مدل‌های رگرسیون لجستیک و پروبیت فضایی برای تحلیل داده‌های یخ‌زدگی گیاهان در استان مازندران

وحید رضائی‌تبار، محسن محمدزاده[۱]

گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۸۸/۳/۵ تاریخ پذیرش: ۱۳۹۱/۳/۱۶

چکیده: تحلیل رگرسیون‌های لجستیک و پروبیت که در مدل‌بندی متغیرهای پاسخ دودویی به کار می‌روند 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 و لازم است همبستگی آن‌ها در تحلیل رگرسیون لجستیک و پروبیت منظور گردد. معمولا در آمار فضایی، پیش‌گویی فضایی برای داده‌های دودویی با استفاده از کریگیدن نشانگر صورت می‌پذیرد. در این روش لازم است ساختار همبستگی داده‌ها از طریق تغییرنگار نشانگر تعیین شود که به نوبه خود دشوار و موثر در کارایی نتایج است. در این مقاله پیش‌گویی داده‌های دودویی فضایی با استفاده از رگرسیون لجستیک و پروبیت فضایی با فرض این‌که مشاهدات روی یک مستطیل توری واقع شده و خطاهای مدل خودهمبسته کامل باشند با دو رهیافت بسامدی و بیزی انجام پذیرفته است. سپس کارایی پیش‌گویی توسط مدل‌های ارائه شده و روش کریگیدن نشانگر در یک مطالعه شبیه‌سازی مورد استفاده قرار گرفته است. در انتها نحوه کاربست آن‌ها برای داده‌های دما در ۲۴ مرکز هواشناسی استان مازندران بررسی و مدل های مناسب ارائه گردیده است.

واژه‌های کلیدی: داده های فضایی دودویی، رگرسیون لجستیک و پروبیت فضایی، مشبکه مستطیلی

رده‌بندی ریاضی: ۶۲J۱۲ و ۶۲J۰۲

۱-آدرس الکترونیکی نویسنده مسئول مقاله: محسن محمدزاده mohsen_m@modares.ac.ir

۱- مقدمه

در تحلیل رگرسیون با این فرض که خطاها نابسته و دارای توزیع نرمال با میانگین صفر و واریانس ثابت هستند، ارتباط بین متغیرهای پاسخ و تبیینی مورد بررسی قرار می‌گیرد. در مسائلی که متغیر پاسخ، دودویی با دو مقدار ۰ و ۱ است می‌توان از مدل های خطی تعمیم یافته، استفاده نمود [۱]. در این مدل، تابعی از میانگین متغیر پاسخ با متغیرهای تبیینی رابطه خطی دارد، که تابع پیوند[1] نامیده می‌شود. وقتی متغیر پاسخ دارای توزیع برنولی با احتمال موفقیت p باشد و تابع پیوند، لوجیت اختیار شود، مدل لجستیک حاصل می‌شود و چنانچه تابع پیوند، تابع توزیع تجمعی نرمال استاندارد باشد، مدل خطی تعمیم‌یافته را مدل پروبیت نامند. در عمل با موارد زیادی 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. معمولا داده‌های دودویی فضایی با استفاده از کریگیدن نشانگر[2] تحلیل می‌شوند، که اولین بار توسط مرجع [۲] به عنوان یک روش ناپارامتری برای پیش‌گویی غیرخطی از مشاهدات مورد استفاده قرار گرفت. باسو و رینسل [۳] با فرض آن‌که داده‌ها بر روی یک مستطیل توری مشاهده شده باشند و خطاهای مدل در هر موقعیت فقط وابسته به موقعیت‌های قبلی خود باشند، با روش ماکسیمم درست‌نمایی پارامترهای مدل را برآورد کردند. این مدل‌ها توسط رینسل و چینگ [۴] برای داده‌های فضایی بسط و توسعه داده شدند. همچنین رهیافت بیزی این مدل‌ها اولین بار در مطالعه انجام شده از سوی بست [۵] مورد توجه قرار گرفت و توسط بانرجی و همکاران [۶] توسعه داده شد.

یکی از مراحل مهم پیش‌گویی داده‌های فضایی دودویی، تعیین ساختار همبستگی داده‌ها از طریق تغییرنگار نشانگر است [۷]. در مقابل در مدل‌های رگرسیون لجستیک و پروبیت داده‌ها به‌صورت خودهمبسته در نظر گرفته می‌شوند و نیاز به تعیین ساختار همبستگی داده‌ها از طریق تغییرنگار ندارند. لذا هدف این مقاله معرفی مدل‌های رگرسیون لجستیک و پروبیت فضایی و مقایسه آن‌ها با کریگیدن نشانگر به لحاظ میزان دقت پیش‌گوهای حاصل از آن‌ها است. این مهم توسط مطالعه‌ای شبیه‌سازی صورت گرفته و نشان داده شده است برازش مدل-های رگرسیون لجستیک و پروبیت نه تنها ساده‌تر از کریگیدن نشانگر هستند بلکه برای داده-های فضایی دودویی که بر شبکه‌ای منظم مشاهده شده باشند از دقت بالاتری نیز برخوردارند.

در این مقاله روش باسو و رینسل برای مدل‌بندی داده‌های دودویی فضایی با رگرسیون‌های لجستیک و پروبیت فضایی برای حالتی که خطاهای مدل خودهمبسته کامل باشند با دو

1- Link Function
2- Indictor Kriging

رهیافت بسامدی و بیزی توسعه داده شده است. در نهایت، کارایی این مدل‌ها و کریگیدن نشانگر در یک مطالعه شبیه‌سازی مورد مقایسه قرار گرفته و نحوه کاربست آنها درخصوص داده‌های دما و یخزدگی گیاهان در ۲۴ مرکز هواشناسی استان مازندران بررسی شده است. بخش ۴ فقط به منظور ارائه نحوه تحلیل مدل‌های رگرسیون لجستیک و پروبیت با رهیافت بیزی ارائه شده و بدیهی است در شرایطی که فرض‌های اولیه مسئله برقرار باشند همواره این رهیافت نتایج بهتری از روش بسامدی ارائه می‌کند. در عین حال دقت پیش‌گوهای حاصل از روش‌های بسامدی و بیزی با روش اعتبارسنجی متقابل مورد مقایسه قرار گرفته و مدل برتر معرفی شده است.

