مقاله پژوهشی

روشی نوین در براورد ناپارامتری تابع شدت فرایندهای نقطه‌ای پواسون فضایی و کاربرد آن در براورد شدت رویش درختان اینگاساپیندوئیدس

میترا هاشمی‌نیا

(نویسنده مسئول)، دانشجوی دکتری آمار، دانشکده آمار، ریاضی و کامپیوتر، دانشگاه علامه طباطبایی، تهران، ایران. mitra_hasheminia94@atu.ac.ir

رضا پورطاهری

دانشیار گروه آمار، دانشکده آمار، ریاضی و کامپیوتر، دانشگاه علامه طباطبایی، تهران، ایران. pourtaheri@atu.ac.ir

چکیده: مدل‌سازی و براورد تابع شدت یک الگوی نقطه‌ای، یکی از مسائل مقدماتی و در عین حال اساسی در استنباط فرایندهای نقطه‌ای است و به‌عنوان پیش‌نیاز بسیاری از مسائل دیگر به شمار می‌رود و از دیدگاه‌های مختلفی به آن پرداخته شده‌است. با پیشرفت سریع فناوری‌های جمع‌آوری داده، طیف گسترده‌ای از داده‌ها تولید شده‌است و در نظر گرفتن متغیرهای کمکی، گام بزرگی در جهت پیشرفت نظریه‌ی فرایند نقطه‌ای بوده‌است و تاکنون عمدتاً از منظر پارامتری به آن پرداخته شده‌است.

در این مقاله ما روشی جدید برای براورد ناپارامتری تابع شدت یک فرایند نقطه‌ای پواسون ناهمگن که تابعی نامعلوم از چندین متغیر کمکی فضایی مستقل است، معرفی می‌کنیم. در روش پیشنهادی، با بهره‌گیری از تکنیک تقریب توابع پایه شعاعی برای توابع چند متغیره، مدل ناپارامتری تابع شدت به یک مدل لگ خطی تبدیل می‌شود. از آن‌جایی که دقت تقریب تابع چندمتغیره به‌طور مستقیم بر دقت براورد تابع شدت تاثیر دارد، بنابراین با بهینه‌سازی پارامتر شکل تابع پایه شعاعی از طریق کمینه کردن ملاک اطلاع بیزی، کیفیت براورد ناپارامتری تابع شدت فرایندهای نقطه‌ای پواسون فضایی را ارتقا می‌دهیم.

عملکرد روش پیشنهادی با استفاده از یک مطالعه‌ی شبیه‌سازی شده و داده‌های واقعی که به بررسی شدت یک گونه از درختان به نام "اینگاساپیندوئیدس" در جنگل‌های گرمسیری می‌پردازد، ارزیابی شد. نتایج نشان داد که روش پیشنهادی ما قادر است براورد ناپارامتری تابع شدتِ فرایندهای نقطه‌ای پواسون فضایی را با دقت مناسبی ارائه دهد. این رویکرد جدید در مقایسه با روش‌های ناپارامتری موجود که در تعداد متغیرهای کمکی مورد استفاده محدودیت دارند، از مزیت‌های اساسی برخوردار است.

واژگان کلیدی: براورد حداکثر درستنمایی تقریبی، تقریب تابع چند متغیره، شمارش پیکسل‌ها، فرایند نقطه‌ای پواسون فضایی ناهمگن، متغیرهای کمکی فضایی مستقل.

۱. مقدمه

فرایندهای نقطه‌ای، فرایندهای تصادفی هستند که برای مدل‌بندی رویدادهایی که در فضا یا زمان و یا به‌طور همزمان در فضا و زمان اتفاق می‌افتند، استفاده می‌شوند. بنابراین فرایند نقطه‌ای به سه شاخه‌ی فرایندهای نقطه‌ای فضایی، زمانی و فضایی-زمانی تقسیم می‌شوند. در این مقاله فرایند نقطه‌ای فضایی مورد بررسی قرار خواهد گرفت که مدلی مفید برای یک الگوی تصادفی از نقاط در فضای d بُعدی است که معمولاً $d = 2$ در نظر گرفته می‌شود. با توجه به این‌که فرایندهای نقطه‌ای فضایی، شاخه‌ای از آمار فضایی است که هدف اصلی آن مطالعه‌ی ساختار هندسی الگوهای ایجاد شده توسط اشیایی (به‌نام پیشامدها) است که به‌طور تصادفی در فضا توزیع می‌شوند، بنابراین این مدل‌ها می‌توانند به‌طور مستقیم برای مدل‌بندی و تجزیه و تحلیل داده‌هایی که به شکل یک الگوی نقطه‌ای مانند نقشه‌ی مکان‌های درختان یا لانه‌های پرندگان (در زمینه‌ی بوم‌شناسی و جنگل‌داری، ایلیان و همکاران [۱]؛ لاو و همکاران [۲]؛ رنر و همکاران [۳] و اشتویان و پنتینن [۴] را ببینید.)، موقعیت ستارگان و کهکشان‌ها (در زمینه‌ی اخترشناسی و نجوم، بابو و فیگلسون [۵]؛ بدلی و همکاران [۶] را ببینید)، مکان نورون‌ها در بافت مغز (در زمینه‌ی علوم

Corresponding author: mitra_hasheminia94@atu.ac.ir
https://doi.org/10.48313/jqem.2023.199824

تاریخ دریافت: ۱۴۰۲/۱۲/۲۹ تاریخ پذیرش: ۱۴۰۳/۰۳/۱۵
دوره ۱۳/ شماره ۳
صفحات ۳۱۷-۳۳۴

اعصاب، جعفری ممقانی[۷] را ببینید.)، یا آدرس منزل افراد مبتلا به یک بیماری نادر (در زمینه‌ی اپیدمیولوژی، دیگل [۸]، [۹]؛ گاترل و همکاران [۱۰] را ببینید) هستند، مورد استفاده قرار گیرند.

فرایندهای نقطه‌ای انواع مختلفی دارند که در آن میان، فرایند پواسون یک فرایند نقطه‌ای ساده است که در حوزه‌ی فرایندهای نقطه‌ای نقشی اساسی دارد و بسیاری از فرایندهای نقطه‌ای پیشرفته‌تر براساس فرایند پواسون ساخته می‌شوند. بنابراین در بسیاری از مطالعات و کاربردها روی این فرایند تمرکز شده است و از آن به‌عنوان فرایند مرجع استفاده می‌شود. فرایند نقطه‌ای پواسون براساس "تابع شدت" به دو رده‌ی "فرایند نقطه‌ای پواسون همگن" و "فرایند نقطه‌ای پواسون ناهمگن" تقسیم می‌شود. فرایند پواسون همگن فرایند پواسونی است که تابع شدت آن مقداری ثابت باشد، در غیر این‌صورت آن را فرایند پواسون ناهمگن می‌نامند.

منظور از "شدت"، چگالیِ مورد انتظار نقاط در واحد سطح است که به‌طور معمول به‌عنوان نرخِ وقوع، فراوانی یا بروز پیشامدهای ثبت شده در الگوی نقطه‌ای تعبیر می‌شود. بنابراین تعریف زیر را برای تابع شدت λ می‌توان ارائه داد:

$$\lambda(s) = \lim_{|ds|\to 0}\left\{\frac{E(N(ds))}{|ds|}\right\}$$

تابع شدت λ یک مشخصه‌یِ توصیفی اصلی از یک فرایند نقطه‌ای است. بنابراین مدل‌بندی و براورد تابع شدت فرایند نقطه‌ای، از اهداف اصلی در نظریه فرایند نقطه‌ای فضایی است. این تابع احتمال وقوع یک نقطه (یا یک پیشامد) را در یک توپ بینهایت کوچک در اطراف یک مکانِ مشخص نشان می‌دهد و معمولاً فرض می‌شود که ناهمگن است، به این معنا که توسط یک تابع شدت متغیر مشخص می‌شود. این مسئله از منظرهای مختلف از جمله روش‌های پارامتری و ناپارامتری مورد بررسی قرار گرفته‌است. در روش پارامتری با در نظر گرفتن یک مدل ریاضی برای تابع شدت که وابسته به مقادیر پارامتر است، براورد براساس یافتن مقادیر پارامتر صورت می‌پذیرد در حالی‌که در روش ناپارامتری با این فرض که اطلاعی از توزیع فرایند در دسترس نیست، تابع شدت با استفاده از روش‌های تقریب‌زدن تابع به صورت احتمالاتی براورد می‌شود.

از طرفی در نظر گرفتن متغیرهای کمکی فضایی که بر وقوع پیشامدهای مورد علاقه در فضا تأثیر می‌گذارند، یکی از موضوعات مهم و جذاب در تحلیل الگوهای نقطه‌ای است. برای مثال، در بررسی الگوهای فضایی وقوع آتش‌سوزی جنگل‌ها درکاتالونیا اسپانیا متغیرهای کمکی فضایی مانند نوع پوشش گیاهی، ویژگی‌های توپوگرافی (ارتفاع، شیب زمین)، شرایط هواشناسی (دما، رطوبت، سرعت باد)، فعالیت‌های انسانی (نزدیکی به جاده‌ها، تراکم جمعیت) در نظر گرفته‌شد. محققان علاقه‌مند بودند بدانند آیا الگوهای فضایی وقوع آتش‌سوزی‌ها در جنگل‌ها، با عوامل محیطی و انسانی مرتبط هستند یا خیر؟ علاوه براین نتایج نشان داد که استفاده از این متغیرهای کمکی، توانایی مدل فرایند نقطه‌ای در درک الگوهای فضایی وقوع آتش‌سوزی در جنگل را بهبود می‌بخشد. (دیاز دلگادو [۱۱] را ببینید.)

از سال ۱۹۷۰ مدل‌های پارامتری بسیاری برای بررسیِ تاثیر متغیر(های) کمکی فضایی بر شدت فرایند نقطه‌ای مطرح شده‌اند که ساده‌ترین و متداول‌ترین مدل پارامتری، مدل لُگ خطی است که عبارت است از:

$$\lambda(s) = exp\{\theta_0 + \theta Z(s)\},\ \ s \in D \subset \mathbb{R}^2, \qquad (۱)$$

بعدها مطالعات متعددی جهت توسیع مدل (۱) به حالت چند متغیره انجام شد که از جمله می‌توان به واگه پترسون [۱۲]، چورالدین و همکاران [۱۳] و [۱۴] اشاره کرد.

