مدل‌بندی و تحلیل داده‌های فضایی گم‌شده با استفاده از مدل‌های اتورگرسیو فضایی و روش ماکسیمم درست‌نمایی برداری‌شده

سمیرا زحمتکش، محسن محمدزاده*
گروه آمار، دانشگاه تربیت مدرس

دریافت: ۹۹/۰۵/۱۱ پذیرش: ۹۹/۰۷/۰۲

چکیده

گاهی مجموعه داده‌ها، به عنوان تحقق‌های یک میدان تصادف فضایی، شامل مقادیر گم‌شده هستند. مقادیر مشاهده شده وگم‌شدهٔ فضایی که در همسایگی یکدیگر قرار دارند می‌توانند حاوی اطلاعات مفیدی باشند. بازیابی این اطلاعات از دست رفته به روشی مناسب موجب دستیابی به نتایج معتبر و دقیق‌تری خواهد شد. به منظور مدل‌بندی داده‌های فضایی گم‌شده می‌توان ازمدل‌های اتورگرسیو استفاده کرد. آن‌چه که در استفاده از این مدل‌ها حائز اهمیت است استفاده از روش مناسب برای دستیابی به برآوردی از پارامترهای مدل و در نتیجه پیشگویی در موقعیت‌های فاقد مشاهده است. بررسی‌ها نشان داده است که در استفاده از این مدل‌ها، برآورد ماکسیمم درستنمایی پارامترهای مدل منجر به محاسبات زمان‌بر و ماکسیمم‌های موضعی می‌شود. در این مقاله روش جایگزین "ماکسیمم درستنمایی برداری‌شده" معرفی و نحوهٔ تحلیل مدل‌ها در حضور مقادیر گم‌شده مورد بررسی قرار می‌گیرد. همچنین مطالعات شبیه‌سازی و مثال کاربردی برای ارزیابی عملکرد روش تحت مطالعه ارائه خواهد شد.

واژه‌های کلیدی: داده‌های فضایی گم‌شده، مدل‌های اتورگرسیو فضایی، ماکسیمم درستنمایی برداری‌شده، ماتریس وزن.

رده‌بندی ریاضی (۱۰۲۰): 91B72،91D25 ،62F15.

۱. مقدمه

یکی از روش‌های مدل‌بندی داده‌های فضایی استفاده از مدل‌های اتورگرسیو فضایی است، که در آن از طریق یک ماتریس وزن فضایی مناسب، وابستگی فضایی مشاهدات در نظر گرفته می شود. در این مدل‌ها شدت وابستگی بین مشاهدات از طریق ضریب اتورگرسیو فضایی توصیف می‌شود. اولین بار [۲۰] با فرض این که متغیرهای پاسخ و تبیینی روی یک مستطیل توری مشاهده شده‌اند و با فرض اینکه خطاهای اندازه‌گیری در هر موقعیت تنها وابسته به موقعیت‌های قبل و بعد و بالا و پایین موجود درهمسایگی مشاهدات هستند، وابستگی فضایی را از طریق یک مدل اتورگرسیو مدل‌بندی کرد. سپس افرادی مانند [۲، ۴ و ۵] به تحلیل داده‌های فضایی با استفاده از مدل‌های اتورگرسیو پرداختند. [۲] نشان داد برآورد مدل‌های اتورگرسیو به روش کمترین توان‌های دوم خطا اریب و ناسازگار هستند. برآوردیابی پارامترهای این نوع از مدل‌های فضایی عموماً شامل عملیات محاسبات پیچیده مانند استخراج مقادیر ویژه، معکوس کردن و محاسبهٔ دترمینان ماتریس کواریانس است که حجم این محاسبات به میزان قابل توجه با زیاد شدن مشاهدات افزایش می‌یابد. در

*نویسندهٔ مسئول مقاله mohsen_m@modares.ac.ir

روش ماکسیمم درستنمایی نیاز به روش‌های بهینه‌سازی غیرخطی با استفاده از مشتقات و سایر محاسبات تقریبی است که خود یک مسأله محسوب می‌شود. متأسفانه چنین مسائلی سبب می‌شوند که کاربر به جای یک مقدار بهینه به مقادیر بهینه موضعی دست پیدا کند و حتی از خطاهایی که در حین بهینه‌سازی رخ می‌دهد بی‌خبر باشد. [۱۸] مثالی را مطرح کرد که در آن چندین مقدار بهینه برای یک پارامتر در یک مسأله، درستنمایی حاصل می‌شد. در واقع یک برآوردگر ایده‌آل باید قابلیت به کار گرفتن داده‌هایی با حجم بالا و استنباط سریع داشته باشد و نباید متکی به الگوریتم‌های بهینه‌سازی غیرخطی باشد که تنها یک مقدار موضعی و نه فراگیر را به عنوان مقدار بهینه در اختیار می‌گذارد. [۹] از درستنمایی نیم‌رخ[۱] برای برآورد پارامترها استفاده کرد که در آن به منظور کاهش تعداد پارمترها ابتدا تعدادی از پارامترها بر حسب یکی از پارامترها در تابع درستنمایی جایگذاری می‌شوند و سپس تابع درستنمایی نیم‌رخ نسبت به آن پارامتر ماکسیمم می‌شود و پس از برآورد آن و جایگذاری در تابع درستنمایی سایر پارامترها برآورد می‌شوند. [۸] سازگاری برآوردهای حاصل از ماکسیمم کردن تابع درستنمایی نیم‌رخ را به اثبات رساند. [۱۵] روشی برای محاسبهٔ سریع برآوردها وقتی متغیر وابسته از یک فرایند اتورگرسیو فضایی تبعیت می‌کند ارائه دادند که به "ماکسیمم درستنمایی برداری‌شده[۲]" شهرت دارد. از دیدگاه محاسباتی برداری کردن مسأله از افزایش هزینه‌های استفاده از بهینه‌گر غیرخطی، که به طور معمول همراه تکرار هستند، جلوگیری می‌کند.

وجود گمشدگی در داده‌های فضایی اغلب امری طبیعی است. واضح است که با بازسازی داده‌ها و جانهی مقادیر گم‌شده می‌توان به نتایج و استنباط‌های دقیق‌تری دست یافت. در داده‌های فضایی، شدت وابستگی بین مشاهداتی که در همسایگی یکدیگر قرار دارند قوی‌تر از وابستگی مشاهدات دور است. این ویژگی بر مدل‌بندی و استنباط‌های آماری داده‌ها تأثیرگذار است. بنابراین مقادیر گم‌شده‌ای که در فواصل مکانی نزدیک به هم یا مشاهدات قرار دارند شامل اطلاعات مفیدی هستند که به کار گرفتن آن‌ها می‌تواند منجر به نتایج دقیق‌تری در تحلیل داده‌ها شود. در نظر گرفتن فرضی مناسب در مورد سازوکار گم‌شدگی در داده‌ها حایز اهمیت است. [۱۹] فرایندهای مختلف بروز گم‌شدگی در داده‌ها را از هم متمایز کرد و سازوکار گم‌شدگی را به سه دسته گم‌شدگی کاملاً تصادفی[۳] (MCAR)، گم‌شدگی تصادفی[4] (MAR) و گم‌شدگی غیرتصادفی[۵] (MNAR) تقسیم‌بندی کرد. وقتی فرایند گمشدگی مستقل از داده‌های مشاهده شده و مشاهده نشده باشد، در این صورت MCAR رخ می‌دهد، MAR فرض ضعیف‌تری است و زمانی رخ می‌دهد که فرایند گم‌شدگی تنها به مقادیر مشاهده شده بستگی داشته باشد. به طور کلی این دو حالت از گم‌شدگی را قابل چشم‌پوشی می‌نامند، به این معنی که تحت فرض MCAR و MAR استفاده از مدل‌های رایج مربوط به داده‌های کامل و همچنین روش‌های مبتنی بر درستنمایی بدون درنظر گرفتن فرایندی که منجر به گم‌شدگی شده است، اغلب منجر به نتایج معتبر قابل قبول خواهند شد [۱۲]. وقتی فرایند گم‌شدگی هیچ یک از دو دسته MCAR و MAR نباشد، گم‌شدگی MNAR رخ داده است، در این حالت گم‌شدگی هم به داده‌های مشاهده شده و هم به داده‌های مشاهده نشده بستگی دارد و اصطلاحاً گم‌شدگی غیرقابل چشم‌پوشی است. از جمله روش‌های معمول برای مقابله با داده‌های گمشده "جانهی

[1] Profile Likelihood
[2] Vectorized Maximum Likelihood
[3] Missing Completely At Random
[4] Missing At Random
[5] Missing Not At Random