۲- مدل‌های آماری

فرض کنید متغیر پاسخ دودویی y_{ij}، در مکان $s=(i,j)$ روی مشبکه مستطیلی $A_{mn}=\{(i,j): i=1,...,m, j=1,...,n\}$ دارای توزیع برنولی با پارامتر p_{ij} مشاهده شده باشد به‌طوری که

$$y_{ij}=x'_{ij}\beta+z_{ij} \qquad (i,j)\in A_{mn} \tag{۱}$$

که در آن، x_{ij} بردار k بعدی متغیرهای تبیینی، z_{ij} عبارت خطا و β بردار k بعدی ضرایب رگرسیون هستند. باسو و رینسل [۳] با فرض آن‌که خطاها در مدل (۱) خودهمبسته ناقص به‌صورت $z_{ij}=\alpha_1 z_{i-1,j}+\alpha_2 z_{i,j-1}+\alpha_3 z_{i-1,j-1}$ باشند، پارامترهای مدل را به‌روش ماکسیمم درست‌نمایی برآورد کردند. در این مقاله روش باسو و رینسل برای خطاهای خودهمبسته فضایی کامل به فرم‌های

$$z_{ij}^{(1)}=\alpha_1 z_{i-1,j}+\alpha_2 z_{i,j-1}+\alpha_3 z_{i+1,j}+\alpha_4 z_{i,j+1}+\varepsilon_{ij} \tag{۲}$$

$$\begin{aligned} z_{ij}^{(2)}=&\alpha_1 z_{i-1,j}+\alpha_2 z_{i,j-1}+\alpha_3 z_{i+1,j}+\alpha_4 z_{i,j+1}+\alpha_5 z_{i-1,j-1}\\ &+\alpha_6 z_{i+1,j-1}+\alpha_7 z_{i-1,j+1}+\alpha_8 z_{i+1,j+1}+\varepsilon_{ij}, \end{aligned} \tag{۳}$$

$$\begin{aligned} z_{ij}^{(3)}=&\alpha_1 z_{i-1,j}+\alpha_2 z_{i,j-1}+\alpha_3 z_{i+1,j}+\alpha_4 z_{i,j+1}-\alpha_1\alpha_2 z_{i-1,j-1}\\ &-\alpha_2\alpha_3 z_{i+1,j-1}-\alpha_1\alpha_4 z_{i-1,j+1}-\alpha_2\alpha_4 z_{i+1,j+1}+\varepsilon_{ij}, \end{aligned} \tag{۴}$$

توسعه داده می‌شود. برای این کار مدل (۱) با خطاهای خودهمبسته فضایی کامل را می‌توان به‌صورت کلی

$$y_{ij} = x'_{ij}\beta + z_{ij}^{(w)}, \quad w = ۱,۲,۳ \qquad (۵)$$

با امید ریاضی و واریانس زیر نوشت.

$$p_{ij}^{(w)} = E\left(y_{ij}\right) = x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}, \quad w = ۱,۲,۳ \qquad (۶)$$

$$Var\left(y_{ij}\right) = p_{ij}^{(w)}\left(۱ - p_{ij}^{(w)}\right).$$

مدل لجستیک فضایی

با انتخاب تابع پیوند لوجیت، مدل رگرسیون لجستیک فضایی به صورت

$$\text{logit}\left(p_{ij}^{(w)}\right) = \ln\frac{p_{ij}^{(w)}}{۱ - p_{ij}^{(w)}} = x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}, \quad w = ۱,۲,۳ \qquad (۷)$$

است. برای نمونه تصادفی $y = \left(y_{۱۱},...,y_{mn}\right)$، که در آن y_{ij} از توزیع برنولی با پارامتر $P\left(y_{ij} = ۱\right) = p_{ij}$ پیروی می‌کند، با قرار دادن $p = \left(p_{۱۱},...,p_{mn}\right)$ تابع درست‌نمایی به صورت

$$\begin{aligned} L(p;y) &= p\left(y_{۱۱} \mid y_{۱۲},...,y_{mn};p\right)p\left(y_{۱۲} \mid y_{۱۳},...,y_{mn};p\right)...p\left(y_{mn};p\right) \\ &= p_{۱۱}^{y_{۱۱}}\left(۱-p_{۱۱}\right)^{۱-y_{۱۱}}\sum_{y_{۱۱}} p\left(y_{۱۱} \mid y_{۱۲} \mid y_{۱۳},...,y_{mn};p\right)... \sum_{y_{۱۱},...,y_{mn-۱}} p\left(y_{۱۱},...,y_{mn};p\right) \end{aligned}$$

خواهد بود که به‌دلیل پیچیدگی، ماکسیمم کردن آن دشوار است. لذا از لگاریتم تابع شبه-درست‌نمایی یعنی ضرب توابع چگالی کناری شرطی به‌صورت

$$\ell(p,y) = \sum_{(i,j)\in A_{mn}} y_{ij}\ln\frac{p_{ij}}{۱-p_{ij}} + \sum_{(i,j)\in A_{mn}} \ln\left(۱-p_{ij}\right) \qquad (۸)$$

استفاده می شود [۸]. اکنون با جایگذاری (۷) در رابطه (۸) لگاریتم تابع شبه‌درست‌نمایی برای β به‌صورت

$$\ell(\beta,\alpha,y) = \sum_{(i,j)\in A_{mn}} y_{ij}c_{ij}^{(w)} - \sum_{(i,j)\in A_{mn}} \ln\left(۱+e^{c_{ij}^{(w)}}\right) \qquad (۹)$$

حاصل می‌شود، که در آن

$$c_{ij}^{(w)} = x'_{ij}\beta + z_{ij}^{(w)}, \qquad w = ۱,۲,۳.$$

اکنون با مشتق گرفتن از رابطه (۹) نسبت به پارامترهای α و β و حل معادلات به روش‌های عددی برآورد ضرایب مدل، امید ریاضی و واریانس آن‌ها قابل محاسبه‌اند.

