
بسم الله الرحمن الرحیم

دانشکده علوم ریاضی

رشته آمار، گرایش آمار ریاضی

پایان‌نامه کارشناسی ارشد

یک مدل منعطف برای پیش‌گویی فضایی پاسخ‌های نرخ به کمک توزیع بتا ـ دوجمله‌ای

نگارنده: لیلا عابدین پور لیارجدمه

استاد راهنما

دکتر حسین باغیشنی

استاد مشاور

دکتر نگار اقبال

بهمن ۱۳۹۶

شماره:
تاریخ:

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم **لیلا عابدین پور لیارجدمه** با شماره دانشجویی **۹۴۱۱۷۲۴** رشته آمار گرایش آمارریاضی تحت عنوان **یک مدل منعطف برای پیش‌گویی فضایی پاسخ‌های نرخ به کمک توزیع بتا-دوجمله‌ای** که در تاریخ **۹۶/۱۱/۹** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

قبول (با درجه: عالی) ☒ **مردود** ☐

نوع تحقیق: نظری ■ **عملی** ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱- استادراهنمای اول	دکتر حسین باغیشنی	استادیار	
۲ـ استادراهنمای دوم			
۳- استاد مشاور	دکتر نگار اقبال	استادیار	غایب
۴- نماینده تحصیلات تکمیلی	دکتر احمد نزاکتی رضازاده	دانشیار	
۵- استاد ممتحن اول	دکتر داود شاهسونی	دانشیار	
۶-ـ استاد ممتحن دوم	دکتر محمد آرشی	دانشیار	

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده: ۹۶/۱۱/۹

دانشکده ریاضی

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می‌تواند پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم به:

پدر و مادر عزیزم

به پاس عاطفه سرشار و گرمای امید بخش وجودشان

و خواهر و برادر مهربانم

به پاس محبت های بی دریغ شان

و همه آنان که اندیشیدن را به من آموختند، نه اندیشه ها را

سپاس گزاری...

سپاس خداي را كه سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت‌هاي او ندانند و كوشندگان، حق او را گزاردن نتوانند. سلام و دورد بر محمّد و خاندان پاك او، طاهران معصوم، هم آنان كه وجودمان وام‌دار وجودشان است و نفرين پيوسته بر دشمنان ايشان تا روز رستاخيز...

بدون شک جایگاه و منزلت معلم، اجّل از آن است که در مقام قدردانی از زحمات بی‌شائبه او، با زبان قاصر و دست ناتوان، چيزي بنگاريم. اما از آن‌جایی که تجلیل از معلم، سپاس از انسانی است که هدف و غایت آفرینش را تامین می‌کند، وظیفه خود می‌دانم از زحمات بی‌دریغ استاد راهنمای خود، جناب آقای دکتر حسین باغیشنی که همواره با شکیبایی و درایت فراوان، مرا مستفیض راهنمایی‌های بی‌دریغ خود از لحاظ علمی و اخلاقی نمودند، صمیمانه تشکر و قدردانی نمایم که قطعا بدون راهنمایی‌های ارزنده ایشان این مجموعه به انجام نمی‌رسید. از استاد گرامی، سرکار خانم دکتر نگار اقبال، که زحمت مشاوره این پایان‌نامه را تقبل فرمودند، کمال تشکر را دارم.

از اساتید محترم آقایان دکتر شاهسونی، دکتر آرشی، دکتر ربیعی و دکتر نزاکتی، که در زمان تحصیل از محضرشان سود برده و از تجربیات ایشان استفاده کردم، سپاسگزاري می‌نمایم و امیدوارم همواره در پناه لطف الهی محفوظ و موفق باشند.

همچنین تشکر و قدردانی فراوان خود را از خانواده عزیزم که با حمایت‌های خویش، همواره مشوق من بوده‌اند ابراز می‌نمایم.

در پایان نیز از دانشگاه علوم پزشکی گیلان و مرکز آموزشی درمانی رازی رشت جهت ارایه داده‌های مربوط به سرطان معده تشکر و قدردانی می‌کنم.

لیلا عابدین پور لیارجدمه
بهمن ۱۳۹۶

تعهد نامه

اینجانب **لیلا عابدین پور لیارجدمه** دانشجوی کارشناسی ارشد رشته **آمار علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **یک مدل منعطف برای پیش‌گویی فضایی پاسخ‌های نرخ به کمک توزیع بتا- دوجمله‌ای** ، تحت راهنمایی دکتر **حسین باغیشنی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده‌اند.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام '' دانشگاه صنعتی شاهرود '' یا " Shahrood University of Technology " به چاپ خواهند رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده‌اند.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

لیلا عابدین پور لیارجدمه
بهمن ۱۳۹۶

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته‌شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نیست.

چکیده

مدلی که به‌طور معمول برای برازش پاسخ‌های شمارشی فضایی در جوامع متناهی مورد استفاده قرار می‌گیرد، مبتنی بر توزیع دوجمله‌ای است. در مباحث کاربردی، به‌دلیل وجود بیش‌پراکنش معمول در داده‌ها، استفاده از این پذیره توزیعی چندان مناسب نیست و می‌تواند به استنباط ناکارا در برآورد پارامترها و پیش‌گویی فضایی منتهی شود. به‌منظور رفع نقایص مدل دوجمله‌ای، یک رهیافت جانشین را برای مدل‌سازی نسبت‌های همبسته فضایی مبتنی بر توزیع بتا_دوجمله‌ای معرفی می‌کنیم. انعطاف‌پذیری توزیع بتا_دوجمله‌ای می‌تواند آن را به گزینه منتخب برای برازش پاسخ‌های بیش‌پراکنده در این جوامع تبدیل کند. در این پایان‌نامه، به پیش‌گویی فضایی با استفاده از مدل منعطف بتا_دوجمله‌ای و تعمیم آن برای تحلیل فضایی داده‌های مشبکه‌ای در چارچوب استنباط بیزی می‌پردازیم. همچنین با مطالعه شبیه‌سازی، کارایی آن را نمایش می‌دهیم. برای نمایش عملکرد این مدل‌ها، دو مجموعه داده بارندگی استان سمنان و تعداد افراد مبتلا به سرطان معده در استان گیلان را تحلیل می‌کنیم.

کلمات کلیدی: داده‌های شمارشی فضایی، فرآیند بتا_دوجمله‌ای، بیش‌پراکنش، پیش‌گویی فضایی.

فهرست مقاله‌های مستخرج از پایان‌نامه

۱. عابدین پور، ل.، باغیشنی، ح و اقبال، ن.، (۱۳۹۶)، ''پیش‌گویی فضایی در فرآیندهای بتا_دوجمله‌ای''، مجموعه مقالات یازدهمین سمینار احتمال و فرآیندهای تصادفی، ۸۷-۹۷، قزوین.

۲. عابدین پور، ل.، باغیشنی، ح و اقبال، ن.، (۱۳۹۶)،''پیش‌گویی فضایی پاسخ‌های نرخ: مقایسه مدل‌های کریگینگ بتا_دوجمله‌ای و هم‌کریگینگ دوجمله‌ای''،مجموعه مقالات دومین سمینار آمار فضایی و کاربردهای آن، (۱۷۹-۱۶۷)، شاهرود.

۳. عابدین پور، ل.، باغیشنی، ح و اقبال، ن.، (۱۳۹۶)، ''پیش‌گویی نرخ بارندگی در استان سمنان با مدل بتا_دوجمله‌ای فضایی''،مجموعه مقالات دومین کنفرانس ملی و اولین کنفرانس بین‌المللی محاسبات نرم، (۱۰۸۷-۱۰۸۱)، رودسر.

۴. مدل اتوبتا_دوجمله‌ای برای تحلیل نرخ سرطان معده در استان گیلان، مجله دانشگاه علوم پزشکی گیلان، ارسال شده.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

داده‌های شمارشی همبسته فضایی

۱.۱ مقدمه

در بسیاری از علوم مانند علوم محیطی، پزشکی، ژنتیک و زمین‌شناسی با داده‌هایی مواجه می‌شویم که مستقل از یکدیگر نیستند و وابستگی آن‌ها ناشی از موقعیتشان در فضای مورد مطالعه است، به‌طوری که ساختار وابستگی آن‌ها تابعی از فاصله مشاهدات از یکدیگر است. این‌گونه داده‌ها، داده‌های فضایی[۱] نامیده می‌شوند. به‌عنوان مثال‌هایی از این نوع داده‌ها می‌توان میزان بروز یک بیماری در مناطق مختلف، غلظت گازهای آلاینده هوا در میدان‌های مختلف یک شهر و درصد خلوص طلا در نقاط مختلف یک معدن را نام برد.

دو ویژگی مهم داده‌های فضایی، نمایش هر داده با موقعیت آن در فضای مورد مطالعه و وابستگی فضایی این داده‌هاست. معمولا داده‌های فضایی که از موقعیت‌های مجاور جمع‌آوری می‌شوند، وابستگی بیش‌تری دارند و با افزایش فاصله بین موقعیت داده‌ها، این وابستگی کاهش می‌یابد. بدیهی است که به دلیل عدم برقراری شرط استقلال در داده‌های فضایی، تحلیل آماری این گونه داده‌ها با روش‌های آماری معمول مقدور نیست. از این رو، شاخه آمار فضایی برای تحلیل این گونه مشاهدات پدید آمده و در حال توسعه می‌باشد. در حقیقت آمار فضایی شاخه‌ای از علم آمار است که به تحلیل داده‌های فضایی می‌پردازد. این شاخه از آمار می‌کوشد بین مقادیر مختلف یک متغیر و فاصله و جهت قرارگیری آن‌ها نسبت به هم ارتباطی برقرار کند. این ارتباط

[۱] Spatial data

فضایی، ساختار فضایی[۲] نام دارد. از جمله مباحث مهم در تحلیل داده‌های فضایی، بررسی و مطالعه ساختار وابستگی فضایی و برآورد آن با استفاده از مشاهدات و همچنین پیش‌گویی در موقعیت‌های فاقد مشاهده و تعیین خطای پیش‌گویی است.

اگرچه گسترش آمار فضایی اغلب توسط آماردانان صورت گرفته است، اما پیدایش و کاربرد عملی این شاخه از آمار صرفا از طریق آماردانان انجام نشده است. توسعه این شاخه به وسیله زمین‌شناسان تحت عنوان زمین‌آمار[۳] صورت گرفته است. به همین سبب واژه‌پردازی و اصطلاحات آن با ادبیات آمار کلاسیک قدری متفاوت است. نخستین بار به دنبال روند تکاملی ذخایر معدنی که از قبل از سال ۱۹۶۰ آغاز شده بود، ماترون با انتشار مقاله بنیادین خود در سال ۱۹۶۲ توانست شاخه جدیدی از علم آمار تحت عنوان زمین‌آمار را پایه‌گذاری کند. زمین‌آمار به شاخه‌ای از علم آمار گفته می‌شود که خواص داده‌های فضایی را تحلیل می‌کند. در این شاخه، ناحیه جغرافیایی یک ناحیه چگال[۴] است، به این معنی که می‌توان بین هر دو موقعیت مکانی یک مکان دیگر را در نظر گرفت.

در شکل‌گیری آمار فضایی افراد مختلفی سهیم بوده‌اند و به‌طور همزمان با ماترون فعالیت داشته‌اند. از جمله گاندین (۱۹۶۳) بحثی به نام تحلیل عینی[۵] را با مفاهیم بسیار نزدیک به تعاریف ماترون و دیگران در زمینه هواشناسی پیشنهاد کرد و از این طریق روش‌هایی برای تحلیل داده‌های فضایی ابداع نمود. ماترون (۱۹۶۳) پیش‌گویی فضایی کریگینگ[۶]، که منتسب به دی.جی.کریگ است را مطرح نمود. کریگ یک مهندس معدن آفریقای جنوبی بود که در معادن طلا کار می‌کرد و به تجربه فهمید که از اطلاعات بلوک‌های مجاور هم می‌توان برای پیش‌گویی عیار طلا استفاده کرد. دیوید (۱۹۷۷) از زمین‌آمار برای برآورد منابع ذخیره طلا استفاده کرد. جورنل و هیوجبرگ (۱۹۷۸) به طور مفصل روش‌های زمین‌آماری را با ادبیات خاص شرح دادند. ریپلی (۱۹۸۱) در کتابی تحت عنوان ’’ آمار فضایی’’ پیش‌گویی فرآیندهای تصادفی را مطرح کرد. کرسی (۱۹۹۳) با انتشار کتاب خود با عنوان ’’ آمار برای داده‌های فضایی’’، موجب پیشرفت گسترده این شاخه از آمار گردید.

داده‌های شمارشی همبسته فضایی در بسیاری از تحقیقات کاربردی مانند پزشکی، اقتصاد، زیست‌شناسی و علوم اجتماعی مورد توجه محققان قرار دارند. اولین روش گسترده تحلیل داده‌های فضایی استفاده از روش کریگینگ بود. کریگینگ یک روش درونیابی است که ویژگی کلیدی آن وابستگی به مکان می‌باشد. در این روش به مشاهدات نزدیک‌تر، به دلیل وابستگی بیش‌تر، وزن بزرگ‌تر و به مشاهدات دورتر وزن کوچک‌تری تخصیص داده می‌شود. در ۵۰ سال گذشته، از زمانی که کریگینک معرفی شد، تعمیم‌هایی به‌منظور بهبود بخشیدن به

---

[۲] Spatial structure
[۳] Geostatistics
[۴] Dense
[۵] Objective analysis
[۶] Kriging

پیش‌گویی‌ها مطرح شده‌اند، که این تعمیم‌ها شامل کریگینگ نشانگر[۷] (سولو، ۱۹۸۶؛کرسی، ۱۹۹۳)، کریگینگ معمولی[۸] (کرسی، ۱۹۸۸؛ جورنل و روسی، ۱۹۸۹)، کریگینگ عام[۹] (لیسوسکین و دیزایشکر، ۲۰۰۶؛ دیل زیمرمن و همکاران، ۱۹۹۹)، و هم‌کریگینگ[۱۰](گورث، ۱۹۹۸) می‌باشند.

اگرچه کریگینگ سال‌های زیادی پذیره این‌که مدل زیربنایی داده‌ها مانا[۱۱] (معمولا مانای مرتبه دوم) و یک فرآیند تصادفی گاوسی است را حفظ کرد، اما تحلیل داده‌های فضایی با این پذیره‌ها پیچده‌تر و با مشکل مواجه شدند. در بسیاری از موقعیت‌های کاربردی با داده‌هایی مواجه می‌شویم که نرمال نیستند. یک نگرانی اصلی، داده‌های شمارشی فضایی و نسبت‌های فضایی بود که هر دو اساسا غیر نرمال هستند. به‌همین دلیل دو رده مدل‌های آمیخته خطی تعمیم‌یافته[۱۲](GLMMs) و مدل‌های سلسله‌مراتبی بیزی در سال ۱۹۹۰ ظهور کردند. مدل‌های آمیخته خطی تعمیم‌یافته به‌عنوان تعمیمی از مدل‌های خطی تعمیم یافته[۱۳](GLMs) برای توزیع‌های مختلف معرفی شدند. گات وی و استروپ (۱۹۷۷) دریافتند که مدل‌های آمیخته خطی تعمیم یافته فضایی[۱۴] می‌توانند برای پیاده‌سازی ساختار همبستگی فضایی، با توزیع پاسخ غیرنرمال استفاده شوند. تاکنون مدل‌های گوناگونی برای تحلیل داده‌های فضایی غیرنرمال معرفی شده‌اند. نخستین مدل هم‌کریگینگ دوجمله‌ای، توسط لاجونی (۱۹۹۱) برای پیش‌گویی نرخ فضایی بر اساس نسبت نمونه‌ای مشاهده‌شده از یک بیماری نادر، پیشنهاد شد. الیور و همکاران (۱۹۹۸)، از این مدل برای تحلیل نرخ سرطان کودکان استفاده کردند. مونستیز و همکاران (۲۰۰۶) استفاده از مدل کریگینگ پواسون را برای داده‌های شمارشی همبسته فضایی پیشنهاد دادند.

در مواردی که ساختار ناهمگن موجود در داده‌های شمارشی یک جامعه متناهی فضایی موجب بیش‌پراکندگی می‌شود، عدم انعطاف توزیع دوجمله‌ای استفاده از آن را محدود می‌کند. برای مرتفع کردن این مساله از رهیافت جانشین آن مبتنی بر توزیع بتا_دوجمله‌ای استفاده می‌کنیم. انعطاف‌پذیری توزیع بتا_دوجمله‌ای باعث شده است که این توزیع در برازش مجموعه داده‌های بیش‌پراکنده سهم به‌سزایی را ایفا کند. در این پایان‌نامه، به پیش‌گویی فضایی با استفاده از توزیع بتا_دوجمله‌ای می‌پردازیم که انعطاف توزیع بتا_دوجمله‌ای موجب می‌شود تا به نتایج دقیق‌تری در استنباط برسیم. در ادامه، ابتدا به معرفی برخی از مفاهیم کلیدی آمار فضایی می‌پردازیم. بخش‌هایی از این فصل از کتاب محمدزاده (۱۳۸۹) با عنوان ”آمار فضایی و کاربردهای آن“، برداشت شده است.

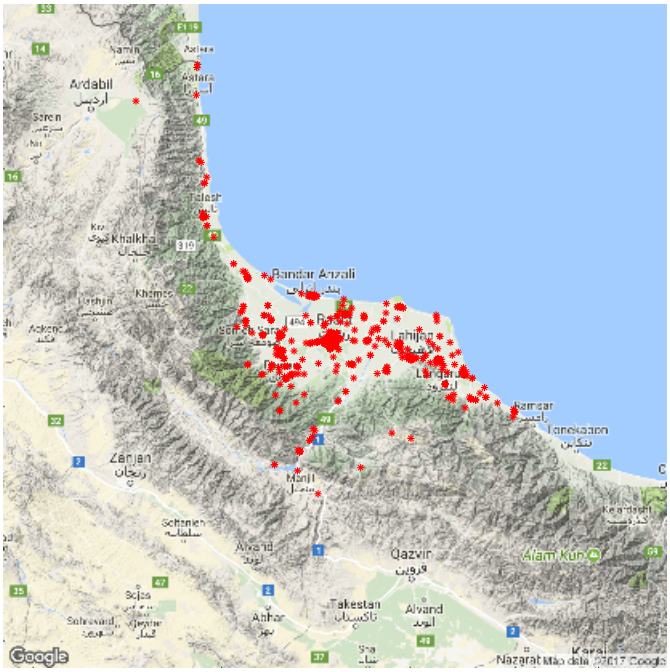
---

[۷] Indicator kriging
[۸] Ordinary kriging
[۹] Universal kriging
[۱۰] Cokriging
[۱۱] Stationary
[۱۲] Generalized linear mixed models
[۱۳] Generalized linear models
[۱۴] Spatial generalized linear models

۱.۱.۱ داده‌های فضایی

داده‌های فضایی مشاهداتی هستند که وابستگی آن‌ها ناشی از موقعیت و مکان قرار گرفتن آن‌ها در فضای (جغرافیایی) مورد مطالعه است. داده‌های فضایی بر اساس این‌که مکان‌های فضایی و متغیرهای تصادفی مرتبط به چه صورت باشند، به سه گروه داده‌های زمین‌آماری[۱۵]، داده‌های مشبکه‌ای[۱۶] و الگوهای نقطه‌ای[۱۷] تقسیم می‌شوند.

داده‌های زمین‌آماری: این نوع داده‌ها در موقعیت‌های ثابت و مشخص در ناحیه‌ای چگال مشاهده می‌شوند. یعنی بین دو موقعیت مفروض، ممکن است موقعیت‌های دیگری هم وجود داشته باشند. متغیر مورد بررسی ممکن است گسسته یا پیوسته باشد. به‌عنوان مثال می‌توان به مقدار ریزش باران ثبت‌شده در ایستگاه‌های هواشناسی، تعداد نوعی جانور دریایی در یک مجموعه از مکان‌های نمونه‌گیری‌شده و محل سکونت افراد مبتلا به سرطان معده در استان گیلان (شکل ۱.۱) اشاره کرد. هدف اصلی از تحلیل این گونه داده‌ها، پیش‌گویی پاسخ در مکان‌های مشاهده‌نشده است.



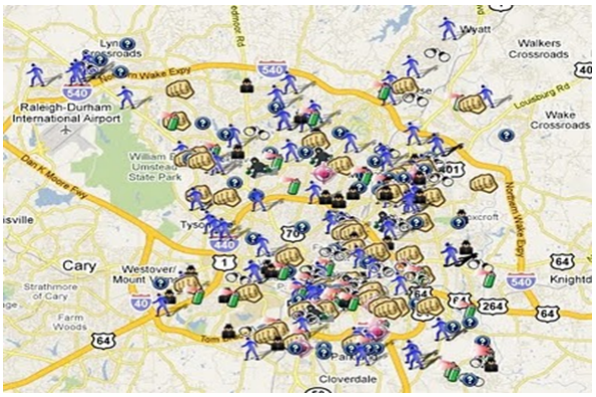
شکل ۱.۱: محل سکونت افراد مبتلا به سرطان معده در استان گیلان

داده‌های مشبکه‌ای: این نوع داده‌ها مربوط به نواحی غیرچگال هستند، یعنی بین مکان‌های مشاهده‌شده مکان دیگری وجود ندارد. این مکان‌ها ممکن است منظم یا نامنظم باشند. تعداد افراد مبتلا به سرطان معده در شهرهای مختلف استان گیلان و تعداد حوادث رانندگی

[۱۵] Geostatistical data
[۱۶] Lattice data
[۱۷] Point patterns

یا نرخ رشد اقتصادی هر استان مثال‌هایی از داده‌های مشبکه‌ای هستند. در تحلیل داده‌های مشبکه‌ای هدف عمده مدل‌بندی احتمالاتی مشاهدات است.

الگوهای نقطه‌ای: در این نوع داده‌ها موقعیت مشاهده‌شده، خود یک متغیر تصادفی است. این داده‌ها به سه دسته به‌طور کامل تصادفی فضایی[۱۸](CSR)، منظم[۱۹] و خوشه‌ای[۲۰] تقسیم و به مدل‌بندی آن‌ها اقدام می‌شود. به‌عنوان مثال می‌توان به موقعیت گونه‌ای از درختان در یک ناحیه جنگلی و محل تخلفات رانندگی در یک مسابقه رالی (شکل ۲.۱) اشاره کرد.



شکل ۲.۱: محل تخلفات رانندگی در یک مسابقه رالی

۲.۱.۱ مدل‌های آماری فضایی: میدان تصادفی

برای تحلیل داده‌های فضایی ابتدا باید یک مدل آماری در نظر گرفته شود. در آمار فضایی معمولا مدل‌بندی داده‌ها با تعریف یک میدان تصادفی[۲۱] صورت می‌گیرد. بنابراین ابتدا به معرفی میدان تصادفی و برخی از ویژگی‌های آن که در ادامه مورد استفاده قرار می‌گیرند، می‌پردازیم.

تعریف ۱.۱.۱. (میدان تصادفی). میدان تصادفی مجموعه‌ای از متغیرهای تصادفی مانند

$$Z(\cdot) = \left\{ Z(s); s \in D \subseteq R^d, d \geqslant ۱ \right\}$$

است که در آن $z(s)$ متغیر تصادفی در اندیس s و D یک مجموعه اندیس‌گذار است.

تعریف ۲.۱.۱. (توابع میانگین، واریانس و کوواریانس میدان تصادفی). برای میدان تصادفی $Z(\cdot)$ میانگین در موقعیت s و کوواریانس در موقعیت‌های $s_۱$ و $s_۲$ به‌ترتیب به صورت

$$\mu(s) = E\left[Z(s)\right], s \in D$$

$$C(s_۱, s_۲) = Cov\left[Z(s_۱), Z(s_۲)\right] = E\left[\left(Z(s_۱) - \mu(s_۱)\right)\left(Z(s_۲) - \mu\left(s_۲\right)\right)\right], s_۱, s_۲ \in D$$

[۱۸] Complete spatial randomness
[۱۹] Regular
[۲۰] Cluster
[۲۱] Random field

تعریف می‌شوند. برای $s_1 = s_2 = s$، واریانس میدان تصادفی $Z(\cdot)$ در مکان s به‌صورت

$$Var\left[Z(s)\right] = E\left[Z\left(s\right) - \mu\left(s\right)\right]^2 = C\left(s,s\right)$$

حاصل می‌شود. هر میدان تصادفی $Z(\cdot)$ را می‌توان به‌صورت

$$Z(s) = \mu\left(s\right) + \delta\left(s\right), s \in D$$

تجزیه کرد که در آن $\mu(\cdot)$ تغییرات بزرگ‌مقیاس[۲۲] یا روند و $\delta(\cdot)$ فرآیند خطا یا تغییرات کوچک‌مقیاس[۲۳] میدان تصادفی نامیده می‌شوند. تغییرات کوچک‌مقیاس ممکن است از خطای اندازه‌گیری و تغییرات بزرگ‌مقیاس از تغییرپذیری بین موقعیت‌های مشاهده‌شده ناشی شده باشند.

به طور معمول، تحلیل داده‌های فضایی بر اساس مشاهدات نمونه دشوار است، اما در نظر گرفتن برخی ویژگی‌های میدان تصادفی موجب ساده‌تر شدن مساله خواهند شد که در ادامه به معرفی آن‌ها می‌پردازیم.

تعریف ۳.۱.۱. (مانای ذاتی). میدان تصادفی $Z\left(\cdot\right)$ را مانای ذاتی[۲۴] گوییم هرگاه

۱. میانگین میدان تصادفی ثابت یا مستقل از s باشد، یعنی $E\left(Z\left(s\right)\right) = \mu$

۲. اختلاف واریانس در دو نقطه s_1 و s_2 فقط تابعی از فاصله موقعیت‌های s_1 و s_2 باشد. یعنی

$$Var\left(Z\left(s_1\right) - Z\left(s_2\right)\right) = 2\gamma\left(s_1 - s_2\right), s_1, s_2 \in D.$$

که در آن $\gamma(.)$ تابعی از نوع فاصله است.

تعریف ۴.۱.۱. (مانای مرتبه دوم). میدان تصادفی $Z\left(\cdot\right)$ را مانای مرتبه دوم[۲۵] گوییم هرگاه علاوه بر مانایی در میانگین، یعنی $E\left(Z\left(s\right)\right) = \mu$، کوواریانس $Z(s_1)$ و $Z(s_2)$ فقط تابعی از فاصله موقعیت‌های s_1 و s_2 باشد، یعنی

$$Cov(Z(s_1), Z(s_2)) = C(s_1 - s_2), \qquad s_1, s_2 \in D \subset R^d.$$

بدیهی است که تحت مانایی مرتبه دوم

$$Var(Z(s)) = Cov\left(Z\left(s\right), Z\left(s\right)\right) = C(0) = \sigma^2$$

یعنی واریانس میدان تصادفی به موقعیت فضایی وابسته نیست و تغییرپذیری میدان در همه‌جا یکسان است.

---

[۲۲] Large scale variation
[۲۳] Small scale variation
[۲۴] Intrinsic stationary
[۲۵] Second order stationary

تعریف ۵.۱.۱. (مانای قوی). میدان تصادفی $Z(\cdot)$ مانای قوی نامیده می‌شود، هرگاه برای تمامی موقعیت‌های $s_1, \cdots, s_n$ و $h \in R^d$، $(Z(s_1), \ldots, Z(s_n)) \stackrel{D}{=} (Z(s_1+h), \ldots, Z(s_n+h))$ باشد. به بیان دیگر در صورت انتقال موقعیت‌های $s_1, \cdots, s_n$ در راستای h، توزیع توام تغییر نمی‌کند.

تعریف ۶.۱.۱. (میدان تصادفی گاوسی). میدان تصادفی $Z(\cdot)$ گاوسی نامیده می‌شود، هرگاه برای هر $m \geq 1$ توزیع توام $(Z(s_1), \cdots, Z(s_n))$ نرمال چندمتغیره باشد.

قبل از ورود به بخش بعدی، با توجه به استفاده از نرم اقلیدسی در ادامه، ابتدا این تابع را تعریف می‌کنیم.

تعریف ۷.۱.۱. (تابع نرم اقلیدسی). فرض کنید $1 \leq p < \infty$ باشد، تابع $p-$نرم، برای بردار n بعدی $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ از مقادیر حقیقی، به‌صورت زیر تعریف می‌شود:

$$||x||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

اگر مقدار $p = 2$ باشد، نرم حاصل را اقلیدسی می‌نامند.

۳.۱.۱ ساختار همبستگی فضایی

در آمار فضایی، مشابه مفاهیم واریانس و کواریانس در آمار کلاسیک، ساختار همبستگی فضایی داده‌ها توسط تابع تغییرنگار[۲۶] و هم‌تغییرنگار[۲۷] تعیین و در تحلیل داده‌های فضایی مورد استفاده قرار می‌گیرد، که به‌ترتیب به تعریف آن‌ها می‌پردازیم.

تعریف ۸.۱.۱. (تغییرنگار). واریانس تفاضل بین مقادیر میدان تصادفی در دو موقعیت s و $s+h$ را تغییرنگار می‌نامند که به‌صورت

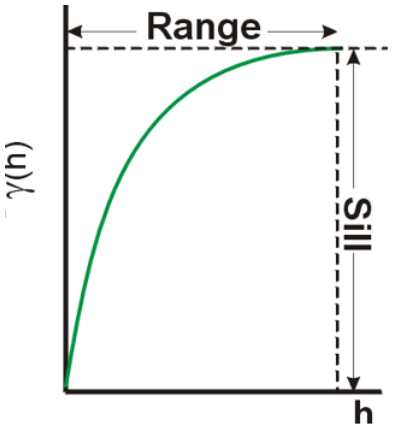
$$2\gamma(h) = Var(Z(s+h) - Z(s)) \qquad (1.1)$$

تعریف می‌شود. تابع $\gamma(h)$ نیم‌تغییرنگار[۲۸] نامیده می‌شود.

کوچک بودن تغییرنگار، نشانگر وابستگی زیاد و بزرگ بودن آن، نشانگر وابستگی کم میدان تصادفی است. تغییرنگار را می‌توان به‌عنوان معیاری برای نشان دادن تاثیرگذاری کمیت یک موقعیت بر کمیت موقعیت‌های دیگر در فضای مورد مطالعه به‌کار برد. نمودار تابع تغییرنگار در شکل ۳.۱ نشان داده شده است. اگر تغییرنگار برای تمام مقادیر h ثابت

[۲۶] Variogram
[۲۷] Covariogram
[۲۸] Semi-variogram

باشد، داده‌ها همبستگی فضایی ندارند یا همبستگی بین آن‌ها بسیار ضعیف است. برعکس شیب نمودار تغییرنگار بیانگر همبستگی فضایی داده‌هاست. تغییرنگار سه پارامتر دامنه[۲۹]، ازاره[۳۰] و اثرقطعه‌ای[۳۱] دارد که به‌ترتیب آن‌ها را تعریف می‌کنیم.



شکل ۳.۱: یکی از حالت‌های تابع تغییرنگار

تعریف ۹.۱.۱. (دامنه). فاصله‌ای که بعد از آن همبستگی داده‌های فضایی ناچیز و حتی مستقل از یکدیگر می‌شود، دامنه تغییرنگار نامیده می‌شود. تغییرنگار از این فاصله به بعد تغییر چندانی ندارد و به‌صورت تخت در می‌آید. از این رو داده‌های واقع در خارج از دامنه با روش‌های کلاسیک آماری قابل تحلیل هستند.

تعریف ۱۰.۱.۱. (ازاره). در صورتی که تغییرنگار به کران بالایی منتهی شود، آن کران ازاره تغییرنگار نامیده می‌شود و به‌صورت c در زیر تعریف می‌شود:

$$\lim_{||h||\to\infty} \gamma(h) = c$$

تعریف ۱۱.۱.۱. (اثر قطعه‌ای). مقدار تغییرنگار در مبدا مختصات اثر قطعه‌ای نامیده می‌شود و به‌صورت c_0 در

$$\lim_{h\to 0} \gamma(h) = c_0$$

تعریف می‌شود. اثر قطعه‌ای معمولا به دلیل خطاهای اندازه‌گیری یا تغییرات شدید کمیت مورد بررسی در موقعیت‌های نزدیک به هم ظاهر می‌شود.

در رابطه با تغییرنگار می‌توان قضیه زیر را عنوان کرد.

[۲۹] Range
[۳۰] Sill
[۳۱] Range

قضیه ۱.۱.۱. . برای هر میدان تصادفی مانای ذاتی، تابع تغییرنگار همیشه منفی شرطی است، یعنی برای هر دنباله اعداد ثابت $\{a_i\}_{i=1}^m$ که $\sum_{i=1}^m a_i = 0$ ، $\sum_{i=1}^m \sum_{j=1}^m a_i a_j \gamma(s_i - s_j) \leq 0$ است (کرسی، ۱۹۹۳).

تعریف ۱.۱.۱۲. (هم تغییرنگار). کوواریانس دو متغیر $Z(s)$ و $Z(s+h)$ که به‌صورت

$$C(h) = Cov(Z(s), Z(s+h))$$

تعریف می‌شود را هم‌تغییرنگار گویند. اگر میدان تصادفی مانای مرتبه دوم باشد، آن‌گاه هم‌تغییرنگار به‌صورت $C(h) = E[Z(s)Z(s+h)] - \mu^2$ ساده می‌شود. علاوه‌براین رابطه

$$\gamma(h) = C(0) - C(h)$$

بین تغییرنگار و هم‌تغییرنگار برقرار است، زیرا

$$\begin{aligned} 2\gamma(h) &= Var[Z(s+h) - Z(s)] \\ &= Var(Z(s+h)) + Var(Z(s)) - 2Cov[Z(s), Z(s+h)] \\ &= C(0) + C(0) - 2C(h). \end{aligned}$$

تابع هم‌تغییرنگار، تابعی معین مثبت است.

تعریف ۱.۱.۱۳. (همبستگی‌نگار). ضریب همبستگی بین $Z(s)$ و $Z(s+h)$ با شرط $C(0) > 0$ همبستگی‌نگار نام دارد و به‌صورت زیر تعریف می‌شود:

$$\rho(s, s+h) = \rho(h) = \frac{C(h)}{C(0)} = 1 - \frac{\gamma(h)}{C(0)}.$$

در حالت کلی توابع برداری تغییرنگار، کوواریانس و همبستگی‌نگار به اندازه و جهت بردار $h \in R^d$ بستگی دارند، اما اگر این سه فقط تابعی از $||h||$، یعنی نرم h، در فضای R^d باشند و به جهت h بستگی نداشته باشند، آن‌ها را به‌ترتیب تغییرنگار، کوواریانس و همبستگی‌نگار همسان‌گرد[۳۲] نامند و در غیر این صورت ناهمسان‌گرد نامیده می‌شوند.

تعریف ۱.۱.۱۴. (میدان تصادفی همسان‌گرد). میدان تصادفی $Z(\cdot)$ همسان‌گرد نامیده می‌شود، هرگاه توابع تغییرنگار، کوواریانس یا همبستگی‌نگار آن همسان‌گرد باشند.

مدل‌های معتبر تغییرنگار

توابعی که در شرط همیشه منفی شرطی صدق می‌کنند را مدل‌های معتبر تغییرنگار گویند. جورنل و هویج‌برگتس (۱۹۷۸) مدل‌های پارامتری مختلفی را معرفی کرده‌اند که از آن جمله می‌توان به مدل‌های تغییرنگار گاوسی، نمایی، کروی، توانی، لگاریتمی، موجی، سهمی‌گون و ماترون اشاره کرد. در ادامه دو مدل گاوسی و کروی را معرفی می‌کنیم.

[۳۲] Isotropic

۱. مدل گاوسی: تابع نیم‌تغییرنگار گاوسی به‌صورت

$$\gamma(h) = c_0 + c\left(1 - e^{\frac{-||h||}{a}}\right), h \in R^d, d \geq 1$$

تعریف می‌شود. به‌طوری که در آن پارامترهای a، c_0 به‌ترتیب دامنه و اثر قطعه‌ای و $c + c_0$ نیز ازاره است.

۲. مدل کروی: تابع نیم‌تغییرنگار کروی به‌صورت

$$\gamma(h) = \begin{cases} c_0 + c\left(\frac{3}{2}\frac{||h||}{a} - \frac{1}{2}\frac{||h||^3}{a^3}\right) & 0 < ||h|| < a, h \in R^d, d = 1,2,3 \\ c + c_0 & ||h|| > a \end{cases}$$

می‌باشد، که دارای سه پارامتر a، c و c_0 است.

۲.۱ پیش‌گویی فضایی

یکی از مسایل مهم در آمار فضایی، پیش‌گویی مقدار نامعلوم میدان تصادفی در موقعیت فاقد مشاهده s_0 بر اساس مشاهدات $Z = (Z(s_1), \cdots, Z(s_n))'$ است. معمولا با پذیرفتن مانایی میدان تصادفی، معلوم بودن ساختار همبستگی فضایی و در نظر گرفتن تابع زیان توان دوم خطا، پیش‌گوی بهینه از می‌نیمم کردن میانگین شرطی $\hat{Z}(s_0) = E(Z(s_0)|Z)$ به‌دست می‌آید. بدون در نظر گرفتن پذیره‌های ساده‌کننده، محاسبه این میانگین شرطی از یک توزیع $(n+1)$ متغیره $(Z(s_0), Z')$ مقدور نیست. پذیره‌ای که معمولا در نظر گرفته می‌شود، گاوسی بودن میدان تصادفی است. در این صورت توزیع توام $(Z(s_i), Z')$ نیز گاوسی است. همچنین پذیره گاوسی بودن میدان موجب می‌شود که پیش‌گوی $E(Z(s_0)|Z)$ نیز خطی شود و فقط به مقادیر $\mu(s_i) = E(Z(s_i))$ و $C(s_i, s_j) = Cov(Z(s_i), Z(s_j))$ وابسته باشد، که در آن $i, j = 1, \cdots, n$. علاوه بر این، پذیره گاوسی بودن، هم‌واریانسی شرطی، یعنی عدم وابستگی $Var(Z(s_0)|Z)$ به Z، را نیز فراهم می‌سازد. در این صورت تغییرات پیش‌گوی $E(Z(s_0)|Z)$ که با میانگین توان دوم خطای شرطی اندازه‌گیری می‌شود، از میانگین توان دوم خطای غیرشرطی، $E\left(Z(s_0) - \hat{Z}(s_0)\right)^2$، نامتمایز شده و محاسبه آن آسان می‌شود.

در آمار فضایی تعیین بهترین پیش‌گوی خطی نااریب برای $Z(s_0)$ کریگینگ نام دارد، که توسط ماترون (۱۹۶۳) به افتخار دی. جی. کریگ نام‌گذاری شده است. به‌عبارت دیگر کریگینگ یک روش پیش‌بینی در آمار فضایی به ویژه در زمین‌آمار می‌باشد، که معادل بهترین پیش‌گوی نااریب خطی است. در این روش برای پیش‌گویی در یک موقعیت مشخص، به مشاهدات نزدیک‌تر وزن بزرگ‌تر و به مشاهدات دورتر وزن کوچک‌تری داده می‌شود. هم‌چنین وزن‌ها به‌گونه‌ای انتخاب می‌شوند که واریانس پیش‌گو کمینه شود. ویژگی دیگر کریگینگ این

است که در هر موقعیت فضایی واریانس پیش‌گو ارایه می‌شود و می‌توان توزیع خطاها را در فضای مورد بررسی به‌دست آورد.

فرض کنید پیش‌گویی $Z(s_0)$ بر اساس مشاهدات $Z(s_1),\cdots,Z(s_n)$ در موقعیت‌های $s_1,\cdots,s_n$ مورد نظر باشد. پیش‌گوی مکانی به‌صورت ترکیب خطی $\hat{Z}(s_0)=\sum_{i=1}^{n}\lambda_i Z(s_i)+k$ است، که در آن ضرایب λ_i و k طوری تعیین می‌شوند که $\hat{Z}(s_0)$ ناااریب و دارای کم‌ترین واریانس در بین همه پیش‌گوهای خطی ناااریب باشند. چنانچه میانگین میدان تصادفی $Z(s)$ ثابت، معلوم یا نامعلوم باشد، شرط ناااریبی کریگینگ مستلزم آن خواهد بود که $k=0$ و $\sum_{i=1}^{n}\lambda_i=1$ باشند. بنابراین بسته به این‌که میانگین میدان تصادفی معلوم، ثابت نامعلوم یا تابعی از موقعیت‌ها باشد، کریگینگ به انواع کریگینگ ساده، معمولی و عام دسته‌بندی می‌شود، که در ادامه به معرفی آن‌ها می‌پردازیم.

۱.۲.۱ کریگینگ معمولی

کریگینگ معمولی توسط ماترون (۱۹۷۱)، تحت شرط‌های زیر معرفی شده است:

۱. میدان تصادفی به‌صورت

$$Z(s)=\mu+\delta(s) \qquad (۲.۱)$$

است، که در آن $\mu\in\Re$ مقداری است ثابت اما نامعلوم و $\delta(s)$ میدان تصادفی مانای ذاتی با میانگین صفر است.

۲. پیش‌گو برای $Z(s_0)$، خطی و به صورت

$$\hat{Z}(s_0)=\sum_{i=1}^{n}\lambda_i Z(s_i)$$

با قید $\sum_{i=}^{n}\lambda_i=1$ است، که در آن λ_iها وزن‌های کریگینگ و $Z(s_i)$ مشاهدات می‌باشند. شرط $\sum_{i=1}^{n}\lambda_i=1$ ناااریبی پیش‌گو را تضمین می‌کند، یعنی با استفاده از (۲.۱)

$$E\left[\hat{Z}(s_0)\right]=\mu$$
$$\sum\lambda_i EZ(s_i)=\mu\sum\lambda_i=\mu$$

بنابراین

$$\begin{aligned}
E\left[\hat{Z}(s_\circ) - Z(s_\circ)\right] &= E\left[\sum_{i=1}^{n} \lambda_i Z(s_i) - Z(s_\circ)\right] \\
&= E\left[\sum_{i=1}^{n} \lambda_i Z(s_i)\right] - E[Z(s_\circ)] \\
&= \sum_{i=1}^{n} \lambda_i E[Z(s_i)] - E(Z(s_\circ)) \\
&= \sum_{i=1}^{n} \lambda_i \mu - \mu \\
&= \circ
\end{aligned}$$

از طرفی پیش‌گوی $Z(s_\circ)$ بهینه است هرگاه توان دوم خطای پیش‌گو، یعنی

$$\sigma_e^2 = E\left[Z(s_\circ) - \hat{Z}(s_\circ)\right]^2$$

در رده همه پیش‌گوهای خطی $\sum_{i=1}^{n} \lambda_i Z(s_i)$، با قید $\sum_{i=1}^{n} \lambda_i = 1$ می‌نیمم باشد.

فرض کنید در مدل (۲.۱)، $E[\delta(\cdot)] = \circ$ و شرط مانایی مرتبه دوم برقرار و مدل دارای هم‌تغییرنگار معلوم $C(h)$ باشد. در این صورت توان دوم خطای پیش‌گو به‌صورت

$$\begin{aligned}
\sigma_e = E\left[Z(s_\circ) - \hat{Z}(s_\circ)\right]^2 &= E\left[Z(s_\circ) - \sum_{i=1}^{n} \lambda_i Z(s_i)\right]^2 \\
&= E\left[(Z(s_\circ) - \mu) - \left(\hat{Z}(s_\circ) - \mu\right)\right]^2 \\
&= E[Z(s_\circ) - \mu]^2 - 2E\left[(Z(s_\circ) - \mu)\left(\hat{Z}(s_\circ) - \mu\right)\right] + E\left[\hat{Z}(s_\circ) - \mu\right]^2 \\
&= C(\circ) - 2E\left[(Z(s_\circ) - \mu)\left(\sum_{i=1}^{n} \lambda_i Z(s_i) - \mu\right)\right] + \\
&\quad E\left[\left(\sum_{i=1}^{n} \lambda_i Z(s_i) - \mu\right)\left(\sum_{j=1}^{n} \lambda_i Z(s_j) - \mu\right)\right] \\
&= C(\circ) - 2\sum_{i=1}^{n} \lambda_i E[(Z(s_\circ) - \mu)(Z(s_i) - \mu)] \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j E[(Z(s_i) - \mu)(Z(s_j) - \mu)] \\
&= C(\circ) - 2\sum_{i=1}^{n} \lambda_i C(s_\circ, s_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i, s_j)
\end{aligned}$$

حاصل می‌شود. حال با اضافه کردن ضریب لاگرانژ تابع هدف A را تشکیل می‌دهیم:

$$A = \sigma_e + ۲\mu \left(\sum_{i=۱}^{n} \lambda_i - ۱ \right).$$

با استفاده از مشتق‌های جزئی خواهیم داشت

$$\begin{cases} \frac{\partial A}{\partial \lambda_i} & = ۰ - ۲C(s_۰, s_i) + ۲\sum_{i=۱}^{n} \lambda_j C(s_i, s_j) + ۲\mu = ۰. \\ \frac{\partial A}{\partial \mu} & = ۰ \end{cases}$$

بنابراین می‌توان دستگاه زیر را تشکیل داد:

$$\begin{cases} \sum_{i=۱}^{n} \lambda_j C(s_i, s_j) + \mu & = C(s_۰, s_i) \\ \sum_{i=۱}^{n} \lambda_i & = ۱ \end{cases}$$

دستگاه معادلات بالا را می‌توان به‌صورت ماتریسی زیر نوشت:

$$\begin{bmatrix} C(s_۱, s_۱) & C(s_۱, s_۲) & \dots & C(s_۱, s_n) & ۱ \\ C(s_۲, s_۱) & C(s_۲, s_۲) & \dots & C(s_۲, s_n) & ۱ \\ \vdots & \vdots & \ddots & & \\ C(s_n, s_۱) & C(s_n, s_۲) & \dots & C(s_n, s_n) & ۰ \end{bmatrix} \begin{bmatrix} \lambda_۱ \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} C(s_۰, s_۱) \\ C(s_۰, s_۲) \\ \vdots \\ C(s_۰, s_n) \end{bmatrix}.$$

پیش‌گو و واریانس به‌ترتیب به‌صورت

$$\hat{Z}(s_۰) = \lambda' Z$$
$$\sigma_k^۲(s_۰) = C(۰) - \lambda' C + \mu$$

نتیجه می‌شوند، که در آن‌ها بردار $\boldsymbol{\lambda} = (\lambda_۱, \cdots, \lambda_n)'$ و ضریب μ به‌صورت

$$\lambda' = \left[c + ۱_n \frac{\left(۱ - ۱_n' \Sigma^{-۱} c \right)}{۱_n' \Sigma^{-۱} ۱_n} \right]' \Sigma^{-۱}$$
$$\mu = -۱ \frac{\left(۱ - ۱_n' \Sigma^{-۱} c \right)}{۱_n' \Sigma^{-۱} ۱_n}$$

محاسبه می‌شوند به‌طوری که

$$\Sigma = \begin{bmatrix} C(s_۱, s_۱) & C(s_۱, s_۲) & \dots & C(s_۱, s_n) \\ C(s_۲, s_۱) & C(s_۲, s_۲) & \dots & C(s_۲, s_n) \\ \vdots & \vdots & \ddots & \\ C(s_n, s_۱) & C(s_n, s_۲) & \dots & C(s_n, s_n) \end{bmatrix}$$

$$c = (C(s_۰ - s_۱), \cdots, C(s_۰ - s_n))$$

و $\mathbf{1}_n$ برداری است که تمام n عنصر آن یک هستند.
می‌دانیم که بین تغییرنگار و هم‌تغییرنگار رابطه

$$\gamma(h) = C(\circ) - C(h)$$

برقرار است. بنابراین دستگاه معادلات بر حسب تغییرنگار به‌صورت زیر قابل بازنویسی است:

$$\begin{bmatrix} \gamma(s_1,s_1) & \gamma(s_1,s_2) & \dots & \gamma(s_1,s_n) & 1 \\ \gamma(s_2,s_1) & \gamma(s_2,s_2) & \dots & \gamma(s_2,s_n) & 1 \\ \vdots & \vdots & \ddots & & \\ \gamma(s_n,s_1) & \gamma(s_n,s_2) & \dots & \gamma(s_n,s_n) & \circ \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(s_\circ,s_1) \\ \gamma(s_\circ,s_2) \\ \vdots \\ \gamma(s_\circ,s_n). \end{bmatrix}$$

واریانس پیش‌گو نیز به‌صورت زیر به‌دست می‌آید:

$$\sigma^2_{s_\circ} = \sum_{i=1}^{n} \lambda_i \gamma(s_\circ - s_i) + \mu.$$

۲.۲.۱ کریگینگ ساده

کریگینگ ساده حالت خاصی از کریگینگ معمولی است، که در آن میانگین مقداری ثابت و معلوم است. بنابراین شرط ناااریبی $\sum_{i=1}^{n} \lambda_i$ نیاز نیست. لازم به ذکر است که در دنیای واقعی میانگین نامعلوم است و کریگینگ ساده کاربردی ندارد.

