Journal of Statistical Sciences, Spring and Summer, 2025
Vol. 19, No. 1, pp 29-55
DOI: 10.61186/jss.19.1.8

Analysis of Housing Prices in Mashhad City with a Two-Stage Heterogeneous Spatial Modeling Framework

Beheshty, A.[1], Baghishani, H.[2], Behzadi, M. H.[1], Yari, G.[3], Turk, D.[4]

[1]Department of Statistics, Science and Research Branch, Islamic Azad University, Tehran, Iran.
[2]Department of Statistics, Shahrood University of Technology, Shahrood, Iran.
[3]Faculty of Mathematical Sciences, Iran University of Science and Technology, Tehran, Iran.
[4]Department of Mathematics, Lafayette College, Easton, USA.
Corresponding author: H. Baghishani, hbaghishani@shahroodut.ac.ir

Received: 19/11/2024 **Revised:** 14/3/2025 **Accepted and Published Online:** 17/3/2025.

Introduction

Housing prices often exhibit spatial correlation and heterogeneity as key economic and financial indicators. Spatial econometric models are widely used to analyze dependencies but struggle to model spatial heterogeneity effectively. The traditional geographically weighted regression is used to evaluate local heterogeneity; however, it can become quite complex and computationally intensive, especially in areas with homogeneous subregions. Such a challenge emphasizes finding a more efficient and effective solution. This study introduces a two-stage spatial econometric modeling framework for housing price analysis. In the first stage, a spatial clustering algorithm is used to identify homogeneous regions. In the second stage, spatial econometric models are fitted separately for each cluster. This approach employs the integrated nested Laplace approximation as an approximate Bayesian method for conducting statistical inference and analysis.

Material and Methods

The proposed methodology consists of two main steps. First, the adaptive weighted smoothing algorithm is used to identify homogeneous regions, and this identification process is iteratively optimized to differentiate between heterogeneous structures. Second, Bayesian spatial econometric models are

fitted separately for each identified cluster, ensuring that spatial dependencies and heterogeneity are appropriately addressed. Bayesian inference uses the integrated nested Laplace approximation method, which is more computationally efficient than traditional approaches such as Markov chain Monte Carlo. A simulation study was conducted to evaluate the effectiveness of the proposed methodology. Following this validation, the approach was applied to analyze housing prices in Mashhad City.

Results and Discussion
The simulation study results show that the proposed method effectively identifies spatial heterogeneous structures while preserving spatial dependence. Additionally, it outperforms the geographically weighted regression model in terms of accuracy and computational efficiency. An analysis of the Mashhad housing price dataset reveals two distinct clusters, each with different regression coefficients for significant covariates. Furthermore, the spatial dependence structure differs between these clusters: one cluster shows a positive spatial correlation, while the other indicates a negative spatial correlation. These findings emphasize the necessity of simultaneously modeling spatial heterogeneity and dependence for a more comprehensive housing market analysis.

Conclusion
This study introduces a Bayesian two-stage framework for modeling spatially heterogeneous data. By integrating spatial clustering algorithms with Bayesian spatial econometric models, the approach improves analytical accuracy, decreases computational complexity, and offers more interpretable insights into the effects of covariates. Future research could include incorporating spatially varying coefficient models for a more detailed analysis of local heterogeneity and applying machine learning methods in spatial clustering to enhance the detection of homogeneous regions.

Keywords: Spatial econometrics, Bayesian inference, Spatial clustering, Housing market analysis.

Mathematics Subject Classification (2010): 62H11, 62J12.

مجله علوم آماری، بهار و تابستان ۱۴۰۴
جلد ۱۹، شماره ۱، ص ۲۹ -- ۵۵
DOI: 10.61186/jss.19.1.8
مقاله پژوهشی

تحلیل قیمت مسکن شهر مشهد با یک چارچوب مدل‌بندی فضایی ناهمگن دومرحله‌ای

علیرضا بهشتی[۱]، حسین باغیشنی[۲]، محمدحسن بهزادی[۱]، غلامحسین یاری[۳]، دنیل تورک[۴]

[۱]گروه آمار، شاخه علوم و تحقیقات دانشگاه آزادی اسلامی، تهران
[۲]گروه آمار، دانشگاه صنعتی شاهرود، شاهرود
[۳]دانشکده علوم ریاضی، دانشگاه علم و صنعت ایران، تهران
[۴]گروه ریاضی، کالج لافایته، ایستون، آمریکا

نویسنده مسئول : حسین باغیشنی، hbaghishani@shahroodut.ac.ir

تاریخ دریافت: ۱۴۰۳/۸/۲۹ تاریخ بازنگری: ۱۴۰۳/۱۲/۲۴ تاریخ پذیرش و انتشار: ۱۴۰۳/۱۲/۲۷

چکیده: داده‌های حاصل از اندازه‌گیری شاخص‌های مالی و اقتصادی، مانند قیمت مسکن، عموما به‌طور فضایی همبسته و ناهمگن هستند. مدل‌های اقتصادسنجی فضایی برای لحاظ کردن وابستگی موجود در این داده‌ها پرطرفدار هستند. اما مدل‌بندی کارای ناهمگنی فضایی هنوز مورد سوال است. معمولا از رگرسیون وزنی جغرافیایی برای مدل‌بندی ناهمگنی موضعی داده‌های فضایی استفاده می‌شود. این رده از مدل‌ها برای داده‌های فضایی همگن در چند زیرناحیه، بیش از حد پیچیده هستند. در این مقاله، از یک رهیافت مبتنی بر خوشه‌بندی فضایی برای شناسایی زیرنواحی همگن استفاده می‌شود. سپس، در هر زیرناحیه، مدل‌های اقتصادسنجی فضایی بیزی به داده‌ها برازش داده می‌شوند. با توجه به پیچیدگی توزیع پسین مدل‌های پیشنهادی و دوری از مشکلات الگوریتم‌های MCMC، از روش تقریب لاپلاس آشیانه‌ای جمع‌بسته استفاده می‌شود. آنگاه در یک مطالعه شبیه‌سازی، عملکرد رهیافت پیشنهادی ارزیابی و نحوه کاربست رهیافت دومرحله‌ای پیشنهادی برای تحلیل داده‌های قیمت مسکن در شهر مشهد ارائه خواهد شد.

واژه‌های کلیدی: استنباط بیزی تقریبی، خوشه‌بندی فضایی، مدل‌های اقتصادسنجی، ناهمگنی فضایی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 91B84، 62J12.

۱ مقدمه

داده‌های فضایی حاصل از اندازه‌گیری متغیرهای هدف، در یک ناحیه جغرافیایی، هر چه به هم نزدیک‌تر باشند، تمایل بیشتری به شبیه بودن دارند (قانون اول توبلر). این ویژگی، همبستگی فضایی نامیده شده است. برای وارد کردن همبستگی فضایی، مدل‌های آماری مختلفی مانند فرآیندهای تصادفی گاوسی برای داده‌های زمین‌آماری و اتورگرسیو شرطی برای داده‌های مشبکه‌ای (کرسی، ۱۹۹۳؛ استاین، ۲۰۲۲؛ محمدزاده، ۱۳۹۸) و مدل‌های اقتصادسنجی فضایی (لی‌سگ و پیس، ۲۰۰۹؛ رسولی، ۱۳۹۰) برای داده‌های حاصل از شاخص‌های مالی و اقتصادی پیشنهاد شده‌اند. وجود ساختار وابستگی برای متغیرهای اقتصادی مانند قیمت مسکن، ثابت‌شده است. ناهمگنی[1] فضایی ویژگی دیگری است که معمولا در داده‌های فضایی مالی و اقتصادی نمایان است (قانون دوم توبلر). عبارت ناهمگنی، در آمار، نشان می‌دهد یک یا چند مشخصه مورد علاقه در همه زیرمجموعه‌های جامعه آماری یکسان نیست. در مواردی که ناحیه تحت مطالعه بزرگ یا از نظر اقتصادی–اجتماعی متنوع باشد یا وضوح فضایی ناحیه بالا باشد، وجود ناهمگنی فضایی محتمل است. نادیده گرفتن ناهمگنی فضایی، منبعی برای ایجاد اریبی در نمونه‌گیری فضایی، پیدایش اختلاط[2] در مدل‌بندی فضایی و کاهش دقت برآوردهای ضرایب رگرسیونی و پیش‌گویی فضایی است.

تا کنون برای تشخیص وجود هم‌زمان ناهمگنی و همبستگی فضایی، آزمون‌های متعددی طراحی شده‌اند. به‌عنوان نمونه می‌توان به آزمون‌های بریچ پیگان[3] (بریچ و پیگان، ۱۹۷۹)، لاگرانژ چندگانه مشترک (انسلین، ۱۹۹۸)، آزمون چو[4] (انسلین، ۱۹۹۰)، آزمون ناهمسانی واریانس کلجیان –رابینسون (کلجیان و رابینسون، ۱۹۹۸) و ضریب لاگرانژ (مور و همکاران، ۲۰۰۸) اشاره کرد. چالش جدی بعد از معلوم شدن وجود ناهمگنی فضایی، تعیین ساختار ناهمگنی است. برای مدل‌بندی ناهمگنی فضایی، رده مدل‌های رگرسیون موزون جغرافیایی[5] (GWR) اولین انتخاب هستند. این مدل‌ها، با بهره‌گیری از ماتریس‌های وزن فضایی برای متغیرهای پاسخ و تبیینی، ناهمگنی داده‌ها را به‌طور موضعی لحاظ می‌کنند. با این حال، در مواردی که اطلاعات کافی درباره رفتار ناهمگنی داده‌ها در دسترس نیست، تعیین پهنای باند توابع هسته وزن‌دهی و ماتریس‌های وزن متناظر با چالش‌های جدی مواجه می‌شوند که منتهی به ناتوانی در مدل‌بندی کارای ناهمگنی فضایی خواهد شد. افزون بر این، مدل‌های GWR به‌طور موضعی برای هر مکان اثرات متفاوتی را برای متغیرهای تبیینی در نظر می‌گیرد. این در حالی است که در کاربردهای متنوع اقتصادی، برخی از مکان‌ها در ناحیه تحت مطالعه همگن هستند و اندازه اثر یکسانی برای متغیرهای تبیینی در آن‌ها قابل تدوین است. به همین دلیل، در این نوع کاربردها، به‌کارگیری رهیافت‌های مبتنی بر خوشه‌بندی فضایی می‌تواند مفیدتر باشد و ما را از پیچیدگی بیش از اندازه مدل‌های GWR و سختی تفسیر نتایج حاصل از آن‌ها دور کند.

از جمله پژوهش‌هایی که تمرکز آن‌ها بر روی روش‌های خوشه‌بندی فضایی است، می‌توان به موارد زیر اشاره

---

[1] Heterogenity
[2] Confounding
[3] Breuch-Pagan
[4] Chow test
[5] Geographical Weighted Regression

کرد. بیله و همکاران (۲۰۱۷) از یک رویکرد دومرحله‌ای برای مدل‌بندی وابستگی و ناهمگنی فضایی استفاده کردند. در این رویکرد، ابتدا رژیم‌های فضایی[1] شناسایی می‌شوند و در مرحله دوم، در یک چارچوب درستنمایی، پارامترهای مدل برای هر رژیم برآورد می‌شوند. جنیوکس و همکاران (۲۰۱۸) از یک روش دومرحله‌ای برای تحلیل داده‌های فضایی ناهمگن استفاده کردند. لی و سانگ (۲۰۱۹) یک مدل با ضرایب متغیر خوشه‌ای فضایی را برای شناسایی الگوهای خوشه‌بندی فضایی و برآورد ضرایب با در نظر گرفتن یک تابع جریمه ارایه کردند. سوگاساوا (۲۰۲۱) یک روش ترکیبی رگرسیون و خوشه‌بندی را با تقسیم مکان‌های جغرافیایی به تعداد محدودی از مناطق پیشنهاد کردند. در چارچوب بیزی، لاوسون و همکاران (۲۰۱۴) خوشه‌بندی ضرایب رگرسیونی را با مجموعه‌ای از توزیع‌های پیشین گسسته پیشنهاد کردند. ژیهوا و همکاران (۲۰۲۰) یک مدل خوشه‌بندی فضایی بیزی را معرفی کردند که با استفاده از پیشین فرآیند دیریکله (فرگوسن، ۱۹۷۳) ساختار وابسته به مکان ضرایب رگرسیونی را در نظر می‌گیرد و اثرات متغیرهای تبیینی را به‌طور همزمان برآورد و خوشه‌بندی می‌کند. همچنین ژائو و همکاران (۲۰۲۳) با استفاده از یک مدل رگرسیون خوشه‌بندی فضایی و بر اساس رهیافت بیزی آمیخته‌های متناهی به تحلیل ناهمگنی فضایی ضرایب رگرسیونی با پاسخ‌های شمارشی پرداختند. این مدل امکان تشخیص تعداد خوشه‌های ضرایب رگرسیون و سپس مدل‌بندی با استفاده از اطلاعات همسایگی جغرافیایی را فراهم می‌کند.