چندگانه" و "وزن‌دهی احتمال معکوس" هستند. [۱] رویکرد سومی را معرفی کردند که ترکیبی از دو روش جانهی چندگانه و وزن‌دهی احتمال معکوس است. با توجه به نتایج حاصل از مطالعه شبیه‌سازی روش ترکیبی مزایای بیشتری نسبت به سایر گزینه‌ها داشت. [۲۱] در تحلیل داده‌های فضایی گمشده با استفاده از تکنیک مدل پارامتر اشتراکی، به مدل‌بندی توأم فرایند اندازه‌گیری فضایی و فرایند گمشدگی در یک چارچوب بیزی پرداختند، که با استفاده از آن بخشی از اطلاعات از دست رفته قابل بازیابی است. مدل حاصل تحت سازوکارهای مختلف گمشدگی به ویژه وقتی گمشدگی غیرتصادفی است به درستی عمل می‌کند و در نتیجه اثرات سوء مقادیر گمشده تعدیل می‌شود. [۱۰] به مدل‌بندی داده‌های فضایی گم‌شده از طریق مدل‌های اتورگرسیو فضایی پرداختند به طوری‌که به منظور برآورد پارامترهای مدل و پیشگویی از اطلاعات مقادیر گم‌شده و مشاهده‌شده و میانگین شرطی داده‌های گم‌شده به شرط داده‌های مشاهده‌شده استفاده کردند. برای برآورد این مدل‌ها می‌توان از روش ماکسیمم درستنمایی یا رهیافت بیزی استفاده کرد. در این‌جا داده‌های فضایی گم‌شده از طریق مدل‌های اتورگرسیو فضایی تحت فرض گم‌شدگی قابل چشم‌پوشی مدل‌بندی می‌شوند. بر اساس [۱۰] و تحت فرض گم‌شدگی قابل چشم‌پوشی ابتدا داده‌های گم‌شده برآورد می‌شوند و یک مجموعه داده بازسازی‌شده تولید خواهد شد [۱۲، ۱۶]، سپس با روش ماکسیمم درستنمایی برداری‌شده پارامترهای مدل برآورد می‌شوند. نشان داده می‌شود استفاده از این روش نه تنها باعث افزایش دقت پیشگویی‌ها می‌شود بلکه برآورد پارامترها نسبت به روش کم‌ترین توان‌های دوم که ویژگی فضایی داده‌ها را نادیده می‌گیرد نیز بهبود بخشیده می‌شود. پس از بررسی عملکرد این روش به کمک آزمایش‌های مونت کارلویی، نتیجه می‌شود که برآوردها و استنباط‌هایی که از این روش به دست می‌آیند به اندازهٔ دقت پیشگویی مدل‌هایی است که مبنی بر اطلاعات داده‌های کامل‌اند.

در این مقاله ابتدا انواع مدل‌های اتورگرسیو فضایی معرفی می‌شوند. سپس نحوهٔ برازش این نوع از مدل‌ها به داده‌های فضایی گم‌شده و برآورد ماکسیمم درستنمایی برداری شده در حضور مقادیر گم‌شده معرفی خواهد شد. سرانجام در مطالعه‌ای شبیه‌سازی عملکرد روش ماکسیمم‌درستنمایی برداری شده ارزیابی خواهد شد. همچنین داده‌های انتخابات آمریکا به عنوان مثال واقعی، مورد تحلیل قرار می‌گیرد.

۲. مدل‌های اتورگرسیو فضایی

در یک میدان تصادفی فضایی، مدل‌های اتورگرسیو فضایی، پیشگویی متغیر پاسخ را بر اساس رگرسیون معمولی $y = X\beta + \varepsilon$ از طریق میانگین موزون مقادیر در همسایگی مشاهدات و با در نظرگرفتن ضریب اتورگرسیو فضایی در مدل تصحیح می‌کنند. پرکاربردترین مدل‌های اتورگرسیو فضایی مدل تأخیر فضایی[۶] (SLM)، مدل خطا فضایی[۷] (SEM) و مدل داربین فضایی[۸] (SDM) هستند و به ترتیب به صورت

$$y = \rho W y + X\beta + \varepsilon \qquad (۱)$$

$$y = X\beta + v; \quad v = \rho W v + \varepsilon \qquad (۲)$$

$$y = \rho W y + X\beta + WX\gamma \qquad (۳)$$

[6] Spatial Lag Model
[7] Spatial Error Model
[8] Spatial Durbin Model

تعریف می‌شوند، که در آن‌ها $y = (y_1, \dots, y_n)$ بردار متغیر پاسخ، X ماتریس طرح $n \times k$ بعدی متشکل از k متغیر تبیینی، $\beta = (\beta_1, \dots, \beta_k)$ بردار $k \times 1$ ضرایب رگرسیونی، ρ ضریب مدل اتورگرسیو و نشان‌دهندهٔ شدت وابستگی فضایی، ε جمله خطا با توزیع $N(0, \sigma^2 I_n)$ و $W = (w_{ij})$ ماتریس وزن فضایی $n \times n$ بعدی است. در ماتریس وزن برای دو مشاهده همسایه i و j، w_{ij} مقداری مثبت و برای مشاهدات غیرهمسایه صفر است، همچنین $w_{ii} = 0$، یعنی وزن همسایگی هر مشاهده با خودش برابر صفر است. برای تعریف همسایگی و ساخت ماتریس W روش‌های مختلفی وجود دارد که در بخش بعدی به تفصیل بیان خواهد شد. به عنوان یک قرار داد فرض می‌شود ماتریس W استاندارد شده سطری است، یعنی جمع عناصر هر سطر آن برابر ۱ است [۳]. در برخی ماتریس‌های وزن همسایگی که استاندارد سطری نیستند از تقسیم وزن‌های هر سطر بر مجموع آن سطر، می‌توان به ماتریس وزن استاندارد شده سطری دست یافت، در این‌صورت مجموع عناصر هر سطر برابر با ۱ می‌شود و Wy میانگین موزون مشاهدات خواهد بود. به علاوه پارامتر همبستگی فضایی، ρ صرف نظر از مدل مورد استفاده و به منظور معتبر ماندن ماتریس کواریانس، محدود به بازه $(-\frac{1}{\lambda_{\min}}, \frac{1}{\lambda_{\max}})$ است، که در آن $\lambda_{\min}$ و $\lambda_{\max}$ به ترتیب کوچکترین و بزرگترین مقادیر ویژه ماتریس W هستند [۱۰]. وقتی W استانداردشده سطری یا ستونی باشد، $\lambda_{\max} = 1$ و $|\rho| < 1$، [۶]. چنان‌چه $\rho = 0$، همبستگی فضایی وجود ندارد. لازم به ذکر است که اغلب در به کارگیری مدل‌های اتورگرسیو فضایی علاقه‌مند به وجود همبستگی فضایی مثبت هستیم به همین دلیل در عمل فرض می‌شود $\rho \in (0,1)$. در مدل SLM تغییرات y بر اساس ترکیب خطی از همسایگی‌ها و متغیرهای تبیینی است. در مدل SEM فرض می‌شود که در خطاها همبستگی فضایی وجود دارد و در مدل SDM متغیر پاسخ نه تنها به وزن پاسخ‌ها در همسایگی و متغیرهای تبیینی وابسته است بلکه به وزن متغیرهای تبیینی در همسایگی‌ها نیز وابسته است و γ شدت این ارتباط را توصیف می‌کند. با توجه به اینکه در تمامی مدل‌های اتورگرسیو فضایی درایه‌های ε دارای توزیع $N(0, \sigma^2)$ هستند، لگاریتم تابع درستنمایی به صورت

$$\ell(\sigma^2, \varepsilon) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{\sigma^2}\varepsilon'\varepsilon \qquad (۴)$$

است، که با استفاده از آن تابع درستنمایی برای مدل‌های اتورگرسیو فضایی ساخته می‌شود. مدل SLM را می‌توان به صورت

$$y = (I_n - \rho W)^{-1} X\beta + \varepsilon'; \quad \varepsilon' \sim MVN(0, \sigma^2 (I_n - \rho W)^{-1}(I_n - \rho W')^{-1})$$

نیز نوشت. به منظور برآورد ماکسیمم درستنمایی پارامترهای مدل SLM لگاریتم تابع درستنمایی برای مدل تأخیر فضایی با تغییر متغیر $\varepsilon = (I_n - \rho W)y - X\beta$ به صورت

$$\ell(\beta, \rho, \sigma^2, y) = -\frac{n}{2}\ln(2\pi\sigma^2) + \ln|I - \rho W|$$
$$-\frac{1}{2\sigma^2}((I_n - \rho W)y - X\beta)'((I_n - \rho W)y - X\beta) \qquad (۵)$$

حاصل می‌شود. همچنین مدل SEM را می‌توان به صورت

$$y = X\beta + \varepsilon'; \quad \varepsilon' \sim MVN(0, \sigma^2 (I_n - \rho W)^{-1}(I_n - \rho W')^{-1})$$

نوشت، که در آن $\varepsilon' = (I_n - \rho W)^{-1}\varepsilon$. به این ترتیب مدل SEM به صورت یک رگرسیون خطی عام با ماتریس کواریانس غیرخطی برای جمله خطا تبدیل می‌شود. لگاریتم تابع درستنمایی برای مدل خطا فضایی با تغییر متغیر $\varepsilon = (I_n - \rho W)(y - X\beta)$ به صورت

$$\ell(\beta, \rho, \sigma^2, y) = -\frac{n}{2}\ln(2\pi\sigma^2) + \ln|I - \rho W|$$
$$-\frac{1}{2\sigma^2}[(I_n - \rho W)(y - X\beta)]'[(I_n - \rho W)(y - X\beta) \qquad (۶)$$

به دست می‌آید. مدل SDM را می‌توان به صورت مدل SLM نیز بیان کرد به طوری‌که خواهیم داشت:

$$y = \rho W y + X^*\beta' + e; \quad X^* = (X, WX), \quad \beta' = (\beta, \gamma)$$

و

$$y = (I_n - \rho W)^{-1}X^*\beta' + e'; \quad e' \sim MVN(0, \sigma^2(I_n - \rho W)^{-1}(I_n - \rho W')^{-1}),$$

به این ترتیب لگاریتم تابع درستنمایی مشابه مدل SLM به‌دست می‌آید.

