مجله مدل‌سازی پیشرفته ریاضی
سال ۱۳۹۹، دوره ۱۰، شماره ۱، ص ۳۹-۶۱.

شاپا چاپی: ۸۰۸۸-۲۲۵۱
DOI:.10.22055/JAMM.2020.28698.1692

مدل‌بندی داده‌های فضایی-زمانی با گمشدگی غیرقابل چشم‌پوشی

سمیرا زحمتکش، محسن محمدزاده[۱]

گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۷/۱۲/۳ تاریخ پذیرش: ۱۳۹۸/۹/۹

چکیده: اغلب داده‌های فضایی و فضایی-زمانی به‌واسطه شرایطی که تحت آن اندازه‌گیری‌ها صورت می‌گیرد حاوی مقادیر گمشده هستند. مقادیر گمشده‌ای که در فواصل مکانی یا زمانی نزدیک به مشاهدات قرار دارند شامل اطلاعات مفیدی هستند که در نظر گرفتن آن‌ها می‌تواند منجر به نتایج دقیق‌تری شود. بنابراین لازم است حضور داده‌های گمشده در مجموعه داده‌های فضایی موردتوجه و بررسی دقیق قرار گرفته و در تحلیل و مدل‌بندی داده‌ها لحاظ شود. برای این منظور می‌توان با مدل‌بندی توأم فرایندی که منجر به گمشدگی داده‌ها می‌شود و فرایند اندازه‌گیری فضایی یا فضایی-زمانی، برخی اطلاعات از دست رفته را بازیابی کرده و مورداستفاده قرار داد. در این مقاله، با استفاده از تکنیک مدل پارامتر اشتراکی، به مدل‌بندی توأم فرایند اندازه‌گیری فضایی-زمانی و فرایند گمشدگی در یک چارچوب بیزی می‌پردازیم تا اثرات سوء مقادیر گمشده تعدیل شود. همچنین از طریق یک میدان تصادفی پنهان فضایی-زمانی در مدل، بین دو فرایند ارتباط ایجاد خواهیم کرد. به‌منظور برآورد پارامترهای مدل و پیشگویی‌ها، روش بیز تقریبی INLA به همراه راهکار SPDE به کار بسته شده است. سپس بر اساس مدل توأم، داده‌های دمای سطح آب دریای خزر مدل‌بندی و تحلیل شده و عملکرد مدل نیز مورد ارزیابی قرار گرفته است. در انتها نیز به بحث و نتیجه‌گیری پرداخته شده است.

واژه‌های کلیدی: داده‌های فضایی- زمانی، داده‌های گمشده، INLA، راهکار SPDE.

رده‌بندی ریاضی (۲۰۱۰): 91B72, 91D25, 62F15

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

۱- مقدمه

اکثر داده‌های فضایی و فضایی-زمانی مانند داده‌های بارندگی، دمای هوا یا آلودگی هوا، در محیط‌های غیر آزمایشگاهی و خارج از کنترل به دست می‌آیند. این امر سبب بروز گمشدگی در داده‌ها می‌شود. به‌عنوان مثال در اندازه‌گیری‌های ماهواره‌ای وجود ابر، باران‌های سنگین یا ناصاف بودن آسمان می‌توانند از جمله دلایل ایجاد گمشدگی باشند. همچنین خرابی دستگاه‌ها و ابزارهای اندازه‌گیری نیز می‌تواند منجر به تولید مقادیر گمشده در داده‌ها شود. رابین در سال ۱۹۷۶ فرایندهای مختلف گمشدگی را از هم متمایز نمود و سازوکار گمشدگی را به سه دسته گمشدگی کاملاً تصادفی[1] (MCAR)، گمشدگی تصادفی[2] (MAR) و گمشدگی غیرتصادفی[3] (MNAR) تقسیم‌بندی کرد [۱]. وقتی فرایند گمشدگی مستقل از داده‌های مشاهده‌شده و مشاهده نشده باشد، در این صورت MCAR رخ می‌دهد. MAR فرض ضعیف‌تری است و زمانی رخ می‌دهد که فرایند گمشدگی تنها به مقادیر مشاهده‌شده بستگی داشته باشد. وقتی فرایند گمشدگی در هیچ‌یک از دو دسته MCAR و MAR نباشد، گمشدگی MNAR رخ داده است، در این حالت گمشدگی هم به داده‌های مشاهده‌شده و هم به داده‌های مشاهده نشده بستگی دارد و اصطلاحاً گمشدگی غیرقابل چشم‌پوشی است، به‌طوری‌که لازم است برای دستیابی به نتایج معتبر، فرایند گمشدگی به همراه فرایند اندازه‌گیری مدل‌بندی و در تحلیل داده‌ها از آن استفاده شود.

در داده‌های فضایی-زمانی با توجه به وجود وابستگی بین مشاهدات برحسب موقعیت فضایی و زمان مشاهده شدن هر داده، مقادیر گمشده‌ای که در فواصل فضایی یا زمانی نزدیک‌تر نسبت به مشاهدات قرار دارند می‌توانند شامل اطلاعات مفیدی باشند به‌طوری‌که استفاده از آن‌ها منجر به نتایج دقیق‌تری شود. بنابراین لازم است وجود و نوع گمشدگی در داده‌ها موردتوجه و بررسی دقیق قرار گیرد. مطالعه و بررسی داده‌های گمشده فضایی-زمانی اخیراً موردتوجه محققان متعددی قرار گرفته است. اسمیت و همکاران [۲] روشی از تحلیل داده‌های فضایی-زمانی گمشده ارائه کردند که بر اساس تجزیه توابع مشخصه ناپارامتری به‌صورت ترکیب خطی از متغیرهای تبیینی و مؤلفه‌های تصادفی فضایی است به‌طوری‌که مدل حاصل برای جانهی مقادیر گمشده استفاده می‌شود. کندراشو و قیل [۳] از تحلیل طیفی تکین برای جانهی مقادیر گمشده در انواع مختلف داده‌های فضایی و فضایی-زمانی استفاده کردند. چنگ و لو [۴] روشی مؤثر برای جانهی مقادیر گمشده در داده‌های فضایی-زمانی ارائه کردند که در آن الگوی گمشدگی داده‌ها در نظر گرفته می‌شود و اثر مخرب الگوهایی که شامل گمشدگی مداوم هستند خنثی می‌شود. همچنین

1- Missing Completely At Random
2- Missing At Random
3- Missing Not At Random

بائی و همکاران [۵] و یانگ و همکاران [۶] در بررسی داده‌های ترافیک بر اساس روش کریگیدن تکنیک‌های متفاوتی از جانهی مقادیر گمشده ارائه کردند که در آن‌ها از همبستگی فضایی-موجود در داده‌ها بهره گرفته می‌شود. گربر و همکاران [۷] الگوریتمی منعطف برای جانهی مقادیر گمشده در داده‌های فضایی-زمانی سنجش از راه دور ارائه کردند. در اکثر روش‌هایی که تاکنون معرفی شده‌اند، تحت فرض MAR به مطالعه داده‌های فضایی-زمانی گمشده پرداخته شده است. ازآنجاکه در عمل نمی‌توان مستقیماً از روی داده‌ها به سازوکار گمشدگی پی برد، تحت فرض MNAR می‌توان فرایند گمشدگی را توأماً با فرایند اندازه‌گیری مدل‌بندی کرد و از این طریق برخی از اطلاعات از دست رفته را بازیابی نمود. در بسیاری موارد توزیع توأم دو فرایند فرم مشخصی ندارد و لازم است که از روش‌های تجزیه توزیع توأم استفاده کرد، یکی از این روش‌ها مدل پارامتر اشتراکی[1] (SPM) است [۸]، که در آن فرض می‌شود دو فرایند به‌شرط یک متغیر پنهان تصادفی از هم مستقل هستند به‌طوری‌که می‌توان از طریق این متغیر پنهان بین این دو فرایند ارتباط ایجاد کرد. فالمن و وو [۹]، وونش و همکاران [۱۰] و دنیل و هوگان [۱۱] از SPM در تحلیل داده‌های طولی شامل سانسور آگاهی‌بخش استفاده کردند. تکنیک مشابهی توسط دیگل و منزز [۱۲] برای مدل‌بندی توأم فرایند فضایی و طرح نمونه‌گیری استفاده شد تا از استنباط‌های گمراه کننده که به‌واسطه طرح‌های نمونه‌گیری ترجیحی ممکن است حاصل شود جلوگیری به عمل آید. همچنین پتی و همکاران [۱۳] مدل‌های زمین‌آماری را در نظر گرفتند که در آن‌ها موقعیت‌های فضایی مشاهدات با به‌کارگیری SPM آگاهی‌بخش خواهند بود. در این مدل‌بندی در یک چارچوب بیزی موقعیت‌های فضایی با استفاده از یک فرایند گاوسی کاکس و متغیر پاسخ به‌شرط موقعیت‌های فضایی از طریق یک فرایند گاوسی تصادفی فضایی مدل‌بندی می‌شود. استینسلند و همکاران [۱۴] از مدل‌بندی توأم فرایند اندازه‌گیری و فرایند گمشدگی غیرقابل چشم‌پوشی با استفاده از SPM در بررسی یک مجموعه داده ژنتیکی استفاده کردند، که در آن از مدل‌های خطی آمیخته تعمیم‌یافته برای مدل‌بندی دو فرایند استفاده کردند.

