کیفیت پیشگویی مدل از ملاک نمره لگاریتمی اعتبارسنجی متقابل[۱۱](log score) (گنتینگ و رفتری، ۲۰۰۷) استفاده می‌شود که هرچه مقدار این ملاک کوچکتر باشد، مدل بهتر است. میانگین پسین، انحراف استاندارد و صدک‌های ۰/۰۲۵ و ۰/۹۷۵ برای هر سه مدل محاسبه و نتایج در جدول ۱ ارائه شده‌اند. مقادیر DIC و log score مربوط به هر سه مدل نیز در جدول ۲ آورده شده است.

کاندالا و کنیب (۲۰۰۱) دلایل محکمی برای غیرخطی بودن اثر متغیر سن بر میزان رنجوری کودکان مطرح کردند. به منظور بررسی این فرض مدل دوم به‌صورت

$$\eta_i = \mu + z_i^T \beta + f_۱(agc_i) + f_s(S_i) + f_u(S_i), \qquad i = ۱, \ldots, n_d \qquad (۵)$$

در نظر گرفته شده است، که در آن $\{f_۱(\cdot)\}$ یک قدم زدن تصادفی ذاتی مرتبه دوم با دقت $\tau_۱$ است. برای شناسایی پذیر بودن μ قید خطی پنهانی برای $f_۱(\cdot)$ منظور می‌شود. میدان پنهان و بردار ابر پارامترها برای این مدل به ترتیب به صورت $\boldsymbol{x} = \{\mu, \beta_k, f_s(\cdot), f_u(\cdot), f_۱(\cdot), \eta_i\}$ و $\theta = \{\tau_z, \tau_u, \tau_s, \tau_۱\}$ است. نمودار اثر سن بر میزان رنجوری کودکان در شکل ۱ بیانگر غیرخطی بودن آن است. همچنین با درنظر گرفتن اثر غیر خطی سن، همان‌طور که در جدول ۲ ملاحظه می‌شود، مقدار DIC

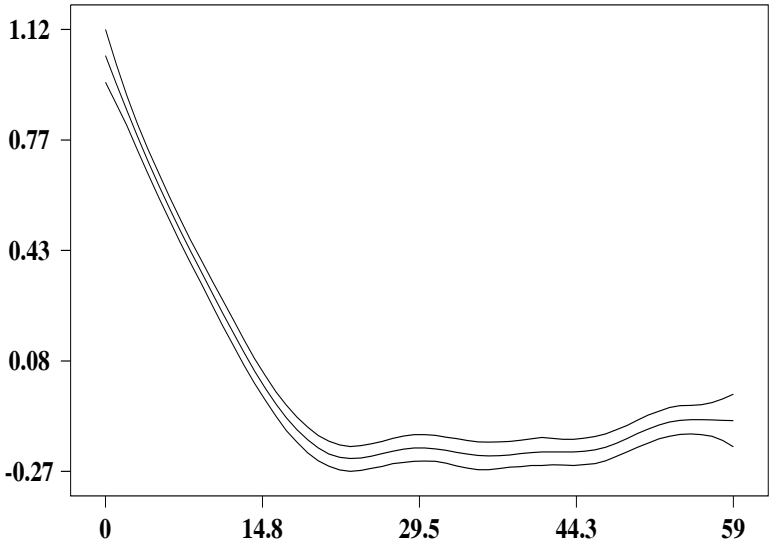
[۱۱] Cross-Validated Logarithmic Score

جدول ۱: برآورد بیزی پارامترهای سه مدل با استفاده از روش INLA

مدل	متغیرهای کمکی	میانگین	انحراف استاندارد	صدک ۰/۰۲۵	صدک ۰/۹۷۵
	μ	-۰/۰۲۴	۰/۱۰۰	-۰/۲۲۱	۰/۱۷۲
	β_{age}	-۰/۰۱۵	۰/۰۰۰	-۰/۰۱	-۰/۰۱۳
	β_{edu1}	-۰/۰۶۲	۰/۰۲۶	-۰/۱۱۵	-۰/۰۰۹
۱	β_{edu2}	۰/۲۲۹	۰/۰۴۷	۰/۱۳۷	۰/۳۲۲
	β_{tpr}	۰/۱۰۱	۰/۰۲۲	۰/۰۵	۰/۱۴۵
	β_{sex}	-۰/۰۵	۰/۰۱۳	-۰/۰۸	-۰/۰۳۳
	β_{bmi}	۰/۰۲۳	۰/۰۰۴	۰/۰۱	۰/۰۳۱
	μ	-۰/۴۳۰	۰/۰۹۶	-۰/۶۱۹	-۰/۲۴۰
	β_{edu1}	-۰/۰۶۱	۰/۰۲۶	-۰/۱۱۲	-۰/۰۱۰
۲	β_{edu2}	۰/۲۳۰	۰/۰۴۵	۰/۱۴۸	۰/۳۲۷
	β_{tpr}	۰/۰۹۰	۰/۰۲۱	۰/۰۴۸	۰/۱۳۴
	β_{sex}	-۰/۰۵۰	۰/۰۱۲	-۰/۰۸۳	-۰/۰۳۳
	β_{bmi}	۰/۰۲۱	۰/۰۰۴	۰/۰۱۳	۰/۰۲۹
	μ	-۰/۳۷۲	۰/۰۹۶	-۰/۵۶۱	-۰/۱۸۳
	β_{edu1}	-۰/۰۶۱	۰/۰۲۶	-۰/۱۱۲	-۰/۰۱۰
۳	β_{edu2}	۰/۲۳۳	۰/۰۴۵	۰/۱۴۴	۰/۳۲۳
	β_{tpr}	۰/۱۰۰	۰/۰۲۳	۰/۰۵۴	۰/۱۴۶
	β_{sex}	-۰/۰۵۸	۰/۰۱۲	-۰/۰۸۴	-۰/۰۳۳

جدول ۲: ملاک‌های ارزیابی سه مدل برای داده‌های زامبیا

	مدل		
	۱	۲	۳
$\overline{D}$	۱۳۰۳۱/۷	۱۲۶۸۵/۷	۱۲۶۷۸/۴
انحرافات از میانگین	۱۲۹۹۲/۴	۱۲۶۳۹/۹	۱۲۶۲۰/۲
تعدادپارامترهای موثر	۳۹/۲	۴۵/۸	۵۸/۲
DIC	۱۳۰۷۱/۱	۱۲۷۳۱/۶	۱۲۷۳۶/۷
log score	۱/۳۴۸۴	۱/۳۱۳۴	۱/۳۱۳۹



شکل ۱: برآورد میانگین و صدک‌های ۰/۰۲۵ و ۰/۹۷۵ سن در مدل ۲

برای این مدل کاهش یافته است. اثرات فضایی برآورد شده در این مدل مشابه با مدل (۴) است.

یک فرض پیشنهادی به منظور توضیح تغییر پذیری اثر فضایی می‌تواند اثر یک متغیر تبیینی مانند حجم بدن مادر (bmi) باشد که برای هر منطقه ضریبی متفاوت دارد که با فرض همواری اثر فضایی همه مناطق و غیر خطی بودن اثر سن بر میزان رنجوری کودکان مدل سوم به‌صورت

$$\eta_i = \mu + z_i^T \beta + f_1(age_i) + bmi_i f_2(s_i), \qquad i = 1, \ldots, n_d \qquad (۶)$$

در نظر گرفته شده است. متغیر تبینی bmi به عنوان وزن برای $f_2(\cdot)$ در نظر گرفته شده است. با مقایسه مقادیر log score و DIC سه مدل در جدول ۲ ملاحظه می‌شود مقادیر این دو ملاک بری مدل دوم از دو مدل دیگر کمتر است. به همین دلیل این مدل به عنوان مدل برتر انتخاب می‌شود. همچنین با مقایسهٔ ملاک‌های ارزیابی دو مدل دیگر، مدل (۶) بر مدل (۴) ارجحیت دارد. نتایج حاصل از روش تقریب لاپلاس آشیانی ترکیبی و الگوریتم‌های MCMC مدل (۵) در جدول ۳ ارائه شده

است. همان طور که ملاحظه می‌شود اختلاف ناچیزی بین نتایج حاصل از این دو روش وجود دارد، اما زمان محاسبات در رایانه (تحت سیستم عامل ۶۴ بیت با حافظهٔ ۴ گیگابایت) با روش INLA، ۱۳ ثانیه و با الگوریتم MCMC یک دقیقه و ۵۷ ثانیه به طول انجامید، یعنی سرعت محاسبات با INLA حدود ده برابر افزایش یافت.

جدول ۳: برآورد پارامترهای مدل ۲ با روش‌های MCMC و INLA

	متغیرهای کمکی	میانگین	انحراف استاندارد	صدک ۰/۰۲۵	صدک ۰/۹۷۵
INLA	β_0	-۰/۴۳۰	۰/۰۹۶	-۰/۶۱۹	-۰/۲۴۰
	β_{edu1}	-۰/۰۶۱	۰/۰۲۶	-۰/۱۱۲	-۰/۰۱۰
	β_{edu2}	۰/۲۳۰	۰/۰۴۵	۰/۱۴۸	۰/۳۲۷
	β_{tpr}	۰/۰۹۰	۰/۰۲۱	۰/۰۴۸	۰/۱۳۴
	β_{sex}	-۰/۰۵ ۰	۰/۰۱۲	-۰/۰۸۳	-۰/۰۳۳
	β_{bmi}	۰/۰۲۱	۰/۰۰۴	۰/۰۱۳	۰/۰۲۹
MCMC	β_0	-۰/۴۳۰	۰/۰۹۶	-۰/۶۱۸	-۰/۲۴۰
	β_{edu1}	-۰/۰۶۱	۰/۰۲۵	-۰/۱۱۲	-۰/۰۱۰
	β_{edu2}	۰/۲۴۰	۰/۰۴۶	۰/۱۴۶	۰/۳۳۲
	β_{tpr}	۰/۰۹۰	۰/۰۲۱	۰/۰۴۸	۰/۱۳۳
	β_{sex}	-۰/۰۶۰	۰/۰۱۳	-۰/۰۸۲	-۰/۰۳۱
	β_{bmi}	۰/۰۲۱	۰/۰۰۴	۰/۰۱۳	۰/۰۲۹

بحث و نتیجه‌گیری

همان طور که در مثال تحلیل داده‌های میزان رنجوری کودکان در زامبیا ملاحظه شد نتایج حاصل از روش INLA با نتایج حاصل از نرم افزاری که مبنای محاسبات آن الگوریتم‌های MCMC است اختلاف بسیار کمی داشت اما سرعت محاسبات با INLA به مراتب بیشتر است. در دنیای امروز که افزایش سرعت محاسبات به همراه دقت کافی خواست همهٔ محققان در زمینه‌های گوناگون است معرفی و استفاده از این روش می‌تواند مورد توجه قرار گیرد.

مراجع

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H., (2008), Gaussian Predictive Process Models for Large Spatial Data Sets. *Jornal of Royal Statistical Society* B, **70**, 825-848.

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

Chu, W. and Ghahramani, Z., (2005), Gaussian processes for Ordinal Regression. *Jornal of Machine Learning Research*, **6**, 1019-1041.

Dempster, A. P., (1974), The Direct Use of Likelihood for Significance Testing, *Proceedings of Conference on Foundational Questions in Statistical Inference*, Department of Theoretical Statistics: University of Aarhus, 335-352.

Dey, D. K., Ghosh, S. K. and Mallick, B. K., (2000), Generalized Linear Models: A Bayesian Perspective. Boca Raton: Chapman & Hall, London.

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

Gneiting, T., Raftery, A., (2007), Strictly Proper Scoring Rules, prediction and estimation. *jornal of American Statistical Association*, Serries B, **102**, 359-378.

Kndala, N., B., Lang, S., Klasen, S. and Fahrmeir, L., (2001), Semiparametric Analysis of the Socio-Demographic and Spatial Determinants of

Undernutrition in Two African Countries. Research in Official Statistics, **1**, 81100.

Kneib, T., Lang, S. and Brezger, A., (2004), Bayesian Semiparametric Regression Based on MCMC Techniques.

Lang, S. and Brezger, A., (2004), Bayesian P-splines. *Jornal of Computationl and Graphical Statistcs*, **13**, 183-212.

Martino, S., Rue, H., (2010), Implementing Approximate Bayesian Inference Using Integrated Nested Laplace Approximation: a manual for the INLA program.

Rue, H., 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., Chopin, N., (2009), Approximation Baysian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations, *Jornal of the Royal Statistical Society*, **71**, pp. 319-392.

Spiegelhalter, D., Best, N. and Carlin, B., (1998), Baysian Deviance the Effective Number of Parameters and the Comparison of Arbitrarily Complex Models.

West, M. and Harrison, J., (1997), Bayesian Forecasting and Dynamic Models, 2nd edn. New York: Springer.

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱
مجموعه مقالات، ص ۱۰۷-۱۱۶

مدل اتوچندجمله‌ای برای تحلیل داده‌های مشبکه‌ای فضایی چند متغیره

امیر کاوسی[۱]، محمدرضا مشکانی[۲]، محسن محمدزاده[۳]، افشین فلاح[۴]

[۱] *گروه علوم پایه، دانشگاه علوم پزشکی شهید بهشتی*

[۲] *گروه آمار، دانشگاه شهید بهشتی*

[۳] *گروه آمار، دانشگاه تربیت مدرس*

[۴] *گروه آمار، دانشگاه بین‌المللی امام‌خمینی*

چکیده : داده‌های مشبکه‌ای نوعی از داده‌های فضایی هستند که در بسیاری از فعالیت‌ها تحقق پیدا می‌کنند. برای تحلیل داده‌های فضایی مشبکه‌ای که مستلزم مدل‌بندی احتمالی آنها است، در حالت یک و چند متغیره پیوسته مدل اتو گاوسی مورد مطالعه قرار گرفته است. اما در حالت گسسته تنها برای مسائل تک متغیره مدل‌های اتو دوجمله‌ای، اتو پواسون و اتو دوجمله‌ای منفی مطرح شده‌اند. در این مقاله، مدل اتو چندجمله‌ای به عنوان تعمیم مدل اتو دوجمله‌ای برای تحلیل داده‌های فضایی مشبکه‌ای چند متغیره گسسته ارائه و برآورد پارامترهای آن مورد بررسی قرار گرفته است. در پایان در یک مطالعه شبیه‌سازی دقت این مدل با مدل چند جمله‌ای کلاسیک مورد ارزیابی قرار گرفته است.

آدرس الکترونیک مسئول مقاله: امیر کاوسی، kavousi@sbmu.ac.ir

کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲H۱۱

واژه‌های کلیدی : داده‌های فضایی مشبکه‌ای، اتو دوجمله‌ای، اتو چندجمله‌ای

۱ مقدمه

در آمار فضایی[۱] با یک میدان تصادفی $\{Z(s) : s \in D \subset R^{d\geq 1}\}$ مواجه هستیم، که مشاهدات تحقق‌های این میدان هستند. اگر D زیر مجموعه‌ای شمارش‌پذیر (معمولاً محدود) از R^d، $d \geq 1$ باشد داده‌ها را مشبکه‌ای[۲] نامند، که در آن D ممکن است به طور منظم[۳] یا نامنظم[۴] شبکه‌بندی شده باشد. در این صورت میدان تصادفی مشبکه‌ای را می‌توان به صورت $\{Z(A_i) : A_i \in \{A_1, \cdots, A_n\}\}$ نمایش داد، که در آن $\{A_1, \cdots, A_n\}$ یک افراز از مشبکه D هستند، یعنی $\bigcup_{i=1}^{n} A_i = D$ و $A_i \bigcap A_j = 0, i \neq j$. برای درک بهتر، فرض کنید D مساحت کشور ایران و $A_1, \cdots A_{31}$ استان‌های کشور باشند. در این صورت $Z(A_1), \cdots, Z(A_{31})$ را می‌توان تعداد حوادث رانندگی، تعداد اتباع خارجی غیر مجاز ساکن یا نرخ رشد اقتصادی هر استان در نظر گرفت. در تحلیل داده‌های مشبکه‌ای عمدتاً هدف مدل‌بندی مشاهدات است. همانطور که در سری‌های زمانی هر مقدار جدید در آینده بر اساس رابطه آن با مشاهدات گذشته و نزدیک پیشگویی می‌شود، برای میدان تصادفی فضایی مشبکه‌ای نیز هر موقعیت جدید از ناحیه بر اساس داده‌های همبسته با آن یعنی مشاهدات واقع در نواحی همسایه نزدیک، مدل‌بندی و بعضاً پیشگویی می‌شود. تعیین ساختار همبستگی فضایی داده‌ها اولین گام در تحلیل داده‌های مشبکه‌ای است، که در آن همسایگی[۵] نقش مهمی را ایفا می‌کند، زیرا بیشترین اطلاعات در مورد موقعیت دلخواه $s_0 \in D$ توسط موقعیت‌های نزدیک و همسایه ارائه می‌شود. بنا براین یک مدل آماری مناسب برای تحلیل داده‌های مشبکه‌ای می‌تواند میدان تصادفی مارکوفی باشد. برای داده‌های گسسته رده‌ای از مدل‌ها مانند اتولوژستیک، اتودوجمله‌ای، اتوپواسن و برای داده‌های

---

۱ Spatial statistics

۲ Lattice Data

۳ Grid

۴ Polygon

۵ Neighborhood

پیوسته مدل اتوگاوسی معرفی شده‌اند (کرسی، ۱۹۹۳). کاربرد این مدل‌ها بیشتر در مدل‌بندی میزان بروز ناحیه‌ای بیماری‌های خاص در پزشکی (کاوسی و همکاران ۲۰۰۷، لاوسون ۲۰۰۱)، مکان‌یابی مناسب برای ساخت و ساز مسکن در شهرسازی، بازاریابی و تعیین مکان‌های مناسب برای ایجاد فروشگاه‌های زنجیره‌ای و غیره است. در این مقاله مدل اتو چندجمله‌ای به عنوان تعمیم مدل اتودوجمله‌ای برای تحلیل داده‌های فضایی مشبکه‌ای چند متغیره گسسته توسعه داده شده و برآورد پارامترهای آن‌ها ارائه شده است. در پایان در یک مطالعه شبیه‌سازی دقت این مدل با مدل چندجمله‌ای کلاسیک مورد ارزیابی قرار گرفته است.

در بخش ۲ مدل اتودوجمله‌ای معرفی می‌شود. تعمیم آن تحت عنوان مدل اتوچندجمله‌ای در بخش ۳ ارائه می‌شود. بخش ۴ به شبیه‌سازی یک مثال و مقایسه مدل چندجمله‌ای کلاسیک و اتوچندجمله‌ای پرداخته می‌شود و بخش پایانی به بحث و نتیجه‌گیری اختصاص داده شده است.

۲ مدل اتو دوجمله‌ای

مدل اتودوجمله‌ای توسط کرسی (۱۹۹۳) به صورت

$$\begin{aligned} p(Z(s_i)|Z_{-i}) &= \binom{Z(s_i)}{n_i} {p_i}^{Z(s_i)} (1-p_i)^{n_i - Z(s_i)} \\ & Z(s_i) = 0, 1, \cdots, n_i \, ; i = 1, \cdots, n, \end{aligned} \qquad (1)$$

تعریف شده است، که فرم نمایش نمایی آن به صورت

$$p(Z(s_i)|Z_{-i}) = exp\{Z(s_i) Logit(p_i) + \ln \binom{Z(s_i)}{n_i} + n_i \ln(1-p_i)\},$$

است، که درآن $p_i = p_i(Z_{-i})$، $A_i(Z_{-i}) = \text{Logit}(p_i) = \ln \frac{p_i}{1-p_i}$ و $B_i(Z(s_i)) = Z(s_i)$. بنابراین معادله

$$Logit(p_i) = \alpha_i + \sum_{j=1}^{n} \theta_{ij} Z(s_j); \; i = 1, \cdots, n$$

که همان معادله اتو لوژستیک است، حاصل می‌شود. با حل این معادله مقدار p_i به صورت

$$p_i = \frac{e^{\alpha_i + \sum_{j=1}^{n} \theta_{ij} Z(s_j)}}{1 + e^{\alpha_i + \sum_{j=1}^{n} \theta_{ij} Z(s_j)}};\ i = 1, \cdots, n \qquad (2)$$

به دست می‌آید، که در آن θ_{ij}ها همبستگی فضایی و α_i تاثیر موقعیت iام است و می‌تواند به عنوان روند مدل یا تاثیر متغیرهای تبیینی در نظر گرفته شود. سایر مدل‌ها یعنی مدل‌های اتولوژستیک، اتو پواسن و مدل اتوگاوسی را می‌توان در کرسی (۱۹۹۳) دید.

۳ مدل اتو چندجمله‌ای

برای تعمیم مدل اتوچندجمله‌ای، فرض کنید در موقعیت iام بردار تصادفی $Z_{-i} \equiv \{\underline{Z}(s_j)\ ;\ j \neq i\} \equiv N_i$ به شرط $\underline{Z}(s_i) = (Z_1(s_i), \cdots, Z_k(s_i))'$ دارای توزیع چندجمله‌ای

$$P(\underline{Z}(s_i)|Z_{-i}) = \binom{n_i}{Z_1(s_i), \cdots, Z_k(s_i)} p_{i1}^{Z_1(s_i)} \cdots p_{ik}^{Z_k(s_i)}\ ;\ i = 1, \cdots, n \quad (3)$$

باشد، که در آن $t = 1, \ldots, k$، $p_{it} \equiv p_{it}(Z_{-i})$ توابعی از موقعیت‌های مجاور هستند. در این صورت این مدل را اتوچندجمله‌ای نامیده و برای یافتن ساختار مناسب احتمال‌های چندجمله‌ای $\underline{p}_i = (p_{i1}, \cdots, p_{ik})$ بر اساس اطلاعات موقعیت‌های همسایه و متغیرهای تبیینی، ابتدا از توزیع چند متغیره برنولی استفاده می‌شود. سپس با استفاده از آماره‌های بسنده توزیع چندجمله‌ای تعیین می‌شود. بنابراین بردار k متغیره برنولی را در مکان iام برای نمونه jام به صورت $(D_{ij1}, \cdots, D_{ijk}) : j = 1, \cdots, n_i; i = 1, \cdots, n$ در نظر بگیرید، که در آن اگر jامین نمونه دارای صفت mام در مکان iام باشد،مقدار D_{ijm} برابر ۱ و در غیر این صورت صفر است.به علاوه رابطه $\sum_{m=1}^{k} D_{ijm} = 1$ برقرار است. برای این بردار تصادفی در هر موقعیت $i = 1, \cdots, n$ به طور مستقل n_i بار نمونه‌گیری نموده و فرض می‌شود، نمونه در هر موقعیت مستقل ازموقعیت دیگر است. با این شرایط در

هر موقعیت از مجموع n_i بردار برنولی مستقل یک k جمله‌ای حاصل می‌شود. برای مدل‌بندی ساختار توزیع برنولی چند متغیره در موقعیت i، D_{ijm} را متغیر mام در موقعیت iام برای نمونه jام تعریف می‌شود، که مقدار مشاهده شده آن d_{ijm} برابر صفر یا یک و $j = ۱, \cdots, n_i$ و $i = ۱, \cdots, n$ است. با مبنا قرار دادن متغیر اول، $k - ۱$ لوجیت توسط $U_{im}, m = ۲, \cdots, k,$ همانند مدل اتو دوجمله‌ای به صورت تابعی از اطلاعات مکان‌های همسایه نزدیک به صورت

$$\begin{aligned}
\ln\left(\frac{P(D_{ij۲}=۱|N_i)}{P(D_{ij۱}=۱|N_i)}\right) &= \alpha_{i۲} + \sum_{t\in N_i}\theta_{t۲}Z_۲(s_t) \equiv U_{i۲} \\
\ln\left(\frac{P(D_{ij۳}=۱|N_i)}{P(D_{ij۱}=۱|N_i)}\right) &= \alpha_{i۳} + \sum_{t\in N_i}\theta_{t۳}Z_۳(s_t) \equiv U_{i۳} \\
&\vdots \\
\ln\left(\frac{P(D_{ijk}=۱|N_i)}{P(D_{ij۱}=۱|N_i)}\right) &= \alpha_{ik} + \sum_{t\in N_i}\theta_{tk}Z_k(s_t) \equiv U_{ik} \qquad (۴)
\end{aligned}$$

ساخته می‌شوند، که در آن، N_i همسایه‌های موقعیت iام، $m = ۱, \cdots, k$، $Z_m(s_t) = \sum_{j=۱}^{n_t} d_{ijm}$، θ_{tm} پارامتر ساختار فضایی و α_{im} اثر موقعیت را نشان می‌دهد، که می‌تواند به صورت تابعی از متغیرهای تبیینی نوشته شود. با حل معادلات (۴) نسبت به $P(D_{ijm}=۱|N_i)$، $m = ۱, \cdots, k$ با رعایت شرط $\sum_{m=۱}^{k} P(D_{ijm}=۱|N_i) = ۱$، خواهیم داشت

$$\begin{aligned}
P(D_{ij۱}=۱|N_i) &= \frac{۱}{۱+\sum_{m=۲}^{k} e^{U_{im}}} = p_{i۱} \\
P(D_{ij۲}=۱|N_i) &= \frac{e^{U_{i۲}}}{۱+\sum_{m=۲}^{k} e^{U_{im}}} = p_{i۲} \\
&\vdots \\
P(D_{ijk}=۱|N_i) &= \frac{e^{U_{ik}}}{۱+\sum_{m=۲}^{k} e^{U_{im}}} = p_{ik} \qquad (۵)
\end{aligned}$$

در معادلات (۵) باید تعداد $n \times k + \sum_{i=۱}^{n} k \times |N_i|$ پارامتر به صورت $t \in N_i$، $m = ۱, \cdots, k$، $i = ۱, \cdots, n$، θ_{tm}، α_{im} برآورد شوند، که در آن $|N_i|$ تعداد همسایه‌های مکان i است. توزیع توام چندجمله‌ای در موقعیت iام برای نمونه jام به

صورت

$$P(D_{ij۱} = d_{ij۱}, \cdots, D_{ijk} = d_{ijk}|N_i) \propto p_{i۱}^{d_{ij۱}} \cdots p_{ik}^{d_{ijk}}$$

است، که با قرار دادن $D_{ij.} = (D_{ij۱}, \cdots, D_{ijk})$ توزیع توام $(D_{i۱.}, \cdots, D_{in_i.})$ با توجه به استقلال نمونه‌ها در هر موقعیت به صورت

$$P(D_{i۱.}, \cdots, D_{in_i.}|N_i) \propto \prod_{j=۱}^{n_i} p_{i۱}^{d_{ij۱}} \cdots p_{ik}^{d_{ijk}}; \quad i = ۱, \cdots, n$$