۳. مدل اتورگرسیو فضایی در حضور داده‌های گم‌شده

در داده‌های فضایی با مقادیر گم‌شده، اگر n برابر با تعداد واحدهای نمونه باشد، تنها برای زیرمجموعه‌ای از آن، یعنی $n_{obs} < n$ متغیر پاسخ y مشاهده شده است. به این ترتیب محاسبه برآورد پارامترهای مدل اتورگرسیو فضایی دشوار خواهد بود، زیرا ماتریس وزن همسایگی برای تمام واحدهای نمونه تعریف شده است، در حالی‌که y تنها برای n_{obs} واحد نمونه مشاهده شده است. در بسیاری از مطالعات فضایی که در داده‌ها گم‌شدگی وجود دارد تنها از اطلاعات موجود در داده‌های مشاهده‌شده استفاده می‌شود. در این‌جا به منظور برآورد پارامترهای مدل اتورگرسیو فضایی و پیشگویی، تحت فرض گم‌شدگی قابل چشم‌پوشی از اطلاعات موجود در داده‌های مشاهده‌شده و گم‌شده استفاده خواهد شد تا دقت پیشگویی‌ها بهبود و کارایی برآوردها افزایش یابد. قبل از معرفی روش‌های مناسب برآورد مدل‌های اتورگرسیو فضایی لازم است که هر یک از مدل‌های مورد نظر به دو بخش مربوط به مشاهدات و گم‌شدگی‌ها تجزیه شوند. به عنوان نمونه مدل SEM در نظر بگیرید. ساختار مدل را می‌توان به صورت

$$\begin{pmatrix} y_{obs} \\ y_{mis} \end{pmatrix} = \begin{pmatrix} X_{obs} \\ X_{mis} \end{pmatrix}\beta + \begin{pmatrix} v_{obs} \\ v_{mis} \end{pmatrix} \qquad (۷)$$

$$\begin{pmatrix} v_{obs} \\ v_{mis} \end{pmatrix} = \rho \begin{pmatrix} W_{oo} & W_{om} \\ W_{mo} & W_{mm} \end{pmatrix}\begin{pmatrix} v_{obs} \\ v_{mis} \end{pmatrix} + \begin{pmatrix} \varepsilon_{obs} \\ \varepsilon_{mis} \end{pmatrix} \qquad (۸)$$

تجزیه کرد، که در آن‌ها اندیس‌های obs و mis به ترتیب بردارها و ماتریس‌های مرتبط با داده‌های مشاهده‌شده و گم‌شده را نشان می‌دهند، W_{oo} یک ماتریس $n_{obs} \times n_{obs}$ بعدی و شامل وزن‌های همسایگی مربوط به داده‌های مشاهده‌شده، ماتریس W_{om} از بعد $n_{obs} \times n_{mis}$ است که در آن n_{mis} تعداد واحدهای گم‌شده در نمونه است و به همین ترتیب W_{mo} و W_{mm} تعاریف مشابهی خواهند داشت. همچنین ماتریس X به طور کامل مشاهده شده است و اندیس‌های obs و mis در تجزیه آن، صرفاً بردارهای متناظر با مقادیر گم‌شده و مشاهده‌شده در متغیرهای پاسخ را نشان می‌دهد. فرض کنید موقعیت تمام داده‌های مشاهده‌شده و گم‌شده معلوم هستند. در واقع معلوم بودن موقعیت مکانی

مشاهدات، این امکان را میسر می‌سازد که ماتریس وزن W برای کل نمونه ساخته شود. ماتریس کواریانس خطاها به صورت $\Gamma = \sigma^2[(I-\rho W)(I-\rho W)']^{-1}$ و ماتریس دقت که معکوس ماتریس کواریانس است به صورت

$$\Omega = \Gamma^{-1} = (\frac{1}{\sigma^2})(I-\rho W)(I-\rho W)'$$

خواهد بود. تابع درستنمایی داده‌های مشاهده‌شده به صورت

$$L(\beta, \sigma^2, \rho | y_{obs}) = \int f(y|\beta, \sigma^2, \rho) dy_{mis} \qquad (۹)$$

حاصل می‌شود که برای مدل SEM داریم

$$f(y|\beta, \sigma^2, \rho) = (2\pi\sigma^2)^{-\frac{n}{2}} |I-\rho W| \exp[-(\frac{1}{2})(y-X\beta)'\Omega(y-X\beta)]. \qquad (۱۰)$$

اگر قرار داده شود $e = y - X\beta$، $e_{obs} = y_{obs} - X_{obs}\beta$ و $e_{mis} = y_{mis} - X_{mis}\beta$ آن‌گاه عبارت

$(y-X\beta)'\Omega(y-X\beta)$ در (۱۰) به صورت

$$e'\Omega e = e'_{obs}\Omega_{oo}e_{obs} + e'_{mis}\Omega_{mo}e_{obs} + e'_{obs}\Omega_{om}e_{mis} + e'_{mis}\Omega_{mm}e_{mis} \qquad (۱۱)$$

تجزیه خواهد شد، که در آن Ω_{oo}، Ω_{mo}، Ω_{om} و Ω_{mm} اجزای ماتریس Ω هستند که مشابه ماتریس وزن W تجزیه شده است. عبارت $e_{mis} = y_{mis} - X_{mis}\beta$ در رابطهٔ (۱۱) اطلاعات مربوط به مقادیر گم‌شده متغیرها را وارد تابع درستنمایی می‌کند، از طرفی اطلاعات مربوط به ماتریس دقت فضایی و همچنین ارتباط مقادیر گم‌شده و مشاهده‌شده از طریق Ω_{mo} و Ω_{om} وارد تابع درستنمایی می‌شود. روش مشابهی برای تجزیه مدل SLM قابل اجراست و قابل تعمیم به مدل SDM است، تنها کافیست به جای X در مدل SLM، $X^* = [X\ WX]$ و به جای β، $(\beta, \gamma)'$ جانهی شوند. در تابع درستنمایی داده‌های مشاهده‌شده که در (۹) آمده است، برای مدل SLM عبارت

$$f(y|\beta, \sigma^2, \rho) = (2\pi\sigma^2)^{-\frac{n}{2}} |I-\rho W| \times \exp[-\frac{1}{2}(y-(I-\rho W)^{-1}X\beta)'\Omega(y-(I-\rho W)^{-1}X\beta)]$$

تابع درستنمایی داده‌های کامل است.

۴. برآورد مدل اتورگرسیو فضایی

[۷] در تحلیل داده‌های مستقل با مقادیر گم‌شده از روش کم‌ترین توان‌های دوم عادی (OLS)[9] استفاده کردند، که ابتدا با روش‌های برآورد مقادیر گم‌شده، مجموعه داده را بازسازی کردند [۱۲، ۱۶]. سپس با روش OLS پارامترهای مدل مورد نظر را برآورد کردند که نتایجی مشابه برآورد OLS حاصل از داده‌های کامل به دست آمد. از آن‌جا که در این حالت مشاهدات مستقل از هم هستند، اطلاعات مربوط به یک واحد نمونه روی اطلاعات مربوط به سایر واحدهای نمونه تأثیر ندارند. اما در داده‌های فضایی واحدهایی که در همسایگی یکدیگر قرار دارند روی یکدیگر تأثیرگذار هستند، از این‌رو لازم است مقادیر گم‌شده‌ای که در همسایگی مشاهدات رخ می‌دهند مورد توجه قرار گیرند [۱۰]. به منظور برآورد ماکسیمم درستنمایی مدل‌های اتورگرسیو فضایی تحت فرض گم‌شدگی قابل چشم‌پوشی از اطلاعات موجود در داده‌های گم‌شده و

[9] Ordinary Least Square

مشاهده‌شده استفاده کردند به طوری‌که نتایج حاصل از روش آن‌ها نشان می‌هد که نسبت به روش OLS که تنها بر اساس داده‌های مشاهده‌شده اجرا می‌شود و همچنین وابستگی فضایی داده‌ها را نیز نادیده می‌گیرد، خطاهای پیشگویی تا حد زیادی کاهش می‌یابد.