در کارهای بیزی اشاره‌شده، پژوهشگران برای برازش مدل و اجرای استنباط بیزی، از الگوریتم‌های MCMC برای تقریب توزیع پسین پارامترهای مدل استفاده کرده‌اند. اگرچه روش‌های نمونه‌گیری MCMC به ابزاری استاندارد برای استنباط در این نوع مدل‌ها تبدیل شده‌اند، اما معمولا دارای محاسباتی پیچیده و زمان‌بر هستند و از مشکلات آمیختگی ضعیف و همگرایی کند رنج می‌برند. برای دوری از این مشکلات، رو و همکاران (۲۰۰۹) روش تقریب لاپلاس آشیانه‌ای جمع‌بسته[2] (INLA) را برای رده مدل‌های گاوسی پنهان[3] (LGM) معرفی کردند. این روش با ترکیب کارای تقریب لاپلاس و انتگرال‌گیری عددی، با دقتی مشابه روش‌های MCMC، تقریبی برای توزیع‌های پسین کناری کمیت‌های هدف فراهم می‌کند. سرعت محاسبه تقریب توزیع‌های پسین کناری با روش INLA در مقایسه با الگوریتم‌های MCMC بسیار قابل توجه و جذاب است، به‌طوری که محاسبات چندین ساعته روش‌های MCMC را با چند ثانیه جایگزین می‌کند. از ویژگی‌های دیگر روش INLA فراهم کردن ابزارهای ارزیابی مدل است که در حوزه انتخاب مدل حیاتی است. رو و همکارانش با ایجاد بسته نرم‌افزاری R-INLA در محیط برنامه‌نویسی R، امکان به‌کارگیری روش تقریبی پیشنهادی خود را، به سادگی، برای حوزه وسیعی از پژوهشگران و به منظور برازش مدل‌های کاربردی متنوعی که در رده LGM قرار می‌گیرند، فراهم کرده‌اند. اگرچه رده LGM طیف وسیعی از پرکاربردترین مدل‌های آماری را شامل می‌شود، اما برازش مدل‌های اقتصادسنجی فضایی با روش INLA تا همین چند سال پیش ممکن نبود. به‌طور دقیق‌تر، مولفه پنهانی که ساختار مدل‌های اقتصادسنجی فضایی را در قالب یک مدل گاوسی پنهان تعبیه کند، وجود نداشت. اولین بار گومز و همکاران (۲۰۲۱) مولفه پنهانی که یک مدل تأخیر فضایی[4] (SLM) را

[1] Spatial Regimes
[2] Integrated Nested Laplace Approximation
[3] Latent Gaussian Models
[4] Spatial Lag Model

در قواره یک LGM درمی‌آورد، معرفی کردند. با این کار، سایر مدل‌های اقتصادسنجی فضایی پرطرفدار نیز قابل تعریف هستند.

رهیافت مبتنی بر خوشه‌بندی پیشنهادی در این مقاله، ترکیبی از روش معرفی‌شده در بیله و همکاران (۲۰۱۷)، روش هموارسازی موزون سازوار[1] (AWS) و روش INLA است. در مرحله اول، با ترکیب رهیافت بیله و همکارانش و روش AWS در یک مدل GWR، خوشه‌بندی فضایی انجام می‌شود. این نوع خوشه‌بندی یک روش تکراری است که با ترکیب الگوریتم هموارسازی موزون سازوار و GWR، به شناسایی نواحی همگن یا رژیم‌های فضایی می‌پردازد. این ترکیب نه تنها تلاش می‌کند نواحی همگن را شناسایی کند، بلکه با در نظر گرفتن رژیم‌های فضایی در یک مدل فراموضعی، می‌تواند ضرایب رگرسیون را به‌طور مجزا برای هر رژیم برآورد کرده و نارسایی روش GWR را در نمایش بیش از اندازه پیچیده ساختار ناهمگنی مدل رفع کند. استفاده از روش AWS نیز به دلیل توانایی آن در کاهش نوفه و هموارسازی بهتر، به افزایش دقت در فرآیند خوشه‌بندی کمک می‌کند. در مرحله دوم، مدل‌های اقتصادسنجی فضایی در هر خوشه با تنظیم یک چارچوب استنباط بیزی به کمک روش INLA برازش داده می‌شوند.

نوآوری این مقاله دو نما دارد: نمای اول، تعمیم رهیافت پیشنهادی بیله و همکاران (۲۰۱۷) در بخش مدل‌بندی رژیم‌های فضایی با به‌کارگیری مدل‌های اقتصادسنجی بیزی به‌جای کلاسیک است. این جنبه امکان بهره‌گیری از اطلاعات کمکی و مرتبط با مساله مورد هدف را از طریق تعیین توزیع‌های پیشین فراهم می‌آورد. از طرفی، با توجه به انتخاب روش تقریبی INLA برای برازش مدل‌های پیشنهادی، کارایی محاسباتی چشم‌گیر در مقایسه با استفاده از روش‌های نمونه‌گیری MCMC پدیدار می‌شود. این کارایی محاسباتی به‌ویژه در شرایطی که داده‌ها حجم بالایی دارند، یک نقطه قوت اساسی محسوب می‌شود. نمای دوم، تعمیم کار گومز و همکاران (۲۰۲۱) است که همزمان با برازش مدل‌های اقتصادسنجی بیزی با روش INLA، ذات ناهمگنی فضایی داده‌ها بر پایه خوشه‌بندی آن‌ها نیز لحاظ می‌شود. این دو تعمیم، امکان شناسایی نواحی همگن، برآورد کارای ضرایب رگرسیونی و تحلیل سریع داده‌های فضایی ناهمگن خوشه‌ای را ایجاد می‌کند. استفاده از رهیافت INLA برای مدل‌بندی داده‌های فضایی خوشه‌ای در یک چارچوب بیزی، توسعه جدیدی است که توسط نویسندگان این مقاله پیشنهاد می‌شود. در ادامه، در بخش دوم، روش پیشنهادی مقاله شامل معرفی الگوریتم خوشه‌بندی و برازش مدل‌های اقتصادسنجی فضایی با روش INLA تشریح می‌شود. در بخش سوم، با طرح یک مطالعه شبیه‌سازی، عملکرد روش پیشنهادی مقاله ارزیابی می‌شود. در بخش چهارم، مدل‌بندی و تحلیل داده‌های قیمت مسکن در شهر مشهد ارایه می‌شود. در پایان بحث و نتیجه‌گیری مطرح خواهد شد.

۲ رهیافت پیشنهادی

رگرسیون موزون جغرافیایی (فودرینگهام و همکاران، ۲۰۰۲) این امکان را فراهم می‌آورد که پارامترها (به جای فراموضعی) ناحیه‌ای برآورد شوند. در مدل GWR برای هر نقطه در فضا یک معادله رگرسیونی جدا تعریف می‌شود

[1] Adaptive Weighted Smoothing

و برای برآورد پارامترهای مدل در هر نقطه، از مشاهدات همسایه آن نقطه استفاده می‌شود به‌طوری که مشاهدات نزدیک وزن بیشتری نسبت به نقاط دورتر دریافت می‌کنند. در حالتی که پاسخ نرمال باشد، می‌توان مدل خطی

$$\boldsymbol{y} = (\boldsymbol{\beta} \odot \boldsymbol{X})\mathbf{1} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma_{\varepsilon}^{2}\mathbf{I})$$

را برای برآورد موضعی اثرات متغیرهای تبیینی نوشت، که در آن $\boldsymbol{y}$ بردار ستونی با بعد n و $\boldsymbol{\beta}$ ماتریس $n \times (p+1)$ با سطر iام $\boldsymbol{\beta}_i = (\beta_{i0}, \ldots, \beta_{ip})^{\mathsf{T}}$ است. عملگر $\odot$ ضرب هادامارد و ماتریس طرح $\boldsymbol{X}$ شامل مشاهدات متغیرهای تبیینی و هم‌بعد با $\boldsymbol{\beta}$ است. بردار $\boldsymbol{\varepsilon}$ نیز n بعدی و $\mathbf{1}$ یک بردار $(p+1)$ بعدی از ۱ها است. برای برآورد پارامترهای رگرسیونی در موقعیت i می‌توان با استفاده از روش کمترین توان‌های دوم، تابع هدف

$$Q(\boldsymbol{\beta}_i) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_i)^{\mathsf{T}}\boldsymbol{W}_i(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_i) = \boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{W}_i\boldsymbol{\varepsilon}, \quad i = 1, \ldots, n$$

را در زیرمجموعه‌ای از نقاط که نزدیک به موقعیت i هستند کمینه کرد، که در آن $\boldsymbol{W}_i$ یک ماتریس قطری با ابعاد n است که درایه‌های روی قطر اصلی آن وزن هر مشاهده برای موقعیت i است. در مدل فضایی GWR، این وزن‌ها با توجه به فواصل جغرافیایی بین مشاهدات تعیین می‌شوند. به این ترتیب، برآوردگر کمترین توان‌های دوم موزون موضعی به صورت

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}_i\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}_i\boldsymbol{y}, \quad i = 1, \ldots, n$$

نتیجه می‌شود. برای ایجاد ماتریس‌های وزن $\boldsymbol{W}_i$ از توابع هسته استفاده می‌شود. توابع هسته معمول برای این کار نمایی، گاوسی، توان دوم و توان سوم هستند. اگرچه همه این توابع، کم و بیش، مشابه هم به مشاهدات نزدیک وزن بیشتر و به مشاهدات دور وزن کمتر نسبت می‌دهند، اما با توجه به نوع پهنای باند از یکدیگر متمایز می‌شوند. در یک مدل GWR، نرخی که در آن وزن‌های مرتبط با هر موقعیت با افزایش فاصله کاهش می‌یابد توسط پارامتر پهنای باند کنترل می‌شود که مقدار آن باید بهینه شود.

انتخاب پهنای باند و در نتیجه انتخاب شعاع همسایگی مناسب، در برآورد ضرایب رگرسیون موضعی بسیار موثر است. برای انتخاب پهنای باند بهینه (b^{opt}) معمولا از روش‌های اعتبارسنجی متقابل[1] استفاده می‌شود (سیلورمن، ۱۹۸۴). این روش‌ها به دنبال پهنای باندی برای تابع هسته هستند که باقی‌مانده‌های حاصل از مدل برای پیشگویی $\boldsymbol{y}_i$ ها را توسط یک زیرمجموعه از داده‌ها می‌نیمم کند. این رهیافت توسط سلولند (۱۹۷۹) معرفی و برای برآورد تابع هسته توسط بومن (۱۹۸۴) پیشنهاد شد. یک رهیافت دیگر برای انتخاب پهنای باند بهینه، که بر پایه مبادله بین پیچیدگی (درجه آزادی) و نیکویی برازش مدل پایه‌ریزی شده است، استفاده از معیارهای انتخاب مدل مانند معیار اطلاع آکائیک[2] (AIC) است (آکائیک، ۱۹۷۶). در این روش مقدار بهینه پهنای باند از طریق می‌نیمم کردن AIC

[1] Cross Validation
[2] Akaike Information Criterion

به‌دست می‌آید. این معیار در متون مختلف با صورت‌های مختلف ارائه شده است. فودرینگهام و همکاران (۲۰۰۲) یک نسخه تصحیح‌شده از AIC را برای GWR، بر اساس کار هورویج و همکاران (۱۹۹۸)، به صورت

$$\mathrm{AIC}_c = ۲n\ln(\hat{\sigma}_\varepsilon) + n\ln(۲\pi) + n\frac{n+\mathrm{tr}(\boldsymbol{S})}{n-\mathrm{tr}(\boldsymbol{S})-۲} \qquad (۱)$$

معرفی کردند که پهنای باند بهینه از می‌نیمم کردن آن نتیجه می‌شود، که در آن $\mathrm{tr}(\boldsymbol{S})$ مجموع درایه‌های روی قطر اصلی ماتریسی با سطرهای $\boldsymbol{s}_i = \boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{W}_i\boldsymbol{X})^{-۱}\boldsymbol{X}^\mathsf{T}\boldsymbol{W}_i$ است و $\hat{\sigma}_\varepsilon$ برآورد انحراف معیار جمله خطا است. برآورد $\hat{\sigma}_\varepsilon$ در یک چارچوب درستنمایی ماکسیمم توسط فودرینگهام و همکاران (۲۰۰۲) ارائه شده است.

