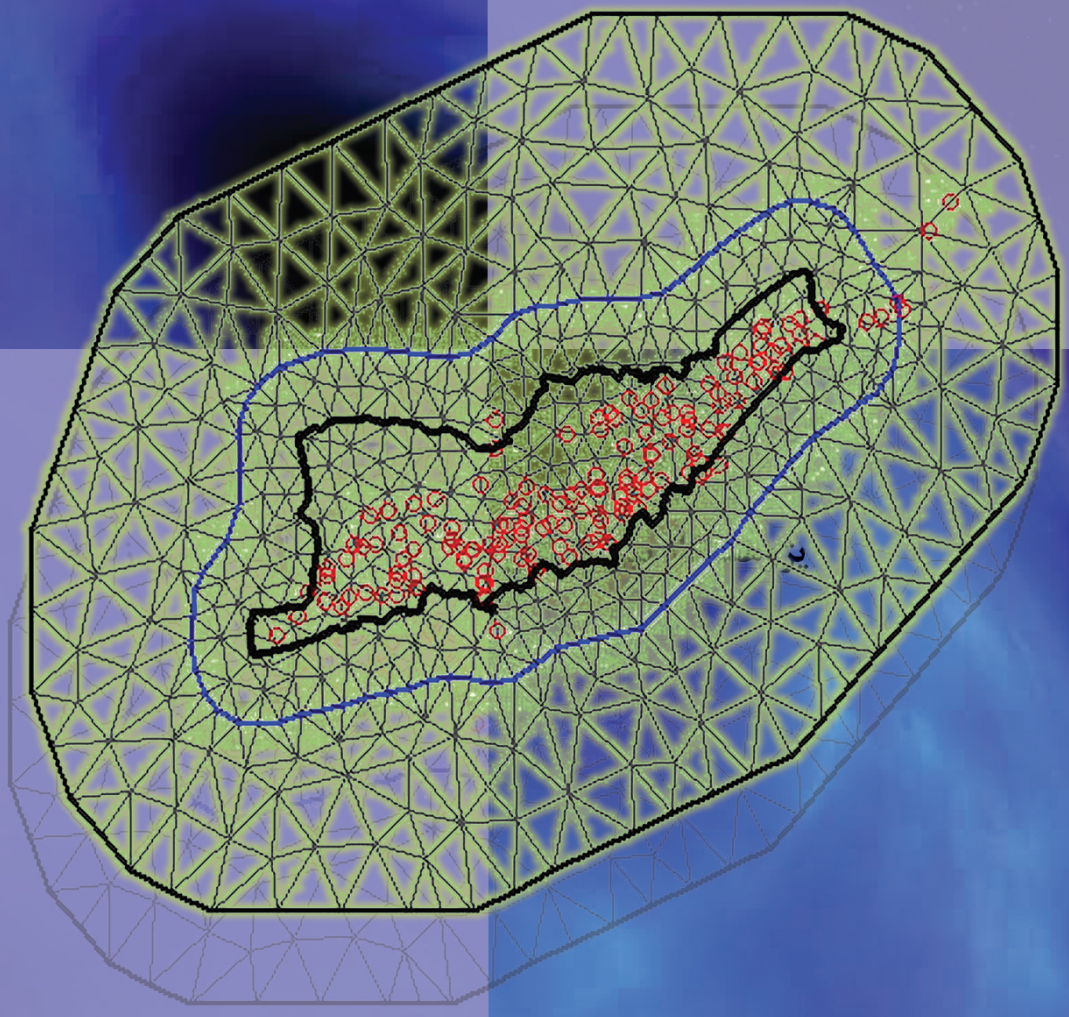
پنجمین سمینار آمار فضایی و کاربردهای آن

۳-۴ آبان ماه ۱۴۰۲
دانشگاه بین‌المللی امام خمینی، قزوین

5th Seminar on Spatial Statistics and Its Applications

Oct. 2023 25-26
Imam Khomeini International University



مجموعه مقالات

دانشگاه نیشابور
دانشگاه صنعتی امیرکبیر
دانشگاه رازی
دانشگاه تهران
دانشگاه صنعتی اصفهان

پژوهشگاه فوه فضا بیه
مؤسسه ژئوفیزیک
دانشگاه صنعتی قم
دانشگاه سمنان
پژوهشکده آمار
مرکز آمار ایران

به نام خدا

پنجمین سمینار آمار فضایی و کاربردهای آن

مجموعه مقالات

آبان ۱۴۰۲

دانشگاه بین‌المللی امام خمینی

مجموعه مقالات پنجمین سمینار آمار فضایی و کاربردهای آن

تدوین کننده: محسن محمدزاده

چاپ: اول آبان ماه ۱۴۰۲

آدرس دبیرخانه: قزوین، دانشگاه بین‌المللی امام خمینی

تلفن: ۳۳۹۰۱۳۷۹ – ۰۲۸

پست الکترونیک: spatial5@ikiu.ac.ir

آدرس سایت: conf.ikiu.ac.ir/spatial5

پیش‌گفتار

توسعه روزافزون جوامع بشری در زمینه‌های مختلف هر روز شتاب بیشتری می‌گیرد. برای غلبه بر این رشد فزاینده و کنترل آن، نیاز به روش‌های پیشرفته مدل‌سازی پدیده‌های مختلف اهمیت دو چندان می‌یابد . اغلب پدیده‌های تجربی دارای یک سری متغیرهای مستقل و وابسته هستند. کشف و مدل‌سازی وابستگی این متغیرها نقش حیاتی در درک بهتر مبتنی بر واقعیت از آن پدیده‌ها دارد. علوم آماری و روش‌های جدید علم داده نقش کلیدی در این زمینه ایفا می‌کنند و همکاری‌های میان رشته‌ای را ارتقا می‌دهند. آمار فضایی ابزار قدرتمندی است که با تحلیل داده‌های فضایی و زمانی، همبستگی‌های آنها را بررسی می‌کند. با در نظر گرفتن این موضوع، روش‌های آمار فضایی را می‌توان در طیف وسیعی از زمینه‌ها مورد استفاده قرار داد. از جمله در علوم و مهندسی زلزله، مهندسی ریسک، مدیریت بحران، علوم جوی و هواشناسی، منابع آب، محیط زیست، زمین شناسی، معدن، برنامه‌ریزی شهری و منطقه‌ای، ترافیک، حمل و نقل، سنجش از دور، بهداشت و درمان، همه‌گیرشناسی، پزشکی قانونی، علوم اجتماعی، نفت و گاز، اقتصاد و بیمه کاربردهای گسترده ای دارند. به منظور ایجاد فرصت تبادل نظر متخصصان در علوم مرتبط با آمار فضایی، پنجمین سمینار آمار فضایی و کاربردهای آن از تاریخ ۳ تا ۴ آبان ۱۴۰۲ به میزبانی دانشگاه بین‌المللی امام خمینی و با همکاری قطب علمی تحلیل داده‌های همبسته فضایی_زمانی دانشگاه تربیت مدرس و انجمن آمار ایران برگزار می‌شود. این سمینار فرصت بی‌نظیری را برای دانشگاهیان، متخصصان، سازمان‌های دولتی، بخش خصوصی و سایر نهادهای فعال در زمینه‌های مختلف مرتبط با آمار فضایی فراهم می‌کند تا با ارائه آخرین دستاوردهای علمی، به تبادل نظر و ارائه نتایج تحقیقات خود بپردازند. با تشکر از کارشناسان محترم داخل و خارج از کشور در زمینه‌های مختلف که با ارائه مقالات ارزشمند خود در ثمربخشی علمی این سمینار سهیم هستند و از داوران محترم، کمیته علمی و کمیته اجرایی که برای برگزاری این سمینار تلاش فراوانی کردند، تشکر می‌کنم. امیدواریم با حضور و مشارکت فعال شما در این سمینار بتوان به اهداف پیش‌بینی شده آن مانند سمینارهای موفق قبلی دست یافت.

دبیر کمیته علمی سمینار
دکتر محسن محمدزاده
آبان ۱۴۰۲

همکاران و حامیان سمینار:

این سمینار به میزبانی دانشگاه بین‌المللی امام خمینی و با همکاری قطب علمی تحلیل داده‌های وابسته فضایی و فضایی‌ـ زمانی دانشگاه تربیت مدرس و انجمن آمار ایران و همچنین حمایت و پشتیبانی سازمان‌ها و موسسات فهرست شده زیر برگزار می‌گردد. بدین وسیله نهایت قدردانی خود را از همه افراد و سازمان‌هایی که سمینار را مورد حمایت‌های مختلف قرار دادند، ابراز می‌داریم.

اعضای کمیته علمی

۱. دکتر محسن محمدزاده (دبیر کمیته علمی)	دانشگاه تربیت مدرس
۲. دکتر افشین فلاح	دانشگاه بین‌المللی امام خمینی
۳. دکتر علی آقامحمدی	دانشگاه زنجان
۴. دکتر حسین باغیشنی	دانشگاه صنعتی شاهرود
۵. دکتر مرتضی بسطامی	پژوهشگاه بین‌المللی زلزله‌شناسی و مهندسی زلزله
۶. دکتر فاطمه حسینی	دانشگاه سمنان
۷. دکتر رامین کاظمی	دانشگاه بین‌المللی امام خمینی
۸. دکتر امید کریمی	دانشگاه سمنان
۹. دکتر موسی گل‌علی‌زاده	دانشگاه تربیت مدرس
۱۰. دکتر کیومرث مترجم	دانشگاه تربیت مدرس

اعضای کمیته برگزاری

۱. دکتر افشین فلاح (دبیر سمینار)	دانشگاه بین‌المللی امام خمینی
۲. دکتر صدیقه زمانی مهریان	دانشگاه بین‌المللی امام خمینی
۳. دکتر الیاس شیوانیان	دانشگاه بین‌المللی امام خمینی
۴. دکتر مریم درگاهی	دانشگاه بین‌المللی امام خمینی
۵. دکتر رامین کاظمی	دانشگاه بین‌المللی امام خمینی
۶. دکتر آرزو حاج‌رجبی	دانشگاه بین‌المللی امام خمینی
۷. دکتر مهسا نادی‌فر	دانشگاه پرتوریا، آفریقای جنوبی

فهرست مقالات

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

دیداری سازی داده‌های بعد بالا از طریق درخت نشانیدنی همسایگی تصادفی

فاطمه اروجلو[1]، موسی گلعلی‌زاده
گروه آمار، دانشگاه تربیت مدرس

چکیده: تحلیل مجموعه داده‌های با ابعاد بالا که در آن‌ها تعداد ویژگی‌ها معمولا بیشتر از تعداد مشاهدات است، در زمینه‌های بسیار متفاوتی شامل تصویرسازی دیجیتالی، تحقیقات بیولوژیکی، ژنتیکی و برخی علوم مبتنی بر تکنولوژی، کاربرد دارد. رویکردهای سنتی آماری در تحلیل و دیداری‌سازی داده‌های با ابعاد بالا اغلب به شکست برمی‌خورند. به همین منظور، در تحلیل داده‌های با ابعاد بالا به کار بردن روش‌های کاهش ابعاد و سپس بهره‌برداری از روش‌های نوین آماری از اهمیت بسزایی برخوردار است. روش t-SNE و خوشه‌بندی سلسله‌مراتبی دو روش رایج تحلیل خوشه‌بندی داده‌ها هستند. از طرفی دیگر، ایجاد درخت بر اساس رویکرد SNE نیز روشی برای تفکیک مجموعه داده‌ها با رویکرد همبستگی مبتنی بر همسایگی است. نکته حائز اهمیت این است که درخت SNE از توانایی سرعت بالای t-SNE و سادگی اجرای خوشه‌بندی سلسله مراتبی بهره برده و آن‌ها را بر اساس اصول علمی به طریق مناسب با هم ترکیب می‌کند. در تحقیق حاضر روش درخت SNE به طور خلاصه مورد بررسی قرار گرفته و چگونگی به کار بردن این روش برای دیداری‌سازی و شفاف‌سازی ساختارهای سلسله مراتبی مجموعه داده‌های با ابعاد بالا نشان داده می‌شود.

واژه‌های کلیدی: دیداری‌سازی، خوشه‌بندی سلسله مراتبی، t-SNE، نشانیدنی، ارقام دست نویس.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H99، 62H30.

---

۱ مقدمه

چالشی که در هنگام تحلیل داده‌های بعد بالا وجود دارد، روبه‌رو شدن با مشقت بعد چندی[1] است. این اصطلاح به پدیده‌های گوناگونی که هنگام تحلیل داده‌ها در فضای با ابعاد بالا روی می‌دهد، اطلاق می‌شود (بلمن، ۱۹۵۷). بنابراین در تحلیل و دیداری‌سازی داده‌های بعد بالا، به کارگیری روش‌های کاهش ابعاد[1] ــ به منظور مقابله با مسئله‌ی مشقت بعد چندی ــ و سپس بهره‌برداری از روش‌های نوین آماری از اهمیت بسزایی برخوردار خواهد بود.

[1] Curse of dimensionality
[1] Dimension reduction

[1] نام و ایمیل ارائه دهنده مقاله: فاطمه اروجلو، mohsen_m@modares.ac.ir@modares.ac.ir

الگوریتم نشانیدنی همسایگی تصادفی[۲] (SNE) یک روش کاهش بعد غیر خطی است که مشاهدات موجود در فضای با ابعاد بالا را به منظور حفظ ساختار محلی بهینه و دیداری‌سازی آن‌ها، در یک فضای کاهش بعد یافته می‌نشاند (هینتون و رویز، ۲۰۰۲). روش SNE از یک معیار احتمال شباهت مبتنی بر چگالی توزیع نرمال استفاده می‌کند، به این صورت که دو احتمال نزدیکی را یک بار بین جفت نقاط در فضای با ابعاد بالا و بار دیگر بین جفت نقاط در فضای کاهش بعد یافته، در نظر می‌گیرد. بر اساس این معیار مشاهده i به ازای $i = ۱, ..., n$ مشاهده‌ی j به ازای $j = ۱, ..., n$ $(i \neq j)$ را به عنوان همسایه خود در فضای کاهش بعد یافته انتخاب خواهد کرد اگر این دو مشاهده در فضای با ابعاد بالا شبیه به هم باشند یا به عبارتی فاصله‌ی آن‌ها از هم کم باشد. از هر دو منظر ریاضی و آمار، هدف از نشانیدنی این است که واگرایی بین دو توزیع نرمال در فضای با ابعاد بالا و فضای با ابعاد پایین کمینه شود. این امر با به حداقل رساندن یک تابع هزینه[۳] حاصل می‌شود که برابر مجموع واگرایی‌های کولبک–لیبلر[۴] بین دو توزیع نرمال مورد اشاره است (کولبک، ۱۹۹۷). هنگامی که واگرایی KL به کمترین مقدار خود برسد، SNE به یک نمایش مناسبی از داده‌های با ابعاد بالا در یک فضای کاهش بعد یافته منجر می‌شود (هینتون و رویز، ۲۰۰۲).

ون‌درماتن و هینتون (۲۰۰۸) نسخه بهبود یافته‌ای از روش SNE به نام نشانیدنی همسایگی تصادفی توزیع شده بر اساس توزیع t-استیودنت[۵] (t-SNE)، را معرفی کردند. روش t-SNE نیز شبیه روش SNE دو توزیع احتمال را مدنظر قرار می‌دهد اما یک تفاوت جدی با روش SNE دارد. روش t-SNE، یک توزیع نرمال بر روی جفت نقاط در فضای با ابعاد بالا و توزیع t دیگری بر روی جفت نقاط در فضای بعد پایین در نظر می‌گیرد، طوری که به مجموعه نقاط مشابه هم احتمال زیاد و نقاط غیرمشابه احتمال کم‌تری اختصاص یابد. آنگاه t-SNE، واگرایی KL بین دو توزیع احتمال مورد اشاره را کمینه می‌کند تا به یک نمایش تصویری مناسب از داده‌های اولیه برسد.

نکته قابل توجه این است که روش SNE از یک تابع هسته نرمال برای تبدیل فاصله جفت مشاهدات از هم به احتمال شباهت استفاده می‌کند، ولی روش t-SNE از یک توزیع t با درجه آزادی یک استفاده می‌کند. دم‌های سنگین‌تر توزیع t در مقایسه با توزیع نرمال به کاهش مشکل ازدحام کمک می‌کند طوری که حباب‌های متمایز از هم در نشانیدنی بعد پایین ظاهر شوند (رابنسون و پیرس–هافمن، ۲۰۲۰).

درخت SNE[۶] یک روش خوشه‌بندی سلسله‌مراتبی[۷] مبتنی بر t-SNE یک بعدی است طوری که با استفاده از آن امکان دیداری‌سازی و شفاف‌سازی ساختارهای سلسله مراتبی داده‌های با ابعاد بالا فراهم می‌شود. روش درخت SNE با دیداری‌سازی داده‌ها در سطوح مختلف می‌تواند هم برای کشف ساختارهای ذاتا سلسله مراتبی درون داده‌ها و هم کشف داده‌های جدید مفید باشد. دیداری‌سازی و تحلیل داده‌ها با درخت SNE اجازه می‌دهد که یک محقق تصمیم بگیرد که کدام مقیاس برای خوشه‌بندی و نمایش تصویری داده‌های او بهتر است. این کار با نشان دادن یک دیدگاه عمیق نسبت به سازماندهی سلسله مراتبی داده‌ها در مقایسه با t-SNE صورت می‌گیرد (رابنسون و پیرس–هافمن، ۲۰۲۰).

در ادامه روش درخت SNE به طور خلاصه شرح داده خواهد شد. سپس نتایجی از کاربست این روش نوین برای دیداری‌سازی دو مجموعه داده ارقام دست‌نویس نشان داده خواهد شد. در آخر در بخش نتیجه‌گیری به طور مختصر علل تمایز روش درخت SNE از دیگر روش‌های مشابه شرح داده خواهد شد.

---

[2] Stochastic Neighbor Embedding
[3] Cost function
[4] Kullback-Leibler
[5] t-Stochastic Neighbor Embedding
[6] Tree-SNE
[7] Hierarchical clustering

۲ خلاصه‌ای از درخت SNE

روش t-SNE یکی از محبوب‌ترین روش‌های کاهش ابعاد غیرخطی برای دیداری‌سازی داده‌های با ابعاد بالا است (ون‌درماتن و هینتون، ۲۰۰۸). رویکرد t-SNE این کار را با مدل‌سازی هر نقطه با ابعاد بالا در یک فضای دو یا سه بعدی انجام می‌دهد، طوری که نقاط مشابه نزدیک به یکدیگر و نقاط غیرمشابه، دورتر مدل‌سازی شوند. برای انجام این کار، t-SNE دو توزیع احتمال می‌سازد، یکی بر روی جفت نقاط در فضای با ابعاد بالا و دیگری بر روی جفت نقاط در فضای بعد پایین، طوری که به مجموعه نقاط مشابه هم احتمال زیاد و نقاط غیرمشابه احتمال کم‌تری اختصاص می‌یابد. آنگاه t-SNE، واگرایی KL بین دو توزیع احتمال مورد اشاره را کمینه می‌کند تا به یک نمایش تصویری مناسب از داده‌های اولیه برسد.

دلیل استفاده از توزیع t برای تبدیل فواصل به احتمالات این است که مشاهداتی که خیلی دورتر از مشاهده خاصی باشند با استفاده از توزیع t کماکان شانسی (حتی ناچیز) برای حضور هنگام اجرای روش t-SNE دارند. به کارگیری توزیع t با توجه به پهن‌تر بودن دم‌های آن نسبت به توزیع نرمال، کمک می‌کند تا خوشه‌های متمایز (در صورت وجود) در فضای تصویر پخش شوند و این موضوع به شناسایی راحت‌تر آن‌ها کمک می‌کند (رابنسون و پیرس ـ هافمن، ۲۰۲۰).

کوباک و همکاران (۲۰۱۹)، با الهام از لیندرمن و همکاران (۲۰۱۹) و با استفاده از روش تبدیل فوریه روشی برای پیاده‌سازی یک نسخه سریع از t-SNE برای درجات آزادی کسری و کوچک‌تر ارائه کرده‌اند. آن‌ها از یک تابع هسته مقیاس‌بندی شده استفاده کرده‌اند که به صورت

$$k(d) = \frac{۱}{(۱ + \frac{d^۲}{\alpha})^{\alpha}} \qquad (۱.۲)$$

تعریف می‌شود. درجه آزادی توزیع t در اینجا برابر با $v = ۲\alpha - ۱$ است. چون ضروری است که درجات آزادی اعدادی مثبت باشند ($v > ۰$)، پس α هم مقداری مثبت است ($\alpha > ۰$). کوباک و همکاران (۲۰۱۹) دریافتند که در حالت $\alpha < ۱$ (که هم‌ارز نامساوی $v < ۱$ است)، t-SNE نمودارهایی با استحکام و با حباب‌های کوچک‌تر و دانه‌ریزتر نسبت به t-SNE استاندارد با $\alpha = ۱$ تولید می‌کند. رابنسون و پیرس ـ هافمن (۲۰۲۰)، با استفاده از ایده روش‌های t-SNE و خوشه‌بندی سلسله مراتبی، درخت SNE که یک روش خوشه‌بندی سلسله مراتبی مبتنی بر t-SNE یک بعدی با مقادیر کاهشی α و سرگشتگی در هر سطح (لایه) است را ارائه کردند. درخت SNE امکان دیداری‌سازی و شفاف‌سازی ساختارهای سلسله مراتبی داده‌های با ابعاد بالا را فراهم می‌سازد. این کار با ایجاد نشانیدنی‌های t-SNE با دم‌های سنگین، برای آشکار کردن ساختارهای دانه‌ریزتر (حباب‌های بیشتر) و سپس انباشتن آن‌ها روی هم انجام می‌شود تا در نهایت ساختاری درخت مانند را ایجاد کند. هنگامی که α کاهش می‌یابد، تعداد خوشه‌ها افزایش خواهد یافت. لازم به اشاره است که در مرحله دیداری‌سازی‌های درخت SNE، معمولا محور عمودی نمایانگر شماره لایه (سلسله مراتبی) و محور افقی بیانگر مختصات نشانیدنی t-SNE یک بعدی خواهد بود.

برای ایجاد نشانیدنی‌های درخت SNE، الگوریتم ابتدا با نشانیدنی استاندارد t-SNE یک بعدی $\alpha = ۱$ و با سرگشتگی[۱] بالا، که به طور پیش فرض جذر تعداد نقاط داده است، شروع می‌شود. بنا به ون‌درماتن و هینتون (۲۰۰۸)، شروع الگوریتم با سرگشتگی بالا شانس یافتن تعداد همسایگی‌های موثر استفاده شده توسط t-SNE را افزایش می‌دهد. این بدان معنا است که تشکیل خوشه‌های بزرگ‌تر تسریع شده و داده‌های بیشتری در خوشه‌ها قرار گرفته و ساختارهای کلی بزرگ‌تری را در داده‌ها به وجود می‌آورد.

رابنسون و پیرس ـ هافمن (۲۰۲۰) ملاحظه کردند که شروع الگوریتم با استفاده از یک سرگشتگی بزرگ باعث می‌شود

[۱] Perplexity

که، درخت SNE بتواند کل طیف سازماندهی داده‌ها، از ساختارهای کلی در بدنه درخت گرفته تا ساختارهای بسیار ریز برگ‌ها در قسمت‌های انتهایی را نمایش دهد. انتخاب مقدار سرگشتگی پیش‌فرض اولیه $\sqrt{N}$ برای تحلیل سرگشتگی در اجرای روش t-SNE توسط اسکولکوف (۲۰۱۹) پیشنهاد شد که در عمل منجر به نتایج مطلوبی می‌شود.

کوباک و برنز (۲۰۱۹) اعتقاد دارند که در بسیاری از آزمایش‌های علمی انتخاب مقدار بسیار بزرگی برای پارامتر سرگشتگی خیلی مهم نیست، زیرا روش t-SNE نسبت به تغییرات کوچک در مقدار سرگشتگی نسبتا مقاوم است. با تغییر مقادیر پارامترهای سرگشتگی و α تعداد زیادی از نشانیدنی‌های t-SNE یک بعدی (معمولا با ۳۰ یا ۱۰۰ بار تکرار) تولید می‌شوند. آنگاه این تعداد در یک نمودار دو بعدی روی هم قرار می‌گیرند. در این نمودار، محور عمودی شماره تکرار الگوریتم t-SNE (شماره لایه درخت) و محور افقی مختصات به دست آمده از این الگوریتم است. در نهایت، به وسیله این نمودار، دیداری‌سازی درخت SNE با اولین لایه در پایین طرح ایجاد شده و سپس α و سرگشتگی در هر کدام از لایه‌های بعدی کاهش می‌یابند. بنا به ون‌درماتن و هینتون (۲۰۰۸) برای تولید هر سطح متوالی، α استفاده شده در سطح قبلی در یک عامل ثابت r که $0 < r < 1$ (در عمل r بسیار نزدیک به ۱ است)، ضرب می‌شود. چنین ایده‌ای می‌تواند به صورت $\alpha_{n+1} = r\alpha_n$ بیان شود، که α_n به مقدار α استفاده شده در نشانیدنی لایه n اشاره دارد. هم‌چنین اگر p_n نشان دهنده سرگشتگی در سطح n باشد، آنگاه $p_{n+1} = p_n^r$. به این ترتیب، α به صفر و سرگشتگی به ۱ نزدیک می‌شود. با چنین ساختاربندی، می‌توان خوشه‌ها را حاوی تنها یک نقطه منفرد در نظر گرفت، زیرا هیچ محدودیتی برای نقاط در تشکیل خوشه‌ها وجود ندارد. بر این اساس مشاهده می‌شود که رفته رفته با نزدیک شدن α به صفر و سرگشتگی به یک، خوشه‌ها کوچک‌تر شده و تعدادشان بیش‌تر می‌شود و در هر لایه به سمت بالا حرکت می‌کنند.

هنگام ایجاد هر لایه، نشانیدنی t-SNE با نشانیدن از سطح قبلی به جای یک مقدار اولیه استاندارد شروع می‌شود. این کار به گونه‌ای انجام می‌شود که هر لایه یک پالایش از خوشه‌بندی یافت شده در لایه قبلی با خوشه‌های بزرگ‌تری که به خوشه‌های کوچک‌تر در سطوح بعدی درخت تقسیم می‌شوند، است. در نتیجه مسیر یک مشاهده یا گروهی از مشاهدات را می‌توان به صورت عمودی از طریق دیداری‌سازی درخت حاصل دنبال کرد.

۳ کاربست درخت SNE روی مجموعه ارقام دست‌نویس

مجموعه داده ارقام دست‌نویس «اصلاح شده موسسه ملی استاندارد و فناوری»[۲] (MNIST)، دارای یک مجموعه آموزشی شامل ۶۰۰۰۰ نمونه و یک مجموعه آزمایشی شامل ۱۰۰۰۰ نمونه است. این مجموعه داده، زیر مجموعه یک مجموعه بزرگتر است که پایگاه داده «موسسه ملی استاندارد و فناوری»[۱] (NIST) ایجاد کرد. هر تصویر انتخاب شده از NIST طی دو مرحله، با حفظ نسبت تصویر، یک بار در جعبه‌های ۲۰ در ۲۰ پیکسل مرکزیت داده شده و سپس دوباره در جعبه‌های ۲۸ در ۲۸ پیکسل جای گرفته‌اند. همچنین در مرحله دوم عددی که در هر تصویر قرار دارد با محاسبه مرکز ثقل در میانه تصویر قرار می‌گیرد. در واقع ابعاد هر تصویر دارای طول ۲۸ پیکسل و عرض ۲۸ پیکسل است که در مجموع ضریب ۷۸۴ پیکسل را تشکیل می‌دهد. هر پیکسل دارای یک مقدار عددی است که نه تنها مقدار پیکسل را نشان می‌دهد بلکه نحوه انتساب خود به پیکسل را معرفی می‌کند. این مقدار نشان دهنده درجه رنگ هر پیکسل در طیف رنگی سفید مطلق تا سیاه مطلق است. مقادیر آن نیز یک عدد صحیح بین صفر تا ۲۵۵ است طوری که عدد صفر نشان دهنده سفید مطلق و عدد ۲۵۵ نشان دهنده سیاه مطلق است. با تقسیم این مقادیر بر ۲۵۵، طیف رنگی از صفر تا ۱ تشکیل شده که عدد صفر سفید مطلق و عدد ۱ سیاه مطلق است.

لازم به اشاره است که مجموعه داده‌های آزمایشی، ۷۸۵ ستون دارد. اولین ستون آن که برچسب (label) نام دارد،

[2] Modified National Institute of Standards and Technology

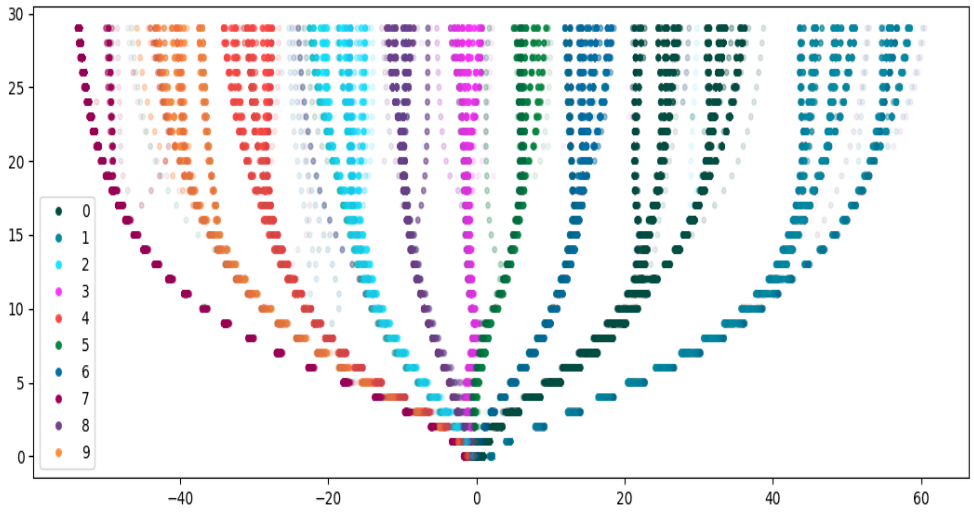
[1] National Institute of Standards and Technology

رقمی است که توسط کاربر ترسیم شده است. مابقی ستون‌ها حاوی مقادیر پیکسل تصویر مربوطه هستند.

درخت SNE سه پارامتر مهم نسبت r، سرگشتگی اولیه (p) و تعداد لایه‌ها (n) دارد. همانطور که شرح داده شد سرگشتگی اولیه جذر تعداد نقاط داده در نظر گرفته می‌شود. از آنجایی که نشانیدنی‌های درخت SNE تمایل دارند ساختارهای بسیار منشعبی را ایجاد کنند، کران پایین α روی ۰/۰۱ تنظیم می‌شود. با توجه به مقدار ثابت کران پایین α، نسبت r مورد نیاز برای کاهش α از ۱ به ۰/۰۱، در تعداد معینی از لایه‌های n از طریق رابطه

$$r = \exp(\frac{\log ۰/۰۱}{n}) \qquad (۱.۳)$$

به دست می‌آید. با استفاده از این پیش فرض‌ها نسبت r و سرگشتگی (p) به صورت خودکار برای مجموعه داده در هر سطح از درخت تعیین می‌شوند. استفاده از تعداد بیشتری از لایه‌ها برای دیداری‌سازی درخت SNE ایده‌آل است چون یک طرح با تفکیک‌پذیری بالا ایجاد می‌کند. به عنوان مثال برای ۱۰۰ لایه $r = ۰/۹۵۵$ و برای ۳۰ لایه $r = ۰/۸۵۸$ خواهد بود.



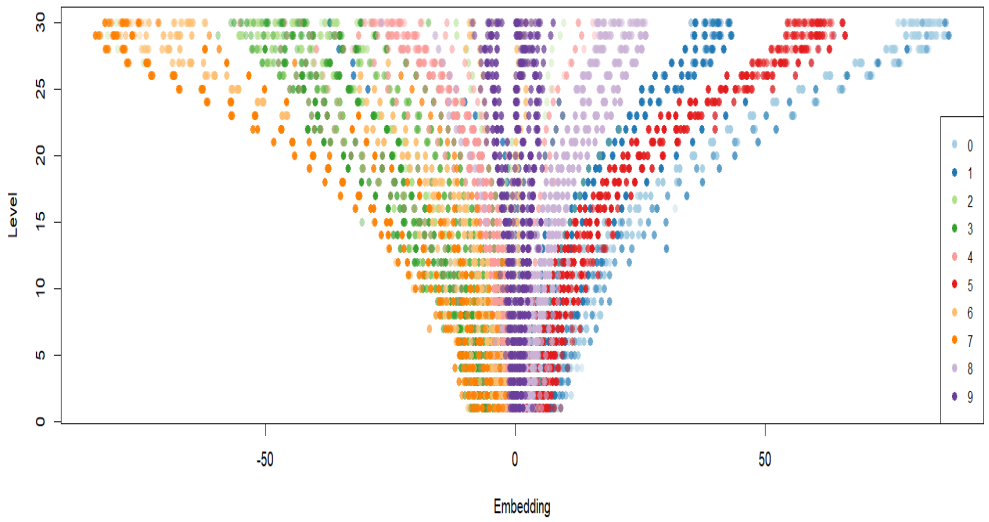
شکل ۱ : دیداری‌سازی درخت SNE از مجموعه داده MNIST که با برچسب‌های رقم واقعی رنگ‌آمیزی شده است.

مجموعه ارقام دست‌نویس هدی، مجموعه‌ی بزرگی از ارقام دست نویس فارسی مشتمل بر ۱۰۲۳۵۳ نمونه دست‌نوشته سیاه‌ـ سفید است. این مجموعه طی انجام یک پروژه کارشناسی ارشد در دانشگاه تربیت مدرس برای بازشناسی فرم‌های دست نویس تهیه شده است. داده‌های این مجموعه از حدود ۱۲۰۰۰ فرم ثبت نام آزمون سراسری کارشناسی ارشد در سال ۱۳۸۴ و آزمون کاردانی پیوسته دانشگاه جامع علمی کاربردی در سال ۱۳۸۳استخراج شده است. فرم‌های مورد پذیرش در این پروژه از طریق شرکت هوش مصنوعی هدی سیستم تهیه شده است. این مجموعه داده شامل ۶۰۰۰ نمونه آموزشی از هر کلاس (کلاس ارقام ۰ تا ۹) و ۲۰۰۰ نمونه آزمایشی از هر کلاس است. شکل‌های ۱ و ۲ به ترتیب کاربست درخت SNE روی مجموعه داده‌های MNIST و هدی را نشان می‌دهند.

بحث و نتیجه‌گیری

از آنجایی که تعیین سطح بهینه دانه‌بندی[۱] برای مشاهده‌ها یا خوشه‌بندی داده‌های بدون برچسب می‌تواند دشوار باشد، دیداری‌سازی و تحلیل داده‌ها با درخت SNE در مقابل t-SNE اجازه می‌دهد که یک محقق تصمیم بگیرد که کدام مقیاس

[۱] دانه‌بندی به تقسیم یک سیستم به ریزترین و جزئی‌ترین اجزای آن اطلاق می‌شود. گاهی می‌توان جزئیات سیستم یا مشخصات آن را نیز دانه‌بندی



شکل ۲ : دیداری‌سازی درخت SNE از مجموعه داده هدی که با برچسب‌های رقم واقعی رنگ‌آمیزی شده است.

برای خوشه‌بندی و نمایش تصویری داده‌های او بهتر است. این کار با نشان دادن یک دیدگاه عمیق نسبت به سازماندهی سلسله مراتبی داده‌ها در مقایسه با t-SNE صورت می‌گیرد. در t-SNE استاندارد، تلاش برای ارزیابی داده‌ها در چند مقیاس با استفاده از تنظیمات مختلف سرگشتگی و یا درجات آزادی امکان‌پذیر است. با این حال نکته قابل تامل این است که اغلب t-SNE ساختار محلی را در نمایش بعد پایین حفظ کرده، اما معمولا ساختار کلی (سراسری) را حفظ نمی‌کند (کوباک و برنز، ۲۰۱۹). در عمل این بدان معنا است که می‌توان نمونه‌هایی را که در نمایش پایانی به یکدیگر نزدیک هستند، شبیه به یکدیگر تفسیر کرد، اما نمی‌توان به‌راحتی بیان نمود که کدام دسته از نمونه‌ها، شبیه به سایر خوشه‌های نمونه‌ها در داده‌های اصلی هستند. بنابراین نمی‌توان نمودار t-SNE را از یک مقیاس به مقیاس‌های متفاوت دیگری ربط داد، زیرا سازماندهی کلی داده‌ها لزوما حفظ نمی‌شوند. در نتیجه، همانطور که رابنسون و پیرس ـ هافمن (۲۰۲۰) اشاره کردند، استفاده از یک رویکرد دیداری‌سازی سلسله مراتبی داده‌ها برای توانایی مقایسه بین مقیاس‌ها ضروری است.

مراجع

Bellman, R. (1957), Dynamic Programming, Princeton Univ, *Press Princeton, New Jersey*.

Hinton, G. E. and Roweis, S. (2002), Stochastic Neighbor Embedding, *Advances in Neural Information Processing Systems*, **15**, 857–864.

Hobbs, J. R. (1990), Granularity, In *Readings in Qualitative Reasoning about Physical Systems*, Elsevier, pp. 542–545.

Kobak, D. and Berens, P. (2019), The Art of Using t-SNE for Single-cell Transcriptomics, *Nature Communications*, **10**, 1–14.

Kobak, D., Linderman, G., Steinerberger, S., Kluger, Y., and Berens, P. (2019), Heavy-tailed Kernels

کرد. هابز و جری (۱۹۹۰) منبع مفیدی برای مطالعه‌ی بیشتر راجع به این موضوع است.

Reveal a Finer Cluster Structure in t-SNE Visualisations, In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 124–139.

Kullback, S. (1997), *Information theory and statistics*, Courier Corporation.

Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., and Kluger, Y. (2019), Fast Interpolation-based t-SNE for Improved Visualization of Single-cell RNA-seq Data, *Nature Methods*, **16**, 243–245.

Oskolkov, N. (2019), user, *Towards Data Science, Online: https://towardsdatascience. com/h ow-to-tune-hyperparameters-of-tsne-7c0596a18868*.

Robinson, I. and Pierce-Hoffman, E. (2020), Tree-sne: Hierarchical Clustering and Visualization Using t-sne, *ArXiv Preprint ArXiv:2002.05687*.

Van der Maaten, L. and Hinton, G. (2008), Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

بررسی پراکنش مکانی ذخیره کربن و نیتروژن خاک در جنگل‌های میانبند مناطق معتدله

علی بالویی، سید محمد حجتی، حامد اسدی، مریم اسدیان [۱]
علوم و مهندسی جنگل، دانشکده منابع طبیعی، دانشگاه علوم کشاورزی و منابع طبیعی ساری

چکیده: پژوهش حاضر در مناطق میان‌بند مازندران و در جنگل آموزشی و پژوهشی دارابکلا ـ ساری انجام گرفته است. در این پژوهش با پیاده‌سازی ۱۶۳ قطعه نمونه به روش سیستماتیک تصادفی و نمونه‌برداری از خاک در این قطعات نمونه، اندازه‌گیری و محاسبه مقدار ذخیره کربن و نیتروژن خاک انجام و سپس از واریوگرافی به منظور تعیین و تشریح ساختار مکانی این داده‌ها استفاده شده است. در ادامه با استفاده از کریگینگ، اقدام به درون‌یابی، برآورد زمین‌آماری و سپس تهیه نقشه‌های پهنه‌بندی از پراکنش ذخیره کربن و نیتروژن خاک در سطح عرصه مطالعه شده است. نتایج تحلیل تغییرنگار دو مشخصه ذخیره کربن و نیتروژن خاک نشان داده است که ذخیره کربن خاک با مدل نمایی، دامنه تاثیر متوسطی از خود نشان داده است (ساختار مکانی = ۵۰/۰۰ درصد). به عبارت دیگر این مشخصه از ساختار مکانی متوسطی برخوردار بوده و امکان ارائه نقشه پهنه‌بندی قابل استناد از آن وجود دارد. مشخصه ذخیره نیتروژن خاک نیز با استفاده از مدل کروی، ساختار مکانی متوسطی از خود نشان داده است (ساختار مکانی = ۳۰/۲۱ درصد). براساس نتایج این پژوهش، هر دو مشخصه ذخیره کربن و نیتروژن خاک، دارای ساختار مکانی می‌باشند. به عبارت دیگر، مقادیر ذخیره کربن و نیتروژن خاک، با داده‌های مکانی ارتباط داشته و با تغییرات مختصات، تغییر می‌کنند. نقشه‌های پهنه‌بندی تهیه شده از این دو عنصر، می‌تواند دید دقیق‌تری از پراکنش آن‌ها ارائه داده و همچنین به تفسیر روابط میان کربن و نیتروژن خاک، با سایر عناصر، تغییرات ارتفاع و پوشش گیاهی کمک شایانی نماید.

واژه‌های کلیدی: جنگل، کربن و نیتروژن خاک، زمین آمار، نقشه پهنه بندی
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 62H11، 62G08.

---

۱ مقدمه

کاهش مقدار دی اکسید کربن اتمسفری، از جمله مهم‌ترین راه کارها جهت مقابله با تغییر اقلیم به‌حساب می‌آید. با اینکه پوسته زمین، اقیانوس و هیدرات‌های گازی مخازن بسیار بزرگتری از خاک برای کربن درنظر گرفته می‌شوند، اما انسان به سادگی قادر به دستکاری آن‌ها با هدف افزایش ذخیره کربن نیست. یکی از بهترین راه حل‌ها جهت کاهش غلظت این گاز در

[۱] نام و ایمیل ارائه دهنده مقاله: مریم اسدیان maryam.asadiyan23@gmail.com

اتمسفر، افزایش میزان ترسیب کربن در خاک با استفاده از سیستم گیاه — خاک است (میر و همکاران ، ۲۰۲۳). بهترین بستر در این سیستم، جنگل‌ها هستند. بر طبق آمار پن و همکاران (۲۰۱۱) جنگل‌های جهان تقریبا ۸۶۲ گیگاتن کربن در خود ذخیره کرده‌اند (۴۴ درصد در یک متر اول خاک، ۴۲ درصد در زی‌توده رو و زیرزمینی، هشت درصد در بقایای مرده و پنج درصد در لاشبرگ‌ها). با توجه به این آمار، خاک بیشترین ظرفیت ذخیره کربن در یک بوم‌سامانه جنگلی را به خود اختصاص می‌دهد. با توجه به رابطه موجود بین کربن و نیتروژن، می‌توان گفت که نیتروژن به صورت غیرمستقیم بر ذخیره کربن در زی‌توده جنگل تاثیر گذار است. بنابراین وجود نیتروژن کافی (بعنوان اصلی‌ترین جزء این نسبت) برای افزایش ترسیب کربن خاک جنگلی ضروری است. مقدار ذخیره نیتروژن خاک به مقدار ورودی، سرعت تغییر شکل و مقدار خروجی آن وابسته است (اسدیان و همکارن ، ۱۴۰۲). این عوامل نیز تحت تاثیر بافت خاک، رطوبت و دمای خاک ، ارتفاع از سطح دریا، نوع پوشش گیاهی و دمای محیط می‌باشند(سردار و همکاران ، ۲۰۲۳). تغییر ارتفاع، با تاثیر بر نوع پوشش گیاهی مستقر، به صورت غیرمستقیم ویژگی‌های فیزیکی و شیمیایی خاک را نیز تحت تاثیر قرار می‌دهد (لنکا و همکاران ، ۲۰۱۳). طبق پژوهش آتشگاهی و همکاران (۱۳۸۸) به‌دنبال تغییرات ارتفاع از سطح دریا، تغییر در آب‌وهوا رخ می‌دهد که با تاثیر بر فرآیندهای شیمیایی، فیزیکی و زیستی خاک و همچنین با تاثیر بر پراکنش گونه‌های گیاهی، خصوصیات خاک را نیز تحت تاثیر قرار خواهند داد. با توجه به نقش ارتفاع از سطح دریا و تغییرات جوامع و تیپ‌های گیاهی در ایجاد تغییرات در ذخایر کربن و نیتروژن خاک و عدم دسترسی به مناطق صعب‌العبور در ارتفاعات بالاتر، و عدم امکان نمونه‌برداري از تمامي نقاط، استفاده از راهکاری مناسب جهت تعمیم نتایج حاصل از نقاط اندازه‌گیري شده به سایر نقاط ضروری می‌باشد. برای رسیدن به این هدف استفاده از علم زمین آمار[1] بسیار راهگشا است. استفاده از این تکنیک برای تهیه نقشه‌های دقیق پراکنش مکانی کربن و نیتروژن بسیار مفید است و می‌تواند اطلاعات دقیقی در خصوص روابط بین این عناصر و همبستگی آن‌ها با سایر پارامترها و درنهایت دیدی کلی از فاکتورهای متنوع و روابط بین آن‌ها برای اتخاذ تصمیمات درست برای احیاء پوشش گیاهی فراهم آورد. به همین منظور مطالعه حاضر با هدف بررسی الگوی پراکنش مکانی ذخیره کربن و نیتروژن خاک در جنگل آموزشی– پژوهشی دارابکلا ساری انجام شده است.

۲ مواد و روش

این پژوهش در سری یک جنگل دارابکلا به مساحت ۲۶۱۲ هکتار انجام گرفت. جنگل آموزشی و پژوهشی دارابکلا، تحت مدیریت دانشگاه علوم کشاورزی و منابع طبیعی ساری در محدوده‌ی طول جغرافیایی ′۱۴ °۵۲ تا ′۳۱ °۵۲ و عرض جغرافیایی ′۲۸ °۳۶ تا ′۳۳ °۳۶ قرار گرفته است. کمینه ارتفاع در این ناحیه ۱۸۷ و بیشینه ارتفاع ۸۷۸ متر بالای سطح دریا است. جنگل دارابکلا دارای هفت جامعه گیاهی طبیعی انجیلی — ممرزستان، بلوط، راش، تاج‌ریزی جنگلی ـ راشستان، ممرز، لرگ و فرفیون جنگلی ـ راشستان و یک توده جنگل‌کاری افرا به مساحت ۵۴ هکتار است. در این تحقیق تمامی توده‌های طبیعی و دست‌کاشت شناسایی شده توسط اسدی و همکاران (۱۴۰۰) مد نظر قرار گرفت. به منظور بررسی تاثیر جوامع گیاهی، شبکه آماربرداری به صورت تصادفی سیستماتیک و ابعاد شبکه ۴۰۰ در ۴۰۰ متر پیاده‌سازی شده است و بعد از حذف نقاط حاشیه‌ای، ۱۶۳ قطعه نمونه به ابعاد ۴۰۰ مترمربع (۲۰ × ۲۰ متر) درنظر گرفته شد. همچنین، برای درنظر گرفتن تاثیر ارتفاع بر ذخیره کربن و نیتروژن خاک، منطقه مورد مطالعه به سه ناحیه ارتفاعی کمتر از ۴۰۰ متر (ناحیه اول)، بین ۴۰۰ تا ۶۰۰ متر (ناحیه دوم) و بیشتر از ۶۰۰ متر بالای سطح دریا (ناحیه سوم) تقسیم شده است. جهت

[1] Geostatistics
[2] Walkey-Black
[3] Kjeldahl
[4] Clod

نمونه‌برداری از خاک در مرکز هر قطعه نمونه، یک نمونه از عمق صفر تا ۱۰ سانتی‌متری با استفاده از بیلچه برداشت شد. نمونه‌ها پس از برداشت، درون کیسه پلاستیکی به آزمایشگاه منتقل شدند. سپس بعد از اندازه گیری کربن (به روش والکی و بلاک[۲])، نیتروزن خاک (به روش کجلدال[۳]) و چگالی ظاهری (به روش پارافین[۴]) ذخیره کربن و نیتروژن خاک محاسبه شد (رازاکاماناریوو و همکاران ، ۲۰۱۱).

نرمال بودن داده‌ها با استفاده از آزمون شاپیرو–ویلک[۵] و همگنی واریانس با آزمون لون[۶] مورد بررسی قرار گرفت. سپس به‌منظور مقایسه ذخیره کربن و نیتروژن خاک بین جوامع گیاهی مختلف و همچنین بین نواحی ارتفاعی از تجریه واریانس یک طرفه و از نرم‌افزار Studio R ورژن (۴.۱.۲) استفاده شد (تیم توسعه‌دهنده آر ، ۲۰۲۱). برای مقایسه میانگین از آزمون HSD Tukey و با استفاده از بسته نرم‌افزار agricolae انجام شد (چمبرز و همکاران ، ۱۹۹۲؛ روی‌استون ، ۱۹۹۵؛ جیسون و چپ‌من هال ، ۱۹۹۶؛ هیل و گاچ ، ۱۹۸۰).

در بررسی‌های آماری، نمونه‌هایی که از کل جامعه به منظور شناخت مشخصه‌های موردنظر برداشت می‌شوند، فاقد داده‌های موقعیت مکانی هستند. درحالی که در زمین‌آمار، افزون بر مقدار یک کمیت معین در یک نمونه، موقعیت مکانی نمونه نیز مورد توجه قرار می‌گیرد. در زمین‌آمار با کاربرد داده‌های یک کمیت در مختصات معلوم، می‌توان مقدار همان کمیت را در نقطه‌ای با مختصات معلوم دیگر، واقع در دامنه‌ای که ساختار مکانی حاکم است تخمین زد (پاتریک و همکاران ، ۲۰۲۳).

از تغییرنگار[۷] به منظور تعیین و تشریح ساختار مکانی داده‌ها استفاده می‌شود. اجزای تغییرنگار عبارتند از دامنه[۸] تاثیر ، حد آستانه[۹] یا سقف و اثر قطعه‌ای[۱۰] می باشند. دامنه تاثیر، بیشینه فاصله‌ای است که پس از آن ساختار مکانی دیگر وجود ندارد و تغییرنگار ثابت می‌شود و افزایش فاصله تاثیری در تغییر مقدار تغییرنگار ندارد. به عبارت دیگر، وقتی تغییرنگار به مقدار ثابتی می‌رسد، ارتفاع تغییرنگار برابر حد آستانه یا سقف تغییرنگار، یعنی برابر مجموع واریانس تصادفی و ساختاردار است. اغلب در عمل، عرض تغییرنگار از مبداء به گونه‌ای است که اثر قطعه‌ای نامیده می‌شود که بیانگر واریانس تصادفی و بدون ساختار است. (کومار و سینا ، ۲۰۱۸).

واریوگرافی[۱۱] اولین قدم برای مدل‌سازی ساختار مکانی به منظور استفاده در کریگینگ[۱۲] است. نسبت واریانس ساختاردار به حد آستانه، ساختار مکانی تغییرنگار نامیده می‌شود. اولین قدم در درون‌یابی[۱۳] کریگینگ، برازش مدلی بر تغییرنگار تجربی است. کریگینگ، روش درون‌یابی و برآورد زمین‌آماری است که قادر است بر اساس مدل برازش شده بر تغییرنگار تجربی و نمونه‌های اندازه‌گیری شده در جامعه، نقاط نمونه‌برداری نشده را بدون اریب و با کمینه واریانس برآورد کند (کومار و سینا ، ۲۰۱۸).

۳ ارزیابی و نتایج

در مطالعات زمین‌آماری بایستی صحت تمام فرضیات و روش‌ها به گونه‌ای کنترل شود. کنترل اعتبار درواقع تخمین هر نقطه نمونه‌برداری شده در یک ناحیه با استفاده از مقادیر نمونه همسایه (بدون درنظر گرفتن خود آن نمونه) با روش‌های درون‌یابی می‌باشد. بدین منظور بعد از برازش مدل به تغییر نما و تعیین پارامترهای مدل، کنترل اعتبار تغییر نما به همراه نمودارهای

[5] Shapiro-Wilk
[6] Levenne
[7] Variogram
[8] Range
[9] Nugget
[10] Sill
[11] Variography
[12] Kriging
[13] Interpolation

تخمین برای متغیرهای مورد بررسی با استفاده از روش ارزیابی متقاطع و با درنظر گرفتن دو پارامتر آماری میانگین انحراف معیار خطا[۱۴] (MBE) و میانگین مطلق خطا[۱۵] (MAE) در نرم‌افزار +GS صورت گرفت. طبق نتایج، میان مقادیر ذخیره نیتروژن خاک، در تیمارهای جوامع گیاهی اختلاف معنی‌داری در سطح اطمینان ۹۹ درصد دیده شده است. حال آنکه تفاوتی میان ذخیره کربن خاک دیده نشده است . بیشترین مقدار ذخیره نیتروژن در جامعه فرفیون جنگلی_ راشستان و کمترین مقدار آن در جامعه تاج‌ریزی جنگلی_ راشستان مشاهده شد (جدول ۱).

جدول ۱: مقایسه میانگین ± اشتباه معیار ذخیره نیتروژن خاک در جوامع گیاهی

نوع مشخصه	ذخیره نیتروژن (تن در هکتار)
جنگل‌کاری افرا	۸۷/۲۴ ± ۳/۵۳^{a}
انجیلی_ ممرزستان	۹۶/۰۷ ± ۵/۴۸^{a}
زیرجامعه تیپیک بلوط	۸۴/۵۲ ± ۵/۸۰^{a}
زیرجامعه راش	۶۹/۴۱ ± ۴/۲۸^{ab}
تاج‌ریزی جنگلی_ راشستان	۵۹/۷۰ ± ۲/۶۱^{b}
زیرجامعه ممرز	۹۶/۸۱ ± ۱۴/۵۰^{a}
زیرجامعه لرگ	۹۵/۳۳ ± ۷/۹۴^{a}
فرفیون جنگلی_ راشستان	۷۵/۴۹ ± ۴/۰۷^{a}

حرف مختلف نشان‌دهنده تفاوت بین جوامع گیاهی در سطح اطمینان ۹۵ درصد روش توکی (HSD) می‌باشند

نتایج نشان داد که اختلاف معنی‌دار آماری در سطح اطمینان ۹۹ درصد برای ذخیره کربن و نیتروژن خاک میان تیمارهای ارتفاعی وجود دارد. همچنین هردو مشخصه کربن و نیتروژن خاک با تغییرات ارتفاع رابطه معکوس دارند. به طوریکه با افزایش ارتفاع از سطح دریا، مقادیر ذخیره کربن و نیتروژن خاک کاهش می‌یابد (جدول ۲).

جدول ۲: مقایسه میانگین ± اشتباه معیار ذخیره کربن و نیتروژن خاک در نواحی ارتفاعی

	ناحیه اول (< ۴۰۰)	ناحیه دوم (۴۰۰ ـ ۶۰۰)	ناحیه سوم (> ۶۰۰)
ذخیره کربن (تن در هکتار)	۷۴/۱۸ ± ۳/۳۲^{a}	۷۰/۵۶ ± ۳/۲۹^{b}	۴۹/۵۴ ± ۱/۸۸^{b}
ذخیره نیتروژن (تن در هکتار)	۹۴/۱۹ ± ۳/۵۱^{a}	۷۵/۴۵ ± ۲/۳۰^{b}	۶۴/۷۱ ± ۳/۴۲^{c}

حروف مختلف نشان‌دهنده تفاوت بین نواحی در سطح اطمینان ۹۵ درصد روش توکی (HSD) می‌باشند

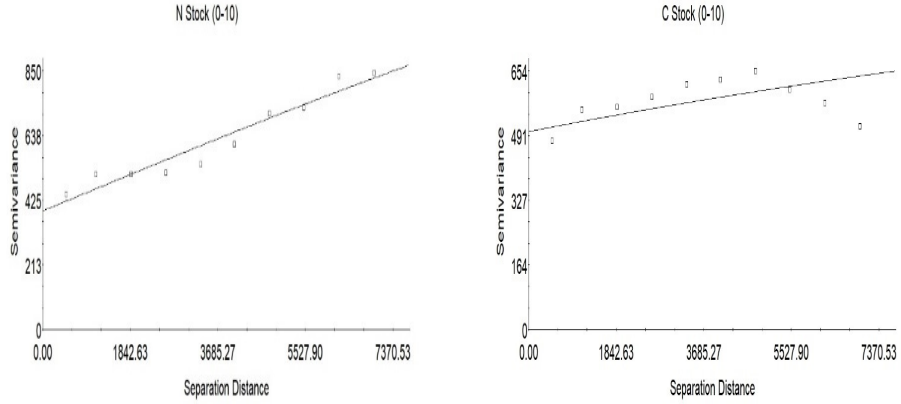
در پژوهش حاضر برای تشخیص پدیده همسانگردی[۱۶] از تغییرنگار سطحی استفاده شده است. در این مطالعه ناهمسانگردی[۱۷] برای تمامی متغیرهای مورد بررسی کنترل شده است. با توجه به تقارن تغییرنگار سطحی، تمامی متغیرها همسانگرد هستند. مدل‌های تغییرنگار تجربی به همراه مدل‌های برازش داده شده به آن‌ها و پارامترهای اعتبارسنجی شده آن‌ها همراه با معیارهای اعتبارسنجی برای مشخصه های مورد بررسی در پژوهش حاضر در ذیل ارائه شده است. متداول‌ترین مدل‌هایی که بیشترین کاربرد را در مطالعات محیط زیستی دارند، مدل‌های نمایی و کروی هستند. طبق نتایج بدست آمده مقدار گزارش شده برای مشخصه دامنه تاثیر، بیش ترین فاصله ای است که پس از آن ساختار مکانی دیگر وجود نخواهد داشت و تغییر نگار ثابت می شود. همچنین مقدار عددی گزارش شده برای مشخصه سقف تغییر نگار یا حد آستانه مقداری که در آن تغییر نگار به مقدار ثابتی خواهد رسید. با توجه به نتایج حاصل از واریوگرافی اگر ساختار مکانی ۷۵ درصد و یا بیشتر باشد، نشان دهنده ساختار قوی، بین ۲۵ تا ۷۵ درصد ساختار متوسط و کمتر از ۲۵ درصد نشان دهنده ساختار ضعیف برای متغیر مورد بررسی است. مشخصه‌ی ذخیره کربن خاک با مدل نمایی دامنه تاثیر متوسطی داشته است. همچنین مشخصه ذخیره نیتروژن خاک نیز با مدل کروی ساختار مکانی متوسطی از خود نشان داده است. مقادیر MBE و MAE

[14] Mean Bias Error
[15] Mean Absolute Error
[16] Isotropic
[17] Anisotropic

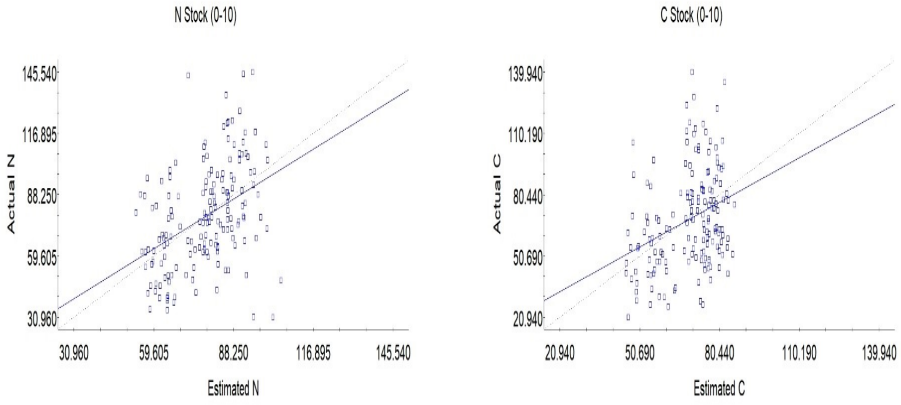
میزان اریبی را نشان می‌دهند و در حالت ایده‌آل باید مساوی صفر باشند. مقادیر مثبت و منفی آن‌ها به ترتیب نشان دهنده برآورد بیشتر یا کمتر از مقادیر واقعی‌اند. با توجه به نتایج بدست آمده مشاهده شد که مقدار خطا و اریبی برآوردها زیاد نبوده و کریگینگ توانسته است براساس مدل‌های انتخابی، برآوردهای نسبتا دقیقی داشته باشد (جدول ۲). نتایج تغییرنگار، برازش مدل و نقشه‌های پهنه‌بندی مشخصه‌های کربن و نیتروژن خاک به ترتیب در اشکال ۱ و ۲ نشان داده شده است.

جدول ۳: تغییر نما، انتخاب مدل و کنترل کریگینگ برای ذخیره کربن و نیتروژن خاک

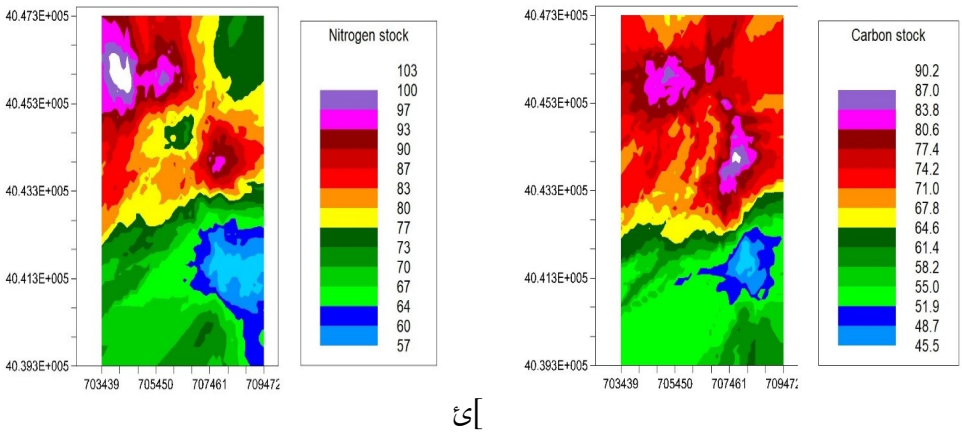
متغیر	مدل	اثر قطعه‌ای	سقف تغییر نگار (حد آستانه)	دامنه تاثیر (متر)	ساختار مکانی (درصد)	میانگین انحراف خطا (MBE)	میانگین مطلق خطا (MAE)
ذخیره نیتروژن	کروی	۳۹۰/۰۰	۱۲۹/۹۰	۲۰۷۳۰/۰۰	۳۰/۲۱	−۰/۰۹	۱۷/۳۳
ذخیره کربن	نمایی	۵۰۱/۰۰	۱۰۰۲/۱۰	۲۱۱۰۰/۰۰	۵۰/۰۰	−۰/۶۴	۱۸/۳۵



شکل ۱: نتایج تغییرنگار مربوط به ذخیره کربن و نیتروژن خاک



شکل ۲: نتایج برازش مقادیر برآورد شده و مقادیر واقعی مربوط به ذخیره کربن و نیتروژن خاک



]ئ

شکل ۳: نقشه پهنه‌بندی ذخیره کربن و نیتروژن خاک

بحث و نتیجه‌گیری

در پژوهش حاضر مشخصه ذخیره کربن با مدل نمایی و ذخیره نیتروژن با مدل کروی ساختار متوسطی دارند. در نقشه پهنه‌بندی ذخیره نیتروژن ناحیه شمالی محل پراکنش زیرجامعه راش و ناحیه جنوبی محل پراکنش جامعه تاج‌ریزی جنگلی‌ـ راشستان است. نواحی محل پراکنش جامعه انجیلی‌ـممرزستان و زیرجامعه ممرز بیشترین ذخیره نیتروژن خاک را دارند (بالویی و همکاران ، ۱۴۰۲). مطابق نقشه، ذخیره کربن خاک از شمال به جنوب نقشه روند کاهشی به خود گرفته است. این موضوع با تاثیر منفی ارتفاع بر ذخیره کربن که در جدول (۲) نشان داده شد همخوانی دارد. با این حال، نتایج این پژوهش نشان داده که جوامع مورد مطالعه تفاوتی به لحاظ ذخیره کربن خاک با یکدیگر ندارند. این پژوهش نشان داد که استفاده از تکنیک زمین آمار در تهیه نقشه‌های پهنه‌بندی از مشخصه‌های خاک می‌تواند مفید واقع شود. نقشه‌های پهنه‌بندی تهیه شده، با توجه به موقعیت مکانی هریک از جوامع گیاهی در جنگل دارابکلا و میانگین مشخصه‌های مورد مطالعه برای هر جامعه، گواه این ادعا است که نقشه‌های پهنه‌بندی با اندازه‌گیری‌های انجام شده انطباق خوبی دارند. در نهایت باید گفت که با توجه به اهمیت عناصر کربن و نیتروژن در بوم‌سامانه جنگل، شناخت هرچه بیشتر الگوی پراکنش مقداری و مکانی این دو عنصر و عوامل تاثیرگذار بر آن ها می‌تواند نقش بسزایی در اتخاذ تصمیمات اصولی در مدیریت جنگل داشته باشد.

مراجع

اسدی، ح.، جلیلوند، ح.، و مسلمی سید محله، س. م.، (۱۴۰۰)، *طبقه‌بندی جوامع گیاهی و ارتباط آنها با عوامل فیزیوگرافیک در جنگل دارابکلای استان مازندران. مجله علمی پژوهشی اکولوژی کاربردی*، **۱۰**(۳): ۱۷ـ۳۳.

اسدیان، م.، حجتی، س.م، و محمدزاده، م.، نادی، م.، (۱۴۰۲)، ارزیابی پاسخ بوم سازگان به تغییر کاربری زمین با استفاده از شاخص کیفیت خاکـ جنگل الندان، *مجله جنگل ایران*، **۱۵**(۱): ۱۷ـ۳۴.

آتشگاهی، ز.، اجتهادی، ح.، و زارع، ح.، (۱۳۸۸)، معرفی فلور، شکل زیستی و پراکنش جغرافیایی گیاهان در جنگل های شرق دودانگه ساری استان مازندران، *مجله زیست شناسی ایران*، **۲۰۳**: ۲۲ـ۱۹۳.

بالویی، ع.، حجتی، س.م.، و اسدی، ح.، اسدیان، م.، (۱۴۰۲)، تاثیر جوامع گیاهی جنگل آموزشی – پژوهشی دارابکلا بر ذخیره کربن در خاک و زی توده رو زمینی، مجله پژوهش و توسعه جنگل، پذیرش چاپ.

Chambers, J. M., Freeny, A and Heiberger, R. M. (1992), *Analysis of Variance; Designed Experiments*, Chapter 5 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth and Brooks/Cole.

Hill, M.O. and Gauch, H.G. (1980), Detrended correspondence analysis: *An Improved Ordination Technique*, Vegetation **42**, 47–58.

Jason C. Hsu. Chapman Hall (1996) , *Multiple Comparisons Theory and Methods*, Department of statistics the Ohio State University, USA, 1996. CRC.

Kumar, N., Sinha, N. K. (2018) Geostatistics: *Principles and Application in Spatial Mapping of Soil Properties*, Geospatial Technologies in Land Resources Mapping, Monitoring and Management, 143-159.

Lenka, N. K., Sudhishri, S., Dass, A., Choudhury, P. R., Lenka, S., Patnaik, U. S. (2013), Soil Carbon Sequestration as Affected by Slope Aspect under Restoration Treatments of a Degraded Alfisol in the Indian Sub-Tropics, *Geoderma*, **204**, 102–110.

Mir, YH., Ganie, MA., Shah, TI., Aezum, AM., Bangroo, SA., Mir, SA., Dar, SR., Mahdi, SS., Baba, ZA., Shah, AM., Majeed, U., Minkina, T., Rajput, VD., Dar, AA.,(2023), *Soil Organic Carbon Pools and Carbon Management Index* under different land use systems in North western Himalayas, PeerJ **11**:e15266 https://doi.org/10.7717/peerj.15266

Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A., Phillips, O. L., Shvidenko, A., Lewis, S. L., Canadell, J. G. (2011), *A Large and Persistent Carbon Sink in the World's Forests*, Science, **333** (6045), 988–993.

R Core Team (2021). R: *A language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Razakamanarivo, R. H., Grinand, C., Razafindrakoto, M. A., Bernoux, M., Albrecht, A. (2011), Mapping Organic Carbon Stocks in Eucalyptus Plantations of the Central Highlands of Madagascar: *A Multiple Regression Approach*, Geoderma, **162** (3–4), 335–346.

Royston, P., (1995), Remark AS R94:*A Remark on Algorithm AS 181: The W test for normality*, Applied Statistics, **44**, 547–551, doi: 10.2307/2986146.

Sardar, MF., Younas, F., Farooqi, ZUR., YL, Li., (2023), Soil nitrogen dynamics in natural forest ecosystem: *A Review*, Front. For. Glob. Change 6:1144930. doi: 10.3389/ffgc.2023.1144930

Patriche, C. V., Rosca, B., Pirnau R. G., Vasiliniuc 1., (2023) *Spatial Modelling of Topsoil Properties* in Romania using geostatistical methods and machine learning. PLos ONE **18** (8): 30289286.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

تحلیل فضایی بیزی داده‌های بقای گسسته صفر آماسیده

سپیده اسعدی[۱]**، محسن محمدزاده**
گروه آمار، دانشگاه تربیت مدرس

چکیده: در این مقاله، به تحلیل داده‌های بقای گسسته صفر آماسیده شامل مشاهدات راست سانسوریده، که از یک ساختار همبسته فضایی نیز پیروی می‌کنند، با استفاده از توزیع وایبول گسسته پرداخته شده است. در مطالعات شبیه‌سازی عملکرد مدل بقای فضایی با اثرات تصادفی مورد ارزیابی و مقایسه عددی قرار می‌گیرد، سپس نشان داده خواهد شد که چگونه می‌توان با رهیافت بیزی مدل پیشنهادی را برای تحلیل داده‌های بقای گسسته صفر آماسیده فضایی مورد استفاده قرار داد.

واژه‌های کلیدی: داده‌های بقای گسسته صفر آماسیده، توزیع وایبول گسسته، داده‌های راست سانسوریده .
کد موضوع‌بندی ریاضی (۲۰۱۰): 62N02، 62H11، 62M30.

۱ مقدمه

متغیر اصلی در تحلیل داده‌های بقا، زمان بقا یا مدت پیگیری تا وقوع یک رویداد مورد علاقه برای یک واحد است. معمولاً این زمان پیوسته و با یک عدد حقیقی نا منفی اندازه‌گیری می‌شود. اما در عمل، گاهی با مواردی مواجه می‌شویم که داده‌های بقا به صورت بازه‌های زمانی گسسته گروه‌بندی شده یا در نقاط زمانی مجزا ثبت می‌شوند، در این صورت زمان بقا یک متغیر تصادفی گسسته است. علاوه براین، زمان بقا برای برخی از واحدهای جامعه کمتر از واحد زمان است، که در حالت گسسته، معمولاً مقدار آن صفر در نظر گرفته می‌شود. در این صورت مجموعه داده‌های بقا شامل تعداد زیادی صفر و اعداد طبیعی مثبت خواهد بود، که "داده‌های بقای گسسته صفر آماسیده" نامیده می‌شوند. یک مدل آماسیده در صفر وقتی به وجود می‌آید که جرم احتمال در نقطه صفر از حد مجاز در خانواده پارامتری استاندارد توزیع‌های گسسته فراتر رود. زمانی که مجموعه داده‌های شمارشی دارای فراوانی بیش از حد در عدد صفر باشند، مدل صفر آماسیده یک پارامتر احتمالی اضافی را برای مقدار صفر که نمی‌تواند به طورکامل توسط فرض مدل برآورد شود، معرفی می‌کند. از آنجایی که در تحلیل داده‌های بقای فضایی

[1] Zero Inflated Discrete Weibull

[۱] نام و ایمیل ارائه دهنده مقاله: سپیده اسعدی، sepideh.asadi@modares.ac.ir

وجود داده‌های سانسور شده در بین مشاهدات و همبستگی فضایی می‌توانند منجر به تغییر در میزان پراکندگی موجود در داده‌ها شود. در این مقاله از توزیع وایبول گسسته صفر آماسیده[۱] (ZIDW) که تعمیمی از توزیع محبوب وایبول پیوسته در مطالعات بقا است و توانایی بازتاب انواع پراکندگی در هر دو حالت داده‌های صفر و غیرصفر را دارد، به عنوان توزیع والد در مدل‌بندی داده‌های بقای صفر آماسیدهانتخاب شده است. از آنجا که مدل‌های آمیخته صفر آماسیده دارای دو مؤلفه هستند: یک مؤلفه دوحالتی که احتمال صفر شدگی یا احتمال شروع[۲] را مدل‌بندی می‌کند و یک مؤلفه شمارشی غیر صفر شدگی که تعداد موارد در یک جمعیت مستعد را مدل‌بندی می‌کند(نیاندوی و همکاران ، ۲۰۲۰)، بنابراین در مدل‌های شرطی فضایی دو مدل رگرسیونی تعمیم‌یافته خطی متناظر با این دو مؤلفه تعریف می‌شود و همبستگی فضایی برای هردو مؤلفه مدل با انتخاب دو اثر تصادفی با توزیعی از خانواده توزیع‌های اتورگرسیو شرطی لحاظ می‌گردد. نیلون و همکاران (۲۰۱۵) نشان دادند، که در مدل‌های فضایی صفر آماسیده برای داده‌های شمارشی همبسته در نظر گرفتن این دو اثر تصادفی با توزیع پیشین دو متغیره اتورگرسیو شرطی ذاتی BICAR نسبت به فرض استقلال آن‌ها با توزیع تک متغیره ICAR نتایج بهتری را در برآورد پارامترهای مدل شاهد خواهیم بود. ما در بخش ۲ این مقاله به تحلیل فضایی داده‌های بقای صفر آماسیده با مشاهدات راست سانسوریده با توزیع $CZIDW$، از طریق مدل‌های دو بخشی می‌پردازیم. سپس با انجام مطالعات شبیه‌سازی بخش ۳ کارایی مدل‌های فضایی با اثرات تصادفی همبسته ارزیابی می‌گردد، در نهایت در بخش ۴ به تحلیل مجموعه داده‌های طول مدت زندگی مشترک با انباشتگی بیش از حد انتظار در بازه صفر تا ۵ سال می‌پردازیم.