۱. ۴. برآورد ماکسیمم درستنمایی

از آن‌جا که اطلاعات مربوط به داده‌های گم‌شده y_{mis} ، در دسترس نیستند محاسبهٔ برآورد پارامترها از طریق ماکسیمم کردن تابع درستنمایی میسر نیست. بر اساس [۱۰] می‌توان این مقادیر گم‌شده را با میانگین شرطی روی داده‌های مشاهده‌شده، y_{obs}، جانهی کرد. با فرض اینکه در مدل SEM خطاها دارای توزیع نرمال هستند، $E(y_{mis}|y_{obs})$ به صورت

$$E(y_{mis}|y_{obs}) = \mu_{mis} + \Gamma_{mo}\Omega_{oo}(y_{obs} - \mu_{obs}) \qquad (۱۲)$$

است، که در آن $\mu_{mis} = X_{mis}\beta$ و $\mu_{obs} = X_{obs}\beta$. همچنین در مدل، $E(y_{mis}|y_{obs})$ به صورت

$$E(y_{mis}|y_{obs}) = \mu_{mis} + \Gamma_{mo}\Omega_{oo}(y_{obs} - \mu_{obs}) \qquad (۱۳)$$

حاصل می‌شود، که در آن $\mu_{mis} = B_{mm}X_{mis}\beta + B_{mo}X_{obs}\beta$ و $\mu_{obs} = B_{oo}X_{obs}\beta + B_{om}X_{mis}$. اگر

$$A = \begin{pmatrix} I_{obs} - \rho W_{oo} & -\rho W_{om} \\ -\rho W_{mo} & I_{mis} - \rho W_{mm} \end{pmatrix} \qquad (۱۴)$$

آن‌گاه $B = A^{-1}$. B_{oo}، B_{om}، B_{mo} و B_{mm} تجزیه‌ای از ماتریس B و مشابه تجزیهٔ ماتریس وزن W است، که در آن ماتریس B تنها تابعی از پارامتر ρ است.

توابع امید شرطی در (۱۲) و (۱۳) شامل پارامترها و اطلاعات نمونهٔ مشاهده‌شده هستند، می‌توان مقادیر گم‌شده برآورد شده از طریق امید شرطی را در تابع درستنمایی جانهی کرد و پس از آن به منظور دستیابی به برآوردی برای تمامی پارامترها، تابع درستنمایی را ماکسیمم کرد. تابع درستنمایی برای داده‌های کامل به صورت

$$f(y_{obs}, y_{mis}|X_{obs}, X_{mis}, \beta, \sigma^2, \rho) = f(y_{mis}|y_{obs}, X_{mis}, \beta, \sigma^2, \rho) f(y_{obs}|X_{obs}, X_{mis}, \beta, \sigma^2, \rho)$$

است، که در آن جمله دوم سمت راست، تابع درستنمایی داده‌های مشاهده شده است. به این ترتیب می‌توان از تابع درستنمایی داده‌های کامل نسبت به داده‌های گم‌شده، y_{mis} انتگرال گرفت و تنها $f(y_{obs}|X_{obs}, X_{mis}, \beta, \sigma^2, \rho)$ را نسبت به پارامترها ماکسیمم کرد و به این ترتیب برآورد ماکسیمم درستنمایی برای داده‌های ناکامل به دست خواهد آمد. چنانچه پیشگویی y_{mis} مورد نظر باشد، $E(y_{mis}|y_{obs}, X_{obs}, X_{mis}, \beta, \sigma^2, \rho)$ تابعی از پارامترهای مدل است. برای ماکسیمم کردن لگاریتم تابع درستنمایی نیاز است معکوس ماتریس $n \times n$ بعدی A محاسبه شود که معمولاً از یک روش تقریبی برای جلوگیری از انجام محاسبات پیچیده استفاده می‌شود.

در روش OLS مقادیر گم‌شده در متغیر پاسخ به صورت $E(y_{mis}) = X_{mis}\hat{\beta}$ برآورد می‌شوند که در آن برآورد β بر اساس اطلاعات نمونه مشاهده‌شده به صورت $\hat{\beta} = (X'_{obs}X_{obs})^{-1}X'_{obs}y_{obs}$ به دست می‌آید. در حالی که در برآورد پاسخ‌های گم‌شده بر اساس امید شرطی مقادیر گم‌شده روی مقادیر مشاهده‌شده، از اطلاعات موجود در X_{obs} و

X_{mis} به طور همزمان استفاده می‌شود که در پیشگویی متغیر پاسخ نیز اثر آن وارد می‌شود. به علاوه ساختار کواریانس موقعیت‌های فضایی مقادیر گم‌شده و مشاهده‌شده به شکل Γ_{mo} و Ω_{oo} وارد می‌شود که به آن‌ها می‌توان به عنوان ضریب همبستگی چندگانه نگاه کرد که همبستگی بین y_{mis} و $E(y_{mis}|y_{obs})$ را نشان می‌دهد [۱۴]. این روش از لحاظ نظری به راحتی قابل بیان است، ولی عملیات محاسباتی ماتریس‌های با بعد بالای Γ_{mo} و Ω_{oo} پیچیده و دشوار خواهند بود. چون این ماتریس‌ها تابعی از پارامتر ρ هستند در هر بار عملیات ماکسیمم کردن تابع درستنمایی برای برآورد پارامترهای β، σ^2 و ρ باید محاسبه شوند. همچنین در طی ماکسیمم کردن تابع درستنمایی، لگاریتم دترمینان جمله $|I-\rho W|$ باید محاسبه شود که در صورت سر و کار داشتن با داده‌های حجیم، عملیات محاسباتی نیاز به حافظه‌ای با حجم بالا برای ذخیره اطلاعات خواهد داشت.

از طرفی در روش ماکسیمم درستنمایی سنتی نیاز به روش‌های بهینه‌سازی غیرخطی با استفاده از مشتقات و سایر محاسبات تقریبی است که خود یک مسأله محسوب می‌شود. متأسفانه چنین مسائلی سبب می‌شوند که کاربر به یک مقدار بهینه فراگیر دست پیدا نکند و حتی از خطاهایی که در حین بهینه‌سازی رخ می‌دهد بی‌خبر باشد. [۱۸] مثالی را مطرح کرد، که در آن چندین مقدار بهینه برای یک پارامتر در یک مسأله درستنمایی حاصل می‌شد. در واقع یک برآوردگر ایده‌آل فضایی باید قابلیت به کار گرفتن داده‌هایی با حجم بالا و استنباط سریع داشته باشد و نباید متکی به الگوریتم‌های بهینه‌سازی غیرخطی باشند که تنها یک مقدار را به صورت موضعی و نه فراگیر به عنوان مقدار بهینه در اختیار می‌گذارند. در زیربخش بعدی روشی کارآمد از لحاظ محاسباتی به منظور برآورد پارامترهای مدل اتورگرسیو فضایی به ویژه برای داده‌هایی با حجم بالا، ارایه می‌شود.

۲. ۴. برآورد ماکسیمم درستنمایی برداری‌شده

در مدل‌های اتورگرسیو فضایی، بسیاری از مشاهدات بر یکدیگر تأثیری ندارند و تنها مشاهدات در هر همسایگی بر یکدیگر تأثیر می‌گذارند، در واقع این موضوع باعث ایجاد یک ماتریس وزن تنک خواهد شد که به معنای شیوع درایه‌های صفر ماتریس وزن است. [۱۵] روشی برای محاسبه سریع برآوردها وقتی متغیر وابسته از یک فرایند اتورگرسیو فضایی تبعیت می‌کند ارایه دادند، که در آن با استفاده از ویژگی تنکی ماتریس وزن و بازنویسی محاسبات مربوط به روش ماکسیمم درستنمایی می‌توان برآوردها را با هزینه کم‌تری به دست آورد.

مدل SLM در رابطهٔ (۱) را در نظر بگیرید که به صورت $(I-\rho)y+\rho(I-W)y=X\beta+\varepsilon$ قابل بازنویسی است. در واقع هدف دستیابی به ترکیب محدب بهینه‌ای از y و $(I-W)y$ است. همچنین لگاریتم تابع درستنمایی در رابطهٔ (۵) برای مدل SLM به صورت

$$L(\beta,\rho,\sigma^2)=c+\ln|I-\rho W|-\frac{n}{2}(SSE) \qquad (۱۵)$$

قابل بازنویسی است، که در آن c مقداری ثابت و SSE مجموع توان دوم خطاها است. در این روش ابتدا با ثابت در نظر گرفتن پارامتر ρ بردار ضرایب رگرسیونی β، با توجه به همبسته بودن خطاها در مدل‌های اتورگرسیو فضایی، از طریق

برآورد کم‌ترین توان‌های دوم تعمیم یافته[10](GLS) برآورد می‌شود. در یک مدل ساده رگرسیونی به صورت $y = X\beta + \varepsilon$ چنانچه فرض استقلال خطاها برقرار نباشد، یعنی $\mathrm{Var}(\varepsilon) = \sigma^2\Sigma$، که در آن پارامتر مقیاس σ^2 نامعلوم و Σ (همبستگی و واریانس نسبی بین خطاها) معلوم است، برآوردگر GLS عبارت $(y - X\beta)'\Sigma^{-1}(y - X\beta)$ را مینیمم می‌کند. در نتیجه بردار ضرایب رگرسیونی β به صورت

$$\hat{\beta}_{gls} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \qquad (۱۶)$$