در این مقاله SPM را برای داده‌های فضایی-زمانی گمشده تعمیم خواهیم داد. برای این منظور با در نظر گرفتن یک فرایند پنهان فضایی-زمانی به‌عنوان عامل مشترک در مدل‌بندی فرایند اندازه‌گیری فضایی-زمانی و فرایند گمشدگی بین این دو فرایند ارتباط برقرار خواهیم کرد. همچنین برای تقریب توزیع‌های پسینی حاشیه‌ای پارامترهای مدل و متغیرهای پنهان در یک چارچوب بیزی، از تقریب لاپلاس آشیانه‌ای جمع‌بسته[2] (INLA) استفاده نموده‌ایم [۱۵]. INLA روش تقریبی سریع و با دقت برای استنباط بر اساس پسینی‌های حاشیه‌ای است و جایگزینی برای الگوریتم‌های زنجیر مارکف مونت‌کارلویی برای مدل‌های گاوسی پنهان است. در اینجا از

1- Shared Parameter Model
2- Integrated Nested Laplace Approximation

راهکار متفاوتی که شامل ارائه یک میدان تصادفی گاوسی[1] (GRF) با تابع کوواریانس مترن، به‌صورت یک میدان تصادفی گسسته، یعنی میدان تصادفی گاوسی مارکفی[2] (GMRF) است، استفاده کرده‌ایم [۱۶]. بر اساس لینگرن و همکاران [۱۷] ارتباط بین GRF و GMRF از طریق معادلات دیفرانسیل جزئی تصادفی[3] (SPDE) ارائه می‌شود، در ادامه از این راهکار تحت عنوان راهکار SPDE نام برده می‌شود. در استفاده از GMRF به‌جای تابع کوواریانس فضایی-زمانی و ماتریس کوواریانس از ماتریس دقت استفاده می‌شود که بر اساس ساختار همسایگی به دست می آید و ماتریسی تنک است، به‌این‌ترتیب محاسبات به میزان قابل‌توجهی تسهیل خواهند شد. کملتی و همکاران [۱۸] به مدل‌بندی داده‌های پیوسته فضایی-زمانی مربوط به غلظت یک آلاینده مشخص در منطقه پایمونت ایتالیا پرداختند، به‌طوری‌که از یک مدل اتورگرسیو پویای فضایی-زمانی مرتبه اول برای مدل‌بندی متغیر پاسخ استفاده کردند و با استفاده از INLA و راهکار SPDE به تحلیل داده‌ها پرداختند. در این مقاله داده‌های فضایی-زمانی که شامل مقادیر قابل‌توجهی گمشدگی هستند را موردتوجه قرار داده‌ایم. از مدل اتورگرسیو پویای فضایی-زمانی مرتبه اول برای مدل‌بندی هر یک از فرایند اندازه‌گیری فضایی-زمانی و فرایند گمشدگی استفاده نموده‌ایم. سپس با استفاده از SPM بین این دو فرایند ارتباط ایجاد کرده‌ایم تا اطلاعات از دست رفته به دلیل وجود گمشدگی از این طریق بازیابی شوند.در بخش کاربرد عملکرد مدل توأم پیشنهادی که تحت فرض MNAR بنا شده است را با عملکرد مدلی دیگر تحت فرض MAR مقایسه کرده‌ایم.

ساختار مقاله به این صورت است که در بخش ۲ به نحوه مدل‌بندی داده‌های فضایی-زمانی می پردازیم. در بخش ۳ نحوه به‌کارگیری راهکار SPM برای تشکیل مدل توأم را مطرح خواهیم کرد. در بخش ۴ روش INLA با استفاده از SPDE را معرفی می‌نماییم و در بخش ۵ کاربردی از مدل توأم برای تحلیل داده‌های دمای سطح آب دریای خزر را ارائه خواهیم کرد.

۲- مدل‌بندی داده‌های فضایی-زمانی در دامنه پیوسته فضایی

داده‌هایی که هم از نظر موقعیت فضایی و هم موقعیت زمانی وابسته باشند داده‌های فضایی-زمانی نامیده می‌شوند. به‌منظور تحلیل این‌گونه داده‌ها لازم است که ساختار همبستگی فضایی-زمانی آن‌ها از طریق تابع کوواریانس فضایی-زمانی مدل‌بندی شود. یک میدان تصادفی فضایی-زمانی به‌صورت

1- Gaussian Random Field
2- Gaussian Markove Random Field
3- Stochastic Partial Differential Equations

$$Y(s,t)=\{y(s,t),(s,t)\in D\subset R^{۲}\times R\}$$

تعریف می‌شود. فرض کنید مشاهدات در d موقعیت فضایی و در T نقطه از زمان جمع‌آوری شده‌اند و تحققی از فرایند فضایی-زمانی در موقعیت s_i، و در زمان t، باشد. مدل فضایی-زمانی به‌صورت

$$y(s_i,t)=\mathbf{z}(s_i,t)\boldsymbol{\beta}+\xi(s_i,t)+\varepsilon(s_i,t) \qquad (۱)$$

را در نظر بگیرید، که در آن $\mathbf{z}(s_i,t)=(z_۱(s_i,t),...,z_p(s_i,t))$ بردار مشاهدات p متغیر تبیینی در موقعیت s_i و زمان t، $\boldsymbol{\beta}=(\beta_۱,...,\beta_p)'$ بردار ضرایب، $\varepsilon(s_i,t)$ خطای اندازه‌گیری از توزیع $N(۰,\sigma_\varepsilon^{۲})$ و $\xi(s_i,t)$ تحققی از یک فرایند پنهان فضایی-زمانی است. به‌عنوان مثال در بررسی کیفیت هوا، این فرایند پنهان می‌تواند سطح واقعی مشاهده نشده از آلودگی هوا باشد. فرض می‌شود $\xi(s_i,t)$ در طی زمان بر اساس یک مدل پویای اتورگرسیو مرتبه اول، AR(۱)، به‌صورت

$$\xi(s_i,t)=a\xi(s_i,t-۱)+\omega(s_i,t) \qquad (۲)$$

تغییر کند، که در آن $|a|<۱$ و $\xi(s_i,۱)$ دارای توزیع $N(۰,\frac{\sigma_\omega^{۲}}{۱-a^{۲}})$ و $\omega(s_i,t)$ یک میدان تصادفی گاوسی مانا با میانگین صفر و تابع کوواریانس فضایی-زمانی

$$Cov(\omega(s_i,t),\omega(s_j,t'))=\begin{cases}۰ & t\neq t'\\ \sigma_\omega^{۲}C(h) & t=t'\end{cases} \qquad (۳)$$

است، که در آن تابع همبستگی فضایی $C(h)$ تنها از طریق $h=\|s_i-s_j\|\in R$ به موقعیت‌های فضایی s_i و s_j وابسته است. به‌این‌ترتیب فرض می‌شود که میدان تصادفی $\omega(s,t)$ مانای مرتبه دوم و همسانگرد است [۱۹]. طبق (۳) برای هر s_i و t، $Var(\omega(s_i,t))=\sigma_\omega^{۲}$. تابع همبستگی فضایی $C(h)$ عموماً به دلیل انعطاف‌پذیری، تابع مترن به‌صورت

$$C(h)=\frac{۱}{\Gamma(\nu)۲^{\nu-۱}}(\kappa h)^{\nu}\mathrm{K}_\nu(\kappa h) \qquad (۴)$$

در نظر گرفته می‌شود، که در آن K_ν تابع بسل تعدیل‌یافته نوع دوم از مرتبه $\nu>۰$ است [۲۰]، پارامتر ν که اغلب ثابت در نظر گرفته می‌شود درجه همواری فرایند را نشان می‌دهد، $\kappa>۰$ پارامتر مقیاس مرتبط با دامنه r است، دامنه فاصله‌ای است که در آن همبستگی فضایی کوچک

و نزدیک به صفر می‌شود و رابطه $r=\frac{\sqrt{8\nu}}{\kappa}$ برای آن تعریف می‌شود [۱۷]. اگر $\mathbf{y}_t=(y(s_1,t),...,y(s_d,t))'$ بردار مشاهدات متغیر پاسخ در d موقعیت فضایی و زمان t باشد، مدل‌های (۱) و (۲) را می‌توان به‌صورت

$$\mathbf{y}_t=\mathbf{z}_t\boldsymbol{\beta}+\xi_t+\boldsymbol{\varepsilon}_t \qquad \boldsymbol{\varepsilon}_t\sim N(0,\sigma^2 I_d) \tag{۵}$$

$$\xi_t=a\xi_{t-1}+\omega_t \qquad \omega_t\sim N(0,\Sigma=\sigma_\omega^2\tilde{\Sigma}) \tag{۶}$$

نوشت، که در آن I_d ماتریس همانی از بعد d است، $\mathbf{z}_t=(\mathbf{z}(s_1,t)',...,\mathbf{z}(s_d,t)')'$ و $\xi_t=(\xi(s_1,t),...,\xi(s_d,t))'$ به‌طوری‌که ξ_1 یک فرایند AR(۱) از توزیع مانای $N(0,\frac{1}{1-a^2}\Sigma)$ است. $\tilde{\Sigma}$ ماتریس همبستگی از بعد d با عناصر $C\left(\left\|s_i-s_j\right\|\right)$ است، که در آن $C(\cdot)$ تابع مترن در (۴) است. همچنین فرض می‌شود $\mathbf{y}=(\mathbf{y}_1,...,\mathbf{y}_T)$ بردار مشاهدات متغیر پاسخ و $\xi=(\xi_1,...,\xi_T)$ تحققی از میدان تصادفی پنهان فضایی-زمانی در d موقعیت فضایی و T زمان است. به‌این‌ترتیب $\theta=(\beta,\sigma_\varepsilon^2,a,\sigma_\omega^2,\kappa)$ بردار پارامترهای مدل است که باید برآورد شود. با در نظر گرفتن توزیع پیشینی مستقل $\pi(\theta)=\Pi_{i=1}^{\dim(\theta)}\pi(\theta_i)$ برای پارامترهای مدل و با فرض استقلال شرطی $\mathbf{y}$ روی ξ در زمان و اینکه فرایند پنهان از یک مدل پویای زمانی مارکفی پیروی می‌کند، توزیع پسینی توأم پارامترهای مدل به‌صورت

$$\begin{aligned}\pi(\theta,\xi\mid\mathbf{y})&\propto\pi(\mathbf{y}\mid\xi,\theta)\pi(\xi\mid\theta)\pi(\theta)\\&\propto\Big(\prod_{t=1}^{T}\pi(y_t\mid\xi_t,\theta)\Big)\Big(\pi(\xi_1\mid\theta)\prod_{t=2}^{T}\pi(\xi_t\mid\xi_{t-1},\theta)\Big)\pi(\theta)\end{aligned} \tag{۷}$$