۲ تحلیل فضایی داده‌های بقای گسسته وایبول صفر آماسیده راست سانسوریده

فرض کنید A حوزه فضایی مورد نظر ما است که به یک زیرمجموعه ثابت از نواحی با اشکال منظم یا نامنظم $A_1, \ldots, A_L$ با مرزهای واضح افراز شده است و $n_1, \ldots, n_L$، به ترتیب تعداد واحدهای هر ناحیه است و در مجموع یک نمونه تصادفی به حجم n از داده‌های بقا موجود باشد، اگر $X \sim ZIDW(\pi(Z), q_{i\ell}(X), \beta)$ زمان بقای دقیق واحد iام در ناحیه lام، یک متغیر تصادفی نامنفی شمارشی با انباشتگی بیش از حد انتظار صفر باشد، که از توزیع گسسته وایبول صفر آماسیده پیروی می‌کند و $X = (1, X_1, \ldots, X_{p_1})$ و $Z = (1, Z_1, \ldots, Z_{p_2})$ به ترتیب ماتریس طرح $(P_1 + 1 \times n)$ بعدی شامل P_1 بردار $(x_1, \ldots, x_{P_1})$ متغیرهای تبیینی و $(P_2 + 1 \times n)$ بعدی شامل P_2 بردار $(z_1, \ldots, z_{P_2})$ باشند بطوری که $x_{p1i\ell}$ مقدار متغیر کمکی $p = 1, \ldots, P$ برای واحد i ام در ناحیه ℓ باشد، تابع توزیع آمیخته دو جزیی با پارامتر آمیختگی[۳] و مدل شمارشی والد[۴] $f_T(t)$ بصورت

$$f_T(t_{i\ell}|x_{i\ell}) = \pi(z_{i\ell})I_{(t_{i\ell}|z_{i\ell}=0)} + (1 - \pi(z_{i\ell}))f_1(t_{i\ell}|x_{i\ell}), \quad i = 1, \ldots, n_\ell, \quad \ell = 1, \ldots, L, \quad (۱.۲)$$

است. فرض کنید توزیع والد $f_1(t)$، توزیع گسسته وایبول با تابع جرم احتمال

$$f_1(t_{i\ell} \mid x_{i\ell}) = q(X_{i\ell})^{t_{i\ell}^{\beta}} - q(X_{i\ell})^{(t_{i\ell}+1)^{\beta}}, t_{i\ell} = 1, 2, \ldots \quad (۲.۲)$$

باشد، که در آن $0 < q(x_{i\ell}) < 1$ پارامتر شکل و $\beta > 0$ پارامتر کنترل‌کننده چولگی توزیع است. در واقع در این مدل فرض می‌شود که مشاهده iام، در ناحیه ℓام، توسط دو فرآیند با احتمالات متفاوت حاصل می‌شود، فرآیند اول فقط صفرهایی با احتمال $\pi_{i\ell}$ تولید می‌کند، در حالی که فرآیند دوم، مقادیر شمارشی را از یک توزیع گسسته وایبول با احتمال $(1 - \pi_{i\ell})$ تولید می‌کند. بنابراین مدل‌های رگرسیونی تعمیم‌یافته خطی در داده‌های صفر آماسیده، دوبخشی است که بخش اول ارتباط

[2] Probability of onset
[3] Mixing parameter
[4] Parent count model

متغیرهای تبیینی $z_{i\ell}$را با احتمال صفر بودن (پارامتر آمیختگی) از طریق یک تابع پیوند لوجیت مدل‌بندی می‌کند و بخش دوم به مدل‌بندی شمارش غیر صفر و متغیرهای تبیینی $x_{i\ell}$ از طریق تابع پیوند مکمل لگاریتم لگاریتم "C-L-L"[۵] پارامتر q می‌پردازد. بنابراین در مدل‌بندی فضایی این قسم از داده‌ها دو رویکرد عمده وجود دارد. در رویکرد اول فرض می‌شود همبستگی فضایی متغیر پاسخ در هر دو مؤلفه توسط دو اثر تصادفی فضایی ساختار یافته فضایی $\phi_1 = (\phi_{11}, \ldots, \phi_{1L})$ و $\phi_2 = (\phi_{21}, \ldots, \phi_{2L})$ که از هم مستقل هستند، محاسبه می‌شود. بنابراین ما دو فرآیند فضایی مجزا در مدل داریم و اجزای مدل را می‌توان با برازش دو مدل رگرسیونی تعیم‌یافته خطی جداگانه ($\phi_1 \perp \phi_2$) به صورت

$$\log(-\log(q(x_{i\ell}))) = \mathbf{x'_{i\ell}}\alpha_{i\ell} + \phi_{2\ell}, \quad \Rightarrow \mathbf{q} \equiv \mathbf{q(X)} = \mathbf{e^{-e^{X'\alpha+\phi_2}}}, \tag{۳.۲}$$

$$\text{logit}(\pi_{i\ell}) = \mathbf{x'_{i\ell}}\gamma_{i\ell} + \phi_{1\ell}, \quad \Rightarrow \pi \equiv \frac{\mathbf{e^{X'\gamma+\phi_1}}}{\mathbf{1+e^{X'\gamma+\phi_1}}} = \mathbf{(1+e^{-X'\gamma+\phi_1})^{-1}}, \tag{۴.۲}$$

برآورد نمود، که در آن $\gamma_{i\ell} = (\gamma_{0i\ell}, \ldots, \gamma_{pi\ell})$ و $\alpha_{i\ell} = (\alpha_{0,i\ell}, \ldots, \alpha_{pi\ell})$ بردار ضرایب رگرسیونی متناظر با این دو مدل رگرسیونی هستند، برآورد نمود. برای درج وابستگی فضایی در قالب مدل‌های شرطی ساختاریافته در مدل ZIDW، دو رویکرد خواهیم داشت. در این رویکرد ϕ_1 و ϕ_2 بطور جداگانه از توزیع پیشین CAR یا مدل اتورگرسیو شرطی ذاتی ICAR (بسیج و همکاران ، ۱۹۹۱) پیروی می‌کنند، که در آن نحوه ورود همبستگی فضایی در هنگام تحلیل داده‌های ناحیه‌ای از طریق ماتریس همسایگی یا مجاورت در مدل منظور می‌شود. رویکرد دوم حالتی است که بین مؤلفه‌های دو بردار اثر تصادفی ϕ_1، ϕ_2 همبستگی فضایی وجود دارد. برای تطبیق با این ارتباط بالقوه با توزیع پیشین ICAR فرض می‌کنیم که $\phi_\ell = (\phi_{1\ell}, \phi_{2\ell})$ از یک توزیع اتورگرسیو شرطی دومتغیره (BICAR) بصورت

$$\boldsymbol{\phi}_\ell \mid \boldsymbol{\phi}_{(-\ell)}, \boldsymbol{\Sigma} \sim \text{N}_2\Big(\frac{1}{m_\ell}\sum_{\ell\in\partial_\ell}\boldsymbol{\phi}_\ell, \frac{1}{m_\ell}\boldsymbol{\Sigma}\Big), \tag{۵.۲}$$

است، که در آن m_ℓ نشان دهنده تعداد همسایگان منطقه ℓ است، ∂_ℓ مجموعه همسایگان برای منطقه ℓ و $\boldsymbol{\Sigma}$ یک ماتریس کوواریانس 2×2 ϕ_ℓ مشروط به سایر اثرات تصادفی فضایی، $\boldsymbol{\phi}_{(-\ell)}$ است. برای تحلیل بقای فضاییِ صفر آماسیده با مشاهدات راست سانسوریده CZIDW، فرض کنید، $C_{i\ell}$ زمان سانسوریدگی باشد که از $T_{i\ell}$ مستقل است، در این صورت برای یک واحد سانسورشده از راست تنها اطلاعات در دسترس این است که $T_{i\ell} > C_{i\ell}$. با تعریف متغیر تصادفی $Y_{i\ell} = \min(T_{i\ell}, C_{i\ell})$ و دو متغیر تصادفی نشانه‌ای $\delta_{i\ell} = 1_{y_{i\ell}\geq Ci\ell}$ که شکست واحدها به علت پیشامد نهایی را نشان می‌دهد و $J_{i\ell} = 1_{y_{i\ell}>0}$ تابع درستنمایی مدل CZIDW برای داده‌های گسسته وایبول صفر آماسیده با مشاهدات راست سانسوریده به صورت

$$\begin{aligned} L_{Area} \;=\; & \prod_{i=1}^{n}\Big\{\left[\pi_i + (1-\pi_i)(1-q_i)\right]^{J_i}\left[(1-\pi_i)\left(q_i^{y_i^\beta} - q_i^{(y_i+1)^\beta}\right)\right]^{1-J_i}\Big\}^{1-\delta_i} \\ & \times\Big\{1 - \left[\pi_i + (1-\pi_i)(1-q_i^{C^\beta})\right]\Big\}^{\delta_i} \end{aligned}$$

خواهد بود، که در آن $q_{i\ell}$ و $\pi_{i\ell}$ به ترتیب در رابطه (۴.۲) و (۳.۲) تعریف شده‌اند. برای بدست آوردن برآورد مؤلفه‌های بردار پارامترهای مدل $\theta = (\alpha, \gamma, \beta, \Phi = (\phi_1, \phi_2), \Sigma)$ با فضای پارامتر Θ ما یک رویکرد مدل‌بندی بیز سلسله مراتبی را در پیش می‌گیریم، ما یک توزیع پیشین دو متغیره CAR (BICAR) را برای $\boldsymbol{\phi}_\ell = (\phi_{1\ell}, \phi_{2\ell})^T$ در رابطه (۵.۲) درنظر می‌گیریم. بدلیل اینکه دامنه $\alpha_j \in \Re$ و $\gamma_j \in \Re$ برای هر $j = 1, \ldots, p$ و $\beta \in \Re^+$ است، بترتیب توزیع پیشین نرمال با اَبَر پارامترهای $(\mu_{\alpha_j}, \sigma^2_{\alpha_j})$ و نرمال با اَبَر پارامترهای $(\mu_{\gamma_j}, \sigma^2_{\gamma_j})$ و معکوس گاما با اَبَر پارامترهای (a, b) می‌تواند یک انتخاب مناسب باشد. معمولاً یک انتخاب کلاسیک برای توزیع پیشین ماتریس واریانس و کواریانس Σ که همبستگی

[۵] Complementary Log-Log link function

بین اثرات تصادفی ϕ_1 و ϕ_2 را شامل می‌شود، توزیع مزدوج معکوس ویشارت $\text{IW}(v_0, \boldsymbol{S}_0)$ با پارامترهای $\nu_0 = 3$ ، $S_0 = \text{diag}(2)$ است. از آنجایی که پیشین های مزدوج در دسترس نیستند، توزیع‌های پسین بیزی اغلب به صورت مدل‌هایی پیچیده با تعداد پارامترهای زیاد به دست می‌آیند. در اینجا از توزیع تمام شرطی هریک از پارامترهای مدل برای استفاده از الگوریتم قدم زدن تصادفی متروپلیس هستینگز به منظور برآورد بیزی پارامترهای $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \phi_1, \phi_2, \beta, \Sigma)$ استفاده می‌کنیم.

۳ مطالعات شبیه‌سازی

در این مطالعه شبیه سازی برای ارزیابی و مقایسه نتایج حاصل از برازش دو مدل همبسته فضایی با اثرات تصادفی مستقل و مدل با اثرات تصادفی همبسته فضایی در محیط نرم افزار R انجام شده است. شبیه‌سازی شامل $p = 2$ متغیر تبیینی، ماتریس طرح $n \times (p+1)$ بعدی $X = (1, X_1, \ldots, X_p)$، متناظر با p متغیر تبیینی، عرض از مبدأ و تعداد $n = \sum_{\ell=1}^{L} n_\ell$ مشاهده است. دو متغیر کمکی غیرمکانی $x_{i\ell 1} x_{i\ell 2}$ که به ترتیب از توزیع‌های نرمال $N(0, 1)$ و یکنواخت $U(0, 1/5)$ تولید شده‌اند، نیز در نظر گرفته شده است. دو پاسخ مرتبط با $T|X \sim ZIDW(\pi, q(X), \beta)$ تحت دو مدل خطی تعمیم‌یافته با توابع پیوند $\text{logit}(\pi)$ و $\log(-\log(q))$ کنترل می‌شوند. مقادیر اولیه $\alpha_{real} = (-2, 0/5, 0/3)$، $\beta = 1/2$ و $\gamma_{real} = (1, 1/5, -0/2)$ برای پارامترهای مدل درنظر گرفته شده است. زمان‌های سانسور $C_{i\ell}$ نیز از توزیع $U(0, 2)$ شبیه‌سازی شده‌اند. همچنین اگر زمان وقوع پیشامد $T_{i\ell}$ تولید شده بزرگتر از زمان سانسور تولید شده نباشد، $\delta_{i\ell} = 1$، در غیر این صورت مقدار آن را صفر در نظر می‌گیریم. برای اضافه کردن ویژگی آماسیدگی در صفر به هر پاسخ ابتدا یک بردار تصادفی از توزیع یکنواخت $U = (u_1, \ldots, u_n) \sim U(0, 1)$ به طول n تولید می‌کنیم، در صورتی که $u_{i\ell} \leq \pi_{i\ell}$ باشد، $J_{i\ell} = 0$ و $Y_{i\ell} = 0$ در غیر اینصورت $J_{i\ell} = 1$ و $Y_{i\ell} \sim DW$ در نظر می‌گیریم. برای شبیه‌سازی مدل اول، مدل‌های زیر را که پایه‌ای برای هر سه مدل دیگر است، در نظر می‌گیریم.

$$\text{logit}(\pi_{i\ell}) = \gamma_{0\ell} + \sum_{p=1}^{2} \gamma_{p\ell} x_{p\ell} + \phi_{1\ell}, \qquad \log(-\log(q_{i\ell})) = \alpha_{0\ell} + \sum_{p=1}^{2} \alpha_{p\ell} x_{p\ell} + \phi_{2\ell}. \qquad (1.3)$$

برای ضرایب رگرسیونی $\alpha_0, \alpha_1, \alpha_2$ توزیع‌های پیشین نرمال $N(0, \sigma^2_{\alpha_0}), N(0, \sigma^2_{\alpha_1}), N(0, \sigma^2_{\alpha_2})$ با پارامترهای دقت $\sigma^{-2}_{\alpha_p} \sim T(10^{-5}, 10^{-5})$، $p = 0, 1, 2$ و به طور مشابه برای $\gamma_0, \gamma_1, \gamma_2$ پیشین‌های نرمال $N(0, \sigma^2_{\gamma_0})$، $N(0, \sigma^2_{\gamma_1})$ و $N(0, \sigma^2_{\gamma_2})$ با پارامترهای دقت $\sigma^{-2}_{\gamma_p} \sim T(10^{-5}, 10^{-5})$، $p = 0, 1, 2$ توزیع پیشین گامای $InvG(0/5, 0/5)$ برای پارامتر شکل β در نظر گرفته شده است.ما از ماتریس مجاورت کشور ایران که بر مبنای فاصله اقلیدسی موقعیت جغرافیایی استانهای کشور است، استفاده کردیم. در ابتدا با توجه به آن‌ها را محاسبه نموده‌ایم.

دو اثرات تصادفی فضایی سطح ناحیه جداگانه $\phi_{1\ell}$ و $\phi_{2\ell}$ به مدل‌های (۱.۳) اضافه می‌کنیم. دو توزیع ICAR تک متغیره مستقل برای $\phi_{1\ell}$ و $\phi_{2\ell}$ اختصاص می‌دهیم. برای شبیه‌سازی مدل دوم یک توزیع پیشین دو متغیره BICAR را برای $\boldsymbol{\phi}_\ell = (\phi_{1\ell}, \phi_{2\ell})^T$ و برای ماتریس کواریانس Σ که همبستگی بین اثرات تصادفی ϕ_1 و ϕ_2 را شامل می‌شود، توزیع ویشارت وارون $\text{IW}(v_0, \boldsymbol{S}_0)$ با پارامترهای $\nu_0 = 3$ ، $S_0 = \text{diag}(2)$ در نظر می‌گیریم، که در آن I_2 یک ماتریس 2×2 همانی است. برای جلوگیری از محدودیت عدم شناساپذیری، پارامتر هموارسازی فضایی، را با پیروی از نیلون و همکاران (۲۰۱۵) $S = 1/38761e-16$ اختیار می‌کنیم. برای برآورد بیزی پارامترهای مدل، الگوریتم قدم زدن تصادفی با تکرار ۵۵۰۰۰۰، دوره داغیدن ۱۰۰۰۰ و انتخاب یک نمونه از هر ۱۰ نمونه تولید شده از زنجیره مارکوف، برای هر پارامتر نمونه‌ای تصادفی به حجم ۵۰ در هر ۳۱ ناحیه تولید شد. برای بررسی دقت و میزان اریبی برآوردها با همبستگی بین $\phi_{1\ell}$ و $\phi_{2\ell}$، چهار ماتریس کواریانس با ۴ ضریب همبستگی ۰/۷۵، ۰/۵، ۰/۲۵ و ۰ در نظر گرفته شده

است. مدل‌ها را با استفاده از تک زنجیره برای ساده‌سازی محاسبات پیاده‌سازی کردیم. نمودارهای اثر و چگالی حاشیه‌ای پسین برای برآوردعناصر بردار پارامترهای ضرایب رگرسیونی α و β و γ و عناصر ماتریس Σ بیانگر همگرایی نمونه‌های تولید شده با الگوریتم MCMC به توزیع هدف و سرعت کاهش همبستگی داده‌ها است. در جدول ۳ میزان اریبی برآورد پارامتر ضرایب رگرسیونی به همراه MSE آن‌ها ارائه شده است. مطابق با یافته‌های این جدول دقت مدل رگرسیون فضایی با اثرات تصادفی ناهمبسته نسبت به مدل گرسیون فضایی با اثرات تصادفی همبسته با افزایش ضریب همبستگی کاهش یافته است. با بررسی ملاک انحراف اطلاع DIC و $\hat{p_D}$تعداد پارامترهای موثر می‌توان نتیجه گرفت که با افزایش ضریب

جدول ۱: برآورد پارامترهای سه مدل رگرسیون فضایی با اثرات تصادفی همبسته فضایی با اثرات تصادفی همبسته

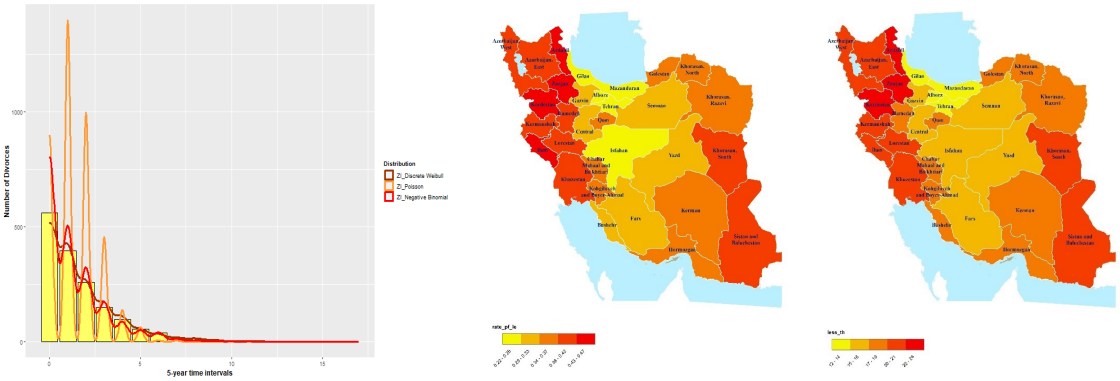
			$\rho = ۰$		$\rho = ۰/۲۵$		$\rho = ۰/۵$		$\rho = ۰/۷۵$	
مدل	پارامتر	مقدار واقعی	اریبی	MSE	اریبی	MSE	اریبی	MSE	اریبی	MSE
مدل فضایی با اثرات تصادفی همبسته	$\alpha_۰$	-۲	$-۰/۲۹$	۵/۳۲	۰/۱۴	۳/۸۳	$-۰/۱۲$	۴/۸۱	$-۰/۲۱$	۵/۱۷
	$\alpha_۱$	۰/۵	$-۰/۰۸$	۲/۳۷	$-۰/۱۴$	۲/۳۳	$-۰/۰۰۳$	۲/۶۳	$-۰/۱۵$	۲/۲۵
	$\alpha_۲$	۰/۳	۰/۰۵	۲/۳۴	۰/۰۲۷	۲/۲۲	۰/۰۱۳	۲/۲۳	۰/۰۲۸	۲/۳۳
	$\gamma_۰$	۱	$-۰/۰۸$	۰/۸۷	$-۰/۱۱۹$	۰/۹۳	$-۰/۰۶۲$	۰/۹۲۹	$-۰/۱۱۴$	۰/۹۶
	$\gamma_۱$	۱/۵	۰/۰۳	۱/۶۵	$-۰/۱۰۹$	۱/۶۸	$-۰/۰۳۶$	۱/۶۷	$-۰/۰۸۸$	۱/۴۴
	$\gamma_۲$	$-۰/۲$	۰/۰۰۱	۱/۸	۰/۰۱	۱/۸۴	۰/۰۹	۱/۷۲۳	$-۰/۱۰۱$	۲/۰۳
معیار ارزیابی مدل همبسته			PD	DIC	PD	DIC	PD	DIC	PD	DIC
			۲۷۹	۶۳۱	۳۳۶	۶۰۶	۳۹۹	۶۷۳	۳۷۲	۵۷۱
مدل فضایی با اثرات تصادفی ناهمبسته	$\alpha_۰$	-۲	۰/۰۳۳	۴/۱۸۶	$-۰/۳۶$	۴/۳۲	$-۰/۱۰۴$	۴/۶۹	۰/۰۳۳	۴/۱۸۶
	$\alpha_۱$	۰/۵	$-۰/۰۱۵$	۲/۲۸	$-۰/۰۱۹$	۲/۵۳	$-۰/۱۲۸$	۲/۲۴	۰/۱۰	۲/۲۸
	$\alpha_۲$	۰/۳	$-۰/۰۲۱$	۲/۱۶۹	$-۰/۰۷$	۲/۱۷	۰/۰۱۰	۲/۲۹	$-۰/۲۱$	۲/۱۶۹
	$\gamma_۰$	۱	۰/۳۱	۱/۲۰	۰/۰۹	۱/۱۵	۰/۲۶	۱/۲۸۹	$-۰/۳۱$	۱/۲۰
	$\gamma_۱$	۱/۵	۰/۰۶	۱/۷۳	۰/۱۱	۱/۰۱۴	$-۰/۰۳$	۱/۵۳	۰/۰۶	۱/۷۳
	$\gamma_۲$	$-۰/۲$	$-۰/۱۶$	۲/۱۹	۰/۰۴۴	۱/۸۵	۰/۰۲۲	۱/۹۵	$-۰/۱۶$	۲/۱۹
معیار ارزیابی مدل مستقل			PD	DIC	PD	DIC	PD	DIC	PD	DIC
			۳۱۴	۷۸۷	۳۷۳	۸۱۵	۳۸۴	۹۷۱	۴۸۵	۱۱۴۵

همبستگی، تناسب بهتر مدل فضایی با اثرات تصادفی هبسته نسبت به مدلی که اثرات تصادفی را جداگانه در نظر می‌گیرد، است.

۴ تحلیل داده‌های طلاق

طلاق به عنوان یک موضوع اجتماعی، انحلال قانونی ازدواج و جدایی زوجین است. تحقیق در مورد «زمان تا انحلال زناشویی » در زمینه تحلیل بقا نقش حیاتی در بررسی این معضل اجتماعی دارد. بررسی‌ها نشان می‌دهد که طول مدت زندگی مشترک در سنوات اخیر در حال کاهش است و زوجین در مدت زمان کوتاهی کمتر از ۵ سال، پس از ازدواج درخواست طلاق می‌دهند. ما در این مطالعه محور زمانی زندگی مشترک را به ۶ دوره ۵ ساله تقسیم کرده‌ایم که به صورت بازه‌های $(۰, ۵]$، $(۵, ۱۰]$، ... و مقادیر زمان بقای گسسته ۰ ، ۱ ، ... را به عنوان نقاط اولیه این بازه‌ها توصیف کند تا بتوانیم انباشتگی بیش از حد انتظار طلاق در بازه $(۰, ۵]$ سال را با مدل‌های صفر آماسیده بررسی نمائیم. از آنجا که مختصات جغرافیایی واحدها به طور دقیق در دسترس نیست و داده‌ها مشبکه‌ای هستند بر اساس روش نمونه‌گیری خوشه‌ای، از هر ۳۱ استان ۵۰ زوج که بین سال‌های ۱۳۵۰ تا ۱۳۹۸ یک یا چند ازدواج را تجربه کرده‌اند، انتخاب شدند. مجموعه داده نهایی شامل ۱۵۵۰ زوج در این مطالعه بود. که در نهایت ۸۷۴ زوج پیشامد طلاق را تجربه کرده‌اند. نمودار بافتنگار تعداد

طلاق برحسب طول مدت زندگی مشترک در پنل (ج) شکل ۱ نشان می‌دهد که درصد قابل توجهی از طلاق (حدود ۳۶ درصد) در پنج سال اول زندگی زناشویی به ثبت رسیده است، بنابراین داده‌ها بالقوه صفر آماسیده هستند. در این نمودار سه توزیع ZIDW ، ZINB و ZIP به داده‌ای زمان رسیدن به پیشامد طلاق برازش داده شده است. توزیع وایبول گسسته صفر آماسیده، یعنی ZIDW، با پارامترهای $\beta = ۱/۱$، $q = ۰/۶$ و با احتمال صفر شدگی $\pi = ۰/۳۶$ برای این داده‌ها مناسب‌تر از دو توزیع دیگر است. با این حال بدلیل آنکه معیار شاخص پراکندگی Dip که عنوان نسبت واریانس به میانگین تعریف می شود، برابر۱/۶۹ بوده است که نشاندهنده بیش پراکندگی در داده‌های زمان رسیدن تا پیشامد طلاق است، می‌توان نزدیکی دو توزیع ZIDW و ZINB را بخوبی در این نمودار مشاهده کرد. زوجینی که تا پایان مدت زمان مطالعه طلاق نداشته باشند از جمله مواردی هستند که هیچ رویدادی ندارند. برای این موارد خاص، فقط می‌توان زمان آن‌ها را سانسور از راست در نظر گرفته‌ایم. از آنجا که داده‌های ما بقای صفر آماسیده هستند، مطابق با شکل ۱ ما همبستگی فضایی مشاهدات را بطور جداگانه برای دو مؤلفه "نسبت طلاق با طول مدت ازدواج کمتر از ۵ سال به کل طلاق در سطح منطقه (احتمال صفر شدگی بقا)" در پنل الف و "میانه تعداد سالهای قبل از طلاق در پنل ب برای افرادی با بقای ازدواج بیشتر از ۵ سال" مشاهده‌کردیم. همچنین از آنجایی که این دو نقشه الگوهای فضایی مشابهی را نشان می‌دهند، بنابراین جای تعجب نیست که بین دو اثر تصادفی همبستگی زیادی وجود دارد. سپس با استفاده از روش انتخاب متغیر در مدل‌های رگرسیونی صفر



(ج) (ب) (الف)

شکل ۱ : شکل الف: احتمال بقای کمتر از ۵ سال، ب: میانگین طول مدت ازدواج بیشتر از سال ۵ زوجین با دوام ازدواج، ج: توزیع صفر آماسیدهطول زندگی مشترک بین زوجین

آماسیده دریافتیم که از میان تمام متغیرهای تبیینی، ماتریس طرح X برای مدل‌بندی احتمال صفر بودن مطابق با رابطه (۴.۲) مشتمل بر متغیرهای کمکی وضعیت اشتغال مرد $x_۱$، مجموع درآمد خانواده $x_۲$، تعداد فرزندان $x_۳$، سابقه ازدواج پیشین یکی از زوجین $x_۴$ و هم‌کفوی $x_۵$ است. همچنین ماتریس طرح Z مشتمل بر متغیرهای کمکی ضعیت اشتغال مرد $z_۱$، مجموع درامد خانواده $z_۲$، تعداد فرزندان $z_۳$، اختلاف سنی زوجین$z_۴$ و تحصیلات زن $z_۵$ است. که برای مدل‌بندی مقادیر شمارشی غیر صفر مطابق با رابطه (۳.۲) در نظر گرفته می‌شوند. همچنین با در نظر گرفتن اثرات تصادفی فضایی$\phi_۱$ و $\phi_۲$ دو مدل رگرسیونی خطی تعمیم یافته به صورت

$$\text{logit}(\pi_{i\ell}) = \gamma_۰ + \gamma_۱ x_{۱i\ell} + \gamma_۲ x_{۲i\ell} + \gamma_۳ x_{۳i\ell} + \gamma_۴ x_{۴i\ell} + \gamma_۵ x_{۵i\ell} + \phi_{۱i},$$
$$\log(-\log(q_{i\ell})) = \alpha_۰ + \alpha_۱ z_{۱i\ell} + \alpha_۲ z_{۲i\ell} + \alpha_۳ z_{۳i\ell} + \alpha_۴ z_{۴i\ell} + \alpha_۵ z_{۵i\ell} + \phi_{۲i},$$

خواهند بود. با وجود اینکه دو نقشه ترسیم شده در پنل (الف) و (ب) شکل ۱ بر همبستگی فضایی داده‌های طلاق و همیستگی میان دو اثر تصادفی فضایی تأکید نموده است، برای اطمینان بیشتر ۳ مدل کلاسیک CZIDW ، مدل فضایی CZIDW با اثرات تصادفی هم‌بسته و مدل فضایی CZIDW با اثرات تصادفی مستقل بر روی داده‌ها برازش داده شد و از طریق ملاک اطلاع انحراف، یعنی مجموع میانگین انحراف و تعداد پارامترهای موثر برآورد شده مورد ارزیابی قرار گرفتند. مقدار ملاک DIC برای مدل رگرسیون فضایی با اثرات تصادفی همبسته فضایی ($\hat{p_D} = ۶۱/۴$) ۱۲۱/۰۸ است که نسبت به ($\hat{p_D} = ۶۴/۹$) ۱۲۷/۸ مدل فضایی با اثرات تصادفی جداگانه کاهش یافته است. مدل رگرسیون فضایی CZIDW با اثرات تصادفی همبسته و مدل رگرسیون فضایی CZIDW با اثرات تصادفی ناهمبسته نسبتاً مثل هم عمل کرده و هر دو نسبت به مدل کلاسیک دارای مقدار DIC کمتری هستند. مدل کلاسیک بدترین تناسب با ($\hat{p_D} = ۸۴/۷$) ۲۳۲/۱۷ را در بین تمام مدل‌های در نظر گرفته شده دارد. سپس برای ارزیابی عملکرد توزیع CZIDW نسبت به دو توزیع CZINB وCZIP در مدل‌های فضایی با اثرات تصادفی هبسته، این معیارها را مطابق با جدول ۲ارزیابی نموده و درنهایت برآورد بیزی (میانگین پسین) از پارامترهای مدل فضایی بقای صفر آماسیدهCZIDW به همراه بازه باورمندی ۹۵ درصدی آن‌ها محاسبه و در جدول ۳ گرآوری شده‌است. میانگین پسین ضرایب رگرسیونی مدل احتمال صفر شدگی برای داده‌های بقای

جدول ۲: مقایسه معیار ارزیابی مدل فضایی با اثرات تصادفی همبسته برای ۳ توزیع گسسته صفر آماسیده

توضیح	متغیر	CZIP	CZINB	CZIDW
ملاک اطلاع انحراف	DIC	۸۳۶	۶۹۶	۶۷۳
تعداد پارامترهای موثر	$\hat{p_D}$	۵۷۴	۳۴۸	۳۹۹

جدول ۳: برآورد پسین ضرایب رگرسیون مدل دو بخشی CZDIW با اثرات تصادفی همبسته

	$\alpha_۰$	$\alpha_۱$	$\alpha_۲$	$\alpha_۳$	$\alpha_۴$	$\alpha_۵$	$\gamma_۰$	$\gamma_۱$
میانگین پسین	۵/۱۶	۱/۶۷	۰/۴۵	۱/۸۶	−۰/۳۳	−۰/۶۷	۱/۲۹	۲/۵۴
بازه باورمندی	(۴/۶۵, ۵/۶)	(۱/۲۸, ۲/۰۹)	(۰/۲۹, ۰/۶۲)	(۱/۶۷, ۲/۰۲)	(−۰/۵۷, −۰/۰۸)	(−۰/۹۳, −۰/۴۱)	(۱/۰۹, ۱/۴۹)	(۲/۳۶, ۲/۷۳)
	$\gamma_۲$	$\gamma_۳$	$\gamma_۴$	$\gamma_۵$	$\Sigma_{۱۱}$	$\Sigma_{۱۲}$	$\Sigma_{۲۲}$	ρ
میانگین پسین	۰/۹۴	۲/۷۹	۰/۰۷	۲/۲۷	۰/۶۲	۰/۳۱	۱/۰۵	۰/۵۵
بازه باورمندی	(۰/۶۹, ۱/۱۸)	(۲/۶۳, ۲/۹۵)	(−۰/۰۸, ۰/۲۳)	(۲/۰۴, ۲/۴۹)	(۰/۶۰, ۰/۶۴)	(۰/۲۲, ۰/۴۲)	(۰/۹۸, ۱/۲۱)	(۰/۴۱, ۰/۵۹)

فضایی با توزیع CZIDW که متغیرهای تبیینی را از طریق تابع پیوند لجیت به احتمال اینکه زمان بقای زوجین کمتر از ۵ سال باشد، مرتبط می‌سازد، در جدول ۲ نشان می‌دهد که تعداد فرزندان در ۵ سال اول زندگی مشترک، وضعیت اشتغال مرد و سابقه ازدواج پیشین یکی از زوجین با برآوردهای پسین به ترتیب ۲/۷۹ ، ۲/۵۴ و ۲/۲۷ بزرگترین ضرایب رگرسیونی در این مدل را به خود اختصاص داده‌اند. همچنین در مدل پیشنهادی ما، بیشترین تاثیر را در مدت زمان دوام ازدواج به شرط آنکه زوجین حداقل ۵ سال با هم زندگی کرده اند، تعداد فرزندان و وضعیت اشتغال مرد دارد. این امر نشاندهنده این است که هرچه تعداد فرزندان بیشتر باشد و شغل مرد با ثبات تر باشد، دوام زندگی زناشویی پس از ۵ سال بیشتر است. همچنین، بازه باورمندی پسین کاملاً باریک و به دور از صفر محدود شده بودند، که نشان می‌دهد مؤلفه‌های واریانس به خوبی شناسایی شده‌اند. برآورد همبستگی اثر تصادفی $\rho = ۰/۵۵$ و بازه باورمندی پسین [۰/۴۱, ۰/۵۹] است که بر ضرروت به کارگیری مدل دو متغیره CZIDW و تناسب آن تأکید می‌کند.

بحث و نتیجه‌گیری

در این مقاله به تحلیل فضایی داده‌های بقای صفر آماسیدهCZDIW با مشاهدات راست سانسوریده پرداخته شده‌است. برای لحاظ کردن اثرات تصادفی فضایی در مدل دو رویکرد وجود داشت. در رویکرد اول اثرات تصادفی مستقل (ناهمبسته) و در رویکرد دوم همبسته در نظر گرفته شده‌اند. با انجام مطالعات شبیه سازی دریافتیم که دقت برآورد پارامترها در مدل فضایی با اثرات تصادفی همبسته بیشتر است و پس ازبررسی ملاک‌های ارزیابی مدل‌های موجود روی داده‌های زمان رسیدن به پیشامد طلاق با انباشتگی در بازه صفر تا ۵ سال دریافتیم، که این مدل برازش بهتری روی داده‌ها دارد.

مراجع

Besag, J, York J. and Mollié, A. (1991), Bayesian Image Restoration with Two Applications in Spatial Statistics, *Annals of the Institute of Statistics and Mathematics*, **43**, 1–59.

Jenkins, S. P. (2005), Survival Analysis, *Unpublished Manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK* .

Neelon, B. H., Ghosh, P. and Loebs P. F. (2015), A Spatial Poisson Hurdle Model for Exploring Geographic Variation in Emergency Department Visits, *Journal of the Royal Statistical Society: Series A*, **176**, 389–413.

Nyandwi, E., Osei, F. B., Amer, S. and Veldkamp, A., (2020), Modeling Schistosomiasis Spatial Risk Dynamics Over time in Rwanda Using Zero-inflated Poisson Regression, *Scientific Reports* , **10**, 1-9.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

تحلیل فضایی توسعه‌یافتگی اشتغال استان‌های کشور

فهیمه برومندی [۱]
سازمان مدیریت و برنامه‌ریزی استان فارس

چکیده: شناخت وضعیت اشتغال و بیکاري و درک نقاط قوت و ضعف آن در برنامه‌ريزي‌هاي اشتغال و توسعه، تأثير به‌سزايي دارد. در این مقاله به منظور تعیین جایگاه استان‌های کشور در موضوع اشتغال‌زایی، شاخص‌های حاصل از اجرای طرح آمارگیری نیروی کار مرکز آمار ایران در دوره چهار ساله ۱۳۹۸ لغایت ۱۴۰۱ مورد استفاده قرار گرفته و بر اساس آن استان‌های کشور با مدل تحلیل خوشه‌ای سلسله مراتبی به پنج خوشه تقسیم شده‌اند. نتایج بررسی نمایش فضایی وضعیت اشتغال‌زایی نشان داد میزان بهره‌مندی در مناطق شمال، شمال‌غرب، شمال شرق و مناطق مرکزی کشور به استثنای استان‌های گلستان، قم و مرکزی مطلوب‌تر و در مناطق جنوب، جنوب‌غرب و جنوب‌شرق کشور وضعیت نامطلوبی دارد.

واژه‌های کلیدی: تحلیل فضایی، اشتغال‌زایی استان‌ها، تحلیل خوشه‌ای سلسله مراتبی، طرح آمارگیری نیروی‌کار
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62G30.

۱ مقدمه

با پیشرفت علم و تکنولوژی، تولید داده‌ها در عصر حاضر به مراتب بیشتر از گذشته است. مدیریت داده‌های خام و استخراج اطلاعات و دانش مفید از آنها، نقش مهمی در تصمیم‌گیری‌ها دارد. تحلیل داده‌ها در اکثر حوزه‌های پژوهش از جمله مدیریت، اقتصاد، مهندسی و سایر رشته‌ها استفاده شده است (رادمهر و علم الهدایی ، ۱۳۹۳). خوشه‌بندي به‌عنوان يكي از روش‌هاي داده‌كاوي توصيفي، به دليل توانايي و قابليت‌هاي بالايي كه در تلخيص اطلاعات و دسته‌بندی آن‌ها دارد مورد توجه محققان و پژوهشگران علوم مختلف قرار گرفته است. اشتغال نيروي انساني ضروری‌ترين هدف برنامه‌ريزي اقتصادي اجتماعي هر كشوري را تشكيل مي‌دهد، به‌طوری که میزان اشتغال افراد در جامعه یکی از مهم‌ترین شاخص‌هاي توسعه‌يافتگی است (آقابخشی و میبدی ، ۱۳۹۲). رسيدن به اهداف چشم‌انداز اشتغال در كشور، از يك طرف بستگي به بررسي و آگاهي از جايگاه مناطق مختلف كشور در زمينه اشتغال و بيكاري و شناخت توانمندي‌ها و محدوديت‌هاي اين مناطق در اين زمينه دارد و از طرف ديگر بستگي تام به تصميمات و عملكرد نهادهاي اقتصادي، اجتماعي و قانوني كشور و همچنين تغيير در قوانين و اجراي آن، در جهت تشويق سرمايه‌گذاري و فعاليت‌هاي كارآفريني دارد (تقدیسی و همکاران ، ۱۳۹۱). در سال‌های اخیر

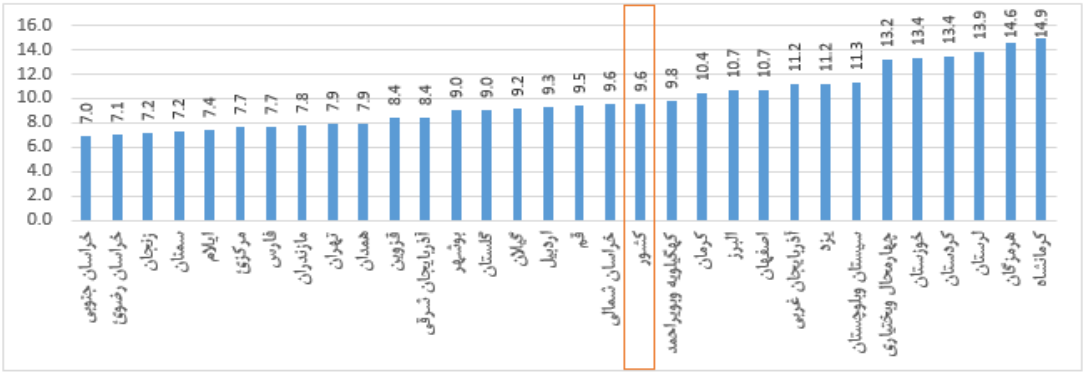
[۱] نام و ایمیل ارائه دهنده مقاله: فهیمه برومندی، fahimeh.boromandi@gmail.com

مقالات متعددی در زمینه بررسی وضعیت اشتغال و ارزیابی و تحلیل شاخص‌های بازار کار نوشته شده است که از جمله آن می‌توان به موارد ذیل اشاره نمود: (صیدایی و همکاران ، ۱۳۹۰) وضع بیکاري و اشتغال کشور از سال ۱۳۳۵ تا ۱۳۸۹ را بررسی نمودند. (رضوانی و همکاران ، ۱۳۹۲) به تحلیل مکانی بیکاري در نواحی شهري و روستایی ایران با رویکرد تحلیل اکتشافی داده‌هاي مکانی پرداختند. (خوچیانی و حسینی ، ۱۳۹۹) به خوشه‌بندی استان‌هاي کشور در ارزیابی و تحلیل نرخ بیکاري با استفاده از روش مبتنی بر چگالی پیش‌بینی پرداختند. (زارعی و برومندی ، ۱۳۹۸) استان‌های کشور را بر مبنای تغییرات شاخص‌های عمده‌ی نیروی کار در دوره ۱۳۹۴ـ۱۳۹۷ با استفاده از روش تاپسیس رتبه‌بندی نمودند. این نکته نیز حائز اهمیت است که کاهش نرخ بیکاری یا حتی تعداد مطلق بیکاران، در صورت افزایش نرخ مشارکت اقتصادی نشانه‌ی توفیق سیاست‌ها است و کاهش یا ثبات نرخ بیکاری با وجود کاهش نرخ مشارکت اقتصادی توفیق در بازار کار محسوب نمی‌شود (حسین‌زاده ، ۱۳۹۹). در این راستا هدف مطالعه حاضر تحلیل فضایی استان‌های کشور براساس شاخص‌های عمده طرح نیروی کار و تعیین مناطق همگن با استفاده از روش خوشه‌بندی سلسله‌مراتبی در دوره ۱۳۹۸ـ۱۴۰۱ جهت شناسایی نقاط قوت و ضعف به منظور ارائه راهبردهای مناسب توسعه اشتغال برای برنامه‌ریزی‌های آتی است.

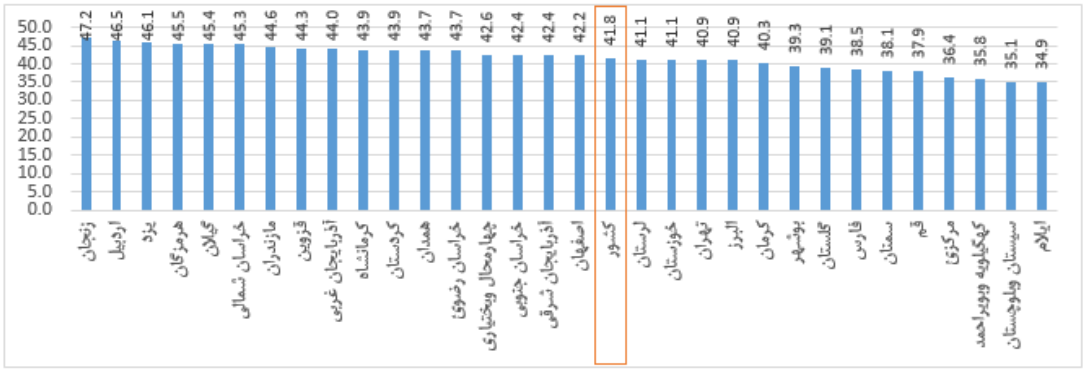
۲ روش پژوهش

با توجه به ماهیت موضوع، رویکرد حاکم بر این پژوهش توصیفی- تحلیلی و از نوع کاربردی است. روش گردآوری داده‌ها در این پژوهش به صورت کتابخانه‌ای و بر اساس مستندات مرکز آمار ایران، به‌ویژه نتایج طرح آمارگیری نیروی کار طی سال‌های ۱۳۹۸ لغایت ۱۴۰۱ می‌باشد. برای بررسی وضعیت اشتغال در استان‌های کشور با توجه به میانگین هندسی شاخص‌های عمده طرح نیروی کار در دوره مذکور که در این پژوهش شاخص‌های اشتغال نامیده می‌شوند و با استفاده از نرم‌افزار R و بهره‌گیری از تحلیل خوشه‌بندی، استان‌های کشور در گروه‌های همگن طبقه‌بندی می‌شوند. نهایتاً نتایج حاصل از تحلیل خوشه‌بندی با استفاده از نرم‌افزار GIS مورد بررسی و نمایش فضایی وضعیت اشتغال‌زایی استان‌های کشور بر اساس سطح بهره‌مندی از شاخص‌های اشتغال ارائه می‌گردد. طرح آمارگیری نیروی کار یکی از مهم‌ترین طرح‌های آمارگیری مرکز آمار ایران است. اطلاعات این طرح که از اعتبارهاي لازم و قابل قبول برخوردار هستند به عنوان منابع اطلاعاتي مهم، در برنامه‌ریزي و برنامه‌هاي توسعه کاربرد فراوان دارند (چکیده‌ی نتایج طرح آمارگیری نیروی کار ، ۱۳۹۹). در این مقاله، برای تعیین جایگاه استان‌های کشور در موضوع اشتغال‌زایی بر روی میانگین هندسی شاخص‌های عمده طرح نیروی کار طی سال‌های ۱۳۹۸ لغایت ۱۴۰۱ متمرکز می‌شویم و شاخص‌های اشتغال مورد استفاده عبارتند از: c_1: میانگین نرخ بیکاری طی سال‌های ۱۳۹۸ـ۱۴۰۱ (بر اساس جمعیت ۱۵ سال و بیشتر) c_2: میانگین نرخ مشارکت اقتصادی طی سال‌های ۱۳۹۸ـ۱۴۰۱ (بر اساس جمعیت ۱۵ سال و بیشتر) c_3: میانگین نسبت اشتغال طی سال‌های ۱۳۹۸ـ۱۴۰۱ (بر اساس جمعیت ۱۵ سال و بیشتر) در یک بررسی توصیفی از لحاظ شاخص اشتغال c_1، کمترین میانگین نرخ بیکاری در این دوره مربوط به استان خراسان‌جنوبی با متوسط نرخ بیکاری ۰/۷ درصد و بیشترین میانگین نرخ بیکاری مربوط به استان کرمانشاه، با متوسط نرخ بیکاری ۹/۱۴ درصد است (شکل ۱). در شکل ۲، ملاحظه می‌شود بیشترین میانگین نرخ مشارکت اقتصادی مربوط به استان زنجان با متوسط نرخ مشارکت ۲/۴۷ درصد و کمترین میانگین نرخ مشارکت نیز مربوط به استان ایلام با نرخ مشارکت ۹/۳۴ درصد است. همچنین بیشترین و کمترین میانگین نسبت اشتغال در این دوره به ترتیب مربوط به استان زنجان با متوسط نسبت اشتغال ۸/۴۳ درصد و استان سیستان و بلوچستان با متوسط نسبت اشتغال ۱/۳۱ درصد است (شکل ۳).

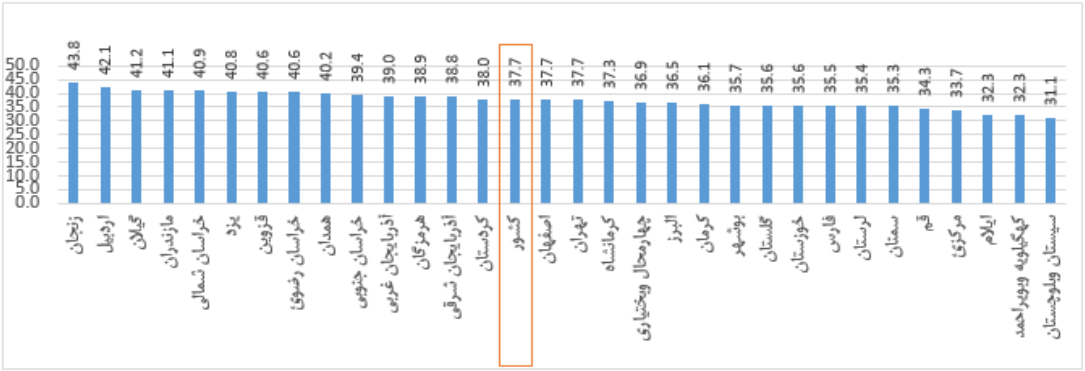
بنابراین با توجه به نتایج بررسی توصیفی شاخص‌های اشتغال می‌توان نتیجه گرفت که بین استان‌های کشور از لحاظ شاخص‌های اشتغال نابرابری و شکاف وجود دارد که در ادامه به تعیین مناطق همگن و همچنین سطح‌بندی استان‌های کشور



شکل ۱: مقایسه استان‌های کشور براساس میانگین نرخ بیکاری طی سال‌های ۱۳۹۸ ـ ۱۴۰۱



شکل ۲: مقایسه استان‌های کشور براساس میانگین نرخ مشارکت اقتصادی طی سال‌های ۱۳۹۸ ـ ۱۴۰۱



شکل ۳: مقایسه استان‌های کشور براساس میانگین نسبت اشتغال طی سال‌های ۱۳۹۸ ـ ۱۴۰۱

بر اساس شاخص‌های اشتغال پرداخته می‌شود.

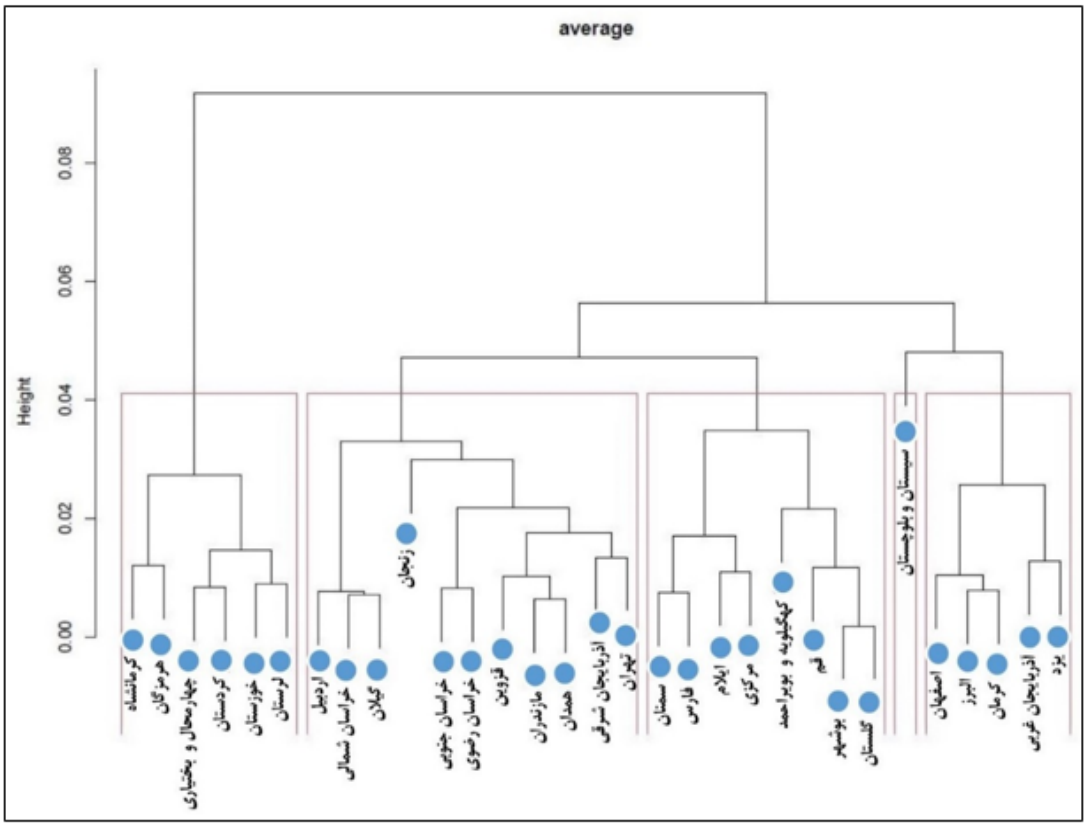
۱.۲ خوشه‌بندی

خوشه‌بندی یکی از روش‌های داده‌کاوی توصیفی است که برای گروه‌بندی مشاهدات (استان‌های کشور) به k خوشه (گروه) با توجه به میزان فاصله آن‌ها از یکدیگر می‌باشد به طوری‌که مشاهداتی که در یک خوشه قرار می‌گیرند بیشترین شباهت و مشاهدات خوشه‌های مختلف بیشترین تفاوت ممکن با یکدیگر را دارند (کافمن و راسیو ، ۲۰۰۵). در این پژوهش با

جدول ۱: میانگین شاخص‌های اشتغال بر اساس خوشه‌های دست آمده

خوشه‌های همگن	استان‌ها	میانگین c_1	میانگین c_2	میانگین c_3
خوشه اول	اردبیل- خراسان شمالی- گیلان- زنجان- خراسان جنوبی- خراسان رضوی- قزوین- مازندران- همدان- آذربایجان شرقی- تهران	۲/۸	۲/۴۴	۶/۴۰
خوشه دوم	اصفهان- البرز- کرمان- آذربایجان غربی- یزد	۸/۱۰	۷/۴۲	۳۸
خوشه سوم	کرمانشاه- هرمزگان- چهارمحال و بختیاری- کردستان- خوزستان- لرستان	۹/۱۳	۴۳	۳۷
خوشه چهارم	سمنان- فارس- ایلام- مرکزی- کهگیلویه و بویراحمد- قم- بوشهر- گلستان	۴/۸	۵/۳۷	۳۴/۳
خوشه پنجم	سیستان و بلوچستان	۳/۱۱	۱/۳۵	۱/۳۱

استفاده از نرم‌افزار R و تحلیل خوشه‌بندی سلسله مراتبی روش تجمیعی و الگوریتم پیوند متوسط استان‌هایی که بیشترین شباهت از نظر میانگین شاخص‌های اشتغال طی سالهای ۱۳۹۸ ـ ۱۴۰۱ را دارند در یک خوشه دسته‌بندی می‌شوند. لازم به ذکر است برای تعیین تعداد خوشه‌ها، در بسته Nbclust در نرم‌افزار R ، از ۲۶ معیار مختلف برای تصمیم‌گیری استفاده می‌شود، در این مقاله با توجه به خروجی بسته Nbclust و بیش‌ترین تفسیرپذیری در برچسب خوشه‌ها، تعداد پنج خوشه پیشنهاد می‌گردد. همانطور که در دندروگرام خوشه‌ها (شکل ۴) و جدول ۱ مشاهده می‌شود استان‌های کشور در زمینه‌های شاخص‌های اشتغال به پنج خوشه‌ی همگن از جمله بسیار بهره‌مند، بهره‌مند، نسبتاً بهره‌مند (متوسط)، محروم و بسیار محروم طبقه‌بندی شده‌اند که در ادامه هریک از خوشه‌ها تشریح می‌شوند.



شکل ۴: خوشه‌بندی استان‌های کشور بر اساس شاخص‌های اشتغال

خوشه اول، بسیار بهره‌مند: در این خوشه، یازده استان اردبیل، خراسان شمالی،گیلان، زنجان، خراسان جنوبی، خراسان رضوی، قزوین، مازندران، همدان، آذربایجان شرقی و تهران با میانگین نرخ بیکاری ۲/۸ درصد، میانگین نرخ مشارکت ۲/۴۴ درصد و میانگین نسبت اشتغال ۶/۴۰ درصد بهره‌مندترین استان‌های کشور به لحاظ شاخص‌های اشتغال هستند.

خوشه دوم، بهره‌مند: استان‌های اصفهان، البرز، کرمان، آذربایجان غربی و یزد با میانگین نرخ بیکاری ۸/۱۰ درصد، میانگین نرخ مشارکت ۷/۴۲ درصد و میانگین نسبت اشتغال ۳۸ به لحاظ بهره‌مندی از شاخص‌های اشتغال در گروه بهره‌مند واقع شده‌اند. بنابراین می‌توان چنین بیان کرد که در مجموع ۶/۵۱ درصد استان‌های کشور در سطح بسیار بهره‌مند و بهره‌مند به لحاظ شاخص‌های اشتغال جای گرفته‌اند. خوشه سوم، نسبتاً بهره‌مند (متوسط): در خوشه سوم، شش استان کرمانشاه، هرمزگان، چهارمحال و بختیاری، کردستان، خوزستان و لرستان با میانگین نرخ بیکاری ۹/۱۳ درصد، میانگین نرخ مشارکت۴۳ درصد و میانگین نسبت اشتغال ۳۷ که حدود ۳/۱۹ درصد استان‌های کشور را شامل می‌شوند به لحاظ بهره‌مندی از شاخص‌های اشتغال در سطح متوسط واقع شده‌اند. خوشه چهارم، محروم : در خوشه‌چهارم، هشت استان سمنان، فارس، ایلام، مرکزی، کهگیلویه و بویراحمد، قم، بوشهر و گلستان که حدود ۸/۲۵ درصد استان‌های کشور را شامل می‌شوند با میانگین نرخ بیکاری ۴/۸ درصد، میانگین نرخ مشارکت ۵/۳۷ درصد و میانگین نسبت اشتغال ۳/۳۴ به لحاظ بهره‌مندی از شاخص‌های اشتغال وضعیت مطلوبی ندارند. خوشه پنجم، بسیار محروم : استان سیستان و بلوچستان با میانگین نرخ بیکاری ۳/۱۱ درصد، میانگین نرخ مشارکت ۱/۳۵ درصد و میانگین نسبت اشتغال ۱/۳۱ درصد وضعیت مطلوبی نداشته و در سطح بسیار محروم واقع شده است. شایان ذکر است برای درک بیشتر تناسب هر یک از برچسب‌های خوشه‌ها توجه به نکات ذیل ضروری است: براساس رتبه‌بندی استان‌های کشور براساس هریک از شاخص‌های اشتغال مورد استفاده (شکل‌های ۳ـ۱)، استان‌های واقع در خوشه بسیار بهره‌مند، متفاوت (بهتر) از استان‌های مورد بررسی هستند. در واقع جایگاه مساعد استان‌های خوشه بسیار بهره‌مند به علت برتری این استان‌ها در همه‌ی شاخص‌های اشتغال است به‌طوری که در هر سه شاخص، این استان‌ها اغلب جزو شش استان برتر کشور هستند. همچنین در رابطه با استان‌های واقع در خوشه بهره‌مند، در دو شاخص میانگین نرخ مشارکت اقتصادی و نسبت اشتغال شرایط مساعدتری را نسبت به استان‌های خوشه‌های پایین داشته‌اند. از آنجایی‌که کاهش نرخ بیکاری لزوماً افزایش نسبت اشتغال را دربرندارد (برای جزئیات بیشتر به مثال تشریحی در مقاله (زارعی و برومندی ، ۱۳۹۸) مراجعه شود) و همچنین با توجه به تعریف نرخ بیکاری که از تقسیم جمعیت بیکار به جمعیت فعال (مجموع بیکار و شاغل) بدست می‌آید افزایش جمعیت شاغل منجر به کاهش نرخ بیکاری می‌شود؛ بنابراین استان‌های خوشه سوم در مقایسه با خوشه چهارم، علی‌رغم اینکه نرخ بیکاری بیشتری دارند ولی از لحاظ متوسط نرخ مشارکت اقتصادی و نسبت اشتغال وضعیت بهتری داشته و برچسب نسبتاً بهره‌مند مناسب این خوشه است. به عنوان مثال در رابطه با قرار گرفتن استان فارس در خوشه محروم می‌توان به جایگاه ضعیف این استان در شاخص‌های میانگین نرخ مشارکت اقتصادی و نسبت اشتغال در مقایسه با استان‌های کشور اشاره کرد. در واقع تعداد کم شاغلین و جمعیت فعال در این استان نسبت به جمعیت کل استان و جمعیت ۱۵ ساله و بیش‌تر، عامل اصلی قرار گرفتن این استان در خوشه محروم است. علاوه‌بر این استان سیستان و بلوچستان در دو شاخص نرخ مشارکت و نسبت اشتغال جزو دو استان ضعیف کشور و در شاخص نرخ بیکاری نیز جزو هفت استان ضعیف می‌باشد بنابراین با توجه به اینکه به لحاظ بهره‌مندی ازهمه‌ی شاخص‌های اشتغال وضعیت نامطلوبی دارد در خوشه بسیار محروم قرار گرفته است. در ادامه برای نمایش بهتر توزیع فضایی و توسعه‌یافتگی استان‌های کشور در موضوع اشتغال‌زایی از نرم‌افزار GIS استفاده شده است که شکل ۴ نتایج آن را نشان می‌دهد.

بحث و نتیجه‌گیری

با بررسی جایگاه استان‌ها در شاخص‌های اشتغال مورد مطالعه در این مقاله مشخص شد که استان‌هایی که در این سه شاخص اغلب جزو شش استان برتر بوده‌اند در خوشه‌ی بسیار بهره‌مند واقع شده‌اند. به صورت خلاصه ۶/۵۱ درصد استان‌های کشور در خوشه‌ی بسیار بهره‌مند و بهره‌مند، حدود ۴/۱۹ درصد در خوشه‌ی نسبتاً بهره‌مند و ۷/۳۸ درصد در خوشه محروم

شکل ۵: نمایش فضایی وضعیت اشتغال‌زایی استان‌های کشور طی سال‌های ۱۳۹۸ لغایت ۱۴۰۱

و بسیار محروم واقع شده‌اند.همچنین میزان بهره‌مندی در مناطق شمال، شمال‌غرب، شمال‌شرق و مناطق مرکزی به استثنای استان‌های گلستان، قم و مرکزی مطلوب‌تر و در مناطق جنوب، جنوب‌غرب و جنوب‌شرق کشور وضعیت نامطلوبی دارد. براساس این پژوهش که از نابرابری فضایی استان‌های کشور در بهره‌مندی از شاخص‌های اشتغال حکایت دارد و با توجه به حساسیت مسأله اشتغال و تأثیر همه جانبه آن بر ابعاد مختلف اقتصادی، اجتماعی و فرهنگی، به برنامه‌ریزان و سیاست‌گذاران پیشنهاد می‌شود که برای تعدیل نابرابری فضایی و توسعه‌ی متوازن فرصت‌های شغلی در بین استان‌های کشور راهکارهای مناسبی اتخاذ نمایند به‌طوری که ضمن حفظ و ارتقاء وضعیت موجود، استان‌های محروم و بسیار محروم را برای توسعه در آینده در اولویت قرار دهند. در واقع در این استان‌ها ضمن توجه به توانمندی‌های هر استان و مطالعات امکان‌ـسنجی در سطوح پایین برنامه‌ریزی(روستا) از طریق اجرای طرح‌های اشتغال‌زایی روستایی، مشاغل خانگی و صنایع تبدیلی وغیره شرایط مساعدی برای اشتغال در این استان‌ها فراهم گردد. برای استان‌های نسبتاً بهره‌مند(متوسط) نیز پیشنهاد می‌گردد که این استان‌ها بعد از استان‌های بسیار محروم و محروم، در اولویت برنامه‌های توسعه اشتغال قرار گیرند و استان‌های بسیار بهره‌مند و بهره‌مند ضمن حفظ وضع موجود این استان‌ها در زمینه‌ی اشتغال با برنامه‌ریزی‌های منسجم و مؤثر، جایگاه این استان‌ها را در زمینه بهره‌مندی از شاخص‌های اشتغال بهبود بخشید.

مراجع

آقابخشی، ح. و رشیدیان میبدی، م. (۱۳۹۲)، *برنامه هاي كاهش بيكاري در برنامه سوم توسعه اقتصادي*، و جایگاه مددكاري اجتماعي در تأمين رفاه، فصلنامه پژوهش اجتماعی، ۵(۲۰)، ۱۳۷ـ۱۵۹.

تقدیسی، ا.؛ جمینی، د.؛ مرادی، ن. (۱۳۹۱)، بررسی و تحلیل روند اشتغال و بیکاری در شهرستان صحنه طی دوره های ۸۵ـ۱۳۷۵، *مجله‌ی علمی تخصصی برنامه‌ریزی فضایی*، ۱(۳)، ۸۱ـ۱۰۶.

چکیده نتایج طرح آمارگیری نیروی کار (۱۳۹۹)، مرکز آمار ایران.

حسین‌زاده، ج.(۱۳۹۹). دلایل و شواهدِ سازگاریِ آمارهای اشتغال و رشد اقتصادی ایران، مرکز آمار ایران.

خوچیانی، ر.؛ حسینی، س. م. (۱۳۹۹)، ارزیابی و تحلیل نرخ بیکاری در سطح استان‌های کشور با استفاده از خوشه‌بندی مبتنی بر چگالی پیش‌بینی، *فصل‌نامه علمی برنامه‌ریزی منطقه‌ای*، **۱۰** (۳۷).

رادمهر، ف.؛ علم الهدایی، س. ح.(۱۳۹۳). خوشه‌بندی: ابزاری برای آنالیز داده‌ها در مطالعات کمی و آمیخته، *روش‌ها و مدل‌های روانشناختی*، **۴** (۱۵)، ۱۳ ـ ۳۶.

رضوانی، م.؛ منصوریان، حسین؛ محمودیان زمانه، م.، حیدریان محمدآبادی، ر. (۱۳۹۲). تحلیل مکانی بیکاري در نواحی شهري و روستایی ایران با *رویکرد تحلیل اکتشافی داده‌هاي مکانی*، فصلنامه برنامه‌ریزی کالبدی ـ فضایی، **۱** (۳)، ۳۷ ـ ۴۸.

زارعی، س.؛ برومندی، ف. (۱۳۹۸). رتبه‌بندی استان‌های کشور بر مبنای تغییرات شاخص‌های عمده‌ی نیروی کار در دوره ۱۳۹۴ ـ ۱۳۹۷ با استفاده از روش تاپسیس، مجله بررسی‌های آمار رسمی ایران، **۳۰**(۱)، ۱۹۳ ـ ۲۰۸.

صیدایي، ا.؛ بهاري، ع.؛ زارعي، ا. (۱۳۹۰). بررسي وضعیت اشتغال و بیکاري در ایران طی سال‌های۱۳۳۵ ـ ۱۳۸۹، راهبرد یاس، **۱** (۲۵)، ۲۱۶ ـ ۲۴۱.

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, NewYork. https://www.amar.org.ir/

Kaufman, L. and Rousseeuw, P.J. (2005): Finding Groups in Data: *An Introduction to Cluster Analysis*, Hoboken, NJ: Wiley.

Rencher, A. (2002), *Method of Multivariate Analysis*, Wiley.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

بررسی عملکرد روندزدایی از داده فضایی با استفاده از رگرسیون بردار پشتیبان

ساره حدادی[۱]، جواد اطمینان
گروه آمار، دانشگاه بیرجند

چکیده: در سال‌های اخیر انواع مدل‌های رگرسیون بردار پشتیبان مورد توجه علوم مختلف به ویژه آمار فضایی قرار گرفته‌اند. به دلیل وجود هسته در مسائل ناخطی و امکان استفاده از توابع مخاطره متفاوت، پژوهشگران با مدل‌های گسترده‌ای در رگرسیون بردار پشتیبان مواجه هستند. از طرفی این روش ویژگی‌های قابل توجهی برای مجموعه داده با حجم کوچک، داده پرت و ابعاد بالا دارد. در این مطالعه، یک مجموعه داده فضایی بدون روند شبیه‌سازی شده و سپس روند به آن اضافه می‌شود. در مجموعه رونددار، مدل‌بندی تابع روند با استفاده از رگرسیون بردار پشتیبان صورت گرفته و سپس داده‌ها روندزدایی می‌شوند. در نهایت برای هر دو مجموعه داده، برآورد تغییرنگار و پیش‌گویی انجام شده و نتایج پیش‌گویی از طریق آزمون فریدمن مقایسه می‌شوند. مدل‌های تغییرنگار حلقوی، مکعبی، نمایی، گاوسی، گنیتینگ، مترن و برای پارامترهای آستانه، اثر قطعه‌ای و دامنه دو سطح و برای روند دو مدل چندجمله‌ای درجه اول و دوم در نظر گرفته شده است.

واژه‌های کلیدی: رگرسیون بردار پشتیبان، تابع هسته، پیش‌گویی فضایی، تغییرنگار، روند.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M20، 62G08، 62J02.

۱ مقدمه

داده‌های فضایی مشاهداتی هستند که وابستگی آن‌ها ناشی از موقعیت‌شان در فضای مورد مطالعه است و این وابستگی معمولاً تابعی از فاصله مشاهدات از یکدیگر است (کرسی، ۱۹۹۳). معمولاً یک میدان تصادفی مانند $\{Z(s) : s \in R^d, d \geq ۱\}$ به‌عنوان مدل آماری در نظر گرفته شده، که به‌صورت $Z(s) = \mu(s) + \delta(s)$ تجزیه می‌شود، که $\mu(\cdot)$ نشان‌دهنده روند و $\delta(\cdot)$ فرایند خطا است. وجود روند در بین مشاهدات باعث ناماناییی میدان تصادفی می‌شود. در صورت وجود ناماناییی در میانگین، برای برازش تغییرنگار و اجرای پیش‌گویی باید روندزدایی از داده‌ها انجام گیرد. یک روش متداول روندزدایی در نظر گرفتن روند به‌صورت یک ترکیب خطی $\mu(s) = \sum_{j=۰}^{p} \beta_j \phi_j(s)$ است که در آن $(\phi_۰(s), \ldots, \phi_p(s))$ بردار تابع‌هایی معلوم بر حسب موقعیت s هستند. بردار ضرایب نامعلوم $\boldsymbol{\beta} = (\beta_۰, \ldots, \beta_p)$ به یکی از روش‌های کمترین توان‌های دوم

[۱] نام و ایمیل ارائه دهنده مقاله: ساره حدادی، sareh_haddadi@birjand.ac.ir

برآورد می‌شود (محمدزاده ، ۱۳۹۸). روش اصلاح میانه روشی ساده برای روندزدایی است، که توسط کرسی (۱۹۹۳) استفاده شده است. در علم اقتصاد روندزدایی به دو روش رگرسیون خطی ساده و تفاضلی مرتبه اول بسیار پر طرفدار است (سارجنت ، ۱۹۶۸). نحوه‌ی کاهش سطح روند متناسب با روش کمترین توان‌های دوم توسط ویرا و همکاران (۲۰۱۰) مطرح شد. حدادی و اطمینان (۱۴۰۲) برای مدل‌بندی تابع روند از روش رگرسیون بردار پشتیبان ناخطی استفاده کردند. آن‌ها جهت مقایسه نتایج حذف روند به روش رگرسیون بردار پشتیبان، همان نحوه‌ی تحلیل کنت و محمدزاده (۲۰۰۰) بر روی داده‌های کلسیم دنبال کرده و در نهایت نتایج حاصل را با هم مقایسه کردند.