مدل پروبیت فضایی:

چنانچه تابع پیوند پروبیت اختیار شود، مدل پروبیت فضایی به‌صورت

$$\text{probit}\left(p_{ij}^{(w)}\right)=\Phi^{-1}\left(p_{ij}^{(w)}\right)=x'_{ij}\beta+z_{ij}^{(w)}-\varepsilon_{ij} \qquad (۱۰)$$

خواهد بود، که در آن $\Phi^{-1}(.)$ معکوس تابع توزیع تجمعی نرمال استاندارد است. در این‌صورت با جایگذاری

$$p_{ij}^{(w)}=\int_{-\infty}^{x'_{ij}\beta+z_{ij}^{(w)}-\varepsilon_{ij}}\frac{1}{\sqrt{2}}e^{\frac{-1}{2}x^2}dx \qquad (i,j)\in A_{mn}$$

در رابطه (۸)، لگاریتم تابع شبه‌درست‌نمایی براساس مدل پروبیت به‌دست می‌آید.

کریگیدن نشانگر: فرض کنیم $N=nm$ و $I=\{I(s_1,z),...,I(s_N,z)\}$ مشاهدات فضایی دودویی از میدان تصادفی $\{I(s_1,z):s\in A_{nm}\}$ با نیم‌تغییرنگار نشانگر $\gamma_z(h)$ باشند، که در آن آستانه z به صورت

$$I(s,z)=\begin{cases}1 & Z(s)\le z\\ 0 & Z(s)>z\end{cases}$$

تعریف می‌شود. در این‌صورت کریگیدن نشانگر این میدان تصادفی در یک موقعیت جدید s_0 براساس مشاهدات I عبارتست از

$$\hat{I}(s_0,z)=\sum_{i=1}^{N}\lambda_i(z)I(s_i,z)$$

به‌طوری که بردار ضرایب آن با شرط $\sum_{i=1}^{N}\lambda_i(z)=1$ از رابطه

$$\lambda'_z=\left(\gamma_z+J\frac{1-J'\Gamma_z^{-1}\gamma_z}{J'\Gamma_z^{-1}J}\right)\Gamma_z^{-1} \qquad (۱۱)$$

به‌دست می‌آید، که در آن $\gamma_z = \left(\gamma_z\left(s_1 - s_0\right),...,\gamma_z\left(s_N - s_0\right)\right)'$، Γ_z ماتریسی $N \times N$ با (i,j) امین درایه $\gamma_z\left(s_i - s_j\right)$ و J یک بردار N تایی با درایه های ۱ است. میانگین توان دوم خطای پیش‌گو نیز به‌صورت زیر تعیین می‌شود [۷]:

$$\sigma^2\left(s_0, z\right) = \gamma_z' \Gamma_z^{-1} \gamma_z - \frac{\left(J' \Gamma_z^{-1} \gamma_z - 1\right)^2}{J' \Gamma_z^{-1} J}$$

۳- تحلیل مدل‌های رهیافت بیزی

پاسکوتو [۹] و بیلی [۱۰] تحلیل داده‌های دودویی فضایی با رهیافت بیزی را مورد مطالعه قرار داده و بانرجی و همکاران [۶] از آن در اپیدمولوژی استفاده کردند، که در این بخش با الهام از روش آن‌ها رگرسیون‌های لجستیک و پروبیت فضایی بیزی ارائه می‌شوند. فرض کنیم متغیر پاسخ y_{ij} در موقعیت $(i,j) \in A_{mn}$ دارای توزیع برنولی با احتمال موفقیت

$$P_{ij}^{(w)} = \text{logit}^{-1}\left(x_{ij}'\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right) = \frac{\exp\left(x_{ij}'\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)}{1 + \exp\left(x_{ij}'\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)}$$

است. با فرض آن که بردار β دارای توزیع پیشین نرمال k متغیره $N_k(\theta, \Sigma)$ باشد، توزیع پسین آن به‌صورت پیچیده‌ی

$$\begin{aligned}
\pi(\beta \mid y) &\propto \phi_k(\beta; \theta, \Sigma)\, p\left(y_{11},...,y_{mn}\right) \\
&= \phi_k(\beta; \theta, \Sigma)\, p\left(y_{11} \mid y_{12},...,y_{mn}\right) p\left(y_{12} \mid y_{13},...,y_{mn}\right)...p\left(y_{mn}\right) \\
&= \phi_k(\beta; \theta, \Sigma)\, p_{11}^{y_{11}}\left(1 - p_{11}\right)^{1-y_{11}} \sum_{y_{11}} \left(y_{11} \mid y_{12} \mid y_{13},...,y_{mn}\right)... \sum_{y_{11},...,y_{mn-1}} p\left(y_{11},...,y_{mn}\right)
\end{aligned}$$

خواهد بود، که در آن $\phi_k(\beta; \theta, \Sigma)$ چگالی توزیع نرمال k متغیره با میانگین θ و ماتریس کواریانس Σ است. با استفاده از تابع شبه‌درست‌نمایی

$$\begin{aligned}
p\left(y_{11},\ldots,y_{mn}; p\right) \approx\; & p\left(y_{11} \mid p_{12},\ldots,y_{mn}; p\right) \\
& \times p\left(y_{12} \mid p_{11}, y_{1,3} \ldots, y_{mn}; p\right) \\
& \vdots \\
& \times p\left(y_{mn} \mid y_{11},\ldots,y_{m}-1_{n}; p\right)
\end{aligned}$$

$$= p_{11}^{y_{11}} \left(1 - p_{11}\right)^{1-y_{11}} \ldots p_{mn}^{y_{mn}} \left(1 - p_{mn}\right)^{1-y_{11}}$$
$$\prod_{ij} \text{logit}^{-1}\left(x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)^{y_{ij}} \left[1 - \text{logit}^{-1}\left(x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)\right]^{(1-y_{ij})}$$

توزیع پسینی به‌صورت

$$\pi(\beta \mid y) \propto \prod_{ij} \text{logit}^{-1}\left(x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)^{y_{ij}} \left[1 - \text{logit}^{-1}\left(x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)\right]^{(1-y_{ij})} \quad (۱۲)$$

خواهد شد. برای نمونه‌گیری از توزیع پسین (۱۲) با الگوریتم متروپلیس– هستینگز می‌توان توزیع t-استیودنت چند متغیره با k درجه ازادی به صورت

$$q(\beta \mid y) \propto \left[1 + k^{-1}(\beta - \beta_0)' \Sigma^{-1} (\beta - \beta_0)\right]^{-\frac{k+p}{2}}$$