اگرچه روش‌های پارامتری بسیار محبوب هستند اما بایستی دقت شود که اگر چنانچه مدل پارامتری مفروض برای تابع شدتِ واقعی مناسب نباشد (یعنی تا حد زیادی از تابع شدت واقعی انحراف داشته باشد)، روش پارامتری ممکن است منجر به براوردهایی نامناسب و غیرقابل اطمینان شود. لذا روش‌های ناپارامتری که دارای انعطاف‌پذیری بیشتر و قابلیت کاربرد گسترده‌تری نسبت به روش‌های پارامتری هستند به‌عنوان جایگزینی مناسب برای روش‌های پارامتری در نظر گرفته می‌شوند.

برای براورد تابع شدت وابسته به متغیرهای کمکی فضایی که به صورت ناپارامتری مدل‌بندی شده‌اند تاکنون کارهای اندکی انجام شده‌است. گوان [۱۵] یک براوردگر هسته‌ی مبتنی بر چندین متغیر کمکی را برای براورد تابع شدت با تعریف فاصله‌ی بین هر دو نقطه به عنوان تفاوت بین مقادیر متغیر کمکی مرتبط با آن‌ها ارائه داد. او ابتدا به کمک تکنیک کاهش بُعدِ "رگرسیون وارون ورقه ورقه شده[1]"، با مسئله‌ی افزایش بُعد مقابله کرد و سپس براوردگر هسته‌ی سازگار خود را معرفی

[1] Sliced inverse regression

کرد. بدلی و همکاران [۱۶]، از اطلاعات یک متغیر کمکی پیوسته برای مدل‌بندی تابع شدت استفاده کردند و براوردگرهای ناپارامتریِ هسته و درستنمایی موضعی را برای تابع شدت پیشنهاد کردند. بوراجو و همکاران [۱۷] مدلی مشابه با بدلی و همکاران در نظر گرفتند و براوردگر هسته‌ی سازگار برای شدت با استفاده از انتخابگرهای پهنای باند بوت استرپ معرفی کردند. آن‌ها همچنین براوردگر پیشنهادی خود را برای فرایندهای نقطه‌ای فضایی-زمانی و همچنین در چارچوب نسخه‌ی چند متغیره‌ی آن، بسط دادند. کورجولی و همکاران [۱۸] یک مدل لگ-پیچش[۲] را برای پارامتری‌سازی شدت فرایند نقطه‌ای فضایی معرفی کردند و وارد و همکارانش [۱۹] براوردی از توابع شدت فرایندهای پواسون را در خمینه‌ی ریمانی[۳] ارائه کردند.

با توجه به این‌که تاکنون به براورد ناپارامتری تابع شدت در حضور متغیرهای کمکی کمتر پرداخته شده‌است لذا در این مقاله با در نظر گرفتن مدل ناپارامتری برای تابع شدت به‌صورت

$$\lambda(s) = exp\left(f(\boldsymbol{Z}(s))\right), \qquad s \in D \subset \mathbb{R}^2, \qquad (۲)$$

که در آن f یک تابع پیوسته‌ی نامعلوم است و $\boldsymbol{Z}(s) = \left(Z_1(s), \dots, Z_p(s)\right)^T : \ D \subset \mathbb{R}^2 \longrightarrow \mathbb{R}^p$، بردار متغیرهای کمکی پیوسته‌ی فضایی p بعدی است که در مکان $s \in D$ اندازه‌گیری شده‌اند، به براورد λ می‌پردازیم.

در روش پیشنهادی ما از توابع پایه شعاعی (RBF)[۴] که تقریبی از تابع چند متغیره‌ی نامعلوم f را ارائه می‌دهد، استفاده می‌شود. دستیابی به براوردهایی با دقت مناسب برای تابع شدت مستلزم استفاده از روش‌های بهینه‌سازی در انتخاب پارامترهای مرتبط با RBFها است که در نهایت منجر به یک نسخه‌ی چند متغیره از مدل (۱) می‌شود.

در ادامه ابتدا تعاریفی از توابع پایه شعاعی و فرایندهای نقطه‌ای پواسون فضایی ارائه می‌دهیم، سپس به توضیح روش گسسته‌سازی به‌کار گرفته‌شده برای تبدیل فرایند نقطه‌ای پواسون به مدل رگرسیونی می‌پردازیم. در نهایت کارایی روش پیشنهادی‌مان بااستفاده از مطالعات شبیه‌سازی و مجموعه داده‌ی واقعی از جنگل‌های گرمسیری نشان داده می‌شود.

۲. روش‌شناسی

۱.۲. تقریب تابع پایه شعاعی

استفاده از روش‌های بدون‌شبکه‌بندی[۵] در مواجه با داده‌های چند بُعدی پیشرفت مهمی در زمینه‌های مختلف علمی است. تکنیک توابع پایه شعاعی یکی از روش‌های بدون‌شبکه‌بندی است که به دلیل توانایی در نمایش توابع غیرخطی پیچیده در زمینه‌های مختلفی مانند یادگیری ماشین، تقریب تابع و درون‌یابی به طور گسترده استفاده می‌شوند. در این توابع، موقعیت‌های نسبی (نه موقعیت دقیق) اهمیت دارد.

تابع پایه شعاعی یک تابع حقیقی مقدار است که مقدار آن تنها به فاصله‌ی نقاط ورودی از نقاط ثابت مشخصی که مرکز نامیده می‌شوند، وابسته است. از نظر ریاضی یک تابع پایه شعاعی عبارت است از $\Phi(\boldsymbol{x}) = \varphi(\|\boldsymbol{x} - \boldsymbol{c}\|)$ که در آن c مرکز RBF است و $\|.\|$ مقدار نُرم روی $\mathbb{R}^p$ است که معمولاً نُرم اقلیدسی در نظر گرفته‌می‌شود (فاشاور [۲۰]). φ می‌تواند از توابع پایه‌ای مختلف استفاده کند، مانند تابع گاوسی $\varphi(r) = exp(-(\gamma r)^2)$، که در آن $\gamma (>0)$ پارامتر شکل نامیده می‌شود و میزان مسطح بودن RBFها را کنترل می‌کند.

همان‌طور که اشاره شد RBFها یکی از ابزارهای کاربردی برای تقریب توابع چند متغیره هستند. تقریب RBF برای تابع چند متغیره‌ی $f(\boldsymbol{x})$ به صورت زیر در نظر گرفته می‌شود:

$$f(\boldsymbol{x}) \approx \sum_{k=1}^{m} \theta_k \varphi_k(s) = \sum_{k=1}^{m} \theta_k \, exp(-(\gamma \|\boldsymbol{x} - \boldsymbol{c}_k\|)^2), \quad \boldsymbol{x} \in \mathbb{R}^p, \qquad (۳)$$

که در آن $\{\boldsymbol{c}_k\}_1^m$ نقاط مرجع (مراکز) و $\{\varphi_k\}_1^m$ توابع پایه شعاعی گاوسی هستند که در نقطه‌ی c_k متمرکز شده‌اند، و تابع $f(\boldsymbol{x})$ به صورت مجموعی از توابع پایه شعاعی که هرکدام به یک نقطه‌ی مرجع $\boldsymbol{c}_k$ متفاوت مرتبط هستند و با ضریب θ_k وزن داده می‌شوند نشان داده‌می‌شود.

برای دستیابی به عملکرد بهینه در تقریب‌های RBF، مهم است که کمیت‌هایی مانند پارامتر شکل γ، تعداد خوشه‌های m و مراکز خوشه $\{\boldsymbol{c}_k\}_1^m$ را با دقت انتخاب کنیم. این انتخاب‌ها می‌توانند به‌طور قابل توجهی بر عملکرد تقریب‌ها تأثیر بگذارند. برای تعیین تعداد بهینه‌ی خوشه‌ها می‌توان از ترکیبی از

[2] Log-convolution model
[3] Riemannian manifold
[4] Radial basis function
[5] Meshfree method

خوشه‌بندی سلسله مراتبی[6] و خوشه‌بندی k-means استفاده کرد. برای این منظور ابتدا با به‌کارگیری خوشه‌بندی سلسله مراتبی دندروگرام‌ها[7] را رسم می‌کنیم و از آن‌ها برای تعیین تعداد بهینه‌ی خوشه‌ها، m، استفاده می‌شود. سپس m را به عنوان ورودی برای خوشه‌بندی داده‌ها به روش k-means استفاده می‌کنیم. در این مطالعه، مقدار بهینه برای γ با کمینه کردن ملاک اطلاع بیزی[8] (BIC) حاصل از مدل‌های رگرسیونی تعیین می‌شود. بحث بیشتر در این مورد در بخش ۴.۲ ارائه خواهد شد.

۲.۲. فرایند نقطه‌ای پواسون

فرض کنید $(\mathcal{S},\mathcal{B})$ فضای‌متریک اندازه‌پذیر باشد. یک فرایند نقطه‌ای روی $\mathcal{S}$، نگاشتی مانند Y از فضای احتمال $(\Omega,\mathcal{A},P)$ به فضای اندازه‌پذیر $(\mathcal{S},\mathcal{B})$ است، هرگاه برای هر مجموعه بورل کراندار B، که در آن $B\in\mathcal{B}$، تعداد عناصر Y که در B قرار می‌گیرند، یعنی $N(B)=N_Y(B)=N(\boldsymbol{Y}\cap B)$، یک متغیر تصادفی باشد.

اندازه شدت یک فرایند نقطه‌ای را با نماد $\mu(B)$ نشان می‌دهیم و عبارت است از:

$$\mu(B)=E[N(B)]$$

همچنین فرض کنید $\lambda:\mathcal{S}\longrightarrow[0,\infty)$ یک تابع انتگرال‌پذیر موضعی روی فضای متریک اندازه‌پذیر $(\mathcal{S},\mathcal{B})$ باشد، به‌طوری‌که برای هر مجموعه‌ی بورل کراندار $B\in\mathcal{B}$،

$$\mu(B)=\int_B \lambda(s)ds \qquad \forall B\in\mathcal{B}$$

$\lambda(.)$ را تابع شدت متناظر با $\boldsymbol{Y}$ می‌نامیم و فرض می‌کنیم تابع شدت وجود دارد.

فرایند نقطه‌ای Y روی $D\subset\mathbb{R}^d$ را یک فرایند نقطه‌ای پواسون با اندازه شدت $\mu(.)$ گویند هرگاه:

۱) به ازای هر مجموعه بورل کراندار B، $B\subset D$، متغیر تصادفی $N(B)$ دارای توزیع پواسون با میانگین $\mu(B)$ باشد.