در این مقاله به طریق شبیه‌سازی، عملکرد روندزدایی با استفاده از روش رگرسیون بردار پشتیبان بر روی مجموعه داده فضایی مطالعه می‌شود. ابتدا یک مجموعه داده فضایی از یک میدان تصادفی گاوسی با میانگین صفر شبیه‌سازی شده و سپس براساس یک مدل، روند تولید و به آن اضافه می‌شود. با استفاده از رگرسیون بردار پشتیبان روند در مجموعه روند‌دار مدل‌بندی و سپس داده‌ها روندزدایی می‌شوند. برای بررسی عملکرد شیوه روندزدایی، برای هر دو مجموعه داده اولیه (بدون روند) و روندزدایی شده، برآورد تغییرنگار و سپس پیش‌گویی به روش کریگیدن برای یک مجموعه موقعیت جدید انجام شده و نتایج از طریق آزمون فریدمن مقایسه می‌شوند. مدل‌های تغییرنگار حلقوی، مکعبی، نمایی، گاوسی، گنیتینگ، مترن و برای پارامترهای آستانه، اثر قطعه‌ای و دامنه دو سطح و برای روند دو مدل چندجمله‌ای درجه اول و دوم در نظر گرفته شده است. برای مطالعه انواع مدل‌های نیم‌تغییرنگار و تعیین پارامترهای آن و پیش‌گویی کریگیدن به کرسی (۱۹۹۳) مراجعه شود.

در بخش ۲ رگرسیون بردار پشتیبان معرفی می‌شود. در بخش‌های ۳ و ۴ به‌ترتیب نحوه‌ی روندزدایی از داده‌های فضایی به کمک رگرسیون بردار پشتیبان و نتایج مطالعه شبیه‌سازی آورده شده است.

۲ رگرسیون بردار پشتیبان

روش ماشین بردار پشتیبان در تحلیل مسائل رگرسیونی، رگرسیون بردارپشتیبان [1] (SVR) شناخته می‌شود. SVR از اصل کمینه‌سازی ریسک ساختاری استفاده می‌کند و در نهایت به یک جواب بهینه کلی منجر می‌شود. این مدل مبتنی بر برازش دو ابرصفحه به مجموعه داده‌ها، به‌گونه‌ای است که ابرصفحه‌ها آنقدر از یکدیگر دور باشند که تمام داده‌ها را شامل شوند و از طرفی کمترین تفاوت ممکن بین مقادیر مشاهده شده و پیش‌گویی شده متغیر پاسخ وجود داشته باشند آیزنمن (۲۰۰۸). در رگرسیون بردار پشتیبان، مدل رگرسیونی برابر ابرصفحه دقیقاً وسط دو ابرصفحه حاشیه‌ای و موازی با آن‌ها می‌باشد. داده‌هایی که روی ابرصفحه‌های حاشیه‌ای قرار دارند، بردارهای پشتیبان گویند. فرض کنید $\{(\boldsymbol{x}_i, y_i), i = ۱, ..., n\}$ مجموعه‌ای شامل n نمونه می‌باشد که $\boldsymbol{x}_i \in R^p$ بردار ورودی و y_i خروجی متناظر است. معادله مدل SVR خطی شبیه رگرسیون خطی به‌صورت $f(\boldsymbol{x}, \boldsymbol{\omega}) = \boldsymbol{\omega}\boldsymbol{x} + b$ است که در این رابطه b عرض از مبدأ، $\boldsymbol{\omega}$ بردار پارامترهای مدل و $\boldsymbol{\omega}\boldsymbol{x}$ ضرب داخلی دو بردار را نشان می‌دهد. اگر مدل ارتباطی بین متغیرها پیچیده باشد به‌گونه‌ای که مدل خطی برای برازش به داده‌ها مناسب نباشد با انتقال فضای ورودی اولیه به فضای با ابعاد بالاتر (فضای ویژگی) از طریق نگاشت ناخطی $\phi : \chi \to \zeta$، مدل خطی به داده‌ها در فضای ویژگی برازش داده می‌شود. به‌طوری که در فضای ویژگی می‌توان از یک مدل رگرسیون خطی به‌صورت $f(\boldsymbol{x}, \boldsymbol{\omega}) = \boldsymbol{\omega}\phi(\boldsymbol{x}) + b$ استفاده کرد. برای به‌دست آوردن مقدار برآورد $\boldsymbol{\omega}$ و b بایستی تابع مخاطره در رابطه زیر کمینه شود:

$$R = \|\boldsymbol{\omega}\|^۲/۲ + C/n \sum_{i=۱}^{n} L_e(y_i, f(\boldsymbol{x}_i))$$

[1] Support vector regression

میزان پیچیدگی مدل توسط عبارت $\|\boldsymbol{\omega}\|^{۲}/۲$ کنترل می‌شود و $C/n \sum_{(i=۱)}^{n} L_e(y_i, f(\boldsymbol{x}_i))$ مخاطره تجربی است که توسط زیان غیر حساس ε به‌صورت زیر اندازه‌گیری می‌شود:

$$L_e(y_i, f(\boldsymbol{x}_i)) = \begin{cases} ۰ & |y_i - f(\boldsymbol{x}_i)| < \varepsilon \\ |y_i - f(\boldsymbol{x}_i)| - \varepsilon & \text{جاهای دیگر} \end{cases}$$

در صورتی که اختلاف بین مقدار واقعی و مقدار پیش‌گویی شده کمتر از ε باشد از خطای پیش‌گویی چشم‌پوشی می‌شود و تابع زیان صفر است.

برازش ناخطی به مجموعه‌ی داده‌ها موجب استفاده از تابع هسته در روش رگرسیون بردار پشتیبان می‌شود. از مهمترین دلایلی که در برازش ناخطی از تابع هسته استفاده می‌شود این است که در فضای ویژگی برای $\phi(\boldsymbol{x})$ فرم صریحی در نظر گرفته نمی‌شود. تابع هسته $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$ ضرب داخلی در فضای ویژگی است و از آن برای مقایسه‌ی زوج مجموعه داده‌ها استفاده می‌شود. در صورتی که $\boldsymbol{D} \subset \zeta$ مجموعه داده شامل $\boldsymbol{x}_i, i = ۱, \ldots, n$ باشد، مقایسه زوجی نقاط $\boldsymbol{D}$ از طریق مقادیر ماتریس مربعی $\boldsymbol{K} = [K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{n \times n}$ صورت می‌گیرد. برای محاسبه تابع هسته از همان نقاط فضای ورودی استفاده می‌شود. هسته خطی، چندجمله‌ای، گاوسی و حلقوی متداول‌ترین توابع هسته هستند که در جدول ۱ نشان داده شده‌اند. برای مطالعه بیشتر راجع به رگرسیون بردار پشتیبان، تابع هسته و نحوه تعیین پارامترهای آن به آیزنمن (۲۰۰۸) مراجعه شود.

جدول ۱: توابع هسته متداول.

تابع هسته	$K(\boldsymbol{x}_i, \boldsymbol{x}_j)$
خطی	$\boldsymbol{x}_i{}'\boldsymbol{x}_j$
چندجمله‌ای	$(c + \gamma \boldsymbol{x}_i{}'\boldsymbol{x}_j)^p$
گاوسی	$e^{-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^{۲}}$
حلقوی	$\tanh(c + \gamma \boldsymbol{x}_i{}'\boldsymbol{x}_j)$

۳ روندزدایی به روش رگرسیون بردار پشتیبان

بیان شد که رایج‌ترین روش مدل کردن روند در داده‌های فضایی، در نظر گرفتن تابعی خطی مانند $\mu(\boldsymbol{s}) = \sum_{j=۰}^{p} \beta_j \phi_j(\boldsymbol{s})$ است. از آن جا که ملاکی برای تعیین بردار $(\phi_۰(\boldsymbol{s}), \ldots, \phi_p(\boldsymbol{s}))$ وجود ندارد، ابتدا مجموعه داده‌ها ترسیم شده و سپس توابعی که معمولاً توابع ساده مانند چندجمله‌ای‌ها هستند، برای برازش به مجموعه داده‌ها انتخاب می‌شوند. بنابراین استفاده از رگرسیون بردار پشتیبان در مدل کردن روند باید پیشنهاد مناسبی باشد. اگر برازش مدل خطی برای مجموعه داده‌ها مناسب نباشد استفاده از توابع هسته مختلف امکان به کارگیری مدل‌های مختلف ناخطی را به کاربر می‌دهد. با برازش مدل‌های مختلف به داده‌ها و مقایسه نتایج می‌توان مناسب‌ترین مدل را تعیین کرد. پس از برآورد روند، با کاستن $\hat{\mu}(\boldsymbol{s})$ از $Z(\boldsymbol{s})$، $\hat{\delta}(s)$ به‌دست می‌آید. تغییرنگار تجربی داده‌های روندزدوده محاسبه و از بین مدل‌های معتبر تغییرنگار، مناسب‌ترین مدل برای برازش به تغییرنگار تجربی انتخاب می‌شود. برای پیش‌گویی در نقاط جدید ابتدا پیش‌بینی به روش کریگیدن ساده بر اساس داده‌های بدون روند انجام شده و سپس روند تعیین شده به روش رگرسیون بردار پشتیبان، به آن اضافه می‌شود.

۴ تحلیل داده شبیه‌سازی شده

برای تولید داده‌ها در محیط R از تابع grf در بسته $geoR$ استفاده شده است. مدل به صورت $Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \delta(\boldsymbol{s})$ تعریف شده که $\boldsymbol{s} = (x, y)$ نشان‌دهنده موقعیت مکانی داده‌ها و $\delta(\boldsymbol{s})$ یك میدان تصادفی گاوسی با میانگین صفر است. شش مدل تغییرنگار حلقوی، مکعبی، نمایی، گاوسی، گنیتینگ و مترن و برای پارامترهای آستانه، اثر قطعه‌ای و دامنه دو سطح و برای روند دو مدل چندجمله‌ای درجه اول و دوم،

$$\mu(\boldsymbol{s}) = ۱ + ۰/۶x - ۰/۳y$$

$$\mu(\boldsymbol{s}) = ۱ + ۰/۴x - ۰/۳y - ۰/۷x^۲ + ۰/۶y^۲ + ۰/۵xy$$

در نظر گرفته شده است. هر مجموعه داده شامل نمونه‌ای تصادفی به حجم ۴۰ از یک ناحیه مربعی به طول ۱ از یک میدان تصادفی گاوسی همسانگرد با میانگین صفر است که روند براساس دو مدل فوق به آن اضافه می‌شود.

روند با استفاده از رگرسیون بردار پشتیبان با هسته گاوسی مدل‌بندی و سپس داده‌ها روندزدایی می‌شوند که برای این منظور از بسته $e۱۰۷۱$ در نرم افزار R استفاده شده است. برای بررسی عملکرد شیوه روندزدایی، برای هر دو مجموعه داده اولیه (بدون روند) و روندزدایی شده، برآورد تغییرنگار و سپس پیش‌گویی به روش کریگیدن برای یک مجموعه موقعیت جدید شامل ۲۵ نقطه،

$$\boldsymbol{D} = \{\boldsymbol{s} = (x, y) | x, y \in \{۰/۱, ۰/۳, ۰/۵, ۰/۷, ۰/۹\}\}$$

انجام شده و تفاوت نتایج پیش‌گویی از طریق آزمون ناپارامتریک فریدمن مقایسه می‌شوند. فرضیه صفر این آزمون رتبه‌ای بیانگر یکسان بودن توزیع دو مجموعه (بدون روند و روندزدوده) است؛ و هدف بررسی اثر تعیین مدل روند از طریق رگرسیون بردار پشتیبان است. در جدول‌های ۲ تا ۷، درصد رد نشدن فرض عدم تفاوت نتایج پیش‌گویی در آزمون فریدمن با توجه به سطح معنی‌داری (α) داده شده در ۲۰۰۰ بار تکرار هر ترکیب آورده شده است.

جدول ۲: نتایج آزمون فریدمن در مدل روند چندجمله‌ای مرتبه اول بر اساس مدل تغییرنگار و پارامتر اثر قطعه‌ای.

				مدل‌های تغییرنگار					
$c+c_۰$	$c_۰$	a	α	مکعبی	گاوسی	گنیتینگ	مترن	حلقوی	نمایی
۰/۷	۰/۰۱	۰/۶	$>$ ۰/۰۵	۰/۶۴۲	۰/۳۵۳	۰/۳۴۵	۰/۳۰۶	۰/۵۹۳	۰/۳۱۴
۰/۷	۰/۰۱	۰/۶	$>$ ۰/۱	۰/۵۳۸	۰/۲۸۷	۰/۲۷۲	۰/۲۴۴	۰/۴۹۱	۰/۲۵۹
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۴۵	۰/۳۵۸	۰/۳۵۳	۰/۳۱۴	۰/۵۹۳	۰/۳۲۳
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۴۲	۰/۲۸۳	۰/۲۷۷	۰/۲۵۱	۰/۴۹۳	۰/۲۶۷
۰/۷	۰/۱	۰/۶	$>$ ۰/۰۵	۰/۶۷۲	۰/۳۷۰	۰/۳۷۲	۰/۳۲۷	۰/۶۲۰	۰/۳۲۳
۰/۷	۰/۱	۰/۶	$>$ ۰/۱	۰/۵۷۱	۰/۲۹۴	۰/۳۰۰	۰/۲۶۳	۰/۵۱۶	۰/۲۶۳

در روند چندجمله‌ای مرتبه اول حالتی که اثر قطعه‌ای ($c_۰$) متغیر و برابر ۰/۱ است ؛ مدل مکعبی و حلقوی در سطح معناداری ۰/۰۵ به‌ترتیب در ۶۷ و ۶۲ درصد مواقع اختلاف پیش‌بینی با داده‌های روندزدوده و بدون روند معنادار نیست. بالاترین درصد رد نشدن عدم تفاوت بین دو مجموعه در حالتی که آستانه ($c+c_۰$) متغیر باشد، در مدل مکعبی و در سطح معناداری ۰/۰۵ دیده می‌شود؛ این مقدار با آستانه ۰/۷ برابر با ۰/۶۴ و با آستانه ۱ برابر با ۰/۶۳ درصد می‌باشد. در حالتی

جدول ۳: نتایج آزمون فریدمن در مدل روند چندجمله‌ای مرتبه اول بر اساس مدل تغییرنگار و پارامتر آستانه .

				مدل‌های تغییرنگار					
$c+c_\cdot$	$c_\cdot$	a	α	مکعبی	گاوسی	گنیتینگ	مترن	حلقوی	نمایی
۰/۳	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۲۹	۰/۳۵۴	۰/۳۶۷	۰/۲۹۷	۰/۵۷۰	۰/۳۰۰
۰/۳	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۲۷	۰/۲۸۲	۰/۲۹۵	۰/۲۴۴	۰/۴۶۴	۰/۲۴۰
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۴۶	۰/۳۵۲	۰/۳۵۳	۰/۳۲۰	۰/۵۹۴	۰/۳۲۵
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۳۶	۰/۲۷۲	۰/۲۷۷	۰/۲۵۴	۰/۵۰۱	۰/۲۶۳
۱	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۳۹	۰/۳۵۶	۰/۳۷۴	۰/۳۲۱	۰/۶۱۲	۰/۳۳۹
۱	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۳۸	۰/۲۷۷	۰/۳۰۴	۰/۲۵۵	۰/۵۰۶	۰/۲۶۸

جدول ۴: نتایج آزمون فریدمن در مدل روند چندجمله‌ای مرتبه اول بر اساس مدل تغییرنگار و پارامتر دامنه .

				مدل‌های تغییرنگار					
$c+c_\cdot$	$c_\cdot$	a	α	مکعبی	گاوسی	گنیتینگ	مترن	حلقوی	نمایی
۰/۷	۰/۰۵	۰/۲	$>$ ۰/۰۵	۰/۸۳۵	۰/۶۷۱	۰/۶۷۱	۰/۵۵۱	۰/۸۱۴	۰/۵۴۶
۰/۷	۰/۰۵	۰/۲	$>$ ۰/۱	۰/۷۴۳	۰/۵۸۳	۰/۵۶۹	۰/۴۳۷	۰/۷۱۹	۰/۴۳۵
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۴۳	۰/۳۵۸	۰/۳۵۵	۰/۳۱۳	۰/۵۸۶	۰/۳۱۳
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۳۹	۰/۲۸۳	۰/۲۸۲	۰/۲۵۰	۰/۴۸۷	۰/۲۴۹
۰/۷	۰/۰۵	۱	$>$ ۰/۰۵	۰/۴۸۴	۰/۲۴۹	۰/۲۳۲	۰/۲۳۷	۰/۴۳۷	۰/۲۵۱
۰/۷	۰/۰۵	۱	$>$ ۰/۱	۰/۳۹۵	۰/۲۱۰	۰/۱۸۲	۰/۱۹۰	۰/۳۶۸	۰/۱۹۳

جدول ۵: نتایج آزمون فریدمن در مدل روند چندجمله‌ای مرتبه دوم بر اساس مدل تغییرنگار و پارامتر اثر قطعه‌ای.

				مدل‌های تغییرنگار					
$c+c_\cdot$	$c_\cdot$	a	α	مکعبی	گاوسی	گنیتینگ	مترن	حلقوی	نمایی
۰/۷	۰/۰۱	۰/۶	$>$ ۰/۰۵	۰/۶۱۸	۰/۳۵۱	۰/۳۴۸	۰/۳۰۲	۰/۵۹۳	۰/۳۲۴
۰/۷	۰/۰۱	۰/۶	$>$ ۰/۱	۰/۵۱۲	۰/۲۸۴	۰/۲۸۵	۰/۲۳۶	۰/۴۸۱	۰/۲۵۵
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۳۶	۰/۳۶۷	۰/۳۶۴	۰/۳۰۳	۰/۵۸۰	۰/۳۲۸
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۳۶	۰/۳۰۴	۰/۳۰۳	۰/۲۴۹	۰/۴۸۴	۰/۲۵۱
۰/۷	۰/۱	۰/۶	$>$ ۰/۰۵	۰/۶۳۸	۰/۳۴۶	۰/۳۴۹	۰/۳۵۳	۰/۶۲۴	۰/۳۶۱
۰/۷	۰/۱	۰/۶	$>$ ۰/۱	۰/۵۳۲	۰/۲۸۶	۰/۲۸۶	۰/۲۸۰	۰/۵۱۱	۰/۲۸۶

که دامنه (a) متغیر و کمترین مقدار (۰/۲) باشد در تمامی مدل‌ها و سطوح معناداری، تفاوت بین دو مجموعه نسبت به سایر دامنه‌ها کمترین است؛ همان طور که در جدول ۴ مشاهده می‌شود، در دامنه کوچک ۰/۲ و در سطح معناداری ۰/۰۵

جدول ۶: نتایج آزمون فریدمن در مدل روند چندجمله‌ای مرتبه دوم بر اساس مدل تغییرنگار و پارامتر آستانه .

				مدل‌های تغییرنگار					
$c+c_0$	c_0	a	α	مکعبی	گاوسی	گنیتینگ	مترن	حلقوی	نمایی
۰/۳	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۲۹	۰/۳۶۸	۰/۳۶۷	۰/۳۰۶	۰/۵۶۹	۰/۳۰۹
۰/۳	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۲۶	۰/۳۰۷	۰/۳۰۸	۰/۲۵۱	۰/۴۷۵	۰/۲۴۹
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۵۳	۰/۳۵۹	۰/۳۵۸	۰/۳۰۱	۰/۶۱۰	۰/۳۲۰
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۳۶	۰/۲۹۴	۰/۲۹۲	۰/۲۳۶	۰/۴۹۶	۰/۲۵۹
۱	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۴۰	۰/۳۶۹	۰/۳۶۶	۰/۳۲۴	۰/۶۱۷	۰/۳۲۲
۱	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۲۸	۰/۲۹۳	۰/۲۹۳	۰/۲۶۳	۰/۵۱۲	۰/۲۵۷

درصد رد نشدن عدم تفاوت بین دو مجموعه برای تمامی مدل‌ها بین بازه (۸۳ − ۵۴) درصد قرار دارد. کمترین و بیشترین درصد به‌ترتیب مربوط به مدل نمایی و مکعبی است. در همین سطح معناداری در صورتی که بیشترین مقدار دامنه ۱ در نظر گرفته شود، درصدها بین بازه (۴۸ − ۲۳) قرار دارد؛ که کمترین و بیشترین درصد به‌ترتیب متعلق به مدل گنیتینگ و مکعبی است.

جدول ۷: نتایج آزمون فریدمن در مدل روند چندجمله‌ای مرتبه دوم بر اساس مدل تغییرنگار و پارامتر دامنه .

				مدل‌های تغییرنگار					
$c+c_0$	c_0	a	α	مکعبی	گاوسی	گنیتینگ	مترن	حلقوی	نمایی
۰/۷	۰/۰۵	۰/۲	$>$ ۰/۰۵	۰/۸۴۳	۰/۶۷۳	۰/۶۷۸	۰/۵۴۹	۰/۸۰۴	۰/۵۳۲
۰/۷	۰/۰۵	۰/۲	$>$ ۰/۱	۰/۷۵۶	۰/۵۷۲	۰/۵۷۱	۰/۴۵۸	۰/۷۲۹	۰/۴۴۵
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۰۵	۰/۶۳۱	۰/۳۵۸	۰/۳۵۶	۰/۲۹۱	۰/۶۰۹	۰/۳۰۴
۰/۷	۰/۰۵	۰/۶	$>$ ۰/۱	۰/۵۳۴	۰/۲۸۳	۰/۲۷۹	۰/۲۳۱	۰/۵۰۳	۰/۲۴۳
۰/۷	۰/۰۵	۱	$>$ ۰/۰۵	۰/۴۹۲	۰/۲۲۴	۰/۲۲۴	۰/۲۷۲	۰/۴۱۳	۰/۲۴۰
۰/۷	۰/۰۵	۱	$>$ ۰/۱	۰/۴۰۱	۰/۱۸۴	۰/۱۸۳	۰/۲۱۹	۰/۳۲۷	۰/۱۹۳

در روند چندجمله‌ای مرتبه دوم در حالتی که اثر قطعه‌ای متغیر و برابر ۰/۱ است مدل مکعبی در سطح معناداری ۰/۰۵ در ۶۳ درصد مواقع بین پیش‌گویی در مجموعه‌ای که در آن روندزدایی از طریق رگرسیون بردار پشتیبان صورت گرفته با مجموعه بدون روند تفاوت نشان نمی‌دهد؛ با تغییر اثر قطعه‌ای به ۰/۰۵ ، مجدداً در همین سطح معناداری و مدل بیشترین درصد مشاهده می‌شود. با تغییر آستانه به ۰/۷ کمتربن درصد رد نشدن عدم تفاوت بین دو مجموعه در سطح معناداری ۰/۱ متعلق به مدل مترن با ۰/۲۳ درصد و بیشترین درصد متعلق به مدل مکعبی در سطح معناداری ۰/۰۵ برابر با ۰/۶۵ است. همچنین در این مدل روند، در نظر گرفتن دامنه کوچک ۰/۲ موجب شده که مدل مکعبی در سطح معناداری ۰/۰۵ در ۰/۸۴ درصد مواقع بین نتایج پیش‌گویی دو مجموعه تفاوتی لحاظ نکند.

به‌طور کلی در روند چندجمله‌ای مرتبه اول و دوم نتایج سطح معنی‌داری آزمون فریدمن نشان می‌دهد، که درصد رد نشدن عدم تفاوت بین دو مجموعه روندزدوده و بدون روند در مدل مکعبی و حلقوی نسبت به سایر مدل‌ها بیشتر است. به عبارتی

رگرسیون بردار پشتیبان در این دو مدل به‌طور نسبتاً موفقتری قادر بوده که مدل مناسب برای روند را تشخیص دهد. از طرفی، تغییرات اثر قطعه‌ای و آستانه در سطوح مختلف معناداری، در نتایج تغییر چشمگیری به وجود نمی‌آورد؛ بالعکس اثر دامنه زیاد است.

بحث و نتیجه‌گیری

برای از بین بردن اثر نامطلوب روند بر برآورد تغییرنگار و پیش‌گویی باید ابتدا روند کشف و مدل شود. هدف از تحلیل مجموعه داده شبیه‌سازی شده بررسی عملکرد رگرسیون بردار پشتیبان در تشخیص و مدل کردن روند است. نتایج شبیه‌سازی با حجم ۴۰ نشان داد که روش رگرسیون بردار پشتیبان در مدل‌های تغییرنگار حلقوی و مکعبی در تشخیص مدل مناسب برای روند موفقتر عمل می‌کنند. زمانی که اثر قطعه‌ای یا آستانه متغیر باشند، در نتایج مدل‌های تغییرنگار در سطوح مختلف تغییر چشمگیری مشاهده نمی‌شود؛ یعنی تغییرنگار رفتار استواری دارد. از طرفی، در نظر گرفتن دامنه کوچکتر منجر به نتایج بهتر برای تعیین روند از طریق رگرسیون بردار پشتیبان می‌شود. لازم به ذکر است که بر طبق نتایج به‌دست آمده شاید خیلی نتوان مناسب بودن روش پیشنهادی را تائید کرد، مشاهده همبستگی ضعیف‌تر داده‌های روندزدوده نسبت به داده‌های اولیه در مدل کردن از طریق رگرسیون بردار پشتیبان، در نظر گرفتن تکرار و حجم نمونه کم و شاید نامناسب بودن آزمون فریدمن برای مقایسه از جمله دلایلی است که مستلزم مطالعات گسترده‌تری در این زمینه است؛ همچنین روش پیشنهادی بایستی با سایر روش‌های روندزدایی مقایسه شود تا بهتر بتوان در مورد کارآیی روش قضاوت کرد.

مراجع

حدادی، س. و اطمینان، ج. (۱۴۰۲)، روندزدایی در آمار فضایی با استفاده از رگرسیون بردار پشتیبان، مجله علوم آماري ایران.

محمدزاده، م. (۱۳۹۸)، آمار فضایی و کاربردهای آن، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

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

Izenman, A.J. (2008), *Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning*, Springer-Verlag, New York.

Kent, J.T. and Mohammadzadeh, M. (2000), Global Optimization of the Generalized Cross-Validation Criterion, *Statistics and Computing*, **10**, 231–236.

Sarjent, T.J. (1968), Interest Rate in the Nineteen-Fifties, *Review of Economics and Statistics*, **50**, 164-172.

Vieira, S.R., de Carvalho, J.R.P., Ceddia, M.A. and Gonzalez, A.P. (2010), Detrending Non Stationary Data for Geostatistical Applications, *Bragantia*, **69**, 1-8.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

تحلیل فضایی گرد و غبار و ارتباط آن با خشکسالی در استان سیستان و بلوچستان

احمد حسینی[۱]
گروه آب و هواشناسی، دانشگاه پیام نور،ایران

چکیده: در مطالعات سازمان هواشناسی جهانی بادها با سرعت بیش از ۱۵ متر برثانیه (۳۰نات) و دید افقی به علت گرد و خاک زیر ۱۰۰۰ متر به عنوان طوفان گرد و غبار SDS یا storm sand and Dust شناخته می‌شود. ابتدا پس از کنترل کمّی و کیفی داده‌ها، از روش IDW با کمترین میانگین مربعات خطا برای روزهای گرد و غباری و از روش کرجینک ساده پدیده خشکسالی استفاده شد.پهنه‌بندی روزهای گرد و غباری نشان داد بیشترین تعداد در ایستگاه زابل و زهک قابل مشاهده است و به سمت نواحی جنوبی استان سیستان و بلوچستان از آن کاسته می‌شود، پهنه‌بندی خکشسالی نشان می‌دهد نواحی مرکزی استان سیستان و بلوچستان مانند ایستگاه ایرانشهر بیشترین خشکسالی‌ها را دارا است اما تعداد روزهای SDS آن نسبت به ایستگاه زابل کمتر است.تحلیل فضایی داده‌ها نشان داد که پدیده خشکسالی در وقوع طوفان‌های گرد و غبار آنچنان اثر گذار نبوده، بلکه خشکی و کمبود رطوبت هوا باعث گسترش پدیده SDS در این منطقه می‌شود و پدیده خشکسالی آن را تشدید می‌نماید.

واژه‌های کلیدی: استان سیستان و بلوچستان، آمار فضایی، خشکسالی، روزهای گرد و غباری.SDS
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 62H11، 62G08.

---

۱ مقدمه

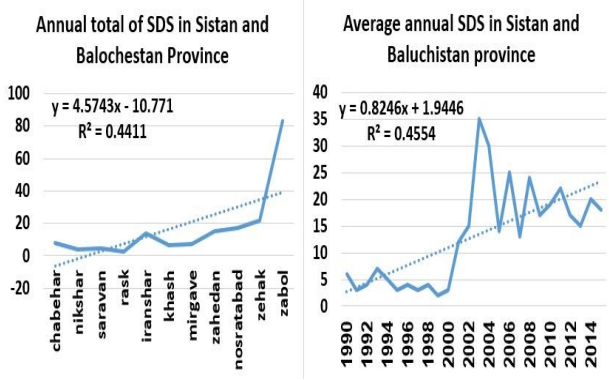
طبق تعریف سازمان هواشناسی جهانی، چنانچه سرعت باد به بیش از ۳۰ نات در ثانیه و دید افقی به‌علت گردوخاک حاصله از آن به کمتر از یک کیلومتر باشد، پدیده موجود طوفان گردوخاک نامیده می‌شود. که به آن پدیده storm Sand گفته می‌شود (سازمان جهانی هواشناسی ، ۲۰۱۷). همچنین چنانچه گردوخاک گسترده‌ای که در اثر طوفان گردوخاک از سایر نقاط دور به ایستگاه آمده باشد و بصورت معلق باعث کاهش دید قائم شده باشد که به عنوان ریز گرد شناخته می‌شود به آن storm Dust گفته می‌شود (سازمان جهانی هواشناسی ، ۲۰۱۵). که از نظر سازمان هواشناسی جهانی به این پدیده‌ها SDS یا storm Dust and sand گفته می‌شود که مکانیزم آن به "مجموعه‌ای از ذرات گردوغبار یا شن و ماسه که توسط یک باد قوی و متلاطم با انرژی به ارتفاعات بالاتر برده می شود" تعریف می شود. مجمع عمومی سازمان ملل متحد و کنگره‌های

[۱] نام و ایمیل ارائه دهنده مقاله: احمد حسینی، ahmad-hossayni@yahoo.com

هواشناسی جهانی طوفان‌های گردوغبار (SDS) را به‌عنوان خطرات شدیدی که می‌تواند بر آب و هوا، محیط زیست، سلامت و اقتصاد، در بسیاری از نقاط جهان اثرگذار است را به عنوان هشدار تلقی می‌کنند (الیفسیریو و همکاران ، ۲۰۲۳) در منطقه خاورمیانه گردوغبار با سرعت زیاد بر فراز عراق، ایران، کویت، پادشاهی عربستان سعودی و بحرین منطقه را تحت تاثیر قرار می‌دهند (سازمان جهانی هواشناسی ، ۲۰۱۸). بنا به گزارش سازمان بهداشت جهانی سالانه ۳/۸ تا ۴/۲ میلیون مرگ زودرس با قرار گرفتن در معرض آلودگی هوای محیطی مرتبط است (سازمان بهداشت جهانی ، ۲۰۱۶) که رقم بالایی را نشان می‌دهد. استان سیستان و بلوچستان بیش از ۵ میلیون هکتار بیابان دارد که ۶ درصد آن یعنی حدود (۸۰۰ هزار هکتار) جزء شن‌زارهای فعال محسوب می‌شود (لطیفی ، ۱۳۸۵). که با هجوم خود به اراضی کشاورزی، انهار، راه‌های ارتباطی، شهرها و روستاها و تأسیات اقتصادی و حیاتی منطقه، خسارات جبران‌ناپذیری به آنها در چند دهه‌ی اخیر وارد نموده است (احمدیان ، ۱۳۷۸). در برآورد صورت گرفته از شدت پتانسیل باد در ۶۰ ایستگاه هواشناسی کشور، ایستگاه زابل بیشترین مقدار فراوانی و سرعت باد را به خود اختصاص داده است که رتبه دوم وقوع طوفان‌های ماسه‌ای در قارة آسیا را دارا است (اداره کل منابع طبیعی استان س و ب ، ۱۳۸۳). ما در این تحقیق به‌دنبال این هستیم که آیا تحلیل فضایی به‌همراه پهنه‌بندی روزهای گردوغبار قادر خواهد بود بحرانی بودن این مساله را نشان دهد یا روش‌های آماری فضایی می‌تواند پراکنش فضایی پدیده گردوغباری را با مساله خشکسالی تبین کند.

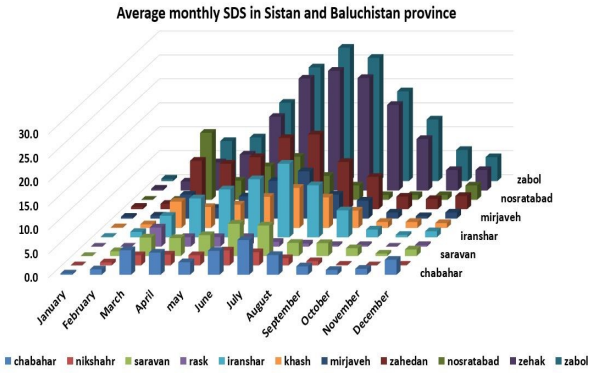
۲ داده‌ها و روش کار

استان سیستان و بلوچستان با ۱۱/۵ درصد مساحت کشور در جنوب شرقی کشور واقع شده است. نمودار شماره یک نشان می‌دهد تعداد روزهای گردوغباری در استان سیستان و بلوچستان از جنوب به شمال افزایش می‌یابد ، سری‌زمانی میانگین سرعت باد سالانه در طول دوره آماری ۱۹۹۰ تا ۲۰۱۵ حاکی از افزایش SDS در دراز مدت می‌باشد.



شکل ۱ : میانگین سالیانه روزهای گرد و غباری در استان سیستان و بلوچستان

همچنین با توجه به نمودار شماره (۲) از نظر فضایی ایستگاه زابل بیشترین روزهای گرد و غباری در اکثر ماه‌های سال را دارد و بطور میانگین در ماه‌های ژوئن، ژولای و اگوست به‌ترتیب با ۲۸ ، ۲۶ و ۱۹ روز با گرد و غبار همراه بوده است و پس از آن ایستگاه زاهدان قرار دارد و ایستگاه های: نیکشهر، راسک، سراوان و چابهار دارای کمترین روزهای گرد و غباری در سطح استان هستند.



شکل ۲: میانگین ماهیانه روزهای گرد وغباری در استان سیستان و بلوچستان

۱.۲ تحلیل فضایی داده‌ها

در این تحقیق پس از اخذ داده‌های تعداد روزهای گرد و غباری SDS و مجموع بارش ماهیانه RM داده‌ها از اداره کل هواشناسی سیستان و بلوچستان تهیه، مورد بررسی واقع شد تا از صحت داده‌ها اطمینان حاصل شود. داده‌های روزانه دریافتی در نرم‌افزار Excel وGIS مورد پردازش قرار گرفت. همچنین جهت مشخص کردن فراوانی روزهای گرد و غباری در طی دوره زمانی مورد مطالعه، از نرم‌افزار SPSS واز روش تحلیل خوشه‌ای با استفاده از تحلیل سلسله مراتبی استفاده شد و تعداد روزهای همراه با طوفان از نظر فراوانی ماهانه طبقه‌بندی شدند. با توجه به کمبود آمار دوره آماری ۲۵ ساله (۱۹۹۰ تا ۲۰۱۵) برای تحلیل روزهای گرد و غباری و از ابتدای تاسیس تا سال ۲۰۱۵ برای بارش انتخاب گردید. جهت انتخاب مناسب‌ترین شاخص بهینه خشکسالی برای تحلیل داده‌های بارش از شاخص دهک‌ها DPI و شاخص بارش استاندارد SPI استفاده شد. مطالعات در این خصوص نشان می‌دهد شاخص دهک‌ها برای دوره‌های کوتاه مدت و شاخص بارش استاندارد برای دوره‌های طولانی‌تر، نتایج قابل قبول‌تری ارائه می‌کند (حجازی زاده ، ۱۳۸۹). اصول کلي در محاسبه دهك‌ها به صورت زير بررسی شد. ۱ ـ مرتب نمودن داده‌هاي بارندگي ماهانه به صورت صعودي ۲ ـ تعيين دامنه دهكي با استفاده از رابطه $D_i = i \times \frac{n+1}{10}$. ۳ ـ برآورد مقادير بارندگي مربوط به هر دهك (حد انتهايي) ۴ ـ تعيين سال‌هاي آماري كه در دهك‌هاي مختلف قرار گرفته‌اند. داده‌هاي اصلي شاخص بارش استاندارد، داده‌هاي مجموع بارندگي ماهانه مي‌باشد. پس از اطمينان از همگن بودن و تصادفي بودن داده‌هاي ماهانه، سري‌زماني در بازه‌هاي زماني ۳، ۶، ۹، ۱۲، ۲۴ و ۴۸ ماهه تشكيل داده شد. هر يك از سري‌هاي زماني با مقياس‌هاي زماني متفاوت با توزيع‌هاي مختلف برازش داده شد و بهترين توزيع گاما و پيرسون شناخته شد و پس از بررسی توزيع آماري گاما برازش خوبي با سري‌زماني اقليمي بارندگی نشان داد توزيع گاما به صورت تابع چگالي احتمال يا فراواني به صورت

$$g(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}}$$

است، كه در آن $a > 0$ پارامتر شكل و $B > 0$ پارامتر مقياس و $x > 0$ مقدار بارندگي و $\Gamma(\alpha)$ تابع گاما است. محاسبه شاخص بارش استاندارد با برازش تابع چگالي احتمال گاما بر توزيع فراواني بارندگي براي تک تک ايستگاه‌ها از طريق نرم‌افزار مربوطه محاسبه شد كه نتايج حاصله از دو روش DPI و SPI در جداول ۴ و ۵ آمده است. سپس جهت تحلیل فضایی تعداد روزهای گرد و غباری و ارتباط آن با میزان خشکسالی، میانگین مربعات خطا در هر کدام از روش‌های زمین‌آماری محاسبه شد که در نهایت روش‌آماری که با کمترین میزان مربعات خطا همراه بود جهت تحلیل فضایی داده‌ها انتخاب شد،

$$MSE = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n}$$

که در آن $(\hat{y}_i - \bar{y})^2$ مقدار مربع خطای هر داده است.

۲.۲ شرح و تفسیر نتایج

جهت تعیین نوع پراکنش داده ها، انواع مدل های زمین آماری جهت پهنه بندی تعداد روزهای گرد وغباری و اثرات خشکسالی بر آن مورد مقایسه قرار گرفتندکه نتایج آن در جدول ۱ آمده است.

جدول ۱: محاسبه حداقل مربعات خطا تعداد روزهای گرد وغباری

SDS	SPI	DPI	mode statistical Spatial
۲۵/۲	۷/۸	۱۲/۳	Weighted Distance Inverse
۳۱/۶	۸/۳۵	۱۳/۷	interpolation polynomial Global
۳۰/۲	۹/۴۴	۱۵	Interpolation Polynomial Local
۳۲/۱	۶/۴	۱۰/۱	Kriging Ordinary
۲۷/۴	۵/۸	۹/۳	kriging Simple
۳۲/۱	۶/۴	۱۰/۸	kriging Universal
۲۷/۹	۶/۶	۱۰/۵	kriging Bayesian Empirical
۳۲	۱۰/۴	۱۶/۴	Smoothing Kernel
۳۵	۶/۹	۱۱/۱	kernel diffusion

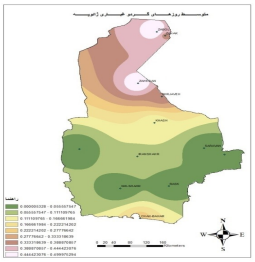
نتایج نشان می‌دهد که روش IDW برای روزهای گرد و غباری SDS و کرجینگ ساده برای دوره‌های خشکسالی DPI و SPI کمترین میزان مربعات خطا را دارد که جهت استفاده از نتایج مطلوب از آن استفاده شد.

که با توجه به شکل‌های ۱ تا ۱۴ (شکل ۳) برای روزهای گرد و غباری نتایج زیر به‌دست آمده است.

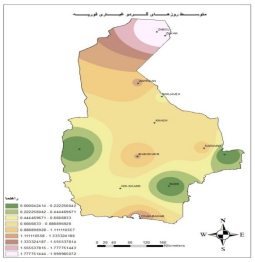
جدول ۲: تحلیل فضایی روزهای گرد و غباری استان سیستان و بلوچستان

ماه	پرکنش فضایی روزهای گرد و غباری
ژانویه	زابل و نواحی جنوبی استان
فوریه	زابل و نواحی جنوبی استان
مارس	زاهدان و زابل و نواحی جنوبی استان
آوریل	زاهدان و زابل و نواحی جنوبی استان
می	زابل و نواحی جنوبی استان
ژوئن	زابل و نواحی جنوبی استان
جولای	زابل و نواحی جنوبی استان
آگوست	زابل و نواحی جنوبی استان
سپتامبر	زابل و نواحی جنوبی استان
اکتبر	زابل و نواحی جنوبی استان
نوامبر	زابل و نواحی جنوبی استان
دسامبر	زابل و نواحی جنوبی استان

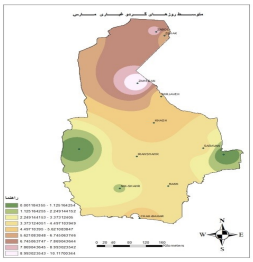
به‌طور کلی بیشترین روزهای گرد و غباری در زابل و نواحی جنوبی آن دیده می‌شود. جدول ۲ نشان می‌دهد که ماه‌های مارس و آوریل زاهدان از زابل پیشی گرفته است. علت آن را می‌توان همزمان با خروج جبهه قطبی و کاهش رطوبت منطقه و در نتیجه افزایش تعداد روزهای گرد و غباری در این منطقه دانست که این موضوع در نواحی شمال غربی ایران خود را تحت عنوان بارش‌های نیسان نشان می‌دهد (علیجانی ، ۱۳۸۰). همچنین متوسط سالیانه تعداد روزهای گرد و غباری نشان می‌دهد که از زابل تا نواحی جنوبی یعنی تا ایرانشهر بیشترین روزهای گرد و غباری را دارد. علت عمده آن را می‌توان کمبود رطوبت و تبدیل شدن سیکلون‌های غربی به مراکز فروبار همراه با رشد طوفان‌های گرد و خاک در این مناطق دانست (علیجانی ، ۱۳۹۰). این موضوع در شکل ۳ شماره (۱۴) با پهنه‌بندی روزهای گرد و غباری بیشتر خود را نشان می‌دهد. به‌طوری که



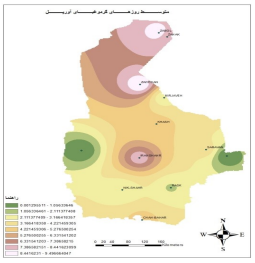
(۱) متوسط روزهای گرد و غباری ژانویه



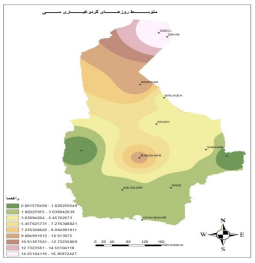
(۲) متوسط روزهای گرد و غباری فوریه



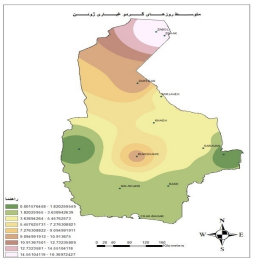
شکل (۳) متوسط روزهای گرد و غباری مارس



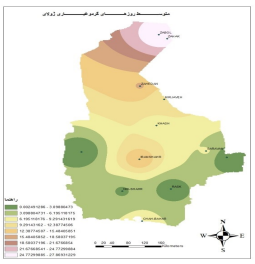
(۴) متوسط روزهای گرد و غباری آوریل



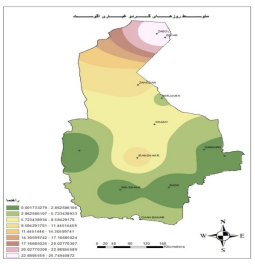
(۵) متوسط روزهای گرد و غباری می



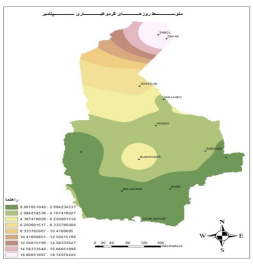
(۶) متوسط روزهای گرد و غباری ژوئن



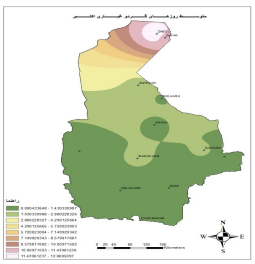
(۷) متوسط روزهای گرد و غباری ژولای



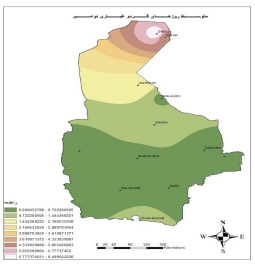
(۸) متوسط روزهای گرد و غباری آگوست



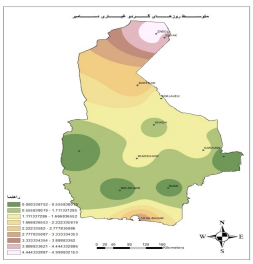
(۹) متوسط روزهای گرد و غباری سپتامبر



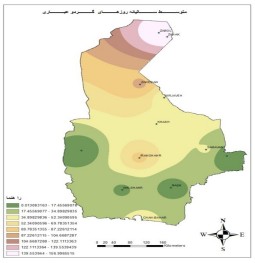
(۱۰) متوسط روزهای گرد و غباری اکتبر



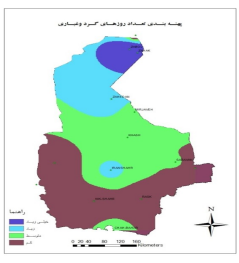
(۱۱) متوسط روزهای گرد و غباری نوامبر



(۱۲) متوسط روزهای گرد و غباری دسامبر



(۱۳) متوسط سالیانه روزهای گرد و غباری



(۱۴) پهنه‌بندی روزهای گرد و غباری

شکل ۳: ۱۴ – ۱

نواحی مرکزی استان سیستان و بلوچستان به‌دلیل نفوذ سامانه رطوبتی مونسون در فصل گرم نسبت به زابل روزهای گرد و غباری کمتری را نشان می‌دهد (علیجانی ، ۱۳۹۰).

۳.۲ بررسی خشکسالی های منطقه

شاخص دهک بارندگی DPI : با توجه به خروجي‌هاي به‌دست آمده می‌توان اذعان داشت که بر اساس شاخص دهک‌ها، تعداد سال‌های بیشتری با خشکسالی ضعیف مواجه بوده‌اند، خشکسالی متوسط نیز در بیشتر مناطق استان تقریبا یکسان بوده است. خشکسالی شدید در ایستگاه‌های ایرانشهر، زابل، زاهدان و سراوان اتفاق افتاده است. در مجموع بیشترین فراوانی وقوع خشکسالی در ایستگاه‌های چابهار،ایرانشهر، خاش، سراوان، زابل و زاهدان دیده می‌شود.

جدول ۳: فراوانی انواع خشکسالی بر اساس شاخص دهک بارندگی (DPI) (تا سال ۲۰۱۵)

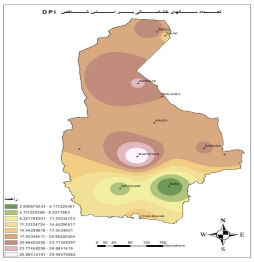
ردیف	ایستگاه	نرمال	خشکسالی ضعیف	خشکسالی متوسط	خشکسالی شدید	خشکسالی خیلی شدید	مجموع خشکسالی
۱	چابهار	۶	۸	۲	–	–	۱۶
۲	ایرانشهر	۱۱	۱۲	۶	۲	–	۳۰
۳	خاش	۸	۷	۳	–	–	۱۸
۴	کنارک	۱	۳	–	–	–	۴
۵	نیکشهر	۲	۴	۱	–	–	۷
۶	راسک	–	۱	۱	–	–	۲
۷	سراوان	۶	۷	۶	۲	–	۲۱
۸	سرباز	۱	–	۱	۱	–	۳
۹	زابل	۶	۶	۸	۳	–	۲۳
۱۰	زهک	۳	۷	۶	–	–	۱۶
۱۱	زاهدان	۴	۷	۱۰	۳	–	۲۴

شاخص بارش استاندارد شده : نتایج به‌دست آمده از شاخص بارش استاندارد (SPI) نشان می‌دهد که خشکسالی ضعیف در تمامی ایستگاه‌ها اتفاق افتاده است. خشکسالی متوسط در ایستگاه‌های ایرانشهر، خاش، زابل و زاهدان دیده می‌شود به طور کلی ایستگاه‌های چابهار، ایرانشهر، خاش، سراوان، زابل و زاهدان بیشترین فراوانی وقوع را دارند.

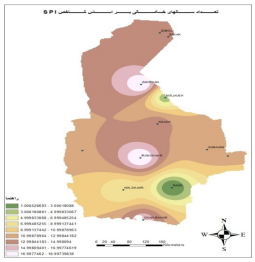
جدول ۴: فراوانی انواع خشکسالی‌های شاخص بارش استاندارد (SPI) (تا سال ۲۰۱۵)

ردیف	ایستگاه	نرمال	خشکسالی ضعیف	خشکسالی متوسط	خشکسالی شدید	خشکسالی خیلی شدید	مجموع خشکسالی
۱	چابهار	۶	۸	–	۲	–	۱۶
۲	ایرانشهر	۵	۷	۷	۱	–	۲۰
۳	خاش	۷	۷	۳	–	–	۱۷
۴	کنارک	۱	۳	–	–	–	۴
۵	نیکشهر	۲	۴	۱	–	–	۷
۶	راسک	۱	۱	–	–	–	۲
۷	سراوان	۷	۶	–	–	–	۱۵
۸	سرباز	۱	۱	۱	۲	–	۳
۹	زابل	۶	۸	۳	–	–	۱۷
۱۰	زهک	۶	۷	–	–	–	۱۳
۱۱	زاهدان	۷	۱۰	۳	–	–	۲۰

در مجموع نتایج شاخص DPI و SPI در منطقه مورد مطالعه مبین آن است که وقوع و تکرار پدیده خشکسالي هر چند سال یک‌بار با توجه به شرایط منطقه دور از احتمال نیست و در تمام نواحی استان، از مناطق پر بارش و رگباری جنوبی (چابهار) تا مناطق کم‌بارش زابل و ایرانشهر دیده می‌شود. به‌طوری که با هر دو روش از لحاظ وقوع پدیده خشکسالی ایستگاه‌های ایرانشهر، زاهدان، زابل، خاش، چابهار و سراوان دارای بیشترین فراوانی بوده‌اند.



(۱۶) تعداد سال‌های خشکسالی بر اساس روش DPI



(۱۷) تعداد سالهای خشکسالی بر اساس روش SPI

بحث و نتیجه‌گیری

به‌طور کلی پهنه‌بندی تعداد روزهای گردوغباری نشان می‌دهد بیشترین روزهای گردوغباری در ایستگاه زابل و زهک قابل مشاهده است که به‌سمت نواحی جنوبی استان سیستان و بلوچستان کشیده و سپس به‌تدریج از آن کاسته می‌شود اما پهنه‌بندی تعداد سال‌های خشکسالی نشان می‌دهد که نواحی مرکزی استان سیستان و بلوچستان مانند ایستگاه ایرانشهر دارای بیشترین خشکسالی‌ها را دارا است ولی تعداد روزهای گردوغباری آن نسبت به ایستگاه زابل کمتر است که می‌توان نتیجه گرفت خشکسالی تنها در وقوع طوفان‌های گردوغبار اثرگذار نبوده، بلکه خشکی و کمبود رطوبت هوا باعث گسترش پدیده SDS در این استان می‌شود، با این وجود بنا به گزارشات اداره کل هواشناسی سیستان و بلوچستان میزان آلودگی حاصله از فرآیند SDS در ۱ فوریه ۲۰۱۳ به ۲۴ هزار میکروگرم در متر مکعب در شهرهای شمالی استان خصوصاً زابل و زهک رسیده است که با توجه به استاندارد جهانی که ۱۵۰ میلی گرم در هر متر مکعب در روز تعیین شده است، این رقم ۸۰ برابر استاندارد بین‌المللی است که ناشی از خشکی منطقه و خشکسالی‌ها و آنومالی‌های شدید اقلیمی است که گه‌گاه سرعت باد در فصل زمستان تا ۹۰ کیلومتر در ساعت می‌رسد در نهایت می‌توان گفت خشکی هوا در بروز پدیده SDS بیشتر اثر گذار بوده و بروز پدیده خشکسالی آن را تشدید نموده است.

مراجع

احمدیان، م .ع، (۱۳۷۸)، بیابان (نگرش سیستمی به فرآیند بیابان زایی و بیابان زدایی)، *فصلنامه تحقیقات جغرافیایی*، شماره پیاپی ۵۲ و ۵۳.

اداره کل منابع طبیعی استان سیستان و بلوچستان طرح اجرای تثبیت شن و بیابان‌زدایی زابل در سال های ۱۳۸۲ و ۱۳۸۳ .

اداره کل هواشناسی استان سیستان و بلوچستان، (۱۳۸۳) آمار روزهای گرد و غباری و بارش ماهیانه.

حجازی زاده،ز و س،جوی زاده ،۱۳۸۹ ، مقدمه‌ای بر خشکسالی و شاخص‌های آن، سازمان مطالعه و تدوین کتب علوم انسانی (سمت) شماره ۲۲۸ .

علیجانی،ب (۱۳۸۰)، *"آب وهوای ایران"*، انتشارات دانشگاه پیام نور .

علیجانی،ب (۱۳۹۰) مبانی آب و هواشناسی، سازمان مطالعه و تدوین کتب علوم انسانی (سمت).

لطیفی، ل، (۱۳۸۵)، *بررسی روند پیشروی تپه‌های ماسه‌ای* با استفاده از تصاویر ماهواره‌ای در طی خشکسالی اخیر در شمال و شرق دشت سیستان، پایان‌نامه کارشناسی ارشد، گروه جغرافیای دانشگاه آزاد اسلامی واحد مشهد

Annex II to the WMO Technical Regulations, (2017) , *Manual on Codes International Codes*, No.306, Volume I.1 Part A – Alphanumeric Codes, CODE TABLES 4677, A–356, **357**.

Annex II to the WMO Technical Regulations, (2015), *Manual on Codes International Codes No*, **306**, Volume I.1 Part B – Binary Codes, Part C – Common Features to Binary and Alphanumeric Code, FM 94 BUFR0 20 003 Present weather, I.2–CODE/FLAG Tables/20—1.

Sand, W., and Advisory, D. S. W. (2018), *WMO Airborne Dust Bulletin*, WMO: Geneva, Switzerland.

World Health Organization (2016), Ambient air pollution: *A Global Assessment of Exposure and Burden of Disease*.

Eleftheriou, A., Mouzourides, P., and Biskos, G., Yiallouros, P., Kumar, P., and Neophytou, M. K. A. (2023), The Challenge of Adopting Mitigation and Adaptation Measures for the Impacts of Sand and Dust Storms in Eastern Mediterranean Region: *A Critical Review*, *Mitigation and Adaptation Strategies for Global Change*, **28**(6), 33.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

تحلیل بیزی تقریبی مدل‌های آمیخته خطی تعمیم‌یافته فضایی با استفاده از یک میدان تصادفی چوله گاوسی مانا

فاطمه حسینی[۱]، امید کریمی،
گروه آمار، دانشگاه سمنان

چکیده: اغلب متغیرهای پنهان که بیانگر همبستگی فضایی در مدل‌های آمیخته خطی تعمیم‌یافته فضایی هستند با استفاده از یک میدان تصادفی گاوسی مدل‌بندی می‌شوند. عدم برقراری فرض گاوسی باعث تاثیر روی دقت پیشگویی‌ها و برآورد پارامترهای مدل می‌شود. در این مقاله با استفاده یک میدان تصادفی چوله گاوسی مانا و به‌کارگیری یک رهیافت بیزی تقریبی، مدل‌های آمیخته خطی تعمیم‌یافته فضایی مدل‌بندی و برآورد می‌شوند. در نهایت در یک مثال شبیه‌سازی به بررسی کارایی رهیافت بیزی تقریبی پرداخته شده است.

واژه‌های کلیدی: تحلیل بیزی، میدان تصادفی مانا، مدل‌های آمیخته خطی تعمیم‌یافته فضایی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62F15، 62J12، 60G15.

۱ مقدمه

از مسائل مهم در مدل های آمیخته خطی تعمیم‌یافته فضایی برآورد پارامترهای مدل و پیشگویی متغیرهای پنهان فضایی است. چون اغلب در این مدل‌ها متغیرهای پاسخ متعلق به یک خانواده نمایی در نظر گرفته می‌شوند و وجود متغیرهای پنهان در این مدل‌ها، برآورد پارامترهای مدل به راحتی میسر نمی‌باشد. همچنین برای سادگی در محاسبات در این مدل‌ها متغیرهای پنهان فضایی با میدان تصادفی گاوسی مدل‌بندی می‌شوند (محمدزاده ، ۱۳۹۴). در این مقاله برای مدل‌بندی متغیرهای پنهان فضایی از یک میدان تصادفی چوله مانا که در برگیرنده میدان تصادفی گاوسی است استفاده می‌شود و با به‌کار بردن یک رهیافت بیزی تقریبی به تحلیل این مدل‌ها پرداخته می‌شود. کیم و مالیک (۲۰۰۴) میدان تصادفی چوله گاوسی را برای تحلیل داده‌های فضایی چوله معرفی نمود. براساس توزیع چوله نرمال بسته آلارد و ناویو (۲۰۰۷)، کریمی و همکاران (۲۰۱۰)، کریمی و محمدزاده (۲۰۱۱، ۲۰۱۲) به معرفی میدان‌های تصادفی چوله گاوسی پرداختند. در مطالعات اشاره شده میدان‌های چوله گاوسی معرفی شده دارای مشکلاتی مثل عدم مانایی است. ریمستاد و امره (۲۰۱۴) با استفاده از میدان تصادفی آلارد و ناویو (۲۰۰۷)، یک میدان تصادفی چوله گاوسی تقریبا مانا تعریف نمودند و کریمی و حسینی (۱۴۰۰) با

[۱] نام و ایمیل ارائه دهنده مقاله: فاطمه حسینی، fatemeh.hoseini@semnan.ac.ir

اصلاح میدان تصادفی چوله گاوسی تقریبا مانا ریمستاد و امره (۲۰۱۴) یک میدان تصادفی چوله گاوسی مانا معرفی نمودند. حسینی و محمدزاده (۲۰۱۲)، حسینی و کریمی (۲۰۲۰، ۲۰۲۱) در مدل‌های آمیخته خطی تعمیم‌یافته فضایی استفاده از توزیع چوله نرمال بسته برای متغیرهای پنهان فضایی را پیشنهاد کردند. یک تحلیل درست‌نمایی مرکب برای داده‌های فضایی با متغیرهای پنهان چوله نیز بیان کردند که از میدان تصادفی چوله گاوسی تقریبا ایستا ریمستاد و امره (۲۰۱۴) استفاده کرده‌اند. حسینی و کریمی (۱۴۰۱) و کریمی (۲۰۲۳) با به کار بردن روش‌های مونت‌کارلویی همیلتونی و الگویتم‌های بیشینه‌سازی امید ریاضی یک الگوریتم جدید برای به‌دست آوردن برآورد ماکسیمم درست‌نمایی پارامترها این مدل‌ها معرفی نمودند. در این مقاله از میدان تصادفی تعریف شده توسط کریمی و حسینی (۱۴۰۰) برای مدل‌بندی متغیرهای پنهان فضایی در مدل‌های آمیخته خطی تعمیم‌یافته فضایی استفاده و سپس با به‌کار گرفتن یک رهیافت بیزی تقریبی، این مدل‌ها برآورد می‌شوند. ساختار مقاله به این‌صورت است که در بخش دوم یک میدان تصادفی چوله گاوسی تقریبا مانا و مانا ارائه می‌شوند. در بخش سوم مدل و تحلیل بیزی مدل مورد بررسی قرار می‌گیرد و در نهایت در بخش چهارم در یک مثال شبیه‌سازی کارایی و دقت مدل بررس می‌شود.

۲ میدان تصادفی چوله گاوسی

بردار تصادفی n بعدی $\boldsymbol{x}$ با تابع چگالی

$$f_{n,q}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Gamma},\boldsymbol{\nu},,\boldsymbol{\Delta}) = k\phi_n(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\ \Phi_q(\boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{\mu});\boldsymbol{\nu},\boldsymbol{\Delta}), \tag{۱.۲}$$

دارای توزیع چوله نرمال بسته چند متغیره با پارامترهای $\boldsymbol{\mu}$، $\boldsymbol{\Sigma}$، $\boldsymbol{\Gamma}$، $\boldsymbol{\nu}$ و $\boldsymbol{\Delta}$ است، که در آن $\Phi_q(\boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{\mu});\boldsymbol{\nu},\Delta)$ تابع توزیع تجمعی q متغیره نرمال با بردار میانگین $\boldsymbol{\nu}$ و ماتریس واریانس کوواریانس $\boldsymbol{\Delta}$ است. $k=[\Phi_q(\boldsymbol{۰};\boldsymbol{\nu},\boldsymbol{\Delta}+\boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^\top)]^{-۱}$، $\boldsymbol{\mu}$ بردار پارامتر مکان، ماتریس $n\times n$ معین مثبت $\boldsymbol{\Sigma}$ ماتریس مقیاس، عناصر ماتریس $\boldsymbol{\Gamma}_{q\times n}$ پارامترهای چولگی هستند. به‌طور خلاصه این توزیع به‌صورت $CSN_{n,q}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Gamma},\boldsymbol{\nu},\boldsymbol{\Delta})$, نمایش داده می شود. وقتی $\boldsymbol{\Gamma}$ ماتریس صفر تعریف شود، تابع چگالی نرمال حاصل می‌شود. گشتاور مرتبه‌ی اول توزیع چوله نرمال بسته به‌صورت $\mathrm{E}(\boldsymbol{X})=\boldsymbol{\mu}+\boldsymbol{\Sigma}\boldsymbol{\Gamma}^\top\boldsymbol{\Psi}$ به‌دست می‌آید که در آن $\boldsymbol{\Psi}=\frac{\Phi_q^*(\boldsymbol{۰};\boldsymbol{\nu},\boldsymbol{\Delta}+\boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^\top)}{\Phi_q(\boldsymbol{۰};\boldsymbol{\nu},\boldsymbol{\Delta}+\boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^\top)}$ است. برای ماتریس معین مثبت $\boldsymbol{\Omega}$، $\Phi_q^*(\boldsymbol{s};\boldsymbol{\nu},\boldsymbol{\Omega})=[\nabla_s\Phi_q(\boldsymbol{s};\boldsymbol{\nu},\boldsymbol{\Omega})]^\top$ که در آن $\nabla_s=(\frac{\partial}{\partial s_۱},\cdots,\frac{\partial}{\partial s_q})^\top$ می‌باشد، ، (گنزالس و همکاران، ۲۰۰۴).

۱.۲ میدان تصادفی چوله گاوسی بسته تقریبا مانا

ریمستاد و امره (۲۰۱۴) میدان تصادفی چوله گاوسی بسته را با تعمیم میدان تصادفی چوله گاوسی تعریف شده توسط آلارد و ناویو (۲۰۰۷) به این‌صورت تعریف کردند که فرض کنید $\boldsymbol{U}(\boldsymbol{s})=\{(U_۱(\boldsymbol{s}),U_۲(\boldsymbol{s}))^\top,\boldsymbol{s}\in\mathcal{D}\subseteq R^d\}$ میدان تصادفی گاوسی دو متغیره باشد و $\boldsymbol{U}_۲=[U_۲(s'_۱),\cdots,U_۲(s'_q)]$ که در آن q ثابت و متناهی است. آنگاه میدان تصادفی چوله گاوسی بسته به‌صورت $\{\boldsymbol{X}(\boldsymbol{s})=[\boldsymbol{U}_۱(s)|\boldsymbol{U}_۲\leq\boldsymbol{۰}\}$ تعریف می‌شود، اگر برای هر مجموعه متناهی $(\boldsymbol{s}_۱,\cdots,\boldsymbol{s}_n)$، $\boldsymbol{X}=(X(\boldsymbol{s}_۱),\cdots,X(\boldsymbol{s}_n))^\top$ دارای توزیع چوله نرمال بسته باشد. آن‌ها نشان دادند که وقتی موقعیت‌های $(\boldsymbol{s}'_۱,\cdots,\boldsymbol{s}'_q)$ به اندازه‌ی کافی از مرزها دور ، $n=q$ و $(\boldsymbol{s}'_۱,\cdots,\boldsymbol{s}'_n)=(\boldsymbol{s}_۱,\cdots,\boldsymbol{s}_n)$ باشند و با در نظر گرفتن شکل توزیع چوله نرمال بسته برای $\boldsymbol{X}$ به‌صورت

$$CSN_{n,n}(\boldsymbol{\mu},\sigma^۲\boldsymbol{C}_\varphi,\frac{\gamma}{\sigma}\boldsymbol{I}_n,\nu\boldsymbol{۱}_n,(۱-\gamma^۲)\boldsymbol{I}_n), \tag{۲.۲}$$

که در آن $\boldsymbol{\mu}$ پارامتر مکان، $\sigma^۲$ پارامتر مقیاس، $|\gamma|<۱$ پارامتر چولگی، $\boldsymbol{C}_\varphi$ ماتریس همبستگی ایستا، $\boldsymbol{I}_n$ ماتریس واحد و $\boldsymbol{۱}_n\in\mathbb{R}^n$ برداری با عناصر یک است، آنگاه $\boldsymbol{X}(\boldsymbol{s})$ تقریبا ایستا است.

۲.۲ میدان تصادفی چوله گاوسی بسته مانا

کریمی و حسینی (۱۴۰۰) نشان دادند در میدان تصادفی تعریف شده توسط ریمستاد و امره (۲۰۱۴) توزیع حاشیه‌ای X_j به ماتریس همبستگی $\boldsymbol{C}_\varphi$ که مربوط به کل شبکه فضایی $(\boldsymbol{s}_1,\cdots,\boldsymbol{s}_n)$ می‌باشد وابسته است، اما ساختار کلی چگالی X_j یک توزیع چوله نرمال بسته است. برای رفع این مشکل یعنی برقراری شرط سازگاری حاشیه‌ای و تعریف یک میدان تصادفی چوله گاوسی مانا، آن‌ها یک تحقق nتایی از میدان تصادفی چوله گاوسی تقریبا مانا به‌صورت $\boldsymbol{X}=(X(\boldsymbol{s}_1),\ldots,X(\boldsymbol{s}_n))^\top$ مطابق رابطه (۲.۲) با تفکیک‌های

$$\boldsymbol{X}_{n\times 1}=\begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{pmatrix},\boldsymbol{C}_\varphi=\begin{pmatrix}C_{11} & C_{12}\\ C_{21} & C_{22}\end{pmatrix},$$

در نظر گرفتند که در آن $\boldsymbol{X}_1\in R^{n_1}$، $\boldsymbol{X}_2\in R^{n_2}$ و $n_1+n_2=n$ است و ماتریس همبستگی فضایی $\boldsymbol{C}_\varphi$ تفکیک‌های متناظر با بردارهای $\boldsymbol{X}_1$ و $\boldsymbol{X}_2$ از ماتریس $\boldsymbol{X}$ هستند. سپس با به‌دست آوردن تابع مولد گشتاور حاشیه‌ای $\boldsymbol{X}_1$ نشان دادند که تابع مولد گشتاور حاشیه‌ای $\boldsymbol{X}_1$ به ماتریس همبستگی $\boldsymbol{C}_\varphi$ برای کل ناحیه فضایی وابسته است و شرط سازگاری حاشیه‌ای برقرار نیست. برای این‌که شرط سازگاری حاشیه‌ای برقرار شود، کریمی و حسینی (۱۴۰۰) پیشنهاد نمودند که پارامترهای یک تحقق از میدان تصادفی چوله گاوسی تقریبا مانا را به‌صورت

$$\boldsymbol{X}\sim CSN_{n,n}\big(\boldsymbol{\mu},\sigma^2\boldsymbol{C}_\varphi,\frac{\gamma}{\sigma}\boldsymbol{C}_\varphi^{-\frac{1}{2}},\nu\mathbf{1}_n,(1-\gamma^2)\mathbf{I}_n\big), \tag{۳.۲}$$

اصلاح شوند، که در آن $\boldsymbol{C}_\varphi^{-\frac{1}{2}}$ عکس ریشه دوم ماتریس $\boldsymbol{C}_\varphi$ است. با محاسبه تابع مولد گشتاور توزیع حاشیه‌ای $\boldsymbol{X}_1$ نشان دادند که توزیع حاشیه‌ای $\boldsymbol{X}_1$ برای میدان تصادفی چوله گاوسی تعریف شده در رابطه (۳.۲) دارای بعد n_1 و از خانواده توزیع‌های چوله نرمال بسته است، پس دارای شرط سازگاری حاشیه‌ای و یک میدان تصادفی خوش تعریف می‌باشد.