حاصل می‌شود، که با توجه به (۵) و (۶) داریم

$$\begin{aligned}\pi(\theta,\xi\mid\mathbf{y})&\propto(\sigma_\varepsilon^2)^{-\frac{dT}{2}}\exp\Big(-\frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^{T}(\mathbf{y}_t-\mathbf{z}_t\beta-\xi_t)'(\mathbf{y}_t-\mathbf{z}_t\beta-\xi_t)\Big)\\&\times(\frac{\sigma_\omega^2}{1-a^2})^{-\frac{d}{2}}\left|\tilde{\Sigma}\right|^{-\frac{1}{2}}\exp\Big(-\frac{1-a^2}{2\sigma_\omega^2}\xi_1'\tilde{\Sigma}^{-1}\xi_1\Big)\\&\times(\sigma_\omega^2)^{-\frac{d(T-1)}{2}}\left|\tilde{\Sigma}\right|^{-\frac{(T-1)}{2}}\exp\Big(-\frac{1}{2\sigma_\omega^2}\sum_{t=2}^{T}(\xi_t-a\xi_{t-1})'\tilde{\Sigma}^{-1}(\xi_t-a\xi_{t-1})\Big)\times\prod_{i=1}^{\dim(\theta)}\pi(\theta_i)\end{aligned}$$

۳- مدل‌بندی توأم فرایند گمشدگی و فرایند اندازه‌گیری فضایی-زمانی

از آنجا که در مطالعات فضایی-زمانی اغلب اندازه‌گیری‌ها در خارج از محیط آزمایشگاه صورت می گیرد و خارج از کنترل محقق است، این گمان می‌تواند وجود داشته باشد که وجود مقادیر گمشده

ممکن است تحت تأثیر عواملی غیرتصادفی باشد، در این صورت نمی‌توان با اطمینان به تحلیل داده‌ها از طریق روش‌های معمول و تحت فرض MAR پرداخت. عموماً تحت فرض MNAR یعنی با فرض اینکه گمشدگی غیرقابل چشم‌پوشی است، مدلی توأم از فرایند اندازه‌گیری (متغیر پاسخ) و فرایند گمشدگی ساخته می‌شود به‌طوری‌که از طریق آن تحلیل حساسیت نیز در نظر گرفته می‌شود.

فرایند گمشدگی $\mathbf{m}=(\mathbf{m}_1,...,\mathbf{m}_T)$ که در آن $\mathbf{m}_t=(m(s_1,t),...,m(s_d,t))$ از طریق یک تابع نشانگر به‌صورت

$$m(s_i,t)=\begin{cases}1 & y(s_i,t)\ is\ missed\\ 0 & y(s_i,t)\ is\ observed\end{cases} \qquad (۸)$$

تعریف می‌شود و از یک توزیع دودویی پیروی می‌کند. احتمال گمشدگی $y(s_i,t)$ به‌صورت $P(m(s_i,t)=1|\varphi)=\pi(s_i,t)$ است، که در آن φ بردار پارامتر مدل فرایند گمشدگی است. $\pi(s_i,t)$ از طریق یک رگرسیون لوجستیک به‌صورت

$$\text{logit}(\pi(s_i,t))=\alpha_0+\alpha_1\xi(s_i,t)+\mathbf{H}(s_i,t)\gamma \qquad (۹)$$

مدل‌بندی می‌شود، که در آن α_0 عرض از مبدأ و بیانگر نسبت ثابتی از گمشدگی است، $\xi(s_i,t)$ تحققی از فرایند پنهان فضایی-زمانی در موقعیت فضایی s_i و زمان t است که در مدل متغیر پاسخ نیز حضور دارد و درواقع بین $\mathbf{y}$ و $\mathbf{m}$ ارتباط برقرار می‌کند و از طریق ضریب α_1 شدت و چگونگی این ارتباط توصیف می‌شود. این فرایند پنهان فضایی-زمانی مشابه (۲) در طی زمان بر اساس اتورگرسیو پویای مرتبه اول تغییر می‌کند. بردار q بعدی $\mathbf{H}(s_i,t)=(H_1(s_i,t),...,H_q(s_i,t))$ متغیرهای تبیینی در ارتباط با احتمال گمشدگی است و $\gamma=(\gamma_1,\cdots,\gamma_q)'$ بردار ضرایب مربوطه است. اکنون با وارد کردن مدل گمشدگی در استنباط‌ها در رابطه (۷) به‌جای توزیع $\pi(\mathbf{y}|\xi,\theta)$، توزیع توأم $\pi(\mathbf{y},\mathbf{m}|\xi,\theta,\varphi)$ جایگزین خواهد شد.

سه چارچوب متفاوت برای مدل‌بندی داده‌ها تحت فرض MNAR وجود دارد که متناظر با تجزیه‌های مختلفی از توزیع توأم متغیر پاسخ و فرایند گمشدگی تعریف می‌شوند. چنانچه توزیع توأم به‌صورت توزیع شرطی فرایند گمشدگی به‌شرط متغیر پاسخ و توزیع حاشیه‌ای متغیر پاسخ به‌صورت

$$\pi(\mathbf{y},\mathbf{m}|\theta,\varphi)=\pi(\mathbf{m}|\mathbf{y},\varphi)\pi(\mathbf{y}|\theta)$$

تجزیه شده باشد، این راهکار مدل گزینش[1] نامیده می‌شود، که نخستین بار در بحث‌های اقتصادی مطرح شد [۲۱]. در اغلب مطالعات، در به‌کارگیری مدل گزینش یک مدل نرمال چند متغیره برای داده‌های کامل، $\pi(\mathbf{y}\,|\,\boldsymbol{\theta})$ و یک رگرسیون لوجستیک یا پروبیت برای فرایند گمشدگی، $\pi(\mathbf{m}\,|\,\mathbf{y},\boldsymbol{\varphi})$ در نظر گرفته می‌شود و اغلب پارامترهای نامعلوم با روش ماکسیمم درستنمایی برآورد می‌شوند[۲۲]. راهکار دوم مدل الگو آمیخته[2] نام دارد که در آن توزیع توأم به‌صورت

$$\pi(\mathbf{y},\mathbf{m}\,|\,\theta,\varphi)=\pi(\mathbf{y}\,|\,\mathbf{m},\boldsymbol{\theta})\pi(\mathbf{m}\,|\,\boldsymbol{\varphi})$$

است و فرض می‌شود پارامترهای دو مدل فرایند اندازه‌گیری و فرایند گمشدگی از هم متمایزند. به‌طورکلی هر دو مدل گزینش و الگو آمیخته دارای مزایا و معایبی هستند. ازآنجاکه فرض‌های سازوکار گمشدگی غیرقابل آزمون کردن هستند، مدل گزینش در برابر این مفروضات استوار نیست و اغلب منجر به درستنمایی مسطح می‌شوند [۲۳] و در مقابل مدل‌های الگو آمیخته شناسایی ناپذیرند. در هر دو مدل لازم است فرض‌هایی درباره داده‌های گمشده در نظر گرفته شود، ولی در نظر گرفتن این مفروضات برای مدل‌های الگوآمیخته ساده‌تر است زیرا پارامترهای توزیع داده‌های کامل از پارامترهای داده‌های گمشده متمایزند. راهکار سوم مدل پارامتر اشتراکی است که در آن با اضافه کردن ضرایب تصادفی به مدل‌های گزینش و الگو آمیخته، مدل‌های گزینش اثرات تصادفی و الگو آمیخته اثرات تصادفی حاصل می‌شوند [۸]. اگر b مجموعه‌ای از اثرات تصادفی، θ_1 پارامترهای توزیع اثرات تصادفی و θ_2 پارامتر مربوط به مدل پاسخ باشد آنگاه $\pi(\mathbf{y},\mathbf{m},b\,|\,\theta_1,\theta_2,\varphi)$ به‌عنوان یک مدل گزینش به‌صورت

$$\pi(\mathbf{y},\mathbf{m},b\,|\,\theta_1,\varphi,\theta_2)=\pi(\mathbf{m}\,|\,\mathbf{y},b,\varphi)\pi(\mathbf{y}\,|\,b,\theta_1)\,f(b\,|\,\theta_2) \qquad (۱۰)$$

و به‌عنوان یک مدل الگوآمیخته به‌صورت

$$\pi(\mathbf{y},\mathbf{m},b\,|\,\theta_1,\varphi,\theta_2)=\pi(\mathbf{y}\,|\,\mathbf{m},b,\theta_1)\pi(\mathbf{m}\,|\,b,\varphi)\pi(b\,|\,\theta_2) \qquad (۱۱)$$

تجزیه می‌شود. فالمن وو [۹] و وو و کارل [۲۴] به نسخه مهم و معروفی از SPM پرداختند که در آن فرض می‌شود متغیر پاسخ و فرایند گمشدگی به‌شرط متغیرهای پنهان از هم مستقل‌اند. به‌طوری‌که توزیع توأم $\pi(\mathbf{y},\mathbf{m},b\,|\,\theta_1,\varphi,\theta_2)$ به‌صورت

$$\pi(\mathbf{y},\mathbf{m},b\,|\,\theta_1,\varphi,\theta_2)=\pi(\mathbf{y}\,|\,b;\theta_1)\pi(\mathbf{m}\,|\,b;\varphi)\pi(b\,|\,\theta_2) \qquad (۱۲)$$