۲) برای هر M مجموعه بورل کراندار و مجزای $B_1,\dots,B_M$، متغیرهای تصادفی $N(B_1),\dots,N(B_M)$ مستقل از یکدیگر باشند. که در آن D دامنه‌ی فضایی مورد علاقه در فضای d–بُعدی است.

همان‌طور که قبلاً اشاره شد، اگر تابع شدت فرایند پواسون، $\lambda(.)$، ثابت باشد، فرایند را همگن و در غیر این صورت، یعنی چنانچه $\lambda(s)$ به s وابسته باشد، فرایند را ناهمگن می‌نامند.

۳.۲. تابع K ناهمگن

تابع K ناهمگن برای شناسایی ساختار الگوی نقطه‌ای فرایندهای ناهمگن به‌کار می‌رود. اگر $\lambda(s)$ تابع شدت واقعی فرایند نقطه‌ای Y باشد، تابع K ناهمگن به صورت زیر تعریف می‌شود:

$$K_{inhom}(r)=\frac{1}{|W|}\,E\left[\sum_{s_i\in Y\cap D}\ \sum_{s_j\in Y\setminus\{s_i\}}\frac{I(\|s_i-s_j\|\le r)}{\lambda(s_i)\lambda(s_j)}\right]$$

که در آن $I(.)$ معرف تابع نشانگر است. اگر الگوی نقطه‌ای از یک فرایند پواسون ناهمگن پیروی کند، مقدار نظری آن برابر با πr^2 است. تابع $\boldsymbol{K}$ می‌تواند به صورت زیر براورد شود:

$$\widehat{K}_{inhom}(r)=\frac{1}{|W|}\sum_{s_i\in Y\cap D}\ \sum_{s_j\in Y\setminus\{s_i\}}\frac{e(s_i,s_j)I(\|s_i-s_j\|\le r)}{\lambda(s_i)\lambda(s_j)}$$

که در آن $e(s_i,s_j)$ فاکتور تصحیح لبه است.

۴.۲. راهبرد محاسباتی

فرض کنید $s_1,\dots,s_n$ نشان دهنده‌ی تحقق‌های مشاهده شده‌ی فرایند نقطه‌ای پواسون $\boldsymbol{Y}$ در D است، که در آن n تعداد نقاط در دامنه‌ی فضایی است که مقداری تصادفی و متناهی دارد. لگاریتم تابع درستنمایی فرایند نقطه‌ای پواسون ناهمگن با تابع شدت λ به‌صورت زیر تعریف می‌شود:

$$\ell(\lambda)=logL(\lambda)=\sum_{i=1}^{n}log\lambda(s_i)-\int_D\lambda(s)\,ds, \quad (۴)$$

که درآن $\lambda(s_i)$ با رابطه‌ی (۲) داده می‌شود و از بخش ثابت $\int_D 1\,ds=|D|$ صرف‌نظر شده‌است.

عبارت فوق شامل انتگرال بر روی دامنه‌ی مشاهده D است که در عمل محاسبه‌ی آن به سختی امکان‌پذیر بوده لذا نیاز به

[6] Hierarchical clustering
[7] Dendrogram
[8] Bayesian information criteria

تقریب عددی است. برای این منظور، استراتژی گسسته‌سازی دامنه‌ی فضایی D به N پیکسل کوچک با مساحت‌های برابر با a را به‌کار می‌بریم (بدلی و همکاران [۲۱]) و متغیرهای شمارشی که تعداد نقاط در هر پیکسل را نشان می‌دهد، در نظر می‌گیریم به‌طوری‌که $\sum_{j=1}^{N} n_j = n$. در نتیجه بخش انتگرال رابطه‌ی (۴)، با مجموع روی پیکسل‌ها تقریب زده می‌شود. همچنین در این روش مکان دقیق نقاط داده را نادیده می‌گیریم و هر نقطه داده را متناظر با مرکز پیکسلی که در آن قرار دارد در نظر می‌گیریم، به این معنا که اگر u_j مرکز پیکسل j–ام باشد، آن‌گاه کلیه نقاط در پیکسل j–ام را متناظر با u_j در نظر می‌گیریم. بنابراین رابطه‌ی (۴) به‌صورت زیر بازنویسی می‌شود:

$$(\lambda) = logL(\lambda) \approx \sum_{j=1}^{N} [n_j log\lambda(u_j) - \lambda(u_j)a], \quad (۵)$$

همان‌طور که دیده می‌شود رابطه‌ی (۵) معادل لگاریتم تابع درستنمایی متغیرهای تصادفی پواسون مستقل N_j با میانگین $a\lambda(u_j)$ است. با استفاده از روابط (۲) و (۳) خواهیم داشت:

$$\mu(u_j) = a\lambda(u_j) \approx a\, exp\left(\sum_{k=1}^{m} \theta_k \varphi_k(u_j)\right)$$
$$= exp(loga + \boldsymbol{\theta}^T \boldsymbol{\varphi}_j), \quad (۶)$$

رابطه‌ی فوق، رگرسیون پواسون لگ خطی با متغیرهای $\boldsymbol{\varphi}_j = \boldsymbol{\varphi}(\|\boldsymbol{Z}(u_j) - \boldsymbol{c}_k\|)$، بردار ضرایب رگرسیون $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_m)^T$ و یک افست که وابسته به مساحت پیکسل‌ها (یعنی $loga$) است، می‌باشد. منظور از افست، عبارتی در پیش‌بینی خطی است که هیچ پارامتری از مدل را شامل نمی‌شود. به این معنا که اثر آن از قبل معلوم و شناخته شده‌است، در این‌جا مساحت پیکسل‌ها به‌عنوان یک متغیر شناخته‌شده که بر میزان شدت رویش درختان اثر دارد تحت عنوان افست در مدل‌ها آمده‌است.

مقدار بهینه‌ی پارامتر شکل γ با کمینه کردن ملاک اطلاع بیزی (BIC) در برازش مدل‌های خطی تعمیم یافته[9] تعیین می‌شود که به‌صورت زیر بیان می‌شود:

$$BIC(\gamma) = -2\ell\left(\hat{\theta}(\gamma)\right) + (m+1)\, log(N), \quad (۷)$$

روش پیشنهاد شده برای براورد ناپارامتری تابع شدت را می‌توان به صورت زیر خلاصه کرد:

الگوریتم براورد:

فرایند نقطه‌ای پواسون $\boldsymbol{Y}(\cdot)$ و متغیرهای مستقل $\boldsymbol{Z}$ بر روی پیکسل‌ها به‌عنوان ورودی در نظ گرفته‌می‌شود:

- با تعیین تعداد بهینه‌ی خوشه‌ها، m، و مراکز آن‌ها $\{\boldsymbol{c}_k\}_1^m$، پیکسل‌ها براساس متغیرهای $\boldsymbol{Z}$ خوشه‌بندی می‌شوند.
- مقدار اولیه برای γ در نظر گرفته‌می‌شود.
- پایه‌های شعاعی $\boldsymbol{\varphi}_j$ به‌عنوان تبدیلی از متغیرهای اصلی با فرمول $\boldsymbol{\varphi}_j = \boldsymbol{\varphi}(\|\boldsymbol{Z}(u_j) - \boldsymbol{c}_k\|)$ محاسبه می‌شوند.
- پنجره‌ی مشاهدات D به شبکه‌های کوچکی از پیکسل‌ها با مساحت a تقسیم می‌شود.
- تعداد نقاط داده n_j که در پیکسل j–ام قرار دارند شمارش می‌شوند.
- بردار متغیر $\boldsymbol{\varphi}_j$ در مرکز پیکسل j–ام محاسبه می‌شوند.
- از نرم افزار آماری استاندارد (مانند تابع glm در R) برای برازش (به روش حداکثر درستنمایی) یک مدل رگرسیون پواسون لگ خطی با پاسخ‌های n_j، متغیرهای $\boldsymbol{\varphi}_j$ (پایه‌های شعاعی) و جزء افست $loga$ استفاده می‌شود.
- ضرایب برازش داده‌شده $\widehat{\boldsymbol{\theta}}$ برای رگرسیون پواسون لگ خطی، براوردهای حداکثر درستنمایی تقریبی $\widehat{\boldsymbol{\theta}}$ برای مدل فرایند نقطه پواسون لگ خطی هستند.
- براورد حداکثر درستنمایی $\hat{\boldsymbol{\lambda}}$ برای مدل فرایند نقطه‌ای پواسون از روی $\widehat{\boldsymbol{\theta}}$ محاسبه می‌شود.
- BIC مدل رگرسیونی برای γ داده‌شده، محاسبه می‌شود.
- مراحل فوق تکرار می‌شوند تا حداقل مقدار BIC حاصل شود. مقدار γ متناظر با آن به‌عنوان مقدار بهینه‌ی پارامتر شکل در نظر گرفته‌می‌شود و γ^* نامیده می‌شود.
- با قرار دادن γ^* در محاسبات، براورد λ به‌دست می‌آید.

[9] Generalized linear models

۳. مطالعه شبیه‌سازی

ما از یک چارچوب مجانبی دامنه‌ی فضایی افزایشی برای ارزیابی عملکرد روش پیشنهادی‌مان استفاده کردیم. دامنه‌های فضایی $[0, 10]^2$، $[0, 20]^2$ و $[0, 30]^2$ در نظر گرفته شد. پنج متغیر مستقل $\boldsymbol{Z}(.)$ با وضوح پیکسل ۱ × ۱ از میدان تصادفی گاوسی با میانگین صفر و تابع کوواریانس نمایی

$$C_{(\sigma^2,\phi)}(h) = cov[Z_i(s), Z_i(s+h)] = \sigma^2 \exp(-\|h\|/\phi), \qquad i = 1, \dots, 5$$

تعریف شد، که در آن h فاصله‌ی اقلیدسی بین دو مکان در D است. σ^2 و ϕ پارامترهایی هستند که به ترتیب واریانس و دامنه‌ی همبستگی فضایی بین مکان‌ها را کنترل می‌کنند و مقدار آن‌ها $\sigma^2 = 0.05$ و $\phi = 0.5$ در نظر گرفته‌شد.