حاصل می‌شود. برای تشکیل تابع درستنمایی نیاز به توزیع توام $(D_{۱..}, \cdots, D_{n..})$ است، که در آن $D_{i..} = (D_{i۱.}, \cdots, D_{in_i.})$. چون توزیع توام فرم بسته‌ای ندارد، از تابع شبه درستنمایی[۶] یعنی حاصل‌ضرب توابع چگالی کناری شرطی به صورت

$$\begin{aligned} L &= \prod_{i=۱}^{n} P(D_{i..}|N_i) \\ &\propto \prod_{i=۱}^{n} \prod_{j=۱}^{n_i} p_{i۱}^{d_{ij۱}} \cdots p_{ik}^{d_{ijk}} \end{aligned}$$

استفاده می‌شود، که لگاریتم آن پس از حذف مقدار ثابت به صورت

$$\ell(\theta) = \sum_{i=۱}^{n} \sum_{j=۱}^{n_i} [d_{ij۱} \ln p_{i۱} + \cdots + d_{ijk} \ln p_{ik}] \qquad (۶)$$

است. با جایگذاری (۴) و (۵) در رابطه (۶) لگاریتم تابع درستنمایی به صورت

$$\begin{aligned} \ell(\theta) &= \sum_{i=۱}^{n} [\sum_{m=۲}^{k} Z_m(s_i)(\alpha_{im} + \sum_{t \in N_i} \theta_{tm} Z_m(s_t)) \\ &- n_i \ln(۱ + \sum_{m=۲}^{k} e^{\alpha_{im} + \sum_{t \in N_i} \theta_{tm} Z_m(s_t)})] \end{aligned} \qquad (۷)$$

حاصل می‌شود. تعداد $q = n \times k + \sum_{i=۱}^{n} k \times |N_i|$ پارامتر برای برآورد وجود دارد، که حتی برای n و k کوچک نیز بسیار زیاد است. بنابراین در عمل باید به طریقی

[۶] Pseudo-Likelihood

تعداد آن‌ها را کاهش داد، تا برآوردهایی کارا حاصل شوند. برای این منظور فرض کنید

$$\alpha_{im} = \beta_m X_i, \;\; \theta_{im} = \gamma_m w_{tm}; \;\; m = ۲, \cdots, k; \; i = ۱, \cdots, n; \; t \in N_i \qquad (۸)$$

که در آن متغیر X_i به عنوان متغیر تبیینی در هر موقعیت و w_{tm} مقداری معلوم به صورت $w_{ij} = \frac{c_{ij}}{c_{i+}}$ است، به طوری که اگر i و j همسایه باشند، $c_{ij} = ۱$ و در غیر این صورت $c_{ij} = ۰$ است. همچنین عدد c_{i+} تعداد همسایه‌های مکان i و $c_{ii} = ۰$ است (کرسی، ۱۹۹۳). بنابراین تعداد پارامترها به $۲(k-۱)$ پارامتر $\underline{\theta} = (\gamma_۲, \cdots, \gamma_k, \beta_۲, \cdots, \beta_k)$ کاهش می‌یابد. برای برآورد پارامترها به روش ماکسیمم درستنمایی نیاز به مشتقات جزئی (۷) نسبت به β_m و γ_m به صورت

$$\begin{aligned} \frac{\partial \ell(\underline{\theta})}{\partial \beta_m} &= \sum_{i=۱}^{n} [Z_m(s_i) x_i \beta_m - \frac{n_i x_i e^{\beta_m x_i + \gamma_m T_m}}{E_i}] \;\; ; \;\; m = ۲, \cdots, k \\ \frac{\partial \ell(\underline{\theta})}{\partial \gamma_m} &= \sum_{i=۱}^{n} [Z_m(s_i) T_m - \frac{n_i T_m e^{\beta_m x_i + \gamma_m T_m}}{E_i}] \;\; ; \;\; m = ۲, \cdots, k \qquad (۹) \end{aligned}$$

است، که در آن $T_m = \sum_{j \in N_i} w_{tm} Z_m(s_i)$ و $E_i = ۱ + \sum_{m=۲}^{k} e^{\beta_m x_i + \gamma_m T_m}$. همان‌طور که ملاحظه می‌شود، سیستم معادلات (۹) به صورت تحلیلی قابل حل نیست و لازم است پاسخ‌ها به روش‌های عددی محاسبه شوند. همچنین چون دقت برآوردگرها به صورت تحلیلی دست نیافتنی است، از روش‌هایی مانند دلتا[۷] یا خودگردانی[۸] برای تقریب آنها استفاده می‌شود. جزئیات استنباط آماری پارامترها و کاربست آن‌هادر یک مثال واقعی توسط کاوسی و همکاران (۲۰۱۱) ارائه شده است.

۴ شبیه‌سازی

در این بخش مدل اتوچندجمله‌ای در یک مطالعه شبیه‌سازی مورد ارزیابی قرار می‌گیرد. برای این منظور یک مشبکه منظم با ابعاد ۱۲ = ۴ × ۳ با موقعیت‌های $s_۱, \cdots, s_{۱۲}$ در نظر گرفته

[۷] Delta Method

[۸] Bootstrap

جدول ۱: MSE برآورد پارامترها در مدل سه‌جمله‌ای

مدل	ایستگاه	$MSE(\hat{p}_{i3})$	$MSE(\hat{p}_{i2})$	$MSE(\hat{p}_{i1})$
	s_1	۰/۰۱۰۳	۰/۰۱۱۱	۰/۰۰۸۵
	s_2	۰/۰۱۳۴	۰/۰۱۰۹	۰/۰۱۲۴
	s_3	۰/۰۰۸۸	۰/۰۰۹۰	۰/۰۰۸۹
	s_4	۰/۰۱۱۱	۰/۰۱۳۴	۰/۰۰۹۷
	s_5	۰/۰۱۲۶	۰/۰۰۹۰	۰/۰۱۱۵
کلاسیک	s_6	۰/۰۱۱۸	۰/۰۱۰۲	۰/۰۱۰۹
	s_7	۰/۰۱۱۷	۰/۰۱۰۱	۰/۰۱۰۵
	s_8	۰/۰۰۸۵	۰/۰۱۲۲	۰/۰۱۲۰
	s_9	۰/۰۱۲۹	۰/۰۱۱۸	۰/۰۰۹۴
	s_{10}	۰/۰۱۳۹	۰/۰۱۰۰	۰/۰۰۸۳
	s_{11}	۰/۰۱۶۴	۰/۰۴۳۱	۰/۰۰۸۶
	s_{12}	۰/۰۱۲۴	۰/۰۱۱۲	۰/۰۱۱۲
	s_1	۰/۰۰۲۲	۰/۰۰۱۲	۰/۰۰۰۹
	s_2	۰/۰۱۱۲	۰/۰۱۰۰	۰/۰۰۵۹
	s_3	۰/۰۰۲۴	۰/۰۰۱۲	۰/۰۰۱۰
	s_4	۰/۱۱۱۴	۰/۰۳۹۱	۰/۰۱۹۷
	s_5	۰/۰۰۰۷	۰/۰۰۰۷	۰/۰۰۰۸
فضایی	s_6	۰/۰۰۳۳	۰/۰۰۱۱	۰/۰۰۱۲
	s_7	۰/۰۰۵۱	۰/۰۰۱۶	۰/۰۰۱۶
	s_8	۰/۰۰۲۲	۰/۰۰۰۸	۰/۰۰۱۲
	s_9	۰/۰۰۲۰	۰/۰۰۱۲	۰/۰۰۱۰
	s_{10}	۰/۰۰۳۹	۰/۰۰۱۵	۰/۰۰۱۳
	s_{11}	۰/۰۱۵۰	۰/۰۳۲۶	۰/۰۰۳۳
	s_{12}	۰/۰۴۴۵	۰/۰۱۶۰	۰/۰۸۱۶

و بردار سه‌جمله‌ای $(Z_1(s_i), Z_2(s_i), Z_3(s_i))$ $i = 1, \cdots, 12$ با استفاده از مدل ۳ با پارامترهای $\beta_2 = 0/002$، $\beta_3 = 0/0014$، $\gamma_2 = 0/0995$، $\gamma_3 = 0/0259$، $p_{i1} = 0/26$، $p_{i2} = 0/32$، $p_{i1} = 0/42$ و متغیر تبیینی $X = (250, 458, 259, 663, 220, 273, 298, 256, 248, 257, 385, 498)$ شبیه‌سازی و پارامترهای p_{i1}، p_{i2} و p_{i3} در دو مدل اتو سه‌جمله‌ای (فضایی) و سه‌جمله‌ای (کلاسیک) برآورد شده‌اند. این عمل را ۱۰۰۰۰ بار تکرار نموده، میانگین توان‌های دوم خطای (MSE) هر یک از برآوردها را محاسبه و نتایج در جدول ۱ درج شده‌اند. مقایسه MSE برآورد پارامتر در جداول مذکور نشان‌گر آن است که دقت برآورد پارامترهای مدل اتوسه‌جمله‌ای معمولاً بیشتر از دقت برآورد پارامترهای مدل سه‌جمله‌ای است، به غیر از دو موقعیت s_4 و s_{12} که به دلیل نزدیکی آن‌ها به کناره‌های مشبکه نتایجی متفاوت حاصل شده است.

۱.۴ بحث و نتیجه‌گیری

نتایج حاصل از شبیه‌سازی در حالت کلی نشان‌گر دقت بیشتر مدل اتو چندجمله‌ای نسبت به چند جمله‌ای کلاسیک است. اما همان طور که نتایج شبیه‌سازی نشان می‌دهد، دقت مدل اتو چندجمله‌ای در موقعیت‌های کناری مشبکه پایین‌تر از مدل چند جمله‌ای کلاسیک است، که این یکی از مشکلات مرسوم اغلب تحلیل‌های فضایی است و نیاز به مطالعه و بررسی جداگانه دارد. روش فضایی از اطلاعات موقعیت‌های همسایه در پیش‌گویی یک موقعیت استفاده می‌کند. لذا زمانی که در یک موقعیت مشاهده (تکرار) کم باشد، دقت اتوچندجمله‌ای به مراتب از حالت چندجمله‌ای کلاسیک بیشتر است. اگر مقدار یک متغیر برابر صفر مشاهده شود، در روش کلاسیک، برآورد احتمال و انحراف معیار آن نیز صفر حاصل می‌شود، که نادرست است. اما در روش فضایی بر اساس اطلاعات همسایگی مقداری مخالف صفر به آن اختصاص داده می‌شود. به‌علاوه در روش فضایی نقاط اوج تعدیل می‌شود، یعنی روش فضایی به عنوان یک هموار کننده عمل می‌کند و به نوعی نسبت به نقاط دورافتاده استوار است. با توجه به اطلاعات همسایگی‌ها در روش

فضایی می‌توان نقاط فاقد مشاهده را پیش‌گویی نمود، در حالی که این مساله برای روش غیرفضایی امکان‌پذیر نیست.

مراجع

Besag, J. (1974), Spatial Interaction and Statistical Analysis of lattice System (with discussion). *J. Roy, Statist. Soc.*, Ser. B, **36**, 192-236.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, New york, Wiley. Combridge University Press.

Kavousi, A. Meshkani, M. R. Mohammadzadeh, M. (2007), Spatial Analysis of Relative Risk of Lip Cancer in Iran: A Bayesian Approach, *Environmetrics*, **20**, 347-359.

Lawson, A. B. (2001), *Statistical Methods in Spatial Epidemiology*, UK, Wiley.

Kavousi, A. Meshkani, M. R. Mohammadzadeh, M. (2011), Spatial Analysis of Auto-multivariate Lattice Data, *Statistical Papers* **52**, 937-952.

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱
مجموعه مقالات، ص ۱۱۷-۱۳۲

تحلیل بیزی داده‌های فضایی با استفاده از توزیع چوله‌نرمال بسته

امید کریمی [۱]، محسن محمدزاده [۲]

[۱] *دانشگاه سمنان، گروه آمار*
[۲] *دانشگاه تربیت مدرس، گروه آمار*

چکیده : در اغلب تحلیل‌های آمار فضایی فرض بر این است که داده‌ها تحققی از یک میدان تصادفی گاوسی هستند، اما مشخصه‌های ناگاوسی مانند متغیرهای تصادفی نامنفی با توزیع چوله در اکثر زمینه‌های علمی دیده می‌شوند. مدل‌بندی این نوع داده‌ها با استفاده از یک فرم گسسته‌ای از میدان تصادفی چوله گاوسی بسته، که براساس توزیع چوله نرمال بسته چند متغیره تعریف شده و از انعطاف‌پذیری بیشتری برخوردار است، صورت می‌پذیرد. در این مقاله خانواده توزیع‌های چوله نرمال بسته که نسبت به ترکیبات خطی و توزیع‌های شرطی بسته است، برای تحلیل داده‌های فضایی چوله ارائه می‌گردد. سپس پیشگویی فضایی بیزی برای میدان تصادفی چوله گاوسی بسته بیان شده و یک مطالعه شبیه‌سازی برای بررسی مناسب بودن مدل صورت پذیرفته است.

آدرس الکترونیک مسئول مقاله: امید کریمی، omid.karimi@profs.semnan.ac.ir
کد موضوع‌بندی ریاضی (۲۰۰۰): /////

واژه‌های کلیدی : توزیع چوله‌نرمال بسته، پیشگویی فضایی بیزی، میدان تصادفی چوله گاوسی بسته.

۱ مقدمه

پیشگویی فضایی یک مساله مهم در علوم محیطی مانند هواشناسی، زمین شناسی، جغرافیا، کشاورزی و غیره است. وقتی داده‌های فضایی ناگاوسی هستند اما تبدیلی از آنها گاوسی باشد، اولیویرا و همکاران (۱۹۹۷) و محمدزاده و خالدی (۱۳۸۳) کریگیدن گاوسی تبدیل یافته را برای پیشگویی فضایی مورد مطالعه قرار دادند. اما در عمل تبدیل نرمال ساز داده‌ها نامعلوم است و نه تنها تعیین آن تحلیل داده‌ها را با مشکلاتی مواجه می‌سازد بلکه گاهی تفسیر داده‌های تبدیل یافته نسبت به داده‌های اصلی از دشواری بیشتری برخوردار است (آزالینی و کاپیتانیو، ۱۹۹۹).

وقتی توزیع مجموعه‌ای از داده‌ها واجد اکثر خواص توزیع نرمال باشد اما متقارن نباشند، به‌عبارت دیگر دارای چولگی باشند، توزیع چوله نرمال[۱](SN) می‌تواند برای مدل بندی این‌گونه داده‌ها مورد استفاده قرار گیرد. توزیع چوله نرمال چند متغیره توسط آزالینی و دالاواله (۱۹۹۶) معرفی گردید. آزالینی و کاپیتانیو (۱۹۹۹) خصوصیات این توزیع را در مسائل کاربردی بیان کردند. سپس گوپتا و همکاران (۲۰۰۴) این توزیع را به حالت کلی‌تری بسط دادند. کلاس جدیدی از توزیع‌ها تحت عنوان چوله نرمال بسته[۲](CSN) توسط دامینگوس وهمکاران (۲۰۰۳) معرفی شد که اکثر توزیع‌های چوله نرمال معرفی شده را در بر می‌گیرد و خصوصیات این توزیع همانند بسته بودن تحت تبدیلات خطی، حاشیه‌ای و شرطی کردن به‌صورت جامع توسط گنزالس و همکاران (۲۰۰۴) ارائه شده است.

در آمار فضایی گاهی با مواردی مواجه می‌شویم که داده‌ها نامتقارن و چوله هستند، مانند داده‌های مالی (کوزوبوسکی، ۱۹۹۹)، داده‌های بارندگی (کیم و مالیک، ۲۰۰۴) یا داده‌های ارتعاشی[۳] (کریمی و همکاران، ۲۰۱۰). اولین بار کیم و مالیک (۲۰۰۲)

---

۱ Skew Normal

۲ Closed Skew Normal

۳ Seismic

تحلیل داده‌های فضایی برای یک میدان تصادفی چوله گاوسی[۴](SG) را مورد بررسی قرار دادند، همچنین کیم و مالیک (۲۰۰۴) پیشگویی فضایی بیزی را در یک مثال کاربردی برای این میدان تصادفی به‌کار گرفتند و پس از آنها آلارد و ناویو (۲۰۰۵) نحوه شبیه‌سازی یک میدان تصادفی CSG را برای داده‌های فضایی ارائه کردند. با توجه به اینکه توزیع CSN از توزیع چوله نرمال کلی‌تر و دارای خواص بسته بودن تحت تبدیلات خطی و شرطی کردن است، مطالعه یک میدان تصادفی CSG می‌تواند شرایط ساده‌تری برای پیشگویی فضایی دقیق‌تر فراهم نماید.

برای این منظور، در این مقاله با استفاده از یک میدان تصادفی چوله گاوسی بسته[۵](CSG) به تحلیل داده‌های فضایی چوله پرداخته می‌شود. مفاهیم اولیه توزیع چوله نرمال بسته در بخش ۲ ارائه می‌گردد. در بخش ۳ پیشگویی بیزی برای میدان تصادفی CSG معرفی و نحوه کاربست آن در یک مثال شبیه‌سازی نشان داده می‌شود. در بخش ۴ مدل معکوس گاوسی بیزی به مدل معکوس چوله گاوسی بسته بیزی تعمیم داده شده و ضمن مقایسه دقت این دو مدل، نحوه بکارگیری مدل پیشنهادی در تحلیل داده‌های ارتعاشی نشان داده شده است.

۲ توزیع چوله نرمال بسته

در این بخش توزیع چوله نرمال بسته که برای اولین بار توسط دامینگوس وهمکاران (۲۰۰۳) ارائه شد، بیان می‌گردد. دلیل اصلی نامگذاری این توزیع به چوله نرمال «بسته» این است که این توزیع نسبت به تبدیلات خطی، کناری و شرطی کردن بسته است. یعنی هر تبدیل خطی روی مولفه‌های این خانواده نیز عضوی از این خانواده است. برای آشنایی بیشتر می‌توان به دامینگوس و همکاران (۲۰۰۳)، گنزالس و همکاران (۲۰۰۴)، دامینگوس و همکاران (۲۰۰۷) و کریمی و محمدزاده (۲۰۰۹) مراجعه کرد.

توزیع چوله نرمال برای اولین بار توسط روبرتس (۱۹۶۶) به‌دست آمد، اما اولین فرم رسمی آن توسط آزالینی (۱۹۸۵، ۱۹۸۶) معرفی شد. حالت چند متغیره این توزیع

---

۴ Skew Gaussian

۵ Closed Skew Gaussian

توسط آزالینی و دالاواله (۱۹۹۶) با یک تابع چگالی به فرم

$$f(\boldsymbol{y};\boldsymbol{\mu},\Sigma,\lambda) = ۲\phi_p(\boldsymbol{y};\boldsymbol{\mu},\Sigma)\Phi(\boldsymbol{\lambda}'\Sigma^{-\frac{۱}{۲}}(\boldsymbol{y}-\boldsymbol{\mu}))$$

معرفی شد، که در آن $\phi_p(\cdot;\boldsymbol{\mu},\Sigma)$ چگالی نرمال p-متغیره با میانگین $\boldsymbol{\mu}$ و ماتریس واریانس کوواریانس Σ، $\Phi(\cdot)$ تابع توزیع تجمعی نرمال استاندارد و $\boldsymbol{\lambda}$ یک بردار p-بعدی از پارامترهای چولگی است. توزیع چوله نرمال بعنوان تعمیمی از توزیع نرمال برای مدل بندی چولگی ارائه شده است. در واقع این خانواده از توزیع‌ها علاوه بر اینکه شامل خانواده توزیع‌های نرمال هستند و در بعضی از خواص این توزیع‌ها مشترک هستند، دارای یک پارامتر چولگی می‌باشند که میزان چولگی توزیع را کنترل می‌کند.

تعریف ۱ (دامینگوس و همکاران، ۲۰۰۳) فرض کنید $p \geq ۱$، $q \geq ۱$، $\boldsymbol{\mu} \in \mathbb{R}^p$، $\boldsymbol{\nu} \in \mathbb{R}^q$، Γ یک ماتریس دلخواه $q \times p$، Σ و Δ ماتریس‌های معین مثبت به ترتیب با بعدهای $p \times p$ و $q \times q$ باشند. چگالی توزیع CSN به‌صورت

$$f_{p,q}(\boldsymbol{y};\boldsymbol{\mu},\Sigma,\Gamma,\boldsymbol{\nu},\Delta) = K\phi_p(\boldsymbol{y};\boldsymbol{\mu},\Sigma)\Phi_q[\Gamma(\boldsymbol{y}-\boldsymbol{\mu});\boldsymbol{\nu},\Delta], \qquad (۱)$$

تعریف می‌شود، که در آن

$$K^{-۱} = \Phi_q(\mathbf{o};\boldsymbol{\nu},\Delta+\Gamma\Sigma\Gamma'), \qquad (۲)$$

$\Phi_q(\cdot;\boldsymbol{\eta},\Psi)$ تابع توزیع تجمعی q-بعدی نرمال با میانگین $\boldsymbol{\eta}$ و ماتریس واریانس کوواریانس Ψ و $\mathbf{o}$ یک بردار q-بعدی با مولفه‌های صفر است. متغیر تصادفی $\boldsymbol{y}$ که دارای این توزیع باشد به‌صورت $\boldsymbol{y} \sim CSN_{p,q}(\boldsymbol{\mu},\Sigma,\Gamma,\boldsymbol{\nu},\Delta)$ نمایش داده می‌شود.

۳ میدان تصادفی چوله گاوسی بسته

یک میدان تصادفی CSG مشابه میدان تصادفی گاوسی به‌صورت زیر تعریف می‌شود.

تعریف ۲ میدان تصادفی $\{Z(s), s \in U \subseteq \mathbb{R}^d\}$ CSG نامیده می‌شود هرگاه برای هر عدد متناهی $m \geq ۱$، $Z(s_۱),\ldots,Z(s_m)$ دارای توزیع توام CSN چند متغیره

باشند.

برای تعریف فرم گسسته‌ای از یک میدان تصادفی CSG فرض کنید $W(s_1),\ldots,W(s_p)$ متغیرهای تصادفی از یک میدان تصادفی به‌صورت

$$W(s) = \boldsymbol{f}'(s)\boldsymbol{\beta} + E_1(s) \quad \forall s \in U = \{s_1,\ldots,s_p\}, \tag{۳}$$

باشند، که در آن U متعلق به فضای اقلیدسی $\mathbb{R}^d$، $d \geq 1$، $\beta \in \mathbb{R}^r$ ضرائب رگرسیونی، $\boldsymbol{f}(s) = [f_1(s),\ldots,f_r(s)]'$ بردار توابعی معلوم از موقعیت‌ها و $\{E_1(s), s \in U\}$ یک میدان تصادفی گاوسی با میانگین صفر و تابع کوواریانس $C(s,s') = \mathrm{Cov}(E_1(s), E_1(s'))$ باشد. همچنین به ازای هر $q \geq 1$، بردار تصادفی $\boldsymbol{E}_2 = (E_{21},\ldots,E_{2q})'$ را در نظر بگیرید که دارای توزیع $N_q(\mathbf{o},\Delta)$ و مستقل از $\boldsymbol{E}_1$ است، که در آن Δ ماتریس معین مثبت است. بردار تصادفی $\boldsymbol{V} = (V_1,\ldots,V_q)'$ را طوری تعریف کنید که j–امین مولفه آن به‌صورت

$$V_j = -\nu_j + \sum_{i=1}^{p} \gamma_j(s_i)E_1(s_i) + E_{2j}, \quad j = 1,\ldots,q, \tag{۴}$$

باشد، که در آن $\nu_j \in \mathbb{R}$ و $\gamma_j(s_i)$ها توابع حقیقی مقدار هستند. در این‌صورت میدان تصادفی $\{Z(s), s \in U\}$ که به‌صورت $Z(s) = [W(s)|\{V \geq 0\}]$ تعریف می‌شود فرم گسسته‌ای از یک میدان تصادفی CSG است، زیرا با قرار دادن $\boldsymbol{W} = (W(s_1),\ldots,W(s_p))'$ می‌توان نوشت:

$$\boldsymbol{W} = \boldsymbol{\mu} + \boldsymbol{E}_1,$$
$$\boldsymbol{V} = -\boldsymbol{\nu} + \Gamma\boldsymbol{E}_1 + \boldsymbol{E}_2,$$

که در آن $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_p)'$، $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_q)'$ و $\Gamma = [\gamma_j(s_i)]_{q\times p}$ است. در این‌صورت برای مقادیر ثابت p و q توزیع شرطی $[\boldsymbol{W}|\boldsymbol{V} \geq \mathbf{o}]$ به‌صورت $CSN_{p,q}(\boldsymbol{\mu},\Sigma,\Gamma,\boldsymbol{\nu},\Delta)$ است.