۳.۲.۱ کریگینگ عام

کریگینگ عام، تعمیم‌یافته کریگینگ معمولی است که توسط ماترون (۱۹۶۹)، به‌عنوان یک پیش‌گوی ناااریب، هنگامی که مقدار میانگین متغیر و نامعلوم است، معرفی شد.

۳.۱ مدل‌های خطی تعمیم‌یافته

مدلی که به طور معمول در رگرسیون مورد استفاده قرار می‌گیرد، مدل خطی[۳۳] (LM) است. در یک مدل خطی با متغیر پاسخ Y و ماتریس طرح X، تابع رگرسیونی، که همان میانگین شرطی $Y|\boldsymbol{X} = x$ است، یک ترکیب خطی از X برای $E(Y|X = x)$ به‌صورت $\boldsymbol{x'\beta}$ در نظر گرفته می‌شود. در این مدل، بردار β، ضرایب رگرسیونی نامعلوم یا به‌عبارتی اثرات ثابت[۳۴] نامیده می‌شوند. زمانی می‌توان از مدل‌های خطی استفاده کرد که تمام پذیره‌های آن برقرار باشد. یکی از این پذیره‌ها، توزیع نرمال برای متغیر پاسخ است. زمانی که این پذیره برای متغیر

[۳۳] Liner model
[۳۴] Fixed effects

پاسخ برقرار نباشد، نلدر و وردربرمن (۱۹۷۲) مدل‌های خطی تعمیم‌یافته را معرفی نمودند. مک‌کلا و نلدر (۱۹۸۹) مدل‌های خطی تعمیم‌یافته را برای مدل‌بندی متغیرهای پاسخ گسسته پیشنهاد دادند. در واقع این مدل‌ها تعمیمی از مدل‌های خطی بر اساس توزیع نرمال هستند که برای تحلیل پاسخ‌های عضو خانواده توزیع‌های نمایی گسترش یافته‌اند، که این تعمیم از دو قسمت تشکیل شده است. اول، توزیع‌هایی به‌جز توزیع نرمال در صورتی که متعلق به خانواده توزیع‌های نمایی باشند، می‌توانند توزیع پاسخ GLM را شامل شوند. دوم، میانگین پاسخ به‌طور مستقیم مدل‌سازی نمی‌شود بلکه با استفاده از تبدیلی از میانگین پاسخ به‌نام تابع پیوند[۳۵] به متغیرهای تبیینی رگرسیونی مرتبط می‌شود.

تعریف ۱.۳.۱. (خانواده توزیع‌های نمایی). خانواده‌ای را عضو خانواده توزیع‌های نمایی متعارف برای متغیر تصادفی Y گویند، هرگاه بتوان تابع چگالی احتمال آن را به‌صورت زیر نوشت:

$$f_Y(y) = \exp\left\{\frac{y\lambda - \psi(\lambda)}{a(\phi)} - c(y,\phi)\right\} \qquad (۳.۱)$$

به‌طوری که ϕ و λ پارامترهای خانواده توزیع و $c(\cdot)$، $\psi(\cdot)$ و $a(\cdot)$ توابعی معلوم هستند.

در یک خانواده توزیع‌های نمایی امید ریاضی و واریانس به‌صورت زیر خواهند بود:

$$E(Y) = \frac{\partial\psi(\lambda)}{\partial\lambda} = \mu \qquad Var(Y) = a(\phi)\frac{\partial^2\psi\lambda}{\partial(\lambda)^2}. \qquad (۴.۱)$$

از جمله توزیع‌های نمایی معروف می‌توان به توزیع‌های نرمال، دوجمله‌ای، پواسون، نمایی، گاما، بتا و گاوسی معکوس اشاره کرد. در یک مدل خطی تعمیم‌یافته، η ترکیبی خطی از متغیرهای تبیینی و پارامترهای رگرسیونی، به‌نام پیش‌گوی خطی[۳۶] بوده به‌طوری

$$g(\mu) = \eta = \boldsymbol{x}'\beta$$

$g(\cdot)$ یک تابع پیوند است. مدل‌های خطی تعمیم‌یافته دارای سه مولفه هستند:

۱. **مولفه تصادفی:** توزیع شرطی متغیر پاسخ را مشخص می‌کند. توزیع پاسخ، معمولا از خانواده توزیع‌های نمایی مانند نرمال، دوجمله‌ای، بتا و پواسون انتخاب می‌شود، که انتخاب مولفه تصادفی بنا به ساختار متغیر پاسخ متفاوت است. برای مثال، اگر ساختار متغیر پاسخ به‌صورت متغیر شمارشی متناهی باشد، می‌توان توزیع دوجمله‌ای را به‌عنوان توزیع مولفه تصادفی انتخاب کرد.

۲. **مولفه منظم:** این مولفه نقش متغیرهای تبیینی را در قالب یک ترکیب خطی از متغیرهای تبیینی و پارامترهای رگرسیونی، η با نام پیش‌گوی خطی مشخص می‌کند. در واقع

[۳۵] Link function
[۳۶] Linear predictor

پیش‌گوی خطی به‌صورت

$$\eta = \boldsymbol{x}'\beta$$

تعریف می‌شود.

۳. **تابع پیوند:** تابع پیوند به‌منظور ارتباط دادن مولفه منظم به مولفه تصادفی استفاده می‌شود. در واقع داریم

$$g(\mu) = \eta$$

یا به‌طور معادل

$$g^{-1}(\eta) = \mu.$$

تابع پیوند انواع مختلفی دارد. همان‌طور که عنوان شد موقعیت‌های کاربردی بسیاری وجود دارند که پذیره نرمال بودن متغیر پاسخ حتی به‌طور تقریبی برقرار نیست. به‌عنوان مثال، فرض کنید متغیر پاسخ یک متغیر گسسته شمارشی است. در مواردی که در بازه زمانی مشخصی حداکثر تعداد n مشخص نباشد از توزیع پواسون استفاده می‌کنیم که تابع پیوند آن لگاریتمی و به‌صورت زیر قابل تعریف است:

$$\eta = g(\mu) = \ln(\mu).$$

بنابراین

$$\mu = e^{\eta} > 0.$$

در بسیاری از موارد با شمارش بیمارانی که دارای یک بیماری خاص هستند، مواجه می‌شویم. در چنین مواردی که متغیر پاسخ شمارشی و حجم جامعه متناهی و مثلا برابر n است، مدل معمول به‌صورت دوجمله‌ای تعریف می‌شود و از رگرسیون لجستیک برای مدل‌بندی داده‌ها استفاده می‌شود که تابع پیوند آن به‌صورت زیر می‌باشد:

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right).$$

۱.۳.۱ مدل‌های آمیخته خطی تعمیم‌یافته

یکی از پذیره‌های اساسی مدل‌های خطی تعمیم‌یافته استقلال بین مشاهدات است. اما در مواردی مانند داده‌های فضایی که پاسخ‌ها به هم وابسته هستند، این پذیره برقرار نیست. در این موارد، تعمیمی از این مدل‌ها معروف به مدل‌های آمیخته خطی تعمیم‌یافته مورد استفاده قرار می‌گیرد. در واقع در این‌گونه موارد که بین مشاهدات همبستگی وجود دارد، معمولا پذیره استقلال مشاهدات به استقلال شرطی تعدیل و همبستگی بین آن‌ها با اضافه کردن اثرات تصادفی از طریق متغیرهای پنهان به مدل لحاظ می‌شود. به‌عبارتی با افزودن اثرات

تصادفی به پیش‌گوی خطی در یک GLM، یک خانواده از مدل‌های منعطف که در موقعیت‌های عملی مختلفی کاربرد دارند، تشکیل می‌شود. متغیر پاسخ در GLMM مستقل شرطی بوده و دارای توزیعی از خانواده توزیع‌های نمایی با تابع چگالی احتمال (۳.۱) می‌باشد و به‌طور مشابه میانگین و واریانس آن از رابطه (۴.۱) به‌دست خواهند آمد. اگر میانگین شرطی برابر با $E(Y|X,Z,\alpha)=\mu$ باشد، آن‌گاه تابع پیوند روی این میانگین به‌عنوان ترکیبی خطی از هر دو عامل ثابت β و تصادفی α به مدل معرفی خواهد شد، یعنی

$$g(\mu)=X\beta+Z\alpha$$

که در آن Z ماتریس طرح متناظر با اثرات تصادفی α است.

۲.۳.۱ مدل‌های آمیخته خطی تعمیم‌یافته فضایی

باتوجه به عدم برقراری پذیره استقلال در بین مشاهدات غیرنرمال، از مدل‌های آمیخته خطی تعمیم‌یافته استفاده خواهیم کرد. در صورتی که همبستگی از نوع فضایی باشد، مدل‌های آمیخته خطی تعمیم‌یافته به مدل‌های آمیخته خطی تعمیم‌یافته فضایی تبدیل می‌شوند. برسلو و کلایتون (۱۹۹۳) از این مدل‌ها در مطالعات پزشکی استفاده کردند. دیگل و همکاران (۱۹۹۸) این مدل‌ها را برای تحلیل متغیرهای فضایی گسسته در یک جامعه ناحیه پیوسته به‌کار بردند.

مدل مورد نظر در این پایان‌نامه یک مدل آمیخته خطی تعمیم‌یافته فضایی با پاسخ شمارشی در یک جامعه متناهی است.

۳.۳.۱ رگرسیون دوجمله‌ای

در برخی از موقعیت‌ها، وضعیت‌هایی وجود دارند که در آن متغیر پاسخ Y تنها دو برآمد دارد. مثلا فشار خون بالا یا فشار خون پایین، یا پاسخ دادن یا پاسخ ندادن یک بیمار به یک دارو. در این قبیل موارد برآمد Y را می‌توان به‌صورت صفر یا یک کدگذاری کرد. در واقع می‌خواهیم برآمد یا احتمال آن را برای یک یا چند مقدار از متغیرهای تبیینی X پیش‌بینی کنیم. در این موارد از رگرسیون لجستیک استفاده می‌کنیم. به‌طور کلی رگرسیون لجستیک با رگرسیون خطی دو تفاوت دارد:

۱. در مدل رگرسیون خطی، کمیت کلیدی $E(Y|X=x)$ توسط یک رابطه خطی از X یا توسط تبدیلات این متغیرها بیان می‌شود. مانند $E(Y|X=x)=\beta_0+\beta_1 x$. در این صورت مقادیر $E(Y|X=x)$ می‌تواند در بازه $-\infty$ تا $+\infty$ باشند. در مقابل، وقتی مقادیر پاسخ از نوع کیفی (اسمی و دودویی) باشند، میانگین که به معنای نسبت است، می‌تواند برآوردی برای $E(Y|X=x)$ ایجاد کند، لذا $0\le E(Y|X)\le 1$. بنابراین

$$E(Y|X=x)=P(Y=1|X=x)\times 1+P(Y=0|X)\times 0.$$

الگویی که بین صفر و یک محدود است و به‌طور مجانبی به صفر و یک میل می‌کند، توسط تابع پیوند لجیت به‌دست می‌آید و به‌صورت زیر بیان می‌شود:

$$p(x) = E(Y|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

این الگو را با تبدیل ساده زیر می‌توان به الگوی خطی تبدیل کرد:

$$\text{logit(p)} = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x.$$

۲. دومین اختلاف مربوط به توزیع شرطی متغیر پاسخ است. مدل رگرسیون خطی به‌صورت $y = x'\beta+\epsilon$ نوشته می‌شود که در آن $\epsilon \sim N\left(0, \sigma^2\right)$ در نظر گرفته می‌شود. اما این مطلب در مورد پاسخ‌های دودویی برقرار نیست. در این وضعیت مقدار متغیر پاسخ را با معلوم بودن x، به‌صورت $y = \pi(x) + \epsilon$ در نظر می‌گیریم. اگر $y = 0$ باشد، آن‌گاه $\epsilon = -\pi(x)$ با احتمال $1 - \pi(x)$ است. بنابراین می‌توان گفت ϵ توزیعی با میانگین صفر و واریانس $\pi(x)[1 - \pi(x)]$ دارد، یعنی توزیع خطا به جای نرمال دوجمله‌ای است. بنابراین توزیع شرطی متغیر پاسخ از یک توزیع دوجمله‌ای با احتمال $\pi(x)$ پیروی می‌کند. پارامترهای β_0 و β_1 نیز معمولا با روش ماکسیمم درستنمایی برآورد می‌شوند.

مدل رگرسیون لجستیک را می‌توان برای بیش از یک متغیر تبیینی نیز تعمیم داد. اگر متغیر پاسخ از نوع دودویی باشد و $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$ متغیرهای تبیینی را تشکیل دهند، آن‌گاه مدل رگرسیون لجستیک به‌صورت زیر نوشته می‌شود:

$$P(Y = 1|X = x) = E(Y|X = x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots, \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots, \beta_n x_n)}$$

به‌طوری که

$$\ln\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 x_1 + \cdots, \beta_n x_n.$$

۴.۳.۱ رگرسیون پواسون

رگرسیون پواسون مدلی بسیار پرکاربرد برای تحلیل داده‌های شمارشی است. مدل رگرسیون پواسون نوعی مدل خطی تعمیم یافته است که در آن متغیر پاسخ شمارشی بوده و از توزیع پواسون پیروی می‌کند. زمینه‌های کاربردی بسیاری برای رگرسیون پواسون وجود دارد. این کاربردها وقتی پیش می‌آیند که پاسخ طبیعی در مساله تنها مقادیر صحیح نامنفی $0, 1, 2, 3, \cdots$ باشند. به‌عنوان مثال‌هایی از رگرسیون پواسون می‌توان به تعداد دفعات اهدای خون یا تعداد دفعات ترک اعتیاد اشاره کرد.

توزیع احتمال پواسون برای متغیر تصادفی Y به‌صورت زیر است:

$$f(y, \mu) = \frac{\mu^y e^{-\mu}}{y!} \qquad y = 0, 1, 2, \cdots$$

در مدل رگرسیون پواسون تابع پیوند طوری انتخاب می‌شود که معکوس آن یک تابع مثبت باشد. بنابراین تابع پیوند لگاریتمی یک انتخاب طبیعی است. یعنی

$$\ln(\mu) = \boldsymbol{x'\beta}$$

که در آن $\mu = (\boldsymbol{x^{(1)}\beta})$، $\boldsymbol{\beta}$ و $\boldsymbol{x}$ به‌ترتیب بردارهای k بعدی از پارامترها و متغیرهای تبیینی هستند. لگاریتم تابع درستنمایی این مدل برای n مشاهده به‌صورت زیر است:

$$\log L(\beta) = \sum_{i=1}^{n} \{-\exp\left(x_i'\beta\right) + y_i x_i'\beta - \log(y_i!)\}.$$

محاسبه برآورد پارامترها مستلزم حل معادلات غیرخطی $\sum_{i=1}^{n}\{y_i - \exp\left(x_i'\hat{\beta}\right)\}x_i = 0$ توسط روش کم‌ترین توان‌های دوم تکراری وزنی[۳۷] (IWLS) است. می‌توان نشان داد که برآورد پارامتر β توسط معادله $(X'WX)^{-1}(X'Wy^*)$ به‌صورت روش تکراری IWLS و به‌صورت عددی محاسبه می‌شود که در آن ماتریس وزنی W قطری با اعضای

$$w_{ij} = \frac{1}{var(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

و $y^* = x'\hat{\beta} + (y-\mu)\frac{\partial \eta}{\partial \mu}$ بردار پاسخ اصلاح‌شده هستند. برآورد پارامترها، تحت شرایط نظم، به‌طور مجانبی دارای توزیع نرمال با میانگین β و ماتریس واریانس A^{-1} است. به‌عبارت دیگر $\sqrt{n}\left(\hat{\beta}_{ML} - \beta\right) \longrightarrow N\left(0, A^{-1}\right)$ که در آن

$$A = lim_{n\to\infty} E\left[\frac{1}{n}\sum_i \exp\left(x_i'\beta\right) x_i x_i'\right].$$

علاوه بر این ویژگی سازگار بودن برآوردگرهای پارامترها نیز برقرار است.

در مسائل کاربردی ممکن است با داده های شمارشی روبرو شویم که برآورد واریانس مشاهده‌شده از واریانس مورد انتظار بزرگ‌تر باشد. این ویژگی، بیش پراکنش[۳۸] نامیده می‌شود. مدل پواسون برای برازش داده‌های شمارشی بیش‌پراکنش مناسب نیست (حسن زاده و کاظمی، ۱۳۹۰).

۴.۱ کریگینگ پواسون

مونستیز و همکاران (۲۰۰۶) مدل کریگینگ پواسون را برای مدل‌سازی مشاهدات شمارشی همبسته فضایی معرفی کردند. در ادامه به معرفی مدل و معادلات این نوع کریگینگ می‌پردازیم.

[۳۷] Iterated weighted least squre

[۳۸] Overdispersion

۱.۴.۱ مشخصات مدل

میدان تصادفی $\{Z = Z(s);\ s \in D \subset \Re^{۲}\}$ را در نظر بگیرید. فرض کنید

$$Z(s)|Y(s) \sim pois(t(s)Y(s))$$

به‌طوری که در آن Z داده شمارشی مشاهده‌شده در مکان s، $Y(s)$ میدان فضایی پنهان تولید کننده پارامتر نرخ پواسون و $t(s)$ نیز زمان مشاهدات در مکان s است. علاوه بر این، فرض کنید میدان تصادفی $Y(s)$ مثبت و مانای مرتبه دوم با میانگین m و واریانس$\sigma_Y^{۲}$ باشد. تابع کوواریانس نیز فقط تابعی از فاصله مشاهدات به‌صورت $C_Y(||s-s'||)$ است. همچنین تابع کوواریانس

$$C_Y(||s-s'||) = Cov[Y(s), Y(s')]$$

می‌تواند در رابطه تغییرنگار یعنی، $\gamma_Y(s-s') = \frac{۱}{۲}E\left[(Y(s)-Y(s'))^{۲}\right]$ بازچینی شود. پذیره دیگری بر روی میدان تصادفی پنهان به جز این‌که $Y \geq ۰$ است وجود ندارد.

در ادامه، به منظور سادگی $Z(s)$، $Y(s)$ و $t(s)$ را به‌ترتیب با Z_s، Y_s و t_s نمایش می‌دهیم. همچنین در معادلات کریگینگ $C_{ss'}$ نشان‌دهنده تابع کوواریانس $C_Y(s-s')$ است. بنابراین با توجه به تعاریف بالا می‌توان نوشت

$$E[Z_s|Y_s] = t_sY_s, \qquad E[Z_s] = t_sE[Y_s] = mt_s$$
$$Var[Z_s|Y_s] = t_sY_s, \qquad E\left[(Z_s)^{۲}|Y_s\right] = t_sY_s + (t_sY_s)^{۲}$$
$$Var[Z_s] = t_s^{۲}\sigma_Y^{۲} + mt_s.$$

کوواریانس فضایی در دو مکان s_i و s_j نیز برابر است با

$$\begin{aligned} E[Z_sZ_{s'}|Y] &= Cov[Z_s, Z_{s'}|Y] + E[Z_s|Y_a]E[Z_{s'}|Y_{s'}] \\ &= \delta_{ss'}t_sY_s + t_st'_sY_sY'_s \end{aligned}$$

به‌طوری که در آن δ_{ij} دلتای کرونکر است، یعنی

$$\delta_{ij} = \begin{cases} ۱ & i = j \\ ۰ & i \neq j \end{cases}$$

امید ریاضی شرطی و کناری $\frac{Z_s}{t_s} - \frac{Z_{s'}}{t_{s'}}$ به‌صورت

$$\begin{aligned} E\left[\frac{Z_s}{t_s} - \frac{Z_{s'}}{t_{s'}}|Y\right] &= \frac{۱}{t_s}E[Z_s|Y_s] - \frac{۱}{t_{s'}}E[Z_{s'}|Y_{s'}] \\ &= Y_s - Y_{s'} \end{aligned}$$

و

$$E\left[\frac{Z_s}{t_s} - \frac{Z_{s'}}{t_{s'}}\right] = m - m = ۰$$

هستند. امید ریاضی شرطی و کناری اختلاف توان دوم نسبت‌ها نیز به‌ترتیب به‌صورت

$$E\left[\left(\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}\right)^2|Y\right]=\frac{1}{t_s^2}E\left[(Z_s)^2|Y_s\right]-\frac{1}{t_{s'}^2}E\left[Z_{s'}^2|Y_{s'}\right]-\frac{2}{t_s t_{s'}}E\left[Z_s Z_{s'}|Y\right]$$
$$=\frac{Y_s}{t_s}-\frac{Y_{s'}}{t_{s'}}-2\delta_{ss'}\frac{Y_s}{t_s}+(Y_s-Y_{s'})^2$$

و

$$E\left[\left(\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}\right)^2\right]=\left[\frac{Y_s}{t_s}+\frac{Y_{s'}}{t_{s'}}-2\delta_{ss'}\frac{Y_s}{t_s}+(Y_s-Y_{s'})^2\right]$$
$$=m\left(\frac{t_s+t_{s'}}{t_s t_{s'}}\right)-2\delta_{ss'}\frac{m}{t_s}+2\gamma_Y\left(Y_s-Y_{s'}\right)$$

نتیجه می‌شوند. بنابراین برای واریانس می‌توان نوشت

$$Var\left[\left(\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}\right)\right]=E\left[\left(\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}\right)^2\right]-E\left[\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}\right]^2.$$

بر اساس تعریف نیم‌تغییرنگار، خواهیم داشت

$$\frac{1}{2}Var\left[\left(\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}\right)\right]=\frac{m}{2}\left(\frac{t_s+t_{s'}}{t_s t_{s'}}\right)-\delta_{ss'}\frac{m}{t_s}+\gamma_Y\left(s-s'\right). \qquad (۵.۱)$$

برای $s\neq s'$ داریم

$$\gamma_Y\left(s-s'\right)=\gamma_Z\left(s-s'\right)-\frac{m}{2}\left(\frac{t_s+t_{s'}}{t_s t_{s'}}\right). \qquad (۶.۱)$$

اگر $s=s'$ آن‌گاه معادله (۵.۱) به‌صورت زیر تبدیل می‌شود:

$$\gamma_Y\left(0\right)=\gamma_Z(0)=0. \qquad (۷.۱)$$

علاوه بر این

$$Var\left[\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}|Y\right]=E\left[\left(\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}\right)^2|Y\right]-E^2\left[\left(\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}\right)|Y\right]$$
$$=\frac{Y_s}{t_s}-\frac{Y_{s'}}{t_{s'}}+(Y_s-Y_{s'})^2-(Y_s-Y_{s'})^2$$
$$=\frac{Y_s}{t_s}+\frac{Y_{s'}}{t_{s'}}$$

$$E\left[Var\left[\frac{Z_s}{t_s}-\frac{Z_{s'}}{t_{s'}}|Y\right]\right]=E\left[\frac{Y_s}{t_s}+\frac{Y_{s'}}{t_{s'}}\right]$$
$$=m\left(\frac{t_s+t_{s'}}{t_s t_{s'}}\right).$$

۲.۴.۱ تغییرنگار کریگینگ پواسون

حال n مشاهده Z_i، $i = ۱, \cdots, n$، با اندازه نمونه n از $Z(s_i)$ در زمان مشاهده‌شده t_i را در نظر بگیرید. برآورد تعدیل‌یافته برای تغییرنگار به‌صورت زیر به‌دست می‌آید:

$$\gamma_Y^*(h) = \frac{۱}{N(h)} \sum_{i=۱}^{n} \sum_{j=۱}^{n} \frac{t_i t_j}{t_i + t_j} \left[\frac{۱}{۲} \left(\frac{Z_i}{t_i} - \frac{Z_j}{t_j} \right)^۲ - \frac{m^*}{۲} \left(\frac{t_i + t_j}{t_i t_j} \right) \right] I_d(i,j) \sim h$$

که در آن منظور از $I_d(i,j) \sim h$ تابع نشانگر برای نقاط s_i و s_j ای است که در فاصله h از هم قرار دارند. عدد $N(h)$ نیز تعداد نقاط همسایه در فاصله h و ثابت نرمال‌ساز است که برای همگن کردن تمام مکان‌های فضایی به‌صورت زیر تعریف می‌شود:

$$N(h) = \sum_{i,j} \frac{t_i t_j}{t_i + t_j} I_d(i,j) \sim h.$$

در معادله تغییرنگار m^* نیز برآورد میانگین Y است. عبارت تصحیح اریبی، $-\frac{m^*}{۲}\left(\frac{t_i+t_j}{t_i t_j}\right)$، به‌طور مستقیم از رابطه (۶.۱) به‌دست می‌آید. حالت ساده‌شده تغییرنگار به‌صورت زیر حاصل می‌شود:

$$\gamma_Y^*(h) = \frac{۱}{۲N(h)} \times \sum_{ij} \left(\frac{t_i t_j}{t_i + t_j} \left(\frac{Z_i}{t_i} - \frac{Z_j}{t_j} \right)^۲ - m^* \right) I_d(i,j) \sim h.$$

۳.۴.۱ معادلات کریگینگ پواسون

برای هر مکان فضایی جدید $s_۰$ مقدار پیش‌گویی $Y_۰$ می‌تواند به‌صورت ترکیب خطی از داده‌های مشاهده‌شده در زمان t_α با وزن‌های λ_i، $i = ۱, ۲, \cdots, m$ نوشته شود.یعنی

$$Y_۰^* = \sum_{i=۱}^{n} \lambda_i \frac{Z_i}{t_i}.$$

بهترین وزن‌های λ_i باید به‌گونه‌ای انتخاب شوند که مقادیر پیش‌گویی در مکان جدید نااریب باشند و کم‌ترین واریانس پیش‌گو را داشته باشند. بنابراین با در نظر گرفتن شرط نااریبی $Y_۰^*$، امید ریاضی شرطی و حاشیه‌ای در مکان جدید عبارتند از

$$E[Y_۰^*|Y] = \sum_{i=۱}^{n} \frac{\lambda_i}{t_i} t_i Y_i = \sum_{i=۱}^{n} \lambda_i Y_i$$

$$E[Y_۰^*] = \sum_{i=۱}^{n} \lambda_i E[Y_i] = m \sum_{i=۱}^{n} \lambda_{i=۱}$$

$$\sum_{i=۱}^{m} \lambda_i = ۱.$$

امید ریاضی شرطی وکناری توان دوم خطای پیش‌گو به‌صورت

$$E\left[(Y_\circ^* - Y_\circ)^۲ \,|Y\right] = E\left[\left(\sum_{i=۱}^{n} \lambda_i \frac{Z_i}{t_i} - Y_\circ\right)^۲ |Y\right]$$
$$= \sum_{i=۱}^{n}\sum_{j=۱}^{n} \frac{\lambda_i}{t_i}\frac{\lambda_j}{t_j} E\left[Z_i Z_j|Y\right] + Y_\circ^۲ - ۲Y_\circ \sum_{i=۱}^{n} \frac{\lambda_i}{t_i} E\left[Z_i|Y\right]$$
$$= \sum_{i=۱}^{n}\sum_{j=۱}^{n} \lambda_i\lambda_j Y_i Y_j + \sum_{i=۱}^{n} \frac{\lambda_i^۲}{t_i} Y_i + Y_\circ^۲ - ۲Y_\circ \sum_{i=۱}^{n} \lambda_i Y_i$$
$$E\left[(Y_\circ^* - Y_\circ)^۲\right] = E\left[\sum_{i=۱}^{n}\sum_{j=۱}^{n} \lambda_i\lambda_j Y_i Y_j + \sum_{i=۱}^{n} \frac{\lambda_i^۲}{t_i} Y_i + Y_\circ^۲ - ۲Y_\circ \sum_{i=۱}^{n} \lambda_i Y_i\right]$$
$$= \sum_{i=۱}^{n}\sum_{j=۱}^{n} \lambda_i\lambda_j C_{ij} + \sum_{i=۱}^{n} \frac{\lambda_i^۲}{t_i} m + \sigma_Y^۲ - ۲\sum_{i=۱}^{n} \lambda_i C_{io} + ۲m^۲ - ۲m^۲$$
$$= \sigma_Y^۲ + \sum_{i=۱}^{n} \frac{\lambda_i^۲}{t_i} m + \sum_{i=۱}^{n}\sum_{j=۱}^{n} \lambda_i\lambda_j C_{ij} - ۲\sum_{i=۱}^{n} \lambda_i C_{io}$$

است. بنابراین با توجه به قید ناریبی، خواهیم داشت

$$Var\,(Y_\circ^* - Y_\circ)^۲ = E\left[(Y_\circ^* - Y_\circ)^۲\right].$$

با می‌نیمم کردن واریانس خطای پیش‌گو، معادله کریگینگ پواسون به‌صورت زیر حاصل می‌شود:

$$\sum_{j=۱}^{n} \lambda_j C_{ij} + \lambda_i \frac{m}{t_i} + \mu = C_{iO} \qquad i = ۱, \cdots, n$$
$$\sum_{i=۱}^{n} \lambda_i = ۱.$$

همچنین شکل نهایی واریانس خطای پیش‌گو با توجه به واریانس کریگینگ معمولی به‌صورت

$$Var\,(Y_\circ^* - Y_\circ) = \sigma_Y^۲ - \sum_{i=۱}^{n} C_{ij} - \mu$$

نتیجه می‌شود.

در بخش‌هایی از این پایان‌نامه به منظور سادگی $Z(s_i)$ و $Y(s_i)$ به ترتیب با Z_i و Y_i نمایش داده شده‌اند.

۵.۱ توزیع‌ها

در این بخش، دو توریع آماری دیگر را که در این پایان‌نامه از آن‌ها استفاده خواهیم کرد، معرفی می‌کنیم.

۱.۵.۱ توزیع دوجمله‌ای

هرگاه متغیر تصادفی X تعداد موفقیت‌ها در n بار تکرار مستقل یک آزمایش برنولی با احتمال موفقیت π در هر آزمایش باشد ($۰ < \pi < ۱$) آن‌گاه X را یک متغیر تصادفی دوجمله‌ای گویند و آن را با نماد $bin(n, \pi)$ نشان می‌دهند. تابع جرم احتمال توزیع دوجمله‌ای به صورت زیر است:

$$f_X(x) = \binom{n}{x}\pi^x(۱-\pi)^{n-x}, \qquad x = ۰, ۱\cdots, n$$

در این توزیع $E(X) = n\pi$ و $Var(X) = n\pi(۱-\pi)$ است.

۲.۵.۱ توزیع بتا

اگر متغیر تصادفی X دارای تابع چگالی احتمال

$$f_{\alpha,\beta}(x) = \frac{۱}{B(\alpha,\beta)}x^{\alpha-۱}(۱-x)^{\beta-۱}; \qquad ۰ < x < ۱, \qquad \alpha, \beta > ۰$$

باشد، آن‌گاه X را دارای توزیع بتا با پارامترهای α و β نامیده و آن را با نماد $X \sim Beta(\alpha, \beta)$ نمایش می‌دهند. همچنین

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

و

$$E(X) = \frac{\alpha}{\alpha+\beta}$$

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^۲(\alpha+\beta+۱)}.$$

فصل ۲

مدل بتا_دوجمله‌ای فضایی

۱.۲ مقدمه

پاسخ‌های شمارشی همبسته فضایی در جوامع متناهی در بسیاری از تحقیقات کاربردی در زمینه‌هایی مانند پزشکی، اقتصاد، علوم زیست‌شناسی، علوم اجتماعی، بهداشت و فیزیک مورد توجه محققان قرار دارند. ماهیت پاسخ در بسیاری از موارد به‌صورت، نرخ، نسبت یا درصد است. به‌عنوان نمونه، ممکن است نسبت افراد مبتلا به سرطان معده، پدیده مورد علاقه باشد. در این موارد که حجم کل جامعه عدد مشخصی است، مدل متداول مبتنی بر توزیع دوجمله‌ای تعریف می‌شود. ویژگی مهم این توزیع، کوچک‌تر بودن واریانس از میانگین است. اما در موارد متعددی، این پذیره برقرار نیست و با مساله بیش‌پراکنش مواجه می‌شویم. بنابراین، در چنین شرایطی توزیع دوجمله‌ای نمی‌تواند ماهیت داده‌ها را به‌خوبی بیان کند و استفاده از آن برای مدل‌بندی داده‌ها منجر به استنباط‌های آماری نادرست در برآورد پارامترها، برآورد ساختار وابستگی و پیش‌گویی فضایی داده‌ها می‌شود. با این مقدمه، مدل بتا_دوجمله‌ای، به دلیل وجود دو پارامتر شکل در توزیع بتا، جانشینی بهتر برای مدل دوجمله‌ای است. این توزیع تمام حالت‌های بیش‌پراکنش، کم‌پراکنش و پراکنش متعادل را برای داده‌ها در نظر می‌گیرد.

مبانی ایده مدل بتا_دوجمله‌ای توسط پیرسون مطرح شد و به صورت گسترده توسط اسکلام آن در آزمون هوش (هوین، ۱۹۷۹؛ لرد، ۱۹۶۵؛ ویلکاکس، ۱۹۸۱)، آزمایش سم‌شناسی[۱]

[۱] Toxicology

(ویلیامز، ۱۹۷۵)، همه‌گیرشناسی[۲] (گریفتس، ۱۹۷۳)، نمایش رسانه‌ها (گرین، ۱۹۷۰) و رفتار خریداران (مسی و همکاران، ۱۹۷۰) استفاده کرد.

ویلیامز (۱۹۷۵) داده‌های مربوط به تاثیر دارویی خاص بر روی حیوانات آزمایشگاهی را مورد بررسی قرار داد. در آزمایش وی متغیر پاسخ، تعداد توله‌هایی بود که در دوره شیرخوارگی زنده مانده‌اند. توزیع دوجمله‌ای با پارامتر ثابت π گزینه معمول برای این نوع داده‌ها بود، اما نکته قابل توجه در آزمایش او آن بود که نسبت بقای توله‌ها در یک زایمان، از یک زایمان به دیگری تغییر می‌کرد. به عبارت دیگر داده‌ها بیش‌پراکندگی داشتند و مدل دوجمله‌ای نمی‌توانست آن‌ها را به‌خوبی مدل‌بندی کند. ویلیامز از توزیع بتا_دوجمله‌ای به‌عنوان یک مدل انعطاف‌پذیر برای تغییرپذیری بین نمونه‌ها استفاده کرد. در مطالعه اثر زایمان روی مدل‌بندی دز پاسخ در تراریختگی، کوپر و همکارانش (۱۹۸۶) نتیجه گرفتند که عدم به‌کارگیری تاثیر زایمان می‌تواند باعث کم‌برآوردی واریانس مرتبط با برآوردهای پارامترها باشد، به‌طوری که استفاده از درستنمایی دوجمله‌ای برای مدل‌بندی داده‌های تراریختگی[۳] به نظر معقول نیست. آن‌ها توزیع بتا_دوجمله‌ای را برای معرفی درجه تغییر همبستگی درون زایمان پیشنهاد کردند. پل (۱۹۸۲) در تحلیل نسبت‌های جنین تاثیرپذیر در آزمایش‌های تراریختگی دریافت که مدل بتا_دوجمله‌ای نسبت به مدل دوجمله‌ای بهتر عمل می‌کند. پک (۱۹۸۶) به این نتیجه رسید که مدل بتا_دوجمله‌ای نسبت به مدل‌های معادل مانند دوجمله‌ای همبسته کوپر و هاسمن (۱۹۷۸) بهتر است. تارون (۱۹۸۲) برای بسیاری از انواع غده‌ها مشاهده کرد که نرخ کنترل عمر غده‌ها بسیار تغییرپذیرتر از آن است که پذیره دوجمله‌ای برای آن‌ها مناسب باشد، و توزیع بتا_دوجمله‌ای را برای نرخ عمر غده‌ها برازش داد. شوکرز (۲۰۰۳) نیز از توزیع بتا_دوجمله‌ای در ارزیابی کارایی دستگاه تشخیص هویت زیست‌سنجی[۴] استفاده کرد.

در این فصل ابتدا به معرفی مدل بتا_دوجمله‌ای و خواص آن می‌پردازیم. در ادامه مدل کریگینگ بتا_دوجمله‌ای که شامل معادلات کریگینگ و واریانس پیش‌گو می‌باشد را بیان می‌کنیم.

۲.۲ توزیع بتا_دوجمله‌ای

فرض کنید متغیر تصادفی شرطی $X|\pi$ نشان‌دهنده تعداد موفقیت‌ها در n آزمایش برنولی با احتمال موفقیت π باشد. اگر متغیر تصادفی π دارای توزیع پیشین بتا با پارامترهای شکل α و β باشد، به‌طوری که احتمال موفقیت π برای هر رخداد متغیر است، آن‌گاه متغیر تصادفی X دارای توزیع کناری بتا_دوجمله‌ای با پارامترهای n، α و β است که تابع احتمال آن به‌صورت زیر به‌دست می‌آید:

---

[۲] Epidemiology
[۳] Teratology
[۴] Biometrics

فرض کنید

$$X|\pi \sim bin(n,\pi)$$

به‌طوری که

$$\pi \sim B(\alpha,\beta).$$

بنابراین تابع احتمال توام به‌صورت زیر نتیجه می‌شود:

$$f(x,\pi) = f(\pi)\cdot f(x|\pi) = \frac{\binom{n}{x}}{B(\alpha,\beta)}\pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}$$

که در آن

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

با انتگرال‌گیری از توزیع توام نسبت به π، داریم

$$f(x) = \int_0^1 \frac{\binom{n}{x}}{B(\alpha,\beta)}\pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}d\pi.$$

با ضرب $B(\alpha+x,\beta+n-x)$ خواهیم داشت

$$f(x) = \frac{\binom{n}{x}B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}\int_0^1 \frac{1}{B(\alpha+x,\beta+n-x)}\pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}d\pi.$$

با توجه به این‌که

$$\int_0^1 \frac{1}{B(\alpha+x,\beta+n-x)}\pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}d\pi = 1$$

پس تابع احتمال کناری به‌صورت

$$f(x) = \binom{n}{x}\frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}, \qquad x = 0,1,\cdots,n$$

حاصل می‌شود که همان توزیع بتا_دوجمله‌ای است و آن را با نماد $X \sim beta-bin(n,\alpha,\beta)$ نشان می‌دهند.

با استفاده از خواص توزیع گاما می‌دانیم

$$\Gamma(\alpha+x) = \Gamma(\alpha)\alpha(\alpha+1)\cdots(\alpha+x-1).$$

بنابراین با توجه به ویژگی‌های ذکر شده، تابع احتمال بتا_دوجمله‌ای به‌صورت زیرقابل بازنویسی است:

$$f(x) = \binom{n}{x}\frac{\alpha(\alpha+1)\cdots(\alpha+x-1)\beta(\beta+1)\cdots(\beta+n-x-1)}{(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+n-1)}.$$

۱.۲.۲ میانگین و واریانس توزیع بتا_دوجمله‌ای

میانگین و واریانس توزیع بتا_دوجمله‌ای را می‌توان به‌صورت زیر محاسبه کرد:

$$\begin{aligned}E(X) &= \sum_{x=۰}^{n} x\binom{n}{x}\frac{(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}\\ &= \frac{xn!}{x!(n-x)!}\frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}\\ &= \sum_{x=۱}^{n} n\binom{n-۱}{x-۱}\frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}\\ &= n\sum_{x=۰}^{n-۱}\binom{n-۱}{x}\frac{B(\alpha+x+۱,\beta+n-x-۱)}{B(\alpha,\beta)}\end{aligned}$$

اکنون با ضرب $B(\alpha+۱,\beta)$ می‌توان نوشت

$$E(X) = \frac{nB(\alpha+۱,\beta)}{B(\alpha,\beta)}\sum_{x=۰}^{n-۱}\binom{n-۱}{x}\frac{B(\alpha+x+۱,\beta+n-x-۱)}{B(\alpha+۱,\beta)}.$$

از آن‌جایی که

$$\sum_{x=۰}^{n-۱}\binom{n-۱}{x}\frac{B(\alpha+x+۱,\beta+n-x-۱)}{B(\alpha+۱,\beta)} = ۱$$

می‌توان نوشت

$$\begin{aligned}E(X) &= \frac{nB(\alpha+۱,\beta)}{B(\alpha,\beta)} = \frac{n\Gamma(\alpha+۱)\Gamma(\beta)}{\Gamma(\alpha+\beta+۱)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\\ &= n\alpha\frac{\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.\end{aligned}$$

در نتیجه

$$E(X) = \frac{n\alpha}{\alpha+\beta}.$$

برای محاسبه واریانس طبق تعریف داریم

$$Var(X) = E(X^{۲}) - E^{۲}(X). \qquad (۱.۲)$$

از آن‌جا که

$$E(X^{۲}) = E(X(X-۱)) + E(X)$$

بنابراین

$$\begin{aligned}E(X(X-۱)) &= \sum_{x=۰}^{n}(x)(x-۱)\frac{n(n-۱)(n-۲)}{x(x-۱)(x-۲)(n-x)!}\frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}\\ &= n(n-۱)\sum_{x=۲}^{n}\binom{n-۲}{x-۲}\frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}\end{aligned}$$

با ضرب $B(\alpha+۲,\beta)$ داریم

$$E(X(X-۱))=\frac{n(n-۱)B(\alpha+۲,\beta)}{B(\alpha,\beta)}\sum_{x=۰}^{n-۲}\binom{n-۲}{x}\frac{B(\alpha+x+۲,\beta+n-x-۲)}{B(\alpha+۲,\beta)}.$$

باتوجه به

$$\sum_{x=۰}^{n-۲}\binom{n-۲}{x}\frac{B(\alpha+x+۲,\beta+n-x-۲)}{B(\alpha+۲,\beta)}=۱$$

نتیجه می‌شود

$$E(X(X-۱))=\frac{n(n-۱)B(\alpha+۲,\beta)}{B(\alpha,\beta)}.$$

بنابراین

$$\begin{aligned}E(X^۲)&=n(n-۱)\frac{B(\alpha+۲,\beta)}{B(\alpha,\beta)}+\frac{n\alpha}{\alpha+\beta}\\&=\frac{\Gamma(\alpha+۲)\Gamma(\beta)}{\Gamma(\alpha+\beta+۲)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\\&=\frac{(\alpha+۱)\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta+۱)(\alpha+\beta)\Gamma(\alpha+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\\&=\frac{n(n-۱)\alpha(\alpha+۱)}{(\alpha+\beta+۱)(\alpha+\beta)}+\frac{n\alpha}{\alpha+\beta}.\end{aligned}$$

با جای‌گذاری در رابطه (۱.۲) خواهیم داشت

$$\begin{aligned}Var(X)&=\frac{n^۲\alpha^۲+n^۲\alpha+n\alpha\beta}{(\alpha+\beta)(\alpha+\beta+۱)}-\frac{n^۲\alpha^۲}{(\alpha+\beta)^۲}\\&=\frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^۲(\alpha+\beta+۱)}\end{aligned}$$

۲.۲.۲ توجیه جعبه و مهره برای مدل بتا_دوجمله‌ای

توزیع بتا_دوجمله‌ای را می‌توان به‌صورت یک مدل جعبه برای مقادیر حقیقی مثبت α و β بیان کرد.