۳ مدل

در مدل‌های خطی تعمیم‌یافته چون متغیر پاسخ لزوما کمی نیست و می‌تواند کیفی هم باشد، مک‌کلاح و نلدر (۱۹۸۹) پیشنهاد دادند برای متغیر پاسخ یک خانواده نمایی که شامل اکثر توزیع‌های آماری گسسته و پیوسته از جمله توزیع نرمال می‌باشد، در نظر گرفته شود. در این مقاله نیز متغیرهای پاسخ فضایی متعلق به خانواده نمایی فرض می‌شوند. قرار دهید $\boldsymbol{Y}^\top=(y_1,\cdots,y_k)$ بردار متغیرهای پاسخ فضایی در موقعیت‌های $\{s_1,\cdots,s_k\}$، $k\leq n$ باشد که متعلق به یک خانوده توزیع نمایی هستند. پس چگالی $\pi(\boldsymbol{y}|\boldsymbol{x})$، به‌طور شرطی مستقل و از یک خانواده‌ی توزیع نمایی به‌صورت $\pi(y_i|x_i)=\exp\{y_ix_i-b(x_i)+c(y_i)\},\ \ i=1,\ldots,k$ فرض می‌شود، که مطابق مدل‌های خطی تعمیم یافته $\mathrm{E}(Y_i|x_i)=g^{-1}(x_i)$ باشد که در آن $g(\cdot)$ یک تابع پیوند معلوم است. همچنین فرض کنید $\boldsymbol{X}=(x_1,\cdots,x_n)^\top$ یک تحقق از میدان تصادفی چوله گاوسی بسته در n موقعیت $\{s_1,\cdots,s_n\}$ با دامنه $\mathcal{D}\subseteq\mathbb{R}^n$ و دارای توزیع چوله نرمال به‌صورت (۳.۲) با پارامتر مکان $\boldsymbol{\mu}=\boldsymbol{H}\boldsymbol{\beta}$ باشد، که در آن $\boldsymbol{H}$ ماتریسی با بعد $n\times(p+1)$ شامل متغیرهای کمکی و $\boldsymbol{\beta}$ بردار پارامترهای رگرسیونی است. برای ماتریس $\boldsymbol{C}_\varphi$ ساختار فضایی نمایی همسانگرد فرض می‌شود. از رابطه‌ی (۱.۲) تابع چگالی $\boldsymbol{X}$ به‌صورت

$$\pi(\boldsymbol{x}|\boldsymbol{\eta})\ \ =\ \ \phi_n(\boldsymbol{x};\boldsymbol{H}\boldsymbol{\beta},\sigma^2\boldsymbol{C}_\varphi)\times\frac{\Phi_n(\frac{\gamma}{\sigma}C_\varphi^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\beta}),\nu\mathbf{1}_n,(1-\gamma^2)\mathbf{I}_n)}{\Phi_n(\mathbf{0};\nu\mathbf{1}_n,\mathbf{I}_n)}, \tag{۱.۳}$$

به‌دست می‌آید، بنابراین بردار پارامترهای مدل به‌صورت $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \sigma, \varphi, \nu, \gamma)^\top$ در نظر گرفته می‌شود که معمولا $\boldsymbol{\nu}$ را معلوم و صفر در نظر می گیرند. اکنون تابع درست‌نمایی را می‌توان به‌صورت

$$L(\boldsymbol{\eta} \mid \boldsymbol{y}) = \int \exp\{\sum_{i=1}^{k}(y_i x_i - b(x_i) + c(y_i)) - \frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{H\beta})^\top \boldsymbol{C}_\varphi^{-1}(\boldsymbol{x}-\boldsymbol{H\beta})\} \times \frac{\Phi_n(\frac{\gamma}{\sigma}\boldsymbol{C}_\varphi^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{H\beta}), \nu\mathbf{1}_n, (1-\gamma^2)\mathbf{I}_n)}{\Phi_n(\mathbf{0}; \nu\mathbf{1}_n, \mathbf{I}_n)} dx,$$

نوشت که تابعی پیچیده است و دارای شکل بسته‌ای نیست. لذا به‌دست آوردن برآوردها به شکل درستنمایی به راحتی امکان‌پذیر نیست در بخش بعد یک رهیافت بیزی تقریبی برای به‌دست آوردن پارامترهای مدل ارائه می‌شود.

۴ برآورد بیزی مدل

قضیه ۱.۴. *فرض کنید متغیرهای پنهان فضایی در مدل آمیخته خطی تعمیم‌یافته فضایی تعمیم‌یافته دارای توزیع چوله نرمال بسته، $(\boldsymbol{X}|\boldsymbol{\eta}) \approx CSN_{n,n}(\boldsymbol{H\beta}, \sigma^2\boldsymbol{C}_\varphi, \frac{\gamma}{\sigma}\boldsymbol{C}_\varphi^{-\frac{1}{2}}, \nu\mathbf{1_n}, (1-\gamma^2)\mathbf{I_n})$ باشد و $\boldsymbol{X}$ به‌صورت $\boldsymbol{X} = (\boldsymbol{X}^{obs\top}, \boldsymbol{X}^{pred\top})^\top$ افراز شده باشد. همچنین فرض کنید متغیرهای پاسخ گسسته فضایی متعلق به خانواده نمایی*

$$\pi(y_i|x_i) = \exp\{y_i x_i - b(x_i)\},\ i = 1, \cdots k$$

باشند. آنگاه با در نظر گرفتن $\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\eta})$ و با خطی‌سازی قسمت درست‌نمایی $\boldsymbol{X}^{obs} = A\boldsymbol{X}$، $A = [I_{k\times k}|\mathbf{0}_{k\times n-k}]$ حول یک مقدار ثابت $\boldsymbol{x}$، توزیع $\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\eta}$ به طور تقریبی چوله نرمال بسته با پارامترهای

$$(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\eta}) \approx CSN_{n,n}(\boldsymbol{\mu}_{x|y,\eta}, \boldsymbol{\Sigma}_{x|y,\eta}, \frac{\gamma}{\sigma}\boldsymbol{C}_\varphi^{-\frac{1}{2}}, \nu_{x|y,\eta}, (1-\gamma^2)\mathbf{I_n}) \qquad (۱.۴)$$

است، که در آن، $\boldsymbol{\mu}_{x|y,\eta} = \boldsymbol{H\beta} + \boldsymbol{C}_\varphi \boldsymbol{R}(z(\boldsymbol{y}, \boldsymbol{x}^{obs}) - \boldsymbol{AH\beta})$، $z_i(y_i, x_i) = [y_i - b'(x_i) + x_i b''(x_i)]/b''(x_i)$ و $i = 1, \ldots, k$ خطی‌سازی قسمت درست‌نمایی در مقدار ثابت از $\boldsymbol{x}$ است. ماتریس $\boldsymbol{R} = \boldsymbol{A}^\top(\boldsymbol{AC}_\varphi\boldsymbol{A}^\top + \frac{1}{\sigma^2}\boldsymbol{P})^{-1}$، $P(i,i) = 1/b''(x_i)$، $\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{x})$ یک ماتریس قطری با عناصر روی قطر و $\boldsymbol{\Sigma}_{x|y,\eta} = \sigma^2\boldsymbol{C}_\varphi(\mathbf{I_n} - \boldsymbol{RAC}_\varphi)$ و $\boldsymbol{\nu}_{x|y,\eta} = \nu\mathbf{1_n} - \sigma\gamma\boldsymbol{C}_\varphi^{\frac{1}{2}}\boldsymbol{R}(\mathbf{z}(\boldsymbol{y}, \mathbf{x^{obs}}) - \boldsymbol{AH\beta})$ می‌باشد، (حسینی و کریمی، ۱۴۰۱).

برای به‌دست آوردن توزیع تقریبی (۱.۴) می‌توان ابتدا یک مقدار اولیه برای $\boldsymbol{x}^{(0)}$ در نظر گرفت و $m = 0$ قرار داده شود و برآورد $\hat{\pi}(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\eta}) = CSN_{n,n}(\boldsymbol{\mu}_{x|y,\eta}(\boldsymbol{x}^{(m)}), \boldsymbol{\Sigma}_{x|y,\eta}(\boldsymbol{x}^{(m)}), \frac{\gamma}{\sigma}\boldsymbol{C}_\varphi^{-\frac{1}{2}}, \nu_{x|y,\eta}(\boldsymbol{x}^m), (1-\gamma^2)\mathbf{I_n})$ محاسبه و میانگین این توزیع تقریبی به عنوان مقدار جدید $\boldsymbol{x}^{(m)}$ منظور و $m = m + 1$ قرار داده شود و الگوریتم تا رسیدن به همگرایی تکرار شود. با در نظر گرفتن توزیع‌های پیشین برای پارامترهای مدل به‌صورت $\boldsymbol{\beta} \sim N(\boldsymbol{a}, B)$، $\sigma \sim IG(\alpha, \tau)$، $\varphi \sim \Gamma(\lambda, \omega)$ و $\gamma \sim U(-1, 1)$ و فرض استقلال پیشین‌ها، $\pi(\boldsymbol{\eta}) = \pi(\boldsymbol{\beta})\pi(\sigma)\pi(\varphi)\pi(\gamma)$، توزیع پسین به‌صورت

$$\begin{aligned} \pi(\boldsymbol{x}, \boldsymbol{\eta}|\boldsymbol{y}) &= \pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta})\pi(\boldsymbol{\eta}|\boldsymbol{y}) \\ &= \frac{\pi(\boldsymbol{y}|x)\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})}{\pi(\boldsymbol{y})} \end{aligned} \qquad (۲.۴)$$

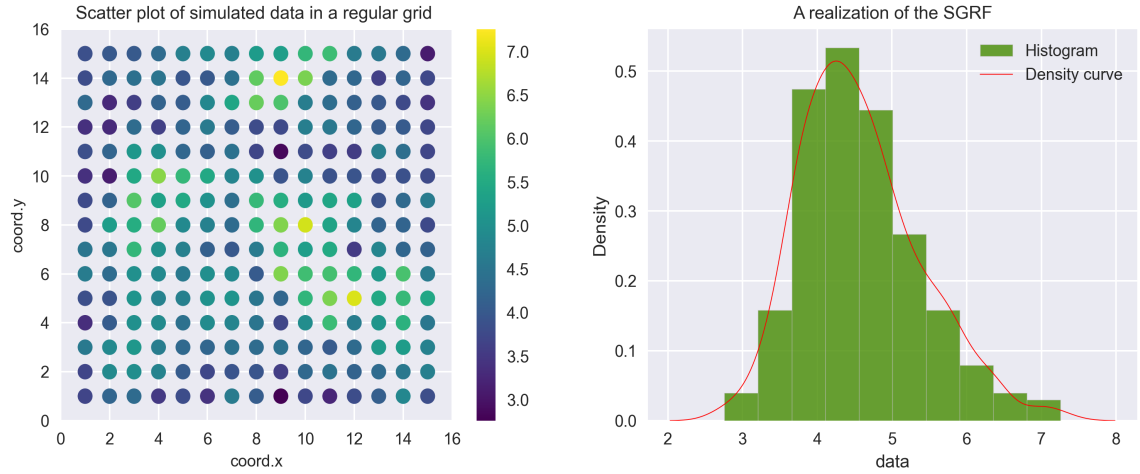
است و بنابراین از (۲.۴) و برآورد توزیع شرطی کامل تقریبی $\hat{\pi}(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\eta})$ توزیع تقریبی حاشیه‌ای پسین به‌صورت

$$\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})}{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta})}$$

به‌دست می‌آید. با در نظر گرفتن توزیع پیشین نرمال برای پارامترهای رگرسیون ثابت می‌شود توزیع شرطی کامل برای این پارامتر شکل بسته‌ای از توزیع چوله نرمال بسته دارد اما توزیع شرطی کامل سایر پارامترها شکل مشخصی ندارند و با ستفاده از الگوریتم‌ها گیبز و متروپلیس هاستینگس برآورد بیزی پارامترها به‌دست آورده می‌شود.

۵ مطالعه شبیه‌سازی

ابتدا یک مشبکه‌ی منظم ۱۵ × ۱۵ در نظر گرفته و تعداد ۲۲۵ موقعیت برروی آن تولید شده است. با در نظر گرفتن ساختار همسانگرد نمایی برای $\boldsymbol{C}_\varphi$ و $\beta_0 = ۲$، $\beta_1 = ۱$، $\sigma^2 = ۱$، $\varphi = ۵$ و $\gamma = ۰/۸۵$ متغیرهای پنهان از توزیع چوله نرمال به‌صورت $(\boldsymbol{x}|\boldsymbol{\eta}) \sim CSN_{n,n}(\beta_0 + \beta_1 \boldsymbol{h}, \sigma^2 \boldsymbol{C}_\varphi, \frac{\gamma}{\sigma} \boldsymbol{C}_\varphi^{-\frac{1}{2}}, \nu \mathbf{1_n}, (1-\gamma^2)\mathbf{I_n})$ استخراج شد، که در آن $\boldsymbol{h}$ بردار مقادیر استاندارد شده عرض جغرافیایی است که به عنوان متغیر کمکی وارد مدل می‌شود. متغیر پاسخ به شرط متغیرهای پنهان، $y_j, j = ۱, \cdots, ۲۲۵$ از توزیع پواسون به‌صورت $y_j \sim Poiss(p_j),\ p_j = \exp(x_j)$ تولید و فرایند شبیه‌سازی ۱۰۰ بار تکرار شد. یک نمونه از موقعیت‌ها و داده‌های تولید شده در شکل ۱ رسم شده است.



شکل۱. تحققی از میدان تصادفی چوله نرمال بسته مانا: راست) هیستوگرام مقادیر متغیرهای پنهان شبیه‌سازی شده، چپ) نمودار پراکنش متغیرهای پنهان روی مشبکه منظم ۱۵ × ۱۵.

برای ۱۰۰ مجموعه داده‌ی تولید شده و مدل چوله گاوسی مانا و مدل گاوسی، رهیافت بیزی تقریبی معرفی شده در مقاله اجرا شد و خلاصه نتایج به شرح جدول ۱ ارائه شده است. برای ارزیابی برآورد بیزی پارامترها از معیار مجذور میانگین توان دوم خطاها

$$RMSE(\hat{\eta}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\eta}_i - \eta_i)^2},\ \ m = ۱۰۰,$$

با ۱۰۰ مجموعه داده شبیه‌سازی استفاده شده است، که در آن m تعداد مجموعه داده‌های شبیه‌سازی برای هر حالت و $\hat{\eta}_i$ برآورد بیزی پارامترهای میدان تصادفی برای مجموعه داده iام است.

جدول۱. نتایج شبیه‌سازی براساس ۲۱۵ × ۱۰۰ مجموعه داده‌ی تولید شده.

گاوسی			چوله گاوسی				
RMSE	Sd.	برآورد	RMSE	Sd.	برآورد	واقعی	Par.
۰/۷۷۱	۰/۷۹۴	۱/۷۶۷	۰/۶۳۲	۰/۶۵۳	۲/۱۹۱	۲	β_0
۰/۰۶۱	۰/۰۵۳	۰/۹۴۱	۰/۰۳۹	۰/۰۴۳	۱/۰۲	۱	β_1
۰/۱۳۱	۰/۱۲۶	۰/۸۱۱	۰/۰۳۳	۰/۰۲۹	۰/۹۹۰	۱	σ^2
۱/۲۱۶	۰/۹۱۱	۴/۰۳۱	۰/۹۰۱	۰/۷۷۱	۵/۱۲۸	۵	φ
–	–	–	۰/۰۰۲	۰/۰۰۱	۰/۹۳۲	۰/۸۵	γ

نتایج نشان می‌دهد که با به‌کار بردن میدان تصادفی مانا و رهیافت بیزی معرفی شده میانگین توان دوم خطاهای برآورد همه‌ی پارامترها برای مدل چوله گاوسی به خوبی عمل نموده است. با توجه به این‌که داده‌ها از یک میدان تصادفی چوله نرمال بسته تولید شده‌اند نتایج نشان می‌دهد که دقت برآورد پارامترهای مدل تحت تاثیر فرض مدل در نظر گرفته شده روی متغیرهای پنهان فضایی می‌باشد و در صورت عدم برقراری فرض گاوسی بودن، از دقت برآوردها کاسته می‌شود.

بحث و نتیجه‌گیری

در این مقاله از یک میدان تصادفی چوله نرمال بسته مانا برای مدل‌بندی متغیرهای پنهان فضایی در مدل‌های آمیخته خطی تعمیم‌یافته استفاده و یک رهیافت بیزی برای برآورد این مدل‌ها معرفی شد. برای به‌دست آوردن برآوردهای بیزی پارمترهای مدل از یک رهیافت بیزی تقریبی و الگوریتم‌های مونت‌کارلویی گیبز و متروپلیس هاستینگز استفاده شد. نتایج شبیه‌سازی نشان می‌دهد که مدل معرفی شده براساس میدان تصادفی چوله گاوسی مانا و رهیافت تقریبی بیزی معرفی شده برای به‌دست آوردن برآوردهای مدل به خوبی عمل نموده است. به دلیل پیچیدگی مدل معرفی شده تحلیل بیزی یا به‌کار بردن روش‌های مونت‌کارلویی زمان بر می‌باشد و گاهی همگرایی‌ها به خصوص برای داده‌های حجیم به راحتی حاصل نمی‌شود. به عنوان پیشنهاد می‌توان برای به‌دست آوردن برآورد پارامترها به روش بیزی از رهیافت $INLA$ یا الگوریتم‌های مونت‌کارلویی همیلتونی استفاده کرد که باعث افزایش سرعت محاسبات می‌شوند.

مراجع

حسینی، ف. و کریمی، ا. (۱۴۰۱) برآورد مدل‌های آمیخته خطی تعمیم‌یافته فضایی با میدان تصادفی چوله گاوسی مانا، *اندیشه آماری*، **۱:۲۷** ۷۳ـ۷۹.

کریمی، ا. و حسینی، ف. (۱۴۰۰)، معرفی یک میدان تصادفی مانای چوله گاوسی، *مجله علوم آماری ایران*، **۱۵**، **۵۴۹**ـ ۵۶۶.

محمدزاده، م.، (۱۳۹۸)، *آمار فضایی و کاربردهای آن*، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،

Allard, D. and Naveau, P. (2007), A New Spatial Skew-Normal Random Field Model, *Communications in Statistics— Theory and Methods*, **36**, 1821-1834.

Gonzalez-Farias, G., Dominguez-Molina, A. and Gupta, A. K. (2004), The Closed Skew Normal Distribution. In: *Genton M. G., ed. Skew-elliptical distributions and their applications: A journey beyond normality*. Boca Raton, FL: Chapman and Hall CRC, 25-42

Hosseini, F., Mohammadzadeh, M., (2012). Bayesian prediction for spatial GLMM's with Closed Skew Normal latent variables, *Australian & New Zealand Journal of Statistics*, **54**, 43-62.

Hosseini, F., Karimi, O. (2020). Approximate likelihood Inference in Spatial Generalized Linear Mixed Models with Closed Skew Normal latent Variables, *Communication in Statistics- Simulation and Computation* **49**, 121-134.

Hosseini, F., Karimi, O., (2021). Approximate pairwise likelihood inference in SGLM models with skew normal latent variables, *Journal of Computational and Applied Mathematics*, **398**, 113692.

Kim, H.M., Mallick, B.K., (2004). A Bayesian prediction using the skew Gaussian distribution. *Journal of Statistical Planning and Inference*, **120**, 85–101.

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

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

Karimi, O., Mohammadzadeh, M., (2012). Bayesian spatial regression models with closed skew normal correlated errors and missing observations, *Statistical Papers* **53(1)**, 205-218.

Karimi, O., (2023). A Hamiltonian Monte Carlo EM algorithm for generalized linear mixed models with spatial skew latent variables, *Statistical Papers*, doi 10.1007/s00362-023-01419-y

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

Rimstad, K., and Omre, H. (2014). Skew-Gaussian Random Fields, *Spatial Statistics* **10**, 43-62.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

تحلیل داده‌های فضایی در حضور داده‌های پرت با رگرسیون چندکی خودهمبسته فضایی

طیبه سارانی [۱]**، محسن محمدزاده**
گروه آمار، دانشگاه تربیت مدرس

چکیده: اگر موضوع مورد مطالعه ما، تحت تأثیر عوامل متنوعی با اثرات فضایی، قرار گیرد، برای مدل‌بندی متغیرهای پاسخ و تبیینی معمولا از مدل خودهمبسته فضایی استفاده می‌شود. از طرفی وجود داده‌های پرت فضایی نیز مدل‌بندی داده‌ها را تحت تاثیر قرار می‌دهد. از آنجا که در مدل رگرسیون چندکی، تلاش می‌شود مانده‌های موزون، به حداقل رسانده شود، می‌توان از ترکیب مدل‌های خودهمبسته فضایی و رگرسیون چندکی، مدلی به دست آورد، که از مدل‌های متعارف کارایی بهتری داشته باشد. در این مقاله، ضمن معرفی مدل رگرسیون چندکی خودهمبسته فضایی، نحوه مدل‌بندی و تحلیل داده‌های فضایی در حضور داده‌های پرت ارائه می‌شود. آنگاه داده‌های سطح بیکاری باز با استفاده از مدل معرفی شده تحلیل می‌شوند. نتایج بیانگر آن است که عملکرد مدل رگرسیون چندکی خودهمبسته فضایی در مقایسه با مدل خودهمبسته فضایی در برخورد با وابستگی داده‌ها و تنوع در مدل‌بندی داده‌های فضایی بهتر است و به راحتی تحت تاثیر داده‌های پرت نیز قرار نمی‌گیرد.

واژه‌های کلیدی: مدل خودهمبسته فضایی، رگرسیون چندکی، داده پرت
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 62H11، 62G08.

۱ مقدمه

بیکاری یکی از معضل‌هایی است که در همه کشورهای در حال توسعه پدید می‌آید. عدم تعادل بین تعداد جویندگان کار نسبت به فرصت‌های شغلی موجود در یک منطقه، را می‌توان به عنوان یکی از مشخصه‌های بیکاری نام برد. بوجود آمدن یک روند اجتماعی می‌تواند باعث افزایش یا کاهش نرخ بیکاری در جامعه شود. برای اندازه‌گیری نرخ بیکاری در یک منطقه، از شاخص سطح بیکاری باز[1] (OUL) ، یعنی نسبت تعداد بیکاران به کل نیروی کار در آن منطقه، استفاده می‌شود. ویژگی‌های مشترک بین مناطق معمولاً به عنوان عوامل مؤثر بر OUL دخیل هستند و به عنوان اثرات فضایی شناخته می‌شوند (دای و جین ، ۲۰۲۱). اثرات فضایی به دو بخش وابستگی فضایی و تنوع فضایی تقسیم می‌شوند. وابستگی فضایی به دلیل ارتباط بین مناطق رخ می‌دهد، در حالی که تنوع فضایی به دلیل تفاوت بین یک منطقه و مناطق

[1] Open Unemployment Level
[2] Spatial Autorehressive

[۱] نام و ایمیل ارائه دهنده مقاله: طیبه سارانی، t.sarani@modares.ac.ir

دیگر رخ می‌دهد. در این صورت برای مدل‌بندی داده‌ها از مدل رگرسیون خودهمبسته[۲] (SAR) فضایی استفاده می‌شود (ور هوف و همکاران ، ۲۰۱۸). گاهی وجود داده‌های پرت نیز بر روش مدل‌سازی داده‌ها تأثیر می‌گذارد (یانوار و همکاران ، ۲۰۱۹). برای مدل‌بندی داده‌های حاوی اثرات فضایی و نقاط پرت از مدل رگرسیون چندکی خودهمبسته فضایی[۳] (SARQR) استفاده می‌شود (دای و همکاران ، ۲۰۲۰).

۲ مدل‌های آماری

برای اعمال روش SAR ، چندین مرحله آزمایش لازم است. چندخطی بودن یک شرط اساسی در خصوص همبستگی بین متغیرهای تبیینی در مدل رگرسیونی است. میزان چند خطی بودن در یک مدل رگرسیونی چندگانه را می‌توان با استفاده از مقدار عامل تورم واریانس[۴] $VIF_\ell = \frac{۱}{۱-R_\ell^۲}$ اندازه‌گیری کرد، که در آن $R_\ell^۲$ برای $\ell = ۱, \ldots, p$، ضریب تعیین تعدیل نشده برای رگرسیون متغیر تبیینی ℓام بر روی بقیه متغیرها است (زو و هوانگ ، ۲۰۱۵). مواقعی که $R_\ell^۲$ برابر با صفر است، مقدار VIF برابر یک خواهد بود و متغیرتبیینی ℓام با بقیه متغیرها همبستگی ندارد. وقتی $VIF = ۱$، متغیرهای تبیینی همبستگی ندارند. وقتی مقدار VIF بین ۱ تا ۵ باشد، متغیرهای تبیینی همبستگی متوسطی دارند. مقدار VIF بیشتر از ۵ باشد، متغیرهای تبیینی همبستگی بالایی دارند.

۱.۲ اثرات فضایی

برای تعیین وجود وابستگی فضایی بین موقعیت‌ها یا مناطق، از آزمون شاخص موران استفاده می‌شود. شاخص موران به صورت (تریبووانشوار و همکاران ، ۲۰۲۲)

$$I = \frac{\sum_{i=۱}^{n}\sum_{j=۱}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{S^۲ \sum_{i=۱}^{n}\sum_{j=۱}^{n} w_{ij}} \qquad (۱.۲)$$

تعریف می‌شود، که در آن n تعداد داده‌ها، y_i مشاهده iام متغیرپاسخ، $\bar{y}$ میانگین مشاهدات، w_{ij} درایه (i,j) ماتریس وزن فضایی W و $S^۲$ واریانس نمونه است. اگر مقدار I مثبت باشد، نواحی مجاور مقادیر مشابهی دارند و الگوی داده‌ها تمایل به خوشه‌بندی دارد.اگر مقدار I منفی باشد، نواحی مجاور مقادیر متفاوتی دارند و الگوی داده‌ها تمایل به پخش شدن دارد. اگر I برابر صفر باشد، هیچ همبستگی فضایی وجود ندارد.(هوانگ و همکاران ، ۲۰۱۰). برای بررسی وجود وابستگی فضایی به متغیر پاسخ، از آزمون ضریب لاگرانژ با آماره (آنسلین ، ۱۹۸۸)

$$LM_{\text{lag}} = \frac{\left(\frac{U'WY}{S^۲}\right)^۲}{T + \frac{(WX\beta)'M(MX\beta)}{S^۲}} \qquad (۲.۲)$$

استفاده می‌شود، که در آن $T = \text{trace}[(W+W')W]$، $M = I - X(X'X)^{-۱}X'$، $S^۲ = \frac{U'U}{n}$، U بردار مانده‌ها، X ماتریس متغیرهای تبیینی با اندازه $n \times (k+۱)$ و β بردار ضرایب رگرسیونی است. اگر $LM_{\text{lag}} > \chi^۲_{\alpha,۱}$، بین متغیرهای تبیینی و متغیر پاسخ وابستگی فضایی وجود دارد. در این‌صورت، مدل‌سازی با مدل خودهمبسته فضایی انجام می‌شود. برای بررسی ناهمسانی واریانس در مدل رگرسیون خطی از آزمون بروش–پاگان (بروش و پاگان ، ۱۹۷۹) با آماره $BP = \frac{۱}{۲}\mathbf{f}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-۱}\mathbf{X}'$ استفاده می‌شود که برای بررسی تنوع فضایی نیز قابل استفاده است. اگر $BP > \chi^۲_{\alpha,k-۱}$ باشد، بین مشاهدات مناطق مختلف تغییرات فضایی معنی‌داری در سطح α وجود دارد، که در آن k تعداد متغیرهای تبیینی درگیر در مدل را نشان می‌دهد.

[3] Spatial Autoregressive Quantile Regression

[4] Variance Inflation Factor

۲.۲ مدل رگرسیون چندکی

در روش رگرسیون چندکی از رویکرد جداسازی داده‌ها به گروه‌های چندکی استفاده می‌شود که ممکن است مقادیر برآورد متفاوتی داشته باشند (یانوار و همکاران، ۲۰۲۳). مدل رگرسیون خطی برای چندک τ به صورت $Y = XB_\tau + U$ است ، که در آن Y بردار n بعدی متغیرهای وابسته ، X ماتریس $n \times (k+1)$ بعدی متغیرهای تبیینی، B_τ عامل $(k+1)$ بعدی ضرایب رگرسیون چندکی مرتبط با چندک τ و U بردار n بعدی مانده‌ها است. برآورد پارامترهای مدل رگرسیون چندکی با مینیمم کردن مجموع قدر مطلق خطاها با وزن‌های τ برای خطاهای مثبت و وزن‌های $(1-\tau)$ برای خطاهای منفی به دست می‌آید و به صورت $\arg\min_{\beta\in\theta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i\beta_i)$ فرمول‌بندی می‌شود، که در آن تابع زیان $\rho_\tau(\cdot)$ به صورت زیر تعریف می‌شود:

$$\rho_\tau(U) = \begin{cases} \tau U & U > 0 \\ (\tau - 1)U & U \leq 0 \end{cases}$$

۳.۲ مدل خودهمبسته فضایی

مدل خودهمبستگی فضایی (SAR) یک مدل خطی با متغیرهای پاسخ همبسته فضایی است و به صورت (آنسلین ، ۱۹۸۸)

$$Y = \lambda WY + X\beta + U, \quad U \sim N(0, \sigma^2 I)$$

بیان می‌شود، که در آن λ ضریب خودهمبستگی فضایی بین مناطق، W ماتریس وزن فضایی $n \times n$ بعدی و β بردار ضرایب رگرسیون چندکی $(k+1)$ بعدی است. پارامترهای β و σ^2 در مدل SAR با روش ماکسیمم درستنمایی به ترتیب به صورت $\bar{\beta} = (X'X)^{-1}X'^{(1-\lambda W)}Y$ و $\bar{\sigma}^2 = \frac{1}{n}((I-\lambda W)Y - Y\beta)'((I-\lambda W)Y - X\beta)$ برآورد می‌شوند، برآورد پارامتر λ را نیز می‌توان با روش‌های عددی به دست آورد.

۴.۲ مدل رگرسیون چندکی خودهمبسته فضایی

SARQR ، از ترکیب دو مدل خودهمبسته فضایی و رگرسیون چندکی حاصل می‌شود. توسعه مدل SAR بر روی چندک τام به صورت $Y = \lambda_\tau WY + X\beta_\tau + U$ تعریف می‌شود (لوم و گلفاند ، ۲۰۱۲). مقدار ضریب خودهمبسته فضایی در مدل SARQR بزرگی وابستگی فضایی بین مناطق مجاور را نشان می‌دهد. در مدل رگرسیون چندکی از متغیر ابزاری [۵] (IVQR) برای برآورد پارامترهای مدل SARQR استفاده می‌شود (یو و همکاران ، ۲۰۲۱). مفروضات مورد استفاده به شرح زیر است (ژانگ و همکاران ، ۲۰۲۰):

۱. $P(u_i \leq 0) = \tau$ برای همه $i = 1, \ldots, n$،

۲. $\sup_{n\geq 1} \max_{1\leq i\leq n} E(u_i) \leq \mu < \infty$،

۳. $U \sim N(0, \lambda^2 I)$.

برآورد پارامترهای λ_τ و β_τ با کمینه کردن رابطه $\arg\min E = [\rho_\tau(y_i - \lambda_\tau \sum_{i=j}^{n} \omega_{ij}y_j - X_i'\beta_\tau - g(X_i, Z_i))]$ نسبت به هر پارامتر به دست آورده می‌شود، که در آن $g(X_i, Z_i)$ یک تابع خطی به عنوان رگرسیون چندکی متغیر ابزاری است. برآورد پارامترها با روش IVQR در مدل SARQR طی مراحل زیر انجام می‌شود (سو و یانگ ، ۲۰۱۱):

۱. مقداری برای λ اختیار نموده و مدل‌سازی در چندک τام به صورت $(\bar{\beta}_\tau(\lambda), \bar{\gamma}_\tau(\lambda)) = \arg\min_{\beta\gamma} Q_\tau(\beta, \lambda, \gamma)$ انجام شود.

[۵] Instrumental Variable Quantile Regression

۲. برآورد λ براساس IVQR با به حداقل رساندن بردار ابزار متغیر $\bar{\gamma}_\tau(\lambda)$ به صورت $\hat{\lambda}_\tau = \arg\min_\lambda \bar{\gamma}_\tau(\lambda)\hat{A}(\hat{\gamma}_\tau(\lambda))'$ حاصل می‌شود، که در آن $\hat{A} = A + O_p(۱)$ و A یک ماتریس معین مثبت است.
۳.برآوردگر β نیز به صورت $\hat{\beta}_\tau = \hat{\beta}_{\tau-۱}\hat{\lambda}_{\tau-۱}$ بدست می‌آید.
مراحل بالا برای هر چندک τ تکرار می‌شود. در هر یک از چندک‌ها، برآوردهای متفاوتی برای پارامترها به دست می‌آید.

۳ تحلیل داده‌های سطح بیکاری باز

با رسم نمودار جعبه‌ای داده‌ها، ملاحظه می‌شود که در مقادیر جمعیت $(X_۱)$، تولید ناخالص داخلی منطقه‌ای $(X_۴)$ و سطح بیکاری باز (Y) داده پرت وجود دارد. بنا بر نتایج آزمون چند خطی در جدول ۱، همه متغیرها دارای مقدار VIF کمتر از ۵

جدول ۱: نتایج آزمون چندخطی مستقل با عامل تورّم واریانس

متغیر	توضیح	VIF
$X_۱$	درصد جمعیت	۱/۷۶
$X_۲$	درصد جمعیت کم‌درآمد	۲/۶۹
$X_۳$	سطح مشارکت نیروی کار	۱/۰۸
$X_۴$	تولید ناخالص داخلی منطقه‌ای	۱/۶۹
$X_۵$	درصد شاخص توسعه انسانی	۲/۷۲

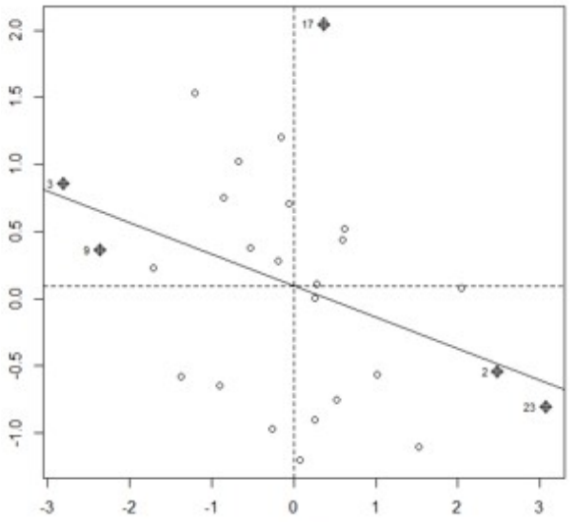
هستند، بنابراین مشکل چندخطی بین متغیرهای تبیینی وجود ندارد. ناهمسانی در مدل نیز با استفاده از آزمون بروش‌ـ پاگان بررسی می‌شود. بر اساس داده‌های مورد استفاده در این مطالعه، مقدار آماره آزمون بروش‌ـ پاگان برابر ۲/۹۵۳۸ با pـ مقدار ۰/۰۶۱ است. این نشان می‌دهد که واریانس OUL در بین مناطق مختلف یکسان است. به دلیل وجود یک اثر تصادفی وابسته فضایی به متغیر پاسخ، از مدل SAR استفاده می‌شود که نتایج برآورد پارامترهای آن در جدول ۲ ارائه شده است. براساس این جدول متغیرهای تبیینی که تأثیر معنی‌داری بر OUL دارند، فقط $(X_۲)$ و $(X_۳)$ با pـ مقدار کمتر از سطح

جدول ۲: برآورد و خطای استاندارد پارامترهای مدل SAR

متغیر	برآورد	خطای استاندارد	آماره-Z	p مقدار
ثابت	۱۳/۰۴۱۱	۹/۹۷۹۳	۱/۳۰۷۲	۰/۱۹۱۲
$X_۱$	۰/۱۴۳۲	۰/۱۳۴۲	۱/۰۶۵۴	۰/۲۸۷۳
$X_۲$	۰/۲۹۵۱*	۰/۱۵۶۴	۱/۸۸۵۳	۰/۰۵۹۳
$X_۳$	−۰/۳۰۲۲*	۰/۰۷۶۳	−۳/۹۴۹۱	۰/۰۰۰۱
$X_۴$	۰/۰۰۵۲	۰/۰۰۵۳	۰/۸۶۹۲	۰/۳۸۵۰
$X_۵$	۰/۱۴۲۱	۰/۰۹۳۲	۱/۵۳۳۲	۰/۱۲۵۰
λ	۰/۳۲۲۲	۰/۱۷۰۱	۱/۸۸۸۲	۰/۰۵۷۲

* معنی‌دار در سطح معنی‌داری $\alpha = ۰/۱$، $Z_{\alpha/۲} = ۱/۶۴۵$

معنی‌داری $\alpha = ۰/۱$ هستند. در این میان، $(X_۱)$، $(X_۴)$ و $(X_۵)$ تأثیر معنی‌داری بر سطح بیکاری باز ندارند. آزمایش اثر فضایی مجدد بر روی مدل SAR انجام شد. در شکل ۱ همانطور که ملاحظه می‌شود پنج داده پرت با علامت ستاره نشان داده شده است. این نقاط پرت بر نتایج پارامترها تأثیر می‌گذارند و دقت مدل را کاهش می‌دهند. بنابراین، مدل OUL



شکل ۱: نقاط پرت فضایی در مدل SAR

به دست آمده براساس روش SAR قابل قبول نیست. در این صورت از مدل SARQR برای مقابله با اثرات فضایی و داده‌های حاوی نقاط پرت فضایی برای به دست آوردن مدل قابل قبول‌تر استفاده می‌شود.

جدول ۳: برآورد پارامترهای مدل SARQR در چندک های مختلف

τ	ثابت‌ها	X_1	X_2	X_3	X_4	X_5	λ_τ
۰/۱۵	۱۴/۳۴۵	−۰/۲۴۹	۰/۱۶۴	−۰/۳۰۴*	−۰/۰۰۰	۰/۱۲۶	۰/۰۲۳
۰/۲۵	−۲/۳۴۲	۰/۰۲۰	۰/۴۳۶	−۰/۱۵۷	۰/۰۰۹	۰/۱۳۹	۰/۰۱۵*
۰/۳۵	۵/۶۰۱	۰/۱۲۵	۰/۴۱۰	−۰/۲۱۶	۰/۰۰۶	۰/۰۹۲	−۰/۰۱۹
۰/۴۵	۱۴/۶۵۹	۰/۱۹۷	۰/۴۲۶*	−۰/۳۳۶*	۰/۰۰۵	۰/۱۰۴	−۰/۰۰۲*
۰/۵۵	۱۲/۱۴۵	۰/۱۲۳	۰/۴۸۷	−۰/۳۳۹*	۰/۰۰۶	۰/۱۴۵	−۰/۰۰۵
۰/۶۵	۱۳/۰۵۵	۰/۳۱۶*	۰/۳۰۹	−۰/۲۸۸*	−۰/۰۰۱	۰/۱۳۱	۰/۰۲۴
۰/۷۵	۱۱/۱۲۸	۰/۳۵۷*	۰/۲۷۸	−۰/۲۳۳*	−۰/۰۰۵	۰/۰۹۳	−۰/۰۱۸
۰/۸۵	−۲/۳۵۷	۰/۳۸۵	۰/۲۶۲	−۰/۲۲۹	−۰/۰۱۶*	۰/۲۵۵	۰/۰۱۸*
۰/۹۵	−۱/۳۸۱	۰/۳۸۰	۰/۲۳۸	−۰/۲۳۶	−۰/۰۱۷*	۰/۲۵۲	۰/۰۱۸*

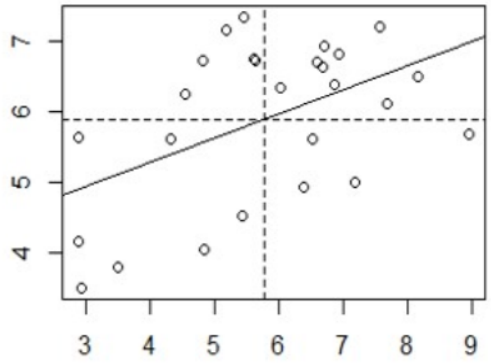
* معنی‌دار در سطح $\alpha = ۰/۱$

برآورد ضرایب برای هر چندک در جدول ۳ ارائه شده است. همانطور که ملاحظه می‌شود ضرایب X_1، X_2، X_3 و X_4 در کمیت انتخاب شده معنی‌دار هستند. ضرایب مدل SARQR در چندک‌های مختلف متفاوت هستند، برخی مثبت، برخی دیگر منفی و تعدادی نزدیک به صفر هستند. بهترین مدل براساس کوچکترین مقدار معیار اطلاع آکائیکه (AIC) (یاسین و همکاران ، ۲۰۲۰)، ارائه شده در جدول ۴ انتخاب می‌شود. همانطور که ملاحظه می‌شود تمام pـ مقدارها برای آزمون وابستگی فضایی و آزمون تنوع فضایی بزرگتر از ۰/۰۵ است. در نتیجه در مدل SARQR ، هیچ وابستگی فضایی به متغیر پاسخ برای هر چندک وجود ندارد و SARQR می‌تواند بر مشکل وابستگی فضایی غلبه کند. علاوه بر این، برای هر مدل pـ مقدار بیشتر از ۰/۰۵ بیانگر آن است که دیگر مشکلی با تنوع فضایی بین منطقه‌ای وجود ندارد و مدل SARQR در همه چندک‌ها دارای مقدار AIC کمتری نسبت به مدل SAR است، AIC در چندک ۰/۴۵ دارای کمترین مقدار است.

جدول ۴: مقایسه مدل‌های SAR و SARQR

مدل	τ	آزمون وابی فضایی		آزمون چندگونگی فضایی		AIC
		مقدار LM	p مقدار	مقدار BP	p مقدار	
SAR	–	۵/۸۶۰۱	۰/۰۱۵۵	۳/۱۶۷۴	۰/۰۷۵۱	۱۰۵/۲۲۰۰
	۰/۱۵	۰/۰۰۵۲	۰/۹۴۲۶	۱/۴۹۵۸	۰/۲۲۱۳	۷۲/۳۵۹۶
	۰/۲۵	۲/۷۳۴۳	۰/۰۹۸۲	۳/۰۶۳۱	۰/۰۸۰۱	۷۲/۰۶۵۳
	۰/۳۵	۲/۲۴۸۴	۰/۱۳۳۸	۰/۷۱۷۹	۰/۳۹۶۸	۷۱/۱۲۶۰
	۰/۴۵	۳/۲۹۴۷	۰/۰۶۹۵	۰/۷۲۳۸	۰/۳۹۴۹	۶۳/۲۱۱۳
SARQR	۰/۵۵	۲/۴۸۹۳	۰/۱۱۴۶	۰/۹۰۲۰	۰/۳۴۲۳	۷۶/۹۱۲۳
	۰/۶۵	۰/۰۳۰۰	۰/۸۶۲۵	۱/۴۱۳۶	۰/۲۳۴۵	۸۸/۵۳۰۶
	۰/۷۵	۲/۶۱۶۴	۰/۱۰۵۸	۱/۱۵۳۸	۰/۲۸۲۷	۸۸/۱۳۳۳
	۰/۸۵	۱/۸۳۴۳	۰/۱۷۵۶	۱/۰۷۳۶	۰/۳۰۰۱	۷۷/۷۹۸۹
	۰/۹۵	۲/۳۴۷۸	۰/۱۲۵۵	۰/۹۴۵۷	۰/۳۳۰۸	۷۸/۸۹۱۰

به عبارت دیگر نقاط پرت فضایی در این چندک شناسایی می‌شوند. شکل ۲ با چندک ۰/۴۵ هیچ نقطه پرت فضایی وجود ندارد، به این معنی که دیگر نقاط پرت در این مدل SARQR وجود ندارد.



شکل ۲: آزمون نقاط پرت فضایی براساس مدل SARQR در چندک ۰/۴۵

بحث و نتیجه‌گیری

مقایسه دو مدل SARQR و SAR براساس معیار AIC ، بیانگر برتری مدل SARQR برای پیش‌گویی مقدار OUL است. همچنین مدل SARQR با اثر فضایی سروکار دارد و به راحتی تحت تأثیر داده‌های پرت فضایی قرار نمی‌گیرد، این مدل می‌تواند اطلاعات را برای هر یک از چندک‌های انتخاب شدهٔ توزیع پاسخ ارائه دهد، در حالی که مدل SAR فقط می‌تواند میانگین پاسخ را پیش‌گویی کند. برای تحقیقات بیشتر، بررسی، ارزیابی و مقایشه کارایی روش‌های دیگر برآورد پارامترها مانند روش شبه ماکسیمم درستنمایی، روش گشتاوری تعمیم‌یافته و روش کمترین توان‌های دوم دو مرحله‌ای توصیه می‌شود.

مراجع

Anselin, L. (1988), The Scope of Spatial Econometrics, *Spatial Econometrics: Methods and Models*, **3**(4), 7–15.

Breusch, T. S. and Pagan, A. R. (1979), A Simple Test for Heteroscedasticity and Random Coefficient Variation, *Econometrica*, **47**(5), 1287-1294.

Dai, X. and Jin, L. (2021), Minimum Distance Quantile Regression for Spatial Autoregressive Panel Data Models with Fixed Effects, *Plos One* , **16**(12), e0261144.

Dai, X., Yan, Z. Tian, M. and Tang, M. (2020), Quantile Regression for General Spatial Panel Data Models With Fixed Effects, *Journal of Applied Statistics*, **47**(1), 45–60.

Huang, H., Abdel Aty, M. A. and Darwiche, A. L. (2010), County Level Crash Risk Analysis in Florida: Bayesian Spatial Modeling, *Transportation Research Record*, **2148**(1), 27–37.

Jin, L., X. Dai, A. Shi, and L. Shi (2016), Detection of Outliers in Mixed Regressive-spatial Autoregressive Models, *Communications in Statistics Theory and Methods*, **45**(17), 5179–5192.

Lum, K. and Gelfand, A. E. (2012), Spatial Quantile Multiple Regression Using the Asymmetric Laplace Process, *Bayesian Analysis*, **7**, 235–258.

Su, L. and Yang, Z. (2011), Instrumental Variable Quantile Estimation of Spatial Autoregressive Models, *Research Collecti School Of Economics*, **7**, 1–35.

Tribhuwaneswari, A. B., Hapsery, A. and Rahayu, W. K. (2022), Spatial Autoregressive Quantile Regression as A Tool for Modelling Human Development Index Factors in 2020 East Java, *AIP Conference Proceedings, Vol. 2668*, https://doi.org/10.1063/5.0112828

Ver Hoef, J. M., Peterson, E. E. Hooten, M. B. Hanks, E. M. and Fortin, M. J. (2018), Spatial Autoregressive Models for Statistical Inference from Ecological Data,*Ecological Monographs*, **88**(1), 36–59.

Xu, P. and Huang, H. (2015), Modeling Crash Spatial Heterogeneity: Random Parameter Versus Geographically Weighting, *Accident Analysis and Prevention*, **75**, 16–25.

Yanuar, F., Deva, A. S. and Zetra, A. (2023), Length of Hospital Stay Model of COVID-19 Patients with Quantile Bayesian with Penalty LASSO, *Communications in Mathematical Biology and Neuroscience*, Article–ID 23.

Yanuar, F., Zetra, A. Muharisa, C. Devianto, D. Putri, A. R. and Asdi, Y. (2019), Bayesian Quantile Regression Method to Construct the Low Birth Weight Model, *Journal of Physics, Conference Series* 1245, 012044.

Yasin, H., Hakim, A. R. and Warsito, B. (2020), Development Life Expectancy Model in Central Java using Robust Spatial Regression with M-estimators, *Communications in Mathematical Biology and Neuroscience*, **69**, 1–16.

Yu, T., Gao, F. Liu, X. and Tang, J. (2021), A Spatial Autoregressive Quantile Regression to Examine Quantile Effects of Regional Factors on Crash Rates, *Sensors*, **22**(1), 5.

Zhang, J., Lu, Q. Guan, L. and Wang, X. (2021a), Analysis of Factors Influencing Energy Efficiency Based on Spatial Quantile Autoregression: Evidence From the Panel Data in China, *Energies*, **14**(2), 504; https://doi.org/10.3390/en14020504.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

مقایسه عملکرد گندم در نواحی آب و هوایی مختلف با ماشین بردار پشتیبان

معصومه شکیب‌فر[۱]، علی محمدیان مصمم
گروه آمار، دانشکده علوم، دانشگاه زنجان

چکیده: در این مقاله، روشی برای تحلیل و مقایسه عملکرد گندم در چهار ناحیه آب و هوایی مختلف با استفاده از الگوریتم ماشین بردار پشتیبان ارائه شده است. ماشین بردار پشتیبان از توابع هسته برای تبدیل داده‌ها به فضای برداری چند بعدی و تعیین مرزهای جداساز استفاده می‌کند. در این مطالعه، انواع مختلف هسته‌های ماشین بردار پشتیبان مورد ارزیابی قرار گرفته‌اند تا مرزهای جداساز غیرخطی بهینه برای تفکیک نواحی مختلف به دست آید. نتایج نشان داد ماشین بردار پشتیبان می‌تواند الگوهای پنهان موجود در داده‌های عملکرد گندم را شناسایی کند.

واژه‌های کلیدی: ماشین بردار پشتیبان، هسته بازآفرین، طبقه‌بندی، داده‌های ژنتیک، آمار فضایی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 68Q32، 68T05.

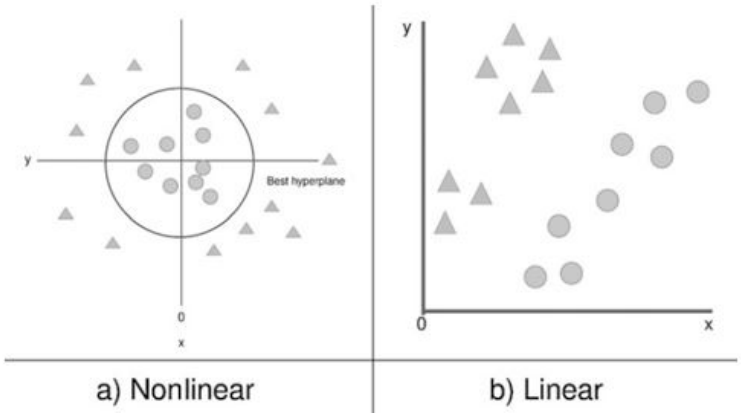
۱ مقدمه

در یادگیری ماشینی، دو مسئله بنیادین به نام‌های طبقه‌بندی و رگرسیون مورد بررسی قرار می‌گیرند. طبقه‌بندی به تخصیص داده‌ها به دسته‌های مختلف می‌پردازد در حالی که رگرسیون سعی در پیش‌بینی یا تخمین مقادیر پیوسته دارد. از روش‌های متعددی برای حل این دو مسئله استفاده می‌شود، اما یکی از روش‌های قدرتمند و پرکاربرد در این زمینه، روش ماشین بردار پشتیبان[۱] (SVM) است. SVM یک الگوریتم یادگیری ماشینی است که با تحلیل دقیق داده‌ها، مرزهای بهینه بین دسته‌ها را تشخیص می‌دهد. این روش از توابع هسته برای تبدیل داده‌های ورودی به فضای برداری با ابعاد بالا استفاده می‌کند و با بهره‌گیری از تابع هزینه و روش بهینه‌سازی، بهترین مرزهای تصمیم‌گیری را استخراج می‌نماید (شکل ۱). روش SVM از اصول ریاضیاتی گرایش داشته و در دهه ۱۹۹۰ توسط واپنیک معرفی و پیشنهاد شده است. این اصول شامل مفاهیمی چون حاشیه حداکثر، بردار پشتیبان و ترفندهای هسته می‌شوند (شکل ۲). این ایده‌ها و مفاهیم باعث بهبود فرآیندهای تصمیم‌گیری و پیش‌بینی در زمینه‌های مختلف از جمله کشاورزی و تحلیل عملکرد محصولات می‌شوند. در این مقاله، ما

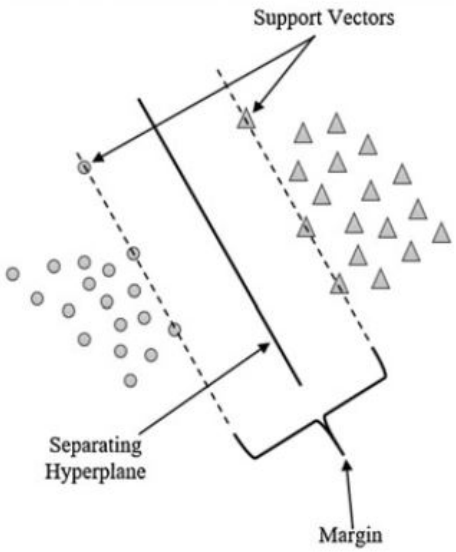
[1] Support Vector Machines

[۱] نام و ایمیل ارائه دهنده مقاله: معصومه شکیب‌فر، shakibfar@gmail.com

به بررسی و مقایسه عملکرد گندم در مناطق مختلف با توجه به شرایط آب و هوایی می‌پردازیم و از روش SVM برای این اندازه‌گیری‌ها استفاده می‌کنیم. تحلیل و نتایج این مطالعه می‌توانند به بهبود روش‌های کشاورزی و افزایش بهره‌وری در تولید گندم در شرایط محیطی مختلف کمک کنند (مونتسینوس ، ۲۰۲۲).



شکل ۱: تبدیل یک مسئله غیرخطی به یک مسئله خطی



شکل ۲: حاشیه حداکثر و بردارهای پشتیبان

۲ روش کار SVM

روش SVM برای طبقه‌بندی داده‌ها و یافتن خط جداکننده بهینه بین دسته‌ها یا خط رگرسیون، از تابع هسته[۲] استفاده می‌کند. برای درک بهتر نحوه کار SVM، می‌توان مراحل عملکرد آن را به طور خلاصه بررسی کرد:

۱. تبدیل داده‌ها به فضای ویژگی بالاتر: ابتدا داده‌ها را از فضای ورودی به یک فضای ویژگی با بعد بالاتر نگاشت می‌دهیم. این تبدیل می‌تواند با استفاده از تابع هسته انجام شود. تابع هسته نقش تابع شباهت بین داده‌ها را ایفا می‌کند و

[2]Kernel Function

مقدار آن نشان می‌دهد که داده‌ها در فضای ویژگی چقدر به هم نزدیک هستند.

۲. جدا کردن دسته‌ها یا پیدا کردن خط رگرسیون: در فضای ویژگی، SVM سعی می‌کند با استفاده از یک خط جداکننده، دسته‌ها را از هم جدا کند یا خط رگرسیون را پیدا کند. خط جداکننده باید حاشیه بین دسته‌ها را بیشینه کند، به این معنی که فاصله بین خط جداکننده و نزدیک‌ترین نقاط داده به خط، حداکثر باشد.

۳. بردارهای پشتیبان: نقاطی که نزدیک‌ترین نقاط به خط جداکننده هستند، بردارهای پشتیبان[3] نامیده می‌شوند. بردارهای پشتیبان نقش مهمی در تعیین خط جداکننده و محاسبه حاشیه (مساحت بین خط جداکننده و نقاط داده) ایفا می‌کنند.

۴. حل مسئله بهینه‌سازی: برای یافتن خط جداکننده بهینه، باید یک مسئله بهینه‌سازی را حل کنیم. این مسئله بهینه‌سازی شامل مینیمم کردن یک تابع هدف به صورت

$$
\begin{aligned}
&L(\beta,\beta_0,\alpha)=\frac{1}{2}\|\beta\|^2-\sum_{i=1}^{n}\alpha_i[y_i(\beta_0+x_i^T\beta)-1]\\
&\frac{\partial L(\beta,\beta_0,\alpha)}{\partial\beta}=\beta-\sum_{i=1}^{n}\alpha_i y_i x_i=0\Rightarrow\beta=\sum_{i=1}^{n}\alpha_i y_i x_i\\
&\frac{\partial L(\beta,\beta_0,\alpha)}{\partial\beta_0}=\sum_{i=1}^{n}\alpha_i y_i=0\Rightarrow\sum_{i=1}^{n}\alpha_i y_i=0\\
&\alpha_i[y_i(\beta_0+x_i^T\beta)-1]=0,\quad i=1,\dots,n\\
&\Rightarrow\alpha_i=0,\quad y_i(\beta_0+x_i^T\beta)=1
\end{aligned}
$$

است، که در آن $\alpha=(\alpha,\dots,\alpha_n)^T$ متغیرهای مثبت ضرایب لاگرانژ و β بردار پارامترها هستند.

۵. طبقه‌بندی داده‌های جدید: بعد از حل مسئله بهینه‌سازی و یافتن پارامترهای بهینه w و b با استفاده از رابطهٔ تابع تصمیم، می‌توانیم داده‌های جدید را در فضای ویژگی طبقه‌بندی کنیم. به‌طور کلی، اگر نقطهٔ جدید را با x نشان دهیم، تابع تصمیم SVM بررسی می‌کند که x بر روی کدام سوی خط جداکننده قرار می‌گیرد. اگر خروجی تابع تصمیم برابر با $+1$ باشد، نقطه x در دستهٔ مثبت قرار می‌گیرد. اگر خروجی تابع تصمیم برابر با -1 باشد، نقطه x در دستهٔ منفی قرار می‌گیرد (بیشاپ ، ۲۰۰۶؛ برلینت ، ۲۰۱۱؛ مونتسینوس ، ۲۰۲۲؛ آبی، ۲۰۰۵).

۳ معرفی تابع هسته

هسته[4] یک تابع متقارن و نیمه معین مثبت است که دو نمونه را به یک مقدار عددی تبدیل می‌کند. این تبدیل نشان دهنده شباهت یا فاصله میان دو نمونه است.یعنی هسته $k(x_i,x_j)$ به ما می‌گوید دو نمونه x_i و x_j چقدر به هم شبیه هستند. اگر $k(x_i,x_j)$ بزرگ باشد، نمونه‌ها به هم نزدیک‌تر هستند و اگر کوچک باشد، دورترند. هسته مناسب به مدل‌های یادگیری ماشینی کمک می‌کند تا الگوهای پیچیده‌تر و غیرخطی را شناسایی کنند (شکل ۳). بسیاری از الگوریتم‌های یادگیری ماشینی مانند ماشین‌های بردار پشتیبان و شبکه‌های عصبی از هسته‌ها استفاده می‌کنند. تابع هسته ضرب داخلی دو تابع بصورت

$$k(x_i,x_j)=\varphi(x_i)^T\varphi(x_j)$$

است. برخی از انواع توابع هسته عبارتند از (مونتسینوس ، ۲۰۲۲؛ برلینت ، ۲۰۱۱؛ آبی، ۲۰۰۵؛ رابرت ، ۲۰۱۴):

[3] Support Vectors
[4] Kernel

۱. هسته خطی (Liniear Kernel):

$$K(x, y) = x^T y$$

۲. هسته چند جمله‌ای (Polynomial Kernel):

$$K(x, y) = (x^T y + c)^d$$

که در آن c یک پارامتر ثابت و a نیز درجه چندجمله‌ای است.

۳. هسته گاوسی (Gaussian Kernl):

$$K(x, y) = \exp\left(-\frac{\|x - y\|^۲}{۲\sigma^۲}\right)$$

که در آن σ یک پارامتر مثبت است.

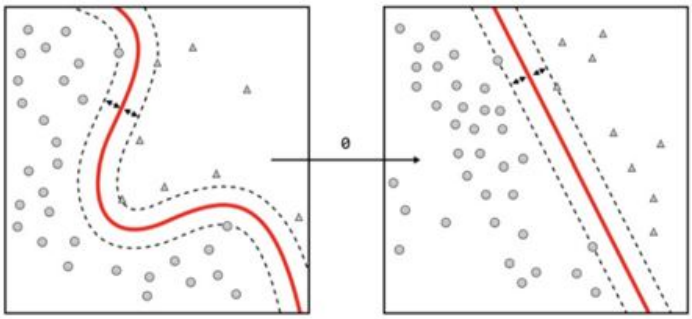
۴. هسته سیگوئیدی (Sigmoid Kernel) :

$$K(x, y) = \tanh(\alpha x^T y + c)$$

که در آن α و c پارامترهای تنظیم‌پذیر هستند.

۵. هسته توابع پایه شعاعی (Radial Basis Function Kernel):

$$K(x, y) = \exp(-\gamma\|x - y\|^۲)$$



شکل ۳: تبدیل یک مسئله پیچیده غیرخطی به یک مسئله خطی با استفاده از توابع هسته

۴ تحلیل داده‌ها

مجموعه داده واقعی مورد تحلیل حاوی اطلاعات از ۹۵۵خوشه گندم حاصل از برنامه جهانی گندم CIMMYT است. CIMMYT یک مرکز بین المللی توسعه گندم وذرت در مکزیک است. این مرکز بین‌المللی آزمایش های بسیاری را در میان محیط‌های کشت گندم متنوع انجام داده است. محیط های نشان داده شده در این آزمایش ها به چهار مجموعه هدف اصلی گروه بندی شده اند که شامل ۴ ناحیه آب و هوایی است. اخیراً ۷۴۴۱ ژنوتیپ خوشه گندم با استفاده از تکنولوژی تنوع

جدول ۱: پیش‌بینی برحسب میانگین مربعات خطا GY در چهار محیط (Env) تحت پنج روش هسته

محیط	هسته نمایی	هسته گاووسی	هسته سیگموئید	هسته چندجمله‌ای	هسته خطی
۱	۰/۸۹۱۵۷۶۲۵۵	۰/۸۹۳۰۵۳۶۵۸	۰/۷۶۱۷۸۳۱۱۴	۱/۰۰۱۱۶۸۱۰۵	۱/۰۰۹۱۲۳۴۹۴
۲	۰/۹۱۰۵۲۶۷۹۴	۰/۹۲۳۵۴۳۷۲۸	۰/۷۷۱۸۸۳۲۹	۰/۹۹۴۹۲۳۱۲۵	۰/۹۱۷۷۶۳۵۲۷
۴	۰/۹۴۸۵۰۳۴۲۲	۰/۹۴۱۰۶۶۹۷۲	۰/۸۶۰۳۶۲۱۶۴	۱/۰۰۴۳۴۷۰۱۷	۰/۹۸۱۸۹۱۹۰۳
۵	۰/۹۴۰۰۰۷۳۳۲	۰/۹۰۵۶۲۲۳۷۲	۰/۸۱۵۹۵۰۸۳۳	۱/۰۰۱۶۲۲۷۷۵	۱/۰۱۰۴۲۰۶۳

آرایه (DArT) تولید شده است. این مجموعه داده شامل اطلاعات فنوتیپی (wheat:Y) و ژنوتیپی (wheat:X) و شجره (wheat:A) مربوط به ۹۵۵خوشه گندم است. ماتریس Y یک ماتریس ۵۹۹ × ۴ بعدی است که حاوی متوسط محصول گندم (GY) دو ساله هر کدام از این خوشه ها در هر یک از ۴محیط است (مونتسینوس ، ۲۰۲۲). روی این مجموعه داده ۵ هسته مختلف را بررسی کرده و بر اساس معیار MSE مقایسه انجام شده است . طبق جدول ۱ هسته سیگموئید برای این داده نتیجه بهتری را ارائه می‌دهد و هسته چند جمله‌ای مناسب این داده نیست.

مراجع

Abe, S. (2005). *Support Vector Machines for Pattern Classification*, Springer.

Robert, C. (2014). *Machine Learning, a Probabilistic Perspective*, Taylor & Francis.

Berlinet, A. and C. Thomas-Agnan (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer Science & Business Media.

Bishop, C. M. and N. M. Nasrabadi (2006). *Pattern Recognition and Machine Learning*, Springer.

Gareth, J., et al. (2013). *An Introduction to Statistical Learning: With Applications in R*, Springer.

Montesinos López, O. A., Montesinos López, A., and Crossa, J. (2022). *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer Nature.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

مطالعه نرخ جرم تحت تأثیر پروتکل‌های بهداشتی کووید-۱۹

راضیه صابری[۱]
دانشکده حقوق، الهیات و علوم سیاسی، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران

چکیده: شیوع کووید-۱۹ از سال ۲۰۲۰ جهان را به‌شدت متأثر کرده و شواهد اولیه نشان می دهد که یکی از این تأثیرات کاهش یا افزایش نرخ جرم است.به‌نظر می‌رسد دلیل اصلی این تغییر، دستور ماندن در خانه توسط‌دولت‌ها بود که زندگی و فعالیت‌های روزمره را تحت تأثیر خود قرار داد. ازآنجاکه صدور و انجام این دستورات در زمان‌ها و به روش‌های مختلف بر شیوه وسبک زندگی مردم‌تأثیرگذاشته ،یک آزمایشگاه-طبیعی به‌وجود آمد تا امکان آزمون برخی نظریه‌های جرم‌شناختی مرتبط برای تبیین تغییر نرخ جرایم ارتکابی فراهم‌شود. در نوشتار حاضر تأثیر اجرای پروتکل بهداشتی، به ویژه فاصله گذاری اجتماعی بر نرخ جرایم کودک‌آزاری، سرقت تعزیری، کیف‌زنی و جیب‌بری، منازعه وکلاهبرداری شبکه ای(رایانه ای)در استان‌های تهران، مازندران، خراسان رضوی، فارس،آذربایجان شرقی،کردستان، سیستان و بلوچستان با بررسی تعدادی از نظریات جرم شناختی مورد مطالعه قرار گرفته است. داده های جرایم فوق مربوطبه دو دوره قبل و بعد از شیوع کووید۱۹ است که از طریق مرکزآمار و فناوری اطلاعات قوه قضاییه به دست آمد. برای تجزیه و تحلیل داده ها از روش سری‌های زمانی منقطع استفاده شده‌است. یافته های پژوهش تایید می کند که شیوع کووید۱۹ و فاصله گذاری اجتماعی الگوی فعالیت های روزمره را تغییر داده و ساختارهای فرصتی ایجاد شده در اثر این تغییرات باعث تغییر در نرخ جرایم کودک‌آزاری،کیف‌زنی و جیب‌بری و منازعه،برخلاف جرایم‌کلاهبرداری‌شبکه‌ای و سرقت تعزیری، طي دو دوره قبل و بعد از شیوع کووید۱۹ شده است.

واژه‌های کلیدی: کووید۱۹ ، پروتکل های بهداشتی، نرخ جرم، نظریه فعالیت روزمره
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62P25،.

۱ مقدمه

مقاله حاضر درباره تأثیر شیوع کووید-۱۹ در سال ۲۰۲۰ بر جوانب مختلف جهانی مانند تغییر فرهنگ، اقدامات دولتی، نرخ جرم و جنایت، اقتصاد، سیاست و تعاملات اجتماعی می‌پردازد. دستورهای رعایت فاصله اجتماعی و ماندن در خانه تأثیرات عمده‌ای داشته‌اند. تغییرات ممکن است نرخ جرم و جنایت را تحت تأثیر قرار دهند، از جمله جرائم مالی. نظریات جرم‌شناسی نیز می‌توانند در تفهیم این تغییرات کمک کنند. این مقاله به تأثیر اقدامات فاصله‌گذاری اجتماعی بر نرخ جرم

[۱] **نام و ایمیل ارائه دهنده مقاله: راضیه صابری،** *raziehsaberi۱۲۳@gmail.com*

در ایران می‌پردازد و تغییرات در جرائم مختلف را در استان‌های مختلف بررسی می‌کند. هدف از این مطالعه، تعیین تأثیر تغییرات در فرصت‌های جرمی به علت فاصله‌گذاری اجتماعی و بررسی نظریات جرم‌شناسی در این زمینه است. این مطالعه به درک بهتر الگوهای جرمی در شرایط فاصله‌گذاری اجتماعی کمک می‌کند و در مدیریت بحران و برنامه‌ریزی آینده نیز مفید است.

۲ مبانی نظری

در بخش مبانی نظری، تأثیر شیوع کووید-۱۹ بر رفتار جمعی و جرم در شرایط استثنایی مورد بررسی قرار می‌گیرد. توضیحات نظریه‌های مختلفی در این زمینه آورده شده است: نظریه تقویت بعد دینداری: این نظریه معتقد است که بحران‌ها مانند شیوع بیماری کووید-۱۹ می‌توانند دینداری افراد را تقویت کنند و به کاهش جرم‌ها منجر شوند (حاجی‌ده‌آبادی، محمدعلی ، ۱۳۹۹). نظریه انسجام اجتماعی (نوع دوستی): این نظریه تاکید دارد که در شرایط استثنایی، افراد به یکدیگر کمک می‌کنند و از رفتارهای دوستانه بیشتری بهره می‌برند که ممکن است باعث کاهش جرایم شود (زهران و همکاران ، ۲۰۰۹؛ بارتون ، ۱۹۶۹؛ هاجکینسون و آندرسن ، ۲۰۲۰؛ کریمر ، ۲۰۱۰).

نظریه بی‌نظمی اجتماعی: این نظریه معتقد است که در شرایط استثنایی، اختلال در نظم اجتماعی می‌تواند به افزایش جرم و جنایت منجر شود، زیرا برخی افراد ممکن است به ندرت به نظم پایبند باشند (حسین‌زاده‌خرمی، مهدی ، ۱۳۹۹؛ هاجکینسون و آندرسن ، ۲۰۲۰).

نظریه اقتصادی جرم: نظریه اقتصادی جرم معتقد است که بحران‌ها و رکود اقتصادی می‌توانند باعث افزایش برخی از جرایم شوند. در این شرایط، افزایش بیکاری و کاهش فرصت‌های اقتصادی ممکن است افراد را به رفتارهای غیرقانونی ترغیب کند (اوپال ، ۲۰۲۰).

نظریه فعالیت روزمره: این نظریه به تأثیر تغییرات در فرصت‌های جرم‌انگیز در شرایط استثنایی توجه دارد و تغییرات در فعالیت‌های روزمره را به عنوان عاملی مؤثر در تغییر نرخ جرم معرفی می‌کند (کوهن و فلسون ، ۱۹۷۹؛ لوبو ، ۲۰۰۲؛ هارپر و فریلینگ ، ۲۰۱۲؛ شان وارانو و همکاران ، ۲۰۱۰؛ پل کرامول و همکاران ، ۱۹۹۵). نظریه فعالیت روزمره به نظر می‌رسد که بیشترین توجیه را برای تغییرات در نرخ جرم‌ها در شرایط استثنایی دارد، زیرا به تغییر فرصت‌های جرم‌انگیز و ساختار فعالیت‌های روزمره توجه می‌کند.

۳ روش شناسی

در بخش روش شناسی، پژوهش به بررسی تأثیر شیوع کووید-۱۹ و اجرای فاصله‌گذاری اجتماعی بر نرخ جرایم گزارش‌شده به پلیس در استان‌های مختلف می‌پردازد. از روش‌های توصیفی و استنباطی برای سنجش اثرات استفاده می‌شود و برای تحلیل آماری از نرم‌افزار R و بسته‌ی forecast استفاده می‌شود. تغییرات در نرخ جرایم در دوره‌های قبل و بعد از شیوع کووید-۱۹ با استفاده از مدل‌های رگرسیون سری زمانی منقطع و با توجه به تاریخ شروع شیوع کرونا و دستور ماندن در خانه مورد بررسی قرار می‌گیرد. برای جمع‌آوری داده‌ها، نمونه‌های استان‌ها به نحوی انتخاب شده‌اند که هر استان نماینده‌ای از نقاط جغرافیایی مختلف باشد. این انتخاب به تعمیم نتایج به استان‌های دیگر از طریق اعتماد به نتایج کمک می‌کند. برای تحلیل تأثیرات شیوع کووید-۱۹ و فاصله‌گذاری اجتماعی بر نرخ جرایم، داده‌های دوره قبل و بعد از شیوع کووید-۱۹ مقایسه می‌شوند. از داده‌های گزارش شده به پلیس از اسفند ۱۳۹۸ تا اسفند ۱۳۹۹ استفاده شده و با داده‌های مشابه از فروردین ۱۳۹۸ تا بهمن ۱۳۹۸ مقایسه می‌شوند. برای تحلیل داده‌ها و برآورد تأثیرات، از مدل‌های رگرسیون سری زمانی منقطع استفاده می‌شود. این مدل‌ها قادر به تخمین روندهای خطی تعداد جرم قبل از دستور ماندن در خانه، تغییر در

جدول ۱: درصد کاهش یا افزایش جرم‌به‌تفکیک استان طی‌دو دوره قبل و بعد ازشیوع بیماری‌کووید۱۹

	آذربایجان‌شرقی	تهران	خراسان‌رضوی	سیستان وبلوچستان	فارس	کردستان	مازندران	کل
کودک‌آزاری	−۰/۵۱	−۸/۷۱	−۱۰/۸	−۳۱/۶۷	−۲۷/۶۹	۳۹/۵۳	−۸/۳۸	−۸/۷۵
کلاهبرداری شبکه‌ای	۱۱/۷۷	۲۵/۶۴	−۴۳/۴۴	−۵۳/۵۲	۷۰/۱۵	۴۱/۳۳	−۳۷/۲۷	۱۶/۰۳
سرقت تعزیری	۱۸/۱۰	۸/۸۶	−۵/۱۵	−۱۳/۷۶	۲/۳۶	۴/۷۳	−۱۱/۳	۳/۳۴
کیف‌زنی وجیب‌بری	−۴۶/۸۸	−۳۱/۹۶	−۷۰/۵۴	−۱۰۰	−۶۳/۶۴	۳۰۷/۱۴	۲/۵۶	−۲۶/۰۱
منازعه	۱/۲۳	−۲/۷۵	−۱۹/۸۴	−۴۵/۶۸	۲۲/۹۱	−۱۴/۰۴	۳۰/۱	−۸/۲۹

تعداد جرم در زمان شیوع کووید_۱۹، و روند خطی تعداد جرم پس از دستور ماندن در خانه هستند. نرم‌افزار R و بسته‌ی forecast برای برازش مدل‌ها به داده‌ها استفاده می‌شود. در کل، این روش با استفاده از مدل‌های آماری سعی دارد تا تأثیرات شیوع کووید_۱۹ و اجرای فاصله‌گذاری اجتماعی بر نرخ جرایم را در استان‌های مختلف مورد بررسی و ارزیابی قرار دهد و تغییرات در طول زمان را از قبل تا بعد از شیوع کرونا مورد ارزیابی قرار دهد.