۴ پیشگویی بیزی برای میدان تصادفی چوله گاوسی بسته

فرض کنید $z = (z(s_1),\ldots,z(s_n))$ مشاهداتی از میدان تصادفی $\{Z(s), s \in U \subseteq \mathbb{R}^d\}$ در n موقعیت $\{s_1,\ldots,s_n\}$ باشند. برای پیشگویی

مقدار $Z(s_0)$ در موقعیت جدید s_0 بر اساس مشاهدات z و ارائه پیشگوی فضایی بیزی بر اساس میدان تصادفی CSG، $\boldsymbol{Z}^*$ را به‌صورت $\boldsymbol{Z}^* = (Z(s_0), \boldsymbol{Z}')'$ در نظر می‌گیریم، که در آن $\boldsymbol{Z} = (Z(s_1), \ldots, Z(s_n)')$. طبق تعریف ۲ $\boldsymbol{Z}^* \sim CSN_{n+1,q}(F^*\boldsymbol{\beta}, C^*, \Gamma^*, \boldsymbol{\nu}, \Delta)$ که در آن $F^* = (\boldsymbol{f}(s_0)\ ,\ F')'$ یک ماتریس طرح رتبه کامل ستونی با درایه‌های $\boldsymbol{f}(s_0) = (f_1(s_0), \ldots, f_r(s_0))'$ و $F = [f_j(s_i)]_{n\times r}$ است و f_jها توابعی معلوم از موقعیت‌ها هستند. همچنین

$$C^* = \begin{pmatrix} C(s_0, s_0) & \boldsymbol{c}' \\ \boldsymbol{c} & C \end{pmatrix},\ \boldsymbol{c} = [C(s_0, s_i)]_{n\times 1},$$

و $\Gamma^* = [\Gamma, \gamma]$ یک ماتریس $q \times (n+1)$ بعدی شامل پارامترهای چولگی است، ه در آن $\Gamma = [\gamma_j(s_i)]_{q\times n}$ و $\gamma = [\gamma_j(s_0)]_{q\times 1}$ است. در مکانیزم نظریه تصمیم، $L(Z(s_0), \hat{Z}(s_0))$ نشان دهنده میزان زیانی است که از پیشگویی $Z(s_0)$ بوسیله $\hat{Z}(s_0)$ حاصل می‌شود. یک پیشگوی بهینه با مینیمم کردن $\mathrm{E}[L(Z(s_0), \hat{Z}(s_0))|\boldsymbol{Z}]$ به‌دست می‌آید. بنابراین پیشگوی بهینه تحت تابع زیان توان دوم خطاها[۶]، $\mathrm{E}(Z(s_0)|\boldsymbol{Z})$ خواهد شد و برای محاسبه آن به توزیع شرطی $[Z(s_0)|\boldsymbol{Z} = z]$ نیاز است. با استفاده از خواص شرطی کردن توزیع CSN، چگالی شرطی $[Z(s_0)|\boldsymbol{Z} = z]$ به‌صورت

$$f_{1,q}(z(s_0)|\boldsymbol{z}) = K_1\phi(z(s_0); \mu_0, \sigma_0) \times \Phi_q(\boldsymbol{d}'(z(s_0) - \mu_0); \boldsymbol{\nu}_0, \Delta), \qquad (۵)$$

به‌دست می‌آید، که در آن

$$\begin{aligned} \mu_0 &= \boldsymbol{f}'(s_0)\boldsymbol{\beta} + \boldsymbol{c}'C^{-1}(\boldsymbol{z} - F\boldsymbol{\beta}),\ \ \sigma_0 = C(s_0, s_0) - \boldsymbol{c}'C^{-1}\boldsymbol{c} \\ \boldsymbol{\nu}_0 &= \boldsymbol{\nu} - \Gamma_1(\boldsymbol{z} - F\boldsymbol{\beta}),\ \ \Gamma_1 = \Gamma + \boldsymbol{\gamma}\boldsymbol{c}'C^{-1} \end{aligned}$$

و K_1 ثابت نرمال‌ساز است. با توجه به رابطه امید ریاضی توزیع CSN، برای تابع زیان توان دوم خطاها پیشگوی بهینه به‌صورت

$$\hat{Z}(s_0) \ = \ E(Z(s_0)|\boldsymbol{z})$$

[۶] Square-error loss

$$= \quad \boldsymbol{f}'(s_\circ)\boldsymbol{\beta} + \boldsymbol{c}'C^{-1}(\boldsymbol{z} - F\boldsymbol{\beta}) + \sigma_\circ \boldsymbol{\psi}'\boldsymbol{\gamma},$$

خواهد شد، که در آن $\psi = \frac{\Phi_q^*(\mathbf{0};\boldsymbol{\nu}_\circ,\Delta+\sigma_\circ\gamma\gamma')}{\Phi_q(\mathbf{0};\boldsymbol{\nu}_\circ,\Delta+\sigma_\circ\gamma\gamma')}$ است. در عمل ضرائب رگرسیونی، ساختار همبستگی فضایی و پارامترهای چولگی نامعلوم هستند. بنابراین یکی از مسائل مهم برای پیشگویی، برآورد پارامترهای نامعلوم مدل است.

فرض کنید ν و Δ معلوم هستند و کوواریانس فضایی ایستا به‌صورت $C(h) = \sigma^2\rho(h;\theta)$ باشد، که در آن $\rho(\cdot;\theta)$ یک تابع همبستگی معلوم و پارامترهای $(\sigma^2,\theta) \in (\circ,\infty)^2$ نامعلوم هستند. θ پارامتر همبستگی فضایی و σ^2 واریانس است. همچنین برای سادگی تفسیر چولگی و کاهش بعد پارامتر چولگی در مدل، می‌توان ناحیه فضایی U را به k ناحیه، هر یک با چولگی ثابت تقسیم‌بندی کرد، در آن‌صورت $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k)'$ خواهد شد، که در آن $\Gamma^* = (\alpha_1 J_1 \ , \ \ldots \ , \ \alpha_k J_k)$ برداری از پارامترهای چولگی و J_iها ماتریس‌های $q \times n_i$ بعدی با درایه‌های یک و $\sum_{i=1}^{k} n_i = n+1$ هستند.

با فرض عدم حتمیت پارامترهای مدل، $\boldsymbol{\eta} = (\beta,\sigma^2,\theta,\alpha)$، رهیافت بیزی برای برآورد آن‌ها و نهایتا محاسبه پیشگوی بهینه $Z(s_\circ)$ براساس داده‌های $\boldsymbol{z}$ قابل استفاده است. یک مدل بیزی شامل تابع درستنمایی $f(\boldsymbol{z}|\boldsymbol{\eta})$ و توزیع پیشین $\pi(\boldsymbol{\eta})$ است.

مطابق خاصیت خانواده توزیع‌های CSN، چگالی کناری $\boldsymbol{Z}$ به شرط $\boldsymbol{\eta}$ به‌صورت

$$f_{n,q}(\boldsymbol{z}|\boldsymbol{\eta}) = K_2\phi_n(\boldsymbol{z};F\boldsymbol{\beta},\sigma^2 R_\theta)\Phi_q(D_1(\boldsymbol{z}-F\boldsymbol{\beta});\boldsymbol{\nu},\Delta_z), \qquad (۶)$$

به‌دست می‌آید، که در آن $\Delta_z = \Delta + \sigma_\circ\gamma\gamma'$، $R_\theta = (\rho(s_i,s_j;\theta))$ و K_2 ثابت نرمال ساز است. با فرض استقلال پیشین، توزیع توام پیشینی به‌صورت

$$\pi(\boldsymbol{\eta}) = \pi(\boldsymbol{\beta},\sigma^2,\theta,\boldsymbol{\alpha}) = \pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\theta)\pi(\boldsymbol{\alpha})$$

و توزیع پسینی متناسب با

$$f(\boldsymbol{z}|\boldsymbol{\eta})\pi(\boldsymbol{\eta}) = f(\boldsymbol{z}|\boldsymbol{\beta},\sigma^2,\theta,\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\theta)\pi(\boldsymbol{\alpha}),$$

خواهد شد. لیزو و لوپرفیدو (۲۰۰۳) نحوهٔ انتخاب توزیع‌های پیشین برای توزیع چوله نرمال را مورد بحث قرار دادند. همچنین لیزو و لوپرفیدو (۲۰۰۶) پیشین‌های

مرجع[۷] و جفریز[۸] را برای پارامتر چولگی توزیع چوله نرمال به‌دست آوردند و نشان دادند این پیشین‌ها سره هستند. بنابراین به‌خاطر اینکه توزیع پسین سره شود، برای همه پارامترهای مدل، پیشین‌های سره در نظر می‌گیریم. پیشین‌های متداول برای $\boldsymbol{\beta}$، $N_r(\boldsymbol{\beta}_\circ, \Sigma_\circ)$ و برای σ^2، $IG(\varphi, \tau)$ اتخاذ شده‌اند. با در نظر گرفتن پیشین‌های دلخواه برای θ و $\boldsymbol{\alpha}$ توزیع پسین به‌صورت

$$\pi(\boldsymbol{\eta}|z) \propto f(z|\boldsymbol{\eta})\phi_r(\boldsymbol{\beta};\boldsymbol{\beta}_\circ,\Sigma_\circ)\frac{\tau^\varphi}{\Gamma(\varphi)}(\frac{1}{\sigma^2})^{\varphi+1}exp\{-\frac{\tau}{\sigma^2}\}\pi(\theta)\pi(\boldsymbol{\alpha}).$$

به‌دست می‌آید، که دارای فرم پیچیده‌ای است. بنابراین از روش‌های MCMC برای شبیه‌سازی از توزیع پسین استفاده می‌شود. برای به کارگیری الگوریتم نمونه‌گیر گیبس توزیع‌های شرطی کامل[۹] به‌صورت

$$\begin{aligned}
\pi(\boldsymbol{\beta}|z,\sigma^2,\theta,\boldsymbol{\alpha}) &\propto \phi_n(z;F\boldsymbol{\beta},\sigma^2R_\theta)\phi_r(\boldsymbol{\beta};\boldsymbol{\beta}_\circ,\Sigma_\circ)\Phi_q(\Gamma_1(z-F\boldsymbol{\beta});\boldsymbol{\nu},\Delta_z),\\
\pi(\sigma^2|z,\boldsymbol{\beta},\theta,\boldsymbol{\alpha}) &\propto IG(\varphi_{\sigma^2},\tau_{\sigma^2})\Phi_q(\Gamma_1(z-F\boldsymbol{\beta});\boldsymbol{\nu},\Delta_z)\Phi_q^{-1}(\mathbf{o};\boldsymbol{\nu},\Delta_c),\\
\pi(\theta|z,\boldsymbol{\beta},\sigma^2,\boldsymbol{\alpha}) &\propto \frac{1}{|R_\theta|^{\frac{1}{2}}}exp\{-\frac{1}{2\sigma^2}(z-F\boldsymbol{\beta})'R_\theta^{-1}(z-F\boldsymbol{\beta})\}\\
&\quad\times\Phi_q(\Gamma_1(z-F\boldsymbol{\beta});\boldsymbol{\nu},\Delta_z)\Phi_q^{-1}(\mathbf{o};\boldsymbol{\nu},\Delta_c)\pi(\theta),\\
\pi(\boldsymbol{\alpha}|z,\boldsymbol{\beta},\sigma^2,\theta) &\propto \Phi_q(\Gamma_1(z-F\boldsymbol{\beta});\boldsymbol{\nu},\Delta_z)\Phi_q^{-1}(\mathbf{o};\boldsymbol{\nu},\Delta_c)\pi(\boldsymbol{\alpha}).
\end{aligned}$$

به‌دست می‌آیند، که در آن‌ها $\varphi_{\sigma^2} = \frac{n}{2}+\varphi$، $\tau_{\sigma^2} = \frac{1}{2}(z-F\boldsymbol{\beta})'R_\theta^{-1}(z-F\boldsymbol{\beta})+2\tau$ و $\Delta_c = \Delta_z + \sigma^2 D_1 R_\theta D_1'$ است. براساس نمونه‌های تولید شده از توزیع‌های شرطی کامل، کمیت‌های توزیع پیشگوی بیزی

$$f(z(s_\circ)|z) = \int f(z(s_\circ)|z,\boldsymbol{\eta})\pi(\boldsymbol{\eta}|z)d\boldsymbol{\eta},$$

را می‌توان برای پیشگویی $Z(s_\circ)$ استفاده کرد، که در آن $f(z(s_\circ)|z,\boldsymbol{\eta})$ به‌صورت (۵) تعریف می‌شود.

---

۷ Reference Priors

۸ Jeffreys

۹ Full Conditional Distributions

قضیه ۱ (کریمی و محمدزاده، ۲۰۱۱) فرض کنید در مدل CSG چگالی $f(z|\boldsymbol{\eta})$ به فرم (۶) و پارامتر $\boldsymbol{\beta}$ دارای توزیع پیشین $N_r(\boldsymbol{\beta}_0, \Sigma_0)$ باشد. در این‌صورت $\boldsymbol{\beta}$ دارای توزیع شرطی کامل

$$[\boldsymbol{\beta}|z, \sigma^2, \theta, \boldsymbol{\alpha}] \sim CSN_{r,q}(\boldsymbol{\mu}_\beta, \Sigma_\beta, -\Gamma_1 F, \boldsymbol{\nu} - \Gamma_1(z - F\boldsymbol{\mu}_\beta), \Delta_z), \qquad (۷)$$

است، که در آن $\Sigma_\beta = (\frac{1}{\sigma^2} F R_\theta^{-1} F + \Sigma_0^{-1})^{-1}$ و $\boldsymbol{\mu}_\beta = \Sigma_\beta(\frac{1}{\sigma^2} F R_\theta^{-1} z + \Sigma_0^{-1}\boldsymbol{\beta}_0)$ می‌باشد.

طبق قضیه ۱، توزیع شرطی کامل $\boldsymbol{\beta}$ دارای فرم خاصی از توزیع CSN است، بنابراین تولید نمونه از آن به راحتی انجام می‌گیرد. توزیع‌های شرطی کامل α، θ و σ^2 دارای فرم خاصی از توزیع‌های شناخته شده نیستند. بنابراین از الگوریتم متروپولیس–هستینگس برای تولید نمونه از آن‌ها استفاده می‌شود. در ادامه، قضیه‌ای برای بدست آوردن تقریبی از توزیع شرطی کامل $\boldsymbol{\alpha}$ ارائه می‌شود، که از آن می‌توان به‌عنوان توزیع نامزد[10] در الگوریتم متروپولیس–هستینگس استفاده نمود.

قضیه ۲ (کریمی و محمدزاده، ۲۰۱۱) فرض کنید در مدل CSG چولگی در کل ناحیه فضایی ثابت و $f(z|\boldsymbol{\eta})$ به فرم (۶) باشد. اگر $\Delta_z \approx \Delta$، $\Delta_c \approx \Delta$ و α دارای توزیع پیشین $N(\alpha_0, \sigma_\alpha^2)$ باشد، آن‌گاه توزیع شرطی کامل α تقریباً به‌صورت

$$CSN_{1,q}(\alpha_0, \sigma_\alpha^2, A, \boldsymbol{\nu} - A\alpha_0, \Delta), \qquad (۸)$$

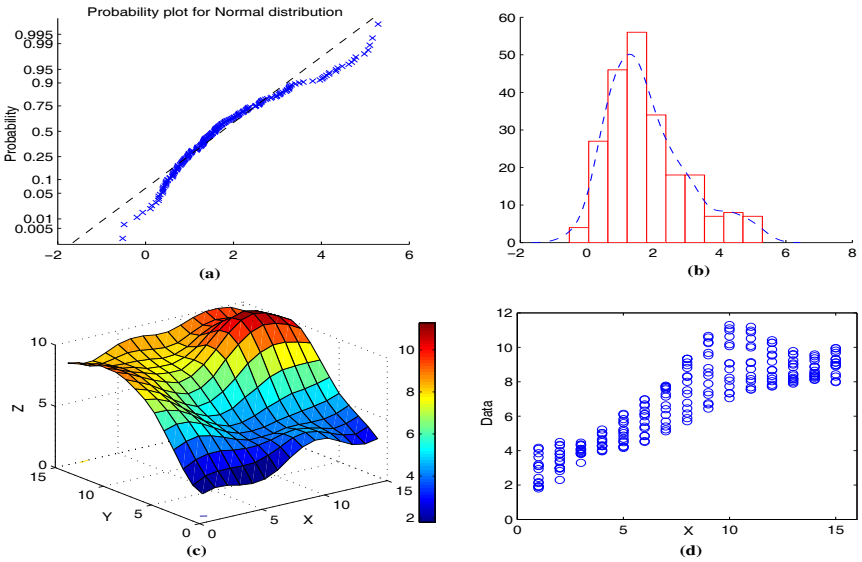
است، که در آن $A = (J_{qn} + \frac{1}{\sigma^2}\mathbf{1}_q \boldsymbol{c}' R_\theta^{-1})(z - F\boldsymbol{\beta})$، J_{qn} یک ماتریس $q \times n$ با عناصر یک و $\mathbf{1}_q$ یک بردار از یک‌ها با بعد q است.

توزیع شرطی کامل σ^2 را می‌توان به‌صورت $\pi(\cdot) \propto g(\cdot)\Psi_q(\cdot)$ نوشت، که در آن $g(\cdot)$ چگالی توزیع گامای معکوس[11]، $IG(\frac{n}{2} + \varphi, \frac{1}{2}[(z - F\boldsymbol{\beta})' R_\theta^{-1}(z - F\boldsymbol{\beta}) + 2\tau])$ است و می‌توان آن را به‌عنوان توزیع نامزد در نظر گرفت. بنابراین احتمال پذیرش در الگوریتم متروپولیس–هستینگس به‌صورت $r(x, y) = \min\{\frac{\Psi_q(y)}{\Psi_q(x)}, 1\}$ خلاصه می‌شود. چون پارامتر θ باید مثبت باشد، توزیع گاما را می‌توان به‌عنوان توزیع نامزد در نظر گرفت.

---

[10] Proposal Distribution

[11] Inverse Gamma Distribution



شکل ۱: (a) نمودار احتمال نرمال؛ (b) بافت‌نگار داده‌های شبیه‌سازی شده بعد از حذف روند. (c) نمودار رویه داده‌ها. (d) نمودار پراکنش داده‌ها در جهت محور x.

۱.۴ *شبیه‌سازی*

در این بخش یک مطالعه شبیه‌سازی برای بررسی اعتبار مدل CSG روی داده‌های فضایی چوله انجام می‌گیرد. ابتدا پارامترهای مدل با رهیافت بیزی برآورد شده، سپس چگونگی پیشگویی بیزی در یک موقعیت جدید دلخواه ارائه می‌شود. تولید نمونه از توزیع پیشگوی $f(z(s_0)|z)$ نیز طی مراحل زیر انجام می‌پذیرد.

۱- $\boldsymbol{\eta}^*$ را از توزیع $\pi(\boldsymbol{\eta}|z)$ با استفاده از الگوریتم‌های MCMC تولید نمایید.

۲- $z(s_0)$ را از توزیع $[z(s_0)|\boldsymbol{\eta}^*, z] \sim CSN_{1,q}$ با چگالی ارائه شده در (۵) تولید کنید.

مشاهداتی از میدان تصادفی CSG، $\{Z(s), s \in U \subseteq R^2\}$ روی شبکه منظم 15×15 با روند خطی $\boldsymbol{f}'(s_i)\boldsymbol{\beta} = \beta_0 + \beta_1 x_i$ شبیه‌سازی شده است، که در آن $s = (x, y)$ و پارامترهای مدل به‌صورت $\beta_0 = 1/5,\ \beta_1 = 0/5,\ \boldsymbol{\nu} = (0, 0)',\ \Delta = I,\ q = 2$ در نظر گرفته شده‌اند. در این شبیه‌سازی فرض شده است A_1 و A_2 افرازهایی از ناحیه فضایی U باشند، که در آن‌ها توزیع داده‌ها دارای پارامترهای چولگی متفاوت

به ترتیب با مقادیر $\alpha_1 = 5$ و $\alpha_2 = 2$ هستند. در ناحیه U، ۲۲۵ موقعیت وجود دارد که موقعیت‌های $s_1, \ldots, s_{112}$ در ناحیه A_1 و بقیه موقعیت‌های $s_{113}, \ldots, s_{225}$ در ناحیه A_2 قرار دارند. در این‌صورت ماتریس چولگی Γ را می‌توان به‌صورت

$$\Gamma = \begin{pmatrix} \alpha_1 \mathbf{1}'_{112} & \mathbf{0}'_{113} \\ \mathbf{0}'_{112} & \alpha_2 \mathbf{1}'_{113} \end{pmatrix}_{2\times 225},$$

تجزیه کرد، که در آن $\mathbf{0}_k$ بردار k بعدی با عناصر صفر و $\mathbf{1}_{k'}$ بردار k' بعدی با درایه‌های یک است. برای ساختار فضایی داده‌ها مدل نمایی همسانگرد با پارامترهای $\theta = 4$ و $\sigma^2 = 5$ در نظر گرفته شده است. نمودار بافت‌نگار و احتمال نرمال داده‌های شبیه‌سازی شده بدون منظور کردن روند خطی در شکل ۱ (a)-(b) نشان می‌دهند که داده‌ها ناگاوسی و چوله به راست هستند. نمودار رویه و پراکنش داده‌های شبیه‌سازی در شکل ۱ (c)-(d) رسم شده‌اند که نشان می‌دهند داده‌ها در جهت محور xها دارای روند خطی هستند. برای برآورد بیزی پارامترهای مدل، پیشین‌های

$$\boldsymbol{\beta} = (\beta_0, \beta_1)' \sim N_2(\mathbf{0}, 10I),\ \sigma^2 \sim IG(2, 10),$$
$$\theta \sim \Gamma(2, 5),\ \boldsymbol{\alpha} = (\alpha_1, \alpha_2)' \sim N_2(\mathbf{0}, 10I),$$

منظور شده‌اند. پارامترهای توزیع پیشین با توجه به بررسی حساسیت پیشین نسبتا با واریانس بالا اتخاذ گردیده‌اند. سپس الگوریتم‌های MCMC با ۵۰۰۰ تکرار اجرا و تحلیل همگرایی الگوریتم‌ها بیانگر حدود ۳۰۰۰ مقادیر داغیدن[۱۲] برای همه پارامترها است. رهیافت بیزی برای مدل‌های گاوسی و CSG اجرا و نتایج در جدول ۱ ارائه شده‌اند. برآوردهای بیزی پارامترها در مدل CSG نسبت به مدل گاوسی به مقادیر واقعی نزدیکتر و از انحراف استانداردهای کوچکتری برخوردارند.

همچنین ملاک اعتبارسنجی متقابل برای مقایسه دو مدل محاسبه و مقادیر آن‌ها در جدول ۲ برتری مدل CSG را نسبت به مدل گاوسی بیان می‌کند. به‌عنوان مثال پیشگویی بیزی در دو موقعیت s_0 و $s_{225} = (15, 15)$ برای هر دو مدل به همراه مقدار واقعی آن در جدول ۲ ارائه شده است. همان‌طور که ملاحظه می‌شود

[۱۲] Burn-in

جدول ۱: برآورد بیزی و انحراف استاندارد پارامترها برای دو مدل گاوسی و CSG .

مدل		CSG		گاوسی	
پارامتر	مقدار واقعی	برآورد	انحراف استاندارد	برآورد	انحراف استاندارد
β_0	۱/۵	۱/۸۷۴۲	۰/۷۴۱۴	۲/۶۸۵۹	۱/۰۰۴۹
β_1	۰/۵	۰/۵۰۹۹	۰/۰۸۳۴	۰/۴۴۱۱	۰/۹۹۳۳
σ^2	۵	۵/۱۳۰۴	۰/۷۲۴۹	۴/۰۵۵۶	۲/۱۶۹۳
θ	۴	۳/۷۷۷۸	۰/۷۹۷۲	۳/۰۵۶۵	۰/۶۵۲۹
α_1	۵	۵/۰۳۹۴	۱/۱۷۱۷	-----	-----
α_2	۲	۲/۱۷۰۳	۱/۱۱۲۷	-----	-----

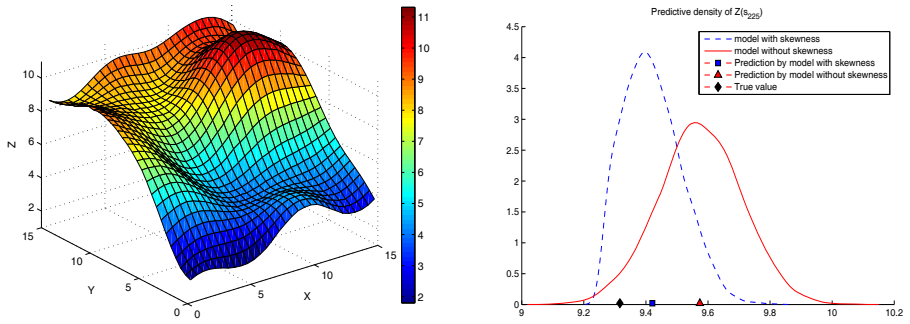
جدول ۲: پیشگویی بیزی و ملاک CVMSE برای دو مدل گاوسی و CSG.

مدل		CSG		گاوسی	
پارامتر	مقدار واقعی	پیشگویی	انحراف معیار پیشگویی	پیشگویی	انحراف معیار پیشگویی
$Z(s_0)$	-----	۲/۹۵۹۸	۰/۰۲۳۱	۲/۹۰۹۴	۰/۰۳۷۹
$Z(s_{225})$	۹/۳۱۶۱	۹/۴۲۰۴	۰/۰۱۱۹	۹/۵۷۵۸	۰/۱۳۴۴
$CVMSE$		۰/۰۰۲۱		۰/۰۹۲	

پیشگویی در موقعیت قدیمی s_{225} نسبت به موقعیت جدید فاقد مشاهده s_0 برای مدل CSG از دقت بیشتری برخوردار هستند. با توجه به اینکه s_{225} یک نقطه مرزی است، می‌توان نتیجه گرفت که مدل CSG بهتر از مدل گاوسی در این نقاط عمل می‌کند. و در حالت کلی ملاک CVMSE پیشگویی با مدل CSG به میزان قابل ملاحظه‌ای نسبت به مدل گاوسی کاهش یافته است. همچنین چگالی‌های پیشگوی $Z(s_{225})$ برای هر دو مدل در شکل ۲ (راست) رسم شده و بطور کلی تمام تحلیل‌های انجام شده نشان می‌دهند پیشگوی بیزی با مدل CSG به مقدار واقعی نزدیکتر است. در نهایت نمودار رویه پیشگویی برای مدل چوله به‌وسیله ۶۷۵ موقعیت جدید روی کل ناحیه فضایی در شکل ۲ (چپ) رسم شده است که یک الگوی پیشگویی با رویهٔ هموار و واضح‌تر نسبت به رویه داده‌ها برای کل ناحیه ارائه می‌دهد.

۵ بحث و نتیجه‌گیری

وقتی داده‌های فضایی چوله هستند، کریگیدن نمی‌تواند بهترین پیشگوی فضایی را فراهم سازد. کریگیدن گاوسی تبدیل یافته نیز حتی اگر تبدیل نرمال‌ساز موجود



شکل ۲: (راست) چگالی پیشگوی $Z(s_{۲۲۵})$ با روش اعتبار سنجی متقابل برای هر دو مدل، (چپ) نمودار رویه پیشگویی برای مدل CSG روی شبکه منظم ۱۵ × ۱۵.

باشد، دارای ضعف‌هایی است. توزیع چوله‌نرمال ممکن است یک مدل مناسب برای مدل‌بندی داده‌های فضایی چوله فراهم نماید. اما توزیع چوله نرمال علاوه بر این که تحت شرطی کردن بسته نیست، نمی‌توان در نواحی مختلف چولگی‌های متفاوتی همراه با همبستگی فضایی توسط این توزیع تعریف نمود. از این‌رو در این مقاله از توزیع چوله نرمال بسته برای مدل‌بندی داده‌های فضایی چوله استفاده شد، که کلاس بزرگتری از توزیع چوله نرمال است و تحت تبدیلات خطی و شرطی کردن بسته است. همچنین پیشگوی فضایی بیزی برای میدان تصادفی چوله گاوسی بسته ارائه و توزیع‌های شرطی کامل برای برآورد پارامترهای مدل توسط الگوریتم‌های MCMC محاسبه شده‌اند. معیار MSE اعتبارسنجی متقابل نشان می‌دهد که مدل CSG مناسب‌تر از مدل چوله گاوسی و کریگیدن گاوسی تبدیل‌یافته برای پیشگویی داده‌های فضایی چوله عمل می‌کند.