تجزیه می‌شود. توزیع توأم متغیر پاسخ و فرایند اندازه‌گیری از انتگرال‌گیری نسبت به هر یک از روابط (۱۰)، (۱۱) و (۱۲) به‌صورت

1- Selection Model
2- Pattern Mixture Model

$$\pi(\mathbf{y},\mathbf{m}\mid b,\theta_1,\theta_2,\varphi)=\int\pi(\mathbf{y},\mathbf{m},b\mid\theta_1,\theta_2,\varphi)db \qquad (۱۳)$$

به دست می‌آید. در مدل‌بندی توأم فرایند اندازه‌گیری و فرایند گمشدگی، ξ به‌عنوان متغیر پنهان فضایی زمانی مشترک در نظر گرفته می‌شود و با فرض استقلال شرطی آن دو روی ξ توزیع توأم آن‌ها بر اساس SPM به‌صورت

$$\pi(\mathbf{y},\mathbf{m}\mid\xi,\theta,\varphi)=\pi(\mathbf{y}\mid\xi,\theta)\pi(\mathbf{m}\mid\xi,\varphi) \qquad (۱۴)$$

تجزیه می‌شود و در تحلیل بیزی این مدل توأم جایگزین $\pi(\mathbf{y}\mid\xi,\theta)$ در رابطه (۷) خواهد شد. اگر $\eta=\{\theta,\varphi\}$، آنگاه توزیع پسینی توأم پارامترهای مدل توأم به‌صورت

$$\pi(\eta,\xi\mid\mathbf{y})\propto(\sigma_\varepsilon^2)^{-\frac{dT}{2}}\exp\Big(-\frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^{T}(\mathbf{y}_t-\mathbf{z}_t\beta-\xi_t)'(\mathbf{y}_t-\mathbf{z}_t\beta-\xi_t)\Big)$$
$$\times\prod_{t=1}^{T}\prod_{i=1}^{d}\Big(\frac{e^{\alpha_0+\alpha_1\xi(s_i,t)+\mathbf{H}'(s_i,t)\gamma}}{1+e^{\alpha_0+\alpha_1\xi(s_i,t)+\mathbf{H}'(s_i,t)\gamma}}\Big)^{m(s_i,t)}\Big(\frac{-2e^{\alpha_0+\alpha_1\xi(s_i,t)+\mathbf{H}'(s_i,t)\gamma}}{1+e^{\alpha_0+\alpha_1\xi(s_i,t)+\mathbf{H}'(s_i,t)\gamma}}\Big)^{1-m(s_i,t)}$$
$$\times\Big(\frac{\sigma_\omega^2}{1-a^2}\Big)^{-\frac{d}{2}}\left|\tilde{\Sigma}\right|^{-\frac{1}{2}}\exp\Big(-\frac{1-a^2}{2\sigma_\omega^2}\xi_1\tilde{\Sigma}^{-1}\xi_1\Big)$$
$$\times(\sigma_\omega^2)^{-\frac{d(T-1)}{2}}\left|\tilde{\Sigma}\right|^{-\frac{(T-1)}{2}}\exp\Big(-\frac{1}{2\sigma_\omega^2}\sum_{t=2}^{T}(\xi_t-a\xi_{t-1})'\tilde{\Sigma}^{-1}(\xi_t-a\xi_{t-1})\Big)\times\prod_{i=1}^{\dim(\eta)}\pi(\eta_i)$$

بازنویسی می‌شود که فرم بسته‌ای ندارد. برای تولید نمونه از توزیع‌های پسینی حاشیه‌ای و برآورد پارامترهای مدل و پیشگویی‌ها، به‌طورمعمول از الگوریتم‌های زنجیر مارکف مونت‌کارلویی استفاده می‌شود [۲۵، ۲۶]. در اینجا به‌منظور تسریع محاسبات بیزی از روش بیزی INLA استفاده شده است که در بخش بعدی به شرح آن خواهیم پرداخت.

۴- روش INLA با استفاده از راهکار SPDE

مدل سلسه‌مراتبی که در (۵) و (۶) تعریف شد به رده مدل‌های گاوسی پنهان تعلق دارد که قابل برآورد از طریق INLA است و روشی برای دستیابی سریع‌تر به تقریبی از توزیع‌های حاشیه‌ای پسینی برای متغیرهای پنهان و پارامترها است. در این روش بر اساس رو و همکاران [۱۵] برای میدان تصادفی یک GMRF در نظر گرفته می‌شود. در این بخش چگونگی ارائه یک GRF با تابع کوواریانس مترن به‌صورت GMRF با استفاده از راهکار SPDE را مطرح و روش INLA را معرفی می‌نماییم.

۴-۱- میدان تصادفی گاوسی مارکفی

یک GMRF میدانی تصادفی از یک فرایند فضایی است که وابستگی فضایی داده‌ها را در دامنه‌های فضایی مختلف مدل‌بندی می‌کند [۱۶]، فرض کنید $X=(x_1,...,x_n)$ از توزیع $N(\mu,Q^{-1})$ یک GMRF با بعد n باشد، که در آن Q ماتریس دقت معین مثبت متقارن و معکوس ماتریس کوواریانس است. به‌این‌ترتیب

$$\pi(X)=(2\pi)^{(-\frac{n}{2})}\,|Q|^{\frac{1}{2}}\exp\Big(-\frac{1}{2}(X-\mu)'Q(X-\mu)\Big).$$

لازم به ذکر است که یک GMRF را می‌توان از طریق توزیع‌های شرطی هر مؤلفه به‌شرط سایر مؤلفه‌ها مشخص کرد. از طرفی خاصیت مارکفی وابسته به تعریف یک ساختار همسایگی است به‌طوری‌که توزیع شرطی کامل x_i تنها به تعداد کمی از سایر مؤلفه‌ها وابسته است. اگر مجموعه همسایگی واحد i ام با δ_i نشان داده شود، آنگاه

$$\pi(x_i \mid x_{-i})=\pi(x_i \mid x_{\delta_i}),$$

که در آن x_{-i} تمام عناصر $\mathbf{x}$ غیر از x_i است. بر اساس هلد و رو [۱۶] این استقلال شرطی به‌صورت

$$x_i \perp x_{-\{i,\delta_i\}} \,\big|\, x_{\delta_i}, \qquad i=1,...,n$$

نمایش داده می‌شود. ویژگی استقلال شرطی مستقیماً به ماتریس دقت Q وابسته است و به‌طورکلی برای زوج i و j با فرض $i \neq j$

$$x_i \perp x_j \,\big|\, x_{-\{i,j\}} \Leftrightarrow Q_{ij}=0,$$

یعنی عناصر غیرصفر Q از طریق ساختار همسایگی تعریف می‌شوند به‌طوری‌که $Q_{ij}\neq 0$ اگر $j\in\{i,\delta_i\}$. درواقع مزیت استنباط بر اساس GMRF ویژگی محاسباتی خوب است که ناشی از تنک بودن ماتریس دقت Q است، به‌طوری‌که عملیات جبری را می‌توان از طریق روش‌های عددی برای ماتریس تنک، اجرا کرد. بعلاوه محاسبات براساس GMRF ها در استنباط بیزی از طریق INLA بهبود داده می‌شوند [۱۵].

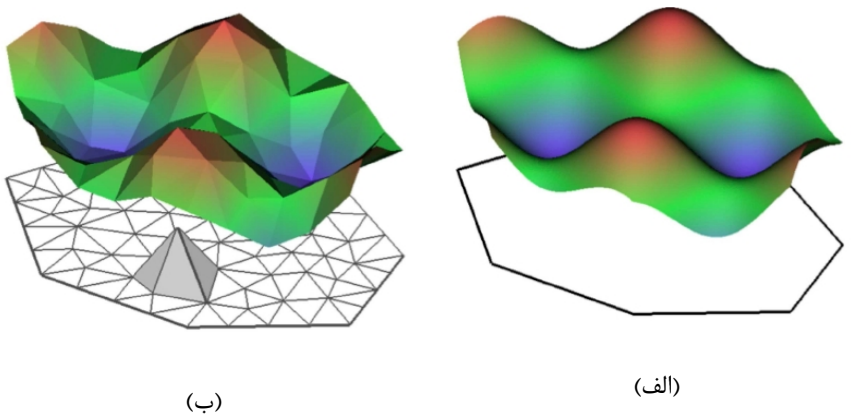
۴-۲-راهکار SPDE

فرض کنید $X(s)=\{x(s), s\in D\subseteq R^2\}$ یک میدان تصادفی گاوسی مانای همسانگرد با تابع کوواریانس مترن (۴) باشد، هدف از راهکار SPDE پیدا کردن GMRF با یک همسایگی و ماتریس دقت تنک Q است که بهترین ارائه از میدان مترن باشد به‌طوری‌که مشکل n بزرگ که

هنگام کار با ماتریس کوواریانس GRF رخ می‌دهد برطرف شود [۲۷] . اساساً راهکار SPDE از یک ارائه عنصر متناهی به‌منظور تعریف میدان مترن به‌عنوان ترکیب خطی از توابع پایه در یک توری مثلثی از دامنه D استفاده می‌کند [۲۸]. این کار شامل تقسیم D به مجموعه‌ای از مثلث‌های مجزا است که حداکثر در یک ضلع یا زاویه مشترک هستند. برای تشکیل توری مثلثی ابتدا هر یک از d موقعیت به‌عنوان یک رأس مثلث در نظر گرفته می‌شوند، سپس سایر رئوس اضافه و مثلث‌ها ساخته می‌شوند. با فرض معلوم بودن توری مثلثی، میدان مترن X(s) به‌صورت ترکیب خطی از توابع پایه به‌صورت

$$X(s)=\sum_{\ell=1}^{n}\psi_{\ell}(s)\omega_{\ell} \qquad (۱۵)$$

تقریب زده می‌شود، که در آن n تعداد رئوس مثلث‌ها در توری، $\psi_{\ell}(s)$ توابع پایه و ω_{ℓ} وزن‌هایی از توزیع گاوسی هستند. توابع $\psi_{\ell}(s)$ به‌گونه‌ای انتخاب می‌شوند که روی هر مثلث خطی باشند، به این صورت که $\psi_{\ell}(s)$ در رأس ℓ برابر ۱ و در سایر رئوس صفر است. به‌این‌ترتیب وزن‌های ω_{ℓ} در هر رأس مقادیر میدان تصادفی را تعیین می‌کنند و مقادیر داخلی هر مثلث از طریق درون‌یابی خطی، محاسبه می‌شود. شکل ۱ یک میدان تصادفی فضایی پیوسته و ارائه عنصر متناهی آن‌که روی یک دامنه مثلثی شده تعریف شده است را نشان می‌دهد.