برآورد می‌شود. در مدل‌های اتورگرسیو فضایی چنانچه مقدار ρ معلوم باشد، برآوردگر GLS، β را به طور نااریب برآورد می‌کند و مجموع توان دوم خطا برای مدل SLM به صورت

$$SSE = (y - \rho Wy - X\beta_\rho)'(y - \rho Wy - X\beta_\rho) \qquad (۱۷)$$

است. با توجه به (۱۶)، مقدار بهینه β بر حسب ρ به صورت

$$\beta_\rho = (X'X)^{-1}X'(I - \rho W)y = \beta_o - \rho\beta_w \qquad (۱۸)$$

است، که در آن $\beta_w = (X'X)^{-1}X'Wy$ و $\beta_o = (X'X)^{-1}X'y$. با جانهی (۱۸) در رابطهٔ (۱۷) داریم

$$SSE = (y - X\beta_o - \rho Wy + \rho X\beta_w)'(y - X\beta_o - \rho Wy + \rho X\beta_w)$$

$$= (e_o - \rho e_w)'(e_o - \rho e_w) = e_o'e_o - 2\rho e_w'e_o + \rho^2 e_w'e_w \qquad (۱۹)$$

که در آن e_o و e_w به ترتیب باقی‌مانده‌های حاصل از برآورد OLS رگرسیون y روی X و رگرسیون yW روی X هستند. وقتی برآوردگر GLS به یک پارامتر قابل برآورد وابسته باشد، تبدیل به برآوردگر کم‌ترین توان‌های دوم تعمیم‌یافته برآوردشده خواهد شد. با توجه به روش ماکسیمم درستنمایی لگاریتم تابع درستنمایی قابل بازنویسی از طریق عبارت SSE در رابطهٔ (۱۹) است که با جاگذاری آن در (۱۵) و پس از کنار گذاشتن مقدار ثابت به صورت

$$L(\beta, \rho, \sigma^2) = \ln|I - \rho W| - \frac{n}{2}\ln(e_o'e_o - 2\rho e_w'e_o + \rho^2 e_w'e_w) \quad (۲۰)$$

خواهد بود. اکنون می‌توان لگاریتم تابع درستنمایی را نسبت به ρ ماکسیمم کرد به این صورت که برداری به طول m از مقادیر ρ در بازهٔ [۱,۰) به صورت $\rho_v = (\rho_1, \dots, \rho_m)$ انتخاب می‌شود. سپس مقادیر لگاریتم تابع درستنمایی در هر یک از این مقادیر محاسبه می‌شود، به این ترتیب خواهیم داشت:

$$\begin{bmatrix} \ln(\beta, \rho_1, \sigma^2) \\ \ln(\beta, \rho_2, \sigma^2) \\ \vdots \\ \ln(\beta, \rho_m, \sigma^2) \end{bmatrix} \propto \begin{bmatrix} \ln|I_n - \rho_1 W| \\ \ln|I_n - \rho_2 W| \\ \vdots \\ \ln|I_n - \rho_m W| \end{bmatrix} - \frac{n}{2}\begin{bmatrix} \ln(\Phi(\rho_1)) \\ \ln(\Phi(\rho_2)) \\ \vdots \\ \ln(\Phi(\rho_m)) \end{bmatrix} \qquad (۲۱)$$

که در آن $\Phi(\rho_i) = e_o'e_o - 2\rho_i e_w'e_o + \rho_i^2 e_w'e_w$. با معلوم بودن مقادیر اسکالر $e_o'e_o$، $e_w'e_o$، $\rho^2 e_w'e_w$ و بردار لگاریتم دترمینان مقادیری که تابعی از ρ_v هستند، تعیین مقدار تابع درستنمایی (۲۱) بسیار ساده می‌شود. [۱۵]

[10] Generalized Least Square

استفاده از الگوریتم‌های مستقیم ماتریس تنک مانند تجزیه چولسکی را به منظور محاسبهٔ لگاریتم دترمینان روی مشبکه‌ای از مقادیر پارامتر ρ پیشنهاد دادند. درایه‌ای از بردار ρ_v را که بزرگ‌ترین مقدار لگاریتم درستنمایی را نتیجه می‌دهد با ρ_{ML} نشان می‌دهیم. چنان‌چه مقدار ρ بسیار نزدیک به صفر برآورد شود، لازم است مجدداً مشبکه‌ای از مقادیر ρ در بازه‌ای که شامل مقادیر منفی نیز است تشکیل شود و عملیات ماکسیمم‌سازی روی این مشبکه انجام شود. مدل SDM مشابه مدل SLM برداری می‌شود و کافی است به جای X ماتریس $[X \quad WX]$ و به جای β بردار $(\beta,\gamma)'$ جایگزین شود. از دیدگاه محاسباتی با برداری کردن مسأله از هزینه‌هایی که ممکن است با استفاده از بهینه‌گر غیرخطی، که به طور معمول به همراه تکرار هستند، متحمل شویم جلوگیری می‌شود. در محیط‌های محاسباتی چنین عملیات‌های همراه با تکرار، کارایی را به طور چشم‌گیری کاهش می‌دهند. با استفاده از تعداد متناهی مقادیر ρ در انتخاب $\rho_{m\ell}$ ، در مقایسه با روش ماکسیمم‌سازی از طریق بهینه‌گیر غیرخطی، به میزان اندکی از دقت برآورد کاسته می‌شود ولی تعیین مقدار تابع لگاریتم درستنمایی بر مشبکه‌ای شامل مقادیر ρ بر میزان استواری برآوردگر می‌افزاید.

۳. ۴. برآورد ماکسیمم درستنمایی برداری‌شده در حضور داده‌های گم‌شده

وقتی در داده‌های فضایی گم‌شدگی وجود داشته باشد، قبل از پرداختن به مسأله ماکسیمم‌سازی لازم است داده‌های گم‌شده با مقادیر مناسبی جانهی شوند. با تکیه بر روش [۱۰]، مقدار $E(y_{\mathrm{mis}}|y_{\mathrm{obs}})$ برآورد مناسبی از مقادیر گم‌شده است به طوری‌که با جایگذاری آن در مجموعه داده، بردار y بازسازی خواهد شد. بردار تکمیل شده y قابل استفاده در عبارت برداری شده (۲۱) است تا برآورد ML پارامتر ρ محاسبه شود. پس از دست‌یابی به ρ_{ML} پارامتر β به صورت $\beta = \beta_o - \rho_{\mathrm{ML}}\beta_w$ برآورد می‌شود. همچنین برآورد پارامتر σ^2 بر اساس SSE و به صورت

$$\hat{\sigma}^2 = \frac{1}{n_{obs}}(e_o^{obs} - \rho_{ml}e_w^{obs})'(e_o^{obs} - \rho_{ml}e_w^{obs}) \qquad (۲۲)$$

محاسبه می‌شود، که در آن نماد obs برای مانده‌های e_o و e_w به این معنی است که در روش OLS برآورد پارامترهای مدل و در نتیجه محاسبهٔ مانده بر اساس داده‌های مشاهده‌شده انجام شده است [۱۲]. پس از محاسبهٔ برآورد پارامترها، می‌توان از $E(y_{\mathrm{mis}}|y_{\mathrm{obs}})$ برای پیشگویی مقادیر گم‌شده استفاده نمود، طوری که بردار y بازسازی خواهد شد و با تکرار فرایند برآورد ML می‌توان به مقدار کارآمدی از $E(y_{\mathrm{mis}}|y_{obs})$ دست یافت. همچنین در طی این تکرارها برآورد پارامترها بر اساس روش ماکسیمم درستنمایی برداری شده بهبود بخشیده می‌شود.

۵. کاربرد

ابتدا با فرض وجود گم‌شدگی قابل چشم‌پوشی در متغیر پاسخ فضایی، بر اساس مفروضاتی که بیان شد، از یک مدل اتورگرسیو فضایی شبیه‌سازی می‌شود و برآورد پارامترهای مدل مفروض با روش ماکسیمم درستنمایی برداری شده کارایی روش بازسازی مقادیر گم‌شده، مورد بررسی قرار می‌گیرد. سپس یک مدل اتورگرسیو فضایی به داده‌های انتخابات ریاست جمهوری در کشور آمریکا که در آن متغیر پاسخ حاوی مقادیر گم‌شده است برازش داده شده است و عملکرد روش جانهی پیشنهادی و دقت برآوردها ارزیابی شده است .

۱. ۵. شبیه‌سازی مدل اتورگرسیو فضایی در حضور مقادیر گم‌شده

مطالعه شبیه‌سازی با در نظر گرفتن دو پیکربندی فضایی مختلف از نحوه قرارگیری داده‌های مشاهده‌شده و گم‌شده در همسایگی یکدیگر انجام شده است. ساختار همبستگی فضایی داده‌های مشاهده‌شده و گم‌شده که در تحلیل داده‌های فضایی بسیار مهم است در این اینجا مورد توجه قرار گرفته است. در حالت اول، فرض می‌شود داده‌های گم‌شده و مشاهده‌شده تداخل فضایی بسیار جدی داشته باشند. به این ترتیب تعداد زیادی از عناصر ماتریس‌های w_{om} و W_{mo} غیرصفر خواهند بود و حتی ممکن است تعداد این عناصر غیرصفر برابر یا بیشتر از عناصر غیرصفر ماتریس‌های W_{oo} و W_{mm} باشند. این حالت تحت عنوان "وابستگی زیاد" مورد مطالعه قرار می‌گیرد.