۲.۱ خوشه‌بندی فضایی

روش پیشنهادی مقاله برای خوشه‌بندی فضایی مبتنی بر کار بیله و همکاران (۲۰۱۷) و با سازوکار الگوریتم AWS است که ابتدا توسط پولزهل و اسپوکونی (۲۰۰۰) در حوزه پردازش تصویر برای تمیز کردن داده‌های نوفه‌دار معرفی شد. الگوریتم AWS از یک روش مبتنی بر همسایگی و یک مجموعه از وزن‌ها برای همسایه‌های هر نقطه در داده‌ها استفاده می‌کند. وزن‌ها بر اساس فاصله هر نقطه با نقاط همسایه خود تعیین می‌شوند. در این روش، ابتدا وزن‌های اولیه برای هر نقطه i به صورت $w_{ij}^{\circ} = K(d_{ij};b)$ در نظر گرفته می‌شوند، که در آن $K(d_{ij};b)$ یک تابع هسته دلخواه با پهنای باند b و d_{ij} فاصله بین موقعیت i و سایر نقاط مشاهده‌شده است. سپس با انتخاب مقدار پهنای باند بهینه، مثلا به کمک می‌نیمم کردن AIC_c در رابطه (۴)، مدل GWR برازش داده می‌شود و برآوردهای ضرایب رگرسیونی برای همه نقاط مشاهده‌شده به صورت ماتریس

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_{۱\circ} & \hat{\beta}_{۱۱} & \cdots & \hat{\beta}_{۱p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_{n\circ} & \hat{\beta}_{n۱} & \cdots & \hat{\beta}_{np} \end{bmatrix}$$

نتیجه می‌شود. بنابراین هر مشاهده i، یک بردار $(p+۱)$ بعدی از ضرایب رگرسیونی دارد. برای تعیین نواحی همگن یا به عبارتی مشاهداتی که رفتار و ضرایب مدل در آن‌ها مشابه است، باید بردارهای ضرایب رگرسیونی آن‌ها دو به دو با هم مقایسه شوند. با به‌روزرسانی وزن‌ها در صورت وجود شباهت، مشاهدات مربوطه در یک خوشه قرار می‌گیرند. پس از به‌روزرسانی وزن‌ها، مدل GWR با استفاده از ماتریس وزن‌های جدید بر روی تمام مشاهدات دوباره برازش داده می‌شود. تکرار این فرآیند تا زمانی که شرط $\max\left|w_{ij}^{(\ell-۱)} - w_{ij}^{\ell}\right| < \omega$ برای مقادیر کوچک ω برقرار شود، ادامه می‌یابد و وزن‌ها در هر تکرار به‌روز می شوند. برای مقایسه مشابهت ضرایب رگرسیونی برآوردشده دو موقعیت

i و j، بیله و همکاران (۲۰۱۷) از آماره والد به صورت

$$\chi_{ij}^{\ell} = (\hat{\boldsymbol{\beta}}_i^{\ell} - \hat{\boldsymbol{\beta}}_j^{\ell})^{\mathsf{T}} (\Sigma^{\ell})^{-1} (\hat{\boldsymbol{\beta}}_i^{\ell} - \hat{\boldsymbol{\beta}}_j^{\ell}) \qquad (۲)$$

استفاده کردند، که در آن بردارهای $\hat{\boldsymbol{\beta}}_i^{\ell}$ و $\hat{\boldsymbol{\beta}}_j^{\ell}$ برآوردهای ضرایب مدل GWR برازش‌شده در تکرار ℓام هستند و ماتریس واریانس ادغامی Σ^{ℓ} از ترکیب دو ماتریس واریانس مربوط به بردارهای $\hat{\boldsymbol{\beta}}_i^{\ell}$ و $\hat{\boldsymbol{\beta}}_j^{\ell}$ در تکرار ℓام، به شکل $\Sigma^{\ell} = \frac{\mathrm{tr}(\boldsymbol{W}_i^{\ell})\Sigma_i^{\ell} + \mathrm{tr}(\boldsymbol{W}_j^{\ell})\Sigma_j^{\ell}}{\mathrm{tr}(\boldsymbol{W}_i^{\ell}) + \mathrm{tr}(\boldsymbol{W}_j^{\ell})}$ به دست می‌آید، به‌طوری که ماتریس Σ_i^{ℓ} برای هر تکرار به صورت

$$\Sigma_i^{\ell} = \left[(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}_i^{\ell}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}_i^{\ell}\right]\left[(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}_i^{\ell}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}_i^{\ell}\right]^{\mathsf{T}}\hat{\sigma}_{\epsilon_i}^{2}$$

محاسبه می‌شود. آماره χ_{ij}^{ℓ} یک تابع درجه دوم از فاصله دو بردار از ضرایب رگرسیونی برآوردشده است و با توجه به نرمال بودن بردار ضرایب، این آماره دارای توزیع کی‌دو با درجه آزادی $p+1$ است. مقادیر بزرگ χ_{ij}^{ℓ} نشان‌دهنده وجود اختلاف معنی‌دار بین $\hat{\boldsymbol{\beta}}_i^{\ell}$ و $\hat{\boldsymbol{\beta}}_j^{\ell}$ است. این اختلاف بین ضرایب رگرسیونی دو موقعیت فضایی i و j نشان‌دهنده وجود ناهمگنی فضایی در ناحیه مورد بررسی است. بنابراین اگر $\chi_{ij}^{2} > \chi_{(p+1)}^{2}(1-\alpha)$ آن‌گاه موقعیت‌های i و j نمی‌توانند در یک خوشه همگن قرار گیرند.

در فرآیند خوشه‌بندی فضایی اشاره‌شده، درایه‌های ماتریس‌های وزن W_i^{ℓ} در هر تکرار، با توجه به نتایج آزمون والد ساخته‌شده توسط آماره آزمون (۲) و در مقایسه با مقادیر قبلی به‌روزرسانی می‌شوند. سازوکار این روش وزن‌دهی بر اساس الگوریتم AWS طراحی شده است. فرآیند همگرایی وزن‌ها به مقادیر منطقی صفر و یک، به کندی انجام می‌شود. برای اطمینان از همگرایی، تابع وزن w_{ij}^{ℓ} با ضرب تابع هسته $K(d_{ij};b)$در یک هسته که تابعی از آماره آزمون χ_{ij} در تکرار ℓام است، ساخته می‌شود. بنابراین

$$w_{ij}^{\ell} = K(d_{ij}^{\ell};b)K(\chi_{ij}^{\ell};\tau) \quad \ell = 1,2,\ldots \qquad (۳)$$

که در آن، $d_{ij}^{\ell} = \frac{d_{ij}}{\ell}$ و $K(\chi_{ij}^{\ell};\tau) = \exp(-\frac{1}{2}(\chi_{ij}^{\ell}\tau)^{2})$. پولزهل و اسپوکونی (۲۰۰۰) انتخاب مقدار $\frac{1}{Q_\alpha(\chi^2_{(p+1)})}$ را برای پارامتر وزن‌دهی τ پیشنهاد کردند، که در آن $Q_\alpha(\chi^2_{(p+1)})$ چندک مرتبه α توزیع کی‌دو با $(p+1)$ درجه آزادی است. آن‌ها به‌روزرسانی دوباره وزن‌های حاصل از (۳) را پیشنهاد کردند، که ترکیبی خطی از وزن‌های به‌دست آمده در تکرار ℓام و مقادیر متوسط محاسبه‌شده آن‌ها در تکرار قبلی، $\bar{w}_{ij}^{(\ell-1)}$، است که به شکل $w_{ij}^{\ell} = (1-\eta)\bar{w}_{ij}^{(\ell-1)} + \eta w_{ij}^{\ell}$ تعریف می‌شود، که در آن $\eta \in (0,1)$ پارامتر کنترلی است که پولزهل و اسپوکونی (۲۰۰۰) آن را پارامتر حافظه نامیدند. در طی فرآیند همگرایی و مشخص شدن خوشه‌ها، کافی بودن مشاهدات هر خوشه بسیار مهم است. اگر مشاهدات یک خوشه تعیین‌شده از تعداد پارامترهای مدل کمتر باشد، مشاهدات آن خوشه به‌عنوان داده‌های پرت از محاسبات خارج می‌شوند. معیاری برای سنجش همگرایی، تابع تغییرات وزن $d(w) = \max_{i,j} |\bar{w}_{ij}^{(\ell-1)} - w_{ij}^{\ell}|$ است که هم‌زمان با تثبیت شدن فرآیند همگرایی، به صفر میل می‌کند.

۲.۲ مدل‌های اقتصادسنجی فضایی

در این بخش، پرکاربردترین مدل‌های اقتصادسنجی فضایی معرفی می‌شوند. فرض کنید $\boldsymbol{y}$ بردار nبعدی مشاهدات پاسخ در موقعیت‌های جغرافیایی مختلف ناحیه تحت مطالعه و $\boldsymbol{W}$ ماتریس همسایگی با بعد n برای نشان دادن مجاورت بین دو موقعیت با سطرهای استانداردشده باشد. مدل خطاهای فضایی[۱] (SEM) بر پایه یک عبارت خطای اتورگرسیو فضایی، به صورت

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} = \rho_{\text{Err}}\boldsymbol{W}\boldsymbol{u} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{\text{۲}}\mathbf{I}_n)$$

ساخته می‌شود، که در آن $\boldsymbol{X}$ و $\boldsymbol{\beta}$ مشابه بالا تعریف می‌شوند، $\mathbf{I}_n$ ماتریس همانی با بعد n است و $\boldsymbol{u}$ جمله خطا با ساختار اتورگرسیو و دارای توزیع نرمال چندمتغیره است. این مدل را می‌توان به صورت

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + (\mathbf{I}_n - \rho_{\text{Err}}\boldsymbol{W})^{-\text{۱}}\boldsymbol{e}, \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{\text{۲}}\mathbf{I}_n)$$

بازنویسی کرد، که در آن ρ_{Err} پارامتر خودهمبستگی فضایی مبتنی بر جمله خطا است. مدل تأخیر فضایی یک ساختار اتورگرسیو را برای متغیر پاسخ در نظر می‌گیرد و به صورت

$$\boldsymbol{y} = \rho_{\text{lag}}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{\text{۲}}\mathbf{I}_n)$$

تعریف می‌شود، که در آن ρ_{lag} پارامتر خودهمبستگی فضایی مبتنی بر متغیر پاسخ است. مدل تأخیر فضایی را نیز می‌توان به صورت

$$\boldsymbol{y} = (\mathbf{I}_n - \rho_{\text{lag}}\boldsymbol{W})^{-\text{۱}}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}), \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{\text{۲}}\mathbf{I}_n)$$

بازنویسی کرد. مدل سوم که به‌طور گسترده در اقتصادسنجی فضایی استفاده می‌شود، مدل دوربین فضایی[۲] (SDM) است که توسط انسلین (۱۹۹۸) به صورت

$$\boldsymbol{y} = \rho_{\text{lag}}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{\text{۲}}\mathbf{I}_n)$$

معرفی شده است، که در آن $\boldsymbol{\gamma}$ بردار ضرایب رگرسیونی متغیرهای تبیینی با تأخیر فضایی است. این مدل معادل

$$\boldsymbol{y} = (\mathbf{I}_n - \rho_{lag}\boldsymbol{W})^{-\text{۱}}(\boldsymbol{X}^*\boldsymbol{\beta}' + \boldsymbol{e}), \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{\text{۲}}\mathbf{I}_n) \qquad (\text{۴})$$

[1] Spatial Error Model
[2] Spatial Durbin Model

است، که در آن $\boldsymbol{X}^* = [\boldsymbol{X}, \boldsymbol{W}\boldsymbol{X}]$ ماتریس متغیرهای تبینی با و بدون تأخیر فضایی و بردار ضرایب رگرسیونی متناظر $\boldsymbol{\beta}' = [\boldsymbol{\beta}, \gamma]$ است. مدل دیگر نزدیک به SDM، مدل خطای دوربین فضایی[1] (SDEM) است، که در آن مولفه خطا دارای ساختار اتورگرسیو است. این مدل به‌صورت

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} = \rho_{\text{Err}}\boldsymbol{M}\boldsymbol{u} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{۲}\mathbf{I}_n)$$

تعریف می‌شود، که در آن $\boldsymbol{M}$ یک ماتریس مجاورت فضایی برای خطا است و می‌تواند با $\boldsymbol{W}$ متفاوت باشد. این مدل به صورت

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\beta}' + (\mathbf{I}_n - \rho_{\text{Err}}\boldsymbol{M})^{-۱}\boldsymbol{e}, \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{۲}\mathbf{I}_n)$$

قابل بازنویسی است. یک مدل ساده‌شده بدون پارامترهای خودهمبستگی مکانی و متغیرهای تأخیری، معروف به مدل X با تأخیر فضایی[2] (SLX)، نیز معرفی شده است که به‌صورت

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\beta}' + \boldsymbol{e}, \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{۲}\mathbf{I}_n)$$

تعریف می‌شود (گومز و همکاران، ۲۰۲۱). این پنج مدل از مدل‌های استاندارد در اقتصادسنجی فضایی هستند که بر سه محور کلیدی خطاهای همبسته فضایی، پاسخ‌های همبسته اتورگرسیو فضایی و متغیرهای کمکی با تأخیر فضایی تمرکز دارند.