فرض کنید جعبه‌ای شامل α گوی سفید و β گوی سیاه باشد. یک گوی را به تصادف از جعبه خارج می‌کنیم. اگر گوی سفید مشاهده شود، آن را همراه با گوی سفید دیگری به جعبه بر می‌گردانیم. برای گوی سیاه نیز به‌همین صورت عمل می‌کنیم. اگر این عمل را n بار تکرار کنیم، آن‌گاه احتمال مشاهده x گوی سفید دارای توزیع بتا_دوجمله‌ای با پارامترهای n، α و β می‌باشد که به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned}P(X=x)&=\binom{n}{x}\frac{\alpha\,(\alpha+۱)\cdots(\alpha+x-۲)\,(\alpha+x-۱)}{(\alpha+\beta)\,(\alpha+\beta+۱)\cdots(\alpha+\beta+x-۱)}\frac{\beta\,(\beta+۱)\cdots(\beta+n-x-۲)\,(\beta+n-x-۱)}{(\alpha+\beta+x)\cdots(\alpha+\beta+n-۱)}\\&=\binom{n}{x}\frac{\Gamma\,(\alpha+x)\,\Gamma\,(\beta+n-x)\,\Gamma\,(\alpha+\beta)}{\Gamma\,(\alpha+\beta+n)\,\Gamma\,(\alpha)\,\Gamma\,(\beta)}\\&=\binom{n}{x}\frac{B\,(\alpha+x,\beta+n-x)}{B\,(\alpha,\beta)}.\end{aligned}$$

لازم به ذکر است که اگر گوی‌ها را به‌صورت تصادفی و با جای‌گذاری از جعبه خارج کنیم و دوباره به جعبه برگردانیم، و این عمل را n بار تکرار کنیم، آن‌گاه احتمال مشاهده x گوی سفید دارای توزیع دوجمله‌ای با پارامترهای n و $\pi = \frac{\alpha}{\alpha+\beta}$ می‌باشد. همچنین اگر گوی‌ها را به‌صورت تصادفی و بدون جای‌گذاری خارج کنیم، آن‌گاه احتمال مشاهده x گوی سفید دارای توزیع فوق هندسی است.

۳.۲.۲ فرآیند فضایی بتا_دوجمله‌ای

فرض کنید میدان تصادفی $\{Z = Z(s);\ s \in D \subset \Re^{۲}\}$ که در m مکان s_i، $i = ۱,۲,\cdots m$، مشاهده شده است، دارای توزیع شرطی $Z_i|Y_i \sim bin(n_i, Y_i)$ باشد، به‌طوری که احتمال موفقیت Y_i دارای توزیع بتا با پارامترهای α و β و ساختار وابستگی Σ_Y است. در واقع در این حالت میدان تصادفی Y_i همبسته فضایی و مدل زیربنایی، فرآیند بتا_دوجمله‌ای محسوب می‌شود. به عبارت دیگر وابستگی فضایی فقط در سطح Y_i وجود دارد. علاوه بر این، فرض کنید میدان تصادفی Y_i مانای مرتبه دوم باشد. یعنی میانگین توزیع بتا مقدار ثابت π و تابع کوواریانس فقط تابعی از فاصله مشاهدات به‌صورت $C_Y(||s_i - s_j||)$ است. افزون بر این، تابع تغییرنگار مدل را با $\gamma_Y(||s_i - s_j||)$ نشان می‌دهیم. بنابراین می‌توان نوشت

$$E(Y_i) = \frac{\alpha}{\alpha+\beta} = \pi \qquad Var(Y_i) = \frac{\alpha\beta}{(\alpha+\beta)^{۲}(\alpha+\beta+۱)} = \sigma_Y^{۲}.$$

با این ویژگی‌ها متغیر تصادفی Z_i دارای توزیع کناری بتا_دوجمله‌ای با تابع احتمال

$$P(Z_i = z_i) = \frac{\mathrm{B}(\alpha + z_i, \beta + n_i - z_i)}{\mathrm{B}(\alpha,\beta)}\binom{n_i}{z_i} \qquad z_i = ۰,\cdots,m$$

می‌باشد. با استفاده از قاعده مکرر امید ریاضی و واریانس شرطی، امید و واریانس توزیع بتا_دوجمله‌ای به‌صورت زیر نتیجه می‌شوند:

$$E(Z_i) = E[E(Z_i|Y_i)] = n_i E(Y_i) = n_i\pi$$
$$Var(Z_i) = Var[E(Z_i|Y_i)] + E[Var(Z_i|Y_i)] = n_i\sigma_Y^{۲}(n_i - ۱) + n_i\pi(۱-\pi).$$

این روابط تاثیر میدان تصادفی پنهان بتا را با وارد کردن پارامترهای شکل α و β در میانگین و واریانس مدل بتا_دوجمله‌ای نسبت به دوجمله‌ای نشان می‌دهند. میانگین و واریانس توزیع دوجمله‌ای به‌ترتیب $n_i\pi$ و $n_i\pi(۱-\pi)$ است. بنابراین عبارت $n_i\sigma_Y^{۲}(n_i - ۱)$ در واریانس توزیع بتا_دوجمله‌ای به‌عنوان عامل افزاینده واریانس توزیع دوجمله‌ای عمل کرده و مساله بیش‌برازش داده‌ها را لحاظ می‌کند.

برای محاسبه کوواریانس شرطی بین متغیرهای Z_i و Z_j می‌توان نوشت

$$\begin{aligned} E(Z_iZ_j|Y) &= Cov(Z_i, Z_j|Y) + E(Z_i|Y_i)E(Z_j|Y_j) \\ &= \delta_{ij}n_iy_i(۱-y_i) + n_in_jy_iy_j \end{aligned}$$

که در آن δ_{ij} دلتای کرونکر است. یعنی

$$\delta_{ij} = \begin{cases} ۱ & i = j \\ ۰ & i \neq j \end{cases}$$

در این صورت $Cov\left(Z_i, Z_j | Y\right) = ۰$ هرگاه Z_i و Y_i ناهمبسته باشد و هم‌چنین اگر $i = j$ آن‌گاه

$$Cov\left(Z_i, Z_j | Y\right) = Var\left(Z_i | Y_i\right).$$

هدف پیش‌گویی فضایی پاسخ‌های نرخ، با استفاده از توزیع بتا_دوجمله‌ای است. بنابراین در این مرحله، مدل از داده‌های شمارشی مشاهده‌شده بتا_دوجمله‌ای به نسبت‌های نمونه‌ای $\frac{Z_i}{n_i}$ تغییر می‌یابد. در ادامه ابتدا روابطی را بین نسبت‌های نمونه‌ای در دو مکان فضایی s_i و s_j بیان می‌کنیم که در نهایت این روابط منجر به تغییرنگار نظری خواهند شد.

امید ریاضی شرطی اختلاف نسبت‌های نمونه‌ای شرطی $\frac{Z_i}{n_i} - \frac{Z_j}{n_j}$ به‌صورت

$$\begin{aligned} E\left[\frac{Z_i}{n_i} - \frac{Z_j}{n_j} | Y\right] &= \frac{۱}{n_i} E\left(Z_i | Y_i\right) - \frac{۱}{n_j} E\left(Z_j | Y_j\right) \\ &= Y_i - Y_j \end{aligned}$$

است و امید ریاضی کناری آن عبارت است از

$$\begin{aligned} E\left[\frac{Z_i}{n_i} - \frac{Z_j}{n_j}\right] &= E\left(E\left[\left(\frac{Z_i}{n_i} - \frac{Z_j}{n_j} | Y\right)\right]\right) \\ &= E\left(Y_i - Y_j\right] \\ &= \pi - \pi = ۰. \end{aligned}$$

به‌طور کلی فرض می‌کنیم که میدان تصادفی پنهان بتا مانا و دارای میانگین ثابت است. بنابراین اختلاف امید ریاضی حاشیه‌ای در دو مکان s_i و s_j صفر خواهد شد.

حال امید ریاضی شرطی اختلاف توان دوم نسبت‌ها را به‌دست می‌آوریم. در این‌جا فرض

می‌کنیم که تمام مکان‌های فضایی مجزا هستند، یعنی $s_i \neq s_j$. بنابراین

$$\begin{aligned}
E\left[\left(\frac{Z_i}{n_i}-\frac{Z_j}{n_j}\right)^{۲}|Y\right] &= E\left[\left(\frac{Z_i^{۲}}{n_i^{۲}}+\frac{Z_j^{۲}}{n_j^{۲}}-۲\frac{Z_iZ_j}{n_in_j}\right)|Y\right] \\
&= \frac{۱}{n_i^{۲}}E\left(Z_i^{۲}|Y_i\right)+\frac{۱}{n_j^{۲}}E\left(Z_j^{۲}|Y_i\right)-\frac{۲}{n_in_j}E\left(Z_iZ_j|Y\right) \\
&= \frac{۱}{n_i^{۲}}\left[n_iY_i+n_i^{۲}Y_i^{۲}-n_iY_i^{۲}\right]+\frac{۱}{n_j^{۲}}\left[n_jY_j+n_j^{۲}Y_j^{۲}-n_jY_j^{۲}\right]- \\
&\frac{۲}{n_in_j}\left[n_in_jY_iY_j\right] \\
&= \left[\frac{Y_i}{n_i}+Y_i^{۲}-\frac{Y_i^{۲}}{n_i}\right]+\left[\frac{Y_j}{n_j}+Y_j^{۲}-\frac{Y_j^{۲}}{n_j}\right]-۲Y_iY_j \\
&= \frac{Y_i}{n_i}-\frac{Y_i^{۲}}{n_i}+\frac{Y_j}{n_j}-\frac{Y_j^{۲}}{n_j}+\left(Y_i-Y_j\right)^{۲} \\
&= \frac{۱}{n_i}Y_i\left(۱-Y_i\right)+\frac{۱}{n_j}Y_j\left(۱-Y_j\right)+\left(Y_i-Y_j\right)^{۲}.
\end{aligned}$$

امید ریاضی کناری اختلاف توان دوم نسبت‌ها عبارت است از

$$\begin{aligned}
E\left[\left(\frac{Z_i}{n_i}-\frac{Z_j}{n_j}\right)^{۲}\right] &= E\left[E\left(\left[\frac{Z_i}{n_i}-\frac{Z_j}{n_j}\right]^{۲}|Y\right)\right] \\
&= E\left[\frac{۱}{n_i}Y_i\left(۱-Y_i\right)+\frac{۱}{n_j}Y_j\left(۱-Y_j\right)+\left(Y_i-Y_j\right)^{۲}\right] \\
&= \frac{۱}{n_i}\pi-\frac{۱}{n_i}\left(\sigma_Y^{۲}+\pi^{۲}\right)+\frac{۱}{n_j}\pi-\frac{۱}{n_j}\left(\sigma_Y^{۲}+\pi^{۲}\right)+۲\gamma_Y\left(s_i-s_j\right) \\
&= \pi\frac{n_i+n_j}{n_in_j}-\frac{n_i+n_j}{n_in_j}\left(\sigma_Y^{۲}+\pi^{۲}\right)+۲\gamma_Y\left(s_i-s_j\right) \\
&= \frac{n_i+n_j}{n_in_j}\left(\pi-\sigma_Y^{۲}-\pi\right)+۲\gamma_Y\left(s_i-s_j\right).
\end{aligned}$$

در نهایت روابط فوق را برای بیان نسبت‌های نمونه‌ای بتاـ دوجمله‌ای $\frac{Z_i}{n_i}$ در شرایط تغییرنگار میدان تصادفی پنهان بتا، γ_Z، به‌صورت

$$\begin{aligned}
Var\left[\frac{Z_i}{n_i}-\frac{Z_j}{n_j}\right] &= E\left[\left(\frac{Z_i}{n_i}-\frac{Z_j}{n_j}\right)^{۲}\right]-E\left[\frac{Z_i}{n_i}-\frac{Z_j}{n_j}\right]^{۲} \\
&= \pi\left(\frac{n_i+n_j}{n_in_j}\right)-\frac{n_i+n_j}{n_in_j}\left(\sigma_Y^{۲}+\pi^{۲}\right)+۲\gamma_Y\left(s_i-s_j\right)
\end{aligned}$$

می‌توان بازنویسی کرد. طبق تعریف، تغییرنگار به‌صورت $\frac{۱}{۴}Var\left[\frac{Z_i}{n_i}-\frac{Z_j}{n_j}\right]$ می‌باشد. بنابراین با ضرب معادله فوق در $\frac{۱}{۴}$، صورت نهایی تغییرنگار به‌صورت زیر حاصل می‌شود:

$$\frac{۱}{۲}Var\left[\frac{Z_i}{Z_j}-\frac{Z_j}{n_j}\right]=\frac{۱}{۲}\left(\frac{n_i+n_j}{n_in_j}\right)\left(\pi\left(۱-\pi\right)-\sigma_Y^{۲}\right)+\gamma_Y\left(s_i-s_j\right).$$

به‌عبارت دیگر

$$\gamma_Z\left(s_i-s_j\right)=\frac{۱}{۲}\left(\frac{n_i+n_j}{n_in_j}\right)\left(\pi\left(۱-\pi\right)-\sigma_Y^{۲}\right)+\gamma_Y\left(s_i-s_j\right).$$

بنابراین تنها اختلاف تغییرنگار نظری نسبت‌های دوجمله‌ای مشاهده‌شده و توزیع بتا پنهان، توسط تابع ثابت پارامترهای بتا پنهان و اندازه نمونه است. پس

$$\gamma_Y(s_i - s_j) = \gamma_Z(s_i - s_j) - \frac{1}{2}\left(\frac{n_i + n_j}{n_i n_j}\right)\left(\pi(1-\pi) - \sigma_Y^2\right). \qquad (۲.۲)$$

علاوه بر این، امید ریاضی شرطی اختلاف توان دوم واریانس نیز برابر با

$$\begin{aligned} E\left[Var\left(\left[\frac{Z_i}{n_i} - \frac{Z_j}{n_j}\right]^2 |Y\right)\right] &= E\left[\frac{1}{n_i}Y_i(1-Y_i)\right] + E\left[\frac{1}{n_j}Y_j(1-Y_j)\right] \\ &= \frac{1}{n_i}\pi - \frac{1}{n_j}\left(\sigma_Y^2 + \pi^2\right) + \frac{1}{n_j}\pi - \frac{1}{n_j}\left(\sigma_Y^2 + \pi^2\right) \\ &= \frac{n_i + n_j}{n_i n_j}\left(\pi(1-\pi) - \sigma_Y^2\right) \end{aligned}$$

است.

۳.۲ برآوردگر تغییرنگار بتا_دوجمله‌ای

اکنون m مشاهده Z_i، $i = 1, \cdots, m$، از توزیع دوجمله‌ای با اندازه نمونه n_i در مکان s_i را در نظر بگیرید. یک برآورد تعدیل‌یافته برای تغییرنگار با استفاده از رابطه (۲.۲) به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned} \gamma_Y(h) \quad = \quad & \frac{1}{N(h)}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{n_i n_j}{n_i + n_j}\left[\frac{1}{2}\left(\frac{Z_i}{n_i} - \frac{Z_j}{n_j}\right)^2 - \right. \\ & \left.\frac{1}{2}\left(\frac{n_i + n_j}{n_i n_j}\right)\left(\hat{\pi}(1-\hat{\pi}) - \hat{\sigma}_Y^2\right)\right] I_d(i,j) \sim h \end{aligned} \qquad (۳.۲)$$

که در آن منظور از $I_d(i,j) \sim h$ تابع نشانگر برای نقاط i و j ای است که در فاصله h از هم قرار دارند. عدد $N(h)$ نیز تعداد نقاط همسایه در فاصله h است که بر حسب وزن‌های $\frac{n_i n_j}{n_i + n_j}$ برای همگن کردن تمام مکان‌های فضایی به‌صورت زیر تعریف می‌شود:

$$N(h) = \sum_{i,j}\frac{n_i n_j}{n_i + n_j} I_d(i,j) \sim h.$$

به‌عبارت دیگر این وزن تغییر تعداد مشاهدات در هر مکان فضایی را نشان می‌دهد. اگر همه مکان‌های فضایی اندازه نمونه یکسانی را داشته باشند، یعنی برای $i \neq j$، $n_i = n_j$، آن‌گاه به تمام نقاط مکانی وزن ثابت n داده می‌شود. مولفه تصحیح اریبی، یعنی

$$\left(\frac{n_i + n_j}{n_i n_j}\right)\left(\hat{\pi}(1-\hat{\pi}) - \sigma_Y^2\right)$$

به‌طور مستقیم از رابطه (۲.۲) تغییرنگار به‌دست می‌آید. همچنین $\hat{\pi}$ و $\hat{\sigma}_Y^2$ به‌ترتیب برآوردهای میانگین و واریانس توزیع بتای پنهان را نمایش می‌دهند.

در بسیاری از برنامه‌های کاربردی دانستن شکل توزیع بتا زیر بنایی (متقارن، چوله به چپ یا چوله به‌راست) مد نظر می‌باشد. بنابراین با جای‌گذاری پارمترهای شکل α و β در عبارت $\left(\hat{\pi}\left(۱-\hat{\pi}\right)-\hat{\sigma}_Y^۲\right)$، خواهیم داشت

$$\begin{aligned}\pi\left(۱-\pi\right)-\sigma_Y^۲ &= \pi-\pi^۲-\sigma_Y^۲\\ &= \frac{\alpha}{\alpha+\beta}-\frac{\alpha^۲}{\left(\alpha+\beta\right)^۲}-\frac{\alpha\beta}{\left(\alpha+\beta\right)^۲\left(\alpha+\beta+۱\right)}\\ &= \frac{\alpha\left(\alpha+\beta\right)\left(\alpha+\beta+۱\right)-\alpha^۲\left(\alpha+\beta+۱\right)-\alpha\beta}{\left(\alpha+\beta\right)^۲\left(\alpha+\beta+۱\right)}\\ &= \frac{\alpha^۲\left(\alpha+\beta+۱\right)+\alpha\beta\left(\alpha+\beta+۱\right)-\alpha\beta-\alpha^۲\left(\alpha+\beta+۱\right)}{\left(\alpha+\beta\right)^۲\left(\alpha+\beta+۱\right)}\\ &= \frac{\alpha\beta\left(\alpha+\beta\right)}{\left(\alpha+\beta\right)^۲\left(\alpha+\beta+۱\right)}\\ &= \frac{\alpha\beta}{\left(\alpha+\beta\right)\left(\alpha+\beta+۱\right)}.\end{aligned}$$

بنابراین عبارت نهایی برآوردگر تغییرنگار با استفاده از پارامترهای شکل توزیع بتای پنهان به‌صورت زیر تبدیل می‌شود:

$$\begin{aligned}\gamma_Y(h) \;=\; &\frac{۱}{۲N(h)}\sum_{i=۱}^{m}\sum_{j=۱}^{m}\left[\frac{n_in_j}{n_i+n_j}\left(\frac{Z_i}{n_i}-\frac{Z_j}{n_j}\right)^۲-\right.\\ &\left.\left(\frac{\hat{\alpha}\hat{\beta}}{\left(\hat{\alpha}+\hat{\beta}\right)\left(\hat{\alpha}+\hat{\beta}+۱\right)}\right)\right]I_d(i,j)\sim h.\end{aligned}\qquad (۴.۲)$$

۴.۲ معادلات کریگینگ بتا_دوجمله‌ای

کریگینگ فضایی یک روش درونیابی است که در آن مقادیر پیش‌گویی ترکیب‌های خطی از مقادیر مشاهده‌شده، با وزن‌های تعریف‌شده به‌صورت تابعی از فاصله بین مکان مشاهده‌شده و مکان پیش‌گویی‌شده هستند. برای هر مکان فضایی جدید $s_۰$ مقدار پیش‌گویی فرآیند بتای پنهان، $Y_۰=Y(s_۰)$، می‌تواند بر حسب یک ترکیب خطی از نسبت‌های نمونه‌ای دوجمله‌ای $\frac{Z_i}{n_i}$ که در مکان‌های s_i مشاهده شده‌اند با وزن‌های λ_i، $i=۱,\ldots,m$، پیش‌گویی شود. یعنی

$$Y_۰^*=\sum_{i=۱}^{m}\lambda_i\frac{Z_i}{n_i}.$$

وزن‌های λ_i باید به‌گونه‌ای انتخاب شوند که پیش‌گوی $Y_۰$ یک پیش‌گوی نااریب با کم‌ترین واریانس باشد. برای اطمینان از ناریبی پیش‌گو باید قید $\sum_{i=۱}^{m}\lambda_i=۱$ را روی ضرایب اعمال کنیم. اگر مقدار پیش‌گویی $Y_۰^*$ ناریب باشد، آن‌گاه $E\left(Y_۰^*\right)=\pi$ که π میانگین توزیع بتای

پنهان است. بنابراین امید ریاضی شرطی در مکان جدید به‌صورت زیر است:

$$\begin{aligned} E\left[Y_{\circ}^{*}|Y\right] &= E\left[\sum_{i=1}^{m}\lambda_i\frac{Z_i}{n_i}|Y\right] \\ &= \sum_{i=1}^{m}\frac{\lambda_i}{n_i}E\left[Z_i|Y\right] \\ &= \sum_{i=1}^{m}\frac{\lambda_i}{n_i}\left[n_iY_i\right] = \sum_{i=1}^{m}\lambda_iY_i. \end{aligned}$$

لذا مقدار امید ریاضی $Y_{\circ}^{*}$ برابر است با

$$\begin{aligned} E\left[Y_{\circ}^{*}\right] &= E\left[\sum_{i=1}^{m}\lambda_iY_i\right] \\ &= \sum_{i=1}^{m}\lambda_iE\left(Y_i\right) \\ &= \pi\sum_{i=1}^{m}\lambda_i. \end{aligned}$$

در نتیجه، ناااریبی کریگینگ مستلزم آن خواهد بود که $\sum_{i=1}^{m}\lambda_i = 1$ باشد. بنابراین اولین شرط کریگینگ بتا_دوجمله‌ای آن است که $\sum_{i=1}^{m}\lambda_i = 1$ باشد. برای بررسی شرط دوم کریگینگ، ابتدا امید ریاضی شرطی توان دوم خطای پیش‌گو را به‌صورت زیر به‌دست می‌آوریم:

$$\begin{aligned} E\left[\left(Y_{\circ}^{*}-Y_{\circ}\right)^{2}|Y\right] &= E\left[\left(\sum_{i=1}^{m}\lambda_i\frac{Z_i}{n_i}-Y_{\circ}\right)^{2}|Y\right] \\ &= E\left[\left\{\left(\sum_{i=1}^{m}\lambda_i\frac{Z_i}{n_i}\right)\left(\sum_{j=1}^{m}\lambda_j\frac{Z_j}{n_j}\right)-2Y_{\circ}\sum_{i=1}^{m}\lambda_i\frac{Z_i}{n_i}+Y_{\circ}^{2}\right\}|Y\right] \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\frac{1}{n_in_j}E\left[Z_iZ_j|Y\right]-2Y_{\circ}\sum_{i=1}^{m}\frac{\lambda_i}{n_i}E\left[Z_i|Y\right]+Y_{\circ}^{2} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\frac{1}{n_in_j}\left[\delta_{ij}n_iY_i\left(1-Y_i\right)+n_in_jY_iY_j\right]-2Y_{\circ}\sum_{i=1}^{m}\lambda_iY_i+Y_{\circ}^{2}. \end{aligned}$$

امید ریاضی شرطی توان دوم خطای پیش‌گو را می‌توان به‌صورت زیر ساده کرد:

$$E\left[\left(Y_{\circ}^{*}-Y_{\circ}\right)^{2}|Y\right] = \sum_{i=1}^{m}\frac{\lambda_i^{2}}{n_i}Y_i\left(1-Y_i\right)+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_jY_iY_j-2Y_{\circ}\sum_{i=1}^{n}\lambda_iY_i+Y_{\circ}^{2}.$$

بنابراین

$$\begin{aligned}
E\left[(Y_{\circ}^{*}-Y_{\circ})\right] &= E\left[\sum_{i=1}^{m}\frac{\lambda_i^{2}}{n_i}Y_i\left(1-Y_i\right)+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j Y_iY_j-2Y_{\circ}\sum_{i=1}^{m}\lambda_iY_i+Y_{\circ}^{2}\right]\\
&=\sum_{i=1}^{m}\frac{\lambda_i^{2}}{n_i}E\left[Y_i\left(1-Y_i\right)\right]+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_jE\left[Y_iY_j\right]-2\sum_{i=1}^{m}\lambda_iE\left[Y_{\circ}Y_i\right]+E\left[Y_{\circ}^{2}\right]\\
&=\sum_{i=1}^{m}\frac{\lambda_i^{2}}{n_i}\left[E\left(Y_i\right)-E\left(Y_i^{2}\right)\right]+\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\left[Cov\left(Y_i,Y_j\right)+E\left(Y_i\right)E\left(Y_j\right)\right]\\
&-2\sum_{i=1}^{m}\lambda_i\left[Cov\left(Y_{\circ},Y_i\right)+E\left(Y_{\circ}\right)E\left(Y_i\right)\right]+\left[Var\left(Y_{\circ}\right)+E\left(Y_{\circ}^{2}\right)\right]\\
&=\sum_{i=1}^{m}\frac{\lambda_i^{2}}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^{2}\right]+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\left[C_Y\left(\left|s_i-s_j\right|\right)+\pi^{2}\right]\\
&-2\sum_{i=1}^{m}\lambda_i\left[C_Y\left(\left|s_{\circ}-s_i\right|\right)+\pi^{2}\right]+\sigma_Y^{2}+\pi^{2}\\
&=\sum_{i=1}^{m}\frac{\lambda_i^{2}}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^{2}\right]+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_jC_Y\left(\left|s_i-s_j\right|\right)+\pi^{2}\\
&-2\sum_{i=1}^{m}\lambda_iC_Y\left(s_{\circ}-s_i\right)-2\pi^{2}+\sigma_Y^{2}+\pi^{2}\\
&=\sum_{i=1}^{m}\frac{\lambda_i^{2}}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^{2}\right]+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_jC_{ij}-2\sum_{i=1}^{m}\lambda_iC_{\circ i}+\sigma_Y^{2}
\end{aligned}$$

که در آن

$$C_{ij}=C_Y\left(\left|s_i-s_j\right|\right).$$

در نتیجه واریانس خطای پیش‌گو به‌صورت

$$\begin{aligned}
Var\left(Y_{\circ}^{*}-Y_{\circ}\right) &= E\left[\left(Y_{\circ}^{*}-Y_{\circ}\right)^{2}\right]-E\left[Y_{\circ}^{*}-Y_{\circ}\right]^{2}\\
&=\left[\sum_{i=1}^{m}\frac{\lambda_i^{2}}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^{2}\right]+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_jC_{ij}-2\sum_{i=1}^{m}C_{\circ i}+\sigma_Y^{2}\right]-0
\end{aligned}$$

به‌دست می‌آید. می‌دانیم که $Y_{\circ}^{*}$ نا اریب است، یعنی $E\left[Y_{\circ}^{*}-Y_{\circ}\right]^{2}=0$، بنابراین واریانس خطای پیش‌گو برابر است با

$$Var\left(Y_{\circ}^{*}-Y_{\circ}\right)=\sum_{i=1}^{m}\frac{\lambda_i^{2}}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^{2}\right]+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_jC_{ij}-2\sum_{i=1}^{m}\lambda_iC_{\circ i}+\sigma_Y^{2}.$$

اکنون با استفاده از روش لاگرانژ، واریانس خطای پیش‌گو را می‌نیمم می‌کنیم. با مشتق‌گیری نسبت به λ_i داریم

$$\frac{\partial}{\partial\lambda_i}Var\left(Y_\circ^*-Y_\circ\right)=\frac{\partial}{\partial\lambda_i}\{\sum_{i=1}^{m}\frac{\lambda_i^2}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^2\right]+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_jC_{ij}-$$
$$2\sum_{i=1}^{m}\lambda_iC_{\circ i}+\sigma_Y^2+2\mu\left[\sum_{i=1}^{m}\lambda_i-1\right]\}$$

با مساوی قرار دادن مشتق با صفر دستگاه معادلات

$$\begin{cases}\circ &= \frac{2\lambda_i}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^2\right]+2\sum_{i=1}^{m}\lambda_jC_{ij}-2C_{\circ i}+2\mu\\ \circ &= \frac{\lambda_i}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^2\right]+\sum_{j=1}^{m}\lambda_jC_{ij}-C_{\circ i}+\mu\end{cases}$$

حل می‌کنیم. بنابراین معادله کریگینگ بتاـ دوجمله‌ای به‌صورت زیر نتیجه می‌شود:

$$\frac{\lambda_i}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^2\right]+\sum_{j=1}^{m}\lambda_jC_{ij}+\mu=C_{\circ i}\qquad i\in 1,2,\cdots,n$$
$$\sum_{i=1}^{m}\lambda_i=1.$$

معادلات کریگینگ را به راحتی می‌توان به‌صورت ماتریسی نوشت. در معادله بالا داریم

$$\boldsymbol{C}\underline{\boldsymbol{\lambda}}=\underline{\boldsymbol{C}}_\circ$$

که در آن C ماتریس ضرایب، $\underline{\boldsymbol{\lambda}}=[\lambda_1,\cdots,\lambda_n,\ \mu]'$ بردار وزن‌های کریگینگ و μ نیز ضریب لاگرانژ است. همچنین

$$\underline{\boldsymbol{C}}_\circ=[\boldsymbol{C}_{\circ 1},\cdots,\boldsymbol{C}_{\circ\boldsymbol{n}},\ 1]'$$

بردار معادلات کریگینگ است.

معادله کریگینگ معمولی به‌صورت

$$\sum_{i=1}^{n}\lambda_iC_{ij}+\mu=C_{\circ i}\qquad i\in 1,2,\cdots,n$$
$$\sum_{i=1}^{m}\lambda_i=1$$

را در نظر بگیرید. ماتریس ضرایب کریگینگ بتاـ دوجمله‌ای در مقایسه با کریگینگ معمولی پیچیده‌تر است، که این پیچیدگی اضافی از عبارت $\frac{\lambda_i}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^2\right]$ ناشی شده است. در واقع در کریگینگ بتاـ دوجمله‌ای، این جمله به عناصر قطر اصلی ماتریس ضرایب اضافه شده

است. بنابراین

$$C=\begin{bmatrix} \hat{C}_{11}+\left[\frac{\pi(1-\pi)-\sigma_Y^2}{n_1}\right] & \hat{C}_{12} & \dots & \hat{C}_{1n} & 1 \\ \hat{C}_{21} & +\hat{C}_{22}+\left[\frac{\pi(1-\pi)-\sigma_Y^2}{n_2}\right] & \dots & \cdot & 1 \\ \vdots & \vdots & \ddots & & \\ \hat{C}_{n1} & \vdots & \dots & \hat{C}_{n_m}+\left[\frac{\pi(1-\pi)-\sigma_Y^2}{n_m}\right] & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix}$$

به‌طوری‌که $\hat{C}_{ij}=\hat{\gamma}_Y(|s_i-s_j|)$ نشان‌دهنده تابع کوواریانس برآوردشده برای نقاط i و j می‌باشد. وزن‌های کریگینگ بتاـ دوجمله‌ای با حل معادله ماتریسی $C\underline{\lambda}=\underline{C}_0$ نتیجه می‌شوند. بنابراین

$$\underline{\boldsymbol{\lambda}}=\boldsymbol{C}^{-1}\underline{\boldsymbol{C}}_0.$$

برای انجام دادن کریگینگ بتاـ دوجمله‌ای باید مجموعه معادلات برای هر مکان پیش‌گویی حل شوند، بنابراین زمانی‌که پیش‌گویی بیش از یک شبکه فضایی متراکم مد نظر باشد، ممکن است که محاسبات سنگین و زمان‌بر باشند.

همچنین واریانس خطای پیش‌گو را می‌توان در شرایط واریانس کریگینگ معمولی یعنی

$$Var_{OK}=\sigma^2-\sum_{i=1}^{m}\lambda_i C_{0i}-\mu$$

بازنویسی کرد. بنابراین با بازنویسی و ساده کردن عبارات داریم

$$\begin{aligned} Var\left(Y_0^*-Y_0\right) &= \sigma_Y^2-2\sum_{i=1}^{m}\lambda_i C_{0i}+\sum_{i=1}^{m}\frac{\lambda_i^2}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^2\right]+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j C_{ij} \\ &= \sigma_Y^2-\sum_{i=1}^{m}\lambda_i C_{0i}+\sum_{i=1}^{m}\frac{\lambda_i^2}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^2\right]+\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j C_{ij} \\ &\quad -\sum_{i=1}^{m}\lambda_i C_{0i} \\ &= \sigma_Y^2-\sum_{i=1}^{m}\lambda_i C_{0i}+\sum_{i=1}^{m}\lambda_i\left\{\frac{\lambda_i}{n_i}\left[\pi\left(1-\pi\right)-\sigma_Y^2\right]+\sum_{j=1}^{m}\lambda_j C_{ij}-C_{0i}\right\} \\ &= \sigma_Y^2-\sum_{i=1}^{m}\lambda_i C_{0i}-\sum_{i=1}^{m}\lambda_i\mu \\ &= \sigma_Y^2-\sum_{i=1}^{m}\lambda_i C_{0i}-\mu \end{aligned}$$

از رابطه بالا می‌توان چنین استنباط کرد که پیچیدگی اضافی نسبت‌های نمونه‌ای توزیع بتاـ دوجمله‌ای، از واریانس خطای پیش‌گو ناشی نشده و این پیچیدگی فقط در ماتریس کوواریانس وجود دارد.

در فصل بعد عملکرد برآوردگر تغییرنگار و کریگینگ بتا_دوجمله‌ای را مورد ارزیابی قرار می‌دهیم.

فصل ۳

دقت برآوردیابی و پیش‌گویی مدل بتا_ دوجمله‌ای فضایی

۱.۳ مقدمه

در این فصل ابتدا به تشریح نحوه تولید داده‌های بتا_ دوجمله‌ای همبسته فضایی با استفاده از روش تبدیل معکوس می‌پردازیم. سپس سه روش مختلف را برای برآورد پارامترهای تغییرنگار معرفی کرده، که پس از بررسی دقت آن‌ها بهترین روش برای برآورد پارامترهای کریگینگ بتا_ دوجمله‌ای را انتخاب می‌کنیم. علاوه بر این، دقت مدل در پیش‌گویی را با معیارهای مختلف مورد ارزیابی قرار می‌دهیم.

۱.۱.۳ روش تبدیل معکوس برای شبیه‌سازی داده‌ها

یاها و شیمولی (۲۰۰۷)، روش تبدیل معکوس را به‌عنوان روشی کارا برای تولید متغیر تصادفی پواسون چند متغیره معرفی کردند. در بعضی از متون این روش را تحت عنوان تبدیل NORTA یعنی''نرمال به هر چیزی'' می‌نامند. اساس این روش تبدیل از توزیع نرمال چندمتغیره به هر توزیع احتمالی دیگر با استفاده از تابع توزیع تجمعی است.

فرآیند شبیه‌سازی روش تبدیل معکوس شامل سه مرحله به‌صورت زیر است:

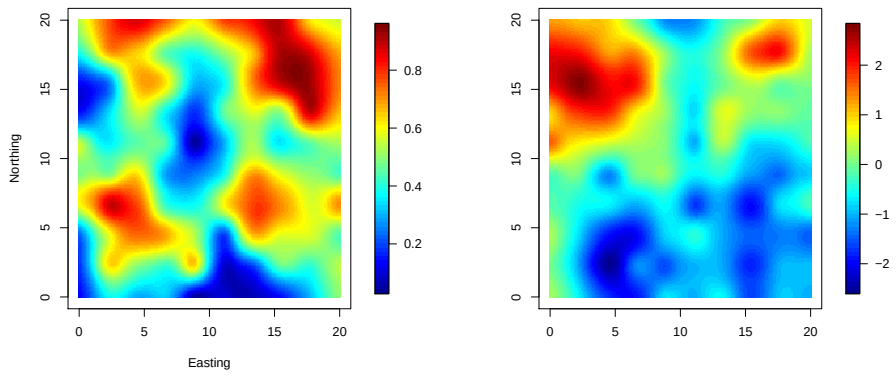
۱. ابتدا بردار توزیع نرمال p بعدی $\boldsymbol{X^N}$، با میانگین $\boldsymbol{\mu} = 0$، و ماتریس همبستگی R^N را تولید می‌کنیم.

۲. تابع توزیع تجمعی را برای هر مقدار بردار توزیع نرمال X_i^N، $i \in 1,2,\cdots,p$، به‌صورت $\Phi(X_i)$ محاسبه می‌کنیم.

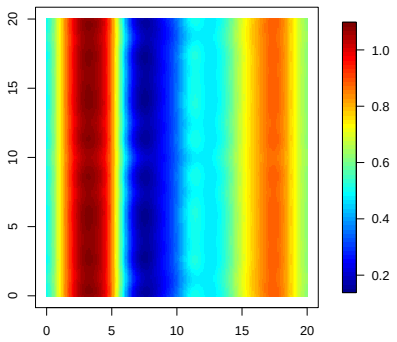
۳. در نهایت معکوس تابع توزیع تجمعی (تابع چندک) توزیع دلخواه، مثلا پواسون با پارمتر نرخ λ_i، را به صورت

$$X_i^D = \Xi^{-1}\left(\Phi\left(X_i^N\right)\right)$$

به‌دست می‌آوریم. در نتیجه $\boldsymbol{X^D}$ یک بردار تصادفی p بعدی از توزیع دلخواه با ماتریس همبستگی R^D است.



(آ) میدان تصادفی نرمال (ب) میدان تصادفی بتا



(ج) میدان نسبت‌های دوجمله‌ای

شکل ۱.۳: میدان‌های شبیه‌سازی‌شده با روش تبدیل معکوس

روش تبدیل معکوس در مقایسه با سایر روش‌ها، ساده و انعطاف‌پذیر است. در این روش

هیچ گونه محدودیتی بر روی ماتریس همبستگی داده‌های نرمال R^D وجود ندارد. همچنین یاها و شیمولی پیشنهاد کردند از متغیرهای تصادفی نرمال استاندارد استفاده شود که علت این انتخاب فقط سادگی آن می‌باشد. مزیت دیگر این روش آن است که بر روی اکثر توزیع‌ها به‌خوبی عمل می‌کند. البته این روش وقتی که توزیع مورد نظر پیوسته باشد، عملکرد بهتری دارد. همچنین یاها و شیمولی اشاره کردند که سختی‌هایی در تعریف معکوس تابع توزیع تجمعی توزیع‌های گسسته مثل پواسون وجود دارند و راه حل‌هایی را برای آن‌ها پیشنهاد دادند. برای اطلاعات بیش‌تر به یاها و شیمولی (۲۰۰۷) مراجعه کنید.

۲.۱.۳ شبیه‌سازی میدان فضایی بتا_دوجمله‌ای

همان‌طور که در بخش قبل عنوان شد، از روش تبدیل معکوس برای تولید نمونه از توزیع‌های مختلف می‌توان استفاده کرد. بنابراین از این روش برای تولید داده‌های بتا_دوجمله‌ای همبسته فضایی استفاده می‌کنیم که مراحل آن به‌صورت زیر هستند:

۱. ابتدا میدان تصادفی نرمال $X(s)$ با ماتریس کوواریانس فضایی $\boldsymbol{\Sigma_X}$ و تابع کوواریانس فضایی γ_X را شبیه‌سازی می‌کنیم. در این مطالعه شبیه‌سازی، نقاط مکانی، گره‌های یک مشبکه فضایی ۲۰ × ۲۰ در نظر گرفته شدند. برای تابع کوواریانس فضایی نیز یک مدل همسان‌گرد کروی با پارامترهای واقعی اثر قطعه‌ای $\tau_X^2 = 0$، دامنه فضایی $\phi_X = 5, 10, 15$ و ازاره فضایی $\sigma_X^2 = 1$ در نظر گرفته شد. بنابراین

$$\boldsymbol{X(s)} \sim N_m(\boldsymbol{\mu_X} = 0, \boldsymbol{\Sigma_X})$$

$$\gamma_X(h) = \begin{cases} \tau_X^2 + \sigma_X^2 \left(\frac{3h}{2\phi_X} - \frac{h^3}{2\phi_x^3}\right) & 0 < h < \phi_X \\ \tau_X^2 + \sigma_X^2 & h \geq \phi_X \end{cases}$$

۲. برای هر نقطه مکانی در مشبکه فضایی، تابع توزیع تجمعی نرمال استاندارد، یعنی $\Phi_X(X_i)$، را برای تبدیل میدان تصادفی نرمال به میدان تصادفی بتا محاسبه می‌کنیم. اگر تابع چندک (معکوس تابع توزیع تجمعی) توزیع بتا $Y_i \sim \text{Beta}(\alpha, \beta)$ را با $\Xi_Y^{-1}(\cdot)$ نشان دهیم، آن‌گاه

$$\Xi_Y^{-1}(\Phi(X_i)) = Y_i.$$

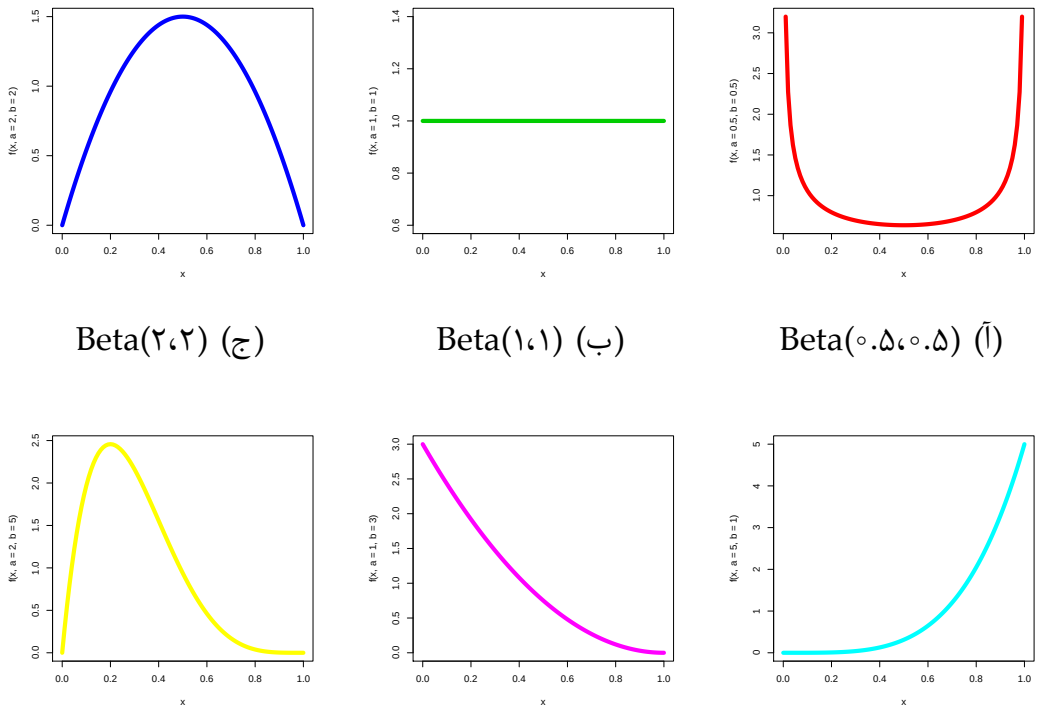
چون هر دو توزیع پیوسته هستند، تبدیل انجام‌شده یک به یک خواهد بود. توزیع بتا دارای ساختار وابستگی $\boldsymbol{\Sigma_Y}$ است، که این ساختار هم از یک مدل کروی با اثر قطعه‌ای $\tau_Y^2 = 0$ پیروی می‌کند. پارامتر دامنه نیز مشابه مدل میدان تصادفی نرمال است. فقط پارامتر ازاره فضایی برای میدان تصادفی بتا برابر با واریانس توزیع بتا به‌صورت زیر خواهد بود.

$$\sigma_Y^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

۳. یک زیرمجموعه m تایی از مکان‌های فضایی با اندازه نمونه تصادفی n_i را به‌صورت تصادفی انتخاب می‌کنیم. سپس در هر مکان نمونه‌گیری، نمونه تصادفی دوجمله‌ای $Z_i|Y_i \sim bin\,(n_i, Y_i)$ تولید می‌شود. بنابراین متغیر تصادفی Z_i دارای توزیع بتا_دوجمله‌ای همبسته فضایی است. شکل ۱.۳ فرآیند تولید داده‌های بتا_دوجمله‌ای همبسته فضایی را نمایش می‌دهد.

۳.۱.۳ مشخصات داده‌های شبیه‌سازی شده بتا_دوجمله‌ای فضایی

برای ارزیابی عملکرد مدل بتا_دوجمله‌ای پیشنهادی در برآورد پارامترها و پیش‌گویی فضایی، از یک مطالعه شبیه‌سازی استفاده کردیم. به‌منظور نمایش انعطاف توزیع‌های بتا که شامل توزیع‌های متقارن و چوله می‌باشد، در این پایان‌نامه شش توزیع بتا مختلف را انتخاب کردیم. توابع چگالی احتمال توزیع‌های انتخاب شده در شکل ۲.۳ نشان داده شده‌اند. همچنین جدول۱.۳ میانگین و واریانس این شش توزیع بتا را نشان می‌دهد.