مراجع

محمدزاده، م.، جعفری خالدی، م. (۱۳۸۳)، پیشگویی فضایی بیزی برای یک میدان تصادفی تبدیل یافته، مجله علوم دانشگاه تهران، جلد ۳۰، شماره اول، ۱۳۳-۱۴۴.

Allard, D. and Naveau P. (2005). Modeling Skewness in Spatial Data Analyis without Data Transformation. *Quantitative Geology and Geostatistics*, **14**, 929-937.

Azzalini, A. (1985). A Class of Distributions which Includes the Normal Ones. *Scandinavia Journal of Statistics*, **12**, 171-178.

Azzalini, A. (1986). Further Results on a Class of Distributions Which Includes the Normal Ones. *Statistica*, **46**, 199-208.

Azzalini, A. and Capitanio, A. (1999). Statistical Applications of the Multivariate Skew Normal Distributions. *Journal of the Royal Statistical Society, Series B*, **61**, 579-602.

Azzalini, A. and Dalla Valle, A. (1996). The Multivariate Skew-normal Distribution. *Biometrika*, **83**, 715-726.

Buland, A. and Omre, H. (2003). Bayesian Linearized AVO Inversion. *Geophysics*, **68**, 189-198.

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

De Oliveira, V., Kedem, B. and Short, D.S. (1997). Bayesian Prediction of Transformed Gaussian Random Fields, *Journal of The American Statistical Association*. **92**, 1422-1433.

Dominguez-Molina, J., Gonzalez-Farias, G. and Gupta, A. (2003). The Multivariate Closed Skew Normal Distribution. *Technical Report 03-12*, Department of Mathematics and Statistics, Bowling Green State University.

Dominguez-Molina, J., Gonzalez-Farias, G., Ramos-Quiroga, R. and Gupta, A. (2007). A Matrix Variate Closed Skew-Normal Distribution with Applications to Stochastic Frontier Analysis. *Communications in Statistics-Theory and Methods*, **36**, 1691-1703.

González-Farías, J., Dominguez-Molina, A. and Gupta, A. K. (2004). The Closed Skew Normal Distribution. in M. G. Genton, (ed), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Boca Raton, FL: Chapman and Hall, 25-42.

Gupta, A., Gonzalez-Farias, G. and Dominguez-Molina, J. (2004). A Multivariate Skew Normal Distribution. *Journal of Multivariate Analysis*, **89**, 181-190.

Karimi, O. and Mohammadzadeh, M. (2011). Bayesian Spatial Prediction for Discrete Closed Skew Gaussian Random Field. *Mathematical Geosciences*, **43**,565-582.

Karimi, O., Omre, H. and Mohammadzadeh, M. (2010). Bayesian Closed-Skew Gaussian Inversion of Seismic AVO Data for Elastic Material Properties. *Geophysics*, **75**, R1-11.

Kim, H.-M. and Mallick, B. K. (2002). Analyzing Spatial Data Using Skew-Gaussian processes. *In Spatial Cluster Modelling*, A. B. Lawson, D. G. T. Denison,(eds.), 163-173.

Kim, H-M., Ha, E. and Mallick, B. (2004). Spatial Prediction of Rainfall Using Skew-normal Processes. In Genton, M. G., editor, *Skew-elliptical Distributions and their applications: a Journey Beyond Normality*, Chapter 16, 279-289. Chapman and Hall/CRC.

Kim, H-M. and Mallick, B. (2004). A Bayesian Prediction Using the Skew Gaussian Distribution. *Journal of Statistical Planning and Inference*, **120**, 85-101.

Kozubowski, T.J. (1999). Geometric Stable Laws: Estimation and Applications. *Mathematical and Computer Modelling*, **29**, 241-253.

Liseo, B. and Loperfido, N. (2003). A Bayesian Interpretation of the Multivariate Skew-normal Distribution. *Statistics and Probability Letters*, **61**, 395-401.

Liseo, B. and Loperfido, N. (2006). A Note on Reference Priors for the Scalar Skew-normal Distribution. *Journal of Statistical Planning and Inference*, **136**, 373-389.

Roberts, C. (1966). A Correlation Model Useful in the Study of Twins. *Journal of American Statistical Association*, **61**, 1184-1190.

Stolt, R.H. and Weglein, A.B. (1985). Migration and Inversion of Seismic data. *Geophysics*, **50**, 2458-2472.

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱

مجموعه مقالات، ص ۱۳۳-۱۴۶

روش‌های استوار تحلیل داده‌های فضایی

انور محمدی، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده : حضور داده‌های دورافتاده در مشاهدات اثر تخریبی در برآورد تغییرنگار و سایر بخش‌های تحلیل داده‌های فضایی همچون پیشگویی فضایی و برآورد پارامترهای روند دارد. در این مقاله برآوردگرهای استوار تغییرنگار معرفی شده و کارکرد آنها در یک مطالعه شبیه‌سازی مورد ارزیابی و مقایسه قرار می‌گیرد. سپس برآورد استوار پارامترهای روند و نحوه پیشگویی فضایی استوار ارائه می‌گردد. در پایان ضمن ارائه یک فرآیند کاری برای تحلیل استوار داده‌های فضایی، نحوه بکار گیری آنها در تحلیل استوار متوسط دمای هوای سالانه ۱۷۰ شهر ایران مطرح و نقشه‌های آماری آنها ارائه می‌گردند.

واژه‌های کلیدی : د اده‌های فضایی، تحلیل استوار، تغییرنگار، روند، پیشگویی فضایی.

آدرس الکترونیک مسئول مقاله: انور محمدی، amstat@gmail.com

کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲H۱۱

۱ مقدمه

در آمار فضایی ساختار همبستگی فضایی داده‌ها توسط تابع تغییرنگار[۱] یا هم‌تغییرنگار[۲] تعیین و در تحلیل داده‌های فضایی مورد استفاده قرار می‌گیرد. اما در عمل این توابع نامعلوم‌اند و باید براساس مشاهدات برآورد شوند. ماترون (۱۹۶۲) یک برآورد گشتاوری برای تغییرنگار ارائه داد، که به شدت تحت تاثیر داده‌های دورافتاده قرار دارد. برای رفع این مشکل کرسی و هاوکینز (۱۹۸۰)، جنتون (۱۹۹۸)، لارک (۲۰۰۰) و محمدی و محمدزاده (۱۳۹۰) برآوردگرهایی استوار برای تغییرنگار پیشنهاد دادند، که قادرند اثر داده‌های دورافتاده را کنترل کنند. از طرفی وجود روند در داده‌ها موجب اریبی برآورد تغییرنگار می‌شود. لذا ضروری است ابتدا روند داده‌ها برآورد و پس از حذف روند، تغییرنگار برآورد شود. به علت اثر داده‌های دورافتاده در برآورد روند، تعیین روشی استوار برای این منظور نیز ضروری است. میلیتینو و همکاران (۲۰۰۳) روشی برای برآورد استوار پارامترهای یک روند خطی فضایی پیشنهاد دادند. پیشگویی فضایی هدف اصلی تحلیل داده‌های فضایی است و روش کریگیدن نیز به شدت تحت تاثیر داده‌های دورافتاده قرار دارد. لذا هاوکینز و کرسی (۱۹۸۴)، آرمسترانگ و بوفاسا (۱۹۸۸)، برک (۲۰۰۱)، فورنیر و فورر (۲۰۰۵) و محمدی (۱۳۹۰) به بررسی روش‌های پیشگویی فضایی استوار پرداختند.

در این مقاله ابتدا برآوردگرهای استوار تغییرنگار و توانایی آنها در برآورد صحیح تغییرنگار در حضور حجم‌های متفاوت داده‌های دورافتاده فضایی مورد ارزیابی قرار می‌گیرد. سپس روش برآورد استوار روند داده‌ها ارائه و بررسی می‌گردد تا بتوان برآورد استوار تغییرنگار را براساس داده‌های فاقد روند محاسبه و شرایط لازم برای پیشگویی استوار داده‌های فضایی را فراهم نمود. در انتها روش‌های استوار ارائه شده برای تحلیل داده‌های متوسط دمای کشور به کار گرفته شده و به بحث و نتیجه‌گیری پرداخته می‌شود.

---

۱ Variogram

۲ Covariogram

۲ برآورد استوار تغییرنگار

داده‌های فضایی مشاهدات $Z(s_1),\ldots,Z(s_n)$ از میدان تصادفی $\{Z(s): s\in D\}$ در موقعیت‌های $s_1,\ldots,s_n$ واقع در ناحیه $D\subset R^2$ هستند. با فرض آنکه میدان مانای ذاتی باشد، ساختار همبستگی فضایی آن توسط تغییرنگار $2\gamma(h)=Var[Z(s)-Z(s+h)],\ s,h\in D$ تعیین می‌شود. برآوردگر گشتاوری تغییرنگار توسط ماترون (۱۹۶۲)، به فرم

$$2\hat{\gamma}_M(h)=\frac{1}{N_h}\sum_{N(h)}(Z(s_i)-Z(s_j))^2 \quad h\in R^d$$

پیشنهاد شد، که در آن N_h تعداد اعضای $N(h)=\{(s_i,s_j): s_i-s_j=h\}$ می‌باشد. این برآوردگر ناریب است، اما به علت وجود عبارت توان دوم $Z(s_i)-Z(s_j)$، به شدت تحت تاثیر داده‌های دورافتاده قرار دارد. کرسی و هاوکینز (۱۹۸۰) یک برآوردگر استوار برای تغییرنگار به صورت

$$2\hat{\gamma}_{CH}(h)=\left(\frac{1}{N_h}\sum_{N(h)}|Z(x_i)-Z(x_j)|^{1/2}\right)^4/(0/457+\frac{0/494}{N_h}) \quad h\in R^d$$

ارائه نمودند، که به علت استفاده از ریشه دوم $Z(x+h)-Z(x)$، نسبت به برآوردگر کلاسیک ماترون، حساسیت کمتری به داده‌های دورافتاده دارد. اما همچنان مشکل حساسیت به حجم‌های بالای داده‌های دورافتاده پابرجاست.

برای ارزیابی میزان استواری یک برآوردگر، شاخص‌های گوناگونی مورد استفاده قرار می‌گیرد. نقطه فروریزش[۳] (ϵ^*) اندازه توانایی مقاومت برآوردگر در برابر حضور داده‌های دورافتاده را نشان می‌دهد (هابر، ۱۹۸۱). تابع تاثیر[۴] (هامپل و همکاران، ۱۹۸۶) نیز اثر وجود آلودگی بی‌نهایت کوچک در یک مشاهده بر روی برآورد را اندازه‌گیری می‌کند. نقطه فروریزشِ بالا و تابع تاثیرِ کراندار از خواص مثبت یک روش استوار به شمار می‌آیند. برای بررسی کراندار بودن تابع تاثیر معمولاً از شاخص حساسیت به خطاهای ناخالص[۵] (γ^*)، که سوپریمم تابع تاثیر است

---

۳ Breakdown Point

۴ Influnce Function

۵ Gross-Error Sensitivity

استفاده می‌شود.

جنتون (۱۹۹۸) نشان داد برآوردگرهای $2\hat{\gamma}_M(h)$ و $2\hat{\gamma}_{CH}(h)$ دارای حساسیت به خطاهای ناخالص نامتناهی $\gamma^* = \infty$ می‌باشند، یعنی تابع تاثیر کراندار ندارند. بعلاوه این برآوردگرها دارای نقطه فروریزش $\varepsilon^* = 0$ هستند (هوبر، ۱۹۸۱)، یعنی از نظر نقطه فروریزش نیز استوار نیستند.

میدان تصادفی نموها در تاخیر h، یعنی $V(h) = Z(s+h) - Z(s)$، را که دارای میانگین صفر و واریانس $2\gamma(h)$ می‌باشد، در نظر بگیرید. اگر $\{V_1(h), ..., V_{N_h}(h)\}$ نمونه‌ای از $V(h)$ حاصل از نمونهٔ $\{Z(s_1), ..., Z(s_n)\}$ باشد، یافتن برآوردگری برای واریانس $V(h)$ معادل یافتن برآوردگری برای تغییرنگار در تاخیر h خواهد بود. با تأمل در رابطه برآوردگر ماترون (۱۹۶۲)، می‌توان آنرا به فرم $2\hat{\gamma}_M(h) = \frac{1}{N_h}\sum_{i=1}^{N_h} V_i(h)^2,\ h \in R^d$ نوشت، که در واقع برآوردگر کلاسیک واریانس نمونه $\{V_1(h), ..., V_{N_h}(h)\}$ است. در نتیجه یک راهکار برای برآورد استوار تغییرنگار، استفاده از برآوردگرهای استوار مقیاس در برآورد تغییرنگار تجربی است.

۱.۲ *برآوردگرهای استوار مقیاس*

برآوردگرهای استوار گوناگونی برای مقیاس معرفی شده‌اند، که شاخص‌ترین آنها برآوردگر MAD [۶] (هامپل، ۱۹۶۸)، S_n (روسیو و کروکس، ۱۹۹۲)، P_n (شامس، ۱۹۷۶) و Q_n (روسیو و کروکس، ۱۹۹۲) هستند. این چهار برآوردگر که به همراه خصوصیات آنها در جدول ۱ معرفی شده‌اند، همگی دارای تابع تاثیر کراندار و نقطه فروریزش بالا هستند، یعنی در برابر حجم‌های بالای داده دورافتاده تخریب نمی‌شوند. هرچند زمان محاسبات P_n و Q_n نسبت به S_n و MAD طولانی‌تر است، اما در وضعیت‌های مختلف این برآوردگرها عملکردهای مختلفی دارند.

عملکرد برآوردگرهای استوار را باید در دو وضعیت متفاوت سنجید. اولاً این برآوردگرها باید در وضعیتی که داده دورافتاده‌ای در میان مشاهدات دیده نمی‌شود،

[۶] Median Absolute Deviation

جدول ۱: برآوردگرهای استوار مقیاس

برآوردگر	رابطه	ε^*	γ^*	کارایی نسبی
MAD	$۱/۴۸۲۶med_i\|x_i - med_j(x_j)\|$	۰/۵۰	۱/۱۷	۳۶/۷۴
S_n	$۱/۱۹۲۶med_i\{med_j\|x_i - x_j\|\}$	۰/۵۰	۱/۶۳	۵۸/۲۳
P_n	$۱/۰۴۸۳med\{\|x_i - x_j\|;\ \ i<j\}$	۰/۲۹	۲/۳۹	۸۶/۰۰
Q_n	$۲/۲۱۹۱\{\|x_i - x_j\|;\ \ i<j\}_{(k)}$ *	۰/۵۰	۲/۰۷	۸۲/۲۷

* $k = \binom{h}{۲} \approx \frac{۱}{۴}\binom{n}{۲}$ ،$h = [\frac{n}{۲}] + ۱$

بتوانند کارایی نسبی مناسبی در مقایسه با برآوردگرهای معمول داشته باشند. ثانیاً در مواجهه با داده‌های دورافتاده تاثیرپذیری کمی داشته باشند. محمدی و محمدزاده (۱۳۹۰) در یک مطالعه شبیه‌سازی عملکرد برآوردگرهای استوار MAD، S_n، P_n و Q_n را با انحراف معیار نمونه‌ای مقایسه نمودند. طبق نتایج آنها در وضعیت داده‌های پاک برآوردگر P_n عملکرد مناسبی از خود نشان داده است که با توجه به کارایی نسبی بالای آن دور از انتظار نیست. اما برآوردگر Q_n با وجود کارایی نسبی بالا برای حجم‌های نمونه‌ای کوچک نتایج مناسبی از خود نشان نمی‌دهد (محمدی و محمدزاده، ۱۳۹۰ و راندال،۲۰۰۸). در مقابل و در حضور داده‌های دورافتاده، تمامی برآوردگرهای استوار بهتر از انحراف معیار نمونه‌ای عمل کرده‌اند. محمدی و محمدزاده (۱۳۹۰) پیشنهاد داده‌اند در شرایطی که حجم داده‌های دورافتاده زیاد نیست از برآوردگر P_n، و در مواجهه با حجم‌های بالاتر داده‌های دورافتاده از MAD و Q_n استفاده گردد. در ضمن استفاده از Q_n را تنها برای حجم‌های نمونه‌ای بالا مناسب دانسته‌اند. بطور کلی نتایج فوق بیانگر آن است که میزان تخریب برآوردگر SD حتی برای آلودگی‌های کم چشمگیر است و استفاده از جایگزین‌های استوار بسیار بهتر خواهد بود.

۲.۲ *برآورد استوار تغییرنگار*

دیدیم که برآوردگرهای تغییرنگار از جنس برآوردگرهای مقیاس برای نمونه تصادفی $\{V_۱(h), ..., V_{N_h}(h)\}$ هستند و از استواری چندانی برخوردار نمی‌باشند. لذا محمدی و محمدزاده (۱۳۹۰) با الهام از برآوردگرهای مقیاس مطرح شده در بخش

قبل، چهار برآوردگر استوار جدید برای تغییرنگار به صورت

$$\begin{aligned}
2\hat{\gamma}_{MAD}(h) &= (MAD(V_1(h),\ldots,V_{N_h}(h))^2 \\
2\hat{\gamma}_{S}(h) &= (S_{N_h}(V_1(h),\ldots,V_{N_h}(h))^2 \\
2\hat{\gamma}_{P}(h) &= (P_{N_h}(V_1(h),\ldots,V_{N_h}(h))^2 \\
2\hat{\gamma}_{Q}(h) &= (Q_{N_h}(V_1(h),\ldots,V_{N_h}(h))^2
\end{aligned}$$

پیشنهاد دادند، که همانند برآوردگرهای استوار مقیاس دارای توابع تاثیر کراندار هستند. علیرغم اینکه داده‌های دورافتاده فضایی چندین بار در محاسبه تغییرنگار حضور دارند، امّا همچنان این برآوردگرها از نقطه فروریزش بالایی برخوردارند. محمدی و محمدزاده (۱۳۹۰) برآوردگرهای استوار تغییرنگار را در یک مطالعه شبیه‌سازی مقایسه نموده و توانایی آنها در کنترل اثر داده‌های دورافتاده را تشریح نمودند. طبق نتایج آنها در وضعیت داده‌های پاک، برآوردگر ماترون بهترین نتایج را ارائه نموده است و نتایج مربوط به برآوردگرهای استوار $\hat{\gamma}_{CH}(h)$ و $\hat{\gamma}_{P}(h)$ و در مرحله بعد $\hat{\gamma}_{S}(h)$ نیز تا حد زیادی قابل قبول بوده است. در مواجهه با حجم پایین آلودگی برآوردگر ماترون $\hat{\gamma}_{M}(h)$ کاملاً تحت تاثیر آلودگی داده‌ها قرار گرفته است و برآوردگرهای $\hat{\gamma}_{CH}(h)$ و $\hat{\gamma}_{MAD}(h)$ بسیار خوب عمل کرده‌اند. با افزایش آلودگی نتایج برآوردگر ماترون بسیار نامناسب‌تر شده است، اما $\hat{\gamma}_{S}(h)$ و $\hat{\gamma}_{MAD}(h)$ بهترین نتایج را داشته‌اند. در پایان برآوردگر $\hat{\gamma}_{MAD}(h)$ برای حجم آلودگی‌های بالا و $\hat{\gamma}_{P}(h)$ و $\hat{\gamma}_{CH}(h)$ برای حجم‌های پایین داده دورافتاده توسط آنها پیشنهاد شده است.

۳ برآورد استوار روند

وجود روند در داده‌های فضایی باعث اریبی برآورد تغییرنگار می‌گردد. لذا باید قبل از برآورد تغییرنگار، روند موجود برآورد و حذف گردد. برای برآورد پارامترهای روند، که بصورت یک مدل خطی فضایی است، به علت وابستگی خطاها باید از روش کمترین توان‌های دوم تعمیم‌یافته (GLS) استفاده کرد. ولی این روش خود

نیازمند برآوردی از ماتریس کوواریانس است. لذا در عمل یک روش سه مرحله‌ای مورد استفاده قرار می‌گیرد: ابتدا به روش کمترین توان‌های دوم معمولی (OLS)، پارامترهای مدل برآورد شده و مانده‌های آن محاسبه می‌شوند. سپس با استفاده از این مانده‌ها تابع کوواریانس و به کمک آن ماتریس کوواریانس برآورد می‌گردد. در مرحله سوم برآوردهای کمترین توان‌های دوم تعمیم‌یافته (GLS) با استفاده از ماتریس کووایانس برآورد شده، بدست می‌آیند. این برآوردگر، هم در مرحله اول یعنی برازش به روش OLS و هم در مرحله سوم‌کاملاً تحت تاثیر داده‌های دورافتاده قرار دارد.

برای مدل‌های رگرسیونی با خطاهای ناهمبسته و هم‌توزیع، برآوردگرهای استوار زیادی ارائه شده است. برای استفاده از این برآوردگرها، لازم است مدل خطی فضایی به یک مدل با خطاهای ناهمبسته تبدیل شود. برای این منظور می‌توان تجزیه چولسکی ماتریس کوواریانس را بکار گرفت، که میلیتینو و همکاران (۲۰۰۳) با استفاده از آن، پارامترهای روند مدل خطی فضایی را برآورد نمودند. اگر Σ ماتریسی با درایه‌های $\mathrm{Cov}(Z(s_i), Z(s_j));\ s_i, s_j \in D$ باشد، تجزیه چولسکی ماتریس کوواریانس بفرم $\Sigma = LL' = LU$ می‌باشد، که در آن L یک ماتریس پایین مثلثی با درایه‌های قطری نامنفی و U ترانهاده آن است. در اینصورت مدل خطی فضایی $Z = F\beta + \epsilon$ را که در آن $Z_{n\times 1}$ بردار مشاهدات، $F_{n\times m}$ ماتریس توابع رگرسیونی، $\beta_{m\times 1}$ بردار پارامترهای مدل و $\epsilon_{n\times 1}$ بردار خطاهای همبسته هستند، می‌توان با ضرب طرفین در L^{-1} بصورت $\tilde{Z} = \tilde{F}\beta + \tilde{\epsilon}$ تبدیل کرد، که در آن $\tilde{Z} = L^{-1}Z$، $\tilde{F} = L^{-1}F$ و $\tilde{\epsilon} = L^{-1}\epsilon$ است. در این مدل $\mathrm{Cov}(\tilde{\epsilon}, \tilde{\epsilon}') = L^{-1}LUU^{-1} = I_{n\times n}$ خواهد بود. یعنی مدل تبدیل یافته دارای خطاهای ناهمبسته با واریانس ثابت ۱ است و در نتیجه می‌توان از تمام روش‌های برآورد استوار مدل‌های رگرسیونی برای آن استفاده نمود.

۴ پیشگویی استوار فضایی

در آمار فضایی پیشگویی میدان تصادفی $Z(\cdot)$ در موقعیت فاقد مشاهده $s_0 \in D$ براساس مشاهدات $Z = (Z(s_1), ..., Z(s_n))$ توسط کریگیدن انجام می‌شود. این پیشگو از دو طریق تحت تاثیر داده‌های دورافتاده قرار دارد. یکی تاثیر داده‌های دورافتاده در برآورد تغییرنگار است که ضرائب کریگیدن را تعیین می‌کند. و دیگری تاثیر مستقیم هر داده دورافتاده در محاسبه کریگیدن است. با بکارگیری برآوردگرهای استوار تغییرنگار کرسی و هاوکینز (۱۹۸۰)، جنتون (۱۹۹۸) و محمدی و محمدزاده (۱۳۹۰) می‌توان بخش اول تاثیرات داده‌های دورافتاده را کنترل نمود، اما برای پیشگویی استوار لازم است اثر مشاهدات آلوده نیز کنترل شود. هاوکینز و کرسی (۱۹۸۴) برای این کار اصلاح مشاهدات از طریق وینزوری کردن آنها را پیشنهاد نمودند. به این ترتیب که هر مشاهده $Z(s_i)$ به صورت

$$Z^{(e)}(s_i) = \begin{cases} Z^{@}_{-i}(s_i) + c\sigma_{-i}(s_i) & if \;\; Z(s_i) - Z^{@}_{-i}(s_i) > c\sigma_{-i}(s_i) \\ Z(s_j) & if \;\; |Z(s_i) - Z^{@}_{-i}(s_i)| \leq c\sigma_{-i}(s_i) \\ Z^{@}_{-i}(s_i) - c\sigma_{-i}(s_i) & if \;\; Z(s_i) - Z^{@}_{-i}(s_i) < -c\sigma_{-i}(s_i) \end{cases}$$

اصلاح می‌شود، که در آن ثابت c سطح وینزوری کردن، $Z^{@}_{-i}(s_i)$ میانه وزنی[۷] و $\sigma_{-i}(s_i)$ واریانس کریگیدن می‌باشند. با فراهم شدن مقادیر ویرایش شده مشاهدات، وزن‌های پیشگویی از برآورد استوار تغییرنگار محاسبه شده و مقدار پیشگویی به صورت $\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z^{(e)}(s_i)$ بدست می‌آید.

ایراد اصلی این روش تغییر و دستکاری مقادیر واقعی مشاهدات است که بی‌شک سبب کاهش دقت نتایج و کاهش کارایی پیشگویی می‌گردد. محمدی (۱۳۹۰) استفاده از پیشگویی استوار پیراسته را به عنوان راهکار مناسب پیشنهاد نمود که عملکرد بهتری از پیشگوی فوق دارد و در یک مطالعه شبیه‌سازی عملکرد مناسب آن را تشریح نمود.

---

۷ Weighted Median

۵ فرآیند تحلیل استوار داده‌های فضایی

به این ترتیب با جایگزین کردن روش‌های استوار، فرآیند کاری زیر توسط محمدی و محمدزاده (۱۳۹۰) برای تحلیل استوار داده‌های فضایی پیشنهاد شد، که در تمام مراحل آمادگی مقابله با داده‌های دورافتاده را دارد:

۱) *برآورد اولیه پارامترهای روند:* توسط یکی از برآوردگرهای استوار کمترین توان‌های دوم پیراسته[۸] (LTS) (روسیو، ۱۹۸۴)، ام-برآوردگرهای تعمیم‌یافته[۹] (GM-برآوردگرها) (مالوز، ۱۹۷۵، کواکلی و هتمنزپرگر، ۱۹۹۳)، پارامترهای روند برآورد شوند.

۲) *برآورد ماتریس کوواریانس:* تحت فرض ایستایی مرتبه دوم، با استفاده از مانده‌های مرحله قبل و توسط برآوردگرهای استوار تغییرنگار، ماتریس کوواریانس برآورد شود.

۳) *برآورد نهایی پارامترهای روند:* با استفاده از برآورد ماتریس کوواریانس برآوردهای استوار پارامترهای روند محاسبه شوند.