در طراحی مدل تأخیر فضایی، ممکن است خودهمبستگی فضایی هم در متغیر پاسخ و هم در جمله خطا وجود داشته باشد. به‌عبارت دیگر، شرایط مدل تأخیر فضایی و مدل خطاهای فضایی به‌طور همزمان برقرار می‌شوند. در این حالت، می‌توان ترکیب این دو مدل را با ماتریس‌های فضایی $\boldsymbol{M}$ و $\boldsymbol{W}$ و پارامترهای وابستگی فضایی ρ_{lag} و ρ_{Err} به صورت

$$\boldsymbol{y} = \rho_{\text{lag}}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} = \rho_{\text{Err}}\boldsymbol{M}\boldsymbol{u} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \text{MVN}(\circ, \sigma^{۲}\mathbf{I}_n)$$

نوشت. این مدل در متون اقتصادسنجی با نام مدل ترکیبی اتورگرسیو فضایی[3] (SAC) یا مدل اتورگرسیو فضایی با خطاهای اتورگرسیو[4] (SARAR) شناخته می‌شود (کلجیان و پروچا، ۲۰۱۰). این مدل زمانی مناسب است که هر دو آزمون لاگرانژ برای ضریب خطا و تأخیر فضایی معنی‌دار باشند. در حالتی که $\boldsymbol{M} = \boldsymbol{W}$ و $\rho_{\text{Err}} = \circ$، مدل SAC به مدل تأخیر فضایی ساده تبدیل می‌شود.

---

[1] Spatial Durbin Error Model
[2] Spatial Lag X Model
[3] Spatial Autoregressive Combined
[4] Spatial Autoregressive Model with Autoregressive Distubances

۲.۳ تقریب لاپلاس آشیانه‌ای جمع‌بسته

همان‌طور که در مقدمه بیان شد، برای دوری از مشکلات برازش مدل‌های اقتصادسنجی فضایی بیزی با الگوریتم‌های MCMC، رو و همکاران (۲۰۰۹) روش INLA را معرفی کردند. جزییات این روش در منبع اصلی رو و همکاران (۲۰۰۹) قابل مشاهده است. به‌طور مشخص، آن‌ها روش پیشنهادی خود را برای رده LGM که یک زیررده از مدل‌های رگرسیون با ساختار جمعی هستند، با پذیرفتن مارکوفی بودن میدان تصادفی گاوسی توسعه دادند. رو و همکارانش فرض کردند توزیع مشاهدات y_i، برای $i = ۱, \ldots, n$، از یک خانواده (اغلب) توزیع‌های نمایی پیروی می‌کند که در آن میانگین توزیع، μ_i، و پیشگوی جمعی مدل، ζ_i، به کمک تابع پیوند $g(\cdot)$ به صورت $g(\mu_i) = \zeta_i$ مرتبط می‌شوند، به‌طوری که

$$\zeta_i = \beta_۰ + \sum_{k=۱}^{p} \beta_k x_{ik} + \sum_{j=۱}^{J} f_j(v_{ij}) + \epsilon_i \qquad (۵)$$

که در آن $\beta_۰$ عرض از مبدا، f_jها اثرات ناپارامتری ساختارمند با برخی وابستگی‌های ساختاری مشخص و ضرایب β_kها اثرات ثابت متغیرهای تبیینی مدل هستند. مولفه ϵ نیز خطای غیرساختاری مدل است. این رده از مدل‌ها به دلیل انعطاف در انتخاب مولفه‌های f_j شامل بسیاری از مدل‌های پرکاربرد آماری است. به‌عنوان نمونه، مولفه‌های مذکور می‌توانند شامل اثرات تصادفی، فضایی، زمانی، فضایی–زمانی، و اثرات ناخطی متغیرهای تبیینی باشند. فرض می‌شود بردار مولفه‌های پنهان $\mathbf{x} = \{\beta_۰, \{\beta_k\}, \{f_j\}, \zeta_i\}$ در مدل (۵)، گاوسی با ابرپارامترهای $\boldsymbol{\theta}$ است.

در چارچوب مدل‌بندی بیزی با روش INLA ساخت یک f_j که بتواند ساختار یک SLM را تعبیه کند، اولین بار توسط گومز و همکاران (۲۰۲۱) انجام شد. این مولفه پنهان، برای بردار مشاهدات با بعد n، در مدل (۵) به صورت یک اثر تصادفی با صورت

$$\mathbf{x}_j = (\mathbf{I}_n - \rho \boldsymbol{W})^{-۱}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}) \qquad (۶)$$

وارد می‌شود، که در آن به‌جای در نظر گرفتن ماتریس کوواریانس $\sigma^۲\mathbf{I}_n$ برای بردار $\boldsymbol{e}$ از ماتریس دقت $\nu\mathbf{I}_n$ استفاده می‌شود که در ادبیات روش INLA کار کردن با ماتریس دقت به‌جای ماتریس کوواریانس با توجه به فرض کردن ساختار مارکوفی برای بردار $\mathbf{x}$، کلیدی است. توزیع‌های پیشین پیش‌فرض پارامترهای مولفه‌های پنهان (۵) در بسته نرم‌افزاری `R-INLA` برای بردار ضرایب $\boldsymbol{\beta}$، گاوسی با بردار میانگین صفر و ماتریس کوواریانس قطری با درایه‌های بزرگ مانند ۱۰۰۰ است، برای $\text{logit}(\rho)$، گاوسی با میانگین صفر و واریانس ۱۰ است و برای $\log(\nu)$، لگ‌گاما با پارامترهای ۱ و ۵×۱۰^{-۵} است. امکان استفاده از توزیع‌های پیشین دیگر، که در مورد مجموعه داده هدف مطلع باشند، نیز وجود دارد.

با پیاده‌سازی مولفه پنهان مختص SLM، می‌توان سایر مدل‌های اقتصادسنجی فضایی را نیز در قالب یک

LGM تعبیه کرد. به‌طور مشخص، SEM یک حالت خاص از مدل (۴) با $\boldsymbol{\beta} = 0$ است. همچنین SDM با جایگزین کردن ماتریس $\boldsymbol{X}^*$ که در (۴) معرفی شد، به‌جای $\boldsymbol{X}$ در (۶) نتیجه می‌شود. مدل SDEM از دو بخش تشکیل شده است: بخش اول آن یک عبارت خطی معمولی شامل $\boldsymbol{X}^*$ و بخش دیگر آن یک اثر SLM با پارامتر $\boldsymbol{\beta} = 0$ است. مدل SLX نیز به سادگی قابل پیاده‌سازی است. برای اطلاعات بیشتر به گومز و همکاران (۲۰۲۱) مراجعه کنید. برای برازش مدل‌های اقتصادسنجی فضایی اشاره‌شده باید ماتریس‌های همسایگی $\boldsymbol{W}$ و $\boldsymbol{M}$ تعیین شوند. لسیج و همکاران (۲۰۱۱) برای به دست آوردن یک ماتریس همسایگی برای داده‌های چگال از روش نزدیک‌ترین همسایه‌ها استفاده کردند. آن‌ها با برازش مدل‌های با شعاع‌های همسایگی مختلف بر مبنای کوچک‌ترین معیار اطلاع انحراف[1] (DIC)، تعداد بهینه نزدیک‌ترین همسایه‌ها و در نتیجه ماتریس‌های مذکور را تعیین کردند. اجرای روش پیشنهادی آن‌ها به دلیل بار سنگین محاسباتی الگوریتم‌های MCMC، زمان‌بر و در مواردی ناممکن است. در مقابل روش INLA به دلیل سرعت بالای برازش مدل‌ها، می‌تواند فرآیند تعیین تعداد همسایه‌های بهینه و محاسبه ماتریس‌های وزن را با محاسبه معیارهای DIC و درستنمایی کناری خیلی ساده و سریع انجام دهد.

۲.۳.۱ انتخاب مدل

در روش‌شناسی INLA برای انتخاب بهترین مدل، معیارهای متنوعی معرفی شده‌اند و توسط بسته نرم‌افزاری R-INLA به راحتی قابل محاسبه هستند. به‌عنوان نمونه می‌توان معیارهای DIC، لگاریتم امتیاز[2] و مختصات پیشگوی شرطی[3] (CPO) را نام برد. اما استفاده از این معیارها برای انتخاب مدل برتر در مجموعه مدل‌های معرفی‌شده در این مقاله بی‌نتیجه است. گومز و همکاران (۲۰۲۱) تشریح کردند که چون مقدار واریانس درستنمایی گاوسی در مدل‌های اقتصادسنجی فضایی بیان‌شده کوچک در نظر گرفته می‌شود، مقدار معیارهای معرفی‌شده برای تمام مدل‌ها یکسان است. به همین دلیل، برای انتخاب مدل برتر فقط امکان استفاده از درستنمایی کناری مدل‌های برازش‌شده وجود دارد. با توجه به این‌که درستنمایی کناری پیچیدگی مدل‌ها را در نظر نمی‌گیرد و فقط کیفیت برازش را می‌سنجد، بدیهی است که مقادیر بزرگتر آن (که بیانگر برازش بهتر مدل متناظر هست) به سمت مدل‌های پیچیده‌تر اریب می‌شوند. به همین دلیل، در این مقاله ابتدا مدل‌ها بر اساس بالاترین معیار درستنمایی کناری مرتب می‌شوند و در مرحله دوم به‌صورت موردی از بین چند مدل برتر که اختلاف درستنمایی کناری آن‌ها کوچک است، بدون توجه به اولویت، مدلی برگزیده می‌شود که کمترین تعداد پارامتر را داراست.