Beta(۰.۵،۰.۵) (آ) Beta(۱،۱) (ب) Beta(۲،۲) (ج)

Beta(۵،۱) (د) Beta(۱،۳) (ه) Beta(۲،۵) (و)

شکل ۲.۳: توابع چگالی بتا

جدول ۱.۳: میانگین و واریانس توزیع‌های بتا

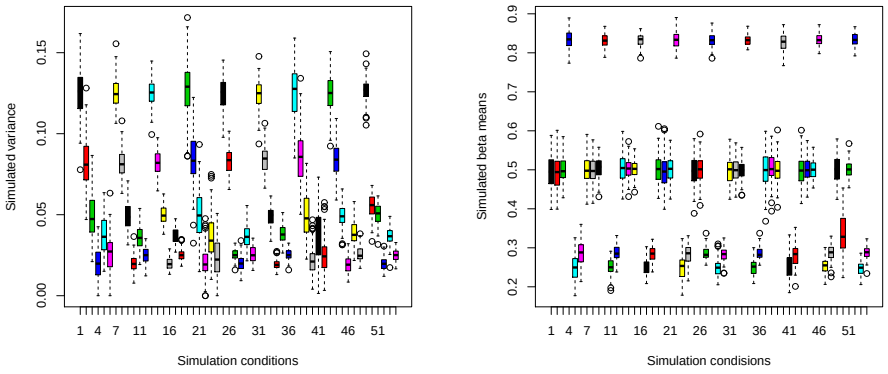
موارد	پارامترهای بتا	میانگین	واریانس
۱	$Beta(۰/۵, ۰/۵)$	۰/۵	۰/۱۲۵
۲	$Beta(۱, ۱)$	۰/۵	۰/۰۸۳
۳	$Beta(۲, ۲)$	۰/۵	۰/۰۵۰
۴	$Beta(۵, ۱)$	۰/۸۳۳	۰/۰۲۰
۵	$Beta(۱, ۳)$	۰/۲۵	۰/۰۳۸
۶	$Beta(۲, ۵)$	۰/۲۸۶	۰/۰۲۶

برای مطالعه شبیه‌سازی، ۵۴ سناریوی مختلف تولید داده را تعریف کردیم؛ برای هر توزیع بتا سه اندازه نمونه در نظر گرفتیم: محدوده بین ۲:۵ برای اندازه نمونه کوچک، محدوده بین ۱۰:۲۰ برای اندازه نمونه متوسط و محدوده بین ۵۰:۱۰۰ برای اندازه نمونه بزرگ تعریف کردیم. پارامتر دامنه فضایی را نیز سه مقدار ۵، ۱۰ و ۱۵ در نظر گرفتیم. مجموع ۵۴ سناریو شبیه‌سازی‌شده در جدول ۲.۳ نشان داده شده‌اند. تمامی داده‌ها را بدون اثر قطعه‌ای، به‌صورت هموار، بر روی مشبکه منظم ۲۰ × ۲۰ تولید کردیم. بنابراین داده‌ها باید توزیع بتا واقعی را نمایش دهند و هرگونه اثر قطعه‌ای که در نسبت‌های نمونه‌ای دوجمله‌ای مشاهده شود فقط از نمونه‌گیری دوجمله‌ای ناشی شده است. همچنین برای هر سناریو ۱۰۰ مجموعه داده شبیه‌سازی کردیم.

جدول ۲.۳: سناریوهای مختلف مطالعه شبیه‌سازی

شماره شبیه‌سازی	توزیع بتا	دامنه	اندازه نمونه
۱	$Beta(۰/۵, ۰/۵)$	۵	small(۲:۵)
۲	$Beta(۱, ۱)$	۵	Small(۲:۵)
۳	$Beta(۲, ۲)$	۵	Small(۲:۵)
۴	$Beta(۵, ۱)$	۵	Small(۲:۵)
۵	$Beta(۱, ۳)$	۵	Small(۲:۵)
۶	$Beta(۲, ۵)$	۵	Small(۲:۵)
۷	$Beta(۰/۵, ۰/۵)$	۵	Medium(۱۰:۲۰)
۸	$Beta(۱, ۱)$	۵	Medium(۱۰:۲۰)
۹	$Beta(۲, ۲)$	۵	Medium(۱۰:۲۰)
۱۰	$Beta(۵, ۱)$	۵	Medium(۱۰:۲۰)
۱۱	$Beta(۱, ۳)$	۵	Medium(۱۰:۲۰)
۱۲	$Beta(۲, ۵)$	۵	Medium(۱۰:۲۰)
۱۳	$Beta(۰/۵, ۰/۵)$	۵	Large(۵۰:۱۰۰)
۱۴	$Beta(۱, ۱)$	۵	Large(۵۰:۱۰۰)
۱۵	$Beta(۲, ۲)$	۵	Large(۵۰:۱۰۰)
۱۶	$Beta(۵, ۱)$	۵	Large(۵۰:۱۰۰)
۱۷	$Beta(۱, ۳)$	۵	Large(۵۰:۱۰۰)
۱۸	$Beta(۲, ۵)$	۵	Large(۵۰:۱۰۰)
۱۹	$Beta(۰/۵, ۰/۵)$	۱۰	Small(۲:۵)
۲۰	$Beta(۱, ۱)$	۱۰	Small(۲:۵)
۲۱	$Beta(۲, ۲)$	۱۰	Small(۲:۵)
۲۲	$Beta(۵, ۱)$	۱۰	Small(۲:۵)
۲۳	$Beta(۱, ۳)$	۱۰	Small(۲:۵)
۲۴	$Beta(۲, ۵)$	۱۰	Small(۲:۵)
۲۵	$Beta(۰/۵, ۰/۵)$	۱۰	Medium(۱۰:۲۰)
۲۶	$Beta(۱, ۱)$	۱۰	Medium(۱۰:۲۰)
۲۷	$Beta(۲, ۲)$	۱۰	Medium(۱۰:۲۰)
۲۸	$Beta(۵, ۱)$	۱۰	Medium(۱۰:۲۰)
۲۹	$Beta(۱, ۳)$	۱۰	Medium(۱۰:۲۰)
۳۰	$Beta(۲, ۵)$	۱۰	Medium(۱۰:۲۰)
۳۱	$Beta(۰/۵, ۰/۵)$	۱۰	Large(۵۰:۱۰۰)
۳۲	$Beta(۱, ۱)$	۱۰	Large(۵۰:۱۰۰)
۳۳	$Beta(۲, ۲)$	۱۰	Large(۵۰:۱۰۰)
۳۴	$Beta(۵, ۱)$	۱۰	Large(۵۰:۱۰۰)
۳۵	$Beta(۱, ۳)$	۱۰	Large(۵۰:۱۰۰)
۳۶	$Beta(۲, ۵)$	۱۰	Large(۵۰:۱۰۰)
۳۷	$Beta(۰/۵, ۰/۵)$	۱۵	Small(۲:۵)
۳۸	$Beta(۱, ۱)$	۱۵	Small(۲:۵)
۳۹	$Beta(۲, ۲)$	۱۵	Small(۲:۵)
۴۰	$Beta(۵, ۱)$	۱۵	Small(۲:۵)
۴۱	$Beta(۱, ۳)$	۱۵	Small(۲:۵)
۴۲	$Beta(۲, ۵)$	۱۵	Small(۲:۵)
۴۳	$Beta(۰/۵, ۰/۵)$	۱۵	Medium(۰:۲۰)
۴۴	$Beta(۱, ۱)$	۱۵	Medium(۱۰:۲۰)
۴۵	$Beta(۲, ۲)$	۱۵	Medium(۱۰:۲۰)
۴۶	$Beta(۵, ۱)$	۱۵	Medium(۱۰:۲۰)
۴۷	$Beta(۱, ۳)$	۱۵	Medium(۱۰:۲۰)
۴۸	$Beta(۲, ۵)$	۱۵	Medium(۱۰:۲۰)
۴۹	$Beta(۰/۵, ۰/۵)$	۱۵	Large(۵۰:۱۰۰)
۵۰	$Beta(۱, ۱)$	۱۵	Large(۵۰:۱۰۰)
۵۱	$Beta(۲, ۲)$	۱۵	Large(۵۰:۱۰۰)
۵۲	$Beta(۵, ۱)$	۱۵	Large(۵۰:۱۰۰)
۵۳	$Beta(۱, ۳)$	۱۵	Large(۵۰:۱۰۰)
۵۴	$Beta(۲, ۵)$	۱۵	Large(۵۰:۱۰۰)

نمودارهای جعبه‌ای شکل ۳.۳ میانگین و واریانس ۱۰۰ مجموعه داده شبیه‌سازی‌شده را تحت سناریوهای مختلف شبیه‌سازی نمایش می‌دهند. همان‌طور که در شکل مشاهده می‌شود هر یک از سناریوهای توزیع بتا به میانگین واقعی (۰٫۲۸۶, ۰٫۲۵, ۰٫۸۳۳, ۰٫۵, ۰٫۵, ۰٫۵) نزدیک هستند. همچنین با افزایش دامنه فضایی در میانگین توزیع بتای شبیه‌سازی‌شده، تغییرات بیش‌تری ایجاد می‌شود. وابستگی فضایی شدید می‌تواند مقادیر میدان تصادفی را افزایش یا کاهش دهد. در واقع مکان‌های فضایی با وابستگی شدید، مکان‌های همسایه را بیش‌تر تحت تاثیر قرار می‌دهند، که این پدیده بیش‌تر برای توزیع‌های بتای متقارن تجلی پیدا می‌کند. با توجه به این که نمونه‌های دوجمله‌ای هنوز گرفته نشده‌اند بنابراین هیچ‌گونه اختلاف قابل تشخیصی بین اندازه‌های نمونه مشاهده نمی‌شود. به‌عنوان مثال، سناریو شبیه‌سازی ۷،۱ و ۱۳ در یک سطح از مدل موثر هستند. همچنین واریانس توزیع بتای شبیه‌سازی شده در این سه سناریو به واریانس واقعی توزیع بتا نزدیک هستند. مشابه میانگین، واریانس نیز به دامنه فضایی وابسته است، به‌طوری که با افزایش دامنه فضایی تغییرات واریانس نیز بیش‌تر می‌شود.



(آ) میانگین توزیع بتا شبیه‌سازی‌شده (ب) واریانس توزیع بتا شبیه‌سازی‌شده

شکل ۳.۳: میانگین و واریانس شبیه‌سازی‌شده

۴.۱.۳ برآورد پارامترهای توزیع بتا

برای مواردی که تعداد مشاهدات در تمامی مکان‌ها یکسان هستند، چندین مطالعه خوب برآورد برای پارامترها شامل $\hat{\pi}$ و $\hat{\sigma}_Y^2$ یا $(\hat{\alpha}, \hat{\beta})$ انجام شده‌اند. با توجه به این‌که پارامترهای شکل α و β به‌طور مستقیم در معادله (۴.۲) تغییرنگار بتا ـ دوجمله‌ای (فصل دوم) مورد استفاده قرار می‌گیرند، بنابراین برآورد دقیق پارامترهای شکل برای برازش مدل بتا ـ دوجمله‌ای ضروری است. با توجه به این که اندازه نمونه n_i در تمامی مکان‌ها یکسان نیست، بنابراین برآوردگرهایی که پیش از این پیشنهاد شده‌اند، مناسب نیستند. تریپتی (۱۹۹۴) برای اندازه نمونه نابرابر استفاده از روش‌های ماکسیمم درستنمایی را پیشنهاد داد، که برای ارزیابی مدل مورد استفاده

قرار خواهد گرفت. کریگینگ بتا_دوجمله‌ای یک مزیت منحصر به فرد را برای مدل‌سازی نمونه‌های کوچک فراهم می‌کند. برآورد ماکسیمم درستنمایی پارامترهای توزیع بتا، برای اندازه نمونه نابرابر مورد بررسی قرار نگرفته است. برای اطمینان از مناسب بودن این روش، یک مطالعه شبیه‌سازی را برای ارزیابی پارامترهای شکل برآوردشده انجام دادیم. نسبت نمونه‌ای $\frac{Z_i}{n_i}$ را به‌عنوان مقادیر مشاهده‌شده برای برآورد $\hat{\alpha}$ و $\hat{\beta}$ در نظر گرفتیم.

برای هر شش توزیع بتا، ۱۰۰۰ نمونه تصادفی بتا_دوجمله‌ای را در ۱۰۰ مکان نمونه‌گیری، روی مشبکه ۲۰×۲۰ تولید کردیم. برآورد IWLS با استفاده از بسته VGAM در نرم‌افزار R برای برآورد پارامترهای توزیع بتا به‌دست آمدند. جدول‌های ۳.۳ ، ۴.۳ و ۵.۳ به‌ترتیب شبیه‌سازی‌ها را برای دامنه‌های فضایی ۵، ۱۰ و ۱۵ نشان می‌دهند. با توجه به نتایج جدول‌ها مشاهده می‌کنیم که پارامترهای برآوردشده بسیار نزدیک به پارامترهای شبیه‌سازی واقعی برای تمام توزیع‌های بتا هستند. البته برای نمونه‌های بزرگ اندکی کم‌برآوردی[1] وجود دارد که خیلی قابل توجه نیست، و ممکن است علت آن وجود اریبی کوچک در میانگین و میانه برآورد پارامترها باشد.

جدول ۳.۳: برآورد پارامترهای شکل $\hat{\alpha}$ و $\hat{\beta}$ برای هر یک از توزیع‌های بتا با دامنه ۵

توزیع بتا		$\hat{\alpha}_{Small}$	$\hat{\beta}_{Small}$	$\hat{\alpha}_{Medium}$	$\hat{\beta}_{Medium}$	$\hat{\alpha}_{Large}$	$\hat{\beta}_{Large}$
$Beta(۰/۵, ۰/۵)$	میانه	۰/۴۹۹	۰/۴۹۹	۰/۴۹۳	۰/۴۹۳	۰/۴۹۴	۰/۴۹۴
	میانگین	۰/۵۰۳	۰/۵۰۲	۰/۴۹۴	۰/۴۹۴	۰/۴۹۷	۰/۴۹۷
	انحراف معیار	۰/۰۵۸	۰/۰۵۷	۰/۰۳۲	۰/۰۳۳	۰/۰۲۲	۰/۰۲۲
$Beta(۱, ۱)$	میانه	۰/۹۸۵	۰/۹۷۸	۰/۹۹۳	۰/۹۹۴	۰/۹۹۵	۰/۹۹۵
	میانگین	۰/۹۹۸	۰/۹۹۸	۰/۹۹۵	۰/۹۹۴	۰/۹۹۷	۰/۹۹۷
	انحراف معیار	۰/۱۲۴	۰/۱۲۶	۰/۰۵۸	۰/۰۵۷	۰/۰۲۸	۰/۰۲۸
$Beta(۲, ۲)$	میانگین	۲/۰۱۱	۱/۹۹۹	۱/۹۹۰	۱/۹۹۲	۱/۹۹۸	۱/۹۹۸
	میانه	۲/۰۳۸	۲/۰۳۶	۲/۰۰۱	۱/۹۹۹	۱/۹۹۹	۱/۹۹۹
	انحراف معیار	۰/۳۱۸	۰/۳۱۶	۰/۱۲۲	۰/۱۲۱	۰/۰۵۰	۰/۰۵۲
$Beta(۵, ۱)$	میانه	۴/۹۵۳	۰/۹۹۸	۴/۹۷۳	۰/۹۹۸	۴/۹۸۹	۰/۹۹۶
	میانگین	۵/۱۹۷	۱/۰۳۹	۵/۰۲۳	۱/۰۰۵	۵/۰۰۵	۱/۰۰۰
	انحراف معیار	۱/۳۵۱	۰/۲۶۶	۰/۵۱۳	۰/۱۰۰	۰/۲۶۵	۰/۰۴۹
$Beta(۱, ۳)$	میانه	۱/۰۰۵	۳/۰۰۴	۰/۹۹۰	۲/۹۷۵	۰/۹۸۹	۲/۹۶۸
	میانگین	۱/۰۳۲	۳/۰۹۳	۱/۰۰۱	۳/۰۰۲	۰/۹۹۴	۲/۹۸۷
	انحراف معیار	۰/۲۲۴	۰/۶۶۷	۰/۱۰۵	۰/۳۰۸	۰/۰۴۷	۰/۱۴۸
$Beta(۲, ۵)$	میانگین	۲/۰۰۷	۵/۰۳۹	۲/۰۰۸	۵/۰۱۹	۱/۹۹۶	۴/۹۹۳
	میانه	۲/۰۹۵	۵/۲۳۸	۲/۰۱۷	۵/۰۴۱	۲/۰۰۱	۵/۰۰۵
	انحراف معیار	۰/۵۲۱	۱/۳۱۷	۰/۱۷۶	۰/۴۴۴	۰/۰۶۷	۰/۱۷۰

[1] Underestimation

جدول ۴.۳: برآورد پارامترهای شکل $\hat{\alpha}$ و $\hat{\beta}$ برای هریک از توزیع‌های بتا با دامنه ۱۰

توزیع بتا		$\hat{\alpha}_{Small}$	$\hat{\beta}_{Small}$	$\hat{\alpha}_{Medium}$	$\hat{\beta}_{Medium}$	$\hat{\alpha}_{Large}$	$\hat{\beta}_{Large}$
$Beta(۰/۵, ۰/۵)$	میانه	۰/۴۸۳	۰/۴۸۳	۰/۴۸۳	۰/۴۸۳	۰/۴۸۵	۰/۴۸۵
	میانه	۰/۴۸۹	۰/۴۹۰	۰/۴۸۸	۰/۴۸۷	۰/۴۸۹	۰/۴۸۹
	انحراف معیار	۰/۰۷۰	۰/۰۷۰	۰/۰۴۱	۰/۰۴۲	۰/۰۲۸	۰/۰۲۸
$Beta(۱, ۱)$	میانه	۰/۹۷۳	۰/۹۷۵	۰/۹۸۱	۰/۹۷۹	۰/۹۸۷	۰/۹۸۹
	میانگین	۰/۹۸۸	۰/۹۸۷	۰/۹۸۵	۰/۹۸۵	۰/۹۹۱	۰/۹۹۱
	انحراف معیار	۰/۱۳۰	۰/۱۲۹	۰/۰۶۳	۰/۰۶۳	۰/۰۳۰	۰/۰۳۰
$Beta(۲, ۲)$	میانه	۱/۹۶۸	۱/۹۶۷	۱/۹۷۷	۱/۹۷۷	۱/۹۹۱	۱/۹۸۷
	میانه	۲/۰۰۳	۲/۰۰۲	۱/۹۸۳	۱/۹۸۴	۱/۹۹۱	۱/۹۹۱
	انحراف معیار	۰/۳۳۵	۰/۳۲۸	۰/۱۲۴	۰/۱۲۶	۰/۰۵۱	۰/۰۵۰
$Beta(۵, ۱)$	میانه	۴/۹۵۹	۰/۹۹۱	۴/۹۵۸	۰/۹۹۰	۴/۹۳۴	۰/۹۸۸
	میانگین	۵/۲۷۷	۱/۰۵۳	۵/۰۲۳	۱/۰۰۵	۴/۹۷۶	۰/۹۹۵
	انحراف معیار	۱/۴۹۶	۰/۲۹۰	۰/۶۴۳	۰/۱۲۶	۰/۳۱۱	۰/۰۵۸
$Beta(۱, ۳)$	میانه	۱/۰۱۰	۳/۰۲۹	۰/۹۸۳	۲/۹۴۸	۰/۹۸۸	۲/۹۶۷
	میانگین	۱/۰۴۱	۳/۱۱۱	۰/۹۹۵	۲/۹۸۶	۰/۹۹۵	۲/۹۸۵
	انحراف معیار	۰/۲۱۰	۰/۶۱۹	۰/۰۹۶	۰/۲۸۲	۰/۰۴۵	۰/۱۴۲
$Beta(۲, ۵)$	میانه	۲/۰۳۰	۵/۰۵۰	۱/۹۹۱	۴/۹۸۸	۱/۹۹۲	۴/۹۸۰
	میانگین	۲/۱۰۱	۵/۲۵۶	۲/۰۰۴	۵/۰۰۸	۱/۹۹۳	۴/۹۸۴
	انحراف معیار	۰/۵۱۸	۱/۳۱۲	۰/۱۹۱	۰/۴۷۱	۰/۰۷۳	۰/۱۸۹

جدول ۵.۳: برآورد پارامترهای شکل $\hat{\alpha}$ و $\hat{\beta}$ برای هریک از توزیع‌های بتا با دامنه ۱۵

توزیع بتا		$\hat{\alpha}_{Small}$	$\hat{\beta}_{Small}$	$\hat{\alpha}_{Medium}$	$\hat{\beta}_{Medium}$	$\hat{\alpha}_{Large}$	$\hat{\beta}_{Large}$
$Beta(۰/۵, ۰/۵)$	میانه	۰/۴۶۹	۰/۴۷۳	۰/۴۷۳	۰/۴۷۲	۰/۴۸۱	۰/۴۸۲
	میانگین	۰/۴۷۸	۰/۴۷۶	۰/۴۷۶	۰/۴۷۸	۰/۴۸۶	۰/۴۸۶
	انحراف معیار	۰/۰۷۵	۰/۰۷۰	۰/۰۴۵	۰/۰۴۶	۰/۰۳۰	۰/۰۳۱
$Beta(۱, ۱)$	میانه	۰/۹۵۶	۰/۹۵۵	۰/۹۷۱	۰/۹۷۲	۰/۹۸۲	۰/۹۸۲
	میانگین	۰/۹۷۱	۰/۹۷۰	۰/۹۷۶	۰/۹۷۹	۰/۹۸۷	۰/۹۸۷
	انحراف معیار	۰/۱۳۵	۰/۱۳۶	۰/۰۶۹	۰/۰۶۸	۰/۰۳۳	۰/۰۳۳
$Beta(۲, ۲)$	میانه	۲/۰۱۱	۱/۹۹۹	۱/۹۹۰	۱/۹۹۲	۱/۹۹۸	۱/۹۹۸
	میانه	۲/۰۳۸	۲/۰۳۶	۲/۰۰۱	۱/۹۹۹	۱/۹۹۹	۱/۹۹۹
	انحراف معیار	۰/۳۱۸	۰/۳۱۶	۰/۱۲۲	۰/۱۲۱	۰/۰۵۰	۰/۰۵۲
$Beta(۵, ۱)$	میانه	۵/۰۷۷	۱/۰۱۱	۴/۹۳۹	۰/۹۹۸	۴/۹۰۴	۰/۹۷۹
	میانگین	۵/۳۲۰	۱/۰۶۵	۵/۰۲۹	۱/۰۰۵	۴/۹۵۰	۰/۹۹۰
	انحراف معیار	۲/۰۶۶	۰/۳۹۰	۰/۷۱۱	۰/۱۴۰	۰/۳۲۹	۰/۰۶۱
$Beta(۱, ۳)$	میانه	۰/۹۸۹	۲/۹۵۸	۰/۹۸۴	۲/۹۵۲	۰/۹۸۲	۲/۹۵۴
	میانه	۱/۰۲۳	۳/۰۵۹	۱/۰۰۰	۲/۹۹۹	۰/۹۹۱	۲/۹۷۴
	انحراف معیار	۰/۲۲۲	۰/۶۴۶	۰/۱۰۷	۰/۳۱۹	۰/۰۴۷	۰/۱۵۲
$Beta(۲, ۵)$	میانه	۲/۰۱۲	۵/۰۳۴	۱/۹۷۲	۴/۹۳۵	۱/۹۸۸	۴/۹۷۲
	میانگین	۲/۱۰۶	۵/۲۶۸	۱/۹۸۷	۴/۹۶۳	۱/۹۹۲	۴/۹۸۱
	انحراف معیار	۰/۵۶۱	۱/۳۹۸	۰/۱۹۲	۰/۴۷۵	۰/۰۷۳	۰/۱۹۱

۵.۱.۳ ویژگی‌های تغییرنگار بتا_دوجمله‌ای

با توجه به این‌که متغیر پاسخ در این پایان‌نامه به‌صورت نسبت است، بنابراین هر یک از پارامترهای تغییرنگار شامل دامنه، ازاره و اثر قطعه‌ای را با توجه به نسبت‌های مشاهده شده $\frac{Z_i}{n_i}$ بررسی می‌کنیم. همان‌طور که مطرح کردیم توزیع دوجمله‌ای شمارشی دارای توزیع کناری بتا_دوجمله‌ای است، یعنی

$$Z_i \sim beta - bin\left(n_i, Y_i, \alpha, \beta\right).$$

پیش از ارزیابی عملکرد مدل کریگینگ بتا_دوجمله‌ای، برخی از ویژگی‌های روابط فضایی نسبت‌های نمونه‌ای مشاهده‌شده $\frac{Z_i}{n_i}$ باید اثبات شوند. با توجه به این‌که هر دو توزیع نرمال و بتا پیوسته هستند، بنابراین هیچ‌گونه تغییر اضافی در فرآیند تبدیل از میدان تصادفی نرمال به میدان تصادفی بتا به مدل اضافه نمی‌شود و هرگونه تغییر اضافی در اثر قطعه‌ای در سطح نسبت‌های دوجمله‌ای رخ داده است. برای اندازه نمونه ثابت در هر مکان فضایی، روابط بین واریانس میدان تصادفی بتای پنهان، $Var\left(Y\right)$، و واریانس میدان تصادفی بتا_دوجمله‌ای، $Var(Z)$، به‌صورت زیر می‌باشد:

$$Var\left(Y\right) = \frac{\alpha\beta}{\left(\alpha+\beta\right)^{۲}\left(\alpha+\beta+۱\right)}$$
$$Var\left(Z\right) = \frac{n\alpha\beta\left(\alpha+\beta+n\right)}{\left(\alpha+\beta\right)^{۲}\left(\alpha+\beta+۱\right)}.$$

بنابراین واریانس نسبت‌های نمونه‌ای بتا_دوجمله‌ای برابر با

$$\begin{aligned} Var\left(\frac{Z}{n}\right) &= \frac{۱}{n^{۲}} Var\left(Z\right) \\ &= \frac{\alpha\beta\left(\alpha+\beta+n\right)}{n\left(\alpha+\beta\right)^{۲}\left(\alpha+\beta+۱\right)} \end{aligned}$$

خواهد بود. از طرفی $Cov\left(Z_i, Z_j|Y\right) = ۰$، بنابراین ازاره فضایی نسبت‌های نمونه‌ای، باید نسبت به ازاره توزیع بتا تفاوت چندانی نداشته باشد، یعنی $\sigma_Y^{۲} = Var\left(Y\right)$ باشد. البته خطای برآورد ناچیزی در آن وجود دارد. تغییرپذیری اضافی از ساختار بتا_دوجمله‌ای می‌تواند به‌عنوان اثر قطعه‌ای در نظر گرفته شود، یعنی $\tau^{۲} = Var\left(\frac{Z}{n}\right) - Var\left(Y\right)$. همچنین دامنه فضایی در سراسر تبدیل نباید تغییر چندانی داشته باشد، چون همان‌طور که عنوان شد تبدیل از میدان تصادفی گاوسی به میدان تصادفی بتا یک به یک است. همچنین از مکان‌های نمونه‌ای مشابه در هر نقطه استفاده شده است. البته ممکن است نوسان ناچیزی در برآورد دامنه فضایی در سطح نمونه‌گیری دوجمله‌ای رخ دهد، اما به طور کلی باید ثابت باشد. بنابراین پارامترهای تغییرنگار برای نسبت‌های نمونه‌ای طبق پارامترهای شکل برابر هستند با

$$\sigma_Z^{۲} = \sigma_Y^{۲} = \frac{\alpha\beta}{\left(\alpha+\beta\right)^{۲}\left(\alpha+\beta+۱\right)}$$

$$\phi_Z = \phi_Y$$

$$\tau_Z^2 = Var\left(\frac{Z}{n}\right) - Var\left(Y\right)$$
$$= \frac{\alpha\beta\left(\alpha+\beta+n\right)}{n\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)} - \frac{n\alpha\beta}{n\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)}$$
$$\frac{\alpha\beta\left(\alpha+\beta\right)}{n\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)}$$
$$= \frac{\alpha\beta}{n\left(\alpha+\beta\right)\left(\alpha+\beta+1\right)}.$$

با توجه به این‌که این فرمول اندازه نمونه نابرابر را نشان نمی‌دهد، با این وجود نکته بسیار مهمی را در مورد اندازه نسبی نسبت‌های نمونه‌ای بتا ـ دوجمله‌ای بیان می‌کند. اثر قطعه‌ای τ_Z^2 توسط عامل $\frac{1}{n}$ تغییر می‌کند، یعنی با افزایش حجم نمونه در هر مکان فضایی، اثر قطعه‌ای به‌سرعت کاهش می‌یابد. بنابراین تصحیح ارایه‌شده توسط کریگینگ بتا ـ دوجمله‌ای زمانی بیش‌تر مورد توجه قرار دارد که تعداد مشاهدات نسبتا کوچکی در هر مکان فضایی وجود داشته باشد. در حالت کلی می‌توان واریانس مجموعه داده‌های فضایی را در دو مولفه اثر قطعه‌ای و ازاره فضایی تجزیه کرد. یعنی

$$Var\left(\frac{Z}{n}\right) = \tau_Z^2 + \sigma_Z^2$$

که مقادیر آن برای توزیع‌های مختلف بتا در جدول ۶.۳ آورده شده‌اند. همچنین جدول ۷.۳ مقادیر ازاره و اثر قطعه‌ای را برای توزیع‌های مختلف به‌ازای n های متفاوت نشان می‌دهد. با توجه به مقادیر جدول، با افزایش n اثر قطعه‌ای به سرعت کاهش می‌یابد و از بین می‌رود. این کاهش در شکل ۴.۳ نیز با افزایش اندازه نمونه برای توزیع‌های مختلف بتا مشهود است. در نمونه‌های نسبتا بزرگ، تغییرنگار تجربی نسبت‌های نمونه‌ای باید عملکرد خوبی داشته باشند. عبارت قطعه‌ای اضافی موجود به‌عنوان کسری از تغییرپذیری کلی، نسبتا کوچک خواهد بود. واضح است که برای نمونه‌های کوچک استفاده از نسبت‌های نمونه‌ای برای برآورد تغییرنگار میدان بتای پنهان مناسب نخواهد بود.

جدول ۶.۳: واریانس واقعی نسبت‌های نمونه‌ای دوجمله‌ای تحت توزیع‌های بتا در شرایط شبیه‌سازی با n=۳

موارد	پارامترهای بتا	τ_Z^2	σ_Z^2	$Var\left(\frac{Z}{n}\right)$
۱	$Beta(۰/۵, ۰/۵)$	۰/۰۴۲	۰/۱۲۵	۰/۱۶۷
۲	$Beta(۱, ۱)$	۰/۰۵۶	۰/۰۸۳	۰/۱۳۹
۳	$Beta(۲, ۲)$	۰/۰۶۷	۰/۰۵۰	۰/۱۱۷
۴	$Beta(۵, ۱)$	۰/۰۴۰	۰/۰۲۰	۰/۰۶۰
۵	$Beta(۱, ۳)$	۰/۰۵۰	۰/۰۳۸	۰/۰۸۸
۶	$Beta(۲, ۵)$	۰/۰۶۰	۰/۰۲۶	۰/۰۸۵

جدول ۷.۳: پارامترهای فضایی واقعی برای نسبت نمونه‌ای $\frac{Z}{n}$ با اندازه نمونه برابر

توزیع بتا	(۰٫۵, ۰٫۵)	(۱,۱)	(۲,۲)	(۵,۱)	(۱,۳)	(۲,۵)
σ_Z^2	۰/۱۲۵	۰/۰۸۳	۰/۰۵۰	۰/۰۲۰	۰/۰۳۸	۰/۰۲۶
$\tau_Z^2\,(n=1)$	۰/۱۲۵	۰/۱۶۷	۰/۰۲۰۰	۰/۱۱۹	۰/۱۵۰	۰/۱۷۹
$\tau_Z^2\,(n=2)$	۰/۰۶۳	۰/۰۸۳	۰/۱۰۰	۰/۰۶۰	۰/۰۷۵	۰/۰۸۹
$\tau_Z^2\,(n=5)$	۰/۰۲۵	۰/۰۳۳	۰/۰۴۰	۰/۰۲۴	۰/۰۳۰	۰/۰۳۶
$\tau_Z^2\,(n=10)$	۰/۰۱۳	۰/۰۱۷	۰/۰۲۰	۰/۰۱۲	۰/۰۱۵	۰/۰۱۸
$\tau_Z^2\,(n=20)$	۰/۰۰۶	۰/۰۰۸	۰/۰۱۰	۰/۰۰۶	۰/۰۰۸	۰/۰۰۹

۲.۳ روش‌های مختلف برآورد تغییرنگار

در متون آمار فضایی (کرسی، ۱۹۸۵)، برای برآورد تغییرنگار سه استراتژی معمول برای تعیین تعداد نقاط همسایگی و تابع زیان به‌صورت زیر پیشنهاد شده‌اند:

۱.

$$N(h) = \sum_{i,j} I_d(i,j) \sim h \qquad L_1(\theta) = \sum_k N(h)\left[\gamma_Y(k) - \gamma_\theta(k)\right]$$

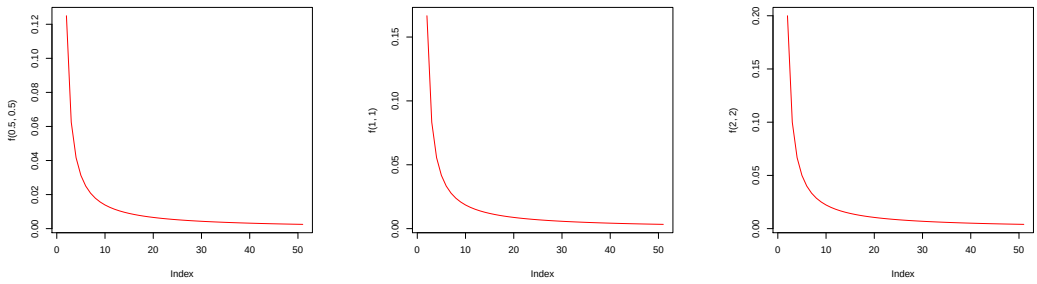
۲.

$$N^*(h) = \sum_{i,j} \frac{n_i n_j}{n_i + n_j} I_d\,(i,j) \sim h \qquad L_2(\theta) = \sum_k N^*(h)\left[\gamma_Y(k) - \gamma_\theta(k)\right]$$

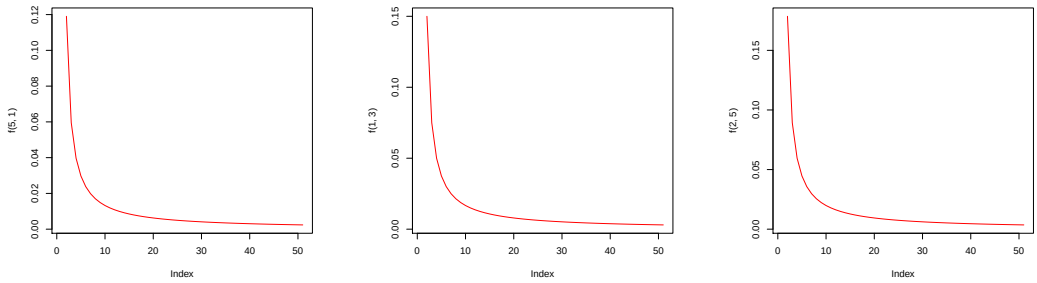
۳.

$$N(h) = \sum_{i,j} I_d(i,j) \sim h \qquad L_3(\theta) = \sum_k \left[\frac{\gamma^*(k) - \gamma_\theta(k)}{\gamma_\theta(k)}\right]^2$$

به‌طوری که در آن $\boldsymbol{\theta}$ بردار پارامترهای تغییرنگار و $\gamma_Y^*\,(k)$ مقدار تغییرنگار بتا_دوجمله‌ای آن‌ها برای فاصله k است. توابع زیان استراتژی اول و دوم یکسان هستند و تنها اختلاف این دو استراتژی در تعداد مشاهدات در هر فاصله است. استراتژی اول فاقد وزن است و از تعداد جفت مشاهدات در هر کلاس فاصله استفاده می‌کند. استراتژی دوم از وزن تعداد مشاهدات یعنی $\frac{n_i n_j}{n_i+n_j}$ برای (i,j) استفاده می‌کند. در حقیقت این عبارت وزن‌دار، اطلاعات مکان‌های مختلف را همگن می‌کند و به جفت مکان‌ها با اندازه نمونه بیش‌تر، وزن بیش‌تری را اختصاص می‌دهد. استراتژی سوم نیز فاقد وزن است.



(آ) $Beta(۲,۲)$ (ب) $Beta(۱,۱)$ (ج) $Beta(۰/۵,۰/۵)$

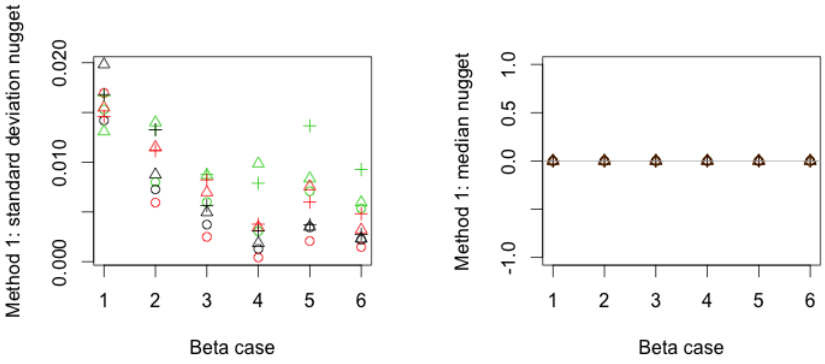


(د) $Beta(۲,۵)$ (ه) $Beta(۱,۳)$ (و) $Beta(۵,۱)$

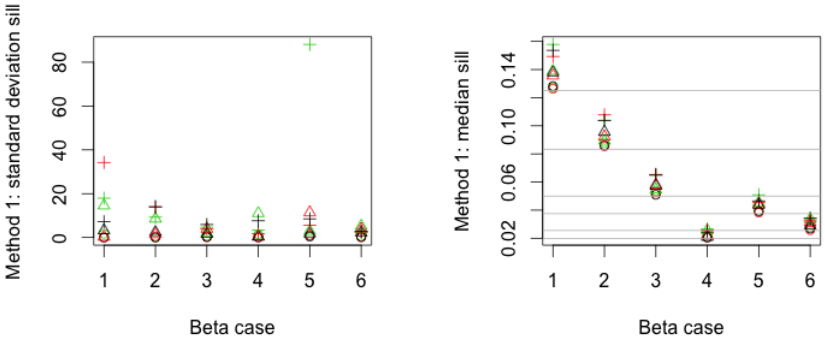
شکل ۴.۳: اثرقطعه‌ای نظری توزیع‌های بتا

برای هر یک از ۵۴ سناریو شامل شش توزیع بتا، سه اندازه نمونه مختلف و سه دامنه فضایی، ۱۰۰ مجموعه داده شبیه‌سازی کردیم. سپس برای تمام مجموعه‌های تولیدشده هر سه استراتژی برآورد پارامتر را اجرا کردیم. شکل‌های۵.۳، ۶.۳ و ۷.۳ برآورد میانه و انحراف معیار را برای ۵۴ سناریو با استرانژی‌های مختلف نشان می‌دهند. در این شکل‌ها دامنه ۵ با ○، ۱۰با △ و ۱۵ با + نشان داده شده‌اند. همچنین اندازه نمونه کوچک با رنگ سبز، اندازه نمونه متوسط با رنگ قرمز و اندازه نمونه بزرگ با رنگ سیاه نمایش داده شده‌اند. با توجه به نتایج شبیه‌سازی‌ها، استراتژی سوم نسبت به دو استراتژی دیگر بدترین عملکرد را نشان می‌دهد. در این روش اثر قطعه‌ای برای توزیع‌های متقارن به‌خوبی برآورد شده است، اما تقریبا برای تمام توزیع‌های نامتقارن بیش‌برآورد شده است. انحراف معیار اثر قطعه‌ای استراتژی سوم نسبت به سایر روش‌ها بزرگ‌تر است. از طرفی در این استراتژی بیش‌ترین اثر قطعه‌ای در اندازه نمونه کوچک وجود دارد، در حالی که دو استراتژی دیگر اثر قطعه‌ای را برای هر سه اندازه نمونه به‌خوبی برآورد می‌کنند. همچنین در این استراتژی ازاره فضایی برای تمام سناریوها بیش‌تر برآورد شده است. ازاره فضایی برای توزیع‌های متقارن با کاهش دامنه کم‌تر برآورد می‌شود. کرسی (۱۹۹۳) استفاده از استراتژی سوم را برای توزیع‌های گاوسی پیشنهاد داد؛ بنابراین عملکرد نامناسب این روش برای داده‌های دوجمله‌ای شمارشی کاملا طبیعی است.

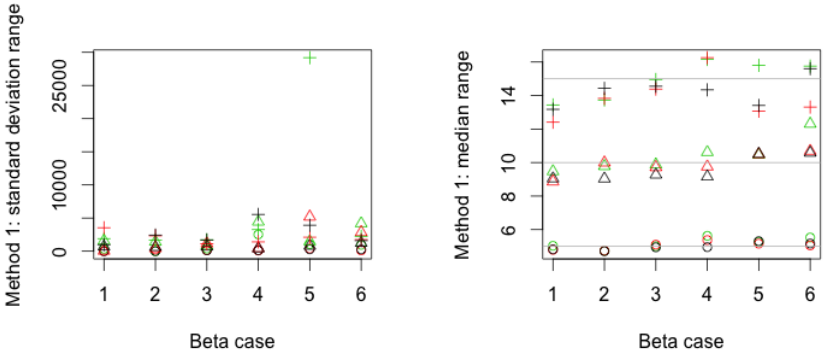
حال به مقایسه دقت برآوردهای استراتژی اول و دوم می‌پردازیم. هر دو استراتژی اثر قطعه‌ای را برای تمام سناریوها به‌خوبی برآورد می‌کنند. به‌طور متوسط می‌توان گفت انحراف معیار اثر قطعه‌ای برای روش دوم تا اندازه‌ای کوچک‌تر است. در روش دوم دو سناریو با انحراف معیار بالا وجود دارند که ممکن است به علت وجود چولگی در داده‌ها رخ داده باشند. برآورد میانه ازاره، تقریبا در هر دو روش یکسان است. استراتژی دوم دامنه فضایی را کمی بهتر برآورد می‌کند. روش اول در دامنه فضایی ۱۰ توزیع‌های ۶-۵ را کمی بیش‌تر و همچنین ۳-۱ را کم‌تر برآورد می‌کند. در استراتژی اول مشکل بیش‌برآوردی برای توزیع‌های متقارن وجود دارد، اما برای توزیع‌های چوله دقیق است. همچنین این استراتژی دامنه فضایی ۱۵ را برای توزیع‌های متقارن کم‌برآورد می‌کند و برای توزیع‌های چوله کاملا متغیر است. روش دوم نیز توزیع بتا ۱ و ۲ را بیش‌برآورد می‌کند، اما برای سایر سناریوهای شبیه‌سازی به‌ویژه در اندازه نمونه بزرگ دقیق است. بر اساس شبیه‌سازی‌ها، استراتژی دوم برآوردهای دقیق‌تری از پارامترهای کوواریانس را ارایه می‌دهد. تعدیل برای تعداد مشاهدات در هر فاصله h یعنی $N^*(h)$ به‌صورت تصادفی انتخاب نشده است. در حقیقت این همان تعدیل ارایه‌شده در کریگینگ بتا_دوجمله‌ای است. این تعدیل نشان می‌دهد که همگن کردن اندازه نمونه n_i در مکان‌های مختلف برای برآورد پارامتر نیز مهم است. هنگام برخورد با نسبت‌ها، دقت نسبت‌های دوجمله‌ای به ویژه در اندازه نمونه کوچک به اندازه نمونه وابسته است. این واقعیت نه تنها در مرحله برآورد تغییرنگار بلکه در مرحله برآورد پارامترهای کوواریانس مدل نیز باید مورد توجه قرار گیرد.