۴ یافته های پژوهش

در این بخش، داده‌های مربوط به نرخ ماهانه جرایم مختلف از جمله کودک‌آزاری، سرقت تعزیری، کیف‌زنی و جیب‌بری، منازعه، و کلاهبرداری شبکه‌ای (رایانه‌ای) در استان‌های تهران، مازندران، خراسان رضوی، فارس، آذربایجان شرقی، کردستان، و سیستان و بلوچستان، از فروردین ماه سال ۱۳۹۸ تا اسفند ماه سال ۱۳۹۹ مورد بررسی قرار گرفته است. در جدول شماره ۱، نتایج بررسی درصد کاهش یا افزایش جرایم مورد مطالعه در دوره‌های قبل و بعد از شیوع بیماری کووید_۱۹ بر اساس استان‌ها نمایش داده شده است. درست است که میانگین کلی استان‌ها در خصوص جرایم کودک‌آزاری، منازعه و کیف‌زنی و جیب‌بری کاهش را نشان می‌دهد، اما وقتی استان‌ها به‌صورت جداگانه مورد بررسی قرار می‌گیرند، تفاوت‌هایی در روند افزایش یا کاهش جرایم مشاهده می‌شود. به‌عبارت دیگر، اگرچه به‌طور کلی نتیجه‌گیری می‌شود که جرایم در دوره پس‌از شیوع کووید_۱۹ کاهش یافته‌اند، اما زمانی که استان‌ها به‌تفکیک جرم و نوع جرم مورد نظر قرار می‌گیرند، این تفاوت‌ها مشخص می‌شود که در برخی استان‌ها ممکن است روند افزایشی داشته باشند.

د ر مورد جرایم سرقت تعزیری و کلاهبرداری شبکه‌ای، میانگین کلی استان‌ها در دوره پس‌از شیوع کووید_۱۹ افزایش نشان می‌دهد. اما با توجه به تفکیک جرم بر اساس استان‌ها، در برخی استان‌ها مثل خراسان‌رضوی، سیستان و بلوچستان و مازندران، تعداد جرایم این دو نوع کاهش یافته است. به‌طور کلی، نتایج این جدول نشان می‌دهند که تفاوت‌های معنی‌داری در روند افزایش یا کاهش جرایم در استان‌ها و نوع‌های جرم مختلف در دوره پس‌از شیوع کووید_۱۹ وجود دارد.

در جدول ۲ و ۳، به‌منظور بررسی معنی‌داری پارامتر $\hat{\beta}_1$ از پی_مقدار گزارش‌شده استفاده می‌شود. معنی‌داری این پارامتر نشان‌دهنده تأثیر معنی‌دار بیماری کووید_۱۹ بر جرائم مذکور با در نظر گرفتن سطح اهمیت پی_مقدار است. در واقع، هر چه پی_مقدار کوچک‌تر باشد، تأثیر بیشتری از شیوع کووید_۱۹ بر رخداد جرائم نشان داده می‌شود. از جدول ۲ مشخص می‌شود که اجرای سیاست ماندن در خانه به همراه تغییرات معنی‌داری در نرخ رخداد جرم کودک‌آزاری در استان‌های آذربایجان شرقی، تهران، خراسان رضوی، سیستان و بلوچستان، فارس و مازندران همراه نبوده است. با این حال، این تغییرات در استان کردستان مشاهده می‌شود که نرخ کودک‌آزاری پس از شیوع کرونا افزایش یافته است. باید توجه داشت که در کل استان‌های مورد بررسی، به جز کردستان، تعداد گزارشات کودک‌آزاری به‌طور توصیفی کاهش یافته است.

هم‌چنین، از پی_مقادیر در جدول ۲ برای جرم کلاهبرداری شبکه‌ای استفاده شده است. این جدول نشان می‌دهد که در استان‌های تهران، فارس و کردستان تغییرات معناداری در نرخ رخداد این جرم وجود دارد. با اینکه در استان‌های آذربایجان شرقی، خراسان رضوی، سیستان و بلوچستان و مازندران تعداد گزارشات کلاهبرداری تقریباً ثابت مانده است، اما با توجه

جدول ۲: نتایج برازش مدل (۱) به داده‌های تعداد رخداد جرائم در استان‌ها

نوع جرم	استان	$\hat{\beta}_1$	پی‌ـ‌مقدار
كودك‌آزاري	آذربایجان	–	۱
	تهران	–	۱
	خراسان‌رضوی	–	۱
	سیستان وبلوچستان	–	۱
	فارس	–	۱
	کردستان	۲/۸	۰/۰۷
	مازندران	–	۱
	کل	–	۱
کلاهبرداری شبکه ای	آذربایجان	۵۶/۶۲	۰/۱۳
	تهران	۲۵/۷۴	۰/۰۲
	خراسان‌رضوی	–	۱
	سیستان وبلوچستان	–	۱
	فارس	۵۶/۷۴	۰/۰۱
	کردستان	۲/۸	۰/۰۷
	مازندران	–	۱
	کل	۵۴/۸۸	۰/۱۴
سرقت تعزیری	آذربایجان	۲۹۵/۶۹۲	۰/۰۱
	تهران	۱۶۰۷/۶۵	۰/۰۵
	خراسان‌رضوی	–	۱
	سیستان وبلوچستان	–	۱
	فارس	۰/۰۸	۰/۵۳
	کردستان	۴۸/۱	۰/۲۷
	مازندران	–	۱
	کل	۱۱۵۰/۲۴	۰/۳۸
کیف زنی و جیب بری	آذربایجان	–	۱
	تهران	–	۱
	خراسان‌رضوی	–	۱
	سیستان وبلوچستان	–	۱
	فارس	–	۱
	کردستان	۹/۵۵	۰/۰۰۳
	مازندران	۱/۳	۰/۴۵
	کل	–	۱

جدول ۳: نتایج برازش مدل (۱) به داده‌های تعداد رخداد جرائم در استان‌ها (ادامه)

نوع جرم	استان	$\hat{\beta}_1$	پی‌ـ‌مقدار
منازعه	آذربایجان	–	۱
	تهران	–	۱
	خراسان‌رضوی	–	۱
	سیستان وبلوچستان	–	۱
	فارس	۱۲/۷۱	۰/۰۲
	کردستان	–	۱
	مازندران	۳/۳۵	۰/۴
	کل	–	۱

به پی‌ـ مقادیر گزارش شده، تأثیر معنی‌داری از شیوع کووید‌ـ۱۹ بر رخداد این جرم وجود ندارد.

در مورد سرقت تعزیری، نیز مشاهده می‌شود که تغییرات معنی‌داری در تعداد رخدادها در استان‌های آذربایجان و تهران رخ داده‌اند. با اینکه در سایر استان‌ها تعداد گزارشات سرقت تعزیری کاهش یافته است، اما افزایش قابل توجه‌ای در استان‌های آذربایجان و تهران دیده می‌شود. با این حال، بر اساس پی‌ـ مقادیر گزارش شده، این تغییرات به‌طور آماری معنی‌دار نیستند.

همچنین در مورد منازعات، در برخی استان‌ها تغییرات معنی‌داری در تعداد رخدادها مشاهده نمی‌شود، اما در استان فارس تغییرات معنی‌داری در تعداد گزارشات منازعه رخ داده است. همچنین، تعداد گزارشات منازعه در استان کردستان نیز افزایش معنی‌داری داشته و میانگین گزارشات در ماه پس از شیوع کووید‌ـ۱۹ افزایش یافته است. در مجموع، نتایج نشان می‌دهند که تأثیرات مختلفی از شیوع کووید‌ـ۱۹ بر جرائم در استان‌های مختلف وجود دارد، و در برخی از استان‌ها تغییرات معنی‌داری در تعداد رخدادها دیده می‌شود.

بحث و نتیجه‌گیری

در این مطالعه، تأثیر شیوع کووید‌ـ۱۹ بر نرخ جرایم مورد بررسی قرار گرفته است. نتایج نشان می‌دهند که نرخ جرم کودک‌آزاری در ماه‌های ابتدایی شیوع بیماری به دلیل سیاست‌های ماندن در خانه کاهش یافته و پس از کاهش این سیاست‌ها و افزایش تعاملات اجتماعی، دوباره افزایش یافته است. این تغییرات می‌توانند به عوامل متعددی نظیر تحرک اجتماعی، کاهش نیروی پلیس، مشکلات گزارش‌دهی، تغییرات اقتصادی و روانشناختی بازگردانده شوند. به طور خلاصه، تحلیل نشان می‌دهد که تأثیرات متنوعی، از جمله محدودیت‌های اجتماعی و تغییرات اقتصادی و روانشناختی، بر روند جرم کودک‌آزاری در دوره‌های مختلف شیوع کووید‌ـ۱۹ تأثیر داشته‌اند.

در دوره شیوع کووید‌ـ۱۹، جرم کلاهبرداری شبکه‌ای در کل کشور و در برخی استان‌ها نظیر تهران، فارس، کردستان و آذربایجان شرقی افزایش یافته است. این نتایج با فرضیه اصلی تحقیق که افزایش جرم کلاهبرداری شبکه‌ای در دوره شیوع بیماری را پیش‌بینی می‌کند، سازگار است. این فرض مبتنی بر تأثیر محدودیت‌های اجتماعی و افزایش استفاده از اینترنت در ایجاد زمینه‌ای برای افزایش جرایم آنلاین است. به عبارت دیگر، افزایش حضور بزه‌دیدگان بالقوه، بزه‌کاران با انگیزه و بدون نگهبانی‌توانمند در فضای مجازی، منجر به افزایش جرم کلاهبرداری شبکه‌ای می‌شود. از این رو، می‌توان نتیجه گرفت که افزایش جرم کلاهبرداری شبکه‌ای در دوران پس از شیوع بیماری کووید در ایران با توجه به نظریه فرصت قابل توجیه است.

در دوره شیوع کووید‌ـ۱۹، جرم سرقت تعزیری در ایران به طرز مشابه کلاهبرداری شبکه‌ای افزایش یافته است، به ویژه در استان‌های تهران، فارس، کردستان و آذربایجان شرقی. تحلیل این جرم با پیچیدگی مواجه است، زیرا داده‌های تفکیکی بر انواع سرقت‌ها وجود ندارد. نظریه فرصت می‌پیش‌بیند که سرقت‌های مسکونی به دلیل حضور دائمی افراد در منازل افزایش یابند، و همچنین سرقت‌های اتومبیل و مغازه‌ها به دلیل نظارت ناکافی در دوره شیوع بیماری افزایش خواهند داشت. با وجود کمبود داده‌ها، نظریه‌های مبتنی بر بی‌نظمی‌اجتماعی یا بحران اقتصادی نمی‌توانند به طور کامل توجیه افزایش جرم سرقت تعزیری در ایران در دوره پس از شیوع بیماری را ارائه دهند. زیرا در دوره‌های اولیه شیوع، با تاکید دولت بر سیاست ماندن در خانه، جرایم سرقت تعزیری کاهش یافته است. همچنین، در ماه‌هایی که اجرای جدی سیاست ماندن در خانه مجبور به خانه‌نشینی مردم کرده، جرایم سرقت تعزیری کاهش می‌یابند. اما در ماه‌هایی که اجرای سیاست ماندن در خانه نادیده گرفته می‌شود، افزایش جرم مشاهده می‌شود. به‌طور کلی، نظریه فرصت بهترین توجیه را برای افزایش جرم سرقت تعزیری در دوره پس از شیوع بیماری ارائه می‌دهد و ارتباط معکوس بین اجرای جدی سیاست ماندن در خانه و نرخ جرایم تعزیری این نظریه را تقویت می‌کند.

در دوره شیوع کووید‌ـ۱۹، جرم‌های جیب‌بری و کیف‌زنی در اکثر استان‌ها به طور کلی کاهش یافته است. این کاهش

به تغییر در روند فعالیت‌های روزمره و احتمال ابتلا به بیماری بازمی‌گردد؛ زیرا تماس‌های اجتماعی و نزدیکی‌ها کاهش یافته و افراد از ارتکاب این جرم‌ها منصرف می‌شوند. با این حال، در برخی مناطق نظیر کردستان و مازندران، جرم‌های جیب‌بری و کیف‌زنی افزایش یافته است؛ این تغییرات ممکن است به دلیل مشکلات اقتصادی و فقر بیشتر، کاهش ناظرین، و کم‌شدن حضور اجتماعی در نظر گرفته شود.

به طور کلی، نظریه‌ی فرصت با درنظر گرفتن تأثیرات مختلف اقتصادی، تحرک افراد، و اجرای پروتکل‌های بهداشتی، می‌تواند تفسیری مناسب برای تغییرات در نرخ جرم‌های جیب‌بری و کیف‌زنی در دوره کووید_۱۹ ارائه دهد.

در دوره شیوع کووید_۱۹، نرخ جرم منازعه در کشور و برخی استان‌ها کاهش یافته است، همچنین در برخی استان‌ها افزایش یافته است. این تغییرات با تحلیل نظریه فرصت در دوران بحران همخوانی دارد؛ به‌عبارت دیگر، کمتر شدن فرصت‌های ارتکاب جرم به دلیل تغییر در فعالیت‌های روزمره و محدودیت‌های اجتماعی و بهداشتی در دوره کووید_۱۹ است.

نتایج نشان می‌دهند که افت نرخ جرم منازعه در دوره کووید_۱۹ به‌طور عمده به تغییرات در روند فعالیت‌های روزمره و کاهش فرصت‌های ارتکاب جرم برمی‌گردد. همچنین، تفسیرها نشان می‌دهند که در برخی مناطق افزایش نرخ جرم منازعه نیز ممکن است با توجه به تشدید محدودیت‌ها و تأثیرات روانی رخ داده باشد.

مراجع

حاجی‌ده‌آبادی، محمدعلی. (۱۳۹۹). «از بحران کرونا تا بحران سیاست‌جنایی»، مجله حقوق اسلامی، ۶۴، ۱۱۱ ـ ۱۳۳.

رحمانی، جبار. (۱۳۹۹). «سبک زندگی‌مراقبتی در شرایط بحران‌کرونا و سهم‌اصحاب علوم‌انسانی در آن»، جستارهایی در ابعاد فرهنگی و اجتماعی بحران‌ویروس کرونا در ایران، تهران، پژوهشکده مطالعات فرهنگی و اجتماعی وزارت علوم تحقیقات و فناوری، ۱۷۱ ـ ۱۸۷.

حسین‌زاده‌خرمی، مهدی. (۱۳۹۹). «کرونا مثابه آزمایش نقض‌کننده (جستاری در گسست واقعیت‌های اجتماعی جامعه ایرانی»، جستارهایی در ابعاد فرهنگی و اجتماعی بحران ویروس کرونا در ایران، تهران، پژوهشکده مطالعات فرهنگی و اجتماعی وزارت علوم تحقیقات و فناوری، ۹۱ ـ ۹۹.

Zahran, Sammy, Tara O'Connor Shelly, Lori Peek, and Samuel Brody. (2009). "Natural Disaster and Social Order: Modelling Crime Outcomes and Disasters in Florida", *International Journal of Mass Emergencies and Disasters*, 27, 26–52.

Barton, A. H. (1969). *Communities in Disaster: A Sociological Analysis of Collective Stress Situations*. Garden City, NY, Doubleday & Company.

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

Hodgkinson, Tarah, and Martin Andresen. (2020). "Show Me a Man or a Woman Alone and I'll Show You a Saint: Changes in the Frequency of Criminal Incidents During the COVID-19 Pandemic", *Journal of Criminal Justice*, 69, 1-13.

Uppal, Pradyumna. (2020). "COVID-19 Will Lead to Inceased Crime Rates in India", *International Journal of Research – Granthaalayah*, 8, 72-78.

Cohen, L. E., and M. Felson. (1979). “Social Change and Crime Rate Trends: A Routine Activity Approach”, *American Sociological Review*, 44, 588–608.

Craemer, T. (2010). “Evaluating Racial Disparities in Hurricane Katrina Relief Using Direct Trailer Counts in New Orleans and FEMA Records”, *Public Administration Review*, 70, 367–377.

LeBeau, J. L. (2002). “The impact of a Hurricane on Routine Activities and on Calls for Police Service: Charlotte, North Carolina, and Hurricane Hugo”, *Crime Prevention and Community Safety*, 4, 53–64.

Harper, D. W., and K. Frailing. (2012). *Crime and Criminal Justice in Disaster*. NC, Carolina University Press.

Varano, Sean, Joseph Schafer, Jeffrey Cancino, Scott Decker, and Jack Greene. (2010). “A tale of Three Cities: Crime and iDsplacement After Hurricane Katrina”, *Journal of Criminal Justice*, 38, 42–50.

Cromwell, Paul, Roger Dunham, Ronald Akers, and Lonn Lanza-Kaduce. (1995). “Routine Activities and Social Control in the Aftermath of a Natural Catastrophe”, *European Journal on Criminal Policy*, 3, 56–69.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

کاربست رهیافت مدل بیزی در داده‌های گسسته فضایی–زمانی

الناز عباسی[1]**، علی م. مصمم**
گروه آمار، دانشگاه زنجان

چکیده:

در تحلیل بیزی داده‌ها ی فضایی–زمانی جرم و جنایت گاهی با داده‌های گسسته‌ای مواجه می‌شویم که به دلیل ناگاوسی بودن توزیع متغیر پاسخ و وجود تعداد زیادی متغیر پنهان در مدل تحت بررسی شکل بسته‌ای برای توزیع پسینی وجود ندارد. در این شرایط استفاده از روش‌های مونت‌کارلوی زنجیر مارکف $(MCMC)$ با چالش‌هایی نظیر، وجود پارامترهای زیاد در ساختار سلسله مراتبی، محاسبات سنگین، شبیه‌سازی گسترده، طولانی و زمان‌بر بودن محاسبات به‌ویژه زمانی که بعد میدان تصادفی بزرگ است و عدم همگرایی توزیع پسینی مواجه هستیم. در این مقاله در یک مطالعه موردی به روش تقریبی لاپلاس آشیانی جمع بسته INLA تحلیل داده‌های جرم و جنایت کشور کانادا می‌پردازیم. این روش قادر است برآوردهایی از منظر وقوع جرم وجنایت در مکان و زمان معین ارائه کرده و همچنین نواحی با رفتار غیر معمول را تشخیص دهد.

واژه‌های کلیدی: تقریب لاپلاس آشیانی جمع بسته، تحلیل سلسله مراتبی بیزی، آمار فضایی–زمانی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 33C45، 62H11.

۱ مقدمه

تحلیل بیزی داده‌های جرم و جنایت معمولاً به صورت استنباط بیزی الگوهای فضایی محض یا الگوهای زمانی محض انجام می‌گیرد. اگر مشاهدات به‌گونه‌ای باشند که مشاهدات نزدیک به هم وابسته‌تر و مشاهدات دورتر وابستگی کمتری داشته باشند، این‌گونه مشاهدات داده‌های فضایی نامیده می‌شوند. بدیهی است که تحلیل آماری داده‌های فضایی با روش‌های آماری معمول مقدور نیست، زیرا شرط اساسی استقلال داده‌ها تحقق پیدا نمی‌کند. با این حال درمورد داده‌های جرم و جنایت چنین تحلیل‌های بیزی صرفاً فضایی یا زمانی به دلیل اینکه اثر متقابل فضا و زمان را در نظر نمی‌گیرند مدل‌های مناسبی نیستند. در این مقاله سعی شده است که با استفاده از مدل‌های بیزی فضایی–زمانی به تحلیل و تفسیر این اثر نیز پرداخته شود. مشاهداتی که هم از نظر موقعیت فضایی و هم از نظر موقعیت زمانی وابسته باشند، داده‌های فضایی–زمانی نامیده می‌شوند. در آمار فضایی–زمانی به طور معمول یک میدان تصادفی به عنوان مدل آماری داده‌های فضایی–زمانی در نظر گرفته می‌شود.

[1] نام و ایمیل ارائه دهنده مقاله: الناز عباسی، elnaz.Abbasi@znu.ac.ir

میدان تصادفی فضایی_زمانی مجموعه‌ای از متغیرهای تصادفی مانند $Z(\cdot,\cdot) = \{Z(s,t); (s,t) \in D \times T\}$ است، که در آن s موقعیت فضایی در مجموعه $D \subseteq R^d$، $d \geq ۱$ و t لحظه زمانی در مجموعه $T \subseteq R$ است.

روش مرسوم برای محاسبات بیزی استفاده از روش‌های مونت‌کارلوی زنجیر مارکف $(MCMC)$ مشکلات فراوانی به دلایلی نظیر وابستگی شدید بین مشاهدات و بعد زیاد ابرپارامترها در محاسبات توزیع‌های پسینی را تجربه کرده است. لذا هدف این مقاله استفاده از روش تقریب لاپلاس آشیانی جمع بسته $(INLA)$ برای تحلیل چنین داده‌هایی است. روش $INLA$ اولین بار توسط (رو و همکاران ، ۲۰۰۹) ارائه شد و سپس توسط (مارتین و همکاران ، ۲۰۱۳) و (رو و همکارن ، ۲۰۱۷) توسعه پیدا کرد.

ادامه ساختار مقاله به صورت زیر می‌باشد. در بخش ۲ روش $INLA$ معرفی می‌شود. در بخش ۳ به تحلیل اکتشافی داده ها ی جرم و جنایت می پردازیم. در بخش ۴ داده‌های تماس به پلیس برای گزارش جرم و جنایت مدل بندی می‌شوند، پارامترهای مدل برآورده شده و با استفاده از معیار بیزی اطلاع انحراف DIC بهترین مدل انتخاب می‌شود.

۲ تقریب لاپلاس آشیانی جمع بسته

اغلب مدل‌های بیزی در رگرسیون، آمار فضایی و آمار فضایی_زمانی دارای ساختار گاوسی پنهان می‌باشند. علاوه بر این در کاربست INLA نیازمند یک میدان تصادفی مارکف گاوسی [۱] می‌باشیم. $GMRF$ یک میدان تصادفی گاوسی با خاصیت استقلال شرطی است یعنی x و x' به شرط سایر مؤلفه‌ها مستقل هستند اگر و تنها اگر i و jامین مؤلفه ماتریس دقت Q صفر هستند. برای توصیف مفاهیم فوق مدل سلسله مراتبی سه مرحله‌ای زیر را در نظر بگیرید، فرض کنید که در مرحله اول متغیر پاسخ $y = (y_۱, \cdots, y_n)'$ به‌طور شرطی مستقل و دارای توزیع نمایی خاصی باشد

$$y|x, \theta_۱ \sim \Pi_{i=۱}^{n} P(y_i|x_i; \theta_۱). \qquad (۱.۲)$$

در مرحله دوم x را یک میدان تصادفی گاوسی پنهان با تابع چگالی

$$P(x|\theta_۲) \propto |Q_{\theta_۲}|_{+}^{۱/۲} \exp(-۱/۲ x' Q_{\theta_۲} x) \qquad (۲.۲)$$

در نظر بگیرید که در آن ماتریس دقت $Q_{\theta_۲}$ یک ماتریس معین مثبت با پارامتر $\theta_۲$ و $|Q_{\theta_۲}|_{+}$ حاصل‌ضرب مقادیر ویژه غیر صفر آن می‌باشد و وارون $Q_{\theta_۲}$ همان ماتریس واریانس کواریانس می‌باشد. در مرحله نهایی فرض کنید که پارامتر $\theta = (\theta_۱, \theta_۲)$ دارای یک توزیع پیشینی $\pi(\theta)$ است. در نتیجه پسینی پارامترهای مدل سلسله مراتبی به صورت

$$\begin{aligned} \pi(\eta, \theta|y) &\propto \pi(\theta)\pi(\eta|\theta_۲)\Pi_{i=۱}^{n}\pi(y_i|\eta_i; \theta_۱) \\ &\propto \pi(\theta)|Q_{\theta_۲}|_{+}^{۱/۲} \exp\{-۱/۲ \eta' Q_{\theta_۲}\eta + \Sigma_{i=۱}^{n} \log P(y_i|\eta_{i;\theta_۱})\} \end{aligned} \qquad (۳.۲)$$

می‌باشد. توزیع‌های پسینی حاشیه‌ای برای متغیرهای پنهان و ابر پارامترها به صورت

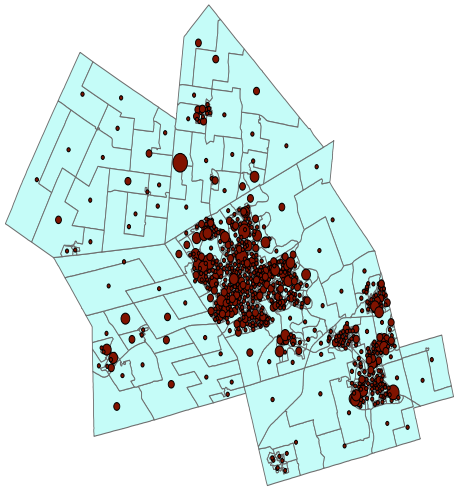
$$\begin{aligned} \pi(\eta_i|y) &= \int \pi(\eta_i|\theta, y)\pi(\theta|y)d\theta \\ \pi(\theta_i|y) &= \int \pi(\theta|y)d\theta_{-i} \end{aligned} \qquad (۴.۲)$$

به‌دست می‌آیند. (رو و همکارن ، ۲۰۱۷) یک تقریب لاپلاس برای توزیع‌های حاشیه‌ای پسینی ارائه کردند. با استفاده از روش تقریب لاپلاس در INLA توزیع‌های پسینی هر یک از پارامترهای نامعلوم به‌دست می‌آید.

[1] Gaussian Markov Random Fields (GMRF)

۳ تحلیل اکتشافی داده ها ی جرم وجنایت کانادا

منطقه آنتاریو در ایالت واترلو شامل شهرهای واترلو، کیچینر، کمبریج و ۴ منطقه روستایی می‌باشد. جمعیت این منطقه در سال ۲۰۱۱، ۵۰۶۱۰۷ نفر بوده که در ۷۵۵ ناحیه‌ی سرشماری توزیع شده است. در کل ۲۹۰۲۷ تماس به پلیس برای گزارش جرم و جنایت در طول ۹۰۶۰ واحد فضاـ زمان ثبت شده است. این داده‌ها برای سال ۲۰۱۱ در ۱۲ دوره ۲ ساعته طبقه‌بندی شده‌اند (لوان و همکاران ، ۲۰۱۶). برای تحلیل فضایی جرائم ابتدا توزیع نقطه ای جرائم در محدوده مورد مطالعه را نمایش می دهیم سپس با استفاده از مدل های گرافیکی و آماری الگوی فضایی این جرائم در منطقه مورد مطالعه استخراج کرده و مناطق جرم خیز (لکه های داغ) را شناسایی می‌کنیم. با استفاده از مدل‌های گرافیکی شامل آزمون



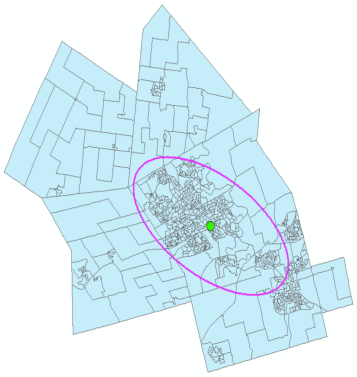
شکل ۱ : توزیع نقطه ای جرائم

میانگین مرکزی و توزیع جهت دار (بیضی انحراف استاندارد)، توزیع فضایی و مرکز ثقل جرائم را ارزیابی می‌کنیم. میانگین مرکزی مشابه میانگین معمولی در آمار است و به همان صورت نیز محاسبه می‌شود. این تحلیل، مرکز ثقل مجموعه ای از عوارض را شناسایی می کند، بیضی انحراف استاندارد این امکان را فراهم می‌کند تا جهت توزیع عوارض را به طور آماری و دقیق شناسایی کنیم.

در شکل ۲ میانگین مرکزی و بیضی انحراف استاندارد جرائم در محدوده مورد مطالعه نشان داده شده است. با توجه به نقشه بدست آمده الگوی جهت عوارض و مرکز ثقل جرائم بدین شرح است: مرکز ثقل جرائم تقریبا بر مرکز محدوده مورد مطالعه منطبق است با توجه به این امر احتمال وقوع جرم در محله های نزدیک به مرکز محدوده مورد مطالعه بیشتر است همچنین بیضی انحراف استاندارد در امتداد جنوب شرقی و شمال غربی کشیده شده است، علت این کشیدگی جرائم ارتکابی در محله های جنوب شرقی است.

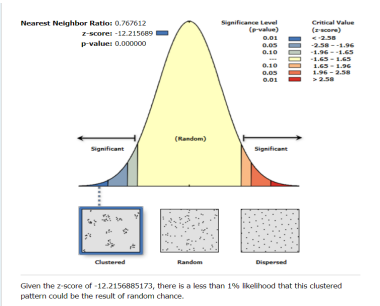
۱.۳ آزمون خوشه بندی

کشف روندهای موجود در داده های فضایی و شناخت الگوها در آمار فضایی از اهمیت ویژه‌ای برخوردار است، زیرا در تشخیص چگونگی توزیع داده‌ها در فضا و همچنین شناخت الگوی پیروی شده توسط توزیع این داده‌ها به ما کمک می‌کند. ازجمله مهمترین آزمون‌ها در شناسایی الگوهای فضایی آزمون نزدیکترین همسایگی است با انجام این آزمون می‌توان پی برد



شکل ۲: میانگین مرکزی و بیضی انحراف استاندارد

که آیا توزیع داده ها خوشه ای است یا پراکنده. در این آزمون اگر نسبت میانگین نزدیکترین همسایگی کمتر از یک باشد داده های مورد مطالعه دارای الگوی خوشه ای و اگر شاخص محاسبه شده بزرگتر از یک باشد داده ها دارای الگوی پراکنده هستند (عسگری ، ۱۳۹۰). نتیجه ی آزمون به صورت عددی و گرافیکی است که هر دو نتیجه در شکل ۳ و جدول ۱.۳ آورده شده است.



شکل ۳: آزمون نزدیک ترین همسایگی

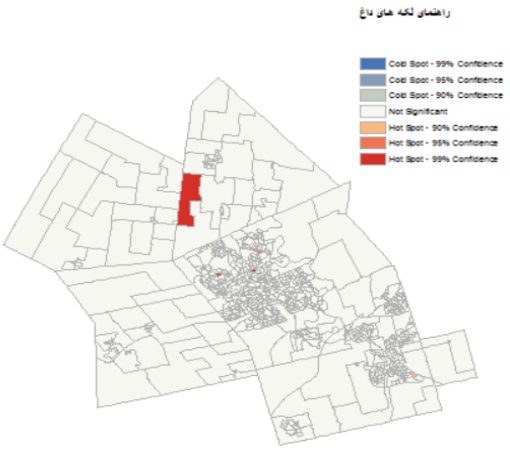
جدول ۱ : آزمون نزدیک ترین همسایگی

میانگین فاصله مشاهده شده	۵۷۴/۶۲۲۸
مینگین فاصله مورد انتظار	۷۴۸/۵۸۴۹
نسبت نزدیک ترین همسایه	۰/۷۶۷۶۱۲
امتیاز استاندارد	−۱۲/۲۱۵۶۸۹
$p - value$	۰

نتایج بدست آمده این فرضیه را که داده های جرم به طور تصادفی در فضا پراکنده شده اند، رد می‌کند. نسبت نزدیک ترین همسایه ی به دست آمده نشان می‌دهد که داده ها به صورت خوشه ای در فضا پراکنده شده اند . مقدار بزرگ $z - score$ و مقدار کوچک $p - value$ از نظر آماری این فرضیه را تایید می‌کند.

۲.۳ شناسایی لکه های داغ و سرد

برای شناسایی خوشه های عوارض با مقادیر زیاد (لکه های داغ) و خوشه های عوارض با مقادیر کم (لکه های سرد) از شاخص $Getis-OrdGi$ (گتیس و اورد ، ۱۹۹۲) استفاده می‌کنیم. این شاخص به برنامه ریزان و تحلیگران کمک می‌کند تا مکان‌هایی را که تعداد زیادی جرم در آن‌ها رخ می‌دهد و به اصلاح مکان های جرم خیز نامیده می‌شوند شناسایی کنند تا با تخصیص درست منابع، سریعتر و موثرتر نسبت به پیشگیری از جرم و کشف آن اقدام کنند. نقشه لکه های داغ در شکل ۴ آورده شده است با توجه به نقشه لکه های داغ محدوده های قرمز محدوده هایی هستند که ۹۹ درصد احتمال دارد درآن



شکل ۴: شناسایی نقاط داغ

مناطق جرم به وقوع بپیوندد محدوده های نارنجی مناطقی هستند که با احتمال ۹۵ درصد در آن ها تعداد زیادی جرم متمرکز شده است.

۴ مدل بندی تماس برای خدمات پلیس

در این بخش برای مدل‌سازی فضایی_زمانی بیزی از مدل سلسله مراتبی سه مرحله‌ای و رهیافت INLA استفاده می‌شود. در مرحله اول چون تعداد تماس‌های جرم و جنایت کم می‌باشد لذا تعداد تماس‌ها برای خدمات پلیس در منطقه $i = ۱,\cdots,۷۵۵$ و دوره زمانی $t = ۱,\cdots,۱۲$، دارای توزیع پواسن به صورت

$$y_{it}|\mu_{it} \sim POI(\mu_{it}) \quad i = ۱,\cdots,۷۵۵, \quad t = ۱,\cdots,۱۲ \tag{۱.۴}$$

فرض شده است. میانگین توزیع تعداد تماس‌ها را به صورت

$$\log(\mu_{it}) = \log(E_{it}) + \alpha + u_i + s_i + \gamma_t + \phi_t + \psi_{it} \tag{۲.۴}$$

در نظر بگیرید که در آن E_{it} تعداد مورد انتظار تماس‌ها بوده و ریسک نسبی به ریسک کلی یا اثر ثابت (α)، اثرات تصادفی فضایی $(u_i + s_i)$، اثرات تصادفی زمانی $(\gamma_t + \phi_t)$ و اثر تصادفی فضایی_ زمانی (ψ_{it}) تجزیه شده است.

در مرحله دوم فرض کنید

$$\begin{aligned} u_i &\sim Normal(۰.\sigma_u^۲), \quad i = ۱, \cdots, ۷۵۵ \\ s_i &\sim ICAR(W, \sigma_S^۲) \\ \gamma_t &\sim Normal(۰, \sigma_\gamma^۲), \quad t = ۱, \cdots, ۱۲ \\ \phi_t &\sim ICAR(P, \sigma_\phi^۲) \end{aligned} \tag{۳.۴}$$

که در آن $\sigma_u^۲$، $\sigma_S^۲$، $\sigma_\gamma^۲$ و $\sigma_\phi^۲$ ابر پارامترها می‌باشند. همانگونه که ملاحظه می‌گردد اثرات تصادفی فضایی و اثرات تصادفی زمانی به دو مؤلفه بدون ساختار و ساختاریافته تجزیه شده‌اند که مدل‌های ساختاریافته شامل مدل $ICAR$ می‌باشند که به ترتیب همبستگی فضایی و زمانی این اثرات را مدل بندی می‌کنند. $ICAR$ یک توزیع پیشینی مرسوم در آمار فضایی برای پارامترهای اثرات تصادفی می‌باشد که در آن میانگین مورد انتظار s_i برابر میانگین همسایگی آن می‌باشد.

برای مرحله سوم تحلیل بیزی سلسله مراتبی توزیع پیشینی تمام ابر پارامترها $Gamma(۰/۵, ۰/۰۰۰۵)$ در نظر می‌گیریم. برای تحلیل بیزی سلسله مراتبی از روش INLA استفاده می‌کنیم. چهار مدل متفاوت برای تحلیل بیزی داده‌ها در نظر گرفته شده است: ۱) ψ_{it}ها مستقل و هم توزیع هستند.

۲) ψ_{it}ها دارای همبستگی فضایی و ناهمبستگی زمانی هستند.

۳) ψ_{it}ها دارای ناهمبستگی فضایی و همبستگی زمانی هستند.

۴) ψ_{it}ها دارای همبستگی فضایی و زمانی هستند.

۱.۴ انتخاب مدل

ورودی‌های مدل‌سازی در نرم‌افزار R با روش INLA با ۴ مدل فوق انجام شده است. تابع INLA و نقشه‌ها با نرم‌افزار GIS تهیه شده‌اند. DIC برای انتخاب مدل بهتر از معیار DIC استفاده می‌شود. DIC معیار بیزی برای نیکویی برازش و تعیین پیچیدگی مدل است. (اشپیگل و همکاران، ۲۰۰۳) پیشنهاد دادند که امید ریاضی پسین آماره انحراف $\bar{D}$ و تعداد پارامترهای مؤثر در مدل p_D اطلاعات مدل را خلاصه می‌کنند. بنابراین معیار اطلاع انحراف DIC به صورت $DIC = \bar{D} + p_D$ است: مدلی که کمترین DIC را نسبت به بقیه مدل‌ها داشته باشد مدل بهتری است. مقادیر DIC و تعداد پارامترهای مؤثر برای ۴ مدل فضایی–زمانی و تعامل فضا–زمان در جدول ۱ آورده شده است. طبق جدول ۱.۴ مدل دوم با کمترین DIC بهترین مدل انتخاب شده است.

جدول ۲: مقایسه DIC مدل‌ها

تعامل فضا–زمان	پارامترهای تعامل	اطلاع انحراف	پارامتر موثر
مدل اول	γ_t و u_i	۵۴۲۷۱/۳۶	۵۰۵۴
مدل دوم	ϕ_t و u_i	۵۳۶۸۱/۴۰	۴۰۷۲
مدل سوم	γ_t و s_i	۵۴۷۷۱/۰۷	۳۹۶۰
مدل چهارم	ϕ_t و s_i	۵۴۷۷۶/۷۱	۳۶۸۰

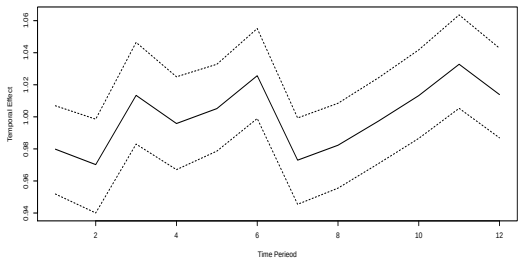
۲.۴ برآورد پارامترهای مدل

برآورد پارامترهای مدل دوم با روش INLA در جدول ۳ آورده شده است. نمودار ۵ نشان‌دهنده‌ی اثرات تصادفی زمانی

جدول ۳: برآورد پارامترهای مدل دوم

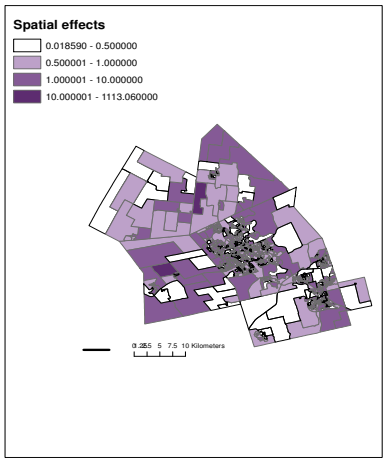
پارامترهای مدل	میانگین	انحراف پسینی	چارک اول	میانه	چارک سوم	مد
σ_u^2	۰/۹۱۹	۰/۴۸۹	۰/۸۲۷	۰/۹۱۷	۱/۰۱۹	۰/۹۱۴
σ_γ^2	۱۷۸۹/۲۹۰	۹۸۷/۱۵۱	۵۴۴/۹۷۶	۱۵۷۴/۷۰۸	۴۲۸۹/۶۶۲	۱۲۰۲/۱۶۲
σ_ψ^2	۱۳/۰۳۴	۰/۳۵۵	۱۲/۳۱۵	۱۳/۰۲۸	۱۳/۷۱۳	۱۳/۰۵۴

است. براساس این نمودار بیشترین تأثیرات اصلی زمانی در دوره زمانی بین ۵۹ : ۱۱ – ۰۰ : ۱۰ و ۵۹ : ۲۱ – ۰۰ : ۲۰ است. بنابراین تخصیص منابع انسانی پلیس، باید در این دوره زمانی بیشتر باشد.



شکل ۵: اثرات تصادفی زمانی همراه با فاصله باورمندی ۹۵ درصد

شکل ۶ نشان‌دهنده‌ی اثرات تصادفی فضایی است. مناطقی با اثرات فضایی بالا، به‌طور مداوم تعداد زیادی تماس خدمات دارند و مناطقی هستند که منابع پلیس باید برای تمام دوره‌های زمانی به کار گرفته شود. در حالت کلی اثرات اصلی فضایی در مراکز منطقه مورد مطالعه بیشترین و در مناطق اطراف روستایی به‌ویژه در شمال غربی کمترین میزان است.



شکل ۶: اثرات تصادفی فضایی

بحث و نتیجه‌گیری

در این مقاله، در یک مطالعه موردی به تحلیل فضایی_زمانی سلسله مراتبی بیزی داده‌های جرم و جنایت کشور کانادا با رهیافت INLA پرداخته شد. که وابستگی فضایی_زمانی را لحاظ کرده و موجب انعطاف پذیرتر بودن مدل برای تشخیص الگوهای غیر معمول می گردد. سپس به تحلیل اکتشافی داده ها پرداخته می شود که می تواند به تشخیص نواحی با رفتار غیر معمول در طول زمان منجر شود. چهار مدل مختلف به داده‌ها برازش و بهترین مدل با استفاده از معیار DIC انتخاب گردید.

مراجع

عسگری، ع.، (۱۳۹۰)، *تحلیل های آمار فضایی و با $ArcGIS$*، چاپ اول، انتشارات سازمان فناوری اطلاعات و ارتباطات شهرداری تهران، تهران.

Luan, H., Quick, M., & Law, J. (2016). Analyzing Local Spatio-tEmporal Patterns of Police Calls-for-Service Using Bayesian Integrated Nested Laplace Approximation. ISPRS *International Journal of Geo-Information*, **5**(9), .162

Lindgren, F., Rue, H., & Lindström, J. (2011). An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: the Stochastic Partial Differential Equation Approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(4), 423-498.

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian Computing with INLA: New Features, *Computational Statistics & Data Analysis*, **67**, .68-83

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian Inference for l Latent Gaussian Models by Using Integrated Nested Laplace Approximations, *Journal of the royal statistical society: Series b (statistical methodology)*, **71**(2), .319-392

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian Computing with INLA: a Review, *Annual Review of Statistics and Its Application*, **4**, 395-421.

Spiegelhalter, D., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2003). Bayesian Measures of Model Complexity and Fit, *Quality Control and Applied Statistics*, **48**(4), 431-432.

Getis, A., & Ord, J. K. (2010). The Analysis of Spatial Association By use of Distance Statistics, *In Perspectives on Spatial Data Analysis*, pp. 127-145, Springer, Berlin, Heidelberg.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

تحلیل فضایی کیفیت آب‌های زیرزمینی شهرستان لنجان در اصفهان

مبینا علی‌بابایی[۱]، **نصراله ایران‌پناه**
گروه آمار، دانشگاه اصفهان

چکیده: در مطالعات محیطی، داده‌ها معمولاً از نظر فضایی وابسته هستند. تعیین ساختار همبستگی فضایی داده‌ها و پیش‌بینی دو مسئله مهم در تحلیل آماری داده‌های فضایی است. برای تعیین ساختار فضایی داده‌ها، یک مدل تغییرنگار پارامتری اغلب به تغییرنگار تجربی داده‌ها برازش داده می‌شود. سپس این مدل‌ها با استفاده از پیشگویی فضایی کریگیدن بر روی داده‌ها اعمال می‌شوند. هدف از این مقاله، تحلیل فضایی هدایت الکتریکی، نیترات و کربن آلی کل به منظور تهیه نقشه پراکندگی آلودگی و پیش‌بینی منطقه مورد بررسی در شهرستان لنجان استان اصفهان است. نتایج این مطالعه نقشه‌های پراکندگی و پیش‌بینی آلودگی را با استفاده و مقایسه کریگیدن و هم‌کریگیدن برای کیفیت آب‌های زیرزمینی این منطقه نشان می‌دهد.

واژه‌های کلیدی:، هدایت الکتریکی، نیترات، کربن آلی کل، کریگیدن، هم‌کریگیدن.
کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲H۱۱، ۶۲M۴۰ .

۱ مقدمه

امروزه کیفیت منابع آبی شامل آب‌های سطحی و زیرزمینی از اهمیت بالایی برخوردار است. از طرف دیگر، با توجه به اثرات سرطان‌زای برخی مواد شیمیایی کشاورزی، بخشی از این مواد شیمیایی ممکن است به آب‌های زیرزمینی نفوذ کنند. از جمله پارامترهای کیفیت آب آشامیدنی و کشاورزی، نیترات ($NO۳$) به‌عنوان ماده معدنی، کربن آلی کل (TOC) به‌عنوان شاخص ترکیبات آلی و هدایت الکتریکی (EC) به‌عنوان شاخص کمیت ناخالصی است. آلودگی در آب‌های سطحی و زیرزمینی از مدفوع تجزیه شده انسان و حیوان، محصولات صنعتی مانند کودهای نیتروژن‌دار و رواناب‌های کشاورزی منشاء می‌گیرد. استفاده سالانه از کودهای نیتروژن‌دار و سایر شیوه‌های مدیریت محصول، منبع قابل توجهی از نیترات‌ها را فراهم می‌کند که ممکن است در مناطقی با خاک‌های در معرض خطر و هیدروژئولوژی به آب‌های زیرزمینی نفوذ کند. نیترات شایع‌ترین آلاینده شیمیایی است که برای سلامت انسان در سفره‌های آب زیرزمینی جهان خطرناک است (دی و همکاران ، ۲۰۰۲). اندازه‌گیری TOC برای بهره‌برداری از تصفیه خانه‌های آب و فاضلاب و سنجش آلودگی آب‌های زیرزمینی اهمیت

[۱] نام و ایمیل ارائه دهنده مقاله: 93Alibabaeimobina@gmail.com

زیادی دارد. به منظور جلوگیری از آلودگی منابع ارزشمند آب شیرین، بسیاری از کشورها و شهرها ارزیابی آسیب‌پذیری آب‌های زیرزمینی را به‌عنوان بخشی از برنامه‌ریزی توسعه شهری خود انجام می‌دهند (سینسرو ، ۲۰۰۰). اندازه‌گیری EC توانایی آب برای هدایت جریان الکتریکی است. EC آب‌ها با ورود نمک‌های مختلف بسته به نوع و مقدار آن‌ها به منابع آب افزایش می‌یابد (قاسیم و همکاران ، ۲۰۰۲).

پیش‌بینی میزان عناصرمتفاوت در آب‌های زیرزمینی به‌طور قابل ملاحظه‌ای مورد توجه قرار گرفته است. این پیش‌بینی‌ها با توجه به تاثیری که عناصرمختلف درکیفیت آب چاه‌ها دارند، در تشخیص موقعیت مناسب برای حفر چاه کاربرد دارند. در تحلیل فضایی داده‌ها قبل از انجام هرگونه تحلیل باید ازطریق تحلیل اکتشافی داده‌های فضایی، ماهیت اولیه مشاهدات از نظر شناسایی فرض‌های اولیه، وجود داده‌های پرت، مانایی، همسانگردی و وجود روند مورد بررسی قرار گیرد. همچنین پیش‌بینی بر اساس روش‌های کریگیدن و هم‌کریگیدن بر روی منطقه مورد مطالعه انجام می‌شود (محمدزاده ، ۱۳۹۸).

شهرستان لنجان با ۱۰۹۳ کیلومتر مربع وسعت یکی از شهرهایی است که در ۳۵ کیلومتری شهر اصفهان و در جنوب غربی استان اصفهان و دشت رودخانه زاینده رود واقع شده است. زمین مورد استفاده در این منطقه عمدتاً کشاورزی و نزدیک به صنایعی مانند نساجی اکریلیک، چغندرقند و کاغذ و مقوا است. از این رو برای بررسی کیفیت آب چاه‌های این منطقه از ۲۵ حلقه چاه آب نمونه‌برداری و متغیرهای مورد نظر اندازه‌گیری شدند. امین و همکاران (۲۰۰۹) تحلیل فضایی سه متغیر ژئوشیمی را در آب‌های زیرزمینی اصفهان انجام دادند.

در بخش ۲ مقدمه‌ای بر تحلیل فضایی ارائه می‌شود. تحلیل اکتشافی داده‌های مورد نظر چاه‌های آب در بخش ۳ ارائه می‌شود. در بخش ۴ دو پیشگویی فضایی کریگیدن و هم‌کریگیدن ارائه و دقت آن‌ها مقایسه می‌شوند. در نهایت در بخش ۵ بحث و نتیجه‌گیری ارائه می‌شود. تحلیل‌های انجام شده در این مقاله با استفاده از نرم‌افزار GS+ است.

۲ تحلیل فضایی

وقتی شرط اساسی استقلال داده‌ها برقرار نباشد و وابستگی بین مشاهدات برحسب مکان مشاهدات باشد از شاخه‌ای از آمار تحت عنوان آمار فضایی استفاده می‌شود. داده‌های فضایی علاوه بر مقادیر مربوط به متغیرهای مورد بررسی، اطلاعات مربوط به مکان مشاهدات را نیز شامل می‌شوند. برای تحلیل داده‌های فضایی لازم است یک مدل آماری در نظر گرفته شود. در آمار فضایی معمولا یک میدان تصادفی به عنوان مدل آماری داده‌های فضایی در نظر گرفته می‌شود. میدان تصادفی مجموعه‌ای از متغیرهای تصادفی مانند $\{Z(s) : s \in D\}$ است به‌طوری که $d \geq ۱, D \subseteq R^d$ می‌باشد (کرسی ، ۱۹۹۳).

در آمار فضایی مفاهیمی شبیه به واریانس و کوواریانس را می‌توان تعریف کرد که بیانگر ساختار همبستگی داده‌ها باشند. متوسط اختلاف بین مقادیر میدان تصادفی را می‌توان معیار خوبی برای بیان این وابستگی در نظر گرفت. واریانس اختلاف بین مقادیر میدان تصادفی در دو موقعیت s و $s+h$ تغییرنگار نامیده می‌شود و به‌صورت $۲\gamma(h) = Var[Z(s+h) - Z(s)]$ تعریف می‌شود. معمولا تحلیل داده‌های فضایی از روی اطلاعات نمونه بسیار دشوار است، اما گاهی فرض‌هایی مانند انواع مانایی موجب ساده‌سازی مسئله می‌شوند، مهم‌ترین آن‌ها مانای ذاتی است. هرگاه میانگین میدان تصادفی ثابت یا مستقل از s باشد یعنی $\mu = E(Z(s)); s \in D$ همچنین واریانس عبارت $(Z(s_۱) - Z(s_۲))$ فقط تابعی از فاصله موقعیت‌های $s_۱$ و $s_۲$ باشد، یعنی $Var[Z(s_۱) - Z(s_۲)] = ۲\gamma(s_۱ - s_۲)$، آنگاه میدان تصادفی مانای ذاتی است. در حالتی که میدان تصادفی $Z(۰)$، مانای ذاتی باشد، یک برآوردگر برای تغییرنگار بر اساس روش گشتاوری توسط ماترون (۱۹۶۳) به‌صورت

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

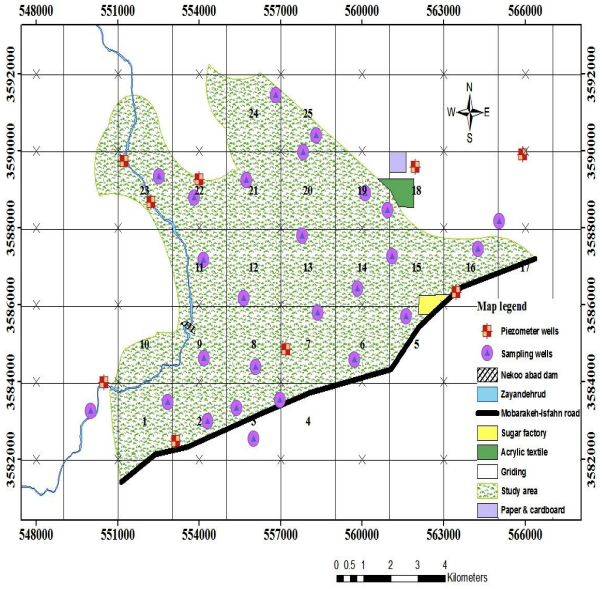
است، که در آن $N(h)$ مجموعه تمام زوج موقعیت‌هایی است که در فاصله h از یکدیگر قرار دارند.

جورنل و هویج برگتس (۱۹۸۷) مدل‌های پارامتری مختلفی را برای تغییرنگار معرفی کردند. به‌طور معمول مدل‌های تغییرنگار دارای سه پارامتر دامنه، ازاره و اثر قطعه‌ای می‌باشد. بازه‌ای که در خارج از آن، تغییرنگار به حالت افقی درآمده و ثابت می‌ماند دامنه (a) نامیده می‌شود. به‌طور معمول تغییرنگار تابعی صعودی است و ممکن است به کران بالا منتهی شود، چنین کرانی ازاره (c) نام دارد. اثر قطعه‌ای (c_0) مقدار تغییرنگار در مبدأ مختصات است. برای انتخاب بهترین مدل تغییرنگار بر اساس تحلیل اکتشافی داده‌های فضایی به ایران‌پناه و همکاران (۲۰۰۹) مراجعه شود. یکی از مدل‌های پرکاربرد تغییرنگار پارامتری مدل کروی است که تابع آن به‌صورت زیر تعریف می‌شود:

$$۲\gamma(h)=\begin{cases} ۰ & ||h||=۰ \\ c_0+c(\frac{۳}{۲}\frac{||h||}{a}-\frac{۱}{۲}(\frac{||h||}{a})^۳) & ۰<||h||\le a \\ c_0+c & ||h||\ge a \end{cases}$$

۳ تحلیل اکتشافی داده‌ها

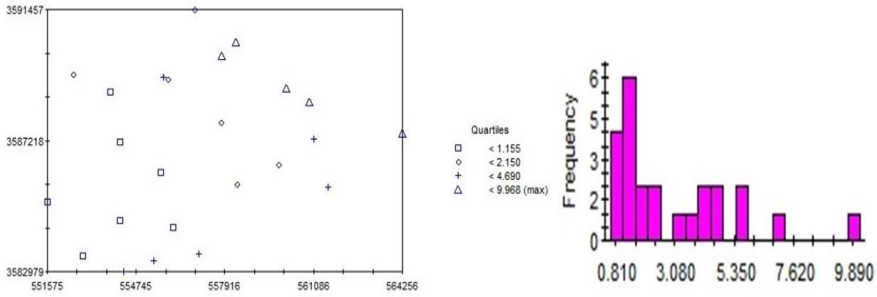
در تحلیل فضایی داده‌ها قبل از انجام هرگونه تحلیل باید ازطریق تحلیل اکتشافی داده‌های فضایی ماهیت اولیه مشاهدات از نظر شناسایی فرض‌های اولیه، وجود داده‌های پرت، مانایی، همسانگردی و وجود روند مورد بررسی قرار گیرد.



شکل ۱: موقعیت چاه‌های نمونه‌برداری در منطقه شهرستان لنجان

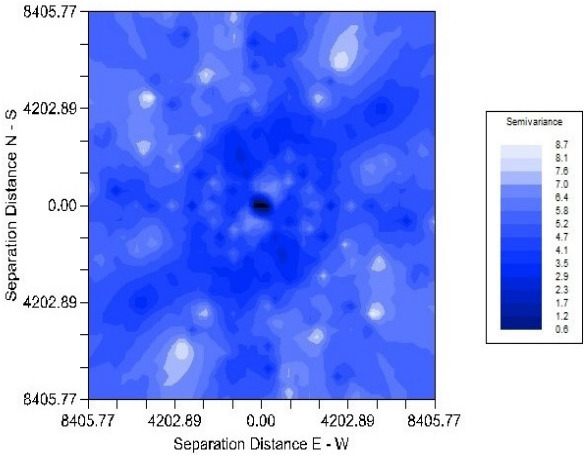
در شکل ۱ می‌توان نقشه جغرافیایی منطقه مورد مطالعه را مشاهده کرد که نشان دهنده موقعیت چاه‌های آب می‌باشد. نمونه‌های به‌دست آمده از چاه‌های آب در آزمایشگاه‌های بهداشت محیط طبق روش‌های استاندارد مورد بررسی قرار گرفتند. متغیرهای مورد بررسی در این نمونه‌ها عبارت‌اند از EC (هدایت جریان الکتریکی) $NO۳$ (نیترات که یک نوع ماده معدنی می‌باشد) TOC (مقدار اتم‌های کربنی است که در ترکیبات آلی در یک نمونه آب وجود دارد). در ادامه به تحلیل و بررسی متغیر EC که در میان این سه متغیر دارای اهمیت بالاتر است، می‌پردازیم.

در شکل ۲ نمودار بافت‌نگار مشاهدات مربوط به متغیر EC رسم شده است. همچنین نمودار پراکندگی مشاهدات رسم شده که موقعیت مشاهدات را با توجه به ۲۵حلقه چاه آب موجود در منطقه مورد مطالعه نشان می‌دهد. توزیع داده‌ها بر



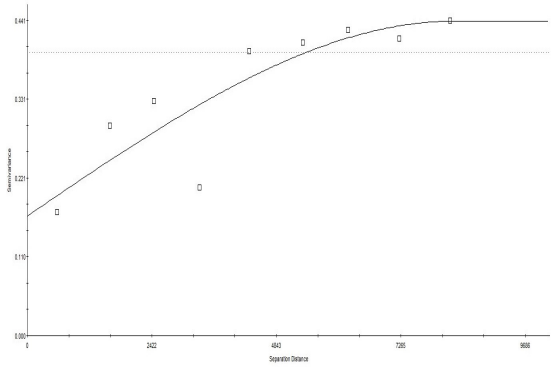
شکل ۲: نمودار بافت‌نگار و پراکنش موقعیت جغرافیایی مشاهدات متغیر EC

اساس بافت‌نگار مشاهدات غیرنرمال چوله به راست است. همچنین وجود روند از جهت غرب به شرق برحسب موقعیت داده‌ها در نمودار پراکنش مشاهده می‌گردد. برای این داده‌ها با استفاده از روش اصلاح میانه روند در داده‌ها حذف گردید.



شکل ۳: رویه تغییرنگار متغیر EC

شکل ۳ رویه تغییرنگار در متغیر EC را نشان می‌دهد. با توجه به نمودار رویه متقارن است و نشان می‌دهد داده‌های جدید فاقد روند هستند و مانایی در میانگین برقرار است، همچنین تغییرنگار همسانگرد می‌باشد.



شکل ۴: تغییرنگار تجربی و برآورد مدل پارامتری کروی متغیر EC

در شکل ۴ تغییرنگار تجربی و برازش تغییرنگار پارامتری کروی متغیر EC ارائه شده است. باتوجه به این شکل وجود همبستگی فضایی در داده‌ها کاملا مشهود است.

جدول ۱: انتخاب مدل پارامتری کروی تغییرنگار و برآورد پارامترهای آن

مدل	اثر قطعه‌ای (c_0)	ازاره (c)	دامنه (a)	R^2	RSS
کروی	۰/۱۶۷۵۰	۰/۴۴۰۰۰	۸۳۲۰/۰۰	۰/۷۴۶	۰/۰۲۰۳
نمایی	۰/۱۸۹۰۰	۰/۹۸۰۲۰	۱۹۳۸۰/۰۰	۰/۷۲۸	۰/۰۲۱۷
خطی	۰/۲۰۰۲۰۱	۰/۴۶۸۶۲	۸۲۲۳/۳۴	۰/۷۱۴	۰/۲۶۱۰
خطی با ازاره	۰/۲۰۴۰۰	۰/۸۲۲۰۰	۱۹۲۷۰/۰۰	۰/۷۱۴	۰/۰۲۲۸
گاوسی	۰/۲۰۸۴۰	۰/۴۴۵۸۰	۴۲۷۰/۰۰	۰/۷۳۱	۰/۰۲۱۵

در جدول ۱ برآورد پارامترهای اثر قطعه‌ای (c_0)، ازاره (c) و دامنه (a) همچنین معیارهای R^2 و میانگین توان دوم باقیمانده‌ها (RSS) برای ۵ مدل تغییرنگار پارامتری کروی، نمایی، خطی، خطی با ازاره و گاوسی ارائه شده است. با توجه به مقادیر دو معیار R^2 و RSS بهترین مدل تغییرنگار پارامتری کروی در بین ۵ مدل تغییرنگار مورد استفاده قرار می‌گیرد.

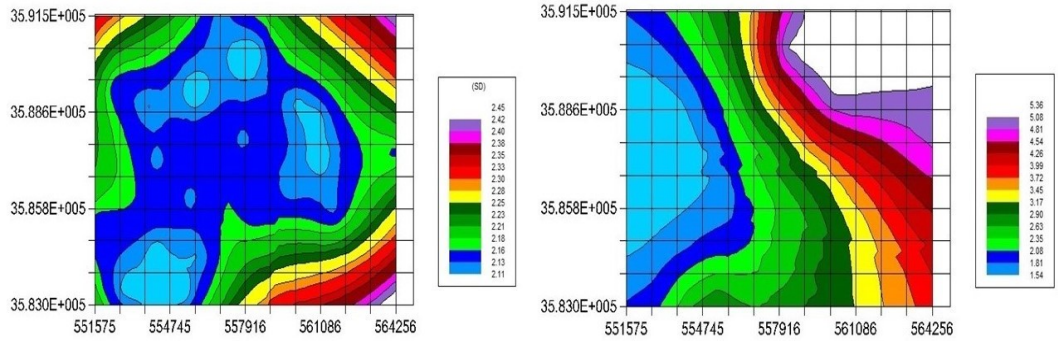
$$2\gamma(h) = \begin{cases} 0 & ||h|| = 0 \\ 0/167 + 0/273(\frac{3}{2}\frac{||h||}{8320} - \frac{1}{2}(\frac{||h||}{8320})^3) & 0 < ||h|| \leq 8320 \\ 0/167 + 0/273 & ||h|| \geq 8320 \end{cases}$$

۴ پیشگویی فضایی کریگیدن و هم‌کریگیدن داده‌ها

در این بخش با استفاده از روش‌های پیشگویی فضایی کریگیدن برای متغیر EC و روش هم‌کریگیدن و به کارگیری متغیر کمکی $NO3$ مقدار متغیرEC در نقاط مختلف محاسبه و دقت آن‌ها مورد مقایسه قرار میگیرند.

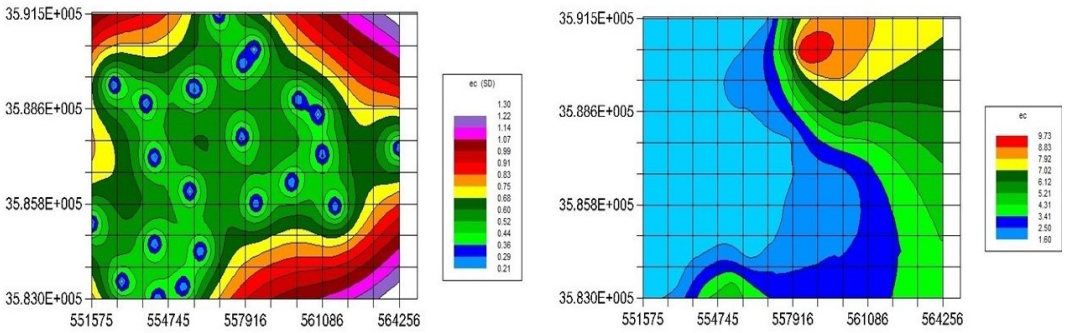
در تحلیل فضایی بهترین پیشگوی خطی نااریب برای $Z(s_0)$ کریگیدن نام دارد. کریگیدن روش پیشگویی است که وقتی هدف پیشگویی $Z(s_0)$ براساس مشاهدات $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))^T$ باشد، بهترین پیشگوی خطی نااریب را به‌صورت $\hat{Z}(s_0) = \lambda^T \mathbf{Z}\mathbf{Z}$ با واریانس کریگیدن $\sigma_k^2(s_0) = E[\hat{Z}(s_0) - Z(s_0)]^2 = \gamma^T\Gamma^{-1}\gamma - (\mathbf{1}^T\Gamma^{-1}\gamma - 1)^2(\mathbf{1}^T\Gamma^{-1}\mathbf{1})^{-1}$ ارائه می‌دهد، که در آن $\gamma = (\gamma(s_0 - s_1), \ldots, \gamma(s_0 - s_n))^T$ و Γ یک ماتریس $n \times n$ که مولفه (i, j) ام آن $\gamma(s_i - s_j)$ است. گاهی در هر موقعیت فضایی $\lambda^T = \gamma + \mathbf{1}((\mathbf{1} - \mathbf{1}^T\Gamma^{-1}\gamma)(\mathbf{1}^T\Gamma^{-1}\mathbf{1})^{-1})^T\Gamma^{-1}$، چند خصیصه مرتبط با هم اندازه‌گیری می‌شوند و تحلیل چندمتغیره داده‌های فضایی مطلوب نظر است. به‌علاوه در مواردی که از یک میدان تصادفی مشاهدات کمی در اختیار باشد و نتوان به اندازه کافی مشاهده به‌دست آورد، ممکن است پیشگویی آن از دقت مناسب برخوردار نباشد. در این‌گونه موارد با در نظر گرفتن ارتباط متغیر مورد نظر و متغیرهای دیگری که از آن به اندازه کافی مشاهده قابل دسترس باشد، می‌توان پیشگویی را به نحوی بهینه و با دقت بیش‌تر انجام داد. این روش پیشگویی، هم‌کریگیدن نامیده می‌شود.

در شکل ۵ نمودار پیشگویی کریگیدن و رویه انحراف معیار کریگیدن متغیر EC نشان داده شده است. همان‌طور که در نمودار پیشگویی کریگیدن مشاهده می‌گردد طیف رنگی بیانگر افزایش مقدار متغیر EC با تغییر رنگ از آبی به بنفش است و نشان می‌دهد مقدار هدایت الکتریکی EC در چاه‌های آب لنجان از جهت غرب به شرق افزایش می‌یابد. علت این افزایش وجود زمین‌های کشاورزی و باغداری در غرب شهرستان لنجان و همچنین وجود کارخانجات صنعتی و صنایع کوچک در



شکل ۵: نمودار پیشگویی کریگیدن و رویه انحراف معیار کریگیدن متغیر EC

غرب این شهرستان می‌باشد. همچنین رویه انحراف معیار کریگیدن برای اندازه‌گیری دقت پیشگویی فضایی کریگیدن نشان داده شده است. در نواحی آبی رنگ که تراکم مشاهدات بیش‌تر است به دلیل انحراف معیار کم‌تر دقت پیشگویی بالاتر است.



شکل ۶: نمودار پیشگویی هم‌کریگیدن و رویه انحراف معیار هم‌کریگیدن متغیر EC

در شکل ۶ نمودار پیشگویی هم‌کریگیدن و رویه انحراف معیار هم‌کریگیدن متغیر EC با استفاده از متغیر کمکی $NO۳$ نشان داده شده است. همان‌طور که در نمودار پیشگویی هم‌کریگیدن مشاهده می‌شود طیف رنگی بیانگر افزایش مقدار متغیر EC با تغییر رنگ از آبی به قرمز است که همان نتیجه مربوط به پیشگویی فضایی کریگیدن به‌دست می‌آید. همچنین رویه انحراف معیار هم‌کریگیدن برای اندازه‌گیری دقت پیشگویی فضایی هم‌کریگیدن نشان داده شده است. در نواحی آبی رنگ که تراکم مشاهدات بیش‌تر است به دلیل انحراف معیار کم‌تر دقت پیشگویی بالاتر است.

جدول ۲: مقایسه دو روش پیشگویی کریگیدن و هم‌کریگیدن برای متغیر EC

Est-SD(cokrig)	Est-SD(krig)	Z-Est(cokrig)	Z-Est(krig)	Y-Coordinate	X-Coordinate
۰/۶۶	۱/۰۷	۲/۰۰	۲/۴۱	۵۵۷۷۳۹	۳۵۸۴۵۶۹
۰/۶۴	۱/۰۶	۱/۹۳	۲/۳۸	۵۵۷۷۳۹	۳۵۸۴۷۴۵
۰/۶۱	۱/۰۴	۱/۸۶	۲/۴۰	۵۵۷۷۳۹	۳۵۸۴۹۲۲
۰/۵۸	۱/۰۳	۱/۸۲	۲/۳۸	۵۵۷۷۳۹	۳۵۸۵۰۹۹
۰/۵۵	۱/۰۲	۱/۷۹	۲/۳۷	۵۵۷۷۳۹	۳۵۸۵۲۷۵

در جدول ۲ با استفاده از ۵ موقعیت با مختصات مورد نظر، پیشگویی فضایی مقدار EC و همچنین انحراف معیار آن در دو روش کریگیدن و هم‌کریگیدن محاسبه شده است. همانگونه که در جدول بالا ملاحظه می‌شود زمانی که از پیشگویی هم‌کریگیدن استفاده شده است، انحراف معیار کم‌تر از حالتی است که پیشگویی کریگیدن مورد استفاده قرار گرفته است. بنا به رابطه عکسی که بین انحراف معیار پیشگویی و دقت پیشگویی وجود دارد، پیشگویی هم‌کریگیدن دقیق‌تر از پیشگویی کریگیدن می‌باشد. می‌توان نتیجه گرفت برای پیش‌بینی دقیق‌تر مقدار متغیرها در مختصات‌های مشخص بهتر است از پیشگویی هم‌کریگیدن استفاده کنیم زیرا عملکرد بهتری نسبت به روش کریگیدن دارد.

۵ بحث و نتیجه‌گیری

در این مقاله با استفاده از داده‌های مربوط به متغیر EC برای ۲۵ حلقه چاه در منطقه لنجان تحلیل فضایی داده‌ها انجام شد. در ابتدا تحلیل اکتشافی داده‌ها شامل نمودار موقعیت داده‌ها، بافت‌نگار متغیر EC جهت تعیین توزیع داده‌ها و مشاهدات پرت و همچنین رویه تغییرنگار برای بررسی مانایی و همسانگردی داده‌ها انجام شد. در ادامه، با استفاده از الگوریتم اصلاح میانه روند در داده‌ها حذف شد. سپس با استفاده از تغییرنگار تجربی، مدل تغییرنگار پارامتری کروی به‌همراه پارامترهای آن برآورد شد. در انتها رویه‌های پیشگویی کریگیدن و هم‌کریگیدن بر اساس متغیر کمکی $NO۳$ به‌همراه رویه‌های خطای معیار آن‌ها رسم گردید. با مشاهده رویه‌های خطای معیار نشان داده شد، پیشگویی فضایی هم‌کریگیدن از انحراف معیار کم‌تر در نتیجه دقت بالاتری برخوردار است. مقایسه دو روش پیشگویی کریگیدن و هم‌کریگیدن بر اساس پیشگویی فضایی برای ۵ موقعیت هم به‌طور مشابه انجام شد که دقت بالاتر پیشگویی فضایی هم‌کریگیدن را به علت استفاده از متغیر کمکی $NO۳$ نشان می‌دهد.

مراجع

محمدزاده، م.، (۱۳۹۸)، *آمار فضایی و کاربردهای آن*، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،

Amin, M. M, Ebrahimi, A., Hajian, M., Iranpanah, N. and Bina, B. (2010), Spatial Analysis of Three Agrichemicals in Groundwater of Isfahan Using GS+, *Iranian Journal of Environmental Health Science & Engineering***7**, 71-80.

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

Di, H. J., Cameron, K. C., Hendry, T., Moore, S. and Smith, N. P. (2002), Nitrate Leaching and Pasture Production from Different Nitrogen Sources on a Shallow Stoney Soil Under Flood-Irrigated Dairy Pasture, *Australian Journal of Soil Research*, **40**, 317–334.