[1] Deviance Information Criterion
[2] Log Score
[3] Conditional Predictive Ordinates

۳ مطالعه شبیه‌سازی

برای ارزیابی عملکرد روش پیشنهادی، مشاهدات از یک مدل SLM با ساختار

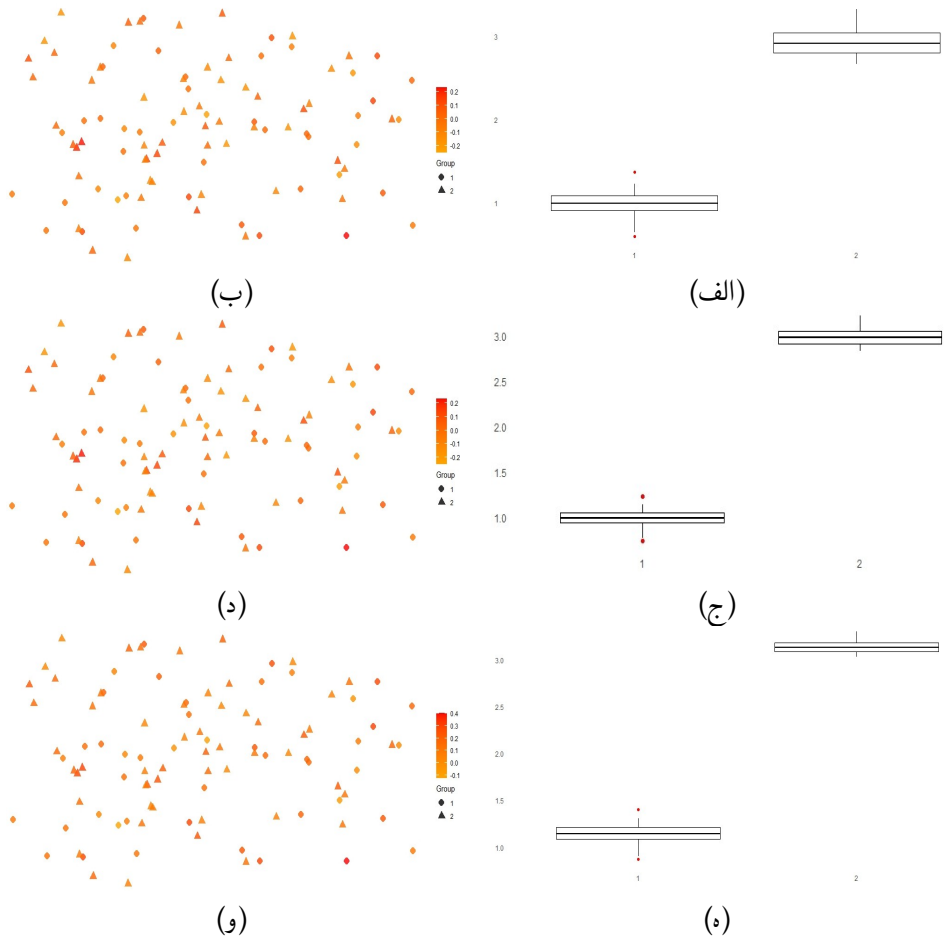
$$\boldsymbol{y} = (\mathbf{I_n} - \rho_{\text{lag}}\boldsymbol{W})^{-1}(\mathbf{x}\beta + \boldsymbol{e}), \quad \boldsymbol{e} \sim \text{MVN}(0, \sigma^2\mathbf{I_n}) \tag{۷}$$

و با $n = ۱۰۰$ شبیه‌سازی شدند. مشاهدات متغیر تبیینی از توزیع نرمال استاندارد تولید شدند. برای تولید موقعیت‌های مکانی، از روش پیشنهادی فودرینگهام و همکاران (۲۰۱۷) و لی و همکاران (۲۰۲۰) استفاده شد. به‌طور مشخص، مختصات موقعیت‌های مکانی برای n مشاهده با تعریف مختصات

$$u_i = ۱۲٫۵ + ۱۲٫۵\sqrt{r_i}\cos(\theta_i); \quad v_i = ۱۲٫۵ + ۱۲٫۵\sqrt{r_i}\sin(\theta_i)$$

تولید شدند، که در آن r_iها از توزیع یکنواخت استاندارد و θ_iها از توزیع یکنواخت در فاصله $[0, 2\pi]$ تولید می‌شوند. مقدار واقعی ضریب رگرسیونی β برای ۴۵ مشاهده اول برابر ۱ و برای ۵۵ مشاهده دوم برابر ۳ انتخاب شد. به عبارتی مشاهدات به دو خوشه متفاوت اختصاص داده شدند. ماتریس وزن‌های فضایی $\boldsymbol{W}$ با روش k–نزدیک‌ترین همسایگی با شعاع $k = ۱۰$ محاسبه شد و برای ارزیابی شدت وابستگی فضایی در عملکرد روش پیشنهادی، برای پارامتر خودهمبستگی فضایی سه مقدار واقعی مختلف $\rho_{\text{lag}} = ۰٫۳, ۰٫۵, ۰٫۷$ تعیین شدند که به‌ترتیب نشان‌دهنده وابستگی فضایی ضعیف، متوسط و قوی هستند. مقدار واقعی واریانس جمله خطا نیز برابر ۰٫۵ انتخاب شد. برای اجرای روش خوشه‌بندی فضایی پیشنهادی، از یک تابع هسته گاوسی با پهنای باند بهینه انتخاب‌شده بر اساس معیار معرفی‌شده در (۴) استفاده شد. برای به‌روزرسانی وزن‌ها نیز از مقادیر $\omega = ۰٫۰۰۰۱$ برای وزن اولیه، $\tau = ۰٫۰۰۱$ برای پارامتر توزین و $\eta = ۰٫۵$ برای پارامتر حافظه استفاده شد. شبیه‌سازی به تعداد $R = ۱۰۰$ بار تکرار شد.

با اجرای الگوریتم خوشه‌بندی، بعد از همگرا شدن، مشاهدات بدون خطا به دو خوشه همگن خود شامل ۴۵ و ۵۵ مشاهده تقسیم شدند. این نتیجه نشان می‌دهد که روش پیشنهادی، در شرایط کنترل‌شده شبیه‌سازی، قادر به شناسایی خوشه‌های همگن با دقت بالا است. می‌توان انتظار داشت عملکرد مشابهی در موقعیت‌های عملی رخ دهد. با این حال، برای مجموعه داده‌های واقعی، به دلیل وجود چالش‌هایی مانند خطای اندازه‌گیری و پیچیدگی‌های بالقوه ساختارهای وابستگی و ناهمگنی فضایی، عملکرد روش خوشه‌بندی ممکن است تحت تاثیر قرارگرفته و همراه با خطا باشد. برازش مدل‌ها برای هر خوشه با روش INLA انجام شد و برای هر دو خوشه با پیروی از استراتژی انتخاب مدل که در بخش قبلی تشریح شد، مدل SLM مدل منتخب شد. شکل ۱ نمودار جعبه‌ای مقادیر برآوردشده پارامتر β در هر خوشه به همراه نمودارهای پهنه‌بندی فضایی اختلاف مقادیر واقعی و برآوردشده β در ناحیه شبیه‌سازی‌شده را به تفکیک مقادیر مختلف ρ_{lag} نمایش می‌دهد. تفاوت ناچیز بین مقادیر واقعی و برآوردشده ضریب رگرسیونی در نقاط مختلف مکانی و تعیین درست خوشه‌های هر کدام از مشاهدات، نشان می‌دهند روش پیشنهادی در شناسایی ساختار ناهمگنی ناشی از وجود خوشه‌های فضایی که ناشی از ناهمگنی در تابع روند است کاملا کارا است و رهیافت برازش مدل نیز برآوردهای مطلوبی ارایه می‌دهد. این نتیجه در هر سه حالت وابستگی ضعیف تا قوی پابرجاست که



شکل ۱. نمودار جعبه‌ای مقادیر برآورد β و تفاضل مقادیر واقعی β با $\hat{\beta}$ به ازای $\rho_{\text{lag}} = ۰/۳$ (الف–ب)، $\rho_{\text{lag}} = ۰/۵$ (ج–د)، $\rho_{\text{lag}} = ۰/۷$ (ه–و)

برای مدل‌بندی هم‌زمان کارای وابستگی و ناهمگنی فضایی بسیار مهم است. جدول ۱ نیز مقادیر متوسط برآوردهای ρ_{lag} و دو معیار متوسط میانگین خطای مطلق $\text{MAB}_{\boldsymbol{\beta}}$ و متوسط میانگین توان‌های دوم خطا $\text{MMSE}_{\boldsymbol{\beta}}$ را که به صورت

$$\text{MAB}_{\boldsymbol{\beta}} = \frac{۱}{R}\sum_{r=۱}^{R}\frac{۱}{n}\sum_{i=۱}^{n}|\hat{\beta}_{r,i} - \beta_r|, \quad \text{MMSE}_{\boldsymbol{\beta}} = \frac{۱}{R}\sum_{r=۱}^{R}\frac{۱}{n}\sum_{i=۱}^{n}(\hat{\beta}_{r,i} - \beta_r)^{۲}$$

تعریف می‌شوند، نشان می‌دهد. نتایج گزارش‌شده در جدول ۱ دقت خوب برآوردها را، هم‌راستا با نتایج شکل ۱، تایید می‌کنند. برای ارزیابی عملکرد مدل در برآورد ضرایب رگرسیونی در هر دو خوشه به‌طور همزمان، مقادیر معیارها در جدول با میانگین گرفتن هر دو مقدار محاسبه‌شده معیارها به تفکیک دو خوشه گزارش شده‌اند. بر اساس مقادیر جدول، می‌توان کارآمدی رهیافت پیشنهادی را در مدل‌بندی داده‌های ناهمگن فضایی، به‌ویژه در موقعیت‌هایی که وابستگی فضایی متوسط و قوی وجود دارد، نتیجه گرفت. شناسایی و برآورد کارای ساختارهای ناهمگنی و وابستگی فضایی برای تحلیل دقیق داده‌ها بسیار مهم است و روش پیشنهادی این مقاله، در مواردی که ساختار ناهمگنی به صورت خوشه‌ای است، موفق عمل می‌کند. در جدول ۱ متوسط زمان برازش مدل در هر دو خوشه نیز (بر حسب ثانیه) گزارش شده است. سرعت بالای برازش مدل‌ها با استفاده از روش INLA در مقایسه با زمان‌های برازش چند ساعته با الگوریتم‌های MCMC کاملا چشمگیر و جذاب است.

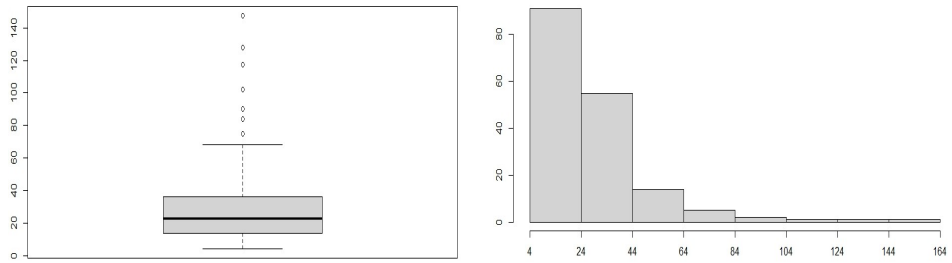
جدول ۱. مقادیر معیارهای برآورد ضریب رگرسیون، متوسط برآوردهای پارامتر خودهمبستگی فضایی در هر دو خوشه به ازای مقادیر مختلف ρ_{lag} و طول زمان برازش مدل

معیار	$\rho_{\text{lag}} = ۰/۳$	$\rho_{\text{lag}} = ۰/۵$	$\rho_{\text{lag}} = ۰/۷$
MAB_β	۰/۱۴۱	۰/۱۰۲	۰/۱۴۸
MMSE_β	۰/۳۰۱	۰/۱۰۰	۰/۱۲۸
$\hat{\rho}$	۰/۲۵۱	۰/۴۰۲	۰/۶۷۶
متوسط زمان اجرا	۳۰	۳۵	۳۲

۴ تحلیل قیمت مسکن در شهر مشهد

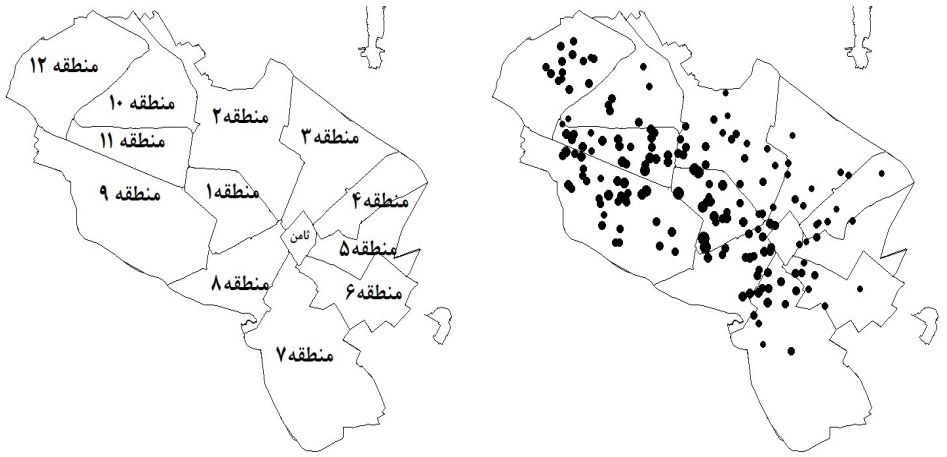
مشهد دومین شهر پرجمعیت ایران است و در شمال شرقی کشور قرار دارد. مجموعه داده قیمت مسکن برای مشهد در وب‌سایت خرید و فروش مسکن https://divar.ir/s/mashhad/buy-residential موجود است. در این مجموعه داده، واحدهای مسکونی بر اساس شرایطشان توسط کارشناسان بانک مسکن رتبه‌بندی و سپس مختصات آن‌ها، شامل طول و عرض جغرافیایی، بر اساس آدرس ثبت‌شده تعیین شده‌اند. شکل ۲ نمودارهای بافتنگار و جعبه‌ای قیمت مسکن را نمایش می‌دهد. با توجه به نمایش توزیع قیمت‌ها در بازه‌های مختلف و نمایش پراکندگی و میانه قیمت بازار مسکن (نمودار جعبه‌ای)، یک درک مقدماتی از وضعیت کلی بازار مسکن مشهد قابل دریافت است. شکل ۳ توزیع مکانی قیمت مسکن را در سیزده منطقه شهرداری مشهد به‌صورت نقطه‌ای بر روی نقشه شهر نمایش می‌دهد. اندازه هر نقطه متناسب با قیمت مسکن در آن منطقه است. این شکل تأکید می‌کند که نمونه‌برداری از همه مناطق شهرداری مشهد انجام شده است و داده‌های این مجموعه برای تحلیل کلان بازار مسکن شهر مشهد می‌تواند موثر باشد.