(آ) میانه اثر قطعه‌ای (ب) انحراف معیار اثر قطعه‌ای

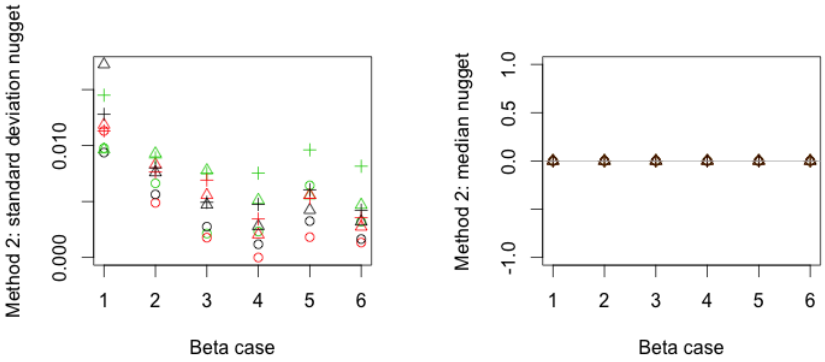


(ج) میانه ازاره (د) انحراف معیار ازاره

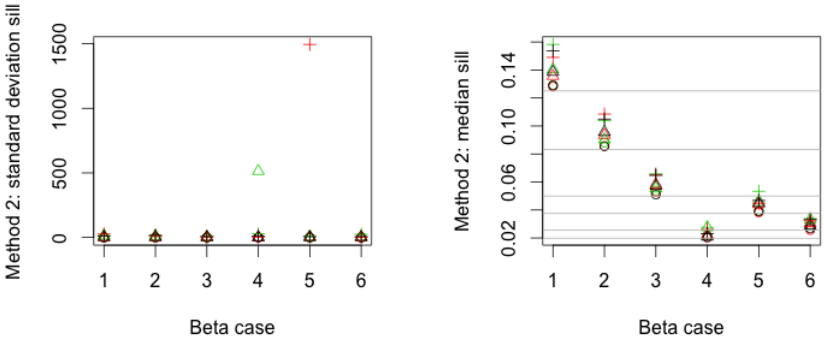


(ه) میانه دامنه (و) انحراف معیار دامنه

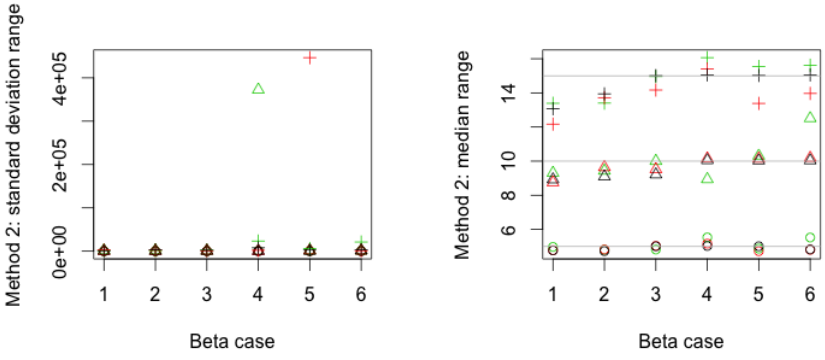
شکل ۵.۳: برآورد میانه و انحراف معیار پارامترهای تغییرنگار برای ۵۴ سناریو شبیه‌سازی بر اساس ۱۰۰ مجموعه داده شبیه‌سازی با استفاده از روش اول



(آ) میانه اثر قطعه‌ای (ب) انحراف معیار اثر قطعه‌ای

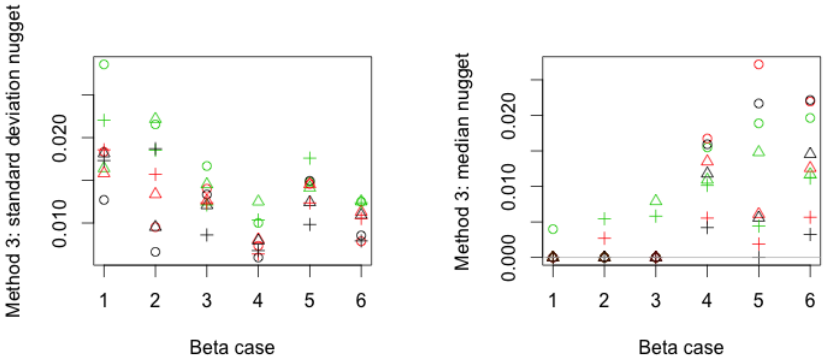


(ج) میانه ازاره (د) انحراف معیار ازاره

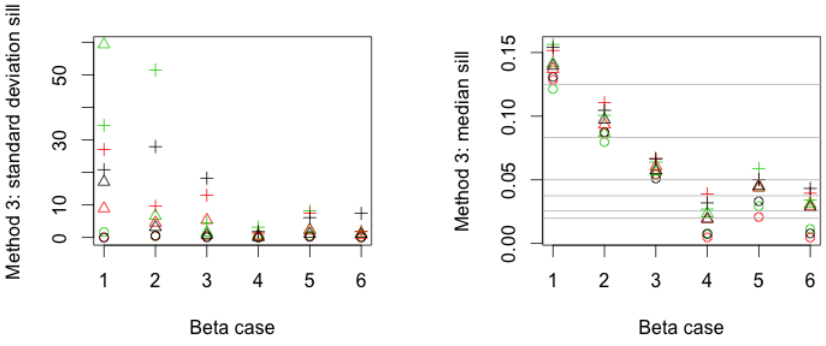


(ه) میانه دامنه (و) انحراف معیار دامنه

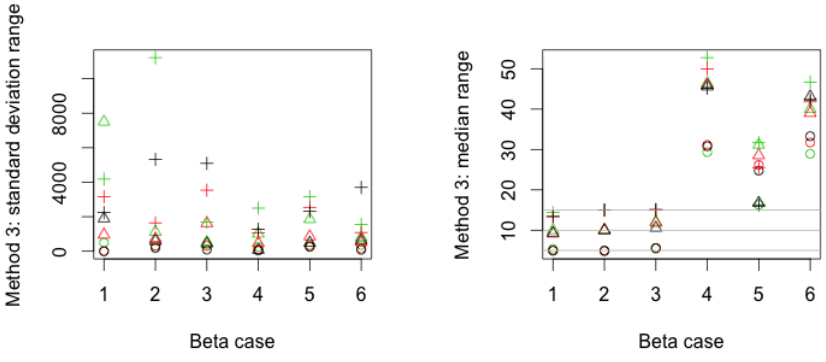
شکل ۶.۳: برآورد میانه و انحراف معیار پارامترهای تغییرنگار برای ۵۴ سناریو شبیه‌سازی بر اساس ۱۰۰ مجموعه داده شبیه‌سازی با استفاده از روش دوم



(آ) میانه اثر قطعه‌ای (ب) انحراف معیار اثر قطعه‌ای



(ج) میانه ازاره (د) انحراف معیار ازاره



(ه) میانه دامنه (و) انحراف معیار دامنه

شکل ۷.۳: برآورد میانه و انحراف معیار پارامترهای تغییرنگار برای ۵۴ سناریو شبیه‌سازی بر اساس ۱۰۰ مجموعه داده شبیه‌سازی با استفاده از روش سوم

۳.۳ برآورد پارامترهای ساختار وابستگی فضایی

برای ارزیابی عملکرد برآورد پارامترها، برای هر سناریو ۱۰۰ مجموعه داده شبیه‌سازی کردیم. همان‌طور که در بخش قبلی بیان شد، با توجه به نتایج حاصل از شبیه‌سازی، برای برآورد پارامترها از تابع زیان $L_۲(\cdot)$ و تعریف تعداد نقاط همسایه $N^*(h)$، در استراتژی دوم، استفاده کردیم. برآوردهای پارامترهای تغییرنگار برای مدل بتا_دوجمله‌ای از می‌نیمم کردن تابع زیان $L_۲(\theta)$، که در تغییرنگار تجربی و تغییرنگار واقعی (کروی) مقداردهی شده است، به‌دست آمده‌اند. نمودارهای جعبه‌ای برآوردهای پارامترهای اثر قطعه‌ای، ازاره و دامنه برای مدل بتا_دوجمله‌ای در شکل‌های ۸.۳، ۹.۳ و ۱۰.۳ نشان داده شده‌اند. در این شکل، ۱۸ سناریوی اول مربوط به دامنه ۵، ۱۸ سناریو دوم مربوط به دامنه ۱۰ و ۱۸ سناریو سوم مربوط به دامنه ۱۵ هستند. همچنین ۶ سناریو اول هر دامنه فضایی، اندازه نمونه کوچک، ۶ سناریو دوم اندازه نمونه متوسط و ۶ سناریو سوم اندازه نمونه بزرگ را نشان می‌دهند. پارامترهای واقعی دامنه فضایی به ترتیب برای این سناریوها ۵، ۱۰، و ۱۵ می‌باشند. اثر قطعه‌ای واقعی صفر است که تقریبا در تمام شبیه‌سازی‌ها به درستی برآورد شده است. پارامترها را برای هر یک از ۱۰۰ مجموعه داده شبیه‌سازی‌شده با استفاده از استراتژِی دوم که در بخش قبل مطرح شد برآورد کردیم. در این بخش مقادیر برآوردشده را برای سه پارامتر فضایی اثر قطعه‌ای، دامنه و ازاره مورد بررسی قرار می‌دهیم.

پارامتر اثر قطعه‌ای یعنی $\tau^۲$ تغییرات ایجادشده توسط نمونه‌های دوجمله‌ای در هر مکان را نشان می‌دهد. در تمام شبیه‌سازی‌ها تغییرپذیری اثر قطعه‌ای برای میدان تصادفی بتا پنهان صفر است، بنابراین هرگونه تغییرپذیری اثر قطعه‌ای در نسبت‌های نمونه‌ای فقط از نمونه‌های دوجمله‌ای ناشی شده است و احتمال زیربنایی در آن تاثیری ندارد. با توجه به این‌که پارامترها را برای ساختار فضایی فرآیند بتای پنهان برآورد می‌کنیم، بنابراین در تمام شبیه‌سازی‌ها میانه اثر قطعه‌ای صفر خواهد بود، که نشان می‌دهد اصلاح برآوردگر تغییرنگار در از بین بردن تنوع اضافی موفق بوده است. با این وجود در برخی از توزیع‌های بتا هنوز هم کمی تغییرپذیری موجود است. به‌عنوان مثال، توزیع بتا نوع یک که نشان‌دهنده توزیع بتا دومدی است بیش‌ترین اثر قطعه‌ای برآوردشده را دارد. به‌طور کلی این توزیع بیش‌ترین تنوع را در داده‌های نمونه‌ای نشان داده است. بنابراین باید انتظار داشت که تنوع اضافی همواره به‌طور کامل نمی‌تواند تعدیل شود. توزیع‌های بتا نامتقارن (۴,۵,۶) تمایل به برآورد نقاط پرت کم‌تری دارند. انحراف معیار اثر قطعه‌ای برآوردشده نیز برای توزیع‌های متقارن تمایل دارند افزایش پیدا کنند. از آن‌جایی که اثر قطعه‌ای در صفر کراندار است، افزایش انحراف معیار تنها به‌دلیل اثر قطعه‌ای مثبت بزرگ است.

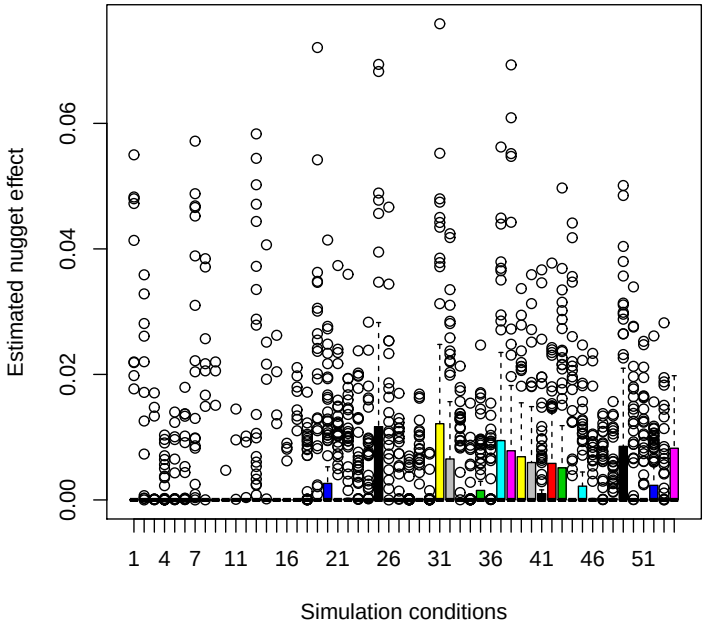
همچنین به‌نظر می‌رسد اندازه اثر قطعه‌ای برآوردشده به دامنه فضایی وابسته است. ۱۸ سناریو اول که دارای کم‌ترین دامنه فضایی، یعنی $\phi = ۵$، هستند نقاط پرت نسبتا کم‌تری

را دارند. سناریوهای ۵۴-۱۹ ازاره فضایی بزرگ‌تر و نقاط پرت به‌مراتب بیش‌تری را نشان می‌دهند. از طرفی با افزایش دامنه فضایی، مقادیر نقاط پرت تمایل دارند افزایش پیدا کنند. می‌توان گفت اندازه نمونه در برآورد اثر قطعه‌ای تاثیری ندارد. در واقع با افزایش اندازه نمونه الگوی قابل تشخیصی در شکل معلوم نیست، که این نشان می‌دهد اصلاح تغییرنگار بتا‌ـ دوجمله‌ای، برآورد پارامترهای فرآیند بتای پنهان را بهبود بخشیده است. با توجه به این‌که در اندازه نمونه کوچک اطلاعات کم‌تری درباره نسبت بتا درست وجود دارد، بنابراین با استفاده از تغییرنگار اصلاح‌نشده در این اندازه نمونه باید نتایج تغییرپذیری بیش‌تری داشته باشند.

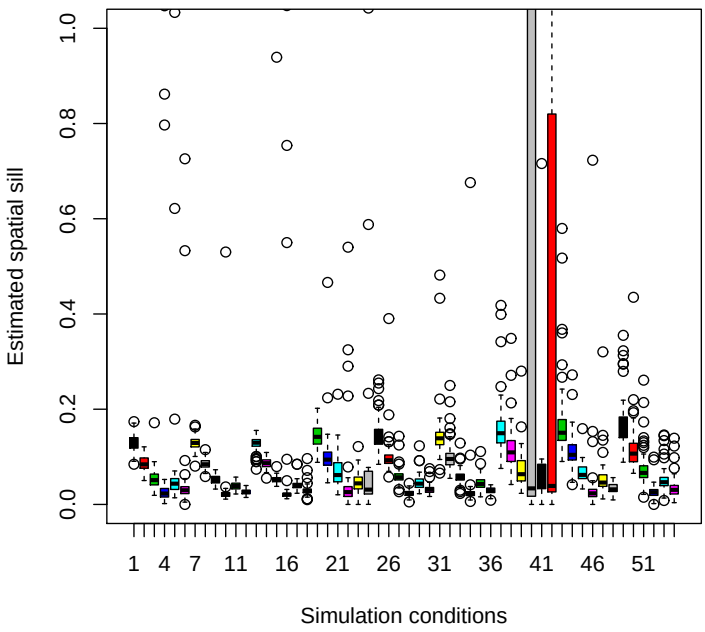
ازاره فضایی مورد انتظار توزیع بتای پنهان، تغییرپذیرتر از توزیع بتای زیربنایی است. نمودار جعبه‌ای برآورد پارامتر ازاره، چوله به راست است و نقاط پرت کم‌تری نسبت به اثر قطعه‌ای دارد. از طرفی، برآوردهای ازاره بزرگ‌تر از یک نیز وجود دارند که با توجه به کراندار بودن توزیع بتا در فاصله صفر و یک طبیعی نیست و نشان می‌دهد که همگرایی برای بعضی از توزیع‌ها ممکن است مشکل‌ساز شود. همچنین با افزایش دامنه فضایی نقاط پرت نیز افزایش می‌یابند.

اندازه نمونه‌های دوجمله‌ای روی میانه ازاره فضایی برای توزیع‌های متقارن تاثیر ندارند. با افزایش دامنه فضایی، نمودار میانه ازاره، ازاره فضایی واقعی را بیش‌تر برآورد می‌کند، به‌طوری که در اندازه نمونه بزرگ این بیش‌برآوردی کاهش می‌یابد اما هنوز هم قابل مشاهده است. تغییرپذیری اضافی می‌تواند به عوامل مختلفی وابسته باشد. یک عامل ممکن می‌تواند برآورد پارامترهای توزیع بتا با استفاده از نسبت‌های نمونه‌ای باشد. با استفاده از $\hat{\alpha}$ و $\hat{\beta}$ برآوردشده برای برآورد تغییرنگار بتای پنهان، سطح دیگری از پیچیدگی اصلاح‌نشده در تغییرنگار اضافه می‌شود که می‌تواند منجر به متورم شدن اثر واریانس شود. توزیع‌های بتا متقارن دامنه گسترده‌تری از مقادیر کلی ممکن دارند و از توزیع‌های بتا چوله متغیرتر هستند.

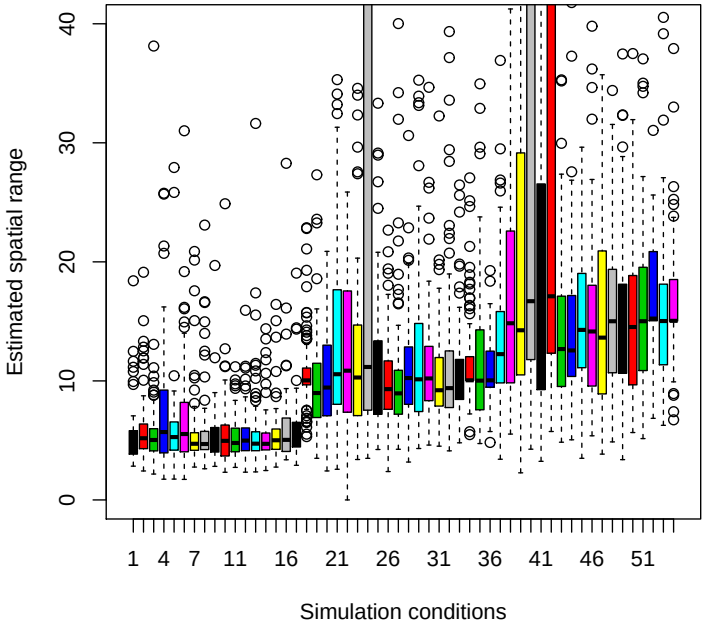
در نهایت دامنه فضایی برآوردشده را در نظر بگیرید. ۱۸ سناریو اول مربوط به کوچک‌ترین دامنه فضایی، یعنی $\phi = ۵$ است. به‌طور کلی برآورد دامنه فضایی در شبیه‌سازی‌ها به‌خوبی بازیابی شده است. اگرچه هنوز هم چولگی کمی در برآورد پارامترها وجود دارد که با افزایش اندازه نمونه کاهش می‌یابد. بر اساس میدان تصادفی فضایی با ابعاد ۲۰×۲۰، دامنه فضایی ۵ با توجه به این‌که فاصله نسبتا کمی با آن دارد تشخیص آن باید آسان باشد. مدل کریگینگ بتا‌ـ دوجمله‌ای، برآورد دامنه فضایی در این اندازه را به‌خوبی انجام می‌دهد. همچنین تغییرپذیری توزیع‌های چوله نیز بیش‌تر از توزیع‌های بتا متقارن است. الگوهای افزایش برآورد دامنه برای توزیع‌های چوله با دامنه $\phi = ۱۰$ و $\phi = ۱۵$ مشابه دامنه $\phi = ۵$ رخ می‌دهد. به‌طورکلی می‌توان گفت، مدل کریگینگ بتا‌ـ دوجمله‌ای برآورد پارامترها را به درستی انجام می‌دهد.



شکل ۸.۳: اثر قطعه‌ای فضایی برآورد شده



شکل ۹.۳: ازاره فضایی برآورد شده



شکل ۳.۱۰: دامنه فضایی برآورد شده

برآورد میدان تصادفی بتای پنهان، دارای معایبی است که باید مرتفع شوند. این معایب شامل موارد زیر هستند:

۱. اطلاعات محدود هستند. در اغلب موارد به‌ویژه در اندازه نمونه کوچک نسبت‌های نمونه‌ای دوجمله‌ای می‌توانند بیش‌تر یا کم‌تر از احتمال زیربنایی باشند.

۲. پارامترها پیچیده‌تر هستند. نه تنها پارامترهای کوواریانس فضایی باید برآورد شوند، بلکه ویژگی‌های میدان تصادفی بتای زیربنایی نیز باید به‌خوبی برآورد شوند. در تحلیل زمین‌آماری، فرض نرمال بودن میدان تصادفی فضایی و ثابت بودن میانگین برای برآورد پارامترهای ساختار وابستگی فضایی ضروری است.

۳. اندازه نمونه ثابت نیست. در هر نقطه سطح متفاوتی از تغییرپذیری در نسبت‌های نمونه‌ای دوجمله‌ای وجود دارد، که این تنوع باید برای عملکرد میدان تصادفی بتای پنهان برآورد و سپس حذف شود.

مدل کریگینگ بتا_دوجمله‌ای به‌عنوان رهیافتی امیدبخش برای مدل‌سازی مسایل پیچیده ظاهر شده است. این مدل برآورد دقیق و پایداری را از ساختار فضایی میدان احتمالی پنهان فراهم می‌کند.

۴.۳ دقت پیش‌گویی

برای بررسی عملکرد مدل بتا_دوجمله‌ای در پیش‌گویی، ۱۰۰ مجموعه داده را (به ازای هر سناریوی شبیه‌سازی) بر روی یک مشبکه ۵۰ × ۵۰ شامل ۲۵۰۰ گره تولید کردیم و با استفاده از روش کریگینگ بتا_دوجمله‌ای (با پارامترهای برآوردشده تغییرنگار) مقادیر نسبت‌های دوجمله‌ای را، برای هر مجموعه داده، پیش‌گویی کردیم. در واقع در این‌جا می‌خواهیم بررسی کنیم که آیا مدل، مقادیر مشاهده‌نشده فرآیند بتای پنهان را به‌خوبی پیش‌گویی کرده است یا خیر؟ برای این منظور از معیارهای مختلفی استفاده می‌کنیم، که در ادامه به تحلیل نتایج این معیارها می‌پردازیم.

نمودارهای شکل‌های ۱۱.۳، ۱۲.۳، ۱۳.۳، ۱۴.۳، ۱۵.۳ و ۱۶.۳ معیارهای مختلف دقت پیش‌گویی را نمایش می‌دهند. در شکل ۱۱.۳ نمودار جعبه‌ای میانگین خطای پیش‌گویی‌شده را برای هر یک از ۱۰۰ مجموعه داده شبیه‌سازی‌شده تحت سناریوهای مختلف نشان می‌دهد. بر اساس همه سناریوهای شبیه‌سازی، میانگین خطای پیش‌گویی بسیار نزدیک به صفر است و روند کمی در آن مشاهده می‌شود. به‌طور کلی با افزایش اندازه نمونه، تغییرات میانگین خطای پیش‌گویی کوچک و تقریبا صفر است. همچنین در دامنه فضایی بزرگ‌تر نیز خطای پیش‌گویی نزدیک به صفر است. از طرفی تعدادی نقاط پرت دیده می‌شوند که نشان می‌دهد حتی اگر برآورد پارامترها برای یک مجموعه داده دشوار باشد مقادیر پیش‌گویی منطقی و دقیق هستند. توزیع بتا ۱ و ۲ که نشان‌دهنده توزیع بتا دومدی و توزیع یکنواخت هستند، به‌ترتیب بیش‌ترین تغییرات میانگین خطای پیش‌گویی را نشان می‌دهند. این توزیع‌های خاص بیش‌تر تمایل دارند در نزدیکی مقادیر صفر و یک که پیش‌بینی با استفاده از مدل کریگینگ بتا_دوجمله‌ای دشوار است، رخ دهند.

میانگین توان دوم خطای پیش‌گویی یک شاخص خوب از دو ویژگی مجموعه‌ای از پیش‌گویی‌ها، یعنی دقت و تنوع است. مجموعه داده‌های پیش‌گویی دقیق باید میانگین توان دوم خطای پیش‌گویی نزدیک به صفر داشته باشند. به‌طوری که افزایش میانگین توان دوم خطای پیش‌گویی در دو حالت رخ می‌دهد: این افزایش می‌تواند به دلیل خطای نسبتا بالا یا وجود نقاط پرت در خطاها باشد. توزیع‌های بتا متقارن در تمام شبیه‌سازی‌ها بیش‌ترین میانگین توان دوم خطای پیش‌گویی را دارند، که علت آن متغیرتر بودن این توزیع‌ها است. همچنین این توزیع‌ها تمایل دارند نسبت‌های شدیدتر را نمایش دهند. توزیع‌های چوله نیز در تمامی شبیه‌سازی‌ها میانگین توان دوم خطای پیش‌گویی کوچک و تقریبا کم‌تر از ۴٪ را دارند. علاوه بر این افزایش اندازه نمونه سبب کاهش میانگین توان دوم خطای پیش‌گویی می‌شود. بنابراین می‌توان گفت، با گرفتن نمونه‌های دوجمله‌ای، پیش‌گویی‌ها تمایل دارند دقیق‌تر شوند که این کاهش بیش‌تر از نمونه‌های کوچک به متوسط مشهود است. این نشان می‌دهد که با مدل کریگینگ بتا_دوجمله‌ای، ده تا بیست مشاهده در هر مکان فضایی ممکن است برای دریافت نتیجه معقول با اطلاعات خوب کافی باشد. با افزایش دامنه فضایی، میانگین توان

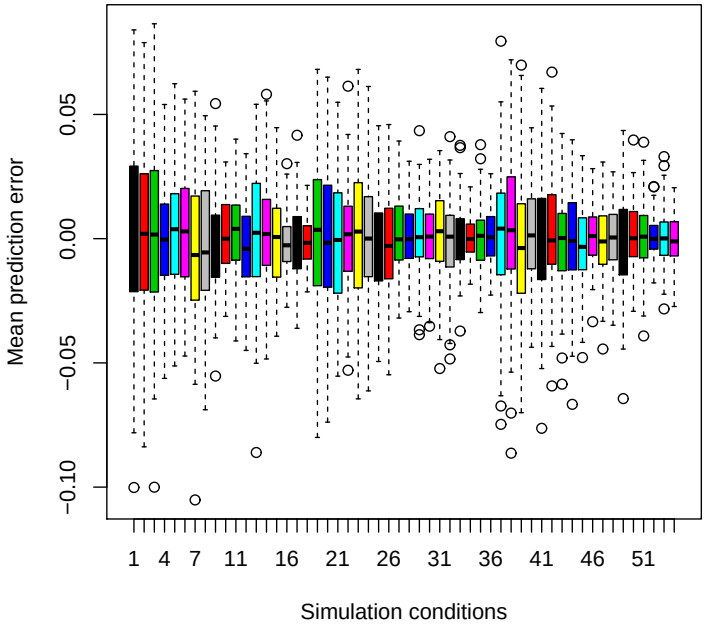
دوم خطای پیش‌گویی کاهش می‌یابد. هرچه وابستگی فضایی قوی‌تر باشد، تغییرات از مکانی به مکان دیگر کم‌تر، و پیش‌گویی نیز می‌تواند بهتر باشد.

معیار دیگر دقت پیش‌گویی، میانگین مقادیر پیش‌گویی است. آیا میانگین مقادیر پیش‌گویی به میانگین واقعی توزیع بتا نزدیک هستند؟ با استفاده از مدل کریگینگ بتا–دوجمله‌ای پاسخ مثبت است، به‌طوری که میانگین مقادیر پیش‌گویی برای شبیه‌سازی‌های بتا به میانگین توزیع بتا بسیار نزدیک هستند. در اندازه نمونه بزرگ میانگین مقادیر پیش‌گویی بسیار سازگار هستند. همان‌طور که انتظار داشتیم توزیع‌هایی با دامنه فضایی بزرگ، مقادیر میانگین پیش‌گویی بسیار پایداری دارند.

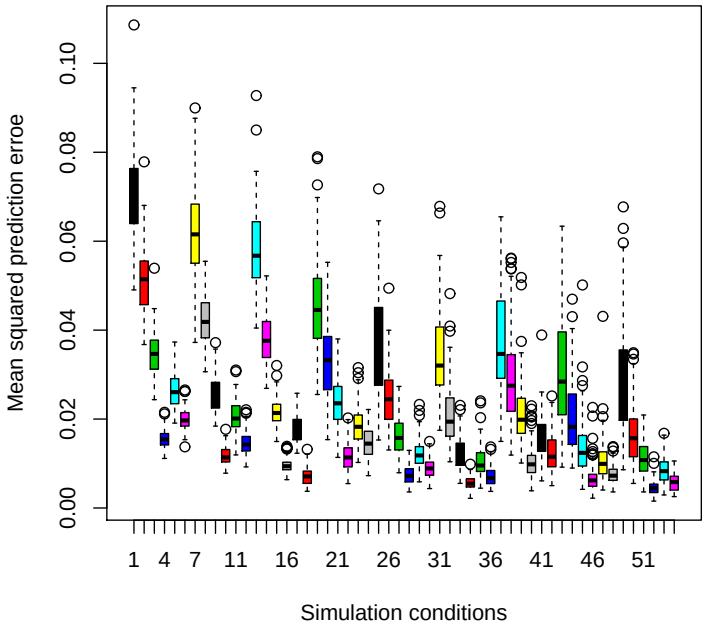
با توجه به شکل ۱۶.۳ مجموعه داده‌ها با دامنه فضایی کوچک مقادیر پیش‌گویی با واریانس کوچک‌تری دارند. به‌طور کلی نمای پیش‌گویی‌ها مسطح است، که این مسطح بودن به خطاهای بالاتر مشاهده‌شده در این شبیه‌سازی برمی‌گردد. سطح پیش‌گویی‌هایی که صاف هستند، گوشه‌ها و شکاف داده‌ها را نمی‌توانند به‌خوبی برآورد کنند. در دامنه فضایی بزرگ‌تر سطوح پیش‌گویی شیب‌دار و واریانس بزرگ‌تر می‌شود. همچنین واریانس پیش‌گویی برای توزیع‌های متقارن بزرگ‌تر است، همان‌طوری که توزیع بتا زیربنایی متغیرتر است. بنابراین پیش‌گویی‌ها نیز باید متغیرتر باشند. با افزایش اندازه نمونه واریانس پیش‌گویی نیز افزایش می‌یابد.

با معیارهای پیش‌گویی که تاکنون در نظر گرفتیم، کاهش برای اندازه نمونه بزرگ به‌خوبی صورت گرفته است. با این وجود، برای واریانس پیش‌گویی در اندازه نمونه بزرگ‌تر باید امیدوار به پیش‌گویی‌های متغیرتر بود. نمونه‌های دوجمله‌ای با n_i بزرگ تمایل دارند به نسبت‌های نمونه‌ای مشاهده‌شده با احتمال درست π_i نزدیک‌تر شوند. همان‌طور که اندازه نمونه افزایش می‌یابد، تعداد قابل توجهی از نمونه‌های مشاهده‌شده نیز دیده می‌شوند. نه‌تنها نسبت‌های نمونه‌ای مورد استفاده برای پیش‌گویی در اندازه نمونه بزرگ دقیق‌تر هستند بلکه آن‌ها انتخاب‌های بیش‌تری دارند. در حقیقت اطلاعات بیش‌تری قابل دستیابی است. اطلاعات کامل‌تر توسط افزایش n_i گرفته‌شده، اجازه می‌دهد که برخی از قله‌ها و دره‌ها بیش‌تر قابل توجه شوند. از این رو در اندازه نمونه بزرگ‌تر پیش‌گویی‌ها متغیرتر خواهند بود.

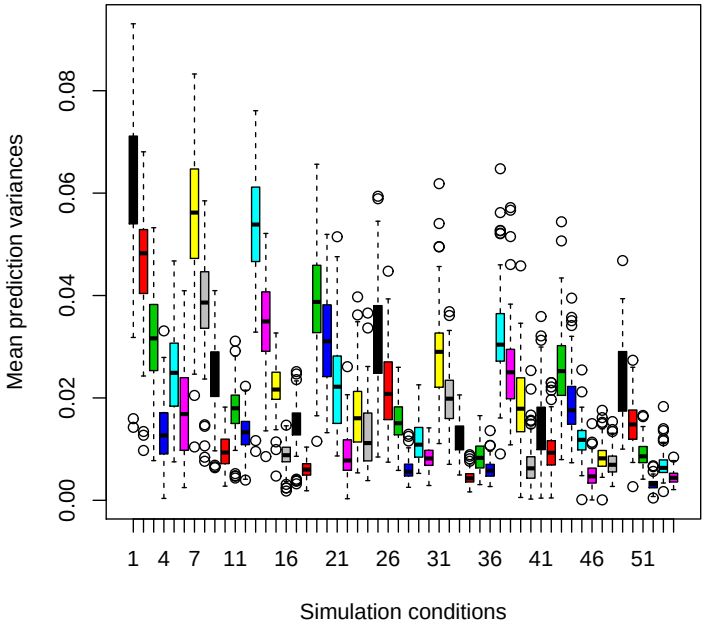
ضریب همبستگی اسپیرمن بین میدان تصادفی بتای پنهان واقعی و مقادیر پیش‌گویی را نیز برای هر شبیه‌سازی محاسبه کردیم. با افزایش اندازه نمونه، همبستگی بین میدان تصادفی بتا و متغیرهای تصادفی پیش‌گویی بیش‌تر می‌شود. ارتباط بالا به‌معنی برآورد بهتر بالا و پایین توزیع است. همچنین با افزایش دامنه فضایی همبستگی بین میدان توزیع بتای پنهان و توزیع بتا پیش‌گویی بیش‌تر می‌شود. این نتیجه نشان می‌دهد مدل کریگینگ بتا–دوجمله‌ای می‌تواند نقاط مرتفع و پست توزیع احتمال زیربنایی را به‌خوبی تشخیص دهد به جز در مواردی که اندازه نمونه کوچک و همبستگی فضایی نسبتا کم باشد.



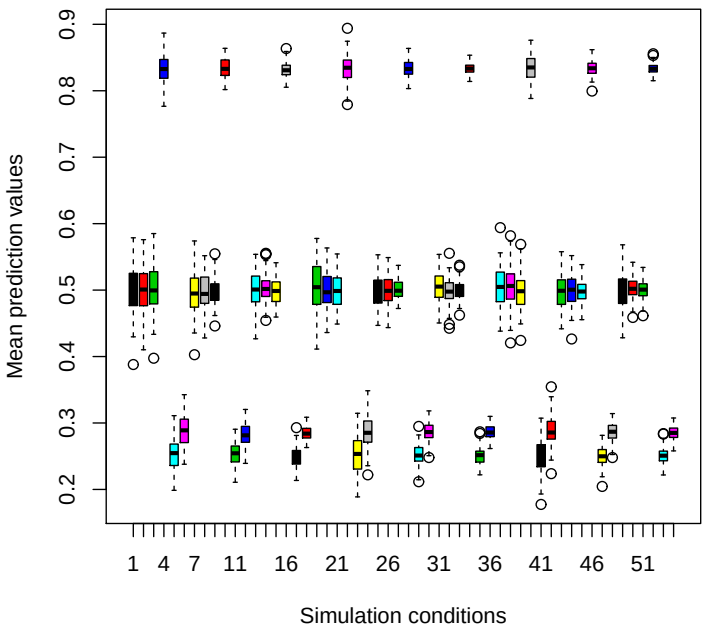
شکل ۱۱.۳: میانگین خطای پیش‌گویی



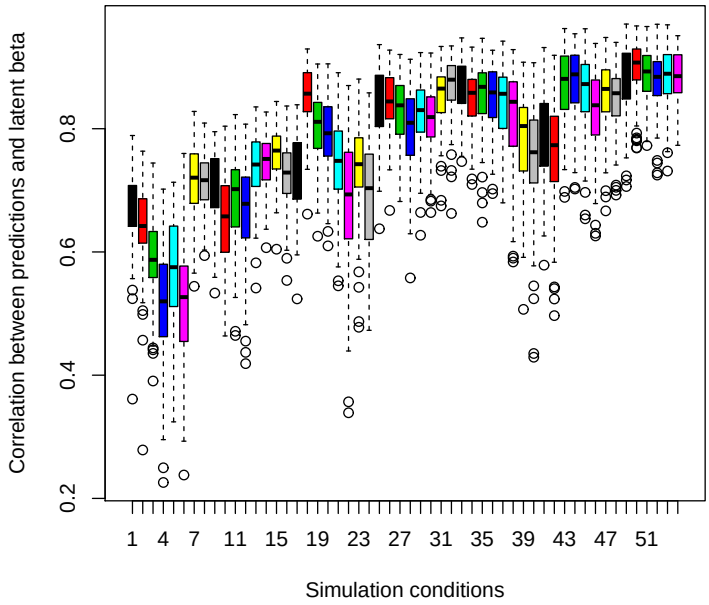
شکل ۱۲.۳: میانگین توان دوم خطای پیش‌گویی



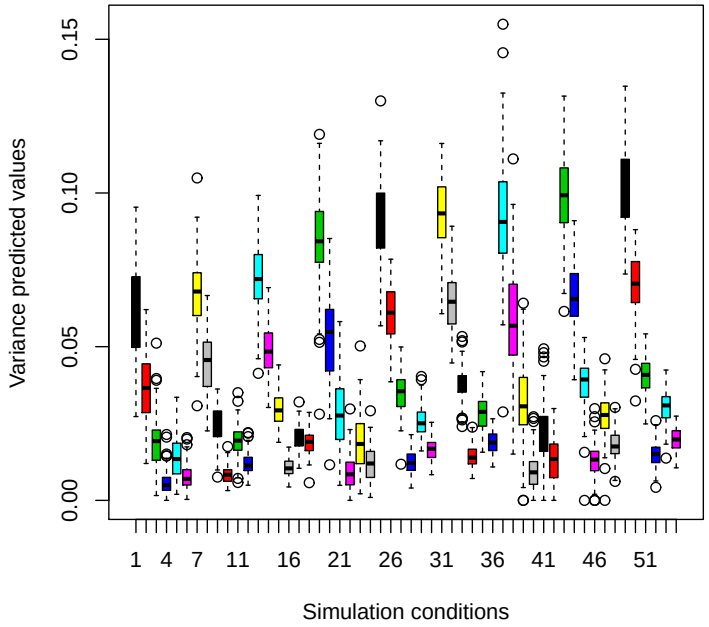
شکل ۱۳.۳: میانگین واریانس پیش‌گویی



شکل ۱۴.۳: میانگین مقادیر پیش‌گویی



شکل ۱۵.۳: همبستگی بین توزیع بتا پیش‌گویی و توزیع بتا پنهان



شکل ۱۶.۳: واریانس مقادیر پیش‌گویی

فصل ۴

کارایی پیش‌گویی مدل بتا_دوجمله‌ای فضایی و مقایسه با روش هم‌کریگینگ دوجمله‌ای

در این فصل یک مدل جایگزین برای کریگینگ بتا_دوجمله‌ای، یعنی مدل هم‌کریگینگ دوجمله‌ای را بر اساس نسبت‌های نمونه‌ای مطرح می‌کنیم. سپس به مقایسه توان هر دو مدل در برآورد پارامترها و پیش‌گویی می‌پردازیم. در انتها کاربست مدل بتا_دوجمله‌ای را برای پهنه‌بندی نرخ بارش در استان سمنان نمایش می‌دهیم.

۱.۴ هم‌کریگینگ دوجمله‌ای

لاجونی (۱۹۹۱) یک روش مشابه برای برآورد احتمال زیربنایی، بر اساس نسبت نمونه‌ای مشاهده‌شده به نام هم‌کریگینگ دوجمله‌ای[۱] معرفی نمود. در این مدل، نسبت‌های نمونه‌ای به‌عنوان متغیرهای کمکی برای میدان تصادفی پنهان مدل‌بندی می‌شوند. دلیل نام‌گذاری هم‌کریگینگ نیز همین واقعیت است. به عبار دیگر با وجود نام هم‌کریگینگ، درحقیقت

[۱] Binomial cokriging

مدل به شکل هم‌کریگینگ نیست: زیرا هیچ متغیر کمکی برای کمک به پیش‌گویی فضایی استفاده نمی‌شود. در واقع بنا به گفته الیور ''از آن‌جایی که ما یک احتمال نامعلوم را از روی نسبت‌های نمونه‌ای مشاهده‌شده برآورد می‌کنیم، بنابراین می‌توان آن را به عنوان یک مدل هم‌کریگینگ در نظر گرفت''. اگرچه این روش به‌صورت گسترده اجرا نشد، اما در سال ۱۹۹۶ در مقاله‌ای توسط الیور و همکاران دوباره ظاهر شد. این مدل توسط لاجونی گسترش داده شد. الیور و همکاران نیز این مدل را برای پیش‌گویی خطر ابتلا به سرطان کودکان در بخش‌هایی از غرب انگلستان برای داده‌های سال‌های ۱۹۸۴-۱۹۸۰ تشریح کردند. برآوردگر تغییرنگار هم‌کریگینگ دوجمله‌ای با استفاده از نمادگذاری تغییرنگار بتا_دوجمله‌ای به صورت زیر است:

$$\hat{\gamma}_Y(h) = \hat{\gamma}_Z(h) - \frac{1}{2}\left(\hat{\pi}(1-\hat{\pi}) - \hat{\sigma}_Y^2\right)\frac{n_i + n_j}{n_i n_j}.$$

شکل کلی برآوردگر تغییرنگار مشابه مدل کریگینگ بتا_دوجمله‌ای است. با این وجود، برخی تفاوت‌های کلیدی در هر دو تغییرنگار و همچنین پذیره‌های دو مدل وجود دارند. به‌طور خاص این تفاوت‌ها عبارتند از

۱. در مدل کریگینگ بتا_دوجمله‌ای به صراحت فرض می‌شود میدان تصادفی پنهان، همبسته فضایی است و از توزیع بتا پیروی می‌کند. اما در مدل هم‌کریگینگ دوجمله‌ای هیچ پذیره توزیعی بر روی مدل زیربنایی وجود ندارد.

۲. در مدل کریگینگ بتا_دوجمله‌ای یک شرط اضافی بر روی پارامترهای برآوردشده میانگین و واریانس، یعنی $\hat{\pi}$ و $\hat{\sigma}_Y^2$، گذاشته می‌شود تا میدان تصادفی پنهان را بین $[0,1]$ محدود کند. در مقابل هم‌کریگینگ دوجمله‌ای چنین شرطی را لازم ندارد و به‌جای آن از میانگین و واریانس نسبت‌های نمونه‌ای برای برآورد $\hat{\pi}$ و $\hat{\sigma}_Y^2$ استفاده می‌شود.

۳. عبارت تصحیح اریبی در هم‌کریگینگ دوجمله‌ای مقداری ثابت است.

۴. در مدل کریگینگ بتا_دوجمله‌ای، تغییرنگار برآوردشده نسبت‌های نمونه‌ای به هر یک از جفت مکان‌ها با توجه به اختلاف واریانس‌ها وزن اختصاص می‌دهد، به‌طوری که با استفاده از این وزن به اختلاف بین جفت مکان‌های فضایی با اندازه نمونه بزرگ‌تر اهمیت بیش‌تری داده می‌شود. در مدل هم‌کریگینگ دوجمله‌ای هیچ گونه وزنی به تغییرنگار نسبت‌های نمونه‌ای اختصاص داده نمی‌شود. این عامل ممکن است سبب شود به نقاطی با اندازه نمونه کوچک اهمیت بیش‌تری داده شود.

با در نظر گرفتن این اختلاف‌ها، انتظار داریم مدل کریگینگ بتا_دوجمله‌ای، برآوردهای دقیق‌تری برای پارامترها و خطای پیش‌بینی کوچک‌تری تولید کند.

۱.۱.۴ معادلات هم‌کریگینگ دوجمله‌ای

برای این‌که مقدار نسبت را در مکان جدید s_0 محاسبه کنیم، دو گزینه داریم. اول، تعمیم هم‌کریگینگ معمولی است. فرض می‌کنیم که میانگین نامعلوم است و مقدار پیش‌گویی در مکان جدید عبارت است از

$$\hat{Y}(s_0) = \sum_{i=1}^{m} \lambda_i \frac{Z_i}{n_i}$$

که در آن m تعداد مکان‌های مشاهده‌شده و λ_i ها وزن‌های هم‌کریگینگ هستند. در این‌جا نیز شرط ناااریبی پیش‌گو $\sum_{i=1}^{m} \lambda_i = 1$ است. با توجه به مانایی مرتبه دوم میدان تصادفی داریم

$$E[Y(s_i)] = E\left[\frac{Z_i}{n_i}\right] = \mu.$$

بنابراین تحت شرط ناااریبی و می‌نیمم کردن واریانس خطای پیش‌گویی، معادلات هم‌کریگینگ دوجمله‌ای به‌صورت

$$\hat{Y}(s_0) = \sum_{i=1}^{m} \lambda_i C_{ij} + \mu = C_{0i}$$

$$\sum_{i=1}^{m} \lambda_i = 1$$

که در آن μ ضریب لاگرانژ، C_{ij} و C_{0i} نیز مشابه نمادهای معادله کریگینگ بتا ـ دوجمله‌ای (فصل دوم) است.

گزینه دوم، برآورد کریگینگ بدون شرط اریبی روی Y است. کریگینگ بدون شرط اریبی وابسته به شرایط زیر است:

$$E\left[\hat{Y}(s_0)\,|Y(s_0)\right] = Y(s_0).$$

لاجونی (۱۹۹۱) نشان داد که

$$E[Y(s_i)\,|Y(s_0)] = \mu + \frac{C_{0i}}{\sigma_Y^2}\{Y(s_0) - \mu\}$$

بنابراین علاوه بر یک بودن مجموع وزن‌ها، برای اطمینان از ناااریبی شرط اضافی

$$\sum_{i=1}^{m} \lambda_i C_{0i} = \sigma_Y^2$$

را داریم. بنابراین دستگاه کامل معادلات کریگینگ عبارت است از

$$\sum_{i=1}^{m} \lambda_i C^{\frac{Z_i}{n_i}}(s_i, s_j) + \psi_1 C^{Y}(s_0, s_i) + \psi_2 = C_{0i}(s_0, s_i)$$

$$\sum_{i=1}^{m} \lambda_i = 1,$$

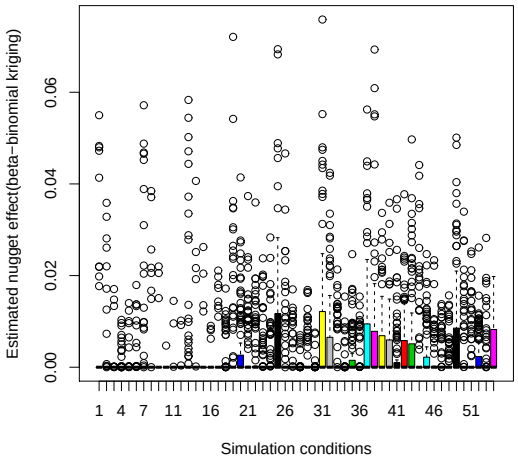
$$\sum_{i=1}^{m} \lambda_i C^{Y}(s_0, s_i) = \sigma_Y^2.$$

۲.۱.۴ برآورد پارامترهای هم‌کریگینگ دوجمله‌ای

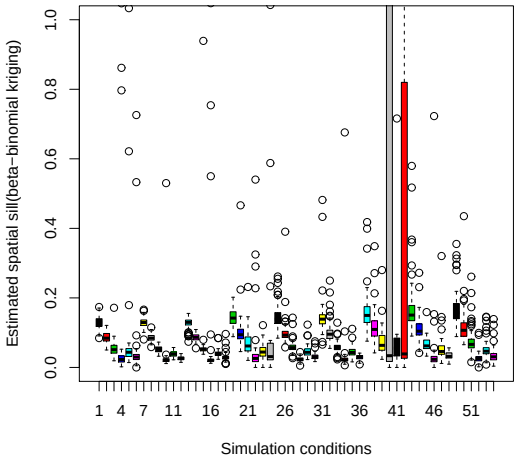
برای ارزیابی عملکرد برآورد پارامترها، برای هر سناریو ۱۰۰ مجموعه داده شبیه‌سازی کردیم. همان‌طور که قبلا بیان شد، با توجه به نتایج حاصل از شبیه‌سازی، برای برآورد پارامترها از تابع زیان $L_۲(\cdot)$ و تعریف تعداد نقاط همسایه $N^*(h)$، در استراتژی دوم، استفاده کردیم. برآوردهای پارامترهای تغییرنگار برای هر دو مدل بتا_دوجمله‌ای و دوجمله‌ای از می‌نیمم کردن تابع زیان $L_۲(\theta)$، که در تغییرنگار تجربی و تغییرنگار واقعی (کروی) مقداردهی شده است، به‌دست آمده‌اند. نمودارهای جعبه‌ای برآوردهای پارامترهای اثر قطعه‌ای، ازاره و دامنه برای مدل بتا_دوجمله‌ای در شکل ۱.۴ و برای مدل دوجمله‌ای در شکل ۲.۴ نشان داده شده‌اند. در هر دو شکل، ۱۸ سناریوی اول مربوط به دامنه ۵، ۱۸ سناریو دوم مربوط به دامنه ۱۰ و ۱۸ سناریو سوم مربوط به دامنه ۱۵ هستند. همچنین ۶ سناریو اول هر دامنه فضایی، اندازه نمونه کوچک، ۶ سناریو دوم اندازه نمونه متوسط و ۶ سناریو سوم اندازه نمونه بزرگ را نشان می‌دهند. پارامترهای واقعی دامنه فضایی به ترتیب برای این سناریوها ۵، ۱۰، و ۱۵ می‌باشند. اثر قطعه‌ای واقعی صفر است که تقریبا در تمام شبیه‌سازی‌ها به درستی برآورد شده است. اثر قطعه‌ای مدل دوجمله‌ای حتی کوچک‌تر از اثر قطعه‌ای کریگینگ بتا_دوجمله‌ای برآورد شده است. همچنین با توجه به دو نمودار جعبه‌ای مربوط به برآورد این پارامتر در هر دو مدل، می‌توان گفت اندازه نمونه در برآورد اثر قطعه‌ای تاثیری ندارد. در واقع الگوی قابل تشخیصی در شکل معلوم نیست که این نشان می‌دهد ۱) اصلاح تغییرنگار بتا_دوجمله‌ای، برآورد پارامترهای فرآیند بتای پنهان را بهبود بخشیده است؛ ۲) اصلاحی مشابه در تغییرنگار دوجمله‌ای نیز بهبودی مشابه را در پی داشته است.