در حالت دوم، داده‌ها به گونه‌ای تولید می‌شوند که موقعیت مکانی داده‌های گم‌شده و مشاهده‌شده ازهم جدا باشند. در این حالت تداخل فضایی بین داده‌های گم‌شده و مشاهده‌شده بسیار کم است و این موضوع باعث شکل گرفتن تعداد کم عناصر غیرصفر در ماتریس‌های وزن فضایی W_{om} و W_{mo} خواهد شد، این ماتریس‌ها ارتباط فضایی مقادیر گم‌شده و مشاهده شده را منعکس می‌کنند. در ادامه این حالت به عنوان "وابستگی کم" مورد بررسی قرار می‌گیرد.

برای ساختن دو پیکربندی فضایی مذکور ابتدا مجموعه مختصات فضایی تصادفی به حجم n=۵۰۰ تولید می‌شود. سپس لازم است مختصات فضایی مرتب شود، این مرتب‌سازی بر اساس بزرگی مجموع طول و عرض جغرافیایی انجام می‌شود. ایده‌های متفاوتی برای ساخت پیکربندی "وابستگی زیاد" وجود دارد، در این‌جا این کار با ایجاد MAR در داده‌ها انجام می‌شود به طوری‌که ابتدا متغیر تصادفی $u = (u_1, \dots, u_n)$ از توزیع یکنواخت U(۰،۱) تولید می‌شود، سپس اگر متغیر نشانگر $m = (m_1, \dots, m_n)$ به صورت

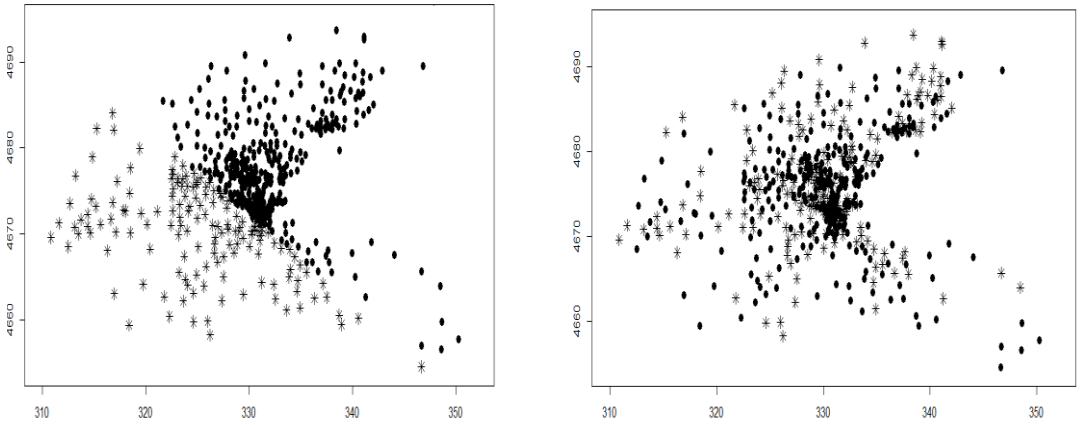
$$m_i = \begin{cases} ۱ & \text{واحد } i \text{ ام گمشده باشد} \\ ۰ & \text{واحد } i \text{ ام مشاهده شده باشد} \end{cases}$$

تعریف شود، آن‌گاه $P(m_i = ۱) = P(u_i < q)$ ، که در آن q یک اسکالر در بازه (۰و۱) است که نسبت تقریبی گم‌شدگی را مشخص می‌کند. با فرض $q = ۰/۳$، به واحدهای نمونهٔ گم‌شدگی تخصیص داده می‌شود. همچنین برای ساخت پیکربندی "وابستگی کم"، چنان‌چه p نسبت مورد نظر از گم‌شدگی باشد، کافی‌ست به نسبت p از واحدهای ابتدایی نمونهٔ گم‌شدگی و به نسبت $p - ۱$ از واحدها، مشاهده شده تخصیص داده شود یا بالعکس. $p = ۰/۳$ در نظر گرفته شده است. در شکل ۱ نحوهٔ قرارگیری داده‌های مشاهده‌شده و گم‌شده درون مختصات فضایی برای این دو حالت نمایش داده شده است. ماتریس وزن W بر اساس همسایگی مرتبهٔ اول ساخته شده است.

در شکل‌های ۲ و ۳ عناصر غیرصفر موجود در W_{oo} و W_{mo} و W_{mm} برای دو حالت مذکور نمایش داده شده است. در حالت "وابستگی زیاد" تعداد عناصر غیرصفر در ماتریس W_{mo} برابر با ۸۴۸ است، همچنین تعداد عناصر غیرصفر در ماتریس‌های W_{oo} و W_{mm} به ترتیب برابر با ۱۴۳۲ و ۵۶۰ است. این در حالی‌ست که برای حالت "وابستگی کم" تعداد عناصر غیرصفر در ماتریس W_{mo} برابر با ۹۳ است که نسبت به تعداد عناصر غیرصفر در ماتریس‌های W_{oo} و W_{mm} که به ترتیب برابر با ۱۷۵۰ و ۱۷۵۲ است، بسیار کم‌تر است. تأثیر پیکربندی‌های متفاوت فضایی که بر اساس ارتباط فضایی نمونه گم‌شده و مشاهده‌شده ساخته شده‌اند در برآورد مقادیر گم‌شده از طریق

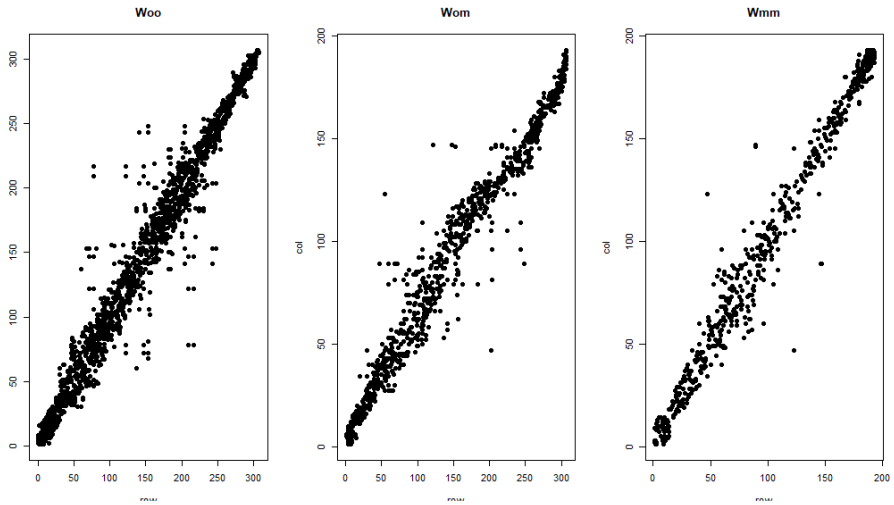
$$E(y_{mis}|y_{obs}) = \mu_{mis} + \Gamma_{mo}\Omega_{oo}(y_{obs} - \mu_{obs}) \qquad (۲۳)$$

مشاهده می‌شود.

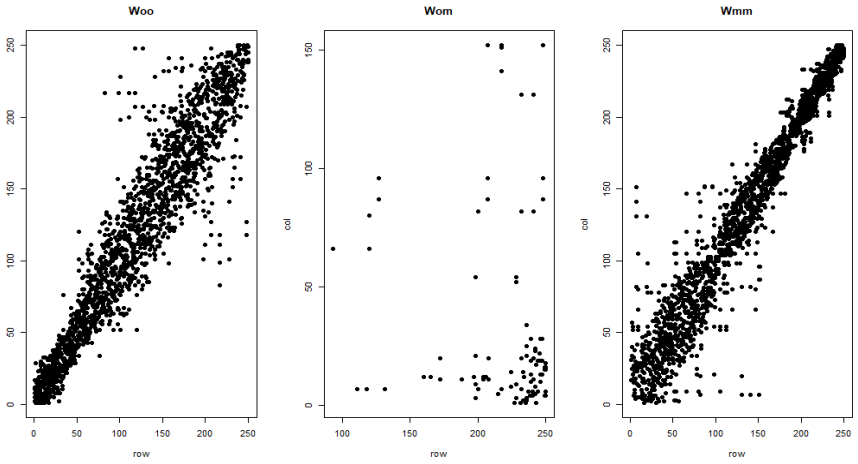


(ب) (الف)

شکل ۱. موقعیت‌های گم‌شدگی (ستاره) و مشاهدات (نقاط سیاه) با پیکربندی الف-وابستگی زیاد و ب- وابستگی زیاد



شکل ۲. درایه‌های غیرصفر موجود در ماتریس‌های W_{mo} W_{oo} و W_{mm} در حالت "وابستگی زیاد"



شکل ۳. درایه‌های غیرصفر در ماتریس‌های W_{mo} W_{oo} و W_{mm} در حالت "وابستگی کم"

در حالت خاص اگر بین نمونه گم‌شده و مشاهده‌شده وابستگی فضایی وجود نداشته باشد، آن‌گاه $W_{mo} = O_{mo}$، که در آن O ماتریس با درایه‌های صفر است. به این ترتیب اطلاعات موجود در خطای پیشگویی که عبارت است از $(y_{obs} - \mu_{obs})$ اهمیت نخواهد داشت، زیرا $E(y_{mis}|y_{obs}) = \mu_{mis}$ و $\Gamma_{mo}\Omega_{oo} = O_{mo}$. در حالتی که وابستگی فضایی بین نمونه گم‌شده و مشاهده‌شده بالا باشد اطلاعات موجود در $(y_{obs} - \mu_{obs})$ اهمیت خواهد داشت و وزن‌های غیرصفر در W_{mo} و W_{om} برای ترکیب اطلاعات موجود در X_{mis} و X_{obs} و $(y_{obs} - \mu_{obs})$ در پیشگویی y_{mis} به کار می‌روند. در این‌جا برای بررسی عملکرد روش پیشنهادی مدل اتورگرسیو SLM در نظر گرفته شده است.