(ب) (الف)

شکل (۱): الف-میدان تصادفی فضایی $X(s)=\cos(s_1)+\sin(s_2)$ ، ب- ارائه عنصر متناهی میدان تصادفی فضایی X(s) در (۱۵).

آنچه که در راهکار SPDE مورد هدف است ارائه عنصر متناهی (۱۵) است که بین میدان تصادفی گاوسی X(s) و GMRF ارتباط برقرار می‌کند. GMRF از طریق وزن‌های گاوسی ω_{ℓ}

تعریف می‌شود که یک ساختار مارکفی با ماتریس دقت Q به‌صورت آنچه در لینگرن و همکاران [۱۷] آمده است، به آن‌ها تخصیص داده می‌شود که تابعی از پارامترهای تابع کوواریانس میدان تصادفی گاوسی است، به‌طوری‌که برای هرگونه مثلث‌بندی حجم محاسباتی به میزان قابل‌توجهی کاهش می‌یابد.

۴-۳- برازش مدل فضایی-زمانی با استفاده از SPDE

در استفاده از راهکار SPDE در زمان $t=1,...,T$ میدان مترن ω_t در (۴) از طریق یک GMRF به‌صورت $\tilde{\omega}_t \sim N(0, Q_s^{-1})$ ارائه می‌شود، که در آن ماتریس دقت Q_s مطابق زیربخش قبل و با توجه به لینگرن و همکاران [۱۷] به دست می‌آید. ماتریس Q_s با توجه به (۳) مستقل زمانی است و بعد آن برابر با تعداد گره‌های توری مثلثی است. بنابراین معادله (۶) به‌صورت

$$\xi_t = a\xi_{t-1} + \tilde{\omega}_t, \qquad \tilde{\omega}_t \sim N(0, Q_s^{-1}), \qquad t = 1,...,T \tag{۱۶}$$

نوشته می‌شود، که در آن $\xi_1 \sim N(0, \frac{Q_s^{-1}}{1-a^2})$. به‌این‌ترتیب توزیع توأم GMRF ، Tn بعدی $\xi = (\xi_1',...,\xi_T')$ به‌صورت $\xi \sim N(0, Q^{-1})$ است، که در آن $Q = Q_T \otimes Q_S$ و

$$Q_T = \begin{pmatrix} \sigma_\omega^2 & -a/\sigma_\omega^2 & \cdots & 0 & 0 \\ -a/\sigma_\omega^2 & (1+a^2)/\sigma_\omega^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & (1+a^2)/\sigma_\omega^2 & -a/\sigma_\omega^2 \\ 0 & 0 & \cdots & -a/\sigma_\omega^2 & \sigma_\omega^2 \end{pmatrix}$$

ماتریس دقت T بعدی فرایند اتورگرسیو زمانی مرتبه ۱ در (۱۶) است [۱۶]، که با توجه به ویژگی مارکفی فرایند دارای ساختار سه قطری است و سایر عناصر ماتریس برابر صفر است. لازم به ذکر است $\otimes$ نشان‌دهنده ضرب ماتریسی تانسوری است. به‌عنوان مثال برای دو ماتریس A و B با ابعاد به ترتیب $n \times n$ و $m \times m$ و با عناصر a_{ij} و b_{lk}، نمایش ماتریسی عملگر $A \otimes B$ به‌صورت

$$\begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix}$$

است. بعلاوه میدان مترن (۵) به‌صورت

$$\mathbf{y}_t = \mathbf{z}_t\beta + B\xi_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2 I_d) \tag{۱۷}$$

قابل بازنویسی است، که در آن ماتریس $d \times n$ بعدی مقدار ξ_t را برای مشاهده $\mathbf{y}_t$ انتخاب می‌کند و یک ماتریس تنک است به‌طوری‌که

$$y(s_i,t) = \mathbf{z}(s_i,t)\beta + \sum_{j=1}^{n} B_{ij}\xi_t + \varepsilon_t(s_i,t)$$

و $B_{ij}=1$ اگر رأس j ام در مکان s_i باشد و در غیر این صورت $B_{ij}=0$. مدل سلسه‌مراتبی که در (۱۶) و (۱۷) تعریف شد به رده مدل‌های گاوسی پنهان تعلق دارد و از طریق INLA قابل برآورد است. با پیروی از رو و همکاران [۱۵]، فرض کنید $\mathbf{x}=\{\xi,\beta\}$ بردار پارامترهای مدل با پیشین‌های مستقل هستند و β دارای توزیع پیشین گاوسی مبهم با ماتریس دقت معلوم و ξ دارای توزیع GMRF است. بنابراین چگالی $\pi(\mathbf{x}|\theta)$ گاوسی با میانگین صفر و ماتریس دقت $Q(\theta_1)$ با ابرپارامتر $\theta_1=(\sigma_\omega^2, a, \kappa)$ است. اگر $\theta_2=\sigma_\varepsilon^2$ و $\theta=(\theta_1,\theta_2)$ مشاهدات $\mathbf{y}=\{\mathbf{y}_t\}$ دارای توزیع نرمال و مستقل شرطی روی $\mathbf{x}$ و θ هستند. توزیع پسینی توأم به‌صورت

$$\pi(\mathbf{x},\theta|\mathbf{y}) = \pi(\theta)\pi(\mathbf{x}|\theta)\prod_{t=1}^{T}\pi(\mathbf{y}_t|\mathbf{x},\theta) \tag{۱۸}$$

است، که در آن $\pi(\mathbf{y}_t|\mathbf{x},\theta) \sim N(\mathbf{z}_t\beta + B\xi_t, \sigma_\varepsilon^2 I_d)$ توزیع شرطی مشاهدات در زمان t طبق (۱۷) است. در اینجا هدف دستیابی به توزیع‌های پسینی حاشیه‌ای میدان پنهان و ابرپارامترها به‌صورت

$$\pi(x_i|\mathbf{y}) = \int \pi(x_i|\theta,\mathbf{y})\pi(\theta|\mathbf{y})d\theta, \quad i=(1,\ldots,T+p)$$
$$\pi(\theta_j|\mathbf{y}) = \int \pi(\theta|\mathbf{y})d\theta_{-j}, \quad j=1,\ldots,4$$

است. تقریب INLA از روش‌های تقریب گاوسی و لاپلاس و انتگرال‌های عددی برای محاسبه توزیع‌های پسینی حاشیه‌ای استفاده می‌کند [۱۵]. در مدل‌بندی توأم فرایند گمشدگی و فرایند اندازه‌گیری فضایی-زمانی، ازآنجاکه میدان تصادفی پنهان فضایی-زمانی ξ به‌عنوان عامل مشترک در مدل (۹) در نظر گرفته شده است، راهکار SPDE که منجر به مدل‌های سلسله مراتبی (۱۶) و (۱۷) شد را به مدل (۹) تعمیم داده و استنباط بیزی با استفاده از INLA و SPDE را اجرا خواهیم کرد.

۵- تحلیل دمای سطح آب دریای خزر

در مطالعات زیست‌محیطی، تحلیل و بررسی شرایط اکولوژیکی دریاها حائز اهمیت است. به‌عنوان مثال استفاده از اطلاعات مربوط به دمای سطح آب دریاها، مدیریت آب‌های سطح زمین و

همچنین پیشگویی‌های هواشناسی را تسهیل می‌کند، زیرا این کمیت الگوهای جابجایی رطوبت و چرخه‌های هیدرولوژیکی را تعیین می‌کند. داده‌های تحت مطالعه بخشی از مجموعه داده‌های حاصل از پروژه ARC-Lake[1] است، که در آن دامنه جمع‌آوری اطلاعات شامل اکثر دریاچه‌های بزرگ جهان است که هدف آن استفاده از مشاهدات ماهواره‌ای برای دست‌یابی به داده‌ها در سطح دریاچه‌ها است و از وب‌سایت www.laketemp.net قابل‌دسترسی هستند. در اینجا از اطلاعات جمع‌آوری شده مربوط به دریای خزر است، استفاده کرده‌ایم. متغیر تحت مطالعه دمای سطح آب دریاچه[2] (LSWT) است که اندازه‌گیری‌های این متغیر به‌صورت روزانه، ماهیانه و فصلی در دسترس است. برای تشکیل یک مجموعه داده فضایی-زمانی مناسب، اندازه‌گیری‌های ماهیانه مربوط به سه ماه زمستانی شامل دسامبر سال ۲۰۱۰، ژانویه و فوریه ۲۰۱۱ و سه ماه بهاری شامل ماه‌های مارس، آوریل و می سال ۲۰۱۱ را در نظر گرفته‌ایم. مشاهده مربوط به متغیر LSWT برای هر ماه، میانگین اندازه‌گیری‌ها در طی روزهای هر ماه است که در انتهای همان ماه گزارش شده است. مشاهدات روی یک توری ۲۱۱×۱۴۶ در سطح دریا و ساحل اطراف دریا گزارش شده است. پس از جدا کردن مرز دریا از ساحل، موقعیت‌های فضایی شامل d=۱۶۷۰۴ نقطه منظم داخل دریا هستند. مختصات جغرافیایی برای این موقعیت‌های فضایی در سیستم UTM استفاده شده است. متغیر LSWT که میانگین ماهانه دمای سطح آب دریا است به‌عنوان متغیر پاسخ در نظر گرفته شده است. درواقع دلیل استفاده از این مجموعه داده وجود مقادیر قابل‌توجهی از گمشدگی در اندازه‌گیری LSWT است. بعلاوه در این مجموعه داده متغیرهای NCloud و NTemp به ترتیب تعداد پیکسل‌هایی را نشان می‌دهند که در آن‌ها آسمان پوشیده از ابر و صاف است و به‌عنوان متغیرهای کمکی در مدل گمشدگی وارد شده‌اند. تصویری از LSWT و متغیرهای NCloud و NTemp برای ماه‌های مختلف فصل زمستان و فصل بهار به ترتیب در شکل‌های ۲ و ۳ آمده است. در ماه‌های دسامبر و ژانویه و فوریه به ترتیب ۱۴/۲، ۲۳/۱ و ۴۲/۶ درصد گمشدگی و در هر سه ماه حدود ۲۵ درصد گمشدگی وجود دارد. همچنین در ماه‌های مارس و آوریل و می به ترتیب ۴/۴، ۴۳ و ۱۶ درصد گمشدگی و در هر سه ماه حدود ۲۰ درصد گمشدگی وجود دارد. اکنون مدل‌بندی فضایی-زمانی برای زمان $T = ۳$ در هر فصل و d موقعیت فضایی انجام می‌شود. فرض کنید $(y(s_۱,t),...,y(s_۱,t))$ تحقق‌های متغیر پاسخ LSWT در موقعیت‌های فضایی $\{s_۱,...,s_d\}$ و زمان t برحسب واحد سلسیوس باشد، که برای آن مدل