در این مجموعه داده، برای هر واحد مسکونی، متغیر پاسخ قیمت واحد مسکونی به میلیارد ریال است و متغیرهای تبیینی عبارتند از ۱) متراژ واحد مسکونی به متر مربع، ۲) متغیر دودویی سیستم مدیریت ساختمان برای



(ب) (الف)

شکل ۲. نمودارهای بافتنگار (الف) و جعبه‌ای (ب) داده‌های قیمت مسکن شهر مشهد



(ب) (الف)

شکل ۳. پراکندگی جغرافیایی داده‌های قیمت مسکن (الف) و مناطق ۱۳ گانه شهرداری (ب) شهر مشهد

کنترل و نظارت تجهیزاتی مانند روشنایی، با دو سطح ۱ برای خانه‌های دارای سیستم مدیریت ساختمان و ۰ برای خانه‌های فاقد این سیستم، ۳) تعداد طبقات مجتمع مسکونی، ۴) تعداد آسانسور، ۵) متغیر دودویی باغ بام[1] با دو سطح ۱ برای واحدهایی که بر روی پشت‌بام آن‌ها فضای سبز یا یک فضای تفریحی وجود دارد و ۰ برای واحدهای بدون این امکان، ۶) تعداد تراس، ۷) تعداد حمام، ۸) سن بنای ساختمان بر حسب سال از زمان صدور پروانه ساخت، ۹) تعداد کل آپارتمان‌های موجود در مجتمع مسکونی، ۱۰) متغیر رسته‌ای نوع سیستم سرمایش با سه سطح ۱ برای واحدهای دارای کولر آبی، ۲ برای کولر گازی و ۳ برای هر دو نوع، و ۱۱) متغیر رسته‌ای نوع سیستم گرمایش با سه سطح ۱ برای واحدهای دارای بخاری، ۲ برای پکیج و ۳ برای هر دو نوع. تعیین قیمت نیازمند

[1] Roof Garden

بررسی دقیق و ارزیابی جامع ویژگی‌های هر واحد مسکونی است. قیمت‌های هر واحد ابتدا توسط فروشندگان و بر اساس ویژگی‌های ملک تعیین می‌شود. سپس این قیمت‌ها توسط کارشناسان بانک مسکن بررسی و در صورت نیاز اصلاح می‌شوند تا یک رویه استاندارد برای تعیین قیمت برقرار باشد. این فرآیند باعث افزایش دقت و اعتبار داده‌ها می‌شود. مجموعه داده مورد استفاده در این پژوهش از اطلاعات اصلاح‌شده موجود در بانک مسکن استخراج شده است. در انتخاب واحدهای مسکونی در این مجموعه داده، تلاش شده است که نمونه‌ای نماینده از همه سیزده منطقه شهرداری مشهد موجود باشد. واحدهای مسکونی منتخب از یک چارچوب داده نمونه‌گیری شده است که شامل تمامی معاملات ثبت‌شده در وب‌سایت‌های خرید و فروش مسکن است. از آن‌جایی که بررسی و استانداردسازی قیمت‌ها توسط کارشناسان فرآیندی پیچیده و زمان‌بر است، تنها برای ۱۷۰ واحد این فرآیند تکمیل و اطلاعات حاصل ثبت شده است. بنابراین مجموعه داده مورد استفاده در این مقاله شامل همین ۱۷۰ واحد است.

برای ارزیابی وجود خودهمبستگی فضایی در داده‌ها، از آزمون‌های موران I (موران، ۱۹۴۸)، ضریب لاگرانژ خطا و ضریب لاگرانژ تأخیر استفاده شد. بر اساس نتایج آزمون‌های خودهمبستگی در جدول ۲، وجود خودهمبستگی فضایی در داده‌ها با هر سه آزمون تایید می‌شود. برای بررسی وجود ناهمگنی فضایی در حضور خودهمبستگی فضایی،

جدول ۲. نتایج آزمون‌های خودهمبستگی فضایی برای داده‌های قیمت مسکن

آزمون‌های خودهمبستگی فضایی	آماره	سطح معنی‌داری
موران I	۳٫۲۶	۰٫۰۰۰۶
ضریب لاگرانژ خطا	۶٫۶۳	۰٫۰۱۰۰
ضریب لاگرانژ تأخیر	۲۲٫۷۹	$<$ ۰٫۰۰۰۶

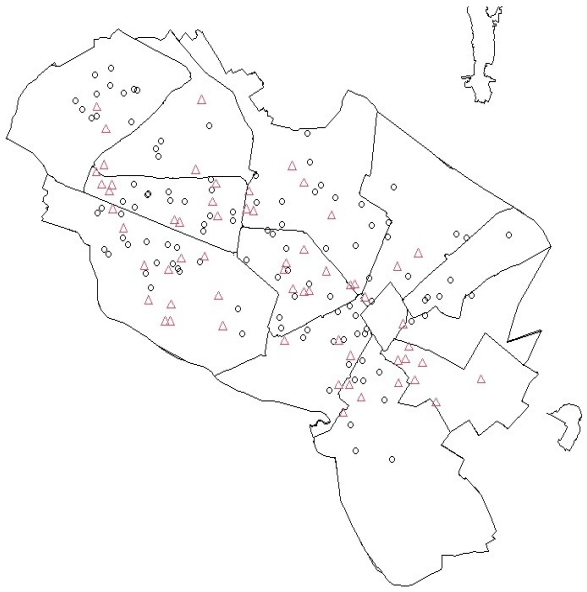
از آزمون بریچ–پیگان استفاده شد. نتایج این آزمون در جدول ۳، معنی‌داری وجود ناهمگنی فضایی را تایید می‌کنند. پس از تایید وجود ناهمگنی فضایی و به منظور مقایسه با روش گومز و همکاران (۲۰۲۱)، مدل‌بندی بدون در نظر

جدول ۳. نتایج آزمون ناهمگنی فضایی بریچ–پیگان برای داده‌های قیمت مسکن

مدل‌های فضایی	آماره بریچ–پیگان	درجه آزادی	سطح معنی‌داری
SLM	۵۶٫۰۳	۱۱	$<$ ۰٫۰۰۰۱
SEM	۵۴٫۹۱	۱۱	$<$ ۰٫۰۰۰۱
SDM	۶۴٫۲۱	۲۲	$<$ ۰٫۰۰۰۱
SAC	۵۵٫۵۱	۱۱	$<$ ۰٫۰۰۰۱
SDEM	۶۰٫۲۳	۲۲	$<$ ۰٫۰۰۰۱
SLX	۶۳٫۶۷	۲۲	$<$ ۰٫۰۰۰۱

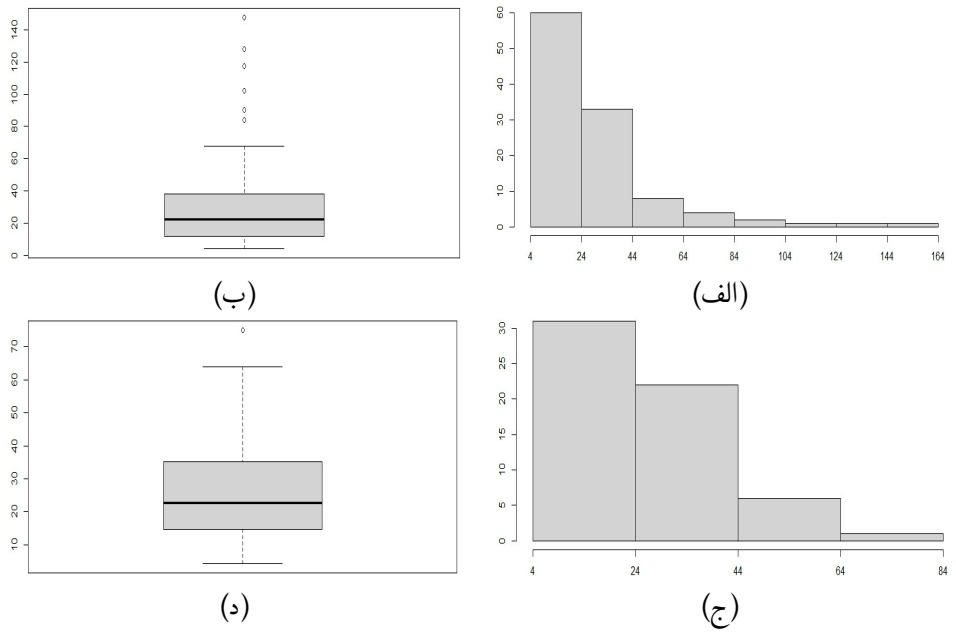
گرفتن خوشه‌بندی انجام شد. نتایج نشان داد در اکثر مدل‌ها، ناهمگنی فضایی با تغییرات قابل توجه در ضرایب مشاهدات مشاهده می‌شود. به‌عنوان دو شاهد، برآورد ضریب رگرسیونی متغیر تبیینی تعداد تراس برای ۱۷۰ خانه از ۰٫۵− تا ۱۵ متغیر است. برآورد ضریب متغیر تبیینی تعداد حمام نیز در بازه (۰٫۷۵, ۱۱) تغییر می‌کند. این تغییرات نشان‌دهنده ناهمگنی فضایی در پارامترهای مدل است، که ناشی از ثابت نبودن رابطه بین متغیر پاسخ و متغیرهای

تبیینی در نقاط مختلف ناحیه تحت مطالعه است. با تایید وجود ناهمگنی فضایی در حضور همبستگی فضایی، رویکرد پیشنهادی در این مقاله برای مدل‌بندی ناهمگنی، در قالب رهیافت خوشه‌بندی، به کار گرفته شد. برای تعیین وزن‌های اولیه در رگرسیون وزنی جغرافیایی، از تابع هسته گاوسی با پهنای باند سازوار استفاده شد که مقدار بهینه آن بر اساس معیار AIC_c عدد ۳۵ به دست آمد. الگوریتم به‌روزرسانی وزن‌ها با تعیین پارامترهای وزن‌دهی $\tau = ۰/۰۰۱$ و حافظه $\eta = ۰/۵$ و $\omega = ۰/۰۰۱$ پس از ۹۴ تکرار به همگرایی رسید و ۱۷۰ مشاهده به ۲ خوشه همگن با حجم‌های $n_۱ = ۱۱۰$ و $n_۲ = ۶۰$ دسته‌بندی شدند. موقعیت مشاهدات هر دو خوشه در شکل ۴ با دو نماد، به ترتیب، ○ و △ نمایش داده شده‌اند. در شکل ۵ نیز نمودارهای جعبه‌ای و بافتنگار مشاهدات پاسخ به تفکیک هر خوشه نمایش داده شده‌اند. تمایز قیمت‌های مسکن در دو خوشه از روی شکل ۵ مشهود است. قیمت واحدهایی که در خوشه اول قرار گرفته‌اند تمایل به مقادیر بزرگتر از قیمت واحدهای خوشه دوم دارند.



شکل ۴. پراکندگی جغرافیایی و خوشه‌های داده‌های قیمت مسکن، نقاط ○: خوشه فضایی اول و نقاط Δ: خوشه فضایی دوم.