متغیر کمکی x از توزیع نرمال با میانگین صفر و واریانس ۲ تولید شده است، فرایند تولید متغیر پاسخ برای مدل SLM به صورت $y = (I_n - \rho W)^{-1}x\beta + (I_n - \rho W)^{-1}\varepsilon$ است، که در آن $\varepsilon \sim N(0, \sigma_\varepsilon^2)$، $\rho = ۰/۶$، $۰/۴ = \beta$ و $\sigma_\varepsilon^2 = ۱$. برای تولید داده در حالات "وابستگی کم" و "وابستگی زیاد" از نمونه یکسان y و x استفاده شده است، به طوری که تنها ترتیب قرار گرفتن n مشاهده در داده‌های مشاهده شده و گم‌شده تغییر می‌کند. برازش مدل SLM به داده‌های شبیه‌سازی شده بر اساس دو مدل تحت عنوان‌های SLM-MISS و SLM-ALL اجرا شده است. بر اساس مدل SLM-MISS برآورد ماکسیمم‌درستنمایی برداری شده پارامترهای مدل‌های اتورگرسیو فضایی بر اساس روشی که ارائه شد محاسبه و پیشگویی‌ها انجام می‌شود. مدل SLM-ALL بر اساس نمونه کامل اجرا می‌شود و برآورد ماکسیمم درستنمایی برداری شده پارامترهای مدل‌های اتورگرسیو فضایی به دست می‌آید. واضح است که این کار در داده‌های واقعی که گم‌شدگی در آن وجود دارد امکان‌پذیر نیست و در این‌جا به عنوان معیاری برای ارزیابی عملکرد دو روش اول استفاده می‌شود. بدیهی است نتیجهٔ ایده‌آل برای روش مبتنی بر مدل SLM-MISS وقتی حاصل می‌شود که دقت پیشگویی در آن معادل یا نزدیک به دقت پیشگویی بر اساس مدل SLM-ALL باشد، که در آن تمام داده‌ها در دسترس هستند و گام جانهی مقادیر گم‌شده در آن صورت نمی‌گیرد.

اکنون مجموعه‌ای از ۱۰۰۰ دنباله از بردار y و x یکسان تولید می‌شود به طوری‌که در هر تکرار ε بازتولید خواهد شد. برای دو مدل مذکور ۱۰۰۰ برآورد و خطای پیشگویی مقادیر گم‌شده، برای دو نوع پیکربندی فضایی به دست می‌آید. برآورد پارامترهای مدل SLM و انحراف استاندارد[11] (SD) بر اساس این دو مدل در جدول ۱ ارائه شده است.

جدول ۱: برآورد پارامترهای دو مدل SLM-MISS و SLM-ALL

پیکربندی فضایی	پارامتر	مقدارواقعی	مدل SLM-MISS		مدل SLM-ALL	
			برآورد	SD	برآورد	SD
وابستگی زیاد	β_1	۰/۴	۰/۳۹۳	۰/۰۶۹	۰/۳۹۷	۰/۰۵۱
	σ_ε^2	۱	۰/۹۹۰	۰/۱۵۵	۰/۹۸۴	۰/۱۰۷
	ρ	۰/۶	۰/۶۲۱	۰/۰۹۷	۰/۶۱۷	۰/۰۹۲
وابستگی کم	β_1	۰/۴	۰/۴۱۹	۰/۱۳۴	۰/۴۴۷	۰/۰۸۶
	σ_ε^2	۱	۱/۳۳۹	۰/۳۰۹	۱/۳۰۹	۰/۲۳۲

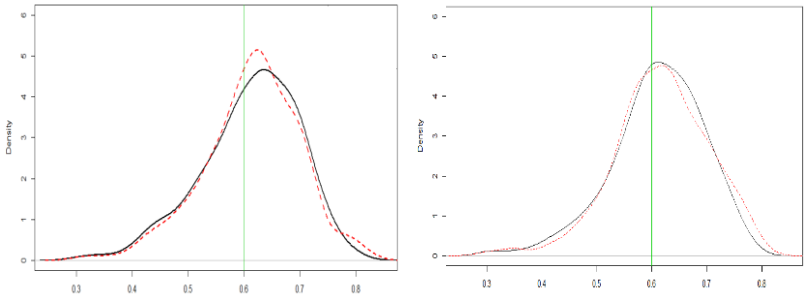
[11] Standard Deviation

ρ	۰/۶	۰/۶۴۷	۰/۱۰۹	۰/۶۱۰	۰/۰۹۲

جدول۲: مقادیر ملاک MAPE برای دو مدل SLM-MISS و SLM-ALL

پیکربندی فضایی	مدل	
	SLM-MISS	SLM-ALL
وابستگی زیاد	۰/۹۱۹	۰/۹۱۰
وابستگی کم	۰/۹۸۵	۰/۹۴۰

بر اساس دو مدل مفروض، پیشگویی درموقعیت‌های گم‌شدگی انجام شده است و میانگین قدرمطلق خطای پیشگویی[12] (MAPE) در جدول ۲ آمده است. در دو حالت "وابستگی زیاد" و "وابستگی کم" دقت پیشگویی مدل SLM-MISS نزدی به مدل SLM- ALL است، که اعتبار روش به کار گرفته شده را تأیید می‌کند. البته در حالت "وابستگی کم" مقدار این کمیت بر اساس مدلSLM-MISS فاصلهٔ بیشتری از مقدار آن براساس مدل SLM-ALL دارد که بیانگر کاهش دقت مدل وقتی است که وابستگی فضایی بین مشاهدات کاهش یافته است. به طور کلی بر اساس نتایج حاصل از آزمایش شبیه‌سازی، مدل SLM-MISS برای نمونه تولید شده عملکرد مناسبی دارد به طوری که چه در برآورد پارامترها و چه در پیشگویی مقادیر گم شده، نتایج بسیار نزدیک به نتایج حاصل از مدل مبتنی بر نمونهٔ کامل،-SLM ALL است.



(ب) (الف)

شکل ۴. توزیع تقریبی برآوردهای ρ در ۱۰۰۰ تکرار برای دو مدل SLM-MISS (خط) و SLM-ALL (خط چین) در دو حالت پیکربندی فضایی الف- وابستگی زیاد و ب- وابستگی کم

۲. ۵. تحلیل فضایی داده‌های انتخابات تحت فرض MAR

برای ارائه کاربردی از مدل‌های اتورگرسیو در مدل‌بندی داده‌های فضایی گم‌شده تحت فرض MAR، از داده‌های مربوط به انتخابات ریاست جمهوری سال ۱۹۸۰ در ۳۱۰۷ شهر آمریکا استفاده شده است. نقطهٔ مرکزی هر شهر به عنوان مختصات فضایی آن شهردر دسترس است. بر اساس وزن‌های فاصلهٔ شعاع، ماتریس وزن W با ابعاد ۳۱۰۷ × ۳۱۰۷ ساخته شده است، که در آن کران مورد نظر برای فاصله، d، میانه فواصل اقلیدسی موقعیت‌های فضایی در نظر گرفته شده است. چهار متغیر تبیین x_1 جمعیت واجد شرایط رأی دادن (بالای ۱۹ سال) در هر شهر، x_2 جمعیت با تحصیلات دانشگاهی در هر شهر، x_3 جمعیت واجد شرایط که دارای مالکیت واحد مسکونی در هر شهر هستند و x_4 متوسط درآمد سرانه افراد

[12] Mean Absolute Prediction Error

واجد شرایط در هر شهر در نظر گرفته شده است. برای کاهش حجم محاسبات، لگاریتم این متغیرها به عنوان متغیرهای تبیینی در مدل در نظر گرفته شده است. همچنین لگاریتم نسبت کل آراء به جمعیت واجد شرایط در هر شهر، یعنی نسبت رأی دهندگان هر شهر، به عنوان متغیر پاسخ منظور شده است. به دلایل نامشخص، حدود ۲۸ درصد از اندازه‌گیری‌های مربوط به متغیر پاسخ گم شده است. برای ارزیابی عملکرد روش ارایه شده در این جا، ابتدا مدل SDM به صورت

$$\text{مدل ۱}: \quad y= \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + \beta_4 \ln(x_4) + W \ln(x_1)\, \beta_5 + W \ln(x_2)\, \beta_6 + W \ln(x_3)\, \beta_7 + W \ln(x_4)\, \beta_8 + \rho W y + \varepsilon$$

با روش ماکسیمم درستنمایی برداری‌شده به داده‌ها برازش داده شده است، سپس به منظور ارزیابی عملکرد مدل۱، مدل ساده رگرسیون به صورت

$$\text{مدل ۲}: \quad y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + \beta_4 \ln(x_4) + \varepsilon$$

با روش OLS به داده‌ها برازش داده شده است. همان‌طور که ملاحظه می‌شود مدل ۱، بر اساس مدل تأخیر فضایی ساخته شده است با این تفاوت که علاوه بر تأخیرهای فضایی متغیر وابسته تأخیر فضایی متغیرهای تبیینی نیز اضافه شده و یک مدل SDM شکل گرفته است. همان‌طور که قبلاً بیان شد مدل SDM حالت خاصی از مدل SLM است.