$$y(s_i,t) = \beta_۰ + \xi(s_i,t) + \varepsilon(s_i,t) \qquad (۱۹)$$

1- ATSR Reproccessing for Climate-Lake
2- Lake Surface Water Temperature

در نظر گرفته شده است. همچنین برای فرایند گمشدگی مدل (۹) به‌صورت

$$\text{logit}(\pi(s_i,t)) = \alpha_0 + \alpha_1\xi(s_i,t) + \text{NCloud}(s_i,t)\gamma_1 + \text{NTemp}(s_i,t)\gamma_2 \qquad (۲۰)$$

است و برای $\xi(s_i,t)$ مدل اتورگرسیو پویای فضایی-زمانی مرتبه اول طبق (۲) در نظر گرفته شده است. برای برآورد پارامترهای مدل و محاسبه توزیع‌های پسینی به روش INLA و راهکار SPDE از بسته R-INLA در نرم‌افزار R استفاده شده است.

توزیع پیشینی پارامترهای مدل همانند پیشینی‌های پیش‌فرض در بسته R-INLA به‌صورت

$$\beta_0 \sim N(0,\infty), \quad \alpha_0,\alpha_1 \sim N(0,0/001^{-1}), \quad \log(\sigma_\varepsilon^{-2}) \sim \text{logGamma}(1/5\times10^{-5}).$$

در نظر گرفته شده است. برای پارامترهای κ و τ نیز از پیشینی‌های پیچیدگی تاوانیده[1] استفاده شده است که اولین بار توسط سیمپسون و همکاران [۲۹] معرفی گردیده است. سپس فاگستاد و همکاران [۳۰] از این ایده استفاده نموده و توزیع پیشینی توأم برای پارامتر دامنه و انحراف استاندار حاشیه‌ای در تابع کوواریانس مترن ارائه کردند. برای این منظور GRF با تابع کوواریانس مترن به‌عنوان تابعی از r و σ^2 بازپارامتری می‌شود و دو احتمال پیشینی به‌صورت

$$P(r < r_0) = p_r, \qquad P(\sigma > \sigma_0) = p_\sigma \qquad (۲۱)$$

در نظر گرفته می‌شوند، که در آن‌ها r_0 و σ_0 مقادیر دمی و p_r و p_σ احتمال‌هایی هستند که توسط کاربر انتخاب می‌شوند. در اینجا r_0 و σ_0 به ترتیب برابر با ۵۰ و ۱۰ و p_r و p_σ برابر با ۰/۱ قرار داده شده‌اند. درواقع با در نظر گرفتن این اطلاع پیشینی، بیان نموده‌ایم که بعید است محدوده دامنه کمتر از ۵۰ باشد و انحراف استاندارد حاشیه از ۱۰ فراتر رود. قابل‌ذکر است برای ضرایب رگرسیونی γ_1 و γ_2 به‌طور پیش‌فرض، توزیع‌های پیشینی گاوسی مبهم با دقت معلوم در نظر گرفته می‌شود. شکل ۴ توری مثلثی ساخته شده در راهکار SPDE را نشان می‌دهد.

در طی بررسی‌های اولیه مدل (۱۹) و (۲۰) و مشاهده معنی‌دار نبودن پارامتر α_0، مدل بدون عرض از مبدأ به فرایند گمشدگی برازش داده شده است و نتایج برآورد سایر پارامترها برای دو فصل زمستان و بهار در جدول ۱ گزارش شده است. ضریب α_1 در دو فصل زمستان و بهار به ترتیب ۰/۵۱۰ و ۰/۲۰۴ برآورد شده است که مقادیری مثبت هستند و نشانه وجود ارتباط مستقیم بین احتمال گمشدگی و دمای سطح آب دریا است. ولی این همبستگی در فصل زمستان شدت بیشتری دارد و این‌گونه می‌توان نتیجه‌گیری کرد که تقریباً در این فصل هرچه LSWT بزرگ‌تر باشد احتمال گمشدگی بیشتر است و همچنین این گمشدگی تحت تأثیر عوامل پنهان فضایی و

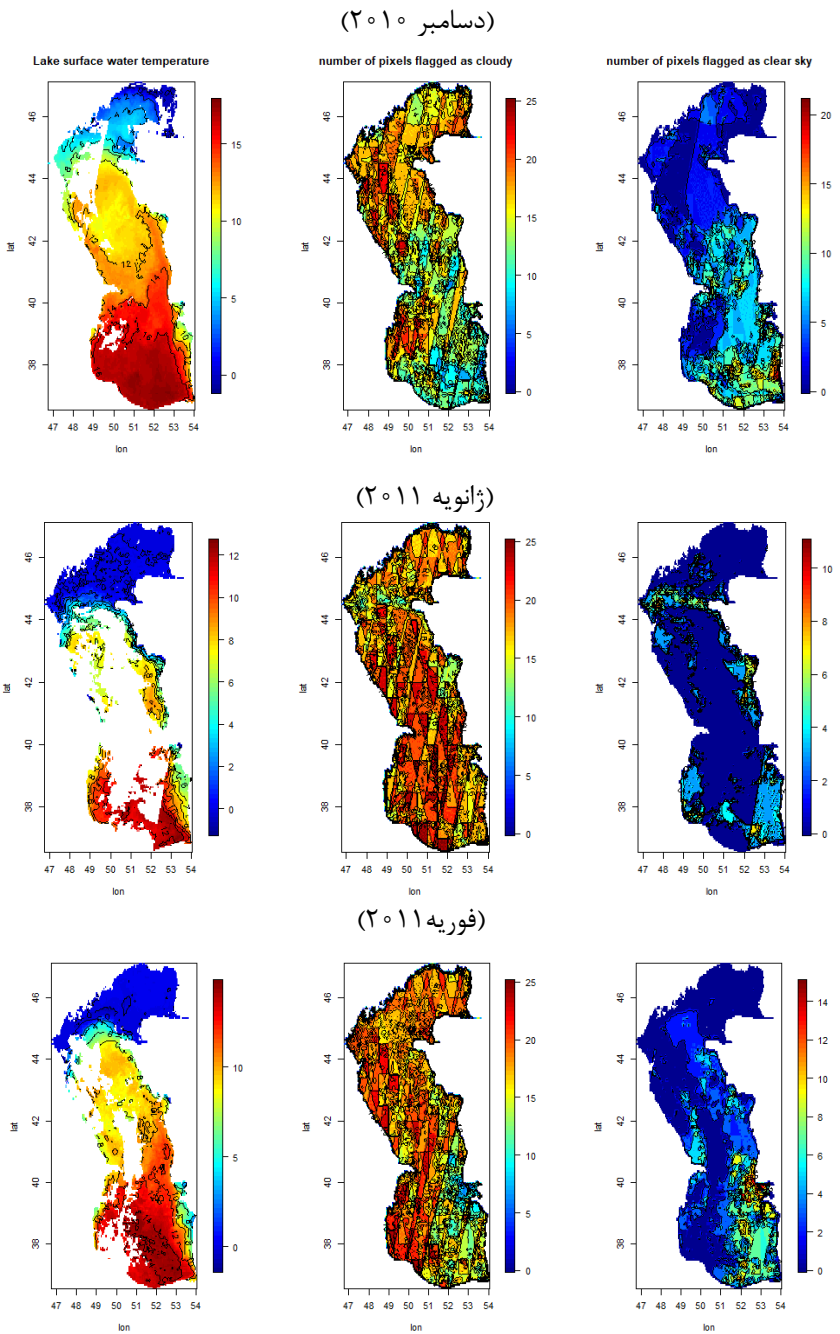
1- Penalised Complexity

فضایی-زمانی است، قابل‌ذکر است چنانچه $\alpha_1 = 0$ برآورد شود نسبت گمشدگی ثابت و بیانگر وجود MAR در داده‌ها است. پارامتر دامنه، r در هر دو فصل مقداری نسبتاً بزرگ برآورد شده است و ازآنجاکه دامنه فاصله‌ای است که در آن همبستگی نزدیک به صفر می‌شود، این مقادیر نشان می‌دهند که همبستگی فضایی قوی بین مشاهدات وجود دارد که با افزایش فاصله به‌کندی کاهش می‌یابد. مقادیر میانگین پسینی برای σ_ω و σ_ε^{-2} نشان می‌دهد که تغییرپذیری بیشتری توسط جمله فضایی در مقایسه با جمله خطای اندازه‌گیری ε توضیح داده می‌شود. میزان a برای دو فصل زمستان و بهار به ترتیب ۰/۴۶۴ و ۰/۲۰۰ برآورد شده است که نشان می‌دهد LSWT در طی زمان در حال تغییر است و این تغییرات در طول زمان، در فصل زمستان بیشتر بوده است. در فصل زمستان مقدار γ_1 برابر با ۰/۰۵۵- برآورد شده است که نشان می‌دهد متغیر تعداد پیکسل‌های ابری با احتمال گمشدگی رابطه معکوس دارد، اگرچه این ضریب مقداری کوچک برآورد شده است و اثر آن کم است. برآورد γ_2 مقدار ۸/۷۵۸- است که علامت منفی و مقدار بزرگ آن نشان دهنده وجود ارتباط معکوس بین متغیر تعداد پیکسل‌های آسمان صاف و احتمال گمشدگی است که با توجه به شکل ۲ این نتیجه منطقی است.