۵۰۰ الگوی نقطه‌ای فضایی از یک فرایند نقطه‌ای پواسون بر اساس مدل‌های شدت لگ خطی که در ادامه می‌آیند، تولید شد:

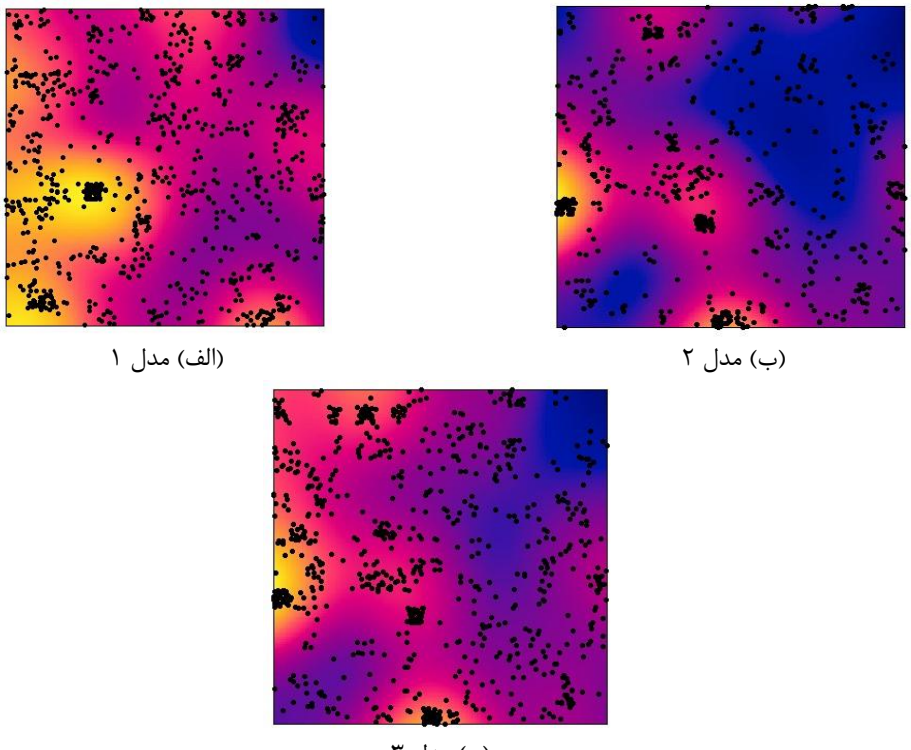
مدل ۱ : $\lambda_1(s) = exp\left(f_1(\boldsymbol{Z}(s))\right)$
$= exp\left(1.5Z_1(s) + 0.5Z_2(s) + 1.1Z_5(s)\right)$

مدل ۲: $\lambda_2(s) = exp\left(f_2(\boldsymbol{Z}(s))\right)$
$= exp(2Z_1(s) - 0.3Z_3^2(s) - Z_4^2(s))$

مدل ۳: $\lambda_3(s) = exp\left(f_3(\boldsymbol{Z}(s))\right)$
$= exp(2Z_2(s)/\{1 + 0.3Z_5^2(s)\})$

شکل ۱ الگوهای نقطه‌ای شبیه‌سازی شده‌ی حاصل از مدل‌های ۱-۳ برای فرایند نقطه‌ای پواسون در دامنه فضایی $[0, 20]^2$ را نشان می‌دهد. همان‌طور که دیده‌می‌شود، انواع مختلف ناهمگنی در نتیجه‌ی استفاده از مدل‌های مختلف برای λ به‌دست آمده است.

هدف ما براورد ناپارامتری تابع شدت است. برای این منظور، ابتدا هر تابع نامعلوم $f_j,\ j = 1,2,3$ را با یک تابع پایه شعاعی گاوسی تقریب می‌زنیم. برای این منظور ابتدا می‌بایست تعداد بهینه‌ی خوشه‌ها و مراکز خوشه‌ها برای هر دامنه‌ی فضایی تعیین شود. با توجه به این‌که مقدار متغیرهای کمکی روی پیکسل‌ها در نظر گرفته شده‌اند (به این معنا که مقدار $\boldsymbol{Z}$ روی هر پیکسل مقدار ثابتی است)، لذا برای خوشه‌بندی داده‌های مربوط به متغیرهای کمکی می‌بایست



(الف) مدل ۱ (ب) مدل ۲

(ج) مدل ۳

شکل ۱. نمونه‌هایی از تحقق فرایندهای نقطه‌ای پواسون روی ناحیه‌ی فضایی $[0, 20]^2$ با توابع شدت معرفی شده در مدل‌های ۱-۳

پیکسل‌ها خوشه‌بندی شوند. با توجه به بزرگ بودن حجم نمونه‌ها (تعداد پیکسل‌ها)، ابتدا نمونه‌ها به زیر گروه‌های کوچکتر (حدود ۱۰۰پیکسل در هر گروه) تقسیم‌بندی شد، سپس با استفاده از روش خوشه‌بندی سلسله مراتبی برای هر زیرنمونه (subsample)، دندروگرام رسم و تعداد خوشه‌ها تعیین شد. در نهایت با در نظر گرفتن بیشینه‌ی مقادیر بین تمام زیرنمونه‌ها، مقدار m بهینه برای هر دامنه‌ی فضایی به‌دست آمد و به‌عنوان مقدار ورودی جهت استفاده از روش خوشه‌بندی k-means استفاده‌شد.

علاوه بر این، مقدار بهینه‌ی پارامتر شکل γ، ۰/۱ در نظر گرفته شد. به‌منظور ارزیابی عملکرد روش به‌کار رفته، از میانگین جمع بسته‌ی توان دوم خطا[10] (MISE) استفاده شد که به صورت زیر تعریف می‌شود:

$$MISE = E\left[\int_D \left(\hat{\lambda}(s) - \lambda_0(s)\right)^2 ds\right], \qquad (۸)$$

که در آن λ_0 و $\hat{\lambda}$ به ترتیب شدت واقعی و براورد آن است.

در جدول ۱ تعداد بهینه‌ی خوشه‌ها و میانگین تعداد نقاط در ۵۰۰ الگوی نقطه‌ای شبیه‌سازی شده برای هر دامنه‌ی فضایی ارائه شده‌است. علاوه براین میزان اریبی و MISE تابع شدت براورد شده برای هر سه مدل تابع شدت در دامنه‌های فضایی مختلف محاسبه شده‌است. یافته‌های جدول ۱ نشان می‌دهد با افزایش دامنه فضایی D، میزان MISE براورد ناپارامتری تابع شدت کاهش می‌یابد. به‌عبارت دیگر، با افزایش دامنه‌ی فضایی، براورد شدت بهبود می‌یابد.

جدول ۱. نتایج شبیه‌سازی: تعداد بهینه‌ی خوشه‌ها، میانگین تعداد پیشامدها، اریبی و MISE برای براوردگر شدت در الگوهای نقطه‌ای فضایی برای دامنه‌های فضایی $[۰، ۱۰]^2$، $[۰، ۲۰]^2$ و $[۰، ۳۰]^2$ با ۵۰۰ تکرار

دامنه‌ی فضایی		$[۰، ۱۰]^2$	$[۰، ۲۰]^2$	$[۰، ۳۰]^2$
تعداد بهینه خوشه‌ها		۷	۱۴	۲۰
مدل ۱	تعداد پیشامدها	۲۶۰/۷	۱۰۱۲/۳	۲۳۰۰/۹
	اریبی	-۰/۰۰۱	-۰/۰۰۳	۰/۰۰۱
	MISE	۰/۸۱۵	۰/۳۹۸	۰/۳۲۵
مدل ۲	تعداد پیشامدها	۱۶۷/۹	۶۷۵/۹	۱۵۳۴/۱
	اریبی	-۰/۰۰۳	-۰/۰۰۹	-۰/۰۰۲
	MISE	۰/۹۹۱	۰/۸۵۲	۰/۲۶۴
مدل ۳	تعداد پیشامدها	۲۲۶/۷	۹۰۶/۴	۲۰۳۹/۷
	اریبی	۰/۰۰۳	-۰/۰۰۲	-۰/۰۰۴
	MISE	۰/۹۲۲	۰/۷۹۳	۰/۵۵۵

همان‌طور که دیده می‌شود، در بزرگترین دامنه‌ی فضایی $[۰، ۳۰]^2$، دقیق‌ترین براورد حاصل می‌شود و در دامنه‌ی فضایی $[۰، ۱۰]^2$، که در آن میانگین تعداد پیشامدها در هر الگوی نقطه‌ای فضایی کم بود، مقدارMISE براوردها نسبت به سایر دامنه‌ها بزرگتر بود.

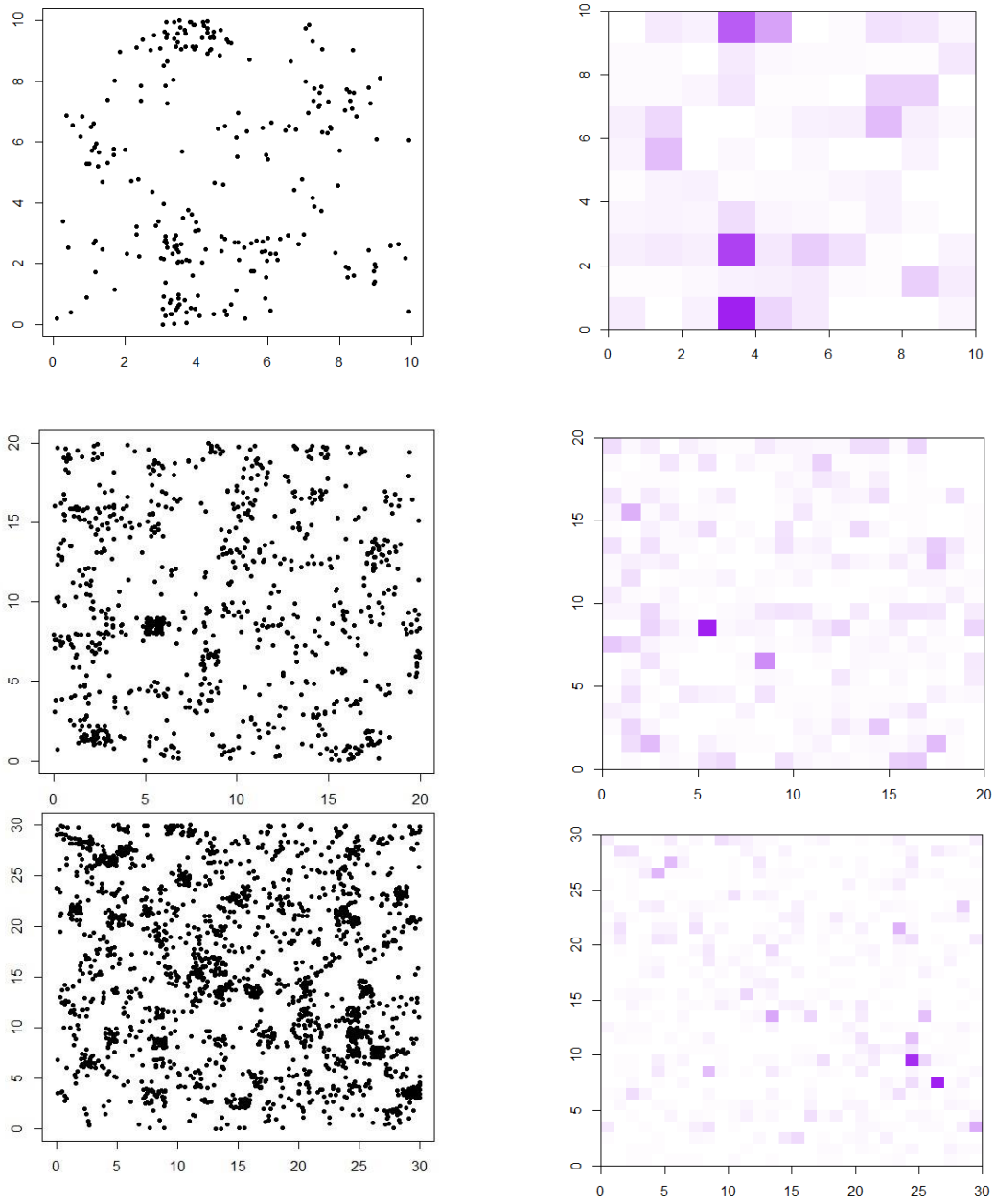
شکل ۲ نمونه‌هایی از تحقق فرایندهای نقطه‌ای پواسون با تابع شدت تعریف شده در مدل ۱، همراه با توابع شدت براورد شده، در دامنه‌های فضایی مختلف را نشان می‌دهد. نتایج عددی و گرافیکی حاصل از مطالعه‌ی شبیه‌سازی نشان می‌دهد، روش ناپارامتری پیشنهادی برای براورد تابع شدت فرایند نقطه‌ای پواسون روش مناسبی است.