را به عنوان تابع چگالی پیشنهادی انتخاب کرد [۱۱]. برای کم کردن حساسیت توزیع پسینی به ابرپارامترها، از توزیع‌های پیشینی زیر استفاده شده است:

$$y_{ij} \sim Ber\left[\text{logit}^{-1}\left(x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)\right]$$
$$\beta \sim N_k(\theta, \Sigma)$$
$$Z_{ij} \mid \sigma^2 \sim N\left(0, 100\sigma^2\right)$$
$$\sigma^2 \sim f\left(\sigma^2\right) = \left(1 + \sigma^2\right)^{-2}.$$

چون اطلاعاتی در مورد پارامترهای توزیع پیشینی در اختیار نیست و از طرفی اگر پیشینی ناسره برای پارامترها در نظر گرفته شود، سره بودن توزیع پسینی باید اثبات شود، برای $Z_{ij} \mid \sigma^2$ توزیع ناآگاهی بخش در نظر گرفته شده است که واریانس آن بسیار زیاد و معادل $100\sigma^2$ است. در این صورت با توجه به فرض استقلال β و σ^2 توزیع پسینی به‌صورت

$$\pi\left(\beta, Z_{ij}, \sigma^2 \mid y_{ij}\right) \propto p\left(y_{ij} \mid \beta, Z_{ij}, \sigma^2\right) \pi\left(\beta, Z_{ij}, \sigma^2\right)$$
$$= p\left(y_{ij} \mid \beta, Z_{ij}, \sigma^2\right) \pi\left(Z_{ij} \mid \sigma^2\right) \pi(\beta) \pi\left(\sigma^2\right)$$
$$\propto \prod_{ij} C\ Ber\left[\log it^{-1}\left(x'_{ij}\beta + z_{ij}^{(w)}\right)\right] \phi_k(\beta; \theta, \Sigma) \phi\left(Z_{ij}; 0, 100\sigma^2\right) \left(1 + \sigma^2\right)^{-2}$$

خواهد بود. برای نمونه‌گیری از توزیع پسینی به منظور برآورد پارامترها از الگوریتم نمونه‌گیری گیبز استفاده می‌شود. برای این منظور توزیع‌های شرطی کامل به صورت زیر به دست آمده‌اند:

$$\pi\left(\beta \mid y_{ij}, Z_{ij}, \sigma^{۲}\right) \propto \prod_{ij} C_{۱} Ber\left[\log it^{-۱}\left(x'_{ij}\beta + z_{ij}^{(w)}\right)\right] \phi_k\left(\beta; \theta, \Sigma\right)$$

$$\pi\left(Z_{ij} \mid y_{ij}, \beta, \sigma^{۲}\right) \propto \prod_{ij} C_{۲} Ber\left[\log it^{-۱}\left(x'_{ij}\beta + z_{ij}^{(w)}\right)\right] \phi\left(Z_{ij}; ۰, ۱۰۰\sigma^{۲}\right)$$

$$\pi\left(\sigma^{۲} \mid y_{ij}, \beta, Z_{ij}\right) \propto \prod_{ij} C_{۲} Ber\left[\log it^{-۱}\left(x'_{ij}\beta + z_{ij}^{(w)}\right)\right]\left(۱ + \sigma^{۲}\right)^{-۲}.$$

برای مدل پروبیت فضایی با فرض‌های

$$y_{ij} \sim Ber\left[\Phi^{-۱}\left(x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)\right],\ w = ۱, ۲, ۳$$

$$\beta \sim N_k\left(\theta, \Sigma\right)$$

توزیع پسینی به صورت

$$\pi(\beta \mid y) \propto \phi_k\left(\beta; \theta, \Sigma\right) \prod_{ij} \Phi^{-۱}\left(x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)^{y_{ij}} \times \left[۱ - \Phi^{-۱}\left(x'_{ij}\beta + z_{ij}^{(w)} - \varepsilon_{ij}\right)\right]^{(۱ - y_{ij})} \quad (۱۳)$$

خواهد شد، که در این حالت نیز برای نمونه‌گیری از توزیع پسینی از الگوریتم متروپلیس–هستینگز با تابع چگالی کاندید t-استیودنت چند متغیره با درجه آزادی k استفاده می‌شود. برای برآورد β با تابع زیان درجه دوم می‌توان از توزیع‌های شرطی کامل نمونه‌گیری کرده، میانگین آن‌ها را به عنوان برآورد پارامترهای مورد نظر استفاده نمود.

۴- ارزیابی مدل‌ها

در این مقاله برای ارزیابی پیش‌گویی فضایی با مدل های رگرسیون لجستیک، پروبیت و کریگیدن نشانگر از ملاک ریشه دوم میانگین توان دوم خطاهای اعتبارسنجی متقابل و ملاک آکائیک[1] (AIC) و برای ارزیابی مدل‌های رگرسیون لجستیک و پروبیت حاصل از توزیع‌های پیشینی مختلف، از ملاک اطلاع شوارتز یا ملاک اطلاع بیزی[2] (AIC) استفاده شده است.

در روش اعتبارسنجی متقابل یک مشاهده را از مجموعه مشاهدات حذف نموده و مقدار آن براساس $N - ۱$ مشاهده باقیمانده پیش‌گویی می‌شود. این عمل را برای تمام N موقعیت تکرار کرده و ملاک ریشه دوم میانگین توان دوم خطاهای اعتبارسنجی متقابل به صورت

1- Akaike Information Criterion
2- Bayes Information Criterion

$$RMSE = \left[\frac{1}{N} \sum_{i=1}^{N} e_{(-i)}^{2} \right]^{\frac{1}{2}} \qquad (۱۴)$$

محاسبه می‌شود، که در آن $e_{(-i)}$ خطای پیش‌گوی محاسبه شده براساس مشاهدات به جز مشاهده i ام است. بدیهی است هر چقدر مقدار این ملاک به صفر نزدیکتر باشد، پیش‌گو از دقت بیشتری برخوردار است.

معیارهای AIC و BIC برای داده‌های فضایی به ترتیب به صورت

$$\mathrm{AIC} = -2\ell p\left(\hat{p}, y\right) + 2k$$

$$\mathrm{BIC} = -2\ell p\left(\hat{p}, y\right) + k \log N$$

می‌باشند، که در آن‌ها $\ell p\left(\hat{p}, y\right)$ ماکسیمم لگاریتم تابع شبه‌درست‌نمایی و k تعداد پارامترها است [۱۲]. هر چه مقدار AIC و BIC برای یک مدل کمتر باشد، برازش مدل به داده‌ها مناسب‌تر و بر مدل‌های رقیب ترجیح دارد.