Iranpanah, N., Mohammadzadeh, M., Vahidi Asl, M. Q. and Yassaghi, A. (2009), Spatial Data Analysis of Finite Strain Data Across a Trust Sheet Using R. *Computer and Geosciences*, **35**, 626-634.

Journel, A. G. and Huijbregts, C. J. (1978), *Mining Geostatistics*, Academic Press, Londen.

Matheron, G. (1963). Principles of Geostatistics, *Economic, Geology*, **58**, 1246-1266.

Qasim, S. R., Motley, E. M. and Zhu, G. (2000), *Water Works Engineering: Planning, Design, and Operation*, Prentice Hall.

Sincero, A. P. (2002), *Physical-Chemical Treatment of Water and Wastewater*, CRC.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

رگرسیون بتای گسسته برای تحلیل داده‌های رتبه‌بندی فضایی

سیده فریناز عمرانی[۱]**، محسن محمدزاده**
گروه آمار، دانشگاه تربیت مدرس

چکیده: چگونگی رفتار با داده‌های رتبه‌بندی حاصل از پاسخ به نظر سنجی‌ها، سالیان زیادی است که توجه محققین را به خود جلب کرده است. متخصصان نیز همیشه به دنبال یک تفسیر کمی و عددی از متغییرهای پاسخ و تاثیر برآورد پارامترها و پیشگویی‌کننده‌ها روی میانگین متغیر پاسخ هستند. گرچه رگرسیون خطی گزینه‌ای مناسب است، اما مفروضات ضمنی آن، یعنی پاسخ‌های نامتناهی ناهمبسته، خطی بودن، همسانی واریانس‌ها، نرمال بودن توزیع داده‌ها، وقتی برای داده‌های رتبه‌بندی همبسته فضایی استفاده می‌شود، غیر واقع‌گرایانه هستند. بنابراین در این مقاله مدل رگرسیون بتای گسسته فضایی معرفی و تعمیم آن به مدل رگرسیون بتای گسسته متورم به منظور لحاظ کردن تأثیر همبستگی متغیرها در مدل ارائه می‌شود. آنگاه با رهیافت بیزی و استفاده از نمونه‌گیری زنجیر مارکوفی مونت کارلو، برآورد پارامترهای مدل و مجموعه‌های باور به دست آورده می‌شوند. سپس نحوه کاربست این مدل در تحلیل داده‌های تداخل درد بیماران نشان داده خواهد شد.

واژه‌های کلیدی: رگرسیون بتای گسسته، داده‌های ترتیبی فضایی، زنجیره مارکوفی مونت کارلو.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62G08، 62J05.

۱ مقدمه

برای تحلیل داده‌های نظرسنجی تعیین نوع داده‌ها تصمیم کلیدی است. در صورتی که داده‌ها بدون نظم باشند، به عنوان داده‌های اسمی در نظر گرفته می‌شوند که مدل‌های لوجیت، چندجمله‌ای و پروبیت برای تحلیل آنها مناسبند (حسن و همکاران، ۲۰۱۶). در مواردی که ترتیب وجود داشته باشد اما تفسیر عددی واضح نباشد، داده‌ها ترتیبی تلقی می‌شوند و مدل‌های مناسب برای تحلیل اینگونه داده‌ها می‌تواند لوجیت ترتیبی و رگرسیون پروبیت باشند (گودریچ و همکاران، ۲۰۱۸). نوع دیگر داده‌ها مشابه داده های ترتیبی هستند، اما تعداد سطوح آنها زیاد است. برای داده‌های رتبه‌بندی، بحث‌های متنوعی در مورد نحوه تفسیر آنها وجود دارد. اگر آنها عددی در نظر گرفته شوند، تفسیر آنها آسان‌تر است اما تعداد سطوح کم آنها ممکن است منجر به نتایجی نامنظم شود (لیدل و کروشکه، ۲۰۱۸). هنگامی که تعداد سطوح زیاد باشد، استفاده از روش‌های

[۱] نام و ایمیل ارائه دهنده مقاله: سیده فریناز عمرانی، farinaz_omrani@modares.ac.ir

عددی با محدودیت‌هایی روبرو می‌شود. در حالت کلی برای تحلیل داده‌های رتبه‌بندی رویکردهای متنوعی، از جمله رگرسیون بتای گسسته که ترکیبی از رگرسیون خطی و رگرسیون ترتیبی وجود دارند، که هر یک مزایا و معایب خود را دارند. رگرسیون بتای گسسته نسخه‌ای تطبیقی از رگرسیون بتا است که از مدل‌های احتمالاتی ترتیبی الهام می‌گیرد و یک متغیر پنهان پیوسته را به متغیر پاسخ گسسته نگاشت می‌دهد (زایلس و همکاران، ۲۰۱۰). با این تفاوت که توزیع بتا را به عنوان توزیع پایه استفاده می‌کند و نقطه برش را به وسیله مقادیر پیشگویی شده بر اساس مشاهدات تعیین می‌کند. این تغییرات باعث کاهش خطر بیش‌برازش نسبت به رگرسیون ترتیبی می‌شود و امکان تفسیر عددی دقیق‌تری از داده‌های رتبه‌بندی را فراهم می‌کند. رگرسیون بتای گسسته شبیه مدل رگرسیون بتای دودویی است (ناجیرا و همکاران، ۲۰۱۸)، با این تفاوت که اولاً، از مفهوم متغیر پنهان در رگرسیون ترتیبی برای توصیف فرآیند انتخاب پاسخ توسط شرکت‌کنندگان نظرسنجی استفاده می‌کند. دوماً، از نرم‌افزار برای محاسبه تورم شمارش مقادیر خارج از محدوده معقول پاسخ استفاده می‌کند با استفاده از عبارات توزیع تجمعی در لگاریتم تابع درستنمایی. این روش بهتر به تطابق با واقعیت فرآیند انتخاب پاسخ‌ها در نظرسنجی کمک می‌کند. رگرسیون بتای گسسته و رگرسیون بتای گسسته متورم نیز مدل‌های مشابهی هستند، اما تفاوت‌های مهمی دارند: اولاً، رگرسیون بتای گسسته از نقاط پایانی منعطف برای برش توزیع بتا استفاده می‌کند، در حالی که رگرسیون بتای گسسته متورم از تابع جرم احتمال در فضای گسسته استفاده می‌کند. دوماً، رگرسیون بتای گسسته تنها روی متغیرهای وابسته رگرسیون را اعمال می‌کند، در حالی که رگرسیون بتای گسسته متورم تأثیر متغیرهای همبسته را نیز مدل می‌کند. سوماً، رگرسیون بتای گسسته از روش نمونه‌برداری گیبز استفاده می‌کند (؟)، در حالی که رگرسیون بتای گسسته متورم از نمونه‌برداری متروپلیس–هستینگز استفاده می‌کند.

۲ مدل رگرسیون بتای گسسته فضایی

تابع چگالی احتمال توزیع بتا با پارمترهای شکل $\alpha, \beta > ۰$ به صورت

$$f(y;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-۱}(۱-y)^{\beta-۱}, \quad ۰ < y < ۱, \tag{۱.۲}$$

است. برای مدل رگرسیون بتا زایلس و همکاران (۲۰۱۰) توزیع بتا را به صورت

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱} \tag{۲.۲}$$

بازپارامتری کردند، به طوریکه پارامترهای میانگین و دقت آن به ترتیب به صورت $\mu = \frac{\alpha}{\alpha+\beta}$ و $\phi = \alpha + \beta$ هستند. این مقادیر بیانگر ناهمگنی مدل است، یعنی با ثابت نگه داشتن پارامتر دقت ϕ، همانطور که میانگین به حدود صفر یا یک نزدیک می‌شود، پراکندگی پاسخ‌ها نیز کاهش می‌یابد و متمایل به فشرده شدن در نزدیکی صفر و یک، یعنی کمینه و بیشینه محدوده متغیر پاسخ می‌شود. با فرض اینکه میانگین توزیع از طریق یک تابع پیوند، ترکیبی خطی از متغیرهای تبیینی است، می‌توان مدل رگرسیون بتا را به صورت $g(\mu) = \mathbf{x}^\top\beta$ بیان کرد، که در آن $\mathbf{x}$ بردار متغیرهای تبیینی، $g(\cdot)$ تابع پیوندی است که بازهٔ $(۰, ۱)$ را به خط واقعی نگاشت می‌کند، به عنوان مثال $g(u) = \log(u/(۱-u))$ تابع پیوند لوجیت است. از آنجا که در بسیاری از مطالعات کاربردی با داده‌های فضایی مواجه می‌شویم که بر حسب موقعیت مکانیشان به یکدیگر وابسته‌اند، در این‌صورت مدل‌بندی پارامتر میانگین μ را با فرض ثابت بودن پارامتر دقت ϕ می‌توان به صورت مدل رگرسیون فضایی $g(\mu) = \mathbf{x}^\top\beta + \eta W y$ در نظر گرفت، که در آن η بردار ضرایب رگرسیونی اثر تصادفی فضایی، W ماتریس وزن فضایی یا مجاورت و $y = (y_۱, \ldots, y_n)$ بردار متغیر پاسخ در موقعیت‌های فضایی یا نواحی $s_۱, \ldots, s_n$ است.

در اکثر رویکردهای مبتنی بر درستنمایی برای برآورد پارامترهای مدل، باید لگاریتم تابع درستنمایی ماکسیمم شود که در رگرسیون بتای گسسته شامل تابع چگالی توزیع بتای (۱.۲) است. قرار دادن $y = ۰$ یا ۱ باعث می‌شود این تابع چگالی

به صفر نزدیک شود و در نتیجه لگاریتم آن به بی‌نهایت میل کند. به همین دلیل، برای y فقط بازهٔ باز $(۰,۱)$ درنظر گرفته می‌شود. بنابراین، قدم اول برای سازگاری رگرسیون بتای گسسته، تبدیل داده‌های خام به محدوده $(۰,۱)$ است. فرض کنید K مقدار یکتا برای متغیر پاسخ y وجود دارد که به صورت صعودی y_k مرتب شده‌اند. یک تبدیل ساده می‌تواند به صورت $z_k = \frac{y_k - y_۱}{y_K - y_۱}$، باشد. اما ممکن است این تبدیل به جای بازهٔ $(۰,۱)$، به بازهٔ $[۰,۱]$ نگاشت شود. به جای آن بافرهای چپ و راست به صورت $b_\ell \equiv (y_۲ - y_۱)/۲$ و $b_r \equiv (y_K - y_{K-۱})/۲$، در نظر گرفته می‌شوند. در واقع محدودهٔ پنهان داده‌ها به سمت چپ تا $y_۱ - b_\ell$ و به سمت راست تا $y_k - b_r$ گسترش داده شده است. این منجر به تبدیل خطی بازنگری شده $z = u(y) \equiv \frac{y-\delta}{r}$ می‌شود، که در آن $y \in \{y_۱, \ldots, y_K\}$ و داریم $\delta \equiv y_۱ - b_\ell$ و $r \equiv y_K - y_۱ + b_\ell + b_r$. در واقع تبدیل فوق داده‌ها را به محدودهٔ $b_\ell / (y_K - y_۱ + b_\ell + b_r) \leq u(y) \leq (y_K - y_۱ + b_\ell) / (y_K - y_۱ + b_\ell + b_r)$ نگاشت می‌کند. برای پیش‌گویی، باید یک تبدیل معکوس اعمال شود تا از انحراف‌های توزیع بتا به محدوده مشاهدات نگاشت شود، سپس یک مرحله گسسته‌سازی انجام شود، به طوری که مشاهده y_k نزدیکترین مقدار به نمونه تولید شده از توزیع بتا با میانگین و پراکندگی داده شده گزارش شود. این تبدیل به صورت $y = u_s^{-۱}(z) \equiv y_k$ تعریف می‌شود. به طوری که

$$|r\hat{x} + \delta - y_k| \leq |r\hat{x} + \delta - y_{k'}|, \quad k' \in \{۱, \ldots, K\}$$

پیش‌گویی نقطه‌ای متغیر پاسخ از محاسبه میانگین نمونه‌های بزرگ تولید شده به روش بالا، حاصل می‌شود.

فرایند گسسته‌سازی باید در تابع احتمال برای برازش مدل لحاظ شود. به عبارت دیگر، اگر مقدار z_k مشاهدهٔ قبل از گسسته‌سازی باشد، نمی‌توان اطمینان داشت که نمونه تولید شده از توزیع بتا همان مقدار z_k است، بلکه تنها می‌دانیم برای هر $۱ < k < K$ مقداری بین $\frac{z_{k-۱}+z_k}{۲}$ و $\frac{z_k+z_{k+۱}}{۲}$ است، که با توابع مرزی به صورت

$$z_\ell(y_k) = \begin{cases} ۰ & k = ۱ \\ \frac{u(y_{k-۱})+u(y_k)}{۲} & ۱ < k \leq K \end{cases}, \quad z_r(y_k) = \begin{cases} \frac{u(y_k)+u(y_{k+۱})}{۲} & ۱ \leq k < K \\ ۱ & k = K \end{cases}$$

خلاصه می‌شود. بنابراین در تابع درستنمایی مشارکت یک نقطه داده با پاسخ y_k به صورت زیر است،

$$P(Y = y_k) = F(z_r(y_k)) - F(z_\ell(y_k)).$$

۱.۲ برآورد بیزی مدل

مدل‌بندی رگرسیون بتای گسسته با رهیافت بیزی و استفاده از نمونه‌برداری زنجیره مارکوفی مونت کارلو (MCMC) برای برآورد تابع چگالی پسین، مزایایی نسبت به روش ماکسیمم درستنمایی دارد. اولاً، نیاز به استدلال‌های آسیب‌پذیر در برابر اطلاعات کم حجم را ندارد و مجموعه‌های باورمندی برای پارامترها و متغیر پاسخ را ارائه می‌دهد. دوماً، این روش امکان رهایی از ماکسیمم موضعی و محاسبه ماکسیمم عام تابع درستنمایی را فراهم می‌سازد، که در مقایسه با الگوریتم‌های مورد استفاده در روش‌های ماکسیمم درستنمایی بهتر عمل می‌کند. سوماً، رهیافت بیزی به کاربران این امکان را می‌دهد که اعتقادات پیشین خود را در مورد مقادیر پارامترها را نیز در محاسبات خود لحاظ کنند. احتمال شرطی پاسخ‌های مشاهده‌شده به صورت

$$p(\mathbf{y} \mid \mathbf{x}; \phi, \mathbf{fi}, b_\ell, b_r) = \prod_{n=۱}^{N} [F(z_r(y_{k[n]}; b_\ell, b_r); g^{-۱}(\mathbf{f}^\top \mathbf{x_n}), \phi) - F(z_\ell(y_{k[n]}; b_\ell, b_r); g^{-۱}(\mathbf{fi}^\top \mathbf{x_n}), \phi)]$$

است، که در آن $z_r(y; b_\ell, b_r)$ و $z_\ell(y; b_\ell, b_r)$ توابعی هستند که هر پاسخ مشاهده‌شده را به دامنه توزیع بتا نگاشت می‌کنند. همچنین $g^{-1}(\mathbf{fi}^\top \mathbf{x})$ تابعی است که میانگین توزیع بتا را با تشکیل پیش‌گوی خطی $\mathbf{fi}^\top \mathbf{x}$ محاسبه کرده و با تابع لوجستیک آن را اجرا می‌کند. بنابراین، تابع لگاریتم توزیع پسین به صورت

$$\begin{aligned}\ell(\phi, \mathbf{fi}, b_\ell, b_r) &= \log[F(z_r(y_{k[n]}; b_\ell, b_r); g^{-1}(\mathbf{fi}^\top \mathbf{x_n}), \phi) - F(z_\ell(y_{k[n]}; b_\ell, b_r); g^{-1}(\mathbf{f}^\top \mathbf{x_n}), \phi)] \\ &+ \Phi(\phi) + \mathbf{B}(\mathbf{fi}) + B_\ell(b_\ell) + B_r(b_r)\end{aligned}$$

به دست می‌آید، که در آن $\Phi(\phi)$، $\mathbf{B}(\mathbf{fi})$، $B_\ell(b_\ell)$ و $B_r(b_r)$ توابع لگاریتمی پیشین مشخص‌شده برای پارامتر دقت توزیع بتا (ϕ)، ضرایب پارامترهای میانگین (β) و بافرهای چپ و راست (b_ℓ, br) هستند.

۳ تحلیل داده‌های تداخل درد

این داده‌ها بر اساس نظرسنجی انجام‌شده در مراکز بهداشتی در ۴۲ ناحیه جغرافیایی انگلستان، مطابق شکل ۱ ـ الف، بر روی تقریباً ۱۰۰۰۰ بیمار در بازه زمانی ۲۰۱۰ تا ۲۰۱۴ جمع‌آوری شده است، تا کیفیت مراقبت‌های دریافتی توسط آن‌ها در هر ناحیه ارزیابی شود. پس از فیلتر کردن و تلفیق داده‌ها، ۱۳۱۸ مشاهده از سه متغیر تبیینی شدت (درد)، مداخله (درد) و سن بیمار جمع‌آوری شده است.

شدت : میانگین ۴ پاسخ، هر کدام در مقیاس ۰ تا ۱۰ (۱۱ سطح) است. این پاسخ‌ها ادراک درد بیماران را در طی مدت ۷ روز قبل از نظرسنجی اندازه‌گیری می‌کنند به صورتی که شدیدترین درد (۱)، کم‌ترین درد (۲)، درد متوسط (۳) و همین الان (۴) را اندازه‌گیری می‌کند.

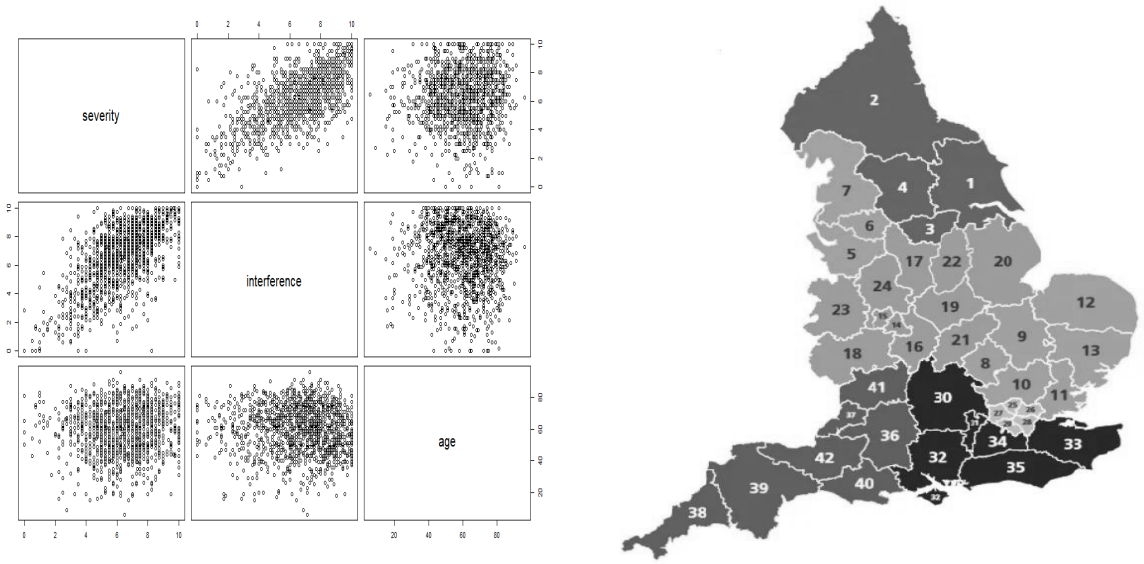
تداخل : میانگین ۷ مقدار، هر کدام در مقیاس ۰ تا ۱۰ (۱۱ سطح) است. این سوالات در طی ۷ روز قبل از نظرسنجی، میزان تداخل در زندگی بیمار به وسیله درد را در ابعاد زیر اندازه‌گیری می‌کنند: (۱) فعالیت‌های عمومی، (۲) خلق و خو، (۳) توانایی پیاده‌روی، (۴) کارهای عادی (خارج از خانه و کارهای خانگی)، (۵) روابط با دیگران، (۶) خواب و (۷) لذت از زندگی.

سن : سن پاسخ‌دهندگان، به سال است.

نمودارهای پراکندگی دو به دوی داده‌ها در شکل ۱ ـ ب و آزمون همبستگی اسپیرمن بیانگر وجود همبستگی مثبت بین مقادیر شدت درد و مداخله درد است. اما تأثیر سن بر مداخله درد کم‌تر واضح است. بعلاوه آزمون اسپیرمن نشان‌دهنده یک همبستگی منفی معنی‌دار بین سن و مداخله درد است.

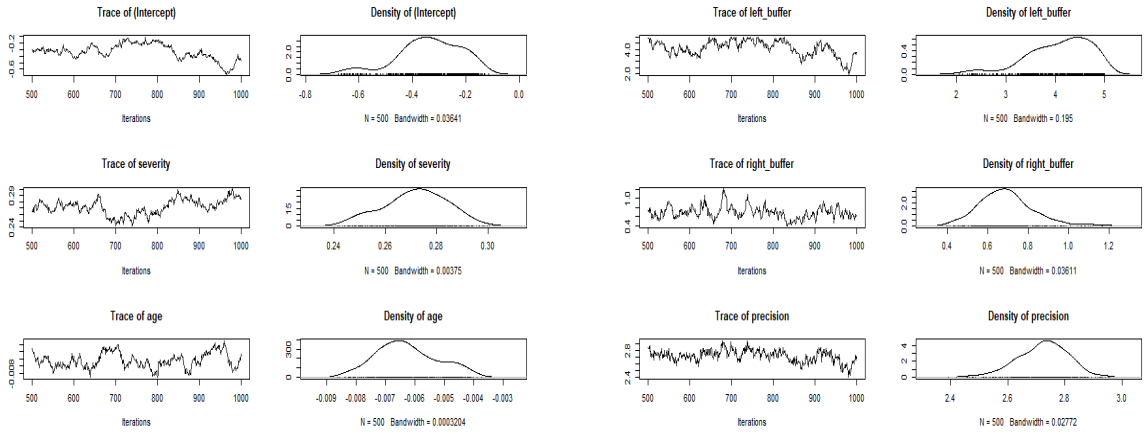
آموزش مدل : تداخل درد (متغیر پاسخ)، بر حسب شدت درد و سن پاسخ‌دهندگان پیش‌گویی می‌شود و ساده ترین فرمول مدل شامل داده‌های آموزشی به صورت $interference = severity + age$ است. در اینجا فرض شده است که مقادیر منحصربفرد مورد انتظار متغیر پاسخ برابر با مقادیر منحصربفرد مشاهده‌شده متغیر پاسخ در داده است. در این مثال، انتظار می‌رود که مقادیر متغیر پاسخ در دامنه ۰ تا ۱۰ با افزایش‌های $\frac{1}{7}$ باشد. با بررسی‌های انجام شده مقادیر $\frac{2}{7}$ و $\frac{3}{7}$ در داده‌های آموزشی رخ نداده‌اند و ممکن است الگوریتم ارائه شده برای محاسبه نقاط برش اشتباه کند که برای اصلاح آن مقادیر منحصربفرد موردانتظار متغییر پاسخ را تغییر می‌دهیم و فرآیند نمونه برداری MCMC را برای برآورد بیزی پارامتر های مدل مورد استفاده قرار می‌دهیم.

ابزارهای تشخیصی در MCMC : با استفاده از نمودارهای روند و تابع خودهمبستگی پس از ۱۰۰۰۰ تکرار، ۵۰۰۰ داغیدن و فاصله ۵، حجم نمونه برابر ۱۰۰۰ در نظر گرفته شده است.



(ب) (الف)

شکل ۱ : الف– نقشه نواحی جغرافیایی انگلستان، ب– نمودار پراکندگی دودویی داده‌های تداخل درد.



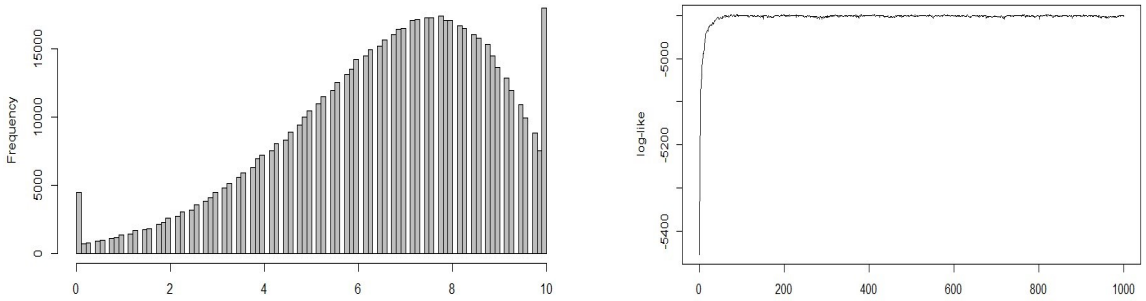
(ب) (الف)

شکل ۲ : نمودارهای روند و چگالی بعد دوره تنظیم و ازافزایش نمونه

باانجام آزمون همگرایی گوک (گوک، ۱۹۹۲)، میتوان ارزیابی کرد که آیا مدل‌های بیزی به درستی به داده‌های مشاهده شده همگرا می‌شوند یا نه.

این آزمون به ارزیابی اینکه آیا بخش‌های ابتدایی و انتهایی زنجیره MCMC میانگین‌های مشابهی دارند یا نه، کمک می‌کند.امتیاز Z یک اندازه‌گیری از تعداد انحراف معیاری است که تفاوت میانگین‌های بخش‌های ابتدایی و انتهایی زنجیره MCMC از صفر فاصله دارد. امتیاز Z نزدیک به صفر (حدود ۲– تا ۲) به معنای همگرایی خوب است و نشان می‌دهد که میانگین‌ها مشابه هستند و تفاوت معناداری وجود ندارد.

برآورد مدل : برای دیدن پراکندگی کامل پیش‌گویی‌ها،از پیشگویی نمونه‌ای استفاده می‌کنیم.نمودار هیستوگرام پیش‌گویی‌های نمونه‌ای به نمودار داده‌های مشاهده شده شباهت بیشتری دارد. با استفاده از آزمون کولموگروف–اسمیرنوف مشاهده می‌شود



(ب) (الف)

شکل ۳: الف– نمودار لگاریتم درستنمایی بعد از دوره داغیدن و افزایش نمونه و ب– هیستوگرام پیش‌گویی تداخل درد

جدول ۱: مقادیر شاخص آزمون همگرایی گوک

سن	شدت	عرض از مبداء	دقت	بافر راست	بافر چپ
−۱/۳۸۶۳	−۰/۹۰۶۹	۱/۵۴۹۸	−۱/۵۸۰	−۴/۶۰۹۴	−۰/۱۱۹۸

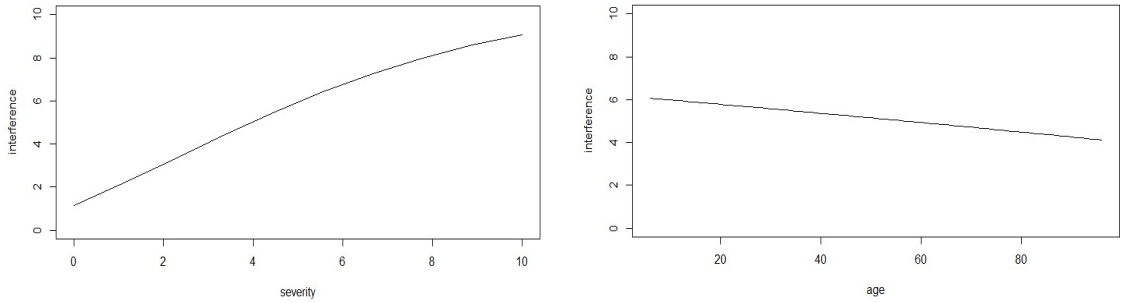
که فرضیه صفر که می‌گوید پیش‌گوی پسین و داده‌های آموزشی دارای توزیع احتمالی یکسان هستند، با p مقدار ۰/۰۸۳۷۲ و اطمینان ۹۵٪ رد نمی‌شود. در جدول ۲ مقادیر چندک‌ها ۰/۰۲۵, ۰/۵, ۰/۹۷۵ در نظر گرفته شده است، که مقدار میانه و همچنین مجموعه باور ۹۵٪ متقارن برای هر ضریب، دقت و پارامترهای بافر چپ و راست را ارائه می‌کند.

جدول ۲: جدول برآورد ضرایب

	بافر چپ	بافر راست	دقت	عرض از مبداء	شدت	سن
۵٪	۲/۸۱۶۸۴۰	۰/۴۸۹۳۲۸۷	۲/۵۵۵۹۴۵	−۰/۵۹۰۵۵۱۱	۰/۲۴۹۹۷۰۵	−۰/۰۰۷۹۰۱۴۳۰
۵۰٪	۴/۱۹۰۴۷۴	۰/۶۸۵۷۰۶۰	۲/۷۲۶۵۲۰	−۰/۳۲۸۲۱۵۸	۰/۲۷۲۷۲۰۳	−۰/۰۰۶۴۰۱۶۹۹
۹۵٪	۴/۹۰۸۶۹۳	۰/۹۳۶۴۸۰۷	۲/۸۴۹۶۲۹	−۰/۱۶۴۵۵۹۰	۰/۲۹۰۳۶۳۷	−۰/۰۰۴۳۹۸۸۹۸

بررسی مجموعه‌های باورمندی نشان می‌دهد که آیا ضرایب با تفسیر بیزی معنی‌دار هستند یا خیر. در این مثال، ملاحظه می‌شود ضرایب هر دو متغیر شدت و سن با اطمینان ۹۵٪ معنی‌دار هستند. [1]اطلاعات فراهم شده، مفید است، اما ضرایب در مدل DBR به دلیل غیرخطی‌بودن ناشی از توزیع بتا و فرآیند گسسته‌سازی، تفسیر معنی‌داری ندارند. به عبارت دیگر، هر تغییر یک واحدی در یک متغیر تبیینی، تغییری در میانگین پاسخ ایجاد نمی‌کند.با تولید شبه ضرایب مدل، نمودارهای رابطه بین یک متغیر تبیینی و پاسخ میانگین، به عنوان یک تابع از مقدار متغیر تبیینی و مشروط به مجموعه ثابتی از مقادیر برای سایر متغیرهای تبیینی در مدل نشان می‌دهند. نمودارهای شکل ۴ نشان می‌دهند که وابستگی میانگین پیش‌گویی شده تداخل درد با متغیر سن رابطه خطی قوی دارد، در حالی که متغیر شدت درد یک رابطه به طور قابل ملاحظه‌ای غیرخطی با میانگین پیش‌گویی شده تداخل درد دارد. این موضوع منطقی است زیرا تأثیر سن بر میانگین پاسخ به محدوده‌ای کوچکتر از متغیر شدت محدود است.

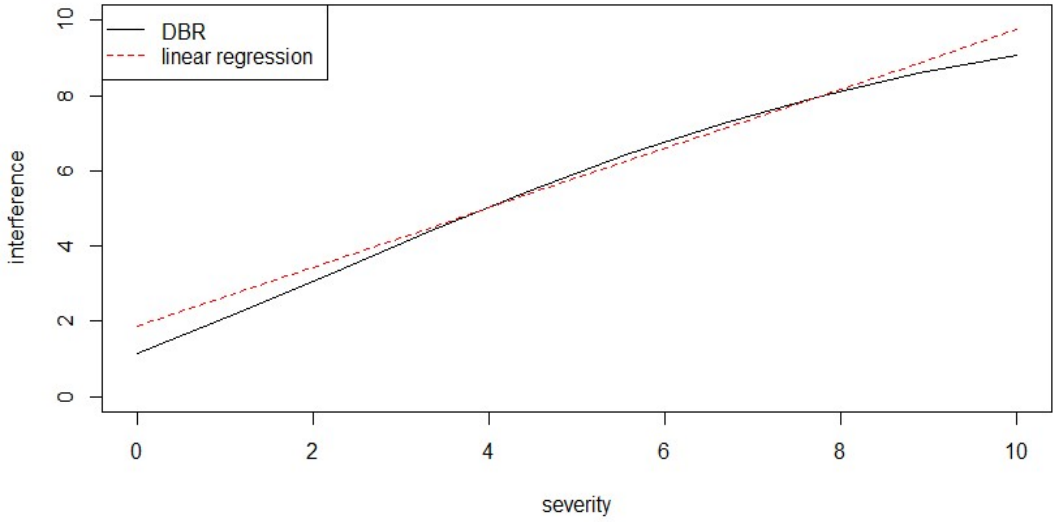
پیش‌گویی‌ها با مدل DBR و رگرسیون خطی در شکل ۵ نشان داده شده است. همانطور که انتظار می‌رفت، وابستگی میانگین مقدار پیش‌گویی شده تداخل به مقدار شدت در رگرسیون خطی به صورت یک خط راست است، در حالی که مدل DBR یک رابطه غیرخطی پیش‌گویی می‌کند. به خصوص، وقتی که مقدار شدت به حداکثر مقدار خود یعنی ۱۰ نزدیک



(ب) (الف)

شکل ۴: نمودار میانگین پیشگویی در مقابل متغیرهای الف_ شدت و ب_ سن

می‌شود، یک شیب کاهنده در میانگین ملاحظه می‌شود. همچنین پیش‌گویی تداخل درد کمتری برای شدت درد با مقدار صفر داریم، نسبت به پیش‌گویی از مدل رگرسیون خطی، که نزدیک‌تر به نتیجه مطلوب پیش‌گویی تداخل صفر برای شدت صفر (عدم درد) است.



شکل ۵: مقایسه رگرسیون بتای گسسته و رگرسیون خطی برای داده‌های تداخل درد

بحث و نتیجه‌گیری

تحلیل داده‌های رتبه‌بندی وابسته فضایی با استفاده از چارچوب تحلیل رگرسیون بتای گسسته فضایی، امکان تعیین تأثیر متغییرهای تبیینی بر پاسخ را، فارغ از فرض‌های غیرواقعی واریانس ثابت و استقلال خطاها در مدل رگرسیون خطی، فراهم می‌کند. بنابراین داده‌های رتبه‌بندی همبسته فضایی را با معلوم بودن ساختار همبستگی داده‌ها می‌توان با رگرسیون بتای گسسته متورم مدل‌بندی نمود. در این مقاله اجرای مدل رگرسیون بتای گسسته فضایی روی داده‌های نظرسنجی درد مورد بررسی قرار گرفت. رهیافت بیزی، همراه با نمونه‌برداری MCMC برای برآورد بیزی، چندین مزیت دارد. به کاربران اجازه می‌دهد از اطلاعات پیشین برای برآورد پارامترهای رگرسیون بهره ببرند. بعلاوه بازه‌های باورمندی بدون نیاز به فرض‌های غیر واقع‌گرایانه درباره رفتار مجانبی لگاریتم تابع درستنمایی محاسبه می‌شوند. هر چند اجرای MCMC برای مجموعه داده‌های بزرگ یا وقتی نیاز به تعداد زیاد داغیدن و نمونه‌برداری برای برآورد دقیق پارامترهای مدل وجود دارد، ممکن است

خیلی زمان‌بر باشد، اما تکنیک‌های متعددی در منابع پیشنهاد شده‌اند که برای افزایش سرعت آن می‌تواند مورد استفاده قرار گیرد. این مزیت شرایط لازم برای تحلیل فضایی داده‌های رتبه‌بندی را فراهم سازد.

مراجع

Hasan, A., Wang, Z. and Mahani, A. S. (2016), Fast Estimation of Multinomial Logit Models: R Package $mnlogit$, *Journal of Statistical Software*, **75**(3), 1–24. doi:10.18637/jss.v075.i03.

Goodrich, B., Gabry, J., Ali, I. and Brilleman, S. (2020), Rstanarm: Bayesian Applied Regression Modeling via Stan, *R Package Version,*, 2.21.1, https://mc-stan.org/rstanarm.

Liddell, T. M. and Kruschke, J. K. (2018), Analyzing Ordinal Data with metric models: What Could Possibly Go Wrong? *Journal of Experimental Social Psychology*, **79**, 328–348.

Zeileis, A., Cribari-Neto, F., Grun, B. and Kos-midis, I. (2010), Beta Regression in R, *Journal of Statistical Software*, **34**(2), 1–24.

Najera-Zuloaga, J., Lee, D. J. and Arostegui I (2018), Comparison of Beta-Binomial Regression Model Approaches to Analyze Health-Related Quality of Life Data, *Statistical Methods in Medical Research*, **27**(10), 2989–3009.

Geweke, J. (1992), Evaluating the Accuracy of Sampling-based Approaches to the Calculations of Posterior Moments. Bayesian Statistics, **4**, 641–649.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

مدل‌سازی صریح وابستگی فضایی برای تحلیل بیزی داده‌های بقا

علیرضا کبورانی [۱]**، موسی گلعلی‌زاده**
گروه آمار، دانشگاه تربیت مدرس

چکیده:

پیش‌بینی و پیشگویی با کمک مدل‌های یادگیری آماری از پراستفاده‌ترین و مهمترین کاربردهای علم آمار است. مدل‌های آماری پس از آموزش به کمک پردازش داده‌های آموزشی، به اهداف متفاوتی مانند تخصیص مشاهدات جدید به رده‌های موجود در مدل‌های رده‌بندی، نحوه عضویت مشاهدات در خوشه‌ها در مدل‌های خوشه‌بندی و موارد بسیار دیگر استفاده می‌شوند. معمولا این امکان وجود دارد که مدل در پیش‌بینی رده یا خوشه‌ی بعضی از مشاهدات خطا داشته باشد و آنها را به اشتباه به رده یا خوشه دیگری تخصیص دهد. در این شرایط مشخص کردن علت خطای مدل آماری به منظور اصلاح روند کارکرد و ارتقای عملکرد آن بسیار حائز اهمیت است. خطاهای مدل ممکن است به علت برچسب‌گذاری اشتباه بالقوه، هم‌پوشانی رده‌ها، وجود زیررده‌ها و موارد اینچنینی باشد. برای یافتن علت خطاهای مدل و دستیابی به بینش بهتر و عمیق‌تر نسبت به نتایج مدل‌های آماری و عملکرد آن‌ها، می‌توان از ابزار ارزیابی مدل‌های آماری مانند ماتریس‌درهم‌ریختگی استفاده کرد. عمده ابزارهای حاضر، بینشی کلی از عملکرد مدل ارائه می‌کنند و بررسی علت خطاها به صورت موردی فراهم نیست. در تحقیق حاضر، نمودار نقشه رده‌ها به عنوان راهکاری شهودی و بسیار پرکاربرد و راهگشا برای شناسایی و یافتن دلایل خطاهای پیش‌بینی مدل‌های آماری و بررسی موردی خطاها در هر یک از نقاط داده معرفی می‌شود. همچنین، نحوه عملکرد نقشه رده‌ها در تحلیل مجموعه اعداد دست نوشته فارسی تشریح می‌شود.

واژه‌های کلیدی: ارزیابی مدل، دیداری‌سازی، نقشه رده‌ها، رده‌بندی، اعداد دست‌نویس.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H30، 62H09، 62H07، 62H99.

۱ مقدمه

آموزش ماشین‌های محاسباتی قدرتمند با کمک مدل‌های آماری برای استنتاج، تصمیم‌گیری، دسته‌بندی و پیش‌بینی پدیده‌ها را شاید بتوان یکی از مهمترین دستاوردهای بشر پس از ظهور ابزارهای تکنولوژی محور دانست. یادگیری ماشین یا یادگیری آماری، در واقع، یک زمینه تحقیقاتی است که به درک و ساخت روش‌هایی اطلاق می‌شود که «یاد میگیرند». به زبان علمی،

[۱] نام و ایمیل ارائه دهنده مقاله: علیرضا کبورانی، mohsen_m@modares.ac.ir@modares.ac.ir

روش‌هایی که از داده‌ها برای بهبود عملکرد در مجموعه‌هایی از وظایف استفاده می‌کنند را روش‌های یادگیری آماری می‌نامند (میشل، ۱۹۹۷).

پس از اجرای تمامی مدل‌های یادگیری آماری از جمله رده‌بندی، همواره هنگام اجرای مدل نهایی بر روی داده‌های آموزشی یا آزمایشی، ممکن است پس از پیشگویی، یک مشاهده در رده‌ای قرار بگیرد که با برچسب داده شده آن متفاوت باشد. گاهی چنین برچسب‌گذاری اشتباه اریبی‌ـ برچسب نامیده می‌شود و این سوال را ایجاد می‌کند که "آیا مشاهده به اشتباه برچسب‌گذاری شده است یا خیر؟" برای پاسخ به این پرسش و سوالات مشابه دیگر، پژوهشگران به دنبال یافتن ابزاری برای بررسی میزان دقت و ارزیابی مدل‌های برازش شده بوده و هستند (راسکا، ۲۰۱۸).

اساسا ابزارهای ارزیابی به هدف یافتن میزان خطای مدل و همچنین یافتن علت خطاها و سپس بهبود مدل‌ها با رفع دلیل خطاها مورد استفاده قرار می‌گیرند. موضوع ارزیابی مدل‌های آماری سال‌های زیادی مورد توجه محققین بسیاری قرار گرفته است. از میان این تحقیقات بی‌شمار می‌توان فعالیت‌های برایر (۱۹۵۰) و معرفی آماره امتیاز بریر[1]، گوود (۱۹۵۲) و معرفی زیان لگاریتمی[2]، کوهن (۱۹۶۰) و معرفی آماره امتیاز کاپا[3]، استهمن (۱۹۹۷) و ایجاد ماتریس درهم‌ریختگی و نسبت‌های ارزیابی مدل را نام برد.

در کنار روش های ارزیابی مدل با کمک آماره‌ها و ضرایب، ابزارهای متنوعی نیز با رویکرد اکتشافی معرفی شده‌اند. از مهمترین این ابزارها می‌توان به نتایج فعالیت‌های هارتیگان و کلینر (۱۹۸۱) و فرندلی (۱۹۹۴) و شکل‌دهی نسخه مصورسازی شده ماتریس درهم‌ریختگی به عنوان نمودار موزاییکی، راسیوف (۱۹۸۷) و معرفی ضریب و نمودار سایه[4]، بردلی (۱۹۹۷) و معرفی AUC-ROC[5] اشاره کرد. گاهی از چنین ابزارهایی برای مقایسه مدل‌های آموزش‌داده شده برای یافتن مدل بهینه استفاده می‌شود. از تمام ابزارهای نام برده شده تا به اینجا می‌توان برای این هدف استفاده کرد اما بعضی از ابزارها تنها برای مقایسه ایجاد شده‌اند که به عنوان شاهد مثال می‌توان از تحقیقات دورفمن (۱۹۷۹) و معرفی ضریب جینی[6] نام برد.

از نکات مهم دیگر در ارزیابی مدل‌های آماری، بررسی تصادفی تعدادی از مشاهداتی است که در فرایند برازش مدل شرایط خاصی دارند. شاهد مثال‌هایی از این موضوع عبارتند از: مشاهدات دورافتاده و یا خطاهای رخ داده توسط آنها که از نگاه پژوهشگر غیرمنطقی است. برای یافتن این مشاهدات و بررسی نوع و میزان خطاهای هر یک از آن‌ها می‌توان از نمودار کاربردی نقشه رده‌ها[7] (رایمکرز و همکاران، ۲۰۲۱) استفاده کرد. معرفی و روش استفاده از این ابزار، موضوع اصلی تحقیق حاضر بوده و در ادامه به طور خلاصه نحوه ساخت آن تشریح و مثالی از کارکرد آن ارائه می‌شود. برای ارائه مطالب مقاله در بخش دوم مفاهیم اولیه نقشه رده‌ها معرفی می‌شوند. بخش سوم دربرگیرنده اجرای نقشه رده‌ها با کمک نرم افزار R برای داده‌های دست نوشته فارسی است. مقاله با بحث و نتیجه‌گیری تکمیل می‌شود.

---

[1] Brier
[2] Logarithmic loss
[3] Kappa score
[4] Silhouette coefficient and Silhouette plot
[5] Area Under the Curve - Receiver Operating Characteristics
[6] Gini coefficient
[7] Class maps
[8] Probability of Alternative Class (PAC)
[9] Farness

۲ نحوه ساخت نقشه رده‌ها

برای تشکیل نمودار نقشه رده‌ها در محور عمودی میزان احتمال تخصیص به رده‌های جایگزین[۸] و در محور افقی میزان دوری[۹] برای هر مشاهده در رده مورد بررسی رسم می‌شوند. نمودارهای نقشه رده‌ها در واقع به صورت جداگانه برای هر مشاهده محاسبه و سپس به تفکیک رده‌های مورد بررسی در یک نمودار رسم می‌شوند. این اقدام به هدف یافتن اطلاعات از ارتباط نقاط‌داده با یکدیگر و قرارگیری تمام نقاط هر رده در کنار یکدیگر انجام می‌شود.

۱.۲ محاسبه احتمالات عضویت نمونه‌ها

کمیت PAC یا احتمالِ پسینی شرطی رده داده شده g_i در مقایسه با بهترین رده جایگزین به صورت

$$PAC(i) = \frac{\tilde{P}r(i)}{\hat{P}r(i,g_i) + \tilde{P}r(i)} \qquad (۱.۲)$$

تعریف می‌شود که در آن مشاهدات درون مجموعه داده هر کدام با برچسبی مثل i که $i = ۱,\ldots,n$ و هر رده با g که $g = ۱,\ldots,G$ مشخص شده باشند. اکثر روش‌های مرسوم رده‌بندی، به همراه نتایج نهایی، احتمال پسینی تخصیص هر مشاهده i به رده g یعنی $\hat{P}r(i,g)$ را با شرط $\sum_{g=۱}^{G} \hat{P}r(i,g) = ۱$ محاسبه می‌کنند. به عنوان مثال رده‌بندی کننده K ـ نزدیکترین همسایگی[۱۰] (فیکس و هاجس، ۱۹۸۹) این احتمالات را به کمک فراوانی رده‌ها در K همسایگی مشاهده i برآورد می‌کند و یا رده‌بندی کننده تحلیل تشخیصی[۱۱] (مک‌لاچلان، ۲۰۰۵) احتمالات پسینی را بر پایه چگالی‌های برآورد شده رده‌ها مشخص می‌کند.

در ادامه مشاهده i به رده‌ای که بیشینه مقدار احتمال پسینی به آن نزدیکتر باشد و با توجه به قانون

$$\hat{g}_i = \text{argmax}_{g\in ۱,\ldots,G}\ \{\hat{P}r(i,g)\} \quad \text{مشاهده } i \text{ می‌بایست به رده } \hat{g}_i \text{ تخصیص یابد که} \qquad (۲.۲)$$

رده‌بندی می‌شود. حال فرض کنید که مشاهده i دارای یک برچسب شناخته شده مانند g_i است. معمولا بررسی می‌شود که برچسب g_i تا چه حد با نتیجه رده‌بندی کننده هم‌خوانی دارد. پس بیشترین مقدار $\hat{P}r(i,g)$ از رده‌ای بجز g_iها به صورت

$$\tilde{P}r(i) = \max_{i\in ۱\ldots G} \hat{P}r(i,g); \quad g \neq g_i \qquad (۳.۲)$$

تعریف می‌شود. اگر $\hat{P}r(i,g_i) > \tilde{P}r(i)$ آنگاه رده g_i بیشترین مقدار احتمال $\hat{P}r(i,g)$ را دارا بوده و رده‌بندی‌کننده مجاب می‌شود که i را به رده با برچسب g_i تخصیص دهد و اگر $\hat{P}r(i,g_i) < \tilde{P}r(i)$ آنگاه مدل مورد مطالعه، مشاهده i را به رده g_i تخصیص نمی‌دهد.

در عبارت (۱.۲) ملاحظه می‌شود که اگر برای مشاهده i، نامساوی $PAC(i) < ۰/۵$ صادق باشد آنگاه رده‌بندی کننده مورد بحث مشاهده i را به رده g_i تخصیص می‌دهد. واضح است درصورتی که نامساوی $PAC(i) > ۰/۵$ رخ دهد آنگاه برای مشاهده i رده‌های جایگزین بهتری نسبت به g_i وجود دارند.

۲.۲ فاصله و همسایگی

دومین مفهوم مورد نیاز برای تشکیل نقشه رده‌ها، میزان دوری است که در واقع نمایانگر میزان دوری مشاهده i از رده g_i در منظر رده‌بندی‌کننده می‌باشد. برای تعریف این مفهوم، فاصله $D(i,g_i)$ که میزان دوری مشاهده i از رده g_i را نشان

[10] K-Nearest Neighbors (KNN)
[11] Discriminant analysis

می‌دهد، مورد نیاز است. از آنجایی که مفهوم فاصله در رده‌بندی‌کننده‌های مختلف متفاوت است، این فاصله با توجه به نوع رده‌بندی‌کننده مورد مطالعه محاسبه می‌شود.

پس از تعریف فاصله، به دلیل رفتار تصادفی مدل‌های متفاوت رده‌بندی، تابع توزیع تجمعی[۱۲] آن برآورد می‌شود. اگر x یک مشاهده تصادفی تولید شده از رده g_i باشد آنگاه می‌توان چنین تابعی را با $D(x, g_i)$ نشان داد. سپس با داشتن فاصله و تابع توزیع تجمعی، می‌توان کمیت دوری را به صورت

$$\text{farness}(i) = Pr[D(x, g_i) \leq D(i, g_i)] \qquad (۴.۲)$$

تعریف کرد. قابل اشاره است که دوری نیز همانند PAC مقادیری بین ۰ تا ۱ را اختیار می‌کند.

در نمودار نقشه رده‌ها برای تشخیص بهتر میزان دوری، عددهای روی محور افقی در واقع چندک‌های توزیع نرمال استاندارد محدود شده به بازه $[۰, ۴]$ هستند و نقطه‌چین عمودی مشخص شده در نقطه ۰/۹۹ به هدف تشخیص داده‌های دورافتاده با اطمینان ۹۹ درصد رسم می‌شود. مقدار دوری کلی به ازای هر مشاهده i به صورت عبارت ریاضی

$$O(i) = \min_{g \in ۱,\ldots,G} \text{farness}(i, g) \qquad (۵.۲)$$

معرفی می‌شود. زمانی که مقدار $O(i)$ از مقدار جداکننده (در اینجا ۰/۹۹) بیشتر باشد مشاهده i به عنوان مشاهده دورافتاده در نظر گرفته می‌شود.

۳ نقشه رده‌ها برای داده‌های اعداد دست‌نویس فارسی

در این بخش با نرم‌افزار R، پس از کاهش بُعد و برازش مدل به مجموعه داده دست‌نویس فارسی (خسروی و کبیر، ۲۰۰۷) نتایج رده‌بندی رده‌های اعداد ۰ و ۱ از این مجموعه داده با کمک نقشه رده‌ها بررسی می‌شود. لازم به ذکر است که پس از آزمون و خطا، بهترین و بهینه‌ترین روش برای کاهش بُعد این مجموعه داده روش Fast-tSNE[۱۳] (، ۲۰۱۹) و کاهش ابعاد مجموعه داده به دو بُعد و سپس برازش مدل K ـ نزدیکترین همسایگی به نتایج این الگوریتم بوده است.

در شکل ۱ در قسمت بالا نقشه رده‌ها برای نتایج رده‌بندی رده‌های ۰ و ۱، و در قسمت پایین تعدادی از دست‌خط‌های شماره‌گذاری شده در نمودار نمایش داده شده است. منظور از رده‌های ۰ و ۱ در واقع مشاهداتی از این مجموعه داده هستند که برچسب حقیقی آن‌ها ۰ و یا ۱ است. نیمه پایینی خاکستری رنگ نمودار، نمایشگر ناحیه‌ای با نقاطی با شرط $PAC(i) < ۰/۵$ است. این نقاط توسط مدل به درستی به رده مربوط به خودشان تخصیص داده شده‌اند. تعدادی از نقاط در بالای منطقه خاکستری قرار گرفته‌اند. پیام این موضوع آن است که ویژگی‌های توام آن نقاط باعث شده‌اند تا تعلق مشاهدات مربوطه به رده‌های جایگزین محتمل‌تر باشد. به همین دلیل مشاهدات مورد بحث به رده‌های جایگزین که با رنگ‌های جز رنگ رده مورد بررسی (در این مثال، آبی کم‌رنگ برای رده ۰ و آبی پررنگ برای رده ۱) مشخص شده است، تخصیص داده شده‌اند.

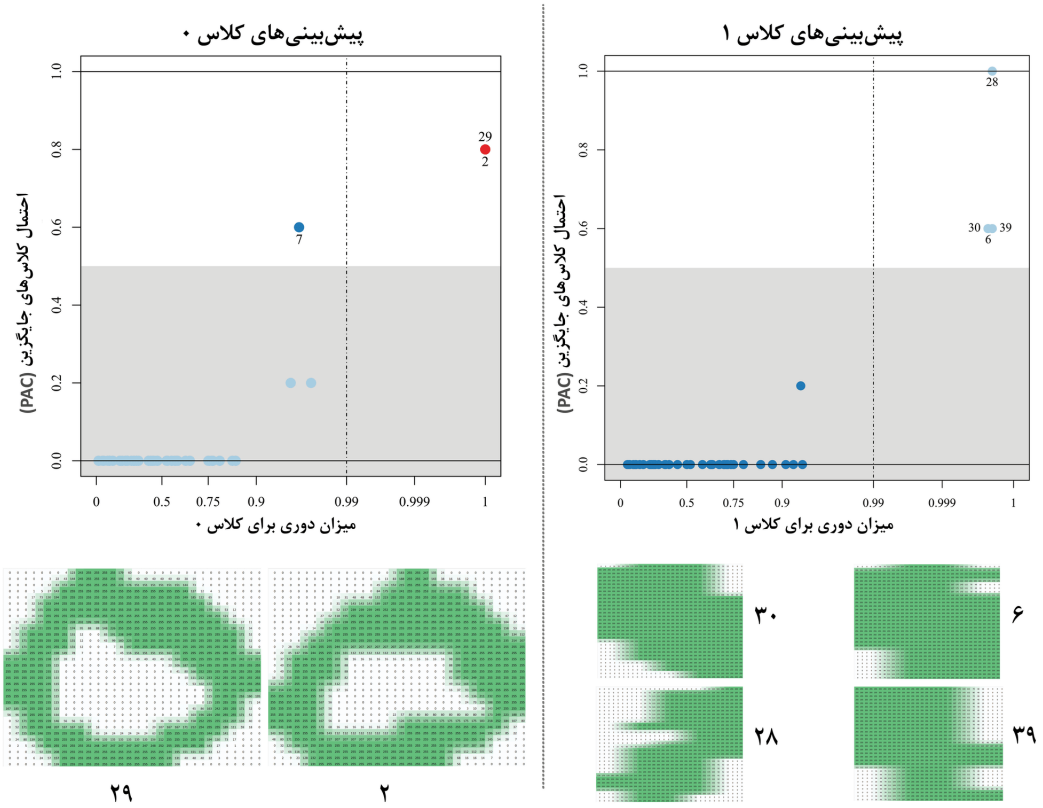
باتوجه به شکل ۱، برای رده ۰ سه نقطه داده و برای رده ۱ چهار نقطه داده مقدار احتمال رده‌های جایگزین بیش از ۰/۵ داشته و به تعبیر رایمکرز و همکاران (۲۰۲۱) این‌ها به عنوان ماهی خارج از آب شناخته می‌شوند، به این معنا که آن‌ها عضو رده تحت بررسی بوده اما به اشتباه در رده دیگری رده‌بندی شده‌اند. نقطه ۲۸ در رده ۱ مقدار PAC نزدیک به ۱ دارد، یعنی مدل با قطعیت این رقم دست‌نوشته را به رده‌ای جز رده ۱ (در اینجا به رده ۰) تخصیص داده است. سه نقطه داده دیگر این رده نیز با مقدار PAC تقریبا برابر در رده اشتباه ۰ قرار گرفته‌اند. با توجه به محور افقی و پارامتر دوری، دونقطه ۲ و ۲۹

[12] Cumulative Distribution Function (CDF)

[13] Fast t-distributed Stochastic Neighbor Embedding

از رده ۰ و سه نقطه ۶، ۳۰ و ۳۹ مشاهدات دورافتاده در نظر گرفته می‌شوند. این مطلب به آن معنا است که ویژگی‌های آنها از مرکز رده خود فاصله معناداری دارد.

باتوجه به تصاویر دست‌خط‌های پیش‌تر ذکر شده، مشخص می‌شود که در رده ۱ سه نقطه با PAC و دوری برابر، شباهت بسیار زیادی داشته و به نظر می‌رسد که با قلم بسیار درشت نوشته شده‌اند و به عدد ۱ شکلی دایره‌ای داده‌اند و به همین دلیل به اشتباه در رده صفر قرار گرفته‌اند. پس از بررسی میانگین تصاویر رده ۱ مشخص شد که نقطه شماره ۲۸ در خلاف جهت سایر ۱ها نوشته شده و مدل توانایی شناسایی این مورد را نداشته است.



شکل ۱: نقشه رده‌ها برای نتایج رده‌بندی رده‌های ۰ و ۱ و تصاویر دست‌خط‌های جالب توجه کشف شده با این نمودار.

در نتایج رده‌بندی رده ۰ با توجه به شکل ۱، نقاط داده‌ی شماره ۲ و ۲۹، میزان دوری و PAC تقریبا برابر دارند و هر دو به اشتباه به رده ۵ تخصیص داده شده‌اند. با بررسی تصاویر این دو دست‌خط به نظر می‌رسد ۰ هایی که نقطه‌ی بالایی آن‌ها نوک تیز نوشته شود از منظر مدل به عنوان عدد ۵ شناسایی می‌شوند. نقشه‌های رده می‌بایست برای هر رده در مجموعه داده‌ای با مشاهدات بیشتر رسم شود تا اگر در نقشه‌ی یک رده نقاط داده‌ای که در یک منطقه تجمع می‌کنند زیاد باشد (مشابه رده ۱ و نوشتارهای با قلم درشت و ۰ های شبیه به ۵)، می‌توان نتیجه گرفت که رده فوق دارای یک زیررده[14] است. با مدنظر قراردادن این مشکل می‌توان در ادامه با اعمال این زیررده‌ها در فرایند برازش مدل، مدل‌های بهینه‌تر و دقیق‌تری را آموزش داد.

[14] Subclass

بحث و نتیجه‌گیری

در این مقاله، باتوجه به دو کمیتی که در تعریف نقشه رده‌ها استفاده شده است، به نظر می‌رسد این ابزار اطلاعات بیشتر و جزئی‌تری از ابزارهای دیگر ارزیابی مدل در اختیار محققین داده قرار می‌دهد. نمودار نقشه رده‌ها، به طور همزمان، احتمال تخصیص مشاهدات به رده‌های جایگزین و میزان فاصله نقاط داده از رده اصلی را نمایش می‌دهد. علاوه بر نکات یاد شده، نقشه رده‌ها اطلاعات مفیدی نیز در مورد ماهیت مجموعه داده مورد بررسی ارائه می‌دهد که در تصمیم‌گیری‌ها و استنباط‌های پژوهشگر داده کمک بسیار زیادی می‌کند.

کمیت PAC علاوه بر نقشی که در نقشه رده‌ها ایفا می‌کند، ابزار بسیارمفیدی برای ارزیابی نتایج انواع رده‌بندی کننده‌ها است. از این ابزار علاوه بر نقشه رده‌ها می‌توان در نمودارها و معیارهای ارزیابی مفید دیگری نیز بهره برد. نکته‌ی قابل توجه در مورد نقشه رده‌ها، موارد استفاده زیاد این نمودار است. در این مقاله تنها به کشف مشاهدات دورافتاده و زیررده‌ها توسط این ابزار اشاره شد. اما از نقشه رده‌ها می‌توان در موارد متعدی از جمله تشخیص دلیل رده‌بندی اشتباه مشاهدات، پیدا کردن هم‌پوشانی رده‌ها، کشف وجود ناهمگنی در رده‌ها و درنتیجه یافتن زیررده‌های جدید درمجموعه داده و مقایسه میان مدل‌های رده‌بندی‌کننده و غیره استفاده کرد. همچنین می‌توان با گسترش ایده این نمودار، ابزار جدیدی برپایه آن با اضافه کردن بُعد سوم برای افزایش اطلاعات مستخرج از این ابزار استفاده کرد. از طرفی باتوجه به ماهیت این ابزار به عنوان نقشه‌ای از نتایج پیش‌بینی، با اعمال همبستگی فضایی در تشکیل اجزای این نمودار می‌توان از این ابزار در موضوعات مرتبط با کریگیدن فضایی و یا مدل‌های رده‌بندی‌کننده‌ی فضایی نیز بهره برد.

مراجع

Bradley, A. P. (1997), The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms, *Pattern recognition*, **30**, 1145-1159.

Brier, G. W. (1950), Verification of Forecasts Expressed in Terms of Probability, *Monthly Weather Review*, **78**, 1-3.

Cohen, J. (1960), A coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, **20**, 37-46.

Dorfman, R. (1979), A Formula for the Gini Coefficient, *The Review of Economics and Statistics*, **61**, 146-49.

Fix, E. and Hodges, J. L. (1989), Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties, *International Statistical Review/Revue Internationale de Statistique*, **57**, 238-247.

Friendly, M. (1994), Mosaic Displays for Multi-Way Contingency Tables, *Journal of the American Statistical Association*, **89**, 190-200.

Good, I. J. (1952), Rational Decisions, *Journal of the Royal Statistical Society: Series B (Methodological)*, **14**, 107-114.

Hartigan, J. A.and Kleiner, B. (1981), Mosaics for Contingency Tables, pp. 268-273.

Khosravi, H. and Kabir, E. (2007), Introducing a Very Large dataset of Handwritten Farsi Digits and a Study on Their Varieties, *Pattern Recognition Letters*, **28**, 1133-1141.

Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., and Kluger, Y. (2019), Fast Interpolation-based t-SNE for Improved Visualization of Single-cell RNA-seq Data, *Nature methods*, **16**, 243-245.

McLachlan, G. J. (2005), *Discriminant Analysis and Statistical Pattern Recognition*, New Jersey: John Wiley & Sons.

Mitchell, T. M. and Mitchell, T. M. (1997), *Machine Learning*, vol. **1**, McGraw-hill, New York.

Raschka, S. (2018), Model Evaluation, Model selection, and Algorithm Selection in Machine Learning, *arXiv preprint arXiv:1811.12808*.

Raymaekers, J., Rousseeuw, P. J., and Hubert, M. (2021), Class Maps for Visualizing Classification Results, *Technometrics*, 1-15.

Rousseeuw, P. J. (1987), Silhouettes: a Graphical Aid to The Interpretation and Validation of Cluster Analysis, *Journal of Computational and Applied Mathematics*, **20**, 53-65.

Stehman, S. V. (1997), Selecting and Interpreting Measures of Thematic Classification Accuracy, *Remote Sensing of Environment*, **62**, 77-89.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

تحلیل بیزی مدل رگرسیون فضایی چوله بر اساس یک زیر کلاس از توزیع CSN

امید کریمی [1]**، فاطمه حسینی**
گروه آمار، دانشگاه سمنان

چکیده: مدل‌های رگرسیون فضایی برای تحلیل پاسخ‌های کمی فضایی براساس روابط خطی و غیرخطی با متغیرهای توضیحی به کار گرفته می‌شوند. معمولا همبستگی فضایی پاسخ‌ها با یک میدان تصادفی گاوسی مدل می‌شوند. اما در عمل با پاسخ‌های چوله مواجه می‌شویم که برای تحلیل آن‌ها از توزیع‌های چوله نرمال استفاده می‌شوند. در این مقاله تحلیل بیز مقداری براساس یک زیر کلاس منعطف از توزیع‌های چوله نرمال بسته ارائه می‌گردد. سپس مدل پیشنهادی روی داده‌های واقعی زمین لرزه‌ای کشور ایران پیاده سازی و مورد تحلیل قرار می‌گیرد.

واژه‌های کلیدی: توزیع چوله نرمال بسته، رهیافت بیز مقداری، داده‌های فضایی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 60G15، 62M30، 60G60.

۱ مقدمه

معمولا در زمینه‌های مختلفی همچون زمین‌شناسی، اپیدمیولوژی، جغرافیا و پزشکی با پاسخ‌های کمی مواجه می‌شویم که دارای همبستگی فضایی هستند. برای مدل‌بندی پاسخ‌های فضایی از مدل‌های رگرسیون فضایی استفاده می‌شود. انسلین (۱۹۹۰) برای لحاظ کردن همبستگی فضایی در مدل‌های رگرسیون روش‌های مختلفی پیشنهاد کرد. اوه و همکاران (۲۰۰۲) رهیافت بیزی سلسله مراتبی را برای این نوع مدل‌ها مورد مطالعه قرار دادند. لی و هوانگ (۲۰۲۲) تاثیر کوید ۱۹ بر بازارهای مسکن ایالات متحده را به‌وسیله مدل‌های رگرسیون فضایی بررسی کردند. معمولا در عمل با پاسخ‌هایی مواجه می‌شویم که دارای چولگی هستند و برای مدل‌کردن آن از توزیع‌های چوله نرمال استفاده می‌شود. کریمی و محمدزاده (۲۰۱۲) تحلیل بیزی برای مدل رگرسیون فضایی چوله بر اساس تحققی از یک میدان تصادفی چوله گاوسی بسته ارائه کردند. در این مقاله تحلیل بیز مقداری [1] روی این مدل‌ها بر اساس یک زیر کلاس منعطف از توزیع چوله نرمال بسته [2] (CSN) بیان می‌شود.

ساختار مقاله به این صورت است که در بخش ۲ یک زیر کلاس از توزیع CSN بیان می‌گردد. مدل رگرسیون فضایی

[1] Variational Bayes (VB)
[2] Closed Skew Normal

[1] نام و ایمیل ارائه دهنده مقاله: امید کریمی، omid.karimi@semnan.ac.ir

چوله به همراه تحلیل بیز مقداری سلسله مراتبی آن در بخش ۳ ارائه می‌گردد. در بخش ۴ روش‌های پیشنهادی در یک مثال واقعی داده‌های زمین لرزه‌ای کشور طی ۱۵ سال اخیر مورد تحلیل قرار می‌گیرند. در نهایت بحث و نتیجه‌گیری ارائه می‌شود.

۲ زیر کلاس توزیع CSN

متغیر تصادفی $\boldsymbol{Y} \sim \mathrm{CSN}_{p,q}(\boldsymbol{\mu}, \Sigma, \Gamma, \boldsymbol{\nu}, \Delta)$ است اگر دارای تابع چگالی (گنزالس و همکاران، ۲۰۰۴) $f_Y(\boldsymbol{y}) = \phi_p(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma)\frac{\Phi_q(\Gamma(\boldsymbol{y}-\boldsymbol{\mu}); \boldsymbol{\nu}, \Delta)}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \Delta+\Gamma\Sigma\Gamma')}$ باشد که در آن $\boldsymbol{\mu} \in R^p$، Σ ماتریس معین مثبت $p \times p$ پارامتر مقیاس، Γ ماتریس $q \times p$ پارامتر چولگی، $\boldsymbol{\nu} \in R^q$ و Δ یک ماتریس $q \times q$ معین مثبت است. ϕ_p و Φ_q به‌ترتیب تابع چگالی و تابع توزیع تجمعی توزیع نرمال چند متغیره هستند. توزیع CSN تحت تبدیلات خطی و مجموع بردارهای تصادفی مستقل بسته است. هر چند این توزیع دارای خواص بسته بودن و انعطاف پذیر است اما تابع چگالی آن دارای پیچیدگی و پارامترهای زیادی است. به عنوان مثال برای q های بزرگ محاسبه Φ_q چالش برانگیز است. به همین خاطر یک زیر کلاس از این توزیع بر اساس ایده مارکز و گنزالس (۲۰۲۲) برای تحلیل مدل رگرسیون فضایی چوله ارائه می‌شود. فرض کنید متغیر تصادفی X_i دارای توزیع CSN به صورت $X_i \sim \mathrm{CSN}_{1,1}(0, 1, \lambda_i, 0, 1)$ و مستقل باشند، آنگاه تابع چگالی به‌صورت $f_X(\boldsymbol{x}) = 2^n\phi_n(\boldsymbol{x}; \mathbf{0}, \mathbf{I_n})\Phi_n(\Gamma\boldsymbol{x})$ به دست می‌آید که در آن $\Gamma = \mathrm{Diag}(\boldsymbol{\lambda})$ و $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$، $\Phi_n(\Gamma\boldsymbol{x}) = \prod_{i=1}^n \Phi(\lambda_i x_i)$ است. در آن صورت $\boldsymbol{X} \sim \mathrm{CSN}_{n,n}(0, I_n, \Gamma, \mathbf{0}, \mathbf{I_n})$ می‌باشد، بنابراین $\mathrm{E}(\boldsymbol{X}) = b\boldsymbol{\delta}$ و $V(\boldsymbol{X}) = \Omega$ است که در آن $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$، $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)'$، $b = (2/\pi)^{\frac{1}{2}}$، $\delta_i = \lambda_i(1 + \lambda_i^2)^{-\frac{1}{2}}$، $\tau_i = (1 - b^2\delta_i^2)^{-\frac{1}{2}}$ و $\Omega = \mathrm{Diag}(\boldsymbol{\tau})$ است. بردار تصادفی $\boldsymbol{X}$ را می‌توان به‌صورت $\boldsymbol{z} = \Omega^{-1}(\boldsymbol{x} - b\boldsymbol{\delta})$ استاندارد کرد، به این ترتیب $\mathrm{E}(\boldsymbol{z}) = \mathbf{0}$ و $\mathrm{V}(\boldsymbol{z}) = I_n$ خواهد شد. در بخش بعد با استفاده از خاصیت بسته بودن توزیع‌های CSN نسبت به تبدیلات خطی مدل رگرسیون فضایی چوله تعریف و تحلیل بیزی آن ارائه می‌شود.

۳ مدل رگرسیون فضایی چوله

با در نظر گرفتن بردار تصادفی استاندارد شده $\boldsymbol{z}$ که در بخش ۲ ارائه شد، مدل رگرسیون فضایی چوله برای بردار پاسخ $\boldsymbol{y}_n$ در n موقعیت فضایی $(s_1, \ldots, s_n)$ را به‌صورت زیر تعریف می‌کنیم:

$$\boldsymbol{y}_n = X\boldsymbol{\beta} + \sigma C\boldsymbol{z}. \tag{۱.۳}$$

که در آن X یک ماتریس $n \times p$ از متغیرهای توضیحی، $\boldsymbol{\beta}$ یک بردار p بعدی از ضرائب رگرسیونی، $\sigma > 0$ پارامتر مقیاس و C یک ماتریس $n \times n$ که از تجزیه ماتریس همبستگی فضایی به‌صورت $\Sigma_n = CC'$ به‌دست می‌آید. در واقع Σ_n ماتریس کوواریانس فضایی بردار تصادفی $\boldsymbol{y}_n$ با مولفه‌های $\sigma_{ij} = \sigma^2\rho(y(s_i), y(s_j); \varphi)$ است که در آن $\rho(\cdot; \varphi)$ یک تابع همبستگی فضایی با پارامتر دامنه فضایی φ است. بنابراین بردار پاسخ‌های فضایی $\boldsymbol{y}_n$ با توجه به خاصیت تبدیلات خطی توزیع CSN دارای توزیع CSN به‌صورت $\boldsymbol{y}_n \sim \mathrm{CSN}_{n,n}(\boldsymbol{\mu}_y, \sigma_y, \frac{1}{\sigma}\Gamma\Omega C^{-1}, \mathbf{0}, \mathbf{I_n})$ است، که در آن $\boldsymbol{\mu}_y = X\boldsymbol{\beta} - b\sigma C\Omega^{-1}\boldsymbol{\delta}$ و $\Sigma_y = \sigma^2 C\Omega^{-2}C'$ است. از این رو تابع چگالی به‌صورت

$$f_\eta(\boldsymbol{y}) = 2^n\phi_n(\boldsymbol{y}; \boldsymbol{\mu}_y, \sigma_y)\Phi_n(\frac{1}{\sigma}\Gamma\Omega C^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y))$$

است، که در آن $\boldsymbol{\eta} = (\boldsymbol{\beta}, \sigma, \varphi, \boldsymbol{\lambda})'$ پارامترهای مدل است. برای برآورد پارامترهای مدل رگرسیون فضایی رابطه (۱.۳) از رهیافت بیزی سلسله مراتبی با توجه به پیچیدگی تابع درست‌نمایی استفاده می‌شود. با در نظر گرفتن پیشین‌های معمول برای

پارامترها به‌صورت $\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_\beta, \Sigma_\beta)$، $\sigma \sim IG(a,b)$، $\varphi \sim G(\alpha,\gamma)$ و $\lambda_i \sim N(\mu_{\lambda_i}, \sigma^2_{\lambda_i})$ توزیع پسین به‌صورت

$$\begin{aligned}\pi(\eta|\boldsymbol{y}) \quad \propto \quad & \phi_n(\boldsymbol{y};\boldsymbol{\mu}_y,\sigma_y)\Phi_n(\frac{1}{\sigma}\Gamma\Omega C^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y))\phi_p(\boldsymbol{\beta};\boldsymbol{\mu}_\beta,\sigma_\beta)\pi(\sigma;a,b)\\ & \times\pi(\varphi;\alpha,\gamma)\Pi_{i=1}^n\phi(\lambda_i;\mu_{\lambda_i},\sigma^2_{\lambda_i})\pi(\boldsymbol{\mu}_\beta)\pi(\Sigma_\beta)\\ & \times\pi(a)\pi(b)\pi(\alpha)\pi(\gamma)\Pi_{i=1}^n\pi(\mu_{\lambda_i})\pi(\sigma^2_{\lambda_i}),\end{aligned}$$

خلاصه می‌شود، که در آن $\pi(\cdot)$ توزیع پیشین ابر پارامترهای $\boldsymbol{\mu}_\beta$، Σ_β، a، b، α، γ، μ_{λ_i} و $\sigma^2_{\lambda_i}$ است. با توجه به پیچیده بودن توزیع پسین لازم است از روش‌های مونت کارلویی و بیز تقریبی استفاده شود. در این مقاله از روش بیز مقداری برای برآورد پارامترها به‌عنوان یک جایگزین مناسب برای روش‌های MCMC استفاده می‌شود.