۴. کاربرد در مجموعه داده‌های جنگلداری

۱.۴. توصیف داده‌ها

مجموعه داده‌های مورد بررسی از مجموعه داده‌ی بزرگی از موقعیت درختان جنگل‌های گرمسیری در پاناما هستند که در محیطی با طول ۱۰۰۰ متر و عرض ۵۰۰ متر جمع‌آوری

[10] Mean integrated square error (MISE)



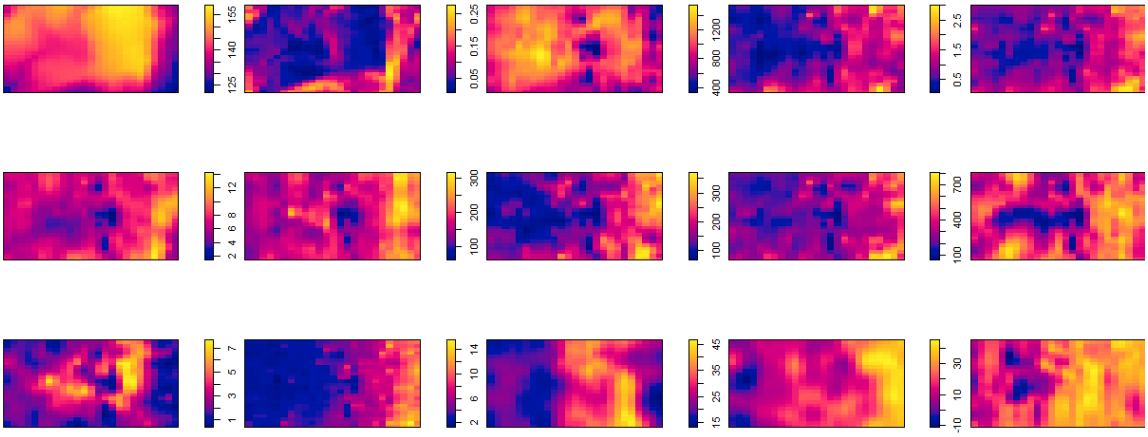
شکل ۲. تحقق فرایندهای نقطه‌ای پواسون براساس تابع شدت مدل ۱ (تعریف شده در بخش ۳) و براورد آن در دامنه‌های مختلف.

شده‌اند. داده‌های موجود شامل بیش از ۳۰۰ گونه از درختان مختلف است (به هابل و فاستر [۲۲]؛ کاندیت و همکاران[۲۳]؛ کاندیت[۲۴] مراجعه کنید).

جمع‌آوری متغیرهای محیطی مانند توپوگرافی و مواد مغذی خاک در مطالعات بوم‌شناختی امری رایج است، زیرا آن‌ها می‌توانند به‌طور قابل توجهی بر توزیع و رشد گونه‌های گیاهی تأثیر بگذارند، بنابراین متغیرهایی مانند ارتفاع، شیب (متغیرهای توپوگرافی)، آلومینیوم، برم، کلسیم، مس، آهن، پتاسیم، منیزیم، منگنز، فسفر، روی، نیتروژن، کانی‌سازی نیتروژن و میزان اسیدیته خاک (مواد مغذی خاک) در این ناحیه به‌صورت شبکه‌ای اندازه‌گیری و جمع‌آوری شده‌اند. مقادیر متغیرهای کمکی در وضوح مختلف پیکسل‌ها، با استفاده از میانگین مقادیر روی پیکسل‌ها محاسبه شده‌اند.

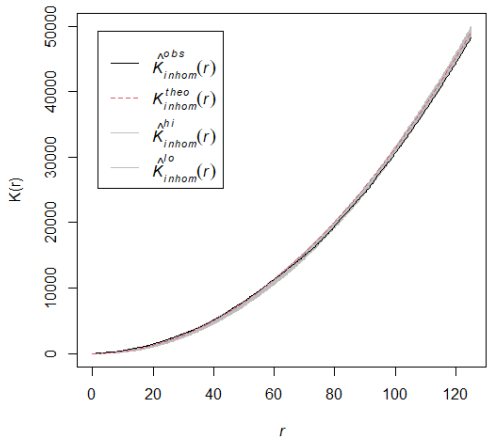
نقشه‌ی متغیرهای محیطی بر روی ناحیه‌ی مشاهده شده در پیکسل‌هایی با ابعاد ۴۰ متر × ۲۰ متر در شکل۳ ارائه شده

است (نقشه‌های متغیرهای کمکی در وضوح پیکسل‌های ۵۰ متر× ۵۰ متر و ۲۰ متر × ۲۰ متر در پیوست آمده‌است).

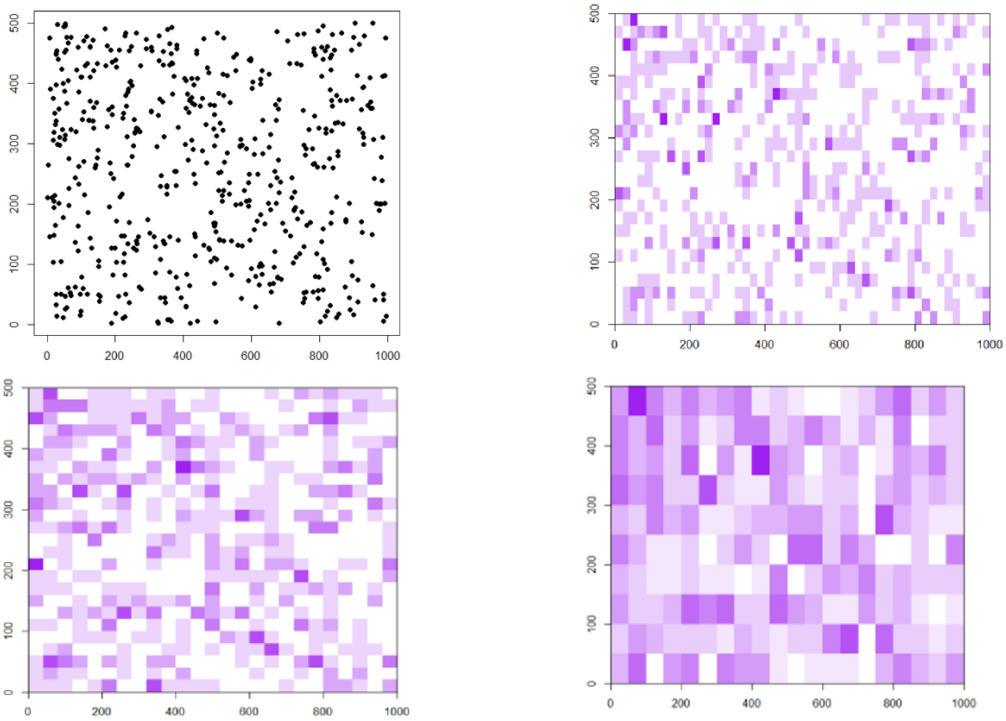


شکل ۳. نقشه‌ی میانگین مقدار متغیرهای کمکی مورد نظر با وضوح پیکسل‌های ۴۰ متر× ۲۰ متر. از چپ به راست: ارتفاع، شیب، آلومینیوم، برم و کلسیم (ردیف ۱)؛ مس، آهن، پتاسیم، منیزیم و منگنز (ردیف دوم)؛ فسفر، روی، نیتروژن، کانی‌سازی نیتروژن و میزان اسیدیته خاک (ردیف سوم). رنگ‌های روشن‌تر نشان دهنده‌ی مقادیر بزرگتر متغیرها هستند.

در این مقاله گونه‌ای از درختان زنده در جنگل گرمسیری با نام "اینگاساپیندوئیدس" (۵۴۵ درخت) مورد بررسی قرار گرفت. شکل ۴ تابع K را به همراه پوش‌های ۹۵٪ بوت استرپ نقطه به نقطه برای مکان درختان اینگاساپیندوئیدس نشان می‌دهد. همان‌طور که در شکل مشاهده می‌شود، الگوی نقطه‌ای موردنظر از فرایند نقطه‌ای پواسون ناهمگن پیروی می‌کند.



شکل ۴. تابع K_{inhom} برای مکان درختان اینگاساپیندوئیدس با پوش‌های ۹۵٪ بوت استرپ نقطه به نقطه. خط نقطه چین قرمز رنگ مقدار K_{inhom} برای فرایند نقطه‌ای پواسون است.