۴) *برآورد تغییرنگار:* با استفاده از برآورد استوار روند، داده‌ها را فاقد روند نموده، برآورد استوار تغییرنگار محاسبه شود.

۵) *پیشگویی فضایی:* با استفاده از برآورد استوار تغییرنگار، پیشگویی استوار هاوکینز و کرسی (۱۹۸۴) یا محمدی (۱۳۹۰) مانده‌ها را بدست آورده و از مجموع روند برازش داده شده در مرحله سوم و پیشگویی استوار مانده‌ها، پیشگویی استوار نهایی محاسبه شود.

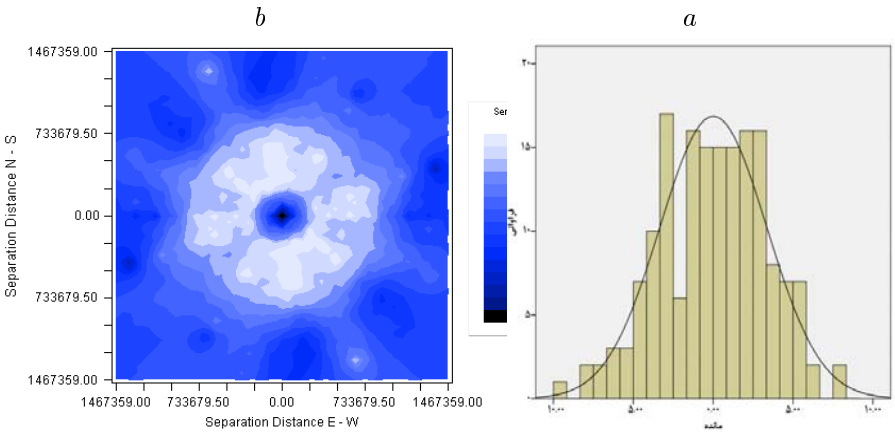
۶ مثال کاربردی

نمودار داده‌های متوسط دمای سالانه ۱۷۰ شهر ایران است (مرادی، ۱۳۸۶) در مقابل طول و عرض جغرافیایی (x و y)، بیانگر وجود روندی معنی‌دار در

[۸] Least Trimmed Squares

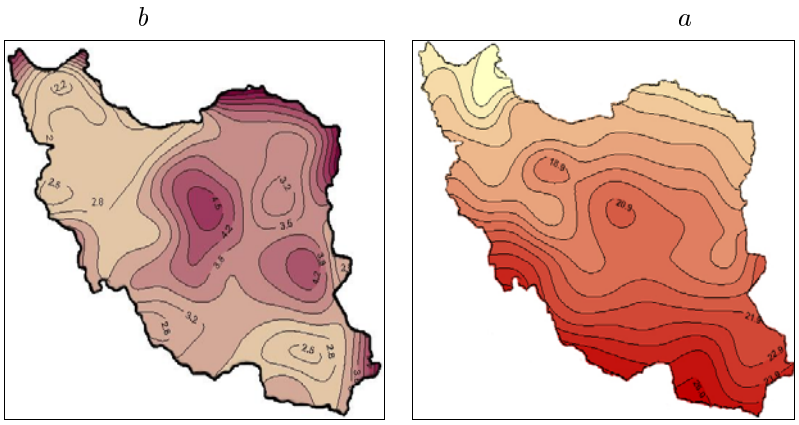
[۹] Generalized M-Estimator

مسیر شمالی-جنوبی است. برازش مدل‌های درجه اول، دوم و سوم به روش پیشرو[۱۰]، مدل $\mu(s) = \beta_0 + \beta_1 y;\ s = (x,y) \in$ با $R^2_{adj} = 0/711$ را به عنوان بهترین مدل نتیجه داده است، که در آن $D \subset R^2$ ناحیه مورد مطالعه است. برای برازش مدل روند مطابق سه مرحله اول فرآیند ارائه شده در بخش ۴، ابتدا برآوردهای کمترین توان‌های دوم پیراسته پارامترهای مدل روند محاسبه و مقادیر $\hat{\beta}_{LTS} = (349/87, -1/07 \times 10^{-5})'$ بدست آمده‌اند. سپس مدل معتبر تغییرنگار کروی به برآوردگر استوار $2\hat{\gamma}_S$ برازش داده و پارامترهای آن به صورت $\theta' = (542781/50, 8/83, 5/27)$ حاصل شده‌اند. آنگاه تحت فرض ایستایی مرتبه دوم، ماتریس کوواریانس با استفاده از برآوردگر استوار تغییرنگار محاسبه و تجزیه چولسکی آن صورت پذیرفته است. برای برآورد استوار پارامترهای روند ابتدا برآوردهای کمترین توان‌های دوم پیراسته (LTS) بصورت $\hat{\beta}_{LTS} = (373/69, -1/15 \times 10^{-5})'$ محاسبه و با استفاده از آنها GM-برآورد $\hat{\beta}_{GM} = (361/90, -1/10 \times 10^{-5})'$ حاصل شده است.



شکل ۱: (a) نمودار هیستوگرام داده‌های بدون روند شده (b) بررسی همسانگردی میدان تصادفی مانده‌ها

[۱۰] Forward



شکل ۲: (a) نقشه پیشگویی استوار داده‌های دما (b) نقشه واریانس پیشگویی داده‌های دما

براساس برآورد استوار روند، داده‌ها فاقد روند شده و با توجه به شکل ۱.a و نتیجه آزمون کولموگروف–اسمیرنوف (p-value=۰/۶۳۰)، مانده‌ها از توزیع نرمال پیروی می‌نمایند. شکل ۱.b نیز بیانگر همسانگرد بودن میدان تصادفی است. در مرحله چهارم، تغییرنگار مانده‌ها با استفاده از برآوردگر استوار $2\hat{\gamma}_S$ برآورد شده و مدل کروی با $\theta' = (۵۴۳۳۹۰/۵۷, ۹/۳۲, ۴/۹۴)$ برازش داده شده است. سپس با فرض گاوسی بودن میدان تصادفی مانده‌ها پیشگویی مقادیر مانده‌ها در هر موقعیت با استفاده از پیشگوی استوار هاوکینز و کرسی (۱۹۸۴) محاسبه شده است. پیشگویی‌های نهایی در هر موقعیت، از مجموع روند برآورد شده و پیشگویی استوار ماندهٔ آن موقعیت حاصل شده است. شکل‌های ۲.a سطح تراز نقشه آماری پیشگویی متوسط دما را در کشور نشان می‌دهد. همانطور که ملاحظه می‌شود متوسط دمای هوا از مناطق شمالی کشور به سمت مناطق کویری و جنوب کشور افزایش یافته است. واریانس‌های پیشگویی‌های کشور در شکل ۲.b نشان دهنده آنستکه پیشگویی‌ها در محدوده دشت کویر و کویر لوت و نقاط مرزی ایران که مشاهدات کمی در اختیار است دارای واریانس بالایی هستند، اما پیشگویی‌ها در سایر نقاط از دقت بالایی برخوردارند.

۷ بحث و نتیجه‌گیری

استفاده از روش‌های استوار برای کنترل اثر داده‌های دورافتاده در تحلیل‌های آماری بسیار مورد توجه است. در این مقاله سعی شد روش‌های استوار مناسب برای تحلیل داده‌های فضایی مورد بحث و بررسی قرار گیرد. برآوردگرهای استوار تغییرنگار در گام اول معرفی و نقاط و ضعف آنها بطور کامل تشریح گردید. روش‌های استوار برازش روند و پیشگویی استوار فضایی نیز مورد بحث قرار گرفتند.

در مجموع توجه به این نکته ضروری است که روش استوار مناسب باید با در نظر گرفتن حجم داده‌های دورافتاده و تطبیق آن با خواص روش‌های استوار موجود صورت گیرد. استفاده از روش‌های استوار با نقطه فروریزش بالا، معادل پایین آمدن کارایی روش‌ها خواهد بود. لذا با یک تحلیل اکتشافی مناسب می‌توان روش‌های استواری را برای تحلیل داده‌های فضایی برگزید که هم‌زمان با کنترل اثر داده‌های دورافتاده کارایی مناسبی را فراهم کنند.

مراجع

محمدی، ا. (۱۳۹۰)، **پیشگویی استوار داده‌های فضایی**، دهمین کنفرانس بین‌المللی آمار ایران، دانشگاه تبریز.

محمدی، ا. و محمدزاده، م. (۱۳۹۰)، **تحلیل استوار داده‌های فضایی در حضور داده‌های دورافتاده**، ارسال برای چاپ.

مرادی، ایوب، (۱۳۸۶)، **مکان‌گزینی مناطق معرف برای مطالعات سنجش از دور، مطالعه موردی: ایران**، پایان‌نامه کارشناسی ارشد، دانشگاه تربیت مدرس.

Armstrong, M. and Boufassa, A. (1988), Comparing the Robustness of Ordinary Kriging and Lognormal Kriging: Outlier Resistance. *Mathematical Geology*, **20**, 447-457.

Berke, O. (2001), Modified Median Polish Kriging and its Application to the Wolfcamp-Aquifer Data. *Environmetrics*, **12**, 731-748.

Coackley, C. W. and Hettmansperger, T. P. (1993). A Bounded Influence, High Brakdown, Efficient Regression Estimator. *Journal of the American Statistical Association*, **88**, 872-880.

Cressie N. A. (1993). Statistics for Spatial Data, Weily, New York.

Cressie N. A. , Hawkins D. (1980). Robust Estimation of the Variogram. *Mathematical Geology*, **12**, 115-125.

Genton, M. G. (1998), Highly Robust Variogram Estimation. *Mathematical Geology*. **130**, 213-221.

Hampel, F. R. (1968), Contributions to the Theory of Robust Estimation. Ph.D. thesis. University of California, Berkeley.

Hawkins, D. and Cressie N. A. (1984), Robust Kriging–A Proposal. *Journal of the International Association for Mathematical Geology*, **16**, 3-18.

Huber, P. J. (1981). *Robust Statistics*, Weily, New York.

Mallows, C. L. (1975), On Some Topics in Robustness. *Technical Memorandum*, Bell Telephone Laboratories. Murray Hill.

Matheron, G. (1962), Traite de Geostatistique Appliquee, Tome I. *Memoires du Bureau de Recherches Geologiques et Minieres*, **14**, Editions: Technip, Paris.

Militino, A. F. and Ugrate, M. D. (1997), A GM Estimation of the Location Parameters in a Spatial Linear Model. *Communications in Statistics, Theroy and Methods*, **26**, 1701-1725.

Militino, A. F. , Palacios, M. B. and Ugrate, M. D. (2003), Robust Trend Parameters in a Multivariate Spatial Linear Model. *Sociedad de Estadistica e lnvestigacion Operativa Test*, **12**, 445-457

Ronald, J. (2008). A Reinvestigation of Robust Scale Estimation in Finite Samples. *Computational Statistics and Data Analysis*, **52**, 5014-5021.

Rousseeuw, P. J. and Croux, C. (1992), Explicit Scale Estimators with High Brekdown Point. L_1*-Statistical Analysis and Related Methods*, editors: Dodge, Y.

Rousseeuw, P. J., and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**, 1273-1283.

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱
مجموعه مقالات، ص ۲۳۱-۱۵۹

تحلیل بیزی مقادیر کرانگین فضایی

بهزاد محمودیان، محسن محمدزاده
گروه آمار، دانشگاه تربیت مدرس

چکیده : در این مقاله مدل فضایی برای تحلیل مقادیر کرانگین با رهیافت بیزی معرفی می‌شود. برای این منظور فرض می‌شود که مقادیر کرانگین مشروط بر پارامتر مکان توزیع مستقل از یکدیگر با توزیع مقدار کرانگین تعمیم‌یافته باشند. سپس میدان تصادفی پنهان با پارامتر مکان توزیع مرتبط شده و تغییرات فضایی بزرگ و کوچک مقیاس مقادیر کرانگین به ترتیب به صورت روند و مولفهٔ خطای همبسته فضایی افراز می‌گردد. همچنین استنباط بیزی و پیشگویی مقدار کرانگین با توجه به مدل فضایی ارائه شده در تحلیل ماکسیمم‌های سرعت باد ایران به کار می‌رود.

واژه‌های کلیدی : مقادیر کرانگین فضایی، توزیع مقدار کرانگین تعمیم‌یافته، میدان تصادفی، استنباط بیزی، پیشگویی.

آدرس الکترونیک مسئول مقاله: نام نویسنده، mahmoudian@modares.ac.ir
کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۰G۳۲، ۶۲M۳۰، ، ۶۲H۱۱

۱ مقدمه

مقادیر کرانگین به مشاهداتی اطلاق می‌شود که دم توزیع را توصیف می‌کنند. در روش ماکسیمم بلوکی پس از تعریف بلوک‌های زمانی به بزرگترین مشاهده در هر بلوک یک ماکسیمم بلوکی اطلاق شده و در صورتی که مشاهدات داخل بلوک بر حسب موقعیت‌های فضایی وابسته باشند این ماکسیمم‌ها دارای همبستگی فضایی شده و مقادیر کرانگین (ماکسیمم‌های بلوکی) فضایی نامیده می‌شوند. تحلیل مقادیر کرانگین فضایی می‌تواند با رویکردی سلسله مراتبی و مرتبط نمودن میدان تصادفی با پارامترهای توزیع مانند مدل‌های آمیختهٔ خطی تعمیم‌یافته طرح‌ریزی شود که پیشگویی مقدار کرانگین در موقعیت‌های فاقد مشاهده هدف اصلی آن است.

در تحلیل مقادیر کرانگین فضایی می‌توان به مدل‌های پیشنهادی توسط کسن و کلز (۱۹۹۹) و کولی و همکاران (۲۰۰۷) اشاره کرد. کسن و کلز (۱۹۹۹) با استفاده از فرایند نقطه‌ای مدل رگرسیونی فضایی را با رهیافت بیزی برای فزونی‌های سرعت باد به کار بردند. کولی و همکاران (۲۰۰۷) مدل رگرسیونی زمین‌آمار را برای فزونی‌های بیش از سرحد که دارای توزیع پارتوی تعمیم‌یافته هستند مشابه با مدل‌های زمین‌آمار دیگل و همکاران (۱۹۹۸) پیشنهاد کردند. در واقع آن‌ها میدان تصادفی گاوسی را با پارامتر مقیاس توزیع پارتوی تعمیم‌یافته مرتبط و به پیشگویی فزونی‌های مقدار بارش پرداختند. در این مقاله میدان تصادفی گاوسی برای پارامتر مکان توزیع مقدار کرانگین تعمیم‌یافته در نظر گرفته می‌شود. مولفه‌های روند و خطا در فرم افراز شده میدان تصادفی به ترتیب با تابع چند جمله‌ای درجه دو از موقعیت‌های فضایی و تابع کواریانس نمایی مدل‌بندی می‌شوند. تحلیل بیزی مدل فضایی با تعیین توزیع‌های پیشین و نحوهٔ شبیه‌سازی از توزیع‌های شرطی کامل بیان می‌گردد. در پایان ضمن برآورد پارامترهای مدل فضایی با رهیافت بیزی، پیشگویی فضایی بیزی ماکسیمم‌های سرعت باد ایران ارائه می‌گردد.

در ادامه در بخش ۲ توزیع مجانبی مقادیر کرانگین به اختصار معرفی می‌شود. در بخش ۳ مدل فضایی برای لحاظ کردن همبستگی‌های فضایی ارائه می‌گردد. تکنیک‌های مونت کارلوی زنجیر مارکوفی به همراه پیشگوی فضایی بیزی در بخش

۴ شرح داده می‌شود. بخش ۵ به نحوهٔ برآورد پارامترها در رهیافت بیزی با تحلیل داده‌های واقعی سرعت باد می‌پردازد. در نهایت بحث و نتیجه‌گیری در بخش ۶ بیان می‌شود.

۲ توزیع مقدار کرانگین تعمیم‌یافته

برخلاف روش‌های کلاسیک آماری که در آن رفتار توزیع در اطراف میانگین بررسی می‌شود، نظریهٔ مقادیر کرانگین به تحلیل رفتار دم توزیع‌ها می‌پردازد. در مدل‌بندی ماکسیمم بلوکی مشاهدات، ابتدا بلوک‌های زمانی به صورت روزانه، سالانه یا بازه‌های زمانی دیگر تعریف و به ماکسیمم بلوک‌ها توزیع مقدار کرانگین تعمیم‌یافته برازش می‌شود. فرض کنید $\{Y_t\}_{t\geq 1}$ دنباله‌ای از متغیرهای تصادفی مستقل و هم‌توزیع با تابع توزیع معلوم $F(y)$ و $M_n = \max_{t=1,\ldots,n} Y_t$ ماکسیمم بلوکی باشند. برای تعیین توزیع پارامتری و مجانبی ماکسیمم‌ها با قبول دو فرض هم‌توزیعی و استقلال، ثابت‌های حقیقی $a_n > 0$ و b_n را طوری می‌یابیم که توزیع $P(\frac{M_n - b_n}{a_n} \leq y)$ در صورت افزایش n $(n \to \infty)$ به تابع توزیع غیرتباهیده $G(y)$ همگرا شود. قضیه انواع کرانگینی (کلز، ۲۰۰۱) نشان می‌دهد که ماکسیمم دنباله‌ای از متغیرهای تصادفی مستقل و هم‌توزیع با ثابت‌های نرمال‌کننده، در صورت وجود، از یکی از توزیع‌های گامبل، فره‌شه و وایبل پیروی می‌کند. برای رسیدن به مدل واحد با پارامترسازی مجدد، توزیع مقدار کرانگین تعمیم‌یافته[۱] (GEV) به صورت

$$G(y) = exp\Big\{ -\Big[1 + \xi(\frac{y-\mu}{\sigma})\Big]_+^{-1/\xi} \Big\}, \qquad (1)$$

به دست می‌آید، که در آن $[x]_+ = \max\{0, x\}$ است. تابع توزیع (۱) دارای سه پارامتر مکان $\mu \in R$، مقیاس $\sigma > 0$ و شکل $\xi \in R$ است. پارامتر شکل ξ رفتار دم توزیع را توصیف می‌کند، $\xi > 0$ توزیع فره‌شه، $\xi < 0$ توزیع وایبل و $\xi = 0$ توزیع گامبل را نتیجه می‌دهد. $\xi > 0$ توزیعی دم کلفت با کاهش پذیری چندجمله‌ای، $\xi = 0$ $(\xi \to 0)$ دم متوسط با کاهش پذیری نمایی و $\xi < 0$ توزیعی با دم باریک و کران بالایی در نقطه $y = \mu + \xi/\sigma$ را مشخص می‌کند.

[۱] Generalized extreme value

تابع توزیع GEV برای مینیمم‌های بلوکی با تعریف $m_n = \min_{t=1,\ldots,n} Y_t$ و رابطهٔ آن با تابع توزیع ماکسیمم‌ها $P(\frac{m_n - d_n}{c_n} \leq y) \to 1 - G(-y)$ بر مجموعهٔ $\{y : 1 - \xi(y-\mu)/\sigma > 0\}$ قابل تعریف است،

$$G(y) = 1 - exp\Big\{ - \Big[1 - \xi(\frac{y-\mu}{\sigma})\Big]_+^{-1/\xi} \Big\}. \qquad (2)$$

۳ مدل فضایی

اگر y_i به ازای $i = 1, \ldots, n$ ماکسیمم مشاهدات سری زمانی در یک بلوک زمانی (مثلاً سال) و موقعیت فضایی s_i باشد، انتظار می‌رود همبستگی فضایی بین مشاهدات در موقعیت‌های فضایی به ماکسیمم‌ها نیز انتقال پیدا کند. این امر موجب می‌شود که پذیرهٔ استقلال مشاهدات در نتایج مجانبی برای ساخت توزیع GEV برقرار نباشد. این نتایج مجانبی می‌توانند با فرض توزیع GEV با پارامتر مکانی که نسبت به موقعیت‌های فضایی تغییر می‌کند و استقلال مشاهدات مشروط بر تغییرات پارامتر مکان دوباره به کار گرفته شوند. به عبارت بهتر با فرض استقلال شرطی ماکسیمم‌های بلوکی مشروط بر پارامتر مکانی که با میدان تصادفی پنهانی مرتبط شده باز هم می‌توان از توزیع GEV استفاده نمود. فرض استقلال شرطی به این مفهوم است که Y_i، ماکسیمم بلوکی در موقعیت i، مشروط بر μ_i از Y_j مشروط بر μ_j برای $i \neq j$ مستقل است. تابع درستنمایی ماکسیمم‌ها در موقعیت‌های $s_1, \ldots s_n$ تحت فرض استقلال شرطی از رابطه (۱) به صورت

$$L(\boldsymbol{\mu}, \sigma, \xi; \boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sigma}\Big[1 + \xi\frac{y_i - \mu_i}{\sigma}\Big]^{-1-1/\xi} exp\Big\{ - \Big[1 + \xi\frac{y_i - \mu_i}{\sigma}\Big]^{-1/\xi} \Big\}, \quad (3)$$

محاسبه می‌شود.

معمولاً برای تحلیل داده‌های فضایی میدان تصادفی گاوسی به دلیل سادگی محاسبات و استفاده از تکنیک‌های کلاسیک زمین‌آمار به کار گرفته می‌شود. یک میدان تصادفی باید در شرایطی چون پایایی تحت جایگشت[۲] و پایداری احتمالی[۳]

[۲] Permutation Invaiance

[۳] Probability Consistency

صدق کند (یاگلم، ۱۹۶۲).

تعریف ۱. میدان تصادفی $\{Z(\boldsymbol{s}); \boldsymbol{s} \in \boldsymbol{D} \subset \boldsymbol{R}^d\}$ گاوسی نامیده می‌شود اگر برای همهٔ موقعیت‌های فضایی $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n) \in \boldsymbol{D} \times \ldots \times \boldsymbol{D}$

$$\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))' \sim N_n(\boldsymbol{m}, \boldsymbol{C}), \qquad (۴)$$

با تابع چگالی احتمال زیر باشد،

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{C}|^{1/2}} exp\Big\{ -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{m})'\boldsymbol{C}^{-1}(\boldsymbol{z}-\boldsymbol{m})\Big\}.$$

در (۴)، $N_n(\boldsymbol{m}, \boldsymbol{C})$ توزیع نرمال چند متغیره با بردار میانگین $\boldsymbol{m}$ و ماتریس کواریانس $\boldsymbol{C}$ را نمایش می‌دهد. بردار میانگین و ماتریس کواریانس معین مثبت توسط توابع میانگین $m(\boldsymbol{s})$ و کواریانس $c(\boldsymbol{s}, \boldsymbol{s}')$ مشخص می‌شوند.

مدل سلسله مراتبی با فرض استقلال شرطی و استفاده از توزیع GEV و میدان تصادفی گاوسی برای تحلیل مقادیر کرانگین قابل طرح است. در مرحلهٔ اول برای مشاهدات توزیع GEV به صورت (۳) با فرض استقلال شرطی در نظر گرفته می‌شود. در مرحلهٔ بعد مدل سلسله مراتبی، مدل

$$\mu(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + W(\boldsymbol{s}), \qquad (۵)$$

برای پارامتر مکان توزیع در نظر گرفته می‌شود، که در آن $\boldsymbol{x}(\boldsymbol{s})$ بردار متغیرهای تبیینی در موقعیت فضایی $\boldsymbol{s}$، $\boldsymbol{\beta}$ بردار ضرایب رگرسیونی و $W(\boldsymbol{s})$ میدان تصادفی گاوسی با میانگین صفر و تابع همبستگی $\rho(\boldsymbol{s}_i - \boldsymbol{s}_j; \phi)$ هستند. براساس مدل (۵) می‌توان پیشین فضایی $N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$ را برای بردار $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))'$ در نظر گرفت، که در آن $\boldsymbol{V}$ ماتریس کواریانس فضایی با درآیه‌های $\boldsymbol{V}_{ij} = \sigma_\mu^2 \rho(\boldsymbol{s}_i - \boldsymbol{s}_j; \phi)$ است. تغییرات بزرگ مقیاس با $\boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta}$ بیان می‌شود که در این مقاله از روند چند جمله‌ای درجه دو برای مدل‌بندی آن با $\boldsymbol{s} = (s_1, s_2)$ به صورت

$$\boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_1^2 + \beta_4 s_1 s_2 + \beta_5 s_2^2, \qquad (۶)$$

استفاده می‌شود. در (۶) متغیرهای تبیینی استاندارد شده‌اند تا موجب مقیاس-پایا شدن توزیع‌های پیشینی پارامترها و همگرایی مناسب الگوریتم مونت کارلوی زنجیر

مارکوفی گردد. توابع همبستگی متفاوتی نظیر کروی، نمایی توانی و ماترن برای مدل‌بندی ساختار همبستگی داده‌های همبستهٔ فضایی پیشنهاد شده است (بانرجی، ۲۰۰۴). این توابع مستقیماً خصوصیات میدان تصادفی تحت مطالعه را تعیین می‌کنند که در عمل سادگی فرم تابعی از نظر تعداد پارامترها و انعطاف‌پذیری آن‌ها ملاک انتخاب قرار می‌گیرد. خانواده توابع ماترن دارای پارامترهای دامنه (مقیاس) و همواری است که دامنهٔ همبستگی و همواری میدان تصادفی توسط این پارامترها کنترل می‌شود. توابع همبستگی نمایی و گاوسی جزو خانواده ماترن هستند، این توابع میدان‌های تصادفی را توصیف می‌کنند که تحقق‌های آن به ترتیب ویژگی مشتق‌پذیری را ندارند و یا بی‌نهایت بار مشتق‌پذیرند. در اینجا برای سادگی تابع همبستگی نمایی به صورت

$$\rho(d_{ij};\phi) = exp(-d_{ij}/\phi),$$

انتخاب شده است، که در آن $d_{ij} = ||\boldsymbol{s}_i - \boldsymbol{s}_j||$ بیانگر فاصلهٔ بین دو موقعیت $\boldsymbol{s}_i$ و $\boldsymbol{s}_j$، $||\cdot||$ فاصلهٔ اقلیدسی و ϕ پارامتر دامنه می‌باشند. پارامتر دامنه سرعت نزول همبستگی با افزایش فاصله بین موقعیت‌های فضایی را کنترل می‌کند.