۵- شبیه‌سازی

در این بخش مدل‌های رگرسیون لجستیک، پروبیت فضایی و کریگیدن نشانگر براساس معیارهای RMSE و BIC مورد مقایسه قرار می‌گیرند. برای این منظور x_1 از توزیع نرمال با میانگین صفر و واریانس‌های ۰/۵، ۱ و ۲ و x_2 از توزیع خی‌دو با ۵ درجه آزادی، به عنوان متغیرهای تبیینی در نظر گرفته شده‌اند. با توجه به رابطه (۷) و اختیار کردن $w = 1$ مقادیر $\{z_{i-1,j}, z_{i,j-1}, z_{i+1,j}, z_{i,j+1}\}$ از توزیع نرمال استاندارد و مقادیر y_{ij} نیز از توزیع برنولی با احتمال موفقیت $P_{ij}^{(1)}$ تولید شده‌اند. با استفاده از الگوریتم نیوتن رافسون و مقادیر اولیه $\beta_0 = 0/05$، $\beta_1 = -0/05$، $\beta_2 = 0/025$، $\alpha_1 = 0/01$، $\alpha_2 = 0/02$، $\alpha_3 = -0/001$ و $\alpha_4 = 0/004$ پارامترهای $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ و $\beta = (\beta_0, \beta_1, \beta_2)$ برآورد شده‌اند. برای پنج حجم نمونه‌ی ۲۵، ۵۰، ۱۰۰، ۲۰۰ و ۵۰۰ شبیه‌سازی ۱۰۰۰۰ بار تکرار شده و مقادیر RMSE محاسبه شده‌اند. از آنجا که مقادیرچارک‌های RMSE بیان‌گر متوسط و کران-های پایین و بالای خطا هستند، مقادیر آن‌ها برای واریانس های ۰/۵، ۱ و۲ محاسبه و در جداول ۱ تا ۳ ارائه شده‌اند. همان‌طور که ملاحظه می‌شود مقادیر RMSE با افزایش حجم نمونه برای هر سه مدل رگرسیون لجستیک، پروبیت فضایی و کریگیدن نشانگر کاهش می‌یابند. ولی مدل‌های رگرسیون لجستیک، پروبیت فضایی نسبت به کریگیدن نشانگر همواره از خطای کمتری برخوردار هستند. این نتایج دور از انتظار نیستند؛ زیرا با توجه به توری بودن داده‌ها

انتظار می‌رود مدل‌های رگرسیون لجستیک، پروبیت فضایی با خطاهای خودهمبسته، برازش مناسب‌تری نسبت به کریگیدن نشانگر داشته باشند.

جدول ۱: مقدار RMSE لجستیک، پروبیت فضایی و کریگیدن نشانگر برای $\sigma^2 = ۰/۵$

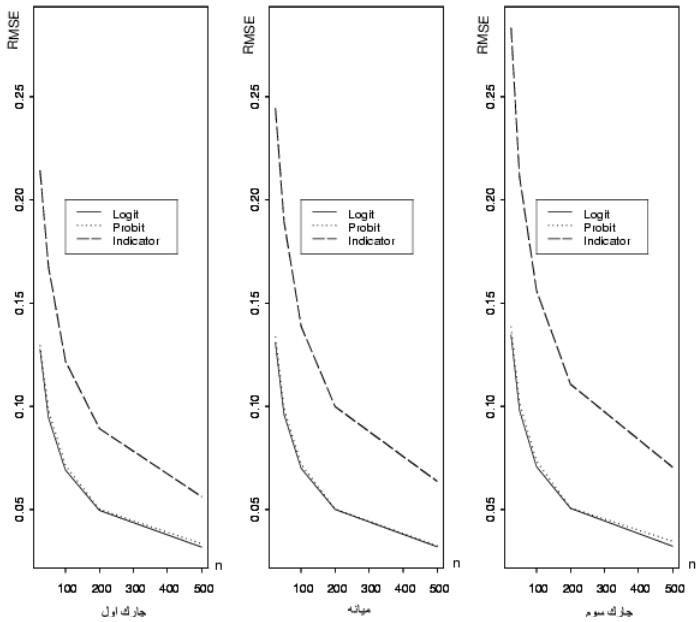
حجم نمونه	لجستیک فضایی			پروبیت فضایی			کریگیدن نشانگر		
N	چارک اول	میانه	چارک سوم	چارک اول	میانه	چارک سوم	چارک اول	میانه	چارک سوم
۲۵	۰/۱۲۳	۰/۱۵۸	۰/۱۹۸	۰/۱۲۵	۰/۱۵۳	۰/۱۸۹	۰/۲۱۶	۰/۲۱۸	۰/۲۱۹
۵۰	۰/۰۹۷	۰/۱۲۴	۰/۱۵۷	۰/۰۹۹	۰/۱۲۵	۰/۱۵۹	۰/۱۹۰	۰/۱۹۱	۰/۱۹۳
۱۰۰	۰/۰۷۳	۰/۰۹۹	۰/۱۲۳	۰/۰۷۹	۰/۰۸۹	۰/۱۱۰	۰/۱۶۷	۰/۱۶۸	۰/۱۷۰
۲۰۰	۰/۰۵۳	۰/۰۶۶	۰/۰۸۱	۰/۰۵۷	۰/۰۷۱	۰/۰۸۸	۰/۱۴۸	۰/۱۴۹	۰/۱۵۱
۵۰۰	۰/۰۳۳	۰/۰۴۱	۰/۰۵۱	۰/۰۳۶	۰/۰۴۵	۰/۰۵۲	۰/۰۷۱	۰/۰۷۵	۰/۰۸۹