همچنین در فصل بهار مقدار γ_1 برابر ۰/۰۲۴ برآورد شده است که اگرچه مقداری کوچک است ولی نشان دهنده ارتباط مستقیم بین تعداد پیکسل‌های ابری و احتمال گمشدگی است، مقدار γ_2 برابر با ۱۲/۶۷۱- برآورد شده است که مقداری بزرگ با علامت منفی است و نشان دهنده تأثیر معکوس و قوی متغیر تعداد پیکسل‌های آسمان صاف و احتمال گمشدگی است، به این معنی که هرچه از وضوح آسمان کاسته شود احتمال گمشدگی در متغیر پاسخ افزایش می‌یابد.

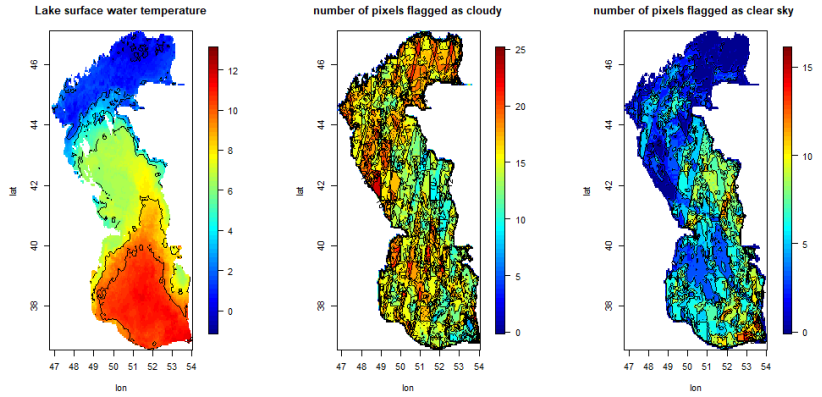
در مطالعات فضایی-زمانی استفاده از مدل مناسب برای پیشگویی در موقعیت‌های فضایی و زمانی جدید از اهمیت بسزایی برخوردار است. با هدف بررسی عملکرد مدل توأم در پیشگویی، مدل دیگری تحت عنوان مدل MAR در نظر گرفته شده است. با در نظر گرفتن این مدل تحت فرض MAR، ابتدا مقادیر گمشده با مقداری مناسب جایگذاری می‌شوند و پس از تشکیل یک مجموعه داده کامل مدل (۱۹) بدون در نظر گرفتن مدل فرایند گمشدگی به داده‌ها برازش داده می‌شود. برای ارزیابی عملکرد پیشگویی این دو مدل از روش اعتبارسنجی متقابل استفاده شده است، که در آن عناصر بردار $\mathbf{y}_{obs}$ به‌صورت تک‌به‌تک از داده‌ها کنار گذاشته شده و پیشگویی در موقعیت مکانی و زمانی مشاهده حذف شده بر اساس سایر داده‌ها انجام گرفته است. سپس ملاک جذر میانگین توان‌های دوم خطای پیشگویی اعتبارسنجی متقابل به‌صورت

$$\mathrm{CVRMSP} = \left[\frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} ((y_{obs})_i - \hat{y}_{-i})^2\right]^{\frac{1}{2}}$$

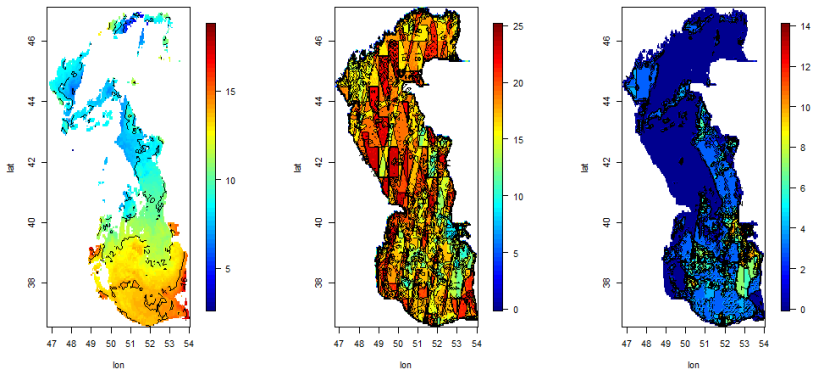


شکل (۲): پهنه‌بندی دمای سطح آب دریای خزر و تعداد پیکسل‌های ابری و آسمان صاف در زمستان

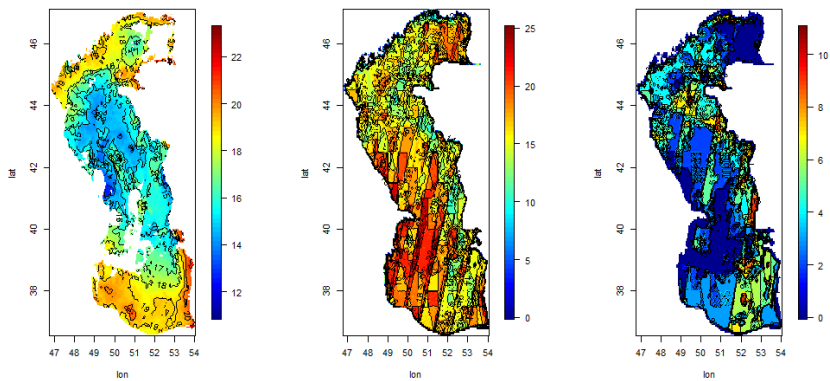
(مارس ۲۰۱۱)



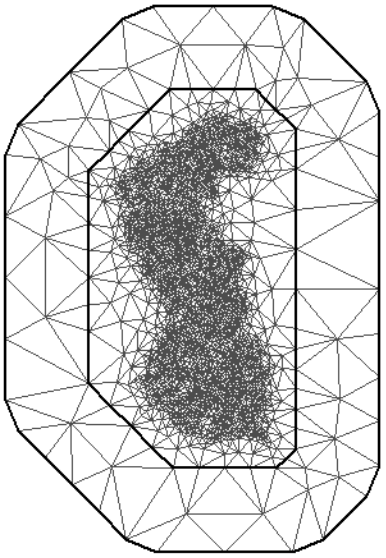
(آوریل ۲۰۱۱)



(می ۲۰۱۱)



شکل (۳): پهنه‌بندی دمای سطح آب دریای خزر و تعداد پیکسل‌های ابری و آسمان صاف در بهار



شکل (۴): توری مثلثی ساخته‌شده در محدوده دریای خزر در استفاده از راهکار SPDE

محاسبه شده است، که در آن $(y_{obs})_i$ عنصر i ام بردار $\mathbf{y}_{obs}$ و $\hat{y}_{-i}$ مقدار پیشگویی آن بر اساس تمام داده‌ها غیر از $(y_{obs})_i$ است. کوچک‌تر بودن این کمیت نشان‌دهنده عملکرد بهتر مدل در پیشگویی است. مقدار این کمیت در فصل زمستان و بهار برای دو مدل توأم و مدل MAR در جدول ۲ ارائه شده است.

در فصل زمستان مقدار CVRMSP برای مدل توأم برابر با ۰/۷۱۰ و برای مدل MAR برابر با ۱/۳۱۴ است. همچنین مقدار این کمیت در فصل بهار برای مدل توأم برابر با ۰/۵۶۱ و برای مدل MAR برابر با ۰/۷۶۸ است. همان‌طور که ملاحظه می‌شود در فصل بهار و زمستان مقدار CVRMSP برای مدل توأم در مقایسه با مدل MAR کوچک‌تر است، ولی در فصل بهار اختلاف این کمیت برای دو مدل حدود ۰/۲ است. با توجه به جدول ۱ این نتیجه برای فصل زمستان و بهار با مقدار برآورد شده برای پارامتر α_1 که به ترتیب برابر با ۰/۵۱۰ و ۰/۲۰۴ است دور از انتظار نیست، زیرا با کوچک‌تر شدن مقدار این پارامتر و نزدیک شدن آن به صفر فرض MAR در داده‌ها تحقق پیدا خواهد کرد، بنابراین در فصل بهار مدل MAR باوجود سادگی و در نظر نگرفتن مدل گمشدگی نتیجه‌ای نزدیک به مدل توأم به همراه دارد. به‌طورکلی نتایج پیشگویی و مقادیر برآورد شده برای دو پارامتر مهم α_1 و σ_ω حاکی از آن هستند که فرایند اندازه‌گیری و فرایند گمشدگی در ارتباط با عاملی پنهان تحت عنوان فرایند تصادفی پنهان فضایی-زمانی با پراکندگی نسبتاً بالایی قرار دارند که یکی از دلایل ایجاد گمشدگی غیرتصادفی در اندازه‌گیری‌ها است و

به‌این‌ترتیب در نظر گرفتن این عامل به‌صورت هم‌زمان در مدل فرایند اندازه‌گیری و فرایند گمشدگی منجر به نتایج بهتر و قابل‌اعتمادتری خواهد شد.