۴ استنباط و پیشگویی

برای برازش مدل با رهیافت بیزی توزیع‌های پیشینی در صورت امکان مزدوج، فاقد اطلاع و سره انتخاب می‌شوند. در این صورت توزیع پسینی پارامترهای مدل (۵) به صورت

$$p(\boldsymbol{\mu},\sigma,\xi,\boldsymbol{\beta},\sigma_\mu^2,\phi|\boldsymbol{y}) \propto L(\boldsymbol{\mu},\sigma,\xi;\boldsymbol{y})p(\boldsymbol{\mu}|\boldsymbol{\beta},\sigma_\mu^2,\phi)p(\boldsymbol{\beta}|\sigma_\mu^2)p(\sigma_\mu^2)p(\sigma)p(\xi)p(\phi),$$

حاصل می‌شود، که در آن $L(\cdot)$ تابع درستنمایی حاصل از توزیع GEV در (۳) است. برای پارامترهای ثابت توزیع مقدار کرانگین، $\log(\sigma)$ و ξ، توزیع پیشینی نرمال با میانگین صفر و واریانس نسبتاً بزرگ انتخاب شده است. لذا توزیع‌های شرطی کامل پارامترهای $\log(\sigma)$ و ξ به ترتیب از حاصل‌ضرب تابع درستنمایی (۳) در توزیع‌های پیشینی مفروض حاصل می‌شوند، که فرم بسته‌ای ندارند. نمونه‌گیری از توزیع‌های

شرطی کامل این پارامترها با الگوریتم متروپولیس–هستینگس قدم‌زدن تصادفی صورت می‌پذیرد.

انتخاب توزیع پیشینی برای پارامترهای ϕ و σ_μ^2 نیاز به دقت خاصی دارد. پیشین‌های ناآگاهی‌بخش موجب عدم همگرایی یا تاخیر در همگرایی الگوریتم مونت کارلوی زنجیر مارکوفی می‌گردد. در نتیجه به جای توزیع پیشینی ناآگاهی‌بخش، توزیع نرمال بریده شده برای پارامتر ϕ و توزیع گامای معکوس مبهم برای σ_μ^2 انتخاب شده است. توزیع نرمال بریده شده با $TN_{(d_{min},d_{max})}(c,e)$ نمایش داده می‌شود، که در آن c با توجه به قاعدهٔ دامنه کاربردی[۴] انتخاب می‌شود. براساس این قاعده انتظار می‌رود که در نصف بیشترین فاصله‌های فضایی، d^*، همبستگی فضایی تقریباً صفر شود. همچنین پارامتر مقیاس e برابر مقدار کوچکی در نظر گرفته می‌شود تا توزیع پیشینی بر روی c تا حدودی متمرکز باشد. پارامترهای d_{min} و d_{max} هم برابر کوچکترین و چارک سوم فاصله‌های فضایی در نظر گرفته می‌شود. توزیع پیشینی گامای معکوس برای σ_μ^2 به صورت $IG(a,b)$ با پارامترهای شکل و نرخ کوچک در نظر گرفته می‌شود. تابع چگالی این توزیع متناسب با $x^{-(a+1)}e^{-bx}$ است و به گونه‌ای پارامترگذاری شده که $E(\sigma_\mu^2)=\frac{b}{a-1}$ و $Var(\sigma_\mu^2)=\frac{b^2}{(a-1)^2(a-2)}$.

با انتخاب توزیع نرمال چند متغیره $N_p(\boldsymbol{\mu}_\beta,\sigma_\mu^2\boldsymbol{\Sigma}_\beta)$، گامای معکوس $IG(a,b)$ و نرمال بریده شده $TN_{(d_{min},d_{max})}(c,d)$ به ترتیب برای ضرایب رگرسیونی $\boldsymbol{\beta}|\sigma_\mu^2$، σ_μ^2 و ϕ توزیع‌های شرطی کامل آن‌ها به صورت

$$
\begin{aligned}
&p(\boldsymbol{\beta}|\boldsymbol{\mu},\sigma_\mu^2,\phi)=N_p(\boldsymbol{K}(\boldsymbol{X}'\boldsymbol{H}(\phi)^{-1}\boldsymbol{\mu}+\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta),\sigma_\mu^2\boldsymbol{K}),\\
&\boldsymbol{K}=(\boldsymbol{X}'\boldsymbol{H}(\phi)^{-1}\boldsymbol{X}+\boldsymbol{\Sigma}_\beta^{-1})^{-1},\\
&p(\sigma_\mu^2|\boldsymbol{\mu},\boldsymbol{\beta},\phi)=IG\big(a^*,b^*\big),\qquad a^*=a+(n+p)/2,\qquad b^*=b+\\
&(\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{H}(\phi)^{-1}(\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta})/2+(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)'\boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)/2,\\
&p(\phi|\boldsymbol{\mu},\boldsymbol{\beta},\sigma_\mu^2)\propto\frac{1}{|\boldsymbol{H}(\phi)|^{1/2}}e^{-\frac{1}{2\sigma_\mu^2}(\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{H}(\phi)^{-1}(\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta})}f_{TN}(\phi),
\end{aligned}
$$

[۴] Practical Range

خواهند بود، که در آن‌ها $\boldsymbol{H}(\phi)$ ماتریس همبستگی فضایی با درآیه‌های $\boldsymbol{H}_{ij} = \rho(\boldsymbol{s}_i - \boldsymbol{s}_j; \phi)$ و $f_{TN}(\cdot)$ بیانگر توزیع نرمال بریده شده هستند. نمونه‌گیری از دو توزیع شرطی کامل اول با گام‌های نمونه‌گیر گیبز و از توزیع $(\phi|\boldsymbol{\mu}, \boldsymbol{\beta}, \sigma_\mu^2)$ با الگوریتم متروپولیس-هستینگس قدم‌زدن تصادفی انجام می‌شود.

برای نمونه‌گیری از توزیع شرطی کامل بردار پارامتر مکان $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))'$ که وابستگی فضایی را مدل‌بندی می‌کنند می‌توان از الگوریتم متروپولیس-هستینگس قدم‌زدن تصادفی استفاده کرد. لذا توزیع شرطی کامل هر یک از مولفه‌های بردار $\boldsymbol{\mu}$ به صورت

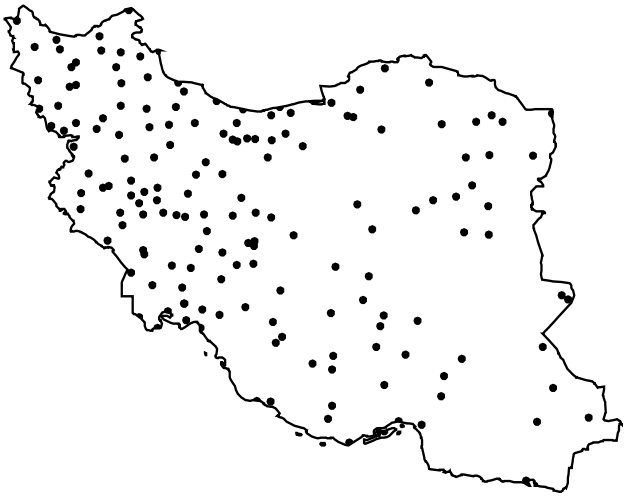
$$\begin{aligned} p(\mu(\boldsymbol{s}_i)|\boldsymbol{\theta}_-, \boldsymbol{y}) &\propto L(\mu_i, \sigma, \xi, y_i) p(\mu(\boldsymbol{s}_i)|\boldsymbol{\mu}_{-i}, \boldsymbol{\beta}, \sigma_\mu^2, \phi) \\ &\propto L(\boldsymbol{\mu}, \sigma, \xi, \boldsymbol{y}) f_N(\mu_{ij}|m_i, s_i), \qquad i = 1, \ldots, n \end{aligned}$$

به دست می‌آید، که در آن $\boldsymbol{\mu}_{-i} = (\mu_1, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n)$، $\boldsymbol{\theta}_-$ بیانگر تمام پارامترها به جز $\boldsymbol{\mu}(\boldsymbol{s}_i)$ و $f_N(\cdot|\mu, \sigma^2)$ بیانگر توزیع نرمال با میانگین μ و واریانس σ^2 می‌باشند. همچنین m_i و s_i توسط میانگین و واریانس توزیع شرطی نرمال چند متغیره قابل حصول است.

فرض کنید پیشگویی در موقعیت $\boldsymbol{s}_0 \in \boldsymbol{D}$ مورد نظر است. در رهیافت بیزی پیشگویی فضایی براساس توزیع پیشگوی

$$p(y(\boldsymbol{s}_0)|\boldsymbol{y}) = \int p(y(\boldsymbol{s}_0)|\boldsymbol{\Theta}, \boldsymbol{y}) p(\boldsymbol{\Theta}|\boldsymbol{y}) d\boldsymbol{\Theta}, \qquad (۷)$$

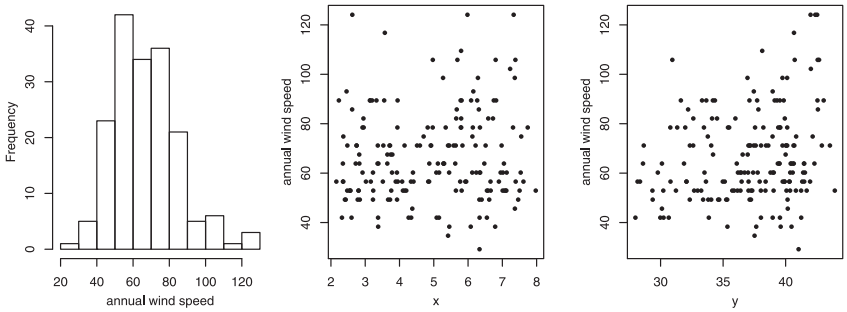
انجام می‌گیرد. محاسبه انتگرال در رابطهٔ (۷) اغلب غیر ممکن یا با دشواری‌هایی همراه است. اگر نمونه‌های مونت کارلوی زنجیر مارکوفی از $p(\boldsymbol{\Theta}|\boldsymbol{y})$ در اختیار باشد، می‌توان با تولید نمونه از توزیع $p(y(\boldsymbol{s}_0)|\boldsymbol{\Theta}, \boldsymbol{y})$ انتگرال رابطهٔ (۷) را با مجموع تقریب زد. نمونه‌گیری از توزیع پیشگوی $Y(\boldsymbol{s}_0)$ با تولید $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \sigma, \xi, \boldsymbol{\beta}, \sigma_\mu^2, \phi)$ از $p(\boldsymbol{\Theta}|\boldsymbol{y})$ با تکنیک‌های مونت کارلوی زنجیر مارکوفی بیان شده و پیشگویی $\boldsymbol{\mu}(\boldsymbol{s}_0)$ از میدان تصادفی گاوسی $\boldsymbol{\mu}(\boldsymbol{s})$ محاسبه می‌شود (دیگل و همکاران ۱۹۹۸).



شکل ۱: موقعیت فضایی ایستگاه‌های سینوپتیک هواشناسی.

۵ تحلیل ماکسیمم‌های سرعت باد ایران

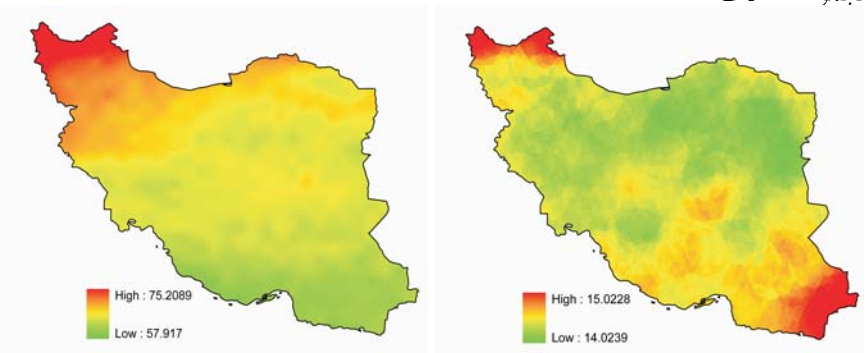
تحلیل داده‌های سرعت باد در بررسی وقوع بادهای طوفانی برای مهندسین در ساخت سازه‌های ساختمانی مثل پل‌ها از اهمیت خاصی برخوردار است (ولشو و اندرسن، ۲۰۰۰). در بعضی از ناحیه‌های جغرافیایی به دلیل وزش تندباد، حرکت قطارها با سرعت بالا با مخاطرات واژگونی و از خط خارج شدن همراه است. این مخاطره به زاویه و سرعت باد بستگی دارد (پیر و کوکن‌هاف، ۲۰۰۴). در برخی از نواحی ایران وزش باد با سرعت بالای ۱۰۰ کیلومتر بر ساعت توسط سازمان هواشناسی ثبت گردیده است. چنین پدیده‌ای می‌تواند به تاسیسات برق‌رسانی مثل پست‌ها یا خطوط انتقال برق آسیب جدی وارد کند. همچنین سرعت بالای باد با به حرکت درآوردن گرد و غبار برای مدت‌ها موجب پایین آمدن میدان دید می‌شود. این امر بخصوص برای پروازهای هوایی در ساعات خاص مهم هستند. تحلیل داده‌های سرعت باد علاوه بر مواردی که ذکر گردید از نقطه نظر کاربردی اهمیت بسیاری دارد. به عنوان مثال فرسایش خاک به دلیل از بین بردن باروری اراضی کشاورزی و تبدیل شدن آنها به بیابان معضل جهانی است. فرسایش بادی فرآیندی است که در آن ذرات خاک از سطح زمینی که قابلیت فرسایش دارد جدا شده و به



شکل ۲: بافت‌نگار و نمودار پراکنش ماکسیماهای سرعت باد در برابر طول و عرض جغرافیایی.

مکان دیگری انتقال می‌یابند. فرسایش بادی یک از مشکلات عمده در مناطق خشک و نیمه خشک است که تحت تأثیر عواملی نظیر سست بودن خاک، ریز و خشک بودن خاکدانه‌ها، فقدان پوشش گیاهی، وجود دشت گسترده و بادهای قوی شکل می‌گیرد. آسیب‌های بادی در دو نوع کلی درون و برون منطقه‌ای موجب کاهش مواد آلی، کاهش ظرفیت نگهداری آب، تغییر بافت، تخریب گیاه در خاک، آلودگی آب‌های آشامیدنی، آلودگی هوا، رسوب ذرات خاک در پشت سدها و کاهش میزان دید در بزرگراه‌ها و فرودگاه‌ها می‌شود. استان‌هایی که در ایران در معرض شدید فرسایش بادی قرار دارند، عبارتند از استان‌های مرکزی، سمنان، خراسان، سیستان‌وبلوچستان، کرمان، هرمزگان، بوشهر، خوزستان، فارس، اصفهان و یزد. بنابراین توانایی پیشگویی دقیق فرسایش بادی خاک برای برنامه‌های حفاظتی، منابع طبیعی و کاهش آلودگی هوا ناشی از طوفان ضروری است. از آنجایی که قدرت فرسایندگی باد بستگی به توان سوم سرعت باد دارد، بررسی و مدل‌بندی داده‌های سرعت باد در هر منطقه حائز اهمیت است (میرزامصطفی و همکاران، ۱۳۸۲).

داده‌هایی که در این مقاله تحلیل می‌شوند به صورت ماکسیمم سرعت باد در ۱۷۷ ایستگاه سینوپتیک هواشناسی ایران برای سال ۱۳۸۹ با واحد کیلومتر بر ساعت می‌باشد. شکل ۱ موقعیت قرار گرفتن ۱۷۷ ایستگاه سینوپتیک و شکل ۲ بافت‌نگار ماکسیمم‌های سرعت باد و نمودار پراکنش ماکسیمم‌ها در مقابل طول و عرض جغرافیایی مختصات مکانی ایستگاه‌ها را نشان می‌دهند. همان‌طور که در شکل ۲



شکل ۳: پهنه‌بندی مقدار کرانگین سرعت باد (سمت راست) و انحراف استاندارد آن (سمت چپ).

ملاحظه می‌شود ماکسیمم‌های سرعت باد توزیع چوله به راست دارند. همچنین مشاهدات دارای روندی در برابر طول و عرض جغرافیایی هستند که از آن‌ها به عنوان متغیرهای تبیینی در روند چندجمله‌ای (۶) استفاده می‌شود. با در نظر گرفتن توزیع‌های پیشینی به صورت

$$\xi, \log(\sigma) \sim N(۰, ۱۰^۴), \quad \phi \sim TN_{(۰/۰۱,۶/۹۲)}(۲/۷۸, ۱),$$
$$\boldsymbol{\beta} \sim N_۶(۰, ۱۰^۸\boldsymbol{I}), \quad \sigma_\mu^۲ \sim IG(۰/۱, ۱۰),$$

برای پارامترها، استنباط بیزی براساس نمونهٔ تصادفی به حجم ۴۰۰۰ با تعداد ۲۰۰۰۰ تکرارها، مرحلهٔ داغیدن ۱۰۰۰۰ تکرار و تاخیر پنجم انجام گرفت. شکل ۳ پهنه‌بندی مقادیر کرانگین سرعت باد ایران را برای مدل فضایی نشان می‌دهد. این پهنه‌بندی براساس پیشگویی مقدار کرانگین سرعت باد در ۲۵۰۰ نقطه (شبکهٔ ۵۰ × ۵۰) با تکنیک‌های بخش ۴ به دست آمده است. واضح است که نواحی شمال غرب و غرب ایران دارای سرعت‌های باد بالایی هستند. جدول ۱ برآورد بیزی پارامترهای مدل و فاصله اطمینان ۹۵٪ بیزی را نشان می‌دهد. با توجه به این جدول توزیع GEV سرعت باد در ایران از دم متوسطی و تغییرپذیری نسبتاً بالایی برخوردار است.

۶ بحث و نتیجه‌گیری

مدل فضایی با رویکرد سلسله مراتبی برای مقادیر کرانگین معرفی گردید که شامل میدان تصادفی گاوسی، توزیع مقدار کرانگین تعمیم‌یافته و فرض استقلال شرطی

جدول ۱: برآورد بیزی پارامترهای مدل.

پارامتر	میانه توزیع پسین	فاصله اطمینان ۹۵٪ بیزی
ξ	-۰/۰۵	(-۰/۱۶و۰/۰۸)
σ	۱۴/۹۲	(۱۲/۹۶و۱۷)
σ_μ^2	۰/۷	(۰/۲۵و۳/۳)
ϕ	۲/۵۸	(۰/۴۱و۴/۶۵)

بود. استنباط بیزی مدل ارائه شده با تعیین توزیع‌های پیشین برای تمام پارامترها همراه بود که در آن عدم حتمیت در برآورد پارامترها لحاظ گرید. با این حال توزیع مقدار کرانگین تعمیم‌یافته توزیعی چوله و دم‌کلفت است در نتیجه میدان تصادفی با توزیع دم کلفت و چوله امکان مدل‌بندی بهتر مقادیر کرانگین را فراهم می‌کند. بنابراین استفاده از میدان تصادفی که میدان تصادفی گاوسی را شامل شده و انعطاف‌پذیری بیشتری را در تحلیل مقادیر کرانگین دارا باشد از جمله مسائل قابل بررسی دیگر خواهد بود.

مراجع

میرزامصطفی، ن.، خلیلی، د.، کمالی، غ.، هادربادی، غ.، دلالیان، م. و افضلی، ف.، (۱۳۸۲)، شبیه‌سازی سرعت باد و جهت باد به منظور پیش‌بینی فرسایش بادی در ایران، سومین کنفرانس منطقه‌ای تغییر اقلیم، دانشگاه اصفهان.

Banerjee, S., Carlin, B.P. and Gelfand, A.E., (2004), Hierarchical modeling and analysis for spatial data, Chapman & Hall/CRC Press.

Casson, E. and Coles, S., (1999), Spatial Regression Models for Sample Extremes, *Extremes*, **1**, 449-468.

Coles, S. G., (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, London.

Cooley, D., Nychka, D. and Naveau, P., (2007), Bayesian Spatial Modeling of Extreme Precipitation Return Levels, *Journal of the American Statistical Association*, **102**, 824-840.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A., (1998), Model-based Geostatistics, *Applied Statistics*, **47**, 299-350.

Payer, T. and Küchenhoff, H., (2004), Modelling Extreme Wind Speeds at a German Weather Station as Basic Input for a Subsequent Risk Analysis for High-speed Trains, *Journal of Wind Engineering and Industrial Aerodynamics*, **92**, 241-261.

Walshaw, D. and Anderson, C. W., (2000), A Model for Extreme Wind Gusts, *Applied statistics*, **49**, 499-508.

Yaglom, A.M., (1962), An introduction to the theory of stationary random functions, Dover Publications, New York.

دومین کارگاه آموزشی آمار فضایی و کاربردهای آن، ۱۰ -۱۱ خرداد ۱۳۹۱
مجموعه مقالات، ص ۱۶۱-۱۷۲

تحلیل داده‌های فضایی با نرم‌افزار SAS

الهام کیوان شکوه، یدالله واقعی
گروه آمار، دانشگاه بیرجند

چکیده : داده‌هایی که برحسب موقعیت (مکان) قرار گرفتن آنها در فضای مورد مطالعه، همبسته باشند و این همبستگی تابعی از فاصله موقعیت آن‌ها در فضای d بعدی $d \geq ۱$ باشد، داده‌های فضایی نامیده می‌شوند. برای تحلیلی ساختار وابستگی داده‌ها در فضای دو بعدی از نرم‌افزارهای $geoR$ و SAS استفاده می‌شود. ما در این مقاله، قابلیت‌های فرمان‌های $KRIGE۲D$، $VARIOGRAM$ و $SIM۲D$ را در نرم‌افزار SAS برای تحلیل داده‌های فضایی دو بعدی، به اختصار توضیح می‌دهیم و در آخر مقایسه نرم‌افزارهای SAS و $geoR$ در تحلیل داده‌های فضایی بیان می‌گردد.

واژه‌های کلیدی : د اده‌های فضایی، براوردگر تغییرنگار، کریگینگ، $geoR$ و SAS.

۱ مقدمه و تعاریف کلی

فرض کنید $\{t \in D \subset R^d\}$ موقعیت یک مشاهده در فضای اقلیدسی d بعدی و $Z(t)$ یک متغیر تصادفی در موقعیت t باشد. حال اگر t روی فضای اندیس D

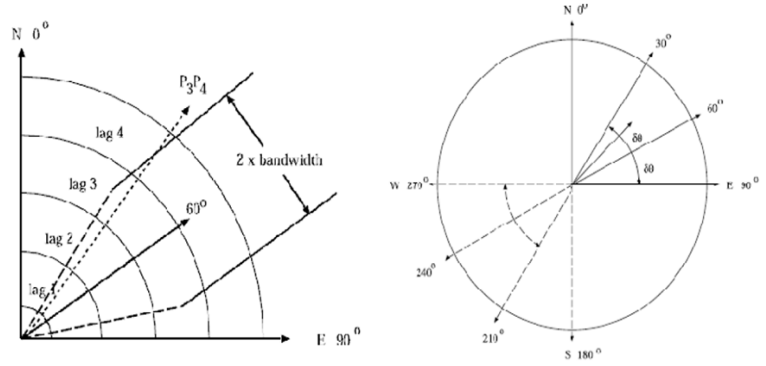
آدرس الکترونیک مسئول مقاله: الهام کیوان‌شکوه، e_keivanshekooh_stat@yahoo.com
کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲H۱۱

تغییر کند، مجموعه‌ای از متغیرهای تصادفی به صورت $\{Z(t); t \in D\}$ به وجود می‌آید که به آن میدان تصادفی می‌گویند. به مشاهدات یا داده‌های حاصل از یک میدان تصادفی داده فضایی گفته می‌شود. برای انجام تجزیه و تحلیل‌های متداول می‌بایست داده‌ها مانا(ایستا) باشند. مانایی بیانگر این نکته است که روابط توزیعی بین یک یا چند داده، با انتقال به اندازه یک بردار ثابت، تغییر نمی‌کند.

برای تحلیل ساختار و وابستگی داده‌ها از توابع تغییرنگار و هم تغییرنگار استفاده می‌شود. برای مدل‌سازی تابع تغییرنگار باید میدان تصادفی مانای ذاتی و برای مدل‌سازی تابع هم تغییرنگار باید مانای مرتبه دوم باشد. برآوردگر کلاسیک تغییرنگار به‌صورت میانگین مربع اختلاف همه زوج داده‌هایی که در بردار فاصله مشخص یا در یک h همسایگی از بردار فاصله h هستند، تعریف می‌شود.

$$۲\hat{\gamma}(h) = \frac{۱}{|N(h)|} \sum_{|N(h)|} (Z(t_i + h) - Z(t_i)) \qquad (۱)$$

چنانچه تغییرنگار به جهت بردار h بستگی نداشته باشد، میدان تصادفی همسانگرد و در غیر این صورت ناهمسانگرد گفته می‌شود. برای برآورد تغییرنگار در جهت خاص معمولاً از ناحیه تحمل زاویه‌ای (شکل ۱ سمت راست) و ناحیه تحمل زاویه‌ای که با قطاع محدود شده (شکل ۱سمت چپ) استفاده می‌شود.



شکل ۱: ناحیه تحمل زاویه‌ای و ناحیه تحمل زاویه و قطاع

مهمترین اهداف تحلیل داده‌های فضایی پیشگویی می‌باشد که روش‌های مختلفی دارد. یکی از این روش‌ها، کریگینگ می‌باشد. کریگینگ، معادل بهترین پیشگوی

نااریب خطی $(BLUP)$ می‌باشد. پیشگوی مکانی به صورت ترکیب خطی از متغیرهای فضایی $\hat{Z}(t_\circ) = \sum_{i=1}^{n} \lambda_i Z(t_i)$ است، که در آن ضرایب λ_i به‌گونه‌ای تعیین می‌شود که اولاً نااریب و ثانیاً در بین همه پیشگوهای خطی نااریب، دارای کمترین واریانس باشد. کریگینگ انواع مختلفی دارد که بسته به شکل و ماهیت توزیعی داده‌ها، یکی از روش‌ها مورد استفاده قرار می‌گیرد. برخی از انواع آن، کریگینگ بر اساس ساختار فضایی است. فرض می‌کنیم $Z(t) = \mu(t) + \sigma(t)$. در حالت کلی $\mu(t) = E(Z(t))$ می‌توان به t بستگی داشته و یا فارغ از t باشد. کریگیدن به دو نوع متداول کریگیدن عادی و کریگیدن عام دسته‌بندی می‌شود. در کریگیدن عادی فرض می‌شود $\mu(t)$ مقداری ثابت و فارغ از t است، اما در کریگیدن عام که حالت کلی‌تر کریگیدن عادی می‌باشد، میانگین $\mu(t)$ به t بستگی داشته و فرض می‌شود یک ترکیب خطی از توابع از قبل تعیین شده است. نوع دیگری از کریگیدن به نام کریگیدن ساده نیز وجود دارد که حالت خاص‌تر کریگیدن عادی است و در مواردی قابل استفاده است که میانگین $\mu(t)$ مقداری ثابت و معلوم باشد. در کارهای واقعی، میانگین نامعلوم است و کریگینگ ساده کاربردی ندارد.