جدول ۲: مقدار RMSE لجستیک، پروبیت فضایی و کریگیدن نشانگر برای $\sigma^2 = ۱$

حجم نمونه	لجستیک فضایی			پروبیت فضایی			کریگیدن نشانگر		
N	چارک اول	میانه	چارک سوم	چارک اول	میانه	چارک سوم	چارک اول	میانه	چارک سوم
۲۵	۰/۱۲۳	۰/۱۶۲	۰/۱۹۴	۰/۱۳۸	۰/۱۶۳	۰/۲۰۱	۰/۲۱۷	۰/۲۱۸	۰/۲۲۰
۵۰	۰/۱۱۰	۰/۱۲۴	۰/۱۵۸	۰/۱۱۱	۰/۱۲۵	۰/۱۵۹	۰/۱۹۲	۰/۱۹۵	۰/۱۹۹
۱۰۰	۰/۰۷۱	۰/۰۹۰	۰/۱۳۲	۰/۰۷۵	۰/۰۹۹	۰/۱۱۴	۰/۱۷۰	۰/۱۷۳	۰/۱۷۸
۲۰۰	۰/۰۵۵	۰/۰۶۹	۰/۰۸۲	۰/۰۵۶	۰/۰۷۱	۰/۰۸۴	۰/۱۴۹	۰/۱۵۵	۰/۱۵۸
۵۰۰	۰/۰۳۷	۰/۰۴۵	۰/۰۵۷	۰/۰۳۸	۰/۰۴۶	۰/۰۵۹	۰/۱۳۳	۰/۱۳۵	۰/۱۳۶

جدول ۳: مقدار RMSE لجستیک، پروبیت فضایی و کریگیدن نشانگر برای $\sigma^2 = ۲$

حجم نمونه	لجستیک فضایی			پروبیت فضایی			کریگیدن نشانگر		
N	چارک اول	میانه	چارک سوم	چارک اول	میانه	چارک سوم	چارک اول	میانه	چارک سوم
۲۵	۰/۱۳۵	۰/۱۳۶	۰/۱۳۸	۰/۱۴۱	۰/۱۷۱	۰/۲۱۴	۰/۲۱۹	۰/۲۴۶	۰/۲۸۳
۵۰	۰/۱۲۰	۰/۱۲۳	۰/۱۳۰	۰/۱۹۴	۰/۱۹۶	۰/۱۹۷	۰/۱۹۷	۰/۱۹۹	۰/۲۱۱
۱۰۰	۰/۰۸۸	۰/۰۹۱	۰/۱۳۵	۰/۰۹۰	۰/۰۹۹	۰/۱۷۱	۰/۱۷۵	۰/۱۸۳	۰/۱۸۹
۲۰۰	۰/۰۵۷	۰/۰۷۵	۰/۰۸۹	۰/۰۶۸	۰/۰۷۸	۰/۰۹۲	۰/۱۵۱	۰/۱۶۱	۰/۱۶۵
۵۰۰	۰/۰۴۰	۰/۰۴۷	۰/۰۵۹	۰/۰۵۱	۰/۰۶۶	۰/۰۶۹	۰/۱۴۰	۰/۱۴۲	۰/۱۵۰



شکل ۱: مقادیر RMSE لجستیک، پروبیت فضایی و کریگیدن نشانگر

همچنین با مقایسه مقادیر جداول ۱ تا ۳ ملاحظه می‌شود که با افزایش واریانس متغیرهای تبیینی، مقادیر RMSE افزایش پیدا می‌کند، که علت آن افزایش تغییرات مقادیر متغیرهای پاسخ و در نتیجه افزایش مقدار خطا به دلیل پراکندگی مقادیر متغیرهای تبیینی است. شکل ۱ که نمودار مقادیر RMSE را برای خطای خودهمبسته $z_{ij}^{(۲)}$ نشان می‌دهد، نشانگر دقت بیشتر مدل‌های لجستیک و پروبیت فضایی نسبت به کریگیدن نشانگر هستند. زیرا این مدل‌ها نسبت به مدل با خطای $z_{ij}^{(۱)}$ کامل‌تر هستند. بنابراین وقتی موقعیت داده‌ها به صورت توری باشند، دو مدل پروبیت و لجستیک فضایی با خطاهای خودهمبسته کامل عمل‌کرد بهتری دارد و مقادیر پیش‌گویی دقیق‌تری نسبت به کریگیدن نشانگر ارائه می‌کنند.

۶- مثال کاربردی

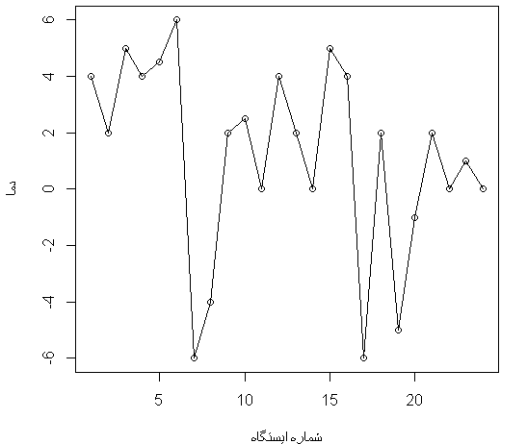
محمدزاده، و کیاپور [۱۳] داده‌های دما در ۲۴ ایستگاه هواشناسی استان مازندران را به عنوان مجموعه‌ای از داده‌های دودویی فضایی مورد مطالعه قرار داده و با استفاده از کریگیدن نشانگر، احتمال یخزدگی در ایستگاه‌ها را به‌دست آورده‌اند. در این بخش با استفاده از مدل‌های

رگرسیون لجستیک و پروبیت فضایی نقشه پیش‌گویی فضایی احتمال یخزدگی در استان مازندران تعیین و دقت آن با نتایج به‌دست آمده توسط مرجع [۱۳] مورد مقایسه قرار می‌گیرد.

از آنجا که احتمال یخزدگی نواحی نزدیک در هر منطقه به یکدیگر مرتبط هستند و این ارتباط در نواحی دورتر از هم کاهش می‌یابد، می‌توان از مدل‌های رگرسیون لجستیک و پروبیت فضایی با خطاهای خودهمبسته کامل نیز برای پیش‌گویی احتمال یخزدگی در هر موقعیت استفاده کرد. شکل ۲ موقعیت قرار گرفتن مشاهدات دمای روز پانزدهم بهمن ماه ۱۳۸۴ را در ۲۴ ایستگاه هواشناسی استان مازندران برحسب سانتی‌گراد نشان می‌دهد. در داده‌های نشانگر، کد یک به ناحیه یخزده (دمای منفی) و کد صفر به ناحیه یخنزده اختصاص داده شده است.