شکل ۵. مکان درختان اینگاساپیندوئیدس (بالا سمت چپ) و تعداد درختان با وضوح پیکسل‌های ۲۰ متر × ۲۰ متر (بالا سمت راست)، ۴۰ متر × ۲۰ متر (پایین سمت چپ)، و ۵۰ متر × ۵۰ متر (سمت راست پایین). رنگ‌های تیره نشان دهنده‌ی مقادیر بزرگتر است.

شکل ۵ مکان این درختان و تعداد آن‌ها را در وضوح مختلف پیکسل‌ها نشان می‌دهد.

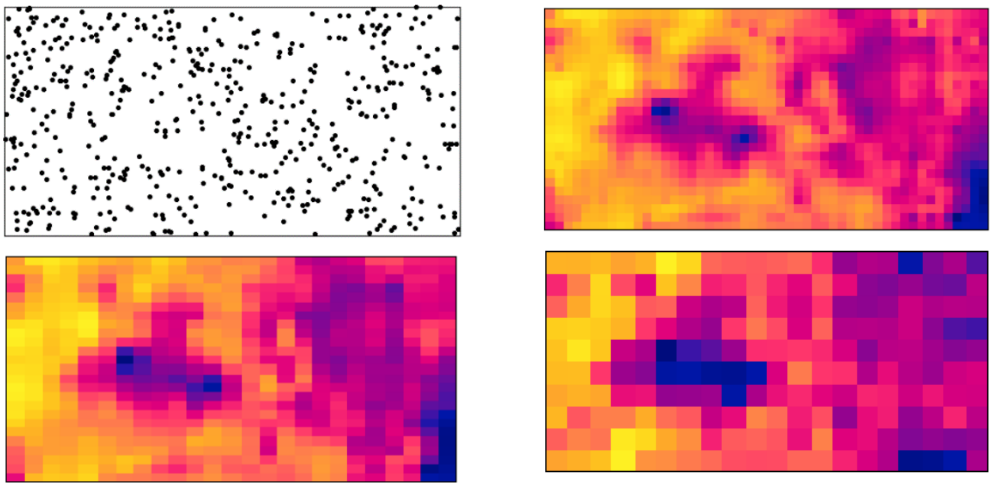
۲.۴. پیاده‌سازی روش پیشنهادی با استفاده از روش تقریب پیکسل‌های کوچک

با جمع‌آوری اطلاعات تفصیلی درباره‌ی عوامل محیطی، پژوهشگران به دنبال مدل‌سازی شدت (یا تراکم) گونه‌های مختلف درختان به‌عنوان تابعی از این متغیرهای کمکی هستند. این رویکرد می‌تواند به درک عوامل محیطی و مکانیزم‌های زیربنایی که توزیع فضایی این گونه‌های گیاهی را در جنگل‌های گرمسیری شکل می‌دهند، کمک کند. در این‌جا دو مدل برای تابع شدت در نظر می‌گیریم و به‌طور خاص تاثیر عوامل محیطی بر توزیع فضایی مکان درختان "اینگاساپیندوئیدس" را بررسی می‌کنیم. مدل اول شامل متغیرهای ویژگی‌های خاک و توپوگرافی است، در حالی‌که مدل دوم تنها شامل ویژگی‌های خاک است، و به کمک آن‌ها و با استفاده از روش پیشنهاد شده به براورد تابع شدت می‌پردازیم.

در گام نخست می‌بایست توابع چندمتغیره را با استفاده از RBF تقریب بزنیم. برای این منظور ابتدا باید خوشه‌بندی پیکسل‌ها را براساس متغیرهای کمکی موجود در پیکسل‌ها انجام دهیم. از آن‌جایی که روش‌های تحلیل خوشه‌ای مبتنی بر فاصله هستند، استانداردسازی باعث کاهش اثر مقیاس و بهبود عملکرد الگوریتم محاسباتی RBF می‌شود و از بروز نتایج گمراه‌کننده و اریب در خوشه‌بندی جلوگیری می‌کند. استانداردسازی برای هر متغیر کمکی با کم کردن میانگین متغیر از مقدار هر داده و تقسیم بر انحراف معیار متغیر موجود انجام شد.

در ادامه با به‌کارگیری تقریب پیکسل‌های کوچک، و شمارش تعداد درختان اینگاساپیندوئیدس در هر پیکسل، مدل فرایند نقطه‌ای را به مدل رگرسیون پواسون تبدیل می‌کنیم و با براورد ضرایب مدل رگرسیونی، $\hat{\boldsymbol{\theta}}$، براورد تابع شدت، $\hat{\boldsymbol{\lambda}}$، را تحت ۲ مدل مذکور به‌دست می‌آوریم. دقت شود با توجه به این‌که براوردها تحت ۳ وضوح مختلف (۲۰ متر × ۲۰ متر، ۴۰ متر × ۲۰ متر و ۵۰ متر × ۵۰ متر) محاسبه می‌شوند، لذا مساحت پیکسل‌ها متفاوت بوده و به‌عنوان افست در مدل‌ها و براوردها لحاظ شده‌است.

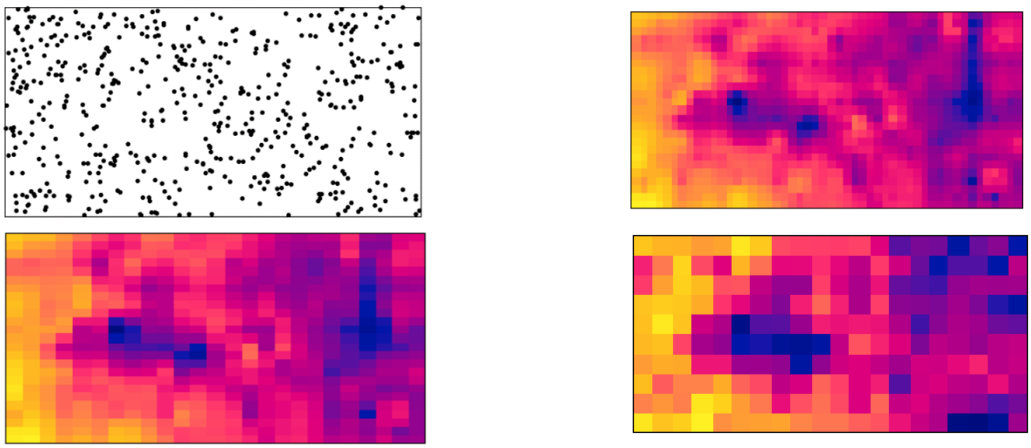
برای ارزیابی میزان دقت تابع شدت براورد شده از روش بصری استفاده شد. نتایج شدت براورد شده حاصل از دو مدل مذکور در شکل ۶ و ۷ ارائه شده‌است. با مقایسه‌ی تصاویر شدت براورد شده (در وضوح مختلف پیکسل‌ها) با الگوی نقطه‌ای واقعی (بالا سمت چپ)، میزان همسویی براوردها با داده‌های واقعی قابل مشاهده است. این تصاویر با استفاده از یک طیف رنگی از آبی (چگالی کم) به زرد (چگالی بالا) رسم شده‌است و همان‌طور که در تصاویر دیده می‌شود نقاط بیشتری در مناطق زرد تصویر و نقاط کمتری در مناطق آبی وجود دارد. لذا روش پیشنهادی برای براورد چگالی درختان اینگاساپیندوئیدس مناسب است.



شکل ۶. داده‌های اینگاساپیندوئیدس مکان ۵۴۵ درخت (بالا سمت چپ)، براورد شدت اینگاساپیندوئیدس به‌عنوان تابعی از متغیرهای توپولوژیکی و ویژگی‌های خاک با استفاده از روش پیشنهادی در پیکسل‌هایی با وضوح ۲۰ متر × ۲۰ متر (بالا سمت راست)، ۴۰ متر × ۲۰ متر (چپ پایین)، و ۵۰ × ۵۰ متر (پایین سمت راست). رنگ‌های روشن‌تر نشان دهنده‌ی مقادیر بزرگتر است.

۵. نتیجه‌گیری

در عصر حاضر با پیشرفت فناوری، میزان داده‌های تولید و جمع‌آوری شده به‌طور چشمگیری افزایش یافته‌است و طیف وسیعی از متغیرهای کمکی برای تحلیل الگوهای نقطه‌ای در دسترس قرار گرفته‌اند. بنابراین، استفاده از روشی که قادر به مواجهه با تعداد زیاد متغیرهای کمکی در برآورد ناپارامتری تابع شدت باشد، ضروری به‌نظر می‌رسد.



شکل ۷. داده‌های اینگاساپیندوئیدس مکان ۵۴۵ درخت (بالا سمت چپ)، براورد شدت اینگاساپیندوئیدس به‌عنوان تابعی از مواد معدنی خاک با استفاده از روش پیشنهادی در پیکسل‌هایی با وضوح ۲۰ متر × ۲۰ متر (بالا سمت راست)، ۴۰ متر × ۲۰ متر (چپ پایین)، و ۵۰ × ۵۰ متر (پایین سمت راست). رنگ‌های روشن‌تر نشان دهنده‌ی مقادیر بزرگتر است.

این در حالی است که روش‌های ناپارامتری موجود برای براورد تابع شدت، محدودیت در تعداد متغیرهای کمکی مورد استفاده را دارند؛ یعنی تنها تعداد کمی از متغیرهای کمکی را می‌توانند در مدل لحاظ کنند. این محدودیت عمدتاً به دلیل مسئله‌ی مشقت بُعد چندی[11] می‌باشد. در این موارد، معمولاً تلاش می‌شود با استفاده از تکنیک‌های کاهش بعد، از پیچیدگی مسائل کاسته و سپس از روش‌های ناپارامتری شناخته شده برای براورد تابع شدت استفاده شود.

روش پیشنهادی ما بدون نیاز به استفاده از تکنیک‌های کاهش بُعد، صرفا با تکیه بر تقریب تابع چندمتغیره‌ی نامعلوم f با کمک توابع پایه شعاعی، به براوردی ناپارامتری برای تابع شدت با دقت قابل‌قبول دست می‌یابد. دقت مناسب در براورد تابع شدت مستلزم تقریب بادقت بالای تابع چند متغیره‌ی f است. برای این‌منظور، باید پارامترهای توابع پایه شعاعی به دقت تعیین و به‌کار گرفته‌شوند. از آن‌جا که روش‌های خوشه‌بندی برای ایجاد خوشه‌ها به الگوریتم‌های مبتنی بر فاصله وابسته هستند، که تحت تاثیر مقیاس متغیرها قرار می‌گیرند، استانداردسازی متغیرهای کمکی ضروری است. پس از استانداردسازی متغیرها، با استفاده از خوشه‌بندی سلسله مراتبی و k-means تعداد بهینه‌ی خوشه‌ها و مراکز خوشه‌ها مشخص می‌شوند. در نهایت با کمینه کردن ملاک اطلاع بیزی در مدل‌های رگرسیونی، پارامتر شکل بهینه تعیین می‌شود.