۲ فرمان‌های نرم‌افزار SAS برای تحلیل داده‌های فضایی

در این بخش فرمان‌ها و گزینه‌های مهم سه فرمان $KRIGE2D$، $VARIOGRAM$ و $SIM2D$ معرفی و با ذکر مثال‌هایی، کاربردهای آن‌ها ذکر می‌گردد. با توجه به گستردگی امکانات فرمان‌های یادشده در این مقاله به بخشی از زیرفرمان‌ها اشاره شده است، به منظور مشاهده کل زیرفرمان‌ها می‌توان به $SAS/STAT$ ۹.۱ مراجعه نمود.

ابتدا لازم است نحوه ورود داده‌های فضایی در نرم‌افزار SAS بیان گردد. در این مقاله ما از داده‌های ضخامت ذغال‌سنگ (متغیر اصلی مورد بررسی) به همراه مشخصات آن به عنوان یک مثال کاربردی استفاده می‌کنیم. این داده‌ها ۷۵ اندازه‌گیری در نقاط مختلف می‌باشد که در سه ستون، به صورت زیر در نرم‌افزار SAS وارد می‌شوند. ستون اول و دوم مختصات مکانی و ستون سوم ضخامت ذغال‌سنگ می‌باشد (شکل ۲).

```
data thick;
   input east north thick @@;
   datalines;
    0.7  59.6  34.1   2.1  82.7  42.2   4.7  75.1  39.5
    4.8  52.8  34.3   5.9  67.1  37.0   6.0  35.7  35.9
    6.4  33.7  36.4   7.0  46.7  34.6   8.2  40.1  35.4
   13.3   0.6  44.7  13.3  68.2  37.8  13.4  31.3  37.8
   17.8   6.9  43.9  20.1  66.3  37.7  22.7  87.6  42.8
   23.0  93.9  43.6  24.3  73.0  39.3  24.8  15.1  42.3
   24.8  26.3  39.7  26.4  58.0  36.9  26.9  65.0  37.8
   27.7  83.3  41.8  27.9  90.8  43.3  29.1  47.9  36.7
   29.5  89.4  43.0  30.1   6.1  43.6  30.8  12.1  42.8
   32.7  40.2  37.5  34.8   8.1  43.3  35.3  32.0  38.8
```

شکل ۲: داده‌های ذغال سنگ

۱.۲ فرمان $VARIOGRAM$

برای محاسبه برآوردگر معمولی، نیرومند، نیم تغییر نگار و همچنین برآورد هم تغییرنگار و برآورد پارامترهای مدل تغییرنگار از این فرمان استفاده می‌گردد.
شکل عمومی فرمان $VARIOGRAM$ به صورت زیر است:

$PROC\ VARIOGRAM\ option;$
$COMPUTE\ computation-option;$
$CORDINATES\ coordinate-variables;$
$DIRECTION\ direction-list;$
$VAR\ analysis-variables-list;$

گزینه‌های زیر فرمان‌های $PROC\ VARIOGRAM$ عبارت‌اند از:
$DATA$ (مشخص‌کننده داده‌های ورودی)، $OUTVAR$ (چاپ شاخص‌های وابستگی فضایی)، $OUTDISTANCE$ (چاپ هیستوگرام اطلاعات داده‌های فضایی) و $OUTPAIR$ (چاپ اطلاعات زوج نقاط داده‌ها).
گزینه‌های زیر فرمان‌های $COORDINATE$ عبارت‌اند از:
$LAGDISTANCE$ (مشخص کننده فاصله لگ‌ها)، $NDIRECTIONS$ (مشخص کننده تعداد زاویه‌ها و جهت‌هایی که می‌خواهیم تغییرنگار را در آن برآورد کنیم)، $ANGLETOL$ (مشخص کننده زاویه تحمل)، $ROBUST$ (مشخص

کننده‌نیم تغییرنگار نیرومند) و $BANDWIDTH$ (مشخص کننده ناحیه تحمل زاویه و قطاع).

مثال ۱ : به منظور نشان دادن نحوه استفاده از زیرفرمان های یاد شده، در این مثال برآوردگرهای همسانگرد گشتاوری و نیرومند تغییرنگار داده های زغال سنگ را در ۷ لگ محاسبه می کنیم.

```
proc variogram data = thick outv = outv;
compute laged = ۷ maxlag = ۱۰ robust;
coordinates xc = east yc = north;
var thick;
run;
proc print data = outv label;
var lag count distance variog rvario;
run;
```

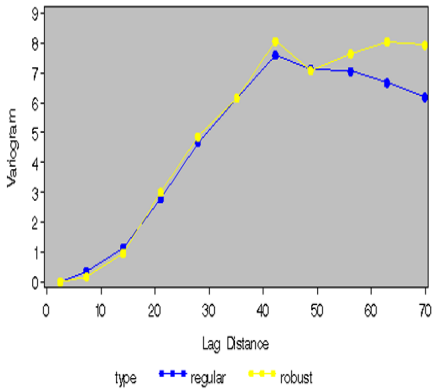
برنامه فوق خروجی ندارد و درصورتیکه فرمان ;proc porint را در انتهای آن اضافه کنیم، خروجی برنامه فوق به صورت زیر است.

خروجی این برنامه یک مجموعه داده است که شامل: نام متغیر($VARNAME$)، شماره لگ‌ها(LAG)، تعداد جفت نقاط ($COUNT$)، نقطه وسط لگ ($DISTANCE$)، میانگین فاصله زوج نقاطی که در لگ موردنظر قرار می‌گیرند ($AVERAGE$)، برآورد نیم تغییرنگار ($VARIOG$)، برآورد هم تغییرنگار ($COVAR$) و نیم تغییرنگار نیرومند ($RVARIO$).

برای رسم نمودار از برنامه زیر استفاده می‌کنیم.

```
data out۲ : set outv;
vari = variog; type =' regular'; output;
vari = variog; type =' robust'; output;
run;
proc gplot data = outv۲;
plot vari*distrance = type/frame
cframe = ligr vaxis = axis۲ haxis = axis۱;
run;
```

شکل زیر تفاوت برآوردگرهای نیرومند و گشتاوری نیم تغییرنگار را برای داده‌های ذغال‌سنگ نشان می‌دهد.



شکل ۳: برآوردگرهای نیرومند و گشتاوری نیم تغییرنگار را برای داده‌های ذغال‌سنگ

۲.۲ فرمان $krige2d$

پیش‌بینی فضایی هر روش، پیش‌بینی است که دارای وابستگی فضایی باشد. روش ساده و عمومی پیش‌بینی فضایی، کریگینگ معمولی است. برای کریگینگ معمولی و بلوکی در فضای دو بعدی در مدل‌های همسانگرد، ناهمسانگرد و نیم‌تغییرنگار از این فرمان استفاده می‌شود. با استفاده از فرمان $KRIGE2D$ می‌توان از چهار مدل گوسی، نمایی، کروی و قدرت و همچنین اثر قطعه برای برآورد کریگینگ استفاده نمود.

شکل عمومی و کلی رویه $krige2d$ به صورت زیر است:

$$
\begin{aligned}
&PROC\ KRIGE2D\ options;\\
&COORDINATES \mid COORD\ coordinate-variables;\\
&GRID\ grid-options;\\
&PREDICT \mid PRED \mid Ppredict-options;\\
&MODEL\ model-options;
\end{aligned}
$$

گزینه‌های زیرفرمان $MODEL$ عبارتند از: $NUGGET$ (مشخص کننده دامنه)، $SCALE$ (مشخص کننده مقیاس)، $ANGLE$ (مشخص کننده زاویه) و $RATIO$(مشخص کننده نسبت شعاع کوچک به شعاع بزرگ در مدل ناهمسانگرد).

مثال ۲ : می‌خواهیم با یک مدل تغییرنگار کروی ($form = s$) که اثر قطعه‌ای آن ۵، آستانه ۲۰، دامنه ۸ و درجهت ۳۵درجه آزیموت دارای بیشترین دامنه است. میزان ذغال سنگ را در یک شبکه منظم پیشگویی کنیم.

```
proc krige2d data = thick outest = est1;
pred var = thick;
model scale = 20 range = 8 nugget = 5 form = s
angle = 35 ratio = 0.7;
coord xc = east yc = north;
grid x = 0 to 20 by y = 0 to 20 by 5;
run;
```

برنامه فوق خروجی ندارد و در صورتی که فرمان $proc\ print;$ را درانتهای آن اضافه کنیم خروجی برنامه در شکل ۴ داده شده است. در جدول ۴ gxc و ygc

Obs	LABEL	VARNAME	GXC	GYC	NPOINTS	ESTIMATE	STDERR
1	Pred1.Model1	thick	15	0	20	44.8085	0.08502
2	Pred1.Model1	thick	15	5	20	44.1551	0.03984
3	Pred1.Model1	thick	15	10	20	43.2197	0.08886
4	Pred1.Model1	thick	15	15	20	42.0987	0.09011
5	Pred1.Model1	thick	15	20	21	40.8532	0.06837
6	Pred1.Model1	thick	15	25	23	39.5745	0.03632
7	Pred1.Model1	thick	15	30	24	38.2638	0.00384
8	Pred1.Model1	thick	15	35	27	37.0465	0.03599

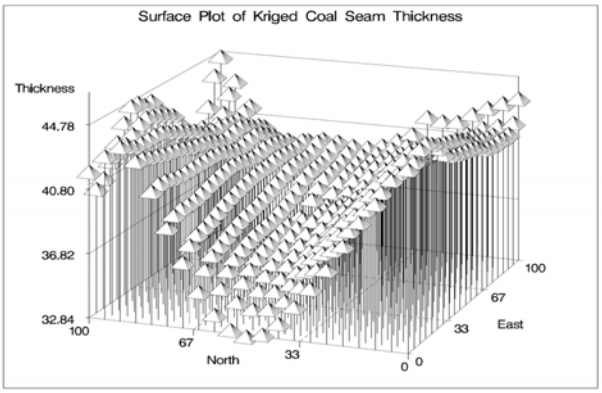
شکل ۴: خروجی برنامه در مثال ۲

مشخصات شبکه و $npoint$ تعدادنقاط درشبکه،$estimate$ برآورد مقدار زغال سنگ در هر شبکه به روش کریگینگ و $stderr$ انحراف معیار خطاست. برای رسم نمودار

سه بعدی از برنامه زیراستفاده می کنیم.

```
proc g3d data = est;
title 'Surface Plot of Kriged Coal Seam Thickness';
scatter gyc * gxc = estimate/grid;
lable gyc =' North'
gxc =' East'
estimate =' Thickess';
run;
```

نمودار سه بعدی مقادیر برآورد شده در زیر آمده است:



شکل ۵: نمودار سه بعدی مقادیر برآورد شده

۳.۲ فرمان $SIM2d$

برای شبیه‌سازی میدان تصادفی گوسی با میانگین و تابع کوواریانس مشخص، از فرمان $SIM2D$ استفاده می‌شود. در این روش ماتریس واریانس-کوواریانس میدان تصادفی $(\Sigma_{n\times n})$ با استفاده ار تابع کوواریانس با پارامترهای معلوم محاسبه و سپس به روش تجزیه چولسکی به صورت $\Sigma = LL'$ تجزیه می‌شود(که L یک ماتریس پایین مثلثی است)، در نهایت از توزیع نرمال n متغیره با بردار میانگین μ و ماتریس کوواریانس $\Sigma = LL'$ یک مشاهده برداری به طول L (برابر تعداد کل داده‌های فضایی که می‌خواهیم شبیه‌سازی کنیم) شبیه‌سازی می‌شود. در این فرمان می‌توان از چهار مدل گوسی، نمایی، کروی و همچنین اثر قطعه برای شبیه‌سازی استفاده نمود.

شکل عمومی فرمان $SIM۲DPROC$ به صورت زیر است:

$PROC\ SIM۲D\ options;$
$COORDINATES\ coordinate-variables;$
$GRID\ grid-options;$
$SIMULATE\ simulate-options;$
$MEAN\ mean-options;$

برخی از گزینه‌های زیرفرمان $SIMULATE$ عبارتند از: $NUMREAL$ (مشخص‌کننده تعداد شبیه‌سازی)، $SEED$ (مشخص‌کننده مقدار آغازین در شبیه‌سازی). سایر گزینه‌ها همانند شناسه‌های زیرفرمان $MODEL$ در فرمان $KRIGE۲D$ است.

مثال ۳ : فرض کنید در چهار نقطه (۱ ,۰)، (۰ ,۴)، (۱ ,۳) و (۳ ,۴) می‌خواهیم از یک مجموعه داده فضایی با میانگین ثابت ۴۰/۱۴ و تابع کوواریانس گوسی با آستانه ۲۰۰، دامنه ۳۰ و مقیاس ۷/۵ شبیه سازی انجام دهیم.

$data\ grid;$
$input\ xc\ yc;$
۱۰
۴۰
۱۳
۴۳
$run;$
$proc\ sim۲d\ outsim = sim۱;$
$simulate\ numeral = ۱\ seed = ۲۰۰$
$scale = ۷.۵\ range = ۳۰.۰\ form = guss;$
$mean\ ۴۰.۱۴;$
$grid\ gdata = grid\ xc = xc\ yc = yc; run; proc\ print; run;$

خروجی برنامه فوق در شکل ۶ داده شده است. در جدول ۶، $ITER$ شمارنده شبیه‌سازی، GXC و GYC مختصات طول و عرض نقاط مطابق داده‌های ورودی و $SVALUE$ مقدار شبیه‌سازی شده متغیر داده شده است.

Obs	LABEL	_ITER_	GXC	GYC	SVALUE
1	Sim1	1	1	0	40.4736
2	Sim1	1	4	0	40.5453
3	Sim1	1	1	3	40.8280
4	Sim1	1	4	3	40.8663

شکل ۶: خروجی برنامه مثال ۳

۳ مقایسه نرم‌افزارهای SAS و $geoR$ در تحلیل داده‌های فضایی

در این قسمت ما نرم‌افزارهای SAS و $geoR$ را در سه بخش برآورد تغییرنگار، پیشگویی و شبیه‌سازی داده‌ها مورد بحث و بررسی قرار می‌دهیم. نرم‌افزار $geoR$ که به عنوان یک کتابخانه از توابع (فرمانها) روی R نصب می‌شود، اختصاصاً برای تجزیه و تحلیل داده‌های داده‌های فضایی دو بعدی ساخته شده است. (ریبه ایرو و دیکل، ۲۰۱۱) دامنه مدل‌هایی که با $geoR$ می‌توان به نیم تغییرنگار برازش داد بیشتر از SAS است. با فرمان $variofit$ یا $likfit$ می‌توان مدل‌های کوشی، کوشی تعمیم‌یافته، مدور، مکعبی، گوسی، نمایی، مترن، کروی، توانی، نایتینک و موجی را برازش داده و پیشگویی را انجام داد.

نرم‌افزار SAS اختصاصاً فرمانی برای برازش مدل به نیم تغییرنگار یا هم تغییرنگار داده‌ها را ندارد و اگر کاربر بخواهد مدل برازش دهد، باید از فرمان‌های عمومی مربوط به برازش مدل‌های رگرسیونی (مانند $NLIN\ PROC$ و $NLP\ PROC$) استفاده کند. در حالیکه با فرمانهای $variofit$ و $likfit$ در نرم‌افزار $geoR$ می‌توان با روش‌های کمترین توان‌های دوم‌های دوم معمولی و موزن (OLSوWLS) و حداکثر درستنمایی (تحت توزیع نرمال چند متغیره) مدل‌های بسیار زیاد و متنوعی به نیم تغییرنگار برآورد شده بوسیله فرمان‌های $variog$ یا $variog$۴ برازش داد.

فرمان $Krige2D$ نرم‌افزار SAS پیشگویی را با روش‌های کریگینک معمولی و بلوکی انجام می‌دهد در حالیکه فرمان $Krige.conr$ در نرم‌افزار $geoR$ قابلیت انجام چهار نوع کریگینگ از جمله کرینگینگ عادی و عام را دارد و علاوه بر آن با فرمان $Krige.bayes$ می‌توان پیشگویی کریگینگ را با روش‌های بیزی نیز انجام داد. نرم‌افزار $geoR$ همانند نرم‌افزار SAS در بخش شبیه‌سازی و پیشگویی امکان تعریف مدل ناهمسانگرد بیضوی (نوع خاصی از ناهمسانگردی هندسی) را داد. یکی از تفاوت‌های نرم‌افزار SAS و $geoR$ این است که در SAS داده‌ها بدون روند در نظر گرفته می‌شود، اگر داده ها روندی داشته باشند، با فرمان $variog$ در $geoR$ و همینطور در فرمان $Krige.com$ برای پیشگویی می‌توان روند را معرفی و در محاسبات منظور نمود.

با فرمان $grf(fieldrandomgoussian)$ در نرم افراز $geoR$ به راحتی می‌توان یک مجموعه‌داده فضایی به به حجم n از یک میدان تصادفی گوسی با مدل نیم تغییرنگار و پارامترهای مشخص (در موقعیتهای معین) شبیه‌سازی نمود. فرمان $image.grf$ می‌تواند یک نمودار تصویری از داده‌های فضایی شبیه‌سازی شده را نشان دهد. بدلیل تنوع تنوع مدل‌ها در نرم‌افزار $geoR$، قابلیت فرمان grf بیشتراز $SIM2D\ PROC$ می‌باشد. از جمله این قابلیت، انجام اعتبارسنجی متقابل-$(cross-validation)$ به منظور ارزیابی دقت پیشگویی و بررسی دقت برازش مدل‌ها با فرمان $xvalid$ می‌باشد.

مراجع

اسماعیلیان، م. (۱۳۸۹)، راهنما ۹.۱ SAS مقدماتی و پیشرفته، انتشارات مؤسسه فرهنگی هنری دیباگران تهران.

حسنی‌پاک، ع، ا. (۱۳۷۷)، زمین آمار (ژئواستاتیستیک)، انتشارات دانشگاه تهران.

رجبیون، ا. (۱۳۹۰)، مدل‌سازی تغییرنگار داده‌های فضایی-زمانی با تابع کوواریانس تفکیک-ناپذیر، پایان نامه کارشناسی ارشد، دانشگاه بیرجند.

مالکی، م. (۱۳۸۸)، رگرسیون غیرخطی و کاربردهای آن در داده‌های فضایی، پایان نامه کارشناسی ارشد، دانشگاه بیرجند.

واقعی، ی. (۱۳۸۱)، تجزیه و تحلیل داده‌های فضایی نامانا و کاربرد آن در همه‌گیری شناسی جغرافیایی بیماریها، رساله دکتر، دانشگاه تربیت مدرس.

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

Ribeiro, P.J. and Diggle, P.G (2012), *geoR: Analysis of geostatistical data version 1.7-2.* http://cran.uma.ac.ir

مجموعه مقالات، ص ۱۸۷-۲۰۲

کاربرد آمار فضایی در هواشناسی

مهدی نقی‌خانی[۱]، مریم زنگنه[۲]

۱. گروه پردازش داده‌ها و اطلاع رسانی، پژوهشکده آمار
۲. مرکز آمار ایران

چکیده : استفاده از اطلاعات و برنامه ریزی در خصوص بهره‌وری از منابع و توسعه پایدار آنها، مستلزم شناخت عوامل حاکم بر چگونگی وضعیت منابع و استعدادهای بالقوه و بالفعل می باشد. پیشرفتهای روز افزون در زمینه فناوری اطلاعات و ارتباطات و جمع آوری داده های آماری و تولید نقشه، سبب گشته که مدیران و برنامه ریزان با حجم وسیعی از اطلاعات مواجه گردند. از آنجایی که ارزش اطلاعات وابسته به زمان ، مکان ، صحت و دقت آن می باشد لذا تشکیل بانکهای اطلاعاتی مورد نیاز و پردازش آنها جهت استخراج اطلاعات مفید کاملا ضروری است. یکی ازاین مسائل میزان بارش است. آگاهی از میزان بارش در کل کشور و یا منطقه‌ای خاص همواره از مسائل حیاتی، پراهمیت و راهبردی بشر است که نقش مهمی در تصمیم گیریهای کوتاه مدت و بلند مدت دارد. در این میان شبکه های سنجش میزان پراکندگی بارش در

[۱] آدرس الکترونیک مسول مقاله: مهدی نقی خانی، naghikhani.m@srtc.ac.ir

نقاط مختلف به منظور پاسخگویی به این نیاز تاسیس شده اند و لازم است با استفاده از این بانکداده ها، اطلاعاتی برای کل سطح استان، منطقه و یا کشور برای برنامه ریزان کشوری و استانی جهت تصمیم گیری در امور اقتصادی، منابع آبی، کشاورزی، تغییرات اقلیمی و... موجود باشد. جهت تعیین مدل بارش و پیش گویی آن مدلهای متفاوتی وجود دارد اما با توجه به اینکه بین دادههای بارش همبستگی مکانی است و به عبارت دیگر داده های که برای یك پارامتر در مناطق مختلف جمع آوری شده اند به همدیگر وابستگی مکانی دارند، تکنیک آمار مکانی یا فضایی یک ابزار قوی آماری در جهت حل مساله مناسب می باشد. در این مقاله با استفاده از اطلاعات موقعیت جغرافیایی (ارتفاع ، طول و عرض جغرافیایی) مربوط به ۲۴۰ ایستگاه سینوپتیك کشور استفاده و نرم افزار **GS$^+$9** ، MINITAB14 و SAS 9.1 مدل میزان بارش کل کشور پیش بینی، منحنی تراز و رویه میزان بارش (و انحراف معیار بارش) بدست می آید و در نهایت بعد از صحه گذاری مدل، میزان بارش در چند نقطه پیش گویی می‌شود.

کلمات کلیدی: کریگیدن (عام و معمولی)، هم کریگیدن ، آمار مکانی (فضایی)، پیش گویی مدل بارش

۱ -مقدمه:

تغییرات اقلیمی در طول چند سال اخیر به عنوان یکی از مهمترین موضوعات زیست محیطی در محافل مختلف مطرح گردیده است. مقولاتی از قبیل آلودگی آب، هوا، بارش و ... اثرات متفاوتی در زندگی بشر روی کره زمین از جمله اسکان، تولیدات کشاورزی و استفاده از انرژی داشته است. با توجه به گسترش علوم این امکان ایجادشده است تا با استفاده از روش‌های مختلف رفتار پدیده‌های مختلف گیتی مدل‌سازی گردد. یکی از این موارد پیش گویی مکانی میزان بارش در مناطقی است که امکان احداث ایستگاه هواشناسی (به دلیل هزینه بالا احداث و نگه داری) نمی‌باشد.لذا با توجه به اینکه اطلاعات مربوط به میزان بارش دارای همبستگی مکانی

می‌باشد می‌توان با استفاده از روش‌های آمار فضایی مدل بارش را بدست آورد و سپس با استفاده از منحنی تراز و رویه میزان پیش گویی‌های مورد نیاز را انجام داد. مجموعه داده‌های مورد بررسی در این مقاله شامل اطلاعات طول، عرض، ارتفاع و میزان بارش مربوط به ایستگاههای۲۴۰ سینوپتیک می‌باشد. هر ایستگاه سینوپتیك، یك مکان مشاهده است که موقعیت آن توسط طول، عرض جغرافیایی مشخص می‌شود. عموماً تحلیل‌های مکانی یك مجموعه داده فضایی یا مکانی شامل سه قسمت: تحلیل توصیفی داده مکانی، مدل‌سازی مکانی (تحلیل تغییرنگار و تعیین مدل تغییرنگار)، پیش‌بینی مکانی متغیر و انحراف معیار با استفاده از تکنیك‌ها و نمودارهای رویه و هم‌تراز می‌باشد. در این تحلیل از نرم‌افزارهای**GS$^+$9** ۱(نرم افزار اصلی) و Minitab 14 و SAS 9.1 (نرم افزارهای کمکی) استفاده شده است. در این زمینه مقالات متعددی وجود دارد ادب و همکارانش (۱۳۸۶) به ارزیابی روش‌های کریگیدن و رگرسیون خطی بر پایه DEM در تهیه نقشه همبارش سالانه خراسان رضوی پرداخته است. یعقوبیان(۱۳۸۷) بارندگی در استان همدان را با روش‌های کریگیدن یش بینی کرده است. اطمینان(۱۳۸۵) و شفعیعی (۱۳۸۷) با استفاده از روش کریگیدن(آمار فضایی و پیشگویی فضایی- مکانی) سطح آب حوزه آبریز دشت بیرجند پیش گویی را کرده است.

فاز اول: بررسی توصیفی داده های مکانی

تحلیل اولیه داده‌های مکانی شامل دو قسمت می‌باشد:تحلیل موقعیت مکان‌های اندازه‌گیری[۲] ۲- تحلیل توصیفی متغیر مورد بررسی (باران). در قسمت اول(تحلیل توصیفی) با استفاده از نمودارهای چندك۳ یا نقشه موقعیت٤ چگونگی توزیع و پراکندگی مقادیر نمونه در حوزه مکانی نشان داده می‌شود. هدف از این نمودارها بررسی

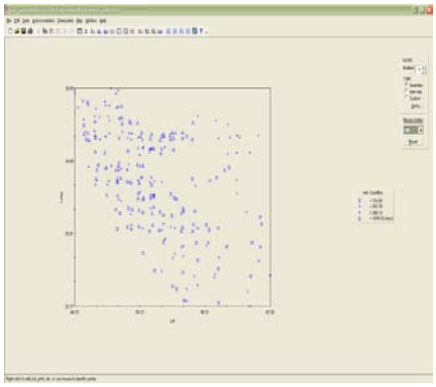
۱. GeoStatistics for the Environmental Sciences (gamma design software)

۲. Measurement locations

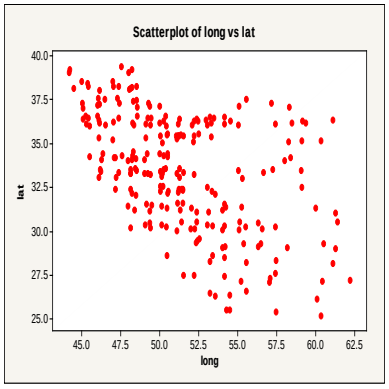
3. Quantile Plots

4. Coordinate Map

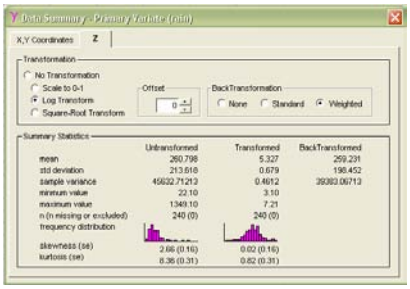
همگنی پراکندگی موقعیت‌های نمونه برداری می‌باشد(عدم همگنی منجر به ایجاد خطا در براورد مدل می گردد). مطابق با شکل۱ و ۲ با کمی دقت متوجه می‌شویم که در ناحیه مرکزی به علت کویری بودن پراکندگی مناسبی از ایستگاه سنجش سینوپتیك ندارد. البته میزان پراکندگی در حاشیه شمالی، شمال غرب، غرب کشور و استان‌های مرکزی کشور مناسب است. قسمت دوم تحلیل توصیفی، تحلیل توصیفی مشاهدات مکانی متغیر اصلی می‌باشد. در مدل تصادفی مکانی معمولی فرض بر این است که مشاهدات از یك توزیع نرمال استخراج شده‌اند و ساختار همبستگی آنها به موقعیت مکان‌های مشاهده بستگی دارد. جهت بررسی موضوع از آزمون نرمالیتی کلموگروف – اسمیرنوف و نمودار احتمال استفاده می‌شود که با توجه با مقدار$p\text{-value} < 0.005$ داده‌های میزان بارش دارای توزیع نرمال نمی‌باشد. لذا با استفاده از تبدیل لگاریتمی داده‌ها دارای توزیع نرمال می‌گردد (شکل ۳).