نمودارهای جعبه‌ای برآورد پارامتر ازاره در هر دو مدل، چوله به راست هستند و در مدل بتا_دوجمله‌ای تعداد نقاط پرت کم‌تری نسبت به اثر قطعه‌ای وجود دارد. با افزایش دامنه فضایی نقاط پرت بیش‌تر هم می‌شوند. واریانس برآورد این پارامتر در مدل دوجمله‌ای خیلی بزرگ‌تر از مدل بتا_دوجمله‌ای است. مقادیر میانه ازاره فضایی تقریبا همیشه کوچک‌تر از تغییرپذیری میدان زیربنایی بتا است. در حقیقت تفاوت بین میانه ازاره و ازاره فضایی واقعی شدید است. این نشان می‌دهد که هم‌کریگینگ دوجمله‌ای ممکن است با همگرایی در تقابل باشد، زیرا در بسیاری از موارد برآوردهای ازاره برای داده‌ها بسیار غیرمنطقی هستند. در بعضی سناریوها، بیش از ۲۵ درصد شبیه‌سازی‌ها مقدار برآوردشده ازاره کاملا بی‌معنی دارند. از طرفی، برای مدل بتا_دوجمله‌ای نیز برآوردهای ازاره بزرگ‌تر از یک وجود دارند که با توجه به کراندار بودن توزیع بتا در فاصله صفر و یک طبیعی نیست و نشان می‌دهد همگرایی برای بعضی از توزیع‌ها ممکن است مشکل‌ساز شود.

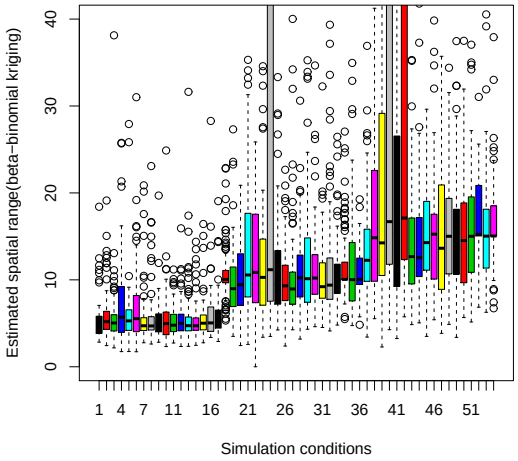
برای پارامتر دامنه فضایی، در مدل بتا_دوجمله‌ای، هر سه مقدار در نظر گرفته‌شده به خوبی بازیابی شده‌اند. روندی تقریبا مشابه برای مدل دوجمله‌ای برقرار است اما مشابه ازاره، واریانس برآوردها نسبت به مدل بتا_دوجمله‌ای خیلی بزرگ‌تر است.



(آ) اثر قطعه‌ای

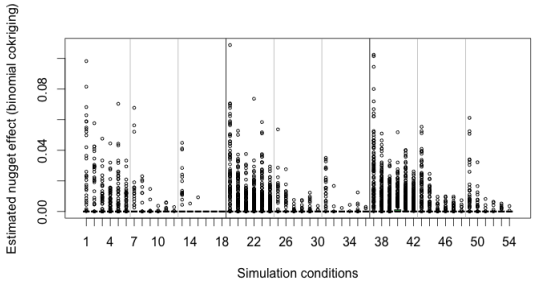


(ب) ازاره فضایی

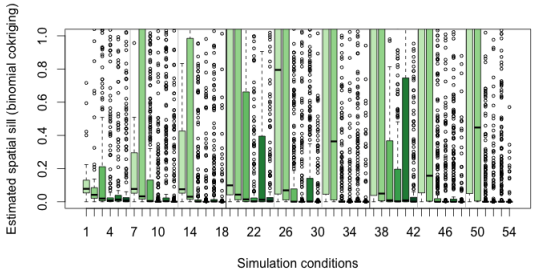


(ج) دامنه فضایی

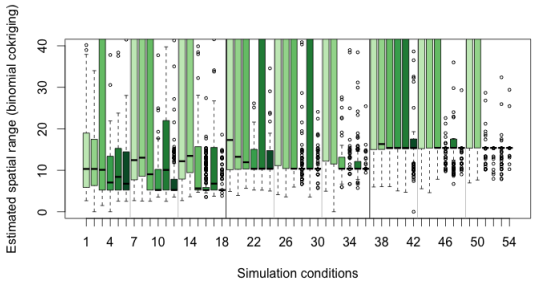
شکل ۱.۴: برآورد پارامترهای تغییرنگار بتا_دوجمله‌ای



(آ) اثر قطعه‌ای



(ب) ازاره فضایی



(ج) دامنه فضایی

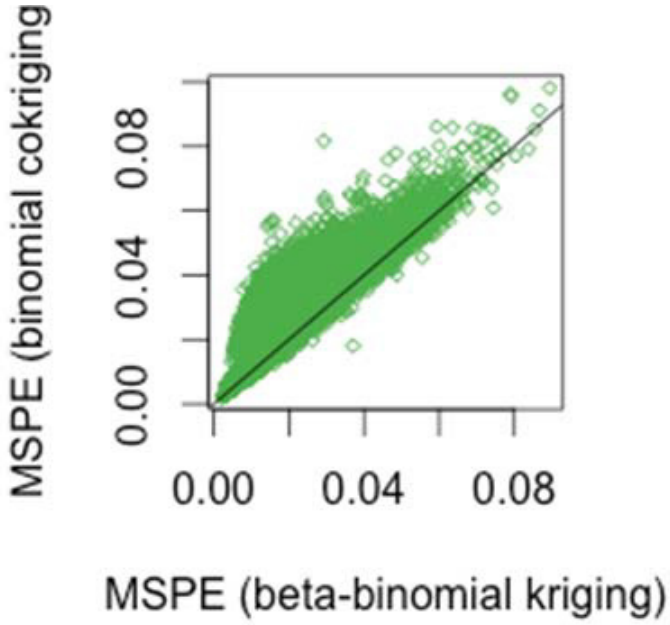
شکل ۲.۴: برآورد پارامترهای تغییرنگار دوجمله‌ای

با توجه به برآوردهای پارامترها در مدل هم‌کریگینگ دوجمله‌ای، می‌توان گفت که نگرانی جدی در مورد استفاده از این مدل برای پیش‌گویی فضایی وجود دارد. به نظر می‌رسد در این مدل، بهترین سناریوی استفاده از آن سناریویی با اندازه نمونه بزرگ، دامنه فضایی نسبتا بزرگ و توزیع زیربنایی بتا با تغییرپذیری کم است.

۳.۱.۴ پیش‌گویی فضایی

برای بررسی عملکرد پیش‌گویی دو روش معرفی‌شده، ۱۰۰ مجموعه داده را (به ازای هر سناریوی شبیه‌سازی) بر روی یک مشبکه ۵۰ × ۵۰ شامل ۲۵۰۰ گره تولید کردیم و با استفاده از دو روش کریگینگ بتا_دوجمله‌ای و هم‌کریگینگ دوجمله‌ای (با پارامترهای برآوردشده تغییرنگار)

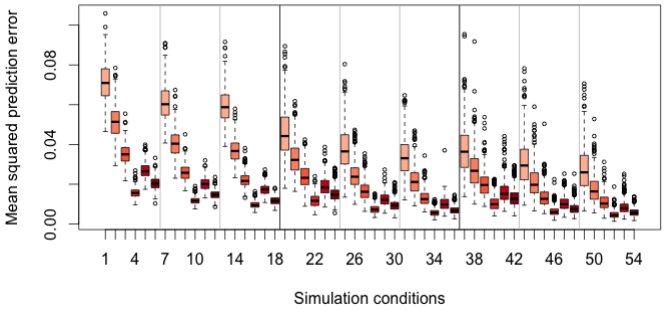
مقادیر نسبت‌های دوجمله‌ای را، برای هر مجموعه داده، پیش‌گویی کردیم. سپس میانگین توان دوم خطای پیش‌گویی[۲] (MSPE) را برای دو مدل بتا–دوجمله‌ای و دوجمله‌ای محاسبه کردیم. شکل ۳.۴ نمودار پراکنش مقادیر MSPE دو روش را نشان می‌دهد. این شکل مقادیر MSPE تجمیع‌شده همه سناریوها را با هم نشان می‌دهد. از این شکل واضح است که مدل کریگینگ بتا–دوجمله‌ای، بر حسب معیار MSPE، عملکرد کاملا برتری نسبت به مدل هم‌کریگینگ دوجمله‌ای دارد. در واقع، در ۷۹/۴ درصد شبیه‌سازی‌ها، کریگینگ بتا–دوجمله‌ای MSPE کم‌تری از هم‌کریگینگ دوجمله‌ای تولید کرده است. در مواردی هم که هم‌کریگینگ دوجمله‌ای MSPE کوچک‌تری دارد، اختلاف نسبتا ناچیز است.



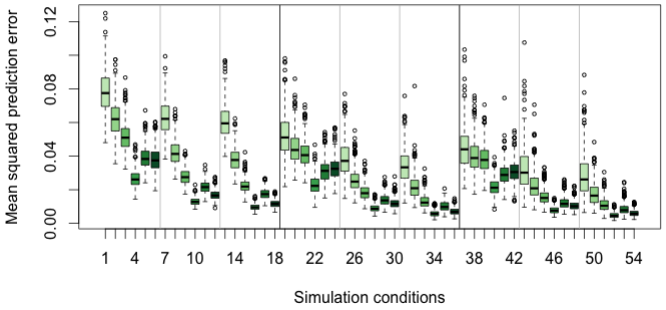
شکل ۳.۴: مقایسه مقادیر MSPE برای دو روش کریگینگ بتا–دوجمله‌ای و هم‌کریگینگ دوجمله‌ای

برای بررسی تاثیر حجم نمونه و اندازه دامنه فضایی بر قدرت پیش‌گویی، مقادیر MSPE را برای دو مدل به تفکیک هر ۵۴ سناریو در شکل ۴.۴ گزارش کرده‌ایم. با توجه به این دو شکل اختلاف توانایی دو روش، به‌ویژه در اندازه نمونه کوچک، مشهود است. ولی دامنه تاثیر چندانی ندارد. بنابراین می‌توان نتیجه‌گیری کرد که برای اندازه‌های نمونه کوچک و متوسط دقت کریگینگ بتا–دوجمله‌ای بیشتر است و باید از آن استفاده کرد.

[۲] Mean Squared Prediction Errors



(آ) میانگین توان دوم خطای پیش‌گویی مدل بتاـ دوجمله‌ای



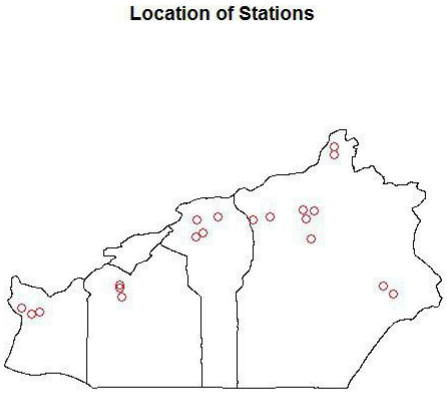
(ب) میانگین توان دوم خطای پیش‌گویی مدل دوجمله‌ای

شکل ۴.۴: میانگین توان دوم خطای پیش‌گویی دو مدل به ازای ۵۴ سناریوی شبیه‌سازی

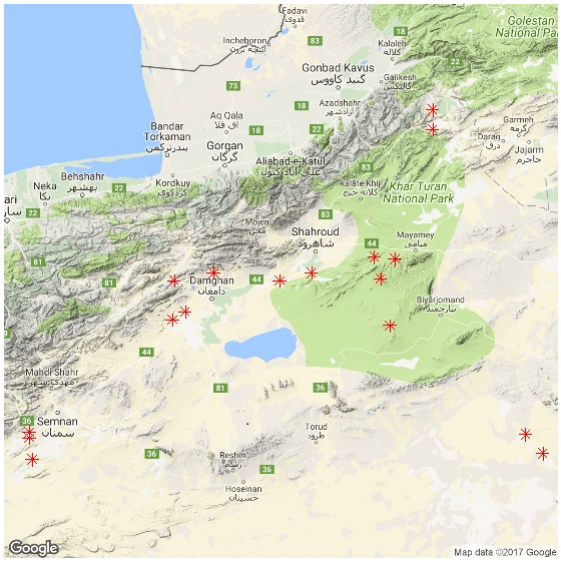
۲.۴ پیش‌گویی و نقشه پهنه‌بندی نرخ بارش در استان سمنان

در این بخش کاربرد روش پیش‌گویی معرفی‌شده را بر روی یک مجموعه داده واقعی مورد بررسی قرار می‌دهیم. مجموعه داده مورد نظر مربوط به تعداد روزهای دارای بارندگی در استان سمنان است که به مدت ۵ ماه از اول آبان تا پایان اسفند سال ۱۳۹۱ طی ۱۴۹ روز در ۲۰ ایستگاه هواشناسی استان سمنان ثبت شده‌اند. با توجه به این‌که مشاهدات تعداد روزهای دارای بارندگی است و داده‌ها به موقعیت قرار گرفتن‌شان در محیط مورد مطالعه بستگی دارند، بنابراین پاسخ‌ها به‌صورت گسسته و شمارشی هستند و ماهیت فضایی دارند. هدف از تحلیل این داده‌ها، تهیه نقشه پهنه‌بندی پیش‌گویی نرخ بارش در کل سطح استان با استفاده از مدل منعطف بتاـ دوجمله‌ای فضایی است. موقعیت ایستگاه‌های هواشناسی بر روی نقشه استان

سمنان در شکل‌های ۵.۴ و ۶.۴ مشخص شده‌اند. با توجه به بارش بیش‌تر در شمال استان، تعداد ایستگاه‌های هواشناسی در کرانه شمالی استان سمنان ۱۸ و در کرانه جنوب شرقی آن فقط ۲ ایستگاه است.



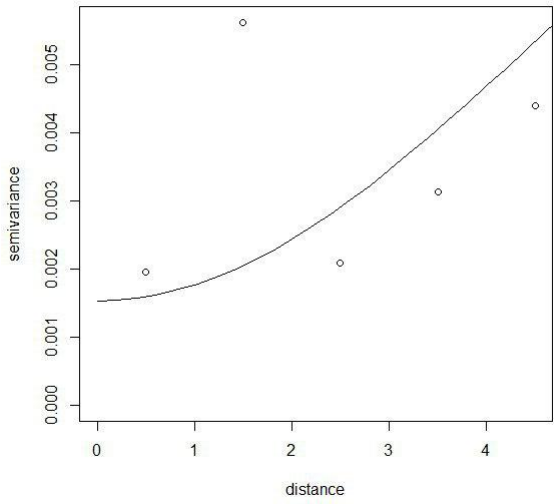
شکل ۵.۴: موقعیت ایستگاه‌های هواشناسی استان سمنان



شکل ۶.۴: موقعیت ایستگاه‌های هواشناسی استان سمنان

برای برآورد ساختار همبستگی روزهای دارای بارش در استان سمنان طی ۱۴۹ روز ثبت‌شده در سال ۱۳۹۱، ابتدا تغییرنگار داده‌های بارندگی را برآورد کردیم. برای استفاده از یک مدل همبستگی پارامتری، مدل‌های مختلف همبستگی فضایی را به تغییرنگار برآورد شده

برازش دادیم. بهترین مدل برازش شده که یک مدل گاوسی است در شکل ۷.۴ نشان داده شده است.

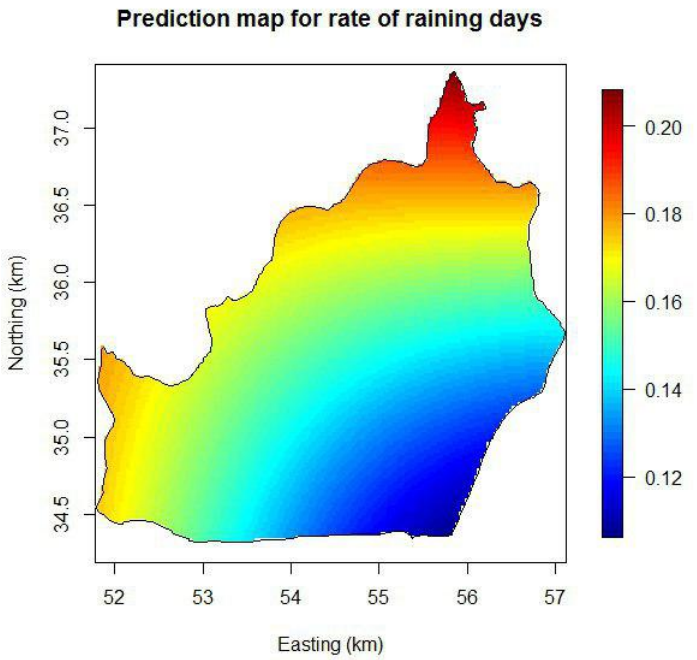


شکل ۷.۴: نیم‌تغییرنگار تجربی و برازش مدل گاوسی به داده‌ها

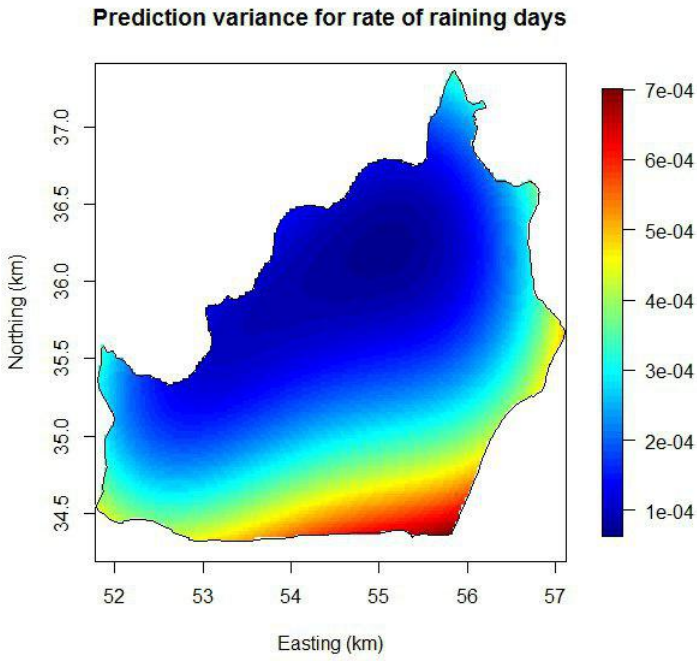
برای پیش‌گویی نقاط فاقد مشاهده در استان از یک مشبکه ۵۰ × ۵۰ که کران‌های استان را کاملا پوشش می‌دهند، استفاده کردیم. تابع تغییرنگار فضایی را نیز مدل گاوسی برازش‌شده بر تابع تغییرنگار تجربی در نظر گرفتیم. متغیر پاسخ، تعداد روزهای دارای بارندگی است که فرض کردیم دارای توزیع دوجمله‌ای شرطی مستقل باشد و متغیر پنهان آن نیز دارای توزیع بتا است. متاسفانه این مجموعه داده شامل متغیرهای رگرسیونی نبود که از اطلاعات آن‌ها برای پیش‌گویی کاراتر بتوان استفاده کرد. شکل ۸.۴ نقشه پهنه‌بندی پیش‌گویی نرخ روزهای دارای بارندگی را روی کل استان سمنان بر اساس مدل بتا_دوجمله‌ای نشان می‌دهد. همان‌طور که از روی شکل واضح است، هر چه به سمت شمال استان حرکت می‌کنیم، نرخ بارندگی نیز افزایش می‌یابد به‌طوری که این نرخ به بیش از ۲۰ درصد نیز می‌رسد. یکی از دلایل این ویژگی همجواری با استان‌های شمالی کشور یعنی گلستان و مازندران است.

شکل ۹.۴ نیز واریانس نرخ بارش پیش‌گویی را نمایش می‌دهد با توجه به آن‌که اکثر اطلاعات در کرانه شمالی استان ثبت شده‌اند، واریانس پیش‌گویی نیز در کرانه شمالی کوچک‌تر از کرانه جنوبی و مرکز استان می‌باشد.

از نقشه پهنه‌بندی پیش‌گویی نرخ بارش، می‌توان برای تصمیم‌گیری‌های مدیریت آب استفاده کرد. این نقشه، وضعیت بارش در سطح استان را برای تمام نقاط فاقد ایستگاه هواشناسی، به‌صورت تقریبی، نمایش می‌دهد.



شکل ۸.۴: نقشه پهنه‌بندی پیش‌گویی نرخ بارش در استان سمنان



شکل ۹.۴: واریانس نرخ پیش‌گویی بارش در استان سمنان

فصل ۵

تحلیل فضایی داده‌های شمارشی مشبکه‌ای با مدل اتوبتا ـ دوجمله‌ای

در این فصل به تعمیم مدل بتا ـ دوجمله‌ای برای تحلیل فضایی داده‌های مشبکه‌ای در چارچوب استنباط بیزی می‌پردازیم. با توجه به پیچیدگی توزیع پسین این مدل، از یک روش تقریبی بیزی به نام تقریب لاپلاس آشیانی جمع‌بسته[۱](INLA) استفاده می‌کنیم. این روش توسط رو و همکاران (۲۰۰۹) برای استنباط در مدل‌های گاوسی پنهان[۲] معرفی شد که مدل‌های فضایی را نیز شامل می‌شود. استفاده از روش INLA در برازش مدل‌های فضایی و فضایی ـ زمانی رو به افزایش است که می‌توان به منابع اشاره‌شده در میرزاجانی (۱۳۹۵) مراجعه کرد. در ادامه داده‌های مربوط به بیماری سرطان معده در استان گیلان را با استفاده از روش INLA تحلیل می‌کنیم. هدف از تحلیل این داده‌ها، تهیه نقشه پهنه‌بندی نرخ سرطان معده در استان گیلان با استفاده از مدل منعطف بتا ـ دوجمله‌ای و همچنین مقایسه آن با مدل دوجمله‌ای است.

---

[۱] Integrated nested Laplace approximation

[۲] Latent Gaussian models

۱.۵ مقدمه

در بسیاری از مطالعات، تحلیل داده‌های فضایی که به‌صورت شمارشی یا متوسط‌گیری روی نواحی دلخواه فضای مورد مطالعه هستند، مورد نظر است. این داده‌ها، داده‌های مشبکه‌ای نامیده می‌شوند. برای مدل‌بندی این داده‌ها از یک میدان تصادفی مانند

$$Z(\cdot) = \{Z(s) : s \in D\}$$

استفاده می‌شود که در آن D زیرمجموعه‌ای ثابت و شمارش‌پذیر (معمولا متناهی) از $\Re^d$ می‌باشد، همچنین D ممکن است به‌طور منظم یا نامنظم شبکه‌بندی شده باشد. تصاویر ماهواره‌ای از سطح زمین که در آن‌ها سطح زمین به تعداد مساوی مستطیل کوچک به‌صورت شبکه منظم در $\Re^2$ هستند، مثالی برای داده‌های مشبکه‌ای منظم است و تعداد افراد مبتلا به سرطان در تمام مناطق خدمات درمانی کشور یک مثال برای داده‌های مشبکه‌ای نامنظم است.

در حقیقت میدان تصادفی مشبکه‌ای را می‌توان به‌صورت $\{Z(s_i) : s_i \in \{s_1, \cdots, s_n\}\}$ نمایش داد که در آن $s_1, \cdots, s_n$ یک افراز از ناحیه D است، یعنی

$$\cup_{i=1}^{n} s_i = D \qquad s_i \cap s_j = \circ \qquad i \neq j.$$

برای درک بهتر، اگر D مساحت استان گیلان و $s_1, \cdots, s_{16}$ شهرستان‌های گیلان باشند، در این‌صورت $Z(s_1), \cdots, Z(s_{16})$ را می‌توان تعداد افراد مبتلا به سرطان معده، تعداد حوادث رانندگی یا نرخ رشد اقتصادی هر شهرستان در نظر گرفت. این نوع داده‌ها روش تحلیل مخصوص خود را دارند. در تحلیل داده‌های مشبکه‌ای، هدف اصلی مدل‌بندی مشاهدات است، در صورتی که در زمین‌آمار پیش‌گویی متغیر در یک موقعیت جدید فاقد مشاهده هدف اصلی محسوب می‌شود. از آن‌جایی که داده‌های مشبکه‌ای در یک زیرمجموعه شمارش‌پذیر از $\Re^d$ امکان تحقق و مشاهده دارند، بنابراین باید دقت شود برای نقاطی می‌توان پیش‌گویی کرد که امکان تحقق دارند و در غیر این‌صورت نتیجه بی‌معنی خواهد بود. در متون مختلف علمی از عبارت‌هایی مانند داده‌های ناحیه‌ای و منطقه‌ای[۳] نیز برای داده‌های مشبکه‌ای استفاده می‌شوند.

برای این‌که ساختار فضایی داده‌های مشبکه‌ای را مدل‌بندی کنیم، دو رهیافت متفاوت وجود دارند که تفاوت این دو رهیافت به مجموعه اندیس‌گذار D در میدان تصادفی

$$\{Z(s) : s \in D\}$$

برمی‌گردد. در رهیافت اول، فرض می‌شود که مشاهدات از یک مجموعه اندیس‌گذار پیوسته آمده‌اند. همچنین فرض می‌شود داده‌ها برای هر ناحیه در مرکز ناحیه مشاهده شده‌اند. در

[۳] Regional data

واقع خلاصه اطلاعات تمام ناحیه در مرکز ناحیه است، یعنی اطلاعات کل ناحیه به یک نقطه منتسب می‌شود. در این رهیافت مدل‌بندی داده‌ها با اندیس پیوسته صورت می‌گیرد، در صورتی که ماهیت داده‌ها مشبکه‌ای است و در هر نقطه امکان تحقق وجود ندارد. در حقیقت در این رهیافت، روش زمین‌آمار که مختص داده‌های با اندیس پیوسته است برای داده‌های با اندیس گسسته به‌کار برده می‌شود. دلیل استفاده از این رهیافت، سر راست بودن تابع کوواریانس فضایی و معمولا تفسیر راحت آن می‌باشد (ملانی، ۲۰۰۲).

در رهیافت دوم، برای مدل‌بندی ساختار فضایی داده‌های مشبکه‌ای، طبیعت گسسته مکان داده‌های مشبکه‌ای در نظر گرفته می‌شود و ساختار همسایگی مشاهدات بر اساس شکل مشبکه‌ها تعریف می‌شود. بنابراین به‌جای اندازه‌گیری فاصله بین مراکز ناحیه‌ها یک روش دیگر بر اساس همسایگی ناحیه‌ها تعریف می‌شود. به‌عنوان مثال از میزان مرز مشترک ناحیه‌ها در ساختار استفاده می‌کنند. در حقیقت وقتی ساختار همسایگی به این صورت تعریف می‌شود، مدل‌هایی مانند اتورگرسیو در سری‌های زمانی تعریف می‌شود. با این دیدگاه ویتل (۱۹۵۴) مدل اتورگرسیو همزمان، هاینینگ (۱۹۷۸) مدل میانگین متحرک، بارتلت (۱۹۷۱) و بیسگ (۱۹۷۴) مدل‌های اتورگرسیو شرطی[۴] (CAR) را برای تحلیل داده‌های مشبکه‌ای ارایه کردند.

۲.۵ داده‌های فضایی مشبکه‌ای

بیرخوف و هامرسلی (۱۹۶۷) و مازارینو (۱۹۸۳) عبارت مشبکه را اولین بار در متون ریاضی مطرح کردند، سپس بیسگ (۱۹۷۵، ۱۹۷۴) از آن در تحلیل داده‌های شبکه‌بندی‌شده استفاده کرد. بیسگ (۱۹۷۴) مدل‌های اتوگاوسی[۵]، اتودوجمله‌ای[۶]، اتولجستیک[۷]، اتوپواسون[۸] را برای تحلیل داده‌های مشبکه‌ای در حالت یک متغیره به‌کار برد، که در همه آن‌ها از میدان‌های تصادفی گاوسی استفاده کرده است. جاکسون (۲۰۰۳) روش‌های مختلف برآورد پارامترها و روش‌های مختلف شبیه‌سازی این مدل‌ها را مورد مطالعه قرار داده و سپس آن‌ها را مقایسه کرده است. چن (۲۰۰۲) تحلیل بیزی مدل اتولجستیک را برای تحلیل توزیع گونه خاصی از گیاهان مطرح کرده است. از بین این مدل‌ها مدل اتوگاوسی یا اتورگرسیو شرطی گاوسی به‌دلیل وجود صورت بسته درستنمایی از لحاظ کاربردی مورد توجه قرار گرفته است. حالت چندمتغیره آن اولین بار توسط ماردیا (۱۹۸۸) برای پردازش تصاویر معرفی گردید و سپس بیلهمیر و همکاران (۱۹۹۷)، پتیت و همکاران (۲۰۰۲)، کارلین و بنرجی (۲۰۰۳)، و گلفند و وناتسو (۲۰۰۳) در مورد آن مطالعات بیش‌تری انجام دادند. تلاش آن‌ها استفاده از رهیافت بیزی به‌ویژه بیزی سلسله‌مراتبی برای برآورد پارامترهای مدل اتوگاوسی چندمتغیره بوده

[۴] Conditional autoregressive
[۵] Auto Gaussian
[۶] Auto binomial
[۷] Auto logistic
[۸] Aoto Poisson

است. ساین و کرسی (۲۰۰۴) مدل اتوگاوسی چندمتغیره را در حالتی که ماتریس همبستگی متقابل فضایی نامتقارن است، ارایه کرده‌اند. اغلب در متون از عنوان میدان تصادفی مارکوفی چندمتغیره[۹] به‌جای مدل اتوگاوسی چندمتغیره استفاده شده است.

همان‌گونه که در سری‌های زمانی رابطه هر مقدار جدید در آینده با مشاهدات گذشته نزدیک مدل‌بندی و مقدار آن پیش‌گویی می‌شود، در میدان تصادفی فضایی مشبکه‌ای نیز هر موقعیت جدید بر اساس داده‌های همبسته با آن، یعنی مشاهدات واقع در نواحی همسایه نزدیک، مدل‌بندی و در بعضی موارد پیش‌گویی می‌شود. بنابراین به منظور مدل‌بندی داده‌های ناحیه‌ای y_i، $i = ۱, \cdots, n$، لازم است اثر سایر همسایه‌ها بر این ناحیه مشخص شود. مجموعه همسایگی موقعیت i با N_i نمایش داده می‌شود و عبارت است از مجموعه موقعیت‌هایی که بیش‌ترین اطلاع را در خصوص مشاهده y_i در اختیار دارند. همسایگی یک ناحیه بر این اساس که منظم یا نامنظم باشد به صورت دو موقعیت هم‌مرز یا دو ناحیه که در یک فاصله معین از هم قرار دارند، تعریف می‌شود. اغلب، همسایگی یک مکان به‌صورت تجربی و بر اساس فاصله تعیین می‌شود. مثال‌های زیر روش‌های مختلف تعیین همسایگی را نشان می‌دهند.

- در مثال معروف داده‌های نشانگان مرگ و میر ناگهانی نوزادان[۱۰] (SIDS) در کارولینای شمالی، نواحی یا چندضلعی‌ها توسط طول و عرض جغرافیایی مشخص شده‌اند و نواحی که در شعاع ۳۰ مایلی ناحیه i قرار دارند، همسایه آن تلقی می‌شوند. منطقه کارولینا مثالی از مشبکه نامنظم با پوشش چند ضلعی است.
- برای بررسی میزان کود شیمیایی در یک زمین کشاورزی که به کرت‌های هم‌اندازه و منظم تقسیم شده است، همسایگی هر خانه ممکن است توسط همسایه‌ها از مرتبه اول (کرت‌های همجوار با هر کرت) یا از مرتبه اول و دوم در نظر گرفته باشند. در جدول ۱.۵ برای مکان $\times$، همسایه‌های مرتبه اول با علامت $\bullet$ و همسایه‌های مرتبه دوم با علامت $\otimes$ مشخص شده‌اند.

			$\otimes$			
			$\bullet$			
	$\otimes$	$\bullet$	$\times$	$\bullet$	$\otimes$	
			$\bullet$			
			$\otimes$			

جدول ۱.۵: مشبکه منظم با پوشش شبکه‌ای از مربع‌ها

[۹] Markov random fields

[۱۰] Sudden infant death syndrome

- در مدل‌های اقتصادی، همسایگی ممکن است بر اساس تبادلات اقتصادی بین دو ناحیه مشخص شود. به‌عنوان مثال اگر متغیر مورد مطالعه، تولیدات اقتصادی کشورهای مختلف باشد، همسایه ناحیه i ام، کشورهایی هستند که با آن مبادلات مثبت تجاری دارند. بنابراین دو کشور همسایه، لزوما مرز مشترک جغرافیایی ندارند.

۱.۲.۵ روش‌های انتساب وزن‌های همسایگی

پس از آن‌که نوع همسایگی را مشخص کردیم، بیان تاثیر این همسایگی دارای اهمیت است. این تاثیر از طریق انتساب وزن به هر یک از همسایه‌ها به‌دست می‌آید. برای نشان دادن وزن‌های همسایگی از ماتریس مجاورت[۱۱] C استفاده می‌کنیم که در آن درایه c_{ij} میزان تاثیر ناحیه i روی ناحیه j را نمایش می‌دهد. همچنین $c_{ii} = ۰$ و $c_{ij} = c_{ji}$. ماتریس C را می‌توان به‌صورت

$$C = \left(\frac{c_{ij}}{c_{i+}}\right)$$

به یک ماتریس استاندارد تبدیل کرد که در آن $c_{i+} = \sum_j c_{ij}$. برای مشخص کردن ماتریس C، می‌توان از ویژگی‌های مختلفی مانند فاصله بین مراکز نواحی، مرز مشترک نواحی، جمعیت هر ناحیه، مساحت نواحی و میزان مبادلات فرهنگی و تجاری نواحی استفاده کرد. ساده‌ترین حالت، انتخاب ماتریس مجاورت به‌صورت

$$c_{ij} = \begin{cases} ۱ & j \in N_i \\ ۰ & i = j \end{cases}$$

است. یکی از راه‌های انتساب وزن‌ها، استفاده از فاصله بین نواحی است. به‌عبارت دیگر، تعریف تابع نزولی بر اساس فاصله به‌گونه‌ای که به همسایه‌های نزدیک، وزن بزرگ‌تر و به همسایه‌های دور، وزن کوچک‌تری اختصاص می‌دهد. مساحت‌های نواحی نیز می‌توانند به این صورت برای تعریف درایه‌های ماتریس C استفاده شوند به‌طوری که $c_{ij} = \frac{A_{ij}}{\sum_{j \in N_i} A_{ij}}$، که در آن A_{ij} مساحت j امین همسایه ناحیه i است. در صورتی که جمعیت نواحی معلوم باشد، می‌توان وزن‌ها را به‌صورت

$$c_{ij} = \frac{A_{ij}}{\sum_{j \in N_i} A_{ij}} \times \frac{n_j}{n_i}$$

تعدیل کرد، به‌طوری که n_i و n_j به‌ترتیب جمعیت ناحیه i ام و جمعیت ناحیه j ام هستند. همچنین با در نظر گرفتن فاصله بین نواحی، می‌توان وزن‌های c_{ij} را به‌صورت

$$c_{ij} = \frac{A_{ij}}{\sum_{j \in N_i} A_{ij}} \times \frac{n_j}{n_i} \times d_{ij}^{\gamma}$$

[۱۱] Adjacency matrix

نوشت که در آن d_{ij} فاصله بین مرکز دو ناحیه و γ پارامتری است که می‌بایست برآورد شود. یک روش دیگر برای تعریف ماتریس C استفاده از مرز مشترک نواحی است، به‌عبارت دیگر، می‌توان نوشت

$$c_{ij} = \left(\frac{L_{ij}}{L_i}\right)^{\tau} \qquad (۱.۵)$$

که در آن L_{ij} طول مرز مشترک ناحیه i با ناحیه j، L_i محیط ناحیه i و $\tau \geq ۰$ پارامتری است که باید برآورد شود. حالت کلی‌تری از روش (۱.۵) در نظر گرفتن درایه‌ها به‌صورت

$$c_{ij} = \frac{L_{ij}}{L_i} d_{ij}^{-\gamma^{\tau}} \qquad (۲.۵)$$

است، که در آن هم از فاصله بین نواحی و هم از طول مرزهای مشترک دو ناحیه استفاده می‌شود.

کرسی (۱۹۹۳) ضمن معرفی تابع همسایگی (۲.۵)، تعریف وزن‌های همسایگی را بر اساس تابع مبتنی بر فاصله که در آن به همسایه‌های دور، وزن کوچک‌تر و به همسایه‌های نزدیک وزن بزرگ‌تری، به‌صورت

$$C_{ij} = \begin{cases} \rho \frac{d_{ij}^{-k}}{c(k)} \frac{n_i}{n_j} & j \in N_i \\ ۰ & j \notin N_i \end{cases}$$

داده می‌شود، پیشنهاد کرد که در آن ρ مقداری ثابت، d_{ij} فاصله بین ناحیه i و همسایه j ام، n_i و n_j جمعیت‌های دو ناحیه i و j، N_i مجموعه همسایگی‌های ناحیه i و

$$c(k) = \max\{d_{ij}^{-k}, j \in N_i, i = ۱, ۲, \cdots, n\}$$

عامل مقیاس است که برای بدون واحد کردن وزن‌های c_{ij} به‌کار می‌رود. همچنین k مقداری ثابت است که معمولا مقادیر ۰، ۱ یا ۲ را برای آن در نظر می‌گیرند.

۳.۵ مدل‌بندی داده‌های مشبکه‌ای

برای مدل‌بندی داده‌های مشبکه‌ای معمولا به دو نوع تغییرات توجه می‌شود. تغییرات بزرگ مقیاس که در آن میانگین به دلیل وابستگی موقعیت‌های فضایی به سایر موقعیت‌های تبیینی تغییر می‌کند و تغییرات کوچک مقیاس که به‌دلیل اثرات متقابل همسایگی‌ها ایجاد می‌شود. معمولا شکل کلی

$$Z(s) = \mu(s) + \epsilon, \quad s \in D$$

برای مدل‌بندی داده‌های مشبکه‌ای تعریف می‌شود، که به آن مدل رگرسیون فضایی نیز می‌گویند و در آن $Z(s)$، مقدار فرآیند تصادفی در موقعیت s است، $\mu(s)$ میانگین میدان تصادفی در موقعیت s است، که ممکن است ثابت یا تابعی از متغیرهای تبیینی باشد و ϵ بیانگر تغییرات

کوچک مقیاس است که معمولا توزیع نرمال $\epsilon \sim N(\circ, \Sigma)$ برای آن در نظر گرفته می‌شود و Σ با ابعاد $n \times n$ ماتریس واریانس متغیر تصادفی در همه موقعیت‌هاست. در واقع تغییرات کوچک مقیاس در Σ منعکس می‌گردد. انتخاب‌های مختلف برای ساختار Σ موجب روش‌های مختلف مدل‌بندی داده‌های مشبکه‌ای به‌صورت اتورگرسیو شرطی CAR، اتورگرسیو همزمان[۱۲] (SAR) و میانگین متحرک MA شده است. در همه این مدل‌ها، پذیره توزیع نرمال چندمتغیره مورد نیاز است و ساختار کوواریانس Σ در هر کدام از آن‌ها به‌صورت زیر در نظر گرفته می‌شود:

$$CAR : \Sigma = (I - \rho W)^{-1} D\sigma^2$$
$$SAR : \Sigma = \left[(I - \rho W)' D (I - \rho W)\right]^{-1} \sigma^2$$
$$MA : \Sigma = (I - \rho W) D (I - \rho W)' \sigma^2$$

که در آن‌ها ρ و σ^2 پارامترهای تغییرات کوچک مقیاس‌اند، W ماتریس وزن‌های همسایگی و D ماتریس قطری برای منظور کردن واریانس ناهمگن توزیع‌های کناری است. ماتریس D گاهی ماتریس وزن نامیده می‌شود، که در این‌صورت نباید با ماتریس W اشتباه شود. دو مدل CAR و SAR مشابه فرآیند اتورگرسیو در سری‌های زمانی تعریف می‌شوند. در مدل‌های مذکور اگر ماتریس Σ متقارن باشد، سبب سادگی بسیاری از محاسبات می‌شود. معمولا از ماتریس D برای متقارن کردن Σ استفاده می‌شود. به‌عنوان مثال اگر

$$Var\left(Z(s_i) | \{Z(s_j); j \in N_i\}\right) = \sigma_i^2 = \frac{\sigma^2}{n_i}$$

باشد، در این‌صورت با انتخاب ماتریس قطری D با عناصر قطری $\frac{1}{n_i}$ ماتریس Σ متقارن می‌شود.

۴.۵ معرفی مدل‌های اتورگرسیو شرطی

یکی از معروف‌ترین مدل‌ها برای تحلیل داده‌های مشبکه‌ای، مدل CAR است که به‌وسیله توزیع‌های احتمال شرطی ساخته می‌شود. موقعیت‌های قرارگیری داده‌ها در مدل‌بندی توزیع احتمال نقش اساسی ایفا می‌کنند. بنابراین لازم است نحوه تاثیر این موقعیت‌ها در مدل‌بندی مورد بررسی قرار گیرد. برای این منظور ابتدا ویژگی‌های زنجیر مارکوف را مورد بررسی قرار می‌دهیم. سپس مدل CAR را معرفی می‌کنیم.

۱.۴.۵ زنجیرهای مارکوف

فرض کنید میدان تصادفی $\{Z(s); s = \circ, ۱, \cdots\}$ در طول زمان به‌دست آمده است و بدون کاستن از کلیت مطلب $Z(\cdot)$ را گسسته در نظر بگیرید. در این صورت زنجیر مارکوف در مدل

[۱۲] Simultaneously autoregressive

احتمالی زیر صدق می‌کند (کرسی، ۱۹۹۳):

$$P(Z(۱),\cdots,Z(n)|Z(۰))=\prod_{s=۱}^{n}Q(Z(s),Z(s-۱)). \qquad (۳.۵)$$

یعنی احتمال توام متغیرها را می‌توان به‌صورت حاصل‌ضرب توابعی وابسته به متغیرهای متوالی نوشت، که به آن ویژگی فاقد حافظه نیز می‌گویند. با توجه به تعریف (۳.۵)

$$\begin{aligned}P(Z(i)|Z(۰),\cdots,Z(i-۱))&=\frac{P(Z(۱),\cdots,Z(i)|Z(۰)}{P(Z(۱),\cdots,Z(i-۱)|Z(۰))}\\&=\frac{\prod_{s=۱}^{i}Q_s\left((Z(s),Z(s-۱)\right)}{\sum_{Z(i)}\prod_{s=۱}^{i}Q\left((Z(s),Z(s-۱)\right)}\\&=\frac{Q_i\left((Z(i),Z(i-۱)\right)}{\sum_{Z(i)}Q\left((Z(s),Z(s-۱)\right)}.\end{aligned}$$

بنابراین هر مشاهده در زمان i فقط به مشاهده قبلی بستگی دارد. در مدل‌بندی داده‌های مشبکه‌ای نیز ویژگی فراموشی زنجیر مارکوف صدق می‌کند، زیرا مقدار مشاهده مکان s_i به‌طور منطقی فقط به موقعیت‌هایی که در همسایگی این مکان قرار دارند، بستگی دارد.