در ادامه یک روش بیز مقداری برای برازش مدل رگرسیون فضایی چوله پیشنهاد می‌شود. هدف بیز مقداری یافتن توزیع مقداری $q(\boldsymbol{\eta})$ است، که توزیع پسین $\pi(\boldsymbol{\eta}|\boldsymbol{y})$ را از طریق مینیمم کردن فاصله احتمال بین توزیع مقداری و توزیع پسین تقریب می‌زند. فاصله احتمال توسط رابطه کولبک ـ لایبلر(KL) به‌صورت

$$\begin{aligned}\mathrm{KL}(q(\boldsymbol{\eta})\|P(\boldsymbol{\eta}\mid\boldsymbol{y})) &= \int \ln\left(\frac{q(\boldsymbol{\eta})}{P(\boldsymbol{\eta}\mid\boldsymbol{y})}\right) q(\boldsymbol{\eta})d\boldsymbol{\eta} = \mathbb{E}_q[\ln q(\boldsymbol{\eta})] - \mathbb{E}_q[\ln P(\boldsymbol{\eta}\mid\boldsymbol{y})]\\ &= \mathbb{E}_q[\ln q(\boldsymbol{\eta})] - \mathbb{E}_q[\ln P(\boldsymbol{\eta},\boldsymbol{y})] + \ln f(\boldsymbol{y}). \qquad (۲.۳)\end{aligned}$$

تعیین می‌شود و با مینیمم کردن رابطه KL به توزیع هدف زیر به‌عنوان تقریبی از توزیع پسین می‌رسد.

$$q^*(\eta) = \arg\min_q \mathrm{KL}(q(\boldsymbol{\eta})\|P(\boldsymbol{\eta}\mid\boldsymbol{y})).$$

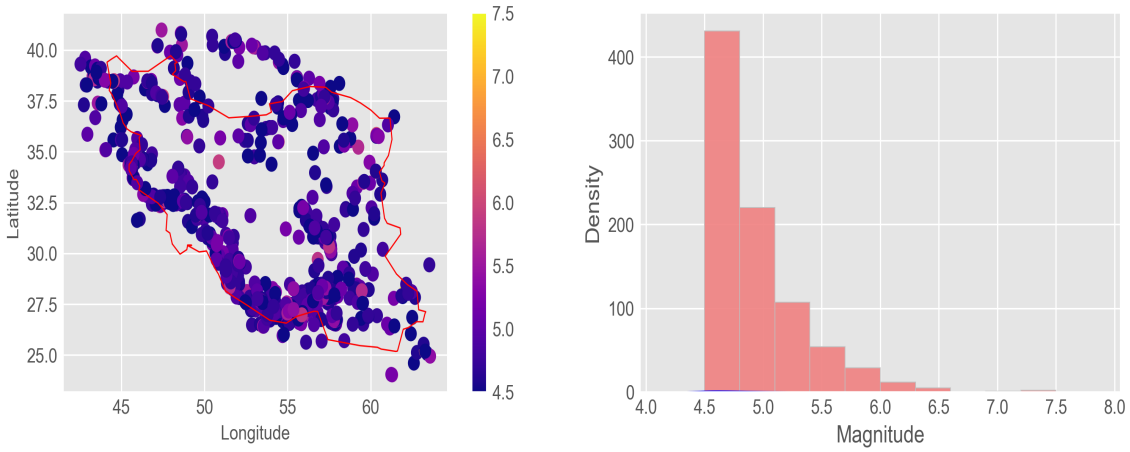
چون $\ln f(y)$ شکل بسته‌ای ندارد، رابطه KL از نظر تحلیلی قابل حل نیست. معادله (۲.۳) را به‌صورت $\mathrm{RLK}(q(\boldsymbol{\eta})) = \ln P(y) - \mathrm{KL}(q(\eta)\|P(\eta\mid y))$ می‌نویسیم. از آنجا که رابطه KL همیشه مثبت است، معادله (۲.۳) نشان می‌دهد که توزیع مقداری بهینه را می‌توان با ماکسیمم کردن رابطه به‌دست آورد. توزیع مقداری باید توسط تحلیلگر انتخاب شود. که معمولا از تقریب های گاوسی استفاده می‌شود.

۴ مثال واقعی: داده‌های زمین لرزه‌ای ایران

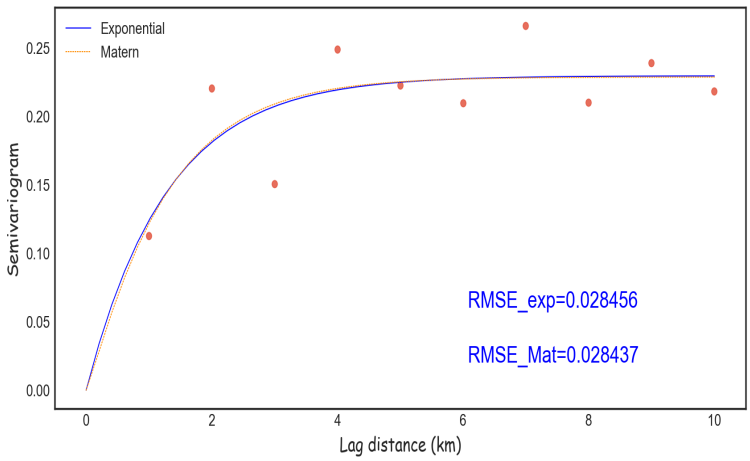
داده‌های زمین لرزه‌ای شامل زمان وقوع زلزله (Date)، عرض جغرافیایی (Lat)، طول جغرافیایی (Long)، عمق (Depth)، شهر (City)، استان (Province) و بزرگی زمین لرزه (Mg) می‌باشد. نقشه‌ی نقاط بزرگی زمین لرزه‌های بالای ۴/۵ در شکل ۱ به همراه هیستوگرام آن‌ها رسم شده است. با توجه به شکل ۱ می‌توان همبستگی مکانی (فضایی) داده‌ها را مشاهده کرد. در ادامه مدل رگرسیون فضایی چوله (۱.۳) به‌منظور ارایه نقشه پیشگویی بزرگی زمین لرزه‌ها در کل نقاط کشور برای مشاهده نقاط پرخطر کشور به‌کار گرفته می‌شود. با در نظر گرفتن بزرگی زمین لرزه به عنوان متغیر پاسخ (Mg_i) برای هر موقعیت s_i و عمق (D_i)، طول $(Long_i)$ و عرض جغرافیایی (Lat_i) زمین لرزه‌ها به عنوان متغیرهای توضیحی مدل رگرسیون فضایی به‌صورت

$$Mg_i = \beta_0 + \beta_1 D_i + \beta_1 Long_i + \beta_2 Lat_i + \psi_i,\ i = 1,2,\ldots, \qquad (۱.۴)$$

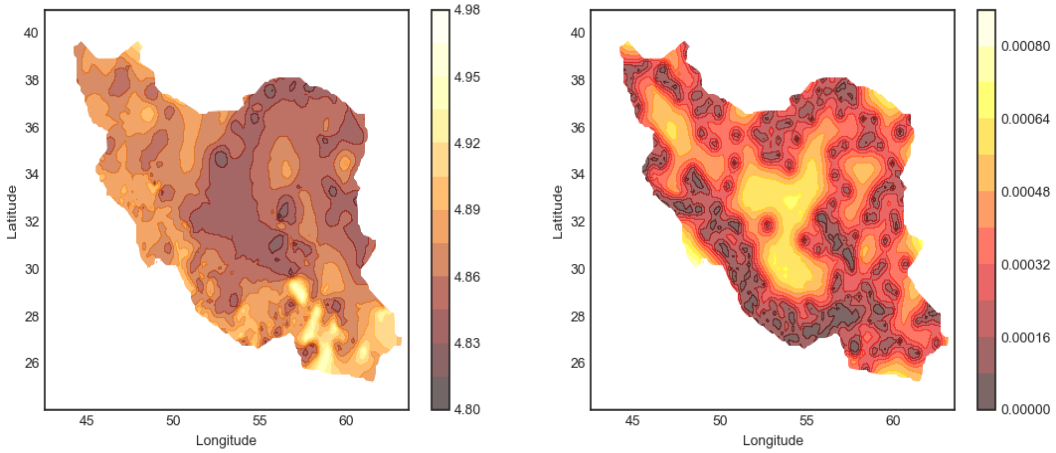
تعریف می‌شود، که در آن ψ_i مولفه iام عبارت $\sigma C\boldsymbol{Z}$ در مدل (۱.۳) است. در این مدل به دلیل شناسایی پذیر بودن پارامتر چولگی به‌صورت $\lambda_i = \lambda$ در نظر گرفتیم. برای تعیین مدل همبستگی فضایی داده‌ها از تغییرنگار تجربی داده‌ها استفاده کرده‌ایم. شکل ۲ نمودار تغییرنگار تجربی به همراه برازش دو مدل تغییرنگار نمایی و ماترن را نشان می‌دهد. همانطور



شکل ۱: (چپ) نمودار هیستوگرام و (راست) نقشه‌ی موقعیت زمین لرزه‌های کشور بیشتر از ۵.۴ طی پانزده سال اخیر



شکل ۲: نمودار تغییرنگار تجربی داده‌های زمین لرزه‌های کشور



شکل ۳: نقشه پیشگویی فضایی و دقت پیشگویی داده‌های زمین لرزه‌های کشور طی پانزده سال اخیر

که مشاهده می‌شود مدل نمایی به خوبی برازش شده است بنابراین این مدل برای تعیین ساختار همبستگی فضایی ماتریس Σ_n در نظر گرفته می‌شود. سپس رهیافت بیز مقداری به همراه روش‌های برآورد پیشنهادی در بخش قبل روی داده‌ها پیاده سازی شد. نتایج برآورد پارامترهای مدل به روش بیز مقداری (VB) در جدول ۱ ارائه شده است که نشان می‌دهد طول و عرض جغرافیایی معنی‌دار نیستند اما عمق لرزه‌ها تاثیر معنی‌داری روی بزرگی لرزه‌ها طی پانزده سال اخیر داشته است. نقشه پیشگویی فضایی و دقت آن در شکل ۳ ارائه شده است که نقاط پر خطر از نظر بزرگی زمین لرزه‌ها را روی نقشه ایران نشان می‌دهد. در واقع نقاط هم‌رنگ از نظر میزان بزرگی لرزه مشابه هستند و به نوعی همبستگی مکانی داده‌ها را نشان می‌دهد.

جدول ۱: نتایج برآورد بیز مقداری پارامترهای مدل رگرسیون فضایی چوله برای داده‌های زمین لرزه‌ای کشور

پارامتر	برآورد	انحراف استاندارد	چندک ۰/۰۲۵	چندک ۰/۹۷۵
β_0	۳/۶۱۹۴	۰/۳۷۲	۲/۸۹۱	۴/۳۴۸
β_1	۰/۰۷۵۲	۰/۰۱۰	۰/۰۵۵	۰/۰۹۵
β_2	−۰/۰۲۳۱	۰/۰۴۹	−۰/۱۱۹	۰/۰۷۲
β_3	۰/۰۴۸	۰/۰۶۴	−۰/۰۷۶	۰/۱۷۴
σ	۰/۲۳۱	۰/۵۶۱	۰/۰۸۱	۱/۱۴۱
φ	۳/۸۶۱	۱/۰۱۷	۰/۲۴۳	۶/۸۶۴
λ	۲/۱۳۱	۱/۵۲۳	۰/۶۸۱	۴/۴۷۲

بحث و نتیجه‌گیری

در این مقاله یک مدل رگرسیون چوله براساس یک زیر کلاس منعطف از توزیع CSN ارائه شد. از رهیافت بیز مقداری سلسله مراتبی با توجه به پیچیدگی تابع درست‌نمایی مدل و بالا بردن سرعت محاسبات در مقایسه با روش‌های MCMC بیان شد. همچنین این مدل پیشنهادی روی داده‌های واقعی زمین لرزه‌ای کشور طی ۱۵ سال اخیر اجرا گردید. که در آن عمق زمین لرزه‌ها تاثیر معنی‌داری در بزرگی زمین لرزه‌های ثبت شده داشت. در نهایت نقشه پیشگویی فضایی روی کل ناحیه فضایی رسم شد که نشان داد بخش‌هایی از جنوب و شرق کشور بیشتر از بقیه نواحی در معرض زمین لرزه‌های به بزرگی ۴/۵ و بیشتر هستند.

مراجع

Anselin, L. (1990), Spatial Dependence and Spatial Structural Instability in Applied Regression Analysis. *J Reg Sci*, **30**, 185–207.

Gonzalez-Farias, G., Dominguez-Molina, A. and Gupta, A. K., (2004), The Closed Skew Normal Distribution. In: *Genton M. G., ed. Skew-elliptical distributions and their applications: A journey beyond normality. Boca Raton, FL: Chapman and Hall CRC,* 2542.

Karimi, O. and Mohammadzadeh, M. (2012), Bayesian Spatial Regression Models with Closed Skew Normal Correlated Errors and Missing, *Statistical Papers* **53(1)**, 205-218.

Márquez-Urbina, O.U., González-Farías, G., (2022). A Flexible Special Case of the CSN for Spatial Modeling and Prediction, *Spatial Statistics*, **47**, 100556.

Lee, J. and Huang, Y. (2022) Covid-19 Impact on US Housing Markets: Evidence from Spatial Regression Models, *Spatial Economic Analysis*, **17**:3, 395-415.

Oh, M., Shina, D.W., and Kim, H.J. (2002) Bayesian Analysis of Regression Models with Spatially Correlated Errors and Missing Observations, *Computational Statistics and Data Analysis (CSDA)*, **39**, 387–400.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

مدل‌سازی صریح وابستگی فضایی برای تحلیل بیزی داده‌های بقا

ساجده مرادنیا[1]، موسی گلعلی‌زاده
گروه آمار، دانشگاه تربیت مدرس

چکیده: در عصر پیشرفت فناوری، حجم غیر قابل تصوری از اطلاعات از جمله داده‌های متنی، تصویری و ویدئویی در پایگاه‌های داده‌ای ذخیره می‌شوند. در این بین، داده‌های تصویر به ویژه تصاویر با وضوح بالا، به عنوان یک نوع مهم از داده‌های فضایی بعد بالا تلقی می‌شوند که متشکل از ماتریسی از پیکسل‌ها (متغیرها) هستند. در این داده‌ها، هر پیکسل نه تنها دارای اطلاعاتی در مورد رنگ و شدت نور تصویر است، بلکه از نظر موقعیت مکانی در تصویر هم دارای اهمیت بالایی است. مدیریت و تحلیل چنین مجموعه داده‌هایی، چالش‌های خاص خود را به دنبال دارد. در این میان، خوشه‌بندی راهنماییده با دخالت دادن متغیر پاسخ، به عنوان یک ابزار قدرتمند، الگوها و ویژگی‌های نهان در داده‌های تصویر را شناسایی کرده و اطلاعات مفیدی را استخراج می‌کند. مقاله حاضر، نحوه اجرای الگوریتم‌های خوشه‌بندی راهنماییده را به طور مختصر تشریح نموده و عملکرد آن‌ها را بر روی مجموعه داده تصاویر دست‌نوشته‌های فارسی بررسی خواهد کرد.

واژه‌های کلیدی: آمار فضایی، داده‌های بعد بالا، خوشه‌بندی راهنماییده، تصاویر دست‌نوشته فارسی، پردازش تصویر.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H30، 62H99.

۱ مقدمه

در آمار فضایی، مفهوم استقلال مشاهدات مانند آمار سنتی نیست، زیرا مشاهدات در اینجا به یکدیگر وابسته هستند و این وابستگی ممکن است بر اساس موقعیت‌های جغرافیایی باشد. داده‌های فضایی به سه دسته اصلی تقسیم می‌شوند: داده‌های زمین‌آماری که در نواحی پیوسته ثبت می‌شوند، الگوهای نقطه‌ای که مکان مشاهده‌ها متغیری تصادفی است و داده‌های مشبکه‌ای که در مکان‌های ناحیه‌ای منظم یا نامنظم قرار دارند (محمدزاده ، ۱۳۹۴). به عنوان مثال، تصاویر ماهواره‌ای نمونه‌ای از داده‌های مشبکه‌ای منظم هستند. در دنیای چند رسانه‌ای امروزی، پردازش تصویر[1] اهمیت زیادی دارد. در این زمینه، تصاویر ورودی تحلیل می‌شوند و خروجی‌هایی مانند تغییرات در نمایش تصویر، تشخیص ویژگی‌ها و حتی فشرده‌سازی

[1]Image processing

[1] نام و ایمیل ارائه دهنده مقاله: ساجده مرادنیا، moradniasajedeh1371@gmail.com

تصویر تولید می‌شوند (پیتاس و ونتسانوپولوس ، ۱۹۹۲). از سوی دیگر، خوشه‌بندی متغیرها به عنوان یک روش در تحلیل داده‌های بعد بالا اطلاعات مفیدی را در مورد همبستگی و تفاوت‌های متغیرها ارائه می‌دهد. در این راستا، خوشه‌بندی راهنماییده رویکرد نوینی است که متغیر پاسخ را در فرآیند خوشه‌بندی مشارکت داده و به تعیین متغیرهای مهم برای پیش‌بینی متغیر پاسخ می‌پردازد (دتلینگ و بوهلمن ، ۲۰۰۲). این روش برای تحلیل داده‌های تصویری که عناصر تشکیل‌دهنده آن‌ها (پیکسل‌ها) به عنوان متغیرها در نظر گرفته می‌شوند نیز قابل تعمیم است. این رویکرد نسبت به روش‌های معمول خوشه‌بندی دارای مزایا و جذابیت‌های خاصی است و می‌تواند در تحقیقات علمی بسیار مفید باشد.

در مقاله حاضر، ابتدا مرور مختصری بر روی الگوریتم‌های خوشه‌بندی راهنماییده از جمله ویلما[۲] (مبتنی بر آماره آزمون ویلکاکسون) و پلورا[۳] (مبتنی بر تابع لگاریتم درستنمایی تاوانیده با استفاده از تاوان ریج) خواهیم داشت. سپس، به معرفی روش پیشنهادی پلاس[۴] که ترکیب الگوریتم پلورا با تاوان لاسو است، خواهیم پرداخت. در ادامه، عملکرد هر یک از این الگوریتم‌ها را بر روی مجموعه داده فضایی تصاویر دست‌نوشته‌های فارسی بررسی و تحلیل خواهیم کرد.

۲ خوشه‌بندی راهنماییده

خوشه‌بندی راهنماییده یک ابزار قدرتمند در تجزیه و تحلیل داده‌های تصویر است. در این روش، با استفاده از اطلاعات موجود در متغیر پاسخ، تصاویر با ویژگی‌های مشابه مانند نوع بافت مشابه، در یک خوشه قرار می‌گیرند. این رویکرد به تشخیص الگوها و ویژگی‌های مهم در تصاویر کمک می‌کند و در تحلیل دقیق‌تر تصاویر تاثیرگذار است. به عنوان مثال می‌توان به تحلیل تصاویر پزشکی در تشخیص نواحی سرطانی موجود در تصاویر سی تی اسکن اشاره کرد. در این موارد، اطلاعات موجود در تصاویر به صورت پیچیده و با ابعاد بالا هستند و با استفاده از خوشه‌بندی راهنماییده، می‌توان نواحی مختلف تصاویر مرتبط با ویژگی‌های خاص (مثلاً نوع بافت سالم و سرطانی) را تشخیص داد. به طور کلی، خوشه‌بندی راهنماییده در تحلیل مجموعه داده‌های فضایی مشبکه‌ای می‌تواند به عنوان ابزاری قدرتمند در تشخیص الگوها، تفاوت‌ها و ویژگی‌های مهم تصاویر به کار گرفته شود. در این بخش به طور خلاصه به مرور الگوریتم‌های خوشه‌بندی راهنماییده پرداخته و در بخش بعدی کاربرد هر یک را بر روی مجموعه داده تصاویر دست‌نوشته‌های فارسی بررسی خواهیم کرد. لازم به ذکر است در این مقاله از واژه "متغیر" به جای "پیکسل" استفاده می‌کنیم.

۱.۲ ویلما

به عنوان اولین الگوریتم خوشه‌بندی راهنماییده برای تحلیل داده‌های بعد بالای تصویر، می‌توان به الگوریتم "ویلما" اشاره کرد. بنا به دتلینگ و بوهلمن (۲۰۰۲)، در رویکرد خوشه‌بندی راهنماییده از یک مدل تصادفی پایه برای داده‌های بعد بالا استفاده می‌شود. در این مدل، متغیر پاسخ رسته‌ای به صورت زوج‌های تصادفی $(\mathbf{X}, Y)$ با مقادیر از داخل مجموعه $\mathbb{R}^p \times \mathbb{Y}$ هستند به طوری که $\mathbf{X} \in \mathbb{R}^p$ بیانگر مقادیر عددی متغیرهای پیشگو هستند که با میانگین صفر و واریانس یک استاندارد شده‌اند. مجموعه $\mathbb{Y}$ برای متغیر پاسخ، مقادیر عددی از مجموعه $\mathbb{Y} = \{۰, ۱, \dots K-۱\}$ هستند که K نمایانگر تعداد حالات متغیر پاسخ است. برای سادگی بحث و فهم آسان مطالب پیش رو، K برابر ۲ در نظر گرفته می‌شود. فرض کنید، تنها تعداد کمی از متغیرها تعیین‌کننده وضعیت متغیر پاسخ هستند.

آنگاه می‌توان احتمال شرطی $P(Y = ۱|\mathbf{X}) = f(\tilde{\mathbf{X}}) = f(\tilde{X}_{C_۱}, \tilde{X}_{C_۲}, \dots, \tilde{X}_{C_q})$ را تعریف کرد که در آن $f(.)$ تابعی غیرخطی از $\mathbb{R}^q$ به $[۰, ۱]$ است. مجموعه $\{C_۱, \dots C_q\}$ جایی‌که $q \ll p$ خوشه‌هایی از متغیرها هستند به‌طوری که

[۲] Wilma
[۳] Pelora
[۴] Pelass

$\{\cup_{j=۱}^{q} C_j\} \subset \{۱, \ldots, p\}$ و به ازای $i \neq j$، $C_i \cap C_j = ۰$. فرض کنید $\tilde{X}_{C_j} \in \mathbb{R}$ نشان‌دهنده نماینده هر خوشه C_j است. ترکیب خطی ساده $\tilde{X}_{C_j} = \frac{۱}{|C_j|} \sum_{g \in C_j} \alpha_g X_g$ بهترین انتخاب برای $\tilde{X}_{C_j}$ است که در آن $|C_j|$ تعداد متغیرهای موجود در خوشه C_j و $\alpha_g \in \{-۱, ۱\}$. با این حال، یافتن زیرمجموعه‌ای از p متغیر و تشکیل خوشه‌های $\{C_۱, \ldots, C_q\}$ با ساختار احتمالاتی دشوار است. برای کمک به این موضوع، دتلینگ و بوهلمن (۲۰۰۲) آماره آزمون ویلکاکسون را به عنوان تابع هدف معرفی کردند. با پیروی از آن‌ها، امتیاز متغیر $j-$ام از بردار $n-$بعدی مقادیر مشاهده‌شده متغیرها یعنی $\xi_i = (x_{۱j}, \ldots, x_{nj})$ به صورت

$$Score(\xi_j) = s(\xi_j) = \sum_{i \in N_۰} \sum_{l \in N_۱} ۱_{[x_{ij} \geq x_{il}]} \qquad (۱.۲)$$

قابل محاسبه است که در آن، x_{ij} مقدار عددی متغیر j در مورد پاسخ i و N_k نشان‌دهنده مجموعه‌ای از $n_k \in \{۱, \ldots, n\}$ حالت‌های مختلف متغیر پاسخ از نوع $k \in \{۰, ۱\}$ است. اگرچه تابع امتیاز تغییریافته دارای قابلیت‌های فراوانی است، اما در برخی مواقع عملکرد خوبی ندارد. لذا، ضروری است تابع امتیاز مد نظر به طریقی اصلاح شود. برای این منظور، دتلینگ و بوهلمن (۲۰۰۲) تابع حاشیه $Margin(\xi_j) = m(\xi_j) = \min_{l \in N_۱}(x_{lj}) - \max_{i \in N_۰}(x_{ij})$ را معرفی کردند که مقیاسی پیوسته برای تفکیک متغیر پاسخ است، که در آن $N_۰$، $N_۱$ و x_{ij} همان نمادها در رابطه (۱.۲) هستند. توجه شود که مقدار تابع حاشیه مثبت خواهد بود اگر و تنها اگر تابع امتیاز صفر باشد و $\tilde{\xi}_j$ به طور کامل و به بهترین نحو، متغیر پاسخ را تفکیک نماید و در غیر این صورت، منفی خواهد بود.

۲.۲ پلورا

با وجود همه مزیت‌های قابل توجه روش ویلما، اما آن دارای محدودیت‌هایی است. دتلینگ و بوهلمن (۲۰۰۴) الگوریتم راهنماییده دیگیری به نام "پلورا" را معرفی کردند. این الگوریتم متکی بر مدل رگرسیون لوژستیک تاوانیده بوده و شامل انتخاب متغیرهای مهم، خوشه‌بندی آن‌ها و طبقه‌بندی مشاهدات است. از نقطه نظر ریاضی و برای رفع مشکل الگوریتم ویلما، دتلینگ و بوهلمن (۲۰۰۴) منفی تابع لگاریتم درستنمایی تاوانیده (مبتنی بر نُرم $L_۲$) را به کار گرفتند و معیار

$$S(\boldsymbol{\theta}) = -\sum_{i=۱}^{n} (Y_i \log p_{\boldsymbol{\theta}}(\tilde{X}_{C_i}) + (۱ - Y_i) \log(۱ - p_{\boldsymbol{\theta}}(\tilde{X}_{C_i}))) + n\frac{\lambda}{۲}\boldsymbol{\theta}^T P \boldsymbol{\theta} \qquad (۲.۲)$$

را براساس احتمالات شرطی $p_{\boldsymbol{\theta}}(\tilde{\mathbf{X}}) = P_{\boldsymbol{\theta}}(Y = ۱|\tilde{\mathbf{X}})$ معرفی کردند و آن را پلورا نامیدند. در رابطه (۲.۲)، $\boldsymbol{\theta}$ برداری از پارامترها است، λ پارامتر تنظیم‌کننده است که میزان تاوان را کنترل می‌کند و P ماتریس تاوان است. از آن‌جایی که هدف محقق از به کارگیری رگرسیون لوژستیک تاوانیده، برآورد $\boldsymbol{\theta}$ از طریق اصل ماکزیمم درستنمایی تاوانیده است، لازم است رابطه (۲.۲) نسبت به $\boldsymbol{\theta}$ مینیمم شود. توجه شود که در رابطه (۲.۲)، ماتریس تاوان (P) به صورت

$$P = \begin{bmatrix} ۰ & ۰ & \ldots & ۰ & ۰ \\ ۰ & Var(\tilde{X}_{C_۱}) & \ldots & ۰ & ۰ \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ ۰ & ۰ & \ldots & Var(\tilde{X}_{C_{q-۱}}) & ۰ \\ ۰ & ۰ & \ldots & ۰ & Var(\tilde{X}_{C_q}) \end{bmatrix} \qquad (۳.۲)$$

در نظر گرفته می‌شود که واریانس نمایندگان هر یک از q خوشه بر روی قطر آن قرار دارد و سایر درایه‌هاش صفر است. برای مینیمم‌کردن رابطه (۲.۲) لازم است از $S(\boldsymbol{\theta})$ نسبت به $\boldsymbol{\theta}$ مشتق گرفت و نتیجه را به صورت

$\tilde{X} = (\mathbf{1}, \tilde{X}_{C_{i\mathbf{1}}}, \ldots, \tilde{X}_{C_{iq}})$ ماتریس طرح است که شامل مراکز خوشه‌ها بوده و $\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \tilde{X}^T(y - \pi_{\boldsymbol{\theta}}) - n\lambda P\boldsymbol{\theta} = \mathbf{0}$ برابر صفر قرار داد. که در آن به ازای $i = \mathbf{1}, \ldots, n$، کمیت $\pi_{\boldsymbol{\theta}} = (p_{\boldsymbol{\theta}}(\tilde{\mathbf{X}}_{\mathbf{1}}), \ldots, p_{\boldsymbol{\theta}}(\tilde{\mathbf{X}}_n))^T$ بردار احتمال شرطی برای تمام n مشاهده است. حاصل این مشتق‌گیری، تعداد $q + \mathbf{1}$ معادله غیر خطی است که پاسخ‌های آن‌ها باید به صورت تقریبی محاسبه شود. بدین منظور، با استفاده از الگوریتم نیوتن‌ـرافسون[5] رابطه $\boldsymbol{\theta}^{new} = (\tilde{X}^T W_{\boldsymbol{\theta}} \tilde{X} + n\lambda P)^{-\mathbf{1}}(\tilde{X}^T(y - \pi_{\boldsymbol{\theta}}) + (\tilde{X}^T W_{\boldsymbol{\theta}} \tilde{X})\boldsymbol{\theta})$ یک روش تکراری برای برآورد بردار پارامتری $\boldsymbol{\theta}$ خواهد بود.

۳.۲ پلاس

در این زیر بخش روش راهنماییده جدیدی به نام "پلاس" را معرفی می‌کنیم. علت چنین نام گذاری این است که روش ما، حاصل ترکیب الگوریتم پلورا با تاوان لاسو است. برای این منظور، خواهیم داشت:

$$S(\boldsymbol{\theta}) = \sum_{i=\mathbf{1}}^{n} (Y_i \log p_{\boldsymbol{\theta}}(X_i) + (\mathbf{1} - Y_i) \log(\mathbf{1} - p_{\boldsymbol{\theta}}(X_i))) - n\frac{\lambda}{\mathbf{2}} P \|\boldsymbol{\theta}\|_{\mathbf{1}}, \tag{۴.۲}$$

که در آن، $\boldsymbol{\theta}^T = (\theta_{\mathbf{0}}, \theta_{\mathbf{1}}, \ldots, \theta_q)$ بردار پارامتری، λ پارامتر تنظیم‌کننده‌ای است که میزان تاوانیدن را کنترل می‌کند و P ماتریس تاوان است. برای حل مسئله بهینه‌سازی، مشکلی که در این جا با آن مواجه هستیم این است که در رابطه (۴.۲)، تابع قدر مطلق $\|\boldsymbol{\theta}\|_{\mathbf{1}}$ در صفر مشتق‌پذیر نیست. این مشکل را می‌توان با استفاده از رویکرد "نزول مختصات چرخه‌ای[6]" برطرف کرد. بدون از دست دادن کلیت مسئله، تابع هدف را به صورت زیر می‌نویسیم:

$$S(\boldsymbol{\theta}) = \underbrace{\sum_{i=\mathbf{1}}^{n} (Y_i \tilde{X}_{C_i}\boldsymbol{\theta} - \log[\mathbf{1} + \exp(\tilde{X}_{C_i}\boldsymbol{\theta})])}_{part\mathbf{1}} - \underbrace{\frac{\lambda n}{\mathbf{2}} P(\boldsymbol{\theta})}_{part\mathbf{2}}, \tag{۵.۲}$$

که در آن، $P(\boldsymbol{\theta}) = \sum_{j=\mathbf{0}}^{q} Var(\tilde{X}_{C_i})|\theta_j|$ است. برای مشتق گرفتن از رابطه (۵.۲) نسبت به $\boldsymbol{\theta}$ و قرار دادن نتیجه برابر با صفر، خواهیم داشت:

$$\begin{aligned} \frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{i=\mathbf{1}}^{n} (Y_i - \frac{\exp(\tilde{X}_{C_i}\boldsymbol{\theta})}{\mathbf{1} + \tilde{X}_{C_i}\boldsymbol{\theta}}) \tilde{X}_{C_i}^T - \frac{n\lambda}{\mathbf{2}} \dot{P}(\boldsymbol{\theta}) \\ &= \tilde{X}_{C_i}^T [Y - g^{-\mathbf{1}}(\tilde{X}_{C_i}, \boldsymbol{\theta})] - \frac{n\lambda}{\mathbf{2}} \dot{P}(\boldsymbol{\theta}) \\ &= \tilde{X}_{C_i}^T [Y - \pi_{\boldsymbol{\theta}}] - \frac{n\lambda}{\mathbf{2}} \dot{P}(\boldsymbol{\theta}) = \mathbf{0}. \end{aligned} \tag{۶.۲}$$

به طوری که در آن، $g^{-\mathbf{1}}(\tilde{X}_{C_i}, \boldsymbol{\theta})$ تابع پیوند و $\pi_{\boldsymbol{\theta}} = (p_{\boldsymbol{\theta}}(\tilde{\mathbf{X}}_{\mathbf{1}}), \ldots, p_{\boldsymbol{\theta}}(\tilde{\mathbf{X}}_n))^T$ بردار احتمال شرطی برای تمام n مشاهده است. از آن جایی که برای وقتی که مشتق وجود نداشته باشد، زیر گرادیان جایگزین مناسبی است، لذا برای مشتق‌گیری از عبارت $part\mathbf{2}$ در رابطه (۵.۲) نسبت به θ_j از مفهوم زیر گرادیان استفاده می‌کنیم که در این جا آن را $\dot{P}(\boldsymbol{\theta})$ نامیدیم. هدف از حل معادله (۶.۲) دستیابی به برآوردی برای $\boldsymbol{\theta}$ است. اما، با توجه به این که از نمایندگان خوشه‌ها ($\tilde{X}_{C_i}$) در این رابطه استفاده کرده‌ایم، $\boldsymbol{\theta}$ حاصل از رابطه (۶.۲) را $\tilde{\boldsymbol{\theta}}$ می‌نامیم. از آن جایی که $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_{\mathbf{0}}, \tilde{\theta}_{\mathbf{1}}, \ldots, \tilde{\theta}_q)$ تابعی از مقادیر غیر خطی است، بنابراین، $q + \mathbf{1}$ معادله غیر خطی در $\lambda \dot{P}(\boldsymbol{\theta}) = \lambda P\tilde{\boldsymbol{\theta}}$ با استفاده از زیر گرادیان حل می‌شوند. به

[5] Newton–Raphson
[6] Cyclic coordinate descent

عبارتی دقیق‌تر می‌دانیم:

$$\tilde{\boldsymbol{\theta}}_j = \begin{cases} -۱ & \tilde{\boldsymbol{\theta}}_j < ۰ \\ [-۱, ۱] & \tilde{\boldsymbol{\theta}}_j = ۰ \\ ۱ & \tilde{\boldsymbol{\theta}}_j > ۰. \end{cases} \qquad (۷.۲)$$

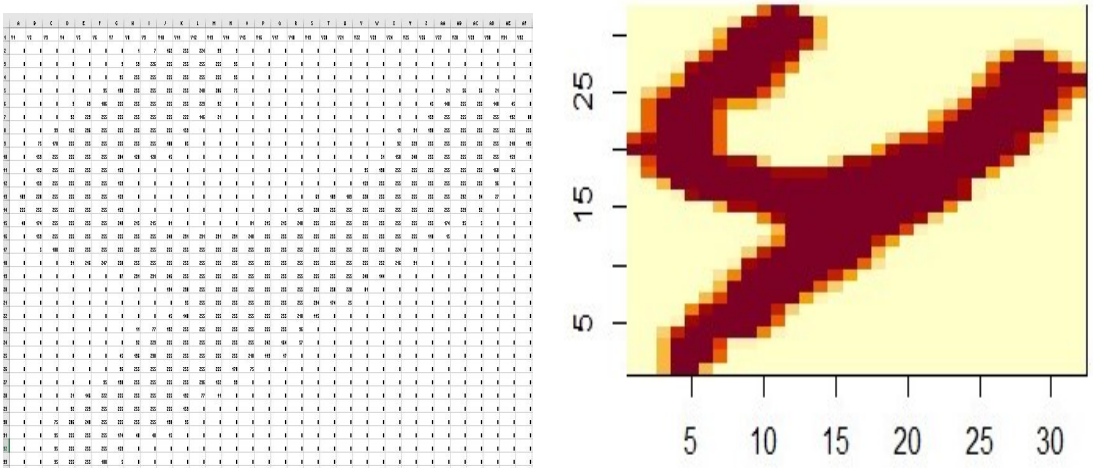
در نتیجه پاسخ نهایی به صورت $\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \tilde{X}_{C_i}^T[Y - \pi_{\boldsymbol{\theta}}] - \frac{n\lambda}{۲}\tilde{\boldsymbol{\theta}}_j$. خواهد بود.

۳ تحلیل مثال واقعی

در این بخش از مقاله حاضر، ابتدا به تعریف مجموعه داده تصاویر دست‌نوشته‌های فارسی که مثالی از داده‌های فضایی مشبکه‌ای است می‌پردازیم. در ادامه، عملکرد الگوریتم ویلما را بر روی آن بررسی خواهیم کرد. ارزیابی روش پلاس و مقایسه نتایج حاصل با نتایج حاصل از الگوریتم پلورا بر روی این مجموعه داده در انتهای بخش خواهد آمد.

۱.۳ شرح مختصری از مجموعه داده تصاویر دست‌نوشته‌های فارسی

مجموعه داده تصاویر دست‌نوشته‌های فارسی که از آن با نام "هدی[۷]" یاد می‌شود، از جمله اولین مجموعه بزرگ تصاویر ارقام دست نویس فارسی است که مشتمل بر ۱۰۲۳۵۲ نمونه است. در گردآوری این مجموعه داده، از فرم‌های ثبت‌نام آزمون سراسری استفاده شده است. تعدادی از فیلدهای عددی در این فرم‌ها با دقت اسکن شده و سپس به صورت دستی تصحیح شدند. این مجموعه داده به دو بخش آموزش و آزمایش تقسیم شده است که هر کدام به ترتیب شامل ۶۰۰۰۰ و ۲۰۰۰۰ نمونه می‌باشند. این مجموعه داده به عنوان مثالی از مجموعه داده‌های فضایی مشبکه‌ای در اختیار محققان قرار گرفته است. در شکل ۱ نمونه‌ای از تصویر رقم "شش" نوشته‌شده توسط کاربر و ماتریس ۳۲ × ۳۲ بعدی تشکیل‌دهنده آن قابل مشاهده است.



(ب) (الف)

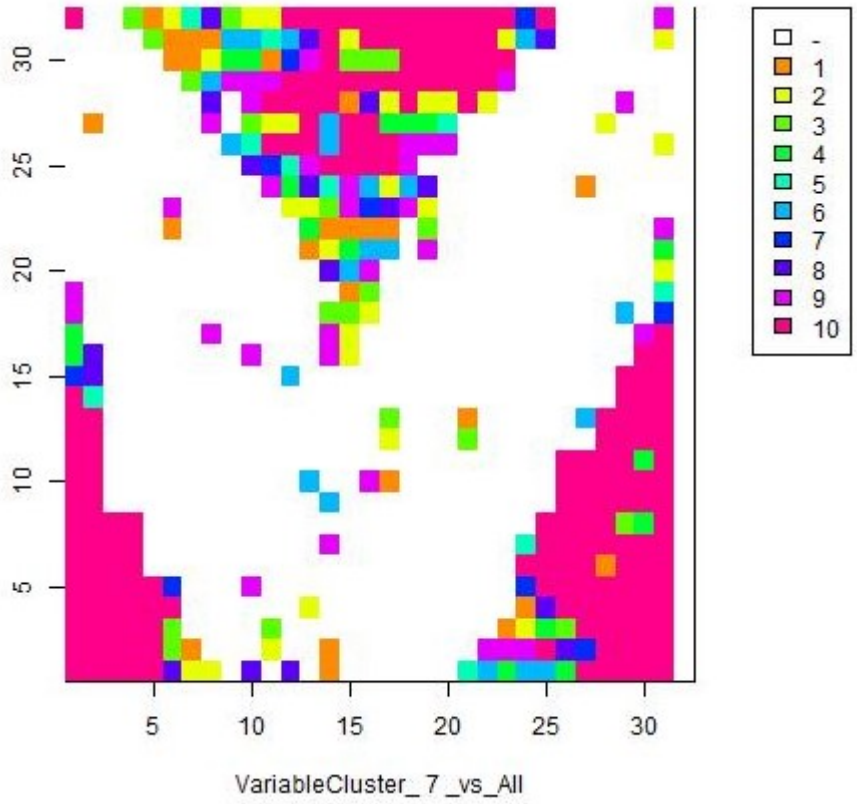
شکل ۱: نمونه رقم ۶ نوشته‌شده توسط کاربر (الف) و ماتریس رقم ۶ نوشته‌شده توسط کاربر (ب).

[7] Hoda

۲.۳ عملکرد خوشه‌بندی راهنماییده بر روی مجموعه داده تصاویر دست‌نوشته‌های فارسی

برای پیاده‌سازی الگوریتم‌های خوشه‌بندی راهنماییده بر روی مجموعه داده هدی در گام اول، ۵۰۰ نمونه اول از مجموعه داده آموزش را انتخاب کردیم. بنا به تیبشیرانی و همکاران (۲۰۰۱) و با استفاده از "آماره فاصله[8]"، تعداد بهینه ۱۰ خوشه برای خوشه‌بندی متغیرهای مجموعه داده هدی مشخص شد. با در نظر گرفتن این تعداد خوشه، الگوریتم ویلما با استفاده از روش "یکی مقابل همه[9]"، برای تشخیص متغیرهای (پیکسل‌های) مهم اجرا شد. به عنوان مثال، الگوریتم ویلما با دقت ۹۷ درصدی، ۴۲۲ متغیر از ۱۰۲۴ متغیر را به عنوان متغیرهای مهم با قرار دادن رقم "هفت" مقابل بقیه، تشخیص داد که به دلیل محدودیت فضای نوشتاری، از ذکر آن‌ها خودداری می‌کنیم. شکل ۲ متغیرهای مهم رقم "هفت" را در مقابل سایر ارقام و خوشه‌های آن‌ها نشان می‌دهد. لازم به ذکر است الگوریتم راهنماییده ویلما از قدرت پیش‌بینی بسیار خوبی برای تمام ارقام مجموعه داده فضایی هدی برخوردار است که به دلیل محدودیت فضا از بررسی جزئیات، صرف نظر می‌کنیم.

برای پیاده‌سازی روش پلورا و پلاس بر روی مجموعه داده هدی، با استفاده از رویکرد اعتبار سنجی متقابل، مقدار بهینه



شکل ۲: متغیرهای مهم رقم "هفت" با استفاده از روش ویلما در مجموعه داده هدی.

پارامتر تنظیم‌کننده (λ) را محاسبه کردیم و در حالت‌های یکسان و متفاوت از نظر تعداد خوشه و پارامتر تنظیم کننده (λ)، به مقایسه نتایج حاصل از این دو روش پرداختیم. نتایج در جدول ۱ گزارش شده است. در این جدول، noc نشان‌دهنده تعداد خوشه است. علاوه بر این، در جدول ۱، حالت‌هایی که روش پلورا عملکرد بهتری نسبت به الگوریتم پلاس دارد، با رنگ سبز مشخص شده است.

با توجه به نتایج جدول ۱، و تعداد کم خانه‌های جدول با رنگ سبز، روش راهنماییده پلاس در مقایسه با الگوریتم پلورا، در

[8] Gap statistic
[9] One-against-all

جدول ۱: مقایسه دقت پیش‌بینی دو رویکرد پلاس و پلورا بر روی مجموعه داده هدی. لازم به ذکر است تمامی مقادیر جدول در ۱۰۰ ضرب شده‌اند.

ارقام	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
Pelass(λ=best, noc= 10)	۸۸	۷۳	۵۷	۷۹	۵۷	۸۵	۶۴	۵۴	۷۶	۷۶
pelora(λ=best, noc= 10)	۶	۶۶	۱۸	۷۵	۷۵	۸۴	۹۰	۶۰	۸۴	۳۰
pelora(λ=1/32, noc=3)	۱۵	۴۸	۲۷	۷۵	۸۱	۶۰	۶۰	۶۶	۸۱	۳۰
pelora(λ=1/32, noc=10)	۱۵	۷۲	۱۸	۶۹	۷۸	۵۱	۸۷	۶۹	۶۰	۳۳
pelora(λ=1/2, noc=3)	۱۲	۹۳	۱۲	۵۴	۷۵	۷۵	۸۴	۴۵	۷۵	۳۹

اکثر موارد عملکرد بهتری بر روی داده‌های تصویر دارد.

بحث و نتیجه‌گیری

می‌توان گفت که روش پیشنهادی ما (پلاس) با استفاده از تاوان لاسو، قابلیت پیش‌بینی خوبی را در مواجه با داده‌های تصویر نشان داده است. علاوه بر این، تاوان لاسو با کاهش بعد داده‌ها، از حجم داده‌ها کاسته و سرعت پردازش را بهبود بخشیده است. با توجه به معیار ارزیابی دقت، می‌توان گفت که با استفاده از تاوان لاسو می‌توان تعداد متغیرها را به شدت کاهش داد و در عین حال دقت بالایی در امر پیش‌بینی به دست آورد. در کاربردهای عملی، می‌توان از روش پلاس در تحلیل داده‌های پزشکی، تشخیص بیماری‌های چندگانه و سایر حوزه‌های داده‌کاوی و یادگیری ماشین استفاده کرد.

مراجع

محمدزاده، م.، (۱۳۹۸)، *آمار فضایی و کاربردهای آن*، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،

Dettling, M., and Bühlmann, P. (2002), Supervised Clustering of Genes, *Genome Biology*, **3**, 0069.1–0069.15.

Dettling, M., and Bühlmann, P. (2004), Finding Predictive Gene Groups from Microarray Data, *Journal of Multivariate Analysis*, **90**, 106-131.

Pitas, I., and Venetsanopoulos, A. N. (1992), Order Statistics in Digital Image Processing, *Proceedings of the IEEE*, **80**, 1893-1921.

Tibshirani, R., Walther, G., and Hastie, T. (2001), Estimating the Number of Clusters in a Data Set via the Gap Statistic, *Journal of the Royal Statistical Society*, B, (Statistical Methodology), **63**, 411-423.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

رویکرد اسپلاین‌های آمیخته کروی جهت تحلیل داده‌های شبه کروی

محمد امین بدیعی، علی محمدیان مصمم[۱]
گروه آمار، دانشگاه زنجان

چکیده: در آمار و زمینه تحلیل داده‌ها، اسپلاین‌های کروی یک رویکرد پیشرو برای مدل‌سازی و تحلیل داده‌های متنوع از جمله داده‌های مرتبط با هواشناسی، نجوم، اپیدمیولوژی و حوزه‌های دیگر است. در این روش، مبحث اسپلاین از فضای مسطح اقلیدسی به فضای کروی تعمیم یافته و با استفاده از انواع مختلف هسته‌ها و فواصل کروی، تلاش می‌کند تا بهترین مدل‌های تطابق داده با کمترین میزان ریشه میانگین مربعات خطا را ایجاد کند. تحقیقات در این حوزه باعث بهبود دقت و اعتبار تحلیل‌های آماری در زمینه‌های مختلف می‌شود و درک بهتری از رفتار داده‌ها و پدیده‌های مورد مطالعه فراهم می‌کند. مزیت اصلی اسپلاین‌های کروی نسبت به مدل‌های مبتنی بر تخمین سطوح کروی با استفاده از اسپلاین های مبتنی بر فضای مسطح اقلیدسی این است که در نتیجه عدم تخمین سطوح کروی با استفاده از صفحات مسطح اقلیدوسی، اطلاعات از بین نرفته و خطای مربوطه از مدل حذف می‌شود. برای تحلیل و مدل‌سازی داده‌های مربوط به سطوح شبه کروی، رویکرد اصلی به صورت تخمین این سطوح با استفاده از سطح کروی است که همچون رویکرد تخمین سطح کره با استفاده از سطوح مسطح اقلیدوسی موجب کاهش دقت و ایجاد خطا در مدل سازی می‌شود. در این مقاله رویکرد نوینی جهت تحلیل و مدل‌سازی داده‌های مربوط به سطوح شبه کروی ارائه شده است.

واژه‌های کلیدی: درون‌یابی اسپلاین، اسپلاین کروی، فضای هیلبرت با هسته بازآفرین.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62M30، 62H11.

۱ مقدمه

با پیشرفت فناوری و جمع‌آوری داده‌ها از طریق ماهواره‌ها از روی سطوح کروی در انواع حوزه‌های علمی، نیاز به روش‌های کارآمدتر برای مدل‌سازی و تحلیل داده‌ها احساس می‌شود. یکی از این روش‌ها برای تحلیل داده‌های استخراج شده از سطوح کروی، استفاده از اسپلاین‌های کروی است.

اسپلاین‌های کروی به عنوان یک مدل آماری کارآمد، در تحلیل داده‌ها و مدل‌سازی فضایی برای داده‌های متنوع استفاده می‌شوند. یکی از ویژگی‌های برجسته اسپلاین‌های کروی توانایی تعمیم بهتر اطلاعات به فضای کروی است که به محققان اجازه می‌دهد به نتایج دقیق‌تری در زمینه‌های مرتبط با مختصات سه‌بعدی یا سطح کروی دست یابند. در این

[۱] نام و ایمیل ارائه دهنده مقاله: علی محمدیان مصمم a.m.mosammam@znu.ac.ir

زمینه پژوهش‌های مختلفی گسترش یافته است که از جمله آن‌ها می توان به فرانک و نیلسون (۱۹۸۰)، وابا (۱۹۸۱) و کلر و بورکوسکی (۲۰۱۹) اشاره کرد. اما همان گونه که تخمین مدل‌های مربوط به سطوح کروی با استفاده از مدل‌های مربوط به سطوح مسطح اقلیدوسی موجب از دست رفتن اطلاعات می‌شود، مدل سازی برای سطوح شبه کروی با در نظر گرفتن آن به عنوان سطوح کروی نیز موجب از دست رفتن اطلاعات و افزایش خطا در تحلیل داده‌های استخراج شده از سطوح شبه کروی خواهد شد. به این منظور در این مقاله، در بخش اول به بررسی اسپلاین روی سطوح مسطح اقلیدوسی، در بخش دوم به بررسی اسپلاین‌های کروی، در بخش سوم به بررسی رویکرد اسپلاین آمیخته جهت مدل سازی اسپلاین برای سطوح شبه کروی و در بخش چهارم نتایج مربوط به شبیه سازی آمده است.

۲ اسپلاین روی فضای مسطح اقلیدسی

برای مجموعه داده‌های $(z_i, x_i, y_i)\,; i = ۱, ..., n.$ که در آن x_i, y_i موقعیت‌های داده‌ها و z_i داده‌های حقیقی مقدار می‌باشد، هدف پیدا کردن تابع هموار f به طوری است که:

$$\sum_{i=۱}^{n} (z_i - f(x_i, y_i))^۲ + \alpha \parallel f \parallel^۲, \tag{۱.۲}$$

مینیمم گردد که در آن:

$$\|f\|^۲ = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left[\left(\frac{\partial^۲ f}{\partial x^۲}\right)^۲ + ۲\left(\frac{\partial f}{\partial x \partial y}\right)^۲ + \left(\frac{\partial^۲ f}{\partial y^۲}\right)^۲ \right] dxdy,$$

است. کلر و بورکوسکی (۲۰۱۹) نشان دادند که که جواب مسئله فوق به صورت:

$$f\,(x, y) = \sum_{i=۱}^{n} \lambda_i r_i^۲ \ln r_i + d_{۰۰} + d_{۰۱} x + d_{۰۱} y,$$

است که به آن اسپلاین صفحه نازک [۱] گفته می‌شود. در این حالت، هسته مربوطه به صورت:

$$K\,(x_i, y_j) = r_{ij}^۲ \ln r_{ij},$$

است که در آن r_{ij} ، فاصله اقلیدسی بین موقعیت داده‌ها می‌باشد. در حالت کلی جواب مسئله اسپلاین به صورت:

$$f\,(x, y) = \sum_{i=۱}^{n} \alpha_i K\,(x_i, y_i)\,,$$

است که در آن $K\,(x_i, y_i)$ هسته مورد نظر می‌باشد.

۳ اسپلاین کروی

فرض کنید $C_{\cdot}^{\infty}\,(S)$ ، مجموعه همه توابع نامتناهی هموار φ بر روی کره با مقدار میانگین برابر با صفر باشد. بدیهی است که:

$$\langle f, g \rangle := \int_S \Delta_s f \cdot \Delta_S g dS, \tag{۱.۳}$$

[1] Thin plate spline

یک ضرب اسکالر در $C^{\infty}(S)$ است که در آن Δ_s عملگر لاپلاس ـ بلترامی [۲] به صورت:

$$\Delta_S = \frac{1}{\sin\vartheta}\frac{\partial}{\partial\vartheta}\left(\sin\vartheta\frac{\partial}{\partial\vartheta}\right) + \frac{1}{\sin^2\vartheta}\frac{\partial^2}{\partial\lambda^2},$$

است. (ϑ_i, λ_i) ، موقعیت داده‌های z_i روی کره می‌باشد. با توجه به (۱.۳) مربع نرم [۳] به صورت:

$$\|f\|^2 = \int_S (\Delta_s f)^2 dS, \tag{۲.۳}$$

تعریف می‌شود و فضای سوبولوف [۴] :

$$H^{2,2}(S) = \{f \in C^{\infty}(S); \|f\| < \infty\},$$

با نرم (۲.۳) ، تشکیل یک فضای هیلبرت با هسته بازآفرین می‌دهد وابا (۱۹۹۰) که شامل هسته بازآفرین [۵] زیر است (کلر و بورکوسکی ، ۲۰۱۹) :

$$\begin{aligned} K(\vartheta_1, \lambda_1; \vartheta_2, \lambda_2) &= \sum_{\ell=1}^{\infty}\sum_{m=-\ell}^{\ell} Z_{\ell,m}(\vartheta_1, \lambda_1)\overline{Z_{\ell,m}(\vartheta_2, \lambda_2)} \\ &= \sum_{\ell=1}^{\infty} \frac{2\ell+1}{\ell^2(\ell+1)^2} P_\ell(\cos\psi), \end{aligned} \tag{۳.۳}$$

اگر هارمونیک‌های کروی سطح کاملا نرمال شده [۶] را با $Y_{\ell,m}$ نشان دهیم، آنگاه تابع $Z_{\ell,m}$ یک مجموعه متعامد یکه [۷] در $H^{2,2}(S)$ بوده و $Z_{\ell,m}$ ، به صورت زیر تعریف می‌شود:

$$Z_{\ell,m} = \frac{1}{\ell(\ell+1)} Y_{l,m}.$$

همچنین ψ ، زاویه کروی بین موقعیت‌ها را نشان می‌دهد:

$$\cos\psi = \cos\vartheta_1\cos\vartheta_2 + \sin\vartheta_1\sin\vartheta_2\cos(\lambda_1 - \lambda_2).$$

لذا همانند اسپلاین روی فضای اقلیدسی، باتوجه به (۱.۲) و (۳.۳) برای اسپلاین کروی، هدف، یافتن تابع هموار f به طوری است که:

$$\sum_{i=1}^{n}(z_i - f(\xi_i, \xi_j))^2 + \alpha\|f\|^2, \tag{۴.۳}$$

را مینیمم نماید. جواب آن را می‌توان به صورت زیر نوشت (وابا ، ۱۹۸۱) :

$$f = \sum_{i=1}^{n}\alpha_i K(\xi_i, \xi_j).$$

در ادامه به معرفی و بررسی رویکرد اسپلاین آمیخته جهت مدل سازی اسپلاین برای سطوح شبه کروی پرداخته شده است.

[2] Laplace–Beltrami operator
[3] Norm
[4] Sobolev space
[5] Reproducing kernel
[6] Fully normalized surface spherical harmonics
[7] Orthonormal

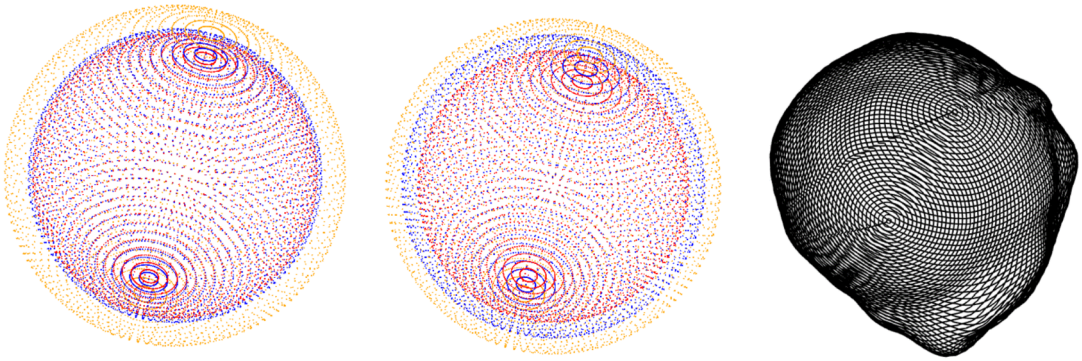
۴ رویکرد اسپلاین آمیخته کروی

در رویکرد اسپلاین‌های کروی آمیخته برای تحلیل داده‌های مربوط به سطوح شبه کروی، سطح شبه کره توسط چند کره مختلف تقریب زده می‌شود. علت تخمین سطح شبه کروی توسط چند کره به منظور ایجاد مدل‌های مختلف برای موقعیت‌هایی با فاصله‌های متفاوت نسبت به مرکز شبه کره است. فرض کنید برای داده‌های $(z_i, \varphi_i, \theta_i, \rho_i)\,; i = ۱, ..., n.$ که در آن $\varphi_i, \theta_i, \rho_i$ مختصات کروی هستند ، سطحی شبه کروی با کمینه شعاع $r_{\min}$ و بیشینه شعاع $r_{\max}$ که در آن $r_{\min} < \rho_i \leq r_{\max}; i = ۱, ..., n$ است وجود دارد، قصد داریم شبه کره مذکور را توسط m کره با شعاع‌های $r_i;\ i = ۱, ..., m$ به‌طوری که $r_{\min} < r_j \leq r_{\max};\ i = ۱, ..., m$ است، تقریب بزنیم.در این‌صورت داریم:

$$\sum_{j=۱}^{m} \alpha_j \sum_{i=۱}^{n} (z_i - f_j(\phi_i, \theta_i))^۲ + \lambda_j \|f_j\|^۲.$$

که در آن $a_j = ۱$ است اگر و فقط اگر $r_j < \rho_i \leq r_{j+۱}$ و در غیر این صورت $a_j = ۰$ است. تعداد m تا زمانی افزایش داده می‌شود که تغییر قابل ملاحظه‌ای در نتایج بدست آمده ایجاد شده باشد. برای این منظور از دو روش تقسیم اختلاف بیشینه و کمینه شعاع شبه کره به m کره با فواصل یکسان و یا تقسیم به صورتی که هر کدام از این فواصل شامل مقدار برابری از مساحت سطح شبه کره باشد استفاده کرد. در شکل ۱ تخمین سطح شبه کروی با سه کره نمایش داده شده است.

لازم به ذکر است استفاده از این رویکرد زمانی دقت و کارایی مورد انتظار را در بر می‌گیرد که میان تغییر فاصله مکان داده‌ها از مرکز شبه کره و اندازه داده‌های مورد بررسی همبستگی وجود داشته باشد.



آ) سطح شبه کروی

ب) تقسیم بیشینه و کمینه شعاع سطح شبه کروی به کره‌هایی با فواصل یکسان

ج) تقسیم بیشینه و کمینه شعاع سطح شبه کروی به کره‌های شامل تعداد مساوی از موقعیت‌ها

شکل ۱: تخمین سطح شبه کروی با سه کره.

۵ مقایسه عملکرد رویکرد اسپلاین آمیخته کروی و اسپلاین کروی

در این بخش به مقایسه عملکرد رویکرد اسپلاین آمیخته کروی و اسپلاین کروی برای تحلیل داده‌های مربوط به سطح شبه کروی پرداخته شده است. برای انجام مقایسه مورد نظر، تعداد ۱۶۲۰۰ داده مربوط به یک شهاب سنگ شبیه‌سازی شده است که در آن چهار متغییر اندازه‌گیری شده و اطلاعات مربوط به آن در جدول ۱ آمده است. در این آزمایش ۱۰ درصد از تعداد کل داده‌ها حذف و با استفاده از اسپلاین کروی و رویکرد اسپلاین کروی آمیخته دوگانه و رویکرد اسپلاین کروی آمیخته سه‌گانه و همچنین با استفاده از ۷۸ گره برای ایجاد مدل‌ها، پیش بینی صورت گرفته است. در انتها همبستگی و

میانگین مربعات خطای داده‌های پیش‌بینی شده و داده‌های اولیه با استفاده از روش‌های مذکور محاسبه و نتایج آن در جدول ۲ نشان داده شده است.

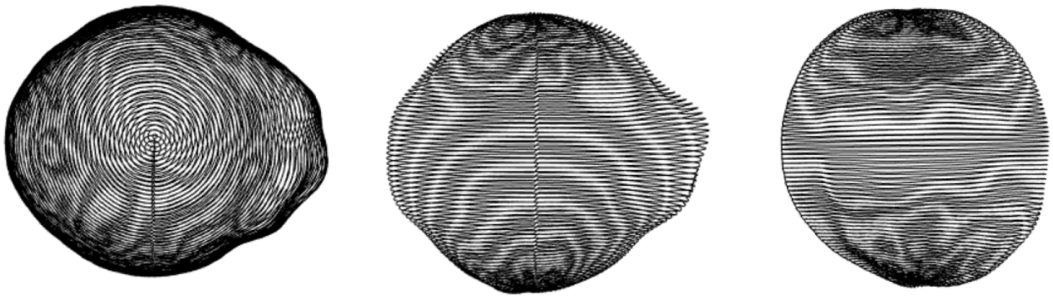
همانگونه که مشاهده می‌شود در مورد متغییرهایی که اندازه همبستگی خطی بزرگتری نسبت به تغییر شعاع شهاب‌سنگ دارند، استفاده از مدل‌های آمیخته جهت پیشبینی داده‌های حذف شده موجب کاهش قابل توجه میانگین مربعات خطا و همچنین افزایش همبستگی میان داده‌های پیشبینی شده با مقادیر حذف شده آنها شده است.

جدول ۱: عملکرد اسپلاین کروی و اسپلاین کروی آمیخته جهت تحلیل داده‌های مربوط به سطح شبه کروی

شعاع (واحد)			همبستگی متغیرها و فاصله از مرکز			
میانگین	بیشینه	کمینه	متغیر اول	متغیر دوم	متغیر سوم	متغیر چهارم
۱۶۰.۵۲	۲۰۱.۸۴	۱۴۹.۶۹	۰.۶۰۹	۰.۰۱۶	ـ۰.۶۰۱	۰.۸۶۴

جدول ۲: مقایسه عملکرد اسپلاین کروی و اسپلاین کروی آمیخته جهت تحلیل داده‌های سطح شبه کروی

متغیر	اسپلاین کروی	اسپلاین آمیخته دوگانه	اسپلاین آمیخته سه‌گانه	معیار
اول	۰.۱۹۷۳	۰.۶۳۳۰	۰.۶۴۳۶	COR
	۲۲۸.۴۷	۱۶۸.۱۴	۱۵۹.۰۳	MSE
دوم	ـ۰.۰۶۳۰	۰.۰۱۰۱	ـ۰.۰۰۵۲	COR
	۴۲۳.۱۹۸	۴۲۲.۹۲	۴۲۳.۰۲	MSE
سوم	۰.۲۱۲۷	۰.۷۲۶۵	۰.۷۳۹۶	COR
	۱۸۸.۶۱	۱۱۸.۶۴	۱۱۷.۵۰	MSE
چهارم	۰.۲۹۱۸	۰.۸۵۸۴	۰.۸۹۱۹	COR
	۱۰۸۱.۳۳	۴۹۲.۴۹	۴۵۴.۳۵	MSE



آ) نمای جانبی ب) نمای جانبی ج) نمای از بالا

شکل ۲: شهاب‌سنگ شبیه سازی شده جهت آزمایش مذکور.

بحث و نتیجه‌گیری

در این مقاله، رویکرد استفاده از اسپلاین‌های کروی آمیخته جهت تحلیل داده‌های مربوط به سطوح شبه کروی معرفی شد که با توجه به نتایج حاصله در بخش قبل، مشاهده می‌شود که در مواردی که همبستگی میان مولفه‌ای همچون شعاع شکل شبه کروی و متغییر مورد بررسی وجود دارد، استفاده از این رویکرد موجب کاهش خطا و افزایش کارایی و دقت مدل می‌شود. لازم به ذکر است که این رویکرد، روش جدید در استفاده از اسپلاین‌های کروی می‌باشد.

مراجع

Franke, R., and Nielson, G. (1980), Smooth Interpolation of Large Sets of Scattered Data, *International Journal for Numerical Methods in Engineering*, **15**(11), 1691-1704.

Keller, W., and Borkowski, A. (2019), Thin Plate Spline Interpolation, *Journal of Geodesy*, **93**(9), 1251-1269.

Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics.

Wahba, G. (1981). Spline Interpolation and Smoothing on the Sphere, *SIAM Journal on Scientific and Statistical Computing*, **2**(1), 5-16.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

خوشه‌بندی سلسله‌مراتبی داده‌های زمین‌آماری

سیده سمیه موسوی[۱]، عادل محمدپور[۱]، اسحاق الماسی[۲]
[۱] گروه آمار، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)
[۲] گروه آمار، دانشگاه رازی

چکیده: این مقاله بر روی خوشه‌بندی فضایی داده‌های زمین‌آماری مدل‌بندی‌شده با میدان‌های تصادفی αـ پایدار زیرگاوسی تمرکز می‌کند. برای این منظور، یک معیار قرابت بر اساس تابع درستنمایی در یک الگوریتم خوشه‌بندی سلسله‌مراتبی معرفی شده‌است که ویژگی‌های فضایی و غیرفضایی و همچنین ساختار فضایی داده‌ها را درنظر می‌گیرد. ایده‌ی روش پیشنهادی به حداکثر رساندن معیار قرابت در انتخاب جفت خوشه‌ها برای ادغام در هر مرحله است. لذا محدودیت تعریف فاصله‌ی دو مجموعه از نقاط در فرایندهای خوشه‌بندی را ندارد. برای ارزیابی روش پیشنهادی، الگوریتمی برای شبیه‌سازی داده‌های زمین‌آماری از میدان تصادفی αـ پایدار زیرگاوسی ارائه می‌کنیم. سپس به بررسی کارایی این روش با استفاده از داده‌های زمین‌آماری شبیه‌سازی‌شده می‌پردازیم.

واژه‌های کلیدی: داده‌های زمین‌آماری، میدان تصادفی αـ پایدار زیرگاوسی، خوشه‌بندی فضایی، معیار قرابت
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 60E07.

۱ مقدمه

در دهه‌های اخیر، شناسایی مناطق همگن یا خوشه‌ها در داده‌های زمین‌آماری به‌طور گسترده مورد بررسی قرار گرفته‌است. در این راستا، خوشه‌بندی فضایی به عنوان ابزار مفیدی برای گروه‌بندی این داده‌ها براساس معیارهای مختلف قرابت فضایی ایجاد شده‌است. معیارهای قرابت (تشابه و عدم‌تشابه) نقش مهمی در فرآیندهای خوشه‌بندی دارند. اکثر الگوریتم‌های خوشه‌بندی استاندارد بر پایه‌ی مینیمم‌کردن عدم‌تشابه بین همه‌ی جفت مشاهدات توسط توابع فاصله عمل می‌کنند. با توجه به همبستگی فضایی داده‌های فضایی، باید معیاری را اعمال کنیم که این ویژگی ذاتی داده‌ها را درنظر بگیرد. از سوی دیگر، یکی از راه‌های خوشه‌بندی داده‌های فضایی، تعریف یک معیار قرابت فضایی در روش‌های خوشه‌بندی استاندارد و کلاسیک است، به‌طوری‌که این معیار ساختار فضایی بین داده‌ها را نیز در نظر بگیرد. کربی (۲۰۰۹) روشی را برای اعمال ساختار فضایی مشاهدات در الگوریتم خوشه‌بندی با به حداکثر رساندن تابع درستنمایی توزیع گاوسی چند متغیره در فرآیند

[۱] نام و ایمیل ارائه دهنده مقاله: سیده سمیه موسوی s.s.mousavi.stat@gmail.com

خوشه‌بندی به صورت سلسله‌مراتبی پیشنهاد کرد. او نشان داد که تابع درستنمایی توزیع‌ها می‌تواند یکی از بهترین معیارهای تشابه برای خوشه‌بندی داده‌های زمین‌آماری باشد، زیرا ساختار فضایی داده‌ها را از طریق ماتریس کوواریانس در نظر می‌گیرد. موسوی (۱۴۰۲) از این ایده استفاده کرده و یک معیار قرابت جدید بر اساس تابع درستنمایی برای یافتن خوشه‌ها در یک فرآیند سلسله‌مراتبی تجمیعی پیشنهاد می‌دهد.

در بسیاری از کاربردها، یک میدان تصادفی گاوسی برای مدل‌بندی داده‌های فضایی انتخاب می‌شود (محمدزاده ، ۱۳۹۸). با این حال، پدیده‌های طبیعی مانند مبالغ خسارت بیمه‌ی طوفان، ثروت اقتصادی، تراکم جمعیت، میزان بارندگی و بافت خاک، ممکن است شامل مقادیر کرانگین یا نقاط دورافتاده باشد که دم‌های سنگینی را در توزیع‌شان نشان می‌دهند. در چنین مواردی، از توزیع‌هایی استفاده می‌شود که در مقایسه با توزیع گاوسی دم‌های سنگین‌تری دارند و نسبت به نقاط دورافتاده استوار هستند. در میان این توزیع‌ها می‌توان به توزیع αـ پایدار زیرگاوسی ($SG\alpha S$) از خانواده‌ی توزیع‌های پایدار، اشاره کرد که می‌تواند انتخاب مناسب‌تری نسبت به توزیع گاوسی برای داده‌های فضایی دم‌سنگین باشد.

ادامه این مقاله به چهار بخش تقسیم شده است. بخش ۲ شامل تعاریف مورد نیاز میدان تصادفی $SG\alpha S$ برای مدل‌بندی داده‌های زمین‌آماری است. بخش ۳ یک معیار قرابت جدید را با استفاده از توابع درستنمایی معرفی می‌کند و یک الگوریتم خوشه‌بندی را بر اساس این معیار ارائه می‌کند. در بخش ۴، عملکرد الگوریتم پیشنهادی و معیار قرابت را برای خوشه‌بندی داده‌های زمین‌آماری شبیه‌سازی‌شده از میدان‌های تصادفی مانای $SG\alpha S$ ارزیابی می‌کنیم. در بخش ۵ به نتیجه‌گیری می‌پردازیم.