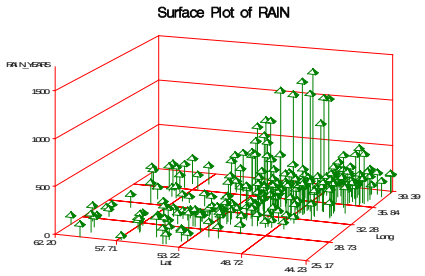
شکل ۱- نمودارهای چندك



شکل ۲- نمودار رویه پاسخ



شکل ۳ – آمار توصیفی



شکل ٤ موقعیت پراکندگی ایستگاههای سینوپتیک

سومین نموداری که در این قسمت استفاده می‌شود نمودار رویه پاسخ است که جهت تشخیص وجود روند و نوع مشخصه عمومی مکانی استفاده می‌شود. با توجه شکل ۴روند رویه‌ای خاصی در مشاهدات مشاهده نمی‌شود. عموما مطابق با این نمودارد دو راهبرد جهت حل مساله ایجاد می‌گردد :۱-عدم وجود روند رویه ای: در این حالت از خود مشاهدات جهت تجزیه و تحلیل استفاده می‌گردد. ۲-وجود روند رویه‌ای خاص: در صورت مشاهده روند رویه‌ای خاصی حتما می‌بایست رویه پاسخ را برازش داد و سپس تحلیل مکانی بروی باقی‌مانده‌ها انجام شود.ویژگی دومی که از این نمودار معین می‌گردد مشخصه عمومی مکانی است که عموما ۳ مشخصه عمومی اغلب در مشاهدات مکانی رخ می‌دهد.

۱ -تغییر کم ، تغییر پذیری بزرگ – مقیاس:

(slowly varying, large-scale variations in the measured values)

۲ -نامنظم – تغییرپذیر مقیاس – کوچك:

(irregular, small-scale variations)

۳ -همسانی اندازه‌گیری‌ها در موقعیت‌های نزدیك به هم دیگر

(similarity of measurements at locations close together)

که در حالت ۱ و ۳ معمولا از روش‌های هموارسازی استفاده می‌شود و در حالت دوم از روش‌های پیش بینی نقشه‌ها استفاده می‌شود. که طبق این نمودار مشاهدات از نوع نامنظم – تغییرپذیر مقیاس – کوچك می‌باشد. اما نکته مهم عدم وجود روند رویه خاص در موقعیت‌های مشاهده است. البته در بعضی نقاط روندی مشاهد می‌گردد که به دلیل وجود نقاط دور افتاده می‌باشد که در قسمت بعد به آن اشاره می‌گردد.

فاز دوم: مدلسازی مکانی

گام بعدی شناخت عوامل بروز تغییرات در متغیر وابسته و میزان اثر هر یك از این عوامل و تشخیص مدل تغییرنگار تجربی است. در مدل مکانی، تغییرات را می‌توان بطورکلی به دو عامل «ساختار میانگین» و «ساختار وابستگی مکانی» نسبت داد، که تغییرات ناشی از اثر متغیرهای توضیحی در «ساختار میانگین» و تغییراتی که صرفا ناشی از وابستگی مکانی هستند، در «ساختار وابستگی مکانی» مدل‌بندی می‌شوند. به عبارت دیگر مدل مکانی تغییرات داده‌ها را به دو صورت زیر تجزیه می‌کند:

تغییرات ناشی از وابستگی مکانی + تغییرات ناشی از ساختار میانگین = تغییرات داده‌ها

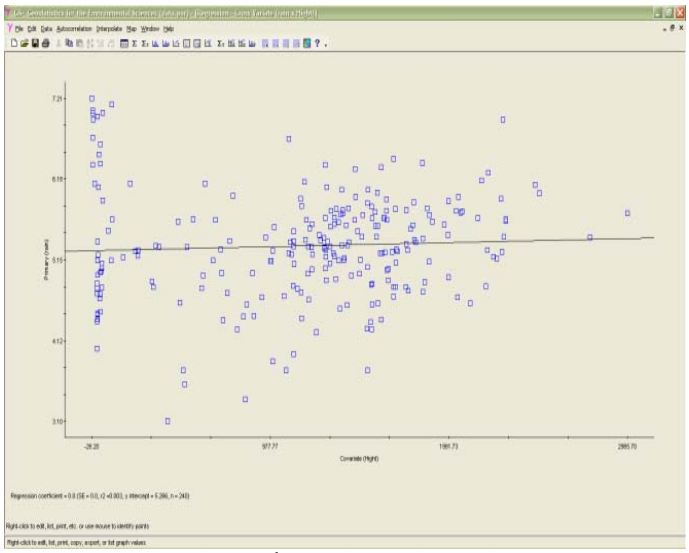
بروز تغییرات کوچک مقیاس ممکن است ناشی از عواملی مثل خطای اندازه گیری یا تغییر پذیری در درون مکان مشاهده باشد. می‌توان تغییرات کوچک مقیاس را به صورت جمله خطا در مدل در نظر گرفت. بروز تغییرات بزرگ مقیاس ممکن است ناشی از تغییر پذیری بین مکان‌های مشاهده باشد. برای داده‌های مکانی ، تغییر پذیری بین مکان‌های مشاهده به صورت تابعی از فاصله بین مکان‌های مشاهده مدل بندی می‌شود چرا که وقتی مکان‌های مشاهده به هم نزدیک هستند مقادیر مشاهده شده در آن مکان‌ها به هم شبیه تر هستند ، این شباهت را وابستگی مکانی می‌نامند.

تعریف مدل خطی باخطای وابسته مکانی

داده‌های مکان $Z(s_1),...,Z(s_n)$ که در مکان های $\{s_1,...,s_n\}$ مشاهده شده اند، به صورت زیر مدل‌بندی می‌شوند:

$$Z(s) = \sum_{l=1}^{q} \beta_l x_l(s) + \delta(s) \qquad s \in D \subset R^d$$

که $\{x_l(.), l = 1,...,q\}$ گردایه‌ای است از q متغیر توضیحی غیر تصادفی که ممکن است به موقعیت مکانی بستگی داشته باشند و $\delta(.)$ فرآیند خطا با میانگین صفر و واریانس محدود است و امکان دارد وابسته مکانی باشد. با توجه به شکل ۲ روند رویه‌ای خاصی دیده نمی‌شود(بجز نقاط دور افتاده). لذا تغییرات داده‌ها ناشی از تغییرات ناشی از وابستگی مکانی است و تغییرات ناشی از ساختار میانگین ندارد. (اما جهت بررسی دقیق‌تر متغیر ارتفاع به عنوان متغیر توضیحی مورد بررسی قرار گرفته می‌شود . اما مطابق با شکل ۵ رابطه رگرسیونی معنی داری وجود ندارد و میانگین ثابت می‌باشد.)



شکل ۵- نمودار رابطه رگرسیونی

پیش گویی مکانی

برای پیش گویی روش‌های گوناگونی ارائه شده است که در مقاله به روش کریگینگ اشاره می شود. کریگینگ که از نام دکتر د. جی. کریگینگ گرفته شده، شیوه‌ای است که

براساس آن می‌توان مقدار متغیر در یک مکان جدید را با استفاده از مقدار متغیر در دیگر مکان‌های مشاهده پیش گویی کرد. این روش بهترین پیش گوی خطی نااریب را که به آن پیش گوی بهینه نیز می‌گویند، به دست می‌دهد. پیش گویی که از روش کریگینگ به دست می آید ، خطی ، نااریب و دارای کوچکترین واریانس در بین تمام پیش گوهای نااریب خطی است. که به دو صورت کریگینگ معمولی و کریگینگ عام می‌باشد.

- کریگینگ معمولی:

$Z(s) = \mu + \delta(s), \quad s \in D, \mu \in R^d$, μ ثابت ونامعلوم

$$p(\Sigma; s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i), \quad \sum_{i=1}^{n} \lambda_i = 1$$

که $p(Z; s_0)$ پیش گوی متغیر Z در مکان s_0 است.

شرط برابری مجموع ضرائب پیش گوی خطی با ۱ ، نااریب بودن پیش گو را تضمین می کند.

- کریگینگ عام

در مدل کریگینگ معمولی $\mu \in R^d$ ثابتی نامعلوم فرض می شد. کریگینگ عام، تعمیم یافته کریگینگ معمولی است. به این ترتیب که $E[Z(s)] = \mu(s)$ دیگر یک مقدار ثابت نیست بلکه ترکیب خطی نامعلومی از توابع معلوم $\{x_1(s),...,x_q(s)\}$ است. (با وجودی که $x_j(s)$ها به صورت تابعی از مکان مشاهده نوشته شده اند، مقدار آنها ممکن است برابر مقداری عددی مثل ۱ یا برابر مقدار یک متغیر توضیحی مربوط به داده مشاهده شده درمکان s باشد.)

کریگینگ عام تحت دو فرض زیر پیش گویی می‌کند.

$$Z(s) = \sum_{l=1}^{q} \beta_l x_l(s) + \delta(s) \qquad s \in D$$

که $\beta = (\beta_1,...,\beta_q)' \in R^q$ برداری نامعلوم از پارامترها و $\delta(.)$ یک فرآیند تصادفی ذاتاً ایستا با میانگین صفر است.

$$p(\sum; s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i), \quad \lambda' X = x'$$

شرط $\lambda' X = x'$ شرط لازم و کافی برای ناااریب بودن پیش گو است.

مدل‌سازی «وابستگی مکانی»

برای تشریح ارتباط مکانی بین مقادیر یک متغیر مکانی در مکان‌های مختلف از تغییرنگار استفاده می‌شود. طبیعی‌ترین راه برای مقایسه دو مقدار $Z(s)$ و $Z(s+h)$ در دو مکان یکی به مختصات s و دیگری $s+h$ (که به فاصله h از نقطه s قرار دارد)، بررسی اختلاف این دو مقدار است. تابع «عدم تشابه» یا «تغییرنگار» دو نقطه به فاصله h به صورت زیر تعریف می شود:

$$2\gamma(h) = E(Z(s) - Z(s+h))^2$$

در حالت کلی، کمیت $2\gamma(.)$ که تابعی از تفاوت($s_i - s_j$) است، «تغییرنگار» نامیده می‌شود:

$$\text{var}(Z(s_i) - Z(s_j)) = 2\gamma(s_i - s_j) \qquad \forall s_i, s_j \in D$$

برآوردگر کلاسیک «نیم تغییرنگار» را به صورت زیر تعریف کرد:

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(s_i) - Z(s_j))^2$$

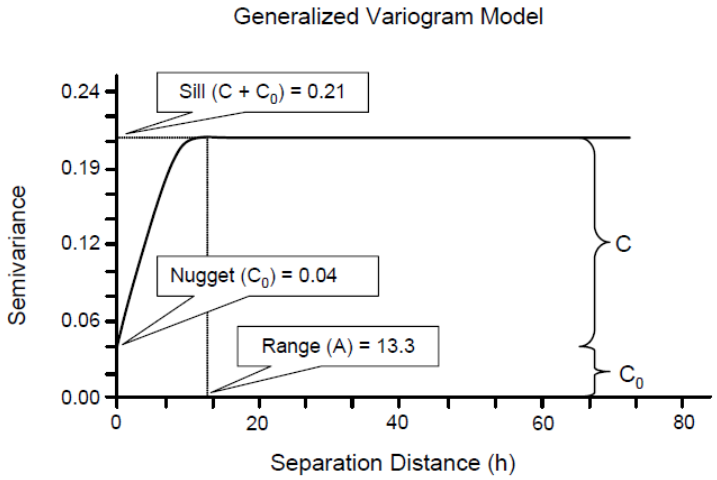
که جمع روی $N(h) \equiv \{(i,j) : s_i - s_j = h\}$ انجام می شود و $|N(h)|$ تعداد عناصر متمایز $N(h)$ است. این برآوردگر ناااریب است اما نسبت به مشاهدات غیرمعمول استوار نیست. عموماً پنج مدل نیم تغییرنگار کروی[1]، نمایی[2] ، گوسی[1] ،

---

[1]. Spherical

[2]. Exponential

توانی[۲] و خطی[۳] وجود دارد(جدول ۱) که با استفاده از روش های حداقل مربعات، درستنمایی ماکزیمم و روش‌های استوار مدل‌های نیم تغییرنگار تجربی برازش داده می‌شود. نمودار تغییر نگار تجربی یك نمودار Semi variance در مقابل Separation Distance می‌باشد که بوسیله ۳ پارامتر واریانس نقطه‌ای[٤]، حد آستانه[٥] و دامنه تعیین می‌گردد. (شکل۶)



شکل ٦ – نمودار تغییر نگار

---

[1]. Gaussian

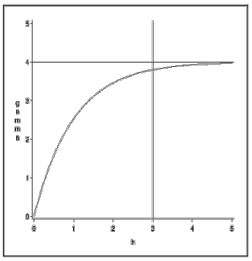
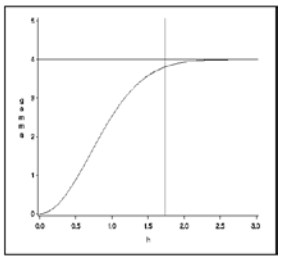
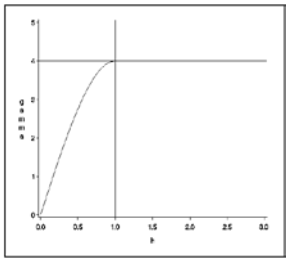
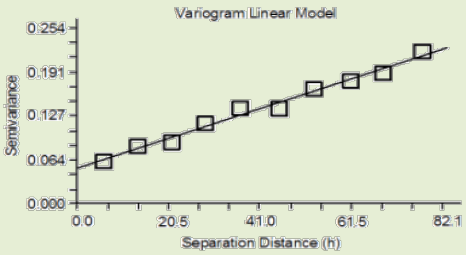
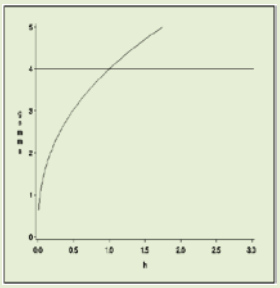
[2] .Power

[3]. Linear

[4]. Nugget Variance

[5] . Sill

جدول ۱ انواع مدل‌های تغییرنگار

مدل نمایی	مدل گوسی	مدل کروی
$\gamma(h)=\left\{c_0+c\left[1-\exp\left(\frac{h}{a_o}\right)\right]\right\}$	$\gamma(h)=\left\{c_0+c\left[1-\exp\left(\frac{h^2}{a_o^2}\right)\right]\right\}$	$\gamma(h)=\begin{cases} c+c_0\left[\frac{3}{2}\frac{h}{a_0}-\frac{1}{2}\left(\frac{h}{a_0}\right)^3\right] & h\le a_0 \\ c+c_0 & h>a_0 \end{cases}$
		
مدل خطی		**مدل توانی**
$\gamma(h)=c_o+\left[h\left(\frac{c}{a_o}\right)\right]$		$\gamma(h)=c_o h^{a_o}$
		

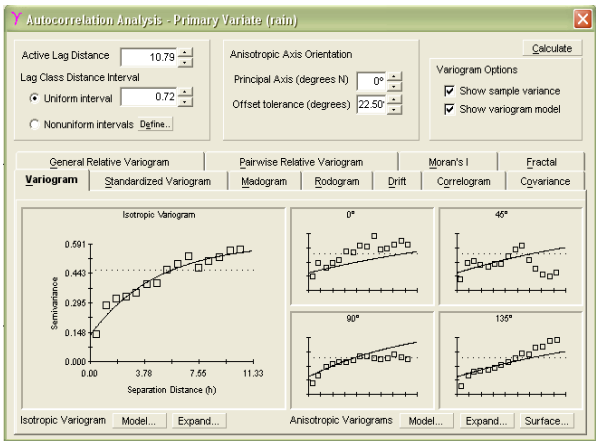
برای برازش مدل مناسب به تغییرنگار تجربی، می‌بایست پارامترهای دامنه، اثر نقطه وآستانه مدل به گونه‌ای انتخاب می‌شوند که تفاوت تغییرنگار تجربی و تغییرنگار برازش داده شده، می‌نیمم شود. مطابق شکل‌های ۷ – ۸ نتایج جدول ۲ استخراج میگردد.

جدول۲- جدول مقایسه انواع مدل

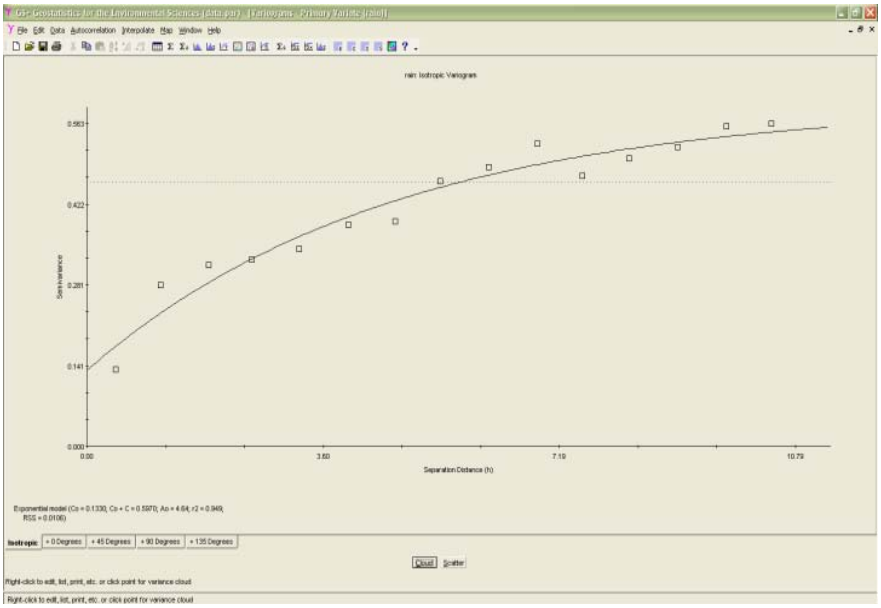
مدل	Nugget Effect(c_o)	Sill($c+c_o$)	Rang	RSS	R^2	$\frac{c}{c+c_o}$
کروی	0.176	0.541	9.700	0.0133	0.935	0.675
نمایی	0.133	0.597	13.920	0.0106	0.949	0.770
گوسی	0.219	0.537	7.9155	0.0181	0.912	0.592

که در نهایت مدل نهایی نیم تغییر نگار با توجه به سه معیار R^2 ، RSS و $\frac{c}{c+c_o}$ بصورت زیر انتخاب می‌شود:

$$\gamma(\mathrm{h}) = 0.597[1-\exp(\mathrm{h}/(13.92)]$$



شکل ۷- نمودار تغییر نگار تجربی



شکل ۸- نمودار تغییر نگار تجربی

فاز سوم: پیش مکانی

بعد از استخراج بهترین مدل نیم تغییرنگار، پیش‌گویی میزان بارندگی برروی نقاطی که میزان وجود ندارد انجام می‌گیرد.

پیش‌گویی کریگ برای مقدار یك متغیر در مکان s_0 و واریانس پیش‌گویی از روابط زیر بدست می‌آیند:

$$p(z(s_0) = \hat{z}(s_0) = \{\gamma + (x'\Gamma^{-1}x)^{-1}(x - x'\Gamma^{-1}\gamma)\}^{-1}\Gamma^{-1}z)$$

$$(\sigma_k^2(s_0) = \gamma' \Gamma^{-1}\gamma - (x - x'\Gamma^{-1}\gamma)'(x'\Gamma^{-1}x)^{-1}(x - x'\Gamma^{-1}\gamma))$$

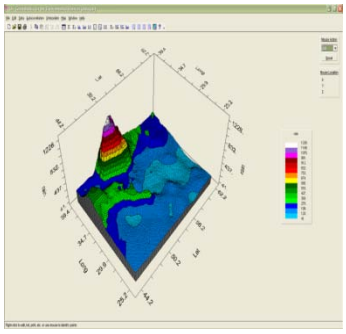
Γ ماتریسی $n \times n$ است که عنصر (i,j) ام آن $\gamma(s_i - s_j)$ می باشد و $\gamma =$ $(\gamma(s_0 - s_0), \dots, \gamma(s_0 - s_n))'$ وبازه پیش گویی ۹۵% برابر است با$\left(\hat{z}(s_0)\right) \mp 1.96\sigma_k^2(s_0)$ شکل ۹ و ۱۰ رویه مکانی برازش شده و شکل ۱۱ و ۱۲ مقادیر برآورد انحراف معیار بصورت دو بعدی و سه بعدی‌ها نشان می‌دهد. جهت صحه‌گذاری مدل پیش‌گویی میزان قدرت پیش‌گو از نمودار Actual Mean در مقابل Estimate Mean و نمودار توالی باقیمانده ها مطابق شکل ۱۳ و ۱٤ استفاده شده است که با توجه به معیارهای زیر مدل بدست آمده تقریبا مناسب می‌باشد و مطابق شکل ۱۲ تصادفی بودن باقیمانده‌ها مورد تایید می‌باشد و اما بعضی مشکوک به دور افتاده بودن می باشد.

$R^2 = 0.84$ MSD=54106.7 SEpred=126.65 SE=0.061 MAPE=0.561 MAD=139.46

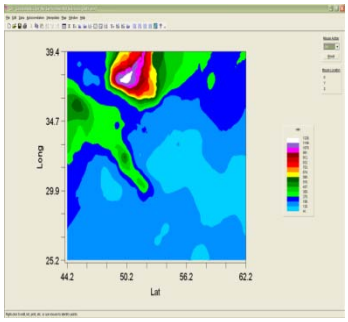
و در نهایت میزان بارش و انحراف استاندارد بعضی از ایستگاههایی سینوپتیک که گزارش نشده است پیش بینی شده است که به شرح جدول ۳ می باشد.

جدول۳- جدول پیش بینی

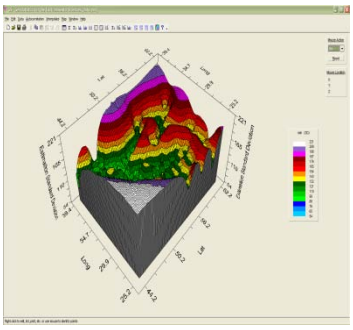
ردیف	نام ایستگاه	طول	عرض	مقدار پیش بینی	انحراف استاندارد
۱	سراباله(ایلام)	33.47	46.34	326	103
۲	برازجان(بوشهر)	29.15	51.1	163	148
۳	کیاشهر(گیلان)	37.25	49.33	1022	87
٤	پل دختر(لرستان)	33.03	47.43	274	102
٥	قروه(همدان)	35.10	47.48	315	132



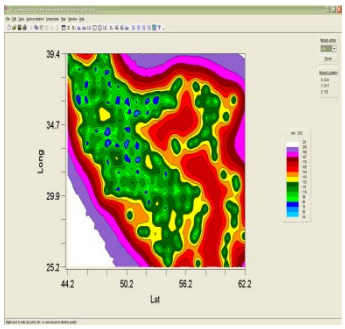
شکل ۹- نمودار سه بعدی رویه بارش ایران



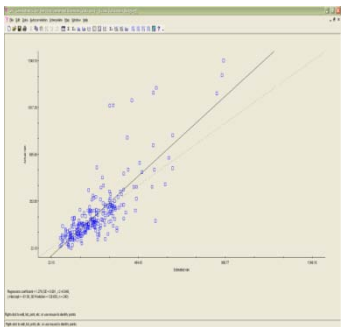
شکل ۱۰- نمودار دو بعدی رویه بارش ایران



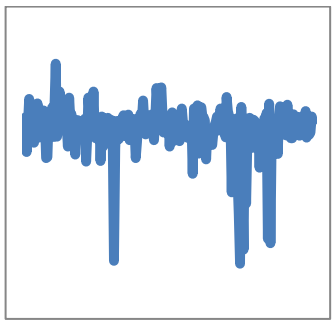
شکل ۱۱ - نمودار سه بعدی SDرویه بارش ایران



شکل ۱۲- نمودار دو بعدی SD رویه بارش ایران



شکل ۱۳- از نمودار Actual Mean در مقابل Estimate Mean



شکل ١٤ - نمودار توالی باقیمانده ها

مراجع:

۱. پیش گویی میزان تراکم آلاینده های HC و NOX در هوا : نظریه و کابرد : یزدانی ، افسانه پایانه کارشناسی ارشد . دانشگاه شهید بهشتی

۲. مبانی زمین آمار، علی مدنی مرکز نشر دانشگاه ایر کبیر واحد تفرش

۳. پیش گویی سطح آب حوزه آبریز دشت بیرجند به روش کریگیدن ، جلال اطمینان هشتمین کنفرانس بین المللی آمار ایران ، تابستان ۱۳۸۵

۴. پیش گویی بارندگی در استان همدان بر اساس داده های فضایی، بهزاد یعقوبیان، نهمین کنفرانس بین المللی آمار ایران ، تابستان ۱۳۸۷

۵. پیش گویی فضایی – زمانی سطح آب های زیر زمینی در دشت بیرجند، علی اصغر شفیعی، نهمین کنفرانس بین المللی آمار ایران ، تابستان ۱۳۸۷

6-Pardo – Iquzquiza.E(1998). Comparison of Geostatistical Methods for Estimation the Average Climatologically fain fall mean using Data on Precipitation and Topography, Int. J. Climatology.V18.P1031-1047

7-Gouvoets A. Auchinicloss, A.U. (2006). Performance Comparison of spatial and Space – time Interpolating techniques for Prediction of air Pollution Concentrations in the los Angeles.

8-kyriadis. P.C and Miller .N.L.(2004). A Spatial Time Series frame work for Simulation dally Precipitation at regional Scales Journal of Hydrology . 207 . 236-25