شکل ۲: موقعیت ایستگاه‌های هواشناسی استان مازندران

با فرض آن‌که دما از مدل خودهمبسته کامل پیروی می‌کند، موقعیت داده‌ها در یک مستطیل توری به صورت $A_{mn}=\{(i,j):i=۱,...,۴,j=۱,...,۴\}$ در نظر گرفته شده است. میزان دما بدون در نظر گرفتن محل قرار گرفتن آن‌ها نیز در شکل ۳ نشان داده شده است. در مدل‌های رگرسیون لجستیک و پروبیت فضایی به ازای $i,j=۱$، مقادیر اولیه‌ای برای دما مورد نیاز است. بنابراین هر چه مرتبه خطا در مدل‌های رگرسیون لجستیک و پروبیت فضایی بالاتر باشد، مقادیر اولیه زیادتری مورد نیاز است. لذا امکان بروز خطا افزایش می‌یابد. بنابراین لزوما افزایش مرتبه یک مدل، به معنی بهتر بودن مدل نیست، بلکه باید به همبستگی آن از لحاظ مکانی با سایر مشاهدات توجه شود. با توجه به شکل ۳ ملاحظه می‌شود که مشاهدات دما در ۲۴ شهر استان مازندران به‌صورت تصادفی در بازه [۶ و ۶-] توزیع شده‌اند. بنابراین یک ایده مناسب برای تولید مقادیر اولیه می‌تواند اختیار یک نمونه تصادفی در این بازه باشد.



شکل ۳: نمودار میزان دما در ۲۴ شهر استان مازندران

جدول ۴: معیارهای ارزیابی مقایسه مدل

مدل	نوع خطا	RMSE	AIC
پروبیت فضایی	۱	۰/۴۰	۲۰/۷۷۸
	۲	۰/۱۴	۱۵/۱۱۸
	۳	۰/۱۸	۱۲/۶۵۴
لجستیک فضایی	۱	۰/۴۳	۲۲/۲۹۱
	۲	۰/۲۳	۱۷/۲۲۰
	۳	۰/۲۶	۱۹/۷۱۷
کریگیدن نشانگر	-	۰/۴۸	۲۵/۳۷۱

مقادیر ریشه دوم میانگین توان دوم خطاها و ملاک آکائیک برای مدل‌های رگرسیون لجستیک و پروبیت فضایی در جدول ۴ ارائه شده است. با توجه به این جدول، مدل مناسب برازنده شده به داده‌های دما، بر اساس RMSE، مدل پروبیت فضایی با خطاهای خودهمبسته $z_{ij}^{(۲)}$ و براساس AIC ، مدل پروبیت فضایی با خطاهای خودهمبسته $z_{ij}^{(۳)}$ است. چون مدل با خطای $z_{ij}^{(۳)}$ از پارامترهای کمتری نسبت به مدل با خطای $z_{ij}^{(۲)}$ برخوردار است، برای تحلیل داده‌های دما مناسب‌تر و به‌صورت زیر است:

$$\begin{aligned}\text{Probit}\left(P_{ij}^{(۳)}\right) = &-۰/۰۱۲۸ - ۰/۰۹۳۴x_{ij} + ۰/۰۵۶۱z_{i-۱,j} - ۰/۱۰۱۲z_{i,j-۱} + ۰/۰۷۶۷z_{i+۱,j} \\ &- ۰/۰۰۰۸z_{i,j+۱} + ۰/۰۰۵۶z_{i-۱,j-۱} + ۰/۰۰۷۷z_{i+۱,j-۱} + ۰/۰۰۰۴z_{i-۱,j+۱} \\ &- ۰/۰۹۰۸z_{i+۱,j+۱}, \qquad i = ۱,...,۴,\ j = ۱,...,۶\end{aligned} \tag{۱۵}$$

با توجه به مدل (۱۵)، آستانه تحمل گیاهان به ازای هر متر افزایش ارتفاع، به‌صورت ضریبی از ۰/۰۹۳۴ کاهش می‌یابد.

برای تحلیل بیزی داده‌های دما با توزیع پیشینی $\beta \sim N_k\left(۰,۱۰^۴ I\right)$، برای تولید نمونه از توزیع پسینی

$$\begin{aligned}\pi(\beta \mid y) = &\frac{۱}{p(y)} \prod_{ij} \log it^{-۱}\left[x_{ij}'\beta + z_{ij}^{(w)} - \varepsilon_{ij}^{(w)}\right]^{y_{ij}} \\ &\times\left[۱ - \log it^{-۱}\left(x_{ij}'\beta + z_{ij}^{(w)} - \varepsilon_{ij}^{(w)}\right)\right]^{(۱-y_{ij})} \phi_k\left(\beta; \alpha, ۰, ۱۰^۴ I\right),\ w = ۱,۲,۳\end{aligned}$$

از الگوریتم متروپلیس- هستینگز استفاده شده است و توزیع t-استیودنت چند متغیره با درجه آزادی $k = ۵$، تعداد متغیرهای $p = ۶$ و $\Sigma = ۱۰^۴ k$ به‌صورت

$$q(\beta \mid y) \propto \left[۱ + k^{-۱}\left(\beta - \beta_۰\right)' \Sigma^{-۱}\left(\beta - \beta_۰\right)\right]^{-\frac{k+p}{۲}}$$

به‌عنوان تابع چگالی کاندید انتخاب شده‌اند، که در آن $\beta_۰$ بردار مقادیر اولیه است. علت این انتخاب را می‌توان کلفت‌تر بودن دم‌های تابع چگالی کاندید نسبت به توزیع نرمال دانست، که با انتخاب درجه آزادی ۵ این امر میسر شده است. همچنین $\hat{p}(y) = \left[\frac{۱}{k}\sum_{i=۱}^{k} \frac{۱}{p(y \mid \beta_i)}\right]^{-۱}$ اختیار شده است. مقادیر BIC_p برای مدل‌های لجستیک، پروبیت فضایی و کریگیدن نشانگر در جدول ۵ بیان‌گر آن است که مدل لجستیک فضایی با خطای $z_{ij}^{(۱)}$ به صورت

$$\begin{aligned}\log it\left(P_{ij}^{(۱)}\right) = &۱/۰۰۴ + ۰/۲۵۴۱x_{ij} - ۰/۰۵۶z_{i-۱,j} + ۰/۰۶۷z_{i,j-۱} \\ &+ ۰/۰۹۲۳z_{i+۱,j} + ۰/۱۰۳۵z_{i,j+۱}\end{aligned} \tag{۱۶}$$

مدل مناسبی برای برازش به داده های دما است. برای مقایسه دقت پیش‌گوهای حاصل از دو مدل بسامدی و بیزی، ملاک ریشه دوم میانگین توان دوم خطای اعتبار سنجی متقابل (RMSE) محاسبه و به‌ترتیب مقادیر ۰/۱۸ و ۰/۱۳ حاصل شدند، که بیان‌گر برتری پیش‌گوی بیزی بر پیش‌گوی بسامدی است.