۲ مدل‌بندی داده‌های زمین‌آماری از K میدان تصادفی SGαS

توزیع‌های αـ پایدار یک رده‌ی وسیع از توزیع‌های احتمالاتی دم‌سنگین هستند. این خانواده از توزیع‌ها با نماد $S(\alpha,\beta,\gamma,\delta)$ یا $S_\alpha(\gamma,\beta,\delta)$ نمایش داده می‌شوند که در آن $\alpha\in(0,2]$ شاخص پایداری، $\beta\in[-1,1]$ پارامتر چولگی، $\gamma\in(0,\infty)$ پارامتر مقیاس و $\delta\in\mathbb{R}$ پارامتر مکان، چهار پارامتر توزیع αـ پایدار هستند.

تعریف ۱.۲. (*میدان تصادفی فضایی $SG\alpha S$، اسپودارف (۲۰۱۵)*). *فرض کنید $X\sim S_{\frac{\alpha}{2}}\left(\left(\cos\frac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}},1,0\right)$ یک توزیع پایدار چوله به راست، یعنی پایدار مثبت است و $\{Z(\boldsymbol{s});\ \boldsymbol{s}\in\mathbb{D}\subset\mathbb{R}^d\}$ یک میدان تصادفی گاوسی با میانگین صفر و ماتریس کوواریانس $\boldsymbol{\Sigma}$، مستقل از X باشد. در این‌صورت $Y(\boldsymbol{s})\overset{D}{=}X^{\frac{1}{2}}Z(\boldsymbol{s})+\delta$ دارای یک میدان تصادفی $SG\alpha S$ با پارامتر بردار مکان $\boldsymbol{\delta}$ و ماتریس پراکندگی $\boldsymbol{\Sigma}$ است.*

تعریف ۲.۲. (*تابع چگالی $SG\alpha S$، نولان (۲۰۱۳)*). *فرض کنید $\boldsymbol{Y}\overset{D}{=}X^{\frac{1}{2}}\boldsymbol{Z}+\boldsymbol{\delta}$ یک بردار تصادفی n بعدی $SG\alpha S$ با شاخص پایداری $\frac{\alpha}{2}$، بردار مکان $\boldsymbol{\delta}$ و ماتریس پراکندگی $\boldsymbol{\Sigma}$ باشد. بردار تصادفی $\boldsymbol{Z}$ را می‌توان به صورت $\boldsymbol{Z}\overset{D}{=}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}_1$ نیز بیان کرد که در آن $\boldsymbol{\Sigma}^{\frac{1}{2}}$ تجزیه‌ی چولسکی ماتریس $\boldsymbol{\Sigma}$ و $\boldsymbol{Z}_1\sim N_n(\mathbf{0},\mathbf{I})$. بنابراین رابطه‌ی فوق را می‌توان به صورت زیر بازنویسی کرد:*

$$\boldsymbol{Y}\overset{D}{=}X^{\frac{1}{2}}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}_1\right)+\boldsymbol{\delta}=\boldsymbol{\Sigma}^{\frac{1}{2}}\left(X^{\frac{1}{2}}\boldsymbol{Z}_1\right)+\boldsymbol{\delta}=\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Y}_1+\boldsymbol{\delta},$$

که در آن $\boldsymbol{Y}_1$ یک بردار تصادفی n بعدی $SG\alpha S$ با بردار مکان $\mathbf{0}$ ماتریس پراکندگی $\mathbf{I}_{n\times n}$ است. نولان (۲۰۱۳) تابع چگالی بردار مشاهده $\boldsymbol{y}$ را به‌صورت زیر نشان داد:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\alpha,\boldsymbol{\delta},\boldsymbol{\Sigma})=|\boldsymbol{\Sigma}|^{-\frac{1}{2}}f_{\boldsymbol{Y}_1}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\delta})\right)=|\boldsymbol{\Sigma}|^{-\frac{1}{2}}H_{\alpha,n}\left(\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\delta})\right\|\right)\qquad(۱.۲)$$

که در آن $|\boldsymbol{\Sigma}|$ بیانگر دترمینان ماتریس $\boldsymbol{\Sigma}$ و $H_{\alpha,n}(r)$ تابع شعاعی است که به‌صورت زیر بیان می‌شود:

$$H_{\alpha,n}(r)=H(r\mid\alpha,n)=\begin{cases}\dfrac{\Gamma(\frac{n}{2})}{2\pi^{\frac{n}{2}}}r^{1-n}f_R\left(r\mid\alpha,\gamma_0=1,n\right); & r>0\\ \dfrac{\Gamma(\frac{n}{\alpha})}{\alpha 2^{n-1}\pi^{\frac{n}{2}}\Gamma(\frac{n}{2})^2}; & r=0\end{cases},$$

$$f_R\left(r\mid\alpha,\gamma_0,n\right)=\frac{2}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\int_0^{\infty}(rt)^{\frac{n}{2}}J_{\frac{n}{2}-1}(rt)e^{-(\gamma_0 t)^{\alpha}}dt.$$

که در آن $\Gamma(\cdot)$ و $J_\nu(\cdot)$ به‌ترتیب توابع گاما و بسل نوع اول از درجه ν هستند.

در ادامه‌ی این بخش دو حالت برای تخصیص داده‌های فضایی دارای توزیع SGαS به K گروه را بررسی می‌کنیم:

۱ـ فرض کنید $\boldsymbol{Y}_k\overset{D}{=}X_k^{\frac{1}{2}}\boldsymbol{Z}_k+\boldsymbol{\delta}_k$، $k=1,\ldots,K$، گروه‌هایی از یک میدان تصادفی فضایی مانای SGαS در $n=\sum_{k=1}^{K}n_k$ موقعیت مکانی باشند که X_kها بیانگر متغیرهای تصادفی پایدار مثبت با پارامترهای پایداری یکسان $\frac{\alpha}{2}$ و $\boldsymbol{Z}_k\sim N_{n_k}\left(\boldsymbol{0},\boldsymbol{\Sigma}_k\right)$ مستقل از X_kها باشند. امکان دارد اختلاف بین گروه‌ها تنها به اختلاف بردارهای مکان گروه‌ها مرتبط باشد و $\boldsymbol{\delta}_{n\times 1}=(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_K)'$، که $\boldsymbol{\delta}_k$ بردار مکان kامین گروه باشد. در صورتی‌که ماتریس پراکندگی به صورت کلی $\boldsymbol{\Sigma}_{n\times n}$ باشد، تابع چگالی توأم مشاهدات K گروه به صورت رابطه‌ی (۱.۲) است. در حالت خاص که ماتریس پراکندگی به فرم بلوکی $\boldsymbol{\Sigma}_{n\times n}=\oplus_{k=1}^{K}\boldsymbol{\Sigma}_k=\mathrm{diag}(\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K)$ نوشته شود که $\boldsymbol{\Sigma}_k$ ماتریس پراکندگی kامین گروه با اندازه‌ی n_k است، تابع چگالی $\boldsymbol{y}=(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_K)'$ به صورت زیر ساده می‌شود:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\alpha,\boldsymbol{\delta},\boldsymbol{\Sigma})=(|\boldsymbol{\Sigma}_1|\cdots|\boldsymbol{\Sigma}_K|)^{-\frac{1}{2}}H_{\alpha,n}(\sqrt{\|\boldsymbol{\Sigma}_1^{-\frac{1}{2}}(\boldsymbol{y}_1-\boldsymbol{\delta}_1)\|+\cdots+\|\boldsymbol{\Sigma}_K^{-\frac{1}{2}}(\boldsymbol{y}_K-\boldsymbol{\delta}_K)\|}) \quad (۲.۲)$$

۲ـ اکنون فرض کنید $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K$ بردارهای تصادفی SGαS فضایی مستقل از هم با پامترهای مکان $\boldsymbol{\delta}_k\in\mathbb{R}^{n_k}$ باشند و X_kها دارای توزیع پایدار مثبت با شاخص‌های پایداری متفاوت $\frac{\alpha_k}{2}$ و $\boldsymbol{X}_k\sim N_{n_k}\left(\boldsymbol{0},\boldsymbol{\Sigma}_k\right)$. بنابراین با فرض اینکه $f_{\boldsymbol{Y}_k}\left(\boldsymbol{y}_k|\alpha_k,\boldsymbol{\delta}_k,\boldsymbol{\Sigma}_k\right)$ تابع چگالی توأم گروه kام است، تابع چگالی توأم کل گروه‌ها به‌صورت زیر نوشته می‌شود:

$$f_{\boldsymbol{Y}}(\boldsymbol{y})=f_{\boldsymbol{Y}_1}\left(\boldsymbol{y}_1|\alpha_1,\boldsymbol{\delta}_1,\boldsymbol{\Sigma}_1\right)\times\cdots\times f_{\boldsymbol{Y}_K}\left(\boldsymbol{y}_K|\alpha_K,\boldsymbol{\delta}_K,\boldsymbol{\Sigma}_K\right) \quad (۳.۲)$$

۳ روش خوشه‌بندی

شاخص قرابت به‌عنوان یک جزء اصلی در فرآیندهای خوشه‌بندی، نزدیکی اشیاء را اندازه‌گیری می‌کند. روش‌های خوشه‌بندی فضایی تلاش می‌کنند تا مشاهداتی را که هم در ویژگی‌های فضایی و هم غیرفضایی بیشترین نزدیکی را دارند شناسایی کرده و سپس آنها را در درون خوشه‌ها دسته‌بندی کنند. در این بخش، یک معیار قرابت جدید با استفاده از تابع درستنمایی و یک الگوریتم خوشه‌بندی فضایی برای گروه‌بندی داده‌های زمین‌آماری دارای توزیع SGαS معرفی می‌کنیم. بخش‌های زیر جزئیات معیار پیشنهادی و الگوریتم خوشه‌بندی را ارائه می‌کنند.

معیار قرابت بر اساس درستنمایی: فرض کنید $\boldsymbol{y}=\left(y\left(\boldsymbol{s}_1\right),\ldots,y\left(\boldsymbol{s}_n\right)\right)'$ داده‌های زمین‌آماری روندزدوده‌شده در n موقعیت فضایی مشخص باشند. تابع درستنمایی آن‌ها به‌صورت زیر تعریف می‌شود:

$$L\left(\boldsymbol{\theta},\boldsymbol{\Sigma}|\boldsymbol{y}\right)=f_{\boldsymbol{Y}}\left(y\left(\boldsymbol{s}_1\right),\ldots,y\left(\boldsymbol{s}_n\right);\boldsymbol{\theta},\boldsymbol{\Sigma}\right),$$

که در آن، $\boldsymbol{\theta}$ و $\boldsymbol{\Sigma}$ به ترتیب بردار پارامتر و ماتریس پراکندگی (یا کوواریانس) هستند. کربی (۲۰۰۹) روشی را با ترکیب ساختار فضایی مشاهدات در الگوریتم خوشه‌بندی سلسله‌مراتبی و با درنظر گرفتن بزرگترین مقدار تابع درستنمایی (LLF) توزیع گاوسی چند متغیره در فرآیند خوشه‌بندی پیشنهاد کرد. با فرض اینکه $\boldsymbol{y}_i = \left(y(\boldsymbol{s}_1^i), \ldots, y(\boldsymbol{s}_{n_i}^i)\right)'$ و $\boldsymbol{y}_j = \left(y(\boldsymbol{s}_1^j), \ldots, y(\boldsymbol{s}_{n_j}^j)\right)'$ دو زیرمجموعه‌ی مجزا از مجموعه داده‌های $\boldsymbol{y}$ هستند، معیار قرابت درستنمایی به صورت زیر تعریف می‌شود:

$$L(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\boldsymbol{y}_i, \boldsymbol{y}_j) = f_{\boldsymbol{Y_i}, \boldsymbol{Y_j}}\left(\boldsymbol{y}_i, \boldsymbol{y}_j; \boldsymbol{\theta}, \boldsymbol{\Sigma}\right)$$

بنابراین، ما این معیار را توسعه داده و یک معیار قرابت جدید بر اساس تابع درستنمایی برای اندازه‌گیری شباهت بین مشاهدات به صورت زیر تعریف کرده‌ایم:

$$PML(\boldsymbol{y}_i, \boldsymbol{y}_j) = \frac{n_i n_j}{(n_i + n_j)} \frac{L(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\boldsymbol{y}_i, \boldsymbol{y}_j)}{L(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\boldsymbol{y}_i) L(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\boldsymbol{y}_j)} \tag{۱.۳}$$

که در آن، n_i و n_j به ترتیب طول بردار مشاهدات $\boldsymbol{y}_i$ و $\boldsymbol{y}_j$ هستند. شایان ذکر است که با نزدیک شدن دو مشاهده، درستنمایی آن‌ها افزایش می‌یابد که شباهت بین دو مشاهده را ایجاد می‌کند.

ایده‌ی روش پیشنهادی به حداکثر رساندن معیار قرابت در انتخاب جفت خوشه‌ها برای ادغام در هر مرحله است. روش جدید یک ویژگی منحصر به‌فرد نسبت به خوشه‌بندی سلسله‌مراتبی کلاسیک بر حسب معیار عدم‌تشابه دارد. در این روش به جای تابع فاصله از تابع متناظر با چگالی نقاط به‌عنوان معیار قرابت استفاده می‌شود که منفی لگاریتم آن، یک تابع زیان مناسب برای برآورد پارامتر مکان است.

روش اتصال خوشه‌ها: از آنجایی که یک روش سلسله‌مراتبی تجمیعی توسط یک ملاک اتصال دو خوشه را در هر مرحله ترکیب می‌کند، شباهت بین دو خوشه باید اندازه‌گیری شود. بنابراین، ما یک ملاک اتصال جدید با استفاده از معیار قرابت PML تعریف کرده‌ایم که در آن شباهت بین دو خوشه با ماکسیمم شباهت بین اعضای دو خوشه اندازه‌گیری می‌شود. با فرض اینکه $\mathcal{C}_i$ و $\mathcal{C}_j$ دو خوشه باشند که به ترتیب شامل بردار مشاهدات $\boldsymbol{y}_i = \left(y(\boldsymbol{s}_1^i), \ldots, y(\boldsymbol{s}_{n_i}^i)\right)'$ و $\boldsymbol{y}_j = \left(y(\boldsymbol{s}_1^j), \ldots, y(\boldsymbol{s}_{n_j}^j)\right)'$ هستند، شباهت بین آن‌ها بر اساس معیار PML به‌صورت زیر محاسبه می‌شود:

$$S(\mathcal{C}_i, \mathcal{C}_j) = \max_{r,l} \left\{ PML\left(y(\boldsymbol{s}_r^i), y(\boldsymbol{s}_l^j)\right);\ \ y(\boldsymbol{s}_r^i) \in \mathcal{C}_i,\ y(\boldsymbol{s}_l^j) \in \mathcal{C}_j \right\} \tag{۲.۳}$$

۱.۳ الگوریتم خوشه‌بندی با استفاده از معیار قرابت PML

برای استفاده از معیار قرابت PML، ابتدا باید ساختار همبستگی فضایی داده‌های روندزدوده شده را مشخص کرده و پارامترها برآورد شوند. سپس الگوریتم خوشه‌بندی پیشنهادی به صورت زیر پیاده‌سازی می‌شود:

۱ـ فرایند با n خوشه شروع می‌شود، یعنی هر مشاهده به‌عنوان یک خوشه درنظر گرفته می‌شود. سپس مقدار معیار PML برای هر یک از $\binom{n}{2}$ جفت از خوشه‌ها محاسبه می‌شود.

۲ـ جفت دارای بیشترین مقدار PML باهم ادغام شده و یک خوشه‌ی جدید ایجاد می‌شود. در این مرحله تعداد خوشه‌ها به $n-1$ خوشه کاهش می‌یابد (یک خوشه دارای دو مشاهده و مابقی خوشه‌ها دارای یک مشاهده هستند).

۳ـ تمام $\binom{n-1}{2}$ گروه‌بندی زوجی ممکن از $n-1$ خوشه ارزیابی می‌شوند. جفت دارای بیشترین مقدار PML باهم ادغام شده و یک خوشه‌ی جدید ایجاد می‌شود. در این مرحله تعداد خوشه‌ها به $n-2$ خوشه کاهش می‌یابد.

۴ـ فرایند تا زمانی ادامه می‌یابد که یک خوشه حاوی کل مجموعه داده ایجاد شود. شایان ذکر است که تعداد کل معیار قرابت محاسبه‌شده برای خوشه بندی $\sum_{i=0}^{n-2} \binom{n-i}{2}$ است.

یک ملاک برای تعیین تعداد مناسب خوشه‌ها، استفاده از معیار آکائیکه (AIC) است. از این رو، در هر مرحله از فرآیند خوشه‌بندی مقدار AIC محاسبه و تعداد خوشه‌های با کمترین مقدار AIC به‌عنوان تعداد بهینه انتخاب می‌شود.

۴ ارزیابی روش خوشه‌بندی پیشنهادی

در این بخش، الگوریتم خوشه‌بندی پیشنهادی با معیار قرابت PML و روش اتصال S و الگوریتم سلسله‌مراتبی تجمیعی با معیارهای عدم‌تشابه اقلیدسی و منهتن و روش‌های اتصال کامل، میانگین، وارد و تک اتصالی برای گروه‌بندی داده‌های زمین‌آماری شبیه‌سازی‌شده از توزیع SGαS را مقایسه و ارزیابی می‌کنیم. از دو معیار اعتبارسنجی خوشه‌ای، خطای خوشه‌بندی (MI) و رند تعدیل‌یافته (ARI) برای مقایسه‌ی نتایج الگوریتم‌های خوشه‌بندی در این داده‌ها استفاده شده‌است.

شبیه‌سازی داده‌های زمین‌آماری از توزیع آمیخته‌ی SGαS

شبیه‌سازی داده‌های فضایی برای ارزیابی کارایی و استواری روش‌های آمار فضایی، تحلیل انواع داده‌های فضایی و الگوریتم‌های خوشه‌بندی فضایی ضروری است. در این بخش، یک الگوریتم برای شبیه‌سازی از توزیع‌های آمیخته‌ی میدان‌های تصادفی SGαS برای داده‌های زمین‌آماری ارائه می‌کنیم که جزئیات آن در جدول **الگوریتم ۱** آمده است.

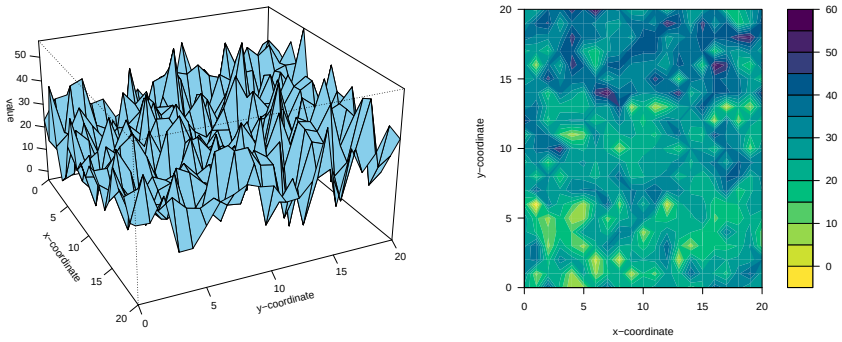
الگوریتم ۱: تولید داده‌های زمین‌آماری از یک توزیع آمیخته‌ی $SG\alpha S$ با K گروه

ورودی: n : تعداد کل مشاهدات
K : تعداد گروه‌ها
α : مقدار شاخص پایداری
$\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$: موقعیت‌های مکانی
$C_k(\|\boldsymbol{h}\|);\ k = 1, \ldots, K$: توابع کواریانس K گروه
$p_k; k = 1, \ldots, K$: پارامترهای هر گروه، $0 \leq p_k \leq 1$، $\sum_{k=1}^{K} p_k = 1$
$\boldsymbol{\delta}_k;\ k = 1, \ldots, K$: مولفه‌های بردار مکان $\boldsymbol{\delta}_{n\times 1} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K)'$

۱. تولید $(n_1, \ldots, n_K)$ از توزیع چندجمله‌ای با پارامترهای $(n, p_1, \ldots, p_K)$؛ $\sum_{k=1}^{K} n_k = n$
۲. انتخاب K زیرمجموعه از $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ به صورت: $\{\boldsymbol{s}_1^1, \ldots, \boldsymbol{s}_{n_1}^1\}, \ldots, \{\boldsymbol{s}_1^K, \ldots, \boldsymbol{s}_{n_K}^K\}$
۳. محاسبه‌ی تأخیرهای فضایی: $\|\boldsymbol{h}_{ij}^k\| = \|\boldsymbol{s}_i^k - \boldsymbol{s}_j^k\|$، $i, j = 1, \ldots, n_k$، $i \neq j$
۴. ایجاد ماتریس‌های کواریانس $\boldsymbol{\Sigma}_k = \left[C_k(\|\boldsymbol{h}_{ij}^k\|)\right]$؛ $k = 1, \ldots, K$، $i, j = 1, \ldots, n_k$
۵. ایجاد ماتریس بلوکی $\boldsymbol{\Sigma}_{n\times n} = \operatorname{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$
۶. قرار دهید: $\boldsymbol{\mu}_{n\times 1} = \boldsymbol{0}$
۷. تولید از توزیع گاوسی چند متغیره $\boldsymbol{z} \leftarrow N_n(\boldsymbol{\mu}_{n\times 1}, \boldsymbol{\Sigma}_{n\times n})$
۸. تولید از توزیع پایدار مثبت $x \leftarrow S_{\frac{\alpha}{2}}\left(\left(\cos\frac{\pi\alpha}{4}\right)^{\frac{1}{\alpha}}, 1, 0\right)$
۹. قرار دهید: $\boldsymbol{\delta}_{n\times 1} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K)'$
۱۰. محاسبه‌ی بردار $\boldsymbol{y} \leftarrow x^{\frac{1}{2}}\boldsymbol{z} + \boldsymbol{\delta}$

خروجی: $\boldsymbol{y}$: یک بردار داده‌های زمین‌آماری از توزیع SGαS با K گروه.

ابتدا $n = 441$ موقعیت مکانی را روی یک شبکه منظم 21×21 در مربع $[0, 20] \times [0, 20]$ ایجاد کردیم. تعداد گروه‌های واقعی را برابر با سه $(K = 3)$ با اندازه‌های مختلف خوشه‌ای انتخاب کرده‌ایم. ساختار همبستگی را مدل نمایی بدون پارامتر اثر قطعه‌ای و با $\sigma = 1$ و پارامتر دامنه‌ی θ درنظر می‌گیریم. برای تعیین اندازه‌ی گروه‌ها از یک توزیع چندجمله‌ای با درصدهای یکسان $p_1 = p_2 = p_3 = \frac{1}{3}$ استفاده کردیم که اندازه‌های $n_1 = 151$، $n_2 = 149$ و $n_3 = 141$ به‌دست آمد. بردارهای مکان سه گروه را به صورت $\boldsymbol{\delta}_1 = 22\ \mathbf{1}_{n_1\times 1}$، $\boldsymbol{\delta}_2 = 28\ \mathbf{1}_{n_2\times 1}$ و $\boldsymbol{\delta}_3 = 33\ \mathbf{1}_{n_3\times 1}$ درنظر گرفتیم. سپس، داده‌ها از توزیع آمیخته‌ی سه جزئی SGαS با مقادیر مختلف α و θ در مدل کوواریانس نمایی، در $n = 441$ موقعیت مکانی تولید شدند. از اینرو، سناریوهای شبیه‌سازی متفاوتی برای پارامترهای α و θ انجام شد. نمودار سه بعدی مشاهدات و نمودار کانتور برای یک نمونه از این شبیه‌سازی‌های در شکل ۱ نشان داده‌شده است.



شکل ۱: (چپ) نمودار سه‌بعدی مشاهدات و (راست) نمودار کانتور برای یک نمونه داده‌ی فضایی از توزیع آمیخته‌ی سه جزئی SGαS با $\alpha = ۰/۷۵$

نتایج خوشه‌بندی‌ها

روش سلسله مراتبی با معیارهای اقلیدسی و منهتن و چهار معیار اتصال، و الگوریتم خوشه‌بندی و معیار قرابت پیشنهادی برای گروه‌بندی همه‌ی مجموعه داده‌های تولید شده پیاده‌سازی شد. مقادیر شاخص‌های MI و ARI برای هر یک از مجموعه داده‌ها پس از ۱۰۰۰ تکرار و انجام این الگوریتم‌های خوشه‌بندی در تمام این داده‌ها با استفاده از مقادیر متفاوت α، θ، معیارهای تشابه و پیوند مختلف، محاسبه و در جدول ۱ ارائه شده‌است. با توجه به این جدول، نتایج زیر حاصل می‌شود:

- معیارهای PML و سپس درستنمایی عملکرد بهتری نسبت به بقیه دارند.
- با افزایش مقادیر پارامترهای α و θ، کارایی همه‌ی الگوریتم‌ها افزایش می‌یابد.
- صرف‌نظر از معیارهای قرابت به‌کار رفته، معیار تک اتصالی عملکرد ضعیف‌تری نسبت به بقیه دارد.

لازم به ذکر است که بسته‌های `stable` و `mclust` در نرم‌افزار R برای اجرای این مثال شبیه‌سازی استفاده شده‌اند.

جدول ۱: مقادیر تجمعی MI و ARI خوشه‌بندی داده‌های شبیه‌سازی‌شده از توزیع‌های SGαS با روش‌های پیشنهادی و سلسله مراتبی

		$\alpha = ۰/۵$				$\alpha = ۱/۵$			
		$\theta = ۰/۵$		$\theta = ۱$		$\theta = ۲/۵$		$\theta = ۱۰$	
معیار قرابت	معیار اتصال	ARI	MI	ARI	MI	ARI	MI	ARI	MI
	کامل	۵.۳۸	۱.۳۲	۵.۴۷	۳.۲۷	۲.۵۱	۴.۲۳	۴.۵۷	۸.۱۹
	تکی	۵.۳۰	۷.۳۸	۷.۳۸	۲.۳۴	۱.۴۲	۱.۲۷	۷.۴۶	۳.۲۲
اقلیدسی	متوسط	۱.۳۷	۶.۳۳	۱.۴۴	۳.۲۹	۲.۴۹	۶.۲۴	۳.۵۴	۶.۲۰
	وارد	۸.۳۹	۸.۲۹	۵.۴۸	۱.۲۵	۲.۵۲	۷.۲۳	۱.۶۰	۸.۱۷
	کامل	۱.۳۹	۶.۳۰	۹.۴۷	۱.۲۵	۴.۵۲	۵.۲۲	۱.۵۹	۲.۱۸
	تکی	۷.۳۱	۹.۳۶	۱.۴۰	۲.۳۳	۵.۴۵	۳.۲۶	۳.۴۹	۴.۲۱
منهتن	متوسط	۴.۳۸	۵.۳۱	۷.۴۵	۳.۲۸	۶.۵۰	۵.۲۲	۲.۵۷	۸.۱۸
	وارد	۷.۴۱	۵.۲۷	۹.۴۸	۲.۲۴	۱.۵۵	۶.۲۰	۸.۶۱	۳.۱۷
درستنمایی	LLF	۴.۵۰	۶.۲۲	۲.۵۳	۱.۲۰	۹.۶۱	۴.۱۷	۷.۶۸	۵.۱۳
PML	S	۲.۵۴	۵.۱۹	۱.۵۷	۳.۱۸	۸.۶۳	۶.۱۴	۲.۶۸	۹.۱۱

		$\alpha = ۱/۵$							
		$\theta = ۰/۵$		$\theta = ۱$		$\theta = ۲/۵$		$\theta = ۱۰$	
معیار قرابت	معیار اتصال	ARI	MI	ARI	MI	ARI	MI	ARI	MI
	کامل	۲.۵۷	۳.۲۰	۱.۵۹	۵.۱۸	۷.۶۵	۸.۱۴	۸.۷۱	۷۰.۹
	تکی	۳.۴۵	۱.۲۶	۶.۴۷	۵.۲۴	۲.۵۵	۴.۲۰	۶.۶۷	۱.۱۶
اقلیدسی	متوسط	۳.۵۶	۸.۲۱	۵.۵۸	۱.۱۹	۷.۶۴	۴.۱۵	۳.۷۰	۵.۱۰
	وارد	۲.۵۹	۵.۱۸	۴.۶۲	۷.۱۶	۵.۶۸	۱.۱۴	۲.۷۳	۳۷.۹
	کامل	۰.۵۸	۷.۱۹	۸.۵۹	۱.۱۷	۱.۶۷	۵.۱۲	۵.۷۳	۱۲.۹
	تکی	۳.۴۶	۱.۲۴	۱.۴۸	۵.۲۲	۴.۵۸	۶.۱۷	۲.۶۸	۳.۱۵
منهتن	متوسط	۴.۵۶	۲.۲۰	۶.۵۸	۹.۱۸	۳.۶۵	۱.۱۵	۲.۷۱	۸۳.۹
	وارد	۷.۶۰	۱.۱۸	۳.۶۴	۱.۱۶	۱.۷۰	۰.۱۳	۳.۷۵	۹۱.۸
درستنمایی	LLF	۷.۶۴	۲.۱۵	۵.۶۸	۱.۱۴	۸.۷۴	۵.۱۱	۲.۸۰	۵۲.۷
PML	S	۵.۶۸	۱.۱۴	۱.۷۰	۰.۱۳	۱.۷۶	۷۰.۸	۷.۸۳	۹۱.۶

بحث و نتیجه‌گیری

هدف این مقاله خوشه‌بندی داده‌های زمین‌آماری حاوی مقادیر کرانگین است که دم‌های سنگینی را در توزیع خود نشان می‌دهند و هیچ تبدیل گاوسی در آنها یافت نمی‌شود. بنابراین، توزیع‌های SGαS مورد بررسی قرار گرفت. ابتدا مدل‌هایی

برای این داده‌ها از K گروه از میدان‌های تصادفی مانای SGαS ارائه کردیم. از آنجایی که تابع درستنمایی، ساختار فضایی داده‌ها را از طریق ماتریس کوواریانس یا پراکندگی در نظر می‌گیرد، می‌تواند معیار قرابت مناسب در خوشه‌بندی داده‌های زمین‌آماری باشد. از سوی دیگر، این معیار تأخیرهای فضایی را نیز در نظر می‌گیرد تا خوشه‌های فضایی متمرکز با بیشترین شباهت در متغیرهای پاسخ ایجاد کند. بنابراین، یک معیار قرابت جدید PML را بر اساس تابع درستنمایی در یک فرآیند خوشه‌بندی سلسله‌مراتبی برای این نوع داده‌ها پیشنهاد دادیم. در نهایت، بر اساس داده‌های شبیه‌شده از توزیع آمیخته‌ی SGαS نشان دادیم که معیارهای PML و درستنمایی بهتر از سایر معیارهای درنظر گرفته‌شده در خوشه‌بندی داده‌های زمین‌آماری دم‌سنگین عمل می‌کنند.

مراجع

محمدزاده، م. (۱۳۹۸)، *آمار فضایی و کاربردهای آن*، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،

موسوی، س. س. (۱۴۰۲)، *خوشه‌بندی داده‌های فضایی دم‌سنگین*، رساله دکتری، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران).

Kerby, A. (2009), *Spatial Clustering Using the Likelihood Function*, University of Nebraska-Lincoln.

Nolan, J. (2013), Multivariate Elliptically Contoured Stable Distributions: Theory and Estimation *Computational Statistics*, **28**, 2067-2089.

Spodarev, E., Shmileva, E., and Roth, S. (2015), Extrapolation of Stationary Random Fields, *Stochastic Geometry, Spatial Statistics and Random Fields*, 321-368.

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲ Oct. 2023 25-26
conf.ikiu.ac.ir/spatial5

مدل‌بندی پاسخ‌های چندمتغیره فضایی در GAMLSS با مفصل

نیما نخعی[۱] [۱]، مهسا نادی‌فر[۲]، حسین باغیشنی[۱]، نگار اقبال[۱]
[۱]گروه آمار، دانشگاه صنعتی شاهرود
[۲]گروه آمار، دانشگاه پرتوریا، آفریقای جنوبی

چکیده: مدل‌های چندمتغیره نوینی که با نظریه مفصل توسعه یافته‌اند، بر مبنای خانواده مدل‌های GAMLSS پایه‌ریزی شده‌اند به‌طوری‌که در آن نه فقط میانگین متغیرهای پاسخ بلکه سایر پارامترهای توزیع‌های پاسخ‌ها به متغیرهای تبیینی رگرسیونی پیوند زده می‌شوند. پارامتر وابستگی مفصل که وظیفه مدل‌بندی وابستگی بین پاسخ‌ها را دارد نیز قابلیت متصل شدن به متغیرهای تبیینی را داراست. در این رهیافت، هم ساختار وابستگی و هم ویژگی‌های توزیعی پاسخ‌ها با ترکیب یک مدل GAMLSS و یک تابع مفصل پارامتری مناسب مدل‌بندی می‌شوند. این امر به ما این امکان را می‌دهد که نه تنها شدت وابستگی بین پاسخ‌ها را اندازه‌گیری کنیم، بلکه امکان شناخت عواملی که باعث وابستگی می‌شوند نیز فراهم می‌شود. در این مقاله، پس از بیان ساختار مدل‌بندی و رهیافت برازش مبتنی بر درستنمایی تاوانیده برای پاسخ‌های فضایی، با مثال‌های واقعی و شبیه‌سازی کاربست مدل را تشریح می‌کنیم.

واژه‌های کلیدی: مدل جمعی تعمیم‌یافته برای مکان، مقیاس و شکل (GAMLSS)، تابع مفصل، بردار پاسخ چندمتغیره فضایی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62M30.

۱ مقدمه

مدل‌های کلاسیک رگرسیونی معمولا به بررسی تاثیر مجموعه‌ای از متغیرهای تبیینی بر یک متغیر پاسخ تمرکز دارند. در این راستا می‌توان به مدل‌های خطی، خطی تعمیم‌یافته (GLM)، جمعی تعمیم‌یافته (GAM) و جمعی تعمیم‌یافته برای مکان، مقیاس و شکل[۱] (GAMLSS) اشاره کرد که در همه آن‌ها معمولا متغیر پاسخ تک‌متغیره است (وود ، ۲۰۰۶). اما در موقعیت‌های کاربردی متنوعی در حوزه‌های مختلف علوم، توسعه یک مدل رگرسیونی با دو یا چند متغیر پاسخ (به‌طور همزمان)، هدف جذاب مطالعه پژوهشی و تحلیل مساله است. علاوه‌بر وابستگی میان پاسخ‌ها، در بسیاری از مسایل خود متغیرهای پاسخ نیز دارای ساختارهای وابستگی شامل وابستگی زمانی، فضایی و فضایی‌ـزمانی هستند. لحاظ کردن

[1]Generalized additive models for location, scale and shape

[۱]نام و ایمیل ارائه دهنده مقاله: نیما نخعی، nnakhaei@shahroodut.ac.ir

این ساختارهای وابستگی در یک مدل با پاسخ چندمتغیره، مدل‌بندی را پیچیده‌تر می‌کند و توسعه رهیافت‌هایی که کارایی محاسباتی و آماری را دارا باشند، بسیار با اهمیت است. نظریه مفصل[۲] یک چارچوب مناسب و عملی را برای مدل‌بندی پاسخ‌های چندمتغیره در حوزه مدل‌های رگرسیونی فراهم کرده است که می‌توان به مطالعات واتر و چاوز_دمولین (۲۰۱۵)، یی (۲۰۱۵)، کلین و همکاران (۲۰۱۵) و مارا و رادیک (۲۰۱۷) اشاره کرد. در این مقاله، رهیافت مبتنی بر نظریه مفصل در چارچوب مدل‌های GAMLSS را برای مدل‌بندی پاسخ‌های چندمتغیره فضایی به‌کار می‌گیریم و با مثال‌های شبیه‌سازی و کاربردی، کارایی آن را ارزیابی می‌کنیم.

۲ چارچوب مدل‌بندی

توسعه مدل‌های رگرسیونی مبتنی بر مفصل در سال‌های اخیر به‌طور قابل توجهی مورد توجه بوده است. مزیت استفاده از نظریه مفصل در رده مدل‌های GAMLSS امکان مدل‌بندی منعطف پارامترهای مختلف توزیع‌های کناری پاسخ‌ها و پارامتر وابستگی مفصل است، به‌گونه‌ای که به پیشگوهای جمعی حاصل از اثرات مختلف، شامل خطی، ناخطی، تصادفی، فضایی و فضایی_زمانی، مرتبط می‌شوند. این امر به ما این امکان را می‌دهد که نه تنها شدت وابستگی بین پاسخ‌ها را اندازه‌گیری کنیم، بلکه امکان شناخت عواملی که باعث وابستگی می‌شوند را داشته باشیم. برای بیان چارچوب مدل‌بندی حالت دومتغیره را در نظر می‌گیریم.

۱.۲ مدل جمعی مفصل برای مکان، مقیاس و شکل

فرض کنید $F(y_1, y_2 \mid \boldsymbol{z}_1, \boldsymbol{z}_2)$ تابع توزیع تجمعی توام دو متغیر تصادفی پیوسته Y_1 و Y_2 به شرط دو مجموعه از متغیرهای تبیینی $\boldsymbol{z}_1$ و $\boldsymbol{z}_2$ باشد. امکان یکی بودن $\boldsymbol{z}_1$ و $\boldsymbol{z}_2$ وجود دارد. با استفاده از تابع مفصل دومتغیره $\mathcal{C}(\cdot, \cdot; \theta)$ می‌توان تابع توزیع را به‌صورت زیر نمایش داد:

$$F(y_1, y_2 \mid \boldsymbol{z}_1, \boldsymbol{z}_2) = \mathcal{C}(F_1(y_1 \mid \boldsymbol{z}_1), F_2(y_2 \mid \boldsymbol{z}_2), \theta) \qquad (۱.۲)$$

که در آن $F_1(y_1 \mid \boldsymbol{z}_1)$ و $F_2(y_2 \mid \boldsymbol{z}_2)$ توابع توزیع کناری Y_1 و Y_2 به شرط $\boldsymbol{z}_1$ و $\boldsymbol{z}_2$ هستند و θ پارامتر وابستگی مفصل است که وابستگی میان دو توزیع کناری را اندازه‌گیری می‌کند. برای توابع مفصل معروف، مانند کلایتون، گاوسی و گامبل، رابطه‌ای بین پارامتر θ و ضریب τ کندال، به‌عنوان یک معیار مناسب سنجش وابستگی، وجود دارد (نلسن ، ۲۰۰۶).

معمولا توزیع‌های کناری متغیرهای پاسخ Y_1 و Y_2 به وسیله توابع توزیع و چگالی پارامتری مشخص می‌شوند که می‌توان آن‌ها را با شکل کلی $F_m(y_m \mid \mu_m, \sigma_m, \nu_m)$ و $f_m(y_m \mid \mu_m, \sigma_m, \nu_m)$، برای $m = ۱, ۲$، نمایش داد، که در آن‌ها μ_m، σ_m و ν_m به ترتیب پارامترهای مکان، مقیاس و شکل هستند. این پارامترها، در چارچوب مدل‌های GAMLSS، به پیشگوهای جمعی η به کمک توابع پیوند یکنوای مناسب مرتبط می‌شوند. نمایش عمومی پیشگوی جمعی، برای n مشاهده، به صورت زیر است:

$$\eta_i = \beta_{۰} + \sum_{k=۱}^{K} s_k(\boldsymbol{z}_{ki}) \qquad i = ۱, \cdots, n \qquad (۲.۲)$$

به‌طوری که $\beta_{۰} \in \mathbb{R}$ پارامتر عرض از مبدا است، $\boldsymbol{z}_{ki}$ زیربردار kام از بردار متغیرهای تبیینی $\boldsymbol{z}_i$، شامل انواع متغیرهای دودویی، رسته‌ای، پیوسته، و فضایی، است. همچنین توابع $s_k(\boldsymbol{z}_{ki})$ بیانگر عمومی اثرات است که وابسته به نوع متغیرهای تبیینی متناظر تعیین می‌شوند. در چارچوب مدل‌های جمعی، هر اثر $s_k(\boldsymbol{z}_{ki})$ را می‌توان با یک ترکیب خطی از J_k تابع پایه

[۲] Copula

$b_{k_{j_k}}(\boldsymbol{z}_{ki})$، بر حسب ضرایب رگرسیونی $\beta_{k_{j_k}} \in \mathbb{R}$ به صورت $\sum_{j_k=1}^{J_k} \beta_{k_{j_k}} b_{k_{j_k}}(\boldsymbol{z}_{ki})$ تقریب زد. در این صورت، با بردار ضرایب $\boldsymbol{\beta}_k = (\beta_{k1}, \cdots, \beta_{kj_k})^\mathsf{T}$ و ماتریس طرح $Z_k[i, j_k] = b_{kj_k}(\boldsymbol{z}_{ki})$، می‌توان $\{s_k(\boldsymbol{z}_{k1}), \cdots, s_k(\boldsymbol{z}_{kn})\}^\mathsf{T}$ را به شکل $\boldsymbol{Z}_k\boldsymbol{\beta}_k$ بازنویسی کرد. بنابراین، برای هر n مشاهده، پیشگوی جمعی در (۲.۲) را می‌توان به شکل زیر نوشت:

$$\boldsymbol{\eta} = \beta_0 \mathbf{1}_n + \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{Z}_K\boldsymbol{\beta}_K \qquad (۳.۲)$$

که در آن $\mathbf{1}_n$ بردار n تایی از یک‌ها است. با قرار دادن $\boldsymbol{Z} = (\mathbf{1}_n, \boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_K)$ و $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^\mathsf{T}, \cdots, \boldsymbol{\beta}_K^\mathsf{T})^\mathsf{T}$، می‌توان معادله (۳.۲) را به شکل $\boldsymbol{\eta} = \boldsymbol{Z}\boldsymbol{\beta}$ نوشت.

برای اعمال هموارى روى اثر $s_k(\cdot)$، معمولا از جریمه درجه دوم به شکل $\lambda_k\boldsymbol{\beta}_k\boldsymbol{D}_k\boldsymbol{\beta}_k$ استفاده می‌شود. پارامتر هموارسازی $\lambda_k \geq 0$ وظیفه کنترل تعادل بین هموارسازی و برازش را دارد و در شناسایی شکل اثر نقش اساسی دارد. درایه‌های ماتریس $\boldsymbol{D}_k$ بر اساس نوع اثر متغیر تبیینی تعیین می‌شوند. برای اثرات ناخطی متغیرهای تبیینی پیوسته، این ماتریس به صورت $D_k = \int \boldsymbol{d}_k(z_k)\boldsymbol{d}_k(z_k)^\mathsf{T} dz_k$ تعریف می‌شود، که در آن j_kامین عضو $\boldsymbol{d}_k(z_k)$ از مشتق مرتبه دوم $b_{kj}(z_k)$ محاسبه می‌شود. برای اثر فضایی مشبکه‌ای برای داده‌های فضایی با M ناحیه، $\boldsymbol{\beta}_k$ بردار اثرات فضایی است و ماتریس طرح متناظر به صورت

$$\boldsymbol{Z}_k[i, m] = \begin{cases} 1 & \text{اگر مشاهده } i \text{ به ناحیه } m \text{ متعلق باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

تعریف می‌شود، که در آن $m = 1, \cdots, M$. همچنین درایه‌های ماتریس $\boldsymbol{D}_k$ به صورت

$$\boldsymbol{D}_k[m, q] = \begin{cases} -1 & m \neq q \bigwedge m \sim q \\ 0 & m \neq q \bigwedge m \nsim q \\ N_m & m = q \end{cases}$$

تعریف می‌شود، که در آن $m \sim q$ به معنی همسایه بودن دو ناحیه m و q است و N_m تعداد کل همسایه‌های ناحیه m را نشان می‌دهد. تفسیری از جریمه متناظر برای اثر فضایی بیان‌شده معادل با این پذیره است که اثرات فضایی $\boldsymbol{\beta}_k$ از یک میدان تصادفی گاوسی مارکوفی پیروی می‌کند. جریمه کل را می‌توان به صورت $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{D}_\lambda\boldsymbol{\beta}$ تعریف کرد که در آن $\boldsymbol{D}_\lambda = \text{diag}(0, \lambda_1\boldsymbol{D}_1, \cdots, \lambda_K\boldsymbol{D}_K)$.

۲.۲ برازش مبتنی بر درستنمایی تاوانیده

اگر f_1 و f_2 توابع چگالی متناظر با F_1 و F_2 باشند، آنگاه تابع چگالی احتمال توام f که از رابطه (۱.۲) نتیجه می‌شود، عبارتست از

$$f(y_{1i}, y_{2i}) = \frac{\partial^2 F(y_{1i}, y_{2i})}{\partial y_{1i}\partial y_{2i}} = \frac{\partial^2 \mathcal{C}(F_1(y_{1i}), F_2(y_{2i}))}{\partial F_1(y_{1i})\partial F_2(y_{2i})} \times \frac{\partial F_1(y_{1i})}{\partial y_{1i}} \times \frac{\partial F_2(y_{2i})}{\partial y_{2i}}$$

که در آن، برای سادگی نمادگذاری، شرطی کردن روی متغیرهای تبیینی حذف شده است. برای توزیع‌های کناری سه پارامتری در رده مدل‌های GAMLSS، می‌توان نوشت

$$\begin{aligned} f(y_{1i}, y_{2i}|\boldsymbol{\delta}) &= \mathrm{c}(F_1(y_{1i}|\mu_{1i}, \sigma_{1i}, \nu_{1i}), F_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i}); \theta_i) \\ &\times f_1(y_{1i}|\mu_{1i}, \sigma_{1i}, \nu_{1i}) \times f_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i}) \end{aligned}$$

که در آن $\mathrm{c}(\cdot,\cdot;\theta)$ چگالی مفصل است و $\boldsymbol{\delta}=(\boldsymbol{\beta}_{\mu_1}^\top,\boldsymbol{\beta}_{\mu_2}^\top,\boldsymbol{\beta}_{\sigma_1}^\top,\boldsymbol{\beta}_{\sigma_2}^\top,\boldsymbol{\beta}_{\nu_1}^\top,\boldsymbol{\beta}_{\nu_2}^\top,\boldsymbol{\beta}_{\theta}^\top)^\top$. بنابراین، تابع لگاریتم درستنمایی مدل عبارتست از

$$\begin{aligned}\ell(\boldsymbol{\delta}) &= \sum_{i=1}^{n}\log\{\mathrm{c}(F_1(y_{1i}\mid\mu_{1i},\sigma_{1i},\nu_{1i}),F_2(y_{2i}\mid\mu_{2i},\sigma_{2i},\nu_{2i});\theta_i)\}\\ &+ \sum_{i=1}^{n}\sum_{m=1}^{2}\log\{f_m(y_{mi}\mid\mu_{mi},\sigma_{mi},\nu_{mi})\}.\end{aligned}$$

در چارچوب مدل‌های GAMLSS پارامترهای $\boldsymbol{\delta}$ به پیشگوهای $\eta_{\mu_{1i}},\eta_{\mu_{2i}},\eta_{\sigma_{1i}},\eta_{\sigma_{2i}},\eta_{\nu_{1i}},\eta_{\nu_{2i}}$ و η_{θ_i} توسط توابع پیوند مناسب، که معکوس آن‌ها روی پیشگوها حافظ دامنه پارامترها باشند، مرتبط می‌شوند.

با توجه به ساختار منعطف پیشگوهای جمعی برای وارد کردن انواع اثرات ناخطی و فضایی، استفاده از تابع درستنمایی اصلی برای برآورد کردن پارامترهای مدل منجر به برآورد شکلی از توابع هموار اثرات می‌شود که می‌تواند دور از روند زیربنایی واقعی داده‌ها باشد (وود ، ۲۰۰۶). به همین دلیل، مارا و رادیک (۲۰۱۷) به‌کار بردن رهیافت ماکسیمم درستنمایی تاوانیده را پیشنهاد کردند که برای مدل هدف مقاله به صورت

$$\ell_p(\boldsymbol{\delta})=\ell(\boldsymbol{\delta})-\frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{S}_\lambda\boldsymbol{\delta} \qquad (۴.۲)$$

تعریف می‌شود، که در آن $\boldsymbol{S}_\lambda=\mathrm{diag}(\boldsymbol{\lambda}_{\mu_1}\boldsymbol{D}_{\mu_1},\boldsymbol{\lambda}_{\mu_2}\boldsymbol{D}_{\mu_2},\boldsymbol{\lambda}_{\sigma_1}\boldsymbol{D}_{\sigma_1},\boldsymbol{\lambda}_{\sigma_2}\boldsymbol{D}_{\sigma_2},\boldsymbol{\lambda}_{\nu_1}\boldsymbol{D}_{\nu_1},\boldsymbol{\lambda}_{\nu_2}\boldsymbol{D}_{\nu_2},\boldsymbol{\lambda}_{\theta}\boldsymbol{D}_{\theta})$ و هر بردار $\boldsymbol{\lambda}$ به صورت $(\lambda_1,\cdots,\lambda_K)^\top$ تعریف می‌شود. برای ماکسیمم کردن درستنمایی (۴.۲)، مارا و رادیک (۲۰۱۷) یک نسخه کارا از الگوریتم ناحیه اعتماد[3] معرفی‌شده توسط رادیک و همکاران (۲۰۱۶) را توسعه دادند که در آن پارامترهای هموارسازی به‌صورت خودکار انتخاب می‌شوند. این الگوریتم دومرحله‌ای است که برای دیدن جزییات آن، خواننده را به مارا و رادیک (۲۰۱۷) ارجاع می‌دهیم.

مرحله اول: در تکرار a، با ثابت نگه داشتن بردار پارامترهای هموارساز در $\boldsymbol{\lambda}^{[a]}$ و به ازای مقادیر مفروض $\boldsymbol{\delta}^{[a]}$، درستنمایی (۴.۲) با استفاده از الگوریتم ناحیه اعتماد ماکسیمم می‌شود. به عبارت دیگر

$$\boldsymbol{\delta}^{[a+1]} = \boldsymbol{\delta}^{[a]}+\underset{\boldsymbol{e}:\|\boldsymbol{e}\|\le\triangle^{[a]}}{\mathrm{argmin}}\ \breve{\ell}_p(\boldsymbol{\delta}^{[a]}) \qquad (۵.۲)$$

که در آن

$$\breve{\ell}_p(\boldsymbol{\delta}^{[a]})=-\left\{\ell_p(\boldsymbol{\delta}^{[a]})+\boldsymbol{e}^\top\boldsymbol{g}_p(\boldsymbol{\delta}^{[a]})+\frac{1}{2}\boldsymbol{e}^\top\boldsymbol{H}_p(\boldsymbol{\delta}^{[a]})\boldsymbol{e}\right\}$$

به‌طوری که $\boldsymbol{g}_p(\boldsymbol{\delta}^{[a]})=\boldsymbol{g}(\boldsymbol{\delta}^{[a]})-\boldsymbol{S}\boldsymbol{\delta}^{[a]}$ و $H_p(\boldsymbol{\delta}^{[a]})=H(\boldsymbol{\delta}^{[a]})-\boldsymbol{S}$. مولفه‌های بردار $\boldsymbol{g}(\boldsymbol{\delta}^{[a]})$ عبارتند از

$$\boldsymbol{g}_{\mu_1}(\boldsymbol{\delta}^{[a]})=\frac{\partial\ell(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}_{\mu_1}}|_{\boldsymbol{\beta}_{\mu_1}=\boldsymbol{\beta}_{\mu_1}^{[a]}},\cdots,\boldsymbol{g}_{\theta}(\boldsymbol{\delta}^{[a]})=\frac{\partial\ell(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}_{\theta}}|_{\boldsymbol{\beta}_{\theta}=\boldsymbol{\beta}_{\theta}^{[a]}}.$$

درایه‌های ماتریس هسیان عبارتند از $H(\boldsymbol{\delta}^{[a]})=\frac{\partial^2\ell(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}_r\partial\boldsymbol{\beta}_h}|_{\boldsymbol{\beta}_r=\boldsymbol{\beta}_r^{[a]},\boldsymbol{\beta}_h=\boldsymbol{\beta}_h^{[a]}}$، که در آن $r,h=\mu_1,\mu_2,\sigma_1,\sigma_2.\nu_1,\nu_2,\theta$. افزون بر این، $\|\cdot\|$ نرم اقلیدسی و $\triangle^{[a]}$ شعاع ناحیه اعتماد است که در هر تکرار الگوریتم تعدیل می‌شود. با به کارگیری بردار امتیاز و ماتریس هسیان تحلیلی، این مرحله دقیق و سریع اجرا می‌شود. در صورت بسته نبودن این دو کمیت از تقریب‌های عددی بردار امتیاز و ماتریس هسیان استفاده خواهد شد.

مرحله دوم: با ثابت نگه داشتن بردار پارامترها در $\boldsymbol{\delta}^{[a+1]}$، کمیت زیر باید حل شود:

$$\boldsymbol{\lambda}^{[a+1]}=\underset{\lambda}{\mathrm{argmin}}\|\boldsymbol{M}^{[a+1]}-\boldsymbol{A}^{[a+1]}\boldsymbol{M}^{[a+1]}\|^2-\breve{n}+2\mathrm{tr}(\boldsymbol{A}^{[a+1]})$$

[3] Trust region algorithm

که در آن $\boldsymbol{M}^{[a+1]} = \sqrt{-\boldsymbol{H}(\boldsymbol{\delta}^{[a+1]})}\boldsymbol{\delta}^{[a+1]} + \sqrt{-\boldsymbol{H}(\boldsymbol{\delta}^{[a+1]})}^{-1}\boldsymbol{g}(\boldsymbol{\delta}^{[a+1]})$ و

$$\boldsymbol{A}^{[a+1]} = \sqrt{-\boldsymbol{H}(\boldsymbol{\delta}^{[a+1]})}(-\boldsymbol{H}(\boldsymbol{\delta}^{[a+1]}) + \boldsymbol{S})^{-1}\sqrt{-\boldsymbol{H}(\boldsymbol{\delta}^{[a+1]})}.$$

همچنین $\mathrm{tr}(\boldsymbol{A}^{[a+1]})$ تعداد درجات آزادی موثر[۴] مدل جریمه‌شده است و چنان‌چه هر دو توزیع کناری سه پارامتر داشته باشند، $\breve{n} = ۷n$. بردار شیب $\boldsymbol{g}$ و ماتریس هسین در مرحله اول به‌دست می‌آیند، هر دو برای محاسبه برآورد به‌صورت مدلار استخراج شده‌اند، به عنوان مثال:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\delta})}{\partial \beta_{\mu_1}} &= \sum_{i=1}^{n}\Big\{\frac{1}{f_1(y_{1i}\mid \mu_{1i},\sigma_{1i},\nu_{1i})}\frac{\partial f_1(y_{1i}\mid \mu_{1i},\sigma_{1i},\nu_{1i})}{\partial \mu_{1i}} \\ &+ \frac{1}{c(F_1(y_{1i}\mid \mu_{1i},\sigma_{1i},\nu_{1i}),F_2(y_{2i}\mid \mu_{2i},\sigma_{2i},\nu_{2i});\theta_i)} \\ &\times \frac{\partial c(F_1(y_{1i}\mid \mu_{1i},\sigma_{1i},\nu_{1i}),F_2(y_{2i}\mid \mu_{2i},\sigma_{2i},\nu_{2i});\theta_i)}{\partial F_1(y_{1i}\mid \mu_{1i},\sigma_{1i},\nu_{1i})}\frac{\partial F_1(y_{1i}\mid \mu_{1i},\sigma_{1i},\nu_{1i})}{\partial \mu_{1i}}\Big\}\frac{\partial \mu_{1i}}{\partial \eta_{\mu_{1i}}}\boldsymbol{Z}_{\mu_{1i}}\end{aligned}$$

که مشتق مرتبه اول $\ell(\boldsymbol{\delta})$ با توجه به β_{μ_2}، β_{σ_1}، β_{σ_2}، β_{ν_1} و β_{ν_2} بسیار شبیه به رابطه بالا است.

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\delta})}{\partial \beta_{\theta}} &= \sum_{i=1}^{n}\Big\{\frac{1}{c(F_1(y_{1i}\mid \mu_{1i},\sigma_{1i},\nu_{1i}),F_2(y_{2i}\mid \mu_{2i},\sigma_{2i},\nu_{2i});\theta_i)} \\ &\times \frac{\partial c(F_1(y_{1i}\mid \mu_{1i},\sigma_{1i},\nu_{1i}),F_2(y_{2i}\mid \mu_{2i},\sigma_{2i},\nu_{2i});\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \eta_{\theta_i}}\Big\}\boldsymbol{Z}_{\theta_i}.\end{aligned}$$

دو مرحله الگوریتم برازش تا زمانی تکرار می‌شوند که معیار

$$\frac{|\ell(\boldsymbol{\delta}^{[a+1]}) - \ell(\boldsymbol{\delta}^{[a]})|}{۰/۱ + |\ell(\boldsymbol{\delta}^{[a+1]})|} < ۱۰^{-۷}$$

برای دو مرحله متوالی برقرار شود. شایان ذکر است که رویکرد الگوریتم ناحیه اعتماد معمولا پایدارتر و سریع‌تر از رقیبانی مانند روش نیوتن–رافسون است. با توجه به توزیع مجانبی نرمال برای برآوردگرهای ماکسیمم درستنمایی، ساخت فواصل اطمینان و انجام آزمون فرضیه برای پارامترهای مدل و اثرات هموار سرراست است.

۳.۲ فرایند ساخت مدل

فرآیند ساخت مدل در چارچوب پیشنهادی، شامل سه مرحله کلیدی است: ۱) انتخاب توزیع‌های کناری مناسب، ۲) انتخاب متغیرهای تبیینی موثر برای ساخت پیشگوهای هر پارامتر و ۳) انتخاب تابع مفصل مناسب که بهترین تطبیق و کمی‌سازی قدرت وابستگی میان توزیع‌های کناری را داشته باشد. انجام هر کدام از این سه مرحله، مبتنی بر معیارهایی است که در ادامه به اختصار آن‌ها را تشریح می‌کنیم.

از مانده‌های چندکی می‌توان برای ارزیابی نیکویی برازش توزیع‌های کناری پاسخ‌ها استفاده کرد. این نوع باقی‌مانده‌ها برای هر توزیع کناری به صورت $\hat{r}_{mi} = \Phi^{-1}\{F_m(y_{mi}\mid \hat{\mu}_{mi},\hat{\sigma}_{mi},\hat{\nu}_{mi})\}$، برای $m = ۱, ۲$ و $i = ۱, \cdots, n$، تعریف می‌شود، که در آن $\Phi^{-1}(\cdot)$ تابع چندک توزیع نرمال استاندارد است. اگر $F_m(y_{m_i})$ به توزیع واقعی نزدیک باشد، چندک‌های $\hat{r}_{mi}$ تقریبا از توزیع نرمال استاندارد پیروی می‌کنند. بنابراین ترسیم نمودار چندک–چندک این باقی‌مانده‌ها برای شناسایی نقصان برازش توزیع‌های کناری روش گرافیکی مناسبی است. البته باید توجه داشت که نیکویی برازش توزیع‌های کناری، در یک مدل چندمتغیره، یک شرط لازم است ولی کافی نیست. انتخاب متغیرهای تبیینی برای پیشگوها بر اساس

[۴] Effective degrees of freedom (edf)

معیارهای انتخاب مدل AIC و BIC قابل اجراست. البته برای انتخاب مدل پایه نه چندان دور از واقعیت، شناخت خوب از مساله مورد بررسی ضروری است. با توجه به توزیع‌های کناری و پارامترهای پیشگو، انتخاب تابع مفصل نیز بر اساس مقادیر بهینه AIC و BIC انجام می‌شود. این دو معیار در مدل پیشنهادی به صورت زیر تعریف می‌شوند:

$$\mathrm{AIC} = -۲\ell(\hat{\boldsymbol{\delta}}) + ۲\mathrm{edf}, \quad \mathrm{BIC} = -۲\ell(\hat{\boldsymbol{\delta}}) + \log(n)\mathrm{edf}$$

که در آن $\mathrm{edf} = \mathrm{tr}(\hat{\boldsymbol{A}})$ و لگاریتم درستنمایی به ازای برآوردهای ماکسیمم درستنمایی تاوانیده مقداردهی شده است.

۳ ارزیابی و کاربست مدل پیشنهادی

برای ارزیابی عملکرد مدل، $n = ۷۰$ مشاهده از بردار پاسخ دومتغیره $(Y_۱, Y_۲)$ را توسط مدل زیر شبیه‌سازی کردیم:

$$\begin{aligned}
(Y_۱, Y_۲)^\top | x_i &\sim N_۲\left((\mu_۱, \mu_۲)^\top, \Sigma\right), \quad \Sigma = \begin{bmatrix} \sigma_۱^۲ & \rho\sigma_۱\sigma_۲ \\ \rho\sigma_۱\sigma_۲ & \sigma_۲^۲ \end{bmatrix} \\
\eta_{\mu_d} &= \beta_{۰_d} + \beta_{۱_d} x_i + \phi_i \\
\log(\eta_{\sigma_d}) &= \gamma_{۰_d} + \gamma_{۱_d} x_i + \phi_i, \quad d = ۱, ۲
\end{aligned}$$

که در آن متغیر تبیینی x از یک توزیع نرمال استاندارد شبیه‌سازی شد، مولفه فضایی ϕ توسط یک میدان تصادفی گاوسی مارکوفی ذاتی (یا همان مدل ICAR) با میانگین صفر و بر اساس نواحی ۷۰ گانه کشور آلمان با مقادیر متفاوت برای پارامتر واریانس $\sigma_\phi^۲$ مدل‌بندی شد و مقادیر $\rho = ۰/۴$، $(\beta_{۰_d}, \beta_{۱_d}) = (۰/۹, ۰/۸۵)$ و $(\gamma_{۰_d}, \gamma_{۱_d}) = (-۰/۵, ۱)$، برای $d = ۱, ۲$، انتخاب شدند. برای برازش مدل از بسته SemiParBIVProbit در نرم‌افزار R استفاده کردیم. به ازای مقادیر مختلف σ_ϕ ابتدا تابع مفصل بهینه بر اساس معیارهای AIC و BIC انتخاب شدند و سپس مدل نهایی با انتخاب توزیع‌های کناری نرمال برازش شدند. نتایج در جدول ۱ گزارش شده‌اند. با بزرگ‌تر شدن مقادیر واریانس اثر فضایی، که به معنی سهم بیشتر این اثر در هر دو پارامتر توزیع‌های کناری است، دقت برآوردها برای ضرایب رگرسیونی پارامترهای واریانس کمتر می‌شود. حتی برای پاسخ دوم ضریب $\gamma_۰$ با علامت اشتباه برآورد شده است. روندی تقریبا معکوس برای ضرایب رگرسیونی پارامترهای میانگین توزیع‌های کناری قابل برداشت است. در مجموع برازش مناسب مدل تحت شرایط مختلف برای میزان سهم اثر فضایی، قابل ملاحظه است.

جدول ۱: نتایج برازش مدل دومتغیره در مثال شبیه‌سازی

σ_ϕ	مفصل	ضرایب $\mu_۱$	ضرایب $\sigma_۱$	ضرایب $\mu_۲$	ضرایب $\sigma_۲$
۰/۱۶	Joe	(۱/۲۰, ۰/۵۸)	(−۰/۴۷, ۰/۸)	(۰/۳۴, ۱/۸۰)	(−۰/۵, ۰/۷۶)
۰/۹۶	Gussian	(۰/۵۵, ۰/۵۷)	(−۰/۵, ۰/۹۳)	(۱/۰۰, ۰/۹۹)	(−۰/۴, ۰/۹۴)
۱/۹۶	Joe	(۰/۸۳, ۰/۶۴)	(−۱/۰, ۰/۷۸)	(۰/۵۰, ۱/۸۰)	(−۰/۵, ۱/۰۰)
۴/۹۶	Frank	(۰/۸۳, ۱/۳۰)	(−۱/۱, ۰/۶۷)	(۰/۵۳, ۱/۷۰)	(−۰/۷, ۰/۸۹)
۸/۹۶	Plackett	(۱/۳۰, ۰/۸۰)	(−۰/۹۹, ۰/۶)	(۱/۱۰, ۰/۸۴)	(۰/۴۴, ۱/۱۰)
مقادیر واقعی $(\beta_{۰_d}, \beta_{۱_d}) = (۰/۹, ۰/۸۵)$ و $(\gamma_{۰_d}, \gamma_{۱_d}) = (-۰/۵, ۱)$					

مجموعه داده واقعی این مثال، تعداد مرگ و میر کودکان هنگام تولد در ۱۰۰ ناحیه ایالت کارولینای شمالی طی دو دوره سال‌های ۷۸ − ۱۹۷۴ و سال‌های ۸۴ − ۱۹۷۹ می‌باشد. در این‌جا تعداد فوتی‌ها در دو بازه زمانی اشاره‌شده را به‌عنوان

جدول ۲: درمقادیر معیارهای ارزیابی مدل برای انتخاب مفصل مناسب در مثال واقعی

مفصل	AIC	BIC
نرمال	۱۸۵۰/۰	۱۸۸۸/۰
فرانک	۱۷۳۷/۵	۱۷۷۵/۵
علی_میخاییل_حق	۱۶۸۳/۶	۱۷۱۷/۲
فارلی_گامبل_مورگنسترن	۱۸۰۳/۵	۱۸۳۷/۱
کلایتون	**۱۶۷۹/۸**	**۱۷۱۳/۷**
گلمبوس	۱۷۲۶/۸	۱۷۶۲/۷
جو	۱۷۶۸/۷	۱۸۰۲/۴
گمبل	۱۶۹۶/۴	۱۷۳۳/۶
کلایتون ۹۰ درجه	۱۹۳۳/۹	۱۹۶۶/۳
گلمبوس ۹۰ درجه	۱۸۹۵/۷	۱۹۳۴/۲
جو ۹۰ درجه	۱۹۸۴/۰	۲۰۲۲/۰
گمبل ۹۰ درجه	۱۹۶۴/۶	۲۰۰۳/۵

مشاهدات دو متغیر پاسخ شمارشی در نظر گرفتیم. با توجه به ماهیت پاسخ‌ها، توزیع پواسن را برای توزیع‌های کناری هر دو پاسخ انتخاب کردیم. این مجموعه داده در بسته spData، در نرم‌افزار R، قابل دسترس هستند. مدل مورد نظر را می‌توان به‌صورت

$$\begin{aligned}(Y_1, Y_2)^\top | \boldsymbol{\mu} = (\mu_1, \mu_2)^\top &\sim \text{BPoisson}(\boldsymbol{\mu}) \\ \log(\mu_d) &= \beta_{0d} + \phi_i \qquad d = 1, 2, \quad i = 1, \cdots, 100 \end{aligned} \tag{۱.۳}$$

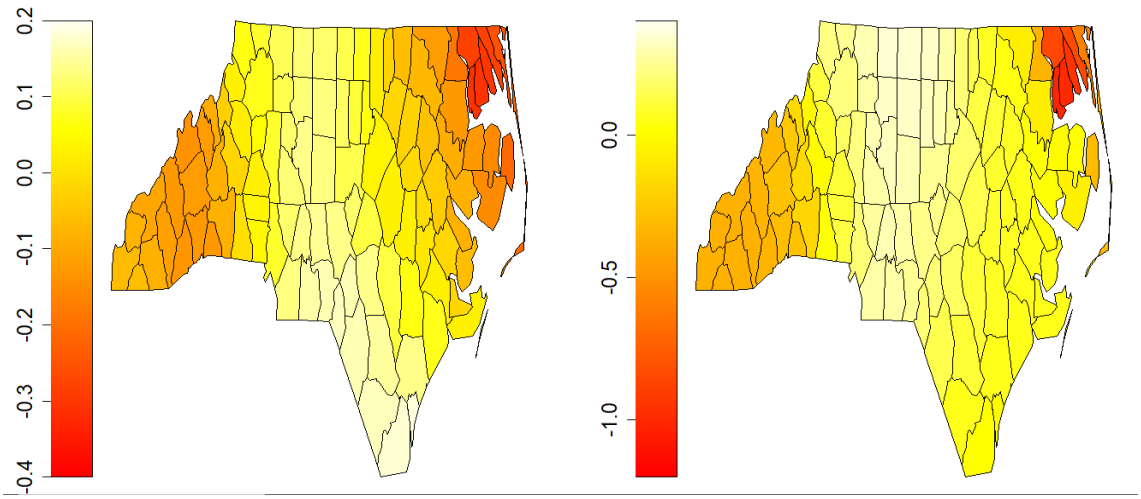
بیان کرد، که در آن متغیر فضایی ϕ توسط یک مدل ICAR با پارامتر واریانس σ^2_ϕ مدل‌بندی شد. به ازای توابع مفصل مختلف، مدل (۱.۳) برازش داده شد که مقادیر معیارهای AIC و BIC آن‌ها در جدول ۲ گزارش شده‌اند. بر اساس نتایج گزارش‌شده، تابع مفصل کلایتون به‌عنوان بهترین مفصل انتخاب و نتایج حاصل از برازش مدل با این مفصل در جدول ۳ گزارش شده‌اند. با توجه به برآورد پارامترهای واریانس اثر فضایی برای دو پاسخ، میزان همواری اثر برای پاسخ دوم بیشتر است. این نتیجه در شکل ۱ که پهنه‌بندی اثرات فضایی برآوردشده برای دو پاسخ را نشان می‌دهد نیز مشهود است. روند تغییرات نظام‌مند فضایی نیز از روی شکل برای هر دو پاسخ کاملا روشن است.

جدول ۳: نتایج برازش مدل پواسون دومتغیره فضایی با مفصل منتخب کلایتون در مثال واقعی

	پارامتر					
مفصل	β_{01}	β_{02}	$\sigma^{(1)}_\phi$	$\sigma^{(2)}_\phi$	AIC	BIC
کلایتون	۱/۸۸	۲/۰۴	۰/۵۷	۱/۰۸	۱۶۷۹/۸	۱۷۱۳/۷

بحث و نتیجه‌گیری

در این مقاله از یک رهیافت مبتنی بر نظریه مفصل در رده مدل‌های GAMLSS برای مدل‌بندی منعطف پاسخ‌های چندمتغیره فضایی استفاده کردیم. نتایج مثال‌های شبیه‌سازی و کاربردی، کارایی روش پیشنهادی را نشان دادند. نتایج برای پاسخ‌های



شکل ۱: پهنه‌بندی اثرات فضایی برآوردشده در مثال واقعی: سمت چپ برای Y_1 و سمت راست برای Y_2

فضایی مشبکه‌ای بیان شد. تعمیم رهیافت مدل‌بندی برای پاسخ‌های زمین‌آماری می‌تواند یک هدف جذاب برای توسعه مدل پیشنهادی باشد. اجرای استنباط برای مدل پیشنهادی در یک چارچوب بیزی نیز مورد توجه است.

مراجع

Klein N., Kneib T., Klasen S., Lang S. (2015), Bayesian Structured Additive Distributional Regression for Multivariate Responses, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **64**, 569-591.

Marra, G., Radice, R. (2017), Bivariate Copula Additive Models for Location, Scale and Shape, *Computational Statistics & Data Analysis*, **112**, pp. 99-113.

Nelsen, R. (2006), *An Introduction to Copulas*, Second edition, Springer, New York.

Radice, R., G. Marra, and M. Wojtys (2016), Copula Regression Spline Models for Binary Outcomes,*Statistics and Computing*,**26**, pp. 981-995.

Vatter T., Chavez-Demoulin V. (2015), Generalized Additive Models for Conditional Dependence Structures, *Journal of Multivariate Analysis*, **141**, 147-167.

Wood, S.N. (2006), *Generalized Additive Models: An Introduction With R*, Chapman and Hall, London.

Yee T. W. (2015), *Vector Generalized Linear and Additive Models, With An Implementation in R*, Springer, New York.

قزوین، دانشـگاه بین‌المللی امام‌خمینی، دانشکده علوم پایه، گروه آمار،
دبیرخانه پنجمین سـمینار آمار فضایی و کاربردهای آن،
تلفن: ۳۳۹۰۱۳۷۹ (۰۲۸) نمابر: ۳۳۷۸۰۰۴۰ (۰۲۸)
پست الکترونیکی spatial5@ikiu.ac.ir