۲.۴.۵ معرفی مدل

بیسگ (۱۹۷۴) مدل اتورگرسیو شرطی را برای مدل‌بندی داده‌های مشبکه‌ای به‌صورت زیر تعریف کرد:

اگر میدان $\{Z(s_i);\quad i=۱,۲,\cdots,n\}$ گاوسی باشد، توزیع شرطی $Z(s_i)$ نسبت به بقیه مشاهدات $Z_{-i}=\{Z(s_j);\quad j\neq i=Z(s_j);j\in N_i\}$ به‌صورت

$$P(Z(s_i)|Z_{-i})=\frac{۱}{\sqrt{۲\pi s_i^۲}}\exp\{-\frac{۱}{۲}\frac{\left(Z(s_i)-\theta_i\{Z_{-i}\}\right)^۲}{s_i^۲}\}\qquad i=۱,\cdots,n \qquad (۴.۵)$$

است که در آن $E\left(Z(s_i)|Z_{-i}\right)=\theta_i\{Z_{-i}\}$ و $Var(Z(s_i)|Z_{-i})=s_i^۲$ به‌ترتیب میانگین و واریانس شرطی هستند. تحت شرایط نظم، میانگین شرطی به‌صورت

$$\theta_i\{Z_{-i}\}=E\left(Z(s_i)|Z_{-i}\right)=\mu_i+\sum_{j\in N_i}c_{ij}\left(Z(s_j)-\mu_j\right) \qquad (۵.۵)$$

خواهد بود، که در آن $c_{ii}=۰$ و برای $j\notin N_i$، $c_{ij}=۰$. با توجه به (۴.۵) و (۵.۵) و قضیه بیسگ، ثابت می‌شود اگر $(I-C)$ وارون‌پذیر باشد، آن‌گاه $\boldsymbol{Z}=(Z(s_۱),\cdots,Z(s_n))$ دارای توزیع نرمال با بردار میانگین $\boldsymbol{\mu}$ و ماتریس واریانس $(I-C)^{-۱}M$ است، یعنی

$$\boldsymbol{Z}\sim Gau(\mu,(I-C)^{-۱}M) \qquad (۶.۵)$$

که در آن $M=diag(s_۱^۲,\cdots,s_n^۲)$، $Var(Z(s_i)|Z_{-i})=s_i^۲$ و $\boldsymbol{\mu}=(\mu_۱,\cdots,\mu_n)$. برای خوش‌تعریف بودن (۶.۵) باید $(I-C)^{-۱}M$ متقارن و همیشه مثبت باشد، که شرط $c_{ij}s_j^۲=c_{ji}s_i^۲$ تقارن $(I-C)^{-۱}M$ را تضمین می‌کند.

در صورتی که در (۵.۵) متغیرهای تبیینی $\boldsymbol{x}$ حضور داشته باشند، تغییرات بزرگ مقیاس را به‌صورت $\mu = \boldsymbol{x}'\beta$ پارامتری می‌کنیم، که در آن $\boldsymbol{x}$ بردار q بعدی از متغیرهای تبیینی و β ضرایب رگرسیونی است. تغییرات فضایی کوچک مقیاس که ماتریس C آن‌ها را در برمی‌گیرد، می‌تواند به‌صورت $C = \phi G$ نوشته شود، که در آن ماتریس معلوم G ساختار همسایگی را تعریف می‌کند و ϕ شدت وابستگی فضایی را کنترل می‌کند. پارامتر ϕ^2 می‌تواند به‌عنوان توان دوم همبستگی شرطی فضایی بین همسایگی $Z(s_i)$ و $Z(s_j)$ تفسیر گردد، به انتخاب G و M بستگی دارد (کرسی، ۱۹۹۳).

با فرض هم‌واریانسی شرطی، برای $i = 1, \cdots, n$، $Var(Z(s_i)|Z_{-i}) = s^2$ و $M = s^2 I$. با ثابت نگه داشتن $(I - C)$، برآورد ماکسیمم درستنمایی β و s^2 به‌صورت

$$\hat{\beta} = \left[X'(I-C)X\right]^{-1} X'(I-C)Z$$
$$\hat{s}^2 = \frac{(Z - X\hat{\beta})' \left[(I-C)(Z - X\hat{\beta})\right]}{n}$$

حاصل می‌شوند. اگر برآوردگرهای فوق در تابع درستنمایی جایگزین شوند، می‌توان با ماکسیمم کردن تابع درستنمایی نسبت به c_{ij} ها برآوردهای متناظر را به‌دست آورد.

در مدل CAR، لگاریتم دترمینان $(I - C)$ که در تابع درستنمایی ظاهر می‌گردد، حائز اهمیت است. این موضوع در مدل‌های شرطی سری زمانی از اهمیت کم‌تری برخوردار است، زیرا در سری‌های زمانی اغلب n بزرگ است و لگاریتم دترمینان ماتریس کوواریانس به صفر میل می‌کند و از آن صرف نظر می‌شود. در مدل‌های فضایی مشبکه‌ای چنین فرضی معمولا وجود ندارد، به‌همین دلیل سعی زیادی در حل مساله توسط مکلود (۱۹۷۷)، دنت و مین (۱۹۷۸) و انسلی (۱۹۷۹) صورت پذیرفته است. موران و بیسگ (۱۹۷۵) با در نظر گرفتن حالت خاص مدل‌های شرطی نرمال به محاسبه دترمینان $(I - C)$ پرداخته‌اند.

۳.۴.۵ بررسی بیش‌تر مدل‌های شرطی

در این بخش مدل‌های شرطی و کاربرد آن‌ها در تحلیل داده‌های مشبکه‌ای را مورد بررسی قرار می‌دهیم. ابتدا قضیه بسیار مهم تجزیه بیسگ (۱۹۷۴) را بیان می‌کنیم، که با استفاده از آن توزیع توام $\boldsymbol{Z}$ در مدل CAR قابل دستیابی است.

قضیه ۱.۴.۵. (تجزیه بیسگ، ۱۹۷۱) اگر $P(\cdot)$ توزیع توام میدان $\{Z(s_i);\ \ i = 1, 2, \cdots, n\}$ باشد که در شرط مثبت بودن صدق می‌کند، آن‌گاه

$$\frac{P(Z)}{P(Y)} = \prod_{i=1}^{n} \frac{P(Y(s_i)|Z(s_1), \cdots, Z(s_{i-1}), Y(s_{i+1}), \cdots, Y(s_n))}{P(Y(s_i)|Z(s_1), \cdots, Z(s_{i-1}), Y(s_{i+1}), \cdots, Y(s_n))}$$

که در آن $\boldsymbol{Z} = (Z(s_1), \cdots, Z(s_n))$ و $\boldsymbol{Y} = (Y(s_1), \cdots, Y(s_n))$ دو تحقق ممکن از میدان $\boldsymbol{Z}$ هستند.

تعریف ۱.۴.۵. (جرگه[۱۳]). مجموعه‌ای از تمام مکان‌های تک عضوی که هیچ همسایه‌ای ندارند یا مکان‌هایی که تمام اعضای آن با هم همسایه‌اند، جرگه نامیده می‌شود. با وجود n موقعیت و بسته به تعریف ساختار همبستگی، اندازه جرگه‌ها ممکن است ۱، ۲ تا n باشد. اگر جرگه‌ها فقط از مرتبه یک باشند، در این صورت موقعیت‌ها باید از همدیگر مستقل باشند. در مباحث کاربردی جرگه‌های حداکثر تا مرتبه ۲ در نظر گرفته می‌شوند.

برای مدل‌بندی داده‌های فضایی بر اساس میدان تصادفی مارکوفی با فرض این‌که رویداد صفر در هر موقعیت شانس وقوع داشته باشد، بیسگ (۱۹۷۴) تابع $Q(\cdot)$ را برای داده‌های گسسته به‌صورت

$$Q(\boldsymbol{Z}) = \ln\left(\frac{P(\boldsymbol{Z})}{P(\circ)}\right)$$

تعریف کرد، که در آن توزیع توام $\boldsymbol{Z}$ به‌صورت

$$P(\boldsymbol{Z}) = P(\circ)\exp\{Q(\boldsymbol{Z})\}$$

قابل بیان است. چون

$$۱ = \sum P(\boldsymbol{Z}) = P(\circ)\sum \exp\{Q(\boldsymbol{Z})\}$$

در نتیجه

$$P(\circ) = \frac{\sum P(\boldsymbol{Z})}{\sum \exp\{Q(\boldsymbol{Z})\}} = \frac{۱}{\sum \exp\{Q(\boldsymbol{Z})\}}.$$

در نتیجه توزیع توام $\boldsymbol{Z}$ به‌صورت

$$P(\boldsymbol{Z}) = \frac{\exp\{Q(\boldsymbol{Z})\}}{\sum \exp\{Q(\boldsymbol{Z})\}} \qquad (۷.۵)$$

به‌دست می‌آید. در حالت پیوسته علامت جمع در مخرج رابطه (۷.۵) با علامت انتگرال تعویض می‌شود. به‌دلیل وجود $\sum \exp\{Q(\boldsymbol{Z})\}$ و $\int \exp\{Q(\boldsymbol{Z})\}$ در روابط بالا در اکثر توزیع‌ها توزیع توام $\boldsymbol{Z}$ صورت بسته ندارد و برای ماکسیمم کردن تابع درستنمایی نیاز به روش‌های عددی است. دقت کنید که میزان اطلاع درباره $Q(\cdot)$ هم‌ارز میزان اطلاع درباره $P(\cdot)$ است.

قضیه ۲.۴.۵. (بیسگ، ۱۹۷۴). تابع $Q(\cdot)$ دارای دو ویژگی زیر است

۱. تابع $Q(\cdot)$ در رابطه

$$\frac{P(Z(s_i)|\boldsymbol{Z}_{-}i)}{P(\circ(s_i)|\boldsymbol{Z}_{-}i)} = \frac{P(\boldsymbol{Z})}{P(\circ)} = e^{Q(\boldsymbol{Z})-Q(\boldsymbol{Z}_i)} \qquad (۸.۵)$$

صدق می‌کند، که در آن $\circ(s_i)$ نشانگر پیشامد $Z(s_i) = \circ$ است،

$$\boldsymbol{Z}_{-i} \equiv \{Z(s_j);\ \ j \neq i\}$$

و $\boldsymbol{Z}_i \equiv (Z(s_۱), \cdots, Z(s_{i-۱}), \circ, Z(s_{i+۱}), \cdots, Z(s_n))$.

[۱۳] Clique

۲. تابع $Q(\cdot)$ به‌صورت یکتا روی تکیه‌گاه $\boldsymbol{Z}$ ها به‌صورت

$$\begin{aligned} Q(\boldsymbol{Z}) = & \sum_{1\le i\le n} Z(s_i)G_i(Z(s_i)) \\ & + \sum\sum_{1\le i\le j\le n} Z(s_i)Z(s_j)G_{ij}(Z(s_i),Z(s_j)) \\ & \vdots \\ & + Z(s_1)\cdots,Z(s_n)G_{1,\cdots,n}(Z(s_1),\cdots,Z(s_n)) \end{aligned} \qquad (۹.۵)$$

بسط داده می‌شود، که در آن $G(\cdot)$ تابع پیوند است. در واقع $Q(\cdot)$ می‌تواند برحسب مجموع توابعی از زیرمجموعه‌های $Z(s_1),\cdots,Z(s_n)$ نمایش داده شود، که این زیرمجموعه‌ها توسط جرگه‌ها مشخص می‌شوند.

قضیه ۳.۴.۵. (هامرسلی‌ـ کلیفورد، ۱۹۷۱). فرض کنید $\boldsymbol{Z}$ یک میدان تصادفی مارکوفی باشد، که در شرط مثبت بودن صدق می‌کند، در این صورت شرط لازم برای بسط (۹.۵) آن است که اگر مکان‌های $i,j,\cdots,s$ در یک جرگه نباشند، آن‌گاه، $G_{ij\cdots,s}(\cdot)=0$ باشد.

قضیه هامرسلی‌ـ کلیفورد باعث شده است که میدان‌های تصادفی مارکوفی کاربردهای بسیار زیادی در متون فضایی پیدا کنند.

۵.۵ توزیع‌های نمایی شرطی

در این بخش خانواده توزیع‌های نمایی برای مدل‌بندی توزیع‌های شرطی

$$P(Z(s_i)|Z(s_j),\ j\in N_i)$$

را معرفی می‌کنیم. این خانواده منعطف، بسیاری از توزیع‌ها مانند توزیع دوجمله‌ای، پواسون، گاوسی و گاما را در بر می‌گیرد. همچنین مدل‌های فضایی که شکل نمایی دارند، به‌راحتی تفسیر می‌شوند (بیسگ، ۱۹۷۴). در حالت گسسته خانواده توزیع‌های نمایی شرطی بر اساس پارامترهای طبیعی به‌صورت

$$P(Z(s_i)|\boldsymbol{Z}_{-i}) = e^{L_i(\boldsymbol{Z}_{-i})M_i(Z(s_i))+N_i(Z(s_i))+V_i(\boldsymbol{Z}_{-i})} \qquad (۱۰.۵)$$

نوشته می‌شوند که در آن $L_i(\cdot)$ و $M_i(\cdot)$ ضابطه‌های معلومی دارند و $N_i(\cdot)$، $V_i(\cdot)$ توابعی از مقادیر مشاهده‌شده در همسایگی‌های موقعیت i هستند. بیسگ (۱۹۷۴) برای خانواده توزیع‌های نمایی شرطی با فرض وابستگی فقط جفتی بین موقعیت‌ها (یعنی $G_a(\cdot)=0$ برای هر a که تعداد عناصر آن سه یا بیشتر باشد) رابطه زیر را اثبات کرد. در حقیقت این نتیجه از قضیه هامرسلی‌ـ کلیفورد نتیجه می‌شود:

$$A_i(\boldsymbol{Z}_{-i}) = \alpha_i + \sum_{j=1}^{n}\theta_{ij}M_j(Z(s_j)\ \ j=1,2,\cdots,n,\ j\neq i \qquad (۱۱.۵)$$

که در آن $\theta_{ij} = G_{ij}(Z(s_i), Z(s_j))$ و $\theta_{ij} = \theta_{ji}$. بنابراین در صورتی که $j \notin N_i$ (همسایه نبودن i و j)، $\theta_{ij} = ۰$. همچنین برای شناسایی‌پذیر بودن مدل شرط $\theta_{ii} = ۰$ در نظر گرفته می‌شود.

با فرض وابستگی فقط جفتی، رده‌ای از مدل‌ها برای داده‌های گسسته مانند اتولجستیک، اتودوجمله‌ای و اتوپواسون و برای داده‌های پیوسته مدل اتوگاوسی معرفی شده‌اند. کاربردهایی از این مدل‌ها را می‌توان در مدل‌بندی میزان بروز ناحیه‌ای بیماری‌های خاص در پزشکی (کاوسی و همکاران، ۱۳۸۵؛ لاوسون، ۲۰۰۱)، مکان‌یابی مناسب برای ساخت و ساز مسکن در شهرسازی، بازاریابی و تعیین مکان‌های مناسب برای ایجاد فروشگاه‌های زنجیره‌ای نام برد. در ادامه مدل‌های معمول برای پاسخ‌های گسسته را معرفی می‌کنیم.

اتولجستیک

فرض کنید متغیر مورد مطالعه دودویی صفر و یک باشد، که وجود یا عدم وجود یک مشخصه را بیان می‌کند. تحت پذیره وابستگی فقط جفتی بین موقعیت‌ها، توزیع شرطی متغیر دودویی $Z(s_i)$ لجستیک است. با قرار دادن $G_i(۱) \equiv \alpha_i$ و $G_{ij}(۱,۱) = \theta_{ij}$ در (۹.۵) می‌توان نوشت

$$Q(\boldsymbol{Z}) = \sum_{i=۱}^{n} \alpha_i Z(s_i) + \sum \sum_{۱ \le i \le j \le n} \theta_{ij} Z(s_i) Z(s_j)$$

با محاسبه

$$Q(\boldsymbol{Z}) - Q(\boldsymbol{Z}_i) = \alpha_i Z(s_i) + \sum_{j=۱}^{n} \theta_{ij} Z(s_i) Z(s_j)$$

و جای‌گذاری در (۸.۵) و استفاده از برابری

$$P(Z(s_i) = ۱ | \boldsymbol{Z}_{-i}) = ۱ - P(۰(s_i) | \boldsymbol{Z}_{-i})$$

و در نظر گرفتن این‌که $Z(s_i) = ۰$ یا $Z(s_i) = ۱$، نتیجه می‌شود

$$\frac{۱ - P(۰(s_i) | \boldsymbol{Z}_{-i})}{P(۰(s_i) | \boldsymbol{Z}_{-i})} = \exp\{\alpha_i + \sum_{j=۱}^{n} \theta_{ij} Z(s_j)\}.$$

بنابراین

$$\frac{۱}{P(۰(s_i) | \{z(s_j) : j \neq i\})} = ۱ + \exp\{\alpha_i + \sum_{j=۱}^{n} \theta_{ij} z(s_j)\}.$$

پس

$$P(۰(s_i) | \{z(s_j) : j \neq i\}) = \frac{۱}{۱ + \exp\{\alpha_i + \sum_{j=۱}^{n} \theta_{ij} z(s_j)\}}$$

یا معادل آن

$$\begin{aligned} \text{logit}(\text{P}(0(\text{s}_\text{i}) | \boldsymbol{Z}_{-\text{i}})) &= \ln\left(\frac{۱ - P(۰(s_i)}{P(۰(s_i) | \boldsymbol{Z}_{-i}}\right) \\ &= \alpha_i + \sum_{j=۱}^{n} \theta_{ij} Z(s_j) \end{aligned}$$

نتیجه می‌شود. در نتیجه مدل اتولجستیک به‌صورت

$$P(Z(s_i)|\boldsymbol{Z}_{-i}) = \frac{\exp\{\alpha_i Z(s_i) + \sum_{j=1}^{n} \theta_{ij} Z(s_i) Z(s_j)\}}{1 + \exp\{\alpha_i Z(s_i) + \sum_{j=1}^{n} \theta_{ij} Z(s_i) Z(s_j)\}}$$

به‌دست می‌آید، که در آن α_i پارامترهای روند فضایی (تغییرات بزرگ مقیاس) و θ_{ij} پارامترهای وابستگی فضایی (تغییرات کوچک مقیاس) هستند.

مدل اتودوجمله‌ای

مدل اتودوجمله‌ای به‌صورت

$$\begin{aligned} &P\left(Z\left(s_i\right)|\{z\left(s_j\right): j \neq i\}\right) \\ &\equiv \binom{n_i}{z\left(s_i\right)} \pi_i\left(\{z\left(s_j\right): j \neq i\}\right)^{z(s_i)} \left(1 - \pi_i\left(\{z\left(s_j\right): j \neq i\}\right)\right)^{n_i - z(s_i)} \quad z\left(s_i\right) = 0, 1, \cdots, n \end{aligned} \tag{۱۲.۵}$$

تعریف می‌شود، که صورت نمایش نمایی آن به‌شکل

$$P\left(Z(s_i)|\boldsymbol{Z}_{-i}\right) = \exp^{Z(s_i)\mathrm{logit}(\pi_i) + \ln\binom{n_i}{z(s_i)} + n_i \ln(1-\pi_i)}$$

است که در آن $\pi = \pi(\boldsymbol{Z}_{-i})$، $L_i(\boldsymbol{Z}_{-i}) = \mathrm{logit}(\pi_i) = \ln\frac{\pi_i}{1-\pi_i}$ و $M_i\left(z(s_i)\right) = z(s_i)$. با توجه به تعریف داریم

$$\begin{aligned} Q(\boldsymbol{Z}) &= \ln\left(\frac{P(z(s_i)|z(s_j): j \neq i)}{P(\circ(s_i)|z(s_j): j \neq i)}\right) \\ &= \ln\left(\frac{\binom{n_i}{z(s_i)} \pi_i^{z(s_i)} (1-\pi_i)^{n_i - z(s_i)}}{(1-\pi_i)^{n_i}}\right) \\ &= \ln\left(\binom{n_i}{z(s_i)} \pi_i^{z(s_i)} (1-\pi_i)^{-z(s_i)}\right) \\ &= \ln\left(\binom{n_i}{z(s_i)} \left(\frac{\pi_i}{1-\pi_i}\right)^{z(s_i)}\right) \\ &= z(s_i)\{\ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln\binom{n_i}{z(s_i)} \\ &= \alpha_i z(s_i) + \sum_{i=1}^{n} \theta_{ij} z(s_i) z(s_j) + \ln\binom{n_i}{z(s_i)} \end{aligned}$$

$$Q(Z) = \sum_{i=1}^{n} \alpha_i z(s_i) + \sum \sum_{n \leq j \leq i \leq 1} \theta_{ij} z(s_i) z(s_j) + \sum_{i=1}^{n} \ln\left(\binom{n_i}{z(s_i)}\right).$$

بنابراین معادله

$$\mathrm{logit}(\pi_i) = \alpha_i + \sum_{j=1}^{n} \theta_{ij} z(s_j) \qquad i = 1, \cdots, n$$

که همان مدل اتولجستیک است، به‌دست می‌آید. با حل آن بر حسب π_i داریم

$$\pi_i\left(\{z(s_j): j\neq i\}\right)=\frac{\exp\{\alpha_i+\sum_{i=1}^{n}\theta_{ij}z(s_j)\}}{1+\exp\{\alpha_i+\sum_{i=1}^{n}\theta_{ij}z(s_j)\}} \qquad (۱۳.۵)$$

که در آن θ_{ij} ها همبستگی فضایی و α_i تاثیر موقعیت i ام است به‌طوری که می‌تواند به‌عنوان روند مدل یا تاثیر متغیرهای کمکی در نظر گرفته شود. از رابطه (۱۲.۵) می‌توان به‌عنوان پیش‌گو و رابطه (۱۳.۵) به‌عنوان برآوردگر در موقعیت i ام استفاده کرد.

مدل اتوپواسون

توزیع شرطی اتوپواسون با پذیره وابستگی فقط جفتی به‌صورت

$$P(Z(s_i)|\boldsymbol{Z}_{-i})=\frac{e^{-\lambda_i(\boldsymbol{Z}_{-i})}\{\lambda_i(\boldsymbol{Z}_{-i})\}}{Z(s_i)!}$$

است. برای این مدل می‌توان نوشت

$$\begin{aligned}Q(\boldsymbol{Z})-Q(\boldsymbol{Z}_i)&=\ln\left(\frac{P(z(s_i)|z(s_j):j\neq i)}{P(\circ(s_i)|z(s_j):j\neq i)}\right)\\&=\ln\frac{\frac{e^{-\lambda_i Z(s_i)}\lambda_i^{Z(s_i)}}{Z(s_i)!}}{e^{-\lambda_i Z(s_i)}}=\ln\frac{\lambda_i^{Z(s_i)}}{Z(s_i)!}\\&=Z(s_i)\ln(\lambda_i)-\ln(Z(s_i)!).\end{aligned}$$

با توجه به (۱۱.۵) خواهیم داشت

$$\begin{aligned}Q(\boldsymbol{Z})&=Z(s_i)\{\alpha_i+\sum_{i=1}^{n}\theta_{ij}Z(s_i)\}-\ln(Z(s_i)!\\&=\alpha_i Z(s_i)-\ln(Z(s_i)!)+\sum_{j\neq i}\theta_{ij}Z(s_i)Z(s_j).\end{aligned}$$

بنابراین

$$Q(\boldsymbol{Z})=\sum_{i=1}^{n}\alpha_i Z(s_i)+\sum\sum_{1\leq i\leq\ j\leq n}\theta_{ij}Z(s_i)Z(s_j)-\sum_{i=1}^{n}\ln(Z(s_i)!).$$

مدل اتودوجمله‌ای منفی

مدل فضایی دوجمله‌ای منفی با پارامترهای میانگین μ و بیش‌پراکندگی β، تحت عنوان اتودوجمله‌ای منفی، برای داده‌های شمارشی به‌صورت

$$P(Z(s_i)|\boldsymbol{Z}_{-i})=\frac{\Gamma\left(Z(s_i)+\mu_i\right)}{Z(s_i)!\Gamma\left(\mu_i\right)}\left(\frac{\beta_i\{\boldsymbol{Z}_{-i}\}}{\beta_i\{\boldsymbol{Z}_{-i}\}+1}\right)^{\mu_i}\left(\frac{1}{\beta_i\{\boldsymbol{Z}_{-i}\}+1}\right)^{Z(s_i)}\quad Z(s_i)=0,1,\cdots$$

تعریف می‌شود. در صورت وابستگی فقط جفتی بین متغیرها

$$\beta_i\{\boldsymbol{Z}_{-i}\}=e^{-\alpha_i-\sum_{j=1}^{n}\theta_{ij}Z(s_j)}-1$$

است.

در ادامه برای معرفی مدل اتوبتا_دوجمله‌ای، ابتدا نسخه بازپارامتری‌شده آن را معرفی می‌کنیم.

۶.۵ توزیع بتا_دوجمله‌ای بازپارامتری‌شده

فرض کنید

$$Z(s_i)|\pi(s_i) \sim bin(n(s_i), \pi(s_i))$$

به‌طوری که

$$\pi(s_i) \sim \mathrm{Beta(a(s_i), b(s_i))}.$$

بنابراین $Z(s_i)$ دارای توزیع کناری بتا_دوجمله‌ای با تکیه‌گاه $\{۰, \cdots, n_s\}$ و تابع احتمال

$$P(Z_i(s) = z(s_i)|a(s_i), b(s_i)) = \binom{n_s}{z(s_i)} \frac{B(a(s_i) + z(s_i), n_s + b(s_i) - z(s_i))}{B(a(s_i), b(s_i))}$$

می‌باشد. میانگین و واریانس $Z_i(s)$ به‌ترتیب برابر هستند با $n_s \frac{a(s_i)}{a(s_i)+b(s_i)}$ و

$$\frac{n_s a(s_i) b(s_i)(n_s + a(s_i) + b(s_i))}{(a(s_i) + b(s_i))^۲(۱ + a(s_i) + b(s_i))}.$$

حال میانگین و واریانس توزیع بتا_دوجمله‌ای را بازپارامتری می‌کنیم.

فراری و سریباری (۲۰۰۴) توزیع بتای بازپارامتری را برای مطالعه نسبت‌ها پیشنهاد کردند. آن‌ها پارامترهای توزیع بتا را به‌گونه‌ای بازپارامتری کردند که مدل رگرسیونی بر اساس میانگین متغیر پاسخ تعیین شود.

فرض کنید متغیر پاسخ Z_i دارای توزیع بتا با پارامترهای $a(s_i)$ و $b(s_i)$ باشد، که در آن $a(s_i), b(s_i) > ۰$. در این‌صورت، میانگین توزیع برابر $\mu(s_i) = \frac{a(s_i)}{a(s_i)+b(s_i)}$ خواهد بود. با قرار دادن $\gamma(s_i) = a(s_i) + b(s_i)$ پارامترهای اصلی به‌صورت $a(s_i) = \gamma(s_i)\mu(s_i)$ و

$$b(s_i) = \gamma(s_i)(۱ - \mu(s_i))$$

می‌توانند بازنویسی شوند. حال با توجه به تعاریف فوق تابع چگالی توزیع بتا را به‌صورت زیر بازنویسی می‌کنیم:

$$f_i(z_i|\mu(s_i), \gamma(s_i)) = \frac{\Gamma(\gamma(s_i))}{\Gamma(\mu(s_i)\gamma(s))\Gamma((۱ - \mu(s_i))\gamma(s))} z_i^{\mu(s_i)\gamma(s_i)-۱} (۱ - z_i)^{(۱-\mu(s_i))\gamma(s)-۱} I_z(۰, ۱)$$

$۰ < \mu(s_i) < ۱, \gamma(s) > ۰$

در این‌صورت

$$Z_i \sim \mathrm{Beta}(\mu(\mathrm{s_i})\gamma(\mathrm{s_i}), (1 - \mu(\mathrm{s_i}))\gamma(\mathrm{s_i}))$$

با میانگین μ_i و واریانس $\sigma^۲ = \frac{\mu(s_i)(۱-\mu(s_i))}{\gamma(s_i)+۱}$ است.

با استفاده از ویژگی‌های رگرسیون بتا و بازپارامتری کردن توزیع بتا_دوجمله‌ای با عبارات

$$\mu(s_i) = \frac{a(s_i)}{\gamma(s_i)}$$

$$\gamma(s_i) = a(s_i) + b(s_i)$$

تابع احتمال $Z(s_i)$ را می‌توان به‌صورت

$$Z(s_i) \sim beta - bin\left(n_s, \mu(s_i)\gamma(s_i), (۱ - \mu(s_i)\gamma(s_i)\right)$$

نوشت. میانگین و واریانس توزیع بتا_دوجمله‌ای بازپارامتری نیز به‌صورت

$$E(Z(s_i)) = n_s\mu(s_i)$$

و

$$Var(Z(s_i)) = n_s\mu(s_i)(۱ - \mu(s_i))\frac{\gamma(s_i) + n_s}{\gamma(s_i) + ۱}$$

به‌دست می‌آیند. لازم به ذکر است پارامتر بیش‌پراکندگی برای مدل بتا_دوجمله‌ای توسط

$$\frac{\gamma(s) + n_s}{\gamma(s) + ۱} \in (۱, n_s)$$

لحاظ می‌شود. همچنین زمانی که $n_s = ۱$ یا $\gamma(s) \to \infty$ واریانس کناری توزیع بتا_دوجمله‌ای به واریانس دوجمله‌ای نزدیک می‌شود.

۱.۶.۵ مدل اتوبتا_دوجمله‌ای

در این بخش به تعمیم مدل بتا_دوجمله‌ای به اتوبتا_دوجمله‌ای می‌پردازیم. مدل اتوبتا_دوجمله‌ای به‌صورت

$$\begin{aligned} &P\left(Z\left(s_i\right)|\{z\left(s_j\right) : j \neq i\}\right) \\ &= \binom{n_s}{z(s_i)} \frac{B(a_i(s) + z(s_i), n_s + b(s_i) - z_i(s))}{B(a(s_i), b(s_i))} \end{aligned}$$

تعریف می‌شود. با توجه به (۸.۵)

$$\begin{aligned} Q(\boldsymbol{Z}) &= \log\left(\frac{P(Z(s_i)|z(s_j) : j \neq i)}{P(\circ(s_i)|P(z(s_j) : j \neq i}\right) \\ &= \log\left(\binom{n_s}{z(s_i)} \frac{B(a(s_i) + z(s_i), n_s + b(s_i) - z(s_i))}{B(a(s_i), b(s_i) + n_s)}\right) \end{aligned}$$

با لگاریتم گرفتن، داریم

$$\begin{aligned} Q(\boldsymbol{Z}) &= \log B(a(s_i) + z(s_i), n_s + b(s_i) - z(s_i)) + \log\binom{n_s}{z(s_i)} - \log B(a(s_i), b(s_i)) \\ &= lgamma(a(s_i) + z(s_i)) + lgamma(n_s + b(s_i) - z(s_i)) + \log\binom{n_s}{z(s_i) + n_s} \\ &- lgamma(a(s_i) + b(s_i) + n) - \log B(a(s_i), b(s_i)) \end{aligned}$$

که در آن منظور از $lgamma$، لگاریتم تابع گاما است. در نهایت تابع احتمال مدل اتوبتا- دوجمله‌ای به صورت زیر حاصل می‌شود:

$$P\left(Z\left(s_i\right)|\{z\left(s_j\right):j\neq i\}\right)=\binom{n_i}{z(s_i)}\frac{B(\mu(s_i)\gamma(s_i)+z_i,n_s+۱-\mu(s_i)\gamma(s_i)-z_i)}{B(\mu(s_i)\gamma(s_i),(۱-\mu(s_i)\gamma(s_i))}$$

۷.۵ استنباط بیزی تقریبی

گاهی در استنباط بیزی توزیع پسین مدل‌ها شکل بسته‌ای ندارند. در چنین مواردی معمولا از الگوریتم‌های نمونه‌گیری مونت کارلویی برای تقریب توزیع استفاده می‌شود. وجود همبستگی در میدان تصادفی پنهان معمولا موجب کاهش کارایی این الگوریتم‌ها، افزایش زمان محاسبات و همگرایی کند الگوریتم می‌شود. برای مرتفع کردن این مشکل، رو و همکاران (۲۰۰۹) روش تقریب لاپلاس آشیانه‌ای جمع‌بسته را معرفی کردند، که در استنباط مدل‌های گاوسی پنهان جانشین مناسبی برای الگوریتم‌های MCMC است. در روش INLA انتگرال‌گیری عددی و تقریب لاپلاس به‌طریقی کارا ترکیب می‌شوند به‌طوری که محاسباتی سریع و با دقت قابل قبول جایگزین شبیه‌سازی‌های سنگین MCMC می‌شوند.

۱.۷.۵ مدل‌های گاوسی پنهان

گام اول در تعریف یک مدل گاوسی پنهان، همراه با چارچوب بیزی، تشخیص توزیع داده‌های مشاهده‌شده $\boldsymbol{Z}=(Z_۱,\cdots,Z_n)$ است. پرکاربردترین روش تشخیص توزیع Z_i با پارامتر μ_i (معمولا میانگین $E(Z_i)$)، تابع پیش‌گوی جمعی η_i از طریق تابع پیوند $g(\cdot)$، مانند $g(\mu_i)=\eta_i$ است. پیش‌گوی خطی جمعی η_i به‌صورت زیر تعریف می‌شود:

$$\eta_i=\alpha+\sum_{m=۱}^{M}\beta_m v_{mi}+\sum_{l=۱}^{L}f_l(t_{li})+\epsilon_i \qquad i=۱,\cdots,n \tag{۱۴.۵}$$

که در آن α عرض از مبدا است، ضرایب $\boldsymbol{\beta}=\{\beta_۱,\cdots,\beta_M\}$ اثرات ثابت خطی متغیرهای تبیینی $\boldsymbol{v}=(v_۱,\cdots,v_M)$ روی متغیر پاسخ است و $\boldsymbol{f}=\{f_۱(\cdot),\cdots,f_L(\cdot)\}$ مجموعه‌ای از توابع است که در جملات بردار متغیرهای تصادفی و پنهان $\boldsymbol{t}=(t_۱,\cdots,t_L)$ تعریف شده‌اند.

همه مولفه‌های پنهان (غیرقابل مشاهده) مورد نظر برای استنباط در مجموعه پارامترها با نام $\boldsymbol{x}$ که به‌صورت $\boldsymbol{x}=\{\alpha,\boldsymbol{\beta},\boldsymbol{f}\}$ تعریف شده، جمع‌آوری می‌شود. به‌علاوه بردار m بعدی ابرپارامترها به‌صورت $\boldsymbol{\theta}=\{\theta_۱,\cdots,\theta_m\}$ مشخص می‌شود.

با فرض استقلال شرطی، توزیع شرطی n مشاهده به‌وسیله تابع درستنمایی زیر به‌دست می‌آید:

$$\pi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})=\prod_{i=۱}^{n}\pi(z_i|x_i,\boldsymbol{\theta}). \tag{۱۵.۵}$$

فرض می‌کنیم پیشین $\boldsymbol{x}$ دارای توزیع نرمال چندمتغیره با بردار میانگین ۰ و ماتریس دقت

$$Q(\boldsymbol{\theta}) = \Sigma_{\boldsymbol{\theta}}^{-1}$$

یعنی $\boldsymbol{x} \sim N(0, Q(\boldsymbol{\theta}))$ با تابع چگالی

$$\pi(\boldsymbol{x}|\boldsymbol{\theta}) = (2\pi)^{-\frac{n}{2}} |Q(\boldsymbol{\theta})|^{\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{x}' Q(\boldsymbol{\theta})\boldsymbol{x}) \qquad (۱۶.۵)$$

باشد. در حالتی که $Q(\boldsymbol{\theta})$ یک ماتریس دقت تنک[۱۴] باشد، مولفه‌های میدان گاوسی پنهان $\boldsymbol{x}$، به‌طور شرطی، مستقل هستند. این ویژگی با عنوان میدان تصادفی مارکوف گاوسی[۱۵] (GMRF) شناخته می‌شود (رو و هلد، ۲۰۰۵). توزیع توام $\boldsymbol{x}$ و $\boldsymbol{\theta}$ به‌وسیله ضرب درستنمایی (۱۵.۵) و چگالی (۱۶.۵) و توزیع پیشین ابرپارامترهای $\pi(\boldsymbol{\theta})$ به دست می‌آید:

$$\begin{aligned}
\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{z}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\theta)\pi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) \qquad (۱۷.۵)\\
&\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} \pi(z_i|x_i, \boldsymbol{\theta})\\
&\propto \pi(\boldsymbol{\theta}) \times |Q(\boldsymbol{\theta})|^{-\frac{n}{2}} \exp(-\frac{1}{2}\boldsymbol{x}' Q(\boldsymbol{\theta})\boldsymbol{x}) \times \prod_{i=1}^{n} \exp(\log(\pi(z_i|x_i, \boldsymbol{\theta}))\\
&\propto \pi(\boldsymbol{\theta})|Q(\boldsymbol{\theta})|^{\frac{n}{2}} exp\left(-\frac{1}{2}\boldsymbol{x}' Q(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i=1}^{n} \log(\pi(z_i|x_i, \boldsymbol{\theta}))\right).
\end{aligned}$$

۲.۷.۵ تقریب لاپلاس آشیانه‌ای جمع‌بسته

برای تحلیل بیزی مدل‌های گاوسی پنهان به روش INLA، لازم است توزیع‌های پسینی کناری متغیرهای پنهان و ابرپارامترها به‌صورت

$$\pi(x_i|\boldsymbol{z}) = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{z})\pi(\boldsymbol{\theta}|\boldsymbol{z})d\boldsymbol{\theta} \qquad i = 1, \cdots, n_d \qquad (۱۸.۵)$$

$$\pi(\theta_j|\boldsymbol{z}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{z})d\boldsymbol{\theta}_{-j} \qquad j = 1, \cdots, m \qquad (۱۹.۵)$$

محاسبه شوند، که در آن $\boldsymbol{\theta}_{-j}$ بردار حاصل از حذف درایه j ام $\boldsymbol{\theta}$ است و n_d بعد میدان تصادفی $\boldsymbol{x}$ است. روش INLA تقریب‌هایی برای چگالی‌های پسین (۱۸.۵) و (۱۹.۵) به‌صورت

$$\tilde{\pi}(x_i|\boldsymbol{z}) = \int \tilde{\pi}(x_i, \boldsymbol{z}, \boldsymbol{\theta})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{z})d\boldsymbol{\theta}, \qquad i = 1, \cdots, n$$

$$\tilde{\pi}(\theta_j|\boldsymbol{z}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{z})d\boldsymbol{\theta}_{-j} \qquad j = 1, \cdots, m$$

[۱۴] Sparse

[۱۵] Gaussian Markov random field

ارایه می‌کند. روش INLA برای محاسبه تقریبی این توزیع‌ها، با استفاده از تبدیل‌های لاپلاس و انتگرال‌گیری عددی، محاسباتی سریع و تقریبی دقیق را جایگزین شبیه‌سازی‌های سنگین الگوریتم‌های MCMC می‌کند. تقریب‌های $\tilde{\pi}(x_i|z)$ و $\tilde{\pi}(\theta_i|z)$ در سه گام به شرح زیر محاسبه می‌شوند:

۱. توزیع پسینی کناری $\pi(\boldsymbol{\theta}|z)$ به‌صورت

$$\tilde{\pi}(\boldsymbol{\theta}|z) = \frac{\pi(\boldsymbol{x},\boldsymbol{\theta},z)}{\tilde{\pi}(\boldsymbol{x}|\boldsymbol{\theta},z)}|_{\boldsymbol{x}=\boldsymbol{x}^{*(\boldsymbol{\theta})}}$$

تقریب زده می‌شود، که در آن $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},z)$ تقریب گاوسی (رو و هلد، ۲۰۰۵) برای توزیع شرطی کامل $\boldsymbol{x}$ و $\boldsymbol{x}^*$ مد توزیع شرطی کامل $\boldsymbol{x}$ است.

۲. محاسبه $\pi(x_i|\boldsymbol{\theta},z)$ که توسط رو همکاران (۲۰۰۹) سه تقریب گاوسی، لاپلاس و لاپلاس ساده‌شده[۱۶] را برای تقریب پیشنهاد دادند. ساده‌ترین روش، تقریب گاوسی و دقیق‌ترین روش، تقریب لاپلاس است. در نهایت، روش تقریب لاپلاس‌ساده شده با کاهش جزئی در دقت از لحاظ محاسباتی آسان‌تر از تقریب لاپلاس است.

۳. گام سوم ترکیب دو گام پیشین و استفاده از روش انتگرال‌گیری عددی برای رسیدن به هدف است. توزیع پسین به‌صورت

$$\tilde{\pi}(x_i|\boldsymbol{z}) = \sum_{k=۱}^{K} \tilde{\pi}(x_i|\boldsymbol{\theta}_k,\boldsymbol{z}) \times \tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{z}) \times \triangle_k$$

به‌دست می‌آید که در آن K تعداد $\boldsymbol{\theta}$ های انتخاب‌شده از تکیه‌گاه و $\triangle_k$ ها وزن‌های این نقاط هستند.

در ادامه از تقریب INLA برای تحلیل داده‌های واقعی سرطان معده بر اساس مدل اتوبتا- دوجمله‌ای استفاده می‌کنیم.

۸.۵ تحلیل داده‌های سرطان معده در استان گیلان

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

[۱۶] Simplifield laplace

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

شرایط جوی و محیطی در هر منطقه زمینه را برای بروز و شیوع برخی بیماری‌ها مساعد می‌کند. سرطان نیز از جمله بیماری‌های کشنده و نگران‌کننده امروزی است که انسان با آن دست و پنجه نرم می‌کند. عوامل عمده تاثیرگذار بر این بیماری را عوامل محیطی و ژنتیکی می‌دانند. میزان بروز انواع آن در نواحی جغرافیایی مختلف، متفاوت است. در سال‌های اخیر در کشور ما این میزان افزایش پیدا کرده است. به‌خصوص در استان گیلان، سرطان معده شیوع فراوان دارد و سالانه جان صدها نفر را به خطر می‌اندازد. استان گیلان از نظر فراوانی سرطان معده مقام اول را در کل کشور دارد (رمضانی، ۱۳۸۷).

نسبت بالایی از سرطان‌ها به‌طور مستقیم با عوامل محیط زیست ارتباط دارند و لازم است که عوامل موثر در این مورد تعیین و مشخص گردد (ویلکاکسون و تنسر، ۱۹۹۹). از هر ۹ مرگ در جهان یک مورد مربوط به سرطان است. سرطان بعد از بیماری‌های قلبی عروقی دومین علت مرگ در جوامع است و سومین علت مرگ بعد از بیماری‌های قلبی عروقی و تصادفات در کشور ما است. با توجه به رابطه بین بروز انواع سرطان و شرایط جغرافیایی منطقه‌ای، به همین علت تفاوت‌های آشکاری را در میزان شیوع و فراوانی هر کدام از سرطان‌ها در مناطق مختلف شاهد هستیم به‌طوری که در عرض بالای ۳۲ درجه جغرافیایی شیوع بیماری سرطان مری بالا است و تاثیر زمینه‌های محیطی و شرایط جغرافیایی و نوع اشتغال را با این بیماری در ارتباط دانسته‌اند.

سرطان معده عبارت است از رشد بدون کنترل سلول‌های بدخیم در معده که در آن بیش‌تر افراد تا مراحل پیشرفته بیماری، علامتی ندارند. سلول‌های سرطانی معمولا به مرور رشد می‌کنند و تبدیل به تومور می‌شوند. این سرطان در استان گیلان بسیار شایع است، به‌طوری که به گفته یکی از مسئولان مرکز تحقیقات گوارش و کبد بیمارستان رازی رشت، هر دو روز یک مورد سرطان جدید معده تشخیص داده می‌شود و از هر سه نفری که بر اثر ابتلا به سرطان فوت می‌شوند یک نفر مبتلا به سرطان معده است.

در مورد روند بهبود برآوردهای پهنه‌بندی بیماری می‌توان به تحقیقات انجام‌شده اشاراتی داشت. ابتدایی‌ترین نقشه را می‌توان به جان اسنو نسبت داد که در سال ۱۸۵۴ و به دنبال همه‌گیری وبا در شهر لندن تهیه کرد. کلایتون و کالدور (۱۹۸۷) مدل‌های سلسله‌مراتبی و استنباط بیز تجربی مرتبط با آن را برای میزان‌های مرگ و میر استانداردشده در حالی که همبستگی فضایی بین مشاهدات در نواحی همسایه نیز لحاظ می‌شود، مطرح کردند.

در تحلیل داده‌های بیماری که غالبا گسسته و شمارشی هستند، می‌توان کایسر و همکارانش

(۱۹۹۷) را نام برد که برای تحلیل فضایی داده‌های دارای توزیع پواسون، مدل توزیع پواسون فضایی را معرفی نمودند. الیور و همکارانش (۱۹۹۸) هم‌کریگینگ دوجمله‌ای را برای تهیه پهنه‌بندی خطر سرطان کودکان در غرب انگلستان به‌کار بردند. مونستیز و همکارانش (۲۰۰۴ـ ۲۰۰۶) استفاده از مدل کریگینگ پواسون را برای داده‌های شمارشی همبسته فضایی پیشنهاد دادند. آن‌ها در مدل پیشنهادی خود از یک طرح وزن‌دار به منظور اختصاص دادن وزن بزرگ‌تر به مشاهدات بزرگ‌تر استفاده کردند.

۱.۸.۵ نرخ سرطان معده در استان گیلان

در این بخش نرخ سرطان معده در استان گیلان را با روش INLA، تحلیل می‌کنیم. در این راستا دو هدف را دنبال می‌کنیم: ۱) تهیه یک نقشه پهنه‌بندی از نرخ سرطان معده در استان گیلان و ۲) بررسی تاثیر متغیرهای تبیینی.