جدول ۵: مقادیر معیار ارزیابی BIC_p

مدل	نوع خطا	BIC_p
پروبیت فضایی	۱	-۴/۶۴۴
	۲	-۲/۴۴۱
	۳	-۱/۸۹۶
لجستیک فضایی	۱	-۱۱/۴۱۱
	۲	-۶/۸۸۹
	۳	-۴/۳۰۷
کریگیدن نشانگر	-	۸/۸۸۸

۷- بحث و نتیجه‌گیری

برای مدل‌بندی داده‌های دودویی فضایی می‌توان همانند کریگیدن نشانگر، از مدل‌های رگرسیون لجستیک یا پروبیت فضایی با خطاهای خودهمبسته فضایی نیز استفاده کرد. در این مقاله با فرض توری بودن ناحیه شامل موقعیت مشاهدات، مدل‌های مذکور برای پیش‌گویی فضایی داده‌های دودویی پیشنهاد گردید. در مطالعه شبیه‌سازی نشان داده شد که برای داده-های دودویی فضایی با موقعیت‌های توری، استفاده از مدل‌های رگرسیون لجستیک و پروبیت فضایی نسبت به کریگیدن نشانگر از خطای کمتری برخوردار است.

ویژگی تحلیل داده‌های واقع بر یک مستطیل توری بی‌نیازی از برآورد ساختار همبستگی داده‌ها در قالب توابع تغییرنگار یا هم‌تغییرنگار است. در این صورت ساختار همبستگی به صورت ساده خودهمبسته در نظر گرفته می‌شود که برازش آن‌ها ساده‌تر از برآورد توابع تغییرنگار یا هم-تغییرنگار است. یکی از مسائل اساسی، نحوه توری کردن ناحیه مورد مطالعه است، که باید به-گونه‌ای انجام شود تا حتی الامکان گره‌ها فاقد مشاهده یا دارای چند مشاهده نباشند. تعیین توزیع پیشین مناسب در رهیافت بیزی، خود مساله‌ای است که باید مورد توجه قرار گرفته و حساسیت پیش‌گوها با انواع توزیع‌های پیشین مورد بررسی قرار گیرد.

تقدیر و تشکر

نویسندگان از داوران محترم به خاطر پیشنهادات ارزنده‌ای که موجب بهبود مقاله گردید، کمال تشکر را دارند. از حمایت قطب علمی داده‌های ترتیبی و فضایی دانشگاه فردوسی مشهد نیز قدردانی می‌شود.

مراجع

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

[2] Journel, A. G. (1983), Nonparametric Estimation of Spatial Distribution, *Mathematical Geology,* **15**, 445-468.

[3] Basu, S. and Reinsel, G. C. (1994), Regression Models with Spatially Correlated Errors, *Journal of the American Statistical Association***, 89**,88-99.

[4] Reinsel, G. C. and Cheng, W. K. (2003), Approximate ML and REML Estimation for Regression Models with Spatial or Time Series AR (1) Noise*, Statistics and Probability Letters,* **62**, 123-135.

[5] Best, N. G. (1999), Bayesian Models for Spatially Correlated Disease and Exposure Data, *Bayesian Statistics* 6, edited by J. M. Bernardo, B. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford University Press*, Oxford, UK*. 131-156.

[6] Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004), Hierarchical Modeling and Analysis for Spatial Data, *Chapman and Hall,* London.

[7] Cressie, N. (1993), Statistical for Spatial Data, *Wiley, New York.*

[8] Besage, J. E. (1974), Spatial Interaction and Statistical Analysis of Lattice System*, Journal of the Royal Statistical Society, Series B*, **36** , 192-236.

[9] Pascutto, C. (2000), Statistical Issues in the Analysis of Disease Mapping Data, *Statistics in Medicine,* **19**, 2493-2519.

[10] Baily, T. (2001), Spatial Statistical Analysis in Health, *Cadernos de Saude Publica*, **17**, 1083-1098.

[11] Marjerison, W, M. (2006), Bayesian Logistic Regression with Spatial Correlation, *MSc Thesis, Worcester Polytechnic Institute*, Worcester, Massachusetts, USA.

[12] Molenberghs, G. (2005), Models for Discrete Longitudinal Data. *Springer Science,* New York.

[۱۳] محمدزاده، م. و کیاپور، آ. (۱۳۸۷)، پیش‌گویی فضایی داده‌های سخت و نرم برای تهیه نقشه احتمال یخزدگی، مجله امیرکبیر، سال نوزدهم، شماره ه-۶۸، ۴۴-۳۷.

Spatial Logistic and Probit Regression Models for Analysis of Frosting Data in Mazandaran Province

Vahid Rezaeitabar and Mohsen Mohammadzadeh

Department of Statistics, Tarbiat Modares University, Tehran, Iran

Abstract

Logistic and Probit regression models are usually used in binary response variable analysis based on independence assumption of the observations. But, in practice, there are many situations in which, due to their different locations in the underling space of study, this assumption dose not satisfies. In spatial statistics, it is generally supposed that the binary observations are analyzed with indicator Kriging. In this paper, we considered the spatial Logistic and Probit regression models with auto correlated errors on a rectangular grid. Also, in a simulation study, the prediction accuracy of the models has been compared. Finally, the implementation of the models for a temperature data set, reported by weather stations in Mazandaran province of Iran, is shown.

Keywords: Spatial binary data, Logistic and Probit regression, Rectangular grids.

Mathematics Subject Classification (2000): 62J12, 62J02