جدول (۱): برآوردهای پسینی پارامترهای مدل توأم فرایند پاسخ و فرایند گمشدگی و انحراف از استاندارد و چهارک‌ها

فصل	پارامتر	Mean	Std	۰/۰۲۵	۰/۹۷۵
	β_0	۲/۶۴۳	۰/۴۰۷	۱/۸۵۴	۳/۴۵۲
	σ_ε^{-2}	۲۴/۵۱۵	۱/۲۸۷	۲۲/۱۳۴	۲۷/۲۰۰
	σ_ω	۴/۲۵۶	۰/۵۳۴	۳/۴۱۸	۵/۴۹۲
زمستان	r	۴۴۳/۱۴۱	۵۷/۸۶۲	۳۵۲/۶۸۸	۵۷۷/۲۵۹
	a	۰/۴۶۴	۰/۰۲۲	۰/۴۱۹	۰/۵۰۸
	α_1	۰/۵۱۰	۰/۰۱۶	۰/۴۷۸	۰/۵۴۲
	γ_1	-۰/۰۵۵	۰/۰۱۲	-۰/۰۷۹	-۰/۰۳۲
	γ_2	-۸/۷۵۸	۱/۰۰۵	-۱۰/۹۹۸	-۷/۰۶۵
	β_0	۴/۵۶۹	۲/۳۹۳	۲/۶۸۳	۶/۴۱۹
	σ_ε^{-2}	۴۱/۰۸۸	۲/۷۶۸	۳۶/۴۷۲	۴۷/۲۵۸
	σ_ω	۷/۲۵۳	۰/۱۴۰	۶/۹۹۸	۷/۵۴۸
بهار	r	۸۷۱/۰۹۸	۱۱/۹۶۰	۸۴۶/۴۸۲	۸۹۵/۲۲۳
	a	۰/۲۰۰	۰/۰۰۶	۰/۱۸۷	۰/۲۱۳
	α_1	۰/۲۰۴	۰/۰۱۳	۰/۱۷۷	۰/۲۳۰
	γ_1	۰/۰۲۴	۰/۰۱۱	۰/۰۰۲	۰/۰۴۵
	γ_2	-۱۲/۶۷۱	۸/۵۴۵	-۳۲/۷۵۸	-۰/۲۸۴

جدول (۲): ملاک CVRMSP برای دو مدل

	مدل	
فصل	توأم	MAR
زمستان	۰/۷۱۰	۱/۳۱۴
بهار	۰/۵۶۱	۰/۷۶۸

بحث و نتیجه‌گیری

در این مقاله، تحت فرض MNAR به مدل‌بندی داده‌های فضایی-زمانی گمشدگی پرداخته شد. با توجه به اینکه در برخی مطالعات فضایی-زمانی گمشدگی می‌تواند متأثر از عوامل پنهان فضایی-زمانی باشد، نمی‌توان به‌سادگی تحت فرض MAR به نتایج قابل‌اعتمادی دست پیدا کرد. در چنین مواردی می‌توان با به‌کارگیری تکنیک‌هایی نظیر مدل پارامتر اشتراکی، فرایند گمشدگی و فرایند اندازه‌گیری را توأماً مدل‌بندی کرد و ارتباط عوامل پنهان فضایی-زمانی و فرایند گمشدگی را کشف و بررسی کرد و به نتایج قابل قبولی نیز دست یافت. بررسی‌های بیشتر در زمینه عملکرد مدل توأم تحت شرایط مختلف از طریق مطالعات شبیه‌سازی در برنامه‌های مطالعاتی آینده قرار گرفته است.

تقدیر و تشکر

نویسندگان از داوران محترم برای پیشنهادهای ارزنده‌ای که موجب ارتقا و ارائه بهتر مقاله شد و از هیئت تحریریه و سردبیر محترم مجله کمال تشکر را دارند. از حمایت قطب علمی تحلیل داده‌های وابسته فضایی و فضایی-زمانی دانشگاه تربیت مدرس نیز قدردانی می‌شود.

منابع

[1] Rubin, D. B. (1976). Inference and missing data, *Biometrika*, **63**, 581-592.

[2] Smith, R. L., Kolenikov, S. and Cox, L. H. (2003). Spatio-temporal modeling of PM2.5 data with missing values, *Journal of Geophysical Research: Atmospheres*, 108(D24).

[3] Kondrashov, D. and Ghil, M. (2006). Spatio-temporal filling of missing points in geophysical data sets, *Nonlinear Processes in Geophysics*, **13**, 151-159.

[4] Cheng, S. and Lu, F. (2017). A two-step method for missing spatio-temporal data reconstruction, *ISPRS International Geo-Information*, **6**, 187.

[5] Bae, B., Kim, H., Lim, H., Liu, Y., Han, L. D. and Freeze, P. B. (2018).Missing data imputation for traffic flow speed using spatio-temporal cokriging, *Emerging Technologies*, **88**, 124-139.

[6] Yang, H., Yang, J., Han, L. D., Liu, X., Pu, L., Chin, S. M. and Hwang, H. L. (2018). A Kriging based spatio-temporal approach for traffic volume data imputation, *PloS one*, **13**, e0195957.

[7] Gerber, F., de Jong, R., Schaepman, M., Schaepman-Strub, G. and Furrer, R. (2018). Predicting missing values in spatio-temporal remote sensing

data, *IEEE Transactions on Geoscience and Remote Sensing*, **56**, 2841-2853.

[8] Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated measures studies, *Journal of the American Statistical association*, **90**,1112-1121.

[9] Follmann, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data, *Biometrics*, 151-168.

[10] Vonesh, E., Greene, T. and Schluchter, M. (2006), Shared parameter models for joint analysis of longitudinal data and event times, *Statistics in Medicine,* **25**, 143-163.

[11] Daniels, M. J. and Hogan, J. W. (2008). *Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis*, Chapman and Hall.

[12] Diggle, P., Menezes, R. and SU, T. (2010). Geostatistical inference under preferential sampling (with discussion). *Applied Statistics*, **59**, 191-232.

[13] Pati, D., Reich, B. J. and Dunson, D. B. (2011). Bayesian geostatistical with informative sampling locations. *Biometrika*, **98**, 35-48.

[14] Steinsland, I., Thorrud Larsen, C., Roulin, A. and Jensen, H. (2014). Quantitative genetic modelling and inference in the presence non-ignorable missing data, *International Journal of Organic Evolution,* **68**, 1735-1747.

[15] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of Statistical Planning and Inference*, **71**, 319-392.

[16] Rue, H. and Held, L. (2005). *Gaussian Markove Random Fields:Theory and Applications*, Chapman and Hall.

[17] Lindgren, F., Rue, H. and Lindstr, O. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, *Journal of the Royal Statistical Society*, **73**, 423-498.

[18] Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, **97**, 109-131.

[19] Cressie, N. (1993). *Statistics for Spatial Data*, Revised Edition, John Wiley, New York.

[20] Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*, Courier Dover Publications.

[21] Vella, F. (1998). Estimating models with sample selection bias: a survey, *Journal of Human Resources*, 127-169.

[22] Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion), *Journal of the Royal Statistical Society*, **43**, 49-93.

[23] Molenberghs, G., Michiels, B., Kenward, M. G. and Diggle, P. J. (1998). Monotone missing data and patternmixture models, *Statistica Neerlandica*, **52**, 153-161.

[24] Wu, M. C. and Carroll R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process, *Biometrics*, 175–188.

[25] Sahu, S., (2011). Hierarchical Bayesian models for space-time air pollution data, *Handbook of Statistics*, **30**, 477-495.

[26] Cameletti, M., Ignaccolo, R. and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in piemonte, *Environmetrics*, **22**, 85-996.

[27] Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H., (2008).Gaussian predictive process models for large spatial datasets, *Journal of the Royal Statistical Society*, **70**, 825-848.

[28] Brenner, S. C. and Scott, R. (2007). *The Mathematical Theory of Finite Element Methods*, New York: Springer.

[29] Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: a principled, practical approach to constructing priors, *Statistical Science*, **32**, 1-28.

[30] Fuglstad, G., Simpson, D., Lingren, F. and Rue, H., (2019). constructing priors that penalize the complexity of Gaussian random fields, *Journal of American Statistical Association*, **114**, 445-452.

JAMM (2020) 10(1): 39-61
DOI:.10.22055/JAMM.2020.28698.1692.

Shahid Chamran University of Ahvaz

ISSN: 2251-8088
Jamm.Scu.ac.ir

Modeling of Spatio-Temporal Data with Non-Ignorable Missing

Samira Zahmatkesh, Mohsen Mohammadzadeh

Department of Statistics, Tarbiyat Modares university, Tehran, Iran

Received: February 22 2019 Accepted for publication: November 30 2019
Corresponding author: mohsen_m@modares.ac.ir

Abstract: Often, due to conditions under which measurements are made, spatio-temporal data contain missing values. Missing data in spatial or temporal vicinity may include useful information. Using this information, we can provide more accurate results, so missing data should be carefully examined. By modeling the missing process and spatio-temporal measurement process jointly, some lost information could be recovered. In this paper, we implement joint modeling in a Bayesian framework using the "shared parameter model" technique, so that the bad effects of missing values will be moderated. Also, we will associate these two processes via a latent spatio-temporal random field. To estimate the model parameters and for predictions, the Bayesian method INLA using SPDE approach is applied. Also, the lake surface water temperature data for Caspian Sea is used to evaluate the performance of the joint model.

Keywords: Spatio-temporal data, Missing data, INLA, SPDE approach.

Mathematics Subject Classification (2010): 91B72, 91D25, 62F15.