مجموعه داده مورد استفاده شامل کلیه داده‌های ثبت‌شده بیماران سرطانی در مرکز آموزشی درمانی رازی رشت از اول خرداد ۱۳۹۱ تا پایان اسفند ۱۳۹۵ می‌باشد. کل بیماران مبتلا به سرطان معده در این مجموعه داده‌ها ۶۴۴ مورد هستند که ۶۲ درصد آن‌ها مرد و ۳۸ درصد زن هستند. در سال ۹۱ کم‌ترین مراجعه صورت گرفته و بیش‌ترین مراجعه نیز در سال ۹۵ است که دلیل آن شاید به‌خاطر ثبت دقیق‌تر اطلاعات برای سال ۹۵ نسبت به سال‌های دیگر باشد. متغیر پاسخ تعداد افراد بیمار در شهرهای مختلف استان گیلان و متغیرهای تبیینی شامل سرانه فضای سبز و مدت اقامت بیماران (تعداد روزهای بستری) هستند. با توجه به این‌که متغیر پاسخ شمارشی و گسسته است و مشاهدات به موقعیت جغرافیایی که در آن قرار دارند وابسته هستند، بنابراین داده‌ها ماهیت فضایی و مشبکه‌ای دارند.

برای تحلیل داده‌های سرطانی با استفاده از مدل بتاـ دوجمله‌ای و دوجمله‌ای، پیش‌گوی خطی

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \qquad i = 1, \cdots, n$$

را در نظر گرفتیم، که در آن $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ بردار ضرایب رگرسیونی، x_1 نرخ بستری بیمار در بیمارستان و x_2 نرخ سرانه فضای سبز است. همچنین $\boldsymbol{u} = (u_1, \cdots, u_{16})'$ اثر تصادفی فضایی CAR برای لحاظ کردن همبستگی فضایی بین شانزده شهرستان استان گیلان، وارد مدل شده است. در مدل CAR در نظر گرفته‌شده برای تحلیل این داده‌ها از یک مدل ذاتی[۱۷] استفاده کردیم که در آن فرض می‌شود $\rho = 1$ است و تنها ابرپارامتر این مدل پارامتر واریانس σ^2 است. البته در روش INLA به‌جای کار کردن با σ^2 و تعریف توزیع پیشین برای آن، با عکس آن که پارامتر دقت محسوب می‌شود مدل‌بندی (و تعریف توزیع پیشین) انجام می‌شود.

برای تکمیل مدل‌بندی بیزی، باید توزیع‌های پیشین پارامترهای β و $\boldsymbol{\theta}$ تعیین شوند. برای این کار، در هر دو مدل دوجمله‌ای و بتاـ دوجمله‌ای، از توزیع‌های پیشین استاندارد روش

[۱۷] Intrinsic

INLA استفاده کردیم؛ برای پارامترهای β توزیع‌های پیشین نرمال با میانگین صفر و واریانس بزرگ ۱۰۰۰ تعیین شدند و برای پارامتر دقت مدل CAR توزیع پیشین گاما با پارامترهای شکل ۱ و مقیاس ۰/۰۰۱ انتخاب شد.

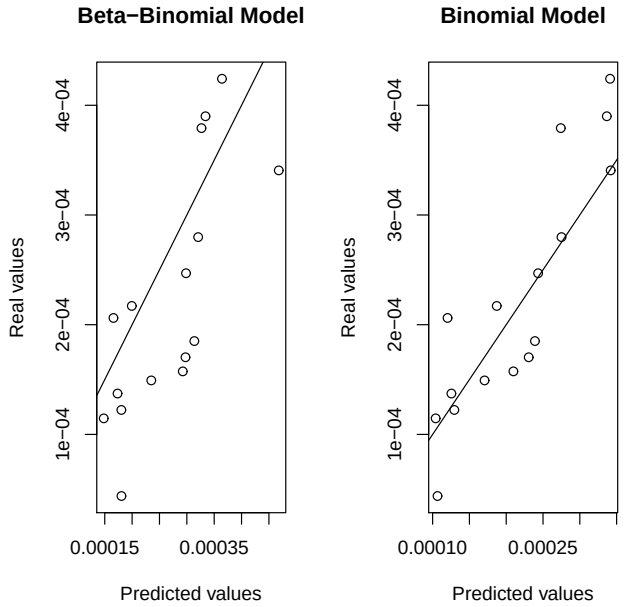
هر دو مدل اتودوجمله‌ای و اتوبتا_دوجمله‌ای با روش INLA و به کمک بسته R-INLA در نرم‌افزار R، بر روی داده‌ها برازش داده شدند. نتایج برازش هر دو مدل در جدول ۲.۵ گزارش شده‌اند. همچنین شکل ۱.۵ نمودارهای پراکنش مقادیر واقعی نسبت بیماران مبتلا به سرطان معده را در مقابل مقادیر برازش‌شده از هر دو مدل، نمایش می‌دهد. با توجه به نتایج جدول و نمودارهای پراکنش مذکور، می‌توان گفت

- هر دو متغیر تبیینی معنی‌دار هستند.
- برآوردهای پارامترها از نظر اندازه اثر به هم نزدیک هستند ولی دقت برآوردها (بر حسب انحراف معیارهای برآوردشده و فاصله‌های اعتبار ۸۰ درصد) اختلاف قابل ملاحظه با هم دارند. با توجه به عدم انعطاف لازم در مساله مدل‌بندی بیش‌پراکنش یا کم‌پراکنش موجود در داده‌ها توسط مدل دوجمله‌ای، انحراف معیارهای کوچک در این مدل می‌تواند نشانه‌ای از کم‌برآورد کردن دقت واقعی برآوردگرها باشد. از این منظر، مدل بتا_دوجمله‌ای توانایی ارایه توصیف بهتری از مدل را دارد. البته مساله نیاز به بررسی بیش‌تر دارد.
- طول مدت اقامت بیماران بر نرخ سرطان معده در شهرستان‌های استان گیلان، تاثیر مستقیم (مثبت) دارد. با توجه به تفسیر پارامترها در مدل‌های لجیت، برآورد ۰/۶۲ برای اثر طول مدت بستری در مدل بتا_دوجمله‌ای، به این معنی است که اگر طول روزهای بستری بیماران ساکن در شهرستانی یک روز افزایش یابد، بخت ابتلا به سرطان معده در آن شهرستان ۸۶ درصد افزایش می‌یابد.
- در مقابل، نرخ سرانه فضای سبز در هر دو مدل تاثیر منفی دارد؛ به این معنی که اگر نرخ سرانه فضای سبز در شهرستانی یک واحد افزایش یابد، بخت ابتلا به سرطان معده در آن شهرستان ۴۶ درصد کاهش خواهد داشت.
- با توجه به نمودارهای پراکنش مقادیر واقعی در مقابل برازش‌شده نسبت ابتلا به سرطان، به سختی می‌توان برازش بهتر نسبی مدل دوجمله‌ای را در مقابل مدل بتا_دوجمله‌ای مشاهده کرد. البته این برازش بهتر می‌تواند منبعی برای بیش‌برازش مدل دوجمله‌ای باشد که در بحث پیش‌گویی فضایی منتج به عملکرد ضعیف خواهد شد. برای بررسی این موضوع مقادیر پاسخ دو شهرستان آستانه اشرفیه و لنگرود را کنار گذاشتیم و با بقیه داده‌ها دو مدل را برازش دادیم. سپس از روی مدل‌های برازش‌شده این دو مقدار را پیش‌گویی کردیم. مقادیر میانگین توان دوم خطای پیش‌گویی MSPE برای دو مدل دوجمله‌ای و بتا_دوجمله‌ای به ترتیب برابر ۰/۰۱۸ و ۰/۰۰۴ به‌دست آمدند. در واقع

مدل بتا–دوجمله‌ای بیش از چهار برابر مدل دوجمله‌ای در پیش‌گویی تواناتر و دقیق‌تر عمل کرده است. همان‌طور که اشاره کردیم، دلیل این نتیجه بیش‌برازش بودن مدل دوجمله‌ای در مقابل مدل بتا–دوجمله‌ای است.

جدول ۲.۵: نتایج برازش دو مدل بتا–دوجمله‌ای و دوجمله‌ای برای داده‌های سرطان معده

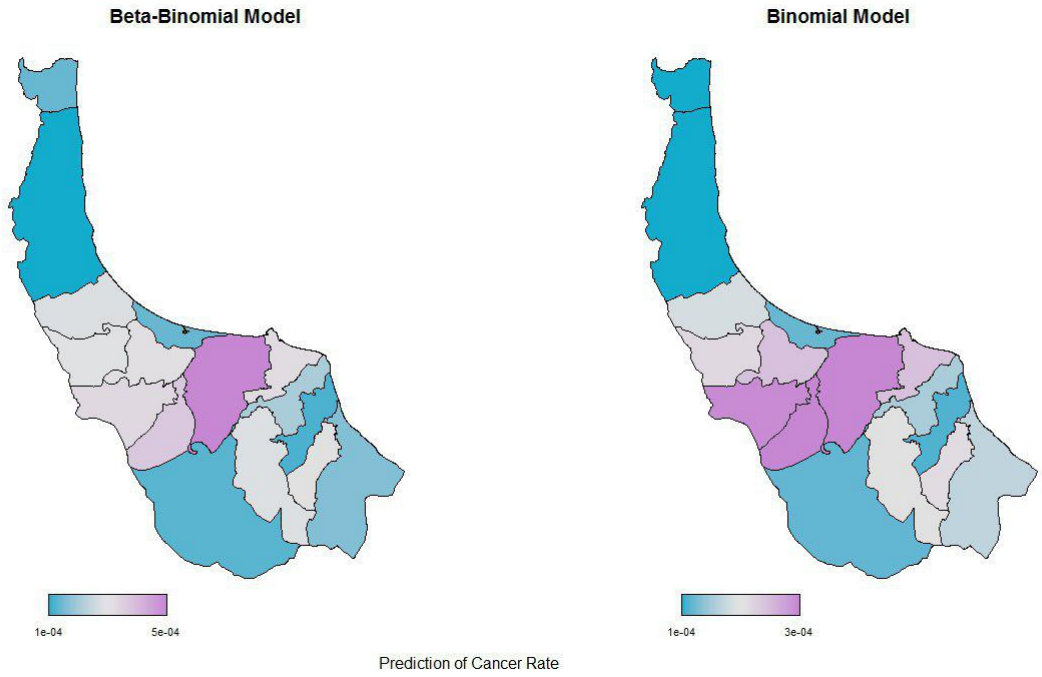
	اثر	برآورد اثر	انحراف معیار	ناحیه اعتبار۸۰٪
مدل بتا–جمله‌ای	ضریب ثابت	−۸/۳۷	۰/۲۳	(−۸/۷۰, −۷/۹۴)
	نرخ مدت اقامت	۰/۶۲	۰/۳۸	(۰/۱۳, ۱/۱۰)
	نرخ فضای سبز	−۰/۶۱	۰/۴۰	(−۱/۱۳, −۰/۱۰)
مدل دوجمله‌ای	ضریب ثابت	−۸/۵۴	۰/۱۷	(−۸/۶۹, −۸/۴۱)
	نرخ مدت اقامت	۰/۷۳	۰/۱۸	(۰/۵۲, ۰/۹۵)
	نرخ فضای سبز	−۰/۶۹	۰/۱۸	(−۰/۹۲, −۰/۴۶)



شکل ۱.۵: نمودار پراکنش مقادیر برازش‌شده در مقابل مقادیر واقعی نرخ سرطان معده در دو مدل دوجمله‌ای (سمت راست) و بتا–دوجمله‌ای (سمت چپ)

نقشه‌های پهنه‌بندی حاصل از دو مدل برای پیش‌گویی نرخ سرطان معده در استان گیلان در شکل ۲.۵ نشان داده شده‌اند. همان‌طور که از نقشه‌های پهنه‌بندی مشخص است، مقادیر

پیش‌گویی نرخ ابتلا به سرطان در اغلب شهرستان‌های استان گیلان توسط هر دو مدل مشابه هستند؛ در شهرستان‌هایی که اختلاف وجود دارد، مدل دوجمله‌ای نرخ بزرگ‌تری را نسبت به مدل بتا_دوجمله‌ای پیش‌گویی کرده است. ریشه ممکن این مساله را نیز در بیش‌برازش بودن مدل دوجمله‌ای مطرح کردیم. بنابراین، با توجه به توانایی بیش‌تر مدل بتا_دوجمله‌ای در پیش‌گویی، بر استفاده از نتایج مبتنی بر مدل بتا_دوجمله‌ای هم در توصیف عوامل موثر بر افزایش یا کاهش نرخ سرطان و هم پیش‌گویی، تاکید می‌کنیم.



شکل ۲.۵: نقشه‌های پهنه‌بندی نرخ سرطان معده برای مدل‌های دوجمله‌ای (سمت راست) و بتا_دوجمله‌ای (سمت چپ)

۹.۵ نتیجه‌گیری و آینده تحقیق

در این پایان‌نامه برای رفع نقص‌های مدل دوجمله‌ای، به معرفی و بررسی عملکرد مدل کریگینگ بتا_دوجمله‌ای به‌عنوان جانشینی برای مدل‌بندی پاسخ‌های نرخ همبسته فضایی پرداختیم. با استفاده از یک مطالعه شبیه‌سازی، دقت این مدل در برآورد پارامترها و پیش‌گویی فضایی پاسخ‌های نرخ را نسبت به مدل هم‌کریگینگ دوجمله‌ای، به‌ویژه برای اندازه‌های نمونه نسبتا کوچک، نشان دادیم. مدل پیشنهادی می‌تواند جانشینی ساده برای مدل‌های پیچیده‌تر فضایی، مانند مدل‌های آمیخته خطی تعمیم‌یافته و مدل‌های سلسله‌مراتبی بیزی برای اندازه نمونه کوچک باشد.

با توجه به مطالبی که در این پایان‌نامه به آن‌ها پرداخته شد، می‌توان به موضوعات زیر

به‌عنوان آینده تحقیق اشاره کرد:

- کریگینگ بتا ـ دوجمله‌ای می‌تواند برای مدل‌بندی داده‌های دودویی به کار گرفته شود. مدلی که به‌طور معمول برای مدل‌سازی داده‌های دودویی استفاده می‌شود، کریگینگ نشانگر است. در کریگینگ نشانگر فرض می‌شود که متغیر تصادفی دودویی، نتیجه یک متغیر آستانه‌ای است نه یک مشاهده دودویی تصادفی. کریگینگ بتا ـ دوجمله‌ای می‌تواند برای تشخیص تغییرات اضافی فرآیند دودویی و روند زیربنایی فضایی مفید باشد. همچنین با توجه به عملکرد مناسب مدل بتا ـ دوجمله‌ای در اندازه نمونه کوچک، ممکن است روش مناسبی برای مدل‌سازی داده‌های دودویی باشد.
- بررسی کریگینگ بیزی مدل بتا ـ دوجمله‌ای فضایی.
- تعمیم مدل کریگینگ بتا ـ دوجمله‌ای به مدل کریگینگ بتا ـ دوجمله‌ای تعمیم‌یافته.

مراجع

[۱] رمضانی ب. و حنیفی ا، (۱۳۸۷) ’’شناخت پراکندگی جغرافیایی شیوع سرطان معده در استان گیلان’’، **علوم و تکنولوژی محیط زیست**، شماره دو، دوره سیزدهم.

[۲] میرزاجانی م، (۱۳۹۵)، پایان نامه ارشد: ’’استنباط بیزی تقریبی مدل‌های رگرسیونی پویا با استفاده از تقریب لاپلاس آشیانه‌ای جمع بسته’’، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود.

[۳] محمدزاده م، (۱۳۹۱)، ’’آمار فضایی و کاربردهای آن’’، چاپ اول، **مرکز نشر آثار علمی دانشگاه تربیت مدرس** ، تهران.

[۴] حسن زاده ف و کاظمی ا، (۱۳۹۰)، ’’مقایسه‌ی انواع مدل‌های رگرسیون در تحلیل داده‌های شمارشی’’، **مجله بررسی‌های رسمی ایران** ، شماره ۱، سال ۲۲، ۳۲-۱۷.

[5] Ansley, C. F. (1979), ”An Algorithm for Exact Likelihood of a Mixed Autoregressive Moving Avrage Process”, **Biometrika**, 66, 59-65.

[6] Bartlett, M. S. (1971), ” Phisical Nearest-Neighbour Models and Non-Liner Time-Series”, **Jornal of Applied Probablity**, 8, 222, 232.

[7] Besag, J. (1974), ” Spatial Interaction and Statistical Analysis of Lattice System (with Discustion)”, **Jornal of the Royal Statistical Society**, B, 36(2), 192-236.

[8] Besag, J. and Moran, P. A. P. (1975), ” On the Estimation and Testing of Spatial Intraction Gaussian Lattice Process”, **Biometrika**, 62, 555-262.

[9] Billheimer, D., Cardoso, T., Freeman, E., Guttorp, P., Ko, H. W., and Silkey, M. (1997) ”Natural Variability of Benthic Species Composition in the Delaware Bay”, **Environmental and Ecological Statistics**, 4(2), 95-115.

[10] Branscum, A. J., Johnson W. O. and Thurmond, M. (2007). ” Bayesian Beta Regression: Applications to Houshold Expenditure Data and Genetic Distance Between Food and Mouth Dieseas Viruses”, **Australian and New Zealand Journal of Statistics**, 49, 287-301.

[11] Breslow, N. E. and Clayton, D. G. (1993), ”Approximate Inference in Generalized Linear Mixed Models ” **Journal of the American Statistical Association**, 88(421), 9-25.

[12] Kaiser, M.S. and Cressie, N. (1997), ”Modeling Poisson variables with positive spatial dependence”, **Statistics and Probability Letters**, 35(4), pp.423-432.

[13] Carlin, B. P. and Benarjee, S. (2003), ”Hierarchical Multivariate CAR Models for Spatio-Temporally Correlated Data”, **In Baysian Statistics 7** (eds. J. M. Bernardo et.al), 45-63, Oxford University Press.

[14] Chen, F. (2002), ” Bayesian Modeling of Multivariate Spatial Binary Data With Application to the Distribution of Plant Spacies” , PHD Thesis, Department of Statistics, Florida State University.

[15] Clyton, D. and Kaldor, J. M. (1987), ”Empirical Bayes Estimates of Age-Standardized Relative Risk for Use in Disease Mapping ” , **Biometrics**, 43, 671-681.

[16] Noel Cressie. (1985), ”Fitting variogram models by weighted least squares”, **Journal of the International Association for Mathematical Geology**,17 (5), 563-586.

[17] Cressie, N. (1990), ”The Origins of Kriging”, **Elsevier**, 22, 239–252.

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

[19] Cressie, N. (1988), ” Spatial Prediction and Ordinary Kriging”, **Mathematical Geology**, 20, 405–421.

[20] David, M. (1977), ” Geostatistical ore Reserve Estimation Elsevier”, **Elsevier**, Amsterdam.

[21] Dent, W. and Min, A. (1978), ” A Monte Carlo Study of Autoregressive Moving Avrage Processes”, **Jornal of Econometriecs**,7, 23-55.

[22] Diggle, P. J., Moyeed, R. A. and Tawn, J. A. (1998), ” Model-Based Geostatistics”, **Journal of the Royal Statistical Society**, 47(3), 299-350.

[23] Gandin, L. S. (1963),” Objective Analysis of Meteorological Fields”, **Gidrometeor, Leningrad**.

[24] Gelfand, A. E, and Vounatsou, P. (2003), ”Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis”, **Biostatistics**, 1, 11-15.

[25] Goovaerts, P. (1998), " Ordinary Cokriging Revisited ", **Mathematical Geology**, 30, 21–42, 1998.

[26] Gotway, A.C ., Carol, A. and Walter, W. S. (1997),"A Generalized Linear Model Approach to Spatial Data Analysis and Prediction" **Journal of Agricultural, Biological, and Environmental Statistics**, 2(2), 157–178.

[27] Green, J. D. (1970), "Personal Media Probabilities", **Journal of Advertising Research**, 10, 12-18.

[28] Griffiths, D. A. (1973), "Maximum Likelihood Estimation for the Beta-Binomial Distributions and an Application to the household Distribution of the Total Number of Cases of a Disease", **Biometrics**, 29, 637–648.

[29] Hainning, R. P. (1978), " The Moving Avrage for Spatial Intraction" **Tranctions, Institute of British Geographers**, N. S. 3, 202-225.

[30] Huynth, H. (1979), "Statistical Inference for Two Reliability Indices in Mastery Testing Based on the Beta-Binomial Model", **Journal of Educational Statistics**, 4, 231-246.

[31] Jakson, M. C (2003), "Spatial Data Analysis for Discrete Data on a Lattice", Thesis of PHD, University of Mariland, Department of Mathematics.

[32] Journel, A. G. and Huijbregts, C. J. (1978), "Mining Geostatistics", **Academic Press**, London.

[33] Journe, A. G. and Rossi, M. E. (1989), "When do we Need a Trend Model in Kriging?", **Mathematical Geology**, 21, 715–739.

[34] Kupper, L. L, and Haseman, J. K. (1978), "The use of a correlated binomial model for the analysis of certain toxicological experiments", **Biometrics** , 69-76.

[35] Kupper, L. L., Portier, C., Hogan, M. D., and Yamamoto, E. (1986), "The Impact of Litter Effects on Dose-Response Modeling in Teratology", **Biometrics**, 42, 85-98.

[36] Lajaunie, C. (1991), "Local Risk Estimation for a Rare Noncontagious Disease Based on Observed Frequencies", **Note N-36/91/G, Centre de Géostatisque, Ecole des Mines de Paris**.

[37] Lawson, A. B., Biggeri, A. B., Boehning, D., Lesaffre, E., Viel, J. F., Clark, A., and Divino, F. (2001)." Disease Mapping Models: an Empirical Evaluation" **Statistics in Medicine**, 2217-2241.

[38] Lesauskiene, E. and Ducinskas, k. (2003), ” Universal Kriging for Spatio-Temporal Data” **Mathematical Modelling and Analysis**, 8(4), 283–290, 2003.

[39] Lord, F. M. (1965), ”A Strong True-Score Theory, with Applications”, **Psychometrika**, 30, 234-270.

[40] Mardia, K. V. (1988), ” Multi-Dimential Multivariate Gaussian Markov Random Fields with A pplications to Image Processing”, **Jornal of Multivariate Analysis**, 24, 265-284.

[41] Massy, William F., David B. Montgomery, and Donald G. Morrison. (1970), ”Stochastic Models of Buying Behavior (Cam-bridge, Mass)”.

[42] Matheron, G. (1962), ”Traite de Geostatistique Appliquee”, **Tome I, Memoires du Bureau de Recherches Geologiques et Minieres**, 14, Paris.

[43] Matheron, G. (1963), ”Principles of Geostatistics, with Applications”, **Economic Geology**, 58(8), 1246-1266.

[44] Mazzarino, M. J., Heinrichs, H., and Folster, H. (1983), ” Holocene Versus Accelerated Actual Proton Consumption in German Forest Soils”, **Effects of Accumulation of Air Pollutants in Forest Ecosystems. Reidel, Dordrecht**, 64, 531-534.

[45] McLoad, A. I. (1977),” Improved Box-Jenkins Estimators”, **Biometrika**, 64, 531-534.

[46] McCullach, P. and Nelder, J. A. (1989),” Generalized Linear Models”, Second Edition, Chapman , Hall, London.

[47] Melanie, M. W. (2002), ” A Close Look at the Spatial Structure Impied by the Car and Sar Models” **Jornal of Statistical Planning and Inference**, 121, 311-324.

[48] Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J. P., and Guinet, C. (2006). ”Geostatistical Modelling of Spatial Distribution of Balaenoptera Physalus in the Northwestern Mediterranean Sea from Sparse Count Data and Heterogeneous Observation Efforts”, **Ecological Modelling**, 193, 615-628.

[49] Nelder, J. A. and Wedderburn, R. W. M. (1972), ” Generalized Linear Models”, **Journal of the Royal Statistical Society**, Series A, 135(3), 370-384.

[50] Nelder, J . A and Baker, R. J. (1972), ” Generalized Linear Models”, **In Encyclopedia of Statistical Sciences**, chapter 4.

[51] Oliver, M.A., Webster, R., Lajaunie, C., Muir, K.R., Parkes, S.E., Cameron, A.H., Stevens, M.C.G. and Mann, J.R.,. (1998), "Binomial Cokriging for Estimating and Mapping the Risk of Childhood Cancer", **Mathematical Medicine and Biology**, 15, 279-297.

[52] Pack, S. E. (1986), "Hypothesis Testing for Proportions with Overdispersion", **Biometrics**, 42, 967-972.

[53] Paul, S. R. (1982), "Analysis of Proportions of Affected Fetuses in Teratology Experiments", **Ecological Modelling**, 38, 361-370.

[54] Pearson, E. S. (1925), " Bayes' theorem, Examined in the Light of Experimental Sampling", **Biometrics**, 17, 388-442.

[55] Pettitt, A. N, Weir, I. S. and Hart, A. G. (2002), "A Conditional Autoregressive Gaussian Process for Irregularly Spaced Multivariate Data with Application to Modeling Large Sets of Data." **Statistics and Computing**, 4, 95-115.

[56] Ripley, B. D. (1981), "**Spatial Statistics**", John Wiley, New York.

[57] Rue, H., Martino, S., and Chopin, N. (2009), " Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations", **Journal of the Royal Statistical Society B**, 71, 319–392.

[58] Sain, R. E, and Cressie, N. (2004), "A Spatial Model for Multivariate Lattice Data", Ohaio State. University.

[59] Schuckers, M. E. (2003), " Using the Beta-Binomial Distribution to Assess Performance of a Biometric Identification Device", **Nternational Journal of Image and Graphics**, 3(03), 523-529.

[60] Skellam, J. G. (1948), " A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as a Variable Between the Sets of Trials", **Journal of the Royal Statistical Society. Series B (Methodological)**, 10, 257-261.

[61] Solow, A. R. (1986), " Mapping by Simple Indicator Kriging", **Mathematical Geology**, 18, 335–352.

[62] Snow, J. (1854), " he cholera near Golden-Square, and at Deptford", **Medical Times and Gazette**, 9, pp.321-322..

[63] Tarone, R. E. (1982), "The Use of Historical Control Information in Testing for a Trend in Proportions", **Ecological Modelling**, 34, 69-76.

[64] Tripathi, R. C. Gupta, R. C and Gurland. J. (1994), "Estimation of Parameters in the Beta Binomial Model", **Annals of the Institute of Statistical Mathematics**, 46(2), 317–331.

[65] Whittle, P. (1954), " On Stationary Processes in the Plance", **Biometrika**, 434-449.

[66] Wilkinson D, Tanser F. (1999) " GIS/ GPS to Document Increased Access to Comminity-Based Treatment for Tuberculosis in Africa.", **The Lancet**, 354, 394-395.

[67] Williams, D. A. (1975), " The Analysis of Binary Responses from Toxicological Experiments Involving Reproductionand Teratogenicity", **Biometrics**, 31, pp. 949–952.

[68] Yahav, I, Shmueli, G. (2007), "An Elegant Method for Generating Multivariate Poisson Random Variables", **rXiv:0710**, 5670 [stat.CO].

[69] Zimmerman, D., Pavlik, C, Ruggles, A., and Armstrong, M. P. (1999), " An Experimental Comparison of Ordinary and Universal Kriging and Inverse Distance Weighting", **Mathematical Geology**, 31, 375–390.

پیوست آ

دستورات نرم افزار R

در این پیوست دستورات مربوط به شبیه‌سازی میدان تصادفی بتاـ دوجمله‌ای، برآورد پارامترها و پیش‌گویی فضایی شبیه‌سازی‌شده در فصل سوم، ارایه شده‌اند.

```
library(RandomFields)
BETA<-c(1,3)
SILL<-1
RANGE<-10
NUGGET<-0
SAMPLE<-c(2,4,8)
N<-1000
#simulate a Gassian random field
lat<- runif(n=N,min=0,max=20)
lat
lon<-runif(n=N,min=0,max=20)
lon
coord<-cbind(lat,lon)
coord
```

```
colnames(coord)<-c('lat','lon')
colnames(coord)
nobs<-length(lat)
nobs
model<- RMspheric(var=SILL, scale=RANGE)+ RMnugget(var=NUGGET)
sim.normal<- RFsimulate(model=model, x=lat, y=lon, grid=FALSE)
sim.normal
model
mean<-mean(sim.normal$variable1)
mean
sd<-sd(sim.normal$variable1)
sim.cdf<-pnorm(sim.normal$variable1,mean=mean, sd=sd)
sim.beta<-qbeta(sim.cdf,shape1=BETA[1],shape2=BETA[2])
sim.beta<-cbind(coord,sim.beta)
colnames(sim.beta)<-c('lat','lon','B')
sim.beta<- as.data.frame(sim.beta)
#Generate binomial
sim.binomial<- matrix(nrow=nobs, ncol=2)
for(I in 1:nobs){
size<-sample(SAMPLE,1)
sim.binomial[I,1]<-rbinom(n=1,size=size,prob=sim.beta$B[I])
sim.binomial[I,2]<-size
}
sim.binomial<-as.data.frame(cbind(coord,sim.binomial))
colnames(sim.binomial)<-c('lat','lon','Z','N')
sim.binomial$P<-sim.binomial$Z/sim.binomial$N
sim.binomial$P
```

دستورات لازم برای برآورد پارامترهای شکل بتا

```
library(RandomFields)
library(geoR)
library(sp)
BETA<-c(5,1)
SILL<-1
```

```
RANGE<-10
NUGGET<-0
SAMPLE<-2:5
N<-500
est.param<-matrix(nrow=1000,ncol=6)
colnames(est.param)<-c('alpha_small','beta_small','alpha_medium','beta_medium',
'alpha_large','beta_large')
for(RUN in 1:1000){
lat<-runif(n=N,min=0,max=20)
lon<-runif(n=N,min=0,max=20)
coord<-cbind(lat,lon)
nobs<-N
colnames(coord)<-c('lat','lon')
model<-RMspheric(var=SILL,scale=RANGE)+RMnugget(var=NUGGET)
sim.normal<-RFsimulate(model=model,x=lat,y=lon,grid=FALSE)
mean<-mean(sim.normal$variable1)
sd<-sd(sim.normal$variable1)
sim.cdf<-pnorm(sim.normal$variable1,mean=mean,sd=sd)
sim.beta<-qbeta(sim.cdf,shape1=BETA[1],shape2=BETA[2])
sim.beta<-cbind(coord,sim.beta)
colnames(sim.beta)<-c('lat','lon','B')
sim.beta<-as.data.frame(sim.beta)
sim.binary.small<-matrix(nrow=nobs,ncol=2)
small<-2:5
sim.binary.medium<-matrix(nrow=nobs,ncol=2)
medium<-10:20
sim.binary.large<-matrix(nrow=nobs,ncol=2)
large<-50:100
for(I in 1:nobs){
size.s<-sample(small,1)
sim.binary.small[I,1]<-rbinom(n=1,size=size.s,prob=sim.beta$B[I])
sim.binary.small[I,2]<-size.s
}
lat<-runif(n=N,min=0,max=20)
```

```
lon<-runif(n=N,min=0,max=20)
coord<-cbind(lat,lon)
nobs<-N
colnames(coord)<-c('lat','lon')
model<-RMspheric(var=SILL,scale=RANGE)+RMnugget(var=NUGGET)
sim.normal<-RFsimulate(model=model,x=lat,y=lon,grid=FALSE)
mean<-mean(sim.normal$variable1)
sd<-sd(sim.normal$variable1)
sim.cdf<-pnorm(sim.normal$variable1,mean=mean,sd=sd)
sim.beta<-qbeta(sim.cdf,shape1=BETA[1],shape2=BETA[2])
sim.beta<-cbind(coord,sim.beta)
colnames(sim.beta)<-c('lat','lon','B')
sim.beta<-as.data.frame(sim.beta)
for(I in 1:nobs){
size.m<-sample(medium,1)
sim.binary.medium[I,1]<-rbinom(n=1,size=size.m,prob=sim.beta$B[I])
sim.binary.medium[I,2]<-size.m
}
lat<-runif(n=N,min=0,max=20)
lon<-runif(n=N,min=0,max=20)
coord<-cbind(lat,lon)
nobs<-N
colnames(coord)<-c('lat','lon')
model<-RMspheric(var=SILL,scale=RANGE)+RMnugget(var=NUGGET)
sim.normal<-RFsimulate(model=model,x=lat,y=lon,grid=FALSE)
mean<-mean(sim.normal$variable1)
sd<-sd(sim.normal$variable1)
sim.cdf<-pnorm(sim.normal$variable1,mean=mean,sd=sd)
sim.beta<-qbeta(sim.cdf,shape1=BETA[1],shape2=BETA[2])
sim.beta<-cbind(coord,sim.beta)
colnames(sim.beta)<-c('lat','lon','B')
sim.beta<-as.data.frame(sim.beta)
for(I in 1:nobs){
size.1<-sample(large,1)
```

```
sim.binary.large[I,1]<-rbinom(n=1,size=size.1,prob=sim.beta$B[I])
sim.binary.large[I,2]<-size.1
}
sim.binary.small<-as.data.frame(cbind(coord,sim.binary.small))
colnames(sim.binary.small)<-c('lat','lon','Z','N')
sim.binary.medium<-as.data.frame(cbind(coord,sim.binary.medium))
colnames(sim.binary.medium)<-c('lat','lon','Z','N')
sim.binary.large<-as.data.frame(cbind(coord,sim.binary.large))
colnames(sim.binary.large)<-c('lat','lon','Z','N')
# Estimate beta parameters
# Depending on computer platform (Mac or PC) , family name
# is 'betabinomial .ab' or 'betabinomialff '
library(VGAM)
fit.small<-vglm(cbind(Z,N-Z)~1,betabinomialff,data=sim.binary.small)
est.param[RUN,1]<-Coef(fit.small)[1]
est.param[RUN,2]<-Coef(fit.small)[2]
fit.medium<-vglm(cbind(Z,N-Z)~1,betabinomialff,data=sim.binary.medium)
est.param[RUN,3]<-Coef(fit.medium)[1]
est.param[RUN,4]<-Coef(fit.medium)[2]
fit.large<-vglm(cbind(Z,N-Z)~1,betabinomialff,data=sim.binary.large)
est.param[RUN,5]<-Coef(fit.large)[1]
est.param[RUN,6]<-Coef(fit.large)[2]
}
```

دستورات لازم برای برآورد پارامترهای ساختار وابستگی فضایی

```
rm(list=ls())
library(RandomFields)
library(RColorBrewer)
library(geoR)
library(sp)
library(VGAM)
set.seed(2387)
STARTING <- read.csv('starting_values.csv')
N <- 100
```

```
NSIM <- 100
for(s in 1:nrow(STARTING)){
BETA <- STARTING[s,1:2]
SILL<-1
RANGE <- STARTING[s,3]
NUGGET<-0
SAMPLESIZE<-STARTING[s,4]:STARTING[s,5]

est.param<-matrix(nrow=NSIM,ncol=14)
colnames(est.param)<-c('beta.tau','beta.sigma','beta.phi','bbk1.tau','bbk1.sigma'
,'bbk1.phi','bbk2.tau','bbk2.sigma','bbk2.phi','bbk3.tau','bbk3.sigma',
'bbk3.phi','alpha','beta')
for(RUN in 1:NSIM){
print(RUN)
#Simulate a dense Gaussian random fields
lat<-seq(from=0,to=20,length=100)
lon<-seq(from=0,to=20,length=100)
coord<-expand.grid(lat,lon)
nobs<-N
colnames(coord)<-c('lat','lon')
model<-RMspheric(var=SILL,scale=RANGE)+RMnugget(var=NUGGET)
sim.normal<-RFsimulate(model=model,x=lat,y=lon,grid=TRUE)
sim.normal<-cbind(coord,sim.normal$variable1)
colnames(sim.normal)<-c('lat','lon','A')
at<-seq(from=-4,to=4,length=100)
col<-colorRampPalette(brewer.pal(9,'Reds'))(100)
coordinates(sim.normal)<-~lat+lon
mean<-mean(sim.normal$A)
sd<-sd(sim.normal$A)
sim.cdf<-pnorm(sim.normal$A,mean=mean,sd=sd)
sim.beta<-qbeta(sim.cdf,shape1=BETA$alpha,shape2=BETA$beta)
sim.beta<-cbind(coord,sim.beta)
colnames(sim.beta)<-c('lat','lon','B')
sim.beta<-as.data.frame(sim.beta)
```

```
at<-seq(from=0,to=1,length=100)
coordinates(sim.beta)<-~lat+lon
gridded(sim.beta)<-TRUE
spplot(sim.beta,zcol='B',main=paste('Simulated beta distribution,alpha=',BETA[1],
'beta=',BETA[2]),at=at,col.regions=col,colorkey=TRUE)
sample.points<-sample(x=1:nrow(coord),size=N)
beta.points<-sim.beta[sample.points,]
coord.points<-coord[sample.points,]
sim.binomial<-matrix(nrow=N,ncol=2)
for(I in 1:N){
size<-sample(SAMPLESIZE,1)
sim.binomial[I,1]<-rbinom(n=1,size=size,prob=beta.points$B[I])
sim.binomial[I,2]<-size
}
sim.binomial<-as.data.frame(cbind(coord.points,sim.binomial))
colnames(sim.binomial)<-c('lat2','lon2','Z','N')
sim.binomial$P<-sim.binomial$Z/sim.binomial$N
fit.ab<-vglm(cbind(Z,N-Z)~1,family=betabinomialff,data=sim.binomial)
est.param[RUN,13]<-Coef(fit.ab)[1]
est.param[RUN,14]<-Coef(fit.ab)[2]
alpha<-Coef(fit.ab)[1]
beta<-Coef(fit.ab)[2]
bias<-alpha*beta/((alpha+beta)*(alpha+beta+1))
#Estimate beta-binomial kriging variogram
varS<-matrix(nrow=N*N,ncol=3)
colnames(varS)<-c('dist','value','mult')
index<-1
for(i in 1:N){
for(j in 1:N){
mult<-(sim.binomial$N[i]*sim.binomial$N[j])/(sim.binomial$N[i]+sim.binomial$N[j])
varS[index,1]<-dist(sim.binomial[c(i,j),1:2])
varS[index,2]<-mult*(dist(sim.binomial$P[c(i,j)])^2)-bias
varS[index,3]<-mult
index<-index+1
```

```
}
}
varS<-as.data.frame(varS)
bin.varP<-variog(data=sim.binomial$P,coords=coord.points,breaks=1:20)
beta.var<-variog(data=sim.beta$B,coords=coord,breaks=1:20)
#Use lag distances as defined in 'variog'
h<-1:20
bin.varS<-matrix(nrow=length(h),ncol=4)
colnames(bin.varS)<-c('nobs','var','h','ave.weights')
for(H in h){
x<-varS[varS$dist<H& varS$dist>=(H-1),]
bin.varS[H,1]<-nrow(x)
bin.varS[H,2]<-max(colSums(x)[2]/(2*colSums(x)[3]),0)
bin.varS[H,3]<-H
bin.varS[H,4]<-mean(1/mult)
}
vario.beta<-beta.var
vario.prop<-bin.varP
vario.bbk<-bin.varP
vario.bbk$v<-bin.varS[,2]
vario.bbk2<-bin.varP
vario.bbk2$v<-bin.varS[,2]
vario.bbk2$n<-bin.varS[,4]*bin.varS[,1]
beta.parm <- variofit(vario=vario.beta, ini.cov.pars=c(0.1,RANGE), cov.model='sph',
fix.nugget=FALSE)
bbk.parm1 <- variofit(vario=vario.bbk, ini.cov.pars=c(0.1,RANGE),cov.model='sph',
fix.nugget=FALSE,weights='npairs')
bbk.parm2 <- variofit(vario=vario.bbk2, ini.cov.pars=c(0.1,RANGE),cov.model='sph',
fix.nugget=FALSE,weights='npairs')
bbk.parm3 <- variofit(vario=vario.bbk, ini.cov.pars=c(0.1,RANGE),cov.model='sph',
fix.nugget=FALSE,weights='cressie')
est.param[RUN,1]<-beta.parm$nugget
est.param[RUN,2]<-beta.parm$cov.pars[1]
est.param[RUN,3]<-beta.parm$cov.pars[2]
```

```
est.param[RUN,4]<-bbk.parm1$nugget
est.param[RUN,5]<-bbk.parm1$cov.pars[1]
est.param[RUN,6]<-bbk.parm1$cov.pars[2]
est.param[RUN,7]<-bbk.parm2$nugget
est.param[RUN,8]<-bbk.parm2$cov.pars[1]
est.param[RUN,9]<-bbk.parm2$cov.pars[2]
est.param[RUN,10]<-bbk.parm3$nugget
est.param[RUN,11]<-bbk.parm3$cov.pars[1]
est.param[RUN,12]<-bbk.parm3$cov.pars[2]
write.csv(est.param,paste('param',s,'.csv'))
}
}
```

دستورات لازم برای پیش‌گویی فضایی

```
library(RandomFields)
library(geoR)
library(sp)
library(VGAM)
# sim.binomial is the simulated binomial sample data
colnames(sim.binomial) <- c('lat','lon', 'Z','N','P')
fit.ab <- vglm(cbind(Z,N-Z)~1,family=betabinomialff,data=sim.binomial)
alpha <- Coef(fit.ab)[1]
beta <- Coef(fit.ab)[2]
bias <- alpha*beta/((alpha+beta)*(alpha+beta+1))
# Estimate beta?binomial variogram
varS <- matrix(nrow=N*N,ncol=3)
colnames(varS) <- c('dist','value','mult')
index<-1
for(i in 1:N){
  for(j in 1:N){
mult <- (sim.binomial$N[i]*sim.binomial$N[j])/
(sim.binomial$N[i]+sim.binomial$N[j])
varS[index,1] <- dist(sim.binomial[c(i,j),1:2])
```

```
varS[index,2] <- mult*(dist(sim.binomial$P[c(i,j)])^2)-bias
varS[index,3] <- mult
index <- index+1
  }
}
varS <- as.data.frame(varS)

bin.varP <- variog(data=sim.binomial$P,coords=coord.points,breaks=1:20)
beta.var <- variog(data=sim.beta$B,coords=coord,breaks=1:20)
# Use lag distances as defined in 'variog '
h <- 1:20
bin.varS <- matrix(nrow=length(h),ncol=4)
colnames(bin.varS) <- c('nobs','var','h','ave.weights')
for(H in h){
  x<-varS[varS$dist<H & varS$dist>=(H-1),]
  bin.varS[H,1]<-nrow(x)
  bin.varS[H,2]<-max(colSums(x)[2]/(2*colSums(x)[3]),0)
  bin.varS[H,3]<-H
  bin.varS[H,4]<-mean(1/mult)
}

vario.beta <- beta.var
vario.prop <- bin.varP
vario.bbk2 <- bin.varP
vario.bbk2$v <- bin.varS[,2]
vario.bbk2$n <- bin.varS[,4]*bin.varS[,1]
beta.parm <- variofit(vario=vario.beta,ini.cov.pars=c(0.08,RANGE),
cov.model='sph',fix.nugget=FALSE)
bbk.parm <- variofit(vario=vario.bbk2,ini.cov.pars=c(0.08,RANGE),
cov.model='sph',fix.nugget=FALSE,weights='npairs')
if(bbk.parm$nugget<0) bbk.parm$nugget=0
##MAKE predictions
A <- diag(c(bias/sim.binomial$N,0))
B <- matrix(rep(1,(N+1)^2),nrow=(N+1),ncol=(N+1))
```

```
for(i in 1:N){
  for(j in 1:N){
dist <- dist(sim.binomial[c(i,j),1:2])
B[i,j] <- bbk.parm$cov.pars[1]*(1-(3*dist/(2*bbk.parm$cov.pars[2])-
dist^3/(2*bbk.parm$cov.pars[2]^3)))
  if(dist>bbk.parm$cov.pars[2]) B[i,j] <- 0
  }
}
B[N+1,N+1] <- 0
C <- A+B
Cinv <- solve(C)
# Set up C0 matrix for each point at location o,
# solve for the weights , and predict
pred <- matrix(nrow=nrow(coord),ncol=4)
colnames(pred) <- c('lat','lon','pred','var')
for(o in 1:nrow(coord)){
C0<-matrix(rep(1,N+1),ncol=1)
for(i in 1:N){
  dist <- dist(rbind(sim.binomial[i,1:2],coord[o,1:2]))
  C0[i,1] <- bbk.parm$cov.pars[1]*(1-(3*dist/(2*bbk.parm$cov.pars[2])-
        dist^3/(2*bbk.parm$cov.pars[2]^3)))
        if(dist>bbk.parm$cov.pars[2]) C0[i] <- 0
  }
lambda <- Cinv%*%C0
pred[o,3] <- max(min(sum(lambda[1:N]*sim.binomial$P),1),0)
pred[o,1] <- coord[o,1]
pred[o,2] <- coord[o,2]
pred[o,4] <- bbk.parm$cov.pars[1]-sum(lambda*C0)
}

pred <- as.data.frame(pred)
pred$beta <- sim.beta$B
pred$error <- pred$pred-sim.beta$B
```

Abstract

The popular model for spatial count responses in finite populations is constructed based on binomial distribution. In many applications, due to overdispersion problem, to use binomial distribution is not appropriate and can result in inefficient inferences regarding the estimation of parameters and spatial prediction. In order to eliminate the defects of a binomial model, an alternative approach for modeling spatially correlated proportions is to use a beta-binomial distribution. The flexibility of beta-binomial distribution could make it the first choice for modeling overdispersed count responses in finite populations. In this thesis, we describe the spatial prediction of rates by using the flexible beta-binomial model and generalize it for spatial analysis of lattice data in the framework of Bayesian inference. We also demonstrate its efficacy by a simulation study. To evaluate the performance of this model, we analyze two real data sets including the number of rainfall days in Semnan province, Iran and the number of people with gastric cancer in Guilan province, Iran.

Keywords: Spatial counts, beta-binomial process, overdispersion, spatial prediction.

Shahrood University of Technology

Faculty of Mathematical Sciences

MSc Thesis in: Mathematical Statistics

A flexible model for spatial prediction of rate responses by using beta-binomial distribution

By: Leila Abedinpour Liiarjadmeh

Supervisor

Dr. Hossein Baghishani

Advisor

Dr. Negar Eghbal

January 2018