سپس با استفاده از تقریب RBF برای تابع نامعلوم f، مدل ناپارامتری تابع شدت به‌صورت یک مدل لگ خطی بیان می‌شود. در ادامه دامنه‌ی فضایی به پیکسل‌های کوچک تقسیم می‌شود و با شمارش نقاط داده در پیکسل‌ها، مسئله‌ی براورد تابع شدت در فرایندهای نقطه‌ای پواسون فضایی به رگرسیون پواسون تبدیل می‌شود و با استفاده از مدل‌های خطی تعمیم‌یافته، به براورد ضرایب θ پرداخته‌می‌شود که از روی آن، براورد ناپارامتری تابع شدت به‌دست خواهد آمد.

ما نتایج روش پیشنهادی را بر روی داده‌های شبیه‌سازی شده و همچنین مجموعه داده‌های واقعی جنگل گرمسیری اعمال کردیم. همان‌طور که در مطالعه‌ی شبیه‌سازی نشان داده‌شد، افزایش دامنه‌ی فضایی منجر به کاهش میانگین جمع بسته‌ی توان دوم خطا (MISE) می‌شود. به‌عبارت دیگر، با افزایش تعداد پیشامدها، روش ما احتمالا براوردگرهایی تولید می‌کند که به همان اندازه کارآمد هستند که گویی مدل واقعی را برازش داده‌ایم.

سپاسگزاری از حمایت مالی:

این پژوهش هیچ کمک هزینه‌ی خاصی از هیچ مؤسسه سرمایه‌گذار در بخش عمومی، تجاری یا غیرانتفاعی دریافت نکرده است.

سپاسگزاری:

نویسندگان مقاله مراتب قدردانی خود را از داوران محترم ابراز می‌دارند. نظرات ارزشمند و سازنده‌ی داوران بر غنای علمی مقاله افزود. همچنین نویسندگان از مدیریت و کارکنان مجله‌ی "مهندسی و مدیریت کیفیت" برای همکاری و حمایت صمیمانه خود تشکر می‌نمایند.

تعارض منافع:

نویسندگان اعلام می‌دارند که هیچ گونه تضاد منافعی در مطالعه‌ی حاضر وجود ندارد.

۶. مراجع

[1] Illian, J. B., Møller, J., & Waagepetersen, R. P. (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environmental and Ecological Statistics*, *16*, 389-405.

[2] Law, R., Illian, J., Burslem, D. F., Gratzer, G., Gunatilleke, C. V. S., & Gunatilleke, I. A. U. N. (2009). Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology*, *97*(4), 616-628..

[3] Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., ... & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*(4), 366-379.

[4] Stoyan, D., & Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical science*, 61-78.

[11] Curse of dimensionality

[5] Babu, G. J., & Feigelson, E. D. (1996). Spatial point processes in astronomy. *Journal of statistical planning and inference*, *50*(3), 311-326.

[6] Baddeley, A., Rubak, E., & Turner, R. (2015). *Spatial point patterns: methodology and applications with R*. CRC press.

[7] Jafari Mamaghani, M., Andersson, M., & Krieger, P. (2010). Spatial point pattern analysis of neurons using Ripley's K-function in 3D. *Frontiers in neuroinformatics*, *4*, 1285.

[8] Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *153*(3), 349-362.

[9] Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press.

[10] Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, 256-274.

[11] Díaz-Delgado, R., Lloret, F., & Pons, X. (2004). Spatial patterns of fire occurrence in Catalonia, NE, Spain. *Landscape Ecology*, *19*, 731-745.

[12] Waagepetersen, R. (2008). Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika*, *95*(2), 351-363.

[13] Choiruddin, A., Coeurjolly, J. F., & Letué, F. (2017). Spatial point processes intensity estimation with a diverging number of covariates. *arXiv preprint arXiv:1712.09562*.

[14] Choiruddin, A., Cuevas-Pacheco, F., Coeurjolly, J. F., & Waagepetersen, R. (2020). Regularized estimation for highly multivariate log Gaussian Cox processes. *Statistics and Computing*, *30*(3), 649-662.

[15] Guan, Y. (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, *103*(483), 1238-1247.

[16] Baddeley, A., Chang, Y. M., Song, Y., & Turner, R. (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and its interface*, *5*(2), 221-236.

[17] Borrajo, M. I., González-Manteiga, W., & Martínez-Miranda, M. D. (2020). Bootstrapping kernel intensity estimation for inhomogeneous point processes with spatial covariates. *Computational Statistics & Data Analysis*, *144*, 106875.

[18] Coeurjolly, J. F., Cuevas-Pacheco, F., & Descary, M. H. (2022). A convolution type model for the intensity of spatial point processes applied to eye-movement data. *Spatial Statistics*, *51*, 100651.

[19] Ward, S., Battey, H. S., & Cohen, E. A. K. (2023). Nonparametric estimation of the intensity function of a spatial point process on a Riemannian manifold. *Biometrika*, *110*(4), 1009-1021.

[20] Fasshauer, G. E. (2007). *Meshfree Approximation Methods with MATLAB*. World Scientific Pub Co Inc.

[21] Baddeley, A., Berman, M., Fisher, N.I., Hardegen, A., Milne, R.K., Schuhmacher, D., Shah, R., and Turner, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. Electron J Stat 4:1151-1201.

[22] Hubbell, S.P., and Foster, R.B. (1983). Diversity of canopy trees in a neotropical forest and implications for conservation. In: Sutton SL, Whitmore TC, Chadwick AC (ed) Tropical Rain Forest: Ecology and Management, Blackwell Scientific Publications, Oxford, pp 25–41.

[23] Condit, R., Hubbell, S. P., & Foster, R. B. (1996). Changes in tree species abundance in a neotropical forest: impact of climate change. *Journal of tropical ecology*, *12*(2), 231-256.

[24] Condit, R. (1998). Tropical Forest Census Plots. Berlin, Germany and Georgetown, Texas: Springer-Verlag and R. G. Landes Company.

A Novel Approach to Nonparametric Estimation of The Intensity Function of Spatial Poisson Point Processes and Its Application in Estimating the Intensity of Inga Sapindoides Trees

Mitra Hasheminia[12]**,**
Ph.D. Student in Statistics, Faculty of Statistics, Mathematics, and Computer University of Allameh Tabataba'i, Tehran, Iran.

Reza Pourtaher
Associate Prof., Department of Statistics, Faculty of Statistics, Mathematics, and Computer, Allameh Tabataba'i University, Tehran, Iran.

Abstract: Modelling and estimating the intensity function of a point pattern is one of the preliminary and fundamental issues in inference of point processes, and it considered as a prerequisite for many other problems. It has been addressed from different perspectives. With the rapid development of data-collection technologies, a wide range of data has been produced, and considering covariates has been a big step forward in the theory of point processes, which has mainly been addressed from a parametric perspective.

In this paper, we introduce a novel approach for nonparametrically estimating the intensity of an inhomogeneous Poisson point process, which is an unknown function of several independent spatial covariates. In the proposed method, using the approximation technique of radial basis function for unknown multivariate functions, the nonparametric model of the intensity function is transformed into a log-linear model. Since the accuracy of the multivariate function approximation directly affects the accuracy of the intensity function estimate, we enhance the nonparametric estimation quality of the intensity function in spatial Poisson point processes by optimizing the shape parameter of the radial basis function through minimizing the Bayesian information criterion.

The performance of the proposed method was evaluated using a simulation study and real data examining the intensity of a tree species, "Inga Sapindoides", from tropical forests. The results showed that our proposed method is capable of providing an accurate nonparametric estimate of the intensity function in spatial Poisson point processes. This novel approach has fundamental advantages over existing nonparametric methods, which are limited in the number of covariates that can be used.

Key Words: Approximate Maximum Likelihood Estimate; Pixel Counts; Inhomogeneous Spatial Poisson Point Process; Independent Spatial Covariates.

Aim and Introduction

Modelling the first-order intensity function is one of the main aims in point process theory, and various approaches have been explored to address this. This work focuses on the nonparametric estimation of the intensity function of an inhomogeneous Poisson point process in the important case where the intensity depends on

[12] *Corresponding Author Email: mitra_hasheminia94@atu.ac.ir

independent covariates. With the advancement of sciences, increasing amounts of data are produced and collected. The inclusion of spatially varying covariates in the models for the intensity function has been a big step forward on point process theory and it has been mostly addressed so far from a parametric perspective.
The simplest and most common parametric model is the log-linear model, which formulates the log-intensity as a linear combination of available covariates; for instance, consider the following model:

$$\lambda(s) = exp\{\theta_0 + \theta Z(s)\}, \;\; s \in D \subset \mathbb{R}^2, \qquad (1)$$

Although parametric methods are very popular and are often used, it should be noted that if the assumed parametric model is not suitable for the true intensity function, inappropriate and unreliable estimates may be obtained. Therefore, nonparametric methods, which have more flexibility and broader applicability than parametric methods, are considered a suitable alternative to parametric approaches.
So far, only a limited number of studies have been conducted to estimate the intensity function based on spatial covariates that are modeled nonparametrically. The main limitation of existing nonparametric approaches is the constraint to consider a small number of covariates, which is primarily due to the well-known "curse of dimensionality" phenomenon.
We assume the intensity function of a spatial Poisson point process that includes several observed independent covariates, which is written as follows:

$$\lambda(s) = exp\left(f\left(\boldsymbol{Z}(s)\right)\right), \qquad s \in D \subset \mathbb{R}^2, \qquad (2)$$

where f is an unknown continuous function and $\boldsymbol{Z}(s)$ is the independent covariate vector that is associated with the spatial location $s \in D$.