برای برازش مدل‌های مختلف با روش INLA، ابتدا باید ماتریس‌های همسایگی $\boldsymbol{W}$ و $\boldsymbol{M}$ تعیین شوند. برای این مجموعه داده، هر دوی این ماتریس‌ها یکسان در نظر گرفته شدند. برای شناسایی ماتریس $\boldsymbol{W}$ روش نزدیک‌ترین همسایه‌ها به کار گرفته شد. برای اجرای این روش ابتدا باید تعداد بهینه همسایه‌ها (K) معلوم شود. در اغلب متون این تعداد عددی بین ۵ تا ۳۵ انتخاب شده است. در این مقاله، تعیین K بهینه توسط معیارهای DIC و لگاریتم درستنمایی کناری انجام شد. برای این کار، از مدل پایه SLM استفاده شد. به‌طور مشخص، مدل SLM با ماتریس‌های همسایگی مختلف که با روش نزدیک‌ترین همسایه‌ها و مقادیر مختلف K به دست آمده بودند، به



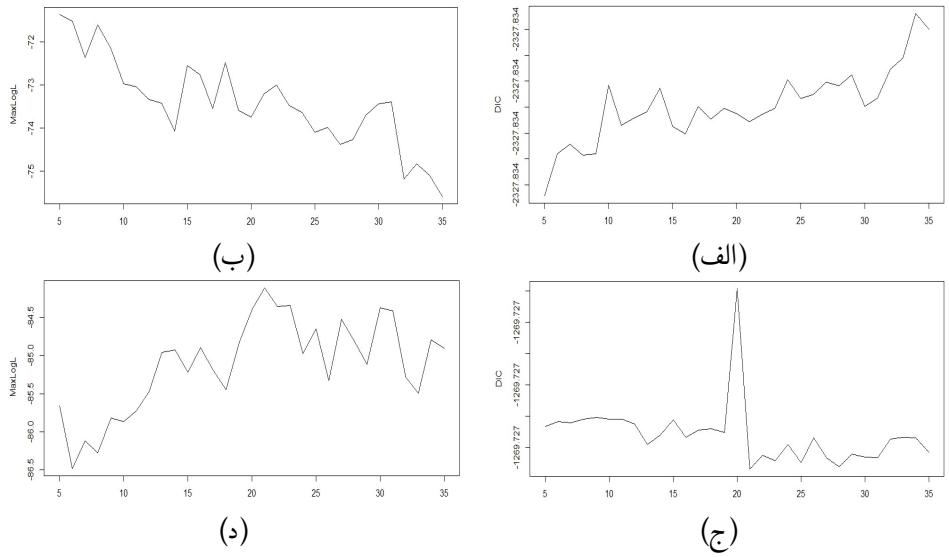
شکل ۵. نمودارهای بافتنگار و جعبه‌ای قیمت مسکن خوشه اول (الف و ب) و خوشه دوم (ج و د)

داده‌های هر خوشه برازش داده شد. سپس مقدار Kی که معیار DIC مدل متناظر آن کمترین یا لگاریتم درستنمایی کناری ماکسیمم‌شده مدل متناظر آن بیشترین بود انتخاب شد. برای مجموعه داده‌های هر دو خوشه هر دو معیار ذکرشده عدد یکسانی را برای K بهینه اعلام کردند. نتایج اجرای این فرآیند در شکل ۶ نمایش داده شده است. بر اساس این نتایج، برای هر دو خوشه ۱ و ۲، به ترتیب، تعداد بهینه همسایه‌ها برابر ۵ و ۲۱ تعیین و ماتریس‌های $\boldsymbol{W}$ متناظر ساخته شدند. با برازش مدل‌های مختلف رقیب، بر اساس معیار ماکسیمم لگاریتم درستنمایی کناری (جدول ۴) و با توجه به لحاظ کردن پیچیدگی مدل‌ها (بر اساس درجه آزادی گزارش‌شده در جدول ۳)، در هر دو خوشه، مدل خطای فضایی، SEM، به عنوان مدل برتر انتخاب شد. دلیل انتخاب این مدل، نزدیکی مقادیر معیار انتخاب به مقادیر بهترین مدل برازش‌شده یعنی SDM و سادگی آن (با درجه آزادی ۱۱) نسبت به مدل SDM (با درجه آزادی ۲۲) است.

جدول ۴. ماکسیمم لگاریتم درستنمایی کناری مدل‌های رقیب خوشه‌ها برای داده‌های قیمت مسکن

خوشه	SLX	SDEM	SDM	SLM	SEM	SAC
۱	−۱۱۷٫۰۱	−۱۱۵٫۳۲	−۴۳٫۳۸	−۷۱٫۶۸	−۵۸٫۹۸	−۷۱٫۶۸
۲	−۱۱۴٫۱۱	−۱۱۴٫۴۰	−۴۲٫۰۹	−۸۵٫۲۹	−۶۶٫۷۱	−۸۵٫۵۱

نتایج برازش مدل منتخب SEM برای هر دو خوشه در جدول ۵ و شکل ۷ گزارش شده‌اند. در جدول ۵ برآوردهای



شکل ۶. مقادیر DIC و ماکسیمم لگاریتم درستنمایی کناری برای خوشه‌های فضایی داده‌های قیمت مسکن

پسینی ضرایب رگرسیونی متغیرهای تبیینی و فواصل اعتبار ۹۵ درصد متناظر آن‌ها گزارش شده‌اند و در شکل ۷ توابع چگالی پسین کناری پارامترهای خودهمبستگی در مدل SEM به تفکیک هر خوشه نمایش داده شده‌اند. در ادامه، تحلیل نتایج حاصل بر اساس اثرات معنی‌دار بیان شده‌اند. در هر دو خوشه مورد مطالعه، تعداد تراس‌ها با اندازه اثری مثبت و قابل توجه روی افزایش قیمت مسکن به‌طور معنی‌داری تاثیر دارد. این نتیجه با طبیعت بازار سازگار و به دلیل توجه ویژه‌ای است که ساکنان به داشتن تراس‌ها دارند، به‌ویژه زمانی که فضای باز محدودی مانند حیاط در اختیار ندارند. متغیر متراژ با اندازه اثری مثبت روی افزایش قیمت مسکن به‌طور معنی‌داری تأثیر دارد. اندازه اثر ۰٫۰۰۸ برای خوشه اول و ۰٫۰۱۱ برای خوشه دوم ممکن است کوچک به نظر برسند و برخلاف معنی‌داری این متغیر، تفسیری بر تاثیر ناچیز و قابل نادیده گرفتن اثر متراژ واحد مسکونی بر قیمت آن ارایه شود. در پاسخ به این ابهام، باید به مقیاس اندازه‌گیری این متغیر (بر حسب متر مربع) توجه کرد که مقادیر عددی آن در مقایسه با سایر متغیرهای تبیینی حاضر در مدل چندین برابر بزرگ‌تر است. این توجه می‌تواند این واقعیت را روشن کند که متراژ واحد مسکونی یک عامل اثرگذار جدی در تعیین قیمت واحدهای مسکونی است. یکی از نتایج جالب و تا حدودی متناقض، اثر معنی‌دار منفی متغیر تعداد حمام‌ها است. اگرچه در اکثر مناطق، تعداد حمام‌ها یکی از عوامل مهم و تاثیرگذار در تعیین قیمت مسکن است، اما در این دو خوشه از داده‌های مشهد حتی تاثیر منفی داشته است. این نتیجه می‌تواند به دلیل ترجیح ساکنان شهر مشهد برای داشتن اتاق‌های بزرگ‌تر به جای حمام‌های اضافی باشد. در هر دو خوشه، سیستم مدیریت هوشمند به دلیل استفاده بهینه از انرژی در تعیین قیمت مسکن معنی‌دار بوده و به نظر می‌رسد ساکنین توجه ویژه‌ای به آن دارند. تعداد آسانسور در ساختمان‌ها نیز یکی دیگر از عوامل معنی‌دار با اثر

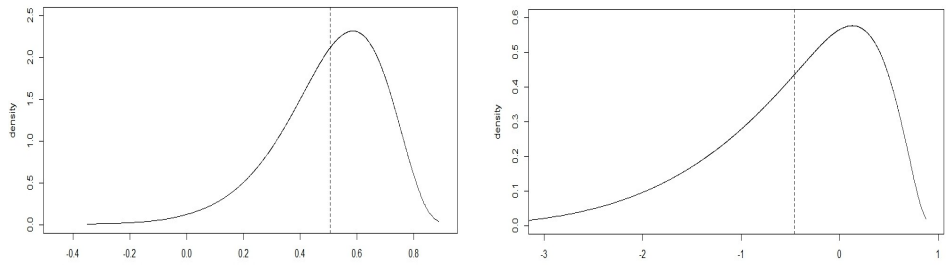
مثبت روی قیمت مسکن است که بسیاری از افراد در زمان تصمیم‌گیری برای خرید ملک به آن توجه می‌کنند. در شهر مشهد، همانند سایر شهرها، داشتن آسانسور نقش مهمی در تعیین قیمت مسکن ایفا می‌کند. متغیرهای تبیینی باغ‌بام و سن ساختمان در خوشه اول تأثیر معنی‌داری بر قیمت مسکن ندارند، اما در خوشه دوم نتیجه متفاوت است. بررسی توزیع متغیر سن ساختمان در دو خوشه نشان می‌دهد که در خوشه اول، سن ساختمان پراکندگی کمتری دارد. این واقعیت می‌تواند منبع اصلی معنی‌داری نشدن سن ساختمان در این خوشه باشد. در صورت وجود پراکندگی کوچک مشاهدات در یک متغیر تبیینی، ممکن است مدل قادر به شناسایی اثرات آن بر متغیر پاسخ نباشد. در مقابل، در خوشه دوم که تنوع بیشتری در سن ساختمان مشاهده می‌شود، اثر این متغیر بر قیمت مسکن مشخص و معنی‌دار شده است. این یافته نشان می‌دهد که میزان تأثیر متغیرها می‌تواند بسته به توزیع آن‌ها در هر خوشه متفاوت باشد. درک این تفاوت‌ها برای تحلیل دقیق‌تر بازار مسکن ضروری است. در خوشه دوم، وجود باغ‌بام احتمالاً به دلیل افزایش جذابیت بصری، فضای سبز و آرامش بیشتر برای ساکنان، تأثیر مثبت و معنی‌داری بر قیمت مسکن داشته است. همچنین، سن ساختمان در خوشه دوم نشان‌دهنده اهمیت ساختمان‌های جدیدتر با امکانات به‌روزتر و نگهداری بهتر است، که مورد توجه ساکنان این خوشه قرار گرفته است. تعداد طبقات، تعداد واحدهای مجتمع، و نوع سیستم سرمایش در هر دو خوشه معنی‌دار نیستند. به عبارت دیگر، ارتفاع ساختمان (تعداد طبقات) به‌عنوان یک ویژگی تأثیرگذار در تصمیم‌گیری خریداران به چشم نیامده است. احتمالاً متغیرهای دیگری مانند امکانات داخلی یا دسترسی به آسانسور نقش بیشتری دارند. از طرفی با وجود این‌که افزایش تعداد واحدها می‌تواند به کاهش سهم زمین هر مالک منجر شود، این متغیر در این دو خوشه تأثیر محسوسی بر قیمت مسکن نشان نداده است. احتمال توجه خریداران در این مناطق به سایر ویژگی‌های مسکن، مانند طراحی داخلی یا امکانات مشترک، بیشتر است. از طرف دیگر، به نظر می‌رسد در شهر مشهد، سیستم سرمایش در اکثر ساختمان‌ها تعبیه شده است و تفاوت زیادی بین ساختمان‌ها در این زمینه وجود ندارد. به همین دلیل، این متغیر در مدل‌های هر دو خوشه معنی‌دار نیست. در مقابل نوع سیستم گرمایشی با اثری مثبت روی قیمت مسکن کاملا معنی‌دار است. دلیل آن هم می‌تواند خطرهای استفاده از بخاری در مقایسه با پکیج باشد.

شکل ۷ نشان می‌دهد که نوع وابستگی فضایی در دو خوشه رفتارهای کاملا متفاوتی دارند. ضریب همبستگی مشاهدات در مدل SEM برای هر دو خوشه اگرچه از توزیع مشابهی پیرو می‌کند ولی برآورد میانگین پسینی آن در خوشه اول مثبت و در خوشه دوم منفی است. این نتیجه می‌تواند تا حدودی دلیل تفاوت نتایج تحلیل اثرات متغیرهای تبیینی در دو خوشه را تشریح کند.

در مجموع، این نتایج نشان می‌دهند که اولویت‌ها و انتظارات ساکنان هر خوشه در رابطه با ویژگی‌های مسکن متفاوت است و تأثیرگذاری متغیرها به نیازها و سلیقه‌های خاص هر گروه بستگی دارد و توجه به آن‌ها می‌تواند به‌عنوان یک نشانگر ارتقاء و پیشرفت در ساخت و سازهای شهری مشهد مطرح شود و نقش مهمی در جذب سرمایه‌گذاری‌های مسکن و توسعه شهری ایفا کند.