جدول ۳: برازش دو مدل SDM و رگرسیون خطی به داده‌های انتخابات

	مدل					
	رگرسیون خطی			SDM		
پارامتر	برآورد	SD	p-مقدار	برآورد	SD	p-مقدار
β_0	۱/۰۶۵	۰/۱۲۸	$۲/۶e^{-۱۴}$	۶/۸۹۹	۰/۰۳۶	۰/۰۰۱
β_1	−۰/۶۲۶	۰/۰۵۴	$<۲/۴e^{-۱۶}$	−۰/۷۲۱	۰/۰۲۳	$۸/۶e^{-۶}$
β_2	۰/۸۶۲	۰/۰۵۲	$<۲/۴e^{-۱۶}$	۰/۴۵۳	۰/۰۳۵	$۲/۴e^{-۱۵}$
β_3	۰/۰۵۲	۰/۰۴۱	$۱/۴e^{-۱۰}$	۰/۱۷۰	۰/۴۷۱	۰/۱۲۰
β_4	−۰/۲۸۳	۰/۰۵۷	$۴/۱e^{-۶}$	−۰/۰۴۸	۰/۳۶۸	۰/۶۳۹
β_5	-	-	-	−۵/۳۰۸	۰/۰۳۵	۰/۰۰۳
β_6	-	-	-	۱/۰۹۳	۰/۷۹۶	۰/۰۰۲
β_7	-	-	-	۵/۲۳۶	۰/۰۲۴	۰/۰۰۱
β_8	-	-	-	−۰/۸۵۰	۰/۵۸۴	۰/۵۲۰

برای آزمون همبستگی موران p_مقدار برابر با $۷/۴۰۹e^{-۷}$ حاصل شده است که بیان‌گر وجود وابستگی فضایی بین داده‌ها است [۱۲]. همچنین پارامتر ρ که شدت وابستگی فضایی را بیان می‌کند در مدل SDM برابر با ۰/۴۸۲ برآورد شده است، که به نوبهٔ خود وجود وابستگی فضایی را نیز تأیید می‌کند. نتایج مربوط به برآورد ضرایب رگرسیونی برای دو مدل رگرسیون خطی و مدل SDM در جدول ۲ ارائه شده است. همان‌طور که ملاحظه می‌شود در هر دو مدل ضریب

متغیر جمعیت واجد شرایط مقداری منفی برآورد شده است که این رخداد به دلیل در نظر گرفتن متغیر پاسخ به صورت تابع لگاریتمی نسبت آراء به جمعیت واجد شرایط امری طبیعی است. اثر هر چهار متغیر تبیینی در مدل رگرسیون خطی معنی‌دار است. اثر متغیر میانگین درآمد سرانهٔ افراد واجد شرایط و متغیر وابسته فضایی آن در مدل SDM معنی‌دار نیست. همچنین با وجود عدم معنی‌داری متغیر جمعیت دارای مالکیت واحد مسکونی در مدل SDM، متغیر تأخیر فضایی آن با ضریب بزرگ و مثبتی معنی‌دار است. به علاوه در مدل SDM متغیرهای تأخیر فضایی جمعیت واجد شرایط رأی دادن و جمعیت با تحصیلات آکادمیک معنی‌دار هستند که در مدل OLS نادیده گرفته شده‌اند و این موضوع می‌تواند از نقاط ضعف احتمالی مدل OLS محسوب شود. ضریب تعیین تعدیل‌یافته برای مدل SDM که با روش ماکسیمم درستنمایی برداری شده برازش داده شده است برابر با ۰/۷۳۴ است در حالی که مقدار این کمیت برای مدل رگرسیون خطی برابر با ۰/۶۲۸ است. همچنین مقدار SSE در روش OLS برابر با ۴۳/۳۰۲ و برای روش ماکسیمم درستنمایی برداری‌شده برابر با ۲۵/۶۴۲ است. در پیشگویی متغیر پاسخ در موقعیت‌های گم‌شدگی، مقدار RMSE برای روش OLS برابر با ۰/۲۲۲ و برای روش ماکسیمم درستنمایی برداری‌شده برابر با ۰/۱۸۴ است. مقادیر این کمیت‌ها حاکی از آن است که مدل تأخیر فضایی که با روش ماکسیمم درستنمایی برداری‌شده برازش داده شده است، با به کارگیری اطلاعات فضایی موجود در داده‌ها، منجر به خطای کم‌تری در پیشگویی متغیر پاسخ شده است.

نتیجه‌گیری

در این مقاله تحت فرض قابل چشم‌پوشی بودن گم‌شدگی، مدل‌بندی داده‌های فضایی با مدل‌های اتورگرسیو فضایی و پیشگویی مقادیر گم‌شده انجام شد. سپس روشی مناسب برای برآورد پارامترهای مدل تحت عنوان برآورد ماکسیمم درستنمایی برداری شده معرفی شد که جایگزینی مناسب برای روش ماکسیمم درستنمایی کلاسیک، در موارد مواجهه با پیچیدگی‌های مربوط به ماکسیمم کردن تابع درستنمایی به ویژه در مدل‌های اتورگرسیو فضایی است، به طوری‌که مشکل دستیابی به ماکسیمم موضعی از طریق آن رفع می‌شود. در این راستا انجام آزمایش شبیه‌سازی و مثالی واقعی مورد هدف قرار داده شد که نتایج حاصل حاکی از آن بود که روش معرفی شده و نحوه مدل‌بندی در چنین داده‌هایی عملکرد قابل قبولی داشته و نسبت به روش رایج OLS نتایج بهتری حاصل می‌شود .

تقدیر و تشکر

نویسندگان از هیأت تحریریه، داوران و ویراستار محترم مجله برای پیشنهادهای ارزنده‌ای که موجب ارائه بهتر مقاله شد و از حمایت قطب علمی تحلیل داده‌های وابسته فضایی- زمانی دانشگاه تربیت مدرس قدردانی می‌نمایند.

References

۱. عثمانی، ف. و راسخی، ع. (۱۳۹۷)، روش‌های وزن دهی احتمال معکوس و جانهی چندگانه برای تحلیل پاسخ در حالت گمشدگی، *مجله علوم آماری*، **۱۲**، ۴۶۹-۴۸۳.

2. Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Klumer Academic Publishers, The Netherlands.

3. Anselin, L., Hudak, S. (1992), Spatial Econometrics in Practice: A Review of Software Options, *Journal of Regional Science and Urban Economics*, **22**, 509-536.

4. Besag, J. (1974), Spatial Interaction and the Statistical Analysis of Lattice Systems (With Discussion), *Journal of Statistical Society*, **36**, 192–225.

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

6. Haining, R. (2003), *Spatial Data Analysis: Theory and Practice*, Cambridge University Press, Cambridge.

7. Knight, J. R., Sirmans, C. F., Gelfand, A. E. and Ghosh, S. K. (1998), Analyzing Real Estate Data Problems Using the Gibbs Sampler, *Real Estate Economics* , **26**, 469-492.

8. Lee, L. F. (2002), Consistency and Efficiency of Least-Squares Estimation for Mixed Regressive Spatial Autoregressive Models, *Econometric Theory*, **18**, 252-277.

9. Lesage, J. P. (1999), *Spatial Econometrics*, The Web Book of Regional Science, Regional Research Institute, West Virginia University, Morgantown, WW.

10. Lesage, J. P. and Pace, R. K., (2004), Models for Spatially Dependent Missing Data, *Journal of Real Estate Finance an Economics*, **29**, 233-254.

11. Li, H., Calder, C.A. and Cressie, N., (2012), One-step Estimation of Spatial Dependence Parameters: Properties and Extensions of the APLE Statistic, *Journal of Multivariate Analysis*, **105**, 68–84.

12. Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, John Wiley and Sons.

13. Moran, P. A. P. (1950), Notes on Continuous Stochastic Phenomena, *Biometrika*, **37**, 17-23.

14. Muirhead, R. J. (1982), *Aspects of Multivariate Statistical Theory*. Wiley: New York

15. Pace R. K., Barry R. (1997), Quick Computation of Spatial Autoregressive Estimators, *Geographical Analysis*, **29**, 232-246.

16. Rao, C. R., Toutenburg, H. (1995), *Linear Models: Least Squares and Alternatives*. Springer-Verlag: New York.

17. Ripley, B. (1981), *Spatial Statistics*, New York: John Wiley.

18. Ripley, B. (1988), *Statistical Inference for Spatial Processes*, Cambridge: Cambridge University Press.

19. Rubin, D. B. (1976), Inference and Missing Data. *Biometrika*, **63**, 581-92.

20. Whittle, P. (1954), On Stationary Processes in the Plane, *Biometrika*, **41**, 434-449.

21. Zahmatkesh, S., Mohammadzadeh, M. (2021), Bayesian Prediction of Spatial Data With Non-ignorable Missingness, *Statistical Papers*, *62*, 2247-2267.