Methodology

In high-dimensional settings, our proposed method provides an approximation of the unknown multivariate function f by employing a radial basis function approach [2]. Substituting this approximation into equation (2) leads to the multivariate version of model (1). The resulting expression can be written as:

$$\lambda_{\boldsymbol{\theta}}(s) = exp\left(f\left(\boldsymbol{Z}(s)\right)\right) \approx exp\left(\sum_{k=1}^{m} \theta_k \varphi_k(s)\right) = exp\left(\boldsymbol{\theta}^T \boldsymbol{\varphi}_j\right). \qquad (3)$$

This model states that the intensity is $\lambda_{\boldsymbol{\theta}}(s)$ where the value of $\boldsymbol{\theta}$ is to be estimated. The loglikelihood for $\boldsymbol{\theta}$ is:

$$\ell(\boldsymbol{\theta}) = logL(\boldsymbol{\theta}) = \sum_{i=1}^{n} log\, \lambda_{\boldsymbol{\theta}}(s_i) - \int_D \lambda_{\boldsymbol{\theta}}(s)\, ds, \qquad (4)$$

The likelihood function of a Poisson point process involves an integral over the spatial window D. This means that the likelihood cannot be computed exactly, but must be approximated numerically [1]. A quadrature strategy for approximating the Poisson process likelihood is to divide the window D into small pixels of equal area a. The integral over the window D is then approximated by a sum over pixels. We also

instead of using the exact locations of the data points $s_1, \dots, s_n$, just we count the number of data points in each pixel. Thus we approximate equation (4) by:

$$\ell(\boldsymbol{\theta}) = logL(\boldsymbol{\theta}) \approx \sum_{j=1}^{N} [n_j log\ \lambda_{\boldsymbol{\theta}}(u_j) - \lambda_{\boldsymbol{\theta}}(u_j)a], \qquad (5)$$

The right-hand side of (5) has the same form as the loglikelihood of independent Poisson random variables N_j with means $a\lambda(u_j)$.

From equation (5), we can write the Poisson regression model as:

$$\mu_{\boldsymbol{\theta}}(u_j) = a\ \lambda_{\boldsymbol{\theta}}(u_j) \approx\ exp\left(\sum_{k=1}^{m} \theta_k \varphi_k(u_j)\right) = exp(loga + \boldsymbol{\theta}^T \boldsymbol{\varphi}_j), \qquad (6)$$

where the radial basis $\boldsymbol{\varphi}_j$ are transformed original variables, and the coefficients $\boldsymbol{\theta}$ need to be estimated. $loga$ is offset term, this means the effect of the pixel area on the intensity is considered a known, fixed factor, rather than a parameter that needs to be estimated by the model. Incorporating known variables as offsets can improve the accuracy of the intensity function estimation, as it allows the model to focus on learning the other, unknown relationships.

By using standard statistical software (such as the glm function in R) we fit (by maximum likelihood) a loglinear Poisson regression model with responses N_j, regression covariates $\boldsymbol{\varphi}_j$, and offsets $loga$. The fitted coefficients $\widehat{\boldsymbol{\theta}}$ for the loglinear Poisson regression are the approximate maximum likelihood estimates for the loglinear Poisson point process model (3). And now we can calculate the approximate maximum likelihood estimates ،$\hat{\boldsymbol{\lambda}}$ for the Poisson point process model from $\widehat{\boldsymbol{\theta}}$.

It should be noted that, since the accuracy of the multivariate function approximation directly affects the accuracy of the intensity function estimate, we enhance the nonparametric estimation quality of the intensity function in spatial Poisson point processes by optimizing the parameters related to RBFs such as, number of clusters, the cluster centers and shape parameter.

When the number of clusters is not known in advance, a combination of hierarchical clustering and k-means clustering can be used. Firstly, hierarchical clustering is utilized to generate a dendrogram and determine the optimal number of clusters, denoted as, m. However, due to the large sample size N (a large number of pixels), in the initial step we divided the samples into smaller subgroups (each containing about 100 pixels per group). Subsequently, a dendrogram was constructed for each subsample to determine the optimal number of clusters. Finally, by aggregating the maximum values from all subsamples, the optimal value for m in the spatial domain was determined and used as an input parameter for implementing the k-means clustering method. In this manner, the optimal values for the number of clusters and cluster centers were determined. It is important to note that we initially standardized the covariates before proceeding with clustering. This step is crucial to ensure that all variables are on a similar scale and to avoid biasing the clustering results towards variables with larger ranges.

In this study, we utilized the Bayesian information criterion (BIC) to compute the optimal value of the shape parameter. This was achieved by minimizing the BIC obtained from fitting a generalized Poisson regression model, as expressed below:

$$BIC(\gamma) = -2\ell\left(\hat{\theta}(\gamma)\right) + (m+1)\ log(N), \qquad (7)$$

Findings

The performance of the proposed method was evaluated using a simulation study and real data examining the intensity of a tree species, "Inga Sapindoides", from tropical forests.
In the simulation study, we considered different types of models for the intensity function. By using the mean integrated square error (MISE) of the estimations, we demonstrated that our proposed method is capable of providing an accurate nonparametric estimate of the intensity function in spatial Poisson point processes. Additionally showed that as the observed window size was increased, the MISE of the estimator decreased.

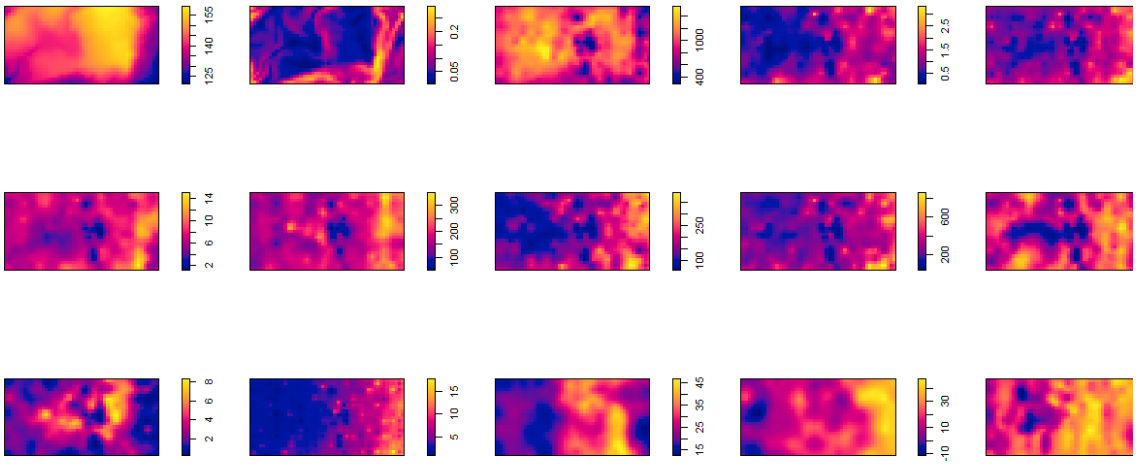
Discussion and Conclusion

This novel approach has fundamental advantages over existing nonparametric methods, which are limited in the number of covariates that can be used. In other words, the available nonparametric methods for estimating the intensity function can only incorporate a small number of covariates, but in the proposed method, without using dimensionality reduction techniques and simply by relying on the approximation of the unknown multivariate function using radial basis functions, we are able to perform a nonparametric estimation of the intensity function of the spatial Poisson point process in the presence of several independent spatial covariates.
Proper accuracy in estimating the intensity function requires a high-accuracy approximation of the unknown multivariate function f. For this purpose, the parameters of the radial basis functions must be carefully determined and used. Since clustering methods depend on distance-based algorithms to create clusters, which are affected by the scale of variables, standardization of covariates is necessary. After standardizing the variables, using hierarchical clustering and k-means, the optimal number of clusters and cluster centers are determined. Finally, by minimizing the Bayesian information criterion in regression models, the optimal shape parameter is determined.

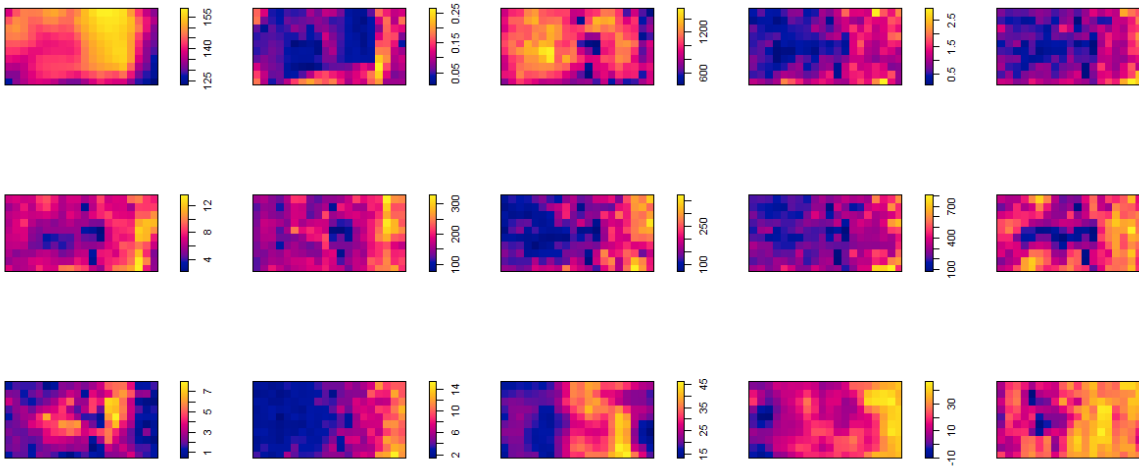
Reference

[1] Baddeley, A., Berman, M., Fisher, N.I., Hardegen, A., Milne, R.K., Schuhmacher, D., Shah, R. and Turner, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. Electron J Stat 4:1151-1201.
[2] Fasshauer, G.E. (2007). Meshfree approximation methods with MATLAB (Vol. 6). World Scientific.

پیوست:



شکل ۸. نقشه‌ی میانگین مقدار متغیرهای کمکی مورد نظر در وضوح پیکسل ۲۰ متر× ۲۰ متر. از چپ به راست: ارتفاع، شیب، آلومینیوم، برم و کلسیم (ردیف ۱)؛ مس، آهن، پتاسیم، منیزیم و منگنز (ردیف دوم)؛ فسفر، روی، نیتروژن، کانی‌سازی نیتروژن و میزان اسیدیته (ردیف سوم). رنگ‌های روشن تر نشان دهنده‌ی مقادیر بزرگتر متغیرها هستند.



شکل ۹. نقشه‌ی میانگین مقدار متغیرهای کمکی مورد نظر در وضوح پیکسل ۵۰ متر× ۵۰ متر. از چپ به راست: ارتفاع، شیب، آلومینیوم، برم و کلسیم (ردیف ۱)؛ مس، آهن، پتاسیم، منیزیم و منگنز (ردیف دوم)؛ فسفر، روی، نیتروژن، کانی‌سازی نیتروژن و میزان اسیدیته (ردیف سوم). رنگ‌های روشن تر نشان دهنده‌ی مقادیر بزرگتر متغیرها هستند.