جدول ۵. برآورد پسینی ضرایب رگرسیونی مدل SEM به تفکیک خوشه‌ها

ضرایب	خوشه ۱			خوشه ۲		
	میانگین	خطای معیار	فواصل اعتبار ۹۵%	میانگین	خطای معیار	فواصل اعتبار ۹۵%
عرض از مبداء	۱٫۵۲۱	۰٫۱۵۴	(۱٫۲۲۱ ، ۱٫۸۲۴)	۰٫۹۸۰	۰٫۲۰۵	(۰٫۵۸۰ ، ۱٫۳۸۴)
متراژ	۰٫۰۰۸	۰٫۰۰۱	(۰٫۰۰۷ ، ۰٫۰۰۹)	۰٫۰۱۱	۰٫۰۰۱	(۰٫۰۰۹ ، ۰٫۰۱۲)
سیستم مدیریت ساختمان	۰٫۱۸۹	۰٫۰۷۲	(۰٫۰۴۸ ، ۰٫۳۳۲)	۰٫۱۹۰	۰٫۰۸۵	(۰٫۰۲۵ ، ۰٫۳۵۹)
تعداد طبقات	−۰٫۰۲۳	۰٫۰۱۹	(−۰٫۰۶۱ ، ۰٫۰۱۶)	۰٫۰۳۱	۰٫۰۲۸	(−۰٫۰۲۳ ، ۰٫۰۸۵)
تعداد آسانسور	۰٫۲۲۷	۰٫۰۵۳	(۰٫۱۲۴ ، ۰٫۳۳۴)	۰٫۱۷۰	۰٫۰۷۸	(۰٫۰۱۸ ، ۰٫۳۲۵)
باغ بام	−۰٫۰۴۸	۰٫۰۵۲	(−۰٫۰۵۳ ، ۰٫۱۵۰)	۰٫۱۴۵	۰٫۰۷۲	(۰٫۰۰۵ ، ۰٫۲۸۹)
تعداد تراس	۰٫۳۲۶	۰٫۰۴۶	(۰٫۲۳۷ ، ۰٫۴۱۸)	۰٫۴۰۷	۰٫۰۶۸	(۰٫۲۷۳ ، ۰٫۵۴۰)
تعداد حمام	−۰٫۰۹۹	۰٫۰۴۰	(−۰٫۱۷۸ ، ۰٫۰۲۲)	−۰٫۱۹۷	۰٫۰۷۳	(−۰٫۳۴۱ ، ۰٫۰۵۳)
سن ساختمان	−۰٫۰۰۲	۰٫۰۰۵	(−۰٫۰۰۸ ، ۰٫۰۱۲)	۰٫۰۱۵	۰٫۰۰۷	(۰٫۰۰۲ ، ۰٫۰۲۹)
تعداد واحدهای مجتمع	−۰٫۰۰۷	۰٫۰۰۵	(−۰٫۰۱۶ ، ۰٫۰۰۲)	−۰٫۰۰۳	۰٫۰۰۹	(−۰٫۰۲۱ ، ۰٫۰۱۵)
سیستم سرمایش	−۰٫۰۰۱	۰٫۰۴۱	(−۰٫۰۸۳ ، ۰٫۰۷۸)	−۰٫۰۰۴	۰٫۰۴۱	(−۰٫۰۸۵ ، ۰٫۰۷۶)
سیستم گرمایش	۰٫۱۰۷	۰٫۰۴۸	(۰٫۰۱۴ ، ۰٫۲۰۳)	۰٫۰۹۹	۰٫۰۵۸	(۰٫۰۱۴ ، ۰٫۲۱۴)



(ب) (الف)

شکل ۷. چگالی‌های پسین کناری پارامترهای خودهمبستگی فضایی با میانگین پسینی (خطوط عمودی) در مدل خطاهای فضایی برای خوشه ۱ (الف) و خوشه ۲ (ب).

بحث و نتیجه‌گیری

در این مقاله، یک رهیافت دومرحله‌ای برای مدل‌بندی و تحلیل داده‌هایی که همزمان ناهمگنی و همبستگی فضایی دارند معرفی شد. در گام اول، خوشه‌بندی فضایی با ترکیب روش رگرسیون موزون جغرافیایی و الگوریتم هموارسازی وزن‌های سازوار انجام شد. این روش به دلیل توانایی آن در گروه‌بندی و شناسایی نواحی همگن فضایی، امکان بهبود دقت تحلیل‌های آماری را فراهم می‌کند. در گام دوم، برای هر خوشه شناسایی‌شده، مدل‌های اقتصادسنجی فضایی بیزی با استفاده از روش INLA برازش داده شدند. استفاده از روش INLA زمان محاسباتی برای تحلیل‌های بیزی را به‌طور چشمگیری کاهش می‌دهد.

دستاورد کلیدی این پژوهش، ترکیب مؤثر خوشه‌بندی فضایی و مدل‌بندی بیزی است که از طریق آن، روش کلاسیک پیشنهادی بیله و همکاران (۲۰۱۷)، در بخش برازش مدل‌ها، به نسخه بیزی تعمیم داده شد و عدم توانایی

رهیافت گومز و همکاران (۲۰۲۱) برای شناسایی و وارد کردن ناهمگنی خوشه‌ای فضایی پوشش داده شد. برخلاف دو مطالعه ذکرشده، رهیافت پیشنهادی در ابتدا داده‌ها را به خوشه‌های همگن تفکیک کرده و سپس تحلیل را بر اساس این خوشه‌ها انجام می‌دهد. این کار نه‌تنها منجر به شناسایی کارای روابط بین متغیرها می‌شود بلکه به کاهش اریبی و افزایش دقت ضرایب رگرسیونی و تحلیل‌ها کمک می‌کند. در مطالعه موردی، نتایج حاصل از مدل‌بندی داده‌های قیمت مسکن در شهر مشهد نشان داد که رهیافت پیشنهادی قادر است همبستگی‌های فضایی و ناهمگنی‌های موجود در داده‌ها را با دقت خوب شناسایی و تحلیل کند. به‌عنوان مثال، خوشه‌بندی فضایی انجام‌شده نشان داد که عوامل تأثیرگذار بر قیمت مسکن در هر خوشه، ممکن است متفاوت از خوشه‌های دیگر باشد. این دست‌آورد می‌تواند در تحلیل‌های سیاست‌گذاری منطقه‌ای و برنامه‌ریزی شهری نیز بسیار مفید باشد.

از منظر آینده تحقیق، تعمیم رهیافت پیشنهادی به روش‌های پیشرفته‌تر خوشه‌بندی نظیر رگرسیون موزون جغرافیایی چندمقیاسی[1] که پهنای باند متفاوتی برای متغیرهای تبیینی در نظر می‌گیرد، یک مسیر جذاب و کاربردی است. همچنین، توسعه نسخه بیزی خوشه‌بندی فضایی با استفاده از رگرسیون موزون جغرافیایی چندمقیاسی، می‌تواند گامی دیگر در جهت بهبود تحلیل‌های فضایی باشد. این مسیرهای تحقیقاتی می‌توانند از منظر دقت بیشتر در تحلیل داده‌های فضایی و کاهش هزینه محاسباتی، بسیار ارزشمند باشند.

تقدیر و تشکر

نویسندگان مقاله از سه داور گرامی، سردبیر محترم و ویراستار مجله که نظرات آن‌ها باعث بهبود چشم‌گیر و ارائه بهتر مقاله شد، قدردانی می‌کنند.

مراجع

رسولی، ح. (۱۳۹۰)، مدل‌های اتورگرسیو فضایی و تحلیل داده‌های معاملات مسکونی شهر تهران، مجله علوم آماری، ۵، ۲۰۲-۱۸۹.

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

Akaike, H. (1976). An Information Criterion (AIC), *Mathematical Sciences*, **14**, 5–7.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Vol. 4, Springer Series, London.

[1] Multiscale GWR

Anselin, L. (1990). Spatial Dependence and Spatial Structural Instability in Applied Regression Analysis, *Journal of Regional Science*, **30**(2), 185–207.

Bille, A. G., Benedetti, R., and Postiglione, P. (2017). A Two-step Approach to Account for Unobserved Spatial Heterogeneity, *Spatial Economic Analysis*, **4**, 81–91.

Bowman, A. W. (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimates, *Biometrika*, **71**(2), 353–360.

Breusch, T. S., and Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation, *Econometrica: Journal of the Econometric Society*, **47**(5), 1287–1294.

Cleveland, W. S. (1979). Robust Locally Weighted Regression on Smoothing Scatterplots, *Journal of the American Statistical Association*, **74**(368), 829–836.

Cressie, N. A. (1993). *Statistics for Spatial Data*, Wiley Series in Probability and Mathematical Statistics, New York.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, **1**(2), 209–230.

Fotheringham, A. S., Brundson, C., and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Chichester, UK: Wiley.

Fotheringham, A. S., Yang, W., and Kang, W. (2017). Multiscale Geographically Weighted Regression (MGWR), *Annals of the American Association of Geographers*, **107**(6), 1247–1265.

Geniaux, G., and Martinetti, D. (2018). A New Method for Dealing Simultaneously with Spatial Autocorrelation and Spatial Heterogeneity in Regression Models, *Regional Science and Urban Economics*, **72**, 74–85.

Gomez-Rubio, V., Bivand, R. S., and Rue, H. (2021). Estimating Spatial Econometrics Models with Integrated Nested Laplace Approximation, *Mathematics*, **9**(17), 2044.

Hurvich, C. M., Simonoff, J. S., and Tsai, C. L. (1998). Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion, *Journal of the Royal Statistical Society, Series B*, **60**(2), 271–293.

Kelejian, H. H., and Robinson, D. P. (1998). A Suggested Test for Spatial Autocorrelation and/or Heteroskedasticity and Corresponding Monte Carlo Results, *Regional Science and Urban Economics*, **28**(4), 389–417.

Kelejian, H. H., and Prucha, I. R. (2010). Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances, *Journal of Econometrics*, **157**(1), 53–67.

Lawson, A. B., Choi, J., and Zhang, J. (2014). Prior Choice in Discrete Latent Modeling of Spatially Referenced Cancer Survival, *Statistical Methods in Medical Research*, **23**(2), 183–200.

LeSage, J. P., and Pace, R. K. (2009). *Introduction to Spatial Econometrics*, Chapman and Hall/CRC, London.

LeSage, J. P., Pace, K. R., Lam, N., Campanella, R., and Liu, X. (2011). New Orleans Business Recovery in the Aftermath of Hurricane Katrina, *Journal of the Royal Statistical Society, Series A*, **174**, 1007–1027.

Li, F., and Sang, H. (2019). Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets, *Journal of the American Statistical Association*, **114**(527), 1050–1062.

Li, Z., Fotheringham, A. S., Oshan, T. M., and Wolf, L. J. (2020). Measuring Bandwidth Uncertainty in Multiscale Geographically Weighted Regression Using Akaike Weights, *Annals of the American Association of Geographers*, **110**(5), 1500–1520.

Moran, P. A. (1948). Some Theorems on Time Series: The Significance of the Serial Correlation Coefficient, *Biometrika*, **35**, 255–260.

Mur, J., López, F., and Angulo, A. (2008). Symptoms of Instability in Models of Spatial Dependence, *Geographical Analysis*, **40**(2), 189–211.

Polzehl, J., and Spokoiny, V. (2000). Adaptive Weights Smoothing With Applications to Image Restoration, *Journal of the Royal Statistical Society, Series B*, **62**, 335–354.

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

Silverman, B. W. (1984). A Fast and Efficient Cross-Validation Method for Smoothing Parameter Choice in Spline Regression, *Journal of the American Statistical Association*, **79**, 584–589.

Stein, A. (2022). The Development of the Journal Spatial Statistics: The First 10 Years, *Spatial Statistics*, **50**, 100576.

Sugasawa, S. (2021). Grouped Heterogeneous Mixture Modeling for Clustered Data, *Journal of the American Statistical Association*, **116**(534), 999–1010.

Zhao, P., Yang, H. C., Dey, D. K., and Hu, G. (2023). Spatial Clustering Regression of Count Value Data via Bayesian Mixture of Finite Mixtures, *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3504–3512.

Zhihua, M., Yishu, X., and Guanyu, H. (2020). Heterogeneous Regression Models for Clusters of Spatial Dependent Data, *Spatial Economic Analysis*, **15**(4), 459–475.